UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ARIANE LAO BALDYKOWSKI
SANDRO ALBERTO MICZEVSKI JUNIOR
SAULO ALVES DE BRITO

MINERAÇÃO DE DADOS DO UNTAPPD E TWITTER: USANDO PREFERÊNCIAS POR CERVEJAS EM TOMADA DE DECISÃO E NO ESTUDO DE DIFERENÇAS CULTURAIS

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA

2017

ARIANE LAO BALDYKOWSKI
SANDRO ALBERTO MICZEVSKI JUNIOR
SAULO ALVES DE BRITO

MINERAÇÃO DE DADOS DO UNTAPPD E TWITTER: USANDO PREFERÊNCIAS POR CERVEJAS EM TOMADA DE DECISÃO E NO ESTUDO DE DIFERENÇAS CULTURAIS

Trabalho de Conclusão de Curso apresentado ao Departamento Acadêmico de Informática da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de "Bacharel em Sistemas de Informação".

Orientador: Thiago Henrique Silva

CURITIBA

2017

Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Câmpus Curitiba
Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Informática
Coordenação do Curso de Bacharelado em Sistemas de Informação

TERMO DE APROVAÇÃO

“MINERAÇÃO DE DADOS DO UNTAPPD E TWITTER: USANDO PREFERÊNCIAS POR CERVEJAS EM TOMADA DE DECISÃO E NO ESTUDO DE DIFERENÇAS CULTURAIS”

por

“ARIANE LAO BALDYKOWSKI
SANDRO ALBERTO MICZEVSKI JUNIOR
SAULO ALVES DE BRITO”

Este Trabalho de Conclusão de Curso foi apresentado no dia 17 de Novembro de 2017 como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação na Universidade Tecnológica Federal do Paraná - UTFPR - Câmpus Curitiba. O(a)(s) aluno(a)(s) foi(ram) arguido(a)(s) pelos membros da Banca de Avaliação abaixo assinados. Após deliberação a Banca de Avaliação considerou o trabalho _____.

_____ **Prof. Thiago Henrique Silva** (Presidente - UTFPR/Curitiba)	_____ **Prof. Alexandre Reis Graeml** (Avaliador 1 - UTFPR/Curitiba)
_____ **Prof. Luiz Celso Gomes Junior** (Avaliador 2 - UTFPR/Curitiba)	_____ **Prof. Leyza E. Baldo Dorini** (Professor Responsável pelo TCC – UTFPR/Curitiba)
_____ **Prof. Leonelo Dell Anhol Almeida** (Coordenador(a) do curso de Bacharelado em Sistemas de Informação – UTFPR/Curitiba)	

“A Folha de Aprovação assinada encontra-se na Coordenação do Curso.”

Aos nossos familiares e amigos que tornaram as pausas entre um parágrafo e outro momentos de incentivo para continuarmos produzindo.

Agradecimentos

Agradecemos a esta instituição pelo excelente ambiente oferecido aos seus alunos e os profissionais qualificados que disponibiliza para nos ensinar, proporcionando desafios, ampliando horizontes e fazendo com que sonhos viessem a se tornar realidade.

Ao professor que orientou este projeto, Thiago Henrique Silva, pelo estímulo, ensino, críticas, dedicação e confiança em nosso trabalho. Ele que sempre nos mostrou a melhor forma de fazer algo grande com coisas simples.

Aos demais professores, que durante estes quatro anos de graduação compartilharam conosco seus conhecimentos.

Nossos familiares, por não medirem esforços para que pudéssemos levar nossos estudos adiante, por todo o amor, incentivo e apoio incondicional.

Por fim, a todos aqueles que de alguma forma contribuíram com nosso desenvolvimento como estudante, profissional e como pessoa, e que torceram pela concretização deste projeto.

“Conhecimento não é aquilo que você sabe,
mas o que você faz com aquilo que você sabe.”

Aldous Huxley

Resumo

Redes sociais online disponibilizam vastas quantidades de dados que agora são facilmente obtidos para vários tipos de análises. Este estudo demonstra a capacidade das redes sociais baseadas em localização de servirem como base para tomadas de decisão e entendimento de preferências regionais. Foram coletados dados do Untappd, uma rede social para apreciadores de cervejas, através do Twitter; esses dados foram utilizados em processos de clusterização e comparados com o caso da possível implantação da "Rua da Cerveja" em Curitiba. É também demonstrado o potencial para explorar as preferências de cerveja no estudo do comportamento Social Urbano, particularmente relacionado à identificação automática de aspectos culturais, uma informação valiosa que pode permitir novos serviços.

Palavras-chaves: Coleta de dados, Redes sociais, Untappd, Twitter, Serviços baseado em localização.

Abstract

Online social networks, make available vast amounts of data which nowadays is easily obtained for various types of analyses. This study demonstrates the ability of location-based social networks to serve as a basis for decision-making and understanding of regional preferences. Data was collected from Untappd, a social network for beer lovers, through Twitter; this data was used in clustering processes and compared to the possible creation of a Craft Beer Street in Curitiba. It is also demonstrated the potential to explore beer preferences in the study of urban social behavior, particularly related to the automatic identification of cultural aspects, a valuable information that can enable new services.

Key-words: Data mining, Social networks, Untappd, Twitter, Location Based Services.

Lista de ilustrações

Lista de tabelas

Lista de Algoritmos

Lista de abreviaturas e siglas

PIB	Produto Interno Bruto
RSP	Rede de Sensoriamento Participativo
API	*Application Programming Interface*
JSON	*JavaScript Object Notation*
URL	*Uniform Resource Locator*
TCC	Trabalho de Conclusão de Curso
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
MST	*Minimum Spanning Tree*
KDD	*Knowledge Discovery in Databases*
IDE	*Integrated Development Environment*
SIG	Sistema de Informação Geográfica
SBRC	Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos
WebMedia	Simpósio Brasileiro de Sistemas Multimídia e Web
SBC	Sociedade Brasileira de Computação
MBR	*Minimum Bounding Rectangle*

Sumário

1 Introdução

A cerveja foi uma importante bebida na história da humanidade, sendo encontrados indícios de sua produção desde desenhos rupestres, passando pela Mesopotâmia, Idade média e até os dias de hoje com o aumento do interesse em cervejas artesanais. Essa importância pode ser explicada pela facilidade do seu processo de produção comparado com outras bebidas, preço mais acessível que o vinho, e pelo conhecimento empírico dessas sociedades que a cerveja era, em muitos casos, mais saudável que a água decorrente da presença de álcool. Os cervejeiros eram considerados importantes artesãos em muitas dessas sociedades, conforme relatado por Silva, Leite e Paula (2015).

No Brasil a bebida foi trazida por imigrantes europeus, e seu primeiro registro histórico foi um anúncio de venda de cerveja brasileira no Jornal do Comércio do Rio de Janeiro, de 27 de outubro de 1836. A primeira cervejaria nacional a produzir em escala industrial surgiu entre 1870 e 1880. Em 2013 o setor cervejeiro foi responsável por 2% do PIB brasileiro, produzindo 14 milhões de litros e gerando R$ 21 bilhões em impostos (SILVA; LEITE; PAULA, 2015).

Como dito por Silva, Leite e Paula (2015, pág. 90):

> A cerveja está enraizada na cultura dos países ocidentais e movimenta vários mercados econômicos no cenário mundial. Por ser uma das bebidas mais consumidas no mundo, desperta interesse de grandes empresas. No Brasil, é considera a bebida de preferência nacional, por ser leve e refrescante, agradando o paladar da maioria dos consumidores de bebidas alcoólicas.

Além das cervejas industrializadas, as cervejas são fabricadas de forma artesanal em vários locais. Essas cervejas são caracterizadas pela melhor qualidade da matéria prima e a adição de produtos regionais, o que gera sabores mais robustos e únicos (SILVA; LEITE; PAULA, 2015). Mesmo sendo produtos de preço e custo elevado, quando comparadas com cervejas industrializadas, essas cervejas têm se tornado uma oportunidade de negócio em crescimento. Segundo o relatório de inteligência do Sebrae, o mercado de cervejas artesanais está em crescimento (SEBRAE, 2015).

Estudar a dinâmica urbana é tradicionalmente desafiador. Geralmente são necessários grandes estudos sociais, incluindo uma grande quantidade de entrevistas, pesquisas e observações, que ainda assim resultam em apenas uma apresentação limitada da realidade (SANTALA et al., 2017; CRANSHAW et al., 2012). Neste trabalho, exploramos o potencial e as possibilidades de usar dados disponíveis em mídias sociais para entendimento da dinâmica urbana do consumo de cerveja e sus diferenças entre diversas cidades.

Existem diversas fontes de dados públicos disponíveis, bastante diversificadas entre si. Cidades em todo o mundo criaram iniciativas para abrir dados públicos, mas as fontes de dados abertas mais utilizadas incluem grandes redes de redes sociais, como Instagram, Foursquare e

Untappd. Esses exemplos de mídias sociais também são chamados de redes sociais baseadas em localização (LSBNs)[1], onde os usuários atuam como uma espécie de sensor usando dispositivos móveis para produzir grandes quantidades de dados relacionados a vários aspectos urbanos e sociais, que podem ser uma fonte rica de informações apoiando a tomada de decisões de indivíduos, empresas e cidades (SILVA et al., 2017).

A quantidade de informação disponível de forma livre e razoavelmente já estruturada sobre o consumo de cerveja foi estudada e descrita por Chorley et al. (2016) por meio de dados retirados do aplicativo *Untappd*[2]. Silva e Graeml (2016) também realizaram um estudo semelhante com foco na realidade brasileira e, de forma específica, analisando o consumo das cervejas artesanais para as cidades de: São Paulo, Rio de Janeiro, Curitiba e Belo Horizonte.

1.1 Motivação

A transformação dos dados coletados na *web* em informações relevantes pode se tornar uma ferramenta fundamental para auxiliar diversos segmentos de mercado. As redes sociais *online* consistem de diversos usuários compartilhando uma grande quantidade de informações. Algumas dessas informações podem ser úteis quando aplicadas a análises direcionadas, como é o caso do gosto dos usuários por cervejas artesanais. Essas análises podem trazer consideráveis benefícios quando utilizadas a fim de encontrar respostas ou padrões comportamentais dos usuários.

A escolha do tema deste trabalho originou-se da possibilidade de utilizar dados de redes sociais para auxiliar questões de planejamento urbano e empreendedores na decisão estratégica do negócio. A preferência por essa área está relacionada principalmente com a necessidade de pesquisar e analisar informações que possam potencialmente contribuir, por exemplo, para um adequado planejamento empresarial, além de identificar outras características a partir dos dados coletados, para auxiliar empresas ou indivíduos que possuam poucos recursos para investir em pesquisas.

1.2 Justificativa

Este estudo apresenta uma oportunidade de entender melhor o consumo das cervejas artesanais, bem como, características de popularidade. Com base nos dados coletados do *Twitter* compartilhados através do *Untappd* em 2012, 2013, 2014 e agora em 2017, uma análise comparativa possibilita identificar quais as possíveis mudanças que ocorreram neste cenário durante este período. Os padrões e características das regiões, bem como informações sobre o índice de consumo de cerveja artesanal, podem beneficiar um planejamento empresarial, urbano, social,

[1] *Location Based Social Network.*
[2] https://untappd.com/

entre outros. Este tipo de informação é muito valiosa para quem busca empreender nesta área de negócio, visto que, o consumo de cerveja artesanal ainda é um hábito pouco estudado, e portanto, podem haver novas oportunidades de pesquisa.

1.3 Objetivos

Esse trabalho tem como objetivo identificar formas de explorar as informações coletadas através das redes sociais, a fim de auxiliar tomadas de decisão no planejamento urbano e aspectos mercadológicos, bem como investigar questões relacionadas às diferenças culturais voltadas ao consumo de cerveja. Este objetivo principal pode ser dividido nos seguintes objetivos específicos:

- Identificar padrões para o consumo de cerveja artesanal, utilizando dados coletados no *Twitter*, compartilhados por usuários através do *Untappd*;
- identificar características sobre o consumo de cerveja;
- procurar padrões que permitam correlacionar ou agrupar as cidades e países com semelhanças no consumo de cerveja artesanal;
- identificar áreas populares para o consumo de cerveja artesanal em algumas cidades brasileiras;
- fornecer uma visualização de como estas informações poderiam auxiliar na tomada de decisão e estudo de diferenças culturais.

1.4 Apresentação do documento

Este documento será organizado da seguinte forma. O Capítulo 2 apresentará o estado da arte do tema escolhido, além de trabalhos correlatos recentes, com o intuito de situar este trabalho no estágio atual do conhecimento. A metodologia utilizada se encontra no Capítulo 3, nele são descritas todas as etapas para o desenvolvimento do projeto. No Capítulo 4 são apresentados os resultados obtidos. No Capítulo 5 encontra-se a conclusão deste trabalho e, por fim, no Capítulo 6 são apresentadas algumas possibilidades de trabalhos futuros.

1.5 Contribuições

Com o objetivo de demonstrar a relevância das informações aqui construídas, neste tópico são citadas as contribuições realizadas a partir dos dados coletados deste trabalho. Portanto, a partir dos resultados obtidos foram produzidos dois artigos publicados em congressos, sendo

eles, respectivamente, o Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos [SBRC][3] e o Simpósio Brasileiro de Sistemas Multimídia e Web [WebMedia][4].

O primeiro artigo publicado foi o *"Making Sense of the City: Exploring the Use of Social Media Data for Urban Planning and Place Branding"*. Conforme informações no *website* do congresso, o SBRC firmou-se como o mais importante evento científico nacional em redes de computadores e sistemas distribuídos, e um dos mais concorridos em informática. O objetivo deste artigo foi de investigar as informações obtidas através do Untappd, no contexto de Planejamento Urbano para a cidade de Curitiba. Através da utilização de um caso real, foi explorada de que forma uma abordagem poderia auxiliar questões de planejamento, considerando aspectos de identidade do local.

O segundo artigo, *"Um brinde ao Untappd! Usando preferências por cervejas para o planejamento urbano e estudo de diferenças culturais."*, foi publicado no WebMedia. Este é um evento promovido anualmente pela Sociedade Brasileira de Computação [SBC][5] e visto como uma excelente oportunidade de intercâmbio científico e técnico entre alunos, pesquisadores e profissionais das áreas de Multimídia, Hipermídia e Web. Este artigo teve como objetivo explorar informações obtidas através do Untappd, no contexto do Planejamento Urbano em Curitiba. Além disso, o artigo mostra o potencial em explorar comportamentos sociais, através da preferência e consumo de cervejas, a fim de identificar aspectos culturais.

[3] https://sbrc2017.ufpa.br/o-evento/apresentacao/
[4] https://www2.sbc.org.br/ce-webmedia/wp/pt/apresentacao/
[5] https://www.sbc.org.br/

2 Levantamento Bibliográfico e Estado da Arte

Nessa seção será descrito o estado da arte do tema escolhido, dividido em quatro itens. Primeiramente sobre o tema de Inteligência coletiva, computação social e urbana, mineração de dados e por fim, trabalhos relacionados ao tema deste estudo.

2.1 Inteligência coletiva

O conceito de inteligência coletiva foi abordado por Pierre Lévy inicialmente na década de 90. Malone (2008) defende que a inteligência coletiva pode ser um grupo de indivíduos que fazem coisas coletivamente que parecem ser inteligentes. Através da Internet foi possível utilizar a inteligência de um grande número de pessoas, que encontram-se conectadas de tantas maneiras diferentes, em uma escala nunca vista antes (MALONE, 2008).

Em 1990 Lévy antecipou grandes mudanças que viriam a acontecer no decorrer dos anos. Naquela época, muitas das pesquisas e publicações na área trataram da teoria sobre a inteligência coletiva e seus temas adjacentes. Lévy (2003) analisa os principais segmentos observados dentro do tema de inteligência coletiva, sendo eles: ética, economia, tecnologia, política e estética. Desta forma, Lévy (2003) direciona e traz reflexões com relação a possíveis campos de exploração e uso da inteligência coletiva.

Para Segaran (2007), os benefícios da inteligência coletiva como, por exemplo, coletar dados de grupos diferentes de indivíduos, já eram possíveis mesmo sem a utilização da Internet. Uma forma básica de realizar pesquisas era através de um censo. Coletar dados de um grupo de pessoas permite extrair algumas conclusões estatísticas desconhecidas das pessoas em sua individualidade (SEGARAN, 2007). Todavia, a ausência de uma linguagem articulada e a falta de correspondência mútua no compartilhamento de conhecimento entre os indivíduos, condição elementar da inteligência coletiva, explicam o insucesso e justificam o fato de não ter atingido anteriormente a inteligência coletiva em sua plenitude(LÉVY, 2003).

A obra Cibercultura (LÉVY, 1999), ainda se mostra atual, por trazer muitos aspectos relevantes do conhecimento do ciberespaço e, consequentemente, os instrumentos da inteligência coletiva. Em sua obra, Lévy, faz algumas reflexões sobre os sistemas de educação e salienta que os profissionais desta área devem ampliar seus conhecimentos com o suporte do ciberespaço, devido à quantidade de oportunidades existentes neste meio. Para Lévy (1999), o meio heterogêneo designado ciberespaço, conhecido atualmente por Internet, possibilita a conexão de vários dispositivos, como, por exemplo, os computadores que transmitem informações através de: correio eletrônico, conferências, documentos compartilhados, entre outros. O ciberespaço é um espaço aberto que consiste em uma realidade multidirecional, em um mundo virtualizado,

em que computadores funcionam como meios de acesso (LÉVY, 1999).

Neste contexto, para Lévy (2003), os indivíduos voltaram a ser nômades. Esta perspectiva não se refere a um local ou território geográfico, mas a um espaço amplo e invisível de conhecimento (LÉVY, 2003). O ciberespaço permite que dois ou mais indivíduos permaneçam interligados, independente do local geográfico em que se encontram (BEMBEM; SANTOS, 2013). Nesse cenário, a via da inteligência coletiva se mostra promissora para Lévy (2003) e, portanto, o autor previa possibilidades de modificações estruturais na constituição do conhecimento no futuro.

A inteligência coletiva pode ser alcançada por meio de uma nova linguagem de comunicação, via digital e nômade, sendo uma inteligência que se expõe, se sustenta e se expande com a utilização de recursos tecnológicos. Portanto, para Lévy (2003), o cenário requer indivíduos que possuam capacidade de navegar no ciberespaço uma vez que:

> A prosperidade das nações, das regiões, das empresas e dos indivíduos depende de sua capacidade de navegar no espaço do saber. A força é conferida de agora em diante pela gestão ótima dos conhecimentos, sejam eles técnicos, científicos, da ordem da comunicação ou derivem da relação "ética" com o outro. Quanto melhor os grupos humanos conseguem se constituir em coletivos inteligentes, em sujeitos cognitivos, abertos, capazes de iniciativa, de imaginação e de reação rápidas, melhor asseguram seu sucesso no ambiente altamente competitivo que é o nosso (LÉVY, 2003, pág. 19).

Diante disto, a contribuição cognitiva do indivíduo pode consumar uma inteligência coletiva bem-sucedida (LÉVY, 2003). Com relação ao espaço do saber, Lévy (2003) aponta para uma infraestrutura das relações humanas, em que indivíduos se reuniriam para compartilhamento de seus saberes.

Neste contexto, Lévy (2003, pág. 99) afirma que:

> Se nossa inteligência pessoal é a alma de um pequeno mundo, os intelectuais coletivos englobam mundos bem maiores e mais variados. Eles enriquecem nosso pensamento quanto maior for a nossa participação, e pensam melhor quanto maior for o número de almas e mundos que envolverem.

Um exemplo bem conhecido são os mercados financeiros, onde um preço não é definido por um indivíduo, mas pelo comportamento de muitas pessoas independentes. Esses mercados combinam as experiências de milhares de pessoas, bem como seus conhecimentos, para criar projeções (SEGARAN, 2007).

Para Bembem e Santos (2013), novas formas de construção e compartilhamento do conhecimento foram possíveis através do trabalho coletivo, que por sua vez, permitiu o desenvolvimento de redes e novas formas de acesso diante de novas tecnologias. Existente desde 1994, o uso dos meios digitais como forma de publicidade em massa ganharam capacidade de mensuração através da utilização de métricas como, por exemplo, quantidade de visitantes totais, impressões, custo por clique, entre outras formas (PALOMINO; ANDRADE, 2013). E com o

surgimento da *Web* 2.0 e o uso real dos conceitos e ideias expressas pela inteligência coletiva, as publicações passaram a ter maior foco no campo prático, em especial a partir de 2009, como relatado por Bembem e Santos (2013). Este ambiente descentralizado e colaborativo da *Web 2.0* deu aos seus usuários a capacidade de gerar conteúdo e colaborar para a geração da própria inteligência coletiva, onde cada indivíduo tem sua parcela de contribuição (BEMBEM; SANTOS, 2013).

Neste contexto, há inúmeras formas de utilizar o conceito de inteligência coletiva. Para Segaran (2007) e Malone (2008) um exemplo é a Wikipédia, criada inteiramente a partir de contribuições de usuários, sendo mantida por vários grupos de pessoas. O resultado disto tudo é uma enciclopédia muito maior do que qualquer outro grupo único foi capaz de realizar. Segaran (2007) também faz menção ao *Google*, mecanismo de busca na Internet mais popular do mundo, que utilizou a inteligência coletiva de forma diferenciada da Wikipédia. O *Google* faz uso de informações do conteúdo já existentes na web. Foi o primeiro a classificar as páginas da Internet levando em consideração informações ditas por usuários (SEGARAN, 2007). Em relação a isto, a partir de 2006, surge o termo *crowdsourcing*, um novo modelo de negócio via web que, de forma inteligente, aproveita as soluções criativas de uma rede distribuída de indivíduos, ou seja, utiliza o potencial da "multidão" para gerar dados, ou até mesmo se apoderar da contribuição de ideias de usuários, a fim de gerar benefícios específicos para organizações (BRABHAM, 2013).

A ideia de ambiente colaborativo é relatado também por Palomino e Andrade (2013), para quem as redes sociais virtuais, como o *Facebook*, *Twitter*, *Youtube*, entre outras existentes, têm adquirido importância devido à auto-geração de conteúdo, onde usuários compartilham experiências de seu cotidiano. Por meio desta cultura participativa, a inteligência coletiva se torna mais forte. Portanto, os processos de coleta de informações se tornam um pilar de grande importância para esta área, pois, a partir dos dados coletados, é possível criar métricas e indicadores para mensurar o alcance das informações sobre diversas perspectivas dos consumidores (PALOMINO; ANDRADE, 2013). Para Segaran (2007), o sucesso das redes sociais tem facilitado cada vez mais o compartilhamento de informações entre os usuários.

Nesta direção, Silva (2012) salienta a importância de resgate dos dados em sistemas *online* [1], onde são gerados milhões de unidades de conteúdo publicados por minuto, que são utilizados por diferentes organizações ou até mesmo indivíduos distintos para compreender, por exemplo, um determinado comportamento, uma vez que dados digitais são basicamente reflexos e rastros de comportamentos, ações e acontecimentos físicos. Além disto, este conteúdo pode ser processado e expandido a fim de responder questões de acordo com o interesse de diferentes empresas ou indivíduos. Esta grande quantidade de dados, resultado de um mundo conectado, possibilita a criação de sistemas de "predição", ou até mesmo identificação de tendências, ainda em pouco uso, devido sua complexidade, porém, com potencial de ganhar maior importância (SILVA, 2012).

[1] Plataforma virtual, como exemplo, as redes sociais, que refletem experiências da vida real entre os indivíduos.

2.2 Computação social e urbana

A Computação Social para Schuler (1994) é descrita como qualquer tipo de aplicação computacional em que *softwares* servem como recurso intermediário ou principal para estabelecer uma relação social. Ainda, a definição adotada em Ala-Mutka et al. (2009) para Computação Social é de que se trata de um conjunto de aplicações abertas, baseadas em *web* e de fácil utilização pelos usuários, em que os próprios indivíduos estão habilitados a assumir novos papéis na criação de conteúdo, disseminação de informação e prestação de serviços. Para Parameswaran e Whinston (2007), a Computação Social é um conceito que permite uma rica troca de informações e surgiu para dominar a *web*. Parameswaran e Whinston (2007) ainda citam que a Computação Social tem o objetivo de empoderar usuários a fim de manifestarem a sua criatividade, engajarem-se em interações sociais, contribuirem com seus conhecimentos, entre outras atividades que envolvam a coletividade. Ainda, muitas organizações podem sofrer mudanças em sua forma de atuação devido às críticas coletivas de seus consumidores (PARAMESWARAN; WHINSTON, 2007). Isso faz com que, conforme Ala-Mutka et al. (2009), a inovação social tome espaço e avance para fora das instituições estabelecidas e práticas de trabalho.

Sendo assim, a Computação Social deve permitir aos usuários conectar-se em rede, compartilhar dados, colaborar e co-produzir conteúdo (ALA-MUTKA et al., 2009). Algumas delas são: Redes sociais, como o *Facebook* e *Linkedin*, que permitem aos usuários compartilhar informações pessoais e profissionais, respectivamente; *blogs* que permitem aos seus usuários expressarem-se e interagir uns com os outros; *sites* de leilão *online*, onde os usuários compartilham opiniões e em conjunto criam um sistema de reputação dos vendedores, como o *eBay*; *sites* de conteúdo colaborativo onde usuários podem compartilhar, produzir e co-produzir conteúdo, como a *Wikipedia*; jogos *multi-player*, como o *World of Warcraft*; entre outros exemplos (ALA-MUTKA et al., 2009).

Para Ala-Mutka et al. (2009), a Computação Social teve uma rápida ascenção tornando-se popular e um fenômeno importante, em termos de alcance, tempo de uso e atividades realizadas. Este rápido crescimento foi motivado pelo aumento na diponibilidade de banda larga e na quantidade de dispositivos acessando à Internet. A utilização dos dispositivos *mobile* tornou-se popular, e em 2009, cerca de 40% dos usuários visitaram as redes sociais por meio de dispositivos móveis, onde as principais atividades foram: verificar comentários, mensagens e postar atualizações de *status* (ALA-MUTKA et al., 2009). Como consequência da rápida ascenção, as redes sociais tornaram-se os maiores sistemas de reputação e identidade do mundo, a exemplo do *twitter* que rapidamente tornou-se um fenomeno global, o qual fornece acesso às informações praticamente em tempo real (ALA-MUTKA et al., 2009).

Considerando este cenário , Kindberg, Chalmers e Paulos (2007) trazem à luz o conceito de Computação Urbana, como sendo a integração da computação com sensores e atuação das tecnologias no dia a dia do ambiente urbano como, por exemplo, ruas, praças, *pubs*, *shoppings*, entre outros. Paulos e Goodman (2004) e Paulos, Anderson e Townsend (2004) introduziram

o termo Computação Urbana para caracterizar os estudos realizados sobre dados heterogêneos coletados em ambientes urbanos, a partir de sensores, veículos e pessoas (SILVA; LOUREIRO, 2015).

Indivíduos constantemente entram e saem de ambientes urbanos, ocupando-os em volumes diferentes durante os vários períodos do dia e, ainda, alterando seus padrões de comportamento entre dia e noite. Entretanto, os ambientes urbanos são pouco explorados, pois são desafiadores quanto à experimentação e implantação de artefatos. Por exemplo, a dificuldade em se instalar sensores nas cidades torna complexa a análise destes ambientes (KINDBERG; CHALMERS; PAULOS, 2007). Ainda, considerando este caso, aproximadamente metade da população mundial vive em ambientes urbanos, onde a maioria das pessoas possui telefones portáteis, os quais possuem diversas funcionalidades, além de simplesmente realizar chamadas de voz. Para Silva e Loureiro (2015), estes usuários atuam como sensores sociais, fornecendo dados que relatam suas experiências de vida diária.

Os dispositivos portáteis despertaram um grande interesse em pesquisadores devido à sua grande capacidade em atuar como sensores. Além disso, os dispositivos portáteis tornaram-se intrinsicamente pessoais, pois são mini-computadores que podem ser transportados no bolso e, por este motivo oferecem um imenso potencial de coleta de dados muito valiosos sobre o ambiente, hábitos e comportamentos das pessoas (LORETO et al., 2016). Para Burke et al. (2006) o sensoreamento participativo irá apropriar-se de dispositivos portáteis para formar sensores interativos e participativos que permitam aos usuários coletar, analisar e compartilhar conhecimento. Neste contexto, Silva e Loureiro (2015) exemplicam que as análises podem ocorrer por meio da utilização de certos tipos de mídias sociais, que podem ser vistas como uma rede de sensoriamento participativo (RSP).

Sendo assim, uma RSP requer participação ativa e voluntária na contribuição de dados sobre diversos aspectos urbanos. Para que as pessoas possam atuar como sensores sociais, estes dados devem necessariamente estar disponíveis de forma pública. Segundo Silva e Loureiro (2015), uma RSP é composta por características de determinada rede social em conjunto com informações espaço-temporais[2]. O *Twitter* é uma mídia social considerada como exemplo de uma RSP, pois os dados fornecidos pelos usuários permitem monitorar acontecimentos quase em tempo real em uma determinada cidade (SILVA; LOUREIRO, 2015).

Neste cenário, a Computação Urbana, pela aplicação de suas tecnologias, oferece serviços avançados aos cidadãos (KAMIENSKI et al., 2016). Como exemplo, Chicago, uma cidade de grande porte nos Estados Unidos, por meio do Smart Grid, [3] está modernizando a infraestrutura de rede elétrica para conceder um serviço inteligente com opções mais econômicas e medição de consumo em tempo real utilizando medidores individuais conectados a uma rede. Para Silva e Loureiro (2015), a Computação Urbana se relaciona com determinadas discipli-

[2] Caracterizado por serviços baseados em localização.

[3] http://www.cityofchicago.org/city/en/progs/env/smart-grid-for-a-smart-chicago.html

nas do domínio da Ciência da Computação, como, por exemplo, sistemas distribuídos, redes de computadores, redes de sensores, sistemas cooperativos, inteligência artificial e interação humano-computador. Neste contexto, Kamienski et al. (2016), fazem menção às possíveis categorias de cenários da Computação Urbana, sendo elas: planejamento urbano, sistemas de transporte, ambiente, segurança, consumo de energia, economia e aplicações sociais.

Kamienski et al. (2016) ainda citam que outras áreas de interesse urbano também necessitam do apoio da Ciência da Computação para analisar dados com o objetivo de obter a melhoria constante da vida dos indivíduos. Para Silva e Loureiro (2015), as análises também têm por objetivo garantir a melhoria dos aspectos urbanos e sociais que são cada vez maiores e crescem juntamente com o tamanho das cidades. Por meio destas análises podem ser desenvolvidos sistemas capazes de melhorar o entendimento sobre a dinâmica das cidades, além de trazer benefícios para seus habitantes (SILVA; LOUREIRO, 2015).

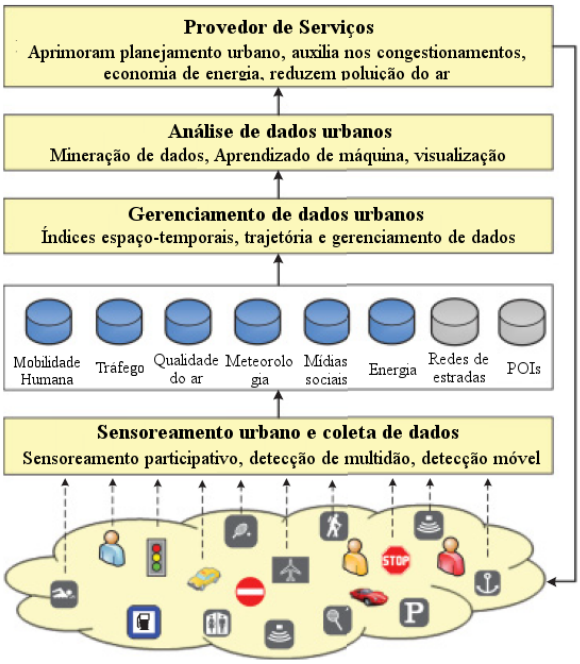


Figura 1: *Framework* geral da Computação Urbana

Fonte: extraído de Zheng et al. (2014).

A Figura 1, extraída de Zheng et al. (2014), descreve o *Framework* geral da Computação Urbana, que reflete seu funcionamento de modo abrangente. Este quadro é composto por quatro camadas, sendo elas: detecção urbana, gerenciamento de dados urbanos, análise de dados e fornecimento de serviços (ZHENG et al., 2014).

Segundo Zheng et al. (2014) na primeira camada, a de detecção urbana, procura-se coletar dados de diversas fontes, seja por meio de sensores dos *smartphones* pessoais ou ainda pelo constante monitoramento das mídias sociais. Conforme citam Silva e Loureiro (2015), esses dados ainda podem ser obtidos a partir de canais oficiais, redes de sensores tradicionais,

infraestrutura das cidades e rede de sensoreamento participativo. A segunda camada, de gerenciamento dos dados, é onde a organização dos dados acontece garantindo a eficiência na análise dos dados. A terceira camada representa a etapa de análise dos dados. Conforme Silva e Loureiro (2015), é onde a edição, execução de códigos e interpretação dos resultados acontece. Finalmente, Zheng et al. (2014) citam que a quarta camada representa a etapa de fornecimento de serviços, desenvolvido a partir do conhecimento obtido na camada inferior. O objetivo desta camada é realizar melhorias no ambiente urbano por meio de constantes iterações sobre essas etapas. Algumas das responsabilidades atribuídas à Computação Urbana incluem aprimorar a mobilidade urbana, amenizar congestionamentos, reduzir o consumo de energia e poluição do ar (ZHENG et al., 2014).

2.3 Mineração de dados

Parte do interesse em mineração de dados é motivado, basicamente, pelas mudanças no ambiente corporativo e os avanços nas pesquisas referentes à essa área (CABENA et al., 1998). Atualmente, as mudanças são fundamentais e influenciam o modo como as organizações veem e planejam aproximar-se de seus clientes. Algumas delas são: padrões comportamentais dos consumidores; situação de mercado; novos nichos de mercado; menor ciclo de vida dos produtos, entre outros (CABENA et al., 1998). Seguindo este contexto, Everitt et al. (2011) exemplificam essas mudanças fazendo menção às pesquisas de mercado, em que agrupar um grande número de entrevistados, de acordo com suas preferências, pode ser útil para identificar um "nicho de produto". Para Fayyad, Piatetsky-Shapiro e Smyth (1996), o processo de extração e análise de bases, a partir de um conjunto de dados, é de suma importância no suporte ao processo decisório.

Devido à crescente quantidade de dados digitais armazenados em bases de dados, tornou-se difícil identificar informações úteis sem o auxílio de teorias e ferramentas computacionais, trazendo em questão o problema de se dispor de muitos dados, mas pouco ou quase nenhum conhecimento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Diante deste problema, existe uma necessidade de aplicar técnicas e ferramentas que permitam a automatização e análise dos dados de forma inteligente, agregando sentido aos dados brutos conforme o objetivo do *Knowledge Discovery in Databases* (KDD) [4] (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Ainda não é consenso a definição dos termos *Knowledge Discovery in Databases* e *Data Mining*. Para Fayyad, Piatetsky-Shapiro e Smyth (1996) e Tan, Steinbach e Kumar (2009), a Mineração de Dados está inclusa em uma das atividades do processo do KDD, mas para Liu (2007), a Mineração de Dados pode ser também denominada KDD. Cabena et al. (1998), por sua vez, definem os termos como sendo sinônimos e explicam que não existe uma única definição

[4] Descoberta de conhecimento nas Bases de Dados.

para a Mineração de Dados que atenda uma aprovação universal. Entretanto, apresenta uma definição aceitável: a Mineração de Dados é o processo de extrair de grandes bases de dados informações previamente desconhecidas, válidas e contestáveis (CABENA et al., 1998).

Para Tan, Steinbach e Kumar (2009), o KDD é uma parte integral da descoberta de conhecimento em bancos de dados, onde os dados brutos são transformados em informações úteis, sendo composto essencialmente de três etapas: pré-processamento, mineração de dados e pós-processamento conforme demonstrado na Figura 2.

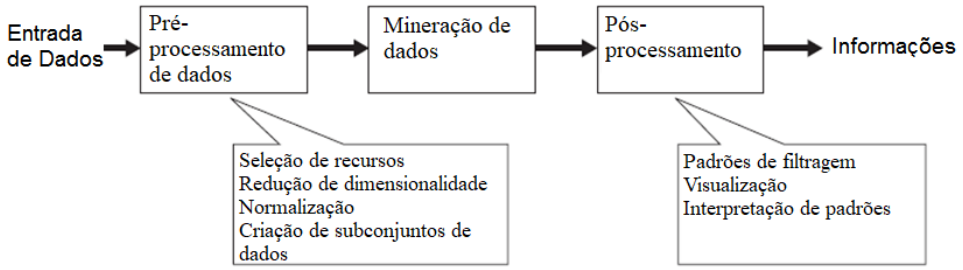


Figura 2: O processo de descoberta de conhecimento em bancos de dados (KDD).

Fonte: extraído de Tan, Steinbach e Kumar (2009).

Os dados de entrada podem ser armazenados de diversas maneiras, desde arquivos simples ou tabelas em um repositório de dados (TAN; STEINBACH; KUMAR, 2009). Na etapa de pré-processamento de dados ocorre a transformação dos dados brutos para um formato que dê suporte para a realização das etapas subsequentes (TAN; STEINBACH; KUMAR, 2009). Liu (2007), explica que nesta etapa os dados brutos normalmente possuem ruídos, anomalias ou atributos irrelevantes, que os tornam inadequados para utilização imediata. O quadro Mineração de Dados, na Figura 2, retrata a etapa em que os dados tratados no processo anterior são alimentados para um algoritmo de Mineração de Dados, que irá produzir padrões ou gerar conhecimento (LIU, 2007). A última etapa, o pós-processamento de dados, envolve padrões de filtragem, interpretação de padrões e, por fim, a visualização dos resultados que permitem aos analistas a obtenção de uma diversidade de pontos de vista (TAN; STEINBACH; KUMAR, 2009). Para Liu (2007), várias técnicas de avaliação e visualização são usadas para tomar decisão. Porém, nem todos os padrões identificados podem ser úteis. Sendo assim, somente esta etapa identifica aqueles que são úteis para utilização.

De acordo com Tan, Steinbach e Kumar (2009), a etapa do pré-processamento dos dados quando comparada com as demais etapas, pode ser considerada como o passo mais trabalhoso e, talvez, o mais demorado no processo geral, devido às diversas formas como os dados podem ser coletados e armazenados.

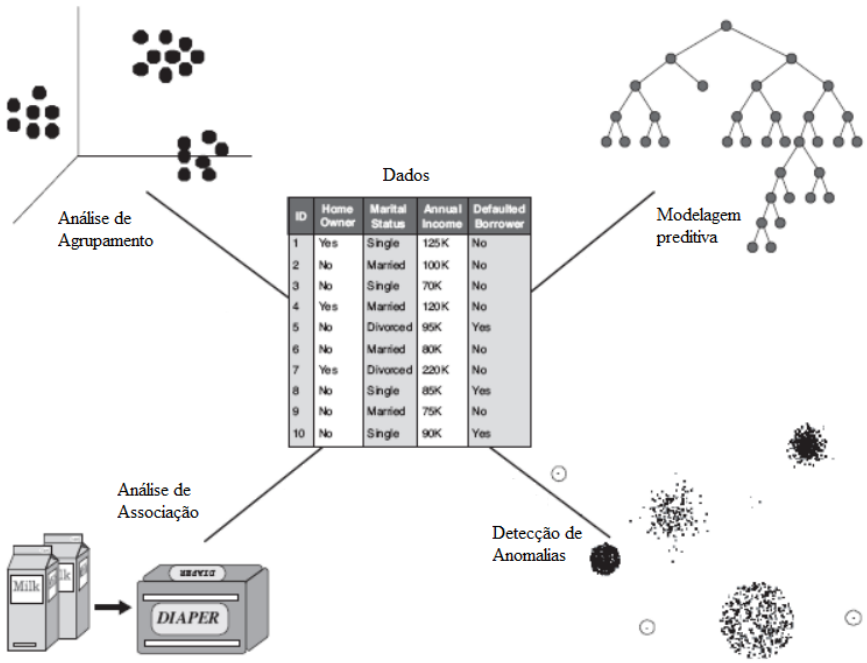
As tarefas de Mineração de Dados geralmente são divididas em duas categorias principais, sendo elas: tarefas de previsão e tarefas descritivas (TAN; STEINBACH; KUMAR, 2009). Conforme Tan, Steinbach e Kumar (2009), as tarefas de previsão têm como objetivo:

> prever o valor de um determinado atributo baseado nos valores de outros atributos. O atributo a ser previsto é comumente conhecido como a variável dependente ou alvo, enquanto que os atributos usados para fazer a previsão são conhecidos como as variáveis independentes ou explicativas.

Ainda para Tan, Steinbach e Kumar (2009), as tarefas descritivas, por sua vez, têm como objetivo:

> derivar padrões (correlações, tendências, grupos, trajetórias e anomalias) que resumam os relacionamentos subjacentes nos dados. As tarefas descritivas da mineração de dados são muitas vezes exploratórias em sua natureza e frequentemente requerem técnicas de pós-processamentos para validar e explicar os resultados.

A Figura 3 ilustra um quadro das tarefas centrais da Mineração de Dados. Sendo elas: Modelagem Previsiva, Análise de Associação, Detecção de Anomalias e Análise de Agrupamentos.



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Figura 3: Quatro das tarefas centrais da mineração de dados.

Fonte: extraído de Tan, Steinbach e Kumar (2009).

Para Cabena et al. (1998), a modelagem previsiva consiste em usar observações para formar um modelo contendo características sobre um fenômeno. No contexto da mineração de dados, é utilizada para analisar dados existentes na base e determinar características essenciais sobre os dados. Entretanto, para que isso ocorra, é necessário que os dados contenham observações consistentes que permitam ao modelo aprender a fazer predições precisas. Tan, Steinbach e Kumar (2009) citam que a modelagem previsiva pode ser dividida em dois tipos: a de classificação, onde as variáveis alvo são discretas, e a regressão, que é usada para variáveis contínuas.

Um exemplo de classificação seria prever se um usuário *Web* fará uma compra em uma livraria *online*. A variável alvo, neste caso, é de valor binário. Por outro lado, prever o preço futuro de uma ação, por exemplo, é uma tarefa de regressão, dado que o preço é um valor contínuo. O objetivo de ambas as tarefas é gerar um modelo que minimize o erro entre o valor previsto e o real da váriável (TAN; STEINBACH; KUMAR, 2009).

A análise de associação, por sua vez, pode ser utilizada para descobrir padrões que descrevam características altamente associadas dentro do conjunto de dados (TAN; STEINBACH; KUMAR, 2009). Sendo assim, Cabena et al. (1998) afirmam que, em contraste com a modelagem previsiva, que procura caracterizar a base de dados como um todo, a análise de associação tem como objetivo estabelecer ligações entre registros individuais ou conjuntos de registros. Considerando a quantidade de dados, ou seja, o tamanho do espaço de busca, o objetivo desta análise é extrair os padrões mais interessantes de uma forma eficiente (TAN; STEINBACH; KUMAR, 2009). Liu (2007) e Tan, Steinbach e Kumar (2009) trazem o exemplo clássico da aplicação da análise de associação, conhecida por análise de cesta de compras, onde procura-se descobrir como os itens comprados por consumidores de um supermercado estão associados.

Ainda, a detecção de anomalias consiste em identificar observações em que as características são significativamente diferentes do resto dos dados (TAN; STEINBACH; KUMAR, 2009). Essas observações são conhecidas como fatores estranhos, anomalias ou ainda *outliers*. Cabena et al. (1998) expressam que analistas utilizam estatística e técnicas de visualização, como a regressão linear, para auxiliar a identificação dos *outliers* nos dados. Tan, Steinbach e Kumar (2009) ainda complementam que um bom detector de anomalias deve ter uma alta taxa de detecção de *outliers* e uma baixa taxa de alarmes falsos. Exemplos de aplicação da detecção de anomalias são: detecção de fraudes, intromissões na rede, padrões incomuns de doenças e perturbações no meio ambiente.

Finalmente, conforme apresentado por Cabena et al. (1998), a análise de agrupamento consiste em particionar base de dados em segmentos onde os registros sejam similares, ou seja, agrupar dados que contenham determinadas propriedades semelhantes entre si. Alguns exemplos de aplicação da análise de agrupamento são: agrupar conjuntos de clientes relacionados, descobrir áreas do oceano que afetem significativamente o clima da Terra, entre outros (TAN; STEINBACH; KUMAR, 2009). Sendo assim, os algoritmos e métodos de agrupamento de dados são descritos com maiores detalhes a seguir.

2.4 Agrupamento ou clusterização

Segundo Liu (2007), a clusterização é uma das técnicas de análise de dados mais utilizadas em quase todos os campos. Entre as possíveis áreas de aplicação, Liu (2007), cita medicina, psicologia, botânica, sociologia, biologia, arqueologia, *marketing*, entre outras. Devido à expansão da *Web*, surgiu a necessidade de explorar conjuntos de dados disponíveis em muitas

áreas da ciência (LIU, 2007).

Levando em consideração essa necessidade de explorar dados, Tan, Steinbach e Kumar (2009) citam que a *clusterização* é utilizada com o objetivo de dividir dados em grupos que tenham sentido ou sejam usuais, ou ainda em grupos que tenham ambas as características. Neste contexto, Witten e Frank (2005) afirmam que as técnicas de *clusterização* - também conhecidas como método de agrupamento - são empregadas quando existem casos que podem ser divididos em grupos naturais. Presume-se que esses *clusters* reflitam semelhanças que ocorrem nestes grupos naturais onde os casos são esboçados.

Ainda de acordo com Witten e Frank (2005), o resultado de uma *clusterização* pode se expressar de diferentes formas. Os grupos identificados podem ser exclusivos, de forma que os casos observados pertençam somente a um dos agrupamentos. Podem ser sobrepostos, quando um dos casos pode fazer parte de diversos grupos. Ainda, podem ser probabilísticos, onde cada caso pertencerá a cada grupo com uma certa probabilidade. E, por fim, podem ser hierárquicos, de tal modo que exista uma divisão entre as instâncias dos grupos. Essas formas de clusterização podem ocorrer conforme a natureza das semelhanças entre os elementos agrupados (WITTEN; FRANK, 2005).

Neste contexto, existem duas principais divisões de categoria nos métodos de agrupamento de dados. A primeira refere-se aos Métodos Particionais que contemplam algoritmos exclusivos e não-exclusivos. A segunda categoria faz referência aos métodos hierárquicos onde algoritmos aglomerativos e algoritmos divisivos estão inseridos (VALE, 2005). Sendo assim, abaixo são apresentadas as principais formas de clusterização.

2.4.1 O algoritmo K-means

O termo *"K-means"* foi inicialmente proposto por MacQueen et al. (1967), embora a ideia seja uma atualização à proposta por Steinhaus (1956).

O algoritmo de clusterização *K-means* requer várias iterações. Cada iteração envolve identificar a distância de *"n"* observações ao centro dos *"k" clusters*, com o objetivo de agrupar as observações que possuem as menores distâncias entre si (WITTEN; FRANK, 2005). Conforme abordado por MacQueen et al. (1967), o algoritmo *K-means* é facilmente programável e ecônomico computacionalmente, de forma que é possível utilizá-lo para processar grandes quantidades de dados.

O primeiro passo a ser dado é definir o valor de *"k"*, que representa a quantidade de *clusters* que serão gerados. Na sequência, o algoritmo seleciona *"k"* pontos aleatoriamente, a fim de torná-los centros de *clusters*. Em seguida, todos os elementos são atribuídos ao centro do *cluster* mais próximo de acordo com o resultado do cálculo da distância. Desta forma, objetiva-se garantir que os elementos sejam agrupados de acordo com a melhor distância na realização do re-cálculo do centróide em cada iteração do algoritmo. O processo é terminado

após uma quantidade pré-estabelecida de iterações ou quando os *clusters* se estabilizam, ou seja, quando os agrupamentos formados permanecem inalterados em determinada iteração do algoritmo (WITTEN; FRANK, 2005), conforme apresentado na Figura 4.

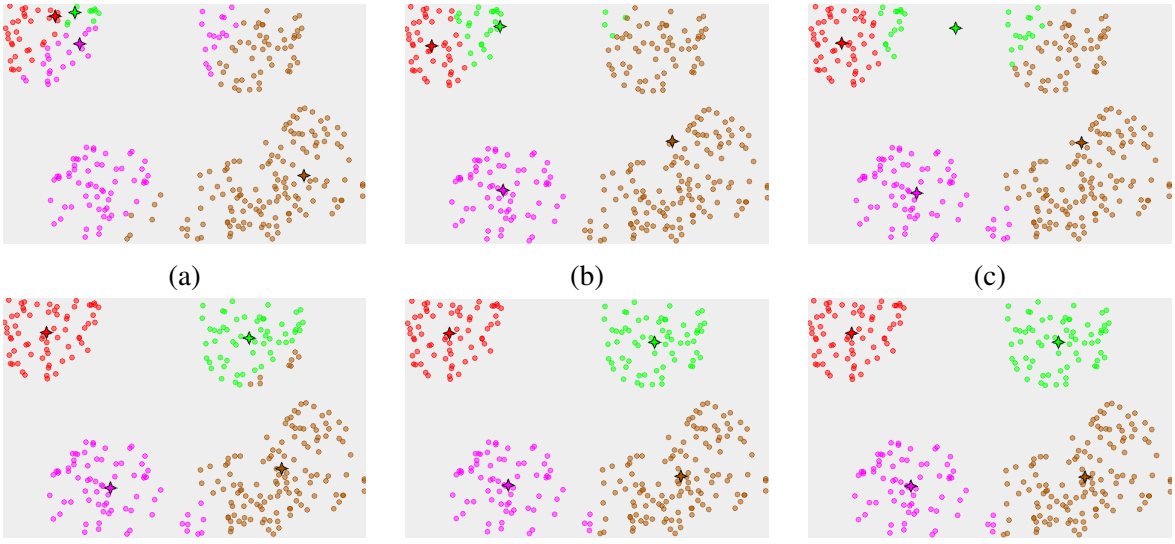


Figura 4: Representação das sucessivas iterações do algoritmo *K-means*

A Figura 4 foi extraída do aplicativo proposto por Mirkes e Leicester (2012) que permite a visualização passo a passo da execução do algoritmo. Nas imagens que compõem a Figura 4 é possível visualizar a sequência de execução deste algoritmo. Sendo assim, 4a demonstra a escolha aleatória de pontos que serão adotados como centro de *cluster*, onde, para esta ilustração *"k"* recebeu o valor quatro. A sequência 4b, 4c, 4d, 4e representa os passos seguintes, onde são recalculados os centros de *cluster* através da média das distâncias entre os pontos. E por fim, é possível perceber a estabilização do algoritmo ao comparar as imagens 4e e 4f, onde não existe alteração entre elas. O pseudo-código é apresentado abaixo em Algoritmo 1.

Algoritmo 1: Algoritmo *K-Means* básico

```
1: Selecione K pontos como centróides iniciais.
2: repita
3:     Forme K grupos atribuindo cada ponto ao seu centróide mais próximo.
4:     Recalcule o centróide de cada grupo.
5: até que os centróides não mudem.
```

Fonte: extraído de Tan, Steinbach e Kumar (2009).

2.4.2 Algoritmo DBSCAN

O *Density Based Spatial Clustering of Applications with Noise* (DBSCAN) é um dos principais algoritmos de *clusterização* baseado em densidade e foi proposto por Ester et al.

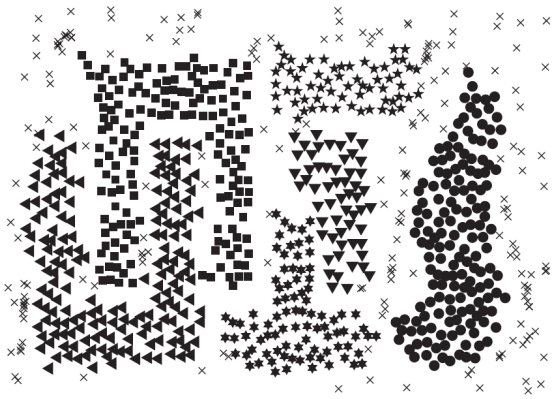
(1996).

Algumas das características que fazem do DBSCAN uma escolha técnica adequada para diversos domínios de aplicação consiste no fato do número de *clusters* ser variável, já que não é necessário informar inicialmente uma quantidade de *clusters* específica. Além disso, permite a formação de *clusters* de forma aleatória e ainda identifica facilmente ruídos ou pontos isolados dos agrupamentos - denominados como *outliers* (ESTER et al., 1996).

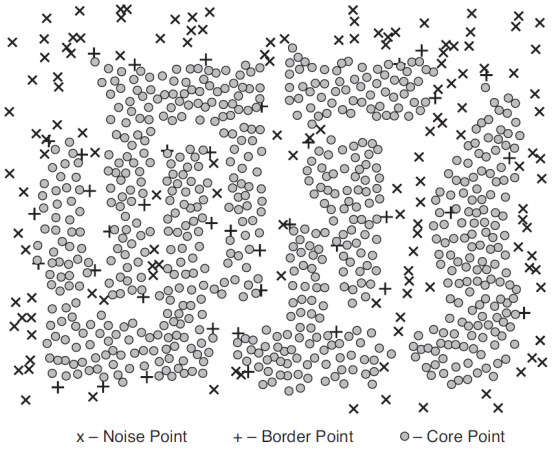
O DBSCAN classifica áreas de densidade, com o objetivo de formar *clusters*, através da distância euclidiana dos vizinhos mais próximos de um determinado ponto. Sendo assim, para que um *cluster* seja formado é necessário que exista uma quantidade mínima de pontos semelhantes dentro de uma determinada área (ESTER et al., 1996). Ainda segundo Ester et al. (1996), além dos dados, são passados dois parâmetros para o algoritmo, são eles:

- Epsilon (ε): raio estabelecido para a procura de pontos semelhantes.
- minPoints: quantidade mínima de pontos semelhantes dentro do raio determinado.

Inicialmente, o algoritmo escolhe aleatoriamente um ponto de ínicio e analisa a sua vizinhança. Se a quantidade de objetos localizados dentro do raio *epsilon* apresentar uma quantidade maior ou igual de pontos que o parâmetro *minPoints*, então, um novo *cluster* é formado. Após isto este mesmo processo é executado nos pontos vizinhos a fim de inserir no *cluster* criado os pontos que atendem os requisitos. Caso, ao fim da execução, ainda existam pontos não pertencentes a algum *cluster*, o processo é reiniciado a partir de um destes pontos até que todos os *clusters* possíveis tenham se formado (ESTER et al., 1996). É possível ver o resultado final do processo executado pelo DBSCAN na Figura 5a.



(a) *clusters* encontrados pelo *DBSCAN*



(b) Identificação de *core*, *border* e *noise points*

Figura 5: *Clusterização* utilizando o *DBSCAN* em 3000 pontos

Fonte: extraído de Tan, Steinbach e Kumar (2009).

Na Figura 5b é possível perceber que a execução do DBSCAN pode marcar os pontos como: *corePoints*, um ponto que possui uma quantidade de vizinhos maior ou igual ao *minPoints*, ou seja, são os pontos pertencentes aos *clusters* formados. Os pontos ainda podem ser classificados como *borderPoints*. Isso ocorre quando o parâmetro *minPoints* não é atendido, mas este ponto é vizinho de um *corePoint*. E, finalmente, um *noisePoint* é marcado quando o número de vizinhos é menor que a quantidade de *minPoints* e não há nenhum *corePoint* em sua redondeza no raio delimitado (ESTER et al., 1996). O pseudo-código é apresentado abaixo em

Algoritmo 2.

Algoritmo 2: Algoritmo DBSCAN

1: Rotular todos os pontos como de centro, de limite ou de ruído.
2: Eliminar os pontos de ruído.
3: Colocar uma aresta entre todos os pontos de centro que estejam dentro da *Eps* uns dos outros.
4: Tornar cada grupo de pontos de centro conectados um grupo separado.
5: Atribuir cada ponto de limite a um dos grupos dos seus pontos de centro associados.

Fonte: extraído de Tan, Steinbach e Kumar (2009).

2.4.3 Métodos hierárquicos

Na classificação hierárquica, os dados são compostos de uma série de partições que podem conter desde um único *cluster*, contendo todos os pontos, a *"n" clusters*, cada um contendo um único ponto (EVERITT et al., 2011). A estrutura utilizada para representar essas partições de dados é denominada dendograma e, conforme Witten e Frank (2005), a denominação possui origem da palavra grega *"dendron"* que significa "árvore". Neste contexto, a estrutura é uma representação gráfica que demonstra a ordem em que os dados foram agrupados (VALE, 2005).

Para Tan, Steinbach e Kumar (2009) e Everitt et al. (2011), os métodos hierárquicos são considerados uma segunda categoria importante dos métodos de *clusterização*. Existem duas abordagens básicas para gerar um agrupamento hierárquico, sendo elas: *"divisiva"* e *"aglomerativa"*. A Figura 6 consiste em um diagrama bi-dimensional e tem por objetivo exemplificar a estrutura de um dendograma. Portanto, representa as fusões efetuadas no agrupamento aglomerativo, bem como, as divisões que ocorrem no agrupamento divisivo.

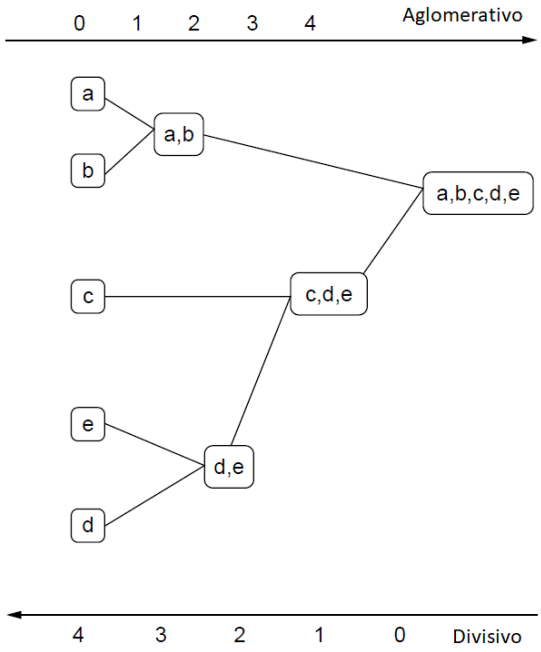


Figura 6: Exemplo de uma estrutura hierárquica (aglomerativa e divisiva)

Fonte: extraído de Everitt et al. (2011).

A definição para os metódos hierárquicos aglomerativos, Everitt (2006) é de que são métodos de análise em *cluster* que iniciam cada ponto em um *cluster* separado, e, após uma sequência de passos, combina pontos com alguma semelhança em *clusters* novos e maiores. Este processo se repete até que seja atingido um último estágio onde todos os pontos são membros de um único grupo. É possível visualizar esta ocorrência na Figura 6, da esquerda para a direita. Conforme Everitt et al. (2011), este método é, provavelmente, um dos mais utilizados para agrupamentos hierárquicos.

As operações básicas que permitem realizar o agrupamento aglomerativo são similares. A cada estágio os métodos procuram realizar a fusão dos pontos ou grupos que possuem semelhanças. As diferenças entre os métodos existentes encontram-se nas formas de definição da distância entre *clusters*, ou seja, a similaridade através da distância euclidiana (EVERITT et al., 2011).

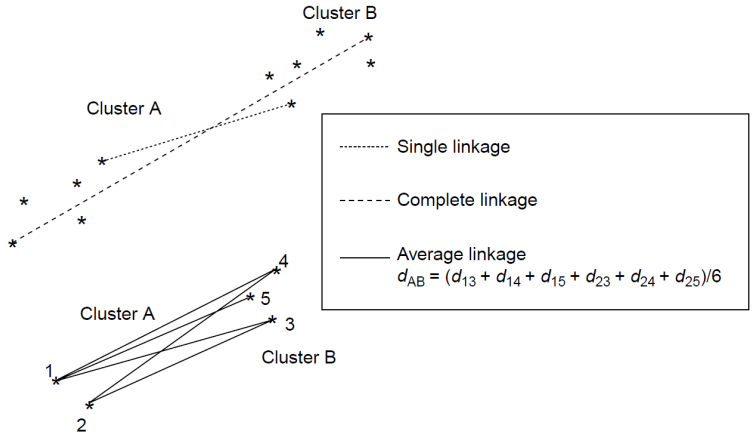


Figura 7: Exemplos dos três métodos de cálculo de distância entre *clusters*: *single*, *complete* e *average*

Fonte: extraído de Everitt et al. (2011).

Conforme descrito por Everitt et al. (2011) *single linkage*, *complete linkage* e *group average* são as operações básicas mais comuns no agrupamento aglomerativo. Os três métodos utilizam uma matriz de proximidade como entrada de dados, ou seja, uma estrutura de dados que contenha as distâncias relativas entre cada ponto no conjunto a ser analisado. O método *single linkage* define a proximidade a partir da distância entre os dois pontos mais próximos, desde que estes estejam em *clusters* diferentes. O *complete linkage*, por sua vez, utiliza os pontos mais afastados para calcular a distância entre os *clusters*. Por fim, o método *group average* define a proximidade entre *clusters* por meio da distância média entre todos os pontos dos *clusters* (TAN; STEINBACH; KUMAR, 2009). De acordo com Everitt et al. (2011), além destas operações, diversas outras são utilizadas, como por exemplo, o método aglomerativo por centróide, *Median linkage* e o método de *Ward*. A Figura 7 exemplifica as três operações descritas acima e o Algoritmo 3, apresenta o pseudo-código.

Algoritmo 3: Algoritmo de agrupamento hierárquico aglomerativo básico

```
1: Calcule a matriz de proximidade, caso necessário.
2: repita
3:     Agrupe os dois grupos mais próximos.
4:     Atualize a matriz de proximidade para refletir a proximidade entre o novo grupo e
       os grupos originais.
5: até que reste apenas um grupo.
```

Fonte: extraído de Tan, Steinbach e Kumar (2009).

Métodos divisivos operam na direção oposta aos métodos aglomerativos, segundo Everitt et al. (2011). Para Tan, Steinbach e Kumar (2009), o processo inicia por um grande *cluster* e sucessivamente ocorre a divisão deste *cluster*. É possível visualizar o funcionamento deste mé-

todo na Figura 6, quando analisado da direita para a esquerda. Ainda para Everitt et al. (2011), os algoritmos divisivos tornam difícil a implementação de forma eficiente, pois costumam ser computacionalmente custosos, devido às sucessivas divisões de *clusters* para formar outros dois *sub-clusters*, em cada estágio, até que restem grupos únicos. Esse é um dos motivos para os métodos divisivos serem utilizados com menor frequência.

Contudo, ainda para Everitt et al. (2011), para dados contendo variáveis binárias, é possível realizar as divisões de um modo computacionalmente eficiente através do *"Monothetic Divisive Method"*, que emprega o uso de uma única variável binária a ser dividida em um determinado estágio. Este método procura dividir *clusters* de modo que, em cada estágio, os *clusters* obtidos contenham membros com uma divisão precisa entre os atibutos presentes e ausentes. Ainda existem os *"Polythetic divisive methods"* que se aproximam dos métodos aglomerativos, uma vez que utilizam as váriaveis simultaneamente e podem utilizar uma matriz de proximidade. Abaixo em Algoritmo 4 é apresentado um pseudo-código do método divisivo.

Algoritmo 4: Algoritmo de agrupamento hierárquico divisivo MST

```
1: Calcule uma árvore de dispersão mínima para o grafo de diferenças.
2: repita
3:         Crie um novo grupo para dividir a conexão correspondente à maior diferença.
4: até que restem apenas grupos únicos.
```

Fonte: extraído de Tan, Steinbach e Kumar (2009).

2.5 Trabalhos correlatos

A seguir, são apresentados alguns trabalhos com temas semelhantes ao proposto no presente estudo. Em Silva e Graeml (2016) são demonstradas algumas das possibilidades em se analisar dados de redes sociais, com exemplo de coleta e análise de dados do *Untappd*. O estudo em questão foi elaborado para auxiliar um empreendedor fictício a tomar suas decisões e planejar estratégias com foco em pequenas ou médias empresas que possuem pouco ou nenhum recurso para investir em pesquisas.

O *Untappd* é uma rede social que permite que usuários compartilhem com seus amigos informações sobre as cervejas e locais onde estão consumindo. Neste contexto, Silva e Graeml (2016) também desenvolveram um estudo onde foram coletados dados das mensagens do *Twitter*, chamadas de *tweets*, que foram disparadas pelo *Untappd*. Os dados foram coletados e analisados, com foco principal nas informações do Brasil, embora representassem apenas 1% de toda a base de dados dos *tweets* coletados. Para a análise foram consideradas quatro cidades, sendo: Curitiba, Belo Horizonte, São Paulo e Rio de Janeiro. Questões interessantes relacionadas aos hábitos de consumo de cerveja foram analisadas por Chorley et al. (2016) também a partir de dados coletados do aplicativo *Untappd*. Esses autores, coletaram dados publicamente

acessíveis, em tempo real, disponibilizados pela API do aplicativo *Untappd*. Durante 112 dias foram coletados dados relacionados ao consumo de cervejas em 40 locais, entre EUA e Europa.

Chorley et al. (2016) fazem referência aos tipos de limitações presentes neste tipo de análise. Uma delas é referente ao fato de não haver uma forma de garantir a veracidade dos *checkins* realizados. Além disso, os dados são coletados somente se o *post* estiver em modo público. Em muitos casos, a fim de garantir a privacidade, os usuários utilizam publicações somente em modo privado - não sendo possível obter acesso a este tipo de dado. Ainda como limitação para a pesquisa, Chorley et al. (2016) relatam o fato de não ser possível afirmar se o usuário fez o *checkin* no exato momento em que estava consumindo a cerveja, como também, não ter acesso a informação com relação a quantidade de bebida consumida. O público para a pesquisa fica restrito somente entre os usuários de *Smartphones* que instalaram o aplicativo *Untappd*.

Apesar de todas as limitações encontradas, Chorley et al. (2016) chegaram à conclusão que, embora os EUA possuíssem maiores quantidades de usuários, os europeus consumiam mais cervejas. Além disso, identificaram que poucos usuários são mais propensos a explorar diferentes tipos de cervejas. Os autores mostram também, com base em uma análise de *ranking*, que os usuários parecem ser relativamente positivos quanto às cervejas consumidas, pois a média de classificação foi em 4 pontos de 5.

Já Barajas, Boeing e Wartell (2017) tiveram como objetivo explorar de que forma pequenas cervejarias artesanais podem alterar a dinâmica de uma região, em questões de desenvolvimento e revitalização, sendo os primeiros a examinar empiricamente o relacionamento entre os bairros e as cervejarias artesanais. Os autores reforçam a importância do estudo realizado para que os planejadores urbanos reconheçam a importância das cervejarias artesanais na revitalização dos bairros, ao mesmo tempo em que protegem os residentes do potencial deslocamento. Para isso, os autores utilizaram uma base de dados contendo informações de início e fim de atividades das cervejarias desde 2004. Primeiramente, são descritas características dos locais onde as cervejarias atuam, utilizando dados de um censo dos Estados Unidos. Em seguida foi avaliado de que forma a decisão de implantação das cervejarias influencia nas mudanças da composição residencial, além de explorar diferenças em escala regional e sub-regional.

Enfim, são feitas algumas sugestões de planejamento urbano, considerando as instalações de cervejarias artesanais. Muitas cidades possuem a iniciativa de incentivar a instalação de cervejarias artesanais pelo fato de haver um grande potencial de desenvolver novos empregos, catalisar os investimentos da região, além da possibilidade de se tornar um atrativo turístico (BARAJAS; BOEING; WARTELL, 2017).

Segundo Karamshuk et al. (2013), dados georreferenciados fornecidos pelos comércios podem viabilizar informações que permitam modelar o valor comercial de certas áreas urbanas. Neste contexto, Karamshuk et al. (2013) coletaram dados do *Foursquare*[5], a fim de entender

[5] Rede social baseada no uso de geolocalização dos aparelhos

como a popularidade de três redes varejistas em Nova York é formada analisando-se pela quantidade de *checkins*. Além disso, Karamshuk et al. (2013) categorizaram as informações coletadas em *"place-geographic"*, que buscam integrar informações sobre os tipos de locais e a interação espacial entre eles, e *"user mobility"*, que procura entender os movimentos dos usuários entre os locais.

A popularidade dos lugares pode ser melhor explicada pela fusão de recursos geográficos e de mobilidade, como a presença de estações de trem ou aeroportos nas proximidades de alguns comércios, atuando como atrativos de usuários. Além disso, as informações baseadas em localização disponibilizadas em redes sociais podem auxiliar na identificação do melhor local para estabelecer uma empresa (KARAMSHUK et al., 2013).

Estudos envolvendo a compreensão da dinâmica das cidades, em larga escala, são desafiadores. Envolvem horas de observação e entrevistas, normalmente resultando em uma noção parcial da realidade. A fim de propor uma solução a este problema, o artigo de Cranshaw et al. (2012) teve como objetivo estudar a dinâmica social da cidade de Pittsburg na Pennsylvania, com base nos dados de mídias sociais gerados pelos moradores. Através dos aspectos geoespaciais dos dados coletados das centenas de milhares de pessoas, os autores desenvolveram um algoritmo capaz de mapear as diversas regiões de uma cidade, com base nos padrões de atividades das pessoas. Para isso, foram coletados aproximadamente 18 milhões de *checkins* pelo *twitter* e, após isso, alinhados com as informações disponibilizadas na API do *Foursquare*.

O estudo de Cranshaw et al. (2012) apresenta uma forma de visualizar e investigar a dinâmica, estrutura e características de uma cidade, assumindo que pessoas e locais definam o caráter de uma cidade. Os resultados apresentam características quase em tempo real dos aspectos sociais das pessoas, nas diferentes partes das vizinhanças. Os autores apresentam algumas limitações e dificuldades encontradas durante o processo, como da amostra obtida por meio do *twitter* e o viés existente nos locais frequentados e do público que utiliza esse sistema. A fim de corrigir os possíveis erros causados por essas limitações, foram elaboradas entrevistas com os moradores, para confirmar a veracidade das informações encontradas.

Entre os resultados encontrados, estão diversos agrupamentos formados na cidade, que permitem diferenciar as áreas existentes. Os agrupamentos formados auxiliam a compreensão da dinâmica da cidade, com base nos dados que as pessoas geram nas mídias sociais. Cranshaw et al. (2012) citam que este estudo não apresenta apenas as divisões conhecidas, mas também revela mudanças nos padrões sociais locais e os efeitos que eles apresentam sobre a cidade.

Conforme demonstrado por Ho (2017) e Jones et al. (2015), o uso de *big data* para planejamento urbano tem chamado atenção de governos e dos próprios cidadãos. Do lado do governo o uso de *big data* tem se mostrado uma forma rápida e barata de obter informações atualizadas durante o processo de tomada de decisões. Já para os cidadãos, a possibilidade de

https://www.foursquare.com

influenciar na tomada de decisões é o principal fator que os levam a participar de estudos ou utilizar aplicativos para este fim.

O aplicativo *MapLocal* permitiu um melhor levantamento de informações básicas para o planejamento urbano de dois bairros de Birmingham na Inglaterra. Existe a possibilidade de pouca utilização dos dados coletados ou da diminuição de sua importância por influência de fatores externos. Ainda assim, a simples demonstração que envolver as pessoas dentro de seu próprio contexto diário com o uso de um aplicativo é prático e possível já se trata de um avanço para o uso mais abrangente da Computação Social (JONES et al., 2015).

3 Metodologia

A primeira etapa para a elaboração desta pesquisa foi o levantamento bibliográfico, onde foi possível adquirir informações de trabalhos relacionados ao tema do presente estudo, assim como garantir o entendimento das ferramentas e métodos de mineração de dados a serem utilizados. Os trabalhos de Silva e Graeml (2016), Chorley et al. (2016) e Karamshuk et al. (2013), já citados anteriormente, serviram como base para o entendimento do uso do *Untappd* e das possíveis informações que podem ser extraídas dos dados gerados pelo seu uso.

Além do levantamento bibliográfico, na primeira etapa do estudo está contida a elaboração dos *scripts* em linguagem Python responsáveis pela coleta de dados da API do *Twitter*. Esta coleta, permitiu uma réplica da análise proposta por Silva e Graeml (2016), que segue uma ideologia semelhante à deste trabalho. Entretanto, a principal motivação desta análise prévia foi de estudar a viabilidade desse tipo de análise para geração de informações relevantes construídas por meio de colaboração coletiva e KDD.

A segunda etapa consistiu na criação das bases de dados utilizadas para a pesquisa, além da análise aprofundada dos dados em si. O que diferencia a análise nesta etapa da anterior são as novas informações, obtidas durante o período de coleta dos dados, além das técnicas de mineração de dados utilizadas.

3.1 Sobre o aplicativo

O *Untappd* foi escolhido como objeto de estudo desta pesquisa dada a sua relevância entre os apreciadores de cervejas. O aplicativo permite que seus usuários compartilhem informações e interajam sobre quais cervejas estão bebendo, bem como, o local, tornando-o assim uma rede social específica para os apreciadores de cervejas. Devido à grande adesão, suas características e especificidade em relação ao tipo de bebida, o *Untappd* se mostrou uma grande fonte de dados relevantes para análise e obtenção de informações a respeito do consumo de cervejas. Apesar de sua maior utilização ser na América do Norte e Europa, o *Untappd* é também muito utilizado no Brasil para interações sociais entre apreciadores de cervejas, o que o torna ideal para coletar informações sobre este tipo de bebida. Este mesmo aplicativo já foi utilizado por Silva e Graeml (2016) e Chorley et al. (2016) em pesquisas semelhantes à do presente estudo.

Além das características citadas anteriormente, o *Unttapd* permite integração com redes sociais como *Facebook*, *FourSquare* e *Twitter* para o compartilhamento dos *checkins* e conquistas (*badges*). Essas informações quando compartilhadas, pelos usuários do *Untappd*, em forma de *Tweets*, pequenos textos de 140 caracteres compartilhadas no *Twitter*, podem ser facilmente coletadas e armazenadas para posterior análise. Isso se deve à facilidade e gratuidade de cadas-

tro na API do *Twitter*. Portanto, assim como no estudo realizado por Silva e Graeml (2016), nesta pesquisa foram coletados e analisados os dados do *Untappd* compartilhados pelo *Twitter*.

A opção inicial de utilizar dados diretamente do *Untappd*, via API própria, foi considerada como a melhor forma de coleta de dados. Entretanto, conforme demonstrado na Figura 8, é informado na página de registro[1] para API que o *Untappd* não disponibiliza este acesso para fins de pesquisa e todas as requisições são avaliadas por sua equipe própria. Foram realizadas duas requisições de acesso com amplas explicações dos objetivos e possíveis ganhos para o próprio *Untappd* e ambas foram negadas.

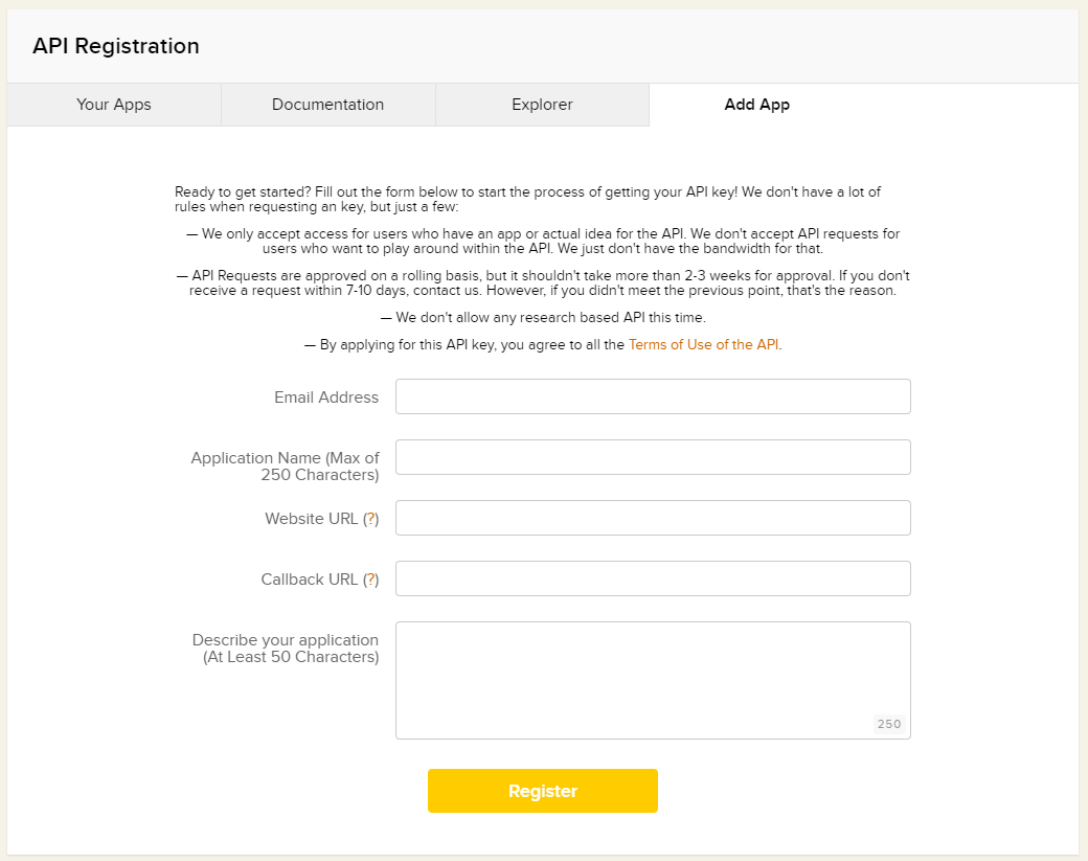


Figura 8: Registro para acesso a API do *Untappd*

Fonte: https://untappd.com/api/register

O uso do *Facebook* foi considerado, durante a etapa inicial desta pesquisa. Entretanto, o uso de ferramentas de busca automatizadas para coleta de dados em tempo real é restrita a editores de mídia, conforme demonstrado na Figura 9.

[1] https://untappd.com/api/register

Figura 9: API de *feed* público do *Facebook*

Fonte: https://developers.facebook.com/docs/public_feed

3.2 Coleta de dados

Embora sejam utilizados apenas os dados disponibilizados pelo *Twitter*, as informações geradas de um subgrupo do total de usuários do *Untappd*, conforme demonstrado por Preoţiuc-Pietro e Cohn (2013), em um estudo sobre os usuários do *Foursquare* em que se utilizam apenas dados do *Twitter*, este subconjunto é relevante para análises de dados. Isto ocorre por englobar uma diversidade de usuários, que se diferenciam em comportamento, localização e faixa etária. Sendo assim, esses dados apresentam uma diversidade suficiente para iniciar estudos de forma rápida e barata com uso de conceitos de Inteligência Coletiva apoiados pelas tecnologias de Computação Social (PREOŢIUC-PIETRO; COHN, 2013). Ainda que não substitua outros estudos específicos e mais aprofundados, o uso desse tipo de dados pode ajudar no planejamento inicial e agregar dados para estudos mais complexos como posicionamento de marca, planejamento urbano, entre outros (GOVERS, 2011; JONES et al., 2015; HO, 2017).

Cada *tweet* coletado tem suas informações representadas em um objeto JSON, uma formatação leve de troca de dados em um objeto de texto que é completamente independente de linguagens de programação, tornando-o fácil para humanos e máquinas entenderem e manipularem. Sendo assim, os objetos JSON são uma forma estruturada de obter e armazenar dados. A coleta dos dados foi feita por meio de um programa desenvolvido em *Python*. Trata-se de uma linguagem de programação orientada a objetos, de código aberto e que encoraja a modularização e reutilização de códigos. A sua escolha também se deu em função de sua simplicidade para a tarefa de coleta de dados utilizando bibliotecas de funções prontas e gratuitas. Foram utilizadas as seguintes bibliotecas:

- *TwitterAPI*: esta biblioteca fornece métodos simplificados para autenticação e uso dos recursos disponibilizados pelo *Twitter* em sua API;
- *Simplejson*: esta biblioteca permite a manipulação dos dados coletados pelo Twitter, visto que cada *tweet* é obtido em formato *JSON*.

O *script* desenvolvido para a coleta de dados utiliza a função *request* da biblioteca *TwitterAPI* que permite especificar parâmetros, a fim de restringir a busca dos *tweets* para receber,

em tempo-real, os *tweets* publicados que se encaixem nos critérios estabelecidos. Com o objetivo de obter uma maior quantidade de dados contendo informações relevantes, foram feitos testes com alguns filtros no *Stream* do *Twitter* API. Alguns deles foram:

- "Untappd": O primeiro critério de busca a ser testado. Embora tenha apresentando uma grande quantidade de dados, não foi o critério adotado, pois falhou em mostrar dados com informações relevantes. Os dados apresentados por esse critério continham, em sua maioria, informações sobre os *Badges* (uma espécie de medalha dada aos usuários pelos seus *checkins*). E, por isso, não apresentou informações relevantes para a elaboração deste estudo.
- "untp it": Por meio de uma análise empírica foi constatado que em alguns dos *tweets* antigos essa era uma cadeia de caracteres relevante para a pesquisa, por se tratar de um *link* disponível em boa parte dos *tweets* enviados a partir do *Untappd*. Ao considerá-la para esta pesquisa, foi identificado que os *tweets* que estavam sendo coletados continham a palavra 'untp' fazendo referência à URL e a palavra 'it', do inglês - usada para definir coisas ou pessoas.
- "untp beer": A partir de uma breve observação de alguns dos resultados obtidos com o filtro citado no tópico anterior, foi identificado que grande parte dos *tweets* enviados através do aplicativo *Untappd* possuíam uma URL no formato 'untp.beer/(...)'. Após a realização de testes com esse critério, foi identificado que seria adequado usá-lo. Pois esse critério atende o objetivo de obter uma quantidade significante de dados que contenham informações relevantes.

Dentro deste parâmetro adotado como filtro estão contidos *tweets* referentes aos *badges* alcançados pelos usuários e anúncio de novas cervejas adicionadas aos estoques das revendas cadastradas no *Untappd*. Estes itens foram salvos separadamente para posterior análise, caso fossem encontrados padrões relevantes.

Conforme abordado por Silva e Graeml (2016), a maioria dos *tweets* enviados por meio do *Untappd* segue um padrão. Um dos *tweets* coletados trazia como mensagem: "Medalha de ouro!!! - Drinking a Cacau Wee by @bodebrown at @riodejaneiro — https://t.co/NQd7fI6LYE #photo". Ao analisar esse *tweet* é possível identificar que após a palavra *Drinking* é informado o tipo de cerveja sendo consumida pelo usuário, após o *by* é informada a empresa fabricante da cerveja e, por fim, após o *at* é informado o local onde o usuário está consumindo a cerveja.

Considerando a estrutura dos *tweets* é possível adquirir informações como: quais são as cervejas consumidas em determinadas cidades, qual a cervejaria que faz mais sucesso em determinadas regiões e explorar características dos usuários do aplicativo. Essas possibilidades foram exploradas em uma análise preliminar, onde o objetivo foi de replicar a análise realizada por Silva e Graeml (2016). Ainda, é possível encontrar nos *tweets* diversas informações como:

data e hora do compartilhamento, coordenadas geográficas, identificação do usuário, entre outros.

Sendo assim, para garantir maior precisão nas análises de geolocalização foram escolhidos apenas os *tweets* que apresentavam coordenada geográfica em seus dados, isto é, incorporada ao JSON. Todos os *tweets* georreferenciados foram então apresentados no Mapa Mundi, em formato de *Heatmap*, para facilitar a identificação do volume de *tweets* de cada região em nível mundial. Isso auxiliou a seleção das regiões relevantes para a análise.

3.3 Demarcação de cidades para o estudo

A partir das observações visuais de cada região, auxiliadas pelos *Heatmaps*, os *tweets* das cidades escolhidas foram separados para a etapa de incremento de informações, conforme descrito na seção à seguir.

Foram consideradas as coordenadas de latitude e longitude que fossem suficientes para englobar cada cidade, evitando abranger a área de outras cidades quando tratou-se de regiões metropolitanas com alta densidade populacional. Cada cidade é representada por dois pontos, sendo o ponto mais ao norte e mais ao oeste dentro das condições estabelecidas; e outro mais ao sul e mais ao leste; descrevendo assim um quadrilátero da região de interesse que delimita a área na qual os *tweets* geolocalizados serão considerados válidos para esta pesquisa. Essa delimitação do espaço geográfico em um retângulo, é também conhecida por *Minimum Bounding Rectangle*, que limita a geometria de uma característica geográfica ou a coleção de geometrias em um conjunto de dados geográficos (KEMP, 2007).

3.4 Incremento e tratamento dos dados

Após a escolha de um par de coordenadas para representar os limites geográficos de cada cidade, foi executado um algoritmo para incremento das informações de cada *checkin* que estivesse dentro dos limites estabelecidos pelo par de coordenadas geográficas.

Após definir as regiões de interesse foram realizadas consultas ao *site* do Untappd, para cada *tweet* coletado, a fim de garantir informações mais precisas dos dados de cada *checkin*. Além disso, as consultas permitiram o enriquecimento de informações a respeito daquele *checkin*, possibilitando a elaboração de análises mais aprofundadas sobre o tema.

Devido à facilidade de manipulação das informações nos objetos JSON, utilizando *scripts* escritos em *Python*, todos os *tweets* coletados foram armazenados em arquivos de texto (*.txt*), o que eliminou a necessidade de um Sistema Gerenciador de Banco de Dados.

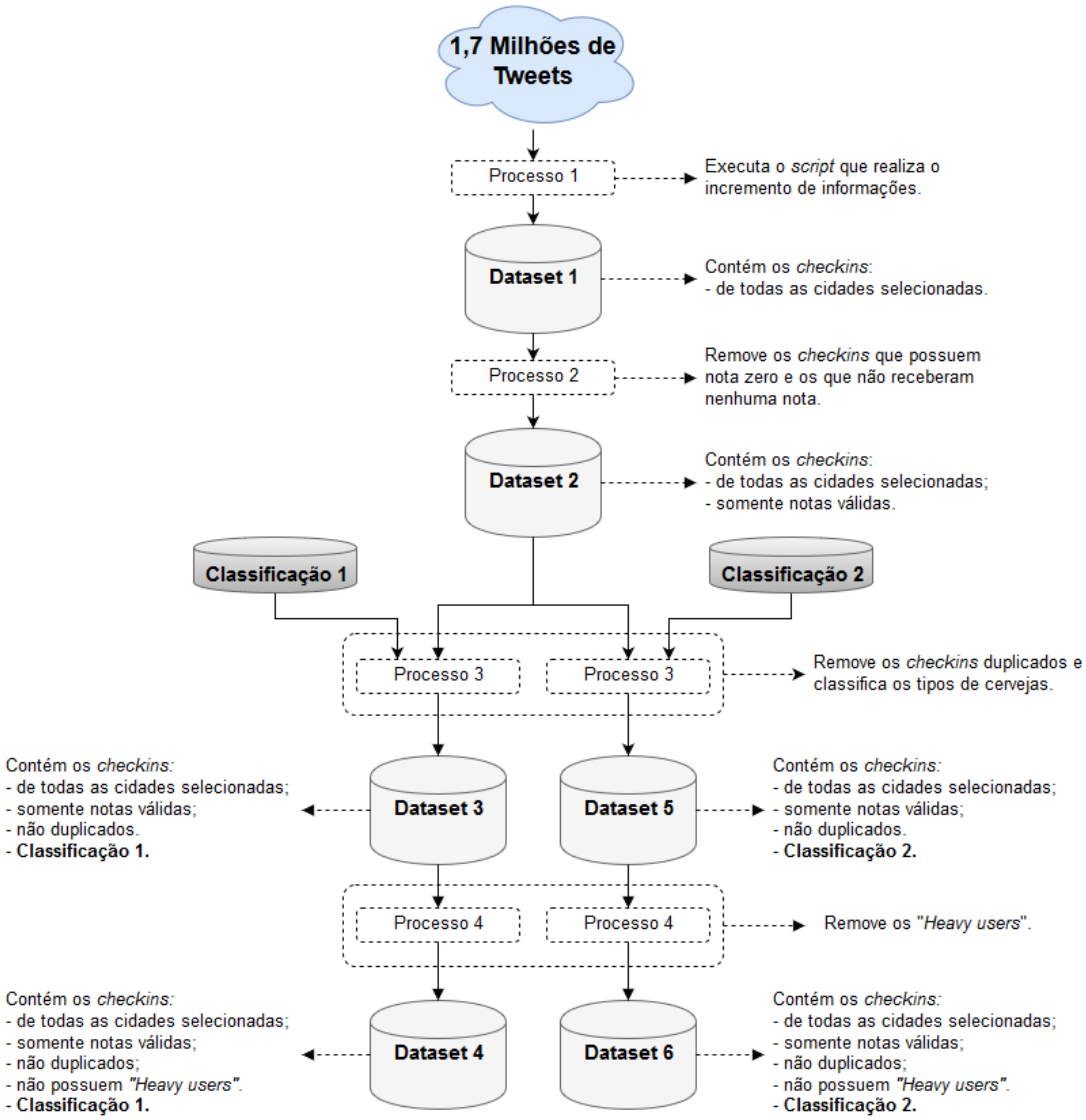


Figura 10: Diagrama do processo de construção das Bases de dados

O diagrama acima é explicado com maiores detalhes nos próximos parágrafos.

Após encerrado o processo de coleta dos dados, a base completa foi submetida a processos de filtragem e classificação com objetivo de diminuir o ruído e gerar informações mais próximas da realidade. O Processo 1 foi realizado sobre a totalidade dos dados armazenados e sua função foi incrementar os dados disponíveis em cada *checkin* que se localizasse dentro das cidades de interesse. Esse incremento de informações se deu através da consulta automática de dados de cada *checkin* do *site* do *Unttapd*. Esse processo gerou um *dataset* temporário aqui chamado de *Dataset* 1.

Identificou-se a necessidade de, primeiramente, eliminar as notas zeradas. Isto deve-se à impossibilidade de identificar quando um usuário realmente atribuiu uma nota zero ou esqueceu de avaliar uma cerveja fornecendo uma nota a ela. Este fato pode vir a gerar falsos positivos, impactando de forma negativa os resultados e análises da pesquisa. Essa filtragem se deu por

meio do Processo 2 sobre o *Dataset* 1 que gerou um novo *dataset* temporário, *Dataset* 2.

Para a classificação das cervejas quanto aos seus tipos, foram utilizadas duas metodologias. A primeira, Classificação 1, foi realizada utilizando todos os tipos de cervejas agrupando os que eram semelhantes entre si. A segunda forma, Classificação 2, foi realizada de acordo com o material disponibilizado pela *Brewers Association* (ASSOCIATION, 2017). Essa segunda metodologia utilizada tem como objetivo agrupar as cervejas por características étnicas, o que pode beneficiar a análise de agrupamento.

Com o objetivo de eliminar dados que caracterizassem favoritismo individual foram eliminados os *checkins* que o usuário repetisse a mesma cerveja. Essa repetição acontece quando a tupla *UsedID*, cerveja e cervejaria se repete dentro da mesma cidade. O Processo 3 também é responsável pelo agrupamento dos tipos de cerveja em grupos, o *Unttapd* possui em seu *site* 177 tipos diferentes de cervejas. As classificações 1 e 2 possuem respectivamente 93 e 10 grupos nos quais os tipos existentes são agrupados. São gerados, após essa filtragem e agrupamento, os *Datasets* 3 e 5 utilizando-se as Classificações 1 e 2 respectivamente.

O Processo 4 executou a função de remover dos *Datasets* 3 e 5 os chamados *"heavy users"*, usuários assíduos do aplicativo que apresentam um número individual de *checkins* muito distante do número individual de *checkins* da maioria dos usuários de cada cidade. O critério adotado foi descartar os usuários que apresentassem um número de *checkins* maior que a soma da mediana e desvio padrão do número de *checkins* por usuário em cada cidade. Ou seja, cada cidade tem um limite específico para considerar um usuário como *"heavy user"*. Este processo gerou a partir dos *Datasets* 3 e 5, respectivamente os *Datasets* 4 e 6 que serão analisados em conjunto com seus predecessores, que possuem os *"heavy users"*, de forma a entender o impacto desses usuários nos modelos de clusterização e na caracterização de preferências de cada cidade.

4 Resultados

Conforme já apresentado, o artigo proposto por Silva e Graeml (2016) tinha como objetivo simular uma pesquisa, onde seria fornecido o suporte para uma micro cervejaria com intuito de encontrar o melhor local para sua instalação. O presente estudo tem por objetivo inicial recriar esta análise comparando resultados, a fim de comprovar as conclusões a respeito do perfil do consumidor de cerveja artesanal encontradas por Silva e Graeml (2016). Contudo, este estudo também se propôs a ir além, visando identificar semelhanças e diferenças entre algumas cidades e países, através de dados mais atualizados e em maior quantidade. Ainda, fornecer informações diversas a respeito do perfil de consumo de cervejas artesanais. Na próxima seção será descrito o trabalho realizado em uma análise preliminar.

4.1 Análise preliminar

Para a realização da análise preliminar foram coletados dados por 15 dias, em Outubro de 2016, totalizando aproximadamente 106 mil *tweets*. Uma quantidade muito próxima à descrita pelos autores Silva e Graeml (2016). Embora uma grande quantidade de *tweets* tenha sido coletada, para efeitos da análise, foram considerados somente os *tweets* que continham informações de geolocalização. A Figura 11 apresenta os locais de todos os *tweets* geolocalizados coletados durante esse período.

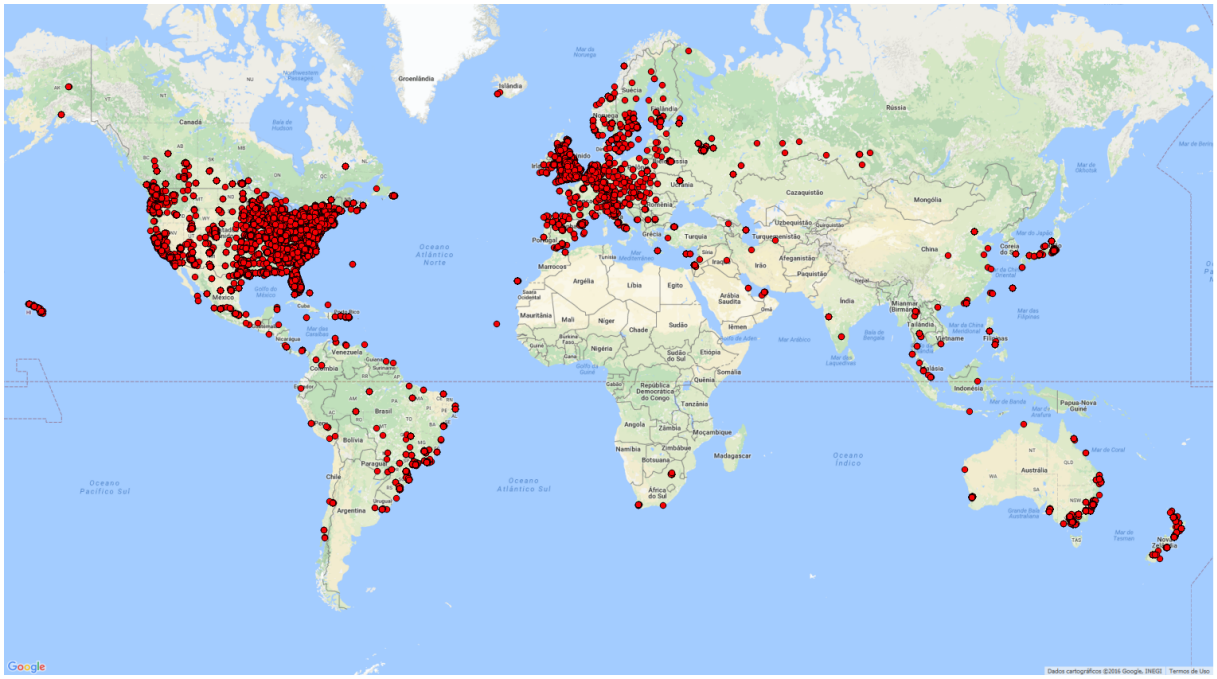


Figura 11: Untappd no mundo - Análise preliminar

Assim como constatado por Silva e Graeml (2016), é possível perceber que o Untappd é um aplicativo muito popular nos Estados Unidos e Europa. Embora o Brasil não forneça uma grande quantidade de dados na base, ainda é capaz de fornecer informações valiosas sobre o perfil do consumidor de cerveja artesanal. Conforme apresentado pelos autores do artigo, os *tweets* não são distribuídos uniformemente pelo país. Grande parte desses dados encontram-se no sudoeste do país, e com maior frequência em grandes capitais como: São Paulo - Capital (38,9%), Rio de Janeiro (11%), Brasília (8,2%), Belo Horizonte (4,4%), Porto Alegre (4,1%) e Curitiba (3%).

Seguindo a mesma técnica adotada por Silva e Graeml (2016), foi possível identificar algumas das preferências dos consumidores de cada uma das quatro cidades através do mapa de palavras. O uso desta ferramenta é recomendado para uma análise preliminar, pois os mapas apresentados a seguir não são proporcionais entre si e somente destacam as palavras que aparecem com maior frequência na base de dados.

(a) Cervejas - Curitiba

(b) Cervejas - Belo Horizonte

(c) Cervejas - São Paulo

(d) Cervejas - Rio de Janeiro

Figura 12: Mapa de palavras com as cervejas

Quando comparadas as quatro cidades com o trabalho realizado por Silva e Graeml (2016), é possível perceber que houveram mudanças na forma de consumo de cerveja. As cidades de Curitiba e Belo Horizonte apresentam menores variedades quanto aos tipos de cervejas, se comparadas a São Paulo e Rio de Janeiro. Isso deve-se ao fato de que as duas primeiras cidades representam apenas 3% e 4,4%, respectivamente, da base de dados referente ao Brasil.

Se comparadas à pesquisa realizada por Silva e Graeml (2016) a cidade de Curitiba

continua chamando atenção por ter um público mais sofisticado, dados os tipos de cervejas contidos nos *tweets* desta região. Por outro lado, Belo Horizonte sofreu pequenas mudanças nos padrões de consumo, apresentando maiores quantidades de cervejas de escala industrial que no estudo anterior.

Ainda, é possível perceber que no Rio de Janeiro, mesmo com algumas das cervejas industriais, uma maior variedade de cervejas artesanais fazem parte do mapa de palavras. Contudo, não é uma mudança significativa no padrão de consumo dos usuários do *Untappd*. Pois no artigo de Silva e Graeml (2016) isso já acontecia. Por fim, São Paulo foi outra cidade que chamou bastante a atenção devido à diversidade de cervejas artesanais ou premium compartilhadas. Ressaltando uma possível mudança na forma de como os usuários estão consumindo as cervejas. No artigo anterior era visível a preferência por cervejas industrializadas, enquanto no presente estudo torna-se claro que as artesanais e premium estão ganhando espaço no mercado cervejeiro de São Paulo.

Uma das possibilidades que chamaram a atenção na análise preliminar foi a de que possivelmente alguns dos *tweets* coletados do Rio de Janeiro venham a ser de turistas passeando naquela região, já que o mapa de palavras apresenta um grande percentual de *checkins* com "Rio de Janeiro" como o local onde os usuários estariam consumindo suas cervejas. Além disso, o Rio de Janeiro foi o local que apresentou maior percentual de locais residenciais. Isto pode indicar que o público do Rio de Janeiro é mais suscetível a consumir suas cervejas em suas próprias residências aos bares da região. Para a análise preliminar essa situação não foi considerada como um problema, portanto, foram considerados os diversos *tweets* coletados sem realizar filtros quanto aos locais de consumo.

Os mapas de palavras, na Figura 13, apresentam os locais onde os usuários marcaram que estavam consumindo suas cervejas. Se combinadas as informações do mapa de palavras de tipos de cervejas e dos locais, torna-se possível perceber que Curitiba e Belo Horizonte possuem uma grande diversidade de locais onde são oferecidas cervejas artesanais. Uma vez que poucos *tweets* mencionando cervejas industrializadas populares no mercado brasileiro foram encontrados. Outra observação interessante é de que possivelmente os consumidores de cerveja artesanal em São Paulo são mais fiéis a locais específicos, entretanto, são propensos a explorar novos tipos de cervejas. Isso é possível pois conforme as informações do site e redes sociais do local mais popular na base de dados de São Paulo, "Empório Alto dos Pinheiros" , possui 33 torneiras de chope e 650 rótulos diferentes. Isso pode ter influenciado a análise, visto que uma grande variedade de cervejas artesanais foi encontrada nos *tweets* desta região, enquanto que uma pequena variedade de locais foi encontrado na base de dados.

(a) Locais - Curitiba

(b) Locais - Belo Horizonte

(c) Locais - São Paulo

(d) Locais - Rio de Janeiro

Figura 13: Mapa de palavras dos locais

Através da análise preliminar, foi possível identificar mudanças significativas no perfil de consumo de cerveja artesanal em todas as cidades analisadas. As mudanças apresentadas talvez não representem todo o mercado cervejeiro, dadas as condições do aplicativo usado para a realização desta análise. Entretanto, foi possível observar que o consumo da cerveja artesanal em alguns locais do Brasil vem aumentando - considerando a comparação entre o período de 2012 à 2016. Sendo assim, essas informações podem apresentar aspectos relevantes em uma análise mais aprofundada.

4.2 Descrição e período da coleta

Para a realização deste estudo foram coletados dados por aproximadamente 6 meses, totalizando 1,7 milhões de *tweets*. O perído de coleta compreendeu meses desde Novembro de 2016 à Abril de 2017. O mês de Novembro resultou em aproximadamente 385 mil *tweets* coletados, Dezembro, por sua vez, 444 mil - sendo o mês com maior número de registros coletados. Em Janeiro de 2017 foram somente 151 mil. Conforme apresentado na Figura 14, os meses de Fevereiro, Março e Abril, respectivamente, 333 mil, 194 mil e 284 mil *tweets* coletados. Vale ressaltar que os dados descritos nesta seção, e posteriormente na análise, não englobam os descritos na seção 4.1.

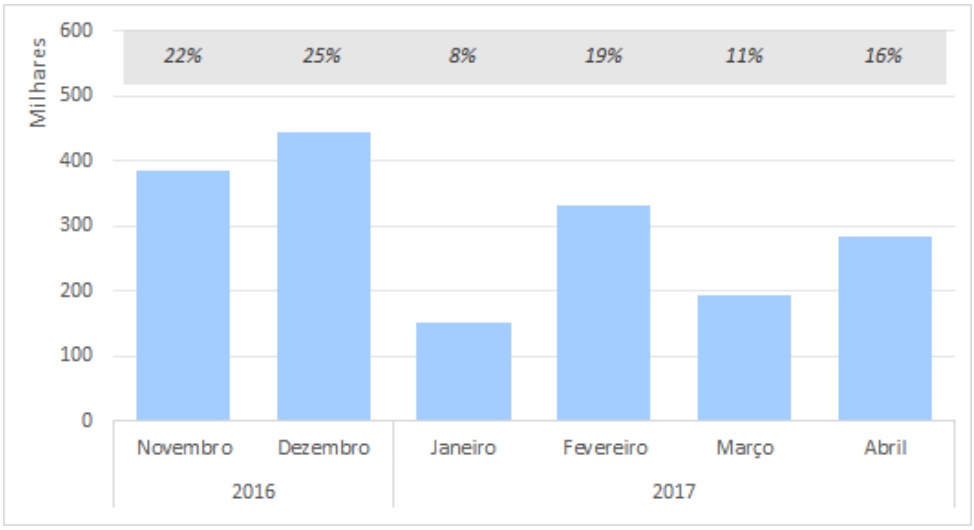
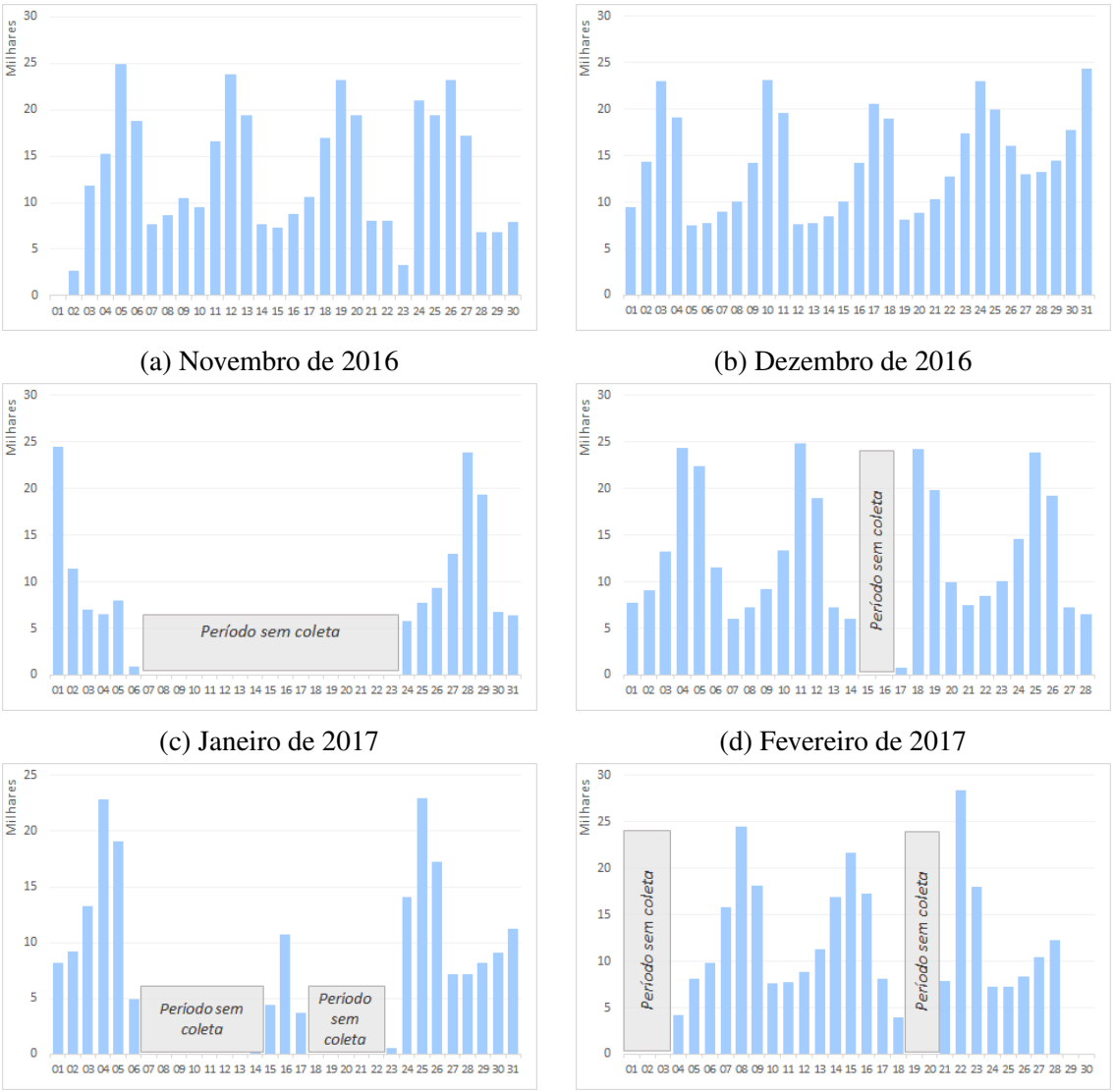


Figura 14: Quantidade de registros coletados por mês.

A baixa quantidade de *tweets* no mês de Janeiro de 2017 é justificada pela interrupção do *script* de coleta devido à falta de espaço no servidor. Problemas semelhantes ocorreram nos meses consecutivos, entretanto, não ocasionado pela necessidade de espaço, mas por outros fatores que resultaram na interrupção da coleta. Os períodos de interrupção podem ser observados abaixo nas Figuras 15c, 15d, 15e e 15f, onde é apresentada a quantidade de registros coletados por dia nos respectivos meses.



(a) Novembro de 2016

(b) Dezembro de 2016

(c) Janeiro de 2017

(d) Fevereiro de 2017

(e) Março de 2017

(f) Abril de 2017

Figura 15: Quantidade de registros coletados diariamente.

Contudo, esse período de coleta resultou em grandes quantidades de dados com diversas informações. Parte desses *tweets* coletados não possuem georreferenciamento e, portanto, foram desconsiderados para esta análise. A fim de auxiliar a visualização dos dados foi elaborada a Figura 16 que apresenta os locais de todos os *tweets* com geolocalização, o que corresponde à aproximadamente 702 mil *tweets* ou 39,2% do total de *tweets* coletados.

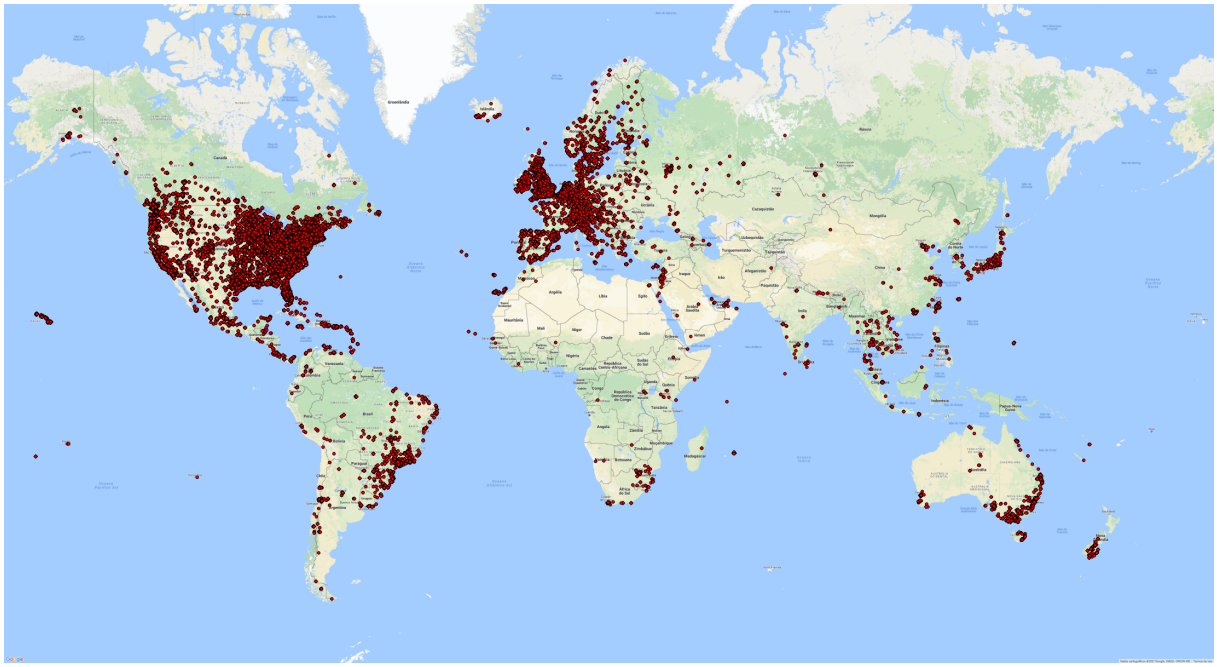


Figura 16: Untappd no mundo

Assim como constatado por Silva e Graeml (2016), é possível perceber que o Untappd é um aplicativo muito popular nos Estados Unidos e Europa. Possivelmente, a baixa adesão ao aplicativo por parte de alguns países pode ser explicado pela interface ser disponibilizada somente na língua inglesa, o que torna necessário o conhecimento deste idioma para conseguir efetivamente utilizá-lo. Sendo assim, alguns países, como o Brasil, não possuem uma quantidade expressiva de dados na base, entretanto, ainda é possível extrair informações valiosas sobre o perfil do consumidor de cerveja artesanal.

A fim de identificar quais regiões apresentavam maior densidade de *checkins* foi construído um *Heatmap* de acordo com parâmetros que possibilitassem gerar uma visualização onde destaca-se apenas as regiões com consumo relevante, viabilizando, assim, o presente estudo. Na Figura 17 é apresentado o *Heatmap* da quantidade de *checkins* coletados do *Twitter* para esta pesquisa em nível mundial.

O *Heatmap* utilizado neste contexto tem por objetivo permitir a visualização da quantidade de *checkins* por região. Sendo assim, é possível considerar a densidade de *checkins* através do tamanho e cor do círculo vermelho, sendo que este, com uma cor mais forte representa maior quantidade de dados naquele local.

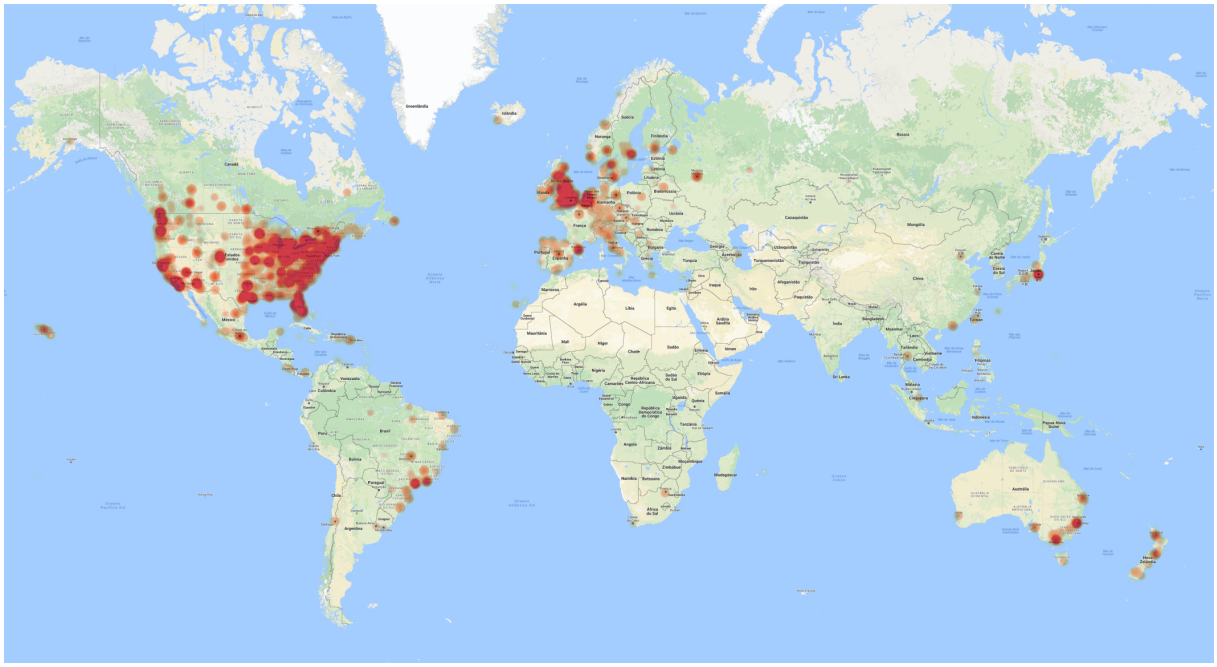


Figura 17: *Heatmap* do Untappd no mundo

Embora existam diversos *tweets* coletados em todo o mundo, conforme visualizado na Figura 16, parte das cidades não apresentam uma densidade considerável de *checkins* para compor uma base de dados confiável, como é possível identificar na Figura 17. Garantir um número mínimo de observações torna-se um passo importante na elaboração da base, pois a quantidade de dados em uma amostra pode vir a interferir na confiabilidade do estudo.

4.3 Cidades selecionadas

Tendo em vista a grande quantidade de dados coletados, torna-se necessário eliminar parte das informações com o objetivo de selecionar somente os conteúdos relevantes para a composição da análise.

Com base no *Heatmap* da Figura 17, utilizado para apoiar a tomada de decisão quanto a seleção de cidades para este estudo. O critério predominante para a seleção está relacionado basicamente à quantidade de *checkins* obtidos em cada região durante o período de coleta. Através da Tabela 1 é possível visualizar a relação de cidades escolhidas pela quantidade de *checkins*.

Continent	**País**	**Cidade**	**Quantidade**
América Central	México	Cidade do México	1.004
América do Norte	EUA	Chicago	6.493
		Los Angeles	2.210
		Nova York	8.080
		Portland	6.451
		São Francisco	3.062
América do Sul	Argentina	Buenos Aires	101
	Brasil	Belo Horizonte	436
		Curitiba	465
		Rio de Janeiro	1.070
		São Paulo	2.555
	Chile	Santiago	146
Ásia	Japão	Osaka	157
		Tokyo	2.764
	Singapura	Singapura	120
	Tailândia	Bangkok	222
Europa	Alemanha	Berlim	771
	Bélgica	Bruxelas	1.160
	Espanha	Barcelona	1.877
		Madrid	770
	França	Paris	383
	Holanda	Amsterdã	1.556
	Irlanda	Dublin	837
	Reino Unido	Londres	6.658
	República Tcheca	Praga	476
	Rússia	Moscou	857
Oceania	Austrália	Melbourne	1.808
		Sydney	2.597

Tabela 1: Tabela da quantidade de *checkins* das cidades escolhidas

A região dos Estados Unidos foi uma das que apresentou grande densidade de *checkins*, totalizando 26.296 entre as cidades selecionadas. Isto deve-se à vasta utilização do aplicativo pelos apreciadores de cervejas artesanais. Em todo o país outras cidades poderiam vir a ser selecionadas, tendo em vista a necessidade de uma quantidade razoável de *checkins* para compor uma análise confiável - é possível visualizar na Figura 34 no anexo A. Entretanto, destacaram-se cinco cidades, são elas: Nova York; Chicago; Portland; São Francisco e Los Angeles. Esta seleção permite a identificação de caracaterísticas específicas das regiões em que se encontram, existindo a possibilidade de caracterizar preferências específicas para cada região.

Outra região que apresentou grandes quantidades de *checkins* foi o continente Europeu - Figura 35 do anexo A. Entretanto, conforme mencionado anteriormente, a grande quantidade de *checkins* coletados nesta região não indica a existência de uma densidade considerável para o estudo. Sendo assim, a seleção foi realizada de forma que atendesse ao critério de quantidade de *checkins* prevalecendo também a diversidade cultural, que é possível visualizar na Tabela 1. Essa seleção resultou em um total de 15.345 *checkins*, composto pelas seguintes cidades: Berlim; Bruxelas; Barcelona; Madrid; Paris; Amsterdã; Dublin; Londres; Praga e Moscou.

Embora existam diversas cidades com quantidades aceitáveis de *checkins*, para a realização de uma análise adequada, estas não foram incluídas na seleção devido à limitação durante a fase de uso do *crawler* além do fato de ter sido priorizada a diversificação na escolha das cidades. Um exemplo é o Reino Unido no qual foi escolhida apenas a cidade de Londres, que em uma análise preliminar já garantiria uma boa representação dessa região.

No continente das Américas, a América do Sul apresenta quantidades relevantes de *checkins* - Figura 36 disponível no anexo A. Conforme apresentado por Silva e Graeml (2016), os *tweets* não são distribuídos uniformemente no Brasil. Grande parte dos dados são encontrados no sudoeste do país e com maior frequência em grandes capitais como: São Paulo (Capital), Rio de Janeiro, Brasília, Belo Horizonte, Curitiba, Florianópolis e Porto Alegre. Foram selecionadas para a composição deste trabalho somente quatro cidades com grandes densidades de *checkins*, além de levar em consideração a mesma seleção descrita por Silva e Graeml (2016), sendo elas: Belo Horizonte, Curitiba, Rio de Janeiro e São Paulo, totalizando 4.526 *checkins*.

Neste contexto, ainda na América do Sul foram considerados para a composição deste estudo os países: Argentina e Chile, respectivamente as cidades de Buenos Aires e Santiago - Figura 36 no anexo A. As cidades em questão foram incluídas inicialmente na seleção pois acreditou-se que viessem a favorecer aspectos culturais na análise, mesmo contendo pequenas quantidades de *checkins*.

Por fim, no continente da Oceania, considerou-se a Austrália como parte dos objetos em análise no presente estudo - Figura 37 do anexo A. Dada a relevância em quantidade de *checkins* foram selecionadas as duas maiores cidades deste país, com um total de 4.405 *checkins*: Melbourne e Sidney.

Para o continente Asiático, nas Figuras 38 e 18, é possível justificar a escolha de quatro cidades dadas as quantidades de *checkins*: Osaka e Tokyo, no Japão; Singapura e Bangkok, capital da Tailândia. Estas quatro cidades resultaram em 3.263 *checkins*.

Figura 18: *Heatmap* da densidade de *checkins* no Japão

A Figura 18 representa o Japão, nela destacam-se três pontos com grandes quantidades de *checkins*. Estes pontos representam as cidades de Tokyo e Osaka e, além dessas, Niigata que inicialmente foi considerada relevante para este estudo. Entretanto, a análise dos dados provenientes desta cidade, constatou que os 180 *checkins* foram gerados por um único usuário. Optou-se por eliminar esta cidade da análise, pois esta poderia vir a causar divergências no resultado levando em consideração apenas o gosto deste usuário.

Para a América Central foi selecionado somente uma cidade, a Cidade do México, totalizando 1.004 *checkins*, conforme a Figura 39 do anexo A.

O objetivo em escolher estas cidades está na expectativa do resultado a ser obtido através dos métodos de agrupamento. Espera-se que nesta etapa, países de um mesmo continente ou cidades com alguma semelhança venham a permanecer próximas umas às outras. Isso deve-se à possibilidade de existir semelhanças nas preferências quanto ao tipo de cerveja consumida. O que pode vir a enfatizar aspectos culturais associados ao consumo de cerveja das determinadas regiões.

4.4 Preparação dos dados para análise

A seguir são mostrados os números comparativos entre os Datasets 3 e 4, conforme o Processo 4 do Diagrama de construção dos Datasets de acordo com a Figura 10. O Dataset 5 passou pelo mesmo Processo 4 resultando no Dataset 6. Esse processo resultou na remoção de 490 *"heavy users"* que representam 8% to total de usuários analisados. Entretanto, há de se notar que esses usuários são responsáveis por 57% dos *checkins* analisados, por isso algumas

das análises dos Datasets 4 e 6 devem ser realizadas com cautela quanto a representatividade real de seus resultados.

Sendo assim, para a remoção dos *"heavy users"* foi adotado um critério específico onde foi determinado um "Limite de corte", conforme representado na Tabela 2. Ou seja, os usuários com quantidades de *checkins* maiores que esse limite foram removidos. Para a realização deste cálculo foi considerada a soma entre a Mediana e Desvio Padrão. Desta forma, através da mediana, seleciona-se o valor central de um conjunto de valores, onde encontram-se os consumidores médios de uma determinada localização. O mesmo não acontece com a média, dado que através desta, seria considerada a quantidade de *checkins* de todos os usuários, incluindo os *"heavy users"*. É possível identificar este efeito quando analisa-se Moscou, por exemplo, em que a média e desvio padrão resultariam em 77,7 contra os 60,7 *checkins* estabelecidos. Portanto, este método foi adotado pois constatou-se que através dele os *Heavy-users* seriam removidos de maneira mais eficaz, contemplando assim, neste cenário, os usuários comuns.

Cidade	**Datasets 3 e 5**			**Limite de Corte**	**Datasets 4 e 6**		
	Média	**Mediana**	**Desvio Padrão**		**Média**	**Mediana**	**Desvio Padrão**
Amsterdã	8,4	2,0	26,4	28,4	3,8	2,0	4,3
Bangkok	6,0	1,0	13,0	14,0	2,7	1,0	3,2
Barcelona	19,0	3,0	49,7	52,7	7,6	3,0	10,6
Belo Horizonte	9,5	5,0	10,5	15,5	4,6	3,0	4,0
Berlim	6,9	3,0	12,0	15,0	4,0	2,0	3,6
Bruxelas	10,3	3,0	24,6	27,6	4,3	3,0	4,6
Buenos Aires	4,4	1,0	7,0	8,0	2,2	1,0	1,9
Chicago	7,4	2,0	18,1	20,1	3,8	2,0	4,0
Cidade do México	8,0	3,0	11,9	14,9	3,6	2,0	3,3
Curitiba	8,6	3,0	14,1	17,1	4,4	2,5	4,2
Dublin	5,7	2,0	8,9	10,9	2,5	2,0	2,1
London	9,4	3,0	25,8	28,8	4,9	3,0	5,4
Los Angeles	6,1	2,0	15,5	17,5	3,2	2,0	3,1
Madrid	11,0	3,0	24,1	27,1	5,2	2,0	5,4
Melbourne	15,2	3,0	29,7	32,7	6,0	2,0	7,5
Moscou	22,0	5,0	55,7	60,7	7,7	4,0	11,8
Nova York	7,8	2,0	18,9	20,9	3,9	2,0	4,0
Osaka	5,2	3,0	6,1	9,1	3,1	2,5	2,3
Paris	4,9	2,0	8,7	10,7	2,8	2,0	2,4
Portland	15,7	3,0	44,5	47,5	6,9	3,0	8,8
Praga	8,5	2,5	12,6	15,1	3,8	2,0	3,8
Rio de Janeiro	9,6	4,0	15,1	19,1	4,6	3,0	4,8
Santiago	4,6	1,0	8,2	9,2	2,3	1,0	2,4
São Francisco	7,2	2,0	16,1	18,1	3,8	2,0	3,9
São Paulo	14,4	3,0	46,6	49,6	6,6	3,0	7,8
Singapura	3,8	2,0	4,7	6,7	2,4	2,0	1,8
Sydney	20,1	4,0	38,3	42,3	5,5	2,0	7,5
Tokyo	21,6	4,5	51,9	56,4	8,7	4,0	11,6
Total	**9,6**	**3,0**	**26,2**	**29,2**	**4,5**	**2,0**	**5,6**

Tabela 2: Dados estatísticos - Heavy users

Conforme Tabela 2 é notada uma redução na média e mediana do número de *checkins* por usuário em todas as cidades, após a execução do processo 4, já mencionado anteriormente. Esse efeito já era esperado pois foram removidos apenas usuários com alto número de *checkins*.

Porém o desvio padrão do número de *checkins* por usuário diminuiu de 26,2 para 5,6 o que representa uma amostra mais uniforme para as análises em questão.

Cidade	**Checkins Originais**	**Checkins Removidos**	**Checkins Restantes**	**Percentual Removido**
Amsterdã	1556	891	665	57,3%
Bangkok	222	130	92	58,6%
Barcelona	1877	1185	692	63,1%
Belo Horizonte	436	275	161	63,1%
Berlim	771	373	398	48,4%
Bruxelas	1160	714	446	61,6%
Buenos Aires	101	58	43	57,4%
Chicago	6493	3434	3059	52,9%
Cidade do México	1004	615	389	61,2%
Curitiba	465	252	213	54,2%
Dublin	837	530	307	63,3%
London	6658	3373	3285	50,7%
Los Angeles	2210	1143	1067	51,7%
Madrid	770	445	325	57,8%
Melbourne	1808	1183	625	65,4%
Moscou	857	588	269	68,6%
Nova York	8080	4332	3748	53,6%
Osaka	157	76	81	48,4%
Paris	383	186	197	48,6%
Portland	6451	3831	2620	59,4%
Praga	476	297	179	62,4%
Rio de Janeiro	1070	626	444	58,5%
Santiago	146	78	68	53,4%
São Francisco	3062	1573	1489	51,4%
São Paulo	2555	1449	1106	56,7%
Singapura	120	50	70	41,7%
Sydney	2597	1999	598	77,0%
Tokyo	2764	1760	1004	63,7%
Total	**55086**	**31446**	**23640**	**57,1%**

Tabela 3: *Checkins* removidos por cidade

A Tabela 3 mostra a quantidade de *checkins* removidos por cidade e o número de *checkins* restantes. Como consequência da remoção dos *"heavy users"* percebeu-se que algumas das cidades passaram a apresentar baixas quantidades de *checkins*. Visto que esse fato poderia prejudicar, em partes, a análise, foi realizada uma análise empírica a fim de determinar a necessidade de estabelecer um número mínimo de *checkins*. Portanto, a fim de garantir a representatividade de cada cidade nas análises, estabeleceu-se como mínimo trezentos *checkins*.

Sendo assim, algumas das cidades incialmente selecionadas não apresentaram a quantidade mínima de *checkins* mesmo antes da remoção dos *"heavy users"*, são elas: Bangkok, Buenos Aires, Osaka, Santiago e Singapura. Portanto, estas cidades foram removidas da análise posterior.

Cidade	Usuários Originais	Usuários Removidos	Usuários Restantes	Percentual Removido
Amsterdã	185	10	175	5,4%
Bangkok	37	3	34	8,1%
Barcelona	99	8	91	8,1%
Belo Horizonte	46	11	35	23,9%
Berlim	111	11	100	9,9%
Bruxelas	113	10	103	8,8%
Buenos Aires	23	3	20	13,0%
Chicago	878	63	815	7,2%
Cidade do México	126	19	107	15,1%
Curitiba	54	6	48	11,1%
Dublin	146	25	121	17,1%
London	711	45	666	6,3%
Los Angeles	361	23	338	6,4%
Madrid	70	7	63	10,0%
Melbourne	119	15	104	12,6%
Moscou	39	4	35	10,3%
Nova York	1036	81	955	7,8%
Osaka	30	4	26	13,3%
Paris	78	8	70	10,3%
Portland	411	31	380	7,5%
Praga	56	9	47	16,1%
Rio de Janeiro	111	15	96	13,5%
Santiago	32	3	29	9,4%
São Francisco	428	31	397	7,2%
São Paulo	177	9	168	5,1%
Singapura	32	3	29	9,4%
Sydney	129	21	108	16,3%
Tokyo	128	12	116	9,4%
Total	**5766**	**490**	**5276**	**8,5%**

Tabela 4: Usuários removidos por cidade

A Tabela 4 apresenta a quantidade de usuários removidos de cada cidade durante o Processo 4, conforme figura 10.

4.5 Análise exploratória

A seção de análise exploratória tem como finalidade obter algumas das características dos dados coletados dos usuários do aplicativo Untappd. Os resultados aqui demonstrados são provenientes dos Datasets 3 e 5, em que são consideradas todas as cidades com seus respectivos *"Heavy Users"*, e dos Datasets 4 e 6, onde os *"Heavy Users"* foram removidos.

Inicialmente, para o desenvolvimento desta seção foi realizada a contagem de *checkins* agrupando-os por região, e classificando em ordem decrescente. Além disso, foi considerada a frequência acumulada para estas quantidades. Esta visão auxilia na compreensão da quantidade de informação adquirida em cada uma das cidades selecionadas.

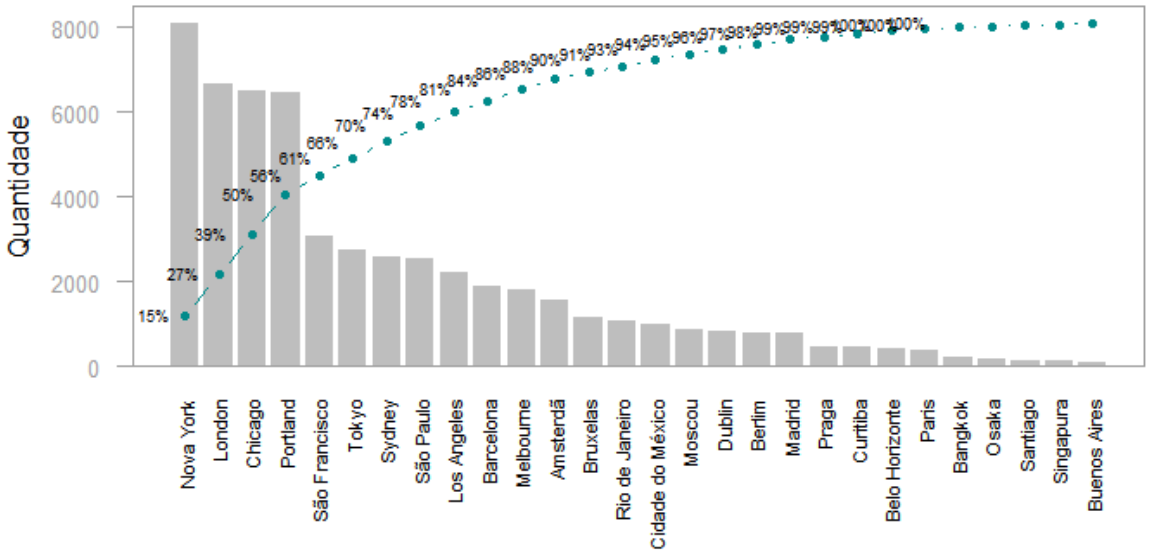


Figura 19: Quantidade e representatividade de *checkins*

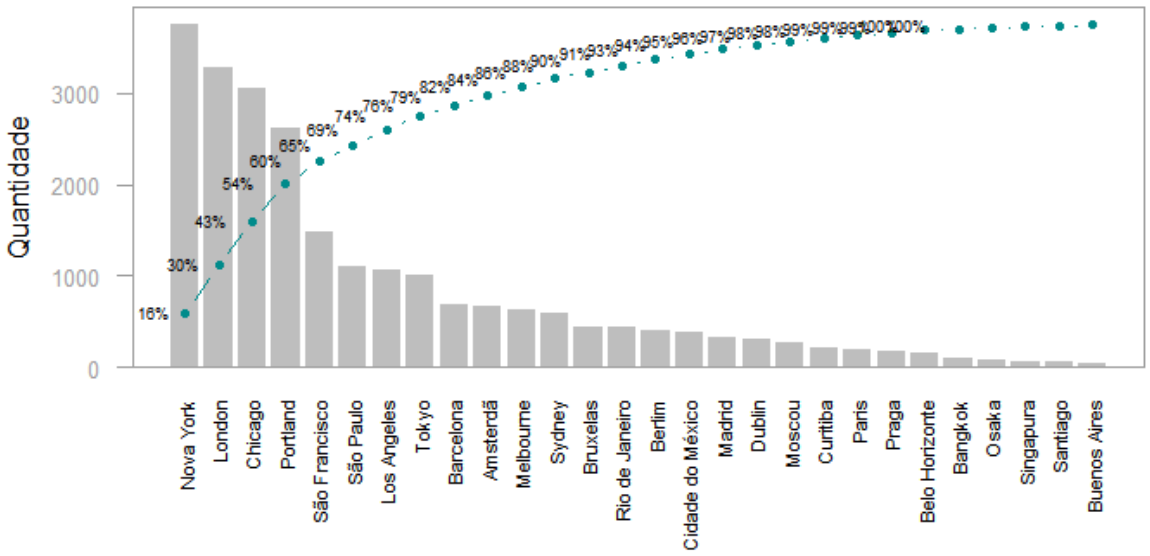


Figura 20: Quantidade e representatividade de *checkins* - Sem *Heavy Users*

Nas Figuras 19 e 20, obtidas através do processamento do Dataset 5 e Dataset 6, é possível visualizar a quantidade de *checkins* coletados em cada cidade. Para estas representações foram utilizados apenas os dados com referenciamento geográfico, visto que estes são essenciais para o decorrer do estudo.

Sendo assim, é possível identificar na Figura 19 que grande parte dos dados considerados para a análise são provenientes dos Estados Unidos. Enquanto que, os demais caracterizam-se por ser de grandes cidades como Tokyo, Sydney e São Paulo. Quando removidos os *Heavy Users* do Dataset 5, obtem-se a distribuição apresentada na Figura 20. Através desta Figura é possível perceber certa homogeneidade nos dados, visto que outras cidades - além dos Estados Unidos - apresentam certa representatividade nos dados analisados. Isso garante a consistência na análise, pois existe uma diversificação quando considerados aspectos étnicos.

Neste contexto, considerando a retirada dos *Heavy Users*, notou-se a necessidade de desconsiderar algumas das cidades da análise - conforme citado na seção 4.4. Neste contexto, torna-se possível avaliar preferências por cervejas especiais, considerando os usuários do aplicativo Untappd. Para analisar tal aspecto, inicialmente, foi construído o histograma de notas destes usuários - também chamados aqui de *Ratings*. Estas notas dadas pelos usuários podem variar de 0 a 5, com intervalos de 0,25.

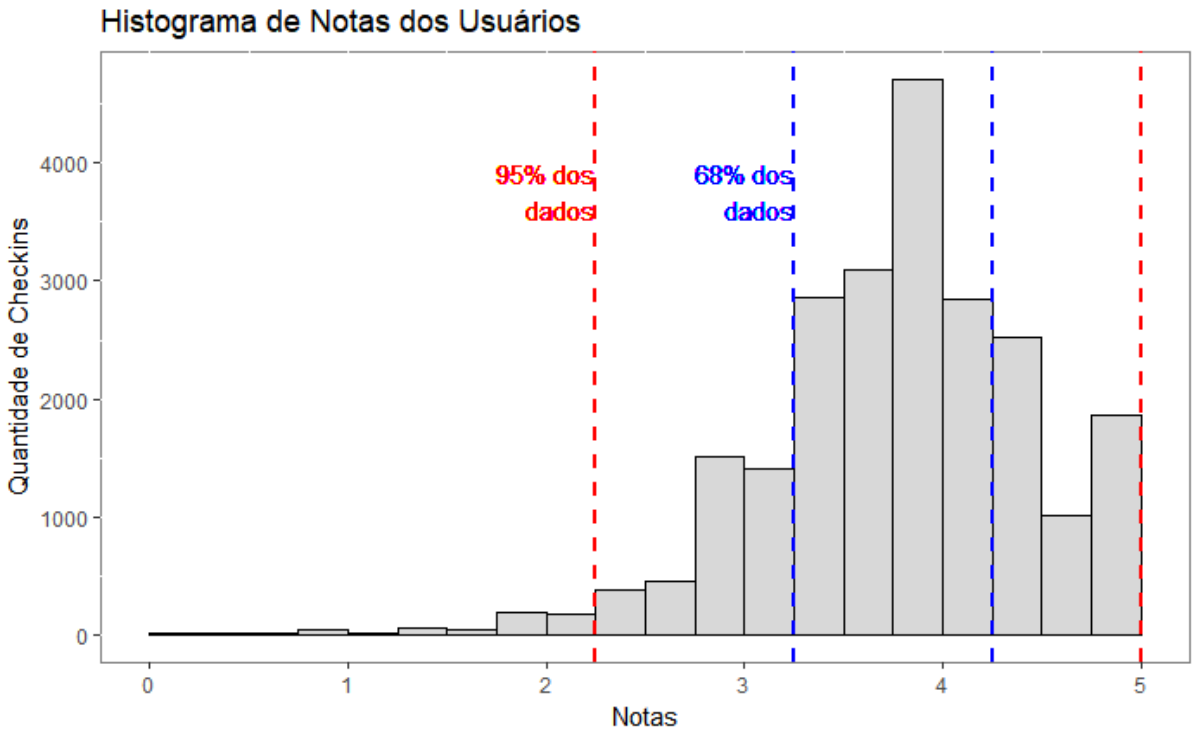


Figura 21: Histograma da nota dada pelos usuários das cidades selecionadas

A partir da Figura 21 é possível notar que 68% das notas atribuídas encontram-se entre 3,25 e 4,25. Enquanto que 95% destas, estão entre 2,25 e 5. Uma das possíveis causas desta concentração de notas superiores à média pode estar associada às formas de identificar um critério qualitativo quanto às notas. Isso deve-se à dificuldade em mensurar o que a nota realmente representa, considerando a escala proposta. Problemas semelhantes podem ser encontrados em pesquisas que utilizam conceitos da escala de Likert (CUMMINS; GULLONE, 2000; HODGE; GILLESPIE, 2003). É possível observar com base no Histograma representado na Figura 21,

que as notas de 4,5 a 4,75 tiveram poucas marcações com relação às faixas adjacentes. Isto pode ser justificado pela falta de cognição, por parte dos indivíduos, do real significado de cada valor desta escala.

A fim de facilitar a visualização das informações, os *Ratings* foram agrupados em intervalos de 0,5 em uma escala de 0 a 5. No gráfico a seguir estão representados os *Ratings* classificados por cidades.

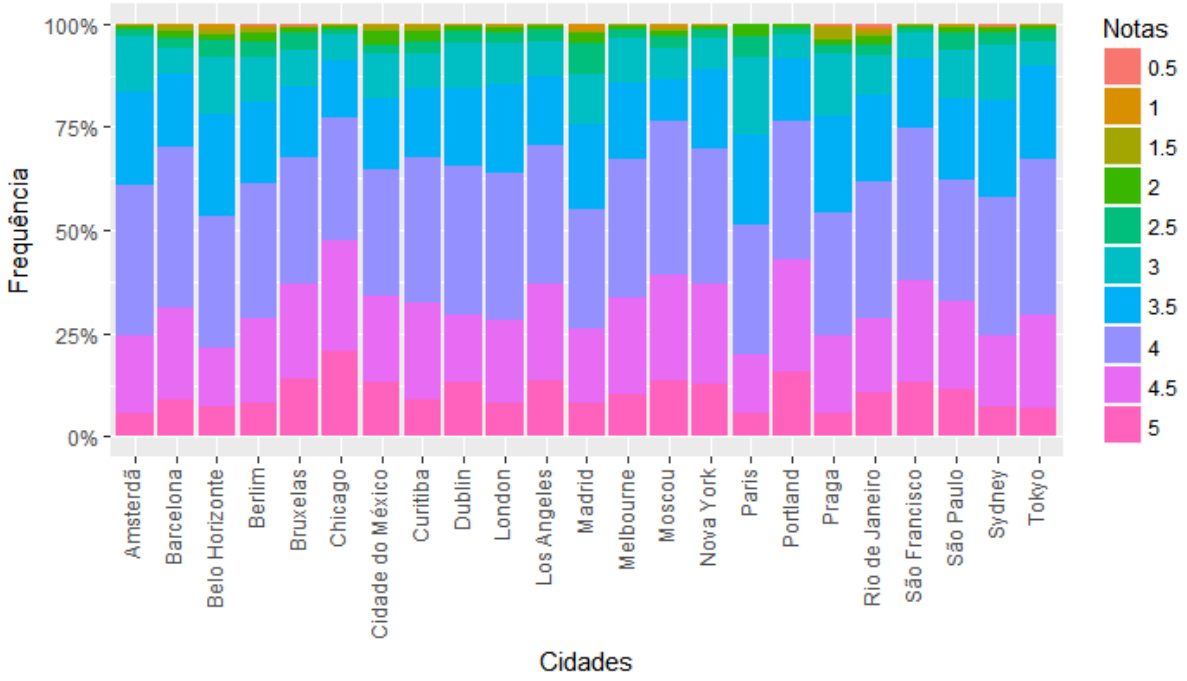


Figura 22: Checkins por notas dos usuários

A partir da Figura 22 é possível obter uma visão geral a respeito da distribuição de *Ratings* por cidades. Onde observa-se que para todas as cidades as notas 4; 4,5 e 5 representam mais de 50% dos *checkins*. Por outro lado, é importante perceber que notas menores que 2 possuem pequena representatividade na grande maioria das cidades. Isso pode indicar uma maior propenção dos usuários do aplicativo em atribuírem notas maiores que 4 para as cervejas de suas preferências, como já relatado sobre as potenciais dificuldades encontradas para este tipo de escala.

Conforme abordado por Mosher (2017), existem diversas características das cervejas que podem ser avaliadas ou medidas, a fim de ditar o sabor de cada estilo de cerveja. Algumas das medidas utilizadas para avaliar as cervejas tem o objetivo de garantir que seus fabricantes consigam replicar uma receita desenvolvida anteriormente, bem como, auxiliar o consumidor a julgar as propriedades da bebida antes mesmo de prová-la. Duas dessas medidas são: o IBU[1] e ABV[2]. Sendo a primeira utilizada para avaliar o amargor da bebida, através da medição

[1] *International Bitterness Units* - ou, traduzido Unidade Internacional de Amargor.
[2] Alcohol by Volume - Traduzido, álcool por volume.

de quantidade de ácidos iso-alpha presentes nela. Enquanto que a segunda caracteriza-se pela quantidade de álcool presente na bebida.

Apropriando-se deste conceito, cogitou-se a possibilidade de existir alguma relação entre o teor alcoólico e amargor quando relacionadas às notas atribuídas a cada cerveja. Desta forma, para realizar esta análise foram considerados os *Heavy Users*, pois neste contexto as informações provenientes destes trazem informações relevantes quanto às preferências por cervejas.

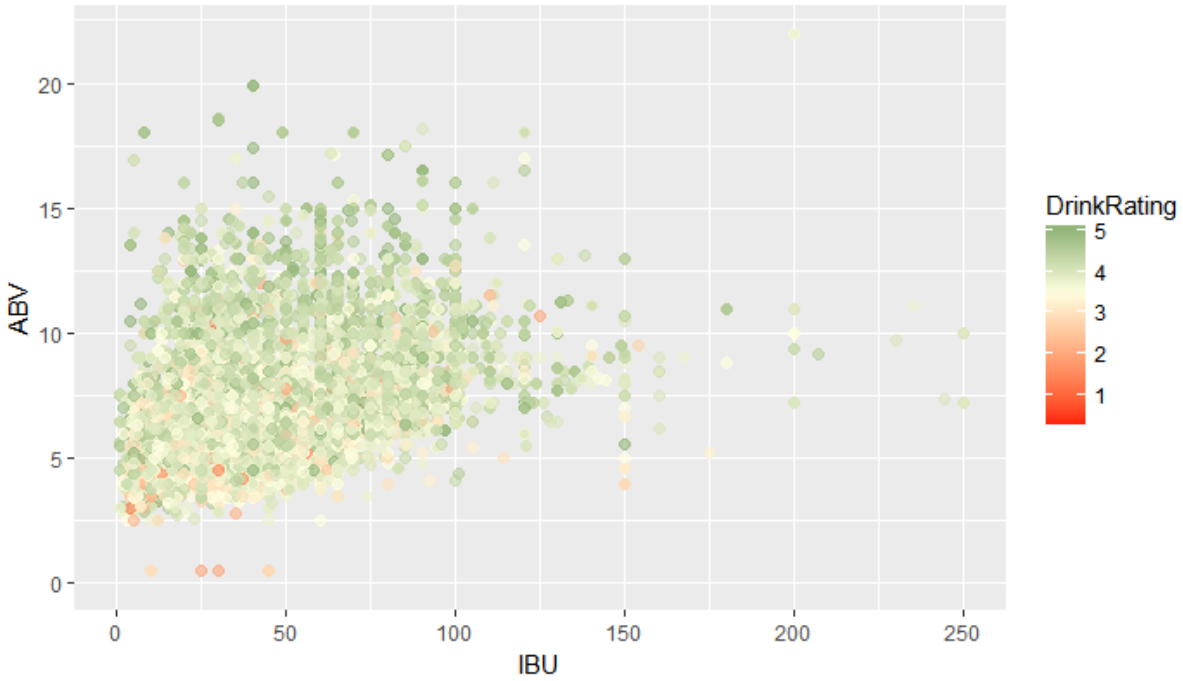


Figura 23: Relação entre ABV por IBU considerando as notas dos usuários

A Figura 23 traz a visualização da relação entre ABV x IBU. Onde, no eixo *"x"* pode ser encontrado o índice referente ao amargor, no eixo *"y"* o teor alcóolico e a escala de cores representa as notas atribuídas pelos usuários às bebidas consumidas. Através da visualização da imagem, é possível perceber que existe uma concentração desta relação entre ABV x IBU. Entre 0 e 120 do eixo *"x"* refere-se ao IBU, onde concentram-se as cervejas consumidas, e entre 2,5 e 12 no eixo *"y"* representa o ABV.

A fim de aprofundar a perspectiva, foi aplicado um filtro nas notas que viriam a classificar cervejas como "muito boas". Tomando como base o histograma da Figura 21, estabeleceu-se um ponto de corte de forma que possibilitasse selecionar as cervejas classificadas como melhores, para aplicar ao Dataset 3, em que é possível obter uma classificação mais abrangente em questão de estilos de cerveja - Consultar no Anexo B a Tabela 5.

Entretanto, a fim de aprofundar esta perspectiva foi aplicado um filtro nas notas do Dataset 3, em que é possível obter uma classificação mais abrangente quando consideradas questões referente aos estilos de cerveja - Consultar no Anexo B a Tabela 5. Este filtro foi

atribuído tomando como base o histograma da Figura 21, onde através deste estabeleceu-se um ponto de corte em que o objetivo era de selecionar as cervejas com melhores classificações. Portanto, a fim de atender este critério, foram selecionados somente os *checkins* em que as notas dos usuários foram superiores ou iguais à 4,25.

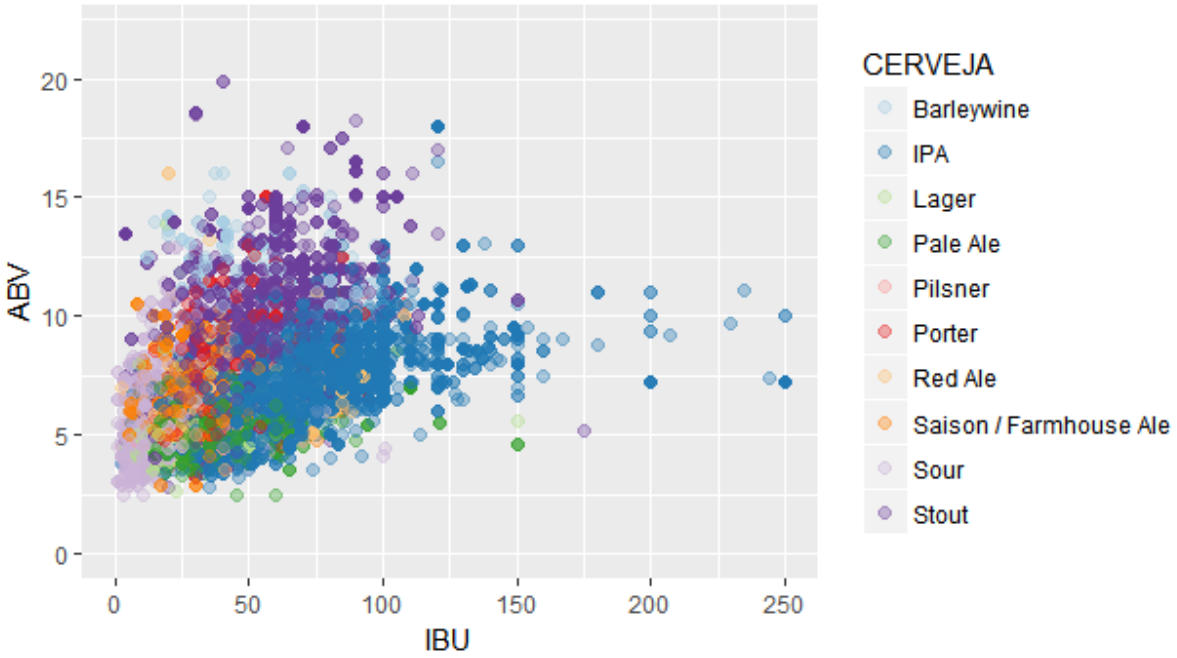


Figura 24: Relação entre ABV por IBU das 10 melhores cervejas - Baseado nos Drink Ratings

A Figura 24, apresenta o resultado do filtro aplicado ao Dataset 3. Porém, o resultado da aplicação deste filtro não apresentou conclusões significativas, devido ao fato de não ocorrerem variações nas concentrações. Isto motivou à seleção, neste Dataset, com as maiores notas, das 10 cervejas mais consumidas, conforme apresentado na figura em questão. É importante notar que as cervejas listadas na legenda da figura estão em ordem alfabética, se considerada a ordem por consumo esta seria: *IPA; Stout; Pale Ale; Sour; Porter: Lager; Saison; Barleywine; Pilsner* e *Red Ale*. Deste modo, torna-se possível identificar alguns padrões relacionados às características das cervejas. Como, por exemplo, as *IPA's* possuírem teor alcóolico menor que as *Stouts* e serem razoavelmente mais amargas ou ainda o baixo teor alcóolico relacionado ao tipo de cerveja *Sour*. Porém, quando consideram-se as notas atribuídas nestes *checkins* não é possível estabelecer uma relação de preferências para o IBU e ABV.

A falta de relação entre as medidas, pode estar relacionada ao fato das notas atribuídas a uma cerveja estarem ligadas às preferências inerente à cada pessoa. Pois, para certos indivíduos uma ótima cerveja pode ser aquela que possui baixo teor alcóolico e leve amargor, enquanto que para outros o oposto é a melhor opção, cervejas altamente alcóolicas e amargor acentuado. Ainda, outra possibilidade é a influência do fator cultural, que será explorado na seção seguinte.

4.6 Identificação de aspectos e diferenças culturais

A compreensão de aspectos culturais pode ser um diferencial quando consideradas as preferências de usuários em determinados nichos de mercado. O entendimento destes aspectos pode vir a favorecer empresas, atuantes em uma localidade, permitindo a adequação de suas atividades através de comparações com a de outros mercados.

Com o objetivo de estudar diferenças culturais foram selecionadas as cidades conforme alguns dos critérios já descritos na Seção 4.3. A fim de demonstrar possíveis aspectos culturais associados ao consumo de cerveja, procurou-se através da seleção abranger cidades diversificadas e populares em diversas partes do mundo para aplicar os métodos de clusterização.

Foi utilizado o Dataset 5, em que as notas zeradas foram eliminadas, restando apenas notas válidas e não duplicadas, de acordo com a chave [UserID, cerveja, cervejaria, cidade], garantindo assim uma maior confiabilidade nos dados. Para esta análise, as preferências dos usuários foram calculadas de acordo com a contagem total das notas informadas e, posteriormente, normalizadas com a maior quantidade de cada cidade.

Ainda, para avaliar o impacto dos usuários assíduos do aplicativo - aqui chamados de *"Heavy users"* - foi construído o Dataset 6 com as mesmas características do Dataset 5, porém, foram retirados os *"Heavy users"*, conforme já relatado no Capítulo 3.

A fim de classificar as preferências dos usuários foi realizada a análise de agrupamentos. Nesta análise foi utilizada a classificação 2 - conforme mencionado na Seção 3.4 e disponível no anexo B - pois considerando os aspectos geográficos, esta veio a favorecer a análise devido aos critérios de classificação utilizados. Para a realização do método de agrupamento é necessário que cada cidade seja representada por um vetor de preferências, que é calculado através do agrupamento das preferências por tipo de cerveja de cada usuário da cidade em questão. Na sequência é calculada a distância entre cada uma das cidades, tendo como base as preferências dos usuários de cada uma delas. Para o cálculo da matriz de distâncias foi utilizado o método Canberra como medida de similaridade, representada pela fórmula a seguir, conforme descrito por Everitt (2006).

$$d_{i,j} = \sum_{k=1}^{q} \left| x_{ik} - x_{jk} \right| / (x_{ik} + x_{jk}) \tag{4.1}$$

Além disso, o critério de agrupamento utilizado foi o *Ward 2*, em que o método para formação dos grupos ocorre através da maximização da homogeneidade entre os grupos formados (SUBHASH, 1996). A medida de homogeneidade é calculada através da soma dos quadrados entre os *clusters*, ou seja, este método tem como finalidade encontrar o menor desvio padrão entre os dados de cada grupo. Desta forma garante que os *clusters* formados deterão o menor erro interno entre seus vetores.

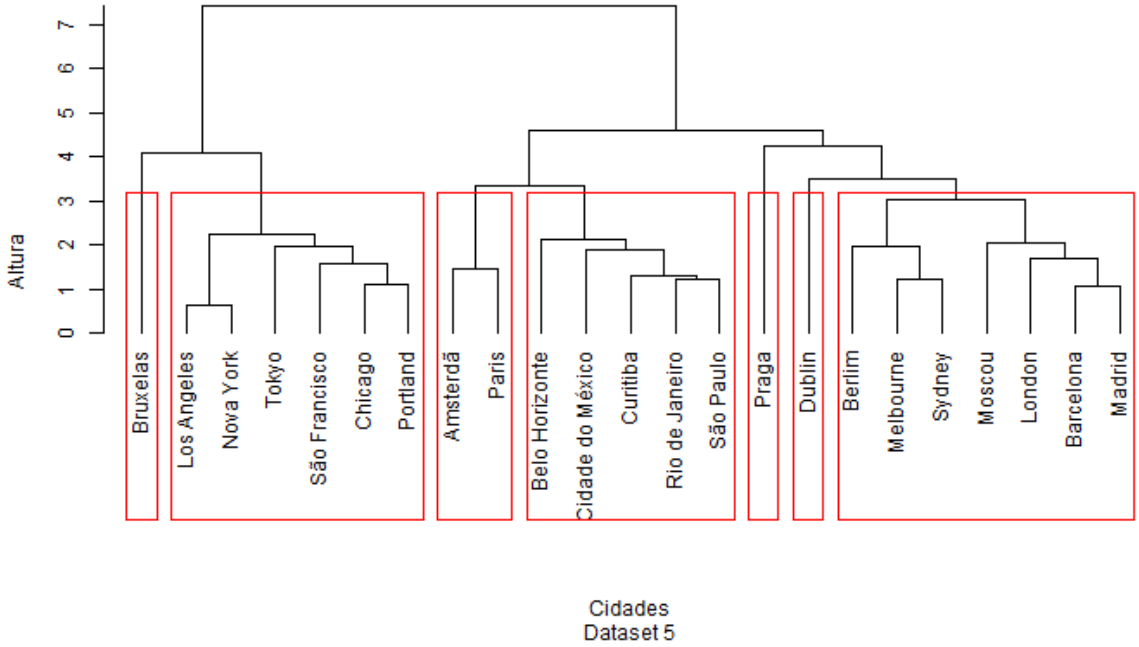


Figura 25: Canberra Ward D2

Através da Figura 25 é possível identificar que o agrupamento favorece as identidades étnicas. De modo que cidades dos Estados Unidos permaneceram unidas em um mesmo *cluster*. Contudo este agrupamento é marcado pela presença de Tokyo. Tendo em vista que o Dataset 5 foi utilizado para a criação deste dendograma, que desconsidera peso das notas, este fato ocorre devido à grande quantidade de *checkins* referentes às cervejas IPAS, classificadas como sendo de origem Norte Americana. Ainda, o método chama a atenção para as cidades do Brasil agrupadas em um mesmo *cluster*, que, por sua vez contém a Cidade do México. Além disso, existe um terceiro grande *cluster* onde estão agrupadas cidades de outros países de colonização Européia.

Neste contexto, é possível perceber que parte das cidades em que esperava-se a formação de um *cluster*, por serem de um mesmo país, indicando similaridade étnica, vieram a se agrupar. Este fato demostra o potencial de exploração dessas características para o estudo de diferenças culturais relacionado ao consumo de cervejas artesanais.

Com o intuito de analisar a possível influência dos *Heavy Users* no agrupamento realizado na Figura 25, executou-se a mesma etapa sobre o Dataset 6. Onde estes são removidos.

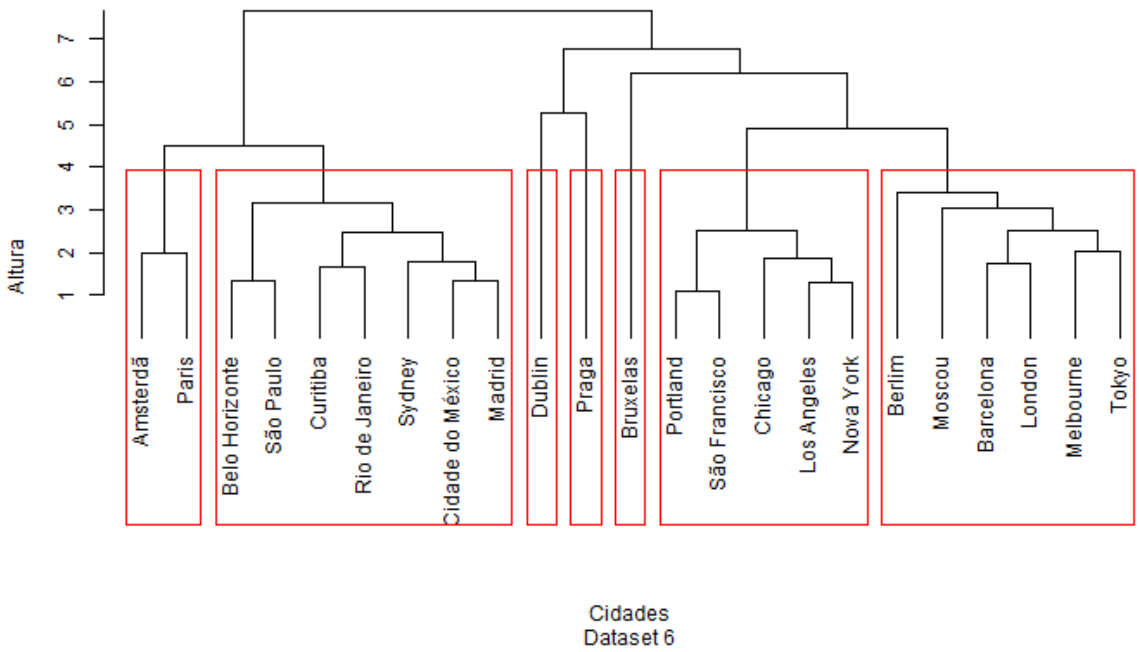


Figura 26: Canberra Ward D2 - Sem *Heavy User*

O resultado do método de agrupamento adotado para o Dataset 6 pode ser visualizado na Figura 26. É possível perceber que a exclusão dos *Heavy Users* trouxe algumas diferenças no agrupamento. Como a diferença no agrupamento das cidades do Brasil, apresentando semelhanças entre os consumidores do Rio de Janeiro e Curitiba, e Belo Horizonte e São Paulo. Ainda, traz à este agrupamento as cidades de Sydney, Cidade do México e Madrid. Entretanto, essa remoção favoreceu o agrupamento das cidades dos Estados Unidos, que permaneceram unidas exclusivamente em um único *cluster*.

Por meio de uma breve análise das informações obtidas para as cidades de Sydney, Cidade do México e Madrid, é possível identificar fatores que permitem esse agrupamento. As três cidades possuem quantidades semelhantes de *checkins* para os tipos de cervejas. Sendo assim, foram selecionadas as quinze cervejas com maiores quantidades de *checkins* nestas cidades e comparadas entre elas. Na Cidade do México e Madrid foram encontradas onze estilos de cervejas semelhantes[3], enquanto que Sydney apresenta um estilo de cerveja a menos e outros dois semelhantes somente à Cidade do México [4].

A análise realizada através da quantidade de *checkins* por região garante a possibilidade de identificar aspectos semelhantes entre os países, se considerada a quantidade de cerveja consumida pelo público médio do aplicativo Untappd. Desta forma, é possível compreender algumas das relações existentes entre os consumidores do público dos países analisados. Bem

[3] Belgian Strong Dark Ale, Blonde Ale, Brown Ale, IPA, Lager, Pale Ale, ,Pilsner, Porter, Red Ale, Sour, Stout.
[4] Não encontra-se entre os quinze mais consumidos Belgian Strong Dark Ale e acrescenta-se Scotch Ale, Golden Ale.

como, identificar relações entre preferências por determinados estilos de cerveja.

Em uma tentativa de melhorar o resultado obtido foi elaborada uma nova forma de cálculo, nomeado por "Hibrido", utilizando as mesmas fontes de dados.

Sendo assim, o método Hibrido foi criado com o objetivo de tentar minimizar o impacto da baixa quantidade de *checkins* influenciado por estilos de cervejas incomuns em determinadas cidades. Para esta análise foram utilizados também o Dataset 5 e o Dataset 6 com o método Canberra para cálculo das distâncias. Porém, as preferências dos usuários foram calculadas de maneira diferente. Primeiramente, foi realizada a soma de todas as notas para cada estilo em cada cidade. Em seguida, foi realizado o cálculo da média de todas as notas para cada estilo em cada cidade. Com esses resultados foi realizado o cálculo do produto, entre a soma e a média. Por fim, estes produtos foram normalizados com o maior valor de cada cidade.

Com a utilização do método Hibrido esperava-se que as cidades similares entre si fossem agrupadas por um mesmo *cluster*, assumindo que isso seja válido para cidades de um mesmo país, além de se esperar um resultado melhor comparado com o método da contagem. É possível perceber na Figura 27, que as cidades Brasileiras permaneceram no mesmo *cluster* junto com a Cidade do México. Além disso, formou-se o *cluster* dos Estados Unidos com todas as suas cidades.

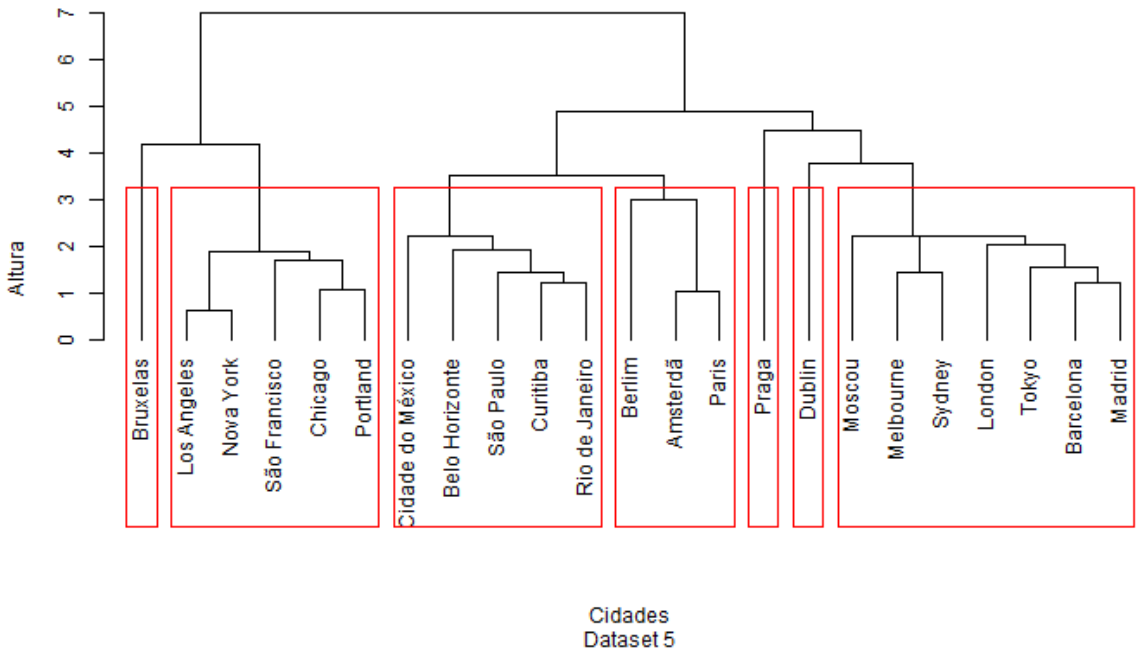


Figura 27: Canberra Ward D2 para o método Hibrido

Nota-se que quando os *"heavy users"* são removidos, ainda utilizando o método Hibrido, agrupamentos diferentes dos apresentados na Figura 27 são formados, o que pode ser identificado através da Figura 28. As cidades Brasileiras permaneceram no mesmo *cluster*, po-

rém, chama atenção a inclusão das cidades de Madrid, Berlim e Sydney. Ainda, as cidades dos Estados Unidos permaneceram unidas em um mesmo *cluster*, somente alternando a posição entre si. Os *clusters* formados que são compostos por somente uma única cidade também permaneceram da mesma maneira, que é o caso de Bruxelas, Praga e Dublin.

De modo geral, comparando os agrupamentos formados pelas Figuras 27 e 28, nota-se que estas mudanças ocorreram após a retirada dos *"heavy users"*. Sendo assim, é possível perceber o impacto que esses usuários - considerados *"outliers"* - podem vir a exercer sobre uma determinada análise, por isso a importância em desconsiderá-los.

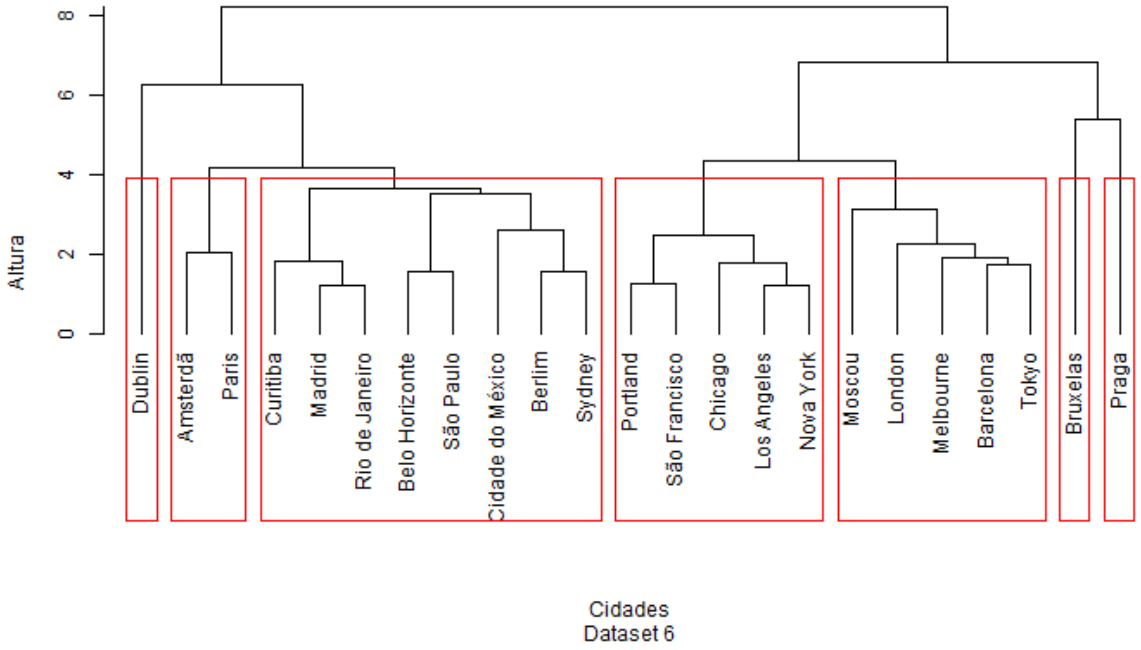


Figura 28: Canberra Ward D2 para o método Hibrido - Sem *Heavy User*

Avaliando os resultados obtidos com o método da Contagem e o método Hibrido é possível perceber algumas alterações nos agrupamentos. No Dataset 5 da contagem, o agrupamento das cidades dos EUA é composto juntamente com a cidade de Tokyo, já no Dataset 5 do Hibrido, o *cluster* formado é composto somente pelas cidades dos EUA. Isso deve-se às características dos métodos adotados para o desenvolvimento da análise.

O método de contagem de *checkin* oferece uma visão relacionada à quantidade de cervejas consumidas. O que permite a observação quanto aos padrões de quantidade de consumo de determinados tipos de cerveja para cada país. Por outro lado, o método Hibrido pretende avaliar as preferências de cada país quanto ao estilo de cerveja, através de um critério de ponderação de notas. Este método oferece vantagens quando procura-se identificar aspectos culturais quando trata-se de preferências por cervejas.

O método Hibrido apresenta maior eficiência no quesito de agrupar de acordo com as

características étnicas e também culturais quanto ao gosto e estilo da cerveja daquela região. Entretanto, o método Hibrido e de Contagem não devem ser comparadas pois apresentam características diferentes entre eles. Enquanto um encarrega-se de considerar aspectos relacionados à quantidade de cerveja consumida, o outro procura identificar preferências relacionadas ao gosto de cada país.

Com o objetivo de exemplificar a utilidade destes métodos, foi elaborado um estudo de caso. A fim de simular um possível sistema de sugestão de destinos turísticos, foram selecionados dez dos usuários com maiores quantidades de *checkins* - tratados como *Heavy Users* - *HU*. Cada usuário é tratado como "HU*x* CIDADE*y*". Onde "CIDADE*y*" representa nome da cidade em que este usuário habita e a letra *"x"* representa um valor numérico para casos em que exista mais de um usuário com a mesma característica.

Cada um destes usuários foi atribuído a um vetor considerando suas respectivas preferências por tipo de cerveja, assim como foi feito para as cidades anteriormente. Em seguida, estes usuários foram incluídos ao processo de agrupamento realizado para o Dataset 6. Deste modo, o resultado do agrupamento permite avaliar como a preferência do *Heavy User* em questão se assemelha à do consumidor médio das demais cidades.

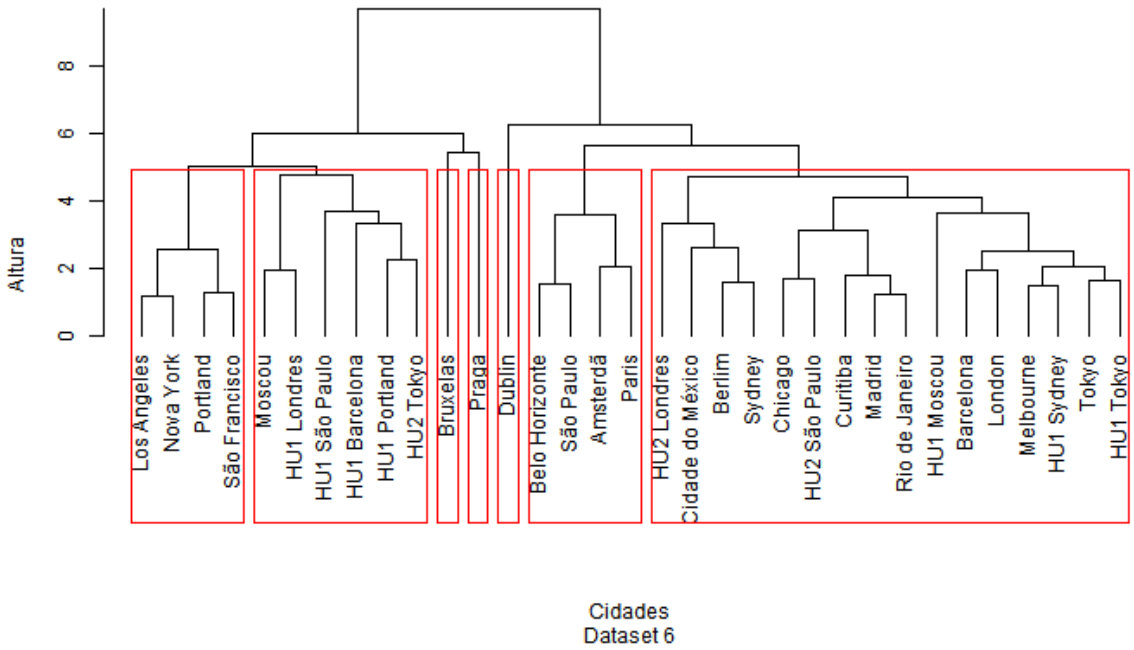


Figura 29: Canberra Ward 2 utilizando o método Hibrido - Com dez usuários

A Figura 29 apresenta o agrupamento realizado para o estudo de caso. Sendo assim, a seleção de cidades recomendadas ocorrerá a partir das cidades próximas à posição em que o usuários se encontra no *cluster* em questão. Podendo estas cidades estar em um *cluster* próximo, desde que estejam em uma mesma derivação do *cluster*.

As possíveis recomendações deste sistema, conforme a Figura 29 consideraria para o usuário "HU1 Londres" o destino de Moscou, devido a proximidade do agrupamento. Ainda, para os usuários "HU1 São Paulo", "HU1 Barcelona", "HU1 Portland", "HU2 Tokyo" um possível destino seriam as cidades dos Estados Unidos - Los Angeles, Nova York, Portland e São Francisco, ainda, sendo possível sugerir também Moscou.

Seguindo o mesmo método, para o "HU2 Londres" as possíveis recomendações seriam Cidade do México, Berlim e Sydney. Enquanto que para o "HU2 São Paulo" as sugestões viriam a ser Chicago, Curitiba, Madrid e Rio de Janeiro. O "HU1 Moscou" poderia receber sugestões de destinos para Barcelona, Londres, Melbourne e Tokyo.

Contudo, um caso interessante apresenta-se com os usuários "HU1 Sydney" e "HU1 Tokyo", pois estes não apresentam preferências diferentes quanto às suas próprias culturas. Entretanto, possíveis sugestões para estes seriam Melbourne, Tokyo, Londres e Barcelona. Este é provavelmente um indicativo de que estes usuários não possuem o hábito de diversificar ou experimentar estilos de cerveja diferentes do comum de sua região de origem.

Este é um dos possíveis exemplos de aplicação prática desta análise. Sendo assim, isso justifica a relevância deste estudo pois a informação apresentada pode auxiliar na tomada de decisão incluindo pequenas empresas em início de atividades a compreender melhor o seu consumidor. Neste exemplo a análise foi aplicada ao turismo, porém, outras áreas podem ser exploradas. Além disso, a cerveja é somente um exemplo de pesquisa que pode vir a ser substituída por outro produto.

4.7 Identificação de áreas populares

Atualmente diversas cidades têm criado iniciativas a fim de tornar alguns dos dados oficiais disponíveis ao público, entretanto, grande parte das informações podem ser extraídas de mídias sociais como Facebook[5], Instagram[6] e Twitter[7]. Estes exemplos de mídias sociais são chamadas de Redes Sociais Baseadas em Localização (LSBNs)[8], onde indivíduos atuam como sensores sociais, através da utilização dos *smartphone's* (SANTALA et al., 2017).

O estudo das dinâmicas urbanas em diferentes escalas espaciais é um ponto tradicionalmente desafiador, pois, normalmente é um processo caro, uma vez que demanda a realização de entrevistas e questionários a um grande número de pessoas, resultando em uma apresentação limitada da realidade. Entender como indivíduos utilizam os espaços urbanos pode auxiliar na compreensão dos moradores das cidades e auxiliar nas decisões de planejamento urbano. Deste modo, através dos dados adquiridos por meios das mídias sociais, explorou-se a possibilidade de identificar a utilização do espaço urbano das cidades.

[5] https://www.facebook.com
[6] https://www.instagram.com
[7] https://twitter.com/
[8] *Location Based Social Network.*

Sendo assim, conforme já abordado anteriormente na Seção 2.4.2, o DBSCAN tem como objetivo formar *clusters* através da distância euclideana dos vizinhos mais próximos de um determinado ponto. Para que isso ocorra é essencial que sejam passados como parâmetros do algoritmo um raio e a quantidade mínima de observações a serem consideradas para este agrupamento. Portanto, nesta seção serão abordados os casos de agrupamentos utilizando o DBSCAN.

Para a realização deste agrupamento foram considerados todos os *checkins* presentes para a determinada região. Ou seja, não foram removidos os *"Heavy Users"* e cervejas duplicadas. Isso oferece uma visão mais aproximada do cenário real encontrado nestas cidades.

Estes agrupamentos foram realizados através do DBScan considerando as coordenadas geográficas presentes na estrutura dos *tweets* coletados. Deste modo é possível realizar o cálculo de distância mínima para identificar a densidade de uma região.

A motivação em considerar esta seção na análise foi a possibilidade de criação da "Rua da Cerveja" - *"Beer Street"* -, anunciada pela prefeitura de Curitiba[9]. Assim, através da aplicação deste método é possível avaliar a viabilidade de implantação deste projeto. Sendo que por meio desta análise torna-se possível compreender algumas das preferências levando em conta aspectos geográficos, como locais de maior movimentação, áreas de maior atração deste público, entre outros.

A partir disto, a análise foi extendida para as demais cidades Brasileiras, sendo elas: Belo Horizonte, Rio de Janeiro e São Paulo. Procurou-se considerar as cidades já analisadas por Silva e Graeml (2016).

[9] http://www.curitiba.pr.gov.br/noticias/prefeitura-articula-implantacao-da-rua-da-cerveja/41348

4.7.1 Áreas populares para Curitiba

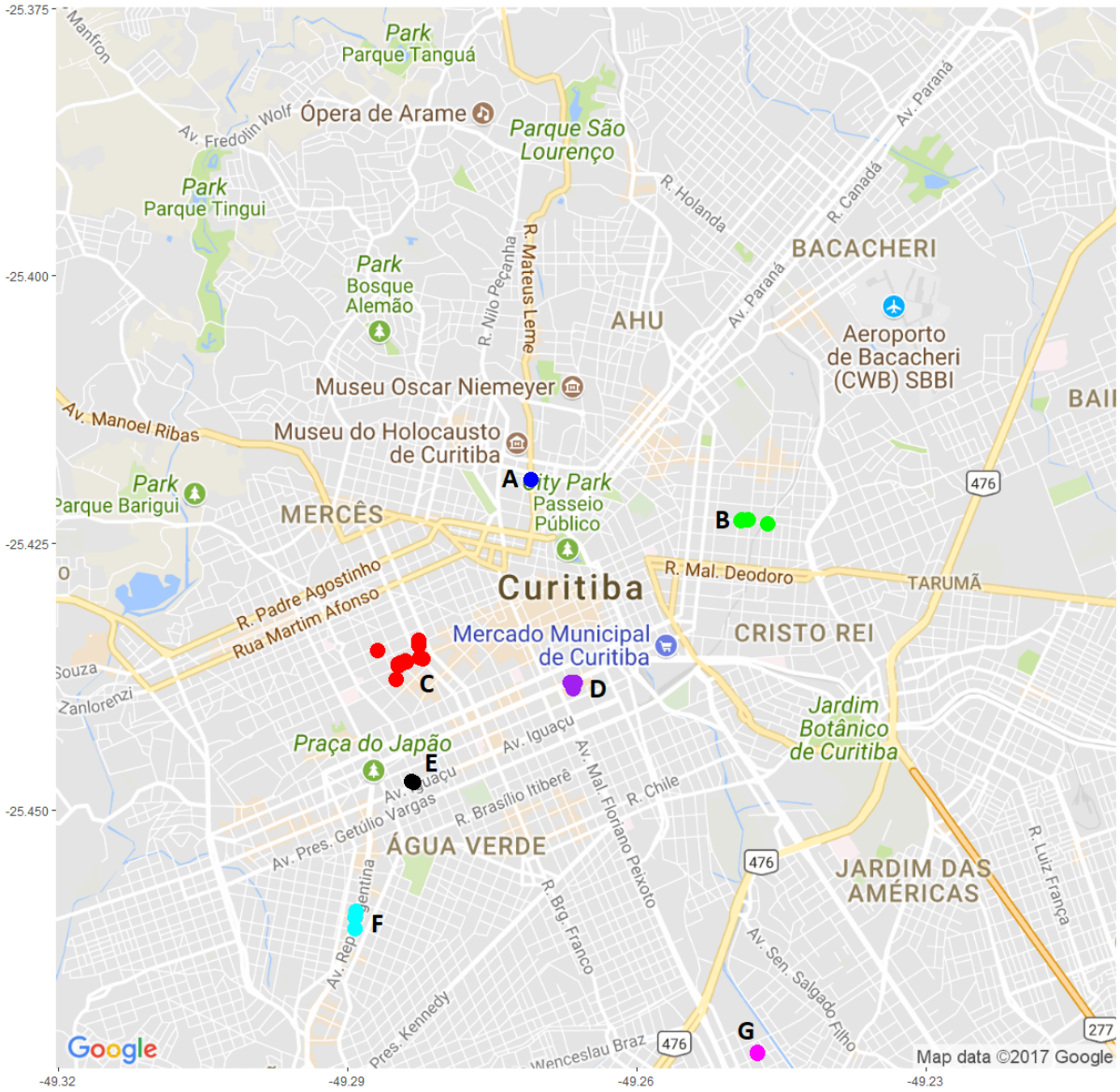


Figura 30: DBScan da cidade de Curitiba

A Figura 30 é o resultado do agrupamento utilizando o DBScan da cidade de Curitiba. Para a aplicação deste método foi definido o parâmetro Eps - raio - como 250 metros e considerado o número mínimo de pontos igual a 20 para todos os conjuntos de dados.

Considerando a possibilidade de criação da "Rua da Cerveja" em Curitiba, essas informações poderiam vir a ser úteis para auxiliar a tomada de decisão quanto à escolha do ponto. Sendo assim, após a aplicação do BDScan foram encontrados sete pontos candidatos, descritos a seguir.

O *cluster* representado pela letra "A" sinalizado pela cor azul é onde localiza-se o *Hop'n Roll*, um conhecido *pub* especializado em cerveja artesanal. A letra "B" possui pontos locali-

zados no mapa na cor verde, sendo representado principalmente pela Masmorra Cervejaria, Crazy4Beer e Mercado 153, localizada junto à Rua Itupava, um centro gastronômico [10].

Os *Clusters* das letras "C" e "E", respectivamente nas cores, vermelha e preta, estão localizados no bairro do Batel, um dos bairros que está entre os mais ricos da cidade de Curitiba, que hospeda muitos restaurantes, *pubs* e casas noturnas. Esses *clusters* também oferecem vários *pubs* especializados em cervejas artesanais, como *The Meatpack House*, *Whatafuck* e Clube do Malte. Além disso, especificamente no ponto "E", a Avenida Iguaçu, é marcada pela presença do *We are Bastards Pub* e o *Crossroads*.

O *Cluster* da letra "D", representado na cor roxa, está localizado o Shopping Estação que hospedou em novembro de 2016 um evento cervejeiro - *Way Craft Beer Soul*[11] - em comemoração a uma das cervejarias existentes da região. Contudo, o local ainda oferece algumas opções para o consumo de cerveja artesanal, bem como, estrutura para eventos como o realizado.

O *Cluster* da letra "F", representado na cor ciano, está localizado no bairro Água Verde. Embora este seja um bairro residencial, o que pode possuir diversos *checkins* caseiros, o *cluster* formado é representado principalmente pelo Império Cervejeiro, um *pub* especializado em cerveja artesanal com grande variedade de rótulos.

Por fim, no ponto representado pela letra "G", na cor rosa, fica situada a *Bodebrown Pub* no bairro Hauer. Uma Micro-cervejaria instalada em um bairro residencial, afastado do centro da cidade, que procura hospedar diversos eventos envolvendo cerveja, comidas e música. Estas são as características que tornam este o local ideal para a escolha do ponto como a "Rua da Cerveja".

Além disso, algumas características fundamentais devem ser levadas em consideração para a criação da "Rua da Cerveja". A primeira característica que deve ser observada (C1) é a possibilidade de bloquear a rua por algumas horas à noite e finais de semana, de forma a não causar impactos significativos no local. A segunda característica (C2) é que esta Rua não deverá ser localizada próximo a locais que exijam silêncio, como hospitais e algumas áreas residênciais. A terceira (C3) exige que as empresas nas proximidades não sejam impactadas negativamente com a escolha. Finalmente, a escolha deve atender à quarta característica (C4) que representa a capacidade de acomodar um público de aproximadamente 10 mil pessoas para expor produtos, comercializar alimentos ou oferecer infra estrutura para eventos ou shows.

Tendo em vista estas características, algumas delas não se aplicam aos *clusters* formados, o que justifica a impossibilidade de escolha destes. Quanto a característica C1, os *clusters* correspondentes às letras "A", "D" e "F" estão situados em uma região central e possuem ruas movimentadas, portanto, não seriam bons para escolha do local. Além disso, o ponto represen-

[10] http://www.curitiba.pr.gov.br/noticias/conheca-dez-roteiros-gastronomicos-de-curitiba/42974
[11] http://revistabeerart.com/news/craft-beer-soul-2016

tado pela letra “F” está localizado em uma área residencial, sendo assim, não atende também a característica C2. Os pontos das letras “B”, “C” e “E”, estão localizados em ruas com comércio ativo, portanto, não seria uma escolha adequada por violar a característica C3. Além disso, nenhum dos pontos correspondentes às letras “A”, “B”, “C”, “D”, “E” e “F”, suportaria um público igual ou maior a 10 mil pessoas sem causar um comprometimento negativo na região. Isso deve-se à questão de escalabilidade das ruas, sendo que existe a possibilidade de causar congestionamentos excessivos nestas regiões. O que, por sua vez, não atende a característica C4 e C1.

O ponto “G”, por sua vez, por estar situado em um local afastado do centro da cidade reserva a possibilidade de bloquear a rua nos finais de semana, sem trazer impactos no local, atendendo a característica C1. Ainda, quanto à C2, apesar da rua possuir algumas residências este é marcado pela forte presença de empresas de pequeno porte nos entornos da cervejaria, o que tornaria propícia a realização de eventos nos finais de semana e à noite. A característica C3 justifica-se pela existência da cervejaria desde 2009 no local, promovendo eventos desde lá. Por fim, a infra estrutra tem capacidade para acomodar uma grande quantidade de pessoas, o que pode vir a atender também a caraterística C4.

4.7.2 Áreas populares para Belo Horizonte

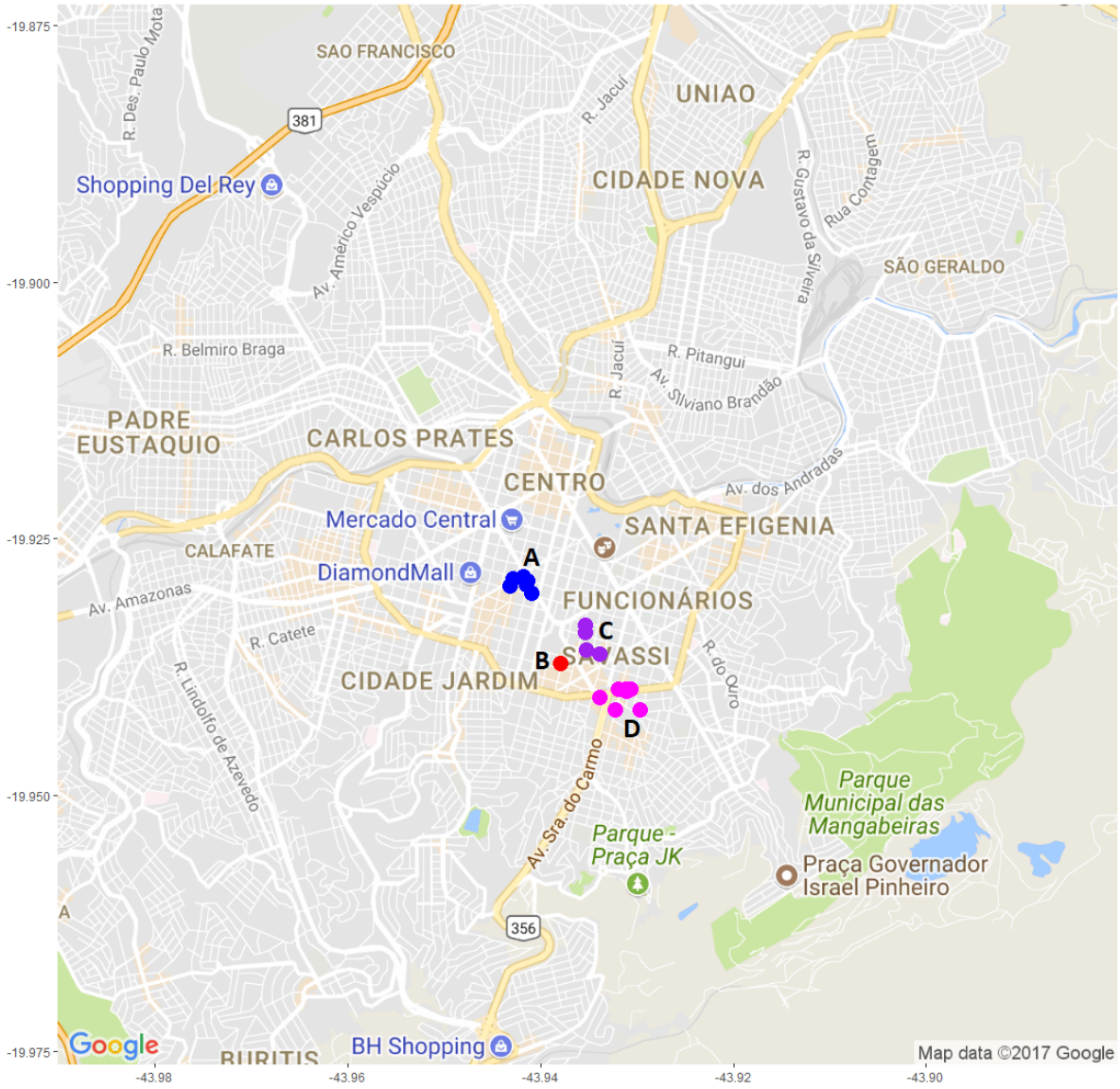


Figura 31: Belo Horizonte

Na Figura 31 está representado o resultado do agrupamento feito para a cidade de Belo Horizonte. Para existir um *cluster* é necessário que existam pelo menos quinze observações em um raio de 250 metros. Dados os critérios, formaram-se quatro *clusters*, sendo eles descritos abaixo.

O *Cluster* em azul, demarcado pela letra "A", é representado pelo "Centro Cervejeiro", localizado no bairro Lourdes. Este comércio é fruto da parceria entre a *Confraria do Malte* e o *Lamas Brew Shop*, onde são encontrados insumos para a fabricação da bebida e chopes artesanais para consumo no local.

O *Cluster* em vermelho, representado pela letra "B", é marcado pelo *"Craft Station"* [12]

[12] https://www.craftstation.com.br/home

- Cervejas Especiais - no bairro Savassi de Belo Horizonte. É uma cervejaria que disponibiliza mais de 350 rótulos de cervejas especiais e ainda seis opções de chopes para consumo no local.

Representado pela cor roxa, o ponto da letra "C", são *checkins* no estabelecimento "Svärten Mugg" [13], localizado também no bairro Savassi. Uma *pub* com temática Nórdica - Viking. Ainda, alguns outros locais foram considerados para a formação deste *cluster*, como um Hotel e duas empresas não relacionadas à área cervejeira.

Os pontos com a letra "D", na cor rosa, também localizado no bairro Savassi, é marcado pela presença de muitos *checkins* no *Stadt Jever*[14] e no *Beb's Bar*[15]. O primeiro é um *pub* com uma temática baseada nos *pubs* Europeus e culinária Alemã. O Segundo, por sua vez, é um bar recente, criado em 2015, que oferece menu diferenciado e opções de cervejas, em sua maioria industrializadas.

Somente o ponto da letra "A", que se diferencia por estar localizado no bairro Lourdes. Contudo, o que impede a criação da "Rua da Cerveja" neste local é a discordância da característica C2. Isto, tendo em vista que este é um local densamente populado em uma área residencial.

Neste contexto, é possível perceber que grande parte dos *checkins* encontrados agrupam-se no bairro Savassi. Este poderia ser um indicativo para a criação de uma possível "Rua da Cerveja". Entretanto, esta é uma região próxima do centro da cidade, com grande fluxo de carros e pessoas. Tendo em vista os quatro critérios estabelecidos anteriormente, a criação desta rua torna-se inviável. Isto porque as características mínimas necessárias para estabelecer este tipo de local não é atendido nos *clusters* encontrados.

[13] http://www.svartenmugg.com.br
[14] http://www.stadtjever.com.br/
[15] http://www.bebsbar.com.br/

4.7.3 Áreas populares para o Rio de Janeiro

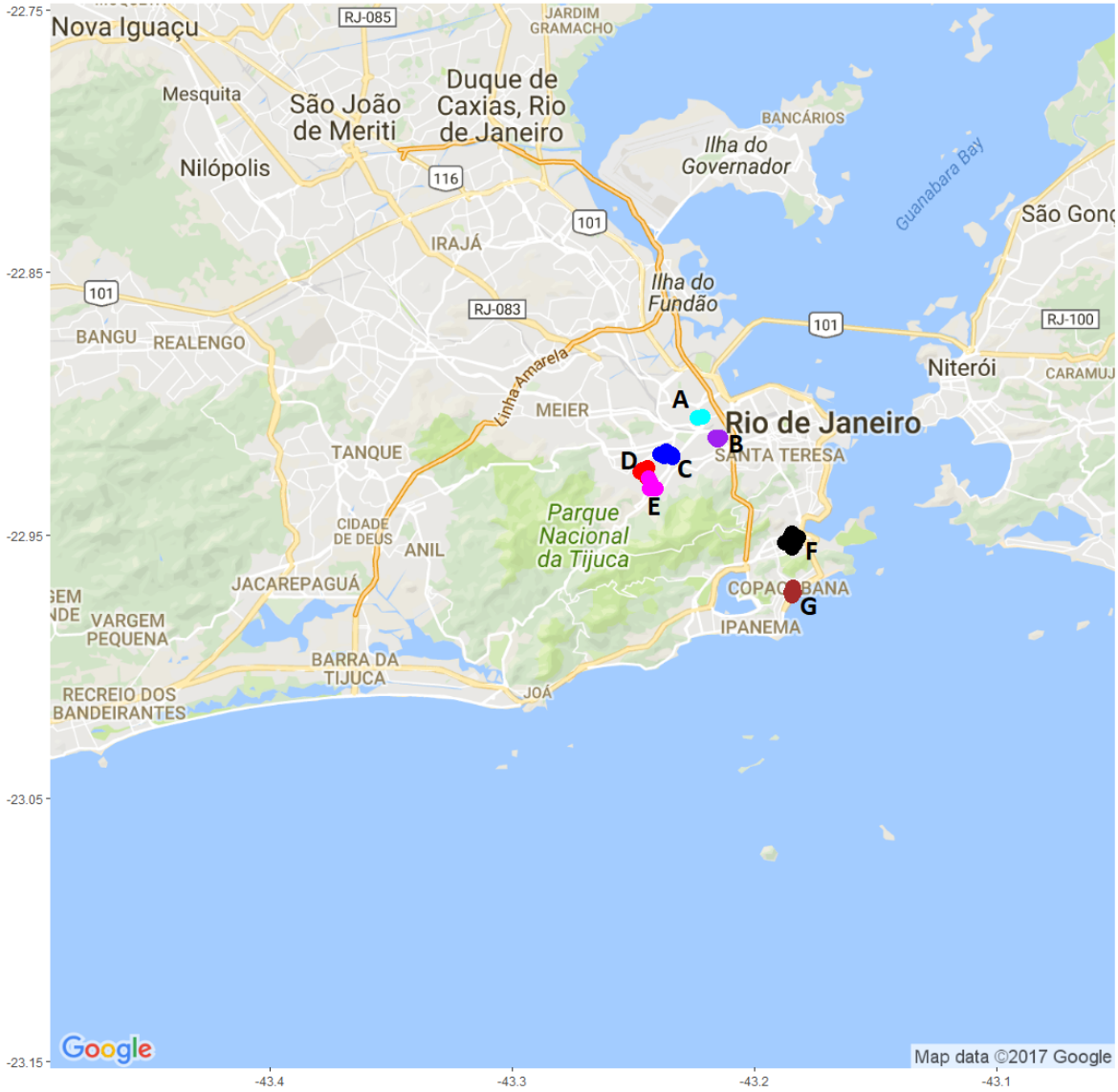


Figura 32: Rio de Janeiro

A Figura 32 representa o resultado do agrupamento para o Rio de Janeiro. Nesta cidade foi considerado o mínimo de vinte e cinco pontos dentro de um raio de 250 metros. Essa escolha foi motivada devido à densidade de dados coletados para esta região, o que poderia formar diversos *clusters* prejudicando a análise desenvolvida.

O ponto com a letra "A", representado na cor ciano, está localizado no bairro São Cristóvão. Os *checkins* coletados neste local são referentes ao evento "Gastro Beer Rio"[16], realizado no parque Quinta da Boa vista, onde está localizado o Zoológico e o Museu Nacional. O evento conta em média com cem rótulos de cervejas entre especiais, nacionais e importadas, além de uma variedade de pratos *gourmet* servidos.

16 https://www.gastrobeerrio.com.br/evento

Na cor roxa, o ponto com a letra "B", no bairro Praça da Bandeira, encontra-se os estabelecimentos *"Hop Lab"* e o *"Botto Bar"*, entre outros. O *"Hop Lab"* [17] conta com trinta torneiras de chope artesanal, servidos em 11 copos diferentes, além de comidas servidas para acompanhamento. Já o *"Botto Bar"* [18], conhecido também por *"Bar do Mestre Cervejeiro"* possui 20 torneiras de chope, onde procura-se servir cerveja com estilos Belga, Alemã, Inglesa e Norte Americana.

Por sua vez, o ponto da letra "C", demarcado na cor azul, encontra-se nas proximidades do estádio do Maracanã. Nesta região encontram-se diversos bares e restaurantes, como o Cerveja Social Clube, On Tap Pub, Hopfen Artesanal e Vikings Steak e Sandwiches. No ponto da letra "D", representado pela cor vermelha, são encontrados *checkins* na "Alquimia do Chopp", Santo Galo e Texas Burger.

Sinalizado na cor rosa, o ponto da letra "E", é representado pelos bares Ge bistro, Bar do Momo e Otto Bar. O ponto da letra "F", com a cor preta, é demarcada pelos bares da região do Botafogo. Por fim, o ponto da letra "G", na cor marrom, situa-se próximo a Orla de Copacabana.

Considerando os agrupamentos encontrados o local mais adequado para uma possível "Rua da Cerveja" seria parque Quinta da Boa vista. Isso deve-se à existência de uma infra estrutura adequada para suportar dez mil pessoas ou mais, comprovado pelos eventos que ocorrem atualmente nesta região. Além de não estar situado em uma área residencial que sofreria grandes impactos com essa implantação. O mesmo aplica-se às áreas comerciais existentes.

Os demais pontos poderiam oferecer recursos adequados para a implantação desta rua, entretanto alguns deles - como "B", "C", "D", "E", e "G" são locais de fluxo intenso de veículos e pessoas, comprometendo a característica C1. Além de possuírem hospitais em seus arredores, como é o caso dos pontos "C" e "D", o que viola a característica C2.

[17] $http://visit.rio/onde\_comer/hop-lab$
[18] http://www.bottobar.com.br

4.7.4 Áreas populares para São Paulo

Figura 33: São Paulo

A Figura 33 é o resultado do agrupamento realizado para a cidade de São Paulo. Para gerar os agrupamentos foram utilizados como parâmetros o número mínimo de pontos igual a 35 e o raio de 250 metros.

O ponto na cor azul escuro, correspondente à letra “A”, é referente à Pizzaria Brascatta e Almada‘s Beer Store, localizado no bairro Alto Da Lapa. Esta segunda caracteriza-se por ser uma loja destinada à venda de cervejas artesanais.

Na região agrupada em cor vermelha, e demarcada com a letra “B”, encontra-se a Rua Augusta [19] no bairro da Consolação. Esta rua é um dos centros Bohemios mais conhecidos da

[19] http://www.cidadedesaopaulo.com/sp/o-que-visitar/atrativos/pontos-turisticos/3308-rua-augusta

cidade de São Paulo, frequentada por jovens desde a década de 1960. Ela reúne diversos bares e baladas e recebe diversas atividades artísticas, como shows e exposições.

O ponto representado pela cor roxa, na letra "C", próximo ao Parque Ibirapuera, refere-se à Cervejoteca no bairro Vila Mariana e na Rua Sena Madureira, onde localiza-se os *checkins* coletados. Entretanto, após o período de coletas, o estabelecimento mudou de endereço, como é possível identificar no *website* [20].

Os pontos próximos à Praça Pôr do Sol, representado pelas letras "D", "E", "F", "G", "H" e "I", respectivamente pelas cores preta, verde escuro, rosa, marrom, ciano e verde claro localizado no Distrito de Pinheiros. Entre os diversos locais que apareceram neste *cluster*, encontram-se o Beerxp Labs, De bruer Bar, São Paulo Tap House, Cervejaria Nacional, Ambar Cervejas Artesanais e, finalmente, o Empório Alto dos Pinheiros, que possui 33 torneiras de chope e 650 rótulos diferentes conforme já descrito na Seção 4.1.

Considerando a possibilidade de criação da "Rua da Cerveja" em São Paulo, um dos locais que possibilitam a implantação desta rua, conforme as características descritas anteriormente, é o ponto "I", onde encontra-se o Empório Alto dos Pinheiros. Desta forma, o que favorece esta escolha são as características da região, pois apesar de apresentar algumas residências nas proximidades, não trás grandes impactos quando considerada a instalação da Rua neste local, o que atende as características C1 e C2 . Na região existem poucos comércios, o que atende a característica C3. Além disso torna-se possível acomodar uma grande quantidade pessoas, atendendo a característica C4.

[20] http://www.cervejoteca.com.br/sobre

5 Conclusões

Neste trabalho foram explorados aspectos referentes à Inteligência Coletiva e Computação Urbana, apropriando-se de conceitos da Mineração de dados, Métodos de Agrupamento e Métodos Hierárquicos. Para isso, foram coletados e tratados dados provenientes do *Twitter*. Sobre estes dados foram realizadas análises que tinham como objetivo compreender algumas das características relacionadas ao consumo de cerveja, bem como, identificar aspectos culturais associados às preferências pelos tipos de cerveja.

Durante a etapa de coleta dos dados foram encontradas algumas dificuldades. Inicialmente, falhas no armazenamento impossibilitaram a coleta dos dados durante alguns dias. Além disso, grande parte dos dados provenientes do *Twitter* não possuíam georreferenciamento, o que levou ao descarte de parte destes dados. Entretanto, nenhum dos problemas trouxeram impactos negativos na análise.

Desta forma, a investigação realizada sobre os aspectos culturais teve como objetivo apresentar como as características étnicas podem estar associadas ao consumo de cervejas. Para a realização desta análise foi aplicado, para o cálculo da matriz de distâncias, a medida de similaridade *Canberra* e o método *Ward 2* para o agrupamento. Nesta análise foram consideradas duas perspectivas, uma para a quantidade de *checkins*, representando a preferência dos usuários em quantidade de cervejas, e outra associando a quantidade à nota atribuida, a fim de avaliar preferências através das notas.

Através da análise de agrupamentos, foram apresentados resultados que indicam possíveis influências culturais no consumo de cervejas. Como o exemplo da Cidade do México agrupada às cidades brasileiras. Isto indica, provavelmente, a influência de aspectos étnicos, dada a próximadade geográfica e ligação à cultura latina. Outro exemplo são as cidades dos Estados Unidos, todas contidas em um único agrupamento, quando consideradas as notas dos usuários.

Ainda, na análise de aspectos culturais, foi proposto um exemplo de aplicação dos métodos de agrupamento utilizado na análise. Este exemplo consistiu de um estudo de caso onde foi simulado um possível sistema de sugestão de destinos turísticos. Para esta simulação foram separados dez usuários assíduos do aplicativo e acrescentados à base de cidades, auxiliado pelos métodos de agrupamentos, os locais mais próximos de onde estes usuários foram agrupados poderiam ser sugeridos como destino à eles. A aplicação deste método levou à sugestão de possíveis destinos turísticos para todos os usuários em questão, considerando suas preferências por cervejas.

Neste contexto, também foi proposta uma abordagem para identificar áreas populares para o consumo de cervejas artesanais, considerando um número mínimo de *checkins* dentro de

um determinado raio. A motivação para esta análise surgiu a partir da possibilidade de criação da “Rua da Cerveja” na cidade de Curitiba.

A partir da aplicação do algoritmo *DBScan* sobre um plano cartesiano, dadas as coordenadas geográficas das quatro cidades brasileiras, formaram-se *clusters*. Estes *clusters* indicam regiões, já conhecidas pela população local, famosas por abrigar bares, *pubs* e restaurantes nestas cidades. Desta forma, foi possível avaliar estes locais e identificar pontos propícios para a instalação de uma "Rua da cerveja"nas cidades analisadas. Auxiliando assim, possíveis tomadores de decisão, quando considerados aspectos de planejamento urbano.

Esta análise exemplifica a eficácia da utilização deste método com a finalidade de identificar os locais em que existam grandes quantidades de usuários de acordo com os parâmetros pré determinados. Além disso, demonstra a importância que este método pode ter no auxílio à tomada de decisão, visto que são informações dos acontecimentos em tempo real. Em que, usuários estão cedendo informações publicamente, gerando uma espécie de popularidade espontânea e potencializando a exploração dos aspectos urbanos e sociais. Isto pode influenciar a forma como as decisões são tomadas em uma cidade, quanto às questões urbanas. Pois, explorando informações como esta, uma prefeitura pode decidir capitalizar ou inviabilizar o aspecto em questão.

Considerando as análises feitas e resultados obtidos, em termos gerais, este trabalho atingiu com sucesso os objetivos propostos. Redes sociais mostram-se muito úteis como ferramenta para extrair informações relevantes sobre características dos usuários. Os dados públicos são de fácil acesso e coleta, e podem beneficiar os responsáveis por tomadas de decisão para os mais diversos fins, entre elaborar estratégias de *marketing*, realizar tendências, personalização, características relevantes para os consumidores, entre outros. Além disso, o grande alcance das redes sociais permite que sejam extraídas informações do mundo inteiro, o que favorece a compreensão de aspectos culturais devido aos dados em larga escala. Vale evidenciar que este tipo de análise pode ser realizado para qualquer objeto de estudo, não sendo restrito necessariamente às cervejas artesanais.

6 Trabalhos futuros

Através do desenvolvimento deste trabalho foi possível identificar diversas características relacionadas ao consumo de cerveja. Entretanto, o escopo foi delimitado a nível de cidades, destinando-se a compreender aspectos étnicos, urbanos e culturais. Sendo assim, este estudo pode ser utilizado como base para a elaboração de outros projetos e análises mais específicas. Na sequência serão sugeridos alguns trabalhos que podem ser desenvolvidos.

Um dos possíveis estudos consiste na identificação de particularidades que tornam um determinado local o mais frequentado, como exemplo: características do ambiente, métodos de pagamento, acessibilidade e ofertas. Esta pesquisa pode ser realizada através de uma análise de algumas características ao redor deste local, como: bares, restaurantes, shoppings, entre outros. Além dessas características é possível identificar qual é a opinião dos usuários quanto ao bairro, comércios nas proximidades e do próprio estabelecimento em foco na análise.

Ainda, nesta direção, o conteúdo disponibilizado em redes sociais pode conter informações das opiniões dos usuários. Sendo assim, é possível avaliar o sentimento dos indivíduos por meio do conteúdo da publicação, através do texto escrito. Ainda, outro parâmetro que poderia ser utilizado para avaliação são os *emotions*, hoje em dia muito usado por diversos usuários de redes sociais. Este tipo de análise pode se tornar relevante para auxiliar em pesquisas de satisfação e compreensão de aspectos que motivam pessoas a ter determinados comportamentos ou frequentar um local específico.

Dados para compor estas análises podem ser obtidos através de serviços baseados em localização, oferecidos em plataformas *mobile*, como o *Foursquare* e o *Google Maps*. Através do *Foursquare*, em um raio de distância previamente estipulado, é possível identificar algumas características da região a fim de detectar aspectos que fazem aquele local ser popular. Além disso, é possível procurar informações dos consumidores a respeito do estabelecimento através do próprio *Foursquare* ou ainda em uma API disponibilizada pelo *Google Maps*.

Outro possível estudo, a partir das informações adquiridas, é o de estabelecer uma análise a nível de usuário. É possível obter informações relacionadas à preferência destes por determinados tipos de cerveja, bem como, os locais onde preferem tomá-las. Através deste método de análise, torna-se possível compreender aspectos culturais relacionados ao consumo de cerveja, bem como, identificar potenciais padrões não explorados antes. Estes aspectos podem vir a ser benéficos quando considerada a necessidade de compreender o mercado de consumidores de cervejas artesanais ou padrões comportamentais relacionados a este público.

Referências

ALA-MUTKA, K. M. et al. *The Impact of Social Computing on the EU Information Society and Economy*. [S.l.], 2009. Disponível em: <http://EconPapers.repec.org-/RePEc:ipt:iptwpa:jrc54327>. Citado na página 21.

ASSOCIATION, B. *Brewers Association 2017 beer style guidelines*. 2017. Citado na página 45.

BARAJAS, M.; BOEING, G.; WARTELL, J. *Neighborhood Change, One Pint at a Time: The Impact of Local Characteristics on Craft Breweries*. 2017. 155-176 p. Citado na página 36.

BEMBEM, A. H. C.; SANTOS, P. L. V. A. da C. Inteligência coletiva: um olhar sobre a produção de pierre lévy. *Perspectivas em Ciência da Informação*, scielo, v. 18, p. 139 – 151, 12 2013. ISSN 1413-9936. Disponível em: <http://dx.doi.org/10.1590/S1413-99362013000400010>. Citado 2 vezes nas páginas 19 e 20.

BRABHAM, D. C. *Crowdsourcing*. [S.l.: s.n.], 2013. Citado na página 20.

BURKE, J. A. et al. Participatory sensing. *Center for Embedded Network Sensing*, 2006. Citado na página 22.

CABENA, P. et al. *Discovering Data Mining: From Concept to Implementation*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1998. (An IBM Press Book Series). ISBN 0-13-743980-6. Citado 4 vezes nas páginas 24, 25, 26 e 27.

CHORLEY, M. et al. Pub crawling at scale: Tapping untappd to explore social drinking. In: *Tenth International AAAI Conference on Web and Social Media*. [s.n.], 2016. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13048>. Citado 4 vezes nas páginas 15, 35, 36 e 39.

CRANSHAW, J. et al. The livehoods project: Utilizing social media to understand the dynamics of a city. In: *International AAAI Conference on Weblogs and Social Media*. [S.l.: s.n.], 2012. p. 58. Citado 2 vezes nas páginas 14 e 37.

CUMMINS, R. A.; GULLONE, E. Why we should not use 5-point likert scales: The case for subjective quality of life measurement. In: *Proceedings, second international conference on quality of life in cities*. [S.l.: s.n.], 2000. p. 74–93. Citado na página 61.

EAPSP. *Empório Alto dos Pinheiros - Home*. 2016. Disponível em: <http://www.eapsp.com-.br/>. Nenhuma citação no texto.

ESTER, M. et al. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1996. (KDD'96), p. 226–231. Disponível em: <http://dl.acm.org/citation.cfm?id=3001460-.3001507>. Citado 2 vezes nas páginas 30 e 31.

EVERITT, B. S. *The Cambridge dictionary of statistics*. [S.l.]: Cambridge University Press, 2006. Citado 2 vezes nas páginas 33 e 65.

EVERITT, B. S. et al. Cluster analysis. John Wiley & Sons, p. 71–110, 2011. Citado 5 vezes nas páginas 24, 32, 33, 34 e 35.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, p. 37–54, 1996. Citado na página 24.

GOVERS, R. From place marketing to place branding and back. *Place Branding and Public Diplomacy*, Springer, v. 7, n. 4, p. 227–231, 2011. Citado na página 41.

HO, A. Big data and evidence-driven decision-making: Analyzing the practices of large and mid-sized us cities. In: *Proceedings of the 50th Hawaii International Conference on System Sciences*. [S.l.: s.n.], 2017. Citado 2 vezes nas páginas 37 e 41.

HODGE, D. R.; GILLESPIE, D. Phrase completions: An alternative to likert scales. *Social Work Research*, Oxford University Press, v. 27, n. 1, p. 45, 2003. Citado na página 61.

JONES, P. et al. Maplocal: use of smartphones for crowdsourced planning. *Planning Practice & Research*, Taylor & Francis, v. 30, n. 3, p. 322–336, 2015. Citado 3 vezes nas páginas 37, 38 e 41.

KAMIENSKI, C. et al. Xxxiv simpósio brasileiro de redes de computadores e sistemas distribuídos (sbrc 2016). In: (UEFS), A. A. T. R. C. (Ed.). *Livro de Minicursos SBRC 2016*. Sociedade Brasileira de Computação (SBC), 2016. cap. II, p. 51–100. Disponível em: <https://www.researchgate.net/profile/Alexandre_Heideker/publication-/303810868_Computacao_Urbana_Tecnologias_e_Aplicacoes_para_Cidades_Inteligentes/links-/576809b808ae8ec97a423e6b.pdf>. Citado 2 vezes nas páginas 22 e 23.

KARAMSHUK, D. et al. Geo-spotting: mining online location-based services for optimal retail store placement. In: ACM. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2013. p. 793–801. Citado 3 vezes nas páginas 36, 37 e 39.

KEMP, K. *Encyclopedia of Geographic Information Science*. SAGE Publications, 2007. ISBN 9781452265605. Disponível em: <https://books.google.com.br/books?id=clF2AwAAQBAJ>. Citado na página 43.

KINDBERG, T.; CHALMERS, M.; PAULOS, E. Guest editors' introduction: Urban computing. *IEEE Pervasive Computing*, IEEE, v. 6, n. 3, p. 18–20, 2007. Citado 2 vezes nas páginas 21 e 22.

LÉVY, P. *Cibercultura [Cyberculture]*. [S.l.: s.n.], 1999. Citado 2 vezes nas páginas 18 e 19.

LÉVY, P. *A inteligência coletiva: por uma Antropologia do ciberespaço.* [S.l.: s.n.], 2003. Citado 2 vezes nas páginas 18 e 19.

LIU, B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2007. (Data-Centric Systems and Applications). ISBN 978-3-540-37881-5. Disponível em: <http://dx.doi.org/10.1007/978-3-540-37882-2>. Citado 4 vezes nas páginas 24, 25, 27 e 28.

LORETO, V. et al. *Participatory Sensing, Opinions and Collective Awareness*. [S.l.]: Springer, 2016. Citado na página 22.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297. Citado na página 28.

MALONE, T. W. What is collective intelligence and what will we do about it. *Collective Intelligence: Creating a Prosperous World at Peace, Earth Intelligence Network, Oakton, Virginia*, p. 1–4, 2008. Citado 2 vezes nas páginas 18 e 20.

MIRKES, E.; LEICESTER, U. of. *K-means and K-medoids applet.* 2012. Disponível em: <http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids-.html>. Citado na página 29.

MOSHER, K. T. a. M. *Brewing Science: A Multidisciplinary Approach*. 1. ed. Springer International Publishing, 2017. ISBN 978-3-319-46394-0,978-3-319-46393-3. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=326914D7CB3E04DCDE687D7572F80DD4>. Citado na página 62.

PALOMINO, P. T.; ANDRADE, L. A. de. Perspectivas em redes sociais: A inteligência coletiva como ferramenta de análise de métricas e indicadores de desempenho. In: *Geminis - Abordagens Multiplataformas*. [s.n.], 2013. Disponível em: <http://www.revistageminis-.ufscar.br/index.php/geminis/article/view/152>. Citado 2 vezes nas páginas 19 e 20.

PARAMESWARAN, M.; WHINSTON, A. B. Social computing: An overview. *Communications of the Association for Information Systems*, v. 19, n. 1, p. 37, 2007. Citado na página 21.

PAULOS, E.; ANDERSON, K.; TOWNSEND, A. Ubicomp in the urban frontier. In: *Workshop at Ubicomp 2004*. [s.n.], 2004. Disponível em: <http://www.paulos.net/papers-/2004/Urban%20Frontier%20Workshop%20(UbiComp%202004).pdf>. Citado na página 21.

PAULOS, E.; GOODMAN, E. The familiar stranger: Anxiety, comfort, and play in public places. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2004. (CHI '04), p. 223–230. ISBN 1-58113-702-8. Disponível em: <http://doi.acm.org/10.1145/985692.985721>. Citado na página 21.

PREOŢIUC-PIETRO, D.; COHN, T. Mining user behaviours: a study of check-in patterns in location based social networks. In: ACM. *Proceedings of the 5th Annual ACM Web Science Conference*. [S.l.], 2013. p. 306–315. Citado na página 41.

PYTHON. *Python.org: What is Python? Executive Summary.* 2016. Disponível em: <https://www.python.org/doc/essays/blurb/>. Nenhuma citação no texto.

SANTALA, V. et al. Making sense of the city: Exploring the use of social media data for urban planning and place branding. 2017. Citado 2 vezes nas páginas 14 e 71.

SCHULER, D. Social computing - introduction to the special section. *Commun. ACM*, v. 37, n. 1, p. 28–29, 1994. Disponível em: <http://doi.acm.org/10.1145/175222.175223>. Citado na página 21.

SEBRAE. *Sebrae : Cervejas Artesanais*. 2015. Disponível em: <https://www-.sebraeinteligenciasetorial.com.br/produtos/relatorios-de-inteligencia/cervejas-artesanais-/55c4ad3614d0c01d007ffeae>. Citado na página 14.

SEGARAN, T. Programming collective intelligence. In: . [S.l.: s.n.], 2007. Citado 3 vezes nas páginas 18, 19 e 20.

SILVA, H. A.; LEITE, M. A.; PAULA, A. R. V. de. Cerveja e sociedade. *Contextos da Alimentação–Revista de Comportamento, Cultura e Sociedade, São Paulo: Centro Universitário Senac*, 2015. Citado na página 14.

SILVA, T. Para entender o monitoramento de mídias sociais. In: . [S.l.: s.n.], 2012. Citado na página 20.

SILVA, T. H.; GRAEML, A. *Exploring Collected Intelligence from Untappd to Support the Location Decision for New SMEs*. 2016. Citado 11 vezes nas páginas 15, 35, 39, 40, 42, 46, 47, 48, 52, 55 e 72.

SILVA, T. H.; LOUREIRO, A. A. F. Computação urbana: Técnicas para o estudo de sociedades com redes de sensoriamento participativo. In: *Anais da XXXIV Jornada de Atualização em Informática*. Sociedade Brasileira de Computação – SBC, 2015. v. 8329, p. 68–122. ISBN 978-85-88442-99-3. Disponível em: <http://homepages.dcc.ufmg.br/~thiagohs/papers-/textoJAI.pdf>. Citado 3 vezes nas páginas 22, 23 e 24.

SILVA, T. H. et al. A large-scale study of cultural differences using urban data about eating and drinking preferences. *Information Systems*, Elsevier, v. 72, p. 95–116, 2017. Citado na página 15.

SIMPLEJSON. *Simple, fast, extensible JSON encoder/decoder for Python*. 2016. Disponível em: <https://pypi.python.org/pypi/simplejson>. Nenhuma citação no texto.

STEINHAUS, H. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, v. 1, n. 804, p. 801, 1956. Citado na página 28.

SUBHASH, S. Applied multivariate techniques. *John Wily & Sons Inc., Canada*, 1996. Citado na página 65.

TAN, P. N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. Ciencia Moderna, 2009. ISBN 9788573937619. Disponível em: <https://books.google.com-.br/books?id=69d6PgAACAAJ>. Citado 10 vezes nas páginas 24, 25, 26, 27, 28, 29, 31, 32, 34 e 35.

TWITTERAPI. *A Python wrapper arround Twitter API*. 2016. Disponível em: <https://github-.com/bear/python-twitter>. Nenhuma citação no texto.

UNTAPPD. *Untappd: About Untappd*. 2016. Disponível em: <https://untappd.com/about>. Nenhuma citação no texto.

VALE, M. N. do. *Agrupamentos de dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos*. Tese (Doutorado) — PUC-Rio, 2005. Citado 2 vezes nas páginas 28 e 32.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical machine learning tools and techniques*. [S.l.: s.n.], 2005. Citado 3 vezes nas páginas 28, 29 e 32.

ZHENG, Y. et al. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM, v. 5, n. 3, p. 38, 2014. Citado 2 vezes nas páginas 23 e 24.

ANEXO A – *Heatmaps* utilizados na escolha das cidades para o desenvolvimento do estudo

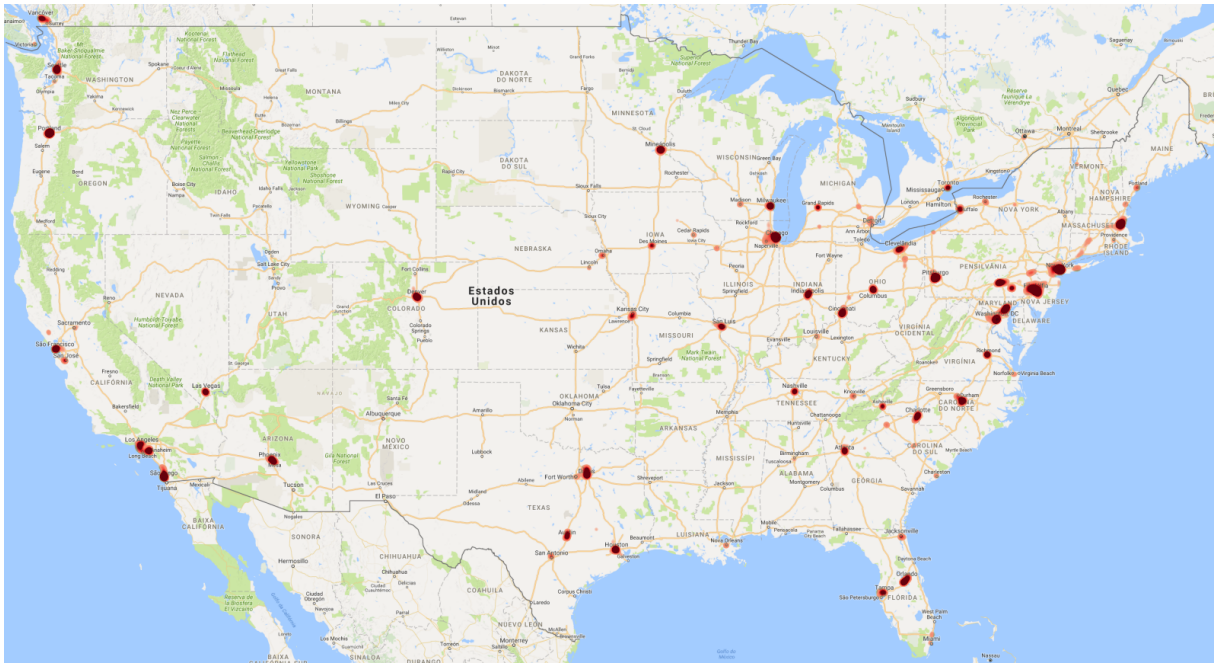


Figura 34: *Heatmap* da densidade de *checkins* nos Estados Unidos

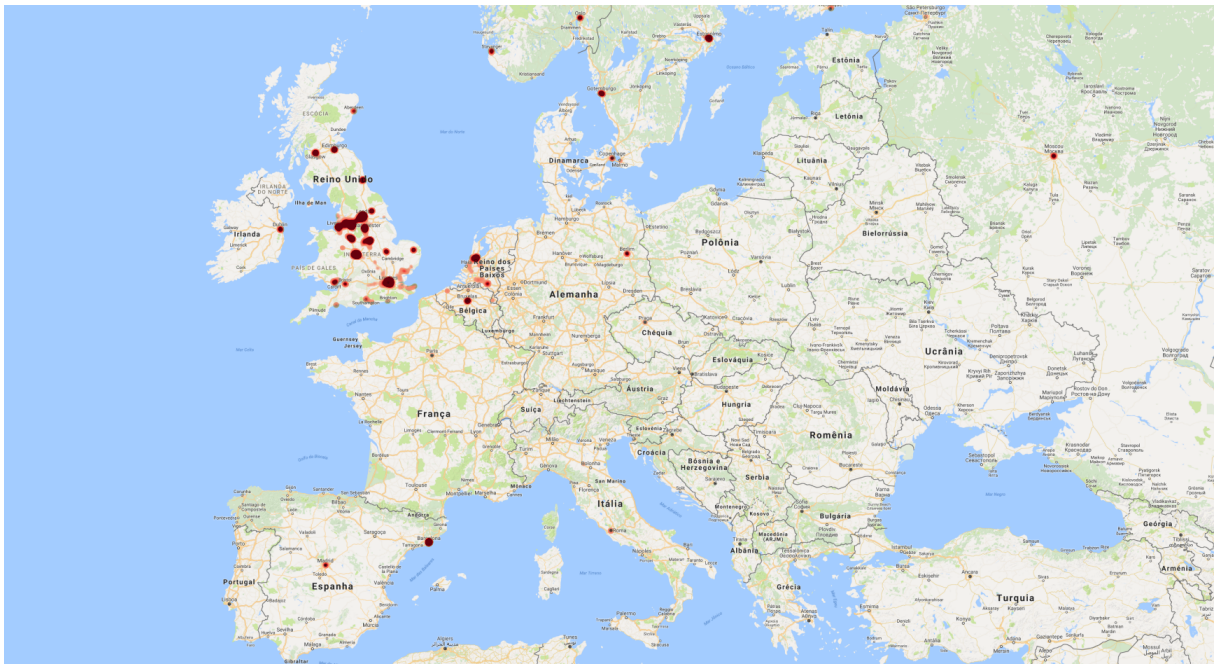


Figura 35: *Heatmap* da densidade de *checkins* no continente Europeu

Figura 36: *Heatmap* da densidade de *checkins* na América do Sul

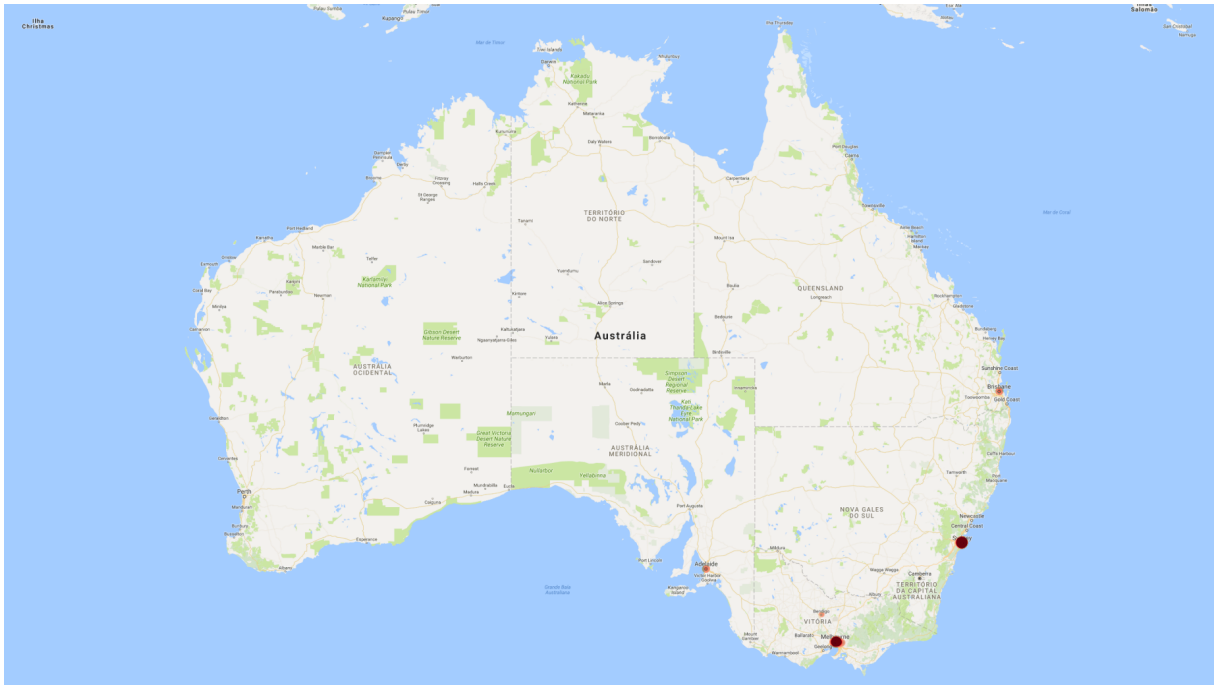


Figura 37: *Heatmap* da densidade de *checkins* na Austrália

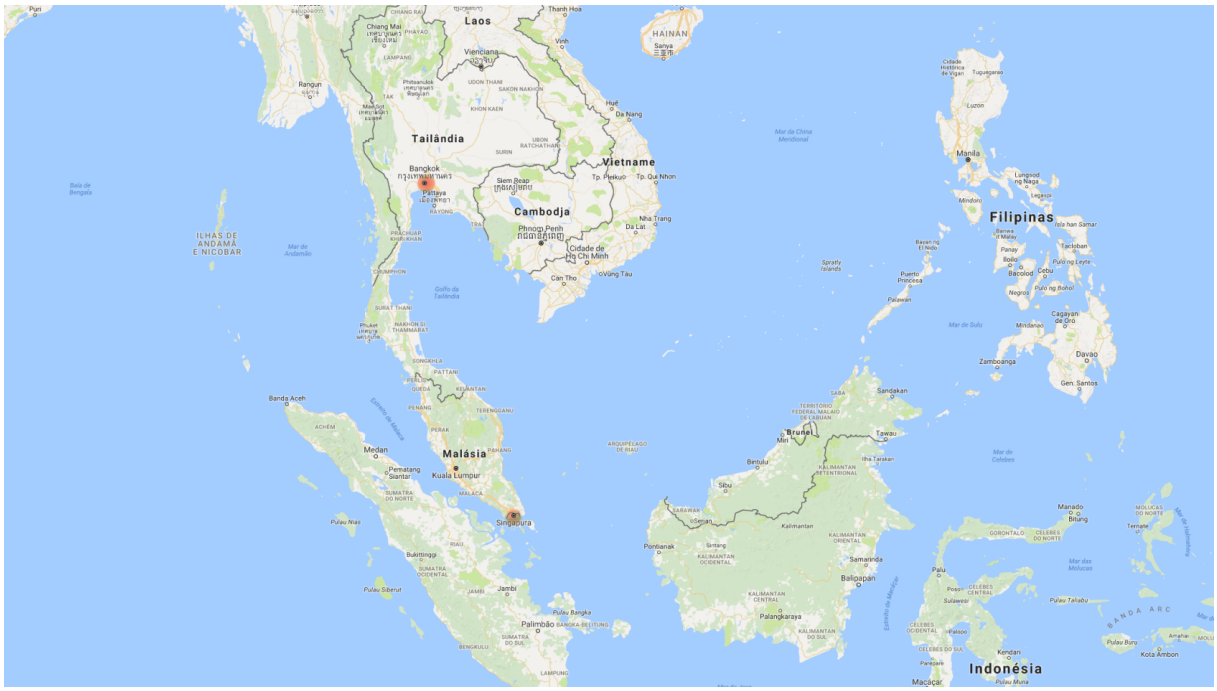


Figura 38: *Heatmap* da densidade de *checkins* na Ásia

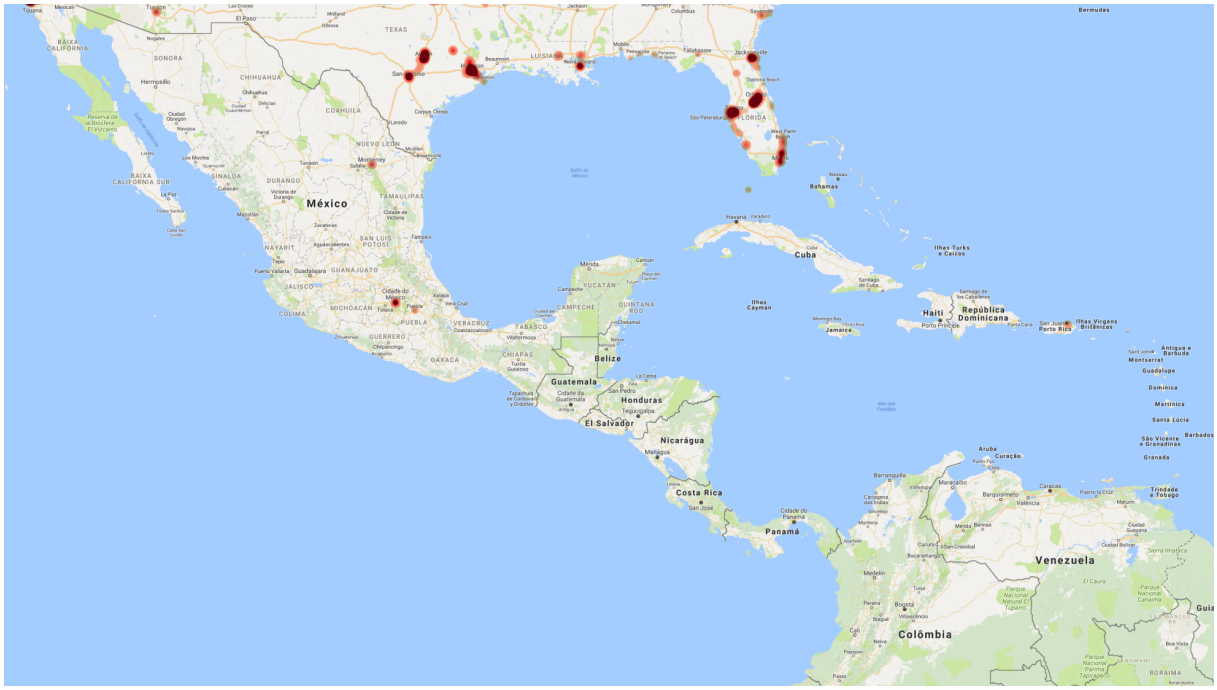


Figura 39: *Heatmap* da densidade de *checkins* na América Central

ANEXO B – Classificações dos tipos de cervejas

ID	CERVEJA
1	Altbier
2	Applewine
3	Australian Sparkling Ale
4	Barleywine
5	Belgian Dubbel
6	Belgian Quad
7	Belgian Strong Dark Ale
8	Belgian Strong Golden Ale
9	Belgian Tripel
10	Bière de Champagne / Bière Brut
11	Bière de Garde
12	Black & Tan
13	Blonde Ale
14	Bock
15	Braggot
16	Brown Ale
17	Burton Ale
18	California Common
19	Cider
20	Cream Ale
21	Cyser
22	Dampfbier
23	Dark Ale
24	Doppelbock
25	Dunkelweizen
26	Eisbock
27	English Bitter
28	English Mid Ale
29	English Pale Ale
30	Extra Special / Strong Bitter
31	Faro

ID	CERVEJA
32	Festbier
33	Fruit Beer
34	Ginger Beer
35	Gluten-Free
36	Golden Ale
37	Grisette
38	Gruit / Ancient Herbed Ale
39	Grätzer / Grodziskie
40	Happoshu
41	Hefeweizen
42	Hefeweizen Light / Leicht
43	IPA
44	Imperial Pale Ale
45	Kellerbier / Zwickelbier
46	Kombucha
47	Kristallweizen
48	Kvass
49	Kölsch
50	Lager
51	Lambic
52	Maibock / Heller (Helles) Bock
53	Malt Beer
54	Malt Liquor
55	Mead
56	Melomel
57	Mumme
58	Märzen
59	Non-Alcoholic
60	Old Ale
61	Other
62	Pale Ale

ID	CERVEJA
63	Pale Wheat Ale - American
64	Patersbier
65	Perry
66	Pilsner
67	Porter
68	Pumpkin / Yam Beer
69	Pyment
70	Rauchbier
71	Red Ale
72	Roggenbier
73	Root Beer
74	Rye Beer
75	Rye IPA
76	Sahti
77	Saison / Farmhouse Ale
78	Schwarzbier
79	Scottish Ale
80	Shandy / Radler
81	Smoked Beer
82	Sour
83	Specialty Grain
84	Spiced / Herbed Beer
85	Stout
86	Strong Ale
87	Traditional Ale
88	Weizenbock
89	Wheat Wine
90	Winter Ale
91	Winter Warmer
92	Witbier
93	Zoigl

Tabela 5: Classificação 1 das cervejas

ID	Classificação 2
1	British Ale
2	Irish Ale
3	North America Ale
4	German Ale
5	Belgian and French Ale
6	Other Ales
7	Germanic Lager
8	North America Lager
9	Other Lager
10	Mixed or Other Types

Tabela 6: Classificação 2 das cervejas - por *Brewers Association*