André Carvalho Machado

ANÁLISE E CORRELAÇÃO DE DADOS: UM ESTUDO DE CASO USANDO O AIRBNB E O TRIPADVISOR EM FLORIANÓPOLIS

Florianópolis – SC

2022

André Carvalho Machado

ANÁLISE E CORRELAÇÃO DE DADOS: UM ESTUDO DE CASO USANDO O AIRBNB E O TRIPADVISOR EM FLORIANÓPOLIS

Proposta de Trabalho de Conclusão de Curso para obtenção do grau de Bacharel no curso de Sistemas de informação na Universidade Federal de Santa Catarina.

Universidade Federal de Santa Catarina – UFSC

Departamento de Informática e Estatística – INE

Graduação em Sistemas de Informação

Orientador: Carina Friedrich Dorneles

Florianópolis – SC

2022

André Carvalho Machado

ANÁLISE E CORRELAÇÃO DE DADOS: UM ESTUDO DE CASO USANDO O AIRBNB E O TRIPADVISOR EM FLORIANÓPOLIS

Trabalho aprovado. Florianópolis – SC, ___ de _____ de 2022:

Carina Friedrich Dorneles
Orientador

Professor Jônata Carvalho
Banca 1

Professor Mateus Grellert
Banca 2

Florianópolis – SC
2022

Agradecimentos

À minha mãe, Rejane, e a minha avó Bete por todo o carinho e amor dedicado durante a minha trajetória e durante esse processo.

À minha namorada, Danielle Nunes, pelo apoio demonstrado ao longo de todo o período de tempo em que me dediquei a este trabalho.

Agradeço à todos, minha família, parentes e amigos que com seu incentivo me fizeram chegar à conclusão do meu curso e começo de uma nova carreira

E a todos que direta ou indiretamente fizeram parte da minha formação, os meus mais sinceros muito obrigado.

Resumo

O setor de hospedagem nacional apresentou um cenário de crescimento conforme o passar dos anos, grande parte devido ao acesso a internet e a facilidade de reservas através de apps. Além disso, com o acesso facilitado, surgiram sites com fóruns de viagens interativos, fornecendo informações e opiniões de viajantes com conteúdos relacionados ao turismo. A quantidade de dados gerados por usuários da web cresce cada vez mais, por outro lado existe a complexidade na extração e obtenção desses dados. Os Web Scrapers que nada mais são do que coletores de dados web, estão dispostos a rastrear grandes conteúdos de uma página web online que são analisados e indexados, e depois disponibilizam o conteúdo aos usuários. Diante dessas ferramentas, o estudo utiliza Web Scraping para a coleta dos dados, com a intenção de criar pontos para análises, através de consultas em modelos dimensionais. As análises buscam trazer gráficos informativos e indicadores, mostrando ao setor de turismo uma relação entre a plataforma AirBNB do setor de hospedagem e da plataforma Tripadvisor do setor de avaliações turísticas através das identificações de relação dos atributos fornecidos pela hospedagem, as avaliações e comentários. Permitindo assim visualizar o impacto das mesmas sobre o setor.

Palavras-chave: Web Scraping, Extração, Turismo, Nacional, Airbnb, Tripadvisor.

Abstract

The national hosting sector presented a growth scenario over the years, largely due to internet access and the ease of booking through apps. In addition, with easier access, websites with interactive travel forums emerged, providing information and opinions from travelers with content related to tourism. The amount of data generated by web users grows more and more, on the other hand there is complexity in extracting and obtaining this data. Web Scrapers, which are nothing more than web data collectors, are willing to crawl large contents of an online web page that are analyzed and indexed, and then make the content available to users. In view of these tools, the study uses Web Scraping to collect data, with the intention of creating points for analysis, through queries in dimensional models. The analyzes seek to bring informative graphs and indicators, showing the tourism sector a relationship between the AirBNB platform of the accommodation sector and the Tripadvisor platform of the tourism evaluation sector through the identifications of the relationship of the attributes provided by the accommodation, the evaluations and comments. Thus allowing to visualize their impact on the sector.

Keywords: Web Scraping, Extraction, Tourism, National, Airbnb, Tripadvisor.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CSV	Comma Separated Values
ETL	Extract Transform Load
BI	Business Intelligence
UFSC	Universidade Federal De Santa Catarina
UDESC	Universidade do Estado de Santa Catarina

Sumário

1 Introdução

Atualmente, a maior parte dos processos de informações turísticas são realizados eletronicamente, incluindo o aluguel de hospedagens. Os clientes deixam suas impressões digitais em grande parte das atividades realizadas tanto no planejamento da viagem como durante e após, mas também através de comentários sobre diferentes plataformas. Consequentemente, uma grande quantidade de dados sobre as necessidades e comportamentos dos clientes, bem como a sua percepção dos serviços, são armazenados em várias fontes (GARCíA et al., 2020). Com isso, surgem as aplicações como o Airbnb, que fornecem as mais variadas opções de hospedagens para os viajantes, trazendo facilidade e substituindo parte dos intermediários do plano, tais como as agências de turismo ou corretores imobiliários.

Os conteúdos criados pelos viajantes são percebidos como altamente confiáveis, credíveis e relevantes, atualizados e envolventes. Assim, as pessoas que planejam viagens geralmente levam em consideração as avaliações geradas por outros viajantes durante o processo de tomada de decisão, pois a intangibilidade das experiências de turismo impossibilita o teste de pré-compra e, portanto, aumenta a necessidade de relatórios de experiência em primeira pessoa (YOO; SIGALA; GRETZEL, 2016). Aplicações como o TripAdvisor, surgem com o intuito de suprir essa necessidade em uma plataforma única, onde é criado um fórum de viajantes, compartilhando as mais variadas experiências.

Enquanto o significado computacional de uma plataforma é uma 'infra-estrutura programável sobre a qual outro software pode ser construído e executado', no discurso público o termo plataforma é cada vez mais usado para descrever empresas que oferecem serviços web 2.0 e oferecem uma oportunidade de comunicação, interação e vendas. As plataformas de turismo estão centradas principalmente na mobilidade, acomodação, alimentação e experiências de viagem. O Airbnb faz parte de um conjunto específico de plataformas digitais que facilitam a troca monetária de acomodações residenciais (casas particulares, quartos e leitos) e experiências turísticas entre os indivíduos (MINCA; ROELOFSEN, 2022). Já o TripAdvisor faz parte de outro conjunto, um site de informações com uma comunidade compartilhando conteúdo de viagens do mundo inteiro e capacitando os usuários a escrever, pesquisar e compartilhar resenhas de viagens (YOO; SIGALA; GRETZEL, 2016).

Com o passar do tempo e a facilidade das reservas das hospedagens, o número de usuários da plataforma vêm mudando, mudando assim os destinos e os padrões que eram comumente utilizados. O Airbnb é uma inovação desafiadora à qual a hospitalidade tradicional tem que se adaptar (OSKAM; BOSWIJK, 2015). A confiança desempenha o

papel principal para a tomada de decisão nas reservas de hospedagens. A plataforma age como intermediária e garante algumas seguranças para os seus usuários, incluindo avaliações dos próprios usuários sobre a experiência com a acomodação. Algo que a plataforma do TripAdvisor também se propõe a fazer, entretanto abrangendo outros segmentos.

Diante dessas informações, o presente estudo busca realizar a implementação de uma aplicação que coleta dados web de sites como o Airbnb e TripAdvisor, utilizando o processo de *Web Scraping*, extraindo os dados e convertendo-os em informações estruturadas. Os dados extraídos devem trazer os valores das hospedagens, a faixa de preços dos restaurantes, as localizações dentro da cidade, o ranking, as avaliações e comentários. Para armazená-los será utilizado um banco de dados.

O município de Florianópolis no estado de Santa Catarina foi selecionado para este estudo por ser um dos mais visitados destinos turísticos do Brasil. É um destino reconhecido mundialmente por suas belezas naturais, cercado de praias, e pela qualidade de vida que proporciona. A cidade é a capital brasileira com maior pontuação no Índice de Desenvolvimento Humano (IDH). A economia da cidade é fortemente baseada em tecnologia da informação, turismo e serviços (PERES; PALADINI, 2021).

Esta aplicação deve atender aos interesses de órgãos responsáveis pelo turismo nacional e empresas relacionadas ao turismo que buscam encontrar a média de preços utilizados nas acomodações, além da faixa de valores para os restaurantes, assim como os comentários e avaliações. Através da implementação de um esquema de visualização das informações, mapas e gráficos combinados o resultados devem ser apresentados através da plataforma PowerBI, e por fim a correlação entre esses dados.

1.1 Objetivo Geral

A proposta do trabalho tem como objetivo principal ser um passo inicial para mostrar a distribuição dos dados presentes no Airbnb e no TripAdvisor. De forma geral, a ideia é trazer mais informação ao setor do turismo nacional, quanto as acomodações e suas características assim como os valores empregados por parte dos anfitriões na plataforma do Airbnb, determinando a relação entre os preços praticados e suas médias, além da relação entre as localizações, avaliações e comentários. E também dos dados da plataforma de viagens TripAdvisor, determinando as faixas de preço, localizações, avaliações e comentários. Por fim a tentativa de traçar uma correlação entre as duas bases de dados.

1.2 Objetivos Específicos

Para atingir o objetivo geral, os seguintes objetivos específicos foram definidos:

i Coletar dados do site Airbnb e do site Tripadvisor, por meio de Web Scraping;

ii Criar representações visuais (gráficos, nuvens de palavras e mapas de calor) que mostrem o comportamento dos dados presentes nas bases de dados coletadas, servindo como indicadores para o ramo de turismo;

iii Identificar os valores das hospedagens na cidade de Florianópolis;

iv Identificar a faixa de preço dos restaurantes na cidade de Florianópolis;

v Identificar a relação entre os comentários e avaliações fornecidos pelas usuários das hospedagens no Airbnb e da plataforma TripAdvisor;

vi Disponibilizar um conjunto de indicadores de correlação entre os atributos;

2 Fundamentação Teórica

Neste capítulo, é discutido o referencial teórico necessário para compreensão da proposta apresentada pelo presente trabalho. Inicialmente, a fonte dos dados *Airbnb* é contextualizada, trazendo sua importância ao disponibilizar diversas acomodações online através de uma plataforma com diversas informações. Depois, a fonte de dados TripAdvisor é contextualizada, fornecendo informações e opiniões de conteúdos relacionados ao turismo. Por fim são apresentados os conceitos de extração de dados.

2.1 Plataforma Airbnb

O Airbnb constrói uma ponte de e-service entre os viajantes e os proprietários para satisfazer a demanda e a oferta neste mercado de dois lados. Ele permite que um anfitrião anuncie uma propriedade, como uma casa ou quarto, para aluguel de curto prazo, enquanto permite que o turista viva como um local. Desde o seu lançamento em 2008, esse tipo de acomodação compartilhada ponto a ponto tornou-se uma força muito disruptiva para o setor de hospitalidade tradicional. Nos Estados Unidos, o Airbnb teve um crescimento de demanda de 30% nos últimos anos, atingindo 5% de participação de mercado com aproximadamente 30% de penetração de mercado (ZHU; KUBICKOVA, 2022).

As tecnologias da Web 2.0 permitiram o modelo de negócios inovador do Airbnb, mas ser disruptivo, deve eventualmente haver demanda por um produto. A demanda por um serviço como o Airbnb não é um dado adquirido, pois o Airbnb é consideravelmente carente em muitas das áreas que são mais importantes para os turistas na escolha do alojamento hoteleiro, como a qualidade do serviço, simpatia da equipe, reputação da marca e segurança. Como foi discutido, no entanto, produtos disruptivos geralmente têm desempenho inferior com diz respeito aos atributos-chave dos produtos predominantes, mas produtos disruptivos também são frequentemente mais baratos e oferecem novos benefícios. Muito pelo contrário, a acomodação do Airbnb é normalmente mais barato do que a acomodação tradicional, e a acomodação do Airbnb introduz benefícios associados à permanência numa residência (GUTTENTAG, 2013).

Além dos preços econômicos, as acomodações do Airbnb também oferecem diversos benefícios advindos da permanência em uma residência. Por exemplo, alguns turistas podem preferir a sensação de estar em uma casa sobre um hotel, e os anfitriões do Airbnb podem fornecer conselhos locais úteis. Os hóspedes do Airbnb também costumam ter acesso a comodidades residenciais práticas, como cozinha completa, máquina de lavar e secadora. A experiência de morar em uma residência também oferece aos hóspedes a chance de ter uma experiência mais local, interagindo com o anfitrião ou vizinhos, e possivelmente

ficar em uma área 'não turística', já que as acomodações do Airbnb tendem a ser mais dispersas do que as acomodações tradicionais (GUTTENTAG, 2013).

A vantagem para os proprietários destes imóveis é que através da plataforma online podem chegar facilmente a um mercado global. Simultaneamente, usando o Airbnb, os visitantes têm acesso a uma gama cada vez maior de opções de acomodação durante a viagem (GUTTENTAG et al., 2017).

O Airbnb e empresas semelhantes enfrentam um ressentimento crescente dos moradores locais que temem que esses empreendimentos, juntamente com muitas outras atividades relacionadas ao turismo, transformem seus bairros residenciais outrora tranquilos em guetos de visitantes. Enquanto isso, os municípios lutam para identificar maneiras de regular o crescimento do Airbnb, seja por meio de tributação ou pela imposição de medidas drásticas destinadas a limitar ou erradicar totalmente os aluguéis de curto prazo. Por exemplo, a expansão fenomenal do Airbnb em Reykjavik levou o governo islandês a impor restrições à transformação de mais casas e quartos em aluguéis de curto prazo. Da mesma forma, Berlim e Amsterdã limitam por quanto tempo as propriedades podem ser alugadas pelo Airbnb (IOANNIDES; RöSLMAIER; ZEE, 2019).

2.2 Plataforma Tripadvisor

Quando falamos dos benefícios de viver na era da informação, há poucos exemplos mais emblemáticos desses benefícios do que o impacto da Internet no turismo e nas viagens. Com a enorme quantidade de informações disponíveis na Internet sobre destinos de viagem e opções de hospedagem, o planejamento de viagens pessoais tornou-se um grande passatempo. Esse aumento na disponibilidade de informações resulta em viajantes mais bem informados, o que, por sua vez, leva a um mercado mais eficiente (CUNNINGHAM et al., 2010).

O Tripadvisor, a maior plataforma de orientação de viagens do mundo, ajuda centenas de milhões de pessoas todos os meses a se tornarem melhores viajantes, desde o planejamento até a reserva e a realização de uma viagem. Viajantes de todo o mundo usam o site e o aplicativo do Tripadvisor para descobrir onde ficar, o que fazer e onde comer com base nas orientações de quem já esteve lá. Com mais de 1 bilhão de avaliações e opiniões de quase 8 milhões de empresas, os viajantes recorrem ao Tripadvisor para encontrar ofertas de acomodações, reservar experiências, reservar mesas em restaurantes deliciosos e descobrir ótimos lugares nas proximidades. Como uma empresa de orientação de viagens disponível em 43 mercados e 22 idiomas, o Tripadvisor facilita o planejamento, independentemente do tipo de viagem (TRIPADVISOR, 2022).

A plataforma oferece vários serviços direcionados a consumidores e empresas e adiciona continuamente novos serviços e recursos para atender às necessidades em evolução

de viajantes e fornecedores de turismo.Entre vários tópicos de conteúdo gerados pelos usuários, os conteúdos relacionados ao turismo são frequentemente os assuntos mais populares compartilhados e consumidos (YOO; SIGALA; GRETZEL, 2016).

É importante ressaltar que o uso das mídias sociais está cada vez mais integrado em todas as fases da experiência turística. No entanto, o TripAdvisor também é um infomediário especializado no campo de 'Big Data' e focado em vincular e atender as necessidades tanto da demanda quanto da oferta turística, fornecendo uma plataforma tecnológica na qual o conteúdo pode ser criado, analisado e distribuído para atender às necessidades de viajantes e empresas de turismo (YOO; SIGALA; GRETZEL, 2016).

2.3 Extração de dados

Crawling é o processo de explorar uma aplicação da web automaticamente. O web crawler visa descobrir na internet páginas de um aplicativo da Web navegando pelo aplicativo. Isso geralmente é feito simulando as possíveis interações do usuário. À medida que a quantidade de informações na web vem aumentando drasticamente, os usuários da web dependem cada vez mais dos mecanismos de busca para encontrar os dados desejados. Para que os motores de busca aprendam sobre os novos dados à medida que se tornam disponíveis na web, ele precisa rastrear e atualizar constantemente o mecanismo de pesquisa da base de dados (MIRTAHERI et al., 2014).

A definição tradicional de um web crawler assume que todos o conteúdo de um aplicativo da web é acessível por meio de URLs. Logo na história do rastreamento na web ficou claro que rastreadores da Web não podem lidar com as complexidades adicionadas por aplicativos da Web interativos que dependem da entrada do usuário para gerar páginas da Web. Esse cenário geralmente surge quando o aplicativo da Web é uma interface para um banco de dados e depende da entrada do usuário para recuperar o conteúdo do banco de dados. O novo campo de Deep Web-Crawling nasceu para resolver esse problema (MIRTAHERI et al., 2014).

Um rastreador da web é um dos principais componentes de motores de pesquisa na web. O crescimento do rastreador da web está aumentando na mesma forma como a web está crescendo. Uma lista de URLs está disponível com o rastreador da web e cada URL é chamado de semente. Cada URL é visitado pelo web crawler. Ele identifica os diferentes hiperlinks na página e os adiciona à lista de URLs a serem visitados. Esta lista é denominado como fronteira de rastreamento. Usando um conjunto de regras e políticas, URLs na fronteira são percorridos individualmente. Páginas diferentes da Internet são coletadas pelo analisador e o gerador é armazenado no sistema de banco de dados da pesquisa motor. Os URLs são então colocados na fila e depois é agendado, e pode ser acessado um a um por o motor de busca, um por um, sempre que necessário. As ligações e

arquivos relacionados que estão sendo pesquisados podem ser disponibilizados sempre que necessário em momento posterior de acordo com os requisitos. Com a ajuda de algoritmos adequados, os rastreadores da Web encontram o links relevantes para os motores de busca e usá-los ainda mais. Bancos de dados são máquinas muito grandes como o DB2, usadas para armazenar grandes quantidade de dados (AHUJA; SINGH; NICA, 2014).

Existem vários usos de rastreadores da web: Os rastreadores também podem ser usados para automatizar tarefas de manutenção em um site, como verificar links ou validar código HTML. Os rastreadores podem ser usados para coletar tipos específicos de informações de páginas da Web, como coletar endereços de e-mail (geralmente para spam). Os mecanismos de pesquisa costumam usar rastreadores da Web para coletar informações que estão disponíveis em páginas da Web públicas. Eles coletam dados para que, quando os internautas inserirem um termo de pesquisa em seus site, eles podem fornecer rapidamente ao surfista sites relevantes. Os linguistas usam rastreadores da web para realizar uma análise textual. Eles percorrem a Internet para determinar quais palavras são comumente usadas hoje (AHUJA; SINGH; NICA, 2014).

3 Trabalhos Relacionados

3.1 Trabalhos de conclusão de curso

A seguir, são descritos alguns trabalhos de conclusão de curso que foram criados ou estendidos para integrar as pesquisas na plataforma Airbnb.

3.1.1 O uso de big data para análise de oferta de imóveis via Airbnb em destinos turísticos

Aplicações como o Airbnb começaram a surgir com o intuito de facilitar a busca por hospedagens. Considerado como uma inovação disruptiva, o aplicativo é simples e intuitivo, e busca diminuir a distância entre hóspede e anfitrião, bem como trazer uma experiência de acomodação transformada, diferente das oferecidas por métodos tradicionais. O estudo busca implementar uma aplicação prática que atenda aos interesses de órgãos responsáveis pelo turismo na região de Florianópolis e aos usuários de plataformas de aluguel por temporada. Propondo um ambiente de análise dos dados sobre oferta de imóveis em Florianópolis disponibilizados no Airbnb aplicando técnicas de extração de conhecimento (WEIGEL, 2019).

Para o levantamento dessas informações, os dados são coletados por mineração de conteúdo na web por meio de um Web Crawler desenvolvido em python, com integrações de bibliotecas como Requests e BeautifulSoup. Os dados coletados através do script, utilizaram apenas o site do Airbnb como fonte e os dados ficam armazenados em um banco de dados Postgresql. Ao final da coleta de dados, os arquivos JSON com os dados sobre disponibilidade do imóvel foram transformados em arquivos CSV. Finalizado o processo de conversão, os dados foram limpos e organizados, tratando valores nulos e normalizando atributos como tipo do quarto e disponibilidade (WEIGEL, 2019).

Concluindo, o resultado do estudo se deu através dos mapas e indicadores gerados no desenvolvimento notando que a procura pelos imóveis disponibilizados via Airbnb na Ilha de Florianópolis, fora da alta temporada, tende a crescer durante os finais de semana e feriados, bem como o número de hóspedes. As habitações com maior ocupação tendem a possuir um menor valor de diária e hospedar menos pessoas, diferente das habitações com maior arrecadação que tendem a ser lugares de alto nível e estarem em regiões de alto padrão da ilha, bem como hospedar um maior número de pessoas. Os indicadores para habitação, número de hóspedes e ocupação propiciaram novos conhecimentos sobre a oferta (WEIGEL, 2019).

3.1.2 Análise de dados Python para obter insights dos hosts do Airbnb

Cada vez mais, os viajantes estão usando o Airbnb em vez de se hospedar em hotéis tradicionais. No entanto, em um mercado de Airbnb tão crescente e competitivo, muitos anfitriões podem achar difícil transformar suas hospedagens atraentes entre tantas outras listadas na plataforma. Ao usar Python para analisar todos os dados e todos os aspectos das listagens do Airbnb, o autor propõe testar e encontrar correlações entre certas variáveis e listagens populares (TIAN, 2021).

Para a leitura dos arquivos CSV utilizou-se Pandas e os códigos de séries, DataFrame, mesclagem e DateTime para ajudar a entender e visualizar melhor os resultados. Outro método em Python que o autor usou é o NLTK, também chamado de Natural Language Toolkit. Antes de entrar profundamente no conjunto de dados, o autor fez uma análise estatística básica para calcular o número de listagens exclusivas, hosts e perímetro básico (média, mediana e desvio padrão). Para a limpeza o autor utilizou a função drop_duplicates() para retirar os duplicados.

A partir disso foi feita uma divisão de duas categorias de anfitriões, sendo elas super anfitriões e apenas anfitriões para compreender melhor a diferença entre cada grupo. Tornar-se um super anfitrião pode trazer muitos benefícios diretos e indiretos. De acordo com as políticas do Airbnb, um anfitrião precisa manter 90% ou mais de taxa de resposta, 1% ou menos de taxa de cancelamento e 4,8% de pontuações de avaliações totais. Para encontrar os padrões específicos do superhost, o autor examinou primeiro o tempo médio de resposta de super e não super anfitriões. Existem quatro tipos de tempo de resposta neste conjunto de dados que estão "dentro de uma hora", "dentro de algumas horas", "dentro de um dia ” e “ dentro de alguns dias” (TIAN, 2021).

Depois de realizar as análises foi possível perceber que os super anfitriões são melhores em todos os tipos de pontuações, mas possuem grandes vantagens em três tipos de pontuação, que são precisão, limpeza e valor. Como pontos de atenção para os anfitriões que não são super melhorarem estão a redução do tempo de resposta com os possíveis convidados, e a obtenção das identidades verificadas por parte dos anfitriões (TIAN, 2021).

3.1.3 Modelo de predição dos preços de acomodações em Nova York

Este artigo tem como objetivo abordar a previsão de preços do Airbnb com diferentes aprendizados de máquina em três cidades da cidade de Nova York, encontrando a variável dependente e as variáveis independentes para calcular as correlações entre elas e prever o modelo.No entendimento dos dados, o artigo faz uso do dataset disponibilizado pelo Kaggle onde são listadas 44317 propriedades em Nova York, 59881 em Paris e 22552 em Berlim (LUO XUANYU ZHOU, 2019).

Para a análise de dados, o cluster é criado e o conjunto de dados adicionado de forma

tabular. A limpeza de dados é realizada, já que existem muitas informações desnecessárias para o modelo, alguns grupos de detalhes são removidos, os valores duplicados e ausentes também.

Na etapa de exploração e visualização dos dados, foram avaliados os valores dos recursos e as interações entre os recursos visualizados. A análise de dados fornece os detalhes em gráficos e tabelas que são facilmente entendidas. Depois de visualizar e encontrar as correlações, algoritmos de aprendizado de máquina são usados para encontrar as previsões.

Após a preparação, análise e finalização da correção para o cálculo da predição, o algoritmo para prever é escolhido. O modelo escolhido para o experimento foi um modelo de regressão, sendo necessário apenas uma precisão razoável. Além disso, com as condições analisadas, os modelos Gradient Boosted Regression e XGBoost também serviram.

Construir o modelo de previsão de mercado para o Airbnb com o melhor desempenho e com base em uma série de recursos, incluindo especificações de propriedade, entrada de hosts e consumidores nas listagens é a parte central da análise. Na comparação dos resultados entre a análise de regressão e Gradient Boost Regression foi possível perceber que a segunda opção teve a pontuação mais alta, portanto foi melhor em prever os resultados das listagens de preços (LUO XUANYU ZHOU, 2019).

3.2 Trabalhos divulgados em blogs

A seguir, são descritos dois trabalhos publicados em blogs que foram criados para analisar os dados e o mercado da plataforma Airbnb.

3.2.1 Desenvolvimento de estratégia de preços: Análise de mercado do Airbnb com Python

A aquisição de dados no exemplo utilizado vem da construção de um web scraper usando a linguagem Python, acessando os dados brutos e curados do Airbnb sob demanda e separando os resultados por faixas de preço. Após a coleta é realizado a limpeza dos dados e nesse ponto o foco foi limpar as variáveis de preço e quartos. Para o preço a estratégia é remover cifras e alterar os valores para inteiro, já no caso da variável quartos os valores strings foram alterados para números inteiros, exemplo "Estúdio"passa a ser "0"(BLAKE, 2021).

Conhecer os percentis de taxa do mercado para a categoria é um ponto importante, com isso a chance de encontrar o melhor valor cresce, sem acabar cobrando caro demais e ficando fora do mercado, cada produto possui sua particularidade e compreensão sobre o seu valor. No estudo em caso utilizou-se o "np.percentile"para encontrar 25 (Q1), 50 (Q2 / mediana) e 75 (Q3), em sequência foi criado um gráfico semelhante a um histograma,

para visualizar a distribuição dos dados. Com o gráfico é possível realizar uma análise de sentimento e descobrir mais sobre as faixas de preço coletando informações como onde estão localizadas, quantos banheiros possuem, entre outros, criando assim um perfil para cada faixa de preço. Entretanto nem sempre os quartis podem ser os melhores modelos de precificação, alguns mercados podem se encaixar melhor em quintis ou assim por diante (BLAKE, 2021).

É importante se concentrar em ofertas alternativas dentro da categoria da propriedade, em vez de apenas acomodações de tamanho semelhante. Observar características diferenciais, já que nem todas as acomodações são criadas e disponibilizadas iguais. Cada perfil de unidade possui seu próprio segmento de clientes e seu próprio nível de sensibilidade ao preço (BLAKE, 2021).

3.2.2 Análise dos dados do Airbnb - Dublin

O dataset analisado foi obtido através do site Inside Airbnb, sendo uma versão resumida com um conjunto de dados de 7894 imóveis e 16 variáveis, dos tipos float, inteiros e objetos. Foram obtidas variáveis como Título do anúncio da propriedade, Nome do bairro, Tipo de acomodação oferecida, Valor do Aluguel, Número de avaliações entre outras (JOÃO, 2021).

Para a limpeza de dados, fez-se necessário a remoção de uma variável de grupo de vizinhos onde os valores eram nulos, e também houve a detecção de outliers nas variáveis de valor do aluguel e no mínimo de noites que foram removidos de modo que não prejudicasse a análise.

Uma das perguntas para a análise é saber a média do mínimo de noites para os aluguéis. De acordo com a extração do estudo, a cidade de Dublin possui uma média de no mínimo 2 noites, indicando que os anfitriões costumam disponibilizar os imóveis fazendo com que os hóspedes passem ao menos o fim de semana hospedados. Outra pergunta interessante constatada no estudo é qual a localidade mais cara de Dublin? no caso em questão a resposta foi Dublin City e Dun Laoghaire-Rathdown sendo a média €88,37 e €85,99 euros, respectivamente. A resposta para os valores elevados se encontra no caso de Dublin possuir uma facilidade de acesso ao centro e aos famosos 'pubs' e destinos turísticos, já Dun Laoghaire-Rathdown é um lugar tranquilo e sossegado e próximo de Dublin (JOÃO, 2021).

A categoria de imóveis mais alugados na cidade de Dublin segundo a análise é de 79 por cento para os quartos privados e Casa/Apartamentos inteiros e apenas 1 por cento são quartos compartilhados. Com essas informações é possível concluir que viajar em família ou em casal, são opções boas considerando que será mais fácil encontrar hospedagens na região. A média de gastos de hospedagem para a cidade é de cerca de €85,22 por noite.

Para o estudo, no caso da cidade de Dublin foi possível inferir e responder algumas perguntas, concluindo que o custo de viajar e se hospedar lá não é realidade para a maioria dos brasileiros (JOÃO, 2021).

3.3 Comparativo entre trabalhos

Com o crescimento da utilização da plataforma Airbnb no decorrer dos anos, vieram a surgir muitos estudos abordando o tema, diversas análises com os mais distintos resultados foram apresentados. A seguir será disponibilizado uma tabela com o comparativo entre os trabalhos utilizados.

Trabalhos	**Foco do trabalho**	**Resultados apresentados**	**Região analisada**	**Técnicas de análise de dados**
Trabalho 3.1.1	Análise de oferta de imóveis via Airbnb em Florianópolis	Indicadores	Florianópolis/SC /Brasil	Mapas de calor
Trabalho 3.1.2	Análise de dados Python para obter insights dos hosts do Airbnb	Insights	Los Angeles/CA /USA	NLTK, Análise de Sentimento, Correlação
Trabalho 3.1.3	Modelo de predição dos preços de acomodações em Nova York.	Previsão	Nova York/NY /USA	Regressão Linear, Regressão Gradient Boost, XGBoost
Trabalho 3.2.1	Análise de mercado do Airbnb com Python.	Percentis	Deauville/Normandia/França	Análise de sentimento, Gráfico do KDE
Trabalho 3.2.2	Análise dos dados do Airbnb - Dublin.	Insights	Dublin/Irlanda	Mapa de calor
Trabalho proposto	Análise e correlação de dados: um estudo de caso usando o AirBnB e o TripAdvisor em Florianópolis	Indicadores	Florianópolis/SC /Brasil	Correlação, Mapa de calor, Análise de sentimento

Tabela 1 – Comparativo dos trabalhos relacionados

Fonte: Elaborado pelo autor

De acordo com os estudos analisados, o autor identificou que não existe nenhum trabalho que aborde e faça uma análise de correlação entre os atributos do Airbnb com os dados da cidade de Florianópolis. Sendo assim, se faz necessário identificar quais as correlações entre os valores de cobrança dos alugueis empregados na cidade com os possíveis atributos atrelados a decisão de estabelecer esse valor.

4 Desenvolvimento

Construir um Web Scraper, antes de tudo, requer o conhecimento de como exatamente os usuários navegam em um site e o que acontece durante esse processo do ponto de vista do processamento de informações.(NEMESLAKI; POCSAROVSZKY, 2011)

Neste capítulo, é proposto um modelo de extração de dados e análise utilizado para coletar os dados disponíveis na plataforma do Airbnb e na plataforma Tripadvisor. Nas subseções seguintes serão descritos a visão geral, os passos da utilização do Web Scraper e do modelo utilizado para o carregamento e a transformação dos dados.

4.1 Visão Geral

Esta seção tem como objetivo apresentar as aplicações desenvolvidas neste trabalho, mostrando seus tipos diferentes de componentes, as interações que ocorrem e os resultados de saída.

Como pode ser visto na Figura 1, o Scaper deve varrer as páginas do Airbnb que contemplem a cidade alvo da busca, no caso Florianópolis. Ao encontra-la deve salvar os dados de cada hospedagem e suas características em um arquivo.

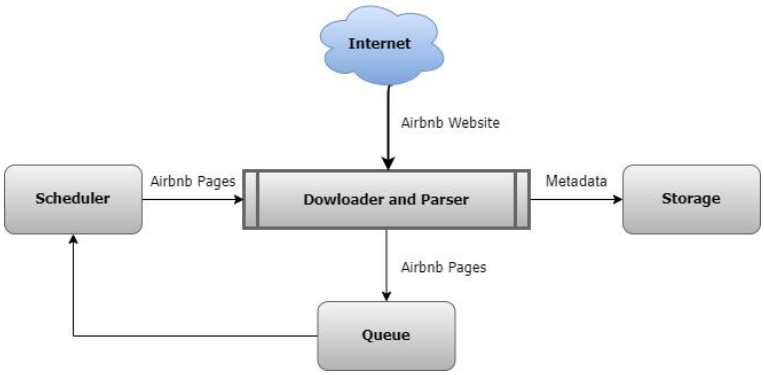


Figura 1 – Arquitetura de um Web Scaper para o Airbnb

Fonte: Adaptado de Ahuja, Singh e nica (2014)

Assim como o Scraper do Airbnb, o Scraper para o Tripadvisor segue o mesmo padrão de fluxo como pode ser visto na Figura 2, o Scraper também deve varrer as páginas do Tripadvisor que contemplem a cidade de Florianópolis. Ao encontra-la deve salvar os dados em um arquivo.

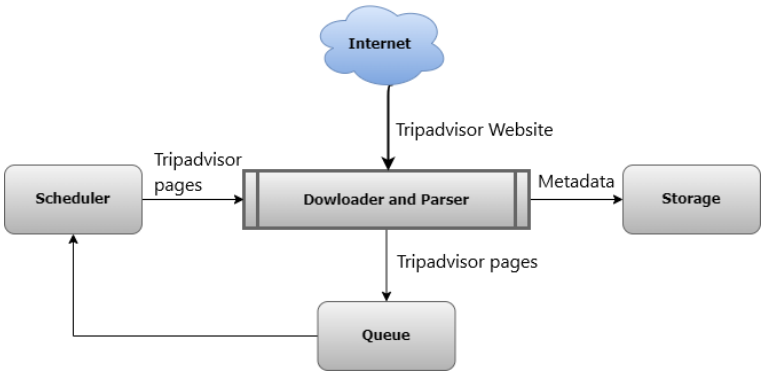


Figura 2 – Arquitetura de um Web Scraper para o Tripadvisor

Fonte: Adaptado de Ahuja, Singh e nica (2014)

Nas subseções seguintes são descritos, a localização escolhida, a organização dos dados, o desenvolvimento e implementação do Web Scraper e a carga de dados.

4.2 Localização dos dados

A cidade de Florianópolis, capital do estado de Santa Catarina na região sul do Brasil, conta com uma população estimada de 516 mil pessoas. A cidade tem seu desenvolvimento econômico baseado no turismo, inovação tecnológica e no setor de serviços. Pode-se notar que Florianópolis é considerada uma das mais importantes cidades inteligentes no Brasil pelo estudo de Sistemas Urbanos (2017) e está posicionada entre as cidades brasileiras mais empreendedoras Santos-Júnior et al. (2019).

A escolha da cidade para o estudo vêm da necessidade de encontrar mais informações sobre as hospedagens hoje disponíveis na plataforma Airbnb. Considerando os valores empregados e as características que determinam esses valores. Além de analisar outros fatores da cidade, como as atrações e restaurantes disponíveis na plataforma do Tripadvisor. Auxiliando os anfitriões das residências de Florianópolis a terem uma noção melhor sobre as avaliações das hospedagens, os valores empregados e os fatores que podem se relacionar a isso.

4.3 Coleta dos dados

A obtenção dos dados da cidade de Florianópolis, se da através de dois Web Scrapers, disponibilizados na plataforma Apify. Para o acesso a essa plataforma é necessário criar uma conta que possui algumas características limitadas, entretando as extrações para o estudo de caso estão dentro do valor disponibilizado gratuitamente. Os dados coletados foram primeiramente armazenados em arquivos de formato CSV.

O Web Scraper relacionado ao Airbnb, está disponibilizado no GitHub por (DUONG, 2021). O Airbnb Scraper é projetado para extrair a maioria dos dados do Airbnb disponíveis publicamente para os anúncios de acomodações. É possível obter todos os dados básicos sobre o anúncio, as avaliações, comentários, preços, detalhes do anfitrião e também do hóspede.

Para o estudo com o Airbnb como fonte de dados. Os atributos a serem coletados são:

- url: url da acomodação identificando o acesso a página;
- nome_acomodacao: nome da acomodação identificando um dos itens listados no site Airbnb;
- tipo_acomodacao: identificador do tipo de acomodação, exemplo: casa, quarto inteiro, quarto compartilhado, etc;
- quantidade_hospedes: a quantidade de hóspedes máxima permitida;
- preco: preço diário de aluguel da acomodação;
- superhost: identificador de um super anfitrião;
- localizacao: endereço contendo o bairro e a cidade da acomodação;
- latitude: latitude relacionada a acomodação;
- longitude: longitude relacionada a acomodação;
- avaliacao: notas de avaliação das acomodações;
- comentario: os comentários realizados pelos hóspedes em relação a acomodação;
- lingua_comentario: a lingua utilizada nos comentários realizados pelos hóspedes em relação a acomodação;

Em relação ao Web Scraper destinado ao Tripadvisor, está disponibilizado no GitHub por (COPELLI, 2022). O Tripadvisor Scraper permite obter dados do Tripadvisor. Ele é adequado para casos de uso onde é necessário coletar avaliações, e-mails, endereços, prêmios e muitos outros atributos de hotéis e restaurantes e atrações das cidades.

No caso do Tripadvisor como fonte de dados alguns atributos disponíveis na extração foram eliminados, os dados relacionados a hotéis por exemplo não são coletados. Os atributos a serem coletados são:

- web_url: url do restaurante ou atração identificando o acesso a página;
- nome: nome do restaurante ou atração identificando um dos itens listados no site Tripadvisor;
- tipo: identificador para definir o tipo: restaurante ou atração;

- premiacao: identificador se o restaurante já foi premiado com o Certificado de Excelência entregue pelo Tripadvisor;
- posicao_ranking: posição que o restaurante esta no ranking de classificação do Tripadvisor ;
- faixa_de_preco: faixa de preço do restaurante, classificada entre $,$$, $$$,$$$$;
- tipo_culinaria: tipo de culinária utilizado no restaurante, exemplo: brasileira, japonesa, etc;
- endereco: endereço contendo a rua, bairro e cidade do restaurante ou atração;
- latitude: latitude relacionada ao restaurante ou atração;
- longitude: longitude relacionada ao restaurante ou atração;
- avaliacao: notas de avaliação do restaurante ou atração;
- comentario: os comentários realizados pelos clientes ou visitantes em relação aos restaurantes ou atrações;
- lingua_comentario: a lingua utilizada nos comentários realizados pelos visitantes em relação aos restaurantes e atrações;

4.4 Limpeza dos dados

A qualidade dos dados é um dos problemas mais importantes no gerenciamento de dados, pois dados sujos geralmente levam a resultados imprecisos de análise de dados e decisões de negócios incorretas.(ILYAS; CHU, 2019) A limpeza de dados inclui todas as metodologias cujo objetivo é "detectar e remover erros e inconsistências dos dados para melhorar a qualidade dos dados. (CALABRESE, 2018)

Após a coleta de dados, os dados foram limpos e tratados. Para a limpeza foi necessário a imputação de valores não preenchidos com a palavra nulo, para padronizar os valores nulos. Houve tratamento nos atributos de preço, sendo necessário a retirada do cifrão, no atributo endereço foi necessário eliminar localidades que não se enquadravam na região de Florianópolis. Além disso foi preciso eliminar as duplicidades.

4.5 Scraper

Inicialmente é necessário percorrer as páginas do site Airbnb como um usuário normal, ele insere o destino, as datas desejadas e clica no botão de pesquisa. O mecanismo de classificação do Airbnb gera uma listagem de diversas acomodações com algumas breves descrições. Ao acessar as listagens é possível obter descrições com maiores detalhes das

acomodações. Sendo assim, o Scraper deve extrair as informações de dois tipos de páginas, as de pesquisa e de detalhes.

Ao extrair as listagens da página de pesquisa foram criadas algumas funções, chamadas, findListings, getListingsSection, addListings, a biblioteca nos permite navegar pela árvore HTML a acessar os elementos, obtendo assim o o texto referente a listagem.

Após a definição para os atributos, entramos no processo de acessar todas as páginas da cidade de Florianópolis. Cada página possui 20 anúncios, e conforme realizada a combinação dos parâmetros de pesquisa o Airbnb fornece acesso a até 300 anúncios por localidade. Com isso foi criado a função findListings na qual tem o objetivo de varrer todas as páginas disponíveis, apenas inserindo o link inicial da pesquisa.

Para extrair os dados do Tripadvisor o método foi semelhante, percorrendo as páginas do site como um usuário, após a inserção da localidade uma lista é gerada com os restaurantes e as atrações disponíveis na região e suas descrições. O Scraper nesse caso deve extrair as informações de dois tipos de páginas, as de pesquisa e de detalhes novamente.

Entretanto as funções utilizadas para o Tripadvisor foram iniciadas através da função buildSearchRequestsFromLocationName, que recebe como parâmetro de entrada a localização desejada. A partir dela é realizado uma requisição para a obtenção da lista com os restaurantes e atrações através da função getRequestListSources.

4.6 Modelo banco de dados

Com o objetivo de salvar os dados em um banco de dados, foi necessário a criação de um modelo dimensional, separando algumas dimensões. A figura a seguir apresenta o esquema do banco de dados para o Airbnb no lado esquerdo e no lado direito para o Tripadvisor:

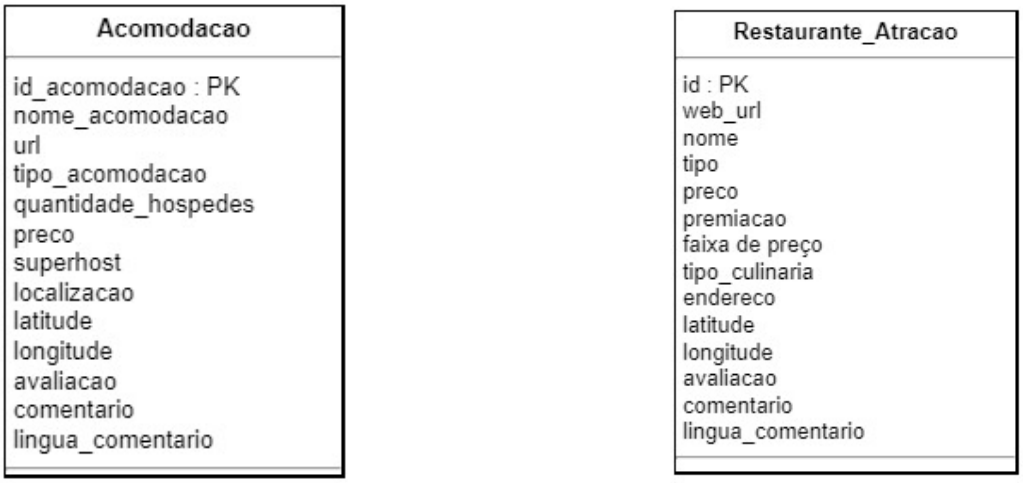


Figura 3 – Modelo Dimensional para o Airbnb e Tripadvisor

Fonte: Elaborado pelo autor (2022)

4.7 Carga dos dados

Para o processo de carga dos dados e ETL, foi utilizado a ferramenta de Data Integration, Pentaho. Com ela foi possível realizar a etapa de limpeza dos dados. Ao fim do processo os dados foram carregados em um banco de dados PostgreSQL de código aberto.

5 Análise dos dados

5.1 Dados do Airbnb

De acordo com os dados coletados, é possível identificar alguns fatores dentro das acomodações encontradas para a região de Florianópolis. A amostra extraída são de 10.241 ofertas disponíveis que equivale a 100% dos dados.

5.1.1 Tipos de propriedades

Diante da região em destaque, foram encontrados alguns tipos diferentes de propriedades disponíveis, são estas: Apartamentos inteiros, Casas inteiras, Quartos inteiros, Quartos compartilhados e Outros que contemplam acomodações peculiares como barcos, contêineres e camper/motorhome.

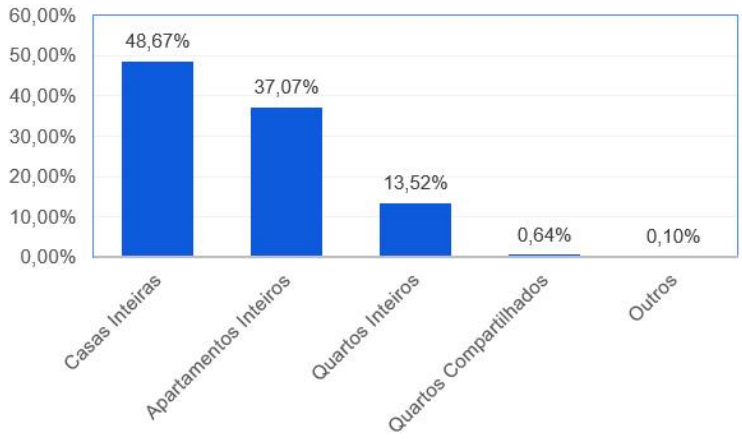


Figura 4 – Tipos de propriedades

Fonte: Elaborado pelo autor (2022)

Em relação aos resultados, foi possível observar como as casas e apartamentos oferecidas por inteiro correspondem a um total de aproximadamente 85,74% dos tipos de propriedades para a região de Florianópolis, um valor muito elevado comparando com os outros tipos disponíveis.

5.1.2 Avaliação

Cada anúncio disponível no site possui um espaço designado a avaliações e comentários realizados pelas pessoas que ali se hospedaram. O anfitrião por sua vez tem um retorno sobre seu serviço, além de que, para os demais consumidores as avaliações

publicadas podem ser um fator decisivo para a escolha do local. É gerado um resultado que pode variar entre 1 a 5 estrelas, sendo 0 a nota mais baixa e 5 a mais alta. Para o estudo será utilizado um arredondamento das notas.

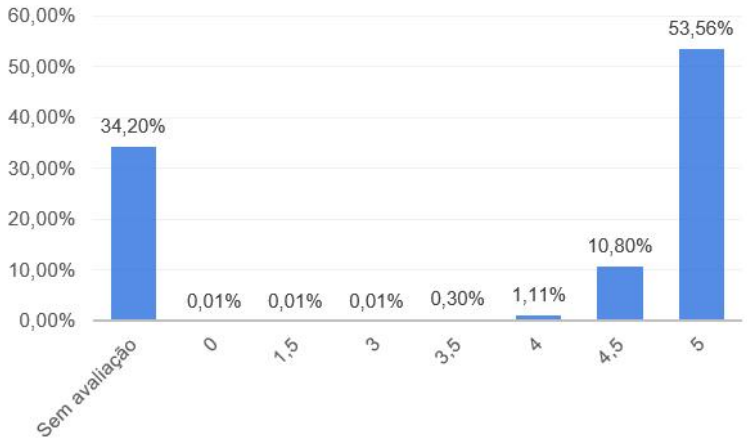


Figura 5 – Classificação Airbnb

Fonte: Elaborado pelo autor (2022)

Conforme observa-se a maior parte das avaliações coletadas estão entre 4,5 e 5, 10,80% para a nota 4,5 e 53,56% para a nota 5, totalizando um percentual de 64,36% nessa faixa de avaliação. Por outro lado 34,20% das acomodações não possuem notas de avaliação, nesse caso podemos considerar que muitos desses anúncios podem ser novos, nunca receberam hóspedes ou simplesmente os hospedes não fizeram suas avaliações. Já no caso das notas altas podemos inferir que quando um hóspede se dedica a fazer uma avaliação, a grande maioria avalia com notas boas as acomodações da região de Florianópolis.

5.1.3 Comentários

Além das notas das avaliações, existem os comentários que podem ser realizados pelos hóspedes. Esses comentários podem servir de base para outros hóspedes tomaram suas decisões em se hospedarem ou não, comunicando se o anúncio segue os padrões da publicação ou se ocorreu alguma situação diferente do esperado.

Uma análise com uma amostra de comentários foi realizada, nesse sentido foram coletados 4.429 comentários em português e 3.497 em inglês de cada acomodação presente no estudo e com isso criado uma nuvem de palavras mais utilizadas nesses comentários. Para o experimento foi necessário a utilização da Ferramenta Microsoft Power BI, com o visual Word Cloud.

Figura 6 – Comentários na língua portuguesa

Fonte: Elaborado pelo autor (2022)

Com esse visual, as palavras mais usadas acabam se tornando maiores. Na imagem para os comentários na língua portuguesa, é possível notar o uso das palavras 'casa', mencionada 1.192 vezes, e 'apartamento', mencionado 704 vezes. Outras palavras que receberam bastante destaque foram 'localização', mencionada 669 vezes e 'praia' mencionada 658 vezes.

Analisando esses valores e as palavras encontradas, observa-se que é comum o uso da descrição do tipo de imóvel, casa ou apartamento, nos comentários. Outro fator é que para os comentários das acomodações em Florianópolis considerou-se muito o fator localização, e relacionado a isso a proximidade com as praias. A palavra, 'limpa', encontrada 402 vezes também traz um significado para a análise, sendo limpeza um outro ponto importante de atenção na hora de avaliar uma acomodação.

Os comentários quando realizados, tendem a ser positivos, como no caso do uso das palavras 'excelente', mencionado 691 vezes, 'ótima', mencionada 499 vezes e 'recomendo', mencionado 467 vezes, além de outras palavras que podem ser encontradas como: boa, maravilhosa, impecável, incrível, etc.

Florianópolis é conhecida por ser uma cidade muito turística, recebendo muitos visitantes de outros países, nesse sentido também houveram avaliações em outras línguas e no caso o filtro utilizado foi a língua inglesa, já que houve uma boa amostragem disponível.

Figura 7 – Comentários na língua inglesa

Fonte: Elaborado pelo autor (2022)

A nuvem de palavras para essa língua mantém alguns padrões parecidos. a palavra 'place' que pode ser traduzida como 'lugar' aparece 1.286 vezes sendo a mais utilizada. A segunda mais usada foi 'stay', traduzindo é interpretada como 'estadia', mencionada 979 vezes. A localização ou 'location' apareceu 810 vezes, e nesse caso pode ser considerado que o significado da utilização de localização tende a ser relacionado ao posicionamento da acomodação na geografia local, diferente da palavra lugar que pode estar mencionando a acomodação em si.

Palavras como apartamento e casa que foram as mais utilizadas em português, continuam sendo muito utilizadas na língua inglesa, 'apartament' ou apartamento, foi usado 719 vezes, e 'house' ou casa foi utilizado 613 vezes. Além da palavra 'beach' que significa praia, utilizada 806.

Entretanto a palavra 'host' que significa anfitrião foi muito mais usada na língua inglesa, 808 vezes, já na língua portuguesa foi utilizado 248 vezes. Nesse caso observa-se que para turistas da língua inglesa o fator anfitrião é muito mencionado nos comentários, sendo esse um ponto de atenção já que o tratamento com esses hóspedes pode ser diferenciado.

Existe uma relação muito grande entre as palavras usadas nos dois diferentes idiomas, sobretudo nos fatores indicados como tipo de acomodação, na localização e no quesito praia, além das palavras positivas representadas de diversas formas. Mas em um contexto geral os comentários para as acomodações em Florianópolis, tendem a ser voltadas positivamente para a localização geográfica e fatores como praia.

5.1.4 Preço

O preço das acomodações é um dos fatores a ser analisado pelos hóspedes. Uma acomodação pode conter várias comodidades, entretanto, o hóspede precisa ter condições de pagar para se hospedar ali.

O Airbnb disponibiliza uma faixa de preço referente à região pesquisada pelos clientes, com a finalidade de facilitar a pesquisa. A faixa de preço é composta pelas seguintes informações: o preço mais baixo e o preço mais alto dos anúncios listados em uma determinada região, além de disponibilizar a informação do preço médio da região. O hóspede então pode filtrar conforme o valor que ele está disposto a pagar por uma diária, portanto, só serão apresentado os anúncios com o preço de acordo com o especificado.

Para a análise de Florianópolis foi observado uma grande diferença com relação ao preço da diária das acomodações. A média de preço encontrada na região foi de R$536,78 reais, sendo a menor diária de R$30,00 reais e a mais alta R$46.971,00 reais por dia.

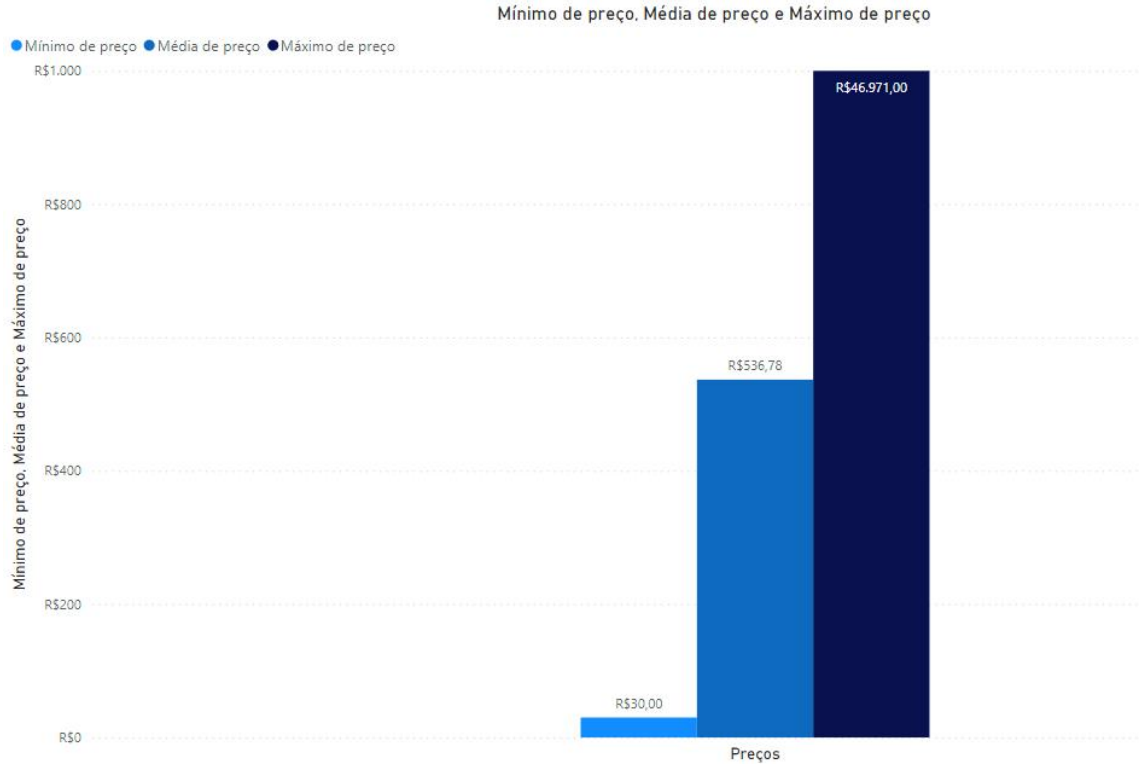


Figura 8 – Mínimo,média e máximo de preços

Fonte: Elaborado pelo autor (2022)

Conforme segmentamos os diferentes tipos de propriedades citados anteriormente (Apartamentos inteiros, Casas inteiras, Quartos inteiros, Quartos compartilhados e Outros) percebemos as diferenças nas médias dos valores diários. As casas inteiras como imaginado possuem o maior valor médio diário, sendo R$818,59 reais. Já os apartamentos inteiros possuem uma média de preço de R$425,07 reais. Os quartos inteiros segue um valor médio

de R$376,12. Para os quartos compartilhados a média segue em R$121,60. E no quesito outros tipos de acomodações o valor médio é de R$3.411,78, entretanto nesse caso foram relacionados apenas 9 propriedades, sendo que duas delas, dois barcos, possuem a diária de R$17.500,00 e R$9.500,00, sem essas acomodações a média cai para R$506,00.

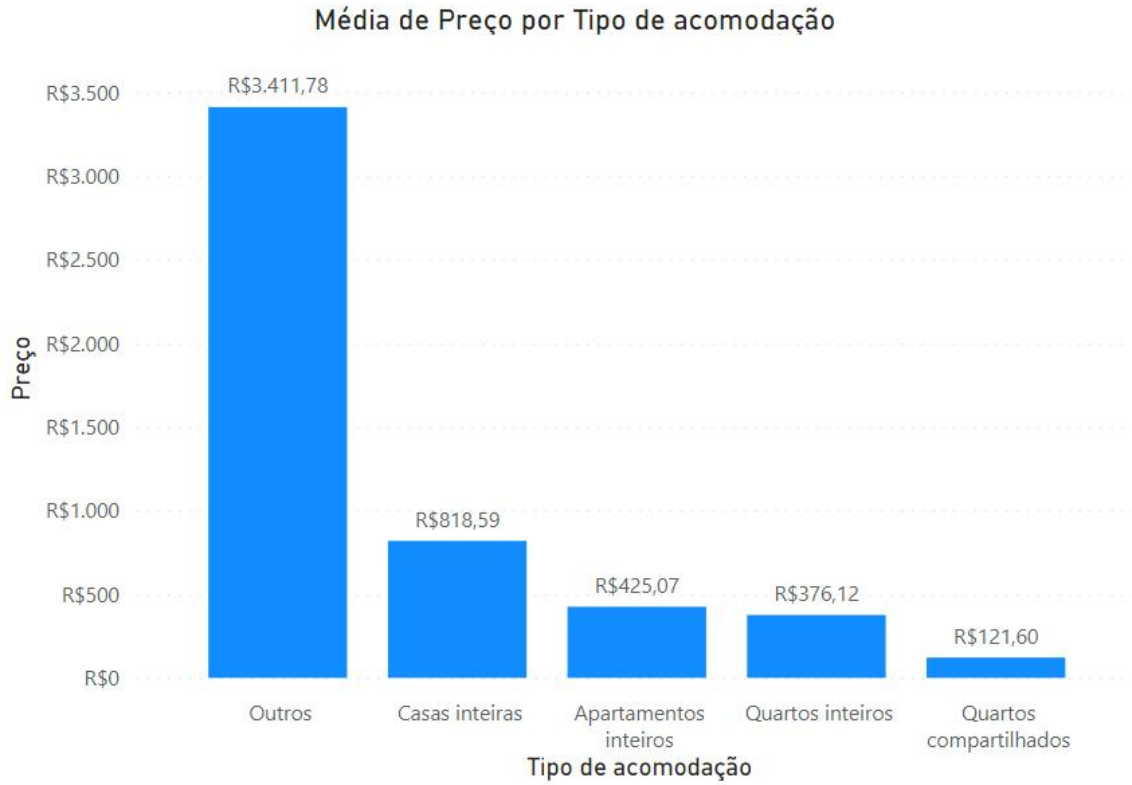


Figura 9 – Média de Preços por acomodação

Fonte: Elaborado pelo autor (2022)

A partir dos resultados obtidos foi possível verificar que o preço da diária pode ser realmente o fator decisivo para o consumidor. Observando uma variação entre os diferentes tipos de acomodações.

5.1.5 Localização

Em relação a localização das acomodações listadas no Airbnb, existem algumas regiões dentro de Florianópolis que podem ser classificadas de acordo com o espaço geográfico, como bairros residenciais, praias e a parte central. Além da divisão Norte, Sul, Centro e Leste da ilha.

A região central da ilha possui a função de estabelecer o comércio mais intenso, concentrando muitas empresas e também sendo referência em arquitetura histórica, concentra o Mercado Público, um dos lugares mais importantes para o crescimento do comércio na cidade antigamente. Os bairros residenciais são as regiões onde existe boa parte da

moradia dos habitantes. Já as praias são regiões onde normalmente concentram-se a parte voltada ao turismo.

Com o mapa de calor a seguir desenvolvido na ferramenta da Microsoft Power BI, com o visual HeatMap, é possível observar a distribuição das acomodações dentro da ilha de Florianópolis:

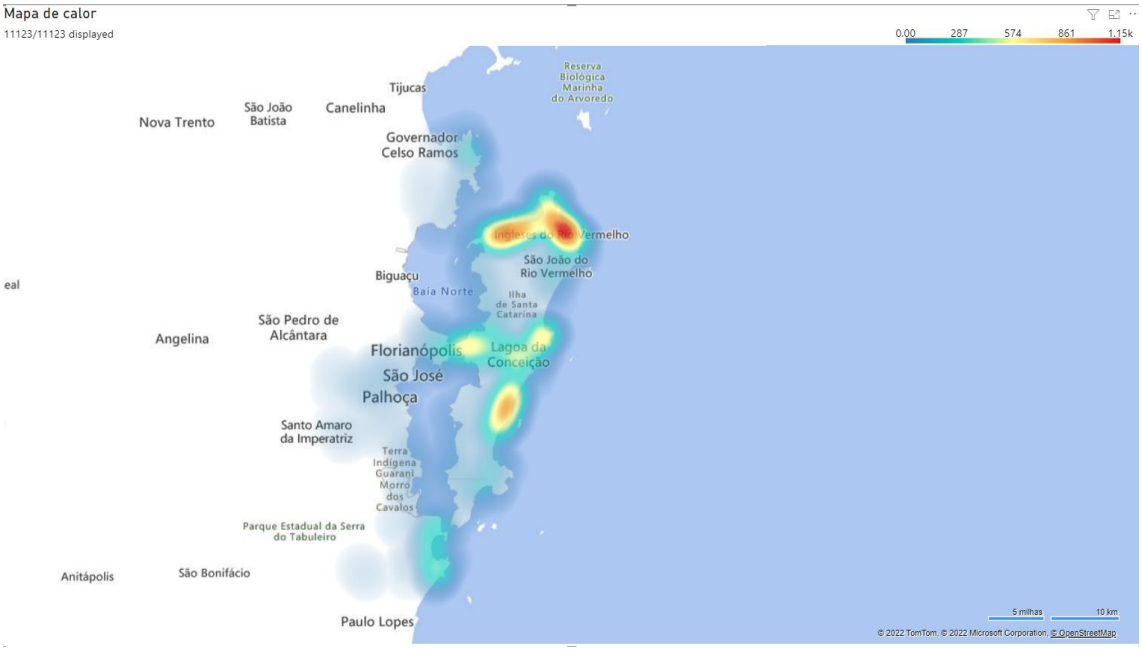


Figura 10 – Mapa de calor Airbnb - Florianópolis

Fonte: Elaborado pelo autor (2022)

No mapa é possível observar a quantidade superior de acomodações disponíveis no norte da ilha, principalmente na região dos Ingleses, onde nos últimos anos vem crescendo muito. Contudo Canasvieiras e Jurerê também apresentam um índice muito alto de acomodações disponíveis. Outra parte da cidade que merece atenção é a parte leste, Lagoa e Campeche se destacam nos bairros com maiores volumes de acomodações disponíveis. Por fim o centro da cidade também ocupa uma boa parte do volume de acomodações disponíveis.

5.2 Dados do Tripadvisor

De acordo com os dados coletados do Tripadvisor, é possível identificar alguns fatores dentro dos estabelecimentos encontrados para a região de Florianópolis. A amostra extraída são de 5.180 restaurantes e 300 atrações que equivalem a 100% dos dados. As atrações podem ser interpretadas como : atividades de turismo, praias, cinema, shoppings, mercados e parques. E os restaurantes englobam também cafés e bares.

5.2.1 Avaliação

Cada restaurante ou atração disponível no site do TripAdvisor possui um espaço designado a avaliações e comentários realizados pelas pessoas que já passaram por ali. O dono do restaurante por sua vez tem um retorno sobre seu serviço, além de que, para os demais consumidores as avaliações publicadas podem ser um fator decisivo para a escolha do local. E a própria prefeitura pode analisar em relação as atrações das cidades modelos de melhorias com base nas avaliações. É gerado um resultado que pode variar entre 1 a 5 estrelas, sendo 1 a nota mais baixa e 5 a mais alta. Para o estudo será utilizado um arredondamento das notas.

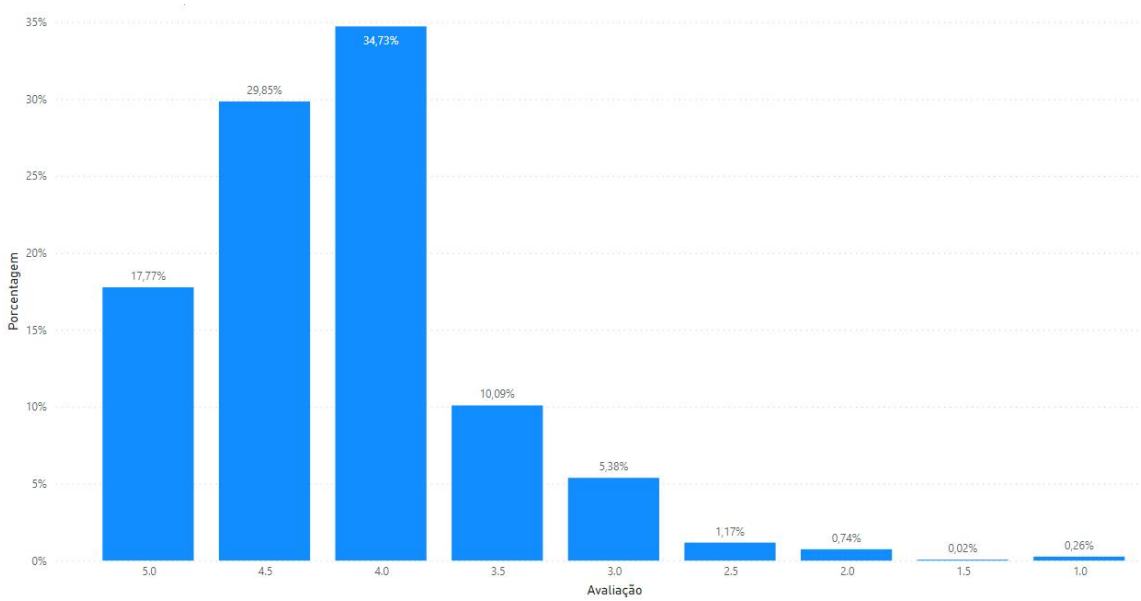


Figura 11 – Classificação TripAdvisor

Fonte: Elaborado pelo autor (2022)

Conforme observa-se a maior parte das avaliações coletadas estão entre 4 e 5, sendo 34,73% para a nota 4, 29,85% para a nota 4,5 e 17,77% para a nota 5, totalizando um percentual de 82,35% nessa faixa de avaliação. Com esses números podemos considerar que grande parte dos restaurantes e atrações possuem uma avaliação média dentro dos padrões acima da nota 4, sendo um valor bem alto. De acordo com os usuários a classificação dos restaurantes de modo geral em Florianópolis é muito boa, um índice alto de gastronomia e boas atrações como diversas praias.

5.2.2 Comentários

Os comentários podem ser realizados pelos usuários da plataforma sobre as atrações e os estabelecimentos disponíveis na cidade. Outros usuários podem se basear nos

comentários para frequentar ou não esse lugar, deixando claro como foi sua experiência e se recomenda para outras pessoas.

Uma análise com uma amostra de comentários foi realizada, nesse sentido foram coletados 3.132 comentários em português e 1.033 comentários em inglês para os restaurantes e com isso foi criado uma nuvem das palavras mais utilizadas.

Figura 12 – Comentários na língua portuguesa no TripAdvisor

Fonte: Elaborado pelo autor (2022)

Com esse visual, as palavras mais usadas acabam se tornando maiores. Na imagem para os comentários na língua portuguesa, é possível notar o uso da palavra 'Atendimento' em destaque sendo mencionada 819 vezes, além da palavra 'bom', mencionado 713 vezes. Com esse conjunto de palavras podemos sugerir que um bom atendimento faz bastante parte dos comentários realizados. Além da menção a palavra 'comida' mencionada 609 vezes.

Analisando esses valores e as palavras encontradas, observa-se que é comum o uso das palavras relacionadas a localização, nos comentários. 'Lugar' aparece 391 vezes e 'Local' por sua vez 344 vezes.

Outro fator a ser considerado é o 'Ambiente', mencionado 384 vezes, que acaba tendo um peso grande no conjunto de palavras utilizadas. Sendo assim muitas pessoas que frequentam restaurantes optam por um ambiente agradável.

Os comentários quando realizados, tendem a ser positivos, como no caso do uso das palavras 'Bom', 'Boa', 'Recomendo','Excelente', além de outras palavras que podem ser encontradas.

Outro filtro foi utilizado para determinar palavras da língua inglesa.

Figura 13 – Comentários na língua inglesa no TripAdvisor

Fonte: Elaborado pelo autor (2022)

A nuvem de palavras para essa língua mantém alguns padrões parecidos. a palavra 'place' que pode ser traduzida como 'lugar' aparece 308 vezes sendo a terceira mais utilizada. A mais usada foi 'good', traduzindo é interpretada como 'bom', mencionada 405 vezes. A palavra 'food' apareceu 387 vezes, e nesse caso pode ser considerado que o significado da utilização de 'comida' atrelado a outra palavras positivas tendem a trazer uma faixa grande de comentários positivos em relações aos restaurantes da cidade de Florianópolis.

Nesse estudo existiu uma relação grande entre as palavras usadas nos dois diferentes idiomas, sobretudo nos fatores indicados como lugar, comida e palavras positivas de diversas formas.

5.2.3 Preço

O segmento preço, segue sendo um dos fatores a ser analisado pelos clientes. Um restaurante pode conter várias propostas, alguns optam por cardápios mais em conta e outros são mais refinados. Em Florianópolis existem muitas variedades, tanto de tipos de culinárias como de propostas de espaços arquitetônicos. No conjunto de dados analisado, foi observado uma grande quantidade de restaurantes com a proposta de culinárias brasileira, japonesa, italiana, observa-se também boa quantidade de restaurantes de frutos do mar, pizzas e fast-foods, além de cafés e bares. Foi observado uma grande diferença com relação aos diversos tipos de culinárias.

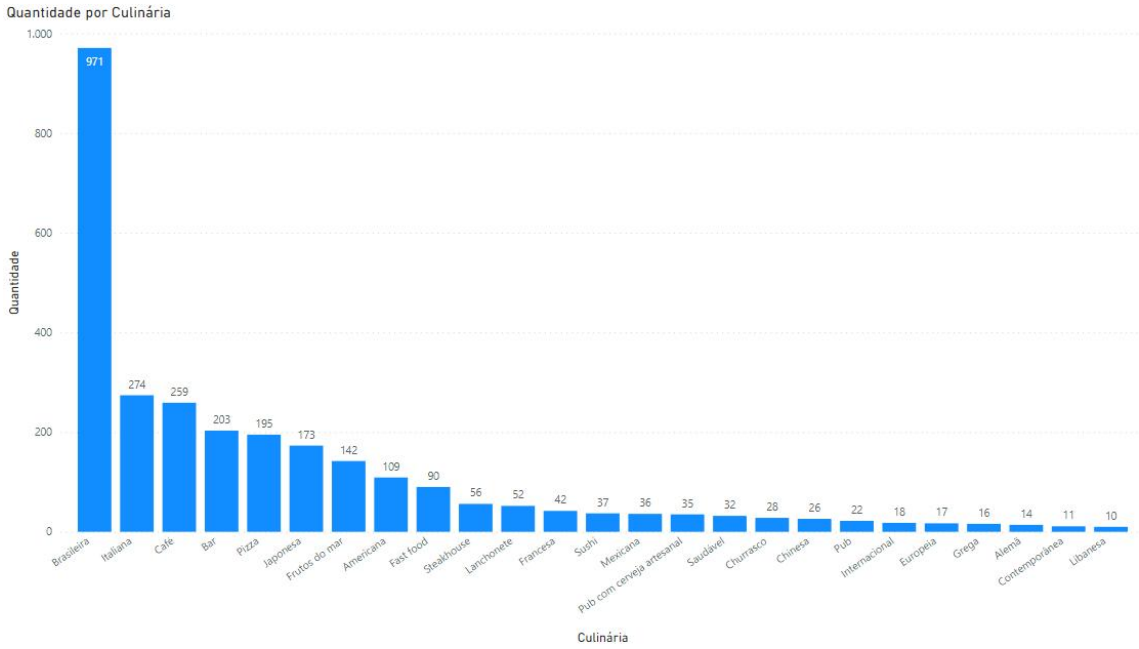


Figura 14 – Culinária quantidade no TripAdvisor

Fonte: Elaborado pelo autor (2022)

O TripAdvisor disponibiliza uma faixa de preço diferente para os usuários, utilizando o cifrão como indicador. Não existem definições exatas de valor para cada faixa de preços, mas o senso comum dos usuários da plataforma poderia classifica-las como sendo:

$: comida com preço baixo;

$$: comida com preços médios;

$$$: comida com preços mais altos;

$$$$: comida com preços bem altos, normalmente cozinhas internacionais com chefes renomados.

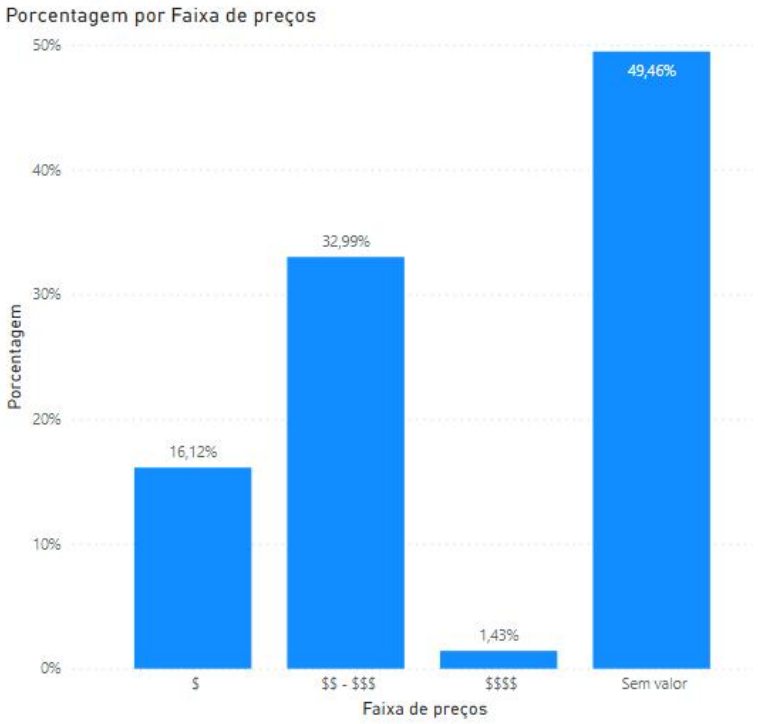


Figura 15 – Faixa de preços no TripAdvisor

Fonte: Elaborado pelo autor (2022)

Conforme visualizamos as diferentes faixas de valores citados anteriormente, percebemos as diferenças nas médias dos valores. É notável a grande quantidade de restaurantes que não possuem o registro de faixas de valores, sendo responsável por uma porcentagem de 49,46%. Grande parte dos restaurantes da ilha de Florianópolis, 32,99% se enquadram na faixa de preços entre valores médios e valores mais altos. É possível observar que 16,12% dos valores são relacionados a um preço baixo, portanto dentro da ilha segundo a análise é mais fácil encontrar restaurantes que tenham o preço mais elevado. Por fim temos a faixa com os valores mais altos em 1,43%, onde se enquadram restaurantes com cozinhas internacionais e chefes renomados, que acaba sendo uma porcentagem bem baixa para uma cidade turística de alto padrão.

A partir dos resultados obtidos foi possível verificar que mesmo com o alto índice de restaurantes sem informações de faixa de preços, uma boa porcentagem está dentro do padrão de preços médios e mais elevados. Portanto podemos considerar que para realizar refeições fora de casa na ilha de Florianópolis, as opções de restaurantes com preços mais elevados serão maiores.

5.2.4 Localização

Em relação a localização das atrações e restaurantes listados no TripAdvisor, conforme mencionado anteriormente a região central da ilha possui a função de estabelecer o

comércio mais intenso, portanto concentra uma grande variedade de restaurantes utilizados no dia-a-dia da população,além do Mercado Público e outras atrações históricas. Algumas atrações e restaurantes estão concentrados nas praias, região voltada ao turismo.

Com o mapa de calor a seguir desenvolvido na ferramenta da Microsoft Power BI, com o visual HeatMap, é possível observar a distribuição das atrações dentro da ilha de Florianópolis:

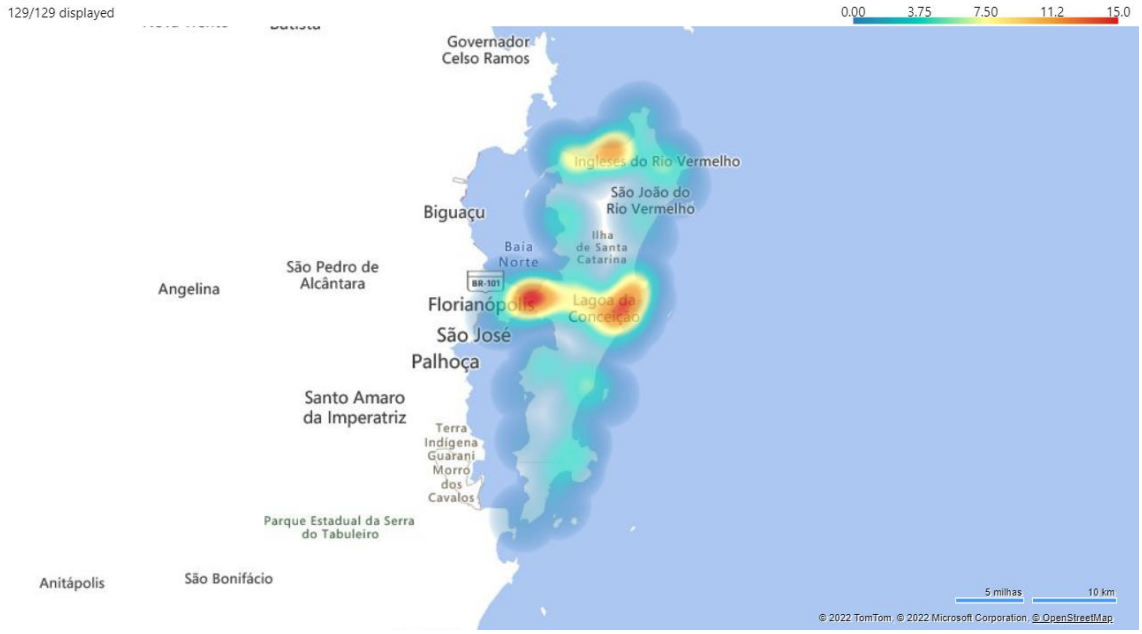


Figura 16 – Mapa de calor das atrações no TripAdvisor

Fonte: Elaborado pelo autor (2022)

É possível observar a quantidade elevada de atrações localizadas no centro da cidade, que devido a sua região histórica e cercada de mercados e centros comerciais possui muitas atrações. A região de Canasvieiras também se destaca, muito pela quantidade de atividades relacionadas ao turismo, já que recebe boa parte de visitantes do nosso país vizinho Argentina. Contudo Lagoa da conceição e parte Leste da ilha também apresentam um índice muito alto de atrações, porém mais distribuídas, a região possui diversas praias e muita prática de esportes ao ar livre, como o surfe, que é destaque e recebe alguns eventos na Praia da Joaquina e na Praia Mole.

Considerando o mapa de calor no mesmo formato, porém agora com o filtro de restaurantes, obtivemos a imagem a seguir:

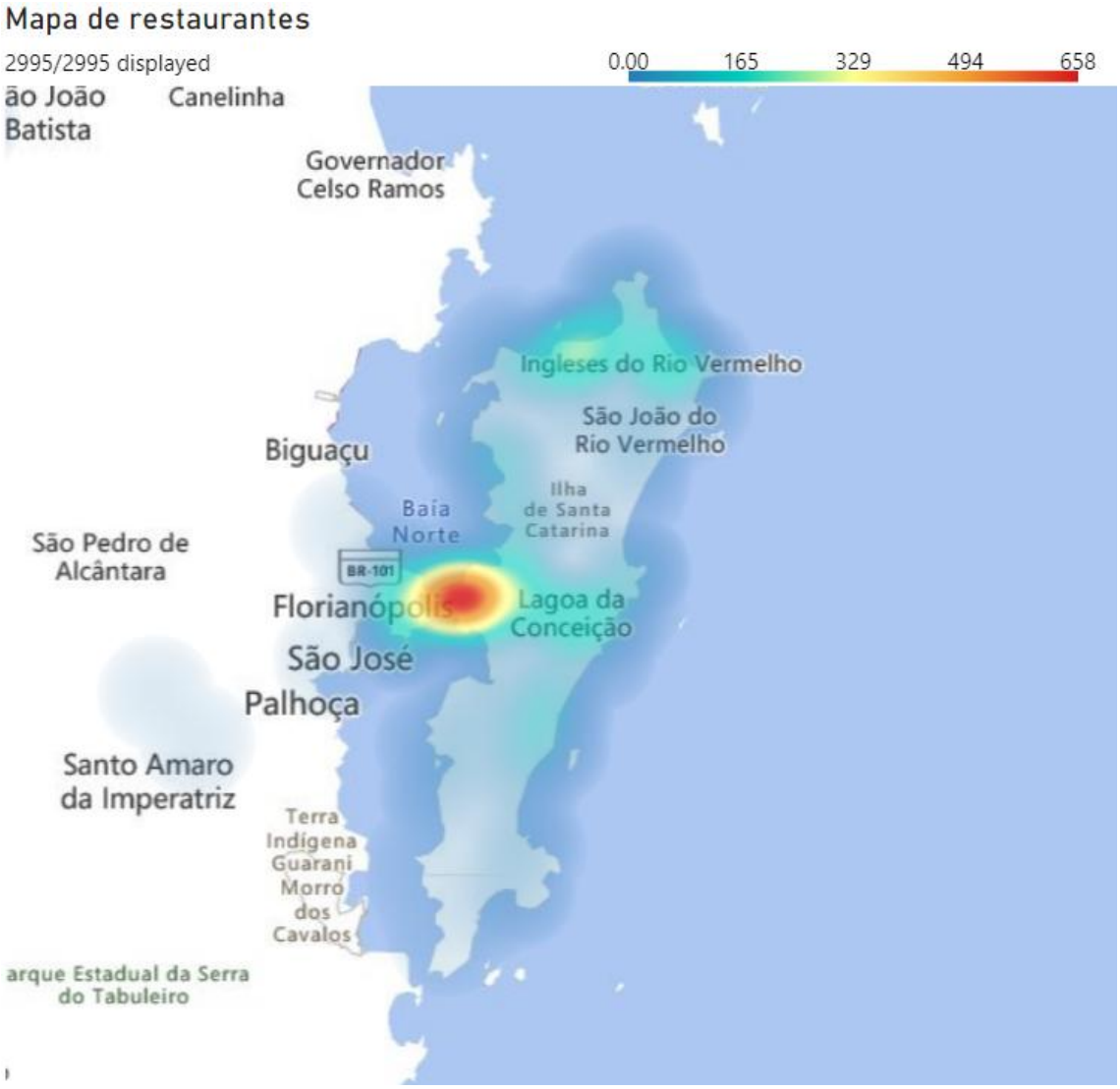


Figura 17 – Mapa de calor dos restaurantes no TripAdvisor

Fonte: Elaborado pelo autor (2022)

Nele nota-se que o centro é o grande foco de restaurantes, bares e cafés disponíveis na ilha. Outras regiões como o Norte da ilha e a parte Leste, também apresentam algumas manchas de calor. Para entendermos melhor a distribuição central foi realizada outra imagem mais aproximada:

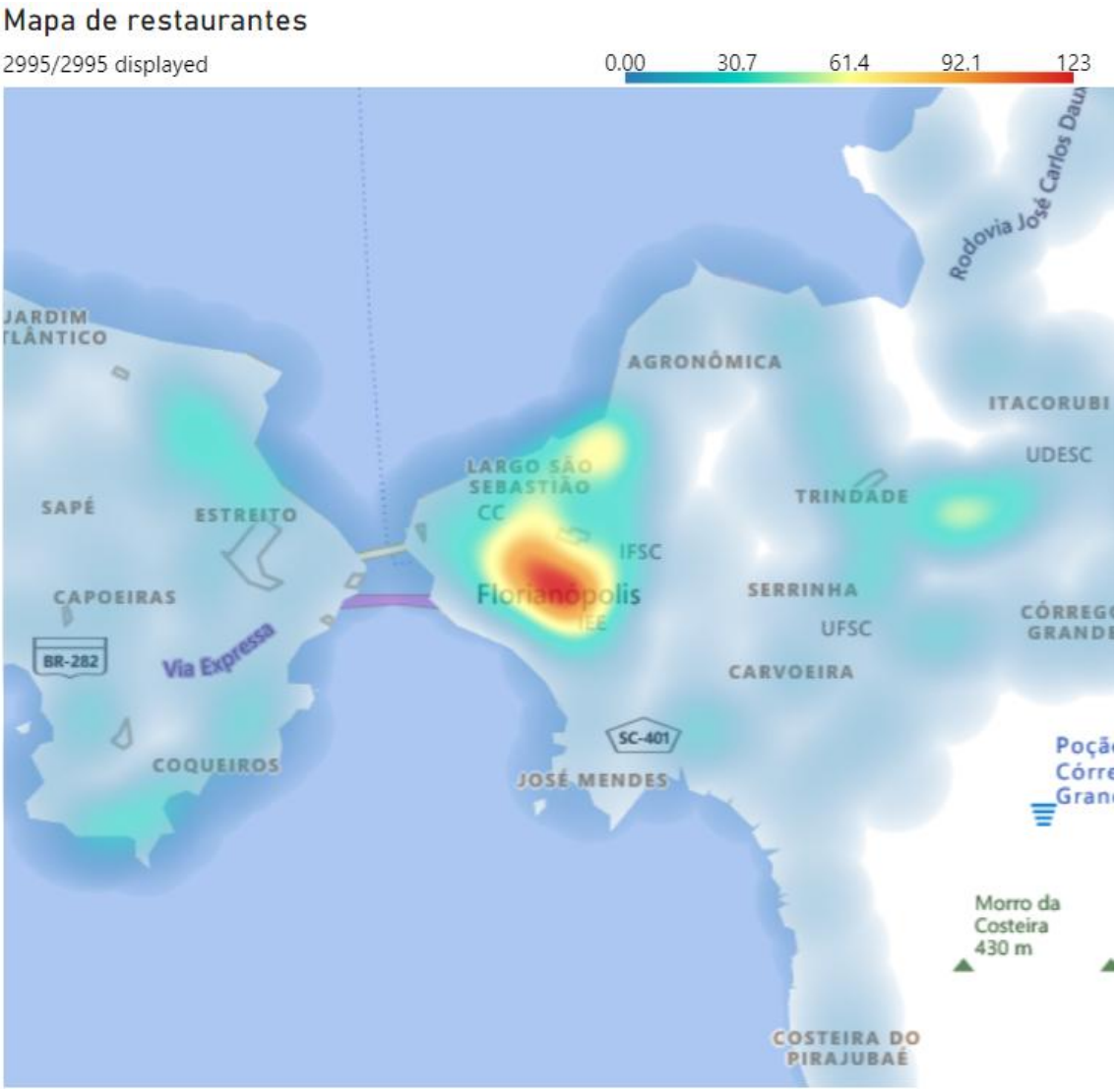


Figura 18 – Mapa de calor dos restaurantes da região central no TripAdvisor

Fonte: Elaborado pelo autor (2022)

Conforme aproximamos, é perceptível que a parte continental também influência na grande relação de restaurantes, assim como a parte da Trindade, arredores da UFSC e UDESC. Regiões onde são famosas por abrigar os mais diversos tipos de culinárias. Coqueiros por ser uma via voltada a gastronomia e Trindade e Santa Mônica por abrigarem muitos estudantes universitários. Já o Centro e a Beira-mar abrigram restaurantes do dia a dia, restaurantes refinados, bares e cafés de diversos gostos.

5.3 Correlação entre os dados do Airbnb e TripAdvisor

Ao falarmos em análise de dados, é necessário entender qual é a associação entre as variáveis. A análise de correlação é uma forma descritiva que mede se há e qual o grau de dependência entre variáveis, ou seja, o quanto uma variável interfere em outra, sendo

que essa relação de dependência pode ou não ser causal. Essa medida de grau de relação é medida através dos coeficientes. O coeficiente de correlação pode variar em termos de valor de -1 a +1, quanto maior for o valor absoluto do coeficiente, mais forte será a relação entre as variáveis. Segundo (COHEN, 1992) os tamanhos de efeito podem se enquadrar em

r = |0,10| -> correlação fraca.

r = |0,30| -> correlação moderada.

r = |0,50| -> correlação forte.

Para a realização da correlação entre as variáveis foi necessário importar as bibliotecas pandas e seaborn. O pacote pandas é uma ferramenta na manipulação de dados, tabelas e *dataframes*. Já o pacote seaborn é excelente para a criação de visualizações gráficas, principalmente em casos de mapeamentos estatísticos.

As duas bases de dados foram transformadas em *dataframes*, através da função read_csv() da biblioteca pandas. Após a transformação foi realizada a função corr() também da biblioteca pandas, onde ela realiza a correlação entre os atributos do dataset. A função possui uma configuração para o uso de diversos métodos de correlação, os métodos utilizados foram Pearson, Kendall, Spearman. O ambiente para utilização do código foi o Google Colaboratory.

Primeiro os dados serão analisados individualmente, começando com os referentes ao Airbnb:

```
df.corr(method='pearson')
```

	latitude	longitude	quantidade_hospedes	preco	avaliacao
latitude	1.000000	0.552864	0.161014	0.067139	-0.051059
longitude	0.552864	1.000000	0.107088	-0.019374	-0.038399
quantidade_hospedes	0.161014	0.107088	1.000000	0.273736	-0.045563
preco	0.067139	-0.019374	0.273736	1.000000	-0.154777
avaliacao	-0.051059	-0.038399	-0.045563	-0.154777	1.000000

Figura 19 – Correlação dos dados do Airbnb utilizando o método de Pearson

Fonte: Elaborado pelo autor (2022)

```
df.corr(method='kendall')
```

	latitude	longitude	quantidade_hospedes	preco	avaliacao
latitude	1.000000	0.386624	0.145968	0.098678	-0.046289
longitude	0.386624	1.000000	0.108903	0.041411	-0.032477
quantidade_hospedes	0.145968	0.108903	1.000000	0.467015	-0.033627
preco	0.098678	0.041411	0.467015	1.000000	-0.153627
avaliacao	-0.046289	-0.032477	-0.033627	-0.153627	1.000000

Figura 20 – Correlação dos dados do Airbnb utilizando o método de Kendall

Fonte: Elaborado pelo autor (2022)

```
df.corr(method='spearman')
```

	latitude	longitude	quantidade_hospedes	preco	avaliacao
latitude	1.000000	0.558787	0.203314	0.146872	-0.058814
longitude	0.558787	1.000000	0.151413	0.062783	-0.041344
quantidade_hospedes	0.203314	0.151413	1.000000	0.604938	-0.040082
preco	0.146872	0.062783	0.604938	1.000000	-0.196255
avaliacao	-0.058814	-0.041344	-0.040082	-0.196255	1.000000

Figura 21 – Correlação dos dados do Airbnb utilizando o método de Spearman

Fonte: Elaborado pelo autor (2022)

Utilizando a biblioteca do seaborn é possível visualizar melhor a força de cada propriedade:

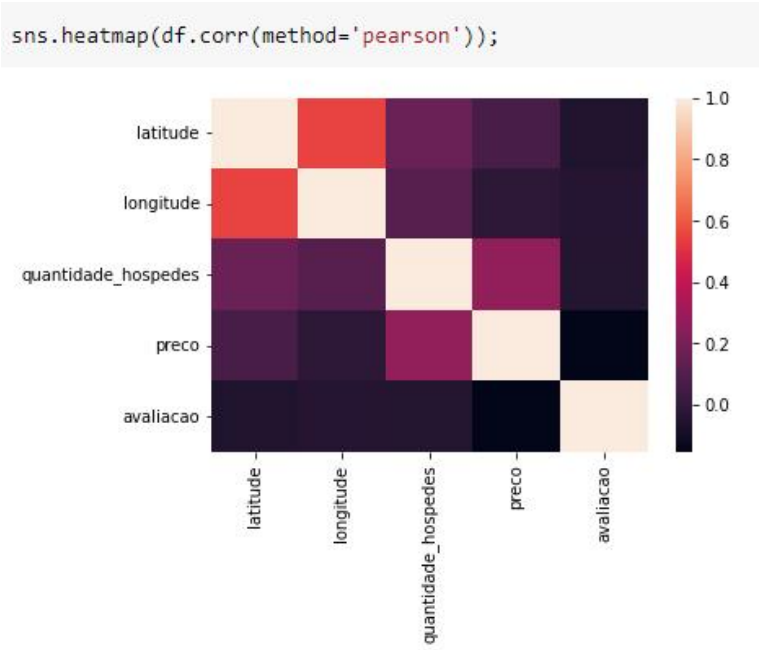


Figura 22 – Correlação dos dados do Airbnb com Heatmap utilizando o método de Pearson

Fonte: Elaborado pelo autor (2022)

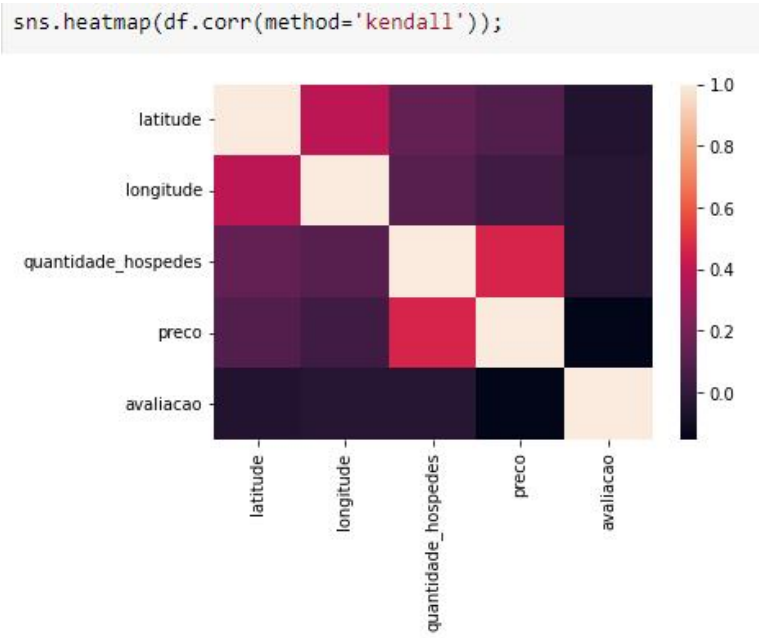


Figura 23 – Correlação dos dados do Airbnb com Heatmap utilizando o método de Kendall

Fonte: Elaborado pelo autor (2022)

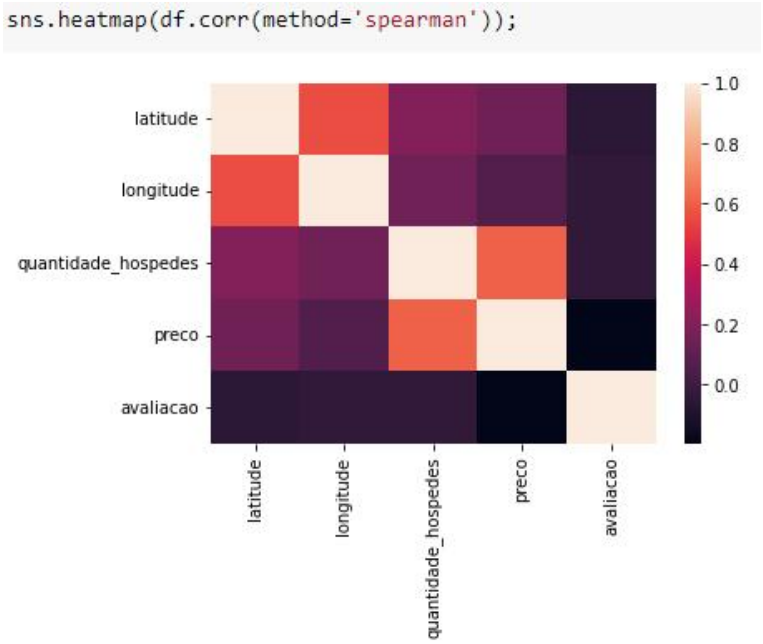


Figura 24 – Correlação dos dados do Airbnb com Heatmap utilizando o método de Spearman

Fonte: Elaborado pelo autor (2022)

A partir da análise dos dados do Airbnb, podemos considerar que com a utilização da metodologia de Pearson, os coeficientes com correlação positiva forte identificados foram latitude e longitude em 0,55. Para o caso de preço e quantidade de hóspedes houve uma correlação positiva moderada em 0,27. Como relação negativa fraca temos o caso de preço e avaliação onde o valor ficou em -0,15.

Para a metodologia de Kendall, latitude e longitude seguem com correlações positivas entretanto de forma mais moderada em 0,38. Já quantidade de hóspedes e preço passam a ter uma correlação moderada com um grau mais forte em 0,46.

Spearman e sua metodologia nos mostra que semelhante a Pearson a correlação segue positiva e forte para latitude e longitude com grau de 0,55. É interessante notar que o grau de correlação positiva entre a quantidade de hóspedes e o preço sobe para 0,60.

Seguindo agora com o conjunto de dados do TripAdvisor, foi realizada outra análise individual para entendimento das relações das variáveis entre si. Os métodos e abordagens foram as mesmas das etapas anteriores:

```
df2.corr(method='pearson')
```

	premiacao	id	latitude	longitude	quantidade_avaliacao	faixa_de_preco	posicao_ranking	avaliacao
premiacao	1.000000	0.234391	0.069901	0.120729	0.482320	0.016482	-0.688184	0.411677
id	0.234391	1.000000	0.058600	-0.032976	-0.121667	-0.021233	-0.022462	0.148679
latitude	0.069901	0.058600	1.000000	-0.684765	-0.006549	-0.009538	0.026543	-0.017588
longitude	0.120729	-0.032976	-0.684765	1.000000	0.005912	0.012358	-0.027072	0.024111
quantidade_avaliacao	0.482320	-0.121667	-0.006549	0.005912	1.000000	0.328763	-0.425784	0.124834
faixa_de_preco	0.016482	-0.021233	-0.009538	0.012358	0.328763	1.000000	-0.647595	0.012639
posicao_ranking	-0.688184	-0.022462	0.026543	-0.027072	-0.425784	-0.647595	1.000000	-0.424899
avaliacao	0.411677	0.148679	-0.017588	0.024111	0.124834	0.012639	-0.424899	1.000000

Figura 25 – Correlação dos dados do TripAdvisor utilizando o método de Pearson

Fonte: Elaborado pelo autor (2022)

```
df2.corr(method='kendall')
```

	premiacao	id	latitude	longitude	quantidade_avaliacao	faixa_de_preco	posicao_ranking	avaliacao
premiacao	1.000000	0.104896	0.095965	0.084020	0.510513	0.060200	-0.632367	0.367343
id	0.104896	1.000000	0.005779	0.033898	-0.137372	-0.015518	0.003224	0.123909
latitude	0.095965	0.005779	1.000000	0.299581	0.035149	0.033240	-0.024728	-0.015459
longitude	0.084020	0.033898	0.299581	1.000000	0.036429	0.026104	-0.045714	0.035812
quantidade_avaliacao	0.510513	-0.137372	0.035149	0.036429	1.000000	0.604219	-0.710952	0.001129
faixa_de_preco	0.060200	-0.015518	0.033240	0.026104	0.604219	1.000000	-0.513703	-0.011910
posicao_ranking	-0.632367	0.003224	-0.024728	-0.045714	-0.710952	-0.513703	1.000000	-0.305994
avaliacao	0.367343	0.123909	-0.015459	0.035812	0.001129	-0.011910	-0.305994	1.000000

Figura 26 – Correlação dos dados do TripAdvisor utilizando o método de Kendall

Fonte: Elaborado pelo autor (2022)

```
df2.corr(method='spearman')
```

	premiacao	id	latitude	longitude	quantidade_avaliacao	faixa_de_preco	posicao_ranking	avaliacao
premiacao	1.000000	0.132504	0.131993	0.113697	0.659026	0.073333	-0.791467	0.422994
id	0.132504	1.000000	0.009982	0.052887	-0.193899	-0.024593	0.008330	0.167747
latitude	0.131993	0.009982	1.000000	0.442533	0.050293	0.041663	-0.037609	-0.021192
longitude	0.113697	0.052887	0.442533	1.000000	0.052657	0.033303	-0.069035	0.048432
quantidade_avaliacao	0.659026	-0.193899	0.050293	0.052657	1.000000	0.730855	-0.852872	-0.001213
faixa_de_preco	0.073333	-0.024593	0.041663	0.033303	0.730855	1.000000	-0.640934	-0.014605
posicao_ranking	-0.791467	0.008330	-0.037609	-0.069035	-0.852872	-0.640934	1.000000	-0.396221
avaliacao	0.422994	0.167747	-0.021192	0.048432	-0.001213	-0.014605	-0.396221	1.000000

Figura 27 – Correlação dos dados do TripAdvisor utilizando o método de Spearman

Fonte: Elaborado pelo autor (2022)

Após o uso da biblioteca pandas, novamente foi usado a biblioteca seaborn para trazer uma melhor visualização:

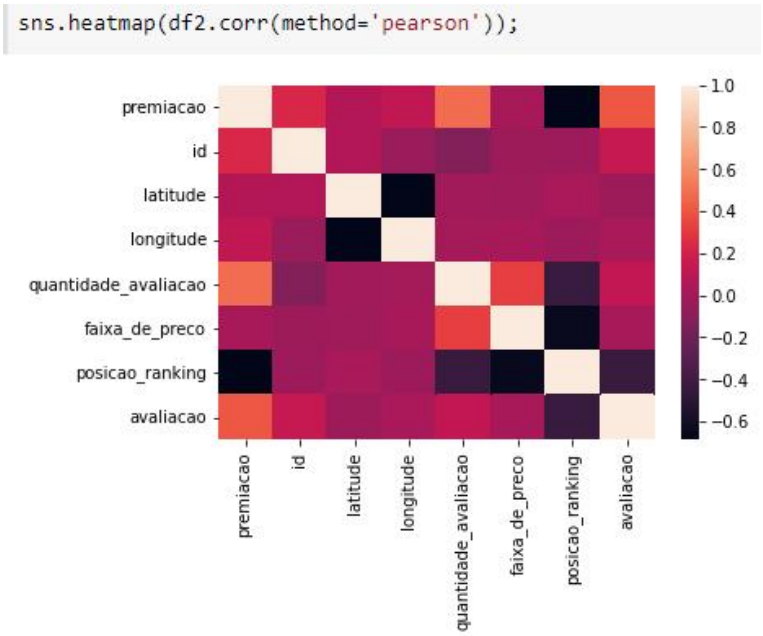


Figura 28 – Correlação dos dados do TripAdvisor com Heatmap utilizando o método de Pearson

Fonte: Elaborado pelo autor (2022)

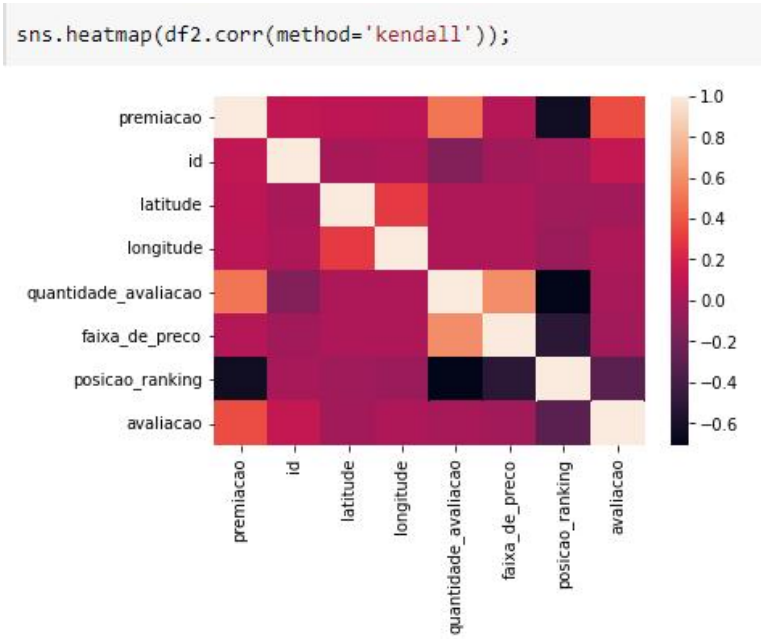


Figura 29 – Correlação dos dados do TripAdvisor com Heatmap utilizando o método de Kendall

Fonte: Elaborado pelo autor (2022)

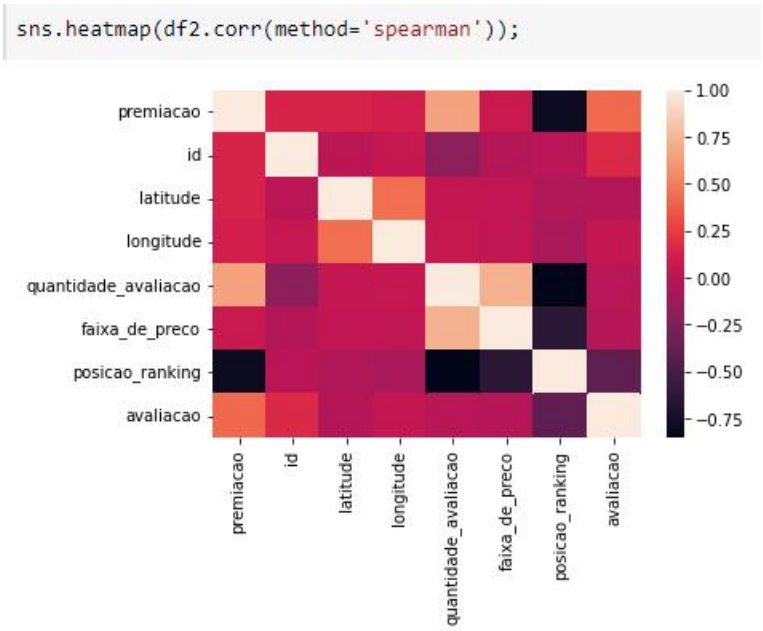


Figura 30 – Correlação dos dados do TripAdvisor com Heatmap utilizando o método de Spearman

Fonte: Elaborado pelo autor (2022)

Na análise dos dados do Tripadvisor, podemos considerar que com a utilização da metodologia de Pearson os atributos de premiação e quantidade de avaliação chegaram a um coeficiente positivo e moderado, premiação e avaliação também obtiveram correlação positiva de forma moderada na faixa de 0,40. Para esse caso latitude e longitude tiveram as correlações de forma negativa forte em -0,68. Quantidade de avaliações e a faixa de preço obtiveram uma correlação de 0,32, positiva e moderada. Posição do ranking obteve uma correlação negativa forte tanto com premiação como com faixa de preços.

Para a metodologia de Kendall, o padrão segue para premiação e as notas de avaliação, entretanto premiação e quantidade de avaliações a correlação se torna positiva mais forte. Latitude e longitude apresentam uma correlação positiva moderada. A quantidade de avaliações possui uma correlação positiva forte com a faixa de preço, chegando a 0,60 e negativa forte com posição no ranking em -0,71.

Spearman e sua metodologia nos mostra que os coeficientes com correlação positiva forte identificados foram: premiação e quantidade de avaliações onde o valor se estabeleceu em 0,65. Premiação também teve uma correlação positiva moderada com as notas de avaliação, em 0,42. Contudo para premiação e posição do ranking houve uma correlação negativa forte de -0,79. Latitude e longitude seguiram padrões parecidos com o do Airbnb onde a correlação entre eles foi positiva e moderada. A faixa de preço e a quantidade de avaliações tiveram uma correlação positiva muito forte, sendo 0,73. Posição do ranking em relação a quantidade de avaliações e faixa de preços obteve uma correlação negativa e

forte.

Portanto nesta etapa seguimos com a correlação entre as duas bases de dados, Airbnb e Tripadvisor, através da biblioteca pandas, concatenando os dois *dataframes* gerados. Para o eixo horizontal está a legenda dos campos pertencentes ao Airbnb e para o eixo vertical está a legenda dos campos pertencentes ao TripAdvisor:

```
pd.concat([df, df2], axis=1, keys=['df', 'df2']).corr(method='pearson').loc['df2', 'df']
```

	latitude	longitude	quantidade_hospedes	preco	avaliacao
premiacao	-0.169627	-0.110177	-0.156733	-0.178402	0.086175
id	0.011409	0.015089	0.029681	-0.018940	0.025946
latitude	0.003090	-0.020971	0.030295	0.025310	-0.024070
longitude	-0.014046	0.005862	-0.011566	-0.015355	0.019559
quantidade_avaliacao	-0.026168	-0.017591	-0.053951	0.020720	0.002739
faixa_de_preco	-0.004198	0.016352	0.015727	0.090105	-0.029313
posicao_ranking	-0.000238	-0.022146	-0.034373	-0.121155	0.037563
avaliacao	-0.002090	0.007302	-0.001865	0.020310	-0.007591

Figura 31 – Correlação dos dados do Airbnb e TripAdvisor utilizando o método de Pearson

Fonte: Elaborado pelo autor (2022)

Seguindo o padrão das análises, foi usado a biblioteca seaborn para trazer uma melhor visualização:

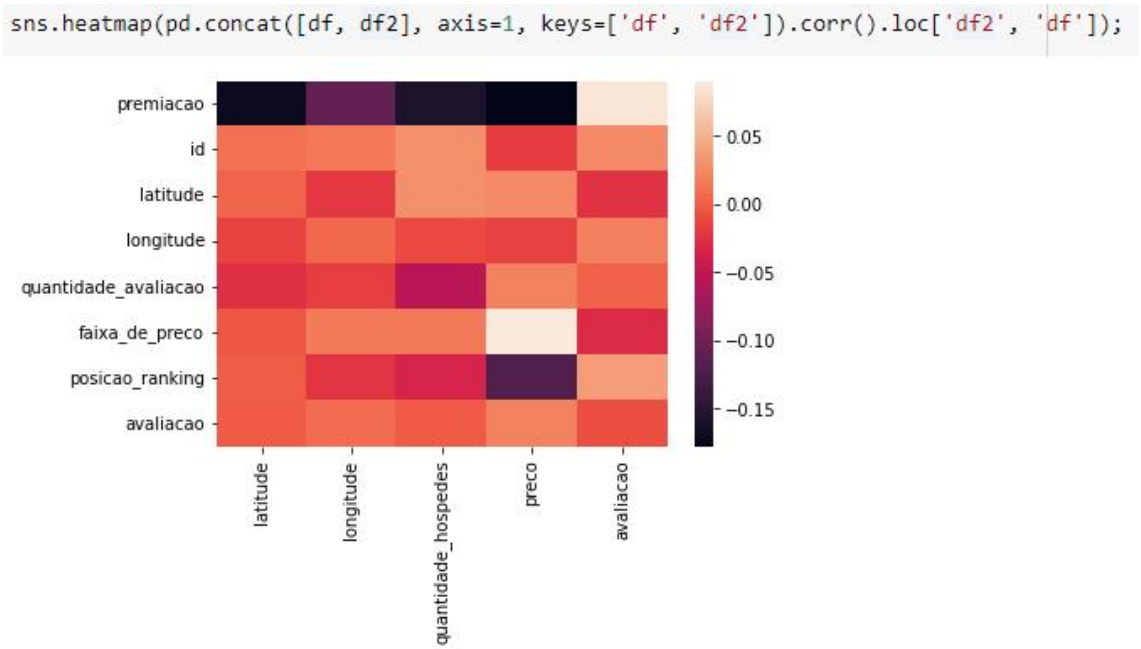


Figura 32 – Correlação dos dados do Airbnb e TripAdvisor utilizando Heatmap e o método de Pearson

Fonte: Elaborado pelo autor (2022)

A partir da concatenação entre os dois *dataframes* propostos para a análise, no eixo Y os atributos pertencentes ao Tripadvisor e no eixo X os atributos pertencentes ao Airbnb, pode ser observado que não houveram correlações positivas e nem negativas com grau de forte e moderada. Os maiores valores positivos foram entre a correlação da premiação do tripadvisor, com as notas das avaliações do airbnb. Além de faixa de preço dos restaurantes, bares e cafés coletados pelo Tripadvisor estarem correlacionadas positivamente porém com uma força fraca com os preços do Airbnb.

6 Considerações Finais

Conforme o número de pessoas utilizando as plataformas onlines de hospedagens cresceu, fez com que esses sites como Airbnb se tornassem um ótimo lugar para fazer coleta de dados e aplicar em pesquisas. Sites relacionados ao turismo surgiram e trouxeram aos viajantes a possibilidade de analisarem os destinos através dos comentários e indicações de outras pessoas da rede. Neste trabalho foram realizadas algumas implementações, a extração das acomodações e os detalhes relacionados a elas disponíveis no Airbnb, a extração de dados relacionados aos restaurantes e acomodações disponibilizadas no Tripadvisor, ambas para a cidade de Florianópolis. Foram utilizados dois Web Scrappers focados em varrer os sites tanto do Airbnb quanto do Tripadvisor e armazenar essas extrações em arquivos de formato CSV. Conforme foram gerados os arquivos, fez-se necessário um tratamento e a limpeza dos dados para uma análise com maior precisão e qualidade. Sendo assim os dados ficaram disponíveis para serem usados a fim de encontrar os conjuntos de indicadores.

Foram desenvolvidas as fases de análises e comparações entre os atributos extraídos e coletados, identificando os atributos com maior correlação através das matrizes de correlação de Pearson, Kendall e Spearman junto com as médias de preços empregadas na cidade de Florianópolis, outros fatores levados em consideração foram as avaliações e comentários deixados pelos usuários. Além da fonte de dados do Airbnb foi acrescentada a fonte de dados do TripAdvisor reunindo dados do contexto de restaurantes e atrações envolvendo a cidade de Florianópolis.

Os objetivos específicos foram atingidos. Através da obtenção dos dados e das análises, foi possível observar os valores empregados e as faixas de preço tanto para as acomodações quanto para os restaurantes e atrações da cidade de Florianópolis. As distribuições geográficas foram apresentadas através de mapas de calor, demonstrando que para os bairros do Norte da Ilha existem muito mais ofertas de acomodações que outras partes da ilha, contudo a parte Central e o Leste da Ilha também apresentam um bom volume. Já em relação as atrações o Centro e o Leste da ilha possuem um índice muito superior ao restante. Ao falarmos de restaurantes, bares e cafés o Centro contém uma larga dominância em opções. Os comentários e avaliações analisados trouxeram algumas visões de que de acordo com a amostra coletada, grande parte dos visitantes e da população, avalia bem as acomodações, os restaurantes e as atrações propostas por Florianópolis, tendo visto que os comentários positivos se destacaram.

Para as correlações observou-se que houveram diferenças nos tamanhos de efeito sugerido por (COHEN, 1992) para os dois conjuntos de dados com as diferentes metodologias

empregadas. Tanto Airbnb como o Tripadvisor tiveram correlações fortes tanto positivamente quanto negativamente em algumas variáveis individuais. Quando os dois conjuntos foram correlacionados, analisou-se que as variáveis de faixa de preço do Tripadvisor e o preço do Airbnb obtiveram forças de correlação fracas.

Para os trabalhos futuros, existe a possibilidade de desenvolver e aprimorar as análises propostas com novas fontes de dados, pode-se trazer outros atributos pertencentes ao Airbnb, como as comodidades das acomodações e as influências nos valores praticados. Outros atributos do Tripadvisor também podem ser explorados, como a possibilidade de usar os dados relacionados aos hotéis presentes na plataforma. Novas métricas e indicadores podem ser implementados para o estudo para que integrem as correlações entre os diversos aplicativos de turismo para a cidade de Florianópolis. Um aprofundamento da análise das distribuições usando estatísticas descritivas que descrevam a tendência central (média e mediana) e dispersão (variância e desvio padrão) de dados com valores numéricos, que acabam adicionando uma camada de detalhes e podem ser utilizadas para fazer comparações com outros conjuntos de dados.

Referências

AHUJA, M.; SINGH, J.; NICA, V. Web crawler: Extracting the web data. *International Journal of Computer Trends and Technology*, v. 13, p. 132–137, 07 2014. Citado 3 vezes nas páginas 18, 25 e 26.

BLAKE, M. *Pricing Strategy Development: Airbnb Market Analysis with Python*. 2021. Disponível em: <https://medium.com/codex/pricing-strategy-development-airbnb-market-analysis-with-python-f873be137346>. Citado 2 vezes nas páginas 21 e 22.

CALABRESE, B. Data cleaning. In: _____. [S.l.: s.n.], 2018. ISBN 9780128096338. Citado na página 28.

COHEN, J. Quantitative methods in psychology. *Psychological Bulletin*, p. 155, 06 1992. Citado 2 vezes nas páginas 46 e 55.

COPELLI, M. *airbnb-analytics*. 2022. Disponível em: <https://github.com/maxCopell/tripadvisor-scraper>. Citado na página 27.

CUNNINGHAM, P. et al. Does tripadvisor makes hotels better? 01 2010. Citado na página 16.

DUONG, T. *Airbnb Scraper*. 2021. Disponível em: <https://github.com/dtrungtin/actor-airbnb-scraper>. Citado na página 27.

GARCíA, J. Álvarez et al. Big data and tourism research: Measuring research impact. *Quality and Quantity*, 09 2020. Citado na página 12.

GUTTENTAG, D. Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, v. 18, p. 1–26, 12 2013. Citado 2 vezes nas páginas 15 e 16.

GUTTENTAG, D. et al. Why tourists choose airbnb: A motivation-based segmentation study. *Journal of Travel Research*, v. 57, p. 004728751769698, 04 2017. Citado na página 16.

ILYAS, I.; CHU, X. *Data Cleaning*. [S.l.: s.n.], 2019. ISBN 9781450371520. Citado na página 28.

IOANNIDES, D.; RöSLMAIER, M.; ZEE, E. van der. Airbnb as an instigator of 'tourism bubble' expansion in utrecht's lombok neighbourhood. *Tourism Geographies*, Routledge, v. 21, n. 5, p. 822–840, 2019. Citado na página 16.

JOÃO, G. *Análise dos dados do Airbnb — Dublin*. 2021. Disponível em: <https://medium.com/data-hackers/an%C3%A1lise-dos-dados-do-airbnb-dublin-32594035f90a>. Citado 2 vezes nas páginas 22 e 23.

LUO XUANYU ZHOU, Y. Z. Y. Predicting airbnb listing price across different cities. 12 2019. Citado 2 vezes nas páginas 20 e 21.

MINCA, C.; ROELOFSEN, M. Becoming airbnbeings: on datafication and the quantified self in tourism. In: _____. [S.l.: s.n.], 2022. p. 95–116. ISBN 9781003265429. Citado na página 12.

MIRTAHERI, S. et al. A brief history of web crawlers. 05 2014. Citado na página 17.

NEMESLAKI, A.; POCSAROVSZKY, K. Web crawler research methodology. 01 2011. Citado na página 25.

OSKAM, J.; BOSWIJK, A. Airbnb: The future of networked hospitality businesses. *Journal of Tourism Futures*, v. 2, p. 22–42, 03 2015. Citado na página 12.

PERES, C.; PALADINI, E. Exploring the attributes of hotel service quality in florianópolis-sc, brazil: An analysis of tripadvisor reviews. *Cogent Business and Management*, v. 8, p. 1926211, 01 2021. Citado na página 13.

SANTOS-JúNIOR, A. et al. Entendiendo la gobernanza de los destinos turísticos inteligentes: el caso de florianópolis - brasil understanding smart tourism destinations' governance: the case of florianópolis - brazil. v. 4, p. 29–39, 05 2019. Citado na página 26.

TIAN, Z. Use python data analysis to gain insights from airbnb hosts. *Advances in Mathematical Physics*, v. 2021, p. 1–10, 10 2021. Citado na página 20.

TRIPADVISOR. *About Tripadvisor*. 2022. Disponível em: <https://tripadvisor.mediaroom.com/us-about-us>. Citado na página 16.

WEIGEL, C. O uso de big data para análise de oferta de imóveis via airbnb em destinos turísticos. 11 2019. Citado na página 19.

YOO, K.-H.; SIGALA, M.; GRETZEL, U. Exploring tripadvisor. In: _____. [S.l.: s.n.], 2016. p. 239–255. ISBN 978-3-642-54088-2. Citado 2 vezes nas páginas 12 e 17.

ZHU, G.; KUBICKOVA, M. From homeowner to airbnb host: The role of trust and perceived value. *Journal of Quality Assurance in Hospitality and Tourism*, p. 1–23, 01 2022. Citado na página 15.

Análise e correlação de dados: um estudo de caso usando o AirBnB e o TripAdvisor em Florianópolis

André C. Machado[1]

[1]Departamento de Informática e Estatística
– INE –
Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 5040 – Trindade – Florianópolis – SC – Brazil

andre_machado123@hotmail.com

Abstract. *The national hosting sector has shown a growth scenario over the years, largely due to internet access and the ease of booking through apps. The amount of data generated by web users grows more and more, on the other hand there is the complexity in extracting and obtaining this data. The analyzes seek to bring informative graphs and indicators, showing the tourism sector a relationship between the AirBNB platform of the accommodation sector and the Tripadvisor platform of the tourism evaluation sector through the identifications of the relationship of the attributes provided by the accommodation, the evaluations and comments. Thus allowing to visualize their impact on the sector.*

Resumo. *O setor de hospedagem nacional apresentou um cenário de crescimento conforme o passar dos anos, grande parte devido ao acesso a internet e a facilidade de reservas através de apps. A quantidade de dados gerados por usuários da web cresce cada vez mais, por outro lado existe a complexidade na extração e obtenção desses dados. As análises buscam trazer gráficos informativos e indicadores, mostrando ao setor de turismo uma relação entre a plataforma AirBNB do setor de hospedagem e da plataforma Tripadvisor do setor de avaliações turísticas através das identificações de relação dos atributos fornecidos pela hospedagem, as avaliações e comentários. Permitindo assim visualizar o impacto das mesmas sobre o setor.*

1. Introdução

Atualmente, a maior parte dos processos de informações turísticas são realizados eletronicamente, incluindo o aluguel de hospedagens. Os clientes deixam suas impressões digitais em grande parte das atividades realizadas tanto no planejamento da viagem como durante e após, mas também através de comentários sobre diferentes plataformas. Consequentemente, uma grande quantidade de dados sobre as necessidades e comportamentos dos clientes, bem como a sua percepção dos serviços, são armazenados em várias fontes [Álvarez García et al. 2020]. Com isso, surgem as aplicações como o Airbnb, que fornecem as mais variadas opções de hospedagens para os viajantes, trazendo facilidade e substituindo parte dos intermediários do plano, tais como as agências de turismo ou corretores imobiliários.

Os conteúdos criados pelos viajantes são percebidos como altamente confiáveis, credíveis e relevantes, atualizados e envolventes. Assim, as pessoas que planejam viagens

geralmente levam em consideração as avaliações geradas por outros viajantes durante o processo de tomada de decisão, pois a intangibilidade das experiências de turismo impossibilita o teste de pré-compra e, portanto, aumenta a necessidade de relatórios de experiência em primeira pessoa [Yoo et al. 2016]. Aplicações como o TripAdvisor, surgem com o intuito de suprir essa necessidade em uma plataforma única, onde é criado um fórum de viajantes, compartilhando as mais variadas experiências.

Enquanto o significado computacional de uma plataforma é uma 'infra-estrutura programável sobre a qual outro software pode ser construído e executado', no discurso público o termo plataforma é cada vez mais usado para descrever empresas que oferecem serviços web 2.0 e oferecem uma oportunidade de comunicação, interação e vendas. As plataformas de turismo estão centradas principalmente na mobilidade, acomodação, alimentação e experiências de viagem. O Airbnb faz parte de um conjunto específico de plataformas digitais que facilitam a troca monetária de acomodações residenciais (casas particulares, quartos e leitos) e experiências turísticas entre os indivíduos [Minca and Roelofsen 2022]. Já o TripAdvisor faz parte de outro conjunto, um site de informações com uma comunidade compartilhando conteúdo de viagens do mundo inteiro e capacitando os usuários a escrever, pesquisar e compartilhar resenhas de viagens [Yoo et al. 2016].

Com o passar do tempo e a facilidade das reservas das hospedagens, o número de usuários da plataforma vêm mudando, mudando assim os destinos e os padrões que eram comumente utilizados. O Airbnb é uma inovação desafiadora à qual a hospitalidade tradicional tem que se adaptar [Oskam and Boswijk 2015]. A confiança desempenha o papel principal para a tomada de decisão nas reservas de hospedagens. A plataforma age como intermediária e garante algumas seguranças para os seus usuários, incluindo avaliações dos próprios usuários sobre a experiência com a acomodação. Algo que a plataforma do TripAdvisor também se propõe a fazer, entretanto abrangendo outros segmentos.

Diante dessas informações, o presente estudo busca realizar a implementação de uma aplicação que coleta dados web de sites como o Airbnb e TripAdvisor, utilizando o processo de *Web Scraping*, extraindo os dados e convertendo-os em informações estruturadas. Os dados extraídos devem trazer os valores das hospedagens, a faixa de preços dos restaurantes, as localizações dentro da cidade, o ranking, as avaliações e comentários. Para armazená-los será utilizado um banco de dados.

O município de Florianópolis no estado de Santa Catarina foi selecionado para este estudo por ser um dos mais visitados destinos turísticos do Brasil. É um destino reconhecido mundialmente por suas belezas naturais, cercado de praias, e pela qualidade de vida que proporciona. A cidade é a capital brasileira com maior pontuação no Índice de Desenvolvimento Humano (IDH). A economia da cidade é fortemente baseada em tecnologia da informação, turismo e serviços [Peres and Paladini 2021].

Esta aplicação deve atender aos interesses de órgãos responsáveis pelo turismo nacional e empresas relacionadas ao turismo que buscam encontrar a média de preços utilizados nas acomodações, além da faixa de valores para os restaurantes, assim como os comentários e avaliações. Através da implementação de um esquema de visualização das informações, mapas e gráficos combinados o resultados devem ser apresentados através da plataforma PowerBI, e por fim a correlação entre esses dados.

2. Fundamentação Teórica

Neste capítulo, é discutido o referencial teórico necessário para compreensão da proposta apresentada pelo presente trabalho. Inicialmente, a fonte dos dados *Airbnb* é contextualizada, trazendo sua importância ao disponibilizar diversas acomodações online através de uma plataforma com diversas informações. Depois, a fonte de dados TripAdvisor é contextualizada, fornecendo informações e opiniões de conteúdos relacionados ao turismo. Por fim são apresentados os conceitos de extração de dados.

2.1. Plataforma Airbnb

O Airbnb constrói uma ponte de e-service entre os viajantes e os proprietários para satisfazer a demanda e a oferta neste mercado de dois lados. Ele permite que um anfitrião anuncie uma propriedade, como uma casa ou quarto, para aluguel de curto prazo, enquanto permite que o turista viva como um local. Desde o seu lançamento em 2008, esse tipo de acomodação compartilhada ponto a ponto tornou-se uma força muito disruptiva para o setor de hospitalidade tradicional. Nos Estados Unidos, o Airbnb teve um crescimento de demanda de 30% nos últimos anos, atingindo 5% de participação de mercado com aproximadamente 30% de penetração de mercado [Zhu and Kubickova 2022].

As tecnologias da Web 2.0 permitiram o modelo de negócios inovador do Airbnb, mas ser disruptivo, deve eventualmente haver demanda por um produto. A demanda por um serviço como o Airbnb não é um dado adquirido, pois o Airbnb é consideravelmente carente em muitas das áreas que são mais importantes para os turistas na escolha do alojamento hoteleiro, como a qualidade do serviço, simpatia da equipe, reputação da marca e segurança. Como foi discutido, no entanto, produtos disruptivos geralmente têm desempenho inferior com diz respeito aos atributos-chave dos produtos predominantes, mas produtos disruptivos também são frequentemente mais baratos e oferecem novos benefícios. Muito pelo contrário, a acomodação do Airbnb é normalmente mais barato do que a acomodação tradicional, e a acomodação do Airbnb introduz benefícios associados à permanência numa residência [Guttentag 2013].

Além dos preços econômicos, as acomodações do Airbnb também oferecem diversos benefícios advindos da permanência em uma residência. Por exemplo, alguns turistas podem preferir a sensação de estar em uma casa sobre um hotel, e os anfitriões do Airbnb podem fornecer conselhos locais úteis. Os hóspedes do Airbnb também costumam ter acesso a comodidades residenciais práticas, como cozinha completa, máquina de lavar e secadora. A experiência de morar em uma residência também oferece aos hóspedes a chance de ter uma experiência mais local, interagindo com o anfitrião ou vizinhos, e possivelmente ficar em uma área 'não turística', já que as acomodações do Airbnb tendem a ser mais dispersas do que as acomodações tradicionais [Guttentag 2013].

A vantagem para os proprietários destes imóveis é que através da plataforma online podem chegar facilmente a um mercado global. Simultaneamente, usando o Airbnb, os visitantes têm acesso a uma gama cada vez maior de opções de acomodação durante a viagem [Guttentag et al. 2017].

O Airbnb e empresas semelhantes enfrentam um ressentimento crescente dos moradores locais que temem que esses empreendimentos, juntamente com muitas outras atividades relacionadas ao turismo, transformem seus bairros residenciais outrora tranquilos em guetos de visitantes. Enquanto isso, os municípios lutam para identificar maneiras

de regular o crescimento do Airbnb, seja por meio de tributação ou pela imposição de medidas drásticas destinadas a limitar ou erradicar totalmente os aluguéis de curto prazo. Por exemplo, a expansão fenomenal do Airbnb em Reykjavik levou o governo islandês a impor restrições à transformação de mais casas e quartos em aluguéis de curto prazo. Da mesma forma, Berlim e Amsterdã limitam por quanto tempo as propriedades podem ser alugadas pelo Airbnb [Ioannides et al. 2019].

2.2. Plataforma Tripadvisor

Quando falamos dos benefícios de viver na era da informação, há poucos exemplos mais emblemáticos desses benefícios do que o impacto da Internet no turismo e nas viagens. Com a enorme quantidade de informações disponíveis na Internet sobre destinos de viagem e opções de hospedagem, o planejamento de viagens pessoais tornou-se um grande passatempo. Esse aumento na disponibilidade de informações resulta em viajantes mais bem informados, o que, por sua vez, leva a um mercado mais eficiente [Cunningham et al. 2010].

O Tripadvisor, a maior plataforma de orientação de viagens do mundo, ajuda centenas de milhões de pessoas todos os meses a se tornarem melhores viajantes, desde o planejamento até a reserva e a realização de uma viagem. Viajantes de todo o mundo usam o site e o aplicativo do Tripadvisor para descobrir onde ficar, o que fazer e onde comer com base nas orientações de quem já esteve lá. Com mais de 1 bilhão de avaliações e opiniões de quase 8 milhões de empresas, os viajantes recorrem ao Tripadvisor para encontrar ofertas de acomodações, reservar experiências, reservar mesas em restaurantes deliciosos e descobrir ótimos lugares nas proximidades. Como uma empresa de orientação de viagens disponível em 43 mercados e 22 idiomas, o Tripadvisor facilita o planejamento, independentemente do tipo de viagem [Tripadvisor 2022].

A plataforma oferece vários serviços direcionados a consumidores e empresas e adiciona continuamente novos serviços e recursos para atender às necessidades em evolução de viajantes e fornecedores de turismo.Entre vários tópicos de conteúdo gerados pelos usuários, os conteúdos relacionados ao turismo são frequentemente os assuntos mais populares compartilhados e consumidos [Yoo et al. 2016].

É importante ressaltar que o uso das mídias sociais está cada vez mais integrado em todas as fases da experiência turística. No entanto, o TripAdvisor também é um infomediário especializado no campo de 'Big Data' e focado em vincular e atender as necessidades tanto da demanda quanto da oferta turística, fornecendo uma plataforma tecnológica na qual o conteúdo pode ser criado, analisado e distribuído para atender às necessidades de viajantes e empresas de turismo [Yoo et al. 2016].

2.3. Extração de dados

Crawling é o processo de explorar uma aplicação da web automaticamente. O web crawler visa descobrir na internet páginas de um aplicativo da Web navegando pelo aplicativo. Isso geralmente é feito simulando as possíveis interações do usuário. À medida que a quantidade de informações na web vem aumentando drasticamente, os usuários da web dependem cada vez mais dos mecanismos de busca para encontrar os dados desejados. Para que os motores de busca aprendam sobre os novos dados à medida que se tornam disponíveis na web, ele precisa rastrear e atualizar constantemente o mecanismo de pesquisa da base de dados [Mirtaheri et al. 2014].

A definição tradicional de um web crawler assume que todos o conteúdo de um aplicativo da web é acessível por meio de URLs. Logo na história do rastreamento na web ficou claro que rastreadores da Web não podem lidar com as complexidades adicionadas por aplicativos da Web interativos que dependem da entrada do usuário para gerar páginas da Web. Esse cenário geralmente surge quando o aplicativo da Web é uma interface para um banco de dados e depende da entrada do usuário para recuperar o conteúdo do banco de dados. O novo campo de Deep Web-Crawling nasceu para resolver esse problema [Mirtaheri et al. 2014].

Um rastreador da web é um dos principais componentes de motores de pesquisa na web. O crescimento do rastreador da web está aumentando na mesma forma como a web está crescendo. Uma lista de URLs está disponível com o rastreador da web e cada URL é chamado de semente. Cada URL é visitado pelo web crawler. Ele identifica os diferentes hiperlinks na página e os adiciona à lista de URLs a serem visitados. Esta lista é denominado como fronteira de rastreamento. Usando um conjunto de regras e políticas, URLs na fronteira são percorridos individualmente. Páginas diferentes da Internet são coletadas pelo analisador e o gerador é armazenado no sistema de banco de dados da pesquisa motor. Os URLs são então colocados na fila e depois é agendado, e pode ser acessado um a um por o motor de busca, um por um, sempre que necessário. As ligações e arquivos relacionados que estão sendo pesquisados podem ser disponibilizados sempre que necessário em momento posterior de acordo com os requisitos. Com a ajuda de algoritmos adequados, os rastreadores da Web encontram o links relevantes para os motores de busca e usá-los ainda mais. Bancos de dados são máquinas muito grandes como o DB2, usadas para armazenar grandes quantidade de dados [Ahuja et al. 2014].

Existem vários usos de rastreadores da web: Os rastreadores também podem ser usados para automatizar tarefas de manutenção em um site, como verificar links ou validar código HTML. Os rastreadores podem ser usados para coletar tipos específicos de informações de páginas da Web, como coletar endereços de e-mail (geralmente para spam). Os mecanismos de pesquisa costumam usar rastreadores da Web para coletar informações que estão disponíveis em páginas da Web públicas. Eles coletam dados para que, quando os internautas inserirem um termo de pesquisa em seus site, eles podem fornecer rapidamente ao surfista sites relevantes. Os linguistas usam rastreadores da web para realizar uma análise textual. Eles percorrem a Internet para determinar quais palavras são comumente usadas hoje [Ahuja et al. 2014].

3. Desenvolvimento

Construir um Web Scraper, antes de tudo, requer o conhecimento de como exatamente os usuários navegam em um site e o que acontece durante esse processo do ponto de vista do processamento de informações.[Nemeslaki and Pocsarovszky 2011]

Neste capítulo, é proposto um modelo de extração de dados e análise utilizado para coletar os dados disponíveis na plataforma do Airbnb e na plataforma Tripadvisor. Nas subseções seguintes serão descritos a visão geral, os passos da utilização do Web Scraper e do modelo utilizado para o carregamento e a transformação dos dados.

3.1. Visão Geral

Esta seção tem como objetivo apresentar as aplicações desenvolvidas neste trabalho, mostrando seus tipos diferentes de componentes, as interações que ocorrem e os resul-

tados de saída.

Como pode ser visto na Figura 1, o Scaper deve varrer as páginas do Airbnb que contemplem a cidade alvo da busca, no caso Florianópolis. Ao encontra-la deve salvar os dados de cada hospedagem e suas características em um arquivo.

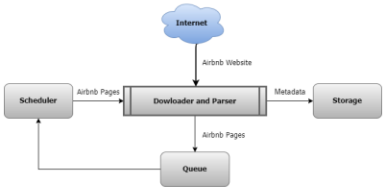


Figure 1. Arquitetura de um Web Scaper para o Airbnb

Assim como o Scraper do Airbnb, o Scraper para o Tripadvisor segue o mesmo padrão de fluxo como pode ser visto na Figura 2, o Scraper também deve varrer as páginas do Tripadvisor que contemplem a cidade de Florianópolis. Ao encontra-la deve salvar os dados em um arquivo.

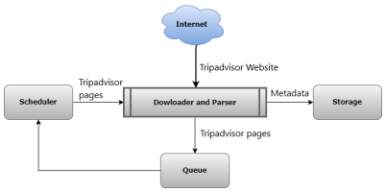


Figure 2. Arquitetura de um Web Scraper para o Tripadvisor

Nas subseções seguintes são descritos, a localização escolhida, a organização dos dados, o desenvolvimento e implementação do Web Scraper e a carga de dados.

3.2. Localização dos dados

A cidade de Florianópolis, capital do estado de Santa Catarina na região sul do Brasil, conta com uma população estimada de 516 mil pessoas. A cidade tem seu desenvolvimento econômico baseado no turismo, inovação tecnológica e no setor de serviços. Pode-se notar que Florianópolis é considerada uma das mais importantes cidades inteligentes no Brasil pelo estudo de Sistemas Urbanos (2017) e está posicionada entre as cidades brasileiras mais empreendedoras Santos2019.

A escolha da cidade para o estudo vêm da necessidade de encontrar mais informações sobre as hospedagens hoje disponíveis na plataforma Airbnb. Considerando os valores empregados e as características que determinam esses valores. Além de analisar outros fatores da cidade, como as atrações e restaurantes disponíveis na plataforma do Tripadvisor. Auxiliando os anfitriões das residências de Florianópolis a terem uma noção melhor sobre as avaliações das hospedagens, os valores empregados e os fatores que podem se relacionar a isso.

3.3. Coleta dos dados

A obtenção dos dados da cidade de Florianópolis, se da através de dois Web Scrapers, disponibilizados na plataforma Apify. Para o acesso a essa plataforma é necessário criar

uma conta que possui algumas características limitadas, entretando as extrações para o estudo de caso estão dentro do valor disponibilizado gratuitamente. Os dados coletados foram primeiramente armazenados em arquivos de formato CSV.

O Web Scraper relacionado ao Airbnb, está disponibilizado no GitHub por [Duong 2021]. O Airbnb Scraper é projetado para extrair a maioria dos dados do Airbnb disponíveis publicamente para os anúncios de acomodações. É possível obter todos os dados básicos sobre o anúncio, as avaliações, comentários, preços, detalhes do anfitrião e também do hóspede.

Para o estudo com o Airbnb como fonte de dados. Os atributos a serem coletados são:

• url: url da acomodação identificando o acesso a página;

• nome_acomodacao: nome da acomodação identificando um dos itens listados no site Airbnb;

• tipo_acomodacao: identificador do tipo de acomodação, exemplo: casa, quarto inteiro, quarto compartilhado, etc;

• quantidade_hospedes: a quantidade de hóspedes máxima permitida;

• preco: preço diário de aluguel da acomodação;

• superhost: identificador de um super anfitrião;

• localizacao: endereço contendo o bairro e a cidade da acomodação;

• latitude: latitude relacionada a acomodação;

• longitude: longitude relacionada a acomodação;

• avaliacao: notas de avaliação das acomodações;

• comentario: os comentários realizados pelos hóspedes em relação a acomodação;

• lingua_comentario: a lingua utilizada nos comentários realizados pelos hóspedes em relação a acomodação;

Em relação ao Web Scraper destinado ao Tripadvisor, está disponibilizado no GitHub por [Copelli 2022]. O Tripadvisor Scraper permite obter dados do Tripadvisor. Ele é adequado para casos de uso onde é necessário coletar avaliações, e-mails, endereços, prêmios e muitos outros atributos de hotéis e restaurantes e atrações das cidades.

No caso do Tripadvisor como fonte de dados alguns atributos disponíveis na extração foram eliminados, os dados relacionados a hotéis por exemplo não são coletados. Os atributos a serem coletados são:

• web_url: url do restaurante ou atração identificando o acesso a página;

• nome: nome do restaurante ou atração identificando um dos itens listados no site Tripadvisor;

• tipo: identificador para definir o tipo: restaurante ou atração;

• premiacao: identificador se o restaurante já foi premiado com o Certificado de Excelência entregue pelo Tripadvisor;

• posicao_ranking: posição que o restaurante esta no ranking de classificação do Tripadvisor ;

• faixa_de_preco: faixa de preço do restaurante, classificada entre $,$$, $$$,$$$$;

• tipo_culinaria: tipo de culinária utilizado no restaurante, exemplo: brasileira, japonesa, etc;

• endereco: endereço contendo a rua, bairro e cidade do restaurante ou atração;

• latitude: latitude relacionada ao restaurante ou atração;

• longitude: longitude relacionada ao restaurante ou atração;

• avaliacao: notas de avaliação do restaurante ou atração;

• comentario: os comentários realizados pelos clientes ou visitantes em relação aos restaurantes ou atrações;

• lingua_comentario: a lingua utilizada nos comentários realizados pelos visitantes em relação aos restaurantes e atrações;

3.4. Limpeza dos dados

A qualidade dos dados é um dos problemas mais importantes no gerenciamento de dados, pois dados sujos geralmente levam a resultados imprecisos de análise de dados e decisões de negócios incorretas.[Ilyas and Chu 2019] A limpeza de dados inclui todas as metodologias cujo objetivo é "detectar e remover erros e inconsistências dos dados para melhorar a qualidade dos dados. [Calabrese 2018]

Após a coleta de dados, os dados foram limpos e tratados. Para a limpeza foi necessário a imputação de valores não preenchidos com a palavra nulo, para padronizar os valores nulos. Houve tratamento nos atributos de preço, sendo necessário a retirada do cifrão, no atributo endereço foi necessário eliminar localidades que não se enquadravam na região de Florianópolis. Além disso foi preciso eliminar as duplicidades.

3.5. Scraper

Inicialmente é necessário percorrer as páginas do site Airbnb como um usuário normal, ele insere o destino, as datas desejadas e clica no botão de pesquisa. O mecanismo de classificação do Airbnb gera uma listagem de diversas acomodações com algumas breves descrições. Ao acessar as listagens é possível obter descrições com maiores detalhes das acomodações. Sendo assim, o Scraper deve extrair as informações de dois tipos de páginas, as de pesquisa e de detalhes.

Ao extrair as listagens da página de pesquisa foram criadas algumas funções, chamadas, findListings, getListingsSection, addListings, a biblioteca nos permite navegar pela árvore HTML a acessar os elementos, obtendo assim o o texto referente a listagem.

Após a definição para os atributos, entramos no processo de acessar todas as páginas da cidade de Florianópolis. Cada página possui 20 anúncios, e conforme realizada a combinação dos parâmetros de pesquisa o Airbnb fornece acesso a até 300 anúncios por localidade. Com isso foi criado a função findListings na qual tem o objetivo de varrer todas as páginas disponíveis, apenas inserindo o link inicial da pesquisa.

Para extrair os dados do Tripadvisor o método foi semelhante, percorrendo as páginas do site como um usuário, após a inserção da localidade uma lista é gerada com os restaurantes e as atrações disponíveis na região e suas descrições. O Scraper nesse caso deve extrair as informações de dois tipos de páginas, as de pesquisa e de detalhes novamente.

Entretanto as funções utilizadas para o Tripadvisor foram iniciadas através da função buildSearchRequestsFromLocationName, que recebe como parâmetro de entrada a localização desejada. A partir dela é realizado uma requisição para a obtenção da lista com os restaurantes e atrações através da função getRequestListSources.

3.6. Modelo banco de dados

Com o objetivo de salvar os dados em um banco de dados, foi necessário a criação de um modelo dimensional, separando algumas dimensões. A figura a seguir apresenta o esquema do banco de dados para o Airbnb no lado esquerdo e no lado direito para o Tripadvisor:

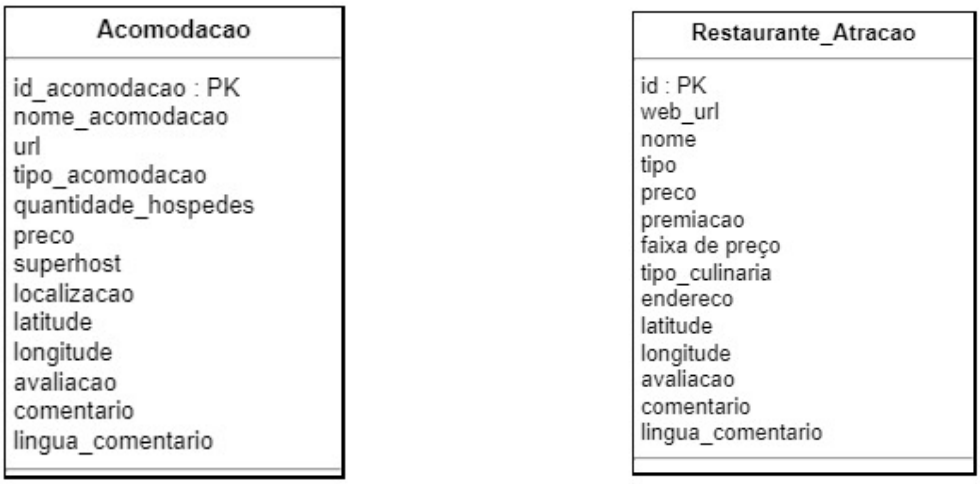


Figure 3. Modelo Dimensional para o Airbnb e Tripadvisor

Fonte: Elaborado pelo autor (2022)

3.7. Carga dos dados

Para o processo de carga dos dados e ETL, foi utilizado a ferramenta de Data Integration, Pentaho. Com ela foi possível realizar a etapa de limpeza dos dados. Ao fim do processo os dados foram carregados em um banco de dados PostgreSQL de código aberto.

4. Análise dos dados

4.1. Dados do Airbnb

De acordo com os dados coletados, é possível identificar alguns fatores dentro das acomodações encontradas para a região de Florianópolis. A amostra extraída são de 10.241 ofertas disponíveis que equivale a 100% dos dados.

4.1.1. Tipos de propriedades

Diante da região em destaque, foram encontrados alguns tipos diferentes de propriedades disponíveis, são estas: Apartamentos inteiros, Casas inteiras, Quartos inteiros, Quartos compartilhados e Outros que contemplam acomodações peculiares como barcos, contêineres e camper/motorhome.

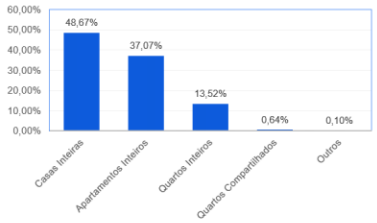


Figure 4. Tipos de propriedades
Elaborado pelo autor (2022)

Em relação aos resultados, foi possível observar como as casas e apartamentos oferecidas por inteiro correspondem a um total de aproximadamente 85,74% dos tipos de propriedades para a região de Florianópolis, um valor muito elevado comparando com os outros tipos disponíveis.

4.1.2. Avaliação

Cada anúncio disponível no site possui um espaço designado a avaliações e comentários realizados pelas pessoas que ali se hospedaram. O anfitrião por sua vez tem um retorno sobre seu serviço, além de que, para os demais consumidores as avaliações publicadas podem ser um fator decisivo para a escolha do local. É gerado um resultado que pode variar entre 1 a 5 estrelas, sendo 0 a nota mais baixa e 5 a mais alta. Para o estudo será utilizado um arredondamento das notas.

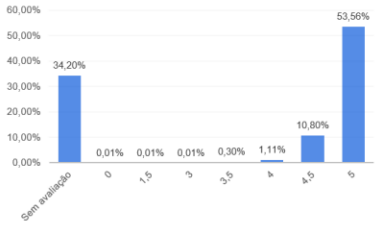


Figure 5. Classificação Airbnb
Elaborado pelo autor (2022)

Conforme observa-se a maior parte das avaliações coletadas estão entre 4,5 e 5, 10,80% para a nota 4,5 e 53,56% para a nota 5, totalizando um percentual de 64,36% nessa faixa de avaliação. Por outro lado 34,20% das acomodações não possuem notas de avaliação, nesse caso podemos considerar que muitos desses anúncios podem ser novos, nunca receberam hóspedes ou simplesmente os hospedes não fizeram suas avaliações. Já no caso das notas altas podemos inferir que quando um hóspede se dedica a fazer

uma avaliação, a grande maioria avalia com notas boas as acomodações da região de Florianópolis.

4.1.3. Comentários

Além das notas das avaliações, existem os comentários que podem ser realizados pelos hóspedes. Esses comentários podem servir de base para outros hóspedes tomaram suas decisões em se hospedarem ou não, comunicando se o anúncio segue os padrões da publicação ou se ocorreu alguma situação diferente do esperado.

Uma análise com uma amostra de comentários foi realizada, nesse sentido foram coletados 4.429 comentários em português e 3.497 em inglês de cada acomodação presente no estudo e com isso criado uma nuvem de palavras mais utilizadas nesses comentários. Para o experimento foi necessário a utilização da Ferramenta Microsoft Power BI, com o visual Word Cloud.

Figure 6. Comentários na língua portuguesa
Elaborado pelo autor (2022)

Com esse visual, as palavras mais usadas acabam se tornando maiores. Na imagem para os comentários na língua portuguesa, é possível notar o uso das palavras 'casa', mencionada 1.192 vezes, e 'apartamento', mencionado 704 vezes. Outras palavras que receberam bastante destaque foram 'localização', mencionada 669 vezes e 'praia' mencionada 658 vezes.

Analisando esses valores e as palavras encontradas, observa-se que é comum o uso da descrição do tipo de imóvel, casa ou apartamento, nos comentários. Outro fator é que para os comentários das acomodações em Florianópolis considerou-se muito o fator localização, e relacionado a isso a proximidade com as praias. A palavra, 'limpa', encontrada 402 vezes também traz um significado para a análise, sendo limpeza um outro ponto importante de atenção na hora de avaliar uma acomodação.

Os comentários quando realizados, tendem a ser positivos, como no caso do uso das palavras 'excelente', mencionado 691 vezes, 'ótima', mencionada 499 vezes e 'recomendo', mencionado 467 vezes, além de outras palavras que podem ser encontradas como: boa, maravilhosa, impecável, incrível, etc.

Florianópolis é conhecida por ser uma cidade muito turística, recebendo muitos visitantes de outros países, nesse sentido também houveram avaliações em outras línguas e no caso o filtro utilizado foi a língua inglesa, já que houve uma boa amostragem disponível.

Figure 7. Comentários na língua inglesa
Elaborado pelo autor (2022)

A nuvem de palavras para essa língua mantém alguns padrões parecidos. a palavra 'place' que pode ser traduzida como 'lugar' aparece 1.286 vezes sendo a mais utilizada. A segunda mais usada foi 'stay', traduzindo é interpretada como 'estadia', mencionada 979 vezes. A localização ou 'location' apareceu 810 vezes, e nesse caso pode ser considerado que o significado da utilização de localização tende a ser relacionado ao posicionamento da acomodação na geografia local, diferente da palavra lugar que pode estar mencionando a acomodação em si.

Palavras como apartamento e casa que foram as mais utilizadas em português, continuam sendo muito utilizadas na língua inglesa, 'apartament' ou apartamento, foi usado 719 vezes, e 'house' ou casa foi utilizado 613 vezes. Além da palavra 'beach' que significa praia, utilizada 806.

Entretanto a palavra 'host' que significa anfitrião foi muito mais usada na língua inglesa, 808 vezes, já na língua portuguesa foi utilizado 248 vezes. Nesse caso observa-se que para turistas da língua inglesa o fator anfitrião é muito mencionado nos comentários, sendo esse um ponto de atenção já que o tratamento com esses hóspedes pode ser diferenciado.

Existe uma relação muito grande entre as palavras usadas nos dois diferentes idiomas, sobretudo nos fatores indicados como tipo de acomodação, na localização e no quesito praia, além das palavras positivas representadas de diversas formas. Mas em um contexto geral os comentários para as acomodações em Florianópolis, tendem a ser voltadas positivamente para a localização geográfica e fatores como praia.

4.1.4. Preço

O preço das acomodações é um dos fatores a ser analisado pelos hóspedes. Uma acomodação pode conter várias comodidades, entretanto, o hóspede precisa ter condições de pagar para se hospedar ali.

O Airbnb disponibiliza uma faixa de preço referente à região pesquisada pelos clientes, com a finalidade de facilitar a pesquisa. A faixa de preço é composta pelas seguintes informações: o preço mais baixo e o preço mais alto dos anúncios listados em uma determinada região, além de disponibilizar a informação do preço médio da região. O hóspede então pode filtrar conforme o valor que ele está disposto a pagar por uma diária,

portanto, só serão apresentado os anúncios com o preço de acordo com o especificado.

Para a análise de Florianópolis foi observado uma grande diferença com relação ao preço da diária das acomodações. A média de preço encontrada na região foi de R$536,78 reais, sendo a menor diária de R$30,00 reais e a mais alta R$46.971,00 reais por dia.

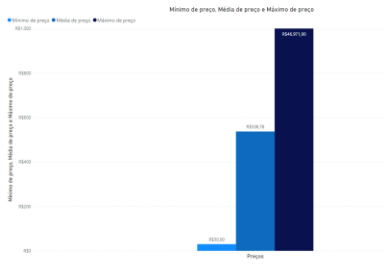


Figure 8. Mínimo,média e máximo de preços
Elaborado pelo autor (2022)

Conforme segmentamos os diferentes tipos de propriedades citados anteriormente (Apartamentos inteiros, Casas inteiras, Quartos inteiros, Quartos compartilhados e Outros) percebemos as diferenças nas médias dos valores diários. As casas inteiras como imaginado possuem o maior valor médio diário, sendo R$818,59 reais. Já os apartamentos inteiros possuem uma média de preço de R$425,07 reais. Os quartos inteiros segue um valor médio de R$376,12. Para os quartos compartilhados a média segue em R$121,60. E no quesito outros tipos de acomodações o valor médio é de R$3.411,78, entretanto nesse caso foram relacionados apenas 9 propriedades, sendo que duas delas, dois barcos, possuem a diária de R$17.500,00 e R$9.500,00, sem essas acomodações a média cai para R$506,00.

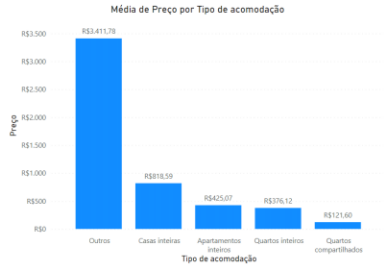


Figure 9. Média de Preços por acomodação
Elaborado pelo autor (2022)

A partir dos resultados obtidos foi possível verificar que o preço da diária pode ser realmente o fator decisivo para o consumidor. Observando uma variação entre os diferentes tipos de acomodações.

4.1.5. Localização

Em relação a localização das acomodações listadas no Airbnb, existem algumas regiões dentro de Florianópolis que podem ser classificadas de acordo com o espaço geográfico,

como bairros residenciais, praias e a parte central. Além da divisão Norte, Sul, Centro e Leste da ilha.

A região central da ilha possui a função de estabelecer o comércio mais intenso, concentrando muitas empresas e também sendo referência em arquitetura histórica, concentra o Mercado Público, um dos lugares mais importantes para o crescimento do comércio na cidade antigamente. Os bairros residenciais são as regiões onde existe boa parte da moradia dos habitantes. Já as praias são regiões onde normalmente concentram-se a parte voltada ao turismo.

Com o mapa de calor a seguir desenvolvido na ferramenta da Microsoft Power BI, com o visual HeatMap, é possível observar a distribuição das acomodações dentro da ilha de Florianópolis:

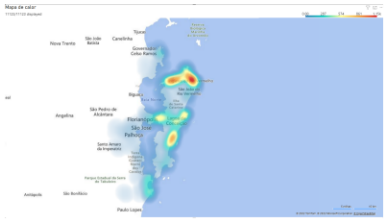


Figure 10. Mapa de calor Airbnb - Florianópolis
Elaborado pelo autor (2022)

No mapa é possível observar a quantidade superior de acomodações disponíveis no norte da ilha, principalmente na região dos Ingleses, onde nos últimos anos vem crescendo muito. Contudo Canasvieiras e Jurerê também apresentam um índice muito alto de acomodações disponíveis. Outra parte da cidade que merece atenção é a parte leste, Lagoa e Campeche se destacam nos bairros com maiores volumes de acomodações disponíveis. Por fim o centro da cidade também ocupa uma boa parte do volume de acomodações disponíveis.

4.2. Dados do Tripadvisor

De acordo com os dados coletados do Tripadvisor, é possível identificar alguns fatores dentro dos estabelecimentos encontrados para a região de Florianópolis. A amostra extraída são de 5.180 restaurantes e 300 atrações que equivalem a 100% dos dados. As atrações podem ser interpretadas como : atividades de turismo, praias, cinema, shoppings, mercados e parques. E os restaurantes englobam também cafés e bares.

4.2.1. Avaliação

Cada restaurante ou atração disponível no site do TripAdvisor possui um espaço designado a avaliações e comentários realizados pelas pessoas que já passaram por ali. O dono do restaurante por sua vez tem um retorno sobre seu serviço, além de que, para os demais consumidores as avaliações publicadas podem ser um fator decisivo para a escolha do local. E a própria prefeitura pode analisar em relação as atrações das cidades modelos de melhorias com base nas avaliações. É gerado um resultado que pode variar entre 1 a 5 estrelas, sendo 1 a nota mais baixa e 5 a mais alta. Para o estudo será utilizado um arredondamento das notas.

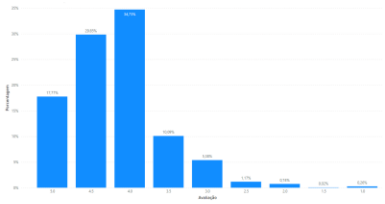


Figure 11. Classificação TripAdvisor
Elaborado pelo autor (2022)

Conforme observa-se a maior parte das avaliações coletadas estão entre 4 e 5, sendo 34,73% para a nota 4, 29,85% para a nota 4,5 e 17,77% para a nota 5, totalizando um percentual de 82,35% nessa faixa de avaliação. Com esses números podemos considerar que grande parte dos restaurantes e atrações possuem uma avaliação média dentro dos padrões acima da nota 4, sendo um valor bem alto. De acordo com os usuários a classificação dos restaurantes de modo geral em Florianópolis é muito boa, um índice alto de gastronomia e boas atrações como diversas praias.

4.2.2. Comentários

Os comentários podem ser realizados pelos usuários da plataforma sobre as atrações e os estabelecimentos disponíveis na cidade. Outros usuários podem se basear nos comentários para frequentar ou não esse lugar, deixando claro como foi sua experiência e se recomenda para outras pessoas.

Uma análise com uma amostra de comentários foi realizada, nesse sentido foram coletados 3.132 comentários em português e 1.033 comentários em inglês para os restaurantes e com isso foi criado uma nuvem das palavras mais utilizadas.

Figure 12. Comentários na língua portuguesa no TripAdvisor
Elaborado pelo autor (2022)

Com esse visual, as palavras mais usadas acabam se tornando maiores. Na imagem para os comentários na língua portuguesa, é possível notar o uso da palavra 'Atendimento' em destaque sendo mencionada 819 vezes, além da palavra 'bom', mencionado 713 vezes. Com esse conjunto de palavras podemos sugerir que um bom atendimento faz bastante parte dos comentários realizados. Além da menção a palavra 'comida' mencionada 609 vezes.

Analisando esses valores e as palavras encontradas, observa-se que é comum o uso das palavras relacionadas a localização, nos comentários. 'Lugar' aparece 391 vezes e 'Local' por sua vez 344 vezes.

Outro fator a ser considerado é o 'Ambiente', mencionado 384 vezes, que acaba tendo um peso grande no conjunto de palavras utilizadas. Sendo assim muitas pessoas que frequentam restaurantes optam por um ambiente agradável.

Os comentários quando realizados, tendem a ser positivos, como no caso do uso das palavras 'Bom', 'Boa', 'Recomendo','Excelente', além de outras palavras que podem ser encontradas.

Outro filtro foi utilizado para determinar palavras da língua inglesa.

Figure 13. Comentários na língua inglesa no TripAdvisor
Elaborado pelo autor (2022)

A nuvem de palavras para essa língua mantém alguns padrões parecidos. a palavra 'place' que pode ser traduzida como 'lugar' aparece 308 vezes sendo a terceira mais utilizada. A mais usada foi 'good', traduzindo é interpretada como 'bom', mencionada 405 vezes. A palavra 'food' apareceu 387 vezes, e nesse caso pode ser considerado que o significado da utilização de 'comida' atrelado a outra palavras positivas tendem a trazer uma faixa grande de comentários positivos em relações aos restaurantes da cidade de Florianópolis.

Nesse estudo existiu uma relação grande entre as palavras usadas nos dois diferentes idiomas, sobretudo nos fatores indicados como lugar, comida e palavras positivas de diversas formas.

4.2.3. Preço

O segmento preço, segue sendo um dos fatores a ser analisado pelos clientes. Um restaurante pode conter várias propostas, alguns optam por cardápios mais em conta e outros são mais refinados. Em Florianópolis existem muitas variedades, tanto de tipos de culinárias como de propostas de espaços arquitetônicos. No conjunto de dados analisado, foi observado uma grande quantidade de restaurantes com a proposta de culinárias brasileira, japonesa, italiana, observa-se também boa quantidade de restaurantes de frutos do mar, pizzas e fast-foods, além de cafés e bares. Foi observado uma grande diferença com relação aos diversos tipos de culinárias.

O TripAdvisor disponibiliza uma faixa de preço diferente para os usuários, utilizando o cifrão como indicador. Não existem definições exatas de valor para cada faixa de preços, mas o senso comum dos usuários da plataforma poderia classifica-las como sendo:

$: comida com preço baixo;

$$: comida com preços médios;

$$$: comida com preços mais altos;

$$$$: comida com preços bem altos, normalmente cozinhas internacionais com chefes renomados.

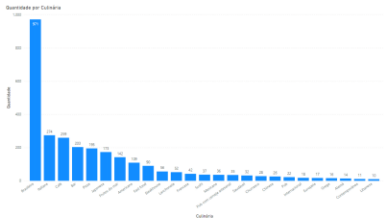


Figure 14. Culinária quantidade no TripAdvisor
Elaborado pelo autor (2022)

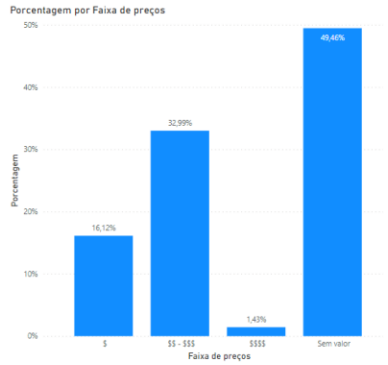


Figure 15. Faixa de preços no TripAdvisor
Elaborado pelo autor (2022)

Conforme visualizamos as diferentes faixas de valores citados anteriormente, percebemos as diferenças nas médias dos valores. É notável a grande quantidade de restaurantes que não possuem o registro de faixas de valores, sendo responsável por uma porcentagem de 49,46%. Grande parte dos restaurantes da ilha de Florianópolis, 32,99% se enquadram na faixa de preços entre valores médios e valores mais altos. É possível observar que 16,12% dos valores são relacionados a um preço baixo, portanto dentro da ilha segundo a análise é mais fácil encontrar restaurantes que tenham o preço mais elevado. Por fim temos a faixa com os valores mais altos em 1,43%, onde se enquadram restaurantes com cozinhas internacionais e chefes renomados, que acaba sendo uma porcentagem bem baixa para uma cidade turística de alto padrão.

A partir dos resultados obtidos foi possível verificar que mesmo com o alto índice de restaurantes sem informações de faixa de preços, uma boa porcentagem está dentro do padrão de preços médios e mais elevados. Portanto podemos considerar que para realizar refeições fora de casa na ilha de Florianópolis, as opções de restaurantes com preços mais elevados serão maiores.

4.2.4. Localização

Em relação a localização das atrações e restaurantes listados no TripAdvisor, conforme mencionado anteriormente a região central da ilha possui a função de estabelecer o comércio mais intenso, portanto concentra uma grande variedade de restaurantes utilizados no dia-a-dia da população,além do Mercado Público e outras atrações históricas. Algumas atrações e restaurantes estão concentrados nas praias, região voltada ao turismo.

Com o mapa de calor a seguir desenvolvido na ferramenta da Microsoft Power BI, com o visual HeatMap, é possível observar a distribuição das atrações dentro da ilha de Florianópolis:

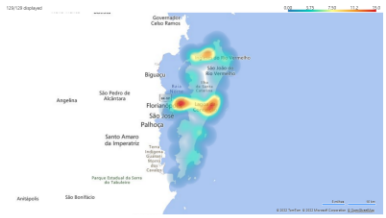


Figure 16. Mapa de calor das atrações no TripAdvisor
Elaborado pelo autor (2022)

É possível observar a quantidade elevada de atrações localizadas no centro da cidade, que devido a sua região histórica e cercada de mercados e centros comerciais possui muitas atrações. A região de Canasvieiras também se destaca, muito pela quantidade de atividades relacionadas ao turismo, já que recebe boa parte de visitantes do nosso país vizinho Argentina. Contudo Lagoa da conceição e parte Leste da ilha também apresentam um índice muito alto de atrações, porém mais distribuídas, a região possui diversas praias e muita prática de esportes ao ar livre, como o surfe, que é destaque e recebe alguns eventos na Praia da Joaquina e na Praia Mole.

Considerando o mapa de calor no mesmo formato, porém agora com o filtro de restaurantes, obtivemos a imagem a seguir:

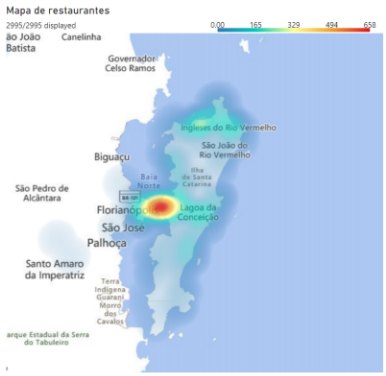


Figure 17. Mapa de calor dos restaurantes no TripAdvisor
Elaborado pelo autor (2022)

Nele nota-se que o centro é o grande foco de restaurantes, bares e cafés disponíveis na ilha. Outras regiões como o Norte da ilha e a parte Leste, também apresentam algumas

manchas de calor. Para entendermos melhor a distribuição central foi realizada outra imagem mais aproximada:

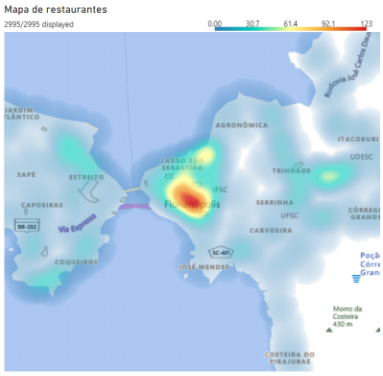


Figure 18. Mapa de calor dos restaurantes da região central no TripAdvisor
Elaborado pelo autor (2022)

Conforme aproximamos, é perceptível que a parte continental também influência na grande relação de restaurantes, assim como a parte da Trindade, arredores da UFSC e UDESC. Regiões onde são famosas por abrigar os mais diversos tipos de culinárias. Coqueiros por ser uma via voltada a gastronomia e Trindade e Santa Mônica por abrigarem muitos estudantes universitários. Já o Centro e a Beira-mar abrigram restaurantes do dia a dia, restaurantes refinados, bares e cafés de diversos gostos.

4.3. Correlação entre os dados do Airbnb e TripAdvisor

Ao falarmos em análise de dados, é necessário entender qual é a associação entre as variáveis. A análise de correlação é uma forma descritiva que mede se há e qual o grau de dependência entre variáveis, ou seja, o quanto uma variável interfere em outra, sendo que essa relação de dependência pode ou não ser causal. Essa medida de grau de relação é medida através dos coeficientes. O coeficiente de correlação pode variar em termos de valor de -1 a +1, quanto maior for o valor absoluto do coeficiente, mais forte será a relação entre as variáveis. Segundo [Cohen 1992] os tamanhos de efeito podem se enquadrar em

r = —0,10— - correlação fraca.

r = —0,30— - correlação moderada.

r = —0,50— - correlação forte.

Para a realização da correlação entre as variáveis foi necessário importar as bibliotecas pandas e seaborn. O pacote pandas é uma ferramenta na manipulação de dados, tabelas e *dataframes*. Já o pacote seaborn é excelente para a criação de visualizações gráficas, principalmente em casos de mapeamentos estatísticos.

As duas bases de dados foram transformadas em *dataframes*, através da função read_csv() da biblioteca pandas. Após a transformação foi realizada a função corr() também da biblioteca pandas, onde ela realiza a correlação entre os atributos do dataset. A função possui uma configuração para o uso de diversos métodos de correlação, os métodos utilizados foram Pearson, Kendall, Spearman. O ambiente para utilização do código foi o Google Colaboratory.

Primeiro os dados serão analisados individualmente, começando com os referentes ao Airbnb:

```
df.corr(method='pearson')
```

	latitude	longitude	quantidade_hospedes	preco	avaliacao
latitude	1.000000	0.552864	0.161014	0.067139	-0.051059
longitude	0.552864	1.000000	0.107088	-0.019374	-0.038399
quantidade_hospedes	0.161014	0.107088	1.000000	0.273736	-0.045563
preco	0.067139	-0.019374	0.273736	1.000000	-0.154777
avaliacao	-0.051059	-0.038399	-0.045563	-0.154777	1.000000

Figure 19. Correlação dos dados do Airbnb utilizando o método de Pearson
Elaborado pelo autor (2022)

```
df.corr(method='kendall')
```

	latitude	longitude	quantidade_hospedes	preco	avaliacao
latitude	1.000000	0.386624	0.145968	0.098678	-0.046289
longitude	0.386624	1.000000	0.108903	0.041411	-0.032477
quantidade_hospedes	0.145968	0.108903	1.000000	0.467015	-0.033627
preco	0.098678	0.041411	0.467015	1.000000	-0.153627
avaliacao	-0.046289	-0.032477	-0.033627	-0.153627	1.000000

Figure 20. Correlação dos dados do Airbnb utilizando o método de Kendall
Elaborado pelo autor (2022)

```
df.corr(method='spearman')
```

	latitude	longitude	quantidade_hospedes	preco	avaliacao
latitude	1.000000	0.558787	0.203314	0.146872	-0.058814
longitude	0.558787	1.000000	0.151413	0.062783	-0.041344
quantidade_hospedes	0.203314	0.151413	1.000000	0.604938	-0.040082
preco	0.146872	0.062783	0.604938	1.000000	-0.196255
avaliacao	-0.058814	-0.041344	-0.040082	-0.196255	1.000000

Figure 21. Correlação dos dados do Airbnb utilizando o método de Spearman
Elaborado pelo autor (2022)

Utilizando a biblioteca do seaborn é possível visualizar melhor a força de cada propriedade:

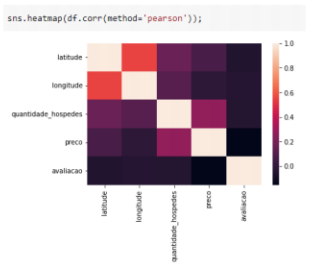


Figure 22. Correlação dos dados do Airbnb com Heatmap utilizando o método de Pearson
Elaborado pelo autor (2022)

A partir da análise dos dados do Airbnb, podemos considerar que com a utilização da metodologia de Pearson, os coeficientes com correlação positiva forte identificados foram latitude e longitude em 0,55. Para o caso de preço e quantidade de hóspedes houve uma correlação positiva moderada em 0,27. Como relação negativa fraca temos o caso de preço e avaliação onde o valor ficou em -0,15.

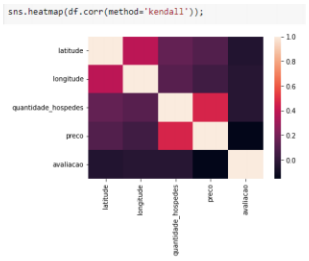


Figure 23. Correlação dos dados do Airbnb com Heatmap utilizando o método de Kendall

Elaborado pelo autor (2022)

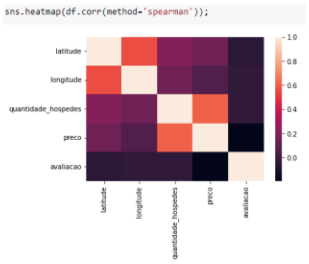


Figure 24. Correlação dos dados do Airbnb com Heatmap utilizando o método de Spearman

Elaborado pelo autor (2022)

Para a metodologia de Kendall, latitude e longitude seguem com correlações positivas entretanto de forma mais moderada em 0,38. Já quantidade de hóspedes e preço passam a ter uma correlação moderada com um grau mais forte em 0,46.

Spearman e sua metodologia nos mostra que semelhante a Pearson a correlação segue positiva e forte para latitude e longitude com grau de 0,55. É interessante notar que o grau de correlação positiva entre a quantidade de hóspedes e o preço sobe para 0,60.

Seguindo agora com o conjunto de dados do TripAdvisor, foi realizada outra análise individual para entendimento das relações das variáveis entre si. Os métodos e abordagens foram as mesmas das etapas anteriores:

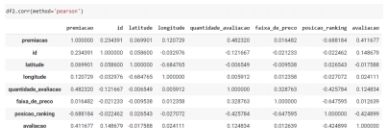
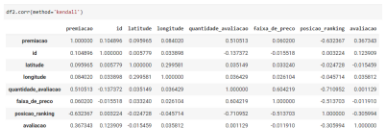


Figure 25. Correlação dos dados do TripAdvisor utilizando o método de Pearson

Elaborado pelo autor (2022)

Após o uso da biblioteca pandas, novamente foi usado a biblioteca seaborn para trazer uma melhor visualização:

Na análise dos dados do Tripadvisor, podemos considerar que com a utilização da metodologia de Pearson os atributos de premiação e quantidade de avaliação chegaram a um coeficiente positivo e moderado, premiação e avaliação também obtiveram correlação positiva de forma moderada na faixa de 0,40. Para esse caso latitude e longitude tiveram



	premiacao	id	latitude	longitude	quantidade_avaliacao	faixa_de_preco	posicao_ranking	avaliacao
premiacao	1.000000	0.104896	0.095965	0.084020	0.510513	0.060200	-0.632367	0.367343
id	0.104896	1.000000	0.005779	0.033898	-0.137372	-0.015518	0.003224	0.123909
latitude	0.095965	0.005779	1.000000	0.299581	0.035149	0.033240	-0.024728	-0.015459
longitude	0.084020	0.033898	0.299581	1.000000	0.056429	0.026104	-0.045714	0.035812
quantidade_avaliacao	0.510513	-0.137372	0.035149	0.056429	1.000000	0.604219	-0.710952	0.001129
faixa_de_preco	0.060200	-0.015518	0.033240	0.026104	0.604219	1.000000	-0.513703	-0.011910
posicao_ranking	-0.632367	0.003224	-0.024728	-0.045714	-0.710952	-0.513703	1.000000	-0.305994
avaliacao	0.367343	0.123909	-0.015459	0.035812	0.001129	-0.011910	-0.305994	1.000000

Figure 26. Correlação dos dados do TripAdvisor utilizando o método de Kendall

Elaborado pelo autor (2022)

	premiacao	id	latitude	longitude	quantidade_avaliacao	faixa_de_preco	posicao_ranking	avaliacao
premiacao	1.000000	0.132504	0.131993	0.113697	0.659026	0.073333	-0.791467	0.422994
id	0.132504	1.000000	0.009982	0.052887	-0.193899	-0.024593	0.008330	0.167747
latitude	0.131993	0.009982	1.000000	0.442533	0.050293	0.041663	-0.037609	-0.021192
longitude	0.113697	0.052887	0.442533	1.000000	0.052657	0.033003	-0.069035	0.048432
quantidade_avaliacao	0.659026	-0.193899	0.050293	0.052657	1.000000	0.730855	-0.852872	-0.001213
faixa_de_preco	0.073333	-0.024593	0.041663	0.033003	0.730855	1.000000	-0.640934	-0.014605
posicao_ranking	-0.791467	0.008330	-0.037609	-0.069035	-0.852872	-0.640934	1.000000	-0.396221
avaliacao	0.422994	0.167747	-0.021192	0.048432	-0.001213	-0.014605	-0.396221	1.000000

Figure 27. Correlação dos dados do TripAdvisor utilizando o método de Spearman

Elaborado pelo autor (2022)

Figure 28. Correlação dos dados do TripAdvisor com Heatmap utilizando o método de Pearson

Elaborado pelo autor (2022)

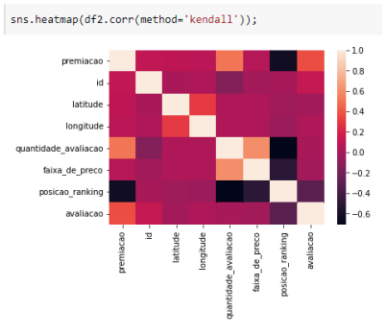


Figure 29. Correlação dos dados do TripAdvisor com Heatmap utilizando o método de Kendall

Elaborado pelo autor (2022)

as correlações de forma negativa forte em -0,68. Quantidade de avaliações e a faixa de preço obtiveram uma correlação de 0,32, positiva e moderada. Posição do ranking obteve uma correlação negativa forte tanto com premiação como com faixa de preços.

Para a metodologia de Kendall, o padrão segue para premiação e as notas de avaliação, entretanto premiação e quantidade de avaliações a correlação se torna positiva

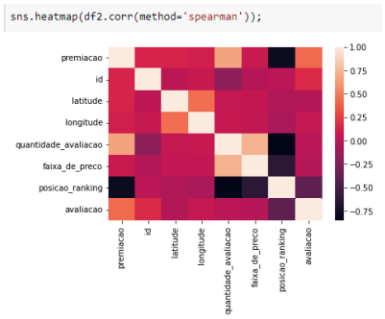


Figure 30. Correlação dos dados do TripAdvisor com Heatmap utilizando o método de Spearman

Elaborado pelo autor (2022)

mais forte. Latitude e longitude apresentam uma correlação positiva moderada. A quantidade de avaliações possui uma correlação positiva forte com a faixa de preço, chegando a 0,60 e negativa forte com posição no ranking em -0,71.

Spearman e sua metodologia nos mostra que os coeficientes com correlação positiva forte identificados foram: premiação e quantidade de avaliações onde o valor se estabeleceu em 0,65. Premiação também teve uma correlação positiva moderada com as notas de avaliação, em 0,42. Contudo para premiação e posição do ranking houve uma correlação negativa forte de -0,79. Latitude e longitude seguiram padrões parecidos com o do Airbnb onde a correlação entre eles foi positiva e moderada. A faixa de preço e a quantidade de avaliações tiveram uma correlação positiva muito forte, sendo 0,73. Posição do ranking em relação a quantidade de avaliações e faixa de preços obteve uma correlação negativa e forte.

Portanto nesta etapa seguimos com a correlação entre as duas bases de dados, Airbnb e Tripadvisor, através da biblioteca pandas, concatenando os dois *dataframes* gerados. Para o eixo horizontal está a legenda dos campos pertencentes ao Airbnb e para o eixo vertical está a legenda dos campos pertencentes ao TripAdvisor:

```
pd.concat([df, df2], axis=1, keys=['df', 'df2']).corr(method='pearson').loc['df2', 'df']
```

	latitude	longitude	quantidade_hospedes	preco	avaliacao
premiacao	-0.169627	-0.110177	-0.156733	-0.178402	0.086175
id	0.011409	0.015089	0.029681	-0.018940	0.025946
latitude	0.003090	-0.020971	0.030295	0.025310	-0.024070
longitude	-0.014046	0.005862	-0.011566	-0.015355	0.019559
quantidade_avaliacao	-0.026168	-0.017591	-0.053951	0.020720	0.002739
faixa_de_preco	-0.004198	0.016352	0.015727	0.090105	-0.029313
posicao_ranking	-0.000238	-0.022146	-0.034373	-0.121155	0.037563
avaliacao	-0.002090	0.007302	-0.001865	0.020310	-0.007591

Figure 31. Correlação dos dados do Airbnb e TripAdvisor utilizando o método de Pearson

Elaborado pelo autor (2022)

Seguindo o padrão das análises, foi usado a biblioteca seaborn para trazer uma melhor visualização:

A partir da concatenação entre os dois *dataframes* propostos para a análise, no eixo Y os atributos pertencentes ao Tripadvisor e no eixo X os atributos pertencentes ao Airbnb, pode ser observado que não houveram correlações positivas e nem negativas com grau de forte e moderada. Os maiores valores positivos foram entre a correlação

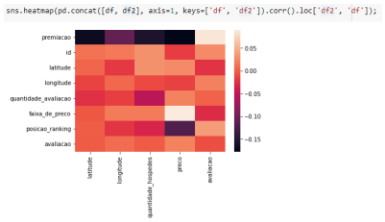


Figure 32. Correlação dos dados do Airbnb e TripAdvisor utilizando Heatmap e o método de Pearson

Elaborado pelo autor (2022)

da premiação do tripadvisor, com as notas das avaliações do airbnb. Além de faixa de preço dos restaurantes, bares e cafés coletados pelo Tripadvisor estarem correlacionadas positivamente porém com uma força fraca com os preços do Airbnb.

5. Considerações Finais

Conforme o número de pessoas utilizando as plataformas onlines de hospedagens cresceu, fez com que esses sites como Airbnb se tornassem um ótimo lugar para fazer coleta de dados e aplicar em pesquisas. Sites relacionados ao turismo surgiram e trouxeram aos viajantes a possibilidade de analisarem os destinos através dos comentários e indicações de outras pessoas da rede. Neste trabalho foram realizadas algumas implementações, a extração das acomodações e os detalhes relacionados a elas disponíveis no Airbnb, a extração de dados relacionados aos restaurantes e acomodações disponibilizadas no Tripadvisor, ambas para a cidade de Florianópolis. Foram utilizados dois Web Scrappers focados em varrer os sites tanto do Airbnb quanto do Tripadvisor e armazenar essas extrações em arquivos de formato CSV. Conforme foram gerados os arquivos, fez-se necessário um tratamento e a limpeza dos dados para uma análise com maior precisão e qualidade. Sendo assim os dados ficaram disponíveis para serem usados a fim de encontrar os conjuntos de indicadores.

Foram desenvolvidas as fases de análises e comparações entre os atributos extraídos e coletados, identificando os atributos com maior correlação através das matrizes de correlação de Pearson, Kendall e Spearman junto com as médias de preços empregadas na cidade de Florianópolis, outros fatores levados em consideração foram as avaliações e comentários deixados pelos usuários. Além da fonte de dados do Airbnb foi acrescentada a fonte de dados do TripAdvisor reunindo dados do contexto de restaurantes e atrações envolvendo a cidade de Florianópolis.

Os objetivos específicos foram atingidos. Através da obtenção dos dados e das análises, foi possível observar os valores empregados e as faixas de preço tanto para as acomodações quanto para os restaurantes e atrações da cidade de Florianópolis. As distribuições geográficas foram apresentadas através de mapas de calor, demonstrando que para os bairros do Norte da Ilha existem muito mais ofertas de acomodações que outras partes da ilha, contudo a parte Central e o Leste da Ilha também apresentam um bom volume. Já em relação as atrações o Centro e o Leste da ilha possuem um índice muito superior ao restante. Ao falarmos de restaurantes, bares e cafés o Centro contém uma larga dominância em opções. Os comentários e avaliações analisados trouxeram algumas visões de que de acordo com a amostra coletada, grande parte dos visitantes e

da população, avalia bem as acomodações, os restaurantes e as atrações propostas por Florianópolis, tendo visto que os comentários positivos se destacaram.

Para as correlações observou-se que houveram diferenças nos tamanhos de efeito sugerido por [Cohen 1992] para os dois conjuntos de dados com as diferentes metodologias empregadas. Tanto Airbnb como o Tripadvisor tiveram correlações fortes tanto positivamente quanto negativamente em algumas variáveis individuais. Quando os dois conjuntos foram correlacionados, analisou-se que as variáveis de faixa de preço do Tripadvisor e o preço do Airbnb obtiveram forças de correlação fracas.

Para os trabalhos futuros, existe a possibilidade de desenvolver e aprimorar as análises propostas com novas fontes de dados, pode-se trazer outros atributos pertencentes ao Airbnb, como as comodidades das acomodações e as influências nos valores praticados. Outros atributos do Tripadvisor também podem ser explorados, como a possibilidade de usar os dados relacionados aos hotéis presentes na plataforma. Novas métricas e indicadores podem ser implementados para o estudo para que integrem as correlações entre os diversos aplicativos de turismo para a cidade de Florianópolis. Um aprofundamento da análise das distribuições usando estatísticas descritivas que descrevam a tendência central (média e mediana) e dispersão (variância e desvio padrão) de dados com valores numéricos, que acabam adicionando uma camada de detalhes e podem ser utilizadas para fazer comparações com outros conjuntos de dados.

References

Ahuja, M., Singh, J., and nica, V. (2014). Web crawler: Extracting the web data. *International Journal of Computer Trends and Technology*, 13:132–137.

Calabrese, B. (2018). *Data Cleaning*.

Cohen, J. (1992). Quantitative methods in psychology. *Psychological Bulletin*, page 155.

Copelli, M. (2022). airbnb-analytics.

Cunningham, P., Smyth, B., Wu, G., and Greene, D. (2010). Does tripadvisor makes hotels better?

Duong, T. (2021). Airbnb scraper.

Guttentag, D. (2013). Airbnb: Disruptive innovation and the rise of an informal tourism accommodation sector. *Current Issues in Tourism*, 18:1–26.

Guttentag, D., Smith, S., Potwarka, L., and Havitz, M. (2017). Why tourists choose airbnb: A motivation-based segmentation study. *Journal of Travel Research*, 57:004728751769698.

Ilyas, I. and Chu, X. (2019). *Data Cleaning*.

Ioannides, D., Röslmaier, M., and van der Zee, E. (2019). Airbnb as an instigator of 'tourism bubble' expansion in utrecht's lombok neighbourhood. *Tourism Geographies*, 21(5):822–840.

Minca, C. and Roelofsen, M. (2022). *Becoming Airbnbeings: on datafication and the quantified Self in tourism*, pages 95–116.

Mirtaheri, S., Dinçktürk, M., Hooshmand, S., Bochmann, G., Jourdan, G.-V., and Onut, I.-V. (2014). A brief history of web crawlers.

Nemeslaki, A. and Pocsarovszky, K. (2011). Web crawler research methodology.

Oskam, J. and Boswijk, A. (2015). Airbnb: The future of networked hospitality businesses. *Journal of Tourism Futures*, 2:22–42.

Peres, C. and Paladini, E. (2021). Exploring the attributes of hotel service quality in florianópolis-sc, brazil: An analysis of tripadvisor reviews. *Cogent Business and Management*, 8:1926211.

Tripadvisor (2022). About tripadvisor.

Yoo, K.-H., Sigala, M., and Gretzel, U. (2016). *Exploring TripAdvisor*, pages 239–255.

Zhu, G. and Kubickova, M. (2022). From homeowner to airbnb host: The role of trust and perceived value. *Journal of Quality Assurance in Hospitality and Tourism*, pages 1–23.

Álvarez García, J., Durán-Sánchez, A., Rama, D., and Simonetti, B. (2020). Big data and tourism research: Measuring research impact. *Quality and Quantity*.