Desafio Seazone análise de dados

Gráficos, dados e análises para o processo seletivo Seazone de Analista de dados Jr.

Rômulo Peixoto

Florianópolis, 2022

Introdução

No dia 11/01/2022 a Seazone, disponibilizou aos seus participantes o desafio prático para o processo seletivo de Analista de dados júnior. Os dados foram disponibilizados em dois arquivos .csv, que podem ser vizualizados utilizando a barra lateral do aplicativo. O objetivo do desafio é analisar os dados de ocupação e preço para responder as seguintes perguntas:

- *Ordenar os bairros em ordem crescente de anúncios.*
- *Ordenar os bairros em ordem crescente de faturamento.*
- *Encontrar correlações entre as características do anúncio e seu faturamento. Citar e explicar.*
- *Medir a antecedência média de reservas. Checar se existe diferença para finais de semana.*

Os códigos estão disponibilziados no repositório Desafio Seazone no GitHub e também na forma de um dashboard que foi criado utilizando o framework Streamlit, mais detalhes no README.md do repositório sobre como acessar o aplicativo.

Metodologia

As análises foram executadas utilizando Python 3.8, com a ferramenta gráfica Seaborn 0.11.1 juntamente com a matplotlib 3.4.2, análise dos dados foi conduzida através da biblioteca pandas 1.3.1 e para criação do dashboard foi utilizado o framework streamlit 1.4.0. As IDE's utilizadas para escrever os códigos foram o Jupyter Notebooks 6.4.0 e o Visual Studio Code 1.63.2 rodando no sistema operacional Ubuntu 20.04.

Para determinação dos faturamentos de cada anúncio, foram utilizados apenas os imóveis que possuem data de agendamento, devido ao formato de locação utilizado pelo Airbnb, onde as reservas são pagas na hora, entretanto o pagamento aos proprietários é feito em até 24 horas após o primeiro check-in dos inquilinos.

Dados

Nesta página você encontra as fontes de dados utilizadas para realizar as análises.

Foram disponibilizados dois conjuntos de dados:

- *desafio_details.csv*
- *desafio_priceav.csv*

Com as suas explicações abaixo:

desafio_details.csv

Este conjunto de dados apresenta todas as características do seus anúncios divididos nas colunas:

- listing_id: Identificador de um anúncio.
- suburb: Bairro do listing.
- star_rating: Nota 1-5 do anúncio.
- is_superhost: Booleano que indica se é superhost ou não.
- number_of_bedrooms: Quantidade de quartos do anúncio.
- number_of_reviews: Quantidade de comentários do anúncio.
- ad_name: Título do anúncio.
- number_of_bathrooms: Número de banheiros do anúncio.

4.691 entradas | 8 colunas

desafio_priceav.csv

Este conjunto de dados apresenta os dados de ocupação e preço das diárias.

- listing_id: Identificador de um anúncio.
- price_string: Preço ofertado.
- occupied: Booleano de ocupação. 1 significa livre e 0 ocupado.
- date: Data a ser alugada.
- booked_on: Data quando “date” foi alugado. Null caso ainda esteja available.

354.520 entradas | 5 colunas

desafio_df

Um terceiro dataframe foi criado, é apenas a junção dos outros dois conjuntos de dados já citados, a sua junção foi feita através dos identificadores iguais de cada linha. Algumas operações foram realizadas para garantir o bom funcionamento desse conjunto, eliminação das linhas duplicadas, preenchimento dos valores númericios faltantes por 0 e modificação dos valores faltantes da coluna 'booked_on' de 'blank' para string vazia.

Por último, as colunas 'booked_on' e 'date' foram convertidas para dados do tipo *datetime[ns]* para futuros cálculos.

289.919 entradas | 13 colunas

O conjunto de dados utilizado para condução da análise 1 foi apenas o desafio_details.csv, as 3 últimas análises foram conduzidas através do desafio_df. Outras formas de organização dos dados foram feitas através do método groupby da biblioteca pandas.

Questão 1:

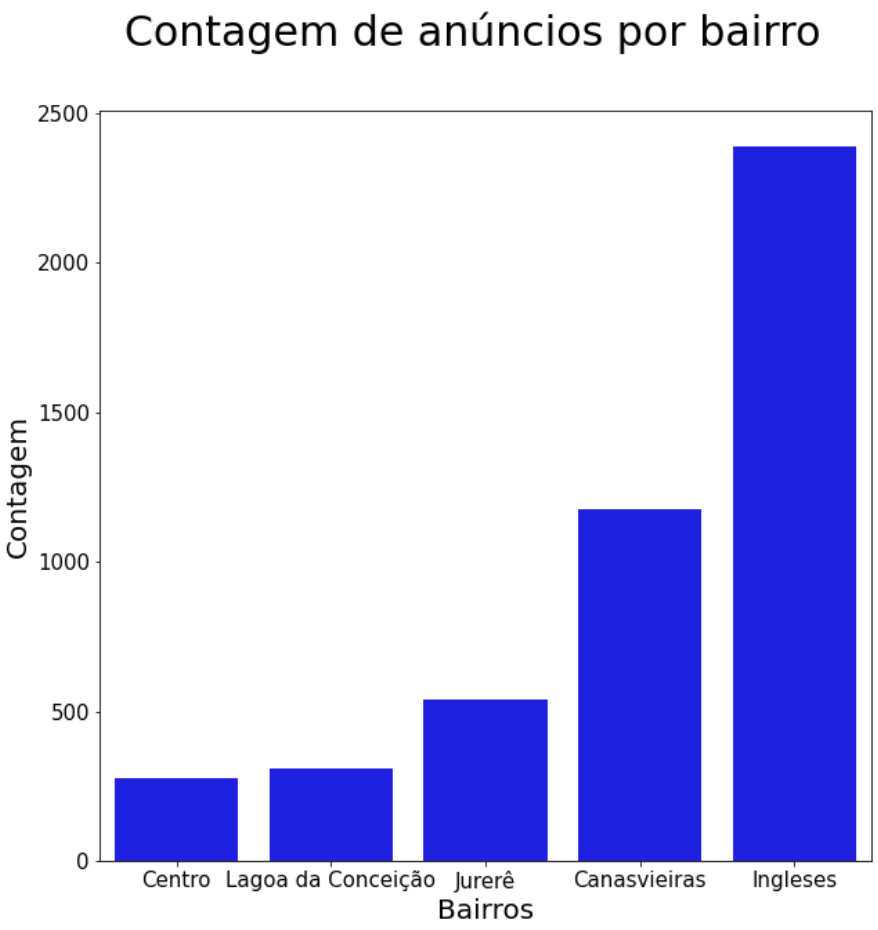
Ordenar os bairros em ordem crescente de anúncios.

Para responder a pergunta os dados foram organizados através do método groupby de modo a contar o número dos identificadores de cada bairro. O gráfico abaixo mostra os

resultados obtidos, sendo o bairro dos Ingleses com o maior número de anúncios, seguido por Canasvieiras.

Essa tendência acontece principalmente devido à busca por imóveis, principalmente no bairro dos Ingleses. Os motivos que levam as pessoas a buscarem cada vez mais imóveis na região acontece pela independência que os bairros vem conquistando ao longo dos anos, devido a sua grande distância do centro da cidade, fator que também leva à preços de aluguéis mais acessíveis.

De acordo com a reportagem da Gaúcha Zero Hora e da ndmais, pessoas que escolheram alugar seus imóveis na região descreveram que a comodidade de encontrar tudo o que precisam no bairro foi o que mais pesou na decisão. Os mesmos fatores se aplicam ao bairro de Canasvieiras, entretanto, segundo moradores, o que diferencia a escolha do bairro é o índice de violência registrado em cada uma das vizinhanças, sendo Canas o mais pacífico entre os dois.



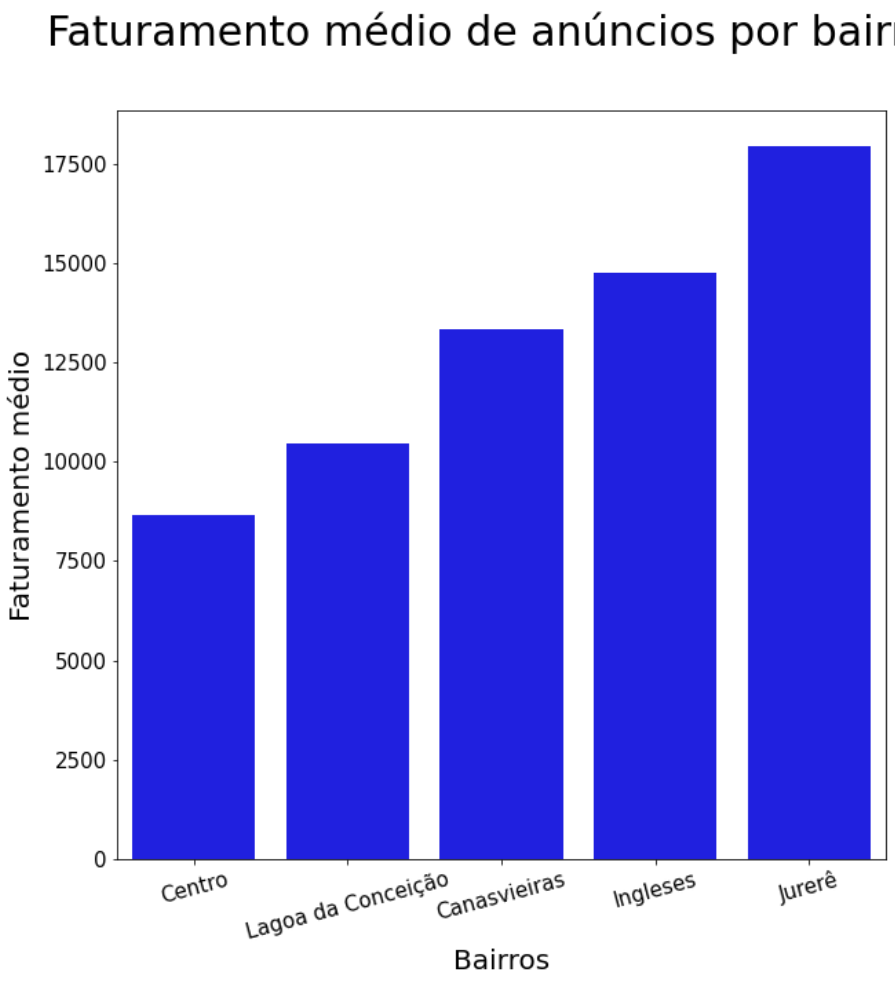
Questão 2

Ordenar os bairros em ordem crescente de faturamento.

Para responder essa pergunta, 4 operações foram utilizadas, a primeira agrupando as linhas que conforme a data de agendamento, em seguida foram filtradas as linhas que não possuiam data de agendamento, depois os valores de diária para cada identificador foram somados e então agrupados conforme os bairros e calculando a média de faturamento.

O bairro com o maior faturamento médio registrado é Jurerê, seguido de Ingleses muito próximo à Canasvieiras. Como dito na questão anterior, os fatores que levam os Ingleses a ser o bairro mais procurado, tornam o seu retorno maior. Porém outro fator que conta ao considerar o faturamento médio de cada bairro é o seu preço médio por metro quadrado.

Analisando os valores disponibilizados pelo sistema COFECI-CRECI em outubro de 2020, Jurerê e Jurerê internacional possuem os valores de R$ 36,67/m2 e R$ 32,79/m2 respectivamente. Entre os valores mostrados se encontra a Lagoa da Conceição com R$ 36,26/m2, mas devido a baixa procura seu faturamento não supera bairros mais acessíveis como Canasvieiras e Ingleses.

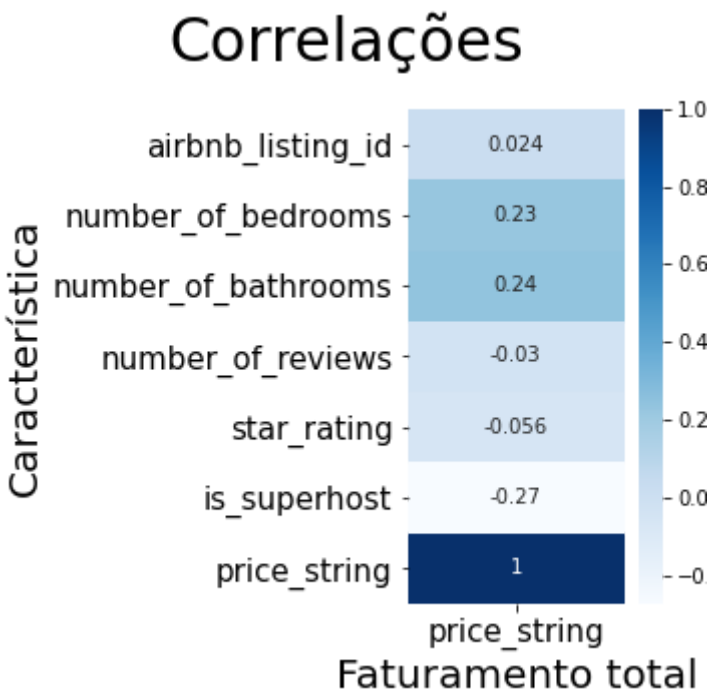


Questão 3

Encontrar correlações entre as características do anúncio e seu faturamento. Citar e explicar.

A resposta dessa questão é extensa e requer várias manipulações dos dados para ser obtida, começando por agrupar os dados conforme as características do anúncio, depois calcular o faturamento total para cada um deles. Depois foi feita a codificação dos bairros para que fosse possível calcular a correlação da característica com o faturamento.

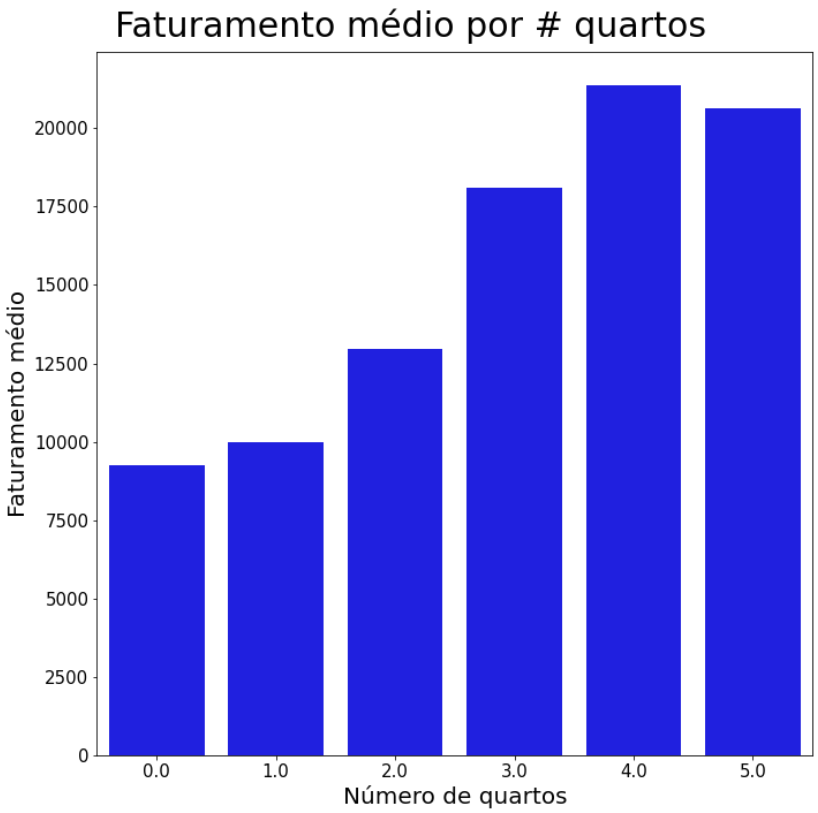
Depois das manipulações foi calculada a matrix de correlação entre as características e isolada apenas a coluna de faturamento, o resultado pode ser observado abaixo:

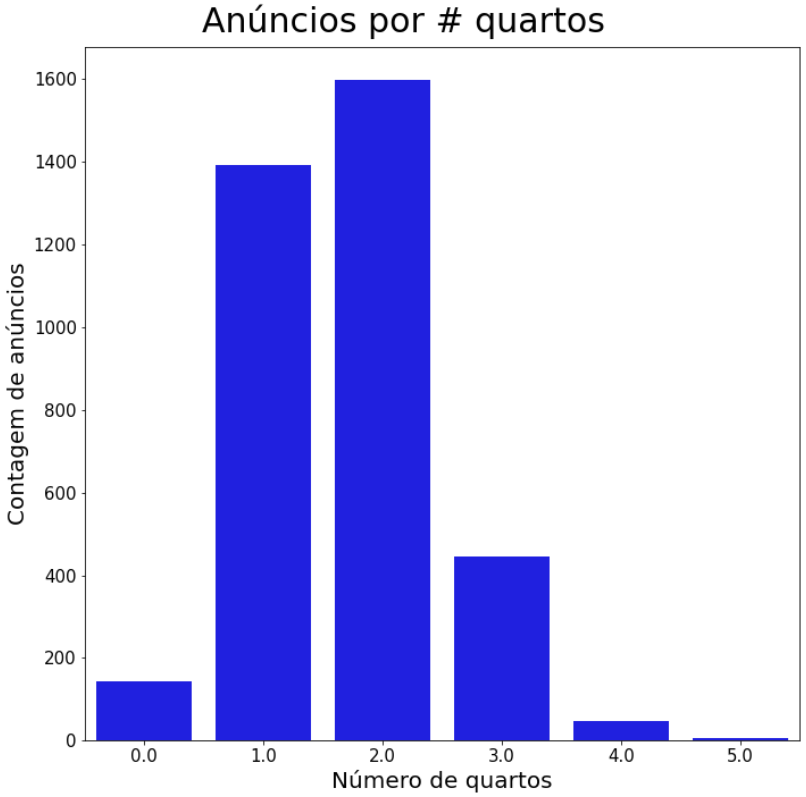


Dado o resultado, é possível perceber uma correlação mais pronunciada entre o número de quartos e de banheiros com o aumento de faturamento enquanto o selo de *superhost* com uma correlação negativa em relação ao faturamento total. As características serão analisadas mais a fundo.

Número de quartos

Para analisar o número de quartos, foi feito dois gráficos que mostram o faturamento médio conforme o número de quartos e um com uma contagem de anúncios conforme o número de quartos.



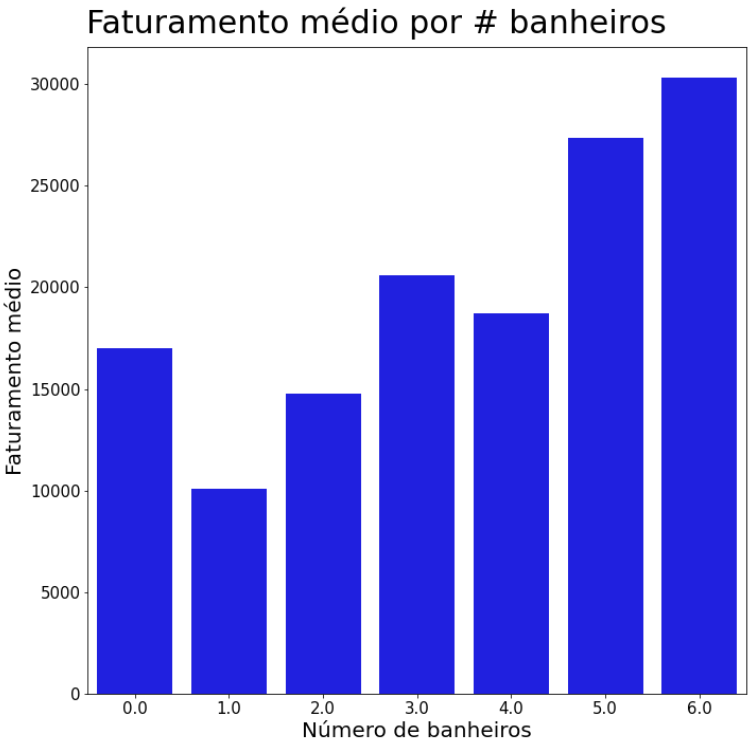


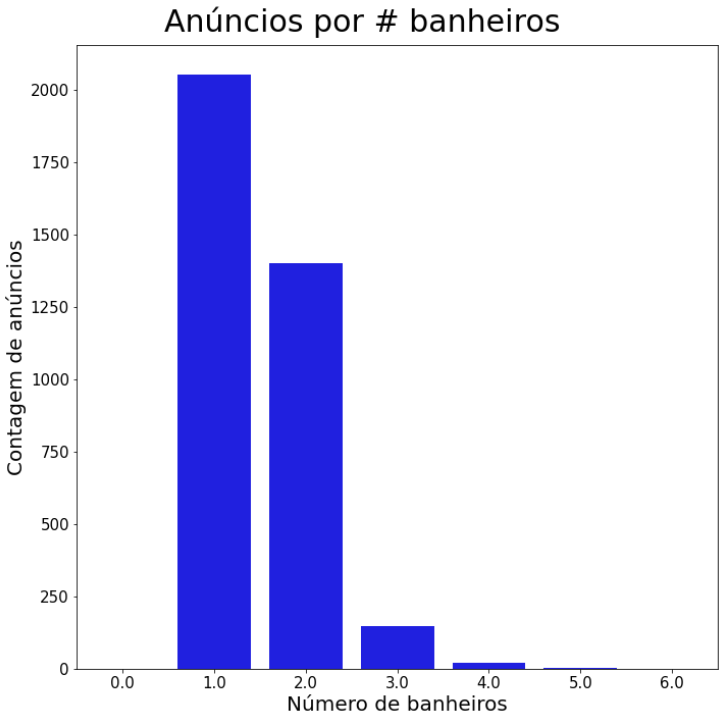
Esses dados mostram que mesmo com uma oferta maior no número de anúncios com 1 e 2 quartos o grosso do faturamento observado ainda se dá pela locação de anúncios com 3 e 4 quartos. A explicação aqui é de que as pessoas que costumam fazer viagens com aluguel

de temporada não estão sozinhas, seja com amigos ou família, as locações ocorrem geralmente para um grupo de pessoas conviver no mesmo ambiente.

Número de Banheiros

A análise do número de banheiros foi da mesma maneira que o número de quartos.





Seguindo a mesma tendência encontrada com o número de quartos, o número de banheiros também possui um maior faturamento para anúncios que possuem maior capacidade de

abrigar pessoas, justificando a correlação positiva das duas características com o faturamento.

Superhost

Para entender como o faturamento se relaciona com o emblema de *superhost* o método groupby possui algumas dificuldades para agrupar e calcular valores booleanos, então uma função foi criada para varrer o agrupamento e disponibilizar os valores de interesse, o resultado pode ser observado abaixo.

	Superhost	Não *Superhost*
Faturamento médio	9626.37	13883.80
Faturamento total	12013711.00	33085098.00
Número total	1248	2383

Dados os valores encontrados é possível analisar algumas características do conjunto de dados, o primeiro é que o número de *superhosts* é um pouco maior que a metade dos não *superhosts*, enquanto o faturamento dos *superhosts* fica abaixo da metade do faturamento total dos outros anuciantes, a diferença mostra que não existe uma tendência realmente vantajosa em se tornar superhost, apesar das garantias de estadia dos inquilinos.

Outra observação que vale ser feita é de que mesmo com o selo, não é garantia de que o faturamento será realmente maior por uma prerência dos hóspedes por contas verificadas e com maiores garantias de uma boa estadia.

Questão 4

Medir a antecedência média de reservas. Checar se existe diferença para finais de semana.

Para esta pergunta, foram selecionados os anúncios que possuíam data de agendamento e foram filtradas as datas de locação selecionando as datas que tem o menor valor por agendamento, depois foram filtradas as datas que não foram agendadas com antecedência para calcular a média e por último, utilizando as datas de locação, foi criada uma coluna com os dias da semana de forma numérica, sendo 0 segunda-feira e 6 domingo.

O resultado das médias pode ser vizualizado abaixo, devido a proximidade dos valores, não é possível extrair muitas informações, além de que a média de antecedência é de 30 dias, para ver mais informações da amostra de dados foi realizada o boxplot para análise dos valores encontrados.

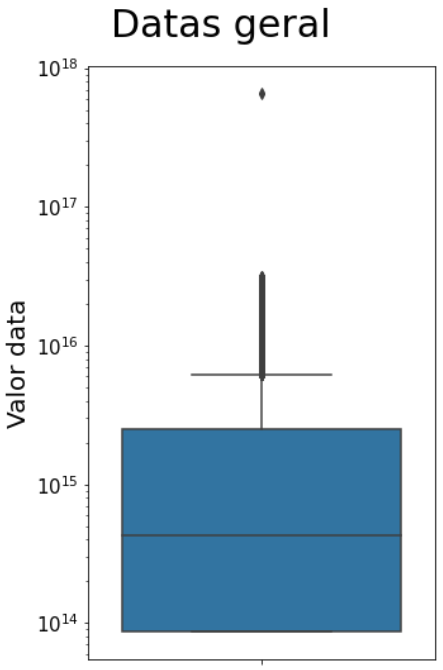
Os valores de data foram convertidos para números em nanosegundos e a escala do gráfico convertida para uma escala logarítmica devido a ordem de magnitude dos valores. É

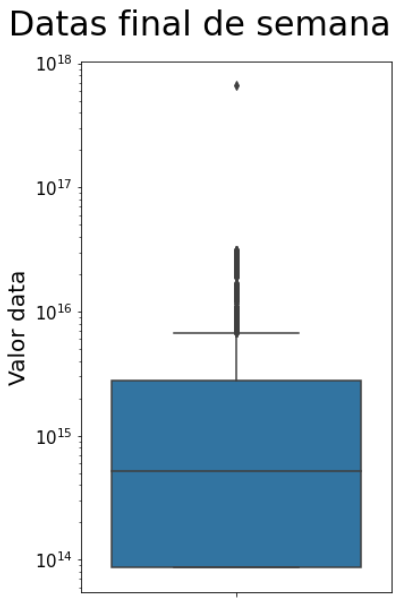
possível perceber que os valores para finais de semana possuem uma quantidade menor de valores atípicos do que locações no geral. Justificando também o aumento de valores dentro da amplitude interquantil observada.

Outros valores que contribuem para a análise são a moda e a mediana, a partir da moda é possível perceber que normalmente as pessoas fazem seus agendamentos com um dia de antecedência, e não trinta dias como previsto em média, isso mostra como os valores que possuem muita antecedência estão acima do normal e também como as reservas são feitas em cima da hora.

Já a mediana mostra que pelo menos 50 porcento dos valores encontrados foram feitos com menos de uma semana de antecedência, o valor corrobora a tese de que das reservas antecedentes a data de locação foram realizadas com um planejamento de curto prazo.

	Todas as datas	Para finais de semana
Média	30 days 11:23:11.511035653	32 days 15:15:34.468085106
Moda	1 days 00:00:00	1 days 00:00:00
Mediana	5 days 00:00:00	6 days 00:00:00





Fontes

https://gauchazh.clicrbs.com.br/geral/noticia/2017/01/ingleses-e-bairro-de-florianopolis-mais-procurado-para-compra-de-imoveis-canasvieiras-para-aluguel-9252582.html

https://ndmais.com.br/economia-sc/bairro-dos-ingleses-uma-cidade-dentro-de-florianopolis/

https://www.creci-sc.gov.br/p/noticias/fipezap-analisa-comportamento-dos-precos-medios-do-aluguel-em-outubro/1402/

Feedback do processo

Eu gostei muito do processo, encontrei desafios ao longo do processo de resposta das perguntas que me deixaram animado para dar o próximo passo em direção para encontrar mais informações escondidas dentro dos dados disponibilizados.

Outro fator que também me deixou empolgado e me motivou a participar foi a liberdade de criação que o processo me permitiu, pude encontrar outras formas de vizualizar os dados que fizeram sentido durante meu progresso.

Consegui aprender muitas coisas ao longo do caminho e também mostrar o que sou capaz de fazer caso seja escolhido para a oportunidade e realizar estas atividades em equipe para aprender ainda mais.