Análise Exploratória dos Dados

Inicialmente, começamos traçando um gráfico representando os valores na coluna de preço. Observando a Figura 1, identificamos a presença de valores discrepantes, o que nos levou a realizar uma depuração dos dados, removendo as entradas com preços superiores a 500 dólares e inferiores a 30 dólares.

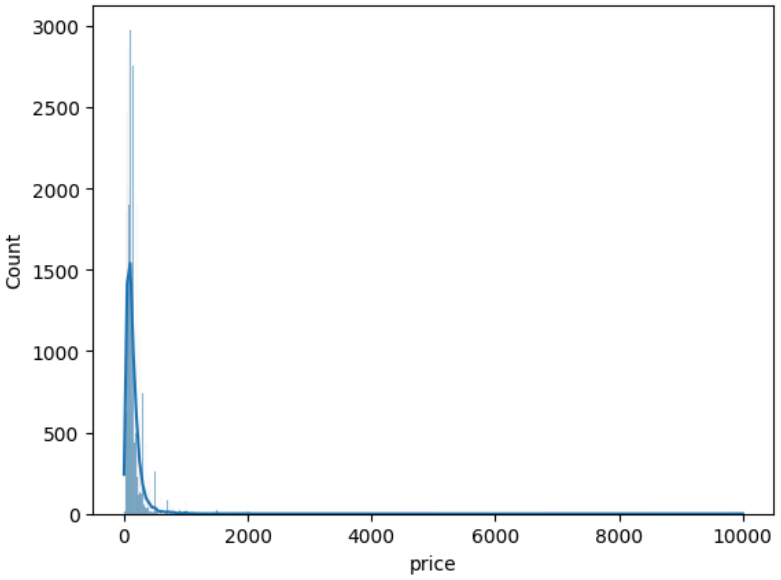

Figura 1 - Histograma da coluna de preços antes da limpeza

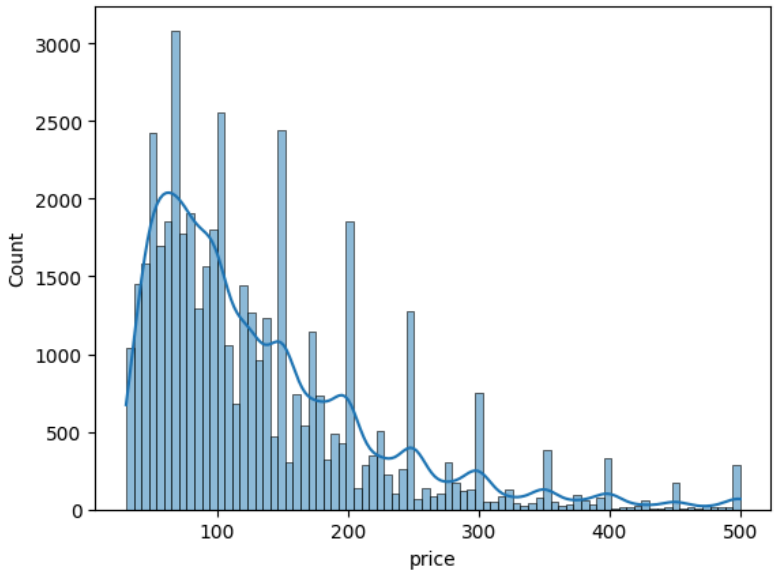

Figura 2 - Histograma da coluna de preços após a limpeza

Da mesma forma que identificamos outliers na coluna de preços, também notamos a presença de dados discrepantes nos bairros, os quais decidimos remover para assegurar a eficácia dos modelos a serem utilizados mais adiante.

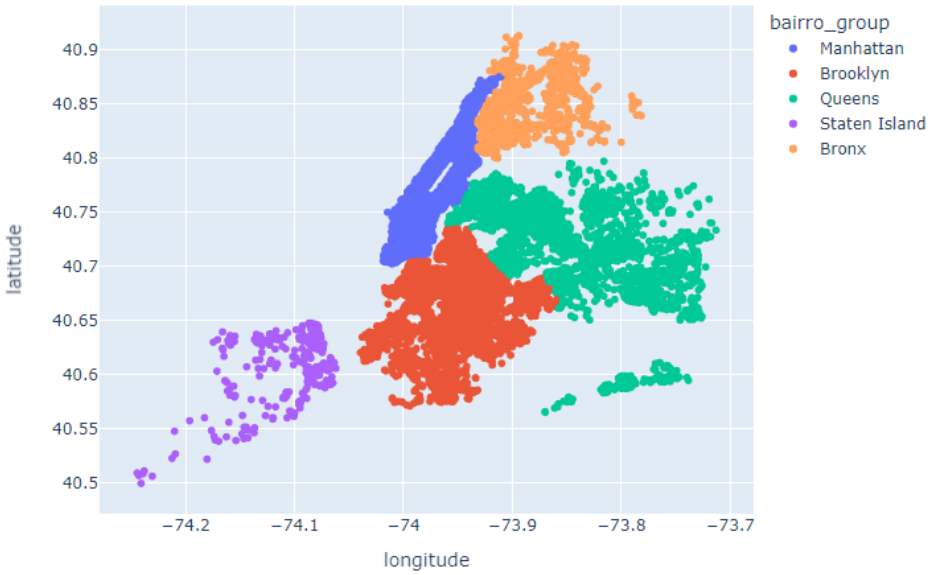

Figura 3 - Plot de mapa utilizando ”bairro_group” com dados de latitude e longitude

Conforme observado na Figura 4, alguns bairros possuem uma frequência muito baixa no conjunto de dados, e optamos por excluir todos aqueles que possuíam menos de 200 registros.

```
Williamsburg           3861
Bedford-Stuyvesant     3640
Harlem                 2617
Bushwick               2420
Hell's Kitchen         1895
                       ...
Graniteville              2
New Dorp                  1
Rossville                 1
Richmondtown              1
Willowbrook               1
Name: bairro, Length: 219, dtype: int64
```

Figura 4 - Contagem de valores da coluna de bairros antes da limpeza

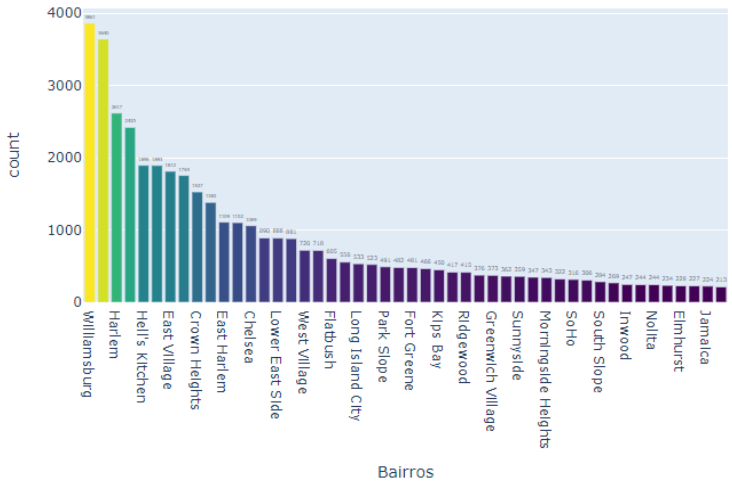

Figura 5 - Gráfico de barra dos bairros após a limpeza

Ao examinarmos as variáveis "room_type", "bairro_group" e "bairro", notamos que exercem influência sobre o preço dos imóveis. A partir da análise da Figura 6, é possível inferir que Manhattan registra o maior preço médio de imóveis. Isso indica uma demanda elevada nesta região.

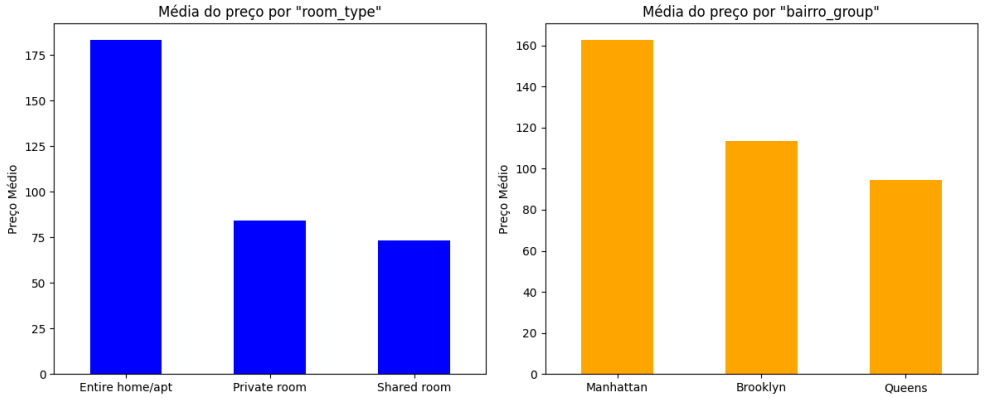

Figura 6 - Gráfico de barra da média de preço para bairros e tipos de quarto

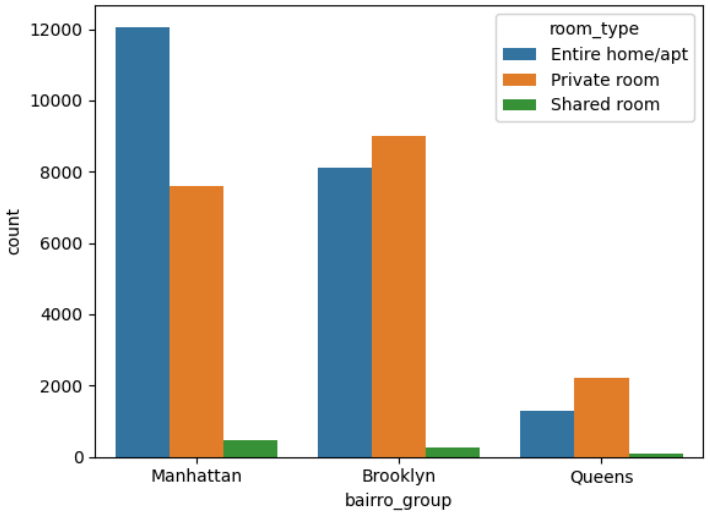

Figura 7 - Correlação entre os agrupamentos de bairros e os tipos de quartos

Ao analisarmos a Figura 7, percebemos que o bairro de Manhattan possui uma maior proporção de quartos do tipo "entire home", enquanto o Brooklyn predominantemente oferece quartos do tipo "private". Considerando que o tipo de quarto "entire home" também exibiu o maior preço médio em comparação aos outros tipos, podemos concluir que esses dois dados estão correlacionados..

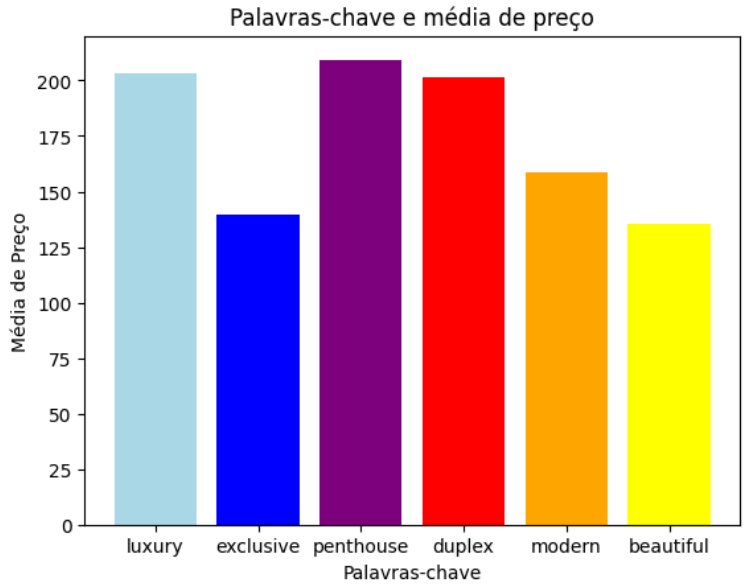

Figura 8 - Gráfico de barra da relação entre palavras-chave e a média de preço

De acordo com a figura acima, ao analisarmos determinadas palavras-chave em conjunto com os preços, percebemos que termos como "luxo", "cobertura" ou "duplex" frequentemente estão relacionados a locais de maior valor. No entanto, é essencial destacar que essas palavras foram escolhidas previamente e não podemos assegurar com total certeza que apenas elas influenciam os preços da mesma maneira que outras variáveis.

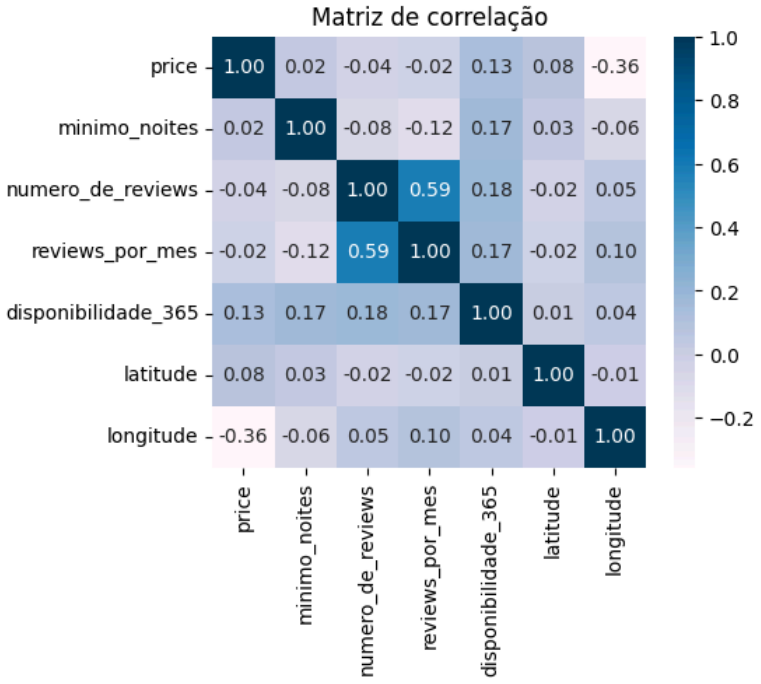

Figura 9 - Matriz de correlação para as variáveis numéricas

Conforme vemos na Figura 9, não há uma correlação linear significativa entre as variáveis e o preço, exceto pela longitude, que exibe uma correlação negativa um pouco mais acentuada. No entanto, é crucial observar que a ausência de uma correlação linear não elimina a possibilidade de existir alguma relação não linear entre as variáveis. Por isso utilizaremos essas variáveis numéricas no teste dos nossos modelos de regressão.

Modelo de Regressão

Primeiramente, definimos os dados de entrada `x = data[['nome', 'bairro_group', 'bairro', 'room_type', 'minimo_noites', 'numero_de_reviews', 'reviews_por_mes', 'calculado_host_listings_count', 'disponibilidade_365']]` e saída `y = data['price']`

Como não conseguimos encontrar uma relação significativa entre o preço e as coordenadas geográficas (latitude e longitude), optamos por utilizar apenas as características de categorias de bairro e tipos de quartos para classificar os imóveis. Estou utilizando uma ferramenta para codificar as informações categóricas. As outras informações nas colunas são numéricas.

O tipo de problema que estamos resolvendo é de regressão, pois estamos prevendo um valor contínuo (o preço dos imóveis). Dentre os modelos que poderiam ser utilizados para esse tipo de problema, optamos por testar vários modelos de regressão, como Linear, Ridge, Lasso, Decision Tree e Random Forest, avaliando-os com base no R2 Score.

No caso de previsão de preços de imóveis, o R2 Score é geralmente preferido como uma métrica de avaliação, especialmente quando se trata de modelos de regressão. O R2 Score fornece uma medida da variabilidade dos dados que é explicada pelo modelo, ou seja, quanto mais próximo de 1 o R2 Score estiver, melhor será a capacidade do modelo em explicar a variabilidade dos preços dos imóveis.

Você pode conferir os resultados do teste abaixo:

- Linear Regression:
  - R-squared score (train): 0.637
  - R-squared score (test): 0.485
- Ridge Regression:
  - R-squared score (train): 0.608
  - R-squared score (test): 0.532
- Lasso Regression:
  - R-squared score (train): 0.275
  - R-squared score (test): 0.276
- Decision Tree Regressor:
  - R-squared score (train): 0.806
  - R-squared score (test): 0.326
- Random Forest Regressor:
  - R-squared score (train): 0.929
  - R-squared score (test): 0.532

Como podemos observar, o modelo de Random Forest foi o mais eficaz entre os modelos avaliados e foi selecionado para aplicação na sugestão de preço com base no conjunto de dados fornecidos. Ao utilizar o modelo proposto para prever o preço do imóvel descrito, obtivemos uma sugestão de preço de $221.35. Este valor está bastante próximo do preço registrado no próprio conjunto de dados ($225), indicando a eficácia do modelo desenvolvido.