1. Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses de negócio relacionadas. Seja criativo!

Algumas hipoteses de negocio levantadas:

1.1. The graphs slightly show a positive correlation tendency with the number of reviews and the availability, even higher when looking at the negative corr. relating the number of reviews with minimum nights per stay, suggesting that properties with higher availability tend to receive more reviews and that the lower is the minimum required nights, the higher the customer is likely to leave a review. This could imply that properties with more open availability and less restrictive booking conditions attract more guests, leading to a higher likelihood of receiving reviews.

1.2. The analysis reveals a market trend of increased property availability towards the end of the year, suggesting a potential alignment with SEASONALITY of the market and higher demand during specific periods;The positive correlation observed between the rolling averages of availability and the number of reviews indicates that as availability increases, the number of reviews tends to follow suit even though only mildly. This connection suggests a inclination on availability and guest reviews, potentially influenced by strategic adjustments made by property owners in response to seasonal variations and increased demand;

NOTE:::: To enhance this understanding, it would be ABSOLUTELY necessary to delve into deeper investigation, segment-specific analyses, considering different property types or locations and years (here I considered only 2018 for being the most recent year that I had the full data (jan-Dec)); Additionally, a cautious approach should be maintained, acknowledging that while correlation implies a relationship, causation requires further exploration of contributing factors and data integrity verification.

1.3. The insights obtained by the geographical distribution gives us a clear vision that Manhattan and Brooklyn are, in that order and by alarge, the neighborhood groups with the most units available for rent. That is due to a higher demand or attractiveness to tourists, most likely influencing pricing and marketing strategies.

1. Responda também às seguintes perguntas:
    1. Supondo que uma pessoa esteja pensando em investir em um apartamento para alugar na plataforma, onde seria mais indicada a compra?

About the prices per neighborhood and taking into account graphs 4 to 6:

2.1.1. For an investor seeking maximum profit probability in NYC real estate, strategic considerations center on neighborhoods with high demand, particularly in Manhattan and Brooklyn, known for consistent popularity;

2.1.2. the avg price analysis reveals potential investment opportunities, with a focus on areas that strike a balance between higher prices and steady demand. Luxury units, like those in Tribeca or exclusive locales such as Seagate in Brooklyn, present opportunities to attract high-end tourists. Seasonal pricing adjustments, diversification across neighborhoods, and local partnerships for added value are key strategies;

2.1.3. HOWEVER, staying informed about market trends by performing a wider, deeper market research, will contribute to long-term success;

2.1.4. Ultimately, understanding the unique selling points of each neighborhood and continuously adapting to market dynamics are crucial for making informed investment decisions with a high profit probability.

1. O número mínimo de noites e a disponibilidade ao longo do ano interferem no preço?

2.2.1. They both do. Minimum nights tend to have a negative correlation with price, where the most expensive units are likely to have a lesser requirement to the minimum nights period of stay. The availability, although doesn’t show a high correlation to price, MIGHT indicate seasonality, which means that an investor that has units available during high season might want to increase its price due to high demand.

1. Existe algum padrão no texto do nome do local para lugares de mais alto valor?

2.3.1. Yes, by using wordcloud it seems that the evaluated data has a pattern in which, terms such as "Central Park" "Beautiful" "Private Room" "Brooklyn" "Spacious" "NYC" "East Village" "Modern" “Manhattan”, etc (check cell 29) are likely popular descriptors for the listed accommodations and contribute to a higher average price, as seen in the analysis. The prominence of these words in the word cloud suggests that they play a role in attracting attention and potentially influencing pricing in the Airbnb listings.

3. Explique como você faria a previsão do **preço** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

3.1. Unfortunately, none of my ML models reached a satisfactory predition and I was unable to improve it enough with the time frame that I had to work with it. However here are my final thoughts and breakdown:

This is a regression problem because the outcome variable (price) is continuous, and we aim to predict its value based on the input features;

The variables chosen include minimum_nights, number_of_reviews, availability_365, and room_type, along with derived variables like has_keyword (indicating if the listing's name contains keywords) and location-based dummies (neighborhood_group_Brooklyn, neighborhood_group_Manhattan, etc.);

R² Score and Root Mean Squared Error (RMSE) were chosen as the performance measures;

The decision to focus on R² Score and RMSE as performance measures was driven by the need to understand both the proportion of variance explained by the model and the average magnitude of the model's prediction errors, respectively, which are crucial for evaluating and improving the model's predictive power.

The Random Forest Regressor emerged as a strong candidate among the models tested, including Linear Regression, Decision Tree, and Ridge Regression.

Pros: Random Forest can handle non-linear relationships without explicit feature engineering for interactions, is less prone to overfitting than Decision Trees due to its ensemble nature, and automatically assesses feature importance.

Cons: It is more computationally intensive than Linear Regression, offers less interpretability due to its complexity, and requires careful hyperparameter tuning to avoid overfitting and ensure optimal performance.

4. Supondo um apartamento com as seguintes características:

```
{'minimum_nights': 1,
 'number_of_reviews': 45,
‘has_keyword’: 0
 'neighborhood_group_Brooklyn': 0,
‘neighborhood_group_Manhattan’: 1,
‘neighborhood_group_Queens’: 0,
‘neighborhood_group_Staten Island’: 0,
‘room_type_Private room’ : 0,
‘room_type_Shared room’ : 0,
‘log_availability_365’: 5.875
 'availability_bins_Medium': 0,
'availability_bins_High': 1,
‘Price’: 225}
```

Qual seria a sua sugestão de preço?

4.1. After creating a df for the described apartment and adapting the columns to match the model pattern, the price I came upon after using the LR model was:

193.90857954894096