**Statistical/Hypothetical Question**

How do Airbnb prices vary across cities in Europe? Which variables have the greatest impact on the price of a property?

**Outcome of EDA**

The aim of this EDA was to explore how Airbnb prices vary across different cities in Europe and understand which variables have the greatest impact on the price of a property. The data was initially collected in multiple CSV files for each city, which were then concatenated to create one data frame for analysis.

The initial EDA involved cross-featuring the longitude and latitude to create the lng-lat variable, which captures the spatial features of each city. The data was then cleaned by dropping columns that were not relevant for the analysis.

Next, the room\_type and host\_is\_superhost variables were transformed into dummy variables, and the column names were changed to better describe the variables. The data was then visualized using histograms to explore the distributions of each variable. The histograms showed that the data was skewed for some variables, and the logarithmic transformation was used to normalize the data.

The correlation matrix was then computed, and a heatmap was generated to visualize the correlations between the variables. The heatmap revealed a strong positive correlation between price and number of bedrooms and satisfaction\_rating and attractions\_rating. In contrast, city\_dist was negatively correlated with price.

Highly correlated variables were then removed from the data to avoid multicollinearity issues in further analysis. The outliers in the price variable were identified using the z-score method, and the data was cleaned by removing the outliers. Finally, the PMF and CDF functions were used to explore the relationships between different variables.

After conducting exploratory data analysis, a hypothesis test was performed to determine whether listings with different room types have the same mean price. The null hypothesis was that the mean price for all room types was the same, while the alternative hypothesis was that private rooms had a higher mean price than shared rooms. The test found a significant difference in mean price between private and shared rooms, with a p-value of 0.0.

A regression analysis was also performed to determine the relationship between various factors and the listing price. Polynomial regression models with degrees ranging from 1 to 4 were used, and the R^2 values and mean squared errors were calculated for each model. The highest R^2 value of 0.54 was obtained with a 4th degree polynomial regression model.

Additionally, p-values were obtained using the statsmodels package to determine the significance of each factor in predicting the listing price. The factors with the lowest p-values, indicating the strongest significance, were the number of bedrooms, attractions rating, location (as indicated by the combined longitude and latitude), and private room status.

**Missed Analysis and Additional Variables**

The EDA covered a wide range of variables, and the initial analysis revealed some strong correlations between the variables. However, there are several additional variables that could have been included in the analysis to better understand the factors that impact the price of a property. For example, variables such as the size of the property, the amenities available, and the proximity to public transport could have a significant impact on the price of a property.

**Incorrect Assumptions**

The EDA did not make any incorrect assumptions, and the analysis was based on the data available. However, it is important to note that the data may be biased due to factors such as the quality of the data collection and the presence of outliers.

**Challenges Faced**

The main challenge faced during the EDA was dealing with the skewed distribution of some variables. The logarithmic transformation was used to normalize the data, but this may not be the best approach for all variables. Additionally, dealing with outliers in the data required careful consideration to avoid removing important information from the dataset.

In conclusion, the EDA provided valuable insights into the factors that impact the price of Airbnb properties in Europe. The analysis revealed some strong correlations between the variables, and additional variables could be included in the analysis to further explore the factors that impact the price of a property. Overall, the EDA provided a good starting point for further analysis and highlighted the importance of carefully selecting variables and cleaning the data to avoid issues such as multicollinearity and outliers.