Introduction to Data Science – Final Paper

1. Introduction

In this project, we sought to understand the relationship between the location of AirBNB properties and their prices, with the hypothesis that location is a key factor influencing price. We hypothesized that properties in more attractive or central areas would command higher prices, as such locations are often perceived as more attractive by tourists and travelers. To test this hypothesis, we used a dataset of AirBNB listings from 10 major cities around the world. By examining these cities, we aimed to answer the broader question: How and when does a property's location affect its price?

"Location" can be interpreted in different ways, so we chose to focus on two key dimensions for this analysis. First, we looked at the property's distance from the city center, assuming that proximity to a central business district or popular tourist areas would make a property more attractive and therefore more expensive. However, we also considered the potential downsides of central locations, such as noise or congestion, which might mitigate this effect. Second, we examined the socioeconomic status of the neighborhood, hypothesizing that properties in wealthier and more upscale neighborhoods would likely command higher prices.

This study is particularly interesting because understanding how location affects price can provide valuable insights for both property owners looking to improve their listings and tourists looking for accommodations that balance price and proximity.

2. Methods

Our work is comprised of the following steps:

a. **General EDA**: In order to run any kind of calculation, we need to format the data. This step made sure that the data frame will include:
   - numerical features only
   - the most relevant features- features that are correlated with the target feature "price", but are minimally correlated between them.
   - no Nan values. This we did by checking what is the distribution of the features that had Nan values. Then we replaced Nan with the median of the feature, if it had skewed distribution. If the feature distributed more symmetrically, we replaced Nans with the mean.

   Additionally, some of the features had to be converted to appropriate units of measure, and normalize to rule out the effect of asset size and city average price. After normalization, the price was converted to the variable *price_per_person*.

b. **1st interpretation of location – Distance**

In this part we suggested a new measure called "distance" which represents the asset's distance from its city's center. We wished to predict the price using this new variable.

We aimed at training a linear regression model on the data, so we had to do further preparations, like normalize per city size.

Then we reviewed graphs to get an impression of the possible relations between distance and price.

- A scatterplot of all the data, showing the relations between distance and price, colored per city.
- Same scatterplot, but per selected cities
- A histogram of Distance

With these graphs in mind (presented in Results), we figured our initial proposal of linear regression models might not fit the data, but we decided to still give it a last chance.

c. Feature selection

This is another preparation before running the model. We chose the most significant features to our model, using-

- Correlation "matrix", showing some strongly correlated features that were removed.
- A simplistic algorithm of feature selection that calculates for each feature its R^2 and the p-value of the score. We kept only the significate features, according to Bonferroni adjustments. We decided not to go for a full backward elimination as we figured the iterative process will exceed the scope of this work.

d. A final representation of the data, using "prices" map in selected cities.
This as well available in Results.

e. Linear Regression model
Originally, we planned to run our models with and without the features we were interested in (i.e. different representations of location), and measure the difference in R^2 as their unique contribution to our target feature price.
However, looking at the graphs and correlations, and after running the initial, most "forgiving" model, we did not continue to the 2nd part.
De facto we did have 2 phases-

- Linear model of all the data
- Linear model trained on the data of each city separately

f. **2nd interpretation of location – Socio-Economic level**

We chose Paris data as a case study for this extra step, assuming location can also affect the area's socio-economic (SE) level. We used Paris neighborhoods as "mediators" of this information, mapping each asset to the SE level of its area.

As with the 1st interpretation, We wished to predict the price using this new variable.

The next steps were similar to the Distance chapter –

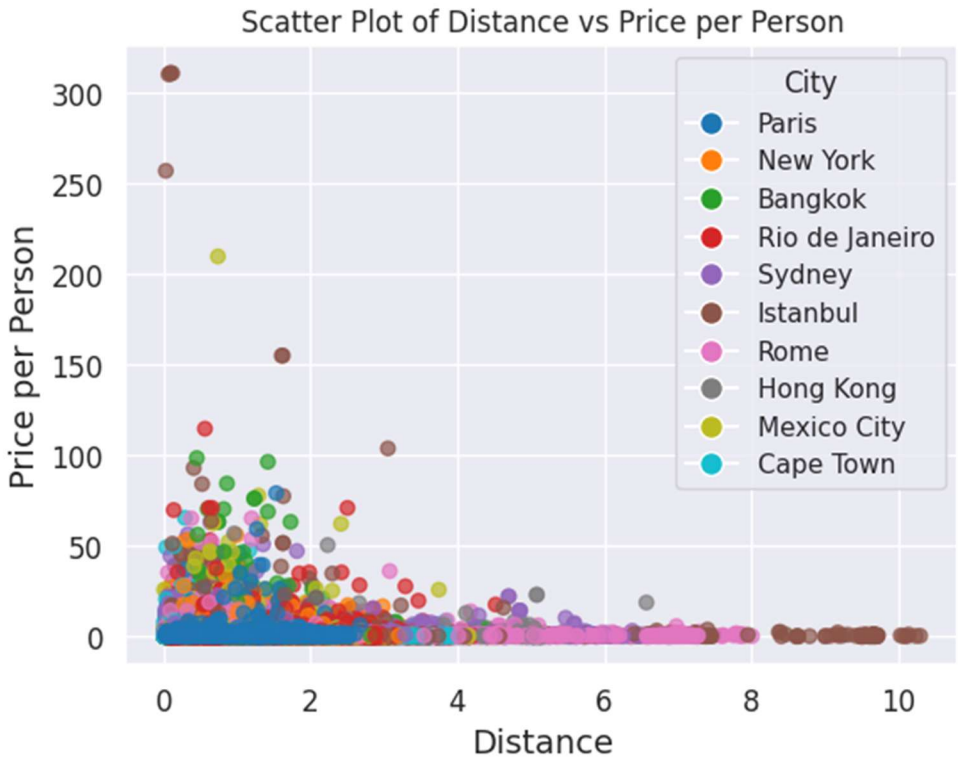
- Review the possible relations using a scatterplot
- Correlation calculation
- Histogram of the assets frequency in each SE level.

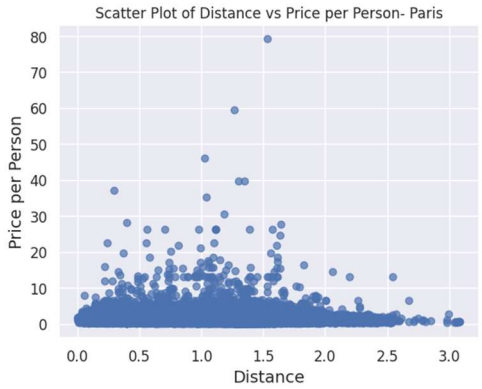
(all available in Results)

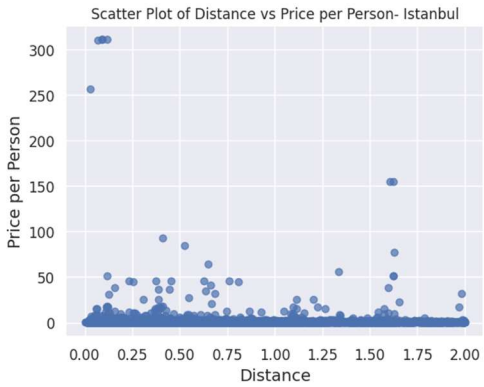
Then we ran another linear regression model.

3. **Results:**

Our initial hypothesis suggested a clear linear relationship between price and location, with properties closer to city centers being more expensive. However, upon analysis, the results did not support this assumption. The relationship between price per person and the features we examined, including distance from the city center, did not fit well with a linear model.

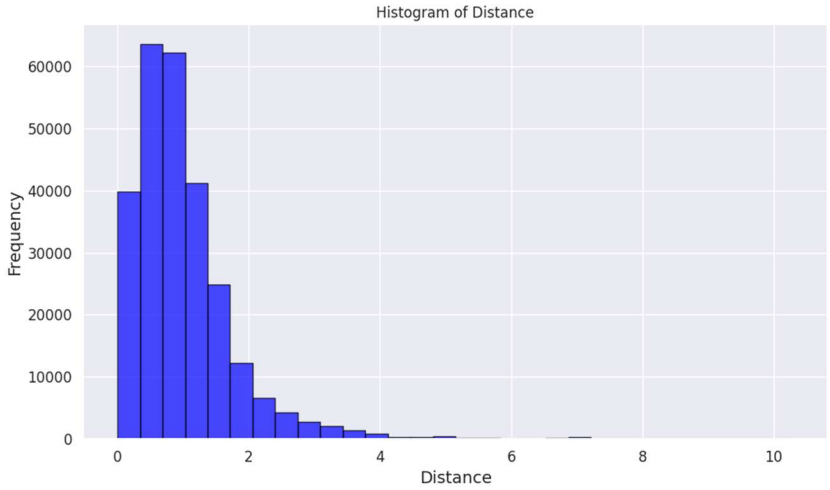






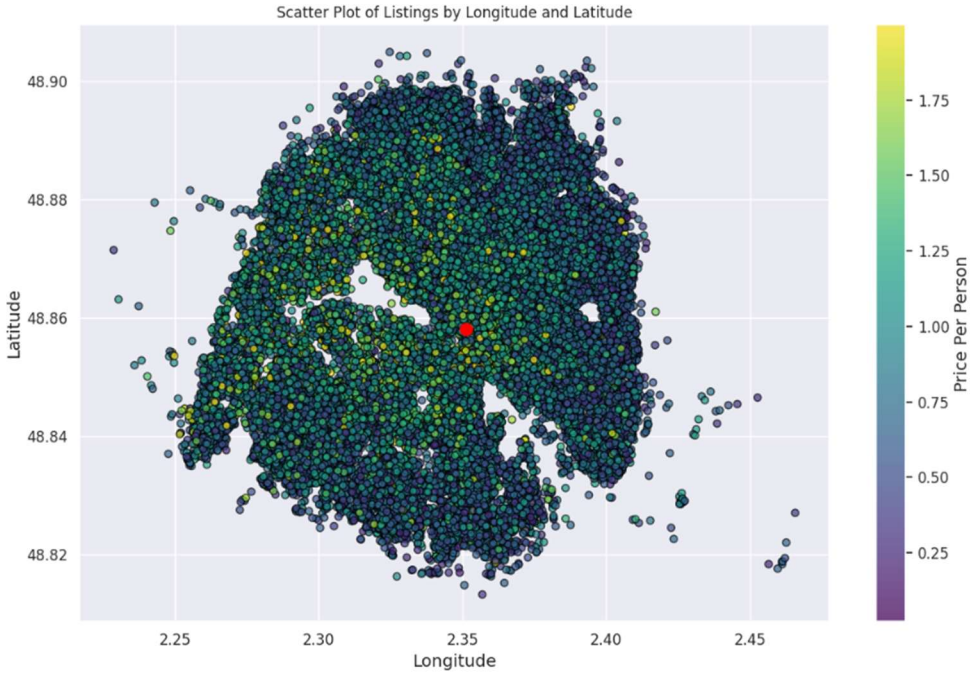
As can be seen from the scatter plot of distance versus price per person-

- Different cities “behave” differently in terms of distance-price relations. This somehow supports our general hypothesis of location-price relations, as cities are different locations technically.
- The relations are non-linear. Even though in first sight some properties located closer to the city center have higher prices, it is probably because the majority of the assets, cheap and expensive, are close to the center anyway.

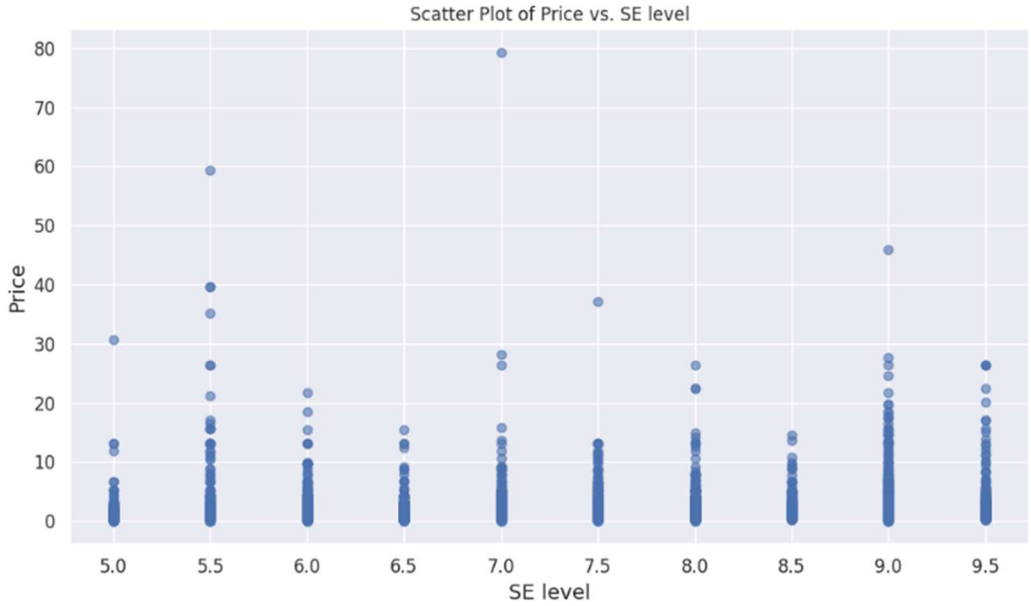


“Playing” with outliers and focusing on cheap/close assets only did not change that (please note that ranges of the 3 cities graphs above).

The graph of Paris assets further shows that while we expected prices to decrease steadily with increasing distance from the city center, the actual pattern is more complex. While some properties further from the city center tend to have lower prices, there are still outliers with very high prices, suggesting that distance alone cannot explain the variance.



This is also true for the 2nd measurement of location – neighborhood SE level in Paris.



The graph implies that here as well the relations are not linear and simple as we initially thought. The low correlation, though significant - 0.18406, also support that.

The results from the linear regression models further support this finding. Across the three different models, the R^2 scores were consistently low, suggesting that the features in the model explain only a small portion of the variance in price. The first model, which analyzed all cities together, returned an R^2 value of 0.035, indicating little explanatory power. The

second model, which analyzed each city individually, produced slightly better but still weak R^2 values across cities, ranging from 0.073 to 0.170, with some cities scoring lower. Finally, the third model, which focused on the Paris dataset alone, produced a similarly weak R^2 value of 0.174.

Model	R^2 Score
Liner model for all country	0.0359

Model Per Country	R^2 Score
Paris	0.1708
New York	0.1104
Bangkok	0.0734
Rio de Janeiro	0.1094
Sydney	0.1702
Istanbul	0.0278
Rome	0.0989
Hong Kong	0.0553
Mexico City	0.1876
Cape Town	0.1876

Model	R^2 Score
Paris SE Level Per Neighborhood	0.1744

While some variables, such as property type and certain amenities, had moderate coefficients, their overall effect on price was not strong enough to form a clear pattern. This suggests that Airbnb pricing is influenced by more complex and possibly nonlinear factors, which the models did not capture.

In conclusion, the results suggest that the relationship between price and features we studied is more complex than we initially assumed. The dispersion in the data and the low predictive power of linear models suggest that other variables—including market demand, seasonal effects, neighborhood characteristics, or even host pricing strategies—play a more important role in determining Airbnb prices. This finding highlights the need for more advanced nonlinear modeling techniques and the inclusion of additional features to better understand the determinants of pricing in the Airbnb market.

4. Discussion

In this project we learned quite a lot, the importance of visualization - far to many times we sat and thought about problems that a simple graph would give us an answer right away, we also learned the proper value of intuitions in data science which don't amount to much, we intuited that the location of a property would most certainly have a direct and linear effect on the price, which simply was not true.

our main finding is that a linear modal is not good enough to predict the price of a property, therefore we assumed that a non-linear modal would do the trick, but one limitation on this assumed implication is that there is a possibility that the airBNB data set does not have enough variables to predict the whole variance of the price, regardless of the modal being linear or not.

If we had more time obviously we would try different nonlinear models, polynomial regressions and random forest to name a few, but mainly we wanted to spend more time trying to find different patterns and sub-populations within the data-set.

There were a lot of difficulties in this project, even something as simple as not forcing the data to give us the answers we wanted was not as simple as one would think, we also struggled with a proper allocation of time, we spent far to much time on the EDA never suspecting that no matter what we do are results are still going to be quiet poor.

5. Learning Points

We learned that data science is not about fitting random data points to some modal, rather it is an iterative process of looking for a pattern, testing it on an appropriate modal, getting our hopes crushed and trying again.