SI 618 Project Part II Report

By: Shivangi Kumar

Motivation

Airbnb, Inc., based in San Francisco, California, operates an online marketplace focused on short-term homestays and experiences. I chose Airbnb as the topic of this project because of my recent Airbnb experiences in the United States as an international student. Airbnb's allow guests to set preferences such as room type and property type when looking up specific rentals in a location. Apart from Airbnb guests' preferences, there are certain aspects that determine the prices of the Airbnb's, and these factors can be helpful in determining which areas or neighborhoods are better suited from a host's perspective. In addition to this, Airbnb features a review system in which guests and hosts can rate and review each other after a stay. The truthfulness and impartiality of reviews may be adversely affected by concerns of future stays because prospective hosts may refuse to host a user who generally leaves negative reviews [2]. Thus, it would be useful to see which properties in different neighborhoods received good reviews in a month as that could help guests choose their rental. Thus, through this project, the overarching goal is to determine the attributes that are the most influential to the prices of listings of the Airbnb's in the US. The analysis questions have been segregated as Airbnb guests' preferences, Airbnb revenue generation, Airbnb Price & property, and Reviews by Airbnb guests. This will help target two to three specific questions within each category.

Key Analysis Questions

Airbnb Guests' Preferences

1. Analyze and plot the number of listings based on their property type to see what people's preferences are of renting an apartment, villa, or a house.
2. Which rentals by location were most reviewed by Airbnb guests?

Airbnb Revenue Generation

3. Calculate the estimated revenue generated for each listing ID by multiplying the price of a property with the "minimum nights" column. What areas are best for bringing in the most money and can be recommended to potential hosts?
4. Plot a bar plot to depict the distribution of the busiest months in terms of number of bookings by months and total estimated revenues generated by the hosts of various properties by months.

Airbnb Price & Property

5. The agenda is to find the variable with highest correlation with price. I will plot a pair plot of selected columns to find correlation of price of the various Airbnb's and other factors such as bedrooms, bathrooms, review scores value, reviews per month and review scores accuracy. In addition, I will perform spearman correlation test on the data to confirm the correlation of price with one other variable having highest correlation.
6. Fit a linear regression model to see how price changes based on the number of accommodates and see if there is any correlation between the two.
7. I will group the properties based on property type to calculate the mean price for each property type. The aim will be to find the most expensive property type by plotting a scatterplot to observe the same.

Reviews by Airbnb Guests

8. In which year (2015, 2016 or 2017) were there the greatest average number of reviews for all the listings and to see the trend in these three years if the number of reviews provided by customers has increased over time?

Data Source

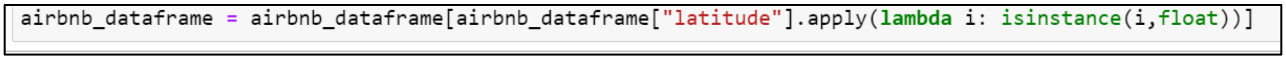
Airbnb Ratings Dataset
Source: https://www.kaggle.com/datasets/samyukthamurali/airbnb-ratings-dataset [1]

In this project, I will investigate the Airbnb ratings dataset which I found on Kaggle. The original dataset has 3 sub-datasets, the LA_listings, NY_listings, and Airbnb_ratings each of which has 59.9k rows and 35 columns. The Airbnb_ratings dataset was filtered to select ratings for only the US and was merged with the other 2 datasets as the columns are the same. The final dataset contains 295,452 rows and 35 columns. The datasets are available to download in CSV format. The data included in this dataset is of the years 2015, 2016 and 2017. The most important attributes used for analysis throughout are latitude, longitude, minimum nights, price, number of reviews, calendar last scraped, property type and room type. An equal number of float data type and object data type variables are present in the dataset that comprise the entire dataset. The data dictionary can be found on the data source link with most of the columns being self-explanatory. For important attributes, the description is as follows:

Listing ID: the ID number of an Airbnb
Host ID: the ID of the host
Longitude: the longitude of the Airbnb
Accommodates: the number of people an Airbnb can accommodate
Bathrooms: number of bathrooms
Bedrooms: number of bedrooms
Price: price of an Airbnb per day
Minimum nights: the minimum number of nights a guest stay
Availability 365: the number of days available in a year
Number of reviews: the total number of reviews
Reviews per month: the number of reviews a host receives per month

Data Manipulation Methods

The first and foremost manipulation that I did was to filter the Airbnb_ratings dataset to limit it to data of the United States and then merge it with the other two listings dataset. In addition, as the latitude column was of object type and the longitude column was of float type, I converted the latitude column to a float data type for further analysis. The latitude column roughly had 300 null values which were disregarded during analysis. I used the lambda function to filter all the float values from the latitude mixed column type as it had string and float values in it.



```
airbnb_dataframe = airbnb_dataframe[airbnb_dataframe["latitude"].apply(lambda i: isinstance(i,float))]
```

Fig 1: Use of lambda function for filtering float values from mixed data type column

Analysis and Visualization

Task 1: Analyze and plot the number of listings based on their property type to see what people's preferences are of renting an apartment, villa, or a house.

In this preliminary data analysis task, I first grouped by the property type and used the "Listing_ID" column to get the frequency of number of listings for each property type. This could be best represented through a bar graph with the property type being on the x-axis and the number of listings which are the count of the listing ids on the y-axis. I also analyzed the prices based on both the room type and property type. As there were 38 types of properties, I took the top 10 properties based on prices. I grouped by both the room type and property type, then formed a bar graph with property type on x-axis, price on y-axis and the room type was given as a hue.

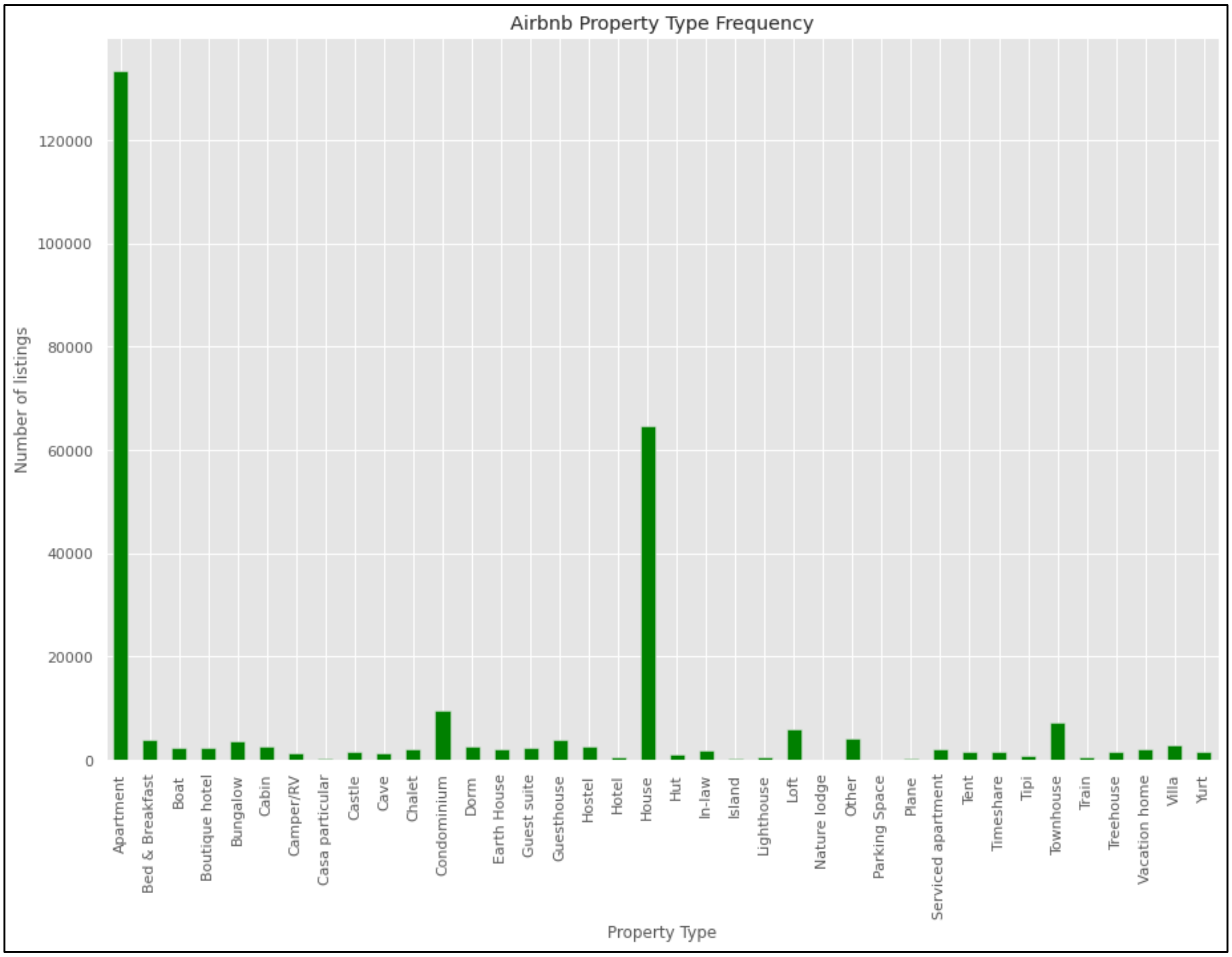


Figure 2: Airbnb Property Type Frequency

Interpretation: It can be seen from this graph that property type plays a key role in the listings of Airbnb. Apartments and houses account for more than 80% of the total listings based on property type as was expected and there are a few instances of other kinds of properties here and there.

Task 2: Which rentals by location were most reviewed by Airbnb guests?

In this task, the "number of reviews" column was used to get the total number of reviews for each neighborhood. The "neighbourhood cleansed" column was grouped to get the total number of reviews for each location. This was plotted as a bar plot with blue markers.

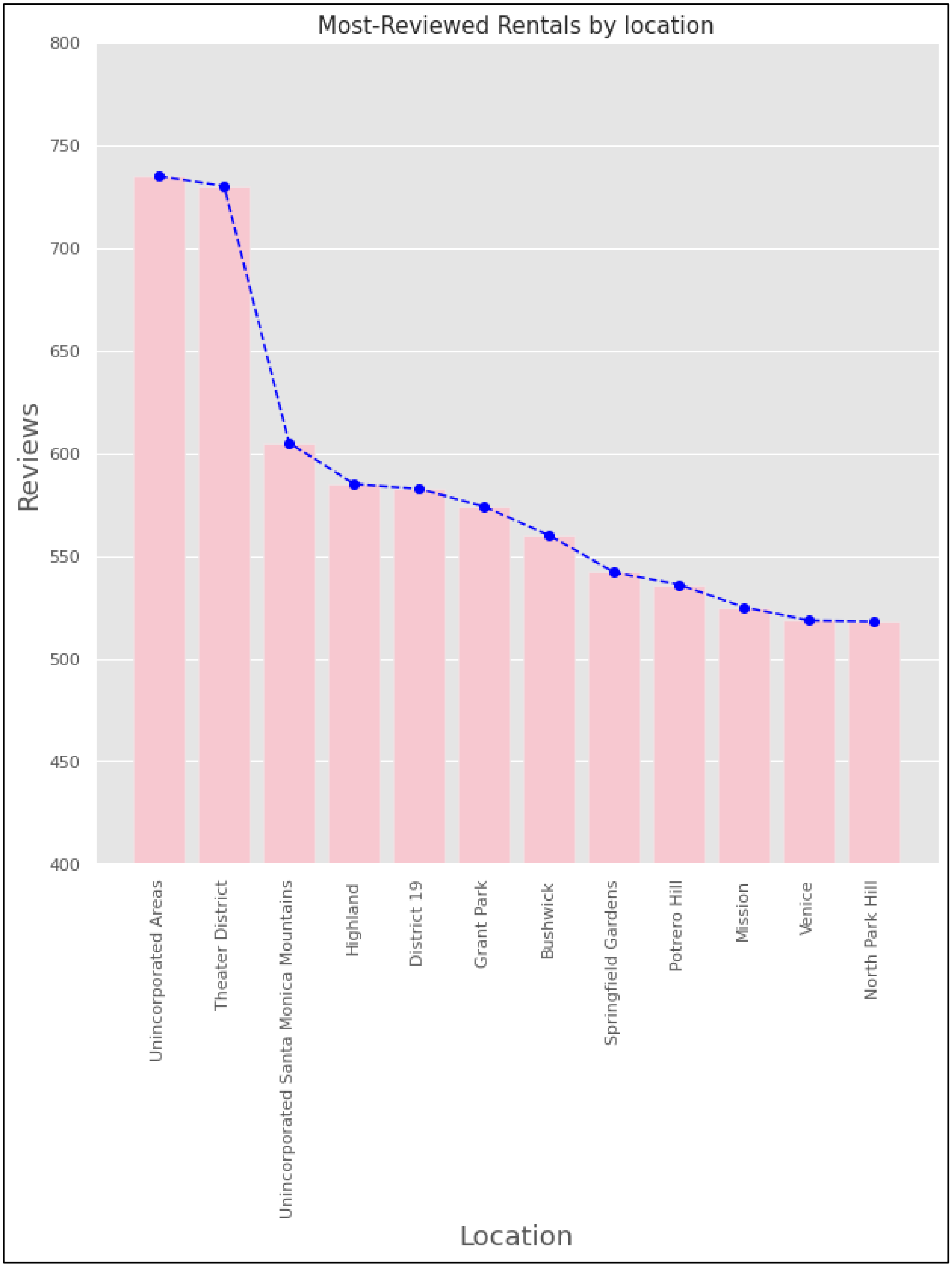


Fig 3: Most reviewed rentals by location

Interpretation: Unincorporated areas and the Theater district have received the highest number of reviews. Most popular Airbnb rentals tend to receive an average of 740 reviews from users. We can also deduce that these rentals were probably most visited or most popular because if they were visited a greater number of times, it might be due to the good reviews of previous guests.

Task 3: Calculate the estimated revenue generated for each listing ID by multiplying the price of a property with the "minimum nights" column. What areas are best for bringing in the most money and can be recommended to potential hosts?

I used the "minimum nights" column and the price of each property to get the estimated revenue for each listing ID. Then I performed a group by on the Listing ID to get the total estimated revenue for each listing. There were certain rows which had price as 0, so I filtered out those rows before doing the analysis. Then I merged this new dataset with the original dataset to get the best neighborhoods for hosts. This was done by taking a mean of the total estimated revenue for each group of neighborhoods. For the top 10 best neighborhoods for potential hosts, a univariate distribution using kernel density estimation for the total estimated revenue was made.

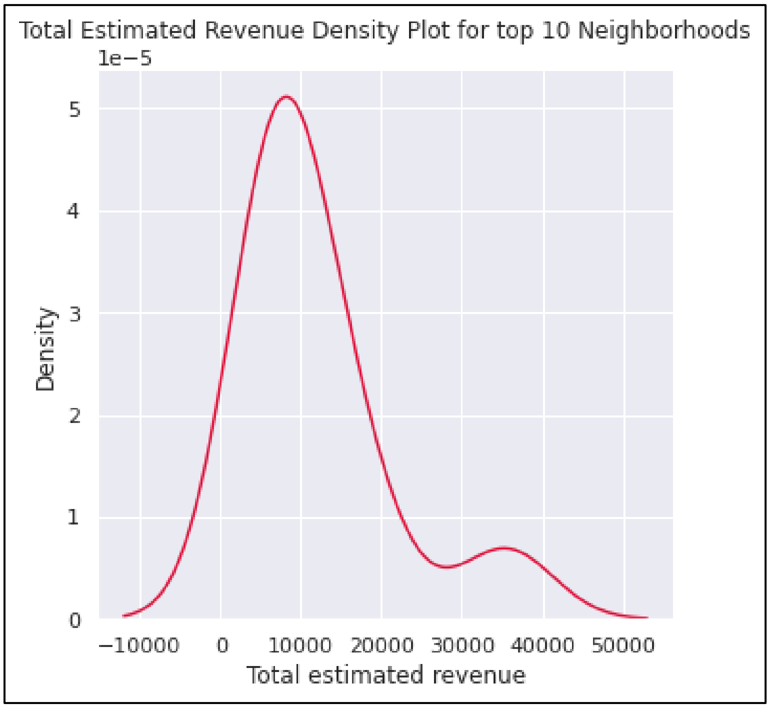


Fig 4: Total Estimated Revenue Density Plot for top 10 Neighborhoods

Interpretation: This is a bimodal distribution of the total estimated revenue as there are two peaks in the density plot. The first peak is highest near the center of the distribution because that's where the most values are located. The second peak is lowest near the ends of the distribution because fewer top 10 neighborhoods have those values of total estimated revenue.

Task 4: Plot a bar plot to depict the distribution of the busiest months in terms of number of bookings by months and total estimated revenues generated by the hosts of various properties by months.

As a part of this task, I tried to project property revenue for prospective hosts in various neighborhoods through a color map where the top n neighborhoods are shown in different color than other neighborhoods. However, the results were not appealing as all the neighborhoods were being shown in a single color. Before creating an approximate map of the top n neighborhoods, I also generated a word cloud to visualize the top n neighborhoods. For this task, I used the "calendar last scraped" column and the "estimated revenue column" to visualize the bookings by month and the revenue by month.

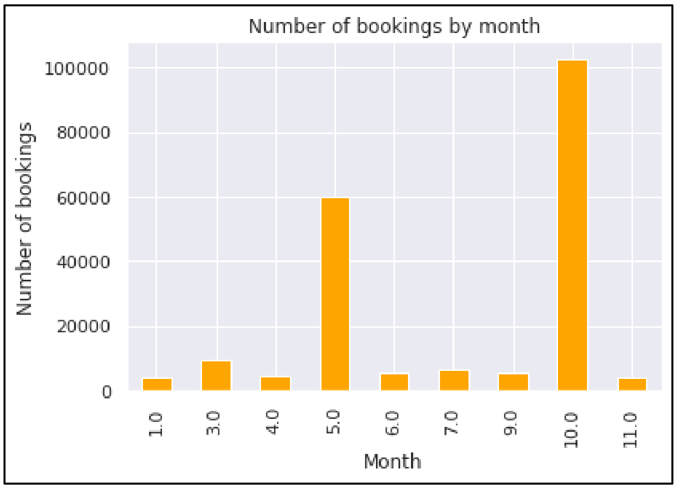


Figure 5: Number of bookings by month

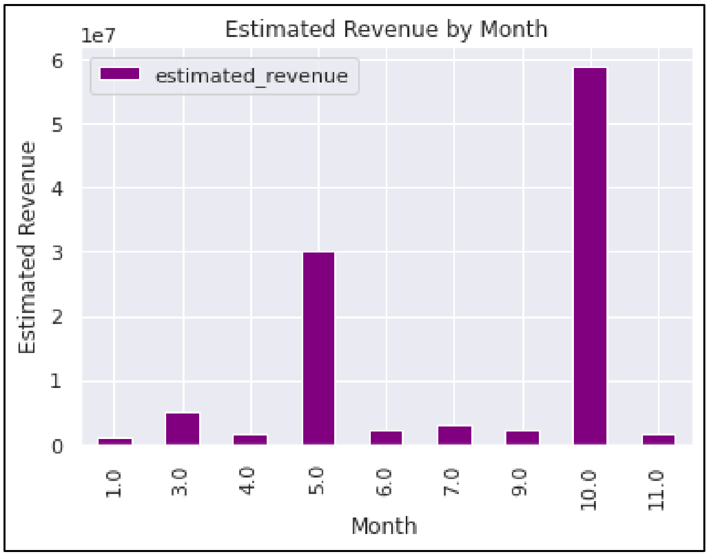


Figure 5: Estimated Revenue by month

Interpretation: We see that estimated revenues are also maximal in the same period as the number of bookings which is during second half of the summer and early fall. Thus, as the estimated revenue depends on the number of bookings for each month, the graphs are similar.

Task 5: The agenda is to find the variable with highest correlation with price. I will plot a pair plot of selected columns to find correlation of price of the various Airbnb's and other factors such as bedrooms, bathrooms, review scores value, reviews per month and review scores accuracy. In addition, I will perform spearman correlation test on the data to confirm the correlation of price with one other variable having highest correlation.

After preliminary data analysis, most of the prices are in the range 0 to 500, bedrooms are generally less than 6, bathrooms are also generally less than 6, the reviews per month are less than 10, the number of reviews is below 60. Thus, I filtered the original dataset based on these conditions. To find the correlation between price and other variables, I plotted a pair plot. From this graph, it was hard to find a linear pattern between 'Price' and other variables, and so I decided to do a correlation test to ensure this assumption. I used a Spearman correlation test and plotted it using a heat map. From the result table, I found that 'Price' and 'Accommodates' have a correlation coefficient of 0.6, which indicates they are moderately correlated and that makes sense as more people a house can accommodate, the more expensive it will be.

Task 6: Fit a linear regression model to see how price changes based on the number of accommodates and see if there is any correlation between the two.

From the analysis in task 5, I decided to fit a linear regression model to see how the price changes based on the number of accommodates and to see if there is any correlation between the two. I grouped the dataset based on the number of accommodates and then calculated the mean price for each set of accommodates. Then I plotted a linear regression model against the actual data.

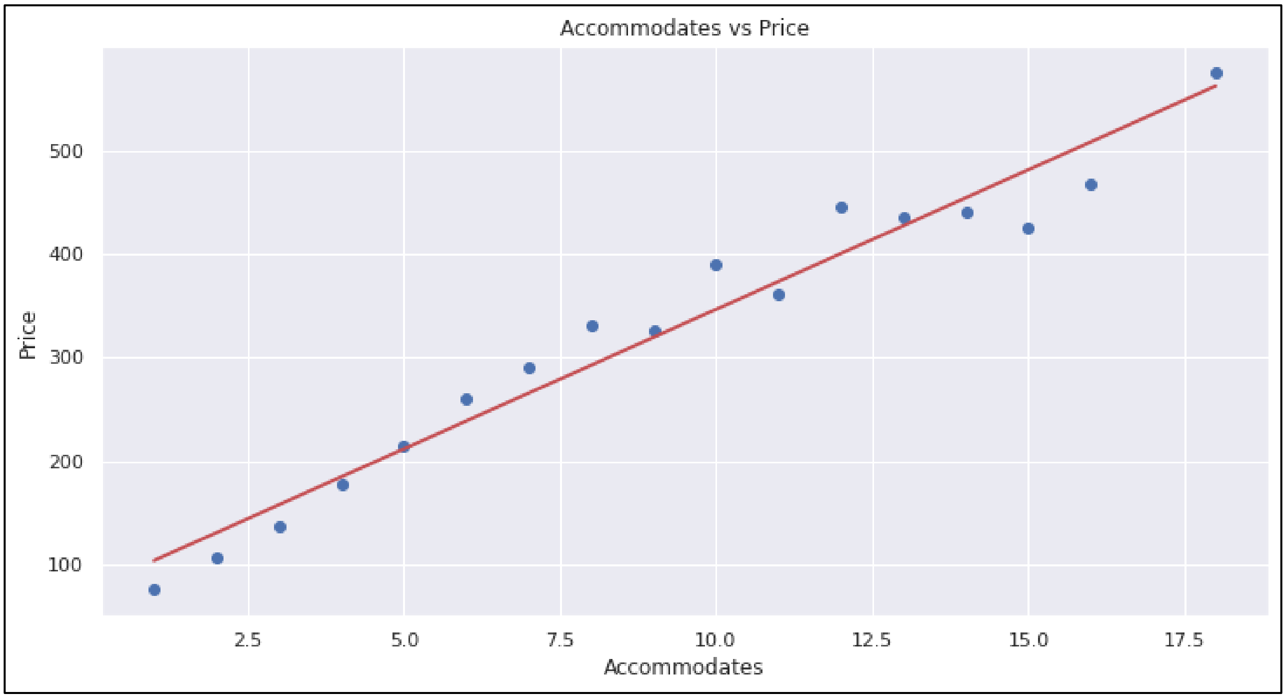


Figure 6: Linear Regression Model – Accommodates vs Price

Interpretation:

From the summary statistics, as R-squared evaluates the scatter of the data points around the fitted regression line, the R-squared value is really good. The larger the R2, the better the regression model fits the observations, so I think that this model fits the observations quite well.

Task 7: I will group the properties based on property type to calculate the mean price for each property type. The aim will be to find the most expensive property type by plotting a scatterplot to observe the same.

After looking at the value counts of the property types, I found that the top 8 property types comprise of most of the listings. Thus, in my analysis, I grouped all the other property types not falling into the top 8 categories as "Other" category. Then, I found the mean price for each property type by grouping by property type and the number of listings for each property type. A scatterplot seemed reasonable as there were 8 property types of which the number of listings and mean price was to be plotted. Thus, on the x-axis is the mean price of the listing, on the y-axis is the number of listings and as a hue I plotted the property type.

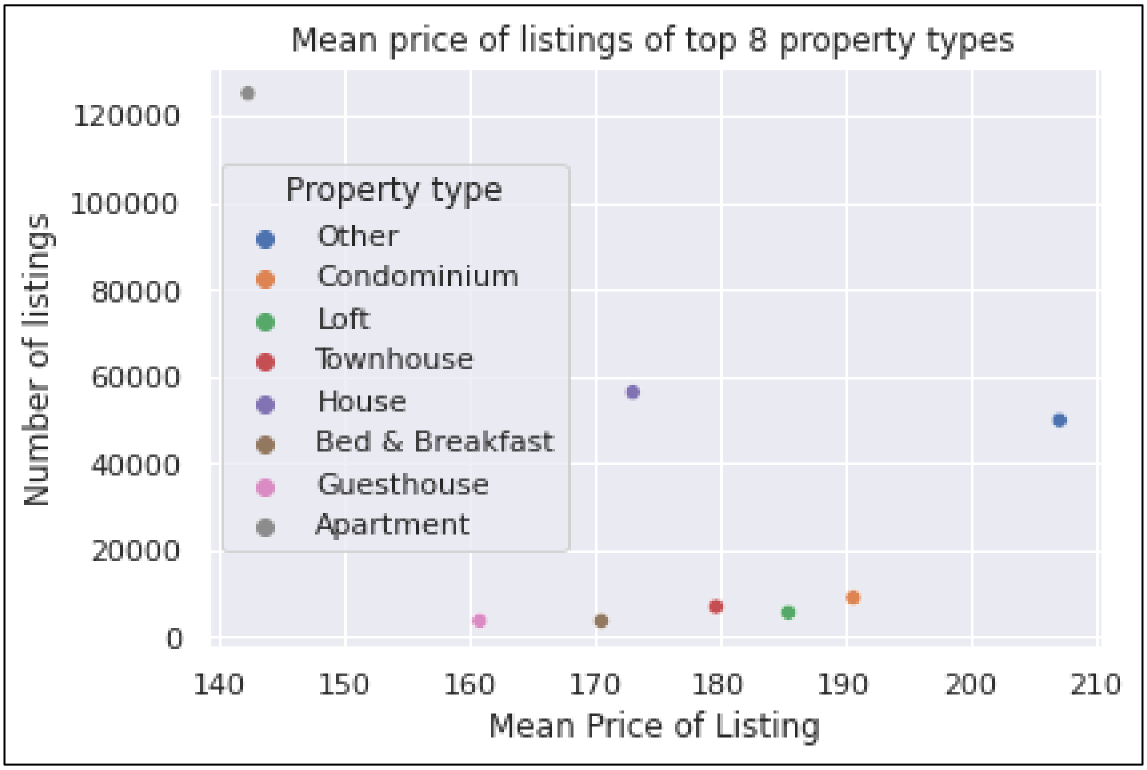


Fig 7: Mean price of listings of top 8 property types

Interpretation: The mean price of an apartment is the cheapest which makes it the property type with maximum number of listings. This is very much valid as more customers would prefer the property type with less cost. The next most popular with a large number of listings is a house which a higher mean price compared to the apartment.

Task 8: In which year (2015, 2016 or 2017) were there the greatest average number of reviews for all the listings and to see the trend in these three years if the number of reviews provided by customers has increased over time?

I created separate columns for Date, Year, and Month based on the "Calendar last scraped" column. Then, I plotted a times series graph based on the date such that the month and year are plotted against the number of reviews. It can be seen that not every year has all the months' data in it which makes it not so appealing.

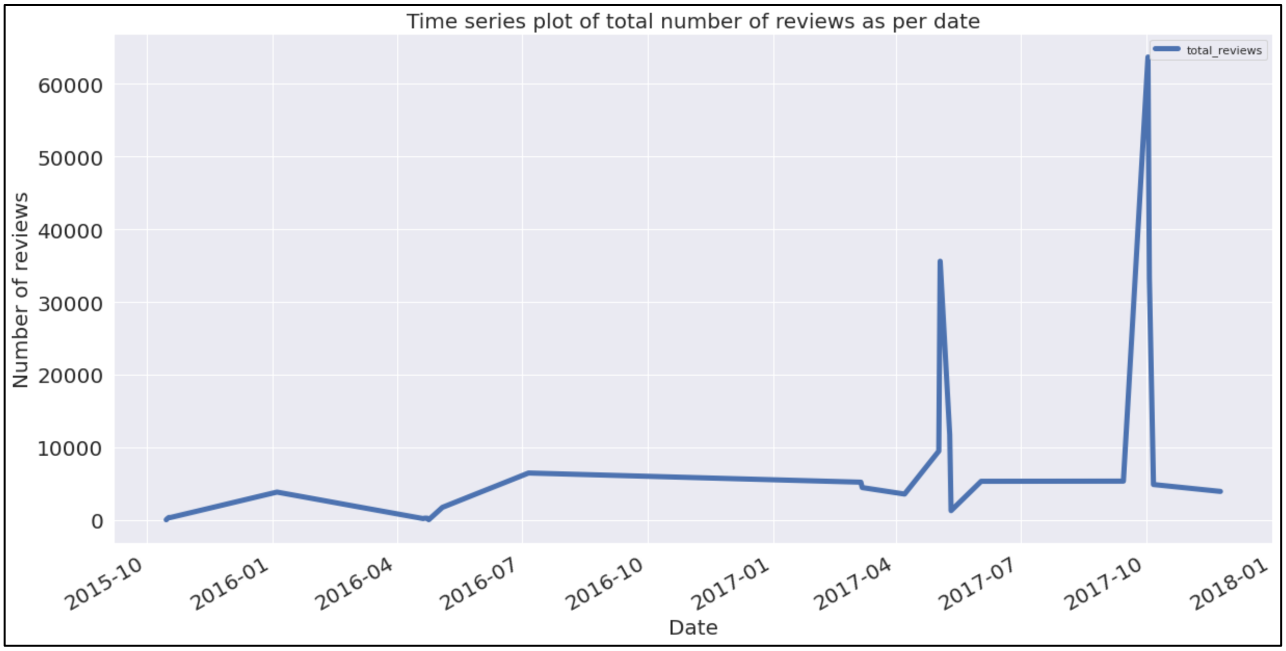


Fig 8: Time series plot of total number of reviews as per date

Interpretation: It can be seen that as time passed by over the years, overall, the number of reviews increased by the year. In April 2017 and October 2017, there can be seen spikes in the number of reviews. This may be due to the increasing popularity of Airbnb and increase in the number of guests due to vacation season.

Challenges

There were three main challenges that I faced throughout this project. I was compelled to drop the rows that did not have values in the latitude column as it was a mixed type of column (float and string) and therefore I had to restrict to the float values in my analysis. In addition, when creating a time series graph for the progression of the number of reviews over the months of 2015,2016, and 2017, I noticed that data was unavailable for a couple of months for the three years which made it difficult to get a bigger picture. Lastly, there were challenges that I faced when creating my color_map function in which I wanted to show the top n neighborhoods based on estimated revenue for potential hosts. I was not able to get the top n neighborhoods in the form of a color map and would like to continue to investigate on it.

References

[1] https://www.kaggle.com/datasets/samyukthamurali/airbnb-ratings-dataset
[2] https://en.wikipedia.org/wiki/Airbnb