Prepared by: Michael Wynn, Scott Miketa, Nishant Singh

Motivation

What accounts for the variability in Airbnb listings?

Airbnb is an online marketplace which provides a platform for home-owners (the host) to rent out their properties to people looking for an accommodation for short-term use. In 2021, there were over 7 million properties (also called listings) on Airbnb and the type of property varies from a shared room to a private room to the entire apartment with property rentals varying from around $60/night to $18,000/night. How are these listing prices decided? Though Airbnb does help the host in setting up a pricing strategy, the price of the property is eventually decided by the host. In a highly competitive market how does the host ensure that they price their properties optimally so as to maximize their earnings? What are the factors which show a strong correlation with the property prices? Is the size of the property the biggest influence on the price? Does the crime rate of a locality influence the property prices? Does proximity to public transport matter?

There are multiple factors that can possibly affect the value of rental listings in each city. For this project, we will be analyzing the variables that have strong relationships with the listing prices in London. We will be exploring many different datasets that are publicly available, such as crime occurrences, population density, local points of interest, and user reviews.

Data Sources

Primary Data Source

The primary data source for the analysis is **Inside Airbnb**. **Inside Airbnb** is an independent website which scrapes and reports the publicly available Airbnb data for over 80 cities around the world.
Source: http://insideairbnb.com/get-the-data.html
From the above source, the following datasets for London were used in the analysis:

S.NO.	File Name	Description	Key variables	Details
1.	Listing	Detailed listing data for all Airbnb properties	▪ Price ▪ Location ▪ Amenities	Format - CSV Size – 152 MB Shape – 67903 x 74

2.	Reviews	Detailed reviews data for the listings in London	▪ Review date ▪ Comment	Format – CSV Size – 309 MB Shape – 1048406 x 6
3.	Neighborhoods	Map of London Boroughs in GeoJSON format	▪ Latitude ▪ Longitude	Format – GeoJSON Size – 1MB Shape – 33 x 3

Secondary Data Sources

Along with the primary data source, the following datasets made up our secondary data sources:

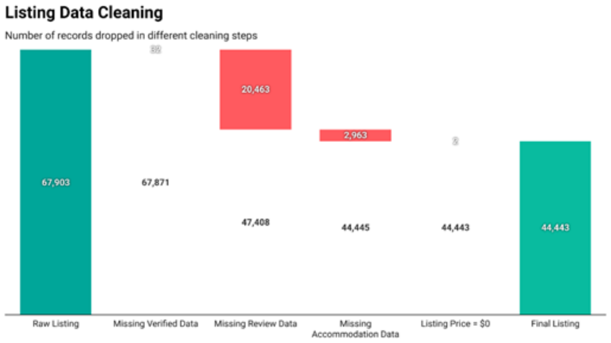
S.NO.	File Name	Description/Source	Key variables	Details
1.	London Underground Data	London Underground Stations with their latitudes and longitudes (London_Underground_maps)	▪ Station name ▪ Latitude ▪ Longitude	Format – HTML table Size – N.A. (Data from webpage) Shape – 307 x 8
2	Property Prices Data	London Borough level property prices (House_price_index)	▪ Borough name ▪ Property price	Format – CSV Size – 2KB Shape – 33 x 4
3.	Crime Data	London Borough level crime data for different crime categories for the period from Nov-19 to Oct-21 (recorded_crime_summary)	▪ Borough name ▪ Crime categories ▪ Crime count	Format – CSV Size – 192 KB Shape – 1555 x 27
4.	Points of Interest data	Top tourist attractions in the UK along with the visitor counts[1] (Tourist_attractions)	▪ Tourist site ▪ Borough name ▪ Lat - Long	Format – CSV Size – 4 KB Shape – 41 x 12
5.	Population data	Land Area, population density and 2011 census data for London Boroughs (Population_density)	▪ Borough Name ▪ Census population	Format – CSV Size – 1.1 MB Shape – 33 x 11

Data Manipulation

We assessed each of the above datasets for inaccurate, irrelevant, incomplete and missing data values. We also identified common keys required to merge datasets. Below is a summary of the data manipulation steps undertaken.

1. Missing value treatment and outlier treatment

[1] . The data for London was copy pasted to a CSV and the latitude and longitude data was added manually


Listing Data Cleaning

Number of records dropped in different cleaning steps

Figure 1. The number of records dropped and the number of records remaining after each cleaning step is figured

Created with Datawrapper

Listing Data is central to our analysis. It captures key features like review score, room type, number of bedrooms etc. of the Airbnb listings. However, for some of the rows in the data, these critical features were missing and hence we decided to drop those records from the final cleaned dataset. As a result, we retained only 44,443 records from the initial 67,903 records. Figure 1 displays the number of records dropped and the reason for dropping the record. For e.g., we dropped 20,463 records because the review scores data were missing.

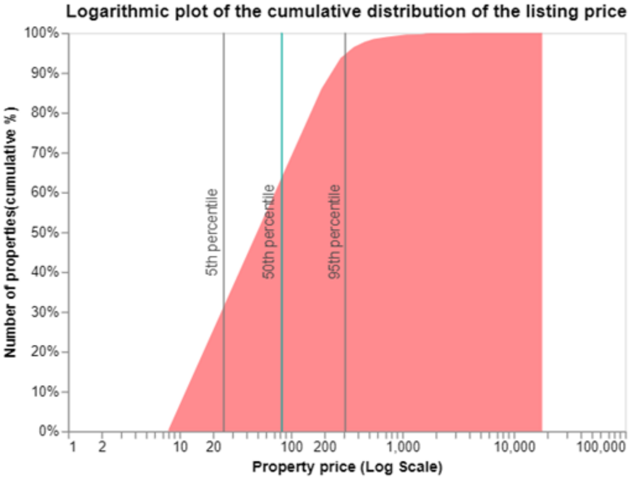
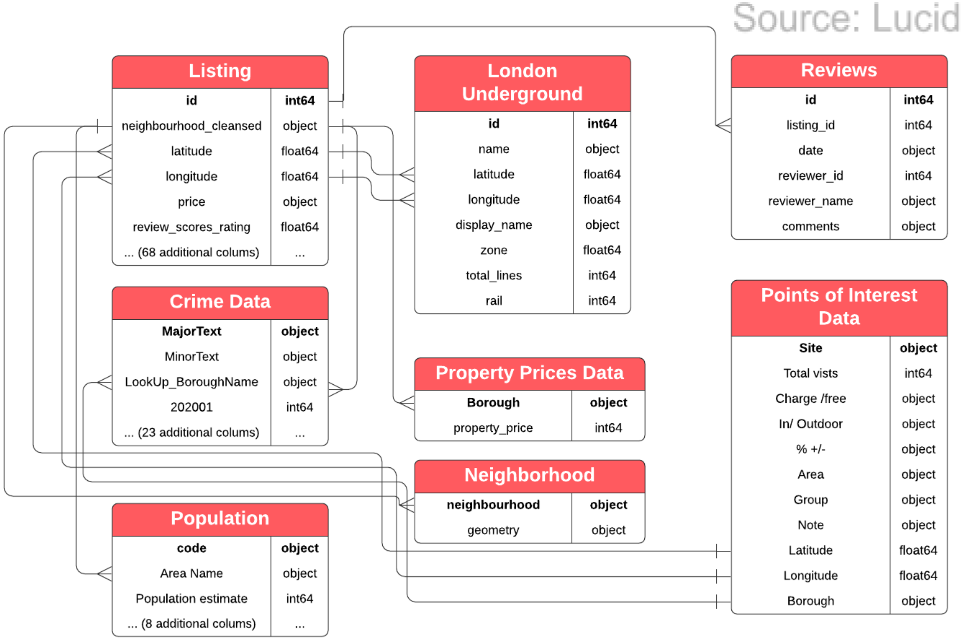


Figure 2. Cumulative distribution of property

In the **Listing data**, we further observed that the distribution for variables such as price, number of bedrooms, number of amenities, number of listings by a host etc. are highly skewed. For e.g., in Figure 2, we observe that the 95th percentile of the listing price is $308 however the maximum value is $18,012. The presence of such outliers would impact the correlation study and the regression analysis. To treat the outliers, we replaced values above the 95th percentile with the 95th percentile and values below the 5th percentile with the 5th percentile value. Besides the Listing Data, none of the other datasets required missing value treatment or outlier treatment.

2. Data joins

Source: Lucidchart.com



Listing	
id	int64
neighbourhood_cleansed	object
latitude	float64
longitude	float64
price	object
review_scores_rating	float64
... (68 additional colums)	...

Crime Data	
MajorText	object
MinorText	object
LookUp_BoroughName	object
202001	int64
... (23 additional colums)	...

Population	
code	object
Area Name	object
Population estimate	int64
... (8 additional colums)	...

London Underground	
id	int64
name	object
latitude	float64
longitude	float64
display_name	object
zone	float64
total_lines	int64
rail	int64

Property Prices Data	
Borough	object
property_price	int64

Neighborhood	
neighbourhood	object
geometry	object

Reviews	
id	int64
listing_id	int64
date	object
reviewer_id	int64
reviewer_name	object
comments	object

Points of Interest Data	
Site	object
Total vists	int64
Charge /free	object
In/ Outdoor	object
% +/-	object
Area	object
Group	object
Note	object
Latitude	float64
Longitude	float64
Borough	object

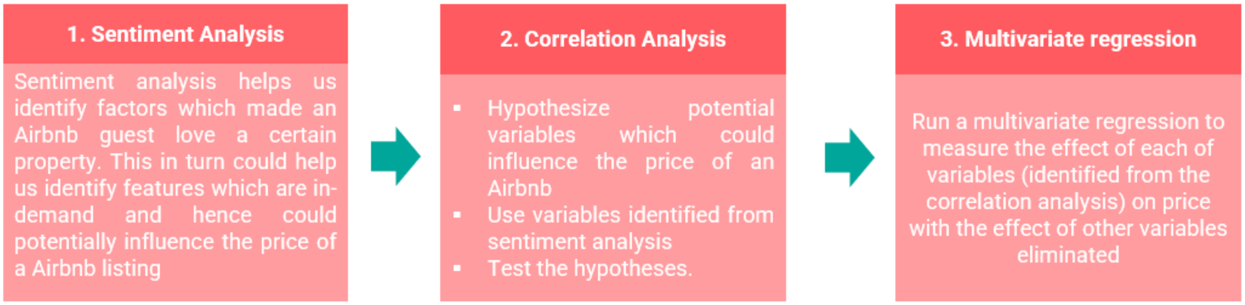
In order to visualize the relationship between our datasets, we have created an entity relationship diagram that shows the connections between the key attributes of our data. Most of the dataset joins were on the borough (neighborhood) level. The other main joins that occurred were on the location level, by comparing their latitude and longitude coordinates.

3. Computation of distance from two latitude-longitude pairs

How does proximity to public transport and proximity to tourist attractions impact the Airbnb prices? To answer the above questions, we decided to use the **London Underground data** and the **Points of interest data** respectively**.** The above datasets provide the coordinates (latitude - longitude) for a point and to compute the distance (in Kms) of every Airbnb listing from these coordinates, we used the Haversine distance. Haversine distance is the shortest distance between two points on the earth (assuming the earth to be spherical).

Analysis and Visualizations

We followed a 3-step approach to the analysis.



1. Sentiment Analysis

Sentiment analysis helps us identify factors which made an Airbnb guest love a certain property. This in turn could help us identify features which are in-demand and hence could potentially influence the price of a Airbnb listing

2. Correlation Analysis

- Hypothesize potential variables which could influence the price of an Airbnb
- Use variables identified from sentiment analysis
- Test the hypotheses.

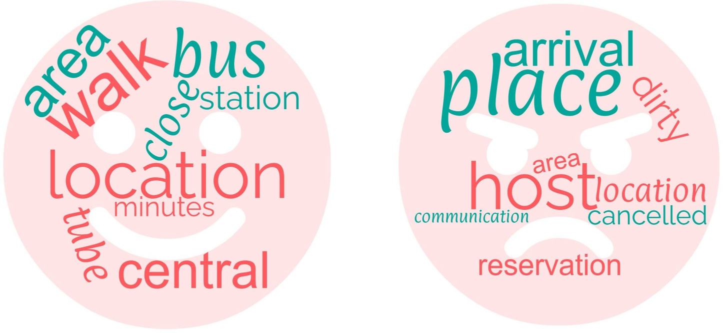
3. Multivariate regression

Run a multivariate regression to measure the effect of each of variables (identified from the correlation analysis) on price with the effect of other variables eliminated

Sentiment Analysis

To begin our investigation, we thought the most useful way to initially explore the driving variables of Airbnb prices would be to hear from customers themselves. While we didn't have access to any Airbnb survey data, we did have access to the reviews that Airbnb users left on hundreds of thousands of properties in London. By analyzing the text of their descriptions, we can see which words appear most frequently in the highly positive and negative listings; there were several frequent words in both of the reviews that pointed us in a few directions.

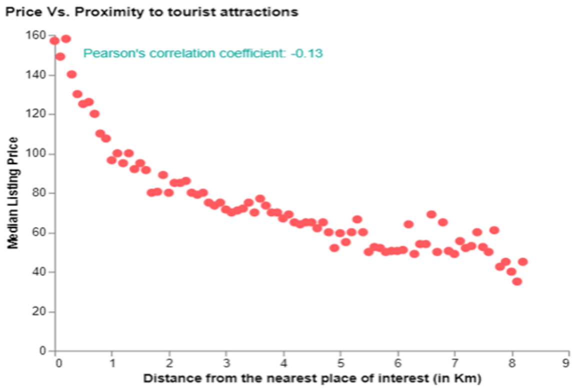
Source: wordclouds.com



These popular 9 words appeared frequently in the highest rated Airbnb listings, which tells us that location, and proximity to certain points of interest seem to be very important to a positive review.

The negative reviews often mentioned negative issues with the host, or the cleanliness of the property, but “location” and “area” were also mentioned.

Listing price Vs. Proximity to tourist attractions

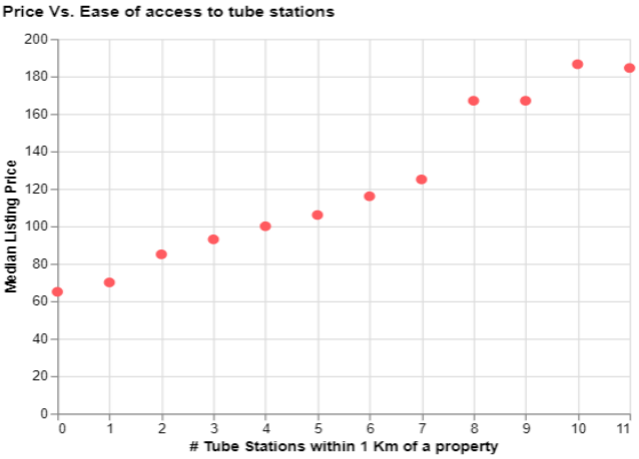


As a popular tourist destination, London has many local points of interest. Because proximity seems to be important to Airbnb visitors, we mapped how proximity to multiple points of interest correlates to the price of listings. There appears to be a negative correlation between the distance of a property from the nearest tourist attraction and the median listing price of the property. Airbnb properties closer to tourist attractions are relatively more expensive.

Listing price Vs. Accessibility to public transport

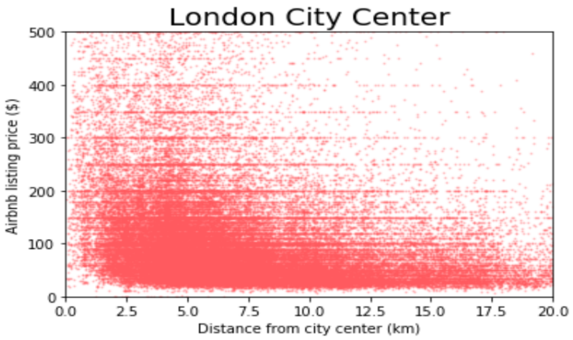
The sentiment analysis revealed that keywords like ‘bus’, ‘station’, ‘walk’ and ‘tube’ are associated with positive reviews, indicating that proximity to public transport is an in-demand feature. One of the most popular public transport systems in London is the London Tube. To measure the correlation between listing price and accessibility to public transport, we analyzed the relationship between listing price and number of tube stations within a kilometer of an Airbnb property.

There appears to be a strong positive relationship between the number of Tube stations within a kilometer of a property and the median property price.



London’s Layout

London is a city that is centered around the River Thames. This means that the city was built with a specific center, and expanded outward from there. Now, London is broken up into 6 primary zones, as a system of concentric circles. We decided to explore the effect of mapping the relationship between Airbnb listing prices and the distance from the city’s center.



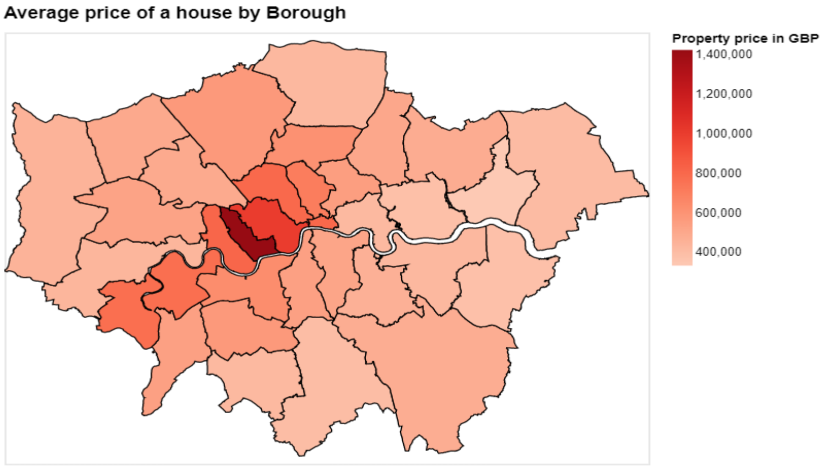
The relationship between the distance from the center of London and the Airbnb listing price had a

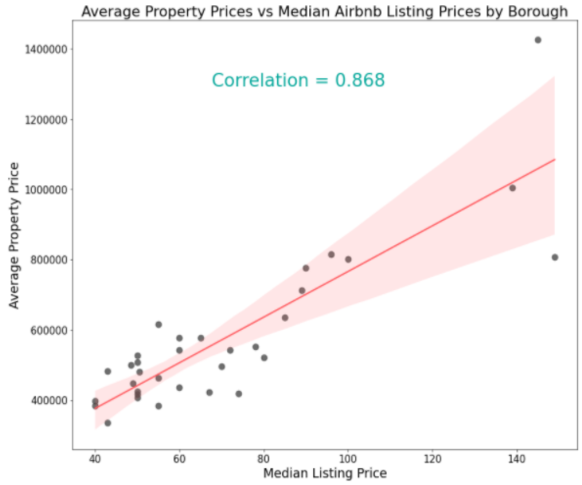
weak negative correlation, a correlation coefficient of only -**0.125**.

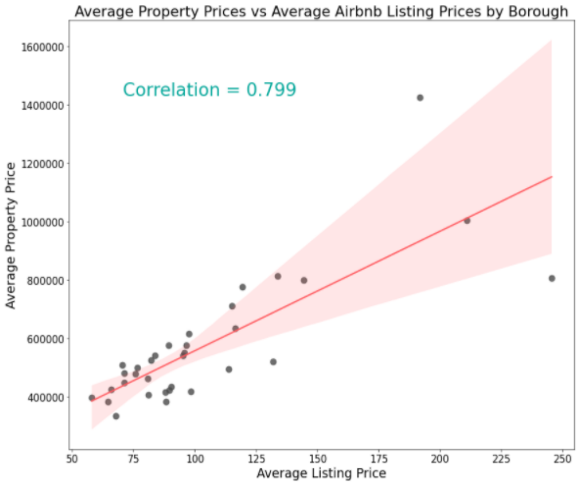
The relationship between listing price and proximity to tourist attraction and proximity to public transportation seems to be more important.

Listing price Vs. Average property price

Property prices across the London boroughs show high variance. While the average price of a house in Barking and Dagenham is £335,000, the average price in Kensington and Chelsea is over £1.4 million. How does the listing price of an airbnb property vary with the property price in the borough? To understand the relationship between property prices and listing prices, we aggregated the data based on borough level and merged primary and secondary data together for further visualization.

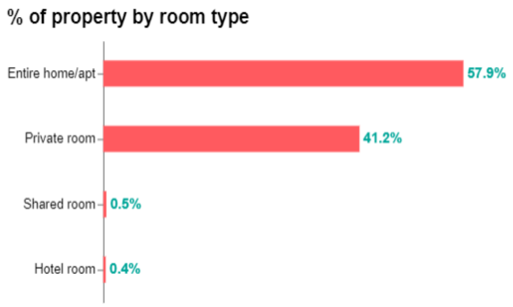

Average price of a house by Borough

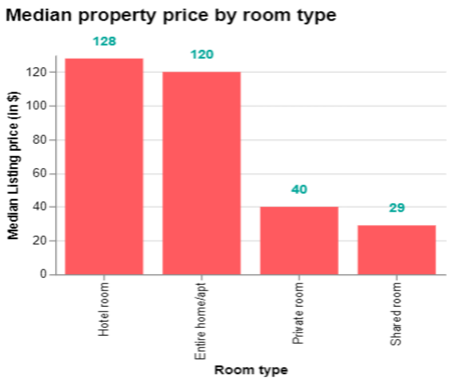

Average Property Prices vs Median Airbnb Listing Prices by Borough


Average Property Prices vs Average Airbnb Listing Prices by Borough

We observe a strong positive correlation between average property prices with both average and median airbnb prices. This observation is consistent with our expectation that listing prices will tend to be higher in most expensive areas, specifically in boroughs like the City of London, Westminster and Kensington and Chelsea. Please note that the Airbnb listing price is in USD and property prices are in GBP.

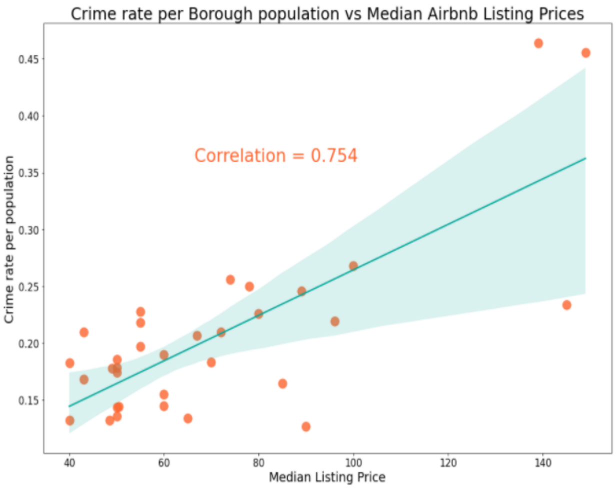
Listing price Vs. Room type





Majority of Airbnb properties in London are either Entire home/apt or Private rooms. Hotel room (median price = $128) and Entire room/apt (median price = $120) are much more expensive that private rooms (median price = $40) and shared rooms (median price = $20)

Listing price Vs. Crime rate by Borough



There is much literature on how criminal behavior imposes direct costs to the victim and indirect costs to society at large. We expected a strong negative correlation between crime rate and airbnb prices. In other words, listing prices in more dangerous neighbourhoods should considerably be lower. However, the data suggested otherwise. The regplot shows that crime rates (normalised by respective borough's 2011 census population) were higher in more expensive boroughs, exhibiting a strong positive correlation with price. There could be a possible confounding variable the analysis is missing such as borough housing vacancy, etc and would require a thorough investigation.

Building the Linear Regression Model (Multivariate Analysis)

Through extensive univariate analysis done in the exploratory stage, we decided on building a linear regression model to inform us of the multivariate relationship between these independent variables in determining their relative influence on airbnb prices.

Dummy variable creation

As the primary dataset consisted of non-numerical representations that include Boroughs, whether host is considered superhost, whether host identity is verified, etc, we needed to transform these categorical data into numeric variables to be used in the model. In order to avoid the scenario in which two or more independent variables are highly correlated (Dummy Variable Trap), the initial base regression equation would be the price of an airbnb listing located in the City of London, that is a private room, with host being superhost, having a profile picture, has host's identity verified and is instantly bookable. All dummy variables are binary and either take the value of 1 or 0.

Quantile-based Flooring and Capping

As the data is highly skewed, we have performed flooring (e.g. the 10th percentile) for the lower price values and capping (e.g. the 90th percentile) for the higher values for both the dependent and numerical independent variables. This will allow us to bring skewness down and increase model accuracy. Consideration of transforming the linear regression model into a log-linear regression was pondered over but such transformation may lead to overfitting (R^2 of 97%) and interpretability is compromised when trying to establish a meaningful multivariate relationship with log(price), instead of price.

Feature Selection Methodology

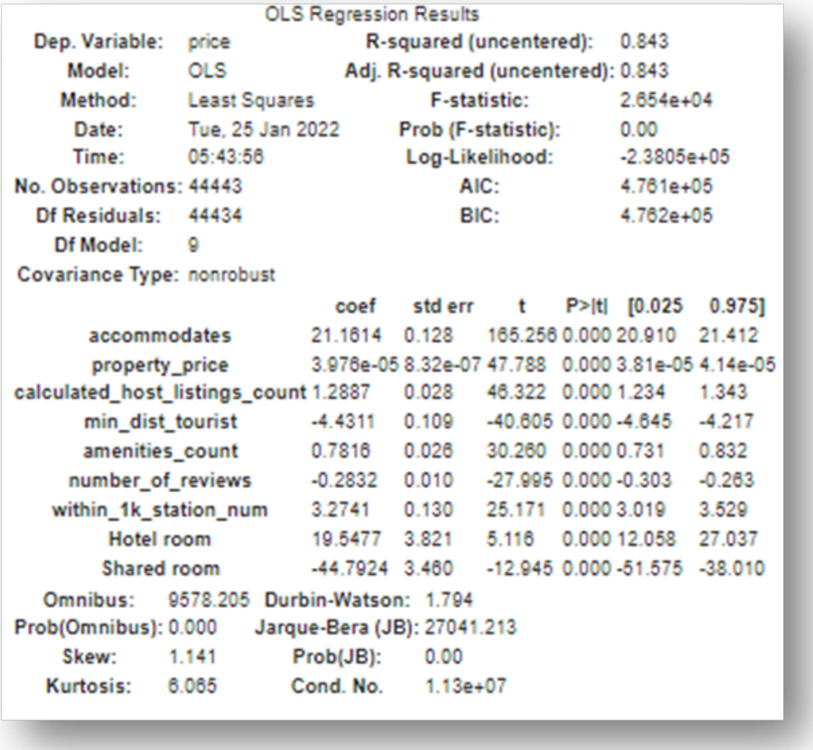
As our model consisted of many variables, there were several tests to be taken to ensure the model is consistent and unbiased. Several iterations of the run consisting of different independent variables guided by p-values, coefficient values, durbin-watson test (serial correlation test), multicollinearity tests using heatmap and Variance Inflation Factor (VIF), and residual probability plot were done.

Regression Model (no-intercept)

The final iteration of the model is used using the following regression equation:

$$\begin{aligned} \mathit{Listing\ price} = {} & \beta_1 \mathit{accomodates} + \beta_2 \mathit{property\_price} + \beta_3 \mathit{Calculated\_host\_listing\_count} \\ & + \beta_4 \mathrm{min}\ \mathit{\_dist\_tourist} \\ & + \beta_5\ \mathit{amenities\_count} + \beta_6 \mathrm{number\_of\_reviews} + \beta_7 \mathrm{within\_1k\_station\_num} \\ & + \beta_8 \mathrm{review\_\ scores\_\ rating} + \beta_9 \mathrm{Dummy}_{\mathit{Hotel\ Room}} + \beta_{10} \mathrm{Dummy}_{\mathit{Shared\ Room}} \end{aligned}$$

Model Summary



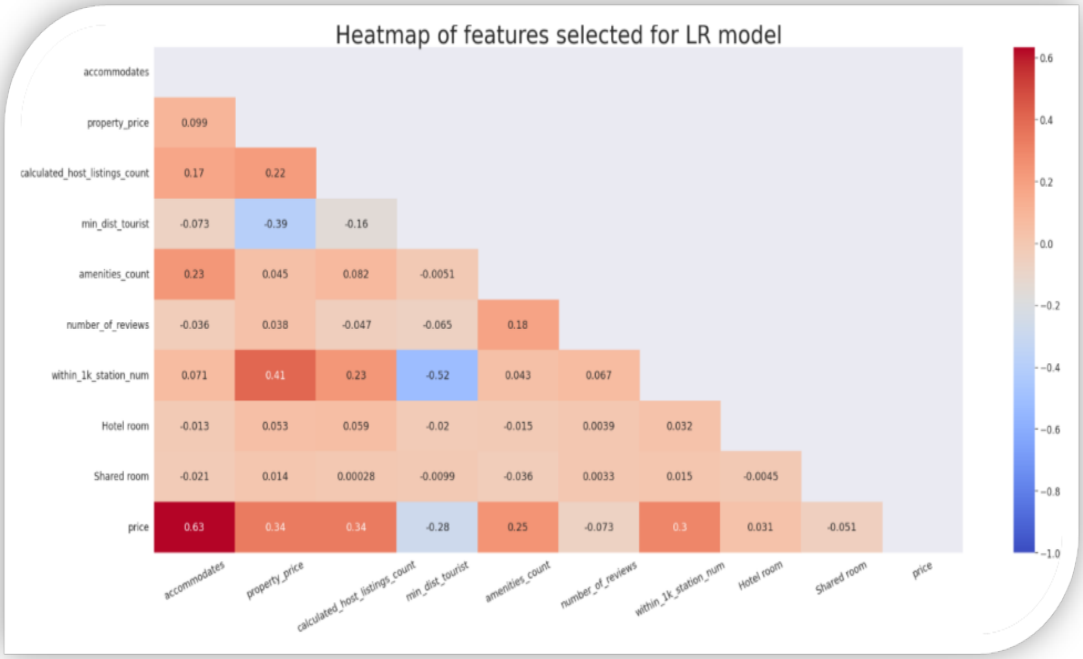
```
                         OLS Regression Results
    Dep. Variable:    price              R-squared (uncentered):      0.843
       Model:         OLS           Adj. R-squared (uncentered): 0.843
      Method:         Least Squares          F-statistic:         2.654e+04
        Date:         Tue, 25 Jan 2022    Prob (F-statistic):      0.00
        Time:         05:43:56            Log-Likelihood:         -2.3805e+05
No. Observations: 44443                        AIC:               4.761e+05
   Df Residuals:      44434                    BIC:               4.762e+05
     Df Model:        9
 Covariance Type: nonrobust
                                  coef     std err    t      P>|t|  [0.025    0.975]
          accommodates            21.1614  0.128    165.256 0.000 20.910    21.412
          property_price          3.976e-05 8.32e-07 47.788 0.000 3.81e-05 4.14e-05
calculated_host_listings_count 1.2887     0.028    46.322  0.000 1.234     1.343
          min_dist_tourist        -4.4311  0.109    -40.605 0.000 -4.645    -4.217
          amenities_count         0.7816   0.026    30.260  0.000 0.731     0.832
        number_of_reviews         -0.2832  0.010    -27.995 0.000 -0.303    -0.263
       within_1k_station_num      3.2741   0.130    25.171  0.000 3.019     3.529
            Hotel room            19.5477  3.821    5.116   0.000 12.058    27.037
           Shared room            -44.7924 3.460    -12.945 0.000 -51.575   -38.010
     Omnibus:      9578.205  Durbin-Watson:     1.794
Prob(Omnibus): 0.000     Jarque-Bera (JB): 27041.213
       Skew:       1.141        Prob(JB):        0.00
     Kurtosis:     6.065        Cond. No.        1.13e+07
```

All independent variables are statistically significant with p-values lower than level of significance even at 1%. In terms of autocorrelation, the **Durbin-Watson statistic** has the model at **1.8** which sits at the acceptable range of 1.5-2.5. There are however, **two main limitations** of the model. The **skewness is > 1 and kurtosis** > **6**. Nevertheless, the independent variables in the multiple regression model generalises the data well, indicating a good fit with an **R-squared value of 84.3%**

Model Interpretation

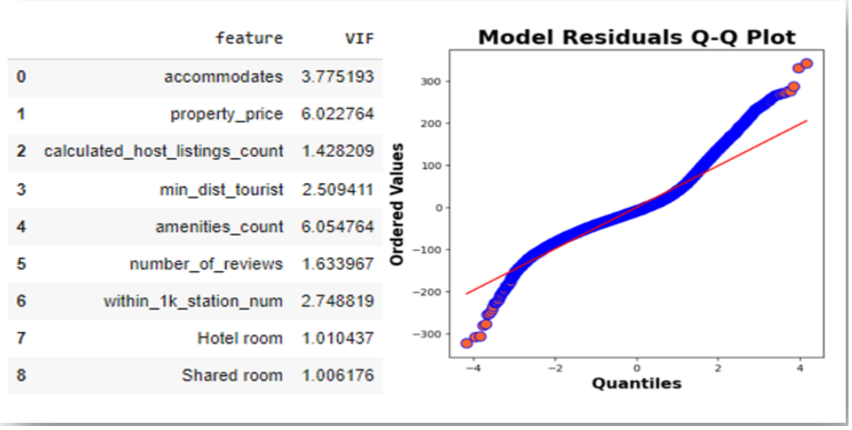
The results of the model show that **the number of people a listing is able to accommodate shares a strong positive correlation with price**, with the highest coefficient value. This reconciles with the expectation that a listing will likely be more expensive the bigger the listing being rented out is. The dummy variable "Hotel" indicates that airbnb properties located in **hotels are more considerably pricier** and are expected, on average, to be $20 higher than other room types. Additionally, airbnb properties with room type as **"Shared room" are cheaper**. We also observe that **listings are more expensive the closer it is to more stations** and **also shares a negative correlation with distance to the closest tourist attraction**. The number of amenities in an airbnb property does not have a huge positive influence on price with a coefficient variable of <1. Calculated host listing count refers to the number of listings a host has. It exerts a comparatively low positive influence on airbnb listing price. Last but not least, we can safely say that the number of reviews for each listing, albeit bearing a negative relationship, has little influence on the listing price.

Multicollinearity Heatmap



As multicollinearity can make it difficult to determine the effect of each predictor on the dependent variable, price, we used a heatmap to investigate that no predictor variables are highly correlated with one another. Otherwise, coefficients of the model may be poor estimations and coefficient signs may be hard to interpret. The heatmap shows that of above absolute values of 0.41.

Variance Inflation Factor (VIF) & Q-Q Residual Plot



	feature	VIF
0	accommodates	3.775193
1	property_price	6.022764
2	calculated_host_listings_count	1.428209
3	min_dist_tourist	2.509411
4	amenities_count	6.054764
5	number_of_reviews	1.633967
6	within_1k_station_num	2.748819
7	Hotel room	1.010437
8	Shared room	1.006176

Using **VIF** as an additional measure to quantify the severity of multicollinearity for our model, **all values are below 10**, indicating an **absence of collinearity**. However, the Q-Q Residual plot proves that despite quantile-based flooring and capping, the **distribution remains heavy-tailed**, as mentioned with the model's summary output having a skewness of 1.14.

Challenges & Conclusion

Exploratory data analysis indicated that there was a large variance in the price of Airbnb listings. Our analyses indicated that there were some factors that had a significant relationship with the price of the listings. Some of these variables include: property listings by borough, proximity to train stations, number of accommodations per listing, as well as the room type.

There were minimal ethical considerations that arose through our exploratory analyses and model building. The only potential ethical issue with the data is on the listing levels. In the Airbnb application, the exact address of the listing is hidden until the booking is confirmed. However, the dataset provides precise latitude and longitude for each listing, granting a much more specific location, rather than offering an area with a moderate radius that contains the location of the listing. The listings' exact location is hidden for security and privacy reasons, that may be undone by analyzing very specific locations.

Statement of Work

Michael Wynn: Setup of Google Colab environment. EDA of primary and secondary data. Multivariate Analysis (Regression Model). Proofing and re-writing of all documents. Source code compilation.
Scott Miketa: Directed Data join setup. Sentiment Analysis. EDA and visualizations of primary and secondary data. Report Design & Formatting. Proofing and editing of all documents.
Nishant Singh: Initiated project with data collection, manipulation, analysis, and visualizations on primary and secondary data. Distance calculation. Report Design & Formatting. Proofing and editing of all documents
All members: Writing content for the project proposal & final report