ST309 Group Project Report

**Analytics of AirBnB**

Candidate numbers: ???

1. **Introduction**
   1. **Preface**

In today’s world, technological advancements have paved the way to digitalise products and services at rates never seen before, transforming the way we live our lives. In the past, one might have needed to book hotel rooms through travel agents or phone calls - today, one can book a friendly stranger’s spare bedroom situated miles away through a mobile application easily and with confidence, with Airbnb as the broker.

Airbnb operates across 220 regions, 100,000 cities, and has almost 3 million hosts[[1]](#footnote-1). With this platform in mind, we have a trove of data with columns of reviews, ratings, amenities available, geographic location etc. across different cities. There are many questions we could explore from the point of view of an Airbnb lister - we will attempt to answer some of these questions, alongside data exploration, with statistical learning procedures we have learnt over the course of ST309.

* 1. **The Business Problem: How to Become a Superhost?**

The motivation behind our interest lies in the fact that there is a sizeable group of people interested to rent their homes through a service such as Airbnb - and a huge majority believes renting their homes on Airbnb is a good money making-strategy[[2]](#footnote-2).

A previous approach regressed prices against various attributes[[3]](#footnote-3) and found that prices positively correlate with locations. Property prices and upkeep vary with location, so it may be difficulty to judge true profit from prices alone. As such, our project will seek to explore another response variable: popularity.

* 1. **The Data**

Our main source of data is from the website [www.insideairbnb.com](http://www.insideairbnb.com), which has detailed scraped information obtained from Airbnb listings across multiple cities. We will use a snapshot of Airbnb listings data from London in 2019. There are 106 columns within a single dataset, so we removed NUMBER irrelevant columns in excel prior to loading the data into RStudio for further cleansing.

While the website is not associated nor endorsed by Airbnb or its competitors, we do note that the website’s provenance holds a somewhat anti-Airbnb stance.

1. **Data Cleansing**

**2.1 Removing Columns (Variables)**

With 106 columns and 86,469 listings available in the London December 2019 snapshot, we had to remove irrelevant variables via Excel to speed up the data loading and analysis process into RStudio. Variables were removed for the following reasons:

* Data not required for analysis
* Data is incomplete; not all listings’ data were successfully scraped

The full list of variables is available in **Appendix 1**.

We are left with 47 columns. Columns 5-15, all of which are character variables that contain text information of listings, were moved to a separate Excel file with a copy of column 1 **id** which serves as a unique identifier. This allows us to work on the text data separately from the numerical data, and greatly reduces the file size of the numerical dataset. Subsequent analysis of text information which requires the numerical data can be done by joining the two data sets on column 1.

**2.2 Removing Rows (Listings)**

Next, we removed records with incomplete information for analysis. These records could lack information due to failures in the data-scraping process. Finally, to reduce the amount of records we intend to analyse for computational reasons, we restricted ourselves to focus on listings that allow short-term stays. We thus remove records where:

* Host related data is blank
* Review scores are missing
* Minimum stay exceeds 3 days

We are left with 39,178 observations out of the original 86,469.

**2.3 Initial Transformations**

One important step was to normalise our listing prices; reason being that cleaning fees significantly[[4]](#footnote-4) increase the true price of listings. We summed the 3-day pro-rata rates with the cleaning fee to obtain a 3-day price per person, which is subsequently divided by the number of guests allowed as per the listing. This value is found to have skewness, so it was normalised with a log transformation. For our exploration and analysis, we will use this newly calculated variable **price\_n** for price-related analysis.

Significant transformation was required for many other variables before they could be used for analysis. These variable transformations are detailed in the markdown file, and usage is explained over the next section whenever the variables appear.

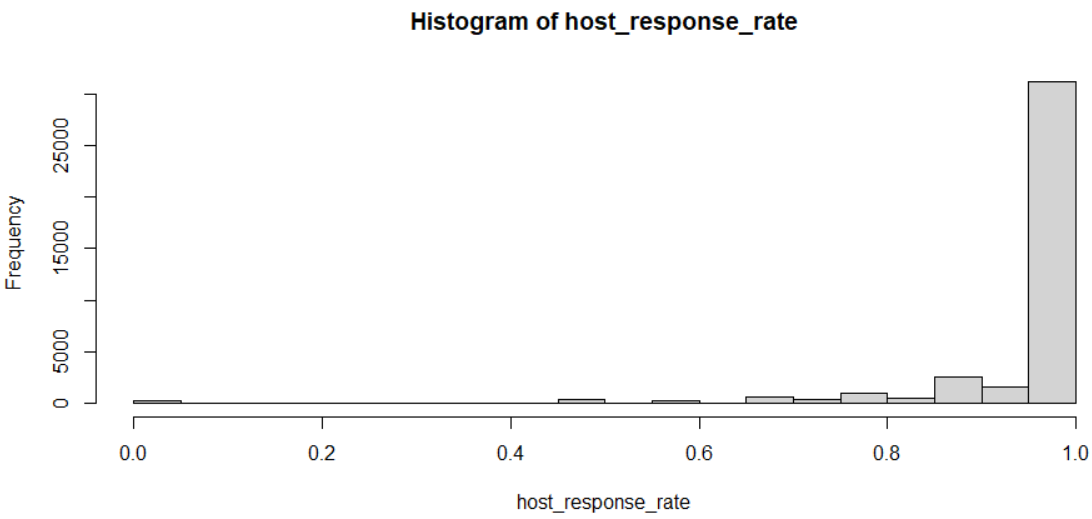
1. **Data Exploration**

We begin our analysis by understanding the listings better.

**3.1 Host-related Data**

Host-related variables allow us to glean into the characteristics of the hosts behind Airbnb listings, such as their response rates and if they were Superhosts. To obtain the number of years the host has operated - **host\_since\_n** was calculated by subtracting **host\_since** from the current date. A factor representing the different response level of hosts is also available as **host\_response\_time\_n**.

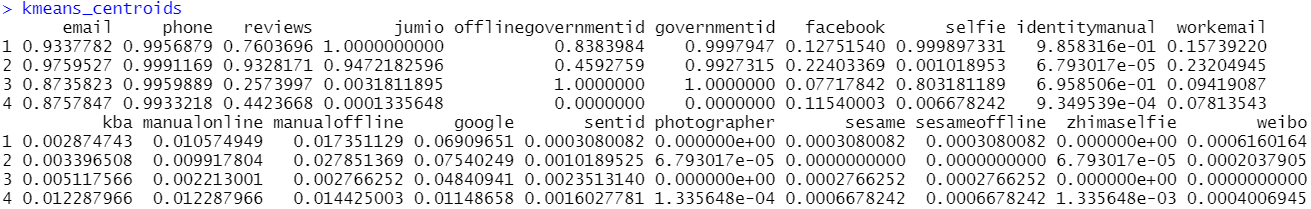
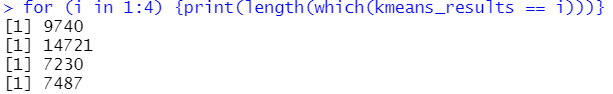
A factor dummy variable representing if hosts have a 100% response rate were also coded from host\_response\_rate as **host\_response\_rate\_n**. We decided to use such a factor because there was a significant number of hosts with a perfect response rate. In fact, the median for host\_response\_rate was at 100%.



**Figure 1: Histogram of host\_response\_rate** *(high frequency of prefect rates)*

Airbnb has a Superhost system where top listers providing consistent positive experiences have the chance of being awarded with the **Superhost** status[[5]](#footnote-5). This was coded as **superhost\_n**. of 29,791 listings, 9,387 were under Superhosts.

As for the types of verifications used by hosts, we had to significantly recode the original character-type data. We proceeded to run a k-means clustering analysis to check for patterns, if any, in verification preferences:



**Figure 2: Host Verification Clusters**

It turns out that verification through email and phone was incredibly popular with all hosts – expected given that these are required when operating an Airbnb. Jumio also happens to be a rather popular platform for identification, where about 24,000 listings’ (out of 39,178) hosts belonged to the clusters that used Jumio. Government ID verification is also popular, but about 7,487 listings’ hosts chose not to verify with their government ID’s. Selfie verification is also utilised in two of the clusters.

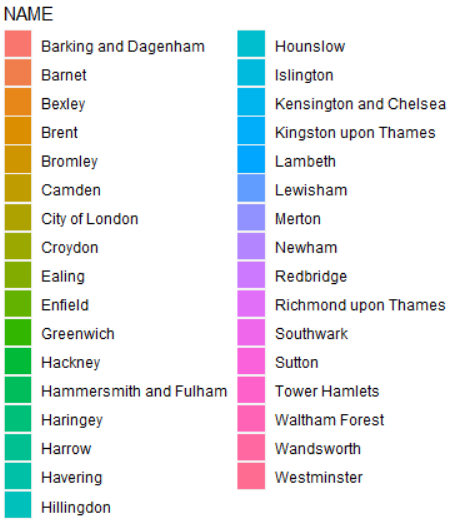
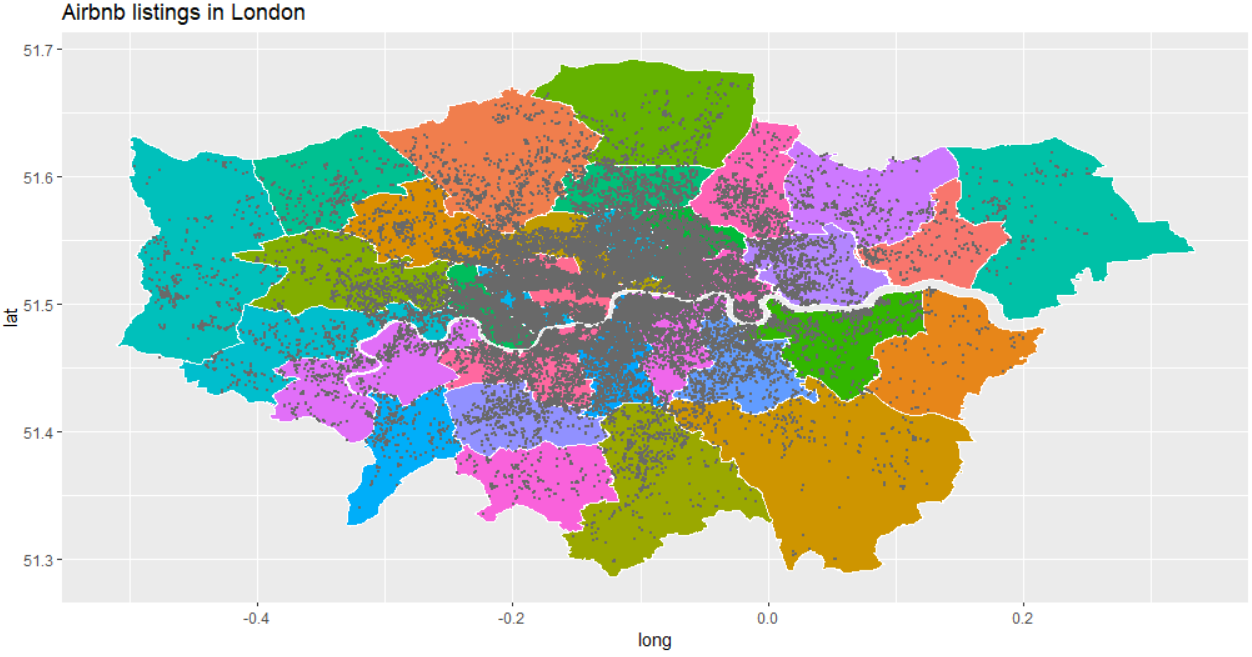
To summarise:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **Email/Phone** | **Jumio** | **Government ID** | **Selfie** |
| 1 (9,740) | ◯ | ◯ | ◯ | ◯ |
| 2 (14,721) | ◯ | ◯ | ◯ | ❌ |
| 3 (7,230) | ◯ | ❌ | ◯ | ◯ |
| 4 (7,487) | ◯ | ❌ | ❌ | ❌ |

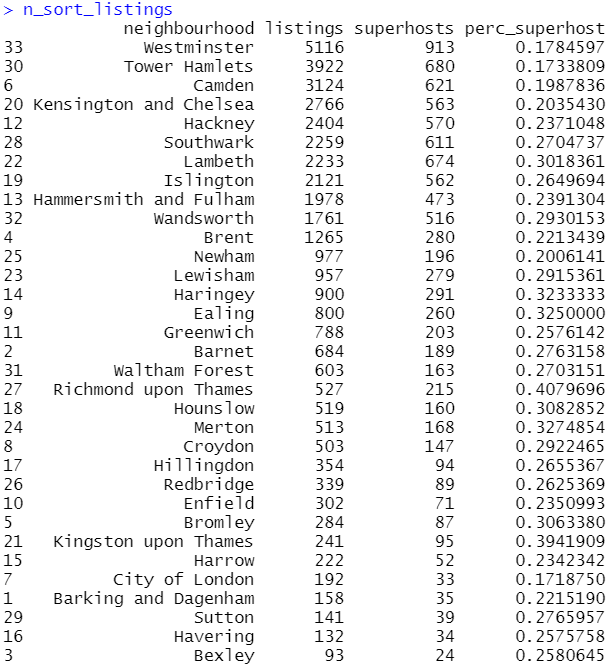
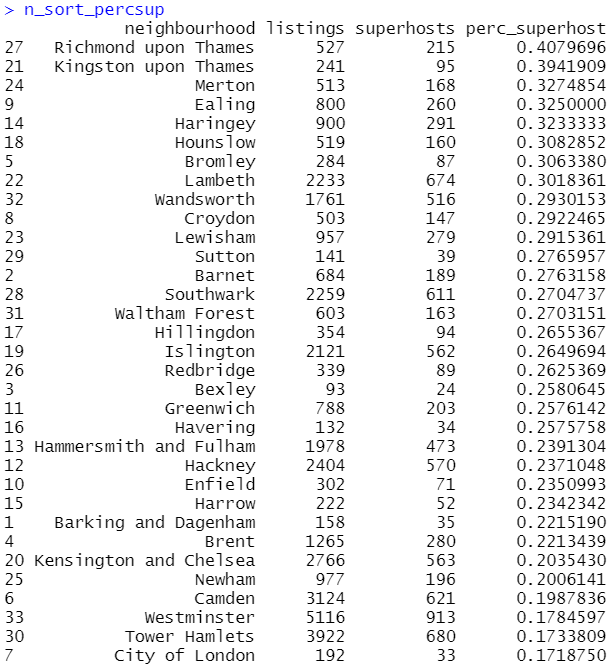
**Figure 3: Host Verification Clusters, Approximated Verification Types**

**3.2 Location Data**

We processed latitudinal and longitudinal data to create a map:



**Figure 4: Map of Airbnb Listings**

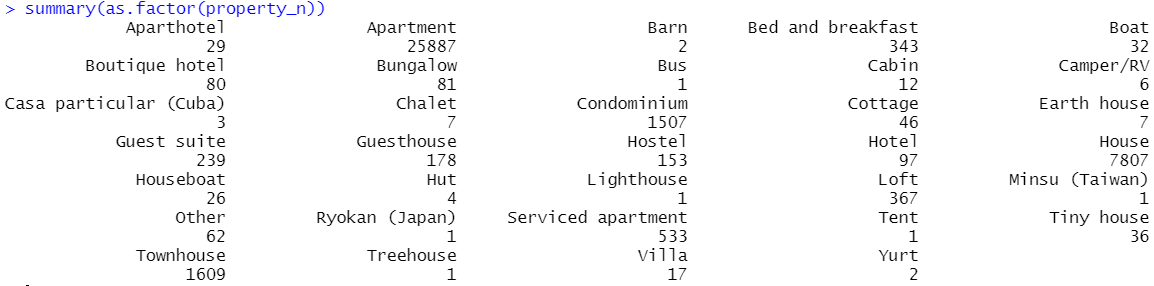


**Figure 5: Boroughs Ranked by Number of Listings and Percentage of Superhosts**

It appears that Westminster, Tower Hamlets, Camden, Kensington and Chelsea, and Hackney are top locations for Airbnb properties. Interestingly, areas with higher Superhost percentages tend to have less listings: Richmond upon Thames and Kingston upon Thames are not ranked high by the number of listings. Perhaps a good amount of engagement and customer satisfaction (hence the Superhost statuses) is required to survive in these somewhat less popular regions further from Central London.

**3.3 Property Data**

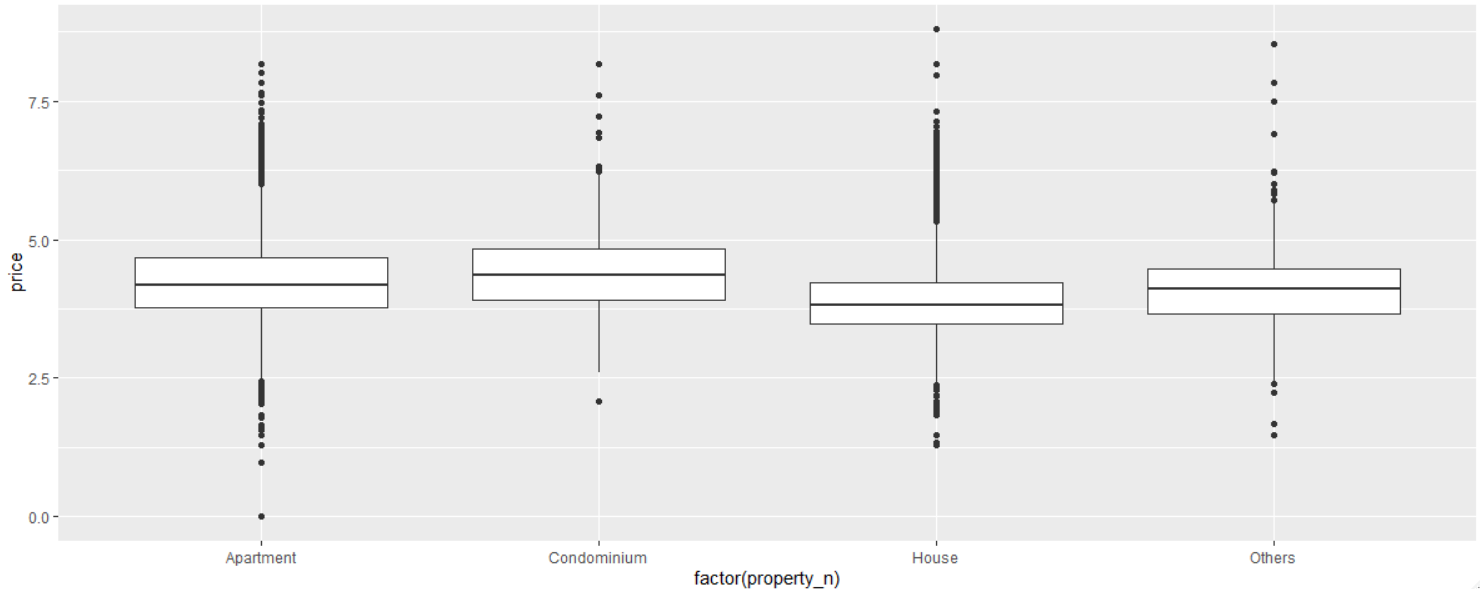
Property types were investigated:



Similar properties were then grouped together to simplify our dataset from 34 types to 4 types:



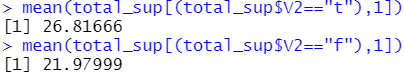
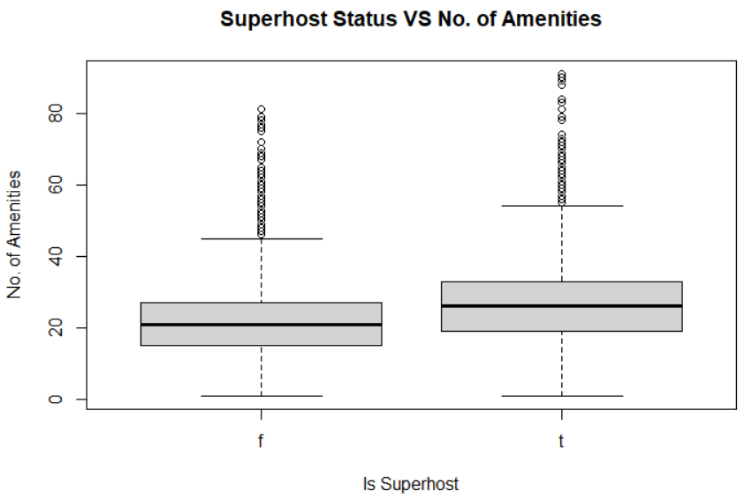
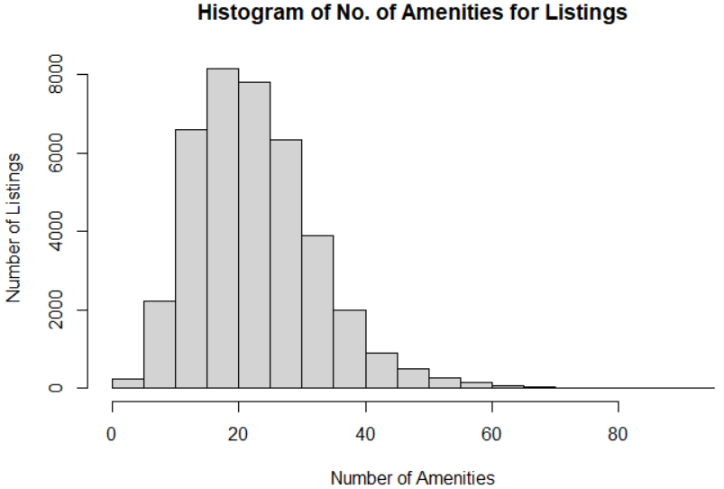
We attempted to check for price differences between the property types. Condominiums tend to have higher prices, which could be explained by their additional amenities and features, whilst houses tend to have lower prices. However, we do note that prices of all property types are within one another’s range – perhaps alternative factors such as location can explain price variation better.



**Figure 6: Box-plot of Normalised Prices by Property Type**

**3.4 Amenities**

In our data, **amenities** contains the a string of amenities marketed in each listing. Thorough data cleaning and transformation was required to recode the data for exploratory analysis:



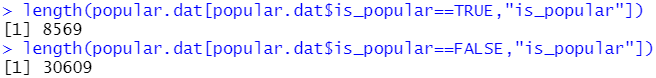
**Figure 7: Distribution of Amenities and Superhost VS Amenities**

It turns out that most listings tend to have about 20 amenities listed, and Superhosts happen to have 5 more listed amenities. Perhaps it would be important to communicate a property’s amenities fully to potential guests.

**3.5 Popularity as a Response Variable**

Since a previous study had explored prices, we aimed to study a different response variable, hoping to have newer and different insights.

Our dataset includes data on number of reviews per month (reviews\_per\_month) and the average review score (review\_scores\_rating) for our listings. We coded a dummy variable **is\_popular** conditional on these two variables, where a listing is popular if both variables are above their respective medians.



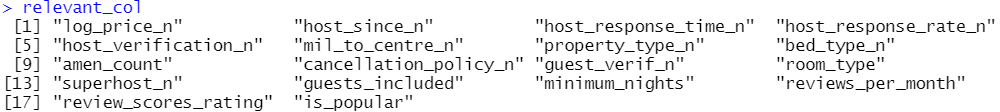
Out of 39,178 listings, a healthy number of 8,569 (21.9% of listings) are considered popular.

1. **Data Analysis with Models**

With exploratory analysis and data transformations performed, we can now try to check relationships between our variables to see if there are any meaningful insights.

We removed non-relevant columns that do not contain information of interest, such as identifiers like **id**, **host\_id**, **host\_name** and character variables like **name, summary, host\_about**, etc. Some variables tend to share characteristics or were transformed, so we have chosen a single variable to represent their similar counterparts to prevent multicollinearity. For example, **mil\_to\_centre\_n** is taken instead of **latitude** and **longitude**, and **price\_n** is used instead of **price** and **cleaning\_fee**. The full selection process is detailed in the markdown file.

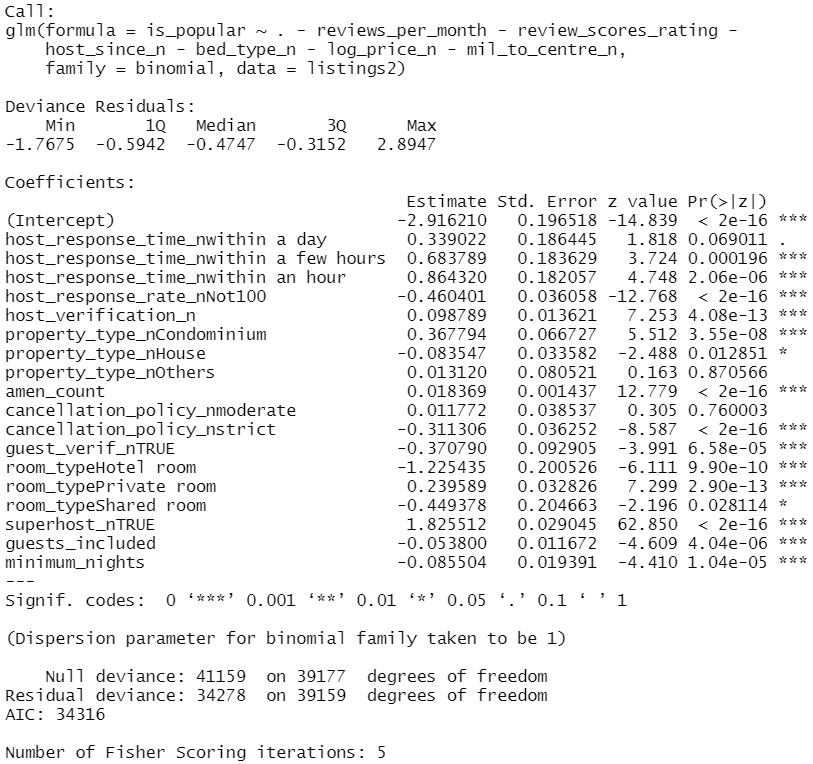
The relevant columns are:



* 1. **Popularity in the Generalised Linear Model**

Our specification sets **is\_popular** as the response variable to the remaining 15 variables – the two variables used to create is\_pouplar are naturally left out of the GLM. We remove statistically insignificant variables one by one, resulting in the removal of **host\_since\_n, bed\_type\_n, log\_price\_n, mil\_to\_centre\_n**.

Jointly-significant variables with multiple factors such as **property\_type\_n** remain in the model, whilst individually insignificant variables such as **log\_price\_n** were removed. The variable removal process is documented in the markdown. The final model is presented here:



**Figure 8: GLM Model 5 Results**

There are some important insights from our model. Responsiveness of the host appears to have a strong positive correlation with popularity – the faster a host reponds (**within an hour** as opposed to **within a few days**, which is the baseline), the more popular their listing. Perfect response rates are also positively correlated with listing popularity, evident from the negative coefficient of not achieving 100% (**host\_response\_rate\_nNot100**).

The more verified the host happens to be, the more popular their listings. This could be due to the fact that guests prefer verified hosts, but we should be wary that it is possible for host effort to be a confounder here; for example, dedicated hosts willing to respond to guests quickly and provide greater service might also put in more effort to verify themselves.

As for the property type, **condominiums** appear to be more popular, while **houses** are the least popular. A greater number of **amenities** also positively correlates with popularity, which does make sense. Even **cancellation** policy appears to matter, as moderate cancellation policy positively correlate with popularity, while strict cancellation policy is correlated with lower popularity.

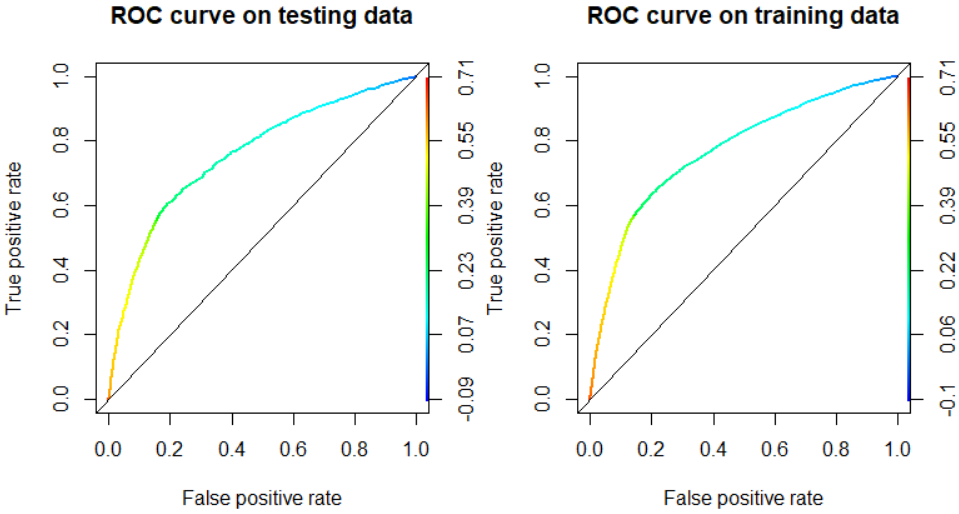
Interestingly, a requirement for **guest verification** is negatively correlated with popularity. It could be that additional requirements for guests to have extensive verifications can be off-putting or troublesome, resulting in lower popularity.

For the room type, it might appear that guests preferred and liked **private rooms** more so than **shared rooms**, **entire home/apt**. The strong negative correlation between popularity and **hotel rooms** is understandable as most guests log into Airbnb to find non-hotel options.

**Superhosts** positively correlate with popularity, which is expected given that the Superhost status is awarded to well-performing Airbnb hosts.

The negative correlations of **guests\_included** and **minimum\_nights** suggest that Airbnb options for smaller groups and shorter stays tend to be more popular. It might be important to note that a smaller **minimum\_nights** might structurally produce more reviews, hence giving listings a higher popularity index. This is because for two properties with the same level of demand, the property available for shorter stays will tend to have more distinct guests over the same period of time; for example, a 3-night minimum property can take 2 guests over a week, while a 1-night minimum property can take 7 guests over a week. Since we designed popularity to be calculated from both review score and number of reviews, we should keep this possibility in mind.

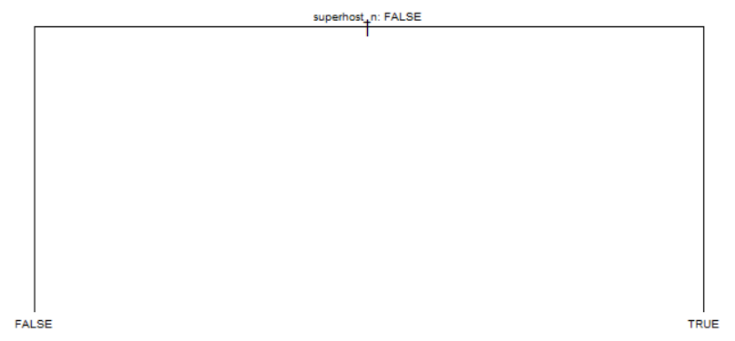
We further evaluated the performance of fa prediction model based on this GLM model’s fitted values with ROC curves. As shown below, the model can have better prediction rate than random guesses (45 degree line).

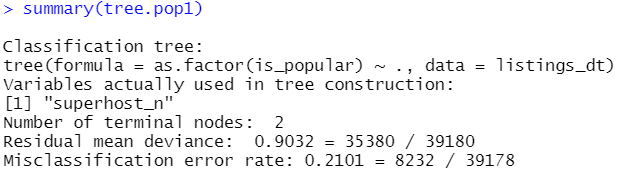


**Figure 9: ROC Curves of GLM Prediction Model**

* 1. **Popularity via Decision Tree**

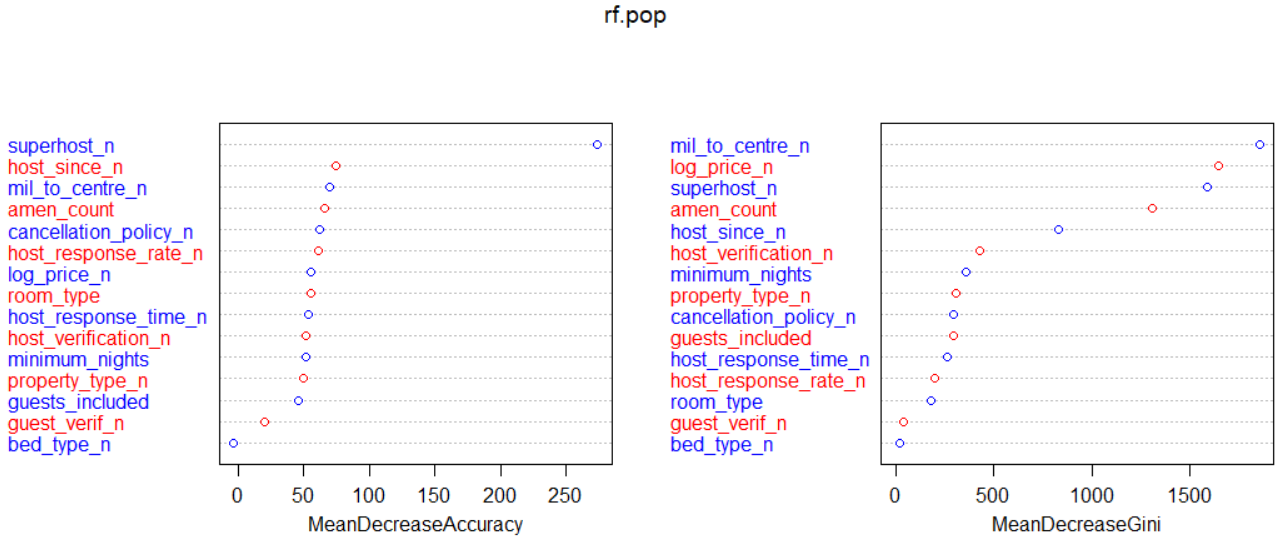
Next, we attempted to classify our listings into whether they were popular or not with a decision tree. Unfortunately, we were left with somewhat trivial results:





**Figure 10: Classification Tree**

With an extremely high residual mean deviance, it was incredibly challenging to classify the listings as popular or not with our current variables. We attempted various ways to improve the decision tree, such as via bagging and random forests, but still failed to achieve a decent hit-rate for popular listings.



**Figure 11: Random Forest**

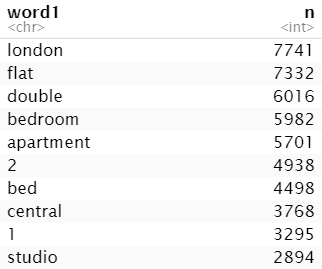
From our random forest analysis, **superhost\_n, mil\_to\_centre\_n, log\_price\_n, amen\_count,** and **host\_since\_n** appear to be relatively more important variables, as excluding these variables either lead to a relatively strong decrease in model accuracy or a decrease in homogeneity of nodes. Unfortunately, the error rate of predicting popular listings remains high at 61.5%.

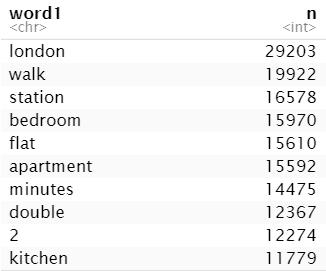
**4.4 Price Analysis with MLR**

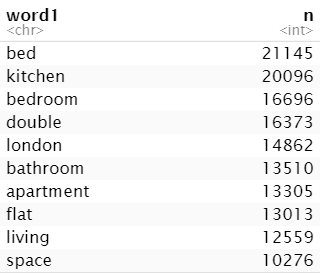
**5 Data Analysis with Text**

Airbnb listing descriptions can be crucial for successful listings[[6]](#footnote-6). With some basic text analysis techniques, we can have a rough understanding of how Airbnb listings’ text descriptions look like. The variables **name**, **summary**, **space**, **description, host\_about** are worth looking into, as these are the blocks of text that users first get to read on any listing.

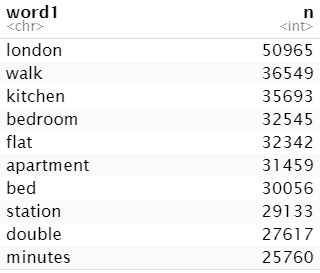
**Name**, **summary**, and d**escription** generally have words linked to apartment and room types. A commonly used word is “minutes”, which suggests that Airbnb listers tend to advertise their listings’ proximities to certain locations, presumably public transportation or areas of interest. Another interesting discovery was how common the word “love” appears in [host\_about], which suggests that it is common for Airbnb hosts to use words to elicit emotional reception (words such as “enjoy”, “happy”, “comfortable” also appear). It is clear that text descriptors need to be colourful and marketable.

**Figure 12a: [Name] Word Cloud and Top Words**

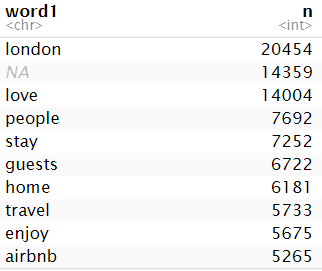
**Figure 12a: [Summary] Word Cloud and Top Words**

**Figure 12c: [Space] Word Cloud and Top Words**

**Figure 12d: [Description] Word Cloud and Top Words**



**Figure 12e: [Host\_about] Word Cloud and Top Words**

We did the same procedure for two groups of listings: popular and non-popular (as previously specified) to see if there might be insightful details about the way hosts write their listing descriptors. The results can be found in the markdown file and are mostly unremarkable, hence not further described in the report. Some intriguing points include:

* In [name], “private” is ranked 6th for in the popular group, but 14th for the non-popular group, which suggests that privacy or private rooms appear to be quite the catch

1. **Concluding Remarks**

**6 Appendix**

**Appendix 1: Table of Variables**

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Column** | **Details** | **Proposed Data Type** |
| 1 | id | Unique identifier | Integer |
| 2 | ~~listing\_url~~ |  | Not required |
| 3 | ~~scrape\_id~~ |  | Not required |
| 4 | ~~last\_scraped~~ |  | Not required |
| 5 | name | Name of listing | Character |
| 6 | summary | Short summary | Character |
| 7 | space | Introduction to space | Character |
| 8 | description | Introduction to listing | Character |
| 9 | experiences\_offered | Not required | Character |
| 10 | neighborhood\_overview | Introduction to neighbourhood | Character |
| 11 | notes | Other notes | Character |
| 12 | transit | Information on transportation | Character |
| 13 | access | Information to get to listing | Character |
| 14 | interaction | Information on how much exposure the host prefers | Character |
| 15 | house\_rules | House rules | Character |
| 16 | ~~thumbnail\_url~~ |  | Not required |
| 17 | ~~medium\_url~~ |  | Not required |
| 18 | ~~picture\_url~~ |  | Not required |
| 19 | ~~xl\_picture\_url~~ |  | Not required |
| 20 | host\_id | Unique identifier | Integer |
| 21 | ~~host\_url~~ |  | Not required |
| 22 | ~~host\_name~~ |  | Not required |
| 23 | host\_since | Date when host joined Airbnb | Date |
| 24 | ~~host\_location~~ |  | Not required |
| 25 | host\_about | Host introduction | Character |
| 26 | host\_response\_time | Host response time | Factor/numeric/dummy |
| 27 | host\_response\_rate | Host response rate | Factor/numeric/dummy |
| 28 | ~~host\_acceptance\_rate~~ |  | Not required |
| 29 | host\_is\_superhost | If host is a superhost | Factor/numeric/dummy |
| 30 | ~~host\_thumbnail\_url~~ |  | Not required |
| 31 | ~~host\_picture\_url~~ |  | Not required |
| 32 | ~~host\_neighbourhood~~ |  | Not required |
| 33 | ~~host\_listings\_count~~ |  | Not required |
| 34 | ~~host\_total\_listings\_count~~ |  | Not required |
| 35 | host\_verifications | Types of verifications host has | Factor/numeric/dummy |
| 36 | host\_has\_profile\_pic | If host has a profile picture | Factor/numeric/dummy |
| 37 | host\_identity\_verified | If host's identity has been verified | Factor/numeric/dummy |
| 38 | ~~street~~ |  | Not required |
| 39 | ~~neighbourhood~~ |  | Not required |
| 40 | neighbourhood\_cleansed | Address details | Character |
| 41 | ~~neighbourhood\_group\_cleansed~~ |  | Not required |
| 42 | ~~city~~ |  | Not required |
| 43 | ~~state~~ |  | Not required |
| 44 | ~~zipcode~~ |  | Not required |
| 45 | ~~market~~ |  | Not required |
| 46 | ~~smart\_location~~ |  | Not required |
| 47 | ~~country\_code~~ |  | Not required |
| 48 | ~~country~~ | Address details | Not required |
| 49 | latitude | Address details | Numeric |
| 50 | longitude | Address details | Numeric |
| 51 | ~~is\_location\_exact~~ |  | Not required |
| 52 | property\_type | Property type; house, apartment etc. | Factor/character/dummy |
| 53 | room\_type | Room type; private room, whole house etc. | Factor/character/dummy |
| 54 | ~~accommodates~~ |  | Not required |
| 55 | ~~bathrooms~~ |  | Not required |
| 56 | ~~bedrooms~~ |  | Not required |
| 57 | ~~beds~~ |  | Not required |
| 58 | bed\_type | Type of bed | Factor/character/dummy |
| 59 | amenities | Amenities available | Factor/character/dummy |
| 60 | ~~square\_feet~~ |  | Not required |
| 61 | price | Daily price | Numeric |
| 62 | ~~weekly\_price~~ |  | Not required |
| 63 | ~~monthly\_price~~ |  | Not required |
| 64 | ~~security\_deposit~~ |  | Not required |
| 65 | cleaning\_fee | Cleaning fee | Numeric |
| 66 | guests\_included | Guests included in the price | Numeric |
| 67 | ~~extra\_people~~ |  | Not required |
| 68 | minimum\_nights | Minimum number of nights per booking | Numeric |
| 69 | ~~maximum\_nights~~ |  | Not required |
| 70 | ~~minimum\_minimum\_nights~~ |  | Not required |
| 71 | ~~maximum\_minimum\_nights~~ |  | Not required |
| 72 | ~~minimum\_maximum\_nights~~ |  | Not required |
| 73 | ~~maximum\_maximum\_nights~~ |  | Not required |
| 74 | ~~minimum\_nights\_avg\_ntm~~ |  | Not required |
| 75 | ~~maximum\_nights\_avg\_ntm~~ |  | Not required |
| 76 | ~~calendar\_updated~~ |  | Not required |
| 77 | ~~has\_availability~~ |  | Not required |
| 78 | ~~availability\_30~~ |  | Not required |
| 79 | ~~availability\_60~~ |  | Not required |
| 80 | ~~availability\_90~~ |  | Not required |
| 81 | ~~availability\_365~~ |  | Not required |
| 82 | ~~calendar\_last\_scraped~~ |  | Not required |
| 83 | number\_of\_reviews | Number of reviews in total | Numeric |
| 84 | number\_of\_reviews\_ltm | Number of reviews in the last twelve months | Numeric |
| 85 | ~~first\_review~~ |  | Not required |
| 86 | ~~last\_review~~ |  | Not required |
| 87 | review\_scores\_rating | Review score; total rating | Numeric |
| 88 | review\_scores\_accuracy | Review score; accuracy | Numeric |
| 89 | review\_scores\_cleanliness | Review score; cleanliness | Numeric |
| 90 | review\_scores\_checkin | Review score; check-in | Numeric |
| 91 | review\_scores\_communication | Review score; communication | Numeric |
| 92 | review\_scores\_location | Review score; location | Numeric |
| 93 | review\_scores\_value | Review score; value | Numeric |
| 94 | ~~requires\_license~~ |  | Not required |
| 95 | ~~license~~ |  | Not required |
| 96 | ~~jurisdiction\_names~~ |  | Not required |
| 97 | ~~instant\_bookable~~ |  | Not required |
| 98 | ~~is\_business\_travel\_ready~~ |  | Not required |
| 99 | cancellation\_policy | Cancellation policy; moderate, strict etc. | Factor/character/dummy |
| 100 | require\_guest\_profile\_picture | If host requires guest to have profile picture | Factor/character/dummy |
| 101 | require\_guest\_phone\_verification | If host requires guest to have verified phone number | Factor/character/dummy |
| 102 | ~~calculated\_host\_listings\_count~~ |  | Not required |
| 103 | ~~calculated\_host\_listings\_count\_entire\_homes~~ |  | Not required |
| 104 | ~~calculated\_host\_listings\_count\_private\_rooms~~ |  | Not required |
| 105 | ~~calculated\_host\_listings\_count\_shared\_rooms~~ |  | Not required |
| 106 | reviews\_per\_month | Number of reviews obtained per month | Numeric |

1. <https://www.stratosjets.com/blog/airbnb-statistics/#:~:text=How%20Many%20Users%20Does%20Airbnb,in%20an%20Airbnb%20every%20night> [↑](#footnote-ref-1)
2. <https://www.cnbc.com/2019/07/03/is-running-an-airbnb-profitable-heres-what-you-need-to-know.html#:~:text=Airbnb%20hosts%20make%2C%20on%20average,and%20the%20services%20you%20provide> [↑](#footnote-ref-2)
3. <https://towardsdatascience.com/how-to-maximize-profits-on-airbnb-data-based-approach-for-hosts-beaf08f26941> [↑](#footnote-ref-3)
4. <https://www.buzzfeednews.com/article/carolineodonovan/why-airbnbs-cost-more-extra-cleaning-fees> [↑](#footnote-ref-4)
5. <https://www.airbnb.com.sg/help/article/828/what-is-a-superhost> [↑](#footnote-ref-5)
6. <https://www.airbnb.com.sg/resources/hosting-homes/a/write-an-appealing-listing-description-13?_set_bev_on_new_domain=1613396592_N2I0ZjE4NjZjZjZm> [↑](#footnote-ref-6)