ST309 Group Project Report

Analytics of Airbnb

Candidate numbers:

15702 50%

18196 50%

ST309 Project Report: Analytics of AirBnB

Introduction

1.1 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]. 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 aspects we could explore, from the point of view of a budding Airbnb lister, through statistical learning procedures gathered over ST309.

1.2 The Business Problem

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].

A previous approach found online regressed prices against various attributes[3] and found that prices positively correlate with locations amongst other factors. Property prices and upkeep vary with location, so it may be difficult to judge true profit from prices alone. Our project will seek to explore another response variable: popularity. We would like to check for possible relationships between popular listings and our variables at hand, and see if there are any insights available.

1.3 The Data

Our main source of data is from the website 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 59 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] https://www.stratosjets.com/blog/airbnb-statistics/#:~:text=How%20Many%20Users%20Does%20Airbnb,in%20an%20Airbnb%20every%20night
[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
[3] https://towardsdatascience.com/how-to-maximize-profits-on-airbnb-data-based-approach-for-hosts-beaf08f26941

2. 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.

2.2 Removing Listings (Rows)

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] 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.

[4] https://www.buzzfeednews.com/article/carolineodonovan/why-airbnbs-cost-more-extra-cleaning-fees

3. 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%.

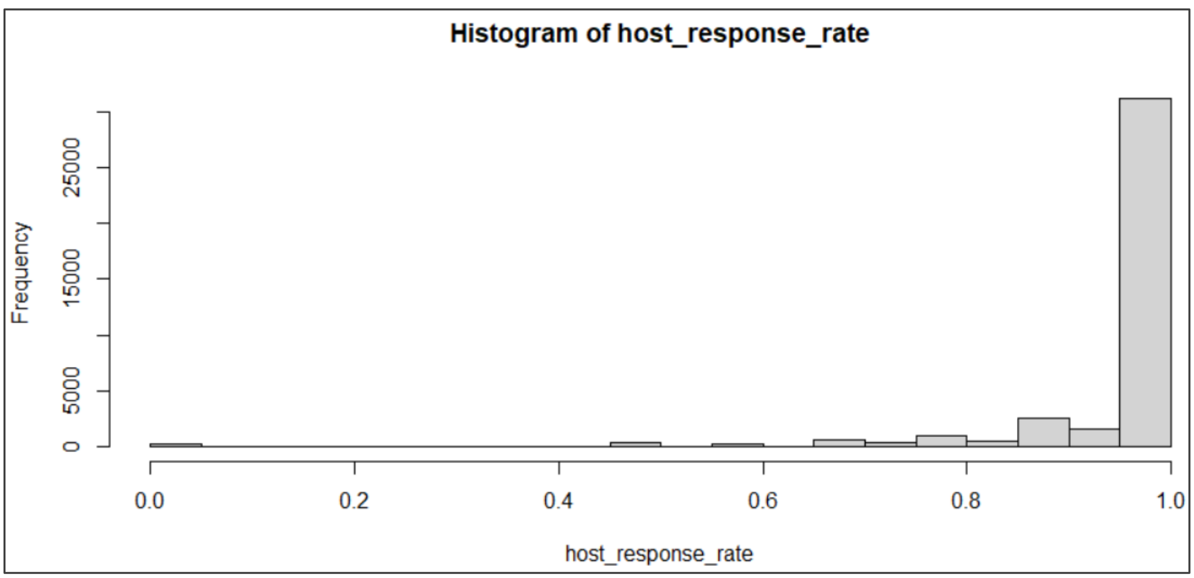


Figure 1: Histogram of host_response_rate *(high frequency of prefect rates)*

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

[5] https://www.airbnb.com.sg/help/article/828/what-is-a-superhost

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:

```
> kmeans_centroids
      email     phone   reviews        jumio offlinegovernmentid governmentid   facebook      selfie identitymanual  workemail
1 0.9337782 0.9956879 0.7603696 1.0000000000           0.8383984    0.9997947 0.12751540 0.999897331   9.858316e-01 0.15739220
2 0.9759527 0.9991169 0.9328171 0.9472182596           0.4592759    0.9927315 0.22403369 0.001018953   6.793017e-05 0.23204945
3 0.8735823 0.9959889 0.2573997 0.0031811895           1.0000000    1.0000000 0.07717842 0.803181189   6.958506e-01 0.09419087
4 0.8757847 0.9933218 0.4423668 0.0001335648           0.0000000    0.0000000 0.11540003 0.006678242   9.349539e-04 0.07813543
          kba manualonline manualoffline     google       sentid photographer       sesame sesameoffline  zhimaselfie        weibo
1 0.002874743  0.010574949   0.017351129 0.06909651 0.0003080082 0.000000e+00 0.0003080082  0.0003080082 0.000000e+00 0.0006160164
2 0.003396508  0.009917804   0.027851369 0.07540249 0.0010189525 6.793017e-05 0.0000000000  0.0000000000 6.793017e-05 0.0002037905
3 0.005117566  0.002213001   0.002766252 0.04840941 0.0023513140 0.000000e+00 0.0002766252  0.0002766252 0.000000e+00 0.0000000000
4 0.012287966  0.012287966   0.014425003 0.01148658 0.0016027781 1.335648e-04 0.0006678242  0.0006678242 1.335648e-03 0.0004006945
```

```
> for (i in 1:4) {print(length(which(kmeans_results == i)))}
[1] 9740
[1] 14721
[1] 7230
[1] 7487
```

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:

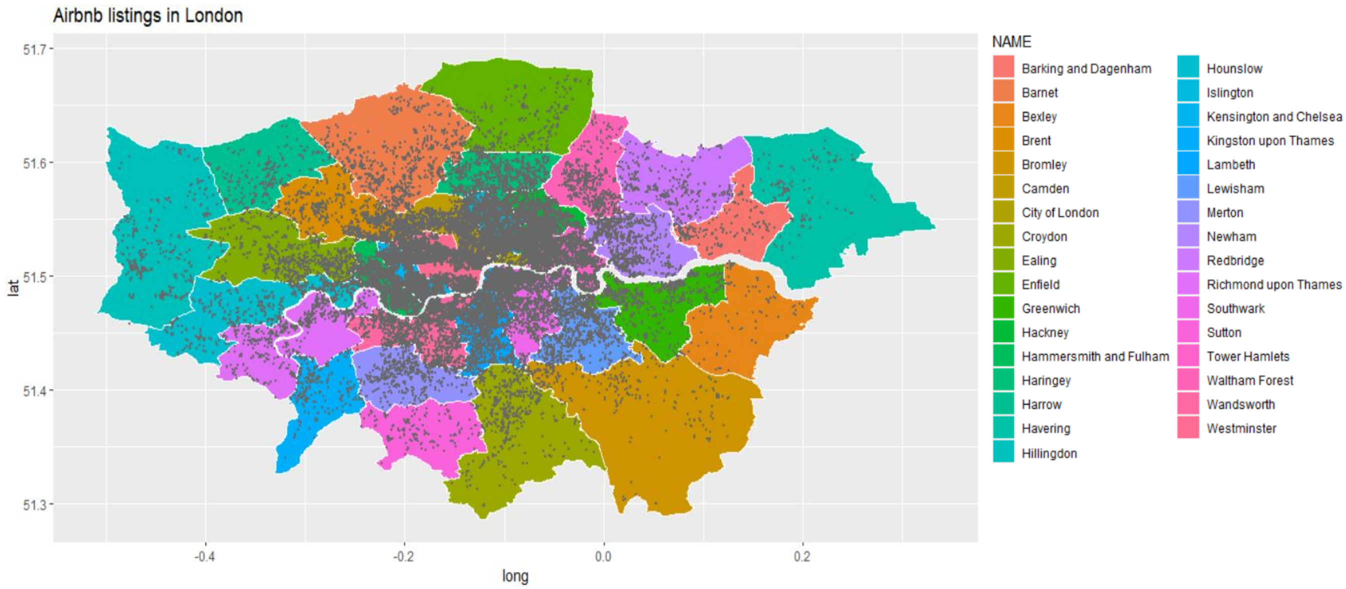


Figure 4a: Map of Airbnb Listings

> n_sort_percsup

	neighbourhood	listings	superhosts	perc_superhost
27	Richmond upon Thames	527	215	0.4079696
21	Kingston upon Thames	241	95	0.3941909
24	Merton	513	168	0.3274854
9	Ealing	800	260	0.3250000
14	Haringey	900	291	0.3233333
18	Hounslow	519	160	0.3082852
5	Bromley	284	87	0.3063380
22	Lambeth	2233	674	0.3018361
32	Wandsworth	1761	516	0.2930153
8	Croydon	503	147	0.2922465
23	Lewisham	957	279	0.2915361
29	Sutton	141	39	0.2765957
2	Barnet	684	189	0.2763158
28	Southwark	2259	611	0.2704737
31	Waltham Forest	603	163	0.2703151
17	Hillingdon	354	94	0.2655367
19	Islington	2121	562	0.2649694
26	Redbridge	339	89	0.2625369
3	Bexley	93	24	0.2580645
11	Greenwich	788	203	0.2576142
16	Havering	132	34	0.2575758
13	Hammersmith and Fulham	1978	473	0.2391304
12	Hackney	2404	570	0.2371048
10	Enfield	302	71	0.2350993
15	Harrow	222	52	0.2342342
1	Barking and Dagenham	158	35	0.2215190
4	Brent	1265	280	0.2213439
20	Kensington and Chelsea	2766	563	0.2035430
25	Newham	977	196	0.2006141
6	Camden	3124	621	0.1987836
33	Westminster	5116	913	0.1784597
30	Tower Hamlets	3922	680	0.1733809
7	City of London	192	33	0.1718750

> n_sort_listings

	neighbourhood	listings	superhosts	perc_superhost
33	Westminster	5116	913	0.1784597
30	Tower Hamlets	3922	680	0.1733809
6	Camden	3124	621	0.1987836
20	Kensington and Chelsea	2766	563	0.2035430
12	Hackney	2404	570	0.2371048
28	Southwark	2259	611	0.2704737
22	Lambeth	2233	674	0.3018361
19	Islington	2121	562	0.2649694
13	Hammersmith and Fulham	1978	473	0.2391304
32	Wandsworth	1761	516	0.2930153
4	Brent	1265	280	0.2213439
25	Newham	977	196	0.2006141
23	Lewisham	957	279	0.2915361
14	Haringey	900	291	0.3233333
9	Ealing	800	260	0.3250000
11	Greenwich	788	203	0.2576142
2	Barnet	684	189	0.2763158
31	Waltham Forest	603	163	0.2703151
27	Richmond upon Thames	527	215	0.4079696
18	Hounslow	519	160	0.3082852
24	Merton	513	168	0.3274854
8	Croydon	503	147	0.2922465
17	Hillingdon	354	94	0.2655367
26	Redbridge	339	89	0.2625369
10	Enfield	302	71	0.2350993
5	Bromley	284	87	0.3063380
21	Kingston upon Thames	241	95	0.3941909
15	Harrow	222	52	0.2342342
7	City of London	192	33	0.1718750
1	Barking and Dagenham	158	35	0.2215190
29	Sutton	141	39	0.2765957
16	Havering	132	34	0.2575758
3	Bexley	93	24	0.2580645

Figure 4b: 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.

We proceeded to see how listing prices varied by location:

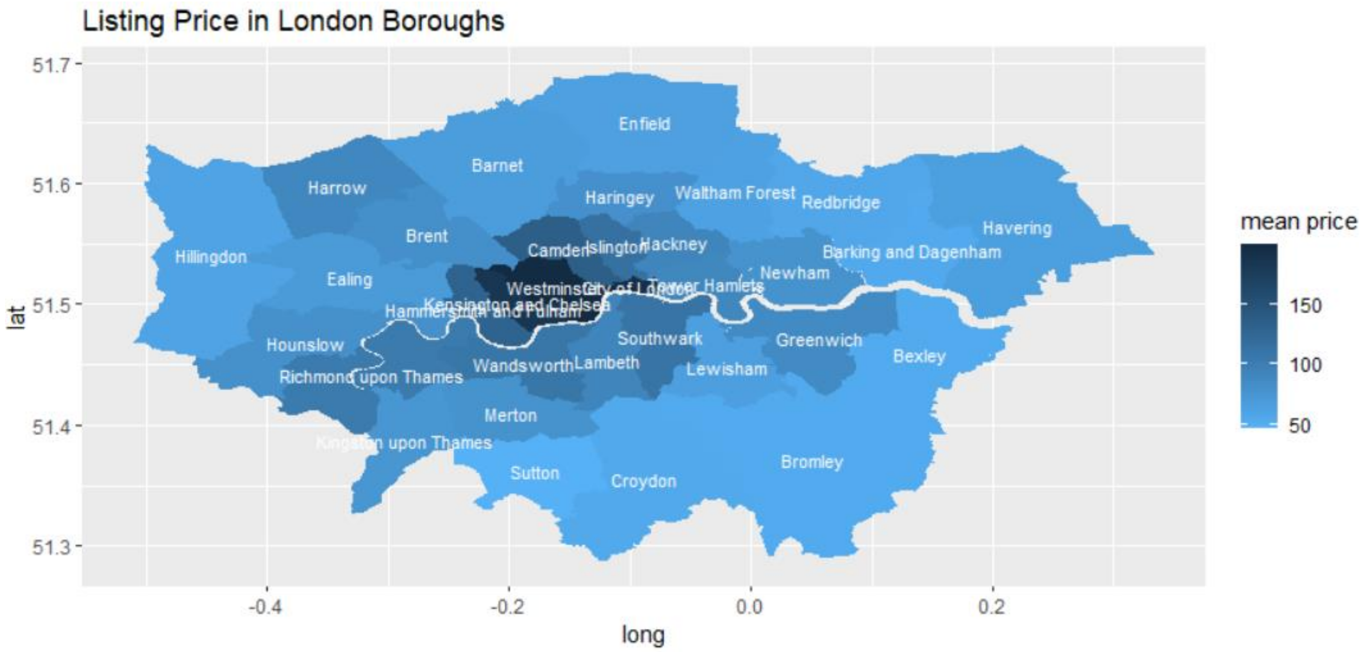


Figure 5: Price Map by Borough

As expected, central locations tend to have higher priced listings.

3.3 Property Data

Property types were investigated:

```
> summary(as.factor(property_n))
          Aparthotel              Apartment                   Barn      Bed and breakfast                   Boat
                  29                  25887                      2                    343                     32
      Boutique hotel               Bungalow                    Bus                  Cabin             Camper/RV
                  80                     81                      1                     12                      6
Casa particular (Cuba)                 Chalet            Condominium                Cottage            Earth house
                   3                      7                   1507                     46                      7
         Guest suite             Guesthouse                 Hostel                  Hotel                  House
                 239                    178                    153                     97                   7807
           Houseboat                    Hut             Lighthouse                   Loft         Minsu (Taiwan)
                  26                      4                      1                    367                      1
               Other          Ryokan (Japan)      Serviced apartment                  Tent             Tiny house
                  62                      1                    533                      1                     36
           Townhouse              Treehouse                  Villa                   Yurt
                1609                      1                     17                      2
```

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

```
> summary(as.factor(property_n))
  Apartment Condominium         House        Others
      26816        1524          9579          1259
```

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.

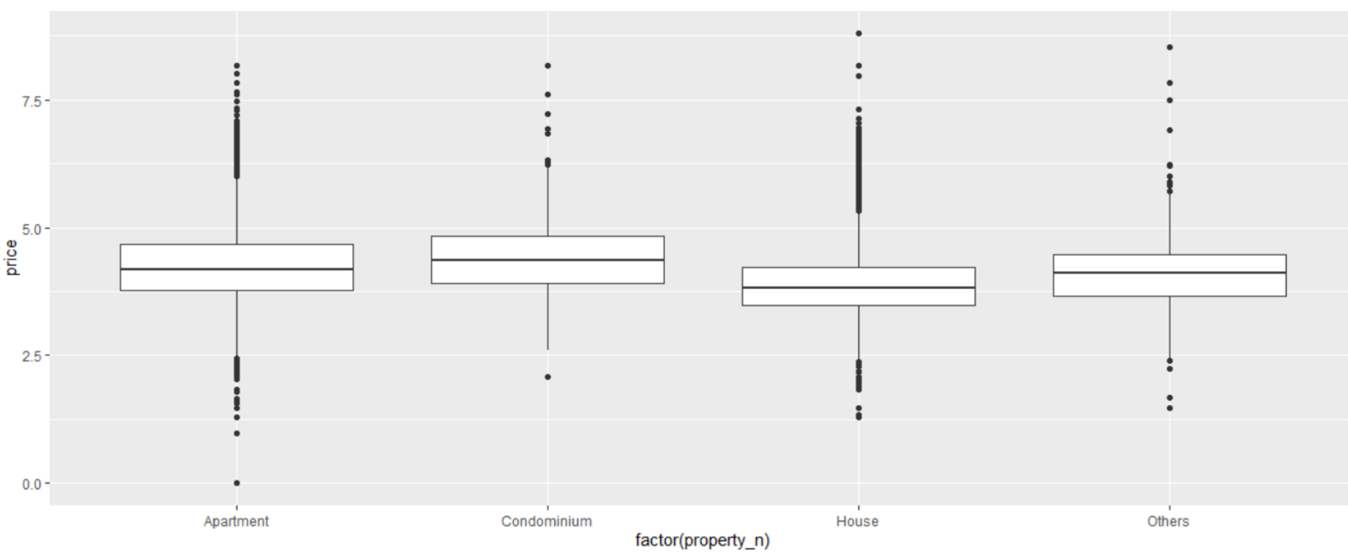
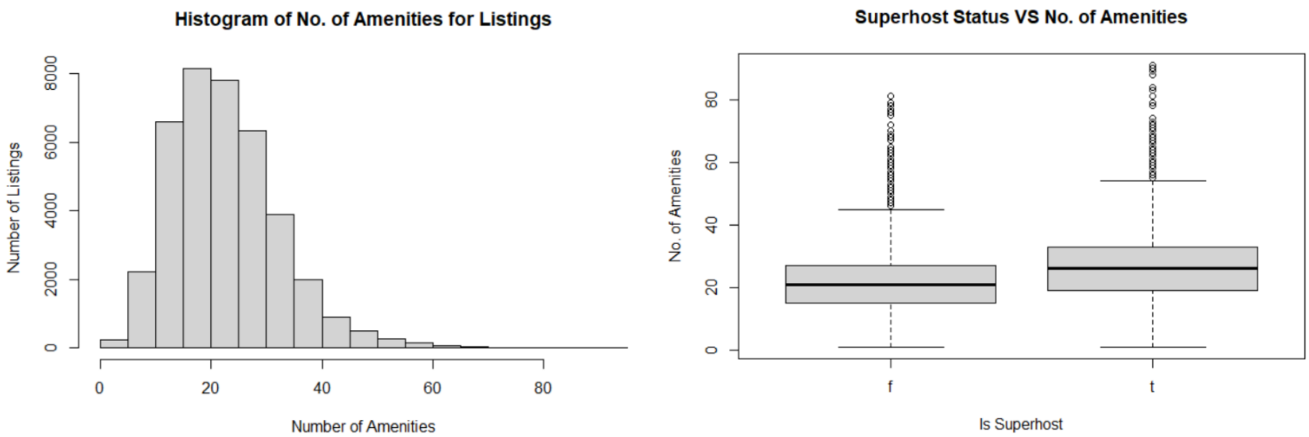


Figure 6: Box-plot of Normalised Prices by Property Type

3.4 Amenities Data

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:




```
> mean(total_sup[(total_sup$V2=="t"),1])
[1] 26.81666
> mean(total_sup[(total_sup$V2=="f"),1])
[1] 21.97999
```

Figure 7a: Distribution of Amenities and Superhost VS Amenities

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

We also generated a list of top 20 amenities found in our listings. Wifi is, unsurprisingly, at the top of the list, followed by some of the common household amenities as expected. We conducted further analysis to see if amenities listed greatly differed between Superhosts and non-Superhosts but found no significant results to report.

item	freq
Wifi	38287
Essentials	37777
Heating	37710
Smoke detector	35115
Kitchen	35002
Hangers	34066
Iron	32747
Washer	32689
Hair dryer	31341
Shampoo	29784
Hot water	27550
TV	27446
Laptop friendly workspace	26628
Carbon monoxide detector	25739
Refrigerator	21472
Dishes and silverware	19506
Bed linens	19026
Oven	18648
Microwave	18070
Cooking basics	17164

Figure 7b: Top 20 Amenities

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.

```
> length(popular.dat[popular.dat$is_popular==TRUE,"is_popular"])
[1] 8569
> length(popular.dat[popular.dat$is_popular==FALSE,"is_popular"])
[1] 30609
```

Out of 39,178 listings, a healthy number of 8,569 (21.9% of listings) are considered popular. This popularity index is further used in the following section.

4. 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:

```
> relevant_col
 [1] "log_price_n"           "host_since_n"          "host_response_time_n"  "host_response_rate_n"
 [5] "host_verification_n"   "mil_to_centre_n"       "property_type_n"       "bed_type_n"
 [9] "amen_count"            "cancellation_policy_n" "guest_verif_n"         "room_type"
[13] "superhost_n"           "guests_included"       "minimum_nights"        "reviews_per_month"
[17] "review_scores_rating"  "is_popular"
```

4.1 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:

```
Call:
glm(formula = is_popular ~ . - reviews_per_month - review_scores_rating -
    host_since_n - bed_type_n - log_price_n - mil_to_centre_n,
    family = binomial, data = listings2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7675  -0.5942  -0.4747  -0.3152   2.8947

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -2.916210   0.196518 -14.839  < 2e-16 ***
host_response_time_nwithin a day       0.339022   0.186445   1.818 0.069011 .
host_response_time_nwithin a few hours 0.683789   0.183629   3.724 0.000196 ***
host_response_time_nwithin an hour     0.864320   0.182057   4.748 2.06e-06 ***
host_response_rate_nNot100            -0.460401   0.036058 -12.768  < 2e-16 ***
host_verification_n                    0.098789   0.013621   7.253 4.08e-13 ***
property_type_nCondominium             0.367794   0.066727   5.512 3.55e-08 ***
property_type_nHouse                  -0.083547   0.033582  -2.488 0.012851 *
property_type_nOthers                  0.013120   0.080521   0.163 0.870566
amen_count                             0.018369   0.001437  12.779  < 2e-16 ***
cancellation_policy_nmoderate          0.011772   0.038537   0.305 0.760003
cancellation_policy_nstrict           -0.311306   0.036252  -8.587  < 2e-16 ***
guest_verif_nTRUE                     -0.370790   0.092905  -3.991 6.58e-05 ***
room_typeHotel room                   -1.225435   0.200526  -6.111 9.90e-10 ***
room_typePrivate room                  0.239589   0.032826   7.299 2.90e-13 ***
room_typeShared room                  -0.449378   0.204663  -2.196 0.028114 *
superhost_nTRUE                        1.825512   0.029045  62.850  < 2e-16 ***
guests_included                       -0.053800   0.011672  -4.609 4.04e-06 ***
minimum_nights                        -0.085504   0.019391  -4.410 1.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41159  on 39177  degrees of freedom
Residual deviance: 34278  on 39159  degrees of freedom
AIC: 34316

Number of Fisher Scoring iterations: 5
```

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 our prediction model based on this GLM model's fitted values with ROC curves. As shown below, the model has a better prediction rate than random guesses (45 degree line).

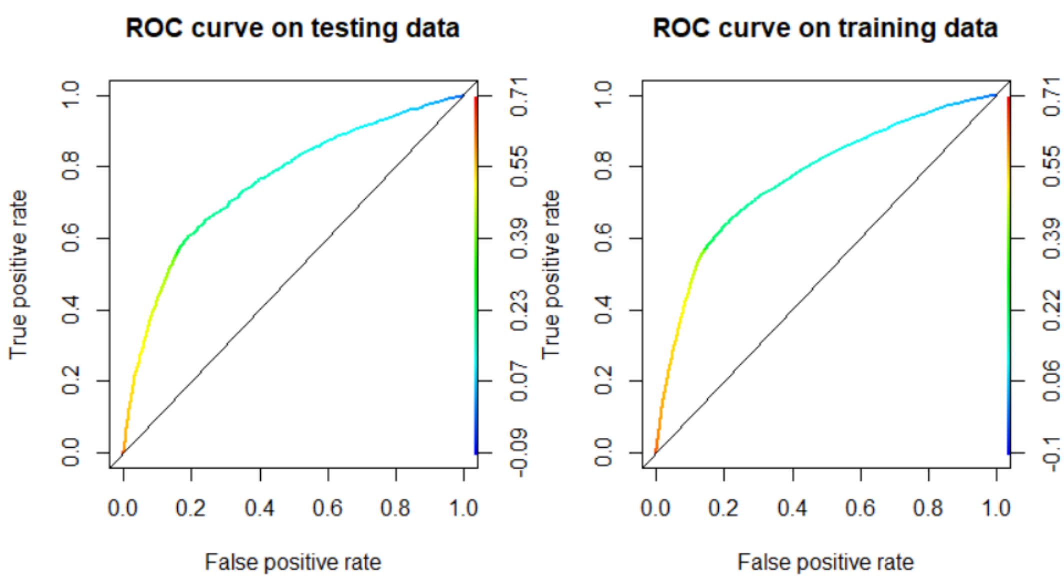
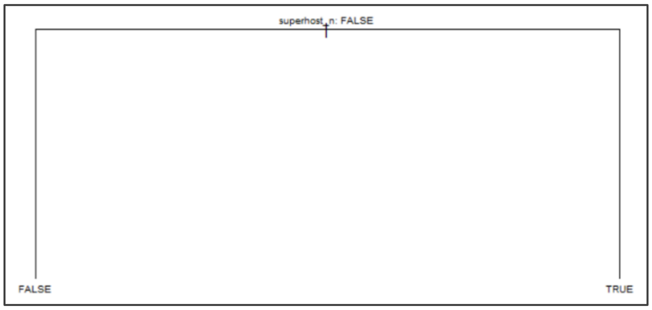


Figure 9: ROC Curves of GLM Prediction Model

4.2 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:




```
> summary(tree.pop1)

Classification tree:
tree(formula = as.factor(is_popular) ~ ., data = listings_dt)
Variables actually used in tree construction:
[1] "superhost_n"
Number of terminal nodes:  2
Residual mean deviance:  0.9032 = 35380 / 39180
Misclassification error rate: 0.2101 = 8232 / 39178
```

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.

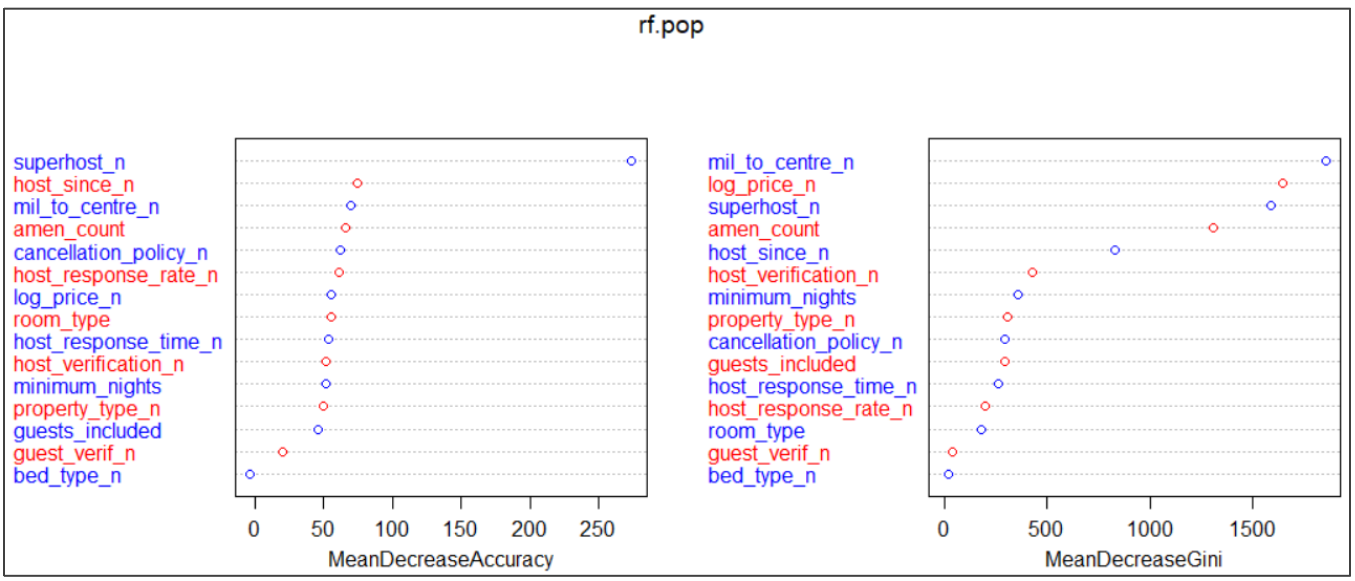


Figure 11: Random Forest

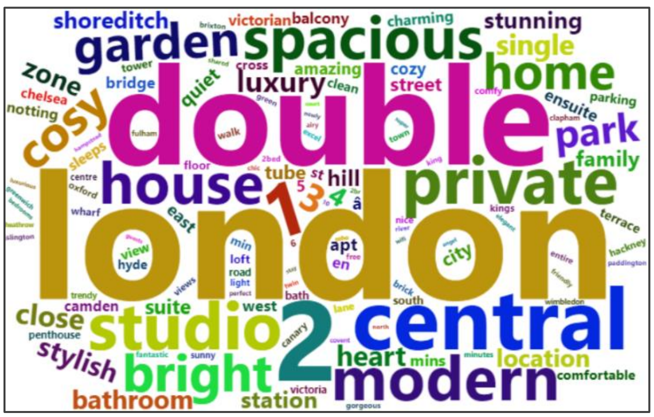
From our random forest analysis, **superhost_n, mil_to_centre_n, log_price_n, amen_count,** and **host_since_n** appear to be relatively more important variables, as excluding these variables either lead to a relatively strong decrease in model accuracy or a decrease in homogeneity of nodes. Unfortunately, the error rate of predicting popular listings remains high at 61.5%.

5. Text Analysis

Airbnb listing descriptions can be crucial for successful listings[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.

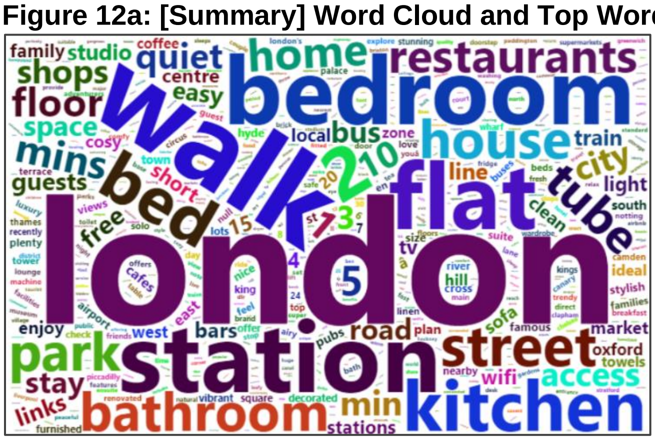
Name, **summary**, and d**escription** generally have words linked to apartment and room types. A commonly used word is "minutes", which suggests that Airbnb listers tend to advertise their listings' proximities to certain locations, presumably public transportation or areas of interest. Another interesting 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). Text descriptors seem to be colourful and marketable.

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



word1 <chr>	n <int>
london	7741
flat	7332
double	6016
bedroom	5982
apartment	5701
2	4938
bed	4498
central	3768
1	3295
studio	2894

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



word1 <chr>	n <int>
london	29203
walk	19922
station	16578
bedroom	15970
flat	15610
apartment	15592
minutes	14475
double	12367
2	12274
kitchen	11779

[6] https://www.airbnb.com.sg/resources/hosting-homes/a/write-an-appealing-listing-description-13?_set_bev_on_new_domain=1613396592_N2I0ZjE4NjZjZjZm

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

word1 <chr>	n <int>
bed	21145
kitchen	20096
bedroom	16696
double	16373
london	14862
bathroom	13510
apartment	13305
flat	13013
living	12559
space	10276

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

word1 <chr>	n <int>
london	50965
walk	36549
kitchen	35693
bedroom	32545
flat	32342
apartment	31459
bed	30056
station	29133
double	27617
minutes	25760

Figure 12e: [Host_about] Word Cloud and Top Words

word1 <chr>	n <int>
london	20454
NA	14359
love	14004
people	7692
stay	7252
guests	6722
home	6181
travel	5733
enjoy	5675
airbnb	5265

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 as the top words remain alike, hence not further described in the report.

6. Concluding Remarks

To conclude our findings, we have consolidated all insights for budding Airbnb listers here:

- **Strive to be a Superhost**: Superhosts' listings are more popular, though we do expect popular listings to have boosted hosts to a Superhost status in the first place

- **Location does matter:** central listings tend to be more expensive, and there are more listings as well, but one can achieve Superhost status despite poorer locations.

- **Popular listings tend to have responsive hosts**: responding faster may be linked to higher probability of closing bookings and better guest satisfaction. With more than half of the hosts having a 100% response rate, getting back to potential guests will put you head on against the competition.

- **Get verified**: verified hosts are more popular, although it could be due to a confounding element of host dedication. Email/phone verification is necessary, while Government ID verification is a great plus (3 out of 4 identified clusters use this).

- **Don't ask too much**: guest verification is found to have a negative relationship against popularity, so only ask for the necessary details.

- **Property types may have impact**: condominiums, others, apartments, houses, in order of popularity. Perhaps consider a condominium unit with great facilities for guests.

- **Room types matter**: guests love private rooms, while shared rooms and entire home/apt also work. If you are a hotel owner, consider another platform for listing your accomodation, as our statistics suggest that hotel rooms remain relatively unpopular.

- **Show them your amenities**: other listers love putting Wifi, Essentials, Heating, Smoke Detector, and Kitchen etc. amongst their list of amenities. If you have more to offer, be sure to declare them - Superhosts are found to offer approximately 25% more amenities.

- **Use the right words**: descriptors tend to state the property's proximity and amenities, whilst host introductions have more "feeling" words to elicit a positive emotional reception. Consider investing time into writing a great caption, although we posit that being responsive to guests is more important, as we found no remarkable differences between popular listings' descriptions and non-popular ones.

While we were able to gather the above correlations, there was unfortunately little explanatory power in the variables under the decision tree model. It is possible that listings with the right level of host responsiveness and other characteristics naturally made their hosts Superhosts, and the same listings became even more popular, resulting in Superhost status to be a main determinant of popularity. It is also possible that our popularity index was not effective in capturing true popularity; monthly review counts may be lower for listings where guests tend to stay longer. Nevertheless, the above are data-backed pointers worth the attention of anyone intending to join Airbnb to rent out their homes.

We envision future renditions of Airbnb data analysis to consider other measures of popularity, such as the number of high scoring reviews weighted by stay-time of guests, or perhaps the click-rate of listings. Having detailed information about the type of guests who stayed at the listings (couples, families, singles) would also help with market segmentation. It would also be interesting to look into COVID-19's impact on Airbnb listings. Lastly, a core factor crucial to Airbnb listings has been left out in our analysis; humans are incredibly visual creatures[7], so we might expect listings' photo quality and presentation to affect popularity of listings. Perhaps with randomised controlled trials or image recognition tools, an interesting aspect to look into deeper would be the impact of listings' photos.

[7] P. Messaris (1999). Visual Persuasion: The Role of Images in Advertising.

7. 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~~		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

8. Bibliography

1. S. Deane (2021). 2021 Airbnb Statistics: Usage, Demographics, and Revenue Growth. https://www.stratosjets.com/blog/airbnb-statistics/#:~:text=How%20Many%20Users%20Does%20Airbnb,in%20an%20Airbnb%20every%20night

2. M. Leonhardt (2019). 82% of people think Airbnb-ing their home is a good money-making strategy—here's what you need to know. 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

3. K. Lauritzen (2019). How to maximize profits on Airbnb? Data-based approach for hosts. https://towardsdatascience.com/how-to-maximize-profits-on-airbnb-data-based-approach-for-hosts-beaf08f26941

4. C. O'Donovan (2018). Here's Why Airbnbs Cost More Than You Think — And What Airbnb Is Doing About It. https://www.buzzfeednews.com/article/carolineodonovan/why-airbnbs-cost-more-extra-cleaning-fees

5. Airbnb (2021). What is a Superhost? https://www.airbnb.com.sg/help/article/828/what-is-a-superhost

6. Airbnb (2020). Write an appealing listing description. https://www.airbnb.com.sg/resources/hosting-homes/a/write-an-appealing-listing-description-13?_set_bev_on_new_domain=1613396592_N2I0ZjE4NjZjZjZm

7. P. Messaris (1999). Visual Persuasion: The Role of Images in Advertising.