Superhost: Airbnb in Paris

Katrin Maliatski

December 12 2019

1 Introduction

15 years ago, the thought of sharing a car or your home with a stranger was unthinkable. The financial crisis of 2008 provided a window of opportunity for creative minds like Joe Gebbia, co-founder of Airbnb, to launch a groundbreaking business. Airbnb is a third party service allowing hosts and travellers to rent accommodation through a free account and acts as an alternative to hotels (Rawes & Coomes, 2019). Just over a century since its inception, Airbnb now boasts over six million properties in over 81,000 cities around the world (Sherwood, 2019).

The sharing economy provided people who are increasingly anti-social and technology-minded with opportunities to interact with strangers and learn to trust them. It is now a common-place act to rent an Airbnb while travelling instead of staying at a hotel. In a sense, it makes people feel as though they are having a more "authentic" experience whilst visiting a new place to stay at the apartment of a local rather than in a sterile, bleak, and overpriced hotel.

Airbnb provides people of all demographics with the option to rent a room, apartment, or an entire house based on their desires, interests, and financial capabilities. The vast quantity and variety of listings make it possible for any person to find what suits their needs. In addition to providing people with the ability to create a genuinely human connection, it gives people the opportunity to earn money while on vacation or help pay their mortgage by renting out an extra room all while retaining the sense of trust (Isaacson, 2017).

The objective of this analysis is to uncover what differs 'Superhosts' from regular Airbnb hosts. The Superhost designation is given to hosts that "provide a shining example for other hosts, and extraordinary experiences for their guests." Every three months, Airbnb checks whether a host meets the criteria of having a high response rate, rating, and numerous stays at their listing(s) (Airbnb, n.d.). Through my analysis, I will explore what additional factors are important to a Superhosts' qualities. I will detail my analysis process of using random forest, principal component analysis, and logistic regression through statistical inference.

2 Data Description

This dataset was found on Kaggle, an online community of data enthusiasts and machine learners. The 'AirBnb Paris' dataset was uploaded in March 2019, and there is no description or further information given on it. The original dataset has 59,882 observations and 94 variables. Many of those variables had missing values or were proved insignificant to my analysis through random forest and logistic regression. The resulting 20 variables that I used include but are not limited to factors such as price, the number of listings, and the number of reviews a host has.

Many of the observations had missing values for response rate and response time. The response rate variable is classified by "within an hour," "within a few hours," within a day," or "within a few days." Observations with a missing response rate variable also had a missing response time variable. As this information is available on the Airbnb listing and each observation has a link to the listing, I manually filled in missing values for existing listings and removed observations where this information was missing or the listing was deleted. The rate of response is a key determining feature to whether or not a host is a Superhost, therefore it was integral to include in my model. As expected, Superhosts tend to respond to guests within an hour than regular hosts, emulating a higher quality of customer service (Appendix 1). The majority of Superhosts respond to guests within an hour, while regular hosts have more variability in response times.

The amenities variable was also of interest to me to see whether certain amenities, such as TV, Wifi, and the host greeting the customer affect the rating. The original amenities variable listed the listing's offerings in one line, therefore I created three new columns with a "True" or "False" value in order to create a dummy variable and determine their relationship to whether a host is a Superhost. Most listings have wifi, however, there is a lot of variance in hosts greeting guests. Superhosts are equally likely to greet guests as they are to not greet them, but they tend to have wifi and the majority have televisions (Appendix 2).

After cleaning my data and removing "N/A" values, I found that my scatter plots showed very low ratings and response times for Superhosts. I went back to my data and manually checked the listings to see whether the hosts were in fact Superhosts. I updated the information for existing listings and deleted those that no longer existed. While some Superhosts did in fact still have a lower rating than 80/100, their response times were still high. After updating my dataset, about 19 percent of my observations are Superhosts. Although there is a big ratio of hosts to Superhosts, this data authentically represents Airbnb as the Superhost designation is only for a select few people, with Superhosts being considered to be the "VIP" of hosts.

I also deemed it important to analyze the Airbnb listings in the 20 Parisian neighbourhoods, or "arrondissements." The data originally had many neighbourhoods, as I assume it is confusing for tourists to judge a neighbourhood by "I Arrondissement" and instead many were labelled "Notre Dame" or other tourist attractions in the neighbourhood. I used the zipcode variable to sort all the observations according to their neighbourhoods in order to make analyzing the neighbourhood variable more manageable. The 18th and 6th Arrondissements have the highest quantity of listings, with the 6th also having the highest quantity of Superhost listings (Appendix 3).

2.1 Collinearity

Collinearity occurs when two or more quantitative variables are highly correlated, making it difficult to determine the impact of each variable individually. It can affect the entire model by giving poor estimates and providing the wrong regression standard errors. To avoid this in my model, I performed a correlation matrix (Appendix 4). As per the rule of collinearity becoming an issue when the absolute correlation is above 0.8, some collinearity is evident between the bedroom and bathroom variables. These variables proved to be insignificant through the random forest importance test as well as irrelevant to my logistic model, therefore they were not utilized in my model.

3 Model Selection

Since the rise in popularity of Airbnb I have been interested in becoming a host myself. Although I do not have a property to list (yet), I know my competitive spirit would be eager to achieve the Superhost designation if I were an Airbnb host.

3.1 Random Forest

In order to better understand what I can do to become one in the future along with high ratings and response rates, I performed a random forest importance test to determine what other variables are most meaningful in a Superhost (Appendix 5). With each tree considering 4 predictors at random (sqrt(15)), the possibility of bias is eliminated due to multicollinearity. The importance visualization tool unsurprisingly shows that a hosts rating is very important to the model. However, the number of listings is an interesting and unique find.

3.2 Principal Component Analysis

To better understand the relationships between the variables and visualize it, I used the principle component analysis (PCA). By reducing variability and the various dimensions into a 2D graph, PCA showed me important variables, warned me of collinearity, and allowed me to see how Superhosts differ from regular hosts (Appendix 6).

My PCA output reveals that a host's number of listings and their rating are negatively correlated, which is a surprising find. The fact that more listings generally result in a host's lower rating could mean that hosts are less involved and attentive to their guests when they have multiple listings to manage.

The first principle component speaks to the quantitative features of a Superhost's listings, such as the amount of people their property accommodates, its price, and the cleaning fee. These three variables tend to move in the same direction, meaning if a property can accommodate more people, it will generally be more expensive, and will have a higher cleaning fee because it is likely a larger space.

The second principle component speaks more to the host's reviews and response rate, which can be attributed to the popularity of a listing. If many people inquire and the host is very responsive, this listing also gets booked more often and generates more reviews compared to other listings. It is evident that Superhosts tend to have few listings, making them more attentive and responsive to their guests. Superhosts do however tend to follow similar patterns to that of regular hosts in terms of accommodation, prices, and cleaning fees.

3.3 Logistic Regression

Using these findings, I created my logistic regression model. Logistic regression is beneficial in creating my model as my response variable, Superhost, is binomial. A host can either be a Superhost (true) or a regular host (false). Using Maximum Likelihood Estimation (MLE), a model that best predicts when a host is a Superhost was developed. I found that the variables that were deemed important by the random forest analysis also made the most positive impact in improving my logistic regression model's r-squared value (Appendix 7).

4 The (Superhost) Model

SupehostModel=lrm(Superhost verified.host+book.instantly+
host.listings.count+price+cleaning.fee+accommodates+security.deposit+
review.scores.rating+neighbourhood+host.greets+has.Wifi+
reviews.per.month+number.of.reviews+minimum.nights+host.response.rate)

This final model predicts the likelihood of a host being a Superhost using 15 variables, including dummy, categorical, and continuous variables.

4.1 RMS

Using the "rms" package, I monitored the r-squared value of my model, with the final model's value being 0.311. This means that about 30 percent of the variance in the Superhost variable is explained by my model.

4.2 K-Fold Test

It was also important to consider the model's error rate through the mean square error (MSE). The MSE shows how closely the fitted line is to the data points. I used a k-fold test to find my model's MSE. A k-fold test splits data into training and test data of equal size and averages the MSE of all the tests. I chose a k value of 50, meaning 49 folds were used as training data, and 1 fold was used at test data. The resulting MSE is 0.12197.

5 Predictions

I input several values of a hypothetical listings into my model to see the odds of the host being a Superhost versus a regular host. This model predicted the host as being a verified host, not allowing instant booking, having a high response rate and responding quickly. These factors help increase the odds of a host being a Superhost. Moreover, greeting guests, having wifi, and the listings being in the II Arrondissement were also used in the prediction and proved to contribute greatly to the likelihood of a host being a Superhost (Appendix 8). A host with these attributes was predicted to be about 80 percent likely to be a Superhost.

6 Suggestions for Hosts

Much like a hotel, the objective of an Airbnb is to provide the guest with a pleasurable and enjoyable stay. It is the host's task to ensure guests are satisfied and if any problems were to arise, to fix them. Dedicated hosts may have the opportunity to become a Superhost, which rewards hosts with higher search placement, higher revenue, and bonuses (Airbnb, n.d.). Superhosts automatically have a heightened sense of trust with their guests as it acts as a form of "pre-screening" for the host's reliability. As such, my data also reveals that Superhosts can charge a slightly higher price than their counterparts (Appendix 9).

Given that Airbnb's guidelines for Superhosts are quite stringent, my analysis revealed that many hosts have high response rates and quick answering times, as well as a high rating yet they are not Superhosts (Appendix 10). This is because a large barrier to becoming a Superhost is the amount of stays at a host's listings, which is also revealed by the number of reviews they have. This means that a host's listing must have lots of availability and be getting booked by guests in order to qualify to be a Superhost. A host that is active and has a lot of availability for their listing is more likely to be a Superhost as they are more experienced and attentive to their guests.

As there is high competition among hosts to achieve the Superhost status, hosts can differentiate themselves by providing elevated customer service. For example, hosts that greet their guests improve their guests' experience. Certain neighbourhoods also have an advantage of being more attractive to visitors, such as the II Arrondissement. Many travel blogs encourage visitors to stay in this neighbourhood as it is considered less "touristy," which is an admirable feature to Airbnb customers who want a more authentic experience. This residential neighbourhood is also conveniently close to many tourist attractions, such as the world-famous Louvre museum that permanently exhibits Leonardo De Vinci's

Mona Lisa. For example, Mathieu, who is a Superhost, has a listing for his “Cosy Flat in the Heart of Paris!” that is just a 17 minute walk to the Louvre. You and three other guests can enjoy this one bedroom and one-and-a-half bathroom apartment for just $144 a night!

Having a property in the heart of Paris is a luxury not many can afford, therefore, hosts elsewhere may have to focus on ensuring their property is more attractive compared to other listings in their area. Having good-quality, well-lit photos would be a great way to differentiate a listing from others and prove your listing more desirable. As previously mentioned, elevating your guests' experiences will also allow a host to differentiate themselves and improve their odds of becoming a Superhost.

7 Limitations

As becoming a Superhost is an exclusive program that is designed for only the best Airbnb hosts, the ratio of Superhosts to normal hosts is approximately 2:10. With lots of factors being shared by Superhosts and normal hosts alike, it makes it rather difficult to pinpoint what other factors contribute to a host becoming a Superhost. My dataset also did not include the host's cancellation rates, which would also be a helpful factor in determining a Superhost.

In order to better understand the ”perks” a Superhost is privy to, it would be useful to survey Airbnb hosts. The questions could ask if they were aiming to become a Superhost and if Airbnb is their only income. It would also be helpful to analyze the financial aspect of being a Superhost, such as determining whether Superhosts do in fact make a higher revenue than their regular host counterparts.

8 Bibliography

Airbnb (n.d.).Airbnb Superhost program details. Retrieved 1 December 2019, from https://www.airbnb.ca/superhost

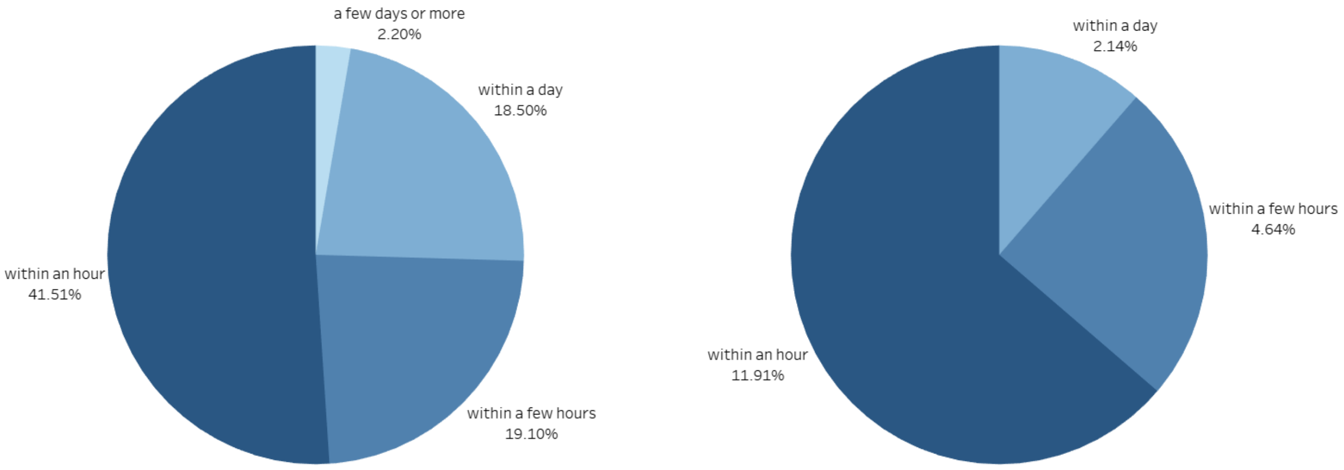
Isaacson, W. (2017). How Uber and Airbnb Became Poster Children for the Disruption Economy. Retrieved 1 December 2019, from https://www.nytimes.com/2017/06/19/books/review/wild-ride-adam-lashinsky-uber-airbnb.html

Rawes, E., Coomes, J. (2019). What is AirBnb? Everything you need to know before using the service. Retrieved 1 December 2019, from https://www.digitaltrends.com/home/what-is-airbnb/

Sherwood, H. (2019). How Airbnb took over the world. Retrieved 1 December 2019, from https://www.theguardian.com/technology/2019/may/05/airbnb-homelessness-renting-housing-accommodation-social-policy-cities-travel-leisure

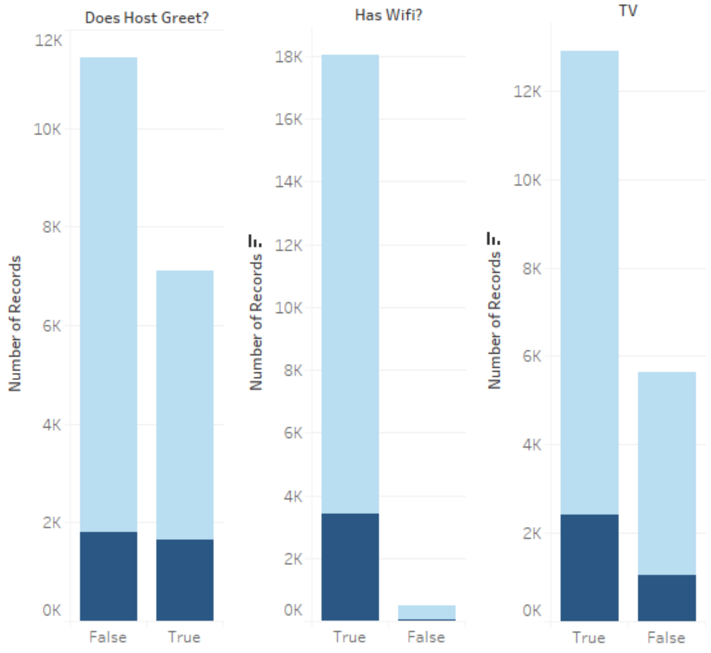
9 Appendix

Appendix 1: Rate of Response



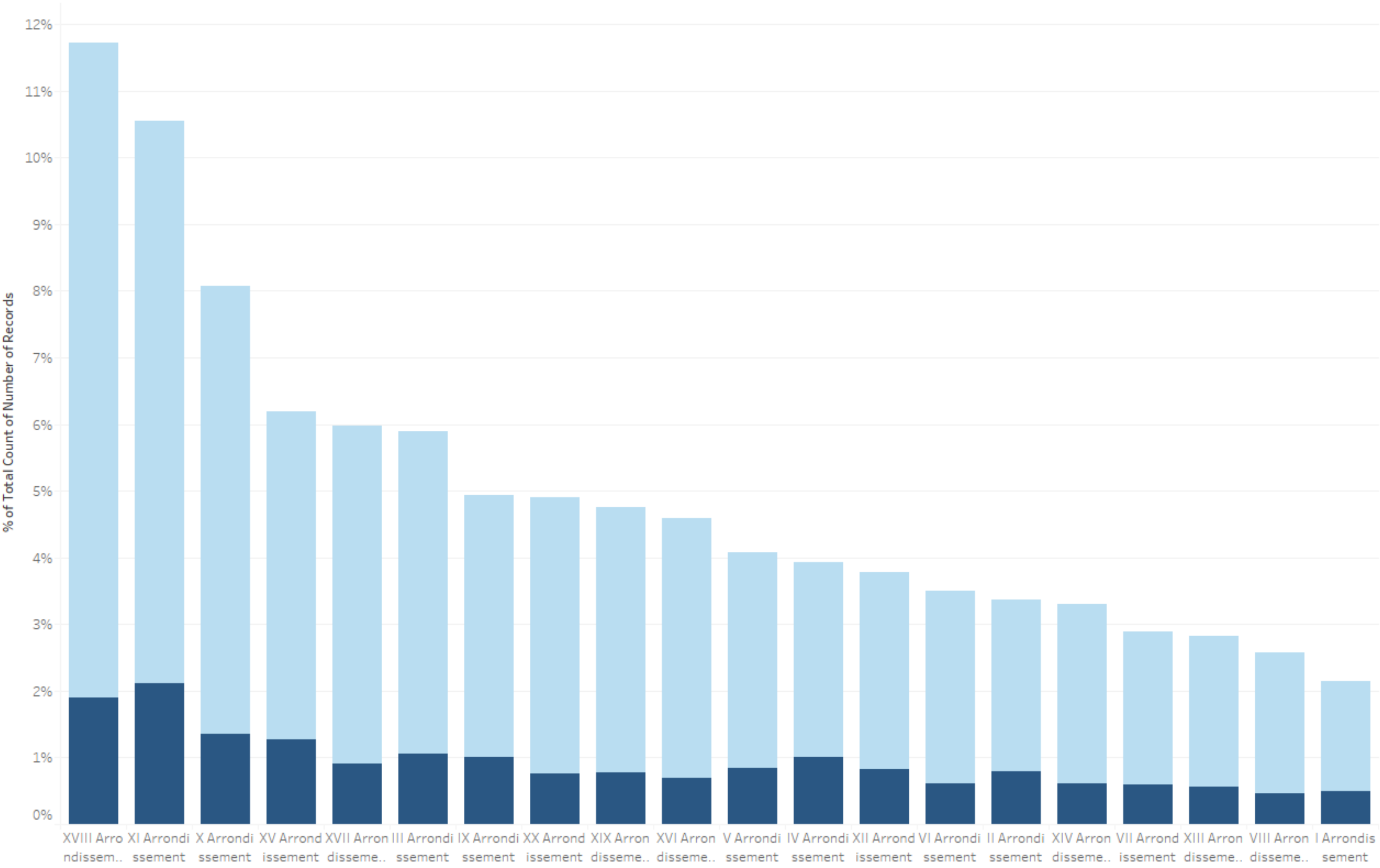
Regular hosts (left) only about half respond within an hour. The vast majority of Superhosts (right) respond within an hour.

Appendix 2: Amenities



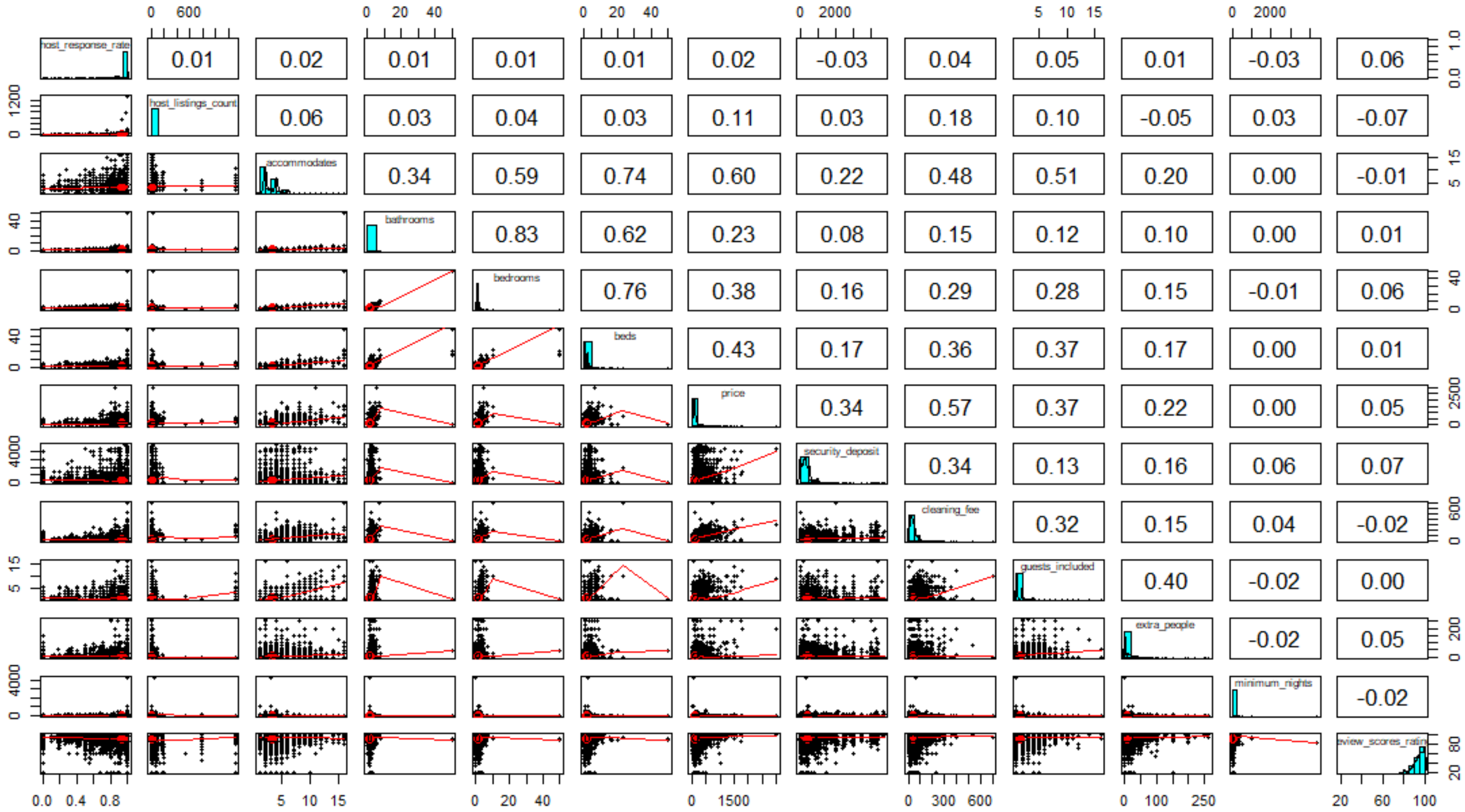
It is clear that Wifi is an essential for an Airbnb, as almost all properties have it. Most properties have televisions, however, it is suprising that almost six thousand do not. There is a much smaller gap between hosts that greet their guests versus those that do not. Interestingly, there is an equal split between greeting and not greeting for Superhosts.

Appendix 3: Neighbourhoods



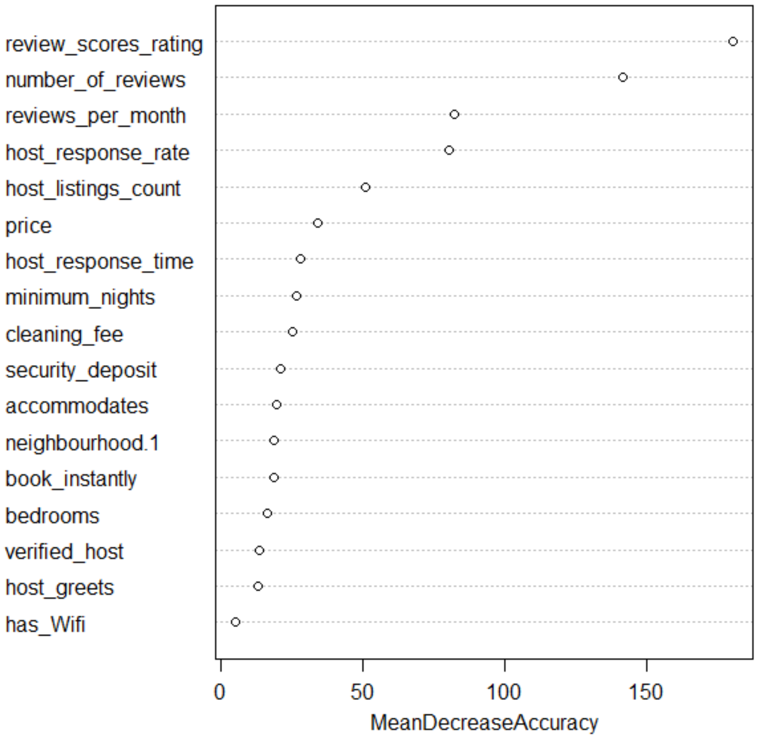
There are 20 Arrondissements or neighbourhoods in Paris, with the 18th and 6th being the most popular. The 1st and 8th Arrondissements are the lest popular.

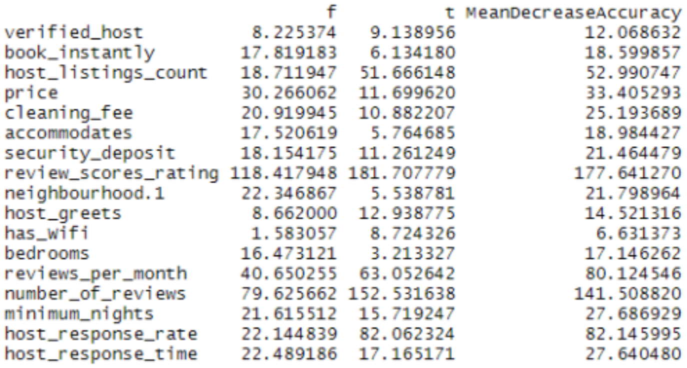
Appendix 4: Collinearity



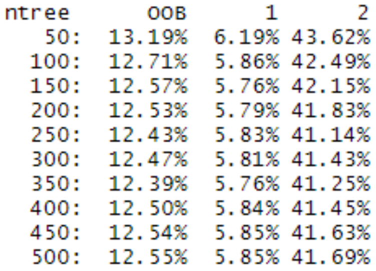
There is collinearity present between the bedroom and bathroom variables. However, they were not used in my final model as neither variable contributed to my model.

Appendix 5: Random Forest





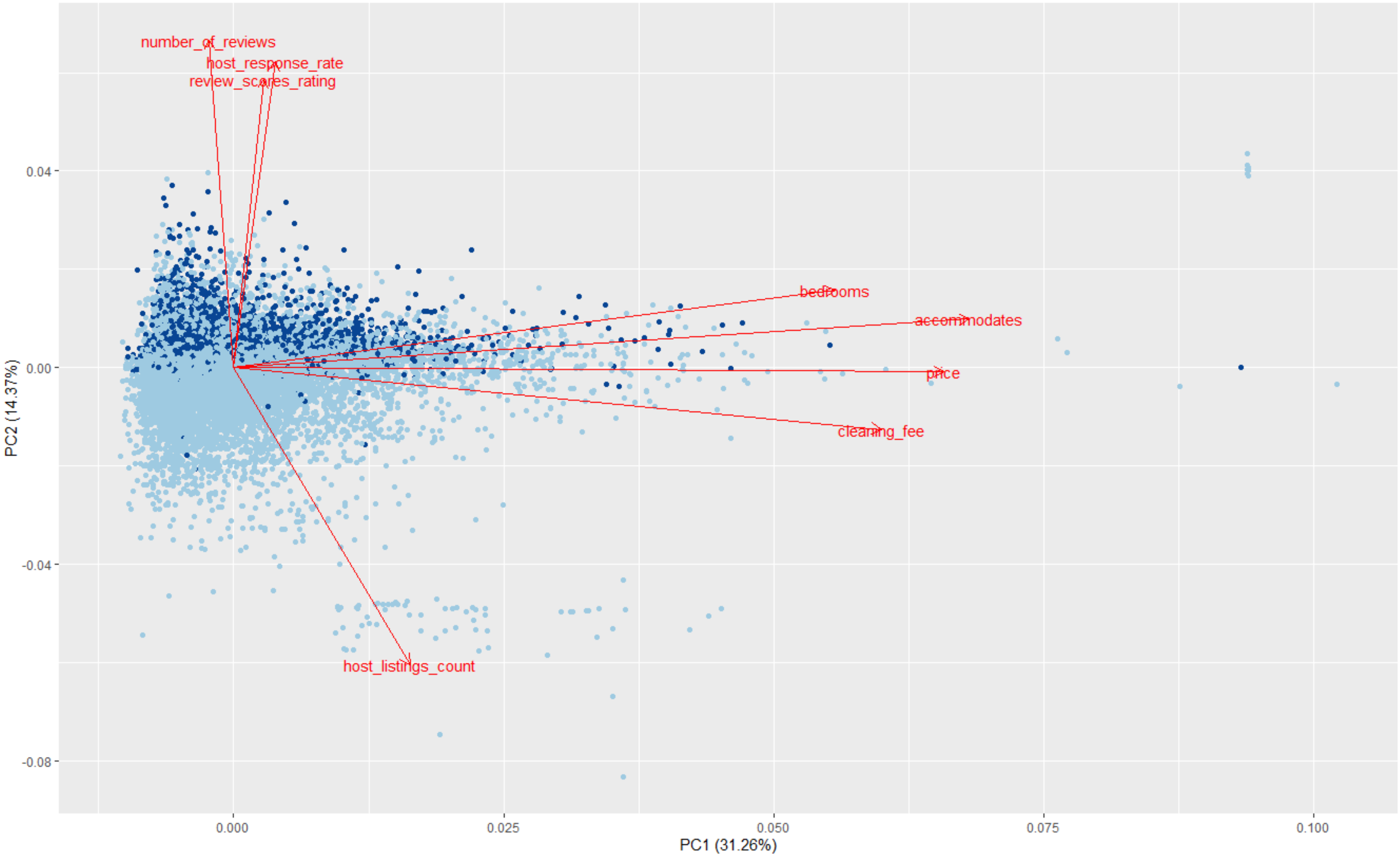
```
                             f          t MeanDecreaseAccuracy
verified_host         8.225374   9.138956            12.068632
book_instantly       17.819183   6.134180            18.599857
host_listings_count  18.711947  51.666148            52.990747
price                30.266062  11.699620            33.405293
cleaning_fee         20.919945  10.882207            25.193689
accommodates         17.520619   5.764685            18.984427
security_deposit     18.154175  11.261249            21.464479
review_scores_rating 118.417948 181.707779          177.641270
neighbourhood.1      22.346867   5.538781            21.798964
host_greets           8.662000  12.938775            14.521316
has_wifi              1.583057   8.724326             6.631373
bedrooms             16.473121   3.213327            17.146262
reviews_per_month    40.650255  63.052642            80.124546
number_of_reviews    79.625662 152.531638           141.508820
minimum_nights       21.615512  15.719247            27.686929
host_response_rate   22.144839  82.062324            82.145995
host_response_time   22.489186  17.165171            27.640480
```



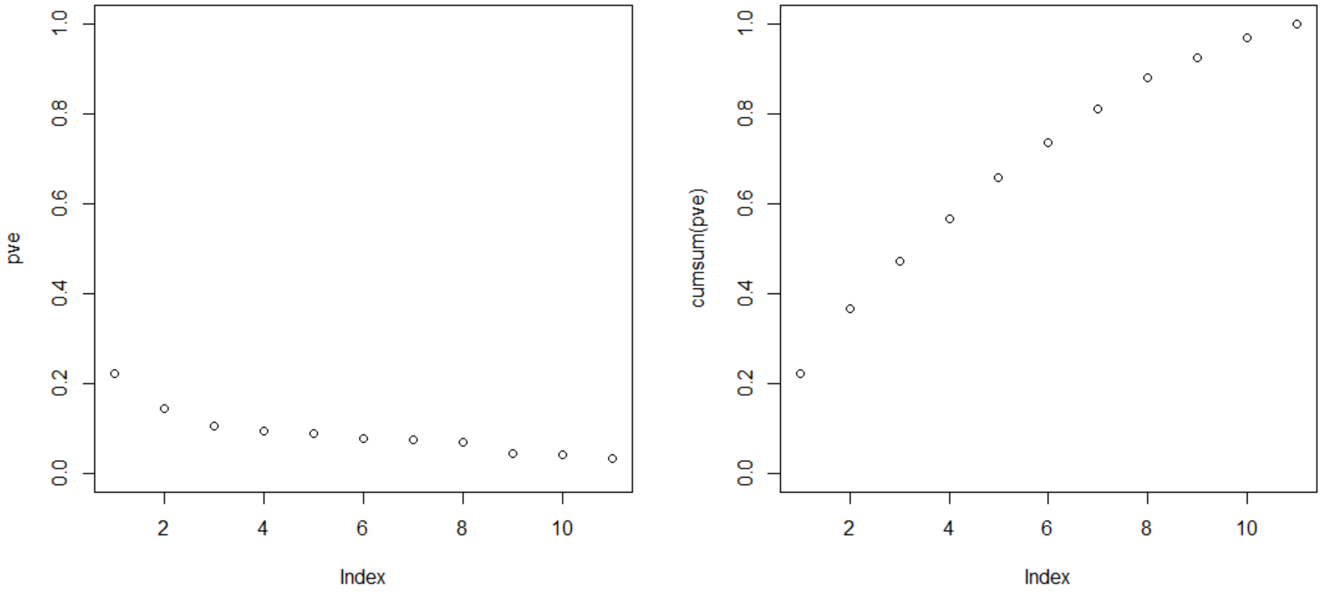
```
ntree      OOB      1      2
   50:  13.19%  6.19% 43.62%
  100:  12.71%  5.86% 42.49%
  150:  12.57%  5.76% 42.15%
  200:  12.53%  5.79% 41.83%
  250:  12.43%  5.83% 41.14%
  300:  12.47%  5.81% 41.43%
  350:  12.39%  5.76% 41.25%
  400:  12.50%  5.84% 41.45%
  450:  12.54%  5.85% 41.63%
  500:  12.55%  5.85% 41.69%
```

The Random Forest test concurred that rating significantly affects whether a host is a Superhost or not. The number of reviews/reviews per month variables are also very important to the model. The listing count was an interesting find. The Out-Of-Bag error was decreased from 13.19% to 12.55% through a 500 tree algorithm.

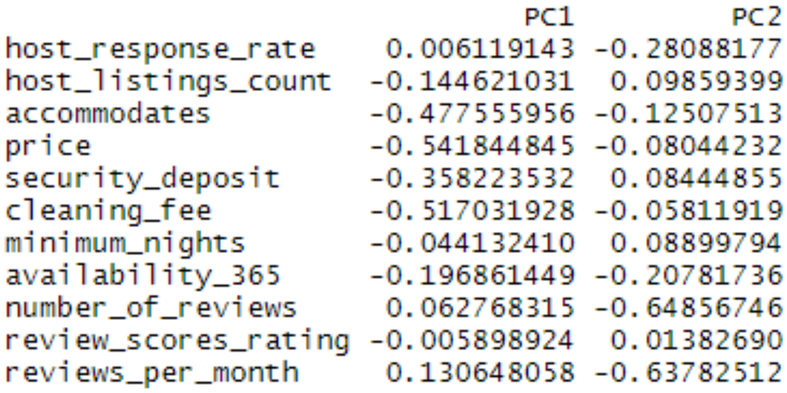
Appendix 6: Principal Component Analysis



This PCA analysis shows that number of reviews, host's response rating and their rating move in the same direction. Similarly, bedrooms, the number of people a property accommodates, its price, and its cleaning fee move in the same direction. However, the listings count is negatively correlated to the rating, response rate, and number of reviews.



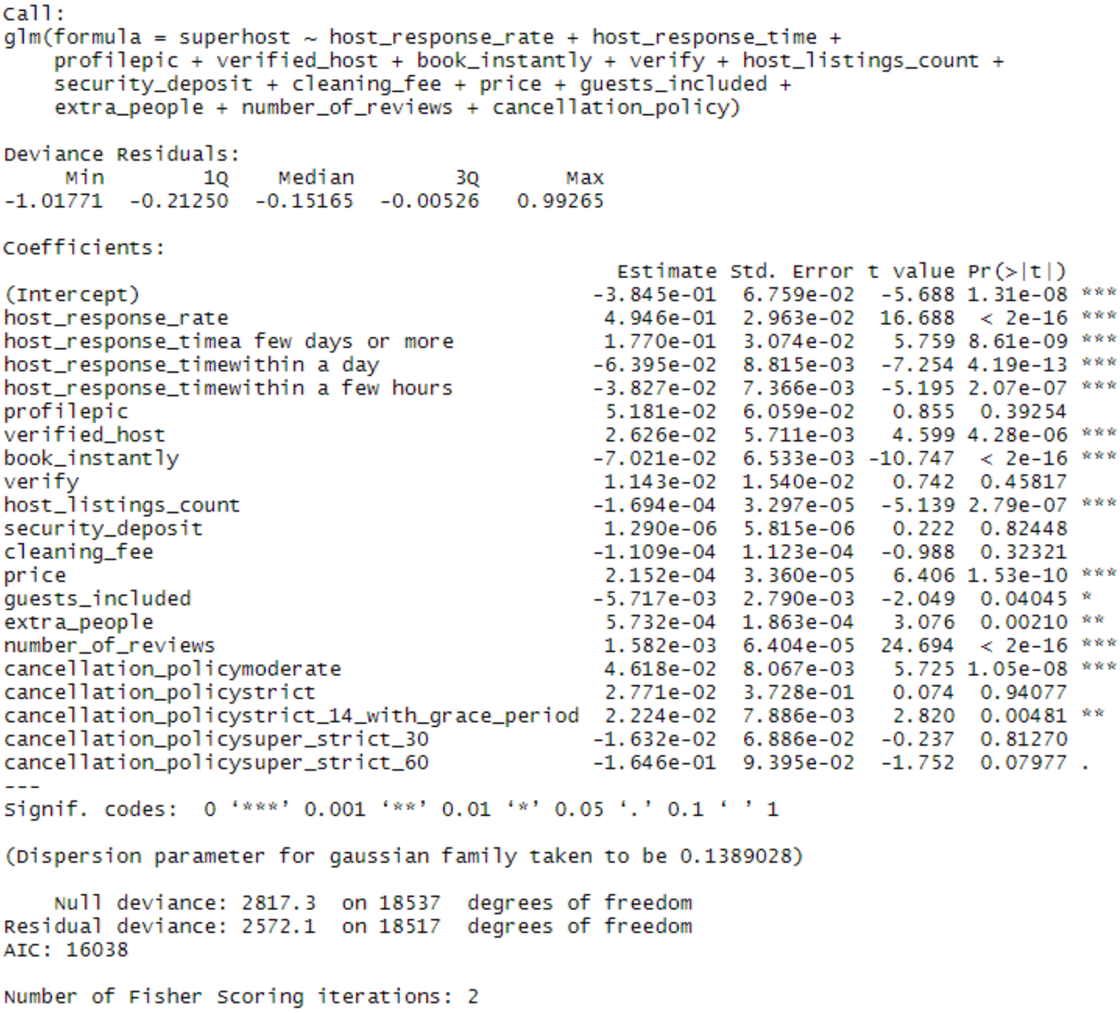
More components increase the accuracy of the model in predicting a Superhost.



```
                              PC1         PC2
host_response_rate    0.006119143 -0.28088177
host_listings_count  -0.144621031  0.09859399
accommodates         -0.477555956 -0.12507513
price                -0.541844845 -0.08044232
security_deposit     -0.358223532  0.08444855
cleaning_fee         -0.517031928 -0.05811919
minimum_nights       -0.044132410  0.08899794
availability_365     -0.196861449 -0.20781736
number_of_reviews     0.062768315 -0.64856746
review_scores_rating -0.005898924  0.01382690
reviews_per_month     0.130648058 -0.63782512
```

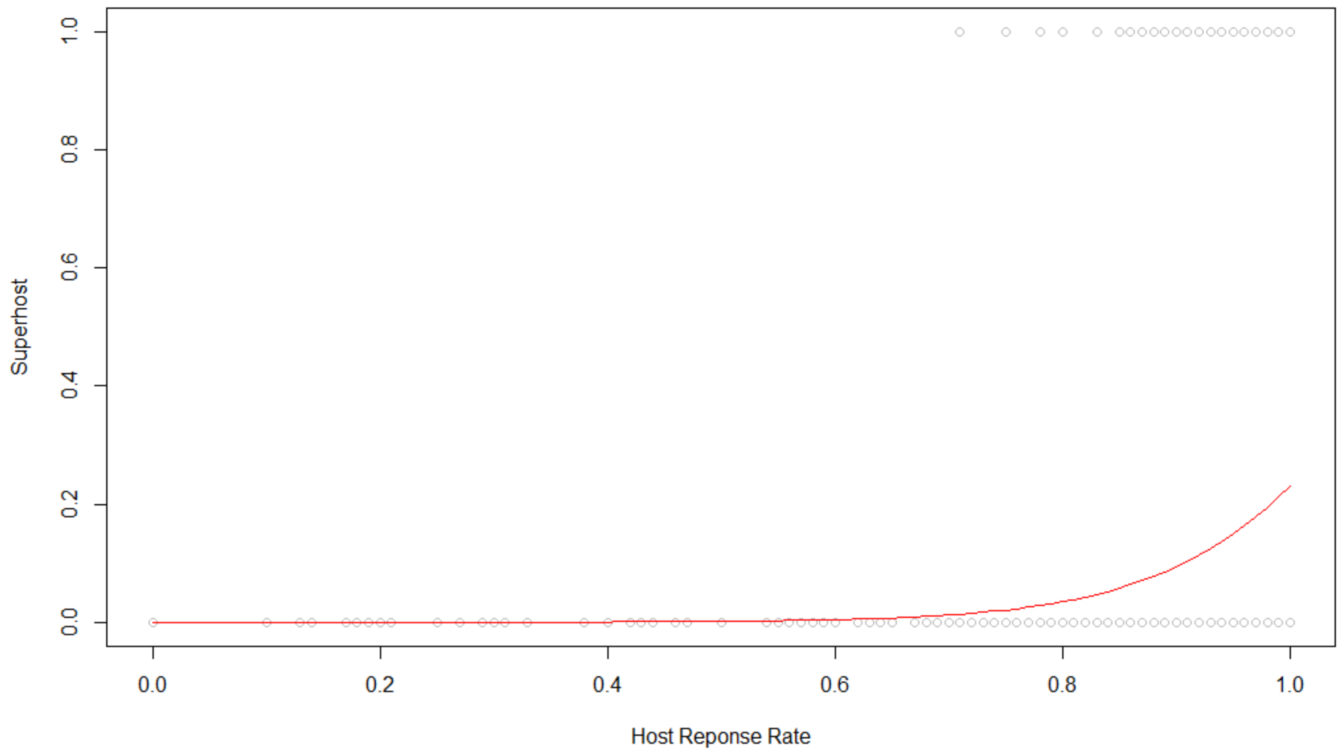
PC1 speaks to the number of people accommodated, price, and cleaning fee. PC2 is explained by the reviews and reviews per month.

Appendix 6: Logistic Regression



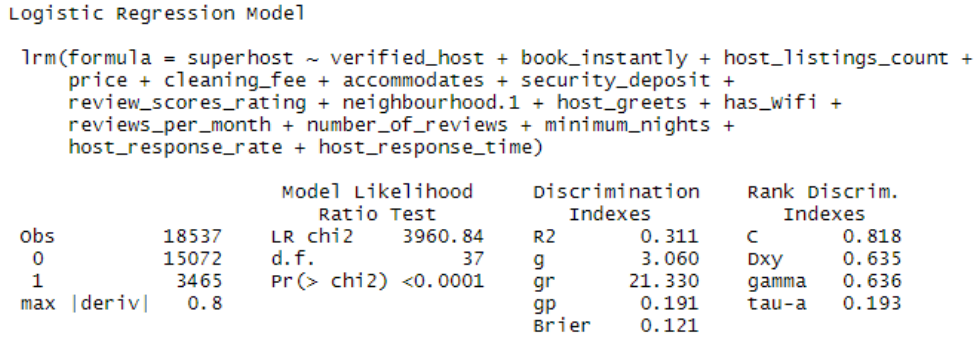
```
Call:
glm(formula = superhost ~ host_response_rate + host_response_time + 
    profilepic + verified_host + book_instantly + verify + host_listings_count + 
    security_deposit + cleaning_fee + price + guests_included + 
    extra_people + number_of_reviews + cancellation_policy)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-1.01771  -0.21250  -0.15165  -0.00526   0.99265  

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                 -3.845e-01  6.759e-02  -5.688 1.31e-08 ***
host_response_rate                           4.946e-01  2.963e-02  16.688  < 2e-16 ***
host_response_timea few days or more         1.770e-01  3.074e-02   5.759 8.61e-09 ***
host_response_timewithin a day              -6.395e-02  8.815e-03  -7.254 4.19e-13 ***
host_response_timewithin a few hours        -3.827e-02  7.366e-03  -5.195 2.07e-07 ***
profilepic                                   5.181e-02  6.059e-02   0.855  0.39254    
verified_host                                2.626e-02  5.711e-03   4.599 4.28e-06 ***
book_instantly                              -7.021e-02  6.533e-03 -10.747  < 2e-16 ***
verify                                       1.143e-02  1.540e-02   0.742  0.45817    
host_listings_count                         -1.694e-04  3.297e-05  -5.139 2.79e-07 ***
security_deposit                             1.290e-06  5.815e-06   0.222  0.82448    
cleaning_fee                                -1.109e-04  1.123e-04  -0.988  0.32321    
price                                        2.152e-04  3.360e-05   6.406 1.53e-10 ***
guests_included                             -5.717e-03  2.790e-03  -2.049  0.04045 *  
extra_people                                 5.732e-04  1.863e-04   3.076  0.00210 ** 
number_of_reviews                            1.582e-03  6.404e-05  24.694  < 2e-16 ***
cancellation_policymoderate                  4.618e-02  8.067e-03   5.725 1.05e-08 ***
cancellation_policystrict                    2.771e-02  3.728e-01   0.074  0.94077    
cancellation_policystrict_14_with_grace_period  2.224e-02  7.886e-03   2.820  0.00481 ** 
cancellation_policysuper_strict_30          -1.632e-02  6.886e-02  -0.237  0.81270    
cancellation_policysuper_strict_60          -1.646e-01  9.395e-02  -1.752  0.07977 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1389028)

    Null deviance: 2817.3  on 18537  degrees of freedom
Residual deviance: 2572.1  on 18517  degrees of freedom
AIC: 16038

Number of Fisher Scoring iterations: 2
```

Many of the varibales such as host response rate, response time, listings count, price have very low p-values and suggest are very significant in the model, as shown previously by the Random Forest test.



Unlike a linear predictive curve, the logistic predictive curve suggest that the likelihood of being a Superhost only increases significantly if a host's response rate is above 80 percent.

Appendix 7: Logistic Regression: rms



```
Logistic Regression Model

 lrm(formula = superhost ~ verified_host + book_instantly + host_listings_count +
     price + cleaning_fee + accommodates + security_deposit +
     review_scores_rating + neighbourhood.1 + host_greets + has_wifi +
     reviews_per_month + number_of_reviews + minimum_nights +
     host_response_rate + host_response_time)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs         18537    LR chi2     3960.84    R2       0.311    C       0.818
  0          15072    d.f.             37    g        3.060    Dxy     0.635
  1           3465    Pr(> chi2) <0.0001     gr      21.330    gamma   0.636
 max |deriv|   0.8                           gp       0.191    tau-a   0.193
                                             Brier    0.121
```

The rms package allowed me to determine the r-squared value of my model, 0.311.

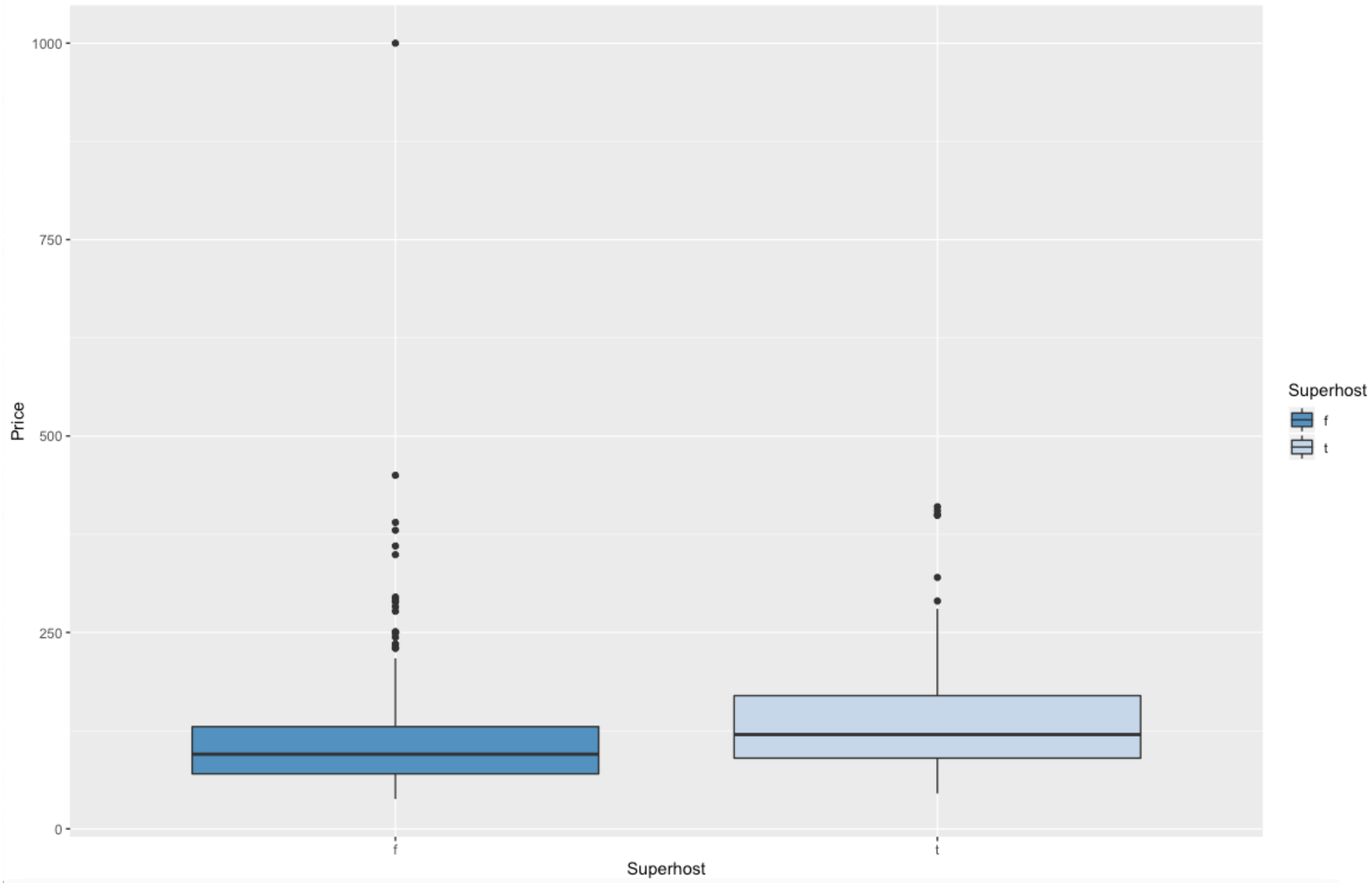
Appendix 8: Predictions

Variable	Predicted Value
Verified Host	1
Instant Book	0
Listings Count	1
Price	300
Cleaning Fee	150
Accomodates	3
Security Deposit	200

Review Rating	100
Neighbourhood	II Arrondissement
Host Greets	1
Has Wifi	1
Reviews per Month	3
Number of Reviews	80
Minimum Nights	5
Host Response Rate	1
Response Time	Within an Hour
RESULT	0.83307

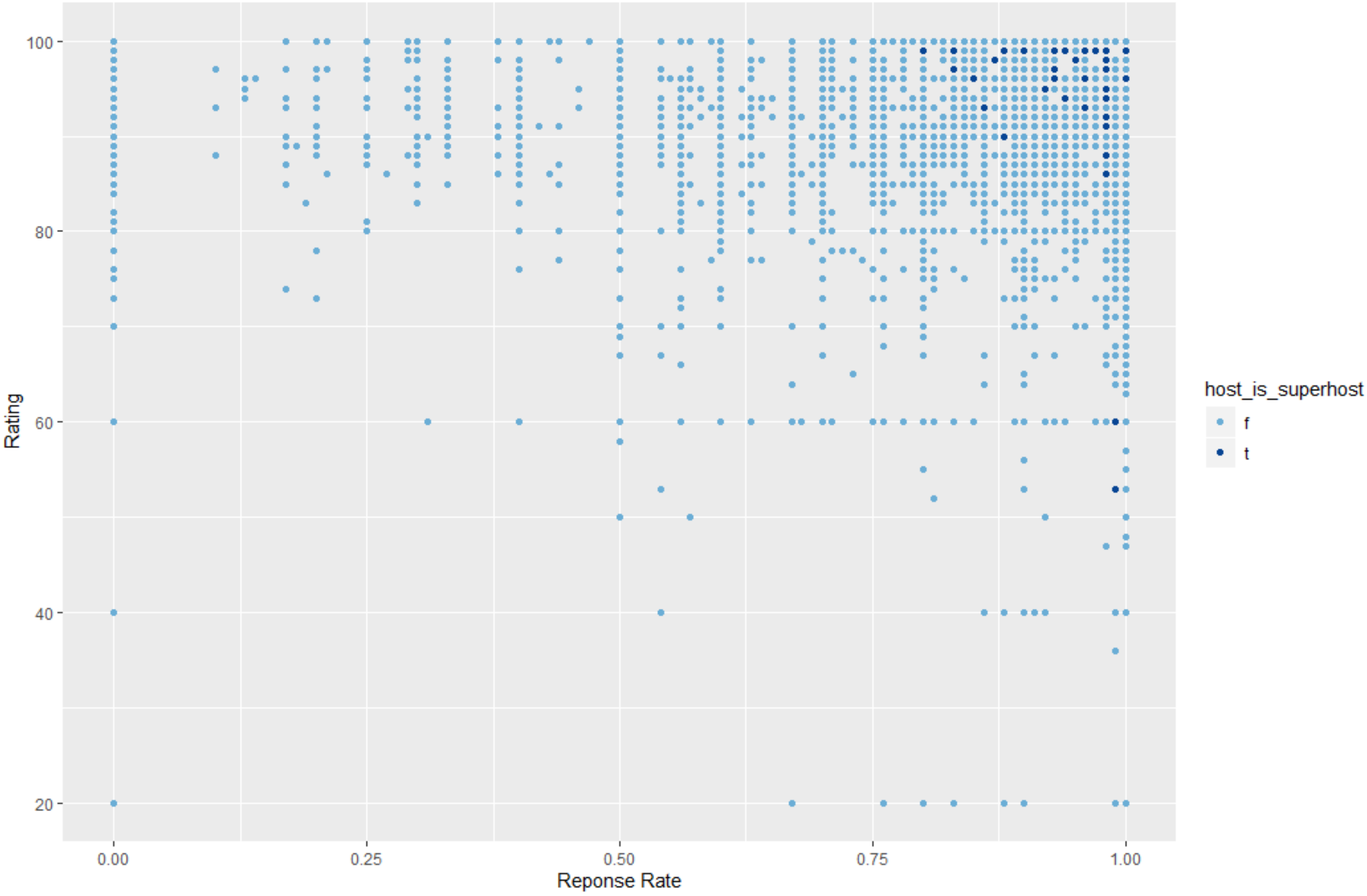
My model predicts that a host has the odds of 83.3% to be a Superhost with the above values.

Appendix 9: Superhost Price vs Regular Host Price



Regular hosts charge an average of $95 a night while Superhosts charge an average of $120 a night, suggesting that being a Superhost has the benefit of being able to charge a higher price due to the added layer of trust and assurance of a good experience.

Appendix 10: Rating & Response Time: Superhosts vs Regular Hosts



This scatterplot shows the trends of a host’s rating and response rate given whether the host is a Superhost or not. There is a clear cluster of Superhosts at the top right of the graph, demonstrating that Superhosts generally have high ratings and fast response rates. However, there are a couple of points that represent hosts with lower ratings, indicating that other factors could contribute to their Superhost designation. This could also suggest that their status as Superhost will be recinded in the next three month period analysis by Airbnb.