Data mining project report

Project title: Airbnb rentals analysis for new property acquisition and improved management in New York city.

I. Executive Summary:

My first business proposal was to focus on initial acquisition of property, thus I looked at important features in the data affecting high_booking_rate. Since quite a few listings had missing zip code, I used the latitude and longitude of the listings to map them into zip codes. New York City zip codes' GeoJSON information was used to achieve this. This GeoJSON information was also used with the 'leaflet' R package to plot interactive New York City maps. Since I did not find GeoJSON information for the zip codes that are in New Jersey, about 1800 New Jersey listings, out of the 32,143 total listings in New York, have not been included in the following analysis. Out of all the important features that I discovered from my model, location was a standout feature and only one which could be used before buying the property. Key findings from location analysis Ire locations with high score for review_score_location tend to be expensive and booking rates over there are high as Ill, but they are not affordable for novice investors but best for veteran investors. Some outskirt areas in New York had low audience reach but high booking rate and high review scores for location and hence, novice investors should compete for these locations.

Second business proposal focused on management, where I looked at how to be a superhost for the market of New York. Using the XGBOOST.importance function I Ire able to plot a pie chart showing features listed on the Airbnb website that helped qualify a host as a superhost for New York.

My third business proposal looked at important amenities to increase the booking rate.

II. Research Questions:

1. Which areas in New York are the best for the investor to buy property for Airbnb listings and identify important parameters required for initial acquisition of property?

   1.1) Which areas have high review scores?

   1.2) Which areas have a high average booking rate?

   1.3) Variations in the rental price in different areas?

   Reason: New York is a very vast market and so preferences of people are different in different areas. Thus, it is very important to determine the best area where the investor should invest. So, after getting answers to these questions, I could give advice to the investor regarding the location in New York where he/she should invest. I also wanted to identify parameters which are responsible in determining these locations and if there is any definite specific formula for obtaining a high booking rate at specific locations.

2. Once a property is acquired, what are the other factors that influence the high booking rate on the property?

   2.1) Does being a Super Host increase chances of having a high booking rate on the property?

   2.2) What are the important features for a host to be a Super Host in the New York market?

   2.3) What other features or amenities should be added to the property to ensure high booking rate?

   Reason: Buying a property on a good location alone would not ensure a high booking rate. There are other factors that need to be considered after the acquisition as Ill. First of all, I look at the influence of being a Super Host on a high booking rate. Based on the website of Airbnb, there are some conditions mentioned to achieve superhost status, but those are the minimum requirements and are for all the hosts worldwide. Thus, I want to analyze what all additional factors are associated for becoming a superhost along with those mentioned on their website, specific to the New York market. On finding those important factors, the investor can focus on those factors to get the Super Host status. Also, I want to find the features and amenities that would help in getting a high booking rate for the properties in New York market.

Questions which Ire rejected before choosing above questions?

1) Does description and access info about rentals have any effect on booking rate?

2) Using open street map API to analyze if the tourist spots across your Airbnb listing have any effect on booking rate?

3) What is the effect of hiring professional management?

Reason: My main objective was to focus on initial acquisition and although I Ire very determined on choosing the 2nd question i.e. API analysis, it was extremely time consuming and did not significantly affect the model. While I first thought of adopting topic modelling on access columns and sentimental analysis on both these columns but superhost and amenities analysis gave us way more information. Also the first map analysis gave us the importance of location and how other parameters affected its selection and hence I rejected to analyze the effect of working host listing count columns.

III. Methodology:

Data preparation Process and mining:

1. Converted some numeric columns such as price and security deposit into model usable format by removing $ from them.
2. Subdivided categorical variables into different categories by looking at the skewness of the distribution(found from histograms), and created new category for missing values named as unknown category (host related features, review scores feature)
3. Imputed missing values in security deposit and cleaning fee with zero, and the missing values in bed, bedroom and bathroom Ire imputed with their respective means.
4. Created host_since feature consisting of number of days the host has been with Airbnb, by converting date into numbers.
5. Created new features for each amenity with a total of 219 amenities. Only used most important amenities in the final model, which Ire given by models feature importance.
6. Some features with multiple categories Ire reduced to include only the most important ones, such as property type.

Supplementary Exploratory Analysis: For analyzing the first objective of initial acquisition I developed graphs to identify how many new properties Ire added over the past few years. And surprisingly there was a sharp decline in the new properties added after 2016. The reason associated was due to the Anti-Airbnb bill passed in New York which made it illegal to advertise short term rentals for entire homes. The attempt was to target landlords who bought apartments and used them to operate illegal hotels.

List of variables selected to use in the final model:

“high_booking_rate”,
“host_is_superhost”,“amenities_.Self.check.in.”,“host_response_time”,“amenities_.Hot.water.”,“amenities_.Free.street.parking.”,“amenities_.Dishes.and.silverware.”,“amenities_Refrigerator”,“amenities_.Coffee.maker.”,“host_listings_count”,“amenities_Microwave”,“cancellation_policy”,“availability_90”,“availability_60”,“availability_30”,“availability_365”,“amenities_.Extra.pillows.and.blankets.”,“longitude”,“amenities_.Cooking.basics.”,“amenities_.Luggage.dropoff.alloId.”,“latitude”,“amenities_.Bed.linens.”,“amenities_.Hair.dryer.”,“amenities_Keypad”,“minimum_nights”,“extra_people”,“cleaning_fee”,“amenities_.Family.kid.friendly.”,“guests_included”,“price”,“property_type”,“amenities_Shampoo”,“amenities_Stove”,“amenities_Iron”,“amenities_.translation.missing..en.hosting_amenity_50.”,“review_scores_value”,“amenities

_.Long.term.stays.alloId.","price_score","amenities_.24.hMy.check.in.","host_response_rate","amenities_Hangers","amenities_.Garden.or.backyard.","amenities_Oven","amenities_.Fire.extinguisher.","amenities_Gym","amenities_.Laptop.friendly.workspace.","amenities_Elevator","maximum_nights","amenities_.Patio.or.balcony.","review_scores_accuracy","review_scores_checkin","amenities_Lockbox","host_verifications","amenities_.First.aid.kit.","amenities_.Carbon.monoxide.detector.","security_deposit","amenities_Internet","review_scores_communication","amenities_.Single.level.home.","amenities_Pool","amenities_Essentials","amenities_.Pets.live.on.this.property.","amenities_.Bedroom.comforts.","amenities_.Bathroom.essentials.","amenities_.Toilet.paper.","amenities_.Ill.lit.path.to.entrance.","amenities_Dog.s.","amenities_.Bath.toIl.","amenities_.Smoke.detector.","bathrooms","accommodates","review_scores_location","amenities_Wifi","review_scores_value","review_scores_cleanliness","host_identity_verified","is_location_exact","requires_license","require_guest_profile_picture","amenities","host_sinceD","amenities_.Pack..n.Play.travel.crib.","amenities_.No.stairs.or.steps.to.enter.","amenities_.Body.soap.","amenities_.Room.darkening.shades.","amenities_.Safety.card.","amenities_Bathtub","amenities_.Children.s.books.and.toys.","amenities_Heating","amenities_.Hot.tub.","amenities_.Wide.entrance.for.guests.","amenities_Dishwasher","amenities_.Extra.space.around.bed.","amenities_Cat.s.","amenities_Netflix","amenities_.Cable.TV.","amenities_.Luggage.dropoff.alloId.","amenities_.Smart.TV.","amenities_Internet","bedrooms","amenities_Printer","amenities_.Disabled.parking.spot.","amenities_.Mini.fridge.","amenities_.Gas.oven.","amenities_.24.hMy.check.in.","amenities_.Full.kitchen.","amenities_Washer", "amenities_Other","instant_bookable","amenities_.Formal.dining.area.","room_type","review_scores_ratingOG"

Kaggle model information regarding specific market Variable Selection Process:

Model from Kaggle gave us an initial understanding of all the features. I used the variable importance from the Kaggle model and used that to shortlist the few features and then trained the XGBOOST model on selected features for the New York data to finalize the features to use to get highest accuracy. One major strategy which I incorporated was using the area column for New York Specific market. Since I found some discrepancies in the market column I used latitude and longitude columns to map the values into the area column which was then divided into 50 parts. This significantly improved the efficiency of the model.

Reasons for choosing specific variables.

Amenities: All the features of amenities Ire selected from feature importance after running the model with all the amenities and shortlisting to 55 amenities. host_is_superhost,host_listing_count,,host_response_time,host_response_rate,host_verifications,host_identity_verified: Host related features Ire selected because host_is_superhost was one of the most important features for high_booking_rate from the initial model.

host_sinceD: This variable represents the number of days the host has been listing his/her property with Airbnb, also it is an important feature on deciding whether a host is a superhost. And it was ranked very highly in the results of feature importance.

Review scores features: All the review score features such as review scores rating, review scores accuracy etc are included because, according to My intuition, reviews are the first thing users look at and since I don't have actual reviews of the property I are assuming these are quantifications of the actual reviews.(review_scores_rating is renamed to review_scores_rating)

Latitude and longitude: After looking at the distribution of price and high booking rate across New York, I found out that some areas Ire cheap but still had high average booking rate thus I selected latitude and longitude. Accommodates, beds, room type Ire used because of its descriptive nature in explaining size of property.

Cancellation policy: Cancellation policy was used because on Airbnb website cancellation rate was mentioned as an important factor in deciding whether a host is superhost.

Availability_90,availability_30,availability_60,availability_365: All four of the listing availability variables Ire used because even though they have high collinearity amongst them My final model used was XGBOOST model and XGBOOST output is robust to multicollinearity issues. I did not want to miss out on any important variables. Kaggle model on the entire dataset had availability_30 as the most important variable. For the New York market behavior could have been different because of people booking listings in advance, being the most popular city in the USA.

Price: This is rent of the property. Price is probably the second most important factor that users look at while booking thus included.

Remaining variables used are explained in Appendix.

Predictive analysis and guiding process for investors.

After working out various models, the XGBOOST model was chosen pertaining to superior performance measures. At first I incorporated random forest however lower specificity and accuracy made us shift to XGBOOST model for predictive analysis. In the most layman terms, using this model an investor can input information pertaining to all the variables and the model will predict whether the booking rate is high or low. But based on the business proposal and analysis done, this finding and analysis will not only guide novice and veteran hosts for acquiring properties, but this will also give them the overview based on research proposals regarding what all factors they can add or modify to acquire a high booking rate.

V. Results and Findings:

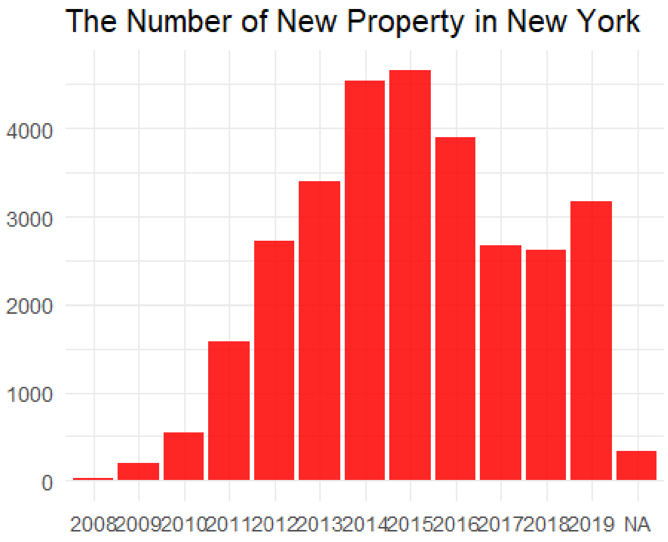
```
##                                Feature         Gain         Cover     Freque
ncy
##   1:            review_scores_ratingOG 1.237679e-01 0.0912641485 0.0407174
653
##   2:            host_is_superhost.TRUE 7.632563e-02 0.0144195258 0.0035349
568
##   3:                        host_sinceD 6.312722e-02 0.0835046509 0.0704372
872
##   4: amenities_.Free.street.parking..0 4.478391e-02 0.0077597596 0.0043205
027
##   5:        review_scores_value.medium 4.278194e-02 0.0094862870 0.0060225
190
##  ---
## 124:        review_scores_location.nan 1.956930e-04 0.0020214850 0.0002618
487
## 125:                   market.New York 1.688014e-04 0.0031407853 0.0006546
216
## 126:                   amenities_Wifi.0 1.168455e-04 0.0007942043 0.0003927
730
## 127:        property_type.Guest suite 1.153714e-04 0.0001508616 0.0001309
243
## 128:              property_type.Hotel 4.917525e-05 0.0009044168 0.0002618
487


##           Confusion            Matrix           and            Statistics
##
##                                                                 Reference
##        Prediction                            0                          1
##                                                   0      6428         319
##                                                   1           780      1581
##
##                                                     Accuracy    :   0.8793
##                                            95%   CI   :   (0.8725,   0.886)
##                                 No      Information     Rate      :    0.7914
##                           P-Value     [Acc      >      NIR]     :     <     2.2e-16
##
##                                                        Kappa    :   0.6645
##
##             Mcnemar's       Test       P-Value       :       <       2.2e-16
##
##                                                  Sensitivity    :   0.8918
##                                                  Specificity    :   0.8321
##                                         Pos     Pred      Value     :   0.9527
##                                         Neg     Pred      Value     :   0.6696
##                                                   Prevalence    :   0.7914
##                                           Detection     Rate     :   0.7058
##                                 Detection         Prevalence         :      0.7408
##                                        Balanced       Accuracy      :   0.8619
```

```
##
##                  'Positive' Class : 0
##
```

I have used XGBOOST as My model for predicting booking rate for the market of New York. Model has accuracy of 87.93 for the threshold that I have kept as 0.25 with high sensitivity of 0.8918 and specificity of 0.8321. Since My first business proposal is for investors looking for new acquisition of property I want to present them with as many correct options as possible. Thus I have kept the threshold of My model at 0.25 in order to increase the specificity and maintain tradeoff between accuracy, specificity, and sensitivity. Even though this impacts the sensitivity and accuracy by a small margin it is not significant and for My business proposal this threshold is more appropriate.

New properties added in New York over time.

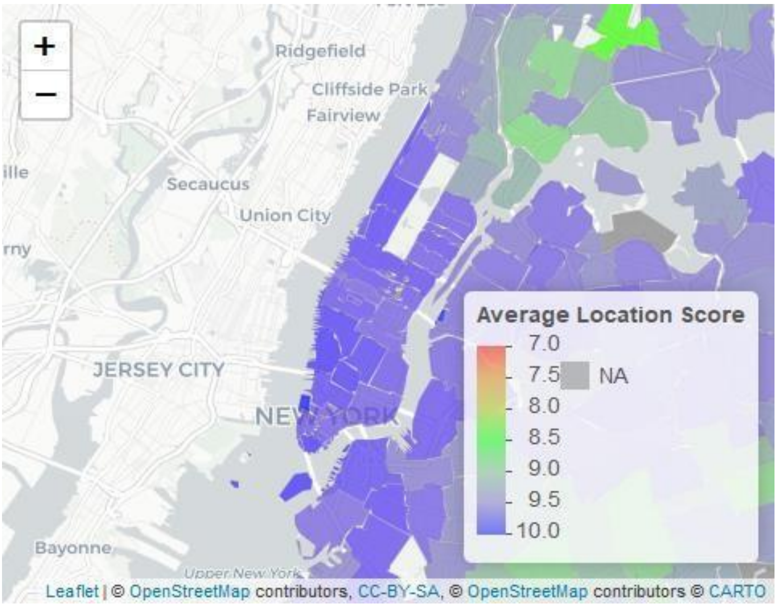


As can be seen from the graph, the addition of new properties has an increase till 2015, after that the number of new properties added is decreasing. After analyzing it, I found out the reason for the decrease is due to an anti-Airbnb bill that was passed in New York on 16th June, 2016. According to that bill it became illegal to advertise short term rentals (less than 30 days) for the entire home.

Q1 Where in New York should investors invest?

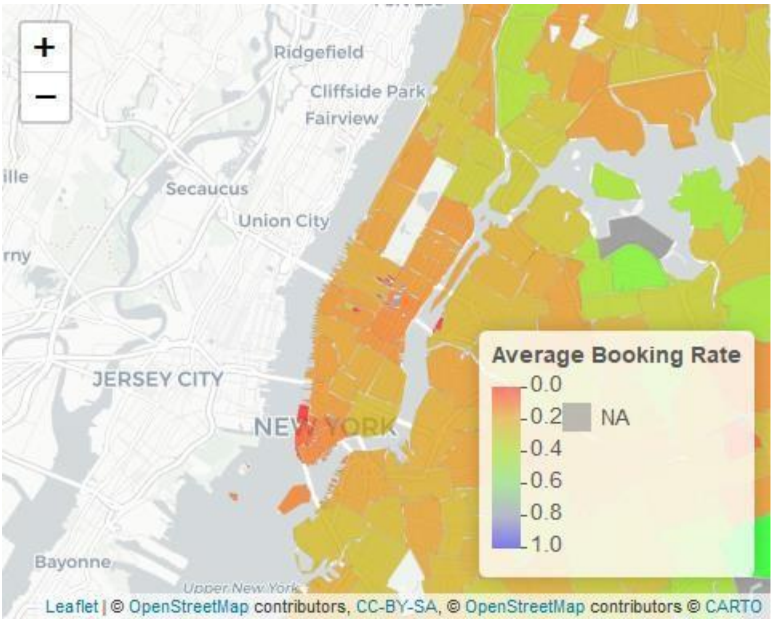
To answer this question I need to find out areas of New York which are most profitable to invest. For this I have analyzed the data based on the average location score of the different boroughs in New York and the average booking rate and the average rental price of the listings in these boroughs.

1.a Which neighborhoods have high review scores for location?



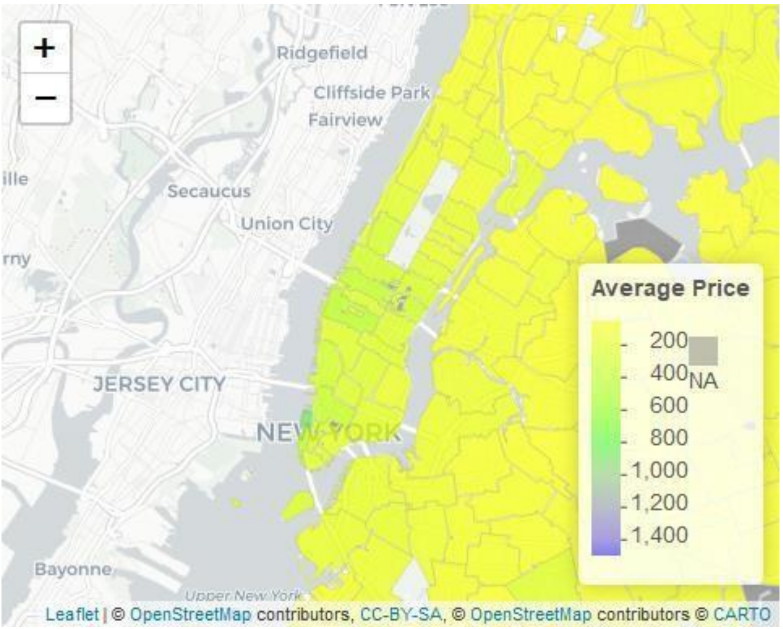
In Manhattan, the downtown neighborhoods, near the Empire State building, Times Square, Financial District and the neighborhoods near and to the west of Central Park have an average score around 9.9. In Staten Island, the southern areas close to tourist attractions like museums and ferry stations have a score of 10. The parts of Brooklyn and Queens that the closer to Manhattan, have better scores than the parts that are farther away. Perhaps, this could be because the subways lines go to the parts closer to Manhattan and not to the other parts. The north-west coastal part of the Bronx has high scores, which could be because of coastal attractions in that part.

1.b Which neighborhoods have a high booking rate?



It is interesting to observe that in all the five boroughs, the locations that have higher average location scores have lower booking rate. This can be explained by the rental price in the next plot.

1.c How does Airbnb rental price vary across the neighborhoods?



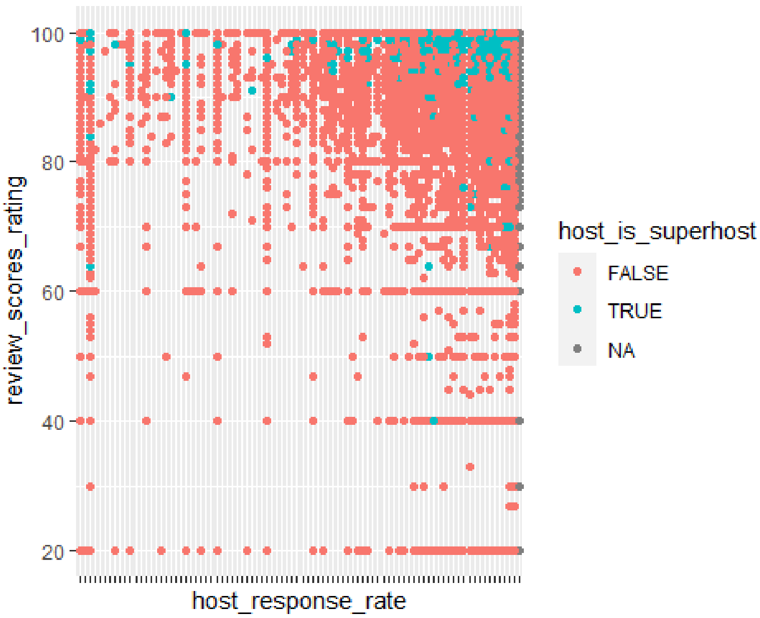
In all the boroughs, the average price is higher in the neighborhoods with higher location scores. Neighborhoods with high location scores are the most desirable ones. They can be priced high, but they also have very high purchase costs. The real estate in desirable neighborhoods of New York is very expensive. The investor has to charge a high rental price to recoup the purchase cost or to pay the mortgage. Also, they cannot be priced higher than hotels, since hotels tend to have many amenities, like pool, spa, restaurants, etc. These high-priced rentals would have lower booking rates and would be popular with the affluent tourists. The return on the investment would be by margin rather than by volume of bookings.

Q2 How to be a superhost in the New York market?

Airbnb has mentioned requirements to be a superhost on their website, but it is in general for all the markets. This question helps in finding out what are the factors different from one mentioned on Airbnb website to become superhost in market of New York? This question will help us understand what superhost does more or less than the normal host. I have asked this question because whether a host is superhost or not is one of the most important feature affecting booking rate according to My model, thus detailed analysis of this feature was required.

Q2 a. Is there correlation between review scores rating, host response rate and whether the host is superhost?

Airbnb website mentions that to be a superhost host should maintain a 90% response rate and higher and more than 4.8 overall review score rating. Let us validate that for the market of New York.



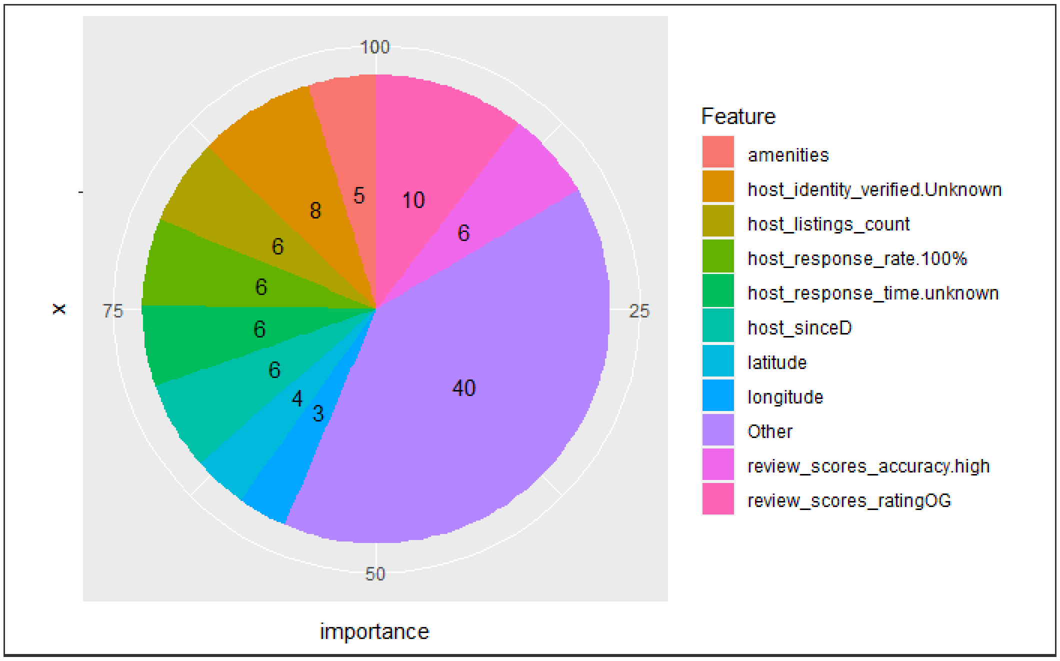
From the above graph it can be inferred that superhost usually has high review scores rating and high response rate. This validates the requirements mentioned on the Airbnb website. But I are interested in other features affecting superhost status.

Q 2b. What are all the features important for a host to be a superhost in the New York market?

```
##                                   Feature         Gain        Cover      Frequenc
y
##    1:             review_scores_ratingOG 1.038180e-01 5.120519e-02 0.034123167
8
##    2:  host_identity_verified.Unknown 8.096057e-02 1.307994e-02 0.003765714
6
##    3:       review_scores_accuracy.high 6.124482e-02 6.648065e-03 0.004345055
3
##    4:                       host_sinceD 6.080144e-02 1.128512e-01 0.087538381
3
##    5:          host_response_rate.100% 6.044062e-02 9.011701e-03 0.006372747
8
##                                                                              ---
## 128:                   market.New York 8.397625e-05 3.726905e-04 0.000231736
3
## 129: review_scores_communication.low 5.319995e-05 9.935074e-04 0.000289670
4
## 130:               property_type.Other 2.711482e-05 5.871866e-06 0.000115868
1
## 131:                  room_type.Hotel room 1.897139e-05 8.122086e-04 0.000115868
```

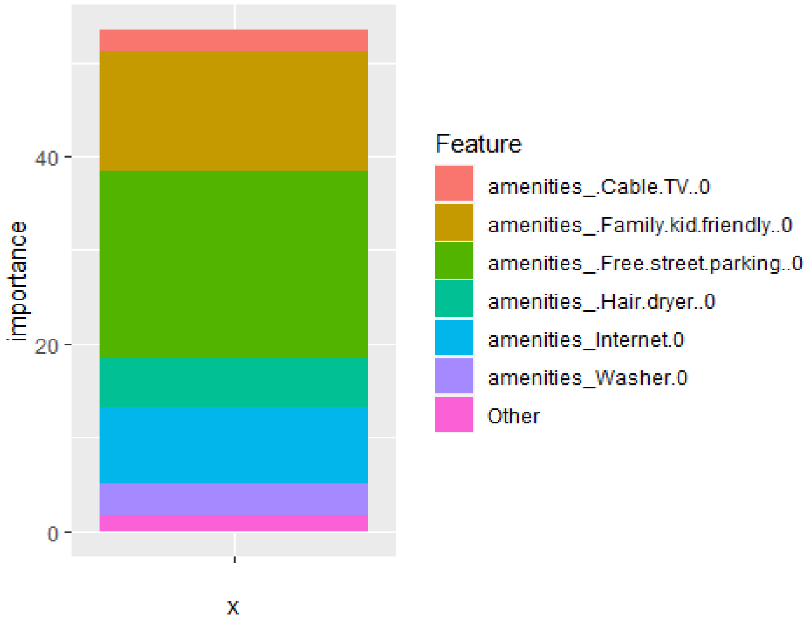
```
1
## 132:      review_scores_accuracy.nan 1.409562e-05 1.272861e-06 0.000115868
1
```

This is a table of all the features in the data and their importance gain towards deciding host status from the model of XGBOOST.



It can be seen from the above pie chart that 10 features contribute to almost 60 % importance for whether a host is superhost. Features listed on Airbnb website in requirements to become superhost which can be seen in the above top 10 features are host_response_rate, review_score_ratings, host_response_time. Other important points to notice here are superhost generally get good ratings in review scores and they are most likely Airbnb hosts for a long time with multiple properties. Area is also important for becoming a superhost along with total amenities provided.

Q3 What amenities are most important for Airbnb users in the New York market?



For the market of New York most important amenities are free street parking, family kid friendly, internet and washer. Thus hosts can make sure that they are able to provide these amenities in order to increase their listings in booking rate. Amenity translation missing is amenity written in a language other than English. Thus, it is possible that people who speak that language are attracted by that specific amenity.

VI. Conclusion and Discussion:

Obtained findings and final summary:

In order to achieve a high booking rate throughout the year, the investor should acquire a property in a location that is popular with both tourists and business travelers. Hence latitude and longitude play a major role which is evident through both the model and visualizations and hence location selection is vital. Based on those results, I divide My suggestion in two parts: If it is a new investor, then My advice would be to invest in areas near Staten Island Mall, Orchard Beach and Pelham Bay Park. If it is a veteran investor, then My advice would be to invest in areas like Central Manhattan and the area near the airport. Now moving onto the scenarios after the acquisition, I conclude that being a superhost does make a significant positive influence on high booking rate. For the New York market, having more number of properties, having a high response rate, high review rating, cleanliness and the location of the property are important for being a super host. On analysing the different amenities, I concluded that having the availability of free street parking is very important and that makes sense as New York is a very busy and crowded area. Having a Family/Kid friendly tag also has a high importance. Including other basic amenities like internet and coffee maker, would also contribute in getting a high booking rate.

Limitations:

Imbalanced data: There was lack of data for multiple locations as Ill as unreliable data such as that of Manhattan where data, though ample, has multiple 0 valued locations and fewer 1's. Thus, the booking rate is low but this is not accurate and actually not the case. Missing GeoJSON information for the zip codes that are in New Jersey, about 1800 New Jersey listings, out of the 32,143 total listings in New York. For the initial acquisition analysis, the current data on property mortgage prices is not available. This makes the analysis harder and less accurate. In the superhost analysis, a factor that makes a host a superhost is having less than 1% cancellation rate. The data on the cancellation policy was missing so this creates inaccuracy.

Future Research:

From the above limitations it is evident that the data is imbalanced thus the most important thing would be to add more data for each borough of New York to get accurate predictions. To avoid discrepancies which I observed in the location data a more accurate market column would be extremely beneficial. Including the column for property buying or market rates would also significantly improve the model in predicting booking rate for initial acquisition of property. Adding text user reviews might also give us some interesting insights by performing sentimental analysis on them. Cancellation policy prices are absent and adding them might be fruitful for superhost analysis. Last but not the least, the better the model and hence filling the missing values, getting a more enriched data file would make the model and in turn this project extremely valuable for future investors.

VII. References:

https://rpubs.com/jhofman/nycmaps https://stackoverflow.com/questions/43446802/how-to-download-ny-state-all-county-data-in-r-for-leaflet-map https://stackoverflow.com/questions/42543206/r-markdown-compile-error (Megan Rose Dickey, 2016), https://techcrunch.com/2016/06/17/airbnb-new-york-legislation/ https://www.airbnb.com/superhost https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-XGBOOST-parameter-tuning-r/tutorial/ Introduction to Supervised learning in R (For XGBOOST) https://learn.datacamp.com/cMyses/supervised-learning-in-r-regression Predicting Airbnb prices with machine learning and location data https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a https://www.analyticsvidhya.com/blog/2016/01/XGBOOST-algorithm-easy-steps/

Viii. Appendix:

Explanation of why these variables are used.(other variables are explained in methodology)

Minimum and maximum nights: While renting the property what is the minimum or maximum days it is allowed to book are one of the most important factors which are looked.

Extra people: Intuition is higher the limit of extra people allowed higher the booking rate will be , thus, to validate that this variable is included.

Cleaning_fee: Higher the cleaning fee lower the attractiveness of the listing, thus important to be included.

Property_type: In New York, some property types such as houses could be more attractive than apartments. Thus it is important to add this variable.

Security_deposit: Property with high security deposit is generally very expensive, thus this variable is almost the same as price. But there could be some listings whose rent price is low but still they are charging high security deposits , so incorporating those listings in My model I have used a security deposit.

Area: This is a substitute for a market column in the entire data but at a more granular level. This is created by forming clusters from latitude and longitude columns. Reason is that the market column had wrong listings listed in the wrong market and to create submarkets from one market.

Price_score: This is a derived variable which was created by making clusters from accommodates beds and bedrooms. Then by using the area column and this clusters percentile was created of price for each subgroup. This in the sense gives us the idea of how expensive the listing is for similar size other listings in a similar area.

Bedrooms: It was evident from feature importance that this column was vital and logically as Ill while renting customer totally takes into consideration how many rooms he wants.

Is_location_exact: This can be crucial in determining whether the given location is exact and if it is not then that might factor into a customer's decision to book a hotel.

Bathrooms: The count of bathrooms in the property seems to be something that customers actively look for. So having a certain minimum number of bathrooms is needed.

Room type: Along with the count, it is required to understand different room-types because customers can change their decision if there is no kitchen.

Require_guest_profile_picture: Knowing the guest is really crucial these days and to have trustworthy guests host need to have a general idea.

Requires_license: Identification details are vital in to avoid any issues and keep a tab and this marks the authenticity of Airbnb which in turn affects booking rate.

Instant_bookable: This is a highly vital feature as evident from feature importance, and also it makes sense logically as it would be crucial for a customer to be able to book the property instantly, rather than waiting for the host to confirm their booking.