INFO5430

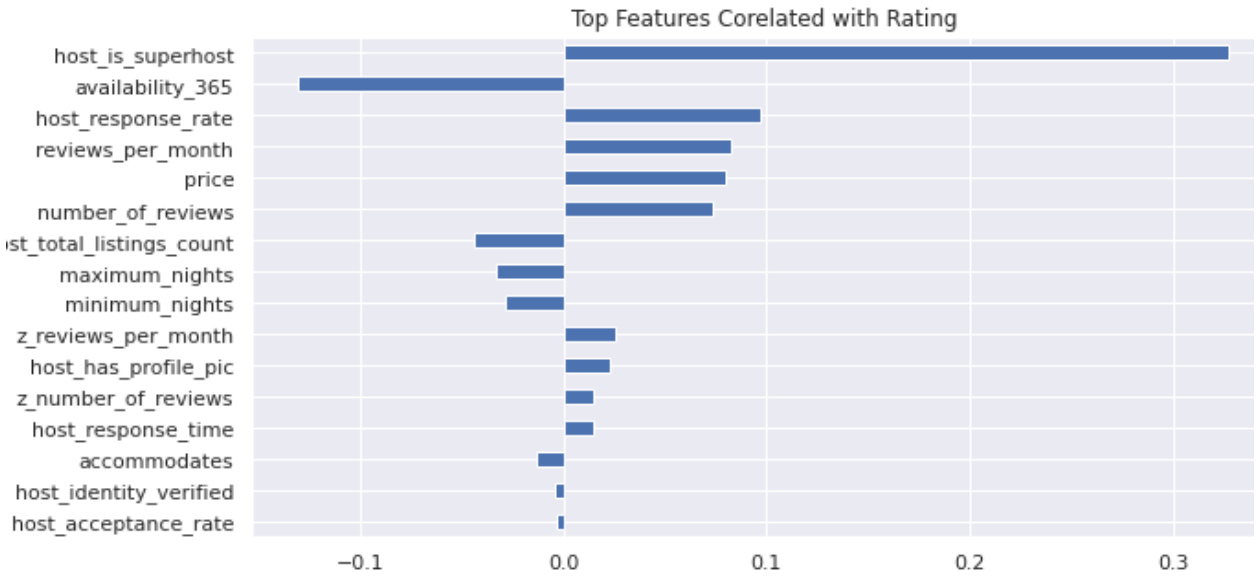
Group member: Linjing Rao , Minyu Huang , Yubang Wu

Staying at an Airbnb instead of a regular hotel has become the new fashion for travelers around the world, the temptation to get a taste of the local lifestyle in the exotic land ranging from a cheap living room coach stay to a luxurious private mansion can be irresistible. This rising demand has attracted many hosts to join the market, New York City even has more Airbnb listings than apartment rentals. But once the host signed up and joined the platform, how to win the 5.0-star rating becomes the ultimate challenge for sustaining the popularity of their airbnb. With the detailed listings and reviews dataset of New York City, provided by Inside Airbnb, our analysis aimed to unlock the secret ingredient for a 5.0 start airbnb. We analyzed the impact of location, airbnb host response rate, the amenities provided and the review comments. The results show that location is not necessarily the most influential factor for a top 5.0 airbnb but good location does make your airbnb more popular. And with more people staying, the more likely you are to get good review comments and achieve that 5.0 Starts.

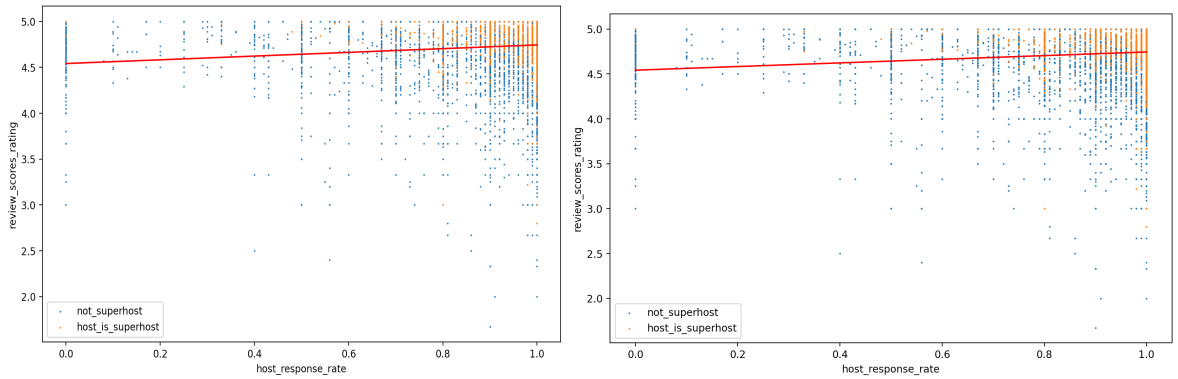
To be more specific, we start with data exploration by first looking at the numeric numbers. We tried to find features that are highly correlated with ratings. Whether or not hosts are super hosts was found to be highly correlated with ratings (a correlation of 0.32), and so is the number of reviews, price, and hosts' response rate.

INFO5430

Group member: Linjing Rao , Minyu Huang , Yubang Wu



Since we want to provide insights for hosts to improve their ratings, simply suggesting ‘become a super host’ or ‘increase your price’ won’t be helping. Instead, we were trying to find actions such as ‘increase your response rate’. We plotted the correlation between response rate and ratings and found that the host who responds more to the tenants tends to have higher listing ratings. We also found that a higher response rate also means more reviews numbers, which could be some icing on the cake.

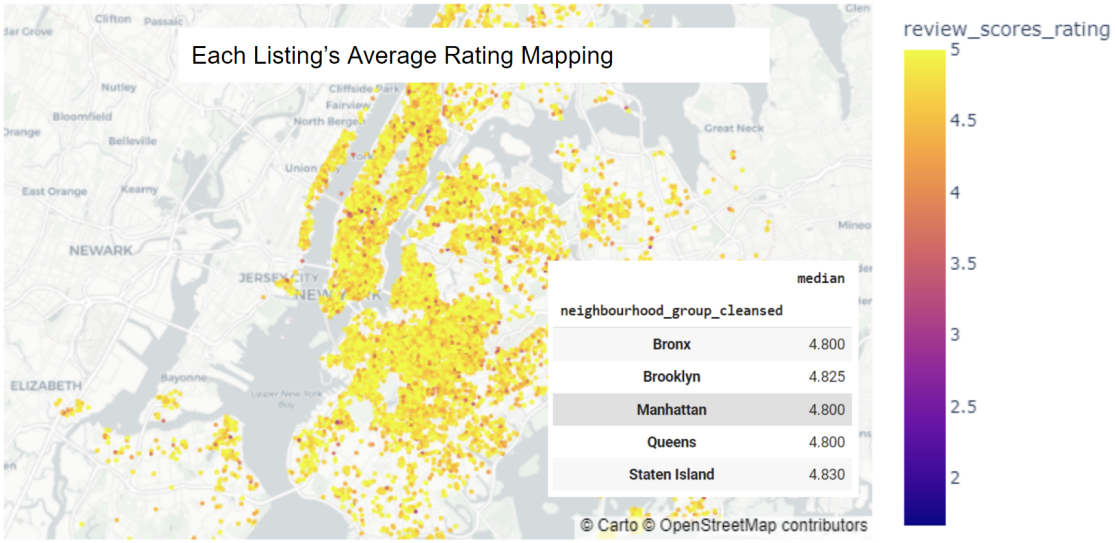


INFO5430

Group member: Linjing Rao , Minyu Huang , Yubang Wu

After analyzing the numeric data, we expand our scope to text-based categorical data which contains more useful information. For example, where does the listing located, what are the amenities provided, and what do customers comment?

First, we ask ourselves where most 5.0-star Airbnb locations and do locations impact the rating. The initial finding is that borough-level location (which part of the city the Airbnb is in) doesn't determine ratings. We found this when mapping each listing's average rating onto the NYC map. Both the high ratings and low ratings are somehow evenly distributed around all boroughs. To prove this, we also calculated the median listing rating of each borough, finding out that there is no statistically significant difference between them.



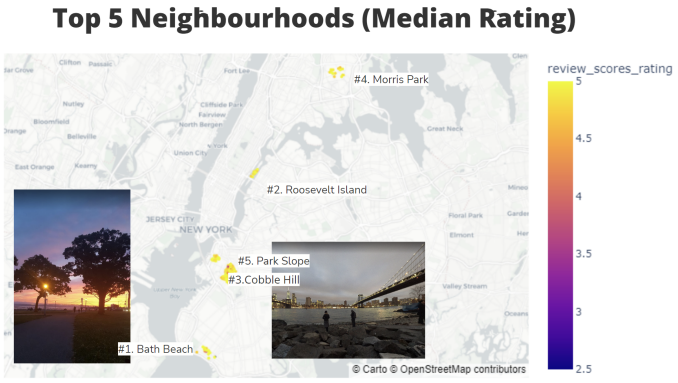
neighbourhood_group_cleansed	median
Bronx	4.800
Brooklyn	4.825
Manhattan	4.800
Queens	4.800
Staten Island	4.830

However, we did spot some clusters of high ratings or low ratings on the map, so we zoomed into the level of the neighborhood and calculated their median ratings (excluding the outliers and neighborhoods with less than 15 listings. We pulled out the top 5 neighborhoods in terms of their

median rating and mapped them on the map. The top neighborhoods in terms of ratings are Bath Beach and Roosevelt Island, all having beautiful scenery.

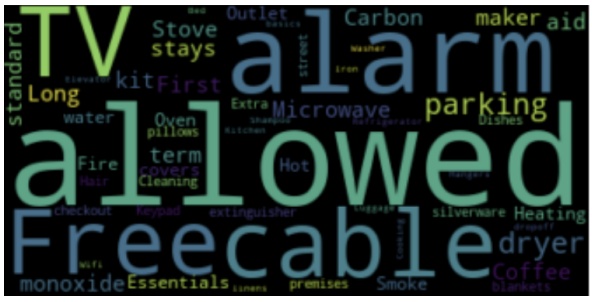
	neighbourhood_cleansed	count	median
5	Bath Beach	18	4.985
183	Roosevelt Island	16	4.955
45	Cobble Hill	17	4.930
142	Morris Park	17	4.930
161	Park Slope	112	4.910
...	...	...	...
227	Williamsbridge	38	4.700
137	Midtown	377	4.690
204	Theater District	65	4.670
146	Mount Hope	24	4.595
167	Port Morris	24	4.520

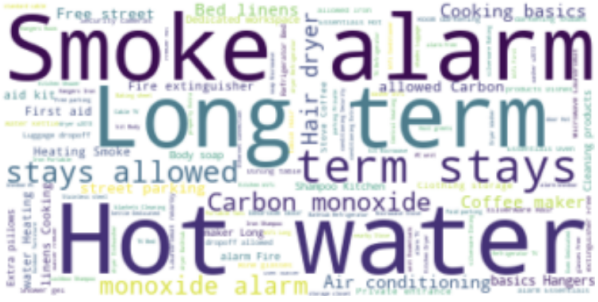
125 rows × 3 columns



Since exhaustive data on amenity is provided in the dataset, we tried to find the correlation between amenities and Airbnb ratings by machine learning. It is interesting to analyze whether the highest number of amenities has a positive correlation with the rating. Are there particular amenities that can largely increase the guest's favorability and are more likely to receive five-star reviews?

First of all, it can be observed from the dataset that the information in the amenities column is expressed through text. Through the world cloud map, we can find some facilities that are significantly related to each other.





INFO5430

Group member: Linjing Rao , Minyu Huang , Yubang Wu

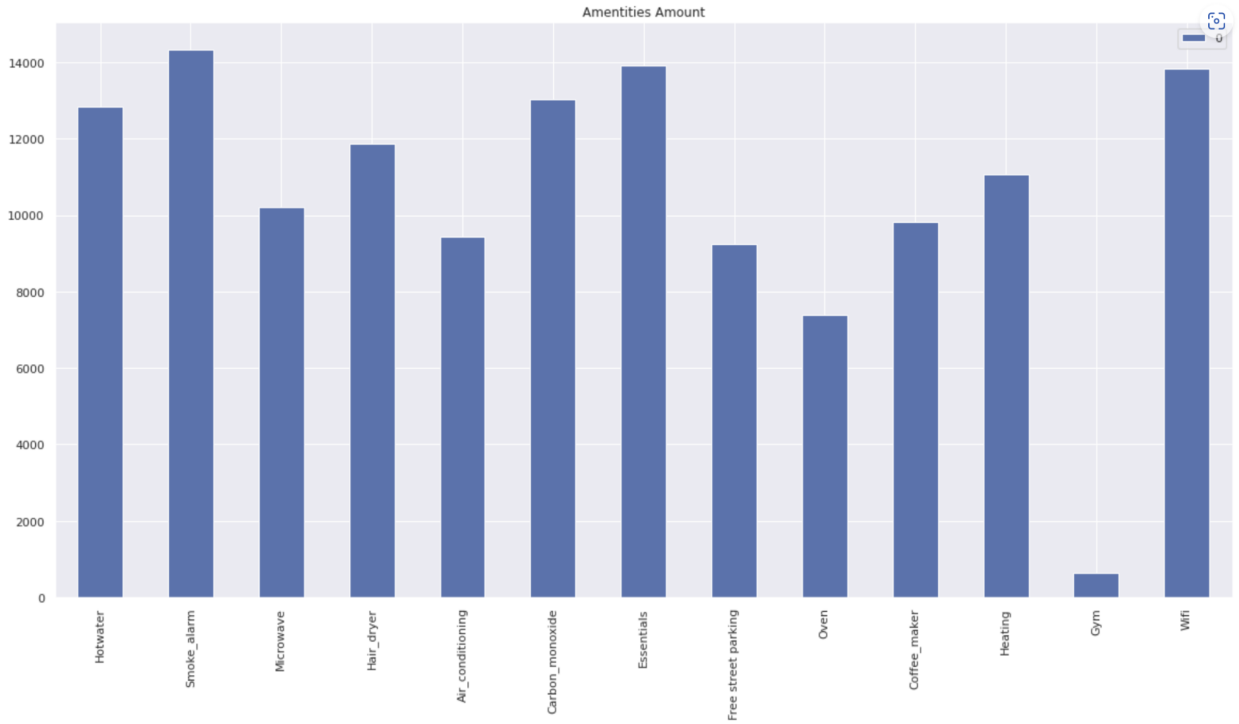
Using the command string split, all the facilities that are significantly related can be distinguished and recode them. The facilities are coded as 1 if they are present and 0 if they are not, thus forming a new and cleaner dataset.Then we plot the sum of the Amenity Amount, we can see that the number of necessities such as hot water, wifi, essentials is the most, while the number of coffee machines, ovens, gyms and so on is less, this data is basically in accordance with the prediction.

	amenities	review_scores_rating
0	["First aid kit", "Microwave", "Stove", "Coffe...	4.89
3	["Hair dryer", "Essentials", "Carbon monoxide ...	4.42
6	["First aid kit", "Coffee maker", "Gym", "Long...	4.90
8	["Essentials", "Oven", "Cable TV", "Refrigerat...	4.34
9	["Coffee maker", "Gym", "Long term stays allow...	5.00
...	...	...
39871	["Microwave", "Coffee maker", "Smart lock", "L...	5.00
39873	["Essentials", "Lock on bedroom door", "Heatin...	5.00
39876	["First aid kit", "Microwave", "Stove", "Coffe...	4.91
39877	["First aid kit", "Microwave", "Laundromat nea...	4.79
39878	["Hair dryer", "First aid kit", "Essentials", ...	4.45

15124 rows × 2 columns

	review_scores_rating	Hotwater	Smoke_alarm	Microwave	Hair_dryer	Air_conditioning	Carbon_monoxide	Essentials
0	4.89	1.0	1.0	1.0	1.0	0.0	1.0	1.0
3	4.42	1.0	1.0	0.0	1.0	1.0	1.0	1.0
6	4.90	1.0	1.0	0.0	1.0	1.0	1.0	1.0
8	4.34	1.0	0.0	1.0	0.0	1.0	0.0	1.0
9	5.00	1.0	1.0	0.0	1.0	1.0	1.0	1.0
...	...	...	...	...	...	...	...	...
39871	5.00	1.0	1.0	1.0	0.0	0.0	1.0	1.0
39873	5.00	0.0	1.0	0.0	0.0	1.0	1.0	1.0
39876	4.91	1.0	1.0	1.0	1.0	1.0	1.0	1.0
39877	4.79	1.0	1.0	1.0	1.0	0.0	1.0	1.0
39878	4.45	1.0	1.0	0.0	1.0	0.0	1.0	1.0

15124 rows × 14 columns



Then, we tried to use machine learning to analyze the relationship between each facility and the review score rating. The correlation between each facility and rating can be expressed by Coefficient. This result also predicts whether the facilities have a positive effect on the guest ratings.

First, Rigid regression was used because the dependent variable, review score rating, is a continuous variable, however, Rigid regression revealed that the model scores were low. Although there was a significant change in Coefficient. However, the overall coefficients were small, and it was speculated that this might be because all the independent variables were only 0 and 1. So we used logistic regression instead. Based on logistic regression, we coded the dependent variable greater than 4.8 as 1 and less than 4.8 as 0. After the fit model, we found that the test score was 57% and the coefficients also changed significantly, so we decided to use this model.

	0
Hotwater	-0.017391
Smoke_alarm	0.261699
Microwave	-0.004840
Hair_dryer	0.215889
Air_conditioning	-0.128408
Carbon_monoxide	-0.015683
Essentials	-0.018540
Free street parking	-0.081925
Oven	0.068705
Coffee_maker	0.405963
Heating	-0.340123
Gym	0.053992
Wifi	-0.462449

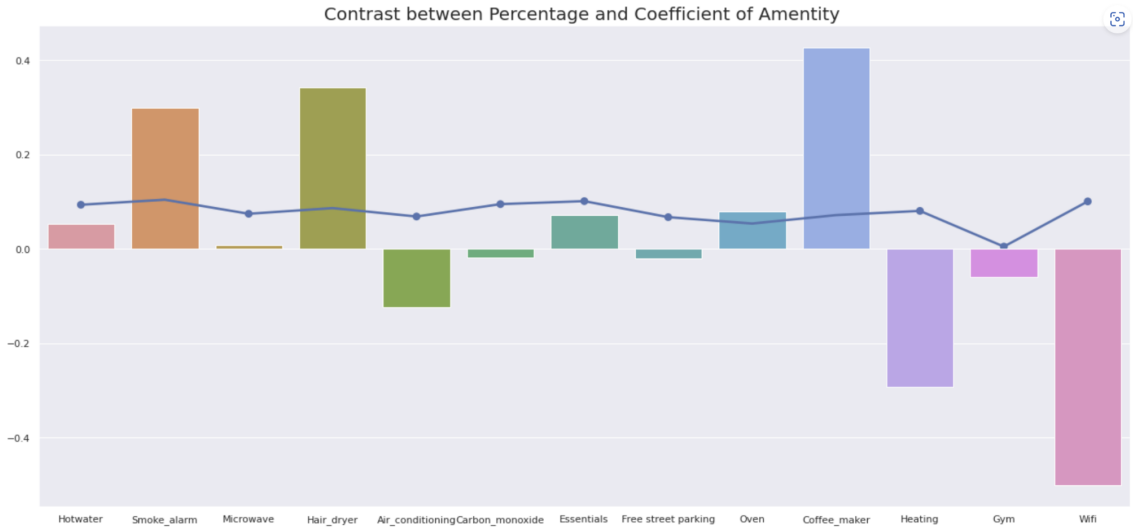
Test Score:8%

	0
Hotwater	-0.017391
Smoke_alarm	0.261699
Microwave	-0.004840
Hair_dryer	0.215889
Air_conditioning	-0.128408
Carbon_monoxide	-0.015683
Essentials	-0.018540
Free street parking	-0.081925
Oven	0.068705
Coffee_maker	0.405963
Heating	-0.340123
Gym	0.053992
Wifi	-0.462449

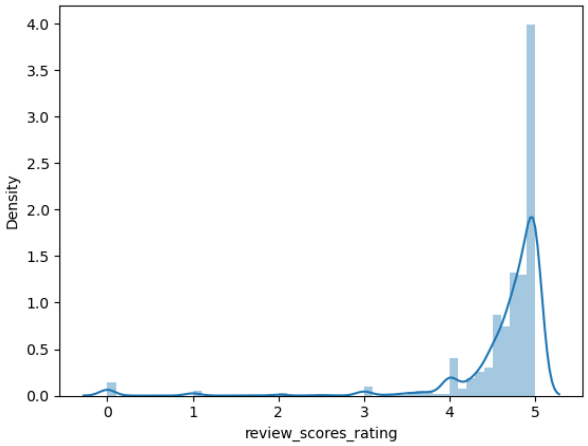
Test Score: 57%

A clear comparison between Percentage and Coefficient of Amenities can be seen in the derived graph. But this result is different from what we expected. The graph shows that common amenities such as hot water and wifi have a strong negative correlation with the rating. This is probably because if these common amenities were not available, it would easily result in a very low rating. In addition, some uncommon amenities such as coffee maker, hair dryer, and oven

have a positive correlation to some extent. This may be because the presence of these less common amenities gives the guest a psychological good feeling.

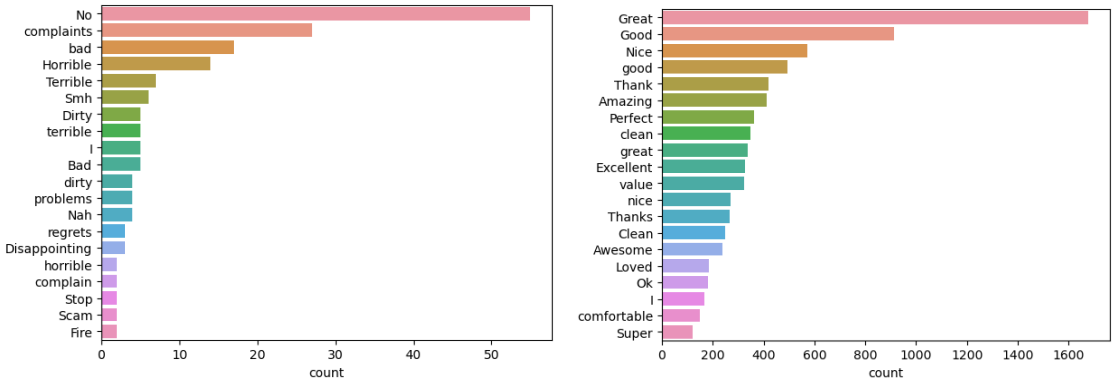


Analyzing what users say about their stay can be a direct way for hacking into finding the key elements that the market demands for a 5.0 star Airbnb. Using the nltk(natural language toolkit) in python to help us with the process of filtering out stopwords, non english sentences and



punctuations, we were able to conduct sentient intensity analysis and calculate the positive, negative, neutral and compound scores for each review. Surprisingly over 95% of the review comments tend to be positive, 3.5% neutral and 1.2% negative.

Using word cloud we found that the most commonly used words are about location and the general positive experience that the user had. The most frequent negative comment word was 'complaints' so most users would write about the location and for positive comments and when for negative it is mostly related to disputes with the host, so keeping a good relationship with the guest is certainly key to maintaining a perfect score.



Since we conduct analysis on the amenities, we also calculate the frequency of occurrences related to amenities in the review comments and found that rarely did people write about them. Specifically we searched for the occurrence of coffee maker since this was the feature that has the highest positive value impact from the amenity analysis, but the result was that none of the comments mentioned coffee maker.

However, this does point out that users tend to write about the good location and complement the stay experience, but they will not write about the specific amenities. Nevertheless, this does not mean that amenities are less important, because the positive correlation indicates that they do help with increasing the odds of getting higher scores and boost the experience.

INFO5430

Group member: Linjing Rao , Minyu Huang , Yubang Wu

Overall, if we were to give airbnb hosts tips to take away from the analysis of what makes a 5.0 start airbnb, we would suggest that they first find a good location so that the airbnb is more likely to be popular and receive more comments, secondly aim to be a superhost who response quicker to guests and earn more trust on the platform. Other than that for the amenities they should consider purchasing elements such as a coffee maker to boost the overall stay experience to the next level, but more importantly they should aim to maintain a satisfactory level stay that will prevent them from getting complaints through review comments.

Bibliography:

Get the Data. (2022). Retrieved 6 December 2022, from http://insideairbnb.com/get-the-data/

Complete Analysis of Airbnb Data - New York City. (2022). Retrieved 6 December 2022, from https://www.kaggle.com/code/scsaurabh/complete-analysis-of-airbnb-data-new-york-city

New York City Airbnb Open Data. (2022). Retrieved 6 December 2022, from https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data