Final Report

Group D

Team Number 1: Hong Seok Kim
Team Number 2: Jeongmoo Kwag
Team Number 3: Huiseong Yoo

Business Problem

Analyze differences in Airbnb rental activity between areas with different property values and economic characteristics in New York. Are there any trends or patterns from listing-specific values(review score) among real-estate or economic values between areas? If so, what can Airbnb do to attract more customers from the competitors?

Business Impact

Background

Fundamentally, Airbnb's utmost competitors are the hotels and accommodations. In terms of the success of the rental business, sustaining the numbers of existing customers for Airbnb and attracting more potential customers from the hotels is the main business goal we want to achieve. The focus of our analysis is on what factors customers care about and how they differ with real-estate or economic values. Unfortunately, even though what customers care about largely depends on their profiles, we often do not get valuable information other than the location they are going to be in, as they search for a place to stay. From our domain knowledge, we know that property values are highly correlated with locations, leading to the necessity of analysis with respect to interactions between property values and customer's satisfaction.

We expect to come up with a model that predicts the probability of having a satisfied customer according to different property values and Airbnb can make use of the model by recommending listings to customers with higher probability of satisfying them derived from the model. We will be able to figure out contributions of listing specific factors to higher ratings and advise landlords to improve on those factors or bring in new landlords that satisfy the factors.

Analytical Context

One of the attributes that we would like to mainly focus on is the review scores of the airbnb listings. Since customers seeking future listings do not have an idea whether the amenities are in fact good, unless they experience it, their decisions tend to depend on high satisfaction and experiences of previous users.

Also, improving on amenities or modification of such different features for customers' priorities is essential to have the upper hand over hotel industries. So, generating predictive models of review scores with reviews and amenities is necessary.

Data

For the exploratory data analysis, We will use three different datasets :

1. listings.csv
2. real_estate.csv
3. calendar.csv

1) The listings.csv is the dataset describing tens of thousands of Airbnb listings in five states. It is approximately 60,000 rows by 29 columns.

   Its strength is the size of the data and many attributes of the listings such as review_scores, price, weekly price, bedrooms, etc. and there is a column for the overall rating of a listing and columns for subcategories of the rating, enabling us to analyze what customers were particularly satisfied with and unsatisfied with.

   The weakness is that there are not many states and the data only covers from the second quarter of 2016 to the second quarter of 2018, which implies time series analysis is not suitable for the dataset, especially when we try to merge the listings dataset with the econ dataset, that is in quarters. And, there is no information on customers other than their ratings, so we have to assume their profile to be the same in all of the listings other than the location customers are looking to stay in.

2) The real_estate.csv is the dataset containing monthly ZHVI and ZRI data for every zip code in the US. ZHVI and ZRI represent home value and rent made by Zillow, an American tech real-estate marketplace company.

The strength of the dataset is that we can gain information about home value and rent prices of every zip code present in the listings dataset, so we can do in depth analysis of review ratings according to home value and rent prices.

The weakness of the dataset is that ZHVI reflects the typical value for homes in the 35th to 65th percentile range in the area, so not every house is taken into account in the ZHVI index.

3) The calendar.csv is the dataset containing dates of listings that were available from and available to and the dataset also contains the price and the metropolitan the listing was in.

The strength of the dataset is that we can gain information about the time frame of the dataset by merging the listings.csv and the calendar.csv.

The weakness of the dataset is that the data span relatively a very short period of time. So, again, some of the analysis methods are not suitable as a result such as Time series analysis, which we can get meaningful inference from.

Methods

Visualizations

For the Visualization, We first want to do the walkthrough of an EDA in order to process on datasets. Since the calendar.csv and listings.csv have both ids for listing we want to merge those datasets and drop unwanted columns. Then we do clean up the datasets by dropping the null values in zip code since we want to run analysis of property values by different areas and a very small portion of data points have null values in zip code, so dropping the null values will not change the distribution of null values significantly. Brief summary will help how the dataset is structured.

With those EDA, we want to visualize histogram, scatterplot, boxplots and heatmap. Histograms are useful when we want to see how data of interest are distributed and their densities. We use histograms to visualize distributions of review scores for listings of Airbnb,amenities, etc.

We also want to use scatter plots and boxplots. Scatterplots are useful for any type of data, whether the data is categorical or numerical. Boxplots are useful for categorical data and a major advantage over scatterplots is that we can gain useful summary statistics such as median and the interquartile range. We are using boxplots in order to see whether the number of bedrooms and room type does affect the review scores of ratings and scatter plots are used to investigate effects of interactions between property values and review scores on overall review scores.

Heatmap is also useful since we want to check out which zip code has a higher home value index. By using heatmaps, we will be able to investigate the difference of Home values according to locations and check whether the home value index does correlate with review scores. From the heatmaps, we conclude that we limit our listings data to only come from New York for a reliable analysis.

After exploratory data analysis, correlation matrix will be plotted for the feature selection before constructing any prediction model. Correlation matrix helps us sort out variables that are correlated, serving as a preliminary solution to alleviating multicollinearity problems we can face in a model.

Upon completing feature selection, we can finally construct models. The models we used generate binary outcomes, 0 and 1, so plotting AUC(area under the curve) makes it easier to evaluate visually how accurate models are. Also, a confusion matrix will also be presented to show the accuracy of a model.

Data wrangling and cleaning

In this part, we will explain how datasets are cleaned and merged and what kind of feature engineering is done and why it is necessary.

As it was mentioned in the visualizations part, calendar.csv and listings.csv were merged to extract the time frame of listings.csv, so that the mean property values can be computed according to the time frame. Then, by merging the merged dataset with real_estate.csv again, each listing in the listings dataset will have a column for property value, so that property value can be included as a predictor when constructing a model.

For null values handling, we fill null values with median values if the proportion of null values we deem is relatively too high to drop. By dropping null values without careful consideration, we might lose valuable information as dropped null values might contain important features. The reason as to why we chose specifically filling null values with median values will be explained in the milestone section.

In the amenities column of the listings dataset, a complete list of amenities in a listing is in a string separated by commas, several methods are applied to make them into categorical variables, so that we can analyze the significance of each amenity on overall review scores rating. First method was to replace all the parentheses to a blank string. Then the CountVectorizer function was applied to split strings by commas and convert it to the vector. After applying this method, the amenities column which was a list of strings is now in the form of a categorical variable.

K-means clustering is done to convert the property value variable, which is numerical when we merge the dataset, into a categorical variable with 3 levels. K-means clustering labels each data with one of K categories by computing the nearest mean.

Having a categorical variable for property values was desired since interpretation when the variable is numerical is not that straightforward. Property values are quite large and we often get meaningful insight in analysis with the interpretation of coefficients. But, when a variable for property values is categorical, interpretation gets much more straightforward, as a unit increase refers to moving from one category to another category. Also, the range of New York's property values is quite wide as we know from our domain knowledge, splitting the property values into equal proportions of 3 categories might not be able to uncover true relationships between property values and review score ratings. So, unsupervised learning fits better and K-means clustering was employed.

Review scores rating was also converted into a categorical variable, because we want to construct a model to estimate whether or not we have a satisfied customer, so binary outcomes are desired rather than a numerical outcome.

Feature selection and hypothesis testing

Feature selection and hypothesis testing need to be performed before we construct a model to determine what variables need to be included and without careful consideration, it is likely that a derived model does not have a desired accuracy and it can cause overfitting or underfitting.

We specifically explored variables that are amenities, since the review scores variables are obviously quite strongly correlated with overall review ratings. We decided to only include variables that appear at least certain amount of times in the amenities column, because amenities with low frequency are likely to be the

ones that are hard to be offered and as the probability of having the amenities is low, when included in a model, the small proportion might cause a bias in the model.

After sorting out amenities by frequency, chi-square test of independence was performed on amenities and overall review ratings. Chi-square test of independence tests whether two categorical variables are independent or not.

H_0: The two variables are independent

H_1: The two variables are not independent

We reject the null hypothesis if the p-value of the test-statistic is less than α, concluding that there is evidence to suggest the two variables are not independent and rejection of the null hypothesis justifies choosing the variables as predictors.

Models

Logistic regression

Preliminary steps required before constructing a model are complete and we need to decide what kind of regression model we need to apply that fits our data(variable) types. Logistic regression has advantages over linear regression models for our data because we converted the overall ratings variable into a categorical variable and the variable is a response variable in our model, and therefore logistic regression was used to fit our data to predict the probability of having a satisfied customer determined by getting the overall rating over 95.

Unlike linear regression, logistic regression does not make assumptions like the response and predictor variables need to be linear and error terms need to be normally distributed and homoscedasticity. So, normalization on variables will not be done to fit a logistic regression model.

One of the assumptions we need to be careful of is multicollinearity. In our analysis, multicollinearity is addressed by dropping highly correlated variables except overall review ratings. Another thing that needs to be checked after constructing a logistic regression model is whether the model is overfitting or underfitting.

Overfitting occurs when we try to fit too many variables as predictors and a model is then fit too closely and starts to explain random noise as well. When a model is

overfit, in logistic regression, the model's accuracy on data that the model was trained on will be relatively high, but the model's accuracy on new data that it was not trained on will fall sharply. Underfitting occurs when a model fails to capture the proper relationships between predictor and response variables, often caused by having too few predictor variables or appropriate predictor variables are not present in the model.

Overfitting and underfitting can be checked by cross validation and AUC(area under the curve) in logistic regression. Cross validation splits data into training and test data sets, feeding the training dataset into our model to train and test our model with the test dataset to test since our model has never been trained with the test dataset. K-fold cross validation was applied for cross validation. K-fold cross validation splits a dataset into K-fold, performing cross validation K times testing with one fold at a time, covering the entire dataset to test eventually. AUC computes the area under the ROC(Receiver operating characteristics) curve. The ROC curve is plotted with TPR(true positive rate) as y-axis and FPR(false positive rate) as x-axis. True positive rate is a rate at which the model predicts the positive class correctly and false positive rate is a rate at which the model predicts the positive class incorrectly. AUC helps us evaluate the accuracy of the logistic regression model and the mean and standard deviation of AUCs generated by K-fold cross validation checks overfitting such that there is a strong sign of overfitting when AUCs give inconsistent results, resulting in large standard deviation.

Several logistic regression models will be constructed in the milestones section. We will explain justifications as to why several models were required in the milestones. After going through several models, it was concluded that we need a model that extends from logistic regression due to low AUC.

Random Forest

RandomForest is one of the most famous and the most used mapping machine learning algorithms. Advantage of using RandomForest is it can be used for both regression and classification. Also, the RandomForest model can handle both categorical and continuous variables. The RandomForest is a collection of decision trees. Each tree of the RandomForest model will make a prediction and values will be averaged like Regression models. Then it will produce a max voted value like classification models.

Even Though the RandomForest model is useful, there are some disadvantages. Since the RandomForest builds numerous decision trees, it definitely needs powerful

computational power and time. When it comes to real-time prediction, RandomForest model might not be the best model to predict the outcome.

For the RandomForest model, there are no formal distributional assumptions. Since the RandomForest model is non-parametric, it can handle ordinal data, non-ordinal data and skewed data.

Neural network

Deep Neural network is a deep learning technique that is particularly good at unraveling underlying non-linear relationships between predictor and response variables. There are input, hidden and output layers. Input layer represents predictor variables and output layer represents response. Activation functions are applied between layers and with hidden layers in between input and output layers, we can add non-linearity to our model and having more hidden layers lead to more complex non-linear models.

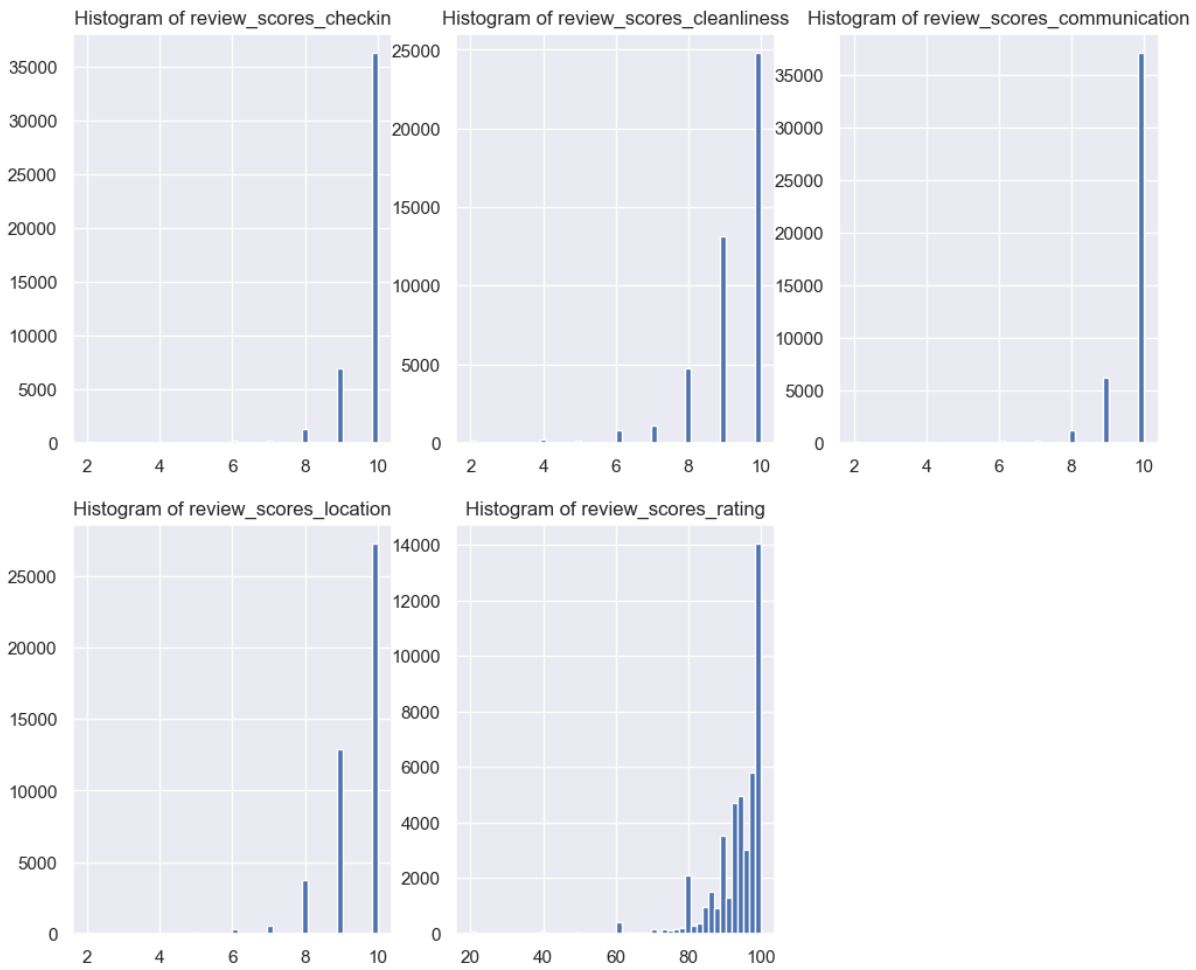
One of the disadvantages is interpretation of a generated model. Unlike logistic regression where beta coefficients of models are given from the model and effects on odds ratio per unit increase can be computed, it is hard to investigate the effect of each individual predictor variable. Overfitting also occurs in deep neural network and it often is hard to get a clear implication as to what needs to be done to fix overfitting when we observe overfitting is present in a model.

Overfitting again will be checked for a neural network model splitting dataset for cross validation and checking the loss curve. In neural network, loss function defines how distant the model and the final answer is and through iterations, a model is constructed to achieve minimal loss. So, overfitting is likely to be there when the loss curve does not seem to stabilize after certain iterations.

Milestones

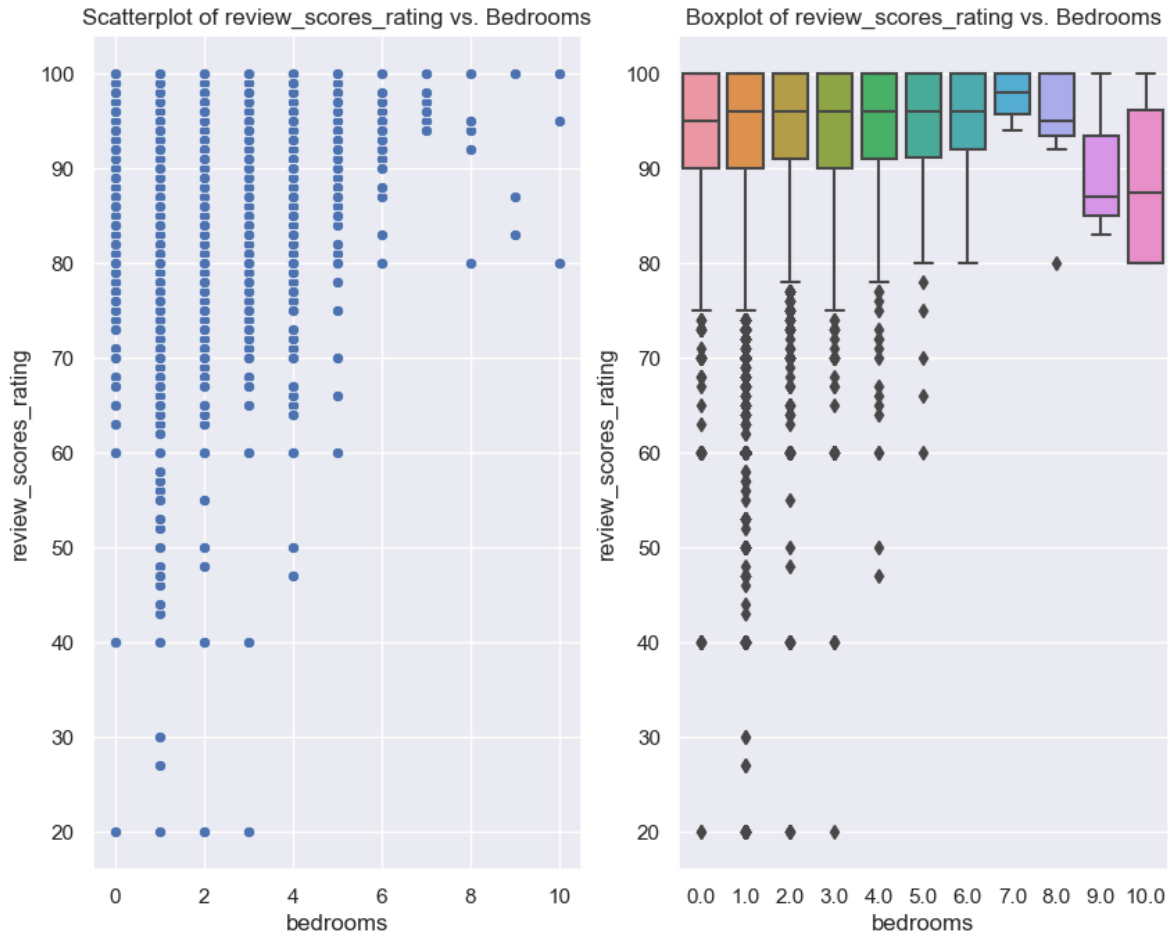
Exploratory data analysis

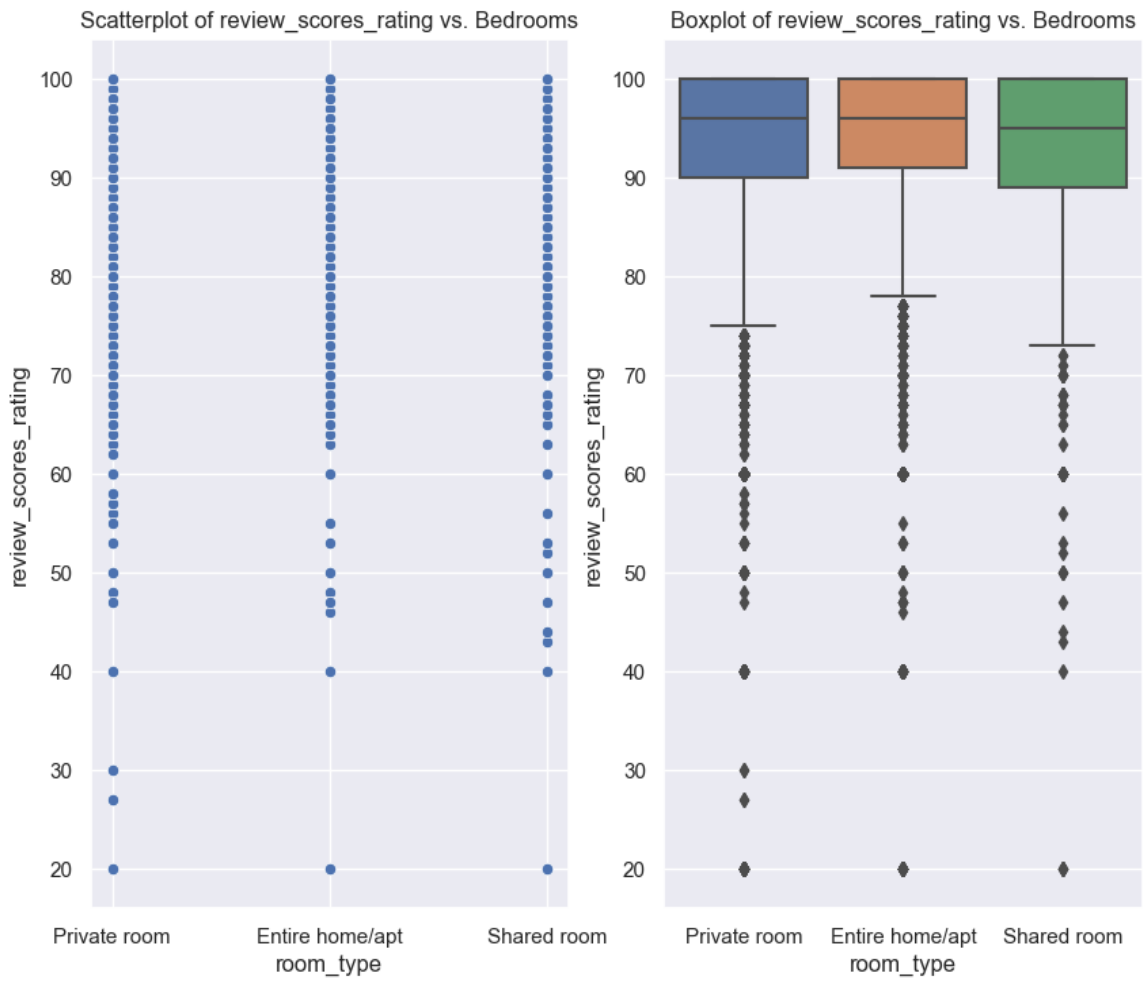
	review_scores_checkin	review_scores_cleanliness	review_scores_communication	review_scores_location	review_scores_rating	review_scores_value
count	45478.000000	45544.000000	45539.000000	45475.000000	45624.000000	45470.000000
mean	9.738555	9.289786	9.758624	9.463002	93.474750	9.384891
std	0.662886	1.066889	0.645440	0.818924	8.240908	0.874256
min	2.000000	2.000000	2.000000	2.000000	20.000000	2.000000
25%	10.000000	9.000000	10.000000	9.000000	90.000000	9.000000
50%	10.000000	10.000000	10.000000	10.000000	96.000000	10.000000
75%	10.000000	10.000000	10.000000	10.000000	100.000000	10.000000
max	10.000000	10.000000	10.000000	10.000000	100.000000	10.000000



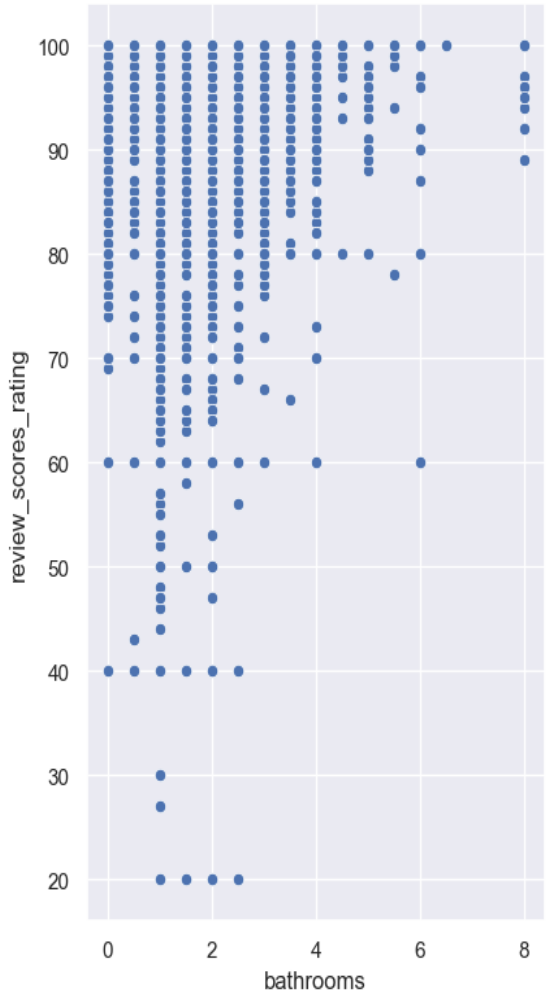
Above are the summary statistics and histogram of review scores of subcategories and overall review scores rating. Note that the overall review scores rating is out of 100 and other review scores are out of 10. All of the ratings tend to be very high as the 50th percentile of review scores of subcategories is 10 and the 50th percentile

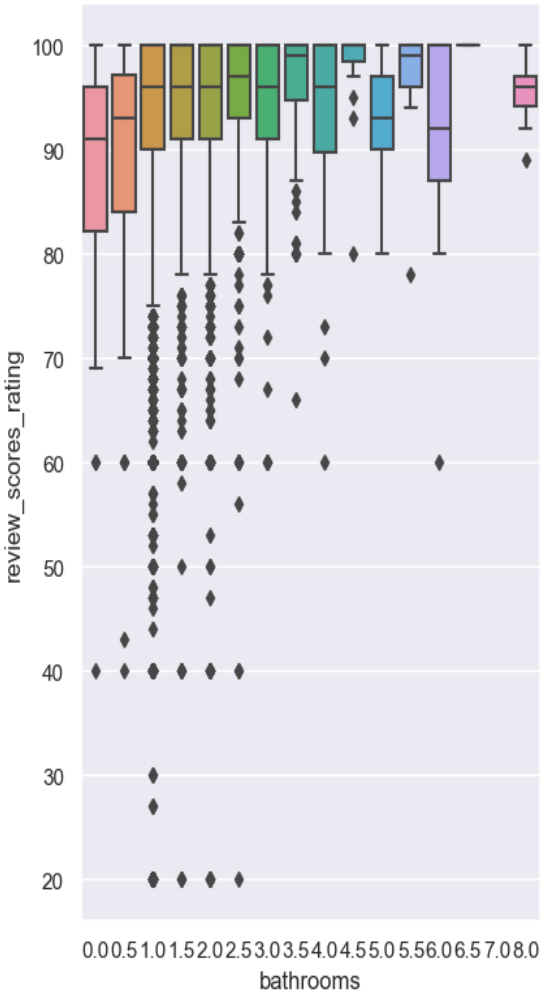
of overall review scores is 96. The mean of cleanliness and value tend to be lower than the other ratings and standard deviation tend to be higher than the other, which can be interpreted as customers on average cared more about cleanliness and value resulting in relatively lower scores and review scores more dispersed.

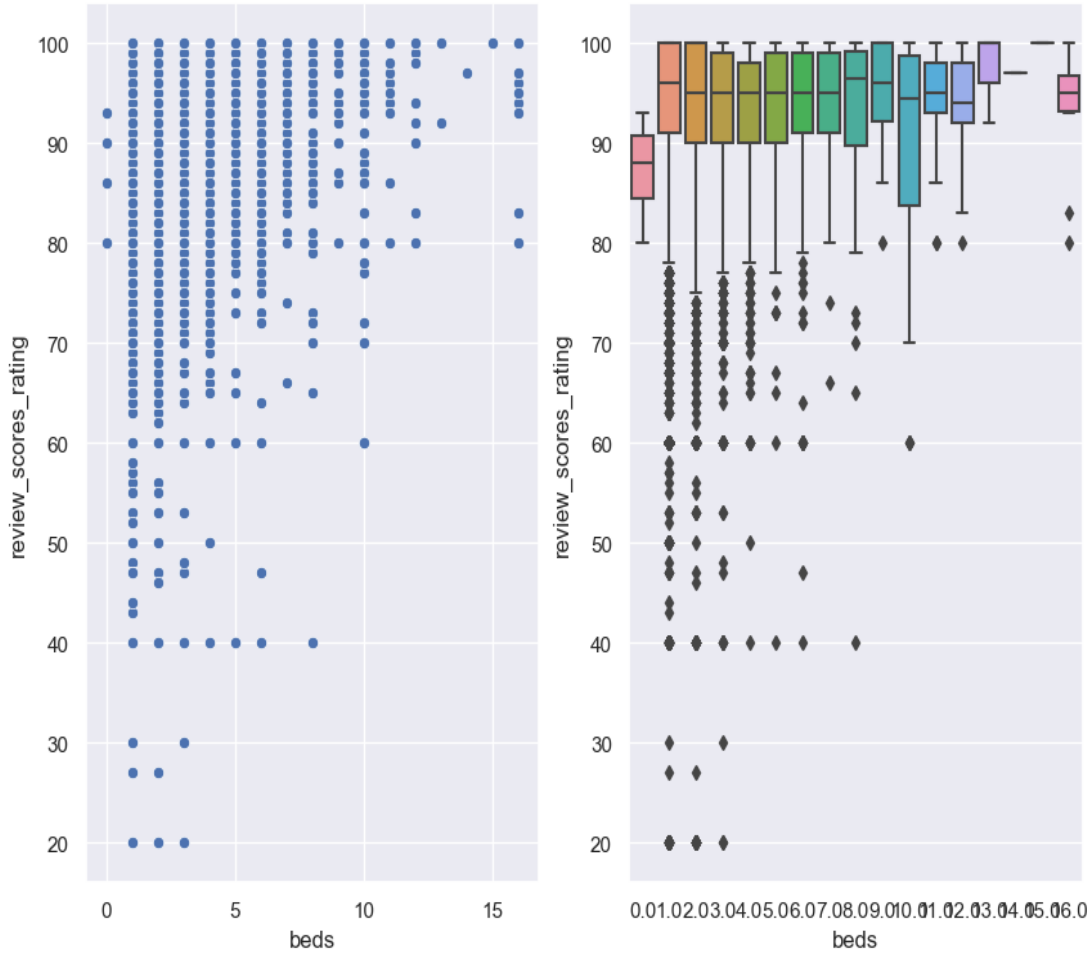




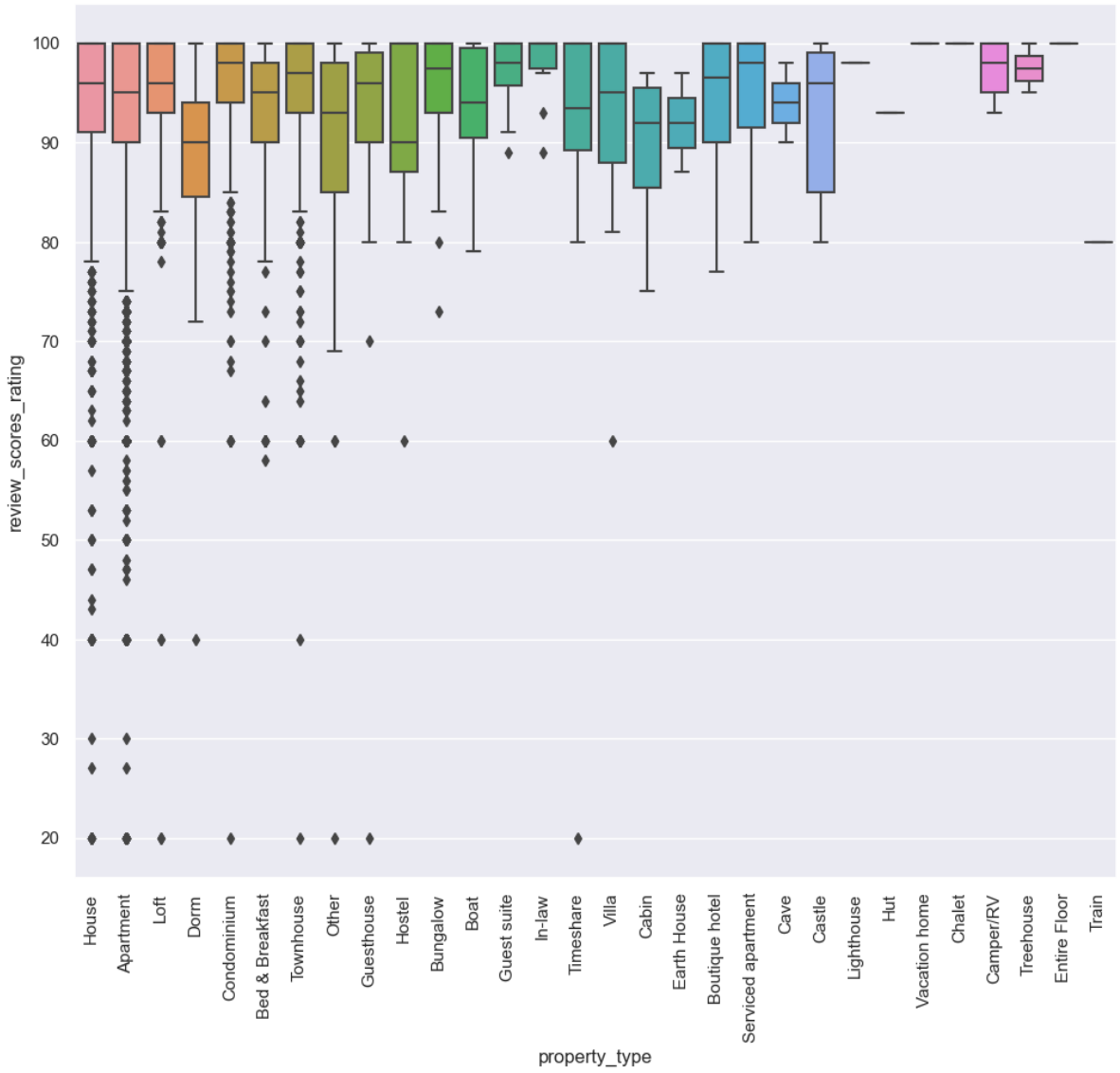
The two scatterplots and boxplots respectively represent review scores rating according to number of bedrooms and type of bedrooms. There is no discernible difference between the number of bedrooms or type of bedrooms.


review_scores_rating
100
90
80
70
60
50
40
30
20
0
2
4
6
8
bathrooms

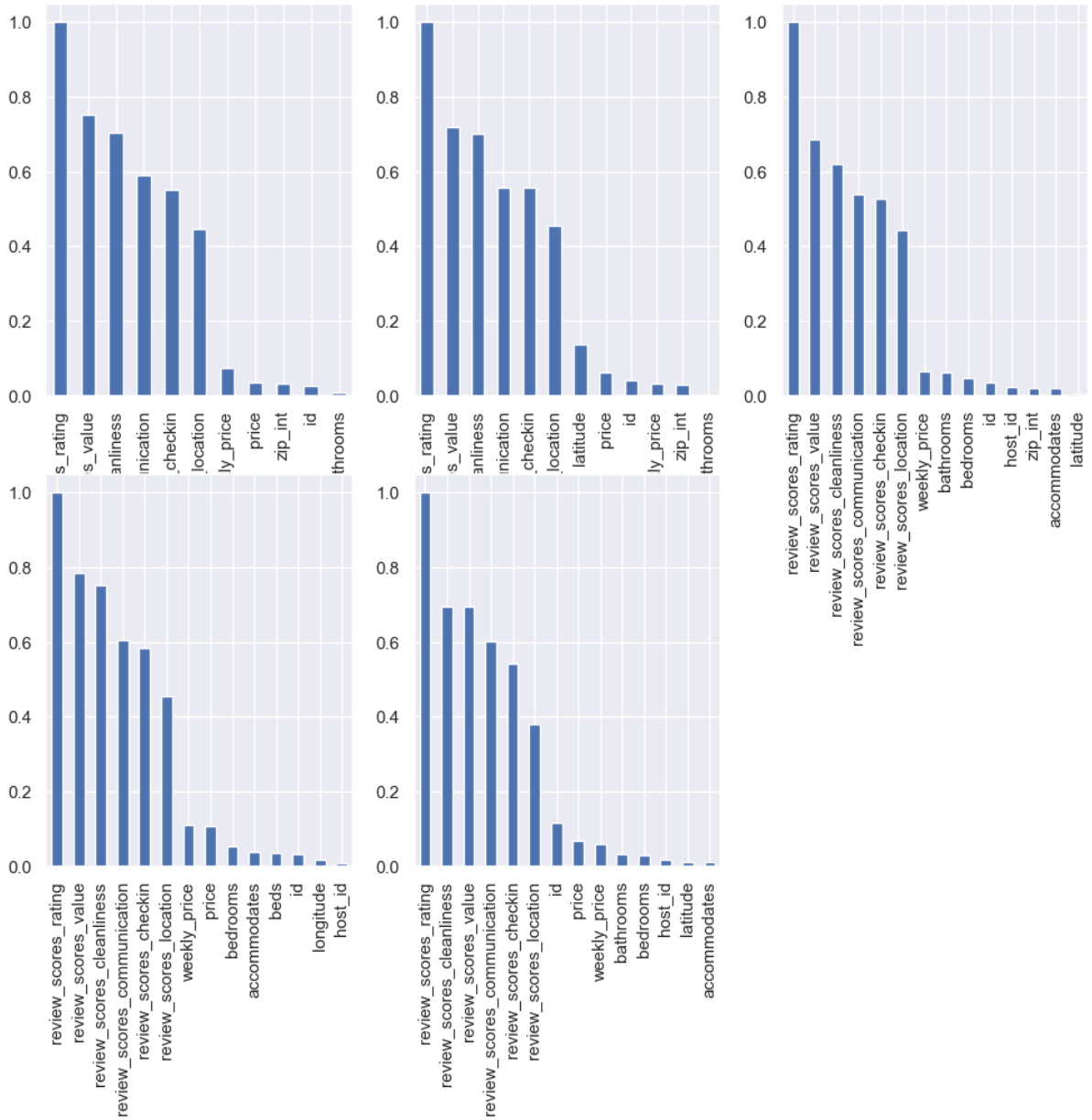

review_scores_rating
100
90
80
70
60
50
40
30
20
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 8.0
bathrooms



The two scatter plots and boxplots above represent overall review rating against number of bathrooms and number of beds. It should be noted that when the number of bathrooms are either 0 or 0.5, the overall rating tends to be noticeably smaller than those with more than one bathroom. Also, when the number of beds is 0, the overall rating tends to be much smaller that the mean seems to be less than 90. There are no noticeable differences between the number of bathrooms or the number of beds when they were larger than or equal to 1. They need to be addressed when we construct prediction models.

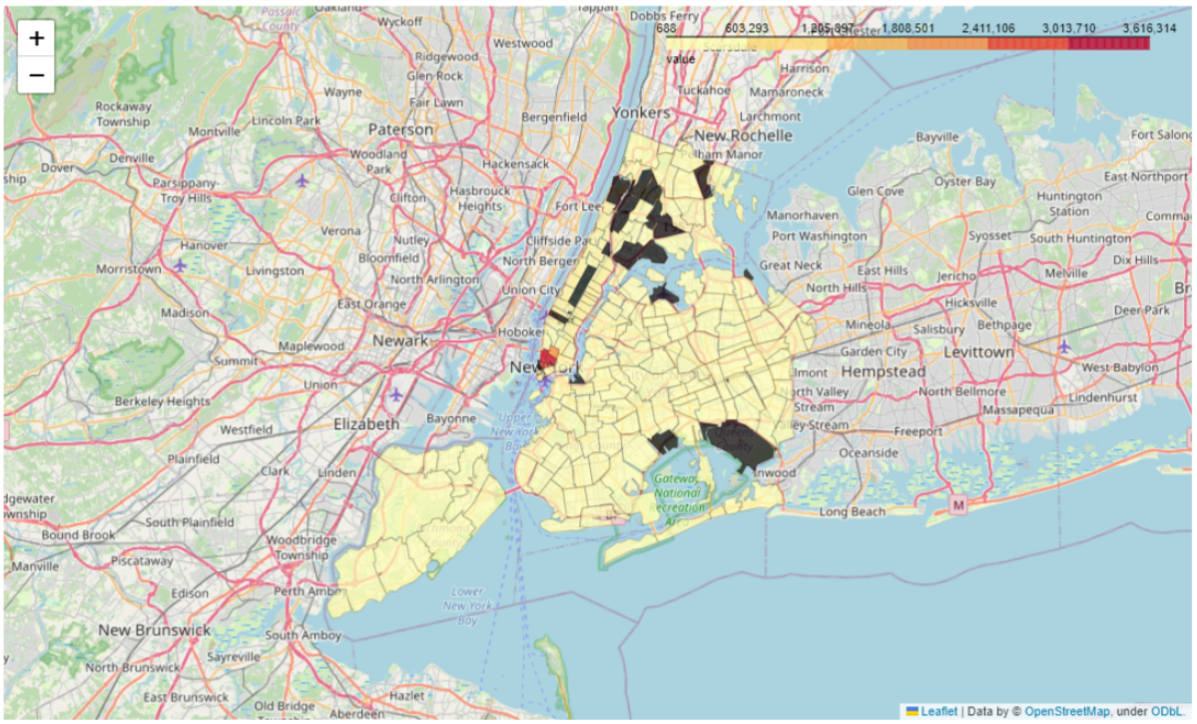


The boxplot above represents the overall review score against different property types. Dorm tends to receive lower ratings and Hostel tends to receive lower ratings and ratings' variance seem to be large.



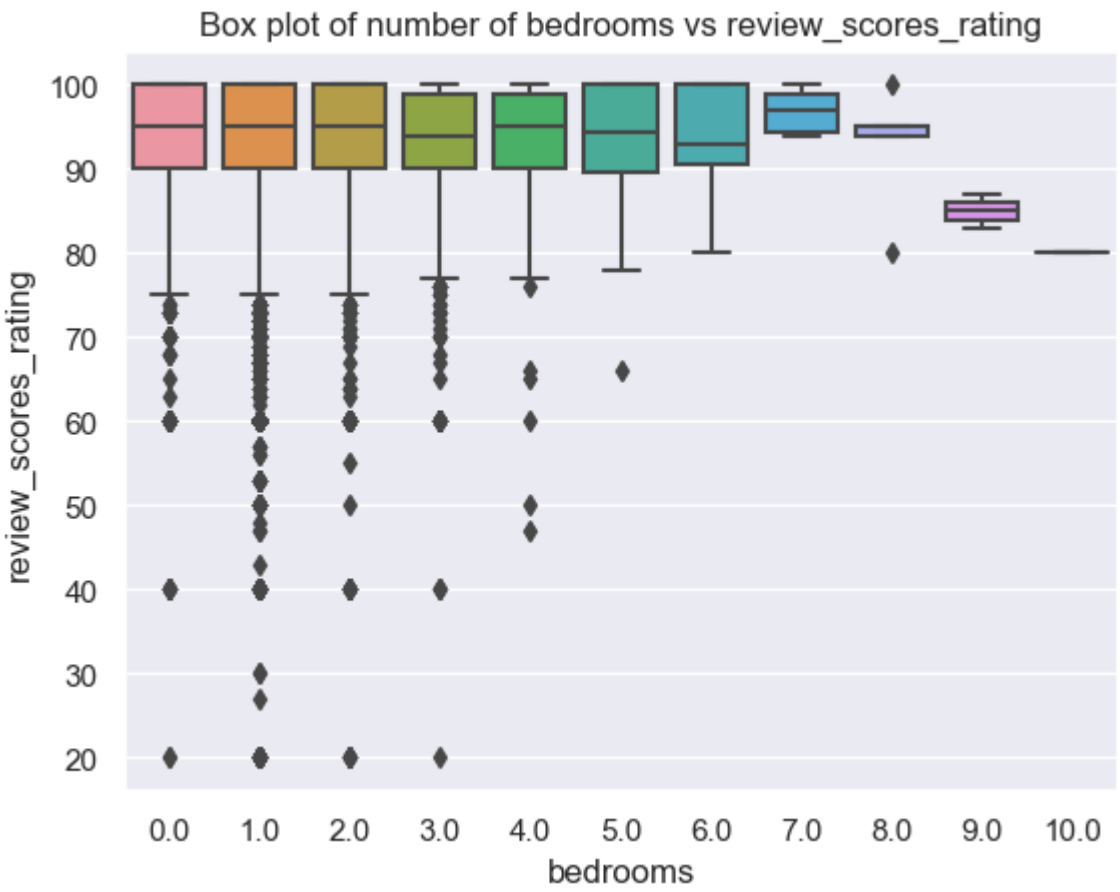
The 5 bar charts above show the correlation between the overall rating and other numerical variables splitting data into 5 different states. As we expected, other sub reviews' correlation coefficients are larger than the other variables. Review scores of value is the most correlated variable in all of the 5 states. Cleanliness was the second most correlated and as we have observed in the summary statistics, their averages were relatively low and variations were relatively large, implying they need to be our focus when constructing models.

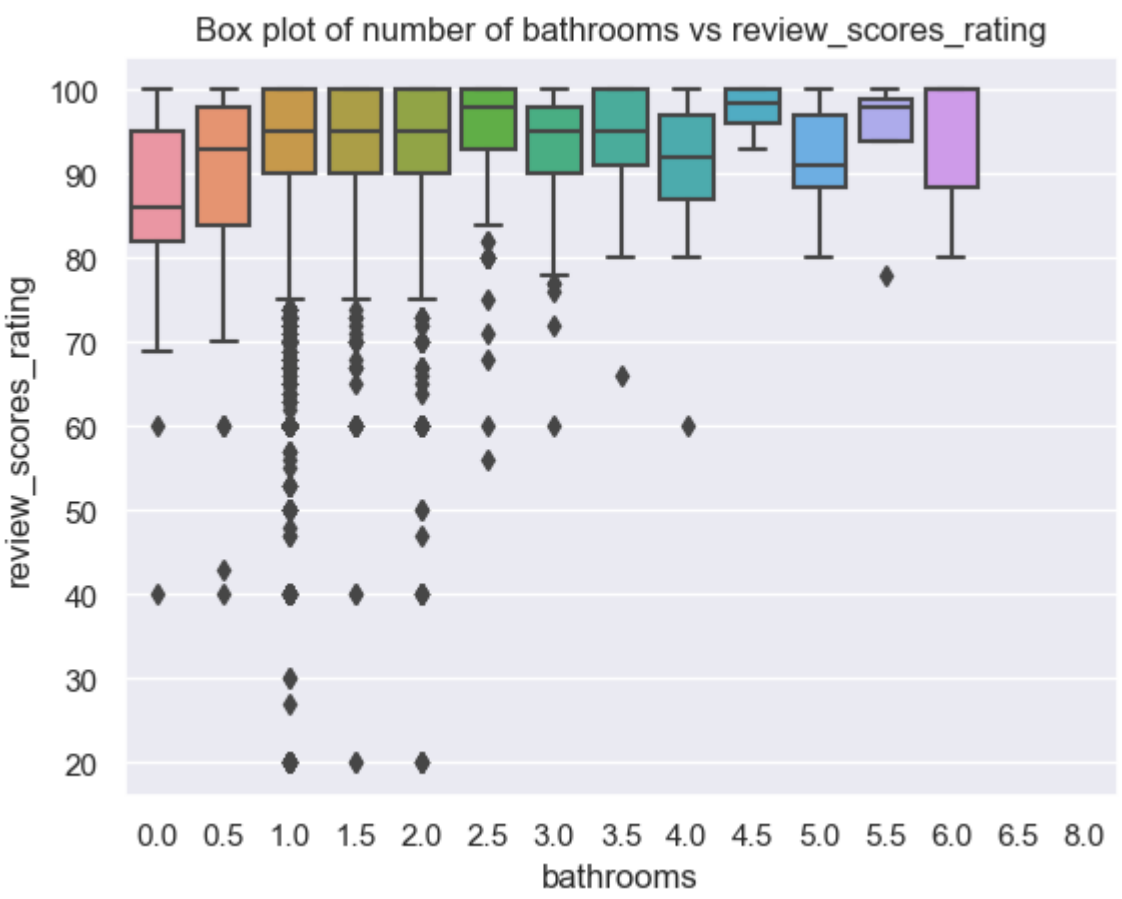
From our domain knowledge, we know that property values vary the most in New York with the most expensive properties in Manhattan and New York's data accounts for approximately 75% of the data. We will focus our analysis on New York to investigate the interactions between listings-specific factors and property values on overall review ratings. Plots for New York are presented below.

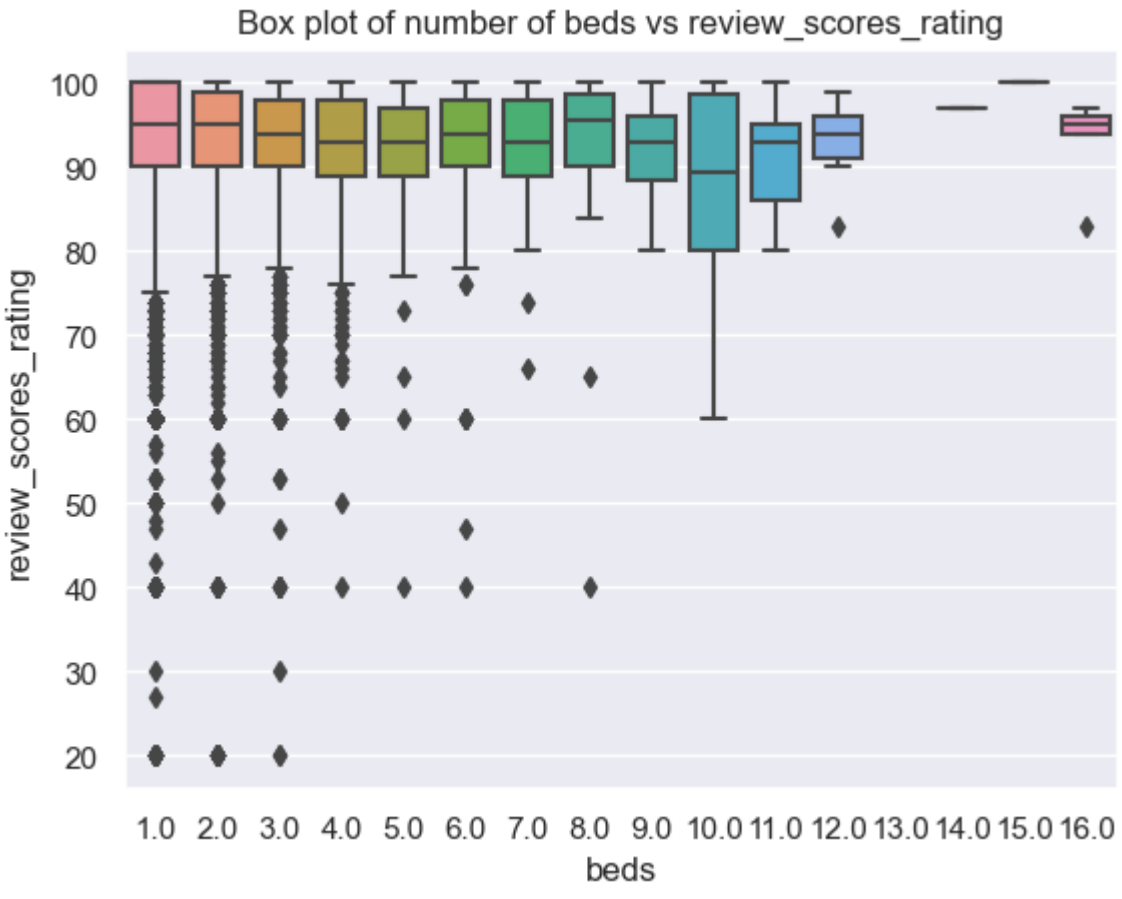


Above is Heat map of New York with respect to Zillow indexes by all the zip codes in the city of New York. The top 5 zip codes that have highest ZHVI values are 11976, 10013, 10007, 11930, 11932 as shown below.

	index	state	type	zipcode	value
372	18454	NY	ZHVI	11976	3.616314e+06
5	18087	NY	ZHVI	10013	3.025386e+06
3	18085	NY	ZHVI	10007	2.691314e+06
344	18426	NY	ZHVI	11930	2.653136e+06
345	18427	NY	ZHVI	11932	2.340550e+06

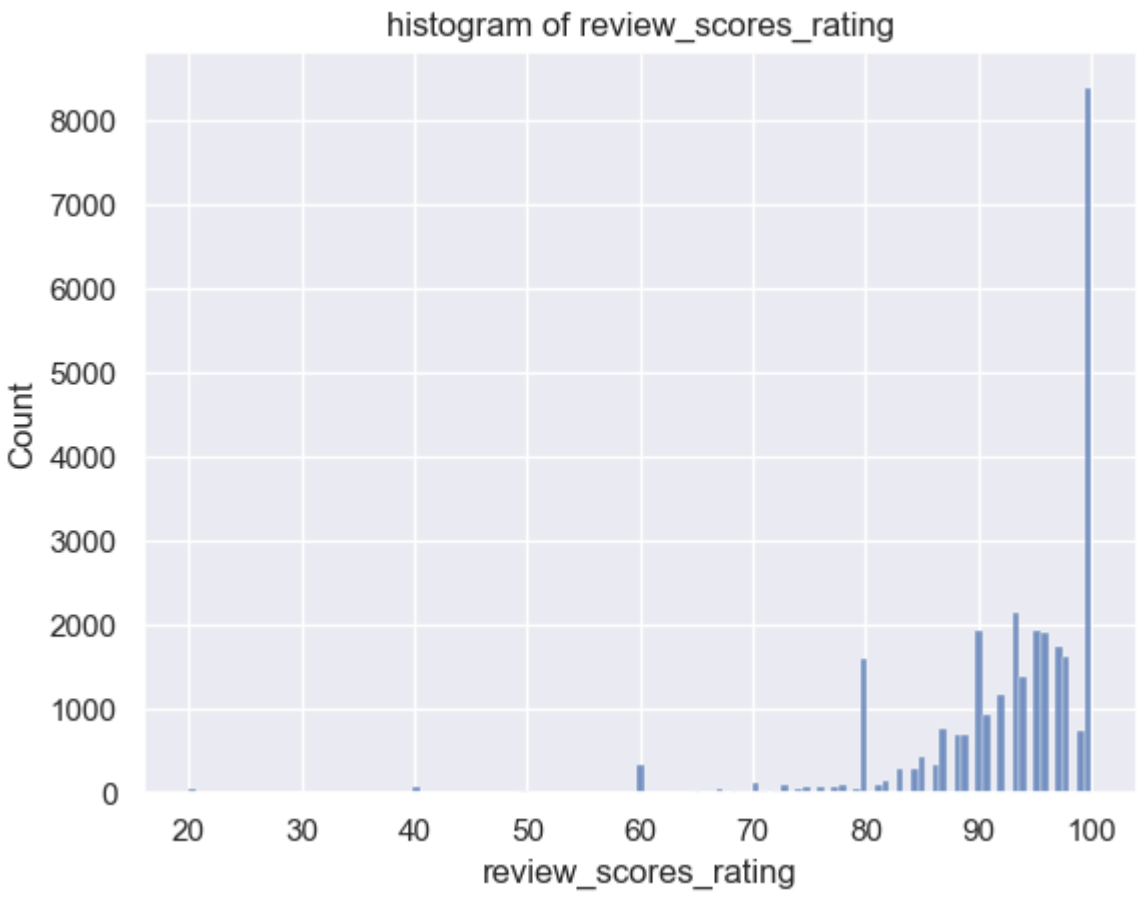

Box plot of number of bedrooms vs review_scores_rating
review_scores_rating
100
90
80
70
60
50
40
30
20
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
bedrooms

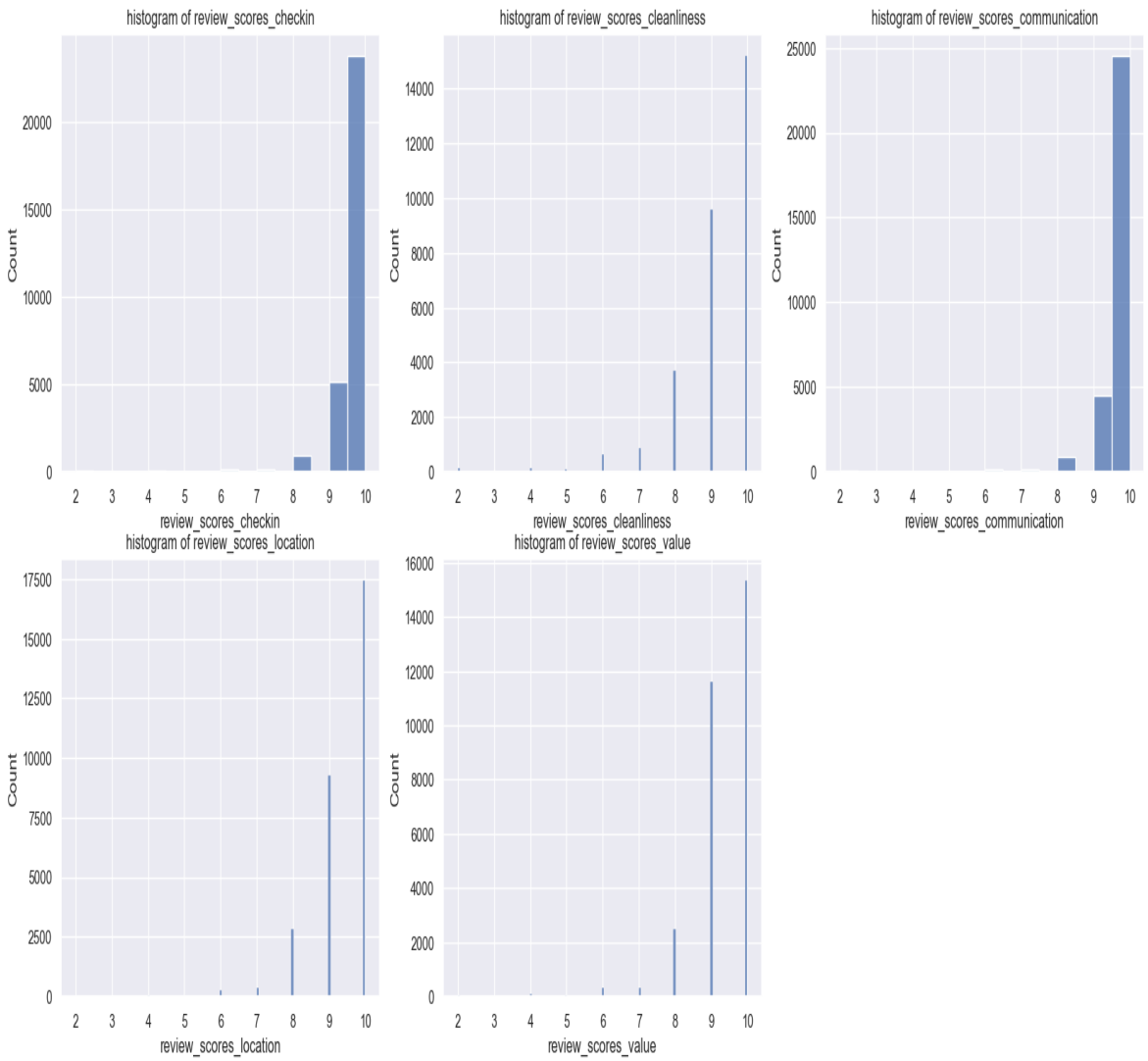

Box plot of number of bathrooms vs review_scores_rating
review_scores_rating
100
90
80
70
60
50
40
30
20
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 8.0
bathrooms


Box plot of number of beds vs review_scores_rating
review_scores_rating
100
90
80
70
60
50
40
30
20
1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0
beds

The boxplots for New York data seem to be very similar to the boxplots of the entire dataset except there are no listings with 0 beds in New York, implying the number of beds do not have significant effect on overall review ratings.

The scatter plot above plots the price of listings and overall review ratings they got. There seems to be no specific relationships to note.


histogram of review_scores_rating
Count
8000
7000
6000
5000
4000
3000
2000
1000
0
20
30
40
50
60
70
80
90
100
review_scores_rating



The plots above plot the histogram of 6 review scores for the listings in New York. There is no notable difference compared to the histogram of 6 review scores we plotted for the entire dataset.

Data cleaning and wranging

As was mentioned in the methodology section, the listings dataset was merged with the calendar dataset to extract the months present in listings dataset because the Zillow property dataset contains monthly property value for each zip code in the United States. So, we also merged the listings dataset with the property value dataset according to corresponding zip codes(zip codes in New York). And, we created a column of mean property value in the listings dataset by calculating the mean of zillow property indexes(ZHVI,ZRI) in the corresponding time frame for each zip code in New York.

accommodates	0.000000
amenities	0.000000
availability_30	0.000000
bathrooms	0.003847
bed_type	0.000000
bedrooms	0.001599
beds	0.001282
cancellation_policy	0.000000
host_id	0.000000
id	0.000000
instant_bookable	0.000000
latitude	0.000000
longitude	0.000000
metropolitan	0.000000
name	0.000799
price	0.000000
property_type	0.000000
review_scores_checkin	0.245937
review_scores_cleanliness	0.244588
review_scores_communication	0.244588
review_scores_location	0.245920
review_scores_rating	0.243039
review_scores_value	0.245970
room_type	0.000000
weekly_price	0.773113
zipcode_x	0.000000
zipcode_modified	0.000000
mean_real_estate	0.000000
reale_group	0.000000
property_val_cat	0.000000

The list above shows the proportion of null values for each variable in the dataset. The proportion of null values in weekly_price is very high with 0.77 and we are dropping the variable in our variable because the proportion of null values is very high and there is price in the dataset, making weekly_price rather redundant.

We can also see that 6 review_scores variables have the proportion of approximately 25 percent for null values, for which we are putting emphasis on our analysis, so just dropping null values can cause serious bias in our analysis as the listings with null values might contain valuable information. As we have seen in the summary statistics and distributions, the review scores' standard deviation seemed to be very small and concentrated, so we are filling null values with the median values, which are quite high, larger than 90. From our domain knowledge, the listings with null values for review scores can be interpreted as customers skipping reviews and customers that are satisfied are more likely to skip reviews than those that were dissatisfied. So, filling median values in is justified.

The proportions of null values in the remaining variables are very small, largest being 0.3 percent in bathrooms, we are simply dropping null values for the remaining variables.

Feature engineering

K-Cluster Means Method

Now in order to run the K-cluster means method, we set the property group as three groups: Low, Mid and High.

By this methodology we obtain mean value with classes property value group which is:
Low property value means are $10881, mid as $644,218 and High as $1,909,449.
We also want to check how many data points are in each group and we have total 38852 data points in the low group, 19963 in mid, and 1241 in high.

```
mid 644224.6873345957
low 10897.322192237678
high 1908841.7849218908

number of data for high: 1241
number of data for mid: 19963
number of data for low: 38852
```

After K-Cluster Means was run, we converted property values into an ordinal variable, property_val_cat, where 0 refers to a listing with low property value, 1 refers to a listing with mid property value, and 2 refers to a listing with high property value.

Logistic Regression Model(Including feature selection)

We now categorize all the amenities in the dataset by tokenizing them in order to run the logistic regression model on amenities and review scores.

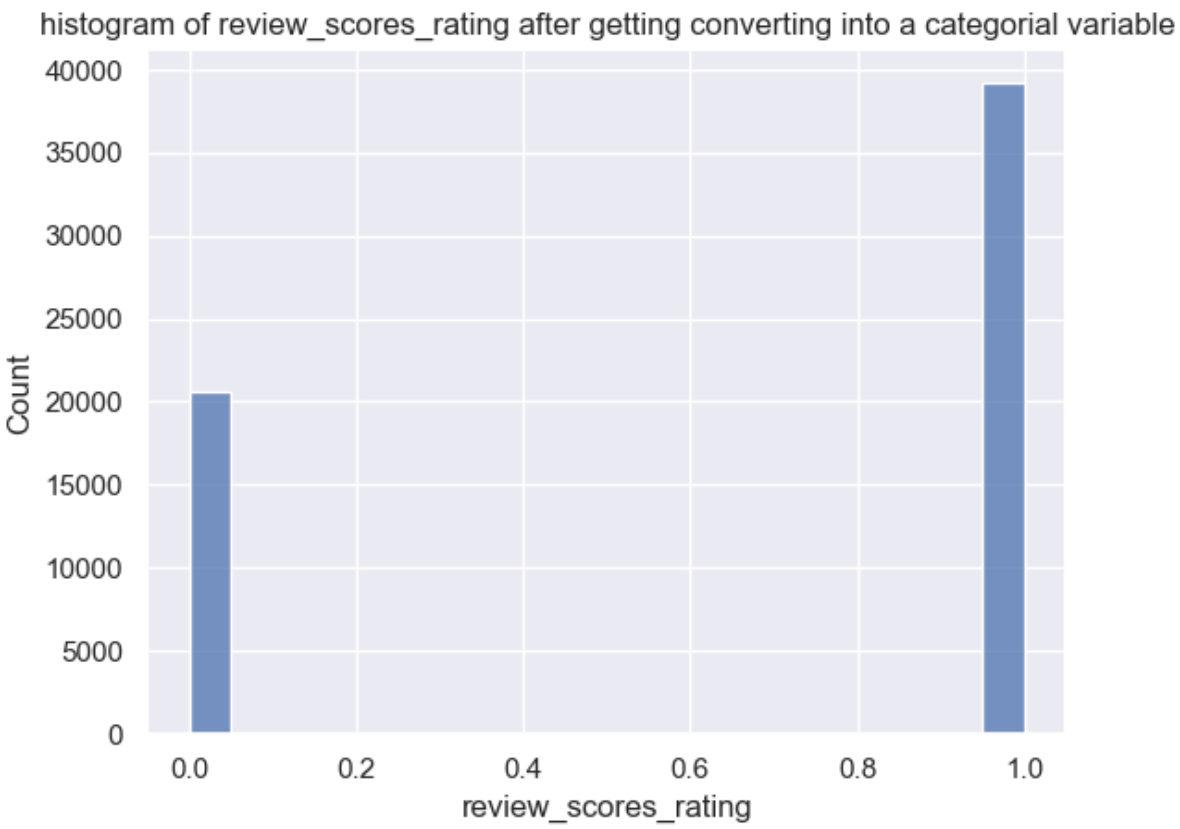
		24-hour check-in	accessible-height bed	air conditioning	baby bath	baby monitor	babysitter recommendations	bathtub	bbq grill	beach essentials	...	tv	washer	washer / dryer	wheelchair accessible	wide clearance to bed
0	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0
1	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	...	1	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	...	1	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	...	1	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
60047	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
60048	0	0	0	1	0	0	0	1	0	0	...	1	0	0	0	0
60049	0	0	0	1	0	0	0	1	0	0	...	1	0	0	0	0
60050	0	0	0	1	0	0	0	0	0	0	...	1	0	0	0	0
60051	0	0	0	1	0	0	0	0	0	0	...	1	0	0	0	0

60052 rows × 96 columns

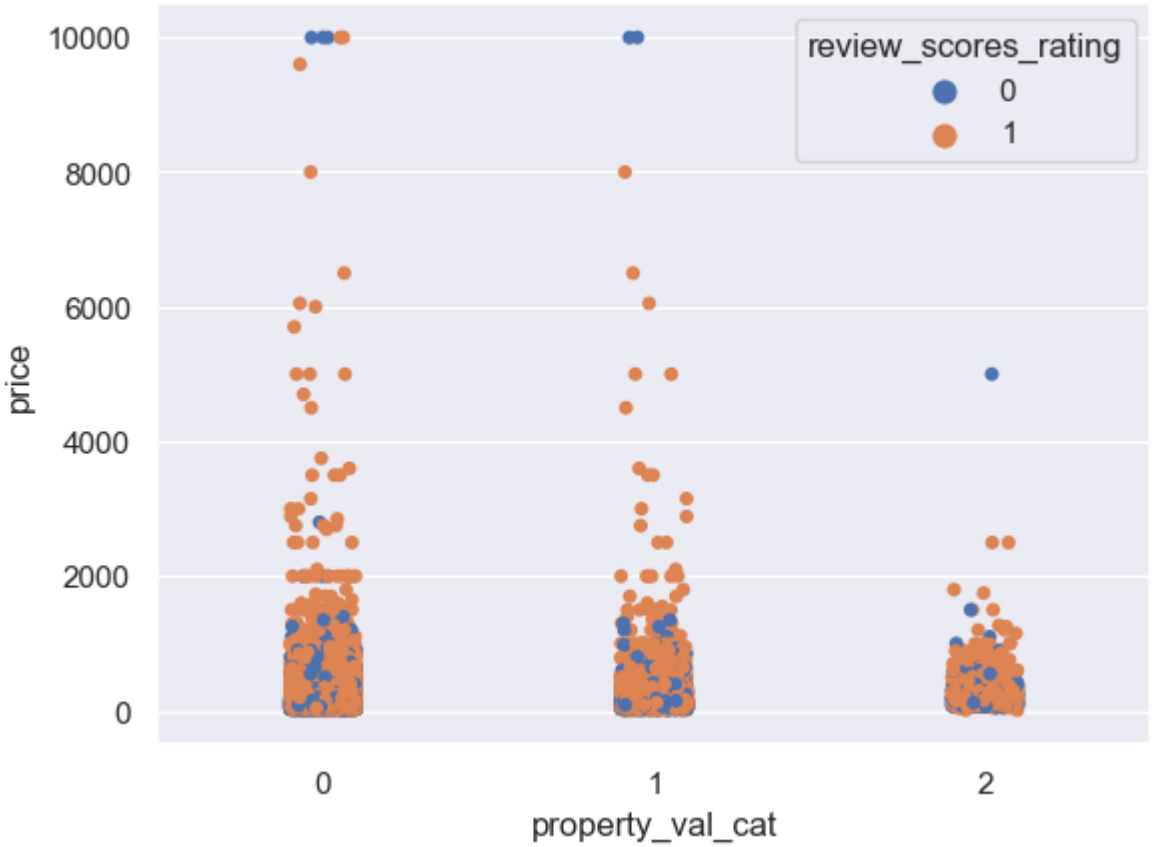
table 2: categorized table by amenities.

Now we find which amenities are frequent and infrequent. We set infrequent amenities by appearing less than 50 times in the dataset and frequent amenities by appearing more than 2000 times. For example, frequent amenities are gym, hot tub, and tv,etc. whereas infrequent amenities are microwave, dishwasher, beach essentials. The total number for frequent amenities are 41 in total.

We now combine these categorical tables with property value with all review scores and also categorize by setting the good review which is overall review score over 95 are 1, else 0. We set the threshold to be 95 because as we have seen in the summary statistics, the mean and the median of review score ratings were very high and the standard deviation was very small. So, we want to set tough standards as to whether a customer was satisfied or not and 95 looked suitable to be set as the threshold, splitting the data into balanced proportions.



As was mentioned in the methodology section, we are only taking amenities and five review scores into account in our model and one might wonder why we are not including price and bathrooms, as price is always an important factor for satisfaction and we were able to see in the box plot, there was some difference in overall review scores between listings with 0 or 0.5 bathrooms and listings with equal to or more than 1 bathroom.



After converting property values and overall review ratings into categorical variables, we were able to plot stripplot of property value categories and price on review score ratings. There seems to be no distinctive difference between property value categories from the plot above and we concluded that there are no significant interactions between price and property value categories on getting overall review ratings over 95. Also, including price in a model will be redundant when we already decided to include review scores value in our model and from our domain knowledge, fierce battles for price are already undergoing between Airbnb and hotels in New York, making it less justifiable to account for in our model since we hope to find alternative actions that can be performed to compete against hotels. Also, we did not include the number of bathrooms either because the proportion of listings with less than one bathroom only accounts for 0.4% of the entire listings.

We now then check the correlation term to see if getting an overall review score over 95 is correlated to each review score and frequent amenities. Below are the positive correlation coefficients for review scores and amenities.

```
review_scores_rating           1.000000
review_scores_value            0.575797
review_scores_cleanliness      0.553090
review_scores_location         0.426918
review_scores_communication    0.379511
review_scores_checkin          0.377296
gym                            0.024253
laptop_friendly_workspace      0.017934
private_entrance               0.017659
washer                         0.017248
dryer                          0.016023
kitchen                        0.013181
elevator_in_building           0.012786
mean_real_estate               0.012560
breakfast                      0.012546
tv                             0.011042
hangers                        0.010134
shampoo                        0.010077
doorman                        0.009565
pets_allowed                   0.009288
property_val_cat               0.009187
hair_dryer                     0.008945
essentials                     0.008261
lock_on_bedroom_door           0.006831
iron                           0.004279
indoor_fireplace               0.004065
wheelchair_accessible          0.004050
smoke_detector                 0.003582
Name: review_scores_rating, dtype: float64
```

We can see that those five section review scores are highly correlated to the review scores of rating and the rest of them have very low coefficient terms. So we can say that those review scores for value, cleanliness are dependent on review score of getting a good review score.

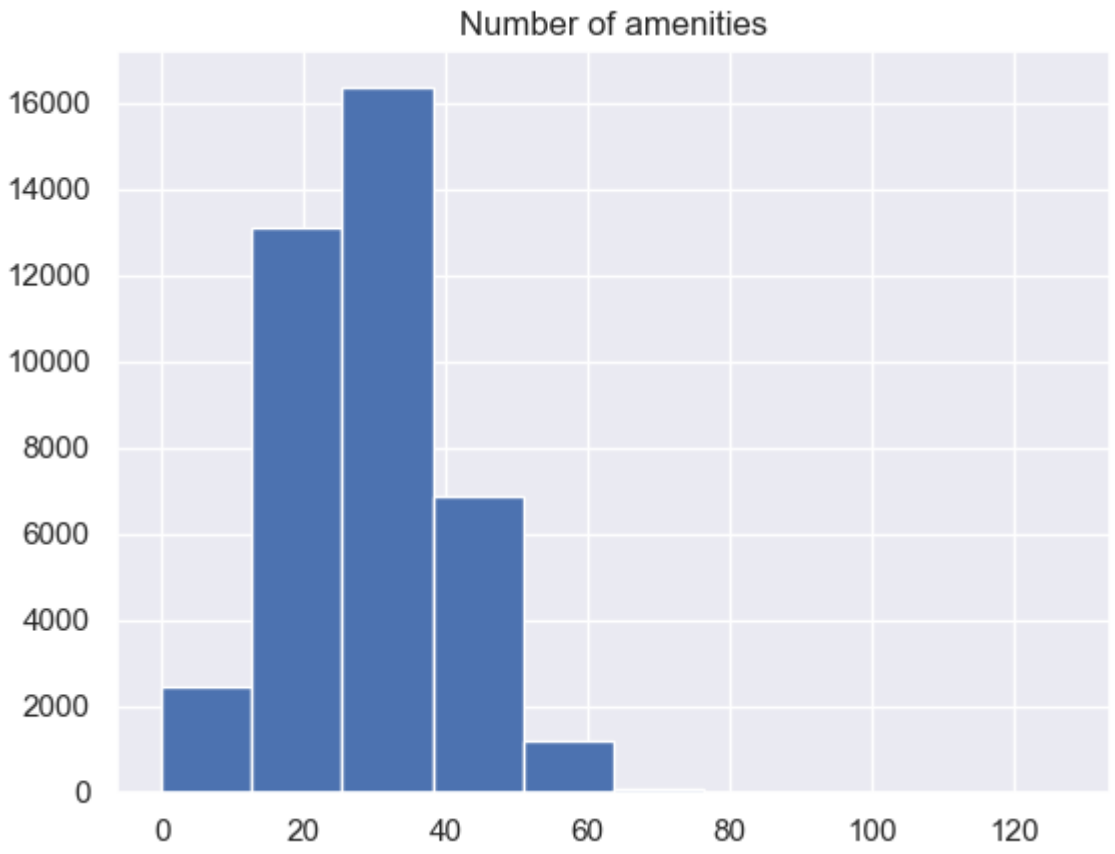
```
internet                     -0.020425
twentyfour_hour_check_in     -0.019242
heating                      -0.011453
free_parking_on_premises     -0.011328
family_kid_friendly          -0.009693
hot_tub                      -0.009068
carbon_monoxide_detector     -0.008612
fire_extinguisher            -0.008544
cat(s)                       -0.007694
cable_tv                     -0.007616
safety_card                  -0.007598
air_conditioning             -0.007534
first_aid_kit                -0.005478
smoking_allowed              -0.004897
pets_live_on_this_property   -0.003975
wireless_internet            -0.003525
buzzer_wireless_intercom     -0.002778
self_check_in                -0.001747
bathrooms                    -0.000548
dog(s)                       -0.000324
suitable_for_events          -0.000282
lockbox                      -0.000186
```

For negative correlation terms, there are internet and 24 hour check in service which is counter-intuitive. These features should have positive correlation which means whenever they are present, the overall rating score must increase.

Chi-square test of independence was performed to see whether amenities and overall ratings are independent. As was explained in the methodology section, not being able to reject the null hypothesis will result in two variables are independent and it is not recommended to predict one from another variable with a regression model. We ran a for loop on 41 amenities and set confidence level 0.1, so amenities with p-value larger than 0.1 were printed out after the for loop.

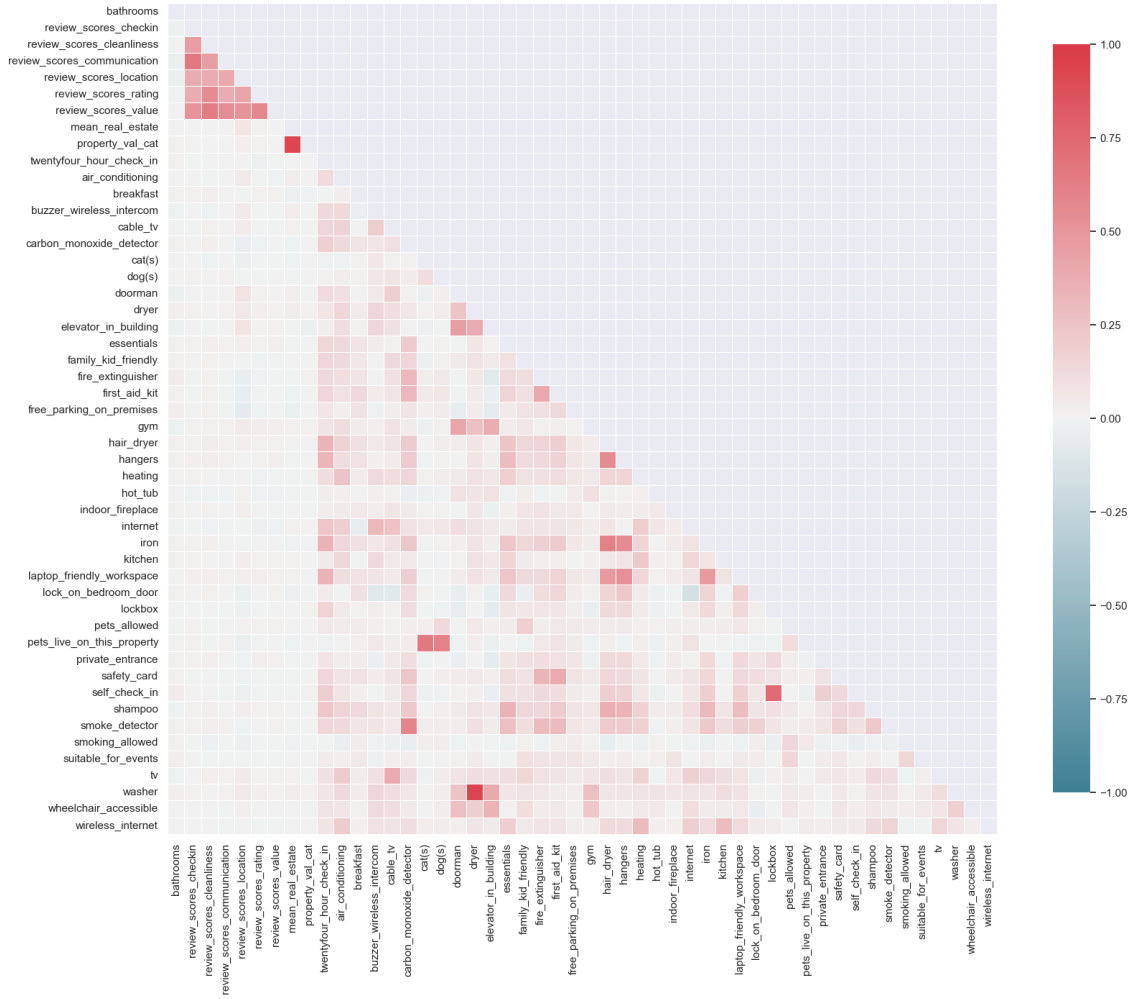
```
['first_aid_kit',
 'smoking_allowed',
 'iron',
 'indoor_fireplace',
 'wheelchair_accessible',
 'pets_live_on_this_property',
 'smoke_detector',
 'wireless_internet',
 'buzzer_wireless_intercom',
 'self_check_in',
 'suitable_for_events']
```

The list above are the amenities that are considered to be independent of overall review ratings based on the confidence level and the amenities that are not on the list are considered to be dependent. We can see that most of the amenities and overall review ratings are dependent, offering reasonable grounds to be included in a model. And, we also decided to include the amenities that were in the list above because of how the average number of amenities are distributed.

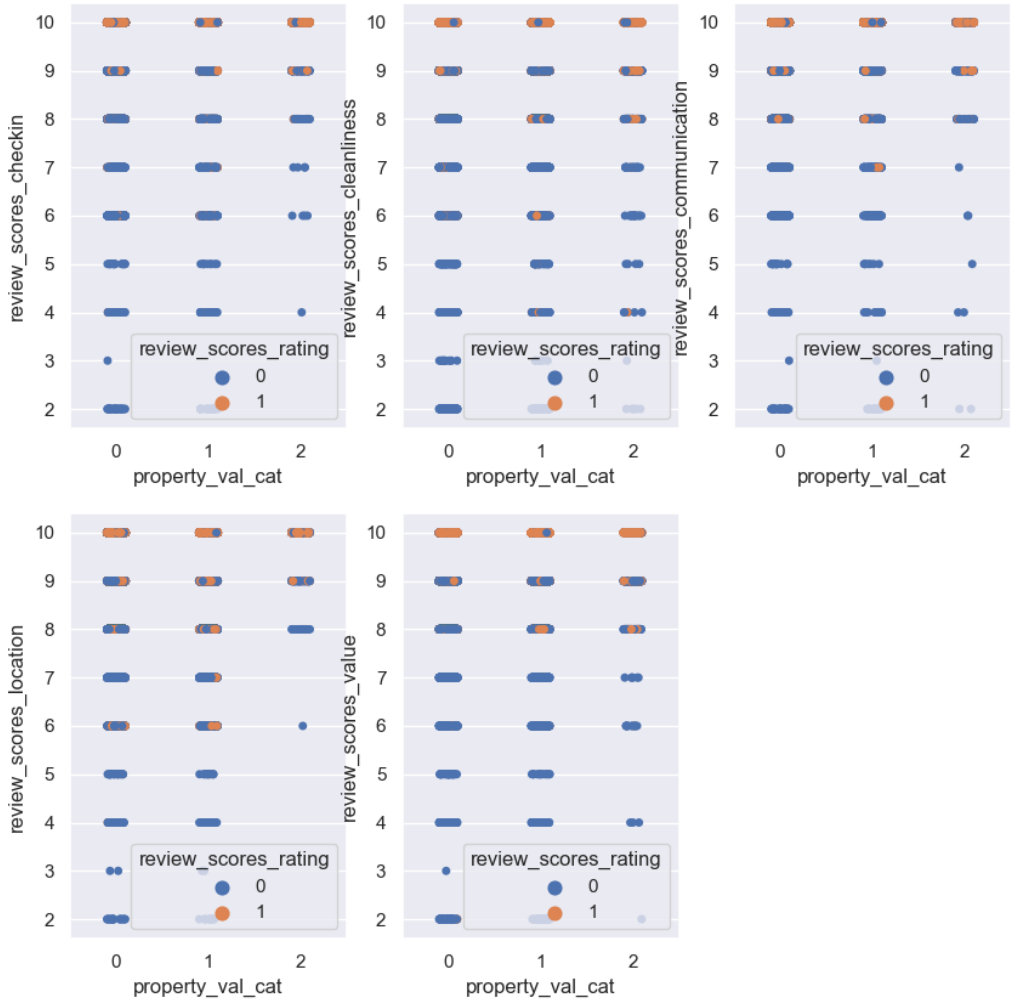


The histogram above plots the number of amenities a listing has. Most of the listings have amenities from approximately 20 to 40 as observed from the histogram. Dropping the list of amenities that are deemed independent of overall review ratings will result in approximately 30 amenities variables present in a model and in our opinion, the model might not be as practical as we desired it to be since the interpretation for listings with more than 30 amenities might not be clear not being able to take at least more than 10 amenities into account. And, complex models like neural network might be able to capture underlying relationships, so we decided to still keep them.

We now want to remove highly correlated variables that might affect our logistic model. We do so by using the heatmap for correlation coefficient matrix. Here, we want to keep the review score term so we want to find the dependent variables to avoid multicollinearity. We see that the dryer, lockbox are clearly in red. Dogs and cats are somewhat red so we also want to remove those dependent variables so we drop those variables. By doing so, we obtain significant amenities. With these feature selection of removing dependent variables, we want to include those significant amenities and review scores in our logistic model.



Here, we also ran a stripplot of review scores of five sections by the property groups to see if there are some patterns in each group of property values. We can all see that when the review score of all sections is 10, the chance that the customers give an overall review score over 95 is very high for each property group. It is also distinct that getting 8 or 9 in review scores leads to desired overall review scores. The stripplot of review score of location do show significance that low and mid property group is likely to have lower review scores less than 8.



If we run the chi-squared test for property group and overall review score rating, we get a very low p value which means there is extremely little difference between observed value and expected value.

```
2.8978895548344445e-07
```

We now first run the logistic regression model with interaction terms; interactions between reviews and property values. We want to set the interaction between review scores between high and low property value for intuitive

comparison. We also included the five most correlated amenities(internet, gym, 24 hour check-in, laptop-friendly, and private entrance) to overall review scores. property_val_cat_1 was dropped in the formula of logistic regression since we first want to see interactions between review scores and property values and we expected including more amenities will not necessarily lead to a more accurate model.

We are setting the mid property values category to be a default category, so terms related to property_val_cat_0 and property_val_cat_2 can be interpreted as comparison between the mid property value category and the category the term refers to.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-49.9691	0.965	-51.808	0.000	-51.859	-48.079
review_scores_checkin	0.5962	0.063	9.491	0.000	0.473	0.719
property_val_cat_0:review_scores_checkin	0.0661	0.078	0.845	0.398	-0.087	0.219
property_val_cat_2:review_scores_checkin	0.8658	0.310	2.791	0.005	0.258	1.474
review_scores_cleanliness	1.3142	0.038	34.720	0.000	1.240	1.388
property_val_cat_0:review_scores_cleanliness	-0.0028	0.047	-0.061	0.951	-0.094	0.088
property_val_cat_2:review_scores_cleanliness	0.2554	0.164	1.554	0.120	-0.067	0.578
review_scores_communication	0.9166	0.076	12.014	0.000	0.767	1.066
property_val_cat_0:review_scores_communication	-0.0077	0.094	-0.082	0.935	-0.192	0.177
property_val_cat_2:review_scores_communication	-0.1101	0.304	-0.362	0.717	-0.706	0.485
review_scores_location	0.7122	0.038	18.544	0.000	0.637	0.787
property_val_cat_0:review_scores_location	-0.0239	0.048	-0.501	0.616	-0.117	0.069
property_val_cat_2:review_scores_location	0.5889	0.334	1.761	0.078	-0.067	1.244
review_scores_value	1.7459	0.048	36.641	0.000	1.653	1.839
property_val_cat_0:review_scores_value	0.0893	0.059	1.521	0.128	-0.026	0.204
property_val_cat_2:review_scores_value	-0.1970	0.191	-1.034	0.301	-0.570	0.177
property_val_cat_0	-1.2263	1.189	-1.031	0.302	-3.557	1.104
property_val_cat_2	-14.1205	5.150	-2.742	0.006	-24.215	-4.026
gym	0.0753	0.055	1.363	0.173	-0.033	0.183
internet	0.0040	0.032	0.124	0.902	-0.059	0.067
twentyfour_hour_check_in	-0.1239	0.037	-3.343	0.001	-0.197	-0.051
laptop_friendly_workspace	0.0469	0.031	1.493	0.135	-0.015	0.108
private_entrance	0.1697	0.074	2.296	0.022	0.025	0.315

Since we do not have many factors, we want to interpret p-values naively around less than 0.15. We can see that all those review scores from the five sections itself are very significant to the overall review score.

For the insignificant variables,

1) Interaction between review score of checkin and low property value are insignificant to overall review score due to high p-value.
2) Interaction between review score of cleanliness and low property values are insignificant to overall review score.
3) Both interaction terms between communication score and low and high property values are very insignificant to overall review score.
4) Interaction term between review score of location in low property values is insignificant to overall review score.
5) Interaction term between review score of value and high property value is insignificant to overall review score.
6) Low property value itself is insignificant to overall review score.
7) Amenities of gym, internet and laptop friendly workspace is seen to be insignificant to overall review score.

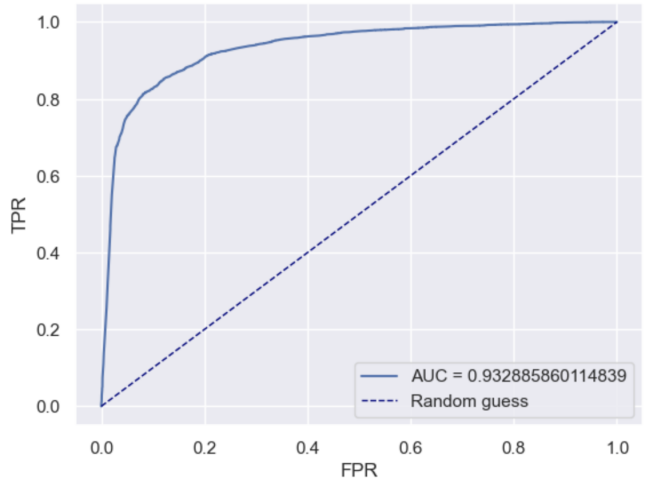
Now for the significant variables, we want to interpret by odds ratio for our logistic model;
Odds Ratio = exp(Beta). Our Beta term would be the coefficient column.
For example,

1) The Odds ratio for review score checkin to overall review score is 1.82. exp(0.5962) = 1.82 .
   This means, as the check-in review score is increasing by one unit, the odds of having overall review scores rating over 95 is increasing by 82%.
2) The Odds ratio for review score check-in at high property value to overall review score is exp(0.8658) = 2.38 .
   As the check-in review score at high property value is increasing by one unit, the odds of having overall review scores rating over 95 is increasing by a factor of 2.38.
3) The Odds ratio for review score of cleanliness to overall review score to overall rating is exp(1.3142) = 3.72.
   As the cleanliness review score is increasing by one unit, the odds of having overall review scores rating over 95 is increasing by a factor of 3.72
4) The Odds ratio for review score of cleanliness at high property value to overall review score is exp(0.2554) = 1.29.

As the cleanliness review score is increasing by one unit at high value property, the odds of having overall review scores rating over 95 is increasing by a factor of 29%.

5) The Odds ratio for review score of communication to overall review score is exp(0.9166) = 2.50.
   As the communication review score is increasing by one unit, the odds of having overall review scores rating over 95 is increasing by a factor of 2.50.
6) The Odds ratio for review score of location to overall review score is exp(0.7122) = 2.04 . As the location review score is increasing by one unit, the odds of having overall review scores rating over 95 is increasing by a factor of 2.04.
7) The Odds ratio for review score of location at high property value to overall review score isexp(0.5889) = 1.80 .
   As the location review score at high property value is increasing by one unit, the odds of having overall review scores rating over 95 is increasing by 80%.
8) The Odds ratio for review score of value to overall review score is exp(1.7459) =5.73 . As the value review score is increasing by one unit, the odds of having overall review scores rating over 95 is increasing by a factor of 5.73.
9) The Odds ratio for review score of value at low property to overall review score is exp(0.0893) =1.09.
   As the value review score at low property is increasing by one unit, the odds of having overall review scores rating over 95 is increasing by 9%.
10) The Odds ratio for review score of value to overall review score is exp(1.7459) =5.73 . As the value review score is increasing by one unit, the odds of having overall review scores rating over 95 is increasing by a factor of 5.73.
11) The Odds ratio for high property value to overall review score is exp(-14.1205) = 7.37131157e-7 . As the property value at high is increasing by one unit($), the odds of having overall review scores rating over 95 is decreasing by a factor of 7.37131157e-7.
12) The Odds ratio for 24 hour check-in to overall review score is exp(-0.1239) = 0.88 . When the 24 hour check-in is present, the odds of having overall review scores rating over 95 is decreasing by a factor of 0.88 than none 24 hour check-in.
13) The Odds ratio for private entrance to overall review score is exp(0.1697) =1.18 . When the private entrance is present , the odds of having overall review scores rating over 95 is increasing by 18% than non private entrance.

If we run the AUC for this current model we obtain an AUC of 0.93 which is a very precise model.



We also run a logistic model of only review score predictors to see how it performs without interaction terms. Here, we have the summary of

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-50.9719	0.559	-91.141	0.000	-52.068	-49.876
review_scores_checkin	0.6568	0.037	17.699	0.000	0.584	0.730
review_scores_cleanliness	1.3179	0.022	60.324	0.000	1.275	1.361
review_scores_communication	0.9118	0.044	20.595	0.000	0.825	0.999
review_scores_location	0.6928	0.023	30.726	0.000	0.649	0.737
review_scores_value	1.8058	0.028	65.600	0.000	1.752	1.860

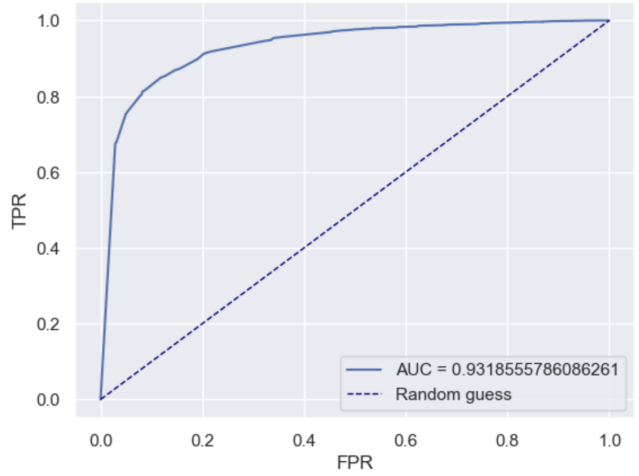
Here, all of the review score predictors are significant.
The Odds Ratio for each predictors to Overall review Scores are:

1) OR(check-in) = exp(0.6568) = 1.93
2) OR(cleanliness) = exp(1.3179) = 3.74
3) OR(communication) = exp(0.9118) = 2.49
4) OR(location) = exp(0.6928) = 2.00
5) OR(value) = exp(1.8058) = 6.08

As the review score is increasing by one unit for each section, the chance of getting an overall review score above 95 is increasing by those stated factors.

If we run the AUC for the review score predictors only, then we have AUC of 0.93185, implying that including interaction terms between property values and review scores do not necessarily increase the accuracy of the model, but it does not decrease the accuracy either. Since some of the interaction terms seemed to be significant according to the previous model, we would still prefer to use the previous model over the model we just generated, where we can take actions on listings depending on property values.

We now want to see if the logistic model with only amenities predictor without interaction terms.

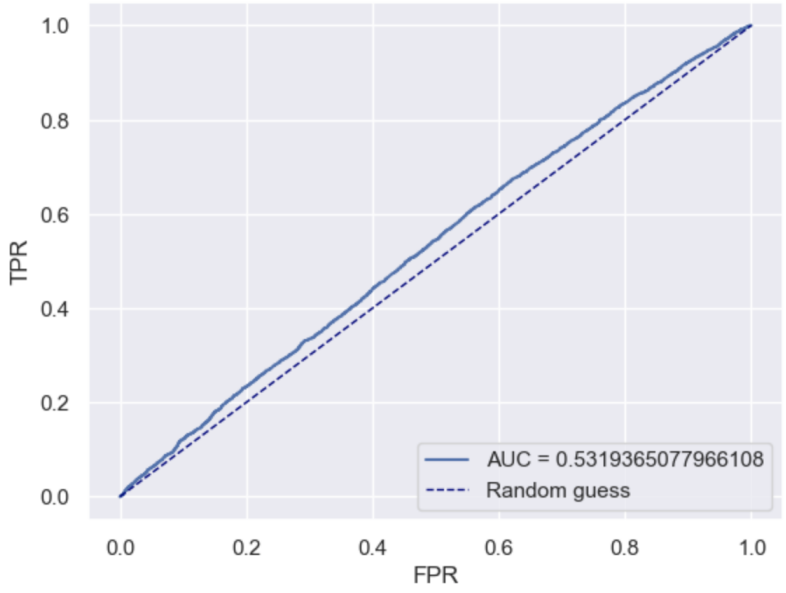


	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.6531	0.049	13.306	0.000	0.557	0.749
gym	0.1188	0.042	2.827	0.005	0.036	0.201
internet	-0.0668	0.023	-2.922	0.003	-0.112	-0.022
twentyfour_hour_check_in	-0.1115	0.026	-4.290	0.000	-0.162	-0.061
laptop_friendly_workspace	0.0931	0.024	3.807	0.000	0.045	0.141
private_entrance	0.2265	0.050	4.525	0.000	0.128	0.325
washer	0.0681	0.022	3.061	0.002	0.024	0.112
kitchen	0.1305	0.041	3.208	0.001	0.051	0.210
elevator_in_building	-0.0078	0.027	-0.286	0.775	-0.061	0.046
breakfast	0.0883	0.037	2.357	0.018	0.015	0.162
heating	-0.1494	0.041	-3.658	0.000	-0.229	-0.069
free_parking_on_premises	-0.0784	0.031	-2.513	0.012	-0.140	-0.017
tv	0.0728	0.022	3.266	0.001	0.029	0.117
hangers	0.0103	0.027	0.382	0.702	-0.042	0.063
shampoo	0.0461	0.023	2.021	0.043	0.001	0.091
family_kid_friendly	-0.0530	0.021	-2.569	0.010	-0.093	-0.013
doorman	0.0702	0.044	1.597	0.110	-0.016	0.156
pets_allowed	0.0928	0.031	3.009	0.003	0.032	0.153
hot_tub	-0.1033	0.045	-2.319	0.020	-0.191	-0.016
hair_dryer	0.0020	0.027	0.073	0.942	-0.051	0.055
carbon_monoxide_detector	-0.0361	0.022	-1.628	0.104	-0.080	0.007
fire_extinguisher	-0.0255	0.024	-1.070	0.285	-0.072	0.021

essentials	0.0113	0.028	0.399	0.690	-0.044	0.067
cable_tv	-0.0449	0.025	-1.824	0.068	-0.093	0.003
safety_card	-0.0436	0.033	-1.323	0.186	-0.108	0.021
air_conditioning	-0.0490	0.027	-1.812	0.070	-0.102	0.004
lock_on_bedroom_door	-0.0194	0.025	-0.790	0.429	-0.068	0.029
first_aid_kit	-0.0009	0.025	-0.037	0.970	-0.049	0.047
smoking_allowed	-0.0814	0.043	-1.911	0.056	-0.165	0.002
iron	0.0111	0.027	0.415	0.678	-0.041	0.064
indoor_fireplace	0.0778	0.052	1.496	0.135	-0.024	0.180
wheelchair_accessible	0.0068	0.042	0.162	0.871	-0.075	0.089

In this model, the insignificant amenities are :
elevator, hangers, doorman, hair dryer, carbon monoxide detector, fire extinguisher, essentials, safety card, lock on bedroom door, first aid kit, iron, indoor fireplace, and wheelchair accessible.
Most significant amenities can be said to be:
gym, internet, 24-hr checkin, laptop friendly workplace, private entrance, washer, kitchen, tv, breakfast, heating, air conditioning, pets allowed, hot tub, smoking allowed, etc.

We want to see whether or not this model is still accurate by solving for AUC. Here, since we obtain AUC of 0.532. This is a very poor model of only amenities without the review score predictors.



Now we want to run a model on amenities with interaction terms between groups of high and low property value.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.7276	0.034	21.686	0.000	0.662	0.793
gym	0.1478	0.062	2.386	0.017	0.026	0.269
property_val_cat_0:gym	-0.0042	0.077	-0.054	0.957	-0.156	0.147
property_val_cat_2:gym	0.7045	0.371	1.897	0.058	-0.023	1.432
internet	-0.1389	0.037	-3.748	0.000	-0.212	-0.066
property_val_cat_0:internet	0.0902	0.045	1.991	0.047	0.001	0.179
property_val_cat_2:internet	0.1844	0.148	1.245	0.213	-0.106	0.475
twentyfour_hour_check_in	-0.0759	0.043	-1.753	0.080	-0.161	0.009
property_val_cat_0:twentyfour_hour_check_in	-0.0487	0.053	-0.917	0.359	-0.153	0.055
property_val_cat_2:twentyfour_hour_check_in	-0.3426	0.165	-2.074	0.038	-0.666	-0.019
laptop_friendly_workspace	0.1319	0.036	3.669	0.000	0.061	0.202
property_val_cat_0:laptop_friendly_workspace	-0.0345	0.044	-0.782	0.434	-0.121	0.052
property_val_cat_2:laptop_friendly_workspace	0.1864	0.143	1.302	0.193	-0.094	0.467
property_val_cat_0	-0.1040	0.041	-2.540	0.011	-0.184	-0.024
property_val_cat_2	-0.3935	0.131	-3.003	0.003	-0.650	-0.137

Here the insignificant values are :

1) gym at low property value listing
2) Internet at high property value listing
3) 24 hour check-on at low property value listing
4) laptop friendly workplace in both low and high property value listing

The pattern is that the presence of amenities in lower property values are more significant in getting an overall 95 rating. The reason could be that people tend to have lower expectations of amenities in high property value listings than low property values since they are more concerned about the cleanliness, price of accommodation in high property value listings.

Now we solve if there is overfitting for our model by using k-fold(k=10) cross validation.If we run the cross validation for our first model with the review score value and the prediction term plus all of the amenities. Since we have a very low standard deviation AUC of 0.00754 and a high mean AUC for this model, we can conclude that this model is accurate and not overfitted.

```
0.931382990815323 0.0075427149561615 94
```

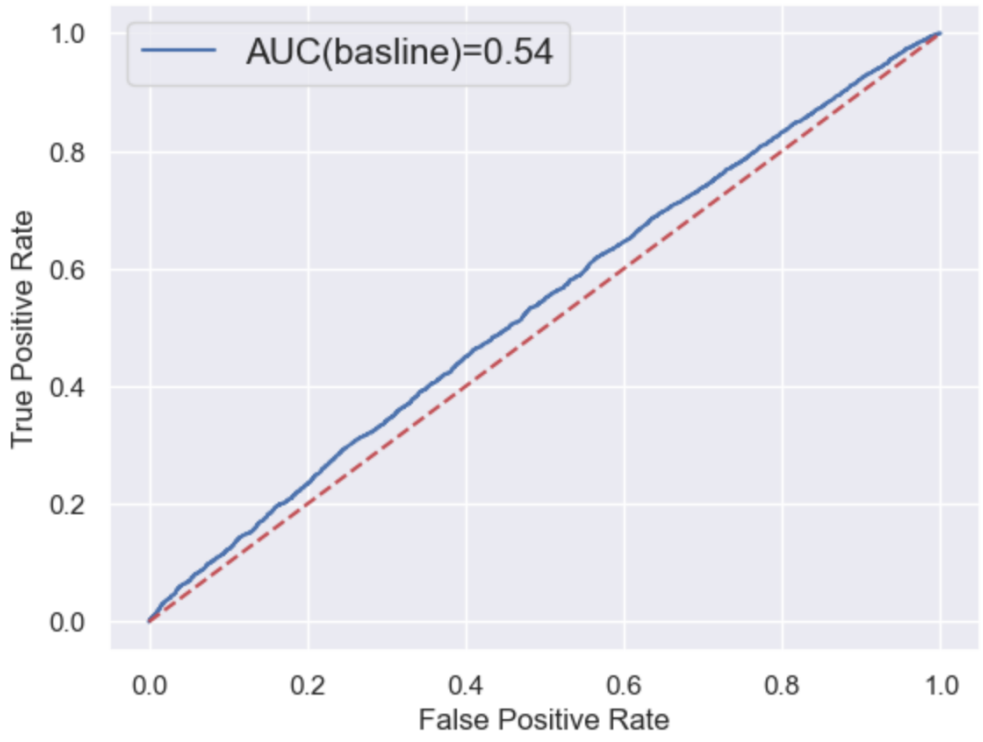
Random Forest

```
RandomForestClassifier(max_depth=100, max_leaf_nodes=5, min_sample
s_split=3,
                       n_estimators=80, random_state=1337)
```

To avoid an overfitted model, it is required to tune the model by stating several conditions.

```
 Training Accuracy :  65.57863501483679
  Testing Accuracy :  65.78617403661289
```

By the result above, it is clear that the model is not overfitted since the accuracy of the training set and testing set are almost same. However, the model is not perfect because the model's accuracy is only 65 percent.



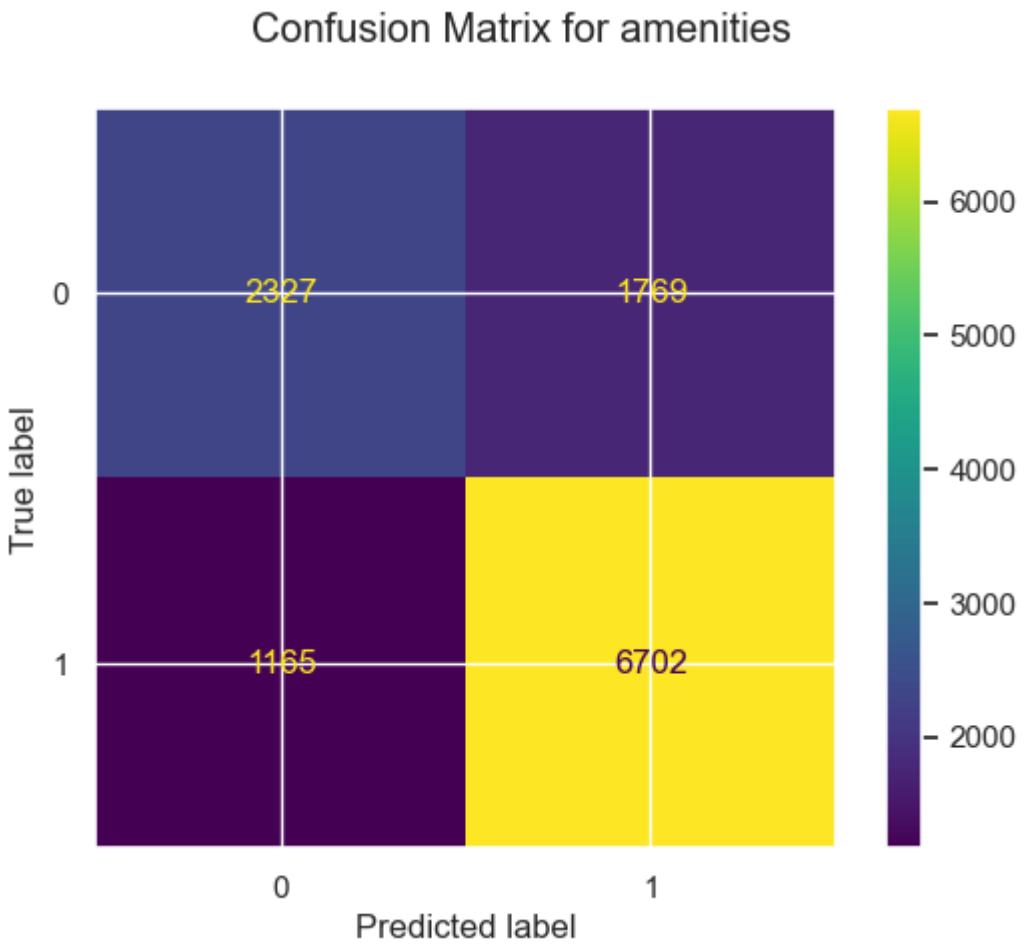
This is the AUC score and the ROC curve for the RandomForest model. The result is similar to the logistic model with all review predictors without interaction term. It is quite difficult to obtain a significant model with RandomForest because we can not add interaction terms. Unlike linear regression and logistic regression, Random

Forest is not a parametric model. Therefore, we decided to move on to the next model which is neural networks.

Neural Network

For neural network, we were unable to use tensorflow due to computing power and we instead used scikitlearn's multilayer perceptron model. Dataset was splitted into training and test data sets and standardized both of them before fitting a model. We set our activation function to be Relu(Rectified linear unit), where the output of the function is 0 when x is less than 0 and x when x is larger than 0. And, the optimization algorithm was set to Adam.

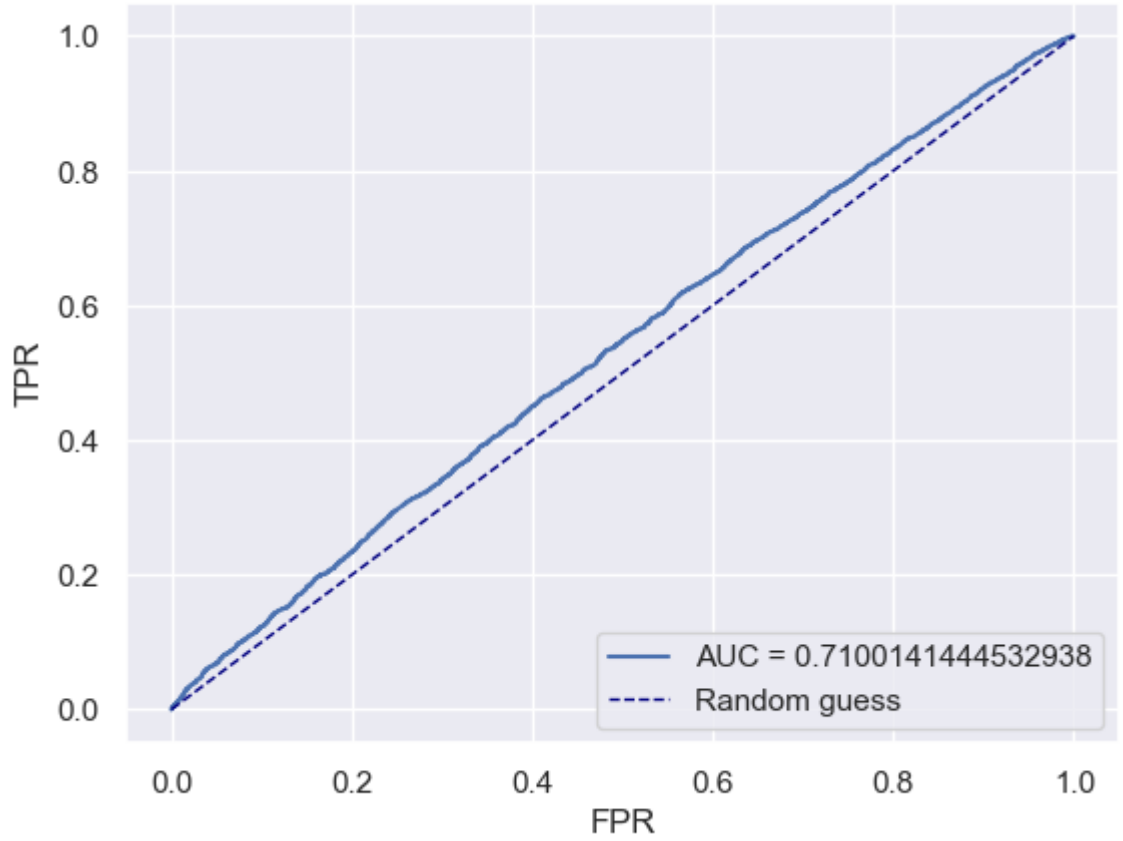
We then got accuracy of our model based on the test dataset. Accuracy was 0.75, which from our perspective, seemed satisfactory.



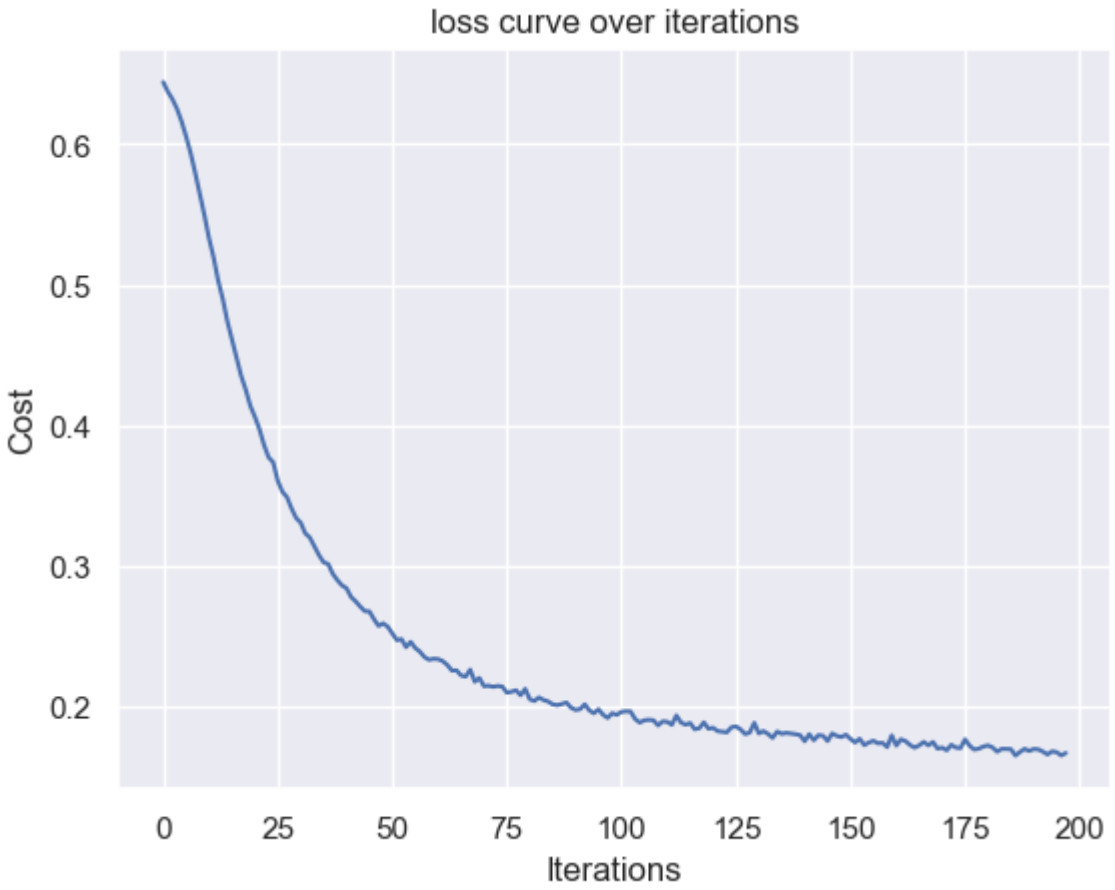
The confusion matrix is a great visualization tool to see whether our model has a desired level of accuracy. The x-axis represents the predicted label and the y-axis represents the true label. So, the yellow area represents the true positive rate and the blue area represents the true negative rate. As we can see from the confusion matrix, the true positive rate seems to be higher than the true negative rate.

	precision	recall	f1-score	support
0	0.67	0.57	0.61	4096
1	0.79	0.85	0.82	7867
accuracy			0.75	11963
macro avg	0.73	0.71	0.72	11963
weighted avg	0.75	0.75	0.75	11963

The summary result precisely tells us that predicting 1 accurately, so the probability of a listing getting more than 95 for the overall review rating when the model expects to, is 0.79, compared to the probability of a listing getting lower than 95 for the overall review rating when the model expects to, which is 0.67. Even though the true negative rate seems relatively low, the high true positive rate is more desired for the analysis as accurately predicting having a satisfied customer from only the list of amenities and property values is what we think is more beneficial for Airbnb.



We then again plotted ROC and computed AUC. AUC for the model is 0.71, which is a huge improvement over the logistic regression model and the random forest model.



The plot above is a loss curve of the model. The loss curve seems to be quite stable and we can conclude that the model is not overfitting since there will be spikes intermittently in a loss curve if a model is overfitting.

So, even though we do not know how each individual amenity and property value affects the prediction of having a satisfied customer, we were able to get a model that predicts to a desired level with neural network, by which we can predict whether a new listing without any prior reviews can get over 95 for overall review rating with only amenities and property values.

Concerns and limitations

We tried to do a time series analysis on listings data based on economic trends derived from econ_state, but as we merged the listings data with calendars data, we found that the listings data ranges from May 2016 to May 2018. GDP and GDP_per_capita are in quarters, which mean there are only 7-8 data points available for analysis per each state and we gave up on time series analysis and went for analysis on listings data based on economic data of states calculated from the mean of GDP and GDP_per_capita from Q2 2016 - Q2 2018 in five states.

The limitation was that we had some overlapping categories in the amenities such as wireless connection, internet. These could have been affecting our results of the logistic regression model in amenities. Speaking of the internet, we found that when the internet is present in the listings, the overall review score 95 actually has decreased which is very counterintuitive so we might need more domain knowledge on such amenities. One of the possible reasons is that they focus more on the internet which distracts their sleep but lack of internet is still inconvenient. The most reasonable background is that the connection level for such listings might not be so fast which customers can feel as inconvenience.

Another major limitation was that we could not find the dataset for the hotel in order to compare them with airbnb listings and thus inability to run a NLP model on these reviews of hotels. All of the dataset was not free and we needed the API to connect it to python which was very problematic.

Conclusion

Finally, success in airbnb would be very crucial to getting a high overall review score. Thus, we would like this business to be successful by attracting more customers with listings that are likely to get high overall review scores. Review score of check-in, cleanliness, communication, location, values were all deeply interacted to overall score getting over 95, which we expected to confirm through our analysis.

According to the different levels of property value in the listings, we are more likely to get an overall review score over 95 when the property values are higher. Regardless of the property values, getting high points in the five sections of review score (from 8-10) will lead to the higher chance of getting an overall rating over 95, especially 10 points.

When the property values are in the low group, the review score of location is likely to be lower than 8 and thus, the chance of failure fulfilling an overall rating over 95 is increasing. So we recommend more decisions have to be made whether the business is accepting house listings in the very low property values groups.

In terms of logistic regression models, we came up that

1) For overall review scores, the most significant factors are review scores of check-in, cleanliness, communication, location, values.
2) Getting high review scores of check-in is very effective in increasing the chance of overall rating over 95 among high property value groups whereas in low group, review score of check-in is insignificant. The reason is that customers do not consider the experience check-in low property listings not important due to low expectation.
3) On the contrary, high review scores of check-in among high property listings is very significant in increasing the chance of overall rating over 95. The interaction between high property values and check-in review score is strong,

so the business must recommend the listings to have a good check-in system in high property value groups.

4) Aside from the check-in review score, the four sections of , cleanliness, location and especially communication values in high or low property value are insignificant to getting an overall rating over 95 if we set our confidence level to be 0.05. Cleanliness and location in higher property values seemed to be relatively significant even though the p-value was higher than 0.05.
5) Among all review scores, the value score had the highest odds ratio of increasing the chance of getting an overall 95 rating by 5. The order of importance is :
   value -> cleanliness -> communication -> location -> check-in.
   Thus, we can see what customers focus most on selecting accommodations are price, and cleanliness. We must have advantages in price over our competitors.
6) It is recommended that from our logistic model, on low property value listings, they should focus on getting a higher value review score. The reason is that customers have low expectations on cleanliness, check-in and location so they tend to evaluate the value for the money in such listings.
7) On the contrary, high property value listings must focus on improving the check-in system, cleanliness and location since they have a certain increase in overall rating.
8) We have concluded that when 24-hour-self check-in amenity is present, the chance of getting an overall 95 rating rather goes down by 12%. The possible reason might be that people do actually feel more comfortable when they check in person. Maybe, there might be a problem in the self check in system in the app of airbnb where customers do feel inconvenience. For example, unclear instruction might lead to inconvenience in check in. We suggest improvement in the self check-in system.
9) Amenities are very insignificant and can be said to have non-linear relationship between overall review score over 95 due to extremely low AUC values.

For such a problem in 9), we ran a neural network and got the predicting overall review over 95 accurately. So the probability of a listing getting more than 95 for the overall review rating when the model expects to, is 0.79 which has a good accuracy of predicting overall review over 95 of listings.
When we accept the new listings, we do not have any information on past review scores but the amenities and property values are given by the locations.
Therefore we can apply this neural network to our new recommendation system. In this way, customers will be recommended such listings with higher probability of satisfying them derived from the model.

Appendix

Timeline

Date	Name	Details
Week 1 (Sep 6 – Sep 12)	Team Formation	
Week 2	Work on idea formation	Formed the team
Week 3	Idea should be finalized	Decided to choose our topic to be economic related since consumer behavior is normally heavily influenced by economic status
Week 4	Datasets sourced	Decided to use listings, econ, property s data for the project
Week 5	Basic EDA	Tried to do Time Series analysis on Listings with Econ data, but found that the time frame is not suitable for analysis
Week 6	FALL READING WEEK	Basic EDA including summary statistics, histogram, boxplot, heatmap, and etc.
Week 7	Basic EDA/ Writing of midterm report	Basic EDA including summary statistics, histogram, boxplot, heatmap, and etc. discussed what models and statistical techniques we can adopt for analysis
Week 8	Basic EDA / Deliver midterm report	Decided to focus on New York since New York accounts for approximately 75% of

		the listings dataset and Econ dataset is not as helpful as we expected. So dropping econ dataset
Week 9	In-depth EDA	Sourced additional dataset for hotels from Yelp
Week 10	In-depth EDA / Statistical analysis	Doing NLP on NY hotels is not possible; data for NY does not exist
Week 11	In-depth EDA / Statistical analysis	Decided to utilize k-means(k=3) algorithm to classify property value by unsupervised learning
Week 12	Statistical analysis / Machine Learning modeling	chi-square test of independence on amenities to decide on explanatory variables to put in the logistic regression model
Week 12	Statistical analysis / Machine Learning modeling	Generated Logistic regression models
Week 14	Writing of final report / presentation video rehearsal	Generated a random forest model and a neural network model
Week 15 (final exams weeks)	Final report and presentation video	Writing final report/presentation