Data Analytics Competition Report

By Power of Girls

November 9th, 2019

Contents

Introduction

Airbnb was built on the idea that everyone should be able to take the perfect trip, including where they stay, what they do, and who they meet. Due to the continuous development of tourism in recent years, Airbnb is getting much more popular. Now when we open the website of Airbnb, we can find various types of accommodation for us to choose. Unlike many traditional hotels under unified management, most of the accommodation types on Airbnb are family hotels.

Boston, MA, is a quintessential blend of colonial history and cutting-edge innovation. Tens of millions of people visit Boston each year to take in its historic sites, diverse neighborhoods, cultural or sporting events.

As Airbnb in Boston is becoming popular during recent years, we are assigned to analyze the data to discover insights behind the information and make recommendations for the future. The dataset was obtained from Airbnb website which covers detailed information of listings and reviews in Boston, MA from January 2019 to September 2019.

In this data analytics report, we looked into Airbnb Boston Data to explore its geography patterns, business growth, availability, price changing trends and etc. We also performed multivariate regression analysis and machine learning to predict pricing. Our goal is to summarize the overall situation of Airbnb in Boston, reveal critical issues may have effects on its future prosperity and make recommendations to improve Airbnb services.

Part I Descriptive Statistics

1.1 Listings of Airbnb in Boston

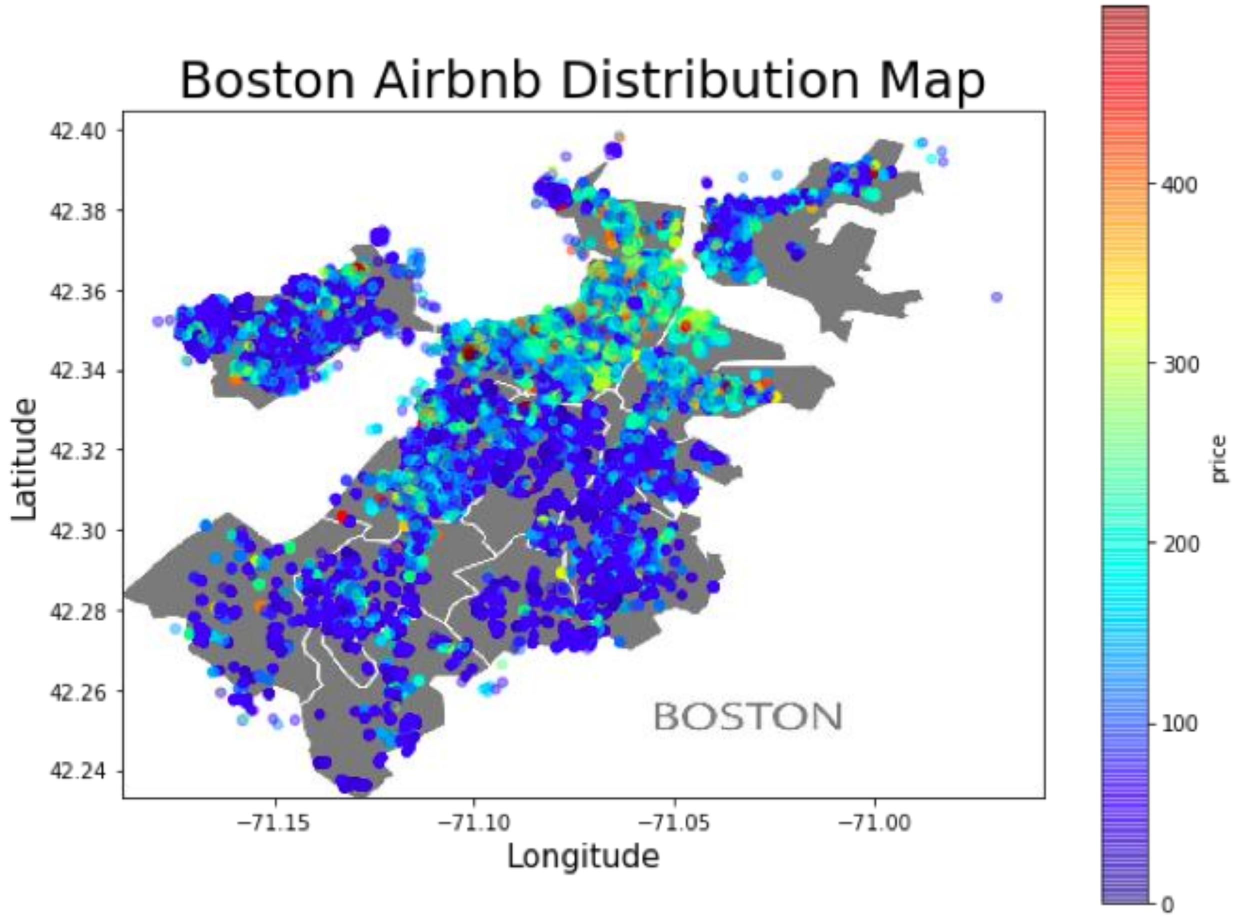


Fig 1.1 Boston Airbnb Distribution Map

Analysis:

We used the data of listings_summary.csv to plot the distribution of Airbnb in Boston with a daily price of no more than 500(delete outliers).

From this graph, we noticed that ***Airbnb is densely distributed in Boston, especially around Back Bay, Allston and Dorchester.*** In general, the denser the distribution, the higher the price.

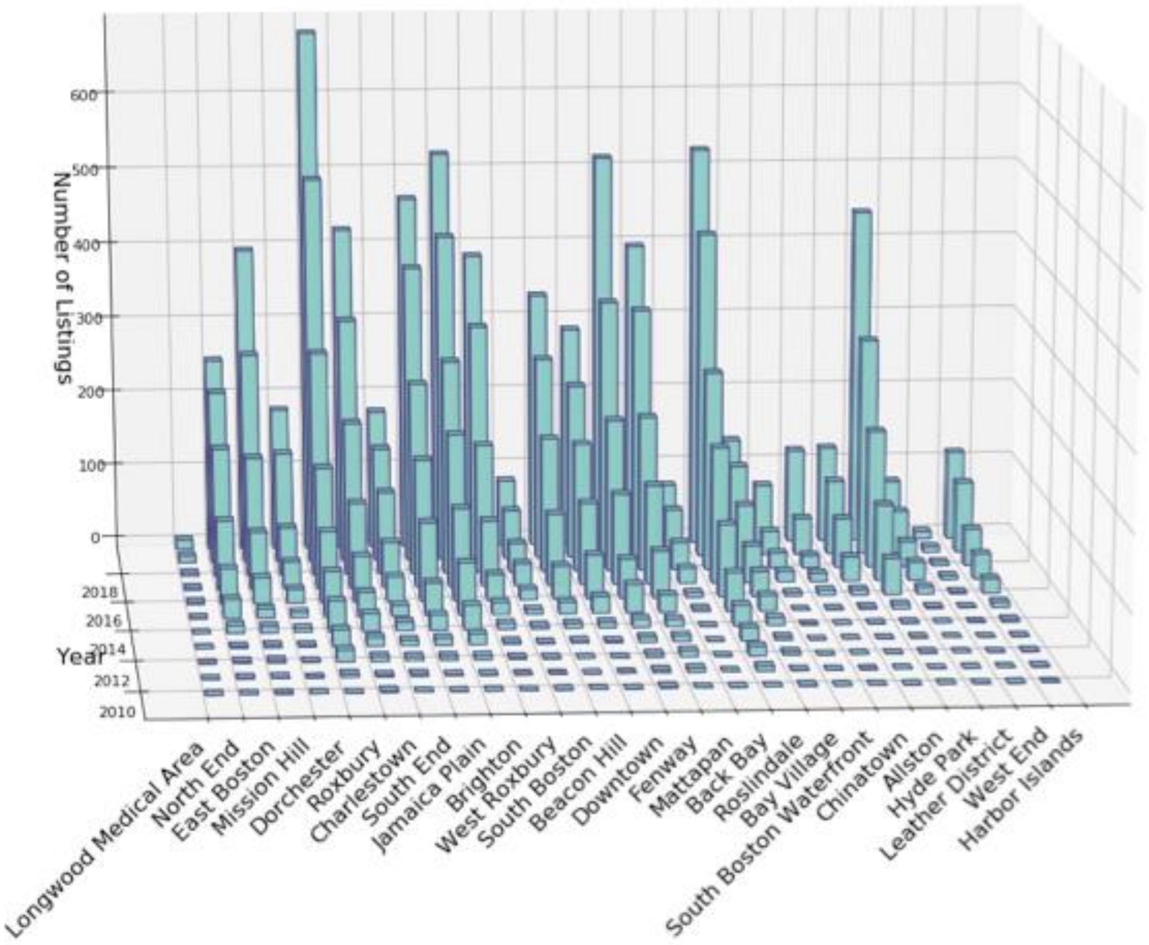


Fig 1.2 Listing Numbers Across Years in Different Neighborhoods (3D Scale)

Analysis:

The 3D bar chart demonstrates each year's listings number in different areas. ***The overall trend is increasing among neighbourhoods***, to be more specific, Dorchester, Back Bay and Jamaica Plain saw a marked growth as early as 2012. The figure for Downtown has quickly caught up since 2015. As for Allston, it wasn't one of the hottest markets for renters before 2016 but has developed dramatically since then. The listings of other neighborhoods such as South End, Fenway and Brighton closely followed, with around 600 listings in 2019.

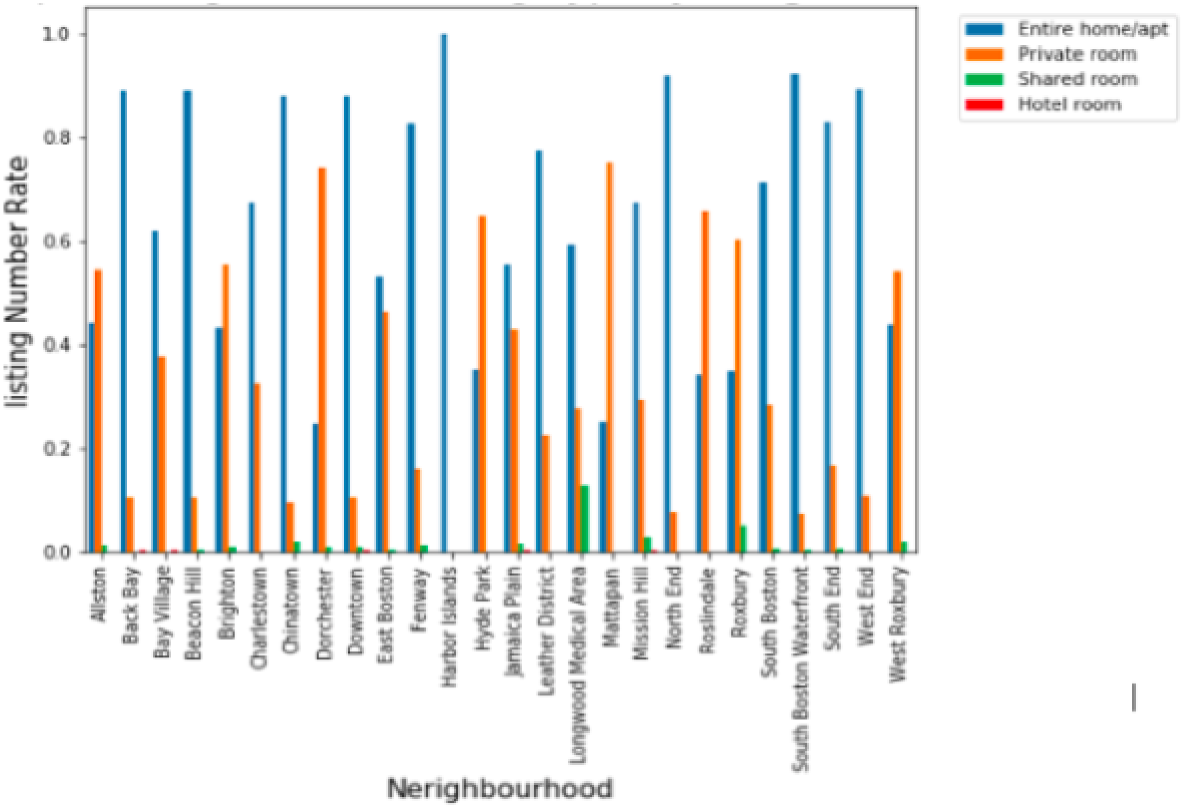


Fig 1.3 Listing Types on Neighbourhoods

Analysis:

We are curious about the distribution of listing number in different neighbourhoods and room types. So, we made a graphic on the percentage of different room types in all neighbourhoods. In general, ***Entire room/apt has the highest listings*** among all four kinds of room types. However, ***Hotel room has the lowest listing number***. This result is in accordance with the reputation that Airbnb is famous for its family hotels rather than traditional hotels.

1.2 Reviews by Customers

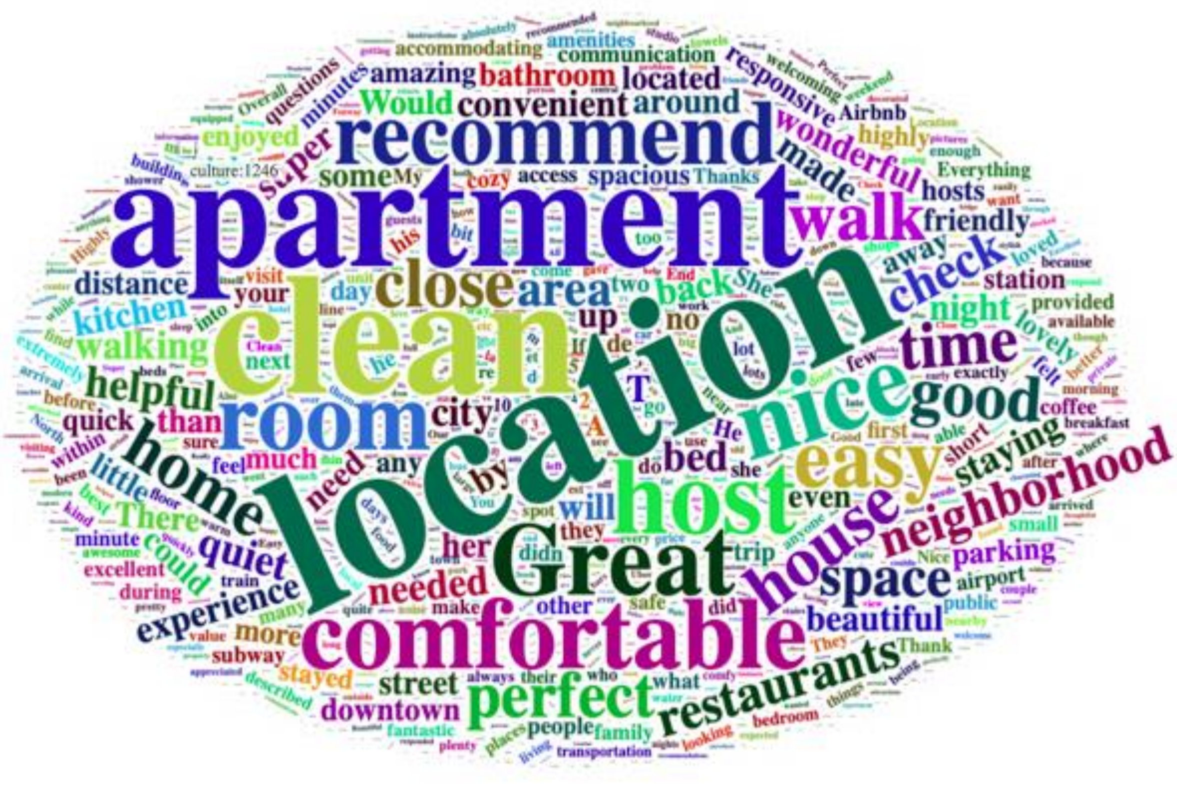


Fig 1.4 WordCloud on KeyWords in Reviews

Analysis:

After investigating the overall listing market of Airbnb in Boston, we turned our attention to the customer part.

From this WordCloud, we can easily see that the main concerns of Boston customers include ***location, area, neighborhood, comfortable, convenient***. Another important aspect is house type, ***apartment*** is mentioned more than ***house****.* Besides these, rooms and cleanliness also have a significant impact on customers as the frequent use of words such as ***clean, bathroom, room, kitchen*** in their review***.***

The factors mentioned above are the major influence on consumer behavior.

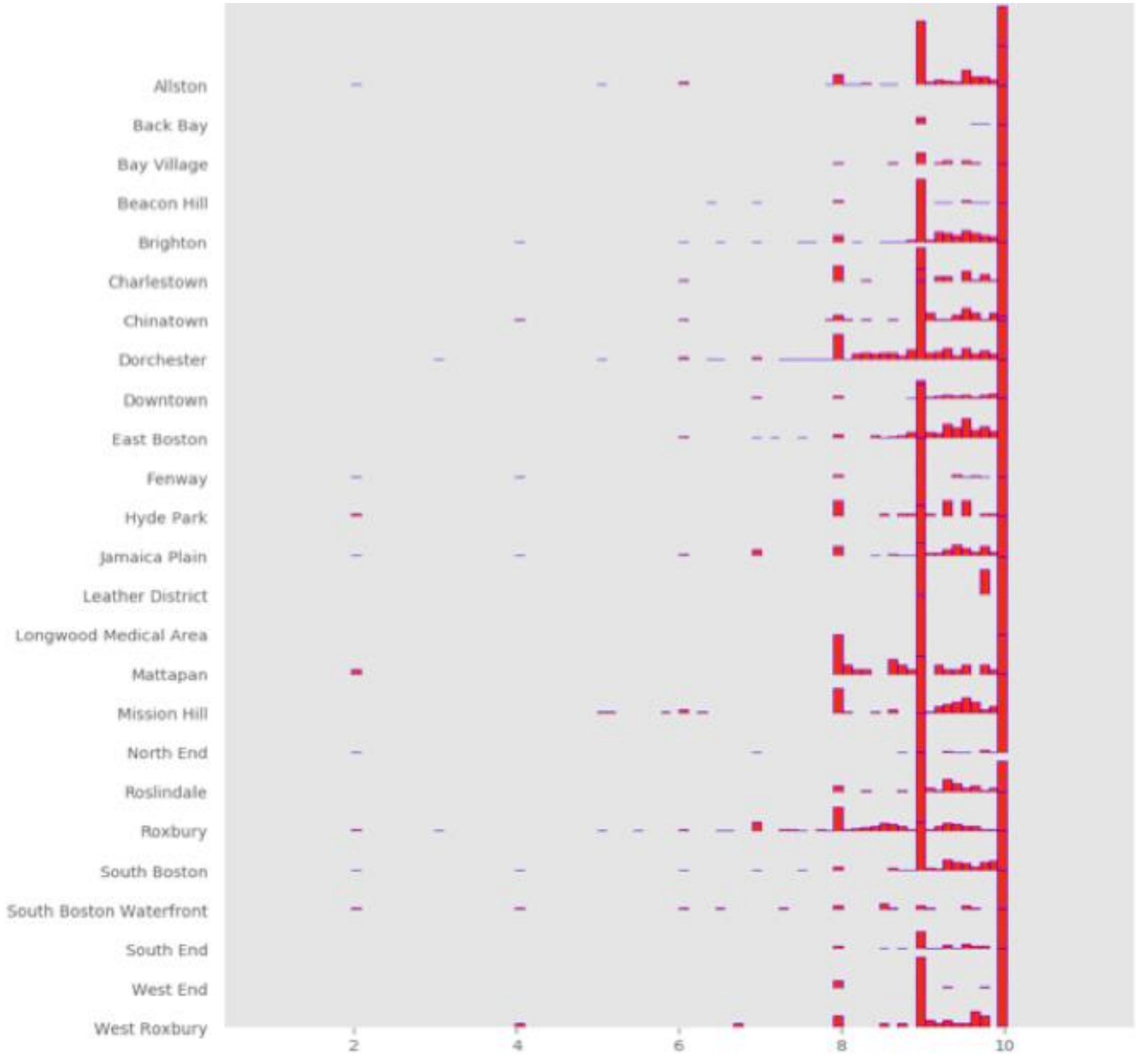


Fig 1.5 Scores for listings by neighborhood

Analysis:

After the initial glance of reviews, we focus on the detail scores among neighbourhoods. Figure 1.5 describes the overview scores on the listings. Most customers tend to give a score higher than 8 but some neighbourhoods have much lower scores than others. Mattapan and Hyde Park have the lowest scores but this might be due to the outliers.

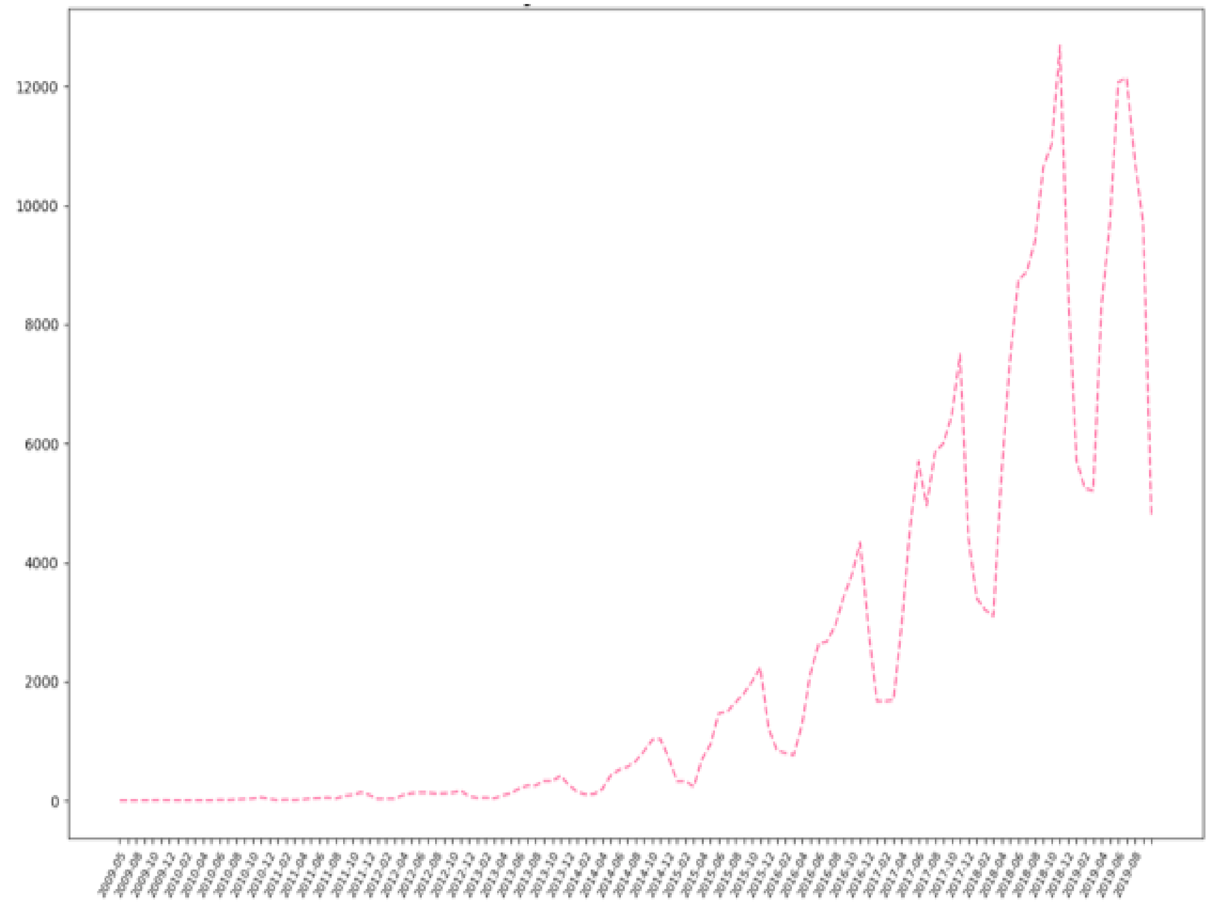


Fig 1.6 Seasonality of Number of Reviews

Analysis:

To see whether seasonality has an influence on review numbers, we draw the graph on month level. We can see that ***Autumn always has the peak of reviews while Winter encounters obviously drop.*** It is probably because of the freezing cold weather in Boston winter

1.3 Price of Airbnb in Boston

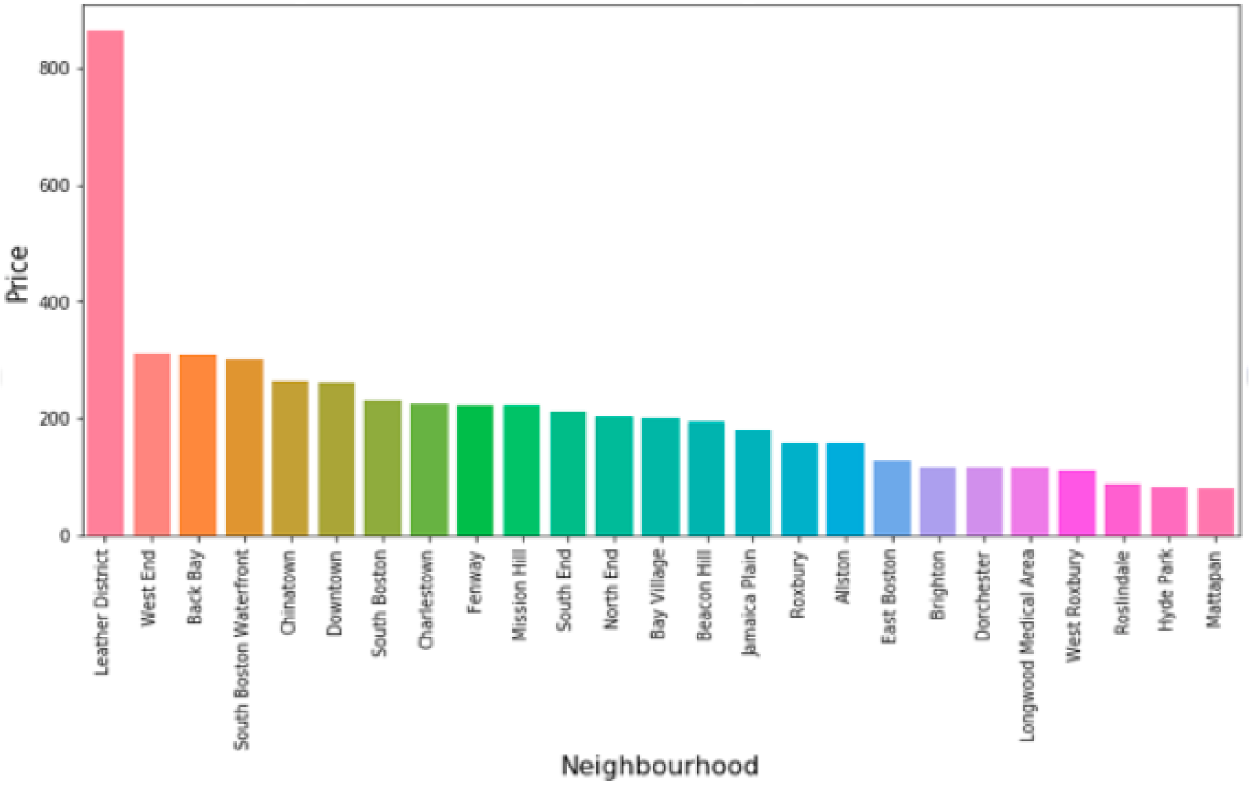


Fig 1.7 Average Prices for Neighbourhoods

Analysis:

After having a general idea about both supply and demand part of Airbnb in Boston, we are trying to explore the most important issue people are concerned about the price. The graph above shows that the average price of Airbnb in Leather Distinct is highest, which reaches more than $800 per day *(although, later on we took those above $800 as outliers)*. Airbnb located in ***West End, Back Bay and South Boston Waterfront have relative higher prices.*** On the contrary, Mattapan has the lowest house rates. Therefore, we come to the conclusion that the influence of geographical location on price is large.

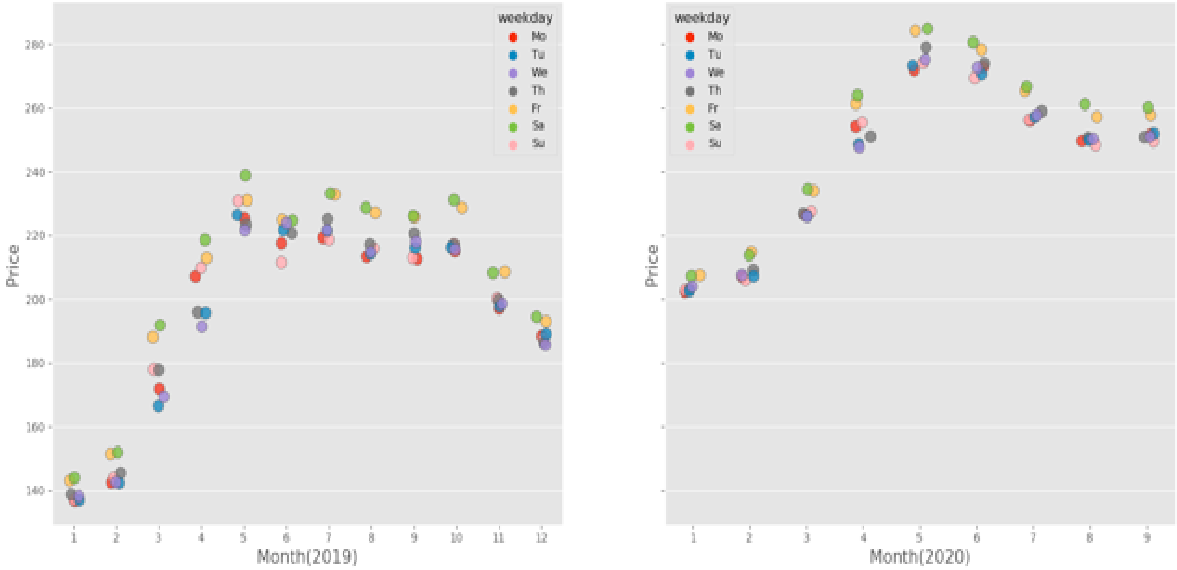


Fig 1.8 Seasonality of Price

Analysis:

We use the Calender.csv to compare the seasonality of price. The average price per month grouped by weekdays is shown above, from which we can see the seasonality in pricing. Within one year, the average price is expected to increase since January, reaching a peak in May, after which the figure shows a steady decline. When putting 2 years together, it is obvious that in overall, the increasing trend of average list price is projected to continue in the future.

Another thing is that, in common sense, the price on weekends can be much higher than workdays. However, the plot implies that the average price in Friday and Saturday manifestly outnumbers that in other weekdays. In contrast, price on Sunday is not significantly different from other workdays.

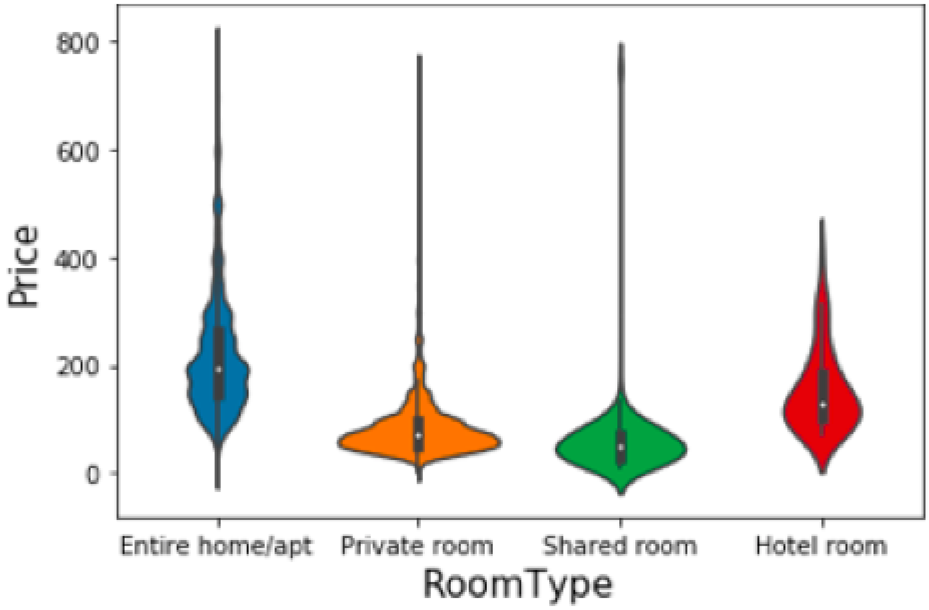


Fig 1.9 Density and Distribution of Prices for Room Types

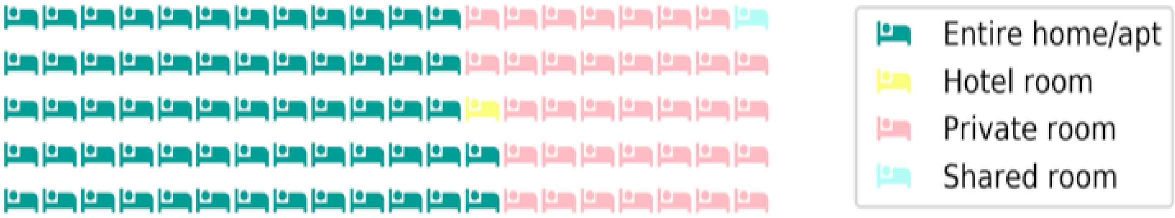


Fig 1.10 Percentage of Room Type

Analysis:

Daily Housing Price is the target variable of further analytics, thus we need to pay more attention to explore what factors affect price. We first connect price to **Room Type** in this graph.

From graph1.10 we can see clearly that ***entire room/apt and private room take up the most percentage***, and entire room/apt is higher. Hotel room and shared room are in similar percentage.

The 1.9 violin plot shows the relationship between Price and RoomType. To prevent the influence of outliers, we only chose the price below $800 per night to plot this graph. From this graph, ***we noticed that shared room has the lowest average price among these four room types.*** Second lowest average price belongs to Private room. ***The price range of Entire home/apt is the largest one which state that price volatility of Entire home/apt is the greatest.***

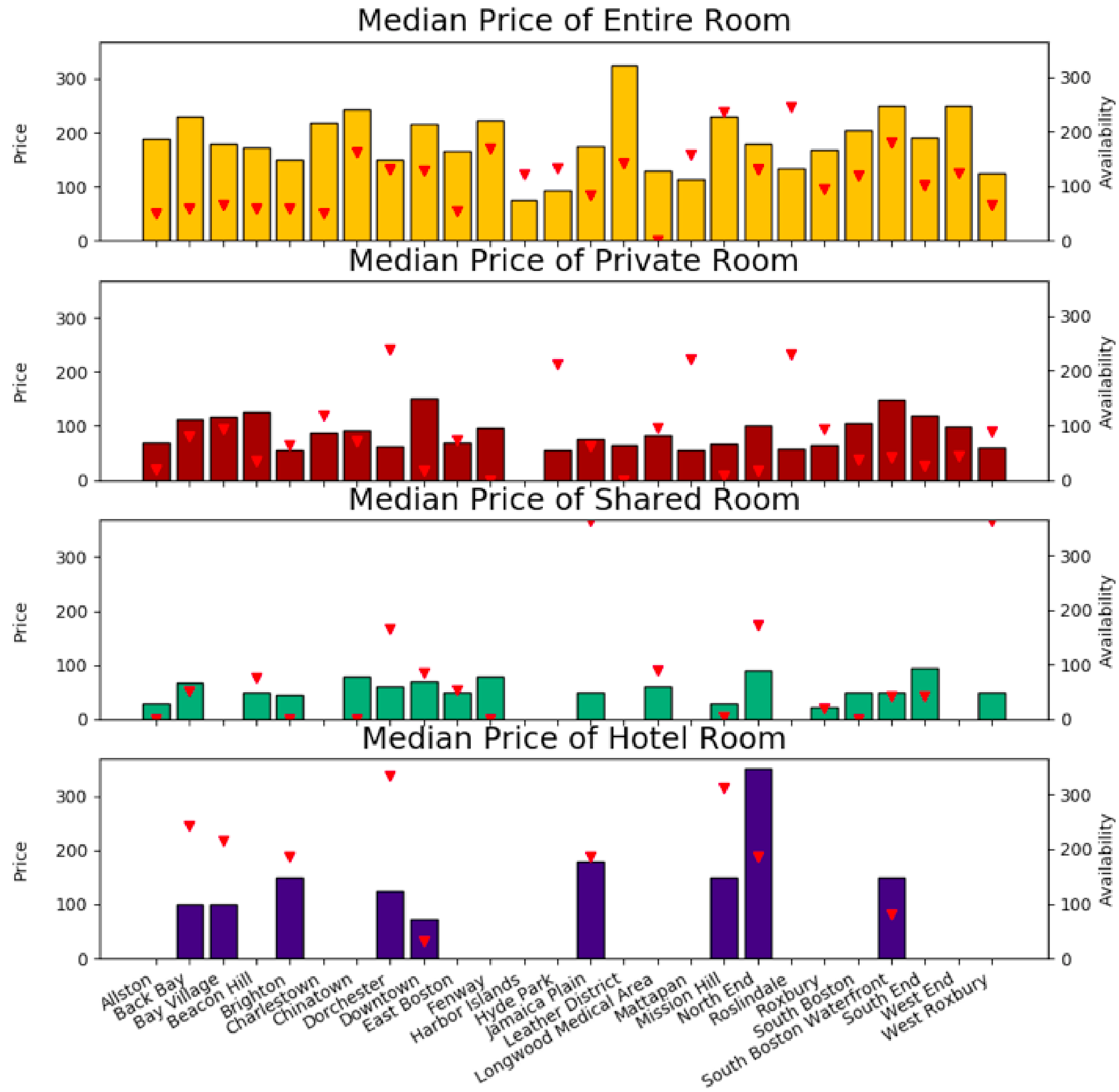


Fig 1.11 Price and Availability for Different Rooms

Analysis:

Then we add neighbourhoods and availability on different room types, and make the plot to see the median price. Generally, ***we can see positive relationship between availability and price for the entire room,but the relationship for private room is not clear.*** It may be because private room has more important variables to explain for price. We can also see the neighbourhood price distribution, it is different among room types.

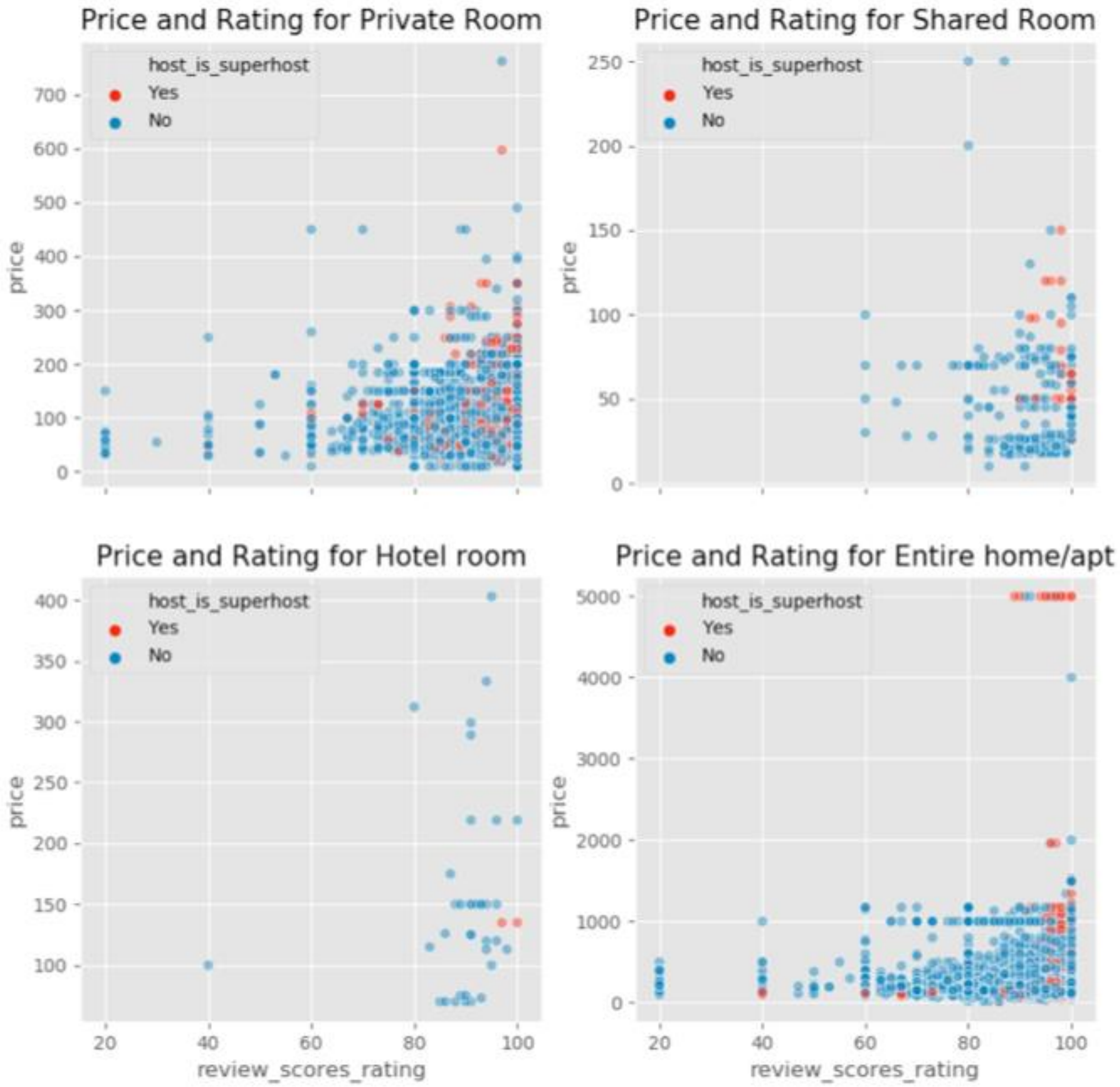


Fig 1.12 Relationship of Scores Rating and Price

Analysis:

Other variables we have interested in are ratings and whether the host is superhost. The scatter plot gives information about the relationship between scores rating received per rental and its price. In these points, the red ones are linked with super hosts.

At first glance, we can tell that there is way more dots for entire home/apt and private room type, indicating that the number of these 2 kinds far outstrips that of hotel room and shared room.

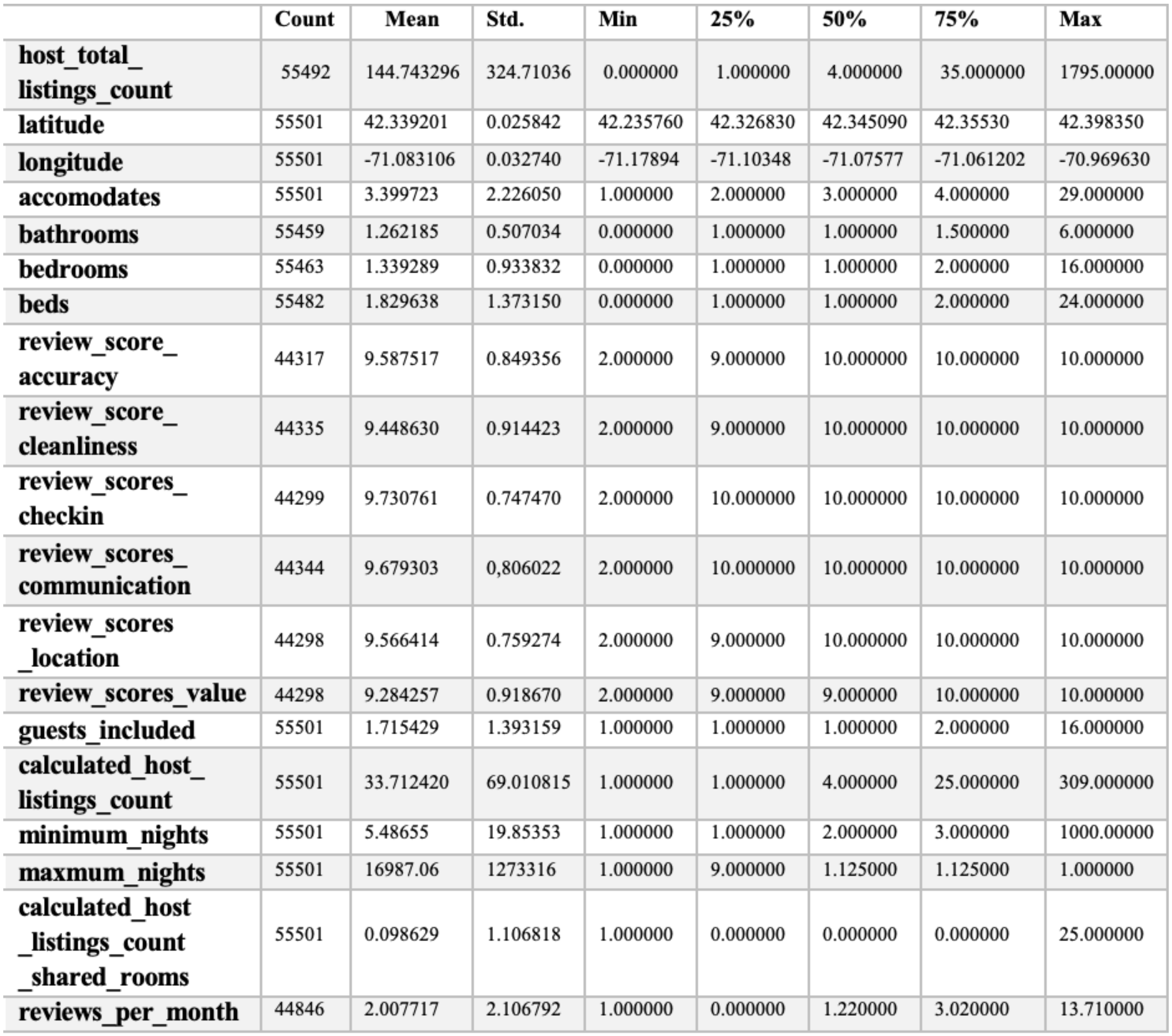
Second, it can be easily observed that on average, the entire home/apt is the most expensive type, with price ranging from 10 to 5000 dollars. On the contrary, the price for shared rooms can be much more affordable, with the highest value around 250 dollars.

Next, we found that within one type of rooms, as the price goes up, the probability of booking an Airbnb with high ratings increases as well. It is possibly because listings with higher price tend to provide better service to satisfy guests.

Besides that, it is also worth noting that listings owned by super hosts are more likely to receive higher scores. In addition, even though their prices vary, generally, the distribution of red dots can be a little bit higher than blue ones. This is especially the case for shared rooms and entire apartments, indicating that super hosts for these 2 types of listings tend to set relatively higher prices.

This plot sheds light on the determinants of price: it is very likely that price is positively correlated with rating scores, room types and the super host title.

1.4 Statistical Distribution of Variables



	Count	Mean	Std.	Min	25%	50%	75%	Max
host_total_ listings_count	55492	144.743296	324.71036	0.000000	1.000000	4.000000	35.000000	1795.00000
latitude	55501	42.339201	0.025842	42.235760	42.326830	42.345090	42.35530	42.398350
longitude	55501	-71.083106	0.032740	-71.17894	-71.10348	-71.07577	-71.061202	-70.969630
accomodates	55501	3.399723	2.226050	1.000000	2.000000	3.000000	4.000000	29.000000
bathrooms	55459	1.262185	0.507034	0.000000	1.000000	1.000000	1.500000	6.000000
bedrooms	55463	1.339289	0.933832	0.000000	1.000000	1.000000	2.000000	16.000000
beds	55482	1.829638	1.373150	0.000000	1.000000	1.000000	2.000000	24.000000
review_score_ accuracy	44317	9.587517	0.849356	2.000000	9.000000	10.000000	10.000000	10.000000
review_score_ cleanliness	44335	9.448630	0.914423	2.000000	9.000000	10.000000	10.000000	10.000000
review_scores_ checkin	44299	9.730761	0.747470	2.000000	10.000000	10.000000	10.000000	10.000000
review_scores_ communication	44344	9.679303	0,806022	2.000000	10.000000	10.000000	10.000000	10.000000
review_scores _location	44298	9.566414	0.759274	2.000000	9.000000	10.000000	10.000000	10.000000
review_scores_value	44298	9.284257	0.918670	2.000000	9.000000	9.000000	10.000000	10.000000
guests_included	55501	1.715429	1.393159	1.000000	1.000000	1.000000	2.000000	16.000000
calculated_host_ listings_count	55501	33.712420	69.010815	1.000000	1.000000	4.000000	25.000000	309.000000
minimum_nights	55501	5.48655	19.85353	1.000000	1.000000	2.000000	3.000000	1000.00000
maxmum_nights	55501	16987.06	1273316	1.000000	9.000000	1.125000	1.125000	1.000000
calculated_host _listings_count _shared_rooms	55501	0.098629	1.106818	1.000000	0.000000	0.000000	0.000000	25.000000
reviews_per_month	44846	2.007717	2.106792	1.000000	0.000000	1.220000	3.020000	13.710000

Fig 1.13 Statistical Distribution of Variables

Analysis:

The listing_details dataset contains 55501 rows and 106 columns.The above three tables shows statistical analysis on some important and relevant numerical variables.This dataset also contains many category-type variables which cannot be perfectly described by statistical analysis but also have a great impact on daily price,such as variables like 'host_is_superhost'.

Our following analysis are mainly based on these important numerical and category-type variables, and we find some interesting insights behind the numbers.

Part II Visualization

2.1 How popular has Airbnb become in Boston?

From Supply Aspect (Listings):

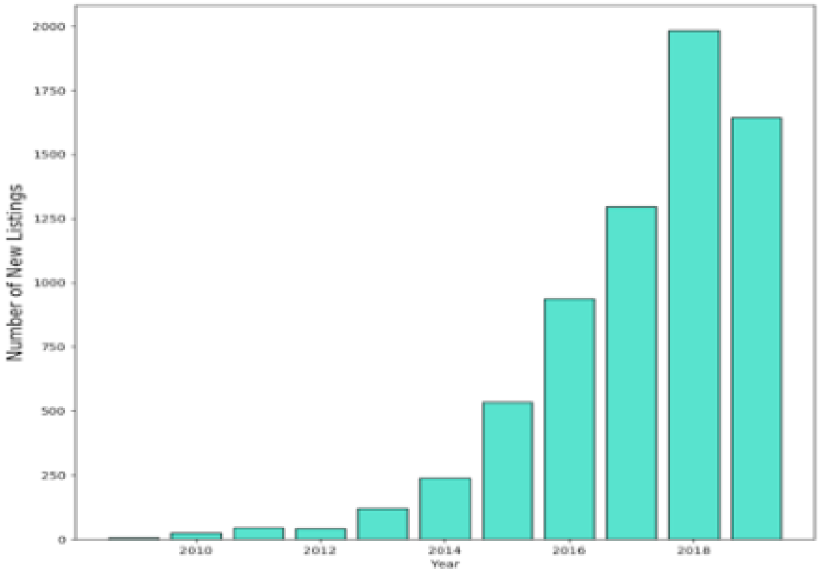


Fig 2.1 Number of New Hosts Each Year

Analysis:

To determine whether there is a significant growth in the Airbnb supply in Boston, 2 columns in *reviews_details.csv* was employed, respectively listing_id and date. Since there's no detailed information about the specific date the place was first listed, we used the year of the first review to estimate the starting year of every listing. By doing so, we were able to calculate the number of new listings per year and closely scrutinize its growth trend.

The bar chart above presents the result. Each green bar shows the figure for yearly new-started rentals, from which we can recognize a boom in the number of newly listed rentals. ***The number of unique listings has experienced an explosive growth since 2009.*** This implies a markedly growth in the supply for Airbnb in Boston market.

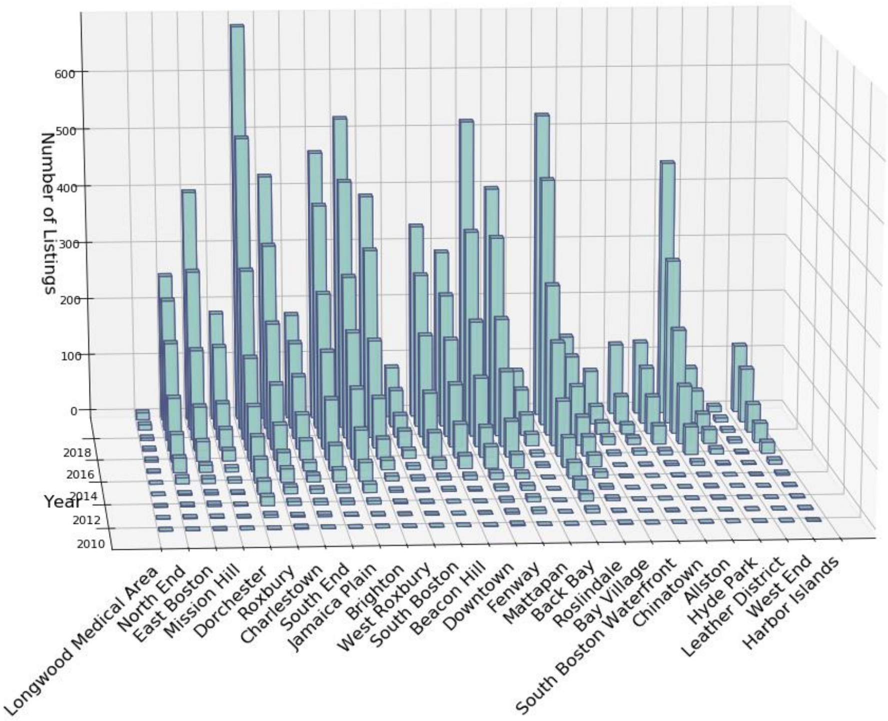


Fig 2.2 Listing Numbers in Different Neighborhoods (3D Scale)

Analysis:

The 3D bar chart suggests each year's listings number in different areas. The overall trend is increasing. To be specific, it turns out that Dorchester, Back Bay and Jamaica Plain saw a marked growth in the supply for Airbnb as early as 2012. The figure for Downtown has quickly catch up since 2015. As for Allston, it wasn't one of the hottest markets for renters until 2016 but has developed dramatically afterwards. The supply of other neighborhoods such as South End, Fenway and Brighton closely followed, with around 600 listings in 2019.

From Demand Aspect(guests):

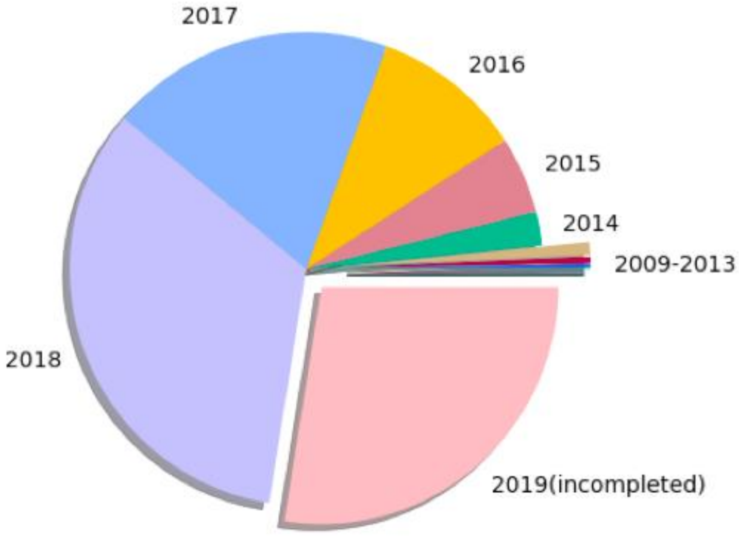


Fig 2.3 Percentage of Reviews Over Time

Analysis:

Apart from the supply analysis, checking the demand is also necessary. In this part, we also used *reviews_details.csv* but from the perspective of guests. Because there are duplicated comments made by the same reviewer, after selecting listing_id, reviewer_id and date from reviews_details, we dropped the duplications. Then the number of reviews each year was computed, which was a rough estimate of how many guests booked Airbnb in that year.

The pie chart compares the figure for different years, suggesting ***a dramatic increase in the number of guests over the years.*** Although the slice of 2019 is slightly smaller than that of 2018, the data of 2019 only contains the first 9 months. To make a better comparison, we extract different years' data for the same season periods to see the changing trend.

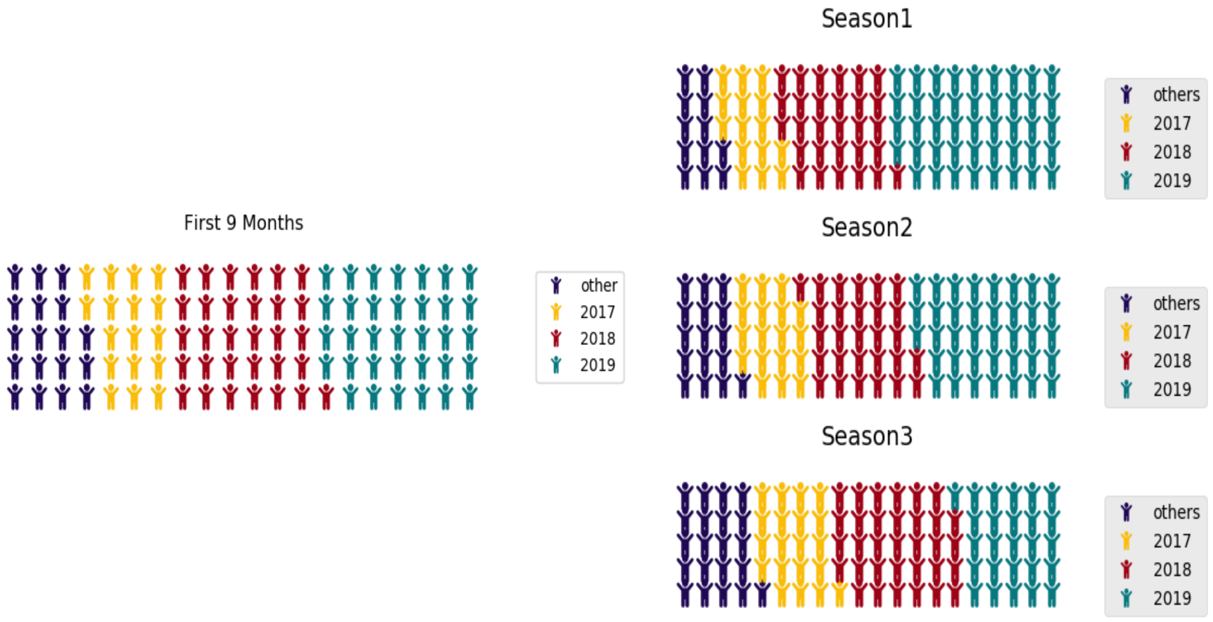


Fig 2.4 Review Numbers Comparison Across Seasons

Analysis:

By comparing the seasons number, we can convince that in overall terms, there is meant to be an obvious rise in the number of guests in 2019. Therefore, the assumption that ***the whole year guest number is increasing across years*** is trustworthy.

But it is still too early to draw the conclusion that more and more people choose Airbnb in Boston. It is possible that a few listings have quit the market, whose reviews and information are not available any longer, thereby leading to a bias when we estimate the number of guests of past years. Thus, in the next step, we need to determine whether the number of reviews received per rentals has generally risen over time. To perform this analysis, we employed joy plot to compare the distribution of how many reviews each listing got across varying years.

Further Demand Analysis:

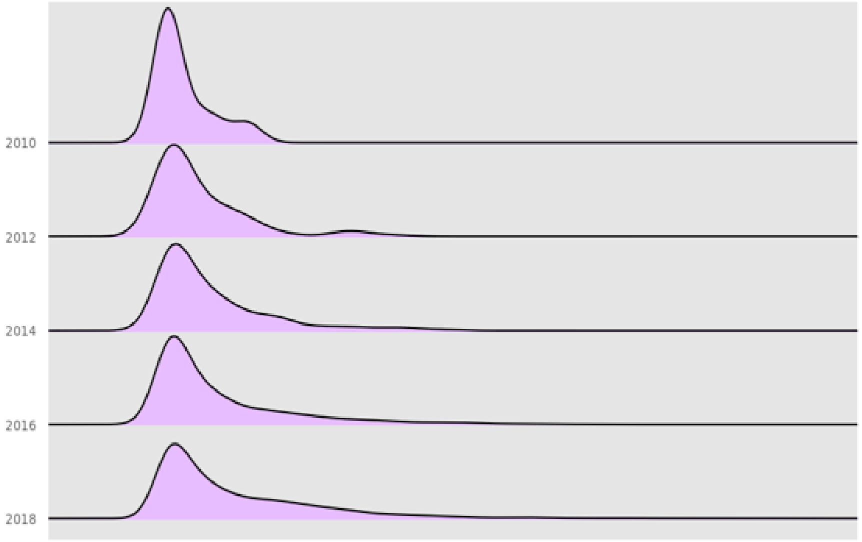


Fig 2.5 Review Numbers Distribution Over Time

Analysis:

There is an obvious trend that as time went by, the distribution has turned out to be flatter and moved to the right (i.e.the probability of getting more reviews has become higher). This phenomenon convinces that, overall, the number of reviews received by each rental has become larger, showing that Airbnb has received more and more customers in Boston market.

This conclusion is still robust when we look at the average number of reviews received by per listing across different areas.

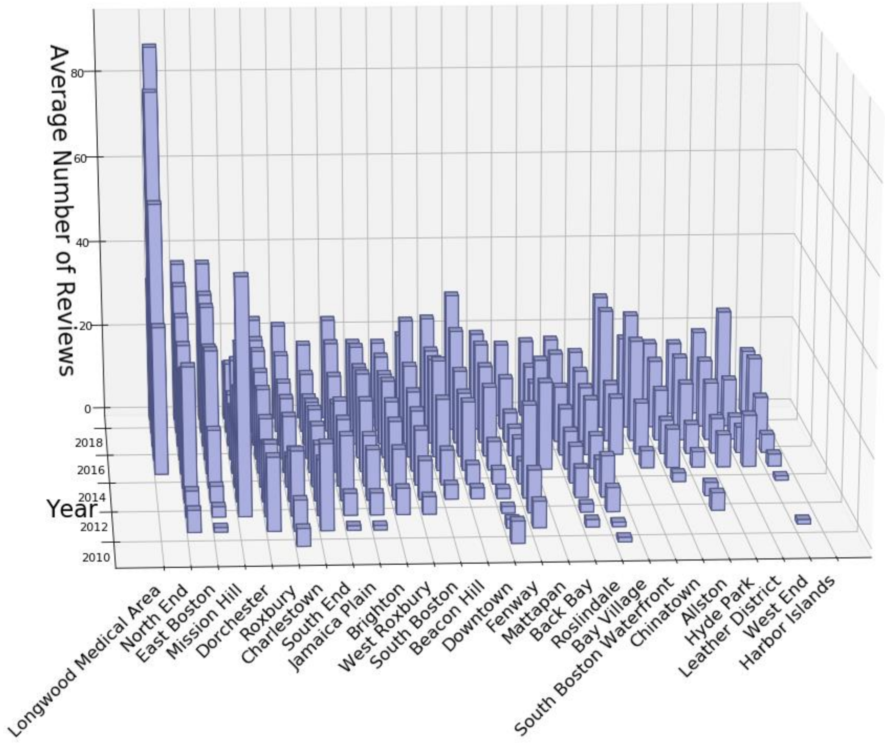


Fig 2.6 Average Number of Reviews (3D Scale)

Analysis:

The 3D bar plot above shows the average number of reviews received per listing in different years, grouped by neighborhood. We can clearly tell that, in general, ***the average figure for most neighborhoods has witnessed an increase over time***, which means on average there has been an increasing demand for Airbnb in different areas of Boston.

Conclusion:

We perform this analysis from both supply and demand side. In both aspects, there has been a remarkable leap. ***We conclude that Airbnb in Boston has been increasingly popular since 2009.*** On the one hand, increasing number of house-owners are willing to transform into a renter, listing their place on the Airbnb website. On the other hand, more travelers regard Airbnb as their premier accommodation provider.

2.2 What geography patterns appear in the Airbnb property listings?

Distribution Map:

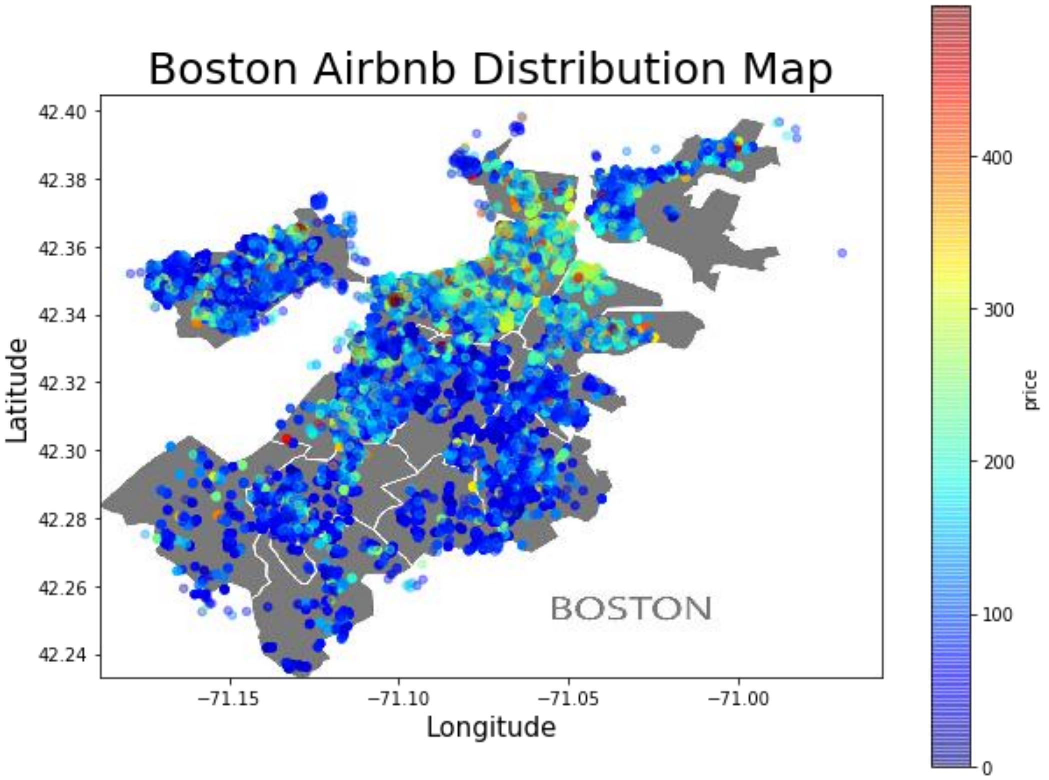


Fig 2.7 Boston Airbnb Distribution Map

Analysis:

We used the data of listings_summary.csv to plot the distribution of Airbnb in Boston with a daily price of no more than 500. Filtering the data whose daily price is more than 500 is because a small part of Airbnb's daily housing prices is more than 500.If all the data is drawn, the display of price differentiation on the graph will be affected by the outliers.

From this graph, we notice that ***Airbnb is densely distributed in Back Bay, Allston and Dorchester.Countrarily, Airbnb is less distributed in the southwest of Boston.*** In general, the denser the distribution, the higher the price, especially in the neighbourhood of Back Bay.

We can see that the locations with ***more airbnb distribution are concentrated in more prosperous areas of Boston, and also some areas with rich tourist attractions.*** This distribution is in accord with the reality that most tourists who come to Boston will choose to live in the center of Boston. Due to the large demand, the number of Airbnb near the center of Boston is larger and prices are higher.

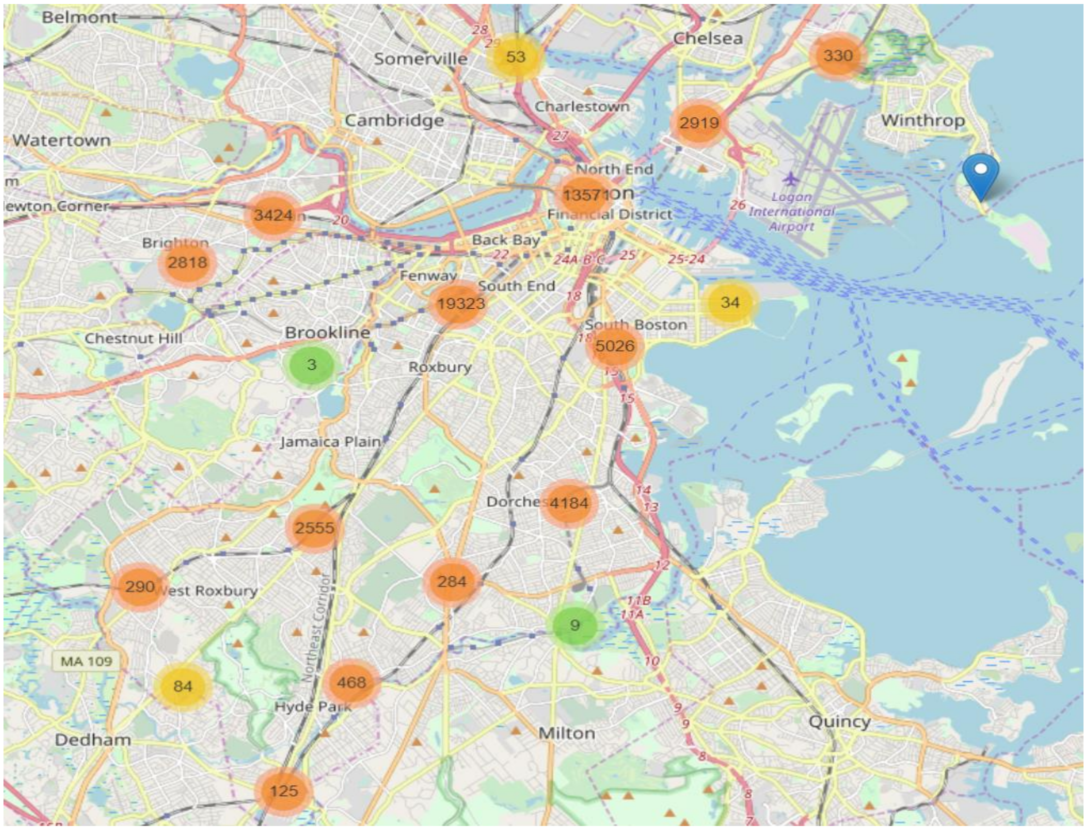


Fig 2.8 Interactive Map of Boston Airbnb Distribution

Analysis:

This is a screenshot of an interactive map we have completed, from which we can notice the distribution of Boston Airbnb in more details. The circles with numbers on the map represent the number of Airbnb in this area.

The conclusion we draw from this map is roughly the same as that of the previous map, but we can know more accurately that most Airbnb located in Fenway, South End and Back Bay. From this interactive map, we can also notice that ***the distribution of Airbnb is closely related to the traffic distribution of Boston.*** There are More Airbnb along the highways and the main roads. We can infer that people prefer to live in places with convenient transportation.

P.S.Interactive map can make the map dynamic change. When you click circles with numbers on the map, this part will be enlarged to show more accurate Airbnb position. However, due to the large size of the Interactive map file, only static screenshots can be displayed in the report. The complete file is Boston_Airbnb_Distribution.html.

Listings Distribution among Neighbourhoods:

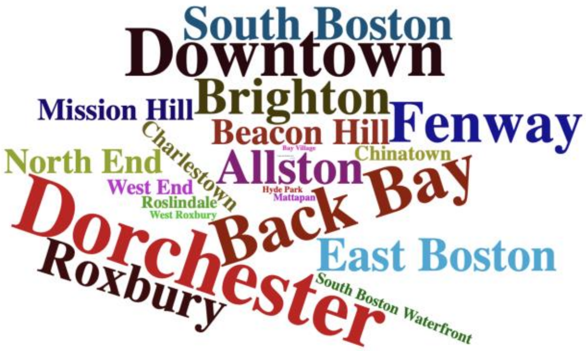


Fig 2.9 WordCloud on Number of Listings

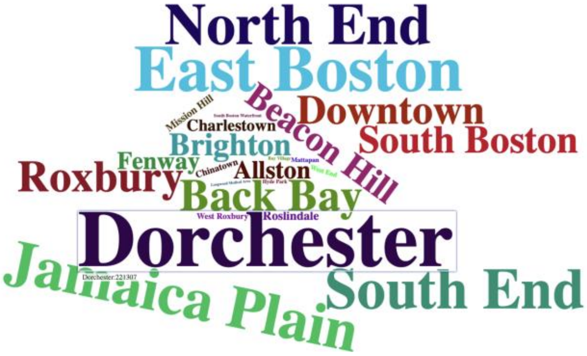


Fig *2.10 WordCloud on Number of Reviews*

Analysis:

These two images geographically suggest the distribution of Airbnb in Boston neighbourhoods and they are also interactive, html files are attached.

The first WordCloud explains geography listing pattern of Airbnb by showing different sizes of words. The bigger size of the word, the more listings are located around the neighborhood. From this graph, it is obvious that ***Dorchester area*** has the **most listings,** ***Downtown*** and ***Back Bay*** **are not far behind**. From the copy file (wordcloud_listings_id.html) we can easily see that there are 4357 listings in Dorchester area. Longwood Medical Area has the least which is only 101 listings.

The second WordCloud concentrates **on the number of reviews**, which shows the area people tend to choose their accommodation. It's clear that Dorchester still on the top followed by Eastern Boston and Jamaica Plain, while Longwood Medical Area still has the least reviews.

Conclusion:

From the geography pattern analysis, we can see that in Boston, ***Airbnb is densely distributed in the areas where are more prosperous, own rich tourist attractions and have convenient transportation.*** Such areas including Dorchester, Back Bay, Downtown and Allston.

2.3 Which neighborhood is the most popular among customers?

Neighbourhood Listing Numbers:

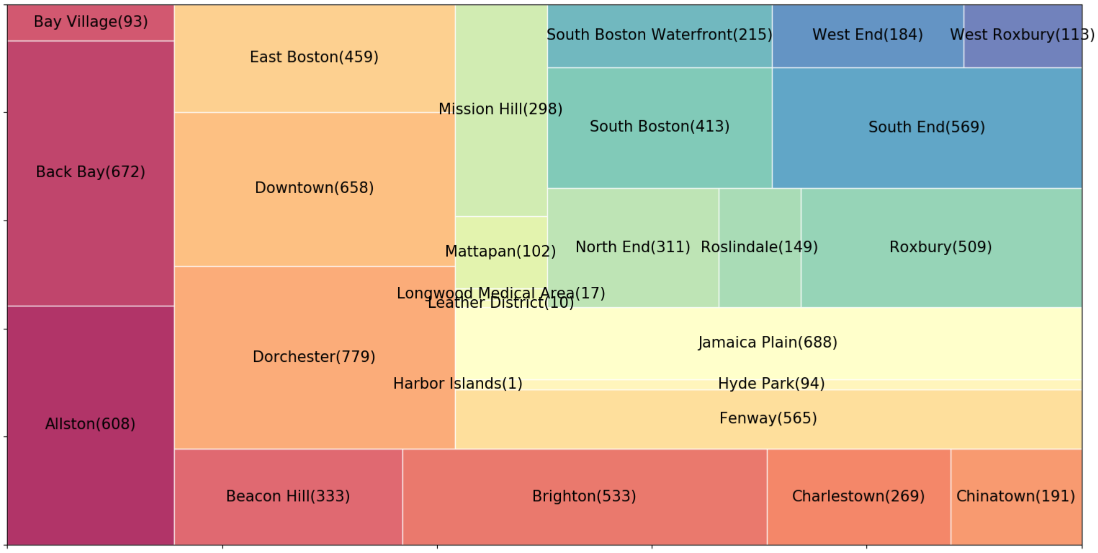


Fig 2.11 Number of Airbnb in Different Neighborhoods

Analysis:

Firstly, we regarded the number of unique listings as the indicator of supply. By simply counting the distinct rentals in each neighborhood, we plotted the graph above. It tells that Dorchester, Jamaica Plain, Back Bay, Downtown and Allston are top 5 largest “Airbnb providers” among 26 neighborhoods in Boston areas. At the same time, Leather District and Longwood Medical Area provide less, with number of listings below 20. It should be mentioned that market size in Harbor Island is close to 0, only having 1 room listed on the website.

Customer Review Numbers:

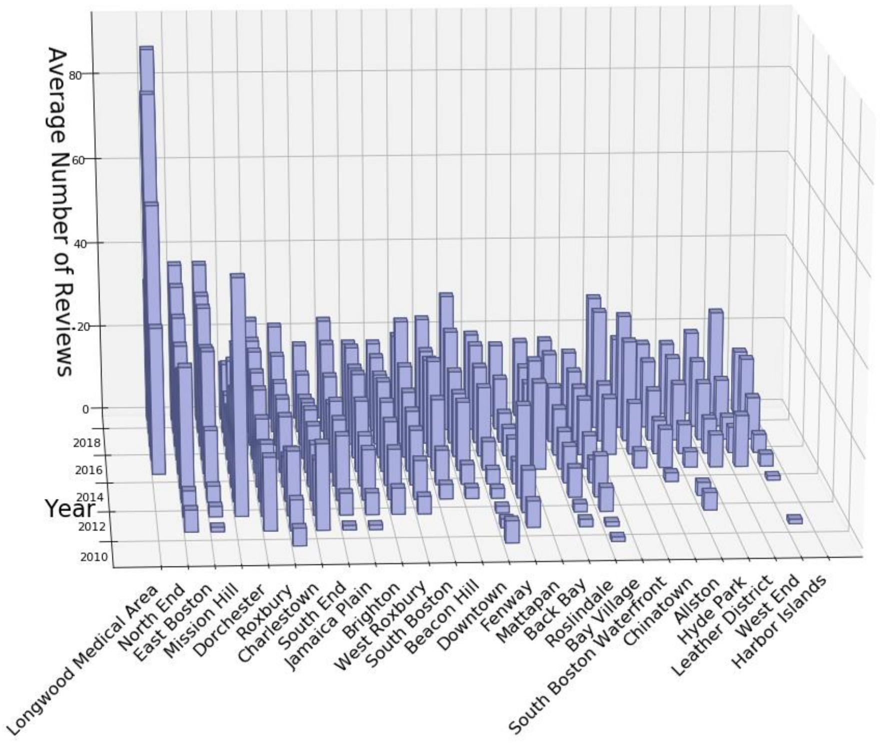


Fig 2.12 Average Number of Reviews (3D Scale)

Analysis:

We went back to the 3D bar chart shown before to discover the popularity (average review numbers) per year in each area.

We found that even though there has only been a slow climb in the supply of ***longwood medical area,*** the average number of guests per listing in this region can be far larger than we expected. It seems to indicate that the demand in this neighborhood has outstripped the supply. It is a similar story when it comes to ***North End*** and ***East Boston***: ***there might be a potential market in these areas waiting to be explored.***

On the other hand, the figure for the neighborhoods with more supply remains stable in recent years. It is possible that the market for these areas is about to be saturated.

Customer Satisfaction (Review Scores):

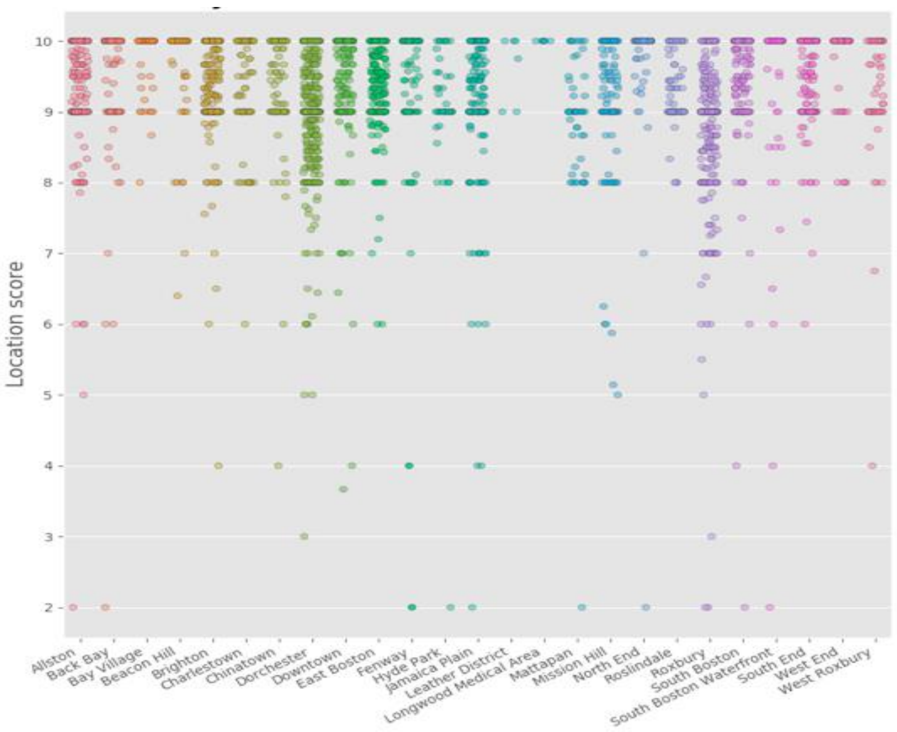


Fig 2.13 Jitter Plot of Location Score

Analysis:

From the jitter plot of the location score, which reflected guests' satisfaction towards the neighborhood. We can see that, among the top 5 neighborhoods providing most Airbnb accommodation, ***the majority of location scores received by Allston, Jamaica Plain, downtown and back bay are distributed within 9 and 10***. The median value for the four areas are higher than 9.5, implying that generally speaking, guests were highly satisfied at these regions.

The scores for Dorchester can be much more separated, with not a few points located between 8 and 9, resulting in a relatively low median score.

North End, East Boston and Longwood Medical Area have also received guests' high satisfaction, while guests in Roxbury were not that pleased with the location.

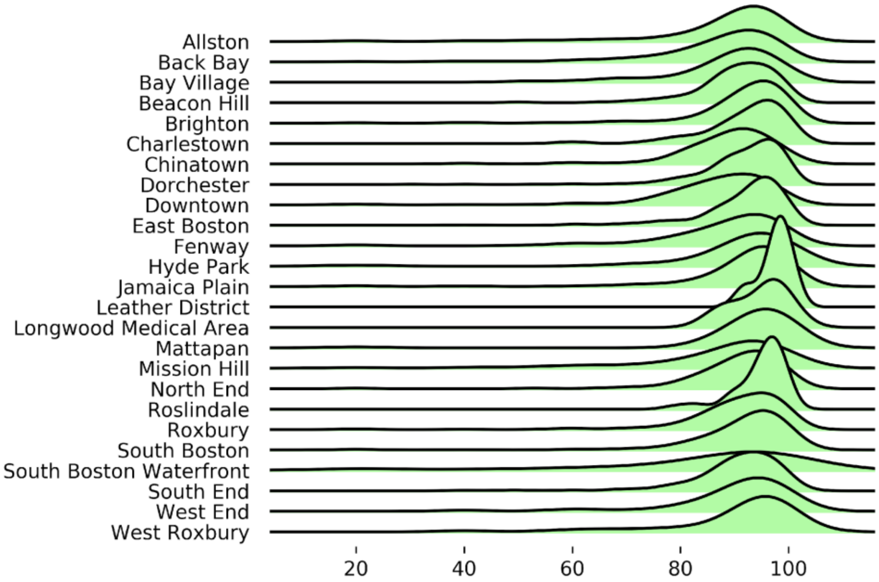


Fig 2.14 Jitter Plot of Review Scores Rating by Neighbourhood

Analysis:

As for the total scores listings received in each neighborhood, things can be quite different. ***Dorchester and Jamaica Plain have a relatively concentrated distribution, having a higher possibility of receiving high scores.*** The distribution of Downtown, Allston and Back Bay can be flat, having a thick left tail, suggesting that they received quite a lot of negative feedback.

East Boston and Longwood medical area did well in terms of the total rating score while rentals in South End failed to receive very high scores on average.

Conclusion:

We applied review size and rating scores as main indicators to estimate the popularity of different neighborhoods among customers. Dorchester, Jamaica Plain, Back Bay, Downtown and Allston are top 5 largest “Airbnb providers” among 26 neighborhoods in Boston areas but the market in these areas are about to be saturated. Among these regions, listings in Jamaica Plain enjoyed both high location scores and high total ratings in general, while listings in other 4 neighborhoods failed to do well in both sides in overall terms.

Besides that, Longwood Medical Area, East Boston and South End are more popular than we expected from the perspective of demand, although their supply size are not so significant. There might be quite a big potential market in the 3 neighborhoods.

Part III Data Mining

3.1 Overall Correlation

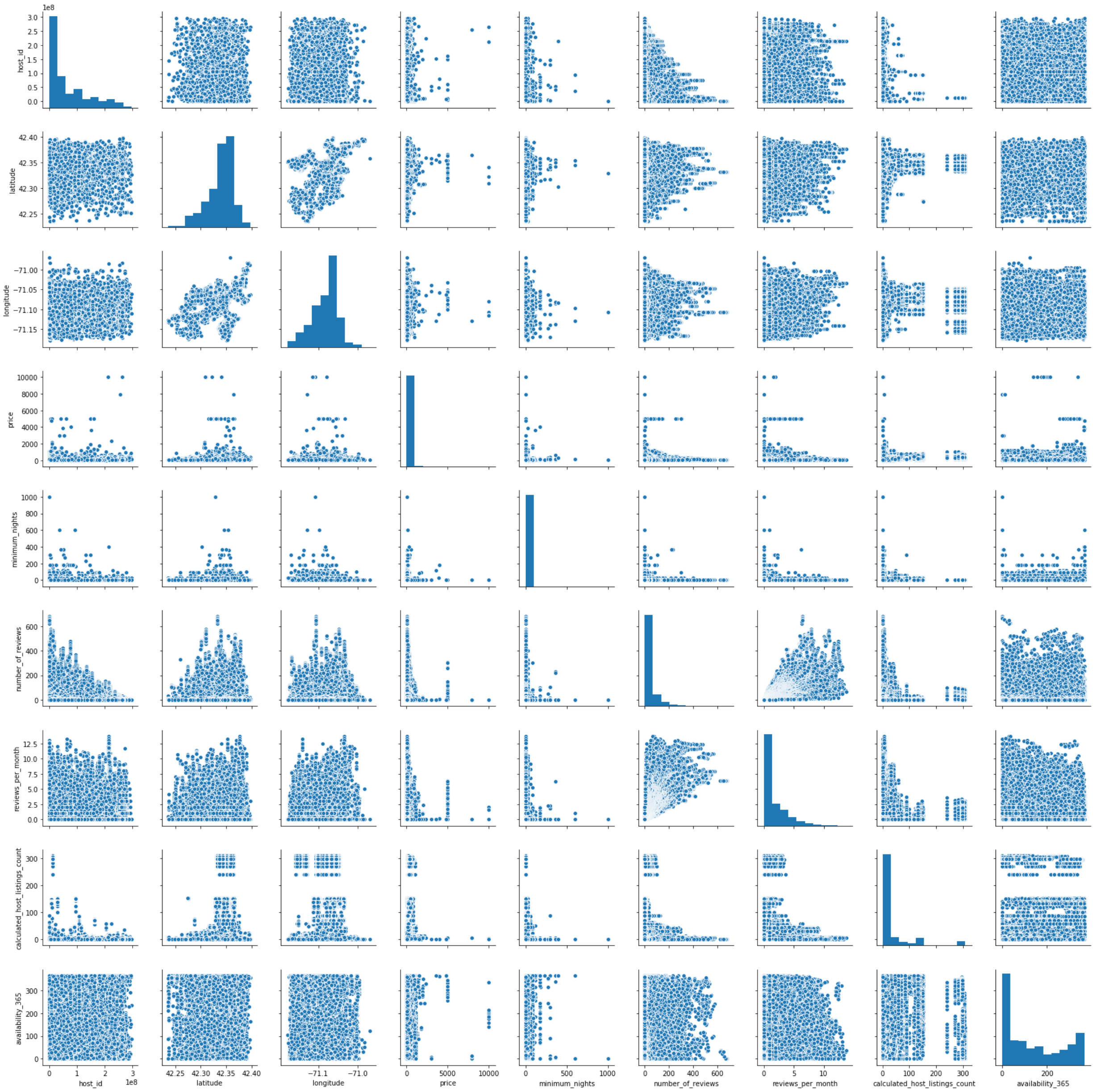


Fig 3.1 Glance of Correlations

We initially clean the data of 'listings_details.csv' and select important variables to see the correlations among them.From the glance of variable relationships, we further build up models to interpret price determinant variables and make pricing prediction.

3.2 Multivariate Regression Analysis

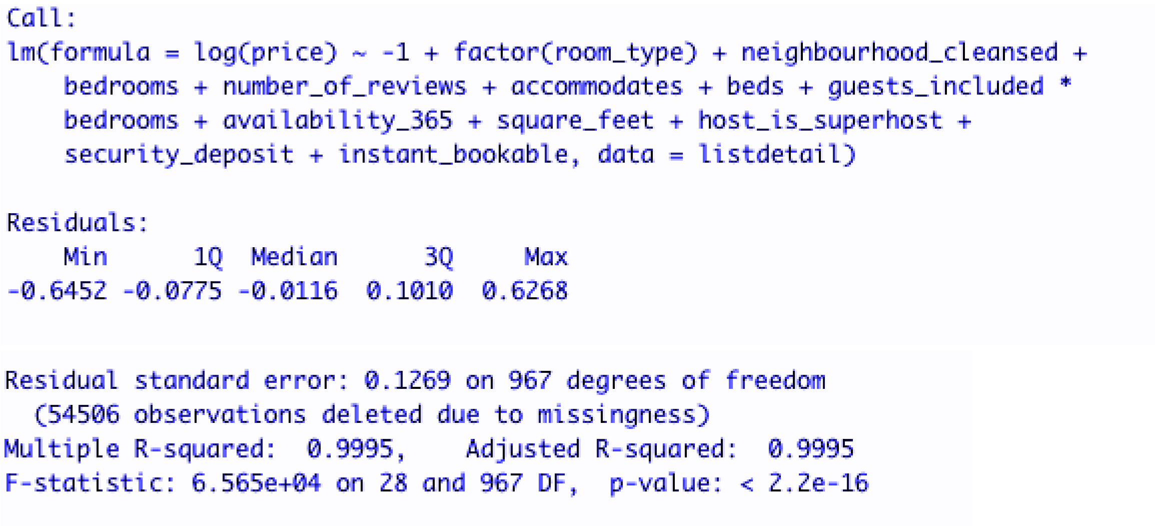
Model Selection:

Target Variable: Daily Housing Price

Step1. Prepare and clean the data
Use ‘listings_details.csv’, which has the most variables. From the previous correlation analysis, we select potential regressors, and change the type of some variables to prepare for regression.

Step2. Use forward-and- backward selection to run regression models
a. Add statistically significant variables, run linear regression.
b. Change the regression to Log-Linear, and find out the adjusted R squared improved by 6%
c. Add factor variables (guests_included : bedrooms)
d. Run log-linear regression with interactions

Result:



```
Call:
lm(formula = log(price) ~ -1 + factor(room_type) + neighbourhood_cleansed +
    bedrooms + number_of_reviews + accommodates + beds + guests_included *
    bedrooms + availability_365 + square_feet + host_is_superhost +
    security_deposit + instant_bookable, data = listdetail)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6452 -0.0775 -0.0116  0.1010  0.6268

Residual standard error: 0.1269 on 967 degrees of freedom
  (54506 observations deleted due to missingness)
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9995
F-statistic: 6.565e+04 on 28 and 967 DF,  p-value: < 2.2e-16
```

Fig 3.2 Results of Regression Model

Model Evaluation:

When we conduct linear regression for the data in r, it automatically deleted all the missingness (observations with NAs), in our model, it deletes 54506 observations.

With the complete observations, our model show that the adjusted R squared is 0.9995, which means the model can explain 99.95% of the data, and almost all the variables are statistically significant at the 5% level.

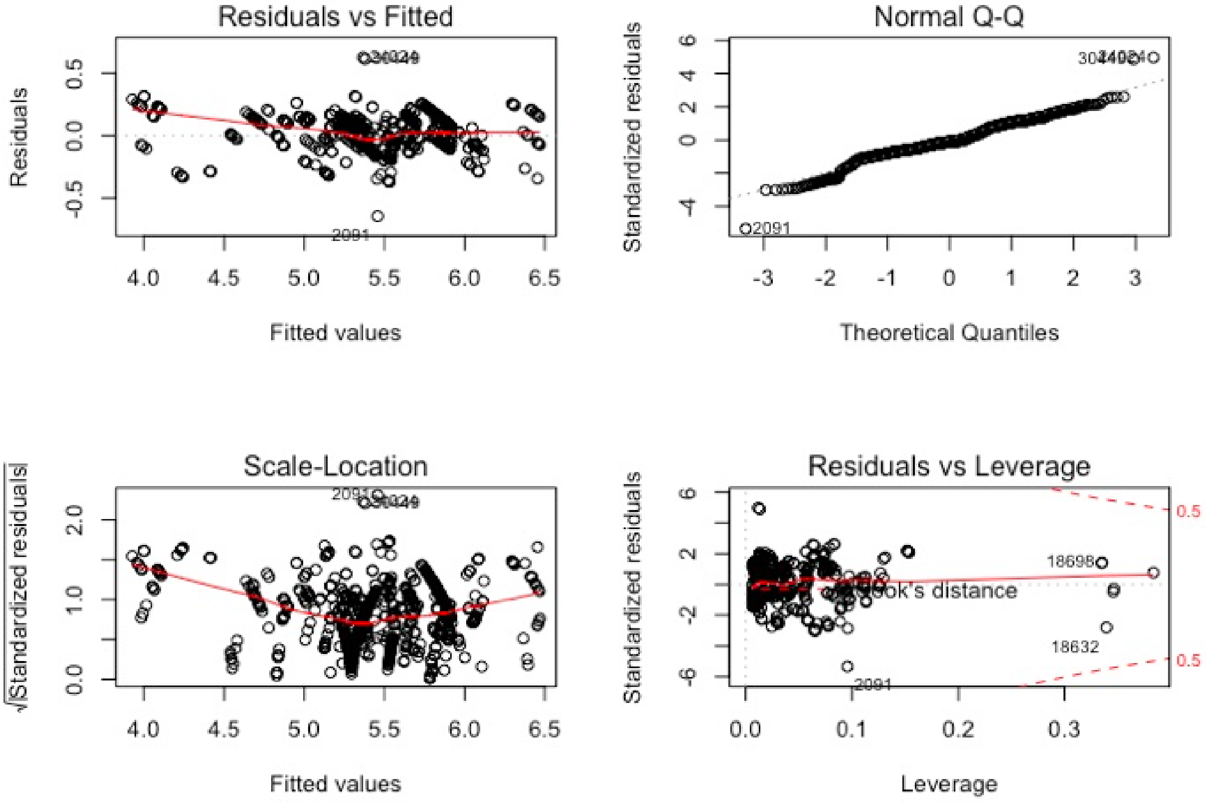


Fig 3.3 Residual Plots

Analysis:

From the plots, we can see four different type of residual information, to better evaluate the model, we make further analysis of the plots.

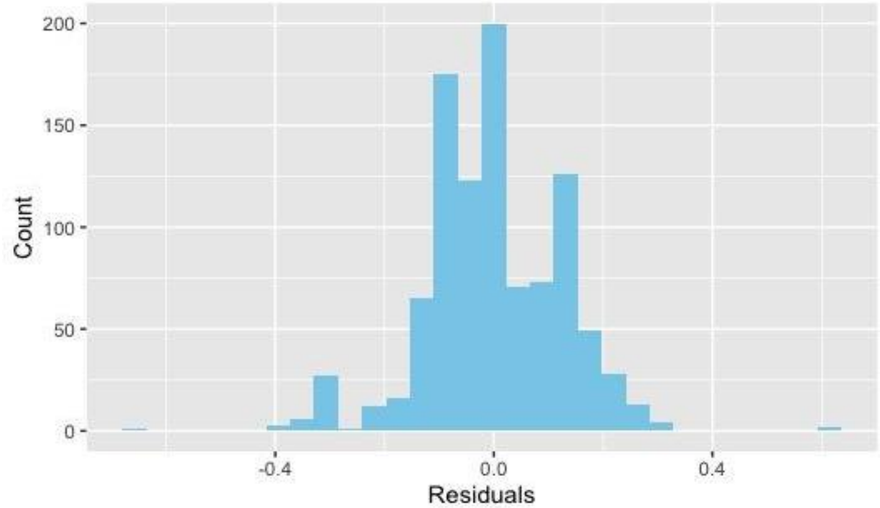


Fig 3.4 Histogram for Residuals

Analysis:

The residuals look Gaussian, a bell- shaped curve, most of them are around 0.

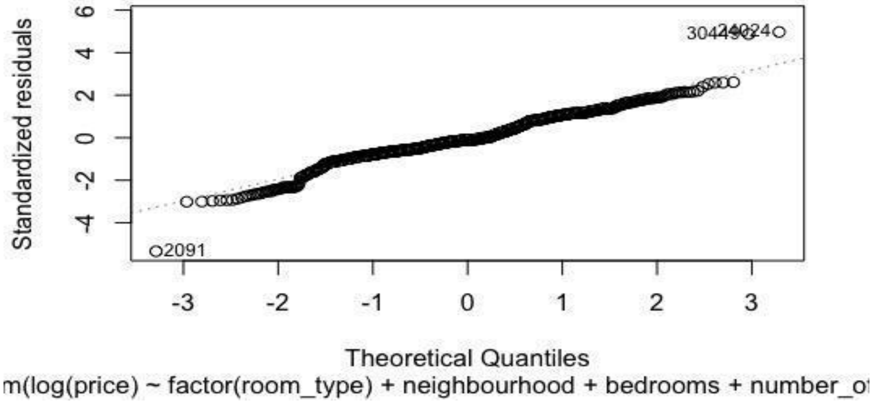


Fig 3.5 Normal Probability Plot of Residual

Analysis:

The normal qq plot helps us determine if our residuals are normally distributed by plotting quantiles (i.e. percentiles) from our distribution against a theoretical distribution.

For our plot, the residuals closely track diagonal line and indicates normal distribution.

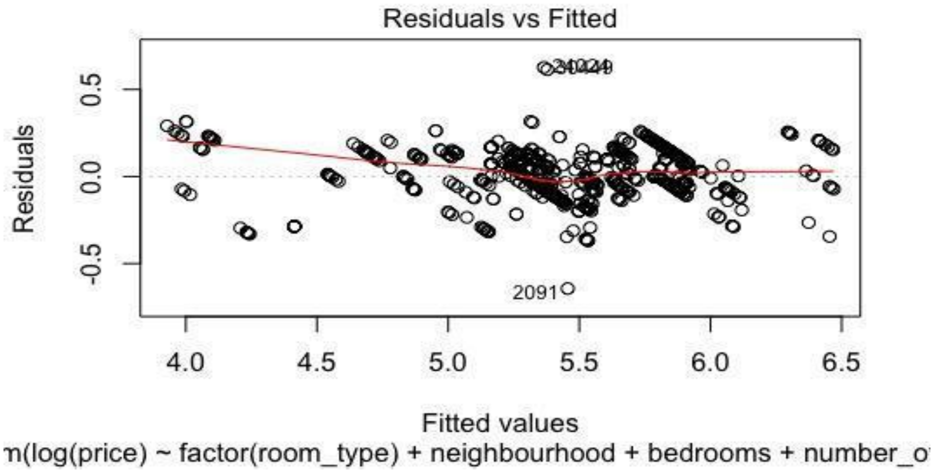


Fig 3.6 Residuals vs. Fitted Values

Analysis:

This plot tests the assumptions of whether the model is linearity and homoscedasticity. For our plot, it's relatively shapeless without clear patterns in the data. There is no obvious outliers, and be generally symmetrically distributed around the zero line without particularly large residuals.

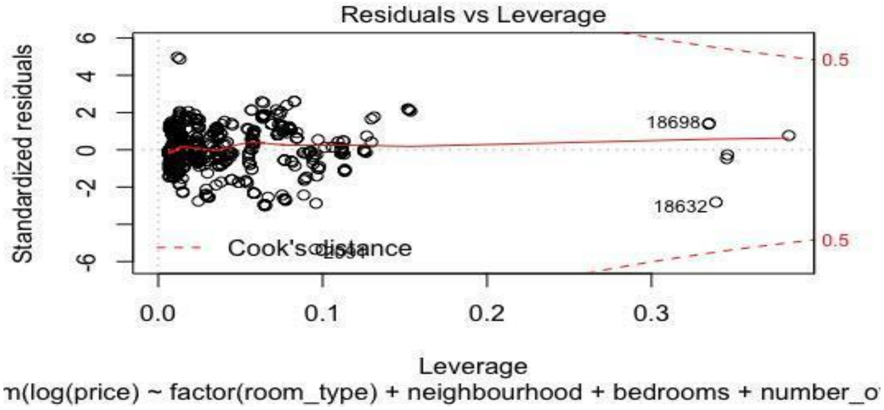


Fig 3.7 Residuals vs. Leverage

Analysis:

The Residuals vs. Leverage plots helps to identify influential data points on the model. The points we're looking for (or not looking for) are values in the upper right or lower right corners, which are outside the red dashed Cook's distance line. Our plot doesn't show any influential cases as all of the cases are within the the dashed Cook's distance line, so there is no influential outliers need to be evaluated.

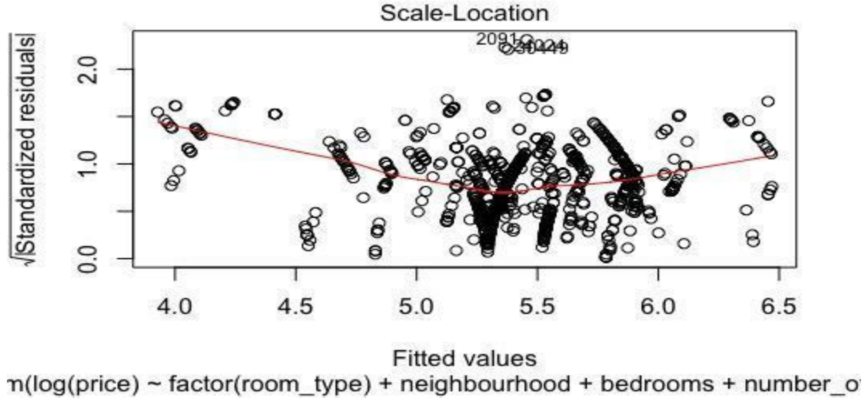


Fig 3.8 Scale-Location Plot

Analysis:

The Scale-Location plot shows whether our residuals are spread equally along the predictor range, i.e. homoscedastic. For our plot, our line starts off horizontal at the beginning of our predictor range, slopes down until reaching 5.3, and then slopes up. In this case, our data is somewhat heteroscedastic. So, we run the heteroscedastic test and make the results more reliable.

Model Interpretation:

Variables(other than factor variables)	Coefficient	Significance Level	Interpretation (given other variables constant)
accommodates	0.083	0%	1 more accommodate, 0.083 increase in daily price
beds	0.133	0%	1 more bed, 0.133 increase in daily price
bedrooms	0.181	0%	Given the bedrooms, 1 more guest included, 0.055 increase in daily price; given guest included, 1 more bedroom, 0.102 increase in daily price
guests_included	0.084	0%	
bedrooms*guests_included	-0.029	0%	
number_of_review	-0.003	0%	Negative relationship between review number and price
availability_365	0.0003	0%	Positive relationship between available days and price
square_feet	-0.0001	5%	Negative relationship between square feet and price
host_superhost	0.198	0%	When the host is super host, the daily price will be 0.198 higher.
security_deposit	0.0001	5%	Positive relationship between security_deposit and price
instant_bookable	0.129	0%	When it is instant bookable, the daily price will be 0.129 higher

Fig 3.9 Variables (other than factor variables) Interpretation

Room_type	Coefficient	Significance Level	Interpretation (given other variables constant)
Entire home/apt	4.509	0%	Among the room types, entire home/apt has the highest daily price, while private room is the lowest.
Hotel room	4.489	0%	
Private room	4.470	0%	

Fig 3.10 Room Type Variables Interpretation

Neighbourhood	Coefficients	Significance Level	Interpretation (given other variables constant)
Beacon Hill	-0.093	Not significant	From the available neighbourhood observations, we can see that East Boston has the highest daily price, West Roxbury has the lowest one. When order the prices of neighborhoods, the top 5 highest are East Boston, South Boston Waterfront, Brighton, South End and South Boston
Brighton	0.265	0%	
Dorchester	0.088	0%	
Downtown	0.039	1%	
East Boston	0.379	0%	
Jamaica Plain	-0.477	0%	
Mission Hill	-0.458	0%	
North End	-0.148	10%	
Roslindale	-0.885	0%	
Roxbury	-0.155	5%	
South Boston	0.225	1%	
South Boston Waterfront	0.295	0%	
South End	0.235	0%	
West Roxbury	-1.13	0%	

Fig 3.11 Neighbourhood Variables Interpretation

Concede:

For our model, many observations are deleted because of variables missingness, so the data is not perfect complete, with more thorough data, the model can be more reliable.

It is possible some other variables still have influence on daily price, such as season, regulation policy, industry trend and etc. We can improve the model further with more information.

There may be some simultaneous causality bias exists, since price may also have effect on some variables, such as review numbers.

Conclusion:

From the model, we can see some variables have positive relationships with daily price of Airbnb:bedrooms, beds, accommodates, guests_included, the more the higher. While review numbers are slightly negatively influenced price. Although availability, square feet and security deposit are correlated with price, the effect is small.

There are two interesting variables that have significant positive relationship with price, which are ***host_is_super host*** and ***instant_bookable***. When the answer is true, the daily price tends to be higher. We can conclude that to be more strongly priced, the service itself is really important.

3.3 Machine Learning

Import Data :

We use listings_details.cvs to do machine learning part. Our **target variable is price**, we need to use existing features to predict **daily prices of Airbnb** in Boston accurately.

After checking the shape of this dataset, we found that this dataset has 55501 rows and 106 columns. There are many features, but not all of them are useful for us to predict prices. Thus, data cleaning is very important.

Clean Data:

Remove “$” sign and “%” sign

We found that price_relatived features (“price”, “weekly_price”, “monthly_price”, “security_deposit” and “cleaning_fee”) have the dollar sign “$” before the number. Therefore, we need to get rid of the dollar sign “$” in order to do statistics. Besides, we need to remove “%” sign for the feature “host_response_rate”.

Remove Outliers:

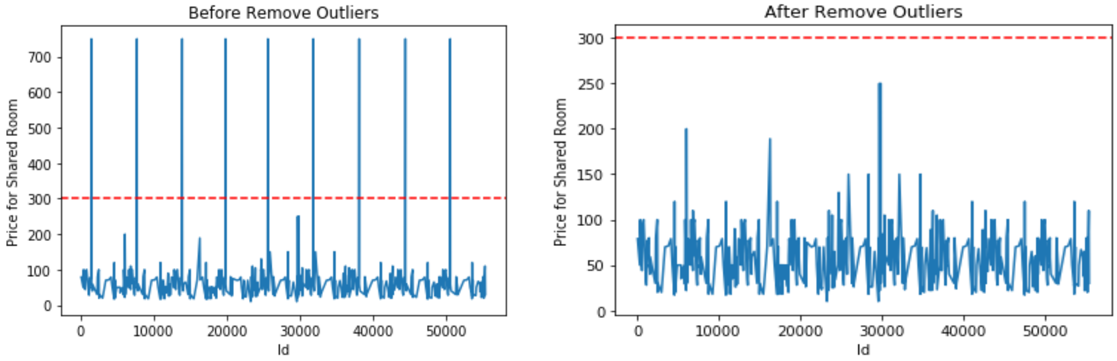


Fig 3.12 Before and After Outliers Removal (shared room)

When we checked the daily price for shared room, we found 9 shared rooms have the daily price $750.This is incredible! We think they are outliers, and we need to remove them.

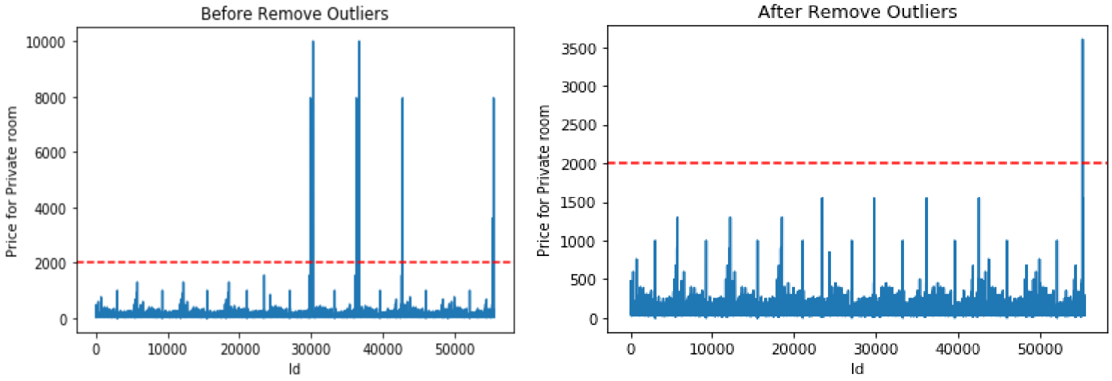


Fig 3.13 Before and After Outliers Removal (private room)

When we checked the daily price for private room, we found 8 shared rooms have the daily price $10000.It is obvious that they are outliers and we need to remove them.

Finally, we noticed some prices are equal to 0, which is irregular. Therefore, we remove them.

Feature Engineering:

Check missing values:

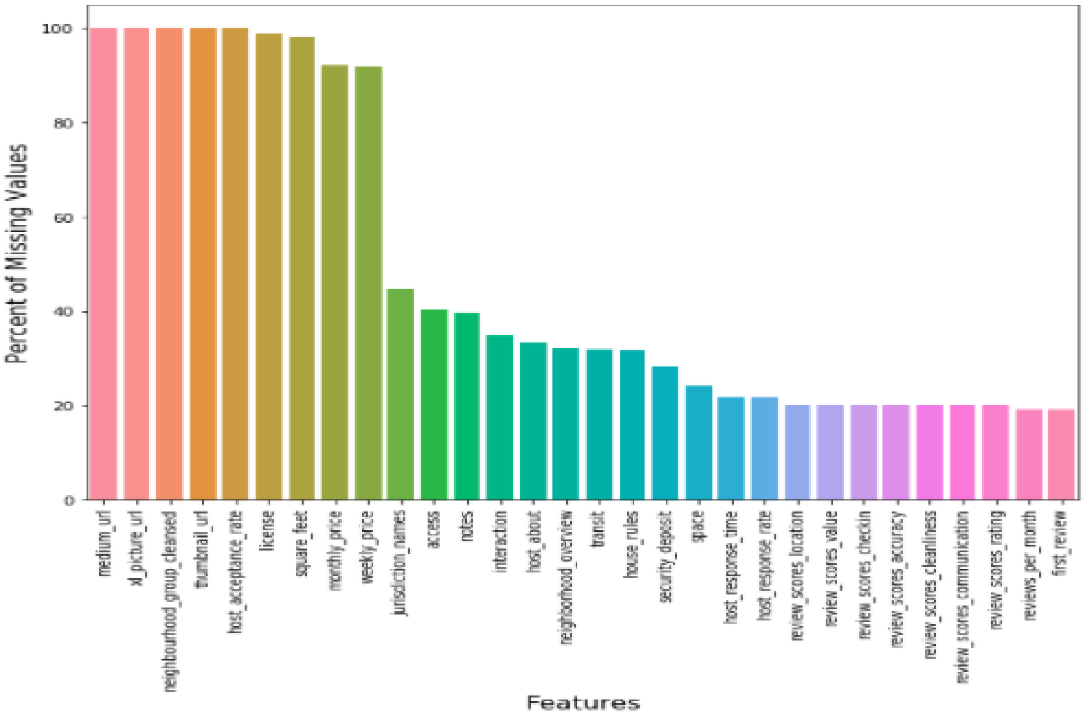


Fig 3.14 Percent of Missing Data by Features

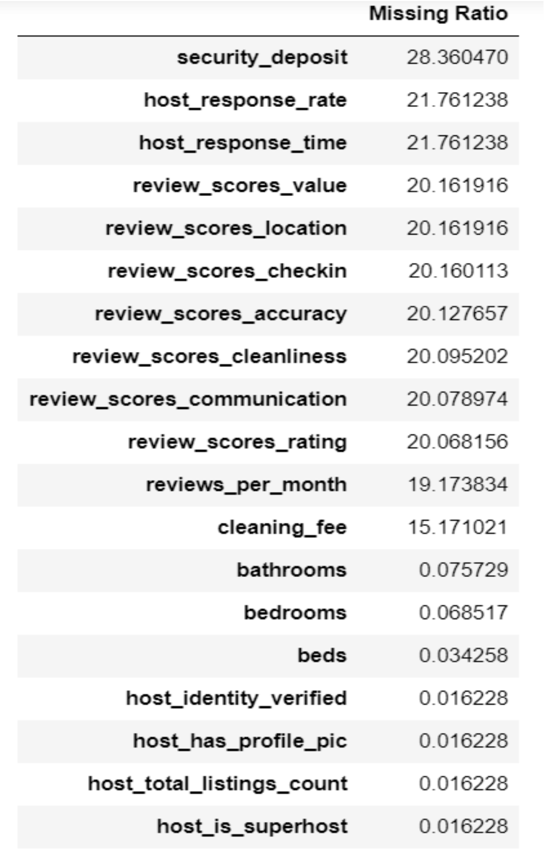
From the above graph, we noticed that some features have 100% of missing values.For these features, we just drop them directly. In addition, some features such as “license”, “square_feet”, “monthly_price” and “weekly_price” have more than 90 percent of missing values. We also drop them directly.

Drop useless features:

1) Drop unimportant variables
"listing_url" "scrape_id" "last_scraped" "name" "thumbnail_url" “Id” etc.
2) Drop variables which have many missing values
"host_acceptance_rate" "license" "weekly_price" "monthly_price" "weekly_price" etc.
3) Drop variables which are all same
"experiences_offered" "is_business_travel_ready"
4) Drop some character description
"summary" "space" "description" "neighborhood_overview" "notes" "transit" etc.

After dropping, we just have 55 features to predict prices.

Imputing missing values:



	Missing Ratio
security_deposit	28.360470
host_response_rate	21.761238
host_response_time	21.761238
review_scores_value	20.161916
review_scores_location	20.161916
review_scores_checkin	20.160113
review_scores_accuracy	20.127657
review_scores_cleanliness	20.095202
review_scores_communication	20.078974
review_scores_rating	20.068156
reviews_per_month	19.173834
cleaning_fee	15.171021
bathrooms	0.075729
bedrooms	0.068517
beds	0.034258
host_identity_verified	0.016228
host_has_profile_pic	0.016228
host_total_listings_count	0.016228
host_is_superhost	0.016228

After dropping some features, we check the missing ratio for our left features and noticed that we still have many missing values. Thus, the next step is to impute missing values.

#security_deposit: use mean security_deposit to fill Na
#host_response_rate: use mean host_response_rate to fill Na
#host_response_time: use most often appeared host_response_time to fill Na
#review_scores_checkin: use mean review_scores_checkin to fill Na
#review_scores_accuracy: use mean review_scores_accuracy to fill Na
#review_scores_cleanliness: use mean review_scores_cleanliness to fill Na
#review_scores_rating: use mean review_scores_rating to fill Na
#review_scores_location: Group by neighborhood and fill in missing value by the median review_scores_location of all the neighborhood
#reviews_per_month: use 0 to fill Na
#cleaning_fee: use mean cleaning_fee to fill Na
#bathrooms: use most often appeared bathrooms to fill Na
#bedrooms: use most often appeared bedrooms to fill Na
#beds: use most often appeared beds to fill Na
#host_identity_verified: use "f" to fill Na
#host_has_profile_pic: use "f" to fill Na
#host_total_listings_count: use most often appeared host_total_listings_count to fill Na
#host_is_superhost: use most "f" to fill Na

Now, we don’t have missing values!!

Data Correlation:

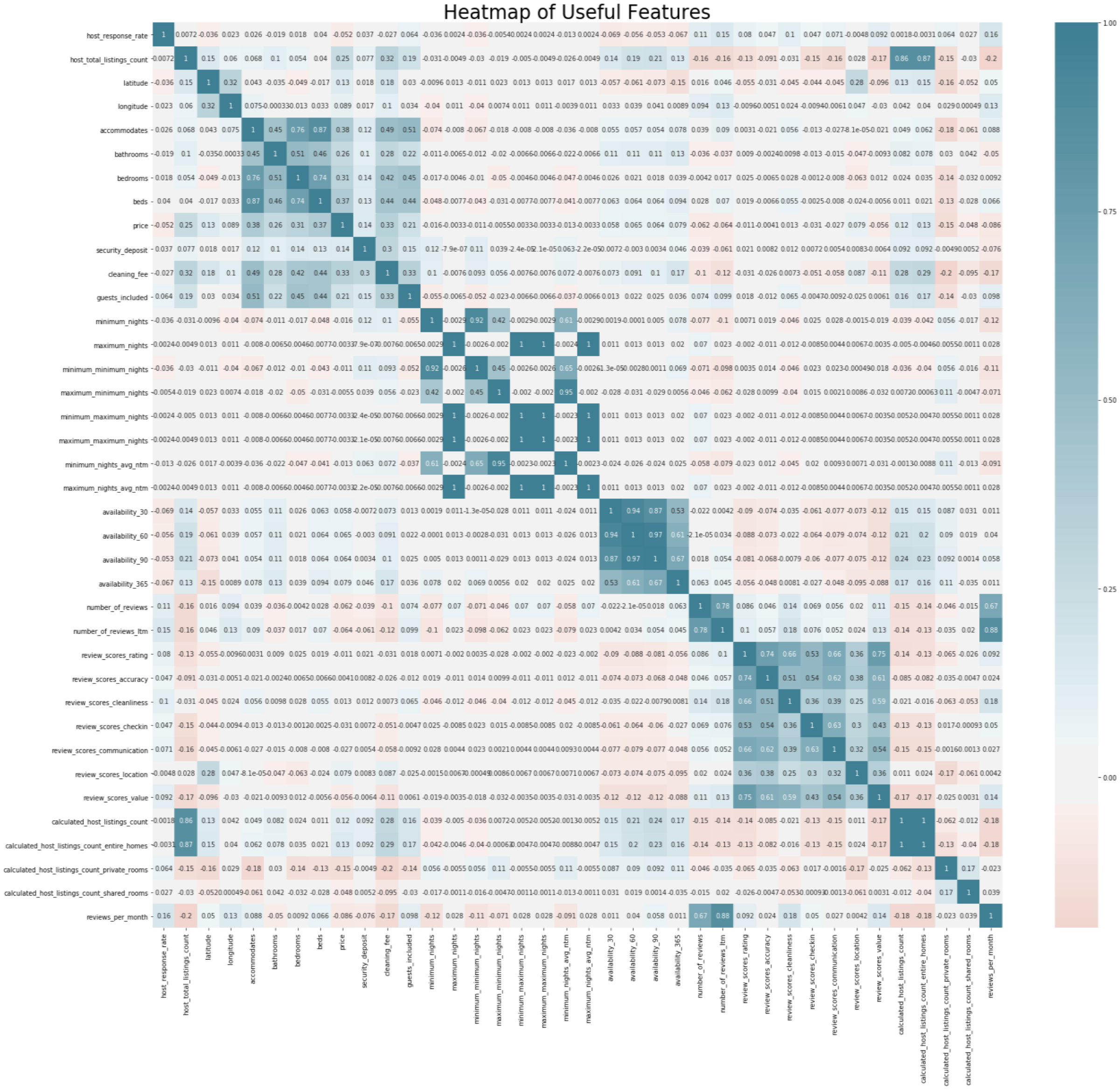


Fig 3.15 Heatmap of Useful Features

More Features Engineering:

Labeling some categorical variables that may contain information in their ordering set should not be ignored. It is very important to apply LabelEncoder to categorical features.

Adding one more important feature:

Since guest_include and the number of bedrooms are closely related with each other,we add one more feature which is guest_include*bedrooms.

Transform Target Variable:

Price is the variable we need to predict. Therefore, let's do some analysis on this variable first.

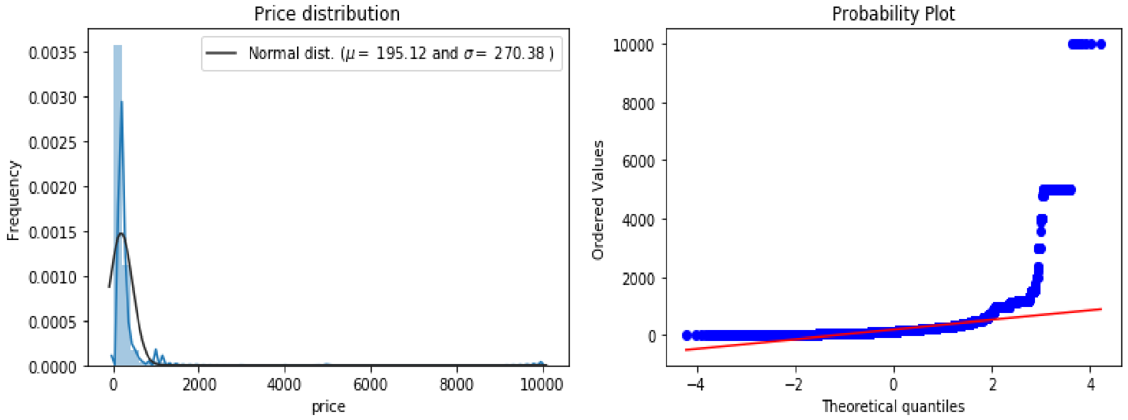


Fig 3.16 Price Distribution and Probability Plot before Transformation

The target variable is right skewed. As (linear) models love normally distributed data, we need to transform this variable and make it more normally distributed.

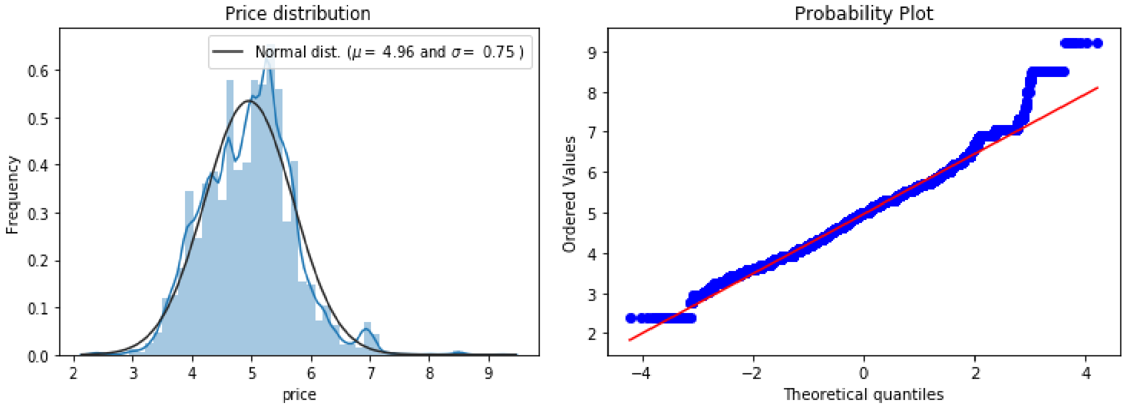


Fig 3.17 Price Distribution and Probability Plot after Transformation

Modelling:

We tried to use different models to predict prices and compare their accuracies.The following is the model we used and their different RMSE.
P.S.For detailed model parameters and codes, please refer to the jupyter notebook file.

RandomForestRegressor:

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
          max_features=None, max_leaf_nodes=None,
          min_impurity_decrease=0.0, min_impurity_split=None,
          min_samples_leaf=1, min_samples_split=2,
          min_weight_fraction_leaf=0.0, n_estimators=200, n_jobs=1,
          oob_score=False, random_state=42, verbose=0, warm_start=False)
```

```
Score for X_train, y_train:0.9953298975624898
Score for X_test, y_test:0.9672106650135358
Root Mean Square Error for test = 0.13564977811328668
```

GBoost:

```
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
            learning_rate=0.05, loss='huber', max_depth=4,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=5, min_samples_split=5,
            min_weight_fraction_leaf=0.0, n_estimators=4000,
            presort='auto', random_state=5, subsample=1.0, verbose=0,
            warm_start=False)
```

```
Score for X_train, y_train:0.9449390847826921
Score for X_test, y_test:0.9199899331803699
Root Mean Square Error for test = 0.15179516637736007
```

XGBoost:

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
      colsample_bynode=1, colsample_bytree=0.2, gamma=0,
      importance_type='gain', learning_rate=0.06, max_delta_step=0,
      max_depth=3, min_child_weight=1, missing=None, n_estimators=5000,
      n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
      reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
      silent=None, subsample=1, verbosity=1)
```

```
Score for X_train, y_train:0.9293100575676068
Score for X_test, y_test:0.9068650462791137
Root Mean Square Error for test = 0.22861732292912815
```

Lightgbm:

```
LGBMRegressor (bagging_fraction=0.85, bagging_freq=2, bagging_seed=2,
       boosting_type='gbdt', class_weight=None, colsample_bytree=1.0,
       feature_fraction=0.2, feature_fraction_seed=2,
       importance_type='split', learning_rate=0.07, max_depth=-1,
       min_child_samples=20, min_child_weight=0.001, min_split_gain=0.0,
       n_estimators=8000, n_jobs=-1, num_leaves=7, objective='regression',
       random_state=None, reg_alpha=0.0, reg_lambda=0.0, silent=True,
       subsample=1.0, subsample_for_bin=200000, subsample_freq=0,
       verbose=-1)
```

```
Score for X_train, y_train:0.9620329331103199
Score for X_test, y_test:0.9344094671625355
Root Mean Square Error for test = 0.1918551064176578
```

Part IV Insight and Recommendation

4.1 How is Airbnb really being used in and affecting the neighborhood?

From the geography pattern in step 2, Airbnb has now widely spread in Boston. As a company belonging to the third sector, Airbnb has been closely connected to the areas they are located. Our first point is that development of Airbnb has brought mutual benefits to both Airbnb and neighborhoods. We can see from the maps, Airbnb listings concentrate on economically developed areas such as Downtown, Back Bay, West End. A convincing reason could be convenience. Due to easy access to transport, restaurants and shopping malls, demand is more in these areas. Mutually, more customers promote the development of other industries in these areas.

The second focus is on those quickly developed areas. After reading the two 3-D graphs above, we can see an interesting thing: Airbnb located in areas like Longwood Medical Area and Mission Hill are less compared to those located in downtown areas but have extremely high average reviews. One possible speculation is that supply falls short of demand in these areas. When there are not many choices for customers, the Airbnb is made full use of which means more people would reserve the same room, which leads to high average reviews. This is a niche for investors who have interests in these areas.

Also, we learned that Airbnb with relatively higher prices are located in West End, Back Bay and South Boston which are coincidently in high overlap with the areas with Airbnb densely distributed in. Combined with steady average reviews, we can conclude that customers are to some degree not sensitive to price. Then we suggest that hosts in these areas could consider improving service quality and expand service scope to be distinguished and win more customers.

4.2 Is there any trend of using Airbnb in Boston over time?

Boston has witnessed a continuous expansion of Airbnb market over time:

First, the demand for Airbnb shows overall increase over the years. Within one specific year, there is an obvious seasonal pattern, with demand increasing from January to October, and then declining slightly in winter season. It reveals that a growing number of visitors across the country and even all over the world who spend time in Boston see Airbnb as prior option. It is probably because travelers value Airbnb’s unconventional home-sharing accommodation, which makes them feel at home.

Second, Airbnb’s supply has also experienced explosive growth since its inception in Boston in 2009 – the number of new listings exponentially swelled each year. Besides the growth in the number of listings, we also see an increase in the diversity of rentals. For example, the offerings of room types like hotel rooms, shared rooms have risen over recent years, targeting at a wide range of visitors from business travelers to families. It can be concluded that over 10 years, Airbnb has continued to attract different homeowners to join, who have constantly provided unique and novel services, all things that traditional hotels fail to compete.

Finally, a growing Boston market can also be inferred from the continuously increasing pricing. According to *calendar.csv*, despite fluctuation, average prices of listings generally increase over 2019, and the upward trend is projected to continue in 2020. Besides, seasonality also exists in pricing, leading to the highest average price in May.

4.3 What recommendation you will make to Airbnb hosts and Airbnb?

For Airbnb hosts:

1) Enter the Unsaturated Neighbourhood

As we mentioned in the 3-D graph (2.12 Average Number of Reviews), it is possible that in areas like longwood medical area, north end and east Boston, supply falls short of demand. Hosts could see this opportunity and try to utilize the market.

2) Be the Super Host!

After analyzing the pricing factors, we find out that super hosts tend to be more competitive in pricing. Super hosts of Airbnb need to satisfy four criteria that covers hosting numbers, response rate, 5-star review numbers and low cancel rate.

In fact, becoming a super host will not only make higher profits, it can also appeal more customer traffic and earn better reputation in the market.

3) Keep Improving Services

We find out that when the house is instant bookable, the price is always higher, and factors such as the availability and security deposit also have positive relationship with price. For Airbnb hosts, they can keep upgrading such kind of services to earn more profits.

Some hosts may hope to attract more customers by lowering prices, nevertheless, the price is not the only determining factor, given the fact that expensive neighbourhoods still attract lots of guests. The quality of the house and the value of the service may serve more important roles in attracting customers. Airbnb hosts should improve the overall value of their service and customer experiences.

For Airbnb:

1) Improve Performance During Slack Season

We see the obvious seasonal fluctuation of Airbnb review numbers in Boston. Although the weather cannot be controlled, Airbnb can make some seasonal discount or hold events during the wintertime. Since there are many festivals in such time period such as Christmas, New Year and Valentines' Day, Airbnb platform can hold festival related activities to motivate hosts and also attract more guests.

2) Provide Unified Cleaning Service

From customer reviews, we notice many customers concern a lot about cleanliness. In real world, lots of Airbnb hosts have to hire third party cleaning services on their own. It may cause two problems. For one thing, customers have to pay extra cleaning fee. For another, the cleaning service is not standard among different hosts and may not satisfy the need of customers.

We wonder if Airbnb can provide the unified cleaning service with lower price to both relieve the pressure of hosts and meet customers' requirements.

3) Propose more evaluation criteria

Airbnb can propose more evaluation criteria, not only for super host, but also for the most popular host, the most interesting host, the most cost-effective host etc. Diversified evaluation standards are conducive to the benign competition of hosts, as well as to improve the popularity and public acceptance of Airbnb. It can also form a mutually beneficial relationship between Airbnb and host.

Part V References

What is Airbnb Super host Status Really Worth?
https://www.airdna.co/blog/airbnb_superhost_status
Visualising Residuals
https://drsimonj.svbtle.com/visualising-residuals
Airbnb Pricing Tool
http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/Airbnb_Pricing_TeamR_MASTER.HTML
Machine Learning
https://www.geeksforgeeks.org/introduction-machine-learning-using-python/
Scikit-learn
https://scikit-learn.org/stable/supervised_learning.html#supervised-learning