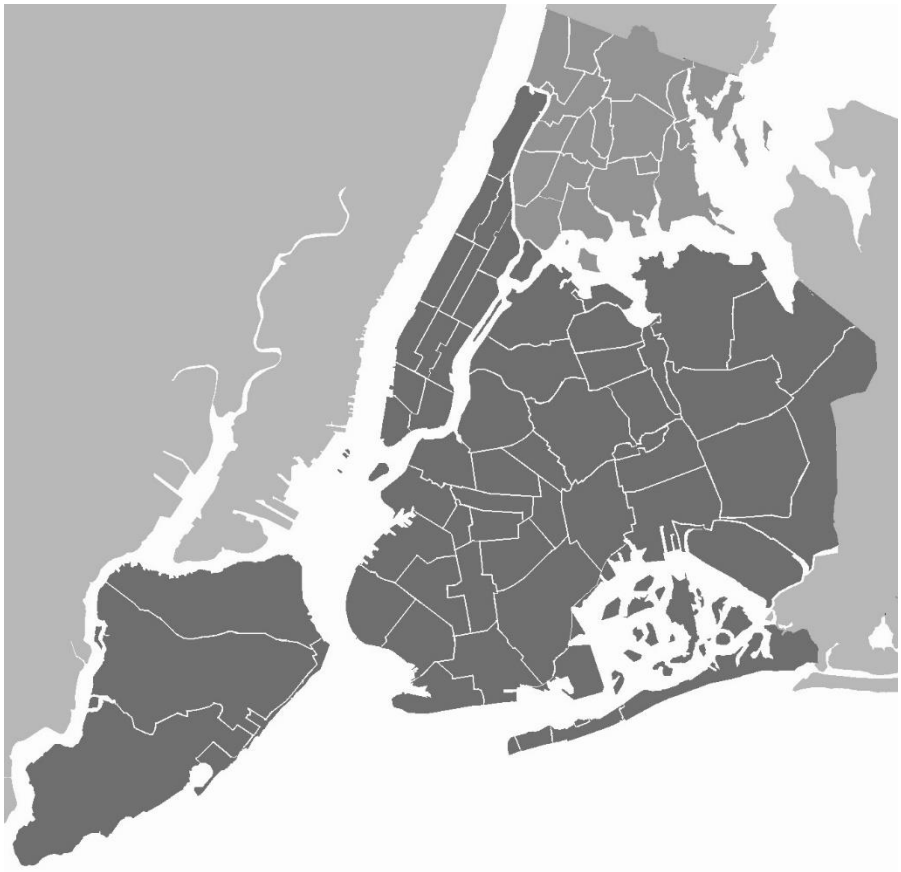
Data Analysis on New York City

Open Airbnb Dataset

Abtin Mahyar



Analysis

1. Data Exploration

The Airbnb dataset has around 50,000 records from different hosts all around New York city with various types of fields which made the dataset valuable and full of hidden information.

The dataset has the following attributes:

- "**id":** identifier of each record (integer)
- **"name":** name of each record (string)
- **"host_id":** identifier of each host (integer)
- **"host_name":** name of each host (string)
- **"neighbourhood_group":** the dataset has split records to five different neighbourhood category depends on its' geographical position which contains some neighbourhood in itself (category)
- **"neighbourhood":** a district within a town (string)
- **"latitude":** geographic coordinate that specifies the north–south position of a point on the Earth's surface (float)
- **"longitude":** a geographic coordinate that specifies the east–west position of a point on the Earth's surface (float)
- **"room_type":** there are three different kind of room in this dataset which will be discussed further (category)
- **"price":** price of each room for a specific amount of time (integer)
- **"minimum_nights":** minimum amount of time that the renter must book in order to stay.
- **"number_of_reviews":** sum of all reviews which is submitted for each record (integer)
- **"last_review":** date of last submitted review (datetime)
- **"calculated_host_listings_count":** number of transaction that each host had in the gathered dataset (integer)
- **"availability_365":** sum of available days for each record during a year (integer)

As it can be suspected based on above attributes, some of the features are useless or illegible, which means they must be discarded during the preprocessing.

Also, dataset has some null values in some of its' features which should be managed during preprocessing. Moreover, the number of null records in *last_review* and *reviews_per_month* are totally equal which is kind of suspicious and it finds out that the records with total number of reviews have null values in these two columns so logically, *reviews_per_month* for these records should be replaced by "0" and the column *last_review* should be dropped due to its' uselessly during our analysis.

FEATURE	SUM OF NULL RECORDS
Name	16
Host_name	21
Last_review	10052
Reviews_per_month	10052

Table 1. Number of null values for each feature

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	48895	48895	48895	48895	48895	48895	48895	38843	48895	48895
mean	19017143.24	67620010.65	40.72894888	-73.9521696	152.720687	7.029962164	23.27446569	1.37322143	7.143982002	112.7813273
std	10983108.39	78610967.03	0.054530078	0.046156736	240.15417	20.51054953	44.55058227	1.680441995	32.95251885	131.6222889
min	2539	2438	40.49979	-74.24442	0	1	0	0.01	1	0
25%	9471945	7822033	40.6901	-73.98307	69	1	1	0.19	1	0
50%	19677284	30793816	40.72307	-73.95568	106	3	5	0.72	1	45
75%	29152178.5	107434423	40.763115	-73.936275	175	5	24	2.02	2	227
max	36487245	274321313	40.91306	-73.71299	10000	1250	629	58.5	327	365

Figure 1. Numerical features description

Figure 2. Numerical features histogram (each outlier is presented by a red shape)

The chart and figure above show that there are also outliers in numerical features which should be eliminated during data preprocessing.

2. Data Preprocessing

The preprocessing done in the "Airbnb_cleaning" and "Airbnb" notebooks consist of the following steps:

1. Null values are eliminated or filled with a specific value. According to **Table 1** there are four columns which have null values. As it is discussed before, feature *reviews_per_month* of records which have no reviews changed to zero and other columns were dropped because those columns were not effective on our analysis.
2. Outliers were dropped from the dataset. The dataset consists of approximately five percent of outliers which means that 95% of the dataset remain unchanged and it can not have a huge influence on our data.
3. Data types are fixed and each attribute convert to its' own data type.
4. Applying binning method on some features due to their skewed distribution.

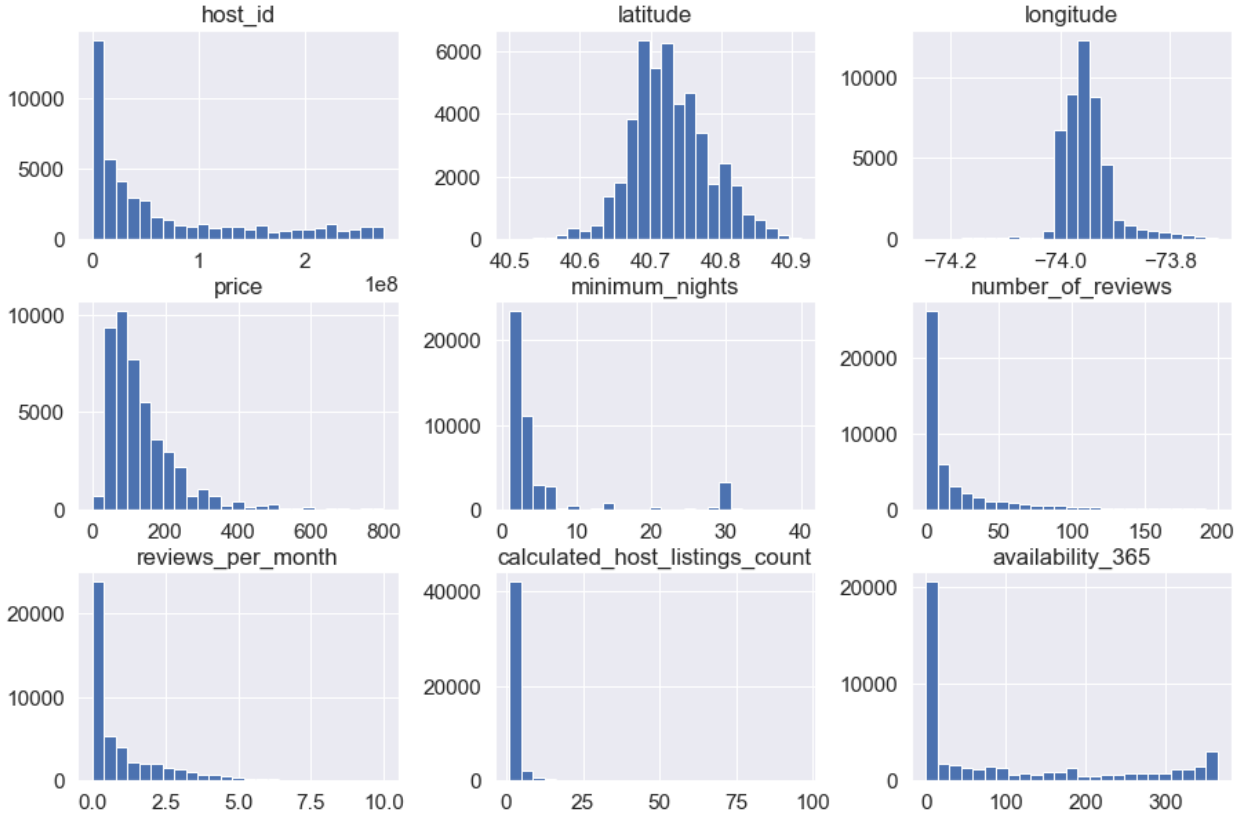


Figure 3. Numerical features distribution after eliminating outliers

The binning method and creating new features was performed with the following steps:

1. *Price* feature was divided to five different categories due to its' distribution with the following presentation and stored in new attribute *price_cat*.

LABEL	INTERVAL (PRICE)	PERCENTAGE OF TOTAL DISTRIBUTION
LOW	(0, 69]	0 - 25%
BELOW_AVERAGE	(69, 102.5]	25 - 50%
ABOVE_AVERAGE	(102.5, 175]	50 - 75%
HIGH	(175, 345]	75 – 95%
VERY_HIGH	(345, 800]	95 – 100%

Table 2. Categories of price_cat

2. *Minimum_nights* feature was divided to three different categories due to its' distribution with the following presentation and stored in new attribute *min_nights_cat*.

LABEL	INTERVAL (MINIMUM_NIGHTS)	PERCENTAGE OF TOTAL DISTRIBUTION
LOW	(1, 2]	0 - 50%
MEDIUM	(2, 10]	50 - 88%
HIGH	(10, 40]	88 – 100%

Table 3. Categories of min_nights_cat

3. *Number_of_reviews* feature was divided to four different categories due to its' distribution with the following presentation and stored in new attribute *reviews_cat*.

LABEL	INTERVAL (PRICE)	PERCENTAGE OF TOTAL DISTRIBUTION
LOW	(0, 5]	0 - 50%
BELOW_AVERAGE	(5, 17]	50 - 70%
ABOVE_AVERAGE	(17, 66]	70 - 90%
HIGH	(66, 200]	90 – 100%

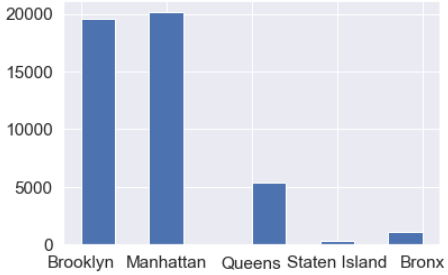
Table 4. Categories of reviews_cat

3. Exploratory Visualization

3.1. Neighbourhoods and Records distribution

The plot shows how records are distributed over different neighbourhood group. At first glance it can be seen that, the majority of hosts and records are located in "Manhattan" and "Brooklyn" and other neighbourhood groups are much smaller than those two in comparing number of records.

Figure 4. Distribution of different record grouped by neighbourhood group



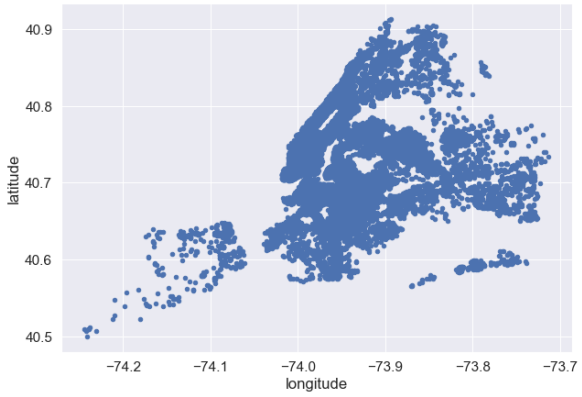


Figure 5. Geographical and distribution of records

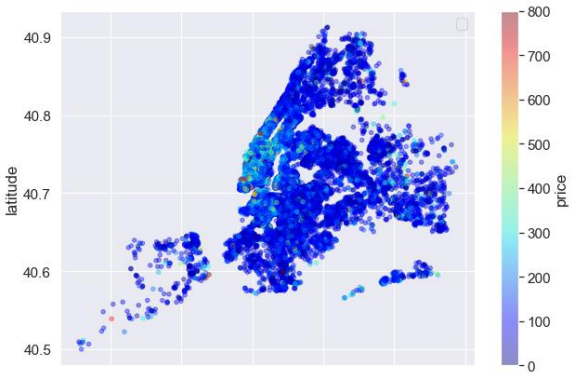


Figure 6. Geographical and distribution of records with their prices

Fig. 5, 6 both plot showing how records and hosts are geographically distributed. Note that majority of records are located around the middle of latitude and longitude intervals. In Figure 6 as it can be seen, records are colored based on their prices and the majority of records with high amount of price are located again around the middle of intervals, so the middle part of the city should be crowded and rich. However, there are some other records on different part of the map which have high price too.

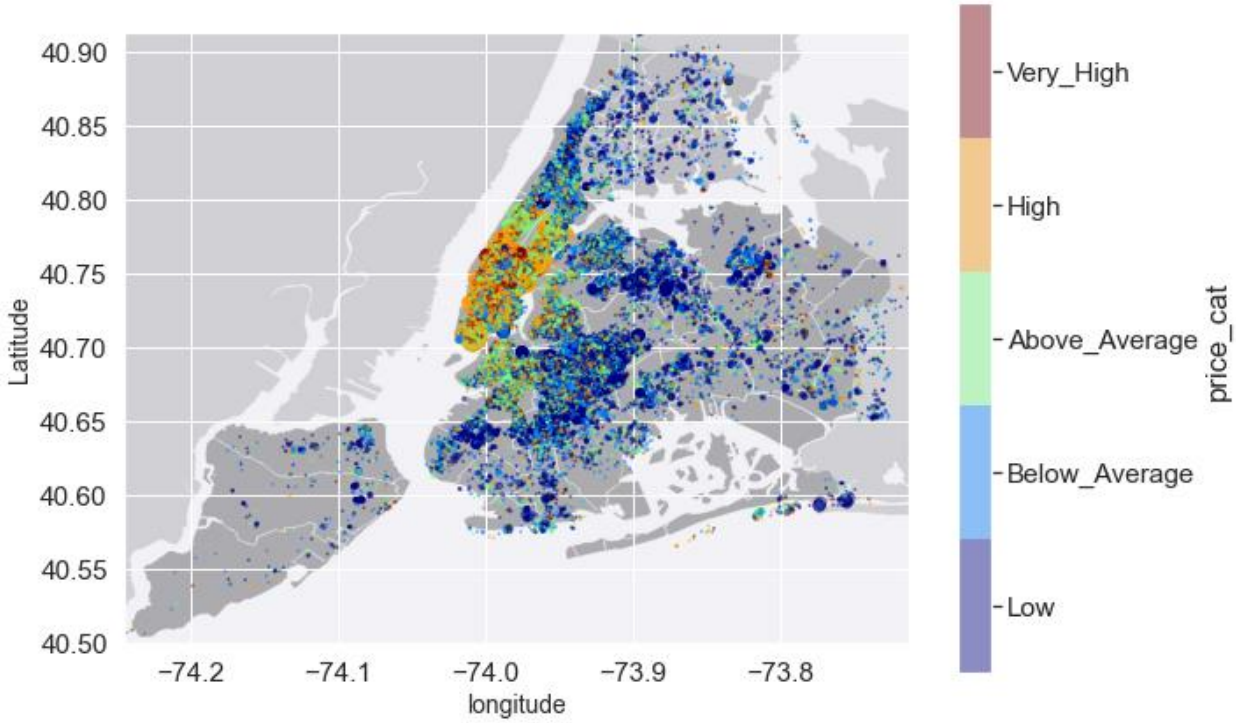


Figure 7. Geographical distribution of records using price_cat and host_listings_count

Fig. 7 The following plot shows how records are distributed around the map. Note that the color of each record represents its' *price_cat* which proofs the same thing that we have discussed before which was the records that are located in the middle of the city are in high and very high

price categories. Also, size (radius) of each record represents its' *calculated_host_listings_count* which helps us to figure it out that the majority of records are located in the middle of the city.

Fig. 8 A plot showing that where are the neighbourhoods are located on the map at how vast they are. The center of each circle comes from applying mean method to longitude and latitude of records in each category and the radius is the mean of standard deviation of longitude and latitude.

So, we can conclude that the majority of people are living in "Manhattan" and "Brooklyn" which are the richest neighbourhoods compare to the others (same result of Figure 4).

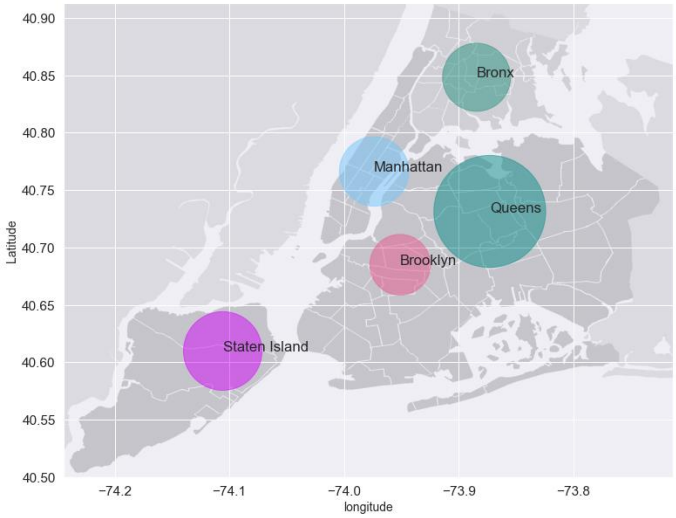


Figure 8. Neighbourhoods groups and its' size on the map

Fig. 9 The following plot shows how records are distributed based on neighbourhood group and quantitative availability variable**.** According to the box plot, we can see that the distributions of records are different from each other. Moreover, Staten Island is the most available neighbourhood compare to the others whilst Brooklyn is the least available neighbourhood among them. Also, Queens and Bronx have the wide range of availability during a year.

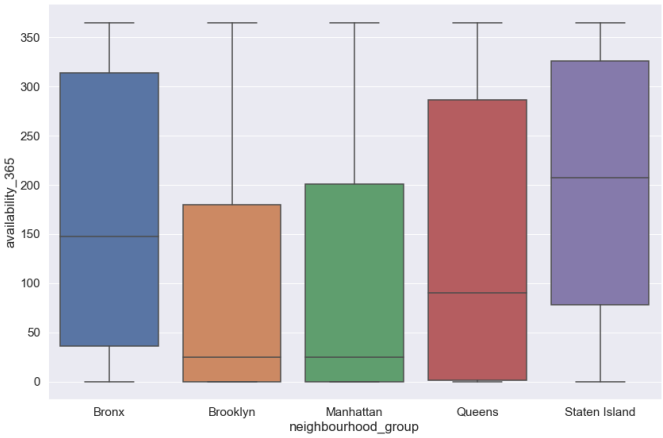


Figure 9. Box plot of records based on neighbourhood group and availability

Fig. 10 A plot showing how records are distributed based on their room type and minimum nights. According to the bar chart, the distributions of room types are different from each other depends on minimum nights which means that these two categorical variables are dependent.

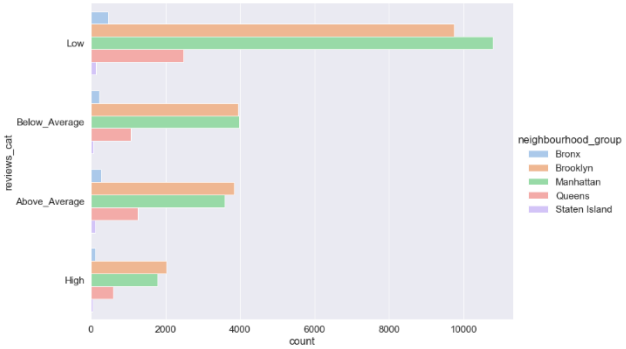


Figure 10. Distribution of records based on neighbourhood group and reviews

Fig. 11 A plot showing how records are distributed based on their neighbourhood group and minimum nights. According to the bar chart, the distributions are different from each other depends on minimum nights which means that these two categorical variables are dependent. As can be seen from the chart, Manhattan has a wide range of number of minimum nights whilst this number for Staten Island is narrower.

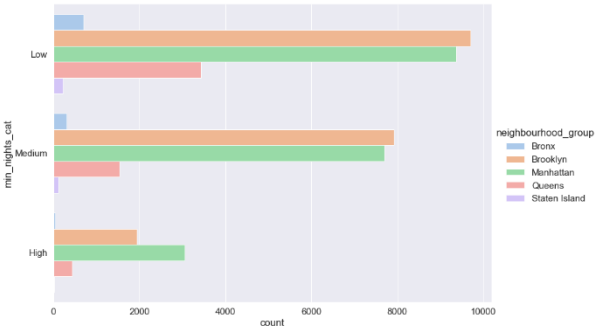


Figure 11. Distribution of records based on neighbourhood group and minimum nights

3.2 Room Type

Fig. 12 The following plot demonstrates the number of total rooms form each category in the dataset. As it can be seen, the number of private and entire home or apartment are much higher than shared room.

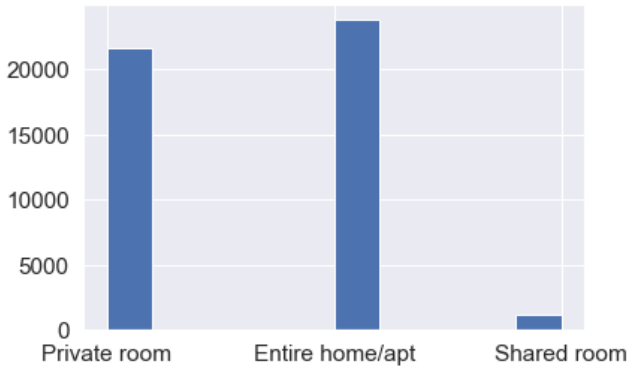


Figure 12. Room Type histogram

Fig. 13 The following plot shows how records are distributed in two categorical variables (neighbourhood group and room type). As it can be seen from the bar chart, the distribution of room type in each city is different from each other and the minority of rooms are as type of shared room which most of them are located in "Manhattan" and "Brooklyn" so we can conclude that these two cities have the wide variety of room types. However, the majority of rooms in "Manhattan" are consists of entire home or apartment whilst the number of private rooms and entire home or apartment are approximately equal. Also, we can guess that room type is dependent on the neighbourhood group which means that the distribution of different kinds of room are different from one city to another; Although, this hypothesis should be tested during statistical analysis.

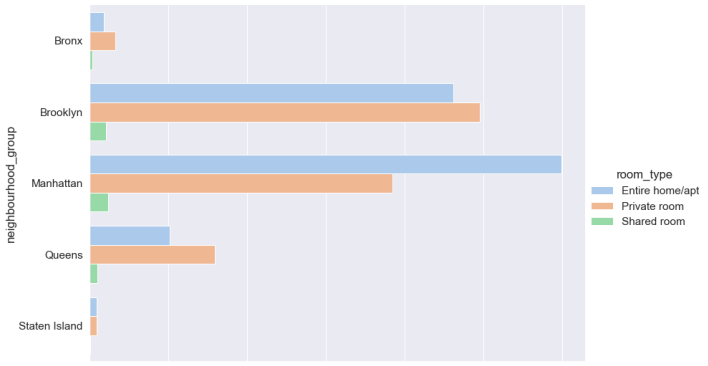


Figure 13. Distribution of records grouped by neighbourhood group and room type

Fig. 14 A plot showing how records are distributed based on their room type and minimum nights. According to the bar chart, the distributions of room types are different from each other depends on minimum nights which means that these two categorical variables are dependent. Also, with increasing number of minimum nights form low to medium or high category, number of entire home or apartment increase and number of private rooms decrease. Moreover, with increasing number of minimum nights, number of shared room and total number of rooms decrease.

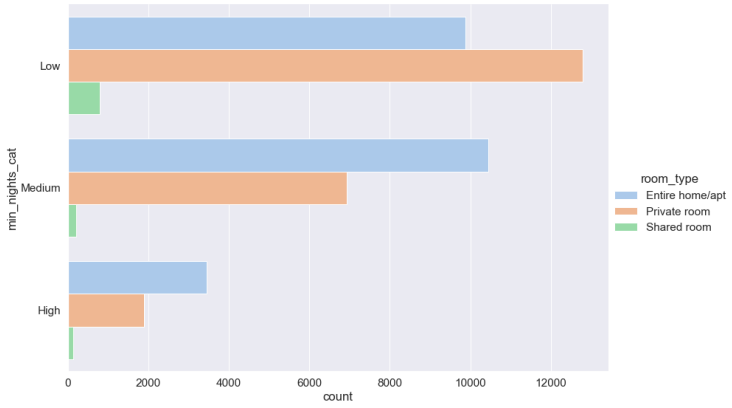


Figure 14. Distribution of records based on room type and minimum nights

Note: I preferred to use *min_nights_cat* which is a made-up categorical variable based on *minimum_nights* instead of its' quantitative variable because this variable is nonparametric and have skewed distribution so the resulted plot with quantitative variable is not useful and effective on our analysis. Also, I do this on some other quantitative further in this paper due to gaining useful and effective plots in our analysis.

Fig. 15 The following plot shows how records are distributed based on room type and quantitative availability variable. According to the box plot, we can see that the distributions of records are different from each other. Although, the distribution of entire home or apartment and private room are similar, shared rooms are more available during a year and has a wider range of availability compare to two other categories.

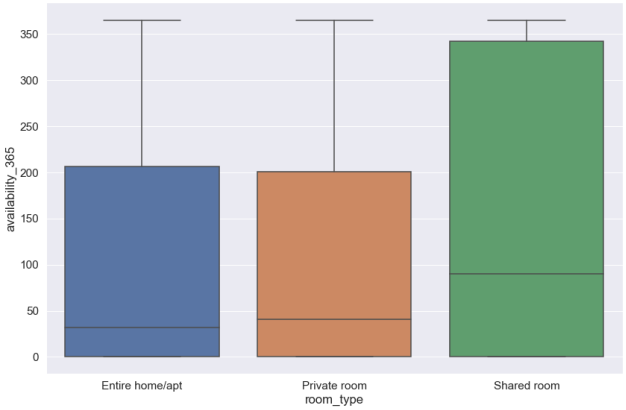


Figure 15. Box plot of records based on room type and availability

3.3. Price

The plot below shows how records are distributed based on their price and neighbourhood group. According to the bar chart, we can see that the distribution of records are different from each other so we can guess that these two categorical variables are dependent. Also, the richest neighbourhood group is "Manhattan" which has highest number of records with room type of above the average and higher. After Manhattan, the richest neighbourhood is Brooklyn. Although, Brooklyn has lots of low-price rooms comparing to other price groups and neighbourhoods. Moreover, Staten Island and Bronx are the poorest neighbourhoods among the others. Furthermore, the very high price rooms are mainly located in Manhattan and Brooklyn; However, the minority of this group are located in Queens which made it a medium neighbourhood in terms of price.

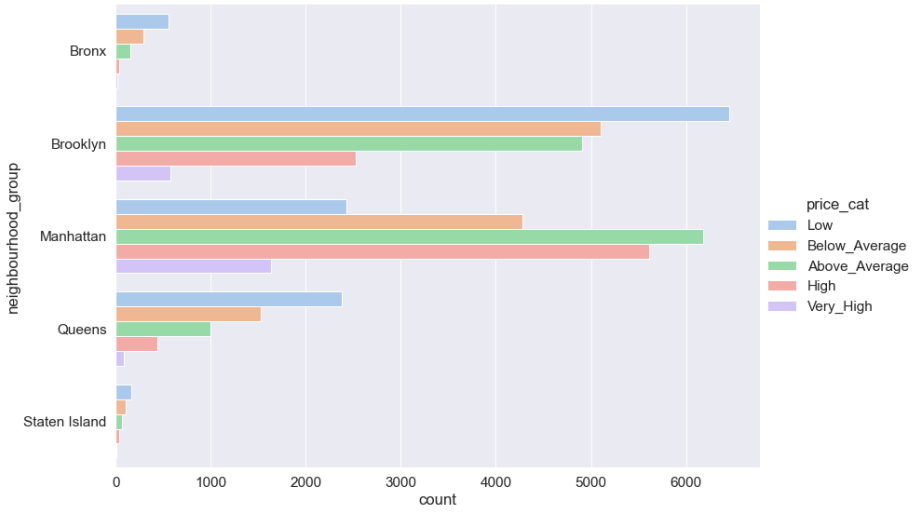


Figure 16. Distribution of records based on price and neighbourhood group

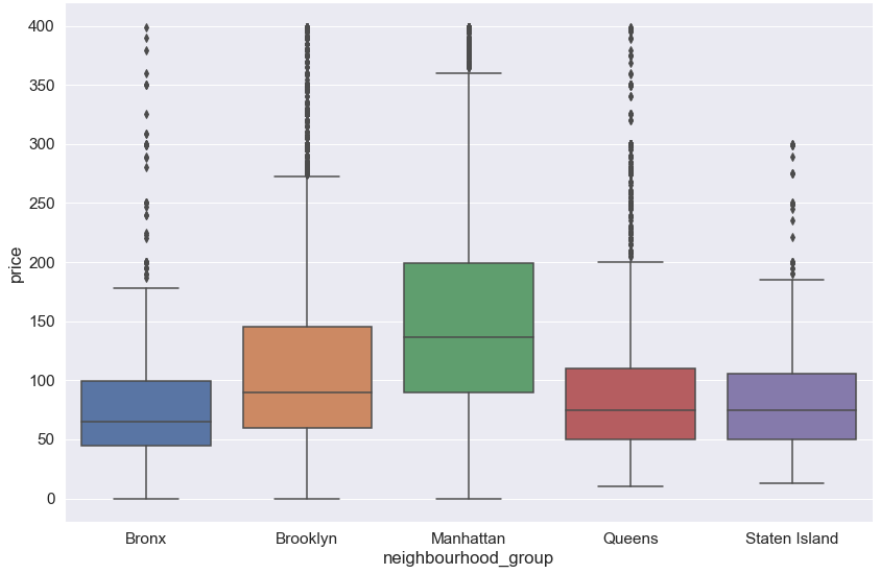


Figure 17. Box plot of records based on their neighbourhood group and price

Fig. 17 The following plot shows how records are distributed based on neighbourhood group and quantitative price variable (because price has skewed distribution, only records with price < 400 are consider in this plot). From this plot we can make more confidence about our hypothesis which have discussed above which was distributions are different form each other. As can be seen in this box plot, Manhattan is the richest neighbourhood group among the others and also, have wide range of different rooms in terms of price. Moreover, Staten Isalnd and Bronx are the poorest neighbourhoods among the others and also, have narrower range of rooms in terms of price.

Fig. 18 A plot showing how records are distributed based on their price and room type. According to the bar chart, we can see that the distributions of records are different from each other so we can guess that these two categorical variables are dependent. Mainly, share rooms and private rooms are cheaper than entire home or apartments which means with increasing of price number of home or apartments are much higher than private or

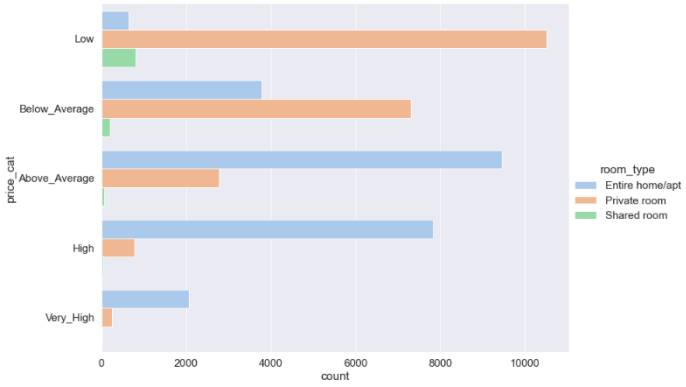


Figure 18. Distribution of records based on price and room type

shared rooms. Also, shared room is the cheapest type of room in comparing with other types.

Fig. 19 The following plot shows how records are distributed in two categorical variables (price and minimum nights). According to the bar chart, we can see that the distributions of records are different from each other so we can guess that these two categorical variables are dependent.

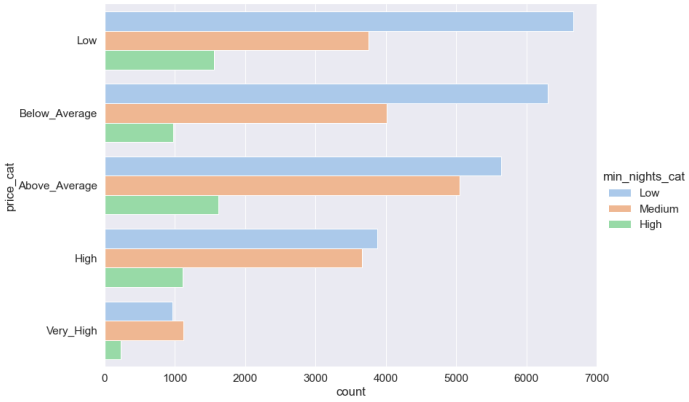


Figure 19. Distribution of records based on price and minimum nights

Fig. 20 A plot showing how records are distributed based on their price and room type. According to the box plot, we can see that the distributions of records are different from each other so we can guess that these two categorical variables are dependent. According to the bar chart, rooms with very high price are more available during the year and this availability decreases when we reduce the price of the room; Therefore, rooms that are below average in terms of price are much less available during the year; although, there is a exception for this which is the rooms with low price which are more available compare to the below average type.

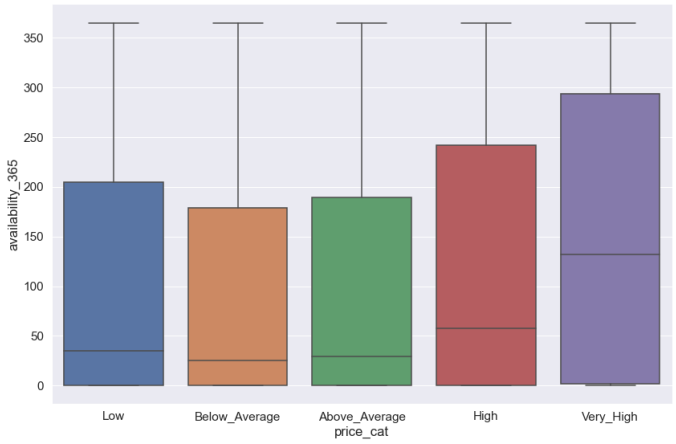


Figure 20. Box plot of records based on price and availability

Fig. 21 The following plot shows how records are distributed in two categorical variables (price and reviews). According to the bar chart, we can see that the distributions of records are different from each other so we can guess that these two categorical variables are dependent.

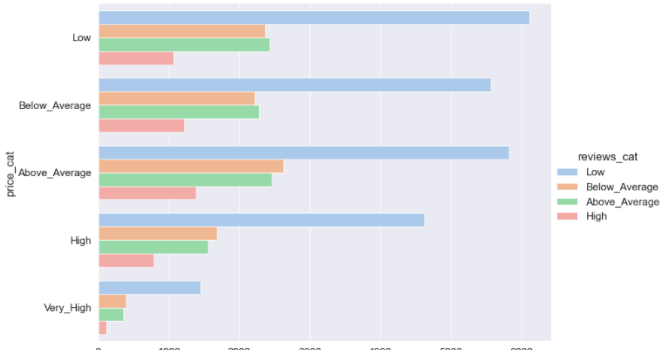


Figure 21. Distribution of records based on price and reviews

4. Statistical Tests and Analysis

4.1. Chi2 Test

Chi-squared is a statistical hypothesis test for checking the dependency of two categorical variable. Since we made lots of hypothesis based on visualization from the last section, we should check our hypothesis of the comparisons between two categorical variables with this test.

For applying Chi-squared test on our dataset, first we should calculate the contingency table (each cell of this table corresponds to the frequency of a specific observation) of each comparison.

H0 (null hypothesis test): Two categorical variable are independent.

After calculating p-value and statistic for each comparison if p-value is less than or equal to significance level (assumed to be 0.05 = 5%) then the H0 will be rejected and categorical variables are dependent. Equally, if the statistic is more than or equal to critical value, then H0 will be rejected.

Results of applying chi-squared test on the dataset comes as follows:

COMPARISON	(STATISTIC, P-VALUE)	RESULT
NEIGHBOURHOOD GROUP VS. REVIEWS	(248.56, < 0.001)	Dependent (reject H0)
NEIGHBOURHOOD GROUP VS. MINIMUM NIGHTS	(882.79, < 0.001)	Dependent (reject H0)
NEIGHBOURHOOD GROUP VS. ROOM TYPE	(1263.05, < 0.001)	Dependent (reject H0)
ROOM TYPE VS. MINIMUM NIGHTS	(1635.19, < 0.001)	Dependent (reject H0)
PRICE VS. NEIGHBOURHOOD GROUP	(5851.42, < 0.001)	Dependent (reject H0)
PRICE VS. ROOM TYPE	(21745.91, < 0.001)	Dependent (reject H0)
PRICE VS. MINIMUM NIGHTS	(704.73, < 0.001)	Dependent (reject H0)
PRICE VS. REVIEWS	(276.01, < 0.001)	Dependent (reject H0)

Table 5. Result and p-values of different categorical comparison using Chi-squared test

As expected, all categorical variables are dependent and our hypothesizes were correct.

4.2. Kruskal-Wallis H Test

Kruskal is a statistical hypothesis test for checking the similarity of distribution of a nonparametric quantitative variable over a categorical variable with more than two sample (it is an alternative test for Friedman test, but since our quantitative variable in each sample does not have same size, we must use this test instead of Friedman). Since we made lots of hypothesis based on visualization from the last section, we should check our hypothesis of the comparisons between a quantitative and a categorical variable using this test.

H0 (null hypothesis test): same distribution of quantitative variable over samples.

After calculating p-value and statistic for each comparison if p-value is less than or equal to significance level (assumed to be 0.05 = 5%) then the H0 will be rejected and quantitative data have different distribution over samples. Equally, if the statistic is more than or equal to critical value, then H0 will be rejected.

Results of applying Kruskal test on the dataset comes as follows:

COMPARISON	(STATISTIC, P-VALUE)	RESULT
NEIGHBOURHOOD GROUP VS. AVAILABILITY	(1056.63, < 0.001)	Different distribution (reject H0)
ROOM TYPE VS. AVAILABILITY	(185.55, < 0.001)	Different distribution (reject H0)
PRICE VS. NEIGHBOURHOOD GROUP	(6173.84, < 0.001)	Different distribution (reject H0)
PRICE VS. AVAILABILITY	(464.05, < 0.001)	Different distribution (reject H0)

Table 6. Results and p-values of different quantitative over categorical comparison using Chi-squared test

As expected, all quantitative data have different distribution over a category variable and our hypothesizes were correct.

4.3. Correlation Matrix

There may be complex and unknown relationship between quantitative variables in our dataset so it's better calculate correlation between different attributes and compare them before applying any linear model.

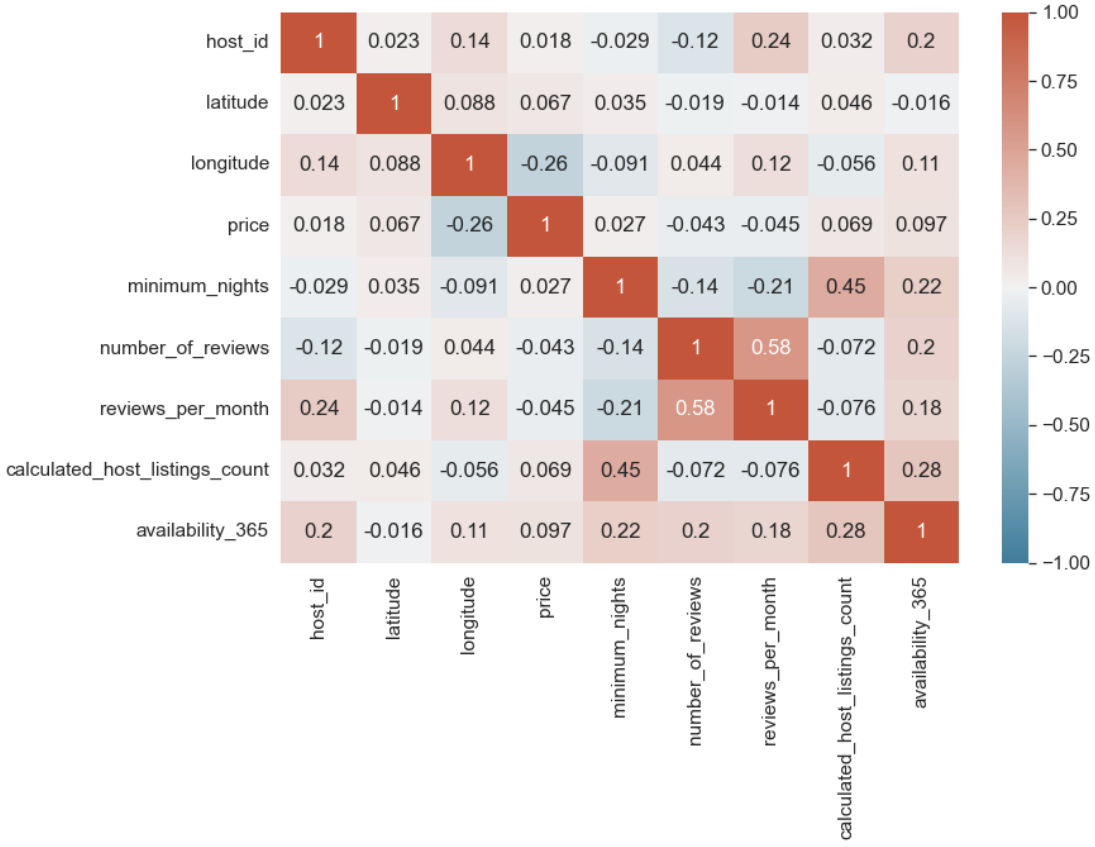


Figure 22. Correlation matrix over quantitative variables

As it can be seen from calculated matrix, there is no strong relationship between our quantitative variables which mainly happened due to skewed distribution of our quantitative variables. As a result, applying a linear or regression model can not be useful and will not give us a valuable information.

4.4. Analysis on Top Hosts

Top hosts are defined as the hosts with highest number of listings; in other words, hosts that repeated in dataset more than the others. Top ten hosts are listed as follows:

HOST_ID	LISTINGS
12243051	96
16098958	96
61391963	90
22541573	87
2E+08	62
7503643	52
1475015	52
1.21E+08	50
2856748	48
1.91E+08	47

Table 7

The table below shows that top hosts are from two richest neighbourhoods (Manhattan and Brooklyn).

NEIGHBOURHOOD_GROUP	COUNT
MANHATTAN	627
BROOKLYN	53
BRONX	0
QUEENS	0
STATEN ISLAND	0

Table 8

The table below shows that top hosts usually provide entire home or apartment or private room and never provide shared room.

ROOM_TYPE	COUNT
ENTIRE HOME/APT	633
PRIVATE ROOM	47
SHARED ROOM	0

Table 9

The table below shows the review category for each host. It can be seen that majority of top hosts dose not get much reviews.

REVIEWS_CAT	COUNT
LOW	586
BELOW_AVERAGE	93
ABOVE_AVERAGE	1
HIGH	0

Table 10

The table below shows that majority of top hosts provide high price rooms.

PRICE_CAT	COUNT
HIGH	304
ABOVE_AVERAGE	276
BELOW_AVERAGE	43
VERY_HIGH	33
LOW	24

Table 11

The table below shows that majority of top hosts provide rooms with higher number of minimum nights.

MIN_NIGHTS_CAT	COUNT
HIGH	640
MEDIUM	38
LOW	2

Table 12

4.5. Analysis on Top Neighbourhoods

Top neighbourhoods are defined as the neighbourhoods that repeated in dataset more than the others. Top ten neighbourhoods are listed as follows:

NEIGHBOURHOOD	COUNT
WILLIAMSBURG	3808
BEDFORD-STUYVESANT	3606
HARLEM	2564
BUSHWICK	2408
UPPER WEST SIDE	1870
HELL'S KITCHEN	1823
EAST VILLAGE	1770
UPPER EAST SIDE	1721
CROWN HEIGHTS	1527
MIDTOWN	1433

Table 13

The table below shows the distribution of room types in top neighbourhoods. According to the table, top neighbourhoods, mainly consists of entire home or apartment or private room.

ROOM_TYPE	COUNT
ENTIRE HOME/APT	11358
PRIVATE ROOM	10685
SHARED ROOM	487

Table 14

The table below shows that top neighbourhoods are from richest and crowded neighbourhood groups (Manhattan and Brooklyn).

NEIGHBOURHOOD_GROUP	COUNT
BROOKLYN	11349
MANHATTAN	11181
BRONX	0
QUEENS	0
STATEN ISLAND	0

Table 15

5. Results

The most important and valuables conclusions that gathered during the processes which were discussed above, are listed as follows:

- The neighbourhoods in this city can be divided into two groups. First group is the crowded and the rich one which consists of rooms with high prices. Manhattan and Brooklyn are placed in this group. The second group is the uncrowded and the poorest one compare to the first group which the majority of rooms that are located in this group are around average or below it. Staten Island, Bronx, and Queens are placed in this group.
- First group (crowded neighbourhoods) have the wider range of room type, price, and minimum night compare to second group. This is probably because of large population who live in the first group which have increased demand for different type of room with different amount of time and different prices. In other words, in first group neighbourhoods, customers can find the desired room that they want easily and faster compare to the second group, and that is why the majority of people are living in the first group neighbourhoods.
- Availability during a year is dependent on type of room, price, and neighbourhood group. Mainly, rooms which located in first neighbourhood group are less available compare to the second group, which is probably because of the majority of people and huge of amount of demand for living in this group. Moreover, rooms with high prices are more available during a year, because the minority of people can pay for these kinds of rooms and demand for this type of rooms are less than other ones. Also, shared room are much more available during a year compare to private rooms or entire apartments or homes due to low frequency and lack of demand for this kind of rooms in the city.
- Price is also dependent on type of room. Usually, price of entire apartment or home is more expensive than private or shared rooms.