Airbnb and Zillow Data Challenge

Arvind Ramkumar
aramkum@asu.edu
480(859)2393

I. Description

1. Problem Description and background
2. Basic Assumptions

II. Data Preprocessing

1. Working with Airbnb Data
2. Working with Zillow Data

III. Data Analysis

1. Number of Airbnb Properties in NYC
2. Number of Reviews Analysis
3. Cost Analysis
4. Revenue Analysis
5. Occupancy Rate Analysis
6. Review on Locations Analysis
7. Return of Investment (ROI) Analysis
8. One Percent Rule
9. Cumulating all Analysis

IV. Conclusion and Recommendations

V. Further Suggestions to Model Improvisation

Description

Analysis to identify the zip codes in NYC that are worthful in investing for rental properties with the aid of Airbnb and Zillow Data

Problem Description and Background

The investors are planning to buy properties in NYC and plan for short term rentals. They have already concluded that two-bedroom flats are more profitable, and we need to make analysis to suggest them on zip codes that could be more profitable and less risk

Assumptions

Given Assumptions

1. The investor will pay for the property in cash
2. The time value of money discount rate is 0%
3. All properties and all square feet within each locale can be assumed to be homogeneous

Assumptions Taken

1. The occupancy rate is taken from availability_360
2. One-night price is considered for analysis, considering that the monthly rate is given by (30* one-night price) discounted at a constant rate
3. Since, 2 bedrooms flats are known to be more profitable we consider only 2-bedroom flats
4. The investor could acquire the property at the latest available cost, and here the acquiring cost is assumed to be the price at 2016 July "2016-07"
5. Reviews that have missing data are given zero

Data Preprocessing

Working with Airbnb Data

1. Consider only two-bedroom flats.
2. Drop all zip codes that are incorrectly entered and also drop the columns that have missing zip codes.
3. Convert all zip codes to 5 digits.
4. Drop the $ symbol in price.

Working with Zillow Data

1. Unify all the zip codes
2. Drop all the zip codes that are out of NYC.
3. Drop all the $ symbol in the cost price

4. Convert Region ID to zip code

Data Quality Insights

1. The Main thing that we are considered is the zip code, price and the cost and there aren't any missing data is price and cost. And relatively a very few missing data in zip code which could be replaced by latitude and longitude, which suggests that the data has a better quality

2. There is so much data about the host, and we could run so much of NLP or sentimental analysis on the host to determine the influence of host on the occupancy. So, we have got abundance of data

3. Almost every time, there is an option for finding out parameters that are not given but are required for us like the occupancy rate and ROI (Discussed later). So, the quality of the data is excellent, and it is up to us to extract the data

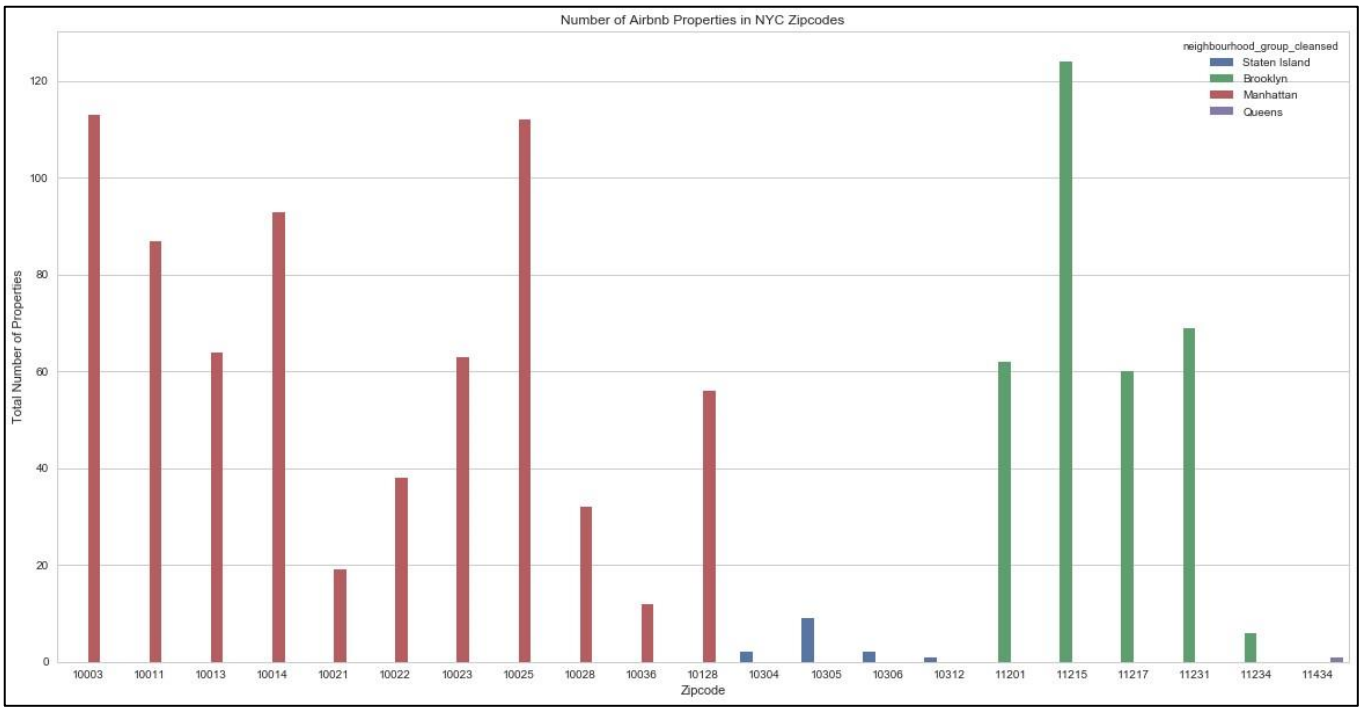
Data Analysis

Number of Airbnb Properties in NYC

This gives us a visual image to think about how many numbers of properties are located in NYC (2 bedroom) under Airbnb. This will give us a heads up in deciding on where to make the purchase.

There could be two interpretation

1. Invest where no one has and venture (High Risk, yet could be more profitable)
2. Invest where everyone has already done (Could be sophisticated with less risk and profitable and uncertain of how high the profit)



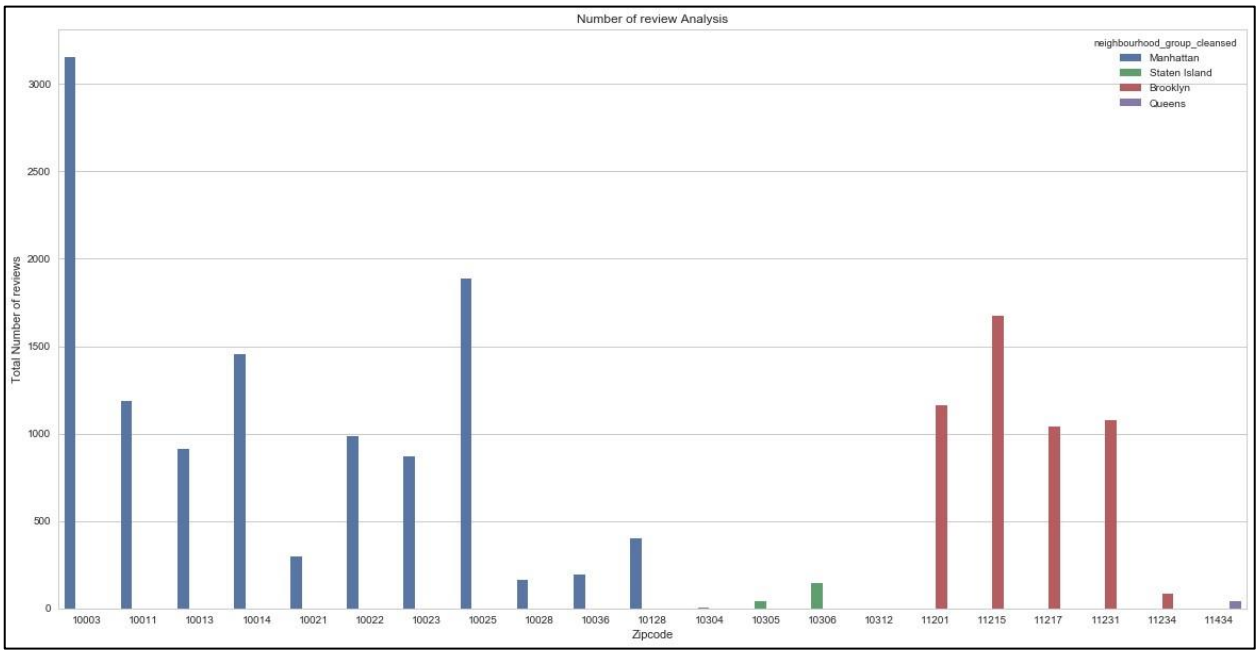
It is so evident that Manhattan has higher number of properties, which means that there is more demand in the Manhattan area.

Number of Reviews Analysis

With a number of factors affecting, the number of reviews might as well play an important role in determining whether or not invest in a property.

There are multiple number of properties in each zip code and there are number of reviews for each of these properties. So, if we could sum up the reviews given in each of the property for a given zip code, it is possible to obtain the total number of reviews per zip code.

Higher the number of reviews is more desirable.

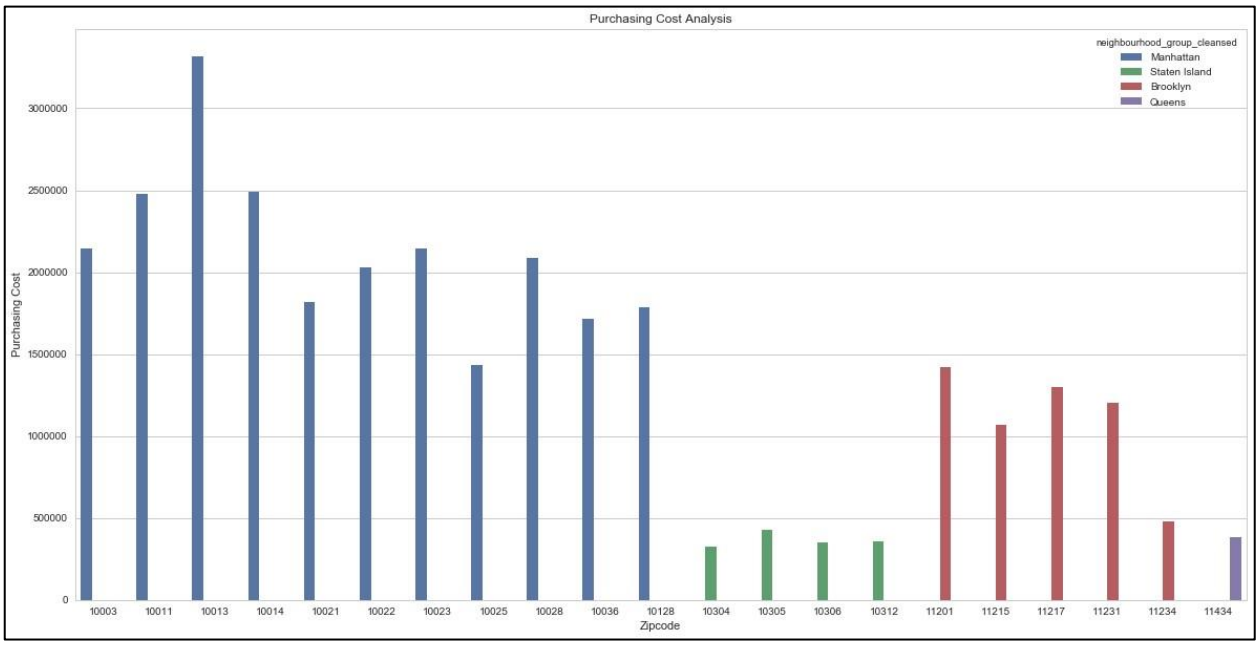


Using the other factors

As mentioned above, there are multiple properties under the same zip code and it is necessary that we need to aggregate it for the measure for one zip code. The aggregation is assumed to be mean, because the mean is a better estimate in this case.

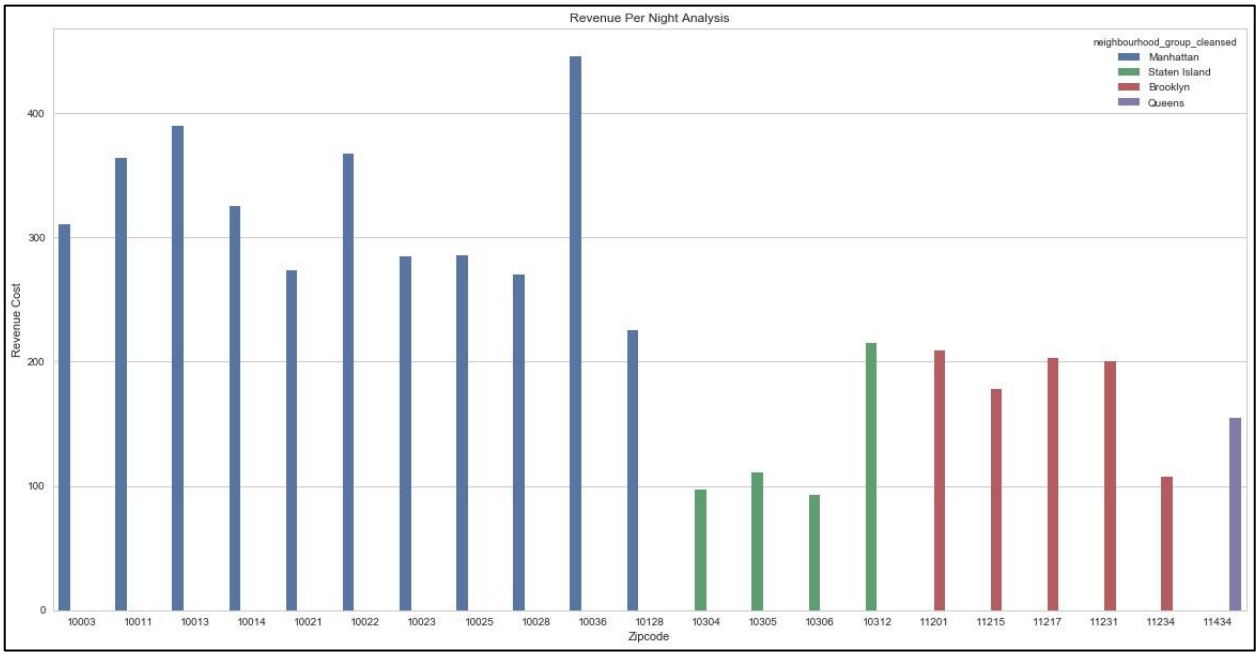
The factors that I think that might work are given below.
1. Price for One Night (Higher the better)
2. Acquiring Cost (Lower the better)
3. Reviews of the location given by customer (Higher the better)
4. Occupancy Rate (Higher the better)
5. ROI (Higher the better)

Cost Analysis



Lower the cost is more desirable

Revenue Analysis



Higher the revenue is more desirable

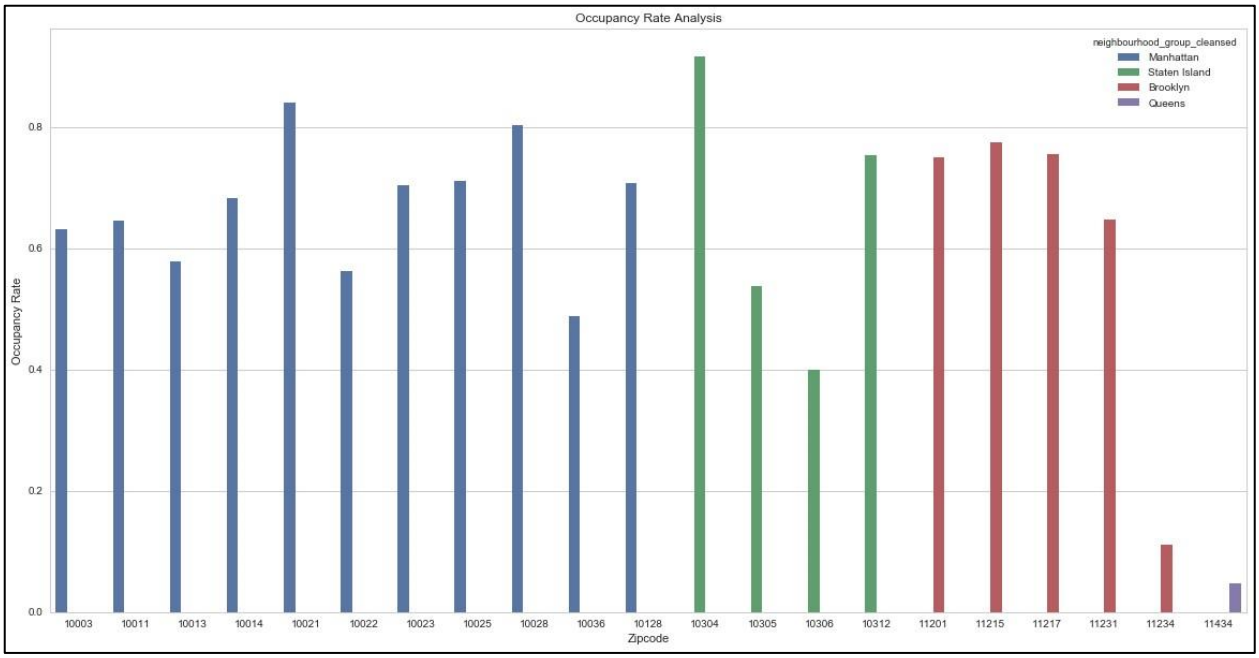
Occupancy Rate Analysis

Although it is given that we can assume the occupancy rate to be 75%, generalizing it would be not that accurate and it is obvious that not all property will have the same occupancy rate.

So, with the available data, there are multiple ways that could be used to model the Occupancy Rate, like using the total number of reviews in a zip code to give the occupancy rate.

But here, I have used the "availability_365" data to define the occupancy rate. It is logical that, if the house is available for the 365 days, then the occupancy is almost zero, and on the contrary, if the house is occupied for all 365 days, then the occupancy rate should be equal to 100%. So,

Occupancy Rate = (1 - (Number of days the house is available in next 365 days/ 365)



Higher the Occupancy Rate is more desired

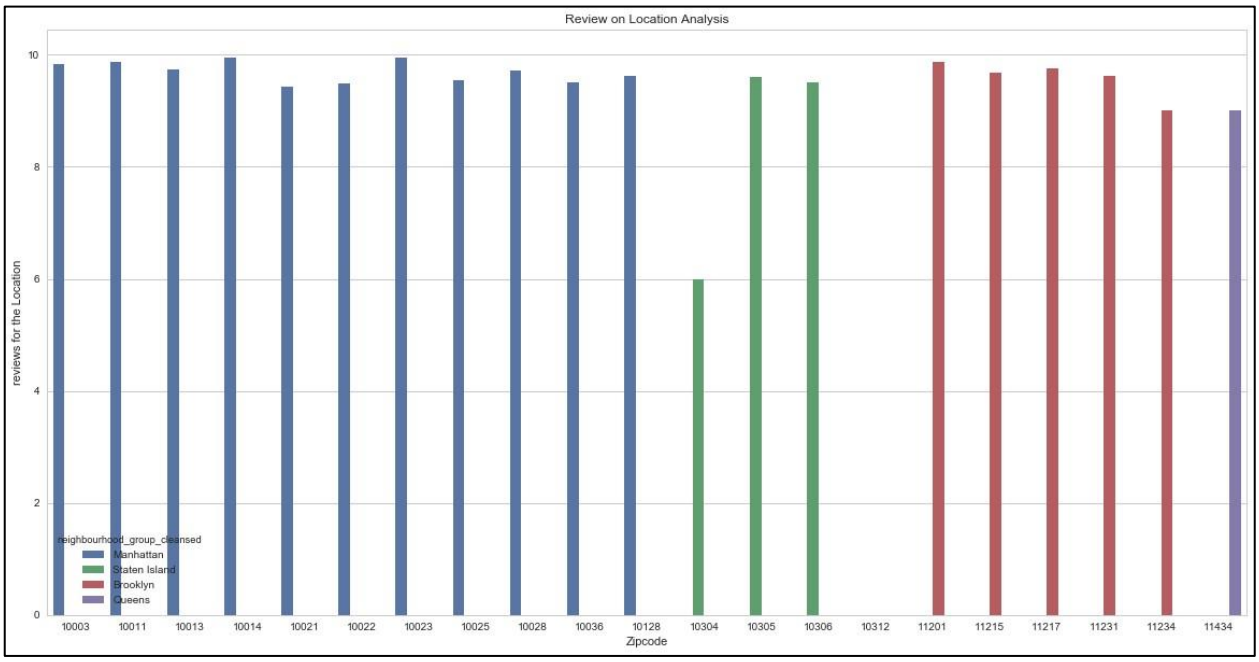
Check whether the occupancy rate and number of reviews are correlated

There are high chances that the Occupancy rate and the number of reviews might be correlated. Because more the people stay, the more the number of reviews.

If they are correlated then, having to work on both these parameters together, will create the problem of multicollinearity. So, we check their correlation.

The Occupancy Rate was just 2% Correlated with the Number of reviews. So, we can have both as a measure.

Review on Location Analysis



Higher Reviews means better reviews, so higher is what necessary

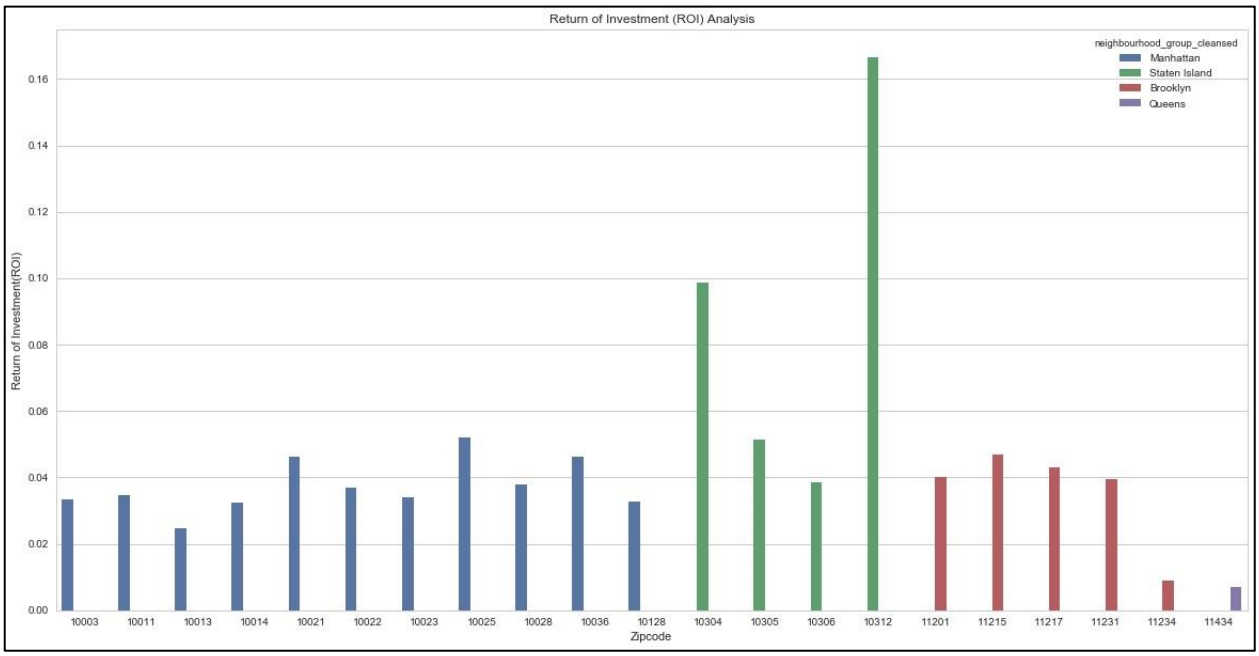
Return of Investment (ROI) Analysis

Return on investment (ROI) is an accounting term that indicates the percentage of invested money that's recouped after the deduction of associated costs.

Here, since we are acquiring the property with no mortgage or interest, those cost and all miscellaneous costs are not considered (No sufficient data). Also, we analyze using the one-night revenue.

ROI = (Price per night * Occupancy rate * 365) / Cost of Acquiring

High ROI is always preferred

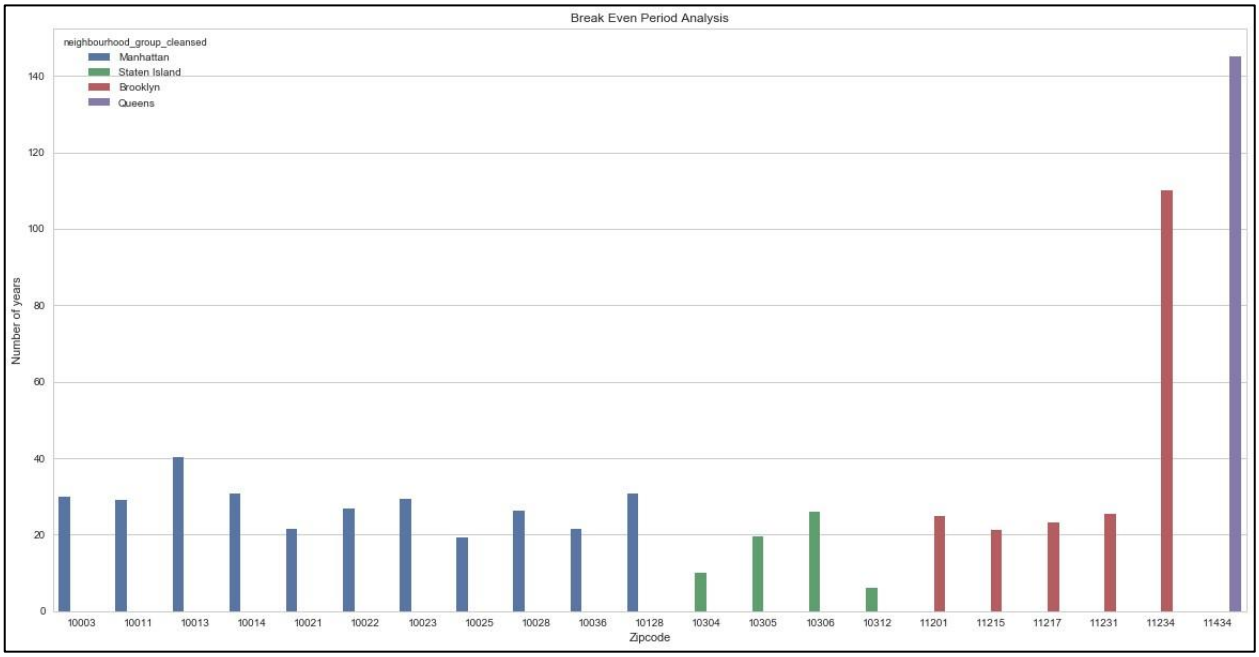


Break Even Period

It is the reciprocal of ROI. It will let us know the details on when (In Years) we can actually get back our investment and further on it would be pure profit.

Lesser the Break-even period is more preferred.
But ROI and Break-even period are both the same things. So, we use ROI for the analysis.

Break even period will help us in realizing in how many years the investor could actually obtain his profit

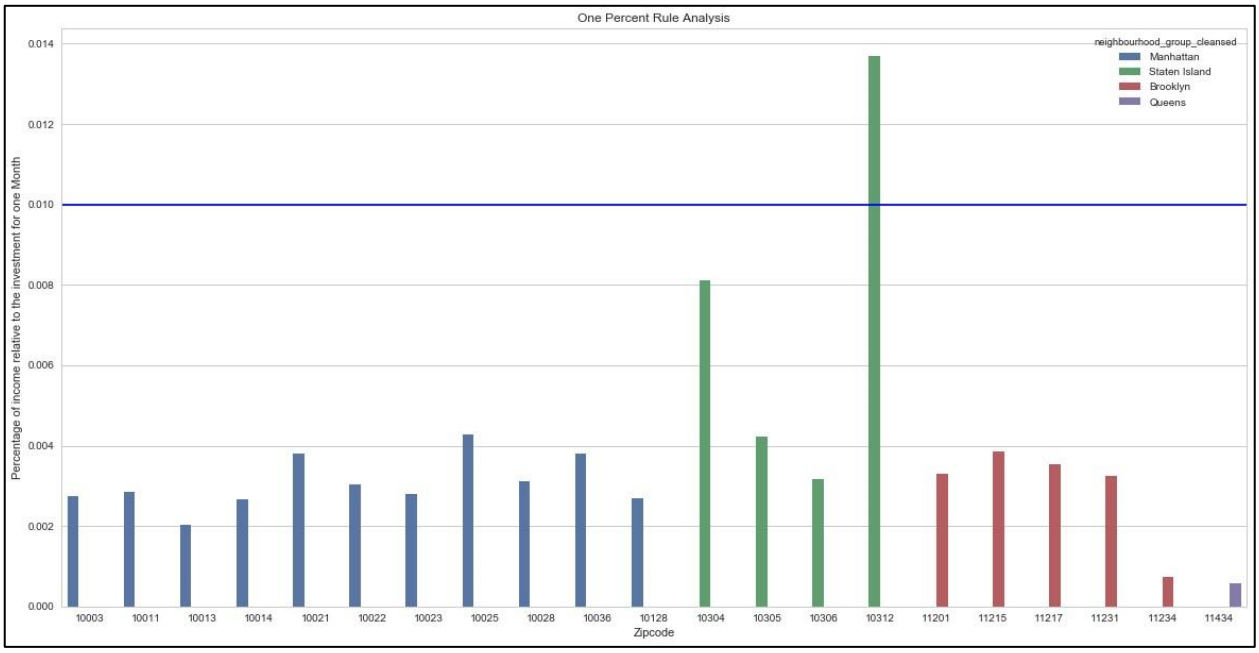


One Percent Rule

This is one of the basic calculations to see whether to invest in a property.

If an investment is not yielding you at least one percent of your investment per month, then there is no point in buying the property. But the main underlying principle behind this is that, at most of the times, people will make an investment, by mortgaging and they need to pay the interest.

This interest will at least be the one percent of the investment every month. This is the reason why one percent rule is applicable. But here, in our case, we have no mortgaging. Yet, this is still a better measure, which is addressed later.

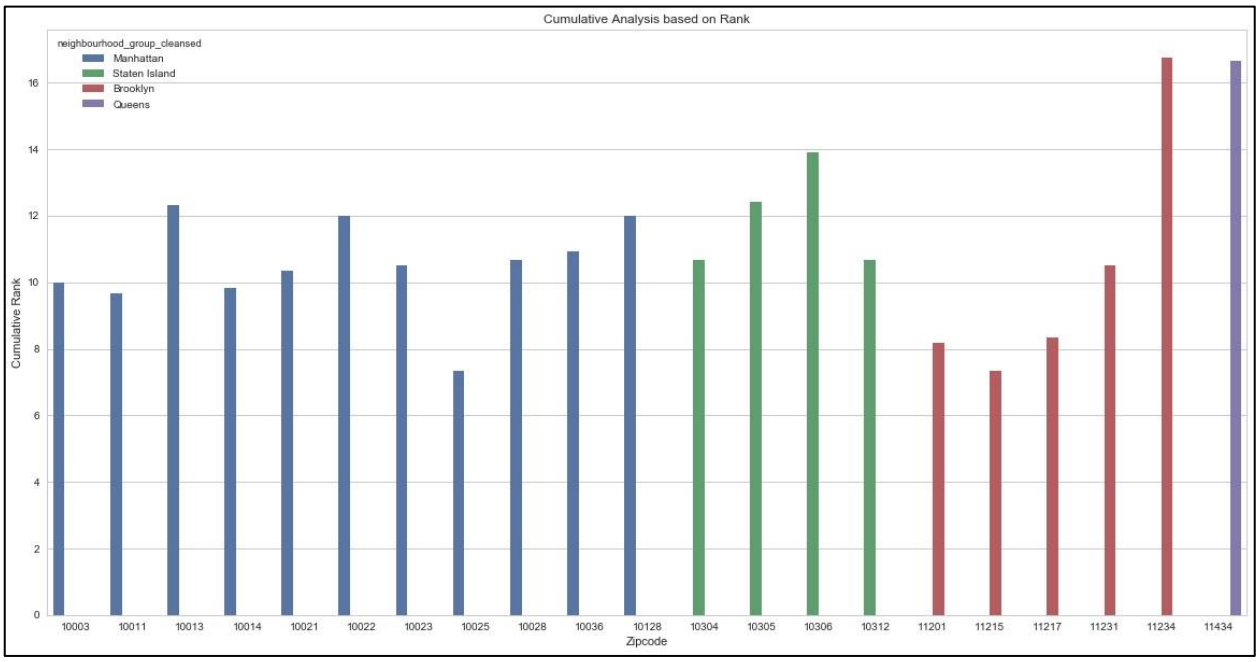


Cumulating All Analysis

Giving Rank to Zip code for all Analysis and Choosing the best location

With the above all preliminary analysis, we have got various different opinion on which could be better on different terms.

So, we need a measure to cumulate all the results to look at the most profitable investment.
Here, I have ranked the zip codes for every other analysis (Price, Cost, Occupancy, ROI, Location Reviews, Number of reviews), the best value is given the minimal rank (Rank 1), and the worst the highest.
So, when we aggregate using mean, the most minimal rank will be our winner.



zipcode	Total Rank Score
10025	7.333333
11215	7.333333
11201	8.166667
11217	8.333333
10011	9.666667
10014	9.833333
10003	10.000000
10021	10.333333
11231	10.500000
10023	10.500000
10304	10.666667
10312	10.666667
10028	10.666667
10036	10.916667
10128	12.000000
10022	12.000000
10013	12.333333
10305	12.416667
10306	13.916667
11434	16.666667
11234	16.750000

Conclusion and Recommendation

According to the Cumulative rank, we can see that the zip codes 10025 and 11215 are more profitable than any other.

But, if we are okay to take risk, it is possible to consider 10312 as well. This is because, the only place the zipcode 10312 fails is with the lack of reviews and lack of reviews about the locations.
10312 has lower investment, higher revenue relative to the investment. It has a descent value of occupancy rate as well. But the fact that it will just fail in the review genre, doesn't make it a bad investment.

10312 is the only zip code that passes in One Percent rule test as well. So, 10312 is also a zip code that the investor might want to consider.

Similarly, 10304 is another zip code, that fails terribly in the reviews. But has the highest occupancy rate. So, for the cost and revenue, relatively, this is not another option that should be ruled out. The number of properties in 10304 is very less but has higher demand. So, investing 10304 is one other options, and ruling it out just because it failed in Reviews, doesn't make sense.

So, (10025, 11215, 10312, 10304) are the possible suggested zip codes, and it is necessary to discuss with the investor as the final decision is purely dependent on amount of risk the investor is ready to take.
As said initially,

There could be two interpretation

1. Invest where not much people have ventured (High Risk, yet could be more profitable) - 10312, 10304
2. Invest where everyone has already done (Could be sophisticated with less risk and although not high profit, still profitable) - 10025, 11215

Meta Data for created data

Field	Description
Occupancy Rate	1 – (Number of days the house is available for the next 365 days/ 365) – It is for each property
Number_of_reviews_x	Sum of all the reviews given to the properties in each of the zip code
Occupancyrate_x	Average of all the occupancy rate given to the properties in each of the zip code
2017-06_x	Average of the cost at which the property is sold in each of the zip code
Price_x	Average cost for renting a property for one night in each of the zip code
Review_scores_location_x	Average of all the reviews given to the properties in each of the zip code
ROI	Gives the return of Investment rate

Breakeven	Gives the time required for getting back the investment (In years)
Onemonth	Amount of average revenue obtained from each zipcode
Onemonthpercent	Ratio of the monthly revenue to the cost of the property
Cost Price Rank	Gives the rank of the zip code in which least cost has the least rank (Least cost – most desirable) and the worst has highest rank
Revenue Price Rank	Gives the rank of the zip code in which highest revenue has the least rank (High Revenue – most desirable) and the worst has highest rank
Number of Reviews	Gives the rank of the zip code in which sum of their reviews is the highest has the least rank (High Review – most desirable) and the worst has highest rank
Location of Review Score	Gives the rank of the zip code in which sum of their reviews is the highest has the least rank (High Review – most desirable) and the worst has highest rank
Occupancy Rate Score	Gives the rank of the zip code in which highest occupancy rate has the least rank (High Occupancy– most desirable) and the worst has highest rank
ROI Rank	Gives the rank of the zip code in which highest ROI has the least rank (High ROI – most desirable) and the worst has highest rank
Total Rank Score	Gives the average rank of all the zip codes in which the least rank corresponds to the worthful investment.

Further Suggestions for model Improvisation

1. Reliability of the data - In general, we are unaware of the reliability of the data. As it is all manually entered data, they could be more error prone. We need to check for the reliability of the data. This could be done by collecting more data, from various other sources.

2. Daily rents of the property - The rental details for the property, are given to be a constant value. But usually, the prices aren't stable and so they fluctuate. For example, over any college breaks, or it could have seasonal dependency. So, the price tends to change.

3. Buying Cost of the property - Time series forecasting can be done based on the available data to determine exactly what would be the actual purchasing cost instead of assuming that they are being brought at the cost that was given in the last known period.

4. Selling Price of the property after the break-even period - If we have more rhetorical data, we can do time series forecasting, to predict what would be the selling price (If they wish to sell it after the break-even period) - Higher the selling price more would be better. And here, the necessity for more rhetorical data is because - the break-even period is approximately 15 years, and using 10 years of data to predict the next 15th year won't be accurate and reliable

5. Area dependency for the price - We do not have any information about the area of the property (It has got 40186 missing data's). If we are able to get that data, then we can analyze the price based on the area dependencies. Same zip code with higher area will have higher rent. And of it is available it is possible to categorize the same zip code based on area and analyze. For example, instead of taking the average of prices of the properties in each zip code, we can categorize like (0-500 sq ft in one category, 500-1000 sq ft in another and so on). So apart from knowing where to invest, also we will know whether to invest in a bigger property

6. Occupancy - Here, we have used the occupancy rate based on the availability for the next 365 days. But the occupancy rate might just not be dependent on that. We need to have other factors to estimate the real occupancy rate. We can collect data from other sources to have more accurate Occupancy Rate

7. Host dependency - From the different genres of the data of Airbnb, we have completely neglected the information about the host, because we are not going to deal with these hosts directly. But the price and occupancy might be highly influential on the categories of data about host. So, we need to use some measure to remove the dependency of the host on price and occupancy rate

8. Heat map of occupancy rate - If we are able to visualize the occupancy rate on the map of NYC, we can visualize if there are anything around the locality that is highly influencing the occupancy rate. The place around it could have a great scenic view, or there could be high occupancy in downtown.

9. Heat map of number of Airbnb properties - Similar to the above mentioned, this will help us to visualize if there are anything specific to the surroundings that is influencing people to have more properties around them.
10. Using Latitude and Longitude to determine the missing zip codes and missing zip codes - Although there are only around 720 missing and incorrect zip codes, replacing it with the help of latitude and longitudes might affect the model and it might help in more accurate visualization

11. Other possible costs need to be considered - We are just considering the revenue that we can get and not considering the possible expenses like maintenance or Property tax etc. Without this the visualization is never accurate. So, we can collect data on this and use this real time to make more accurate analysis