Capital One Data Challenge

Present

By

Chun-Wei Lo

January 14, 2019

Executive Summary

The objective of this report is to find hidden insights from publicly available data of AirBnB and Zillow and further provide recommendations on which zip code have the highest return ratios within New York City and what are the most significant factors contributing to the revenue. Our goal is to help develop a data product and tailor client's data strategies, whose part of its business model relies on short term rentals, to the market demand. Data visualization and predictive statistical models would be employed in the report to better answer our business questions.

1. Introduction

1.1. Business Problem Statement

A real estate company has a niche in purchasing properties with two-bedroom to rent out short term as part of their business model specifically in New York City. They want to learn more about which zip codes are the best to invest in and build a data product with a rational conclusion to find out where their targets are and what factors they should consider when purchasing properties within New York City.

1.2. Dataset Description

The publicity available data from Zellow and AirBnB would be used for our analysis.

- **Zillow Data:**

  Cost data to provide us an estimate of value for two-bedroom properties

- **AirBnB Data:**

  We can treat the data from AirBnB as our revenue data that contains information about properties located in Greater New York area, such as zip code, price per night for each listing, and geographic information.

1.3. Assumption

In order to focus our time and energy on the most critical issues, keeping the following assumptions in mind at this point is important.

1. The investor will pay for the property in cash (i.e no mortgage/ interest rate will need to be account).
2. The time value of money discount rate is 0% (i.e $1 today is worth the same 100 years from now).
3. All properties and all square feet within each locale can assumed to be homogeneous.
4. The occupancy rate is based on the review of location score in AirBnB data.

2. Data Cleaning and Preprocessing

The importance of data integrity would never be overemphasized. Data preparation accounts for 80% of the works of analyst or data scientist. The tasks involve cleaning and organizing data, and we rely heavily on quality data to conduct data analysis. The better the data quality, the more confidence users will have in the output they produce. We will discuss how we preprocessed the cost and revenue data respectively in six data quality dimensions: Accuracy, Integrity, Cleanliness, Correctness, Completeness, and Consistency.

In this section, we would discuss how we deal with messy data of the Zellow and AirBnB data respectively. The scope of data preprocessing involves feature selection, missing value imputation, and data cleaning.

2.1. Revenue Set: Zellow Data

There are approximately 9,000 rows with 262 columns in the raw file. We dropped the properties outside New York City and only keep 25 rows with historical property price information for 2 bedrooms.

Features of our interest	Data Preprocessing
Region Name	1. Rename it to 'zipcode' to line up with cost dataset 2. Unify zip code to 5 digits 3. Check the uniqueness of zip code
City	Drop the properties outside of New York City
Time Series Data	Only keep the historical median price columns for past ten years as we'd like to see time series pattern and trend in ten years (2007/07 – 2017/06)

2.2. Cost Set: AirBnB Data

The AirBnB data set contains information on the listing including location, number of bedrooms, review score and price per night, etc.

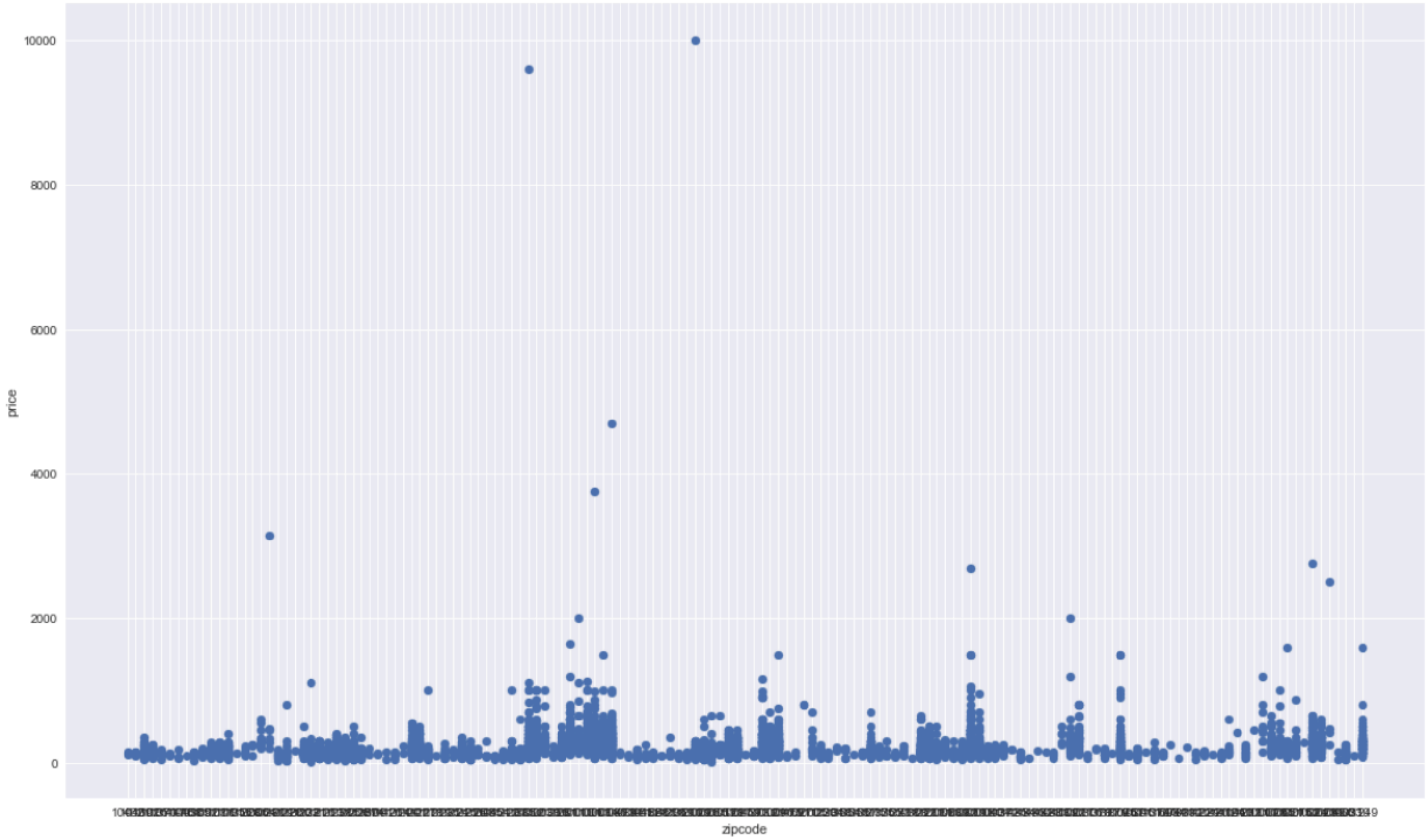
Features of our interest	Data Preprocessing
Zip code	1. Unify zip code to 5 digits 2. Convert latitude and longitude to zip code for missing zip code 3. Check the uniqueness of zip code
Price	Conduct outlier analysis and drop the listing with unusual price with the measure of IQR
Square feet	Remove the listing with unusual size of properties
Bedroom	Select the desired bedroom number only (bedroom = 2) aligned our analysis with client's interest and the niche market

2.2.1. Outlier Analysis

A critical part of data preprocessing is to identify outliers: the observations that are distant from other observations. An outlier can affect the central tendency of a dataset by skewing the results so that the mean is no longer representative of the dataset. In order to detect an outlier, a scatter plot can give us a taste of what they look like.

- **Price**

A scatter plot is the collection of points that show values for two variables. As we can see from the scatter plot below, some of the listing price may look unreasonable since they are too far away from most of the observations. For example, we may doubt the listings with price greater than $2,000 per night may be a mistake or that the hosts input the wrong number when the post their listings.



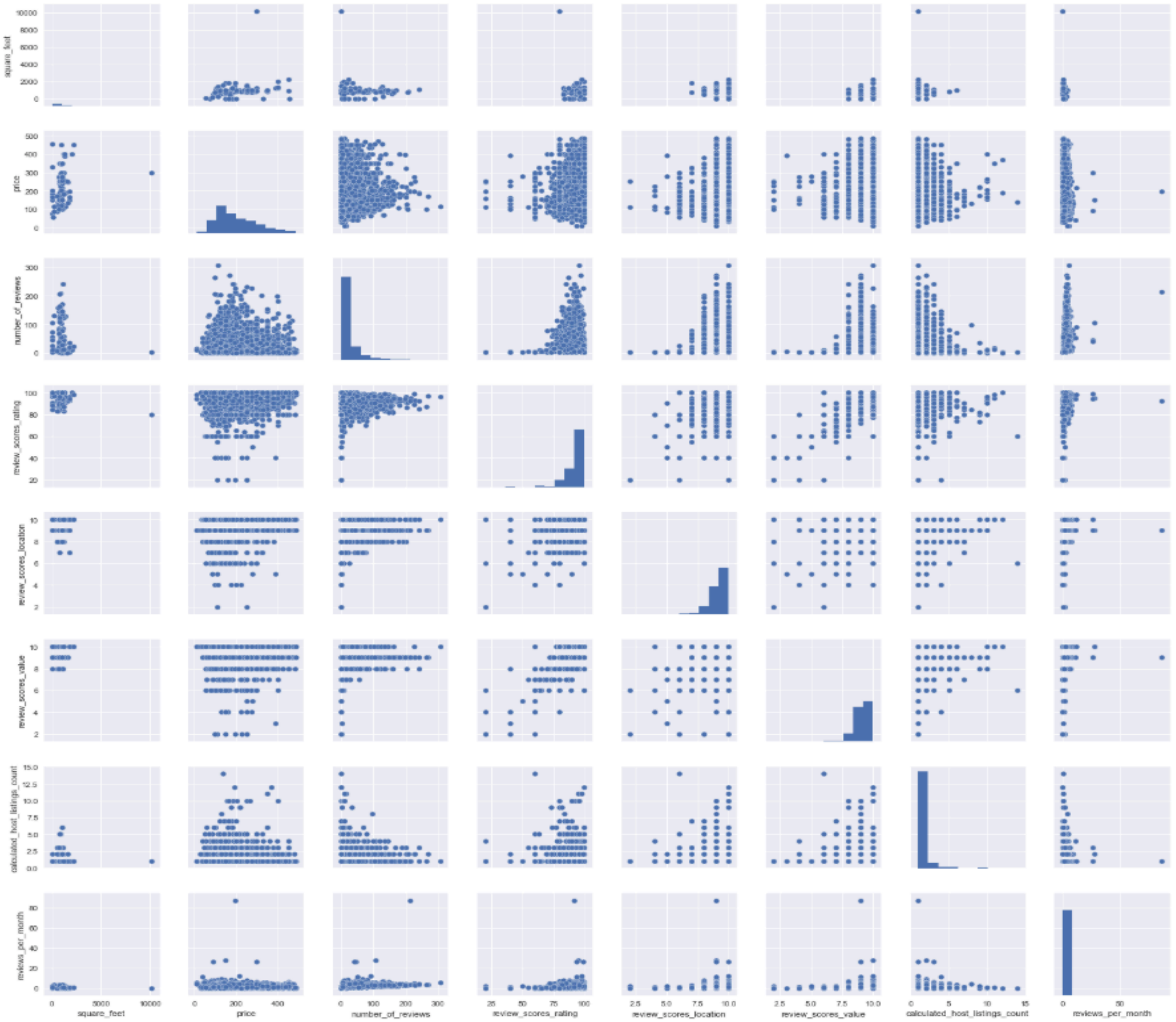
For dealing with outliers, the interquartile range, a commonly used measure of statistical dispersion, is employed to help us remove those abnormal observations. The listing with price that is not in this range would be dropped. We dropped 230 observations from our raw set and there are 4,634 listing left.

$$IQR = \ Q3\ (\ 75th\ percentile) - Q1\ (25th\ percentile)$$

- **Square Feet**

We further built pairs plot to see if any other numerical feature can help us identify outliers. The pairs plot builds on two basic figures, the histogram and the scatter plot. The histogram on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship between two variables. For example, the left-most plot in the second row shows the scatter plot of price and square feet.

It is obvious from the first figure below that there is one point is far away from most of the observations in almost each plot. We further eliminated one listing with square feet of 10,118.



2.3. Exploratory Data Analysis (EDA)

After we removed outliers that may be affecting the quality of our analysis, the next step is to conduct Exploratory data analysis (EDA) to see the main characteristics of the features of our interest and see their relationship between each other.

Exploratory data analysis (EDA) is an approach to analyze date set and summarize their main characteristics with visualization methods. We used correlation matrix to see the relationship between variables and response variables, bar graphs to check the distributions of features, and box plot to enable us to study the distributional characteristics across several groups as well as the level of the scores.

In this section, we will first conduct EDA for revenue and cost dataset separately and then consolidate them into one set to see a bigger picture of the relationship between features in cost set and them in revenue set.

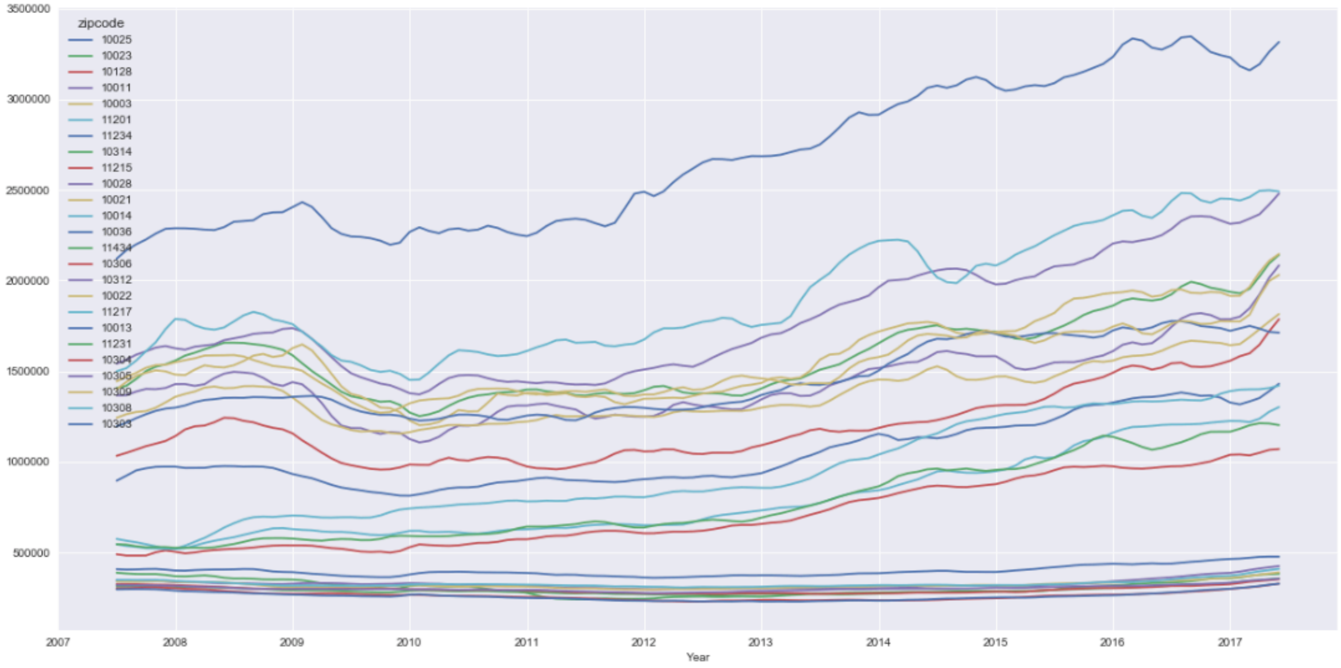
2.3.1. Revenue Set: Zellow Data

2.3.1.1. Time Series Plot

Though we have 20-years historical median price of properties in New York City, the historical data is no complete and there is a substantial amount of missing values in the data prior to 2000. Therefore, we extract the past 10-year median price data and generate a time series plot to see the pattern and trend in the properties price.

The price of properties across all zip codes dropped significantly after 2009. We assumed that it happened due to the real estate market crash after financial crisis in 2008.

However, the property price started to climb in 2012. Specifically, the properties in 10013 has the highest median price. 10013 and 10014 take the second and third place respectively. What's interesting finding from the diagram is that there is a slight decrease in the property price starting in middle 2016.

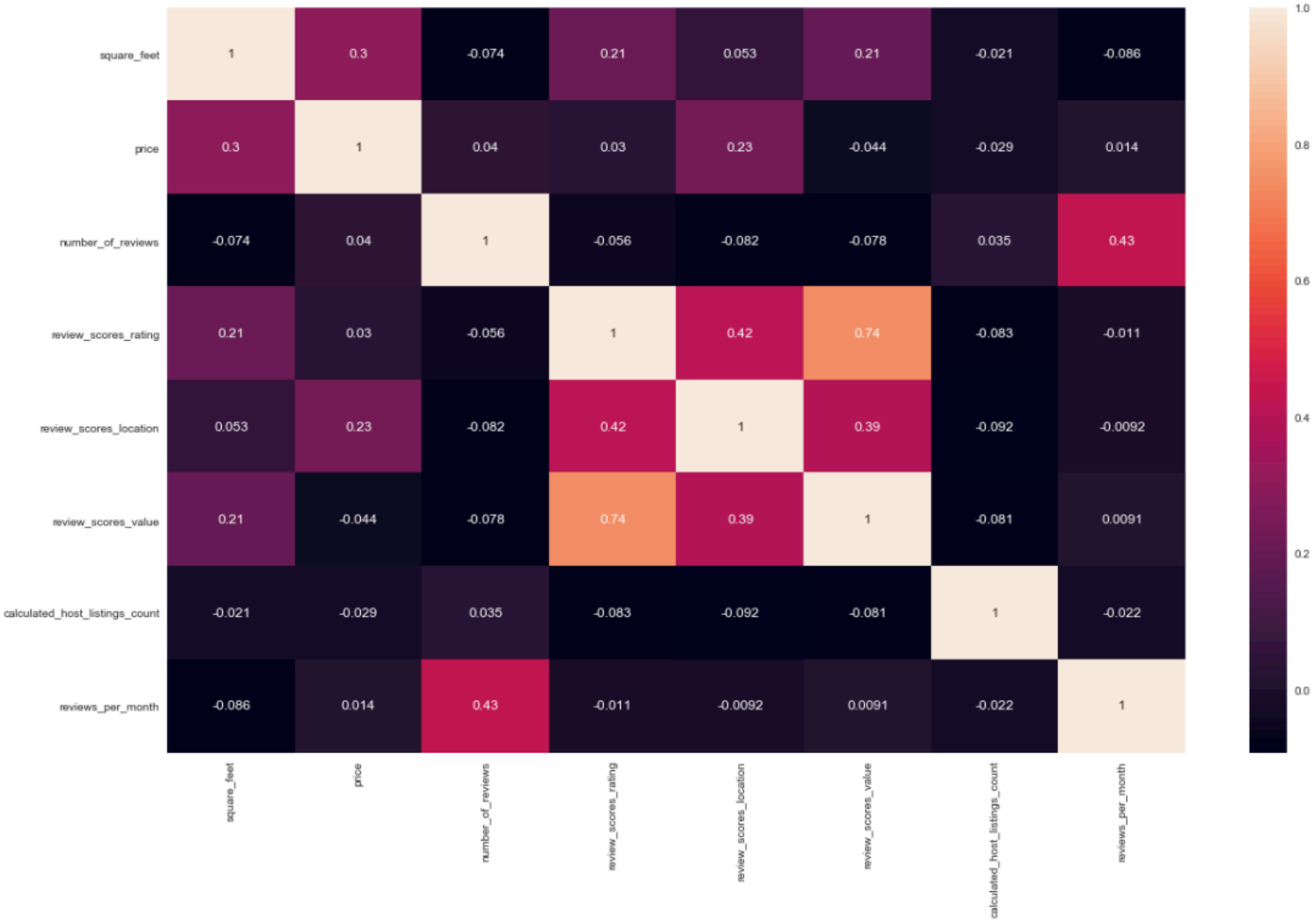


2.3.2. Cost Set: AirBnB Data

2.3.2.1. Correlation heatmap

Firstly, one of the main questions that comes to our mind is which features may have impact on the listing price? From the heatmap, we can observe that *square_feet* and *review_scores_location* have positive relationship with *price* while *review_score_value* and *calculated_host_listings_count* have negative correlation with *price*.

It is interesting to know that *number_of_review* and *review_scoses_rating* seem to be relatively less influential to the listing price.

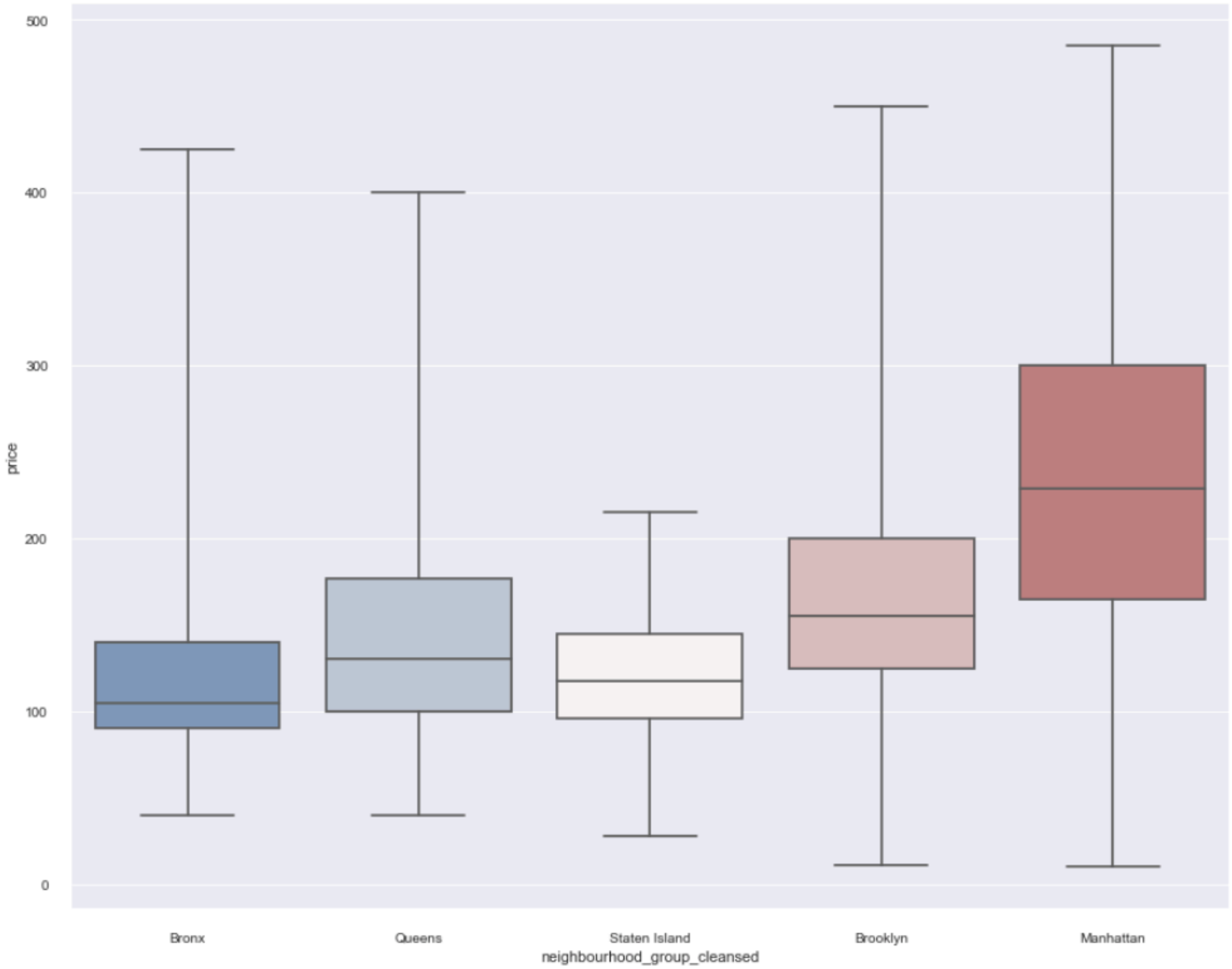


2.3.2.2. Box plots

- Price versus Neighborhood Group

Box plots are used to show overall patterns of response for a group. They provide a useful way to visualize the range and other characteristics of response for a large group. The box plot below suggest that Manhattan neighborhood has the highest median rental price as well as the largest variation in listing price.

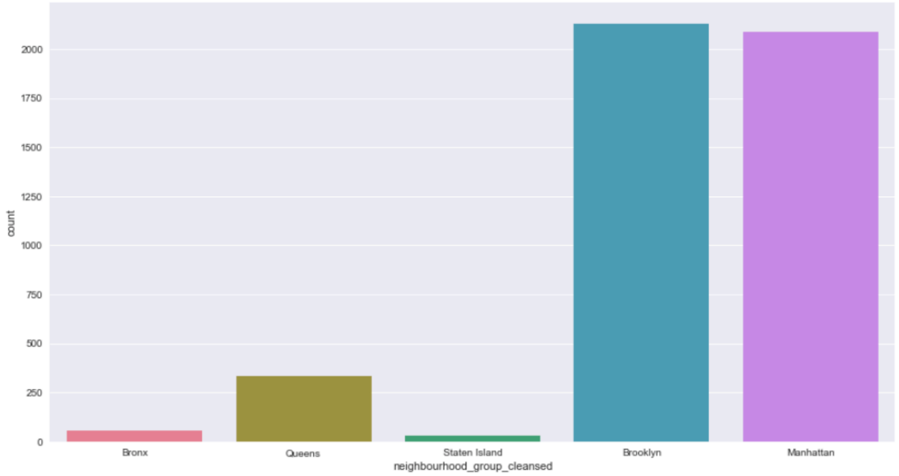
Bronx has the lowest median rental while its variation in price is relatively high among the group.

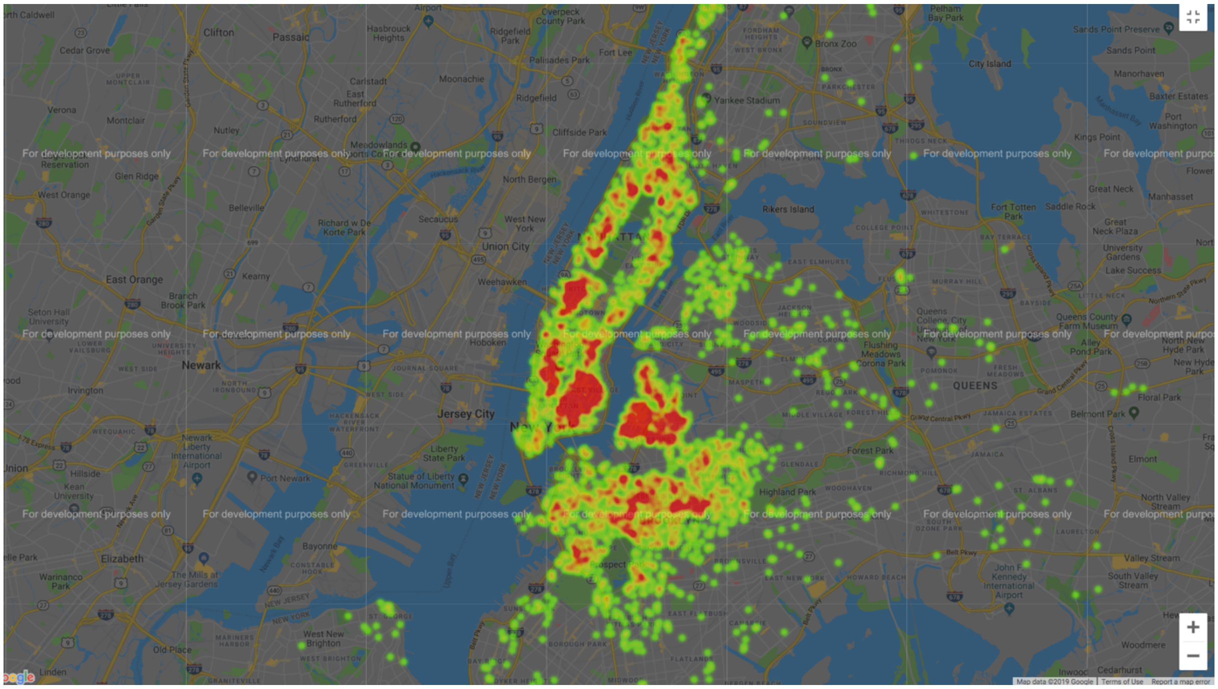


2.3.2.3. GIS Map

We further used gmaps package in python to visualize which area has the most supplies in short-term rental geographically.

The more density the area has, the more listing counts there are. Based on the heat map, it is noticeable that Manhattan area and Brooklyn have the highest density. We can conclude that the hosts in these two areas are in a more competitive market than those in Queens or Bronx.





3. Revenue Cost Analysis

After having an overview of how our revenue and cost data look like, it is necessary to combine these two sets into one master dataset to further do feature engineering. The master dataset was generated by left joining revenue data to cost data based on zip code.

Before we jump into revenue cost analysis, we first need to define how we calculate the metrics used for it.

3.1. Metrics Definition

3.1.1. Expected Return

Expected return is the net income during a period. We calculated the expected return based on such formula.

$$Expected\ Return = Daily\ Listing\ Price\ \times\ Time\ (365\ days)\ \times Occupany\ Rate\ \times 97\%$$

Since AirBnB charges hosts a service fee of 3% to better improve its platform, we need to take that into account in calculating the expected return.

3.1.2. Total Cost

As we mentioned earlier, AirBnB charges hosts a 3% host service fee every time a booking is completed on their platform. We should treat this amount of money as our variable costs.

$$\begin{aligned} Total\ Cost = {} & Property\ Price \\ & + (Daily\ Listing\ Price\ \times\ Time\ (365\ days)\ \times Occupany\ Rate\ \times\ 3\% \end{aligned}$$

3.1.3. Occupancy Rate

The review score of a property has a great impact on occupancy rate. Since there is no real occupancy rate for each property, we took the data from Mashvisor, which is a well-known real estate platform offering metadata in real estate market and build an occupancy rate table based on the *review_scores_location* from revenue data.

Location Score	Occupancy Rate
9.5-10	80%
8.5-9.5	75%
7.5-8.5	70%
6.5-7.5	65%
<6.5	50%

3.1.4. Return on Asset Ratio

Return on Asset Ratio (ROA) is a profitability ratio that measures the net income generated by the total assets in a certain period by comparing net income to the total assets.

$$Return\ on\ Asset\ Ratio\ = \frac{Expected\ Return}{Total\ Assets\ (Cost)}$$

3.1.5. Break-Even Period

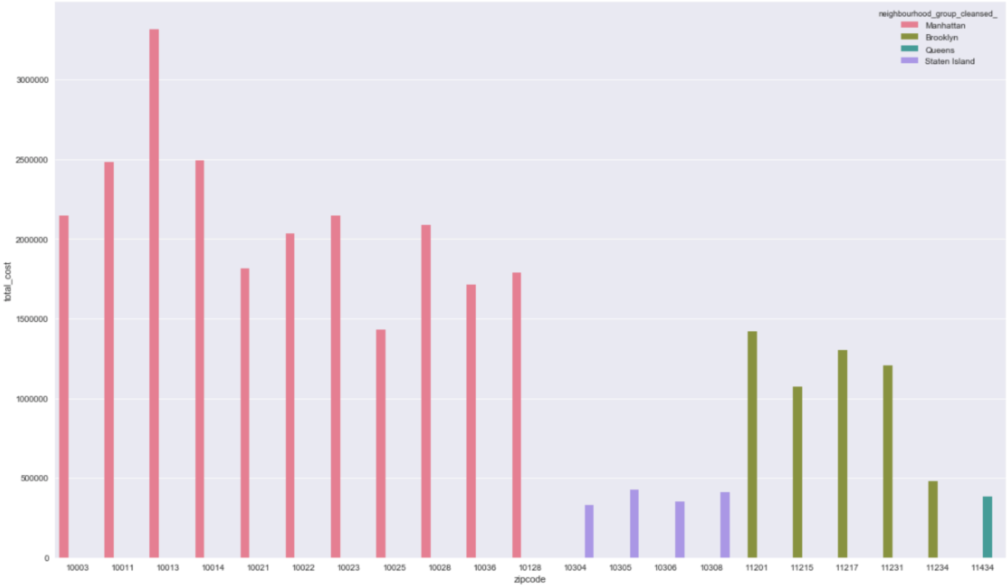
To better understand how long it will take to break even when we buy a property within New York City, we need to decide the break-even point. The definition of the break-even point is the point where total revenues equal to the costs. Since one of the assumptions we have are 0% discount rate and pay the property in cash, we can simply our formula as follow:

$$Breakeven\ Period = \frac{1}{Return\ on\ Asset\ Ratio\ (ROA)}$$

3.2. Return on Asset Ratio and Breakeven Period Analysis

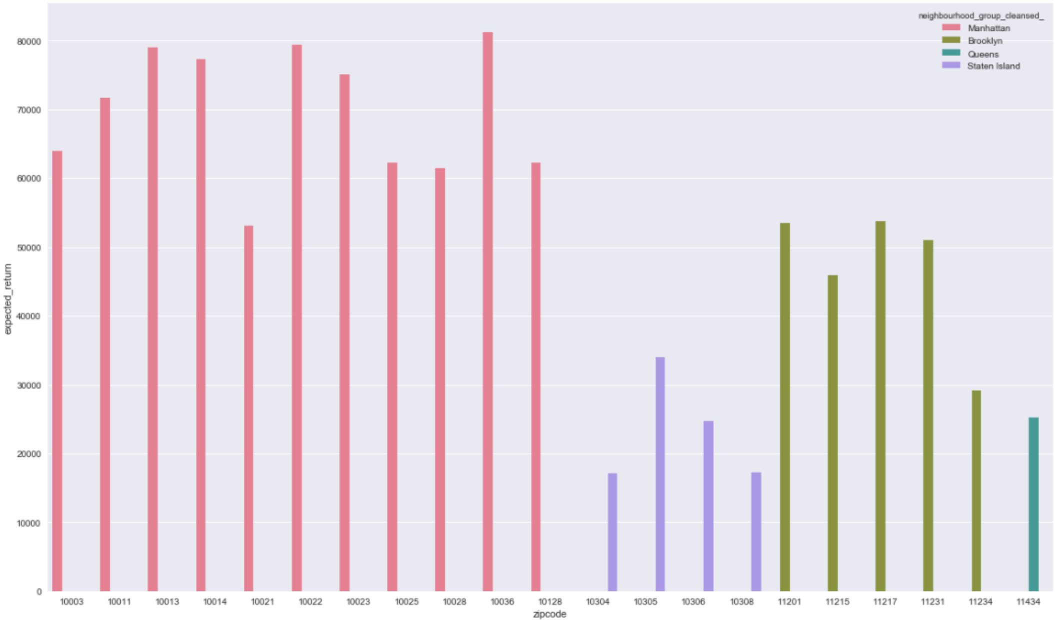
3.2.1. Property cost by zip code and neighborhood

In general, we can find that the median cost of properties in Manhattan is the highest on average among the group, whereas the cost of properties is the lowest in Staten Island. 10013 has the most expensive properties.



3.2.2. Expected Return by zip code and neighborhood

When it comes to expected return, the similar pattern can be found as that of total cost: the expected return of properties in Manhattan is the highest whereas Staten Island has the lowest expected return. However, when we narrow down our lens into zip code, it is surprising that 10006, whose total cost of property is not the most expansive compared to its counterparts, takes the first place in this category followed by 10022 and 10013.



3.2.3. Breakeven Period Analysis

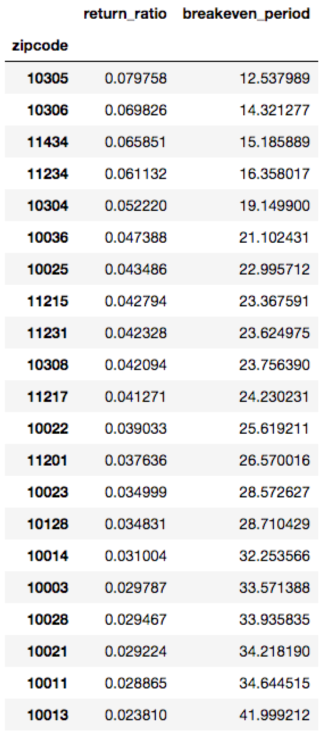
In this section, we'd like to compare return ratios and breakeven points across zip codes within New York City and to know how much time it needs to take to pay off the cost of properties.

The following table shows both *return ratio* and *breakeven period*. In general, properties in Staten Island has the highest return ratios and shortest breakeven periods on average, particularly10305 only needs twelve and a half year to break even.

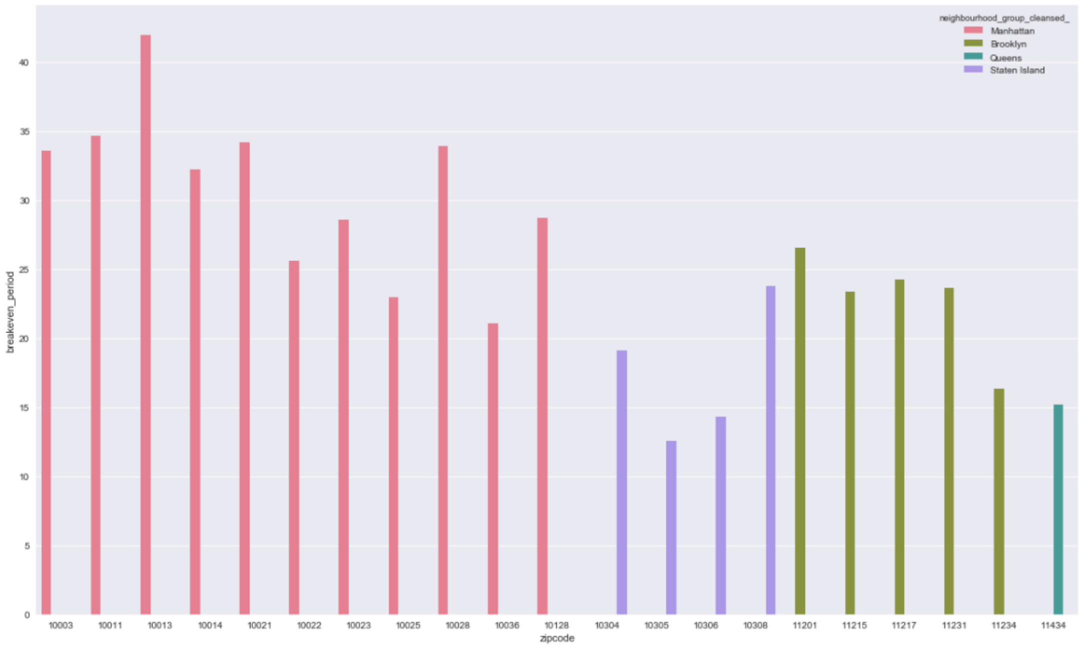
However, the reason for the short break-even points may have something to do with its low property costs. We are not able to determine if they are good targets without detailed information on market demand in that area.

On the other side, when we take a look at properties in Manhattan and Queens where hosts are in relatively fierce competition real estate market, it is worth noting that there are a couple zip codes seem to be good targets, such as 10025 (Manhattan), 10036 (Manhattan) and 11234 (Brooklyn) with breakeven periods of 15 to 21.

Considering the demands in short term rental and popularity, these three zip codes can be added to our list.



	return_ratio	breakeven_period
zipcode		
10305	0.079758	12.537989
10306	0.069826	14.321277
11434	0.065851	15.185889
11234	0.061132	16.358017
10304	0.052220	19.149900
10036	0.047388	21.102431
10025	0.043486	22.995712
11215	0.042794	23.367591
11231	0.042328	23.624975
10308	0.042094	23.756390
11217	0.041271	24.230231
10022	0.039033	25.619211
11201	0.037636	26.570016
10023	0.034999	28.572627
10128	0.034831	28.710429
10014	0.031004	32.253566
10003	0.029787	33.571388
10028	0.029467	33.935835
10021	0.029224	34.218190
10011	0.028865	34.644515
10013	0.023810	41.999212



4. Conclusion & Recommendation

4.1. Conclusion

- Although Staten Island has shortest breakeven periods and highest return ratios, such as 10306 or 11035, the demand for that area is uncertain and we cannot determine if they are good targets until we have sufficient information on the actual occupancy rate of Staten Island.

- 10025 (Manhattan), 10036 (Manhattan) and 11234 (Brooklyn) also performed well in breakeven periods considering their relatively higher return ratios and lower risk of lacking market demands.

- Location and the size of properties seem to be great factors influencing the price of listing from the correlation matrix.

- Market in Manhattan has the fiercest competition in terms of housing price according to the GIS heat map and property cost bar chart. However, we still can find some good investing targets that have significant return ratios.

4.2. Recommendation

- There are a couple of metrics that play critical roles in the price of listing are being simplified in our analysis, including occupancy rate and review score of location. With more detail data in these significant factors can make our analysis more robust.

- We didn’t touch on any predictive models in the analysis while there are lots of opportunities for us to apply predictive model, such as building pricing model for listing and time series model to better know the pattern of property costs.

- Incorporating with people who have intimate domain knowledge and have experiences in finance can also help us with feature selection, interpretability of model and any other relevant factors.

4.3. Tools

Python is the main tool in our analysis. Here are the packages used for data cleaning, data manipulation and data visualization.

Python package	**Application**
pandas, numpy	Data Cleaning, Data Manipulation, Data Aggragation
matplotlib, seaborn	Data Visualization
geopy	GIS heat map

5. Reference:

[1] https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
[2] https://www.mashvisor.com/blog/breaking-even-real-estate-investment-property/
[3] http://www.knowledgedynamics.com/demos/Breakeven/BreakEvenFormula.htm
[4] https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/