NYC REAL ESTATE INVESTMENT ANALYSIS

—— Capital One Data Challenge

Xinran Tao

Applicant for Senior Data Analyst

12/07/2018

Table of Contents

Executive Summary

In this report, I conducted analysis of a real estate investment case to answer the question of which zip codes would generate the most profit on short term rentals within New York City. This analysis report consists of five sections:

- Introduction and Business Context: Define the business problem and objective
- Data Preparation and Quality Check: Introduced the steps to check the validity, accuracy, completeness and uniqueness of our data, engineer variables with problems, select relevant variables and observations and merge datasets by common zip code
- Model Methodology: Break down the problem and lay out analysis plan, make assumptions, determine the metric (Rent-to-Price Ratio) to measure investment profitability
- Model Result: Generate business insights from the cost perspective (property price), the revenue perspective (rent price) and the profitability metrics and provide suggestions for zip code investment
- What Is Next: Some potential further analysis can be performed if sufficient data provided
    - Time Series Analysis and prediction for long-term investment
    - Rent price prediction machine learning model
    - Competitiveness Analysis

1. Business Context

Our client is interested in the investment of two-bedroom properties to rent out short-term in New York City. The objective is to find out the most profitable zip code in NYC.

2. Data Preparation & Quality Check

2.1 Dataset Overview

The datasets consist of the cost data (Zillow dataset), the revenue data (Airbnb dataset), and NYC zip code list. Zillow dataset and Airbnb dataset are provided by our client. NYC zip code list are scrapped from USPS.com.

2.2 Data Quality Check and Data Engineering

There are several dimensions for data quality check. The most common six criteria are:

- Validity: The syntax (format, type, range) conforms to its definition.
- Consistency: The data provided is in accordance with original data source.
- Accuracy: The value correctly describes the real world object or event.
- Completeness: Data is complete, no blank values.
- Uniqueness: No records or features will be duplicated.
- Timeliness: The data was collected timely, not delayed.

Since the two datasets are provided by our client, we can assume the data is consistent with the real data in the Airbnb and Zillow datasets and they were also collected timely. So I will check data quality according to these four dimensions: **Validity, Accuracy, Completeness and Uniqueness.**

We also need to check the relevance of the columns and observations. Because our analysis is based on the investment of the two-bedroom property with zipcode in NYC, so the data should be related to this aspect.

For **Zillow Dataset (Zip_Zhvi_2bedroom.csv)**, we checked data quality and engineered variables in these steps with the corresponding insights:

1. Data Types: The data type correctly reflect the definition of the variables.
2. Duplicates: There are no duplicated rows or columns found. The data is **Unique**.
3. Data Values:

a. I firstly categorized variables into categorical, continuous and discrete variables according to their data types.

b. Then I checked unique values for categorial and discrete variables and checked descriptive analysis for continuous variables.

Note: Because we only care about the data in NYC, also some categorical variables have too many unique values. It would be computationally expensive to check them all. So I just checked the 'state' values and values of other variables within NY State for efficiency.

Result: The values are in the normal range which correspond to the real world situation. So I can conclude the **Validity** and **Accuracy** of this dataset.

4. Missing Values:

a. Since there is no need to care about the missing data whose zip codes are not in NYC, I firstly subset the dataset by choosing the zip code in NYC. I used USPS provided zip code list to make sure the accuracy and authority of the zip code I finally chose.

b. I calculated the missing number and missing ratio for each variable and also arranged them in descending order.

Result: There are only missing values for the earlier years variables (1996-04 ~ 2007-05). Considering the remaining data has unique zip code for each record, it's not good to delete the records with missing values, so I delete the columns with missing values and keep the latest 10 years data, which are adequate for our analysis. The new dataset now has **Completeness**.

5. Outliers: To make sure our analysis result is more representative and unbiased, I checked outliers for the dataset. I detect outliers by drawing box plot, but no outliers found.

The final engineered data frame is named ***zillow_clean***.

For **Airbnb Dataset (listings.csv)**, we checked data quality and engineered variables in these steps with the corresponding insights:

1. Selecting Relevant Columns:

a. There are too many useless columns within the dataset according to the variable description, which are not helpful for our analysis. So it's better to choose only relevant columns from them in the first step.

b. I selected columns and categorized them into the id column, the zipcode column, property location columns, property feature columns, rent price columns and date columns, which are relevant with our analysis.

2. Data Types: 'zipcode' and 'price' has wrong data types because of the special characters within the value. I fixed this by using user-defined functions to delete special characters or extract strings we need.

3. Duplicates: There are no duplicated rows or columns found. The data is **Unique**.

4. Data Values:

a. I also categorized these variables into categorical, continuous and discrete variables according to their data types.
b. I check unique values for categorial and discrete variables and checked descriptive analysis for continuous variables.

Results: Some of the categorical variables are not consistent with the other related variables. For example, a record with location city in New York has 'country' value in Uruguay. Also, in 'State' and 'City' column, the values are very messy. There are several formats to represent the same state or city, making it hard to impute. So I chose to drop these columns to ensure **Validity and Accuracy**.

5. Select Observations with Special Condition: Selected observations with zipcode in NYC using USPS provided zip code list. Select the property which have two bedrooms and are rent in whole home/apartment.
6. Missing Values: Some variables have more than 78% missing ratio, so I just delete these columns to ensure **Completeness**.
7. Outliers: I checked outliers for the dataset by drawing box plot and found it has outliers with larger values. So I used the *mean + 3 * std* as the upper bound to count number of outliers and ratio in total observation. The ratio is less than 0.01 so I just delete them.

The final engineered data frame is named ***airbnb_clean***.

2.3 Aggregating and Merging Datasets

- I aggregated the cleaned airbnb data by zipcode using aggregation function of mean() to generate the mean rent price per zipcode. The processed data frame is named ***airbnb_agg***.
- I merged airbnb_agg and zillow_clean by common zip code and get 22 records. The data frame is named ***df***.

2.4 Visualization of the Processed Datasets

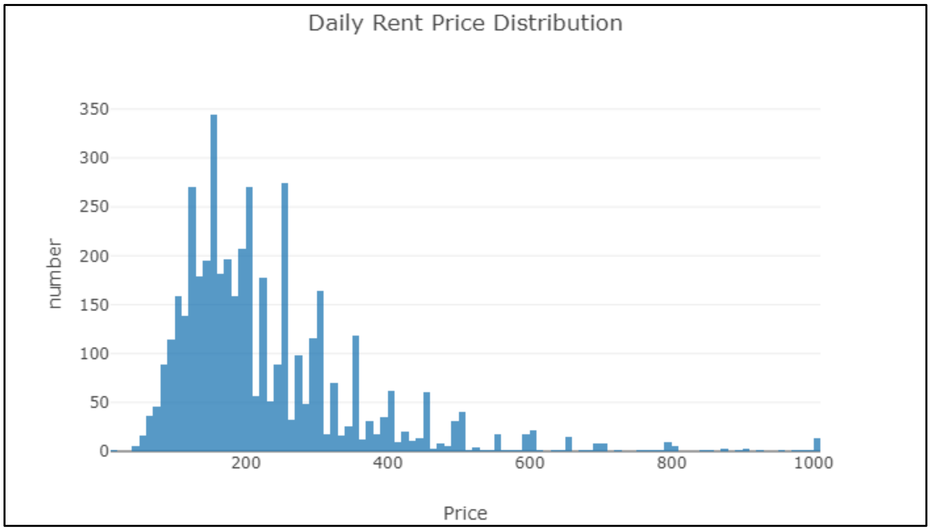

Figure 1 NYC Daily Rent Price distribution

Most rent price are within roughly $80-$300 range. For some properties, it could be as high as $1000 per day.

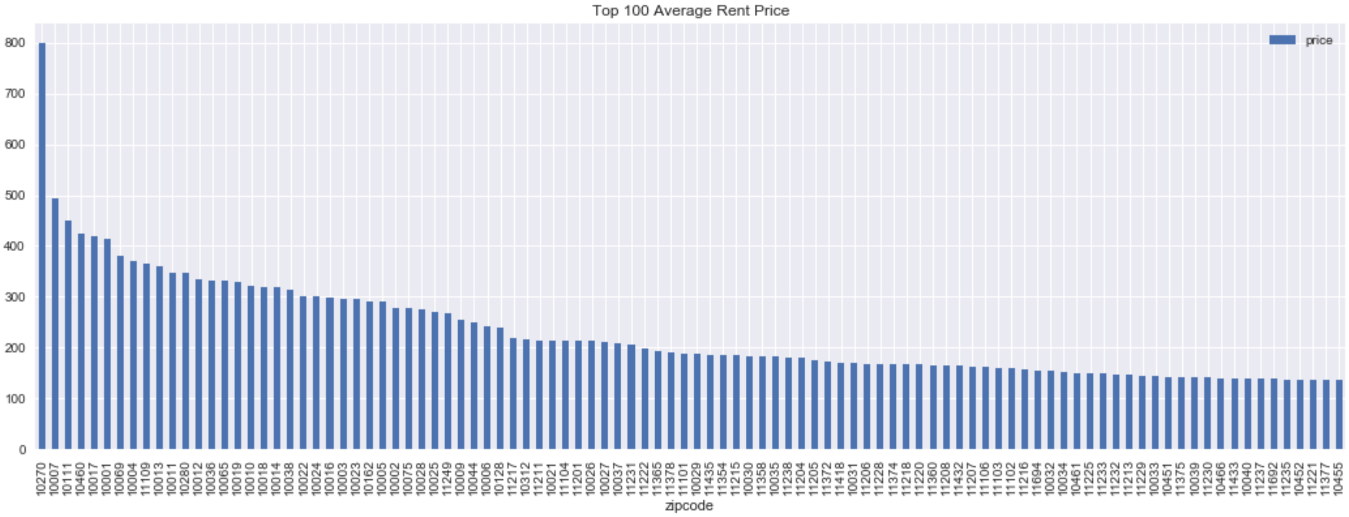

Figure 2 NYC Top 100 Zipcode with Highest Average Daily Rent Price

The rent price of the top 100 zipcode are all above $100 per day. The highest rent lies in zipcode 10270, the place of the Wall Street, which makes sense why the rent is so high. The second highest zipcode in 10007, which is the area of Tribeca, a well-known upper-class district. The third is 10111, a place lies near the Time Square and the Rockefeller Center.

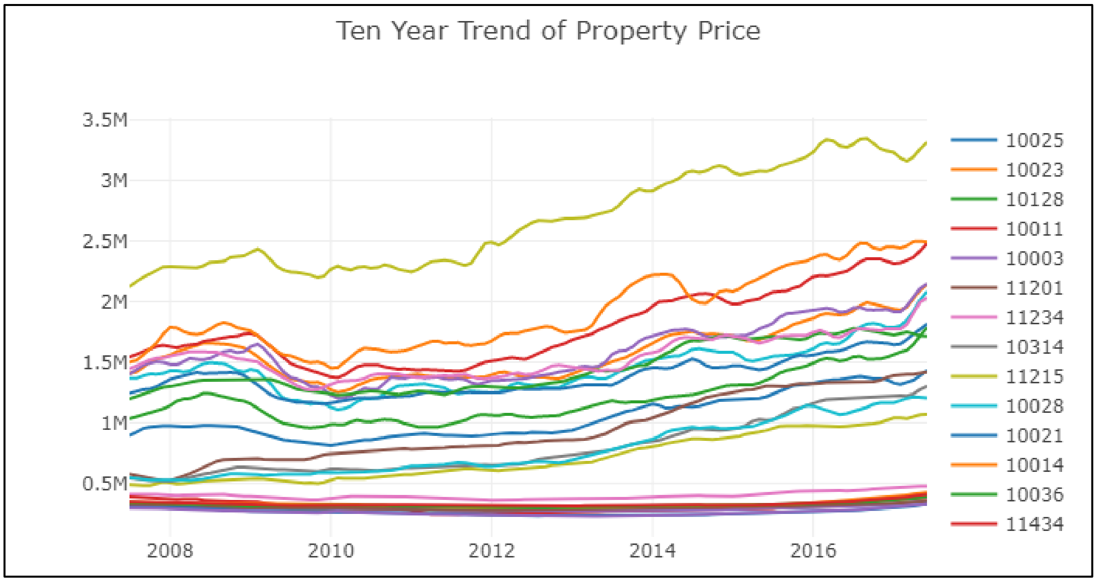

Figure 3 Ten Year Trend of Property Price for all Zipcode

The property price is increasing over the years and for some zipcodes with relatively low property price, the increase seems slower. Notice that there is an obvious decrease around the year of 2010, because of the influence of subprime crisis through 2007 – 2009.

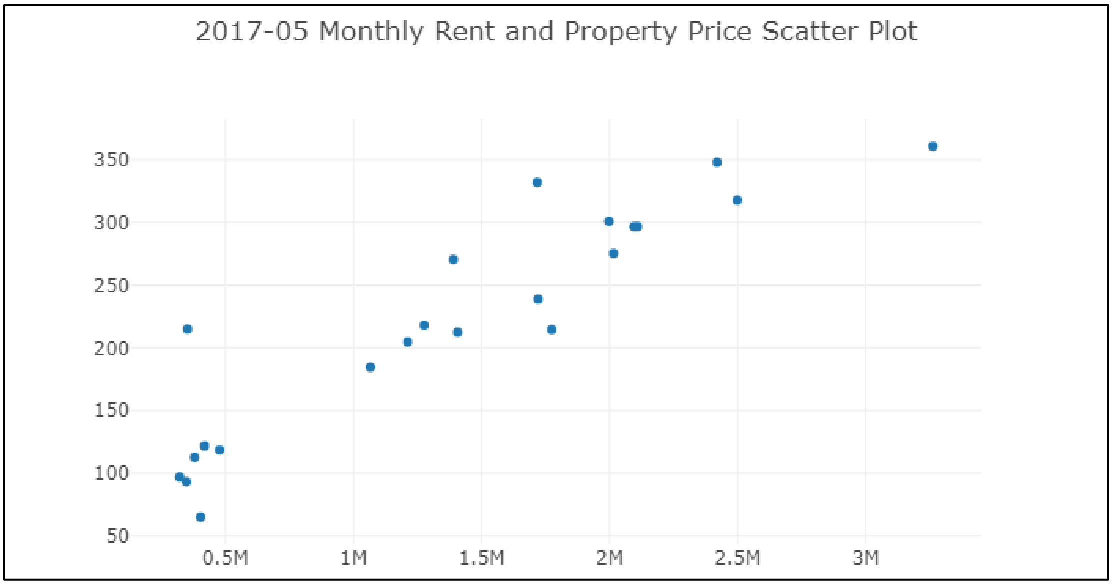

Figure 4 2017-05 Monthly Rent and Property Price

We can see the property price and monthly rent have a positive relationship. Typically, higher property price will have higher rent price.

3. Model Methodology

3.1 Problem Beak-down

Firstly, we need to measure the profitability and define a metric. Because it is short term investment, also the data range we have for the revenue data is in one month 2017-05, the cost data are in 10 years (2007-07 ~ 2017-06) with future data unknown. So we just need to calculate the metric in one month 2017-05 and rank the metric. Then we can get the top 5 zipcodes to bid for our client's investment.

Secondly, I will analyze the property price of the target zipcodes in the long term range to see the potential of the property appreciation within that zipcode area so that we will know the trend of the investment cost.

Thirdly, I will analyze the rent price of the target zipcodes and combine with other demographic factors to see the potential of future rent.

Finally, I will select the zip code with the highest investment worth based on the analysis above.

3.2 Assumptions for Analysis

The first four assumptions come from the client (*AirBnB_Zillow - Data Challenge v6.doc*):

- The investor will pay for the property in cash (i.e. no mortgage/interest rate will need to be accounted for).
- The time value of money discount rate is 0% (i.e. $1 today is worth the same 100 years from now).
- All properties and all square feet within each locale can be assumed to be homogeneous.
- The renting occupancy rate in NYC is 75%.

The other assumptions for the analysis are as follow:

- The renting price and property price are in accordance with the date it shows with no delay or mistake.
- The rent price in AirBnb is representative in all the renting market and the property price in Zillow is representative in all the house purchasing market in NYC.
- The rent price for each property remains same within the month.
- The total revenue for the investment mainly comes from the rent, other earnings from the rent process is not counted in this case.
- The total cost of the investment mainly comes from the purchasing of the property, other costs and losses will not count in this case.
- The days within each month is simplified to 30 days.

3.3 Define Metric

I used Rent-to-Price ratio as the metric to measure how profitable an area is within that month. It was calculated as below:

Rent-to-Price ratio = total revenue within the month/total cost

It means the revenue they earned from monthly rent for each dollar they invested in the property. So the higher the ratio, the higher the profitability.

The *total revenue within the month* is the monthly rent of 2017-05, which is calculated as:

*Monthly Rent = Daily Rent*30*0.75*

The *total cost* is the property price of 2017-05. Therefore, the optimal formula is:

*Rent-to-Price ratio = Daily Rent*30*0.75/Property Price*

4. Model Result

4.1 Insights of Rent-to-Price Rate

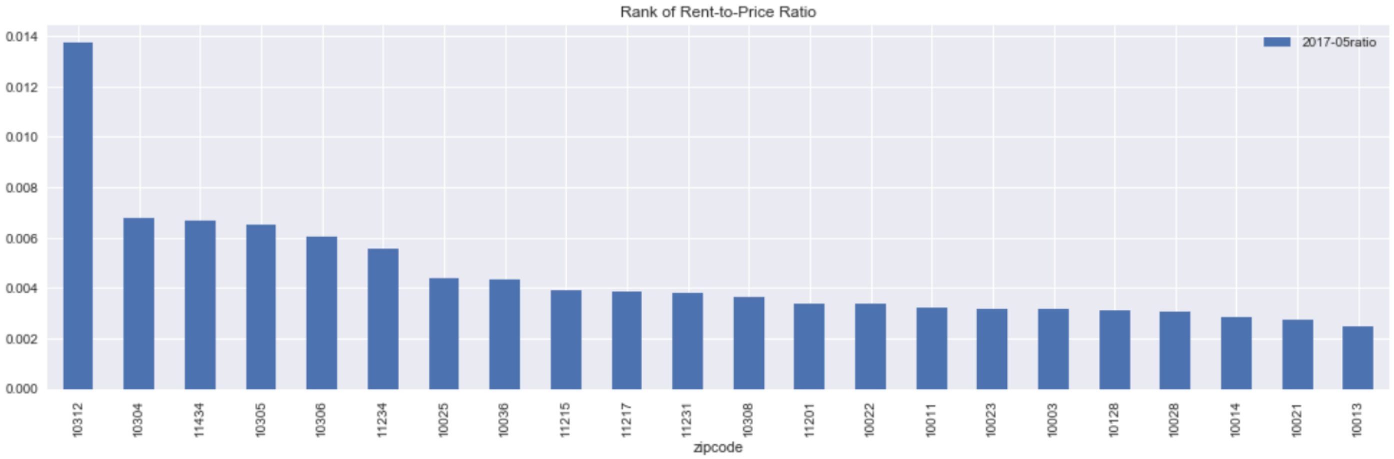

Figure 5 Rank of Rent-to-Price Ratio

The zipcode of 10312, 10304, 11434, 10305 and 10306 are among the top five areas with highest Price-to-Rent Ratio.

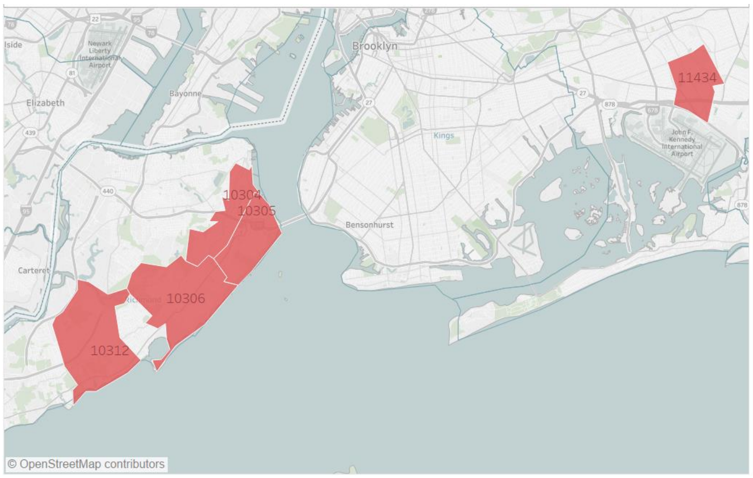

Figure 6 Areas with highest Price-to-Rent Ratio

These five areas are located in the Staten Island and Brooklyn (Text on graph showing the zipcode). They will bid the most profitable area for our client’s investment. Now I will analyze these five zipcode in the property price aspect and rent price aspect.

4.2 Insights of Property Price

From the figures below (Text on graph showing the zipcode), we can see these five areas are among the lowest property price compared with other areas in NYC, which means they have a relatively low investment cost.

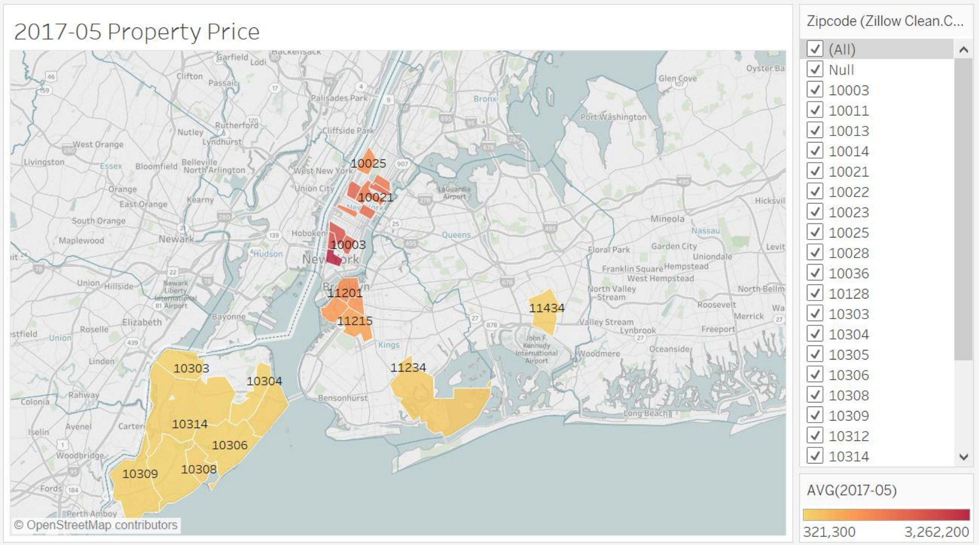

Figure 7 property price visualization

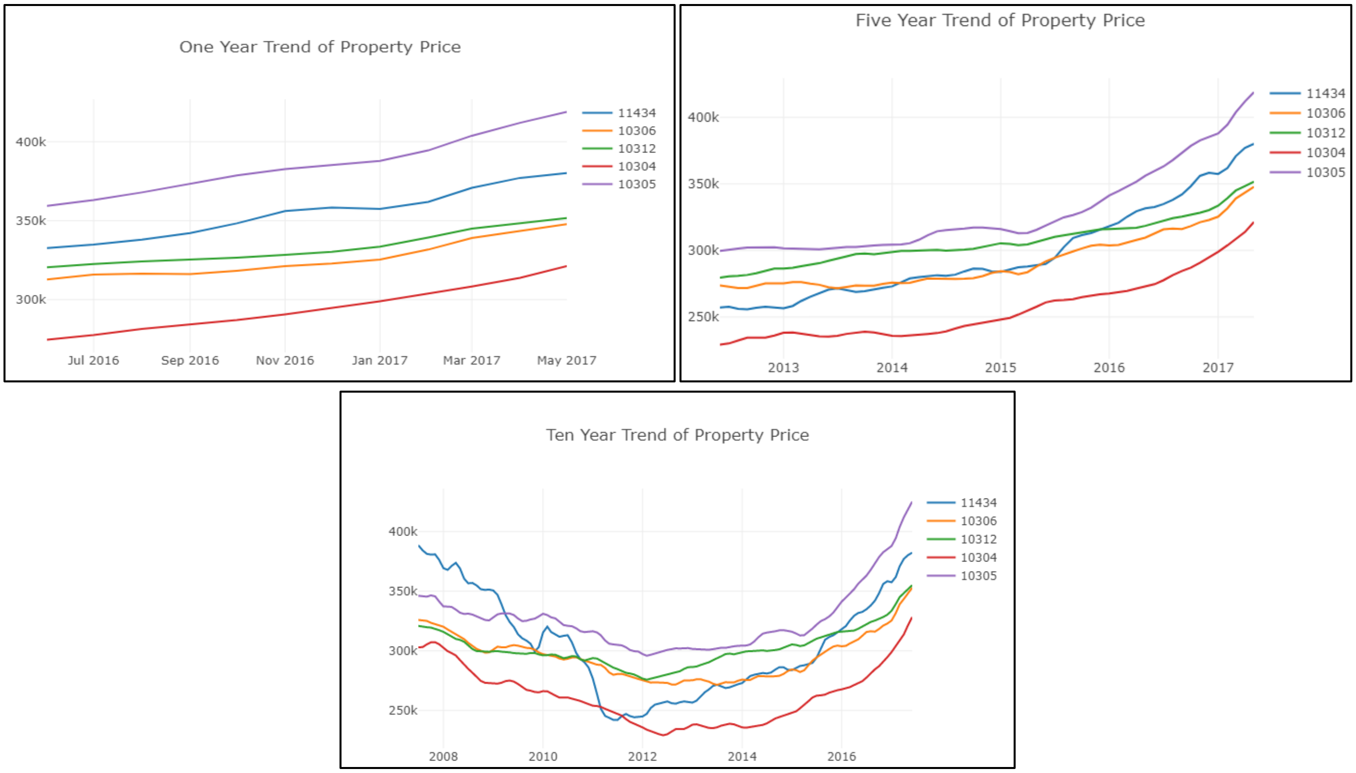

Figure 8 Property price trends for different time horizon

In the one year and five years trend, we can see all the property price within the five zipcode areas are increasing. For ten years, they decrease from 2008 because of the financial crisis.

Comparing the five zipcodes, the property price of 10304 is the lowest among them, but it tends to have a high increasing rate. The property price 10312 is in the middle and it tends to have a lower increasing rate. This characteristic shows good potential to invest 10302 in the both short term and long term.

4.3 Insights of Monthly Rent

From Tableau Results (Text on graph showing the rent price) and Python results (bar plot):

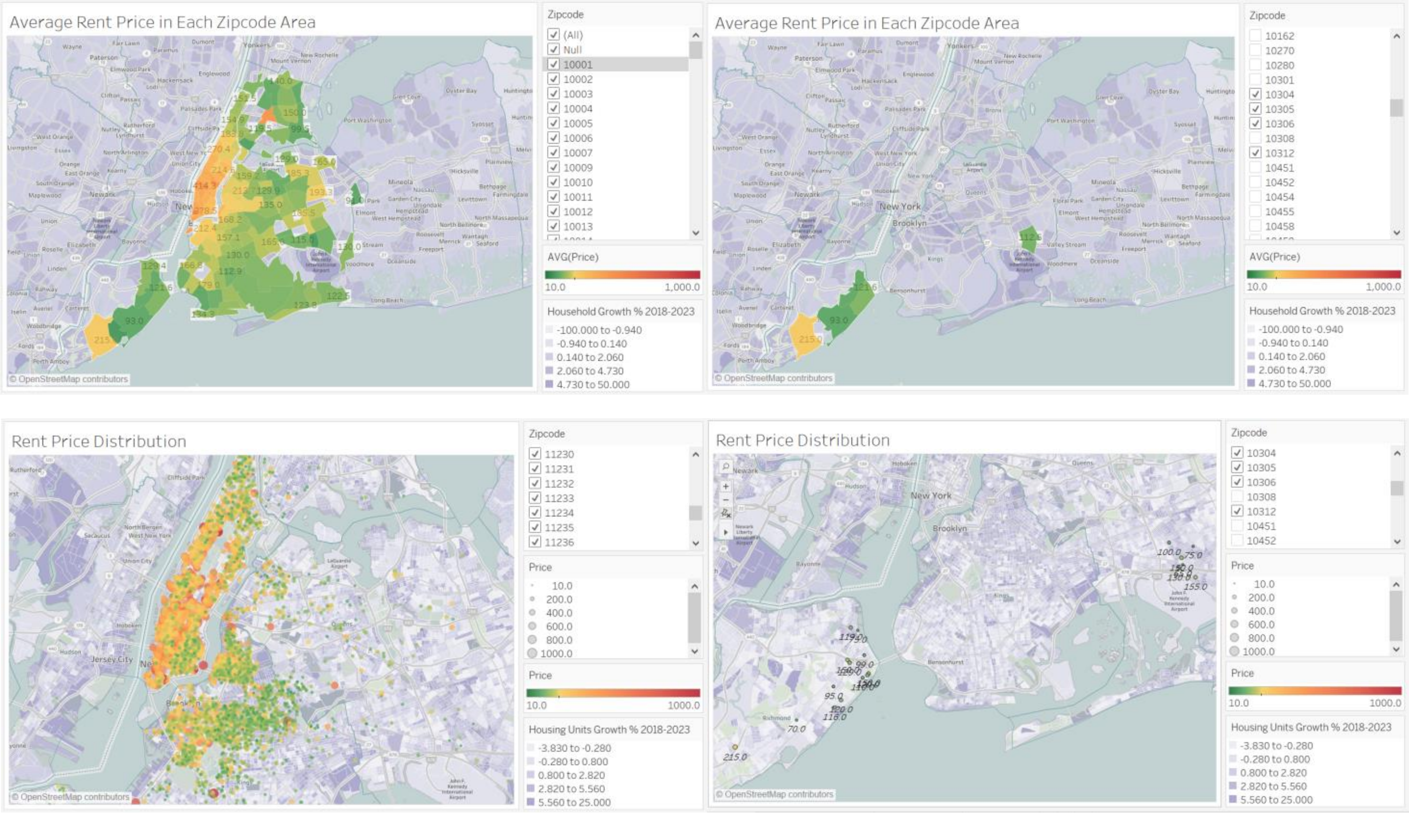

Figure 9 Rent price visualization

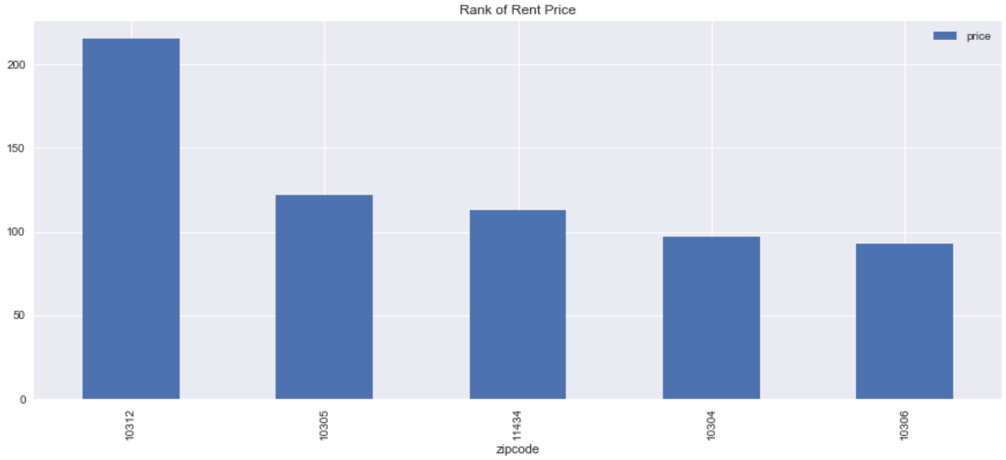

Figure 10 Rank of Rent Price

Zipcode 10312 has the highest rent price (2017-05), much higher than the other four zipcode areas. So in short term, investing in this area will have a higher revenue than other areas. Considering the long term, we see 10304, 11434 are in the area which is potential to have more housing units' growth from 2018-2023. So it is potential to generate more revenue.

4.4 Best Zip code to Invest

Based on our analysis above, we can conclude that zipcode 10312 is the best choice to invest for the short term. Because it has a moderate investment cost with slow increase rate and a relatively high investment revenue compared to other four zipcodes.

In the long run, 10304 seem to be a potential good choice. Because it has the lowest investment cost, moderate investment revenue and high potential getting more housing units in the next five years, which means high potential of getting more profit in future.

Because our client is only interested in short-term investment, I finally chose 10312 as the best zipcode to invest.

5. Further Analysis —— What Is Next

Because there is lack of data and lack of enough time for this analysis case, there can be more interesting work to do if given more data. I will describe some potential further analysis here:

- The rent price data is only limited to one month. If I am able to get the data for the same range as the property price data (2007-07 ~ 2017-06), I am able to do a time series analysis for the rent-to-price ratio rate and analyze its investing potential in a dynamic mode.
- I can build machine learning models for the airbnb dataset to investigate the important factors affecting the renting price and give our client suggestions in improving the renting price in many aspects.
- I can also conduct competitiveness analysis for our client if I am given more data about the investment market and competitors information, so that I can help our client make a better plan for investment.

6. Appendix

Tableau Public: https://public.tableau.com/profile/xinrantao#!/