|  |
| --- |
| Capital One |
| Data Challenge |
| Wenhan Sun |

|  |
| --- |
| wxs368@miami.edu  1-31-2020 |



**NYC Investment: Finding out the most profitable ZIP Codes for short-term house rental in NYC**

**Problem Statement**

A real estate company that has a niche in purchasing properties to rent out short-term as part of their business model specifically within New York City. The real estate company has already concluded that *two-bedroom* properties are the most profitable; however, they do not know which zip codes are the best to invest in.

The real estate company has engaged my firm to build out a data product and provide my conclusions to help them understand which zip codes would generate the most profit on short term rentals within New York City.

**Assumptions**

* 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 are homogeneous (i.e. a 1000 square foot property in a locale such as Bronx or Manhattan generates twice the revenue and costs twice as much as any other 500 square foot property within that same locale)
* The occupancy rate for each property is 75% all year round
* Property tax rate, house insurance fee and utilities are same for each property among different locales

**Tools and Packages**

1) Tools

* Jupyter Notebook

2) Packages

* Basic data clean: pandas, numpy
* Handling outliers: statistics, astropy.stats
* Missing value imputation: sklearn.impute
* Create graphs: matplotlib, seaborn, geopandas, shapely

**Data Selection and Cleaning**

**Data Source**

There are two datasets provided for analysis.

* Zillow dataset - provides median cost for two-bedroom properties in NYC by zip code.
* Airbnb dataset - provides detailed information about property listing, including location, number of bedrooms, reviews, price, etc.

The data were imported from secure sources provided by the company.

**Data Selection**

I did a basic exploration to have a feel of the data and kept the datasets to only columns needed for analysis.

For the Zillow dataset, I selected 81 columns in total, including 3 categorical columns, *RegionName*, *City* and *State*, and *78 numeric columns*, *2010-01 ~ 2017-06*.

For the Airbnb dataset, 12 columns were selected at last, including *city*, *state*, *zipcode*, *bedrooms*, *price*, *latitude*, *longitude*, *is\_location\_exact*, *room\_type*, *property\_type*, *cleaning\_fee* and *neighbourhood\_group\_cleansed*.

**Data Cleaning**

I cleaned the Zillow and Airbnb datasets separately and then joined them together.

At the very beginning, I renamed some columns to make them more readable, for both datasets. Then I checked

* duplicates
* missing values
* outliers
* data type and format of each column

Addition to the process above, I also filtered and saved properties that had only two bedrooms and converted the whole dataset into zip code level for the Airbnb dataset. I saved the cleaned Airbnb dataset into a new dataset named *airbnb\_2bed\_zip*.

The house price columns in the original Zillow dataset were on monthly level. I changed them into yearly level for the purpose of easier interpretation and understanding.

The final shape of the data is as follows:

* Zillow – 25 rows × 101 columns, name of dataset in Jupyter: *zillow*;
* Airbnb – 162 rows × 6 columns, name of dataset in Jupyter: *airbnb\_2bed\_zip*.

**Data ETL**

After changing the messy raw datasets into cleaned ones, I could finally do some data manipulations based on the cleaned datasets.

**a) New variables**

For the *zillow* dataset, I calculated the house prices for each zip code in 2018 and 2019 using the *2017-06* column and the *Case Shiller Index*. The results were saved in two columns named *2018* and *2019* respectively.

For the *joinData2* (refer to the next paragraph) dataset, I created 8 new variables – *total\_revenue*, *ab\_fee*, *net\_revenue*, *closing\_cost*, *investment*, *expenses*, *profit* and *ROI*. [[1]](#footnote-1)

**b) New datasets**

After got the house prices in *2018* and *2019*, I created two subsets of the *zillow* dataset, *zillow1* and *zillow2*. Then, I inner joined *zillow1* and *zillow2* with *airbnb\_2bed\_zip* and got two new datasets, *joinData1* and *joinData2,* respectively. *joinData1* was used for time-series data analysis – used to craft the house prices trend plot - while *joinData2* was used for cross-section data analysis.

**Data Analysis**

**Market Performance**

The goal of the company is to invest and get success in the short-term house rental market in New York City. So, it’s important to get a clear awareness of the performance of the real estate market as well as the short-term rental market in advance.

**a)** **Real estate market performance in NYC**

*Figure1* shows the house prices trend in NYC by zip code from 2010 to 2019. Each line represents a zip code and is colored by the boroughs of NYC.

Purely from the glance at the house prices at different boroughs, I would suggest **the company prioritize the consideration of buying properties in Staten Island or Queens instead of Manhattan or Brooklyn.**

House prices in any zip code in Manhattan are way much higher than those in Staten Island/ Queens, which means the cost is much higher.

The house price difference between Manhattan and Staten Island/Queens is about 5 times in average in 2019. Thus, if the company is going to buy a property in Manhattan, the company has to get 5 times of the profit from the short-term rental activity in order to make that property in Manhattan as profitable as that in Staten Island/Queens, which isn’t that easy.

The same analysis can be applied to Brooklyn and Staten Island/Queens. But the situation is better, since the house price in Brooklyn isn’t that insane as in Manhattan.

Also, from the more and more flatten trend of house price, the company had better not to expect to earn money from house appreciation, at least for the next few years.

A close up of a map

Description automatically generated

Figure 1

**b) Short-term rental market performance on Airbnb in NYC**

Though the cost of buying a property in Manhattan/Brooklyn can be very high, it’s still hard to say that it’s not worth buying one in any zip code in those two boroughs without deeper analysis.

*Figure 2, 3, 4* and *5* show the variance of the cost of buying a house, the annual net revenue, expenses and profit of running a short-term rental business on Airbnb platform on the zip code level respectively.

The lighter the color is, the higher the value is.

*Figure 2* gives the same result that we get from Figure 1 – house prices in Manhattan vary a lot but are still way higher than those in Staten Island/Queens; house prices in Staten Island are not only low but also identical in different zip codes; prices in Brooklyn are at the middle level.

*Figure 3* shows an interesting result. The range of the annual net revenue between Manhattan/Brooklyn and the Staten Island/Queens is much smaller than the range of the house cost. This indicates that Manhattan and the northwest part of Brooklyn may not be as competitive as other areas.

*Figure 4* and *5* further confirm the conclusion. The annual expenses range is larger than the annual net revenue range. High net revenue with high expenses turns into low profit, while relatively high net revenue with low expenses turns into relatively high profit.

We can get a preliminary conclusion that **zip codes in the Staten Island are worth investing, while zip codes in Manhattan may not worth spending money on**. Brooklyn and Queens are in the middle.

|  |  |
| --- | --- |
| Figure 2[[2]](#footnote-2) | Figure 3 |
| Figure 4 | Figure 5 |

**ZIP Codes Selection**

Now that we have got a sense that the company can invest their money in the Staten Island, Queens and Brooklyn and had better not to spend money on Manhattan based on the market performance analysis, it’s time to find out the specific zip codes.

**The method I used is to calculate the annual ROI rate of each zip code and select those zip codes that have an annual ROI rate larger than 4%**.I set up the bottom line at 4% is because that in general, we regard it as reasonable if we can get the investment back within 300 months, which is corresponded to 4% in the form of ROI rate.

**A screenshot of a cell phone

Description automatically generated**

Figure 6

*Figure 7* visualizes the selected zip codes on map. The result is consistent with the former analysis.

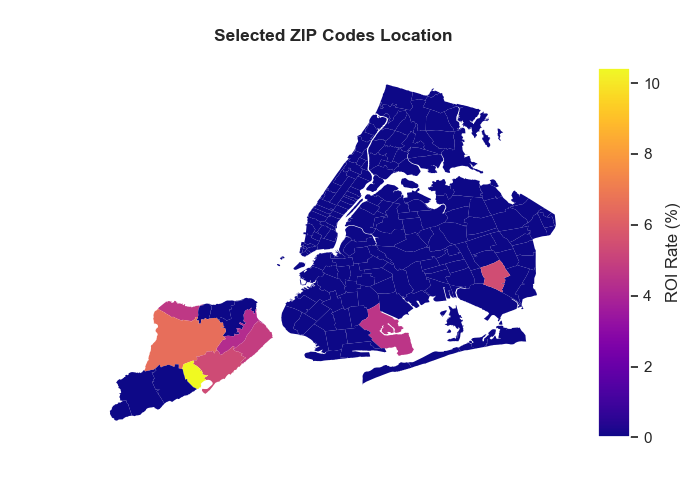
****

Figure 7

**Conclusion**

Based on the analysis above, I suggest the company take 8 zip codes into consideration. (placed by the ROI rate in descending order)

* 10308
* 10314
* 11434
* 10306
* 10305
* 10303
* 11234
* 10304

Six of them locate in the Staten Island, that are 10308, 10314, 10306, 10305, 10303 and 10304; the other two zip codes are in Brooklyn (11234) and Queens (11434) respectively.

None of the selected zip codes is in Manhattan. That is reasonable, since **the house price and annual expenses in Manhattan are so high that counteract the benefit from high net revenue**.

**Next Step**

* Add in more zip codes
* Update the Zillow and Airbnb datasets in terms of time
* Add in more data (the number of data points for each zip code is not even and there are missing values of important columns in both datasets)
* Find the actual occupancy rate for each property ( had better to be on monthly level)
* Determine the type and the number of properties to buy for each zip code
* Take more elements into consideration (such as crime rate, house age, distance from public transportation, distance from downtown area, etc.)

1. The data description and the calculation method of each column were documented in the files “*Metadata – Wenhan Sun – Miami*” and “*codes - Wenhan Sun - Miami.ipynb*” respectively. [↑](#footnote-ref-1)
2. a) The price unit is USD, same in figure 3, 4, 5.

   b) Those dark purple areas are zip codes lack of data information, same in figure 3, 4, 5. [↑](#footnote-ref-2)