Data Challenge

Author: Yunuo Wu

1. Introduction

1.1 Problem Statement

For a real estate company that has a niche in purchasing properties to rent out short-term specifically within New York City, the problem of which zip codes are the best to invest in need to be solved in two perspectives. First, we focus on a high annual return on the investment. Second, the smaller the initial investment, the better.

1.2 Assumptions

- The investor will pay for the property in cash
- The time value of money discount rate is 0%
- All properties and all square feet within each locale can be assumed to be homogeneous
- The Airbnb hosts gain same revenue as the table shows every year
- 70% of people rent once for a day, 20% of people rent once for a week, 10% of people rent once for a month
- Profit maximum in all situations (e.g. customer will bring the maximum number of guests they can to the property)

1.3 Created Metadata

Field	Description
reprice	Price the host is really charging to stay per night considering the discount for weekly reservation and monthly reservation
value	Current Value of Investment, refers to the proceeds obtained from the sale of the investment
cost	The house price in Jan 2020
ROI	A performance measure used to evaluate the efficiency of an investment
PP	The amount of time it takes to recover the cost of an investment
logcost	Apply log to cost

1.4 Other Sources

(1) geoJSON file that has NYC zip code boundary coordinates: https://raw.githubusercontent.com/fedhere/PUI2015_EC/master/mam1612_EC/nyc-zip-code-tabulation-areas-polygons.geojson

2. Quality Check

This part will talk about data cleaning. The goal of data cleaning is to achieve six data quality dimensions which are completeness, uniqueness, timeliness, validity, accuracy and consistency. In this way, qualified data can further be used to conduct analysis.

(1) Airbnb

Field	Quality Dimension	Data Cleaning
zipcode	completeness	Fill null values based on the longitude and latitude given in the dataset
bedrooms	accuracy	Filter two bedrooms properties since they are most profitable
state	validity accuracy	Fill null data Modify one state with only one name
weekly_price	completeness	Fill null values
monthly_price	completeness	Fill null values
Extra_people	validity	Drop dollar sign and make it float

(1) Zillow

Field	Quality Dimension	Data Cleaning
time	consistency	Filter the month in which there are missing values
city	accuracy	Filter the NYC

Then, the two datasets are merged based on the common variable “zipcode”. I find out that the Airbnb dataset does not include host information for zip code 10312. So, after merging, we only have 24 zip codes data in NYC. They are 11215, 11217, 11231, 10023, 11201, 10013, 10011,

10003, 10025, 10128, 10014, 10022, 10036, 10314, 10021, 10028, 10305, 10304, 11234, 10308, 11434, 10303, 10306 and 10309.

Now all the variables in our merged dataset “df_join[cols]” have 1573 values, and the dataset is ready to do further analysis to get insights.

3. Data Munging

3.1 Related Concept

As I mentioned before, the key points to measure the most profitable zip codes are based on two dimensions: annual return and initial investment.

3.1.1 ROI

For annual return, a more direct and accurate way to explain is through the concept of ROI, which is return on investment. It is a performance measure used to evaluate the efficiency of an investment or compare the efficiency of a number of different investments. The formula is as follows:

$$\text{ROI} = \frac{\text{Current Value of Investment - Cost of Investment}}{\text{Cost of Investment}}$$

3.1.2 Payback Period

Another way to measure annual return is through payback period. Payback Period refers to the amount of time it takes to recover the cost of an investment. Unlike other methods of capital budgeting such as net present value (NPV), internal rate of return (IRR), and discounted cash flow, it ignores the time value of money, which exactly matches our assumption. The formula is as follows:

$$\text{Payback Period} = \frac{\text{Cost of Investment}}{\text{Current Value of Investment}}$$

The payback period is the cost of the investment divided by the annual cash flow. So, in this simplified situation, payback period tells the same story as ROI.

$$\text{ROI} = \frac{1}{\text{Payback Period}} \text{-} 1$$

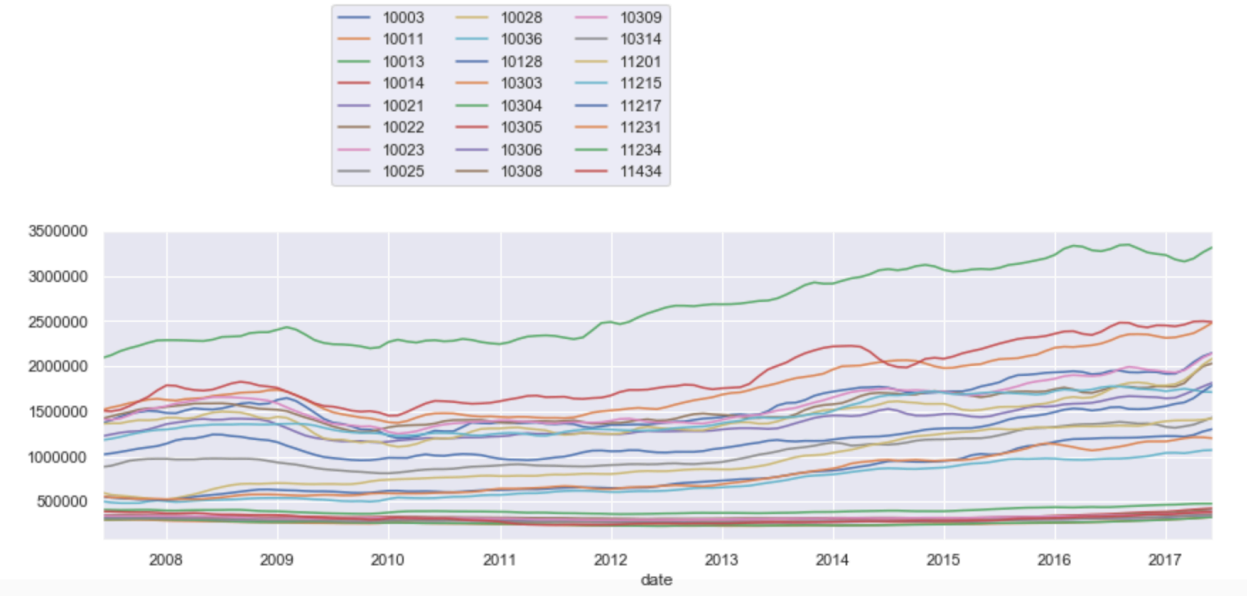
3.1.3 House Price

The initial investment of purchasing properties based on the house price. To make our analysis result updated and accurate, we use time series analysis to predict the house price of different zip codes in NYC in 2020.

3.2 Time series Predicting (house price)

3.2.1 Plot the data

First, we adjust our data frame to let it suitable for time series analysis. Here is the plot of historical house price from June 2007 to June 2017 for the 24 zip codes.



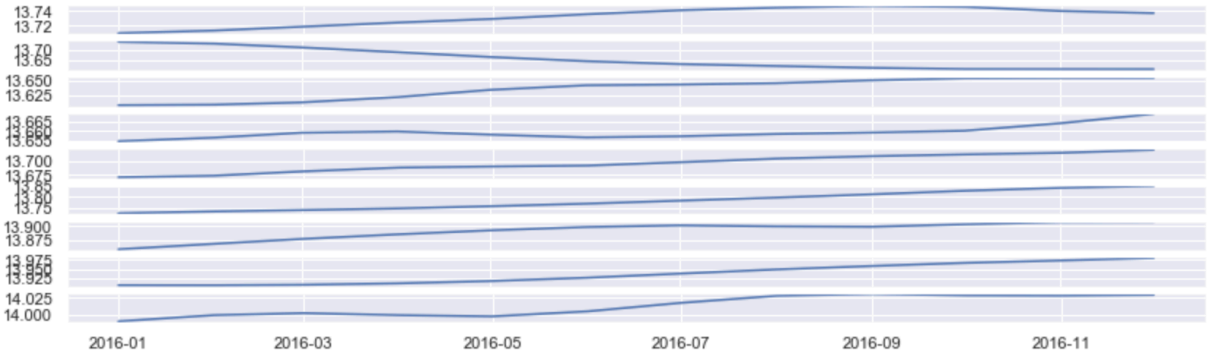
Insights:

- There is an upward trend.
- There are no obvious outliers
- There seems no seasonality, and we will further check it
- It is difficult to judge whether the variance is constant or not, and we will further check it

3.2.2 Seasonality and Constant Variance

Our time series dataset contains historical house price for 24 zip codes. It will be a lot of work if we develop house price predicting models for each zip code. To simplify the case, I will only develop one model for the average house price for all zip codes in NYC. Besides, for the sake of accuracy, I apply log function to house price since these numbers are extremely big.

For seasonality and variance, first I want to check it through a direct way. So I make a plot to show how house price changes in one year. The line plots are aligned vertically for years from Jan 2008 to Jan 2016 to help spot any year to year pattern.

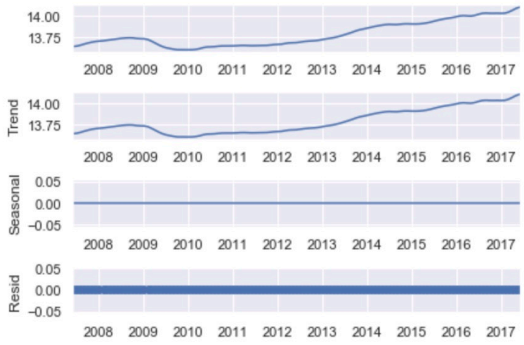


Insights:

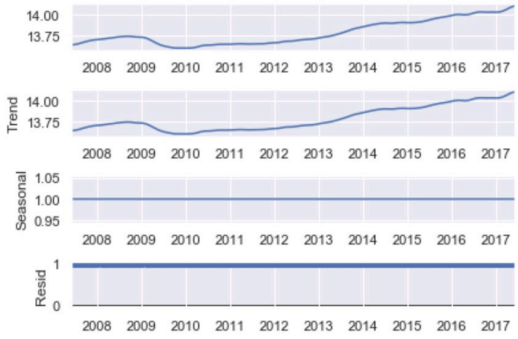
- The pattern appears same in each year
- There is an upward trend: lowest in the beginning of a year, highest in the end of a year
- There is no obvious seasonality
- The variance is constant

Besides, I also use theoretical method to decompose time series data into trend and seasonality from both additive and multiplicative perspectives.

For additive model, the result is:



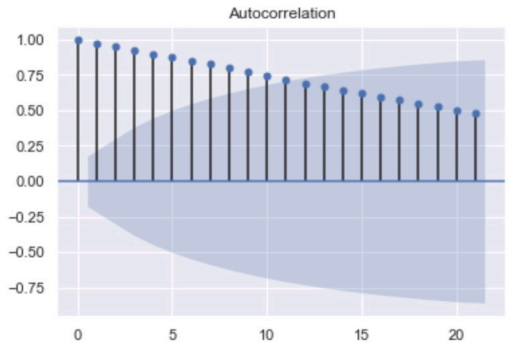
For multiplicative model, the result is:



Both results show that there is no seasonality.

3.2.3 Stationarity

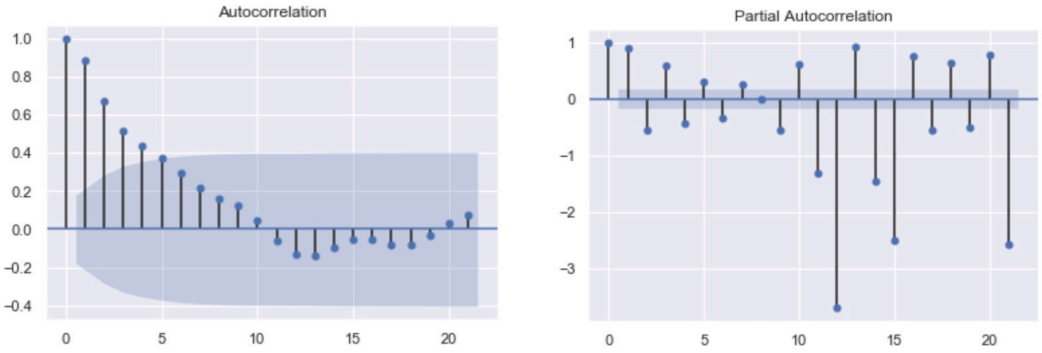
To show whether there is need to do differencing, we need to plot the ACF plot.



Seen from the plot, the autocorrelation decays slowly, which indicates the data is not stationary. After differencing the data fluctuate up and down evenly around the mean, the data is now stationary:

Next, theory method augmented Dickey–Fuller test (ADF) is used to check the stationarity. I come to the conclusion that we cannot reject the null hypothesis that the data is not stationary at the 95% confidence level.

3.2.4 ARIMA Model Parameter Choosing



Insights:

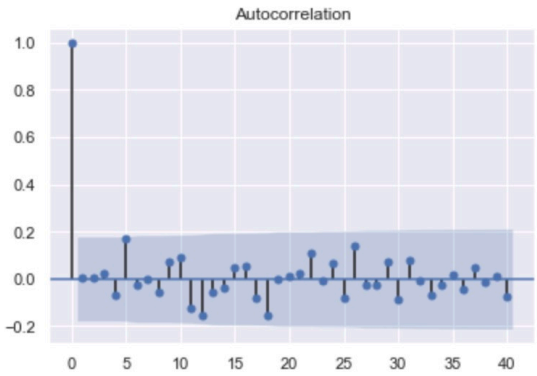
- ACF Plot: the autocorrelation decays fast. It may show a AR(4) process
- PACF Plot: it is difficult to get more information here

  In real life, the values of parameter p and q are usually no more than 4. I write a for loop to test all possible number combinations, and find the combination that carries the smallest BIC. After testing, p=1 and q=3 are the best.

ARIMA Model Results

Dep. Variable:	D.value	**No. Observations:**	120
Model:	ARIMA(1, 1, 3)	**Log Likelihood**	568.002
Method:	css-mle	**S.D. of innovations**	0.002
Date:	Fri, 31 Jan 2020	**AIC**	-1124.004
Time:	21:05:21	**BIC**	-1107.279
Sample:	07-01-2007	**HQIC**	-1117.212
	- 06-01-2017		

The ACF Plot of residuals of the ARIMA(1,1,3) model shows the residuals follow a white noise process. In conclusion, our model is a good fit.



4. Data Application

Using the house price predicting model, we can predict the newest house price in each zip code. With the house price, we get the cost of investment. The current value of investment is equal to annual revenue of renting out. Considering hosts also need to pay 3% of the amount of each transaction as service fee, the formula is as follows:

$$\text{Price} = 0.7*\text{daily\_price} + \frac{0.2*\text{weekly\_price}}{7} + \frac{0.1*\text{monthly\_price}}{30}$$

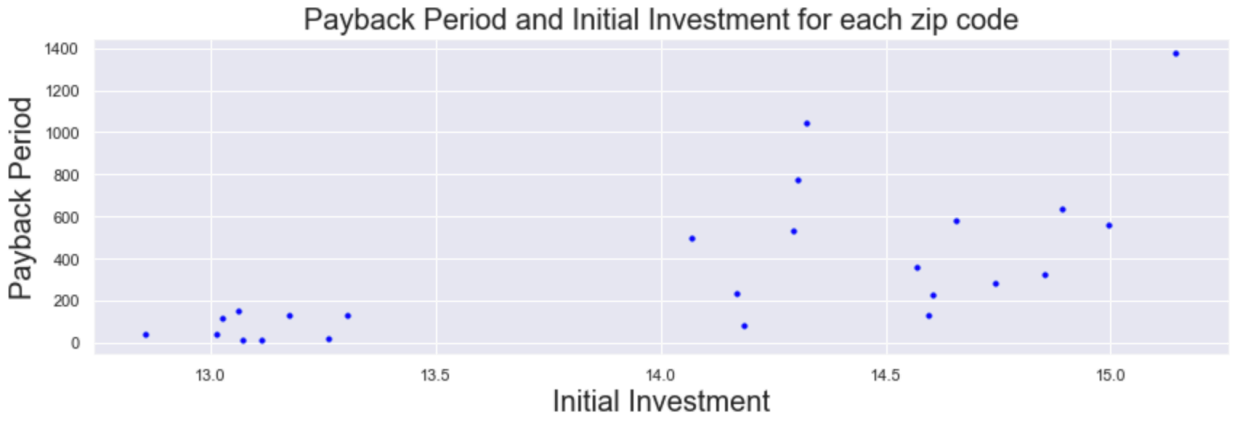
$$\text{Current Value of Investment} = (\text{Price} + \text{extra\_people} * (\text{accommodates} - \text{guests\_included})) * \text{availability\_365}) * 75\% * 97\%$$

$$\text{Current Cost of Investment} = \text{Current House Price}$$

Now, we can calculate the ROI and payback period. In the following comparison, I will only consider payback period as the measure of annual return. In the left is top five zip codes which have the smallest payback period, and in the right is the top five zip codes which have the smallest initial investment.

	zipcode	PP	logcost	ROI	cost
14	10306	11.380011	13.071988	-0.911701	475436
22	11234	15.955811	13.111816	-0.894826	494754
15	10308	22.594862	13.263196	-0.954969	575616
11	10303	38.985886	13.013428	-0.964536	448394
23	11434	43.781869	12.856631	-0.952110	383322

	zipcode	PP	logcost	ROI	cost
23	11434	43.781869	12.856631	-0.952110	383322
11	10303	38.985886	13.013428	-0.964536	448394
17	10314	116.339104	13.026240	-0.991312	454176
12	10304	149.958051	13.060998	-0.985009	470240
14	10306	11.380011	13.071988	-0.911701	475436

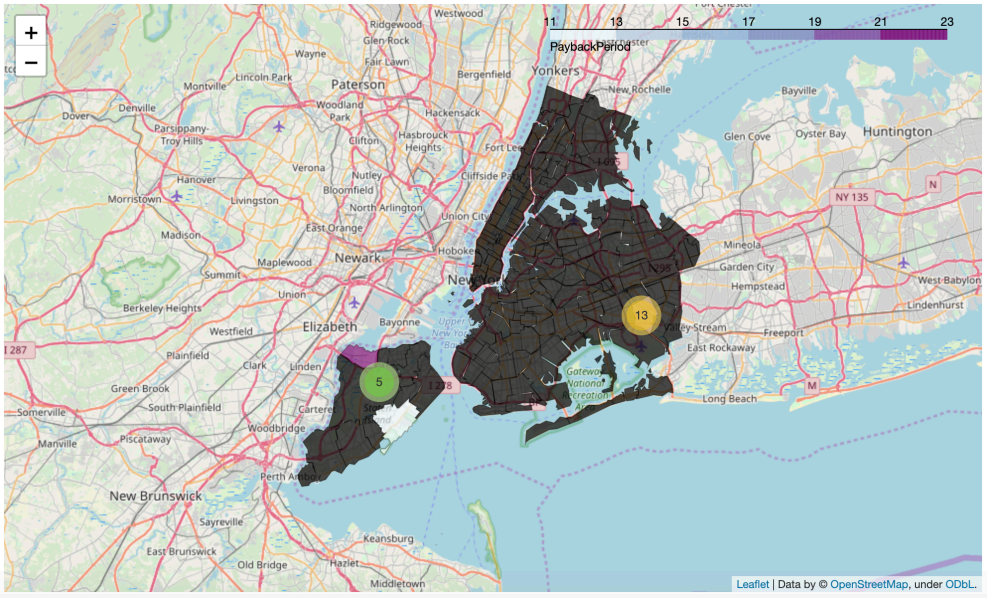


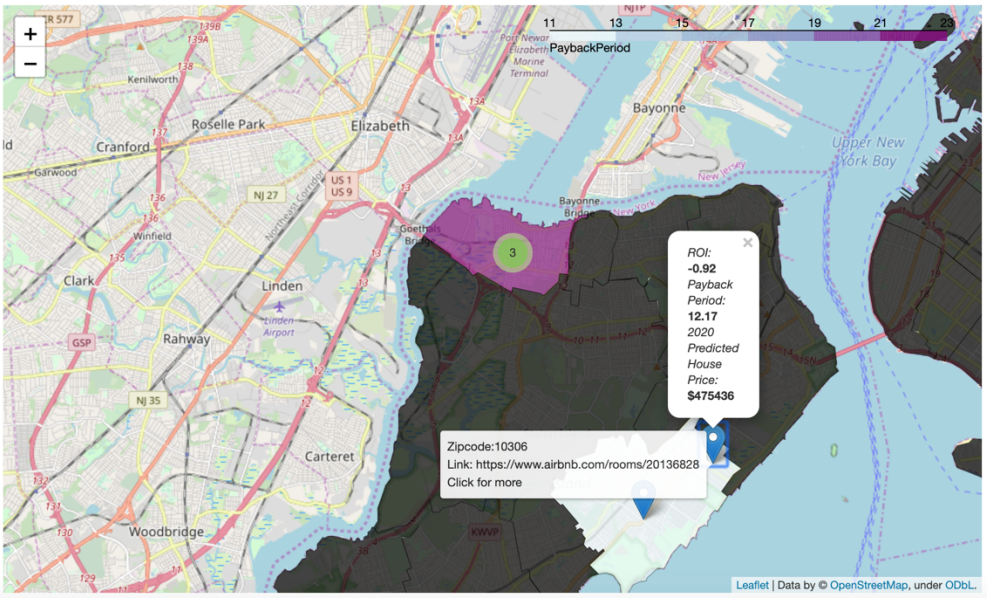
Key Business Insights:

- Taking both factors into consideration, 10306(Staten Island), 10303(Staten Island) and 11434(Queens) are three most profitable zip codes.
- 10306(Staten Island) is located in Richmond County. It has the lowest payback period. So, if you are looking for an investment which can reach to the breakeven point as soon as possible, I strongly recommend you purchase properties in zip code 10306.
- 10303(Staten Island) is located in Richmond County. It has a good balance between initial investment and payback period. They are both controlled below the acceptable level. Basically I will recommend properties in this zip code to every type of investors.
- 11434(Queens) is located in Jamaica City. It has the lowest initial investment. At the same time, however, the payback period is highest among the five. So, if investors' budget is limited, I will recommend them to purchase properties in 1143 zip code.

5. Data Visualization

5.1 Interactive Map





Key Business Insights:

- According to the map, all the properties the hosts provide in these zip codes are close to coastline. As a result, there is a good view from the window. Also, it is more convenient for people to go to the beach(tourist place) considering the transportation. It can explain why these zip codes deserve investing.

- According to the map, these zip codes are close to the airports especially 11434 which is located in Queens County. A short distance to airport and the convenient transportation make them favorite places for tourists in NYC.

5.2 Boxplot

6. What's Next

- The ARIMA model

  The ARIMA model which is built on average house price of all zip codes in NYC may be not suitable for each zip code. In this way, it can lead to inaccurate prediction. The better way to solve the problem is to write a for loop to find the best parameters for each zip code.

- Occupancy rate

  Not all property will have the same occupancy rate, so it may be not reasonable to set 75% as the occupancy rate. Mashvisor 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. We can improve the model based on information Mashvisor provided.