NEW YORK AIRBNB RENTALS – SHORT TERM PROFITABILITY ANALYSIS

Submitted by –

Guru Prasad Kumar

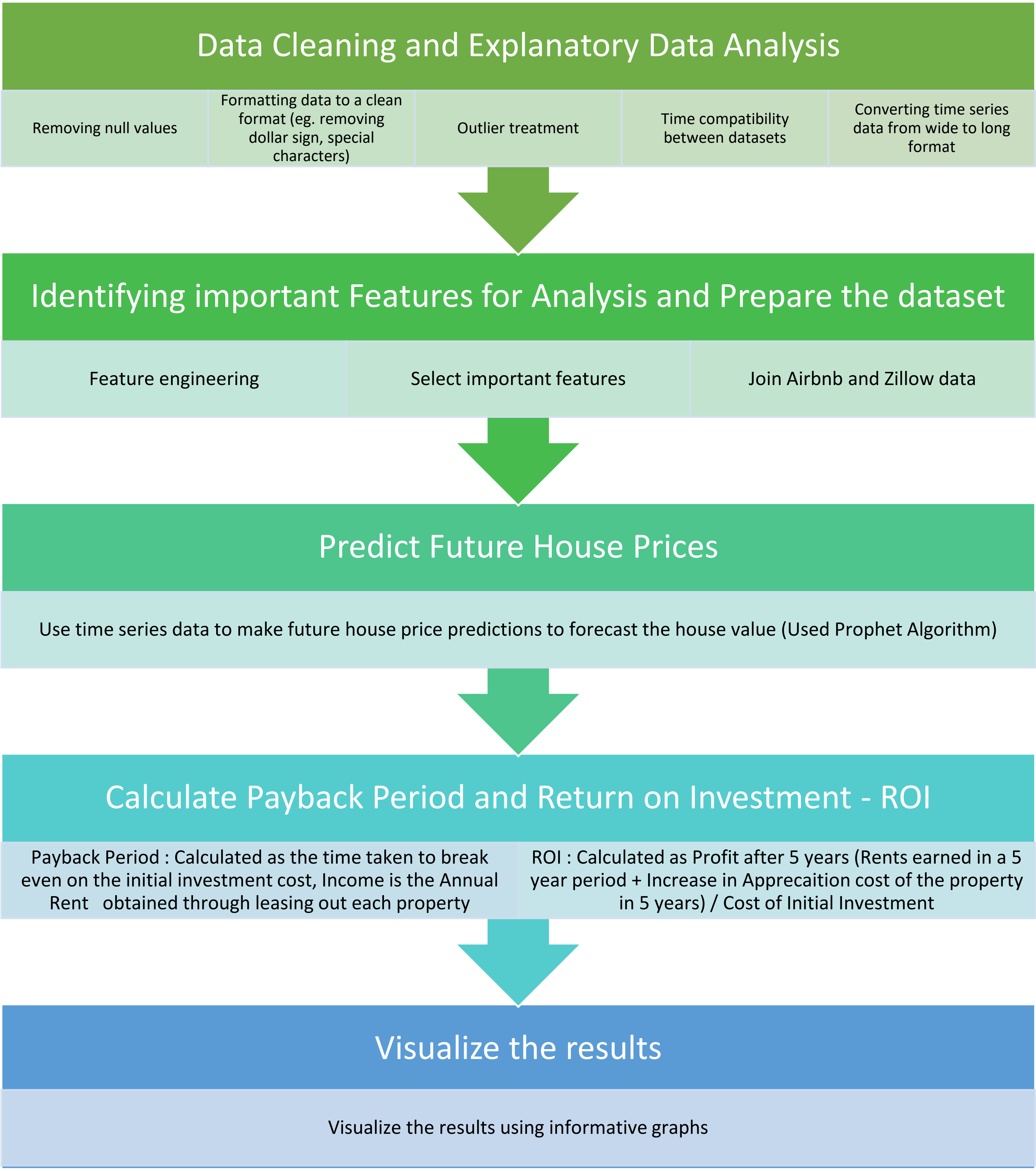
PROBLEM STATEMENT

The Client wants to invest in 2-bedroom properties in New York City and planning to rent them out for short terms. They want to know which the best zip codes are to invest in (zip codes that would generate the most profit on short-term rentals) and they prefer a scalable data product that would help them decide where to invest.

DATA SOURCES

1. Airbnb Data – CSV files containing various attributes of Airbnb listings in NYC.
2. Zillow – Zip code level data which contains the median house prices from 1996 – 2017.

APPROACH



Data Cleaning and Explanatory Data Analysis

- Removing null values
- Formatting data to a clean format (eg. removing dollar sign, special characters)
- Outlier treatment
- Time compatibility between datasets
- Converting time series data from wide to long format

Identifying important Features for Analysis and Prepare the dataset

- Feature engineering
- Select important features
- Join Airbnb and Zillow data

Predict Future House Prices

Use time series data to make future house price predictions to forecast the house value (Used Prophet Algorithm)

Calculate Payback Period and Return on Investment - ROI

Payback Period : Calculated as the time taken to break even on the initial investment cost, Income is the Annual Rent obtained through leasing out each property

ROI : Calculated as Profit after 5 years (Rents earned in a 5 year period + Increase in Appreciaition cost of the property in 5 years) / Cost of Initial Investment

Visualize the results

Visualize the results using informative graphs

ASSUMPTIONS MADE

Occupancy rate of 75% for Airbnb rentals.

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 (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.)

Price listed in the Airbnb data is the final price that the customer pays (i.e. No discounts). Cleaning fee, Security deposit and costs related to extra beds are not considered for the sake of simplicity.

Annual Revenue (Rent) is calculated by, Annual Revenue = Price per night * occupancy rate * availability_365. Rent increases by 5% every year.

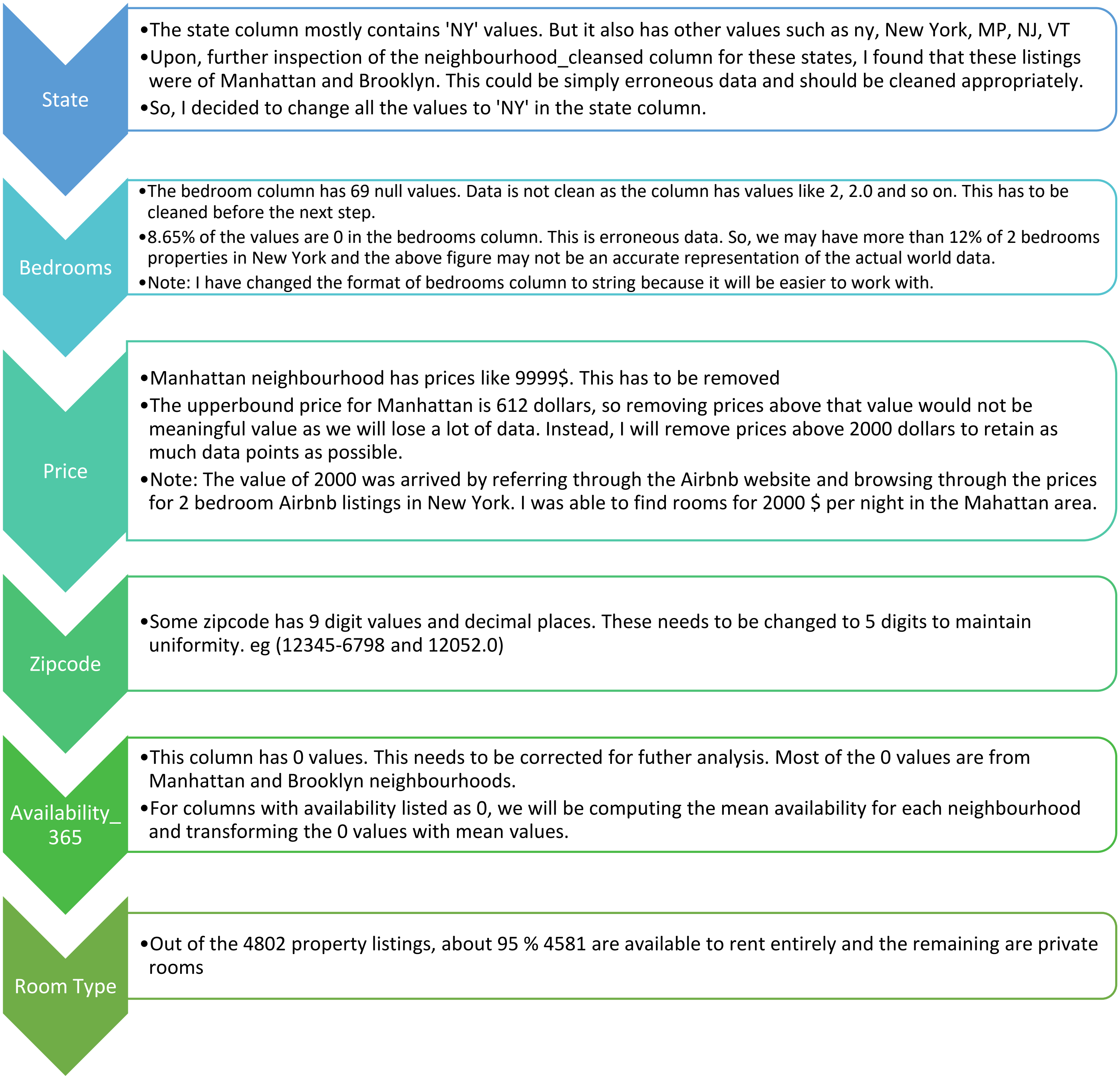
Investor buys the properties in 6th month of 2017 (2017-06). (i.e. Investment cost = Latest Median list price)

Only zipcodes with more than 15 listings are considered for analysis. Zipcodes with more than 15 listings are considered to be zipcodes with high demand.

DATA WRANGLING, EDA AND FEATURE ENGINEERING

1. Airbnb Data:

1.1 Data Cleaning & EDA Insights:



State

- The state column mostly contains 'NY' values. But it also has other values such as ny, New York, MP, NJ, VT
- Upon, further inspection of the neighbourhood_cleansed column for these states, I found that these listings were of Manhattan and Brooklyn. This could be simply erroneous data and should be cleaned appropriately.
- So, I decided to change all the values to 'NY' in the state column.

Bedrooms

- The bedroom column has 69 null values. Data is not clean as the column has values like 2, 2.0 and so on. This has to be cleaned before the next step.
- 8.65% of the values are 0 in the bedrooms column. This is erroneous data. So, we may have more than 12% of 2 bedrooms properties in New York and the above figure may not be an accurate representation of the actual world data.
- Note: I have changed the format of bedrooms column to string because it will be easier to work with.

Price

- Manhattan neighbourhood has prices like 9999$. This has to be removed
- The upperbound price for Manhattan is 612 dollars, so removing prices above that value would not be meaningful value as we will lose a lot of data. Instead, I will remove prices above 2000 dollars to retain as much data points as possible.
- Note: The value of 2000 was arrived by referring through the Airbnb website and browsing through the prices for 2 bedroom Airbnb listings in New York. I was able to find rooms for 2000 $ per night in the Mahattan area.

Zipcode

- Some zipcode has 9 digit values and decimal places. These needs to be changed to 5 digits to maintain uniformity. eg (12345-6798 and 12052.0)

Availability_ 365

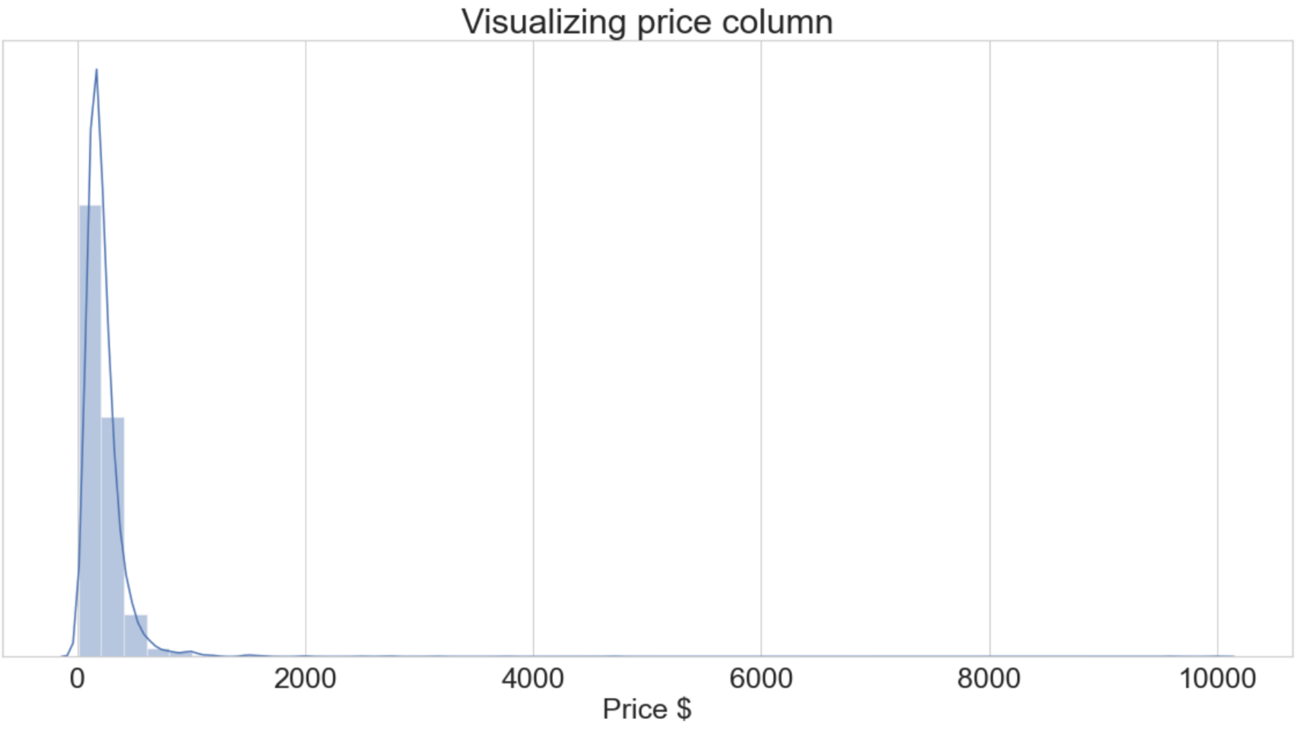
- This column has 0 values. This needs to be corrected for futher analysis. Most of the 0 values are from Manhattan and Brooklyn neighbourhoods.
- For columns with availability listed as 0, we will be computing the mean availability for each neighbourhood and transforming the 0 values with mean values.

Room Type

- Out of the 4802 property listings, about 95 % 4581 are available to rent entirely and the remaining are private rooms

1.2 Explanatory Data Analysis Charts:

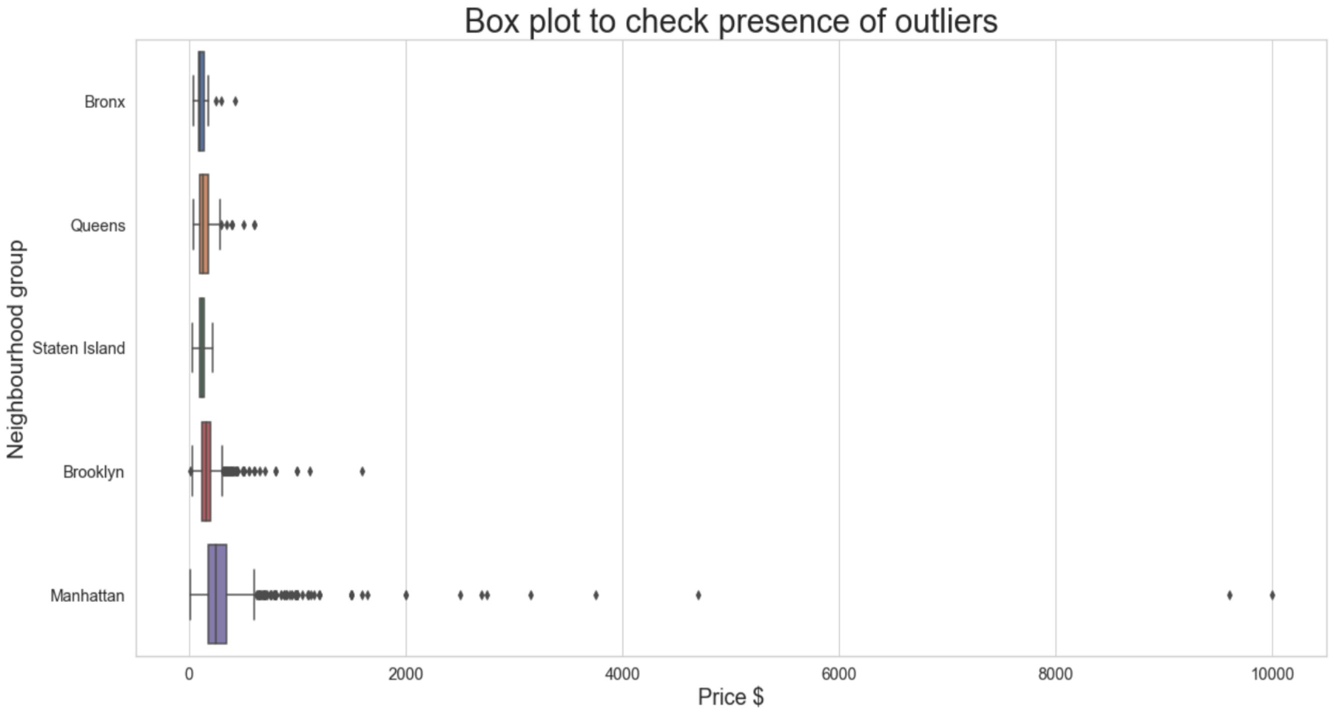
1.2.1 Visualizing price column



Insights

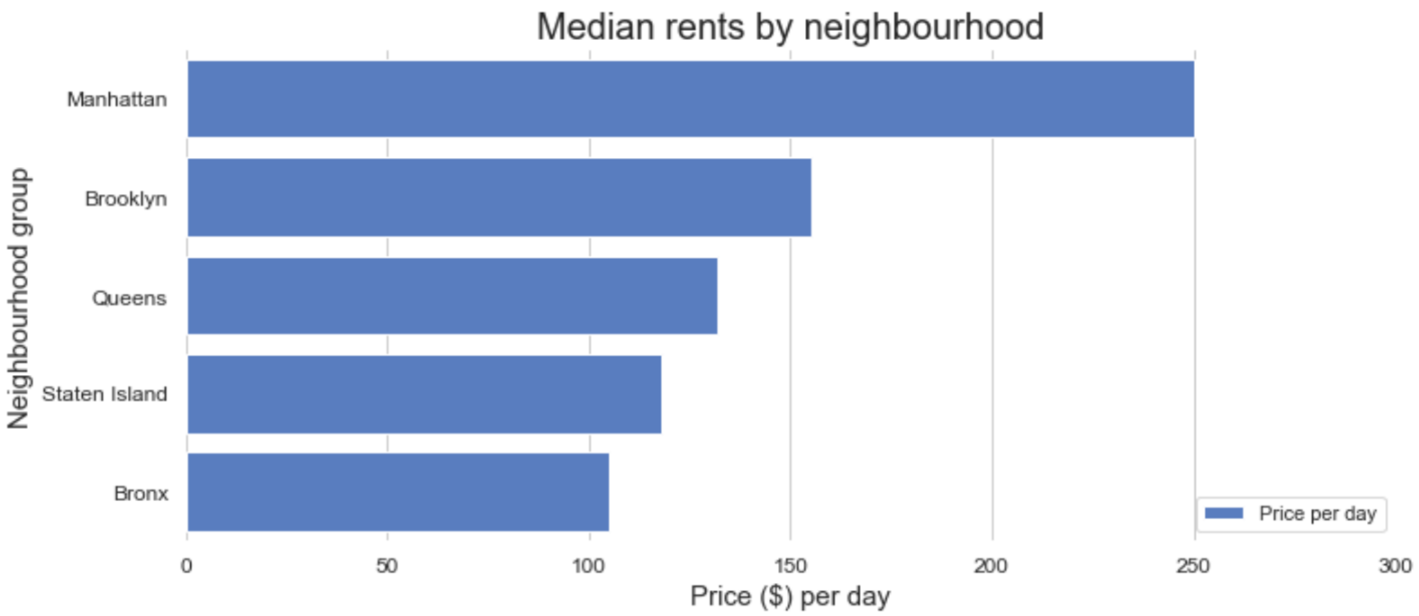
- Price column is positively skewed. In such scenarios, median is a better measure of central tendency than mean since mean is greater than median and it is resistent to outliers compared to mean.

:



Insight

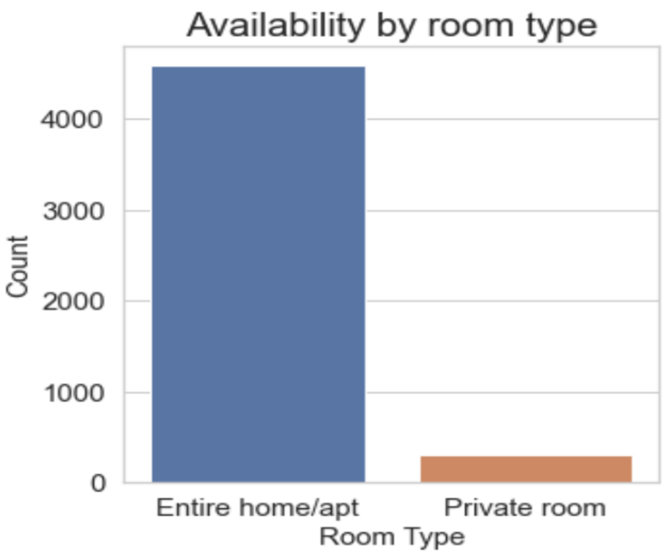
1.Manhattan neighbourhood has prices like 9999$. This could erroneus data and has to be removed since its an outlier



Insights

•Manhattan is the costliest neighbourhood to rent an Airbnb followed by Brooklyn, Queens, Staten Island and Bronx

1.2.2 Visualizing Room Type

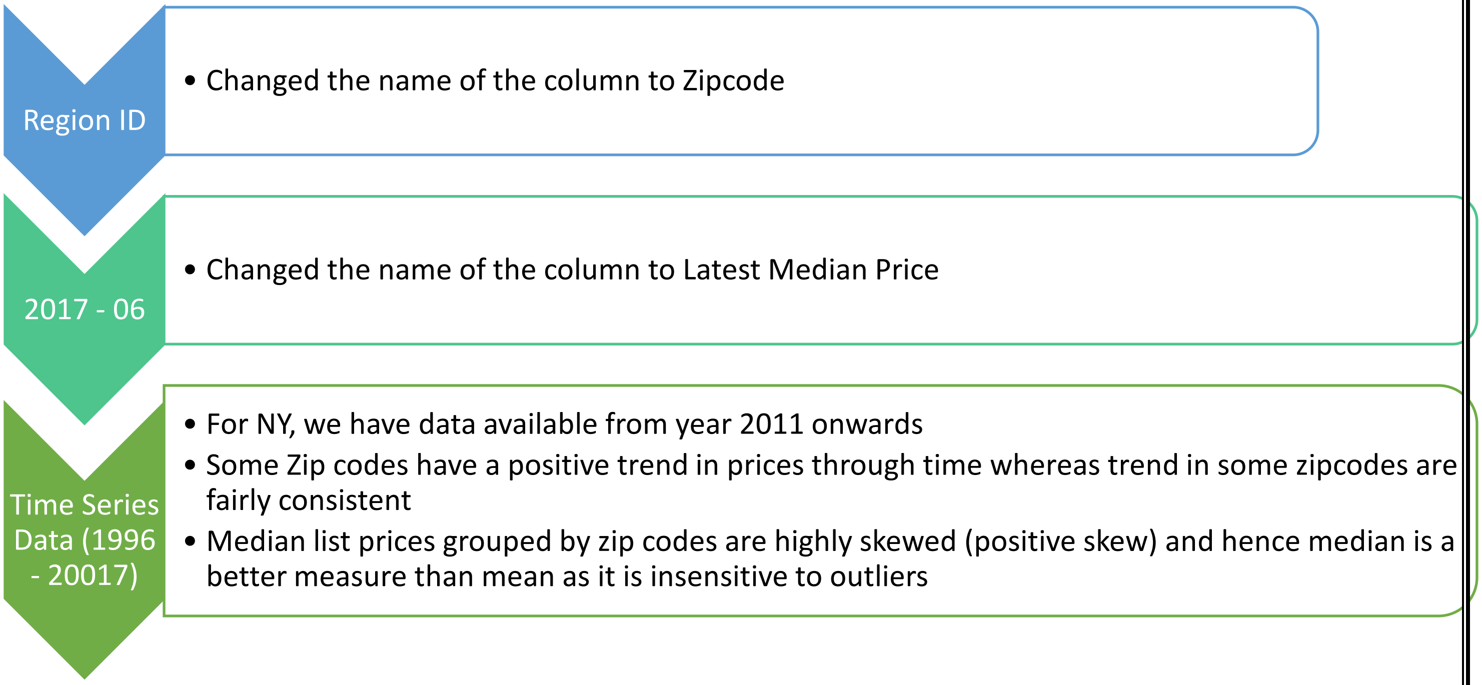


1.3 Feature Selection and Engineering:

S. No	Variables Selected	Variables Created
1.	Neighborhood Cleansed	Annual Revenue
2.	Neighborhood Group Cleansed	
3.	Room Type	
4.	Property Type	
5.	Bedrooms	
6.	City	
7.	State	
8.	Latitude	
9.	Longitude	
10.	Zipcode	
11.	Availability_365	
12.	Price	

2. Zillow Data:

2.1 Data Cleaning and EDA Insights:



Region ID

- Changed the name of the column to Zipcode

2017 - 06

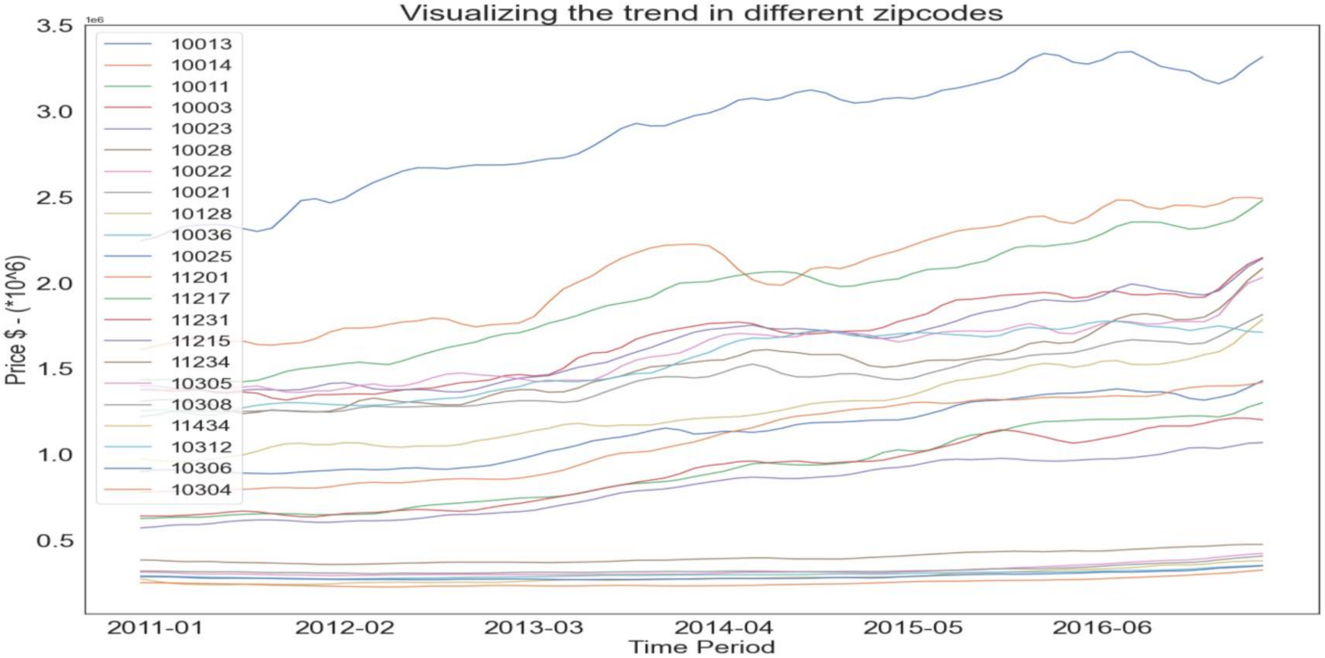
- Changed the name of the column to Latest Median Price

Time Series Data (1996 - 20017)

- For NY, we have data available from year 2011 onwards
- Some Zip codes have a positive trend in prices through time whereas trend in some zipcodes are fairly consistent
- Median list prices grouped by zip codes are highly skewed (positive skew) and hence median is a better measure than mean as it is insensitive to outliers

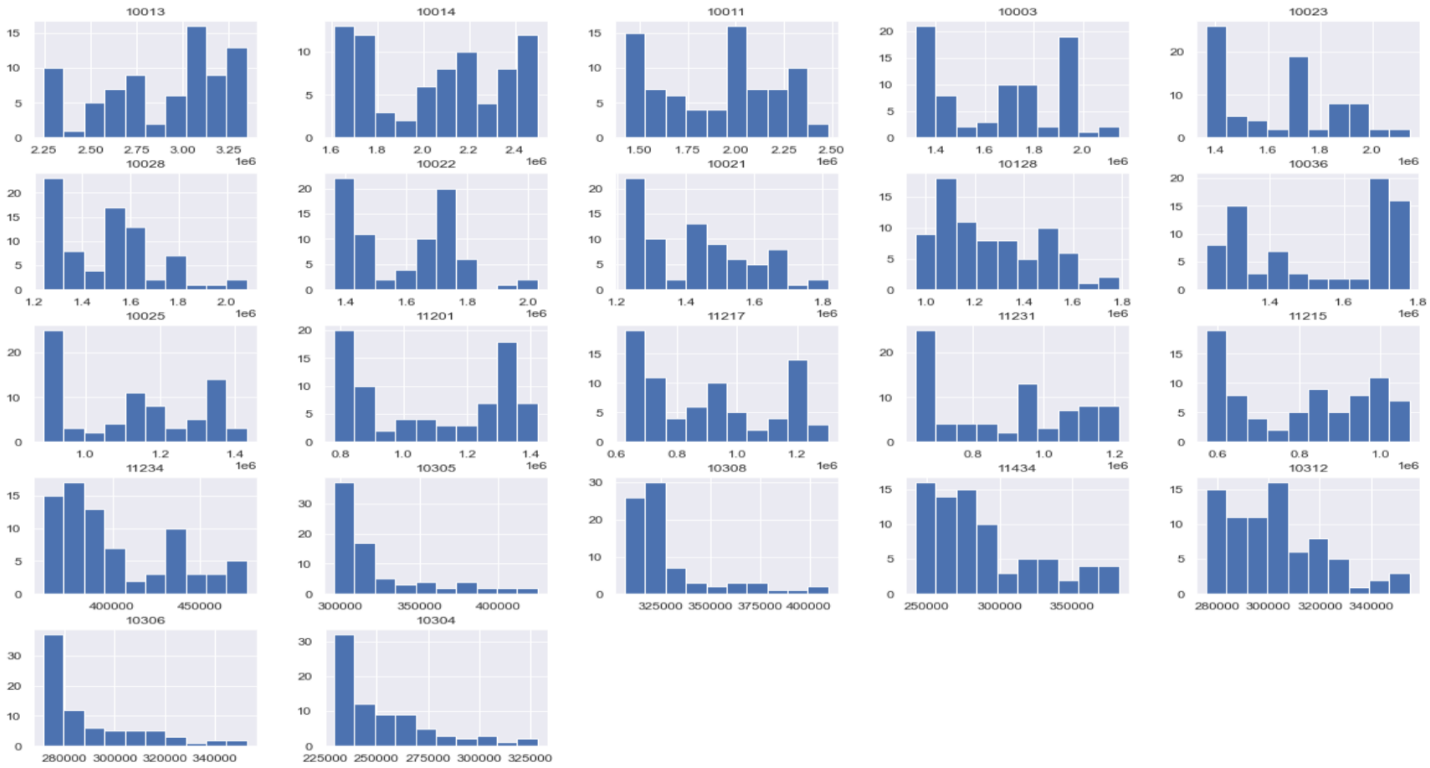
2.2 EDA Charts:

2.2.1 Visualizing Time Series Data



Insights

- For NY, we have data available from year 2011 onwards.
- Some Zipcodes have a positive trend in prices through time whereas trend in some zipcodes are fairly consistent



Insights

- Median list prices grouped by zip codes are highly skewed (positive skew) and hence median is a better measure than mean as it is insensitive to outliers

2.3 Feature Selection and Engineering:

S. No	Variables Selected	Variables Created
1.	Region ID (Renamed to zip code)	NA
2.	2017 – 06 (Renamed as Latest Median Price)	NA
3.	Columns through 2011 – 2017 (Converted to long format for Prophet time series prediction)	

3. Airbnb and Zillow Dataset:

3.1 Data Transformation

a. Joined the Airbnb and Zillow datasets using zip code as the common column.

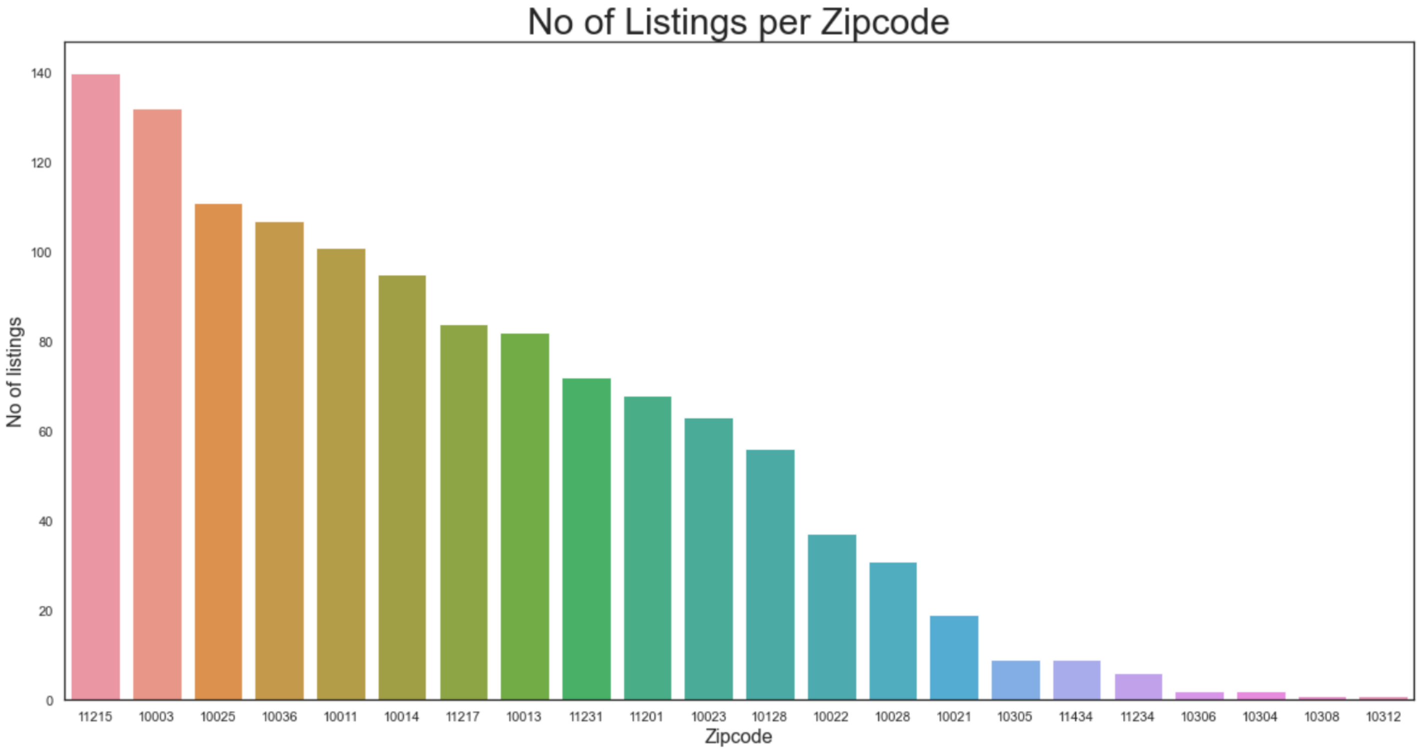
b. Forecasted future house prices from the Prophet model were also joined to this dataset for further analysis.

c. Created separate data frame to hold the listings count value for each zipcode.

3.2 Feature Selection and Engineering:

S.No	Variables Created	Description
1.	Forecast_5yr	Contains the 5 year forecasted value of the house price
2.	Forecast_10yr	Contains the 10 year forecasted value of the house price
3.	Forecast_15yr	Contains the 15 year forecasted value of the house price
4.	Listings_count	Contains the listings count for each zip code

3.3 EDA:



Insights

- Zipcodes like 10308, 10312, 10306, 10304, 11234, 11434, 10305 have lesser than 15 listings whereas Zipcodes like 11215, 10003, 10025, 10036 have more than a 100 listings.
- I have considered zip codes with listings greater than 15 in each zipcode for this analysis.

TIME-SERIES FORECASTING, BREAK-EVEN AND PROFITABILITY ANALYSIS

1. Time Series Forecasting:

1.1 Model Choice – Prophet Algorithm.

a) Prophet algorithm was used to forecast future house prices by using the past time series data given to us.

b) Facebook's Prophet package was chosen because it is extremely handy, open-source and it handles seasonality and outliers very well.

c) Assumptions underlying the Prophet model is that most useful real-world time-series data does not contain any structure beyond trend, seasonality (maybe multiple seasonality like month and year), and causal effects which makes it the better alternative over ARIMA function to forecast time-series data with no major manipulations required to the source data.

d) The prices were forecasted was 4 time periods from the assumed purchase date of the property.

i. 1 year
ii. 5 years
iii. 10 years
iv. 15 years

1.2 Trade-off between the forecasted values:

a)The forecasted prices were then used in further analysis while calculating the Return on Investment (ROI) but I used only the 5-year forecast of the price.

b) Other periods were ignored for this calculation because 10 and 15 years forecasted values may not be the best forecasts as we are trying to predict too far into the future and a 1-year forecast was considered too short for an investment time. Hence, a 5-year forecast was chosen.

2. Break-even and Profitability Analysis:

2.1 Break-even Analysis

a) Break-even analysis was carried out with the help of a custom function that groups the zip codes by median house prices and annual rents for each listing.

b) Calculation:

Payback Period = Initial Investment cost / Income through rent

Where,

Initial Investment cost: It is assumed to be the latest median house price from Zillow data (2017-06 column)

Income through rent: It is calculated by Annual Rent from leasing out the property listing.

Annual Revenue = Price per night * occupancy rate * availability_365

2.2 Profitability Analysis

a) Profitability was measured in terms of ROI on the initial investment cost. The time frame taken for this analysis was 5 years.

b) Custom functions were used to convert the time series data into a long format, predict house prices and calculate ROI.

c) Future house price (the price of the house 5 years after investment) was forecasted using the Prophet time series algorithm.

d) Calculation:

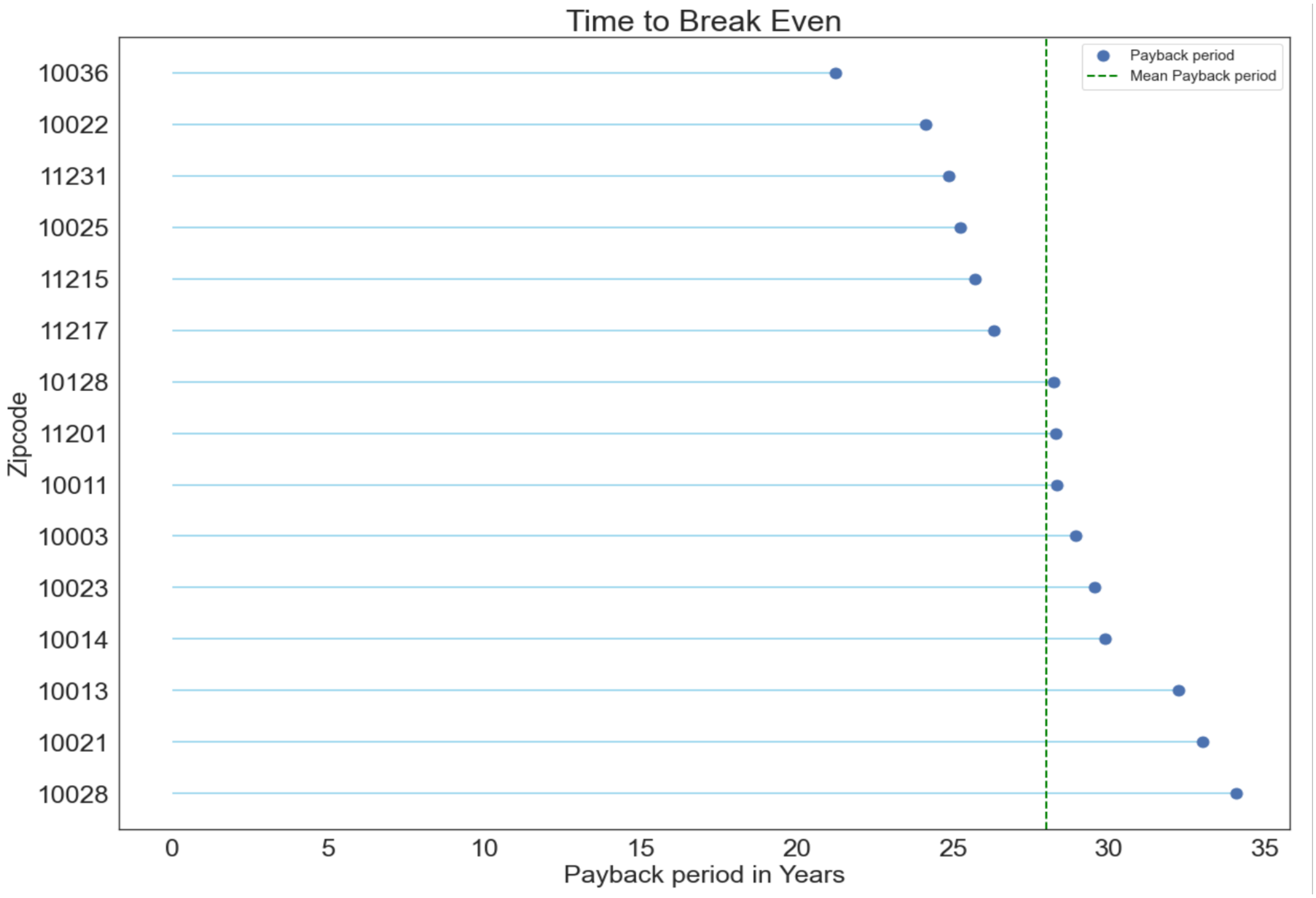
ROI % = Profit / Initial Investment Cost

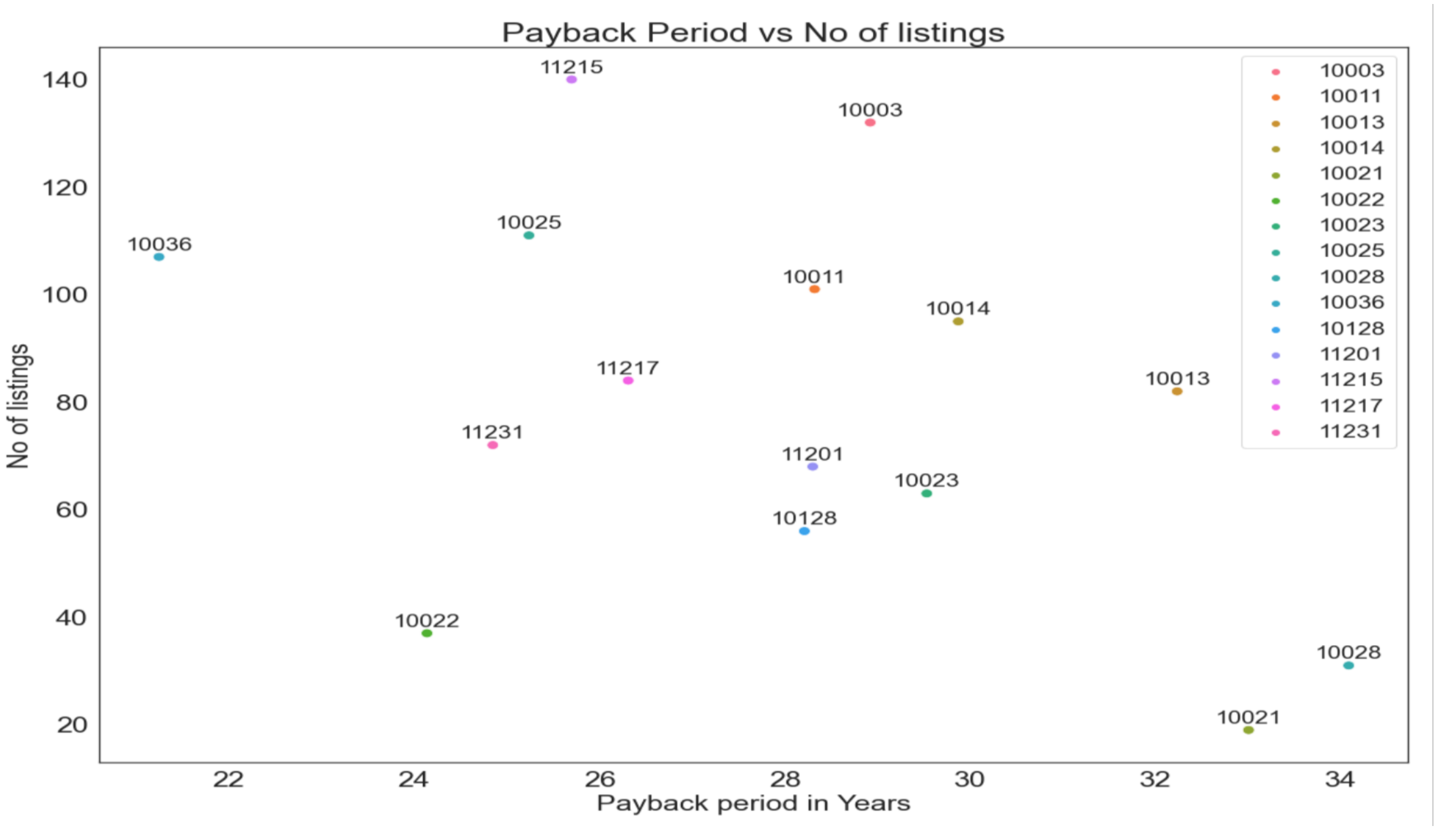
Where,

Profit = Rent for 5 years + Forecasted 5-year house price – Investment Cost

FINDINGS AND CONCLUSION

Based on Payback Period:





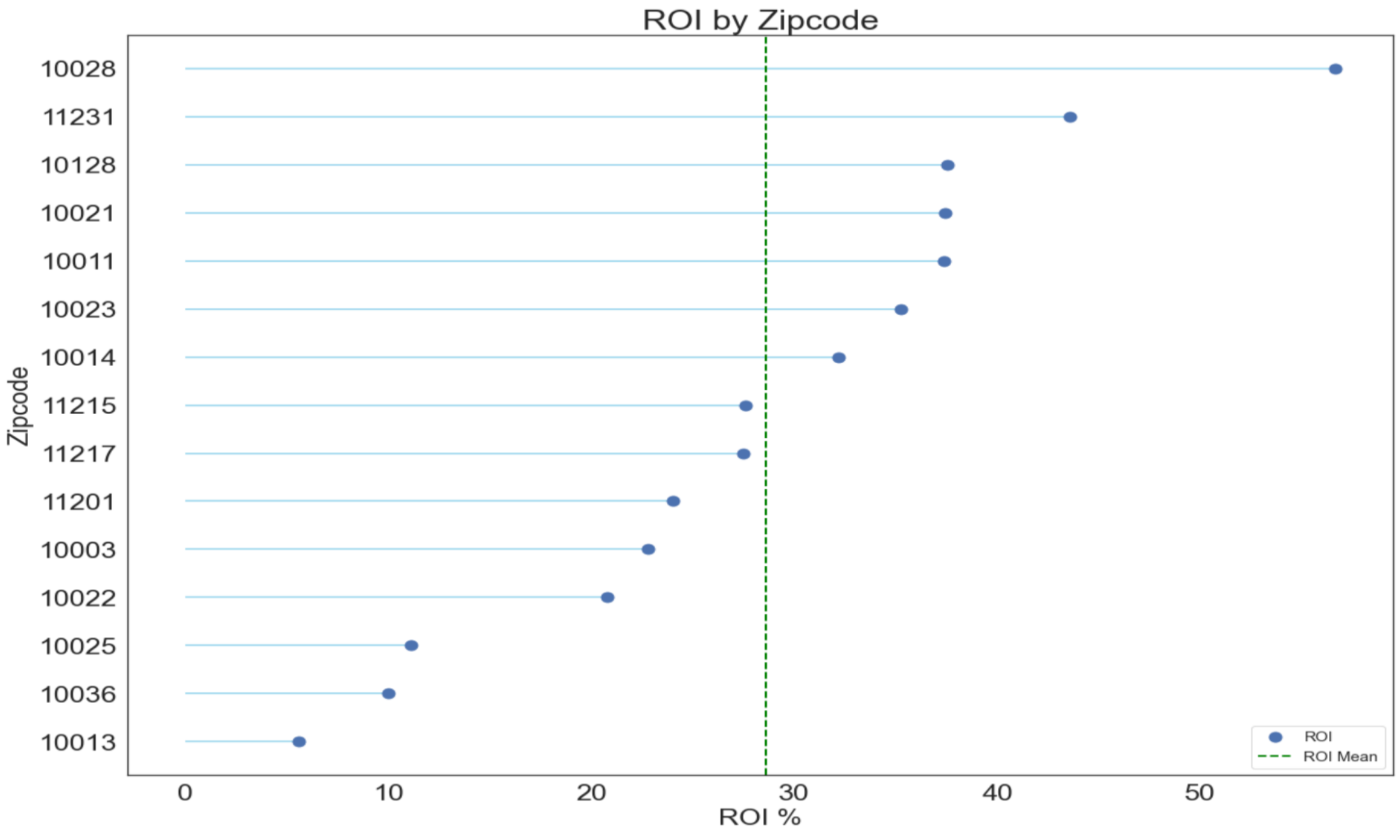
Insights

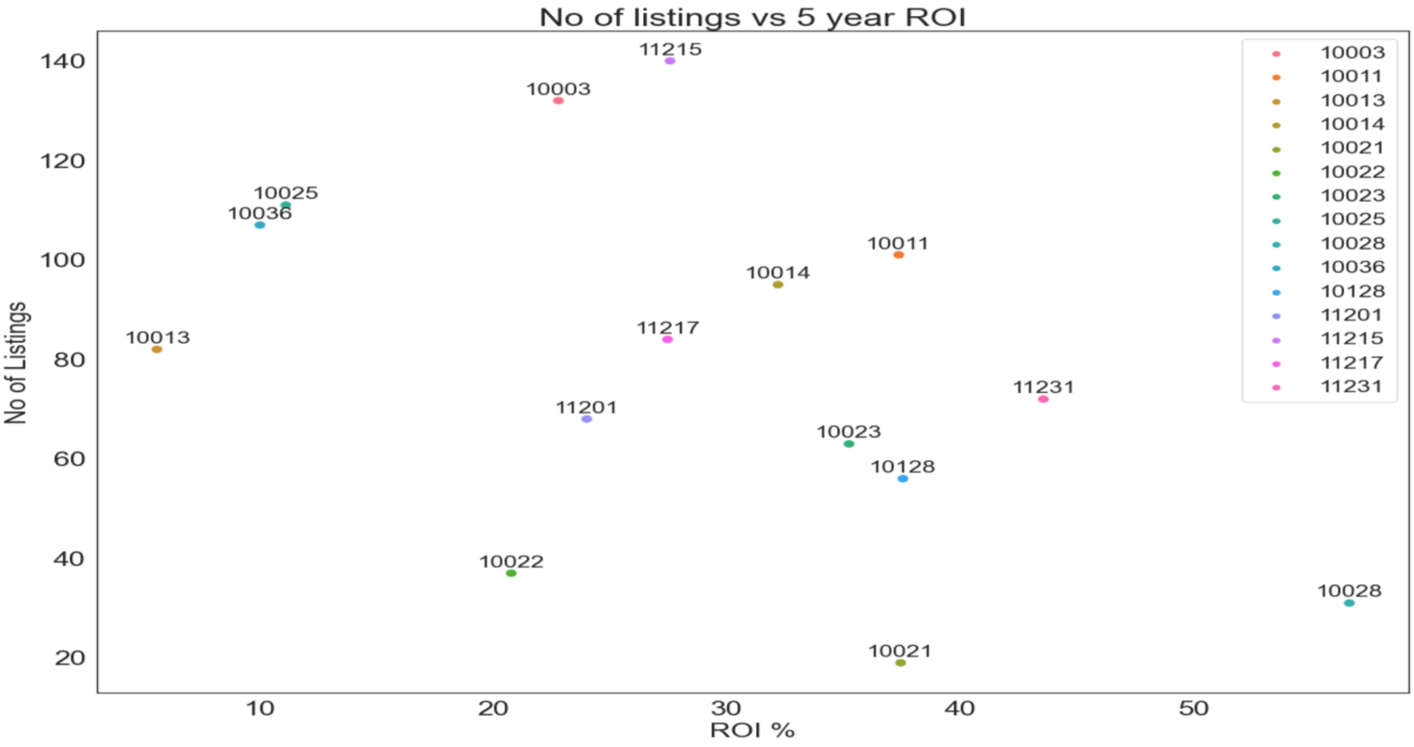
- Average payback period for the NYC listings are 28 years.
- **10036, 10022 (Manhattan)**, **11231 (Brooklyn)**, **10025 (Manhattan)**, **11215, 11217 (Brooklyn)** have quicker break-even time than the average.
- Zip codes with high number of listings seem to achieve break even points quicker than zip codes with lesser listings. This could be because high number of listings have created availability to meet demand in popular locations.

Conclusion

- **10036, 10022 (Manhattan)** and **11231 (Brooklyn)** are the best zip codes based on Payback Period. They are the quickest to break even.

Based on ROI in 5 years:





Insights

- Average payback period for the NYC listings are 28.61 %.
- **10028 (Manhattan)**, **11231 (Brooklyn)**, **10128, 10021, 10011, 10023, 10014 (Manhattan)** produce better returns than the average.
- Zip codes with high number of listings seem to have lesser Return on Investment (ROI) than zip codes with lesser listings.
- Zipcodes like 11215 and 10003 have 100 + listings. Investing in all of them is not a wise choice considering zip codes with higher listings have less ROI. Hence, we need to identify which neighbourhoods perform best in these zipcodes.

Conclusion

- **10028 (Manhattan)**, **11231 (Brooklyn)** and **10128 (Manhattan)** are the best zip codes based on Return on Investment for a 5-year time period.

FUTURE STEPS:

Sentiment analysis of the text columns to understand the attributes of a successful Airbnb listings better.

Square footage information is missing in the dataset. Procuring this data will help us understand price in different neighbourhoods better.

Incorporate weekly prices, multi-day booking discounts, cleaning and other miscellaneous expenses in our analysis to make better conclusions.

Only 24 New York zip codes in Zillow data which restricts our analysis. Securing this data will help us explore other untapped locations in NYC.

We can explore other time series models such as ARIMA by exploring seasonality trends in the data.