Airbnb and Zillow Data Challenge

Author:

Anupam Shukla

MS, Business Analytics(2019-2020)

University of Cincinnati

Table of Contents

Business Problem:

A real estate company plans to purchase 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, the company doesn't know which zip codes are the best to invest in.

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

The publicly available data used for the analysis come from **Zillow** and **Airbnb**.

Cost data : Zillow provides selling cost for 2 bedroom properties in each zipcode for various cities. The cost information is available from April 1996 to June 2017.

Revenue data : Information about property listings in New York including location, number of bedrooms, reviews, price, availability, property description, etc. AirBnB is the medium through which the real estate company plans to lease out their investment property.

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 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.)
- Occupancy rate of 75% throughout the year for Airbnb properties.

Note: Selling cost information of 2 bedroom properties is available only until June 2017. I have performed time series forecasting using ARIMA model to predict price corresponding to data when data was collected for Airbnb

Data overview:

1. The airbnb data contains 48895 rows and 106 columns.

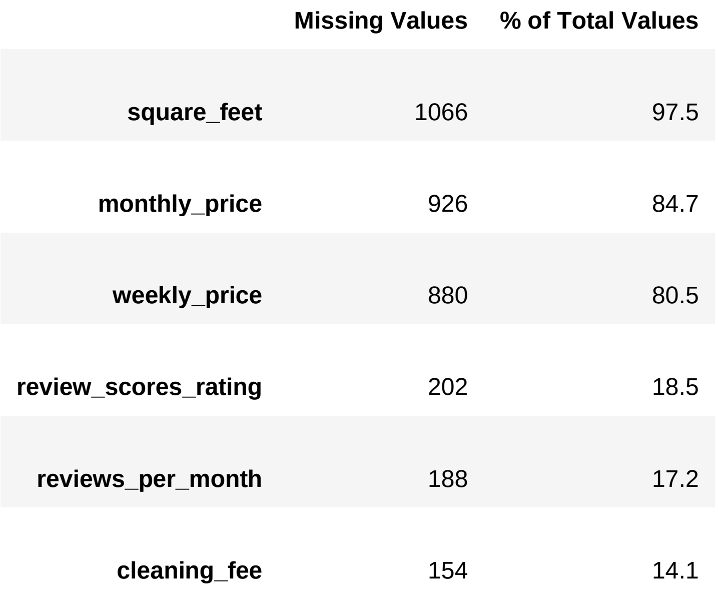
2. The zillow data contains 8946 rows and 262 columns.

3. The combined data from airbnb and zillow contain 1093 records(2-bedroom listings in New York city) and 19 features which is used in further analysis. Number of unique zipcodes in the combined data is 23

Data Quality Issues:

1. Variables price, weekly_price, monthly_price and cleaning_fee contain symbol $ which cannot be used in numerical analysis.Thus, these values need to be cleaned

2. Following variables had the missing values -



	Missing Values	% of Total Values
square_feet	1066	97.5
monthly_price	926	84.7
weekly_price	880	80.5
review_scores_rating	202	18.5
reviews_per_month	188	17.2
cleaning_fee	154	14.1

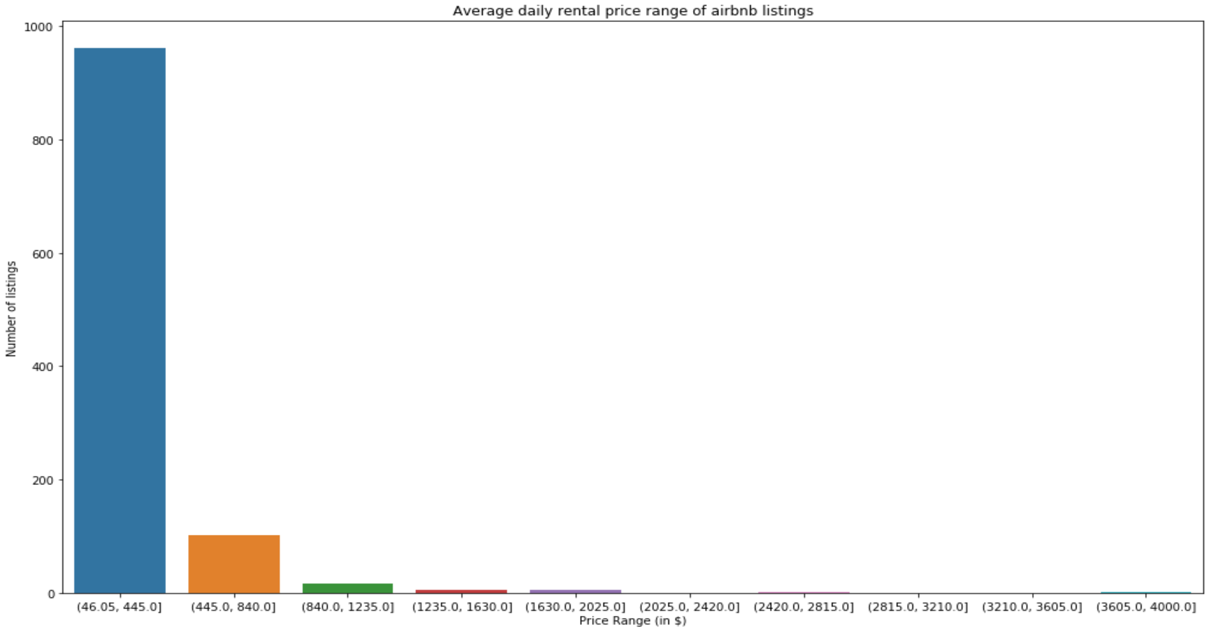
As can be seen, variables square_feet, monthly_price and weekly_price have a lot of missing values so this cannot be used in further analysis. Still, all the values were imputed using IterativeImputer algorithm in python which uses KNN algorithm. The variables review_scores_rating, reviews_per_month and cleaning_fee are used in further analysis.

3. It can be seen that there are two properties with square feet less than 200.In fact these values are 0 and 3 which is incorrect.

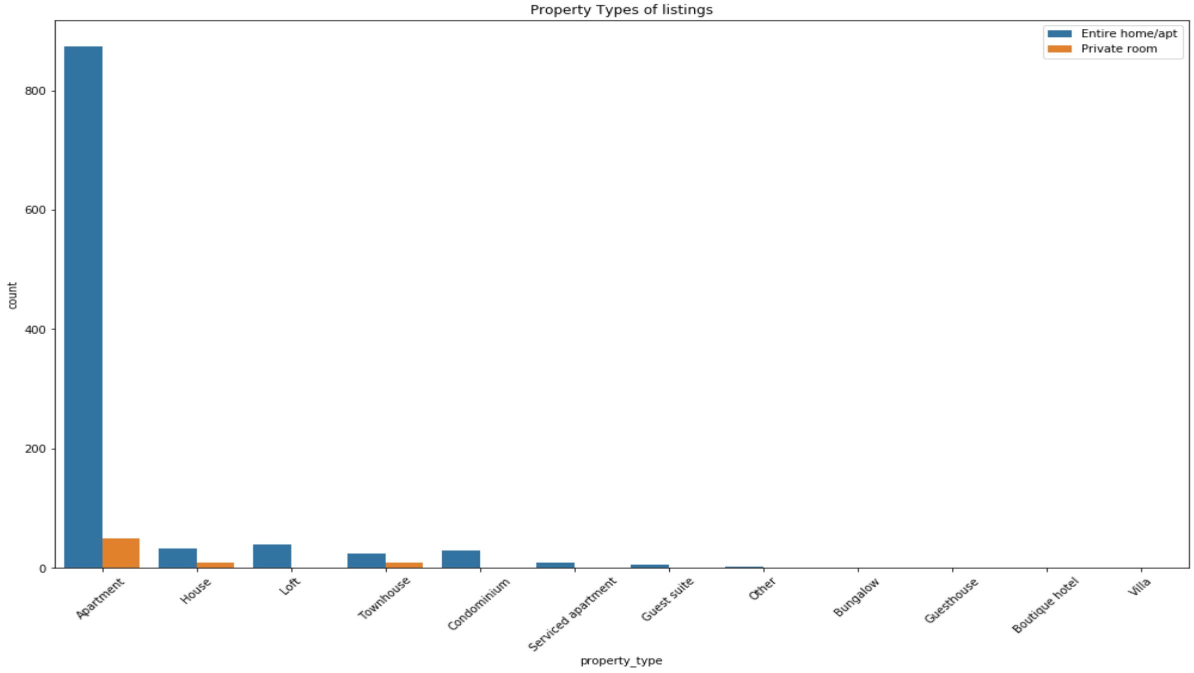
4. Zipcode 10013 was found to be both Manhattan and Brooklyn neighbourhoods in Airbnb listings

Exploratory Analysis and Visualizations:

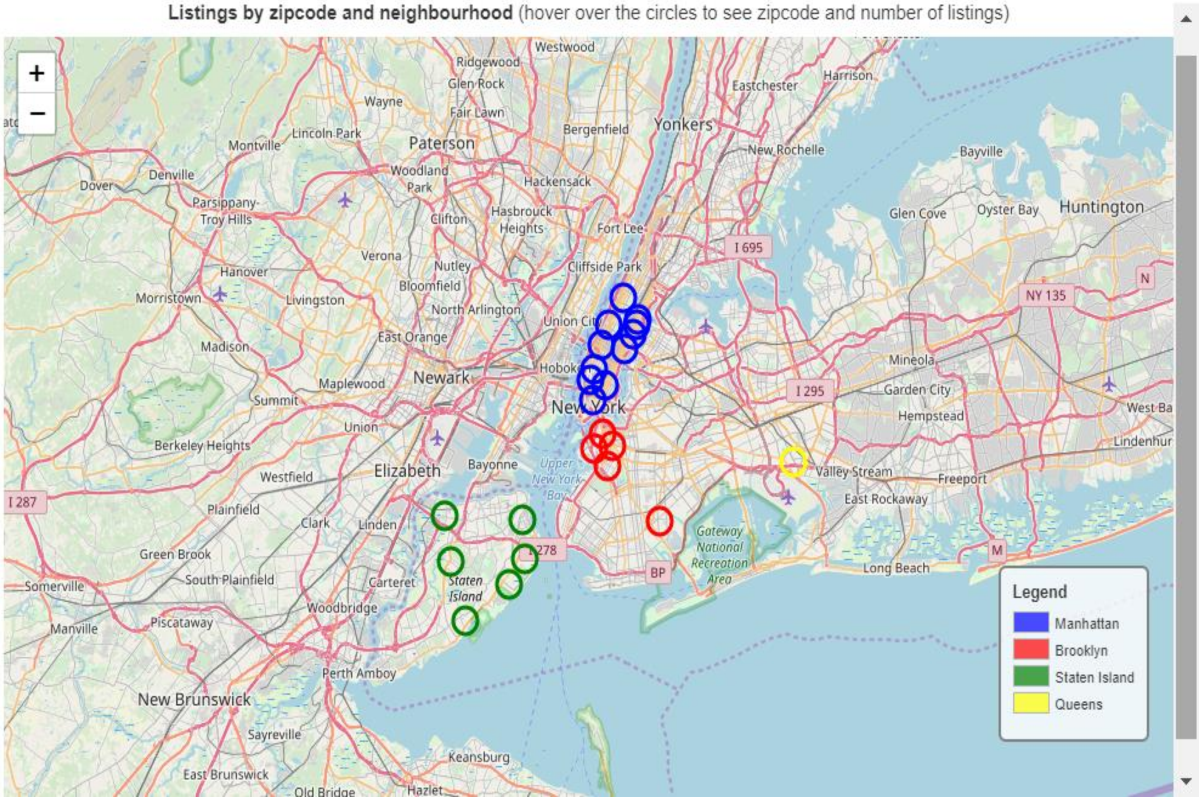
1. Average daily rental price range of airbnb listings



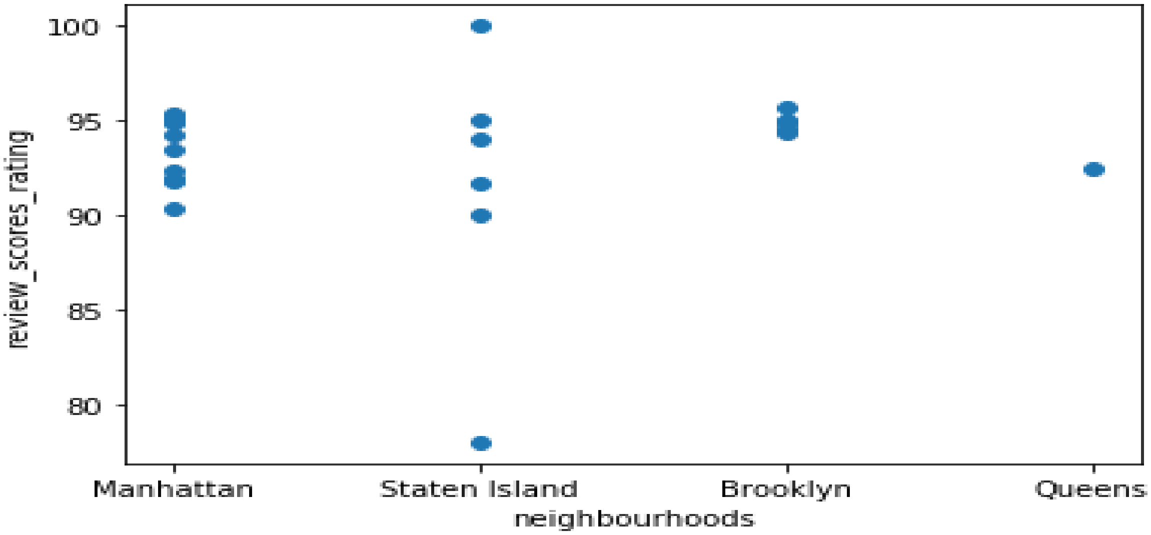
2. Property Types of listings



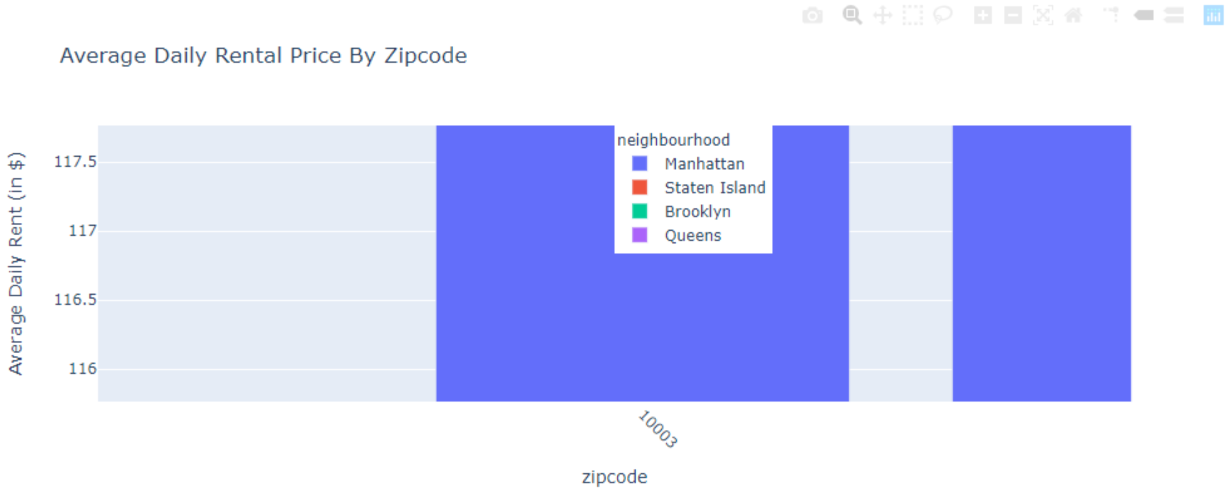
3. Most popular neighbourhoods



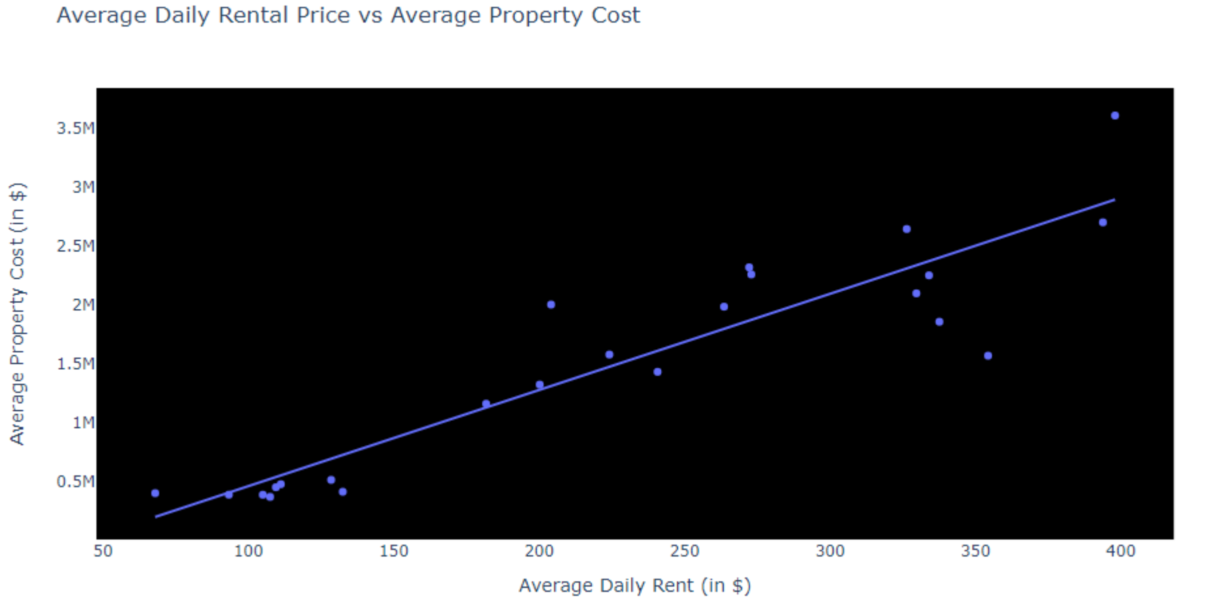
4. Average review ratings by zipcodes in different neighbourhoods



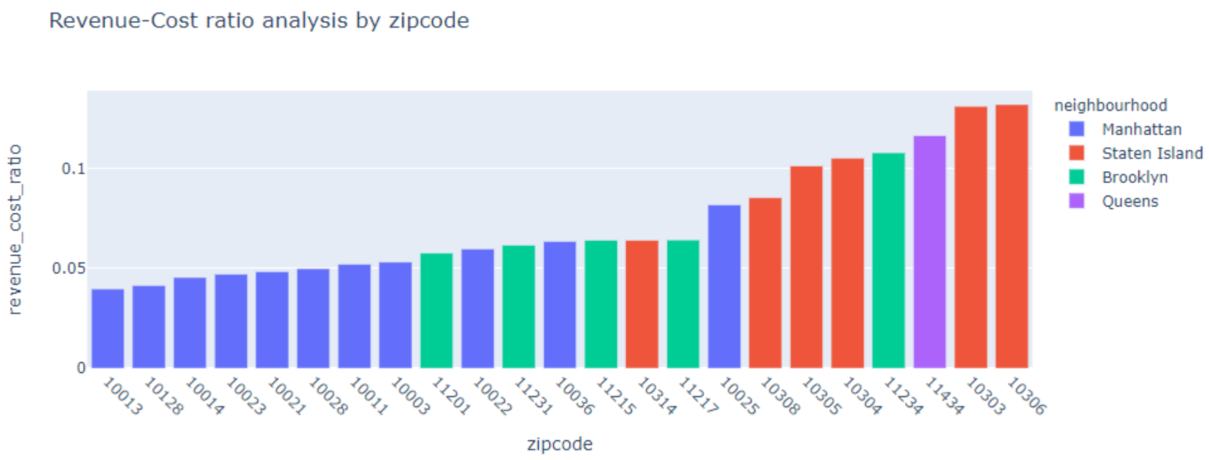
5. Average daily rental price by zipcode



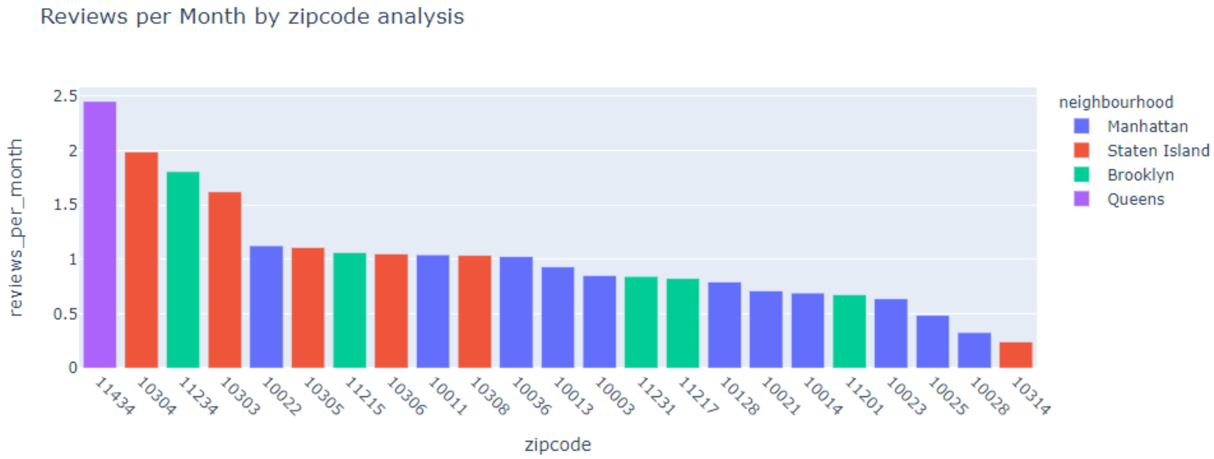
6. Rental prices vs Property Costs



7. Revenue-Cost ratio analysis by zipcode



8. Reviews per month by zipcode analysis



Observations:

1. 23 zipcodes in New york city have atleast 1 listing on airbnb

2. The average lowest daily rent price is 68 and highest is 398

3. There is a zipcode which has average review score rating of 100

4. The zipcode with the highest rating also has the maximum number of reviews

5. The average cheapest property for a zipcode is priced at 371,711 while the most expensive property is priced at 3,682,336

6. The lowest average annual revenue is37,394 while the maximum revenue is 143,307

7. Manhattan and Brooklyn are the most popular neighbourhoods with more zipcodes having airbnb listings. Queens has only one zipcode with airbnb listings

8. Review scores are on average between 90-100 for all zipcodes.

9. Review scores of zipcodes in Manhattan and Brooklyn are between 90 and 95 on average while there is one zipcode in Staten Island where the review rating is less than 80

10. The average rental prices are highest for zipcodes in Manhattan followed by Brooklyn

11. There is a positive linear correlation between the average daily rental price and average property cost which means that the zipcodes with high prices will charge higher rents

12. The revenue to cost analysis reveals that the this ratio is lowest for zipcodes in the Manahattan neighbourhood, which means investing in these neighbourhoods will not be very profitable. On the other hand, zipcodes 10306, 10303, 10304 in Staten island; 11434 in Queens and 11234 in Brooklyn are the top 5 zipcodes in terms of revenue-cost ratio.

13. Even by the measure of number of reviews per month, the zipcodes 10306, 10303, 10304, 11234 and 11434 are in top 7. We have already observed that the ratings for most zipcodes is in the range of 90 - 100, hence it is more likely that these zipcodes will continue to attract customers because of their high ratings.

Conclusion:

In terms of best revenue to cost ratio, lesser breakeven period and good reviews, following zipcodes should be considered for investment -

- 10306, 10303,10304 in Staten Island

- 11234 in Brooklyn

- 11434 in Queens

Metadata:

2 files are included with this submission:

1. **Airbnb_Zillow_data_main.ipynb** – This file contains all the analysis including data cleaning, exploratory data analysis and visualizations, and observations and conclusions

2. **Arima.ipynb** - The file contains code to determine the AutoRegressive(p) and Moving average(q) terms to use in the ARIMA model to forecast current property prices. The values calculated are used in the Airbnb_Zillow_data_main.ipynb

Following additional fields and functions were created to be used in this analysis

Field	Description
price_range	Range of daily rental prices. Created by dividing daily rental price into 10 bins
avg_annual_revenue_per_zipcode	Average annual revenue generated by listings in each zipcode
revenue_cost_ratio	Average revenue to cost ratio for each zipcode
years_to_breakeven	Average breakeven period for each zipcode

Function	Description
currentcost(ser,monthsahead)	Takes time series data and number of months from the last date in time series date for which forecast is to be made
Preprocessdata(zillow,city)	Takes zillow data and city as parameters and returns a dataframe with RegionID,RegionName,SizeRank and CurrentPrice
missing_values_table(df)	Calculates missing values and percentage by column
years_to_breakeven	Average breakeven period for each zipcode

Software used:

Python 3.7 was used to complete this challenge.

Below packages were used –

- numpy - for array manipulation
- pandas – for data exploration
- matplotlib.pyplot – for visualization
- seaborn – for visualization
- KNNImputer (sklearn.impute) – for imputing missing values
- Plotly – for visualization
- Folium – for map visualization