Dale Ashna Martis

AirBnB & Zillow Data Challenge – Summary & MetaData

Problem statement

To build a data solution to assist a real estate company to understand which zipcodes are profitable for short term rentals within New York City

Tools

Python 3.7.4,
Jupyter notebook 6.0.3

Datasets

Airbnb Dataset: This is the revenue dataset that has Airbnb listings with detailed information about each listing like property type, host information, ratings, reviews, price

Zillow Dataset: This dataset gives the cost information of 2-bedroom properties in a particular zip code from the year 1996 to 2017

Given Assumptions

- The investor will pay for the property in cash (i.e. no mortgage/interest rate will need to be accounted).
- 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.)

Assumptions Made

- The price column in the airbnb dataset has extreme values. From the boxplot it can be seen that there are many outliers. From a business point of view, this will be treated as an erroneous input value and is discarded
- Availability for next 365 days is a status quo as far as occupancy in concerned. Lesser than availability, higher the occupancy. Occupancy rate is calculated based on this column
- Weather/Seasonality has little or no impact on number of bookings
- Number of reviews column is indicative of popularity among customers and the reviews are positive
- The zillow dataset has data from 1996-04 to 2017-06, and the airbnb data has been scraped in 2019. In

order to perform the analysis and have the both datasets in the same unit of time frame, ARIMA modelling is used to predict median price for zillow dataset till year 2019

Package loading

Started with loading the packages that will be required for data cleaning and charts

Pandas: Data manipulation

NumPy: Scientific computation

Matplotlib: Plotting charts

Plotly: Interactive visualization

Itertools: Functions creating iterations for efficient looping

Statsmodels: For time series analysis and forecasting

Functions

Created functions for the code that will be repeated and placed it in this section. Following functions are created:

- Function to read data in the data frame
- Function to filter data based on a value
- Function to calculate the percentage of null values in a column
- Function to get zipcode in the correct format
- Function to calculate the occupancy rate
- Function to transpose a dataframe
- Function to create iterative combinations of p,d and q values
- Function to iterate through the p,d,q,s values in order to find out the model with the least AIC value and best fit
- Function to create and train ARIMA model
- Function to predict a variable using ARIMA forecasting technique

Data Setup

As we have to do analysis of zipcodes only for New York city, filtered the Zillow data set only to New York city and removed the unnecessary columns that are not required for the analysis. Also filtered the dataset

from 1996-04 to 2017-06 as they had mostly NULL values and was not significant for our analysis
For the Airbnb dataset, removed the descriptive columns. Also, filtered the Airbnb dataset to 2 bedroom listings, as the Zillow dataset has cost data of 2- bedroom properties.

Forecasting with ARIMA

The last scraped data for the Airbnb data set is of 2019. In order to get both the datasets in the same unit of timeframe, forecasted the cost for months- 2017-06 to 2019-12 for Zillow dataset using ARIMA modelling. Then, filtered the Zillow dataset for the year 2019 only.

Now, both the datasets are for New York city, 2019 year and 2-bedroom properties.

Data Quality Check

After setting up the data, checked for missing values, duplicates, datatypes. Also corrected the price and zipcode formats.

Missing values: Calculated the percentage of the missing values in the respective columns, and the columns with missing values greater than 50% were removed. It was found that the monthly and weekly price columns had 80% missing values, hence they were dropped and 'daily_price' column was selected for the revenue analysis which had 0 missing values. Also, the zipcodes which had 'is_location_exact' column false were dropped.

Datatypes: The zipcode column was integer which was needed to be converted to string as it is supposed to be categorical

Checked for **duplicate** zipcodes in zillow dataset, as it needs to have one cost price for each zipcode

The **zipcode format** in both datasets were incorrect as some of the zipcodes had 10 characters. So created a function to trim the zipcodes to 5 character length for both datasets

Price correction: The price column had dollar and comma sign, which needed to be corrected. Passed the values of price column to a lambda function to get it in the correct format that was needed

Data Merging

Merged the cleaned datasets using matching zipcodes. The merged data set had 24 zipcodes that needed to be analyzed

Outlier Analysis

As the analysis is based on zipcode level, it becomes necessary to average the price related columns zipcode wise.

So before starting with the analysis, this is an important step to check for outliers in the price column. So, plotted the box plot to check for outliers. There were outliers present for every zipcode, which needed to be taken care of.

So first calculated the upper and lower quartiles using the quantile function of pandas

Then calculated the Inter quartile range using the formula **IQR = Q3 -Q1**

Finally calculated the range for the outliers using the formula **Q1 – (1.5*IQR)** and **Q3 + (1.5 * IQR)**. The price values outside this range were then removed

Exploratory and Profit Analysis

The dataset is ready for analysis. To select the best zipcodes to invest, evaluated the zipcodes and plotted charts to check for the following conditions:

- **Number of properties in each zipcode** – Calculated the count of properties in each zipcode using the count method of pandas. Then plotted the count for each zipcode using the bar plot. This chart gives information about which zipcodes have more options available for the customers.
- **Number of reviews received for each zipcode** – Calculated the number of reviews for each zipcode using groupby method of pandas. Then plotted the count of reviews for each zipcode using the bar plot. This chart gives information about which zipcodes are popular among the customers.
- **Correlation Matrix** - Found strong positive correlation between Availablity 30, Availablity 60, Availablity 90 & Availablity 365 columns. This indicates that the booked out apartments tend to remain booked throughout the year, and free apartments tend to remain free throughout the year.

This means we can consider the availability 365 column now, and assume that the availability is going to be consistent throughout the entire 365 day period, and use this to calculate the occupancy rate

- **Occupancy rate for each zipcode** – Calculated the occupancy rate using the formula: days booked/ total days in a year. Then plotted the occupancy rate for each zipcode using the bar plot. This chart gives information about which zipcodes have the maximum properties booked throughout the year.
- **Revenue** – Calculated the revenue using the formula: Price* 365* occupancy rate. Used scatter plot to check for correlation between cost and revenues of the properties in each zipcode.
- **Breakeven period** – Calculated the break-even period using the formula Cost (In Dollars)/Revenue (In Dollars). Plotted the scatter plot of Breakeven period by Revenue to identify the zipcodes in the most profitable quadrants for short term and long term rentals.

Recommendations

- Properties in the zipcode: **11231, 11217 and 11201** from Manhattan, **10036 and 10025** from Brooklyn are best for short term rentals as they reach the breakeven - point sooner. This is because properties in these zipcodes are brought at lower cost and rented out at a high price which results in good revenue.
- The above mentioned zipcodes not only reach the breakeven period early, but are also reliable as these zipcodes have good booking rate throughout the year. Also they have more number of properties and hence more options to choose from. They are also popular as they have received good number of reviews
- From the scatter plot of cost - revenue analysis and profit analysis chart, we can see that the following zipcodes: **10013, 10003, 10011, 10014, 10023** have the highest revenue. The cost price of the properties are higher in these zipcodes and it takes more number of years to reach the break-even point,but once it gets there these properties will generate the highest revenue. So if the real-estate goes for long term rentals, it will be a good idea to invest in these zipcodes

Next Steps

- New york city has around 145 zipcodes, but in the zillow data set there were only 25 zipcodes because of which we had to limit our analysis only to these zipcodes. If we had all the zipcodes for analysis, we could have got more profitable zipcodes for investment.

- Ignored the descriptive columns for airbnb. These coulmns could be considered for sentiment analysis. By using the descriptive data, we could find the probability of a user booking the properties in a zipcode.

- Zillow dataset had only the cost price of the properties in each zip code. It didn't account for the other maintenance prices like cleaning etc. Including all these prices would give a better profit analysis

Column Level Metadata – Information of newly added columns in the dataset

Field	**Description**
Q1	25^{th} percentile of price in each zipcode
Q3	75^{th} percentile of price in each zipcode
IQR (Q3 – Q1)	Inter quartile range of the price column
Occupancy rate (days booked/total days)	Gives the percentage of days the property was occupied in a year
Revenue (Price * Time * Occupancy rate)	It gives the amount of revenue ($) received from each zipcode by renting the houses per year
Break -even period (Cost in dollars/Revenue in dollars)	Gives the number years required to reach the break-even point