Documentation Data Challenge Capital One

• Key Points:

The analysis of this challenge has been done in jupyter notebook with use of interactive components. Please install requirements.txt to install any dependencies on your local system or feel free to run this repository on Binder- an interactive platform to run and view jupyter notebooks.

The order of viewing this project could be:

- Data processing. ipynb
- Decision Point1.ipynb
- Decision Point2.ipynb
- Decision Point3.ipynb

The notebooks contain sufficient explanations intent and insights in the analysis. Please reach out directly at shiva.sync@gmail.com for any issues.

• Assumptions and Considerations:

1) I have restricted this Data Analysis to 2-bedroom housing properties in New York.

2) The houses with room type "Private Room" in Airbnb data in a 2-bedroom setting apartment has a price for only 1 night's stay in room. This has been multiplied by 2 for such properties as if rental management can potentially earn double rent of this property as it has 2 bedrooms. However, the overall rental price per night may not be exactly double as the other room may be smaller or have lesser or more appeal than the room in the listing.

3) Occupancy for the Airbnb listings have been assumed to be 75% throughout the year

4) The rent of each property is assumed to increase every year by the rate at which the value of property appreciates. The calculation of this rate has been discussed section 2.

- # Metrics Created:

1) **Annual_Rent:**
2) This is annual rent collected from a property in Airbnb assuming 75% occupancy using rent price per night from Zillow data. This value is assumed to be fixed.

3) **Payback_Period:**
4) This is the number of years required to earn back the initial cost of investment from annual rental income from property assuming 75% occupancy throughout the year.

5) **Occupancy:**
6) 0.75*365 (Assumed occupancy of all properties in Airbnb)
7) **Rate:**
   So, we all are aware that properties value appreciate/depreciate depending upon the neighborhood and locality. In the Zillow data we have the average housing prices for a zip code for as far back as 1996 but these columns have lot of missing values. However, we do have complete data from 06-2007. The difference between in prices in 06-2007 and 06-2017 divided by 10 has been taken as an average rate by which a property rises in that zip code. Although a naïve approach, it still captures the essence of appreciation/depreciation of property by zip codes over the years and would be useful in adding a location specific factor in Return on Investment calculations.
8) **Roi_without**: This metric is return on investment in 10 years assuming rent appreciates at a constant rate without including the equity value of the property.

   It is calculated as follows:

   (Annual_Rent(1+Rate/100) ^10)/Current Price of the property

9) Roi_10: This metric takes into account the return on investment through rental as well as equity appreciation i.e. the increase in the price of the house is also taken into account to calculate roi. This has been calculated for 10 years as well.

   It is is calculated as:

   A= (Annual Rent(1+Rate/100) ^10)/Price of the property {Rent amount in 10 years}

B= (Price of Property(1+Rate/100) ^10)-Price Of property {Increase in cost of property in 10 Years}

Roi_10= (A+B)/Price of Property

- **Usefulness of these metrics**:

1) Payback_Period : This tells us how fast we can recover investments from our property. This may be one of the conservative estimates to evaluate expected profits from a property.

2) Roi_without : This metric is a good metric to assess profitability from a property. Because of variable rates of appreciation for each zipcode this metric gives a good estimate how location of zipcode determines ROI of a property

3) Roi_10 : This metric would be useful to assess overall value of investment keeping in mind the increase in value of asset and also opens the value analysis in case rental Property decides to sell the current asset after 10 years.
It is expected that Roi_without would be very similar in results in ROI_10 as it has high equity appreciation has same rate of increase as rate of rentals but ROI_10 . But some interesting insights can be drawn from differences in behavior of these two metrics for zipcodes.

Data Quality and Processing

Key Points:

1) select columns of interests in our analysis

2) Handling of missing values in the Data and doing necessary imputations.

3) Subset the data further as per needs for visualization

Methodology:

Airbnb Data:

The biggest concern for our analysis was imputations for missing zipcodes. Once we have data for zipcodes the city and state can easily be inferred from zipcode value. We used pygeocoder a library built on google

maps api to impute missing zipcodes by using latitudes and longitudes. Luckily no missing values on lat and long. After imputation only 2 properties still had missing values as their coordinates were not too specific for api to give a zipcode. They were ignored and removed.

Once we imputed zipcode a quality check was done on zipcodes to ensure proper string length and then a range of zipcodes between 10000 and 20000 were selected as properties in NY.

Next we noticed 22 missing values for bedrooms which were removed from the dataset as their was no way to estimate the number of bedrooms.

finally we subset the 2 bedrooms properties in NY CITY.

This data set was then aggregated to be used for decision point 1 .

Zillow Data :

This dataset was filtered for properties in NY as no missing values were present on city column. Then we selected the columns to be used for analysis.

We took the column 2016-06 as current selling price of the property. In this filtered data set we did not find any missing values.

However missing values were present in certain price years, but we did not use them in our analysis. While calculating rate, the farthest two columns were considered which had complete data in their columns.