Capital One Data Challenge

Airbnb Zillow Data Analysis

By: Kartik Nagarajan

**ABOUT Airbnb Zillow data set**

I was provided with two data set

Cost (Zillow data set): Zillow provides us with an estimate of value for two-bedroom properties

Revenue (Airbnb Data Set): Airbnb is the medium through which the investor plans to lease out their investment property.

Assumption: Occupancy rate in 75% ( I have used a calculation in my R code to do that) which is attached in the file.

**Tools and softwares used**

**Software Required:**

a) Latest version of R language and RStudio for running and executing the markdown file.

b) Latest version of python and its libraries which are mainly used for sentimental analysis

Please follow the below instructions to run the markdown file and generate the data case study output html file

**Step 1**: The main file to run is called **Capital one data challenege.rmd**

Please keep this file in the same folder along with "listings.csv.gz" and "Zip\_Zhvi\_2bedroom.csv" files i.e the folder should contain the data files with the **Capital one data challenege.rmd** file. For listings data directly use the "listings.csv.gz" file without extracting

**Step 2**: Open the **Capital one data challenege.rmd** using RStudio app and then click on knit icon on the top and generate the html file. I am already attaching the knit file in html format so you could open it in the web browser or in R itself

**Step 3**: Although the **Capital one data challenege.rmd** file itself contains code to install and load all the necessary packages but in case there is an error then please make sure below packages are installed and loaded.

* tidyverse
* dplyr
* colorspace
* plotly
* tidyverse
* dplyr
* colorspace
* forecast
* data.table
* Data explorer
* Ggplot2
* MatrixModels
* kableExtra
* ggthemes
* UpserR
* Ggally
* Naniar
* Visdat
* corrplot
* ggpubr
* ggboxplot
* xgboost

Incase if you don’t have the package you can install it with using the command install.package(‘x’);

The html file attached (**Capital one data challenege.html**) describes all the steps taken ranging from Data Cleaning , Missing value handling , missing value imputation , types of plots used to explore and derive insights from data and in the end apply models on the clean data to predict price.

Step 4: I am also attaching a python file **python code- sentimental analysis.html** where I have performed the sentimental analysis after selecting a few descriptive variables. The main reason we performed sentimental and word cloud analysis is explained in another file attached. Following are the libraries that must be installed in order to run the code in system.

* Pandas
* Numpy
* Nltk
* Textblob
* Wordcolud
* Matplotlib

Step 5: Apart from all the visualizations done in R, I have again used the Airbnb and Zillow dataset in order to **develop interactive dashboards in Tableau**. I have made use of Table calculations, Calculation Fields, filtering, Level of Details, Data Cleaning, interactive maps using Tableau Prep and finally performance tuning in order to make sure Dashboard performance in not effected. The main reason to visualize in Tableau in to leverage the strong data visualization engine which provides numerous methods of visualizing data in a more effective manner in order to develop business insights.

The workbook is of the latest Tableau version 2019.4 and cannot be used with older version. Incase if you require an older version Workbook, please intimate me so that I could visualize in that version.

**Conclusion after performing Data analysis in R , Sentimental Analysis in Python and Visualization in Tableau are as follows:**

It can be concluded from the above analysis that When the four neighbourhoods are compared, we can conclude that Manhattan and Brooklyn have higher number of properties listed and have higher number of zipcodes when compared to Queens and Staten Island.

We also observe a huge difference in price per night for properties in Manhattan and have a few outliers. Queens and Staten Island in contrast have a consistent price per night for the properties listed.

Zipcodes in Manhattan have properties with the highest price per night asking.

The mean property cost of properties in Manhattan is very high, and so is the price per night of stay. Hence this neighbourhood has the highest breakeven period.

We observe that Staten Island zipcodes have lower mean cost of properties, but fairly high price per night making it the most profitable neighbourhoods.

We can also conclude from our analysis that Brooklyn zipcodes are doing a pretty decent job in terms of breakeven period. Properties have average cost in this zipcodes and the price per night on Airbnb also are neither too high nor too low

Our sentimental analysis tells us whether the summary provided by the owner for a room/apartment appeals the customers in a positive or a negative way. The word cloud shows us the intensity of the words that tells us about the word frequency. So, in this case we have made word clouds for the positive as well as the negative summary. This would help the owners to change their way towards the marketing side, so that they can use the most used positive words in order to increase their customer staying at their place resulting in more revenue for themselves

**Note:**

Incase if there is an Issue in Opening the Tableau workbook , I have uploaded my visualization on my Tableau Public Profile as Well whose link is <https://public.tableau.com/profile/kartik2802#!/>

**Future Scope:**

1. We can incorporate crime data in NYC to see how crime at a particular area affects Airbnb bookings and the rental price
2. There are many websites that provide weather data, we could also check the effect of different weathers on a rental price of Airbnb in NY city for e.g. how the bookings fare in summer vs winter
3. We could implement holt winter forecasting method using Tableau in order to predict the future prices based on zip codes in order to understand niche areas to invest in .
4. Implement the animation feature of tableau in order to improve the visualization and make it more attractive for the end user to see and understand.
5. We could do data analysis of top competitors of Airbnb and find impactful data insights which could be communicated to stakeholders in order to strategize their next move.
6. Social Media Data- The real estate company can scrape through the social media platformsdata to identify the impressions a particular neighbourhood is having and based on popularity can make a better decision.
7. **Traffic and Ease of Public Transport Data**- If a zipcode is heavily congested with traffic or if it doesnt have proper public transport channels neareby, then there are high chances that toursits will avoid such zipcodes. So analysis of this data needs to be performed.