Coursera Capstone Project

Applied Data Science

Project Title: Analysing Melbourne Housing Market



By: Zin Myint Naung

**This report only includes "Data" section of the final report.

2. Data

In order to analyse, we need to define our data set for Melbourne's housing price. The data set needs to have following columns:

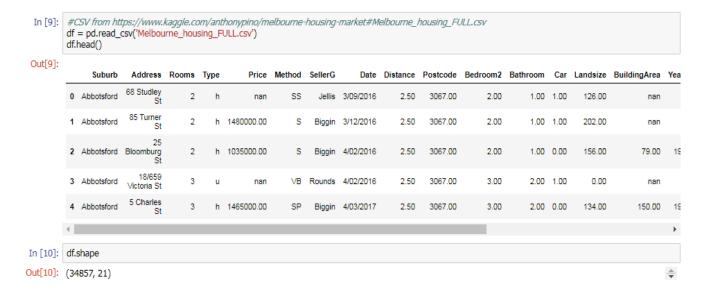
- Suburb (List of neighbourhoods)
- Price (Housing prices, to calculate average price for each neighbourhoods)
- Latitude (To visualize suburb location)
- Longitude (To visualize suburb location)
- Nearby Venues (To fetch using Foursquare API)

2.1 Suburb Data

Melbourne's housing price list can be downloaded from Kaggle's open datasets. This dataset save our time from scraping suburban data from other third party website such as Wikipedia. The dataset also contains latitude and longitude values for further analysis. The URL for the data is as shown below:

https://www.kaggle.com/anthonypino/melbourne-housing-market#Melbourne housing FULL.csv

The csv file includes total of 21 columns with sold price for each address in a suburb. There are total of 34,857 records in the data set.



2.2 Data Wrangling

There are null values in our price columns. In the case of less data set, we should fill those values by taking assumption such as getting average values, etc. However, since we have large dataset, we would simply drop the null values records. Also drop the columns that are not required for our purpose. Therefore, our data frame should look something like below.



2.3 Calculating mean (µ) value

We will calculate the mean value for each suburb using *pandas groupby()* method. Then re-create our data frame using python list as shown below.

