Data Report

The Battle of Neighborhoods  
Real Estate vs Surroundings

short line

This report is for the final course of the Data Science Specialization. A 9-courses series created by IBM, hosted on Coursera platform.

Shreena Parekh  
30th May, 2020

# Data Collection

**New York city neighborhoods were chosen as the observation target due to the following reasons:**

- The availability of real estate prices. Though very limited.

- The diversity of prices between neighborhoods. For example, a 2-bedrooms condo in Central Park West, Upper West Side can cost $4.91 million on average; while in Inwood, Upper Manhattan, just 30 minutes away, it's only $498 thousands.

1. **RealEstate Dataset**

URL:<https://www.cityrealty.com/nyc/market-insight/features/get-to-know/average-nyc-condo-prices-neighborhood-june-2018/18804>

This dataset contains the different Neighborhoods in Newyork with the average estate prices for 1BHK, 2BHK, 3BHK and commercial sites.

I have considered the prices of 2BHK as the average price.

1. **Geographical Data**

URL : <https://geo.nyu.edu/catalog/nyu_2451_34572>

The Geographical data is collected from the geo.nyu website for free. The dataset contains the Neighborhoods of NewYork in different Boroughs along with their geographical latitude and longitude coordinates

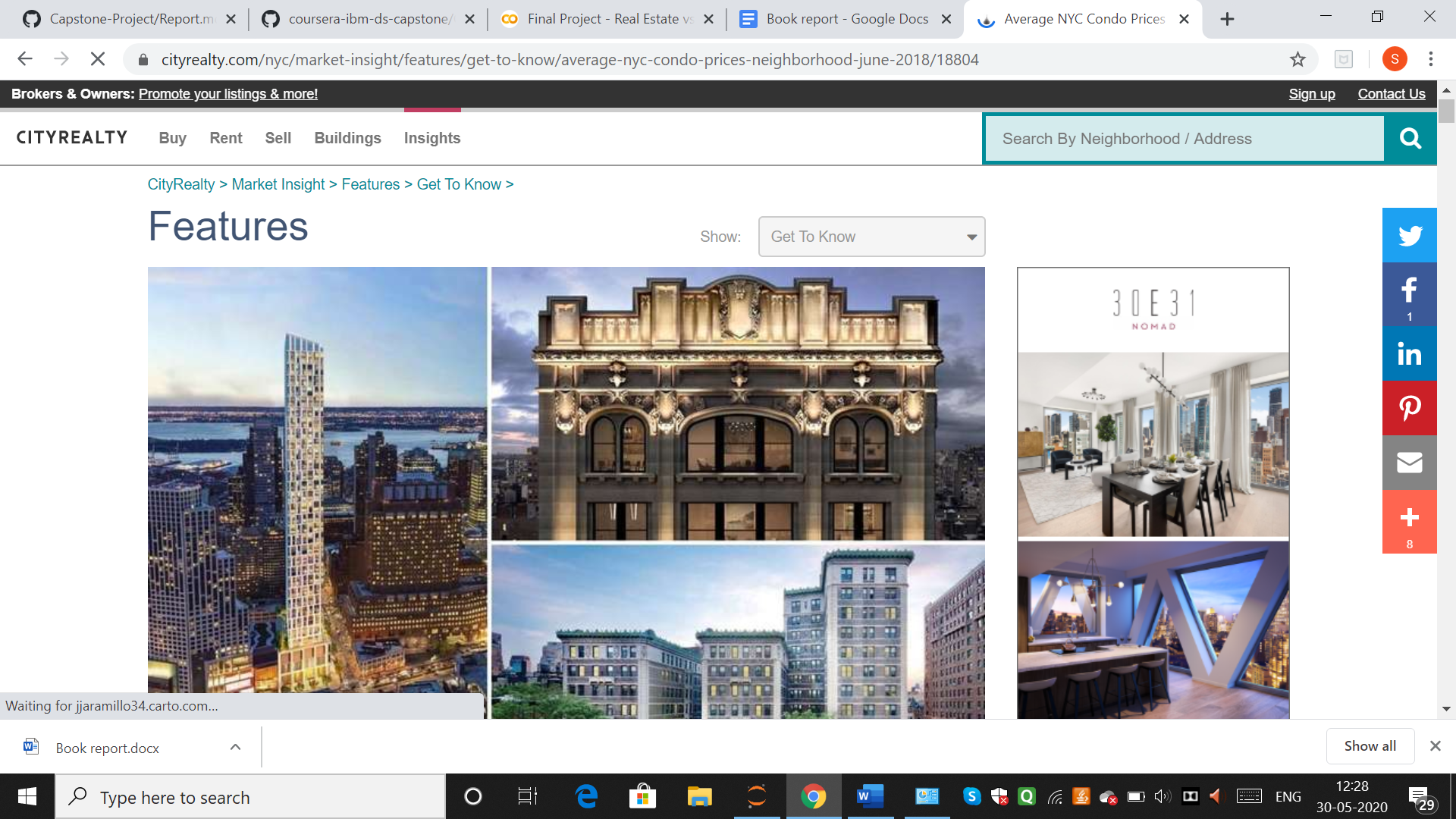
1. **Surrounding Venues**

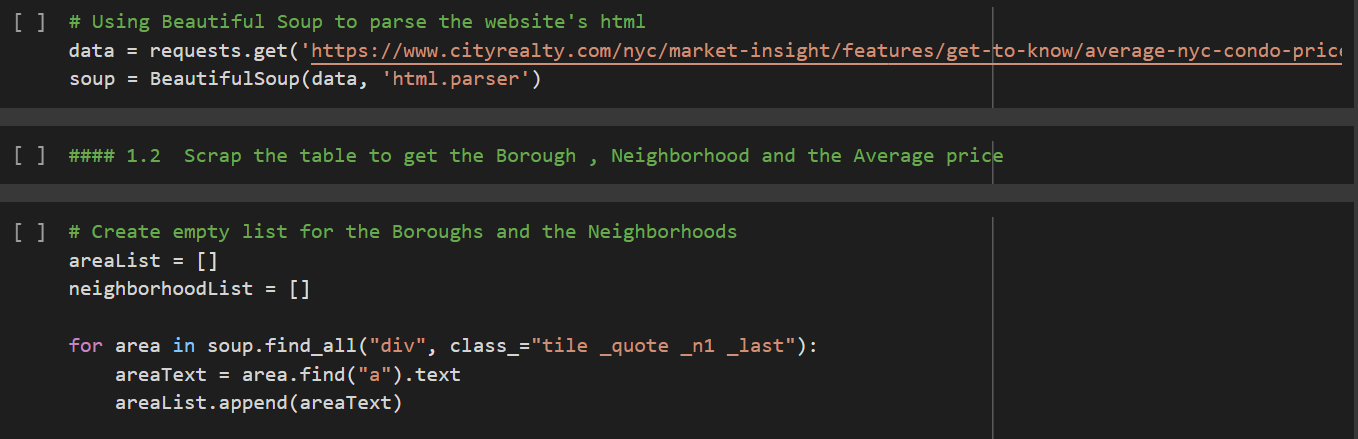
The dataset retrieved is a JSON file containing the details of venues within the range of specified radius from the Neighborhood coordinates. The venue details include the Venue name, category, geographical coordinates etc.

# 1. Real Estate Dataset

The real estate data set is collected using WebScraping using the Beautifulsoup Python module.

<https://www.cityrealty.com/nyc/market-insight/features/get-to-know/average-nyc-condo-prices-neighborhood-june-2018/18804>



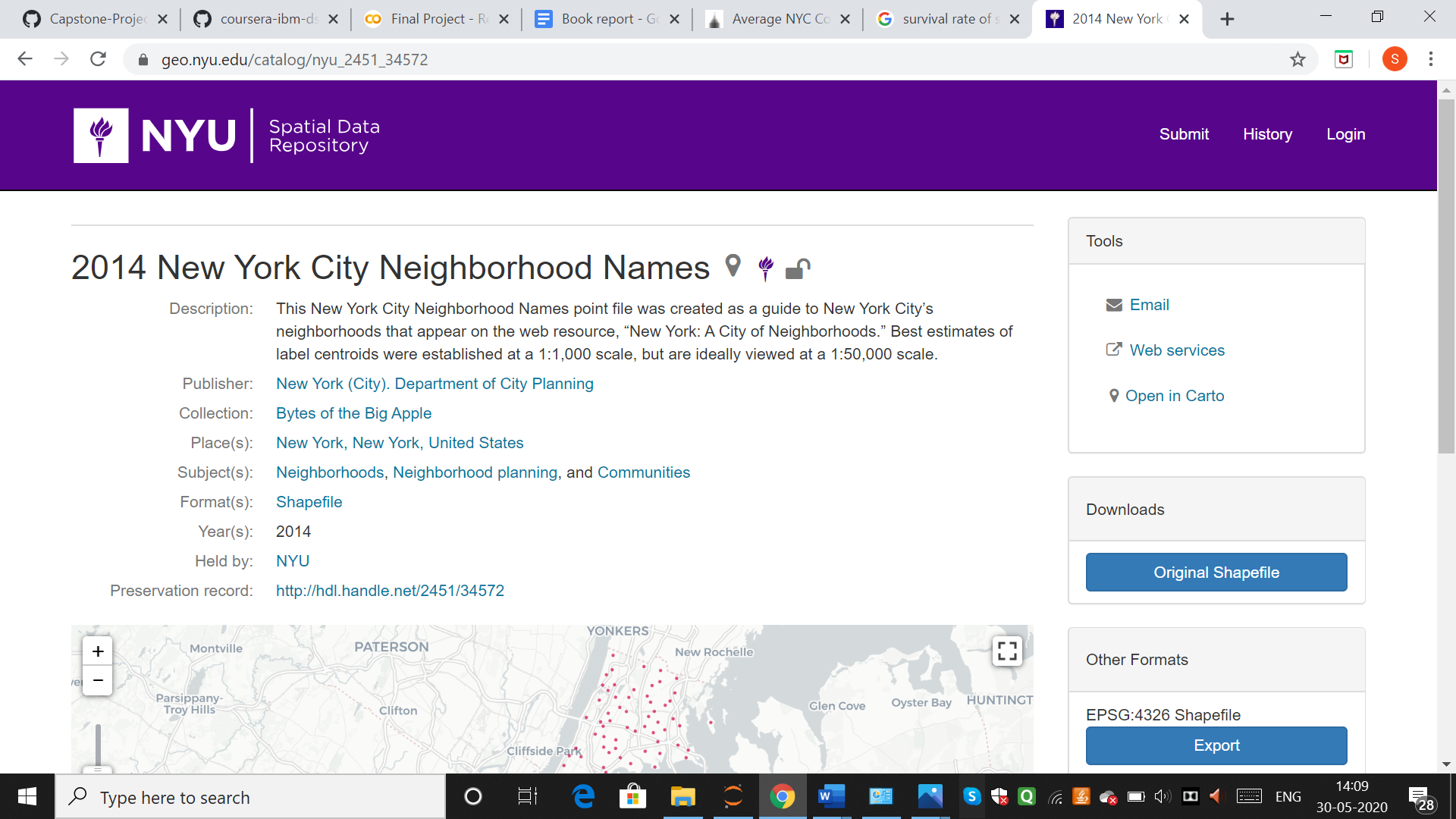
The RealEstate data is scrapped using BeautifulSoup :

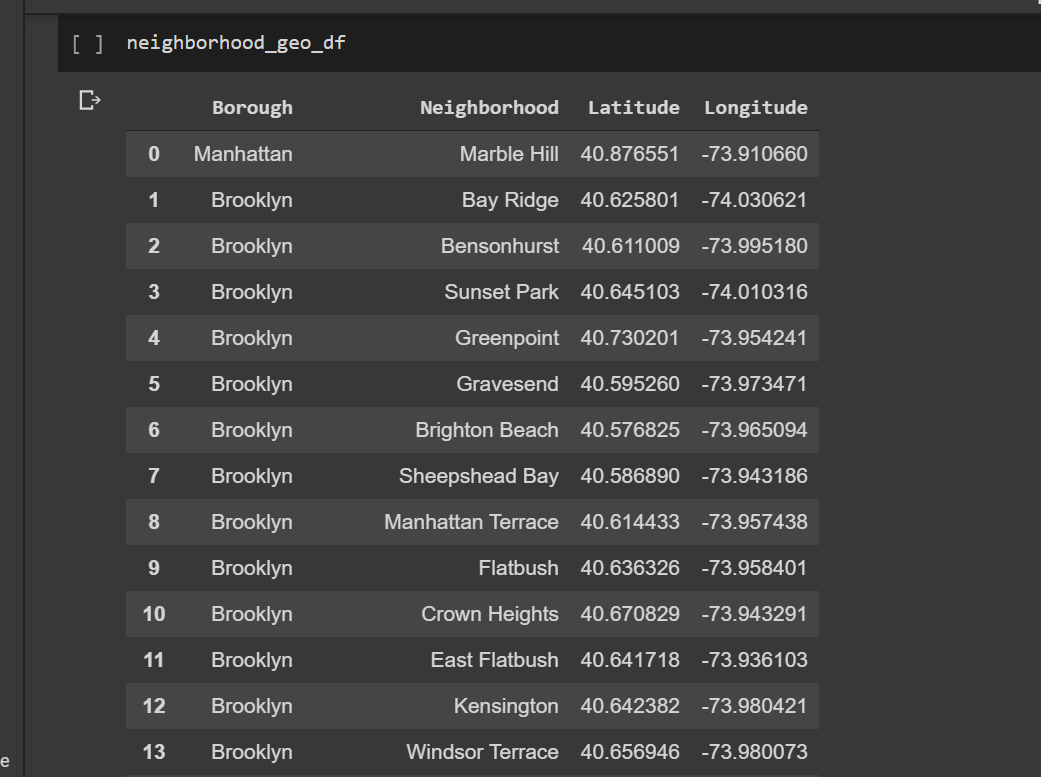
The final dataframe is saved as nyc\_neighborhood\_df :

# 2. Geographical Dataset

The Geographical dataset is collected from the geo.nyu website for free.

URL : <https://geo.nyu.edu/catalog/nyu_2451_34572>



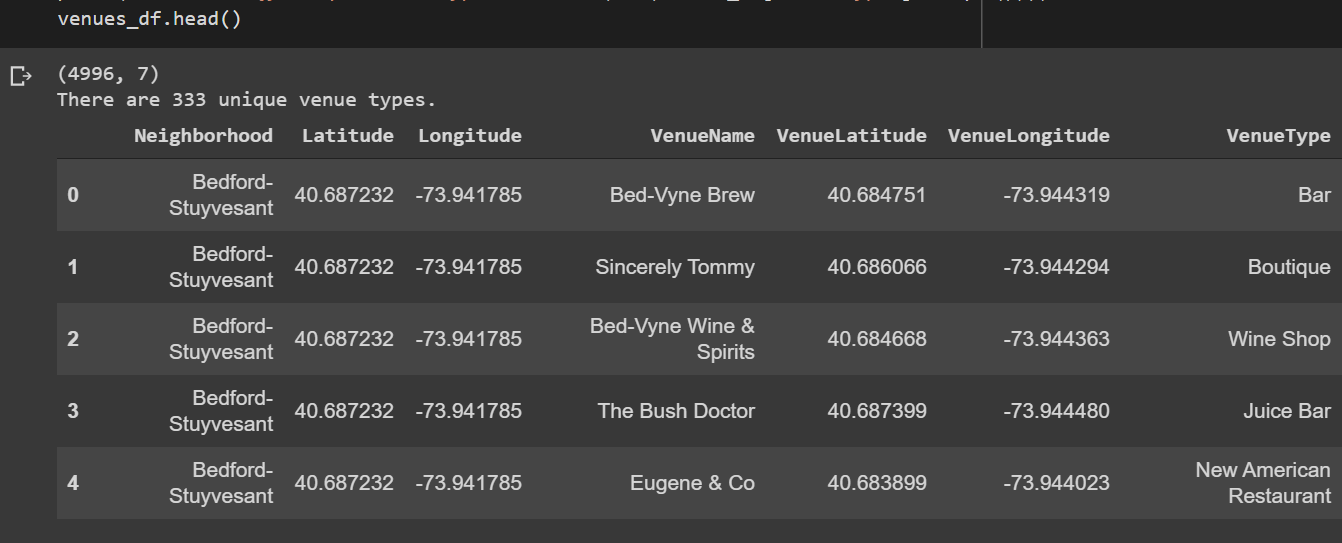
The final dataframe is saved as neighborhood\_geo\_df :

# 3. Surrounding Venues Data

The surrounding location data is retrieved using the FourSquare API call. This was done using a regular GET call and a URL using the explore endpoint. With a private FourSquare Developer account, you can make 99000 regular calls per day.

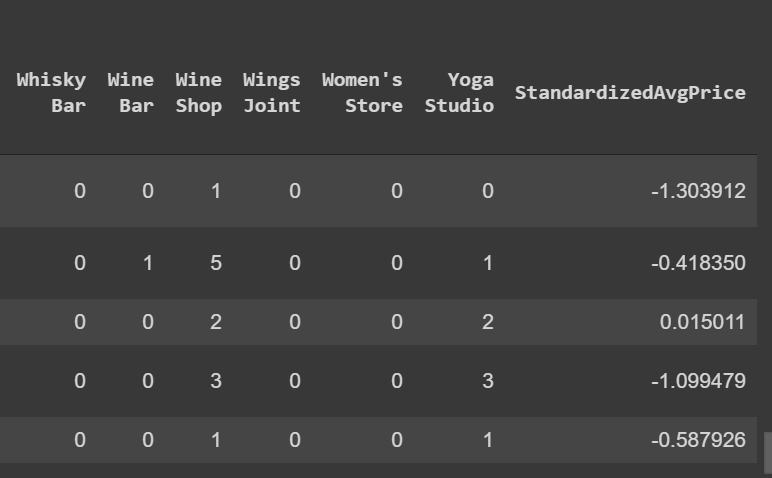
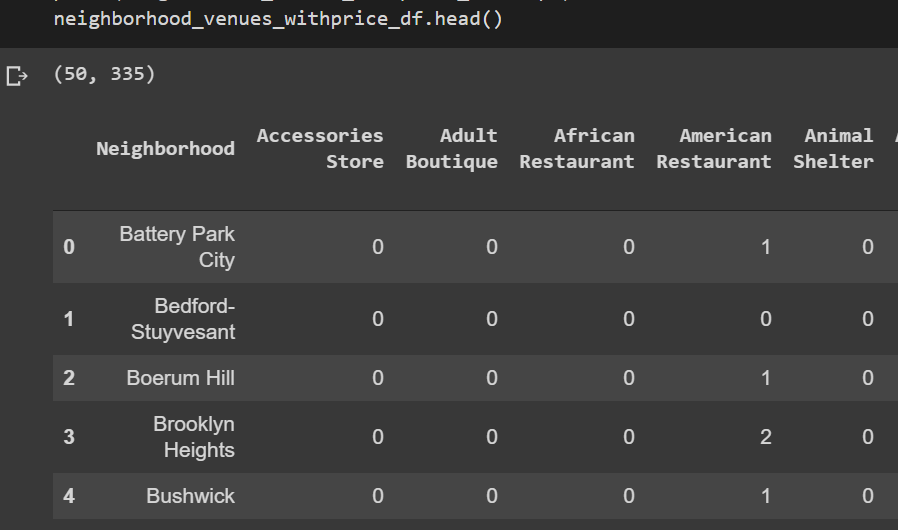


The data is converted into the venues\_df dataset :



# Data Preprocessing

* Merge the RealEstate dataset and the Geographical data into a single dataframe
  + Modify the Neighborhoods names according to their name in the RealEstate dataset
  + Add latitudes and longitudes of missing rows
  + Merge makeups into single Neighborhood
  + Change old names to new ones
  + Convert the names in both the dataframes into same case
* Find the geographic data( polygon type coordinates ) of the neighborhoods. Both their center coordinates and their border.
* Plot Folium and Choropleth Maps
* For each neighborhood, pass the obtained coordinates to FourSquare API. The “explore” endpoint will return a list of surrounding venues in a pre-defined radius.
* Count the occurrence of each venue type in a neighborhood. Then apply one hot encoding to turn each venue type into a column with their occurrence as the value.
* Standardize the average price by removing the mean and scaling to unit variance.

The final dataframe neighborhood\_venues\_withprice\_df looks as :

# Data Visualization

In order to have a first insight of New York city real estate average price between neighborhoods, there is no better way than visualization.

The medium chosen is Choropleth map, which uses differences in shading or coloring to indicate a property’s values or quantity within predefined areas. It is ideal for showing how differently real estate priced between neighborhoods across the New York city map.

# 

The map shows high prices in neighborhoods that are located around Central Park, Midtown and Lower Manhattan. The price reduces further toward North Manhattan or toward Brooklyn.

Manhattan can be considered the heart of New York city. It’s where most businesses, tourist attractions and entertainments are located. So, the venue types that can attract many people are expected to have the most positive coefficients in the regression model.