Project Report

The Battle of Neighborhoods

Real Estate vs Surroundings

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

Project Workflow

1. **Problem Statement**
2. **Data Collection**
    a. Get the Real Estate data - **WebScrapping**
        - Website :
        - Module : **BeautifulSoup**
        - Data file :
    b. Get the Geographic Coordinates
        - Website :
    c. Get the polygon coordinates for Choropleth Map
        - Website :
        - Data File :
    d. Get the surrounding venues - **Foursquare API**
        - URL Group :
        - URL Endpoint :
3. **Data Processing**
    a. Merging the datasets
    b. One Hot encoding to convert Categorical variables to dummies
    c. Feature Scaling
    d. Grouping
4. **Data Modeling** (**PCR - Principal Component Regression**)
    a. PCA : Obtaining the principal components (extracted features)
    b. Linear Regression : Linear Regression Model on extracted features
5. **Data Evaluation**
    a. R2 score
    b. MSE - Mean Squared Error
6. **Conclusion**

Problem Statement

The main goal of this project is to explore the neighborhoods of New York city in order to extract the correlation between the real estate value and its surrounding venues.

The idea comes from the process of a normal family finding a place to stay after moving to another city. It’s common that the owners or agents advertise their properties are closed to some kinds of venues like supermarkets, restaurants or coffee shops, etc.; showing the “convenience” of the location in order to raise their house’s value.

So,the project tries to find out can the surrounding venus affect the price of a house? If so, what types of venues have the most affect, both positively and negatively?

The target audience for this report are:

- Potential buyers who can roughly estimate the value of a house based on the surrounding venues and the average price.
- Real estate makers and planners who can decide what kind of venues to put around their products to maximize selling price.
- Houses sellers who can optimize their advertisements.
- Course's instructors and learners who will grade this project or to anyone who catches this project on GitHub showing how to implement DataScience Python tools for real world problems

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.

2. **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

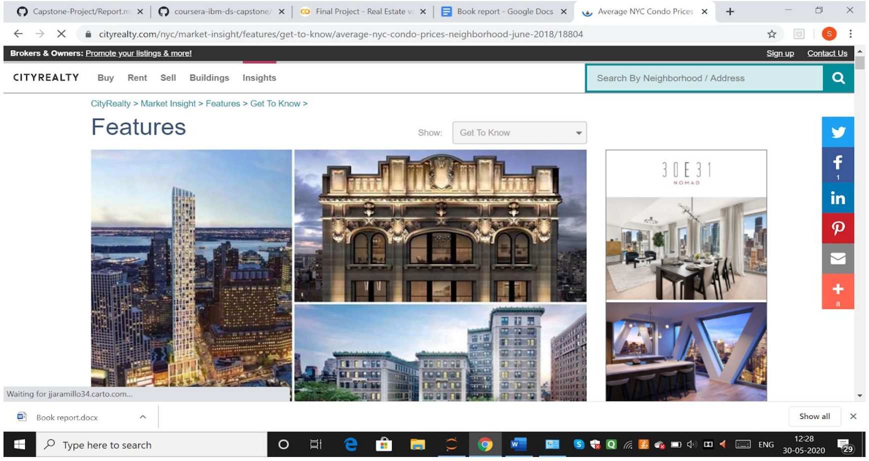
3. **Surrounding Venues**

   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.

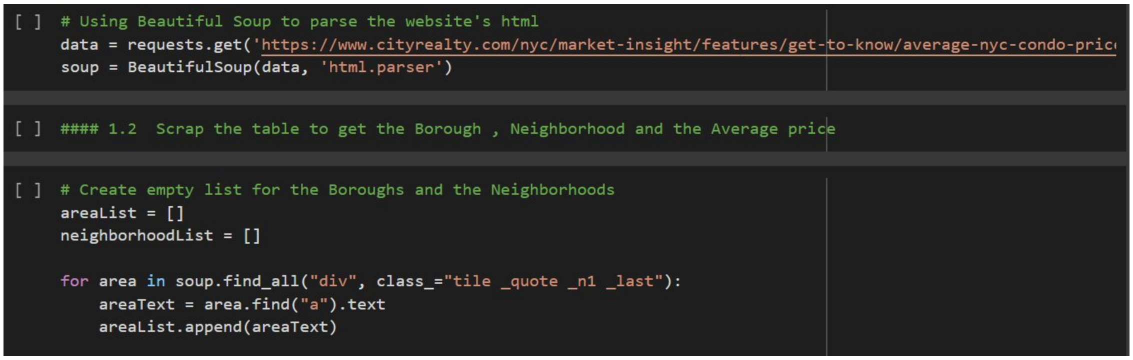
1. Real Estate Dataset

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


Capstone-Project/Report.m
coursera-ibm-ds-capstone
Final Project - Real Estate
Book report - Google Docs
Average NYC Condo Prices
cityrealty.com/nyc/market-insight/features/get-to-know/average-nyc-condo-prices-neighborhood-june-2018/18804
Brokers & Owners: Promote your listings & more!
Sign up
Contact Us
CITYREALTY
Buy
Rent
Sell
Buildings
Insights
Search By Neighborhood / Address
CityRealty > Market Insight > Features > Get To Know >
Features
Show:
Get To Know
30E31
NOMAD
Waiting for jjaramillo34.carto.com...
Book report.docx
Show all
Type here to search
ENG
12:28
30-05-2020

The RealEstate data is scrapped using BeautifulSoup :



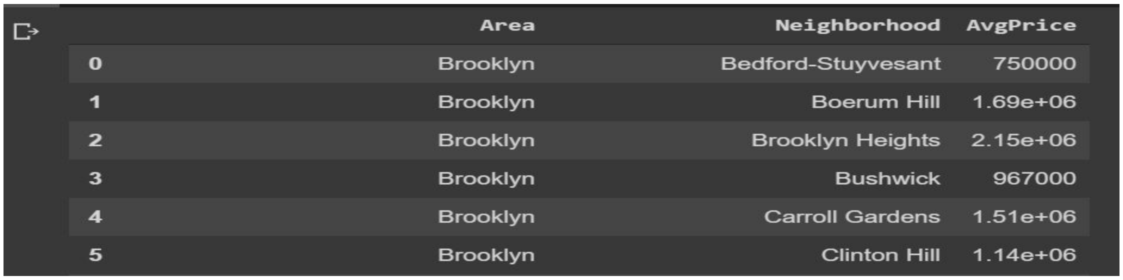
```
[ ] # Using Beautiful Soup to parse the website's html
    data = requests.get('https://www.cityrealty.com/nyc/market-insight/features/get-to-know/average-nyc-condo-pric
    soup = BeautifulSoup(data, 'html.parser')

[ ] #### 1.2  Scrap the table to get the Borough , Neighborhood and the Average price

[ ] # Create empty list for the Boroughs and the Neighborhoods
    areaList = []
    neighborhoodList = []

    for area in soup.find_all("div", class_="tile _quote _n1 _last"):
        areaText = area.find("a").text
        areaList.append(areaText)
```

The final dataframe is saved as nyc_neighborhood_df :

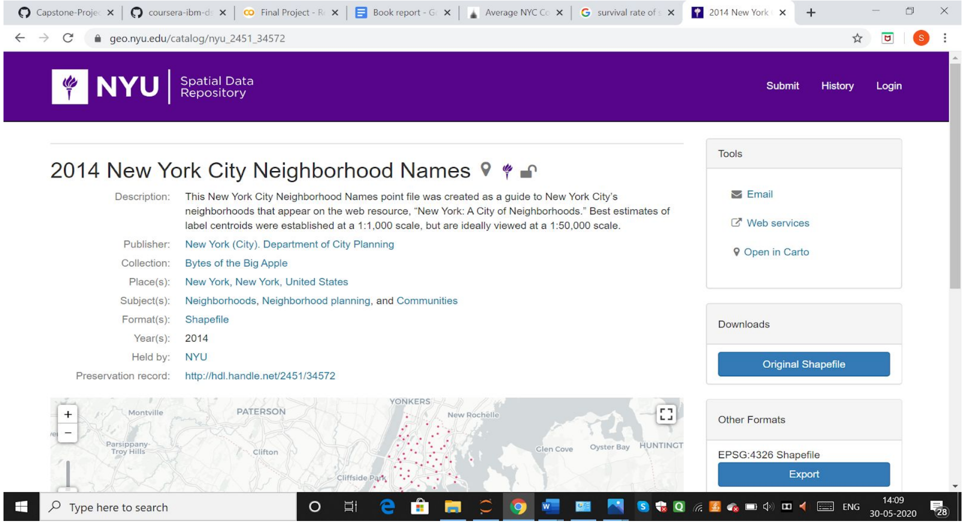


	Area	Neighborhood	AvgPrice
0	Brooklyn	Bedford-Stuyvesant	750000
1	Brooklyn	Boerum Hill	1.69e+06
2	Brooklyn	Brooklyn Heights	2.15e+06
3	Brooklyn	Bushwick	967000
4	Brooklyn	Carroll Gardens	1.51e+06
5	Brooklyn	Clinton Hill	1.14e+06

2. Geographical Dataset

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

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



geo.nyu.edu/catalog/nyu_2451_34572

2014 New York City Neighborhood Names

Description: This New York City Neighborhood Names point file was created as a guide to New York City's neighborhoods that appear on the web resource, "New York: A City of Neighborhoods." Best estimates of label centroids were established at a 1:1,000 scale, but are ideally viewed at a 1:50,000 scale.

Publisher: New York (City). Department of City Planning

Collection: Bytes of the Big Apple

Place(s): New York, New York, United States

Subject(s): Neighborhoods, Neighborhood planning, and Communities

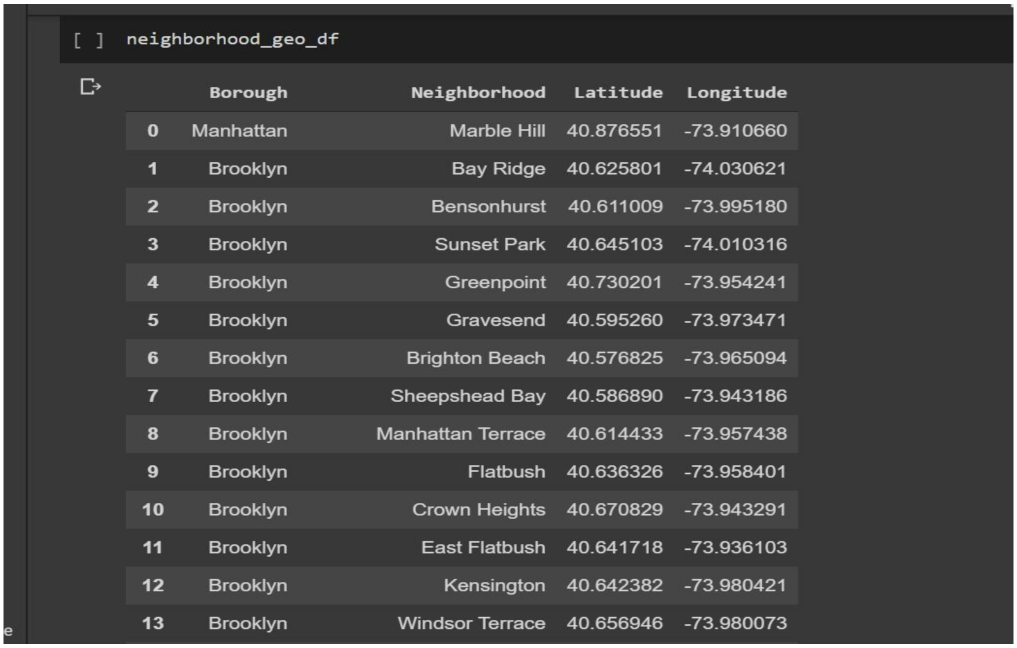
Format(s): Shapefile

Year(s): 2014

Held by: NYU

Preservation record: http://hdl.handle.net/2451/34572

The final dataframe is saved as neighborhood_geo_df :

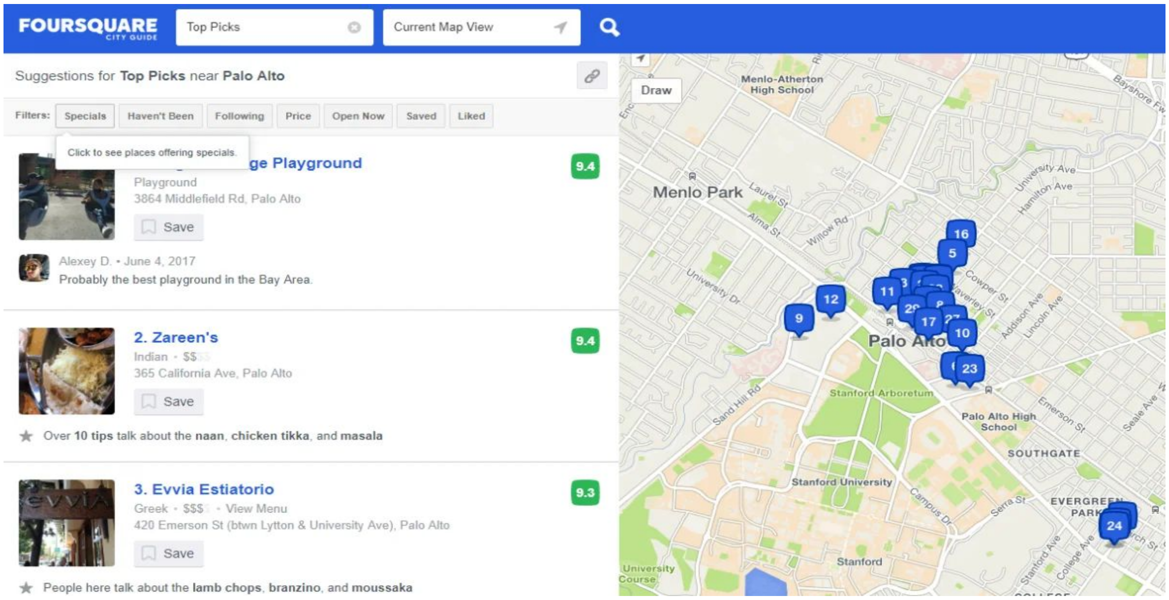


```
[ ]  neighborhood_geo_df
```

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Brooklyn	Bay Ridge	40.625801	-74.030621
2	Brooklyn	Bensonhurst	40.611009	-73.995180
3	Brooklyn	Sunset Park	40.645103	-74.010316
4	Brooklyn	Greenpoint	40.730201	-73.954241
5	Brooklyn	Gravesend	40.595260	-73.973471
6	Brooklyn	Brighton Beach	40.576825	-73.965094
7	Brooklyn	Sheepshead Bay	40.586890	-73.943186
8	Brooklyn	Manhattan Terrace	40.614433	-73.957438
9	Brooklyn	Flatbush	40.636326	-73.958401
10	Brooklyn	Crown Heights	40.670829	-73.943291
11	Brooklyn	East Flatbush	40.641718	-73.936103
12	Brooklyn	Kensington	40.642382	-73.980421
13	Brooklyn	Windsor Terrace	40.656946	-73.980073

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.



FOURSQUARE CITY GUIDE | Top Picks | Current Map View

Suggestions for **Top Picks** near **Palo Alto**

Filters: Specials | Haven't Been | Following | Price | Open Now | Saved | Liked

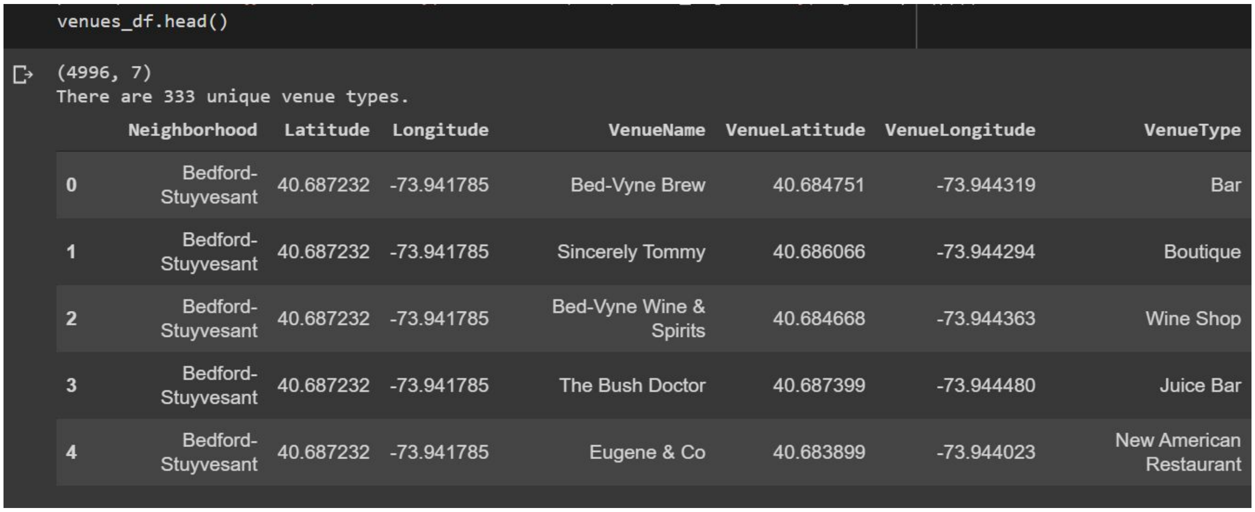
Click to see places offering specials.

ge Playground — 9.4
Playground
3864 Middlefield Rd, Palo Alto
Save
Alexey D. • June 4, 2017
Probably the best playground in the Bay Area.

2. Zareen's — 9.4
Indian • $$
365 California Ave, Palo Alto
Save
★ Over **10 tips** talk about the **naan**, **chicken tikka**, and **masala**

3. Evvia Estiatorio — 9.3
Greek • $$$ • View Menu
420 Emerson St (btwn Lytton & University Ave), Palo Alto
Save
★ People here talk about the **lamb chops**, **branzino**, and **moussaka**

The data is converted into the venues_df dataset :



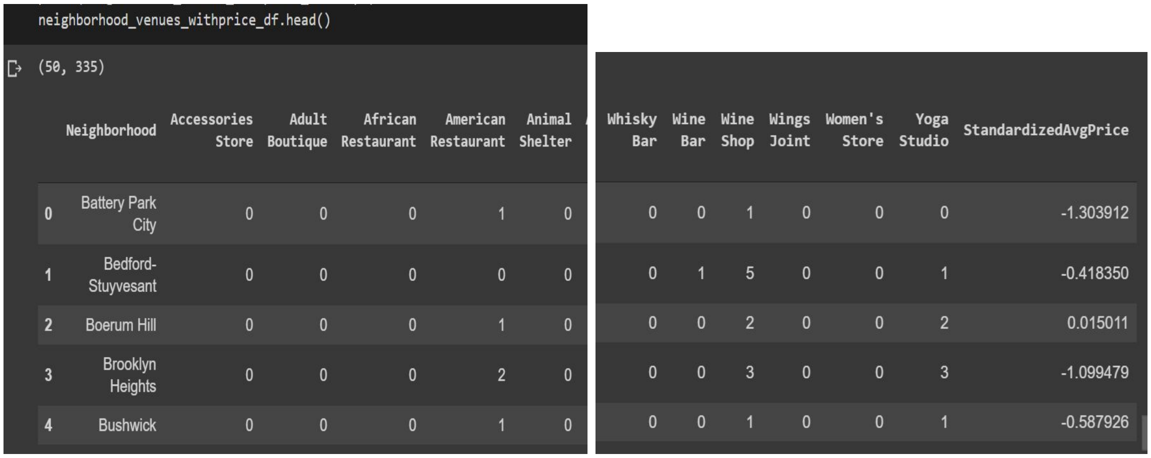
```
venues_df.head()

(4996, 7)
There are 333 unique venue types.
       Neighborhood   Latitude   Longitude          VenueName  VenueLatitude  VenueLongitude        VenueType
0  Bedford-           40.687232  -73.941785      Bed-Vyne Brew      40.684751      -73.944319              Bar
   Stuyvesant
1  Bedford-           40.687232  -73.941785   Sincerely Tommy      40.686066      -73.944294         Boutique
   Stuyvesant
2  Bedford-           40.687232  -73.941785  Bed-Vyne Wine &       40.684668      -73.944363        Wine Shop
   Stuyvesant                                          Spirits
3  Bedford-           40.687232  -73.941785   The Bush Doctor      40.687399      -73.944480        Juice Bar
   Stuyvesant
4  Bedford-           40.687232  -73.941785       Eugene & Co      40.683899      -73.944023     New American
   Stuyvesant                                                                                     Restaurant
```

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 :



```
neighborhood_venues_withprice_df.head()
```

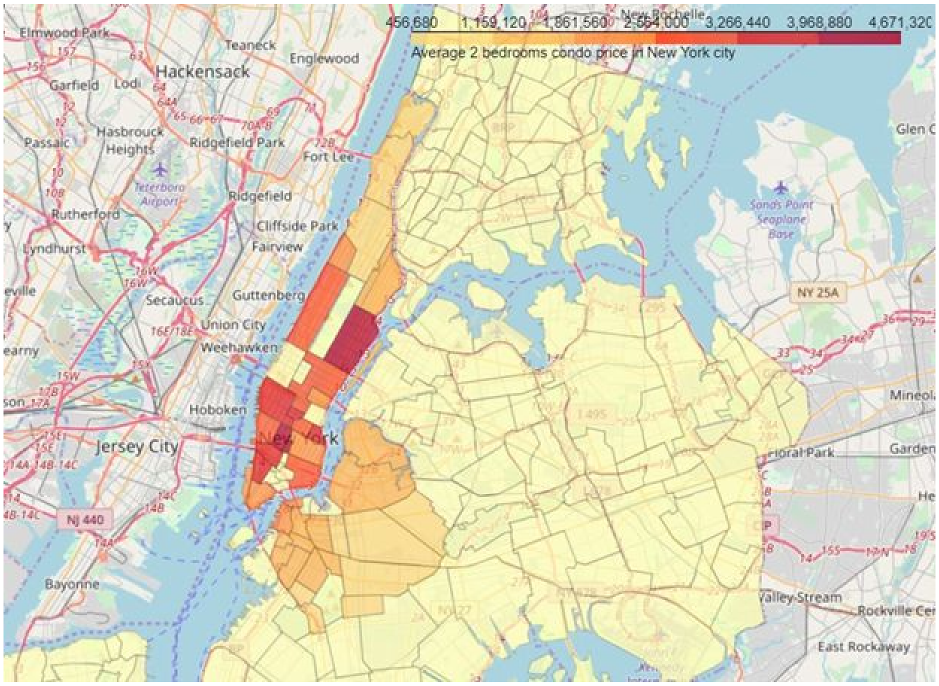
```
(50, 335)
```

	Neighborhood	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Animal Shelter	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	StandardizedAvgPrice
0	Battery Park City	0	0	0	1	0	0	0	1	0	0	0	-1.303912
1	Bedford-Stuyvesant	0	0	0	0	0	0	1	5	0	0	1	-0.418350
2	Boerum Hill	0	0	0	1	0	0	0	2	0	0	2	0.015011
3	Brooklyn Heights	0	0	0	2	0	0	0	3	0	0	3	-1.099479
4	Bushwick	0	0	0	1	0	0	0	1	0	0	1	-0.587926

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.

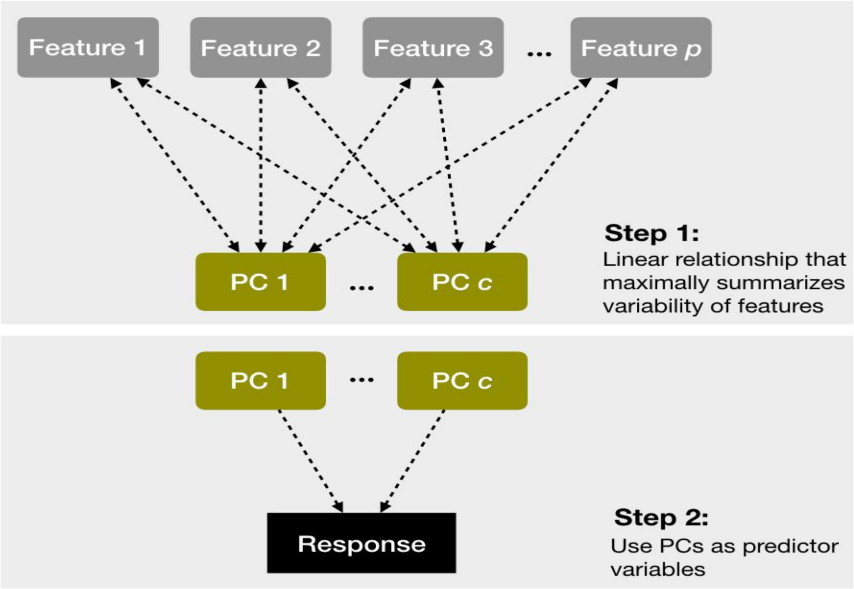
Data Modeling

First, I used a Linear Regression model to fit the data. But the results were not satisfying. The reason behind this could be that the number of features in the dataset were much more greater than the number of samples.

Number of features : 335

Number of samples : 50

So, to avoid this problem and to improvise the model I used PCR(Principal Component Regression):


Feature 1
Feature 2
Feature 3
...
Feature *p*
PC 1
...
PC *c*
Step 1:
Linear relationship that maximally summarizes variability of features
PC 1
...
PC *c*
Response
Step 2:
Use PCs as predictor variables

1. **PCA** : Perform PCA on the features set to obtain the principal components. Then select a subset for the next step.
2. **Regression** : Use regression on the previous subset of principal components to get a list of coefficient correlations.

Principal Component Analysis

PCR can be explained simply as the combination of Principal Component Analysis (PCA) with Linear Regression.

PCR employs the power of PCA, which can convert a set of values of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

As a result, the number of features is reduced while keeping most of the characteristics of the dataset.

Then PCR uses Linear Regression on the converted set to return a coefficient list, just like in normal Regression techniques.

Again, R2 score and MSE are used to see how well the model fit the dataset.

The result is promising as it shows improvement over the simple Linear Regression.

As for the coefficient list, the size has been reduced after performing PCA. So, a dot product with eigenvectors is needed to get it back to the original features size.

The insight is still consistent compared to the Linear Regression's.

Results and Facts

Results:

Even though the scores seem to be improved after applying a more sophisticated method, the model is still not suitable for the dataset. Thus, it can't be used to precisely predict a neighborhood average price.

Explanations for the poor model can be:

- The real estate price is hard to predict.
- The data is incomplete (small sample size, missing deciding factors).
- The machine learning techniques are chosen or applied poorly.

But again, on the bright side, the insight, gotten from observing the analysis results, seems consistent and logical. And the insight is business venues that can serve the needs of most normal people usually situated in pricey neighborhoods.

Facts :

The real challenge is constructing the dataset:

- Usually the needed data isn't publicly available.
- When combining data from multiple sources, inconsistencies can happen. And lots of efforts are required to check, research and change the data before merging.
- For data obtained through API calls, different results are returned with different sets of parameters and different points of time. Multiple trial and error runs are required to get the optimal result.

Conclusions

It’s unfortunately that the analysis couldn’t produce a precise model or showing any strong coefficient correlation for any venue type. But we can still get some meaningful and logical insights from the result.

Based on the observed coefficient correlations, we can say that :

1. Fancy places like restaurants seem to boost real estate's value the most.
2. Neighborhoods that have many restaurants are most likely business areas such as downtown. It's where lots of people go to, lots of activities to enjoy, lots of other businesses.
3. Market and Clothing stores also tend to impact the real estate prices by attracting the crowd.
4. Coffee shops and Lighthouses to a little affect the price as people tend to move in for refreshment.
5. Finally as the result is poor and the correlations are weak, we can say that the real estate prices in NewYork are less according to its surroundings.

Doing this project helps practicing every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.

Some notes on the analysis result:

- The coefficients only show correlation, not causation. So, if your neighborhood average price is low, please don’t go destroying the surrounding bars and food trucks. There might be another reason.
- Toward the person that went through this project, many thanks for the time and patient.