**IBM Data Science Professional Certificate**

**Capstone - The Battle of Neighborhoods (Week 2)**

**REPORT**

**INTRODUCTION**

This is a capstone project for IBM Data Science Professional Certificate. In this project, we will process a hypothetical scenario for opening a new bookstore in Brooklyn/New York City. Bookstores always have an important role for people, especially for educated people and students in large cities like NYC. There are many bookstores in NYC and serious considerations are required to open a new bookstore than it seems. Finding the best location to open a bookstore is one of the most important decisions for investors, and we will complete this project to help stakeholders find the most suitable location.

**BUSINESS PROBLEM**

The objective of this project is to find the most suitable location for the investor to open a new bookstore in Brooklyn/NYC. By using data analysis and tools, and machine learning algorithms such as clustering, this project aims to provide solutions to answer the question; “***If an entrepreneur wants to open a new bookstore, what would be the best location for it in Brooklyn?***”

**TARGET AUDIENCE**

The investor who wants to open a new bookstore in Brooklyn/NYC.

**DATA**

To find the best solution, we will need the following data;

* List of neighborhoods in New York City.
* Latitude and longitude coordinates of these neighborhoods.
* Venue data related to bookstores. This will help us find the neighborhoods which are more suitable to open a bookstore.

**EXTRACTING THE DATA**

* Scrapping data of NYC neighborhoods from <https://cocl.us/new_york_dataset>
* Getting latitude and longitude data of NYC neighborhoods via Geocoder library
* Using Foursquare API to get venue data related to NYC neighborhoods

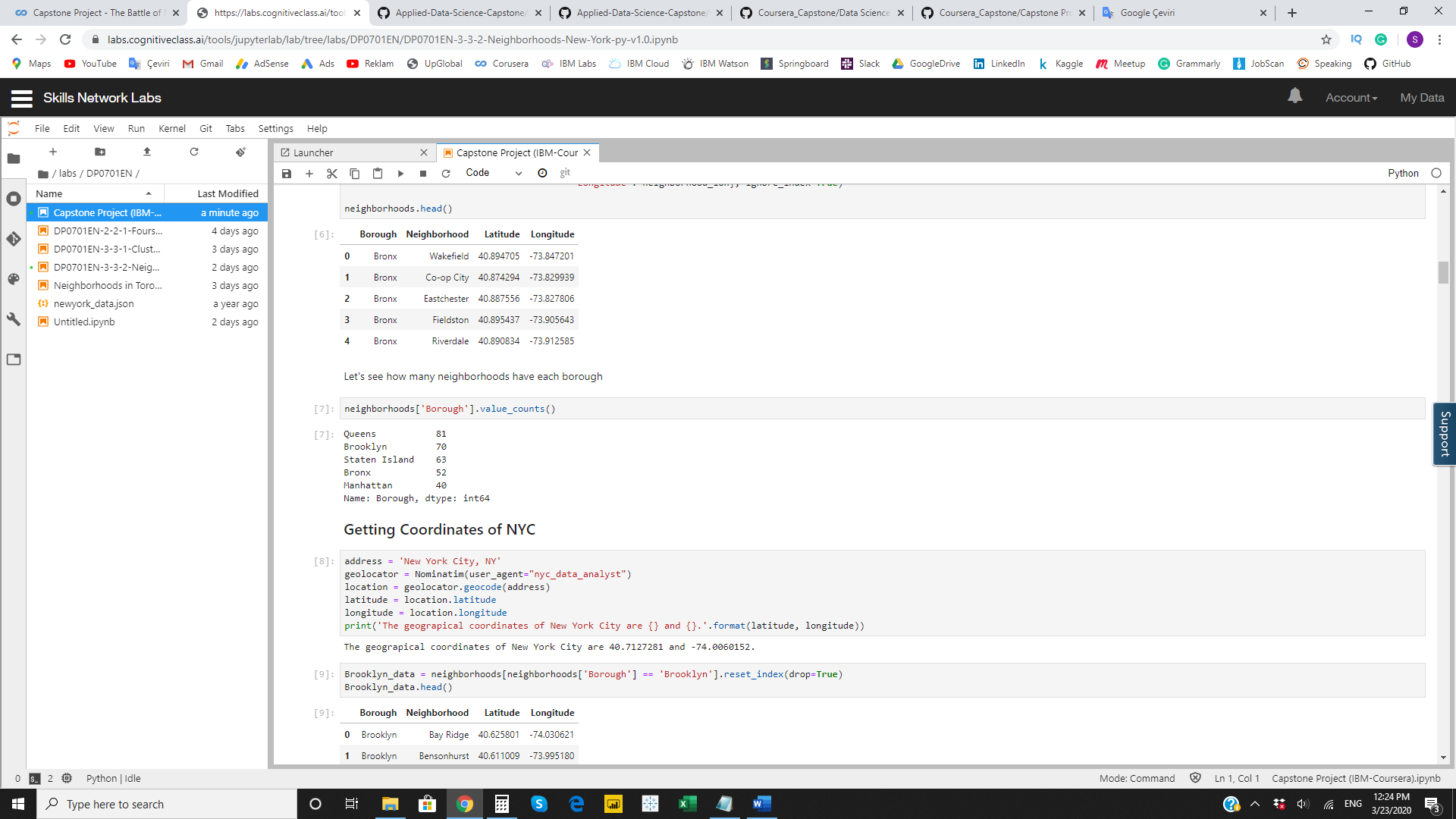
# METHODOLOGY

This project applies various data analysis techniques like;

* Get the data from open sources
* Working with Foursquare API
* Data cleaning
* Data wrangling
* Machine Learning (K-means Clustering)
* Map Visualization

We use the data of NYC neighborhoods from <https://cocl.us/new_york_dataset>

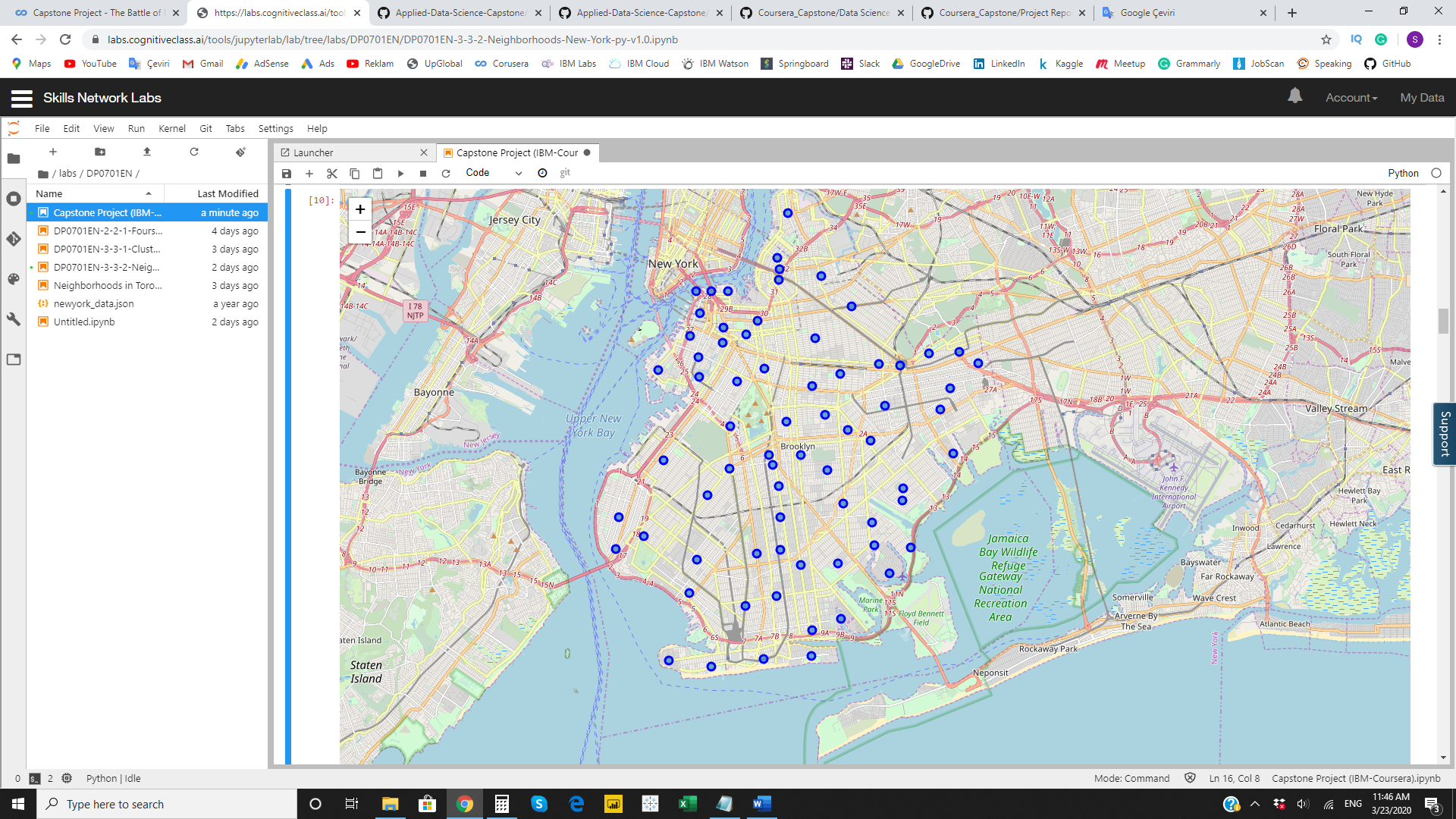
This source has 5 boroughs and 306 neighborhoods of New York City as well as the latitude and longitude data of each neighborhood.



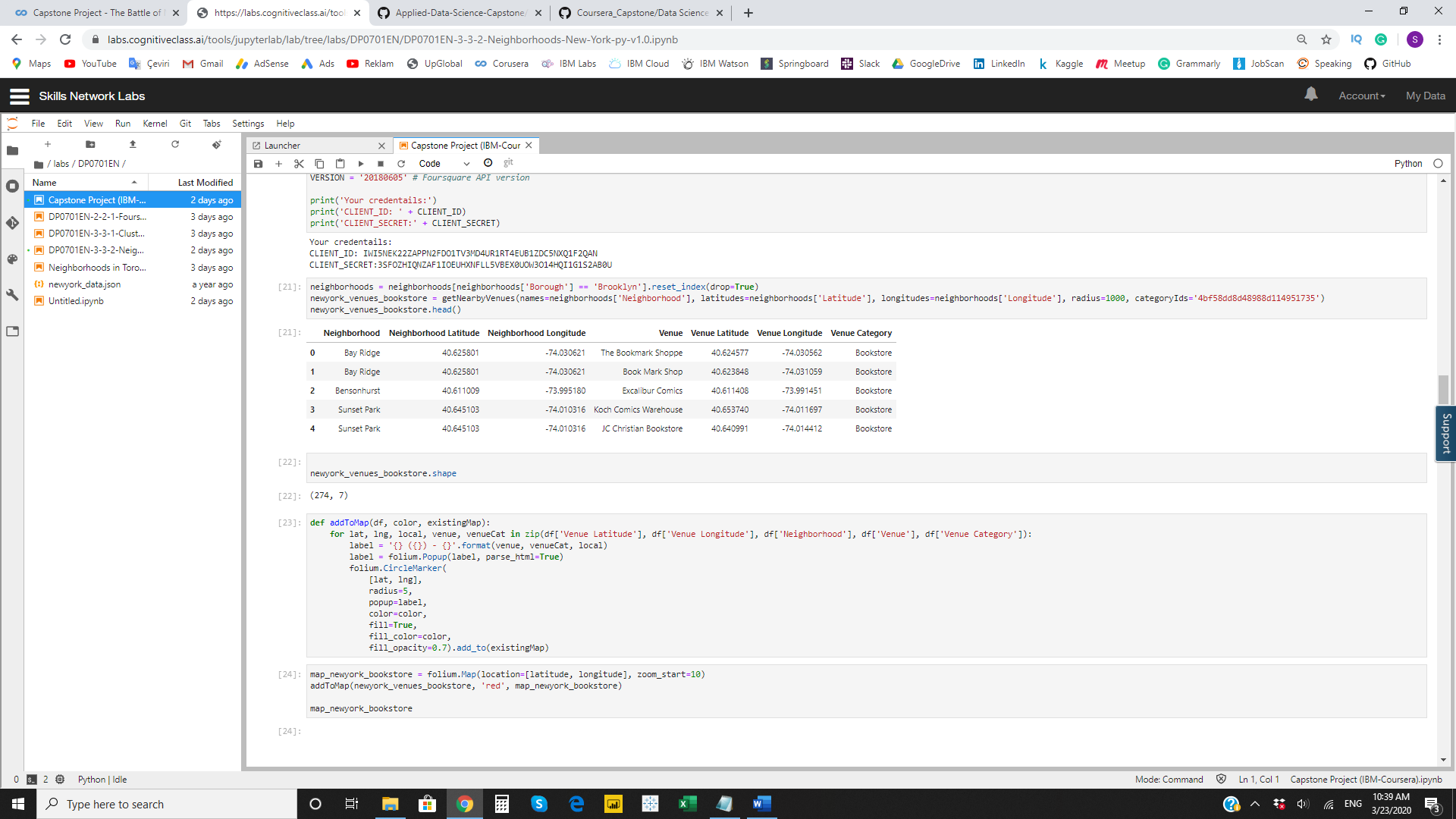
In this project, we focus on the Brooklyn area, and we can see the Brooklyn area has 70 neighborhoods.

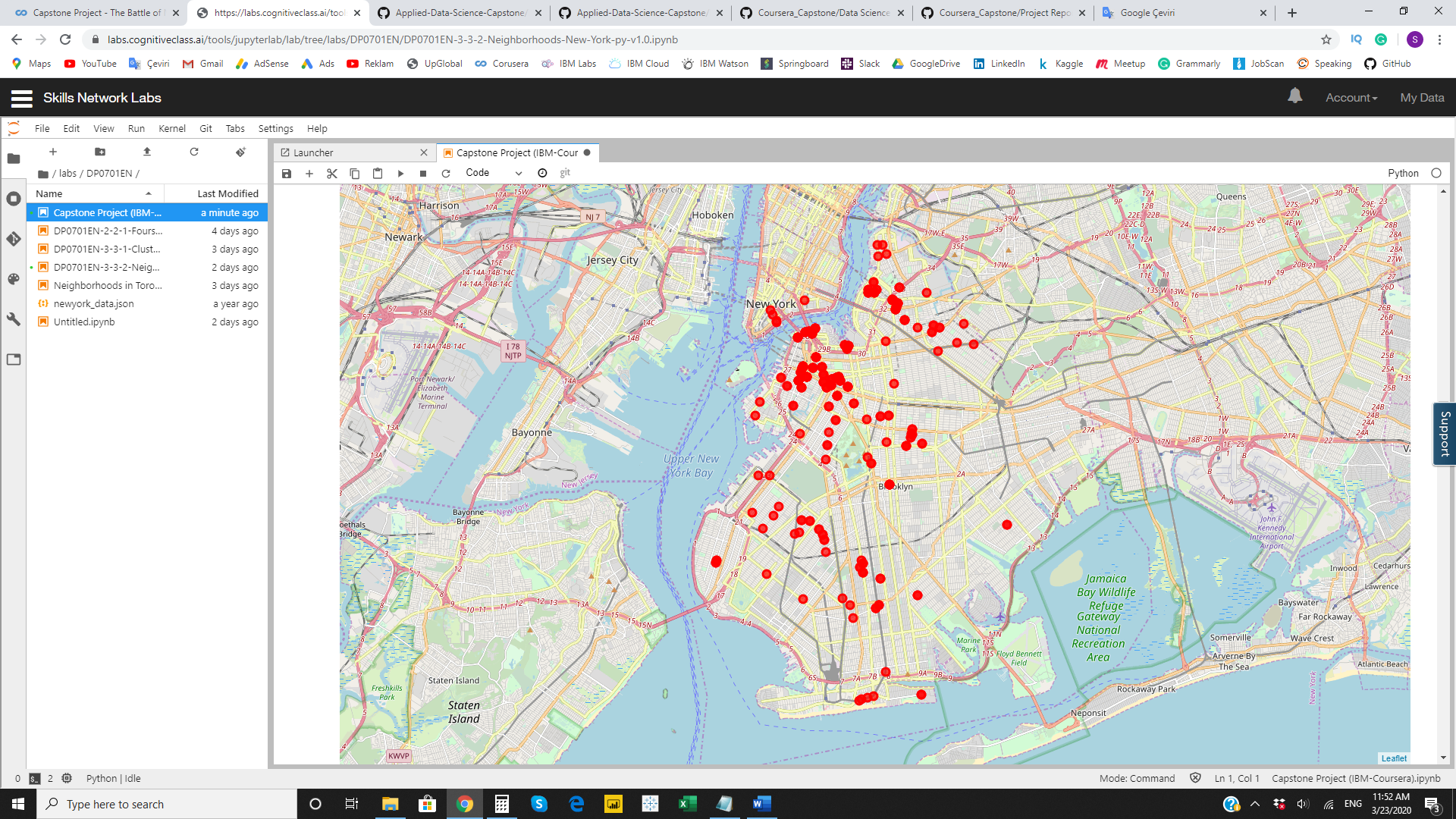
We get the geographical coordinates of these neighborhoods by using Geocoder Python package. This map performs a simple validation of correctness of our data.

In this map, we can see the neighborhoods in Brooklyn;

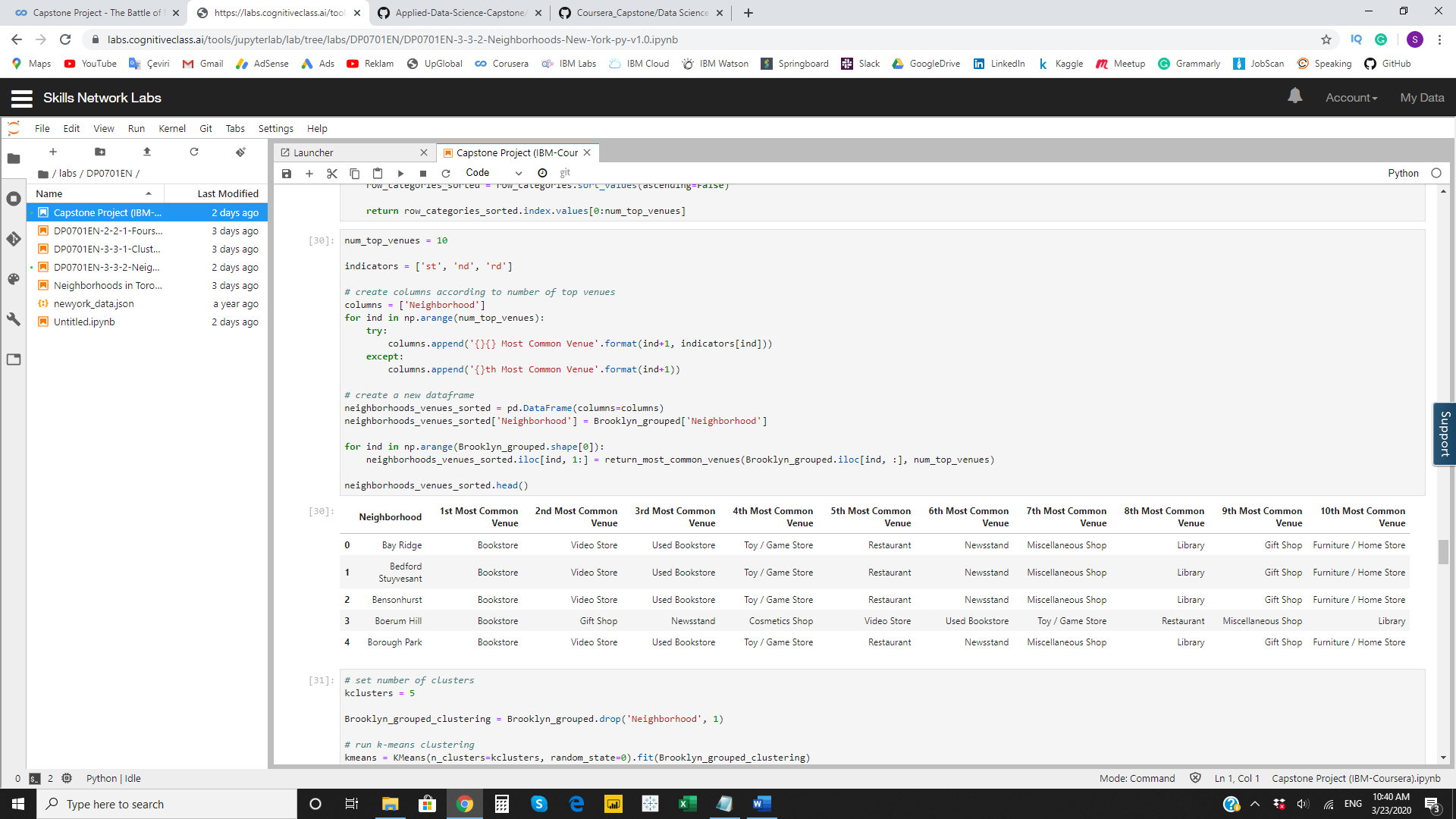


We use Foursquare API to the venues (bookstores) of these neighborhoods in Brooklyn.



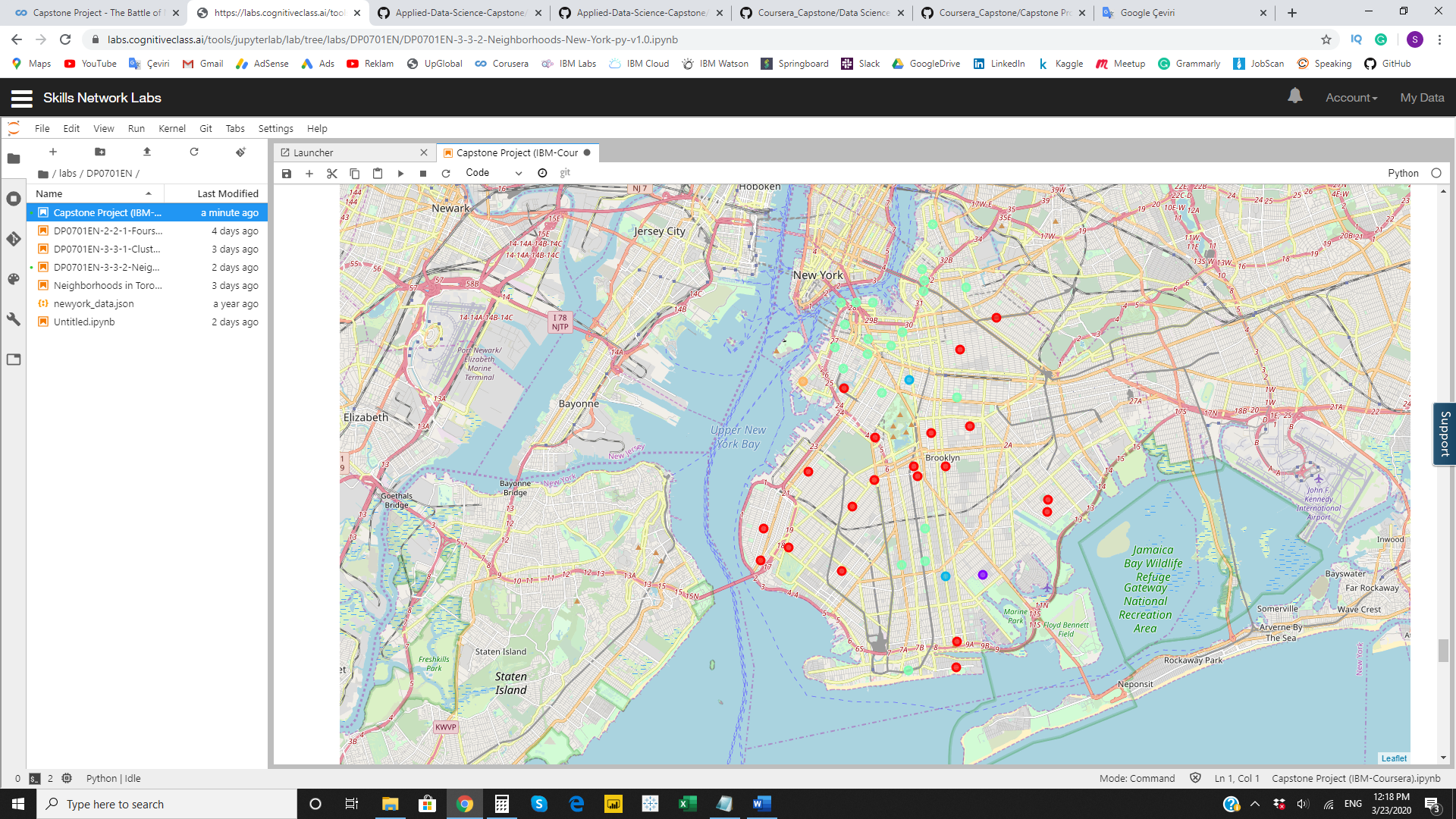


We use this feature to group the neighborhoods into K-means clustering algorithm. And also, the Folium library to visualize the neighborhoods in Brooklyn and its emerging clusters.

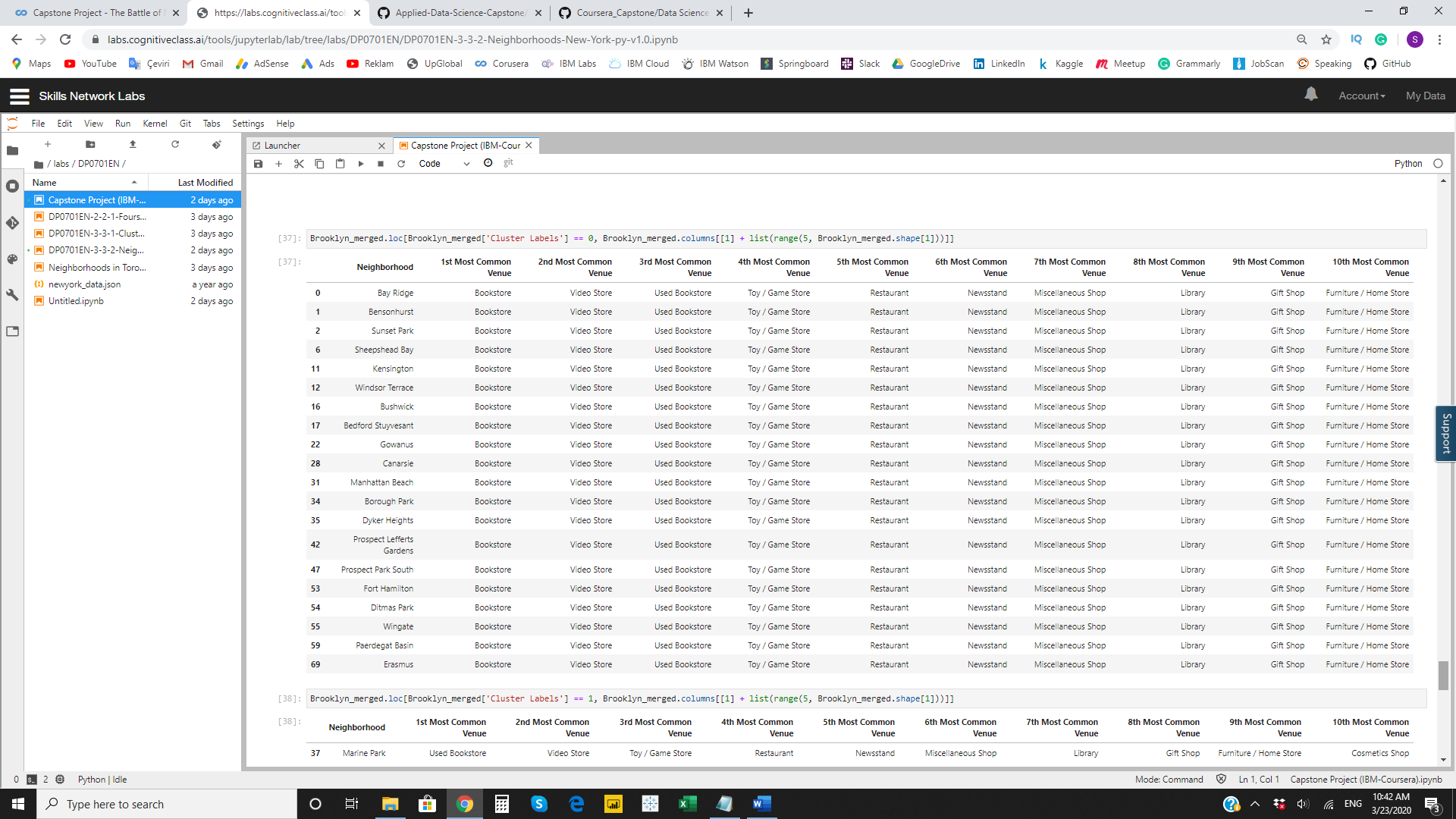


**RESULTS**

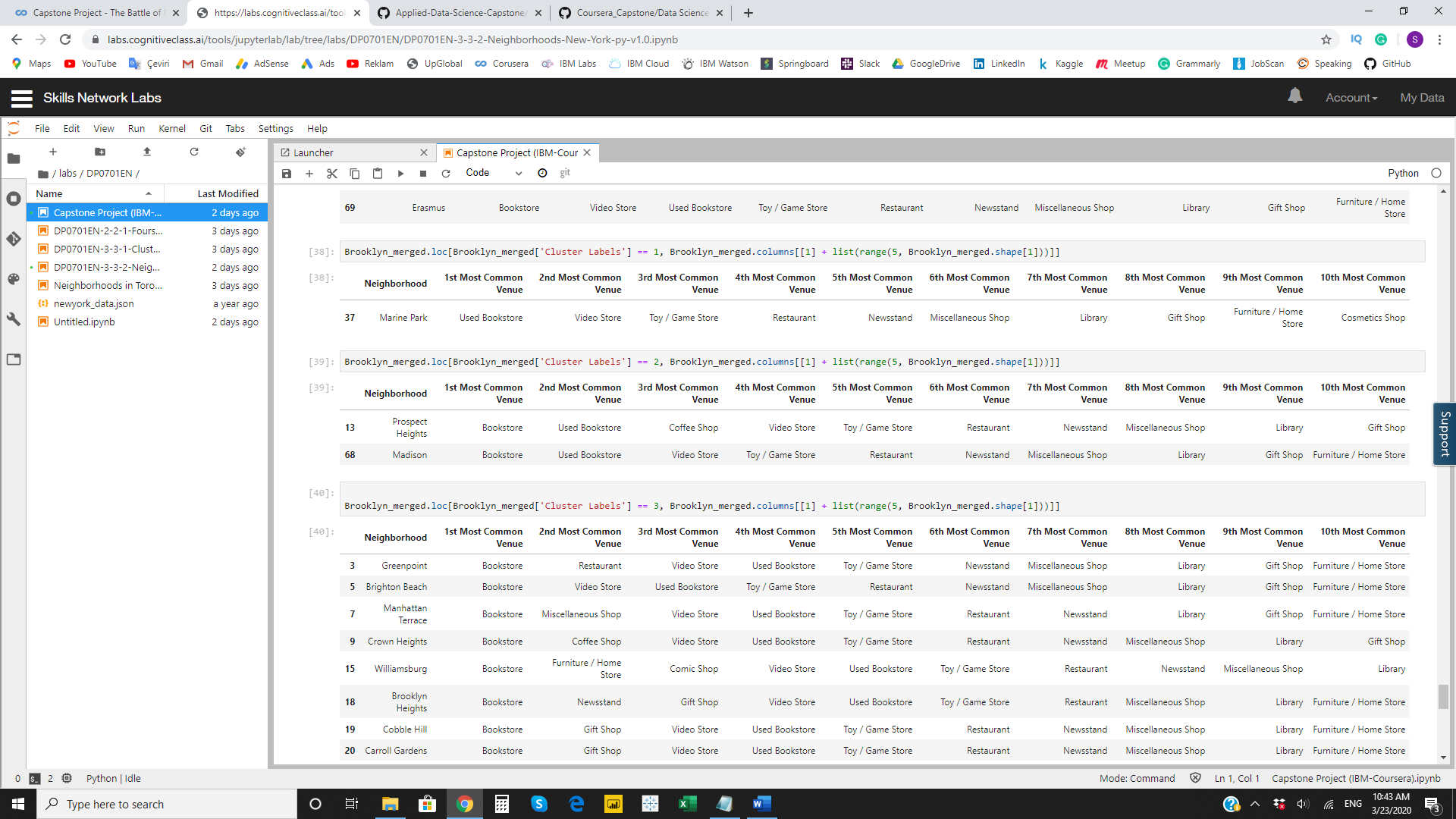
The results from K-means clustering show that we can categorize Brooklyn neighborhoods into 5 clusters. This is clusters on the map;



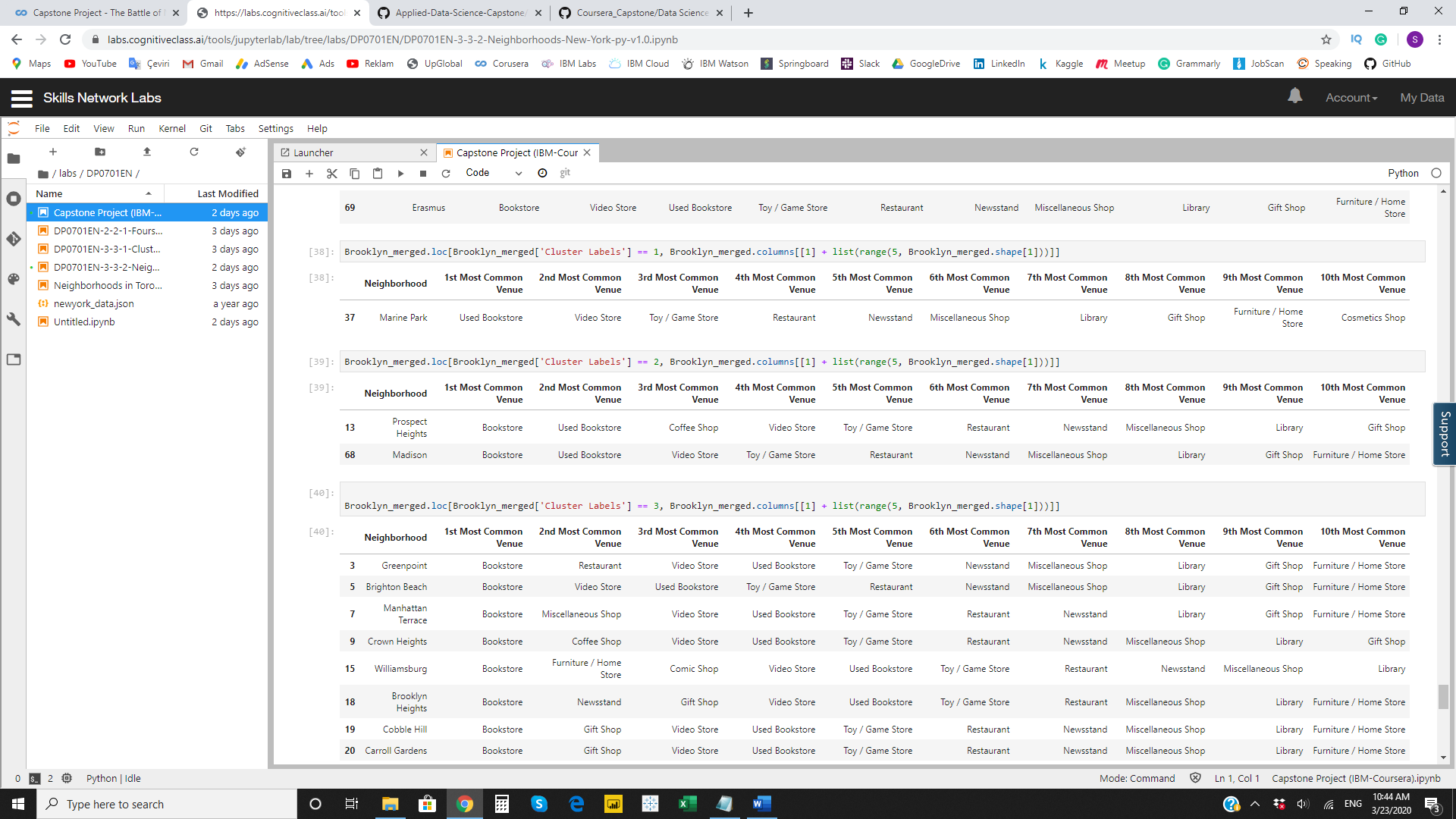
**Cluster 0**



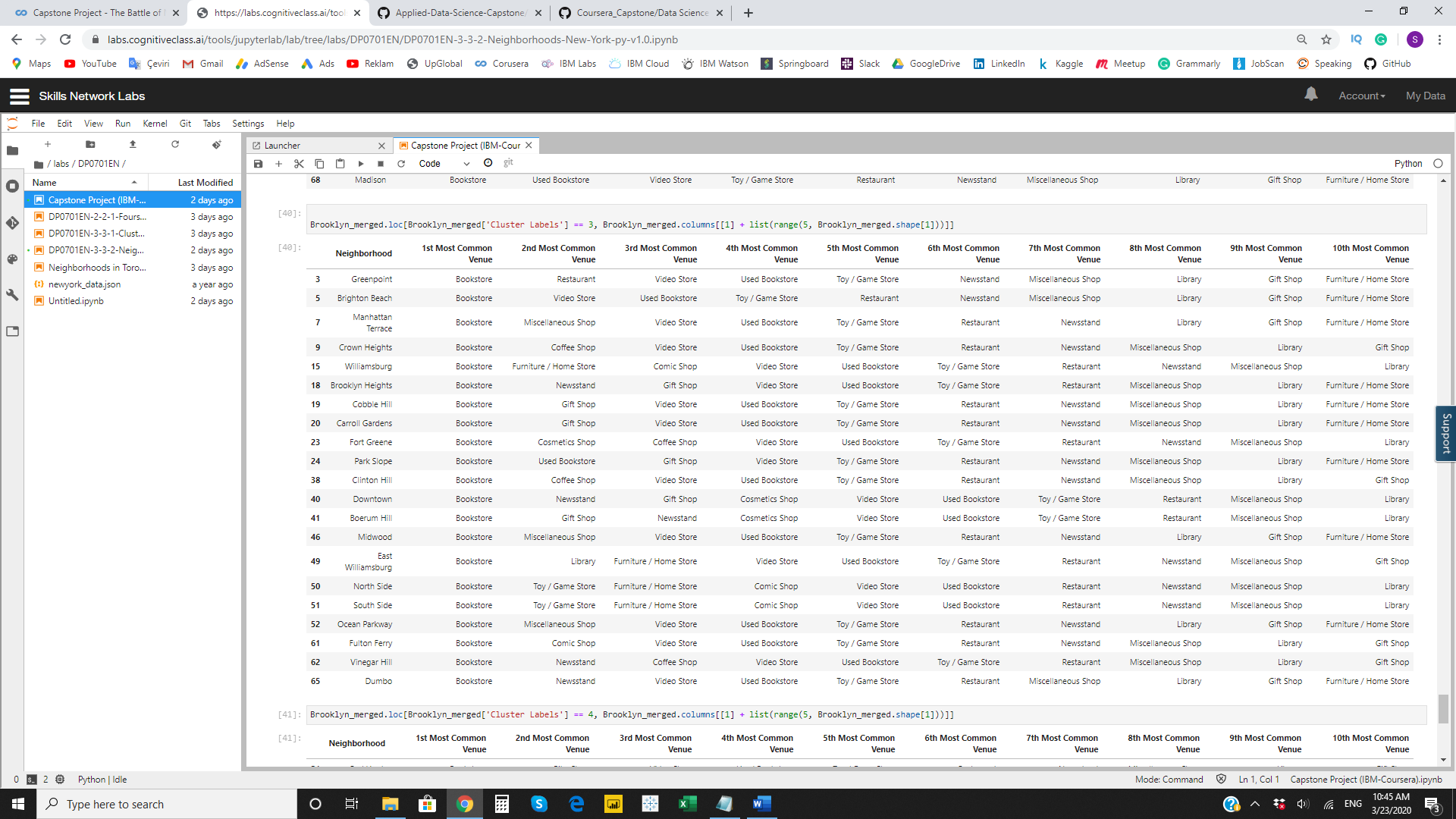
### **Cluster 1**



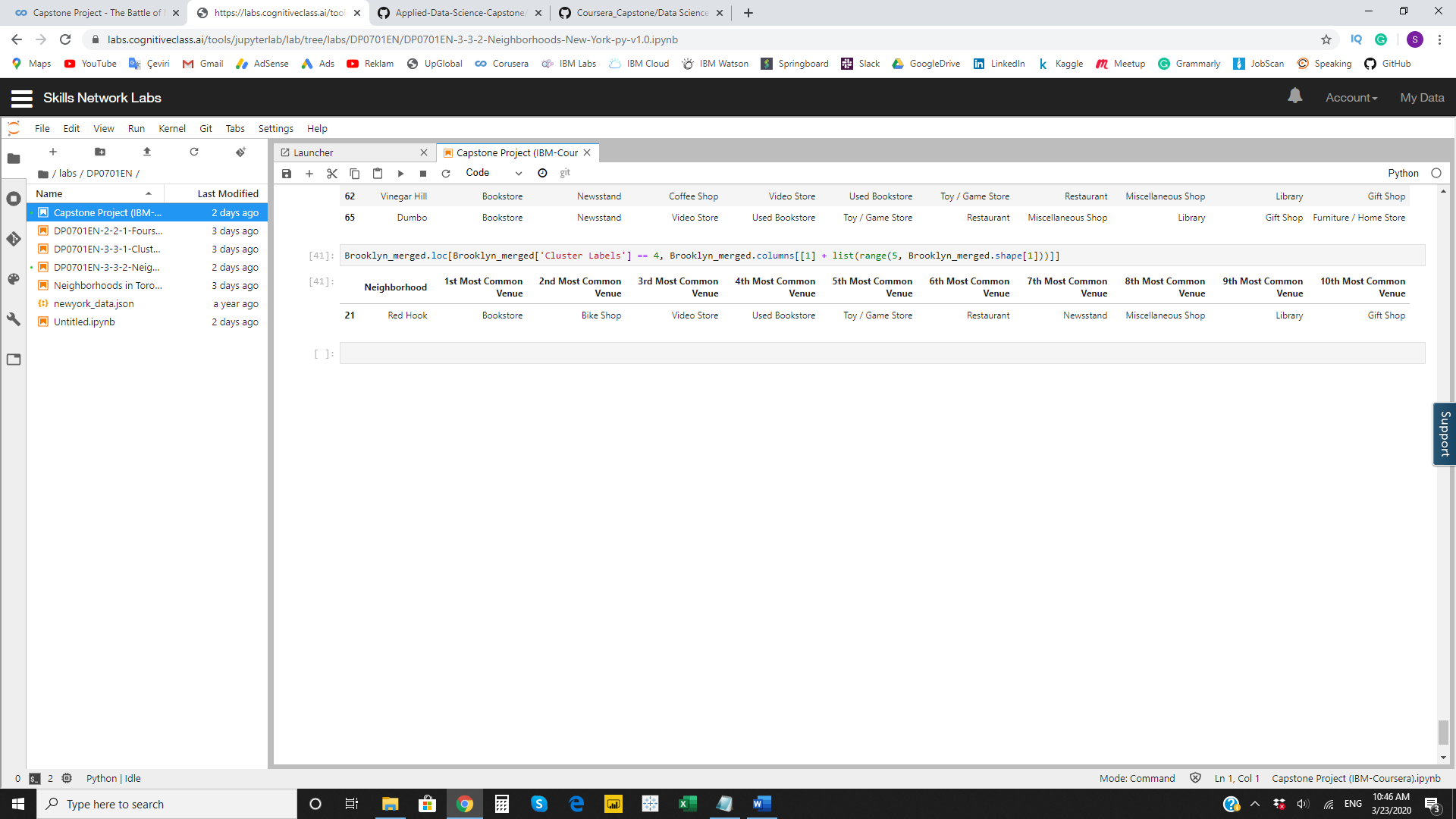
### **Cluster 2**



### **Cluster 3**



**Cluster 4**



**RECOMMENDATIONS**

In this project we only consider the factor of frequency of occurrences of venues to categorize neighborhoods. And this analysis is performed on limited data. It can be augmented be many other factors (e.g. other statistical measures like variance, various percentiles etc. of this number, colleges, schools, public transportation etc.) We are also relying on the data from Foursquare for our analysis.

Under these conditions, from our cluster analysis, it looks that Cluster 1 and Cluster 4 areas are the best choices to open a new bookstore. We would recommend **Red Hook** and **Marine Park** neighborhoods as a potential place for the new bookstore.