### Using Clustering to Choose a Restaurant Location

### Introduction:

The purpose of this project is to show how data science can be used to assist in the decision of where to locate a restaurant. Specifically, this project uses clustering to find out similar areas in New Orleans. Finally, this project will make a recommendation on which neighborhood to locate a restaurant, as well as possible ideas on how to market the restaurant based off of some key data insights.

Many people would stand to gain from having this type of information, restaurant owners (obviously), vendors that services restaurants, or city governments. This project can also be used as a template to solve a number of similar business related problems.

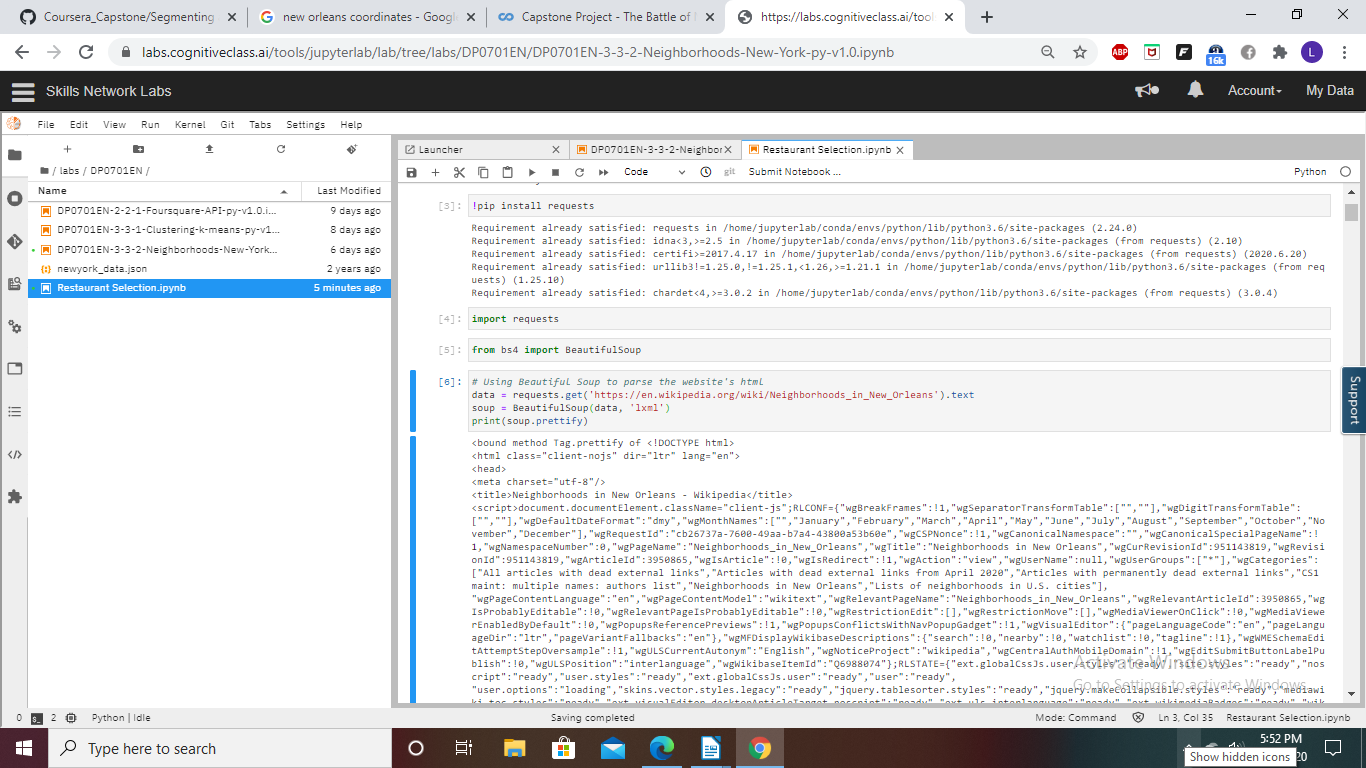
### Data:

We scraped a wikipedia table (https:/en.wikipedia.org/wiki/Neighborhoods\_in\_New\_Orleans) first to obtain the neighborhoods in New Orleans, followed by the geocoder package to find the latitude and longitude. Having the coordinates of all of the neighborhoods allowed us to use Foursquare to explore this problem.

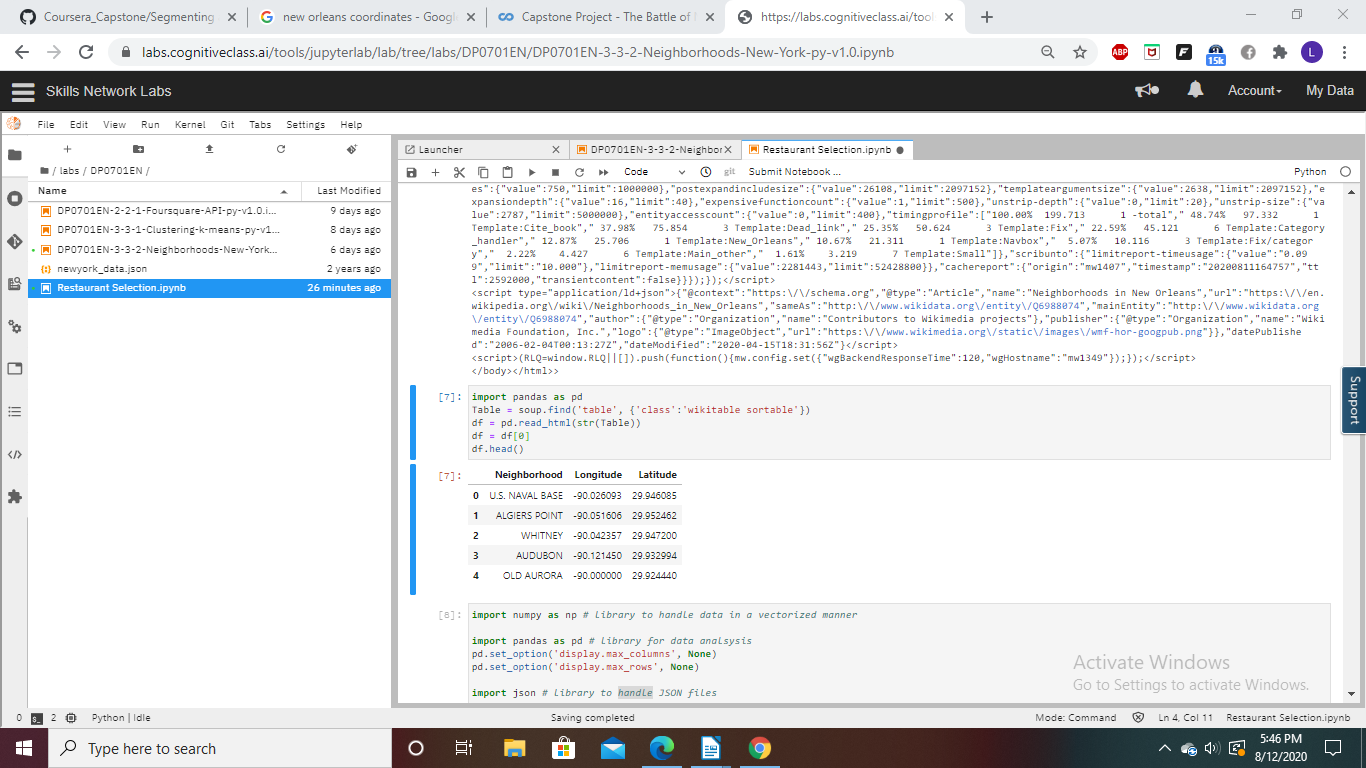
The data that is used in this project comes from Foursquare, one of the largest providers of location data.

### Methodology:

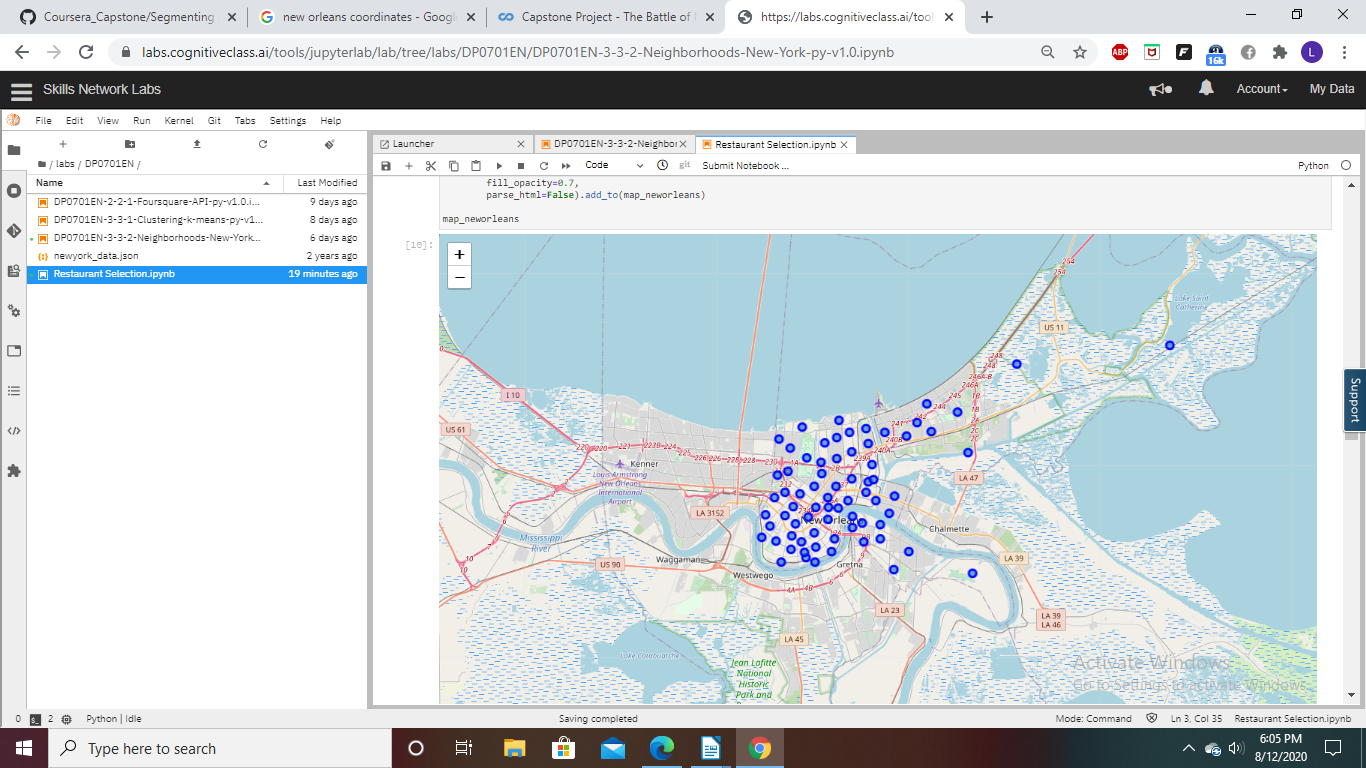
The first thing we did was add all relevant packages. We followed this up by downloading the neighborhood data for New Orleans, which was conveniently provided in table form by Wikipedia.



From this Wikipedia data (too long too show), we then created a data table.

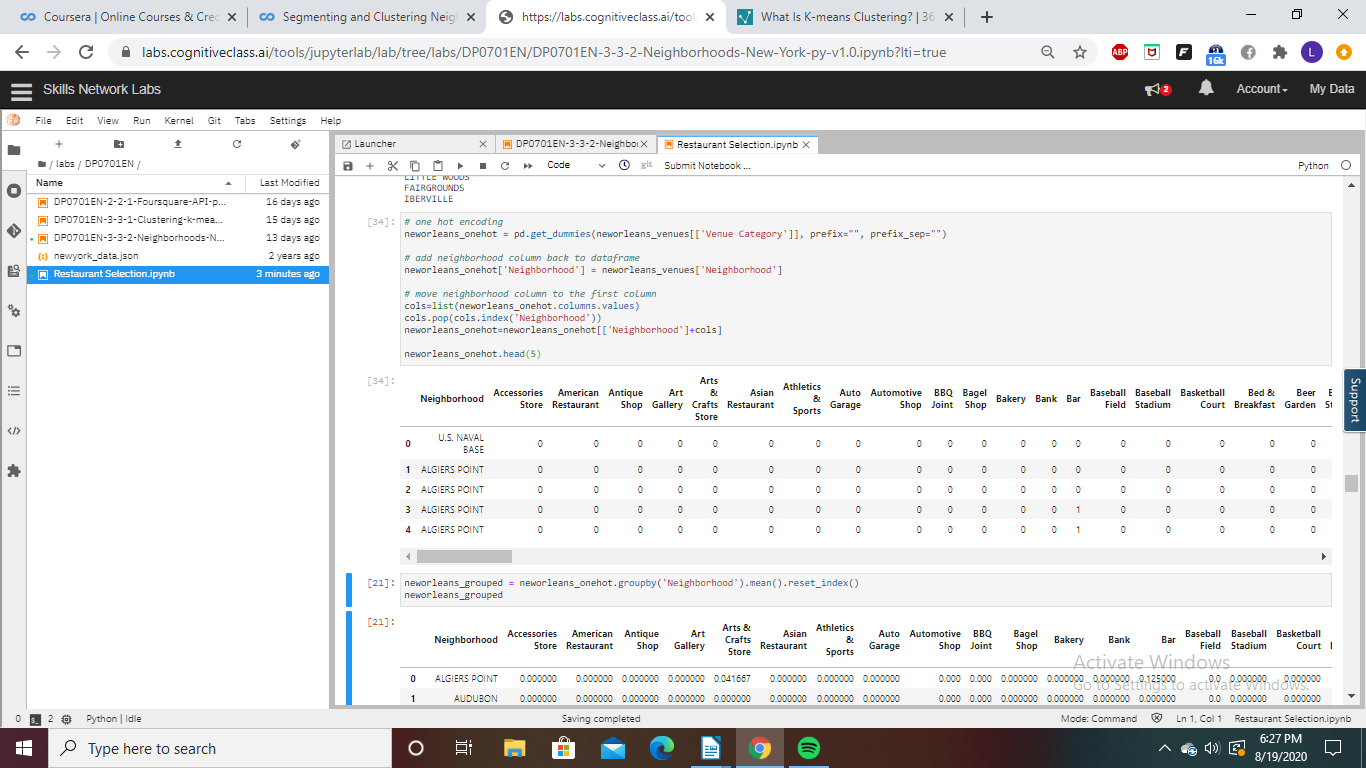


To quickly see where these neighborhoods were located, a map was created.

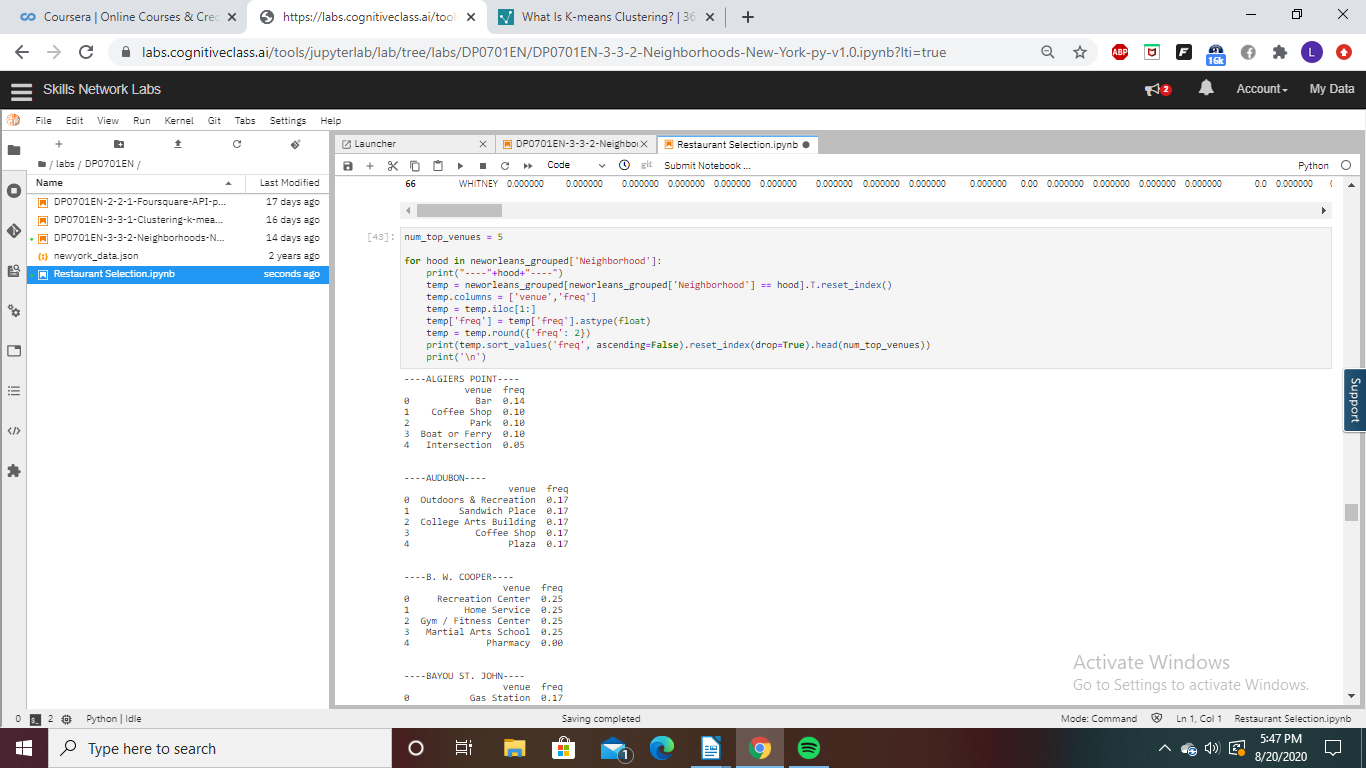


Based off the map and the exploratory data analysis, there appeared to be no missing data or obvious errors.

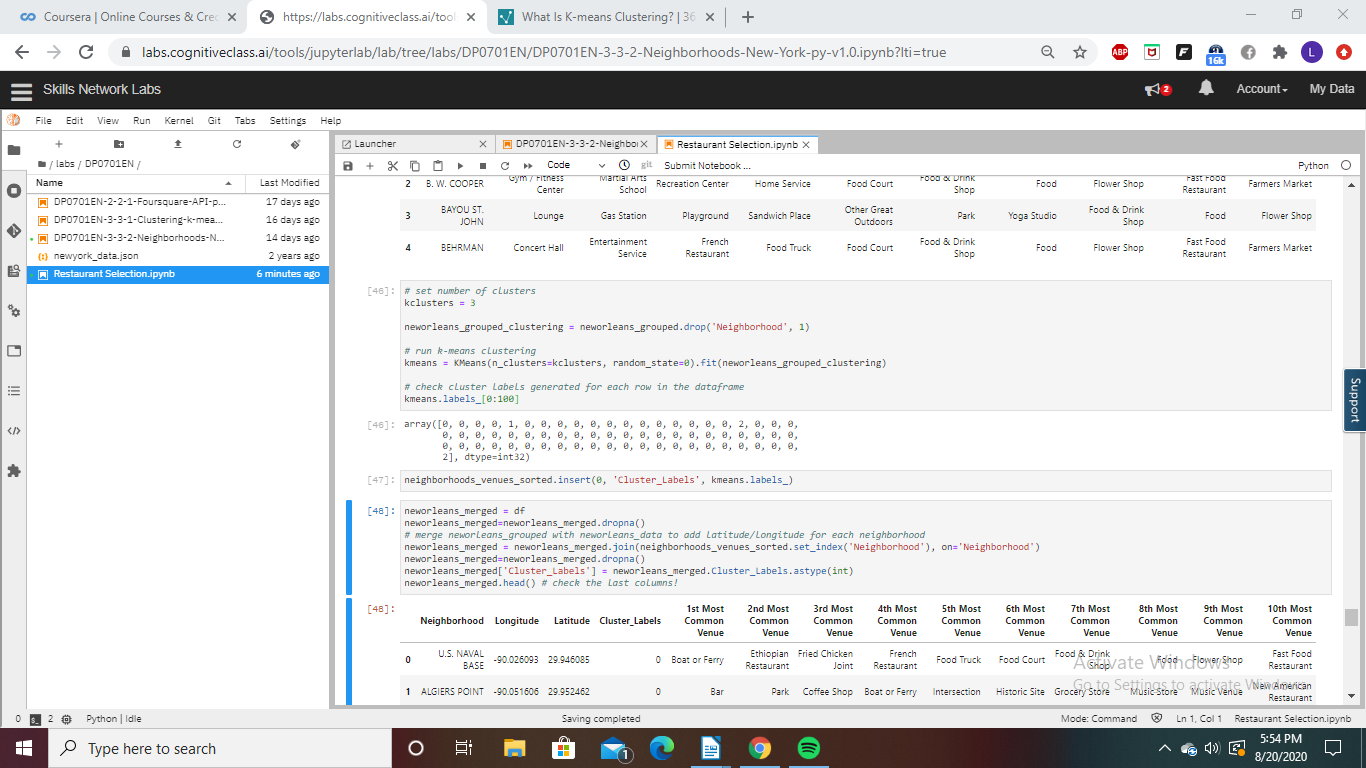
A function was created to find out the nearby venues. This data is qualitative, which doesn’t help out our data science tools, so one-hot encoding was used to put the data into a more usable form.

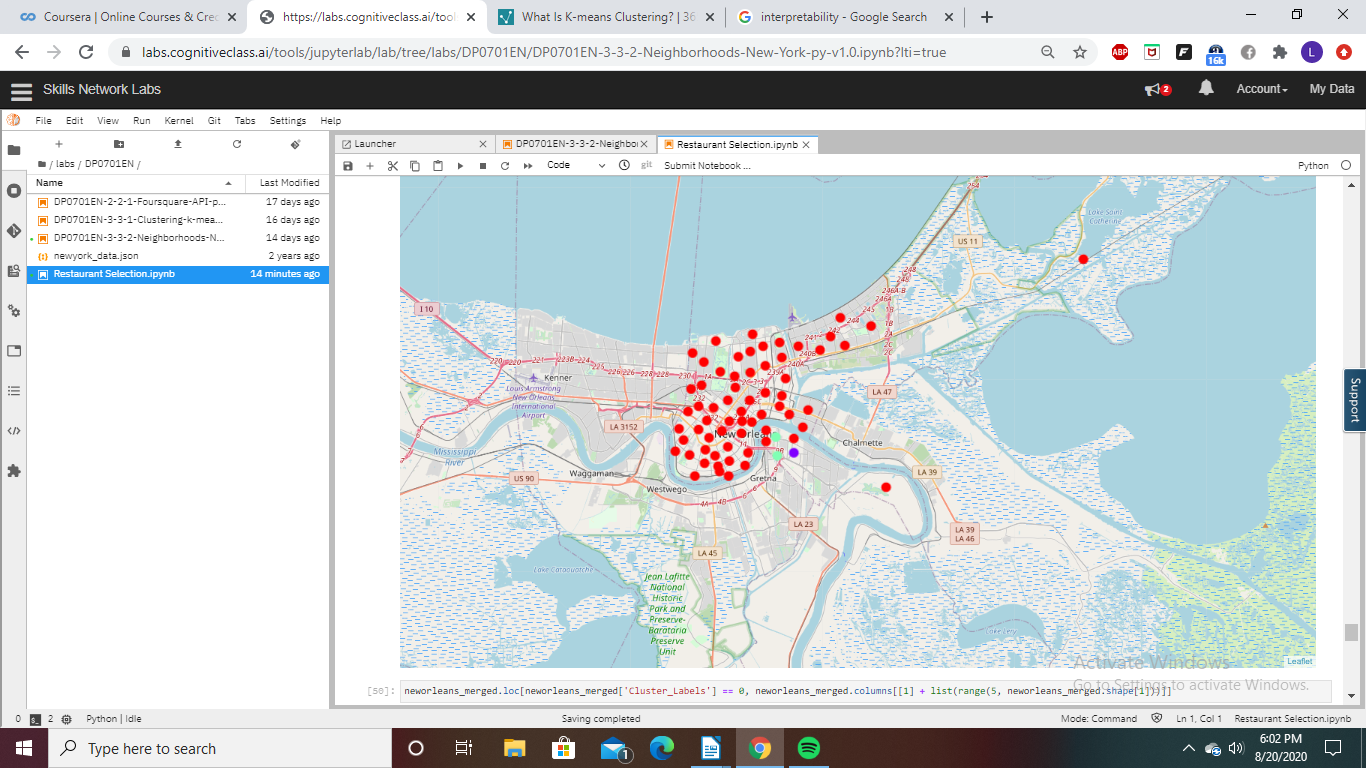


Next a list of the top five venues for each neighborhood was created.



To find out which neighborhoods were similar, K-means Clustering was used. To decide how many clusters we should use, the elbow test would normally be used. For ease of interpretability, I arbitrarily chose to use three clusters.





### Results:

Overall, the three groups that were created were not similar at all in magnitude.

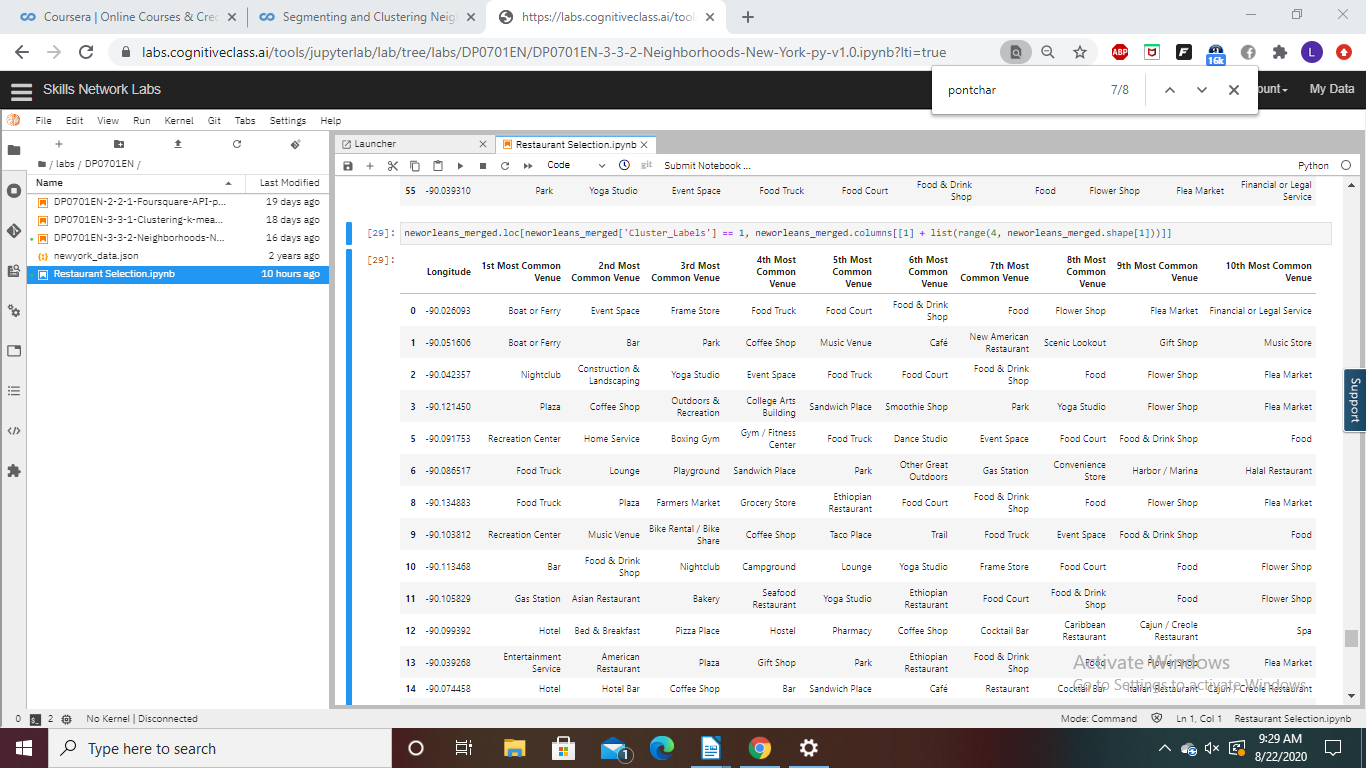
Cluster 0 Includes The Desire Area. ----

|  |  |  |
| --- | --- | --- |
| Desire Area | | |
|  | Venue | Frequency |
| 1 | Skate Park | 1.0 |
| 2 | Accessories Store | 0.0 |
| 3 | New American Restaurant | 0.0 |
| 4 | Nightclub | 0.0 |
| 5 | Nightclub Spot | 0.0 |

This cluster just involves a neighborhood with no businesses and a skate park. The population of this area is 2500.

As the 11th most dangerous neighborhood in new Orleans, with very limited population, it is recommended not to locate a restaurant here. Perhaps a new restaurant would turn this area around, and there would be future opportunities, but that would be a difficult task. Its almost as if our clustering algorithm was warning us to stay away from this neighborhood.

Cluster 1 Includes 71 different neighborhoods.



This cluster is the most promising to locate a restaurant in. In order to explore which of these places is the best to locate a restaurant, a more thorough exploration of the data would have to be done. Some ideas for further exploration are given in the discussion section.

Cluster 2 Includes The Lakeshore- Lake Vista Neighborhood and Pontchartrain Park.

|  |  |  |
| --- | --- | --- |
| Lakeshore- Lake Vista | | |
|  | Venue | Frequency |
| 1 | Harbor/ Marina | 0.5 |
| 2 | Park | 0.5 |
| 3 | Pharmacy | 0.0 |
| 4 | New American Restaurant | 0.0 |
| 5 | Nightclub | 0.0 |

|  |  |  |
| --- | --- | --- |
| Pontchartrain Park | | |
|  | Venue | Frequency |
| 1 | Park | 1.0 |
| 2 | Accessories Store | 0.0 |
| 3 | Pharmacy | 0.0 |
| 4 | New American Restaurant | 0.0 |
| 5 | Nightclub | 0.0 |

### Discussion:

This is a very interesting example that shows how if you are not careful, a data science tool can give results that are completely useless. Our algorithm created three clusters. They can be summarized as follows:

Cluster 0: Skate Park and no businesses.

Cluster 1: 71 Different neighborhoods.

Cluster 2: Parks/ Marinas and no businesses.

Cluster 1 is clearly the most promising area to locate a restaurant. The recommendation is to do a further exploration of the neighborhoods in cluster 1. Some possibilities of variables to explore that would make the selection choice easier are:

1. Crime Rate

2. Population

3. Number of restaurants

Other factors to explore are:

1. Whether the area is growing and shrinking

2. Average income and demographics

3. Failure rate of restaurants in the area

The quantitative analysis of which neighborhood should be combined with a through qualitative analysis of the area.

Conclusion:

The main practical benefit of this project for a business owner is to instruct on how machine learning algorithms can often produce results that are flawed. Data can be used for good, but when misused, data can even be counterproductive to business success. This is why it is so important that a data scientist digs deep into the data and fully understands the possible biases and flaws in his/her data algorithm. A data scientist must have a skeptical mindset and not blindly trust in his/her model without checking for flaws. There are unlimited possibilities for using data science to increase business effectiveness. It is simply a matter of using the right data, the right data science tools, and a dedication to finding the truth.