Battle of the Neighborhoods Capstone

12/23/18

Pete Gil

Introduction

An eager venture capitalist is interested in investing into a hot new restaurant in Brooklyn, NY. This will be a new construction and investment. This will initially focus on finding the best neighborhood based on the number of restaurants within a neighborhood in Brooklyn, NY.

The data will come from a public New York City source and will focus on the borough of Brooklyn. Foursquare data will also be integrated to get coordinates and venue info to establish a better idea in which what area in Brooklyn would be the best to invest in.

Data

Data used is from a public website for New York City. The data consists of all boroughs in New York City and their corresponding coordinates. In order to help solve the problem, data will be used to determine all data for the borough of Brooklyn and coordinates to use in conjunction with Foursquare data will use for extracting venue information.

Data source https://geo.nyu.edu/catalog/nyu_2451_34572

Pre-Processing

The pre-processing of the data was executed by pulling data from the original data source, merged into a Python data frame, and then integrated with Foursquare data to find venues to corresponding neighborhoods and coordinates. The final data set has neighborhood, coordinates, and venue information.

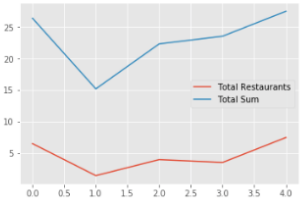
Methodology

With all the pre-processing done, the kmeans machine learning algorithm and clustering on the dataframe is ready for execution. Five clusters were used to determine totals. The end results show Total Restaurants, Total Sum, and the cluster. The end results also show that the K5 cluster has the most restaurants.

Discussion

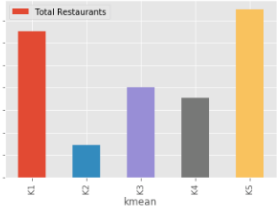
These charts show graphically where the results fall. The charts bring up the discussion of what cluster is the best to focus on for the new restaurant development. It's clear there are many choices - either the new development can focus on the cluster with the most or less totals.

```
In [92]: #let plot this data

         %matplotlib inline

         import matplotlib as mpl
         import matplotlib.pyplot as plt

         mpl.style.use('ggplot') # optional: for ggplot-like style
         print ('Matplotlib version: ', mpl.__version__) # >= 2.0.0

         Matplotlib version:  3.0.2

In [108]: # plot bar chart
          lines = brooklyn_final.plot.line()
```




```
In [118]: # a scatter plot comparing num_children and num_pets
          brooklyn_final.plot(kind='line',x='kmean',y='Total Restaurants',color='red')
          plt.show()

          brooklyn_final.plot(kind='bar',x='kmean',y='Total Restaurants')
```


```
Out[118]: <matplotlib.axes._subplots.AxesSubplot at 0x7f60a14d5a90>
```



Conclusion

Final results show that the final decision of the investment would likely focus on K5, K1, and K2. Between these three, the investment will most likely not focus on an area that has the least amount of restaurants (K2), but with the most (K1, K5). The final recommendation would focus on these neighborhoods in K5.

	Neighborhood	Group
11	Bushwick	5
19	Cypress Hills	5
20	Ditmas Park	5
28	Flatbush	5
30	Fort Greene	5
56	Prospect Park South	5
62	South Side	5
64	Sunset Park	5