Coursera IBM Data Science - Course 9 Capstone Project

Topic: The Battel of Neighborhood

Content:

Note: This report is kept short since most explanation is within the python code, either in Pycharm of the Jupyter Notebook.

1. Topic Description

For the Capstone, I decided to use the New York dataset available in the course and append it with data from geojson.

To get a general overview of New York's Venus Density. We will see that the density of venues and food related venues greatly differs between Queen's neighborhoods.

Interested users might take recommendation from the data as to where to look for resturants in Borough Queen in New York city.

2. Data Description

The New York data for this topic was originally taken from the Capstone course of IBM data science, from Wikipedia and other course sources.

Data is loaded from IBM Shared Box.

https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json

Original source of data is as below

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

Queen borough and neighborhood is fetch into table. Based on that table, Foursquare is used to generate a new data frame that includes the found venues and venues categories for Queen neighborhood.

Based on the processed data frame, 5 clusters are used to categories Queen neighborhood.

The clustering only focuses on food related institution (restaurants, pub etc.), therefore all other venues categories were dropped for this instance.

3. Methods used for data exploration

The methodologies during the Queen's data analysis can be summarized quickly:

- Data wrangling with Python functions, loops, algorithms and regular expressions

- Pandas dataframe manipulation and HTML parsing

- Data visualization with folium and Jupyter Notebooks

- Basic clustering with sklearn and KMeans

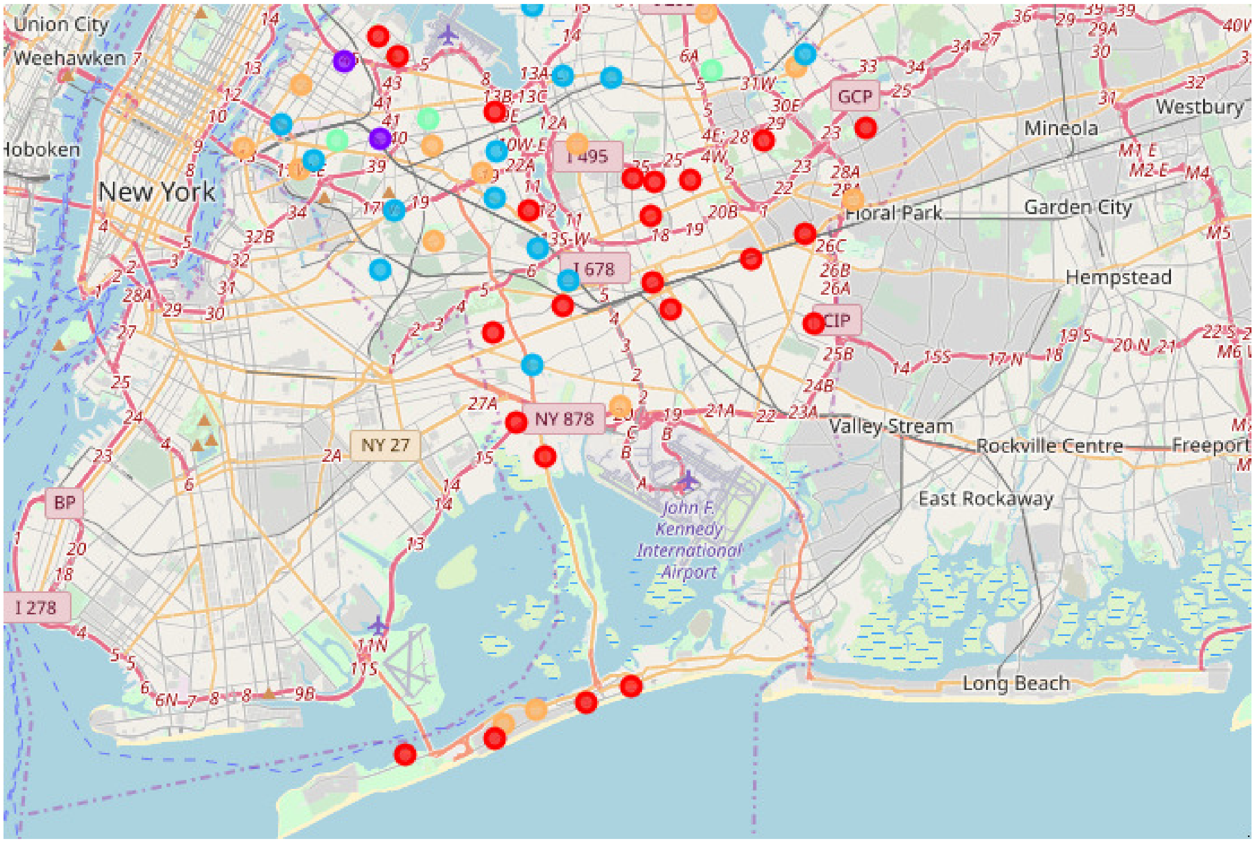
- Basic data analysis and sorting with indexing functions

The import of the data as well as the manipulation and formatting were done with Python “pandas” toolkit to create and manipulate data frames. The Python module “beautifulsoup4” was used to parse data on geojson and import the results into dataframes with pandas.

Functions of pandas such as groupby, aggregate, count etc. were used together to format the data for later analysis during each step of the code. When necessary,

With the Python module “folium”, a map was created to show the overall venue density in Queen with the help of geojson data. It should (again) be noted that these data files and the corresponding maps might not be 100% coordinated with the datasets.

In addition to the visualization, a quick and easy KMeans cluster was created with the Python module “sklearn”. The Toronto neighborhoods were distributed into 4 clusters based on how many venues they each hold.



The statistics aspects here were kept very basic, my focus was on quick outputs and easy data differentiation.

4. Discussion

The results so far show that Queen, as expected from a world metropole, contains a high number of neighborhoods and corresponding venues.

Looking into detail, there is a large focus on food related places in Queen neighborhood since their more common than other venue types. This might be due to Foursquare users being focused on visiting restaurants and giving their respective ratings.

The neighborhood with both the highest amount of single venues and he highest diversity of venue categories would be Sunnyside Gardens with around 76 venues.

From the perspective of a shop or restaurant owner, the competition among Deli / Bodega
seems rather high. Deli / Bodega shops are the most common venue
types.
Dunkin' Donuts is the most common venue in Queen neighborhood. Total 407 venue in neighborhood.

An interesting point is also the large spread of venue amounts between different neighborhoods. The spread tends to rather high overall, with a length of 407 venues on average between the highest density cluster and the least dense while the least dense cluster also inhabits the highest amount of neighborhoods overall.
Results can be seen in detail on the respective Github links above.

5. Conclusion

All in all, the venue distribution in Queen seems rather unequal, but this could also be due to skewed data from Foursquare or the method application.

A future restaurant owner might be inclined to open his restaurant in a less density area of Queen neighborhood to be protected from competition, given that there is also enough population.

Tourists should consider Sunnyside Gardens because of its high venue diversity.