A Neighborhood Recommender for Upcoming Eateries & Restaurants

Week 2 - Part 2: Report

Table of Contents

Introduction

Every business, big or small is a vision backed by hardwork, determination and aspirations. A drive to take an idea from its inception to fruition. Therefore, it becomes all the more neccessary that one's blood and sweat don't go in vain.

Presenting, RESTROLOGY: A Neighborhood Recommender for owners of upcoming Eateries & Restaurants. Restrology analyses local geographical data via Foursquare and recommends restaurant/eatry type(cuisines) for upcoming eateries and restaurants based on one's neighborhood preference for a better shot at success. Restrology analyses all the eateries and restaurants in every neighborhood of a city and then creates a list of top 10 spots (Restaurant/Eatery type) in every neighborhood displayed in percentages of the total restaurants in that particular neighborhood. A prospective business owner can use Restrology to see which cuisine based restaurants are lacking or what type of Restaurants are doing well(due to their high number) in which neighborhoods. He can then make an informed decision and fill the void and have a better chance of establishing a successful business. Restrology is like Astrology for Restaurants.

For example, suppose Mike who is really passionate about food and different cuisines wants to invest his savings in the restaurants business in New York. Since he wants to make sure a steady Return on his Investment, he uses Restrology to analyze the restaurant data of the New York city. He can get the desired data per every neighborhood for the entire city and see which cuisine based restaurants are the least in nuber per neighborhood or he can simply see the same list for his choice of neighborhood. Suppose he is leaning towards opening a restaurant in the 'Midtown' area, he then uses Restrology to see which type of restaurants are prevelant in that particular area and in the adjoining neighborhoods. He observes that among all other cuisines, the area of his choice is lacking 'Steak House' and a 'Sushi' joint or may be he see high concentration of Pizza places due to the neighborhoods proximity to University/Shopping District. Restrology helps individuals like Mike make an informed decision and boost their chance of success.

Data

Restrology, for the purpose of this project, will be restricted to New York City. Therefore, the recommendations made for restaurants/eateries will be from neighborhoods and boroughs of New York city only. We will need a couple of datasets and will integrate them to get the desired outcome which are as follows.

Datasets required for the Project:

1) New York Data (Boroughs + Neighborhoods)

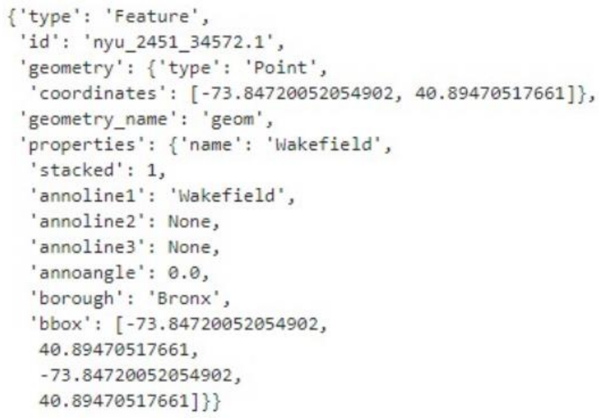
2) Four Square City Guide Data (Venues)

New York Data (Boroughs + Neighborhods)

The first Dataset we will be using would contain all the required geographical data about New York city. Namely, we would be using 'Borough', 'Neighborhood', 'Latitude', 'Longitude' among all the other data elements present in the data. For convenience, we would be using the same data set which was provided to us in Week 3 of this course (Applied Data Science Capstone) https://geo.nyu.edu/catalog/nyu_2451_34572 (https://geo.nyu.edu/catalog/nyu_2451_34572). We will use the same link like we did to load this data from where it is downloaded and hosted https://cocl.us/new_york_dataset (https://cocl.us/new_york_dataset)

New York city has a total of 5 boroughs and 306 neighborhoods. In order to segement the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and logitude coordinates of each neighborhood. The 'Latitude' and 'Longitude' extracted from this dataset will also be pivotal when we use it perform Clustering using K-Means.

All the relevant data is in the features key, which is basically a list of the neighborhoods. If we dive into the elements of this features key, we will find all of its components. Below is a snapshot of what a single item would look like.



```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```

2) Four Square City Guide Data (Venues):

Foursquare City Guide, commonly known as Foursquare, is a local search-and-discovery mobile app which provides search results for its users. The app provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history.

Before we can learn how to retrieve data from the Foursquare database, we need to create a developer account. On May 31st of 2018, Foursquare updated their API, and unlike before, now there are some limitations on how many calls you can make to the API. So when you create a developer account, the default type is the sandbox account, with 950 regular calls per day and 50 premium calls per day, and you can retrieve only one photo and one tip per venue. The account used in this project is a personal account which is still free, and with it you get 99,500 regular calls and 500 premium calls.

That is actually 100 times more calls than the default sandbox account. You'll also get access to over 105 million venues or points of interest, but you still only get two photos and two tips per venue, which is just one more photo and tip compared to the sandbox account. The Foursquare API credentials you need to create are: your Client ID and your Client Secret. You will need to pass these credentials every time you make a call to the API. You can go to the onliine Foursquare documentation and click on API and discover endpoints overview, and there you will find the different groups and endpoints available and whether each endpoint falls under a regular or a premium call.

So using the Foursquare API, we can search for specific type of venues or stores around a given location. it is important to remember that for this data, we make a regular call to the API, and if you have a free personal developer account, you can make up to approximately 99 thousand regular calls per day. We can also learn more about a specific venue or store or shop, like their full address, their working hours, and their menu if they have one, and so on. It's also important to remember that for this data, we would need to make a premium call and with the personal developer account, you can make approximately 500 calls per day. Also with the Foursquare API, we can learn more about a specific Foursquare user, their full name, and any tips or photos that they have posted about venues and stores. For this data, a regular call to the API would be made. Furthermore, we can explore a given location by finding what popular spots exist in the vicinity of the location, and for this data a regular call to the API would be made. And finally, with the Foursquare API, we can explore trending venues around a given location. These are venues with the highest foot traffic at the time this regular call to the API is made.

Exploratory Data Analysis

Exploring Datasets

Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segement the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and logitude coordinates of each neighborhood.

- Extract all the relevant data which are basically a list of Neighborhoods
- Transfer the extracted data into a Data Frame
- Fill the Data Frame with 'Borough', 'Neighborhood', 'Latitude', 'Longitude' data
- For display, create a map of NY city with Neighborhoods superimposed on top
- Segment the neighborhoods of all 5 boroughs
- Visualize all 5 boroughs with all the neighborhoods in it(Brooklyn, Manhattan, Queens, Bronx, Staten Island.
- Explore all 5 Data Frames which are created for eevery borough
- Use Foursquare API to explore venues in all the neighborhoods
- Extract category and clean the json file to produce new Data Frames
- Repeat calling the Foursquare API for all 5 boroughs

Explore Neighborhoods in all 5 Boroughs

- Extract complete neighborhood list for all 5 boroughs
- Create Data Frames covering the venues of all neighborhood
- Group all the returned venues by 'Neighborhood'
- Find out how many unique categories are present in returned 'venues'

Analyze Each Neighborhood of Each Borough

We will analyze each neighborhood of each borough (Brooklyn, Manhattan, Queens, Bronx and Staten Island) and the final product of this stage would be a Data Frmae which top 10 most common Ratautant/Eatry type in each neighborhood which would be repeated for each of 5 boroughs.

- Create One hot ecoding Data Frames for each Borough
- Group rows by 'Neighborhood'
- Take the mean of the frequency of occurrence of each category
- Manually Selecting (Subsetting) Related Features for the Restaurants/Eatries
- Updating the One-hot Encoded DataFrame
- Extracting Top 10 Restaurant/Eatry from 1 Neighborhood in Brooklyn

- Extracting Top 10 Restaurant/Eatry from 1 Neighborhood in Manhattan
- Extracting Top 10 Restaurant/Eatry from 1 Neighborhood in Queens
- Extracting Top 10 Restaurant/Eatry from 1 Neighborhood in Bronx
- Extracting Top 10 Restaurant/Eatry from 1 Neighborhood in Staten Island Let's put that into a pandas dataframe and sort the venues in descending order

Clustering

Now that we have the sorted Data Frames from all 5 boroghs containing all the neighborhoods, it is time we use clustering. We will be using k-means clustering. We would be using k-means to clusterall of our 5 boroughs:

- Brooklyn
- Manhattan
- Queens
- Bronx
- Staten Island

Why k-means?

K-Means can group data only unsupervised based on the similarity of customers to each other. There are various types of clustering algorithms such as partitioning, hierarchical or density-based clustering. K-Means is a type of partitioning clustering, that is, it divides the data into K non-overlapping subsets or clusters without any cluster internal structure or labels. This means, it's an unsupervised algorithm. Objects within a cluster are very similar, and objects across different clusters are very different or dissimilar. So we can say K-Means tries to minimize the intra-cluster distances and maximize the inter-cluster distances.For example you may use Euclidean distance, Cosine similarity, Average distance, and so on. Indeed, the similarity measure highly controls how the clusters are formed, so it is recommended to understand the domain knowledge of your dataset and datatype of features and then choose the meaningful distance measurement.

- Create DataFrames that includes the clusters as well as the top 10 venues for each neighborhood
- Repeat the above step for all 5 boroughs
- Visualize the resulting clusters

Exploring clusters

Examine each cluster in each borough and determine the discriminating venue categories that distinguish each cluster

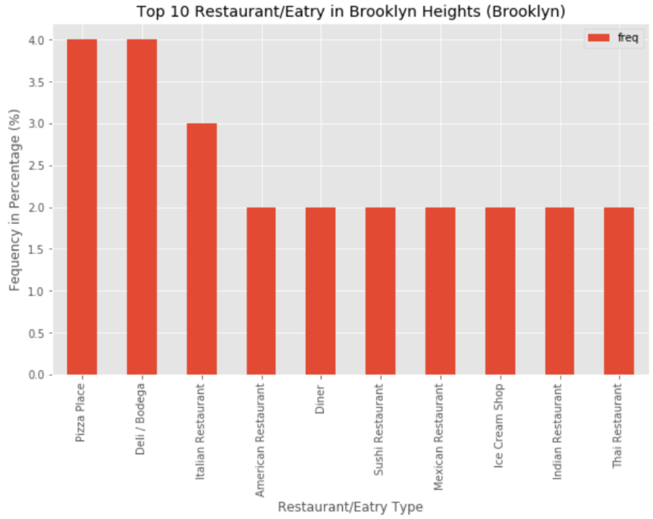
Results

Restrology analyses all the eateries and restaurants in every neighborhood of a city and then creates a list of top 10 spots (Restaurant/Eatry type) in every neighborhood displayed in percentages of the total restaurants in that praticular neighborhood. A prospective business owner can use Restrology to see which cuisine based restaurants are lacking or what type of Restaurants are doing well(due to theirr high number) in which neighnorhoods. He can then make an informed decision and fill the void and have a better chance of establishing a successful business.

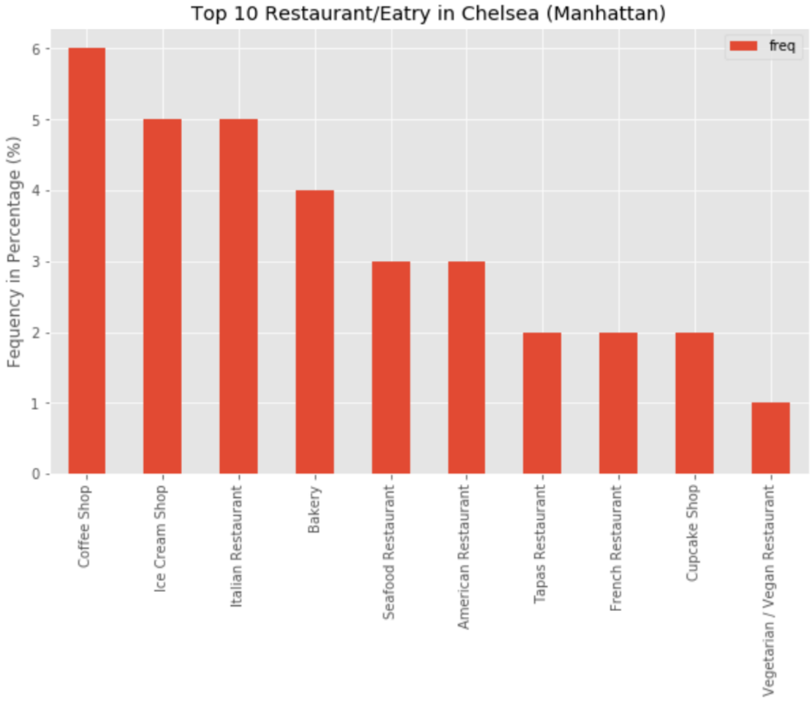
Results in Graphs:

For the sake of scalability and size constraints, we will only present Restrology displaying Top 10 Restaturants/Eatries for only 1 sample per Neighborhood for all the 5 Boroughs (Brooklyn, Manhattan, Queens, Bronx, Staten Island). We will be creating Bar plots using Matplotlib which is a very helpful package for visualization in Python. The style we would be using will be 'ggplot' which is yet another plotting style quite well known in R.

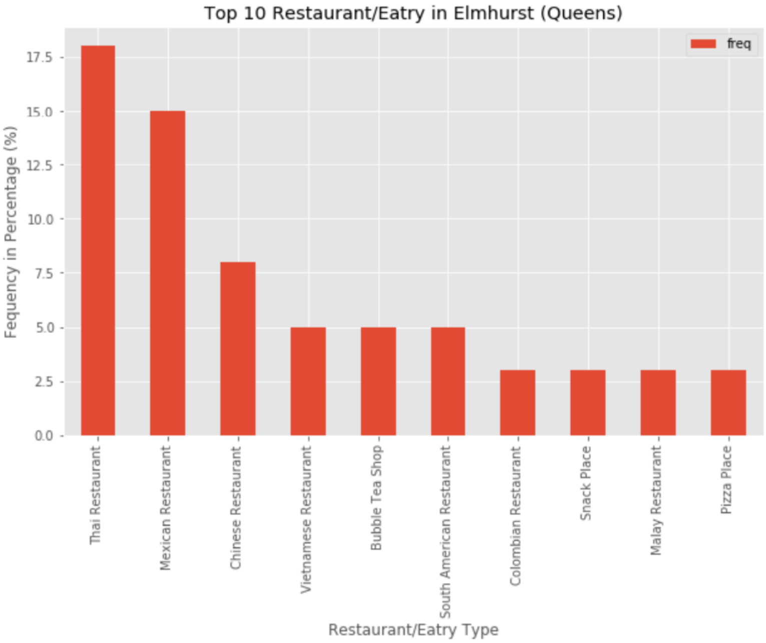
Brooklyn



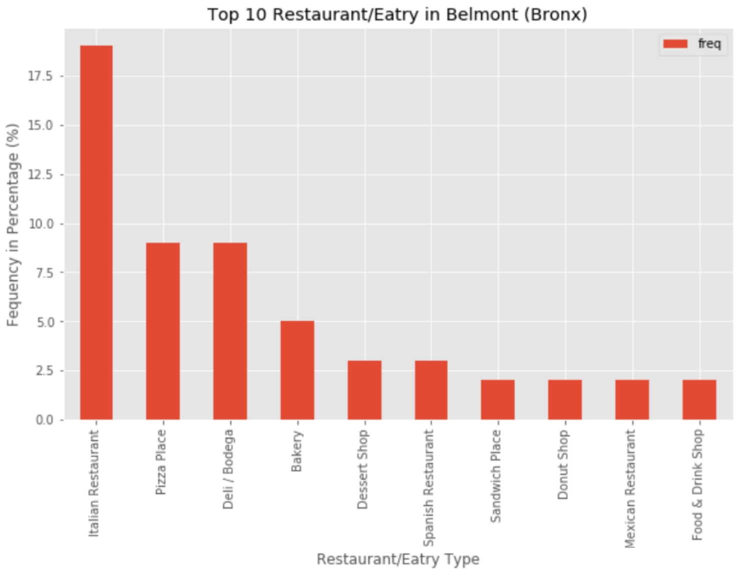
Manhattan



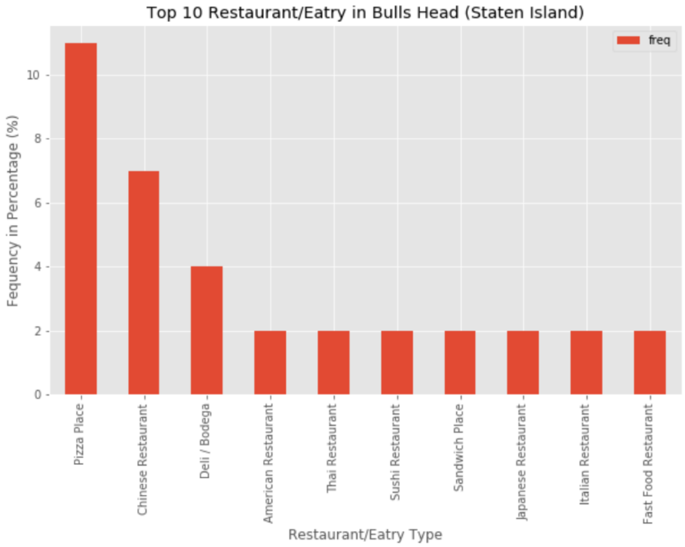
Queens



Bronx



Staten Island



Observations & Recommendations

With this, we have covered pretty much all the steps one needs to take as a Data Scientist or a Data Science Professional. We started with fetching the data, then proceeding to Exploratory Data Analysis before advanding to apllication of Machine Learning algorithms (k-means clustering) in our case. While it is a certainly a significant step in the right direction, there are a few lessons learnt which can now be discussed as Observations & Recommendations and are as follows:

Recommendations:

First, we can integrate other data sets as well, which can add and quantify other factors for Restrology. Like, Store Prices in neighborhood can also be factored in and algorithms can then be applied for the optimal choice.

Similarly, apart from the current data set, we can also integrate some Past data where we have access to closed business's data in the neighborhood. This would help us further strengthen Restrology to somehow quantify and predict the chances for 'Success' for any given Restaurant/Eatry.

We can explore other features from services like Foursquare API to include and factor in parameters like 'Trending' to further enhance Restrology. Something which we were not able to do due to the fact that we have used a free developer account.

Data from nearby cities can be integrated as well to uncover more trends and insights.

Conclusion

This concludes our journey with Restrology in this Capstone Project. During this journey we covered all the gospels in the field of Data Science and Machine Learning. Given the extensive scope of the subject which seems limitless, we have merely made a dent and there are far bigger and tougher horizons for us to conquer. A prospective business owner can use Restrology to see which cuisine based restaurants are lacking or what type of Restaurants are doing well(due to theirr high number) in which neighnorhoods. He can then make an informed decision and fill the void and have a better chance of establishing a successful business.

This was the main purpose of Restrology, to help make sound descisions. To replace haste conclusions with data driven approach. To help prospective Restaurantuers take the plunge with a sense of certainity. Although, the current version might look quite rudimentary, it creates the bedrock for a lot of possibilities which can make Restrology into something revolutionary.