IBM DATA SCIENCE CAPSTONE

WEEK 4

INTRODUCTION

This is a capstone project for IBM Data Science Professional Certificate. My friend has got admission into Harvard University .So, In this project, I am going to help him find a good neighbourhood in Boston for him to live. and I am designing this project to help him find the most suitable location.

BUSINESS PROBLEM

Many People Like my friend who have gone to Boston recently want to find a good Neighbourhood in Boston Which has got different types of venues. The objective of this capstone project Is to find the most suitable location for my friend in Boston. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the question: Finding the Neighbourhoods which has got good restaurants, near to his University.

TARGET AUDIENCE

People who want to Find Neighbourhoods to live in Boston.

DATA

To solve this problem, we will need below data:

- List of neighbourhoods in Boston, USA

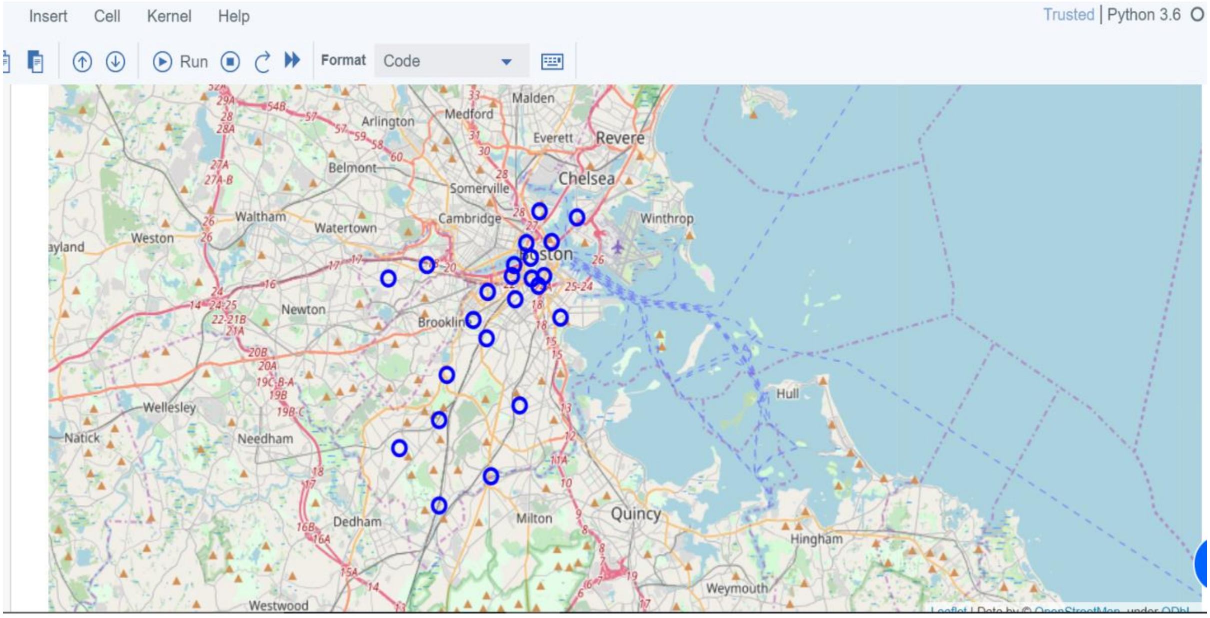
- Latitude and Longitude of these neighbourhoods
- Venue data of These Neighbourhoods.

EXTRACTING THE DATA

- Scrapping of Boston neighbourhoods via Wikipedia
- Getting Latitude and Longitude data of these neighbourhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighbourhoods

METHODOLOGY

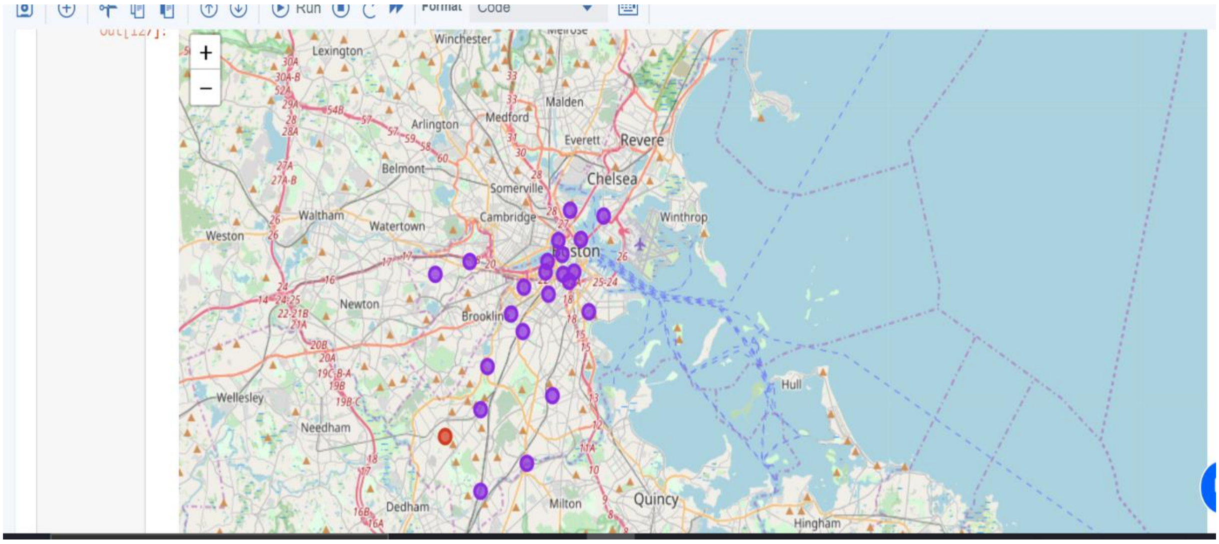
From Wikipedia https://en.wikipedia.org/wiki/Neighborhoods_in_Boston I collected the data which is required for this project. I collected the names of neighbourhoods from wiki page. And I got 23 neighbourhoods after scraping . And then using the Nominatim from geopy.geocoders package I got coordinates i.e. Latitudes and Longitudes of the Neighbourhoods. After that using four square API I collected the 100 venues in the Neighbourhoods of Boston within 600 meters of the Neighbourhood. I used the get_dummies of pandas library on category of venues .And using groupby I grouped the data frame using the neighbourhood column and found the occurrence of each category of venue in each neighbourhood. After that I wrote a function which returns 10 most common venue in each neighbourhood for our understanding purpose. And here I am going use a very famous unsupervised machine learning method which is K Means Clustering. It is an unsupervised machine learning techniques which divided the dataset into k number of clusters. Here I am clustering the neighbourhoods based on the most common places of the neighbourhood. Here’s a map of Boston before Clustering:



I used clustering for the data testing the different values of k and using the elbow test , from the elbow test it is clear that the best value of k is 2 . After that I clustered the data into 2 categories . I assigned labels for each Neighbourhood.

RESULTS:

After clustering the data using the k means clustering I divided the data into 2 clusters .As from the result I got to know that 22 neighbourhoods are similar that is they fall into same cluster, and only neighbourhood is different which is West Roxbury. And here’s the map after clustering the dataset.



Cluster 0 – All the Neighbourhoods in this cluster have mostly restaurants

Cluster 1 – in this Neighbourhoods there are mostly malls theatres and parks

DISCUSSIONS

From the result I observed that all Neighbourhoods have mostly restaurants and only 1 neighbourhood is different which is west Roxbury and which has mostly parks, malls, theatres .So All Neighbourhoods according to the most common venues are same any Neighbourhood will be fine if you are looking for good restaurants.

CONCLUSION

As all neighbourhoods are similar according to most common venue so ,I suggest my friend to live in any neighbourhood other than west Roxbury which is nearer to his Uniiversity