NEW YORK SUBWAY STATION CLUSTERING

CAPSTONE PROJECT

BUSINESS PROBLEM

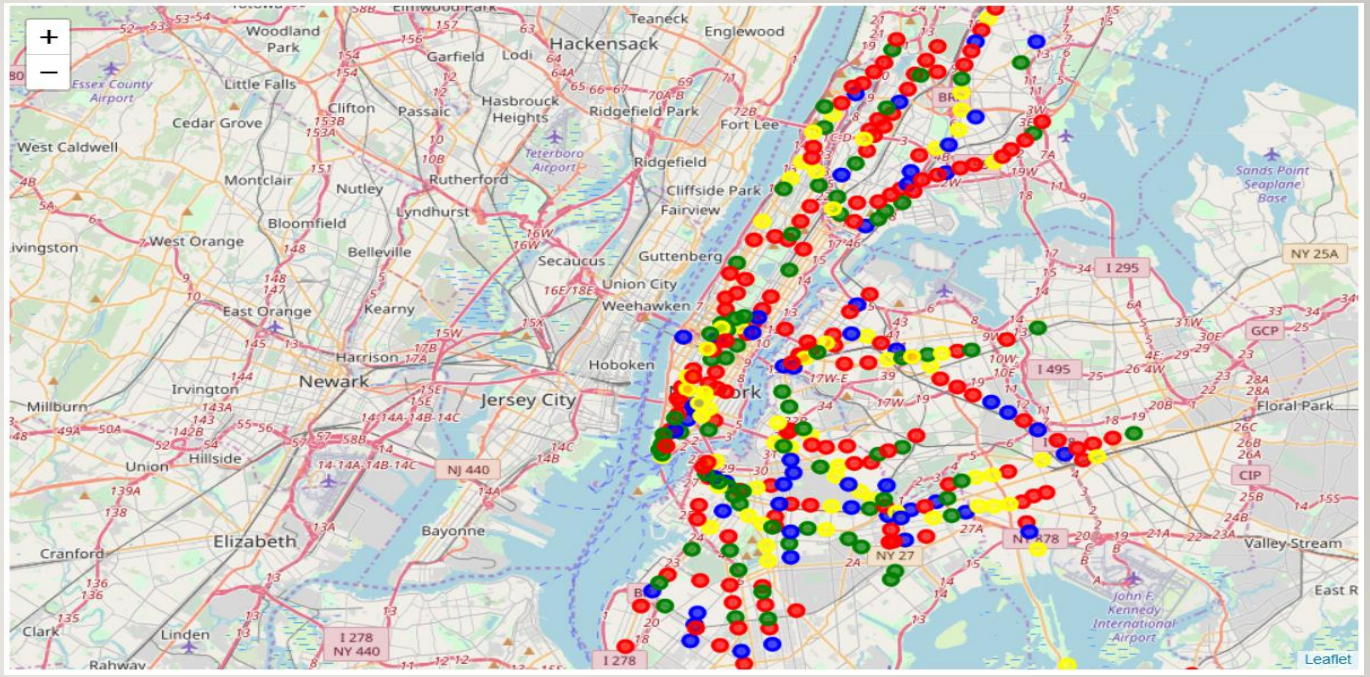
- There are several businesses/POIs around each of the subway stations in New York.
- Similar businesses might be present across stations or around the same station.
- Identifying which area to focus for starting specific business around subway station is difficult.
- A model to categorize the subway stations in different groups so as to make better business decisions

MODELLING APPROACH

- Gathered subway station information from NYC Open Data.
- Identified businesses/POI using Foursquare APIs
- As no historical data is available for training a model, used *K-Means* clustering for creating a model.
- Grouped the subway stations into four clusters based on the most common business/POI in that cluster

RESULTS

- Gathered subway station information from NYC Open Data.
- Identified businesses/POI using Foursquare APIs
- As no historical data is available for training a model, used *K-Means* clustering for creating a model.
- Grouped the subway stations into four clusters based on the most common business/POI in that cluster



CLUSTER 1

subway_cluster1.head()

Out[254]:

Frequency	1st Most Common Venue
15	Coffee Shop
14	Italian Restaurant
10	Park
9	Japanese Restaurant
5	Hotel

CLUSTER 2

subway_cluster2.head()

Out[261]:

Frequency	1st Most Common Venue
10	Discount Store
6	Caribbean Restaurant
6	Pharmacy
6	Pizza Place
5	Fried Chicken Joint

CLUSTER 3

subway_cluster3.head()

Out[262]:

Frequency	1st Most Common Venue
18	Bar
12	Mexican Restaurant
4	Coffee Shop
3	Bakery
3	Latin American Restaurant

CLUSTER 4

subway_cluster4.head()

Out[263]:

Frequency	1st Most Common Venue
19	Pizza Place
5	Caribbean Restaurant
4	Café
3	Bar
3	Coffee Shop

CONCLUSION

- Using *K-Means* clustering, the stations have been grouped into meaningful categories
- The categorization should help business investors to identify potential zones for their businesses.
- Model can be fine tuned to improve the grouping.