New York Subway Station Analysis (Capstone Project)

Business Problem

New York has a well-established subway system that runs through the heart of the city. There are several businesses and places of public interest around each of the subway stations. It is possible that there are multiple options of similar business around specific station area.

This could mean one of the following:

1. There is an opportunity to open businesses that are different in nature.
2. Due to the dynamics of the area, only similar business will flourish and more of the same could be opened.

We need to have a categorization of the top businesses & POIs around different subway stations so it is easier to make an educated determination on what kind of business will work around a specific station.

Value of the analysis

The analysis will provide insight into the top businesses or POIs around each subway station. This data can be used by different groups of people namely:

1. Marketing firms - to know which area has more concentration of a specific product business
2. Building constructors - understand which areas has the need for specific building types.
3. Business investors - they will get insight into what type of business will work in specific areas

The analysis is carried out focusing on the New York subway system. However, the solution is extendable to any other city to create a similar outcome.

Data Section

In this section we will detail out the data needs to perform this analysis.

Data Requirements

In order to perform the analysis, we will be requiring at least the following information:

1. Details of subway stations in New York

2. Geo location coordinates of the subway stations
3. Businesses or POI close to each subway station
4. Details of top businesses or POI

Data Collection

Now that we have identified what data is required, we need to understand how to obtain this data.

1. New York City Open Data (NYC Open Data) website hosts data of all subway stations, their coordinates, route schedule etc. The link for this is https://data.cityofnewyork.us/api/views/kk4q-3rt2/rows.csv?accessType=DOWNLOAD. This is a csv file that can be downloaded from the website.
2. Foursquare Labs store information about businesses and POI around each subway station. We can use Foursquare Venues APIs to gather this relevant information.
3. Identification of top businesses & POI will be part of the modelling process and is not fed from any external source.

Data Understanding & Preparation

The data obtained from NYC Open Data website shows the geo location relation of each subway station. This will help us to map these stations in a map. The geo location information is required by Foursquare APIs to provide relevant business & POI details.

The data obtained from NYC Open data website includes information that are not relevant to this analysis - like station schedule. This will go through a series of data cleanup and data massaging prior to getting used for analysis. Detailed steps involved in massaging the data is described during processing.

Methodology

Python will be the programming language used to perform analysis on data. The code will be edited using Jupyter notebooks in IBM Watson Studio environment.

The data from NYC Open Data website will be loaded to a *pandas* dataframe to perform subsequent analysis. The dataset has 355 subway stations.

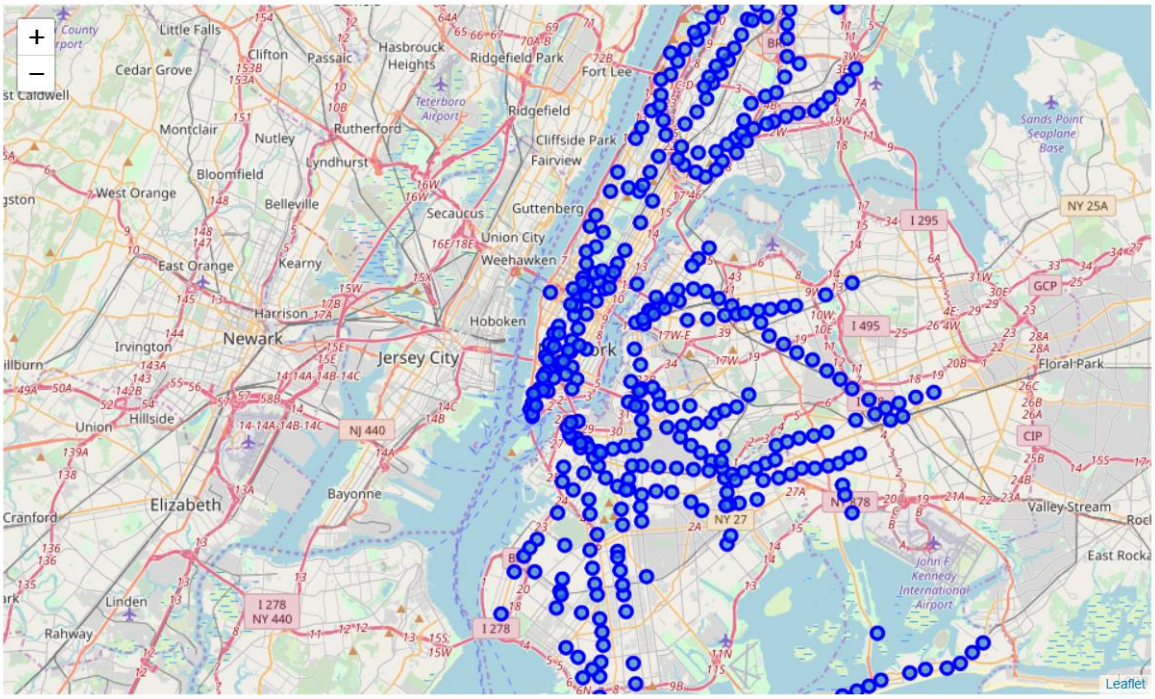
```
subway_df.shape
```

```
Out[28]: (355, 3)
```

Let us first map these stations in a map to see how they look. Each blue spot on the map represent a separate subway station

```
map_newyork
```

Out[31]:



Our objective is to list out the businesses & POI around these 355 stations and try to come up with a categorization. Using the Foursquare API, the nearby POI & businesses are listed out for each subway station based on its geo location coordinates.

Considering the number of subway stations and their closeness to each other in the map, we will get only 10 nearby locations. Also, the analysis here is carried out for all of New York city; however, we could do the same exercise for stations in a specific area. This specific analysis is not in scope of this process.

The Foursquare API returned 318 unique location categories. In order for us to derive a conclusion on the density factor of any POI or business, these 318 unique categories should be broken down.

```
#Check how many venues are there and print the unique categories()
print("Size of subway_venues is: ", subway_venues.shape)
print('There are {} uniques categories.'.format(len(subway_venues['Venue Category'].unique())))
```

```
Size of subway_venues is:  (3523, 7)
There are 318 uniques categories.
```

As a first step, we will create a cross reference between the subway station and the POI using a True/False or 1/0 grid. Then this information is grouped by subway station name so as to identify the Top 3 POI locations for every subway station.

```
neighborhoods_venues_sorted.head()
```

Out[247]:

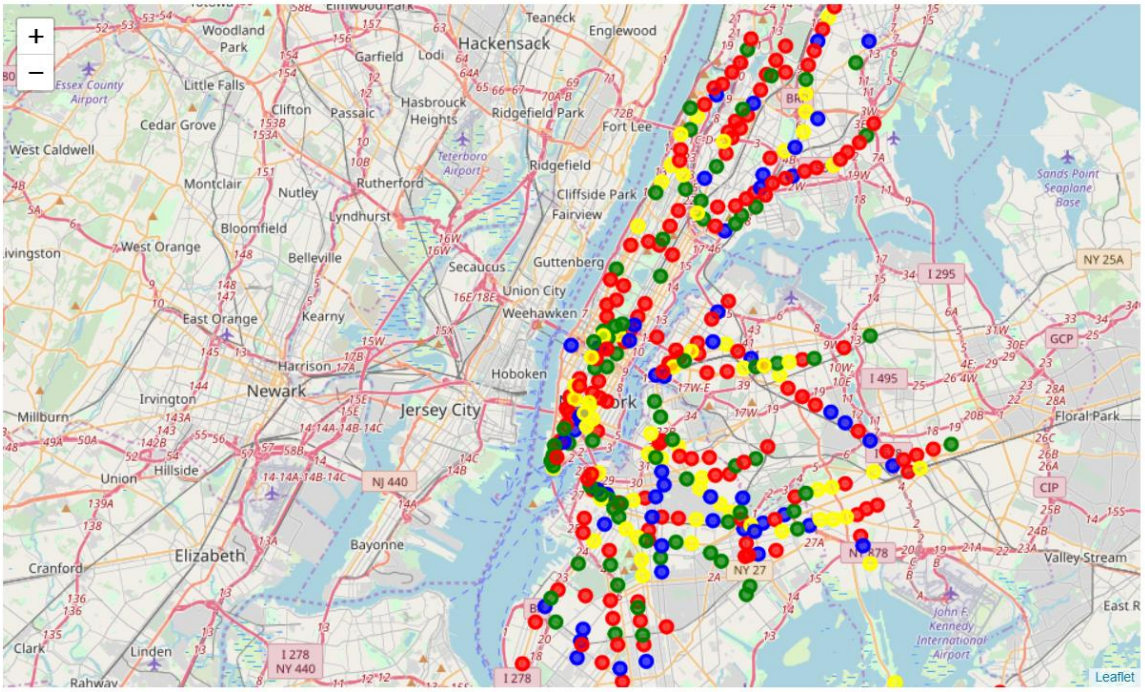
	Subway	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	103rd St	Pizza Place	Yoga Studio	Ice Cream Shop
1	103rd St - Corona Plaza	Latin American Restaurant	Coffee Shop	Deli / Bodega
2	104th St	Pharmacy	Pizza Place	Discount Store
3	104th-102nd Sts	Deli / Bodega	Metro Station	Ice Cream Shop
4	110th St	Steakhouse	Latin American Restaurant	Pet Store

At this point there will be several POIs that are the topmost common places. We do not have any specific historical data to look back and create a model that will tell us which POI combination can be grouped together. Therefore, we will use *K-Means clustering* which is an unsupervised model.

Using an initial K value of 4, let us cluster the data and show that in the map.

```
map_clusters
```

Out[260]:



We now have view of how the different clusters are spread across in the map. We still need to determine what these clusters mean. Let us look at the count of each POI in the cluster to see which ones top the list. We can determine the cluster composition based on that.

Result

Using *K-Means* we successfully clustered the data into four separate clusters. In order for us to derive a meaningful conclusion out of the clustered data, following steps were performed:

- Group the data by the *1st Most Common Venue* and find the count
- Remove all columns except the Venue name to see how many similar venues in that cluster
- Sort the data in descending order of count
- Pick the top 3 venues to determine how that cluster should be categorized.

For each cluster, following are the results:

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

This cluster mainly has *Coffee Shop, Italian Restaurants & Parks*. Let's call this as *Zone 1* and is represented in green on the map.

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

The output clearly indicates this is a *Discount Store* zone. Let's call this as *Zone 2*. This is represented with blue on the map.

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

This zone's top-ranking POIs are *Bar & Mexican Restaurant.* This will be *Zone 3* represented in yellow on the map.

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

Pizza place is the most common POI for this *Zone 4* which is represent in red on the map.

Discussion

Based on the analysis done above, it can be concluded that certain POIs are found in larger numbers closer to specific subway stations than others. The model used here is built using 4 clusters. It is possible to expand on the number of clusters to further break down on the categorization.

Should we have done more than 4 clusters – for the purpose of this exercise, we should be good with 4; however, there are instances where it could be misleading.

Coffee Shop features in the top 5 of three of the clusters. How can we be sure of a business opportunity with coffee shop in those zones? In Zone 4, it looks like *Pizza place* outranks the others. It shows an abundance of pizza joints in that zone. It is better not to start a pizza business in that zone.

Conclusion

Using clustering we were able to group the POIs together and provide a view of the business expansion opportunities around subway stations. The model can be run with more clusters to further refine the grouping.

The final result will address the business problem that was stated initially. The grouping will help users to understand which zone they need to focus for their business opportunities.