Predict the Location of New Amazon Warehouse in Toronto Using Location Data and K-Means Clustering Techniques

This report serves the purposes of using only Geographic information and the K-Means Clustering Machine Learning method to predict the hypothetical new location of a logistic warehouse for Amazon in Toronto, using the data of Manhattan, NY.

The entire report has been divided into the following sections: the summary for the overall analysis, technical details, predictive result, and conclusion. In the technical details section, I will break down the analysis into each step and showing the result as screenshots. In the predictive result section, I will post the result of this analysis as a screenshot, and, in the conclusion, I will state how I arrive at the conclusion and bring up a further recommendation to this analysis.

Analysis Summary

Geographic location data typically refers to latitude and longitude data of a particular location. Depends on the data source, we can also get other relevant information, such as name, business type, reviews, etc., as well. In this case, we are using data provided by Foursquare to analyze how the existing Amazon warehouse in Midtown, Manhattan, NYC is related to other venues around and using the existing relationship to predict a "new" Amazon warehouse in Toronto. The entire analysis will be done by using Jupyter Notebook and the language environment in Python. The packages in Python we will use are listed as following: Numpy, Pandas, Beautifulsoup, XLRD, LXML, Geopy, Sklearn.cluster, and Folium.

The overall process is: first we will import the zip code and coordinate data of Toronto from Wikipedia using the API function in Python and form a data frame. Then we will import the coordinate and location data from Foursquare of Manhattan, use K-Mean Clustering to separate each location into the cluster, and analyze the relationship between each cluster with the Amazon warehouse. Finally, we are going to apply the same procedure to the data of Toronto, except that we will come up with an estimated warehouse location using the result observed from analyzing the Manhattan data.

Technical Details

1. **Data Import**

For the data of Toronto, since it is a table in an HTML page, I utilized the "BeautifulSoup" package in Python to extract the table into Python and loaded it into a data frame. Later I download a CSV file that contains the coordinates data and join with the location data.
For the data of Manhattan, I utilized API in Python to download a JSON file that containing all relevant geolocation information and transformed the data into a data frame.

2. **Data Cleaning**

For the data of Toronto, I use the procedure below to get rid of irrelevant information and to fix the format.

The code is shown below:

```
wiki_df = wiki_df.replace('\n','', regex=True)
wiki_df = wiki_df[wiki_df.Borough != 'Not assigned']
wiki_df.groupby('Postal Code')['Borough','Neighbourhood'].agg(','.join).reset_index()
```

Note: wiki_df is the data frame that contains the location data of Toronto. It contains three columns: Postal Code, Borough, and Neighbourhood.

	Postal Code	Borough	Neighbourhood
0	M1A\n	Not assigned\n	Not assigned\n
1	M2A\n	Not assigned\n	Not assigned\n
2	M3A\n	North York\n	Parkwoods\n
3	M4A\n	North York\n	Victoria Village\n
4	M5A\n	Downtown Toronto\n	Regent Park, Harbourfront\n

Figure 1. Before Cleaning the Data

	Postal Code	Borough	Neighbourhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Figure 2. After Cleaning the Data

For the data of Manhattan, the data is in very good shape and no further cleaning action has been performed.

3. **Data Analysis**

The analysis is separated into three steps (Note: We take the example of the data of Manhattan since the procedure will be the same in analyzing the data of Toronto):

Step 1: Since the data is categorical, to analyze it, we need to transform the data into numerical and reload the transformed data into a new data frame. We used One Hot Encoding, which use 1 or 0 to represent if a given location has a particular or not.

The code is shown below:

```
manhattan_onehot = pd.get_dummies(manhattan_venues[['Venue Category']], prefix="",
prefix_sep="")
manhattan_onehot['Neighborhood'] = manhattan_venues['Neighborhood']
```

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Video Store	Vietnamese Restaurant	Volleyball Court
0	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0
1	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0
2	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0
3	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0
4	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0

Figure 3. One Hot Encoding

Step 2: We first grouped the locations by Neighborhood and calculate the mean of the present frequency of each venue in each Neighborhood and load it into a new data frame.

The code is shown below:

```
manhattan_grouped = manhattan_onehot.groupby('Neighborhood').mean().reset_index()
```

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Video Store	Vietnamese Restaurant	Volleyball Court
0	Battery Park City	0.0	0.0	0.0	0.000000	0.000000	0.0	0.00	0.000000	0.000000	...	0.0	0.000000	0.0
1	Carnegie Hill	0.0	0.0	0.0	0.000000	0.011111	0.0	0.00	0.011111	0.000000	...	0.0	0.011111	0.0
2	Central Harlem	0.0	0.0	0.0	0.065217	0.043478	0.0	0.00	0.000000	0.021739	...	0.0	0.000000	0.0
3	Chelsea	0.0	0.0	0.0	0.000000	0.030000	0.0	0.01	0.000000	0.060000	...	0.0	0.000000	0.0
4	Chinatown	0.0	0.0	0.0	0.000000	0.030000	0.0	0.00	0.000000	0.000000	...	0.0	0.020000	0.0

Figure 4. Mean of Presence Frequency

Step 3: We first defined a function to sort the Venues in descending order, according to the mean we obtained in the previous step. After this, we filtered out the top 10 venues in each neighborhood and loaded them into a new data frame, and display it.

The result is shown below:

	Neighborhood	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	Coffee Shop	Park	Hotel	Clothing Store	Gym	Memorial Site	Playground	Plaza	Shopping Mall	Burger Joint
1	Carnegie Hill	Coffee Shop	Café	Yoga Studio	Bookstore	French Restaurant	Wine Shop	Cosmetics Shop	Pizza Place	Gym	Gym / Fitness Center
2	Central Harlem	Cosmetics Shop	African Restaurant	Chinese Restaurant	Seafood Restaurant	American Restaurant	French Restaurant	Bar	Southern / Soul Food Restaurant	Caribbean Restaurant	Library
3	Chelsea	Coffee Shop	Art Gallery	Bakery	French Restaurant	American Restaurant	Wine Shop	Ice Cream Shop	Seafood Restaurant	Hotel	Park
4	Chinatown	Chinese Restaurant	Bakery	Cocktail Bar	Ice Cream Shop	Bubble Tea Shop	Optical Shop	American Restaurant	Hotpot Restaurant	Spa	Salon / Barbershop

Figure 5. Top 10 Venues in Each Neighborhood

4. Clustering and Visualization

To group Neighborhood into similar groups, we will use the K-Means-Clustering technique, which is an unsupervised machine learning technique. It will group data into separate clusters and make sure that the data within each group has maximum similarity in terms of a particular feature, while the data outside of the cluster is drastically different.

Step 1: We first create five empty clusters and run K-Means Clustering to the data, using the mean data we obtained.

The code is shown below:

```
from sklearn.metrics.pairwise import euclidean_distances
nyc_kclusters = 5

manhattan_grouped_clustering = manhattan_grouped.drop('Neighborhood', 1)
nyc_kmeans = KMeans(n_clusters=nyc_kclusters,
random_state=0).fit(manhattan_grouped_clustering)
```

Step 2: We merged the clustered data with other data.

The result is shown below:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	2	Gym	Discount Store	Sandwich Place	Coffee Shop	Yoga Studio	Pizza Place	Supplement Shop	Steakhouse	Shopping Mall
1	Manhattan	Chinatown	40.715618	-73.994279	0	Chinese Restaurant	Bakery	Cocktail Bar	Ice Cream Shop	Bubble Tea Shop	Optical Shop	American Restaurant	Hotpot Restaurant	Spa
2	Manhattan	Washington Heights	40.851903	-73.936900	3	Café	Bakery	Grocery Store	Spanish Restaurant	Chinese Restaurant	Sandwich Place	Tapas Restaurant	Italian Restaurant	Mobile Phone Shop
3	Manhattan	Inwood	40.867684	-73.921210	3	Mexican Restaurant	Café	Restaurant	Lounge	Pharmacy	Spanish Restaurant	Caribbean Restaurant	Chinese Restaurant	Frozen Yogurt Shop
4	Manhattan	Hamilton Heights	40.823604	-73.949688	3	Pizza Place	Café	Coffee Shop	Mexican Restaurant	Deli / Bodega	Yoga Studio	Sushi Restaurant	Caribbean Restaurant	School

Figure 6. Top 10 Venues in Each Neighborhood with Cluster Label

Step 3: We created a visualization map using the Folium function in Python to visualize the location of each cluster as well as the existing Amazon warehouse location.

The result is shown below:

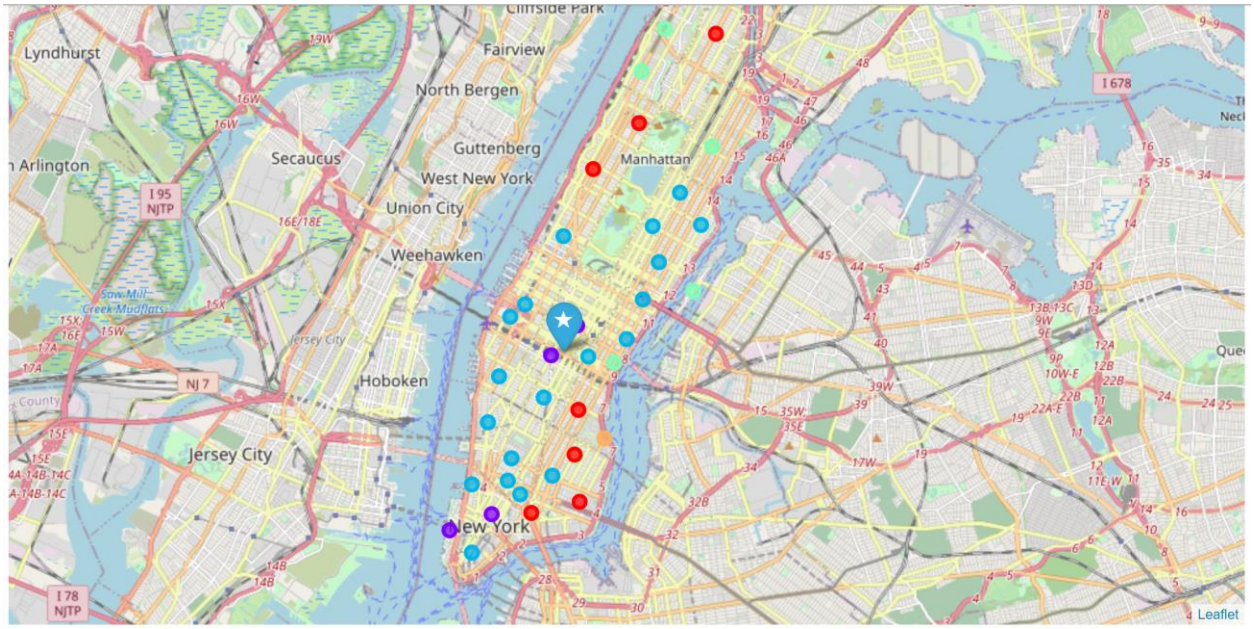

Figure 7. Map Showing the Location of Each Neighborhood in Dots and Amazon Warehouse in Waterdrop Marker

Predictive Result

The Amazon warehouse in Manhattan is located at Midtown, at the intersection of 5th Ave and 34th Street. As the analysis suggests, the warehouse has the shortest distance, 303 meters, to the Midtown South, which belongs to Cluster 1. If we examine the Top 10 venues at Midtown South, we can find that most of the venues are a restaurant, as shown below:

	Neighborhood	1th Most Common Venue	2th Most Common Venue	3th Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
15	Midtown	Hotel	Clothing Store	Coffee Shop	Theater	Gym	Café	Sporting Goods Shop	Bookstore	Bakery	Steakhouse
23	Soho	Clothing Store	Italian Restaurant	Coffee Shop	Boutique	Mediterranean Restaurant	Sporting Goods Shop	Café	Salon / Barbershop	Bakery	Hotel
28	Battery Park City	Coffee Shop	Park	Hotel	Clothing Store	Gym	Memorial Site	Playground	Plaza	Shopping Mall	Burger Joint
33	Midtown South	Korean Restaurant	Hotel	Japanese Restaurant	Dessert Shop	American Restaurant	Hotel Bar	Gym / Fitness Center	Bakery	Burger Joint	Coffee Shop

Figure 8. Examine the Top 10 Venues in the Cluster 2 Neighborhood

If we examine the other two clusters, Murray Hill and Midtown, which are also close to the Amazon warehouse, we can find that the majority of venues are restaurants as well, as shown below:

16	Murray Hill	Sandwich Place	Coffee Shop	Bar	Hotel	American Restaurant	Japanese Restaurant	Pizza Place	Burger Joint	Gym / Fitness Center	Gym

Figure 9. Examine the Top 10 Venues in the Murray Hill

15	Midtown	Hotel	Clothing Store	Coffee Shop	Theater	Gym	Café	Sporting Goods Shop	Bookstore	Bakery	Steakhouse

Figure 10. Examine the Top 10 Venues in the Midtown

In the data of Toronto, as shown below, by directly comparing with the Midtown South Cluster 1 in Manhattan, the closet area will be Downtown Toronto, represented by purple dots at the center of Toronto. The Top 10 venues in this area are restaurants or Café, which occupy 8 spots out of ten.

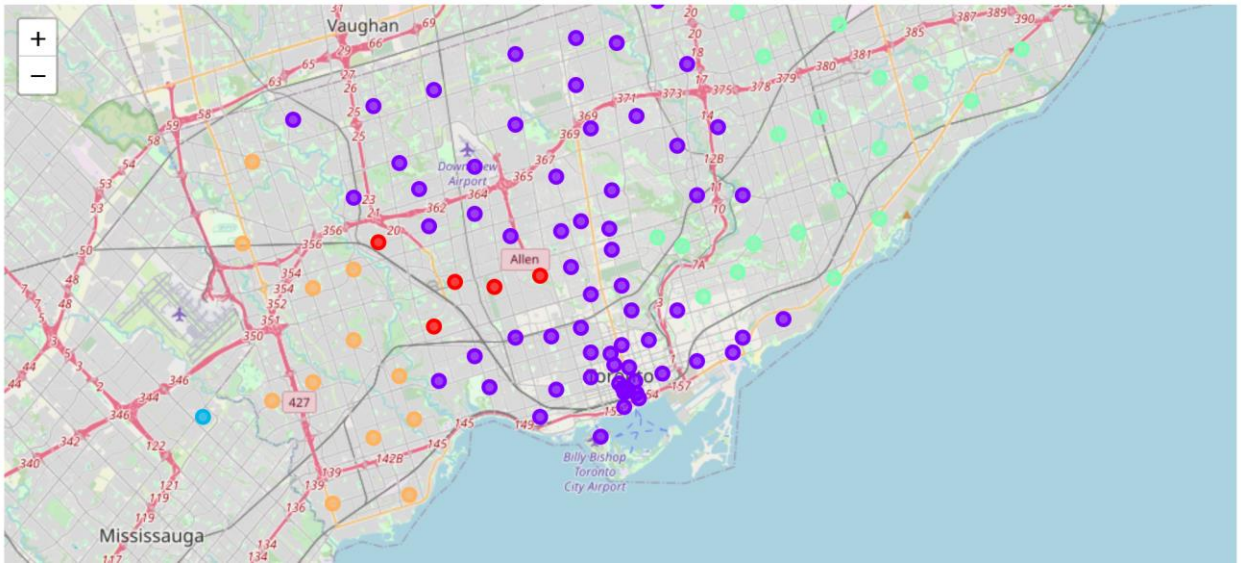

Figure 11. The Clusters in the Toronto

Therefore, without any other information, from this analysis we concluded that the new Amazon warehouse should probably be built in the Downtown Toronto area, surrounded by all kinds of restaurants like the one in Manhattan did.

However, after we researched the real Amazon warehouse locations, we find that this conclusion is not grounded. There are four Amazon warehouses in Toronto, one at the North Toronto, which is the purple dot beside the airport in Figure 11, two at the east Toronto, which at the intersection between purple dots and the light green dots (Cluster 3), the last one at the West Toronto, represented by the red dots, which are Cluster 1. **This suggests that analyzing the location data alone to predict the location of the logistic warehouse is insufficient; and therefore, the prediction from our model is faulty**.

Conclusion and Recommendation

By observing how our analysis deviated from reality, we can think of two possible causes behind this: the first is our model is incomplete and the feature we used to analyze is rather irrelevant.

Another cause is that we cannot directly compare Toronto with Manhattan as they are two different cities in terms of population density, area, and logistic infrastructure. To better enhance this analysis, there two suggestions that we can take into consideration: first, collect more relevant data, for example, we should consider rent price so that after we narrowed down to a specific area, we can use a rent price as another feature to select an optimum location; second, find a city that shares the similar condition with Toronto and run the analysis again.