Applied Data Science Capstone

January 18, 2019

1 Applied Data Science Capstone by Manisha

1.1 Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods (Week 1)

2 Background / Introduction and Business Problem Statement:

2.1 Client - An Indian IT company planning to start up their overseas branch in New York.

Requirement-

The business demand of the IT services and solutions given to the various companies in USA is on rise. Due to new rule the US companies are not able to hire / get on-site the skilled Indian professionals. Thus this company wants to start their subsidiary in US to take care of customer requirements. They have to decide on the location of their new office in New York to open their office. We have been approached to do the study and suggest them an ideal location. Business Problem - As initially they will be sending their Senior Management team from India to establish the subsidy, the concern from the company is to have least issue in settling for these people so that they work can start immediately and, if they needed can stay in on-site location for a longer time. As informed, is the team has to move for a duration of 2-3 years minimum for the subsidy to get stable then following parameters needs to be considered.

1. Population rate

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Alt text

2. Median house price
3. Eating Joints – Vegetarian joints availability is must.
4. Schools / Educational institutes ratings
5. Weather conditions.
6. Crime rates
7. Recreational facilities

This project helps the end user or the stakeholder to achieve the results which will not only recommend but also saves a lot of time in manual search. This will indeed save the time and money of the user.This projects core objective is to study the Neighborhoods with respect to above parameters and provide a detailed analysis to users which can help them decide the Final location of their new office.

This project can be used by the user at the time of renting apartment or buying house in a locality based on the distribution of various facilities available around the neighborhood.This project would compare 2 randomly picked neighborhoods and analyses the top 10 most common venues in each of those two neighborhoods based on the number of visits by people in each of those places. Also, this project uses K-mean clustering unsupervised machine learning algorithm to cluster the venues based on the place category such as restaurants, park, coffee shop, gym, clubs etc. This would give a better understanding of the similarities and dissimilarities between the two chosen neighborhoods to retrieve more insights and to conclude with ease which neighborhood wins over other.

2.1.1 *Alternatively the problem statement can be captured by the Question below*

"How can IT companies decide the location of their new subsidiary to help their overseas staff to choose between which neighbourhoods in NY to live during their 2-3 years stay?"

2.1.2 Trusted Data Sources for the Project ::

The following trusted data sources will be used to conduct Data exploration and other forms of data analysis as part of the methodology for the project. 1. NY Neighbourhood Data – https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json & https://geo.nyu.edu/catalog/nyu_2451_34572

2. Foursquare Location Data This API has a database of more than 105 million places. This project would use Four-square API as its prime data gathering source. Many organizations

:

	name	categories	lat	lng
0	Lollipops Gelato	Dessert Shop	40.894123	-73.845892
1	Rite Aid	Pharmacy	40.896521	-73.844680
2	Carvel Ice Cream	Ice Cream Shop	40.890487	-73.848568
3	Dunkin Donuts	Donut Shop	40.890631	-73.849027
4	SUBWAY	Sandwich Place	40.890656	-73.849192

Alt text

are using to geo-tag their photos with detailed info about a destination, while also serving up contextually relevant locations for those who are searching for a place to eat, drink or explore. This API provides the ability to perform location search, location sharing and details about a business. Foursquare users can also use photos, tips and reviews in many productive ways to add value to the results. (Venues [Restaurants, Community Centres,], TOP tips, Favourites, User Experience, etc.), Educational institutes, Metro (distance) will be used to cluster, segment, target, and position to craft recommendations for the Indian end-user community.

3. Folium- Python visualization library would be used to visualize the neighborhoods cluster distribution of NY city over an interactive leaflet map. Using the data available in the above 3 trusted sources, we will be conducting clustering and neighbourhood based analysis leveraging primarily Foursquare APIs and tools such as KNN and relevant Unsupervised machine learning algorithm K-mean clustering would be applied to form the clusters of different categories of places residing in and around the neighborhoods. Extensive comparative analysis of two randomly picked neighborhoods world be carried out to derive the desirable insights from the outcomes using python's scientific libraries Pandas, NumPy and Scikit-learn. . These clusters from each of those two chosen neighborhoods would be analyzed individually collectively and comparatively to derive the conclusions. Based on this options to the target user community primarily comprising of the Indian Executive's

2.1.3 Python packages and Dependencies:

Pandas - Library for Data Analysis NumPy – Library to handle data in a vectorized manner JSON – Library to handle JSON files Geopy – To retrieve Location Data Requests – Library to handle http requests Matplotlib – Python Plotting Module Sklearn – Python machine learning Library Folium – Map rendering Library Beautifulsoup – Web scraping and data wrangling

Let me also provide below the specific steps that will be deployed to work with the data derived from Foursqaure and the K means approach . We will cover neighborhoods of NY and also the category of "Population Distribution analysis, Median House Price Analysis, School Ratings"

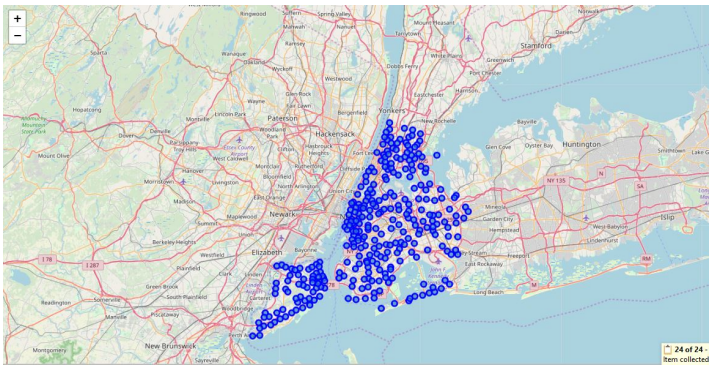
HTTP requests would be made to this Foursquare API server using zip codes of the Newyork city neighborhoods to pull the location information (Latitude and Longitude).Foursquare API search feature would be enabled to collect the nearby places of the neighborhoods. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to

100 and the radius parameter would be set to 500.Leveraging the recommender systems capabilities enabled by Foursquare to conduct compare / contrast the data points to derive the recommendations.

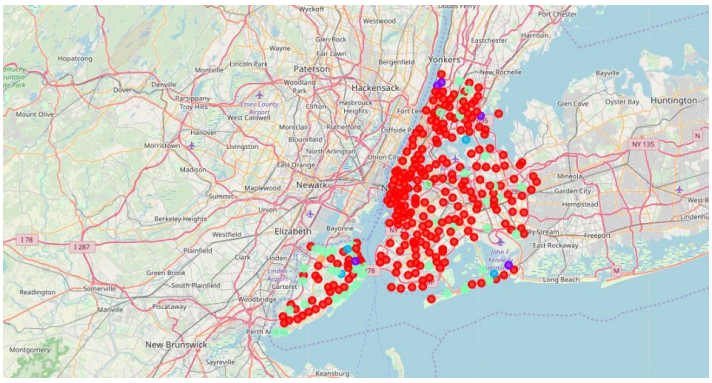
2.1.4 Methodology ::

The methodology employed is aligned with the Data Science 10 steps program that was discussed during an earlier module of the course.

A. The workflow of the project starts with the web scraping and data wrangling. Using the Beautiful Soup library, the postal code and the neighborhood the data is processed to derive the latitude and longitude of the NY neighborhood.
B. With the folium Map, the latitude and longitude of the NY neighborhood provides the choropleth visualization.



C. Obtain relevant places data from Foursquare and clean it for data understanding and grouping etc.
D. Explore the data for clusters & patterns of Neighborhoods in Boroughs.



E. Group the places of high interest into relevant neighborhood and Borough pairings.

Let's confirm the new size

```
NY_grouped.shape
```

```
(301, 431)
```

Let's print each neighborhood along with the top 10 most common venues

```
num_top_venues = 10

for hood in NY_grouped['Neighborhood']:
    print("----"+hood+"----")
    temp = NY_grouped[NY_grouped['Neighborhood'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

F. Generate the markers on NY & maps to highlight the neighborhoods that are in focus.
G. Use Foursquare venue, Categories etc to enable the profiling the primary & focus neighborhoods.
H. FourSquare API and K-means clustering methods are used to retrive the top trend venues of the NY neighborhood.

In this project , decision of buying , rental or setting up market is getting recommended based on the clustered neighborhoods, Population Distribution analysis, Median House Price Analysis, School Ratings.

2.1.5 Results:

With the help of above methodologies, the project can easily help the user to decide which neighbourhood is better to stay based on the factors of number of population (including Indians), localities, schools rating of the particular neighborhood and availability of top trend venues.

2.1.6 Discussion:

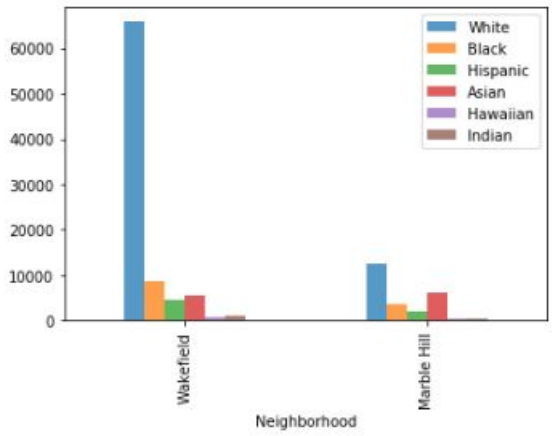
This project is beneficial in many terms, it will save the users time and money. Whenever people are moving to new location they have an anxiety . This project recommend the better places in the intended location of transfer with a very less effort.

2.1.7 Conclusion:

This Analysis concludes that the two places of New York Wakefield , Marble Hill both has great amenities and locality, but out of these two Wakefield has better prospects for buying houses or choose for rental houses. Wakefield has the higher number of Indian population ,good school rating of 9 and a reasonable avg housing price of around 172k , also top 10 common venues shows Wakefield has got a good neighborhood with Laundromat, Pharmacy, Food Truck, Clothing Store , Pizza Place, Donut Shop and many more. Hence Wakefield wins over Marble Hill!

Next comes the critical step of choosing the methods (Unsupervised or Supervised and within each of them the specific methods for machine learning / model dev't and training etc) for data analysis/mining/patterns recognition etc...

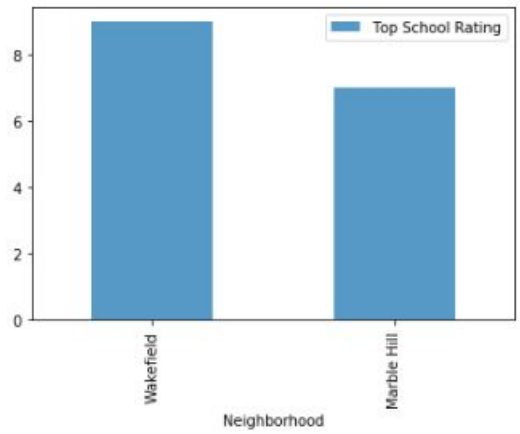
It is also vital that relevant presentation and visualization techniques are used to share and present the findings to the project sponsor / stakeholders / end-users etc. It may be also appropriate to comment any future work that may be valuable to augment the solution and enrich




```
Population_Comparison['Indian']
```

```
Neighborhood
Wakefield      1099
Marble Hill     397
Name: Indian, dtype: int64
```

Alt text




```
]: School_rating_comparison
```

]:

Neighborhood	Top School Rating
Wakefield	9
Marble Hill	7

Alt text

31]:
```
Avg_housing_price_comparison
```

31]:

Neighborhood	Average_Housing_Price
Wakefield	172050.0
Marble Hill	225800.0

Alt text

it.

This study helps user to compare two neighborhood and recommend options with facts.

```
In [ ]:
```