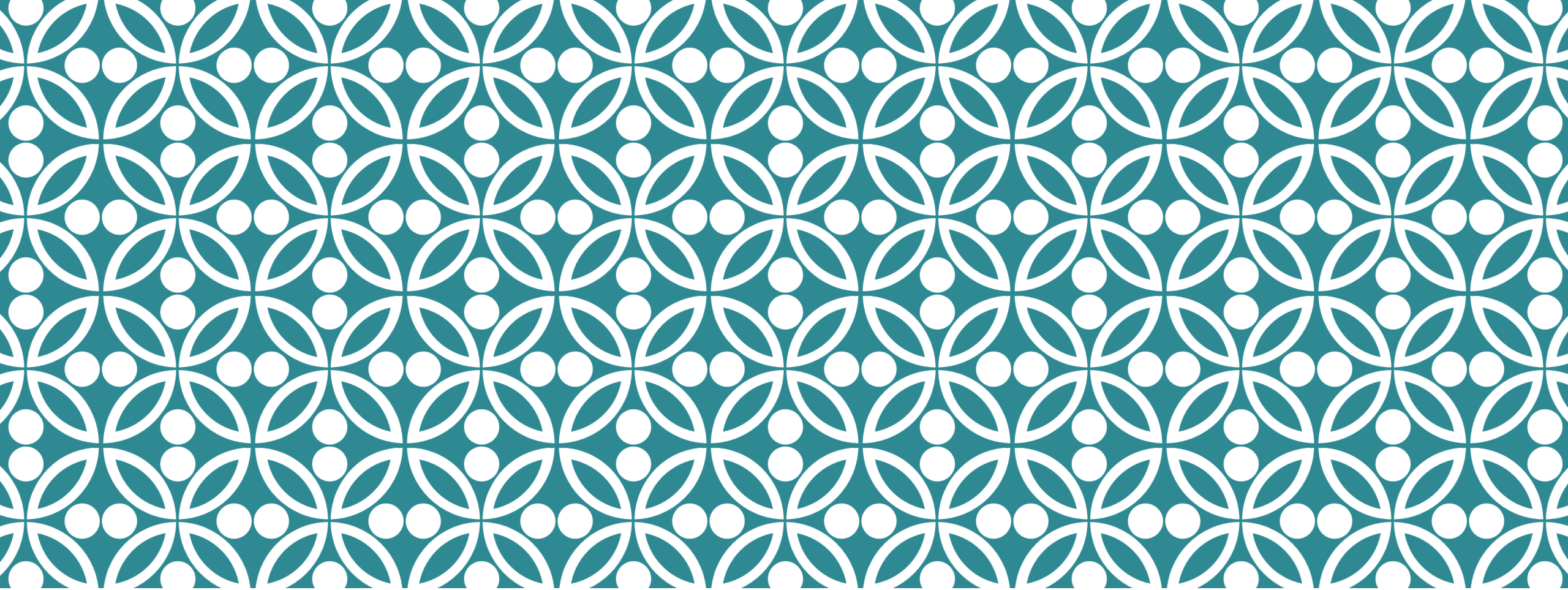


IBM ADVANCED DATA SCIENCE CAPSTONE FINAL PROJECT

Project by:-
Nanduri V P S Anirudh

INTRODUCTION

- The current situation in USA and rather the entire world is in such a way that even if someone says that the whole world is in a standstill, no one would disagree.
- However, one good thing about this time is that a lot of people have some free time on their hands where they are contemplating what can they do to better themselves.
- This also includes many people who will probably upskill themselves and would be looking for a better job when the situation around the world settles a little.
- It also includes quite a few businessmen, and new entrepreneurs who either would like to set up a new company or expand an existing company.
- Real estate agents even if experienced in the field would still need to back up their claims for best suitable option for their clients with some real data.
- In this capstone project, the aim is to help such real estate agents to figure out the types of neighborhoods and then recommend them neighborhoods best suited for their client's needs.

BUSINESS PROBLEM

- The main business problem is to help real estate agents suggest their client's ideal or close to ideal locations/places in accordance to their requirements.
- Some of the questions which can be solved after going through the analysis are "suggest a good place to start my new office", "suggest me a place to move in near my office", etc.

DATA

Data required

- The data needed for this analysis would be information of neighborhoods in New York City, NY with information of their borough and latitude and longitude.
- We would also be needing information of venues in the neighborhood and which type of venues they are.

Data sources

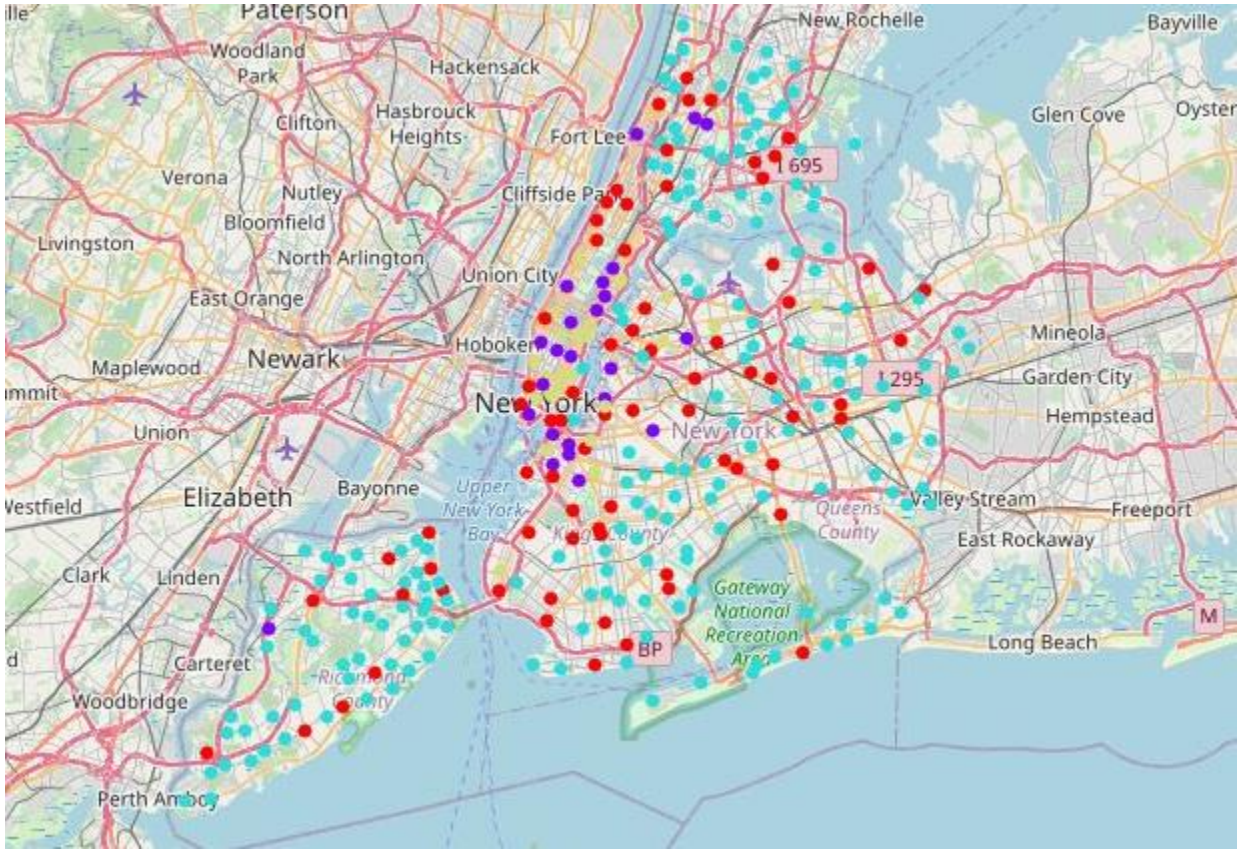
- There are mainly two data required as stated above, one is data containing borough, neighborhood, latitude, and longitude which is taken from json file available here and the information regarding venues will be extracted using Foursquare API while passing required inputs into API.

METHODOLOGY

- Downloading the json file and extract the information of borough, neighborhoods, and their geographic locations.
- Use Foursquare API to fetch venue data which consists of name, category, and their geographic locations.
- Group the categories according to their nature into smaller buckets.
- Group data by neighborhood and taking sum of one-hot encoding of category groups.
- Perform clustering on the data using K-Means clustering.
- Visualizing and understanding clusters based on output of K-Means algorithm.

RESULTS

- The result of the clustering shows that all neighborhoods are classified as one of 4 clusters. This is given in below image which visually depicts the neighborhoods and their assigned clusters.



DISCUSSION

- Neighborhoods in cluster 0 have good health services, transportation, and work spaces near them. Could be considered as a decent place to live in for people who do not care that much about fun and relaxation.
- Neighborhoods in cluster 1 have most services being offered as well as most educational places when compared to other clusters while having a good number of work spaces, hotels, and parks. This would definitely be a good candidate for the customers who can spend a little for their accommodation as the neighborhoods in this cluster are tend to be expensive considering that all factors are available here. This place is also good for people who want to expand business as this place clearly has a good footfall.
- Neighborhoods in cluster 2 seem to be behind all other neighborhoods except for the fact that it has a good transportation facility. It is probably a place which can be developed later and might be a good place for future investment or real estate buys focused on generating rents.
- Neighborhoods in cluster 3 have the most hotels, shops, and pubs. This place is probably more ideal for setting up shops rather than living

FUTURE SCOPE

- Future scope for the project can be to include the real estate pricing data along with the already existing feature set to create new clusters to see which neighborhoods are highly priced with all facilities available and how do neighborhoods which are relatively cheap compare to highly priced ones.
- Another way this can be achieved is by doing a regression model which will quantify the effect of each feature onto the real estate price.
- This will greatly help real estate agents in finding best neighborhood based on the client's choice and budget.

CONCLUSION

- In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 4 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. real estate agents regarding the best neighborhoods to reside in or start a new office space.
- To answer the business question that was raised in the introduction section, the answer proposed by this project is: neighborhoods in cluster 1 are best candidates for both setting up a new office as well as living. Only probable down side in cluster 1 would be possible high rates as most of services and requirements are available.
- The next best place to set up office is in neighborhoods of cluster 3. The second-best place to live is in neighborhoods of cluster 0. It is highly likely that the prices here aren't that high as was the case in neighborhoods of cluster 1.