COURSERA FINAL PROJECT - REPORT

(Clustering and Segmentation)

1.INTRODUCTION

A) BACKGROUND

In today's ever changing world, we humans are exploring new ventures and ways to approach our jobs with new innovative businesses and start-ups and in such times we invest a lot of money and valuable efforts to make sure our shop or for that matter any business venture does good. Irrespective of whether it is a garage , bakery shop or restaurant . A lot is at stake if our ventures or start – ups don't make good profits.

It is statistically proven in a survey performed by the economic times in 2004, that the locality of a business is the 2nd most important feature that decides whether it will make profit or not . And hence if one could by any means get to know about the perfect place to situate their business in, with regards to a specific part of a city, then that would be enhance the probability of their business being a success.

B) THE PROBLEM

As most landlords or property dealers who show you a location or plot of land , won't be equipped with the information or might not be willing to point out the flaws of the neighbourhood and might lack any knowledge of how many similar such shops / businesses run around that region .

They wont even be able to tell whether a particular business will flourish in that neighbouring region or not and will be unabe to give any information about the traits of the people living in that neighbourhoods . They will fail to state how they're (people) outlook towards the new business in town will be .

C) **SOLUTION**

However using Data Science and Machine Learning we can solve this problem and help such businesses know anything they want to about the locality in which they plan on setting up their shop/business .I plan on making a web / mobile application for customers that are planning on opening a store/ start-up in a specific part of the town .Our target Customers will be people or businessman that intend on a start-up or extending their franchise of shops at a specific region .

This application can help them predict and decide if they should open their respective store at a specific location , taking in account the statistics of the most common venues , that people visit in that region and the nearby shops and similar ventures around the neighbourhood and how successful these shops are .

One advantage of this application is the fact that it is available to all customers and businessman irrespective of the store they want to open, it could be a garage , a hospital , mall , restaurant , salon etc

2.DATA

Source :

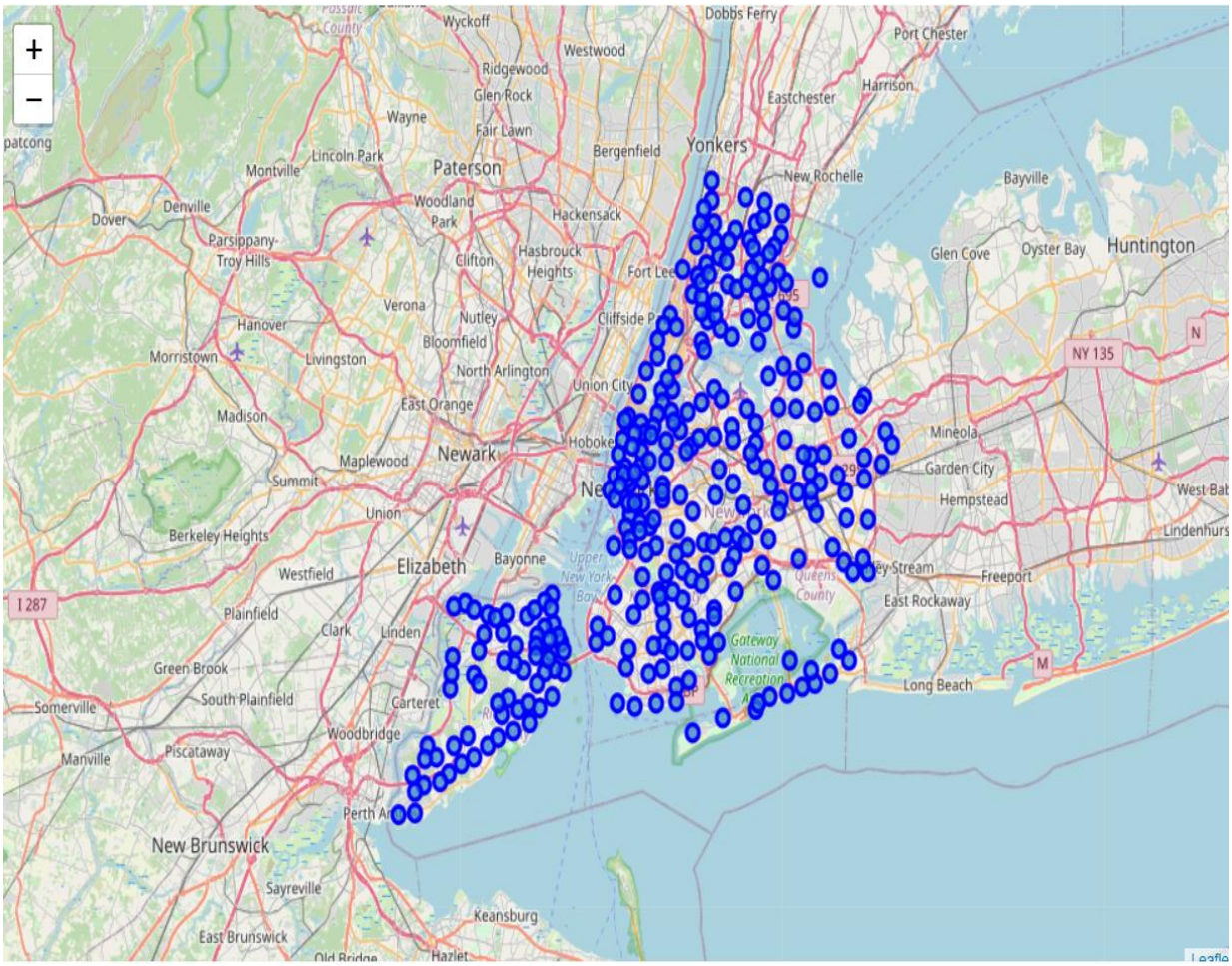
The data used for this notebook is a dummy data that we initially plan on working with first and was obtained from a json file “NYC_data.json” from www.kaggle.com

The data consists of the various borough, latitudes and longitudes and the neighbourhoods in New York City , We choose one such Borough and exercise our analysis on it .

Data obtained by the json file is as follows :

```
In [12]: neighborhoods_data[0]
```

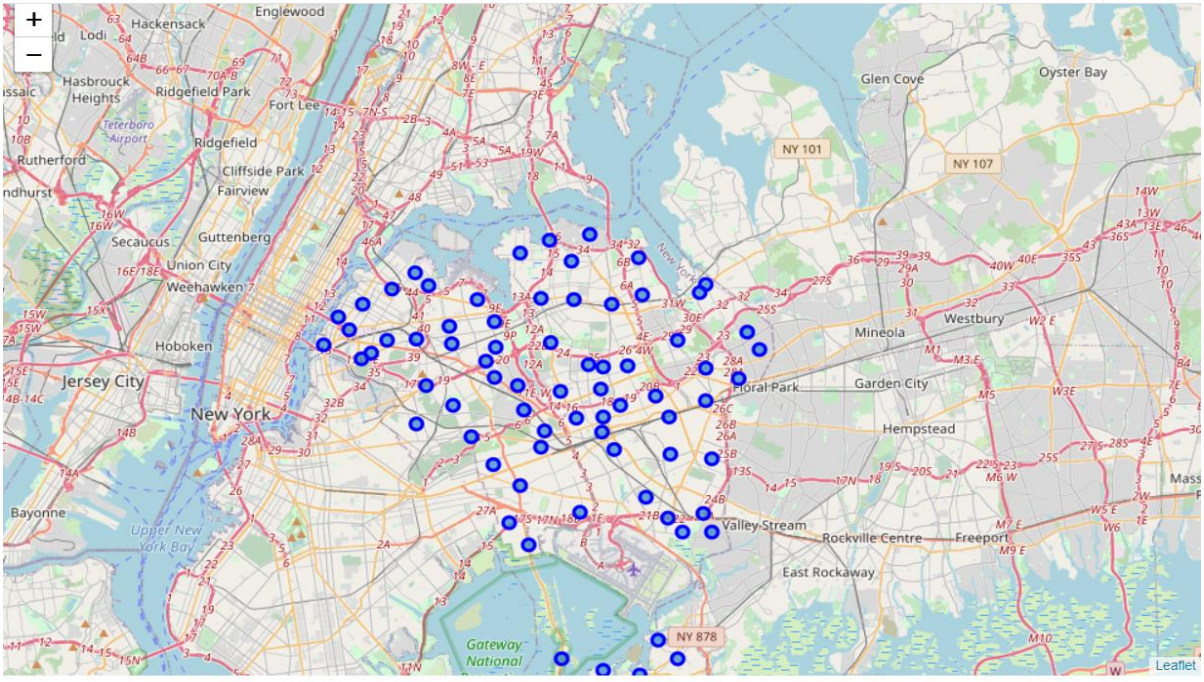
```
Out[12]: {'type': 'Feature',
          'id': 'nyu_2451_34572.1',
          'geometry': {'type': 'Point',
           'coordinates': [-73.84720052054902, 40.89470517661]},
          'geometry_name': 'geom',
          'properties': {'name': 'Wakefield',
           'stacked': 1,
           'annoline1': 'Wakefield',
           'annoline2': None,
           'annoline3': None,
           'annoangle': 0.0,
           'borough': 'Bronx',
           'bbox': [-73.84720052054902,
            40.89470517661,
            -73.84720052054902,
            40.89470517661]}}
```



We further create a map of New York using latitude and longit ude values and add markers to the map , depicting the variou s Boroughs and Neighbourhoods present in the City

3.METHODOLOGY

The main aim of this notebook was to analyse data concerning the various stores and destinations in New York city .For testing and trial purposes the location chosen was Borough : Queens in New York. Changing the Name of the B orough , one could explore various other neighbourhoods in th eir respective boroughs. We then extracted the data from the j son file and cleaned the data and then created a data frame with columns as Borough , Neighbourhood , Latitude and Long itude.We further explore the neighbourhood using Foursquare API, defining the radius as 500 m. Extracting the category of t he venues and group the neighbourhood taking the mean frequency of each category.Then we print each neighbourhood with the 6 most common venues




```
CLIENT_ID = 'N3JZIOYSXFEDAYISCE3ZDMGEYXNBPIQNFUHPWPVEZPOWEQ3E' # your Foursquare ID
```

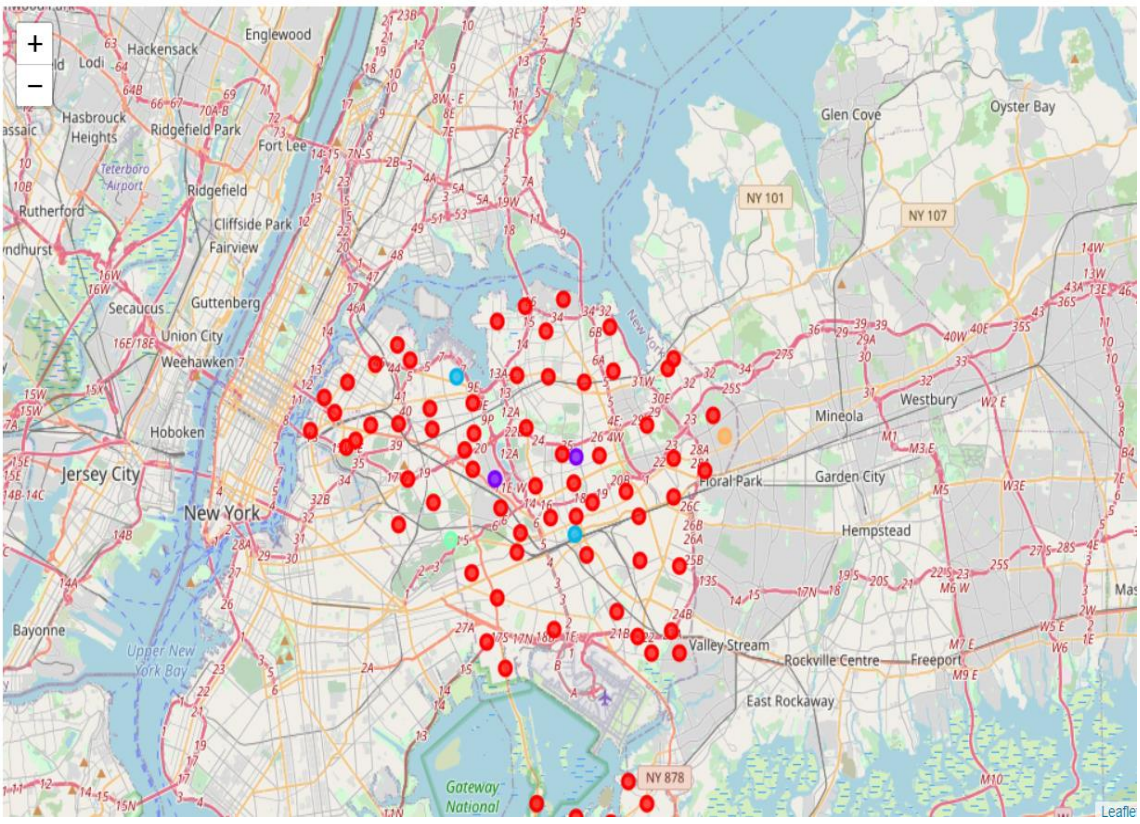
Then we display 10 most common visited venues in each neighbourhood, this information can help the individual or

company to know about the places people in each neighbourh ood like to visit and can properly also predict if the store or b usiness idea will work in that neighbourhood.

CLUSTERING

We further use K Means Clustering to cluster the neighbourho ods and take k=5 as initial number of clusters , then we examine each cluster further analysing the 10 most common venues that people visit in that cluster . This helps the client to chose the cluster where their business will profit the most .

Out[67]:



Out[71]:

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd 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	Queens	Deli / Bodega	Food & Drink Shop	Brewery	Chinese Restaurant	Pizza Place	Intersection	Arts & Crafts Store	Farm	Electronics Store	Empanada Restaurant

In [72]:
```
queens_grouped_clustering.loc[queens_grouped_clustering['Cluster Labels'] == 4, queens_grouped_clustering.
columns[[1] + list(range(5, queens_grouped_clustering.shape[1]))]]
```

Out[72]:

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd 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
52	Queens	Dosa Place	Basketball Court	Indian Restaurant	Grocery Store	Pizza Place	Chinese Restaurant	Falafel Restaurant	Egyptian Restaurant	Electronics Store	Empanada Restaurant

Now reviewing the notebook , one can see for example :

We have a client that wants to open an Indian Restaurant , as we can see that for cluster 4 , the most visited places consists of the Dosa Place and an Indian Restaurant , hence it might not be a good idea to open a restaurant there

However when we see cluster - 1 , we notice that for 6 and 69 , not many restaurants make it to the top 10 and hence these might be good places to open an Indian Restaurant

On the contrary Cluster 0 doesnt seem like a good region to open a resturant in

4.RESULT

As a result of this notebook, we were able to predict suitable r egions in the Borough Queens where a person could open an Indian Restaurant (For the case we have taken currently) .This notebook provides with various ways using folium , machine learning and k means clustering by which we can look into various neighbourhoods of a city , look for the most common venues around 500 m of that neighbourhood followed by which we can analyse neighbourhoods and cluster them further viewing the most visited places among these clusters which helps our customers to decide to buy a plot in t he cluster they finally choose.

5. DISCUSSION

Now we examine clusters

In [68]:
```
queens_grouped_clustering.loc[queens_grouped_clustering['Cluster Labels'] == 0, queens_grouped_clustering.columns[[1] + list(range(5, queens_grouped_clustering.shape[1]))]]
```

Out[68]:

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd 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	Queens	Bar	Middle Eastern Restaurant	Indian Restaurant	Greek Restaurant	Hookah Bar	Mediterranean Restaurant	Café	Bakery	Seafood Restaurant	Deli / Bodega
1	Queens	Grocery Store	Thai Restaurant	Bakery	Bar	Filipino Restaurant	Latin American Restaurant	Donut Shop	American Restaurant	Pub	Deli / Bodega
2	Queens	Latin American Restaurant	Peruvian Restaurant	South American Restaurant	Bakery	Mobile Phone Shop	Mexican Restaurant	Grocery Store	Thai Restaurant	Pizza Place	Spanish Restaurant
3	Queens	Thai Restaurant	Mexican Restaurant	Chinese Restaurant	South American Restaurant	Vietnamese Restaurant	Colombian Restaurant	Snack Place	Food Court	Sushi Restaurant	Bar
4	Queens	Pharmacy	Italian Restaurant	Sandwich Place	Bagel Shop	Fast Food Restaurant	Seafood Restaurant	Supermarket	Sporting Goods Shop	Convenience Store	Mexican Restaurant
5	Queens	Mexican Restaurant	Italian Restaurant	Sandwich Place	Donut Shop	Supermarket	South American Restaurant	Bakery	Empanada Restaurant	Restaurant	Park
7	Queens	Chinese Restaurant	Cosmetics Shop	Pharmacy	Bar	Indian Restaurant	Pizza Place	Bank	Park	Donut Shop	Food

In [69]:
```
queens_grouped_clustering.loc[queens_grouped_clustering['Cluster Labels'] == 1, queens_grouped_clustering.columns[[1] + list(range(5, queens_grouped_clustering.shape[1]))]]
```

Out[69]:

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd 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
6	Queens	Gym / Fitness Center	Gym	Yoga Studio	Pharmacy	Pizza Place	Park	Thai Restaurant	Convenience Store	Video Game Store	Deli / Bodega
69	Queens	Deli / Bodega	Spa	History Museum	Playground	Basketball Court	Bakery	Automotive Shop	Donut Shop	Indie Movie Theater	Ice Cream Shop

In [70]:
```
queens_grouped_clustering.loc[queens_grouped_clustering['Cluster Labels'] == 2, queens_grouped_clustering.columns[[1] + list(range(5, queens_grouped_clustering.shape[1]))]]
```

Out[70]:

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd 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
12	Queens	Donut Shop	Hotel Bar	Coffee Shop	Bus Station	Supermarket	Gas Station	Electronics Store	Bakery	Lake	Rental Car Location
31	Queens	Mobile Phone Shop	Caribbean Restaurant	Shoe Store	Pizza Place	Performing Arts Venue	Department Store	Coffee Shop	Clothing Store	Mexican Restaurant	Sandwich Place

In [71]:
```
queens_grouped_clustering.loc[queens_grouped_clustering['Cluster Labels'] == 3, queens_grouped_clustering.columns[[1] + list(range(5, queens_grouped_clustering.shape[1]))]]
```

Out[71]:

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd 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	Queens	Deli / Bodega	Food & Drink Shop	Brewery	Chinese Restaurant	Pizza Place	Intersection	Arts & Crafts Store	Farm	Electronics Store	Empanada Restaurant

In [72]:
```
queens_grouped_clustering.loc[queens_grouped_clustering['Cluster Labels'] == 4, queens_grouped_clustering.columns[[1] + list(range(5, queens_grouped_clustering.shape[1]))]]
```

Out[72]:

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd 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
52	Queens	Dosa Place	Basketball Court	Indian Restaurant	Grocery Store	Pizza Place	Chinese Restaurant	Falafel Restaurant	Egyptian Restaurant	Electronics Store	Empanada Restaurant

Now reviewing the notebook , one can see for example :

We have a client that wants to open an Indian Restaurant , as we can see that for cluster 4 , the most visited places consists of the

Dosa Place and an Indian Restaurant , hence it might not be a good idea to open a restaurant there

However when we see cluster - 1, we notice that for 6 and 69 , not many restaurants make it to the top 10 and hence these might be good places to open an Indian Restaurant

On the contrary Cluster 0 doesn’t seem like a good region to open a restaurant in

6. CONCLUSION

In this study, using clustering and segmentation I analysed and predicted a suitable location for a customer to set up an Indian Restaurant, in various neighbourhood clusters of Queens New York.