APPLIED DATA SCIENCE CAPSTONE

(COURSERA)

BATTLE OF NEIGHBORHOODS

MOST PREFERED BOROUGH FOR GLUTEN-INTOLERANT PEOPLE

NANCY JENNIFER DSOUZA

24TH MAY 2020

INDEX

Sr.No.	Topics	Page Nos.
1	**Introduction**	5
1.1	Background	5
1.2	Problem	5
1.3	Stack holders	5
2	**Data acquisition and Data cleaning**	5
2.1	Cleaning of the data	6
2.2	Selection of the features	6
3	**Methodology**	7
3.1	Boroughs with gluten-free restaurants	7
3.2	Borough with maximum number of gluten-free restaurants	7
4	**Results**	8
4.1	For the number of gluten-free restaurants in a borough	8
4.2	Maximum Density of gluten-free restaurants in a borough	11
5	**Discussion**	14
6	**Conclusions**	15
7	**Future Scope of Research**	15

LIST OF FIGURES

Figure no.	Figure Name	Page no.
1	Map of New York City showing the borough Bronx	8
2	Map of New York City showing the borough Brooklyn	9
3	Map of New York City showing the borough Manhattan	9
4	Map of New York City showing the borough Queens	10
5	Map of New York City showing the borough Staten Island	10
6	Map of New York City showing the average distance from the average coordinates in Bronx	12
7	Map of New York City showing the average distance from the average coordinates in Brooklyn	12
8	Map of New York City showing the average distance from the average coordinates in Manhattan	13
9	Map of New York City showing the average distance from the average coordinates in Queens	13
10	Map of New York City showing the average distance from the average coordinates in Staten Island	14

LIST OF TABLES

Table No.	Table name	Page no.
1	The table describes the details of the borough dataset along with its respective latitude and longitude values	6
2	Number of gluten-free restaurants in a particular borough arranged in descending order.	8
3	Average distance from average coordinates in a particular borough arranged in ascending order of distances.	11
4	Average distance from average coordinates in a particular borough after outlier removal arranged in ascending order of average distances	14
5	Comparison of the average distances from the average coordinates in a particular borough before and after removal of outliers arranged in ascending order of average distances.	15

1. Introduction

1.1 Background

New York City shortly referred to as NYC is the most popular city in the United States. NYC has been entitled as the financial, cultural and media capital of the world, significantly influencing various fields like business, entertainment, research tech, education, politics, tourism, art, fashion, and sports. The state also comprises of a variety of cuisines since people from all around the world reside and frequently pay visits. Situated at one of the world's largest natural harbors, NYC consists namely of five boroughs which are Bronx, Manhattan, Brooklyn, Queens and Staten Island. All around the year tourists keep visiting NYC for a blissful experience of different art forms and cultures. If a tourist who is gluten-intolerant visits New York, he/she would prefer a borough where gluten-free food is easily available.

To give a brief outline, gluten allergy is medically known as Coeliac Disease. It is described as an immune reaction which occurs by consuming gluten, which is a type of protein found in wheat, barley and rye. This condition is highly common with over 1 to 2 million cases every year. This condition cannot be cured but only treated with medical care. Therefore, tourists suffering from such conditions should be taken care of, as this can not only spoil their vacation but also affect their health in a serious way.

1.2 Problem

This research problem is for such a tourist who wants to visit NYC for a trip and is gluten-intolerant. Henceforth, he or she would appreciate a borough where gluten-free food is easily accessible. The solution is to find a borough which has a maximum number and as well as high density of gluten-free restaurants.

1.3 Stake holders

This research would benefit people who are gluten-intolerant and would like to pay a visit to New York. This will help them choose a borough which would give them access to restaurants which serve gluten-free food.

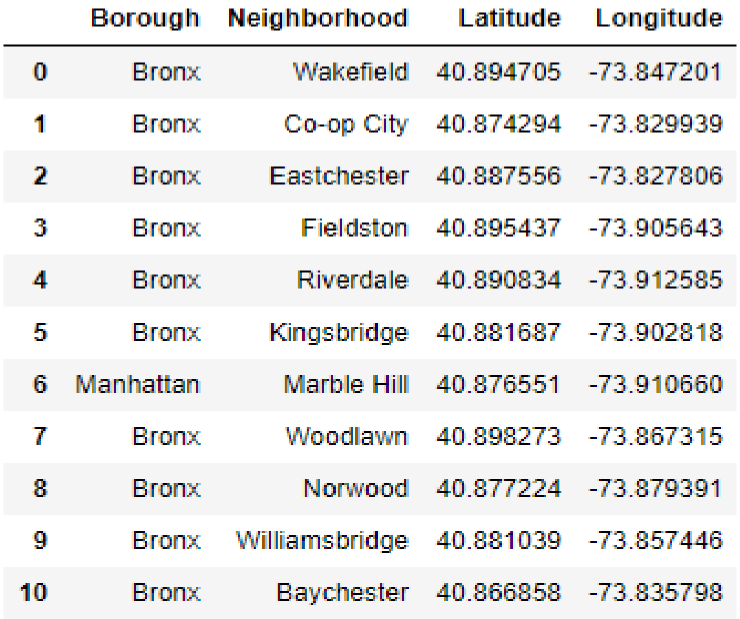
2. Data acquisition

For this research, the following will be required:

a. The number of gluten-free restaurants in all the boroughs
b. Data of the borough and neighborhoods with their respective geographical coordinates (latitude and longitude)

The Geographical Data will be taken by Foursquare API as well as the following links have also been used in the study:

a. https://geo.nyu.edu/catalog/nyu_2451_34572
b. https://cocl.us/new_york_dataset



	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
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391
9	Bronx	Williamsbridge	40.881039	-73.857446
10	Bronx	Baychester	40.866858	-73.835798

Table 1. The table describes the details of the borough dataset along with its respective latitude and longitude values

2.1 *Data Cleaning*

Data cleaning is an essential step since it removes any unnecessary details which might not be useful in the purpose of this research. For this step, we make sure that all of the borough data is available along with its latitude and longitude values and that there are no null values which could be represented by characters like ‘Not Available’, ‘?’, ‘-‘ or ‘NA’ which could be present in the obtained raw form of the data.

2.2 *Selection of the Features*

This step describes the selection of important aspects or ‘features’ which are taken into consideration from the cleaned dataset which could be helpful in attaining the goals of this research. For this step, the following have been selected:

a. The borough dataset consists of 305 rows. From these, features namely, ‘Neighborhood’ which represents the neighborhood name, ‘Latitude’ which represents the latitude value of the respective neighborhood and ‘Longitude’ which represents the longitude value of the respective neighborhood.

b. For the gluten-free restaurants in a particular borough which is obtained by the Foursquare API, the features selected were 'Name' which represents the name of the restaurant, 'Address' which refers to the location of the restaurant, 'Latitude' and 'longitude' which represents the geographical coordinate values of the respective restaurant. Note that, for different boroughs, the size of the dataset varied.

3. Methodology

The methodology implemented for the purpose of this research is explained in the following two steps:

3.1 Obtaining the Borough with the maximum number of gluten-free restaurants

This is the first step from which the borough with highest number of gluten-free restaurants will be obtained. This is done so that the person is not stuck with just a few options, but has a good number of places to try. Greater the number of restaurants, greater is the demand of gluten-free food in that particular borough. This also means more competition between the restaurants which generally refers to better quality of food served by them.

Due to this reason, customer satisfaction would be an important aspect in these restaurants since there are many in a particular borough. In order to complete this task, foursquare API details along with the category ID of the gluten-free restaurants in different boroughs are used as parameters. The foursquare API details include the client ID, client secret and the present version of foursquare.

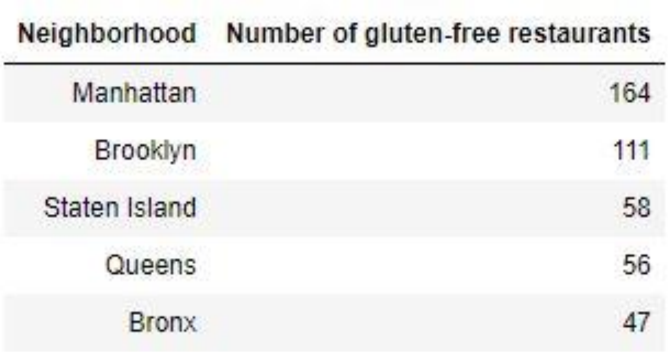
3.2 Obtaining the Borough with the highest density of gluten-free restaurants

Once we have obtained the borough with the maximum number of restaurants, the next step is to make sure that the borough obtained in the previous step consists of densely populated options of restaurants. This essentially means that the restaurants in the obtained borough should be such that they are close to one another so that the tourist does not have to travel long distances just to have a meal.

This step is done by calculating the average of the coordinates with respect to all of the restaurants and furthermore by calculating the average distance of the restaurants from the average of the coordinates. Thus, this step will give the tourist a rough idea of how far each restaurant is from his or her particular spot.

4. Results

4.1 For the number of gluten-free restaurants in a particular borough



Neighborhood	Number of gluten-free restaurants
Manhattan	164
Brooklyn	111
Staten Island	58
Queens	56
Bronx	47

Table 2: Number of gluten-free restaurants in a particular borough arranged in descending order. Note that these are depicting out of top 100 restaurants.

From the table 2, the results clearly depicts that the borough 'Manhattan" has the maximum number of gluten-free restaurants having a count of 164, followed by Brooklyn on 111 and Staten Island with a count of 58.

Maps using the folium package were created in order to visualize the amount of restaurants populated in a particular neighborhood. These maps are shown below in figs 1, 2, 3, 4 and 5. Note that each data point shown on the map as a purple circle represents one gluten-free restaurant in the respective neighborhood.

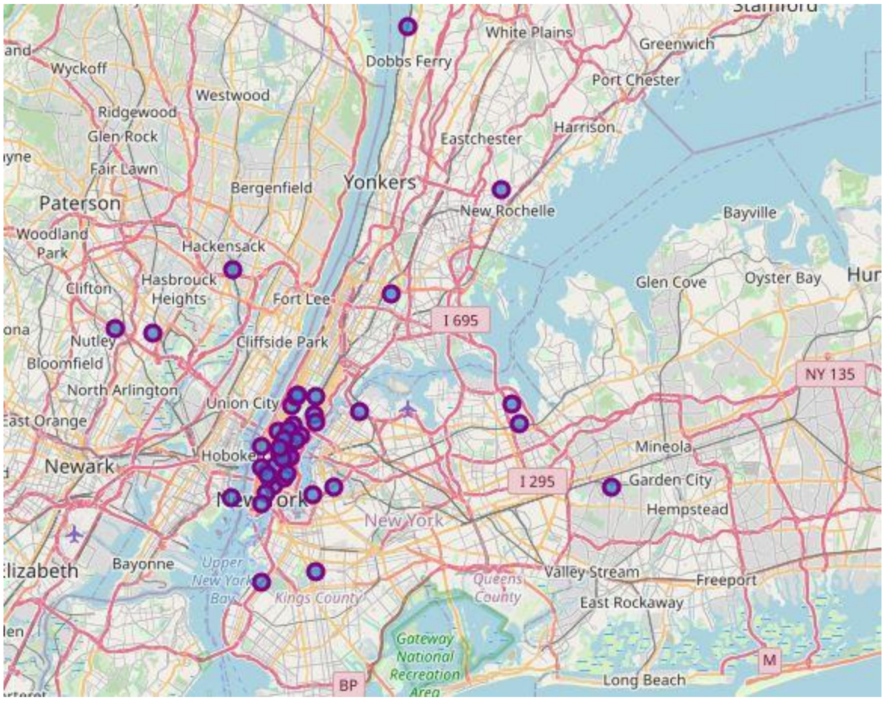


Fig.1.Map of New York City showing the borough Bronx.

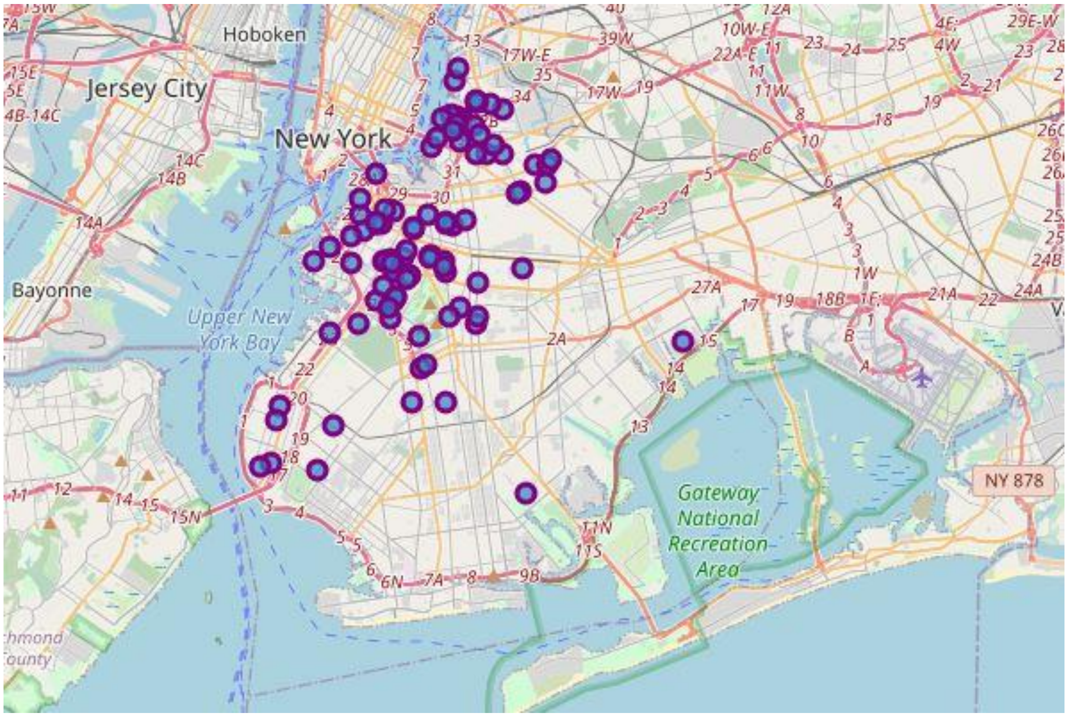


Fig.2.Map of New York City showing the borough Brooklyn.

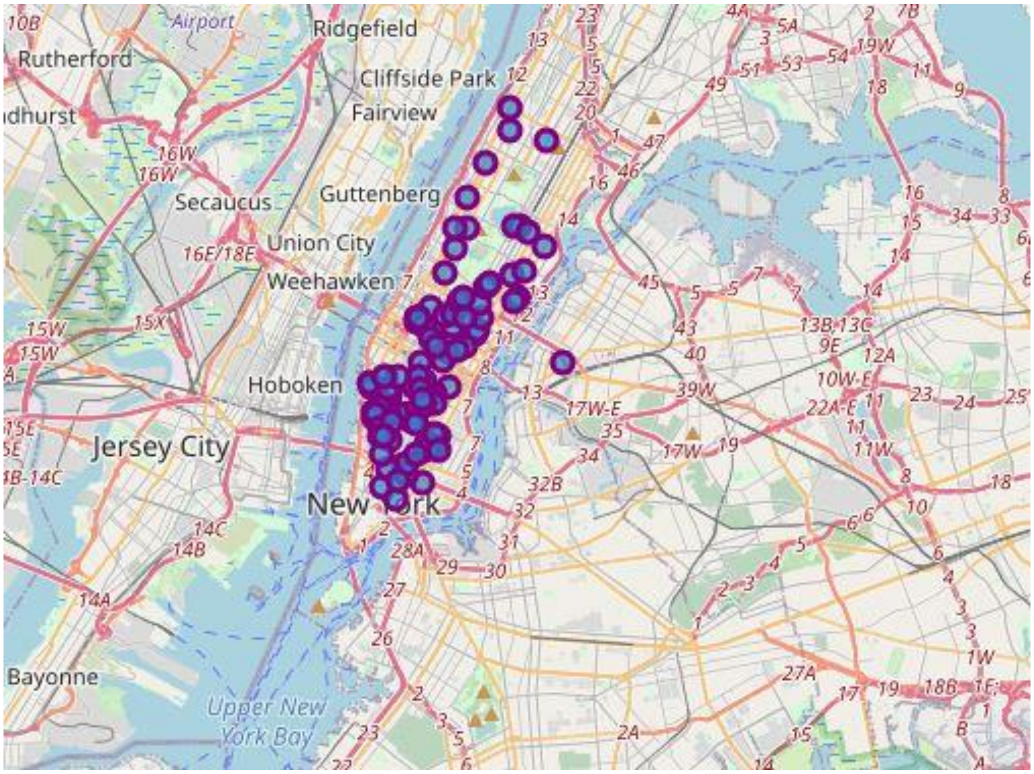


Fig.3.Map of New York City showing the borough Manhattan.

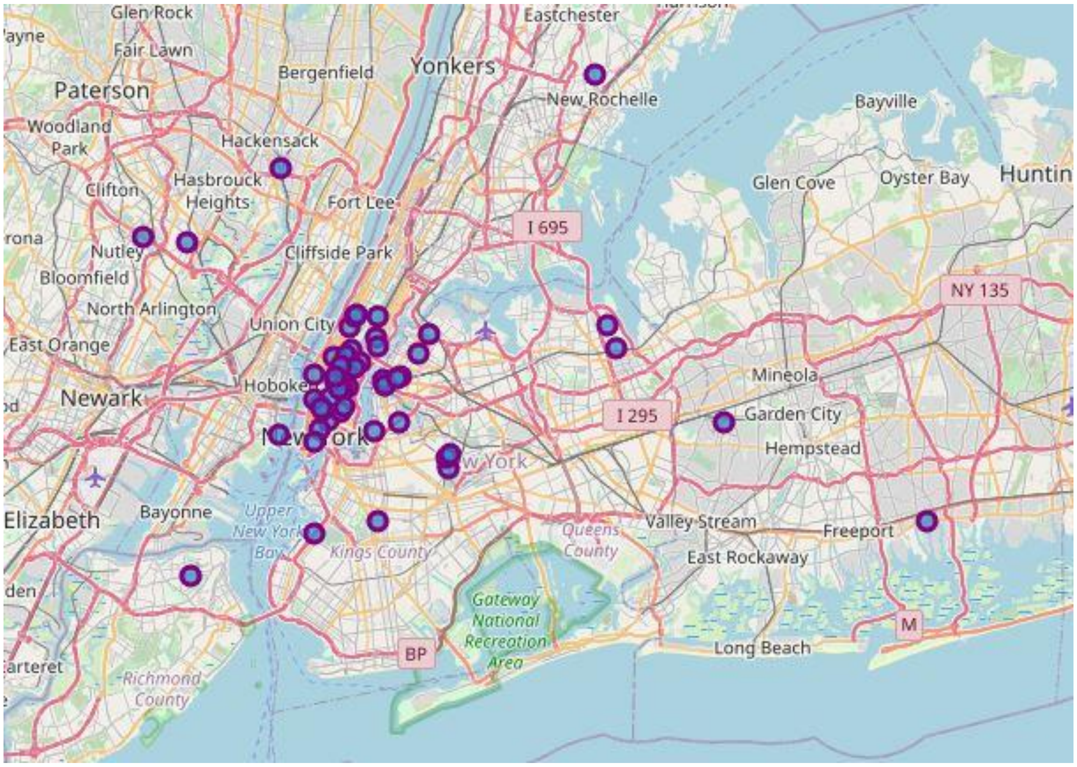


Fig.4.Map of New York City showing the borough Queens.

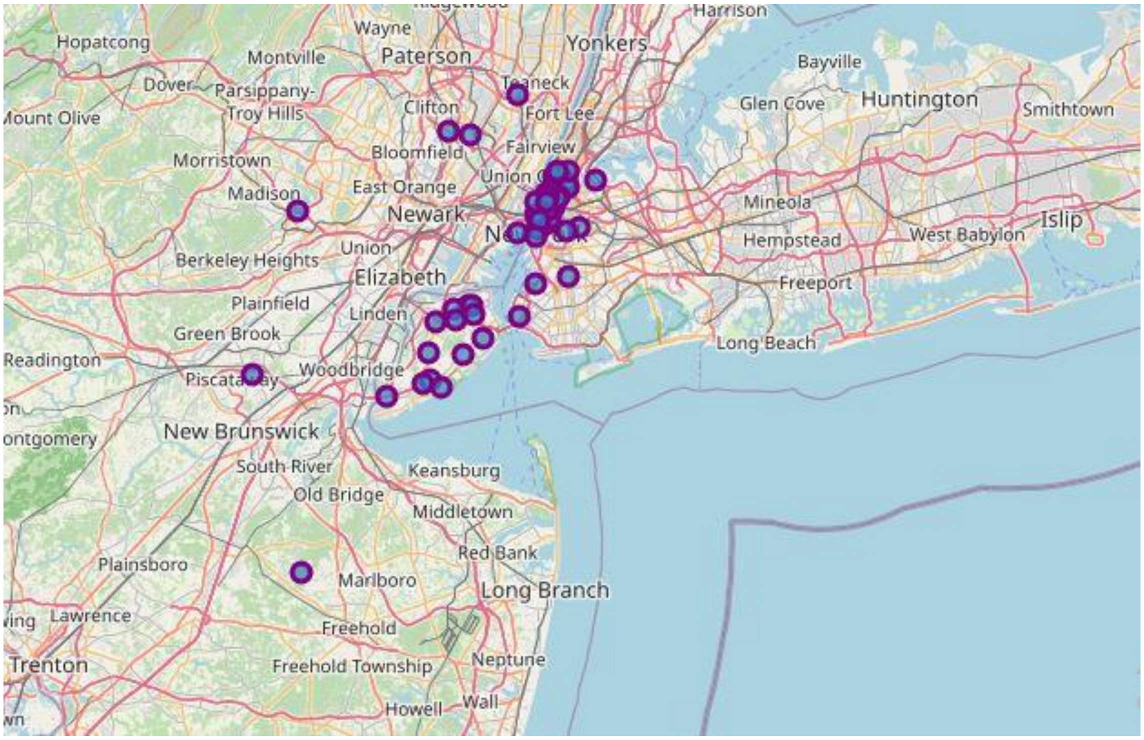
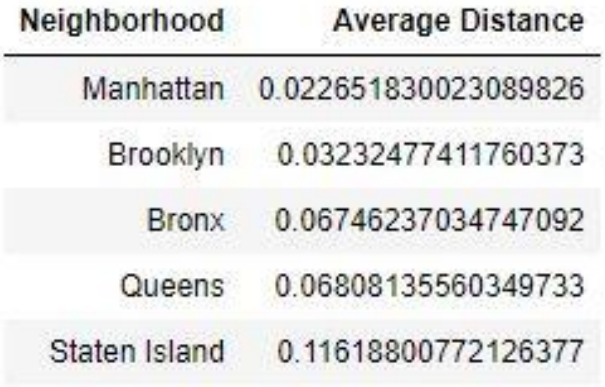


Fig.5.Map of New York City showing the borough Staten Island.

4.2 Maximum Density of gluten-free restaurants in a particular borough



Neighborhood	Average Distance
Manhattan	0.022651830023089826
Brooklyn	0.0323247741176037З
Bronx	0.06746237034747092
Queens	0.06808135560349733
Staten Island	0.11618800772126377

Table 3. Average distance from average coordinates in a particular borough arranged in ascending order of distances.

From the table 3 shown above, it can be clearly seen that the borough “Manhattan” has the most densely populated gluten-free restaurants among the other neighborhoods since the average distance is the least with a value of 0.02265 compared to other neighborhoods.

Using the folium package, maps were made in order to visualize the average distance from the average coordinates. These can be seen below in Figs 6, 7, 8, 9 and 10. Note that each data point shown by a black circle on the map represents one gluten-free restaurant, and the average coordinate is depicted by a blue circle on the map. The distances are shown by red lines on the map and can be interpreted as, bigger the red line, greater is the distance of the restaurant from the average coordinate in the respective neighborhood. The average of all the distances was taken to obtain the values shown in table 3 shown above.

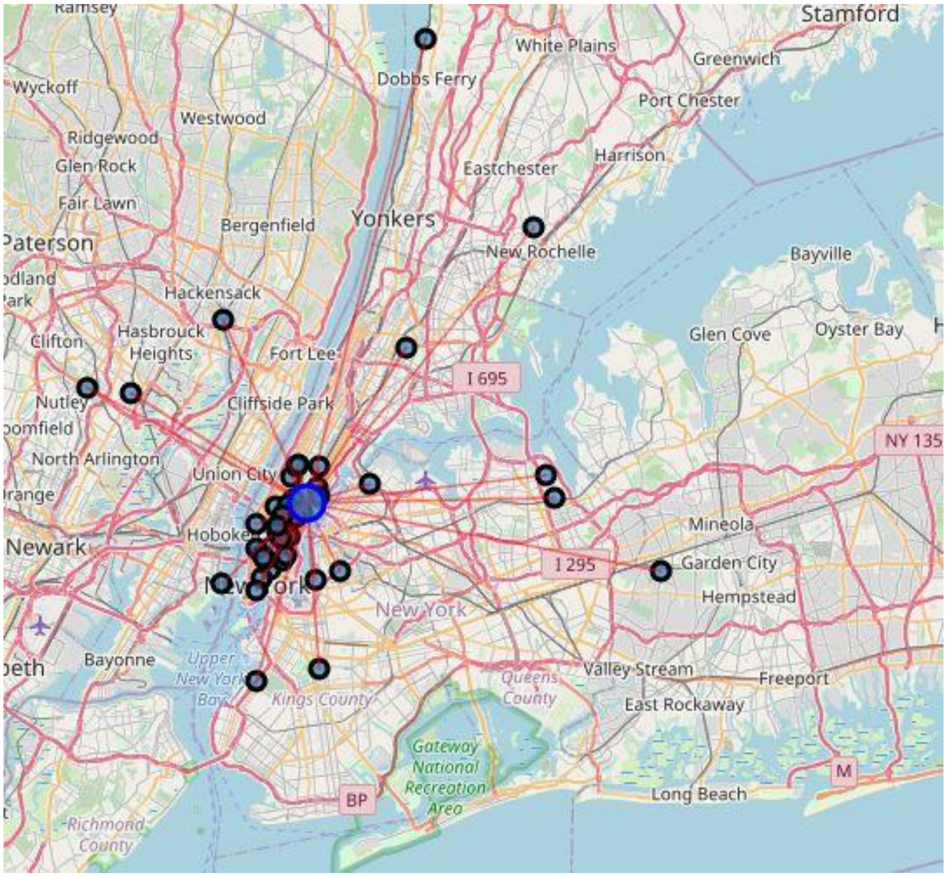


Fig.6. Map of New York City showing the average distance from the average coordinates in Bronx

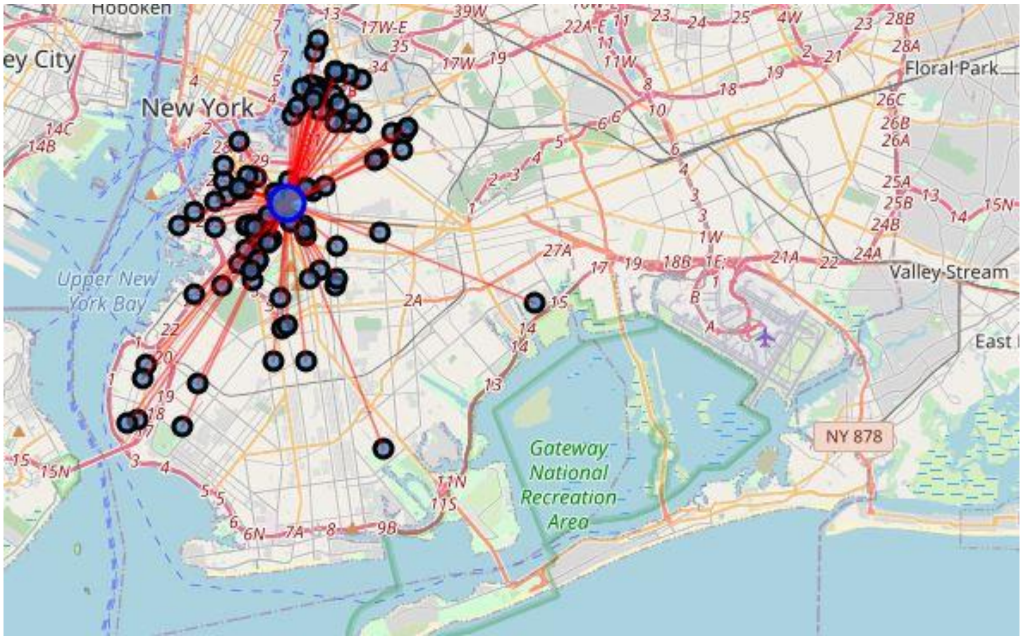


Fig.7. Map of New York City showing the average distance from the average coordinates in Brooklyn.

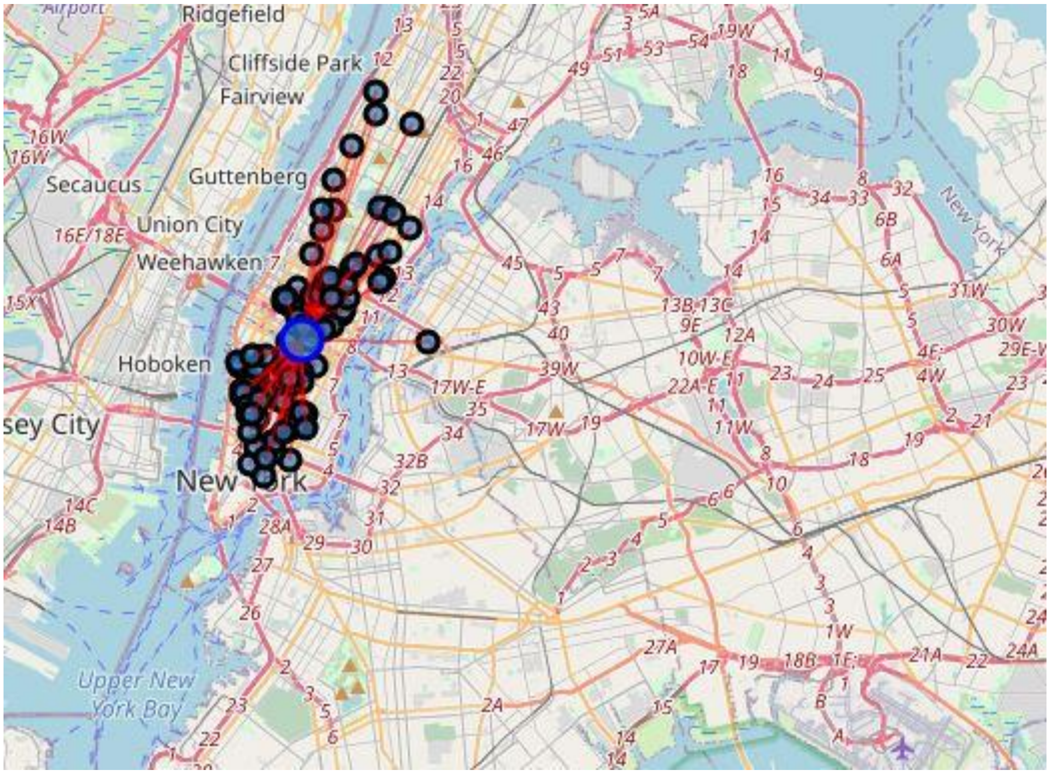


Fig.8. Map of New York City showing the average distance from the average coordinates in Manhattan.

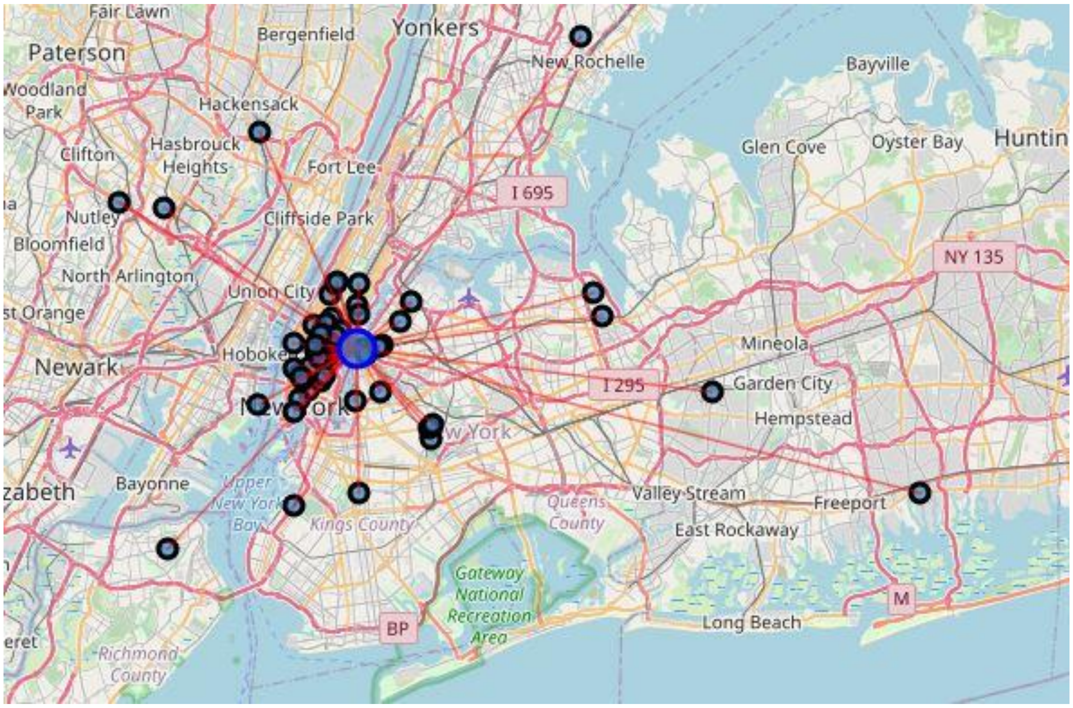


Fig.9. Map of New York City showing the average distance from the average coordinates in Queens.

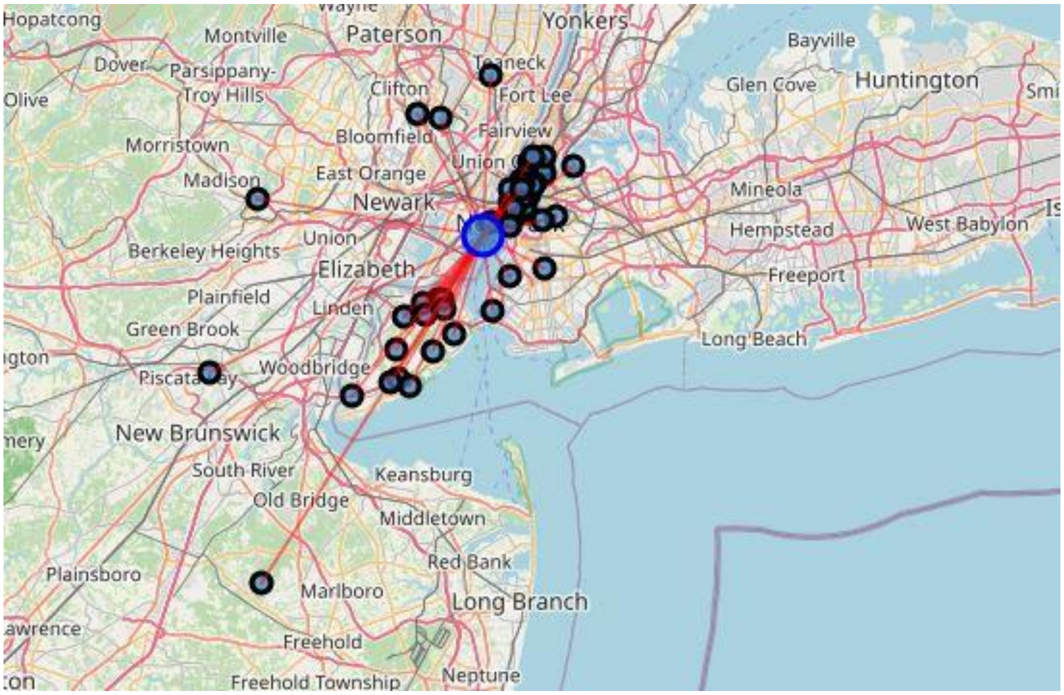
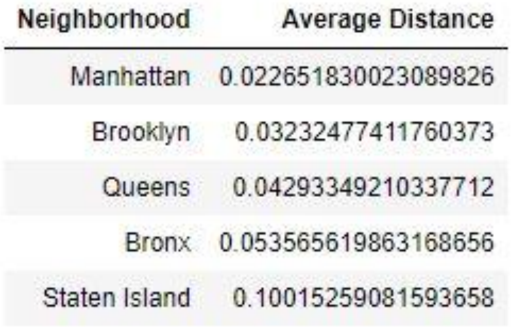


Fig.10. Map of New York City showing the average distance from the average coordinates in Staten Island.

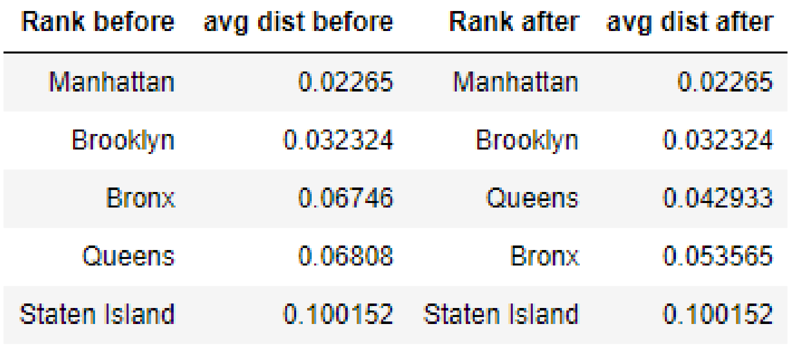
5. Discussion

An essential step in a research study is to make sure that the results obtained are not influenced by the presence of outliers. These can affect the results and the predictions of new cases have a higher rate of being wrong. In this research, we can see that there are a few restaurants in some of the neighborhoods namely, Bronx, Queens and Staten Island, which are located on a farther distance. These are treated as outliers and therefore need to be removed so as to make sure the results obtained are not under the influence of these outliers. From the figures 6, 9 and 10, it is clearly seen that, in the boroughs Bronx, Queens and Staten Island, there are 3, 7 and 3 restaurants respectively which are situated at a much farther distance compared to the others. These outliers might have influenced our result and hence before removing them, a conclusion cannot be made. After the removal of these outliers, the average distance from the average of the coordinates are once again calculated and the results obtained are shown in table 4.



Neighborhood	Average Distance
Manhattan	0.022651830023089826
Brooklyn	0.03232477411760373
Queens	0.04293349210337712
Bronx	0.053565619863168656
Staten Island	0.10015259081593658

Table.4. Average distance from average coordinates in a particular borough after outlier removal arranged in ascending order of average distances



Rank before	avg dist before	Rank after	avg dist after
Manhattan	0.02265	Manhattan	0.02265
Brooklyn	0.032324	Brooklyn	0.032324
Bronx	0.06746	Queens	0.042933
Queens	0.06808	Bronx	0.053565
Staten Island	0.100152	Staten Island	0.100152

Table.5. Comparison of the average distances from the average coordinates in a particular borough before and after removal of outliers arranged in ascending order of average distances.

From tables 4 and 5, it can be clearly seen that outliers affected the distance of each of the three boroughs, Bronx, Queens and Staten Island. After the outliers were removed, the ranking of the boroughs based on the average distances also changed. But this change did not affect our result of the preferred borough which is Manhattan. Since Manhattan is both densely populated and ranks highest in terms of maximum number of gluten-free restaurants, Manhattan should be the preferred choice.

6. Conclusions

From the obtained results, it is clear that Manhattan is the best choice for a gluten-intolerant tourist who wants to visit New York City for a vacation. The outliers influence was present however, the result obtained after the removal of outliers is still significantly high compared to Manhattan. Therefore, the presence of outliers did not affect the affect the final result.

7. Future Scope of Research

This study can be helpful to anyone travelling to NYC and is gluten-intolerant since it is a very common disease. A neighborhood such as Manhattan would be the best choice for such a tourist, since restaurants which serve gluten-free food is easily accessible as compared to other neighborhoods. A limitation in this study would be that, the study does not take into consideration the ratings of the restaurants in each neighborhood, which could be used as an important feature to find a preferred borough which has good quality food and customer service. Along with the ratings, the type of cuisine such as Indian, Chinese or Italian can also be taken into account.