Exploring New York City's Diverse Food Cultures for Tourism Purposes

Shabbir Ahmad

Presentation

for

Applied Data Science Capstone Project

April 29, 2020

Contents

- Introduction/Problem Statement
- Data Description
- Methodology
- Data Analysis
- Discussion
- Conclusion

Introduction / Problem Statement

- New York City (NYC) is the largest metropolitan area in the world
- Over 65 million tourists visited NYC during 2019
- Tourists often find it difficult to find the restaurants and eateries of their choice in those areas.
- Tourists are also overwhelmed by the food choices and the decision-making process may be time consuming.
- This project explores the boroughs of NYC and locate all the possible food options in different neighbourhoods of the city.
- Different types of cuisines are located in the city.

Data Description

Data Requirements

- Data related to boroughs (including neighbourhoods) of New York City is required for this project.
- Specific data regarding different locations are extracted from Foursquare API.
- Libraries like pandas, numpy, matplotlib, folium (for mapping), and pickle are employed for data processing.

Data Description

- **New York City dataset:** This dataset provides the addresses of neighbourhoods of the city. A json file is extracted from this link https://geo.nyu.edu/catalog/nyu_2451_34572

- **Foursquare API:** It is a location data provider and will be used to make API calls to extract data regarding different venues in New York City's neighbourhoods. To access the API, 'CLIENT_ID', 'CLIENT_SECRET' and 'VERSION' will be defined. This link is used. https://developer.foursquare.com/docs/

Stakeholders/Audience

- The audience of this report are mainly local and foreign tourists in New York City who would like to find the eatery of their choice conveniently or explore the diverse food choices available at a particular locality.

- Findings of this report can also be used by the tourism department of the city as well as by tour operators to guide the tourists regarding the locations of the diverse food options.

Methodology

Downloading and exploring the dataset

- The dataset is downloaded from the aforementioned URL.
- The URL returns the .json file containing the dataset in the form of a python dictionary.
- The requisite data i.e. the list of five boroughs and the neighbourhoods within these boroughs, is in the features key.

Methodology

Dataframe ‘neighbourhoods’

- The dictionary is transformed into a dataframe by looping through the data.
- It results into a dataframe with borough, Neighbourhood, Latitude and Longitude details of the city’s neighbourhoods.
- The dataframe has a total of 05 boroughs and 306 neighbourhoods.
- The latitude and longitude of the New York City is found through ‘Geopy’ library. The latitude is 40.71 while the longitude is -74.01.

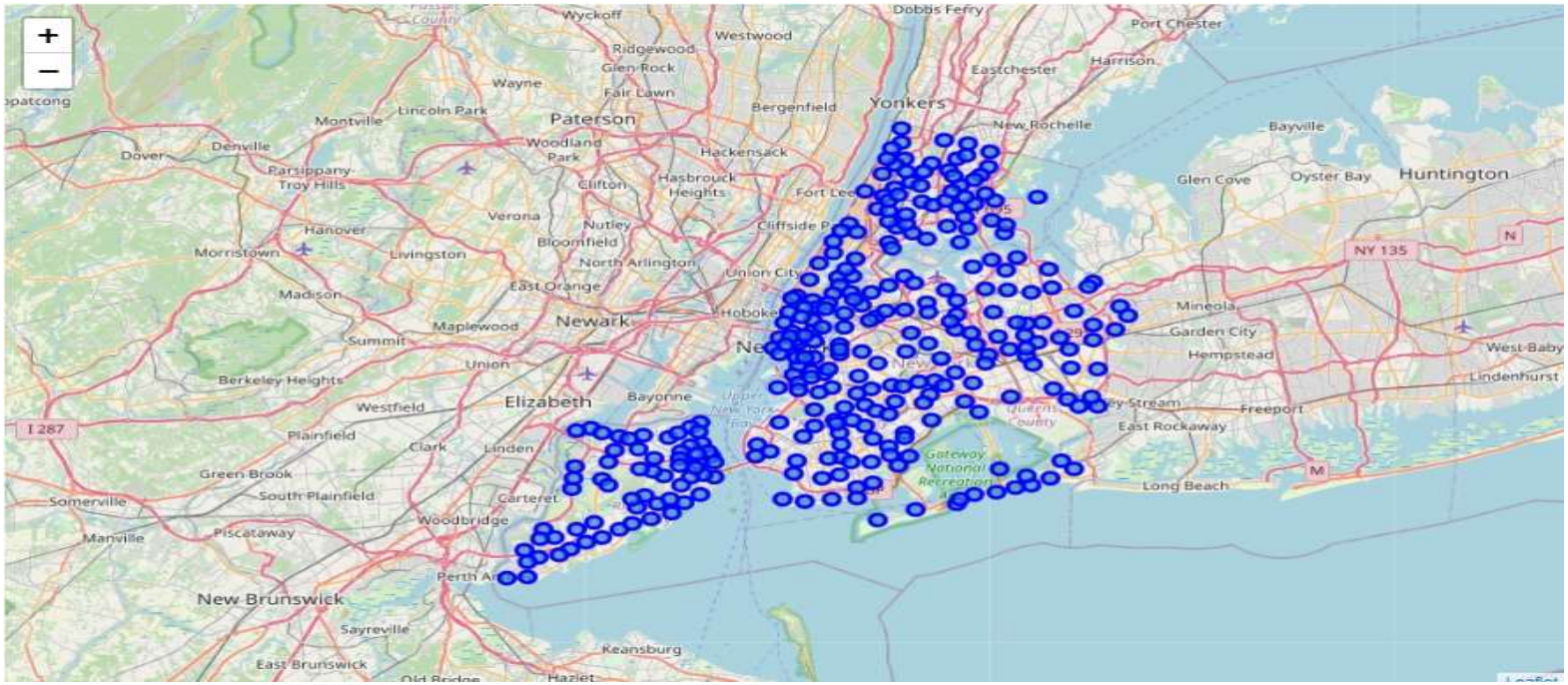
Methodology

First 05 rows of the dataframe

	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

Methodology

The dataframe is further used to generate a map of New York City with neighbourhoods superimposed on top. The map is generated by using the ‘folium’ library.



Methodology

Foursquare API

After extracting Foursquare venue category hierarchy, a list of ten categories of venues is generated but we are only interested in ‘Food’

```
for data in category_list:
    print(data['id'], data['name'])
```

```
4d4b7104d754a06370d81259 Arts & Entertainment
4d4b7105d754a06372d81259 College & University
4d4b7105d754a06373d81259 Event
4d4b7105d754a06374d81259 Food
4d4b7105d754a06376d81259 Nightlife Spot
4d4b7105d754a06377d81259 Outdoors & Recreation
4d4b7105d754a06375d81259 Professional & Other Places
4e67e38e036454776db1fb3a Residence
4d4b7105d754a06378d81259 Shop & Service
4d4b7105d754a06379d81259 Travel & Transport
```

Methodology

Dataframe ‘nyc_venues’

- Instead of using GET request for all neighbourhoods to extract information regarding latitude and longitude, a function ‘getNearbyFood’ is created.
- It creates an API request URL with radius = 500, LIMIT = 100. Limit is set to 100, so that a maximum of 100 venues are returned in the vicinity.
- A new dataframe is created for information regarding all the 306 neighbourhoods.
- The dataframe has 13612 venues and 187 unique sub-categories.

Methodology

Data Cleaning

- Among the 187 categories of food, general categories like ice cream shops, cafes and bars are also included. We need to eliminate these to make our work easier.
- List of general categories is created and then subtracted from all the categories. It leave us only the restaurants and eateries.
- After this process, we have 101 unique categories.

Methodology

One hot encoding

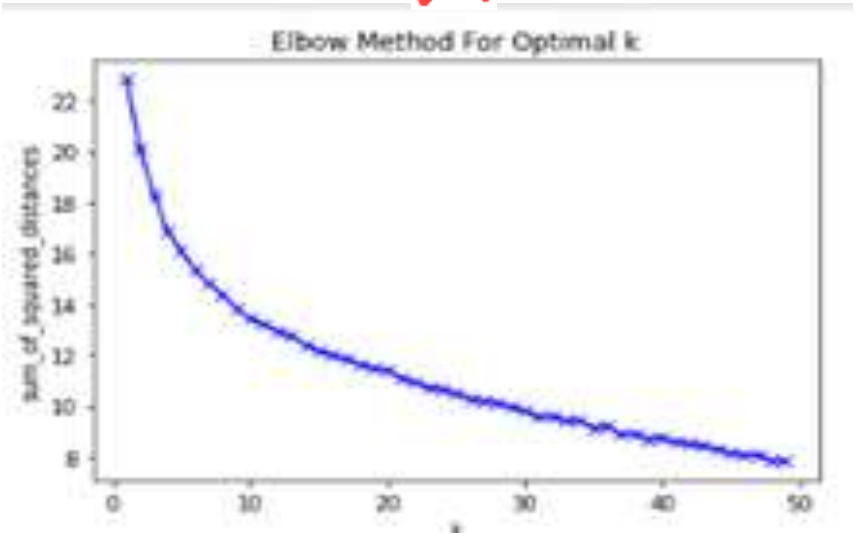
- It is used to analyse each neighbourhood individually to find out the most common cuisine within 500 meters of its radius.
- A new dataframe nyc_onehot is created. ‘Neighbourhood’ column. The size of this dataframe is 6635 data points including all venues for all the neighbourhoods. Now, we can also count the number of venues for each sub-category in every neighbourhood.

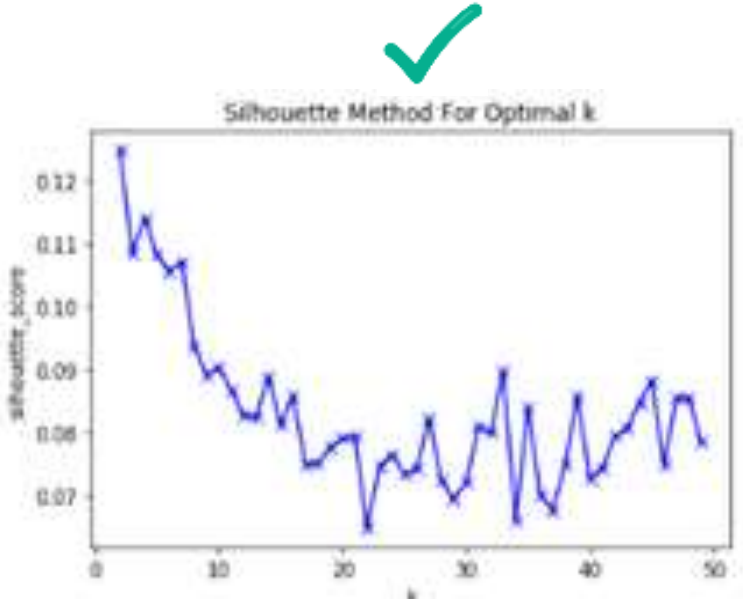
➢ The frequency of occurrence of each category is determined by taking mean of all values. The Foursquare API may return many venues since the limit is set to 100

Methodology

Data Processing - Clustering through K-Means

- To find the optimal number of clusters, The Elbow Method and The Silhouette Methods are used. The latter provides a better solution.
- K = 8





Methodology

Each cluster got the following number of neighbourhoods. Cluster 4 has the highest number (57) while Cluster 3 has the lowest (04).

Setting the number of clusters to 8

```
kclusters = 8

# run k-means clustering
kmeans = KMeans(init="k-means++", n_clusters=kclusters, n_init=50).fit(nyc_grouped_clustering)

print(Counter(kmeans.labels_))
```

```
Counter({4: 57, 2: 55, 6: 54, 5: 51, 7: 33, 1: 26, 0: 22, 3: 4})
```

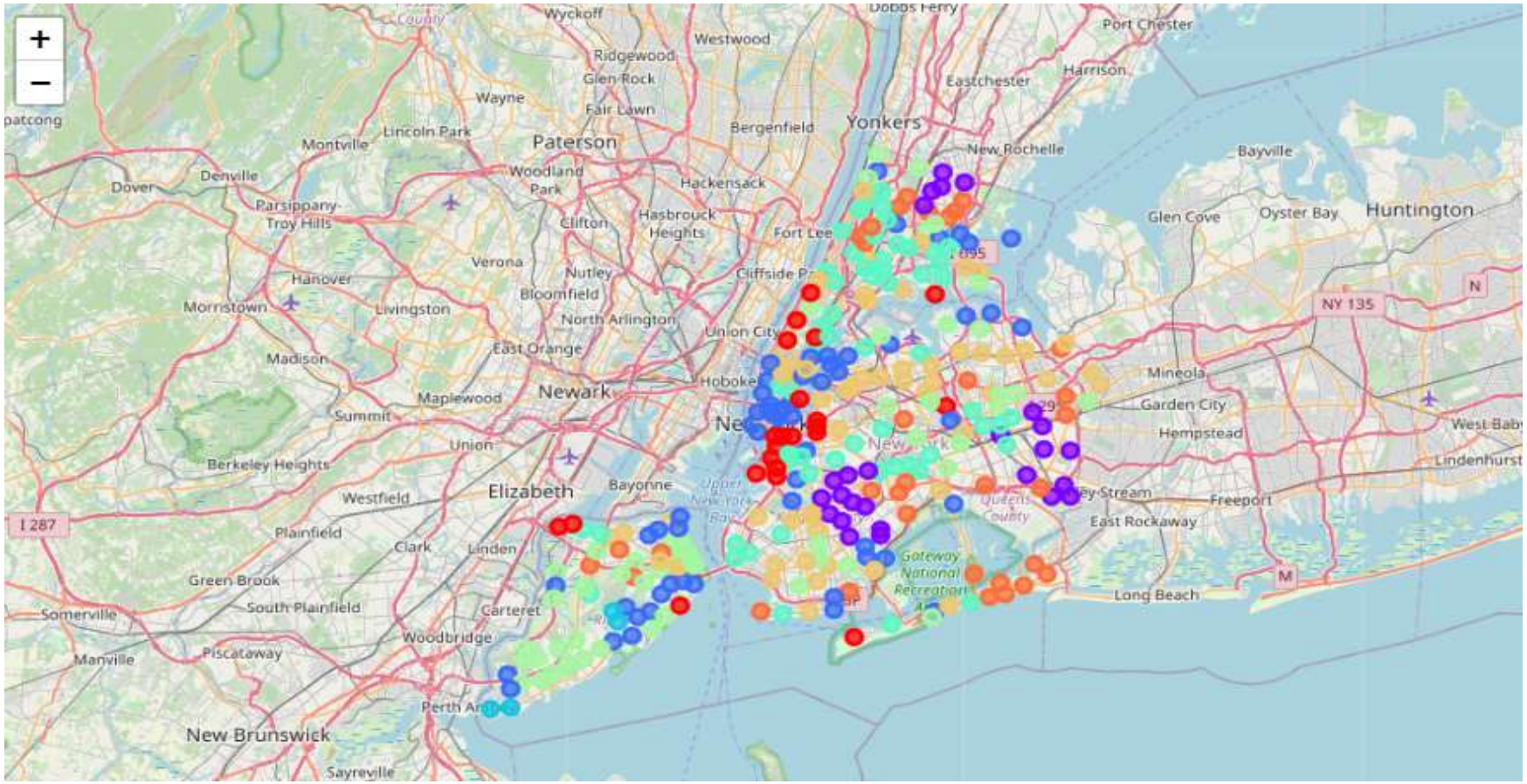
Methodology

- Adding Borough, Latitude and Longitude for each neighbourhood to get a complete set of information regarding the venues available in different neighbourhoods.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
0	7	Allerton	Chinese Restaurant	Pizza Place	Mexican Restaurant	Fast Food Restaurant	Caribbean Restaurant	Bronx	40.865788	-73.859319
1	5	Annadale	Pizza Place	Sushi Restaurant	American Restaurant	Italian Restaurant	Japanese Restaurant	Staten Island	40.538114	-74.178549
2	5	Arden Heights	Pizza Place	American Restaurant	Italian Restaurant	Sushi Restaurant	Mexican Restaurant	Staten Island	40.549286	-74.185887
3	4	Arlington	Pizza Place	Peruvian Restaurant	American Restaurant	Fast Food Restaurant	Spanish Restaurant	Staten Island	40.635325	-74.165104
4	2	Arrochar	Italian Restaurant	Pizza Place	Chinese Restaurant	Steakhouse	Middle Eastern Restaurant	Staten Island	40.596313	-74.067124

Methodology

Neighbourhoods clusters look like this



Results

Cluster – 0

- Pizza Place is the 1st and American Restaurant is the 2nd most common venue.

```
for col in required_column:
    print(cluster_0[col].value_counts(ascending = False))
    print("-----------------------------------------------")
```

```
Pizza Place               12
American Restaurant        8
French Restaurant          1
Seafood Restaurant         1
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------------
American Restaurant            8
Pizza Place                    6
French Restaurant              1
Seafood Restaurant             1
BBQ Joint                      1
Mediterranean Restaurant       1
Mexican Restaurant             1
Italian Restaurant             1
Ramen Restaurant               1
Fast Food Restaurant           1
Name: 2nd Most Common Venue, dtype: int64
-----------------------------------------------
Brooklyn          11
Manhattan          5
Staten Island      3
Queens             2
Bronx              1
Name: Borough, dtype: int64
-----------------------------------------------
```

Results

Cluster – 1

- Caribbean Restaurant is the 1st and Fast Food Restaurant is the 2nd most common venue.

```
for col in required_column:
    print(cluster_1[col].value_counts(ascending = False))
    print("-----------------------------------------------")
```

```
Caribbean Restaurant                    23
Fried Chicken Joint                      1
Pizza Place                              1
Southern / Soul Food Restaurant          1
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------------
Fast Food Restaurant     8
Pizza Place              7
Chinese Restaurant       6
Caribbean Restaurant     3
Indian Restaurant        1
Fried Chicken Joint      1
Name: 2nd Most Common Venue, dtype: int64
-----------------------------------------------
Brooklyn     12
Queens        9
Bronx         5
Name: Borough, dtype: int64
-----------------------------------------------
```

Results

Cluster – 2

- Italian Restaurant is the 1st and Pizza Place is the 2nd most common venue.

Cluster – 3

- Italian Restaurant is the 1st and Asian Restaurant is the 2nd most common venue.

Cluster – 4

- Pizza Place is the 1st and Fast Food Restaurant is the 2nd most common venue.

Results

Cluster – 5

- Pizza Place is the 1st and Italian Restaurant is the 2nd most common venue.

Cluster – 6

- Pizza Place is the 1st and Chinese Restaurant is the 2nd most common venue.

Cluster – 7

- Chinese Restaurant is the 1st and Pizza Place is the 2nd most common venue.

Discussion

- K-mean clustering algorithm is used to segment New York City's neighbourhoods.
- The optimal number of clusters is eight for this project as found out through 'The Silhouette Method.'
- Each of these clusters is analysed through determining the borough count, count of 'First Most Common Venue' and 'Second Most Common Venue.'
- Results for the eights clusters are summarized in the table on the next slide

Discussion

- Pizza is the most common food in New York City. Since it is considered as go-to food, we can set it aside and regard Italian food as the most common cuisine. Caribbean, Chinese, Asian and American foods are also common in different areas.

	Occurrences' Count within the Cluster		
Cluster	1st Most Common Venue	2nd Most Common Venue	Borough
0	Pizza Place	American Restaurant	Brooklyn
1	Caribbean Restaurant	Fast Food Restaurant	Brooklyn, Queens
2	Italian Restaurant	Pizza Place	Staten Island, Manhattan
3	Italian Restaurant	Asian Restaurant	Staten Island
4	Pizza Place	Fast Food Restaurant	Bronx, Brooklyn
5	Pizza Place	Italian Restaurant	Staten Island, Queens
6	Pizza Place	Chinese Restaurant	Queens, Brooklyn
7	Chinese Restaurant	Pizza Place	Queens, Bronx

Conclusion

- Project is aimed at exploring the boroughs of New York City to locate the most common food options in different neighbourhoods.
- Data is collected from ‘New York City Dataset’, processed and cleaned through different python libraries. K-mean clustering algorithm is used to segment the data into clusters.
- After analysis, we find out that Pizza is the most common food.
- Keeping in view the ubiquitous nature of this and other fast foods, it is set aside, and conclusion is drawn on the basis of most common foods next to pizza.
- American, Italian, Caribbean, Chinese and Asian cuisines are most common in the city.