[Problem Statement 1](#_Toc25250217)

[Data Collection 2](#_Toc25250218)

[Data Preprocessing 4](#_Toc25250219)

[Visualization 5](#_Toc25250220)

[Conclusion 10](#_Toc25250221)

This article follows the basic methods that a normal data science project would follow. These are the steps that I took:

* Problem Statement
* Data Collection
* Data Preprocessing
* Visualization
* Machine Learning

# Problem Statement

To check the common food cuisines between Toronto, Canada and New York City, USA.

We will explore the NY and Toronto dataset to see

1. How does the types of cuisines in Toronto and New York vary?
2. Which Cuisines are popular in Toronto and New York?

Trying to find answers to above mentioned question in the Jupiter notebook.

Target audience will be one who has recently shifted in this city OR Existing locality who can get benefit to explore more cuisines in neighborhood

The problem statement for my project was:

‘**Identifying the relation between food in Toronto, Canada and New York City, USA.**’

# Data Collection

**New York Location Dataset** will be collected from web (NYU Spatial Data Repository) which is available in Json format. In Python it will be converted to CSV format. Data will include below columns:

1. Borough

2. Neighborhood

3. Latitude

4. Longitude

**Toronto Location Dataset** Like NY location dataset Toronto dataset will be build with same columns. Dataset will be same as used in previous project for clustering and segmentation. Data will include below columns:

1. Postal Code

2. Borough

3. Neighborhood

4. Latitude

5. Longitude

**Used Toronto Dataset from Module 3 of the project.**

To build a dataset used Toronto dataset used in Week 3 of Capstone project.

While web scraping found NY Data Repository **(**[**link**](https://geo.nyu.edu/catalog/nyu_2451_34572)**)** **Luckily, this dataset exists for free on the web.**

A screenshot of a cell phone

Description automatically generated

CSV needed for Toronto and New York Dataset with boroughs and neighborhoods, latitude and longitude details. Data set for Toronto was provided as a previous assignment that I did in the course. To build this project used **Jupiter Notebook and Python 3.7.3**

Exported Json File data into CSV file for Toronto and NY data set and merging Latitude and longitude data from another NY data repository. the final dataset is generated, which can be seen below.

**Canada (Toronto Dataset)**

A screenshot of a cell phone

Description automatically generated

**NY Dataset**

A screenshot of a cell phone

Description automatically generated

# Data Preprocessing

Now that Data is merged and saved in CSV imported all required packages and started data preprocessing for Toronto Dataset and New York Dataset.

**Approach taken**

- **Folium** is a great visualization library. We have a provision to click on each circle mark to reveal the name of the neighborhood and its respective borough.

- Used Folium to build Toronto map which is interactive map and passing latitude and longitude to city. (same has been done for NY dataset)

- Map of Toronto, with markers of different neighborhoods.

Now next task was to find restaurants in each neighborhood for which Foursquare API was used - In order to use the Foursquare API, first create an account with foursquare site. **(**[**link**](https://developer.foursquare.com/)**)**

**Client Secret, Client ID** and **Version is generated this will be required in code to pass in as argument (explained in Notebook)**

![A screenshot of a cell phone

Description automatically generated]()

Finally using the Foursquare API created datasets, with common restaurants in Downtown Toronto

A close up of text on a white background

Description automatically generated

Same approach was used for New York Data. Here Queens is used as a Borough for analysis. The entire procedure is exactly same as that of Downtown Toronto.

# Visualization

To conclude compared Downtown Toronto and New York restaurants from data frame. In the above table we can see just the 1st Most Common and so on. But we don’t know count of how many of those 1st most common restaurants are present. Hence , used a **count plot** from **Seaborn** package to visualize it.

From the graphs it is clear that Thai Restaurant is 1st Most Common Restaurant in Downtown Toronto Area. However, we can also see another value named just ‘Restaurant’. This can be considered as an inconsistency as we don’t know what kind of food that restaurant serves. We can assume that this restaurant serves majority of cuisines. In the similar way visualizations are generated for Toronto Data.

**Most Common Restaurants in Queens NY**

We can see that Restaurant, likewise if Toronto graph this can be considered as an inconsistency as we don’t know what kind of food that restaurant serves. We can assume that this restaurant serves majority of cuisines. Vietnam and Italian Restaurants lead the 1st Most Common Restaurants in Queens NY.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated