

|  |
| --- |
| Capstone Project- Where to set up a Restaurant in Downtown Toronto?  2021 |
|  |
| **January 9**  Authored by: Venkatesh Saripalli |

Capstone Project- Where to set up a Restaurant in Downtown Toronto?

The Coursera IBM Data Science Professional Certificate course provides a hands-on understanding of Data science, Machine Learning using Cloud Tools: Jupyter / JupyterLab, GitHub, R Studio, and Watson Studio, API such as Foursquare and Libraries: Pandas, NumPy, Matplotlib, Seaborn, Folium, ipython-sql, Scikit-learn, ScipPy, etc.

The Capstone project is final project which carries the highest weightage.

The Project Summary-

The following sections will be covered in the Project Report.

Introduction/Business Problem- Introduction is the section to discuss the business problem and who would be interested in this project.

Data- This section describes the data that will be used to solve the problem and the source of the data.

Methodology- This section represents the main component of the report to discuss and describe any exploratory data analysis, inferential statistical testing performed, and Machine learnings used and why.

Results- This is section to discuss the results.

Discussion- This is section to discuss any observations you noted and any recommendations you can make based on the results.

Conclusion- This is section to conclude the report.

1. Introduction/Business Problem- A description of the problem and a discussion of the background.

An entrepreneur wants to open a restaurant in the Downtown Toronto area. There are several Neighborhoods in the Downtown Toronto area. The ideal location will be the one that already has different types of restaurants and also satisfies other business criteria such as the target customers, type of the restaurant and finances available. The objective is to come up with the top 10 types of Venues in each neighborhood. This analysis will help any entrepreneur to determine the ideal location depending on the target customers, type of the restaurant and finances available for the business that is proposed to be opened.

2. Data- A description of the data and how it will be used to solve the problem.

The purpose of data collection is to enable to decide the top10 Establishments/ Venues in each neighborhood. So, a getting a list of the Venues for each neighborhood in the Downtown Toronto area will enable us to find the top 10 Venues. The following data is identified as necessary to help in achieving the business objective.

The list of postal code and Neighborhoods of Canada along with the Latitude and Longitude- This data will help in using the latitude and longitude to retrieve list of venues in each Neighborhood.

a. The Wikipedia page has a list of Canada postal codes and Boroughs and Neighborhoods. This data will have to be scrapped from the web page. This data will be cleaned up and formatted using one hot encoding techniques and pandas string replacement functions.

b. Retrieve / Get the list of Geo coordinates (Latitude and Longitude) for each postal code. Alternatively, use a csv file that has all the Geo coordinates. This data will be merged with the data retrieved in step1.

The list of venues for each neighborhood- This data will help in using Machine Learning to find patterns in the data by Segmentation and Clustering.

a. Use the **Foursquare** API to get the list of Venues for each Latitude and Longitude.

3. Methodology- Discuss and describe any exploratory data analysis, inferential statistical testing and Machine learning techniques used.

The Wikipedia page has a list of Canada postal codes and Boroughs and Neighborhoods. This data will have to be scrapped from the web page. This data will be cleaned and formatted using one hot encoding techniques and pandas string replacement functions.

Retrieve / Get the list of Geo coordinates (Latitude and Longitude) for each postal code. Alternatively, use a csv file that has all the Geo coordinates. This data will be merged with the data retrieved in step1.

Retrieve a list of Venues in each neighborhood. Use the **Foursquare** API to get the list of Venues for each Latitude and Longitude. A personal account was opened which will enable to get the data required for this purpose.

Construct a normalized Data frame by converting the Categorical Values of the Venue categories to columns having numeric values.

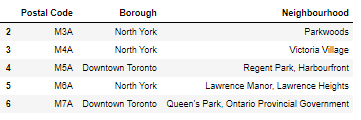
Use Segmentation and Clustering Technique to find the clusters which are similar. **Kmeans** clustering will be used for this purpose as this is one of the effective techniques available for this purpose.

Visualize the clusters by plotting a map. Folium library is a very good library available for this purpose.

These clusters will let the Entrepreneur know the primary activity in that cluster and help him to decide the location. For ex. if the Entrepreneur wants to service office employees, the restaurant can be opened in a neighborhood which has business activity.

### Scrape the data from the Wikipedia URL and create the **pandas** data frame with the Canada postal code and neighborhood data.

# Data frame with Canada postal codes.

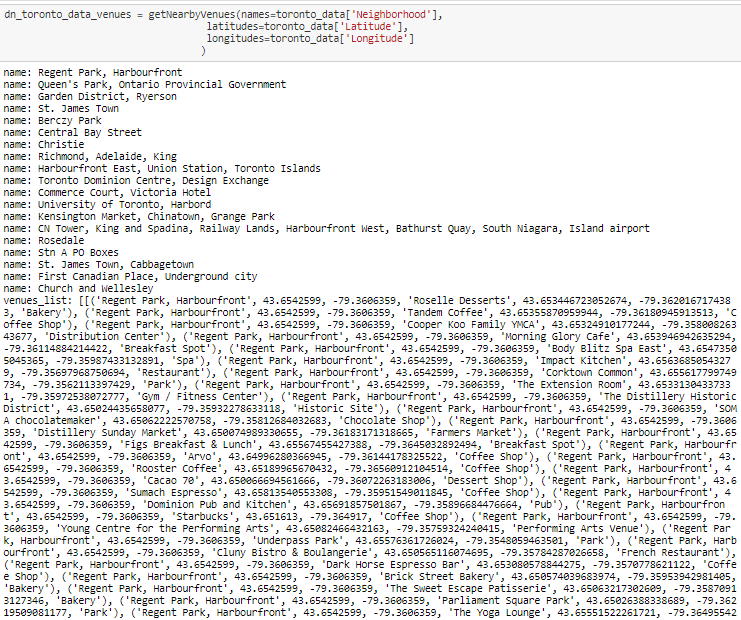


# Data frame with latitude and longitudes of Canada postal codes.

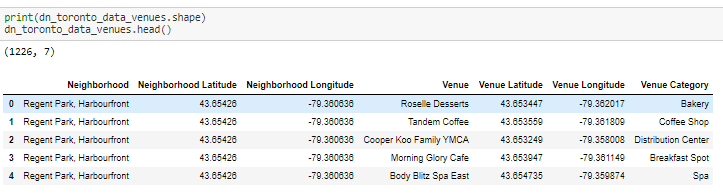


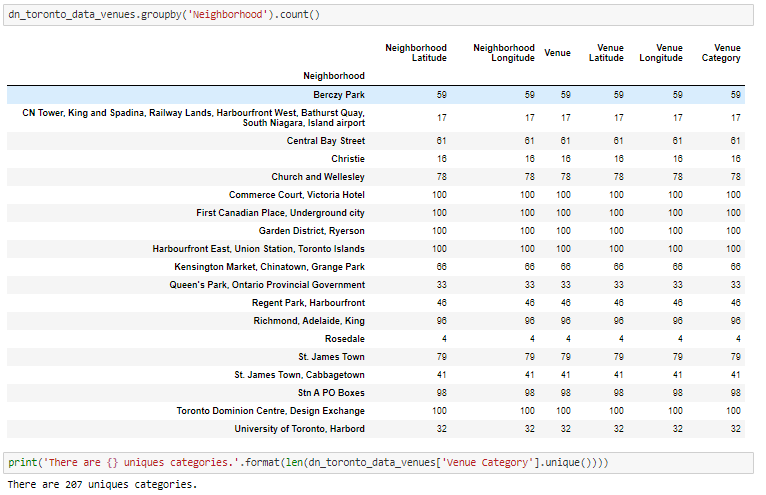
# Get the Venue details of Toronto neighborhoods.





# Data frame with latitude and longitudes and Venue details of Toronto Down town Neighborhoods.





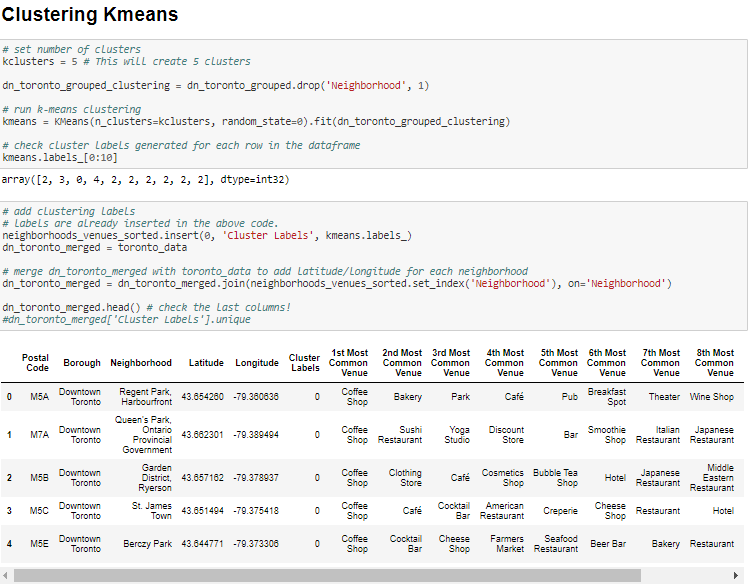
# Data frame with Top10 Venue details of Toronto Down town Neighborhoods.

# 

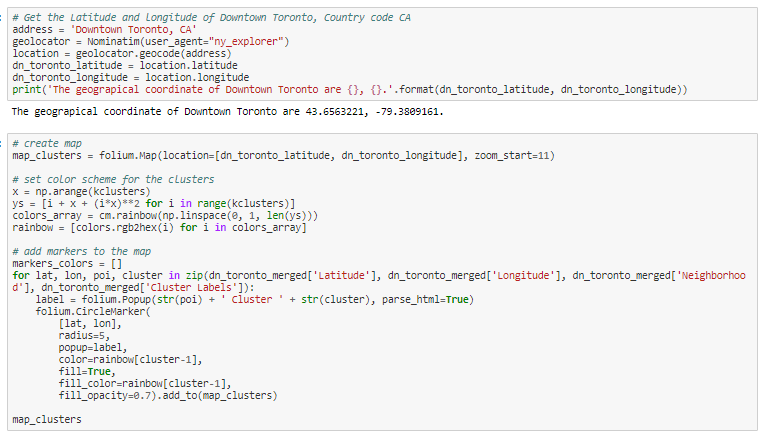
# 

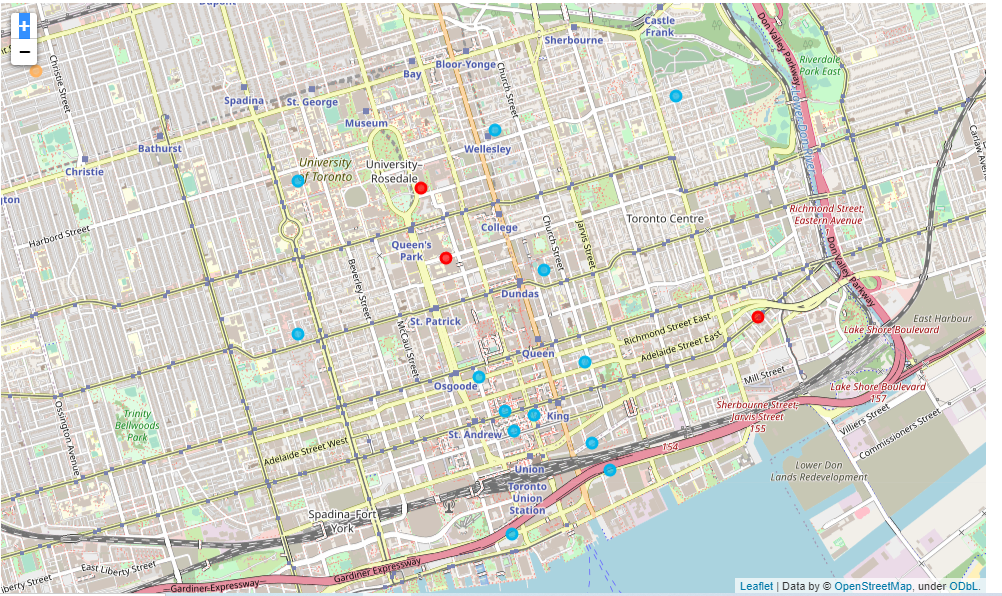
# 

# Clustering Toronto Down town Neighborhoods by using **Kmeans** technique.



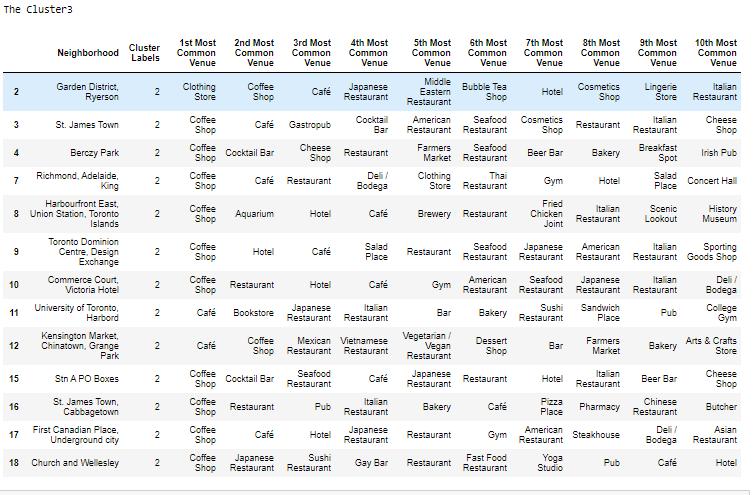
# Map of the Clusters:

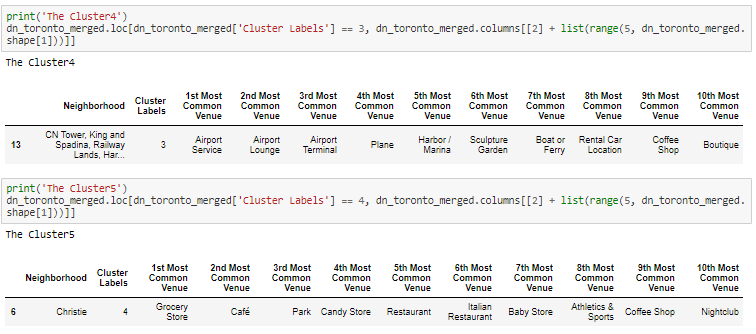




# Print the Clusters:

# 





# 4. Results-

The Results display the top 10 most common venues for each Neighborhood. As can be seen from the results the Cluster3 has restaurants as all the top10 venues. We used K-means clustering to group the Neighborhoods in to Clusters. The map shows that the Cluster3 (blue color circles) has Neighborhoods spread out throughout the Downton Toronto area. The Results were generated by using data on 1/9/2021.

# 5. Discussion-

 As can be seen from the results the Cluster3 has restaurants as all the top10 venues. There are lot of neighborhood choices in Cluster3 to open a restaurant. The other factors that would affect the decision are:

The Type/ Cuisine of the restaurant to be opened.

The customer profile that the restaurant wants to serve, for ex if the restaurant wants to target college students, Cluster4 which has university students will be a suitable location.

The budget of the business, this means factors such as rent and other costs have to be considered.

Finally, some other factors affecting the decision would be familiarity of the Entrepreneur with the Neighborhood, closeness of the Neighborhood for the Entrepreneur.

# 6. Conclusion-

The Report defines the Business problem and then identifies the data necessary that will help in making Recommendations. The analysis is done by formatting and narrowing the data down to top10 venues for each neighborhood and using K-means clustering which enables the identification of similar Neighborhoods. These Results and also other factors mentioned in the Discussion section will be considered in making a Recommendation. The Recommendation will definitely help the Entrepreneur in taking a good decision about the type of Restaurant and the location considering other factors such as the finances available and sets the entrepreneur on the path to **Success!**