THE BEST NEIGHBOURHOOD

Introduction

This is a business proposal to choose the best neighbourhood. The project helps in exploring different criteria while choosing a neighbourhood to live in. This helps people in making smart decisions to select a place to live.

With people migrating to Toronto increasing, better places to live keep decreasing. While moving to a place, there are many factors to consider. This project aims to help in considering those factors while choosing a better place to live. Some of the factors could be shops, schools, malls, food stores, convenience stores, theatres, market etc.

This is a comparative analysis project, that compares neighbourhoods in Toronto and produces a statistical analysis for better understanding. It can be applied to any other place on this planet.

Problem:

The problem that is tried to solve in this project is:
1) Choosing a better neighbourhood to live.
2) Places to visit in a city or state.
3) Rating of various places like plazas, schools etc.

Location:

Downtown, Toronto

Downtown Toronto is the main central business district of Toronto, Ontario, Canada. It is a buzzing area filled with skyscrapers, restaurants, nightlife, and an eclectic mix of neighbourhood. It's also home to iconic attractions like the CN Tower, St. Lawrence Market, and the Royal Ontario Museum, with exhibits on natural history.

Foursquare:

- It is an API
- It is the most trusted, independent location data platform for understanding how people move through the real world.
- This project used Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

Clustering

To compare the similarities of two cities, to explore neighbourhood, segmenting them, and grouping them into clusters is required. It is used to find similar neighbourhood in big cities like New York and Toronto. To be able to do that, clustering data is done, which is a form of unsupervised machine learning: k-means clustering algorithm

Platform - Jupyter Notebook

Libraries used:

Pandas: For creating and manipulating dataframes.

Folium: Python visualization library would be used to visualize the neighbourhoods cluster distribution of using interactive leaflet map.

Scikit Learn: For importing k-means clustering.

JSON: Library to handle JSON files.

XML: To separate data from presentation and XML stores data in plain text format.

Geocoder: To retrieve Location Data.

Beautiful Soup and Requests: To scrap and library to handle http requests.

Matplotlib: Python Plotting Module.

Data Description

This is a comparative analysis project, that compares neighbourhoods in Toronto and produces a statistical analysis for better understanding. It can be applied to any other place on this planet.

Data Source:

The Dataset for this project is obtained from Wikipedia.
Link : https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Data Cleaning:

Data downloaded and scraped from multiple sources were combined into one table. There were a lot of missing values that were removed before analysing. Outliers were checked before proceeding to next step.

Foursquare Data API

Foursquare location data is used to solve the problem.
Data about different venues in different neighbourhood of that specific borough is required. In order to gain that information "Foursquare" locational information is used. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighbourhood
2. Neighbourhood Latitude
3. Neighbourhood Longitude
4. Venue
5. Name of the venue e.g. the name of a store or restaurant
6. Venue Latitude
7. Venue Longitude
8. Venue Category

Methodology:

- First, the data is obtained from the wiki page, scrapped and transformed into a dataframe using pandas.

- Then, the data is modified and cleaned according to the analysis. Outliers are removed. Some columns were dropped.

- The columns latitude and longitude are added to the dataset.

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M1B\n	Scarborough\n	Malvern, Rouge	43.81139	-79.19662
1	M1C\n	Scarborough\n	Rouge Hill, Port Union, Highland Creek	43.78574	-79.15875
2	M1E\n	Scarborough\n	Guildwood, Morningside, West Hill	43.76575	-79.17470
3	M1G\n	Scarborough\n	Woburn	43.76812	-79.21761
4	M1H\n	Scarborough\n	Cedarbrae	43.76944	-79.23892
5	M1J\n	Scarborough\n	Scarborough Village	43.74446	-79.23117
6	M1K\n	Scarborough\n	Kennedy Park, Ionview, East Birchmount Park	43.72582	-79.26461
7	M1L\n	Scarborough\n	Golden Mile, Clairlea, Oakridge	43.71289	-79.28506
8	M1M\n	Scarborough\n	Cliffside, Cliffcrest, Scarborough Village West	43.72360	-79.23496
9	M1N\n	Scarborough\n	Birch Cliff, Cliffside West	43.69510	-79.26466

- Then, the latitude and longitude coordinates of Toronto are found using geocoder.

```
def get_latilong(postal_code):
    lati_long_coords = None
    while(lati_long_coords is None):
        g = geocoder.arcgis('{}, Toronto, Ontario'.format(postal_code))
        lati_long_coords = g.latlng
    return lati_long_coords

get_latilong('M4G')
```

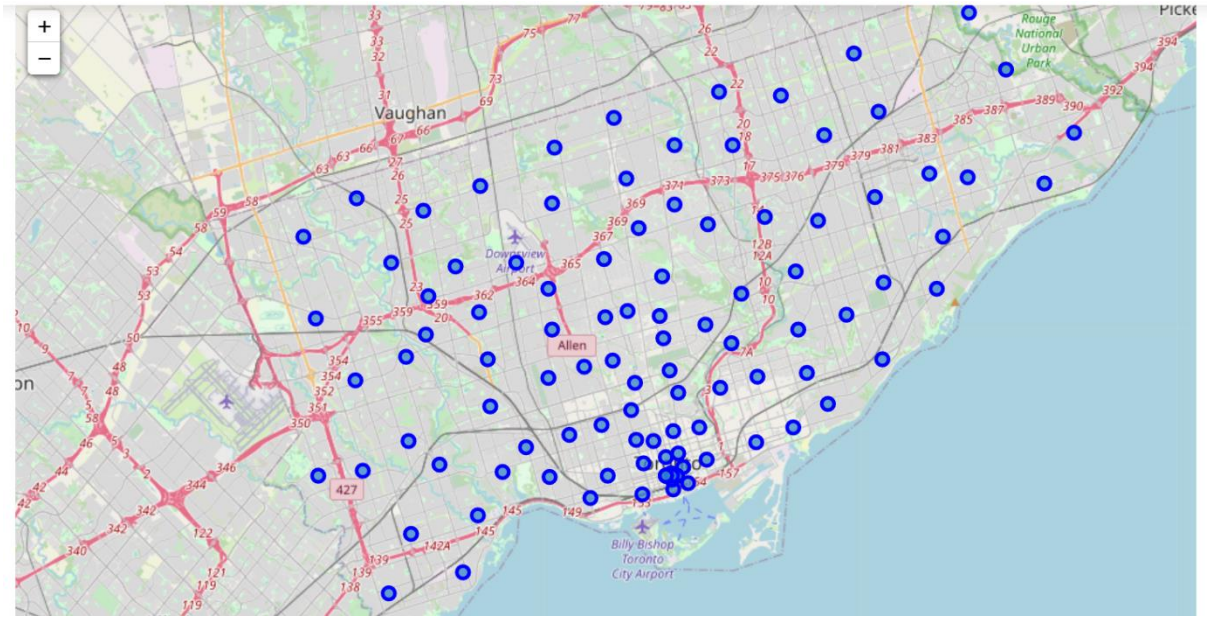
```
[43.709020000000066, -79.36348999999996]
```

- The geographical coordinates of “Downtown, Toronto” are determined using Nominatim.

```
address = 'Downtown Toronto'

geolocator = Nominatim(user_agent="geoapiExercises")
location = geolocator.geocode(address)
latitude_x = location.latitude
longitude_y = location.longitude
print('The Geograpical Co-ordinates of Downtown, Toronto are {}, {}.'.format(latitude_x, longitude_y))
```

```
The Geograpical Co-ordinates of Downtown, Toronto are 43.6563221, -79.3809161.
```

- A map of Downtown is obtained using Folium.



- Nearby venues to Downtown were determined.

	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	UNIQLO ユニクロ	[{'id': '4bf58dd8d48988d103951735', 'name': 'C...	43.655910	-79.380641
1	Silver Snail Comics	[{'id': '52f2ab2ebcbc57f1066b8b18', 'name': 'C...	43.657031	-79.381403
2	Ed Mirvish Theatre	[{'id': '4bf58dd8d48988d137941735', 'name': 'T...	43.655102	-79.379768
3	Blaze Pizza	[{'id': '4bf58dd8d48988d1ca941735', 'name': 'P...	43.656518	-79.380015
4	Yonge-Dundas Square	[{'id': '4bf58dd8d48988d164941735', 'name': 'P...	43.656054	-79.380495

- Different categories of venue could be found.

```
a=pd.Series(nearby_venues.categories)
a.value_counts()[:15]
```

```
5]: Coffee Shop               12
    Hotel                      4
    Japanese Restaurant        3
    Sandwich Place             2
    Restaurant                 2
    Bubble Tea Shop            2
    Bookstore                  2
    Diner                      2
    Furniture / Home Store     2
    Electronics Store          2
    Sushi Restaurant           2
    Seafood Restaurant         2
    Department Store           2
    Burger Joint               2
    Theater                    2
    Name: categories, dtype: int64
```

```
print('There are {} Uniques Categories.'.format(len(Downtown_venues['Venue Category'].unique())))
Downtown_venues.groupby('Neighborhood').count().head()
```

```
There are 306 Uniques Categories.
```

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt	23	23	23	23	23	23
Alderwood, Long Branch	9	9	9	9	9	9
Bathurst Manor, Wilson Heights, Downsview North	12	12	12	12	12	12
Bayview Village	5	5	5	5	5	5
Bedford Park, Lawrence Manor East	24	24	24	24	24	24

Modelling using Machine Learning:

A model is built to determine the top venues, different and number of venues in the neighbourhood and around the neighbourhood. The rating of the venues along with people's reviews could be obtained using this model.

Machine Learning algorithm:

There are different ways an algorithm can model a problem based on its interaction with the experience or environment or whatever we want to call the input data.

1. Supervised Learning

2. Unsupervised Learning.

In this project, as the data is not labelled and does not have a known result, the unsupervised learning type of machine learning algorithm is used to model the data.

Unsupervised Learning:

A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may be through a mathematical process to systematically reduce redundancy, or it may be to organize data by similarity.

Example: K-means.

Clustering Algorithms

Clustering methods are typically organized by the modelling approaches such as centroid-based and hierarchal. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonality.

- Example: k-Means, Hierarchical Clustering.

Categorical data are variables that contain label values rather than numeric values. Each value represents a different category. Some algorithms can work with categorical data directly. Many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric. This means that categorical data must be converted to a numerical form. If the categorical variable is an output variable, you may also want to convert predictions by the model back into a categorical form in order to present them or use them in some application.

One-Hot Encoding

For categorical variables where no such ordinal relationship exists, the integer encoding is not enough. Using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results. A one-hot encoding can be applied to the integer representation. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.

One-hot Encoding for conversion of categorical variables

```
Downtown_onehot = pd.get_dummies(Downtown_venues[['Venue Category']], prefix="", prefix_sep="")

# adding neighborhood column back to dataframe
Downtown_onehot['Neighborhood'] = Downtown_venues['Neighborhood']

# moving neighborhood column to the first column
fixed_columns = [Downtown_onehot.columns[-1]] + list(Downtown_onehot.columns[:-1])
Downtown_onehot = Downtown_onehot[fixed_columns]
Downtown_grouped = Downtown_onehot.groupby('Neighborhood').mean().reset_index()
Downtown_onehot.head(5)
```

0]:

	Zoo Exhibit	Accessories Store	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	A Dealers
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	

Top venues in the neighbourhood is determined.

```
num_top_venues = 5
for hood in Downtown_grouped['Neighborhood']:
    print("---- "+hood+" ----")
    temp = Downtown_grouped[Downtown_grouped['Neighborhood'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

```
---- Agincourt ----
            venue  freq
0   Shopping Mall  0.09
1        Pharmacy  0.04
2    Skating Rink  0.04
3  Clothing Store  0.04
4   Grocery Store  0.04


---- Alderwood, Long Branch ----
            venue  freq
0      Print Shop  0.11
1     Pizza Place  0.11
2     Gas Station  0.11
3        Pharmacy  0.11
4  Sandwich Place  0.11
```

Most common venues were found.

	Neighborhood	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	Agincourt	Shopping Mall	Bubble Tea Shop	Breakfast Spot	Sushi Restaurant	Latin American Restaurant	Chinese Restaurant	Bank	Bakery	Shanghai Restaurant	Clothing Store
1	Alderwood, Long Branch	Print Shop	Gym	Pub	Pizza Place	Coffee Shop	Pharmacy	Sandwich Place	Gas Station	Convenience Store	Hardware Store
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Deli / Bodega	Fried Chicken Joint	Park	Sandwich Place	Sushi Restaurant	Restaurant	Middle Eastern Restaurant	Pizza Place	Mediterranean Restaurant
3	Bayview Village	Flower Shop	Gas Station	Park	Trail	Asian Restaurant	Electronics Store	Elementary School	Escape Room	Ethiopian Restaurant	Event Space
4	Bedford Park, Lawrence Manor East	Pizza Place	Sandwich Place	Coffee Shop	Italian Restaurant	Comfort Food Restaurant	Greek Restaurant	Liquor Store	Juice Bar	Thai Restaurant	Restaurant

K-means clustering:

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

K-means:

```
# Using K-Means to cluster neighborhood into 3 clusters
Downtown_grouped_clustering = Downtown_grouped.drop('Neighborhood', 1)
kmeans = KMeans(n_clusters=3, random_state=0).fit(Downtown_grouped_clustering)
kmeans.labels_
```

34]:
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0,
       2, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 2])
```

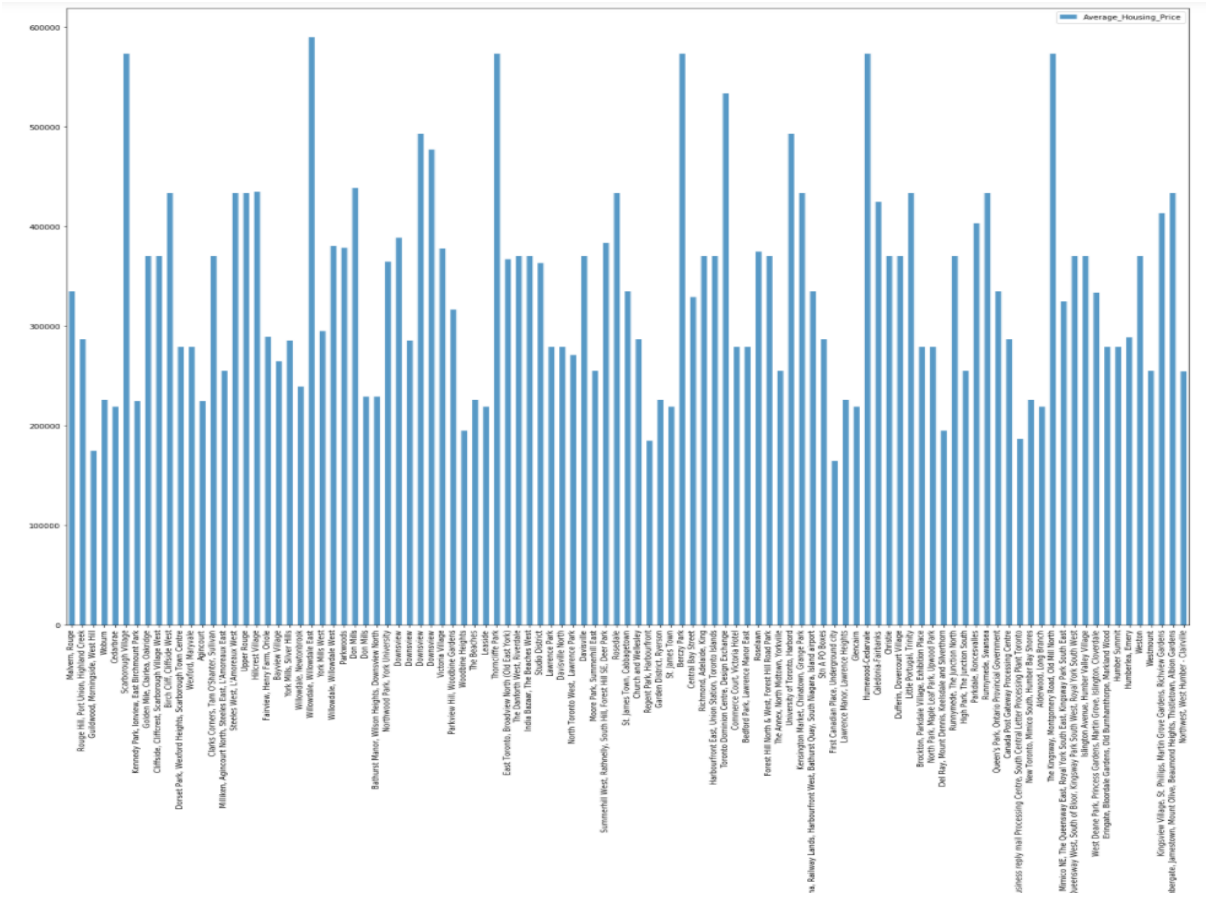
```
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

Downtown_merged =df_2.iloc[:16,:]

# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
Downtown_merged = Downtown_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

Downtown_merged.head()# check the last columns!
```

35]:

	Postalcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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 M Comn Ver
0	M1B\n	Scarborough\n	Malvern,	43.81139	-79.19662	0	Zoo	Fast Food	Paintball	Flea	Fish	Fish & Chips	Filip

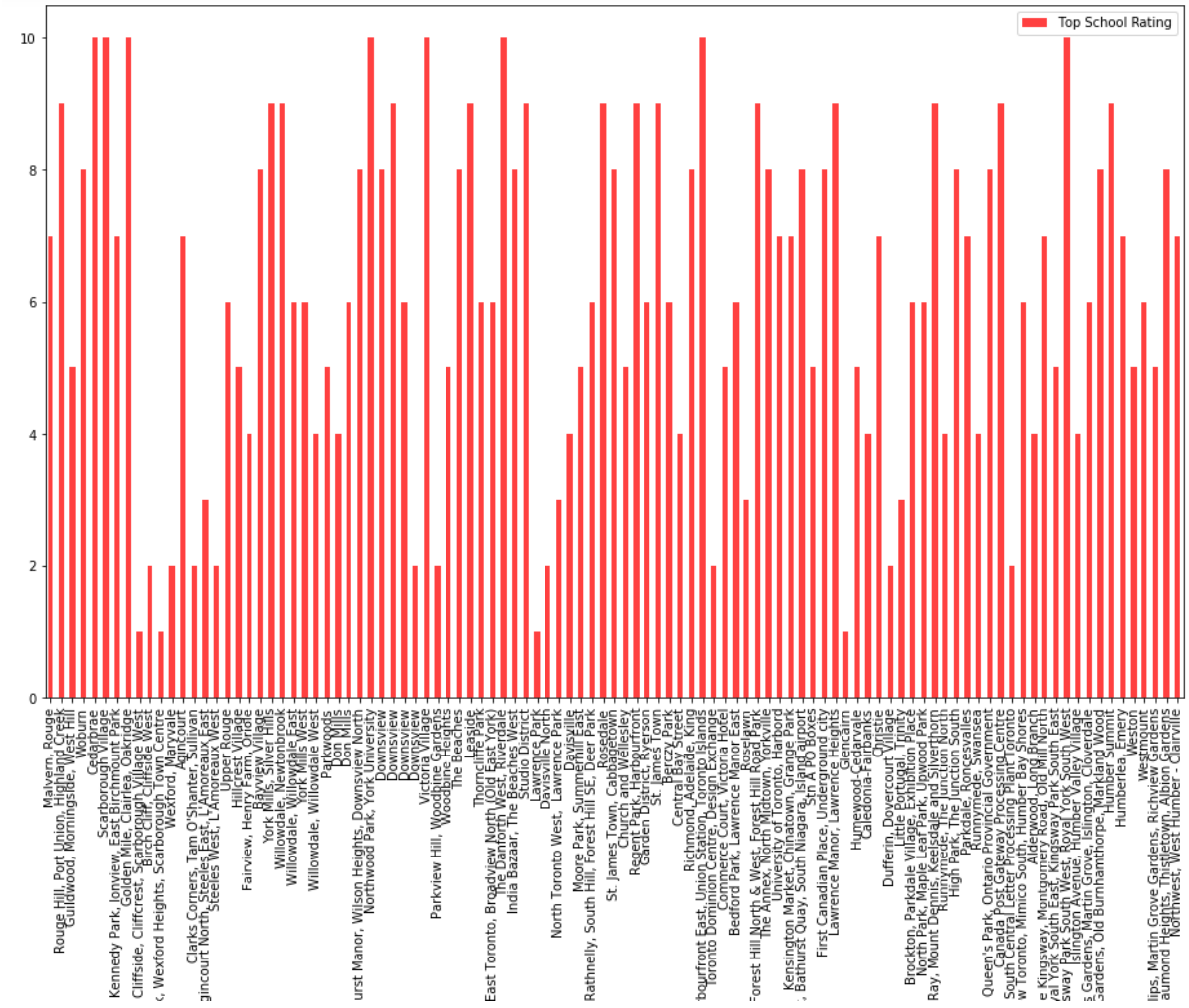
Relationship between Average housing price and Neighbourhood:

To easily understand the relationship a plot between Neighbourhood and average housing price is built.



From the above graph, we could interpret the effect of neighbourhood on housing prices. From the data, the housing prices in Willowdale West is higher than any other area in Downtown.

Relationship between Schools and Neighbourhood:



From the above graph, we could tell that around 8 schools in Downtown have same rating. Most of the schools in Downtown are popular and very few schools have less rating.

Conclusion:

Through this project, it can be concluded that when data is available, it could be used in many ways. In this project the data of Toronto, helped in finding a better neighbourhood in Downtown. The description of venues, housing prices and Schools in Downtown helped in analysing better neighbourhood. The Machine Learning algorithms helped in improvising and predicting data. This is just an example of how useful are the data science concepts, open source data, python libraries, machine learning algorithms in producing statistical results for analysing data.

KRISHNA CHARITHA MANTRIPRAGADA

mantripragadacharitha@gmail.com

LinkedIn: https://www.linkedin.com/in/charitha-mantripragada-3ba179179/

GitHub: https://github.com/charitha-m20