BATTLE OF THE NEIGHBORHOODS

Ran Zou (Karen)
Final presentation of IBM capstone

INTRODUCTION AND BUSINESS PROBLEM

- People accepting job offers in a new city can be overwhelmed by the variety of choices to make regarding housing. All large cities have thriving neighborhoods with a plethora of venues for entertainment, health and well being, education and so on. We need a recommendation engine based on personal preferences and business ratings to help folks choose where to move.

- In this project I'm going to analyze and compare 2 of Toronto's largest Boroughs. I'll be using Foursquare APIs and leverage details about the venues to contrast and compare neighborhoods and ultimately make a recommendation.

DATA AND TOOLS

- For this project, I'll leverage data obtained from a couple of websites. This includes a list of Boroughs, Neighborhoods and their corresponding Latitude and Longitude coordinates. These will be used as input to the Foursquare APIs to collect a list of Venues for the neighborhoods and the venues will be used for API calls to collect details about the venues. The data will then be processed, analyzed and a recommendation will be made.
- Here is a list of websites leveraged for data collection:
    - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
    - http://cocl.us/Geospatial_data
    - https://api.foursquare.com/v2/venues/explore
- Additionally, a variety of Python packages will be used to analyze and display the data to gain insights and eventually provide a recommendation.

METHODOLOGY

- Business Understanding
- Data Collection and Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

BUSINESS UNDERSTANDING

- My character for which I'm developing this solution is someone who got a new job and needs to relocate to an unfamiliar city. With so many choices of venues we need to be able to filter down the neighborhoods according to what the user desires and what the city has to offer. Our character will be asked to provide a top list of desired venues the neighborhood they want to move in would have to have available, along with how strong their preference is. Based on these facts, we'll provide a recommendation of one or more neighborhoods to chose from to move in to.

DATA COLLECTION AND UNDERSTANDING

- First the borough and neighborhood data will be retrieved from Wikipedia. Parsing the page will leverage the BeautifulSoup parsing package. Data will then be organized into a data frame.

- Next, retrieve the latitude and longitude coordinates from the geo special data and merge with the existing data frame.

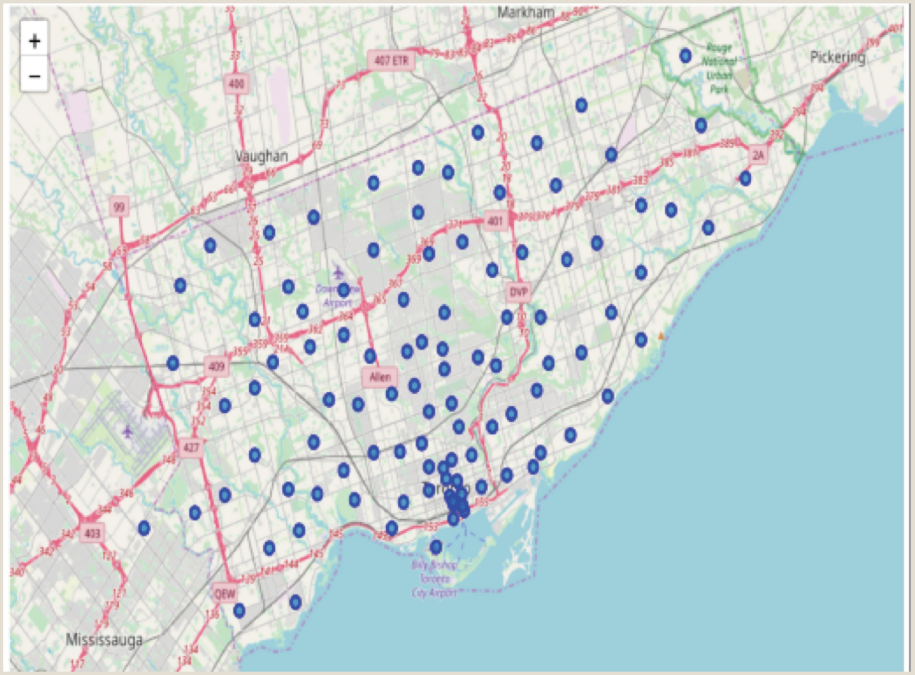
- Finally, for each neighborhood, leverage the Foursquare APIs to retrieve a list of venues.

DATA PREPARATION

- After data collection, we need to prepare the data. We notice that there is both missing and incomplete data in the list of boroughs and neighborhoods. Therefore data cleaning is necessary. We'll remove rows for which the boroughs are "not assigned" and also set the missing neighborhood names to match their corresponding borough names for those rows.
- Most machine learning algorithms utilize only numeric values so our string data needs to be encoded. We'll use one-hot encoding to transform venue data into a 1 and 0 sparse matrix to feed into the K-means clustering algorithm. Before doing that though, we'll group the data by neighborhoods.

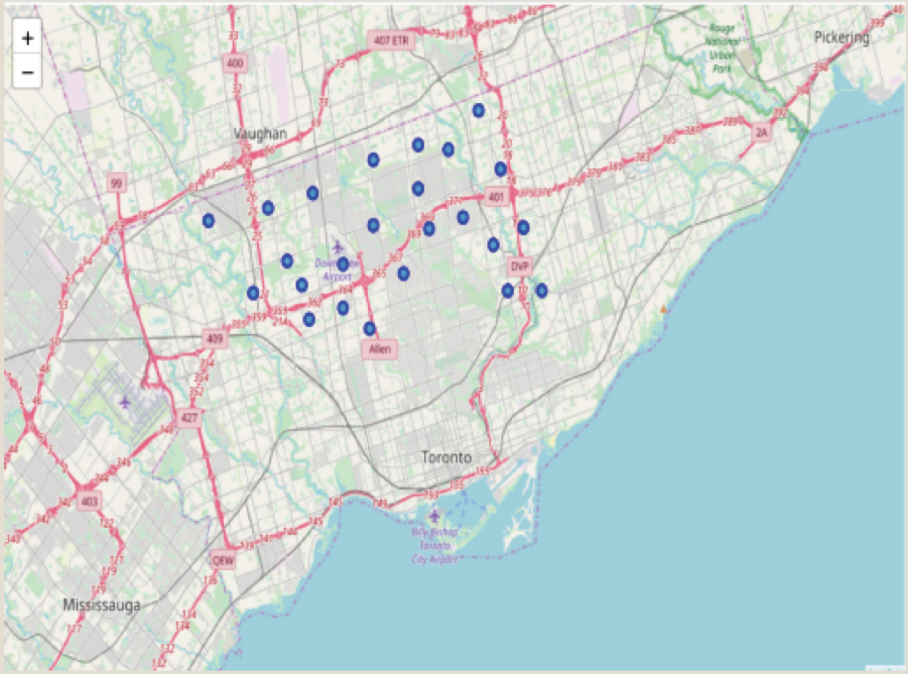
	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M2H	North York	Hillcrest Village	43.803762	-79.363452
1	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556
2	M2K	North York	Bayview Village	43.786947	-79.385975
3	M2L	North York	Silver Hills, York Mills	43.757490	-79.374714
4	M2M	North York	Newtonbrook, Willowdale	43.789053	-79.408493
5	M2N	North York	Willowdale South	43.770120	-79.408493
6	M2P	North York	York Mills West	43.752758	-79.400049
7	M2R	North York	Willowdale West	43.782736	-79.442259
8	M3A	North York	Parkwoods	43.753259	-79.329656
9	M3B	North York	Don Mills North	43.745906	-79.352188
10	M3C	North York	Flemingdon Park, Don Mills South	43.725900	-79.340923

Toronto data frame

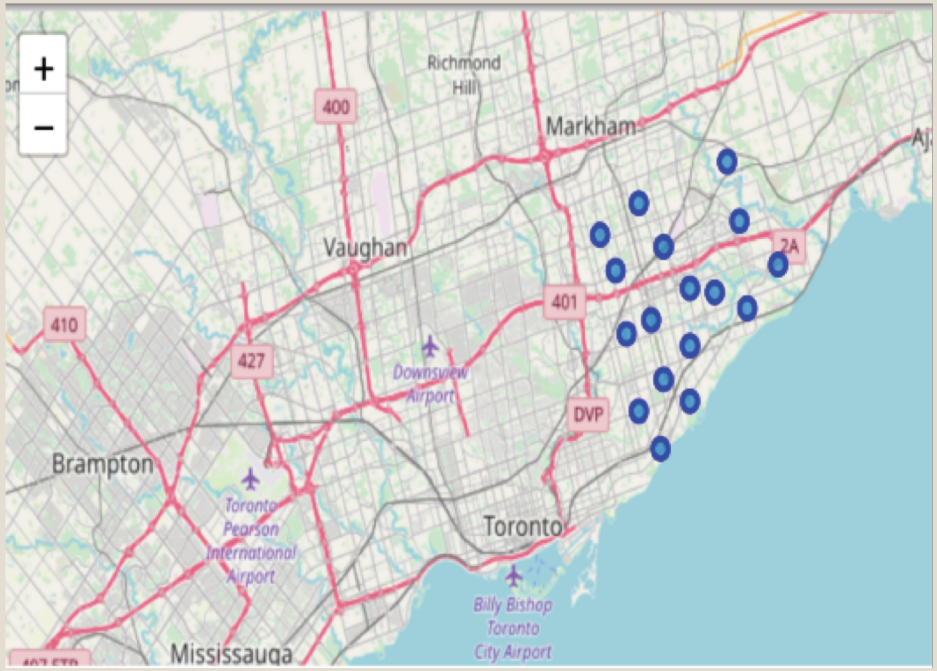


Neighborhoods on map via Folium

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M2H	North York	Hillcrest Village	43.803762	-79.363452
1	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556
2	M2K	North York	Bayview Village	43.786947	-79.385975
3	M2L	North York	Silver Hills, York Mills	43.757490	-79.374714
4	M2M	North York	Newtonbrook, Willowdale	43.789053	-79.408493
5	M2N	North York	Willowdale South	43.770120	-79.408493
6	M2P	North York	York Mills West	43.752758	-79.400049
7	M2R	North York	Willowdale West	43.782736	-79.442259
8	M3A	North York	Parkwoods	43.753259	-79.329656
9	M3B	North York	Don Mills North	43.745906	-79.352188
10	M3C	North York	Flemingdon Park, Don Mills South	43.725900	-79.340923



	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park	43.727929	-79.262029
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848
10	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...	43.757410	-79.273304
11	M1R	Scarborough	Maryvale, Wexford	43.750072	-79.295849
12	M1S	Scarborough	Agincourt	43.794200	-79.262029
13	M1T	Scarborough	Clarks Corners, Sullivan, Tam O'Shanter	43.781638	-79.304302
14	M1V	Scarborough	Agincourt North, L'Amoreaux East, Milliken, St...	43.815252	-79.284577
15	M1W	Scarborough	L'Amoreaux West, Steeles West	43.799525	-79.318389
16	M1X	Scarborough	Upper Rouge	43.836125	-79.205636



MODELING

- First step in the model will be to cluster neighborhoods according to most prevalent venues. We'll be using unsupervised machine learning in the form of K-means clustering.
- Next we'll compute a score for each neighborhood, based on the weighted average of the user preferences and how strong the preference is for each given venue type. The resulting scores will be sorted in descending order and the top ones will be kept.
- Finally, a recommendation will be made providing the user with a list of candidate neighborhoods and a visual aid in the form of a Folium map with markers for the selected neighborhoods.

EVALUATION

- In a real world scenario, we'd collect feedback from the user in terms of their actual choice and input that feedback into our model to further refine it. This is at the heart of the iterative process a data science project goes through.
- Additionally, other data sources would be leveraged, such as housing pricing data, population density, crime rates, traffic congestion and so on. A true recommendation engine would provide the user with multiple choices and a larger array of preferences as input to the model.
- For the purpose of the project, a simple recommendation scheme is chosen.

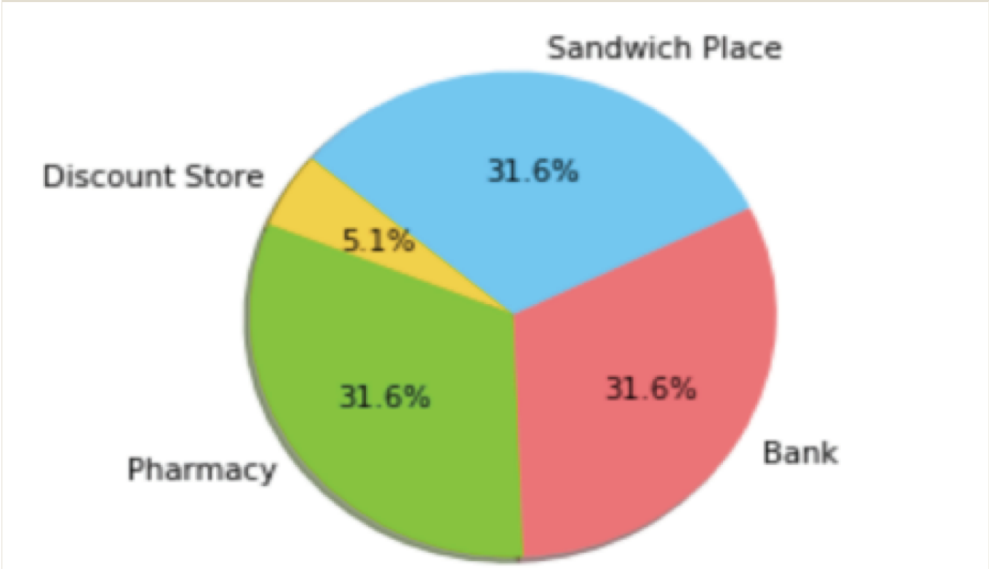
DEPLOYMENT

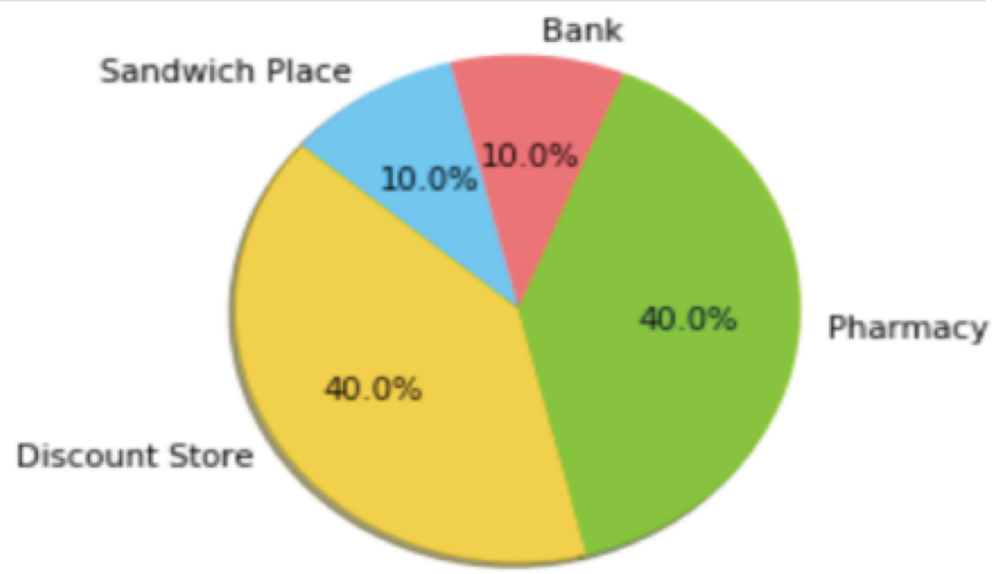
- Since this is a small project on small datasets, my deployment strategy is to use a Jupyter notebook running the various Python packages and custom code locally, on my machine.

- For larger projects and larger datasets, we could use Apache Spark to distribute the RDDs and execute code in parallel on multiple commodity hardware machines.

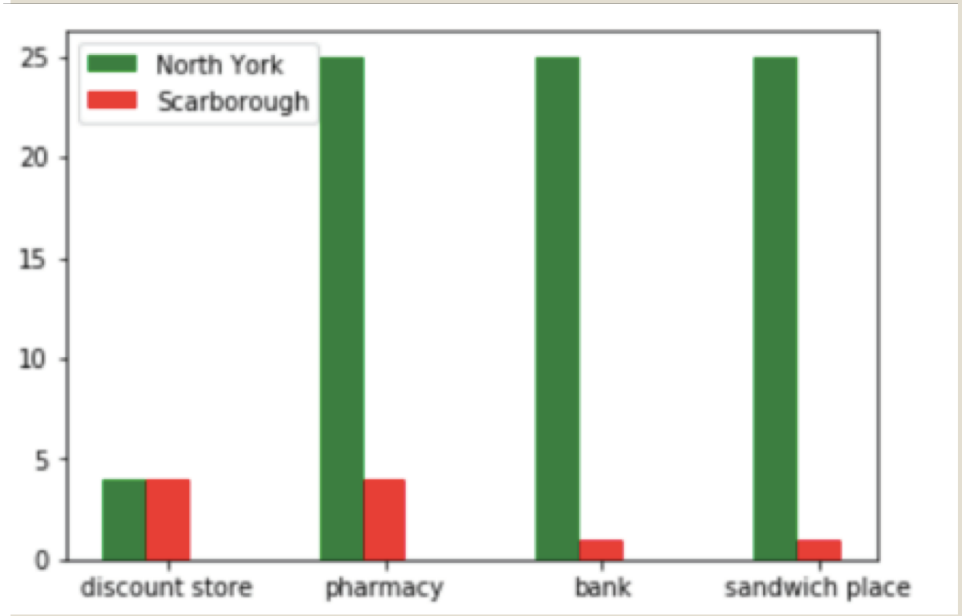
RESULTS

- At first we look at the distribution of values as percent total venues, for each Borough. However, as we'll see on the next slide, once we plug in the weights our user selected for each preference, it will become apparent which one is better.





- While it is visually obvious neighborhoods in North York are a better match according to the user preferences, we visually display to compare them. We also compute a score for each borough.




```
mvNY = showMatchingVenuesCount(northYorkVenues)
```

```
discount store, 4
pharmacy, 25
bank, 25
sandwich place, 25
```

```
mvSC = showMatchingVenuesCount(scarboroughVenues)
```

```
discount store, 4
pharmacy, 4
bank, 1
sandwich place, 1
```

```
def computeScore(di):
    s = 0
    k = 0
    for i in di:
        s = s + di[i] * preferences_weights[k]
        k = k + 1
    return round(s, 2)
```

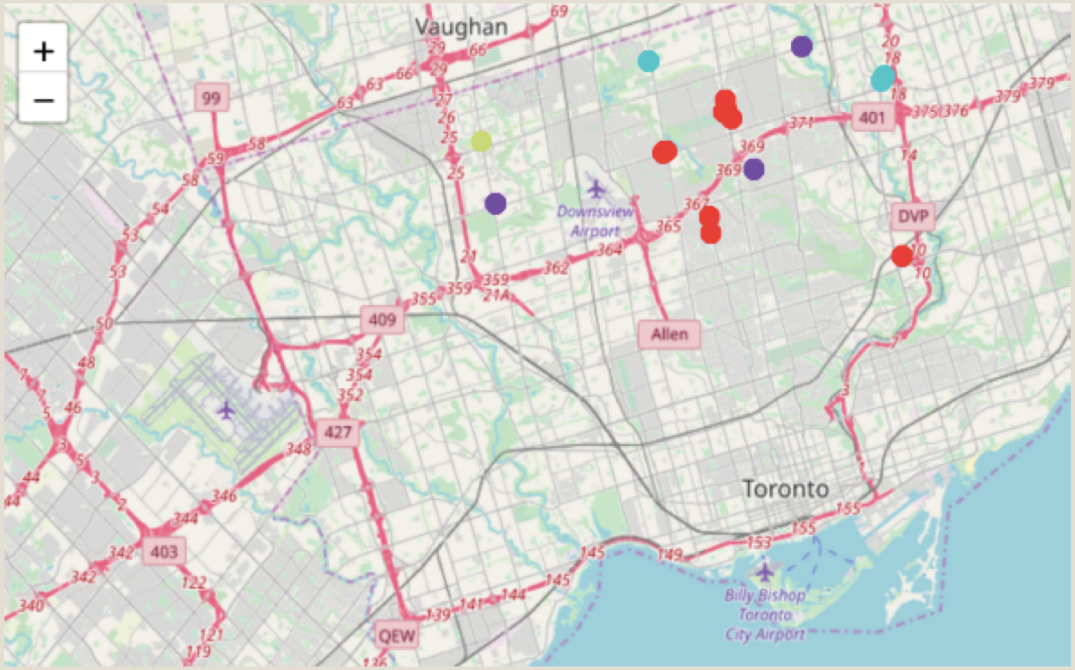
```
computeScore(mvNY)
```

```
18.7
```

```
computeScore(mvSC)
```

```
2.35
```

CONCLUSION



Based on data analysis, we recommend the user move to North York and show a visual representation of the clustered neighborhoods which have venues satisfying the chosen user preferences.