NEW YORK AND TORONTO – AN ANALOGY

SUDHINA S.

CONTENTS

METHODOLOGY

- System Of Data Science Methods Used In The Project

DATA SCIENCE METHODOLOGY

- Stage 1 – Define the Business Problem
- Stage 2- Data Acquisition, Loading and Preparation
- Stage 3 – Data Analysis
- Stage 4 – Result exploration

BUSINESS PROBLEM

- The Question That The Data Science Exercise Is Trying To Solve

COMPARE THE CITIES OF NEW YORK AND TORONTO

- cluster the cities of New York-Toronto, compare the similarities and dissimilarities between the two based on their major venues to answer the questions:
  - How similar are Toronto and New York?
  - which neighbourhoods in New York are more similar (read ' more comfortable') for a person from a given neighbourhood in Toronto?

DATA LOADING AND PREPARATION

- Acquire Relevant Data From Appropriate Sources, Clean The Data Of Any Errors, Format It To Be Ready To Be Passed To The Data Science Algorithm

DATA SOURCES

Toronto

- Wikipedia page containing Canada Postal Code Data
- Geospatial data downloaded as csv file from the course site
- Venue data fetched using Foursquare API

New York

- JSON file from New York city Spatial Data Repository site
- Venue data fetched using Foursquare API

DATA LOADING AND CLEANING - I

Toronto	New York
- Postal code table on Wikipedia page loaded using read_html option in pandas - CSV file containing geospatial data read using read_csv option in pandas - Required dataframe obtained by merging the two dataframes above - Cells with no borough value or value of 'Not Assigned' dropped	- JSON file downloaded from New York city spatial repository site and loaded using json.load command in pandas - Required fields of borough, neighbourhood, latitude and longitude data filtered from 'Features' key and normalized to get required dataframe - No further cleaning done

DATA LOADING AND CLEANING – I (contd.)

Toronto	New York
- For rows with no neighbourhood data, neighbourhood taken to be the same as the borough - Rows grouped by postal code to reduce redundancy - Toronto neighbourhoods isolated from the whole Canada dataframe	

DATA LOADING -II

- For each city, venue information obtained through Foursquare API using the latitude-longitude data per neighbourhood
  - upto 25 venues
  - in each neighbourhood
  - within a radius of 500m

- Name, Latitude and Longitude data of venues filtered from the results, grouped by neighbourhood and appended to existing dataframes for each city

DATA PREPARATION

- The venue columns (all categorical data columns) encoded (one-hot encoding) before passing to the machine learning algorithm
- The unlabelled data sorted and grouped to get a new dataframe with 10 most common venues per neighbourhood for more readability
- Dataframes for combined dataset obtained by concatenating the encoded dataframe and sorted dataframe for the two cities

DATA ANALYSIS

- Apply Data Science Algorithm On The Curated Data To Analyse The Data To Solve The Business Problem

MACHINE LEARNING ALGORITHM

- Major clustering algorithms
  - K-Means Clustering
  - Hierarchical clustering (Agglomerative and Divisive)
  - Mean-Shift Clustering
  - Spectral Clustering
  - Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
  - Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM) and Affinity Propagation
- K-Means Clustering and Hierarchical clustering algorithms - most popular
- K-Means Clustering chosen since
  - the dataset is large
  - cophenetic distance for Hierarchical clustering is low (0.41)

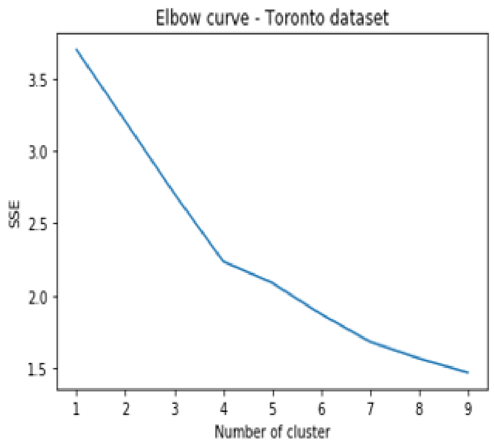
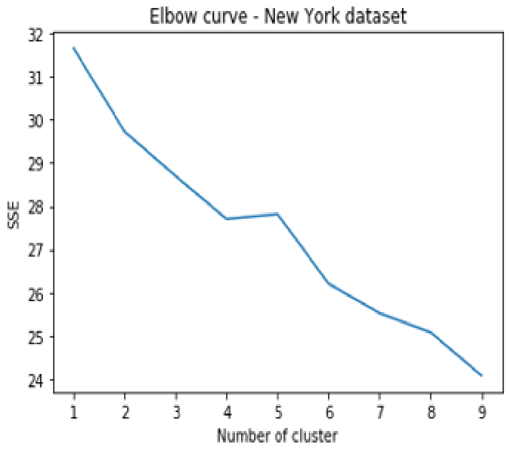
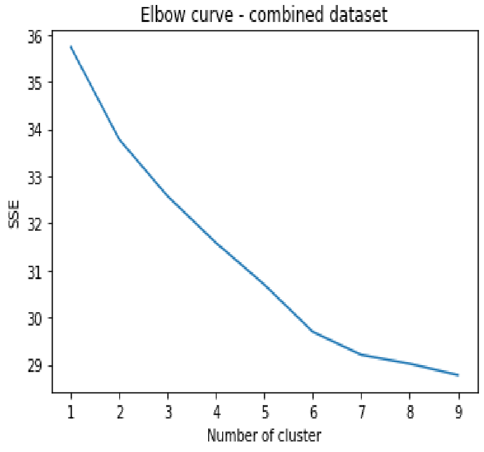
K-MEANS CLUSTERING

- Requires the number of clusters to be pre-set
- Random State parameter set to constant value to make the outputs of various runs more deterministic
- K-Means++ used to initialize the algorithm and hasten the convergence
- After clustering, the cluster labels added back to the venue-sorted dataframe and appended with borough, latitude-longitude data for each neighbourhood

K-MEANS CLUSTERING – Set The Number Of Clusters

- Optimal number of clusters found using
  - Elbow method
  - Silhouette coefficient
- Elbow Method
  - Plot of SSE(sum of squared errors) against number of clusters
  - N-values where the curve dips/bends/flexes are optimal points
- Silhouette Coefficient
  - Measure of logical cohesion of a cluster
  - Values between -1 and 1.
  - Closer the value is to 1, better the clustering

K-MEANS CLUSTERING – Set The Number Of Clusters (Contd.)

Toronto	New York	Combined dataset
		
Toronto dataset - two dips/elbows for the curve at n -values 4,5 and 7	New York dataset - the curve dips at n – values 2,4,5,6,7 and 8	Combined dataset – the curve dips at points 2, 6, 7 and 8

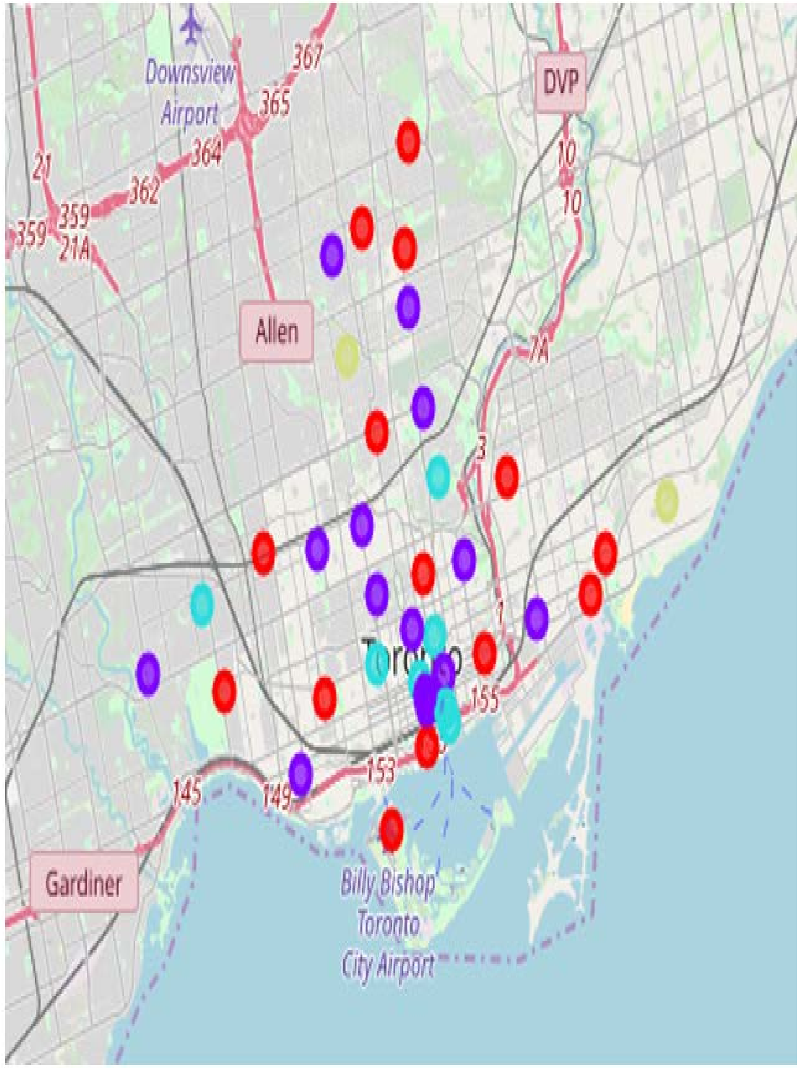
K-MEANS CLUSTERING – Set The Number Of Clusters (Contd.)

Toronto	New York	Combined dataset
Silhouette coefficient - highest for n values of 3, 2, 4… (all Euclidean, in descending order of coefficient magnitude)	Silhouette coefficient - highest for n values of 4, 3, 6, 5… (all Euclidean, in descending order of coefficient magnitude)	Silhouette coefficient - highest for n values of 2, 3, 4, 7… (all Euclidean, in descending order of coefficient magnitude)
Comparing both the measures, the n-value decided as 4	Comparing both the measures, the n-value decided as 4	Comparing both the measures, the n-value decided as 6

RESULTS

- Results Of The Data Analysis Step

RESULTS - TORONTO

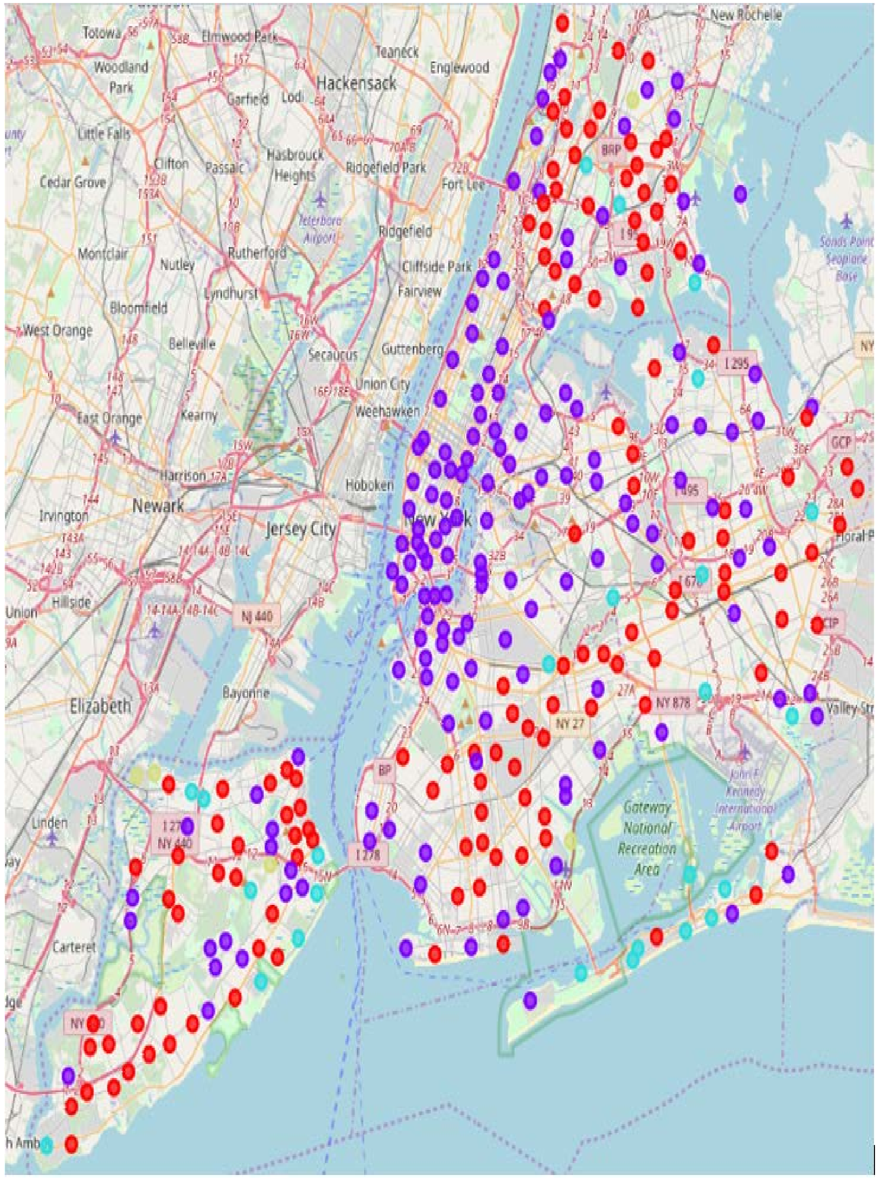


Neighbourhood Clusters – Toronto (Independent run)

Top Venues

- Cluster 1 - Parks, Coffee shops, Pubs/gastropubs, ethnic restaurants (Asian, Greek, Japanese, Indian etc.), Yoga studios, Discount stores, Liquor store/Brewery, etc.
- Cluster 2 - Cafes/Coffee shops/Diners, Gyms, Performing arts venues, Restaurants (Italian, Eastern European, Falafel, Ethiopian etc.), and so on.
- Cluster 3 - Cafes, Farmers markets, Bars (including Beer Bars), Steakhouses, Restaurants (Asian, Mexican, Seafood, Vegan etc.), Clothing stores etc.
- Cluster 4 - Outdoor centres/Parks/Trails/Yoga studios/Dance studios, Eastern European/Ethiopian restaurants, Health food stores among others.

RESULTS – NEW YORK

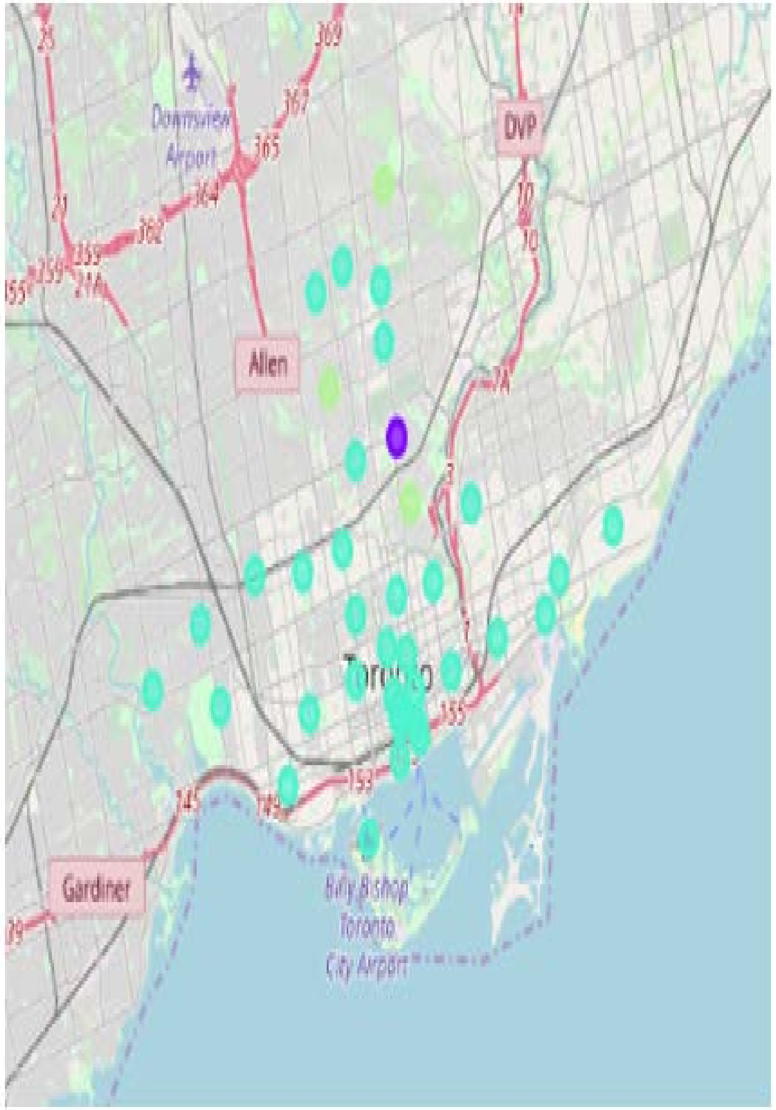


Neighbourhood Clusters – New York (Independent run)

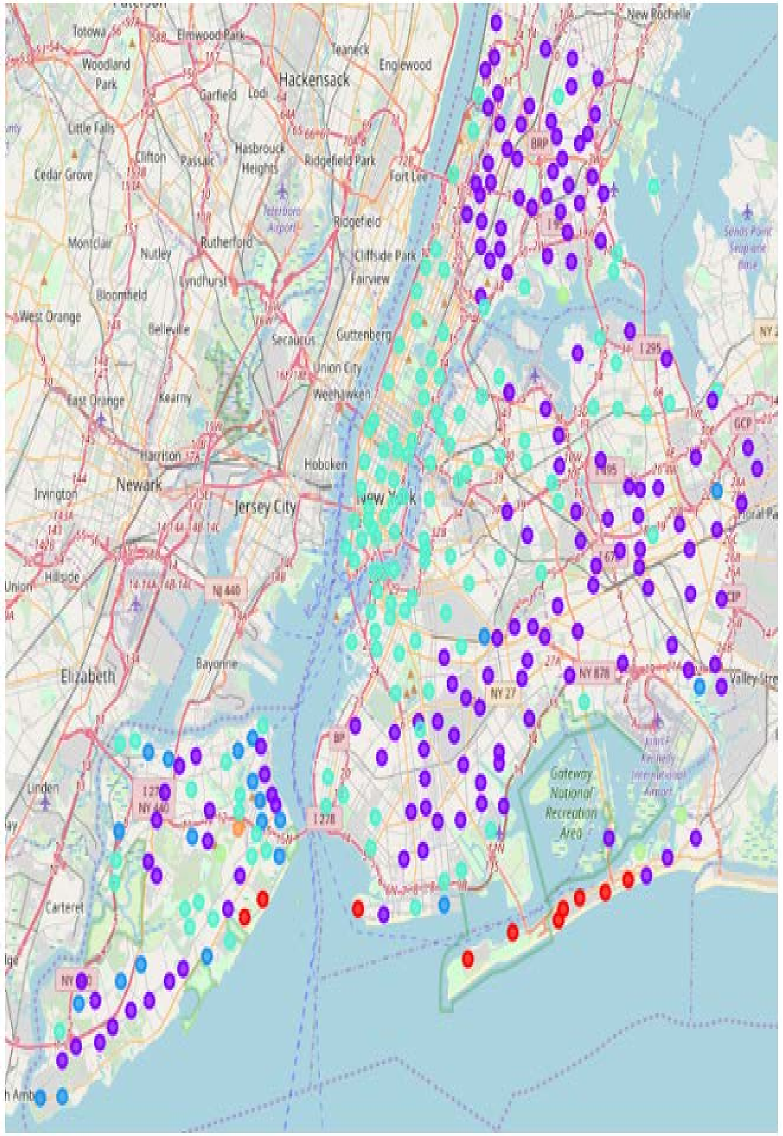
Top Venues

- Cluster 1 –
  - Pizza places, Deli/Bodega's, Bus stations, Banks, Donut shops, Pharmacies, Mobile phone shops, Restaurants (Fast Food, Mexican, Spanish, Caribbean etc.), Fried Chicken Joints, Supermarkets, Bakeries, etc.
  - Also boasts of a number of other categories like Bowling alleys, Skating rings, Race tracks, Hobby shops, Martial Arts Dojo's and many more
- Cluster 2 - Coffee shops/Cafes/Bakeries, Chinese/Italian/Korean/Fast food restaurants, Yoga studios, Farms, Farmer markets, Factories etc.
- Cluster 3 - Deli/Bodega's, Beaches, Parks, Falafel/Italian/Fast Food/Filipino restaurants, Farms, Fields, Farmer markets, Factories etc
- Cluster 4 - Harbor/Marina, Fish and chip shops, Falafel/Fast food/Filipino restaurants, Farms, Fields, Farmer markets, factories etc.

RESULTS – TORONTO AND NEW YORK COMBINED



Neighbourhood Clusters – Toronto (Combined Dataset Clustering)



Neighbourhood Clusters – New York (Combined Dataset Clustering)

RESULTS – TORONTO AND NEW YORK COMBINED (Contd.)

- Total of six clusters
- most of the Toronto neighbourhoods fall in the fourth cluster
- Many New York neighbourhoods fall in the same cluster as well
- prominent venues in the second cluster - Cafes/coffee shops/bakeries, Restaurants (Italian, Sushi, American, Fast food, Mexican etc.), Parks, Yoga studios, Pizza places etc.
- other clusters more or less similar to the New York city clusters

DISCUSSION

- Analyse The Results Obtained From The Data Analysis Stage

RESULT EXPLORATION

- Both the cities similar to a decent degree, both iconic of American culture but in different stages of urban growth
- Multitude of options available to every sect of society – the nature-lovers, the adventurists, the shoppers, the foodies, the sporty, the home-makers and so on, each of these options replete with choices!
- However, New York offers more variety (367 unique categories of venues in comparison with Toronto that offers 177)
- New York- larger, has a greater number of neighbourhoods (306 against 74) and hence exercise higher weightage in deciding the prominent venues in each cluster when clustered together
- Partly biased clustering since venue position is the sole criterion used

RESULT EXPLORATION (Contd.)

- From the user perspective:
  - New York more enticing to a tourist with its plethora of venues but much more costly.
  - Toronto offers similar venues at lesser costs
  - Upto the visitor to choose based on their travel-taste, budget and other preferences
- From business perspective:
  - New York full of life, offering more customers and competition at the same time
- From governing perspective:
  - New York reaching the apex of its growth while Toronto still in the growing phase.
  - Strategic measures to be taken to avoid growth stagnation for New York without overloading the city while Toronto to ensure optimal environment for its growing needs
  - Rectify regional imbalances based on venue concentration information

CONCLUSION

- The cities, Toronto and New York - similar in their iconizing of American culture but different in their degrees of doing so – with New York being a bigger and better torch-bearer
- Toronto, smaller compared to New York with lesser number of neighbourhoods, venues and venue categories
- Migrating between New York and Toronto to be easy to most people, save alone the dimensional difference between them
- With the inclusion of more parameters (like activity data, census data etc.) in the input dataset and cross-validation of more clustering algorithms, the efficiency of the study can be improved.