Clustering-NYC-Toronto-Neighborhoods

By Ayman Altaweel

On 29th, June 2021

LinkedIn: https://www.linkedin.com/in/ayman-altaweel-079774169

OUTLINE

- Introduction
- Data
- Methodology
- Results
- Discussion
- Conclusions

INTRODUCTION

- This project approaches a commonly faced problem by people who have to **move** from **NYC** to **Toronto** or Vice versa.
- The problem they face is that they have to move yet they prefer to move to a **similar place** to where they are, having the same **lifestyle**. (ex: similar places, similar venues, ...etc.)
- So, we **clustered the neighborhoods** within the two cities according to their most common **venues** into **6 clusters**.
- Having clustered them, those people will be able to **identify easily where to go** as simple as looking at the **clustered neighborhoods map** which has been generated and identify similar neighborhoods by **colors or check this** **dashboard**.

Data

- In this project, we used different data sources which are:

- **NYC**-data from **Json** file: data about NYC neighborhoods and their respective latitudes and longitudes.
- **Toronto-data**, web scrapped from Wikipedia: data about Toronto neighborhoods.
- **Geospatial data**: data of respective latitudes and longitudes to Toronto's neighborhoods.

METHODOLOGY

- Data Collection from the previously stated sources.

- Data Cleaning:
  - Dropping unneeded columns.
  - Filtering Canada data to get only Toronto's.
  - Adding longs and lats to Toronto's data.
  - Appending both of Toronto's and NYC's data.
  - One-hot-Encoding venues for further processing (i.e building K-Means clusters)

- ML Modeling:
  - We applied the K-Means algorithm which is a partitioning unsupervised clustering ML model by which we cluster a given set of observations, neighborhoods in our model, according to their features, respective venues appearance likelihood in our model, into non-overlaping clusters without any internal structure.
  - We clustered the neighborhoods into 6 clusters to minimize the within-cluster sum of squares (i.e WCSS) as reasonable as possible.

Results:

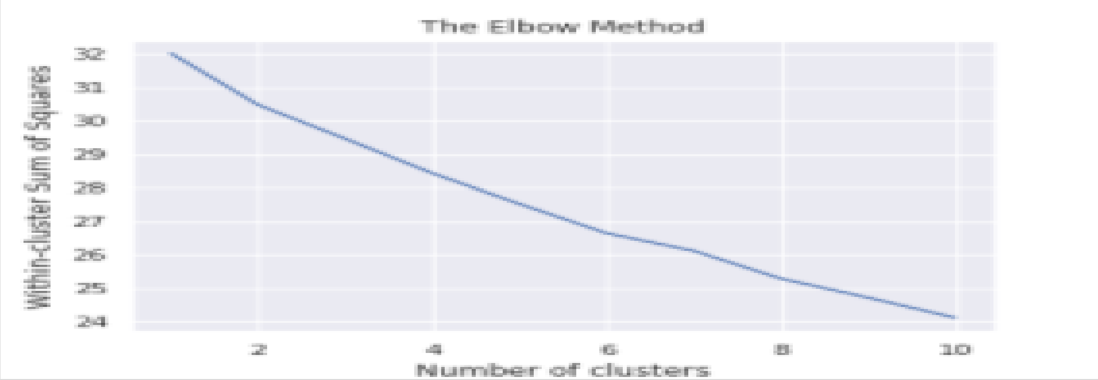
- After building the K-Means clustering model, we got a data frame of all neighborhoods with their respective clusters.
- Clustered data frame

- Also, we made further illustration through this dashboard to add interactivity.

DISCUSSION

- **<u>While clustering, we used the elbow method to get the optimal no. of clusters that minimizes the WCSS as reasonable as possible while keeping clusters have a meaning.</u>**

- **<u>It turned that 6 clusters will be sufficient to generate distinct clusters.</u>**



CONCLUSION

- Here is a summary table that concludes the clusters.

- The vast majority of NYC's neighborhoods fall in the 6th & 3rd clusters.
- The vast majority of Toronto's neighborhoods fall in the 3rd cluster which we gonna describe very soon.
- In Toronto, there are no neighborhoods fall into neither the 2nd nor the 5th clusters.
- (i.e If someone lives either in the 2nd or the 5th clusters in NYC, he/she won't find a similar neighborhood to move to in Toronto.)

Cluster _no	NYC_Cluster_Volume	NYC_Cluster_%	Toronto_Cluster_Volume	Toronto_Cluster_%
1	8	2.62%	1	2.56%
2	1	0.33%	-	-
3	106	34.75%	35	89.74%
4	3	0.98%	2	5.13%
5	21	6.89%	-	-
6	166	54.43%	1	2.56%

THANK YOU...