Capstone project: Finding the closeness between cities based on venue categories

Muhammed Zahit Gök

08.01.2020

Outline

- Introduction
- Data Description and Collection
- Methodology
- Discussion
- Conclusion

Introduction

- The capital cities especially having big population are the important indicator to understand the trend and similarity between the countries.
- The cities that we investigate:
  - New York
  - Toronto
  - Berlin
- The population of each city approximately around 4 million
- The number of venue in order New York, Berlin and Toronto

Introduction

- We aimed at
  - finding the similarities between cities and neighbourhoods
  - comparing the statistics of each cities
  - using clustering and similarity metrics
- As a result, we want to find out which city is closer to New York

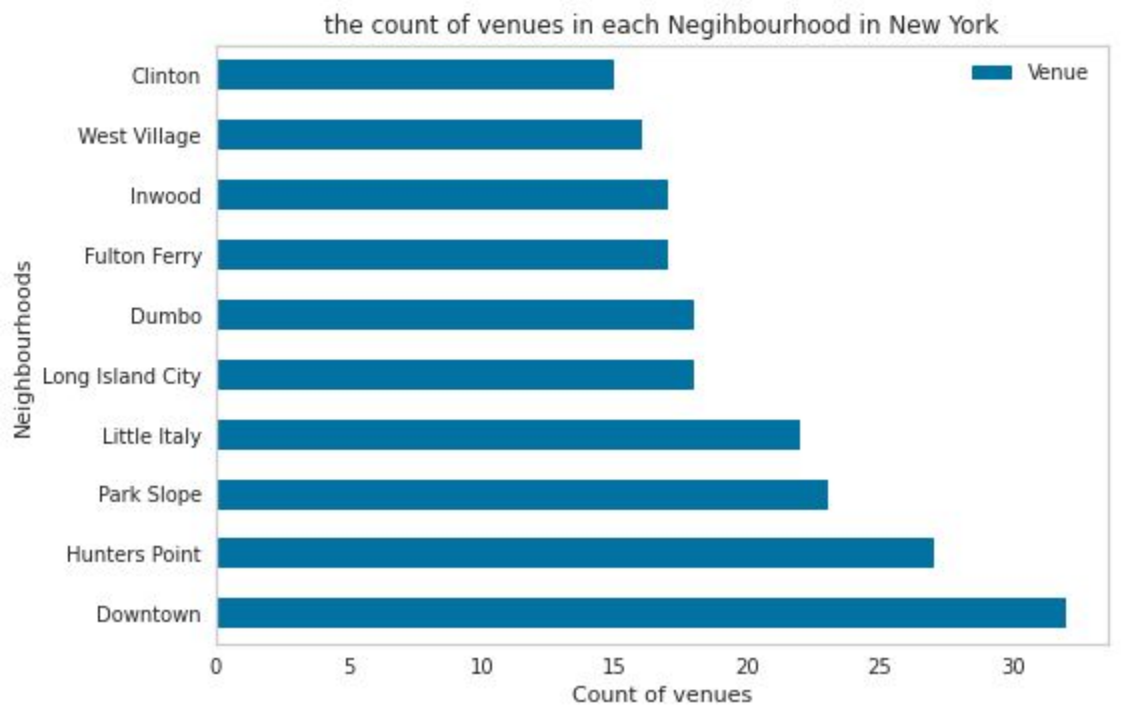
Data Description and Collection

- For collecting data, we used
  - Wikipedia
  - geocode (python library)
  - Foursquare API

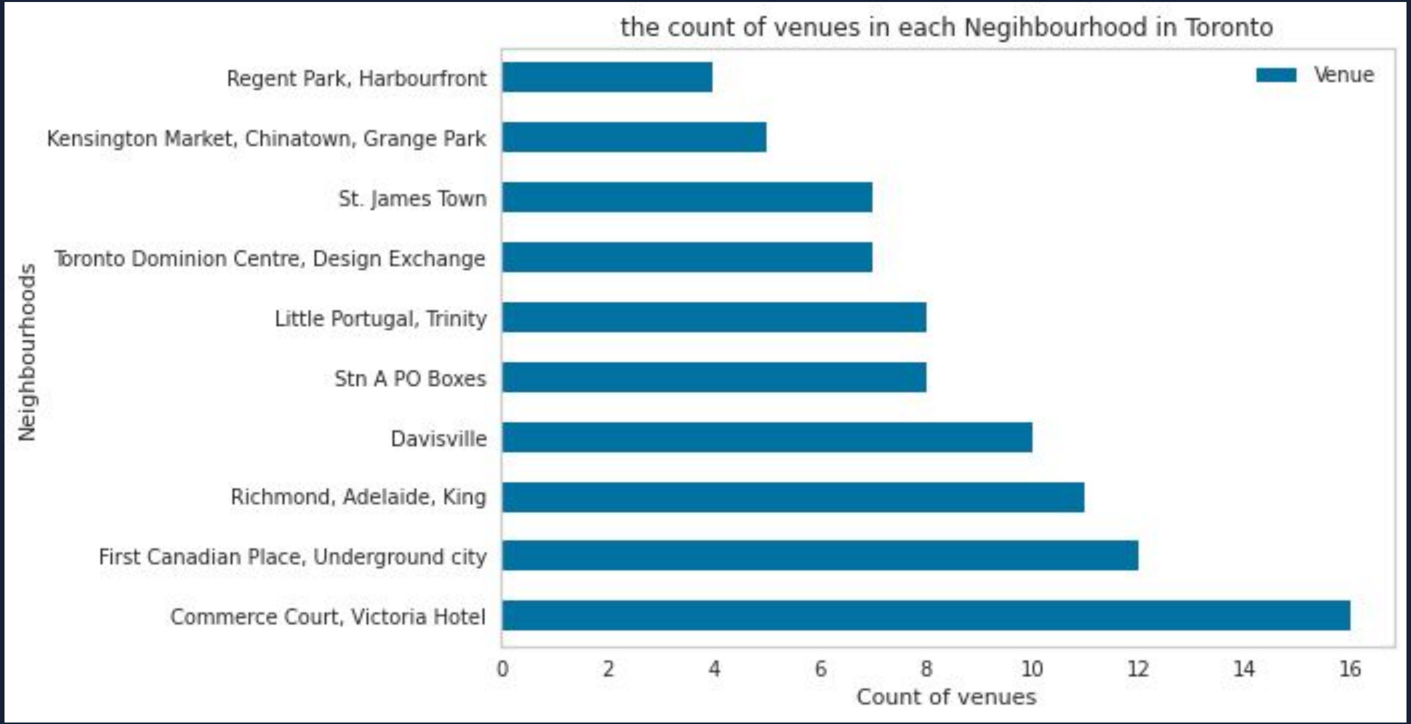
Methodology

- we used
  - k-means clustering
  - cosine similarity metric
  - some statistical methods(such as sum, mean and count)

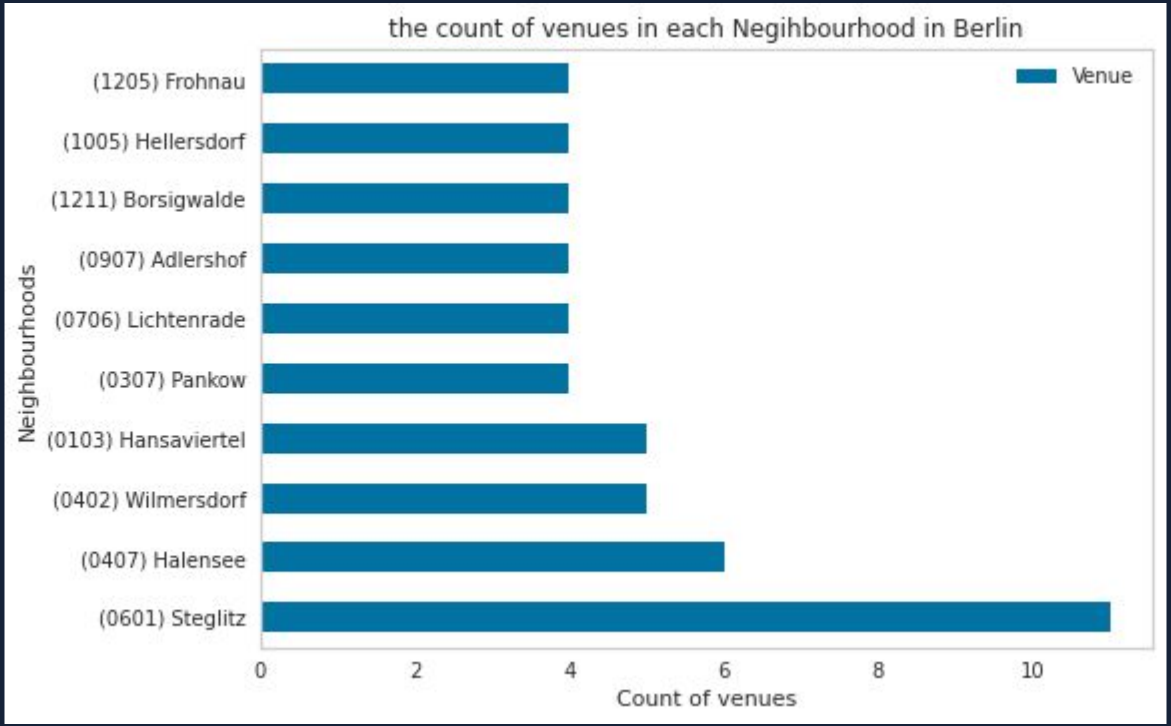
Count of Venues in New York



Count of Venues in Toronto



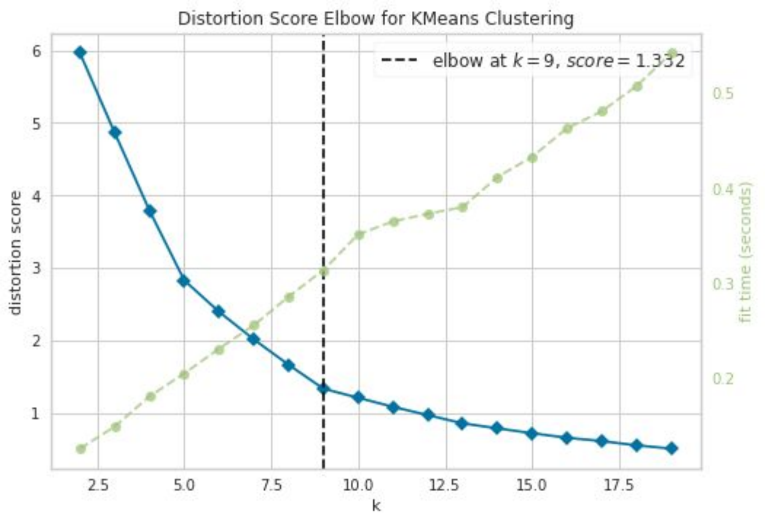
Count of Venues in Berlin



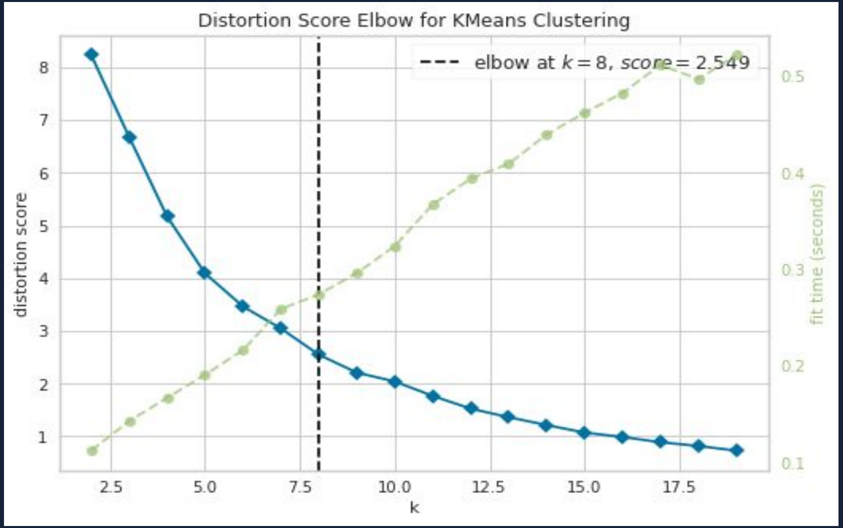
k-means clustering to find how close cities are to each other

finding the best k-value

- We used KElbowVisualizer library to find best k value
  - the nest k for New york and Toronto is 9
  - the nest k for New york and Berlin is 8



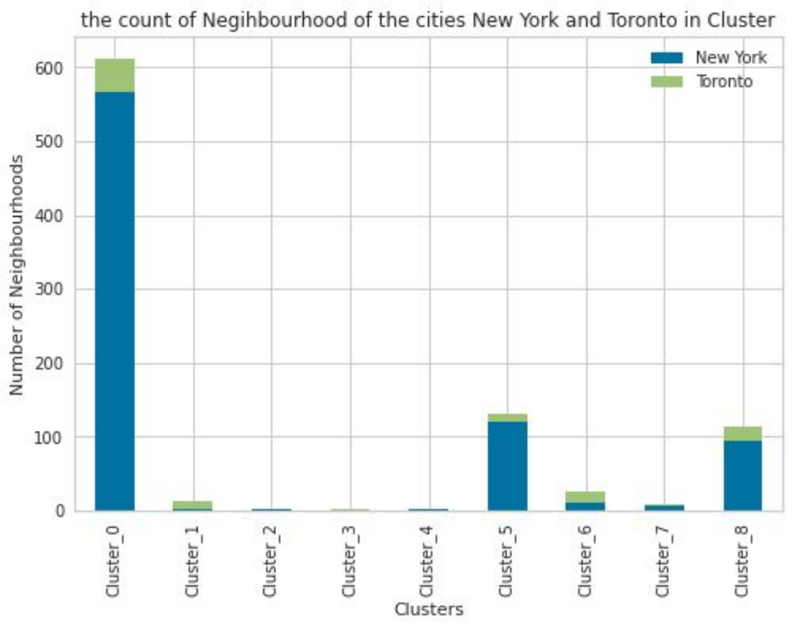
New York and Toronto



New York and Berlin

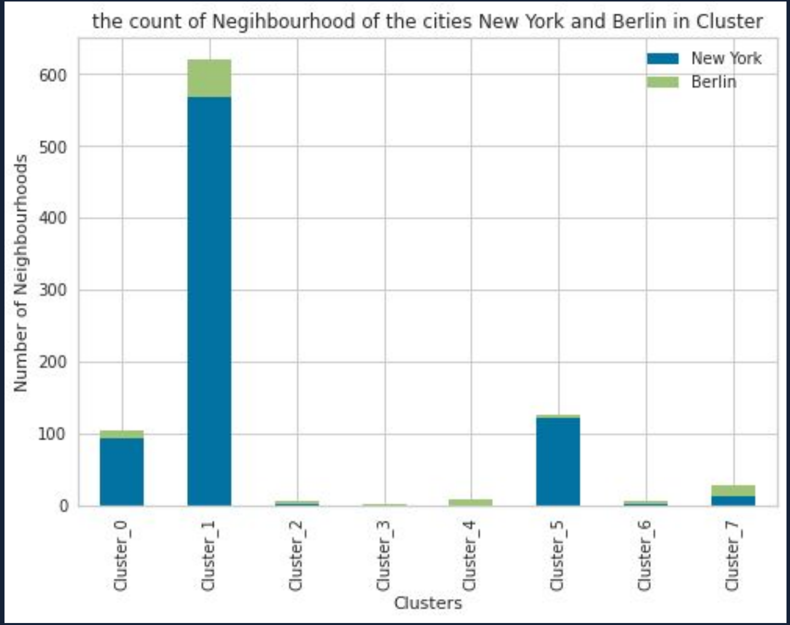
Cluster result of New York and Toronto

- We found some similarities on neighbourhoods locating in different cities.



Cluster result of New York and Berlin

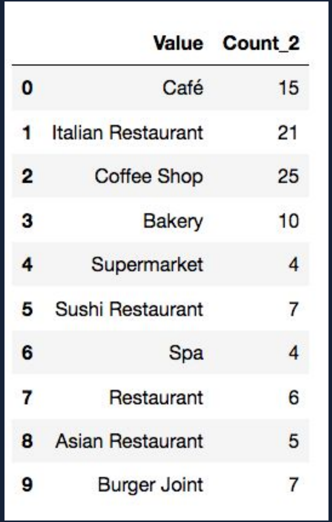
- We also observed some similarities on neighbourhoods in New York and Berlin



Showing the closeness with cosine similarity metric

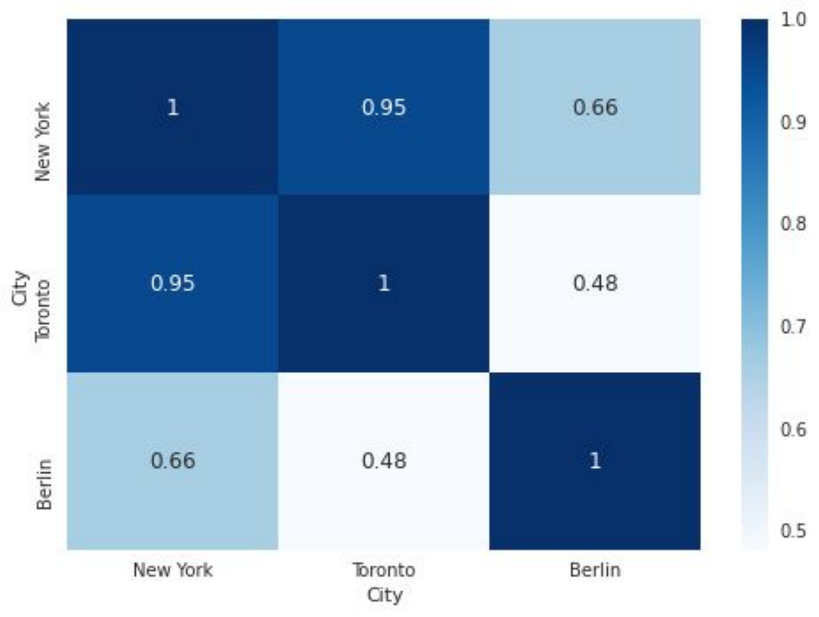
Cosine similarity metric

- We would like to show the closeness in a numerical way.
- We used cosine similarity metric based the count of venu categories.



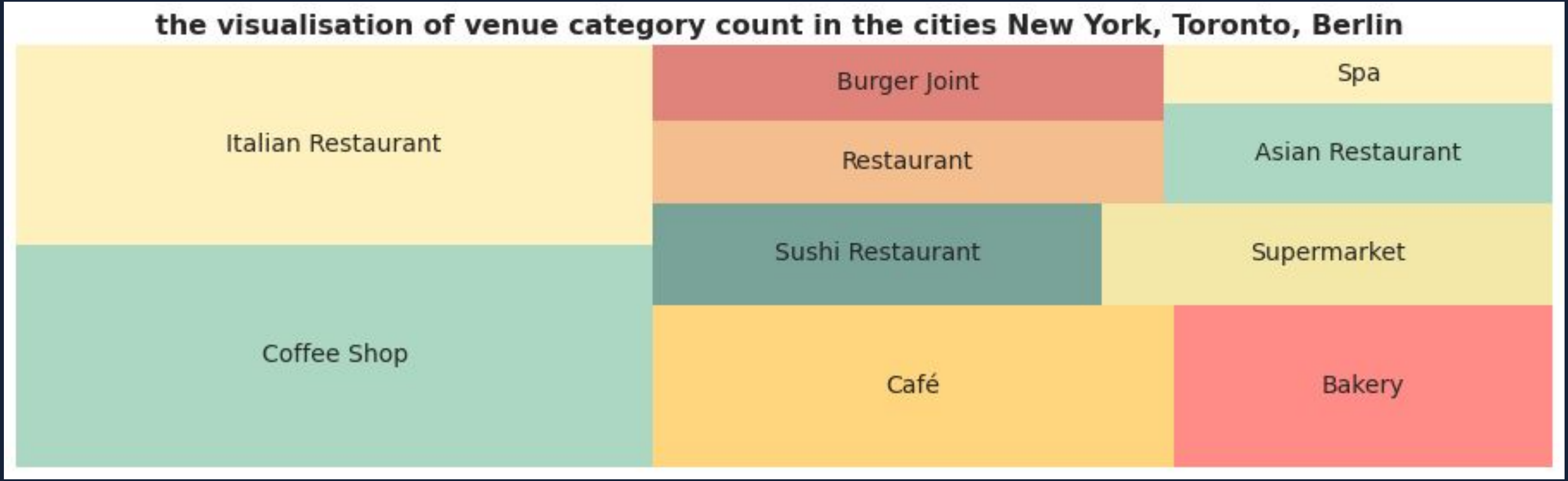
	Value	Count_2
0	Café	15
1	Italian Restaurant	21
2	Coffee Shop	25
3	Bakery	10
4	Supermarket	4
5	Sushi Restaurant	7
6	Spa	4
7	Restaurant	6
8	Asian Restaurant	5
9	Burger Joint	7

Newyork venue category distribution



Total count of venue category based on the cities Newyork, Toronto and Berlin

- We show the tree map of total count of venue categor in three citites



Discussion

- These large and important cities can be a sample to show the trends of the country.
- In this study, we found the similarity between cities by using k-means algorithm and cosine similarity metric.
- We used a certain number of venue category in our approaches. If we increase the category number, the result may be better.
- Moreover, radius which determines the area of search can be increased. Thus, we may find better similarity.

Conclusion

- By performing statistical approaches on the big cities, we can find similarity and solve the problems in the identical cities with these similarities.
- The big cities are the good sample for analysing the countries.