IBM Data Science

CAPSTONE PROJECT: DEMOGRAPHIC CLASSIFICATION OF OTTAWA NEIGHBOURHOODS

Introduction: Business Problem

- We clustered neighbourhoods using venues for various cities.
- Question: **If we cluster neighbourhoods with respect to demographic data, is there any relation?**
- We will look at Ottawa because that is where I live and its interesting.
- Why do we care:
  - Should businesses care about where the venues are, or where the population are?
  - How much does access to venues affect demographics.

Data Acquisition

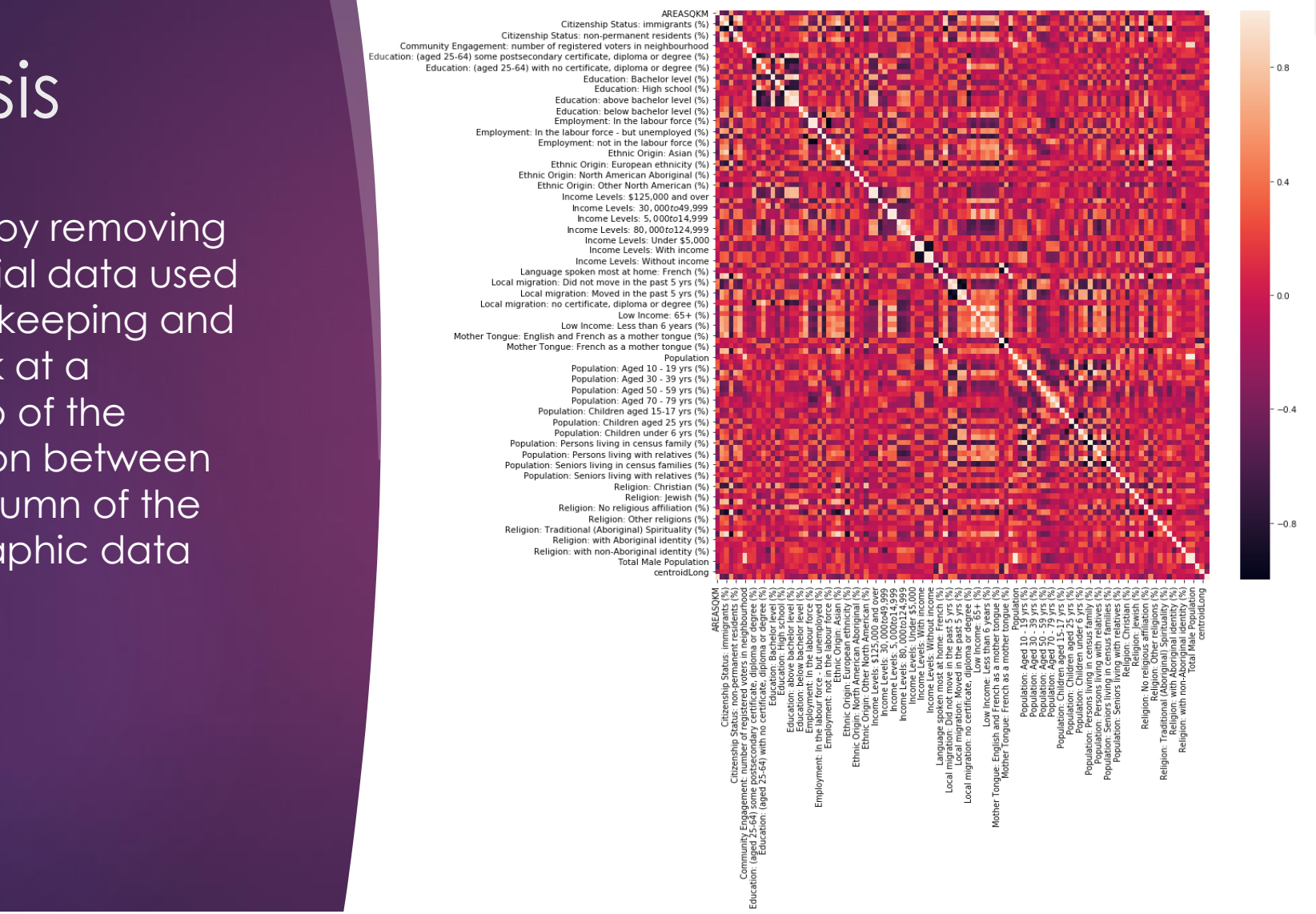
- We will need demographic data:
  - Acquired a GEOJSON for the neighbourhoods of Ottawa which had quite a bit of demographic data including various breakdowns of populations by religion, age, gender, income, religion, etc.
- We will need venue data
  - Acquired using the Foursquare API which had data on the various venues near a location.

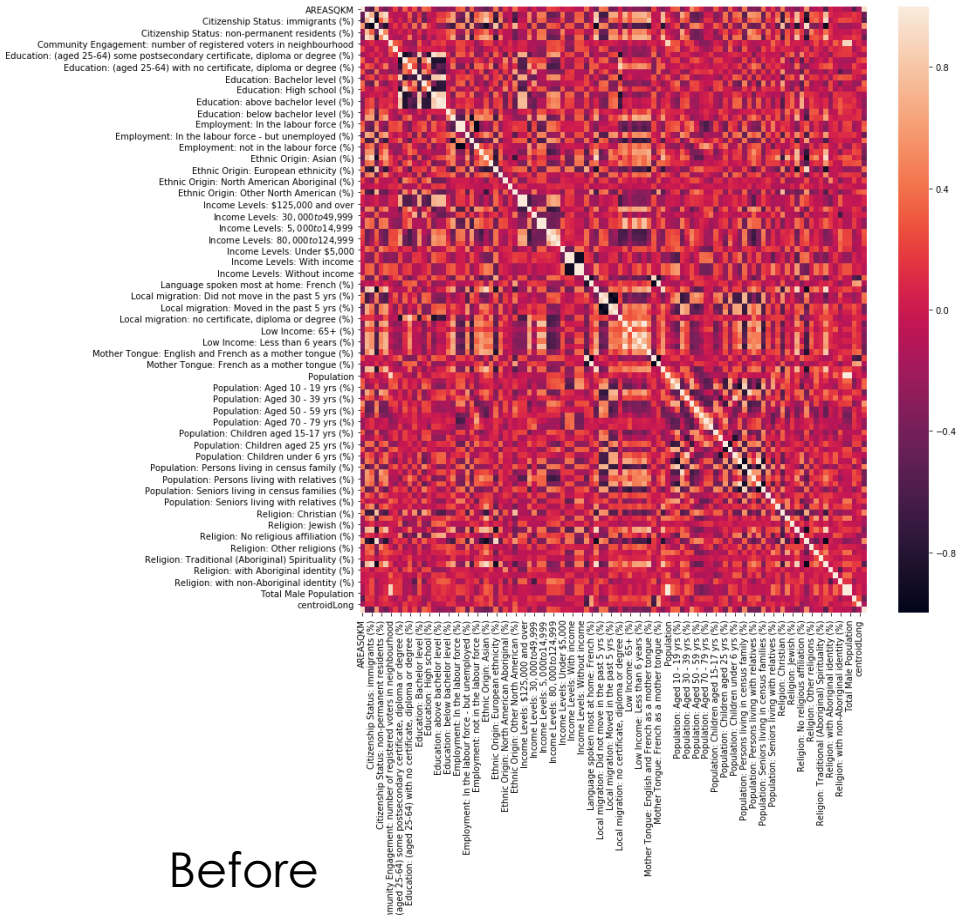
Methodology

- This analysis has 3 main parts:
  - Cleaning the data from the GEOJSON file since it has duplicate and irrelevant data.
  - Clustering the neighbourhoods by demographic data and by venue data using k-means
  - Comparing the results of the two clustering algorithms to see if neighbourhoods tends to cluster together using each method

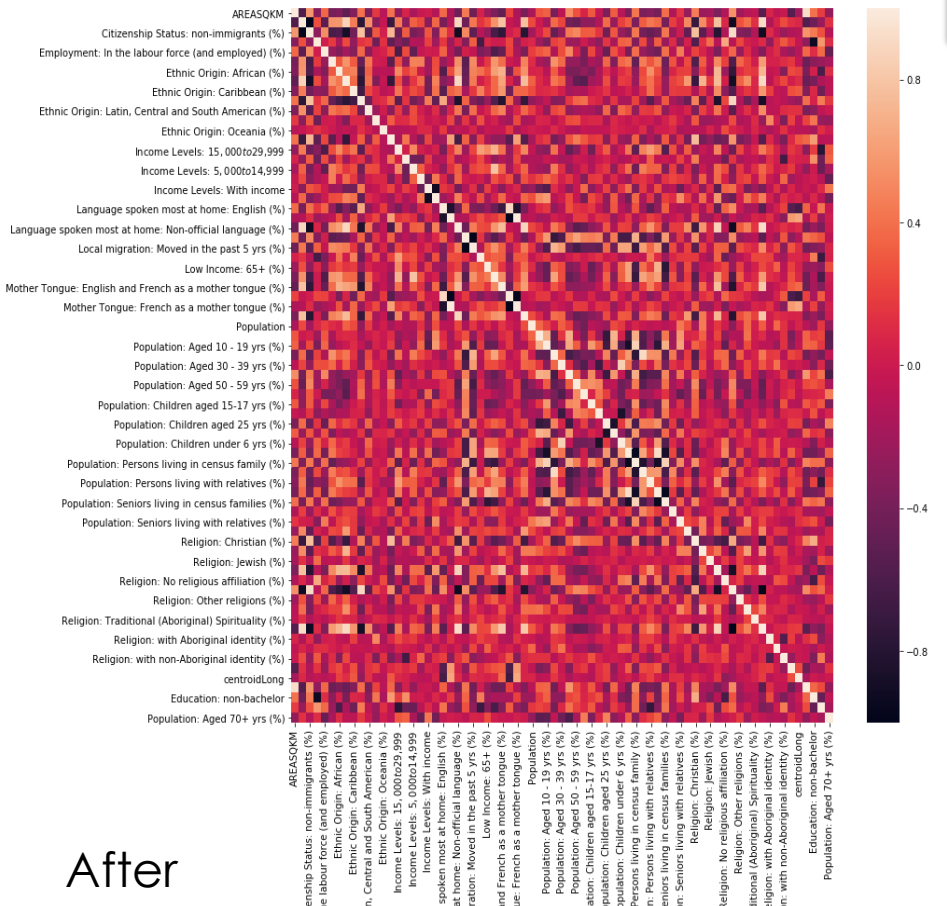
Analysis

- We start by removing some initial data used for book keeping and then look at a heatmap of the correlation between each column of the demographic data





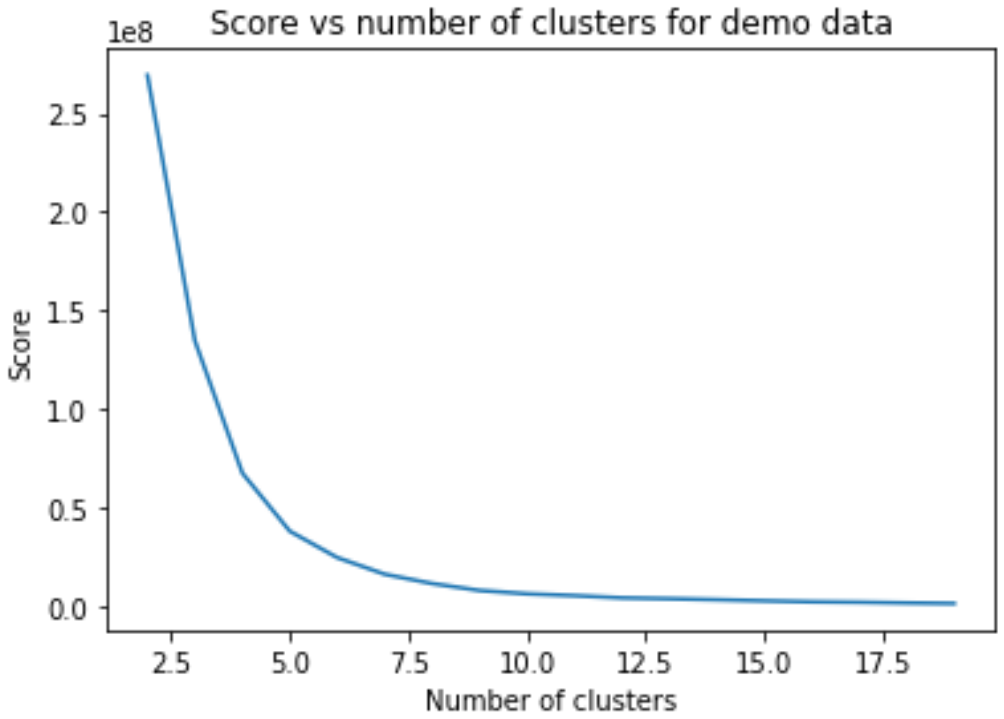
Before



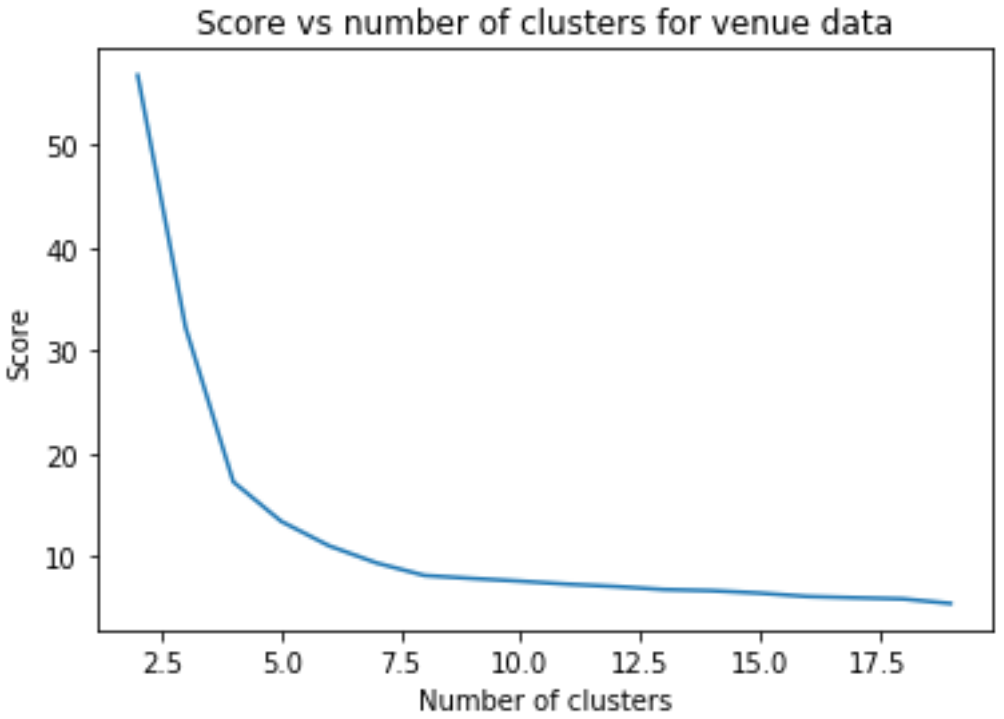
After

Model construction

- Next, we look at the clustering algorithms. We find the number of clusters by looking at a graph of the score of the fit vs the number of clusters and use the elbow method.


Score vs number of clusters for demo data
1e8
2.5
2.0
1.5
1.0
0.5
0.0
Score
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Number of clusters

k=5

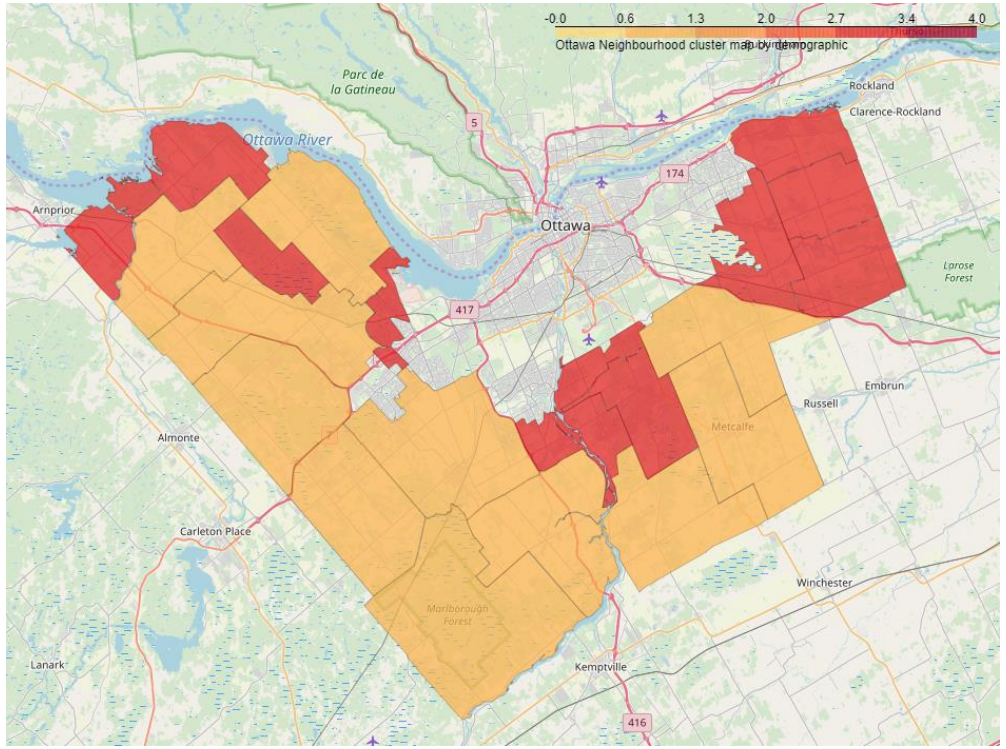

Score vs number of clusters for venue data
50
40
30
20
10
Score
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Number of clusters

k=8

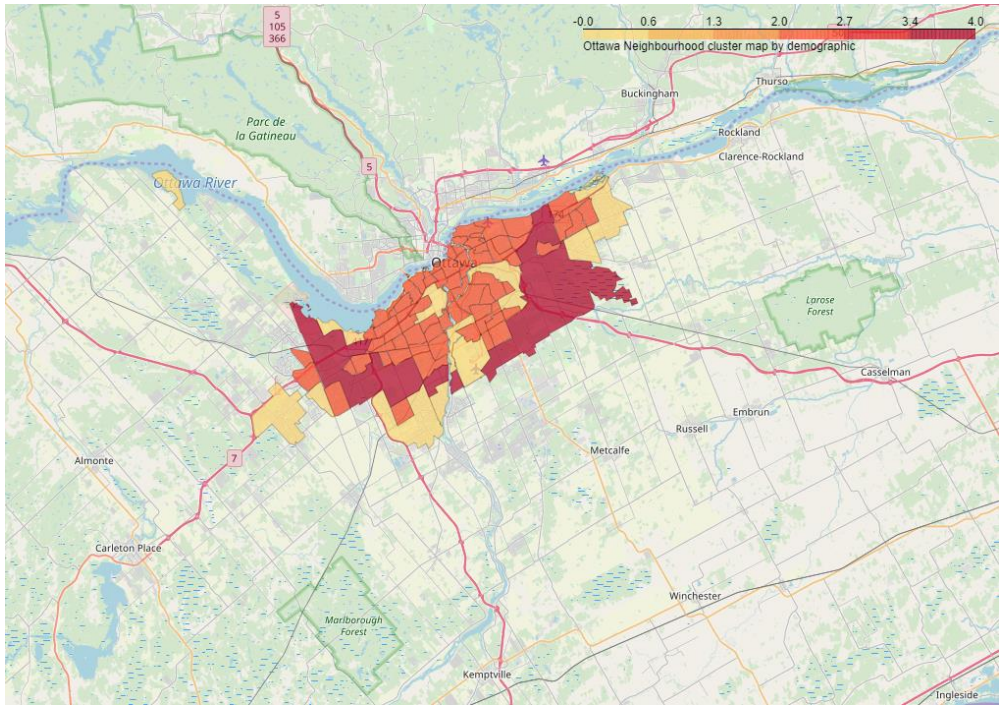
Result of the clustering

- Because folium couldn't handle all the neighbourhoods at once, we split them up into two maps for each clustering algorithm.

Clustering using Demographic data

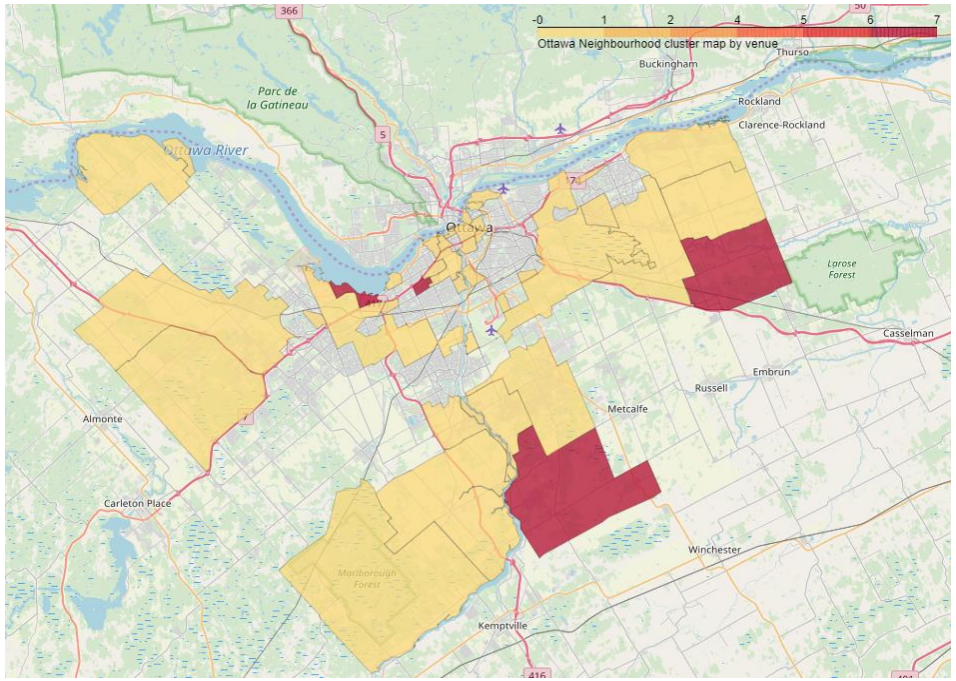


Cluster 1 and 3

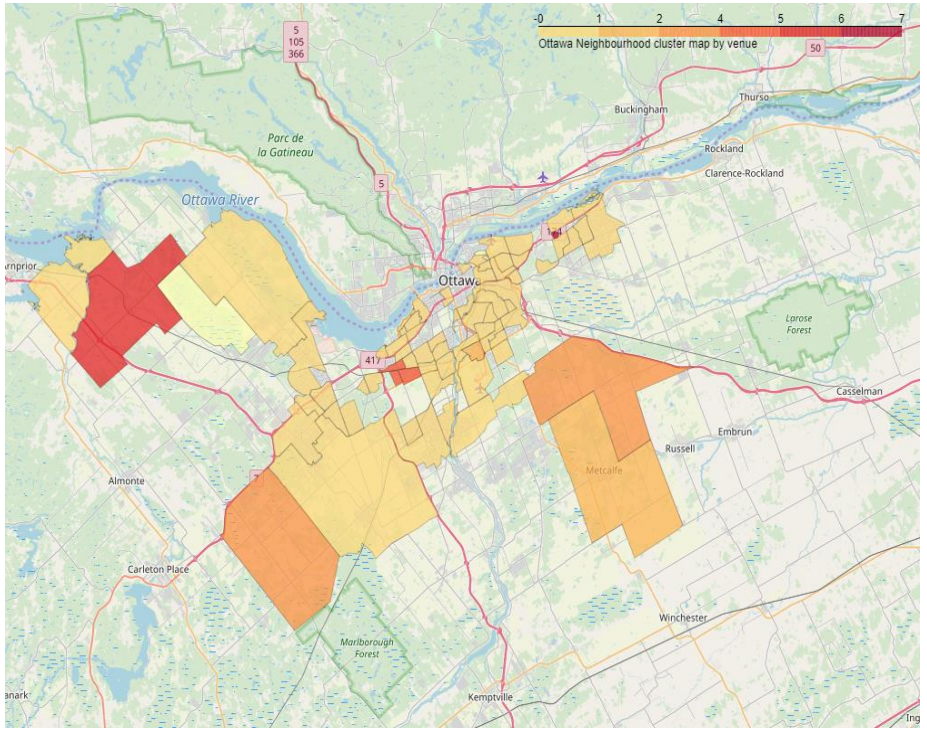


Cluster 0,2 and 4

Clustering using Venue data



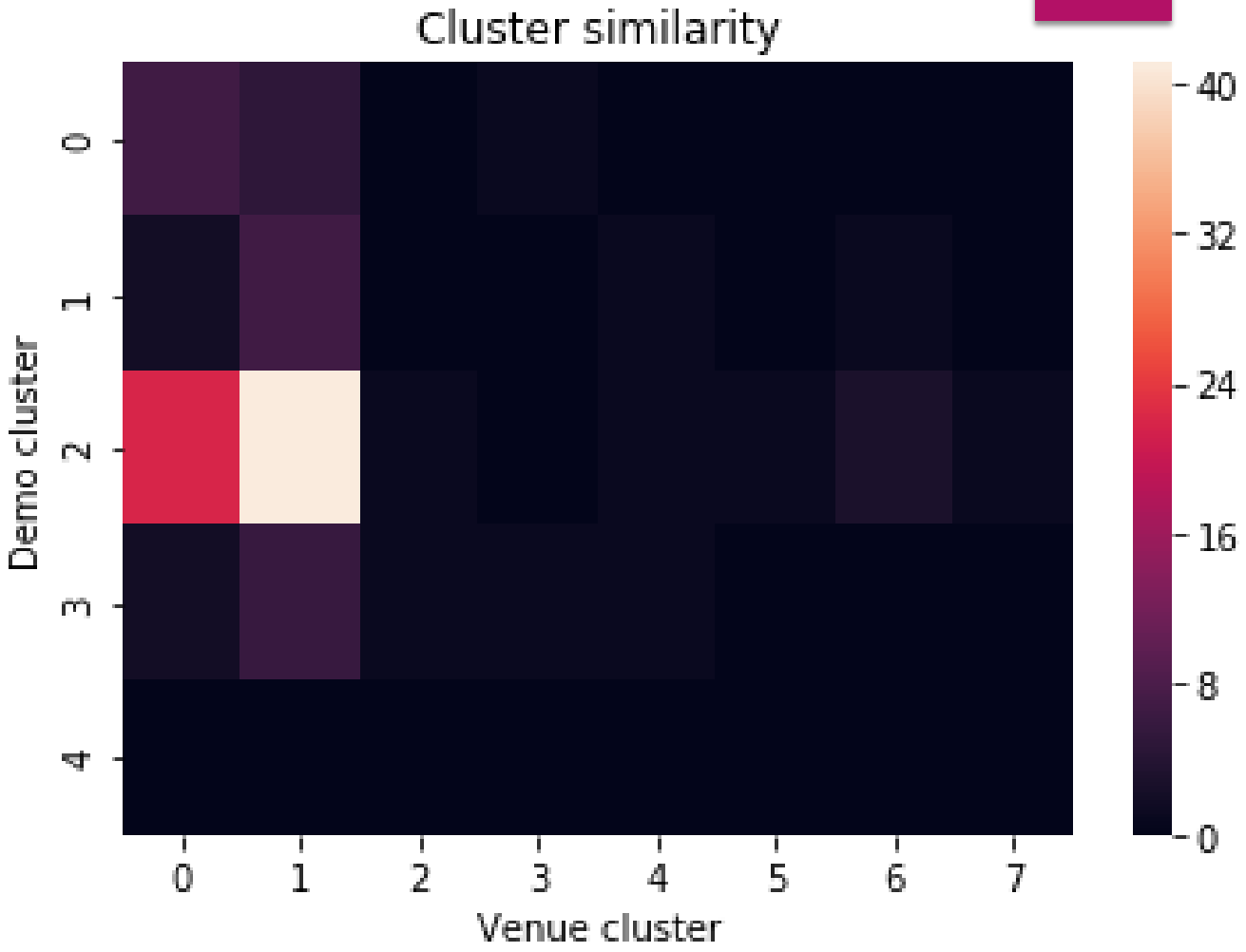
Cluster 1 and 3



Cluster 0,2 and 4

Comparing both clustering algorithms

- Given a neighbourhood, they will have two cluster indices: one for each clustering algorithm.
- We can make a heat map to see if there is a strong connection between each clustering algorithm



Results and Discussion

- The demographic clusters seem to be radial about downtown
- The venue data seems cluster about ethnic ghettos
- Only similar downtown there a large homogeneous part of the city is split up by many small neighbourhoods.

Conclusion

- There does not seem to be a strong connection between demographic information and the venues.
- This study might have been biased by the type of users which contributed to Foursquare.
- Takeaway for business:
  - The consensus seems to be to setup your business in a location with similar businesses, not near the target demographic.
- Takeaway for people:
  - Nearby venues should not be the driving factor when looking for a place to live. Other factors will be more important such as real estate prices for example.