


BATTLE OF THE NEIGHBORHOODS

Name: Carl-Michael Edeling

Email: cmedeling@gmail.com

PROBLEM DESCRIPTION

An Italian Restaurant is looking to open another branch due to their continued success. It is located in Sandton, Johannesburg, South Africa, and is looking to open the second branch in a neighborhood similar to Sandton.

Several white lines of varying lengths and thicknesses are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

DATA ACQUISITION AND MANIPULATION

Venue data was acquired using the Four Square API, and the weather data was acquired using the WeatherBit API.

The venue data was refined to include the suburb names, along with the frequency of venue categories.

The weather data was taken for each suburb, but due to the limited data accessibility from the API license key, it only allowed for 500 calls per day with historical range of 4 years.

The total number of data points per subject for each measure was 8. 2 per year for 4 years.

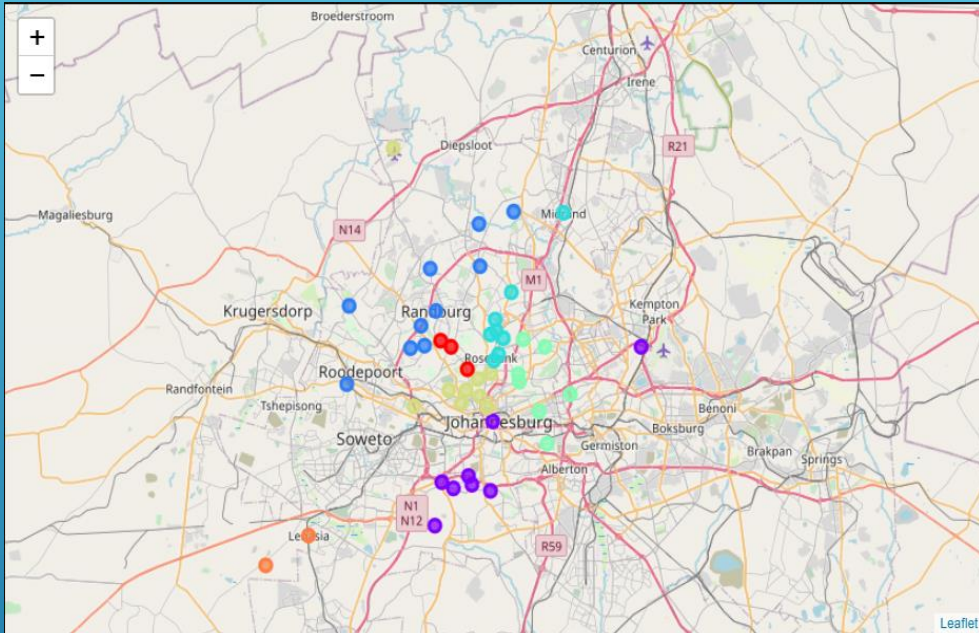


DATA ANALYSIS

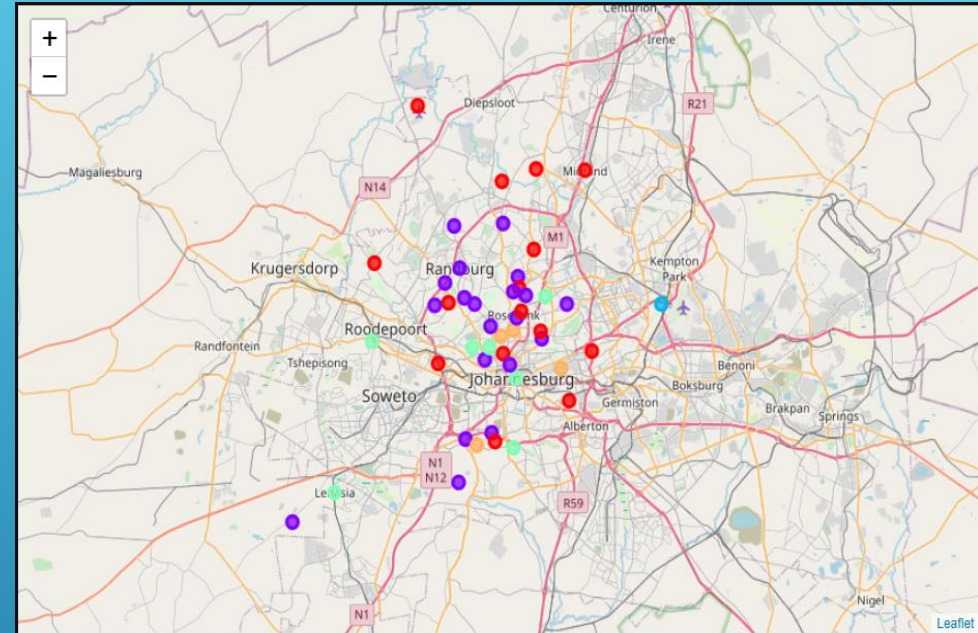
Because the problem involved comparing suburbs and grouping them based on similarity, it was decided to use the unsupervised k-Means Clustering algorithm.

The optimal number of clusters to use was determined by identifying the “elbow-point” on the plot of Inertia vs. Number of Clusters.

RESULTS



Clustering – Venue Data



Clustering – Venue Data and Weather Data

2 k-Means models were generated:

- 1) Using only the venue data
- 2) Using both venue data and weather data

RESULTS

The best results were obtained using only the venue data, which showed 3 suburbs to be similar to Sandton: Illovo, Hyde Park, Sandhurst.

It was decided that the weather data skewed the results too much, resulting in 20 suburbs similar to Sandton.

CONCLUSIONS

The best results were obtained using only the venue data. This however is insufficient to confidently say if the selected suburbs would be an appropriate location to open a second Italian Restaurant.

It is recommended that more data be collected, such as:

- 1) Economic Data
- 2) Demographic Data
- 3) Traffic Data

These datasets were not included due to financial and time limitations.