Clustering Pretoria Neighbourhoods for easier student allocation

February 12, 2021

1 Introduction

Pretoria East is the most vibrant district in the city of Pretoria, South Africa. It boasts a large student population owing to the fact that there are 3 universities within the district and numerous colleges. It is also home to a majority of the most affluent neighborhoods in the city. Accommodation for students that provides a safe and vibrant environment whilst offering access to public transport is notoriously scarce in the city.
New students often end up living in dangerous neighbourhoods that are far from campuses and do not offer places that enhance student life.
This project aimed to cluster the neighborhoods in Pretoria to help students find the ones that best serve their needs. The project made use of the Foursquare API to obtain data on the various neighborhoods in the district. Specifically, the project clustered student-friendly neighbourhoods together so as to suggest these to students seeking accomodation

2 Data

The data used in the project includes a Wikipedia list of all surbubs in Pretoria. Thereafter, the dataframe of Pretoria sububrs was used to geocode each of the suburbs using Python's geopy, geopandas and geocoder. There was no readily available dataset of the geospatial data of Pretoria, therefore we had to scrape Google Maps for the coordinates of the suburbs.
Once the coordinates were obtained for the neighbourhoods, then Foursquare was used to obtain any establishments in each suburb. This presented us with a JSON file of all establishments within a 1km radius of each suburb and these were stored in a dataframe.

3 Features and Processing

We initially had suburbs whose coordinates could not be found on Google Maps. We dropped these suburbs from the dataset as they could not be used on Foursquare.
From the Foursquare JSON file we only needed the venue name, venue category, venue latitude and venue longitude. We then coded the venue categories using Onehot encoding for easier application of scikit-learn algorithms. We then created a dataframe of the mean number of times each category is found in each neighbourhood and found the 10 most common venues per neighbourhood.

4 Model and Technique

After the relevant feature engineering we apply k-means clustering with k=5. Having obtained the cluster for each neighbourhood, we then created a dataframe of cluster labels and the 10 most common venues together with the geospatial data for the neighbourhoods.
We then used folium to create a map of the clusters for further analysis.

5 Results and Discussion

We found that the 87 suburbs in Pretoria can be clustered as shown in the table below;

Cluster Label	Number of Suburbs
0	2
1	2
2	8
3	29
4	46
Total	87

The tables shows us that a majority of suburbs belong to cluster 4. Further analsysis of cluster 4 revealed that it contains campuses for the 3 universities in pretoria and the numerous colleges in the city. This means we can assume that cluster 4 has a high student population.
Cluster 3 is the second most common cluster in the city. The cluster contains suburban and famliy-centric establishments like schools, malls, playgrounds, etc. This cluster mostly caters to the city's working middle class.

6 Conclusion

In conclusion, the objective of the project was to suggest suitable suburbs for students. Based on the analysis conducted, we can suggest that students would be best suited living in suburbs in cluster 4. They are close to campuses and the student community while offering all the convinient venues for a smoother student experience.