Using Cluster Analysis to Investigate Australian Suburb Similarity

Project Summary

The project aims to investigate similarity between various suburbs in Australia by conducting Cluster Analysis on postcodes using Demographics, Income and Rent Data. For the purpose of this project, the "similarity" is defined as overlapping of the aforementioned variables. The project aims to investigate the extent of impact of variable's presence on Suburb Similarity, for instance, analysing the role of Income in grouping various suburbs together. The report further employs demographic variables such as age, gender, marital status and religious affiliation etcetera to investigate impact of each of those on suburb similarity as well as to serve the overarching purpose of this project to generate succinct clusters containing suburbs with similar attributes. The project further aims to use this analysis to find out a suburb in Melbourne most similar to a suburb in Sydney. Lastly, we aim to test various predictive models to achieve a 75% - 80% accuracy in predicting suburb location.

Goals

Our goals for this project in no particular order are as follows:

1. Perform cluster analysis of the demographics of suburbs to determine if they contribute to attracting people to certain postcodes/suburbs, resulting in suburbs with similar characteristics.
2. Conduct cluster analysis and create an algorithm to find similar suburbs across other states based on the similar characteristics they might share.(For e.g.: which suburb in Melbourne is most similar to a suburb in Sydney)
3. Analyse the impact of absence or presence of data set/s on clusters. (For e.g.: would our cluster analysis be different if we were to include only one of the demographics data for our analysis)
4. Perform predictive analysis to accurately predict the location of a suburb and determine whether the suburb is in the capital city, regional area, or rural areas

Data Set Summary

The two sets of data our team will be analysing are ***Demographics dataset*** and ***Income and Rent dataset***. The demographics dataset has been extracted from the 2016 Census Data released by Australian Bureau of Statistics (ABS) using their online table builder tool. Multiple datasets containing 4 different demographics characteristics: age, gender, marital status and religious affiliation of the population in 2472 postcodes of different Australian States have been downloaded in .xlsx format and then merged into a single file 'Demogrpahics.xlsx'. Similarly, Income and Rent dataset has been extracted from Taxation Statistics 2015-2016 released by Australian Taxation Office (ATO) in .csv format. This dataset already contains the income and average rent of the population listed by the postcodes hence, it doesn't require any merging. However, both datasets have missing values for different postcodes, which has been dealt with by using the average values of other postcodes.

2016 Census Data - ABS: https://auth.censusdata.abs.gov.au/webapi/jsf/dataCatalogueExplorer.xhtml
Income and Rent data- ATO: Taxation Statistics 2015-16 - Individuals - Table 25 - data.gov.au

Techniques

The techniques used for this project are cluster analysis and predictive analysis.

Cluster Analysis

Cluster Analysis will be used to generate clusters based on aforementioned variables. Techniques such as K-means and Hierarchical Clustering will be used for experimentation. In case of unsatisfactory results, the team aims to employ techniques such as DB Scan and Louvain Clustering to achieve clusters that can succinctly group similar suburbs.

Predictive Analysis

We aim on experimenting with various Predictive Analytics techniques and models to accurately predict whether a suburb is in the capital city, regional centre or rural area. The techniques in consideration are Principal Component Analysis, Logistic Regression and Random Forest.

Project Plan

The team aims to have achieved the following milestones by 14th and 24th of October respectively:

- Major two to four clusters finalized with key characteristics identified
- A Predictive Model is finalised to accurately predict a postcode's location

Milestone 1:
By the 10th of October, the team aims to have experimented with cluster analysis by plugging in different sets of data to see if the resulting clusters differ.

Experiment 1: Cluster Results using all data sets
Experiment 2: Cluster Results using only the Income and Rent Data
Experiment 3: Cluster Results using only the Demographics Data

By 14th of October, the team wants to finalize the Clusters that best group the data and find the most similar suburbs in Sydney and Melbourne. The success for this milestone is assessed by the team's readiness to use these clusters for predictive analysis and overall readiness to start working on the next milestone.

Milestone 2:
By 24th of October, the team aims to have experimented with a few predictive models detailed earlier in the Techniques Section. The success criteria for this milestone is 75% - 80% accuracy of the predictive model in predicting suburb's location.

Group Members

Rijwa Abbas (45697795)
Mohammed Fardeen (45496277)
Isha Ahsan (45746699)
Prekshya Chand (45487626)