Introduction

Melbourne is subdivided into hundreds of suburbs, which differ in a number of aspects from geography to health care facilities. Understanding the features of different suburbs and their relationships will be beneficial to the local community as well as the government. Therefore, in this project we use the statistical profiles of each suburb downloaded from the DHHS website, analyze them from different perspectives and share some thoughtful insights.

A1

The analysis has been performed from three different perspectives,namely Population for 2012, Top countries of Birth and Land Use. The features for each similarity measure has been kept in the range of 5 to 12 in order to ensure interpretability and depth of the analysis. Three different similarity measures have been used namely, Euclidean Distance, Mahalanobis Distance and Cosine Distance to compare whether the results vary significantly over different measures. To further broaden the range of the analysis , MDS was used for Euclidean distance and PCA was used for Mahalanobis and Cosine distance.

Measure 1: 2012 Population

In this perspective, we have compared the percentage population of 2012 for all age groups from 0-85+. Percentage was used instead of the actual values to be able to consistently compare among different suburbs as the actual would have to be normalised to perform an accurate comparison. The aim of this measure is to try and analyse whether there is a pattern to the living environments chosen by different age groups and gain a deeper insight into the respective reasons for that pattern based on the analysis.

Measure 2: Top countries of birth (Diversity)

For this analysis, the percentage data for the five top countries of birth was used. The aim of this analysis was to understand whether people from specific countries tend to segregate together in surrounding suburbs and whether there are reasons for this diverse segregation and any outlying factors which may have led them to choose to settle in a specific suburb and its surrounding areas. For this measure, the countries are used as features for each suburb by enumerating them for each suburb and setting the columns to the corresponding percentage.

Measure 3: Land Use of Suburbs

In this perspective, like all other measures, percentage was used in order to eliminate the need for normalisation. The aim of this analysis was to determine the land use division ratio among various suburbs, how it impacts the various services provided and the kind of people it attracts(age group, income group and so on).

A2

Figures 1,2 and 3 provide plots for Measure 1 and it can be seen that although the results vary among different similarity measures, some consistent outliers are Parkville and Sorrento. Parkville was an expected outlier as the segregation of students in the city due to proximity of universities has led to this result. Sorrento was a surprise observation, but on further observation, it became clear that being a tourist location and a location close to the beach may have been the reason for the people of older age to choose to stay in that location.

Figure 7 provides the plot for the diversity measure. Fawkner, Springvale and Braybrook are the most obvious outliers. Fawkner stands out because of the high concentration of Italians in that suburb while Braybrook and Springvale are the hubs for Vietnamese, Chinese and Indian people.

Figures 4, 5 and 6 plot the Land Use measure. Moorabbin and Tyabb stand out as obvious outliers which may be due to the high use of land for industrial purpose in Moorabbin and Tyabb containing huge amounts of rural and land for other purposes and limited use of land for residential purposes.

Most of the suburbs remain similar under each measure, this may suggest that the number of features selected for each measure is not large enough to distinguish an adequate number of suburbs. But considering the near consistency of results among all similarity measures, it may also be interpreted that the distribution of features probably conforms to Gaussian distribution and most of them have values centered around the mean.

A3

The geographic proximity is estimated using the distances between suburbs, that is, how many kilometers away a suburb is from another. To address the hypothesis that geographically closed suburbs are similar, we can make the geographical distances between pairs of suburbs as the x axis, and the similarity score(e.g. the Euclidean distance between the age structures of two suburbs)this pair of suburbs achieve as the y axis. In this way, the relationship between similarity measure and geographic proximity can be visualized clearly.

Figures 8, 9 and 10 clearly show that it is improbable to estimate the similarity of suburbs based on their geographical proximity. Although, in some cases, these may have proved to be true . Yet, in other cases, the results are completely contradictory and there is a huge difference in the suburbs despite their geographical similarity. For land use and age structure, the relationship between similarity score and geographic distance seems more obvious with larger slope for linear model, while for diversity, the slope is so small that the relationship can almost be ignored.

Part B

For the purpose of this exploratory analysis, we have performed two different analysis. In the first analysis, an array of features is chosen by intuition from the dataset and their importance on predicting unemployed rate is shown in Figure 11(calculated using varImp function using logistic regression model in R). It can be seen from the plot that "Aged 75+ and lived alone, %" has the most impact on predicting the unemployment rate. Figure 12 is a linear model fitted for this feature and unemployed rate. A possible reason for this is that when counting the unemployed number, old people are excluded since they tend not to work at their age, thus increasing the ratio of employed people. Another interesting phenomenon is the relationship between unemployment rate, language and education. An assumption that was considered was that unemployment rate may be strongly dependent on education but it can be seen from Figure 11 that they are not that important in predicting the outcome. On the contrary, the features related to language("Born in non-English speaking country,%", "Speaks LOTE at home,%", "Poor English proficiency,%") are all ranked high and a linear model is fitted for unemployed rate and born in non-English speaking country (%)(Figure 13) to show the trend. More importantly, if we assume the other high ranked features such as "Personal income <$400/week, %", "Dwellings with no motor vehicle, %" and "Equivalent household income <$600/week, %" as consequences of being unemployed rather than the cause, we may conclude that the major reason people get unemployed is because of poor language proficiency.

In the second analysis, we have tried to observe whether similar set of features among suburbs can be used to analyse the top occupations in a particular suburb. We have used a set of features for this analysis. The English subset contains features such as Poor English Proficiency,Born Overseas, Speaks Language other than English and Born in non- English speaking country. On the other side, we have the social standards and land use subset which checks the Personal income<400$, Dwellings with no Internet,Dwellings with no motor vehicle, industrial and commercial land use. The purpose of this analysis is to determine whether it is possible to predict whether suburbs with similar results for these features have the same occupations as the top occupations. As is seen from figure 14, the suburbs with similar results for a similar set of top occupations tend to cluster together. This has been illustrated in the IPython script. For instance,Mordialloc and Sorrento clustered together and they have the same set of Top 3 occupations.Similar, results were observed with Toorak,Waterways and Windsor.Yet, there were a few contradictions as well as in the case of Mordialloc and Murrumbeena which were parts of different clusters despite having same set of top occupations. On further analysis and comparison, it appears that the same set of occupations in Top 3 usually clustered together with a few occasional inaccuracies which could be improved by further changing the features and adding more suburbs data to improve clustering. Although, this started as a mere analysis, we were surprised with the accuracy of the clusters that we received and hope that it would give us a further insight into how to conduct plausible exploratory analysis to gain deeper insight into data as well as the field of unsupervised learning.

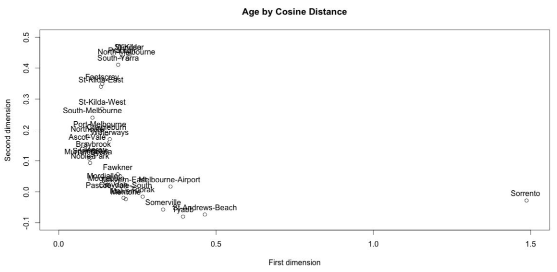


Figure 1. Age Structure by Cosine Distance

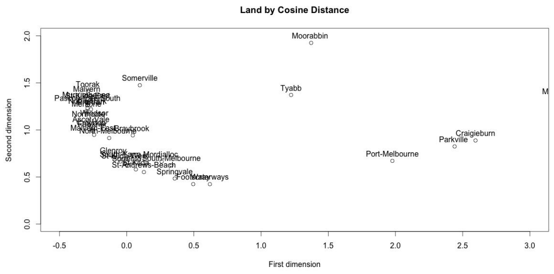


Figure 4. Land Use by Cosine Distance

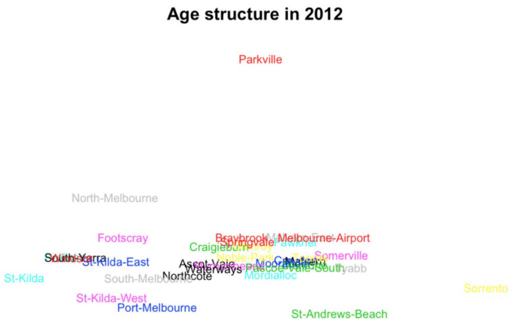


Figure 2. Age Structure by Euclidean Distance

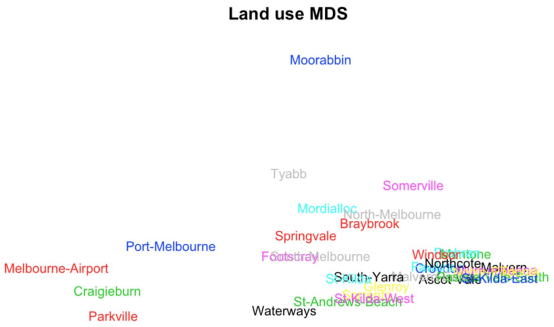


Figure 5. Land Use by Euclidean Distance

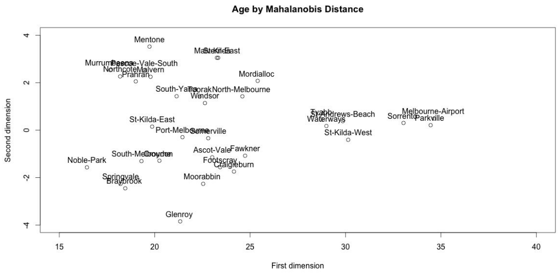
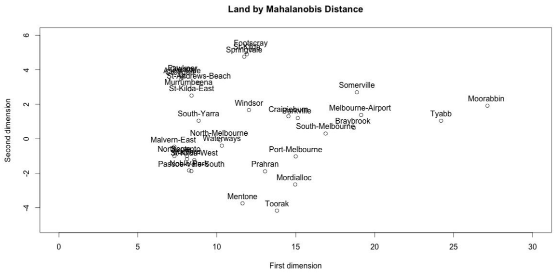
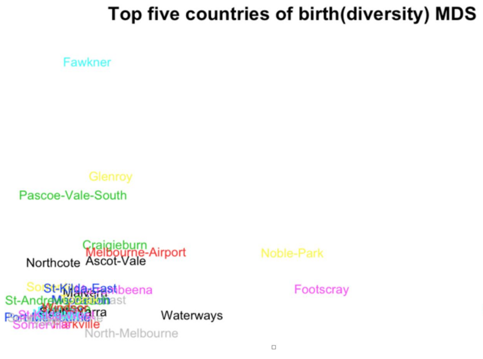


Figure 3. Age Structure by Mahalanobis Distance Figure 6.Land Use by M. Distance





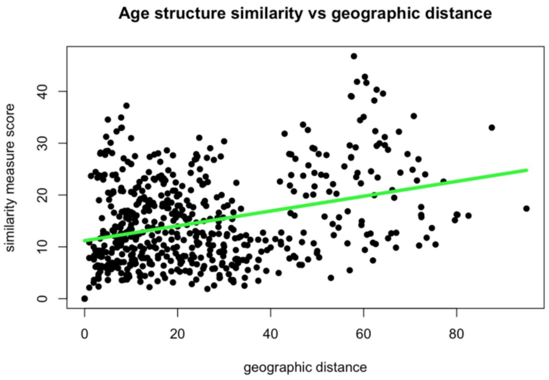
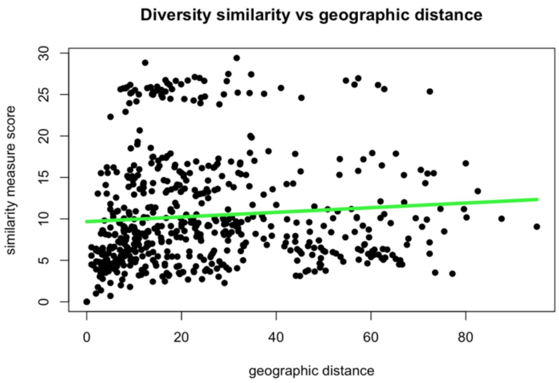


Figure 7. Diversity by Euclidean distance Figure 8. Age similarity & geographic distance



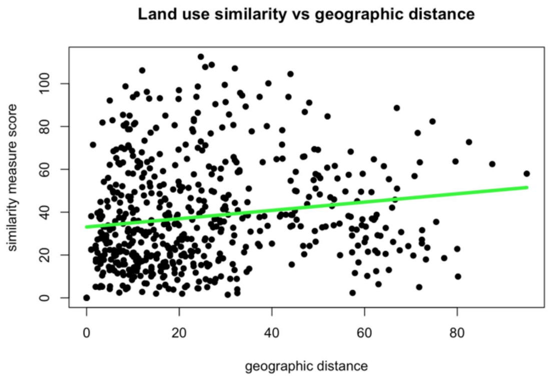


Figure 9.Diversity & geographic distance

Figure 10. Land Use & geographic distance

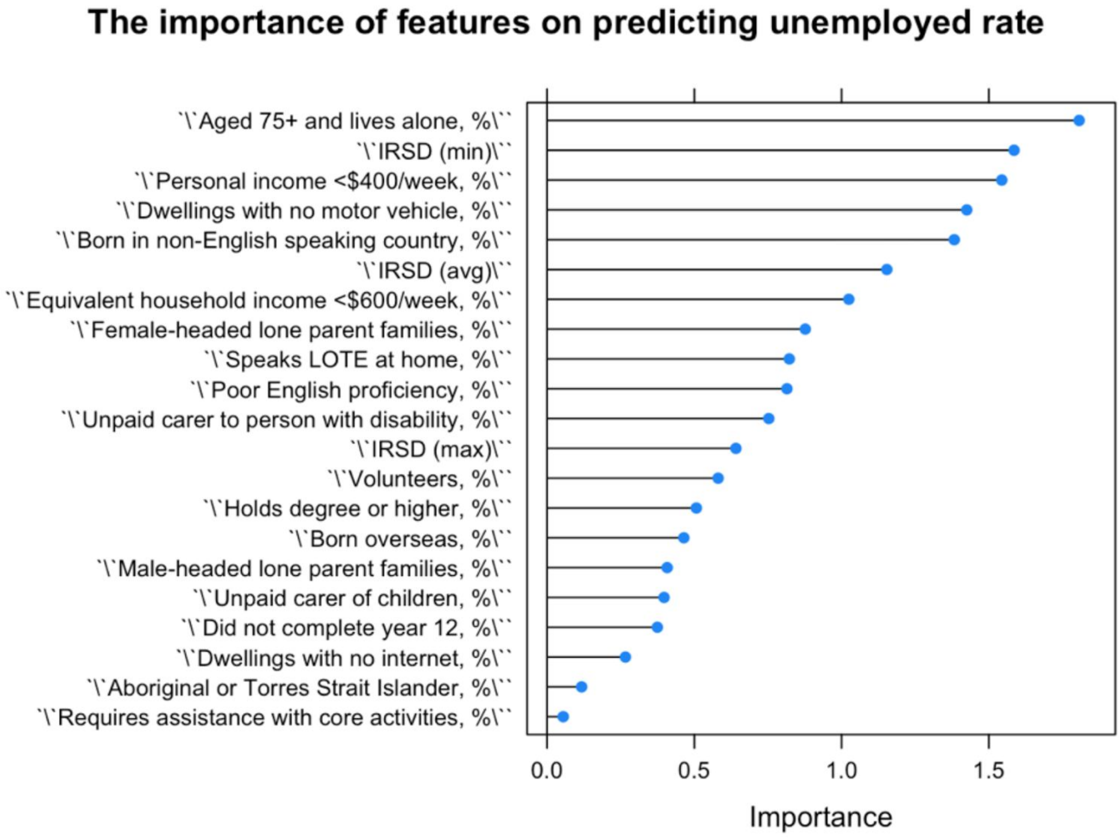
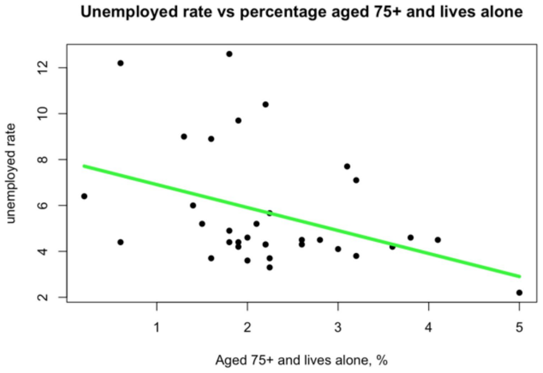


Figure 11. The importance of features on predicting unemployed rate



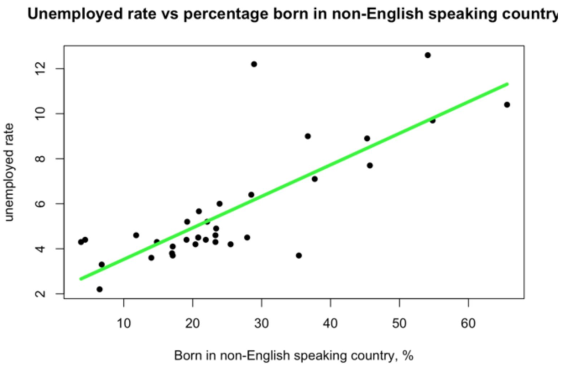
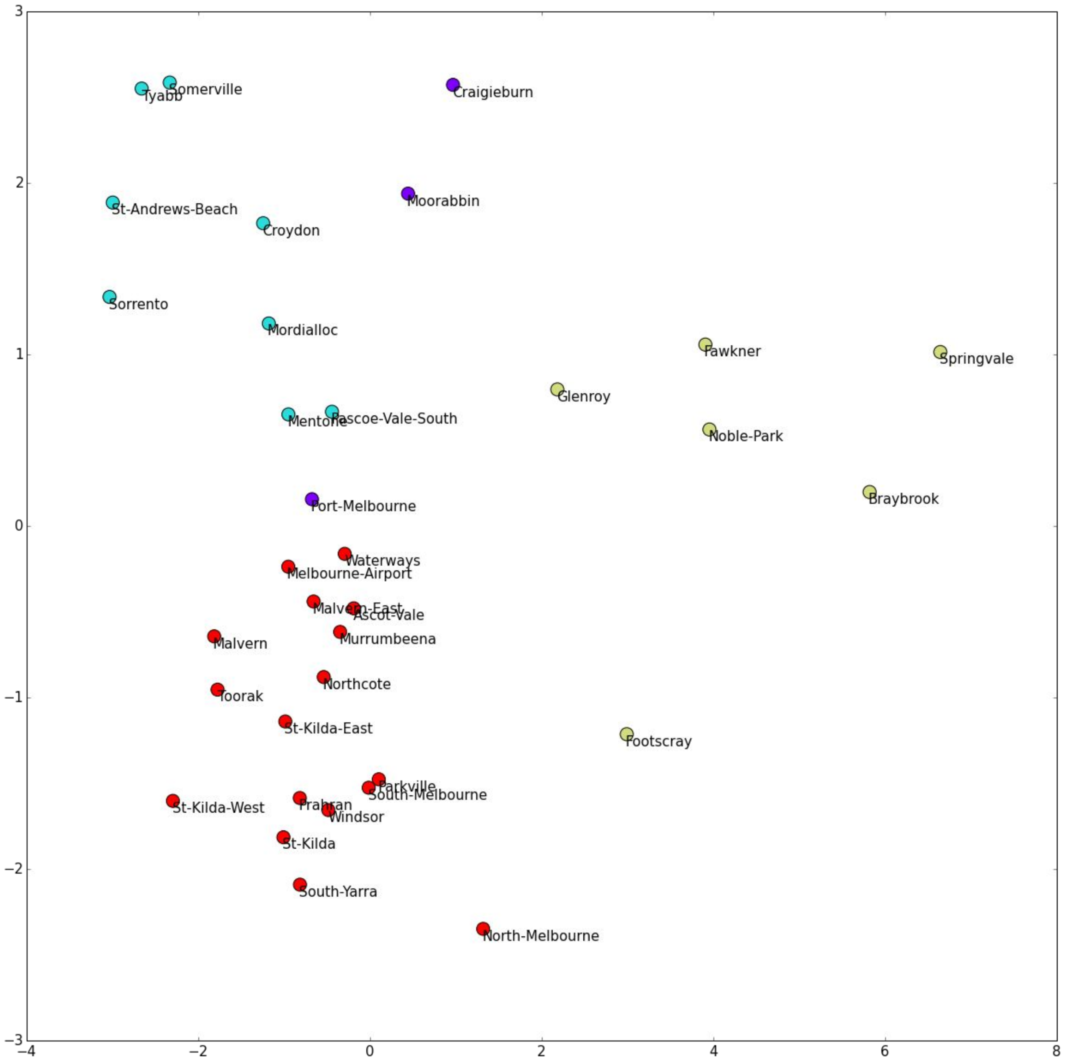


Figure 12. Unemployment & 75+ lives alone(%) Figure 13. Unemployed rate & born in non-English speaking country(%)

Figure 14. Predicting Occupation based on Suburb Information


Tyabb
Somerville
Craigieburn
St-Andrews-Beach
Croydon
Moorabbin
Sorrento
Mordialloc
Fawkner
Springvale
Glenroy
Mentone
Pascoe-Vale-South
Noble-Park
Braybrook
Port-Melbourne
Waterways
Melbourne-Airport
Malvern-East
Ascot-Vale
Murrumbeena
Malvern
Northcote
Toorak
St-Kilda-East
Footscray
Parkville
South-Melbourne
St-Kilda-West
Prahran
Windsor
St-Kilda
South-Yarra
North-Melbourne