CP 318: Data Science for Smart City Applications

Project 2 Report(Team -EPSILON)

Question A1. **Similarity Measures:**

1) **Land Use**: This category tells about how the land usage is distributed between different sectors such as commercial, residential, industrial, etc. As an additional feature, we added the population density and total area of the suburban which we felt would have an immense impact over the land usage.
   Aim: To analyze how the land usage is distributed among suburbs and find any similarity if any.
   Justification: Choice of features is such that, suburbs with similar total area or similar population density will have similar land usage distribution.
   Method: Used Manhattan distance as the similarity measure of comparison. The labels are converted to integer tokens. Null values are filled by zeros.

2) **Services**: This feature talks about the facilities and utility resources present in a suburb. We used all the features in the category.
   Aim: Services points about amenities a place has and therefore also indirectly points towards the cost of living in a suburb. This feature becomes important when we want to build some new facility in the area or want to buy a house in the area. As all data was present in numerical form, so there was no need of explicitly transforming the data.
   Method: Used Euclidean distance as the similarity measure of comparison.

3) **Socio-Demographic**: It talks about how a suburb has developed as a whole and how much land is occupied for what purpose such as Commercial, Industrial, Residential, Rural or something other. As all data was present in numerical form, so there was no need of explicitly transforming the data.
   Aim: This measure tells us proportion of land used for each purpose which can be further helpful in policy development decisions such as there should be more hospitals and schools in suburbs where proportion of Rural and Residential is more than others.
   Method: Used Minkowski distance with p = 3 as the similarity measure of comparison.

Question A2. **Multidimensional Scaling method:**

1. **Land Use**: Figure 1 depicts the MDS scatter plot of Land use category using Manhattan Distance
2. **Services**: Figure 2 depicts the MDS scatter plot of Services category using Euclidean Distance
3. **Socio-Demographic**: Figure 3 depicts the MDS scatter plot of Socio-Demographic category using Minkowski Distance.

- Some suburb pairs look similar in all the MDS such as St Kilda and South Yarra, North Cote and Malvern East, Tyabb and Somerville are found to be of same proximity under all the three MDS
- We expected that the suburbs which have similar land use should have somewhat similar services available but we observed some suburbs conflicting this claim such as Prahran and Windsor, St Andrew Bridge and Port Melbourne and Parkville and Springvale, Prahran and Windsor, Malvern East and Noble Park are near in Land usage but far in Service MDS this could be due to the fact that some areas are more developed as compared to other and therefore providing more services.
- We Expected the suburbs which are near in Socio Demography would be near in Geography Measure but there were some conflicts like Noble Park and Glenroy, St Kilda and South Yarra, Springvale and Braybrook, Northcote and Malvern East as these were found to be far in Geography but near in Socio Demography

Question A3. **Geographical Proximity:**

Figure 4 and 5 depicts the MDS scatter plot of Geographical location of suburbs using two different proximity measures – Euclidean and Minkowski (p = 4) distance.

Method: The location feature is decomposed into two different features – distance and direction. Distance says how far the suburb is from Melbourne (in kms) and Direction feature says the angle in which the suburb is present from Melbourne (in degree). These are scaled using MinMaxScaler.

The distance between suburbs in these plots are compared between the distance between the suburbs obtained in other category. It is found out that the similar suburbs are closer in proximity, which are depicted in the Figure 12.

Question B1.

Feature Selection using Supervised Learning:

We selected most important features from all the features for a given suburb using the Feature selection technique of using Supervised Learning technique – Logistic Regression on the suburb prediction task along with L1 regularizer with large regularizing constant lambda.

Result: We obtained 34 features from the model with lambda = 50.

Important Features: `'Population Density', '2012 ERP age 20-24, persons', '2012 ERP age 25-44, persons', '2012 ERP age 45-64, persons', '2012 ERP, total', '2007 ERP age 20-24, persons', '2007 ERP age 25-44, persons', '2007 ERP age 45-64, persons', '2007 ERP age 85+, persons', '2007 ERP, total', '% change, 2007-2012, age 80-84', '% change, 2007-2012, age 85+', 'Number of Households', 'Occupied private dwellings', 'Population in non-private dwellings', 'Public Housing Dwellings', 'Personal income <$400/week, persons', 'IRSD (min)', 'IRSD (max)', 'IRSD (avg)', 'Holds degree or higher, persons', 'Did not complete year 12, persons', 'Volunteers, persons', 'Unpaid carer of children, persons', 'Born overseas, persons', 'Born in non-English speaking country, persons', 'Speaks LOTE at home, persons', 'Top country of birth, persons', '2nd top country of birth, persons', 'Top language spoken, persons', '2nd top language spoken, persons', 'Public hospital separations, 2012-13', 'Presentations to emergency departments, 2012-13', 'Category 4 & 5 emergency department presentations'`

Unsupervised Clustering using Selected Features:

We found out 5 clusters of suburbs present using the obtained important features which can seen in Figure 6 and Figure 7

Question B2.

Figure 8, 9, 10 and 11 shows that using these 34 features, all the other features and development of a suburb can be found out. The figures show that each cluster having different land use distributions and the suburbs in a cluster are almost from the same neighborhood.

Question B3.

This analysis helps in finding the optimal parameters for development of a suburb. So that the government while doing the city planning can consider these features and plan for the future development by which the infrastructure and service facilities of the suburbs always remain optimal.

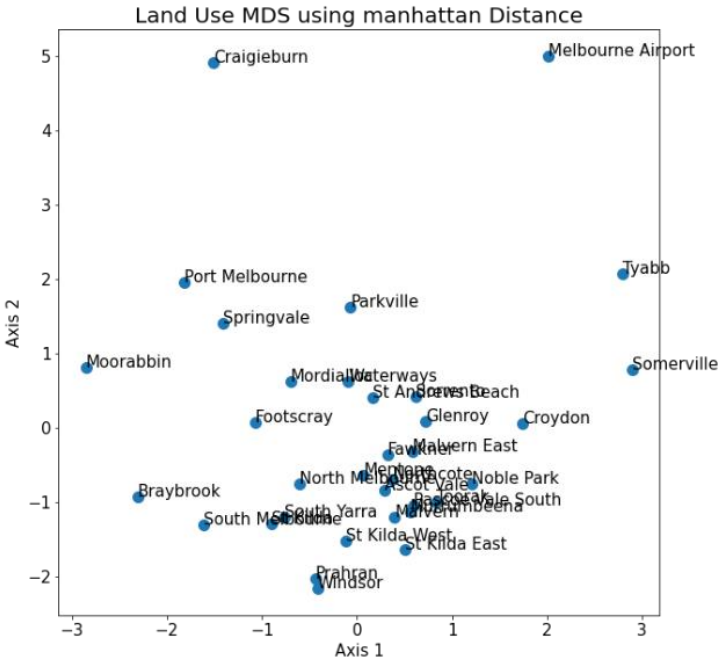

Figure 1

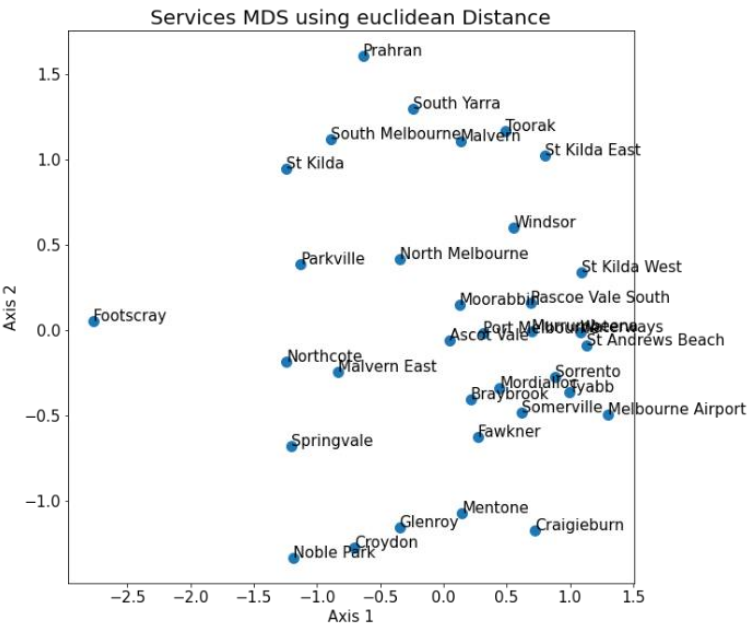

Figure 2

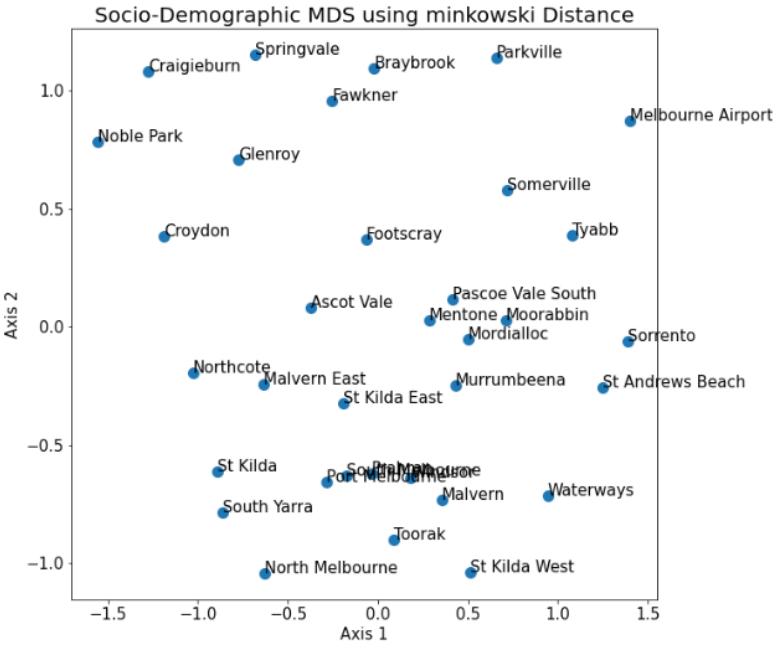

Figure 3

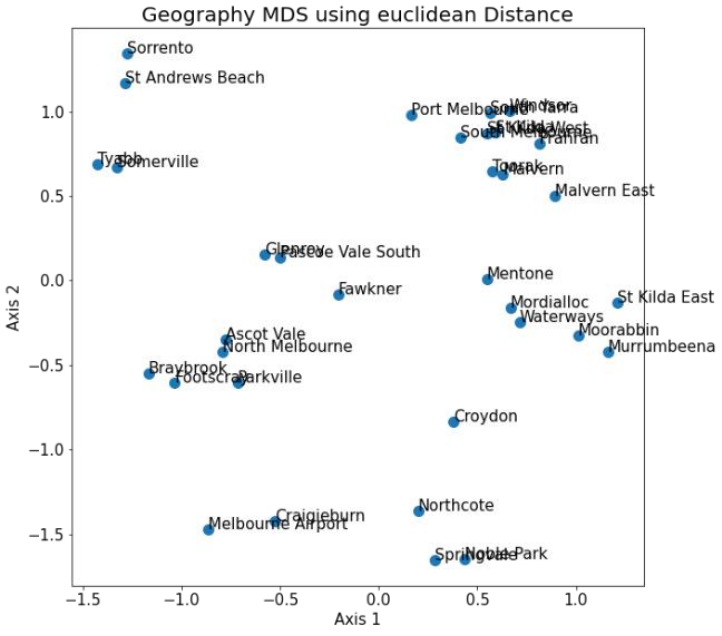

Figure 4

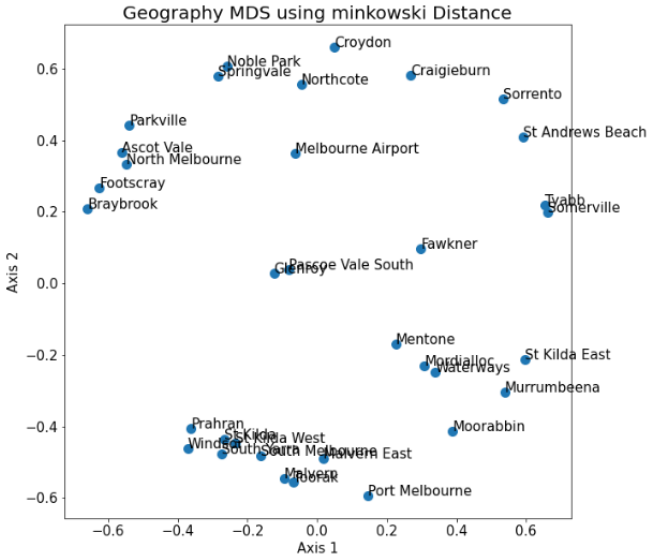

Figure 5

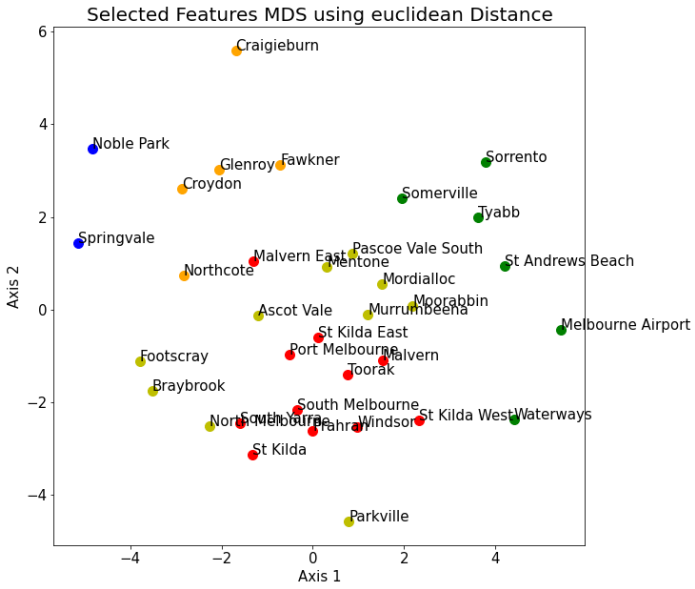

Figure 6

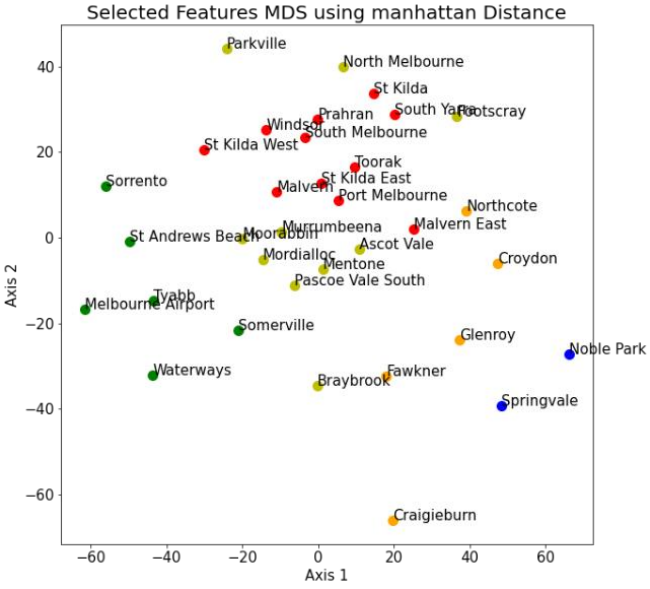

Figure 7

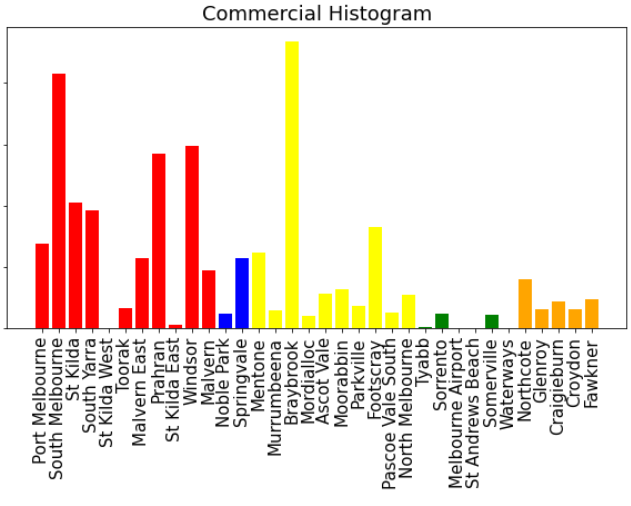

Figure 8

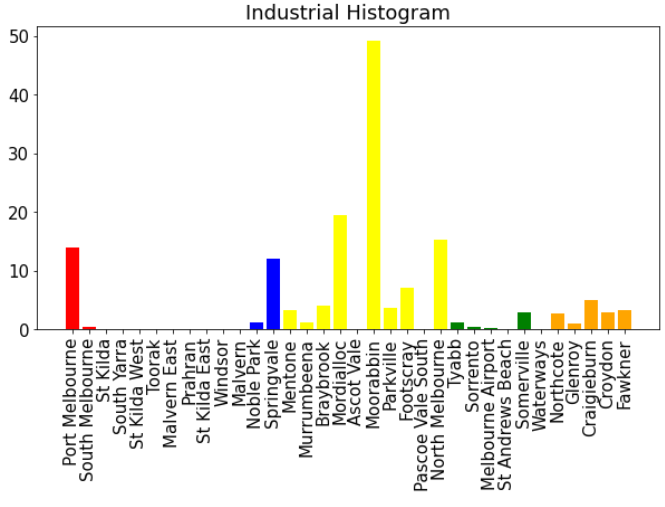

Figure 9

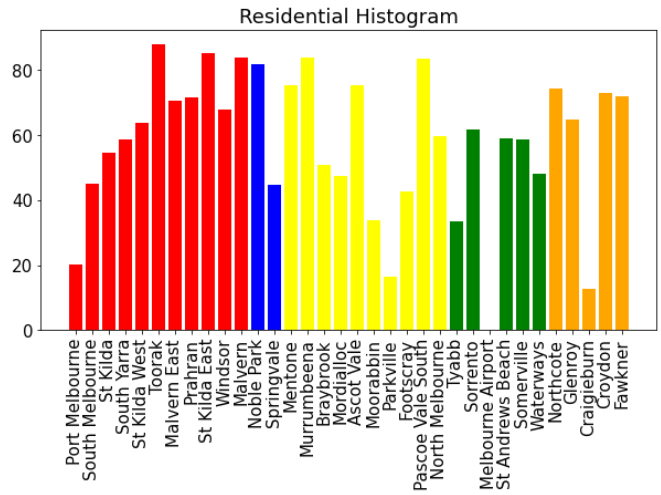

Figure 10

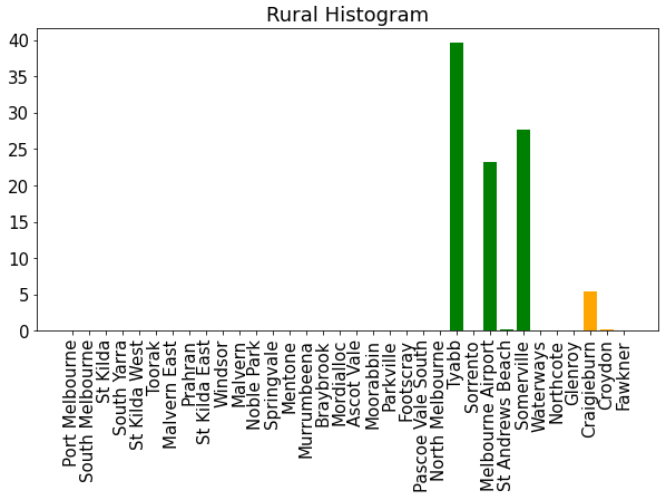

Figure 11

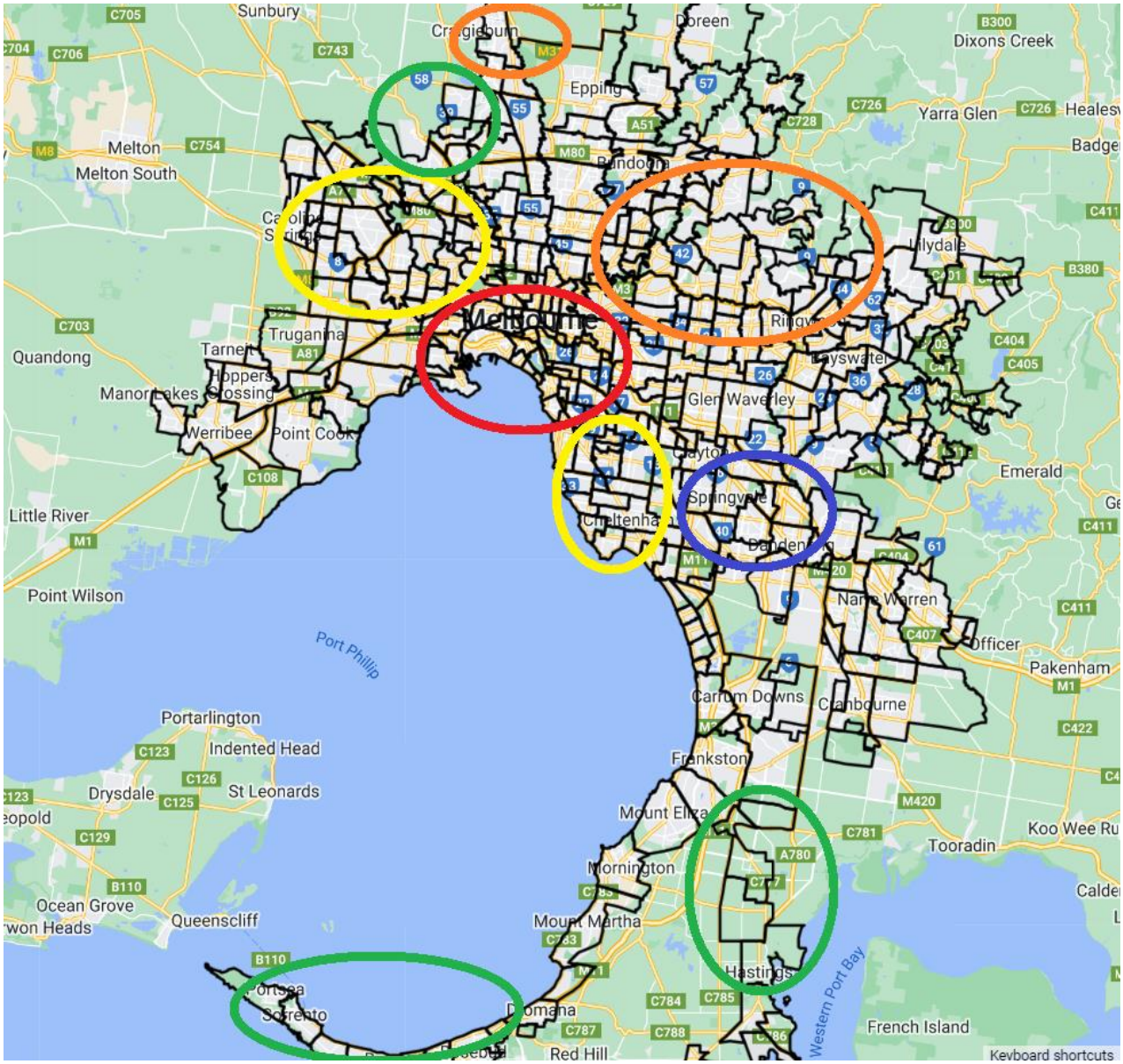

Figure 12