What Factors Influence the Disparities in Obesity Amongst the Municipalities of Metropolitan Melbourne?

Research Question and Relevance to Liveability, Inclusiveness, Health and Sustainability

Our research aims to explore and compare the social and environmental factors that correlate to obesity amongst LGAs of Metropolitan Melbourne. This analysis will assist policymakers in recognising the severity of obesity and potential reform strategies that will assist affected regions. Additionally, this topic addresses liveability and inclusiveness, exploring factors like accessibility to sports and fitness facilities, diet and proximity to CBD.

Datasets Used and How They Are Linked

The report contains datasets from ABS, DHHS and VicHealth. They include ordinal, nominal, discrete and continuous numerical data. They can be found below in the **References** section.

The datasets were chosen as they contained the key attributes of distance to CBD, physical activity rates, proportion of gyms, soft drink consumption and median income. These factors provide strong insights into the variance of obesity within Melbourne.

The key attributes of datasets were linked through social, economical and environmental concerns. For instance, comparisons between income and engagement in fitness centres produced insights into correlations between socioeconomic status and obesity rates. Furthermore, the pairing of gym frequency within an LGA to its geographical area allowed us to contrast the proportion of gyms from different regions without the bias of geographical size.

Wrangling and Analysis Methods

Data cleaning was applied to the collected datasets by using regular expressions for removing noise and multi-valued attributes. Data reduction was performed for removing irrelevant variables beyond the scope of our study as well as filtering excess municipalities beyond metropolitan Melbourne, allowing us to combine relevant attributes into a single CSV file for common use. Moreover, data integration was achieved by merging datasets using LGA names as a primary key. Data transformation was implemented by producing visualisations such as scatter plots, linear regression plots and heatmaps necessary for deducing trends and meaningful interpretations.

Scatter plots were the most versatile. They handled large sets of continuous data and were useful to ascertain trends and similarities among specific LGAs. Histograms and bar graphs did not allow the distinction of specific LGAs based on categorised distances and were not included.

A mutual information model built using machine learning with an 80:20 train-test split was used to identify dependency strengths among variables that showed non-linear correlations. The same split was also applied to linear regression models. Splitting was necessary to mitigate overfitting for obtaining accurate MI and R^2 values.

For linear data sets, Pearson's correlation (R) and linear regression (R^2) were used more frequently than mutual information (MI), as most visualizations suggested linearity between two test variables. These methods, however, were not applicable to non-linear data points. Linear regression was also more visually constructive than Pearson's correlation. This method measured the strength and direction of the relationship between two variables, whilst considering outlier variability via the distance between data points and the regression line.

K-means and hierarchical clustering were two clustering methods used in the study for observing homogenous groups given a test variable. They both analysed the similarity between LGA labelled data points based on euclidean distances. Hierarchical clustered heat maps were used to group LGAs based on shared characteristics. Dissimilarity matrices were considered, but specific LGAs could not be ascertained from them.

All aforementioned techniques were implemented using Scikit-Learn, Matplotlib, Seaborn, Re and Pandas. These tools were necessary for efficiency and accuracy of data processing.

Key Results, Their Significance and Value

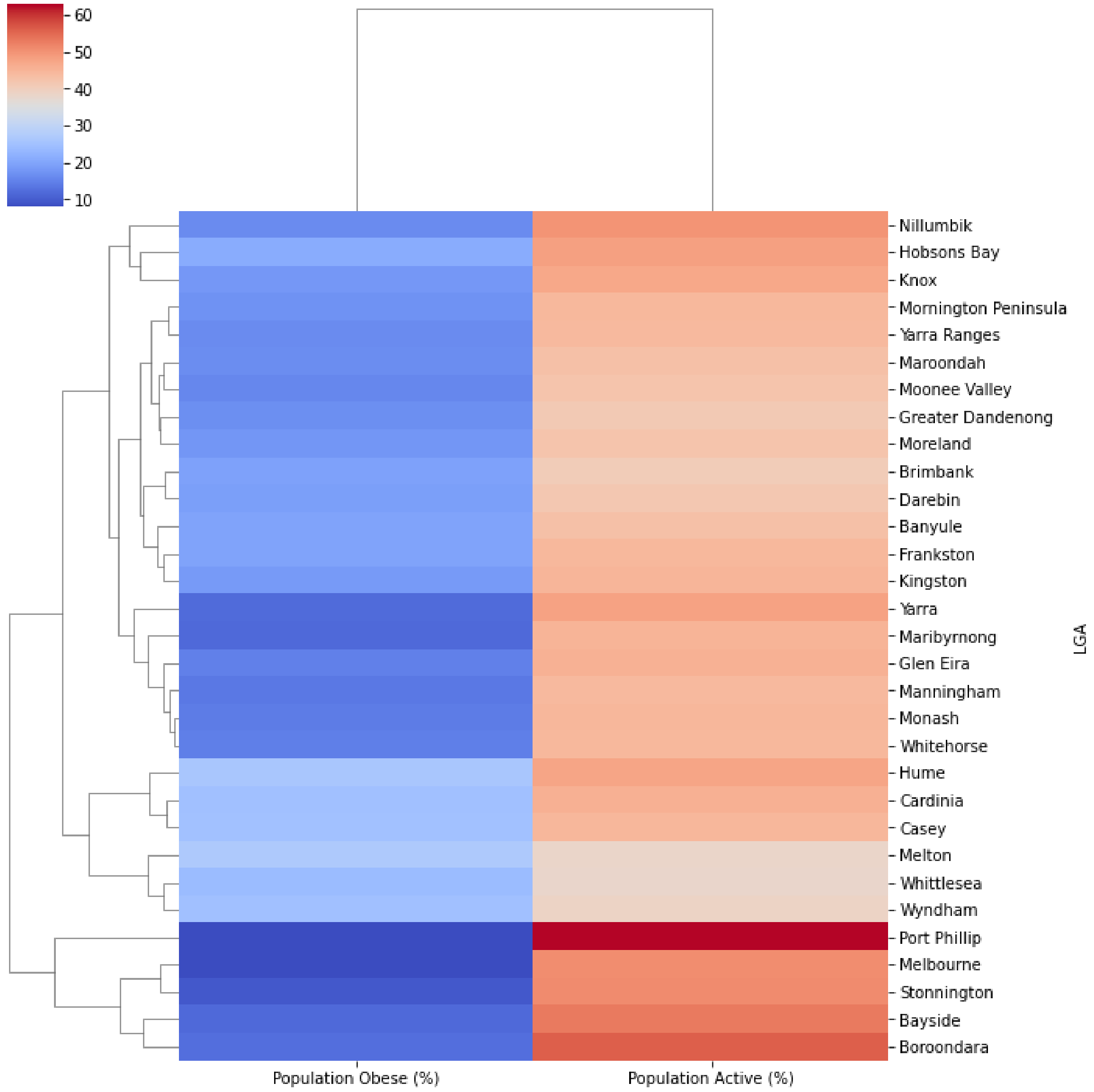

Figure 1. *Hierarchical clustered heatmap for population obese and activity rates.*

Figure 1 compares obesity rates and physical activity amongst the metropolitan LGAs. It illustrates the trend that LGAs closer to Melbourne CBD have the lowest obesity rates and higher activity levels compared to the more distant residential localities. By slicing the dendrogram at level 4, LGAs with higher obesity rates and lower physical activity are clustered at the lower-central region of the heatmap (bright blue, pale orange). LGAs with low obesity percentages and high physical activity rates are grouped at the bottom and the top 80% of the heatmap. They contain subclusters with even lower obesity rates.

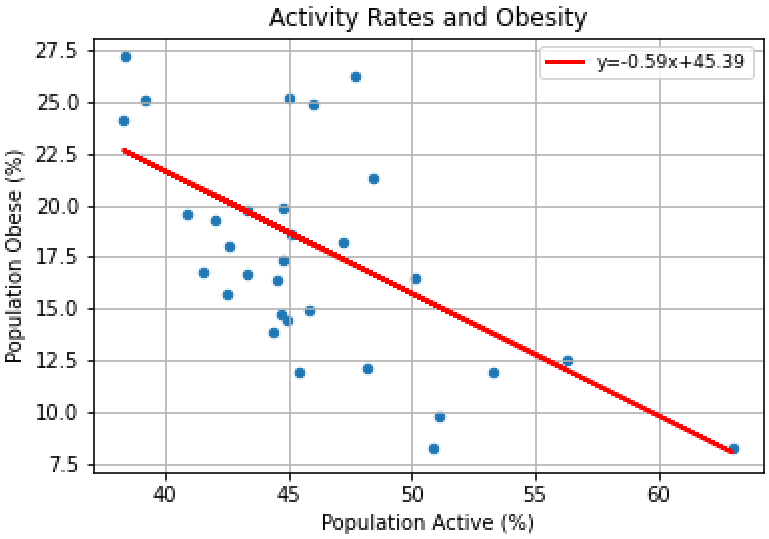

Figure 2. *Linear regression between population activity and obesity rates.*

MI and linear regression suggested a moderately strong, negative correlation between obesity rates and activity rates ($R^2 = 0.37, MI = 0.32$). According to **figure 2**, active LGAs have significantly lower obesity rates than those with lower physical activity. Linear regression predicted physical activity rates below 35% would correspond to a 27.7% obesity rate, in which the outer suburban municipalities are at higher risk of falling within this margin as observed by **figure 3** and the green cluster in **figure 4**.

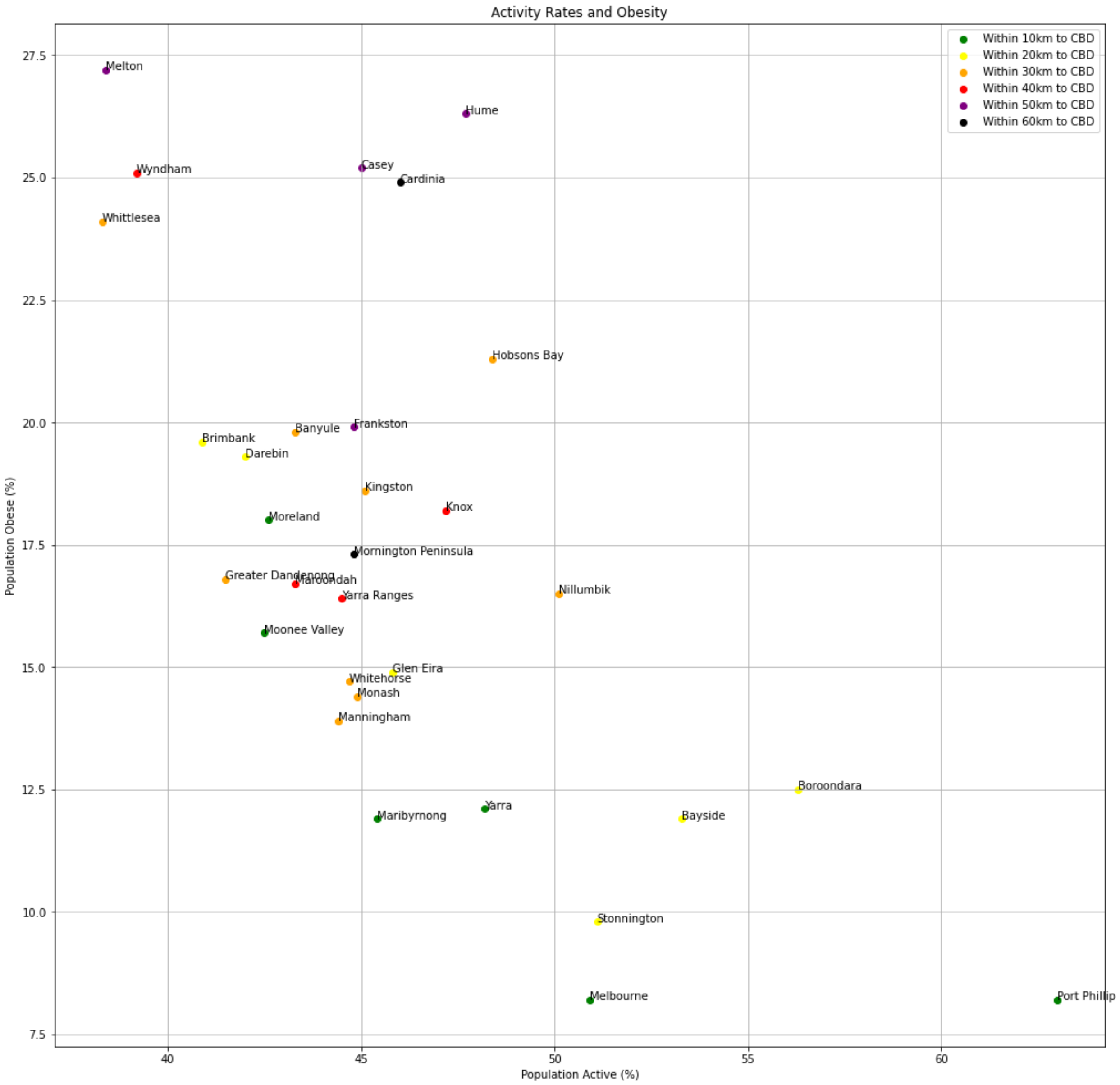

Figure 3. *Scatter plot for population activity and obesity rates organised by distance to CBD.*

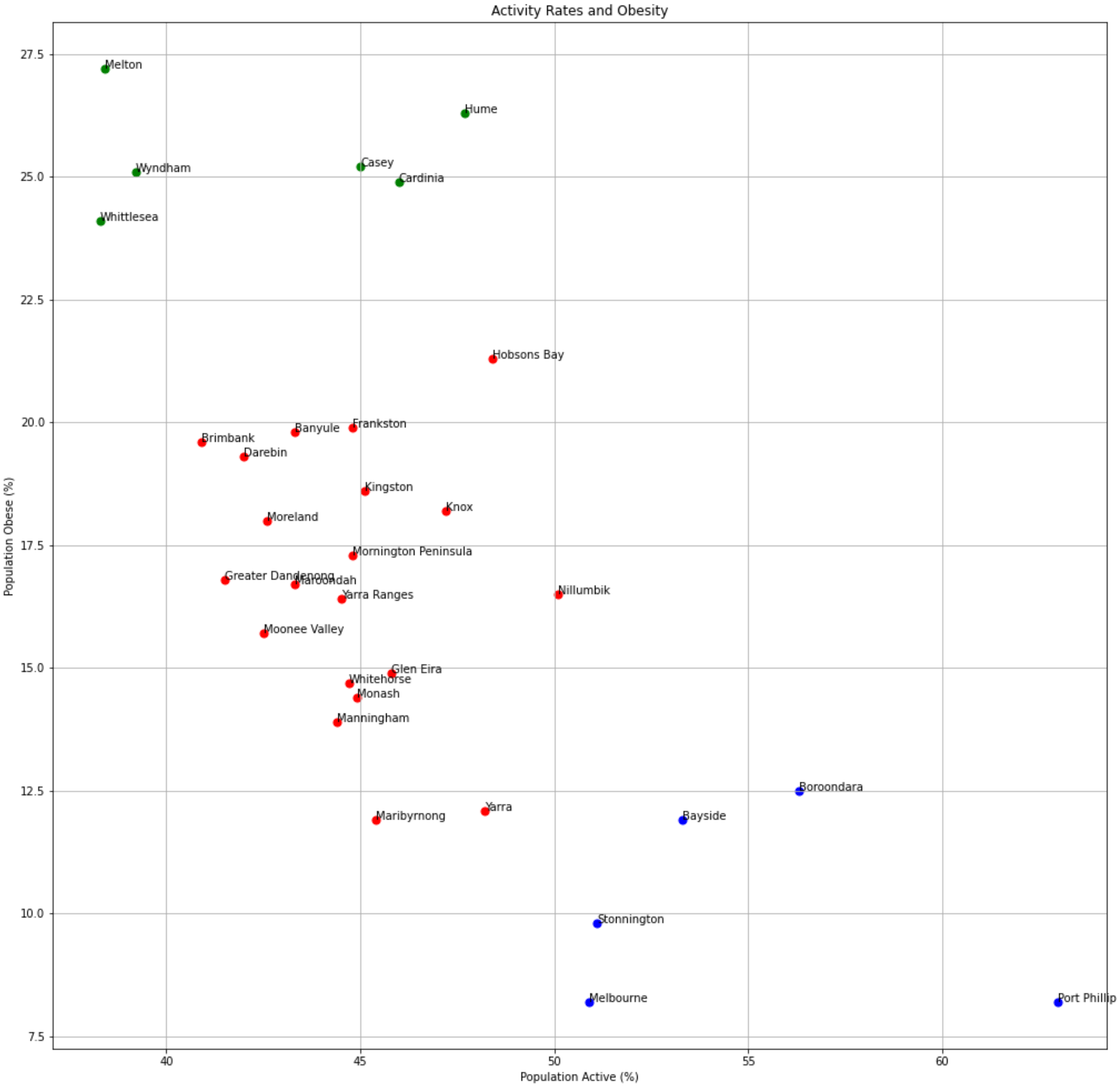

Figure 4. *3-means clustered LGAs for population activity and obesity rates.*

Figure 3 indicates that LGAs closer to the CBD have significantly lower obesity rates and higher activity rates than LGAs further away, an implication that inner municipalities are more active and healthy. LGAs located within 30 kilometers of the CBD are constricted about obesity and activity percentages of 14%, 20%, 40% and 50% respectively. Suburbs within 50 kilometers such as Melton, Cardinia, Casey and Hume have the highest obesity rates in Metropolitan Melbourne.

Cluster	Type
Green	Large residential areas
Red	Inner suburbs
Blue	Affluent townships

Table 1. *Classification of LGA clusters for* **Figure 4***.*

Coordinating **figures 4, 5 and 11**, it is evident that most LGAs within the blue cluster have more gyms; indicative of their ability to afford gym memberships as well as other programs that promote physical activity.

These results are significant because it outlines the obesity rate biases towards large residential LGAs influenced by their low percentages of physical activity. The information is valuable as it drives policymakers to propagate physical activity and to incentivise sporting and gym facilities in LGAs with high obesity risks.

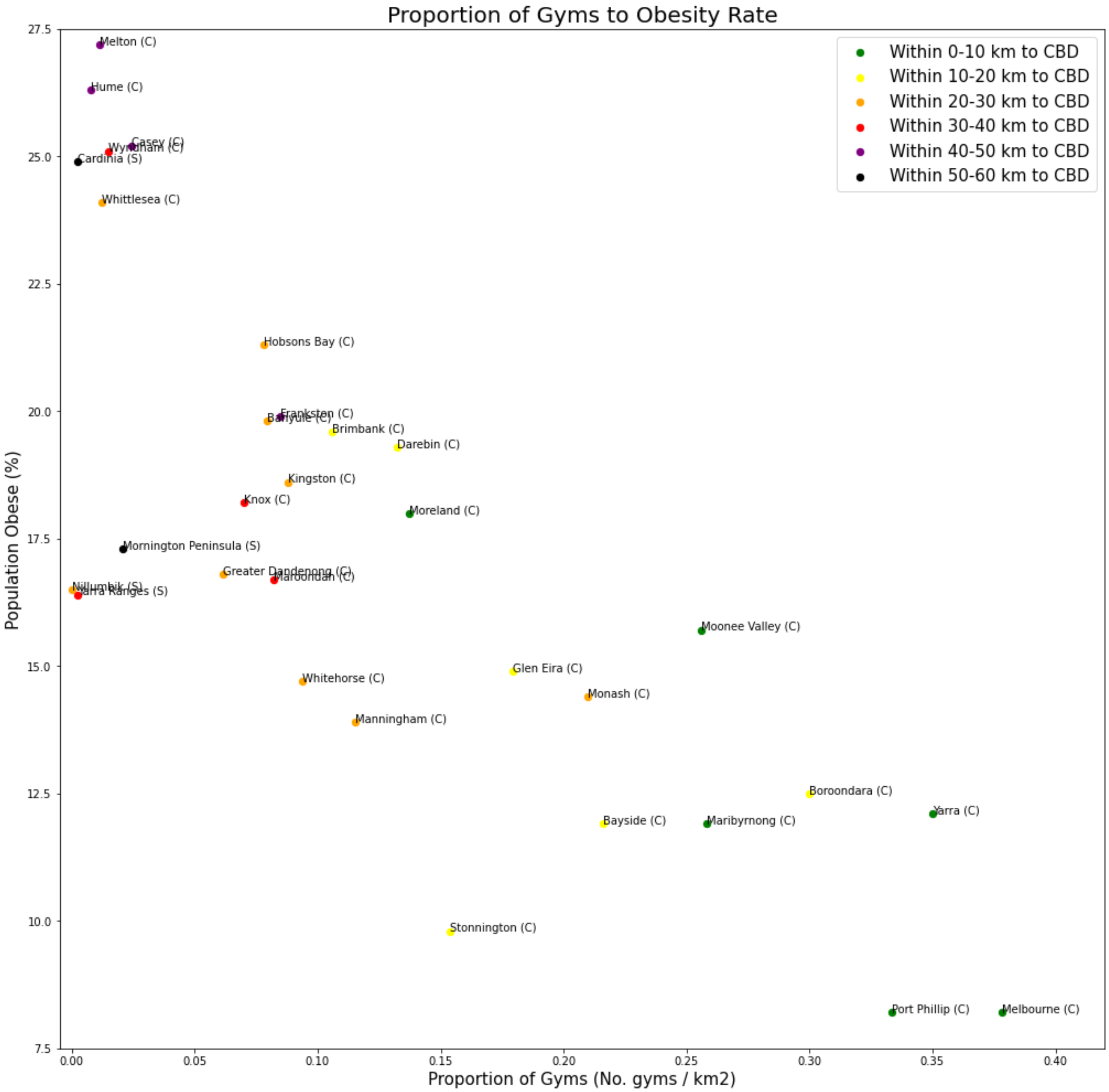

Figure 5. *Scatter plot for proportion of gyms and obesity rates organised by distance to CBD.*

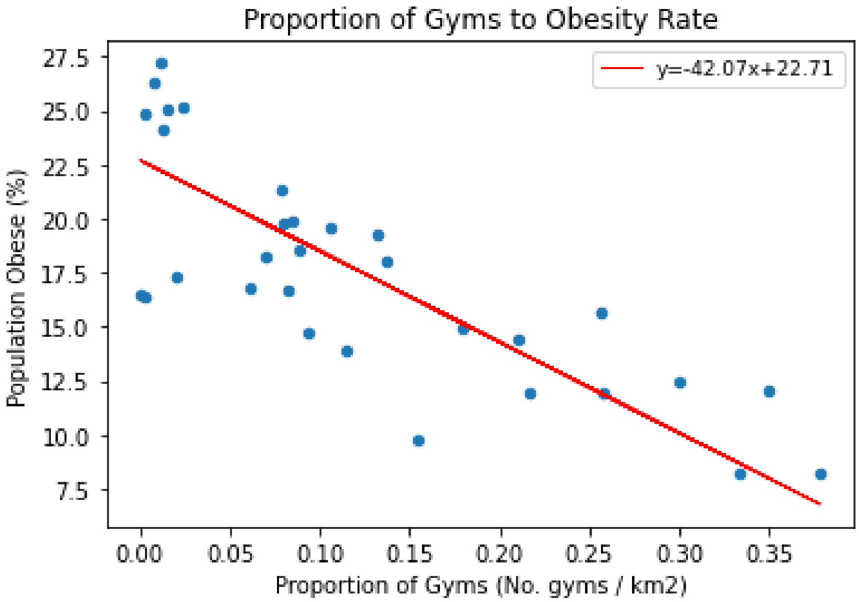

Figure 6. *Linear Regression between proportion of gyms and obesity rates.*

The findings comparing activity rates to obesity are replicated when exploring the proportion of gyms within an LGA as seen in **figure 5**. Again, municipalities closer to the CBD contain a greater number of gyms to land size as indicated by the green and yellow clusters. Inner LGAs have lower obesity rates than the larger LGAs, which contain a lower proportion of gyms indicated by the black and purple data points. **Figure 6** showcases the strong inverse correlation between gym proportion and obesity rate (R^2 = 0.61).

Thus, it can be seen that a greater proportion of gyms corresponds to a high rate of exercise, which in turn correlates with lower obesity rates. The significance of these findings will provide insights to policy makers as guidance to promote fitness and exercise.

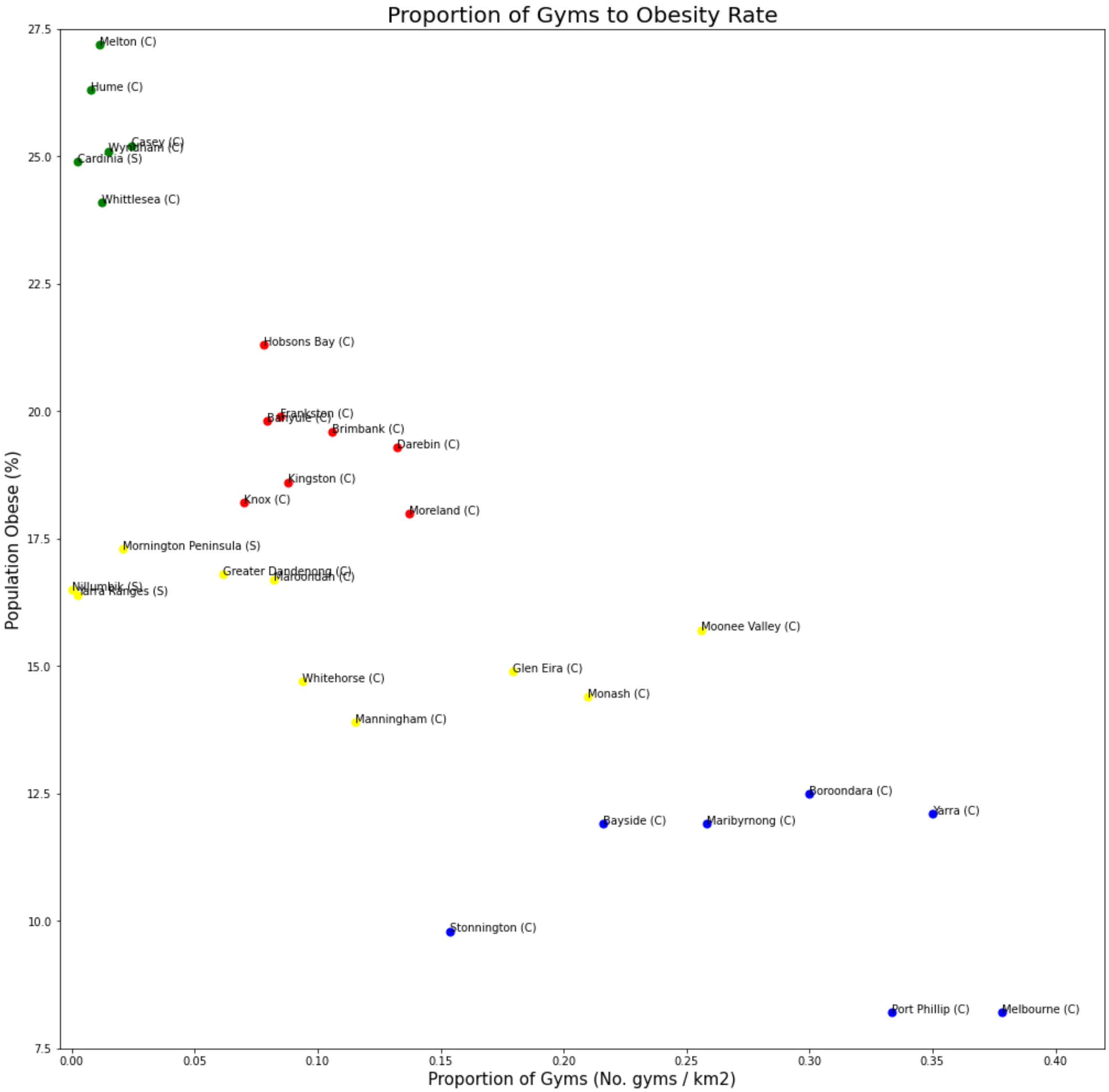

Figure 7. *4-means clustered LGAs for proportion of gyms to obesity rate.*

Clustering analysis of the distribution of gyms across Melbourne LGAs showcases the disparities of regions which feature a smaller proportion of gyms. The high density of the green cluster which centres around the upper-left suggests that regions with minimal gyms collectively have a very high obesity rate. The significance of these insights will assist policymakers in identifying the LGAs which showcase these discrepancies. However, as the number of gyms increase, a greater spread with clusters are observed. This suggests there may be unknown confounders influencing obesity rate.

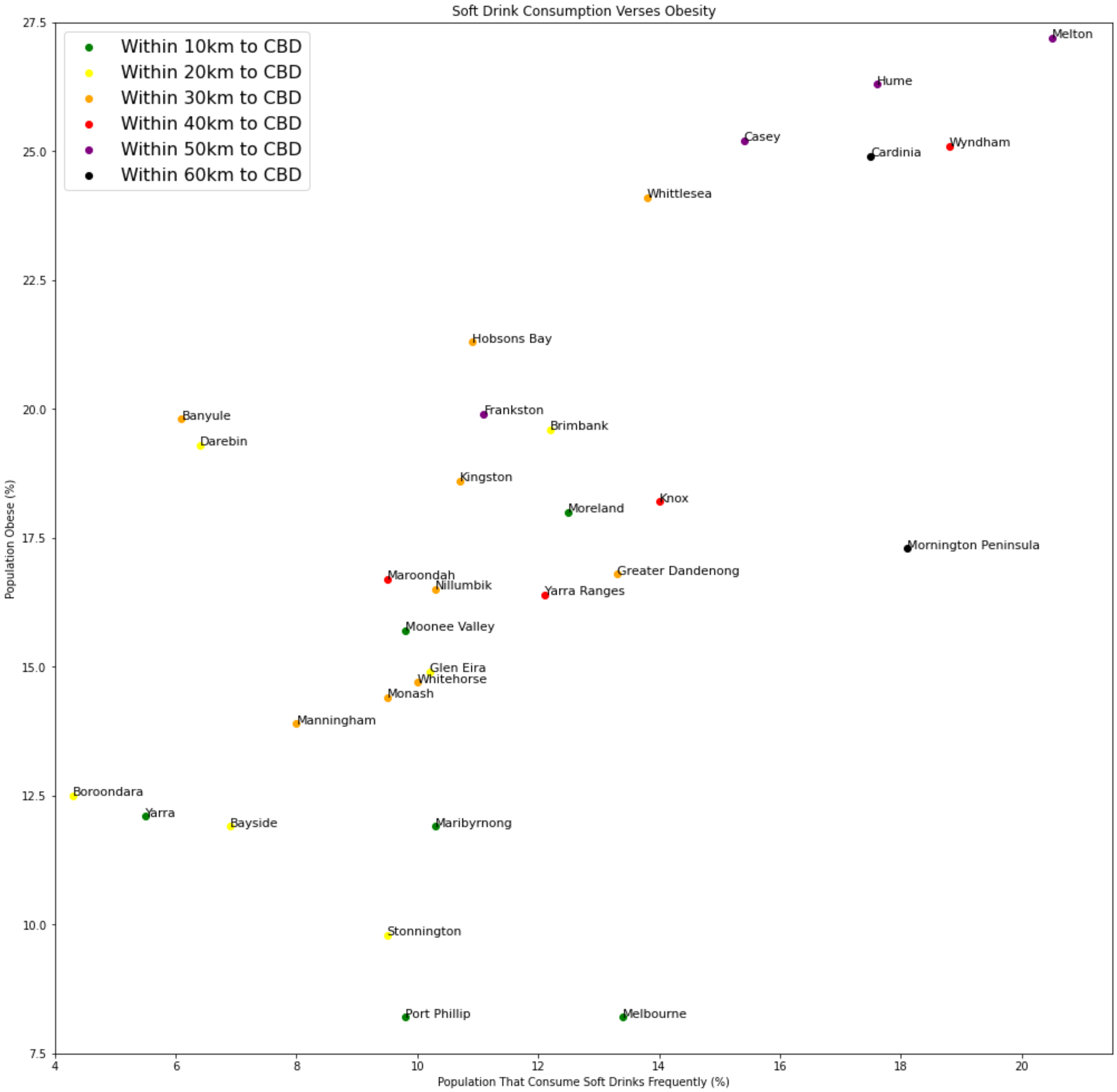

Figure 8. *Scatter plot for soft drink consumption and obesity rates.*

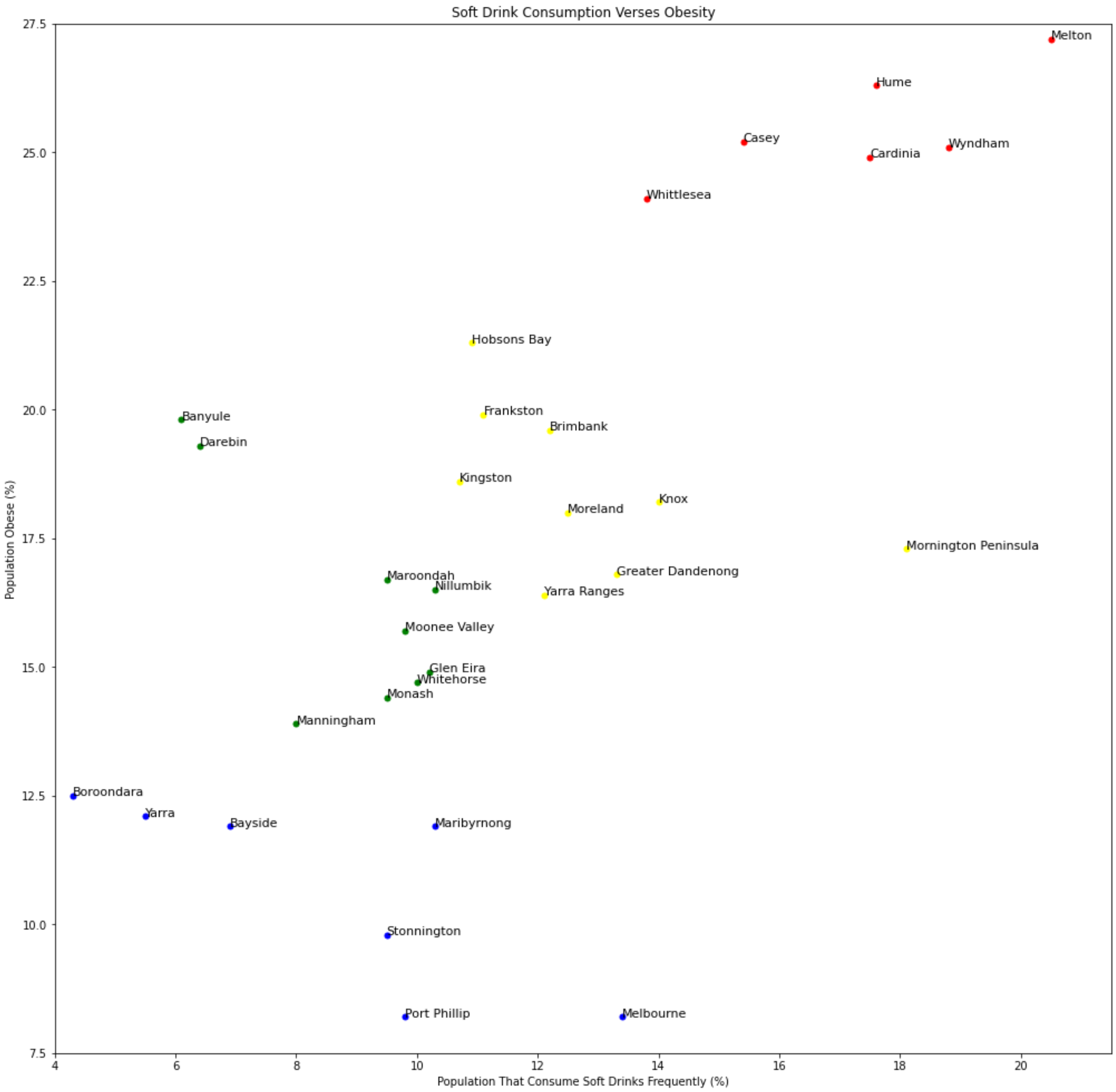

Figure 9. *4-means clustered LGAs for soft drink consumption and obesity rates.*

By observing **figures 8 and 9**, data points for municipalities within 10 and 20 kilometers to the CBD are in the yellow cluster, indicating a healthier index. A large portion of LGAs in the 30 and 40 kilometer radius are located in the blue and green clusters respectively. This underscores that municipalities within the 30 and 40 kilometer ring contain a similar percentage of obese people, but the soft drink consumption from the latter municipalities is greater. LGAs beyond the

50 kilometer radius fall in the red cluster, with the highest percentage of obesity and soft drink consumption.

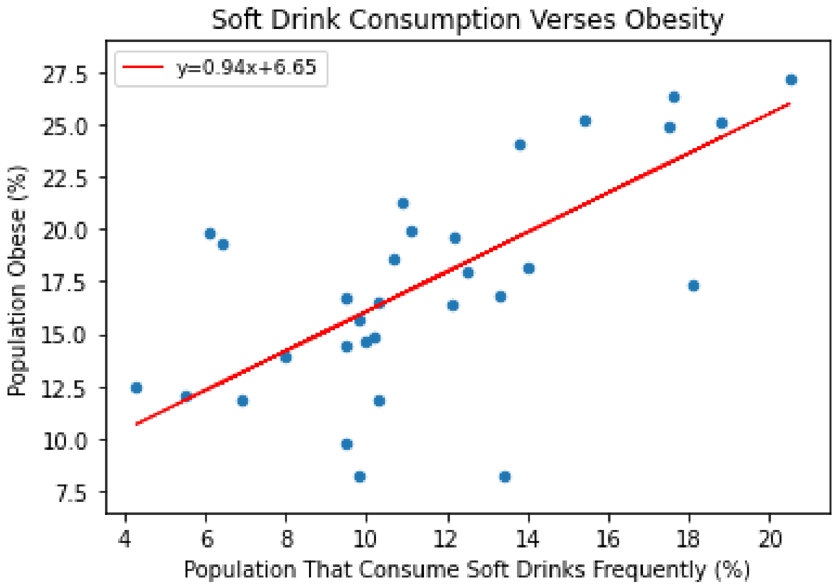

Figure 10. *Linear regression between soft drink consumption and obesity rates.*

By observing **figure 10**, there is a relatively strong, positive relationship between percentage obese and soft drink consumption. Reinforced by a Pearson's correlation of $R = 0.65$, the association between obesity and soft drink consumption shows strong correlation.

These results are significant as it suggests that there is a trend for populations further from the CBD to consume more sugary drinks, resulting in a higher percentage of obesity amongst the outer ring of LGAs.

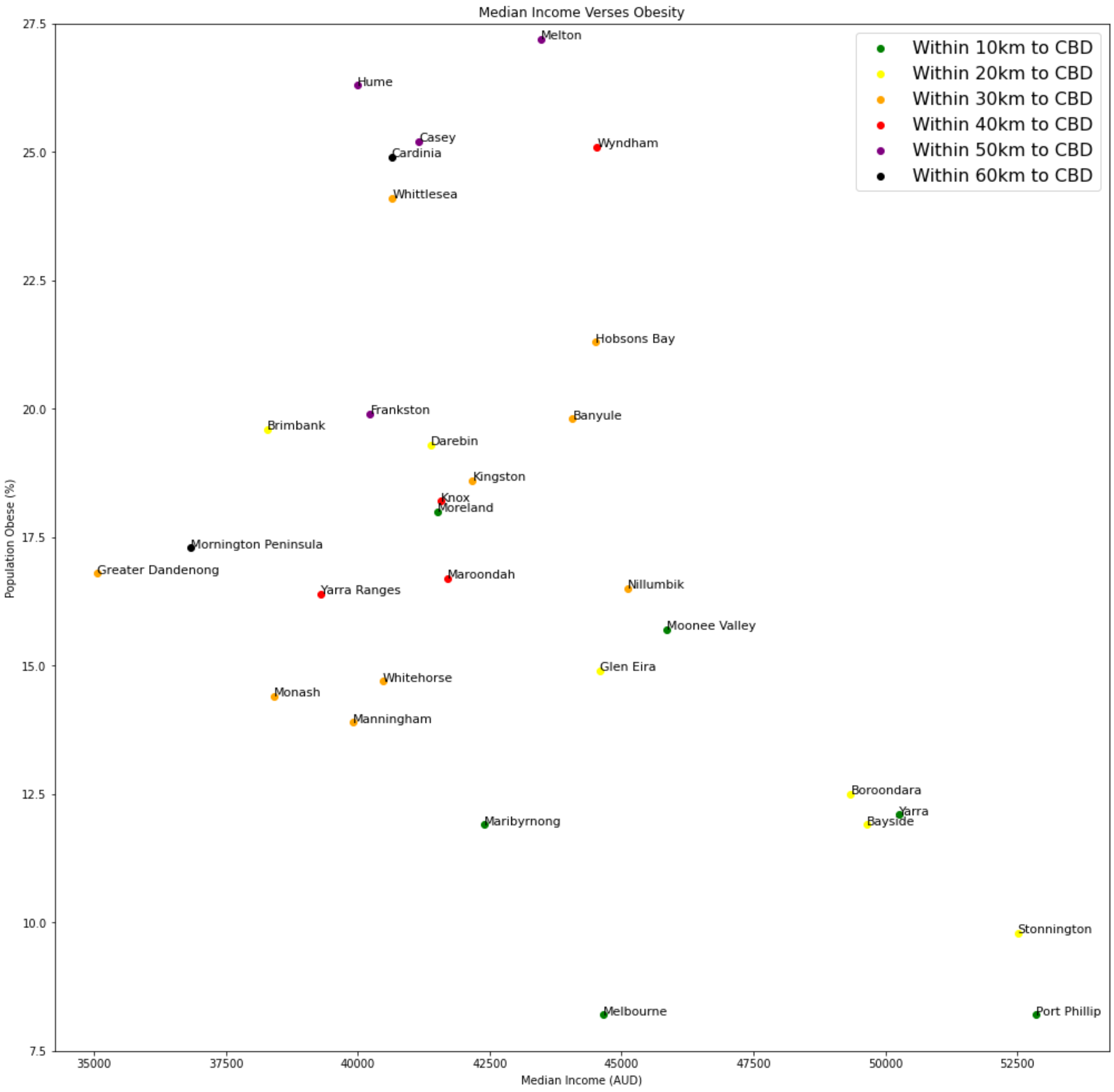

Figure 11. *Scatter plot for median income and obesity rates.*

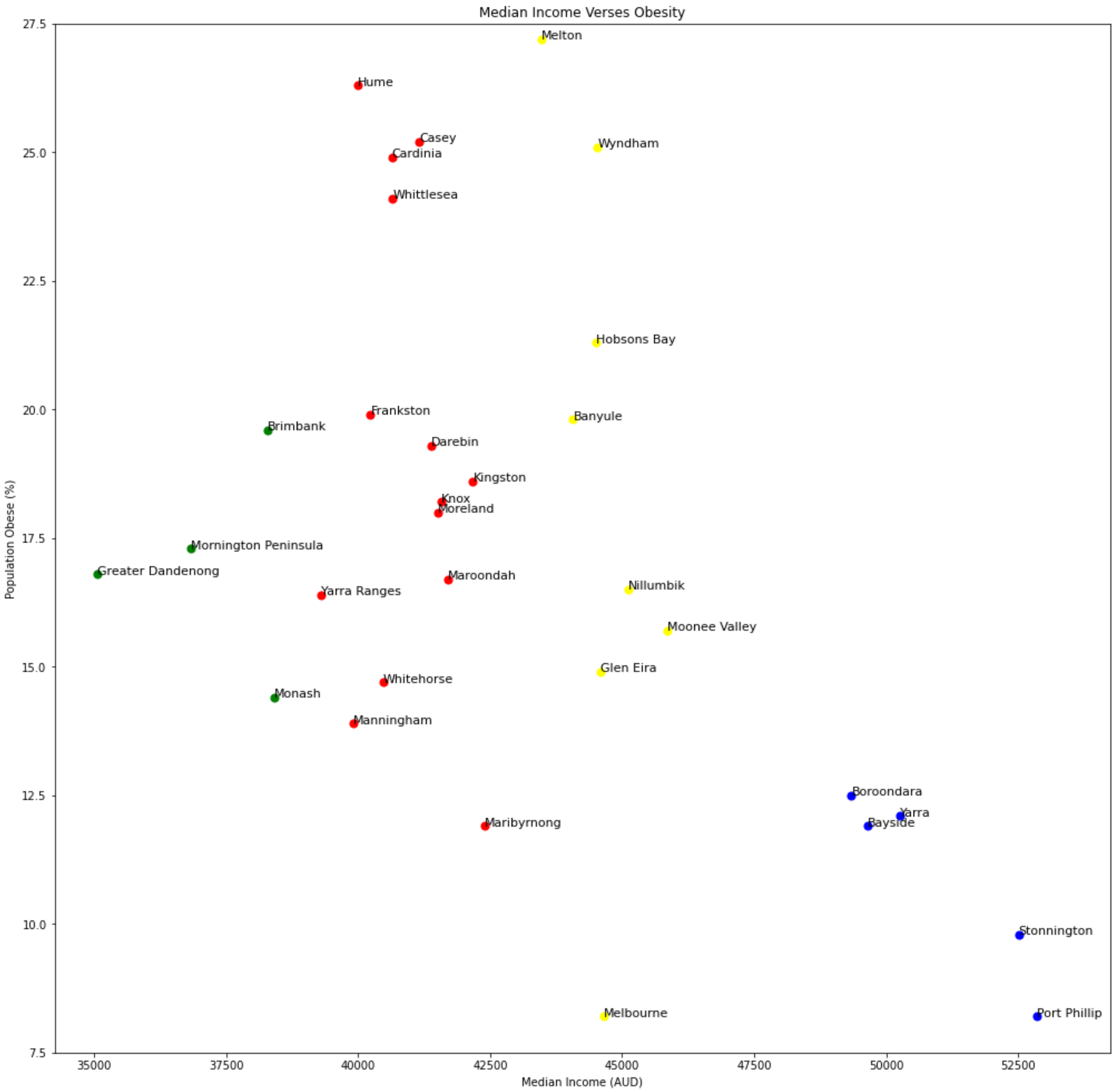

Figure 12. *4-means clustered LGAs for median income and obesity rates.*

By observing **figures 11 and 12**, most LGAs within the 10 and 20 kilometer ring fall within the blue cluster, with the highest median incomes and lowest obesity percentages. With an MI score of 0.29, it suggests that there is minimal correlation between median income and percentage obese. The red and yellow clusters contain LGAs with varying distances to the CBD, indicating that the income range between $40,000 and $45,000 experiences varying obesity rates. Lastly, the green cluster contains LGAs 20 kilometers and beyond the CBD, these data points are more centralised in terms of obesity percentages, but the population has the lowest income.

These results are significant as they suggest that there are higher obesity percentages amongst outer LGAs with median income lower than $47,500. **Figures 3 and 5** show inner municipalities have higher percentages of activity rate and proportion of gyms. Thus, these results could encourage governments to incentivise and subsidise lower-socioeconomic regions, raising awareness and increasing participation in fitness programs.

Limitations and Improvements

Linear regression was implemented for scatter plots that appeared visually linear, however, this assumption is error prone under different conditions of scales and magnitudes. Furthermore, the k-means method performed poorly for some scatter plots with non-spherical shapes, as it generated clusters based on euclidean distance. The outliers from k-means clustering were ignored and labelled as a separate cluster. Moreover, the datasets sourced from trusted organisations were outdated and did not account for the change in obesity rates from population growth overtime.

Future improvements may include utilising non-linear regression models, which circumvents the assumptions for homoscedasticity. Furthermore, the non-linear least squares regression is an alternative that produces accurate estimates of unknown parameters. Fuzzy clustering can be an effective clustering alternative as it mitigates outliers, unlike k-means. Unsupervised learning methods like density clustering can also be used to classify data points based on contiguous regions and allow for more appropriate cluster combinations. A machine learning recommendation system can be used to predict missing or erroneous information through social and environmental attributes that coincide with similar LGAs.

References

Government of the Commonwealth of Australia - Australian Bureau of Statistics. (2018). *LGA Estimates of Personal Income - Income Distribution 2010-2011* [Data set]. AURIN. https://data.aurin.org.au/dataset/au-govt-abs-abs-epi-income-distribution-lga-2010-11-lga2016

Government of the Commonwealth of Australia - Australian Bureau of Statistics. (2020). *ABS - Regional Population (LGA) 2001-2019* [Data set]. AURIN. https://data.aurin.org.au/dataset/au-govt-abs-abs-regional-population-lga-2001-2019-na

Government of Victoria - Department of Health and Human Services. (2017). *Local Government Area (LGA) profiles data 2015 for VIC* [Data set]. AURIN. https://data.aurin.org.au/dataset/vic-govt-dhhs-vic-govt-dhhs-lga-profiles-2015-lga2011

Government of Victoria - Department of Health and Human Services. (2017). *Victoria Sport and Recreation Facility Locations 2015-2016* [Data set]. AURIN. https://data.aurin.org.au/dataset/vic-govt-dhhs-vic-sport-and-recreation-2015-na

Government of Victoria - VicHealth. (2020). *VicHealth - Daily Soft-drink Consumption (LGA) 2011* [Data set]. AURIN. https://data.aurin.org.au/dataset/vic-govt-vichealth-healthy-eating-daily-soft-drink-consumption-lga

Travel Victoria. (2021). *Metropolitan councils.* https://www.travelvictoria.com.au/victoria/metropolitancouncils/#:~:text=31%20municipalities%20cover%20Melbourne%27s%20metropolitan%20area