Group Number: 6

Student IDs, Names & Roles: Sagari Muraliegaran (member) and 4 other members

Contribution Percentage: All members contributed to this work equally.

Big Data and Machine Learning: Assessment 1

1. Executive Summary

This report analyses BMI trends and global obesity utilising a large dataset and machine learning techniques. The dataset includes BMI, obesity prevalence, and socioeconomic factors like years of education and GDP. The dataset was first cleaned to remove missing values and analysed to identify key health patterns and ensure data quality. The exploratory data analysis (EDA) investigated global trends and regional trends, while the scatterplots and boxplots focused on the United Kingdom (UK) assessing the sex-based health disparities. The report key findings:

Sex-based differences: The boxplots revealed males having elevated BMI and systolic blood pressure while females had higher obesity prevalence.

Correlation and Clustering: The heatmaps and clustering techniques reveal strong associations of obesity, urbanisation, BMI, GDP, and blood pressure, with higher-income regions showing higher obesity rates while low-income regions show lower BMI levels which is linked to undernutrition.

Regional Variability: The frequency distribution analysis shows the differences in BMI trends, with higher-income regions having higher BMI and sub-Saharan Africa showing lower BMI due to undernutrition. Indicating greater prevalence of obesity in high-income regions.

The report shows socioeconomic factors, lifestyle, urbanisation, all influence health outcomes, the findings from the report highlight the need for targeted public health interventions to address obesity in developed high-income regions and to fight undernutrition in underdeveloped regions. The analysis showcases the power of big data and machine learning to understand and identify health trends. The data insights can be used by policymakers and public health professionals to create strategies to improve global health conditions.

2. Data Quality

The dataset provides information on obesity and BMI according to various factors such as age, sex, and region. The dataset is stored in a CSV format, and we have carefully examined for any missing values using R. Since the missing values will impact the accuracy of our exploratory data analysis, we will be removing rows containing those values. Since the data is not corrupted, no further actions are necessary.

3. EDA (Exploratory data analysis)

3.1 Summary

Before focusing on the United Kingdom in the dataset, a full analysis was conducted to understand the trends in key health indicators. **Table 1**. showcases the summary statistics for the full dataset,

and **Table 2** shows the UK dataset. The full dataset includes multiple countries, showing more variation in **Table 1** due to the difference in cultural, economic and in the healthcare systems.

Due to the significant variability, the report focuses on the UK data as a subset to identify trends clearer and reduce the complexity.

Statistic	Mean BMI (Adults)	Mean BMI (Children)	Obesity Prevalence (Adults)	Obesity Prevalence (Children)	Systolic Blood Pressure	Years of Education
Global Minimum	17.15	14.59	0.0008601	0.000053	111.7	0.2145
Global 1st Quartile	22.58	17.66	0.0433238	0.008738	123.9	3.6985
Global Median	24.81	18.68	0.1188525	0.028865	127.0	6.6025
Global Mean	24.62	18.62	0.1387562	0.045029	126.8	6.6024
Global 3rd Quartile	26.30	19.51	0.1989284	0.067081	129.7	9.3353
Global Maximum	35.22	24.34	0.6487787	0.324989	137.4	14.5165

Table 1. The summary statistics for key health indicators in the dataset (All countries). The values shown in the table are the distributions for BMI, obesity prevalence, systolic blood pressure, and the years of education. The variables have a large range indicating the variation across the countries in the dataset especially in obesity prevalence and the BMI.

Statistic	Mean BMI (Adults)	Mean BMI (Children)	Obesity Prevalence (Adults)	Obesity Prevalence (Children)	Systolic Blood Pressure	Years of Education
UK Minimum	23.95	18.43	0.09496	0.03066	117.8	10.22
UK 1st Quartile	25.08	19.23	0.14049	0.05258	125.7	11.21
UK Median	26.07	19.86	0.17846	0.07651	127.5	12.04

UK Mean	25.95	19.70	0.18349	0.07333	127.9	11.96
UK 3rd Quartile	26.86	20.17	0.22579	0.09487	131.8	12.77
UK Maximum	27.43	20.52	0.28590	0.10734	133.6	13.36

Table 2. The summary statistics for key health indicators for the United Kingdom. In comparison to the full dataset, the BMI values for both adults and children are higher, this is the same for obesity prevalence in adults, the median and mean are higher compared to the global summary values. The variable years of education was analysed to see the impact on health indicators but the UK dataset shows low variability in education levels, suggesting education is not a key factor for differentiating BMI or obesity prevalence.

3.2 Boxplots

This section presents the boxplots that shows differences by sex in key health indicators, these include BMI, obesity prevalence, and systolic blood pressure. The visualisations show insight into potential differences in health outcomes for both sexes showing distribution and variation helping identify trends in sex-based differences in public health.

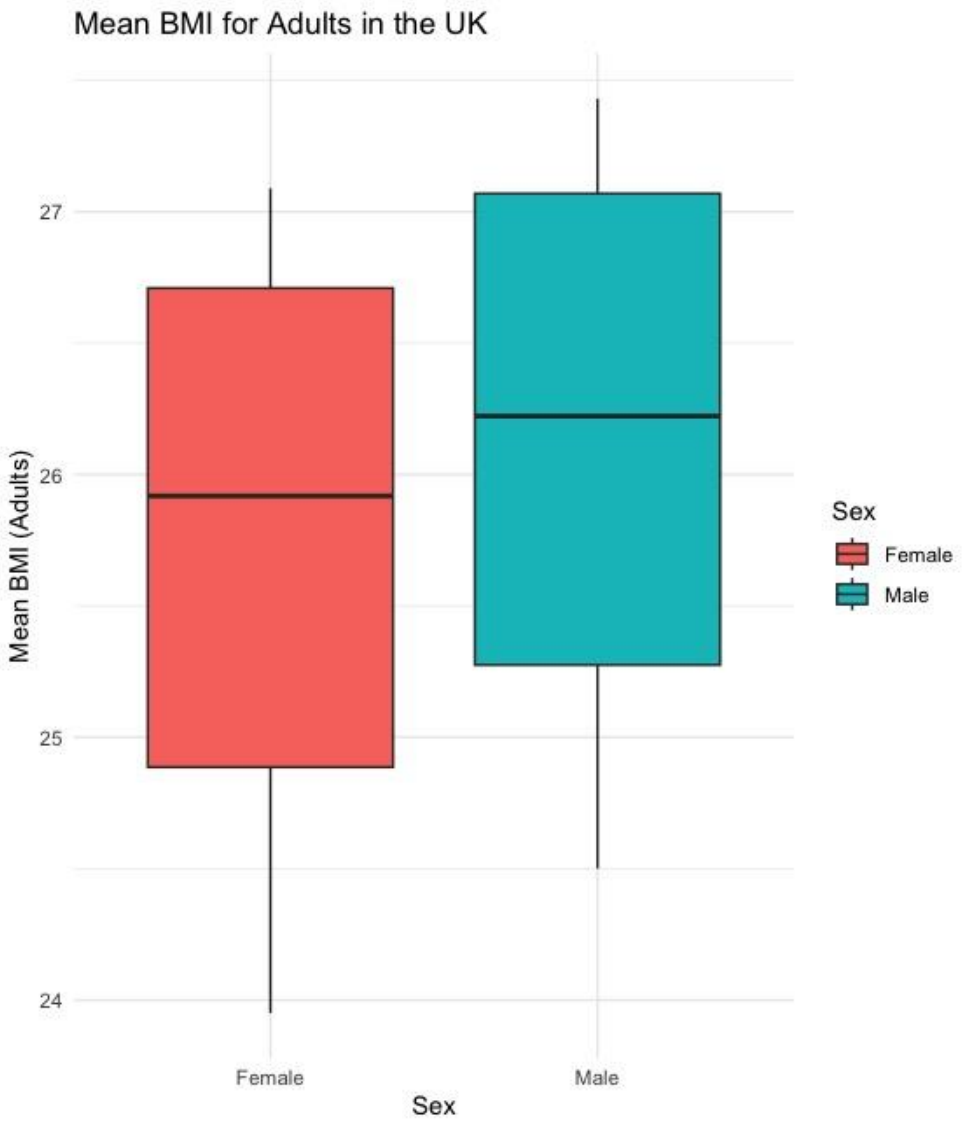

Figure 1. The Mean BMI distribution for adults in the UK by sex. This boxplot shows the distribution of BMI values for females and males in the UK. The median BMI is higher than the females median. The male BMI also has greater variation shown in the boxplot having a wider interquartile range, both distributions fall in the overweight range BMI ≥ 25.

The boxplot in **Figure 1** showcases the mean distribution for mean BMI in UK adults the results shows that males have a higher BMI and more variability in BMI due to the interquartile range. Both groups have median BMI values higher than 25 meaning individuals are overweight according to the World Health Organisation standards. This could show that overweight prevalence is higher in higher income countries The overlap in distribution suggests the BMI differences are not substantial.

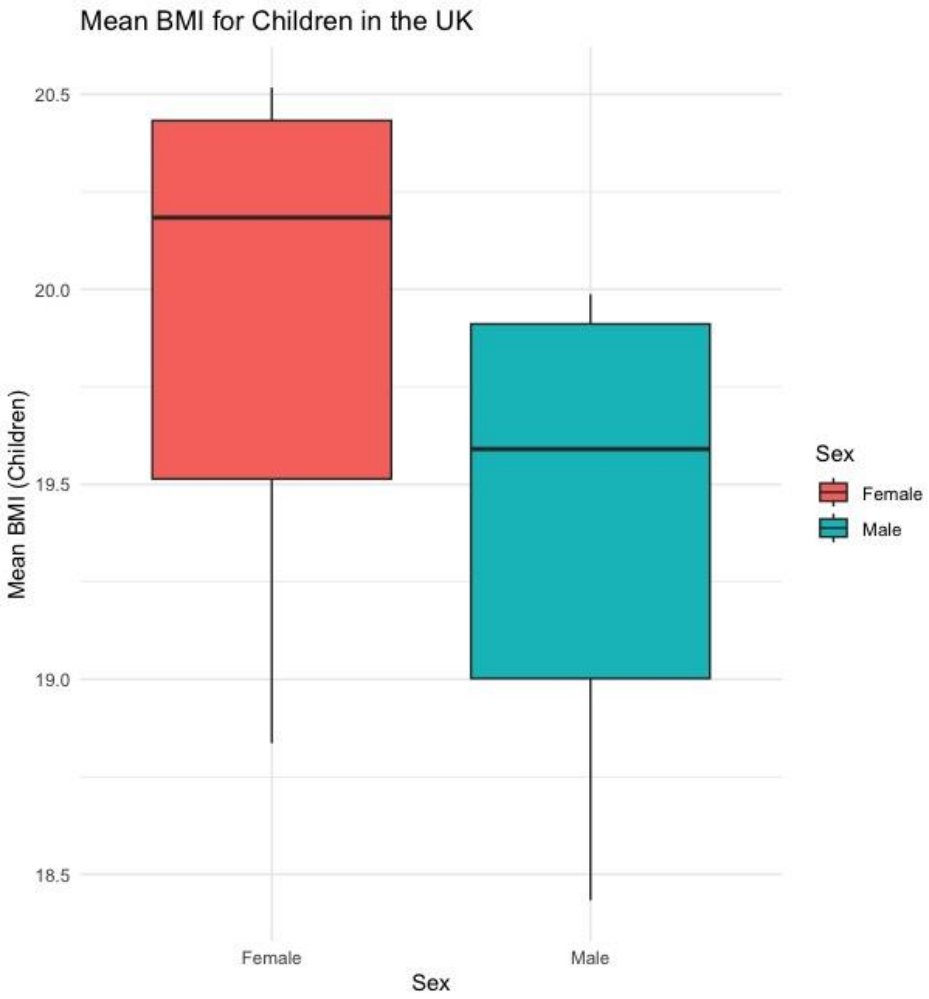

Figure 2. The Mean BMI distribution for children in the UK by sex. The boxplot shows the BMI values distribution in female and male children in the UK. The median BMI for female children is higher than the male children. The female children BMI had a larger interquartile range too.

The boxplot in **Figure 2** shows the difference in the mean BMI for female and male children in the United Kingdom, the median BMI and interquartile range for females is higher compared to the males. This indicates the females having more variability in BMI and having a higher BMI however the overlapping distributions suggest that the BMI difference is not based only on sex but has other factors influencing BMI and differences in BMI are more pronounced in adults as shown in **Figure 1**.

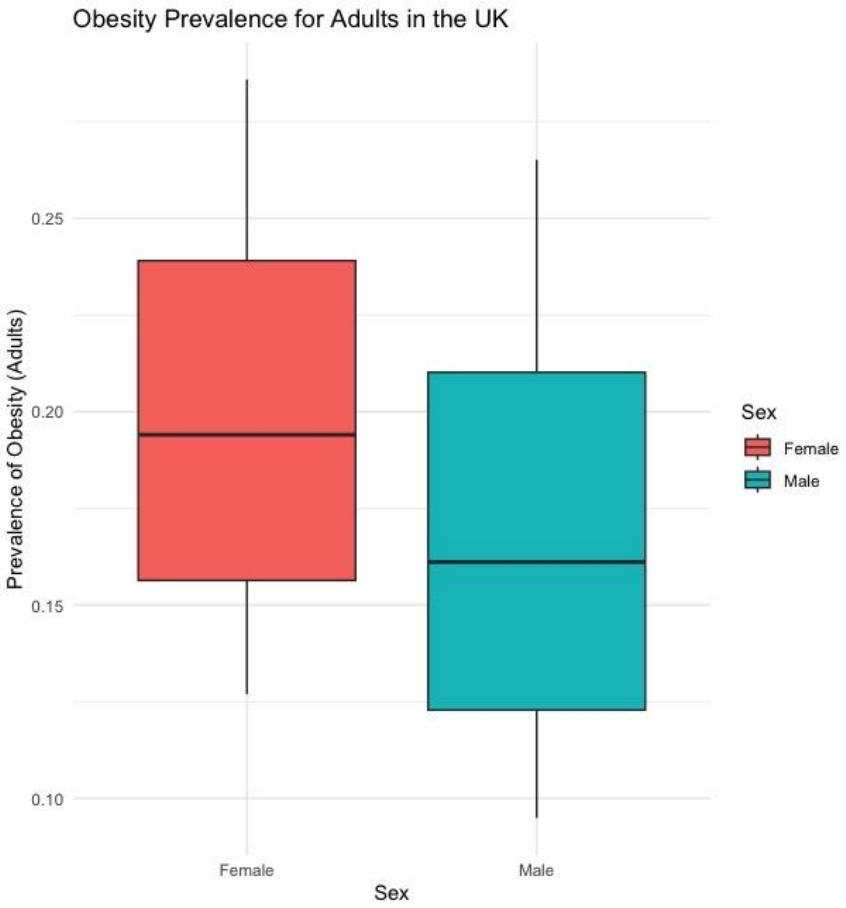

Figure 3. The distribution of Obesity prevalence for adults in the UK by sex. The boxplot shows the distribution of obesity in adult males and females, the median obesity prevalence is higher for females compared to males, both have a similar interquartile range.

The boxplot in **Figure 3** presents the obesity prevalence for adult males and females, the boxplots show females having higher obesity prevalence the IQR is similar and the distributions overlap showing this variable is comparable between females and males. Higher prevalence in females indicated other factors for example metabolic differences or lifestyle affecting obesity rates.

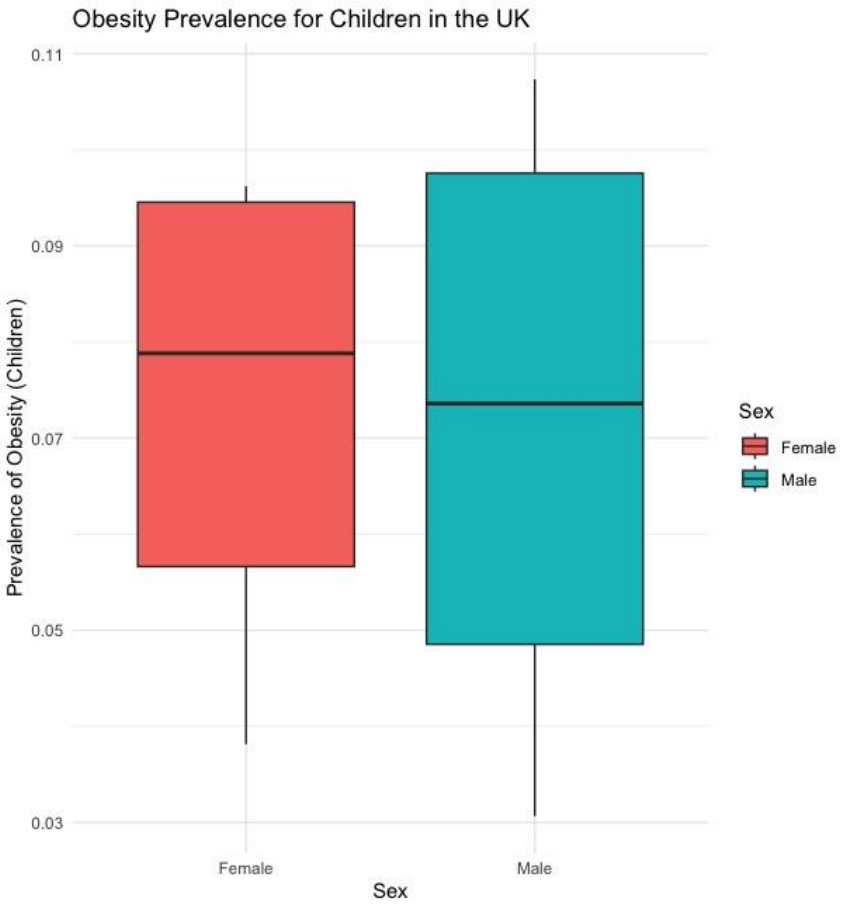

Figure 4. The distribution of Obesity prevalence for children in the UK by sex. The boxplot shows the distribution of obesity prevalence in female and male children. The median is higher for female, and the interquartile range (IQR) and distribution of the values are similar for both.

The boxplot in **Figure 4** shows the distribution of obesity prevalence in male and female children in the UK. The females have a higher median but the overall distribution of the data and IQR are similar and overlap suggesting small sex-based difference in childhood obesity prevalence.

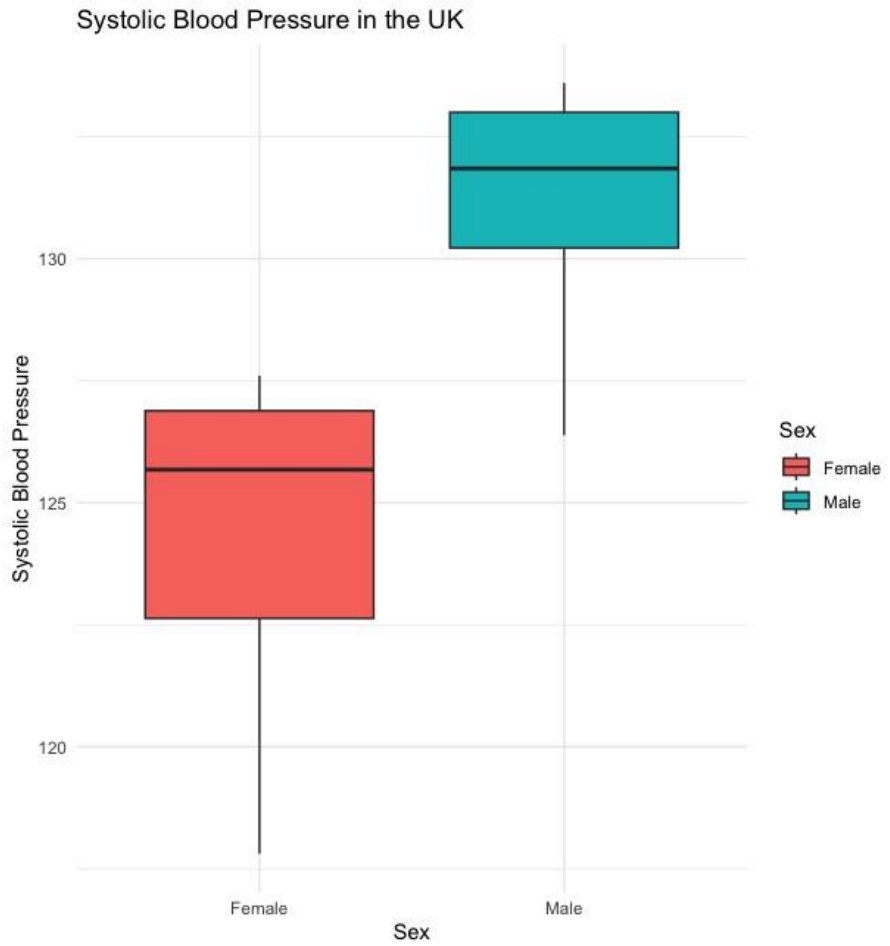

Figure 5. The distribution of systolic blood pressure in the UK by sex. The boxplot shows the differences in males and females systolic blood pressure, the male data has a higher median blood pressure.

The boxplot in **Figure 5** shows systolic blood pressure distribution in males and females in the UK. The results indicate higher blood pressure distribution in males, the IQR for males is also higher indicating high variation in systolic blood pressure compared to females. The elevated blood pressure in males could be due to many other factors.

The dataset included socioeconomic factors such as years in education and GDP however these variables were not included due to the lack of relevance to health outcomes like BMI. Furthermore, the boxplots for both GDP and education in the UK showed minimal variation between males and females shows these factors do not strongly affect health-related variables, so these boxplots were excluded from the report.

3.3. Frequency Distribution of Mean BMI for children by Super Region

This section shows the frequency distribution of mean BMI for children and adults to analyse global variations in BMI across different regions and identify patterns related to nutrition and health disparities.

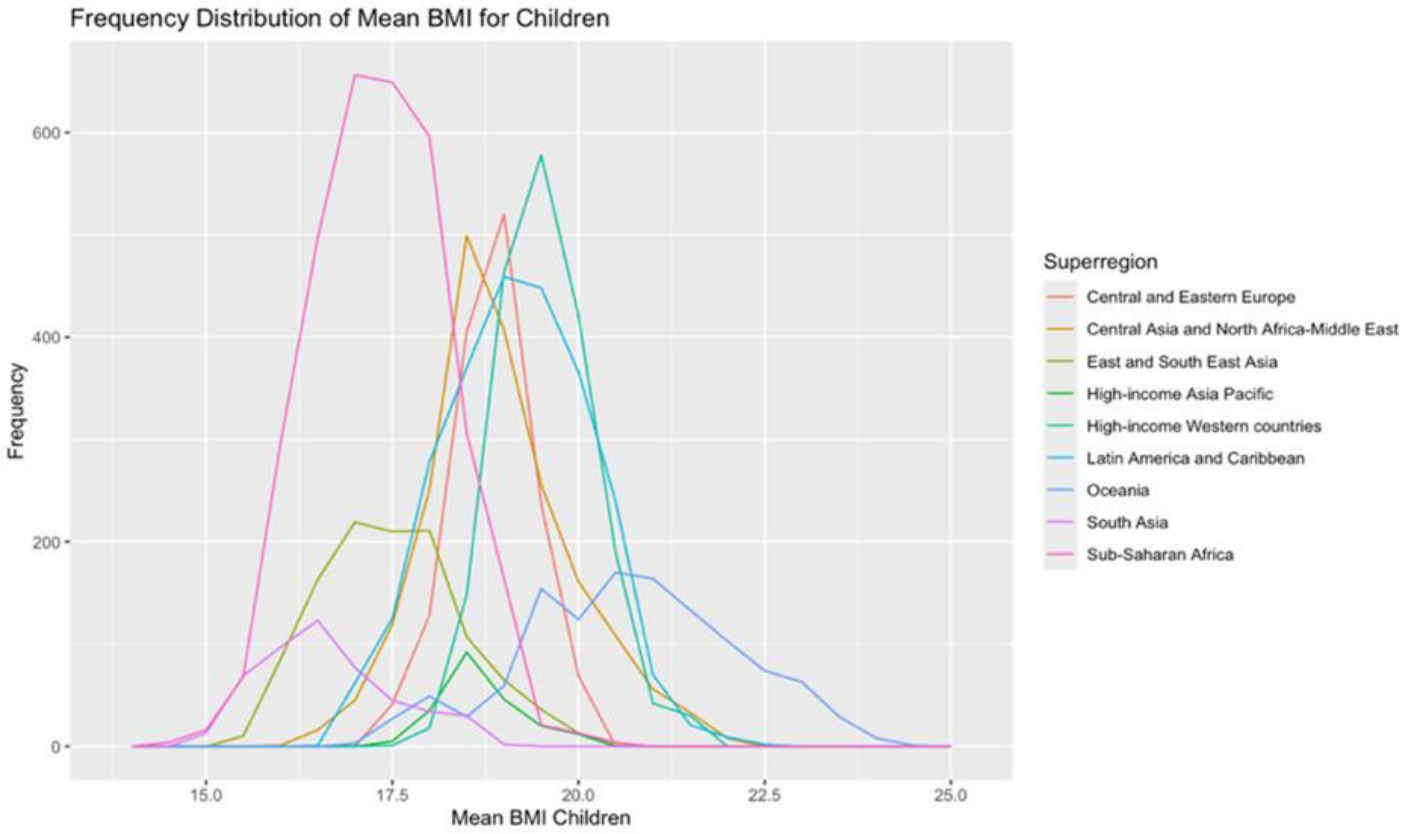

Figure 6. shows the frequency distribution of Mean BMI for Children across Superregions

Figure 6 shows a frequency distribution plot of the mean Body Mass Index (BMI) for children from different super regions. Children's BMI values vary from 14 to 25, with a distribution skewed towards lower BMI values, with most frequencies peaking between 16 and 20. Sub-Saharan Africa (pink line) has the largest peak, indicating that many children in this region have low BMI values, most likely due to undernutrition. In contrast, Latin America, the Caribbean and High-Income Western countries have a larger distribution with higher BMI values, which might indicate greater nutrition access or a higher prevalence of childhood obesity. Other regions, such as South Asia, East and South-East Aisa and Central Asia, have more compact distributions, with BMI values peaking at 18-19, indicating regional differences in nutritional status and dietary habits.

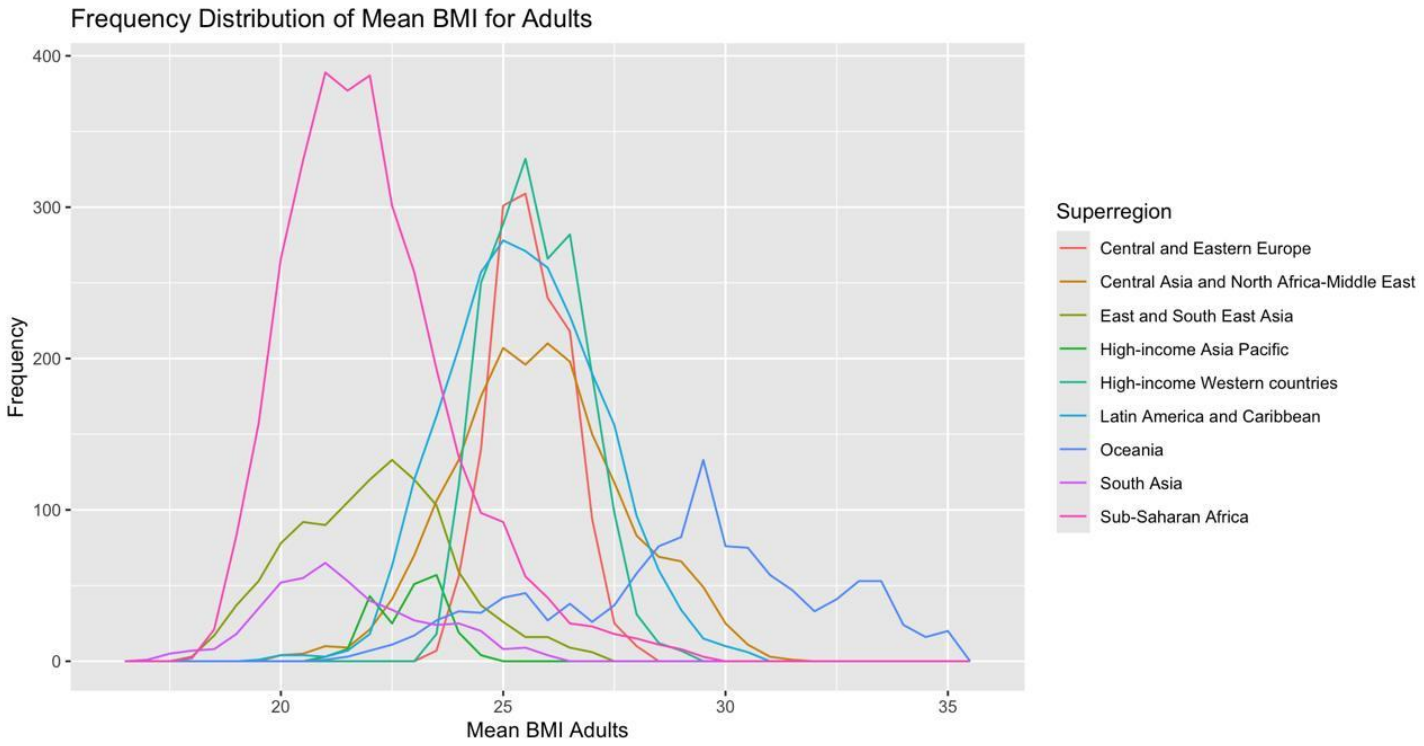

Figure 7. The frequency distribution of Mean BMI for Adults across Super regions.

Figure 7 represents the frequency distribution of Mean BMI for Adults across different super regions. In comparison to the children BMI mean (**Figure 6**), the BMI range for the adults is much higher, ranging from 18 to 35, and their distribution is more normal-shaped, with peaks between 23 and 27. Sub-Saharan Africa (pink line) shows a peak at lower BMI values, but the frequency is lower than in children indicating adults in this region have a little higher BMI than children. The peaks in High-Income Western countries, Latin America and the Caribbean have peaks shifted towards the right, showing that these regions have a higher average BMI among adults. Furthermore, the BMI distribution among adults is larger, with certain locations showing a considerable tail towards higher BMI values, which might represent that obesity is increasing in specific places.

3.4 Correlation heatmaps and scatter plots

This section includes heatmaps and scatterplots that highlight the correlation of various health indicators, demographic factors and economic variables such as BMI, obesity prevalence, urbanisation, GDP, blood pressure and education levels, shedding light on the links between lifestyle, socioeconomic status and public health outcomes.

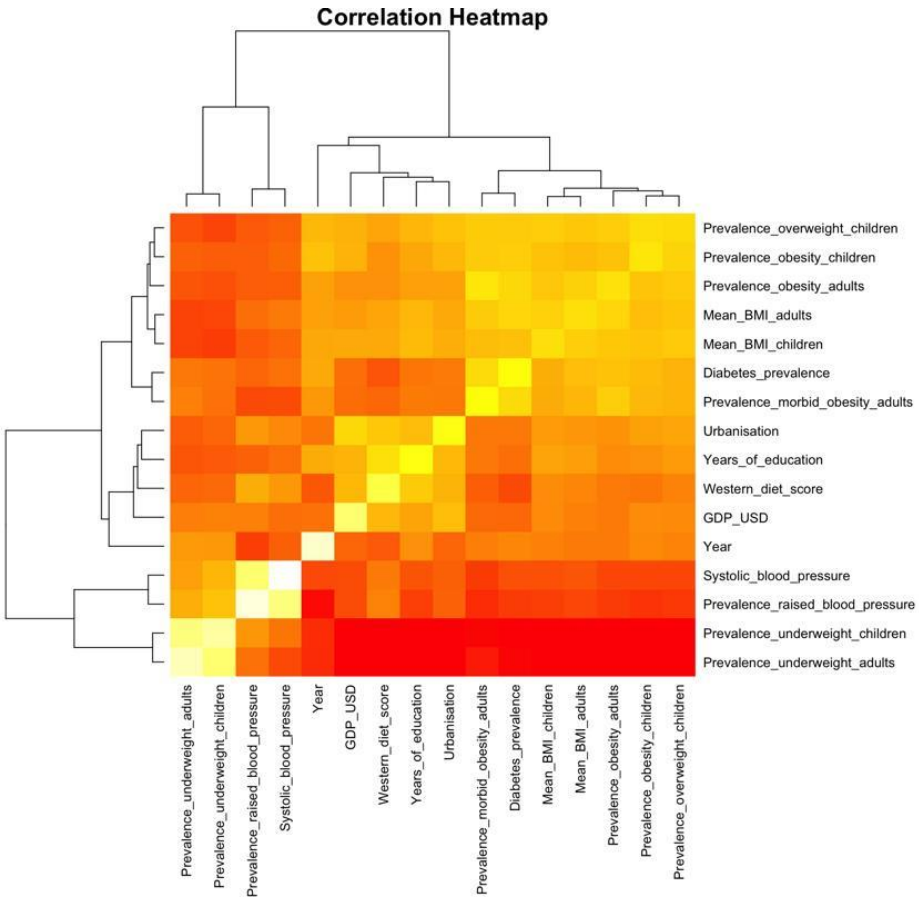

Figure 8. This figure shows a Correlation Heatmap of Health, Demographic and Economic Factors.

This heatmap in **Figure 8** shows the relationship between various health, demographic and economic factors that influence BMI, obesity and public health measures. Red regions indicate a strong correlation whereas the yellow regions indicate a weak correlation. Several relationships appear on the heatmap. Mean BMI in adults has a strong correlation with obesity prevalence, whereas the prevalence of obesity in children corresponds with overweight prevalence in children, indicating that early-life weight habits frequently remain into adulthood. Furthermore, urbanisation and GDP have a substantial relationship with higher BMI levels, indicating the influence of contemporary lifestyles and food habits in more developed countries. A scatterplot will provide a more detailed analysis of the main factors.

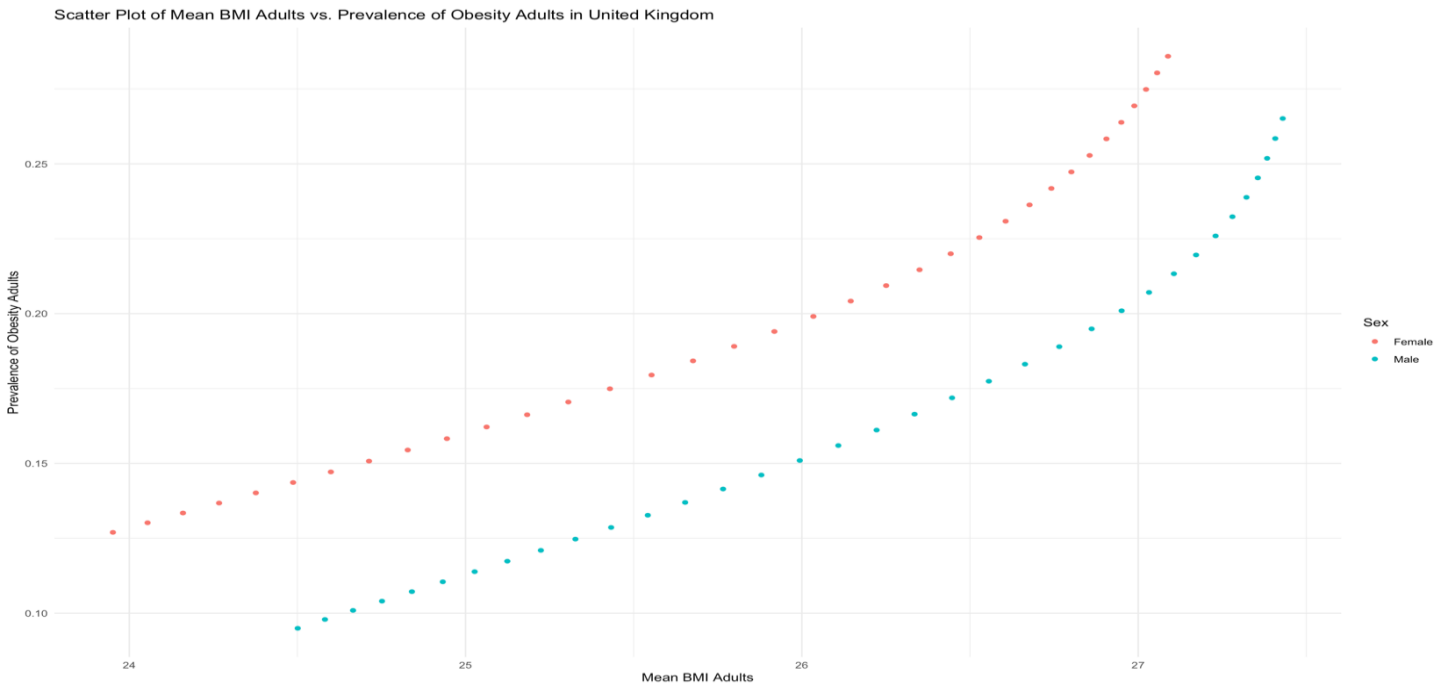

Figure 9. This figure shows a scatterplot highlighting the relationship between Mean BMI in Adults and the Prevalence of Obesity in Adults in the United Kingdom.

Figure 9 highlights that obesity prevalence is increasing as BMI increases. It has a non-linear trend that accelerates at higher BMI levels. There are two lines which emphasise the males and females' comparisons. The plot shows that obesity prevalence starts around 10-15% and the BMI between 24 and 25, where it rapidly climbs to over 25%, which highlights the public health implications of BMI levels. It is also evident that females are experiencing it more compared to males as females have a high mean BMI and prevalence of obesity.

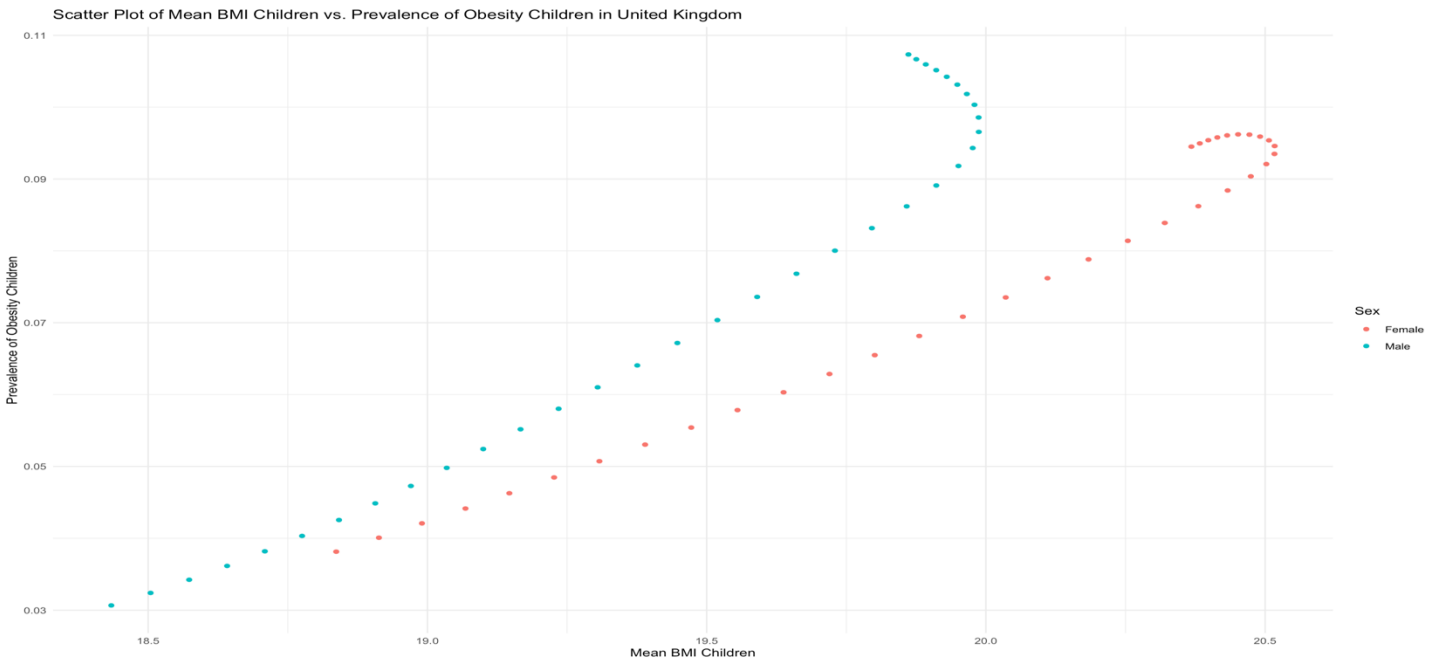

Figure 10. This figure shows a scatterplot highlighting the relationship between Mean BMI in Children and the Prevalence of Obesity in Children in the United Kingdom.

The plot in **Figure 10** has a positive correlation between Mean BMI and Obesity Prevalence, which indicates that as children's mean BMI grows, so does the proportion of obesity. The non-linear trend shows that obesity prevalence increases with increasing BMI values, the two distinct bands in the plot shows the difference between males and females. Obesity prevalence ranges from roughly 3% when the mean BMI is 18.5 to more than 10% when the BMI is 20.5. This trend shows that having a high BMI in children has a strong link to obesity prevalence, with rates reaching the threshold. These findings emphasise the need for early interventions to tackle childhood obesity in the UK.

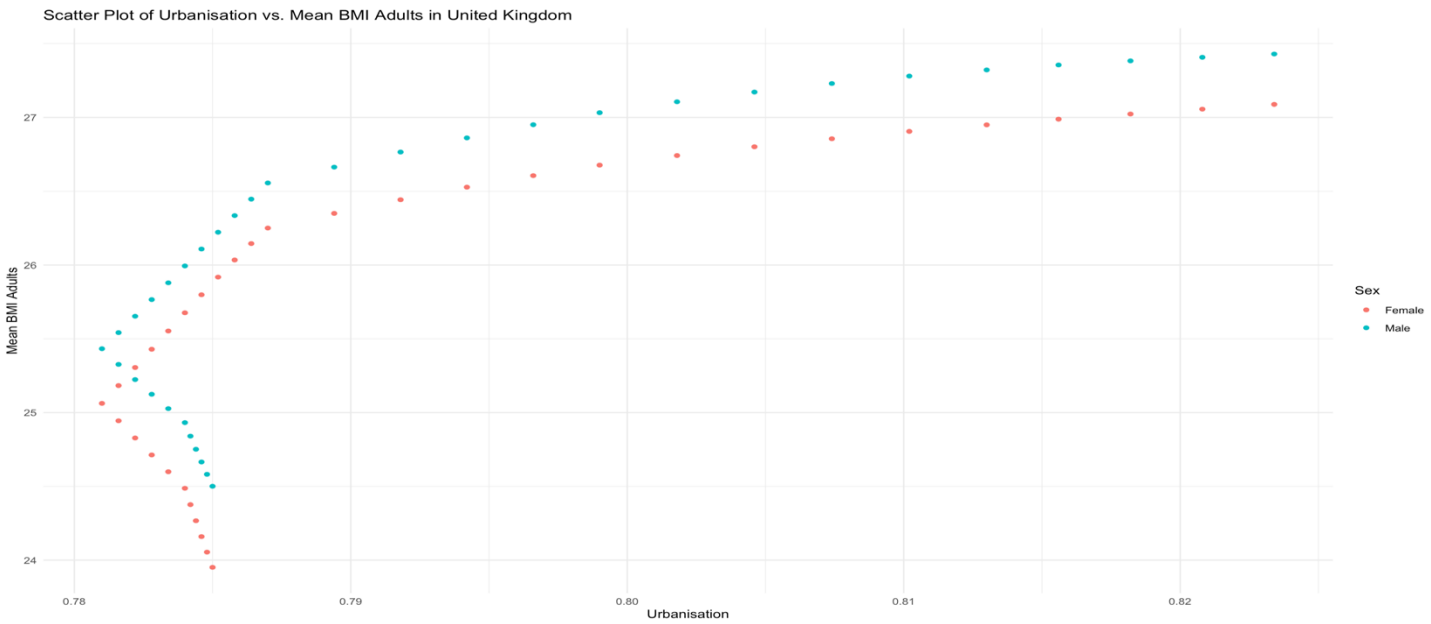

Figure 11. This figure shows a scatterplot highlighting the relationship between Urbanisation and Mean BMI in Adults in the UK (Plot 3).

The plot in **Figure 11** shows both a positive and negative correlation. The negative correlations show that as BMI decreases, urbanisation increases at lower urbanisation levels (around 0.78 – 0.79). There is also a positive correlation, whereas the urbanisation increases and so does BMI. This plot only looks at the UK populations so a negative trend may be due to lifestyle differences in rural and urban areas, where some individuals may be more physically active and may have different dietary habits. Additionally, economic and socio-demographic factors may contribute, as less urbanised regions could have different BMI trends due to healthcare access and food availability. As this plot shows a non-linear relationship it suggests that BMI levels shift depending on urbanisation.

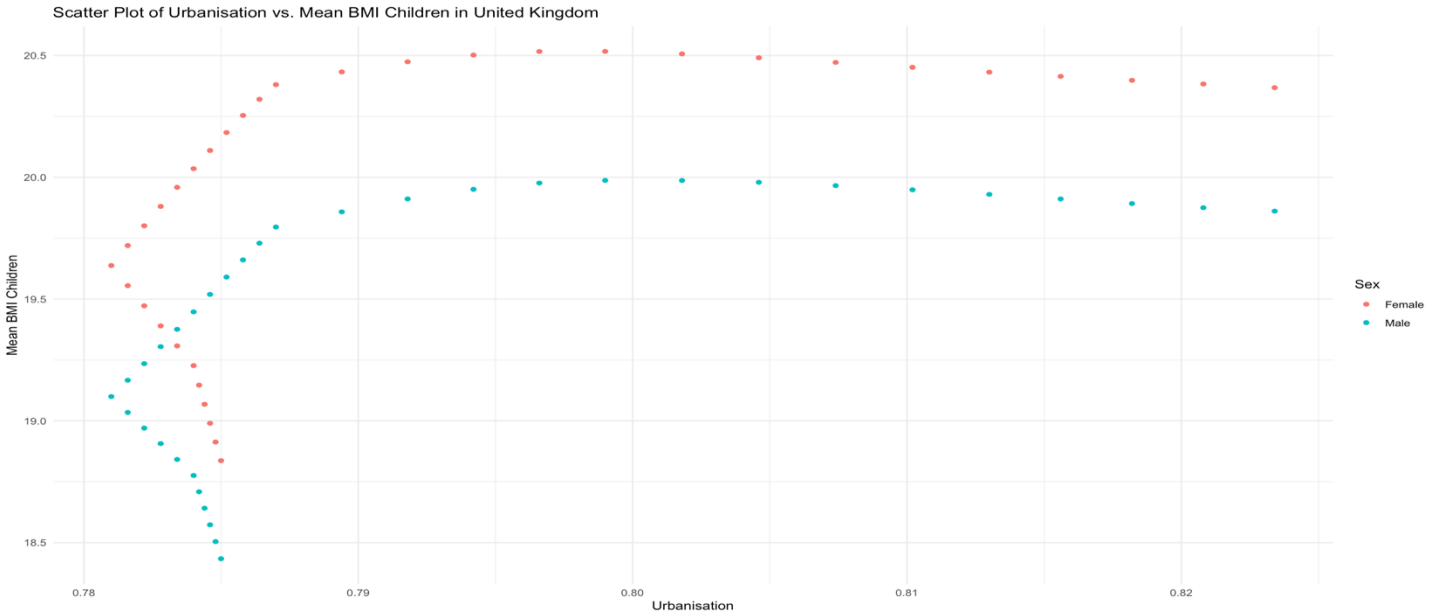

Figure 12. This figure shows a scatterplot highlighting the relationship between Urbanisation and Mean BMI in Children (Plot 4).

The plot in **Figure 12** shows that at low urbanisation levels (0.78 – 0.79), there is a negative correlation, whereas BMI decreases as urbanisation increases. The positive correlation is where BMI increases with urbanisation when it reaches a certain threshold. These two correlations suggest that the relationship between urbanisation and BMI is not linear. This is impacted by factors like lifestyle, dietary habits and socioeconomic conditions. As this plot is only looking at the UK population, the negative correlation may be due to the type of area, where children are having a healthier lifestyle and having different eating habits, which leads to a low BMI. However, as there is an increase in urbanisation in the positive correlation, this would be the opposite of the negative correlation. It is very evident that females are affected more compared to males. This plot emphasises the complex relationship between urbanisation and mean BMI in children, with lower urbanisation initially corresponding with lower BMI, however, if it is above a certain threshold, BMI rises with urbanisation.

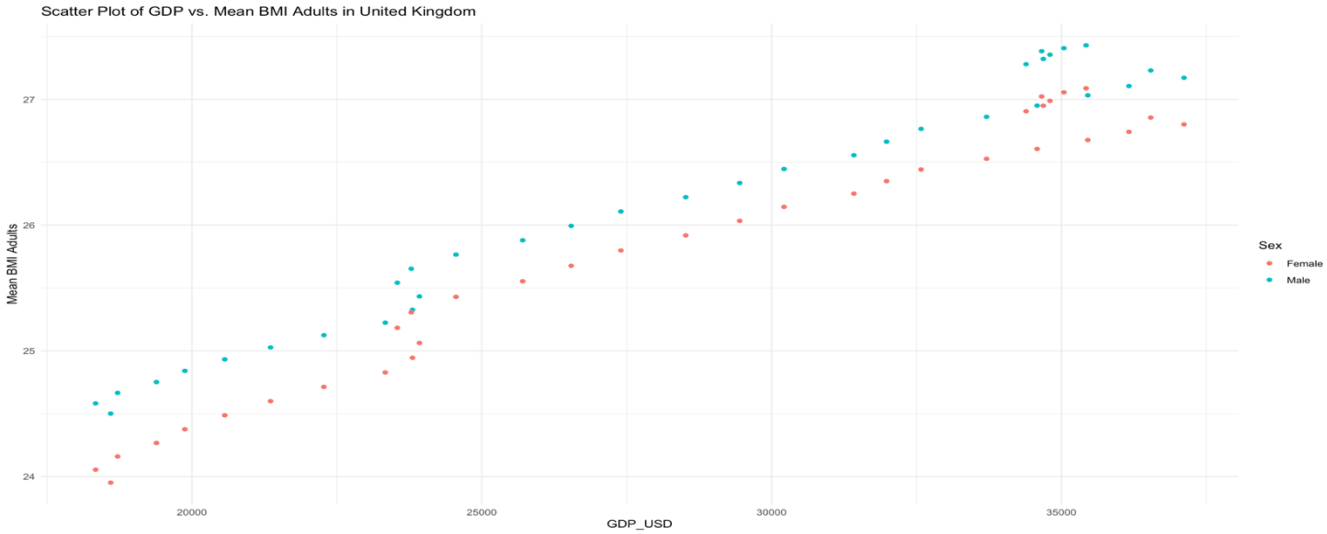

Figure 13. This figure shows the relationship between GDP (in USD) and Mean BMI in Adults in the UK (Plot 5).

This plot in **Figure 13** shows a clear positive correlation. The lower GDP levels (around 18,000 to 22,000 USD) showcase the relationship with a lower BMI between 24 to 25. However, when the GDP increases so does the BMI. This means that high economic development is linked to high BMI levels and varies between the males and females. As this plot is only looking at the UK Adult population, the correlation is probably due to lifestyle and food habits which are linked to economic growth and the males are experiencing it more than females. This highlights the need for public health strategies that include the influence of economic growth on eating habits and physical activity.

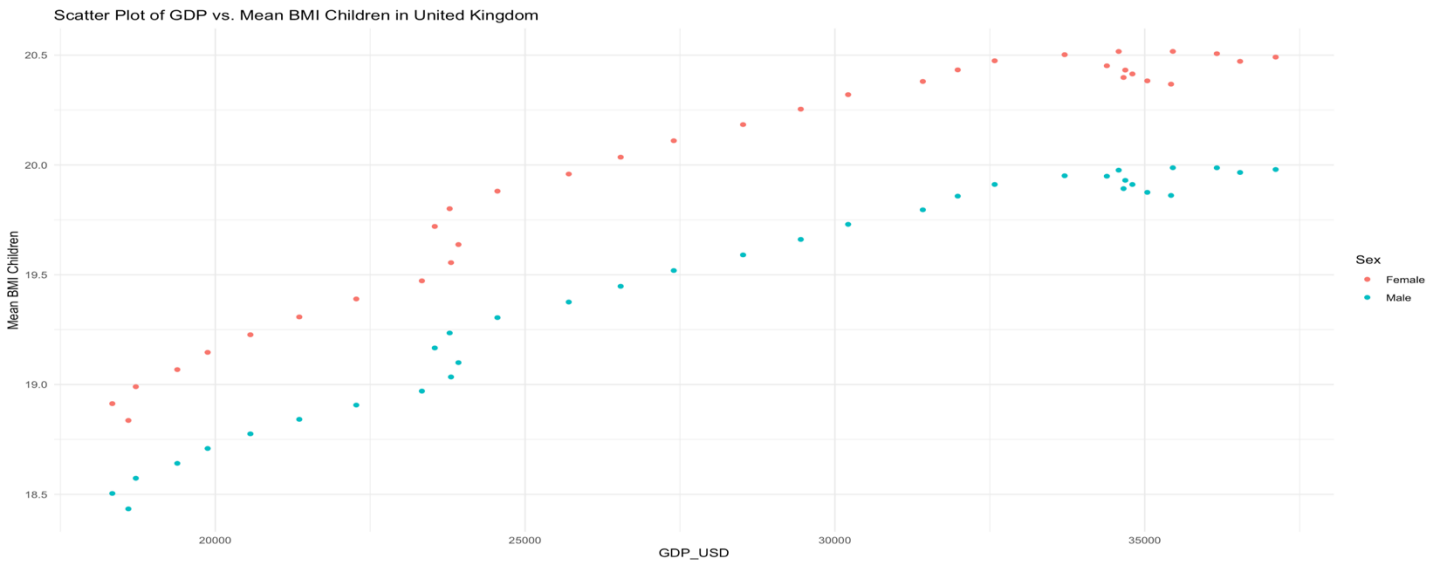

Figure 14. This figure shows the relationship between GDP (in USD) and Mean BMI in Children in the UK (Plot 6).

The plot in **Figure 14** shows a positive correlation which indicates that GDP increases, and the mean BMI of children increases too which varies depending on if they are male or female. This highlights that higher economic growth is a big factor due to BMI levels increasing and this is also in relation to lifestyle and dietary habits.

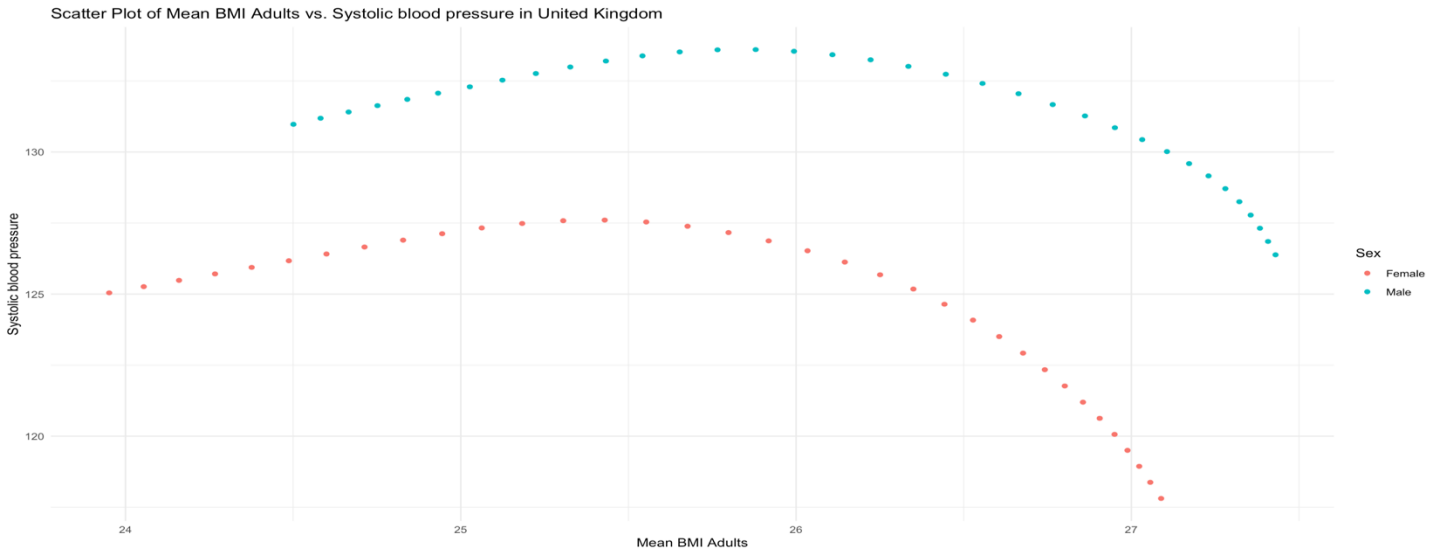

Figure 15. A scatterplot which shows the relationship between Mean BMI in Adults and Systolic Blood Pressure in the UK (Plot 7).

This plot in **Figure 15** shows a non-linear trend between systolic blood pressure and mean BMI in Adults. It shows that systolic blood pressure increases with BMI, peaking at 26-27, however, after the peak it decreases as the BMI level increases. There are two lines expressing males and females,

which allows a comparison between the two genders. Males have consistently higher systolic blood pressure than females at all BMI levels, with a difference of about 5mmHg at equal BMI values. While blood pressure increases with BMI due to fat buildup and circulatory strain, the decline at extreme BMI levels may be caused by increased medication usage, lifestyle changes or physiological responses. There are also physical factors like males having a larger body size, higher muscle mass and hormonal variation.

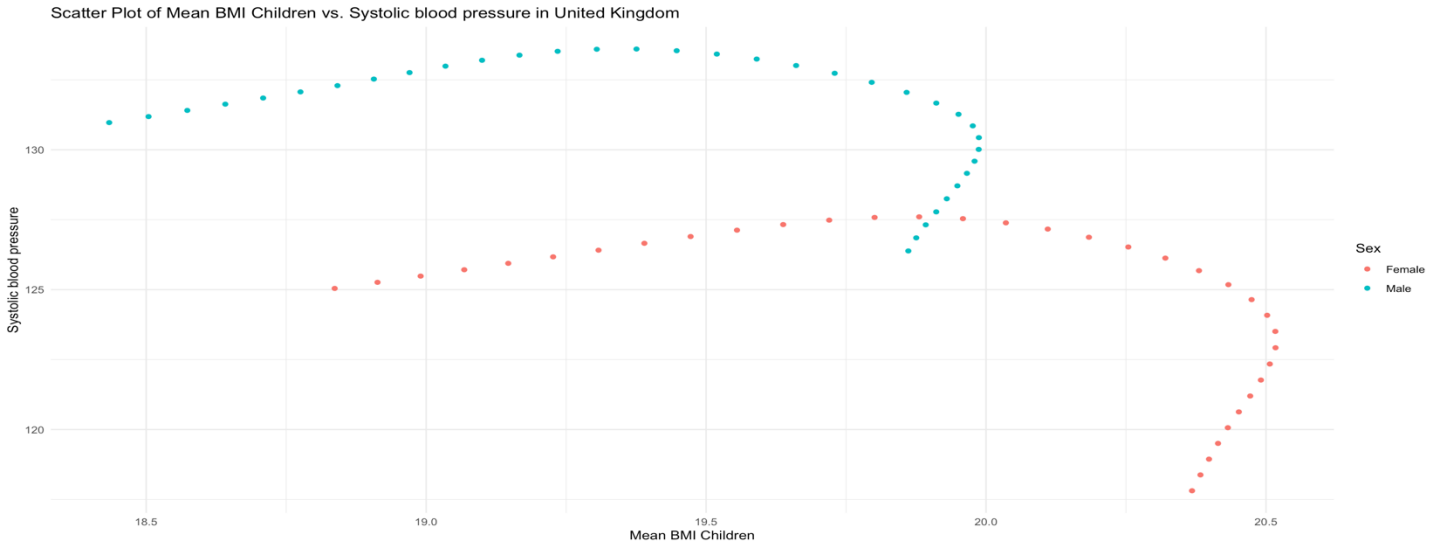

Figure 16. A scatterplot showing the relationship between Mean BMI in Children and Systolic Blood Pressure in the UK (plot 8).

This plot in **Figure 16** shows a non-linear trend, where systolic blood pressure initially increases with BMI, and peaks at a certain BMI range, after peaking it decreases as the BMI value gets higher. The two lines which represent males and females can be compared. Males have higher systolic pressure than females across all BMI levels. As seen for both genders, systolic blood pressure rises as BMI increases, this suggests that higher BMI in children is associated with elevated blood pressure. However, at higher BMI levels (over 20.0 for males and 20.5 for females), blood pressure decreases and generates a curve. The drop may be due to medical treatments, lifestyle changes or physiological adjustments. The observed patterns can be caused by physiological differences since makes have greater blood pressure due to increased muscle mass, larger heart size and hormonal impacts. Furthermore, obesity has a clear influence on blood pressure since extra fat storage might increase cardiovascular strain. The decrease in blood pressure at higher BMI levels might be due to hypertension therapy and lifestyle changes.

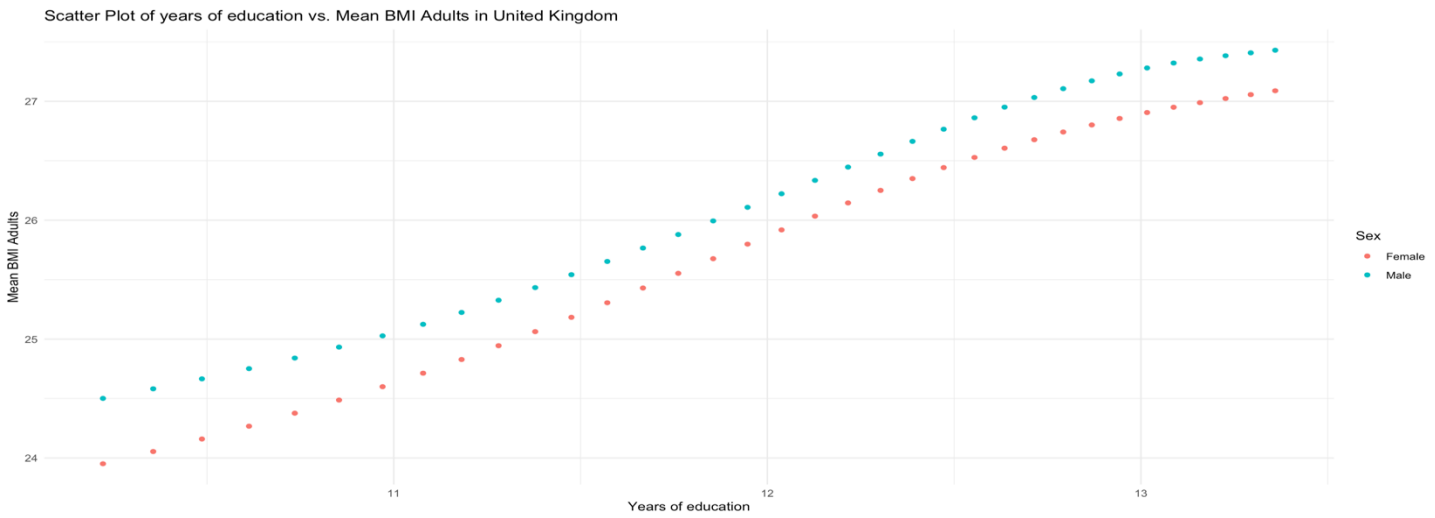

Figure 17. A scatterplot highlighting the relationship between Years in education vs Mean BMI for Adults in the UK, distinguishing between males and females.

The plot shows a positive correlation as mean BMI increases as years of education increase. There are two lines where one represents males and the other represents females. The males have a higher BMI than women across all education levels. For individuals with fewer years of education (10-11 years), their mean BMI is lower compared to an individual who has been in education for more than 12 years, this is also may be affected by who is financially stable and that can affect why they buy for food and lifestyle. Furthermore, men have a higher muscle mass and metabolic rate than women which explains why there is BMI variation.

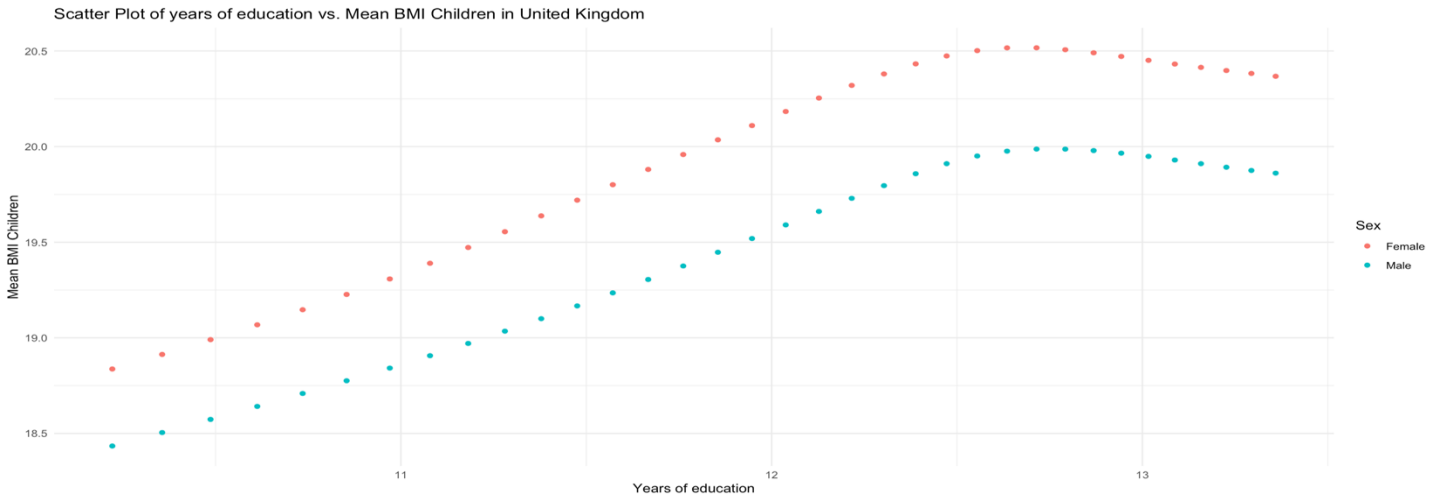

Figure 18. highlights the relationship between Years in education vs Mean BMI for Children in the UK

This plot has a positive correlation, where mean BMI increases as years of education increase. There are two lines which represent the males and females. The females have a higher BMI than the males across all education levels as a child. As there is a gap between the two lines this may be due to

growth and metabolic differences as males have a higher muscle mass which in comparison is higher than females. Additionally, dietary habits and activity also play a role.

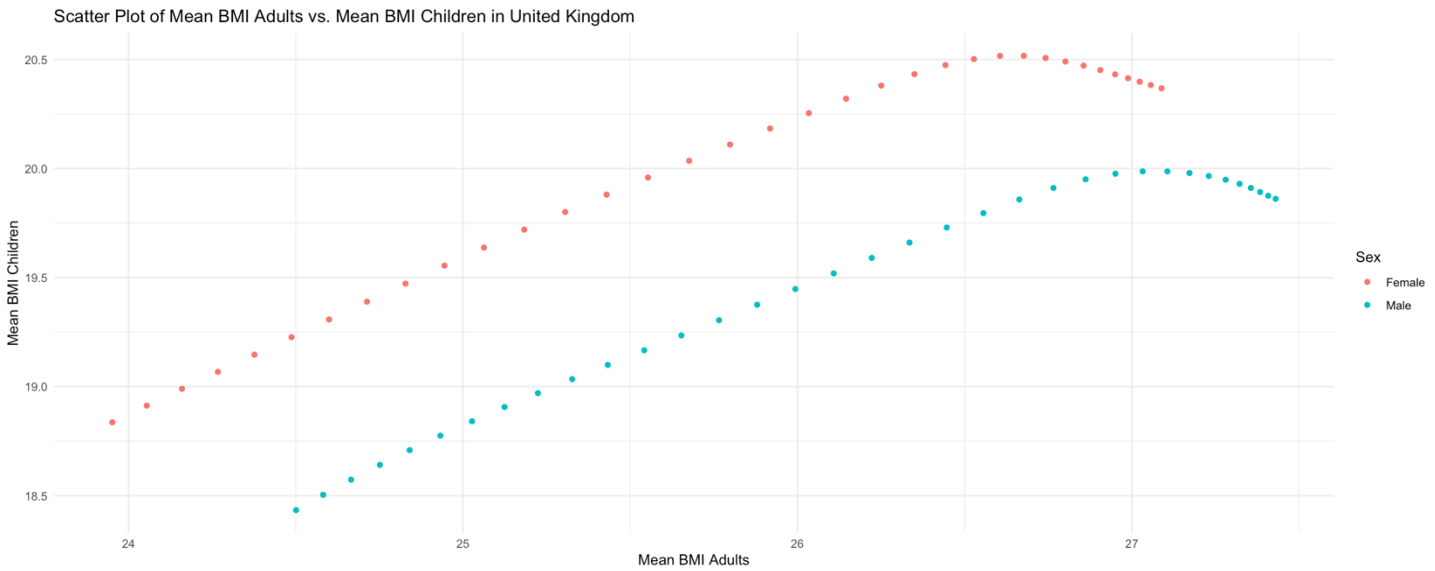

Figure 19. highlights the relationship between Mean BMI for Adults and Mean BMI for Children. This scatterplot shows a positive correlation between adult BMI and children BMI, which suggests that as adult BMI increases, child BMI also increases. This is a non-linear relationship, as can be seen on the plot that it increases at a certain point before declining. Where this can be due to lifestyle and dietary habits. Females shows a higher BMI values compared to males at the same BMI level, which could mean that growth distribution varies by sex. Dietary habits and physical activity could also contribute to the pattern we observed. Higher BMI in adults might correlate to habits that affect children.

4. Clustering Analysis and Distance Matrices

This section of the report utilises clustering techniques and distance matrices to analyse global patterns regarding health – related and socioeconomic/demographic metrics. Clustering allows for the grouping of regions with similar health trends as well as highlighting those with stark differences. Furthermore, the distance matrices provide a visual representation of similarities and dissimilarities between different regions based on health – related and socioeconomic/demographic metrics. It further helps our understanding of regional disparities in relation to those same metrics.

4.1 Super Regions

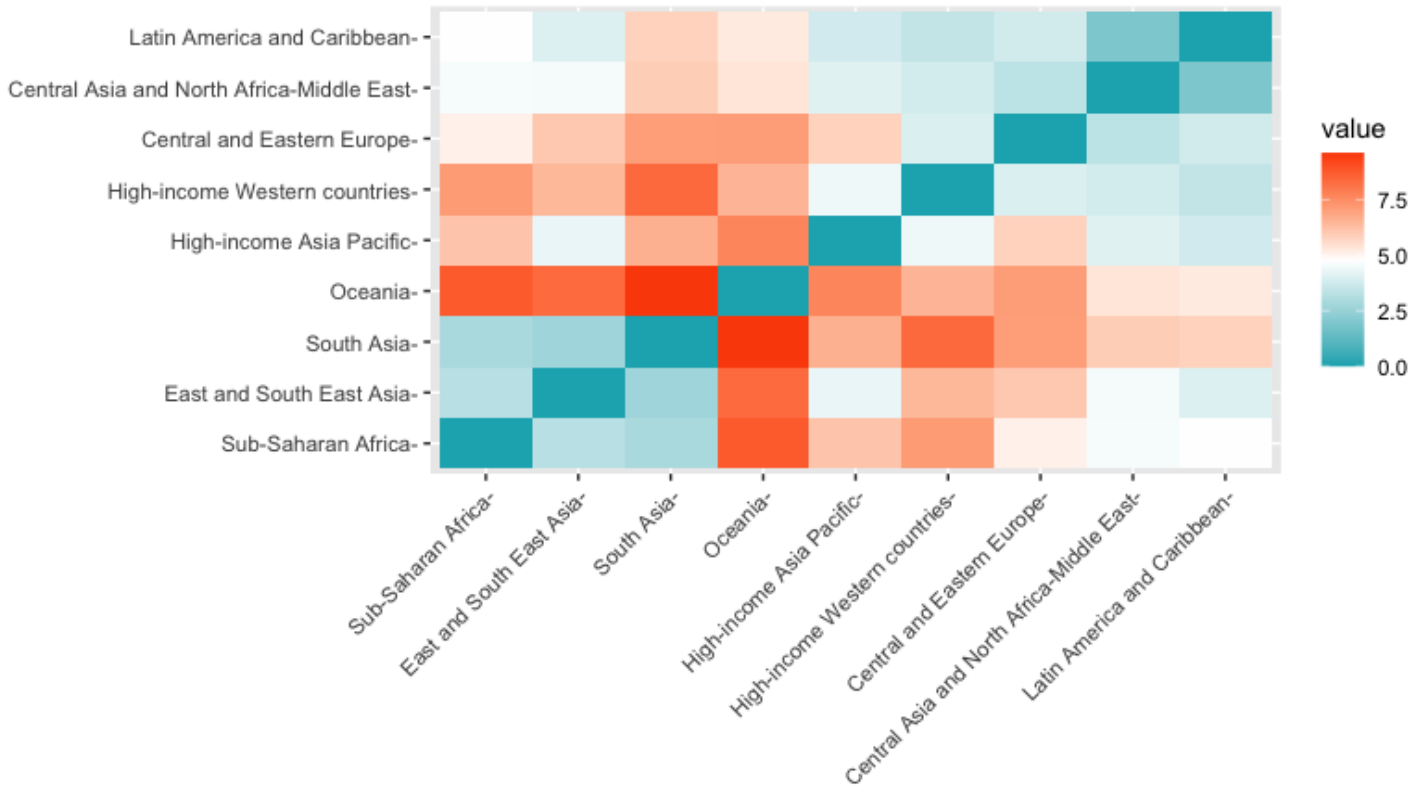

Figure 20. A distance matrix heatmap on the super region dataset. This distance matrix heatmap based on scaled data shows East and South – East Asia and Sub – Saharan Africa have a high similarity regarding the health – related and socioeconomic/demographic metrics as well as South Asia and Sub – Saharan Africa. It also shows a high similarity between Central Asia & North Africa - Middle East and Central and Eastern Europe. Furthermore, it shows High– Income Asia & Oceania, South Asia & High – Income Asia Pacific and Sub – Saharan Africa & High – Income Western Countries show strong dissimilarity. Overall, the distance matrix heatmap illustrates high – income regions (Western countries, Asia Pacific, and Oceania) tend to be more dissimilar from lower – income regions (Sub – Saharan Africa, South Asia). This reflects higher health – related and socioeconomic/demographic metric rates in wealthier regions due to lifestyle differences. It also illustrates geographically close regions (e.g., East & Southeast Asia vs. South Asia) show similarities, suggesting shared regional lifestyle habits.

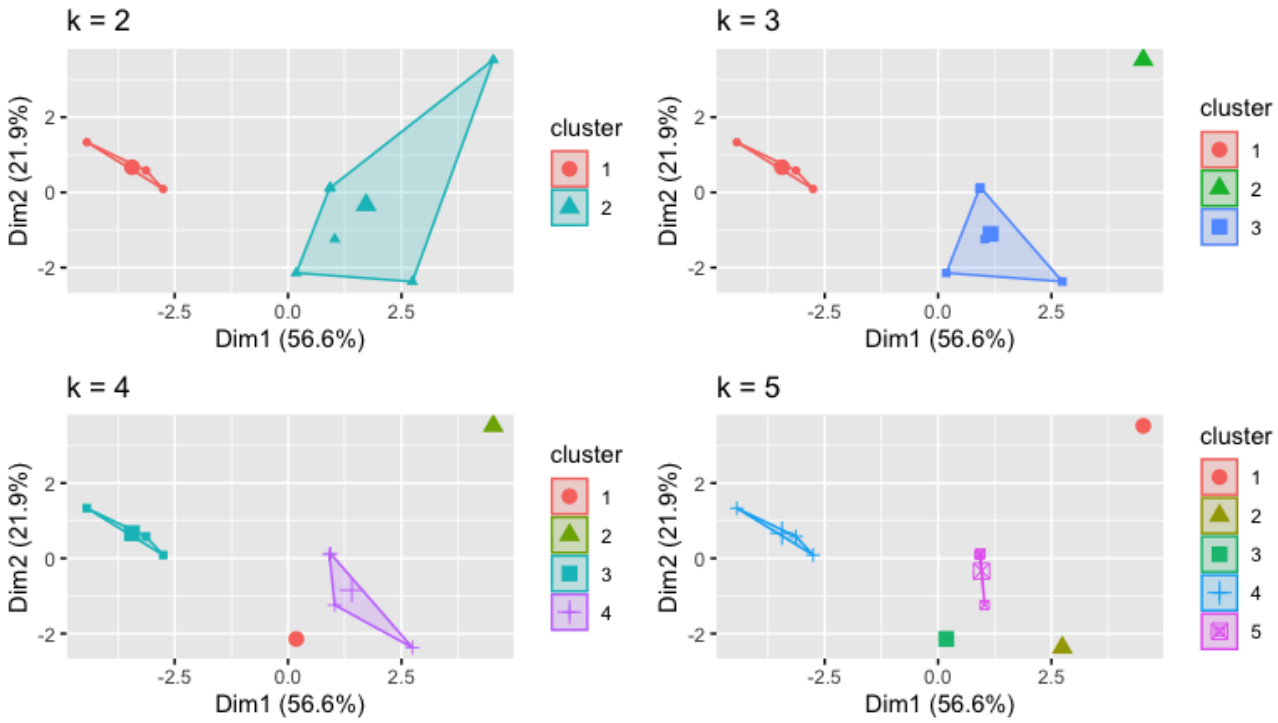

Figure 21. A K – means cluster plot. This figure shows K – means clustering results for different values of K (number of clusters). For K = 5, the dataset is divided into 5 clusters. By this point, some clusters may become too small, potentially leading to overfitting which is to be avoided. This gives us an indication of what our final cluster plot should not look like.

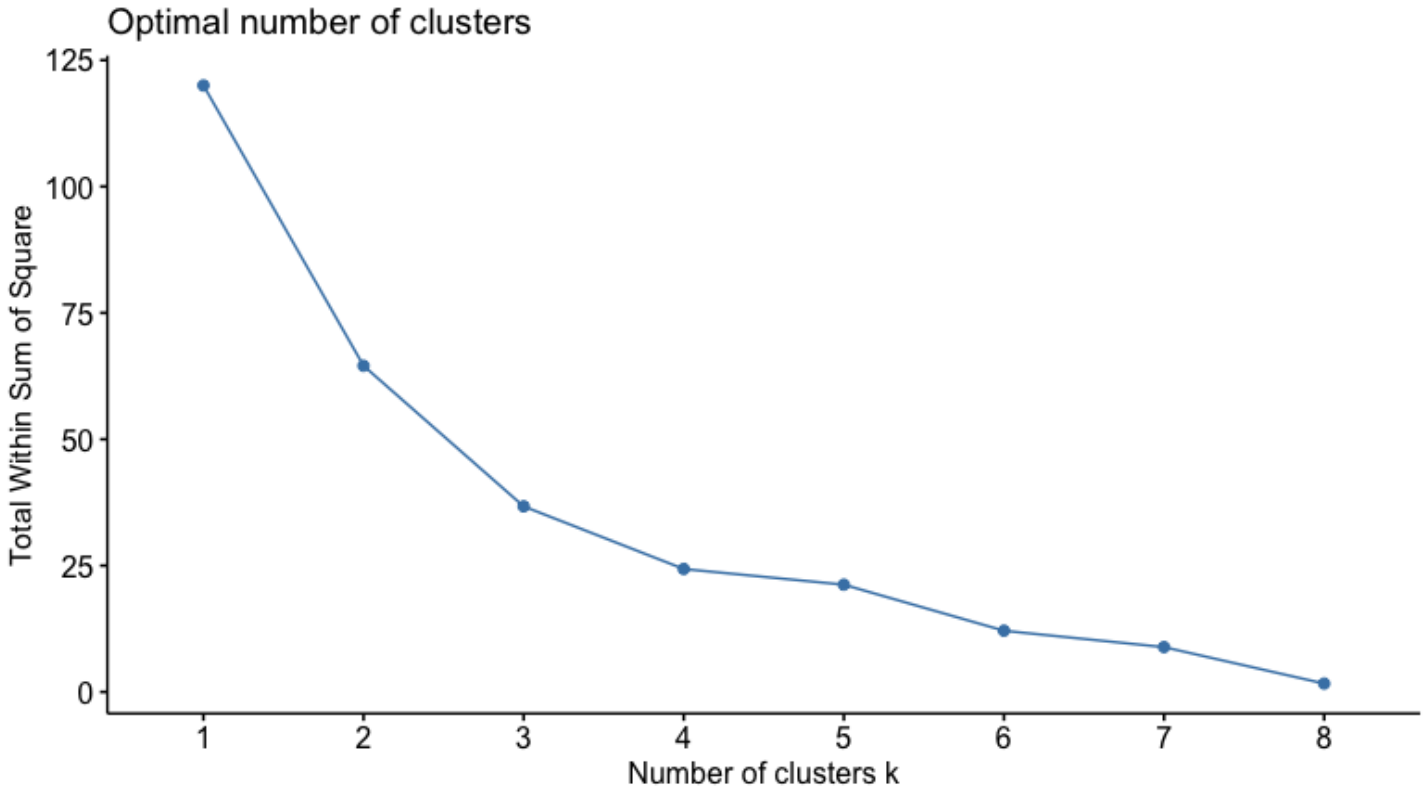

Figure 22. An Elbow method plot. This figure illustrates the Elbow method which helps decide the optimal number of clusters for the dataset. This method shows 2 is the optimal number of clusters for the super region dataset.

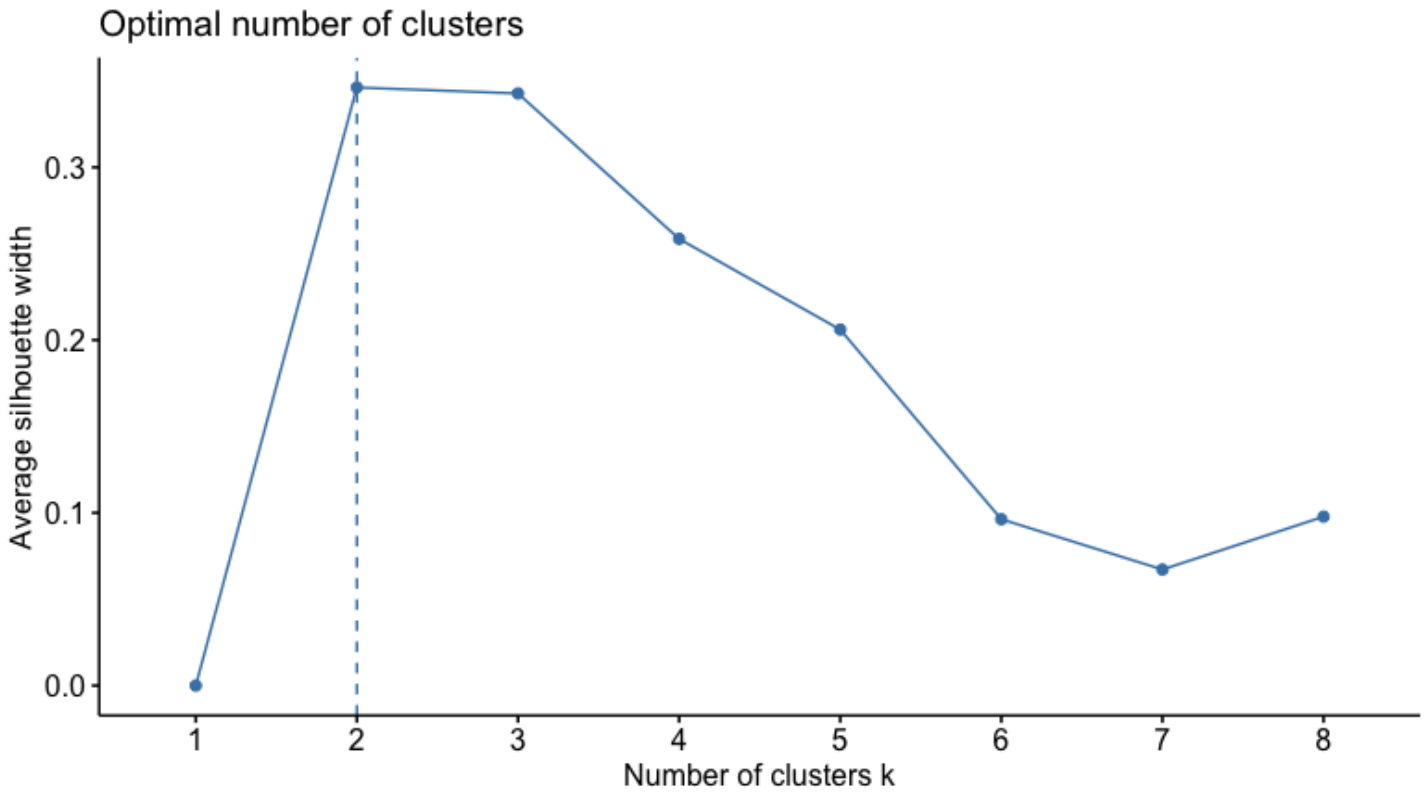

Figure 23. A Silhouette method plot. This figure illustrates the Silhouette method which also helps decide the optimal number of clusters for the dataset. This method also shows 2 is the optimal number of clusters for the super region dataset.

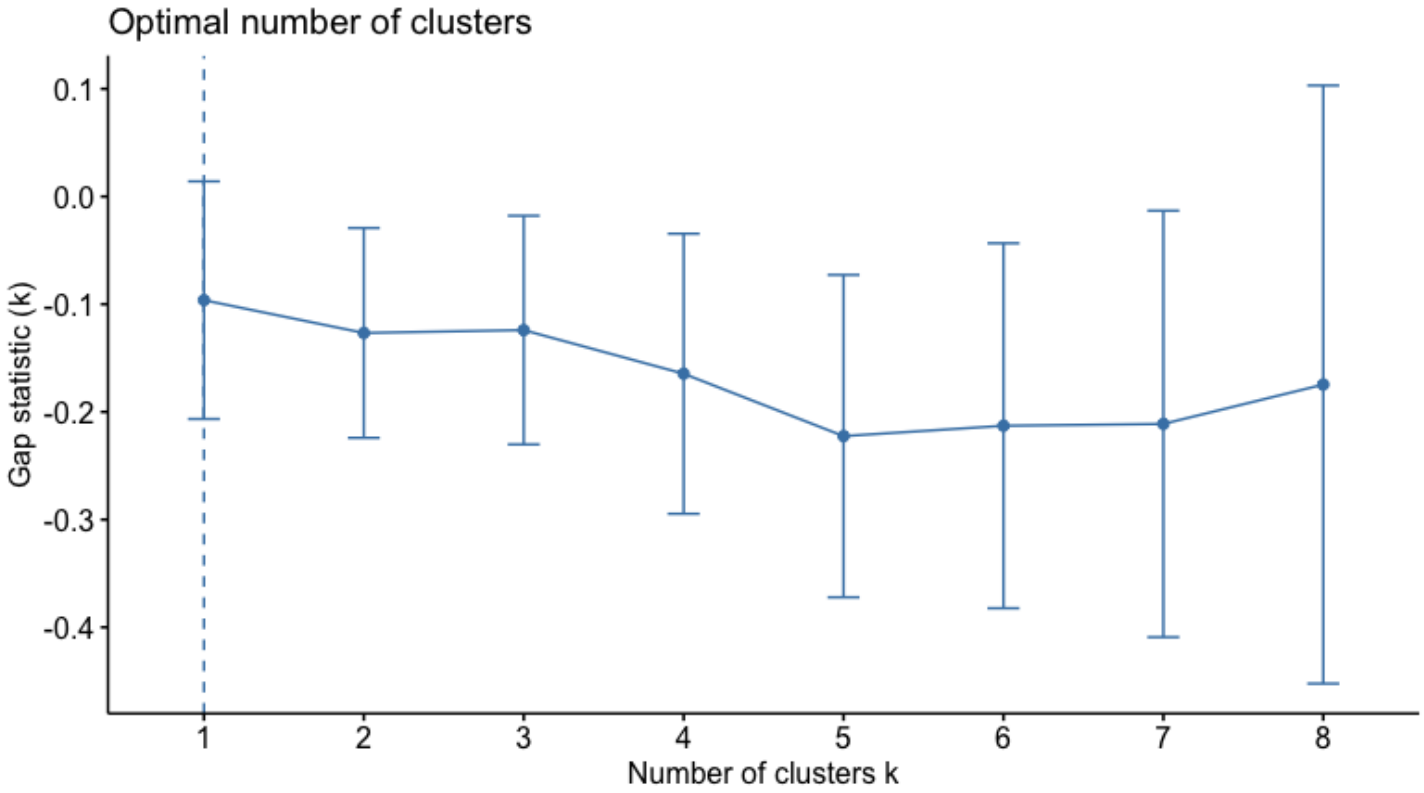

Figure 24. A Gap Statistic method plot. This figure illustrates the Gap Statistic method which also helps decide the optimal number of clusters for the super region dataset. This method shows 1 is the optimal number of clusters. Upon consideration, since the other two methods showed 2 is the optimal number of clusters, it was decided 2 clusters for our final super region cluster plot was best fit.

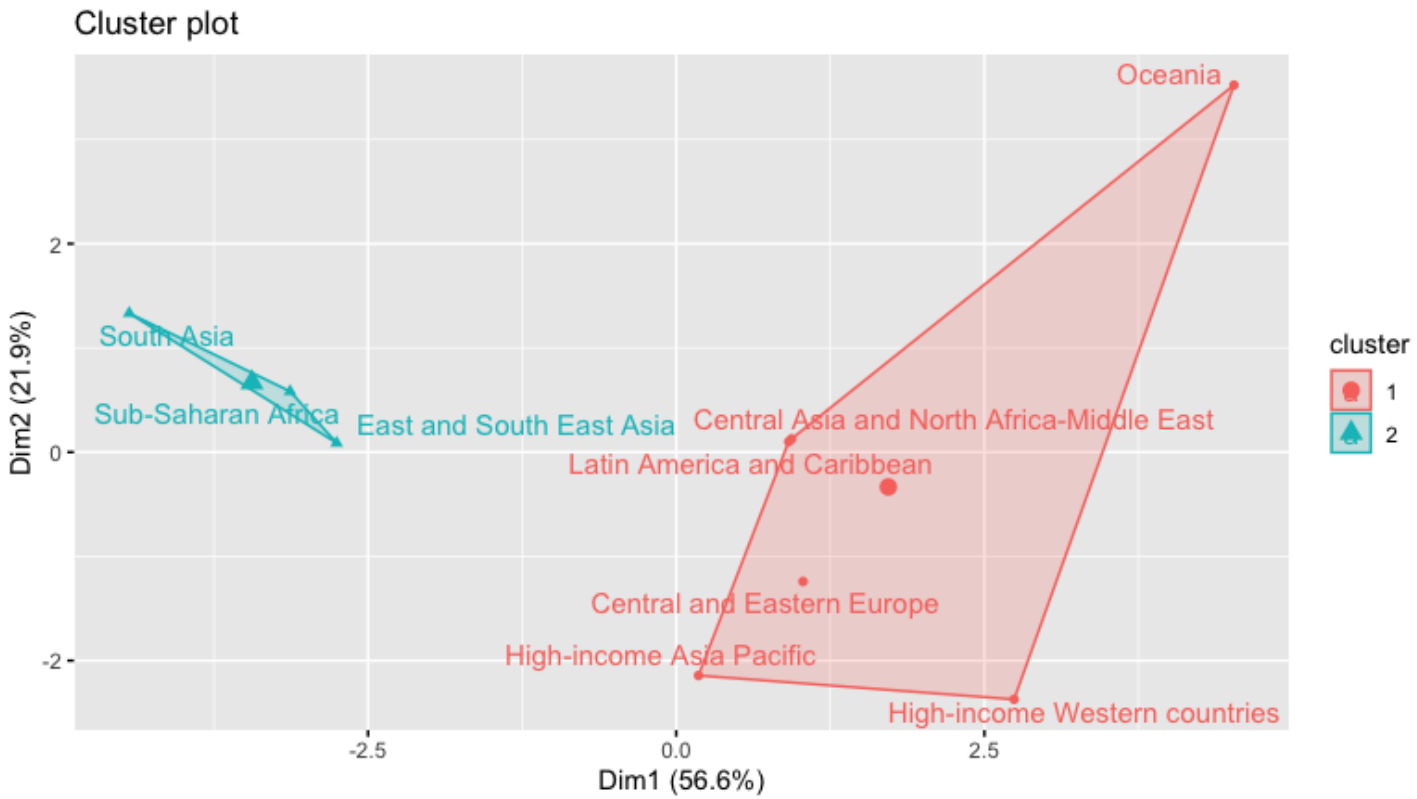

Figure 25. A super region cluster plot. This figure shows the final cluster plot for the super region dataset. The cluster plot groups regions into 2 main clusters. Cluster 1 groups all the regions with higher health – related and socioeconomic/ demographic metric rates and cluster 2 groups all the regions with lower health – related and socioeconomic/demographic metric rates. As expected from the distance matrix heatmap, all the higher income regions were grouped into cluster 1 and the lower income regions grouped into cluster 2.

4.2 Countries

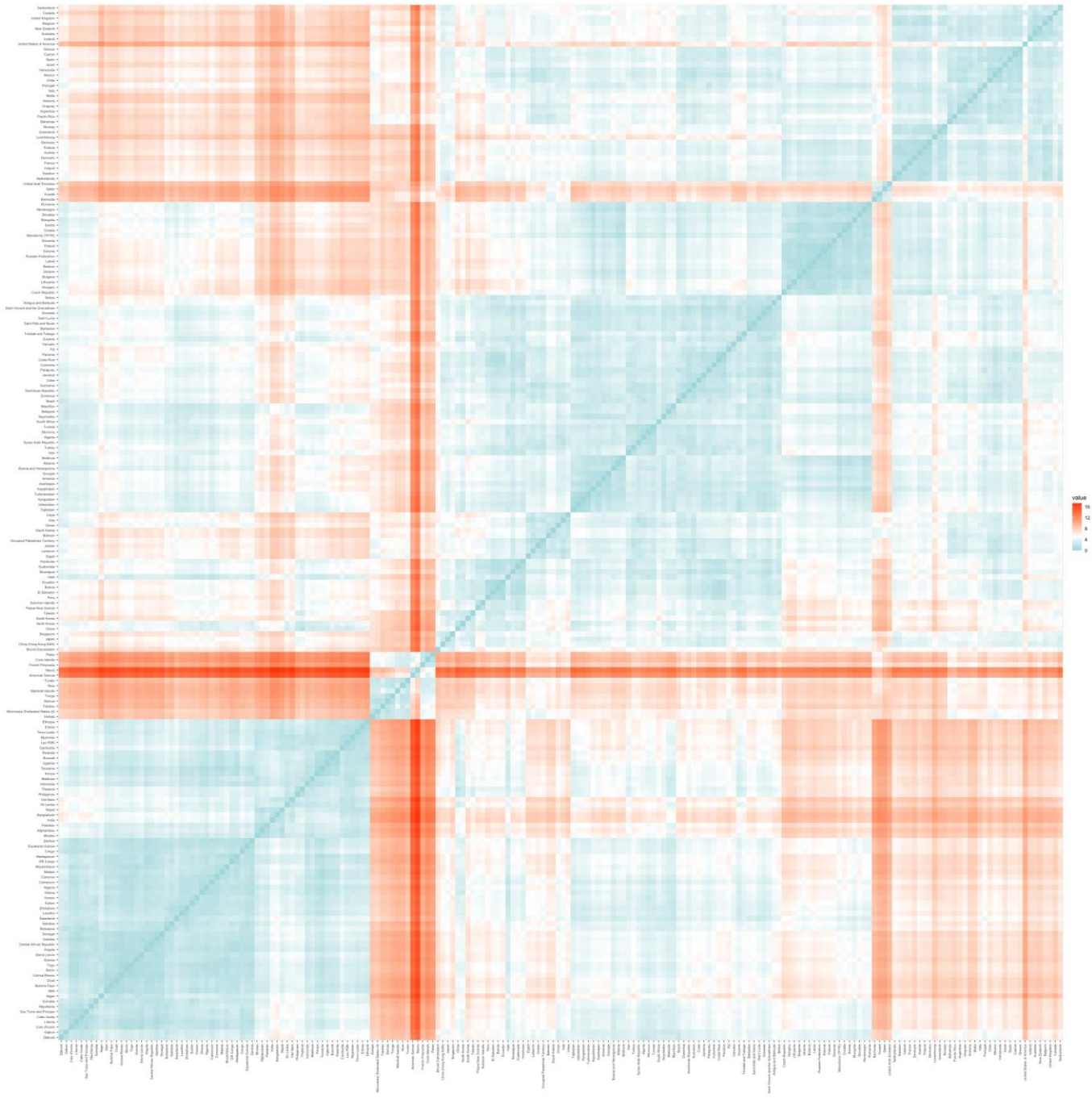

Figure 26. A distance heatmap matrix plot on the country dataset. This figure presents a distance matrix heatmap based on scaled data on countries regarding health related and socioeconomic/demographic metrics. However, given the complexity and breadth of the data, direct interpretation is not immediately informative. To further explore these relationships, cluster analyses will be used to highlight key trends in a more digestible format.

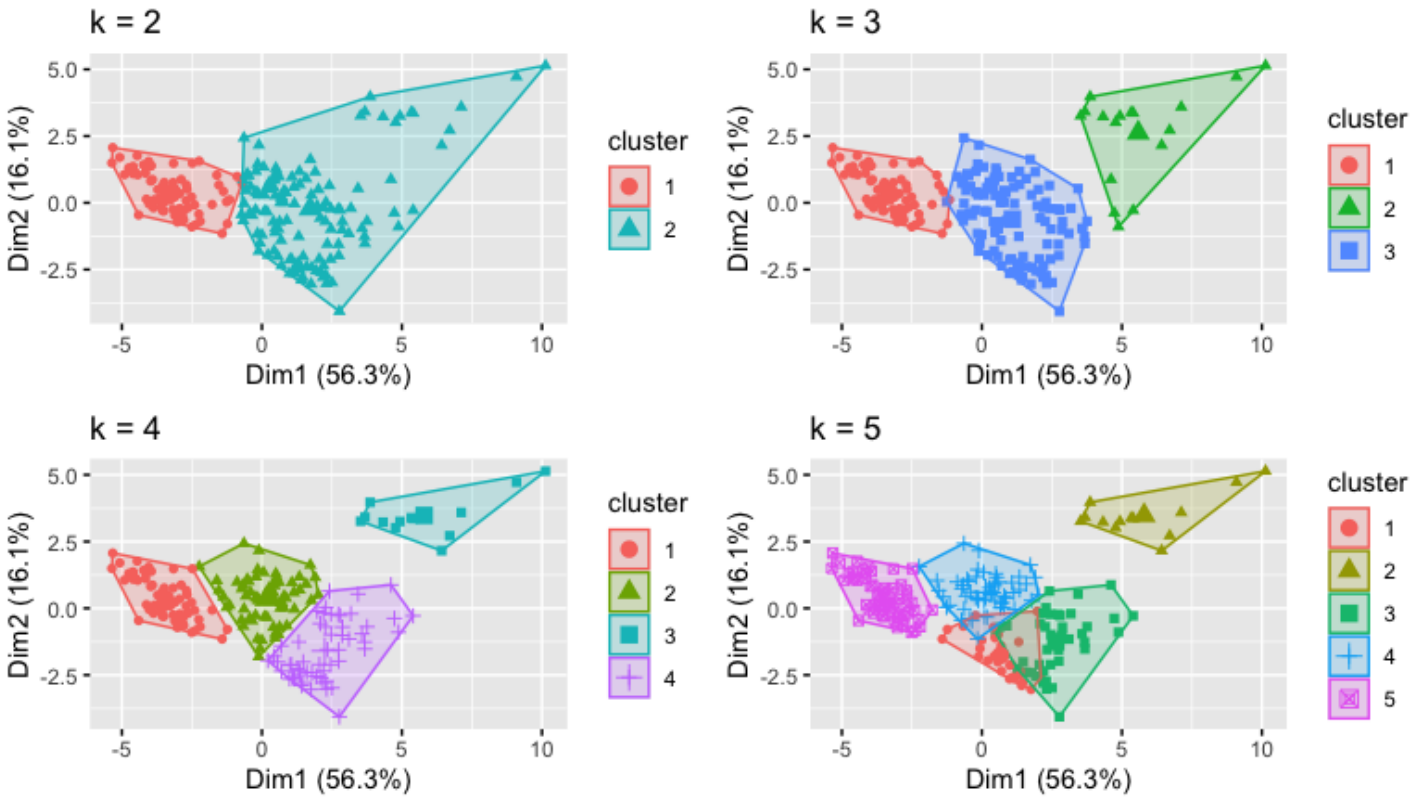

Figure 27. A K – means cluster plot. This figure shows K – means clustering results for different values of K (number of clusters). For K = 5, the dataset is divided into 5 clusters. By this point, some clusters may become too small, potentially leading to overfitting which is to be avoided. This gives us an indication of what our final cluster plot should not look like.

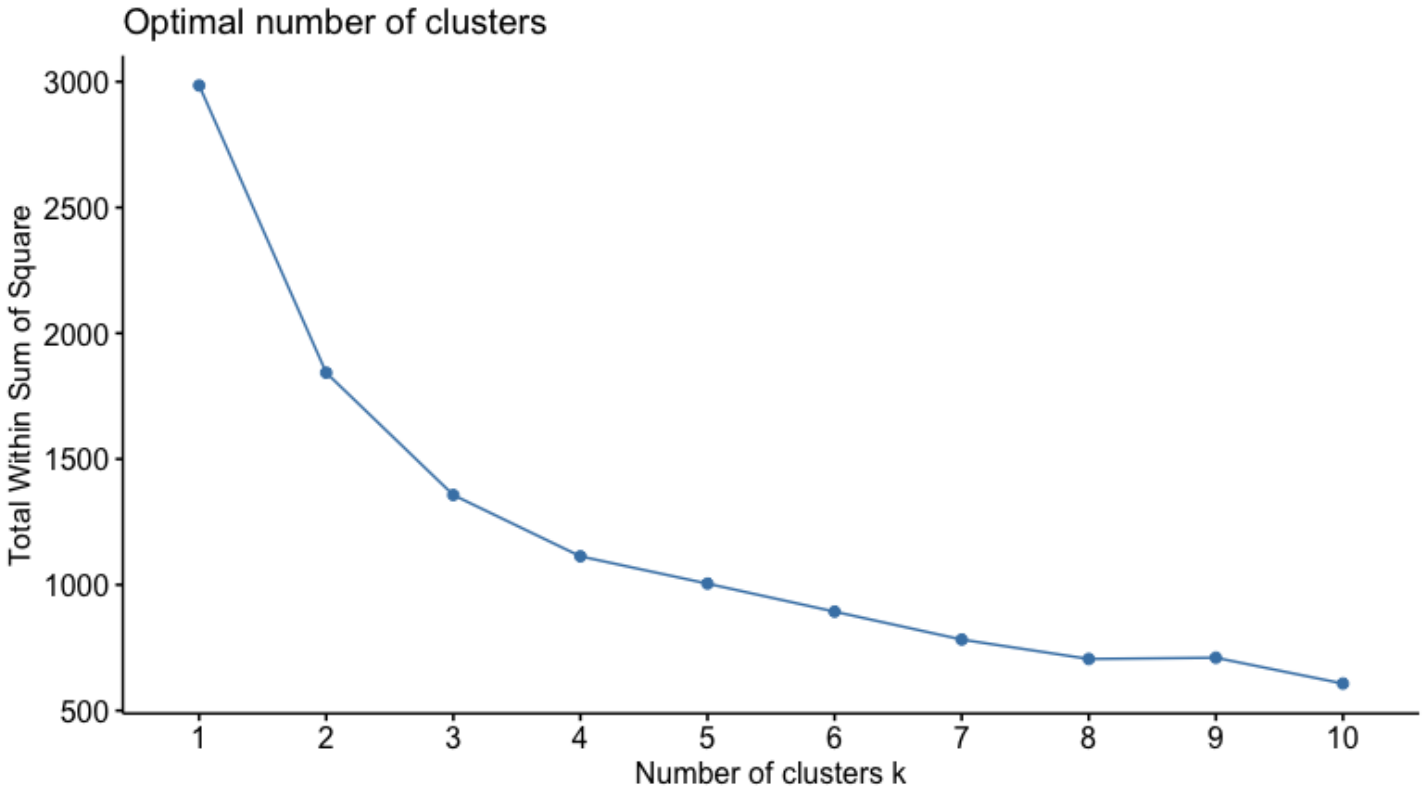

Figure 28. An Elbow method plot. This figure illustrates the Elbow method which helps decide the optimal number of clusters for the dataset. This method shows 2 is the optimal number of clusters for the country dataset.

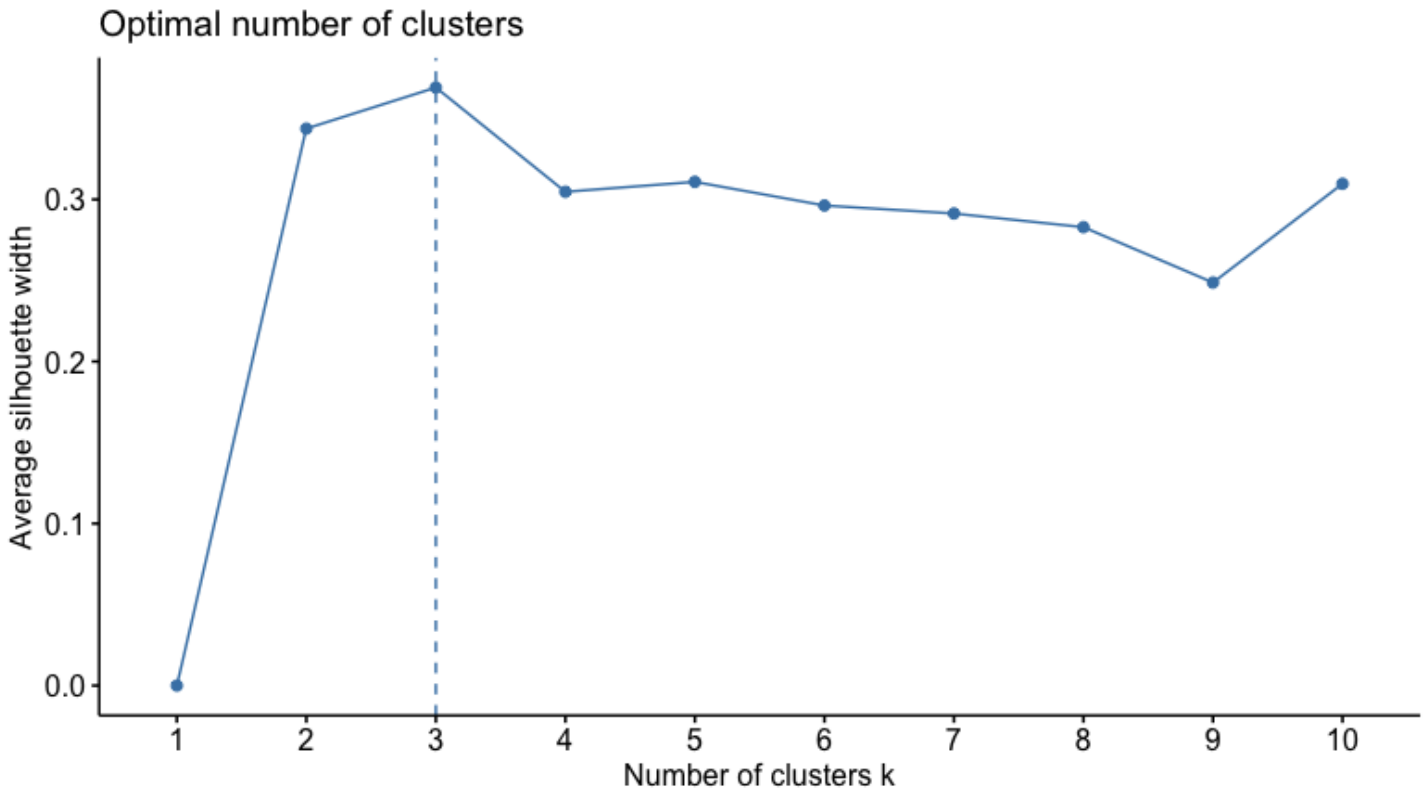

Figure 29. A Silhouette method plot. This figure illustrates the Silhouette method which also helps decide the optimal number of clusters for the dataset. This method differs to the Elbow method and instead shows 3 is the optimal number of clusters for the country dataset.

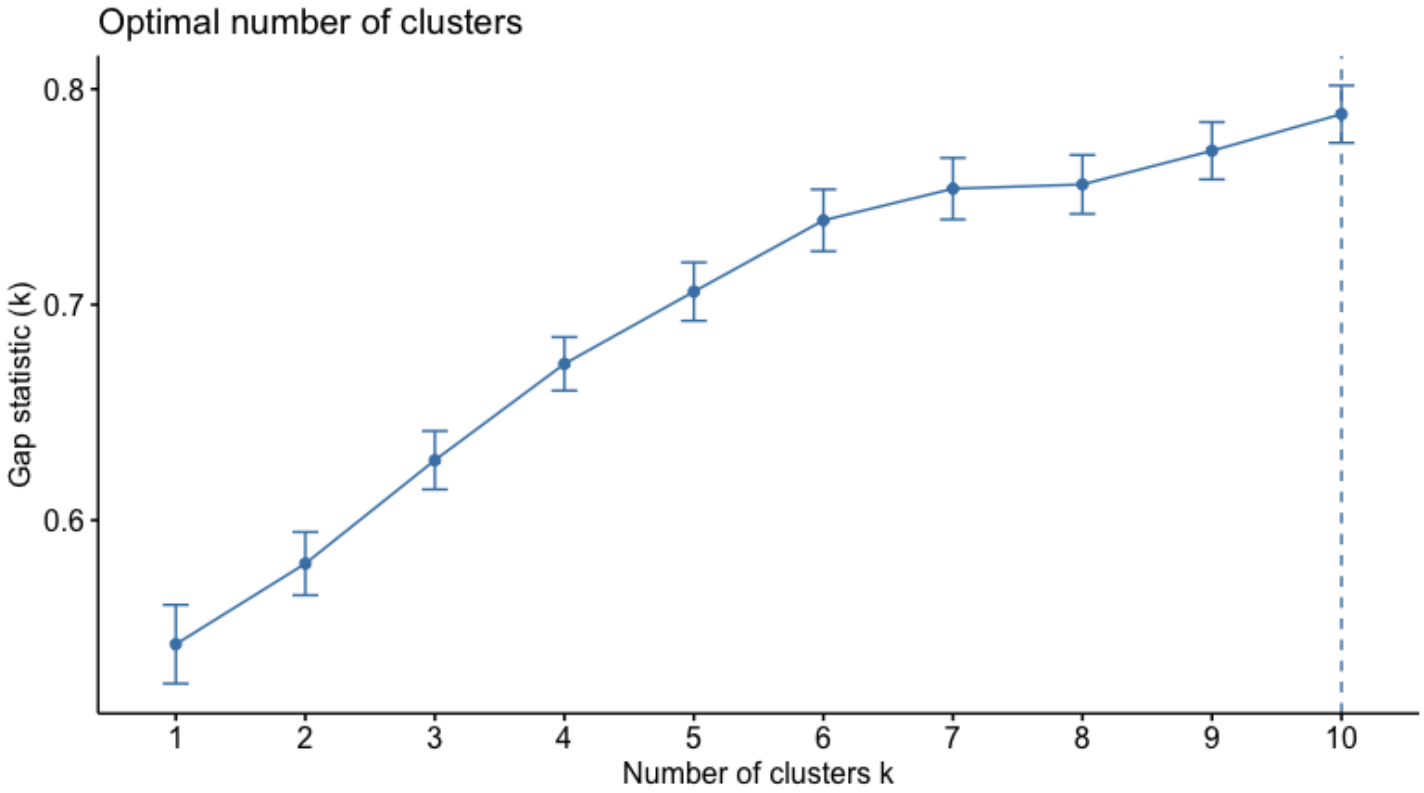

Figure 30. A Gap Statistic method plot. This figure illustrates the Gap Statistic method which also helps decide the optimal number of clusters for the country dataset. This method differs greatly from the last two and shows 10 as the optimal number of clusters for the country dataset.

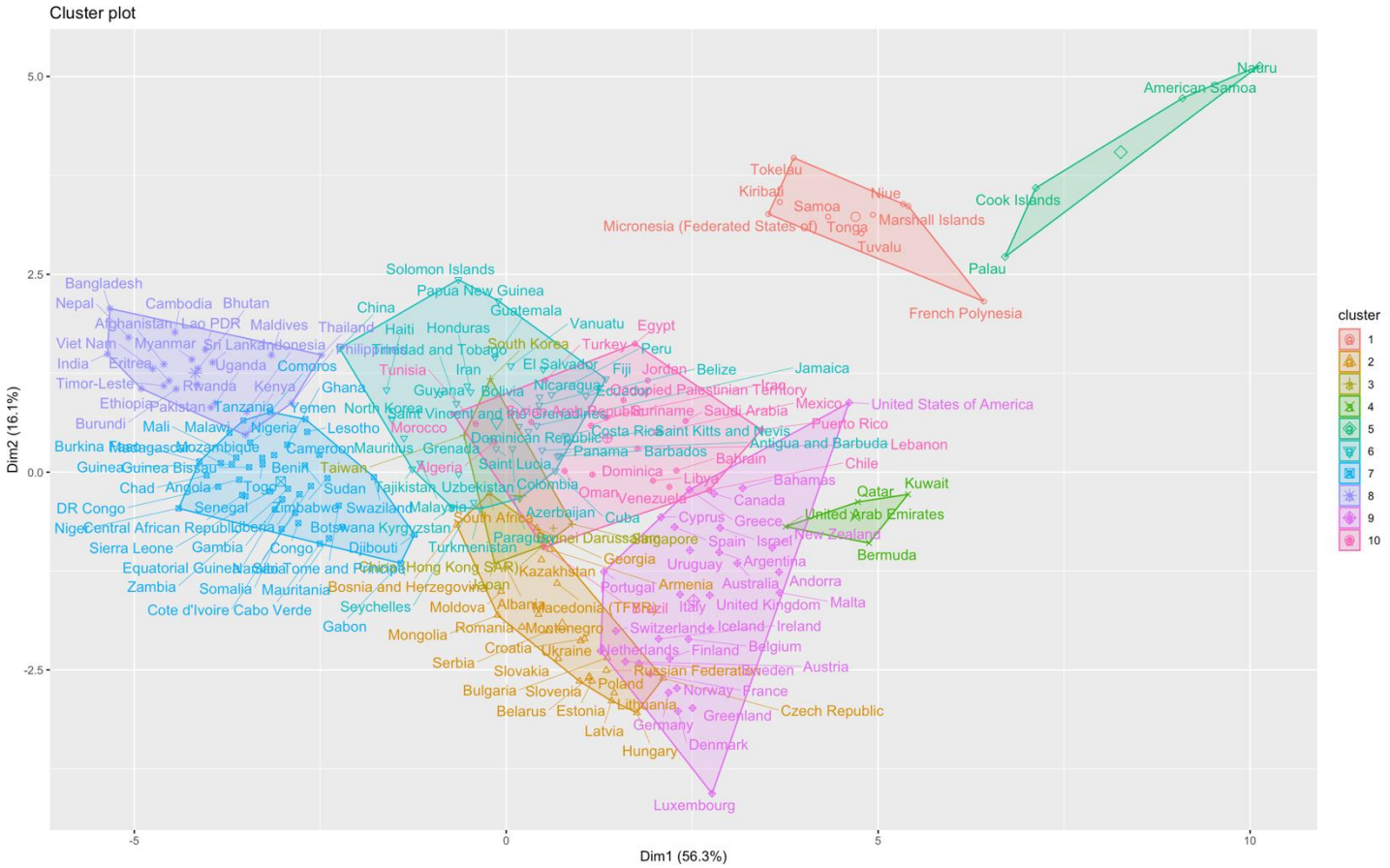

Figure 31. A cluster plot for the country data set. This figure visualises the grouping of counties into 10 clusters according to health related and socioeconomic/demographic metrics. However, despite being recommended by the Gap Statistic method, the cluster analysis appears overfitted, with an excessive number of clusters that reduce interpretability. The overlapping regions make it difficult to extract meaningful insights. As such, a simpler cluster analysis approach was taken below.

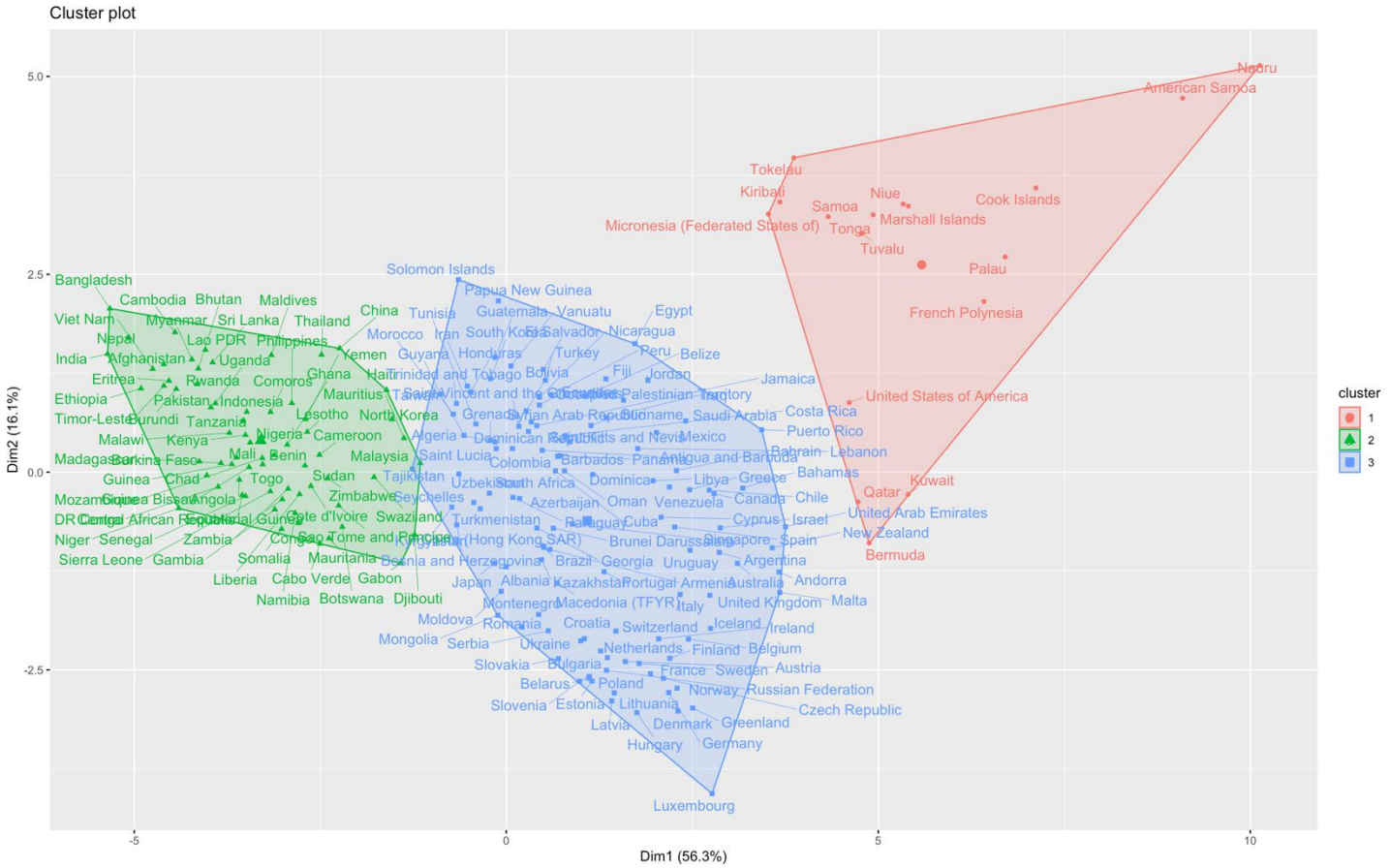

Figure 32. The final cluster plot for the country data set. This plot visualises the grouping of countries into 3 clusters according to health related and socioeconomic/demographic metrics. This country cluster plot validates the broad regional trends observed in the super region analysis. Cluster 1 consists of High-Income Asia & Oceania, the US, and Pacific Island nations (e.g., Samoa, Tonga and Nauru). This highlights that despite their geographic spread, these nations share common metric rates due to high – income lifestyles and trends. Cluster 2 consists mostly of Sub – Saharan African and South Asian countries, reinforcing the strong similarity observed between lower income status regions in the super region analysis. Cluster 3 includes Central and Eastern European, Central Asian, and North African/Middle Eastern nations, aligning with the observed similarities in the super region analysis. To summarise, the country cluster plot analysis cements what is seen in the super region cluster plot analysis. High – income regions (Western countries, Asia Pacific, and Oceania) tend to be highly different from lower- income regions (Sub – Saharan Africa and South Asia) when it comes to health – related metrics. The clustering also supports geographically close regions tend to be more similar due to shared cultural and lifestyle habits e.g., East & Southeast Asia and South Asia.

5. Conclusion

In conclusion, the assignment provides an in-depth analysis of obesity and BMI trends using big data and machine learning techniques. Through data cleaning, exploratory data analysis, correlation methods and clustering, we identified key patterns in health, demographics and socioeconomic

factors influencing BMI levels globally. The analysis highlights significant regional disparities, with high-income regions showing higher BMI and obesity prevalence, while lower-income countries exhibit lower BMI values, which is linked with undernutrition. Clustering further proved these trends, grouping regions based on health-related metrics. The frequency distribution for BMI for children and adults highlights the need for public health efforts to tackle both obesity and malnutrition and promote well-being globally. Additionally, the boxplots analysis of the UK data showed sex-based differences in BMI, obesity prevalence, and systolic blood pressure. The results showed males having higher BMI variability and blood pressure while females have a greater obesity prevalence these results indicate that biological difference and other factors for example lifestyle should be considered to understand health inequalities.

6. Appendix

```
# Load necessary libraries

# Use R version 4.3.3 or higher and the following packages:

# tidyverse (version 2.0.0), tseries (version 0.10-58)

library(tseries)

library(gridExtra) #v2.3

library(tidyverse) #v.2.0.0

library(cluster) #v2.1.4

library(factoextra) #v1.0.7

library(ggplot2) #v3.4.0

library(tidyr) #v1.3.0


# Load data

data <- read.csv("IntOrg_NCD_variables_2024_02_02.csv")


# Remove rows with missing values

data <- na.omit(data)

head(data)


#Summary of the data
```

```
full_summary <-summary(data %>% select(Mean_BMI_children, Mean_BMI_adults,
Prevalence_obesity_children, Prevalence_obesity_adults,
                                       Systolic_blood_pressure, Years_of_education))
print(full_summary)

#Filter the dataset for the United Kingdom
uk_data <- data %>% filter(Country == "United Kingdom")

# Summary statistics for key health indicators in the United Kingdom
uk_summary <- summary(uk_data %>% select(Mean_BMI_children,
Mean_BMI_adults,Prevalence_obesity_children, Prevalence_obesity_adults,
                                         Systolic_blood_pressure, Years_of_education))
print(uk_summary)

#Boxplot: Mean BMI for Adults in the United Kingdom
ggplot(uk_data, aes(x = Sex, y = Mean_BMI_adults, fill = Sex)) + geom_boxplot() + labs(title =
                                                                                         "Mean BMI for Adults in the UK", x = "Sex", y
= "Mean BMI (Adults)") + theme_minimal()

#Boxplot: Mean BMI for Children in the United Kingdom
ggplot(uk_data, aes(x = Sex, y = Mean_BMI_children, fill = Sex)) + geom_boxplot() + labs(title =
                                                                                          "Mean BMI for Children in the UK", x =
"Sex", y = "Mean BMI (Children)") + theme_minimal()

#Boxplot: Obesity Prevalence for Adults in the United Kingdom
ggplot(uk_data, aes(x = Sex, y = Prevalence_obesity_adults, fill = Sex)) + geom_boxplot() + labs(title
=
                                                                                                 "Obesity Prevalence for Adults in the
UK", x = "Sex", y = "Prevalence of Obesity (Adults)") + theme_minimal()
```

```
#Boxplot: Obesity Prevalence for Children in the United Kingdom

ggplot(uk_data, aes(x = Sex, y = Prevalence_obesity_children, fill = Sex)) + geom_boxplot() + labs(title =

                                                                "Obesity Prevalence for Children in the UK", x = "Sex", y = "Prevalence of Obesity (Children)") + theme_minimal()


#Boxplot: Systolic Blood Pressure in the United Kingdom

ggplot(uk_data, aes(x = Sex, y = Systolic_blood_pressure, fill = Sex)) + geom_boxplot() + labs(title =

                                                        "Systolic Blood Pressure in the UK", x = "Sex", y = "Systolic Blood Pressure") + theme_minimal()

ggplot(data = data, mapping = aes(x = Mean_BMI_children, colour = Superregion)) +

  geom_freqpoly(binwidth = 0.5) +

  labs(title = "Frequency Distribution of Mean BMI for Children", x = "Mean BMI Children", y = "Frequency")



# Frequency distribution of Mean BMI for adults by Superregion

ggplot(data = data, mapping = aes(x = Mean_BMI_adults, colour = Superregion)) +

  geom_freqpoly(binwidth = 0.5) +

  labs(title = "Frequency Distribution of Mean BMI for Adults", x = "Mean BMI Adults", y = "Frequency")


# Visualising the correlation using heatmaps

correlation_matrix <- cor(data %>% select_if(is.numeric))

heatmap(correlation_matrix, main = "Correlation Heatmap", col = heat.colors(256), scale = "column", margins = c(14, 10))
```

```
# Plot a scatter plot for Mean_BMI_adults vs Prevalence_obesity_adults in United Kingdom
ggplot(data %>% filter(Country == "United Kingdom"), aes(x = Mean_BMI_adults, y = Prevalence_obesity_adults, color = Sex)) +
  geom_point() +
  labs(title = "Scatter Plot of Mean BMI Adults vs. Prevalence of Obesity Adults in United Kingdom",
       x = "Mean BMI Adults",
       y = "Prevalence of Obesity Adults") +
  theme_minimal()

# Plot a scatter plot for Mean_BMI_children vs Prevalence_obesity_children in United Kingdom
ggplot(data %>% filter(Country == "United Kingdom"), aes(x = Mean_BMI_children, y = Prevalence_obesity_children, color = Sex)) +
  geom_point() +
  labs(title = "Scatter Plot of Mean BMI Children vs. Prevalence of Obesity Children in United Kingdom",
       x = "Mean BMI Children",
       y = "Prevalence of Obesity Children") +
  theme_minimal()


# Plot a scatter plot for Urbanisation vs Mean BMI Adults in United Kingdom
ggplot(data %>% filter(Country == "United Kingdom"), aes(x = Urbanisation, y = Mean_BMI_adults, color = Sex)) +
  geom_point() +
  labs(title = "Scatter Plot of Urbanisation vs. Mean BMI Adults in United Kingdom",
       x = "Urbanisation",
       y = "Mean BMI Adults") +
  theme_minimal()
```

```
# Plot a scatter plot of Urbanisation vs vs. Mean BMI Children in United Kingdom

ggplot(data %>% filter(Country == "United Kingdom"), aes(x = Urbanisation, y = Mean_BMI_children, color = Sex)) +

  geom_point() +

  labs(title = "Scatter Plot of Urbanisation vs. Mean BMI Children in United Kingdom",

       x = "Urbanisation",

       y = "Mean BMI Children") +

  theme_minimal()


# Plot a scatter plot of GDP vs. Mean BMI Adults in United Kingdom

ggplot(data %>% filter(Country == "United Kingdom"), aes(x = GDP_USD, y = Mean_BMI_adults, color = Sex)) +

  geom_point() +

  labs(title = "Scatter Plot of GDP vs. Mean BMI Adults in United Kingdom",

       x = "GDP_USD",

       y = "Mean BMI Adults") +

  theme_minimal()


# Plot a scatter plot of GDP vs.Mean BMI Children in United Kingdom

ggplot(data %>% filter(Country == "United Kingdom"), aes(x = GDP_USD, y = Mean_BMI_children, color = Sex)) +

  geom_point() +

  labs(title = "Scatter Plot of GDP vs. Mean BMI Children in United Kingdom",

       x = "GDP_USD",

       y = "Mean BMI Children") +

  theme_minimal()
```

```
# Plot a scatter plot of Mean BMI Adults vs. Systolic blood pressure in United Kingdom
ggplot(data %>% filter(Country == "United Kingdom"), aes(x = Mean_BMI_adults, y = Systolic_blood_pressure, color = Sex)) +
  geom_point() +
  labs(title = "Scatter Plot of Mean BMI Adults vs. Systolic blood pressure in United Kingdom",
       x = "Mean BMI Adults",
       y = "Systolic blood pressure") +
  theme_minimal()
```

```
# Plot a scatter plot of Mean BMI Children vs. Systolic blood pressure in United Kingdom
ggplot(data %>% filter(Country == "United Kingdom"), aes(x = Mean_BMI_children, y = Systolic_blood_pressure, color = Sex)) +
  geom_point() +
  labs(title = "Scatter Plot of Mean BMI Children vs. Systolic blood pressure in United Kingdom",
       x = "Mean BMI Children",
       y = "Systolic blood pressure") +
  theme_minimal()
```

```
# Plot a scatter plot of years of education vs. Mean BMI Adults in United Kingdom
ggplot(data %>% filter(Country == "United Kingdom"), aes(x = Years_of_education, y = Mean_BMI_adults, color = Sex)) +
  geom_point() +
  labs(title = "Scatter Plot of years of education vs. Mean BMI Adults in United Kingdom",
       x = "Years of education",
       y = "Mean BMI Adults") +
```

```
  theme_minimal()


# Plot a scatter plot of years of education vs. Mean BMI Children in United Kingdom
ggplot(data %>% filter(Country == "United Kingdom"), aes(x = Years_of_education, y = Mean_BMI_children, color = Sex)) +
  geom_point() +
  labs(title = "Scatter Plot of years of education vs. Mean BMI Children in United Kingdom",
       x = "Years of education",
       y = "Mean BMI Children") +
  theme_minimal()


# Plot a scatter plot of Mean BMI Adults vs. Mean BMI Children in United Kingdom
ggplot(data %>% filter(Country == "United Kingdom"), aes(x = Mean_BMI_adults, y = Mean_BMI_children, color = Sex)) +
  geom_point() +
  labs(title = "Scatter Plot of Mean BMI Adults vs. Mean BMI Children in United Kingdom",
       x = "Mean BMI Adults",
       y = "Mean BMI Children") +
  theme_minimal()


# Aggregate data by superregion
superregion_data <- data %>%
  group_by(Superregion) %>%
  summarise(across(where(is.numeric), mean, na.rm = TRUE))


# View the aggregated data
head(superregion_data)
```

```
# Remove non-numeric columns (e.g., Superregion) before scaling
df_scaled <- superregion_data %>%
  select(-Superregion) %>%
  scale()

# Add Superregion back as row names
rownames(df_scaled) <- superregion_data$Superregion

# View the scaled data
head(df_scaled)

# Calculate distance matrix
distance <- get_dist(df_scaled)

# Visualize the distance matrix
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))

# Remove the Year column using base R
df_scaled <- df_scaled[, !colnames(df_scaled) == "Year"]

# Check the result
head(df_scaled)

# Perform k-means clustering for different k values
k2 <- kmeans(df_scaled, centers = 2, nstart = 25)
k3 <- kmeans(df_scaled, centers = 3, nstart = 25)
k4 <- kmeans(df_scaled, centers = 4, nstart = 25)
```

```
k5 <- kmeans(df_scaled, centers = 5, nstart = 25)

# Create cluster plots for each k
p1 <- fviz_cluster(k2, geom = "point", data = df_scaled) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point", data = df_scaled) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point", data = df_scaled) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point", data = df_scaled) + ggtitle("k = 5")

# Arrange the plots in a grid
grid.arrange(p1, p2, p3, p4, nrow = 2)

#elbow method
set.seed(123)
rows <- nrow(df_scaled)
fviz_nbclust(df_scaled, kmeans, method = "wss", k.max = rows - 1)
#elbow is at 2

fviz_nbclust(df_scaled, kmeans, method = "silhouette", k.max = rows - 1)
#silhouette method says 2 is the optimal number of clusters

gap_stat <- clusGap(df_scaled,FUN = kmeans, nstart = 25,K.max = rows - 1, B = 50)
fviz_gap_stat(gap_stat)
#gap statistic says that 1 cluster is optimal

# Perform k-means clustering with 2 clusters
final <- kmeans(df_scaled, centers = 2, nstart = 25)
```

```
# Visualize the clusters with region names (labels)
fviz_cluster(final, data = df_scaled,
             geom = c("point", "text"),  # Show both points and labels
             repel = TRUE,  # Prevent overlapping label
)

# Aggregate data by country
country_data <- data %>%
  group_by(Country) %>%
  summarise(across(where(is.numeric), mean, na.rm = TRUE))

head(country_data)

df_scaled_country <- country_data %>%
  select(-Country) %>%
  scale()

# Add Superregion back as row names
rownames(df_scaled_country) <- country_data$Country

country_distance <- get_dist(df_scaled_country)

png("distance_matrix.png", width = 2000, height = 2000)  # Increase plot size
fviz_dist(country_distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07")) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5, size = 6),  # Rotate and reduce x-axis labels
        axis.text.y = element_text(size = 6))  # Reduce y-axis labels
dev.off()
```

```
# Remove the Year column using base R
df_scaled_country <- df_scaled_country[, !colnames(df_scaled_country) == "Year"]


# Perform k-means clustering for different k values
k2 <- kmeans(df_scaled_country, centers = 2, nstart = 25)
k3 <- kmeans(df_scaled_country, centers = 3, nstart = 25)
k4 <- kmeans(df_scaled_country, centers = 4, nstart = 25)
k5 <- kmeans(df_scaled_country, centers = 5, nstart = 25)

# Create cluster plots for each k
p1 <- fviz_cluster(k2, geom = "point", data = df_scaled_country) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point", data = df_scaled_country) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point", data = df_scaled_country) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point", data = df_scaled_country) + ggtitle("k = 5")

# Arrange the plots in a grid
grid.arrange(p1, p2, p3, p4, nrow = 2)

#elbow method
set.seed(123)
rows <- nrow(df_scaled_country)
fviz_nbclust(df_scaled_country, kmeans, method = "wss")
#elbow is at 3

country_rows <- nrow(df_scaled_country)
```

```
fviz_nbclust(df_scaled_country, kmeans, method = "silhouette")
#silhouette method says 3 is the optimal number of clusters

gap_stat <- clusGap(df_scaled_country,FUN = kmeans, nstart = 25,K.max = 10, B = 50)
fviz_gap_stat(gap_stat)
#gap statistic says that 9 clusters is optimal

# Perform k-means clustering with 9 clusters
final <- kmeans(df_scaled_country, centers = 10, nstart = 25)
# Visualize the clusters with region names (labels)
fviz_cluster(final, data = df_scaled_country,
             geom = c("point", "text"),  # Show both points and labels
             repel = TRUE,  # Prevent overlapping label)
#using 9 has alot of overlapping clusters

# Perform k-means clustering with 3 clusters
final <- kmeans(df_scaled_country, centers = 3, nstart = 25)
# Visualize the clusters with region names (labels)
fviz_cluster(final, data = df_scaled_country,
             geom = c("point", "text"),  # Show both points and labels
             repel = TRUE,  # Prevent overlapping label)
#works very nicely
```