MAST5956 Group Project

Group 10

Members

Luca Gabella (LMG49) (20%) (Member) - Report

Jesse Ohwariovbe (JEO22) (20%) (Member) - Chapter 1

Owen Graham (OG89) (20%) (Leader) - Chapter 2

Godwins Braimoh (GB514) (20%) (Member) - Chapter 3

Victor Taiwo (VT239) (20%) (Member) - Chapter 4

Introduction

For this project we were provided with the IntOrg Non-Communicable Diseases (NCD) variables dataset from 1975 to 2016. The main focus of the dataset was within obesity, BMI, and overweight and underweight prevalence within children and adults separately, also including blood pressure, both raised and systolic, and diabetes prevalence. Finally the dataset included information on the region and Superregion for each datapoint alongside other data to do with diet, urbanisation, education and GDP.

We looked at both health trends and socioeconomic trends individually before comparing how socioeconomic factors affect health trends. Using a collection of different analysis and graph types, and also clustering we were able to plot the data in a readable and understandable presentation for analysis.

Initialising

Assessing Data Quality

We started by checking the amount of NA values contained within the dataset and where they most often occurred, within this we found 6000 NA values of which 3600 occurred in the year 2016, we decided the best action was to omit all data from 2016 from our analysis.

```
sum(is.na(df)) #6000 instances of missing values in the data frame

sum(is.na(filter(df, Year == 2016))) #3600 instances of missing values for rows from 2016
df <- filter(df, Year != 2016)
```

We then checked where the other instances of NA occurred, they were all within the diabetes data, we decided not to omit diabetes data from our analysis but to take into account all the missing data when making said analysis.

```
na_counts <- df %>% summarise_all(~ sum(is.na(.))) #Per column which values have NA, all from Diabetes
```

Chapter 1

Section 1 - BMI and Obesity Trends by Superregion

Mean BMI Trends for Children

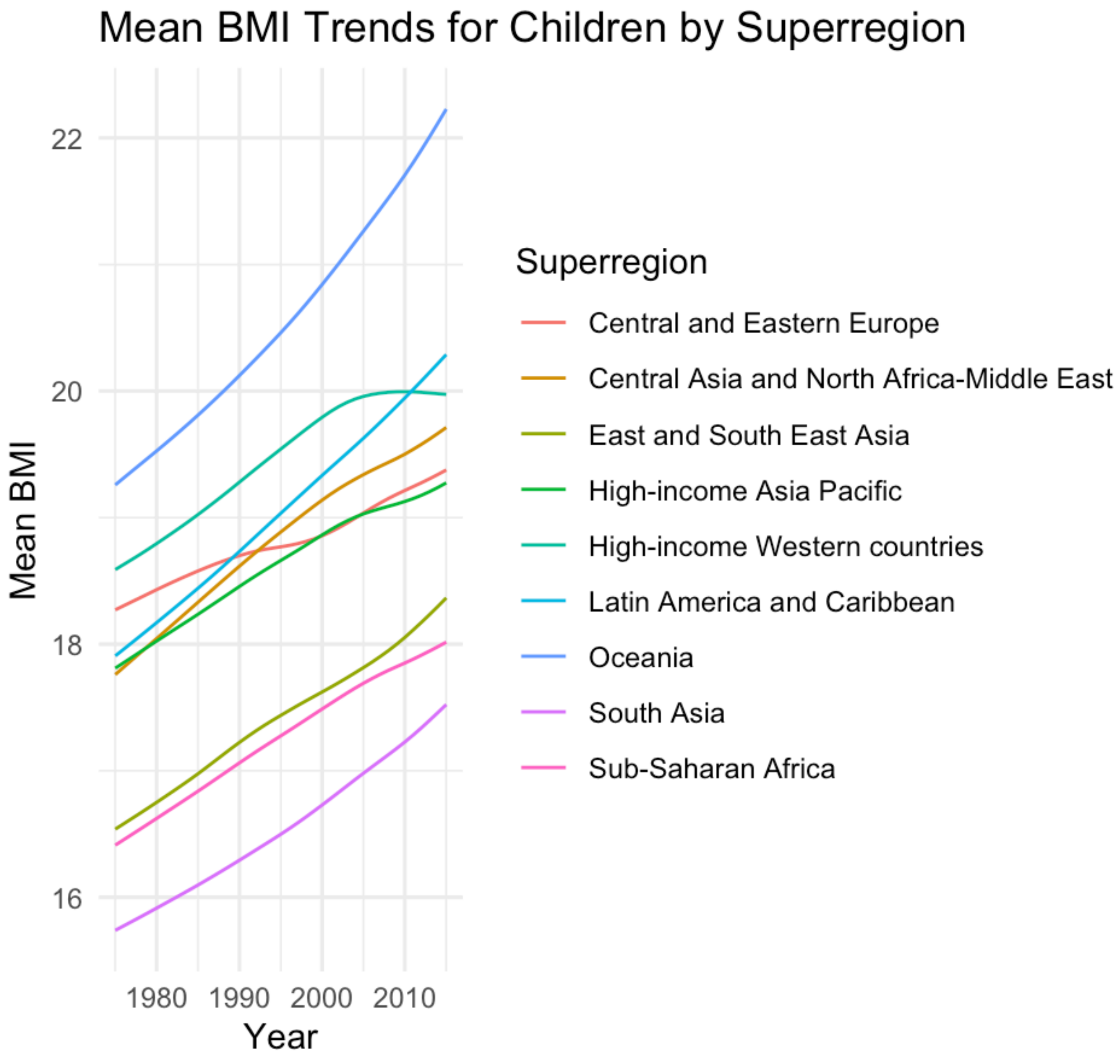

Figure 1.1

Over time, the mean BMI across all superregions has increased, with Oceania seeing the steepest increase and the other regions all seeing relatively similar trends. The main outlier from this data is that although the trend starts similar to all other superregions, in *High-Income Western Countries* the mean BMI trends have started to decrease since around 2010.

Mean BMI Trends for Adults

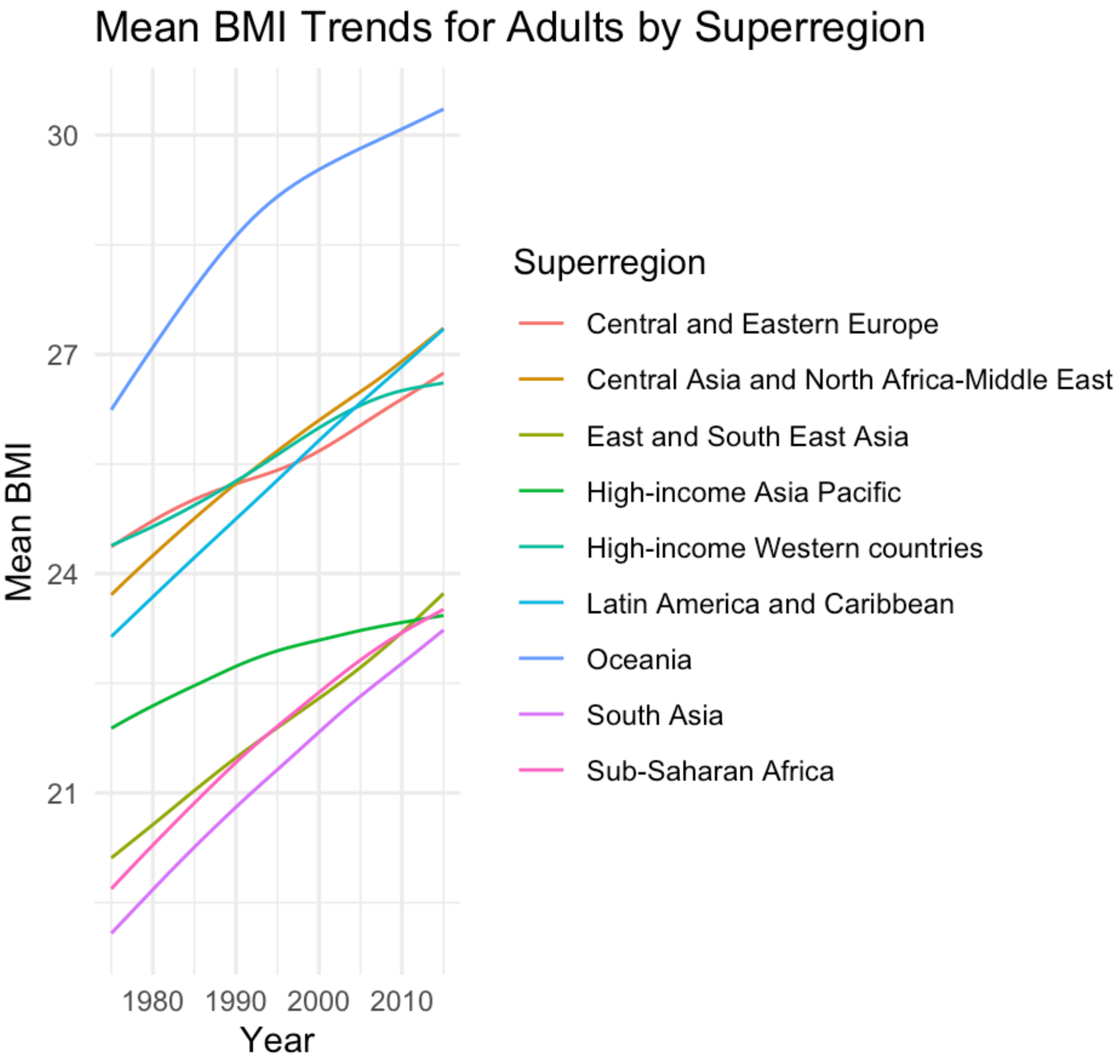

Figure 1.2

Similarly to children, the mean BMI across all superregions for adults has also increased over time although not to the same extent, the gradients of most superregions are similar to that of the graph for children although adults see more than one outlier in *High-Income Asia Pacific*, *High-Income Western Countries*, and *Oceania*. All three of these superregions are beginning to or have already flattened off showing a decrease in adult obesity trends.

Child Obesity Prevalence by Superregion

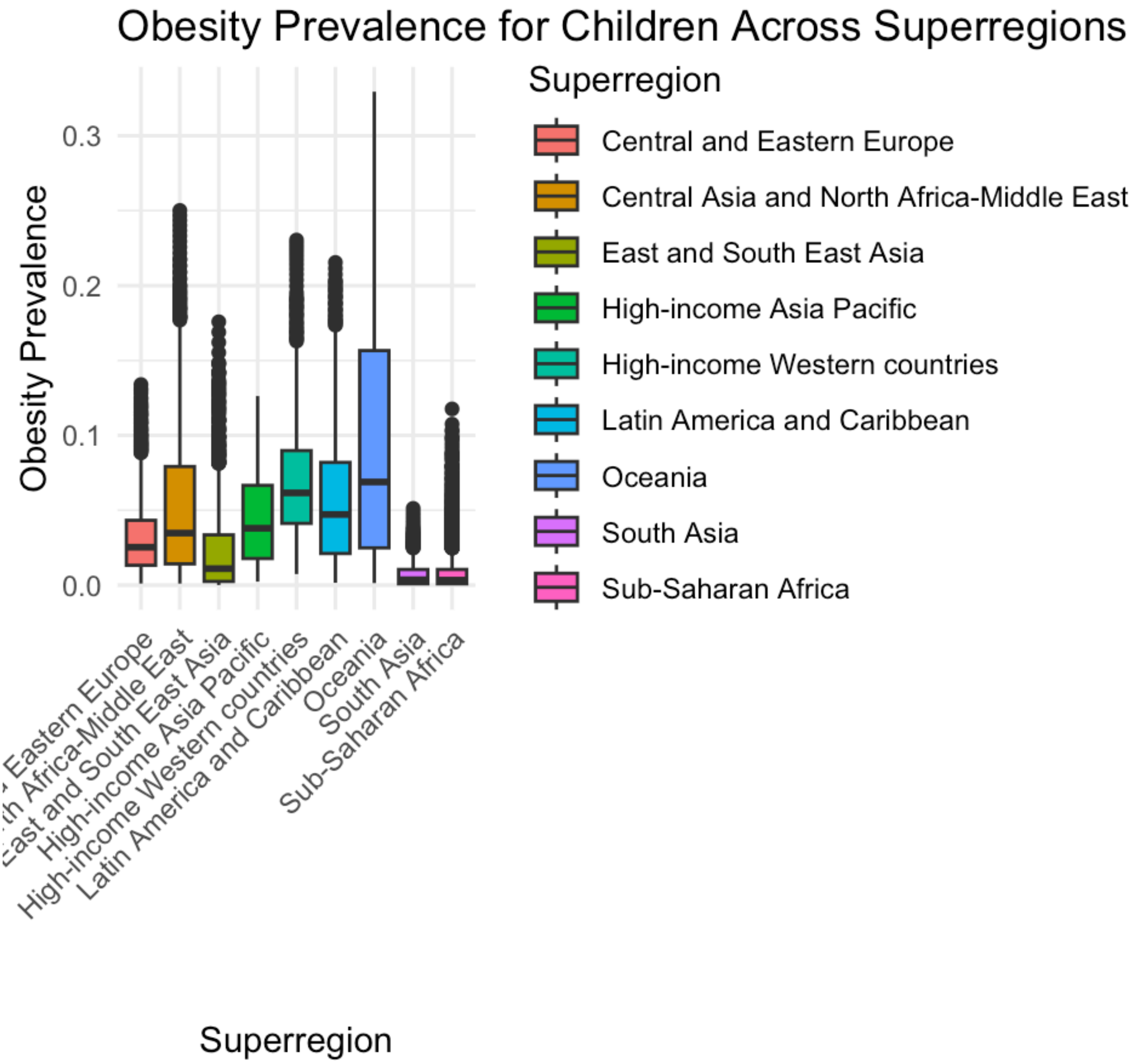

Figure 1.3

Most superregions show very different boxplots, with *South Asia* and *Sub-Saharan Africa* being the two that show the most similarity both with very low medians and small interquartile ranges whereas *Oceania* has the largest interquartile range, largest range and also highest median compared to all the other superregions showing that obesity is most prevalent in *Oceania*.

Adult Obesity Prevalence by Superregion

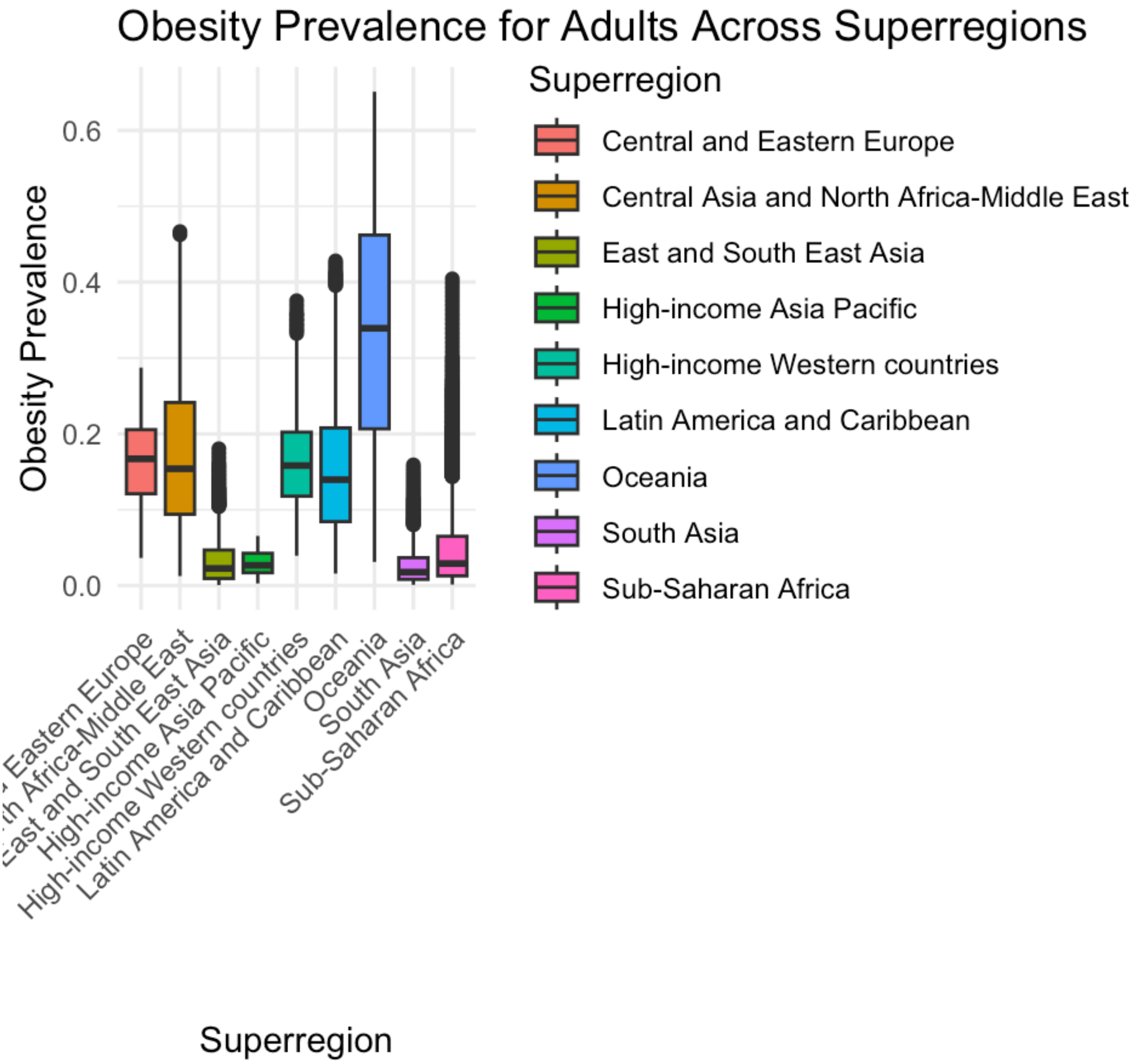

Figure 1.4

Similarly to child obesity plot, *South Asia* and *Sub-Saharan Africa* have the lowest obesity prevalences based on median and interquartile range but are also joined by *High-Income Asia Pacific* and *East and South East Asia*. Also similarly to the child obesity plot *Oceania* has the largest interquartile range, largest range and highest median showing that obesity is also most prevalent in adults in *Oceania*.

Section 2 - Child vs Adult Health Comparison

Child vs Adult Obesity Prevalence

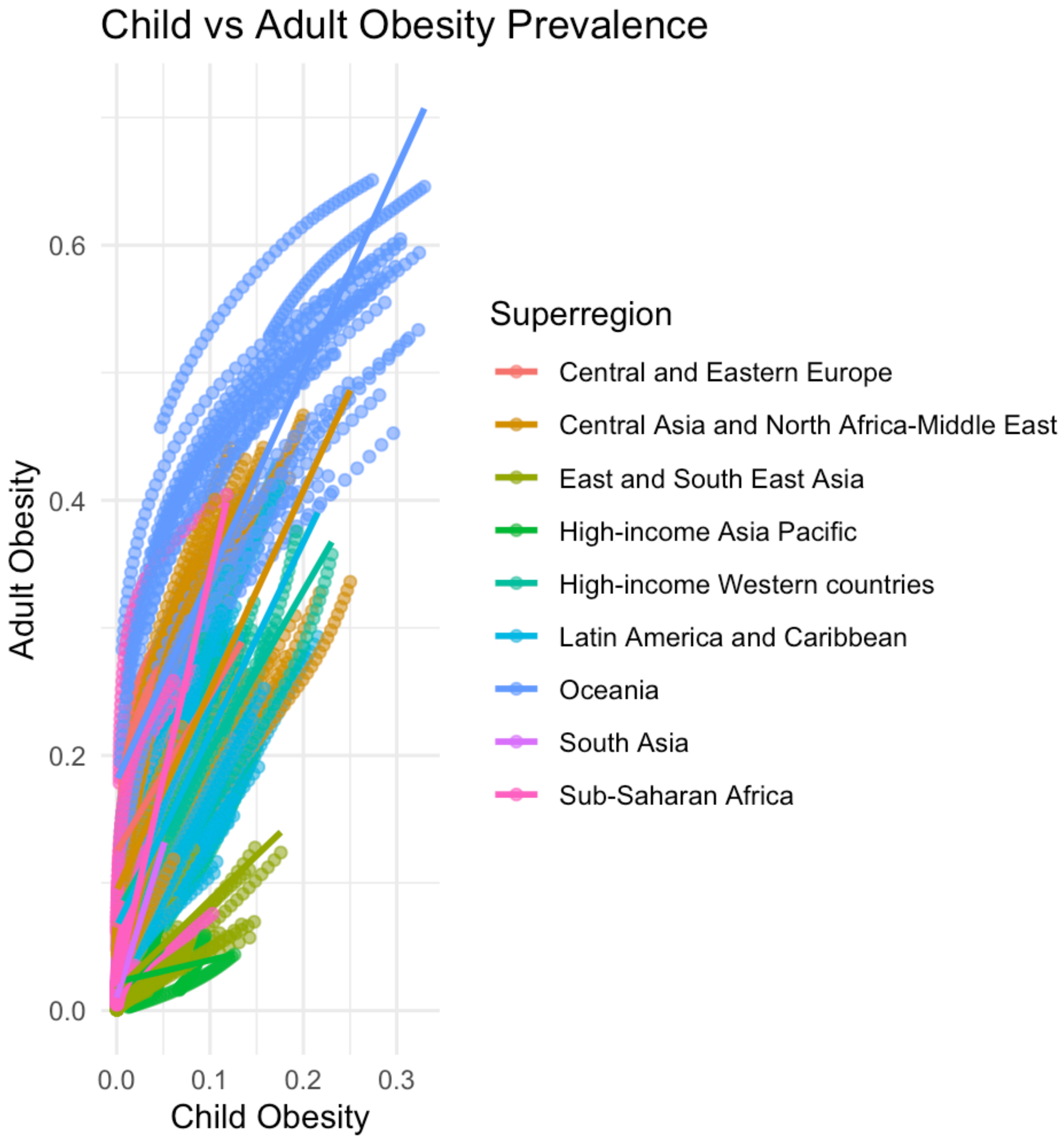

Figure 2.1

This plot shows that obesity tends to be more prevalent within adults than within children due to most of the superregions having relatively steep gradients on their regression lines. This can especially be seen within *Oceania* where obesity is over twice as prevalent in adults as it is in children

Child vs Adult Underweight Prevalence

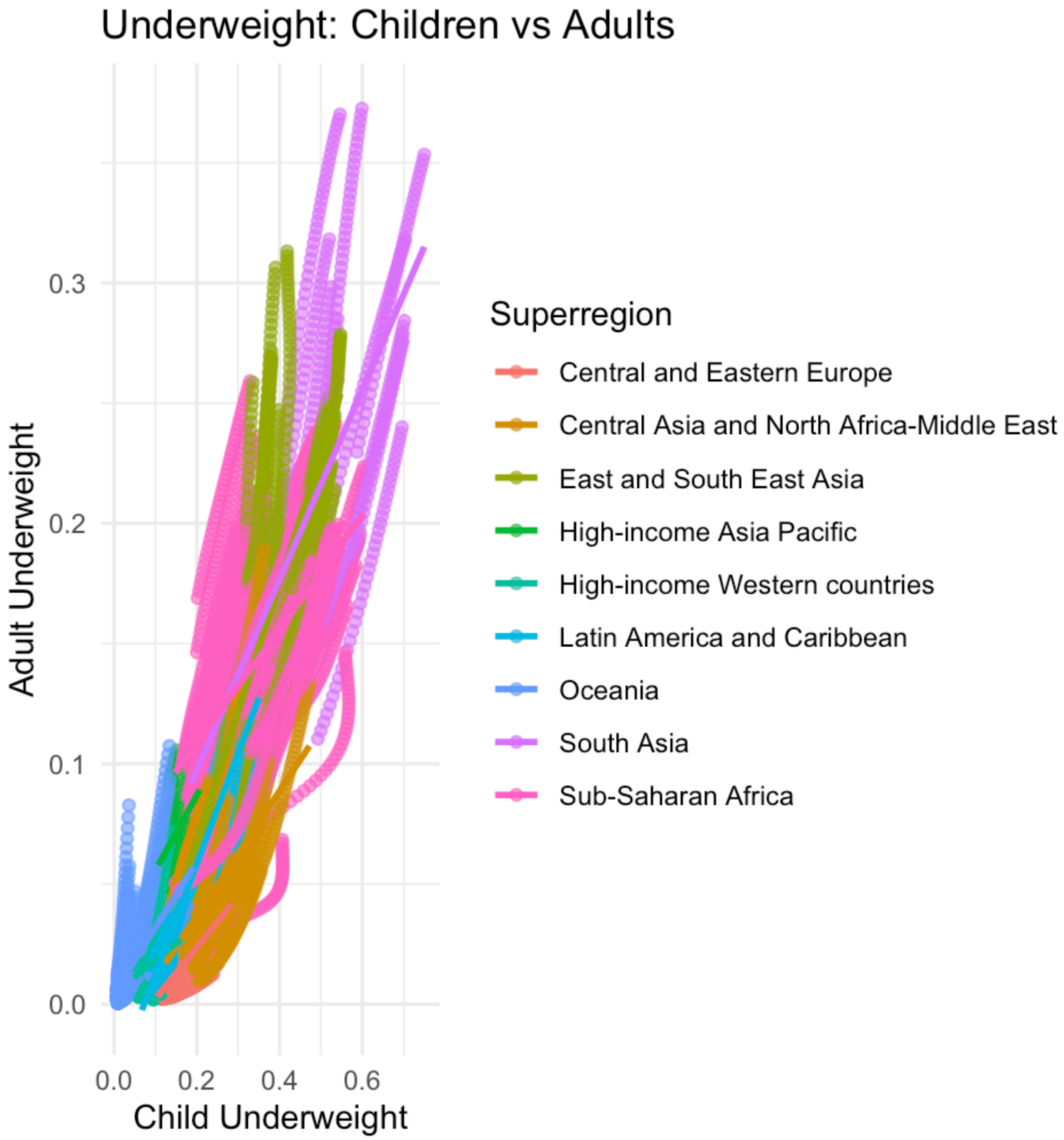

Figure 2.2

The underweight trends are the opposite to those of obesity where it is much more prevalent within children than it is within adults, with *South Asia* having almost double the prevalence of underweight children than that of adults.

Section 3 - Socioeconomic Factors and Health

Child Obesity vs GDP by Superregion

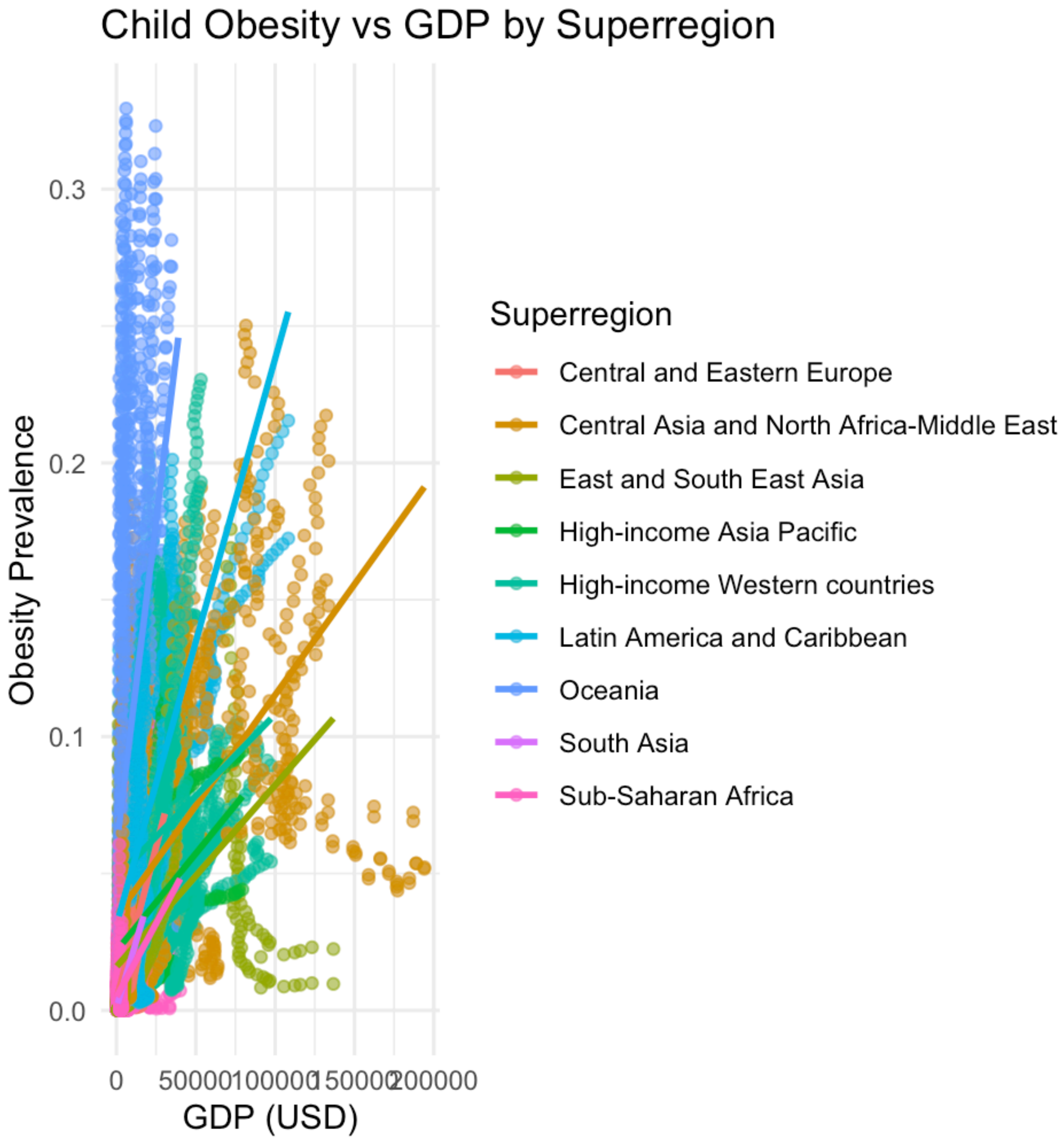

Figure 3.1

This plot shows that where obesity is more prevalent, the GDP tends to be lower and as GDP increases, obesity prevalence lessens within children.

Adult Obesity vs GDP by Superregion

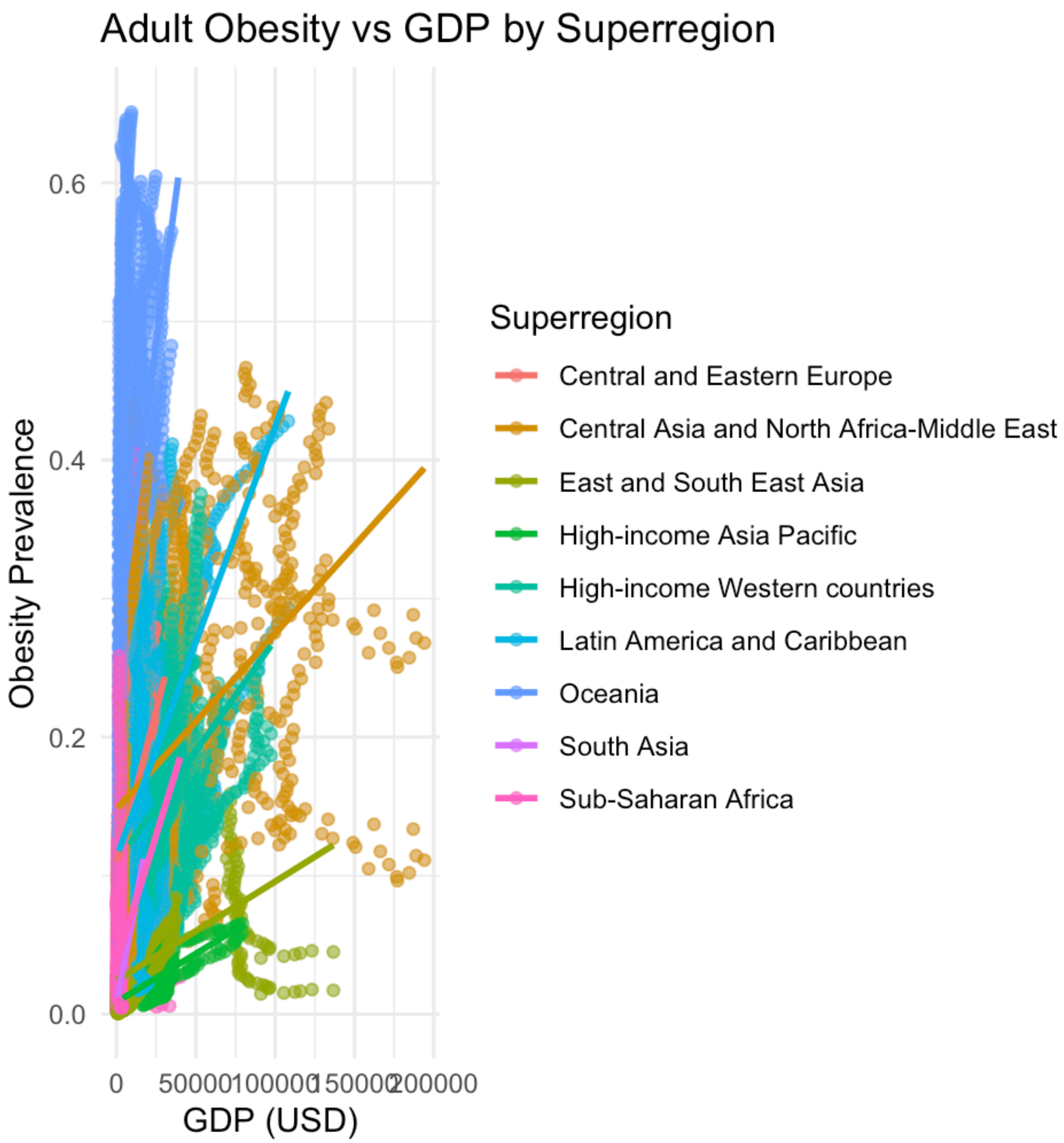

Figure 3.2

This plot shows a similar trend to that of the child obesity vs GDP plot in that GDP is lower in countries with higher obesity prevalence and where obesity is less prevalent in adults, GDP tends to be higher.

Years of Education vs Adult Obesity Prevalence

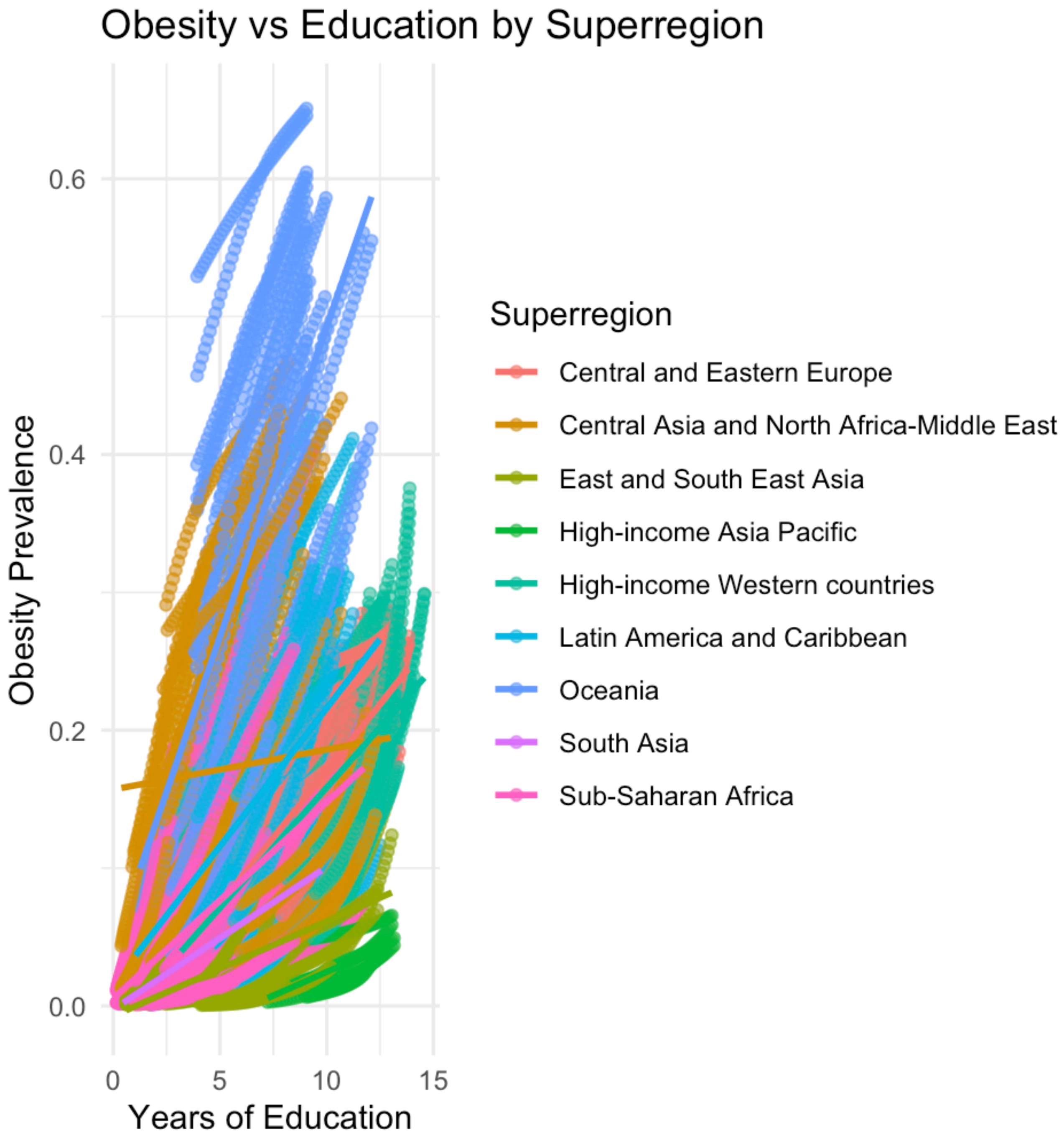

Figure 3.3

This plot shows that there is a positive correlation between obesity prevalence in adults and years of education where those who have more years in education have a lower obesity prevalence than those who have fewer.

Western Diet Score vs Obesity

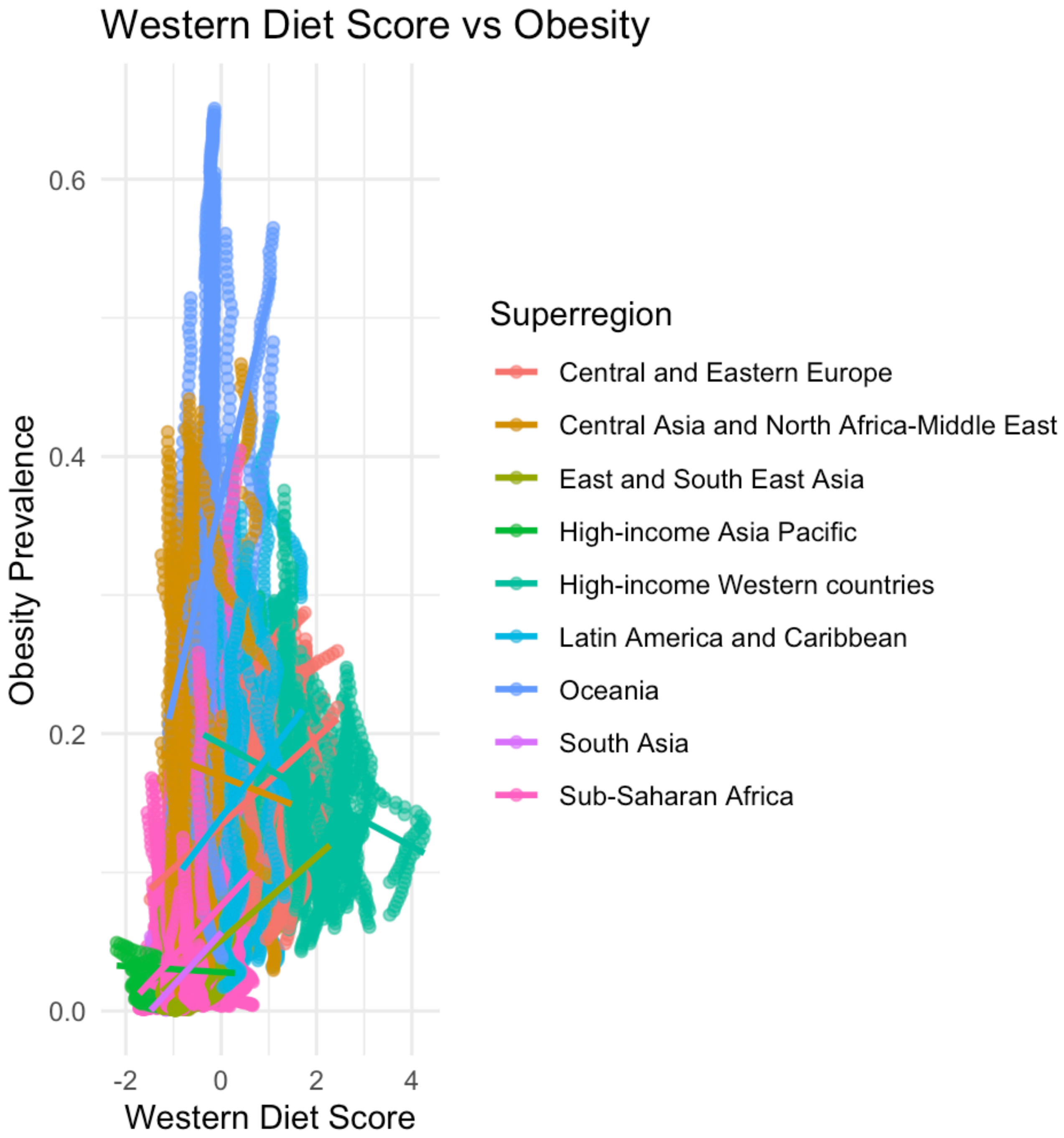

Figure 3.4

This plot shows that as the diet score gets further from 0, the obesity prevalence reduces where those with the highest and lowest western diet scores have the lowest obesity prevalence and those with a diet score of 0 have the highest.

Chapter 2

Section 4 - Correlations between Socioeconomic Factors

Overall Correlation between Years of Education and GDP

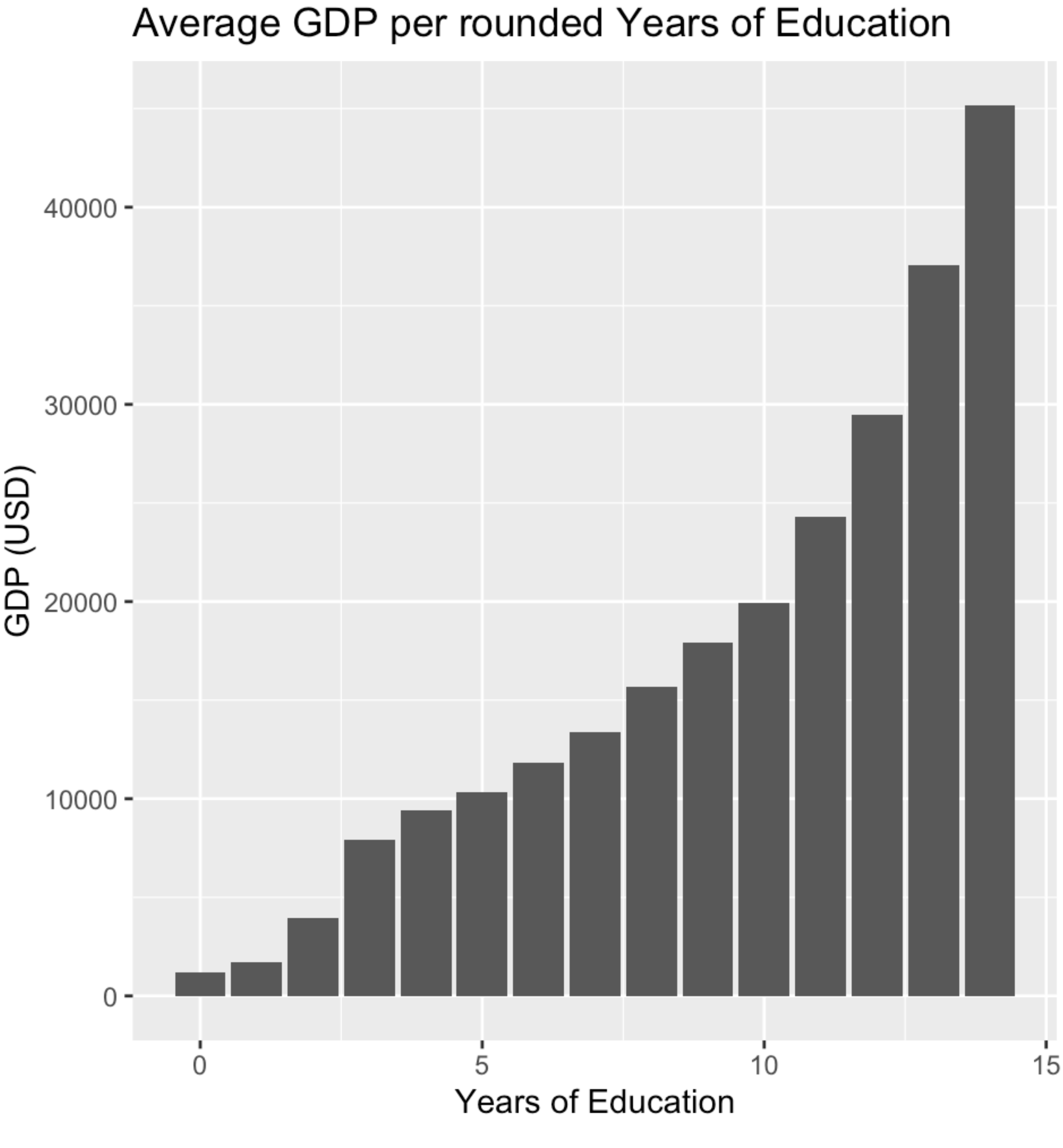

Figure 4.1

This analysis takes all countries across all years and shows that there is a very clear positive correlation that as years of education increase, as does GDP.

Spread of GDPs across the Rounded Years of Education

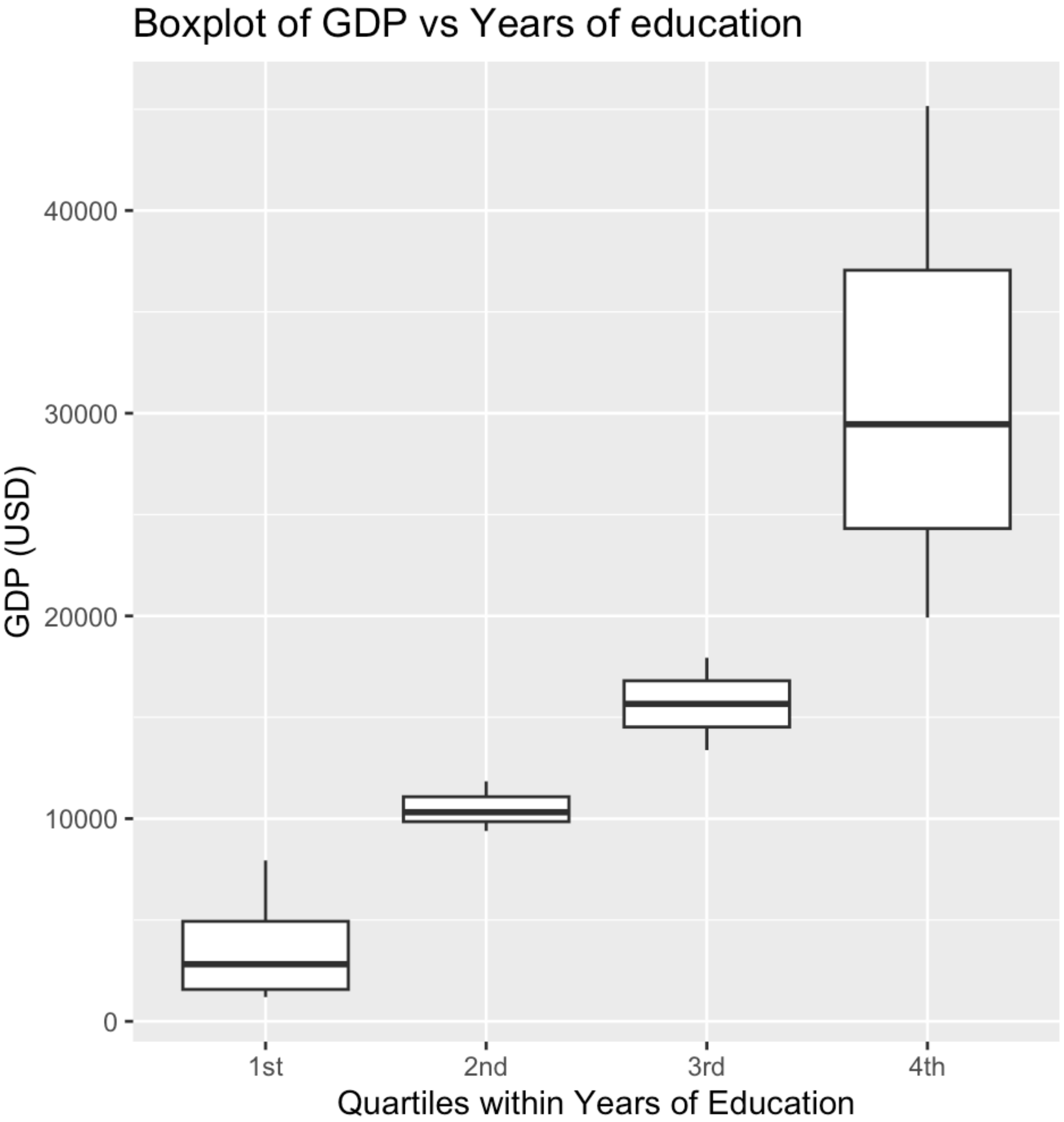

Figure 4.2

This plot shows again that as years of education increase, as does GDP. The highest median is within the 4th quartile but also has the largest interquartile range and largest range. Whereas the 2nd quartile shows the smallest range and interquartile range.

Relationship between Years of Education and GDP

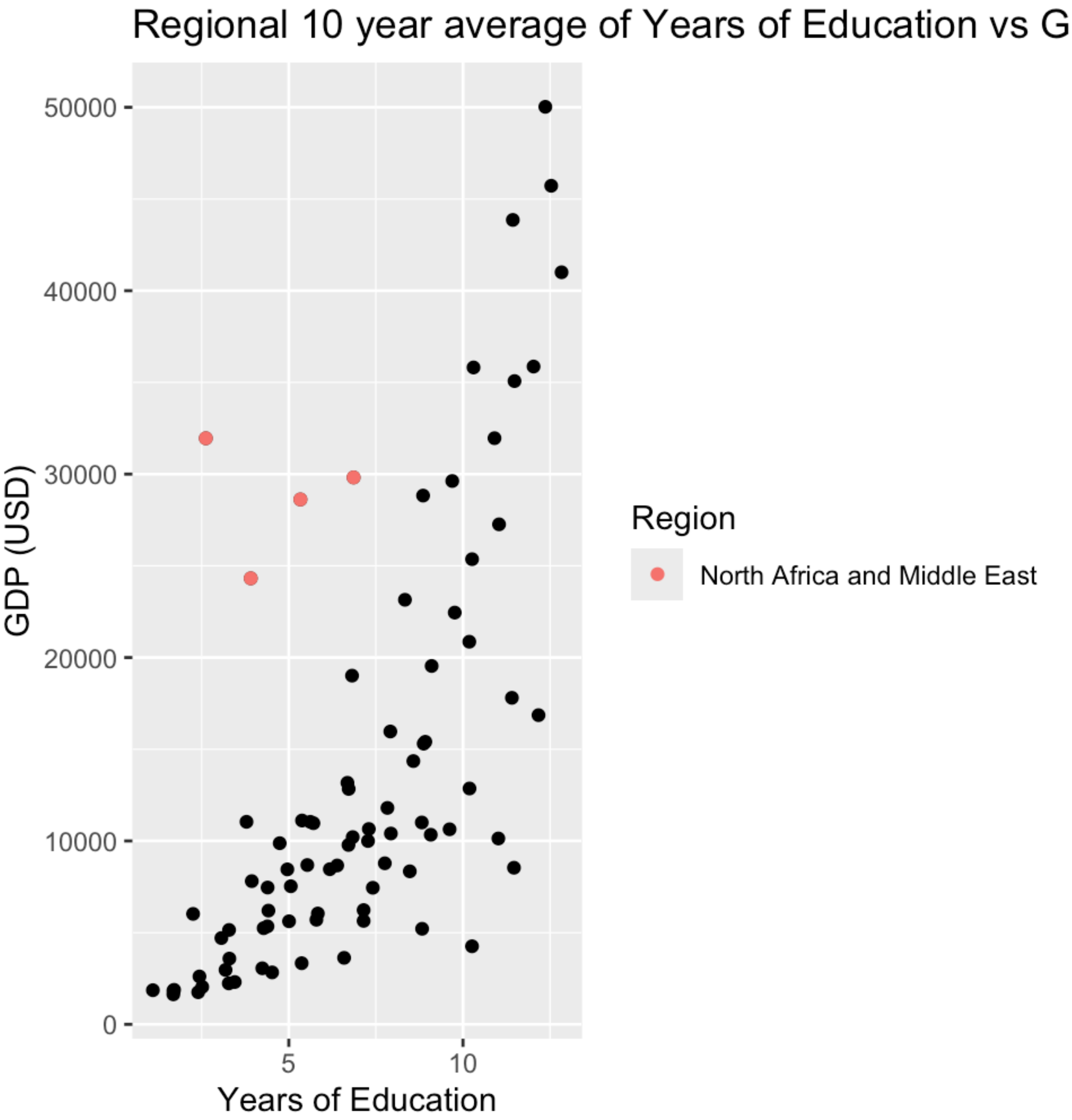

Figure 4.3

This scatter plot shows that there is still a positive correlation between years of education and GDP, each region has been averaged together every 10 years. It also shows an outlier in *North Africa and Middle East* where they show a high GDP but with much lower years of education, these points have been made orange for ease of comparison.

Per Country Percentage Change in Education and Percentage Change in GDP

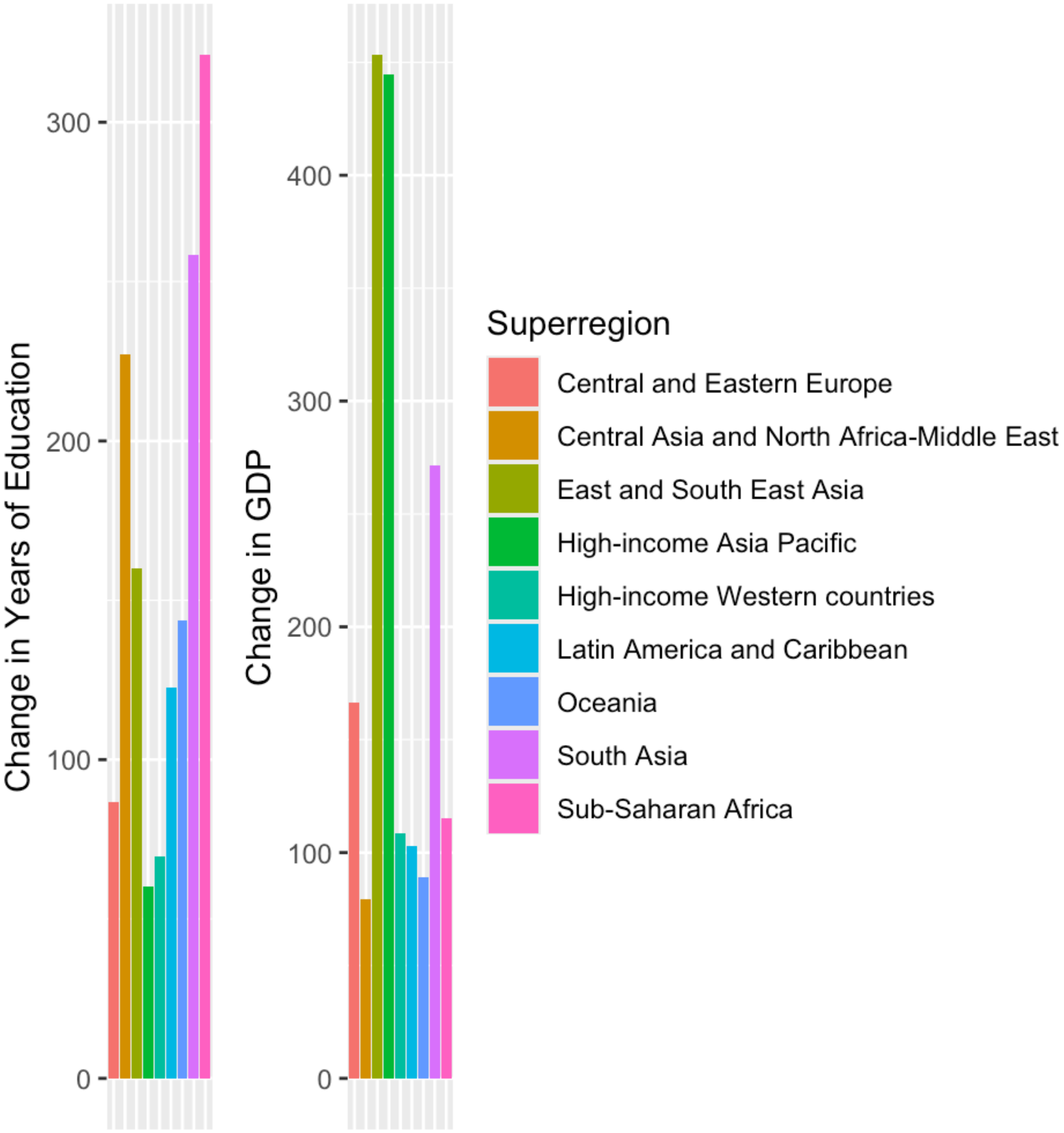

Figure 4.4

This plot shows that all superregions have had a positive percentage increase in both years of education and GDP with *East and South East Asia* showing the highest change in GDP of over and *Sub-Saharan Africa* showing the largest increase in years of education. This graph also shows that there is little to no correlation between the countries with the greatest percentage increase in years of education and those with the greatest percentage increase in GDP. The following scatter plot shows this lack of correlation and that there may be no cause and effect between the variables.

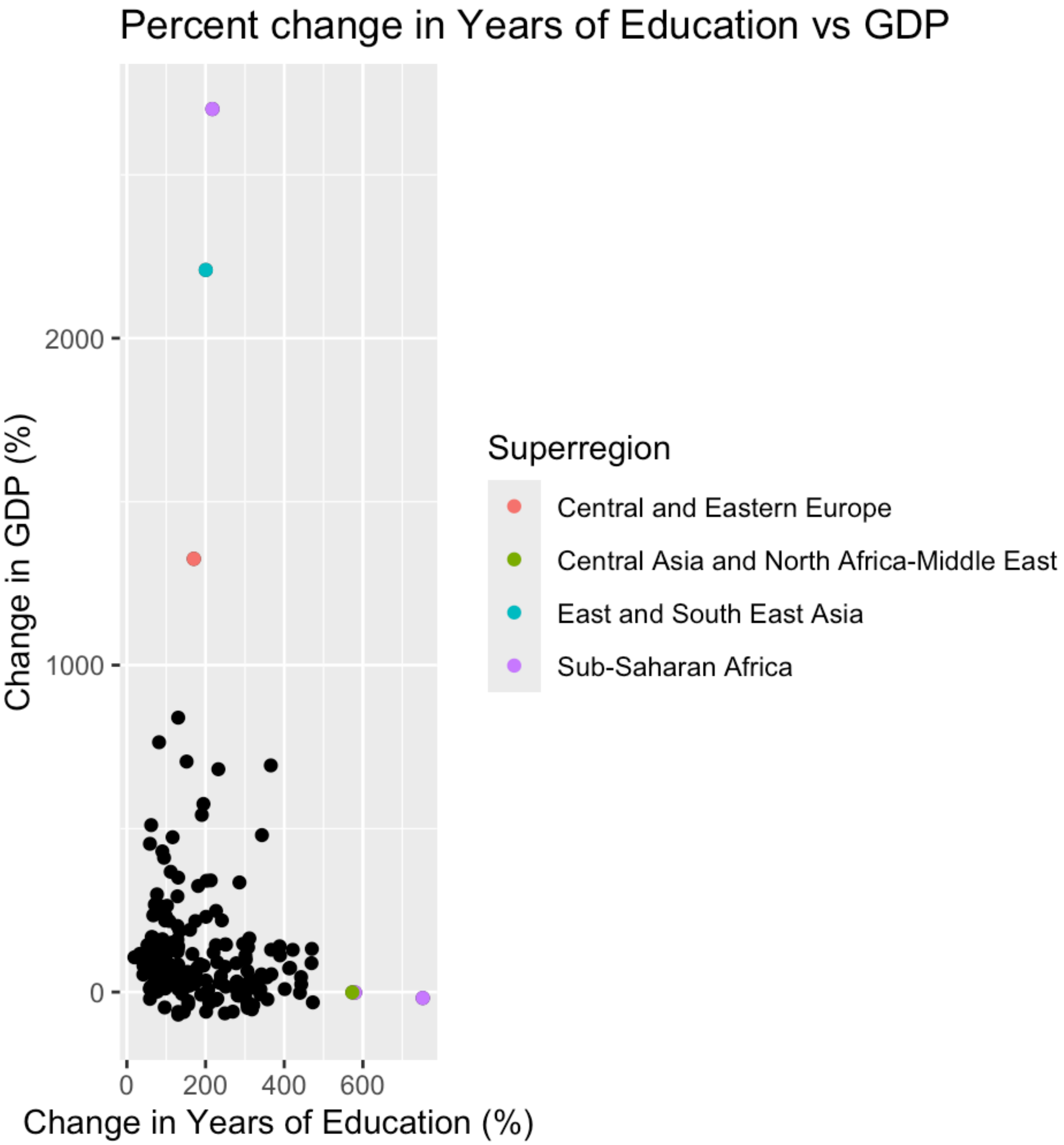

Figure 4.5

Box Plot per Quartile of Percentage change in Years of Education

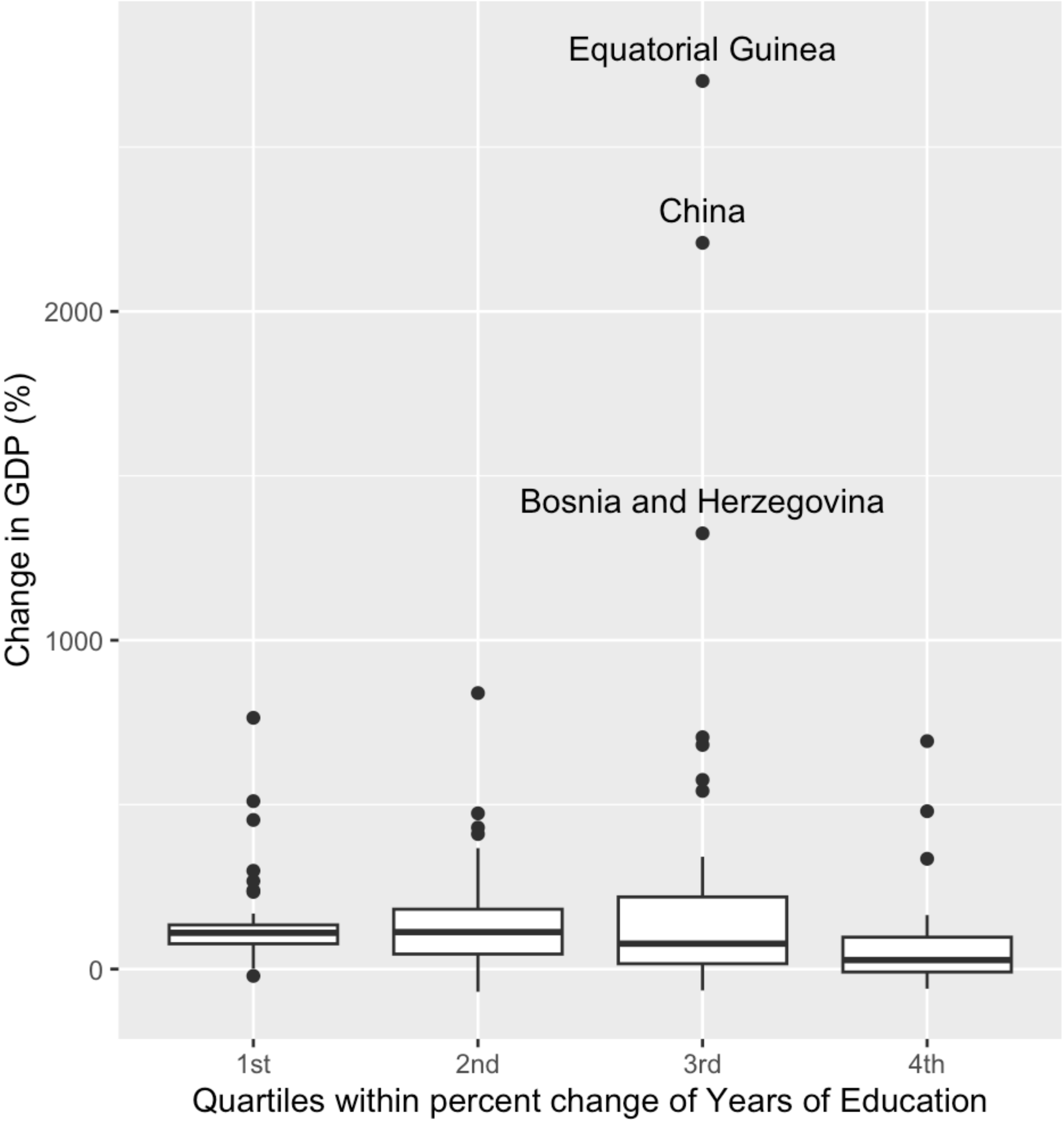

Figure 4.6

This box plot shows that the countries with the greatest percentage in GDP fall in the 3rd quartile rather than the 4th, similarly the box plots and ranges of both the 1st and 4th quartile and very similar. This further adds to the theory that there is no direct correlation between a percentage increase in years of education and of GDP.

Chapter 3

Section 5 - Blood Pressure Clustering

K-Means Clustering

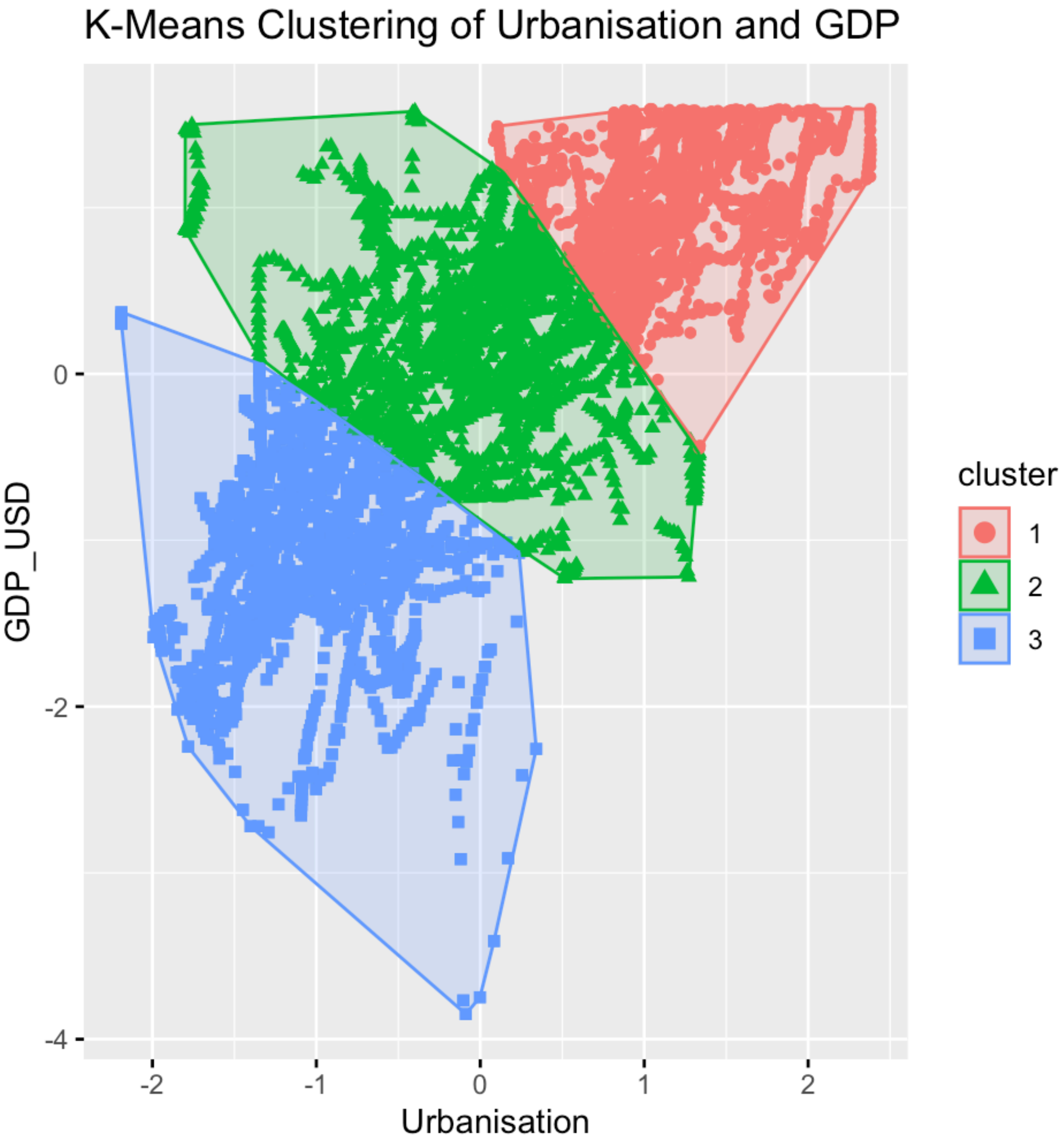

Figure 5.1

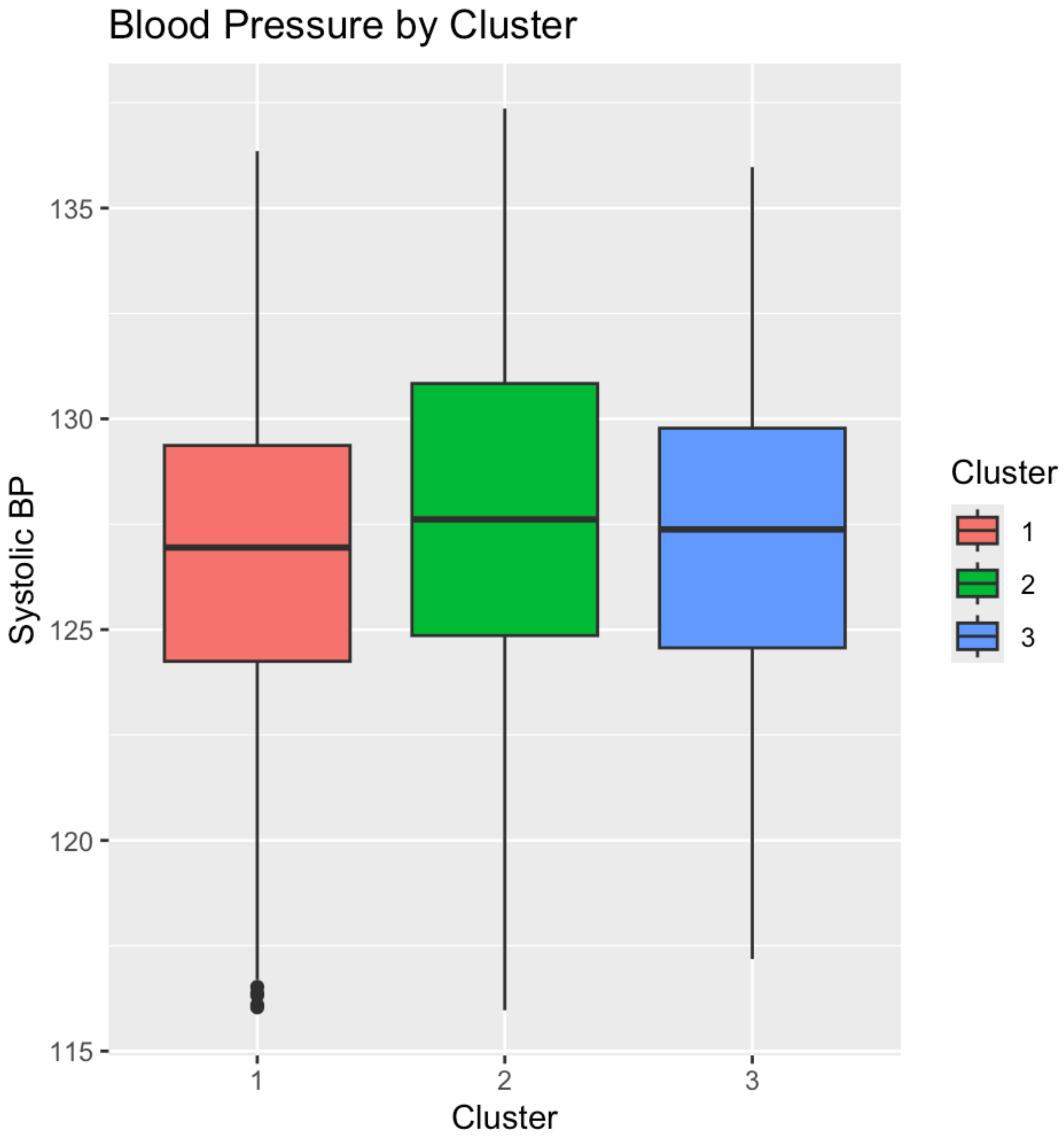

Figure 5.2

In *figure 5.1*, cluster 1 is low urbanisation and low GDP, cluster 2 is medium urbanisation and medium GDP, and cluster 3 is high urbanisation and high GDP. From this and *figure 5.2* we can see that urbanisation makes little to no difference on blood pressure as the box plots for all 3 clusters are similar.

Decision Tree Classification and Regression

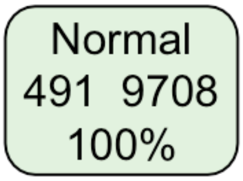

Figure 5.3

```
[1] "Random Forest RMSE: 3.36263822906071"
```

Figure 5.4

Chapter 4

Section 6 - BMI Trends

BMI Trends by Sex and Age Group

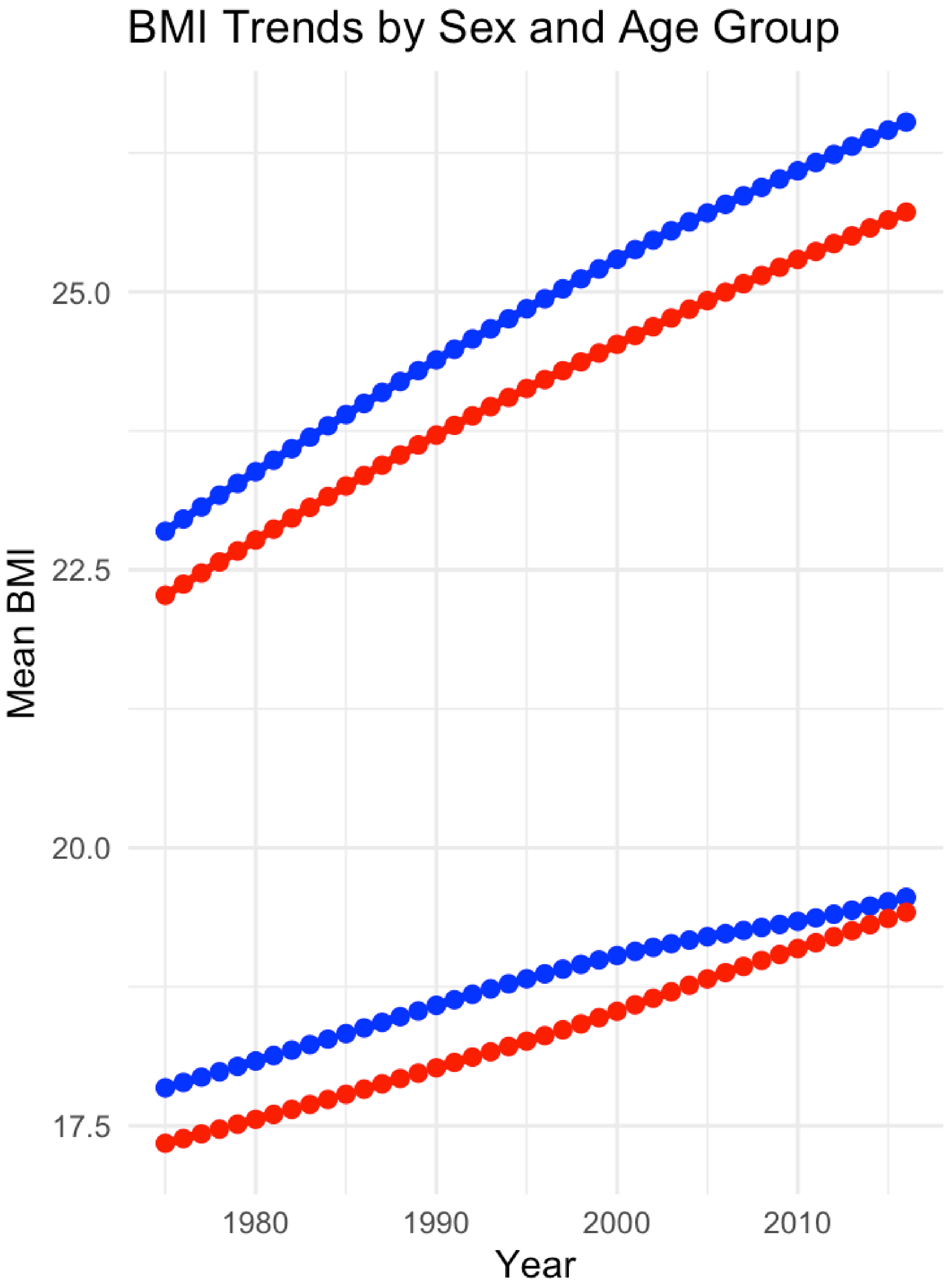

Figure 6.1

This plot shows that the mean BMI for both adults and children have been increasing over time, though adults at a faster rate than children. Similarly in both age groups, males overall have a higher BMI than females do.

Section 7 - Data Sampling

GDP per Year

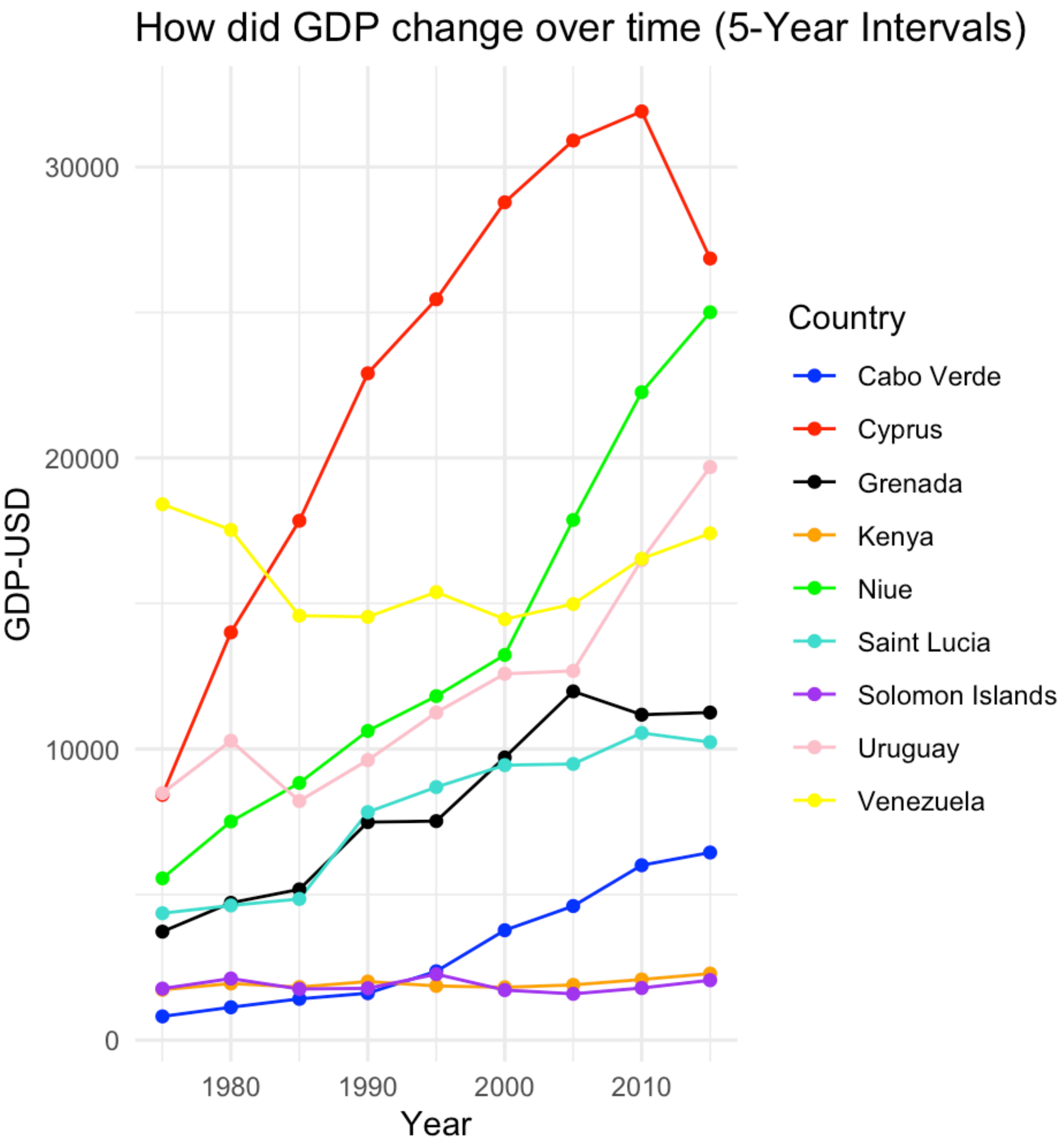

Figure 7.1

This plot shows that over time GDP has increased in the sampled countries, even though overall it has increased, some countries such as *Cyprus* and *Grenada* have begun to see GDP reduce more recently whereas countries such as *Niue* have seen the rate of change increase instead.

Mean BMI vs GDP

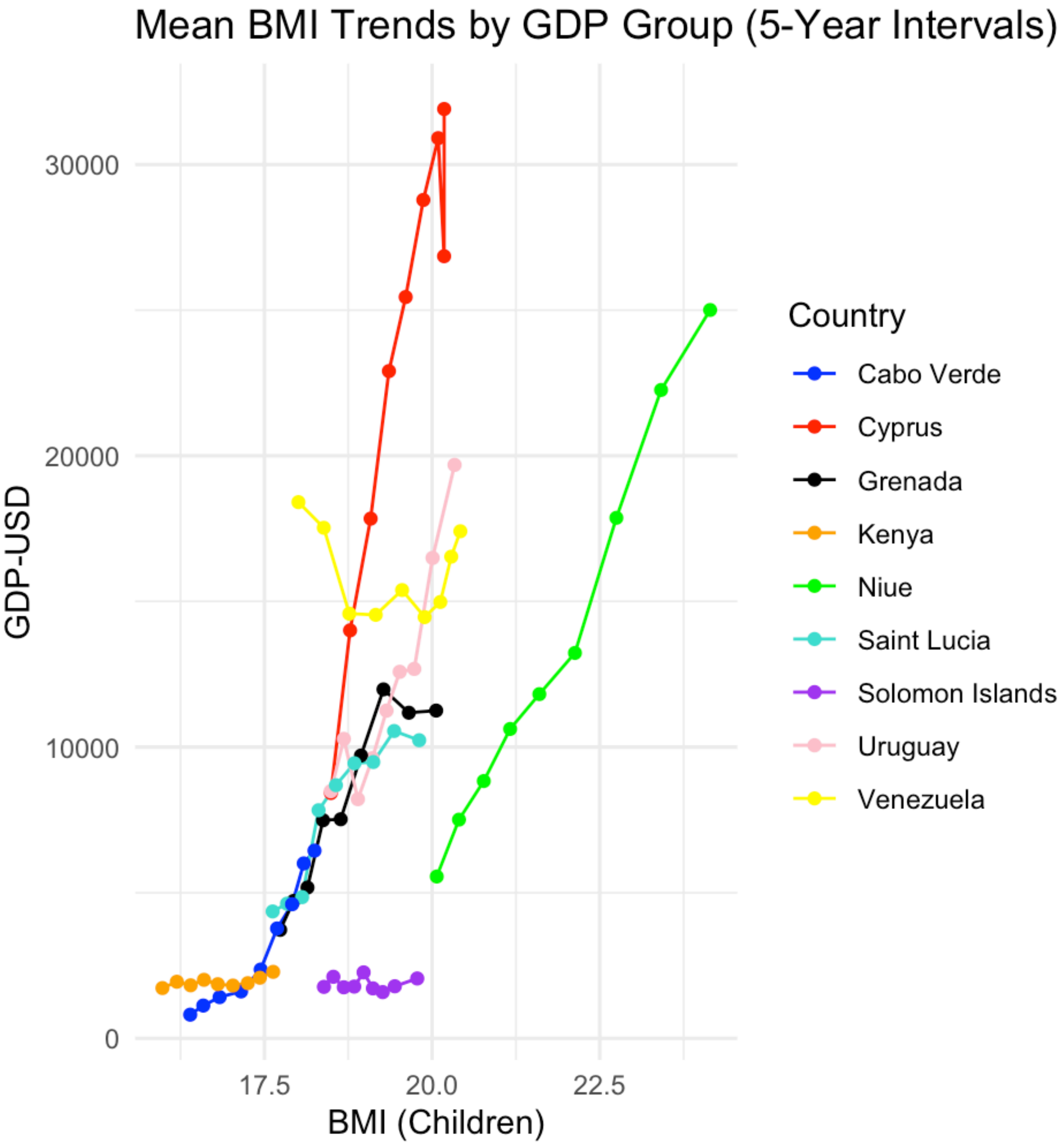
Mean BMI Trends by GDP Group (5-Year Intervals)
GDP-USD
30000
20000
10000
0
17.5
20.0
22.5
BMI (Children)
Country
Cabo Verde
Cyprus
Grenada
Kenya
Niue
Saint Lucia
Solomon Islands
Uruguay
Venezuela

Figure 7.2

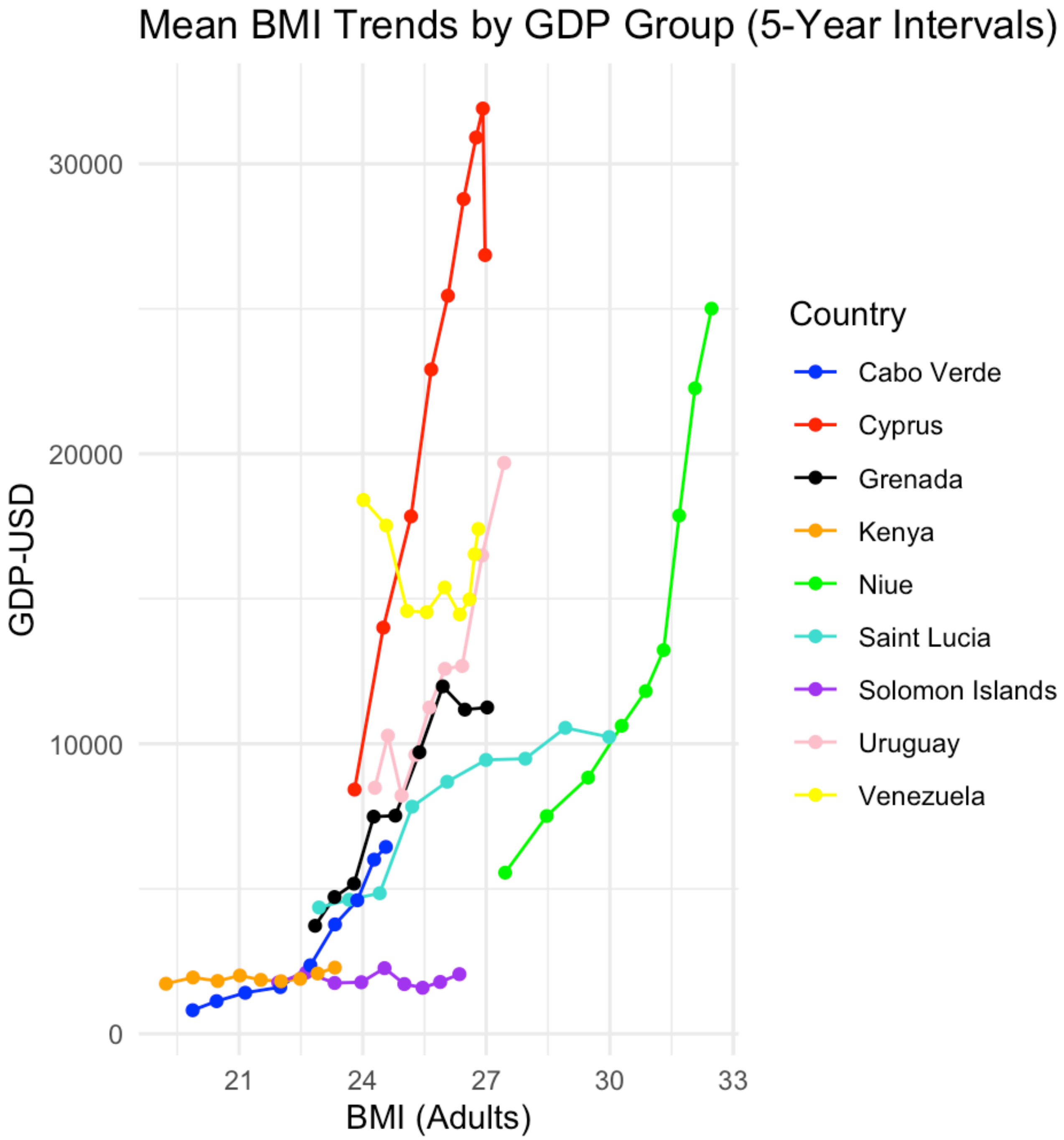
Mean BMI Trends by GDP Group (5-Year Intervals)
GDP-USD
30000
20000
10000
0
21
24
27
30
33
BMI (Adults)
Country
Cabo Verde
Cyprus
Grenada
Kenya
Niue
Saint Lucia
Solomon Islands
Uruguay
Venezuela

Figure 7.3

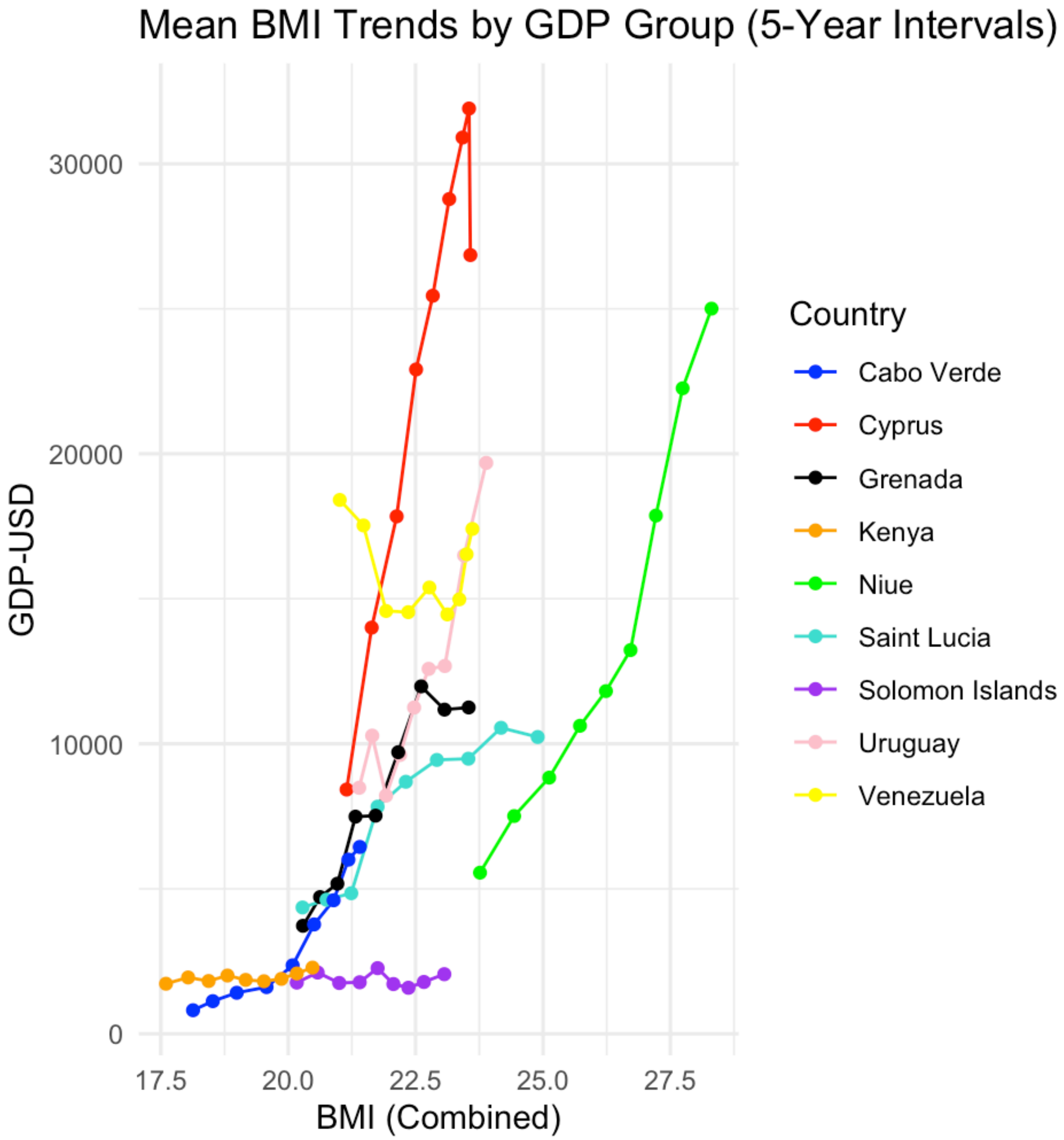

Figure 7.4

These graphs show the mean BMI trends of children, adults and then combined by GDP group in 5 year intervals, all 3 graphs are very similar in nature, for the most part in each country GDP and BMI have both increased over time. In some countries such as the *Solomon Islands* and *Kenya*, BMI has increased a lot whereas GDP has not as opposed to countries like *Cyprus* and *Niue* where there has been a large increase in GDP but not BMI.

Section 8 - Hypothesis Testing

Hypothesis Testing for Child BMI

```
    Pearson's product-moment correlation

data:  data$Mean_BMI_children and data$GDP_USD
t = 55.894, df = 16398, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3871031 0.4128155
sample estimates:
     cor 
0.400038 

    Two Sample t-test

data:  data$Mean_BMI_children and data$GDP_USD
t = -96.848, df = 33198, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13976.88 -13422.37
sample estimates:
  mean of x   mean of y 
   18.55241 13718.17499 
```

Figure 8.1

Hypothesis Testing for Adult BMI

```
    Pearson's product-moment correlation

data:  data$Mean_BMI_adults and data$GDP_USD
t = 48.785, df = 16398, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3425698 0.3693009
sample estimates:
      cor 
0.3560082 

    Two Sample t-test

data:  data$Mean_BMI_adults and data$GDP_USD
```

```
t = -96.807, df = 33198, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13970.96 -13416.45
sample estimates:
  mean of x   mean of y 
   24.46569 13718.17499 
```

Figure 8.2

Hypothesis Testing for Combined BMI

```
	Pearson's product-moment correlation

data:  data.com$Mean_BMI and data.com$GDP_USD
t = 39.871, df = 32798, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2046649 0.2253087
sample estimates:
      cor 
0.2150109 

	Two Sample t-test

data:  data.com$Mean_BMI and data.com$GDP_USD
t = -136.94, df = 66398, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13892.71 -13500.62
sample estimates:
  mean of x   mean of y 
   21.50905 13718.17499 
```

Figure 8.3

Conclusion

To conclude, our data has shown that the more educated an area is, the fewer NCD issues they tend to have. Although there is no direct correlation between percentage increase in GDP and percentage increase in years in education they both impact NCD variables when increased.

Countries with higher GDP and higher years in education tend to have lower mean BMI, less obesity and better health overall.

Appendix 1 - Data Cleaning

Section 1

```
# Remove data from the year 2016, due to lack of sufficient and effective
data
organisation_dataset <- organisation_dataset[organisation_dataset$Year !=
2016, ]
```

Section 2

```
N/A
```

Section 3

```
N/A
```

Section 4

```
sum(is.na(df)) # there are 6000 instances of missing values in the entire
data frame
sum(is.na(filter(df, Year == 2016))) #3600 instances of missing values for
rows from 2016
df <- filter(df, Year != 2016) #One year is not as significant as there
are 40 other years
na_counts <- df %>% summarise_all(~ sum(is.na(.))) #Per column which
values have NA, all from Diabetes

my_info <- df[-c(5:15, 19, 20)] %>%
  filter(Sex == 'Female') %>%
  rename(all_of(c(edu='Years_of_education',gdp='GDP_USD')))
```

Section 5

```
missing_per_column <- colSums(is.na(orgdata))
print(missing_per_column)  # Prints the count of missing values per column

# Remove an unnecessary column ("Diabetes_prevalence"), as it is not
required
orgdata1 <- orgdata[, !names(orgdata) %in% c("Diabetes_prevalence")]

# Remove rows where the row name is "2016" (possibly unnecessary data)
orgdata2 <- orgdata1[!(rownames(orgdata1) %in% "2016"), ]

# Detect and count duplicate rows in the dataset
sum(duplicated(orgdata))

# Function to remove outliers using the Interquartile Range (IQR) method
remove_outliers <- function(df, cols) {
  for (col in cols) {
    Q1 <- quantile(df[[col]], 0.25, na.rm = TRUE)
    Q3 <- quantile(df[[col]], 0.75, na.rm = TRUE)
    IQR <- Q3 - Q1

    # Define lower and upper bounds for outliers
    lower_bound <- Q1 - 1.5 * IQR
    upper_bound <- Q3 + 1.5 * IQR

    # Remove rows where values fall outside the defined bounds
    df <- df[df[[col]] >= lower_bound & df[[col]] <= upper_bound, ]
  }
  return(df)
}

# Select only numeric columns for outlier removal
numeric_cols <- names(orgdata)[sapply(orgdata, is.numeric)]

# Apply outlier removal function to the dataset
orgdata_cleaned <- remove_outliers(orgdata, numeric_cols)

# Compute correlation matrix for blood pressure and socioeconomic factors
cor_matrix <- cor(orgdata[, c("Systolic_blood_pressure",
"Years_of_education",
                              "Urbanisation", "Western_diet_score",
"GDP_USD")],
                  use="complete.obs")

# Visualize correlation matrix
ggcorrplot(cor_matrix, method="circle")

# Histogram: Distribution of key variables
orgdata_cleaned %>%
  select(Systolic_blood_pressure, Years_of_education, Urbanisation,
Western_diet_score, GDP_USD) %>%
  gather(variable, value) %>%
  ggplot(aes(x=value, fill=variable)) +
  geom_histogram(bins=30, alpha=0.6) +
  facet_wrap(~variable, scales="free") +
  theme_minimal() +
  labs(title="Distribution of Key Variables", x="Value", y="Count")
```

Section 6

```
#separating male and female
BMIm <- subset(data, Sex != "Female")
BMIf <- subset(data, Sex != "Male")
#creating a variable of the combined sets
Mean_BMI <- rbind(BMIm,BMIf)
#averaging the data for every year and attributing them to a variable
BMIafe <- aggregate(Mean_BMI_adults ~ Year, data=BMIf, FUN = mean)
BMIama <- aggregate(Mean_BMI_adults ~ Year, data=BMIm, FUN = mean)
BMIcfe <- aggregate(Mean_BMI_children ~ Year, data=BMIf, FUN = mean)
BMIcma <- aggregate(Mean_BMI_children ~ Year, data=BMIm, FUN = mean)
#creating classifiers for men, women, boys and girls so that I can clearly
separate them in the graph
BMIafe$Group <- "Men"
BMIama$Group <- "Women"
BMIcfe$Group <- "Boys"
BMIcma$Group <- "Girls"
#binding the variables together to present the mean BMI for adults and
children
Mean_BMIa <- rbind(BMIafe,BMIama)
Mean_BMIc <- rbind(BMIcfe,BMIcma)
```

Section 7

```
#filter the GDP data and isolate it
data_gdp<-data %>% filter(!is.na(GDP_USD))
#turn the data into quantiles
quantiles <- quantile(data$GDP_USD, probs = c(0.33,0.66),na.rm = TRUE)
#group each country into one of the 3 quantiles
data_gdp <- data.com %>%
  group_by(Country) %>%
  summarise(Avg_GDP=mean(GDP_USD, na.rm = TRUE)) %>%
  mutate(GDP_Level = case_when(
    Avg_GDP <= quantiles[1] ~ "Low",
    Avg_GDP > quantiles[1] & Avg_GDP <= quantiles[2] ~ "Average",
    Avg_GDP > quantiles[2] ~ "High"
  ))
# Randomly select 3 countries from each GDP level
set.seed(42)
selected_countries <- data_gdp %>%
  group_by(GDP_Level) %>%
  sample_n(3) %>%
  pull(Country)
```

Section 8

```
N/A
```

Appendix 2 - Figures

Section 1

Figure 1.1

```
# Plot Mean BMI Trends for Children
ggplot(organisation_dataset, aes(x = Year, y = Mean_BMI_children, color =
Superregion)) +
  geom_line(stat = "summary", fun = "mean") +  # Aggregates mean BMI for
each year
  labs(title = "Mean BMI Trends for Children by Superregion", y = "Mean
BMI", x = "Year") +
  theme_minimal()
```

Figure 1.1

```
# Plot Mean BMI Trends for Adults
ggplot(organisation_dataset, aes(x = Year, y = Mean_BMI_adults, color =
Superregion)) +
  geom_line(stat = "summary", fun = "mean") +
  labs(title = "Mean BMI Trends for Adults by Superregion", y = "Mean
BMI", x = "Year") +
  theme_minimal()
```

Figure 1.1

```
# Boxplot of Child Obesity Prevalence by Superregion
ggplot(organisation_dataset, aes(x = Superregion, y =
Prevalence_obesity_children, fill = Superregion)) +
  geom_boxplot() +  # Shows distribution of obesity prevalence in children
  labs(title = "Obesity Prevalence for Children Across Superregions", y =
"Obesity Prevalence", x = "Superregion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Figure 1.1

```
# Boxplot of Adult Obesity Prevalence by Superregion
ggplot(organisation_dataset, aes(x = Superregion, y =
Prevalence_obesity_adults, fill = Superregion)) +
  geom_boxplot() +
  labs(title = "Obesity Prevalence for Adults Across Superregions", y =
"Obesity Prevalence", x = "Superregion") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Section 2

Figure 2.1

```
ggplot(organisation_dataset, aes(x = Prevalence_obesity_children, y =
Prevalence_obesity_adults, color = Superregion)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Child vs Adult Obesity Prevalence", y = "Adult Obesity", x
= "Child Obesity") +
  theme_minimal()
```

Figure 2.2

```
# Scatter Plot: Underweight Prevalence - Children vs Adults
ggplot(organisation_dataset, aes(x = Prevalence_underweight_children, y =
Prevalence_underweight_adults, color = Superregion)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Underweight: Children vs Adults", y = "Adult Underweight",
x = "Child Underweight") +
  theme_minimal()
```

Section 3

Figure 3.1

```
# Scatter Plot: Child Obesity vs GDP by Superregion
ggplot(organisation_dataset, aes(x = GDP_USD, y =
Prevalence_obesity_children, color = Superregion)) +
  geom_point(alpha = 0.6) +  # Alpha for transparency
  geom_smooth(method = "lm", se = FALSE) +  # Adds linear regression line
  labs(title = "Child Obesity vs GDP by Superregion", y = "Obesity
Prevalence", x = "GDP (USD)") +
  theme_minimal()
```

Figure 3.2

```
# Scatter Plot: Adult Obesity vs GDP by Superregion
ggplot(organisation_dataset, aes(x = GDP_USD, y =
Prevalence_obesity_adults, color = Superregion)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Adult Obesity vs GDP by Superregion", y = "Obesity
Prevalence", x = "GDP (USD)") +
  theme_minimal()
```

Figure 3.3

```
# Scatter Plot: Years of Education vs. Obesity Prevalence (Adults)
ggplot(organisation_dataset, aes(x = Years_of_education, y =
Prevalence_obesity_adults, color = Superregion)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Obesity vs Education by Superregion", y = "Obesity
Prevalence", x = "Years of Education") +
  theme_minimal()
```

Figure 3.4

```
# Scatter Plot: Western Diet Score vs Obesity Prevalence (Adults)
ggplot(organisation_dataset, aes(x = Western_diet_score, y =
Prevalence_obesity_adults, color = Superregion)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Western Diet Score vs Obesity", y = "Obesity Prevalence",
x = "Western Diet Score") +
  theme_minimal()
```

Section 4

Figure 4.1

```
# A plot showing the overall correlation between Years of education and
GDP
# This analysis takes every country across all years
round_edu <- my_info %>%
  filter(edu <= 14) %>% #There are 11 instances (two countries) where the
years of education > 14, not a significant amount in a sample from 8400
  mutate(edu = round(edu, 0)) %>% #Rounded to the nearest integer for
simplification of graphs
  group_by(edu) %>%
  summarise(across(gdp, mean)) %>% #calculates the average GDP for every
grouping of Years of education
  rename_with(.cols = 1, ~"edu")

ggplot(round_edu, aes(edu, gdp)) +
  geom_col() +
  labs( x = 'Years of Education',
        y = 'GDP (USD)',
        title = 'Average GDP per rounded Years of Education')
```

Figure 4.2

```
# A plot that shows the spread of GDPS across the rounded Years of
education
# A boxplot is generated for each Quartile, which are generated from the
overall data (not the rounded data set)
# To find the specific ranges, run the code commented out below
### summary(my_info$edu) ###
round_edu['quartile'] <- sapply(round_edu$edu,
                                assign_quartile, # uses the
assign_quartile function on each row
                                summary = summary(my_info$edu)) #here the
quartile is calculated on every instance from the dataset
ggplot(round_edu, aes(quartile, gdp)) +
  geom_boxplot() +
  labs( x = 'Quartiles within Years of Education',
        y = 'GDP (USD)',
        title = 'Boxplot of GDP vs Years of education')
```

Figure 4.3

```
# A scatter plot showing the relationship between years of education and
gdp
# For the sake of viewing, each region was averaged together every 10
years to create 84 points
my_info['time_frame'] <- sapply(my_info$Year,
                                assign_time_period) # Using the
assign_time_period function on each row
region_10yr_average <- my_info %>%
  group_by(Region, time_frame) %>%
  summarise(across(c(gdp, edu), mean)) #Finds the average of each Region
every 10 years

ggplot(region_10yr_average, aes(edu, gdp)) +
  geom_point() +
  geom_point(data = filter(region_10yr_average, # Used to identify the
bigger outliers
                           gdp > 20000 & edu < 7.5),
             aes(color = Region)) +
  labs(x = 'Years of Education',
       y = 'GDP (USD)',
       title = 'Regional 10 year average of Years of Education vs GDP')
```

Figure 4.4

```
# On a per country basis, showing the percent change in education and
percent change in gdp
# This analysis shows that there might not be a direct cause and effect
relationship

#These two data frames take every instance from the years 1975 and 2015
respectively
early <- filter(my_info, Year == 1975)
late <- filter(my_info, Year == 2015)

differences <- data.frame(unique(my_info$Country),
                          early$Superregion, # The first two lines are to
access the country and superregion values
                          early$edu,
                          late$edu,
                          early$gdp,
                          late$gdp) %>%
  rename(Superregion = 'early.Superregion')
```

```
differences['dif_edu'] <- differences$late.edu - differences$early.edu #
computes the difference between 1975 education and 2015
differences['dif_gdp'] <- differences$late.gdp - differences$early.gdp #
ditto but for gdp
differences['edu_percent'] <- percent_change(differences$late.edu,
differences$early.edu) # computes the percent change in education
differences['gdp_percent']<- percent_change(differences$late.gdp,
differences$early.gdp) # ditto for gdp

region_dif <- differences %>%
  group_by(Superregion) %>%
  summarise(across(c(edu_percent, gdp_percent), mean)) # Calculates the
average percent change for both edu and gdp grouped by superregion

#The code below is a way to remove the x axis from plots where it can
become cluttered
remove_axis <- theme(axis.title.x = element_blank(), axis.ticks.x =
element_blank(), axis.text.x = element_blank())

region_edu <- ggplot(region_dif, aes(Superregion, edu_percent)) + # A bar
plot of each Superregion's % change in education
  geom_bar(stat = 'identity',
           aes(fill = Superregion)) +
  labs(y = 'Change in Years of Education') +
  remove_axis

region_gdp <- ggplot(region_dif, aes(Superregion, gdp_percent)) + # a bar
plot of each Superregions % change in gdp
  geom_bar(stat = 'identity',
           aes(fill = Superregion)) +
  labs(y = 'Change in GDP') +
  remove_axis

ggarrange(region_edu, region_gdp,
          common.legend = TRUE, legend = 'right')
```

Figure 4.5

```
# A simple scatter plot showing edu vs gdp percent change
ggplot(differences, aes(edu_percent, gdp_percent)) + geom_point() +
  geom_point(data = filter(differences, edu_percent > 500 | gdp_percent >
1000),
             aes(color = Superregion)) +  #Identifying some outliers
  labs(x = 'Change in Years of Education (%)',
       y = 'Change in GDP (%)',
       title = 'Percent change in Years of Education vs GDP ')



summary_edu_percent <- summary(differences$edu_percent) #Getting the
quartiles of percent change in edu
differences['edu_cent_quat'] <- sapply(differences$edu_percent,
                                       assign_quartile,
                                       summary = summary_edu_percent)
```

Figure 4.6

```
#Box plot per quartile of percent change in years of education
ggplot(differences, aes(edu_cent_quat,gdp_percent)) + geom_boxplot() +
  geom_text(data = filter(differences, gdp_percent > 1000),
            aes(label = unique.my_info.Country.), nudge_y = 100) +
  labs(x = 'Quartiles within percent change of Years of Education',
       y = 'Change in GDP (%)')
```

Section 5

Figure 5.1

```
# Perform K-Means clustering on Urbanisation and GDP
cluster_data <- orgdata_cleaned %>%
  select(Urbanisation, GDP_USD) %>%
  scale()

# Determine optimal number of clusters using the Elbow Method
fviz_nbclust(cluster_data, kmeans, method = "wss")

# Apply K-Means clustering with 3 clusters
set.seed(123)
kmeans_result <- kmeans(cluster_data, centers = 3, nstart = 25)

# Add cluster labels to the dataset
orgdata_cleaned$Cluster <- as.factor(kmeans_result$cluster)
```

```
# Visualize the clustering results
fviz_cluster(kmeans_result, data = cluster_data, geom = "point", ellipse =
TRUE) +
  labs(title = "K-Means Clustering of Urbanisation and GDP")
```

Figure 5.2

```
# The 'aggregate' function groups the data by 'Cluster'
# and computes the mean Systolic Blood Pressure for each group.
aggregate(orgdata_cleaned$Systolic_blood_pressure,
          by = list(orgdata_cleaned$Cluster),  # Group by Cluster
          FUN = mean)  # Calculate the mean BP for each cluster



# Create a boxplot to compare the distribution of Systolic Blood Pressure
# across different Clusters
ggplot(orgdata_cleaned, aes(x = Cluster,  # X-axis represents clusters
                            y = Systolic_blood_pressure,  # Y-axis
represents BP
                            fill = Cluster)) +  # Fill color based on
cluster group
  geom_boxplot() +  # Generate a boxplot
  labs(title = "Blood Pressure by Cluster",  # Set the title of the plot
       x = "Cluster",  # Label for X-axis
       y = "Systolic BP")  # Label for Y-axis
```

Figure 5.3

```
# Categorize blood pressure into Low, Normal, and High
orgdata_cleaned$BP_Category <-
cut(orgdata_cleaned$Systolic_blood_pressure,
                                   breaks = c(-Inf, 120, 140, Inf),
                                   labels = c("Low", "Normal", "High"))

# Train a Decision Tree model to classify blood pressure categories
bp_model <- rpart(BP_Category ~ Urbanisation + GDP_USD + Cluster, data =
orgdata_cleaned, method = "class")

# Visualize the Decision Tree
rpart.plot(bp_model, type = 4, extra = 101)
```

Figure 5.4

```
# Train a linear regression model to predict systolic blood pressure
model <- lm(Systolic_blood_pressure ~ Urbanisation + GDP_USD + Cluster,
data = orgdata_cleaned)
summary(model)

# Train a Random Forest model for regression
rf_model <- randomForest(Systolic_blood_pressure ~ Urbanisation + GDP_USD
+ Cluster,
                         data = orgdata_cleaned, ntree = 100)

# Predict blood pressure using Random Forest
rf_predictions <- predict(rf_model, newdata = orgdata_cleaned)

# Compute RMSE for Random Forest regression model
rf_rmse <- RMSE(rf_predictions, orgdata_cleaned$Systolic_blood_pressure)
print(paste("Random Forest RMSE:", rf_rmse))
```

Section 6

Figure 6.1

```
# Plot the BMI trends for both Adults and Children, separated by Sex
ggplot(data2, aes(x=Year, y=Mean_BMI, color=Sex, linetype=Age_Group,
group=interaction(Sex, Age_Group))) +
  geom_line(size=1.2) +
  geom_point(size=2) +
  labs(title="BMI Trends by Sex and Age Group",
       x="Year", y="Mean BMI") +
  theme_minimal() +
  scale_color_manual(values=c("blue", "red")) +  # Male = Blue, Female =
Red
  scale_linetype_manual(values=c("solid", "dashed"))  # Children = Solid,
Adults = Dashed
```

Section 7

Figure 7.1

```
# Plot the data for GDP per year
ggplot(data5, aes(x=Year, y=GDP_USD, color=Country, group=Country)) +
  geom_line() +
  geom_point() +
  labs(title="How did GDP change over time (5-Year Intervals)",
       x="Year", y="GDP-USD") +
  scale_color_manual(values=c("blue",
"red","black","orange","green","turquoise","purple","pink","yellow")) +
  theme_minimal()
```

Figure 7.2

```
# Plot the data to see the mean BMI against the GDP (for Children)
ggplot(data3, aes(y=GDP_USD, x=Mean_BMI, color=Country, group=Country)) +
  geom_line() +
  geom_point() +
  labs(title="Mean BMI Trends by GDP Group (5-Year Intervals)",
       x="BMI (Children)", y="GDP-USD") +
  scale_color_manual(values=c("blue",
"red","black","orange","green","turquoise","purple","pink","yellow")) +
  theme_minimal()
```

Figure 7.3

```
# Plot the data to see the mean BMI against the GDP (for Adults)
ggplot(data4, aes(y=GDP_USD, x=Mean_BMI, color=Country, group=Country)) +
  geom_line() +
  geom_point() +
  labs(title="Mean BMI Trends by GDP Group (5-Year Intervals)",
       x="BMI (Adults)", y="GDP-USD") +
  scale_color_manual(values=c("blue",
"red","black","orange","green","turquoise","purple","pink","yellow")) +
  theme_minimal()
```

Figure 7.4

```
# Plot the data to see the mean BMI against the GDP (combined)
ggplot(data5, aes(y=GDP_USD, x=Mean_BMI, color=Country, group=Country)) +
  geom_line() +
  geom_point() +
  labs(title="Mean BMI Trends by GDP Group (5-Year Intervals)",
       x="BMI (Combined)", y="GDP-USD") +
  scale_color_manual(values=c("blue",
"red","black","orange","green","turquoise","purple","pink","yellow")) +
  theme_minimal()
```

Section 8

Figure 8.1

```
#hypothesis testing for child BMI
cor.test(data$Mean_BMI_children, data$GDP_USD, method = "pearson",
conf.level = 0.95)
t.test(data$Mean_BMI_children, data$GDP_USD, var.equal = TRUE)
```

Figure 8.2

```
#hypothesis testing for adult BMI
cor.test(data$Mean_BMI_adults, data$GDP_USD, method = "pearson",
conf.level = 0.95)
t.test(data$Mean_BMI_adults, data$GDP_USD, var.equal = TRUE )
```

Figure 8.3

```
hypothesis testing for combined BMI
cor.test(data.com$Mean_BMI,data.com$GDP_USD, method = "pearson",
conf.level = 0.95)
t.test(data.com$Mean_BMI,data.com$GDP_USD, var.equal = TRUE)
```