THE IMPACT OF ECONOMIC LEVELS AND REGIONS ON THINNES, AND DISEASES ON LIFE EXPECTANCY

PROJECT REPORT SUBMITTED

IN FULFILMENT OF THE REQUIREMENTS FOR COURSE

STAT 467 – MULTIVARIATE ANALYSIS

DEPARTMENT OF STATISTICS OF

MIDDLE EAST TECHNICAL UNIVERSITY

BY

BATUHAN SAYLAM-2429264

NEBİH ŞAHİN-2429298

January 2024

We worked with the dataset, which examines life expectancy, education level, important diseases and special conditions in certain age groups, and has additional features.

First, we tried many methods to find the multivariate normality of our data, such as univariate normal variables or transformation methods. However, we could not examine multivariate normality and we accepted that and started our tests. Then we had to conduct exploratory data analysis before understanding the data better. Based on this, we started our study by mean vector inferences and comparisons of several multivariate means with the part of the data we are interested in. We began applying Principal Component Analysis with the covariance matrix to determine the correlation between variables and the components that represent them. We examined the components that affect the common disease of two different age groups and the effects of the diseases on life expectancy using comparison of several multivariate means, principle components analysis and regression, factor analysis and rotation, clustering and discrimination and classification.

1. Introduction

While using the World Health Organization's updated life expectancy dataset, we first examined how the economic situations of the countries in the dataset and the regions they are located in affect Thinnes in two different age groups or whether there is a relationship. In this process, we used methods such as comparisons of several multivariate means, principal components analysis and regression, and inferences about mean vector, and as a result, we concluded that both of our categorical variables have significant effects on age groups and the common problem they have. Secondly, after conducting exploratory data analysis, we wanted to investigate the effect of the diseases we were curious about on life expectancy and we made a detailed analysis with comparison of several multivariate means, principle components analysis and regression, factor analysis and rotation, clustering and discrimination and classification, and as a result, we have found that factors such as hepatitis B, diphtheria, polio and HIV have an impact on infant deaths and deaths under the age of five, and we have obtained detailed results as a result of our statistical measurements. In the rest of our report, we will explain how we examined these findings step by step.

1.1. Data Description

We found the dataset from Kaggle online data platform. While examining the data consisting of 19 variables and 2864 observations, we only took the data collected for year '2015' and started working on our life expectancy data with 19 columns and 179 rows. Data has 3 categorical variables and 16 numeric variables. Includes information on Population, GDP and Life Expectancy, Measles, Hepatitis B, Polio and Diphtheria vaccines, alcohol consumption, BMI, HIV cases, mortality rates and Thinnes. Information on schooling is also available.

Here is the link of datasets that we use:

(https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated)

In the tables below, we see the first 20 rows of the data;

	Country	Region	Infant_deaths	Under_five_deaths	Adult_mortality
1	Turkiye	Middle East	11.1	13.0	105.8240
2	Spain	European Union	2.7	3.3	57.9025
7	Russian Federation	Rest of Europe	6.6	8.2	223.0000
28	Cameroon	Africa	57.0	88.0	340.1265
44	Gambia, The	Africa	39.7	59.8	261.7065
58	Algeria	Africa	21.6	25.2	95.8155
75	Oman	Middle East	9.6	11.2	89.1875
102	Madagascar	Africa	41.3	59.0	218.4575
111	Norway	Rest of Europe	2.2	2.7	53.8970
113	Vietnam	Asia	17.4	21.8	124.5470
122	Eswatini	Africa	45.2	57.2	434.8210
161	Botswana	Africa	39.6	49.7	248.5950
167	Latvia	European Union	4.3	5.0	160.2370
174	Nepal	Asia	29.2	35.5	154.6040
183	Congo, Dem. Rep.	Africa	73.0	95.9	260.0235
203	Belarus	Rest of Europe	3.1	4.1	163.8720
217	Angola	Africa	57.7	88.1	242.9655
220	Ukraine	Rest of Europe	8.1	9.5	202.8000
242	Costa Rica	Central America and Caribbean	7.7	9.1	85.6715
269	Israel	Middle East	3.2	3.9	56.3370

	Alcohol_consumption	Hepatitis_B	Measles	BMI	Polio	Diphtheria	Incidents_HIV	GDP_per_capita
1	1.32	97	65	27.8	97	97	0.08	11006
2	10.35	97	94	26.0	97	97	0.09	25742
7	8.06	97	97	26.2	97	97	0.08	9313
28	4.55	84	64	24.3	77	84	1.12	1383
44	2.69	97	64	23.9	96	97	0.96	661
58	0.55	95	99	25.5	95	95	0.05	4178
75	0.45	99	98	26.3	99	99	0.05	18445
102	0.86	69	64	21.3	68	69	0.24	467
111	5.97	88	91	26.6	95	95	0.04	74356
113	2.98	97	65	21.7	97	97	0.12	2582
122	7.49	90	83	26.8	84	90	14.30	3680
161	6.14	95	80	24.3	96	95	5.81	6403
167	10.80	94	92	26.6	94	95	0.29	13786
174	0.24	91	65	22.4	90	91	0.17	902
183	4.47	81	64	21.9	78	81	0.35	497
203	9.79	99	99	26.6	99	99	0.22	5967
217	6.53	64	64	23.2	62	64	0.89	3128
220	6.04	22	57	26.5	51	23	0.23	2125
242	3.19	92	90	27.3	92	92	0.19	11643
269	2.74	96	97	27.2	95	95	0.08	35808

Population_mln	Thinness_ten_nineteen_years	Thinness_five_nine_years	Schooling	Life_expectancy	Economy_status
78.53	4.9	4.8	7.8	76.5	Economy_status_Developing
46.44	0.6	0.5	9.7	82.8	Economy_status_Developed
144.10	2.3	2.3	12.0	71.2	Economy_status_Developing
23.30	5.6	5.5	6.1	57.6	Economy_status_Developing
2.09	7.3	7.2	3.4	60.9	Economy_status_Developing
39.73	6.0	5.8	7.9	76.1	Economy_status_Developing
4.27	7.1	6.9	9.5	76.9	Economy_status_Developing
24.23	7.1	7.1	6.1	65.5	Economy_status_Developing
5.19	0.8	0.7	12.5	82.3	Economy_status_Developed
92.68	14.2	14.5	8.0	75.1	Economy_status_Developing
1.10	4.0	4.1	6.5	55.4	Economy_status_Developing
2.12	6.4	6.1	9.2	67.3	Economy_status_Developing
1.98	2.2	2.1	12.8	74.5	Economy_status_Developed
27.02	15.7	16.1	4.7	69.5	Economy_status_Developing
76.24	9.5	9.3	6.4	59.3	Economy_status_Developing
9.46	1.9	2.0	12.2	73.6	Economy_status_Developing
27.88	8.3	8.2	5.0	59.4	Economy_status_Developing
45.15	2.3	2.4	11.3	71.2	Economy_status_Developing
4.85	1.7	1.7	8.8	79.6	Economy_status_Developing
8.38	1.2	1.1	13.0	82.1	Economy_status_Developed

The following is the data structure:

```
##  $ Country                   : Factor w/ 179 levels "Afghanistan",..: 165 149 134 30 61 3 123 98 122 176 ...
##  $ Region                    : Factor w/ 9 levels "Africa","Asia",..: 5 4 8 1 1 1 5 1 8 2 ...
##  $ Infant_deaths             : num  11.1 2.7 6.6 57 39.7 21.6 9.6 41.3 2.2 17.4 ...
##  $ Under_five_deaths         : num  13 3.3 8.2 88 59.8 25.2 11.2 59 2.7 21.8 ...
##  $ Adult_mortality           : num  105.8 57.9 223 340.1 261.7 ...
##  $ Alcohol_consumption       : num  1.32 10.35 8.06 4.55 2.69 ...
##  $ Hepatitis_B               : num  97 97 97 84 97 95 99 69 88 97 ...
##  $ Measles                   : num  65 94 97 64 64 99 98 64 91 65 ...
##  $ BMI                       : num  27.8 26 26.2 24.3 23.9 25.5 26.3 21.3 26.6 21.7 ...
##  $ Polio                     : num  97 97 97 77 96 95 99 68 95 97 ...
##  $ Diphtheria                : num  97 97 97 84 97 95 99 69 95 97 ...
##  $ Incidents_HIV             : num  0.08 0.09 0.08 1.12 0.96 0.05 0.05 0.24 0.04 0.12 ...
##  $ GDP_per_capita            : num  11006 25742 9313 1383 661 ...
##  $ Population_mln            : num  78.53 46.44 144.1 23.3 2.09 ...
##  $ Thinness_ten_nineteen_years: num  4.9 0.6 2.3 5.6 7.3 6 7.1 7.1 0.8 14.2 ...
##  $ Thinness_five_nine_years  : num  4.8 0.5 2.3 5.5 7.2 5.8 6.9 7.1 0.7 14.5 ...
##  $ Schooling                 : num  7.8 9.7 12 6.1 3.4 7.9 9.5 6.1 12.5 8 ...
##  $ Life_expectancy           : num  76.5 82.8 71.2 57.6 60.9 76.1 76.9 65.5 82.3 75.1 ...
##  $ Economy_status            : Factor w/ 2 levels "Economy_status_Developed",..: 2 1 2 2 2 2 2 2 1 2 ...
```

1.2. Research Questions

The research questions are given below:

- Does the mean of the response variables differ significantly from the mean vector we created?
  Responses: "Thinness_five_nine_years" and"Thinness_ten_nineteen_years"
- Does the response variables varies with respect to economy status?
- Does the response variables varies with respect to region?
- Does the response variables varies with respect to both economy status and region?

1.3. Aim of the study

This project analyses the dataset using statistical applications and multivariate analysis components. It tries to draw conclusions by examining the correlation between two different age groups with the same problem according to their regions and economic status.

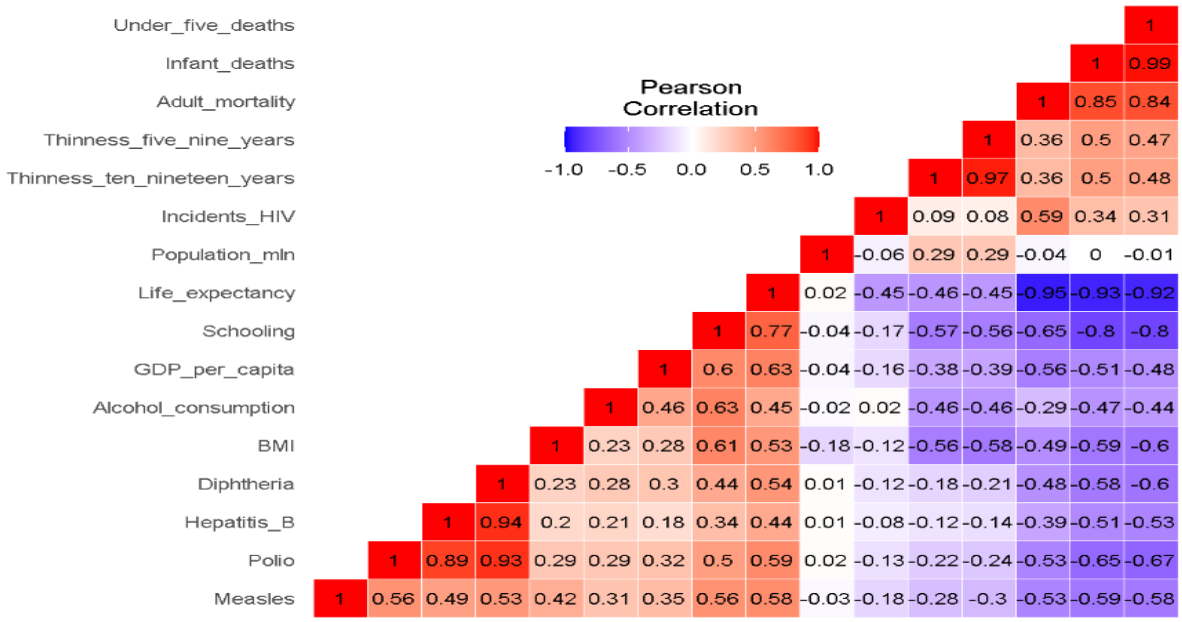
2. Methodology/Analysis

First of all, we examined the relationship between variables using Pearson correlation before conducting more consistent studies on the data set on which we applied exploratory analysis. Then we checked multivariate normality and univariate normality. We used Royston tests for multivariate normality and examined histogram plots for univariate normality. We used the Shapiro-Wilk test to check univariate normality. After that, Royston tests used for bivariate normality. We used Mahalanobis Distance and Adjusted Mahalanobis Distance for Multivariate Outlier detection. Then box-cox transformation for variables that did not provide normality. In the first question, we used Hotelling's T2-test for mean vector comparison of one population. We used Box's m test to test the homogeneity of covariance matrices. Also used Levene's Test to test the homogeneity of variance. In the second question, we used Hotelling's T2-test for mean vector comparison of two independent samples. In the third and fourth question, we used one-way Manova and two-way Manova, respectively, for comparisons of several multivariate means. Principal Component Analysis used to generate new unrelated variables by applying some linear transformations on related variables in the data. Then, we created a Regression model with the Principal Component we obtained. After that we used factor analysis to group variables based on their correlations. In order to identify a linear function that, using linear combination of the variables that exist in the data, describes or classifies at least two groups within the data in order to group on the basis of similarities or distances (dissimilarities) we used Cluster Analysis.

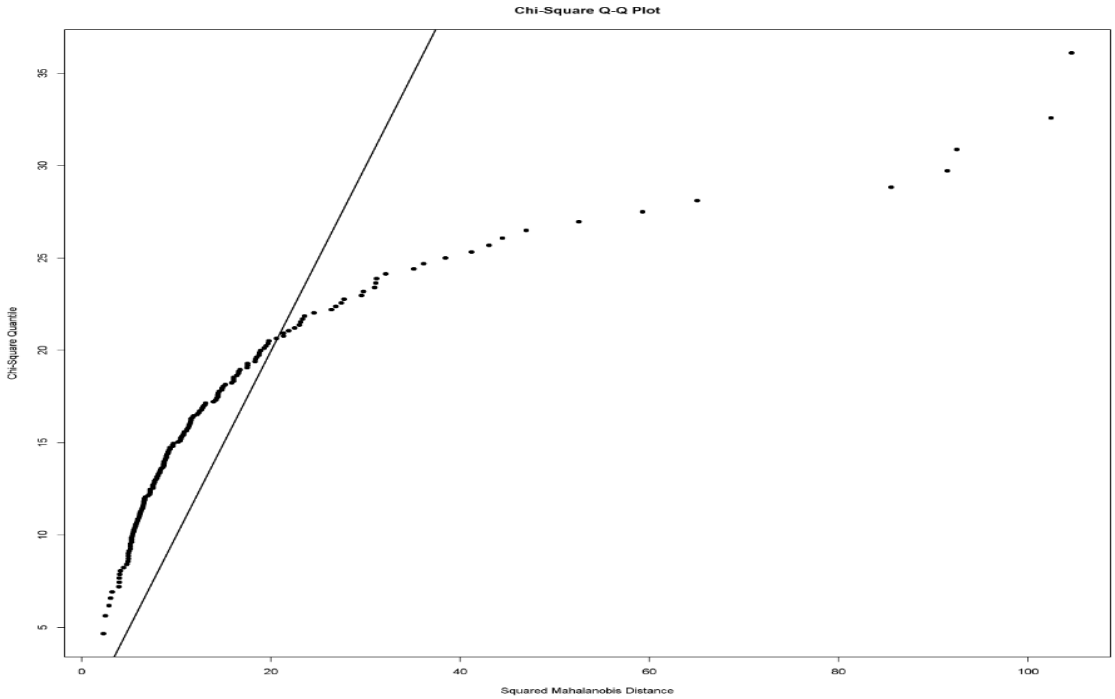
3. Results and Findings

3.1 Explatory Data Analysis

We examined the correlation between variables with Pearson correlation plot.



We checked the multivariate normality of the data with a multivariate QQ plot.



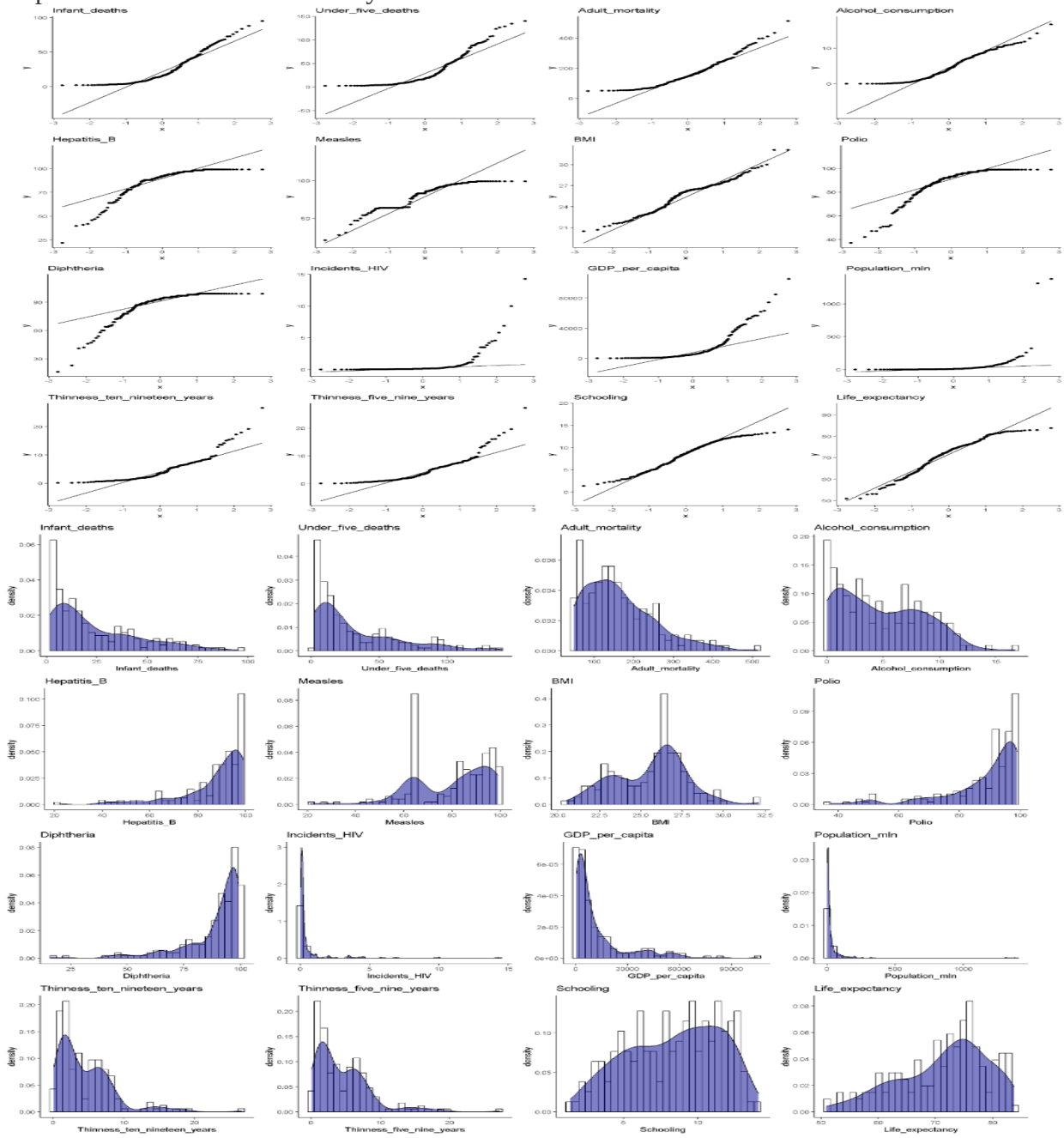
The variables violate the multivariate normality since there are large deviations from the line.

Also, we check the multivariate normality with Royston test.

```
##      Test         H      p value MVN
## 1 Royston 342.1765 2.013471e-70  NO
```

Since p-value is smaller than 0.05, we can reject the null hypothesis which is that the data follows normal distribution.

We performed univariate normality.



As you see, the variables which violates the multivariate normality.

```
##              Test                      Variable Statistic   p value Normality
## 1  Shapiro-Wilk         Infant_deaths              0.8612   <0.001        NO
## 2  Shapiro-Wilk       Under_five_deaths            0.8187   <0.001        NO
## 3  Shapiro-Wilk        Adult_mortality             0.9202   <0.001        NO
## 4  Shapiro-Wilk      Alcohol_consumption           0.9335   <0.001        NO
## 5  Shapiro-Wilk          Hepatitis_B               0.7746   <0.001        NO
## 6  Shapiro-Wilk             Measles                0.8876   <0.001        NO
## 7  Shapiro-Wilk                BMI                 0.9698    6e-04        NO
## 8  Shapiro-Wilk              Polio                 0.7744   <0.001        NO
## 9  Shapiro-Wilk           Diphtheria               0.7283   <0.001        NO
## 10 Shapiro-Wilk         Incidents_HIV              0.3677   <0.001        NO
## 11 Shapiro-Wilk        GDP_per_capita              0.6862   <0.001        NO
## 12 Shapiro-Wilk        Population_mln              0.2339   <0.001        NO
## 13 Shapiro-Wilk Thinness_ten_nineteen_years        0.8216   <0.001        NO
## 14 Shapiro-Wilk  Thinness_five_nine_years          0.8242   <0.001        NO
## 15 Shapiro-Wilk           Schooling                0.9617    1e-04        NO
## 16 Shapiro-Wilk        Life_expectancy             0.9528   <0.001        NO
```

Since the all p-values are smaller than 0.05, the variables which violates the multivariate normality.

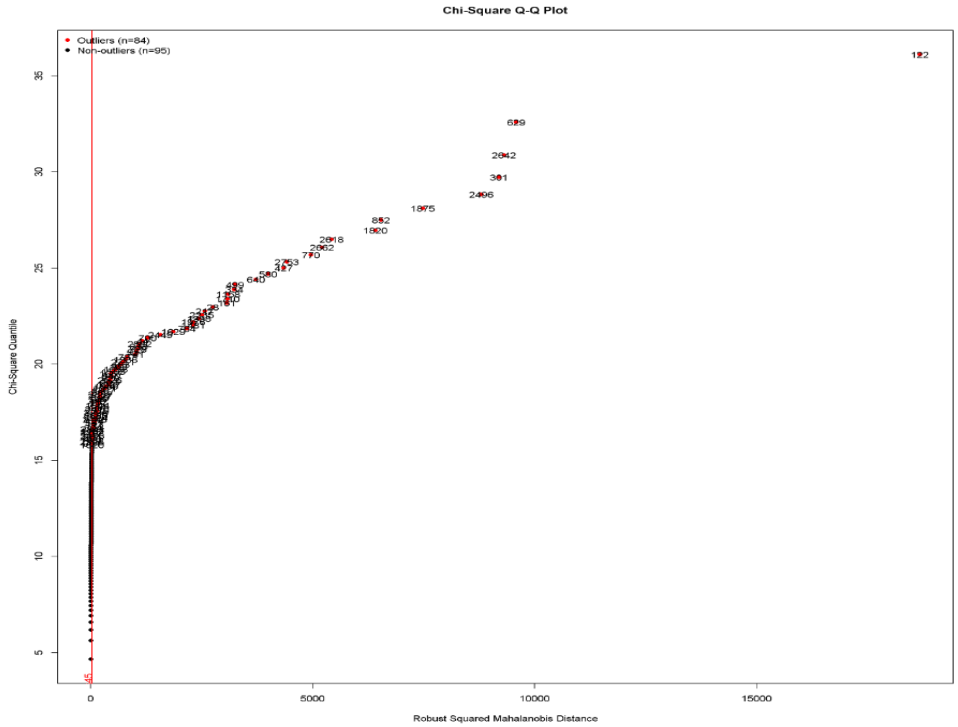
We performed a Royston test on binary variables to check bivariate normality.

```
##      [,1]                    [,2]                    [,3]
##  [1,] "Infant_deaths"         "Under_five_deaths"     "NO"
##  [2,] "Infant_deaths"         "Adult_mortality"       "NO"
##  [3,] "Infant_deaths"         "Alcohol_consumption"   "NO"
##  [4,] "Infant_deaths"         "Hepatitis_B"           "NO"
##  [5,] "Under_five_deaths"     "Adult_mortality"       "NO"
##  [6,] "Infant_deaths"         "BMI"                   "NO"
##  [7,] "Under_five_deaths"     "Alcohol_consumption"   "NO"
##  [8,] "Infant_deaths"         "Diphtheria"            "NO"
##  [9,] "Under_five_deaths"     "Hepatitis_B"           "NO"
## [10,] "Infant_deaths"         "GDP_per_capita"        "NO"
## [11,] "Adult_mortality"       "Alcohol_consumption"   "NO"
## [12,] "Infant_deaths"         "Thinness_ten_nineteen_years" "NO"
## [13,] "Under_five_deaths"     "BMI"                   "NO"
## [14,] "Adult_mortality"       "Hepatitis_B"           "NO"
## [15,] "Under_five_deaths"     "Polio"                 "NO"
## [16,] "Adult_mortality"       "Measles"               "NO"
## [17,] "Alcohol_consumption"   "Hepatitis_B"           "NO"
## [18,] "Adult_mortality"       "BMI"                   "NO"
## [19,] "Under_five_deaths"     "GDP_per_capita"        "NO"
## [20,] "Alcohol_consumption"   "Measles"               "NO"
## [21,] "Under_five_deaths"     "Thinness_ten_nineteen_years" "NO"
## [22,] "Adult_mortality"       "Diphtheria"            "NO"
## [23,] "Alcohol_consumption"   "BMI"                   "NO"
## [24,] "Hepatitis_B"           "Measles"               "NO"
## [25,] "Alcohol_consumption"   "Polio"                 "NO"
## [26,] "Adult_mortality"       "GDP_per_capita"        "NO"
## [27,] "Hepatitis_B"           "BMI"                   "NO"
## [28,] "Alcohol_consumption"   "Diphtheria"            "NO"
## [29,] "Adult_mortality"       "Thinness_ten_nineteen_years" "NO"
## [30,] "Hepatitis_B"           "Polio"                 "NO"
## [31,] "Measles"               "BMI"                   "NO"
## [32,] "Alcohol_consumption"   "GDP_per_capita"        "NO"
## [33,] "Hepatitis_B"           "Diphtheria"            "NO"
## [34,] "Measles"               "Polio"                 "NO"
## [35,] "Hepatitis_B"           "Incidents_HIV"         "NO"
## [36,] "Alcohol_consumption"   "Thinness_ten_nineteen_years" "NO"
## [37,] "Measles"               "Diphtheria"            "NO"
## [38,] "Hepatitis_B"           "GDP_per_capita"        "NO"
## [39,] "BMI"                   "Polio"                 "NO"
## [40,] "Measles"               "Incidents_HIV"         "NO"
## [41,] "BMI"                   "Diphtheria"            "NO"
## [42,] "Alcohol_consumption"   "Life_expectancy"       "NO"
## [43,] "Hepatitis_B"           "Thinness_ten_nineteen_years" "NO"
## [44,] "Measles"               "GDP_per_capita"        "NO"
## [45,] "BMI"                   "Incidents_HIV"         "NO"
## [46,] "Polio"                 "Diphtheria"            "NO"
## [47,] "Hepatitis_B"           "Schooling"             "NO"
## [48,] "BMI"                   "GDP_per_capita"        "NO"
## [49,] "Measles"               "Thinness_ten_nineteen_years" "NO"
## [50,] "Polio"                 "Incidents_HIV"         "NO"
## [51,] "BMI"                   "Population_mln"        "NO"
## [52,] "Polio"                 "GDP_per_capita"        "NO"
## [53,] "Diphtheria"            "Incidents_HIV"         "NO"
## [54,] "BMI"                   "Thinness_ten_nineteen_years" "NO"
## [55,] "Polio"                 "Population_mln"        "NO"
## [56,] "BMI"                   "Thinness_five_nine_years" "NO"
## [57,] "Diphtheria"            "GDP_per_capita"        "NO"
```

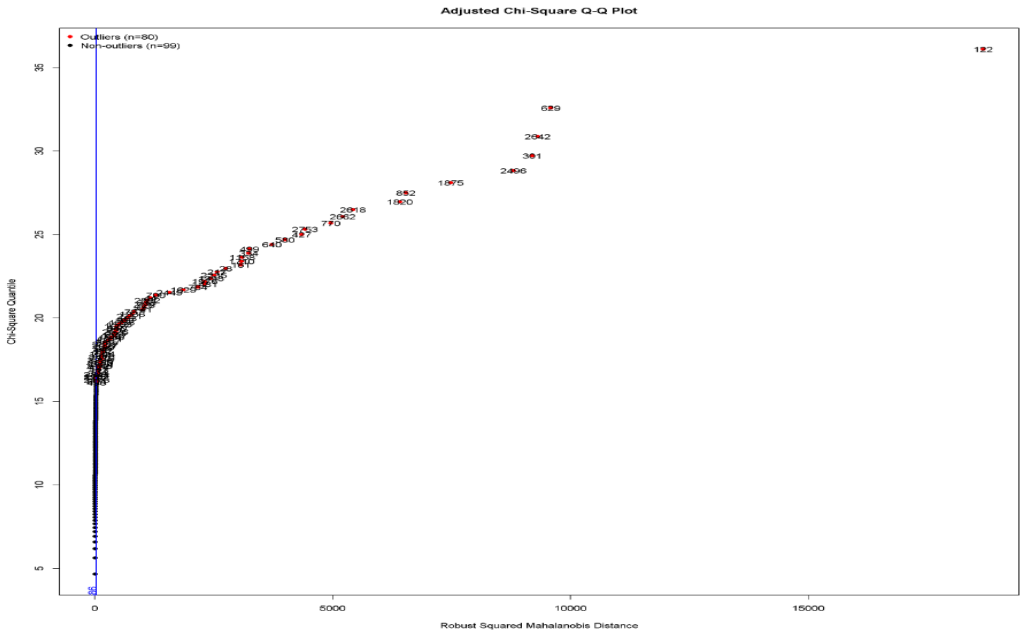
```
## [58,] "Polio"                    "Thinness_ten_nineteen_years" "NO"
## [59,] "BMI"                      "Schooling"                   "NO"
## [60,] "Diphtheria"               "Population_mln"              "NO"
## [61,] "Incidents_HIV"            "GDP_per_capita"              "NO"
## [62,] "Polio"                    "Thinness_five_nine_years"    "NO"
## [63,] "Diphtheria"               "Thinness_ten_nineteen_years" "NO"
## [64,] "Incidents_HIV"            "Population_mln"              "NO"
## [65,] "Diphtheria"               "Thinness_five_nine_years"    "NO"
## [66,] "Polio"                    "Life_expectancy"             "NO"
## [67,] "Incidents_HIV"            "Thinness_ten_nineteen_years" "NO"
## [68,] "GDP_per_capita"           "Population_mln"              "NO"
## [69,] "Diphtheria"               "Schooling"                   "NO"
## [70,] "Incidents_HIV"            "Thinness_five_nine_years"    "NO"
## [71,] "GDP_per_capita"           "Thinness_ten_nineteen_years" "NO"
## [72,] "Diphtheria"               "Life_expectancy"             "NO"
## [73,] "Incidents_HIV"            "Schooling"                   "NO"
## [74,] "GDP_per_capita"           "Thinness_five_nine_years"    "NO"
## [75,] "Population_mln"           "Thinness_ten_nineteen_years" "NO"
## [76,] "Incidents_HIV"            "Life_expectancy"             "NO"
## [77,] "GDP_per_capita"           "Schooling"                   "NO"
## [78,] "Population_mln"           "Thinness_five_nine_years"    "NO"
## [79,] "GDP_per_capita"           "Life_expectancy"             "NO"
## [80,] "Population_mln"           "Schooling"                   "NO"
## [81,] "Thinness_ten_nineteen_years" "Thinness_five_nine_years" "NO"
## [82,] "Population_mln"           "Life_expectancy"             "NO"
## [83,] "Thinness_ten_nineteen_years" "Schooling"                "NO"
## [84,] "Thinness_ten_nineteen_years" "Life_expectancy"          "NO"
## [85,] "Thinness_five_nine_years" "Schooling"                   "NO"
## [86,] "Thinness_five_nine_years" "Life_expectancy"             "NO"
## [87,] "Schooling"                "Life_expectancy"             "NO"
```

In addition to the univariate non-normality of every variable, bivariate non-normality also exhibits non-normality.

We detect outliers.



As can be seen, this dataset contains 84 outlier observations that are proved by the Mahalanobis Distance.



As can be seen, this dataset contains 80 outlier observations that are proved by the Adjusted Mahalanobis Distance.

According to two plots, since the number of the outliers is greater than the ten percent of the dataset, we did not remove the outliers.

3.2 Inferences about mean vector

To find the answer to the first question, is the average of the response variables significantly different from the average vector we created, we chose "Thinness_five_nine_years" and "Thinness_ten_nineteen_years" as the response variables. The mean vector we created is

[4.5 4.5].

The hypotheses are:

H0: Average of the response variables do not significantly different from the average vector we created.

H1: Average of the response variables significantly different from the average vector we created.

Firstly, we checked the multivariate normality with Mardia test.

```
##               Test         Statistic p value Result
## 1 Mardia Skewness 2455.21532300756       0     NO
## 2 Mardia Kurtosis  138.82062093672       0     NO
## 3              MVN              <NA>    <NA>     NO
```

According to result, response matrix does not provide multivariate normality.

```
##          Test                    Variable Statistic   p value Normality
## 1 Shapiro-Wilk  Thinness_five_nine_years      0.8242  <0.001      NO
## 2 Shapiro-Wilk Thinness_ten_nineteen_years    0.8216  <0.001      NO
```

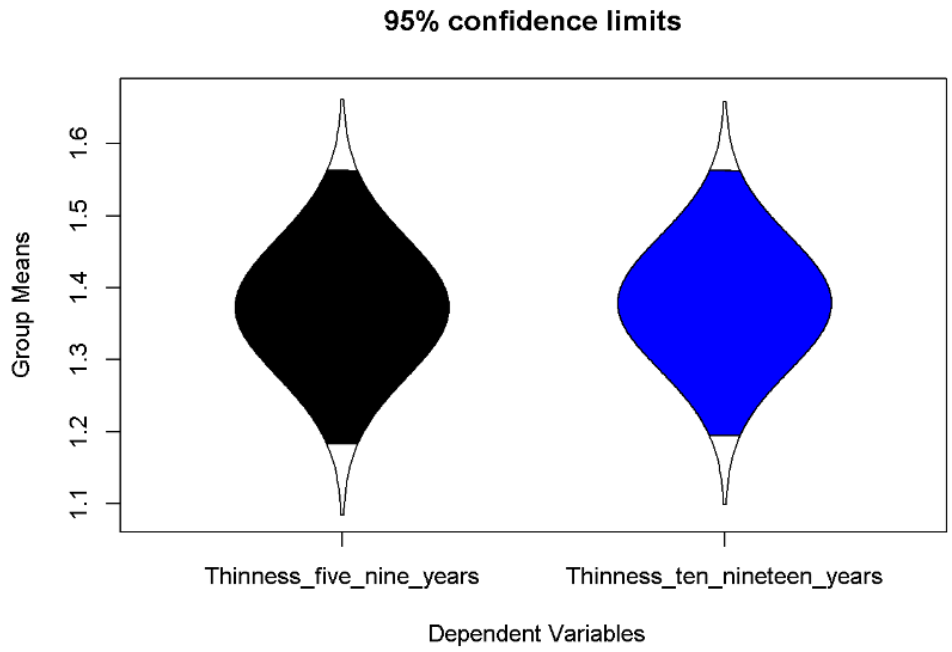
The results of the Shapiro-Wilk tests show us both variables are not normally distributed. Hence, both violate multivariate normality.

Thus, we need to normalize variables by boxcox transformation.

After the transformation,

```
##          Test                    Variable Statistic   p value Normality
## 1 Shapiro-Wilk  Thinness_five_nine_years      0.9879    0.1280    YES
## 2 Shapiro-Wilk Thinness_ten_nineteen_years    0.9859    0.0701    YES
```

The normality is satisfied. Let's see our response matrix before we begin the formal tests.



The means of the groupings do not differ significantly from one another.

Also, the transformed mean vector we created is **[1.785949 1.785949].**

The confidence intervals are (1.18284, 1.562754) for Thinness_five_nine_years and (1.194447, 1.562957) for Thinness_ten_nineteen_years.

As we have seen, both confidence intervals do not include the values of the mean vector we have created. Also, we conducted Hotelling T2- test.

```
##
##  Hotelling's one sample T2-test
##
## data:  datatest
## T.2 = 9.4836, df1 = 2, df2 = 177, p-value = 0.0001223
## alternative hypothesis: true location is not equal to c(1.78594895730098,1.78594895730098)
```

Since p-value is smaller than 0.05, we reject H0. Therefore, we don't have enough evidence to conclude that the transformation of the mean vector equals to transformation of (4.5,4.5).


```
##                                [,1]     [,2]
## Thinness_five_nine_years    1.134499 1.611095
## Thinness_ten_nineteen_years 1.147557 1.609847
```

The values of the mean vector that we created do not appear in the simultaneous confidence intervals for each variable because they do not fall inside the confidence region (that is, the point does not appear inside the ellipse), we reject the null hypothesis.

3.3 COMPARISONS OF SEVERAL MULTIVARIATE MEANS

To find the answer to the second question, is the average of the response variables significantly differ by the economy status, we chose "Thinness_five_nine_years" and "Thinness_ten_nineteen_years" as the response variables.

The hypotheses are:

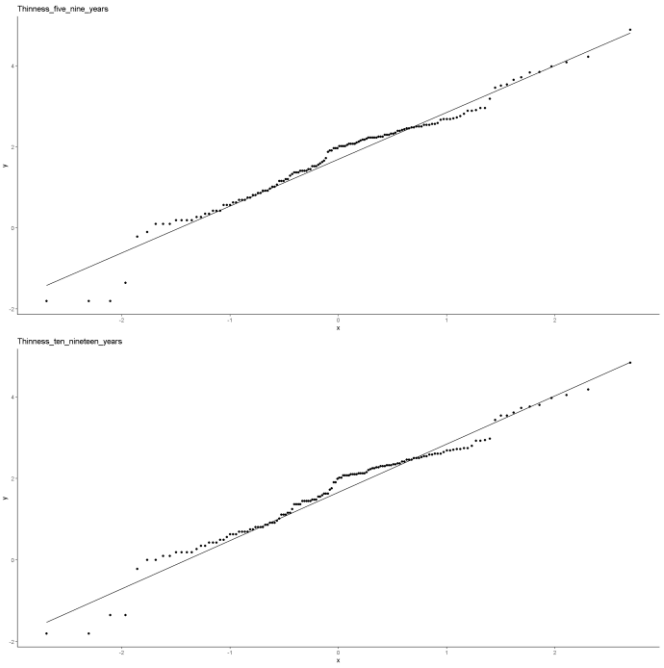
H0: Average of the response variables do not significantly differ by economy status

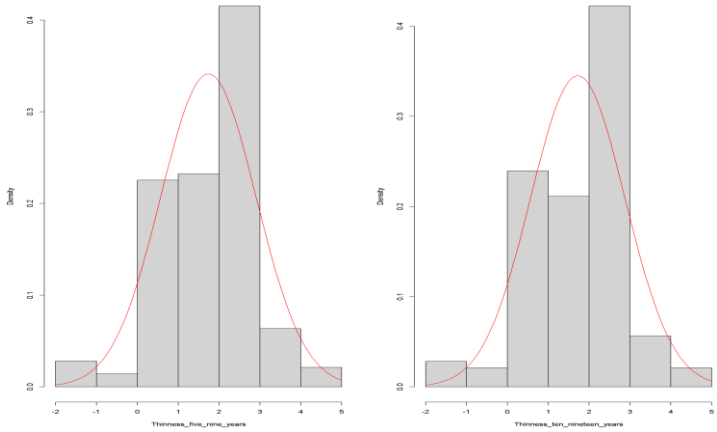
H1: Average of the response variables significantly differ by economy status

Firstly, we checked the univariate normalities of variable grouped by economy status with Shapiro-Wilk test.

```
## # A tibble: 4 × 4
##   Economy_status            variable                  statistic       p
##   <fct>                     <chr>                         <dbl>   <dbl>
## 1 Economy_status_Developed  Thinness_five_nine_years      0.977 0.625
## 2 Economy_status_Developed  Thinness_ten_nineteen_years   0.973 0.501
## 3 Economy_status_Developing Thinness_five_nine_years      0.972 0.00493
## 4 Economy_status_Developing Thinness_ten_nineteen_years   0.974 0.00839
```

The results of the Shapiro-Wilk tests show us the variables grouped by Economy_status_developing are not normally distributed.





Although the p-value is significant for each combination of Economy_status_Developing, we can not reject null hypothesis because of visual inspection leading us to believe that the data is normal.

We check the homogeneity of covariance matrix with Box's M-test.

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  cbind(datatest$Thinness_five_nine_years, datatest$Thinness_ten_nineteen_years)
## Chi-Sq (approx.) = 86.822, df = 3, p-value < 2.2e-16
```

We reject the null hypothesis and come to the conclusion that variance-covariance matrices are not equal for every combination of the dependent variable created by each group in the independent variable since the p-value for Box's M test is significant.

Given that the assumption is false, it would be wise to use Levene's test to verify the homogeneity of variance assumption and determine which variable fails in equal variance.

Levene's Test for Homogeneity of Variance of Thinness_five_nine_years group by Economy_status:

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   1  10.126 0.001727 **
##       177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene's Test for Homogeneity of Variance of Thinness_ten_nineteen_years group by Economy_status:

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1  14.493 0.0001937 ***
##       177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to Levene's tests, both variable fail in homogeneity of variance.

```
##
##  Hotelling's two sample T2-test
##
## data:  cbind(datatest$Thinness_ten_nineteen_years, datatest$Thinness_five_nine_years) by datat
est$Economy_status
## T.2 = 38.137, df1 = 2, df2 = 176, p-value = 1.732e-14
## alternative hypothesis: true location difference is not equal to c(0,0)
```

We reject H0 since the p value is significant. As a result, we have sufficient evidence to prove that the economy status has an impact on the mean of the responses.

Simultaneous Confidence Intervals of Thinness_five_nine_years is (1.155275, 2.332301).

Simultaneous Confidence Intervals of Thinness_ten_nineteen_years is (1.051640, 2.228666).

Bonferroni Confidence Intervals of Thinness_five_nine_years is (1.362527, 2.125050).

Bonferroni Confidence Intervals of Thinness_ten_nineteen_years is (1.258891, 2.021414).

Bonferroni confidence intervals give narrower intervals and both CI for Thinness_ten_nineteen_years and Thinness_five_nine_years does not include 0.

To find the answer to the third question, is the average of the response variables significantly differ by the region, we chose "Thinness_five_nine_years" and "Thinness_ten_nineteen_years" as the response variables.

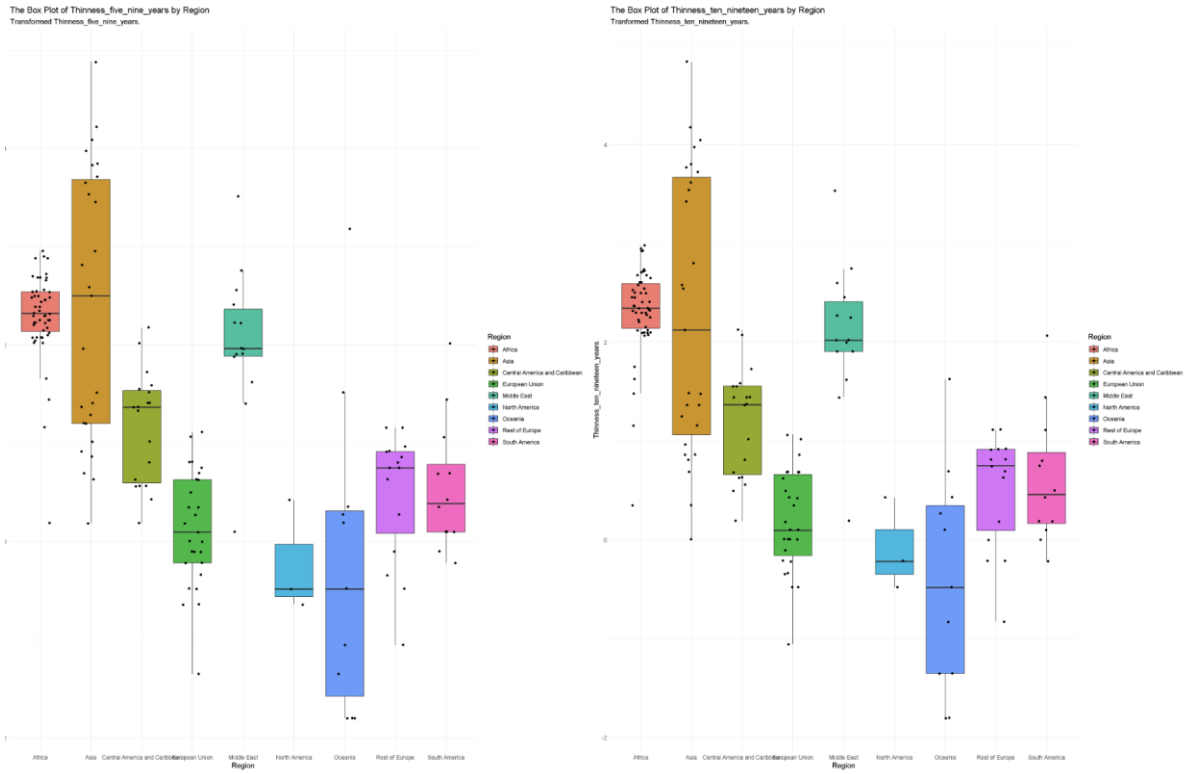
The hypotheses are:

H0: Average of the response variables do not significantly differ by region

H1: Average of the response variables significantly differ by region

Descriptive analysis:

```
## # A tibble: 9 × 6
##   Region                          n mean_9  sd_9 mean_19 sd_19
##   <fct>                       <int>  <dbl> <dbl>   <dbl> <dbl>
## 1 Africa                         51  2.29  0.455  2.32   0.452
## 2 Asia                           27  2.38  1.38   2.28   1.42
## 3 Central America and Caribbean  19  1.17  0.566  1.19   0.556
## 4 European Union                 27  0.131 0.589  0.191  0.516
## 5 Middle East                    14  2.03  0.764  2.07   0.750
## 6 North America                   3 -0.234 0.575 -0.0921 0.466
## 7 Oceania                        11 -0.254 1.58  -0.411  1.12
## 8 Rest of Europe                 15  0.479 0.671  0.505  0.580
## 9 South America                  12  0.554 0.671  0.613  0.666
```



We shall verify the assumptions following the descriptive analysis. Observe the normality. We are unable to employ the mvn once again because we have n<7 for North America.

Firstly, we checked the univariate normalities of variable grouped by economy status with Shapiro-Wilk test.

```
## # A tibble: 18 × 4
##    Region                         variable                    statistic        p
##    <fct>                          <chr>                           <dbl>    <dbl>
##  1 Africa                         Thinness_five_nine_years        0.811   1.33e-6
##  2 Africa                         Thinness_ten_nineteen_years     0.839   6.51e-6
##  3 Asia                           Thinness_five_nine_years        0.922   4.32e-2
##  4 Asia                           Thinness_ten_nineteen_years     0.920   3.99e-2
##  5 Central America and Caribbean Thinness_five_nine_years        0.946   3.35e-1
##  6 Central America and Caribbean Thinness_ten_nineteen_years     0.940   2.68e-1
##  7 European Union                 Thinness_five_nine_years        0.973   6.92e-1
##  8 European Union                 Thinness_ten_nineteen_years     0.971   6.30e-1
##  9 Middle East                    Thinness_five_nine_years        0.913   1.74e-1
## 10 Middle East                    Thinness_ten_nineteen_years     0.920   2.19e-1
## 11 North America                  Thinness_five_nine_years        0.859   2.66e-1
## 12 North America                  Thinness_ten_nineteen_years     0.946   5.51e-1
## 13 Oceania                        Thinness_five_nine_years        0.887   1.28e-1
## 14 Oceania                        Thinness_ten_nineteen_years     0.941   5.37e-1
## 15 Rest of Europe                 Thinness_five_nine_years        0.847   1.58e-2
## 16 Rest of Europe                 Thinness_ten_nineteen_years     0.860   2.39e-2
## 17 South America                  Thinness_five_nine_years        0.906   1.88e-1
## 18 South America                  Thinness_ten_nineteen_years     0.927   3.47e-1
```

All variables except the African region are normally distributed.

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  cbind(dataowa2$Thinness_five_nine_years, dataowa2$Thinness_ten_nineteen_years)
## Chi-Sq (approx.) = 453.64, df = 24, p-value < 2.2e-16
```

We reject the null hypothesis and come to the conclusion that variance-covariance matrices are not equal for every combination of the dependent variable created by each group in the independent variable since the p-value for Box's M test is significant.

Given that the assumption is false, it would be wise to use Levene's test to verify the homogeneity of variance assumption and determine which variable fails in equal variance.

Levene's Test for Homogeneity of Variance of Thinness_ten_nineteen_years group by Region:

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   8   12.91 1.878e-14 ***
##       170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene's Test for Homogeneity of Variance of Thinness_five_nine_years _years group by Region:

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   8  11.044 1.668e-12 ***
##       170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to Levene's tests, both variable fail in homogeneity of variance.

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## Region      8 0.67122   10.734     16    340 < 2.2e-16 ***
## Residuals 170
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is less than alpha, we can thus conclude with 95% confidence that at least one region differs significantly from the others.

To determine which of the two causes the difference, a post-hoc analysis must be conducted if the null hypothesis in the MANOVA is rejected.

```
##  Response Thinness_five_nine_years :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Region        8 175.96 21.9951   31.35 < 2.2e-16 ***
## Residuals   170 119.27  0.7016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Thinness_ten_nineteen_years :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Region        8 172.66 21.5830  34.908 < 2.2e-16 ***
## Residuals   170 105.11  0.6183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is clear from the result above that there is a highly significant difference between the two variables by region.

To find the answer to the fourth question, is the average of the response variables significantly differ by the region and economy status, we chose "Thinness_five_nine_years" and "Thinness_ten_nineteen_years" as the response variables.

The hypotheses are:

H0: Average of the response variables do not significantly differ by region and economy status

H1: Average of the response variables significantly differ by region and economy status

```
## # A tibble: 14 × 2
##    mix                                                      n
##    <fct>                                                <int>
##  1 Africa Economy_status_Developing                        51
##  2 Asia Economy_status_Developed                            1
##  3 Asia Economy_status_Developing                          26
##  4 Central America and Caribbean Economy_status_Developing    19
##  5 European Union Economy_status_Developed                 27
##  6 Middle East Economy_status_Developed                     1
##  7 Middle East Economy_status_Developing                   13
##  8 North America Economy_status_Developed                   2
##  9 North America Economy_status_Developing                  1
## 10 Oceania Economy_status_Developed                         2
## 11 Oceania Economy_status_Developing                        9
## 12 Rest of Europe Economy_status_Developed                  4
## 13 Rest of Europe Economy_status_Developing                11
## 14 South America Economy_status_Developing                 12
```

Also, since we need sample size, which is larger than or equal to 3, we need to remove the data which is smaller than 3. After that,

```
## # A tibble: 9 × 2
##   mix                                                      n
##   <fct>                                                <int>
## 1 Africa Economy_status_Developing                        51
## 2 Asia Economy_status_Developing                          26
## 3 Central America and Caribbean Economy_status_Developing    19
## 4 European Union Economy_status_Developed                 27
## 5 Middle East Economy_status_Developing                   13
## 6 Oceania Economy_status_Developing                        9
## 7 Rest of Europe Economy_status_Developed                  4
## 8 Rest of Europe Economy_status_Developing                11
## 9 South America Economy_status_Developing                 12
```

We shall verify the assumptions following the descriptive analysis. Observe the normality. We are unable to employ the mvn once again because we have n<7 for groups.

```
## # A tibble: 18 × 4
##    mix                                          variable statistic       p
##    <fct>                                        <chr>        <dbl>   <dbl>
##  1 Africa Economy_status_Developing             Thinnes…     0.811 1.33e-6
##  2 Africa Economy_status_Developing             Thinnes…     0.839 6.51e-6
##  3 Asia Economy_status_Developing               Thinnes…     0.926 6.08e-2
##  4 Asia Economy_status_Developing               Thinnes…     0.926 6.23e-2
##  5 Central America and Caribbean Economy_status_Deve… Thinnes…     0.946 3.35e-1
##  6 Central America and Caribbean Economy_status_Deve… Thinnes…     0.940 2.68e-1
##  7 European Union Economy_status_Developed      Thinnes…     0.973 6.92e-1
##  8 European Union Economy_status_Developed      Thinnes…     0.971 6.30e-1
##  9 Middle East Economy_status_Developing        Thinnes…     0.918 2.34e-1
## 10 Middle East Economy_status_Developing        Thinnes…     0.924 2.86e-1
## 11 Oceania Economy_status_Developing            Thinnes…     0.876 1.43e-1
## 12 Oceania Economy_status_Developing            Thinnes…     0.896 2.29e-1
## 13 Rest of Europe Economy_status_Developed      Thinnes…     0.935 6.25e-1
## 14 Rest of Europe Economy_status_Developed      Thinnes…     0.849 2.23e-1
## 15 Rest of Europe Economy_status_Developing     Thinnes…     0.920 3.16e-1
## 16 Rest of Europe Economy_status_Developing     Thinnes…     0.884 1.16e-1
## 17 South America Economy_status_Developing      Thinnes…     0.906 1.88e-1
## 18 South America Economy_status_Developing      Thinnes…     0.927 3.47e-1
```

All variables except the African region are normally distributed.

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  cbind(datatwa$Thinness_five_nine_years, datatwa$Thinness_ten_nineteen_years)
## Chi-Sq (approx.) = 453.64, df = 24, p-value < 2.2e-16
```

We reject the null hypothesis and come to the conclusion that variance-covariance matrices are not equal for every combination of the dependent variable created by each group in the independent variable since the p-value for Box's M test is significant.

```
##                         Df  Pillai approx F num Df den Df    Pr(>F)
## Region                   8 0.69743  11.0431     16    330 < 2.2e-16 ***
## Economy_status           1 0.10480   9.6001      2    164 0.0001141 ***
## Region:Economy_status    4 0.02408   0.5026      8    330 0.8541490
## Residuals              165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is less than alpha, we can thus conclude with 95% confidence that the combination of economy status and region does not differ significantly from the others. However, at least one economy status or region differs significantly from the others.

To determine which of the two causes the difference, a post-hoc analysis must be conducted if the null hypothesis in the MANOVA is rejected.

.

```
##  Response Thinness_five_nine_years :
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
## Region                   8 175.961 21.9951 34.3828 < 2.2e-16 ***
## Economy_status           1  12.085 12.0851 18.8915 2.412e-05 ***
## Region:Economy_status    4   1.634  0.4084  0.6384    0.6358
## Residuals              165 105.552  0.6397
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Thinness_ten_nineteen_years :
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
## Region                   8 172.664 21.5830 37.5591 < 2.2e-16 ***
## Economy_status           1   8.202  8.2017 14.2727 0.0002205 ***
## Region:Economy_status    4   2.091  0.5228  0.9097 0.4597268
## Residuals              165  94.816  0.5746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

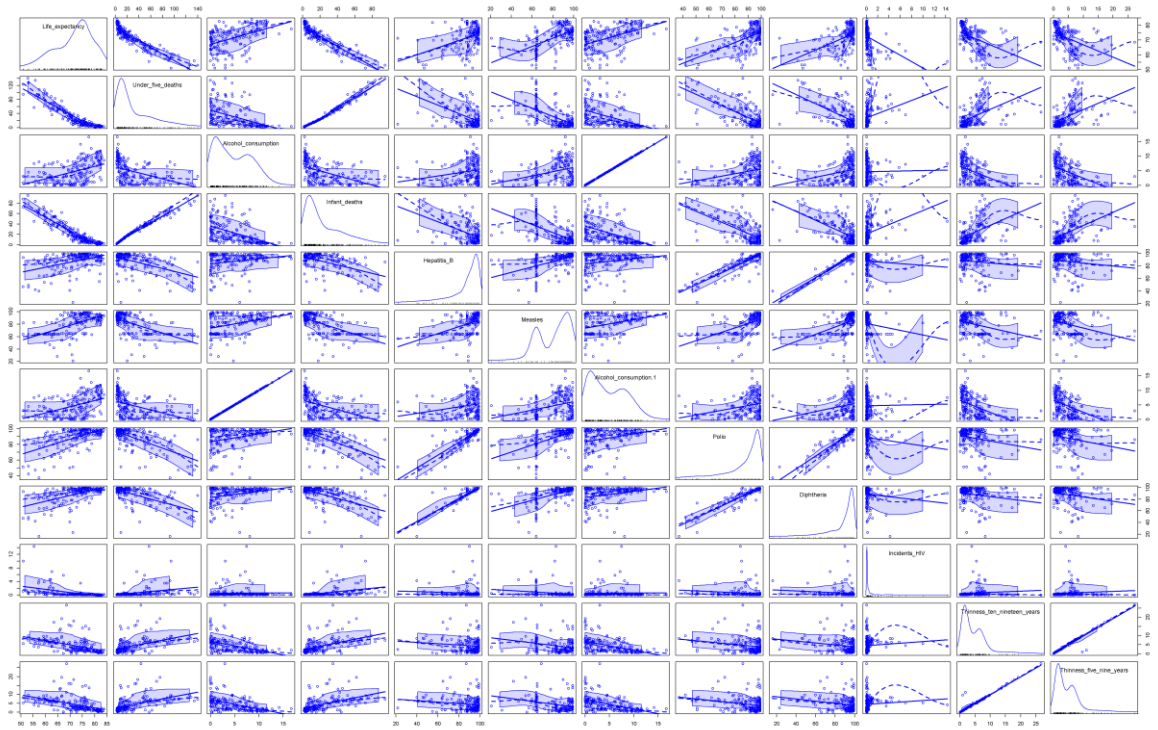
It is clear from the result above that there is a highly significant difference between the two variables by region or economy status since p-values are smaller than 0.05.

3. 4 PRINCIPAL COMPONENTS ANALAYSIS and PRINCIPAL COMPONENTS REGRESSION

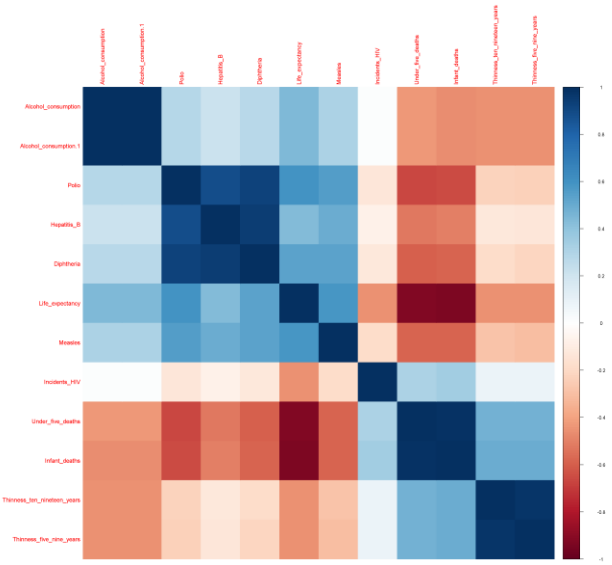
We select the follwing varialbes :

Life_expectancy, Under_five_deaths, Alcohol_consumption, Infant_deaths, Hepatitis_B, Measles, Alcohol_consumption, Polio, Diphtheria, Incidents_HIV, Thinness_ten_nineteen_years, Thinness_five_nine_year'

The scatter and histogram plots of variables:



The pearson correlation plot of data:



Firstly, we need to scale the data. The some of the scaled data:

```
##      Life_expectancy Under_five_deaths Alcohol_consumption Infant_deaths
## 1         0.643020804      -0.579830261        -0.910686853  -0.579809404
## 2         1.447385265      -0.880912648         1.501608764  -0.970734215
## 7        -0.033666758      -0.728819483         0.889852733  -0.789233410
## 28       -1.770072577       1.748126333        -0.047816992   1.556315454
## 44       -1.348738812       0.872814653        -0.544701804   0.751196499
## 58        0.591950045      -0.201149322        -1.116386479  -0.091153391
## 75        0.694091564      -0.635701220        -1.143100717  -0.649617406
## 102      -0.761425079       0.847983116        -1.033572344   0.825658367
## 111       1.383546816      -0.899536300         0.331525176  -0.994003549
## 113       0.464273147      -0.306683354        -0.467230516  -0.286615796
## 122      -2.050961754       0.792112158         0.737581581   1.007159172
## 161      -0.531606662       0.559316499         0.376939379   0.746542632
## 167       0.387667007      -0.828145632         1.621822832  -0.896272346
## 174      -0.250717485       0.118556717        -1.199200615   0.262540485
## 183      -1.553021850       1.993337761        -0.069188382   2.300934141
## 203       0.272757799      -0.856081111         1.352009036  -0.952118748
```

The covariance matrix must be calculated as the initial step in PCA.

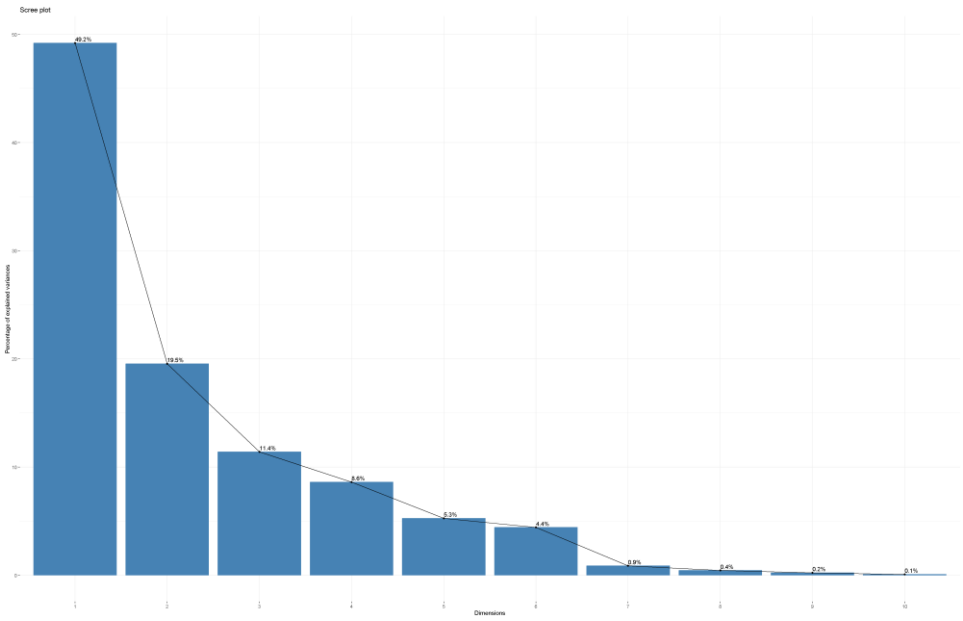
```
##                          Life_expectancy Under_five_deaths
## Life_expectancy                1.0000000       -0.9212079
## Under_five_deaths             -0.9212079        1.0000000
## Alcohol_consumption            0.4487724       -0.4391376
## Infant_deaths                 -0.9305646        0.9899859
## Hepatitis_B                    0.4397991       -0.5286901
## Measles                        0.5838125       -0.5845977
## Alcohol_consumption.1          0.4487724       -0.4391376
## Polio                          0.5910969       -0.6650301
## Diphtheria                     0.5370849       -0.5953535
## Incidents_HIV                 -0.4501293        0.3122778
## Thinness_ten_nineteen_years   -0.4563343        0.4789483
## Thinness_five_nine_years      -0.4522485        0.4719123
##                          Alcohol_consumption Infant_deaths Hepatitis_B
## Life_expectancy                   0.44877242    -0.9305646   0.4397991
## Under_five_deaths                -0.43913759     0.9899859  -0.5286901
## Alcohol_consumption               1.00000000    -0.4692522   0.2102431
## Infant_deaths                    -0.46925224     1.0000000  -0.5063603
## Hepatitis_B                       0.21024309    -0.5063603   1.0000000
## Measles                           0.31445228    -0.5861101   0.4905367
## Alcohol_consumption.1             1.00000000    -0.4692522   0.2102431
## Polio                             0.28825702    -0.6519378   0.8886645
## Diphtheria                        0.27565589    -0.5814566   0.9438458
## Incidents_HIV                     0.01760975     0.3409592  -0.0767645
## Thinness_ten_nineteen_years      -0.45626812     0.4964201  -0.1234409
## Thinness_five_nine_years         -0.45939934     0.4954018  -0.1398797
##                             Measles Alcohol_consumption.1      Polio
## Life_expectancy           0.5838125           0.44877242  0.5910969
## Under_five_deaths        -0.5845977          -0.43913759 -0.6650301
## Alcohol_consumption       0.3144523           1.00000000  0.2882570
## Infant_deaths            -0.5861101          -0.46925224 -0.6519378
## Hepatitis_B               0.4905367           0.21024309  0.8886645
## Measles                   1.0000000           0.31445228  0.5571360
## Alcohol_consumption.1     0.3144523           1.00000000  0.2882570
## Polio                     0.5571360           0.28825702  1.0000000
## Diphtheria                0.5305231           0.27565589  0.9296290
## Incidents_HIV            -0.1843985           0.01760975 -0.1342147
## Thinness_ten_nineteen_years -0.2806768       -0.45626812 -0.2212835
## Thinness_five_nine_years  -0.3048480          -0.45939934 -0.2375340
```

```
##                          Diphtheria Incidents_HIV
## Life_expectancy           0.5370849   -0.45012933
## Under_five_deaths        -0.5953535    0.31227776
## Alcohol_consumption       0.2756559    0.01760975
## Infant_deaths            -0.5814566    0.34095920
## Hepatitis_B               0.9438458   -0.07676450
## Measles                   0.5305231   -0.18439848
## Alcohol_consumption.1     0.2756559    0.01760975
## Polio                     0.9296290   -0.13421473
## Diphtheria                1.0000000   -0.12232658
## Incidents_HIV            -0.1223266    1.00000000
## Thinness_ten_nineteen_years -0.1821165  0.08962638
## Thinness_five_nine_years  -0.2107806    0.08353299
##                          Thinness_ten_nineteen_years
## Life_expectancy                           -0.45633430
## Under_five_deaths                          0.47894832
## Alcohol_consumption                       -0.45626812
## Infant_deaths                              0.49642015
## Hepatitis_B                               -0.12344090
## Measles                                   -0.28067679
## Alcohol_consumption.1                     -0.45626812
## Polio                                     -0.22128352
## Diphtheria                                -0.18211650
## Incidents_HIV                              0.08962638
## Thinness_ten_nineteen_years                1.00000000
## Thinness_five_nine_years                   0.97313553
##                          Thinness_five_nine_years
## Life_expectancy                       -0.45224853
## Under_five_deaths                      0.47191227
## Alcohol_consumption                   -0.45939934
## Infant_deaths                          0.49540181
## Hepatitis_B                           -0.13987969
## Measles                               -0.30484804
## Alcohol_consumption.1                 -0.45939934
## Polio                                 -0.23753399
## Diphtheria                            -0.21078062
## Incidents_HIV                          0.08353299
## Thinness_ten_nineteen_years            0.97313553
## Thinness_five_nine_years               1.00000000
```

Apply the PCA after we need to remove our response variable (Life_expectancy).

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     2.3268 1.4661 1.1195 0.9732 0.76029 0.69642 0.30678
## Proportion of Variance 0.4922 0.1954 0.1139 0.0861 0.05255 0.04409 0.00856
## Cumulative Proportion  0.4922 0.6876 0.8015 0.8876 0.94017 0.98426 0.99281
##                            PC8    PC9    PC10      PC11
## Standard deviation     0.21680 0.1555 0.08864 3.576e-16
## Proportion of Variance 0.00427 0.0022 0.00071 0.000e+00
## Cumulative Proportion  0.99709 0.9993 1.00000 1.000e+00
```

The standard deviation, the portion of variance explained by each principal component, and the cumulative proportion of variance explained are provided by the summary function on the result object. For instance, it is excellent that the first six components account for 98.43% of the data variability.



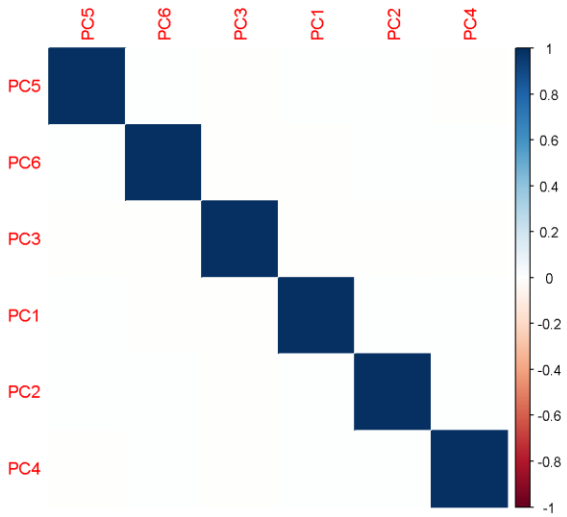
As you can see, 6 components seemed to be sufficient. As we can see from the summary result, the first six components account for over 98.4% of the variability, which is excellent.

Let's now extract the first six components and use them to continue our analysis.

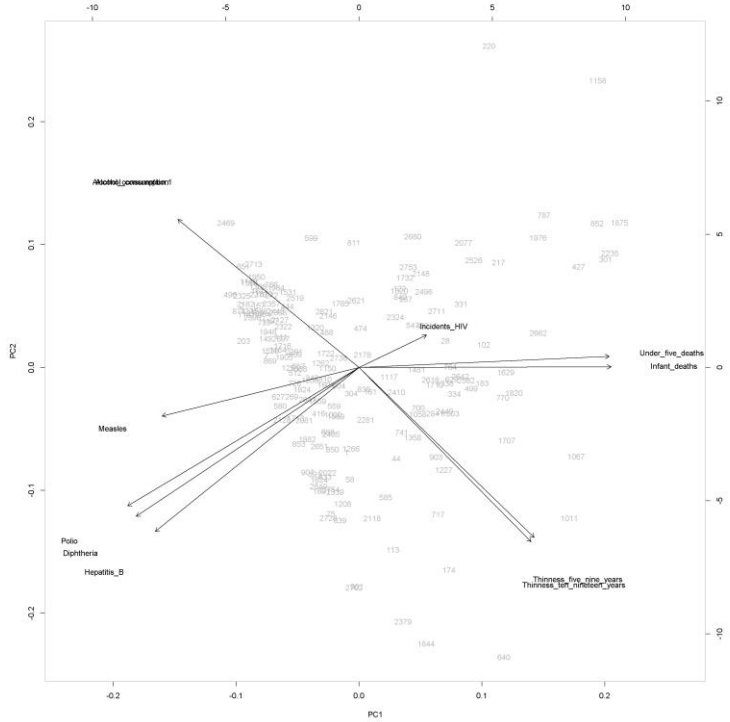
The first six row of PCA:

```
##           PC1        PC2        PC3        PC4        PC5        PC6
## 1  -0.3166815 -1.3499627 -0.8031473 -0.5986859  0.8318938 -1.0141867
## 2  -2.9691284  0.9949661  0.4086677  0.4027019  0.1022927  0.2818884
## 7  -2.3439685  0.1975856  0.1613216  0.3110764 -0.1983731  0.3556468
## 28  2.1876303  0.4295108  0.8815823 -0.3521410  0.9353661  0.8560820
## 44  0.9390288 -1.4496461  0.7014674 -0.5244228  1.1109567  0.0575066
## 58 -0.2264585 -1.7830810 -0.8810037 -0.4240393 -0.7724058  0.4869355
```

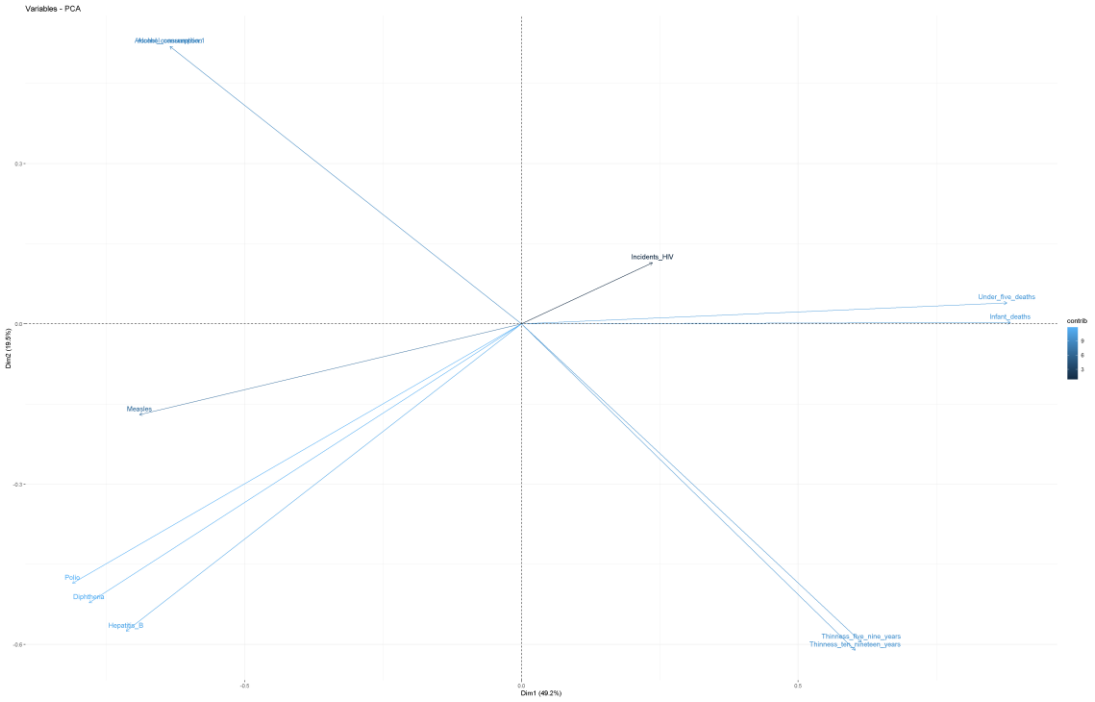
We want to confirm that the components we use should be orthogonal. To put it another way, they need to be linearly independent. Create a correlation plot of the pca's to verify this.



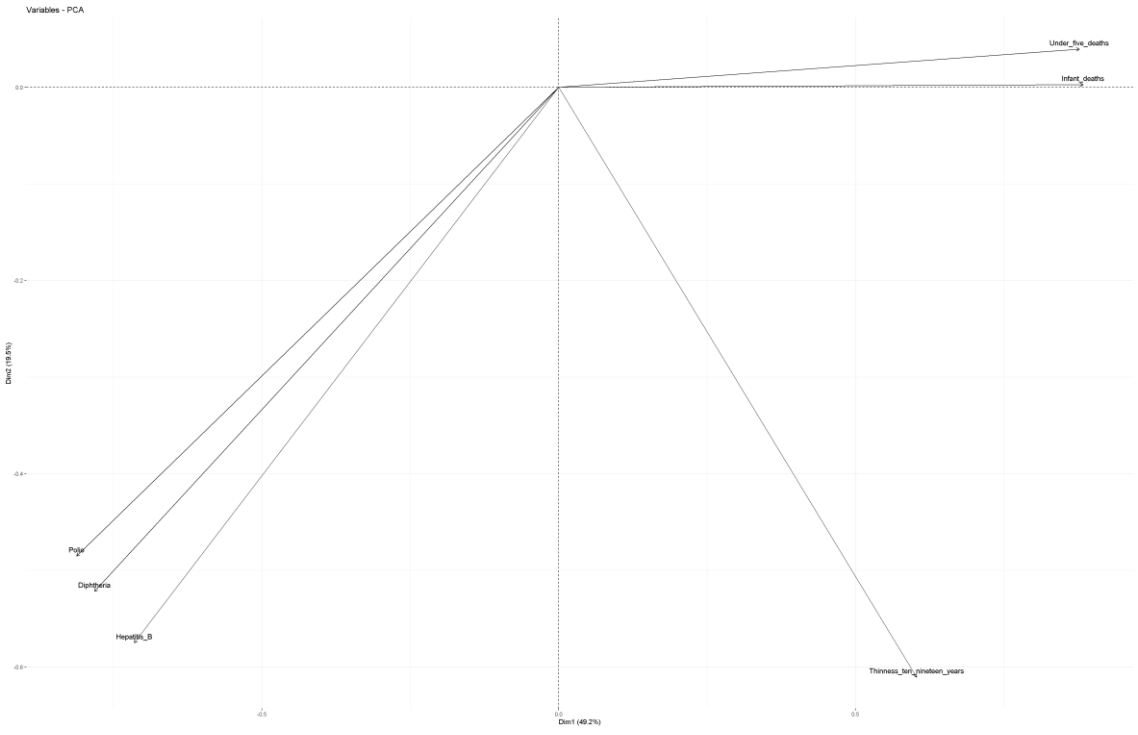
As we have seen, the components are linearly independent.



Within this plot, PC1 is heavily influenced by Incidents_HIV, Under_five_deaths, Infant_deaths, Thinness_five_nine_years, and Thinness_ten_nineteen_years, whereas PC2 is more influenced by the remaining factors.

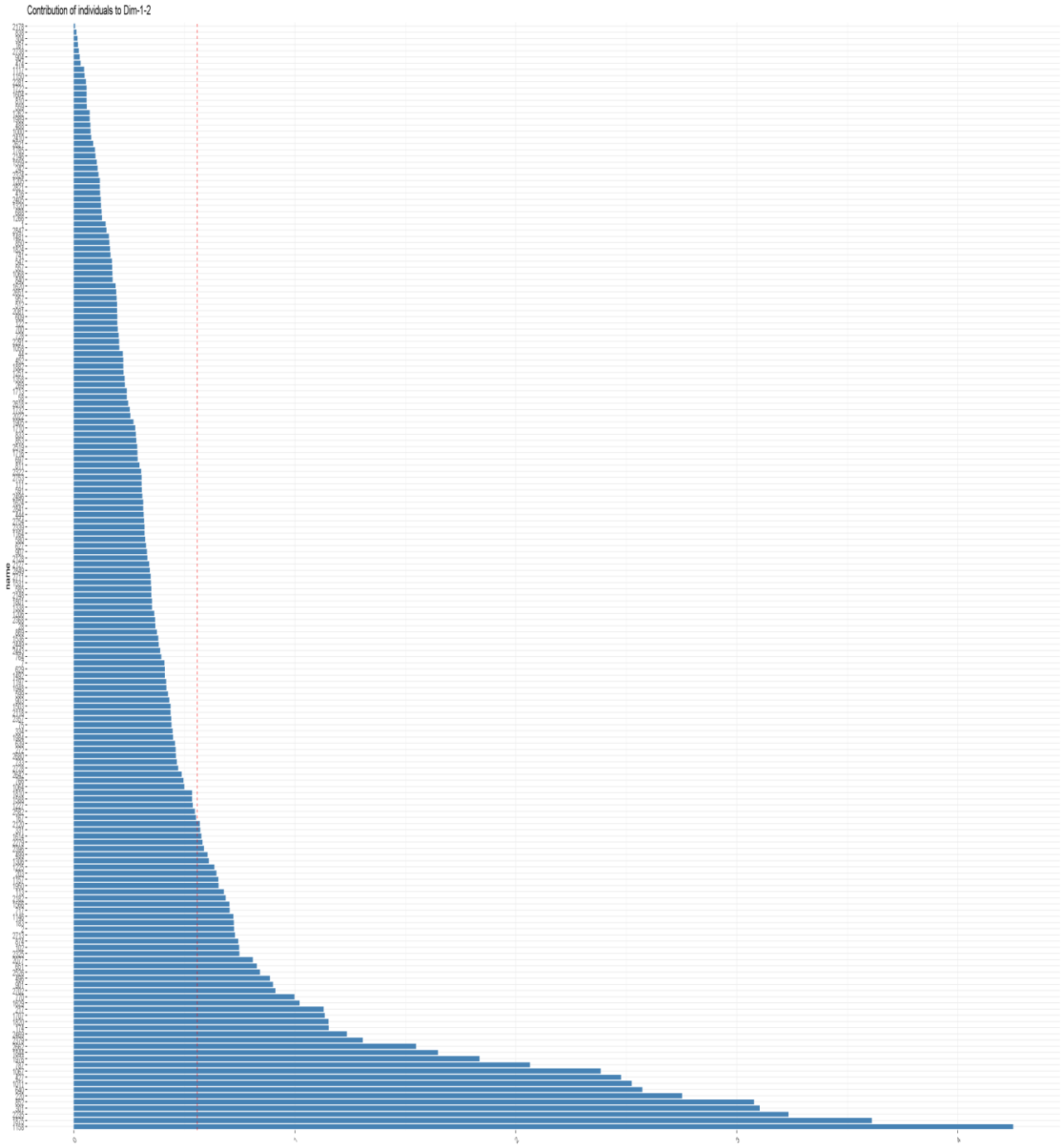


Display the variables that have a significant impact, excluding Infant Deaths and HIV Incidents.



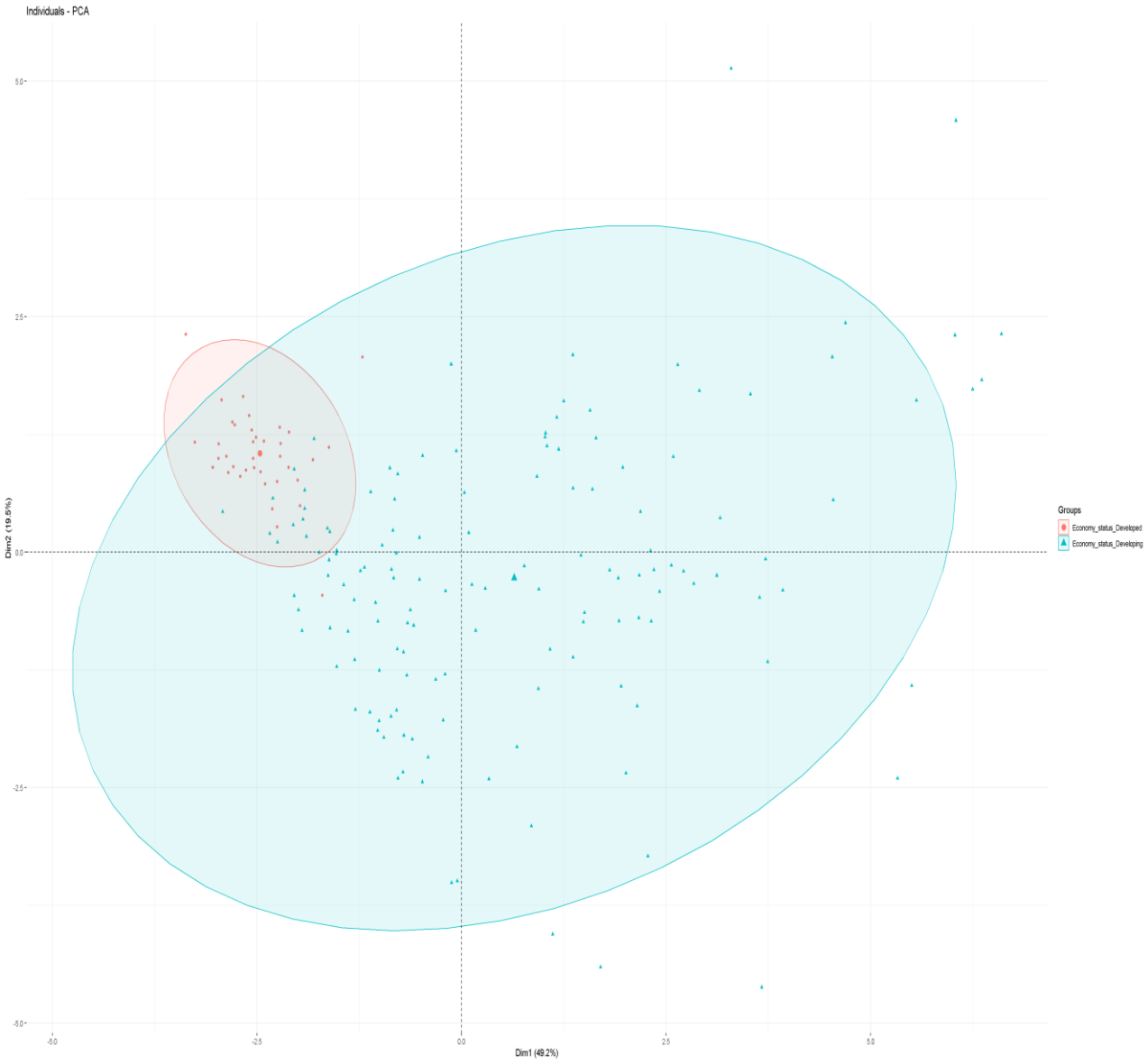
The first two components being most significantly influenced by the first six variables are Polio, Infant_deaths, Hepatitis_B, Under_five_deaths, Diphtheria, and Thinness_ten_nineteen_years

The plot shows which row is explained by PC1 or PC2. Each row number represents the country which is in the row.



The individual contributions to the components can also be seen. Equatorial Guinea (1158th row), for instance, has the largest contribution to the first two components, as can be shown.

The observations can also be classified according to economic status.



Therefore, we are ready for Principal Component Regression.

The response variable is life expectancy and explaratory variables are PC's.

```
##
## Call:
## lm(formula = Life_expectancy ~ ., data = ols.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81568 -0.23490 -0.01441  0.24825  0.76545
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.654e-16  2.490e-02   0.000    1.000
## PC1         -3.555e-01  1.073e-02 -33.122  < 2e-16 ***
## PC2          1.532e-03  1.703e-02   0.090    0.928
## PC3         -2.739e-01  2.231e-02 -12.277  < 2e-16 ***
## PC4          2.461e-01  2.566e-02   9.590  < 2e-16 ***
## PC5         -1.344e-01  3.285e-02  -4.091 6.60e-05 ***
## PC6         -3.108e-01  3.586e-02  -8.668 3.17e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 0.3332 on 172 degrees of freedom
## Multiple R-squared:  0.8927, Adjusted R-squared:  0.889
## F-statistic: 238.6 on 6 and 172 DF,  p-value: < 2.2e-16
```

The model is significant, as you can see. Components explain approximately 89% of y's variability. Furthermore, the sixth component, which provides the most variance, can be considered significant.

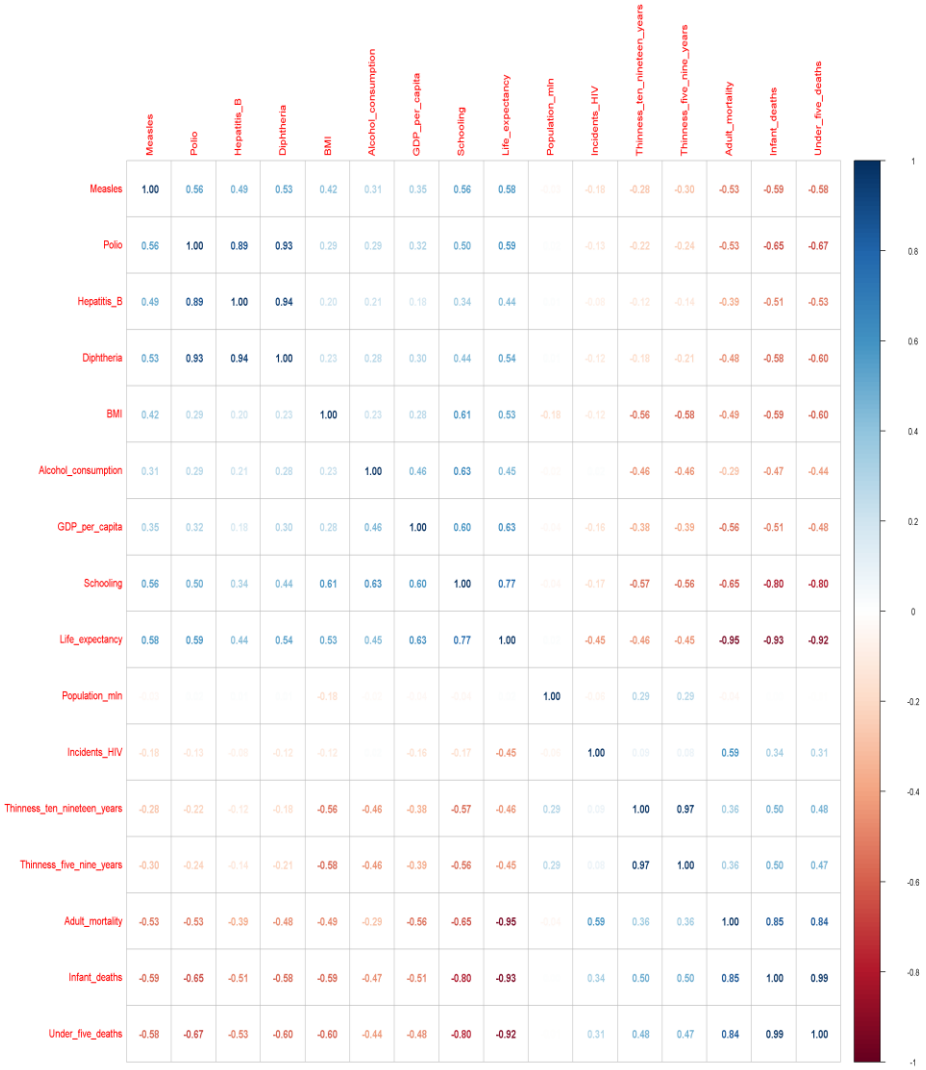
We can use performance metrics like MSE or RMSE to evaluate the model's performance in the data.

MSE:0.1066567

RMSE: 0.3265834

3.5 FACTOR ANALYSIS AND FACTOR ROTATION

We compute and display the data's correlation.



Even though the image above is not very apparent due to the large number of variables, we can still see that some of the variables are correlated.

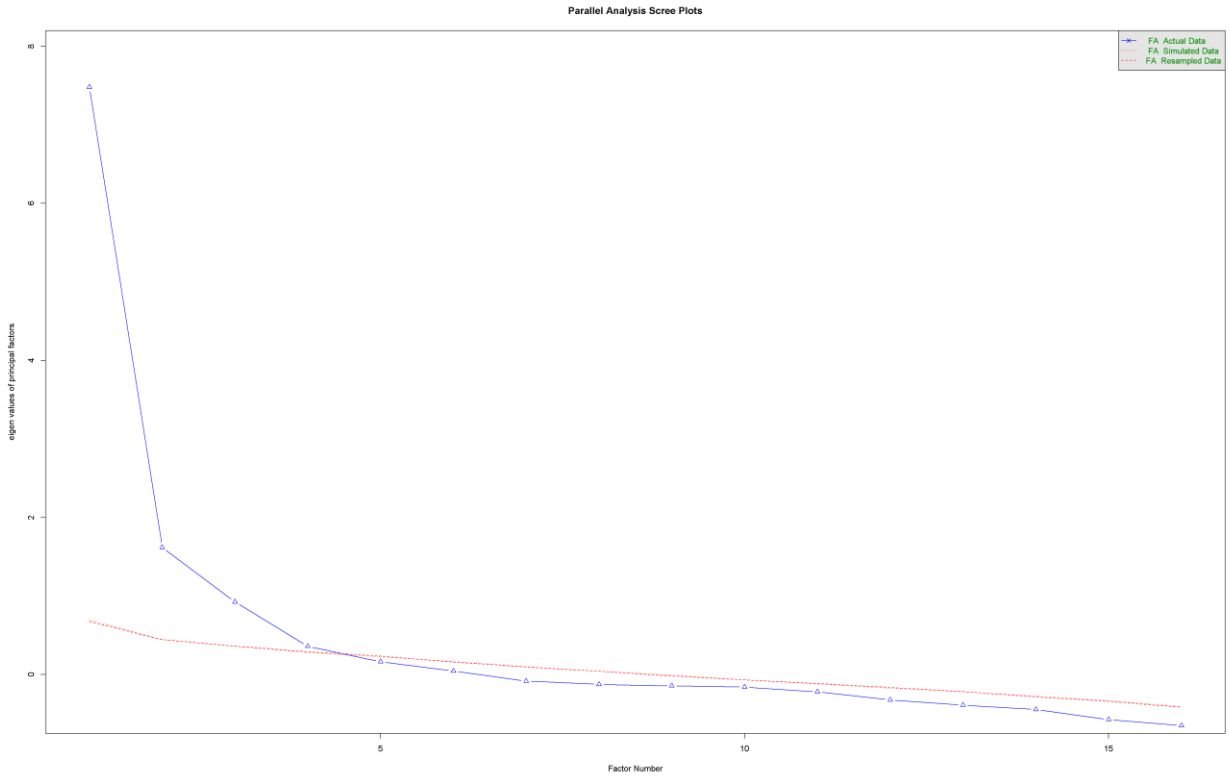
Let's now examine the dataset's variables' factorability. Let's start by creating a new dataset of all the independent variables and running the Kaiser-Meyer-Olkin (KMO) Test on the data.

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cm)
## Overall MSA =  0.84
## MSA for each item =
##               Infant_deaths           Under_five_deaths
##                        0.86                        0.83
##             Adult_mortality         Alcohol_consumption
##                        0.84                        0.83
##                Hepatitis_B                     Measles
##                        0.82                        0.97
##                         BMI                      Polio
##                        0.87                        0.93
##                 Diphtheria               Incidents_HIV
##                        0.80                        0.70
##              GDP_per_capita             Population_mln
##                        0.88                        0.85
## Thinness_ten_nineteen_years    Thinness_five_nine_years
##                        0.73                        0.72
##                   Schooling             Life_expectancy
##                        0.92                        0.87
```

With this data, factor analysis can be performed because MSA > 0.5. Additionally, the Bartlett's test of sphericity ought to be significant.

```
## $chisq
## [1] 3781.936
##
## $p.value
## [1] 0
##
## $df
## [1] 120
```

The suitability of factor analysis was assessed using the Bartletts Test measure of sample adequacy and the Kaiser-Meyer Olkin (KMO) measure. The approximate Chi-square value is 3781.936 with 120 degrees of freedom which is at the 0.05 Level of significance. Additionally, the KMO statistic of 0.84 is large (more than 0.50). Factor analysis is therefore seen as a suitable method for further analysis.



The graph indicates a significant shift in the scree plot's curvature following factor 3. This demonstrates that the total variance changes for decreasing amounts after factor 3.

Since p value which is equal to 9.683779e-78 is less than α (0.05), we reject H0. Let's try a case where we have 7 factors.

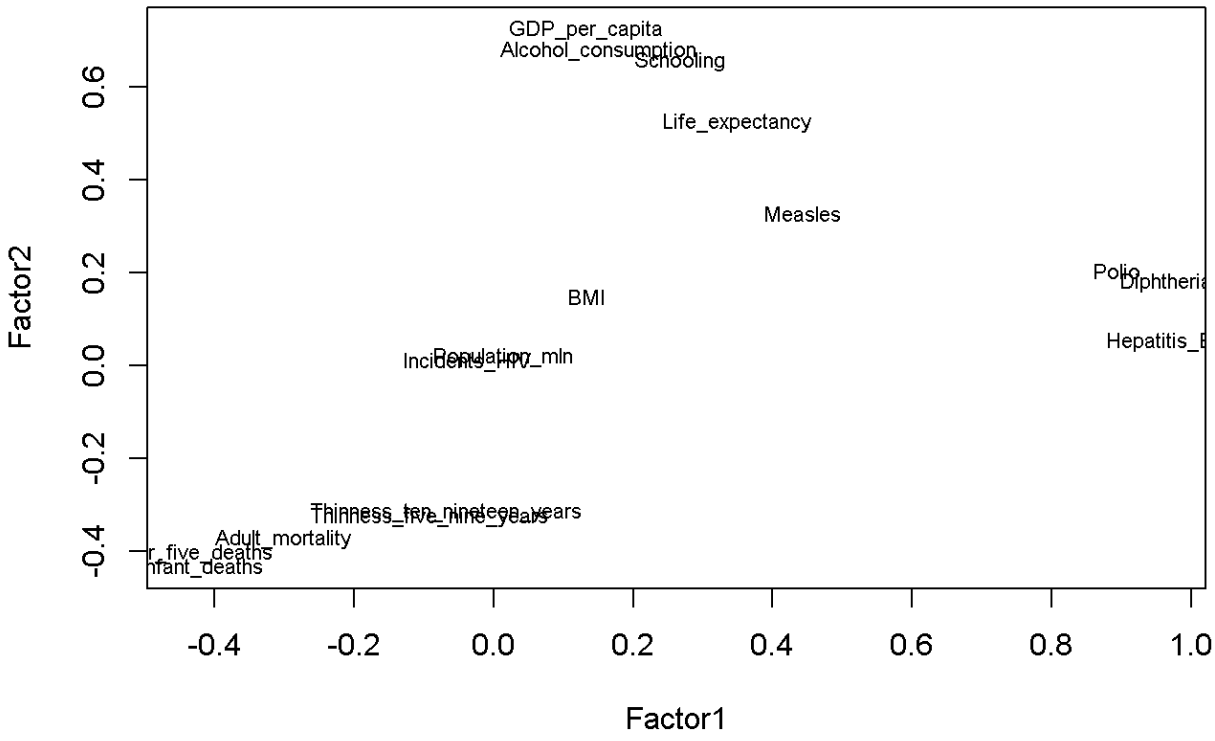
This time, since p value is greatest then the other number factor, but p value is less than α, so we reject null hypothesis and decide on that 7 factors solution is adequate. But we need to continue with 7 factor solution.

```
## Call:
## factanal(x = data2, factors = 7, lower = 0.0122222225)
##
## Uniquenesses:
##               Infant_deaths           Under_five_deaths
##                       0.012                       0.012
##             Adult_mortality         Alcohol_consumption
##                       0.012                       0.421
##                 Hepatitis_B                     Measles
##                       0.072                       0.554
##                         BMI                       Polio
##                       0.012                       0.103
##                  Diphtheria               Incidents_HIV
##                       0.016                       0.478
##              GDP_per_capita              Population_mln
##                       0.348                       0.871
## Thinness_ten_nineteen_years    Thinness_five_nine_years
##                       0.012                       0.012
##                    Schooling             Life_expectancy
##                       0.173                       0.021
##
## Loadings:
##                              Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
## Infant_deaths                -0.418  -0.433   0.433   0.146  -0.297   0.574
## Under_five_deaths            -0.440  -0.402   0.414   0.132  -0.320   0.586
## Adult_mortality              -0.301  -0.373   0.796          -0.245   0.249
## Alcohol_consumption           0.151   0.680          -0.229          -0.190
## Hepatitis_B                   0.955
## Measles                       0.443   0.329  -0.224           0.262  -0.123
## BMI                           0.133   0.149  -0.159  -0.409   0.855  -0.157
## Polio                         0.893   0.204  -0.144                  -0.166
## Diphtheria                    0.965   0.181  -0.117
## Incidents_HIV                                0.721
## GDP_per_capita                0.133   0.724  -0.297  -0.112
## Population_mln                                       0.342
## Thinness_ten_nineteen_years          -0.315   0.134   0.903           0.201
## Thinness_five_nine_years             -0.323   0.125   0.896  -0.131   0.154
## Schooling                     0.268   0.657  -0.174  -0.241   0.343  -0.339
## Life_expectancy               0.349   0.524  -0.623           0.243  -0.353
##                              Factor7
## Infant_deaths
## Under_five_deaths
## Adult_mortality
## Alcohol_consumption
## Hepatitis_B
## Measles
## BMI
## Polio
## Diphtheria
## Incidents_HIV
## GDP_per_capita
## Population_mln
## Thinness_ten_nineteen_years
## Thinness_five_nine_years     0.122
## Schooling
## Life_expectancy
##
##                 Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
## SS loadings       3.564   2.592   2.175   2.086   1.277   1.157   0.025
## Proportion Var    0.223   0.162   0.136   0.130   0.080   0.072   0.002
## Cumulative Var    0.223   0.385   0.521   0.651   0.731   0.803   0.805
##
## Test of the hypothesis that 7 factors are sufficient.
## The chi square statistic is 62.63 on 29 degrees of freedom.
## The p-value is 0.000288
```

Polio, diphtheria, and hepatitis B predominate the first factor. GDP_per_capita is reflected in a second factor. Life expectancy, HIV incidents, and adult mortality dominate the

third factor. BMI is reflected in the fourth factor. The Infant_deaths and Under_five_deaths dominate the sixth element. Thinness_five_nine_years is reflected in the seventh factor.

Furthermore, we observe that nearly 80.5% of the variance is explained by the first 7 components. Effectively, we may reduce the dimensionality from 16 to 7, losing approximately 19.5% of the variance in the process. Factor 2 accounts for 38.5% of the variation and factors 1 for 22.3% of the variance. As mentioned above, 80.5% of the variance in performance can be explained by all 7 components together.



As previously said, GDP_per_capita dominates Factor 2, while polio, diphtheria, and hepatitis B dominate Factor 1.

Now, we can calculate the alpha.

```
##
## Reliability analysis
##  raw_alpha std.alpha G6(smc) average_r S/N   ase mean sd median_r
##       0.89      0.92    0.96      0.67  12 0.011   95 16     0.59
```

The "raw-alpha" value for the overall factor is what matters to us. This indicates the general degree of consistency between the variables inside the factor. This alpha value is 0.89.

3. 6 DISCRIMINATION AND CLASSIFICATION

In order to categorize countries as developed or developing, our aim is to identify a function that divides them according to their values.

Structure of data:

```
## 'data.frame':    179 obs. of  17 variables:
##  $ Infant_deaths              : num  11.1 2.7 6.6 57 39.7 21.6 9.6 41.3 2.2 17.4
...
##  $ Under_five_deaths          : num  13 3.3 8.2 88 59.8 25.2 11.2 59 2.7 21.8 ...
##  $ Adult_mortality            : num  105.8 57.9 223 340.1 261.7 ...
##  $ Alcohol_consumption        : num  1.32 10.35 8.06 4.55 2.69 ...
##  $ Hepatitis_B                : num  97 97 97 84 97 95 99 69 88 97 ...
##  $ Measles                    : num  65 94 97 64 64 99 98 64 91 65 ...
##  $ BMI                        : num  27.8 26 26.2 24.3 23.9 25.5 26.3 21.3 26.6 21.
7 ...
##  $ Polio                      : num  97 97 97 77 96 95 99 68 95 97 ...
##  $ Diphtheria                 : num  97 97 97 84 97 95 99 69 95 97 ...
##  $ Incidents_HIV              : num  0.08 0.09 0.08 1.12 0.96 0.05 0.05 0.24 0.04
0.12 ...
##  $ GDP_per_capita             : num  11006 25742 9313 1383 661 ...
##  $ Population_mln             : num  78.53 46.44 144.1 23.3 2.09 ...
##  $ Thinness_ten_nineteen_years: num  4.9 0.6 2.3 5.6 7.3 6 7.1 7.1 0.8 14.2 ...
##  $ Thinness_five_nine_years   : num  4.8 0.5 2.3 5.5 7.2 5.8 6.9 7.1 0.7 14.5 ...
##  $ Schooling                  : num  7.8 9.7 12 6.1 3.4 7.9 9.5 6.1 12.5 8 ...
##  $ Life_expectancy            : num  76.5 82.8 71.2 57.6 60.9 76.1 76.9 65.5 82.3 7
5.1 ...
##  $ Economy_status             : Factor w/ 2 levels "Economy_status_Developed",..: 2
1 2 2 2 2 2 1 2 ...
```

The values of the loadings of the discriminant functions:

```
## Call:
## lda(Economy_status ~ ., data = train)
##
## Prior probabilities of groups:
##  Economy_status_Developed Economy_status_Developing
##                 0.2255639                 0.7744361
##
## Group means:
##                          Infant_deaths Under_five_deaths Adult_mortality
## Economy_status_Developed       3.573333           4.29000        79.24293
## Economy_status_Developing     29.301942          39.26796       189.34433
##                          Alcohol_consumption Hepatitis_B  Measles      BMI
## Economy_status_Developed            9.727333    91.30000 89.93333 26.44333
## Economy_status_Developing           3.353592    86.12621 76.49515 25.20777
##                            Polio Diphtheria Incidents_HIV GDP_per_capita
## Economy_status_Developed  94.86667   95.23333     0.0750000      36783.267
## Economy_status_Developing 86.25243   85.72816     0.8687379       6456.495
##                          Population_mln Thinness_ten_nineteen_years
## Economy_status_Developed        32.17500                    1.233333
## Economy_status_Developing       50.33204                    5.362136
##                          Thinness_five_nine_years Schooling Life_expectancy
## Economy_status_Developed                 1.173333 12.146667        80.21667
## Economy_status_Developing                5.476699  7.256311        68.95728
##
## Coefficients of linear discriminants:
##                                    LD1
## Infant_deaths                2.035438e-02
## Under_five_deaths           -5.244384e-02
## Adult_mortality              2.219491e-03
## Alcohol_consumption         -2.546864e-01
## Hepatitis_B                  5.255349e-02
## Measles                      1.317721e-02
## BMI                          4.961314e-02
## Polio                       -2.223254e-03
## Diphtheria                  -4.900140e-02
## Incidents_HIV               -8.602644e-03
## GDP_per_capita              -2.145562e-05
## Population_mln               2.557071e-04
## Thinness_ten_nineteen_years  1.158591e-02
## Thinness_five_nine_years     1.781007e-02
## Schooling                   -1.635261e-01
## Life_expectancy             -1.411105e-01
```

According to the LDA output, πEconomy_status_Developing = 0.7744361 and πEconomy_status_Developed = 0.2255639, meaning that 22.6% of the training observations relate to developed countries.

0.02035438$*$Infant_deaths…-0.1411105$*$Life_expectancy is the model. The LDA classifier will predict that the country is developed if this is large, and it will predict that the country is developing if it is little.

Plot of the linear discriminants obtained from the equation:

The classification of each and every combination in the training data set:

Train performance:

```
##                          Actual
## Predicted                 Economy_status_Developed Economy_status_Developing
##   Economy_status_Developed                       28                       2
##   Economy_status_Developing                       2                     101
```

The accuracy of the model is 0.9699248

Test performance:

```
##                          Actual
## Predicted                 Economy_status_Developed Economy_status_Developing
##   Economy_status_Developed                        7                         1
##   Economy_status_Developing                       0                        38
```

The accuracy of the model is 0.9782609

For the test data, the model accurately classifies the countries with a probability of 0.98.

For test data, the misclassification rate is 1−0.98=0.02.

3. 7 CLUSTERING

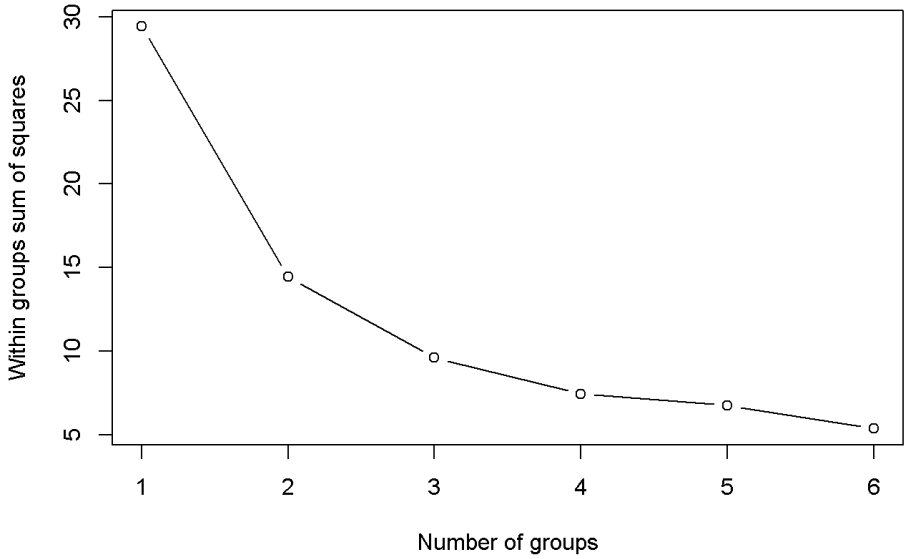
We will now apply k-means clustering to the patient rate data. When we first compute the variances of the patient rates for the various categories of disease, we discover the following:

```
##   Hepatitis_B       Measles         Polio    Diphtheria Incidents_HIV
##    200.720168    261.986567    169.632917    215.908669      2.628673
```

We must standardize the data in some way, and in this case, we standardize each variable by its range. The variances are highly varied, thus applying k-means on the raw data would not make sense. The variances after this standardization are as follows:

```
##   Hepatitis_B       Measles         Polio    Diphtheria Incidents_HIV
##    0.03385397    0.04306157    0.04412927    0.03134108    0.01287277
```

Since the standardized data's variances are so similar, we can move forward with clustering the data. To determine the number of groups, we first plot the within-groups sum of squares for solutions involving one to six groups.



Since there is only "elbow" in the plot for two groups, we will now examine the two-group solution.

```
##   Hepatitis_B  Measles    Polio Diphtheria Incidents_HIV
## 1    92.98551 67.87514 21.67550   88.63017     3.2454691
## 2    68.16661 66.88274 85.45122   49.59271     0.7814634
```

The cluster number for each state is the following:

```
##    1    2    7   28   44   58   75  102  111  113  122  161  167  174  183  203
##    2    2    2    1    2    2    2    1    2    2    2    2    2    2    1    2
##  217  220  242  269  272  301  304  331  334  416  427  444  452  474  488  496
##    1    1    2    2    2    1    2    1    2    2    1    2    1    2    2    2
##  499  512  547  557  559  580  585  591  599  609  627  629  639  640  651  688
##    1    2    1    2    2    2    2    2    2    2    2    2    2    2    2    2
##  697  700  717  728  733  741  764  766  770  787  810  811  833  838  840  850
##    2    2    2    2    2    2    2    2    2    1    2    1    2    2    1    2
##  852  853  869  874  901  903  904  907  957 1000 1011 1058 1064 1067 1068 1117
##    1    2    2    2    2    2    2    2    1    2    1    2    2    1    2    1
## 1146 1150 1157 1158 1164 1197 1205 1206 1225 1227 1251 1262 1266 1306 1320 1338
##    2    2    2    1    2    2    2    2    2    2    2    2    2    2    2    1
## 1358 1481 1492 1503 1531 1536 1559 1566 1588 1604 1614 1620 1624 1629 1707 1710
##    2    2    2    1    2    2    2    2    2    2    2    1    2    1    1    2
## 1713 1716 1722 1732 1785 1801 1810 1820 1824 1844 1875 1882 1905 1948 1950 1976
##    2    2    2    1    2    2    2    1    2    2    1    2    2    2    2    1
## 1984 1989 2022 2077 2081 2118 2120 2127 2128 2146 2148 2178 2182 2235 2279 2281
##    2    2    2    1    2    2    2    2    2    2    1    2    2    1    2    2
## 2291 2322 2324 2325 2339 2357 2368 2379 2396 2405 2410 2443 2449 2469 2496 2519
##    2    2    1    2    2    2    2    2    2    2    2    2    2    2    1    2
## 2526 2582 2618 2621 2642 2651 2662 2680 2702 2711 2713 2728 2738 2753 2754 2821
##    1    1    2    2    1    2    1    1    2    1    2    2    2    1    2    2
## 2841 2847 2849
##    2    2    2
```

4. Discussion/Conclusion

As a result of the analyses carried out in this project, we received effective answers to our questions. We asked 4 questions for more detailed inferences about the data, but these are questions where more than one piece of information is asked. As a result of our questions, we try to understand the reason for the high correlation between Thinness in the 5-9 age group and Thinnes in the 10-19 age group and how it is related to the economic status of regions or countries in the world. Also, we tried to investigate the impact of Measles, Hepatitis B, Polio and Diphtheria vaccines, as well as conditions that negatively affect health such as alcohol consumption, BMI and HIV cases, on life expectancy. First, we tried many methods to find the multivariate normality of our data, such as univariate normal variables or transformation methods. However, we could not examine multivariate normality and we accepted that and started our tests. Our first problem was when we examined the relationship between two age groups by looking at their economic status, we collected sufficient evidence to prove that the economic status had an effect on the average of the answers, and then we continued the comparison of multivariate averages by determining the regions as factors, and in this part we discovered that the region is important. Then, in the continuation of the multivarate analysis, it is clearly seen that there is a significant difference between the two variables with the variance and covariance matrix of the combined effect of the two factors depending on the region or economic situation. We set out to do Principal Component Analysis and Regression. This was a part that we paid attention to in order to understand the impact of diseases and factors those diseases on life expectancy in our data set, which contains many variables, in order to understand the numerical data, we selected among the variables and to transform the data into fewer but essentially representative variables. In summary, we discovered that the first six components account for more than 98.4% of the variability. With a regression-ready inference, we concluded that the model was significant, the components explained approximately 89% of the y variability, and that the sixth component, which provided the most variance, could be considered significant. As a result of factor analysis, the first factor is mostly caused by hepatitis B, diphtheria, and polio. A second factor reflects GDP_per_capita. The third factor is dominated by adult mortality, HIV incidence, and life expectancy. The fourth factor takes BMI into consideration. The sixth factor is dominated by the Infant_deaths and Under_five_deaths. The seventh factor reflects thinness_five_nine_years. Moreover, we note that the first seven factors account for almost 80.5% of the variance. We can effectively reduce the dimensionality from 16 to 7, which will result in a loss of about 19.5% of the process variance. 38.5% of the variance is explained by factor 2, and 22.3% by factor 1. As previously indicated, the combined

effect of all 7 factors explains 80.5% of the variance in performance. Also, the overall level of consistency among the factors' variables is 0.89. In the discrimination and classification, we classify the countries as economically developing or developed by LDA classifier and by using some measurements of the countries. The classifier classifies the countries with a probability of 0.98 accuracy. Finally, we apply k-means clustering to the patient rate data to cluster the countries. We will now examine the two-group solution since there is only "elbow" in the plot for two cluster. 41 out of 179 countries were clustered as cluster 1, while the rest were clustered as cluster 2.

References

Brownlee, J. (2019, August 8). *How to Transform Data to Better Fit The Normal Distribution*. MachineLearningMastery.com. https://machinelearningmastery.com/how-to-transform-data-to-fit-the-normal-distribution/

Coder, R. (2021, December 5). *Box Cox transformation in R*. R CODER. https://r-coder.com/box-cox-transformation-r/

ODTUCLASS 2023-2024 FALL: Multivariate Analysis (n.d.). https://odtuclass2023f.metu.edu.tr/pluginfile.php/521231/mod_resource/content/5/RC8.html

R: Hotelling's T2 Test. (n.d.). https://search.r-project.org/CRAN/refmans/DescTools/html/HotellingsT.html

S. P. (2018, January 19). *MANOVA using R*. YouTube. https://www.youtube.com/watch?v=2lrcdToCOSs

J. (2022, October 23). *Box Cox transformation in R | R-bloggers*. R-bloggers. https://www.r-bloggers.com/2022/10/box-cox-transformation-in-r/