Multivariate Analysis on Life Expectancy Data

Yehya Abdelmohsen

3/24/2022

Contents

Introduction

An important metric that represents the overall health of a population is the life expectancy. Life expectancy is the average age of death. It is a more general metric than infant and child mortality and captures mortality along the entire life course, Max Roser and Ritchie (2013). According to Max Roser and Ritchie (2013) since 1900 the average life expectancy has more than doubled and is currently above 70 years old. Higher life expectancy may also indicate higher quality of life. After a thorough analysis of the data, we find that the average life expectancy in developed countries is higher than developing countries.

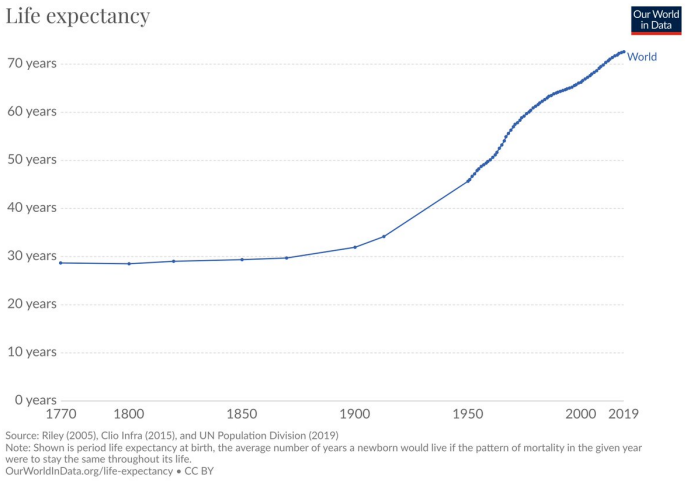


Figure 1: Life expectancy from 1770 to 2019

About the Data

The data presented contains information related to life expectancy for 193 countries. It was collected over 16 years, from 2001-2015. It contains 2938 observations and 22 variables. After an initial look at the data we find some missing values and observations.

The data was collected from the World Health Organization (WHO) and the United Nations (UN) websites with the help of Deeksha Russell and Duan Wang.

Source: www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Variables

- Country - Country name as a string e.g. Afghanistan.
- Year - The year the data was collected. It is a quantitative variable.
- Status - The status of the country. Either developing or developed. It is a categorical variable.
- Life Expectancy - The life expectancy in age. It is a quantitative variable.
- Adult Mortality - The number of adults (between 15 years old and 60 years old) that are expected to die out of a 1000 people population. This is a quantitative variable.

- Infant Deaths - The number of infant deaths out of a 1000 people population. This is a quantitative variable.
- Alcohol - Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
- Percentage Expenditure - Expenditure on health as a percentage of Gross Domestic Product (GDP) per capita. This is a quantitative variable.
- Hepatitis B - Percentage of immunization among 1 year olds for Hepatitis B. This is a quantitative variable.
- Measles - The number of reported cases for measles from a 1000 people population. This is a quantitative variable.
- BMI - The average of the Body Mass Index (BMI) for the whole population. This is a quantitative variable.
- Under Five Deaths - The number of deaths under five years old out of a 1000 people population. This is a quantitative variable.
- Polio - Percentage of immunization among 1 year olds for Polio. This is a quantitative variable.
- Total Expenditure - General government expenditure on health as a percentage of total government expenditure. This is a quantitative variable.
- Diphtheria - Percentage of immunization among 1 year olds for Diphtheria tetanus toxoid and pertussis (DTP3). This is a quantitative variable.
- HIV/AIDS - Deaths per 1000 live births due to HIV/AIDS (0-4 years). This is a quantitative variable.
- GDP - Gross Domestic Product per capita in US Dollars. This is a quantitative variable.
- Population - Population of the country. This is a quantitative variable.
- Thinness 1-19 years - Prevalence of thinness among children and adolescents for age 10 to 19 as a percentage. This is a quantitative variable.
- Thinness 5-9 years - Prevalence of thinness among children for age 5 to 9 as a percentage. This is a quantitative variable.
- Income Composition of Resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1). This is a quantitative variable.
- Schooling - Number of years of schooling. This is a quantitative variable.

Some Remarks Regarding the Data

Below is an initial look at the data.

Country	Year	Status	Life.expectancy	Adult.Mortality	infant.deaths	Alcohol	percentage.expenditure
Afghanistan	2015	Developing	65.0	263	62	0.01	71.27962
Afghanistan	2014	Developing	59.9	271	64	0.01	73.52358

Hepatitis.B	Measles	BMI	under.five.deaths	Polio	Total.expenditure	Diphtheria	HIV.AIDS
65	1154	19.1	83	6	8.16	65	0.1

Hepatitis.B	Measles	BMI	under.five.deaths	Polio	Total.expenditure	Diphtheria	HIV.AIDS
62	492	18.6	86	58	8.18	62	0.1

HIV.AIDS	GDP	Population	thinness..1.19.years	thinness.5.9.years	Income.composition.of.resources	Schooling
0.1	584.2592	33736494	17.2	17.3	0.479	10.1
0.1	612.6965	327582	17.5	17.5	0.476	10.0

In our analysis, we will only focus on observations collected in 2014. Therefore, some countries with no observations in 2014 will be omitted. The remaining observations are $n = 183$. Ten countries of the 193 in the data are not included in the analysis.

In addition, we will omit columns with missing values and the column that contains the year. We are left with $p = 11$. Below is a look at the first three rows of the final data we will be using. Of the 11 variables, 9 are quantitative, 1 is binary, and 1 contains the country name. The binary column Status, which represent whether a country is "Developed" or "Developing" will be converted to 0s and 1s. The original column will be dropped. The column Status is now called Developing.

Country	LifeExp	AdultMort	InfDeaths	PercExp	Measles
Afghanistan	59.9	271	64	73.52358	492
Albania	77.5	8	0	428.74907	0
Algeria	75.4	11	21	54.23732	0

UnderFive	Polio	Diphtheria	HIV.AIDS	Developing
86	58	62	0.1	1
1	98	98	0.1	1
24	95	95	0.1	1

Problem Statement

The data contains multiple variables that may influence life expectancy. Through the analysis of the data we have three goals:

- Identify the variables that have the largest influence on life expectancy.
- Identify countries that may be considered outliers.
- Analyze the effect the variables have on each other.

Descriptive Statistics

Keep in mind that we will repeat the below analysis once our data is free from outliers since the mean and variance-covariance matrix are affected by outliers.

We can see a very clear negative linear relationship between life expectancy and adult mortality, which is expected. As the number of adult deaths decrease, the life expectancy decreases. There doesn't seem to be a relationship between life expectancy and percentage expenditure on health, however, there may still be a relationship after transformation.

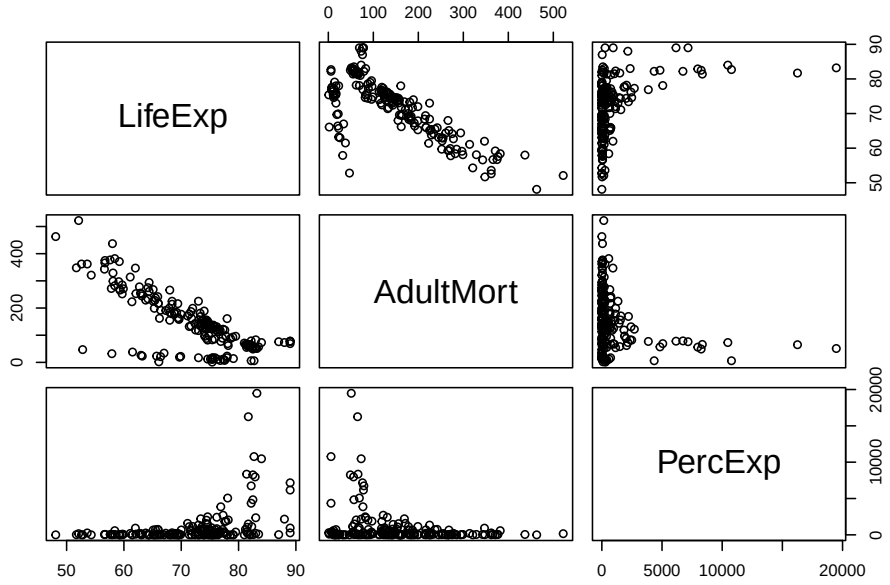


Figure 2: LifeExp vs AdultMort vs PercExp

Now after using the log transformation on PercExp, we observe a linear relationship. We also find that the variance seems to be non constant due to the cone shape of the graph.

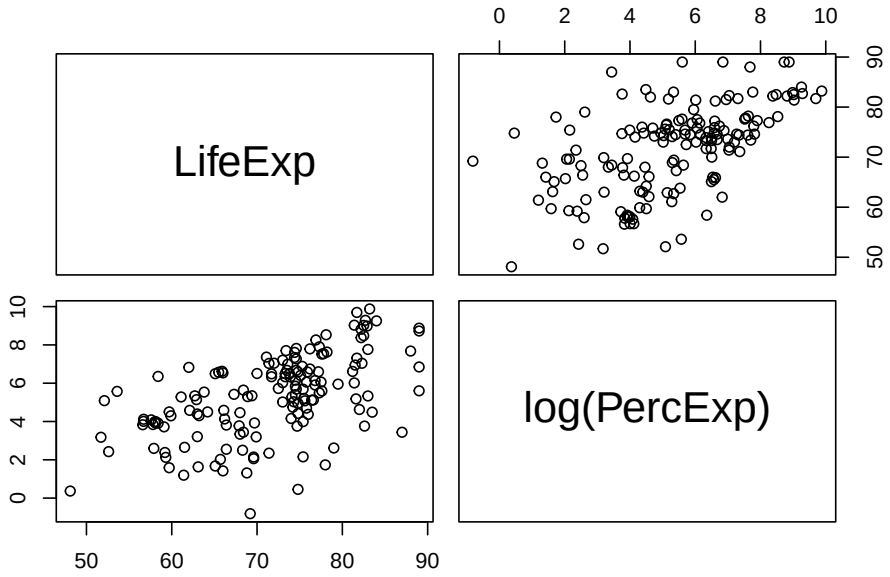


Figure 3: LifeExp vs log(PercExp)

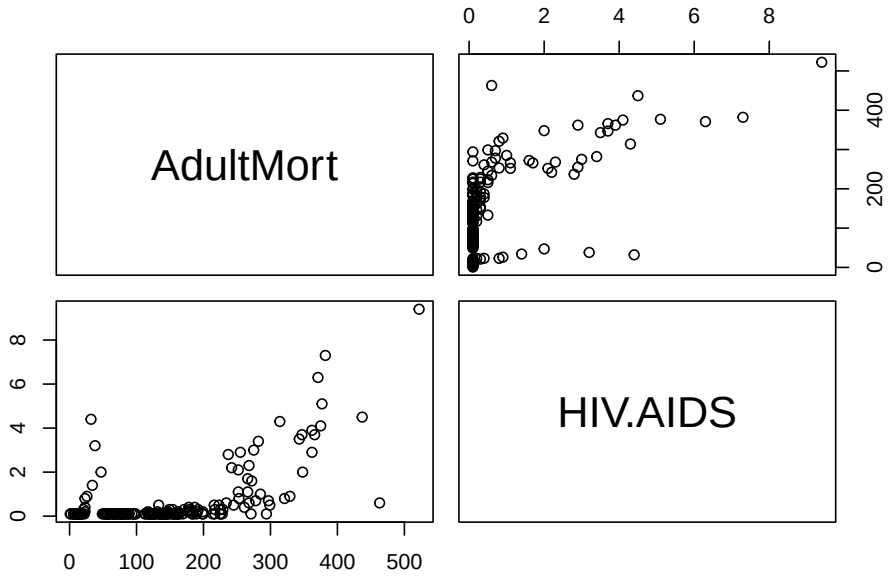


Figure 4: AdultMort vs HIV.AIDS

Firstly, we compute the vector of means. We find that the average life expectancy is around 71.5 years. The average adult mortality is 148.7, meaning that out of a 1000 person population of adults (15-60 years old), we expect 149 to die.

Table 6: Vector of Means

LifeExp	AdultMort	InfDeaths	PercExp	Measles	UnderFive	Polio	Diphtheria	HIV.AIDS	Developing
71.53661	148.6885	24.55738	1001.913	1831.208	32.89071	84.72678	84.08197	0.6819672	0.8251366

The number of infant deaths out of 1000 infants is expected to be 25. The number of deaths out of a population of 1000 people under the age of five is 33. Obviously the mean of deaths for the variable UnderFive is larger than the mean of InfDeaths. This is because infants (0-1 years old) are under the age of five and thus are included in the variable under five. This means that all values of InfDeaths are less than or equal to values of UnderFive, which is the case. In addition, when looking at Figure 5 we see a perfect linear relationship. This is expected since the values are the same in multiple cases and in other cases just shifted to the right by a couple values.

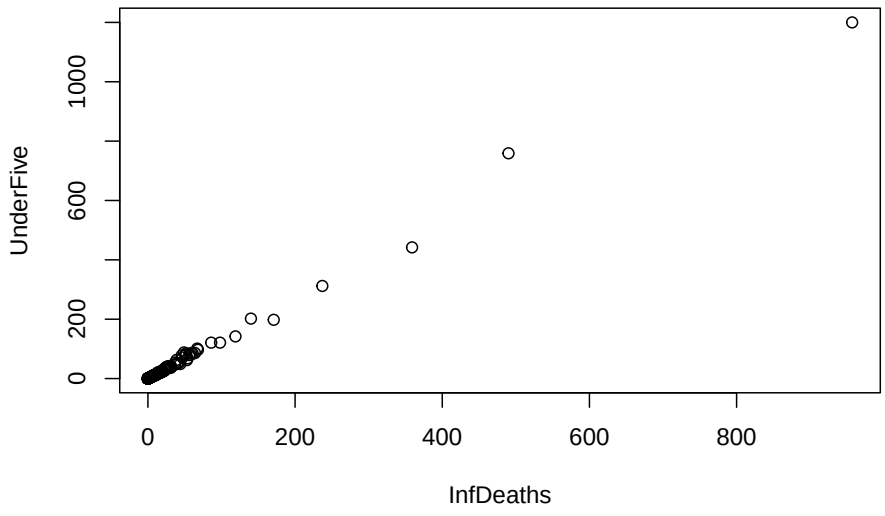


Figure 5: InfDeaths vs Underfive

A peculiar value is the percentage expenditure, which is the expenditure on health as a percentage of Gross Domestic Product (GDP) per capita. This value should be between 0 and 100, however, that is not the case. In addition, the average of the variable measles is over 1000. This should not be the case. The variable measles is defined as the number of reported cases for measles from a 1000 people population, therefore, its average should be less than or equal to 1000.

The variables Polio and Diphtheria represent the percentage of immunization among 1 year olds for these diseases. The values are close to each other. The average for Polio and Diphtheria respectively are 84.7 and 84.1. We can see in the figure Diphtheria vs Polio there seems to be a positive relationship. This may be because the two shots are administered at the same time.

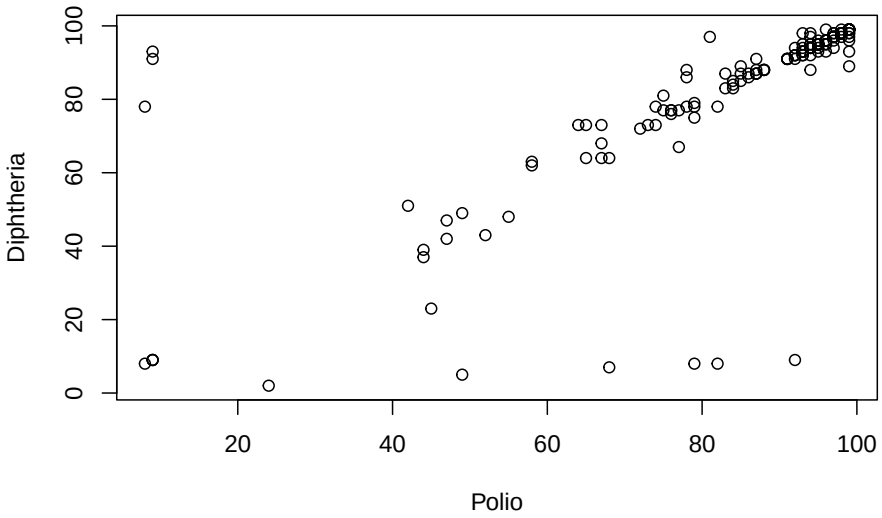


Figure 6: Diphtheria vs Polio

We also look at the mean of the variable HIV.AIDS, which is the deaths per 1000 births due to HIV/AIDS. It is 0.7, which is very small. Finally, the mean of the variable developing is 0.83, which means that around 83% of the countries in the data are developing countries.

Table 7: Correlation Matrix

	LifeExp	AdultMort	InfDeaths	PercExp	Measles	UnderFive	Polio	Diphtheria	HIV.AIDS	Developing
LifeExp	1.00	-0.76	-0.24	0.40	-0.10	-0.26	0.41	0.38	-0.61	-0.52
AdultMort	-0.76	1.00	0.18	-0.25	0.09	0.20	-0.38	-0.30	0.63	0.32
InfDeaths	-0.24	0.18	1.00	-0.10	0.70	1.00	-0.13	-0.13	0.09	0.12
PercExp	0.40	-0.25	-0.10	1.00	-0.07	-0.10	0.08	0.08	-0.13	-0.42
Measles	-0.10	0.09	0.70	-0.07	1.00	0.68	-0.05	-0.10	-0.03	0.09
UnderFive	-0.26	0.20	1.00	-0.10	0.68	1.00	-0.14	-0.14	0.12	0.13
Polio	0.41	-0.38	-0.13	0.08	-0.05	-0.14	1.00	0.75	-0.34	-0.18
Diphtheria	0.38	-0.30	-0.13	0.08	-0.10	-0.14	0.75	1.00	-0.22	-0.18
HIV.AIDS	-0.61	0.63	0.09	-0.13	-0.03	0.12	-0.34	-0.22	1.00	0.19
Developing	-0.52	0.32	0.12	-0.42	0.09	0.13	-0.18	-0.18	0.19	1.00

Looking at the correlation matrix R, we find some obvious as well as some weird values. A value we expect to see is the negative correlation between LifeExp and AdultMort which is -0.76 since we already identified this relationship from Figure 2.

In Figure 7 there seems to be a linear relationship, however, the points close to 0 for HIV.AIDS have a large influence on the regression line. These two variables have a correlation of -0.61. This relationship is obvious and is similar to LifeExp vs AdultMort.

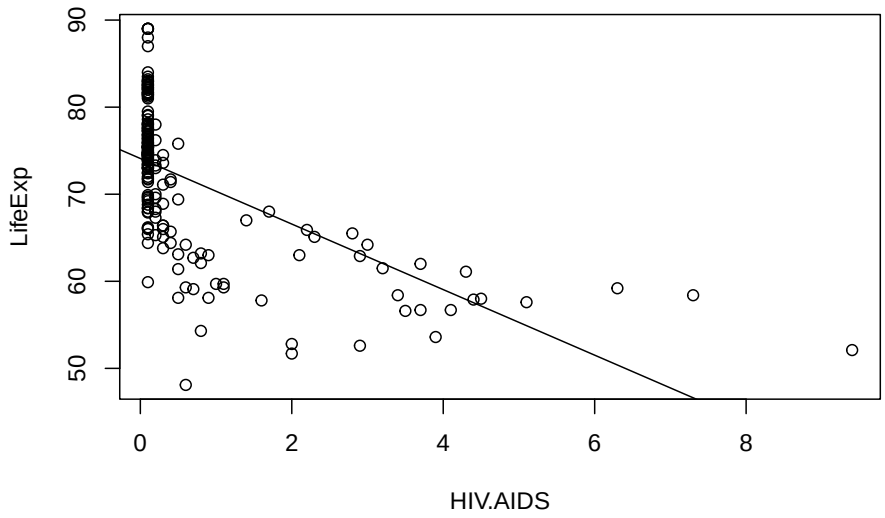


Figure 7: HIV.AIDS vs LifeExp

In Figure 8 we can see that the mean of LifeExp in developing countries is less than that of developed countries. In addition, the number of developed countries is far less than the number of developing countries.

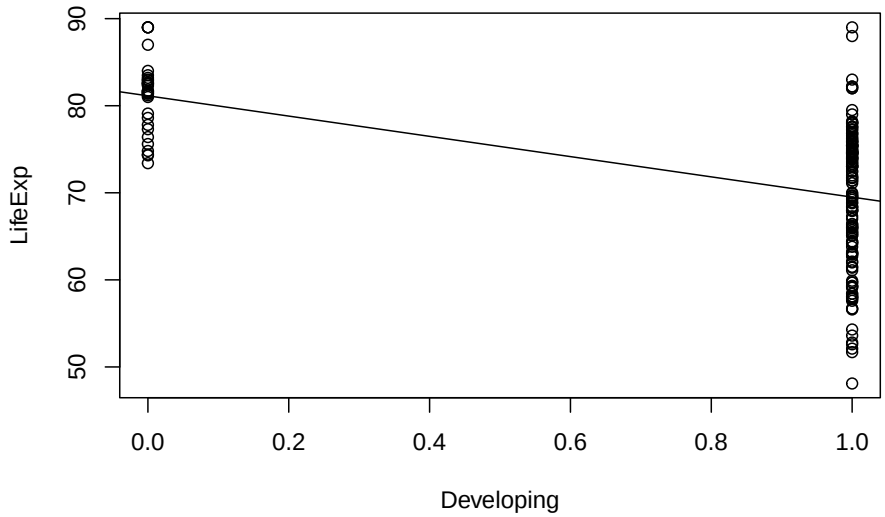


Figure 8: LifeExp vs Developing

The correlation coefficient between Measles and InfDeaths is 0.7. This value is extremely influenced by outliers. One very obvious outlier in Figure 9 is India with 8000 deaths. Keep in mind that this value should be the number of deaths due to measles from a 1000 person population.

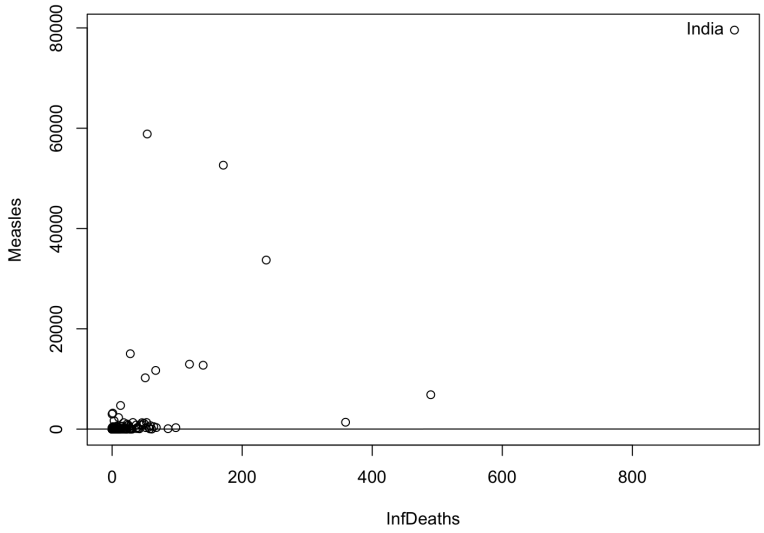


Figure 9: InfDeaths vs Measles

Outliers

Mahalanobis Distances

We will be using the function below to compute the mahalanobis distances for all observations.

```
mahalanobis_dis=function(df){
  dist = c()
  sigma_inv = solve(cov(df))
  xbar = colMeans(df)
  for(i in 1:nrow(df)){
    x_xbar = as.matrix(df[i,]-xbar)
    dist = c(dist,sqrt(x_xbar%*%(sigma_inv)%*%t(x_xbar)))
  }
  return(dist)
}
```

```
MD = mahalanobis_dis(df_2014[,2:11])
df_2014$MD = MD
```

Now we have the mahalanobis distances for all observations as a column in the data. They are sorted and the first 20 largest MDs are shown in Table 8. Not all columns are included in Table 8 for simplicity. We can see that 17 of the 20 largest observations are developing countries. Also, the three developed countries Switzerland, Luxembourg, and Iceland all have a life expectancy over 80 years old. They are also all European countries and are all relatively small. In addition, we observe some countries with very low life expectancy. Countries like Nigeria, Côte d'Ivoire, Sierra Leone, and Lesotho all have a life expectancy of around 50 years. These are all African countries.

Table 8: Data with Mahalanobis Distances

Country	LifeExp	Developing	MD
Nigeria	53.6	1	12.487591
India	68.0	1	11.789939
Philippines	68.4	1	9.409624
Lesotho	52.1	1	8.185775
Switzerland	83.2	0	7.495350
Pakistan	66.2	1	7.113682
China	75.8	1	6.864651
Equatorial Guinea	57.9	1	6.329881
Dominican Republic	73.6	1	6.219381
Luxembourg	81.7	0	6.205190
Indonesia	68.9	1	5.845290
Democratic Republic of the Congo	59.3	1	5.807163
Swaziland	58.4	1	5.613041
Republic of Moldova	71.8	1	5.551212
Iceland	82.5	0	5.455031
Tonga	73.3	1	5.078669
Côte d'Ivoire	52.8	1	4.876560
Gabon	65.5	1	4.787985
Sierra Leone	48.1	1	4.721311
Panama	77.6	1	4.694914

In Figure 10, we can see the ellipse created using the variables AdultMort and LifeExp. We can see clearly two lines created by the scatter of points, one of the lines cuts the ellipse at two extremes and coincides with the major axis. The other line is surrounded by less points and does not coincide with the major axis. These could be two subsets of countries in the data set.

The length of the major axis is $2d\sqrt{\lambda_1} = 2(4.577)\sqrt{11283.708} = 972.3818$. The length of the minor $2d\sqrt{\lambda_2} = 2(4.577)\sqrt{30.993} = 50.96217$.

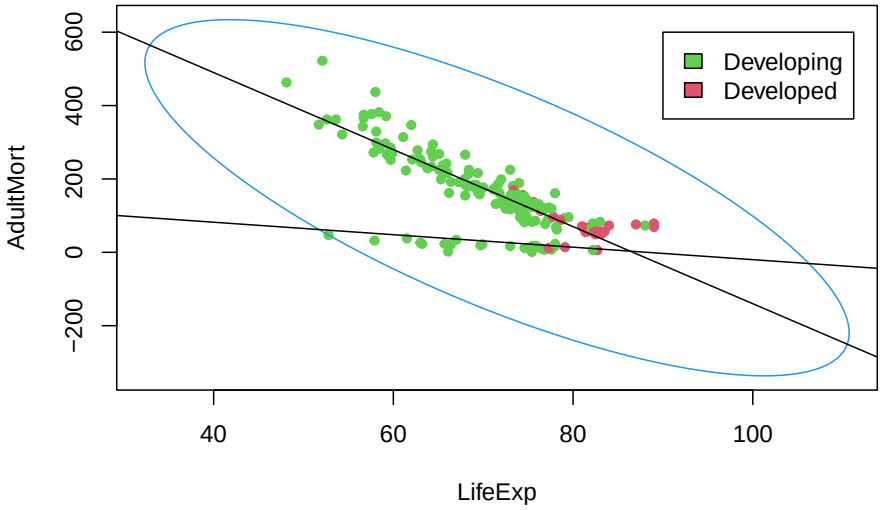


Figure 10: AdultMort vs LifeExp

BACON

Finally, we will use BACON to remove outliers from the data.

```
require(robustX); library(robustbase)
```

```
## Loading required package: robustX
```

```
x = as.matrix(df_2014[,2:10])
output = mvBACON(x)
```

```
## rank(x.ord[1:m,] >= p  ==> chosen m =  36
## MV-BACON (subset no. 1): 36 of 183 (19.67 %)
## MV-BACON (subset no. 2): 99 of 183 (54.1 %)
## MV-BACON (subset no. 3): 100 of 183 (54.64 %)
## MV-BACON (subset no. 4): 100 of 183 (54.64 %)
## MV-BACON (subset no. 5): 99 of 183 (54.1 %)
## MV-BACON (subset no. 6): 99 of 183 (54.1 %)
```

Six iterations of BACON were done and the final subset contains 99 observations out of 183 observations. That is 54.1% the data. We may reach multiple conclusions from this. First, this may mean that our data quality is bad. There may be data points, which were incorrectly recorded, for example. Second, this could mean that our data should be split into two subsets. This means that the world is split into two groups. For example, the two groups could be developed and developing. Note that the variable Developing was

removed before running BACON since it resulted in a singular system.

We can see in Figure 11 that a lot of points are past the critical value.

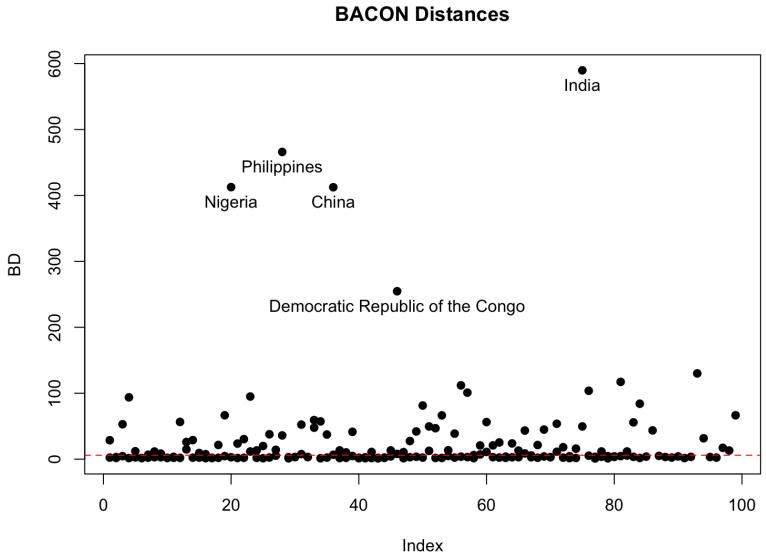


Figure 11: BACON Distances

Below are the observations with the highest bacon distances.

Table 9: Observations with the largest BD

	Country	LifeExp	AdultMort	InfDeaths	PercExp
119	Nigeria	53.6	362	490	263.21110
75	India	68.0	184	957	86.52154
36	China	75.8	86	171	109.87439
127	Philippines	68.4	214	54	31.27232
46	Democratic Republic of the Congo	59.3	266	237	0.00000

Table 10: Observations with the largest BD

	Measles	UnderFive	Polio	Diphtheria	HIV.AIDS	Developing	MD
119	6855	759	49	49	3.9	1	12.487591
75	79563	1200	84	85	0.2	1	11.789939
36	52628	198	99	99	0.1	1	6.864651
127	58848	68	77	67	0.1	1	9.409624
46	33711	312	79	8	1.1	1	5.807163

We can now look at our robust vector of means and correlation matrix to repeat our analysis.

We can see that the average life expectancy increased to 75.5 years, the non-robust was 71.5 years. Adult mortality decreased from 149 deaths out of a 1000 person population of adults to 104. Infant deaths and deaths under five both had a large drop from 25 and 33 to 3 and 4. Therefore, we conclude that these variables were heavily influenced by the outliers.

Table 11: Robust Vector of Means

LifeExp	AdultMort	InfDeaths	PercExp	Measles	UnderFive	Polio	Diphtheria	HIV.AIDS
75.47879	103.6566	2.959596	623.2713	72.29293	3.515151	93.40404	93.60606	0.1323232

Table 12: Robust Correlation Matrix

	LifeExp	AdultMort	InfDeaths	PercExp	Measles	UnderFive	Polio	Diphtheria	HIV.AIDS
LifeExp	1.00	-0.56	-0.17	0.27	0.22	-0.18	0.25	0.28	-0.36
AdultMort	-0.56	1.00	0.04	-0.09	-0.23	0.05	-0.20	-0.19	0.47
InfDeaths	-0.17	0.04	1.00	-0.19	0.27	1.00	-0.02	0.02	-0.01
PercExp	0.27	-0.09	-0.19	1.00	-0.08	-0.19	0.05	0.08	-0.04
Measles	0.22	-0.23	0.27	-0.08	1.00	0.29	0.07	0.07	-0.14
UnderFive	-0.18	0.05	1.00	-0.19	0.29	1.00	-0.03	0.02	0.02
Polio	0.25	-0.20	-0.02	0.05	0.07	-0.03	1.00	0.97	-0.14
Diphtheria	0.28	-0.19	0.02	0.08	0.07	0.02	0.97	1.00	-0.11
HIV.AIDS	-0.36	0.47	-0.01	-0.04	-0.14	0.02	-0.14	-0.11	1.00

Hotelling's T^2

Normality

Obviously we can see in Figure 12 that the majority of variables are not normally distributed, however, we will continue with computing Hotelling's T^2.

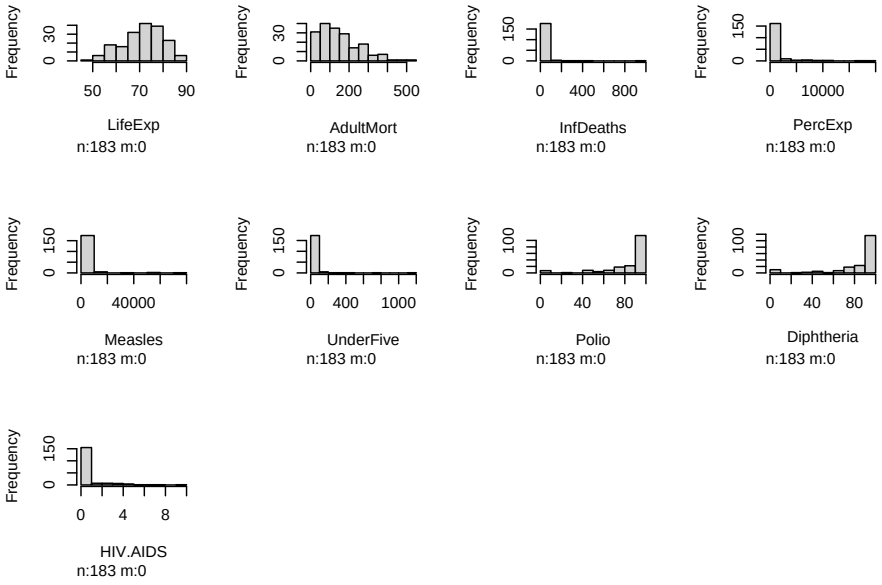


Figure 12: Histograms

With Outliers

We will be comparing two populations for Hotellings T^2. These populations are developed and developing countries.

$H_0 : \mu_{developed} = \mu_{developing}$
$H_1 : \mu_{developed} \neq \mu_{developing}$

```
out1 = ht2(df_2014[df_2014$Developing==1,2:10],df_2014[df_2014$Developing==0,2:10])
```

The number of countries in the two samples are $n_{developing} = 151$ and $n_{developed} = 32$. As can be seen in Table 11 and Table 12 the vector of means for both are very different.

Table 13: Developing Countries Vector of Means

LifeExp	AdultMort	InfDeaths	PercExp	Measles	UnderFive	Polio	Diphtheria	HIV.AIDS
69.50199	164.4305	29.49669	510.0458	2197.371	39.54305	82.98676	82.19205	0.805298

Table 14: Developed Countries Vector of Means

LifeExp	AdultMort	InfDeaths	PercExp	Measles	UnderFive	Polio	Diphtheria	HIV.AIDS
81.1375	74.40625	1.25	3322.909	103.375	1.5	92.9375	93	0.1

```
out1$pv
```

```
##              [,1]
## [1,] 1.660228e-12
```

The p-value is less than the significance level $\alpha = 0.05$. Thus, we reject the null hypothesis $H_0 : \mu_{developed} = \mu_{developing}$.

Without Outliers

```
df_2014_woo = df_2014[output$subset,]
```

```
out2 = ht2(df_2014_woo[df_2014_woo$Developing==1,2:10],df_2014_woo[df_2014_woo$Developing==0,2:10])
```

Table 15: Developing Countries Vector of Means (Without Outliers)

LifeExp	AdultMort	InfDeaths	PercExp	Measles	UnderFive	Polio	Diphtheria	HIV.AIDS
73.912	110.9733	3.386667	587.7759	59.01333	4.026667	92.74667	92.98667	0.1426667

Table 16: Developed Countries Vector of Means (Without Outliers)

LifeExp	AdultMort	InfDeaths	PercExp	Measles	UnderFive	Polio	Diphtheria	HIV.AIDS
80.375	80.79167	1.625	734.1944	113.7917	1.916667	95.45833	95.54167	0.1

```
out2$pv
```

```
##              [,1]
## [1,] 1.713182e-05
```

Again, in the case where there are no outliers the p-value is less than the significance level $\alpha = 0.05$. Thus, we reject the null hypothesis $H_0 : \mu_{developed} = \mu_{developing}$. In conclusion, the vector of means of the developed countries is different from that of the developing countries.

Transformation

```
tf = df_2014
tf$AdultMort = log(df_2014$AdultMort+1)
tf$InfDeaths = log(df_2014$InfDeaths+1)
tf$PercExp = log(df_2014$PercExp+1)
tf$Measles = log(df_2014$Measles+1)
tf$UnderFive = log(df_2014$UnderFive+1)
tf$Polio = (df_2014$Polio)^8
tf$Diphtheria = (df_2014$Diphtheria)^6
#tf = as.matrix(tf[,2:10])
```

```
out3 = ht2(tf[tf$Developing==1,2:10],tf[tf$Developing==0,2:10])
```

```
## Error in solve.default(Sp): system is computationally singular: reciprocal condition number = 9.0539
```

```
tfm = as.matrix(tf[,2:10])
```

```
eig = eigen(t(tfm)%*%tfm)
```

Here I was trying to find out why my transformation did not work, however, I was not able to finish due to time constraints.

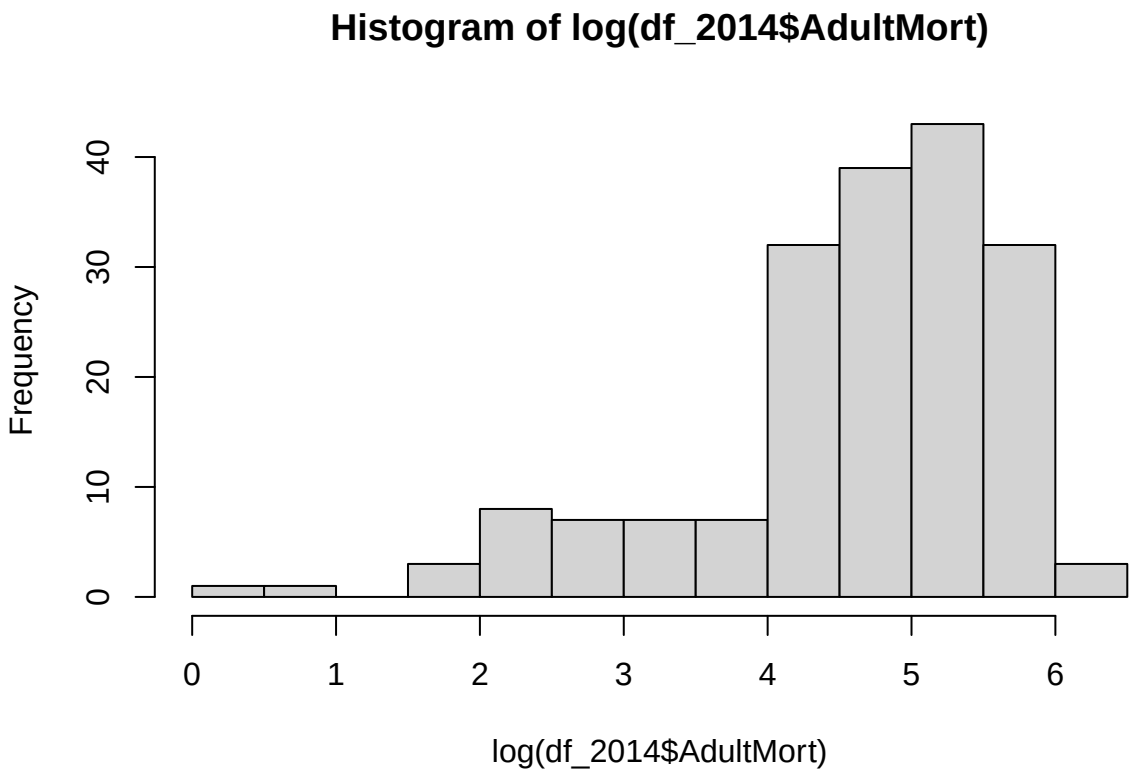
```
cat(sqrt(eig$values[1]/eig$values[2]),sqrt(eig$values[1]/eig$values[3]),sqrt(eig$values[1]/eig$values[4
```

```
## 46729.48 4.955632e+12 1.958732e+14 1.932357e+15 3.830266e+15 3.830266e+15 4.563936e+15 4.563936e+15
```

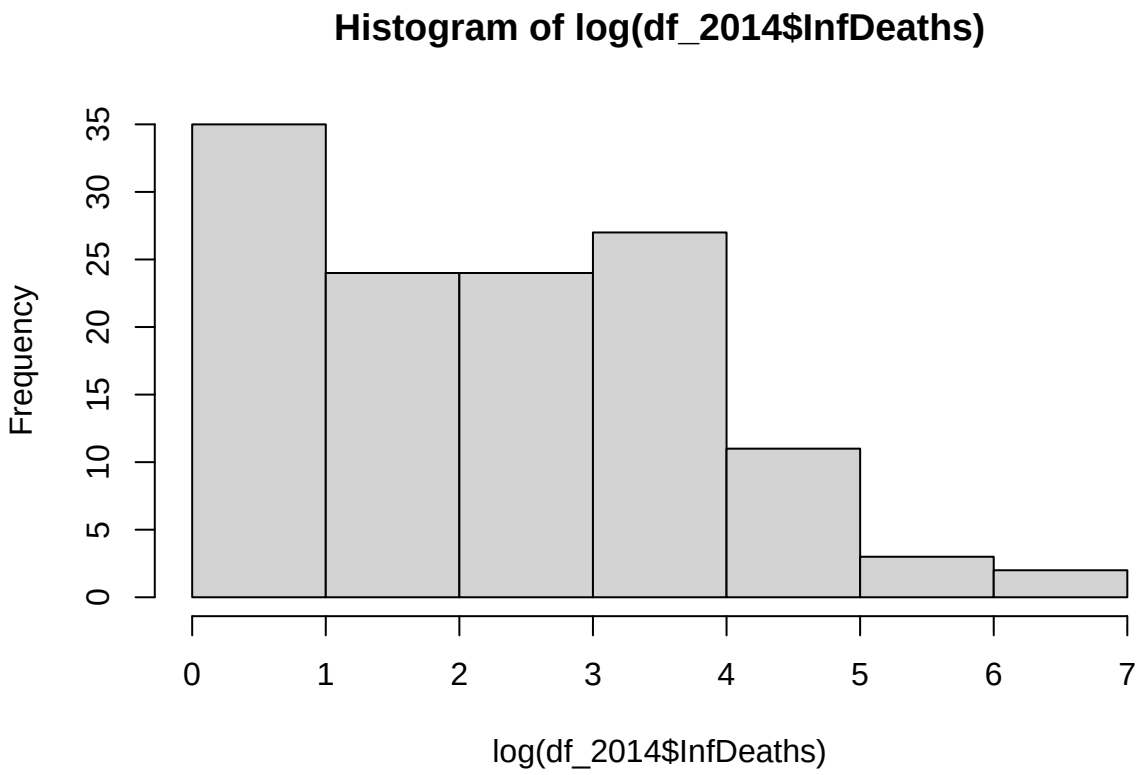
Appendix

Transformations

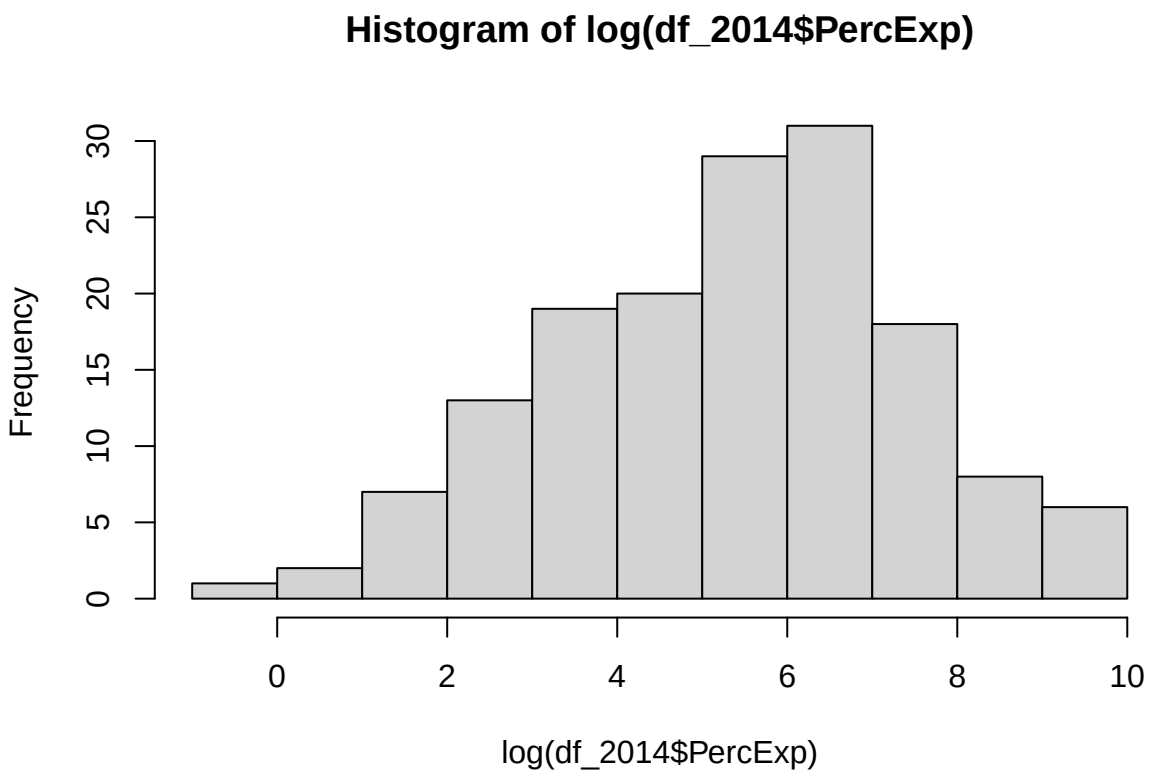
```
hist(log(df_2014$AdultMort))
```



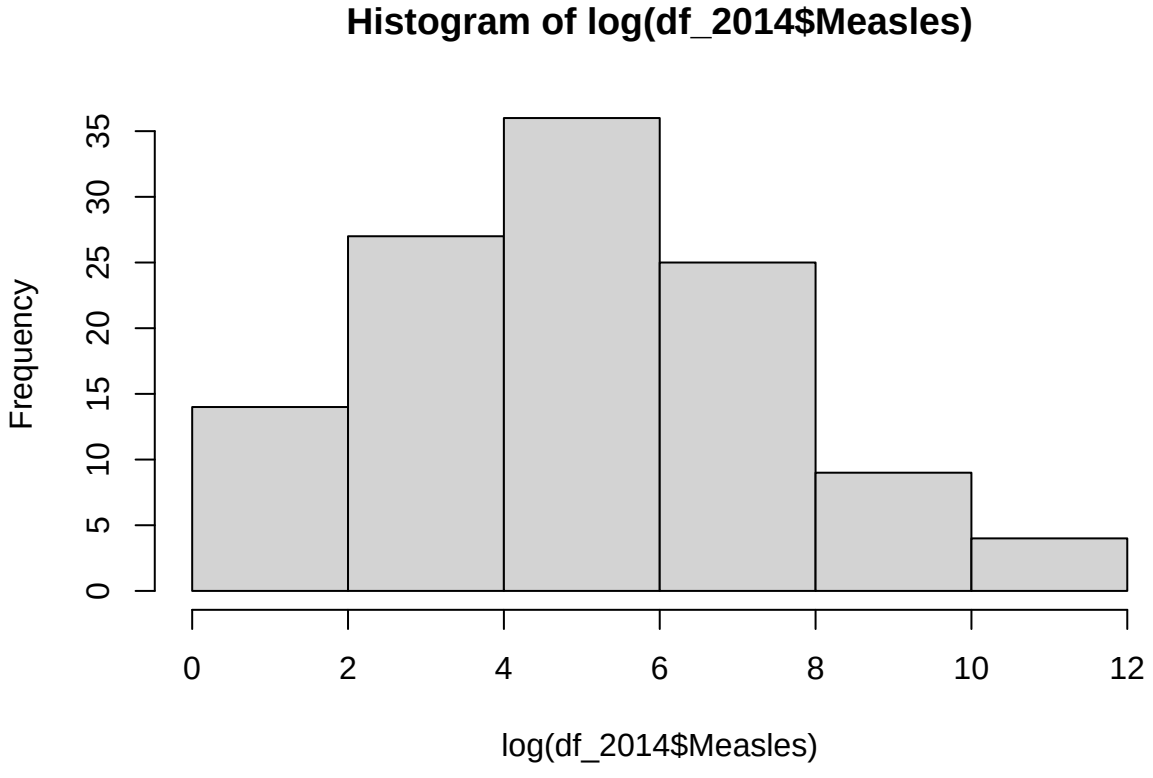

```
hist(log(df_2014$InfDeaths))
```



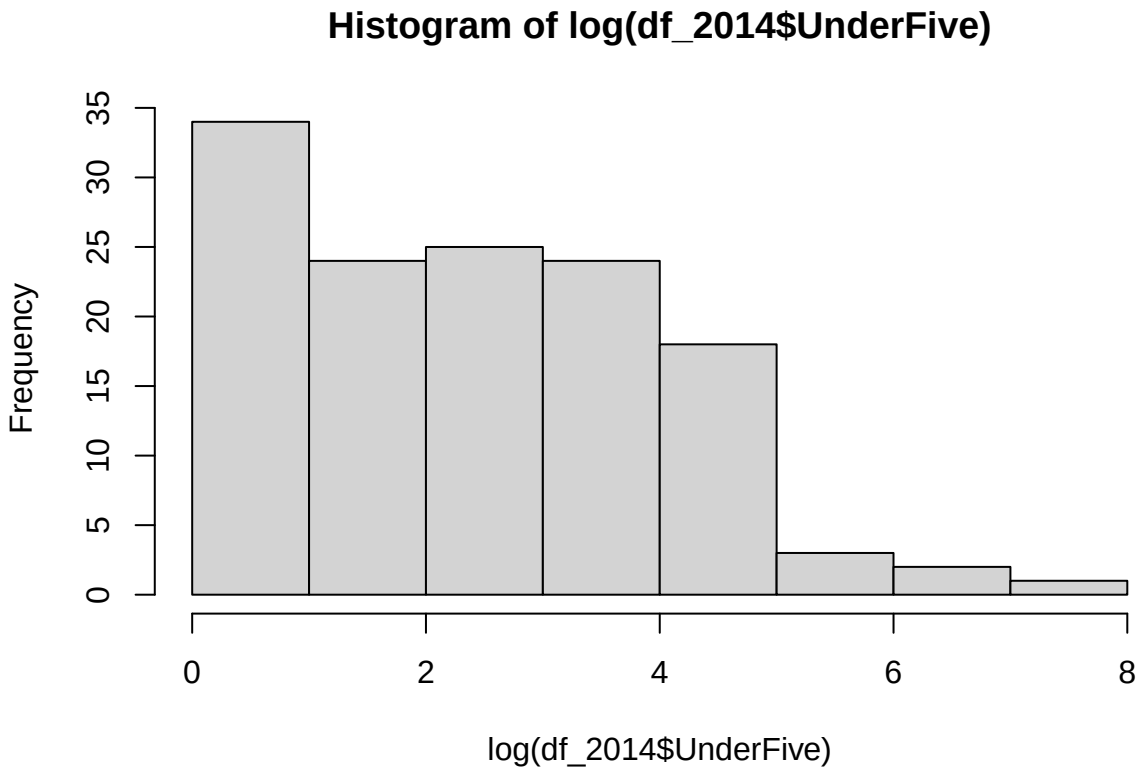
```
hist(log(df_2014$PercExp))
```



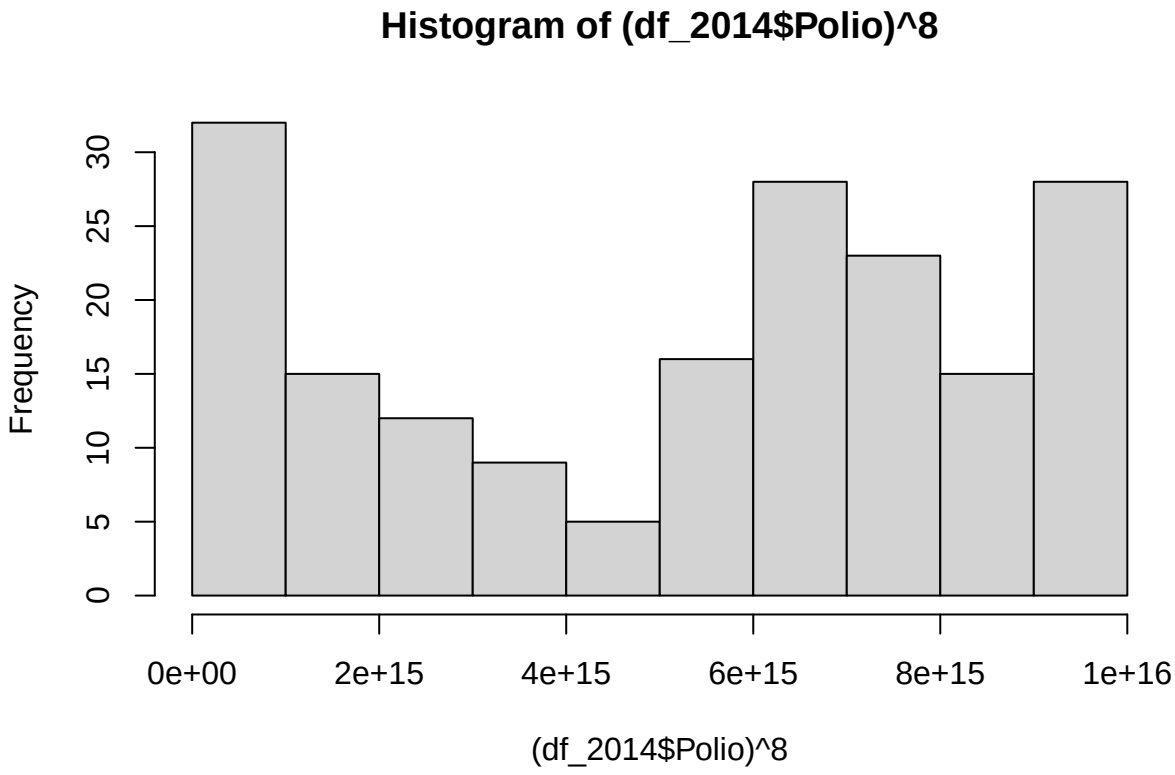

```
hist(log(df_2014$Measles))
```



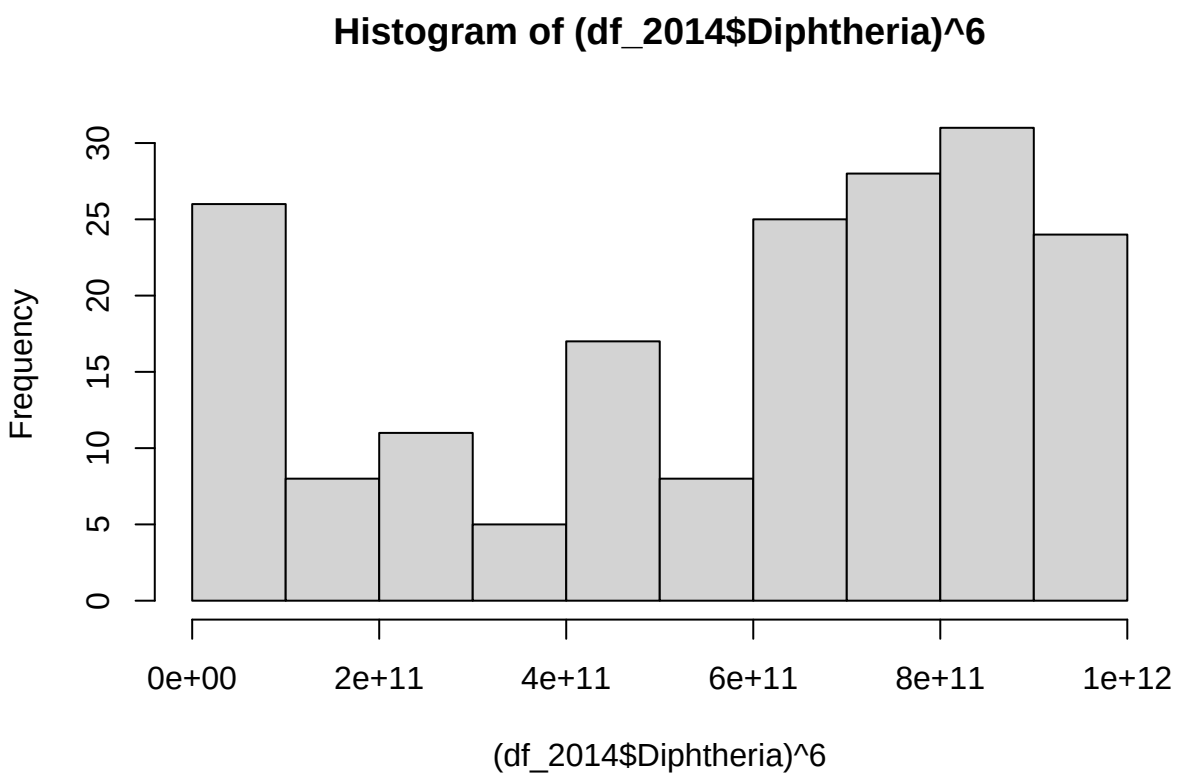
```
hist(log(df_2014$UnderFive))
```




```
hist((df_2014$Polio)^8)
```

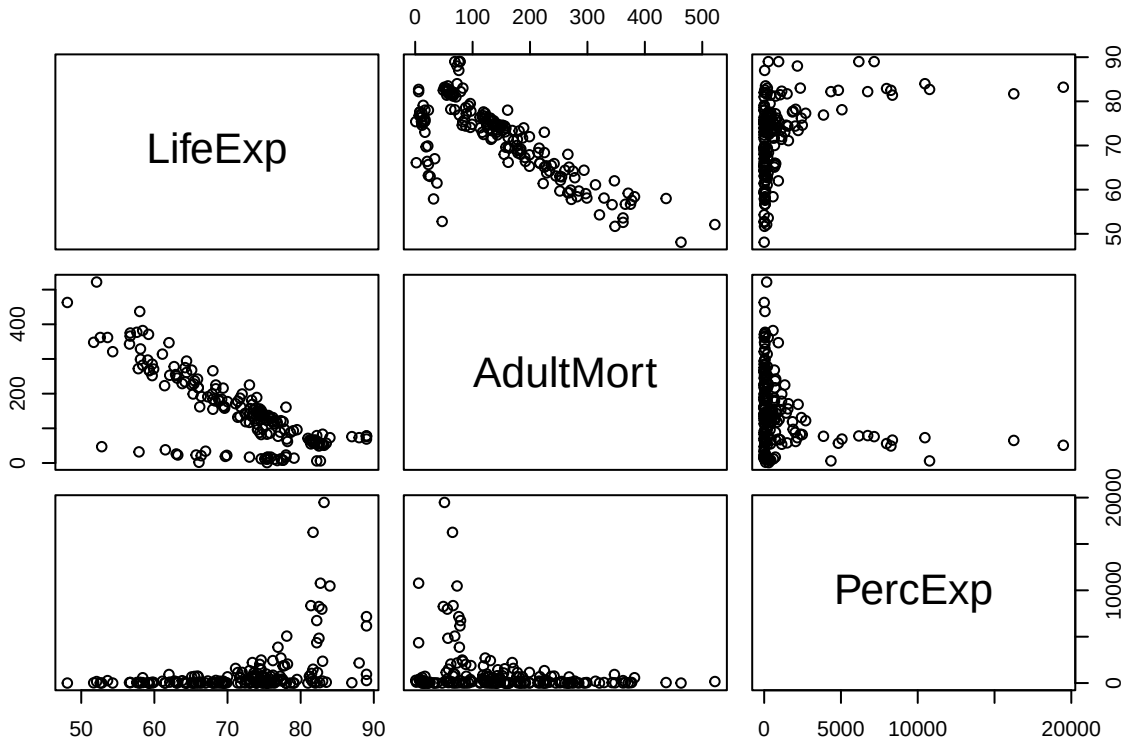


```r
hist((df_2014$Diphtheria)^6)
```

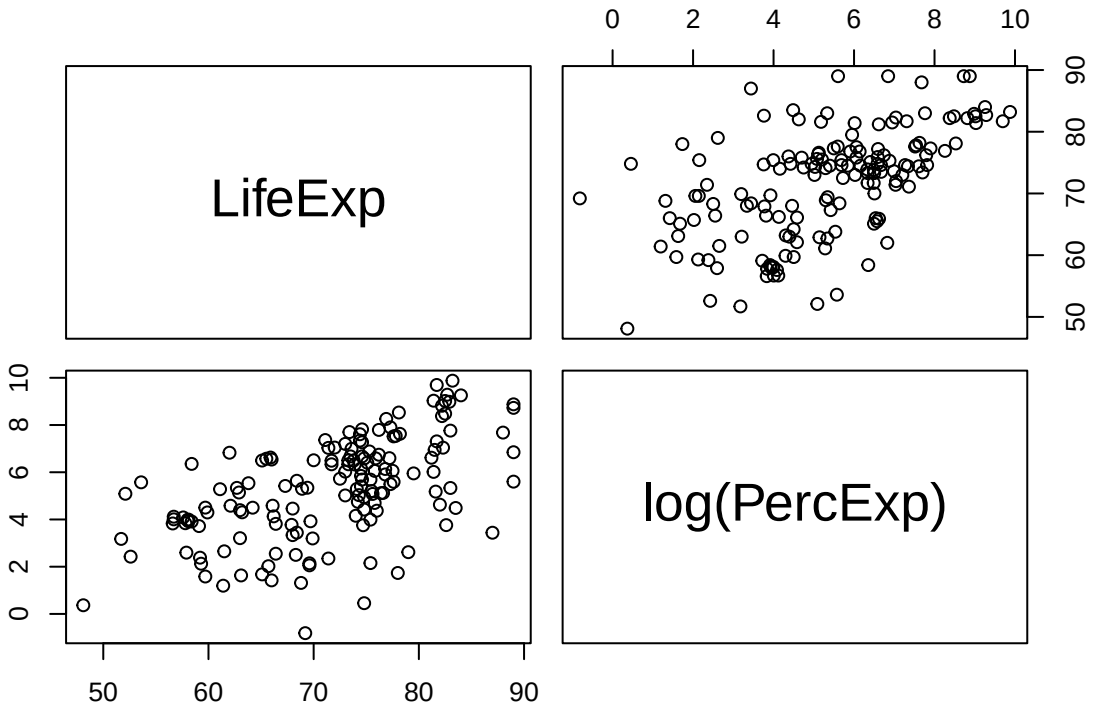


All Code and Output

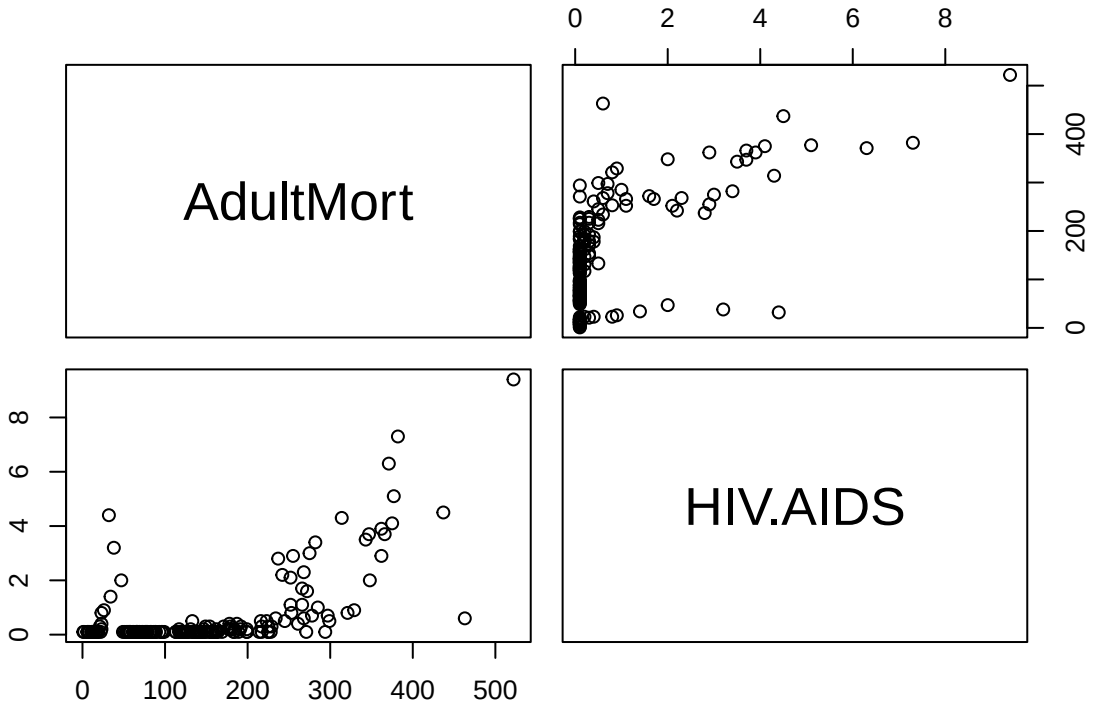
```r
df = read.csv("Life Expectancy Data.csv")

df_2014 = df[df[,2]==2014,] # taking only observations from 2014
df_2014 = df_2014[ , colSums(is.na(df_2014)) == 0] # removing columns with NA
df_2014$Developing = model.matrix( ~ Status - 1, data=df_2014 )[,2] # creating binary variable
df_2014 = df_2014[,-c(2,3)]
colnames(df_2014) <- c("Country","LifeExp","AdultMort","InfDeaths","PercExp","Measles","UnderFive","Poli
row.names(df_2014) <- NULL


pairs(df_2014[,c(2,3,5)]) # LifeExp, AdultMort, PercExp
```




```
pairs(data.frame(df_2014[,2],log(df_2014[,5])),labels = c("LifeExp","log(PercExp)"))
```



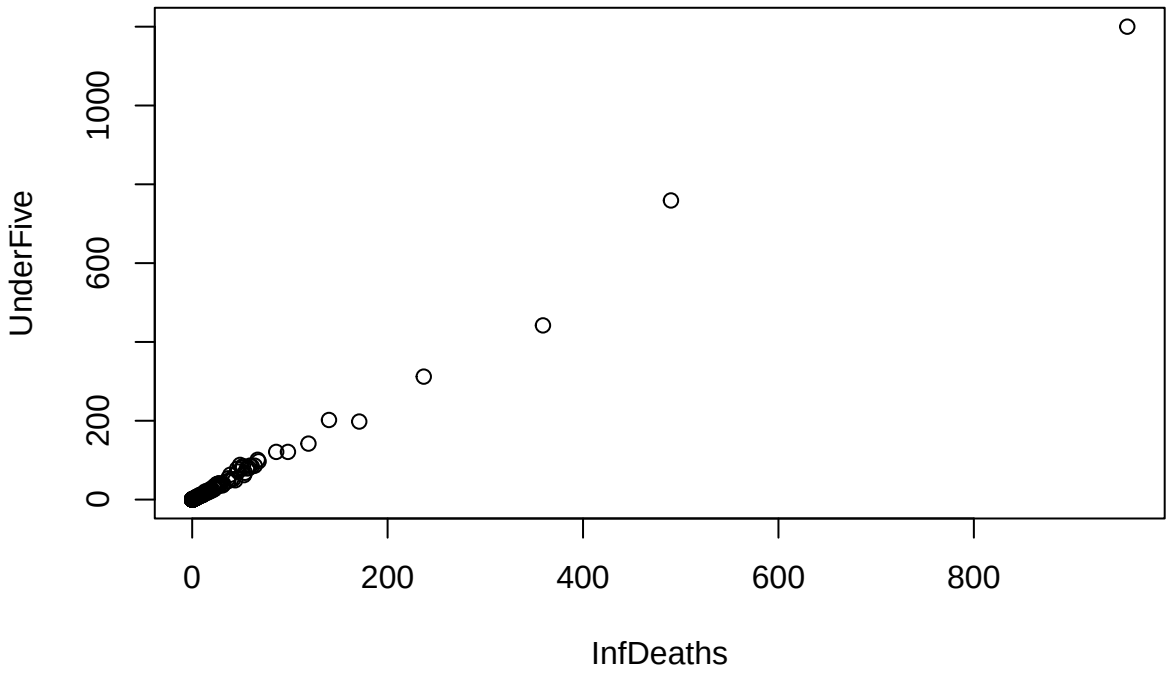
```
pairs(df_2014[,c(3,10)]) #AdultMort vs HIV.AIDS
```



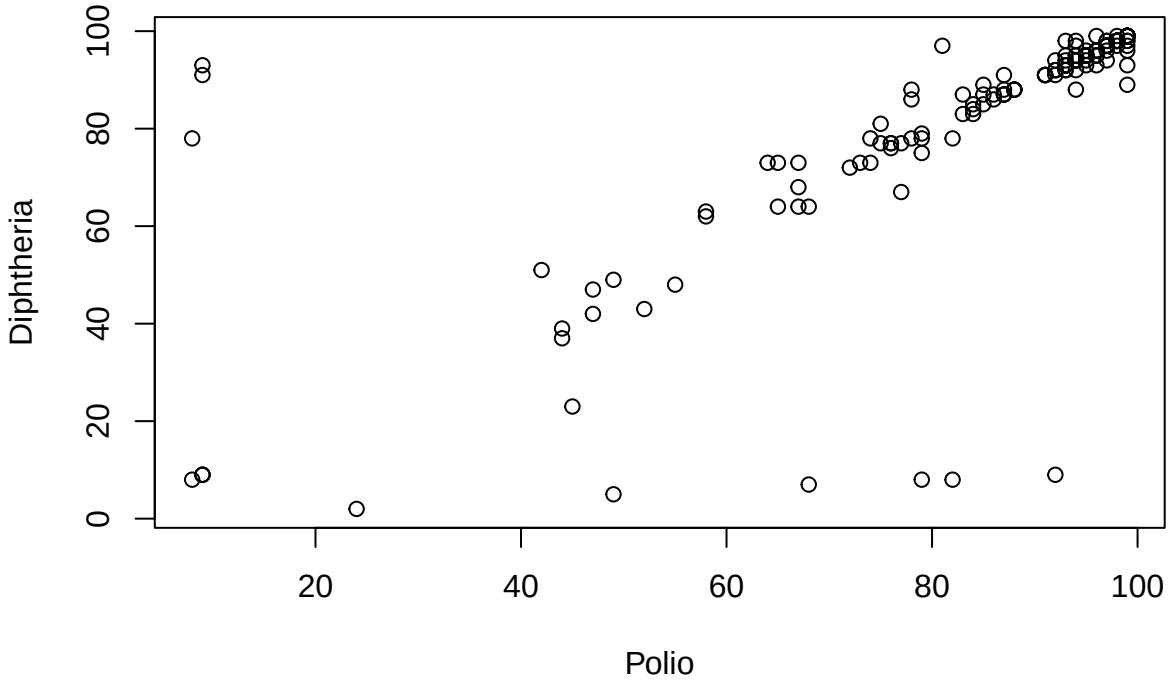

```
colMeans(df_2014[,2:11])
```

```
##      LifeExp    AdultMort    InfDeaths      PercExp      Measles    UnderFive
##   71.5366120  148.6885246   24.5573770 1001.9125498 1831.2076503   32.8907104
##        Polio   Diphtheria     HIV.AIDS   Developing
##   84.7267760   84.0819672    0.6819672    0.8251366
```

```
plot(df_2014[,c(4,7)]) # InfDeaths vs Underfive
```




```
plot(df_2014[,c(8,9)])
```

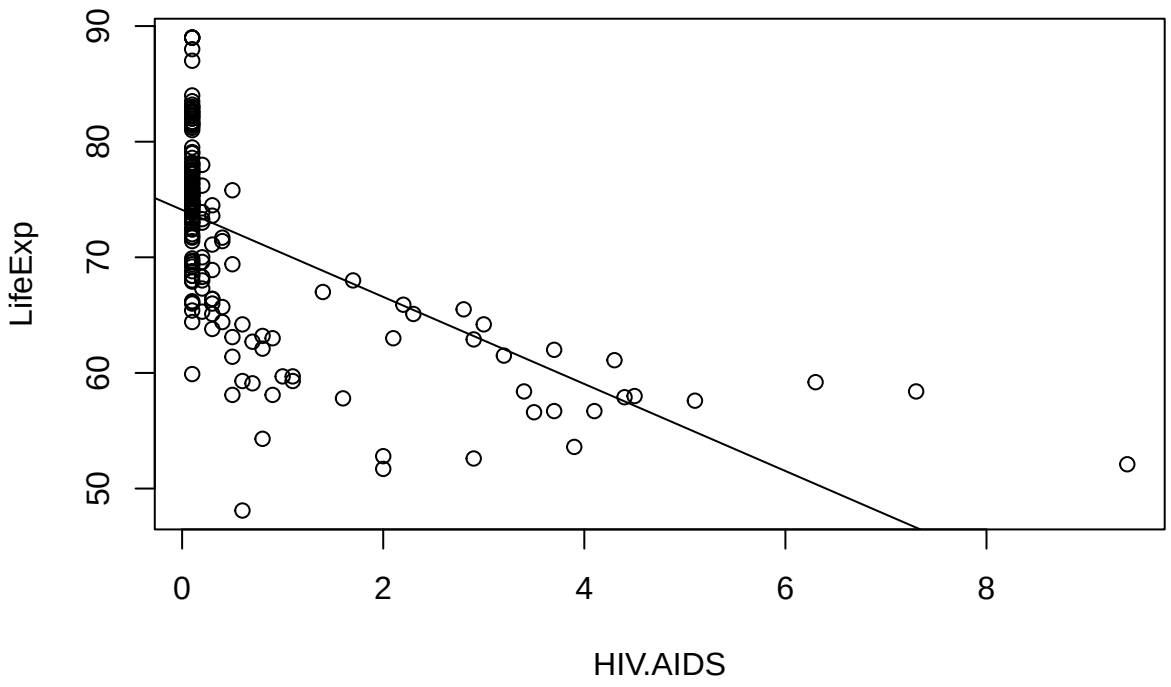



```
cor = round(cor(df_2014[,2:11]),2)
cor
```

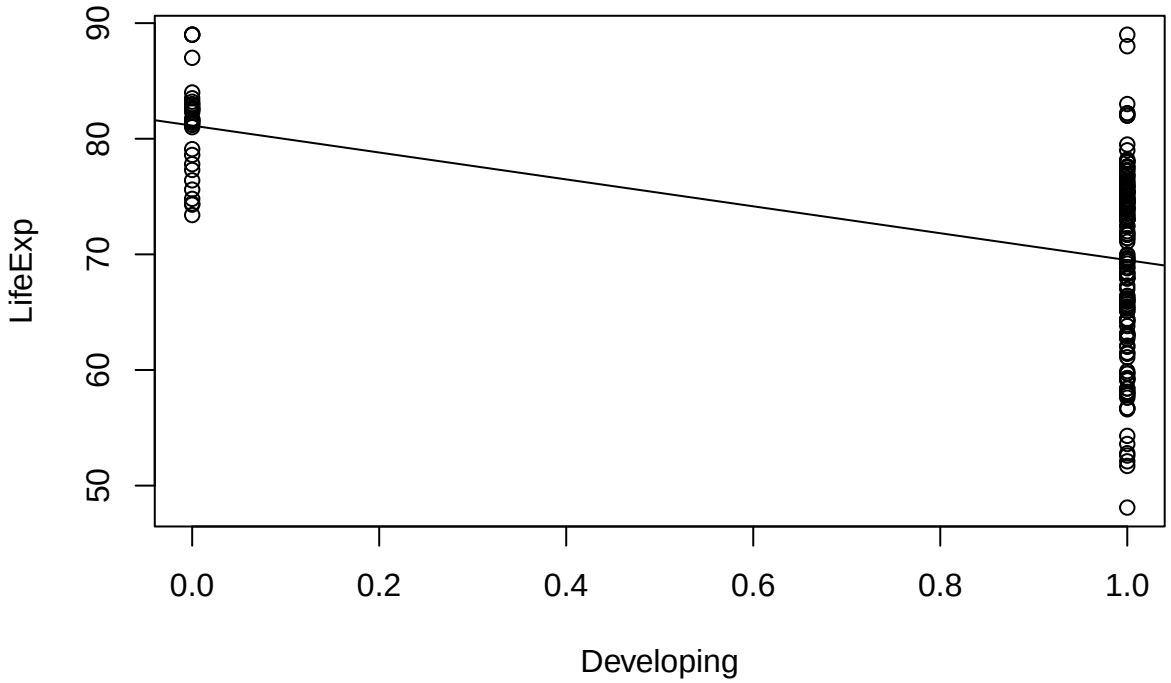
```
##             LifeExp AdultMort InfDeaths PercExp Measles UnderFive Polio
## LifeExp        1.00     -0.76     -0.24    0.40   -0.10     -0.26  0.41
## AdultMort     -0.76      1.00      0.18   -0.25    0.09      0.20 -0.38
## InfDeaths     -0.24      0.18      1.00   -0.10    0.70      1.00 -0.13
```

```
## PercExp       0.40     -0.25     -0.10    1.00   -0.07     -0.10  0.08
## Measles      -0.10      0.09      0.70   -0.07    1.00      0.68 -0.05
## UnderFive    -0.26      0.20      1.00   -0.10    0.68      1.00 -0.14
## Polio         0.41     -0.38     -0.13    0.08   -0.05     -0.14  1.00
## Diphtheria    0.38     -0.30     -0.13    0.08   -0.10     -0.14  0.75
## HIV.AIDS     -0.61      0.63      0.09   -0.13   -0.03      0.12 -0.34
## Developing   -0.52      0.32      0.12   -0.42    0.09      0.13 -0.18
##            Diphtheria HIV.AIDS Developing
## LifeExp          0.38    -0.61      -0.52
## AdultMort       -0.30     0.63       0.32
## InfDeaths       -0.13     0.09       0.12
## PercExp          0.08    -0.13      -0.42
## Measles         -0.10    -0.03       0.09
## UnderFive       -0.14     0.12       0.13
## Polio            0.75    -0.34      -0.18
## Diphtheria       1.00    -0.22      -0.18
## HIV.AIDS        -0.22     1.00       0.19
## Developing      -0.18     0.19       1.00
```

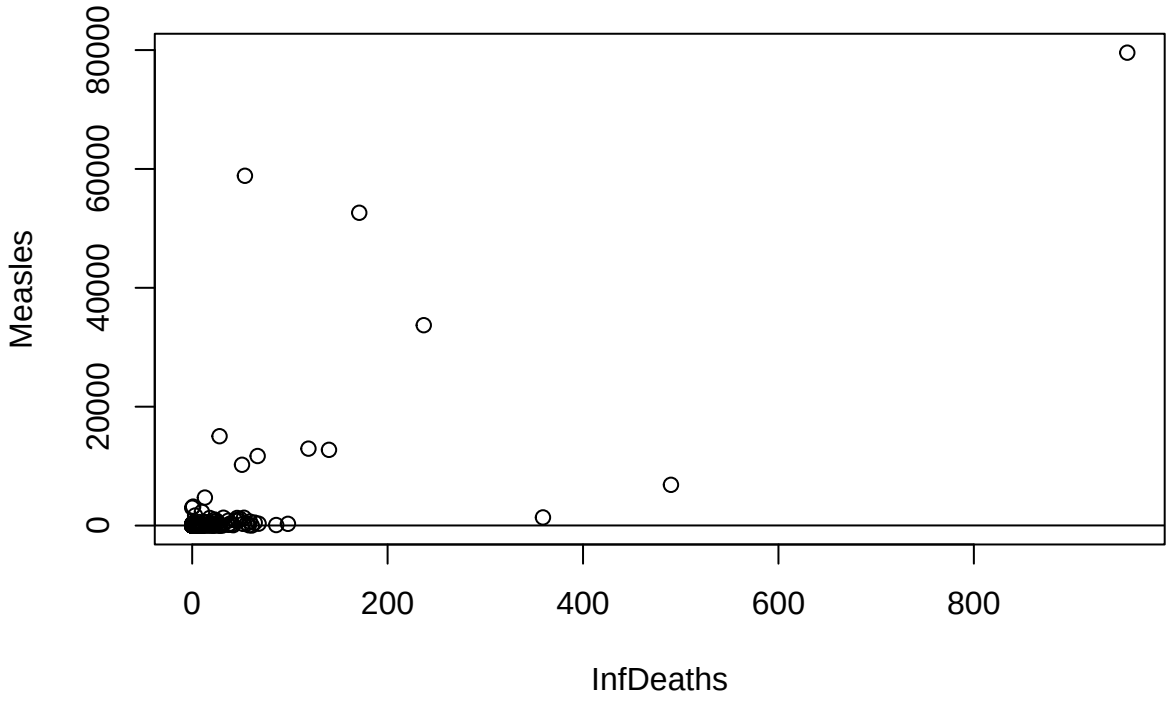
```
reg1 = lm(df_2014$LifeExp~df_2014$HIV.AIDS)
plot(df_2014$HIV.AIDS,df_2014$LifeExp,xlab="HIV.AIDS",ylab="LifeExp")
abline(reg=reg1)
```




```
reg2 = lm(df_2014$LifeExp~df_2014$Developing)
plot(df_2014$Developing,df_2014$LifeExp,xlab="Developing",ylab="LifeExp")
abline(reg=reg2)
```




```
reg3 = lm(df_2014$InfDeaths~df_2014$Measles)
plot(df_2014$InfDeaths,df_2014$Measles,xlab="InfDeaths",ylab="Measles")
abline(reg3)
```




```
# Eliptical Distances
edist = function(x){
  x = as.matrix(x)
  ev = eigen(var(x))
  d = diag(1/sqrt(abs(ev$values)))
```

```
  b = ev$vectors%*%d%*%t(ev$vectors)
  z = x%*%b
  d = dist(z);  return(d) }
ed = edist(df_2014[,2:11])

mahalanobis_dis=function(df){
  dist = c()
  sigma_inv = solve(cov(df))
  xbar = colMeans(df)
  for(i in 1:nrow(df)){
    x_xbar = as.matrix(df[i,]-xbar)
    dist = c(dist,sqrt(x_xbar%*%(sigma_inv)%*%t(x_xbar)))
  }
  return(dist)
}

MD = mahalanobis_dis(df_2014[,2:11])
df_2014$MD = MD

sorted_MD = df_2014
sorted_MD = sorted_MD[order(-MD),]
row.names(sorted_MD) <- NULL
knitr::kable(head(sorted_MD[,c(1,2,11,12)],20),caption = "Data with Mahalanobis Distances") %>%
    kable_styling(latex_options = "HOLD_position")
```

Table 17: Data with Mahalanobis Distances

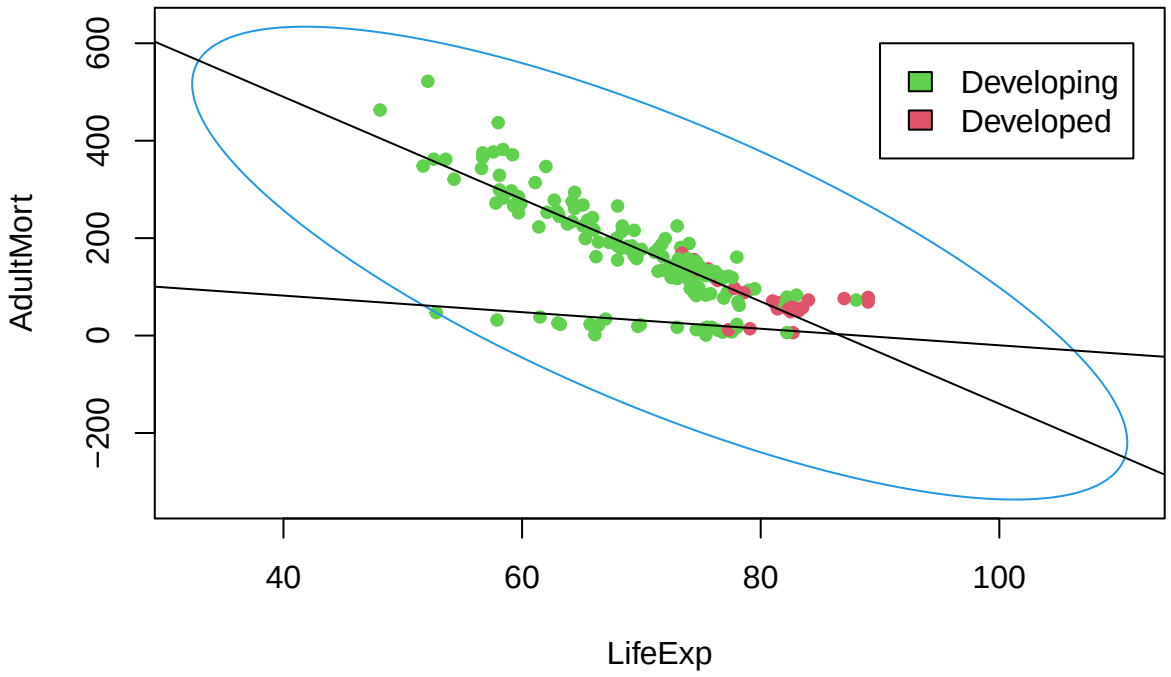
Country	LifeExp	Developing	MD
Nigeria	53.6	1	12.487591
India	68.0	1	11.789939
Philippines	68.4	1	9.409624
Lesotho	52.1	1	8.185775
Switzerland	83.2	0	7.495350
Pakistan	66.2	1	7.113682
China	75.8	1	6.864651
Equatorial Guinea	57.9	1	6.329881
Dominican Republic	73.6	1	6.219381
Luxembourg	81.7	0	6.205190
Indonesia	68.9	1	5.845290
Democratic Republic of the Congo	59.3	1	5.807163
Swaziland	58.4	1	5.613041
Republic of Moldova	71.8	1	5.551212
Iceland	82.5	0	5.455031
Tonga	73.3	1	5.078669
Côte d'Ivoire	52.8	1	4.876560
Gabon	65.5	1	4.787985
Sierra Leone	48.1	1	4.721311
Panama	77.6	1	4.694914

```
e2d = function(loc, cov, dis) {
A = solve(cov)
```

```
eA = eigen(A)
ev = eA$values
lambda1 = max(ev)
lambda2 = min(ev)
eigvect = eA$vectors[, order(ev)[2]]
z = seq(0, 2 * pi, by = 0.01)
z1 = dis/sqrt(lambda1) * cos(z)
z2 = dis/sqrt(lambda2) * sin(z)
alfa = atan(eigvect[2]/eigvect[1])
r <- matrix(c(cos(alfa), -sin(alfa), sin(alfa), cos(alfa)), ncol = 2)
z = t(t(cbind(z1, z2) %*% r) + loc)
}

ellipse=function(x,center,cov,dis,col,type=1){
  if(class(x)== "data.frame") x=as.matrix(x)
  if(class(x)!="matrix") stop("Data is not a matrix.")
  if(ncol(x)!=2) stop("Number of variables should be 2.")
  if(type==1) dis = sqrt(qchisq(dis, 2))
  y=e2d(center,cov,dis); z=rbind(x,y)
  plot(x, pch=19,cex=0.8,col=col,xlim=c(min(z[,1]),max(z[,1])),ylim=c(min(z[,2]),max(z[,2])) )
  lines(y,col=4)
}

x = df_2014[,c(2,3)]
cov = cov(x)
eig = eigen(cov)
mu = colMeans(x)
ellipse(x,mu,cov,col=df_2014[,11]+2,max(mahalanobis_dis(df_2014[,c(2,3)])),type=2)
```

```
## Warning in if (class(x) != "matrix") stop("Data is not a matrix."): the
## condition has length > 1 and only the first element will be used
```

```
legend(90,600,legend=c("Developing","Developed"),fill=c(3,2))
abline(a=910, b=-10.5)
abline(a=150, b=-1.7)
```




```
require(robustX); library(robustbase)
x = as.matrix(df_2014[,2:10])
output = mvBACON(x)
```

```
## rank(x.ord[1:m,] >= p  ==> chosen m =  36
## MV-BACON (subset no. 1): 36 of 183 (19.67 %)
## MV-BACON (subset no. 2): 99 of 183 (54.1 %)
## MV-BACON (subset no. 3): 100 of 183 (54.64 %)
## MV-BACON (subset no. 4): 100 of 183 (54.64 %)
## MV-BACON (subset no. 5): 99 of 183 (54.1 %)
## MV-BACON (subset no. 6): 99 of 183 (54.1 %)
```

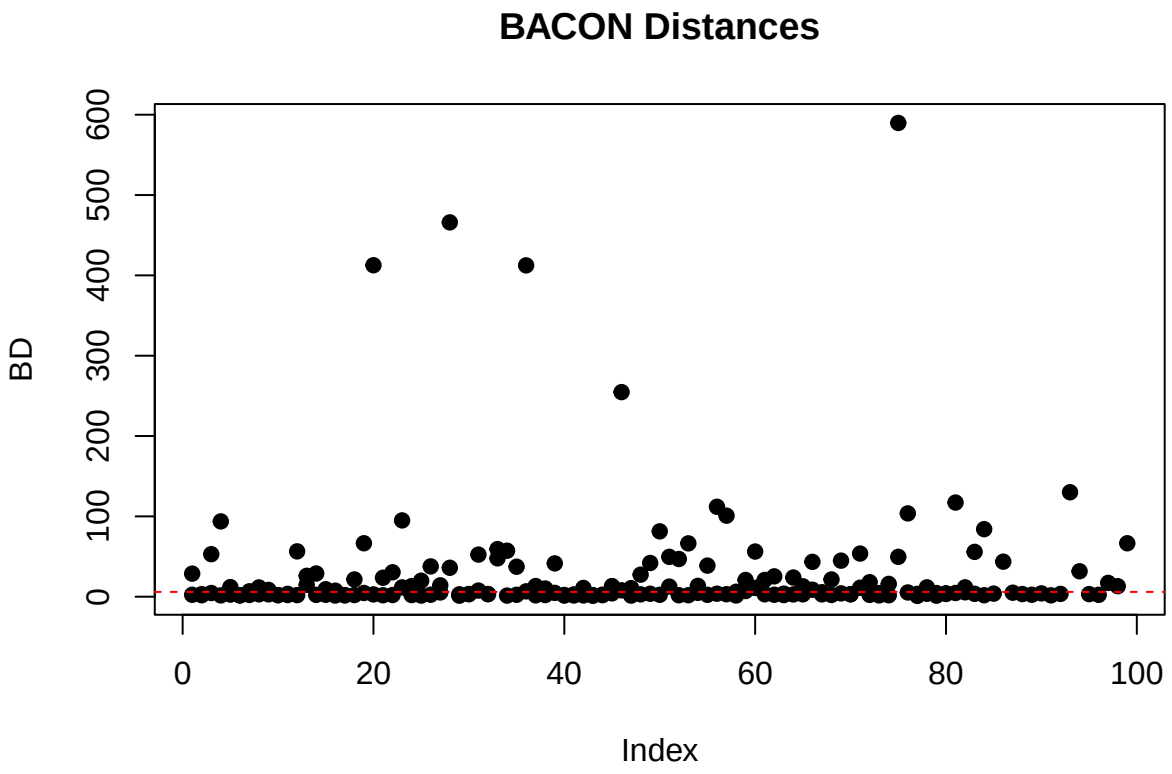
```
df_2014_bd = df_2014
df_2014_bd$bd = output$dis

top_countries = rbind(df_2014[df_2014[,"Country"]=="Nigeria",],df_2014[df_2014[,c("Country")]=="India",]
top_countries
```

```
##                              Country LifeExp AdultMort InfDeaths   PercExp
## 119                          Nigeria    53.6       362       490 263.21110
## 75                             India    68.0       184       957  86.52154
## 36                             China    75.8        86       171 109.87439
## 127                      Philippines    68.4       214        54  31.27232
## 46  Democratic Republic of the Congo    59.3       266       237   0.00000
##     Measles UnderFive Polio Diphtheria HIV.AIDS Developing        MD
## 119    6855       759    49         49      3.9          1 12.487591
## 75    79563      1200    84         85      0.2          1 11.789939
## 36    52628       198    99         99      0.1          1  6.864652
## 127   58848        68    77         67      0.1          1  9.409624
## 46    33711       312    79          8      1.1          1  5.807163
```

```
y = cbind(1:99,df_2014_bd$bd)
```

```
## Warning in cbind(1:99, df_2014_bd$bd): number of rows of result is not a
## multiple of vector length (arg 1)
```

```
colnames(y) <- c("Index","Distance")
plot(y[,1:2],xlab="Index",ylab="BD", pch=19, main = "BACON Distances")
abline(h=output$limit,  col= "red",  lty=2)
```




```
t(output$center)
```

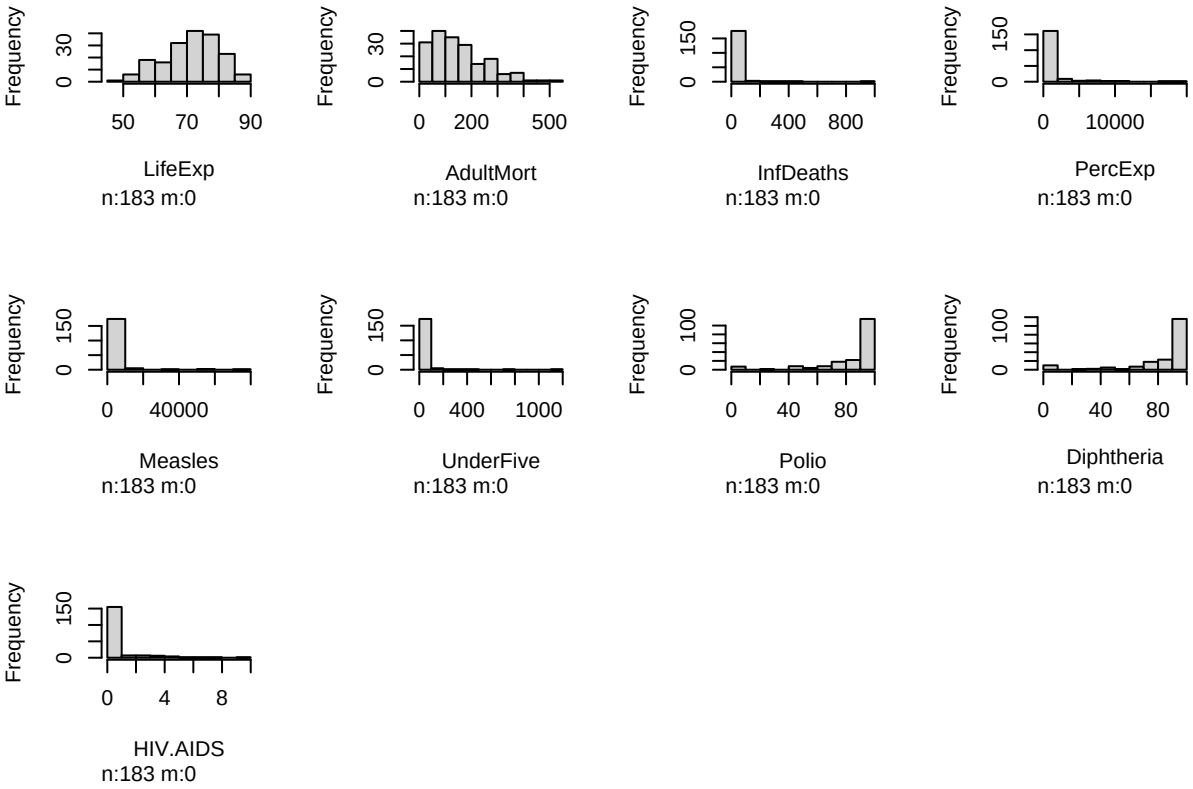
```
##       LifeExp AdultMort InfDeaths  PercExp  Measles UnderFive    Polio
## [1,] 75.47879  103.6566  2.959596 623.2713 72.29293  3.515152 93.40404
##      Diphtheria  HIV.AIDS
## [1,]   93.60606 0.1323232
```

```
robustcor = diag(1/sqrt(diag(output$cov)))%*%output$cov%*%diag(1/sqrt(diag(output$cov)))
colnames(robustcor) <- c("LifeExp", "AdultMort" ,"InfDeaths" ,"PercExp" ,"Measles", "UnderFive", "Polio
row.names(robustcor) <- c("LifeExp", "AdultMort" ,"InfDeaths" ,"PercExp" ,"Measles", "UnderFive", "Polio
round(robustcor,2)
```

```
##            LifeExp AdultMort InfDeaths PercExp Measles UnderFive Polio
## LifeExp       1.00     -0.56     -0.17    0.27    0.22     -0.18  0.25
## AdultMort    -0.56      1.00      0.04   -0.09   -0.23      0.05 -0.20
## InfDeaths    -0.17      0.04      1.00   -0.19    0.27      1.00 -0.02
## PercExp       0.27     -0.09     -0.19    1.00   -0.08     -0.19  0.05
```

```
## Measles       0.22     -0.23      0.27   -0.08    1.00      0.29  0.07
## UnderFive    -0.18      0.05      1.00   -0.19    0.29      1.00 -0.03
## Polio         0.25     -0.20     -0.02    0.05    0.07     -0.03  1.00
## Diphtheria    0.28     -0.19      0.02    0.08    0.07      0.02  0.97
## HIV.AIDS     -0.36      0.47     -0.01   -0.04   -0.14      0.02 -0.14
##            Diphtheria HIV.AIDS
## LifeExp          0.28    -0.36
## AdultMort       -0.19     0.47
## InfDeaths        0.02    -0.01
## PercExp          0.08    -0.04
## Measles          0.07    -0.14
## UnderFive        0.02     0.02
## Polio            0.97    -0.14
## Diphtheria       1.00    -0.11
## HIV.AIDS        -0.11     1.00
```

```
ht2=function(x, y) {
n=dim(x)[1];m=dim(y)[1]; p=dim(x)[2]
xcov=cov(x);
ycov=cov(y)
Sp=(n-1)*xcov+(m-1)*ycov; Sp=Sp/(n+m-2)
xcenter=colMeans(x); ycenter=colMeans(y)
d=xcenter-ycenter
T2=t(d)%*%solve(Sp)%*%d
T2=T2*n*m/(n+m)
F=T2*(n+m-p-1)/(p*(n+m-2))
pv=1-pf(F,p,n+m-p-1)
list(xcenter=xcenter,ycenter=ycenter,xcov=xcov,ycov=ycov, Sp=Sp,T2=T2,F=F,df=c(p,n+m-p-1),pv=pv) }

library(Hmisc)

hist.data.frame(df_2014[,2:11])
```




```
out1 = ht2(df_2014[df_2014$Developing==1,2:10],df_2014[df_2014$Developing==0,2:10])

out1$pv
```

```
##              [,1]
## [1,] 1.660228e-12
```

```
df_2014_woo = df_2014[output$subset,]

out2 = ht2(df_2014_woo[df_2014_woo$Developing==1,2:10],df_2014_woo[df_2014_woo$Developing==0,2:10])

t(out2$xcenter)
```

```
##      LifeExp AdultMort InfDeaths  PercExp  Measles UnderFive    Polio
## [1,]  73.912  110.9733  3.386667 587.7759 59.01333  4.026667 92.74667
##      Diphtheria  HIV.AIDS
## [1,]   92.98667 0.1426667
```

```
t(out2$ycenter)
```

```
##      LifeExp AdultMort InfDeaths  PercExp  Measles UnderFive    Polio
## [1,]  80.375  80.79167     1.625 734.1944 113.7917  1.916667 95.45833
##      Diphtheria HIV.AIDS
## [1,]   95.54167      0.1
```

```
out2$pv
```

```
##              [,1]
## [1,] 1.713182e-05
```

```
tf = df_2014
tf$AdultMort = log(df_2014$AdultMort+1)
tf$InfDeaths = log(df_2014$InfDeaths+1)
tf$PercExp = log(df_2014$PercExp+1)
tf$Measles = log(df_2014$Measles+1)
tf$UnderFive = log(df_2014$UnderFive+1)
tf$Polio = (df_2014$Polio)^8
tf$Diphtheria = (df_2014$Diphtheria)^6
#tf = as.matrix(tf[,2:10])

#out3 = ht2(tf[tf$Developing==1,2:10],tf[tf$Developing==0,2:10])
tfm = as.matrix(tf[,2:10])

eig = eigen(t(tfm)%*%tfm)

cat(sqrt(eig$values[1]/eig$values[2]),sqrt(eig$values[1]/eig$values[3]),sqrt(eig$values[1]/eig$values[4]
```

```
## 46729.48 4.955632e+12 1.958732e+14 1.932357e+15 3.830266e+15 3.830266e+15 4.563936e+15 4.563936e+15
```

References

Max Roser, Esteban Ortiz-Ospina, and Hannah Ritchie. 2013. "Life Expectancy." *Our World in Data*.