1. **Introduction:**
    - Show the data and explain why you selected the data and how the data was collected.

year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Country Name															
Afghanistan M/1000 births	85.4	82.8	80.1	77.4	74.7	72	69.4	80.7	76	71.5	67.3	63.6	60.5	57.8	55.5
Argentina M/1000 births	16.3	15.8	15.3	14.8	14.4	14	13.7	17.4	16.5	15.7	14.9	14.2	13.5	12.8	12.2
Health budget AFG	9.4	8.9	9.8	9.9	10.6	9.9	10.3	9.8	8.6	8.6	7.9	8.8	9.5	10.0	10.2
Health budget ARG	7.7	7.1	7.0	7.6	7.6	7.6	7.7	9.0	8.6	8.4	8.4	8.4	8.2	8.7	7.5

This is data from different countries there are four variables. The rates of mortality for the two countries per 1000 live births and the health budgets for each country. The data represents different mortality rates in the selected two Countries per 1000 live births. This is secondary data and it was obtained from this website https://data.wordbank.org/indicator/SP.DYN.IMRT.IN?end=2018&start=2018&view=bar . I chose the World Bank website since it is more reliable and updated in the word data. The data was collected from different Countries from all over the world but in my case I chose to analyze the data of two countries from the eleven selected countries. My analysis will focus on the Afghanistan and Argentina. This is because the dataset is a large one and I wanted to two countries. I chose Afghanistan due to its political issues and a control experiment country, Argentina.

I t would be interesting to know how these two countries vary in terms of the mortality rates and maybe know what causes the difference in the countries' rates. The mortality data will help in knowing the possible causes of mortality in the children and possibly help in coming up with interventions for different countries to adapt and lower the rates of mortality. The Y variables are represented by the health budget allocations while the X variables are represented by the mortality rates. Health budget allocation is in millions of dollars.

Rationale

With my previous knowledge in data mining and analysis, I believe that the mortality data from the two selected countries will help me in my research and come up with a solid conclusion and recommendation. With the budget allocations data, we will make a conclusion on how the countries deal with the rates of mortality.

Limitations of the data

The presented data above lacks some significance values. The data however does not show the cause of the mortality rates but in my thought, there might be several causes of the mortality rates. If the data all the causes of mortality rates, it would have been easy to make conclusions to a specific cause of the mortality rate.

Data cleaning

There are several steps that the data went through so that it can be used for the analysis. With the several countries data, I chose data for two countries to avoid the long data that could have taken a lot of time in analyzing. There was not complete data from 2000 and this made me use the available data from 2002 to 2016.

Assumptions

My general assumptions of the data basis on the different political situations in the two countries. For instance I assumed that due to the unrests in the politics of Afghanistan, the rates of mortality per 1000 live births is high while the rates for Argentina are low. This is the experiment I would like to carry out in the analysis and know whether it holds.

2. **Describing the data:**
    - Plot histograms of your x variable and the y variable using reasonable intervals

Figure 2.1

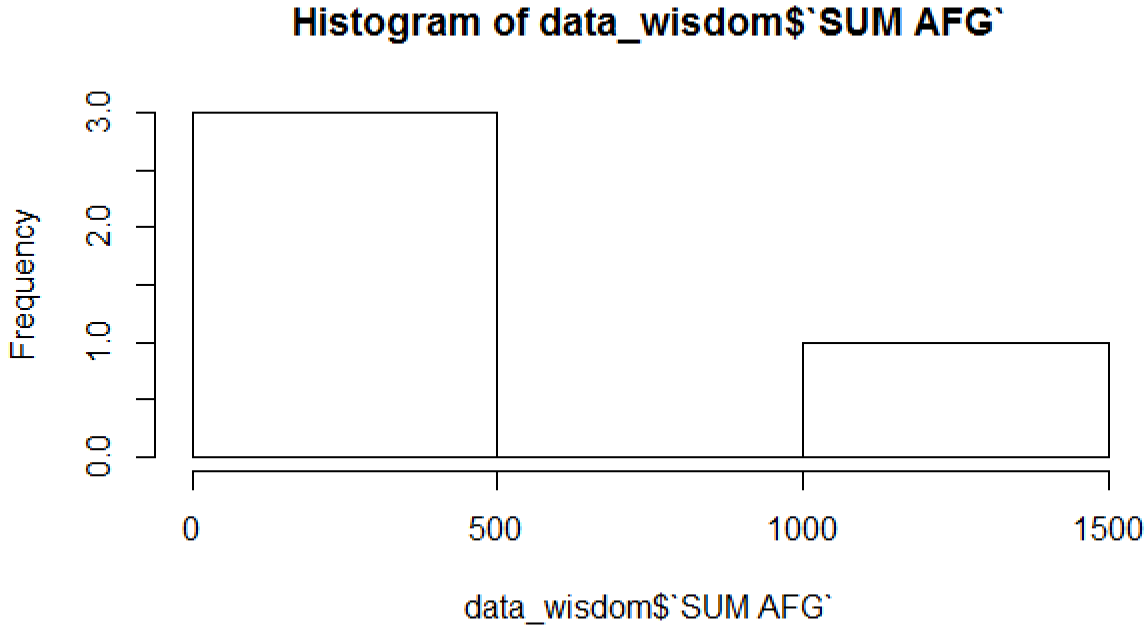


Figure 2.2

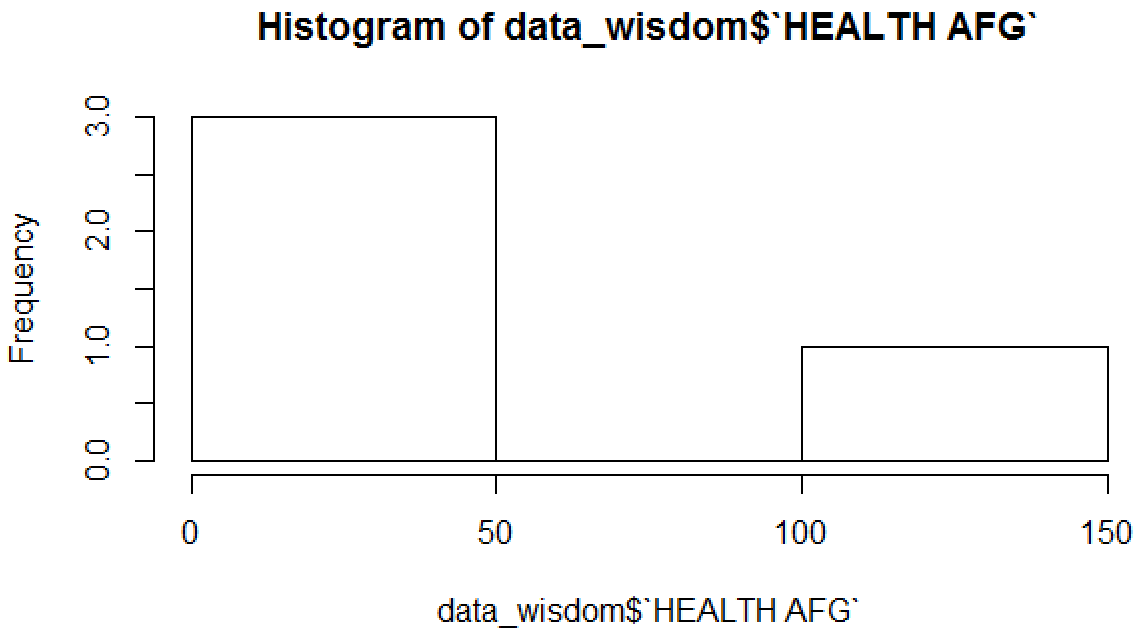


Figure 2.3

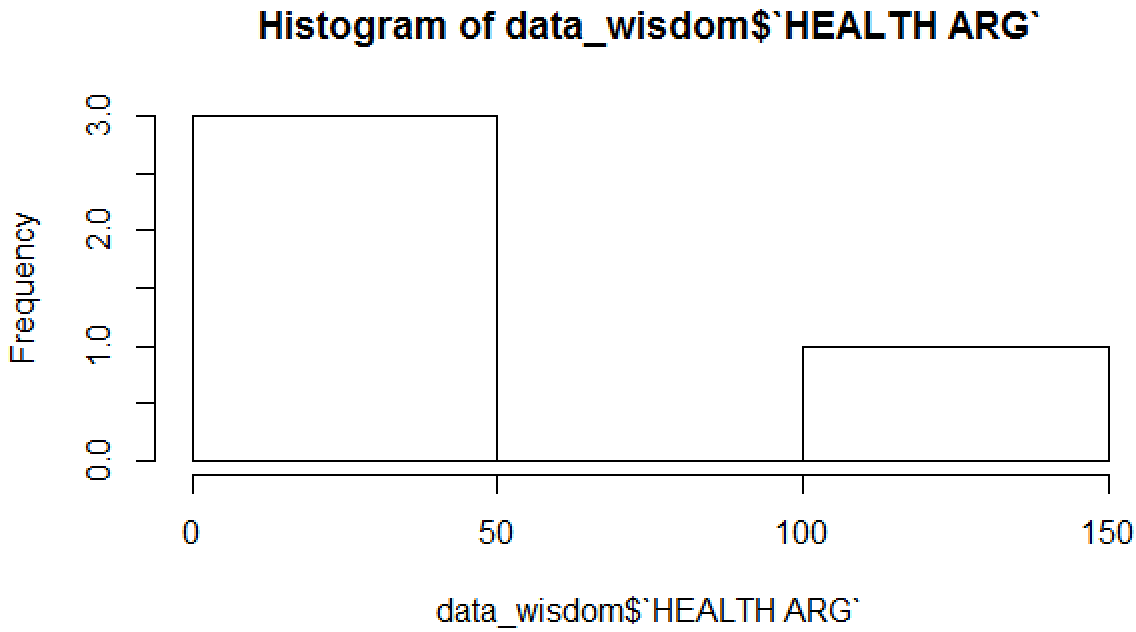


Figure 2.4

Figures 2.1 and 2.2 represent the histograms for the mortality rates in Afghanistan and Argentina respectively while Figure 2.3 and 2.4 represent the histogram for the health budgets in Afghanistan and Argentina respectively. In the histograms above, all look similar. They are all right-skewed and this means that the mean is less than the median of the data.

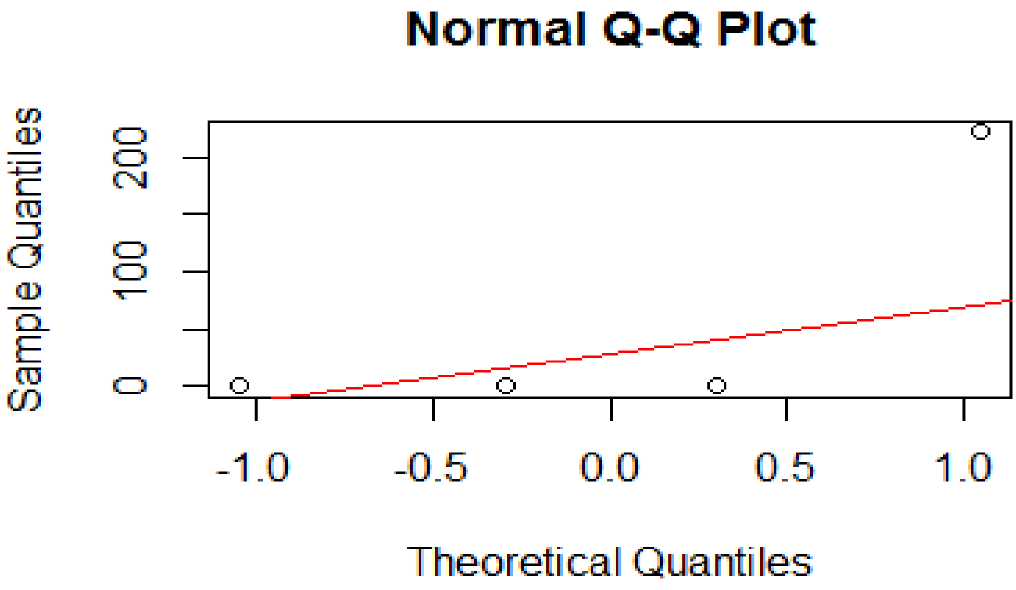
3. Analyze whether the **x and y distributions** satisfy the empirical rule (Yes or No, explain why). **Show details such like the range of within 1 standard deviation, within 2 standard deviation and within 3 standard deviation and the corresponding true percentage falling in these ranges.**

When testing for normality of data, the empirical rule of normal distribution applies. 68 percent of the data should be within the standard deviation of mean, 95 percent of the data must be within the 2 standard deviations and 99.7 percent of the data within the 3 standard deviations. From the above data 75% lies within the 1 standard deviation, 100% of the data lies within the 2 and 3 standard deviations.

This result can be found from the R code below

```
SIGMA2<-(sum(data_wisdom$`HEALTH ARG`>=A-(2*B) & data_wisdom$`HEALTH ARG`<=A+(2*B))/length(data_wisdom$`HEALTH ARG`))*100
> SIGMA2<-(sum(data_wisdom$`HEALTH ARG`>=A-(3*B) & data_wisdom$`HEALTH ARG`<=A+(3*B))/length(data_wisdom$`HEALTH ARG`))*100
> SIGMA3<-(sum(data_wisdom$`HEALTH ARG`>=A-(3*B) & data_wisdom$`HEALTH ARG`<=A+(3*B))/length(data_wisdom$`HEALTH ARG`))*100
> lowerbound<-A-B*3
> UPPERBOUND<-A+B*3
```

The normal Q-Q plot below show the normality of the data.



4. Identify and **list** all outliers in each distribution (**Both X and Y**) using appropriate methodology and explain why they are outliers. If you have more than 10 outliers in either distribution (X or Y) in your dataset, you can just list out the top 10 outliers.

I will employ the boxplots code to test for outliers presence. After the ploting of the boxplots for the x and y variables, the following output came out

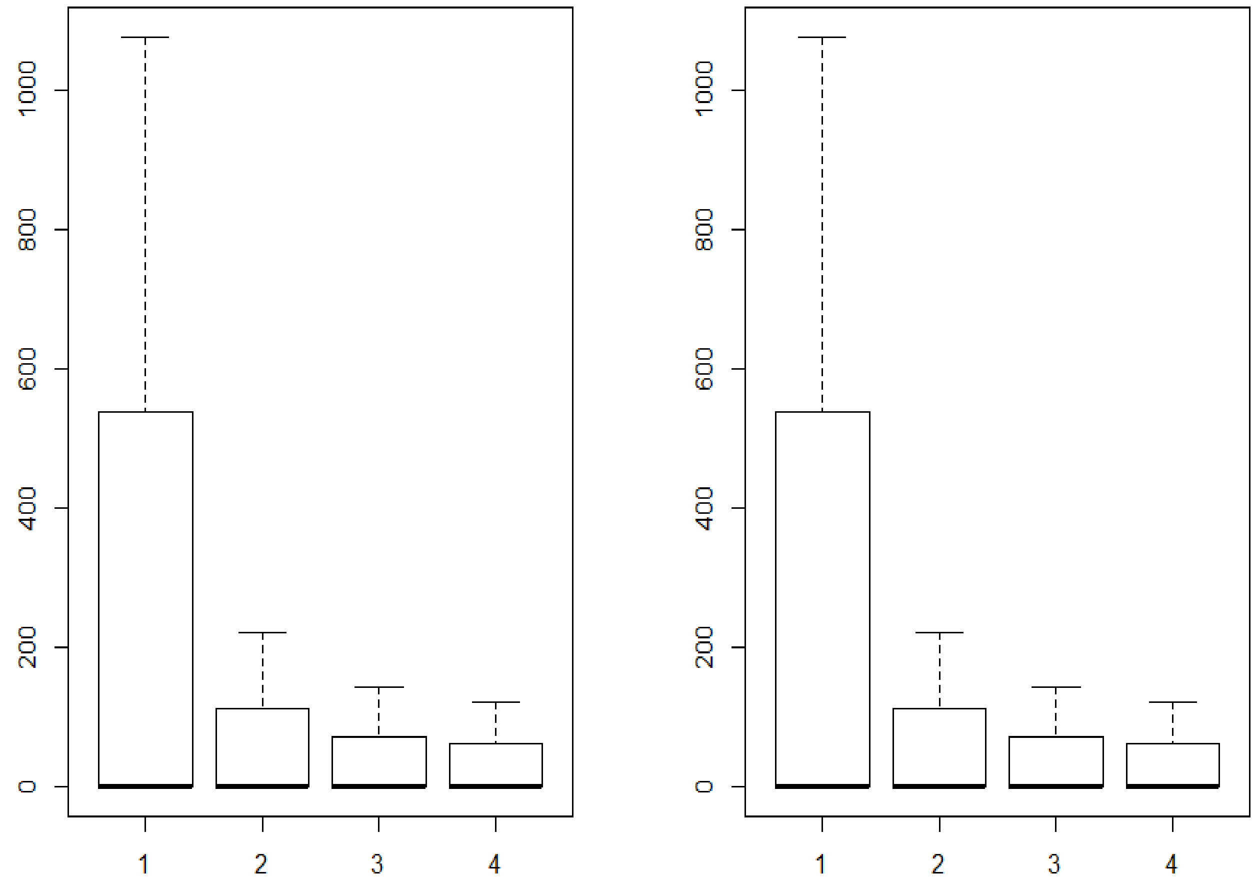


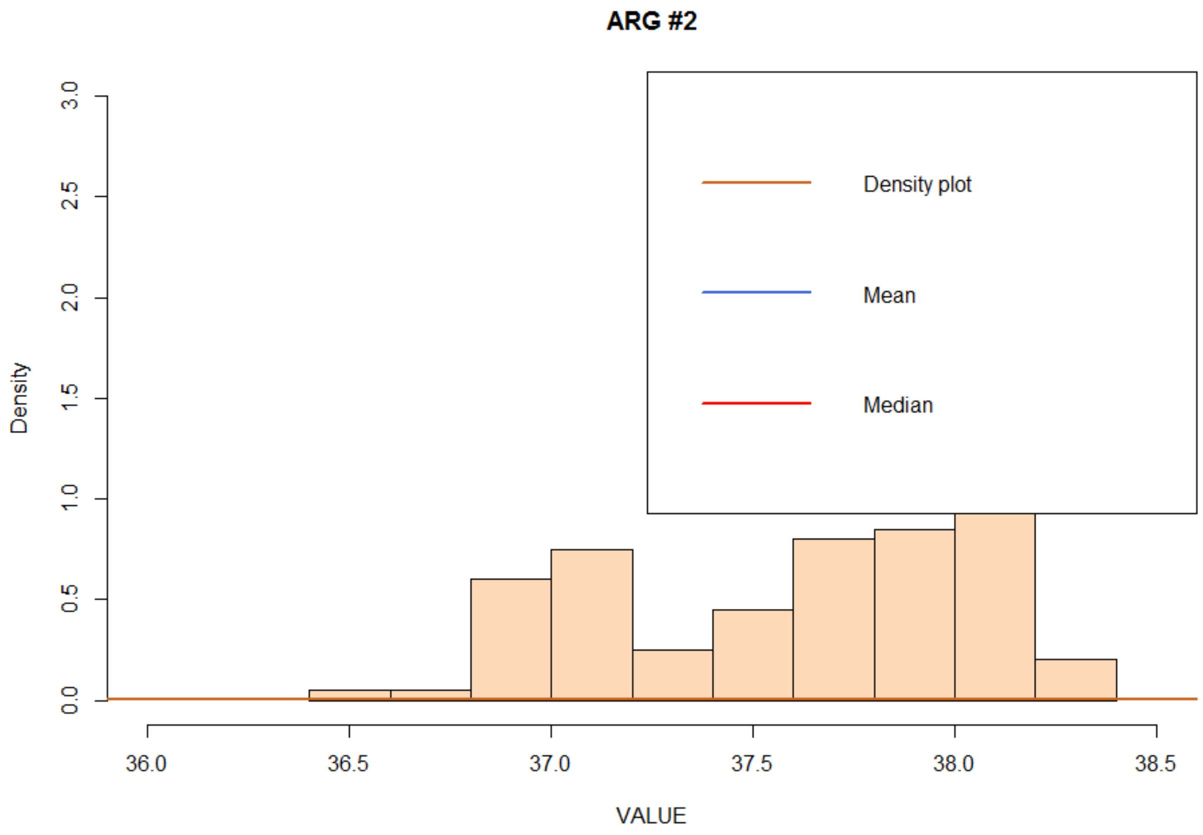
Figure 4
From figure 4 above, all the X and Y variables have an outlier. This is shown by the upper whisker being present while the lower whisker is not present.
From the R code to identify the outliers in the data, the following output shows the respective outliers in the data.

```
$stats
        [,1]   [,2]      [,3]      [,4]
[1,]    0.00   0.00   0.00000   0.00000
[2,]    0.00   0.00   0.00000   0.00000
[3,]    0.00   0.00   0.00000   0.00000
[4,]  537.35 110.75  71.20003  59.80189
[5,] 1074.70 221.50 142.40006 119.60377

$n
[1] 4 4 4 4

$conf
          [,1]     [,2]      [,3]      [,4]
```

```
[1,] -424.5065 -87.4925 -56.24802 -47.24349
[2,]  424.5065  87.4925  56.24802  47.24349

$out
numeric(0)

$group
numeric(0)

$names
[1] "1" "2" "3" "4"
```

5. Calculate the mean, median, and mode and **show where they are on the histogram graph** (you can either edit on the graph in Word ,Excel or PowerPoint, or you can show them by pen on the graph). Finish the following table for the five number summary (Minimum, Q1, median, Q3, maximum) and the z-scores of each.

	x	z-scores	**y**	z-scores
Mean		0		0
Median		0		0
Mode		0		0
Standard Deviation		NA		NA
Min	36.45			
25 percentile				
75 percentile				
Max	38.45			



6. **The Regression**: Show the output and all the plots from Excel from Simple Linear Regression analysis. You can copy and paste from Excel output and plots.

Y=-0.04417x +47.46669 For the Afghanistan mortality rates

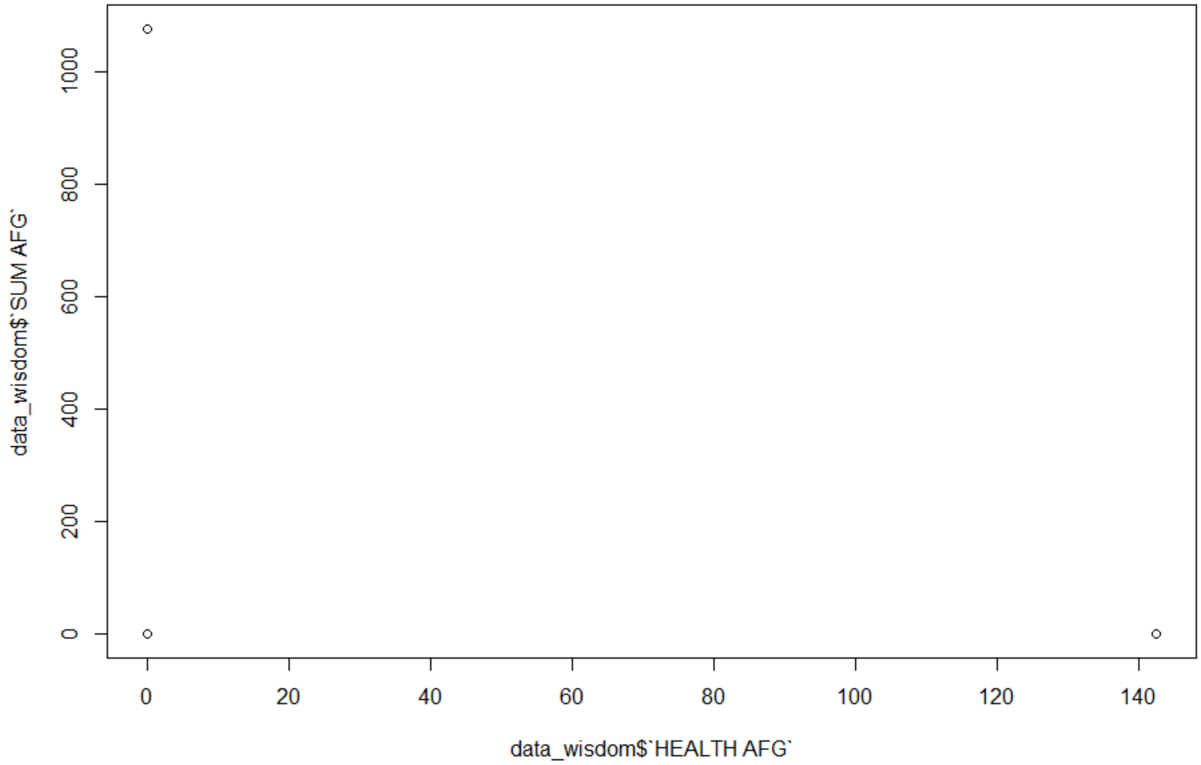
```
lm(formula = data_wisdom$`HEALTH AFG` ~ data_wisdom$`SUM AFG`)

Residuals:
     1      2      3      4 
  0.00 -47.47  94.93 -47.47 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           47.46669   47.46669     1.0    0.423
data_wisdom$`SUM AFG` -0.04417    0.08833    -0.5    0.667

Residual standard error: 82.21 on 2 degrees of freedom
Multiple R-squared:  0.1111,   Adjusted R-squared:  -0.3333 
F-statistic:  0.25 on 1 and 2 DF,  p-value: 0.6667
```

Y=-0.18X+39.87 For the Argentina mortality rates

```
lm(formula = data_wisdom$`HEALTH ARG` ~ data_wisdom$`SUM ARG`)

Residuals:
         1          2          3          4 
-3.987e+01 -5.329e-15 -3.987e+01  7.974e+01 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)              39.87      39.87     1.0    0.423
data_wisdom$`SUM ARG`    -0.18       0.36    -0.5    0.667

Residual standard error: 69.05 on 2 degrees of freedom
Multiple R-squared:  0.1111,	Adjusted R-squared:  -0.3333 
F-statistic:  0.25 on 1 and 2 DF,  p-value: 0.6667
```

7. **The Regression**: Create a scatter plot of your independent variable against the dependent variable using Excel. Make sure your dependent variable is y and your independent is x on the graph. Write a paragraph about your finding in the scatter plot.


data_wisdom$`SUM AFG`
data_wisdom$`HEALTH AFG`