

MORTALITY AND WATER HARDNESS IN ENGLAND AND WALES

PROJECT PROPOSAL:

Is hard water a hazard for health? According to the city of Cambridge (<https://www.cambridgema.gov/>), “Water described as “hard” contains high amounts of dissolved calcium and magnesium. Hard water is not a health risk but is a nuisance because of mineral buildup on plumbing fixtures’ and poor soap and or detergent performance.”

Most of the water supply systems using groundwater as a source, are concerned with water hardness because when the water moves through soil and rock it dissolves small amounts of naturally occurring minerals and as a result those natural minerals contribute into water hardness.

Over the years 1958 – 1964 a data set was collected in 61 large towns in England and Wales to further investigate the environmental cause of mortality due to water hardness in the northern and southern towns. The dataset shows the annual mortality per 100,000 for males and the calcium concentration (in parts per million) for 61 large towns in England and Wales. The higher the calcium concentration, the harder the water.

Water hardness is classified by the U.S. Department of Interior and the Water Quality Association as follows:

Classification	CaCo3 (milli gram per liter)
Soft	0 – 17.1
Slightly hard	17.1 - 60
Moderately hard	60 - 120
Hard	120 - 180
Very hard	>180

Is hard water dangerous to drink or it has a positive effect on the health of its drinkers?

Well, my main goal doing this project was the following two things:

- Are mortality and water hardness related? and
- Do either or both variables differ between northern and southern towns?

Source of dataset: (<https://vincentarelbundock.github.io/Rdatasets/datasets.html>)

EXECUTIVE SUMMARY

Water helps almost every part of the human body to function efficiently. Considering the fact that our bodies are two-third water, so it's important to know the quality of the water we are drinking. People often think that the hard water is a hazard for health, some people try to soften the hard water first and then drink it, because they think that the hard water is not healthy.

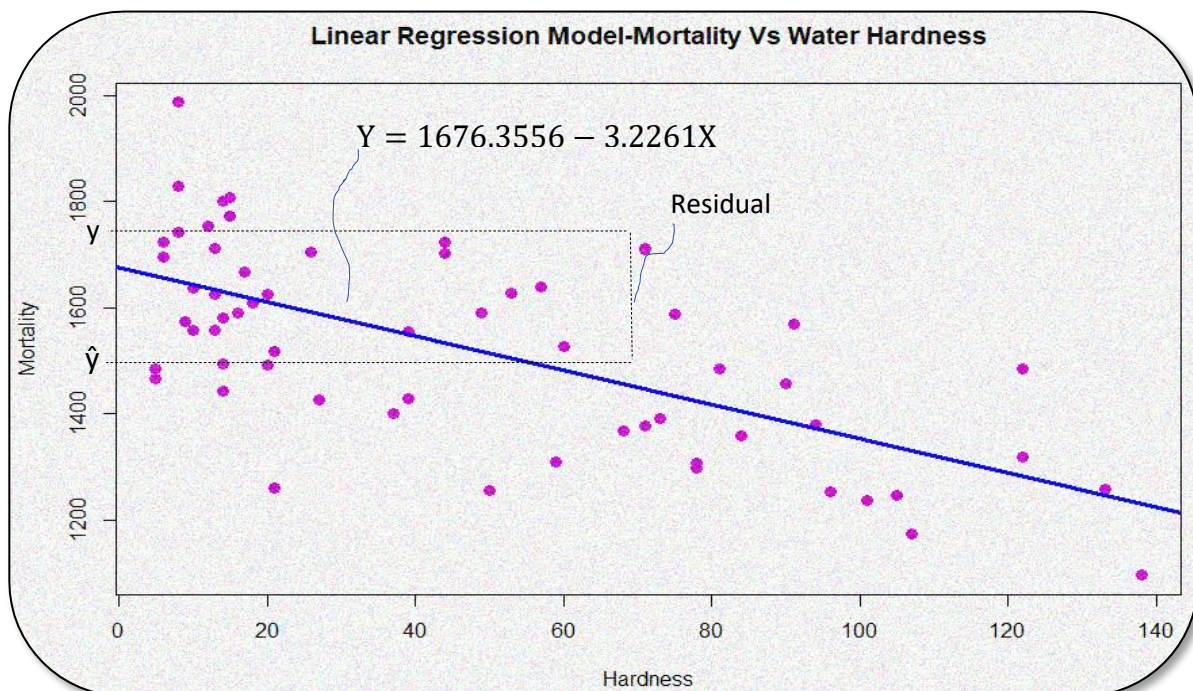
As a water supply engineer who has more than 5 years experience in designing and implementation of water supply networks, I decided to further analyze the relationship of hard water with humans health. In this project I have conducted some statistical analysis on water hardness and the mortalities rates for 61 towns in England and Wales.

Data: The dataset for this project was taken from: <https://vincentarelbundock.github.io/Rdatasets/>

The dataset has 61 towns. Towns are classified by their name and locations (categorical variables), mortality rate and the water hardness (continuous variables).

Methodology: In this project, first I conducted some descriptive analysis on each numerical variable to visualize and see their distribution over the space, estimated the skewness of both numerical variables (*Mortality*, *Hardness*) to see whether they are skewed to the right or left and finally, used the Pearson correlation and the Linear Regression equation for modeling the relationship between **Mortality** as a dependent variable and water **Hardness** as an independent variable. In addition to the linear regression model, I have also visualized the differences in mortality between the northern and southern towns.

After conducting all the statistical analysis, I figured out that there is an inverse (negative) relationship (**-0.654**) between the two numerical variables. A negative correlation also demonstrates a connection between the two variables but in an inverse way. From the figure below, we can see that when water hardness increases, the mortality decreases and when the water hardness decreases the mortality increases. So, in general we can say that the mortality in 61 towns in England and Wale is not associated/caused by water hardness, but rather the water hardness has an inverse relationship with mortality.

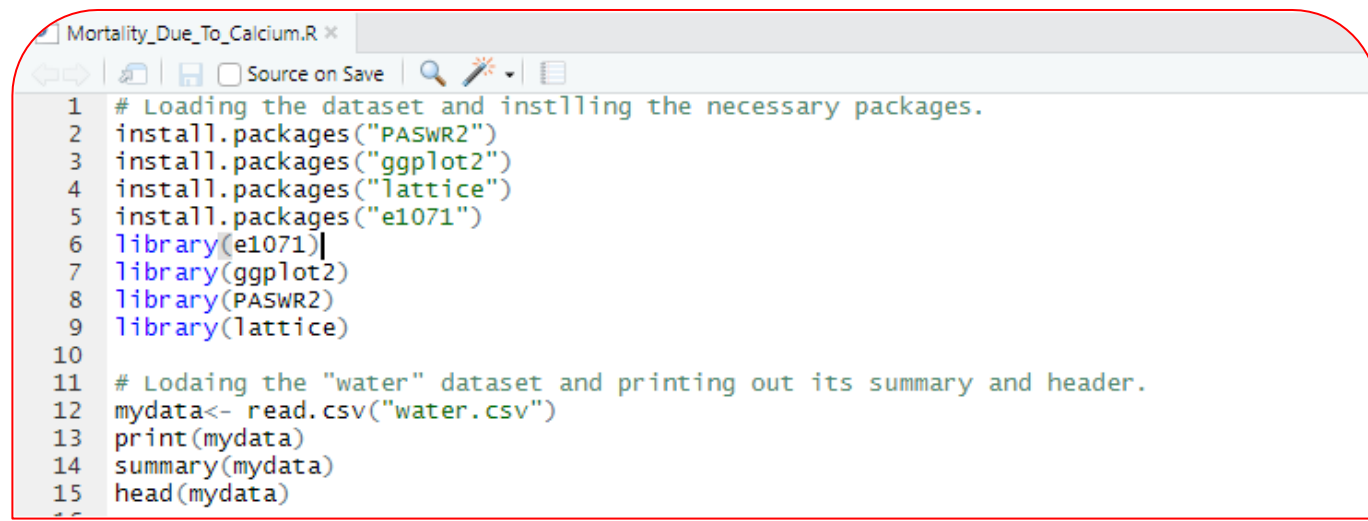


PROCEDURES:

LOADING THE DATASET AND INSTALLING THE NECESSARY PACKAGES:

First let's start with installing some of necessary packages that I might need to load, plot and read my dataset. Since the R code doesn't look nice when one copy and paste it into MS word, so I took various screenshots of the scripts and paste them here as figures. However, the full R code for my project is attached.

R Code:



```
1 # Loading the dataset and installing the necessary packages.
2 install.packages("PASWR2")
3 install.packages("ggplot2")
4 install.packages("lattice")
5 install.packages("e1071")
6 library(e1071)
7 library(ggplot2)
8 library(PASWR2)
9 library(lattice)
10
11 # Loading the "water" dataset and printing out its summary and header.
12 mydata<- read.csv("water.csv")
13 print(mydata)
14 summary(mydata)
15 head(mydata)
```

Output:

`summary(mydata)`

	x	location	town	mortality
Min.	: 1	North:35	Bath : 1	Min. :1096
1st Qu.:	16	South:26	Birkenhead: 1	1st Qu.:1379
Median	:31		Birmingham: 1	Median :1555
Mean	:31		Blackburn : 1	Mean :1524
3rd Qu.:	46		Blackpool : 1	3rd Qu.:1668
Max.	:61		Bolton : 1	Max. :1987
			(Other) :55	
hardness				
Min.	: 5.00			
1st Qu.:	14.00			
Median	: 39.00			
Mean	: 47.18			
3rd Qu.:	75.00			
Max.	:138.00			

`head(mydata)`

	x	location	town	mortality	hardness
1	1	South	Bath	1247	105
2	2	North	Birkenhead	1668	17
3	3	South	Birmingham	1466	5
4	4	North	Blackburn	1800	14
5	5	North	Blackpool	1609	18
6	6	North	Bolton	1558	10

After loading the dataset , reading its summary (1st Qu, Median, Mean, 3rd Qu and max) and the headers, I would like to do some descriptive analysis and study the distribution of each variable visually to see how the observation points are spread over the space.

USING BELOW R CODE:

```
17 # Calculating the Skewness
18 par(mar = rep(2, 4)) # Setting up the margin
19 skewness(mydata$mortality)
20 skewness(mydata$hardness)
21
22 # Plotting the data distribution in histograms.
23 hist(mydata$mortality,
24       main="Distribution of Mortality Variable",
25       ylab = "Frequency",
26       xlab = "Mortality",
27       border="blue", |
28       col="red",
29       )
30
31 hist(mydata$hardness,
32       main="Distribution of Hardness Variable",
33       ylab = "Frequency",
34       xlab = "Hardness",
35       border="blue",
36       col="orange",
37 )
```

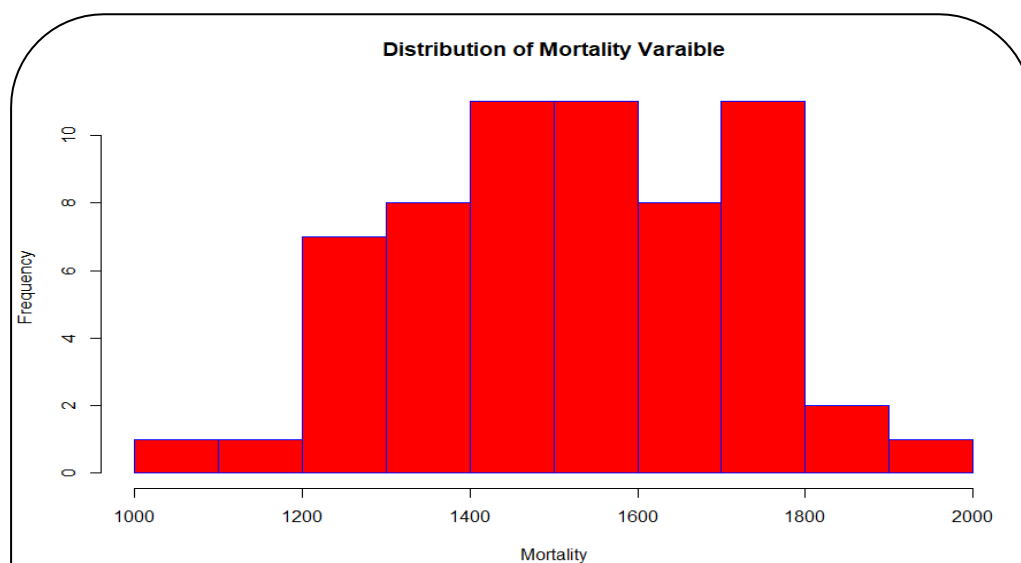
GETTING THIS OUTPUT:

Skewness for Mortality Variable:

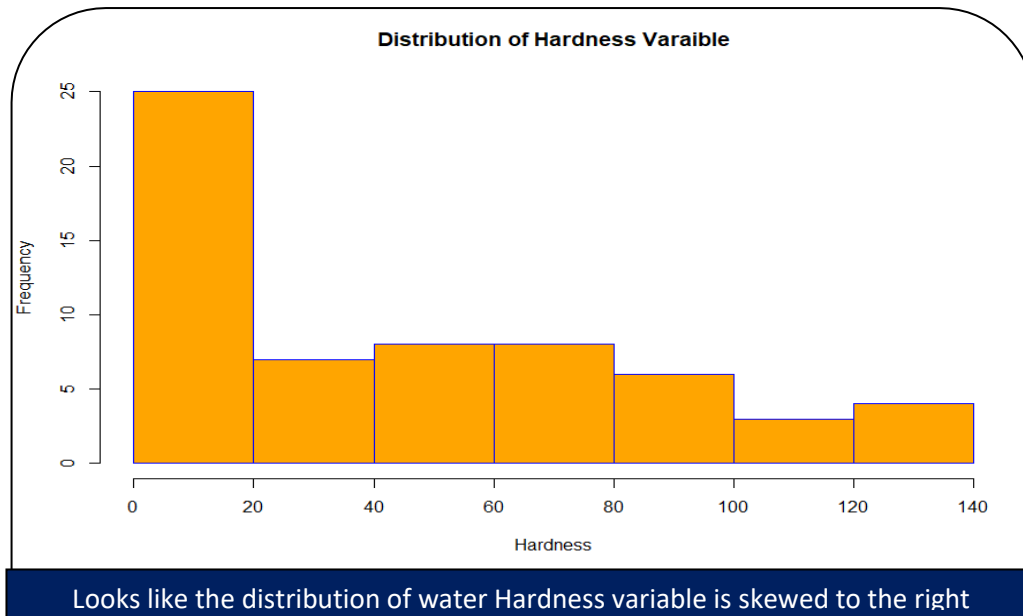
-0.08033603 – This means that the mortality distribution is skewed to the left and the mean and median are less than the mode.

Skewness for Hardness Variable:

0.6585624 - This means that the water hardness distribution is skewed to the right and the mean and median are greater than the mode.

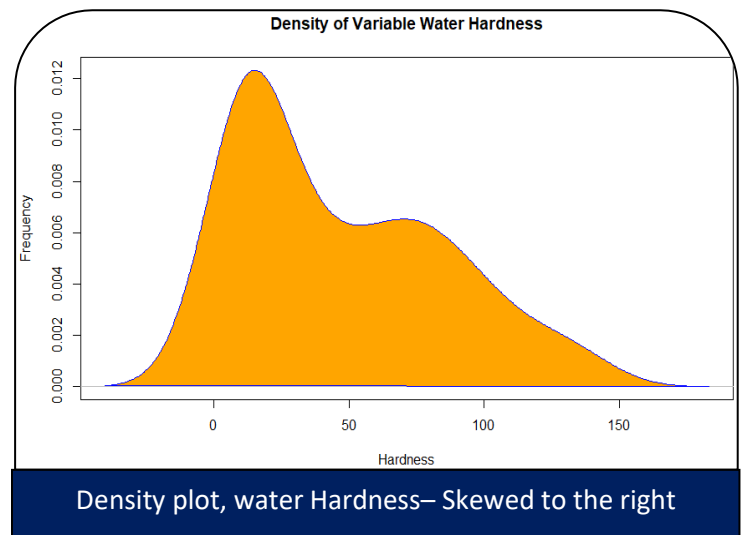
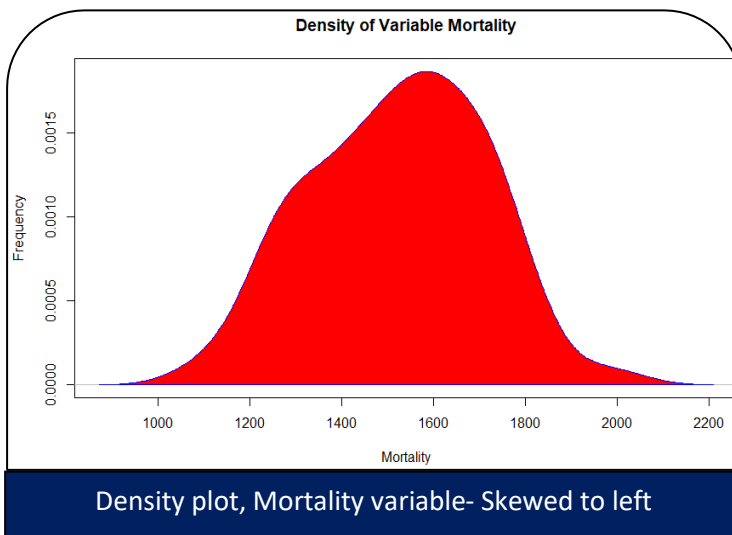


Looks like the distribution of Mortality variable is skewed to the left (negative).



By running the Kernel density equation in R, we can obtain a smoother distribution of both variables as shown in the below figures:

```
2 par(4,4)
53 d <- density(mydata$mortality)
54 d1<- density(mydata$hardness)
55 plot(d,main="Density of Variable Mortality", xlab = "Mortality", ylab = "Frequency")
56 polygon(d, col="red", border="blue")
57 plot(d1, main="Density of Variable Water Hardness", xlab = "Hardness", ylab = "Frequency")
58 polygon(d1, col="orange", border="blue")
```

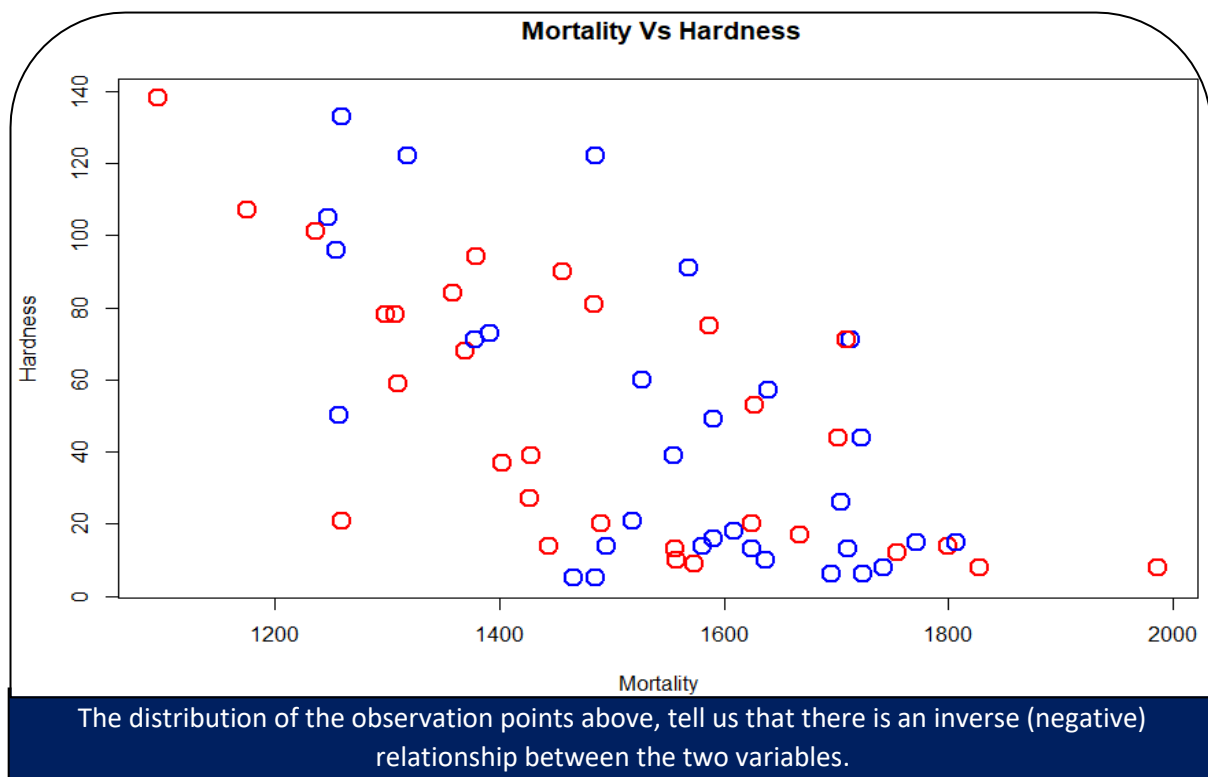


And if we compare the distribution of both variables against each other, the plot will look like this:

USING BELOW R CODE:

```
# Comparing the two variables against each other.
par(4,4)
plot(mydata$mortality, mydata$hardness, main = " Mortality vs Hardness",
     | xlab = "Mortality", ylab = "Hardness", col=c("blue", "red"), cex=2, lwd=2)
```

GETTING THIS OUTPUT:



MODELING THE LINEAR REGRESSION OF THE TWO VARIABLES, MORTALITY AND WATER HARDNESS. MORTALITY AS A DEPENDENT VARIABLE AND WATER HARDNESS AS INDEPENDENT VARIABLE.

USING BELOW R CODE:

```
0 # Modeling the Linera Relationship and Calculating the Pearson correlation
61 cor(Mortality,Hardness)
62
63 Lindear_Model<-lm(Mortality~Hardness)
64 summary(Lindear_Model)
65 plot(Mortality~ Hardness,data=mydata, main = "Linear Regression Model-Mortality vs water Hardness",
66 col="#d00fd6", cex=1.2, pch = 19)
67 abline(lm(Mortality~Hardness, data = mydata))
68 abline(Lindear_Model,col="blue", lwd=3)
```

Pearson correlation:

-0.6548486

Residuals:

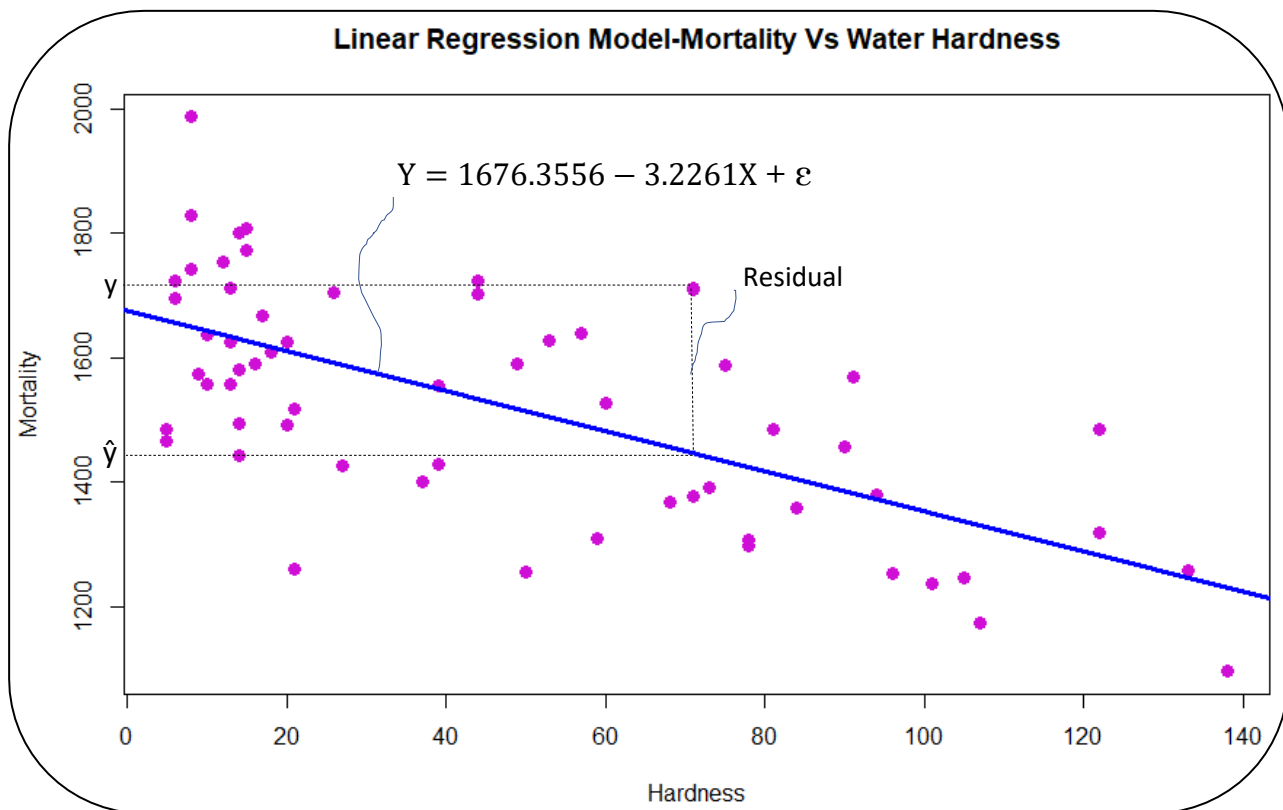
	Min	1Q	Median	3Q	Max
	-348.61	-114.52	-7.09	111.52	336.45

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1676.3556	29.2981	57.217	< 2e-16 ***
Hardness	-3.2261	0.4847	-6.656	1.03e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

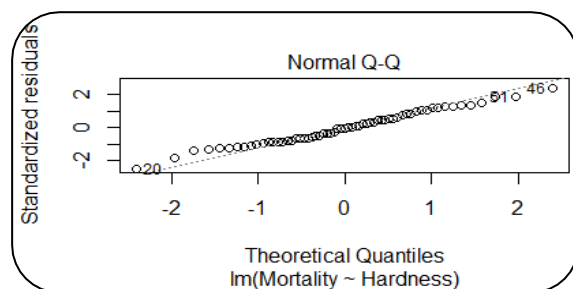
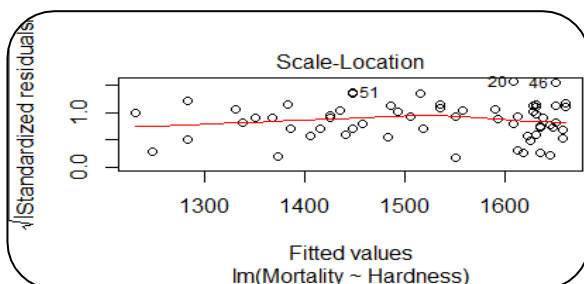
Residual standard error: 143 on 59 degrees of freedom
Multiple R-squared: 0.4288, Adjusted R-squared: 0.4191
F-statistic: 44.3 on 1 and 59 DF, p-value: 1.033e-08



The Pearson correlation coefficient (-0.654) and linear regression model above, tell us that there is an inverse (negative) relationship between the two variables. A negative correlation also demonstrates a connection between the two variables but in an inverse way. In other words, as the water hardness increases, the mortality decreases. In summary, we can say that the mortality in 61 towns in England and Wales is not associated/caused by water hardness, but rather the water hardness has an inverse relationship with mortality. And if we plug in the y-intercept (1676.3556) and the slope (-3.2261) from **Coefficient**, our linear regression equation will look like: $Y = 1676.3556 - 3.2261X$

Discussion: Various studies suggest a correlation between hard water and lower cardiovascular disease mortality; however, no firm conclusions have been made yet. The World Health Organization (WHO) is also trying to coordinate a worldwide study on the effect of cardiovascular disease before and after changes in water hardness. **Source:** (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2265038/>)

From the linear model above and the Pearson correlation coefficient, we can see that our model is not a perfect model, but rather we can see some differences between the independent variable (y) and the predicted value (\hat{y}). By running the R code: *plot (Linear Model)*, we get below graphs:



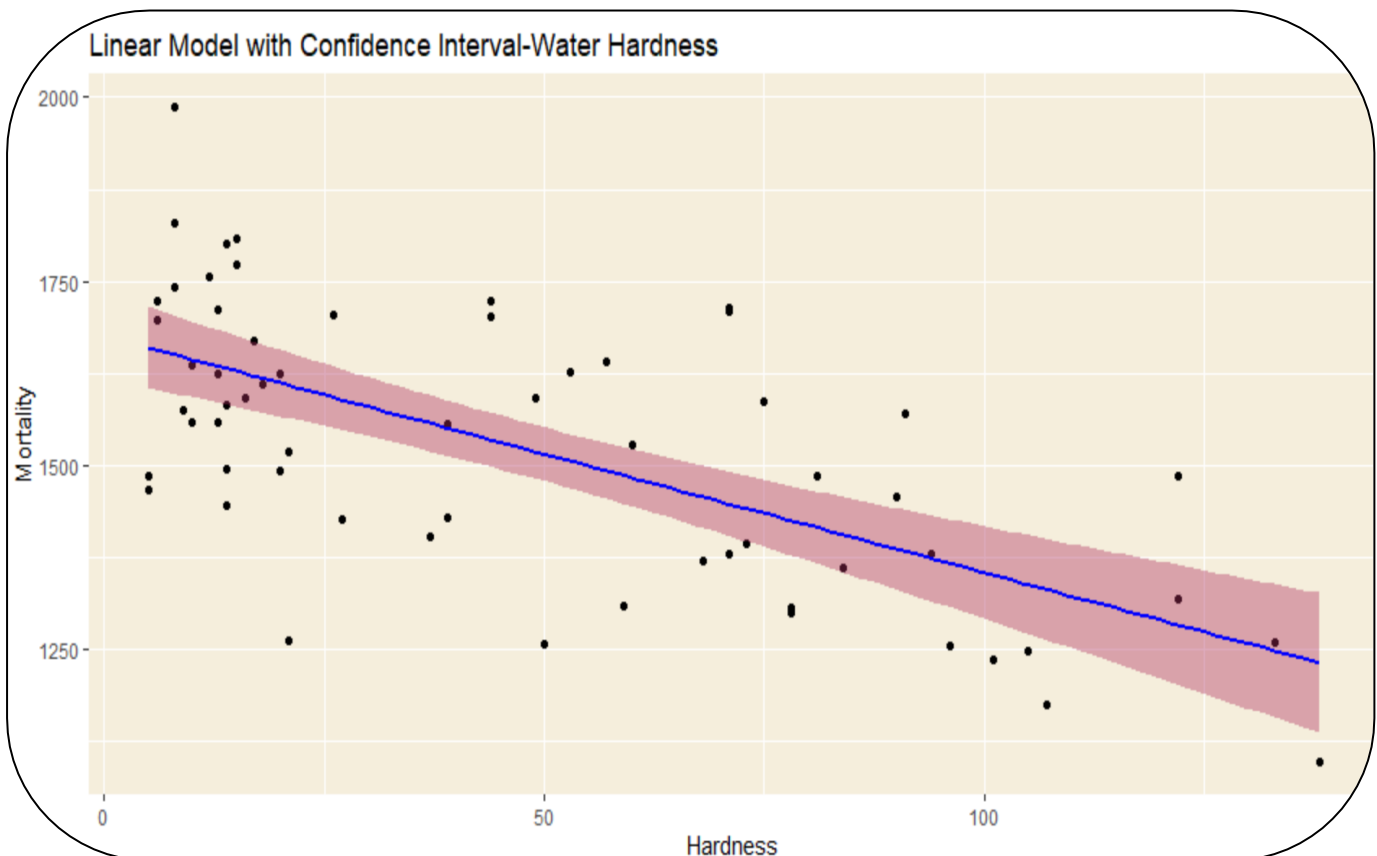
The above two graphs tell us our predicted values are not perfectly fitted on the line. That means, that there is an inverse relationship between the two variables but not a strong one.

Let's make a linear model with 95% confidence interval:

USING BELOW R CODE:

```
ggplot(mydata, aes(x=Hardness, y=Mortality)) +  
  geom_point(shape=19, color="black")+  
  geom_smooth(method=lm, linetype="solid",  
              color="blue", fill="maroon")+  
  ggtitle("Linear Model with Confidence Interval-Water Hardness")+  
  theme(panel.background = element_rect(fill = "#f5eedc"))
```

GETTING THIS OUTPUT:

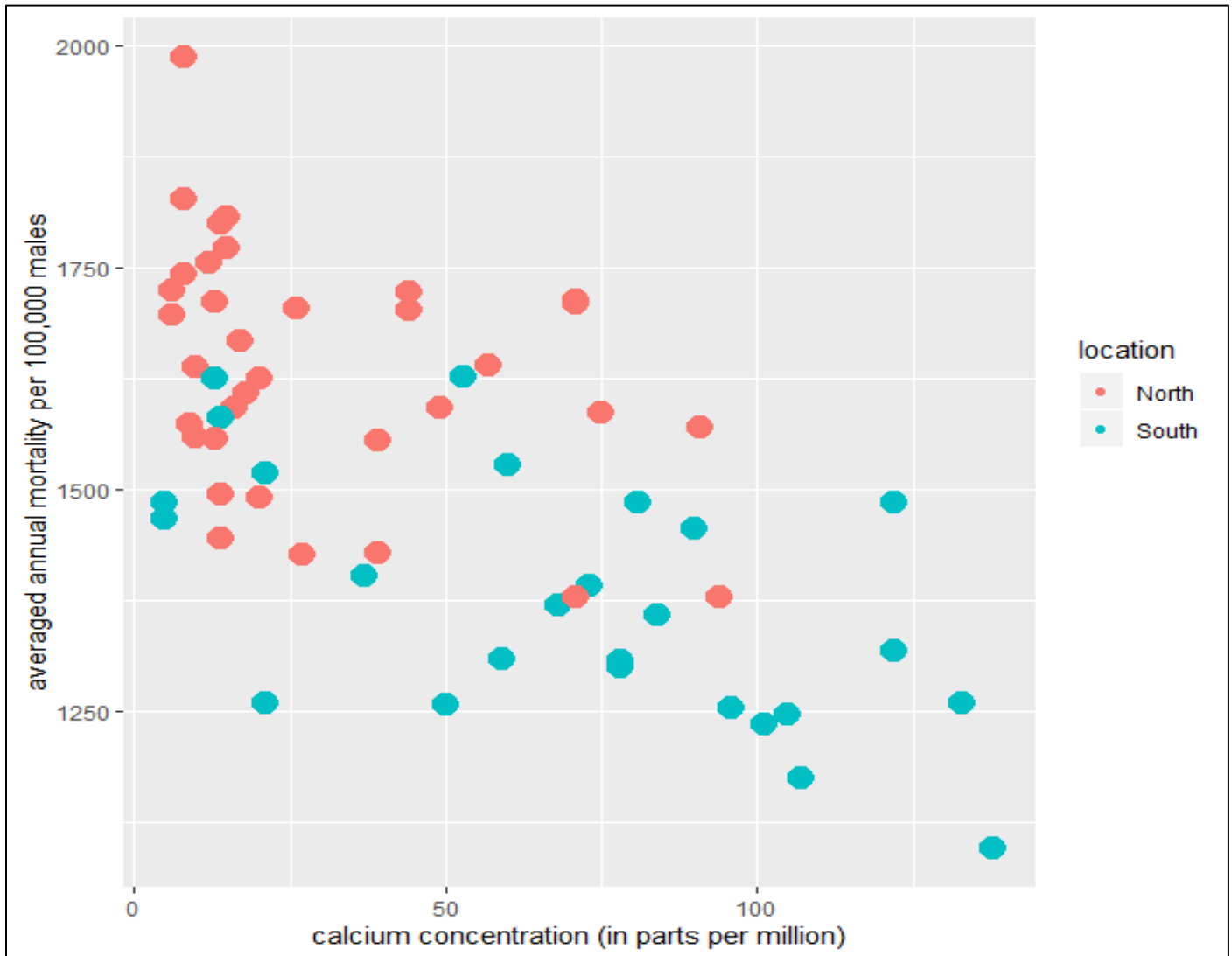


Now we know there is a moderate negative relationship between the two variables (mortality and water's hardness), but let's see how they differ between the northern and southern towns.

USING BELOW R CODE:

```
ggplot(data = mydata, aes(x = hardness, y = mortality, color = location,)) +  
geom_point() + labs( y = "averaged annual mortality per 100,000 males",  
x = "calcium concentration (in parts per million)")
```

GETTING THIS OUTPUT:



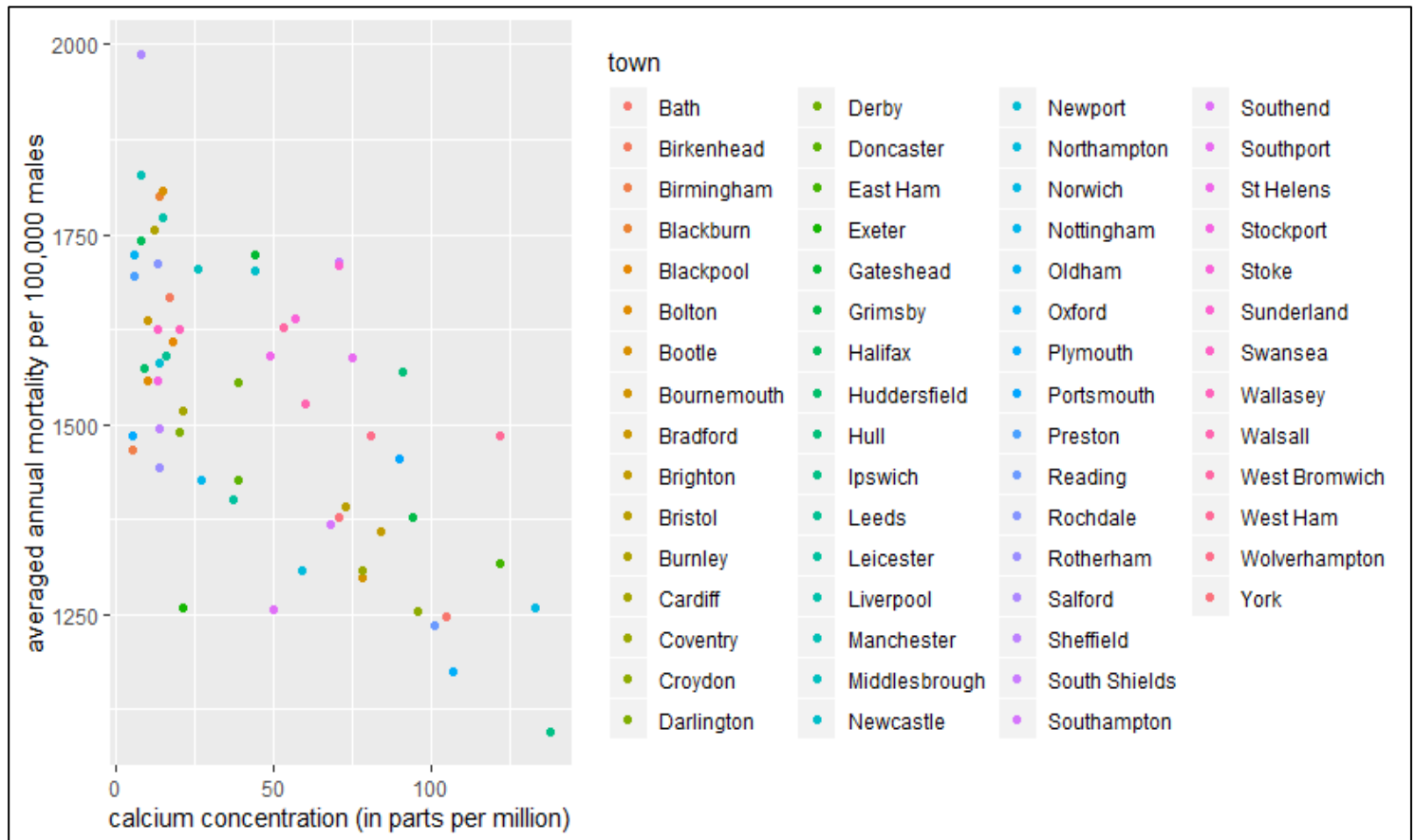
Looks like there is a moderate pattern as such; when the water hardness is increasing, the mortality for the northern towns is decreasing and even it comes to zero, while for the southern towns when the water hardness is increasing the mortality is also somehow increasing. In conclusion we can say that the cause of mortalities in 61 towns in England and Wales is not because of the water hardness, but rather it looks like the water hardness has an inverse relationship with mortalities. In other words, it looks like the water hardness have a positive effect in reducing the mortality.

Let's map the density of mortality by town:

USING BELOW R CODE:

```
# Plotting the two variables by town
ggplot(data = mydata, aes(x = hardness, y = mortality, color = town,)) +
  geom_point() + labs(y = "averaged annual mortality per 100,000 males",
                     x = "calcium concentration (in parts per million)")
```

GETTING THIS OUTPUT:



Again, when the water hardness is zero or close to zero, the mortality rate for most of the towns are high. That means, that the water hardness might be a factor in reducing mortality in those 61 towns. However, more study and research might be needed to come up with a firm conclusion.

Hypothesis Testing:

Let's suppose (*Null Hypothesis*) that the mortality in 61 towns in England and Wales is because of the water hardness. In order to accept or reject the null hypothesis, we need to perform a t-test or hypothesis testing:

USING BELOW R CODE:

```
t.test(Mortality, Hardness, var.equal = TRUE)
```

GETTING THIS OUTPUT:

```
data: Mortality and Hardness
t = 60.239, df = 120, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1428.422 1525.512
sample estimates:
mean of x mean of y
1524.14754  47.18033
```

Since the p value is much smaller than 0.05, so we reject the null hypothesis. The mortality in those 61 towns isn't because of the hard water, but rather it looks like the water hardness has some positive effect in reducing the mortality.