Homework 06 Correlation

Due by 11:59pm, Saturday, March 8, 2025, 11:59pm

S&DS 230/530/ENV 757

To make grading easier, please leave the following line of code in your assignment.

```
set.seed(1234)
```

1) Correlation of 2017 and 2018 New Haven Road Race Times *(18 pts).*

On homework 5 you created a dataset called `nhcombined` that had times for runners who ran in both the 2017 and 2018 New Haven Road Race AFTER removing unusual observations. I've put the final resulting dataset HERE.

(1.1).*(4 pts).* Read the data into an R object called `nhcombined`. Get the dimension and head of this data to make sure it has the same number of rows that you found in homework 5.

```
nhcombined <- read.csv("https://raw.githubusercontent.com/jreuning/sds230_data/refs/heads/main/nhcombine
dim(nhcombined)
```

```
## [1] 982   6
```

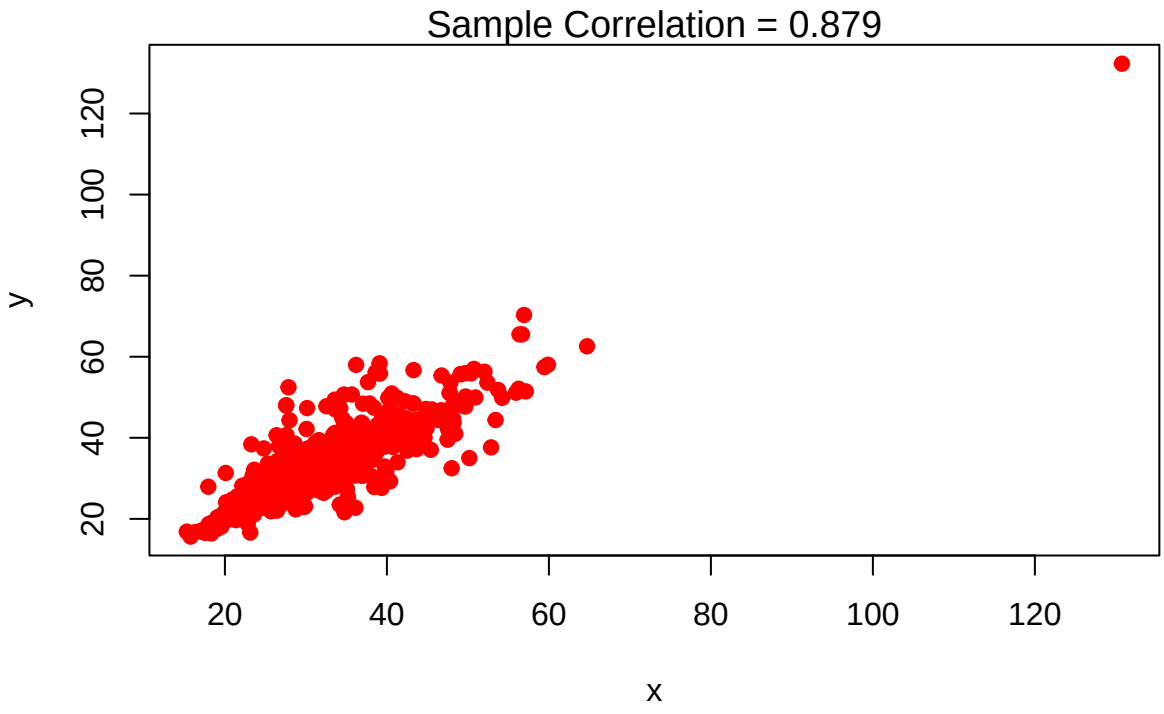
```
head(nhcombined)
```

```
##   X            Name Gender Nettime_2018 Nettime_2017 improvement
## 1 1      Abbey Shaw      F     39.25000     40.25000    1.000000
## 2 2     Abby Dziura      F     39.03333     35.63333   -3.400000
## 3 3      Abby Ganun      F     40.08333     44.65000    4.566667
## 4 4     Abi Hawkins      F     35.86667     27.56667   -8.300000
## 5 5  Abigail Murphy      F     32.88333     34.06667    1.183333
## 6 6 Abraham Cordero      M     29.63333     31.83333    2.200000
```

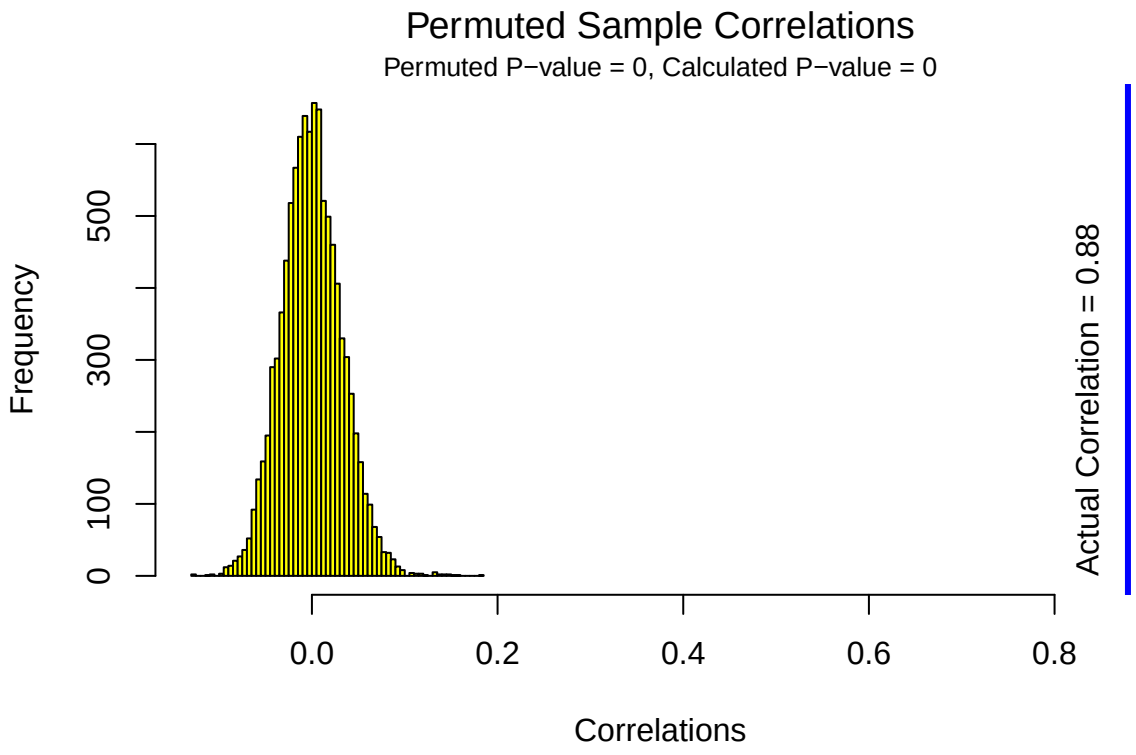
(1.2) *(10 pts)* Find the code that defines the functions `myCor()` (class 11) and `permCor()` (class 12). Paste the code for both functions in a chunk below. Run each function on net times in 2018 and 2017 for runners in your `nhcombined` dataset to make a scatterplot of these times and to run a permutation test on the correlation between times (i.e. we're seeing if times from one year to the next are correlated within runners). Be sure to include the resulting histogram and the p-value of the test. Does it seem appropriate to calculate correlation (think about outliers)?

```
myCor <- function(x, y){
  plot(x, y, pch = 19, col = "red")
  mtext(paste("Sample Correlation =", round(cor(x, y), 3)), cex = 1.2)
}
permCor <- function(x, y, n_samp = 10000, plotit = T){
   corResults <- rep(NA, n_samp)
```

```r
   for (i in 1:n_samp){
      corResults[i] <- cor(x, sample(y))
   }

   pval <- mean(abs(corResults) >= abs(cor(x, y)))

   if (plotit == T){
      #Make histogram of permuted correlations
      hist(corResults, col = "yellow", main = "", xlab = "Correlations", breaks = 50,
           xlim = range(corResults,cor(x,y)))
      mtext("Permuted Sample Correlations", cex = 1.2, line = 1)
      mtext(paste0("Permuted P-value = ", round(pval, 4),", Calculated P-value = ", round(cor.test(x, y
      abline(v = cor(x, y), col = "blue", lwd = 3)
      text(cor(x,y)*0.95, 0, paste("Actual Correlation =", round(cor(x, y), 2)), srt = 90, adj = 0)
   }
   if (plotit == F){
      return(round(pval, 5))
   }
}
myCor(nhcombined$Nettime_2017, nhcombined$Nettime_2018)
```



```
permCor(nhcombined$Nettime_2017, nhcombined$Nettime_2018)
```



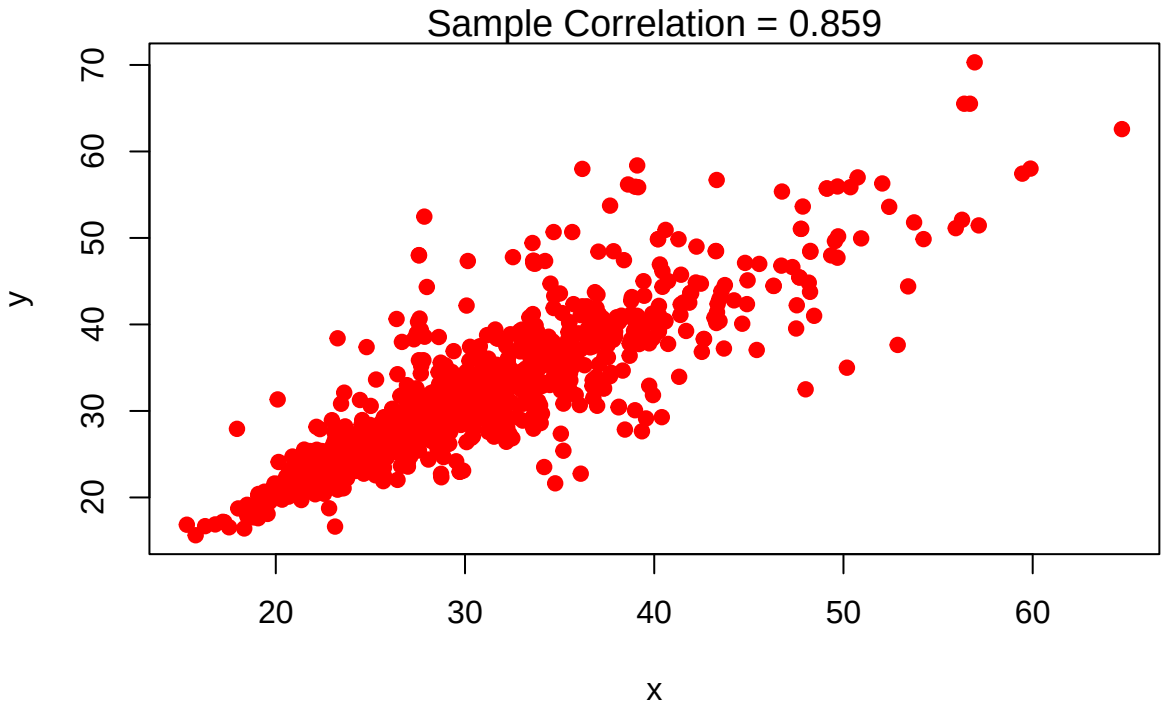

```
permCor(nhcombined$Nettime_2017, nhcombined$Nettime_2018, plotit = F)
```

```
## [1] 0
```

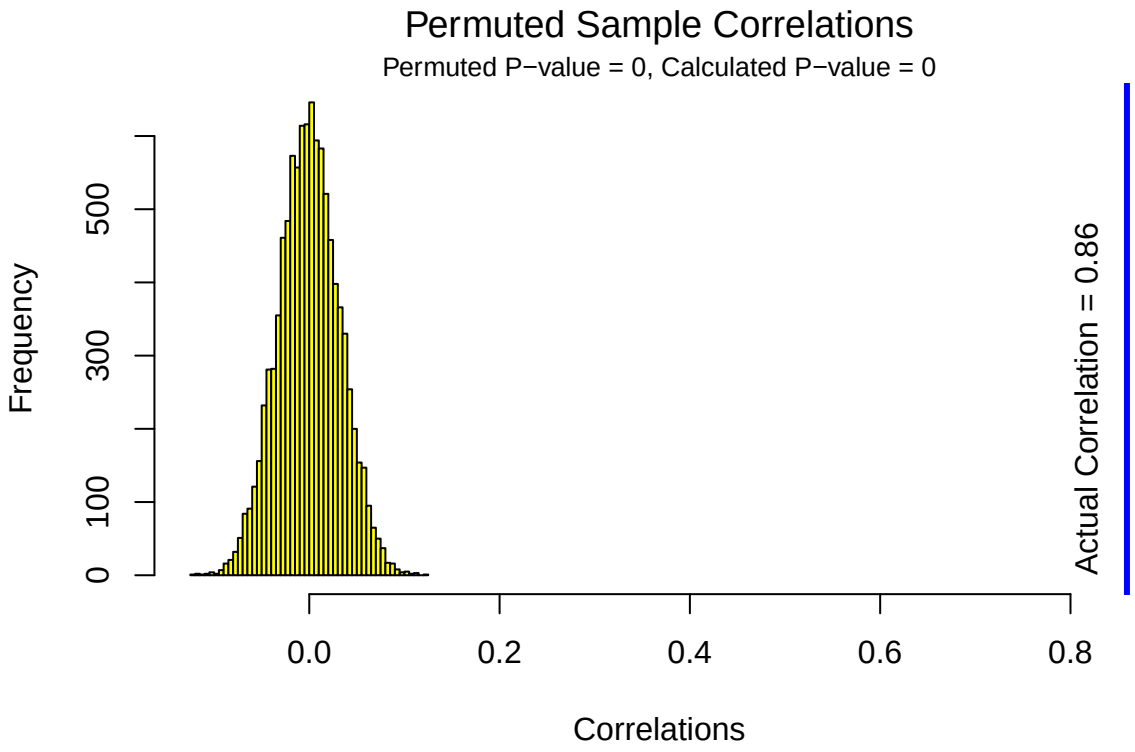
It is not appropriate to calculate correlation here because there is one outlier that can heavily influence the correlation.

(1.3) *(4 pts)* Repeat part 1.2), but first remove the one individual who had times over 120 minutes in both years. Does it seem appropriate to calculation correlation now? How does the correlation change by removing the one outying value? Is the correlation statistically significantly different from zero?

```
nhcombined <- nhcombined[nhcombined$Nettime_2017 < 120,]
myCor(nhcombined$Nettime_2017, nhcombined$Nettime_2018)
```




```
permCor(nhcombined$Nettime_2017, nhcombined$Nettime_2018)
```




```
permCor(nhcombined$Nettime_2017, nhcombined$Nettime_2018, plotit = F)
```

```
## [1] 0
```

It is now appropriate to calculate correlation now that there are no outliers and that the net times in 2017 and 2018 appear to have a linear relationship. The sample correlation decreased from 0.879 to 0.859 after removing the outlier. The sample correlation is statistically significantly different from 0 since the p value is 0 which is less than any significance level.

2) Correlation and the Bootstrap. *(50 pts)*

World Bank Data: CO2 Emissions per capita and Energy Use per capita

A `.CSV` file containing world bank data from 2016 can be downloaded HERE.

The variables we'll consider in this example are:

- `Country` : countryname
- `EnergyUse`: kilograms of oil equivalent usage per capita
- `CO2`: CO2 emissions per capita (metric tons)

In this problem, we will take a closer look at the correlation between Energy Use per capita and CO2 Emissions per capita.

(2.1) *(3pts)* Read the data into an object called `wb`. Then make a dataframe called `wb2` that has complete data but only contains the country name, EnergyUse and CO2. Finally, create two new variables `logEnergy` and `logCO2` which are the natural log of each of these variables, respectively. You only need to retain these 5 columns in `wb2`.

```
wb <- read.csv("https://raw.githubusercontent.com/jreuning/sds230_data/refs/heads/main/WB.2016.csv")
wb2 <- wb[!is.na(wb$EnergyUse) & !is.na(wb$CO2),
          c("Country", "EnergyUse", "CO2")]
wb2$logEnergy <- log(wb2$EnergyUse)
wb2$logCO2 <- log(wb2$CO2)
head(wb2)
```

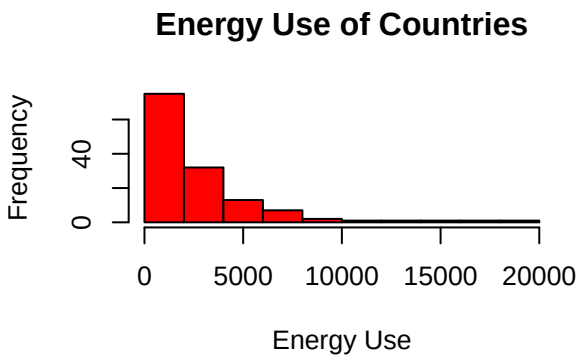
```
##      Country EnergyUse       CO2 logEnergy    logCO2
## 2    Albania  808.4558  1.978763  6.695126 0.6824721
## 3    Algeria 1321.0995  3.717410  7.186220 1.3130272
## 6     Angola  545.0405  1.291328  6.300860 0.2556714
## 8  Argentina 2015.1870  4.746797  7.608467 1.5574702
## 9    Armenia 1018.0712  1.902759  6.925665 0.6433049
## 11 Australia 5328.2240 15.370138  8.580773 2.7324265
```

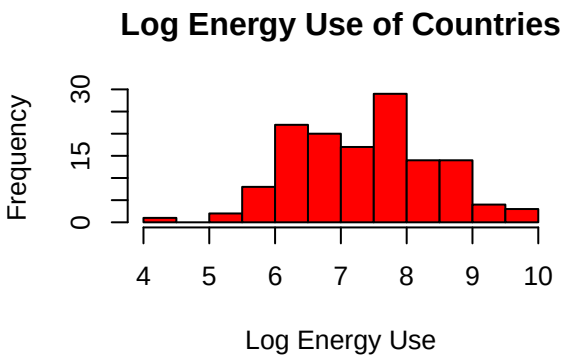
```
dim(wb2)
```

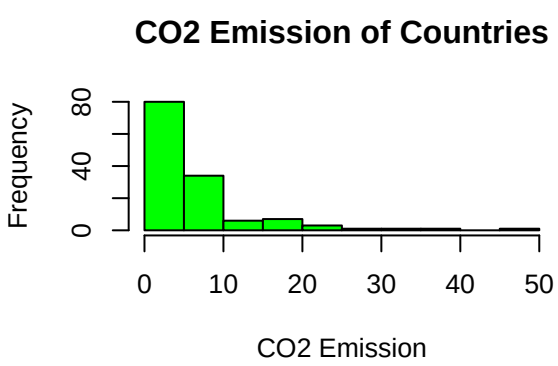
```
## [1] 134   5
```

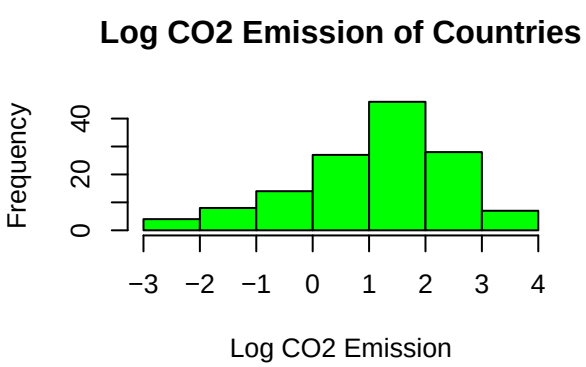
(2.2) *(6 pts)* Make histograms of CO2 and EnergyUse on both the raw scale and the log scale. Then make a scatterplot of CO2 (vertical axis) and EnergyUse (horizontal axis) on both the raw scale and the log scale. Put appropriate labels on all plots and make them look interesting (i.e. use a non-default color at least). Add the correlation to two decival places between variables as a subtitle. Describe in a few words the shape of each distribution. Comment on the appropriateness of calculating correlation on both the raw scale and the log scale and comment on the correlation itself.

```
par(mfrow=c(2, 2))
hist(wb2$EnergyUse,
     xlab = "Energy Use",
     main="Energy Use of Countries",
     col="red")
hist(wb2$logEnergy,
     xlab = "Log Energy Use",
     main="Log Energy Use of Countries",
     col="red")
hist(wb2$CO2,
     xlab = "CO2 Emission",
     main="CO2 Emission of Countries",
     col="green")
hist(wb2$logCO2,
     xlab = "Log CO2 Emission",
     main="Log CO2 Emission of Countries",
     col="green")
```

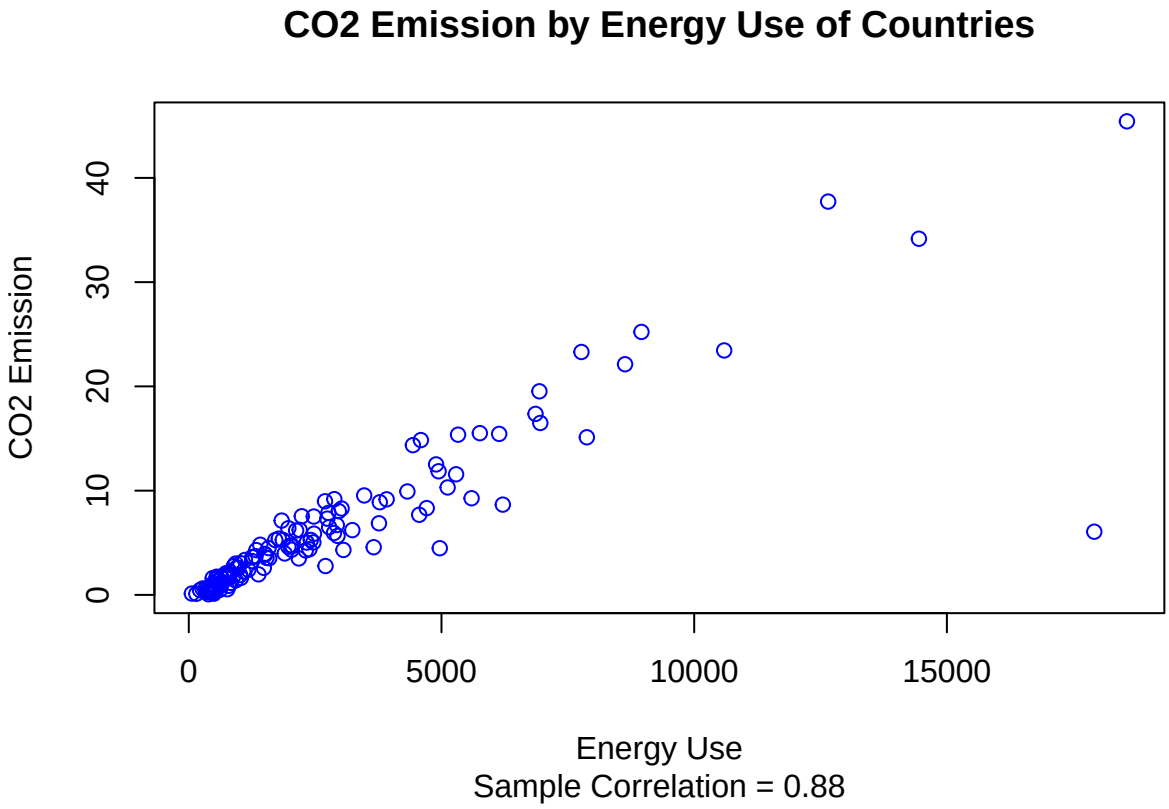




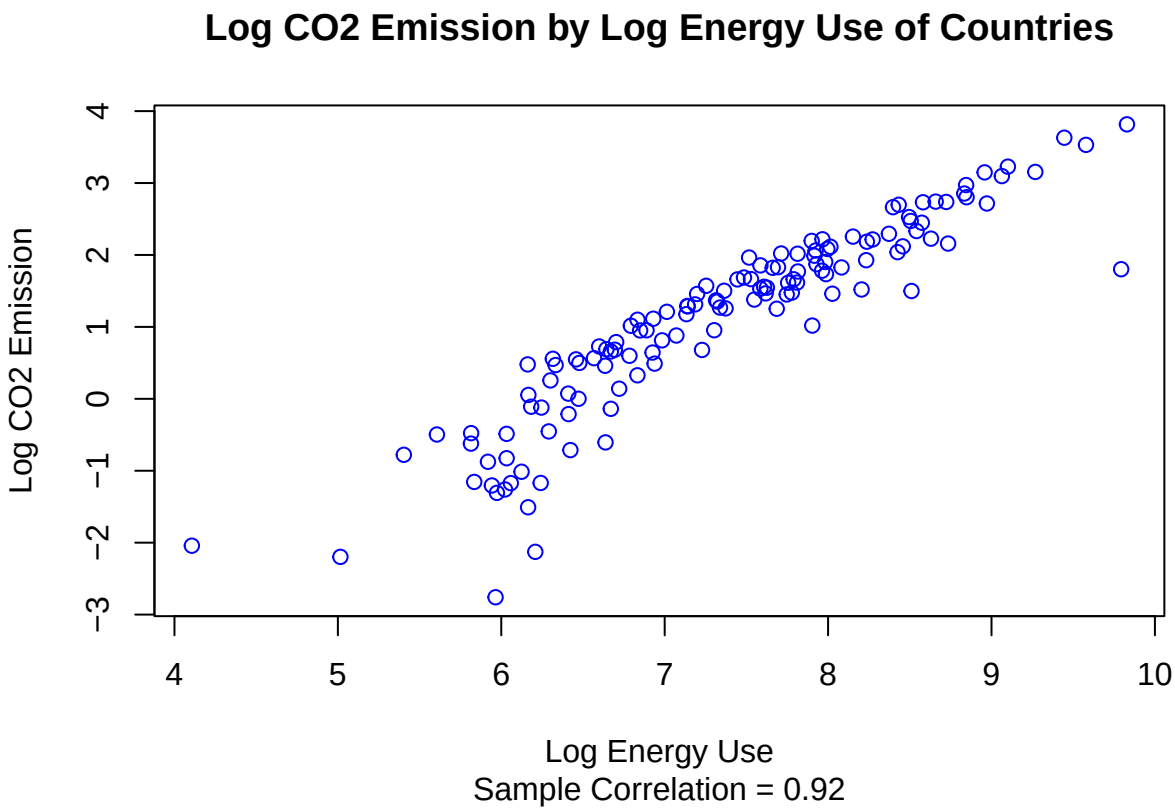





```
par(mfrow=c(1, 1))
plot(wb2$CO2 ~ wb2$EnergyUse,
     col="blue",
     main="CO2 Emission by Energy Use of Countries",
     sub=paste("Sample Correlation =",
               round(cor(wb2$EnergyUse, wb2$CO2), 2)),
     xlab="Energy Use",
     ylab="CO2 Emission")
```




```
plot(wb2$logCO2 ~ wb2$logEnergy,
     col="blue",
     main="Log CO2 Emission by Log Energy Use of Countries",
     sub=paste("Sample Correlation =",
               round(cor(wb2$logEnergy, wb2$logCO2), 2)),
     xlab="Log Energy Use",
     ylab="Log CO2 Emission")
```



Calculating the correlation is not appropriate for the raw scale since the data is more densely packed for lower values than it is for higher value, so it is difficult to see from the graph whether the relationship appears to be linear. Calculating the correlation for the log scale is appropriate since the data is more normally distributed and there appears to be a linear relationship between the two log variables. Under the assumption that both of the relationships are linear, there is a strong, positive, linear correlation between energy use and CO2 emission, and 77.4% of the variation in CO2 emission can be explained by the variation in energy use. There is a strong positive linear correlation between the log energy use and the log CO2 emission, and 84.6% of the variation in log CO2 emission can be explained by the variation in the log energy use.

(2.3) *(4pts)* Identify the one country that is an outlier from the general trend (i.e. country with very HIGH Energy Usage and very LOW Emissions per capita). Give a one sentence reason why this country has the values it does (i.e. why is it an outlier from the trend).

```
wb2[(wb2$EnergyUse > 15000 & wb2$CO2 < 10),"Country"]
```

```
## [1] "Iceland"
```

The majority of Iceland's energy comes from renewable sources that do not have CO2 emissions.

(2.4) *(4 pts)* Calculate (and display) a parametric test of the significance of the correlation between logCO2 and logEnergy, and save the result in an object called `cor1test`. Is the correlation statistically significantly different from zero? What is a 96% confidence interval for the location of the true correlation?

```
cor1test <- cor.test(wb2$logCO2, wb2$logEnergy, method = "pearson", conf.level=0.96)
cor1test
```

```
##
##  Pearson's product-moment correlation
##
## data:  wb2$logCO2 and wb2$logEnergy
## t = 27.739, df = 132, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 96 percent confidence interval:
##  0.8927879 0.9462201
## sample estimates:
##      cor
## 0.923886
```

*The correlation is statistically significantly different from zero since the p value is about 2*10^-16, which is less than any reasonable significance level. The 96% confidence interval is (0.893, 0.946).*

(2.5) *(8 pts)* Get 10000 bootstrap samples from `wb2` following the example given in Class 11. For each sample, fit a regression line predicting log(CO2) based on log(Energy Use) - save results in an object called `bresults`. In addition, calculate the correlation of log(CO2) and log(Energy Use) in your sample - save results in an object called `corResults`.

```
n_bootstrap = 10000
bresults <- c()
corResults <- c()
for (i in 1:n_bootstrap) {
  bootstrap_sample <- wb2[sample(nrow(wb2), replace = TRUE),]
  logCO2_sample <- bootstrap_sample$logCO2
  logEnergy_sample <- bootstrap_sample$logEnergy
  bresults[i] <- coef(lm(logCO2_sample ~ logEnergy_sample))[2]
  corResults[i] <- cor(logEnergy_sample, logCO2_sample)
}
```

(2.6) *(3 pts)* Calculate a 96% bootstrap confidence interval for the true correlation and the true slope based on quantiles. Save results in an objects called `ci_r` (for correlation) and `ci_slope` (for regression slope). Display the results for these intervals.

```
ci_slope <- quantile(bresults, c(0.02, 0.98))
ci_r <- quantile(corResults, c(0.02, 0.98))
ci_slope
```

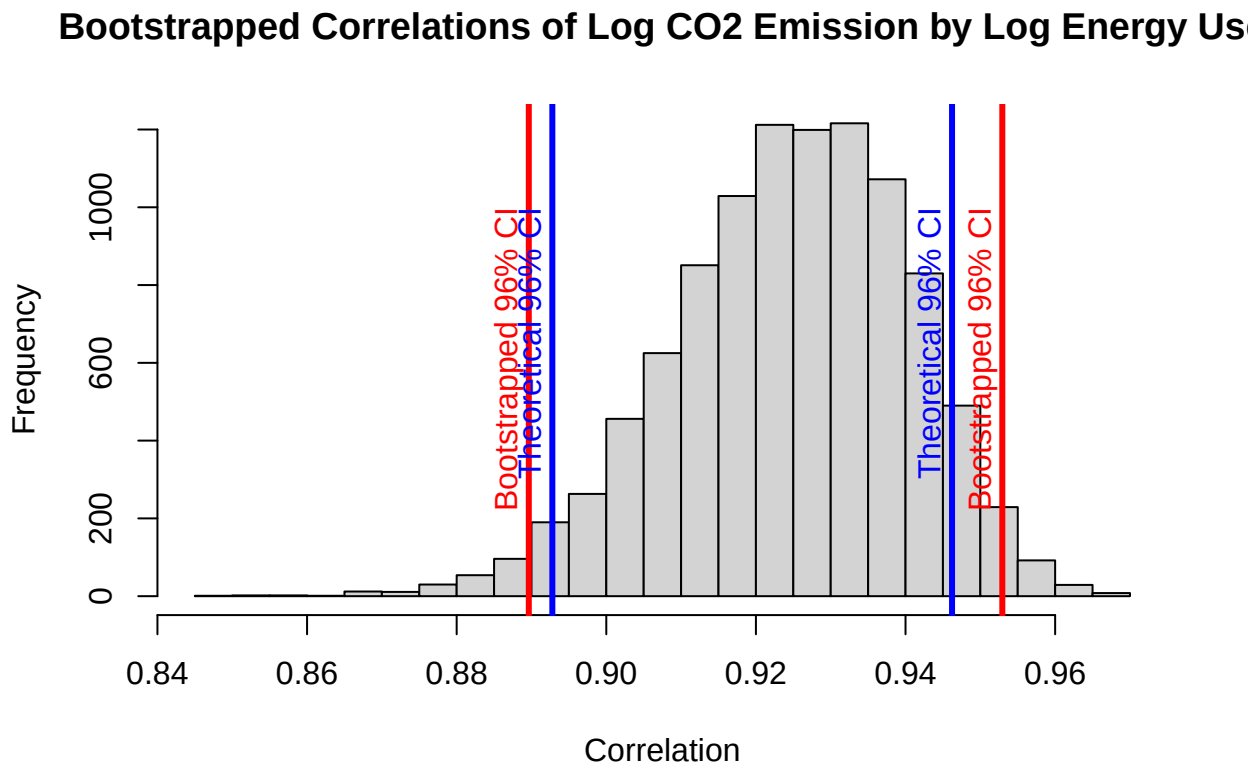
```
##       2%      98%
## 1.047191 1.279183
```

```
ci_r
```

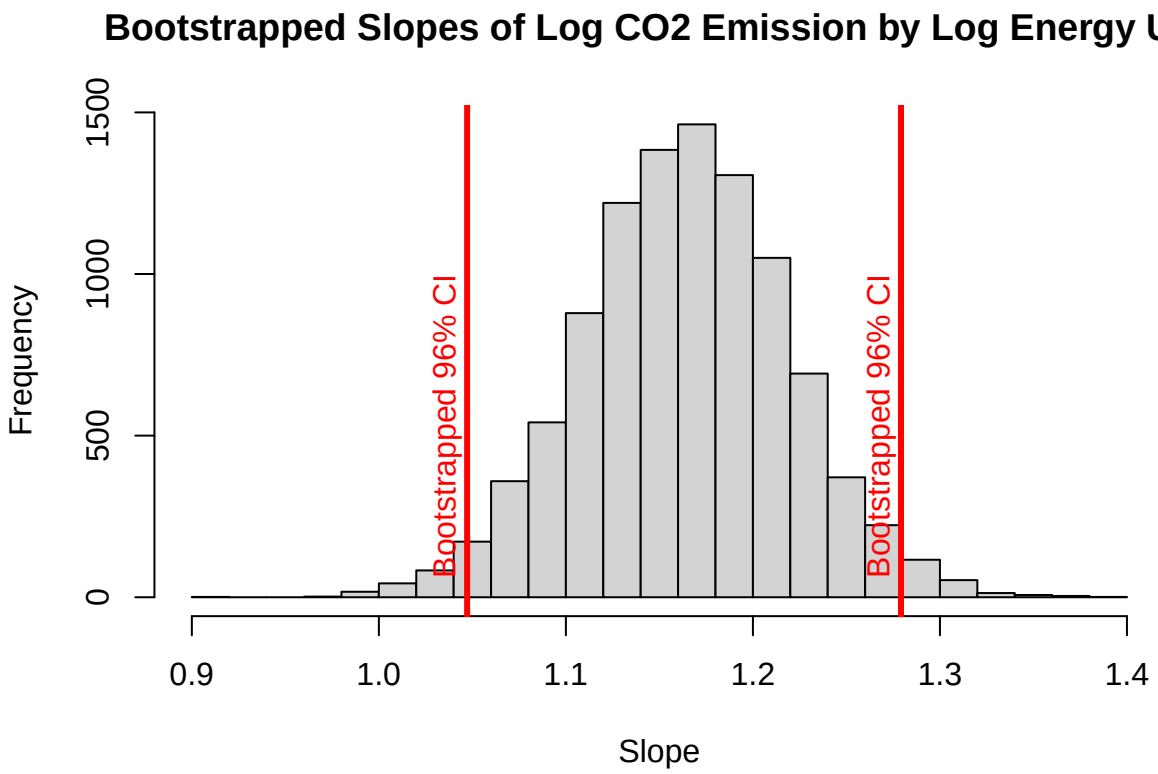
```
##        2%       98%
## 0.8896408 0.9529394
```

(2.7) *(4 pts)* Make a histogram of the results in `corResults`. Add vertical lines for both the theoretical confidence interval and the bootstrapped confidence interval. Make a similar histogram for the regression slopes (i.e. values in `bResults` with corresponding confidence intervals).

```
hist(corResults,
     breaks=20,
     xlab="Correlation",
     main="Bootstrapped Correlations of Log CO2 Emission by Log Energy Use")
abline(v=ci_r, col="red", lwd=3)
text(ci_r, 1000, "Bootstrapped 96% CI", pos = 2, col = "red", srt = 90)
abline(v=cor1test$conf.int, col="blue", lwd=3)
text(cor1test$conf.int, 1000, "Theoretical 96% CI", pos = 2, col = "blue", srt = 90)
```




```
hist(bresults,
     breaks=20,
     xlab="Slope",
     main="Bootstrapped Slopes of Log CO2 Emission by Log Energy Use")
abline(v=ci_slope, col="red", lwd=3)
text(ci_slope, 1000, "Bootstrapped 96% CI", pos = 2, col = "red", srt = 90)
```



(2.8) *(3 pts)* Comment on any differences you observe between the bootstrapped CI and the theoretical CI.

The theoretical CI is slightly tighter than the bootstrapped CI.

(2.9) *(9 pts)* Repeat parts 2.5 through 2.8 but on the RAW SCALE (i.e. use EnergyUse and CO2, not logEnergyUse and logCO2).

```
cor1test <- cor.test(wb2$CO2, wb2$EnergyUse, method = "pearson", conf.level=0.96)
bresults <- c()
corResults <- c()
for (i in 1:n_bootstrap) {
  bootstrap_sample <- wb2[sample(nrow(wb2), replace = TRUE),]
  CO2_sample <- bootstrap_sample$CO2
  Energy_sample <- bootstrap_sample$EnergyUse
  bresults[i] <- coef(lm(CO2_sample ~ Energy_sample))[2]
  corResults[i] <- cor(Energy_sample, CO2_sample)
}
ci_slope <- quantile(bresults, c(0.02, 0.98))
ci_r <- quantile(corResults, c(0.02, 0.98))
ci_slope
```

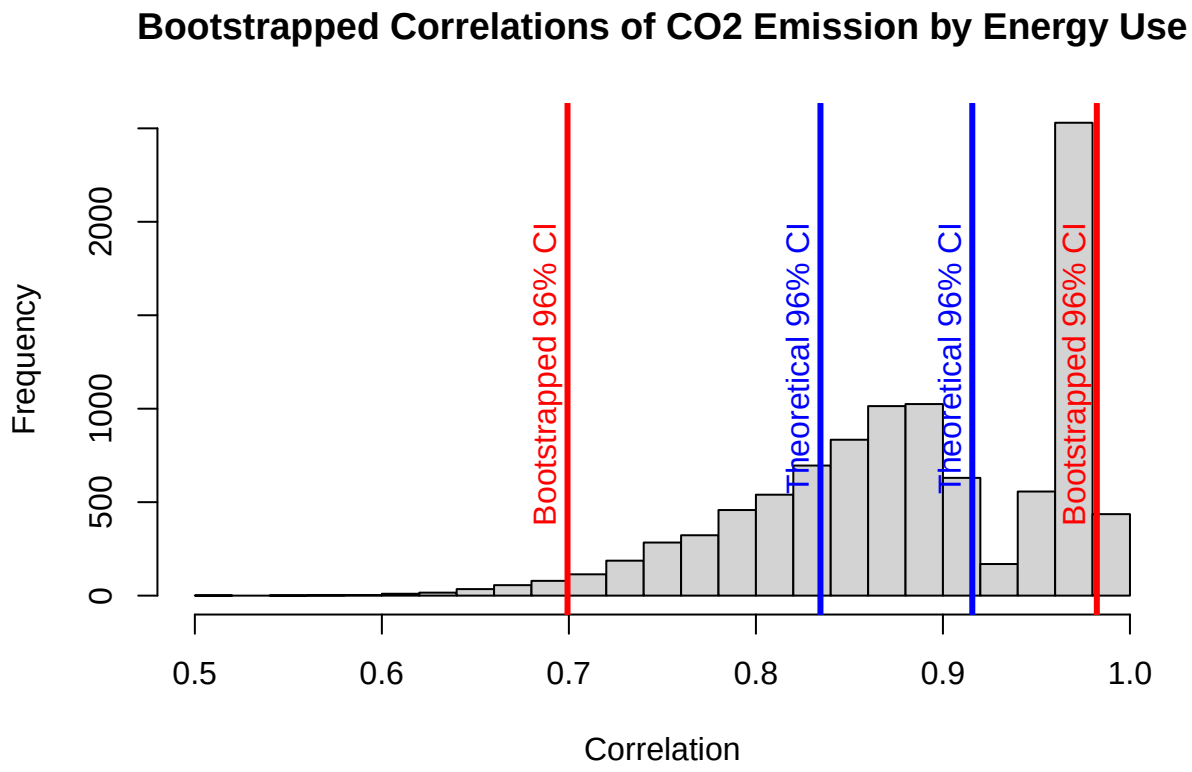
```
##          2%         98%
## 0.001282438 0.002619456
```

```
ci_r
```

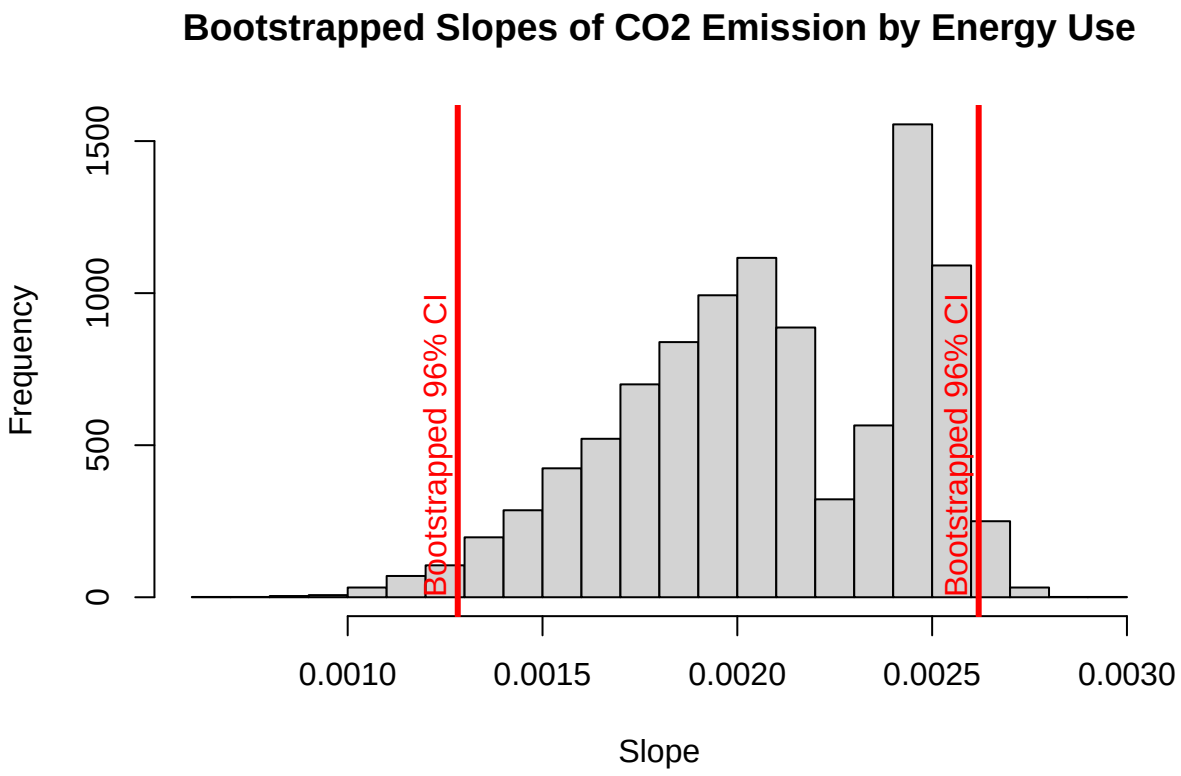
```
##        2%       98%
```

```
## 0.6992708 0.9822926
```

```r
hist(corResults,
     breaks=20,
     xlab="Correlation",
     main="Bootstrapped Correlations of CO2 Emission by Energy Use")
abline(v=ci_r, col="red", lwd=3)
text(ci_r, 2000, "Bootstrapped 96% CI", pos = 2, col = "red", srt = 90)
abline(v=cor1test$conf.int, col="blue", lwd=3)
text(cor1test$conf.int, 2000, "Theoretical 96% CI", pos = 2, col = "blue", srt = 90)
```




```r
hist(bresults,
     breaks=20,
     xlab="Slope",
     main="Bootstrapped Slopes of CO2 Emission by Energy Use")
abline(v=ci_slope, col="red", lwd=3)
text(ci_slope, 1000, "Bootstrapped 96% CI", pos = 2, col = "red", srt = 90)
```




```
# The distribution of bootstrapped correlations is no longer normally distributed
# Instead, it seems to be bimodal with one peak at around 0.88 and another
# peak, this one quite large, at around 0.96. The data is also left skewed.
# The theoretical CI is now much tighter than the bootstrapped CI.
# The bootstrapped slopes seem to follow a similar distribution that is binomal
# and left skewed.
```

2.10) *(8 pts)* Repeat 2.9, but without Iceland. Write a few sentences on what changes.

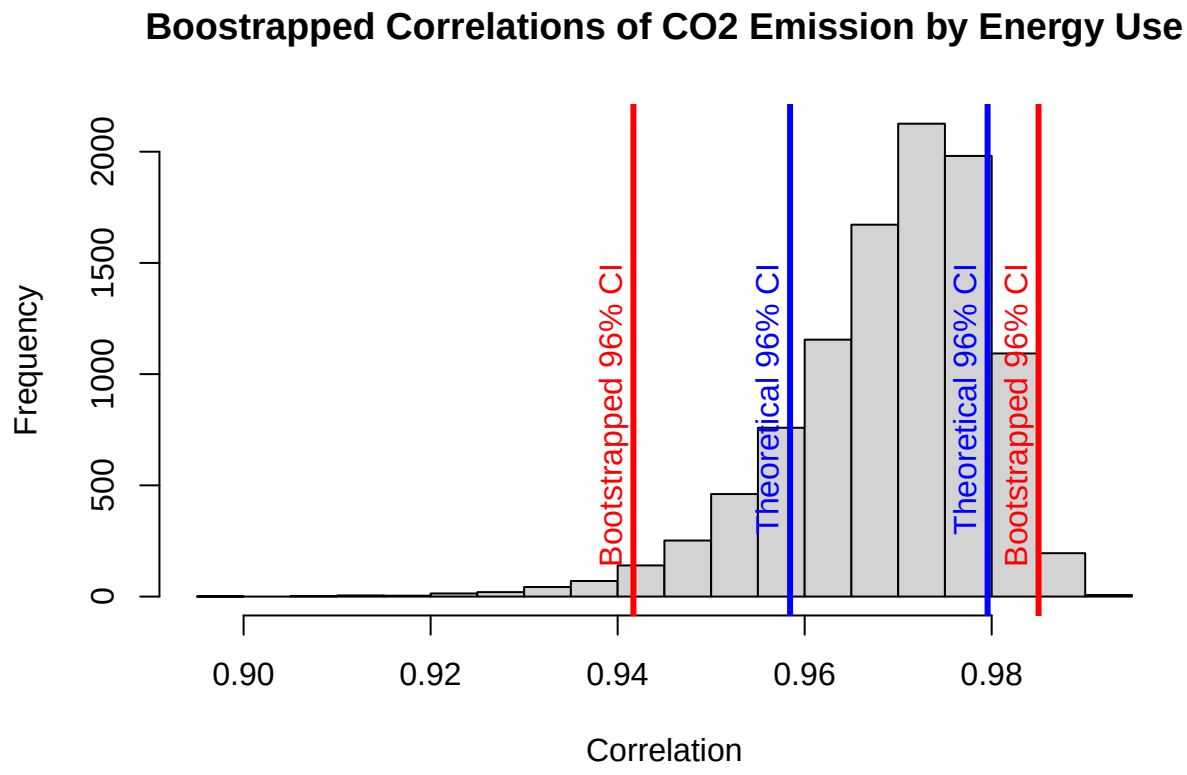
```
bresults <- c()
corResults <- c()
wb2_no_iceland <- wb2[wb2$Country != "Iceland",]
cor1test <- cor.test(wb2_no_iceland$CO2,
                     wb2_no_iceland$EnergyUse,
                     method = "pearson", conf.level=0.96)
for (i in 1:n_bootstrap) {
  bootstrap_sample <- wb2_no_iceland[sample(
    nrow(wb2_no_iceland), replace = TRUE),]
  CO2_sample <- bootstrap_sample$CO2
  Energy_sample <- bootstrap_sample$EnergyUse
  bresults[i] <- coef(lm(CO2_sample ~ Energy_sample))[2]
  corResults[i] <- cor(Energy_sample, CO2_sample)
}
ci_slope <- quantile(bresults, c(0.02, 0.98))
ci_r <- quantile(corResults, c(0.02, 0.98))
ci_slope
```

```
##          2%         98%
## 0.002291248 0.002675959
```

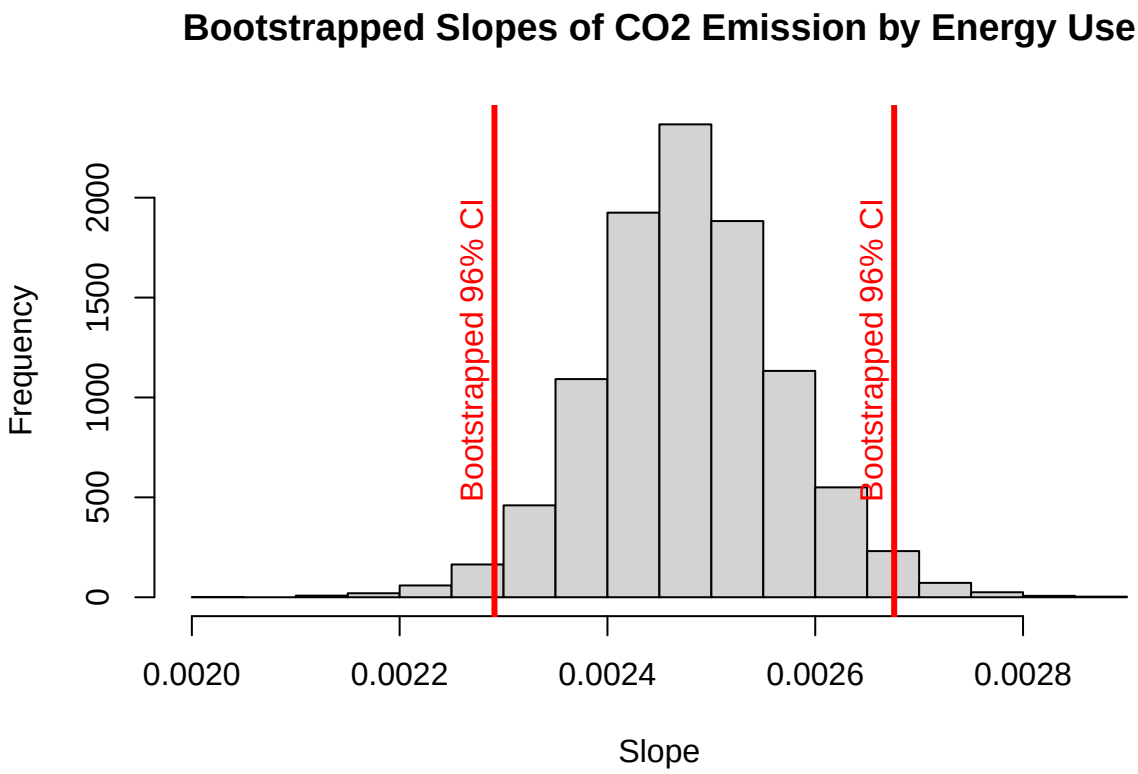
```
ci_r
```

```
##        2%       98%
## 0.9416892 0.9850133
```

```
hist(corResults,
     breaks=20,
     xlab="Correlation",
     main="Boostrapped Correlations of CO2 Emission by Energy Use")
abline(v=ci_r, col="red", lwd=3)
text(ci_r, 1500, "Bootstrapped 96% CI", pos = 2, col = "red", srt = 90)
abline(v=cor1test$conf.int, col="blue", lwd=3)
text(cor1test$conf.int, 1500, "Theoretical 96% CI", pos = 2, col = "blue", srt = 90)
```




```
hist(bresults,
     breaks=20,
     xlab="Slope",
     main="Bootstrapped Slopes of CO2 Emission by Energy Use")
abline(v=ci_slope, col="red", lwd=3)
text(ci_slope, 2000, "Bootstrapped 96% CI", pos = 2, col = "red", srt = 90)
```




```
# The bootstrapped correlations no longer have a bimodal distribution.
# However they are still left skewed, unlike the bootstrapped correlations
# of log CO2 by log Energy Use which has a approx. normal distribution.
# The theoretical CI is still much tighter than the bootstrapped CI. However,
# the bootstrapped slopes now look approximately normally distributed.
```

3) Class Survey Data *(32 pts).*

This questions involves cleaning and making bivariate plots of several variables from the class survey from 2022. I'm going to be less specific of what you need to do - I'll say that at this point you have all the tools you need to clean the data and make interesting plots.

Just run the code below to get the data and remind you of the variables that exist.

```
#Get data
survey <- read.csv("https://raw.githubusercontent.com/jreuning/sds230_data/refs/heads/main/class.survey
names(survey)
```

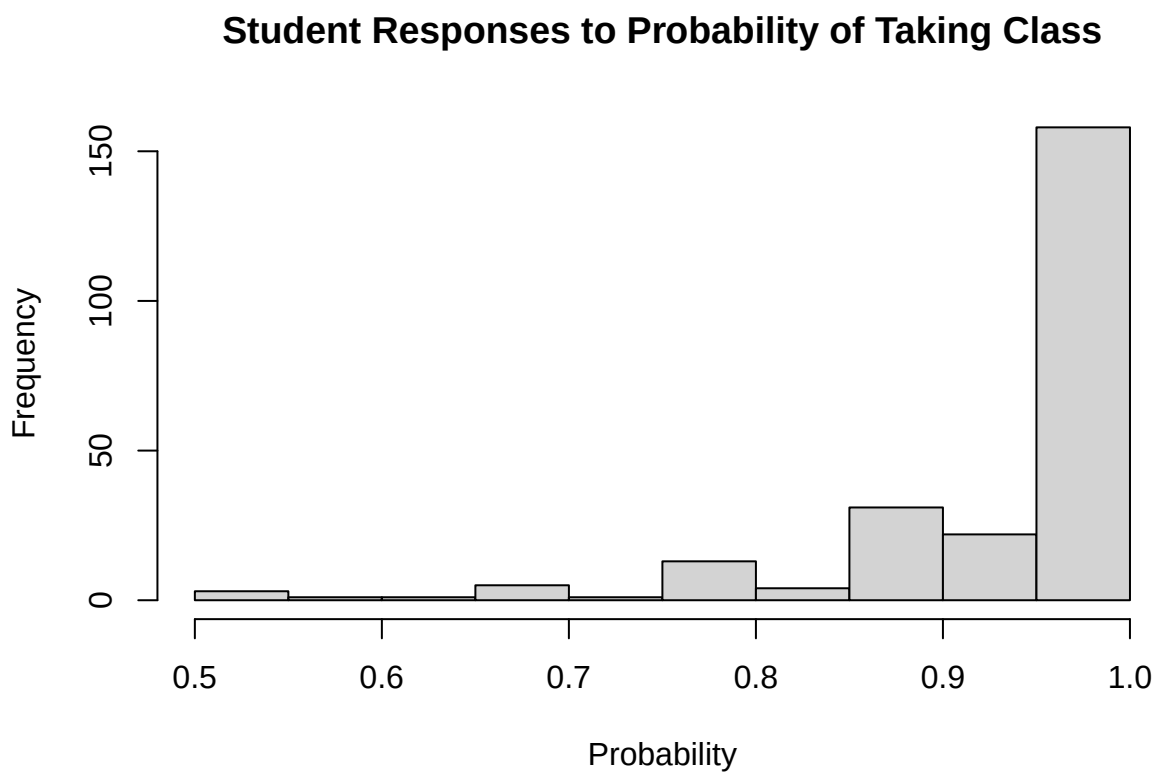
```
##  [1] "ClassProb"  "Status"     "Year"       "Division"   "Gender"
##  [6] "HtCm"       "Hand"       "Haircut"    "Exercise"   "Coursework"
## [11] "Web"        "TV"         "Social"     "Econ"       "Animal"
## [16] "Friends"    "Pulse"
```

```
dim(survey)
```

```
## [1] 239  17
```

(3.1) *(4 pts)* Clean the variable `ClassProb` so that all values are between zero and 1. Assume that any value over 20 is in a percentage and should be changed to a probability. Any value more than 1 but less than 20 should be changed to `NA`. Make a nice looking histogram of the resulting data and comment on the shape of the distribution.

```
survey$ClassProb[survey$ClassProb > 1 & survey$ClassProb < 20] <- NA
survey$ClassProb[survey$ClassProb >= 20] <- survey$ClassProb[survey$ClassProb >= 20] / 100
hist(survey$ClassProb,
     main="Student Responses to Probability of Taking Class",
     breaks=10,
     xlab="Probability")
```



The distribution is unimodal and left skewed. Most of the responses seem to be in the 0.95 - 1.00 range.

(3.2) *(4 pts)* Clean the variables `Social` and `Econ` so that they are integer and numeric on the scale 1 to 7. Make a two-way table of the results. Does it appear there is a relationship between these variables?

```
survey$Social <- as.numeric(gsub(".*([0-9]).*", "\\1", survey$Social))
```

```
## Warning: NAs introduced by coercion
```

```
survey$Econ <- as.numeric(gsub(".*([0-9]).*", "\\1", survey$Econ))
table(survey$Social, survey$Econ)
```

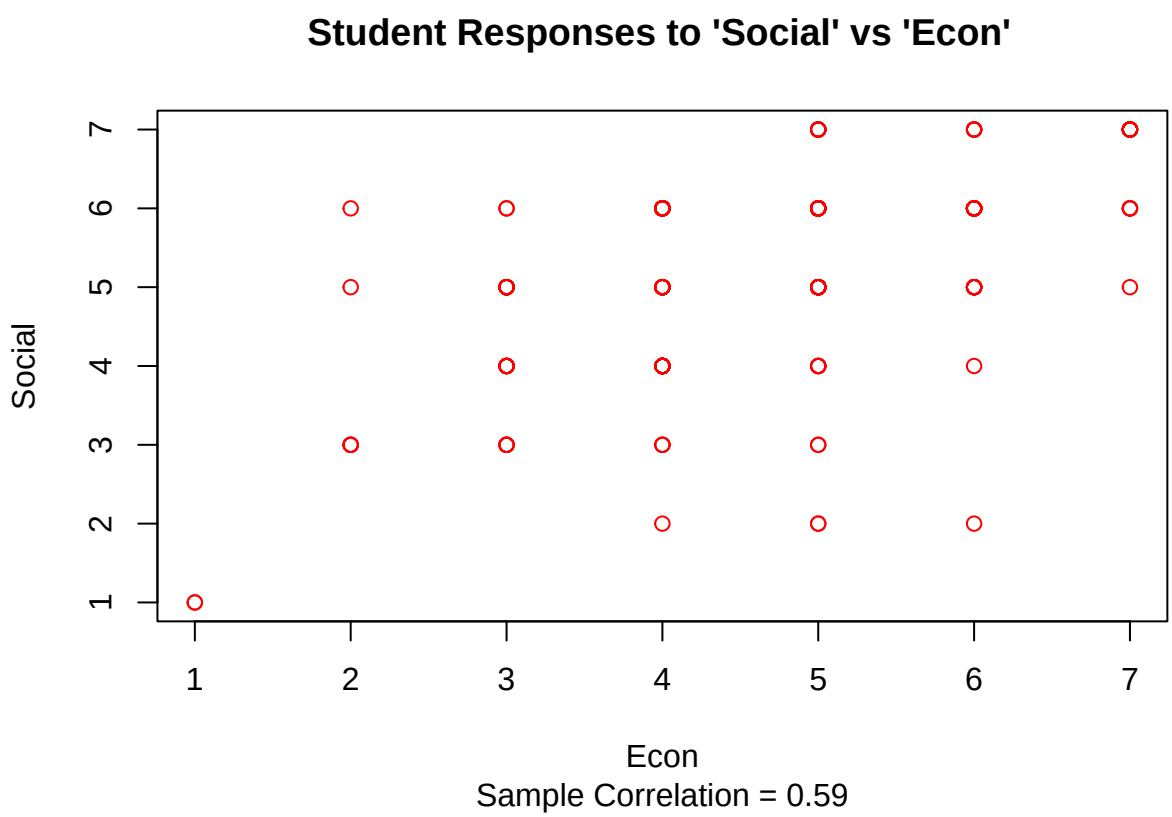
```
##
```

```
##      1  2  3  4  5  6  7
##  1   2  0  0  0  0  0  0
##  2   0  0  0  1  2  1  0
##  3   0  4  5  3  2  0  0
##  4   0  0  9 29  5  1  0
##  5   0  1 13 20 21 10  1
##  6   0  1  2 21 29 22  4
##  7   0  0  0  0  7  5 13
```

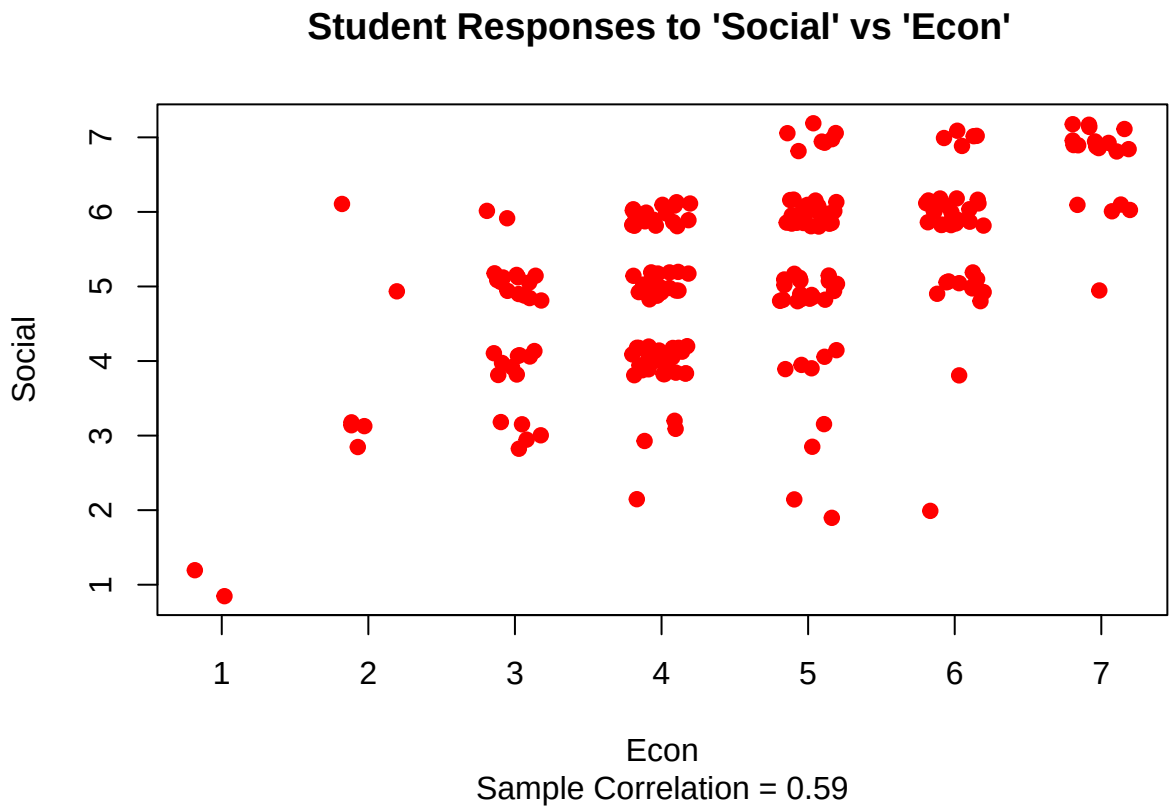
There does appear to be a positive relationship between these two variables; that is, the higher 'Social' is, the higher 'Econ' tends to be.

(3.3) *(12 pts)* Make three scatterplots of `Social` and `Econ` - one on the raw scale, one with jittered results, one with radii proportional to the square root of the frequency. Follow the examples shown in Class 11. Each plot should include the correlation in the labels.

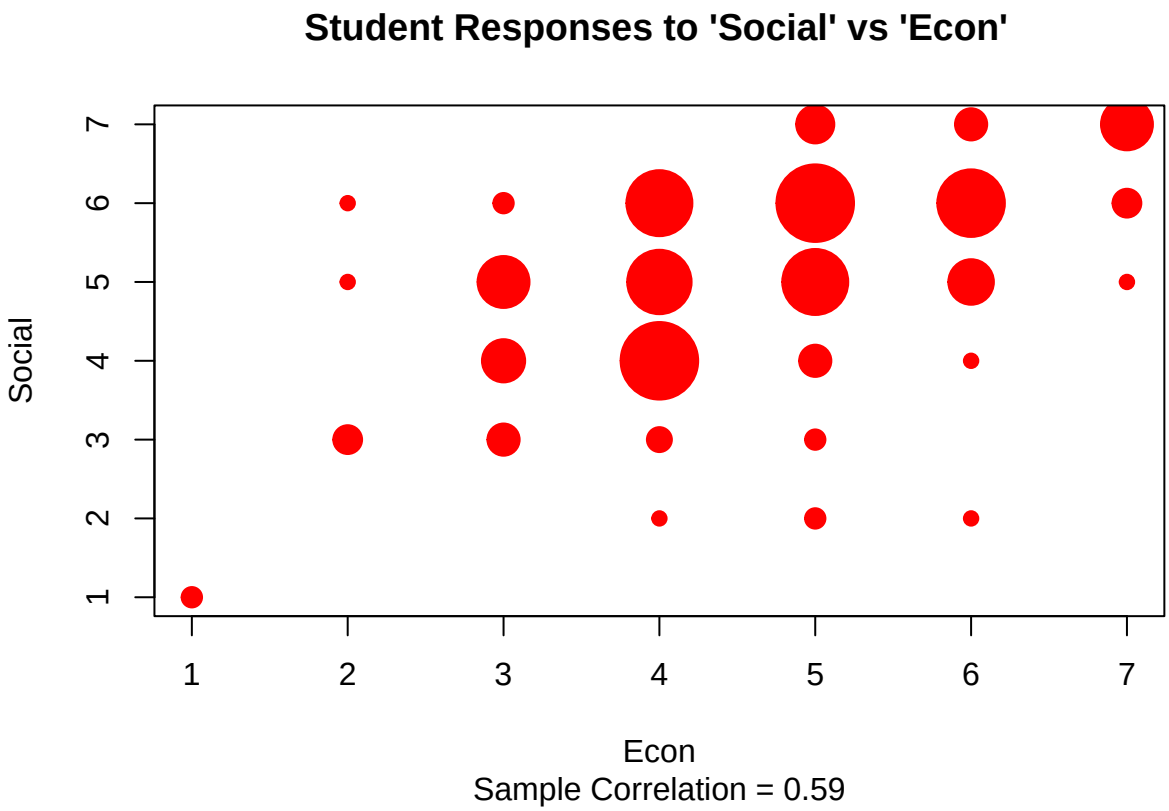
```
sample_corr <- round(cor(survey$Social, survey$Econ, use = "complete.obs"), 2)
plot(survey$Social ~ survey$Econ,
     col="red",
     main="Student Responses to 'Social' vs 'Econ'",
     xlab="Econ",
     sub=paste("Sample Correlation =", sample_corr),
     ylab="Social")
```



```r
plot(x=jitter(survey$Econ, factor=1),
     y=jitter(survey$Social, factor=1),
     col="red",
     main="Student Responses to 'Social' vs 'Econ'",
     xlab="Econ",
     sub=paste("Sample Correlation =", sample_corr),
     ylab="Social",
     pch=19)
```




```r
freq <- c(table(survey$Econ, survey$Social))
x1 <- rep(c(1:7), 7)
y1 <- sort(x1)
plot(x1,
     y1,
     col="red",
     main="Student Responses to 'Social' vs 'Econ'",
     xlab="Econ",
     ylab="Social",
     sub=paste("Sample Correlation =", sample_corr),
     cex = sqrt(freq),
     pch=19)
```

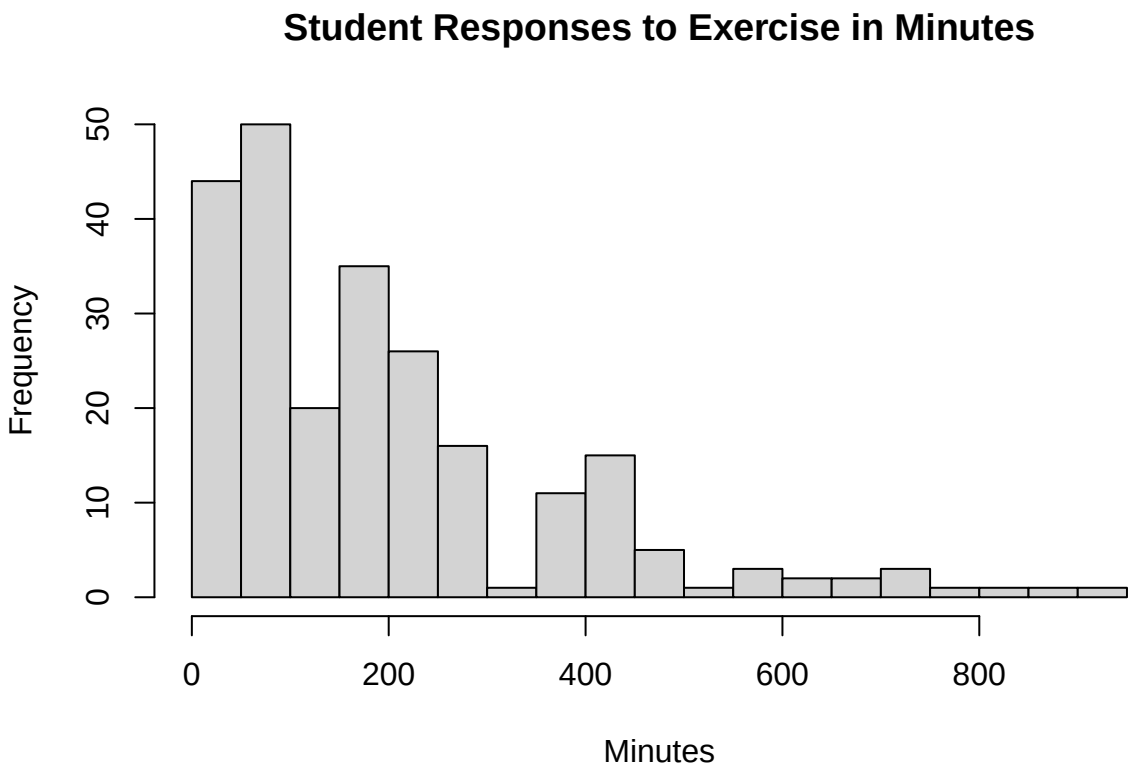


(3.4) *(5 pts)* Write a quick function called `cleanIt()` that cleans the variables `Exercise` and `Coursework` so that they are numeric. Apply this function to these two variables (which by the way are in minutes) and make histograms of the results. In a sentence, describe what you see.

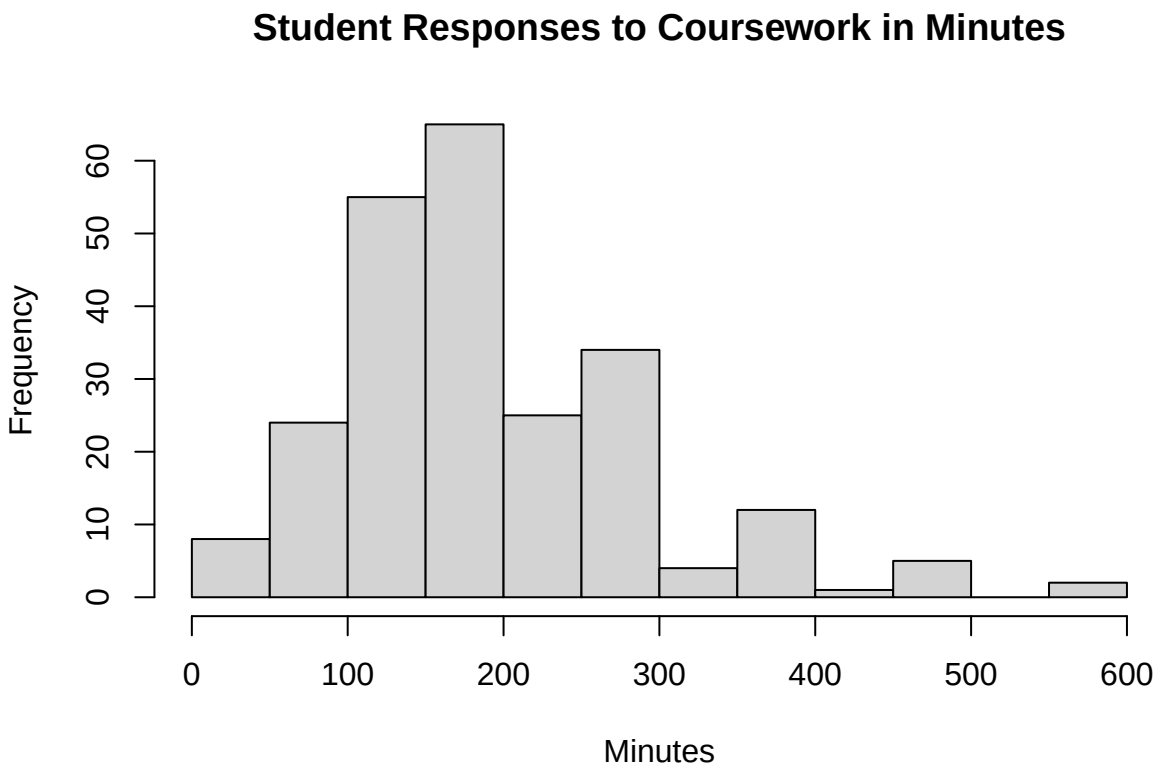
```
cleanIt <- function(x) {
  as.numeric(x)
}
survey$Exercise <- cleanIt(survey$Exercise)
survey$Coursework <- cleanIt(survey$Coursework)
```

```
## Warning in cleanIt(survey$Coursework): NAs introduced by coercion
```

```
hist(survey$Exercise,
     breaks=15,
     main="Student Responses to Exercise in Minutes",
     xlab="Minutes")
```




```
hist(survey$Coursework,
     breaks=15,
     main="Student Responses to Coursework in Minutes",
     xlab="Minutes")
```



Both of the distributions are right skewed, though the exercise distributions specifically appears exponential. For both distributions there is at least one individual who is a complete outlier on the right.

(3.5) *(7 pts)* Following code in classes 11 and 12, create plots `corrplot.mixed` and `chart.Correlation()` for all the variables mentioned in parts 3.1 through 3.4. Write not more than two sentences about what you observe and what you might do next. Note: ignore any warnings you might get, OR just include the cod 'warning = F' at the top of your chunk.

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.4.3
```

```
## corrplot 0.95 loaded
```

```
library(PerformanceAnalytics)
```

```
## Warning: package 'PerformanceAnalytics' was built under R version 4.4.3
```

```
## Loading required package: xts
```

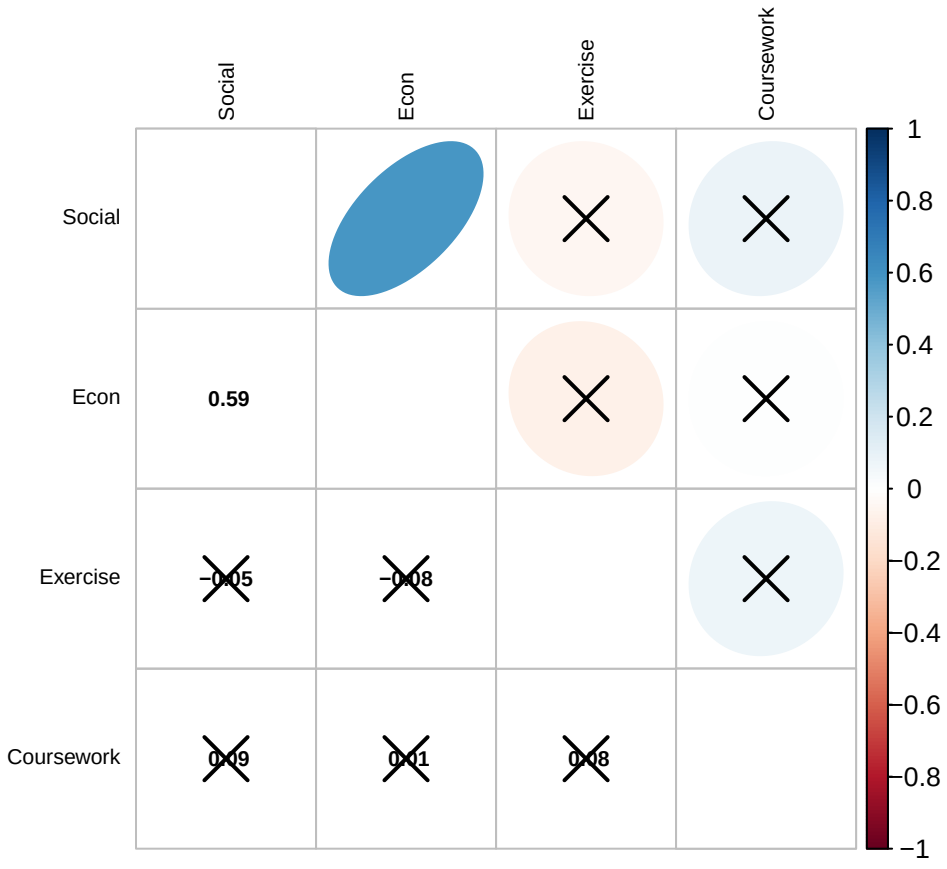
```
## Warning: package 'xts' was built under R version 4.4.3
```

```
## Loading required package: zoo
```

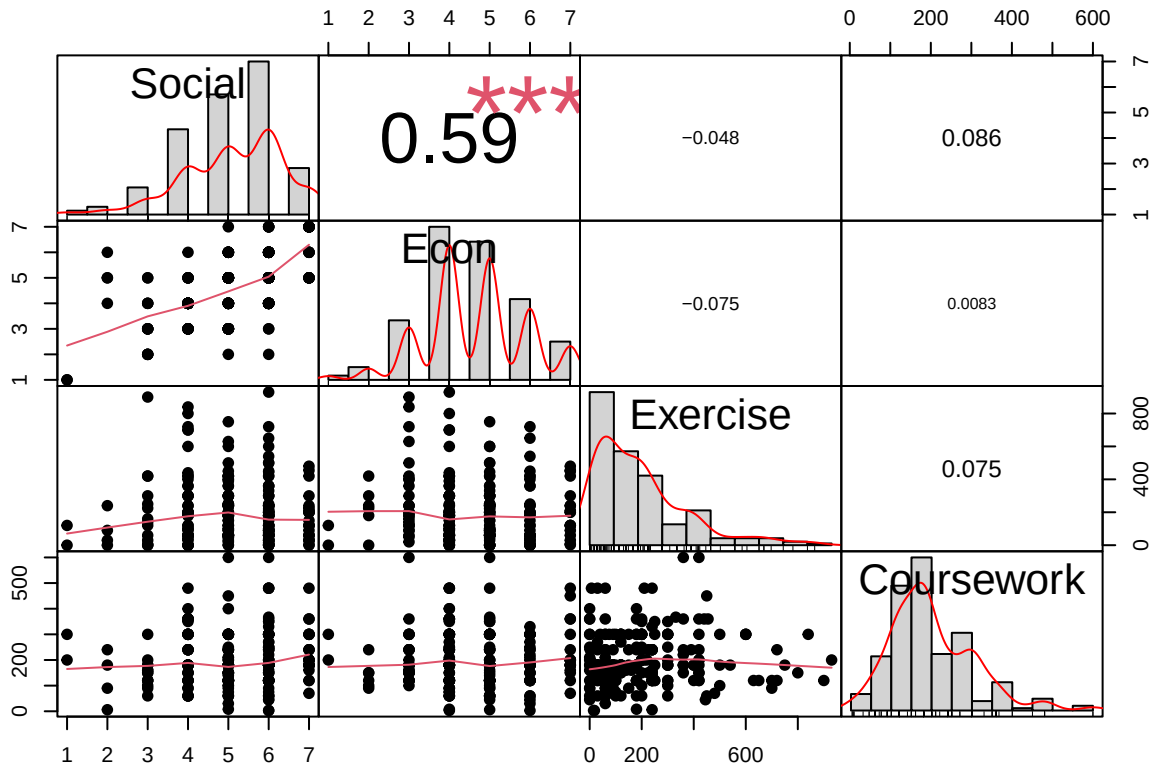
```
## Warning: package 'zoo' was built under R version 4.4.3
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
##
##     legend
```

```
survey2 <- survey[,c("ClassProb", "Social", "Econ", "Exercise", "Coursework")]
survey2 <- survey2[complete.cases(survey2), ]
sigcorr <- cor.mtest(survey2[, -1], conf.level = .95)
corrplot.mixed(cor(survey2[, -1]), lower.col = "black", upper = "ellipse",
               tl.col = "black", number.cex = .7, order = "hclust",
               tl.pos = "lt", tl.cex=.7, p.mat = sigcorr$p, sig.level = .05)
```




```
chart.Correlation(survey2[, -1], histogram = TRUE, pch = 19)
```



It seems that among there 4 variables, there exists some relationship between only one pair of them, between "Econ" and "Social" with a correlation of 0.59. From here, I would do some tests on this correlation, perhaps make some bootstrapped samples for the correlation to determine a confidence interval of the correlation or do a permutation test on the correlation to determine if the value of 0.59 is statistically significantly different from 0.