MA 541 Project

R Markdown

Part 1: Meet the data

Use any software to obtain the sample mean and sample standard deviation for each random variable (column) of the data; the sample correlations among each pair of the four random variables (columns) of the data.

```
df <- read_excel('MA 541 Course Project Data.xlsx')
head(df)
```

```
## # A tibble: 6 x 4
##   Close_ETF      oil     gold       JPM
##       <dbl>    <dbl>    <dbl>     <dbl>
## 1      97.3  0.0392   0.00467  0.0323
## 2      97.8  0.00195 -0.00137 -0.00295
## 3      99.2 -0.0315  -0.00794  0.0257
## 4      99.7  0.0346   0.0146   0.0118
## 5      99.3  0.0136  -0.0114   0.000855
## 6      98.2  0.00320 -0.00550 -0.0390
```

```
colMeans(df)
```

```
##    Close_ETF          oil         gold          JPM
## 1.211530e+02 1.030035e-03 6.628361e-04 5.304110e-04
```

We can see that Oil, Gold, and JPM have sample means of about 0, which is typically expected value for the return of stable stocks over a certain period of time. Close_ETF has a mean of about 121.

```
apply(df,2, FUN = sd)
```

```
##   Close_ETF         oil        gold         JPM
## 12.56979031  0.02109290  0.01128906  0.01101656
```

It is expected that the standard deviation of Close_ETF is a larger value than that of the other 3 due to the nature of the data for that column. Since this variable is actual prices over 1000 observations, we expect that the price would experience large changes, but since the other 3 are returns / price changes, those values are within a much smaller range so the standard deviation should be smaller. To exemplify, a stock price changing by $3 is sensible, but a return changing by 3 is not sensible due to the scale of its values.

```
cor(df)
```

```
##              Close_ETF          oil       gold         JPM
## Close_ETF  1.000000000 -0.009044842 0.02299557  0.03680706
## oil       -0.009044842  1.000000000 0.23565037 -0.12084893
## gold       0.022995570  0.235650372 1.00000000  0.10016984
```

```
## JPM         0.036807058 -0.120848930 0.10016984  1.00000000
```

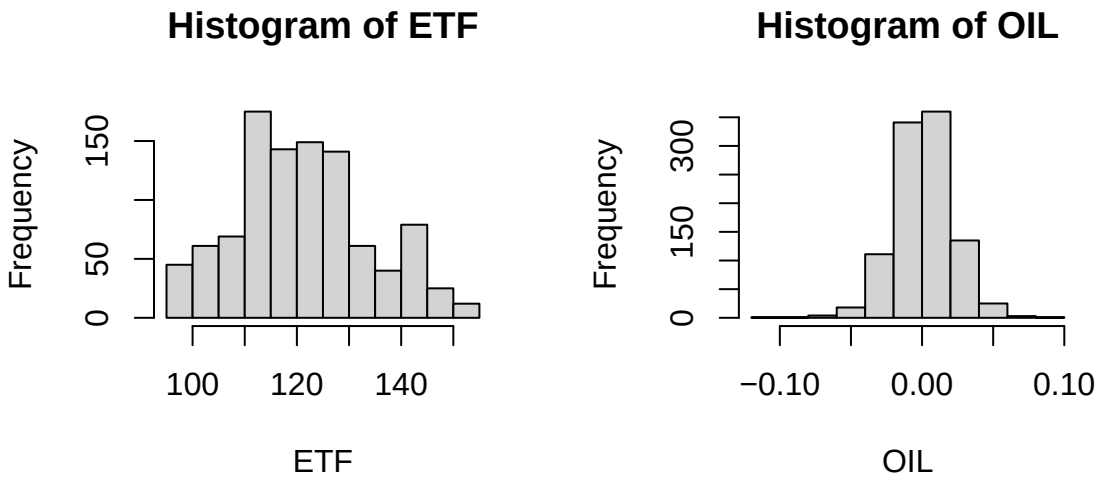
From this we can see that Oil and Gold have the strongest linear relationship with a correlation coefficient of 0.236, and since the correlation is positive, it means that as one increases, the other will increase as well. Only if it was a correlation of 1 (which each variable has with itself) would we see a perfectly linear relationship. The coefficient of 0.236 is still a relatively low coefficient, so these correlations show that none of the variables have strong linear relationships with each other.

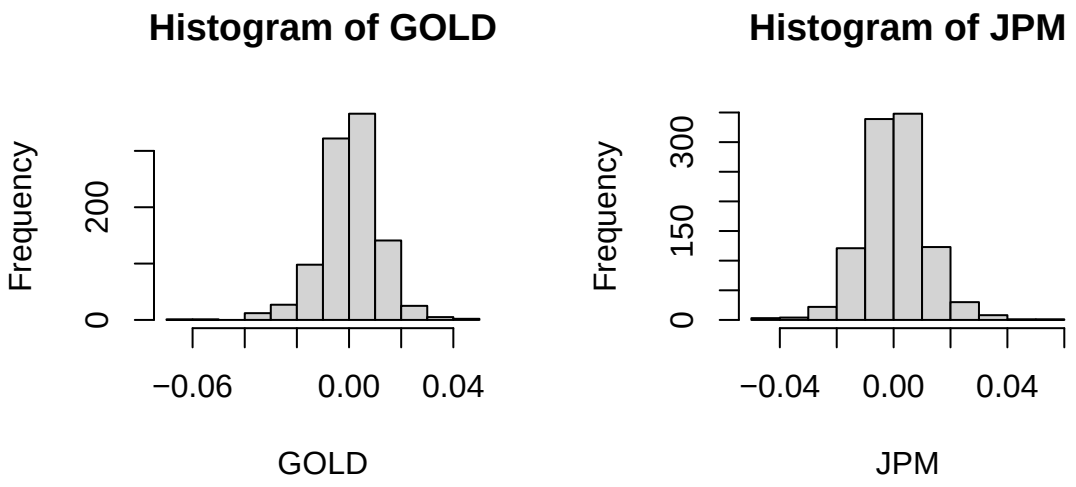
Part 2: Describe your data

1 Histogram for each column

```
ETF <- df$Close_ETF
OIL <- df$oil
GOLD <- df$gold
JPM <- df$JPM

hist(ETF)
hist(OIL)
hist(GOLD)
hist(JPM)
```

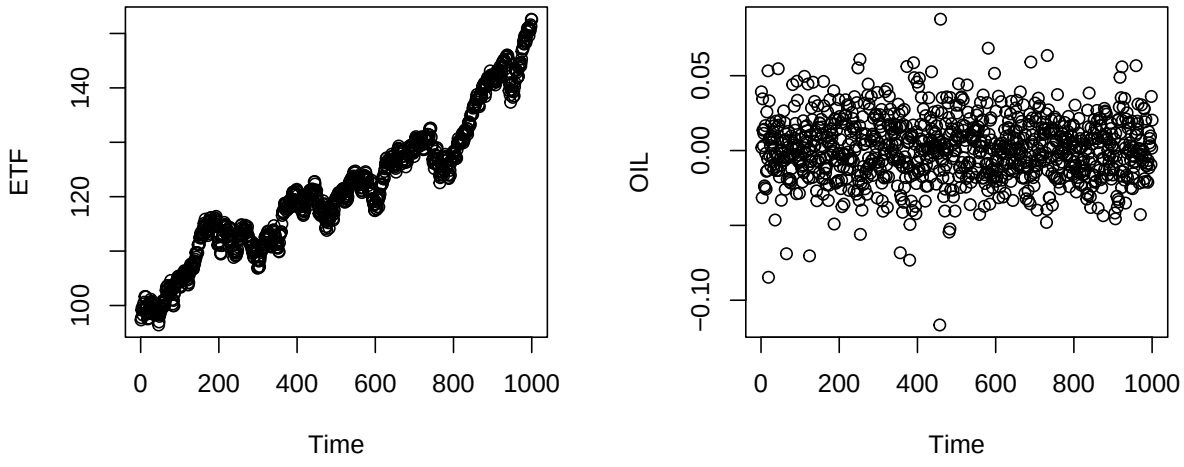


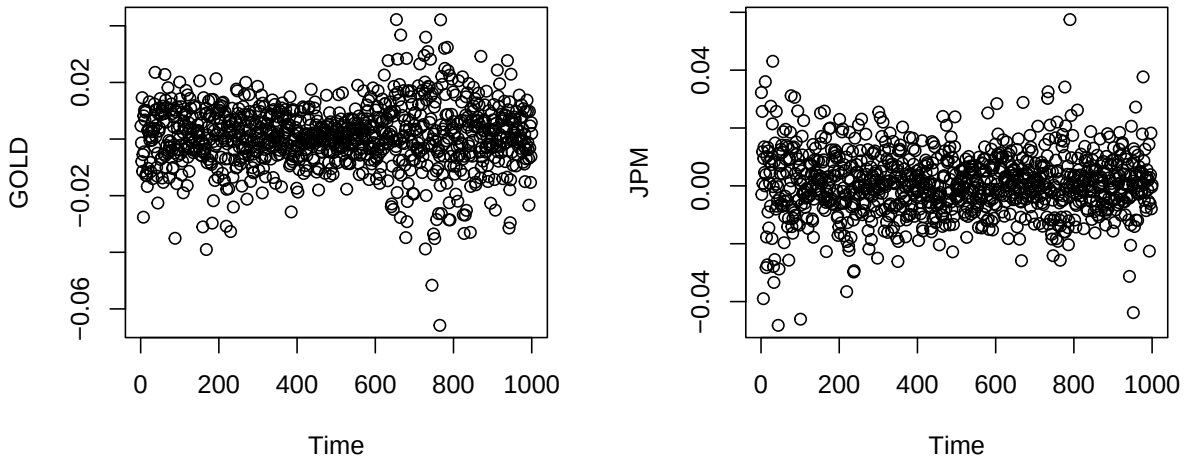


A further analysis into distributions will be completed in the next section, but these histograms depict that the random variables appear to be normally distributed as they are roughly symmetric about the mean, though the Close ETF may be slightly skewed left, and Gold may be slightly skewed right.

2 Time series plot for each column

```
Time <- seq(1,1000)
plot(Time, ETF)
plot(Time, OIL)
plot(Time, GOLD)
plot(Time, JPM)
```

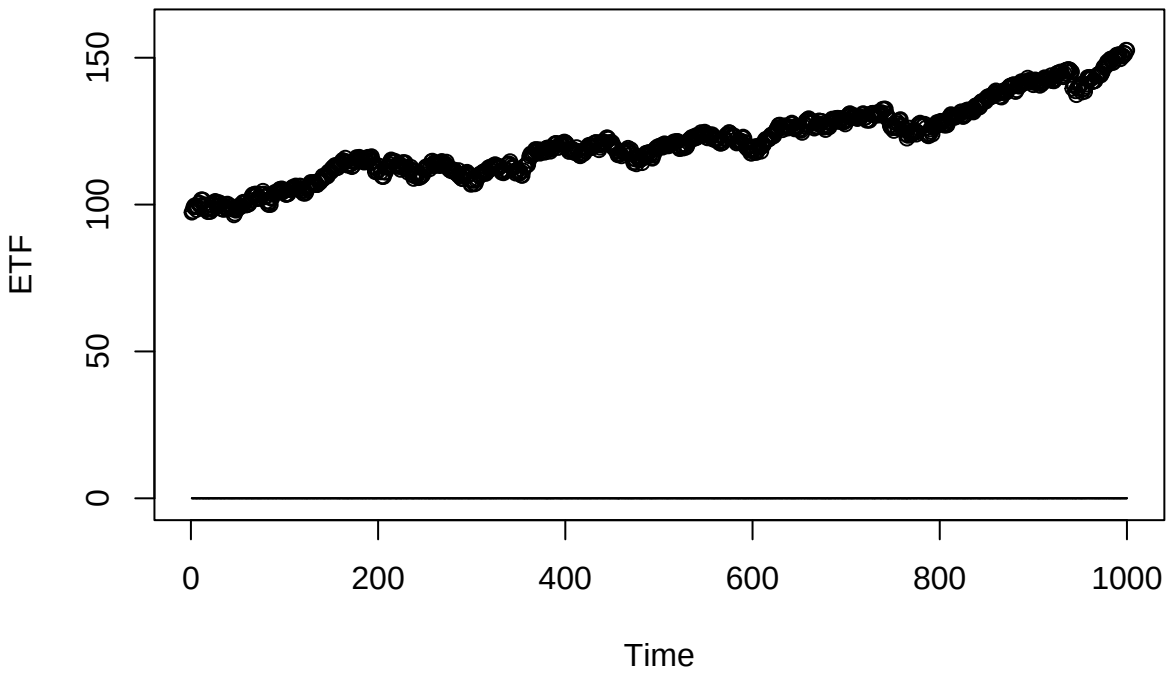




These time series plots further expemplify what was described before with the different types of data for the different columns. ETF Close values range from about 100 to 150, while the others are only on a scale of -0.1 to 0.06. These plots show expected results because we would expect to see some sort of trend in the ETF that shows how the price changes over time, in this case there is a general upward trend, meaning the price has been increasing. For the others, they are more evenly scattered around 0, in line with our sample mean findings. In the Gold graph, you can see a period of time related to high volatility because the values are more extreme at the negative and positive ends of the spectrum. This occurs around days 700 to 800.

3 Time series plot for all four columns

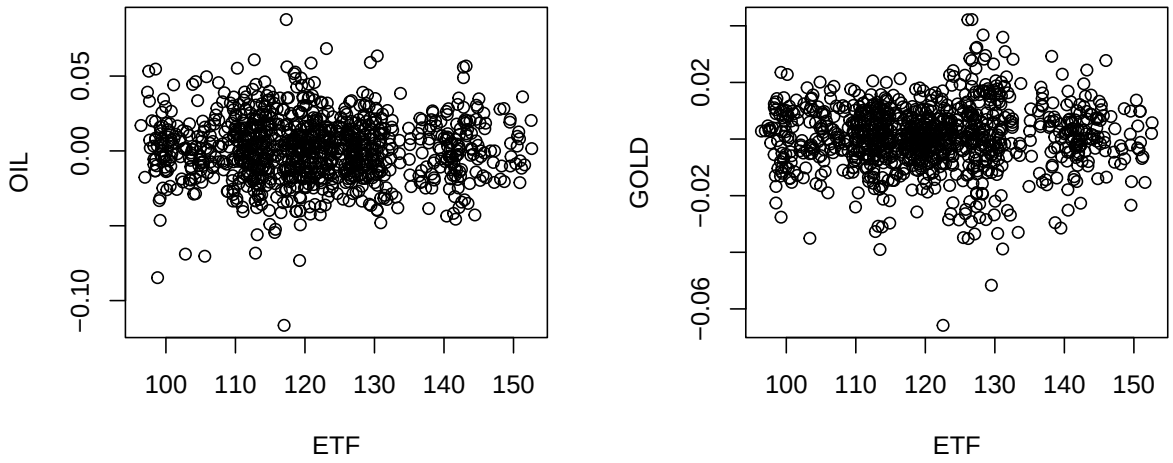
```
plot(Time, ETF, ylim = c(-1,160))
lines(Time, OIL)
lines(Time, GOLD)
lines(Time, JPM)
```

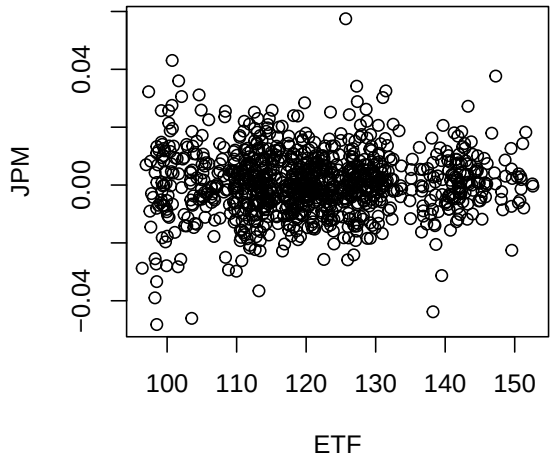


Since Oil, Gold, and JPM are on a different scale as variables as described previously, this graph is expected. All 3 of these variables are coinciding in the horizontal line at the 0 mark since they are all averaged at 0 with small deviations.

4 Three scatter plots to describe the relationships between the ETF column and the OIL column; between the ETF column and the GOLD column; between the ETF column and the JPM column, respectively

```
plot(ETF, OIL)
plot(ETF, GOLD)
plot(ETF, JPM)
```





These 3 scatterplots do not show any linear relationship between the ETF column and the other 3 columns. They show that no matter what the ETF value is, the 3 columns are still scattered around 0.

Part 3: What distribution does your data follow

Basing assumptions off of the histograms, it can be proposed that all of the variables follow normal distributions as explained by their shape previously. To test for this, we will use QQ plots to test for normality.

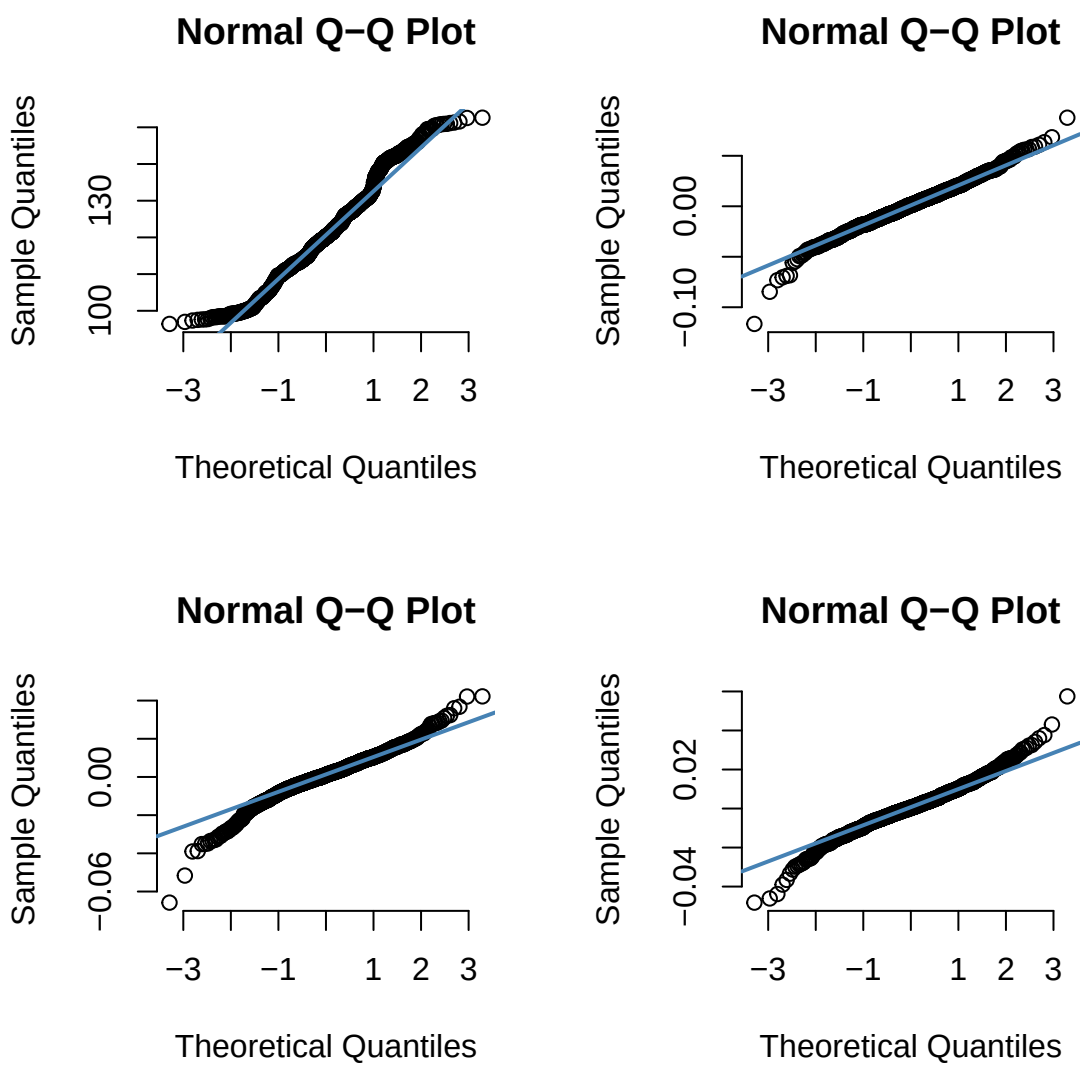
```
qqnorm(ETF, pch = 1, frame = FALSE)
qqline(ETF, col = "steelblue", lwd = 2)

qqnorm(OIL, pch = 1, frame = FALSE)
qqline(OIL, col = "steelblue", lwd = 2)

qqnorm(GOLD, pch = 1, frame = FALSE)
```

```
qqline(GOLD, col = "steelblue", lwd = 2)

qqnorm(JPM, pch = 1, frame = FALSE)
qqline(JPM, col = "steelblue", lwd = 2)
```



Each of the Q-Q Plots fall along a straight line, which indicates that each of the 4 variables are from normal distributions. This supports the initial assumption from looking at the histograms which also appeared to be normally distributed.

Part 4: Break your data into small groups and let them discuss the importance of the Central Limit Theorem

1 Calculate the mean ux and the standard deviation sx of the population

```
mu <- mean(ETF)
sd <- sd(ETF)
```

```
mu
```

```
## [1] 121.153
```

```
sd
```

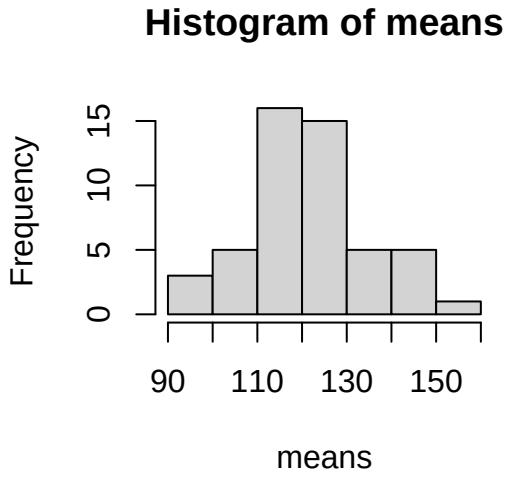
```
## [1] 12.56979
```

2 Break the population into 50 groups sequentially and each group includes 20 values

```
means <- c()
sds <- c()
for(i in seq(1,1000,20)){
  index <- seq(i, 19+i)
  mean <- mean(ETF[index])
  means <- c(means, mean)
  samp_sd <- sd(ETF[index])
  sds <- c(sds,samp_sd)
}
```

3 Calculate the sample mean x of each group. Draw a histogram of all the sample means. Comment on the distribution of these sample means, i.e., use the histogram to assess the normality of the data consisting of these sample means

```
hist(means)
```



Seeing that the histogram is bell-shaped and symmetric about the mean, without fat tails, we can assume that these sample means are normally distributed.

4 Calculate the mean ux and the standard deviation sx of the data including these sample means. Make a comparison between ux and ux, between sx/n and sx. Here, n isthe number of sample means calculated from Item 3) above

```
u_xbar <- mean(means)
u_xbar
```

```
## [1] 121.153
```

```
mu
```

```
## [1] 121.153
```

```r
sd / sqrt(50)
```

```
## [1] 1.777637
```

```r
mean(sds)
```

```
## [1] 1.34401
```

We see that the means are equal, and the standard deviations are relatively similar.

5 Are the results from Items 3) and 4) consistent with the Central Limit Theorem? Why?

The results from 3 and 4 are consistent with the CLT because we see that these samples show consistent results with the population, in terms of the mean, variance, and distribution, which is shown to be normal through the histogram.

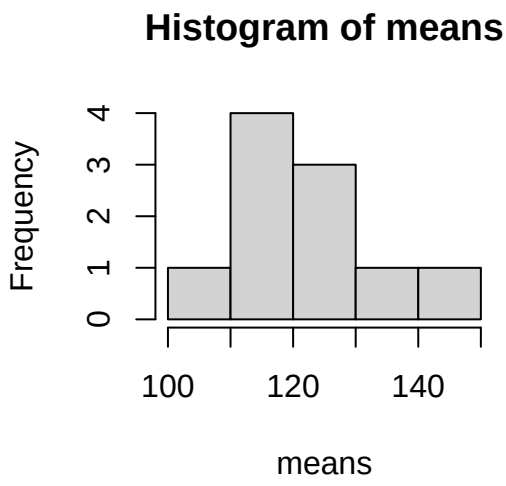
6 Break the population into 10 groups sequentially and each group includes 100 values

```r
means <- c()
sds <- c()

for(i in seq(1,1000,100)){
  index <- seq(i, 99+i)
  mean <- mean(ETF[index])
  means <- c(means, mean)
  samp_sd <- sd(ETF[index])
  sds <- c(sds,samp_sd)
}
```

7 Repeat Items 3) ~ 5).

```r
hist(means)
```



Though we have a small number of samples, the histogram appears to be normally distributed as the values are symmetric about the mean and do not have fat tails.

```r
u_xbar <- mean(means)
u_xbar
```

```
## [1] 121.153
```

```
mu
```

```
## [1] 121.153
```

```
sd / sqrt(50)
```

```
## [1] 1.777637
```

```
mean(sds)
```

```
## [1] 2.994802
```

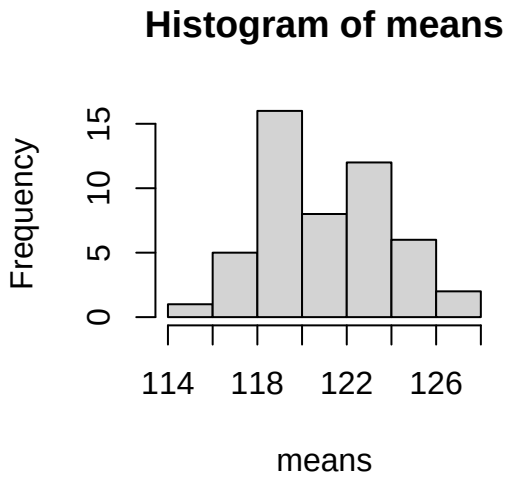
Similarly to the previous result, we see that these results are again consistent with the population, so it is consistent with CLT. We do observe that the standard deviation of the samples is greater than the population standard deviation, which is expected since we have a smaller amount of samples, so there will be more variability among these 10 samples.

8 Generate 50 simple random samples or groups (with replacement) from the population. The size of each sample is 20, i.e., each group includes 20 values

```
means <- c()
sds <- c()
for(i in seq(50)){
  samp <- sample(ETF, 20, replace = TRUE)
  mean <- mean(samp)
  means <- c(means, mean)
  samp_sd <- sd(samp)
  sds <- c(sds,samp_sd)
}
sample2 <- samp
```

9 Repeat Items 3) ~ 5)

```
hist(means)
```



This histogram appears to be normally distributed, possibly with a slightly left skew, but is still relatively symmetric about the mean.

```
u_xbar <- mean(means)
u_xbar
```

```
## [1] 120.9762
```

```
mu
```

```
## [1] 121.153
```

```
sd / sqrt(50)
```

```
## [1] 1.777637
```

```
mean(sds)
```

```
## [1] 12.42309
```

These results are consistent with CLT since we see that the mean of the samples approximates the population mean, and the histogram appears to be normally distributed. Again we see a difference arise in the standard deviation, which is expected since we are using replacement so we will not be using all unique values, and we are only looking at a sample of the population.

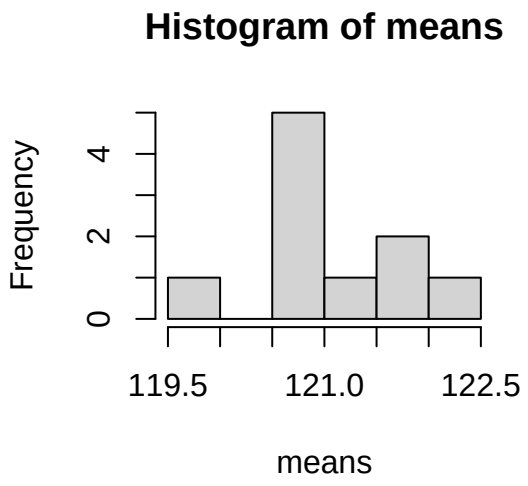
10 Generate 10 simple random samples or groups (with replacement) from the population. The size of each sample is 100, i.e., each group includes 100 values

```
means <- c()
sds <- c()
for(i in seq(10)){
  samp <- sample(ETF, 100, replace = TRUE)
  mean <- mean(samp)
  means <- c(means, mean)
  samp_sd <- sd(samp)
  sds <- c(sds,samp_sd)
}

sample1 <- samp
```

11 Repeat Items 3) ~ 5)

```
hist(means)
```



Though this histogram has gaps due to the sample number of samples, we can still see the general trend of these samples being normally distributed.

```r
u_xbar <- mean(means)
u_xbar
```

```
## [1] 121.0388
```

```r
mu
```

```
## [1] 121.153
```

```r
sd / sqrt(50)
```

```
## [1] 1.777637
```

```r
mean(sds)
```

```
## [1] 12.54534
```

Similarly to the previous section, we see that the sample mean approximates the population mean, and again we see an increase in the standard deviation due to the sampling technique. This is consistent with CLT.

12 In Part 3 of the project, you have figured out the distribution of the population (the entire ETF column). Does this information have any impact on the distribution of the sample mean(s)? Explain your answer.

Since the samples were all taken from a normally distributed population, it is expected that the sample means are also normally distributed. So, by the results that were seen above, we can see that our sample means are in line with what would have been expected.

Part 5: Construct a confidence interval with your data

1 Pick up one of the 10 simple random samples you generated in Step 10) of Part 4, construct an appropriate 95% confidence interval of the mean u

```r
#install.packages("BSDA")
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```r
test1 <- t.test(sample1,conf.level = 0.95)
test1$conf.int
```

```
## [1] 119.5345 124.6589
## attr(,"conf.level")
## [1] 0.95
```

2 Pick up one of the 50 simple random samples you generated in Step 8) of Part 4, construct an appropriate 95% confidence interval of the mean u.

```r
test2 <- z.test(sample2, sigma.x = sd, conf.level = 0.95)
test2$conf.int
```

```
## [1] 118.4486 129.4664
## attr(,"conf.level")
## [1] 0.95
```

In Part 1, you have calculated the mean u of the population (the entire ETF column) using Excel function. Do the two intervals from 1) and 2) above include (the true value of) the mean u? Which one is more accurate? Why?

These two confidence intervals both include the mean mu. Since we have a larger sample size for the first sample, we can conclude that this will be more accurate. As you increase your sample size, you will also be decreasing your standard deviation. This is seen in the confidence interval since it is a smaller range of values centered around the population mean mu, while the second sample of smaller sample size has a larger range despite the same confidence level of 95%.

Part 6: Form a hypothesis and test it with your data

1 Use the same sample you picked up in Step 1) of Part 5 to test H0: u=100 vs. Ha: u != 100 at the significance level 0.05. What's your conclusion?

```
mean(sample1)
```

```
## [1] 122.0967
```

```
t.test(sample1, mu = 100, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  sample1
## t = 17.112, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 100
## 95 percent confidence interval:
##  119.5345 124.6589
## sample estimates:
## mean of x
##  122.0967
```

Since the p-value is less than 0.05, we reject the null hypothesis. This means that the mean of this sample is not equal to 100.

2 Use the same sample you picked up in Step 2) of Part 5 to test H0: u=100 vs. Ha: u != 100 at the significance level 0.05. What's your conclusion?

```
mean(sample2)
```

```
## [1] 123.9575
```

```
t.test(sample2, mu = 100, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  sample2
## t = 8.7061, df = 19, p-value = 4.667e-08
## alternative hypothesis: true mean is not equal to 100
## 95 percent confidence interval:
##  118.1979 129.7171
## sample estimates:
## mean of x
##  123.9575
```

Since the p-value is less than 0.05, we reject the null hypothesis. This means that the mean of this sample is not equal to 100.

3 Use the same sample you picked up in Step 2) of Part 5 to test H0: s = 15 vs. Ha: s != 15 at the significance level 0.05. What's your conclusion?

```
#install.packages("EnvStats")
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
## The following object is masked from 'package:base':
##
##     print.default
```

```
sd(sample2)
```

```
## [1] 12.30652
```

```
varTest(sample2, alternative="two.sided", sigma.squared = 225)
```

```
##
##  Chi-Squared Test on Variance
##
## data:  sample2
## Chi-Squared = 12.789, df = 19, p-value = 0.3016
## alternative hypothesis: true variance is not equal to 225
## 95 percent confidence interval:
##   87.59066 323.08446
## sample estimates:
## variance
## 151.4504
```

Since the p-value is greater than 0.05, we fail to reject the null hypothesis. This means that the true value of the standard deviation is 15.

4 Use the same sample you picked up in Step 2) of Part 5 to test H0: s = 15 vs. Ha: s < 15 at the significance level 0.05. What's your conclusion?

```
varTest(sample2, alternative="less", sigma.squared = 225)
```

```
##
##  Chi-Squared Test on Variance
##
## data:  sample2
## Chi-Squared = 12.789, df = 19, p-value = 0.1508
## alternative hypothesis: true variance is less than 225
## 95 percent confidence interval:
##    0.0000 284.4275
## sample estimates:
## variance
## 151.4504
```

Since the p-value is greater than 0.05, we fail to reject the null hypothesis. This means that the true value of the standard deviation is 15. This is interesting since we see that our actual standard deviation for the

sample is less than 15, but due to the small sample size our results are not as conclusive.

Part 7: Compare your data with a different data set

1 Consider the entire Gold column as a random sample from the first population, and the entire Oil column as a random sample from the second population. Assuming these two samples be drawn independently, form a hypothesis and test it to see if the Gold and Oil have equal means in the significance level 0.05.

```
t.test(OIL, GOLD)
```

```
##
##  Welch Two Sample t-test
##
## data:  OIL and GOLD
## t = 0.48537, df = 1527.9, p-value = 0.6275
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001116768  0.001851167
## sample estimates:
##    mean of x    mean of y
## 0.0010300355 0.0006628361
```

The p-value is greater than 0.05, so we fail to reject the null hypothesis. Therefore, we can conclude that Gold and Oil have equal means. This is also concluded because the 95% confidence interval includes 0.

2 Subtract the entire Gold column from the entire Oil column and generate a sample of differences. Consider this sample as a random sample from the target population of differences between Gold and Oil. Form a hypothesis and test it to see if the Gold and Oil have equal means in the significance level 0.05.

```
diff <- OIL-GOLD
t.test(diff, mu=0)
```

```
##
##  One Sample t-test
##
## data:  diff
## t = 0.54133, df = 999, p-value = 0.5884
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.0009639099  0.0016983087
## sample estimates:
##    mean of x
## 0.0003671994
```

Since the p-value is greater than 0.05, our hypothesis test on sample of differences shows that the true mean equals 0, again showing that Gold and Oil have equal means. This is also concluded because the 95% confidence interval includes 0.

3 Consider the entire Gold column as a random sample from the first population, and the entire Oil column as a random sample from the second population. Assuming these two samples be drawn independently, form a hypothesis and test it to see if the Gold and Oil have equal standard deviations in the significance level 0.05.

```
var.test(GOLD, OIL)
```

```
##
```

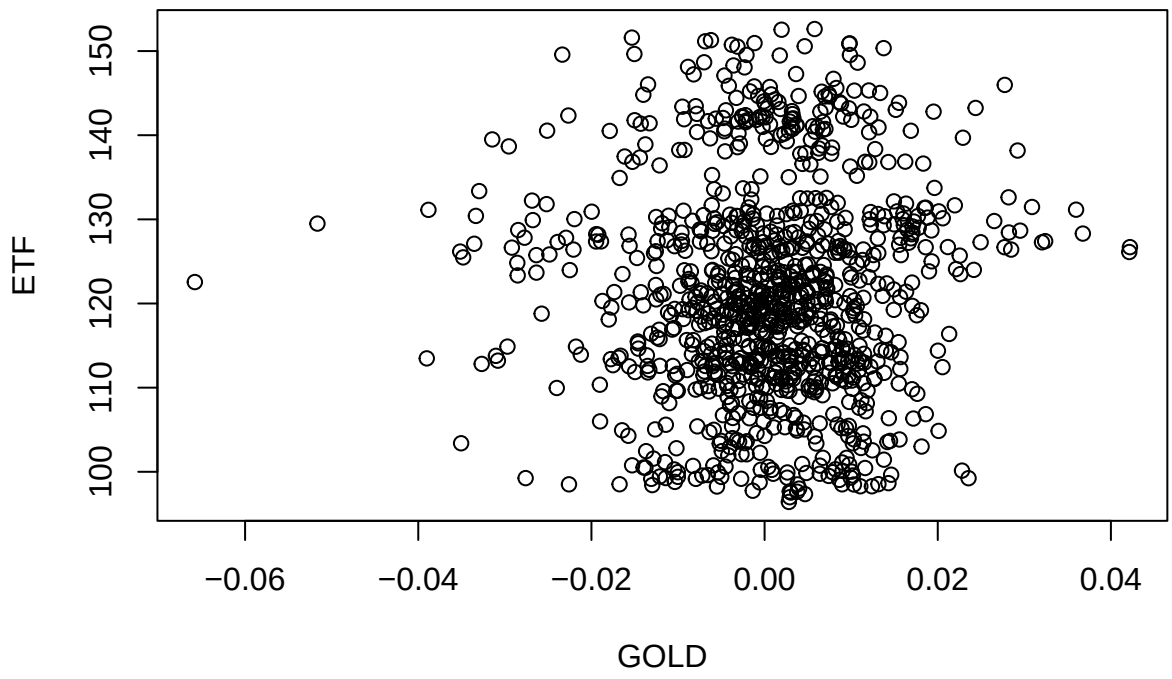
```
##  F test to compare two variances
##
## data:  GOLD and OIL
## F = 0.28645, num df = 999, denom df = 999, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.2530176 0.3242914
## sample estimates:
## ratio of variances
##          0.2864462
```

Since the p-value is less than 0.05, we reject the null hypothesis. Here the alternative hypothesis states "true ratio of variances is not equal to 1", which means that they do not have equal variances, and also not equal standard deviation.

Part 8: Fitting the line to the data

1 Draw a scatter plot of ETF (Y) vs. Gold (X). Is there any linear relationship between them which can be observed from the scatter plot?

```
plot(GOLD,ETF)
```



From this scatter plot, we can see that the values of ETF are randomly scattered around the different Gold values, so no linear relationship is observed.

2 Calculate the coefficient of correlation between ETF and Gold and interpret it

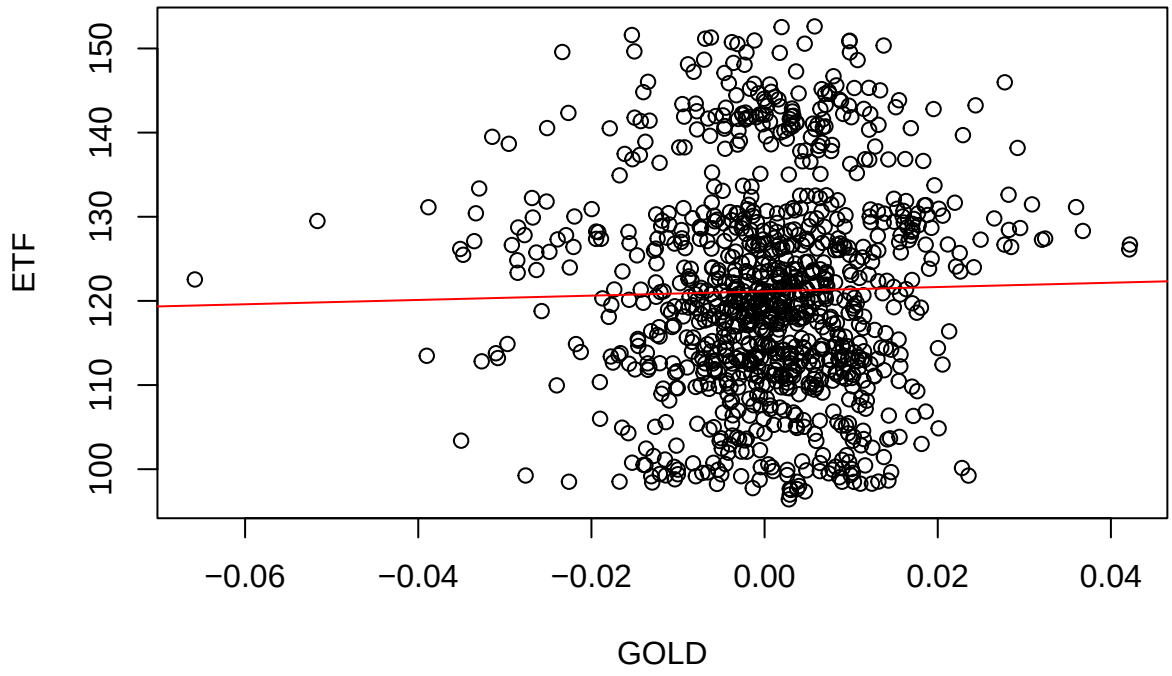
```r
cor(GOLD,ETF)
```

```
## [1] 0.02299557
```

The coefficient of correlation ranges from -1 to 1. A value of -1 means that the variables are perfectly linearly related with a negative relationship, a value of 1 means the variables are perfectly linearly related with a positive relationship, and a value close to 0 means there is no linear relationship. In support of the conclusion from the scatterplot, our coefficient of correlation is close to 0, so there is no linear relationship between ETF and Gold.

3 Fit a regression line (or least squares line, best fitting line) to the scatter plot. What are the intercept and slope of this line? How to interpret them?

```r
plot(GOLD,ETF)
abline(lm(ETF~GOLD), col='red')
```




```r
coef(lm(ETF~GOLD))
```

```
## (Intercept)        GOLD
##   121.13599    25.60439
```

The intercept of the line is 121.13599, which means that when Gold has a value of 0, then the ETF value is 121.14. This value is very close to the mean of the ETF variable. The slope of the line is 25.60439, which means that with an increment of 1 in the Gold variable, on average the ETF variable will increase by 25.60.

4 Conduct a two-tailed t-test with H0: B1 = 0. What is the P-value of the test? Is the linear relationship between ETF (Y) and Gold (X) significant at the significance level 0.01? Why or why not?

```r
summary(lm(ETF~GOLD))
```

```
##
## Call:
## lm(formula = ETF ~ GOLD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.7878  -8.4621  -0.9893   7.5376  31.3537
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 121.1360     0.3983 304.155   <2e-16 ***
## GOLD         25.6044    35.2363   0.727    0.468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.57 on 998 degrees of freedom
## Multiple R-squared:  0.0005288,  Adjusted R-squared:  -0.0004727
## F-statistic: 0.528 on 1 and 998 DF,  p-value: 0.4676
```

We can see that the p-value for B1 (our slope, or coefficient for GOLD), is 0.468. Since this is greater than 0.05, we fail to reject the null hypothesis. This implies that B1 = 0. With B1 = 0, our model simply becomes ETF = Intercept, or ETF = 121.136, roughly the mean of the column. This means that there is no linear relationship between ETF and GOLD since any change in GOLD would not cause a linear change in ETF. If B1 was not equal to 0, that would imply that there is a linear relationship between the two.

5 Suppose that you use the coefficient of determination to assess the quality of this fitting. Is it a good model? Why or why not?

```r
summary(lm(ETF~GOLD))
```

```
##
## Call:
## lm(formula = ETF ~ GOLD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.7878  -8.4621  -0.9893   7.5376  31.3537
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 121.1360     0.3983 304.155   <2e-16 ***
## GOLD         25.6044    35.2363   0.727    0.468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.57 on 998 degrees of freedom
## Multiple R-squared:  0.0005288,  Adjusted R-squared:  -0.0004727
## F-statistic: 0.528 on 1 and 998 DF,  p-value: 0.4676
```

Using the output above, we can see that the R2 value is 0.0005288. This is a very very low value for R2, which means that our model is not a good fit, since the R2 is used to evaluate how well the explanatory variable explains the response variable.

6 What are the assumptions you made for this model fitting?

For this model fitting, there were 4 main assumptions that were made. If this were a multiple linear regression model, we would need to assume no collinearity between independent variables, but in this case we only have one indepdent variable of Gold.

First, it was assumed that ETF and Gold have a linear relationship. Second, it was assumed that both ETF and Gold are normally distributed, which we have shown previously. Third, there is little to no autocorrelation in the data, meaning that the residuals are assumed to be indepedent of each other. Lastly, it was assumed that the data is homoscedastic.

7 Given the daily relative change in the gold price is 0.005127. Calculate the 99% confidence interval of the mean daily ETF return, and the 99% prediction interval of the individual daily ETF return

```
x <- 0.005127
df_gold <-data.frame(GOLD=x)

predict(lm(ETF~GOLD), df_gold,level=0.99,interval='confidence')
```

```
##        fit      lwr      upr
## 1 121.2673 120.1638 122.3707
```

```
predict(lm(ETF~GOLD), df_gold,level=0.99,interval='predict')
```

```
##        fit      lwr      upr
## 1 121.2673 88.80117 153.7334
```

As seen by the fit, this daily relative change in the gold price does not have much impact on the price of the ETF since the value for the ETF is still very close to its mean, which is also very close to the intercept of the model. Our confidence interval shows that the lower and upper limits for the mean daily ETF value are very close to the actual fit. The prediction interval shows a much larger range, meaning that 99% of the times when the relative change in gold price is 0.005127, the close ETF price will be between 88.80117 and 153.7334. It is expected that this interval will be larger than the confidence interval because it is more difficult to predict what the value of ETF price will be just at an individual data point.

Part 9: Does your model predict?

Consider the data including the ETF, Gold and Oil column. Using any software, fit a multiple linear regression model to the data with the ETF variable as the response. Evaluate your model with adjusted R2.

```
mod <- lm(Close_ETF ~ gold + oil, data=df)
summary(mod)
```

```
##
## Call:
## lm(formula = Close_ETF ~ gold + oil, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.6509  -8.5418  -0.9938   7.5909  31.5246
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 121.1427     0.3987 303.856   <2e-16 ***
## gold         29.6226    36.2715   0.817    0.414
## oil          -9.1261    19.4128  -0.470    0.638
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.58 on 997 degrees of freedom
## Multiple R-squared:  0.0007503,  Adjusted R-squared:  -0.001254
## F-statistic: 0.3743 on 2 and 997 DF,  p-value: 0.6879
```

The Adjusted R-squared value is -0.001254. This means that the model is a poor fit and the variables of Gold and Oil are not good explanatory variables for the ETF variable.

Part 10: Checking residuals and model selection

Calculate the residuals of the model fitting you did in Part 9. Check the four assumptions made for the error terms of the multiple regression model using these residuals (mean 0; constant variance; normality; and the independence). You may draw some plots over the residuals to check these assumptions. For example, draw a Normal Probability Plot to check the normality assumption; draw a scatter plot of Residuals vs. Fitted Values to check the constant variance assumption and the independence assumption; and so on. You may refer to the following link https://www.youtube.com/watch?v=4zQkJw73U6I for some hints. In your project report, all the relevant plots and at least one paragraph of summary of checking the four assumptions using those plots must be included. Discuss how you may improve the quality of your regression model according to the strategy of model selection.

```
residuals <- summary(mod)$resid
```

Mean 0 Assumption

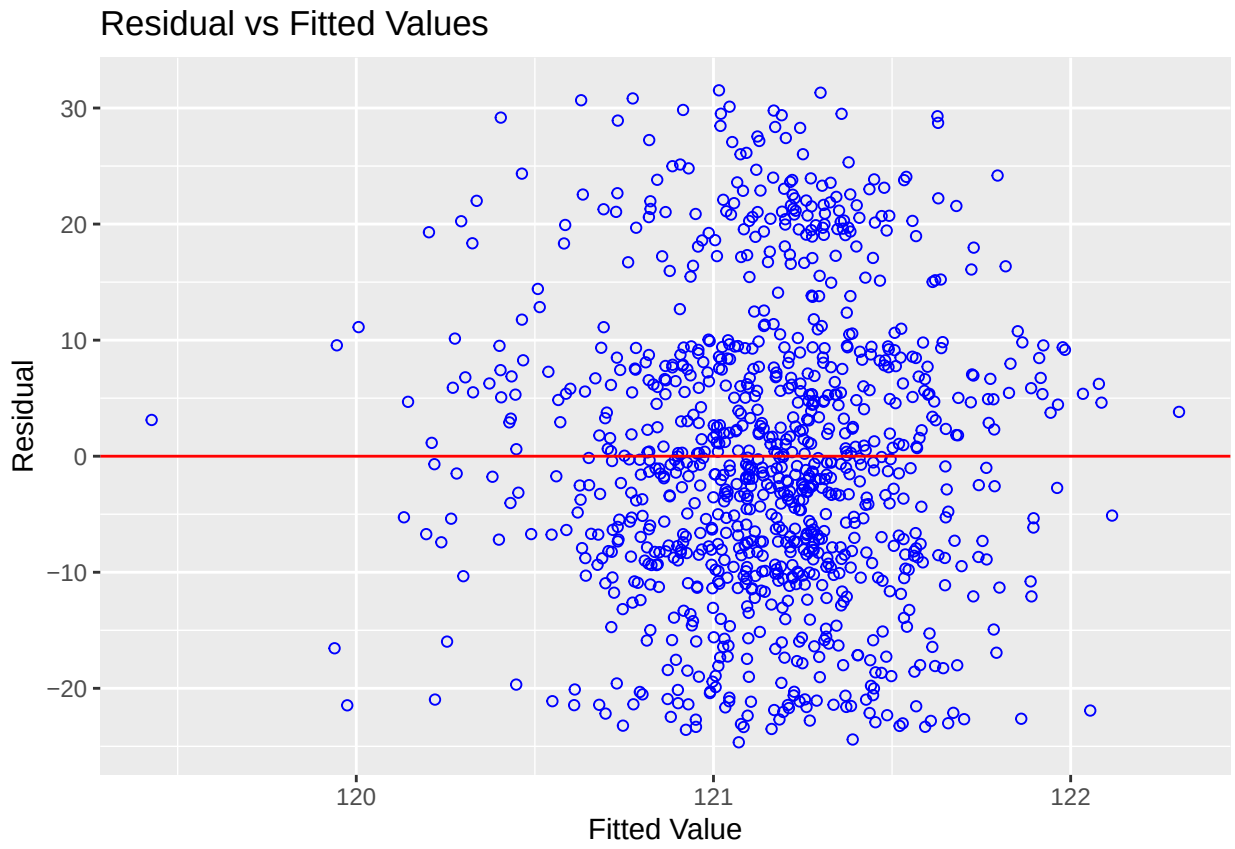
Contstant Variance and Independence Assumptions

```
#install.packages("olsrr")
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

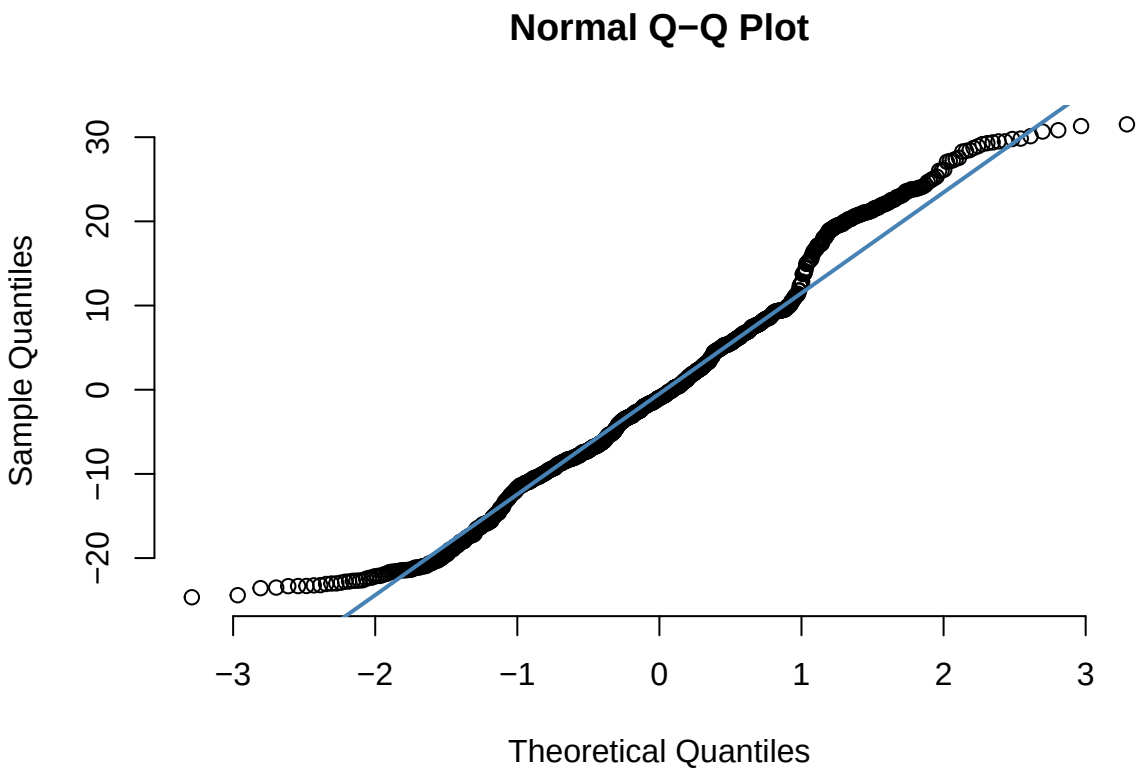
```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_plot_resid_fit(mod)
```



This plot of residuals vs fitted values is very useful to check three of our assumptions: the assumption of mean 0, constant variance, and independence. In order to check for mean 0, we want this plot to show the red line as a horizontal line at 0. We see this in our graph, so it is assumed that our residuals have a mean of 0. We also use this plot to check for constant variance. This is becasue these fitted values are a linear combination of our independent variables, but their impact is removed from this plot. If we saw a cone shape in this graph, that would mean there is not constant variance, so this assumption is met. Finally, we check the independence assumption. This assumption would be violated if there was a clear relationship between the residuals and the fitted values. We do not see this in our plot since the values appeared to be scattered rather randomly, so the independence assumption is met.

```
qqnorm(residuals, pch = 1, frame = FALSE)
qqline(residuals, col = "steelblue", lwd = 2)
```



We can test for normality of residuals using a Q-Q Plot. As described previously, if our Q-Q plot depicts that the values are along a straight line, then we can assume that our data is normally distributed. Therefore, the residuals of this model meet the assumption of normality.

Discuss how you may improve the quality of your regression model according to the strategy of model selection.

We have explored a few different strategies of model selection, including subset selection and stepwise approaches. We could use this to improve the quality of the model, though it likely would not have a large impact since we are dealing with a small number of independent variables. In addition, we have explored the relationships between the variables and have seen that there are not strong linear relationships.