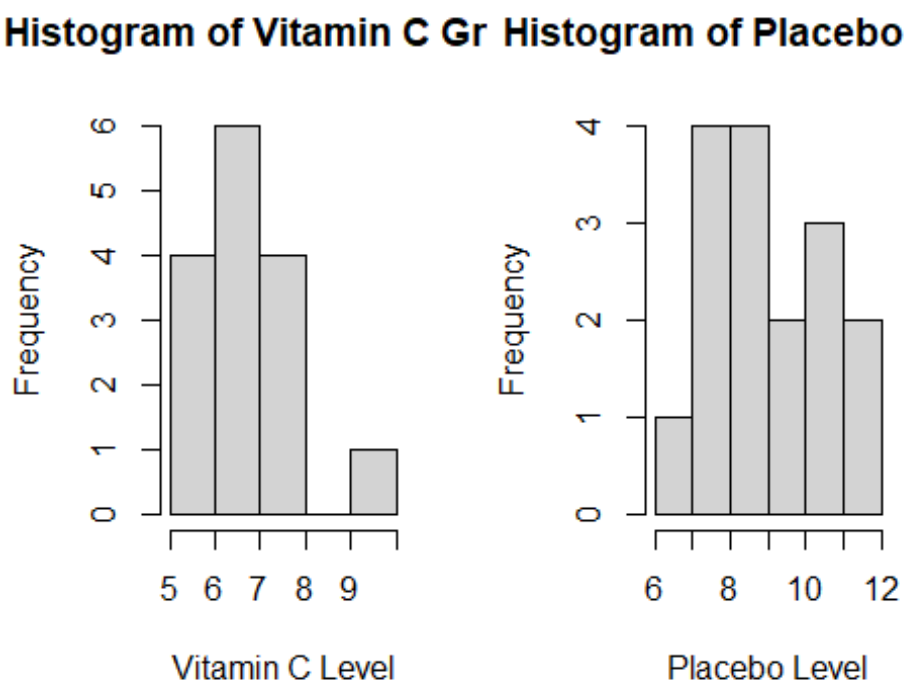
Data 602 - Assignment Four

```
library(MASS)
library(mosaic)

## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order
to add
## additional features.  The original behavior of these functions should not
be affected by this.

##
## Attaching package: 'mosaic'

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following object is masked from 'package:Matrix':
##
##     mean

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum

library(ggplot2)
library(dplyr)
library(resampledata)

##
## Attaching package: 'resampledata'

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

```r
#Question One

#(a) Do these data indicate that the recovery time is quicker with Vitamin C
than without? Carry out the appropriate statistical investigation with a
permutation test. Ensure you show "where" your permutation test statistic
lies on the distribution. Report your (empirical) $P$-value and your
statistical inference. For the number of permutation tests, use 2999.

#Null Hypothesis (H0): No difference in recovery time between the group
receiving Vitamin C and the group receiving a placebo.
#Alternative Hypothesis (Ha): There is a difference in recovery time between
the group receiving the vitamin C and the group receiving the placebo.
Recovery time is quicker with vitamin C than without.
set.seed(123)
vitamin_c <- c(6, 7, 7, 7, 8, 7, 7, 8, 7, 8, 10, 6, 8, 5, 6)
placebo <- c(10, 12, 8, 6, 9, 8, 11, 9, 11, 8, 12, 11, 9, 8, 10, 9)


observed_diff <- mean(vitamin_c) - mean(placebo)


combined_data <- c(vitamin_c, placebo)


num_permutations <- 2999


permuted_diffs <- numeric(num_permutations)

# Permutation
for (i in 1:num_permutations) {
  # Shuffle the combined data
  shuffled_data <- sample(combined_data)


  permuted_vitamin_c <- shuffled_data[1:length(vitamin_c)]
  permuted_placebo <- shuffled_data[(length(vitamin_c) +
1):length(shuffled_data)]
  permuted_diffs[i] <- mean(permuted_vitamin_c) - mean(permuted_placebo)
}

# p-value
empirical_p_value <- mean(abs(permuted_diffs) >= abs(observed_diff))


observed_position <- sum(abs(permuted_diffs) >= abs(observed_diff))
```

```r
# Output results
cat("Observed difference in mean recovery time:", observed_diff, "\n")

## Observed difference in mean recovery time: -2.304167

cat("Empirical p-value:", empirical_p_value, "\n")

## Empirical p-value: 0.000666889

cat("Observed test statistic position in permutation distribution:",
observed_position, "\n")

## Observed test statistic position in permutation distribution: 2

#permutation test p-value of 0.000666889, which is smaller than both the
significance levels of 0.025 (for a two-tailed test) and 0.05 (for a one-
tailed test), it indicates strong evidence against the null hypothesis.
Therefore, we reject the null hypothesis that there is no difference in
recovery time between the Vitamin C and placebo groups. Instead, we conclude
that there is compelling evidence to suggest that participants who received
Vitamin C recovered faster from the common cold compared to those who
received the placebo.

#(b) Re-test your statistical hypothesis in part (a) using the $t$-test. In
doing so, state any assumptions about these data or conditions you are
imposing on these data and conduct the necessary diagnostics to either
confirm or refute such assumptions. Ensure you provide both the $P$-value and
its interpretation related to these data.

#The main assumptions for the t-test include: Normality: The data within each
group should be approximately normally distributed. Homogeneity of variances:
The variances of the two groups should be approximately equal.Independence:
Observations within and between groups are independent of each other.

#Check for normality assumption

# Shapiro-Wilk test for normality
shapiro_test_vitamin_c <- shapiro.test(vitamin_c)
shapiro_test_placebo <- shapiro.test(placebo)

# Print Shapiro-Wilk test results
cat("Shapiro-Wilk Test for Vitamin C Group:\n", "p-value =",
shapiro_test_vitamin_c$p.value, "\n\n")

## Shapiro-Wilk Test for Vitamin C Group:
##  p-value = 0.1534285

cat("Shapiro-Wilk Test for Placebo Group:\n", "p-value =",
shapiro_test_placebo$p.value, "\n\n")

## Shapiro-Wilk Test for Placebo Group:
##  p-value = 0.3611147
```

```r
# Histograms
par(mfrow = c(1, 2))
hist(vitamin_c, main = "Histogram of Vitamin C Group", xlab = "Vitamin C
Level")
hist(placebo, main = "Histogram of Placebo Group", xlab = "Placebo Level")
```




```r
par(mfrow = c(1, 1))

#p-value = 0.1534
#p-value = 0.3611

#Since the p-values are greater than the chosen significance level (e.g.,
0.05), we can conclude that there is no strong evidence against the normality
assumption.

t_test_result <- t.test(vitamin_c, placebo)
print(t_test_result)

##
##  Welch Two Sample t-test
##
## data:  vitamin_c and placebo
## t = -4.445, df = 27.08, p-value = 0.0001345
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.367628 -1.240706
## sample estimates:
```

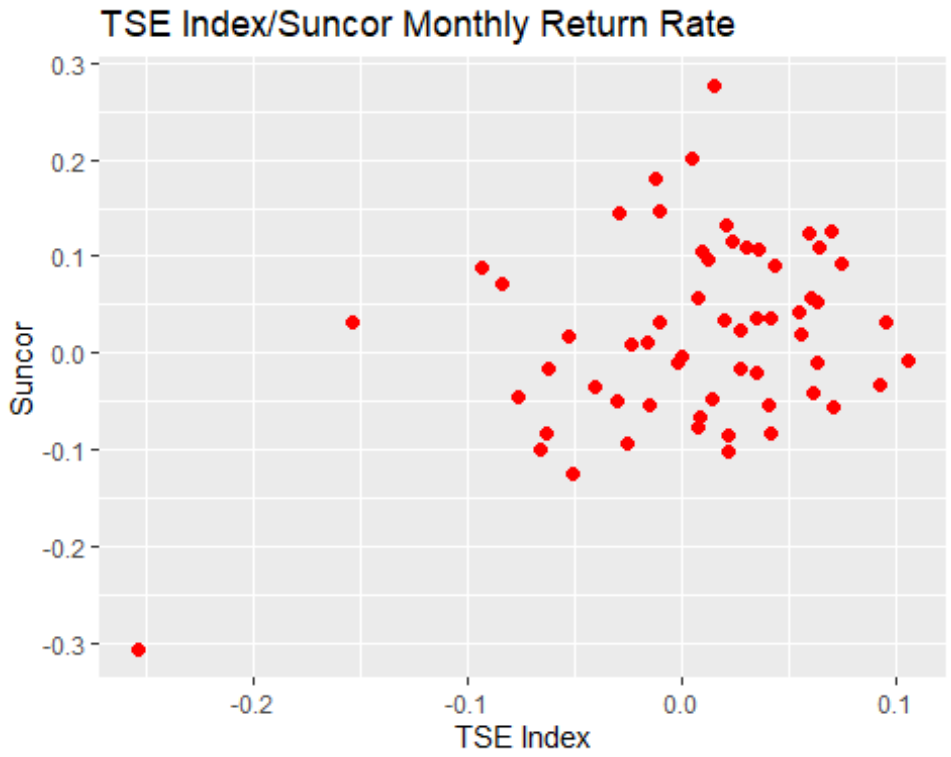
```
## mean of x mean of y
##  7.133333  9.437500

#Since the p-value (0.0001345) is less than the chosen significance level
(0.05) for a two-sided test, and 0.025 for a one-sided test, we reject the
null hypothesis. Therefore, there is sufficient evidence to suggest that the
mean vitamin C level differs significantly from the mean placebo level.

#Question Three

#(a) Pertaining to Question 9: Do these data suggest there is a treatment
effect? Test the existence of a treatment effect using a permutation test. If
a treatment effect is discovered, explain its meaning in the context of these
data. Carry out a permutation test, using 1999 iterations.</br>
#(You will have to carefully consider "what a treatment effect" is in this
scenario.)
data =
read.csv("https://raw.githubusercontent.com/Statman44/Data602/main/chocnochoc
ratings.csv")
head(data)

##   Particpant Class Group GroupName Q9 Overall
## 1          1     1     1 Chocolate  3    2.44
## 2          2     1     1 Chocolate  3    2.56
## 3          3     1     1 Chocolate  5    4.33
## 4          4     1     1 Chocolate  4    4.56
## 5          5     1     1 Chocolate  5    4.56
## 6          6     1     1 Chocolate  5    4.22

#Hypothesis Test:

#H0:u=0 There is no difference in the outcome measure
#H1:u≠0 There is a difference in the outcome measure

chocolate_scores <- data$Overall[data$GroupName == "Chocolate"]
no_choc_scores <- data$Overall[data$GroupName == "NOChoc"]
observed_diff <- mean(chocolate_scores) - mean(no_choc_scores)
all_scores <- c(chocolate_scores, no_choc_scores)
n_perm <- 1999
perm_test_stats <- numeric(n_perm)
# Permutation test
for (i in 1:n_perm) {
  # Permute the group labels
  permuted_groups <- sample(all_scores)

  # Calculate the test statistic for the permuted data
  perm_chocolate <- permuted_groups[1:length(chocolate_scores)]
  perm_no_choc <- permuted_groups[(length(chocolate_scores) +
1):length(permuted_groups)]
  perm_test_stats[i] <- mean(perm_chocolate) - mean(perm_no_choc)
```

```
}

# Calculate the p-value
p_value <- sum(abs(perm_test_stats) >= abs(observed_diff)) / n_perm
# Display the results
cat("Mean Difference:", observed_diff, "\n")

## Mean Difference: 0.2222083

cat("Permutation Test p-value:", p_value, "\n")

## Permutation Test p-value: 0.1150575

#Permutation Test p-value: The p-value of 0.1150575 indicates there is a
11.505% chance of observing a mean difference as extreme as or more extreme
than data. Observed mean difference of 0.222 suggests participants in the
Chocolate group rated experience higher by 0.222 points. In the context of
this data, the treatment effect would be discovering that participants who
were given chocolate scored their experience higher due to receiving the
treat. However, given the significance level of 0.025 (to account for the
two-tailed nature of the test), the p-value of 0.1150575 is greater.
Therefore, we do not have sufficient evidence to conclude that there is a
statistically significant difference in the Overall ratings between the
Chocolate and NOChoc groups.

#(b) Consider the variable **Overall**. Is there a treatment effect with
respect to the professor's overall rating as a teacher? Apply the $t$-test to
the se data. Interpret the meaning of the $P$-value.

t_test <- t.test(chocolate_scores, no_choc_scores)
print(t_test)

##
##  Welch Two Sample t-test
##
## data:  chocolate_scores and no_choc_scores
## t = 1.6616, df = 95.93, p-value = 0.09986
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04324976  0.48766643
## sample estimates:
## mean of x mean of y
##  4.072000  3.849792

#From calculation we get t = 1.6616, df = 95.93, p-value = 0.09986. Since the
P-value is greater than the significance level(stated above), then I fail to
reject the null, and there is no significant difference in the mean overall
scores between the Chocolate and NOChoc groups.

#(c) Consider the test suggested in part (a). Why would a $t$-test not be a
recommended statistical method to carry out the test in part (a)? Explain
```

```
your answer in a few sentences.

#Using a t-test may not be recommended for the test suggested in part (a)
because it relies on certain assumptions about the data that may not hold
true. The t-test assumes that the data are normally distributed and that the
variances of both groups are equal. If these assumptions are violated, the
results of the t-test may be unreliable and could lead to misleading
conclusions. Additionally, while the t-test assumes independent observations
within each group, if this assumption is not met, it could further compromise
the validity of the results.
```

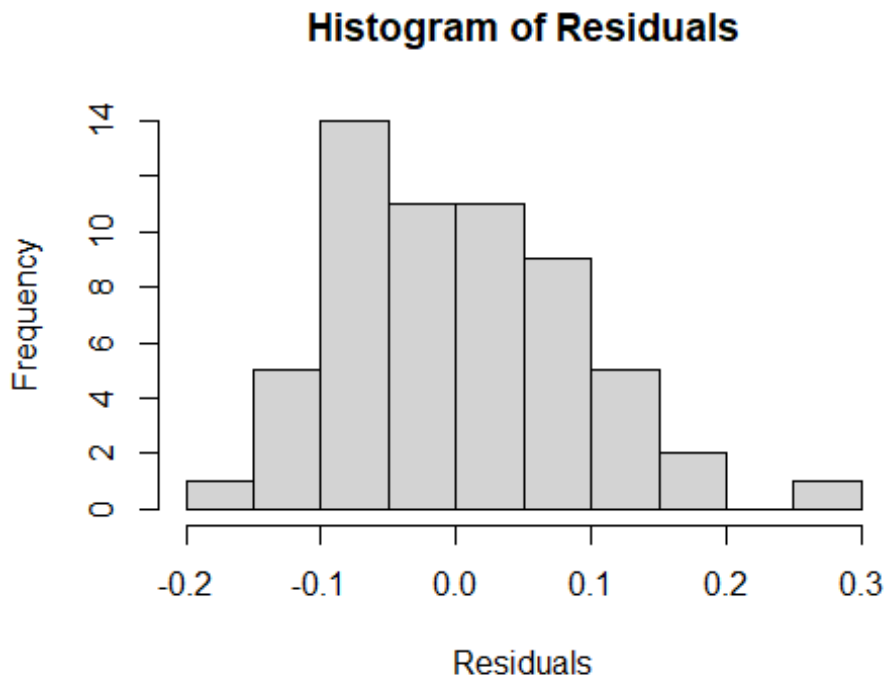
$$R_{StockA,i} = \beta_0 + \beta_1 R_{Market,i} + e_i$$

```
#Question Four


capmdata.df =
read.csv("https://raw.githubusercontent.com/Statman44/Data602/main/capm.csv")
head(capmdata.df)  #to get a sense of what the data look like

##      Suncor TSE.Index
## 1  0.008850 -0.023314
## 2  0.035088  0.041661
## 3 -0.016949  0.027904
## 4  0.043103  0.054967
## 5  0.126722  0.069449
## 6 -0.053790 -0.015124

#(a) Appropriately visualize these data. What can you infer from this
visualization? Provide a brief commentary.

ggplot(capmdata.df, aes(x = TSE.Index, y = Suncor)) + geom_point(col="red",
size = 2) + xlab("TSE Index") + ylab("Suncor") + ggtitle("TSE Index/Suncor
Monthly Return Rate")
```



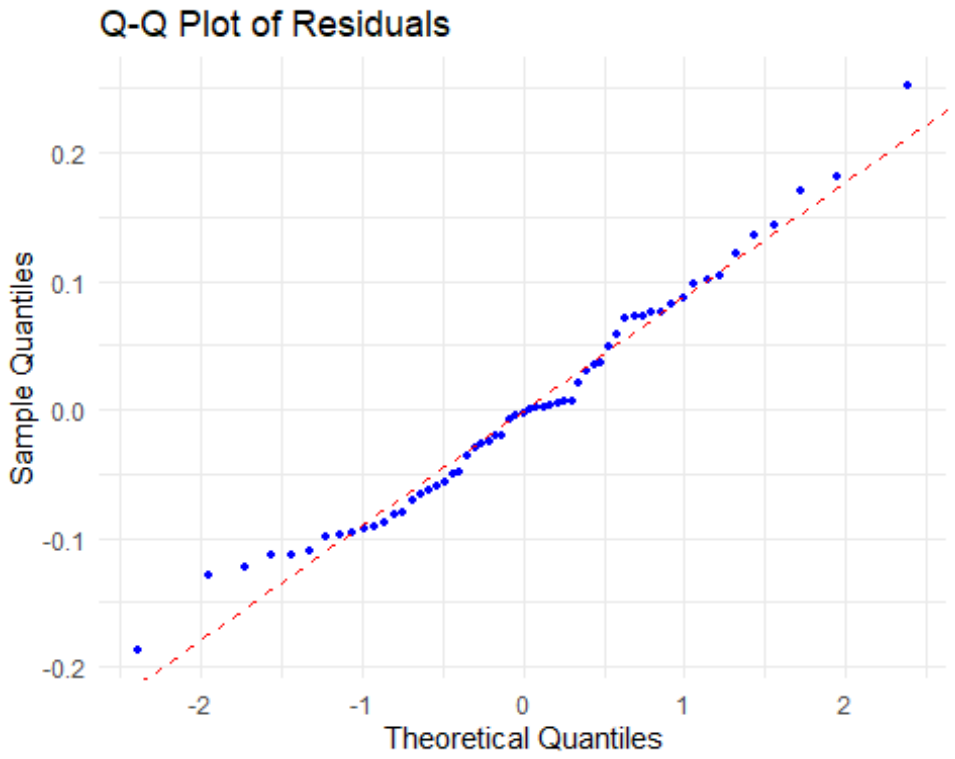

```
#Graph shows a positive relation between the variables TSE.Index and Suncor

#(b) Estimate the model above.

predict.suncorRate = lm(Suncor ~ TSE.Index, data = capmdata.df)
predict.suncorRate$coef

## (Intercept)   TSE.Index
##  0.01664794  0.53869099

#(c) In the context of these data, interpret the meaning of your estimates of
the estimates $\widehat{\beta}_{0}$ and $\widehat{\beta}_{1}$, in the context
of these data.
#As the TSE Index increases, we expect the rate of return of Suncor to
increase on average by the estimated coefficient value, which is
approximately 0.53869099

#(d) Refer to your answer in (b) In a certain month, the rate of return on
the TSE Index was 4%. Predict the rate of return on Suncor stock for the same
month.

predict(predict.suncorRate, data.frame(TSE.Index=0.04))

##           1
## 0.03819558
```

```r
#Predicated rate of return on Suncor stock will be approx: 0.03819558

#(e) Think about the conditions of this model *in the context* of these data. Create the visualizations that inspect each of the two conditions and provide commentary that addresses the validity (or invalidity) of each.

#Need to check for Normality of Residuals and Homoscedasticity

capmdata.df <-
read.csv("https://raw.githubusercontent.com/Statman44/Data602/main/capm.csv")
model <- lm(Suncor ~ TSE.Index, data = capmdata.df)
residuals <- residuals(model)

# Histogram of residuals - This will show the distribution of residuals
hist(residuals, main = "Histogram of Residuals", xlab = "Residuals")
```

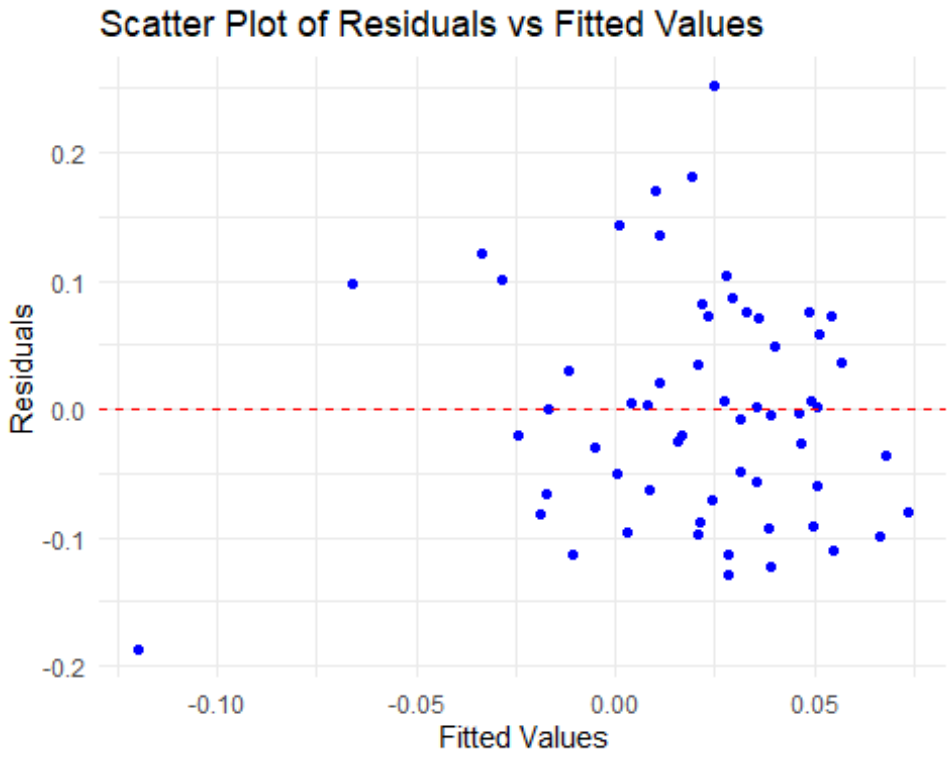



```r
#Q-Q plot of residuals-This will compare the distribution of residuals to a normal distribution

qq_plot <- ggplot(data = data.frame(residuals = residuals), aes(sample =
residuals)) +
  geom_qq(color = "blue", size = 1) +
  geom_abline(intercept = mean(residuals), slope = sd(residuals), color =
"red", linetype = "dashed") +
  labs(title = "Q-Q Plot of Residuals", x = "Theoretical Quantiles", y =
"Sample Quantiles") +
```

```
  theme_minimal()
print(qq_plot)
```




```
#Now test for homoscedasticity with a scatter plot:

model <- lm(Suncor ~ TSE.Index, data = capmdata.df)
residuals <- residuals(model)
fitted_values <- fitted(model)
residuals_df <- data.frame(Fitted_Values = fitted_values, Residuals =
residuals)
ggplot(residuals_df, aes(x = Fitted_Values, y = Residuals)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +  # Add red
line at y=0
  labs(title = "Scatter Plot of Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()
```




```
#(f) From these data, can you infer that the monthly rate of return of Suncor
stock can be expressed as a positive linear function of the monthly rate of
return of the TSE Index? State your statistical hypotheses, compute (and
report) both the test statistic and the $P$-value and provide your decision.

#Hypothesis Test:
#H0: B1<=0 The coefficient (B1) of the TSE Index in the regression model is
less than or equal to zero
#HA:B1>0 The coefficient (B1) of the TSE Index in the regression model is
greater than zero (B1 > 0)

summary(predict.suncorRate)

##
## Call:
## lm(formula = Suncor ~ TSE.Index, data = capmdata.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.187377 -0.068498 -0.003155  0.072103  0.252016
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01665    0.01177   1.414   0.1628
## TSE.Index    0.53869    0.19178   2.809   0.0068 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.08995 on 57 degrees of freedom
## Multiple R-squared:  0.1216, Adjusted R-squared:  0.1062
## F-statistic:  7.89 on 1 and 57 DF,  p-value: 0.006797

#From summary we see: TSE Index (B1) is 0.53869, with a standard error of
0.19178.
#The t-value associated with B1 is 2.809, and the corresponding p-value is
0.0068
(0.53869099-0)/0.19177963 # This is the computed T-test

## [1] 2.808906

#Given that p-value (0.0068) < significance level (0.05), we reject the null
hypothesis.  There is evidence to suggest that the monthly rate of return of
Suncor stock can be expressed as a positive linear function of the monthly
rate of return of the TSE Index.

#(g) Compute a 95% confidence interval for $\beta_{1}$, then interpret its
meaning in the context of these data.

confint(predict.suncorRate, conf.int=0.95)

##                    2.5 %     97.5 %
## (Intercept) -0.006928949 0.04022482
## TSE.Index    0.154658904 0.92272309

#We are 95% confident that the true value of the coefficient for TSE.Index
falls within the interval (0.1547, 0.9227). For each one-unit increase in the
monthly rate of return of the TSE Index, the monthly rate of return of Suncor
stock is expected to increase by an amount between approximately 0.1547 and
0.9227 units.

#(h) Compute a 95% confidence interval for the mean monthly rate of return of
Suncor stock when the TSE has a monthly rate of return of 3%.

predict(predict.suncorRate, newdata=data.frame(TSE.Index=0.03),
interval='conf')

##          fit         lwr        upr
## 1 0.03280867 0.007660256 0.05795708

#Suncor monthly rate will be 95% of the time between 0.00766 and 0.0579


#(i) In a month of September, the TSE Index had a rate of return of 1.16%.
With 95% confidence, compute the September rate of return for Suncor stock.

predict(predict.suncorRate, newdata=data.frame(TSE.Index=0.0116),
interval='conf')
```

```
##          fit           lwr        upr
## 1 0.02289675 -0.0006404417 0.04643395
```

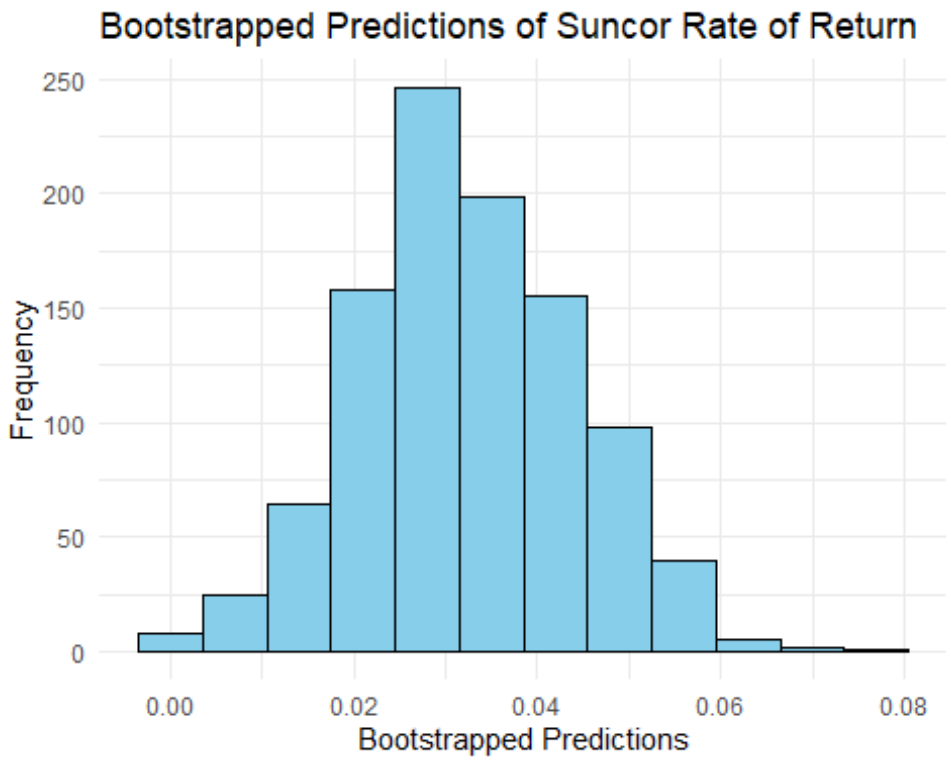
#Suncor stock for that month of September was between -0.00064 and 0.04643

*#(j) Recall the Bootstrap Method. From these data, use the bootstrap method to create a 95% confidence interval for mean monthly rate of return of Suncor stock when the TSE has a monthly rate of return of 3%. Compare your result to your result in part (h). Use 1000 iterations for your bootstrap. *Carefully* consider how you would re sample bivariate data points $(x_{TSE, i}, y_{Suncor, i})$.*

```
Nsims = 1000
TSE = 0.03
boot_predict = numeric(1000)
pair_x_y = dim(capmdata.df)[1]

for(i in 1:Nsims)
{
  index = sample(pair_x_y,replace=TRUE)
  dsample = capmdata.df[index, ]
  temp = lm(Suncor~TSE.Index, data = dsample)
  boot_predict[i] = (coef(temp)[1]) + ((coef(temp)[2])*TSE)
}

reg_bootstrap = data.frame(boot_predict)
head(reg_bootstrap,4)
```

```
##   boot_predict
## 1   0.02142309
## 2   0.04372953
## 3   0.01738620
## 4   0.04396872
```

```
boot_predict_df <- data.frame(boot_predict)
ggplot(boot_predict_df, aes(x = boot_predict)) +
  geom_histogram(binwidth = 0.007, fill = "skyblue", color = "black") +
  labs(title = "Bootstrapped Predictions of Suncor Rate of Return",
       x = "Bootstrapped Predictions",
       y = "Frequency") +
  theme_minimal()
```




```r
lower_bound <- quantile(boot_predict, 0.025)  # 2.5th percentile
upper_bound <- quantile(boot_predict, 0.975)  # 97.5th percentile

cat("95% Confidence Interval for Bootstrapped Predictions: [", lower_bound,
", ", upper_bound, "]\n")
```

```
## 95% Confidence Interval for Bootstrapped Predictions: [ 0.009418191 ,
0.05648683 ]
```

```r
#The confidence interval is very similar to what was fount in part h
```