CDS6224

STATISTICAL DATA ANALYSIS

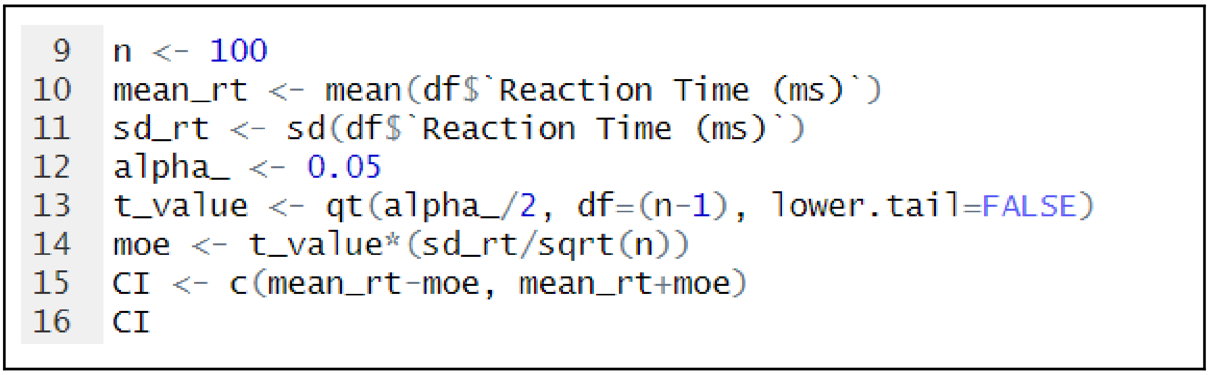
Assignment 2

Tutorial Section: TT2L

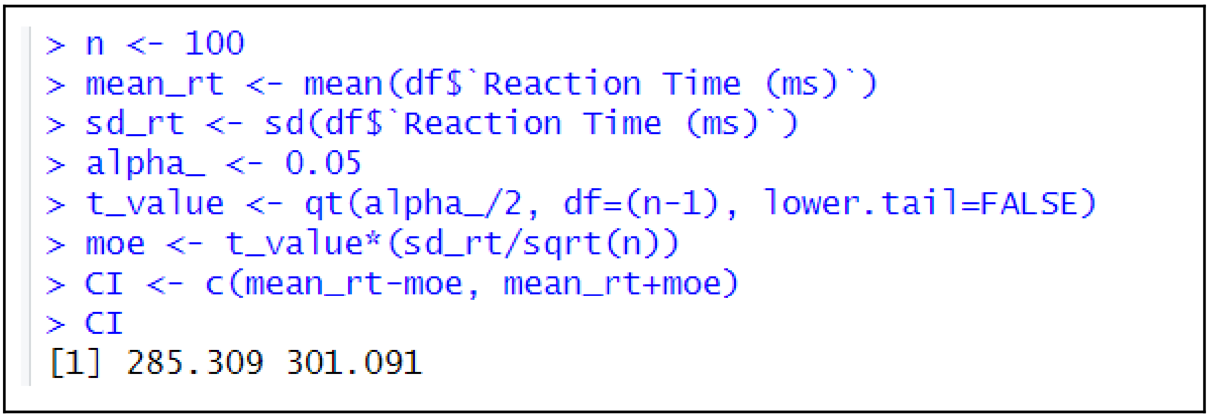
Group Number: G21

ID	Name	Email Address
1211103024	Yap Jack	1211103024@student.mmu.edu.my
1211101186	Tam Li Xuan	1211101186@student.mmu.edu.my
1201203287	Wong Wan Kei	1201203287@student.mmu.edu.my
1211101851	Ang Zhe Jie	1211101851@student.mmu.edu.my

1) The researcher believes that the average reaction time of all citizen in this town is 0.28 seconds.
a) Calculate a 95% confidence interval for the average reaction time.



```
n <- 100
mean_rt <- mean(df$`Reaction Time (ms)`)
sd_rt <- sd(df$`Reaction Time (ms)`)
alpha_ <- 0.05
t_value <- qt(alpha_/2, df=(n-1), lower.tail=FALSE)
moe <- t_value*(sd_rt/sqrt(n))
CI <- c(mean_rt-moe, mean_rt+moe)
CI
```

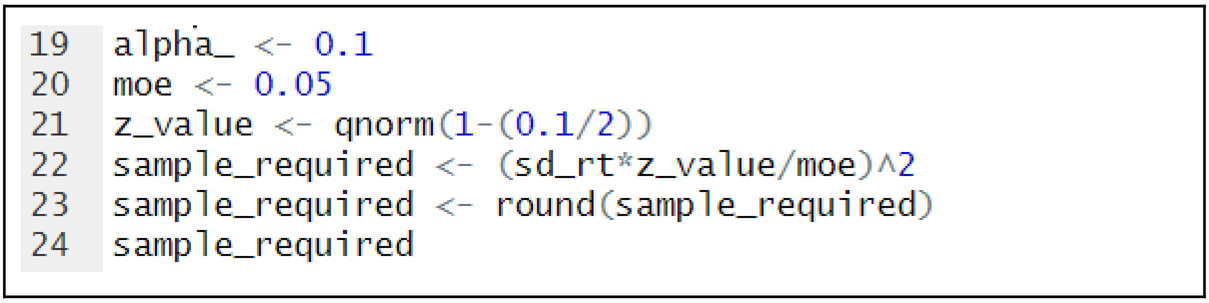


```
> n <- 100
> mean_rt <- mean(df$`Reaction Time (ms)`)
> sd_rt <- sd(df$`Reaction Time (ms)`)
> alpha_ <- 0.05
> t_value <- qt(alpha_/2, df=(n-1), lower.tail=FALSE)
> moe <- t_value*(sd_rt/sqrt(n))
> CI <- c(mean_rt-moe, mean_rt+moe)
> CI
[1] 285.309 301.091
```

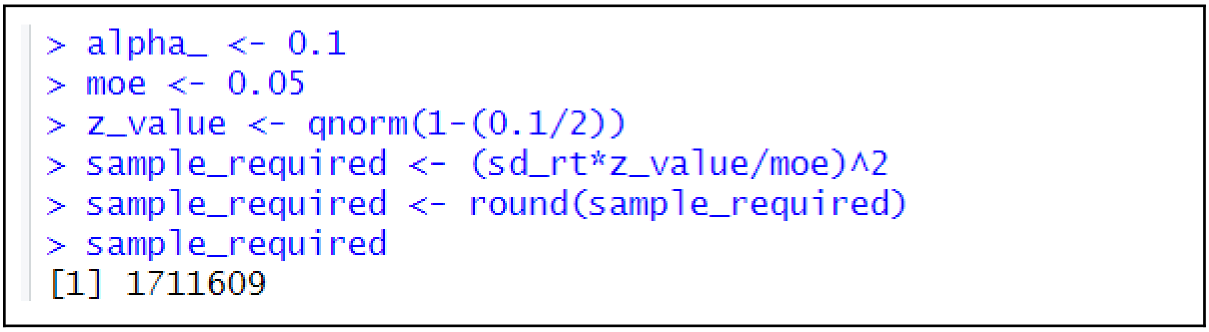
95% confidence interval for the average reaction time is (285.309, 301.091)

b) Do you agree with the researcher? Justify your answer based on your confidence interval in part (a).
Since the confidence interval (285.309, 301.091) does not include 280ms (0.28 seconds), we can conclude that the average reaction time of all citizens is not 280 ms. Therefore, we disagree with the researcher's belief that the average reaction time is 0.28 seconds.

c) Calculate the sample size if we want to be 90% confident that the estimate of population mean is off by at most 0.05.



```
alpha_ <- 0.1
moe <- 0.05
z_value <- qnorm(1-(0.1/2))
sample_required <- (sd_rt*z_value/moe)^2
sample_required <- round(sample_required)
sample_required
```

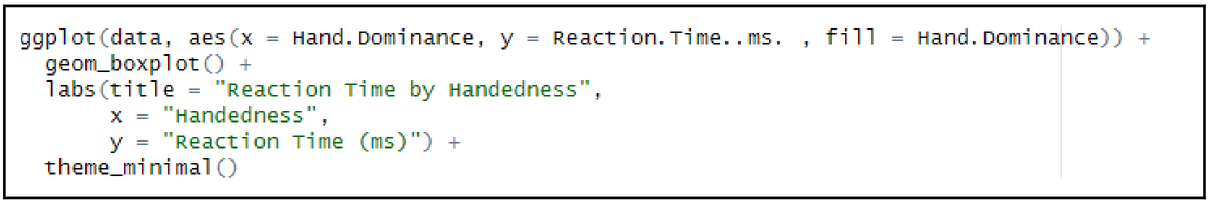


```
> alpha_ <- 0.1
> moe <- 0.05
> z_value <- qnorm(1-(0.1/2))
> sample_required <- (sd_rt*z_value/moe)^2
> sample_required <- round(sample_required)
> sample_required
[1] 1711609
```

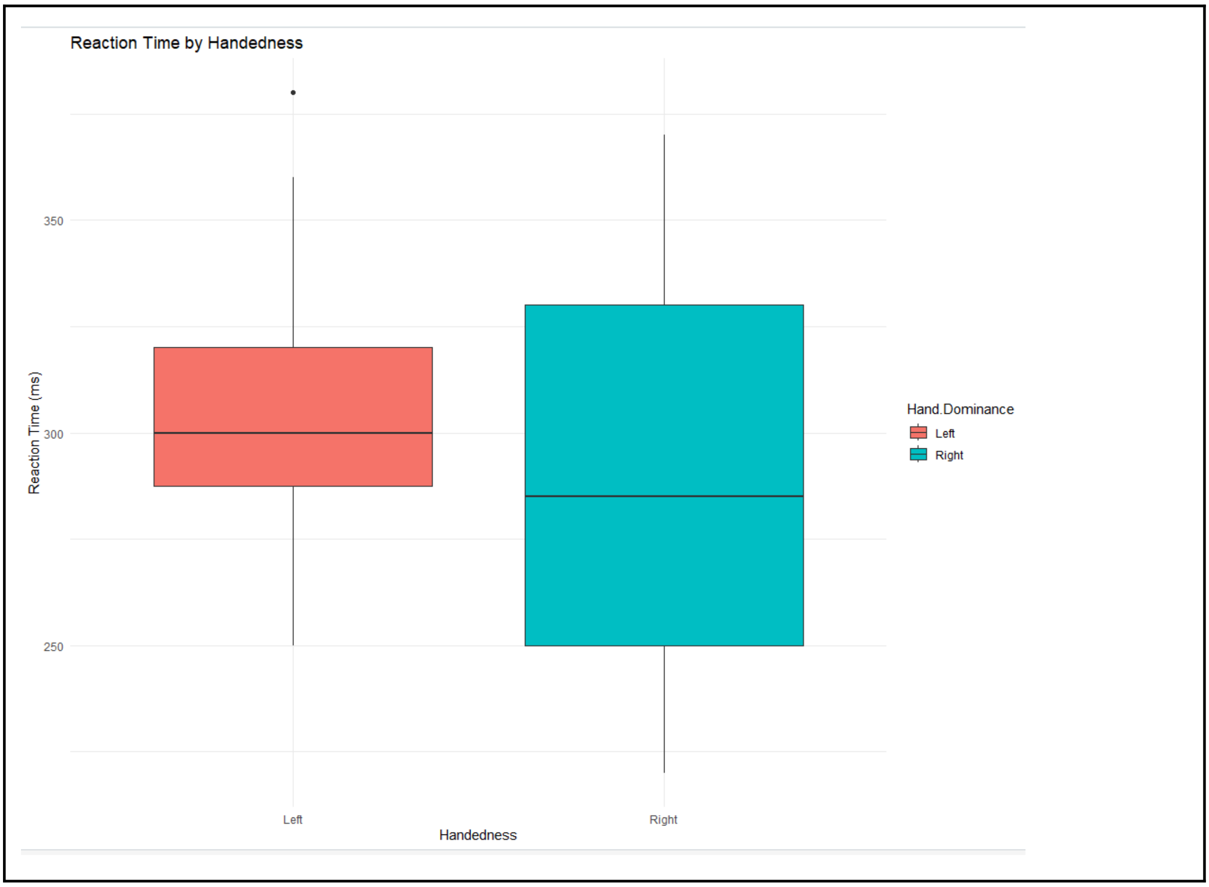
If we want to be 90% confident that the estimate of population mean is off by at most 0.05, the sample size is 1711609.

2) Is there a different in the reaction time for left-handed and right-handed citizen on average? You should

a) Explore the data with appropriate plot. Comment on your plot and answer the question.



```
ggplot(data, aes(x = Hand.Dominance, y = Reaction.Time..ms. , fill = Hand.Dominance)) +
  geom_boxplot() +
  labs(title = "Reaction Time by Handedness",
       x = "Handedness",
       y = "Reaction Time (ms)") +
  theme_minimal()
```



The boxplot and descriptive statistics show that there is only a little difference in reaction times between the left- and right-handed users, with those with right hands having a slightly faster reaction time. Right-handed people have greater variability in their reaction times, shown by the wider IQR of approximately 250 ms to 330ms compared to the IQR of left-handed people, which is approximately 288 ms to 320 ms. Moreover, the median reaction time of left-handed people is slightly higher than the reaction time of right-handed people.

An official hypothesis test, such as an independent t-test, might be used to determine whether the difference is statistically significant. However, based on hand dominance, we detect a modest variance in reaction times for descriptive and visual searches.

b) Calculate an appropriate confidence interval to determine if left-handed and right-handed citizens have different variability in their reaction time. Use a 97% confidence level. You must write all the steps clearly

Calculation of Variance (Steps)

1. **Count the Sample Sizes**
   - Let n_1 be the sample size for left-handed individuals.
   - Let n_2 be the sample size for right-handed individuals.

   Given: n_1=32, n_2=68

2. **Calculate the Sum of Reaction Times**
   - Let $\sum x_1$ be the sum of reaction times for left-handed individuals.
   - Let $\sum x_2$ be the sum of reaction times for right-handed individuals.
3. **Calculate the Mean of Reaction Times**
   - Mean reaction time for left-handed individuals $\bar{x}_1$ is calculated as:

     $\bar{x}_1 = 9690 / 32 = 302.8125$

   - Mean reaction time for right-handed individuals $\bar{x}_2$ is calculated as:

     $\bar{x}_2 = 19630 / 68 = 288.6765$

4. **Calculate the Sum of Squared Differences of Reaction Times**
   - Sum of squared differences for left-handed individuals:

     $\sum(x_1 - 302.8125)^2 = 26846.88$

   - Sum of squared differences for right-handed individuals:

     $\sum(x_2 - 288.6765)^2 = 125380.9$

5. **Calculate Sample Variances**
    - Sample variance for left-handed individuals (s^2_1):

      $s^2_1 = 26846.88 / (32 – 1) = 866.0282$

    - Sample variance for right-handed individuals (s^2_2):

      $s^2_2 = 125380.9 / (68 – 1) = 1871.356$

Calculate F-value

1. **Calculate the Ratio of Both Variances**
    - The ratio of variances (F)

      $s^2_1 / s^2_2 = 866.0282 / 1871.356 = 0.462781$

2. **Determine the Significance Level (α)**
    - Given a 97% confidence level:

      $\alpha=1-0.97=0.03$

    - Thus, $\alpha/2=0.015$
3. **Find the F-value from F-distribution Tables**

      $f_{\mathbf{0.015, 32\text{-}1, 68\text{-}1}} = 0.487951$

      $f_{\mathbf{1\text{-}0.015, 32\text{-}1, 68\text{-}1}} = 1.892772$

Determine the Confidence Interval

- Calculate the Lower and Upper Bounds of the Confidence Interval
    - Lower bound

      lower bound = 0.462781/ 1.892772 = 0.2444990

    - Upper bound:

      upper bound = 0.462781/0.487951 = 0.9484167

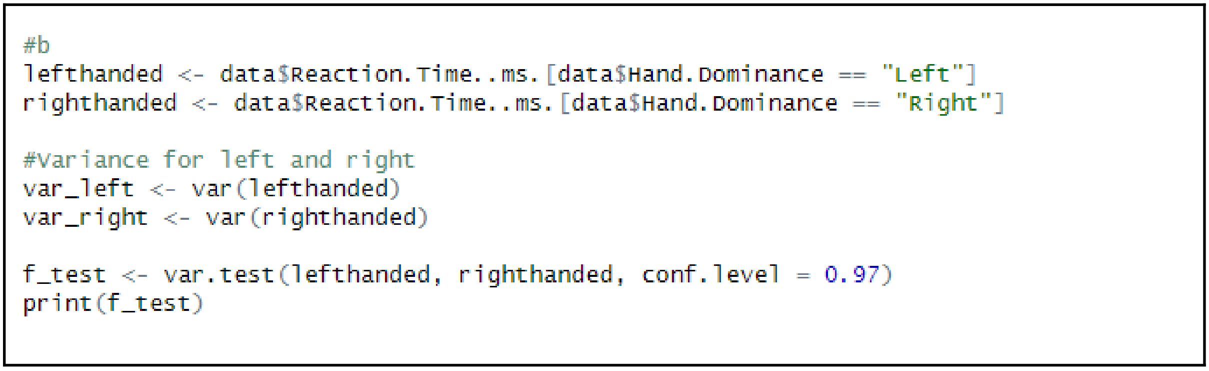
Interpretation

- The 97% confidence interval for the ratio of variances is:

  (0.2444990,0.9484167)

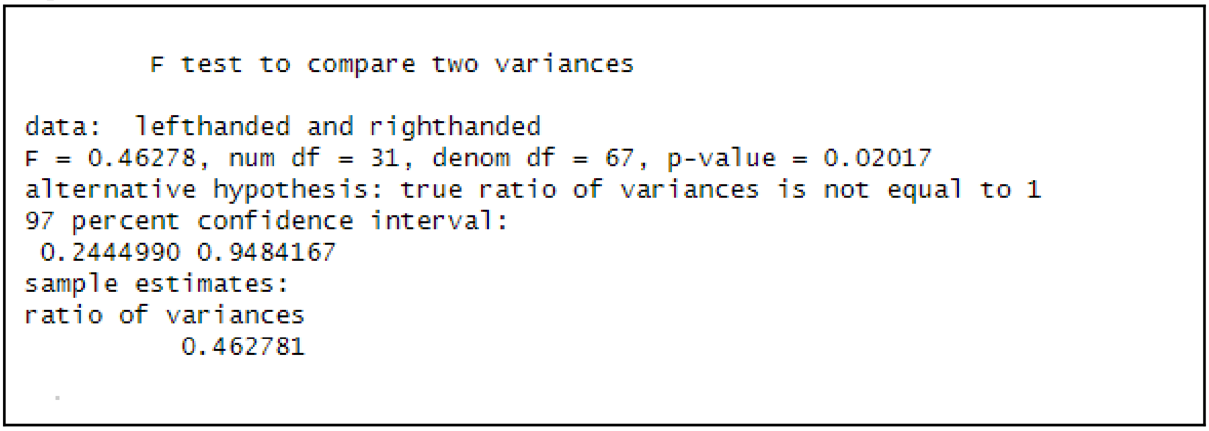
- Since the confidence interval does not include 1, we reject the null hypothesis that the variances are equal. This indicates a statistically significant difference in variability between left-handed and right-handed individuals' reaction times.

Rcode Operation:



```
#b
lefthanded <- data$Reaction.Time..ms.[data$Hand.Dominance == "Left"]
righthanded <- data$Reaction.Time..ms.[data$Hand.Dominance == "Right"]

#Variance for left and right
var_left <- var(lefthanded)
var_right <- var(righthanded)

f_test <- var.test(lefthanded, righthanded, conf.level = 0.97)
print(f_test)
```

Output:



```

	F test to compare two variances

data:  lefthanded and righthanded
F = 0.46278, num df = 31, denom df = 67, p-value = 0.02017
alternative hypothesis: true ratio of variances is not equal to 1
97 percent confidence interval:
 0.2444990 0.9484167
sample estimates:
ratio of variances
          0.462781
```

Using Rcode the F-test was used at a 97% level of trust to determine differences in reaction times between left- and right-handed people. The findings reveal a p-value of 0.02017, which is less than the threshold of 0.03, showing a statistically significant difference in variances. The difference is further supported by the fact of the ratio of variances' confidence interval (0.2444990 to 0.9484167) does not contain 1. It means that people who are right-handed have more variability in their reaction times than people who are left-handed.

c) Construct an appropriate hypothesis test to determine if there a different in the reaction time for left-handed and right-handed citizen on average. Use a 3% level of significance. You must write all the steps clearly.

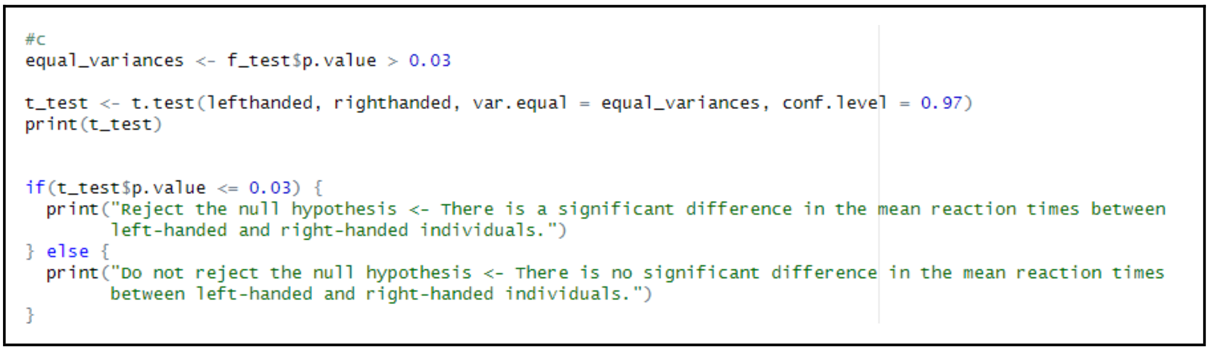
1. **Define the Hypotheses**
    - **Null Hypothesis (H_0)**: There is no difference in the mean reaction times between left-handed and right-handed individuals. ($\mu_1 = \mu_2$)
    - **Alternative Hypothesis (H_a)**: There is a difference in the mean reaction times between left-handed and right-handed individuals. ($\mu_1 \neq \mu_2$)
2. **Sample Means**
    - **Given**
      $n_1 = 32$, $n_2 = 68$
      $\sum x_1 = 9690$, $\sum x_2 = 19630$
    - **Mean reaction time for left-handed individuals:**
      $\bar{x}_1 = 9690 / 32 = 302.8125$
    - **Mean reaction time for right-handed individuals:**
      $\bar{x}_2 = 19630 / 68 = 288.6765$
3. **Difference in means ($\bar{x}_1$ - $\bar{x}_2$):**
    - 302.8125−288.6765=14.136
4. **Sample Variances**
    - $s^2_1 = 866.0282$
    - $s^2_2 = 1871.356$
5. **Standard error (SE):**
   SE ≈ 7.387
6. **Extract the T-value:**
    - T-value:
      $t =(\bar{x}_1 - \bar{x}_2)/SE$
      $t =1.9134$
7. **Degrees of Freedom**
    - **Degrees of freedom (df):**

$$df = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{(s_1^2/N_1)^2}{N_1-1} + \frac{(s_2^2/N_2)^2}{N_2-1}}$$

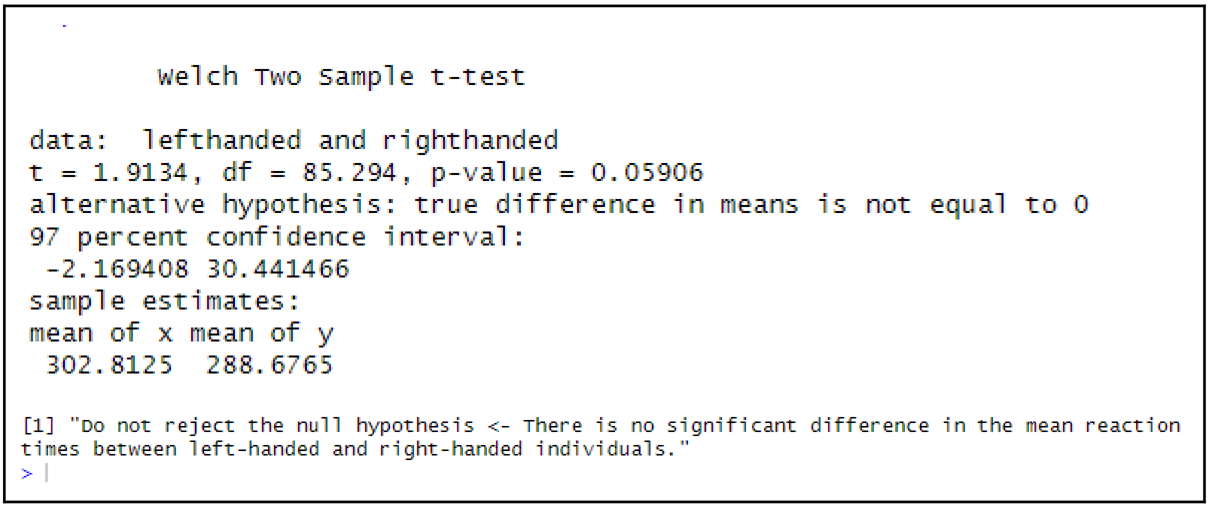
      s^2_1 = 866.0282
      $s^2_2 = 1871.356$
      $n_1 = 32$, $n_2 = 68$
      df ≈ 85.294
8. **Determine the P-value**
    - P-value ≈ 0.05906

Conclusion

- Since the p-value (0.05906) is greater than the significance level (0.03), we fail to reject the null hypothesis.
- The 97% confidence interval for the difference in means is (−2.169408,30.441466), which includes 0, further indicating that there is no statistically significant difference in the mean reaction times between left-handed and right-handed individuals.



```
#c
equal_variances <- f_test$p.value > 0.03

t_test <- t.test(lefthanded, righthanded, var.equal = equal_variances, conf.level = 0.97)
print(t_test)


if(t_test$p.value <= 0.03) {
  print("Reject the null hypothesis <- There is a significant difference in the mean reaction times between
        left-handed and right-handed individuals.")
} else {
  print("Do not reject the null hypothesis <- There is no significant difference in the mean reaction times
        between left-handed and right-handed individuals.")
}
```



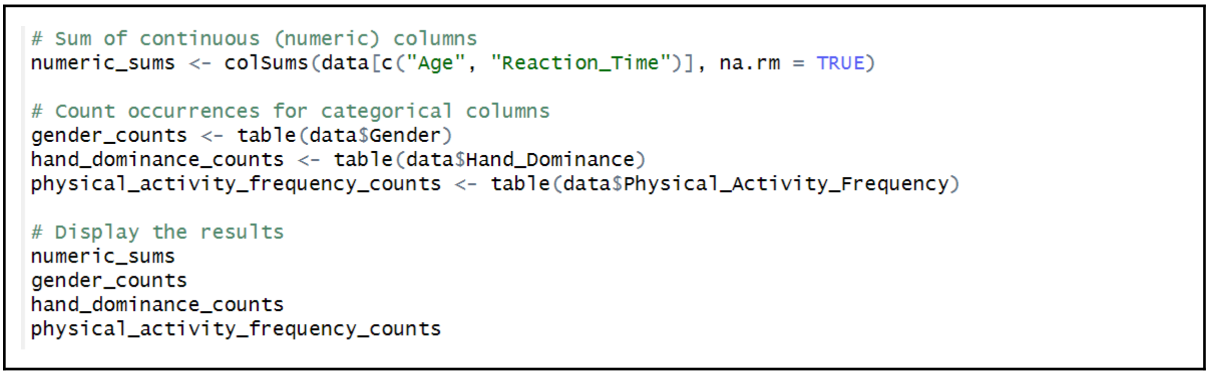
```
	Welch Two Sample t-test

data:  lefthanded and righthanded
t = 1.9134, df = 85.294, p-value = 0.05906
alternative hypothesis: true difference in means is not equal to 0
97 percent confidence interval:
 -2.169408 30.441466
sample estimates:
mean of x mean of y 
 302.8125  288.6765 

[1] "Do not reject the null hypothesis <- There is no significant difference in the mean reaction
times between left-handed and right-handed individuals."
> 
```

By the functions of R Code. We used a Welch Two Sample t-test using a 3% significance (confidence interval of 97%) threshold for determining if left- and right-handed people's response speeds differed. Based on the null hypothesis (H0), there isn't a clear distinction between the two groups' mean reaction times. T-value was 1.9134, df was 85.294, and p was 0.05906 according to test findings. Since p-value exceeds 0.03, there is not enough evidence to reject the null hypothesis. Therefore, there's no discernible difference in the average reaction times of right- and left-handed people.

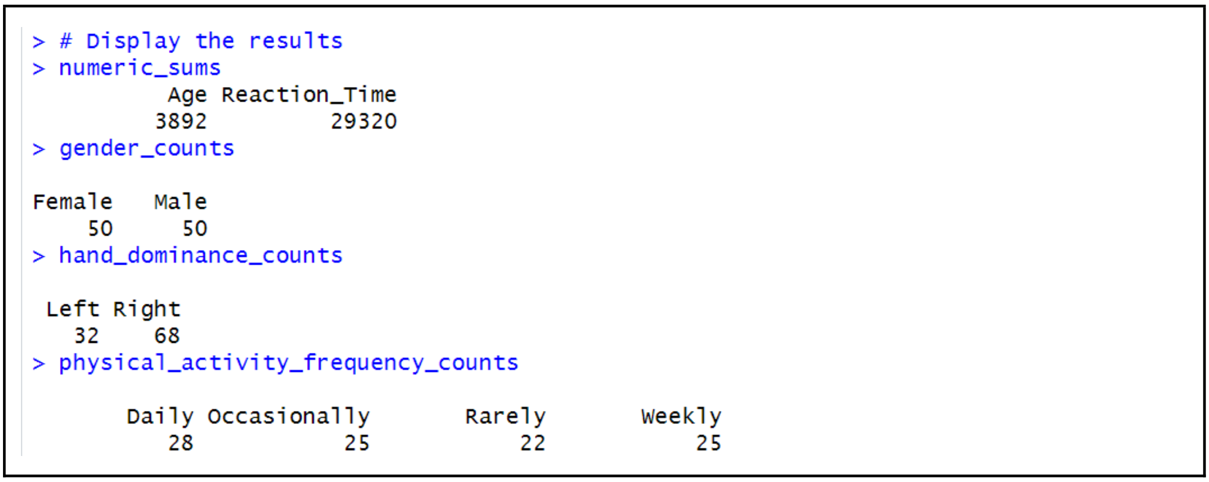
3) The researcher wants to build a regression model for respond time, with the other 4 variables in the dataset as explanatory variables.

a) Construct a linear regression model for respond time on the 4 variables and write your model.

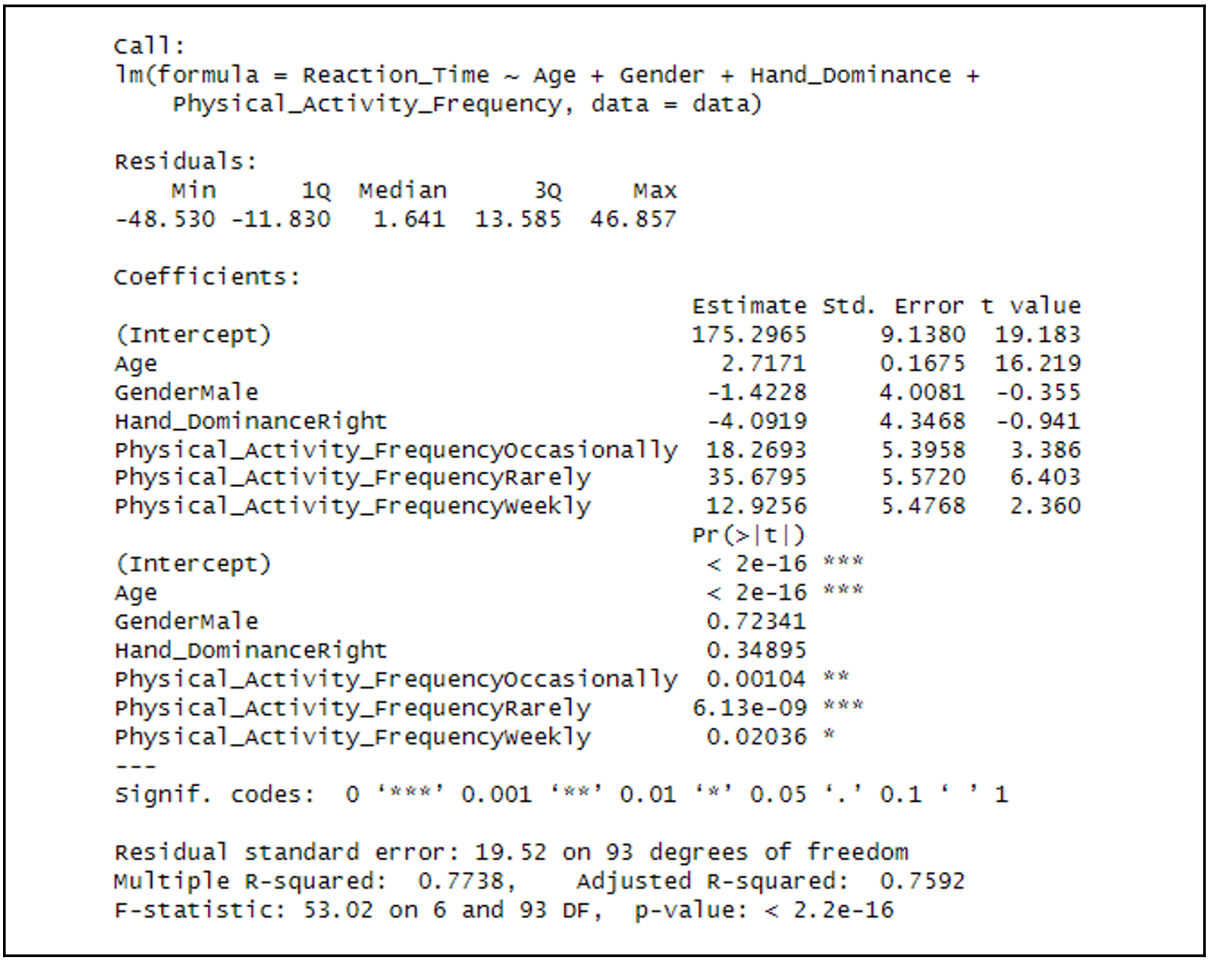
1. Calculate Total for each variables



```
# Sum of continuous (numeric) columns
numeric_sums <- colSums(data[c("Age", "Reaction_Time")], na.rm = TRUE)

# Count occurrences for categorical columns
gender_counts <- table(data$Gender)
hand_dominance_counts <- table(data$Hand_Dominance)
physical_activity_frequency_counts <- table(data$Physical_Activity_Frequency)

# Display the results
numeric_sums
gender_counts
hand_dominance_counts
physical_activity_frequency_counts
```



```
> # Display the results
> numeric_sums
          Age Reaction_Time
         3892         29320
> gender_counts

Female    Male
    50      50
> hand_dominance_counts

 Left Right
   32    68
> physical_activity_frequency_counts

       Daily Occasionally       Rarely       Weekly
          28           25           22           25
```



```
Call:
lm(formula = Reaction_Time ~ Age + Gender + Hand_Dominance +
    Physical_Activity_Frequency, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-48.530 -11.830   1.641  13.585  46.857

Coefficients:
                                       Estimate Std. Error t value
(Intercept)                            175.2965     9.1380  19.183
Age                                      2.7171     0.1675  16.219
GenderMale                              -1.4228     4.0081  -0.355
Hand_DominanceRight                     -4.0919     4.3468  -0.941
Physical_Activity_FrequencyOccasionally  18.2693     5.3958   3.386
Physical_Activity_FrequencyRarely        35.6795     5.5720   6.403
Physical_Activity_FrequencyWeekly        12.9256     5.4768   2.360
                                       Pr(>|t|)
(Intercept)                             < 2e-16 ***
Age                                     < 2e-16 ***
GenderMale                              0.72341
Hand_DominanceRight                     0.34895
Physical_Activity_FrequencyOccasionally 0.00104 **
Physical_Activity_FrequencyRarely      6.13e-09 ***
Physical_Activity_FrequencyWeekly       0.02036 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.52 on 93 degrees of freedom
Multiple R-squared:  0.7738,    Adjusted R-squared:  0.7592
F-statistic: 53.02 on 6 and 93 DF,  p-value: < 2.2e-16
```

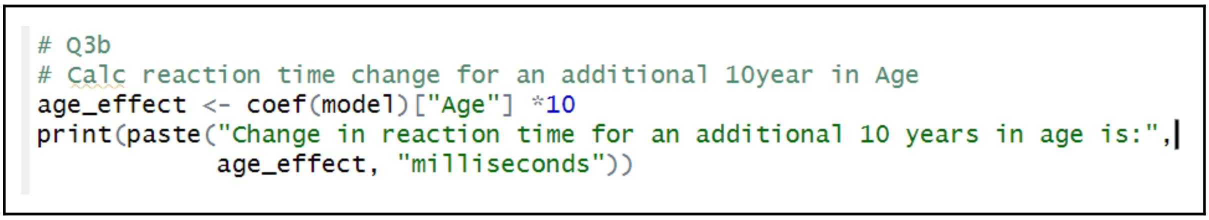
$Multiple\ Linear\ Regression\ Equation$:

$$Reaction\ Time = \beta_0 + \beta_1 \times Age + \beta_2 \times Gender + \beta_3 \times Hand\ Dominance + \beta_4 \times Physical\ Activity\ Frequency$$

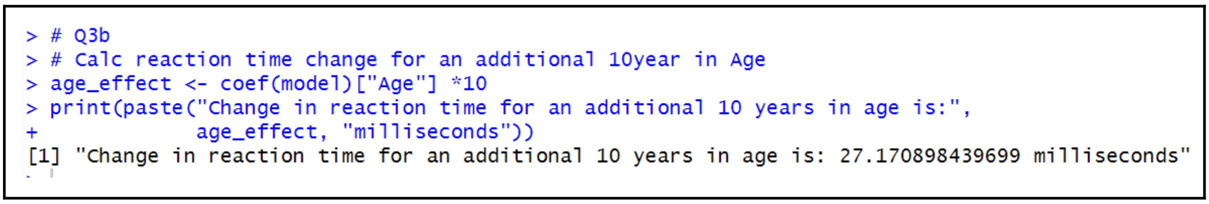
$$Reaction\ Time = 175.2965 + 2.7171 \times Age - 1.4228 \times Gender_{Male} - 4.0919 \times Hand\ Dominance_{Right} + 18.2693 \times Physical\ Activity\ Frequency_{Occasionally} + 35.6795 \times Physical\ Activity\ Frequency_{Rarely} + 35.6795 \times Physical\ Activity\ Frequency_{Weekly}$$

Based on the regression analysis result, the model effectively predicts the reaction time using a combination of age and physical activity frequency. It is because the predictors of age and physical activity frequency (occasionally, rarely, and weekly) are statistically significant, as their p-values are less than 0.05. The changes in these predictors are associated with significant changes in reaction time. Conversely, Gender (Male) and Hand_Dominance (Right) have p-values greater than 0.05, which indicates that 72% of Gender and 35% of Hand Dominance are not meaningful for the regression. Overall, the Adjusted R-squared value shows the model explains approximately 76% of the total variability in reaction time.

b) How is the reaction time change for an additional of 10 years in age? Show all the steps of your calculation.



```
# Q3b
# Calc reaction time change for an additional 10year in Age
age_effect <- coef(model)["Age"] *10
print(paste("Change in reaction time for an additional 10 years in age is:",
            age_effect, "milliseconds"))
```



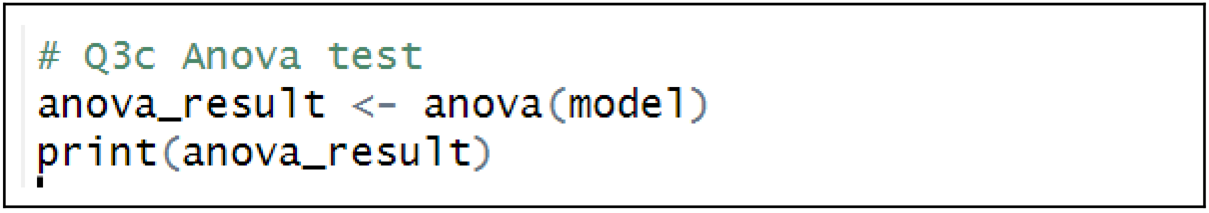
```
> # Q3b
> # Calc reaction time change for an additional 10year in Age
> age_effect <- coef(model)["Age"] *10
> print(paste("Change in reaction time for an additional 10 years in age is:",
+             age_effect, "milliseconds"))
[1] "Change in reaction time for an additional 10 years in age is: 27.170898439699 milliseconds"
```

Change in reaction time = 2.7171 x 10 = 27.171

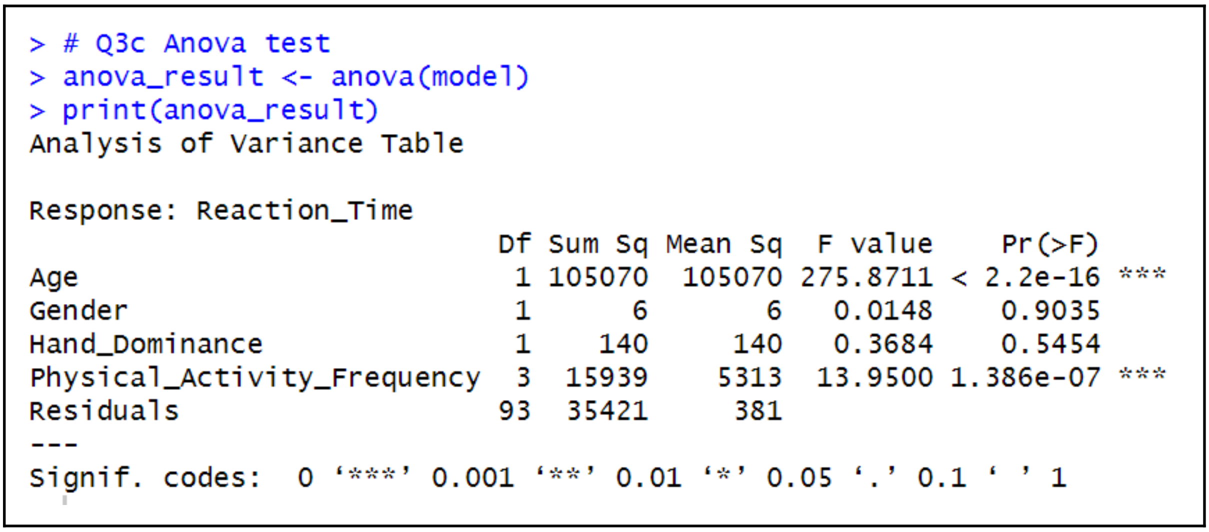
In every additional year, reaction time increases by approximately 27.171 milliseconds. This significant change suggests that as individuals age, their reaction times slow down considerably, possibly indicating ageing-related changes in the body and cognition.

c) Test the significance of your model with the ANOVA approach. You must write all the steps clearly.

R Code Generated ANOVA approach



```
# Q3c Anova test
anova_result <- anova(model)
print(anova_result)
```



```
> # Q3c Anova test
> anova_result <- anova(model)
> print(anova_result)
Analysis of Variance Table

Response: Reaction_Time
                            Df Sum Sq Mean Sq  F value    Pr(>F)    
Age                          1 105070  105070 275.8711 < 2.2e-16 ***
Gender                       1      6       6   0.0148    0.9035    
Hand_Dominance               1    140     140   0.3684    0.5454    
Physical_Activity_Frequency  3  15939    5313  13.9500 1.386e-07 ***
Residuals                   93  35421     381                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1. Age:

   H_0: $\beta_{Age} = 0$ (Age has no effect on Reaction Time)

   H_1: $\beta_{Age} \neq 0$ (Age affects Reaction Time)

Since the p-value is less than 2.2e-16, which is much less than 0.05. Hence, we reject the null hypothesis and conclude that Age has a significant effect on Reaction Time.

2. Gender:
   H_0: $\beta_{Gender} = 0$ (Gender has no effect on Reaction Time)
   H_1: $\beta_{Gender} \neq 0$ (Gender affects Reaction Time)

The p-value of Gender is 0.9035, which is greater than 0,05. Therefore, we do not reject the null hypothesis, this means that Gender does not significantly affect and contribute to Reaction Time.

3. Hand_Dominance:
   H_0: $\beta_{Hand\ Dominance} = 0$ (Hand_Dominance has no effect on Reaction Time)
   H_1: $\beta_{Hand\ Dominance} \neq 0$ (Hand_Dominance affects Reaction Time)

The p-value of Hand Dominance is 0.5454, which is greater than 0.05. Hence, we do not reject the null hypothesis and conclude that Hand Dominance does not significantly affect Reaction Time.

4. Physical_Activity_Frequency:
   H_0: $\beta_{Hand\ Dominance} = 0$ (Physical Activity Frequency has no effect on Reaction Time)
   H_1: $\beta_{Hand\ Dominance} \neq 0$ (Physical Activity Frequency affects Reaction Time)

The p-value of physical activity frequency is 1.386e-07, which is much < 0.05. Hence, we reject the null hypothesis and conclude that Physical Activity. This indicates that Physical Activity Frequency significantly affects Reaction Time.

In conclusion, the overall model is significant, meaning at least one predictor variable affects Reaction Time. This regression model explains a significant portion of the variation in Reaction Time due to Age and Physical Activity Frequency, as their coefficients are not zero. Gender and Hand Dominance do not significantly contribute to the model, indicating no significant effect on Reaction Time.

Hand Written ANOVA approach

SSR: $\sum(\hat{y}_i - \bar{y})^2 = 121155.39$

SSE: $\sum(y_i - \hat{y}_i)^2 = 35420.61$

SST = 120563.443714 + 36012.556286 = 156576

MSR = 121155.39 / 6 = 20192.5654 df(R) = 6

MSE = 35420.61 / 93 = 380.8667 df(E) = 100 - 6 - 1 = 93

f = 20192.5654 / 380.8667 = 53.0174

$f_{0.05, 6, 93} = 2.197679$

ANOVA

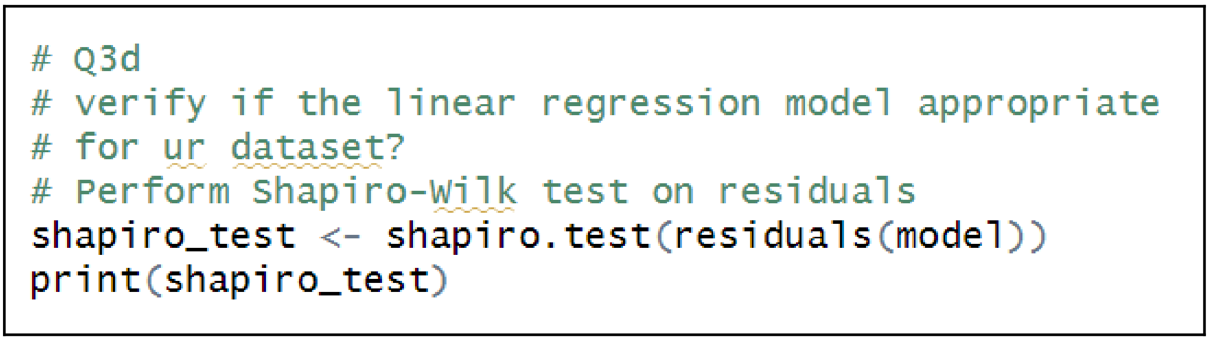
Source of Variation	Sum of Squares	Degree of Freedom (df)	Mean Square	Test Statistic
Regression	121155.39	6	20192.5654	$f = 53.0174$
Error	35420.61	93	380.8667	
Total	156576	99		

$f > f_{0.05, 6, 93}$

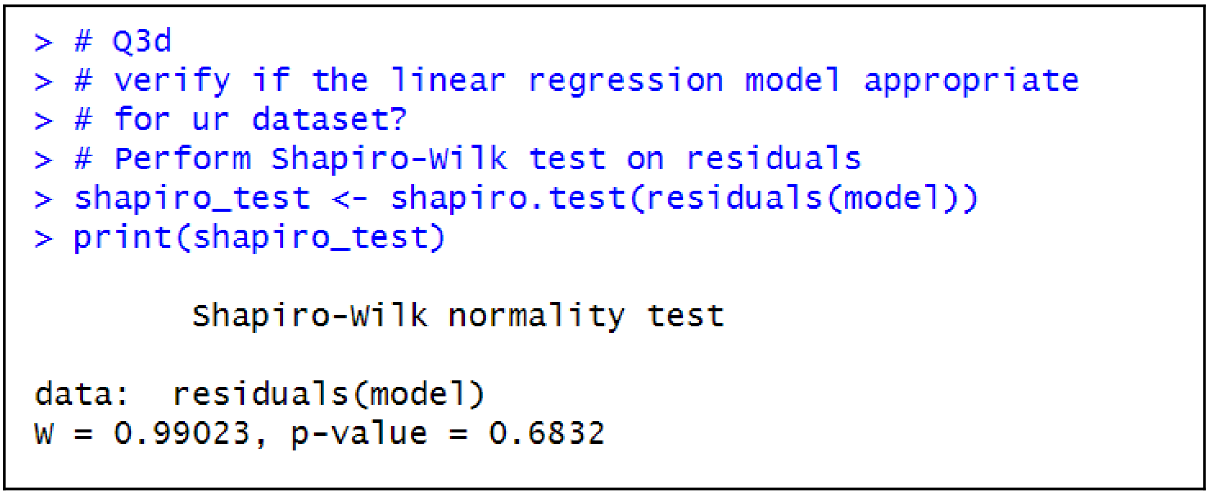
There is enough evidence to reject that the fit of the intercept-only model and our model are equal at 0.05 level of significance

d) Verify if the linear regression model is appropriate for your dataset. You must justify your answer based on appropriate plot. [11 marks]

r^2, f-statistic, residual plot

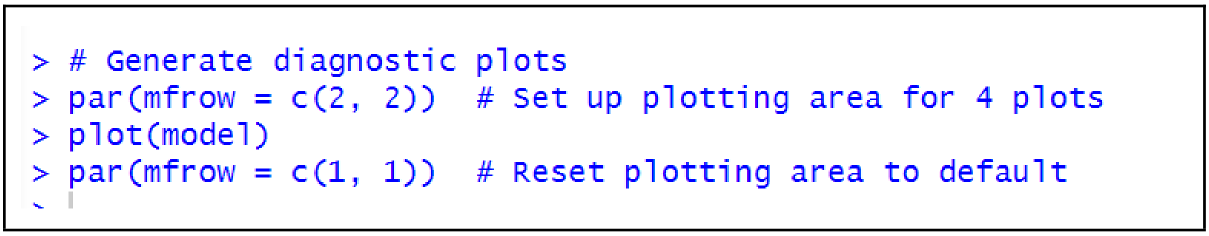


```
# Q3d
# verify if the linear regression model appropriate
# for ur dataset?
# Perform Shapiro-Wilk test on residuals
shapiro_test <- shapiro.test(residuals(model))
print(shapiro_test)
```

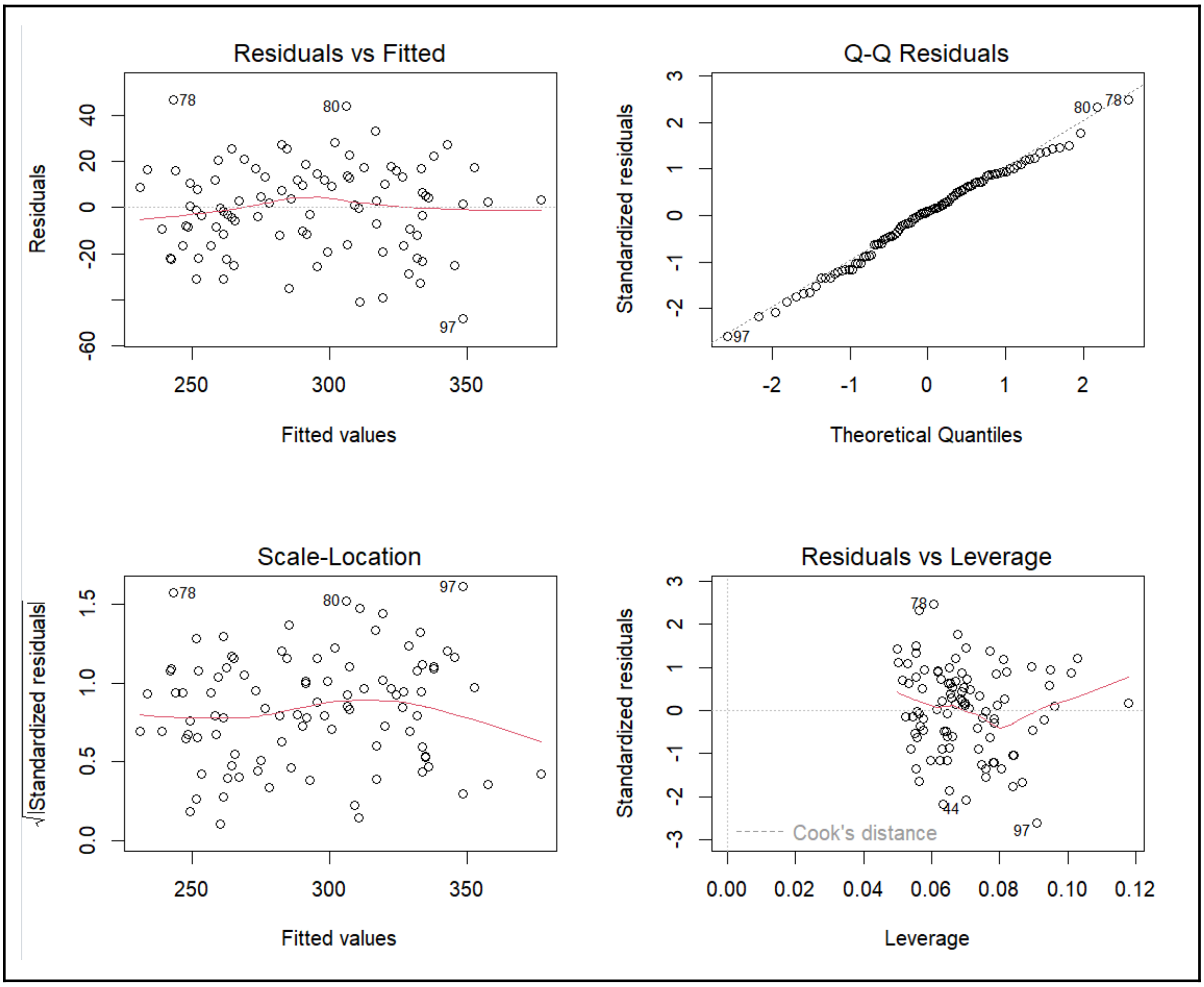


```
> # Q3d
> # verify if the linear regression model appropriate
> # for ur dataset?
> # Perform Shapiro-Wilk test on residuals
> shapiro_test <- shapiro.test(residuals(model))
> print(shapiro_test)

        Shapiro-Wilk normality test

data:  residuals(model)
W = 0.99023, p-value = 0.6832
```

Based on the result of Shapiro-Wilk Test, p-value = 0.6832 which is greater than 0.05. This suggests that the residuals of the linear regression model are normally distributed.



```
> # Generate diagnostic plots
> par(mfrow = c(2, 2))  # Set up plotting area for 4 plots
> plot(model)
> par(mfrow = c(1, 1))  # Reset plotting area to default
```



First, the Residuals vs Fitted Plot shows that the residuals seem to be relatively evenly spread around the horizontal line, indicating no clear pattern and suggesting that the model's linearity assumption is met. However, there are some clustering of residuals, which could suggest minor violations of homoscedasticity, although the overall spread does not indicate significant heteroscedasticity. Points 78, 80 and 97 are identified as outliers in this plot.

The Q-Q Plot shows that most of the residuals fall along the reference line, this indicates the residuals are approximately normally distributed. This suggests that the normality assumption is satisfied, although points 78,80 and 97 diverge at the tails, suggesting possible outliers.

The Scale-Location Plot shows a relatively horizontal red line with an even spread residuals, more evidence to support of the homoscedasticity hypothesis. However, the same outliers (78, 80 and 97) are present, which may need further investigation.

Lastly, the Residuals vs Leverage Plot does not show any values outside Cook's distance lines, indicating no highly influential outliers. Hence, points 78, 80 and 97 have more influence without excessively impacting the model.

In conclusion, the diagnostic plots point to a reasonable fit between the linear regression model and the dataset, given that the assumptions of linearity, residual normality and homoscesdasticity are largely satisfied.