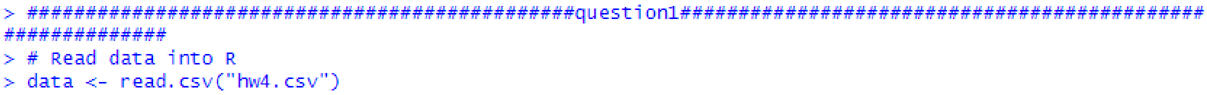
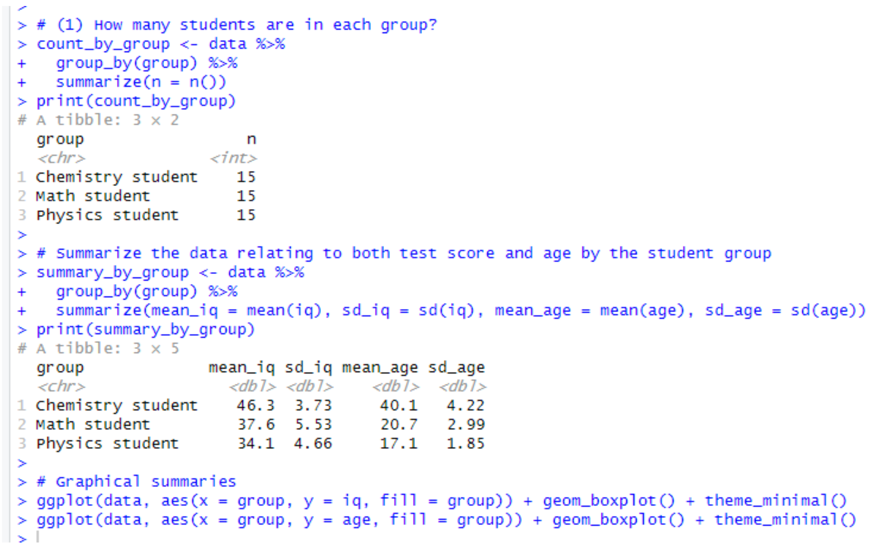
Question 1/2. The data in this document is from 3 groups of students (math, chemistry, and physics) on an IQ related test. Save the data, and read the data into R. Use this data to address the following questions (14 points)

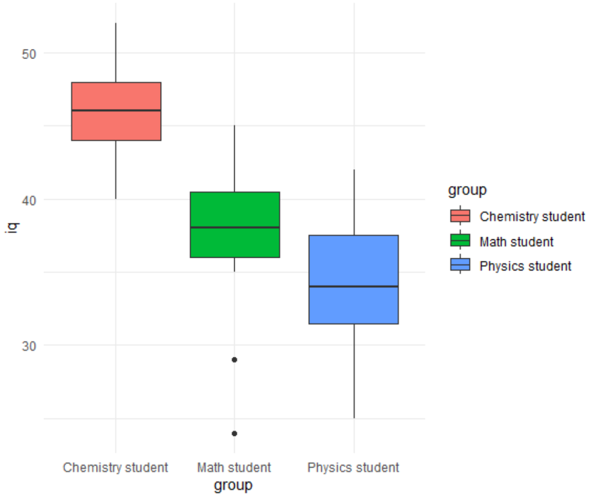


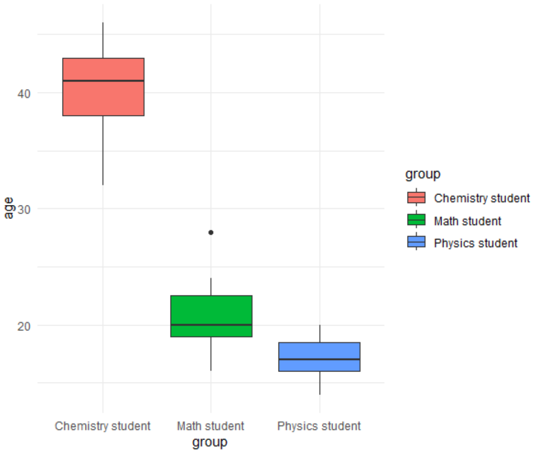
```
> ##################################################question1################################################
###############
> # Read data into R
> data <- read.csv("hw4.csv")
```

(1) How many students are in each group? Summarize the data relating to both test score and age by the student group (separately). Use appropriate numerical and/or graphical summaries. (3 points)

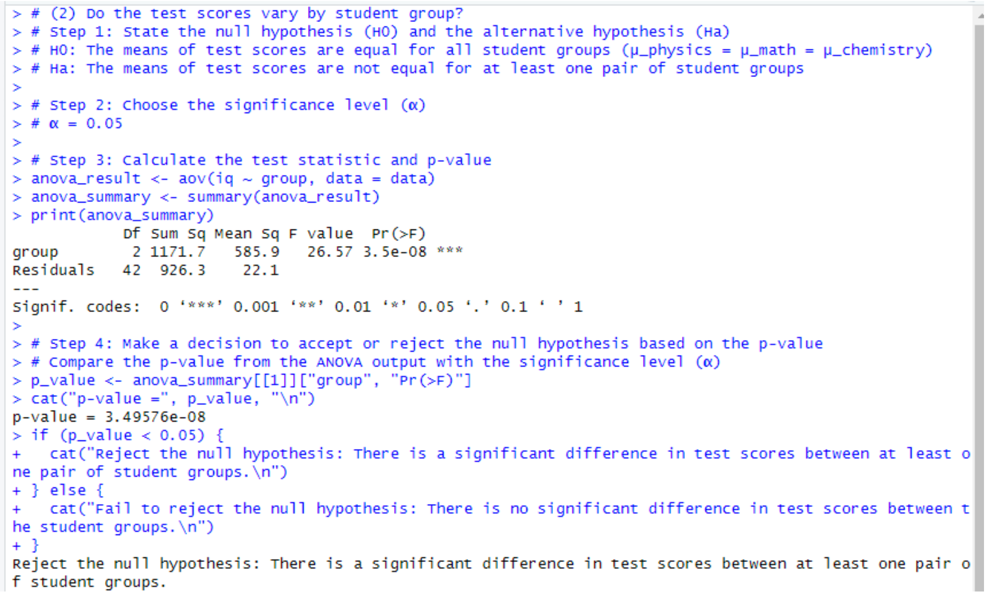


```
>
> # (1) How many students are in each group?
> count_by_group <- data %>%
+   group_by(group) %>%
+   summarize(n = n())
> print(count_by_group)
# A tibble: 3 x 2
  group                 n
  <chr>             <int>
1 Chemistry student    15
2 Math student         15
3 Physics student      15
>
> # Summarize the data relating to both test score and age by the student group
> summary_by_group <- data %>%
+   group_by(group) %>%
+   summarize(mean_iq = mean(iq), sd_iq = sd(iq), mean_age = mean(age), sd_age = sd(age))
> print(summary_by_group)
# A tibble: 3 x 5
  group             mean_iq sd_iq mean_age sd_age
  <chr>               <dbl> <dbl>    <dbl>  <dbl>
1 Chemistry student    46.3  3.73     40.1   4.22
2 Math student         37.6  5.53     20.7   2.99
3 Physics student      34.1  4.66     17.1   1.85
>
> # Graphical summaries
> ggplot(data, aes(x = group, y = iq, fill = group)) + geom_boxplot() + theme_minimal()
> ggplot(data, aes(x = group, y = age, fill = group)) + geom_boxplot() + theme_minimal()
> 
```

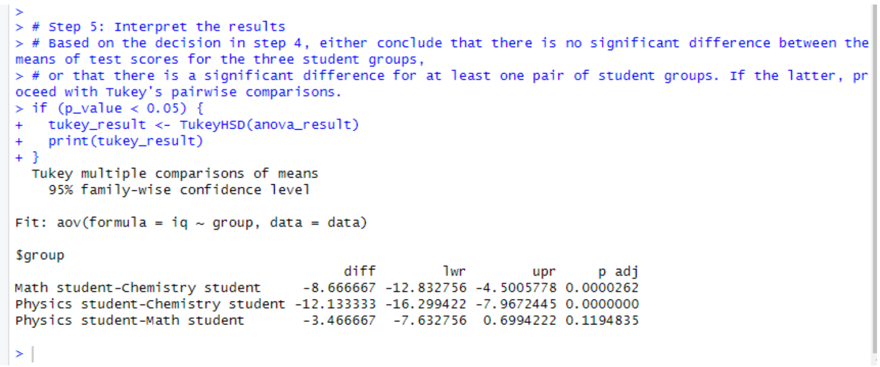




(2) Do the test scores vary by student group? Perform a one way ANOVA using the aov or Anova function in R to assess. Use a significance level of α=0.05. Summarize the results using the 5- step procedure. If the results of the overall model are significant, perform the appropriate pairwise comparisons using Tukey's procedure to adjust for multiple comparisons and summarize these results. (7 points)



```
> # (2) Do the test scores vary by student group?
> # Step 1: State the null hypothesis (H0) and the alternative hypothesis (Ha)
> # H0: The means of test scores are equal for all student groups (μ_physics = μ_math = μ_chemistry)
> # Ha: The means of test scores are not equal for at least one pair of student groups
>
> # Step 2: Choose the significance level (α)
> # α = 0.05
>
> # Step 3: Calculate the test statistic and p-value
> anova_result <- aov(iq ~ group, data = data)
> anova_summary <- summary(anova_result)
> print(anova_summary)
            Df Sum Sq Mean Sq F value  Pr(>F)
group        2 1171.7   585.9   26.57 3.5e-08 ***
Residuals   42  926.3    22.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Step 4: Make a decision to accept or reject the null hypothesis based on the p-value
> # Compare the p-value from the ANOVA output with the significance level (α)
> p_value <- anova_summary[[1]]["group", "Pr(>F)"]
> cat("p-value =", p_value, "\n")
p-value = 3.49576e-08
> if (p_value < 0.05) {
+   cat("Reject the null hypothesis: There is a significant difference in test scores between at least o
ne pair of student groups.\n")
+ } else {
+   cat("Fail to reject the null hypothesis: There is no significant difference in test scores between t
he student groups.\n")
+ }
Reject the null hypothesis: There is a significant difference in test scores between at least one pair o
f student groups.
```



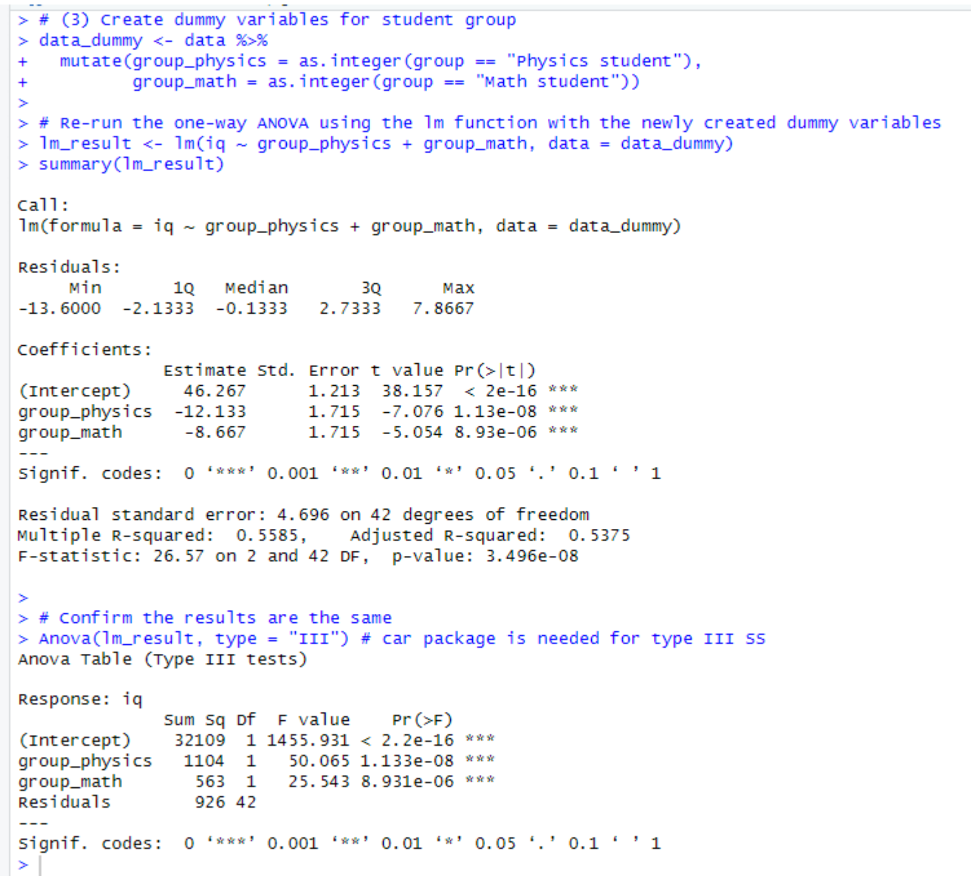
```
>
> # Step 5: Interpret the results
> # Based on the decision in step 4, either conclude that there is no significant difference between the
means of test scores for the three student groups,
> # or that there is a significant difference for at least one pair of student groups. If the latter, pr
oceed with Tukey's pairwise comparisons.
> if (p_value < 0.05) {
+   tukey_result <- TukeyHSD(anova_result)
+   print(tukey_result)
+ }
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = iq ~ group, data = data)

$group
                                       diff        lwr        upr     p adj
Math student-Chemistry student     -8.666667 -12.832756 -4.5005778 0.0000262
Physics student-Chemistry student -12.133333 -16.299422 -7.9672445 0.0000000
Physics student-Math student       -3.466667  -7.632756  0.6994222 0.1194835

> |
```

(3) Create an appropriate number of dummy variables for student group and re-run the one-way ANOVA using the lm function with the newly created dummy variables. Set chemistry students as the reference group. Confirm the results are the same (specifically point out test statistics, pvalues, etc. that show the results are equivalent). What is the interpretation of the beta estimates from the regression model? (4 points)



```
> # (3) Create dummy variables for student group
> data_dummy <- data %>%
+   mutate(group_physics = as.integer(group == "Physics student"),
+          group_math = as.integer(group == "Math student"))
>
> # Re-run the one-way ANOVA using the lm function with the newly created dummy variables
> lm_result <- lm(iq ~ group_physics + group_math, data = data_dummy)
> summary(lm_result)

Call:
lm(formula = iq ~ group_physics + group_math, data = data_dummy)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.6000  -2.1333  -0.1333   2.7333   7.8667 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     46.267      1.213  38.157  < 2e-16 ***
group_physics  -12.133      1.715  -7.076 1.13e-08 ***
group_math      -8.667      1.715  -5.054 8.93e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.696 on 42 degrees of freedom
Multiple R-squared:  0.5585,	Adjusted R-squared:  0.5375 
F-statistic: 26.57 on 2 and 42 DF,  p-value: 3.496e-08

>
> # Confirm the results are the same
> Anova(lm_result, type = "III") # car package is needed for type III SS
Anova Table (Type III tests)

Response: iq
              Sum Sq Df  F value    Pr(>F)    
(Intercept)    32109  1 1455.931 < 2.2e-16 ***
group_physics   1104  1   50.065 1.133e-08 ***
group_math       563  1   25.543 8.931e-06 ***
Residuals        926 42                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> |
```

Intercept: The intercept term is 46.267, which represents the average IQ score for the reference group (the group that is not included in the model as a dummy variable). In this case, the reference group is neither Physics nor Math students.

group_physics: The beta estimate for the Physics student group is -12.133. This indicates that, on average, Physics students have an IQ score that is 12.133 points lower than the reference group (neither Physics nor Math students). This difference is statistically significant with a p-value of 1.13e-08 (less than 0.001), denoted by '***'.

group_math: The beta estimate for the Math student group is -8.667. This means that, on average, Math students have an IQ score that is 8.667 points lower than the reference group (neither Physics nor Math students). This difference is also statistically significant with a p-value of 8.93e-06 (less than 0.001), denoted by '***'.

The model's adjusted R-squared value is 0.5375, which means that about 53.75% of the variability in IQ scores can be explained by the student group variable (Physics and Math students).

The F-statistic (26.57) and the corresponding p-value (3.496e-08) indicate that there is a significant relationship between the student group and IQ scores overall.

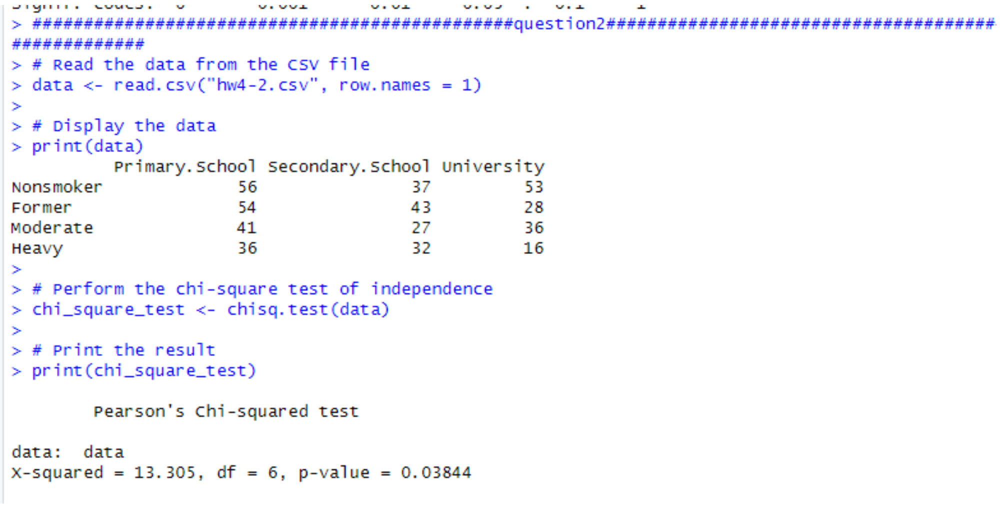
Question 2/2. In the United States, there is a strong relationship between education and smoking: well-educated people are less likely to smoke. Does a similar relationship hold in France? To find out, researchers recorded the level of education and smoking status of a random sample of 459 French men aged 20 to 60 years. The two-way table below displays the data.

#Yes, the results suggest that a similar relationship holds in France. The p-value obtained from the Pearson's Chi-squared test (0.03844) is less than the significance level of 0.05, indicating that there is a significant association between smoking status and educational level among French men aged 20 to 60 years.

Is there convincing evidence of an association between smoking status and educational level among French men aged 20 to 60 years? (Identify the right type of chi-square test, explain why, and check the conditions. State an appropriate pair of hypotheses to test in this setting and carry out the test by calculating the test statistic, degree of freedom, and P-value. What conclusion would you draw at α= 0.05?)

#Yes, there is convincing evidence of an association between smoking status and educational level among French men aged 20 to 60 years. The p-value of 0.03844 from the Pearson's Chi-squared test is less than the significance level of 0.05, which indicates a significant association between the two variables. This implies that the relationship between education and smoking status in France may be similar to the one observed in the United States, where well-educated people are less likely to smoke.

The code is below:



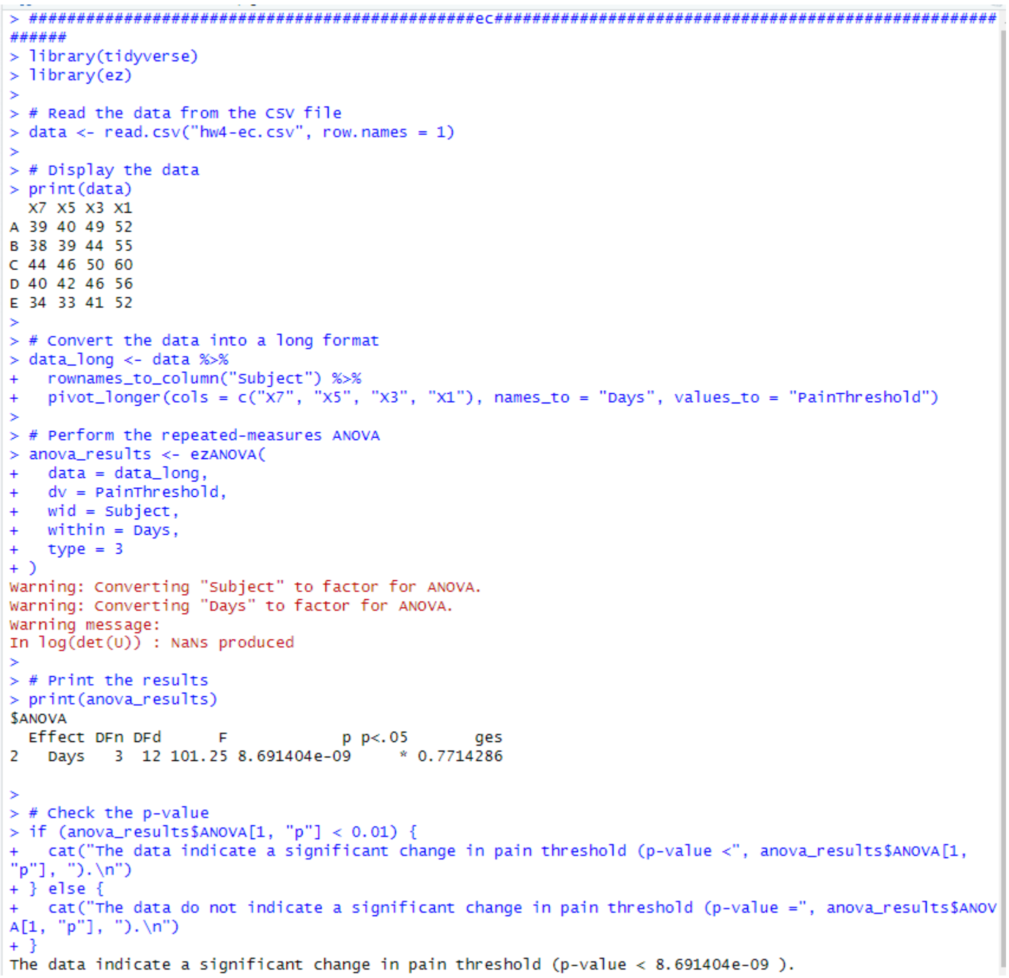
```
Signif. codes:  0        0.001      0.01      0.05  .  0.1     1
> ###############################################question2######################################
#############
> # Read the data from the csv file
> data <- read.csv("hw4-2.csv", row.names = 1)
>
> # Display the data
> print(data)
          Primary.School Secondary.School University
Nonsmoker             56               37         53
Former                54               43         28
Moderate              41               27         36
Heavy                 36               32         16
>
> # Perform the chi-square test of independence
> chi_square_test <- chisq.test(data)
>
> # Print the result
> print(chi_square_test)

        Pearson's Chi-squared test

data:  data
X-squared = 13.305, df = 6, p-value = 0.03844
```

The endorphins released by the brain act as natural painkillers. For example, a study monitored endorphin activity and pain thresholds in pregnant rats during the days before they gave birth. The data showed an increase in pain threshold as the pregnancy progressed. The change was gradual until 1 or 2 days before birth, at which point there was an abrupt increase in pain threshold. Apparently, a natural painkilling mechanism was preparing the animals for the stress of giving birth. The following data represent pain-threshold scores. Do these data indicate a significant change in pain threshold? Use a repeated-measures ANOVA with α =.01

X7 means 7 in the table
X5 means 5 in the table
X3 means 3 in the table
X1 means 1 in the table



```
> ###############################################ec#####################################################
######
> library(tidyverse)
> library(ez)
>
> # Read the data from the CSV file
> data <- read.csv("hw4-ec.csv", row.names = 1)
>
> # Display the data
> print(data)
  X7 X5 X3 X1
A 39 40 49 52
B 38 39 44 55
C 44 46 50 60
D 40 42 46 56
E 34 33 41 52
>
> # Convert the data into a long format
> data_long <- data %>%
+   rownames_to_column("Subject") %>%
+   pivot_longer(cols = c("X7", "X5", "X3", "X1"), names_to = "Days", values_to = "PainThreshold")
>
> # Perform the repeated-measures ANOVA
> anova_results <- ezANOVA(
+   data = data_long,
+   dv = PainThreshold,
+   wid = Subject,
+   within = Days,
+   type = 3
+ )
Warning: Converting "Subject" to factor for ANOVA.
Warning: Converting "Days" to factor for ANOVA.
Warning message:
In log(det(U)) : NaNs produced
>
> # Print the results
> print(anova_results)
$ANOVA
  Effect DFn DFd      F            p p<.05       ges
2   Days   3  12 101.25 8.691404e-09     * 0.7714286

>
> # Check the p-value
> if (anova_results$ANOVA[1, "p"] < 0.01) {
+   cat("The data indicate a significant change in pain threshold (p-value <", anova_results$ANOVA[1,
"p"], ").\n")
+ } else {
+   cat("The data do not indicate a significant change in pain threshold (p-value =", anova_results$ANOV
A[1, "p"], ").\n")
+ }
The data indicate a significant change in pain threshold (p-value < 8.691404e-09 ).
```