2. Load the tidyverse library.

```
library(tidyverse)

js <- read_csv('job_satisfaction1.csv')

## Rows: 106 Columns: 2

## ── Column specification
────────────────────────────────────────────────────────

## Delimiter: ","
## chr (1): education_level
## dbl (1): score

##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

set.seed(31878039)

my_js <- js %>%
  sample_n(104) %>%
  mutate(education_level <- factor(education_level,
 levels = c("school", "college", "university")
))
```

3.Use R’s summarise function (also works as summarize) to get a summary of the number of cases and mean (average) score by education level.

```
my_js %>%
 group_by(education_level) %>%
 summarise(
 count = n(),
 mean_score = mean(score)
 )

## # A tibble: 3 × 3
##   education_level count mean_score
##   <chr>           <int>      <dbl>
## 1 college            35       6.33
## 2 school             33       5.53
## 3 university         36       8.53

summary(my_js)

##  education_level        score             ... <- NULL
##  Length:104         Min.   : 4.480   school    :33
##  Class :character   1st Qu.: 5.772   college   :35
##  Mode  :character   Median : 6.445   university:36
##                       Mean   : 6.837
##                       3rd Qu.: 7.910
##                       Max.   :10.000
```

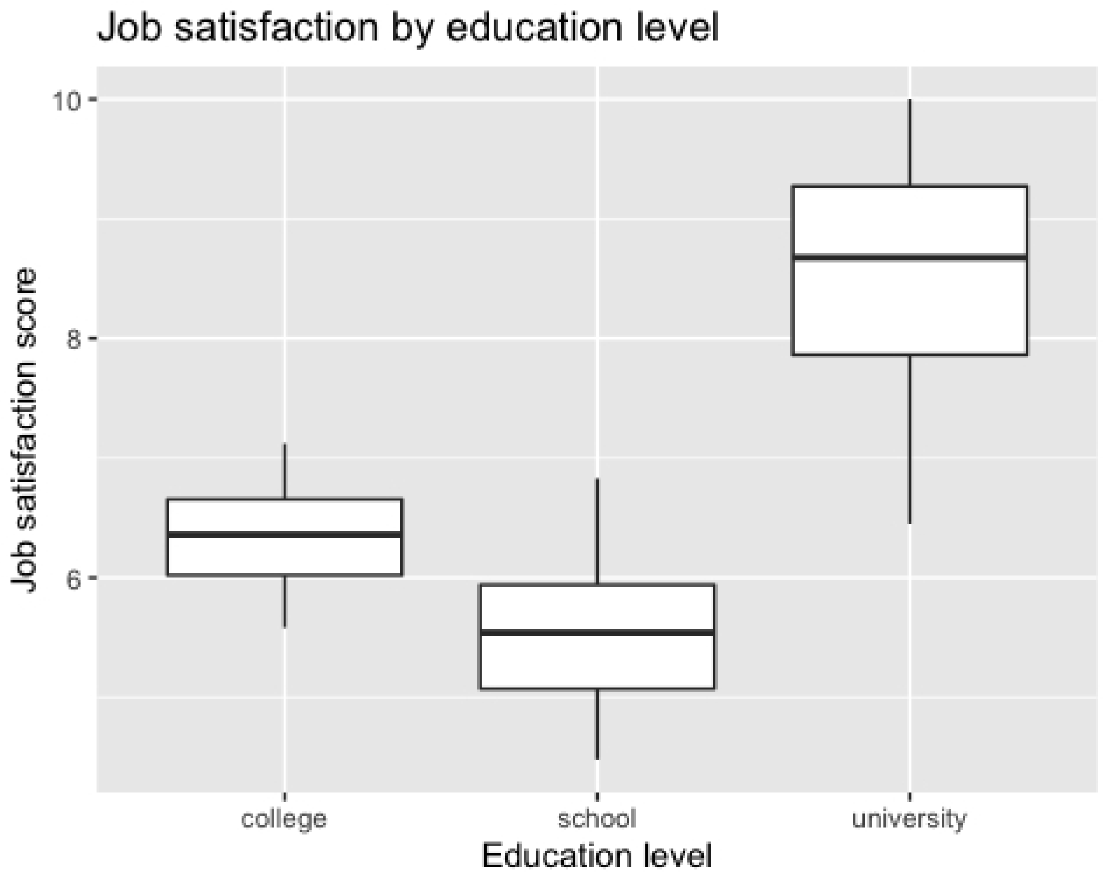
college mean score = 6.326857

school mean score = 5.527273

university mean score = 8.532222

4. Make a boxplot of score by education level using ggplot.

```
my_js %>% ggplot(aes(x = education_level, y = score)) +
 geom_boxplot() +
 labs(
 title = "Job satisfaction by education level",
 x = "Education level",
 y = "Job satisfaction score"
 )
```



5.Fit a model to predict score using education_level. Use summary to display the model summary. Explain what the coefficient values in the model tell you and relate these coefficient values to the summary of the data.

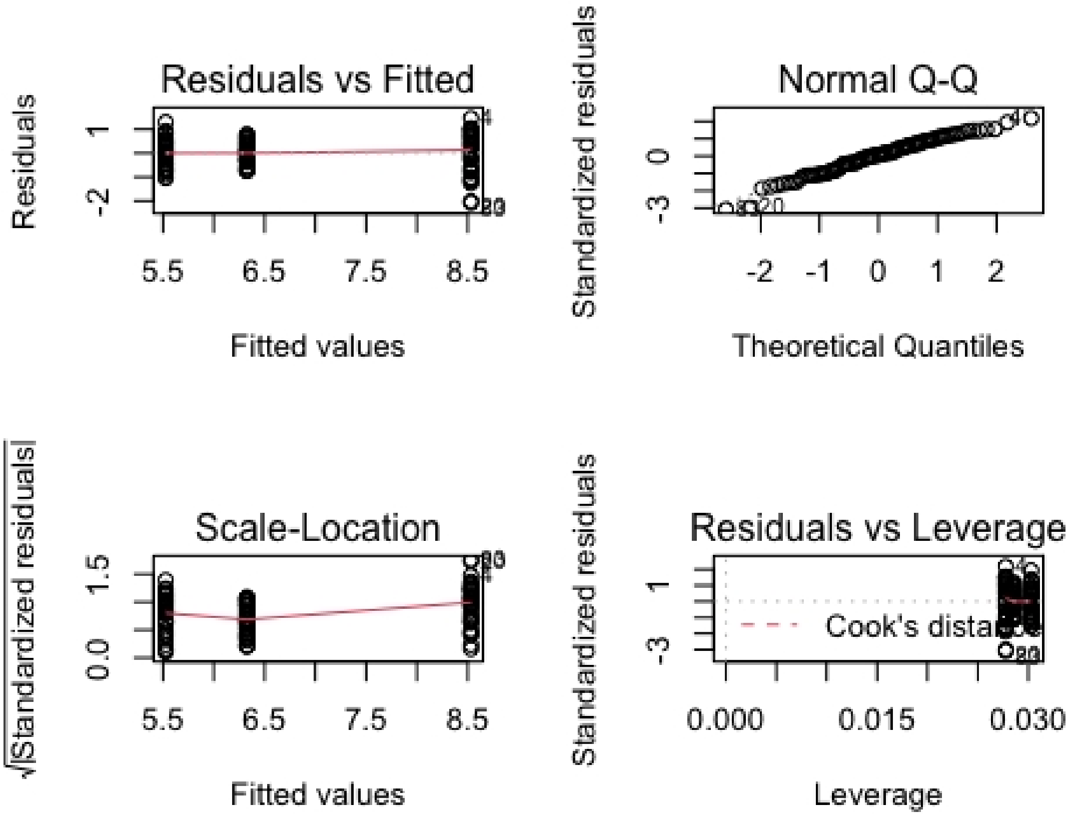
```
## Call:
## lm(formula = score ~ education_level, data = my_js)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08222 -0.45727  0.03546  0.48283  1.46778
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 6.3269     0.1154  54.810  < 2e-16 ***
## education_levelschool      -0.7996     0.1657  -4.825 4.96e-06 ***
## education_leveluniversity   2.2054     0.1621  13.604  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6829 on 101 degrees of freedom
## Multiple R-squared:  0.7822, Adjusted R-squared:  0.7779
## F-statistic: 181.4 on 2 and 101 DF,  p-value: < 2.2e-16
```

The baseline coefficient (education_levelcollege) of 6.3269 implies the sample mean for college score. The two remaining coefficients are the difference from the baseline. The coefficient (education_levelschool) of 6.3269 - 0.7996 = 5.5273 implies the sample mean for school score. The coefficient (education_leveluniversity) of 6.3269 + 2.2054 = 8.5323 implies the sample mean for university score.

These coefficient values are the same as part [3] as both parts show the mean score for each education level.

6.Create at the usual 4 regression diagnostic plots. Include the plots in your report. Explain what assumptions the Scale-Location plot and Normal Q-Q plot give information about and whether these plots indicate anything unusual or wrong for this model.

```
par(mfrow = c(2,2))
plot(m1)
```



Scale-Location plot explains if the residuals follow a constant variance for all values of predictors. In Scale-Location, the residuals are constant and show no pattern so they follow the constant variance assumption.

Normal Q-Q plot explains if the residuals are normally distributed. In Normal Q-Q, the residuals are closer to the line with a slight deviation at the extremities, but they still follow the normality assumption.

Both plots do not indicate anything unusual or wrong for this model.

In Residuals vs Fitted, the residuals are random so they follow linearity.

In, Residuals vs Leverage, there is no influential points.

7.Use R’s anova function to display the analysis of variance for your model. Include the output in your assignment report. State what null and alternative hypotheses the F-test statistic and its associated p value shown in the `anova’ output are testing, in the context of the variables used in this model. Give your conclusion based on this F-test and state what evidence that conclusion is based on.

```
anova(m1)
## Analysis of Variance Table
##
## Response: score
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## education_level   2 169.172  84.586  181.38 < 2.2e-16 ***
## Residuals       101  47.102   0.466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null: No difference in population mean score between college, school and university. Alternative: There is a difference in population mean score between college, school and university. The p-value for the education_level F-test statistic is pr(F > 181.38) = < 2.2e-16.

There is sufficient evidence that there is a difference in population mean score between college, school and university.

8. How many pair-wise comparison tests would you need to make if you tested each group in the education_level factor compared to every other group? Show how the overall Type 1 error rate is calculated if a significance level of 0.05 is used for each test and give the result of this calculation.

Would need to do (3(3-1)/2) = 3 pair_comparisons.

The overall Type 1 error rate is (1 – (1 – 0.05)^3) = 0.14.

9. Use the TukeyHSD function to get adjusted pairwise confidence intervals from this model. Show the output in your report. Explain in about 50 words what the confidence intervals shown in the TukeyHSD function output are confidence intervals for. Explain in about 30 words whether the TukeyHSD confidence intervals show evidence of a difference between any of the education_level group population means and if so, which.

```
TukeyHSD(aov(m1))

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = m1)
```

```
##
## $education_level
##                             diff       lwr        upr    p adj
## school-college     -0.7995844 -1.193746 -0.4054227 1.47e-05
## university-college  2.2053651  1.819749  2.5909810 0.00e+00
## university-school   3.0049495  2.613453  3.3964455 0.00e+00
```

The population mean score of school is lower than that of college by between -1.193746 and -0.4054227. The population mean score of university is higher than that of college by between 1.819749 and 2.5909810. The population mean score of university is higher than that of school by between 2.613453 and 3.3964455.

With 95% confidence, there is an evidence that the population mean score for all comparisons (school-college, university-college, university-school) are different as their intervals do not contain zero.

10. For this and the following questions you will use a subset of the PlantGrowth dataset included in R. This dataset gives results from an experiment to compare yields (as measured by dried weight of plants), variable weight obtained under a control and two different treatment conditions, variable group.

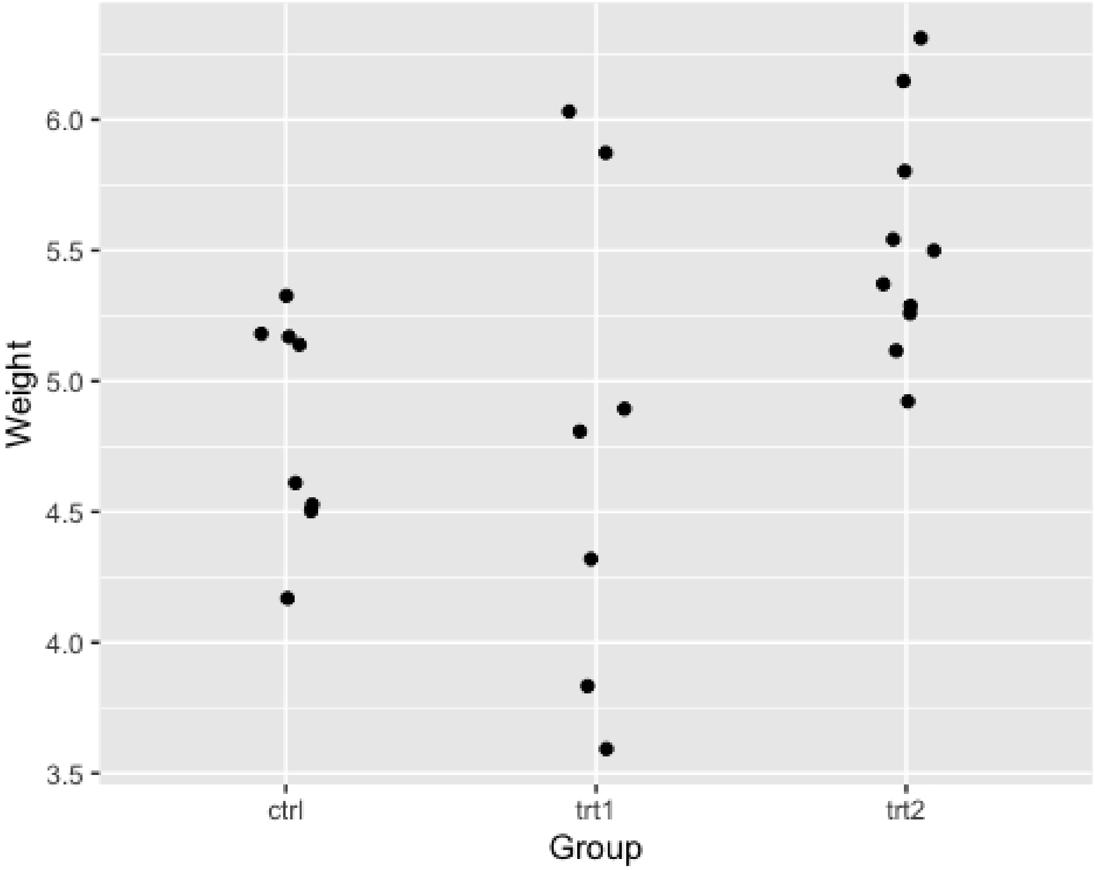
```
set.seed(31878039)
my_plants <- PlantGrowth %>% sample_n(25)
```

11. Create a summary of the counts and mean weights per group. You can use and adapt the code from part [3] above. Also make a plot of the weights for each group in your data. In this case there are not many observations in each group and a boxplot is not an ideal choice.

```
my_plants %>%
 group_by(group) %>%
 summarise(
 count = n(),
 mean_weights = mean(weight)
 )

## # A tibble: 3 × 3
##   group count mean_weights
##   <fct> <int>        <dbl>
## 1 ctrl      8         4.83
## 2 trt1      7         4.76
## 3 trt2     10         5.53

my_plants %>% ggplot(aes(x = group, y = weight)) +
 geom_jitter(width = 0.1) +
 labs(
 x = "Group",
 y = "Weight"
 )
```



I think that there are no differences in the population mean weight by group as they all overlap with each other based on the plot.

However, there could be a difference between in population mean weight between control and treatment1 based on their means, but this needs more statistical investigation to confirm.

12. Create a regression model to predict weight using group. Then use the TukeyHSD function to get adjusted pairwise confidence intervals for this model. Show the summary output for the regression and the TukeyHSD output in your report. Comment in about 100 words on whether there is evidence of a difference between any of the pairs of group population mean weights and if so, which. You should explain clearly what values from the output you are using and how these justify your comments.

```
m2 <- lm(weight ~ group, data=my_plants)
TukeyHSD(aov(m2))

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
```

```
##
## Fit: aov(formula = m2)
##
## $group
##                  diff          lwr       upr     p adj
## trt1-ctrl -0.06589286 -0.865498930 0.7337132 0.9766776
## trt2-ctrl  0.69725000 -0.035601071 1.4301011 0.0641834
## trt2-trt1  0.76314286  0.001765585 1.5245201 0.0493996
```

With 95% confidence, there is an evidence that the population mean weight between treatment2 and treatment1 is different as their confidence interval of between 0.001765585 and 1.5245201 does not contain zero.

The population mean weight between treatment1 and control does not indicate a difference as their interval of -0.865498930 and 0.7337132 contain zero. This implies that zero is a possible difference between. Same case also apply for treatment2 and control with an interval of between -0.035601071 and 1.4301011.