# Assignment6

## Question1, 2

library(tidyverse)

jobs <- read.csv('job\_satisfaction1.csv')

jobs = jobs %>% mutate(education\_level = factor(education\_level,

levels = c("school", "college", "university")))

set.seed(17053777)

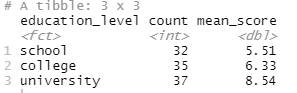
my\_js <- jobs %>% sample\_n(104)

## Question3

my\_js %>% group\_by(education\_level) %>%

summarise( count = n(),

mean\_score = mean(score) )



# mean score of school level is 5.51

# mean score of college level is 6.33

# mean score of university level is 8.54

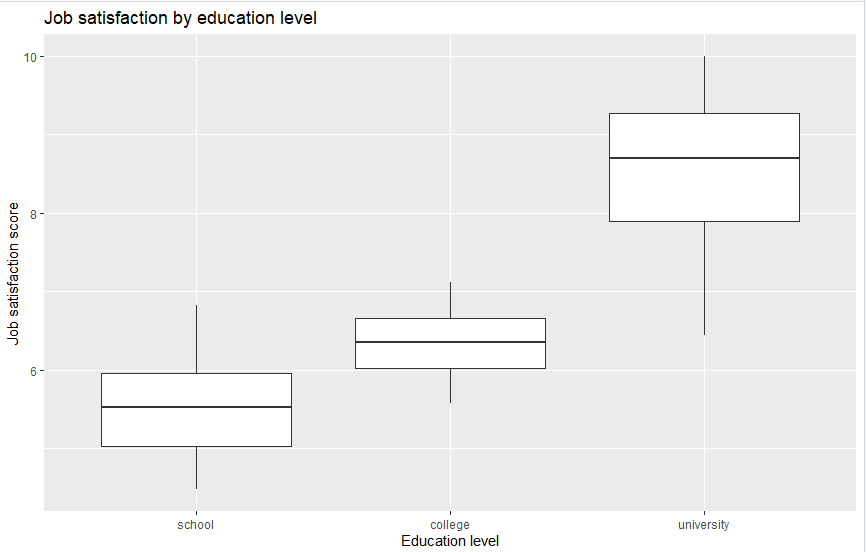
## Question4

my\_js %>% ggplot(aes(x = education\_level, y = score)) +

geom\_boxplot() +

labs( title = "Job satisfaction by education level",

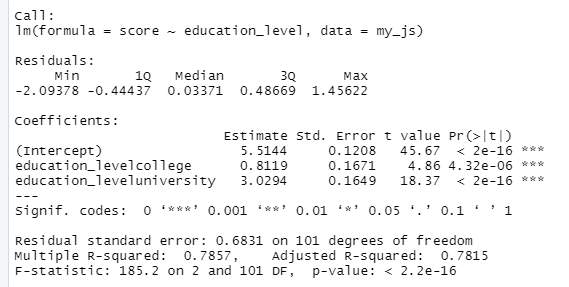
x = "Education level", y = "Job satisfaction score" )



## Question5

m1 <- lm(score ~ education\_level, data = my\_js)

summary(m1)



# the baseline of the score is school, as the mean score of school is 5.51, the intercept here is 5.5144. Here are the coefficient values of the score of college and the score of university.

# score of college= 5.5144 + 0.8119 \* education\_level\_college

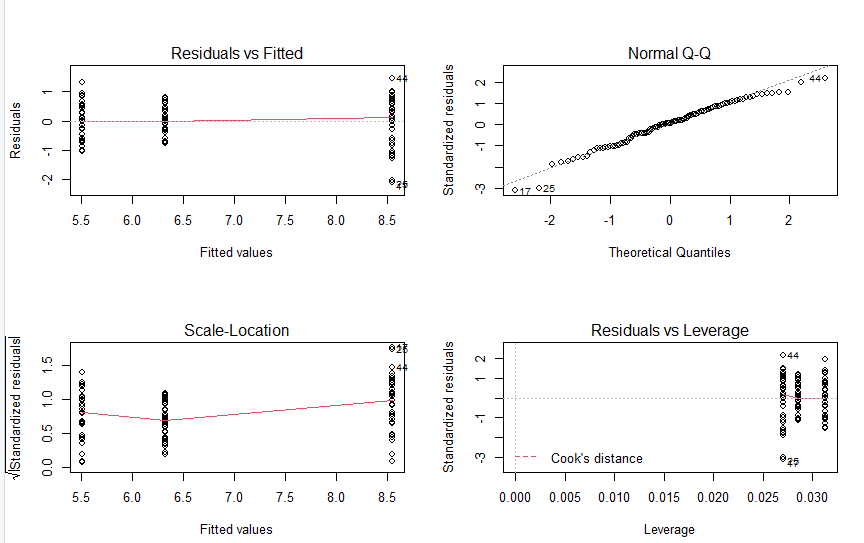
# score of university= 5.5144 + 3.0294 \* education\_level\_university

# From the equations above, we can see when the education level of college is 1, the score of college is 6.3263 which is around the mean score of college 6.33. When the education level of university is 1, the score of university is 8.5438 which is around the mean score of the university 8.54.

## Question6

par(mfrow = c(2, 2))

plot(m1)

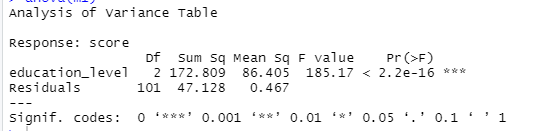


# From Normal Q-Q graph, although after 1 theoretical quantity, there are some deviation between residuals and the line, overall, we can observe that the data are generally normally distributed.

# From Scale-Location plot we can observe that the variance decreases at first then increases, which means the line is not that horizontal. And there seems to be a pattern that is not random, hence, the variance is not equal. It does not meet the equal variance assumption.

## Question7

anova(m1)



# we assume null hypotheses is there is no correlation between score and educational level and alternative hypotheses is there is correlation between score and educational level.

# we can see the p-value is much smaller than 0.05, so we reject null hypothesis.

# with p-value <= 0.05, we can say that there is no difference between the means and conclude that a significant difference does exit. There is a difference between the mean score across the different education level.

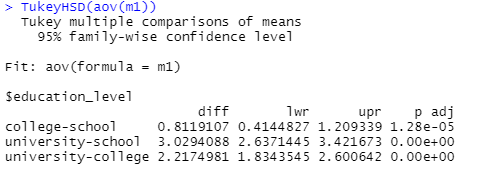
## Question8

# pr(at least 1 Type I error in 3 similar tests) = 1-(1-0.05) ^ 3 = 1-0.8574 = 0.1426

# we are making 3 comparisons as K=3, so using K(K-1)/2 gives us 3.

## Question9

TukeyHSD(aov(m1))



# The confidence intervals are for the differences in population mean scores. We can see 0 is not between intervals of lower and upper in each group pair. There is statistically significant evidence for a college-school, university-school and university-college difference in population mean score.

## Question10

data("PlantGrowth")

set.seed(17053777)

my\_plants <- PlantGrowth %>% sample\_n(25)

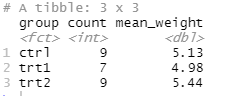
my\_plants

## Question 11

my\_plants %>% group\_by(group) %>%

summarise( count = n(),

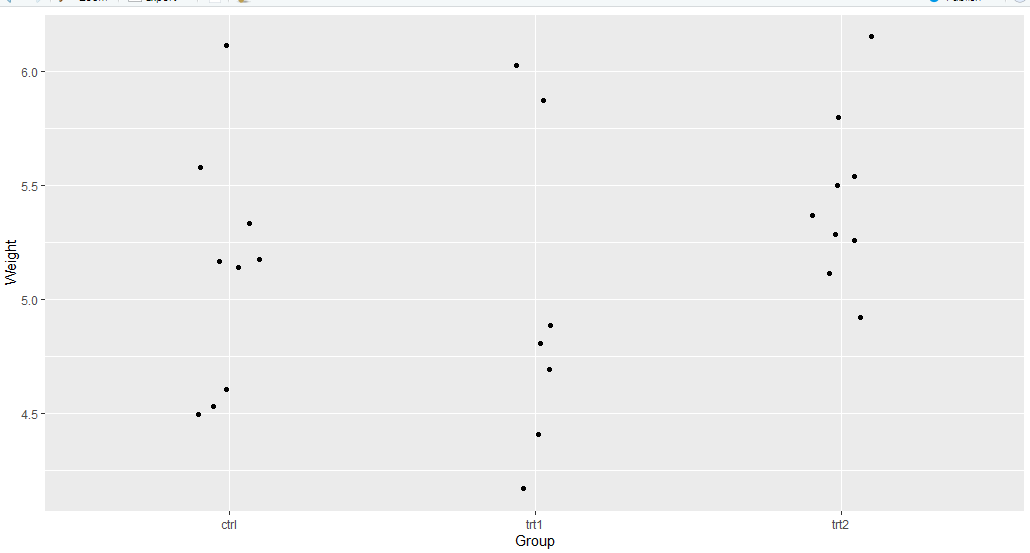
mean\_weight = mean(weight))



my\_plants %>% ggplot(aes(x = group, y = weight)) +

geom\_jitter(width = 0.1) + # width says how big the jitter is

labs( x = "Group", y = "Weight" )



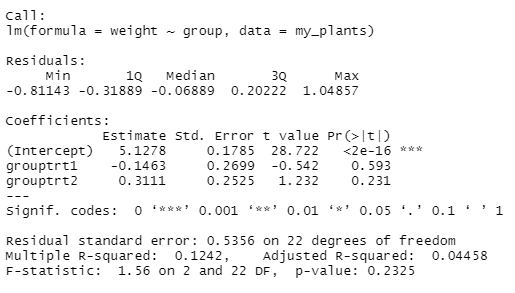
# Yes, there is a difference in population mean weights by group. As we can see the population weights in trt2 is much higher than ctr1 and trt1. The mean of trt2 is around 5.5, but the mean of ctr1 is 5.25, and the mean of trt1 is around 5.0.

# the actual difference isdelf and the size of the sample

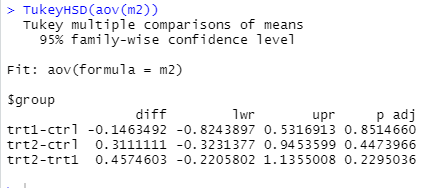
## Question12

m2 <- lm(weight ~ group, data = my\_plants)

summary(m2)



TukeyHSD(aov(m2))



# The confidence intervals are for the differences in population mean weight.

# For group trt1-ctr1, the interval of upper and lower is between -0.8243897 to 0.5316913 which contains 0.

# For group trt2-ctr1, the interval of upper and lower is between -0.3231377 to 0.9453599 which contains 0.

# For group trt2-trt1, the interval of upper and lower is between -0.2205802 to 1.1355008 which contains 0.

# There is no statistically significant evidence for trt1-ctr1, trt2-ctr1 and trt2-trt1 difference in population mean weight, since their intervals contain 0.