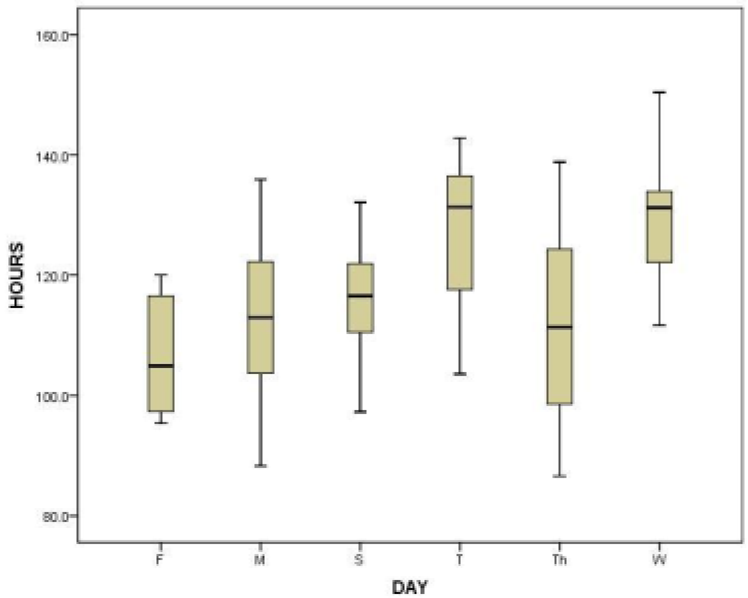
HW5

Name: Naga Satya Silpa Annadevara.

Student ID: A20517818.

let's use the "clerical_Q2.txt" data.
A store manager noticed that the busiest days for clerical staff are Wednesdays and Tuesdays. See enclosed box plot. The manage tries to compare the group means in hours by different days



a). [10] Observe the box plot. Can you confirm that the hours in Tuesday is the highest? Why?

ANSWER:

- From the boxplot, we can observe that there are 6 different groups with days variable on the x-axis and hours variable on the y-axis. For each group, we can observe a 1-box plot.
- From the visualization, we can clearly observe that the variation within the groups (boxplots) are not even/uniform. So, we cannot compare the groups based on their q2 values. Why? Because the within-group variation (IQR) is large. So, we cannot use q2 to compare the groups.
- So, we need to have additional information such as mean values for every group to compare them & give the conclusion.
- As the boxplot cannot provide mean values, we need to use ANOVA technique to say hours on Tuesday is the highest or not.
- So, with the variation difference among the groups & without knowing the mean values for every group, we cannot confirm that hours on Tuesday is the highest.

b). [20] Write down your hypothesis in the ANOVA to compare the group means in hours by different days

ANSWER:

Null hypothesis: All the groups have the same mean. In other words, the means of hours for different days are equal.

$H0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

- μ_1 = average of hours on Monday group
- μ_2 = average of hours on Tuesday group
- μ_3 = average of hours on Wednesday group
- μ_4 = average of hours on Thursday group
- μ_5 = average of hours on Friday group
- μ_6 = average of hours on Saturday group

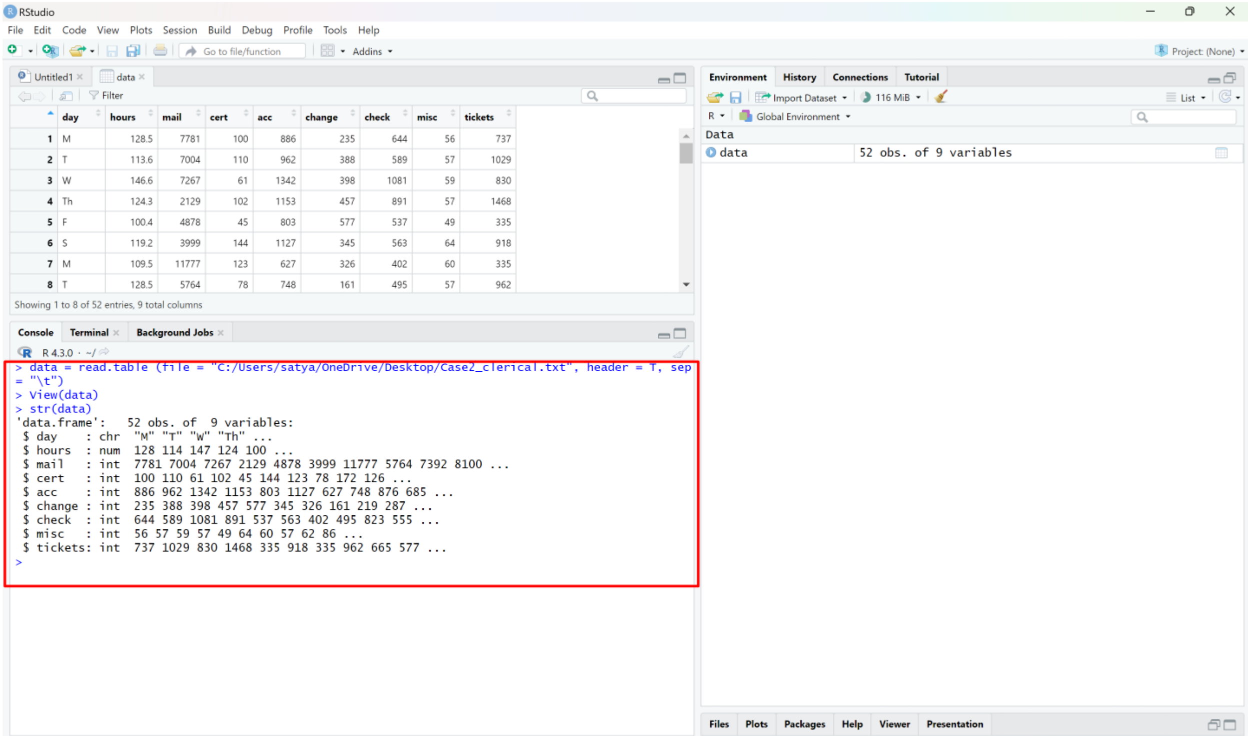
Alternative hypothesis: At least 2 groups (1-pair of the mean) have different means. In other words, at least 2 group's means among the means of hours for different days are not equal.

Ha = Not all the μ_t are equal / $\mu_i \neq \mu_j$

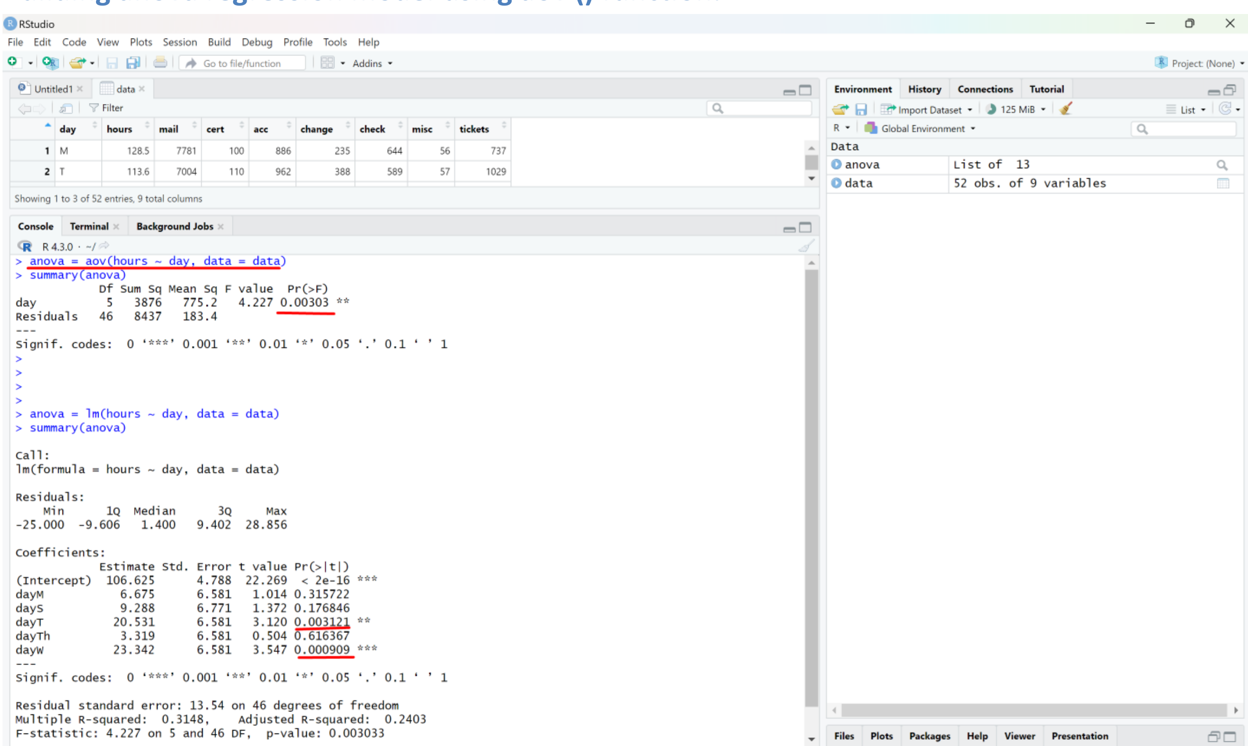
b). [30] Using R to build the ANOVA regression model and help the manager to make the decision whether the group means in hours or different days are the same or not.

ANSWER:

Importing clerical_Q2.txt data into the r-studio:



Building anova regression model using aov () function:



Decision/ conclusion:

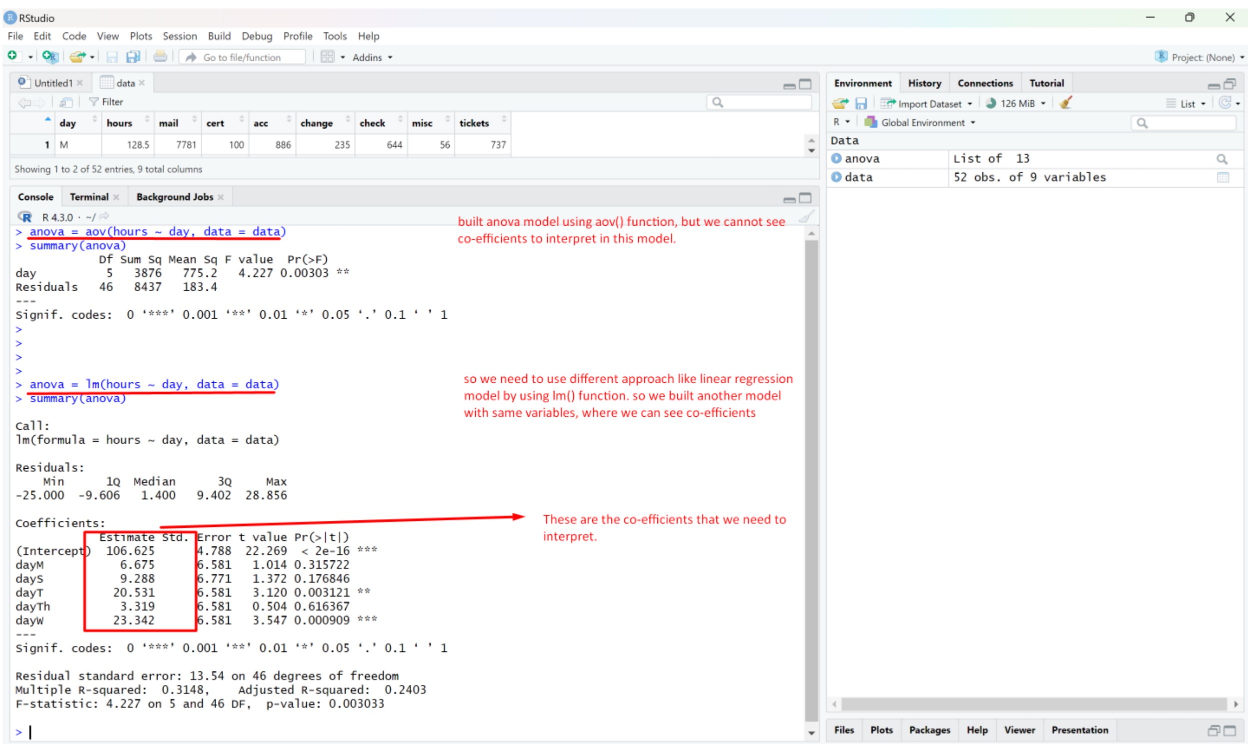
- Using 95% confidence level, as the p-value (0.00303) in F-test is smaller than alpha ($p < 0.05$%) we do have enough evidence to reject null hypothesis.
- In other words, we have enough evidence to reject that all group means in hours or different days are not same. At least 2 group means are different.

- To know which 2 groups means are different, we need to look at the t-test. We can clearly observe that Tuesday and Wednesday group means are not same.

c). [20] Try to interpret the coefficients you got in the ANOVA regression model from part b).

ANSWER:

We cannot see co-efficients in the model which we built using aov () function.
So we need to use other modeling approaches such as linear regression (e.g., lm())
So built model using lm() function with same variables to see the co-effeicients.



Interpretation of co-efficients: **as there is no dayFriday, we are going to use Friday as baseline**.

hours = $\beta 0 + \beta 1 * X1 + \beta 2 * X2 + \beta 3 * X3 + \beta 4 * X4 + \beta 5 * X5 + e$

- Intercept (+106.625): ~ $\beta 0$

The intercept represents the estimated mean value for the reference group (Friday) when all other dummy variables are zero. In this case, it suggests that the estimated mean value for Friday is 106.625.

- dayM (+6.675): β1………….. X1= dayM

  The coefficient for "dayM" (+6.675) represents the difference in mean between Monday and Friday. The positive coefficient indicates that, on average, the mean for Monday is 6.675 units higher than the mean for Friday.

- dayS (+9.288): β2………….. X2= dayS

  The coefficient for "dayS" (+9.288) represents the difference in mean between Saturday and Friday. The positive coefficient indicates that, on average, the mean for Saturday is 9.288 units higher than the mean for Friday.

- dayT (+20.531): β3…………… X3= dayT

  The coefficient for "dayT" (+20.531) represents the difference in mean between Tuesday and Friday. The positive coefficient suggests that, on average, the mean for Tuesday is 20.531 units higher than the mean for Friday.

- dayTh (+3.319): β4……………. X4= dayTh

  The coefficient for "dayTh" (+3.319) represents the difference in mean between Thursday and Friday. The positive coefficient indicates that, on average, the mean for Thursday is 3.319 units higher than the mean for Friday.

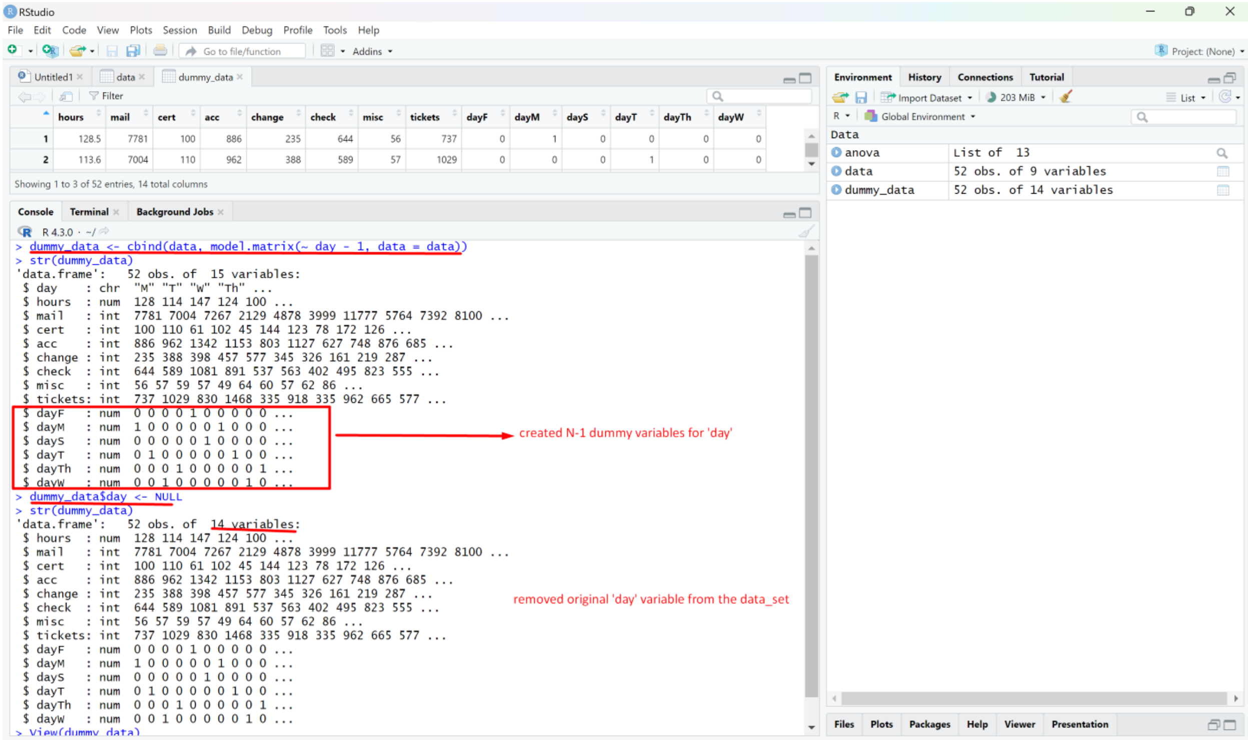
- dayW (+23.342): β5……………. X5= dayW

  The coefficient for "dayW" (+23.342) represents the difference in mean between Wednesday and Friday. The positive coefficient suggests that, on average, the mean for Wednesday is 23.342 units higher than the mean for Friday.

d). [20] practice for data preprocessing: create N-1 dummy variables for the variable ‘DAY’. Convert the variable “mail” to nominal variable by creating 4 groups. Again, paste the codes and snapshots.

ANSWER:

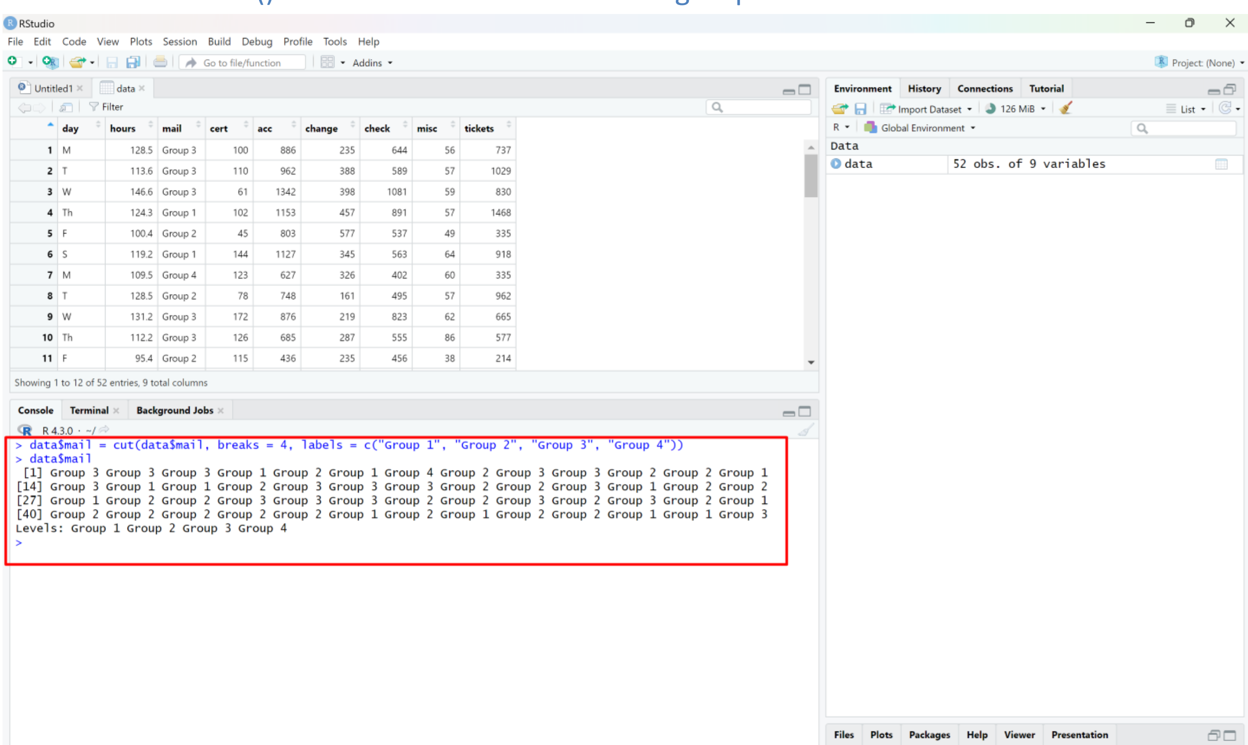
Created N-1 dummy variable for the variable “DAY”:



Note: For creating dummy variable we need to use model.matrix() function , As dummies package was removed from the latest version of R.

Convert the variable "mail" to nominal variable by creating 4 groups.:

We need to use cut() function to cut the mail into 4 groups:



Note: as the group values came into scientific notation, I used labels to improve the readability.