Assignment

Ian Timothy Henry Suarez - 47519843

Question 1

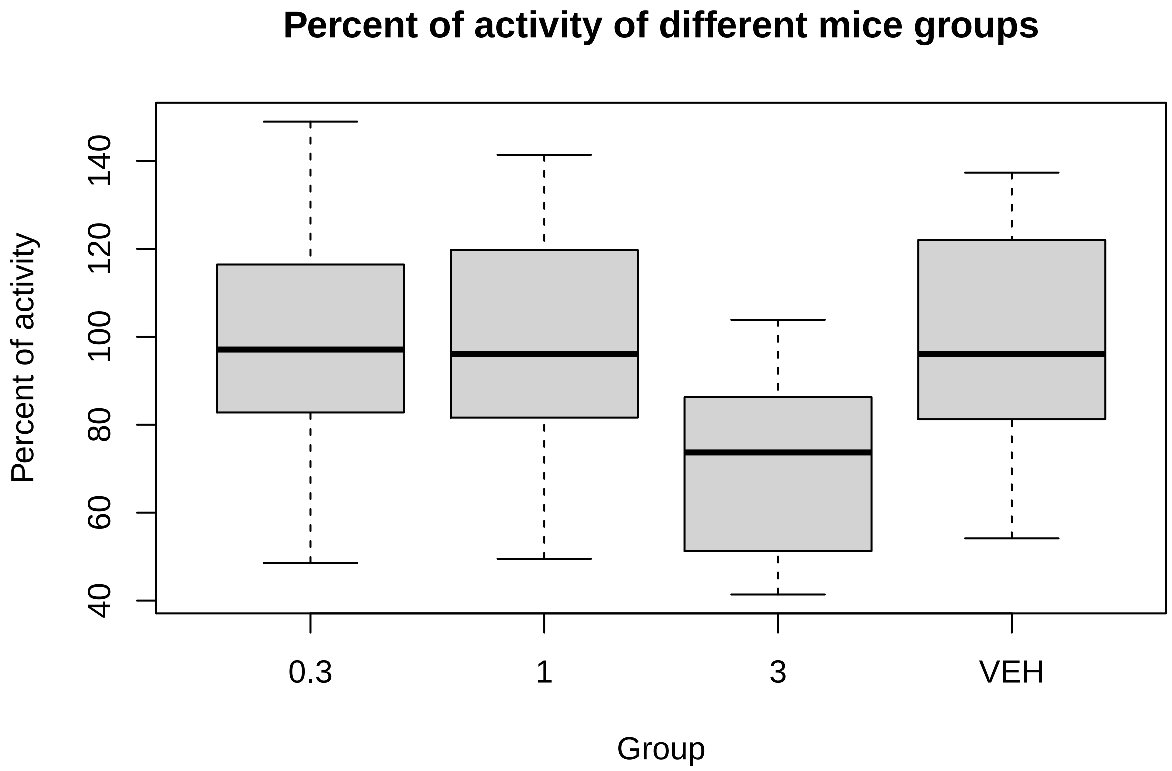
In a study of the effects of marijuana on mice, ordinary mice were divided in four groups. The first group of mice (VEH) received a vehicle only, which is a shot with the same inactive ingredients as the other mice but no THC (the active ingredient in marijuana). The three other groups of mice received different dosages of THC (0.3, 1 and 3 mg/kg respectively). The investigators then measured the locomotor activity of the mice by placing each mouse in a box lined with photocells to measure the total distance covered by that mouse, and the results are reported as the percentage movement relative to the untreated mice.

First, it is needed to import the data.

```
mice = read.table(here::here("data", "mice_pot.txt"), header=TRUE)
```

1. Plot a boxplot by group; specificially, this boxplot should contain four boxes, each of which corresponds to one of the four treatments (VEH, 0.3 mg/kg, 1 mg/kg and 3 mg/kg). Then comment on two interesting features of this boxplot.

```
boxplot(mice$percent_of_act ~ mice$group,
        main = "Percent of activity of different mice groups",
        ylab = "Percent of activity", xlab = "Group")
```



- This boxplot shows that the group of mice that received the treatment of 3 mg/kg of THC has significantly less activity than the other 3 groups.
- This boxplot shows that all 4 groups have a similar spread. This means that the 4 groups have a similar variance.

2. Assuming that all test assumptions are ok, test if there is any difference in activities across the four dosage groups.

- **Hypotheses:** $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$; H_1 : at least one mean is different
- **ANOVA table:**

```
mice.aov <- aov(percent_of_act ~ group, data = mice)
summary(mice.aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)  
## group        3   6329  2109.7   3.126 0.0357 *
## Residuals   42  28344   674.8                 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Test statistic:** $F_{obs} = 3.126$
- **Null distribution:** If H_0 is true, F_{obs} behaves like a distribution $F_{3,42}$

- **P-Value:** $P(F_{3,42} \geq 3.126) = 0.0357 < 0.05$
- **Conclusion:** Since the P-Value is less than the significance level of 5%, there is enough evidence to reject the null hypothesis H_0. This means that at least the mean of one group of mice is different from the rest of the groups.

3. If someone wants to find out whether there is any difference in average activity between the first three dosages (VEH, 0.3 mg/kg, 1 mg/kg) as a whole and the highest dosage (3 mg/kg), form an appropriate contrast to evaluate and test the difference. Remember to state null and alternative hypotheses, give the test statistic value and associated degrees of freedom, the evidence for rejecting or not rejecting the null hypothesis with a conclusion.

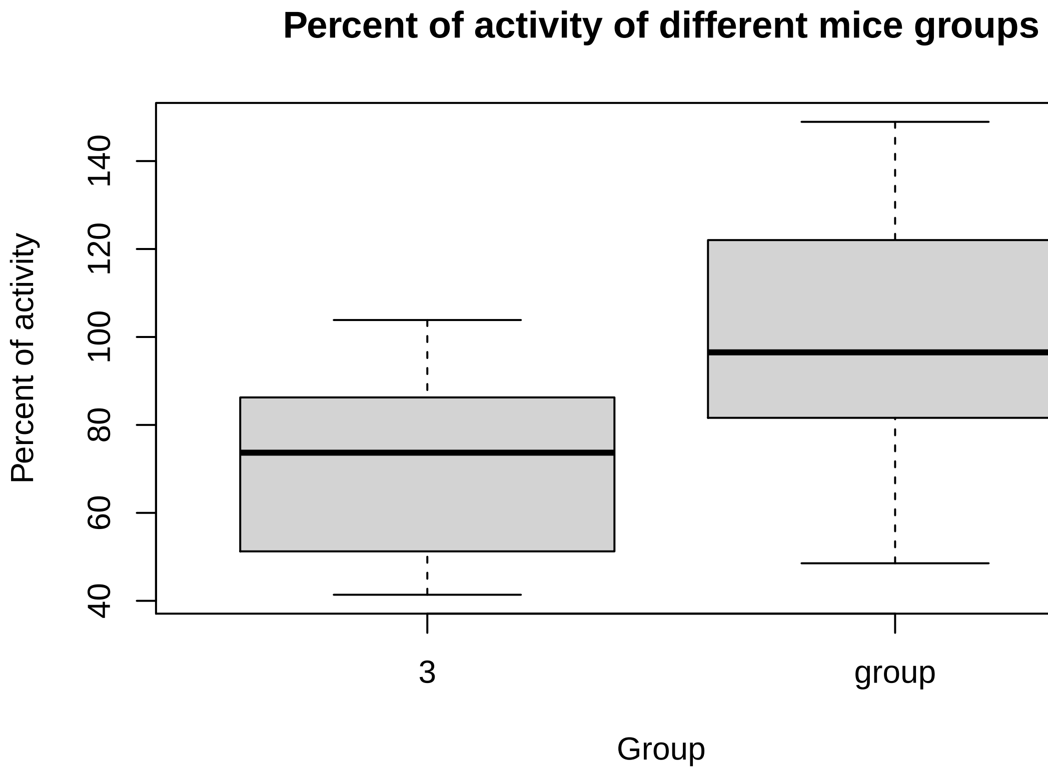
First, the data needs to be processed so the first three dosages (VEH, 0.3 mg/kg, 1 mg/kg) correspond to a single group. The values in the group column which correspond to 0.3 mg/kg and 1 mg/kg will be substituted with the value "group" so that they all correspond to the same group. This whole process will be implemented on a copy of the data

```r
mice2 <- mice
mice2$group[mice2$group == 0.3] <- "group"
mice2$group[mice2$group == 1] <- "group"
mice2$group[mice2$group == "VEH"] <- "group"
```

Now, there are only two groups in the data.

The next boxplot show the variability between the two groups

```r
boxplot(mice2$percent_of_act ~ mice2$group,
        main = "Percent of activity of different mice groups",
        ylab = "Percent of activity", xlab = "Group")
```



It can be seen that the variability between the two groups is similar. Therefore, we assume that they have equal variance.

- **Hypotheses:** $H_0 : \alpha_1 = \alpha_2$; $H_1 : \alpha_1 \neq \alpha_2$

To carry out this hypothesis test, there are two options; analysis of variance or two sample t-test. Both of these methods are equivalent when only two groups are being tested. There is a mathematical identity which states that $F_{1,v} = t_v^2$. The difference is that using a t-test allows for a one-sided test for the two groups.

Both methods will be implemented to be compared.

First, the t-test will be implemented.

```
t.test(mice2$percent_of_act ~ mice2$group, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  mice2$percent_of_act by mice2$group
## t = -3.1222, df = 44, p-value = 0.003169
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -46.64441 -10.04934
## sample estimates:
##    mean in group 3 mean in group group
##           70.66787            99.01474
```

In this case the test statistic value is $t_{obs} = -3.1222$, the associated degrees of freedom are $df = 44$ and the $P - Value = 0.003169$.

Then, the analysis of variance will be implemented.

```
summary(aov(mice2$percent_of_act ~ mice2$group))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## mice2$group  1   6289    6289   9.748 0.00317 **
## Residuals   44  28384     645                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case the test statistic value is $F_{obs} = 9.748$, the associated degrees of freedom are $df = 1, 44$ and the $P - Value = 0.00317$

After comparing both methods it can be verified that $t^2_{44} = F_{1,44}$. This is $(-3.1222)^2 = 9.748133$, which was the result for both methods.

Since the P-Value in both methods was equal, and less than the significance level of 5%, there is enough evidence to reject the null hypothesis H_0. This means that there is enough evidence to state that the mean of the two groups is different, $\alpha_1 \neq \alpha_2$.

Question 2

A car manufacturer wants to study the fuel efficiency of a new car engine. It wishes to account for any differences between the driver and production variation. The manufacturer randomly selects 5 cars from the production line and recruits 4 different test drivers.

```
kml = read.table(here::here("data", "kml.dat"), header=TRUE)
```

1. For this study, is the design balanced or unbalanced? Explain why.

For a balanced design, it is necessary to have the same number of samples for each pair of factors. This can be verified with the following table, which counts the samples for each pair of factors.

```
table(kml$driver, kml$car)
```

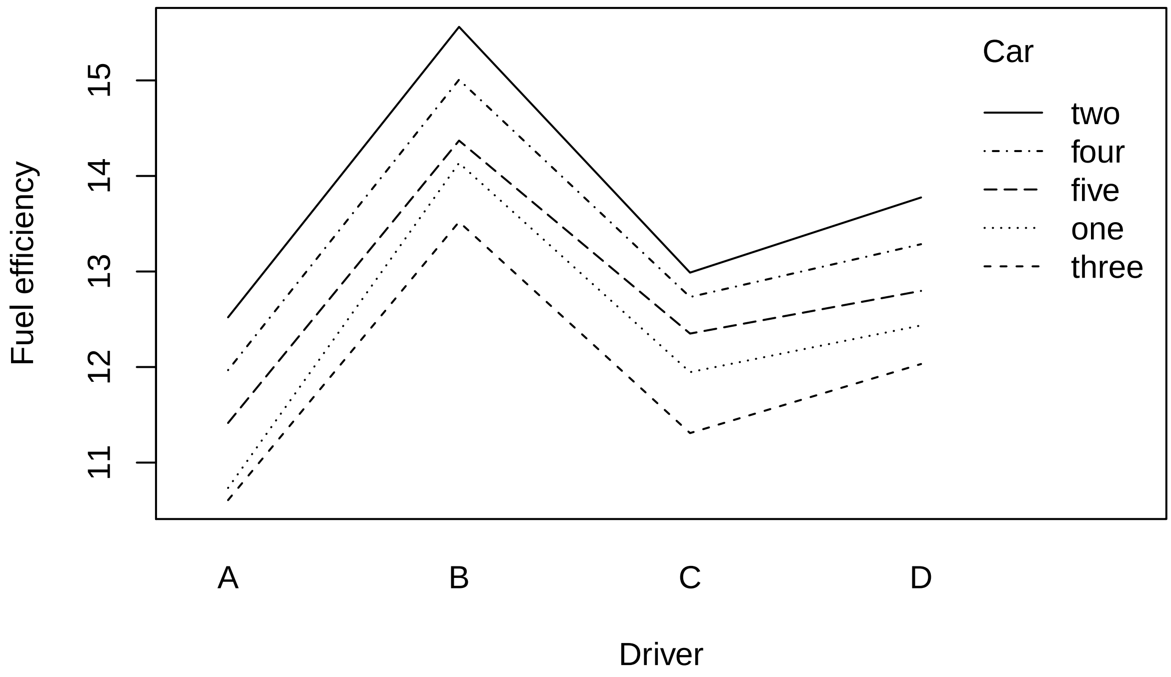
```
##    
##     five four one three two
##   A    2    2   2     2   2
##   B    2    2   2     2   2
##   C    2    2   2     2   2
##   D    2    2   2     2   2
```

This table shows that there are two records of each driver using each of the 5 different cars. This means that there are the same number of samples for each pair of factors, and therefore it is a balanced design.

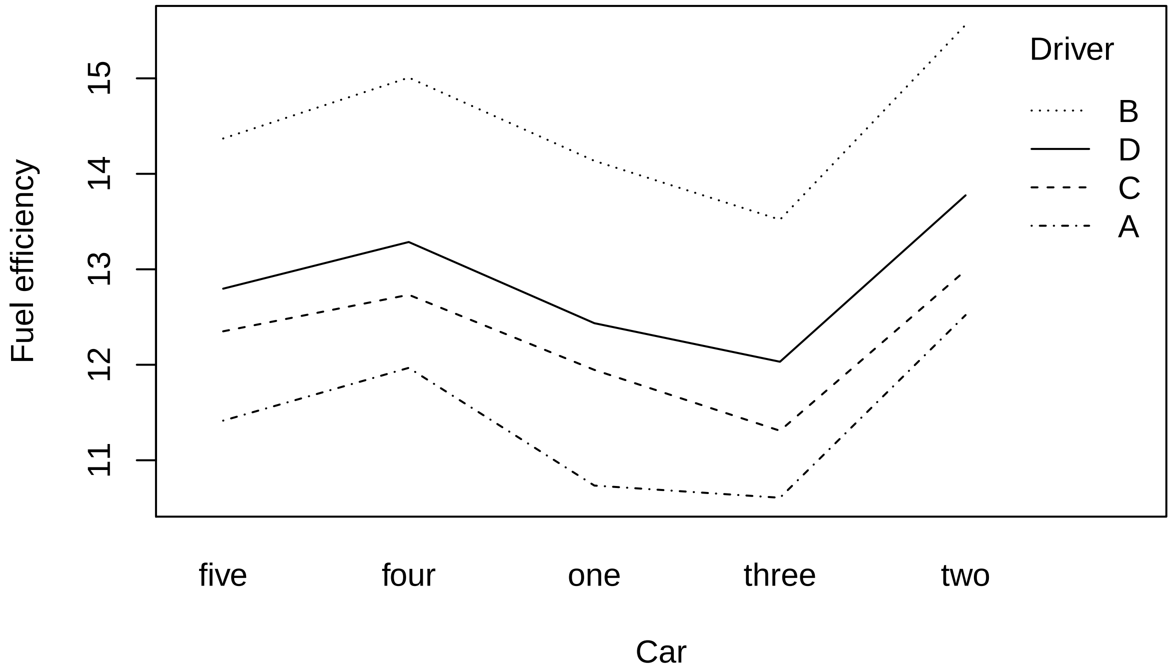
2. Construct two different preliminary graphs that investigate different features of the data and comment.

- Interaction plot: This graphs show how the response variable changes with different levels of the two different factors. It is used to review the interaction between the factors. If there is no interaction between the factors, the change in the response to changes in the first factor won't be affected by changes in the second factor. This can be seen if the slope of the response with respect of the first factor is the same across the second factor. In other words, the lines would be parallel if there is no interaction between the factors.

```r
interaction.plot(kml$driver, kml$car, kml$kmL, trace.label = "Car",
                 xlab = "Driver", ylab = "Fuel efficiency")
```



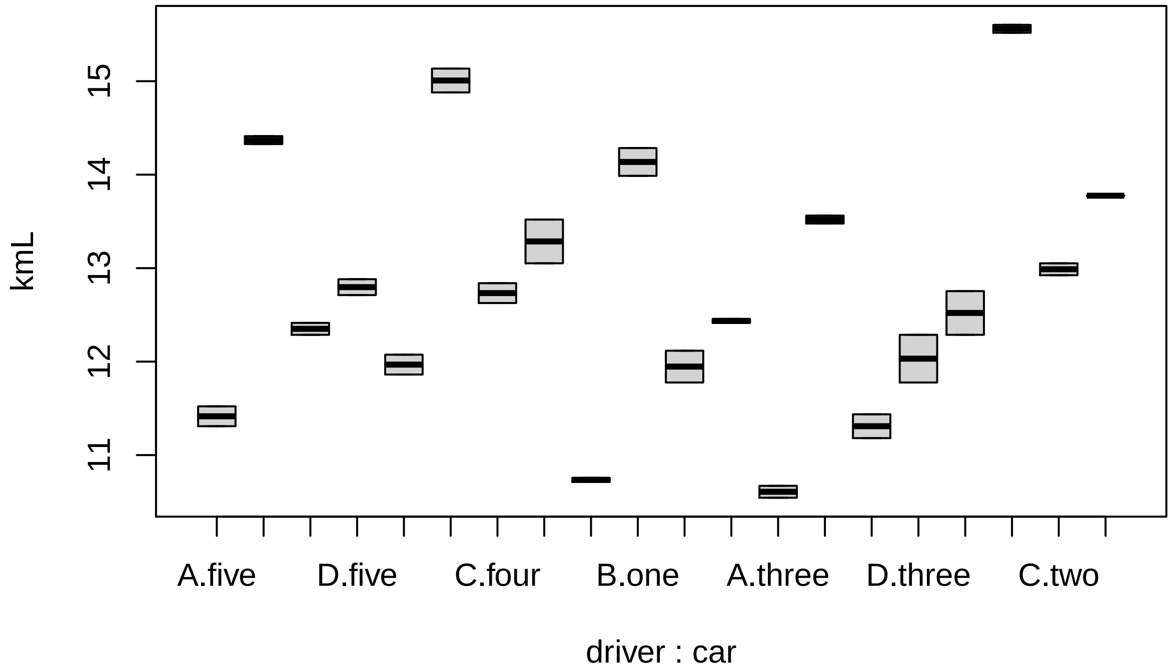
```r
interaction.plot(kml$car, kml$driver, kml$kmL, trace.label = "Driver",
                 xlab = "Car", ylab = "Fuel efficiency")
```



In the first graph, it can be observed that the slope of the response with respect of the first factor is very similar across the second factor. The most notorious change of the slope can be seen in the car 1 across the drivers. In the second graph, it can be observed that the slopes are very similar. The most notorious change can be seen in the driver C. In conclusion, these two graphs show that there is a very small interaction between the two factors.

- Boxplot: This graph gives information about the relative size of the effects on the response of each pair of factors. It shows the variability and the possible outliers in the observed data.

```
boxplot(kmL ~ driver + car, data = kml)
```



It is difficult to interpret this graph because there are 20 different interactions. It is also difficult to review the overall pattern and dynamics. Nevertheless, it can be observed in which pair of factors there is more variability. This can be seen in the wider boxes. The interaction of factors with more variability are driver D with car 4, driver D with car 3 and driver A with car 2.

3. Analyze the data, stating null and alternative hypothesis for each test, and check assumptions.

In this type of problem, three aspects can be tested: the effect of the interaction between the factors, the effect of the first factor and the effect of the second factor. The model for the response variable is the following

$$Y_{ijk} = \mu + \alpha_i + \beta j + \gamma ij + \epsilon ijk$$

where

- Y_{ijk} : kmL response
- μ : overall mean
- α_i corresponds to the effect of the factor driver for $i = 1, 2, 3, 4$
- βj corresponds to the effect of the factor car for $j = 1, 2, 3, 4, 5$
- γij corresponds to the effects of the interaction component between the two factors
- ϵijk corresponds to the random unexplained variation.

- **Hypotheses:**

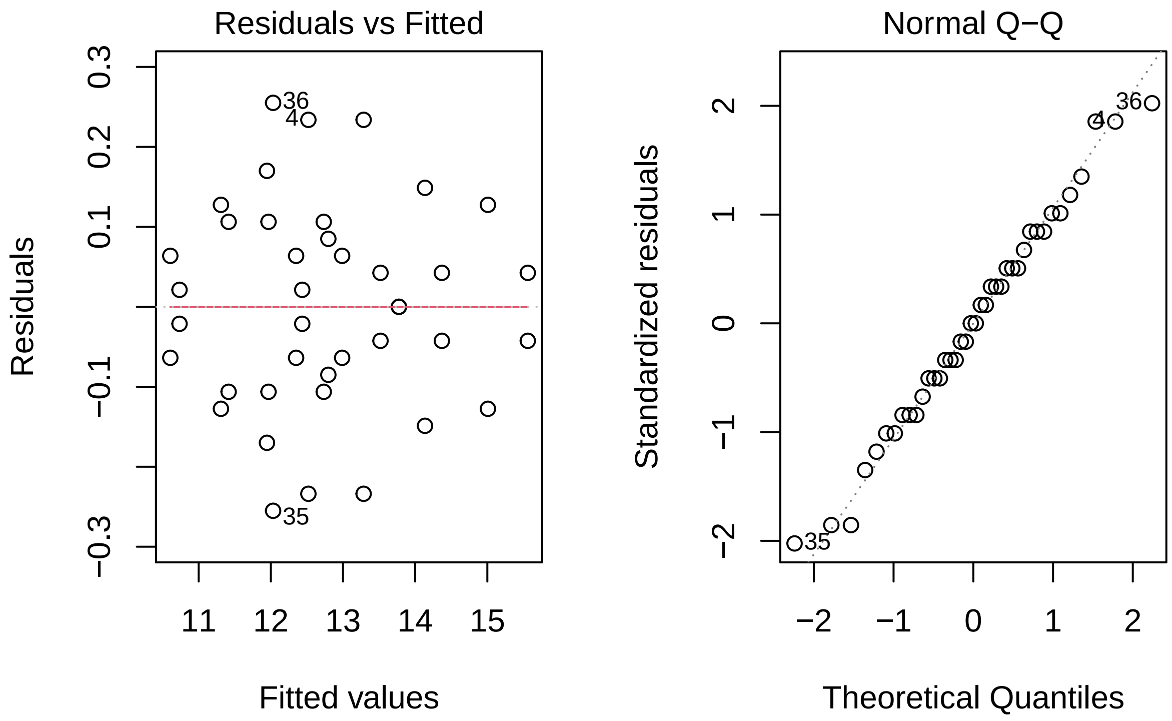
$H_0 : \gamma_{ij} = 0$; H_1 : Not all $\gamma_{ij} = 0$

$H_0 : \alpha_i = 0$; H_1 : Not all $\alpha i = 0$

$H_0 : \beta j = 0$; H_1 : Not all $\beta j = 0$

- **Check assumptions:**

```r
kml.aov <- aov(kmL ~ driver * car, data = kml)
par(mfrow = c(1,2))
plot(kml.aov, which = 1:2)
```



In these graphs, it can be observed that the residuals follow a normal distribution. Nevertheless, in the previous boxplot it can be observed that the variability between the effects is slightly different.

```r
summary(kml.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## driver       3  50.66  16.887  531.60  < 2e-16 ***
## car          4  17.12   4.280  134.73 3.66e-14 ***
## driver:car  12   0.44   0.037    1.16    0.371    
## Residuals   20   0.64   0.032                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After the observing that the interaction between the two factors is insignificant, this term can be removed.

```
kml.aov2 <- aov(kmL ~ driver + car, data = kml)
summary(kml.aov2)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## driver       3  50.66  16.887   501.5 <2e-16 ***
## car          4  17.12   4.280   127.1 <2e-16 ***
## Residuals   32   1.08   0.034
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

4. State your conclusions about the effect of driver and car on the efficiency kmL. These conclusions are only required to be at the qualitative level and can be based off the outcomes of the hypothesis tests in c. and the preliminary plots in b.. You do not need to statistically examine the multiple comparisons between contrasts and interactions.

After the two way anova implementation, there is enough evidence to reject the null hypotheses for α_i and β_j because the P-Value for these two factors is less than the significance level of 5%. This means that not all $\alpha_i = 0$ and not all $\beta_j = 0$. On the other hand, there is not enough evidence to reject the null hypothesis for γ_{ij} because the P-Value for the interaction is greater than the significance level of 5%.

In conclusion, the car engine and the driver by themselves have an impact on the fuel efficiency, but the interaction between these two elements does not have an impact on the fuel efficiency.