STA303 - Assignment 1

Last name: Khambaita
First name: Sunil
Student ID: 1000285924
Course section: STA303H1S-L0101

Jan 28th, 2017

Initializing data once before starting the questions.

```
# Loading ggplot2 library
library(ggplot2)

# Loading in the a1 data set
a1data = read.table("workmandata.csv", sep = ",", header = T)

# Checking type of variables in this data
str(a1data)
```

```
## 'data.frame':    200 obs. of  2 variables:
##  $ workman: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ y      : int  318 289 309 317 286 281 284 288 293 264 ...
```

```
# Converting workman field in the dataset as a factor variable
a1data$workman = as.factor(a1data$workman)
```

Q1 (a-d) - Data 1: mental treatment output

(a) Calculate the means and standard deviations of output for each workman. make a boxplot comparing the part output for the 10 workmen, give a short comment for the boxplot produced.

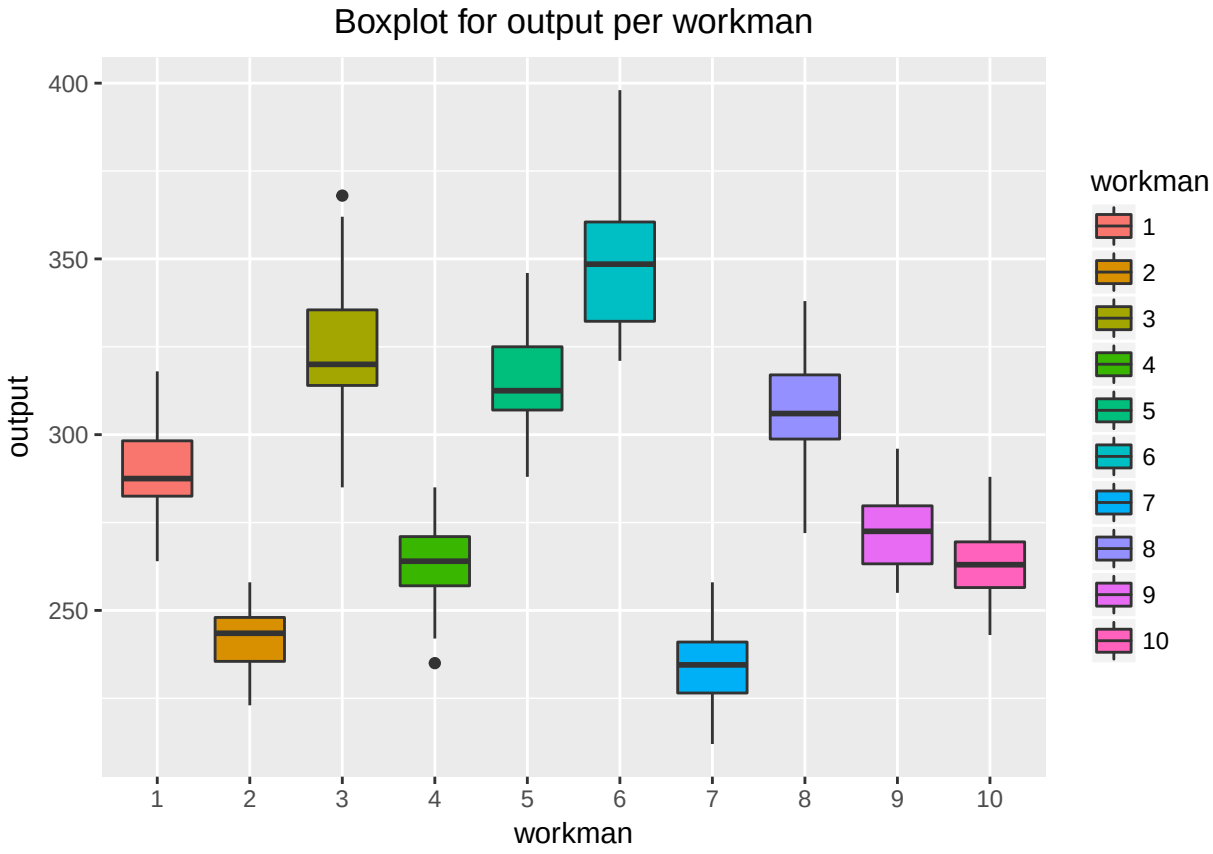
```
# Calculating the means of Y (output) for each workman
with(a1data, tapply(a1data$y, a1data$workman, mean))
```

```
##      1      2      3      4      5      6      7      8      9     10
## 290.25 241.70 324.65 262.85 314.85 350.70 234.80 306.60 273.45 263.90
```

```
# Calculating the standard deviations of Y (output) for each workman
with(a1data, tapply(a1data$y, a1data$workman, sd))
```

```
##         1         2         3         4         5         6         7
## 16.039015  9.608658 21.086975 11.430684 14.676422 23.517295 11.491874
##         8         9        10
## 16.109494 12.407447 12.130432
```

```r
# Creating a boxplot comparing Y (output) for the 10 workmen
ggplot(a1data, aes(x = workman, y = y)) + geom_boxplot(aes(fill = workman)) +
  xlab("workman") + ylab("output") + ggtitle("Boxplot for output per workman") +
  theme(plot.title = element_text(hjust = 0.5))
```



Comments on boxplot : . . .

Analyzing the boxplot, we can see that there is definitely variability on output from all the 10 workmen.

Workman #7 seems to have the lowest output compared to the rest with its inter-quartile range falling under 250 units, the most similar to workman #7's productivity is workman #2.

Workman #6 seems to be the most productive of the group, with his/her inter-quartile range falling above 350 units which is the most in all 10 of the workmen, he/she also has the longest upper whisker in the group meaning that they are able to produce almost close to 400 units on a good time.

Workman #4 and Workman #3 both have outliers which shows that on certain periods one produces slightly less than their usual normal output and the other produces slightly more than their usual normal output, respectively.

(b) Applying one-way ANOVA to this data, testing the equality of the output means for the ten workmen. State the null and alternative hypothesis for the p-value in ANOVA output. How significant is the result ?

```
# Applying one-way ANOVA
summary(aov(a1data$y~a1data$workman, data = a1data))
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## a1data$workman   9 254380   28264   118.5 <2e-16 ***
## Residuals      190  45335     239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null and alternative hypothesis for the F test: ...

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_9 = \mu_{10} \\ H_a : \exists i \neq j, s.t. \mu_i \neq \mu_j \end{cases}$$

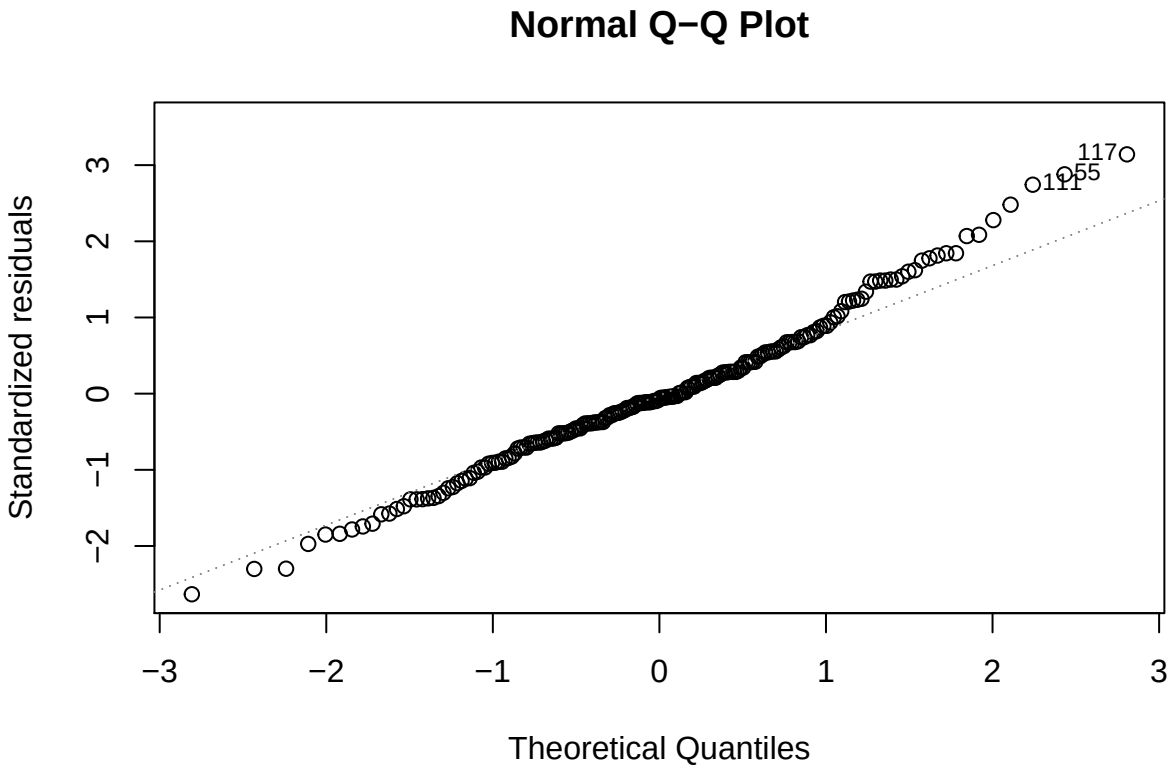
How significant is the result?:...

The p-value we got was less than 2.2×10^{-16}. This is a very very small number indicating that we can reject the null hypothesis with a high degree of significance.

This is also clear to see why by observing the boxplots calculated earlier definitely showed variability in output of each workmen.

(c) ANOVA assumes that the data in each group are distributed normally. This assumption is equivalent to saying that the residuals of the best-fitting model are distributed normally. Check the normality assumption by doing a qqnorm plot in conjunction with qqline based on the residuals from the linear regression model fitting. What conclusion do you have from the plot?

```
# Creating a normal qq plot
plot(lm(a1data$y~a1data$workman), which = 2, caption = list(""), main = "Normal Q-Q Plot",
     sub.caption = "")
```



Comments on normal Q-Q plot: ...

Observing the normal qq plot it appears to be lightly tailed on both sides but it is safe to conclude that the sample looks okay and it satisfies it's normality assumption.

(d) Examine the output variability for the ten workmen using the Bartlett test. What is your conclusion?

```
# Calculating output variability using Bartlett's test.
bartlett.test(a1data$y~a1data$workman)
```

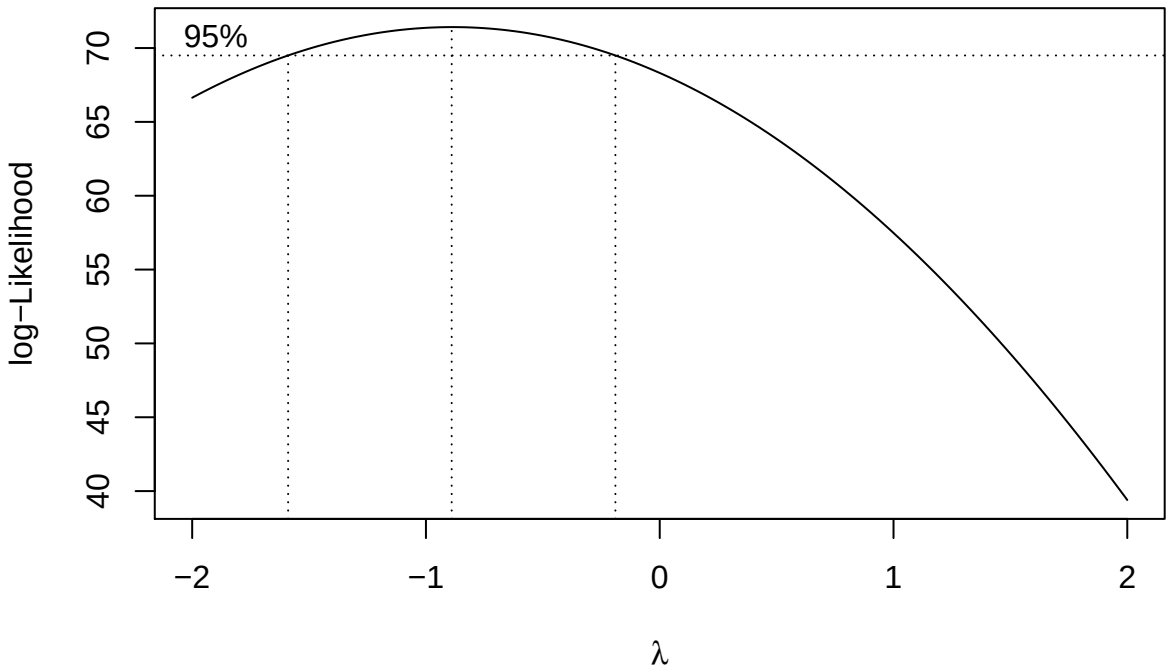
```
##
##  Bartlett test of homogeneity of variances
##
## data:  a1data$y by a1data$workman
## Bartlett's K-squared = 28.792, df = 9, p-value = 0.0007024
```

Conclusion: Using the Bartlett's test we receive a p-value of 0.0007024 which is less than 0.05 and we can conclude by rejecting the null hypothesis, which tells us the variances across the groups are not equal.

Q2 (a-d) - Data 1: donner party output

(a) To stabilizing the variance, we apply Box-cox power transformation, it suggests a simple variance stabilizing of the data. What is the simple transformation on Y suggested from boxcox()?

```
library(MASS)
bc=boxcox(lm(a1data$y ~ a1data$workman),lambda=seq(-2,2,by=0.01))
```




```
bc$x[bc$y==max(bc$y)]
```

```
## [1] -0.89
```

Simple Transformation: For a simple power transformation, I believe that the inverse of the response variable (y) appears to be the most appropriate choice. If we were to analyze the box-cox transformation diagram, we notice that the 95% confidence interval for λ is at the maximum when it's at the value -0.89, which is closer to -1 than -$\frac{1}{2}$ which is my reasoning in choosing the inverse transformation.

(b) Examine the transformed Y (from Q2-a) variability for the ten workmen using the Bartlett test. What is your conclusion? Does it agree or disagree with Q1-d?

```
# Performing the inverse transformation on the response variable
inverse_output = (a1data$y)^(-1)

# Calculating output variability on transformed response using Bartlett's test.
bartlett.test(inverse_output~a1data$workman)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  inverse_output by a1data$workman
## Bartlett's K-squared = 3.5241, df = 9, p-value = 0.9399
```

Conclusion: Using the Bartlett's test on the transformed Y we receive a p-value of 0.9399 which is greater than 0.05 and we can conclude by failing to reject the null hypothesis, as we do not have enough evidence. This tells us the variances across the groups are equal and disagrees with Q1-d.

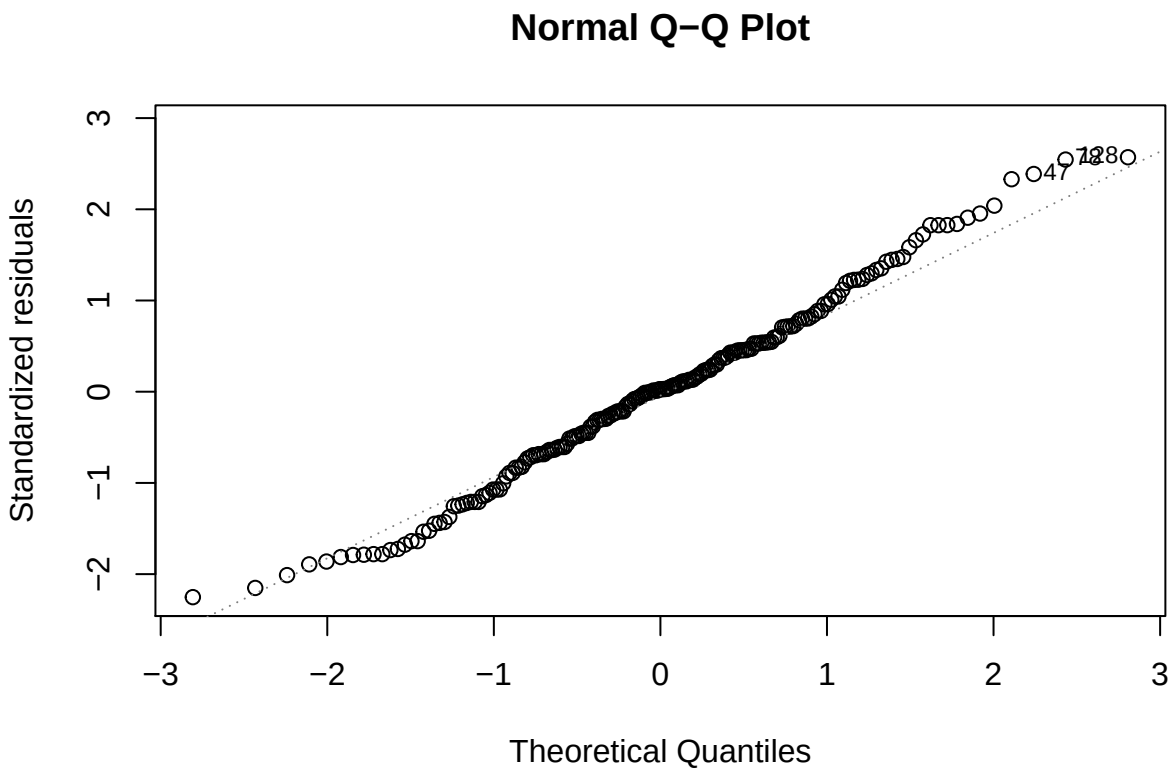
(c) Applying one-way ANOVA to this data, testing the equality of the output means for the ten workmen. How significant is the result? Does it agree with result you have in (Q1-b). Also repeat Q1-c to check the normality assumption for the transformed data, compare to Q1-c, what comment do you have ?

```
# Applying one-way ANOVA
summary(aov(inverse_output ~ a1data$workman))
```

```
##                 Df    Sum Sq   Mean Sq F value Pr(>F)
## a1data$workman   9 3.829e-05 4.254e-06   132.9 <2e-16 ***
## Residuals      190 6.080e-06 3.200e-08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: The p-value we received using the transformed Y was less than 2.2×10^{-16}. This is a very very small number indicating that we can reject the null hypothesis with a high degree of significance. This result agrees with what we received in Q1-b.

```
# Creating a normal qq plot
plot(lm(inverse_output~a1data$workman), which = 2, caption = list(""), main = "Normal Q-Q Plot",
     sub.caption = "")
```



Observing the normal qq plot of the transformed Y, the results are pretty similar to what we found from Q1-c, this qq plot appears to be less lightly tailed compared to the first qq plot so it is safe to conclude that the sample looks okay and it satisfies it's normality assumption.

(d) Why would we want to prefer the second ANOVA over the first one, even though both give roughly the same significance?

We would prefer the second ANOVA over the first one mainly due to the fact that it satisfies all the assumptions of an ANOVA model which are independence, normality (distributions of residuals are normal) and equality of variance.

We notice that the Bartlett's test on the first model caused us to reject the null hypothesis, resulting us to

conclude the variances across the groups are not equal. However, we failed to reject it on the transformed model (second ANOVA) which shows that equality of variance was satisfied.

Q3 (a-c) - Data 2: beers tasting

(a) Find the rating mean for each country and type. Find also the cell mean for each treatment combination (county and type combination).

```
beers = read.table("beers.csv",sep=",",header=T)
str(beers)
```

```
## 'data.frame':    36 obs. of  4 variables:
##  $ name   : Factor w/ 36 levels "1554 Black","60minute",..: 2 28 15 32 3 26 20 25 13 16 ...
##  $ type   : Factor w/ 2 levels "IPA","Lager": 1 1 1 1 1 1 1 1 1 1 ...
##  $ country: Factor w/ 3 levels "Belgium","UK",..: 3 3 3 3 3 3 1 1 1 1 ...
##  $ rating : num  4.09 4.19 4.27 4.22 3.89 4.48 4.21 3.81 3.99 4.04 ...
```

```
# Find the rating mean for each country
with(beers, tapply(beers$rating, beers$country, mean))
```

```
##  Belgium       UK      USA
## 3.654167 3.535833 3.775833
```

```
# Find the rating mean for each type of beers
with(beers, tapply(beers$rating, beers$type, mean))
```

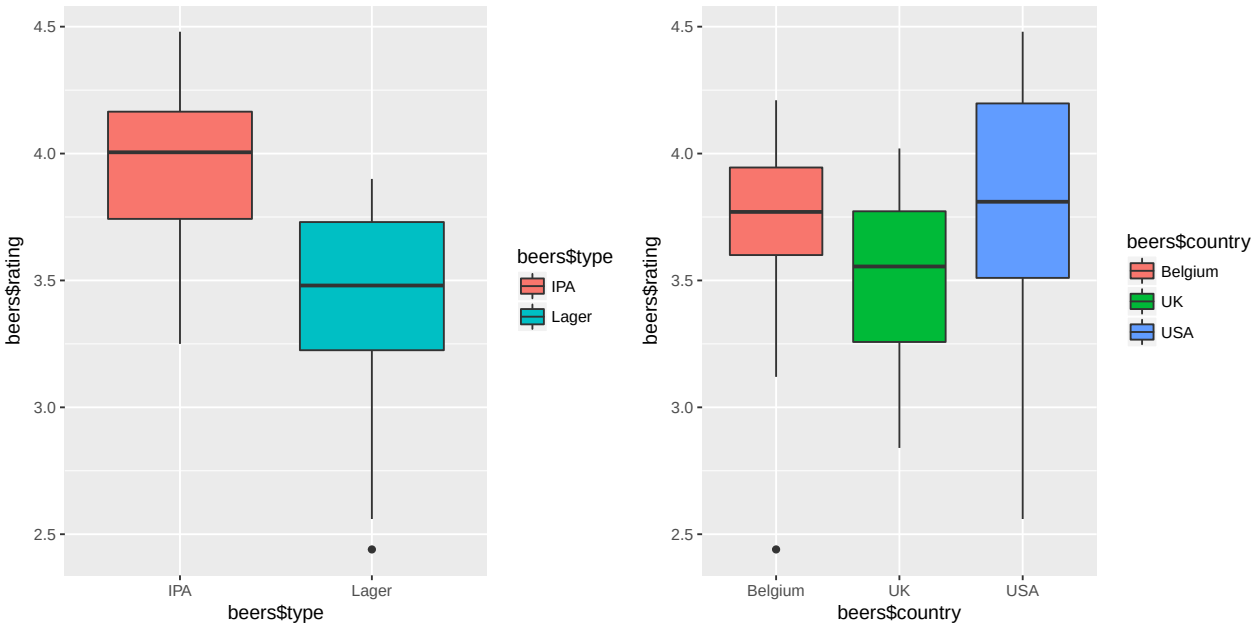
```
##      IPA    Lager
## 3.922778 3.387778
```

```
# Find the cell mean for each treatment combination
aggregate(beers$rating~beers$country+beers$type, data = beers, FUN="mean")
```

```
##   beers$country beers$type beers$rating
## 1       Belgium        IPA     3.950000
## 2            UK        IPA     3.628333
## 3           USA        IPA     4.190000
## 4       Belgium      Lager     3.358333
## 5            UK      Lager     3.443333
## 6           USA      Lager     3.361667
```

(b) Create box-plot of rating with respect to two factors, type and country. What can you say about the difference of rating mean for each factor ?

```
library(gridExtra)
library(ggplot2)

p1 = ggplot(beers, aes(x = beers$type, y = beers$rating, fill = beers$type)) + geom_boxplot()

p2 = ggplot(beers, aes(x = beers$country, y = beers$rating, fill = beers$country)) + geom_boxplot()

grid.arrange(p1, p2, ncol=2)
```



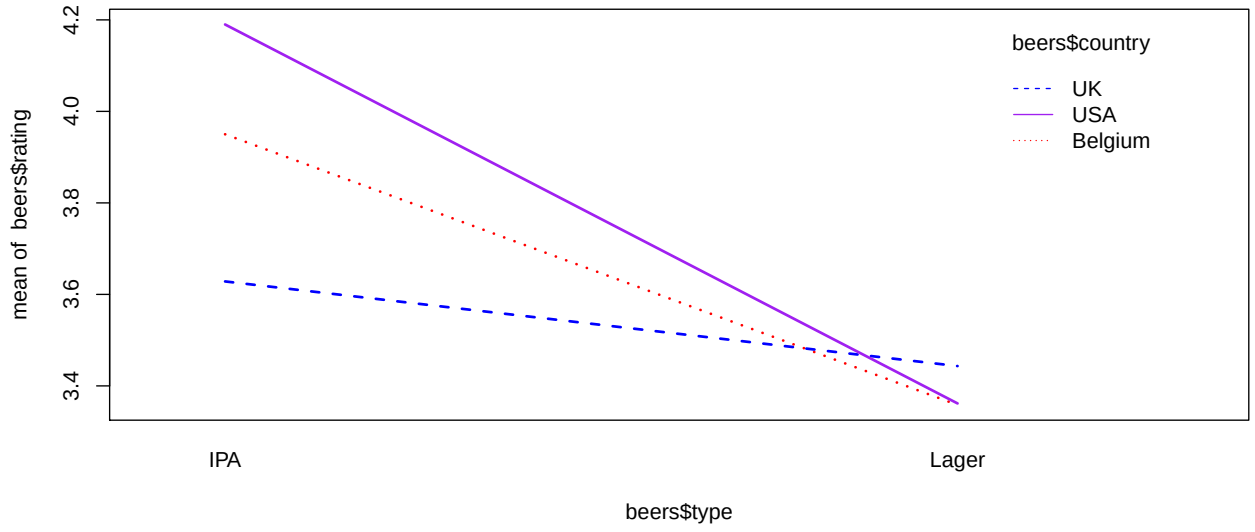
Comments: Observing the two boxplots: We can draw the following conclusion, from the first boxplot of rating by type, it's quite clear to conclude that the main effect of type seems to be significant, as they differ on Lager and IPA.

From the second boxplot of rating by country, although there are some differences in location and spread between the country samples, these differences do not really show a noticeable pattern and do not seem significant. Therefore the main effect of country seem do not seem to be significant.

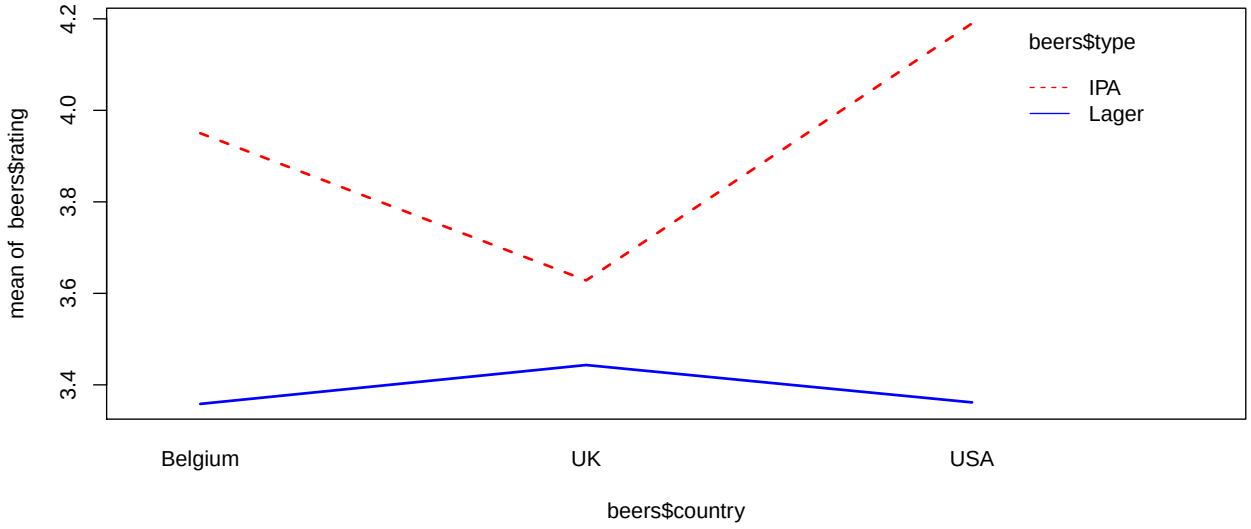
However, tests need to be ran in order to be confirmed.

(c) Create the interaction plot. What could you say about the main effect and interaction effect ?

```
# Creating interaction plots
interaction.plot(beers$type,beers$country,beers$rating,lwd = 2,col=c("red","blue", "purple"))
```



```
interaction.plot(beers$country, beers$type, beers$rating, lwd = 2, col = c("red","blue"))
```



Comments: Observing the interaction plots, we notice that there is an important interaction with country and beer type, we notice that IPA has the highest rating in USA, it also has the 2nd highest rating in Belgium. IPA is the lowest rating in UK. This is inverted when we speak about Lager. While UK has the lowest rating in IPA it has the highest rating in Lager and USA and Belgium in turn have the lowest. Overall, the main effect for country is bigger in beer type IPA than beer type Lager.

We also notice that there is a cross-over interaction when approaching beer type lager, the lines are not parallel but however converge which could suggest that there is a slight interaction at most.

Q4 (a-d) - Data 2: beers tasting

(a) Perform a two-way ANOVA to test the main effect of country and type, and for the interactions upon the rating. What conclusion do you have from this two-way ANOVA analysis ? How does this result connect to Q3-b.

```
# Performing two-way ANOVA
fit <- lm(beers$rating ~ beers$country * beers$type)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: beers$rating
##                           Df Sum Sq Mean Sq F value    Pr(>F)
## beers$country              2 0.3456 0.17281  1.2773 0.2935138
## beers$type                 1 2.5760 2.57602 19.0404 0.0001394 ***
## beers$country:beers$type   2 0.6353 0.31763  2.3477 0.1129074
## Residuals                 30 4.0588 0.13529
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: From our two-way ANOVA. We have the following conclusions:

Country: the p-value we have acquired is 0.2935138 which is greater than 0.05, therefore it's non significant and we fail to reject the null hypothesis which shows Country has no main effect.

Type: the p-value we have acquired is 0.0001394 which is less than 0.05, therefore it is significant and we have enough evidence to reject the null hypothesis which shows Type does have a main effect.

Interaction: the p-value we have acquired is 0.1129074 which is greater than 0.05, therefore it's non significant and we fail to reject the null hypothesis which states that the interaction term has no main effect.

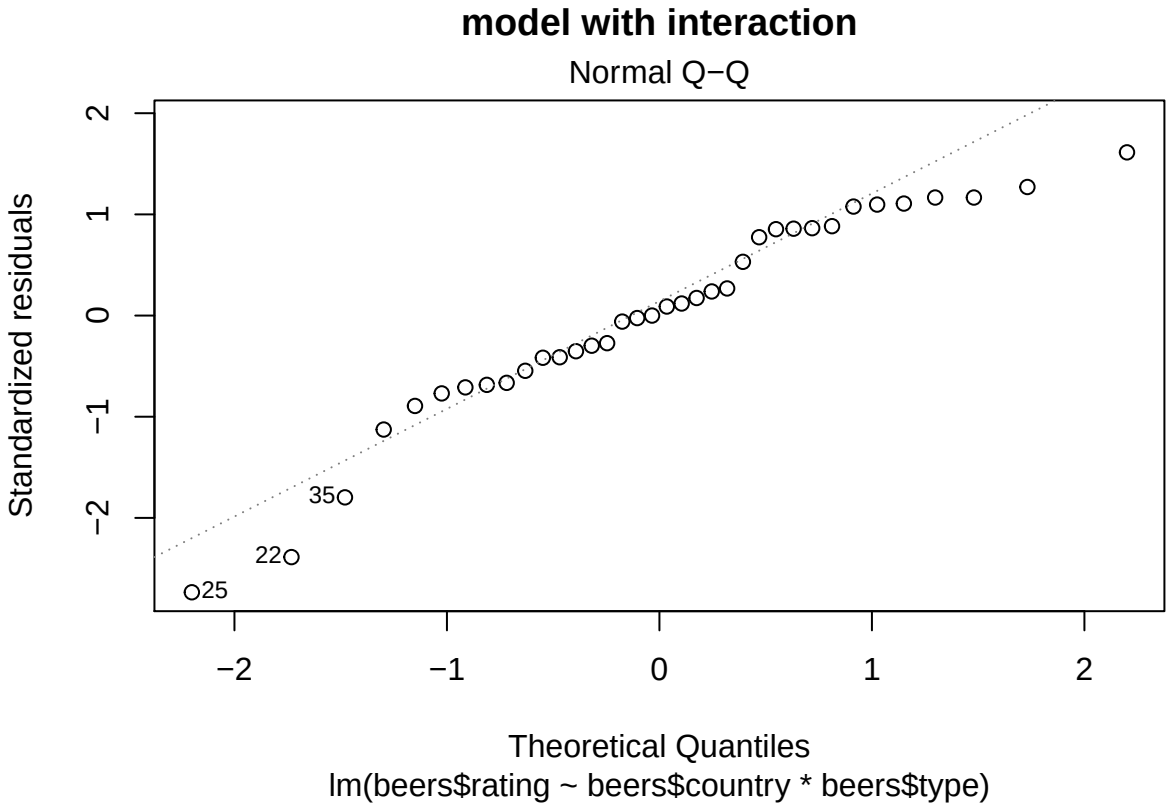
This result, confirms with conclusions

(b) Refit the data with a two-way ANOVA without the interaction term, give the ANOVA output. Checking the normality assumption before and after refitting as in Q1-c and state your conclusion.

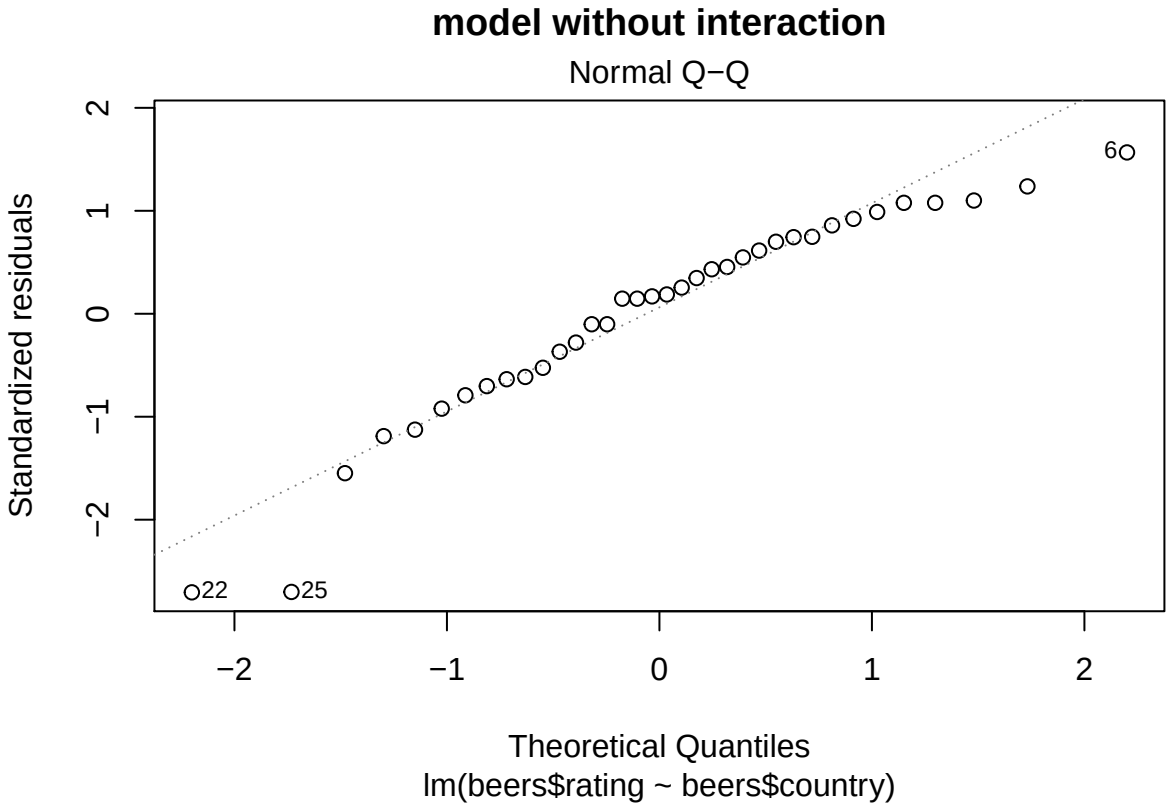
```
# Model with interaction
interaction_fit = lm(beers$rating ~ beers$country * beers$type)

# Model without interaction
fit = lm(beers$rating ~ beers$country)

# Anova fit
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: beers$rating
##               Df Sum Sq Mean Sq F value Pr(>F)
## beers$country  2 0.3456 0.17281  0.7844 0.4647
## Residuals     33 7.2701 0.22031
```

```
# Normal QQ with interaction
plot(interaction_fit, which = 2, main = "model with interaction")
```




```
# Normal QQ without interaction
plot(fit, which = 2, main = "model without interaction")
```



Comments: The p-value we have acquired is 0.4647 which is greater than 0.05 therefore we fail to reject the null hypothsesis stating that all rating means per country should be equal.

Adding on to this, normal qqplots with and without interaction are similar, which shows it's correct to fail to reject the null hypothesis.

(c) Instead of examining the normal qq plot, now we consider to use the Shapiro-Wilk Normality Test (R built-in function: *shapiro.test*()) to evaluate the normality assumption for model without interaction term.

```
# Shapiro test
shapiro.test(beers$rating)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  beers$rating
## W = 0.95431, p-value = 0.1429
```

Comments: The p-value we got is 0.1429 which greater than 0.05 therefore we fail to reject the null hypothesis stating that the sample comes from a population that has a normal distribution.

(d) Find 95% TukeyHSD family-wise confidence interval for the difference of means of county. Try R command $TukeyHSD(aov(rating\ type + country, data = beers), which = "country")$. Does this result agree with the significance you have in the ANOVA output in Q4-b?

```
# TurkeyHSD test
TukeyHSD(aov(rating~type+country,data=beers), which="country")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = rating ~ type + country, data = beers)
##
## $country
##                    diff        lwr       upr     p adj
## UK-Belgium  -0.1183333 -0.5025658 0.2658992 0.7317529
## USA-Belgium  0.1216667 -0.2625658 0.5058992 0.7189598
## USA-UK       0.2400000 -0.1442325 0.6242325 0.2884605
```

Comments: The result we achieved disagrees with the answer in Q4-b. The ANOVA table suggests that country has no effect on rating. This technically means that all the means should be equal. However, Turkey's CI's show that the rating mean per country are different from each other.