STA303 - Assignment 1

Last name: Akad
First name: Dogan
Student ID: 1001386083
Course section: STA303H1S-L0101

Feb 5th, 2017

Q1 (a-d) - Data 1: working 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.

```
library(ggplot2)
work = read.table("/Users/doganakad/Desktop/uoft/second semester/sta303/Assignments/A1/workmandata.csv"
str(work) # check the type of variables in this data
```

```
## '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 ...
```

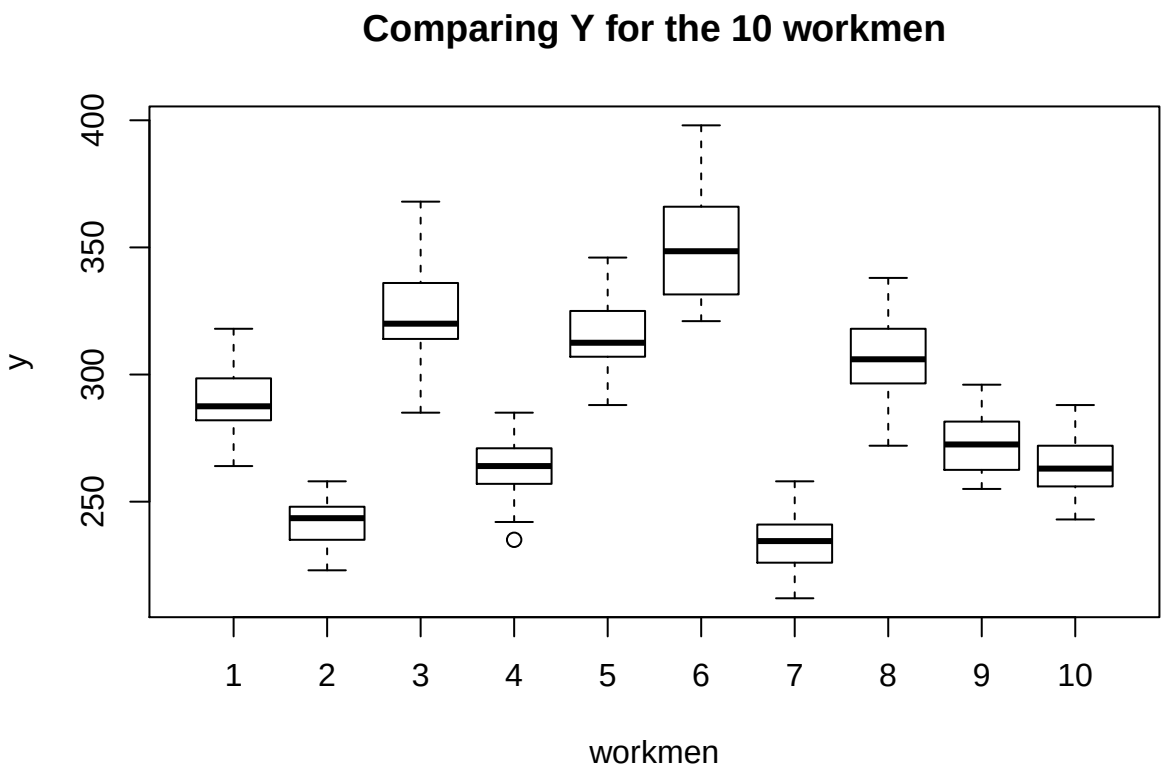
```
work$workman = as.factor( work$workman) # put workman into a factor variable
y = work$y
workman = work$workman

# The means and sd of Y for each workman
with(work ,tapply(y, 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
```

```
with(work, tapply(y, 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
```

```
# the boxplot
boxplot(y~workman,data=work, main="Comparing Y for the 10 workmen", xlab="workmen", ylab="y")
```



Comments on boxplot : Looking at the boxplot we can say that each workmen has a different daily part output value. Groups such as 2 and 7 have the lowest daily part output value whereas group 6 has the highest y value with median 348.5. Group 3's median is closer to the lower quartile which can indicate the distribution of y values can be negatively skewed. There's an outlier at group 4.

(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 ?

```
# one way anova
summary(aov(y~workman, data=work))
```

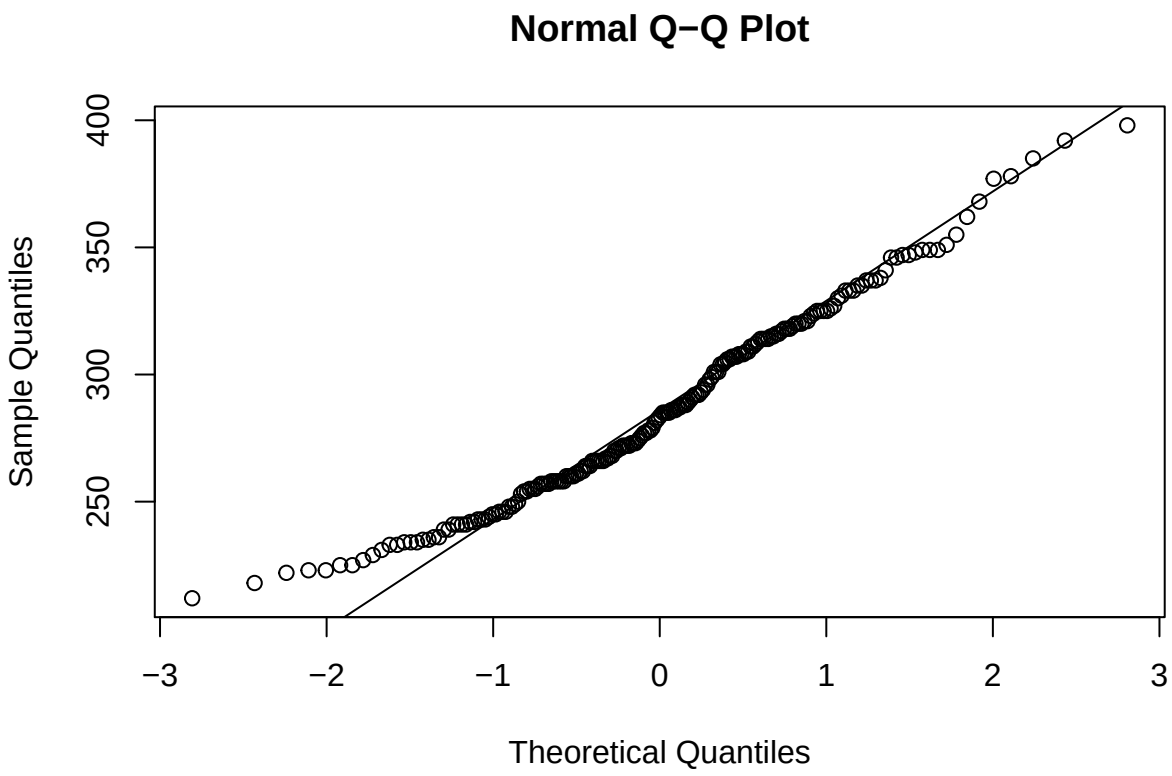
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## workman       9 254380   28264   118.5 <2e-16 ***
## Residuals   190  45335     239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null and alternative hypothesis for the F test: Null hypothesis: Output means for all the ten workmen are the same. Alternative hypothesis: At least two of the workmen have different output means.

How significant is the result?: Because our p value is very low, we have strong significant evidence to reject our null hypothesis. Therefore, at least two of the workmen have different output means and the result is significant.

(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?

```
# qqnorm plot
qqnorm(y); qqline(y)
```



Comments on normal Q-Q plot: In the lower(-3- -1.5) and upper(2-3) quantiles, residuals do not fall on to the straight line. Other than these quantiles, residuals follow a straight line which indicates that between -1 and 2 the residuals follow a normal distribution. Therefore, the residuals in general in this plot are not perfectly normally distributed.

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

```
# bartlett test
bartlett.test(y~workman, data=work)
```

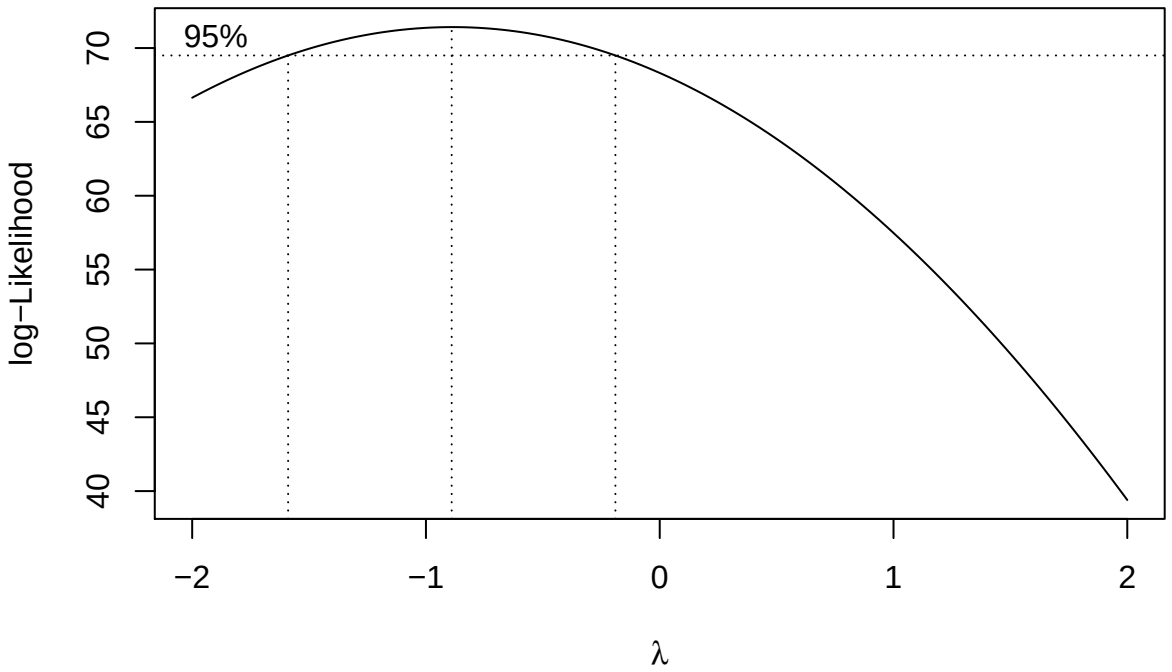
```
##
##  Bartlett test of homogeneity of variances
##
## data:  y by workman
## Bartlett's K-squared = 28.792, df = 9, p-value = 0.0007024
```

Conclusion: With 0.05 significance level, p-value is less than the significance level. Therefore, we strongly reject our null hypothesis and can conclude that 10 groups don't have equal variances.

Q2 (a-d) - Data 1: working 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(y~workman,lambda=seq(-2,2,by=0.01))
```




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

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

Simple Transformation: Simple transformation on Y suggested from boxcox is 1/y.

(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?

```
Yt = (1/work$y)
bartlett.test(Yt~workman, data=work)
```

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

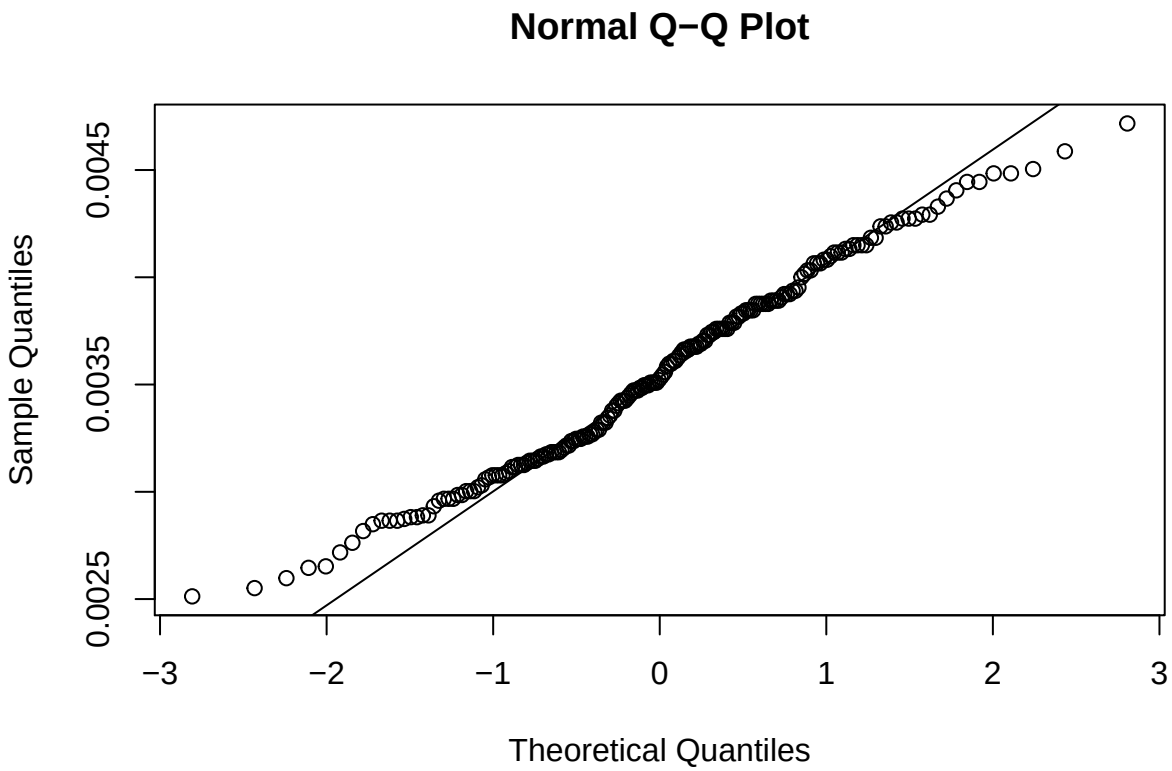
Conclusion: With 0.05 significance level, since the p-value is greater we fail to reject the fact that the variance is same for all groups. It's the opposite of the result we have found in Q1-d therefore it disagrees with it.

(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 ?

```
# one way anova
summary(aov(Yt~workman, data=work))
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## 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
```

```
# normal qqplot
qqnorm(Yt); qqline(Yt)
```



Conclusion: Looking at the anova table, the p-value is extremely small(<2e-16). With 0.05 significance level, since pvalue < significance level, we strongly reject the null hypothesis and therefore can conclude that there are at least two groups with different means. It's the same result what we have found in Q1b. Looking at the normal qqplot, the residuals follow the line very well except at the upper quantile but even with that we can still say that the normality assumption is satisfied.

(d) Why would we want to prefer the second ANOVA over the first one, even though both give roughly the same significance? We prefer the second ANOVA over the first one because the residuals of the second one are normally distributed.

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("/Users/doganakad/Desktop/uoft/second semester/sta303/Assignments/A1/beers.csv",sep=
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 ...
```

```
beers$name = as.factor(beers$name) # put name into a factor variable
beers$type = as.factor(beers$type) # put type into a factor variable
beers$country = as.factor(beers$country) # put country into a factor variable
rating = beers$rating
name = beers$name
type = beers$type
country = beers$country

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

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

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

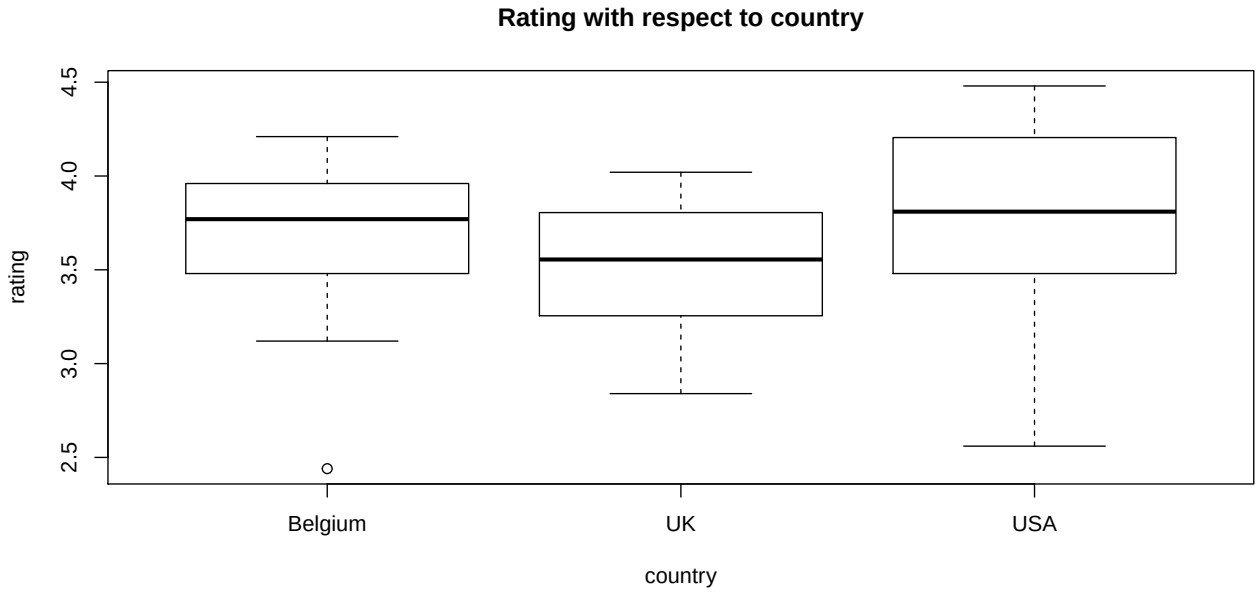
```
##      IPA    Lager
## 3.922778 3.387778
```

```
# Find the cell mean for each treatment combination
with(beers, tapply(rating, list(type, country), mean))
```

```
##        Belgium       UK      USA
## IPA   3.950000 3.628333 4.190000
## Lager 3.358333 3.443333 3.361667
```

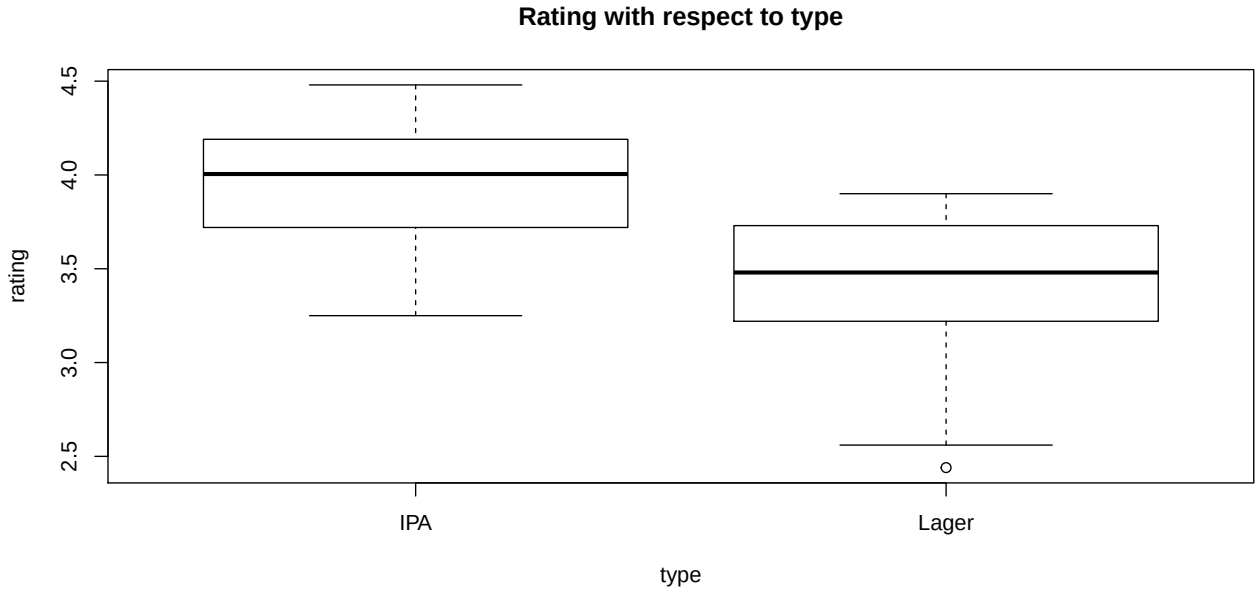
(b) Create box-plot of rating with respect to two factors, type and country. What can you say about the difference of rating mean for each factor ?

```
# boxplot of rating with respect to country
boxplot(rating~country,data=beers, main="Rating with respect to country", xlab="country", ylab="rating"
```



Comments with respect to country: Medians and interquartile ranges for each country is really close to each other.Therefore there is a similar variation in countries. Their means are also close to each other if we look at the result at Q3-a. Combining these factors, we see evidence of low between-group variance, so we have a little evidence that the main effect is significant.

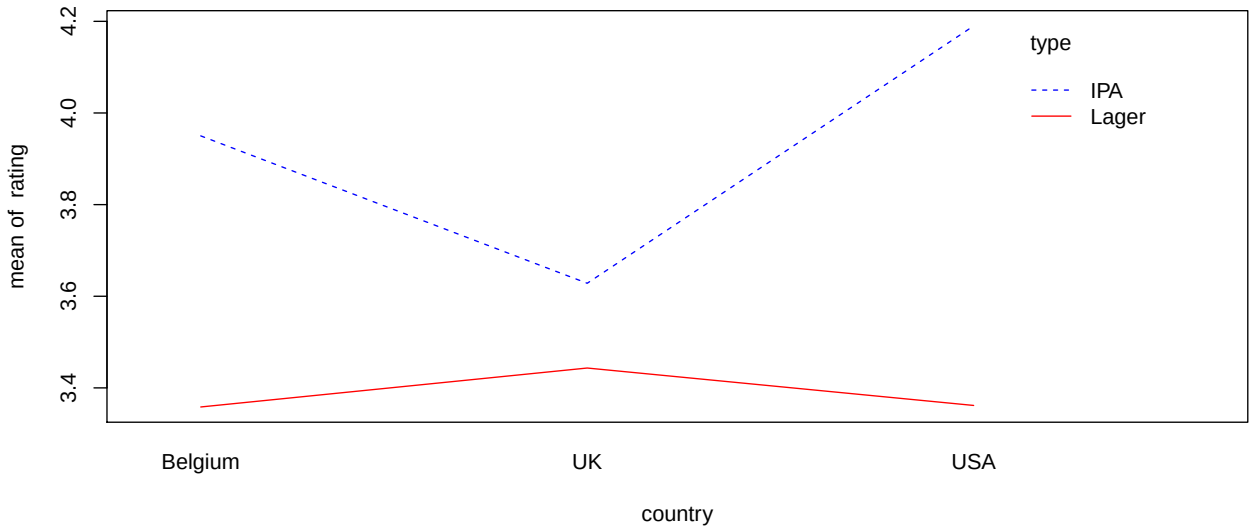
```
# boxplot of rating with respect to type
boxplot(rating~type,data=beers, main="Rating with respect to type", xlab="type", ylab="rating")
```



Comments with respect to type: IPA has higher rating than Lager. Their means are significantly different, not even close to each other if we look at the result from Q3-a. IPA's median is higher than Lager's upper quantile. Combining all of these we can say that effect of beer type on rating is signifiant.

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

```
# interaction plot
 with(beers ,interaction.plot(country, type, rating, col=c("blue","red"), legend = TRUE))
```



Comments: If we change the level of country, IPA has a bigger/stronger change than Lager which indicates that country has a bigger main effect on IPA. Looking at the plot, we can also say that since increasing the level of country from Belgium to UK causes the mean rating of IPA to drop substantially and increasing from UK to USA causes the mean to increase, there is a reinforcement effect. This is the exact opposite for Lager, so the effect of country is surpressed when the type is Lager. Also there is a significant horizontal distance between the lines for IPA and lager which concludes that beer type has a significant effect. Finally, looking at the lines for both types, since increasing country changes the line we can also say that country has a significant effect.

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.

```
# two way anova
ConT = lm(formula = rating ~ country * type, data = beers)
anova(ConT)
```

```
## Analysis of Variance Table
##
## Response: rating
##              Df Sum Sq Mean Sq F value    Pr(>F)
## country       2 0.3456 0.17281  1.2773 0.2935138
## type          1 2.5760 2.57602 19.0404 0.0001394 ***
## country: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: The value of F statistic is large for type, therefore F test for type produces a significant result for type whereas it doesn't produce a significant result for country and interaction term. We can conclude from the F test for type that variances between groups is larger than variances within groups, therefore type has a significant main effect. In contrast to that, country doesn't have a significant main effect and country&type together don't have significant interaction effect because F test for type is not statistically significant.

(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.

```
# two way anova
Wi = lm(formula = rating ~ country+type, data = beers)
anova(Wi)
```

```
## Analysis of Variance Table
##
## Response: rating
##           Df Sum Sq Mean Sq F value    Pr(>F)
## country    2 0.3456 0.17281  1.1781 0.3208682
## type       1 2.5760 2.57602 17.5611 0.0002044 ***
## Residuals 32 4.6940 0.14669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments: The F value is not large, therefore we can say that the F test is not significant. This concludes that between group variances are not significantly higher than inside group variances.

(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(Wi$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Wi$residuals
```

```
## W = 0.95046, p-value = 0.1081
```

Comments: With 0.05 significance level and p value 0.1081 and since the pvalue is greater than our significance level, we fail to reject the null hypothesis. We fail to reject that residuals are normally distributed.

(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?

```
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: All 3 of the confidence intervals include the null hypothesis(0) which indicates that we fail to reject the null. Therefore we can conclude that there is no difference in treatment means among 3 levels. This result is the same with what we have found from the ANOVA output in Q4b.