Q6

Mert Göksel Bilge Özkır Aisuluu Baktybekova

```r
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```r
library(rstatix)
```

```
##
## Attaching package: 'rstatix'

## The following object is masked from 'package:stats':
##
##     filter
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks rstatix::filter(), stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
library(broom)
```

Data Prep:

```r
ap_1 <- c(1000,
          1500,
          1200,
          1800,
          1600,
          1100,
          1000,
          1250)
ap_2 <- c(1500,
          1800,
          2000,
          1200,
          2000,
          1700,
          1800,
          1900)
ap_3 <- c(900,
          1000,
          1200,
          1500,
          1200,
          1550,
          1000,
          1100)
df <- cbind(ap_1, ap_2, ap_3)
```

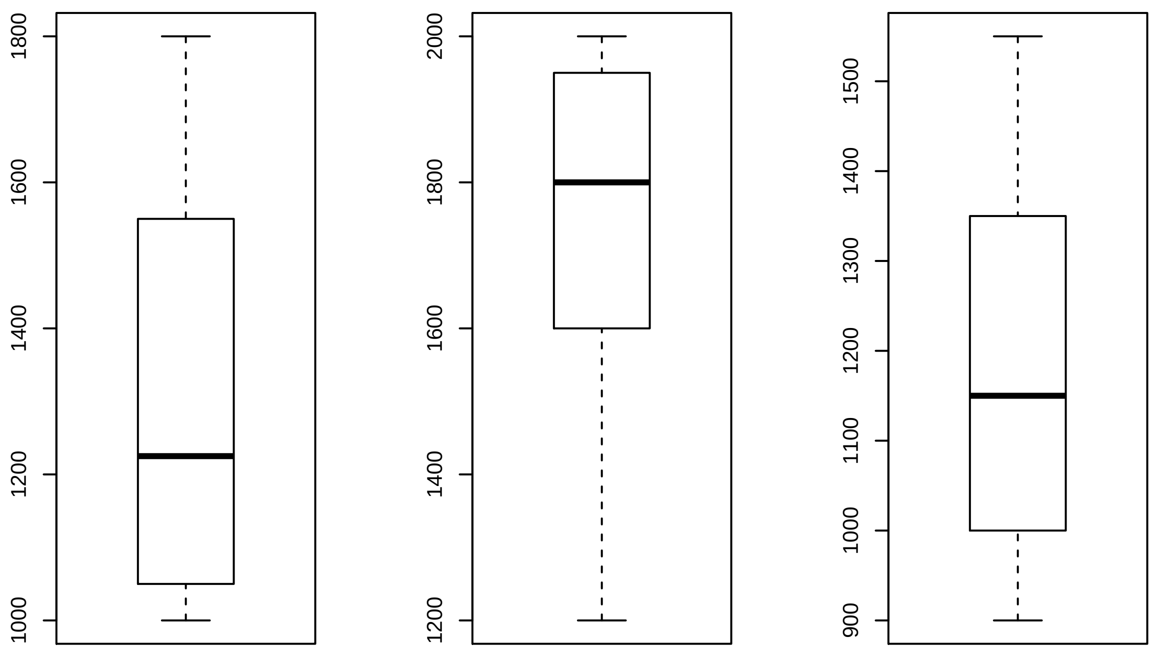
A:

```r
#We will use ANOVA to test to see if the observations
#suggest a difference between results of methods.
#H0: Pop Means are equal for all 3 groups
#H1: Pop Means are not equal for at least one group

#To apply ANOVA we need to see first if these samples can be assumed to be normal
apply(df, 2, shapiro.test)
```

```
## $ap_1
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.91076, p-value = 0.3595
##
##
## $ap_2
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.88411, p-value = 0.206
##
```

```
##
## $ap_3
##
##  Shapiro-Wilk normality test
##
## data:  newX[, i]
## W = 0.89968, p-value = 0.2871
```

```
#All of them seems to be good to use for ANOVA

#Lets see if they have any outliers
par(mfrow = c(1,3))
boxplot(df[,1])
boxplot(df[,2])
boxplot(df[,3])
```




```
#None has outliers

#Lets check homogenity of variances
melt(df) %>% levene_test(formula = value~Var2)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     2    21     0.195 0.825
```

```r
#P value is higher than 0.05 thus there is no evidence for heterogenity of variance

#Finally we can use anova
model <- aov(formula = value~Var2, data = melt(df))
model %>% summary()
```

```
##             Df  Sum Sq Mean Sq F value  Pr(>F)
## Var2         2 1362708  681354    9.41 0.00121 **
## Residuals   21 1520625   72411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

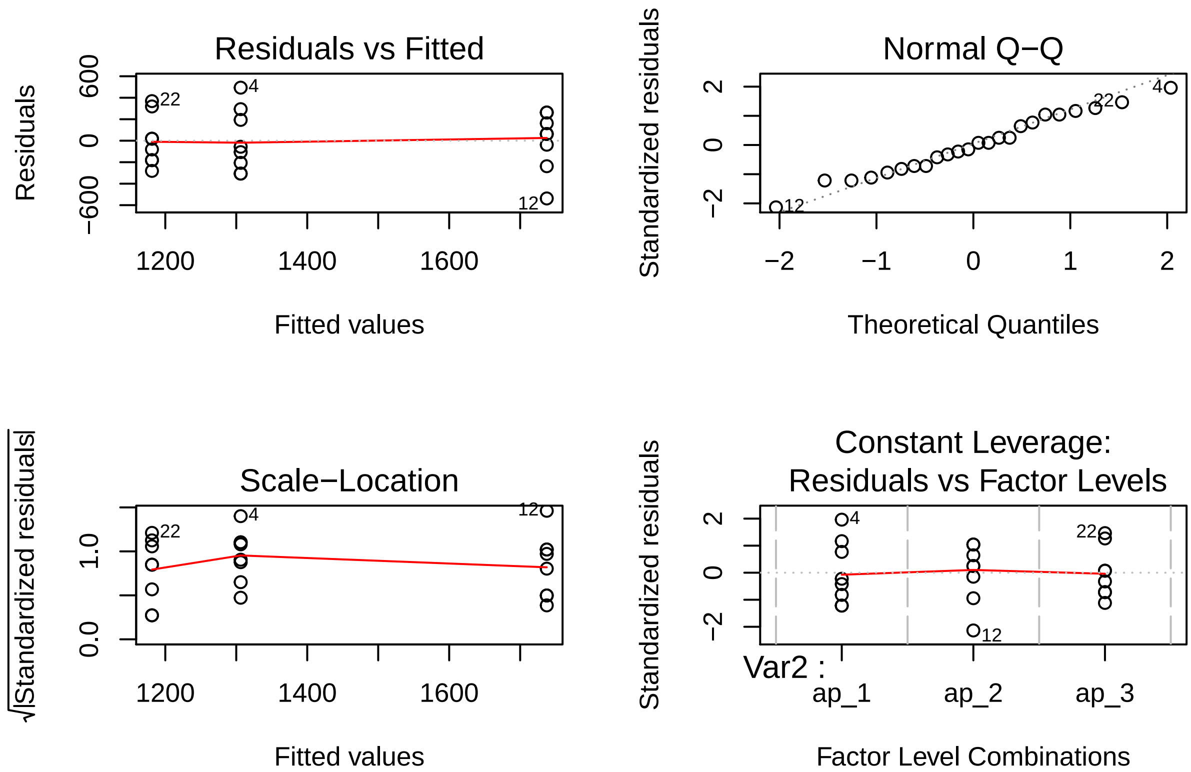
```r
#There are significant differences between groups as p value < 0.05
#But we do not know which of these 2,3 groups are different,
#this p value only tells us that there is a group that
#is different. So we can use Tukey comparison,or do pairwise ttest to each pair
#to see which groups are different from each other.
```

B:

```r
summary(resid(model))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -537.500 -187.500   -9.375    0.000  210.938  493.750
```

```r
#We see that mean is 0.

par(mfrow = c(2,2))
plot(model)
```




```
#First plot is showing how the residuals behave, we see no trend nor difference
#from zero.

#But there are some outliers namely 4, 12, 22.
#This may result in heterogeneity of variances or non normality.
#So we test for these.

#We already applied levenes test to see
#if variances are homogeneous and gotten good results.
#So lets test for normality of residuals. Looking at the qqplot
#it seems this test will be satisfactory, but lets test it anyways.
shapiro.test(resid(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.98067, p-value = 0.9078
```

```
#0.9 p value which is absolutely assumable to be normal.

#Now model adequacy is shown by Rsquared. Rsquared is from linear model.
#Linear model, if applied the same formula (value~Var2), will yield the same result.
#Lets test.
summary(model)
```

```
##              Df  Sum Sq Mean Sq F value  Pr(>F)
## Var2         2 1362708  681354    9.41 0.00121 **
## Residuals   21 1520625   72411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(value~Var2, data = melt(df)))
```

```
## Analysis of Variance Table
##
## Response: value
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## Var2       2 1362708  681354  9.4096 0.001209 **
## Residuals 21 1520625   72411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#See? Now we can use this regression model to get Rsquared

summary(lm(value~Var2, data = melt(df)))
```

```
##
## Call:
## lm(formula = value ~ Var2, data = melt(df))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -537.50 -187.50   -9.37  210.94  493.75
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1306.25      95.14  13.730 5.85e-12 ***
## Var2ap_2      431.25     134.55   3.205  0.00425 **
## Var2ap_3     -125.00     134.55  -0.929  0.36342
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 269.1 on 21 degrees of freedom
## Multiple R-squared:  0.4726, Adjusted R-squared:  0.4224
## F-statistic:  9.41 on 2 and 21 DF,  p-value: 0.001209
```

```
#R^2 = 47.26%. Meaning only 47.26% of total variation is explained by anova model.
#As this is oneway anova we dont need to consider adjusted rsquared.

#We can also find this R^2 with its formula.
tidy_aov <- tidy(model)
tidy_aov
```

```
## # A tibble: 2 x 6
##   term         df    sumsq  meansq statistic  p.value
##   <chr>     <dbl>    <dbl>   <dbl>     <dbl>    <dbl>
## 1 Var2          2 1362708. 681354.      9.41  0.00121
## 2 Residuals    21 1520625   72411.     NA    NA
```

```r
sum_squares_regression <- tidy_aov$sumsq[1]
sum_squares_residuals <- tidy_aov$sumsq[2]
Rs <- sum_squares_regression/(sum_squares_regression+sum_squares_residuals)
Rs
```

```
## [1] 0.4726156
```