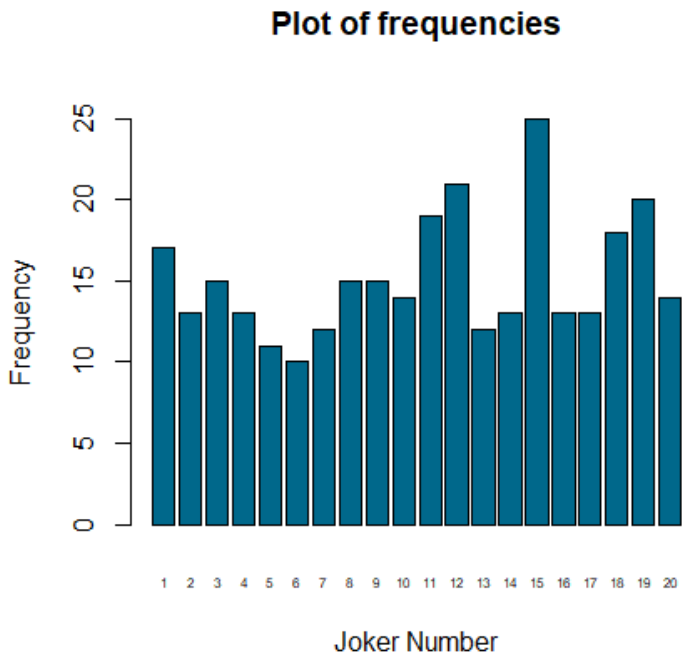
Assignment 3

Exercise 1

1. Import data and visualize

Import data form "joker_numbers.csv" file, which is created by the three files downloaded from Joker webpage. File consists of 1 column, each row being one Joker Number (File is included in assignment folder).

Plot frequency of each number (1 to 20) into a barplot.



2. Check whether lottery is fair or not

Perform Pearson' s chi-squared test to check whether the observed frequency distribution differs from the theoretical distribution or not. In other words, whether the lottery is fair or not. Each one of 20 Joker numbers is considered as categorical variable, to which a frequency was calculated and assigned. The events considered are mutually exclusive and have total probability equal to 1. Pearson's test hypothesis model is:

- H_0: All number' s frequencies are equal
- H_1: At least one number' s frequency is different

```
        Chi-squared test for given probabilities

data:  frequencies
X-squared = 17.858, df = 19, p-value = 0.5319
```

P-value is 0.5319. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected, thus lottery can be considered fair. Any difference between the frequencies arose by chance.

Exercise 2

1. Test the normality assumption of response time within each treatment.

Perform Shapiro Wilk test to check whether observations of response time of each treatment are normally distributed (a stands for response time of treatment A and b for response time of treatment b).

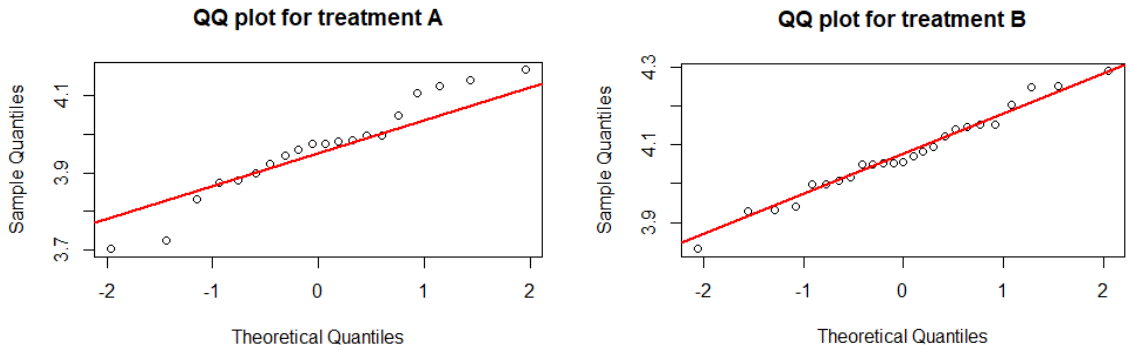
- H_0: observations in a/b are normally distributed
- H_1: observations in a/b are not normally distributed

```
> shapiro.test(a)

        Shapiro-Wilk normality test

data:  a
W = 0.95593, p-value = 0.4662
```

P-value is 0.4662. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected. Thus, observations can be considered normally distributed.

```
> shapiro.test(b)

        Shapiro-Wilk normality test

data:  b
W = 0.98129, p-value = 0.9094
```

P-value is 0.9094. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected. Thus, observations can be considered normally distributed.



Let's perform a second test for normality, in case Shapiro Wilk outcomes are not valid. Perform Anderson Darling test to check whether observations of response time of each treatment are normally distributed (a stands for response time of treatment A and b for response time of treatment b).

- H_0: observations in a/b are normally distributed
- H_1: observations in a/b are not normally distributed

```
> ad.test(a)

        Anderson-Darling normality test

data:  a
A = 0.35519, p-value = 0.4244
```

P-value is 0.4244. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected. Thus, observations can be considered normally distributed.

```
> ad.test(b)

        Anderson-Darling normality test

data:  b
A = 0.22483, p-value = 0.8
```

P-value is 0.8. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected. Thus, observations can be considered normally distributed.

2. Is the variance of response time common between treatments?

Perform F test to compare the two variances.

- H_0: ratio of a, b subset variances is 1
- H_1: ratio of a, b subset variances is not 1

```
        F test to compare two variances

data:  a and b
F = 1.2911, num df = 19, denom df = 24, p-value = 0.5482
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5505266 3.1661168
sample estimates:
ratio of variances
           1.29107
```

F-value is 0.5482. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected. Thus, variances can be considered equal.

Perform additional Bartlett test for equality of variances.

- H_0: variances of a, b subsets are homogeneous
- H_1: variances of a, b subsets are not homogeneous

```
        Bartlett test of homogeneity of variances

data:  list(a, b)
Bartlett's K-squared = 0.3405, df = 1, p-value = 0.5595
```

P-value is 0.5595. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected. Thus, variances can be considered equal.

3. Is there any difference in mean response time between the two drugs?

Perform t-test to check whether means of two subsets are equal. Variances are considered equal.

- H_0: means of a, b subsets are equal
- H_1: means of a, b subsets are not equal

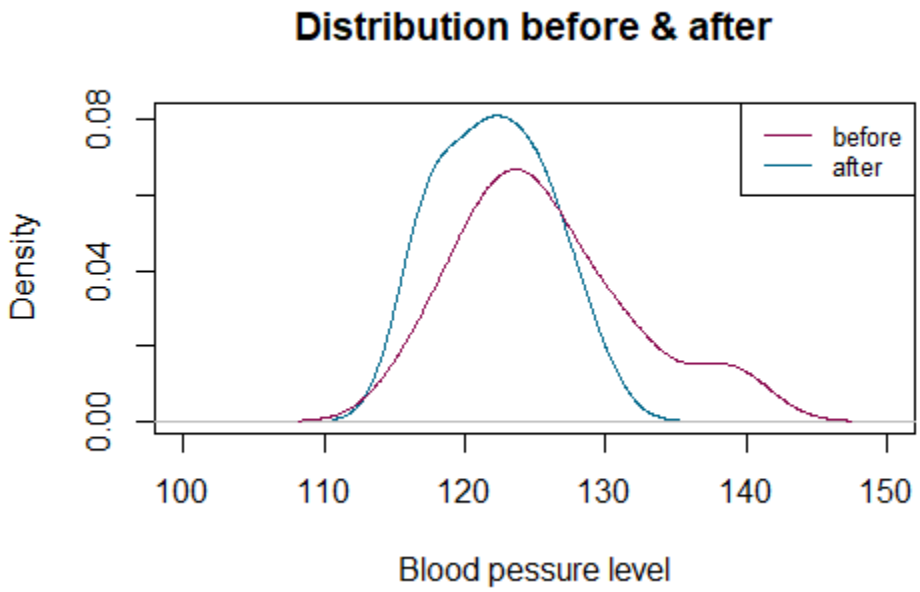
```
        Two Sample t-test

data:  a and b
t = -3.2419, df = 43, p-value = 0.002296
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.18338671 -0.04272819
sample estimates:
mean of x mean of y
 3.961835  4.074892
```

P-value is 0.002296. For a significance level of $\alpha = 5\%$, H_0 is rejected. Thus, means of response time of treatments are not equal.

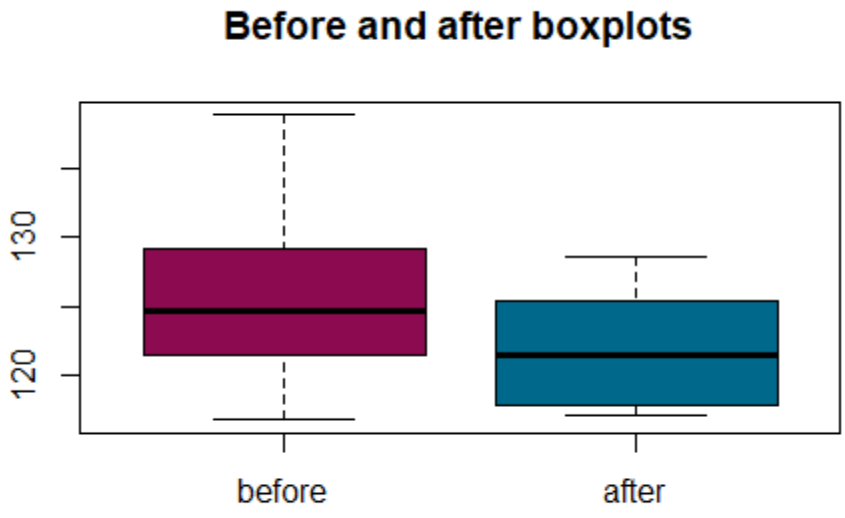
Exercise 3

1. Visualize the data and report summary statistics

Density plots for observations before and after the drug.



Boxplots for observations before and after the drug.



Summary statistics

```
     before          after      
 Min.   :116.8   Min.   :117.2  
 1st Qu.:121.7   1st Qu.:118.8  
 Median :124.7   Median :121.5  
 Mean   :125.7   Mean   :122.1  
 3rd Qu.:128.5   3rd Qu.:125.1  
 Max.   :138.8   Max.   :128.6  
```

Difference in means is already observed. Later we will prove that the deference observed is statistically significant. Min values are about the same, but Max values are quite different (10 units less for after drug observations).

2. Test the claim of the pharmaceutical company

The Pharmaceutical company claims that blood pressure after the drug is significantly lower than before. We will perform Shapiro Wilk test to show that observations are normally distributed and then t-test to prove the significance of the mean difference.

Perform Shapiro Wilk test to check the normality of observations.

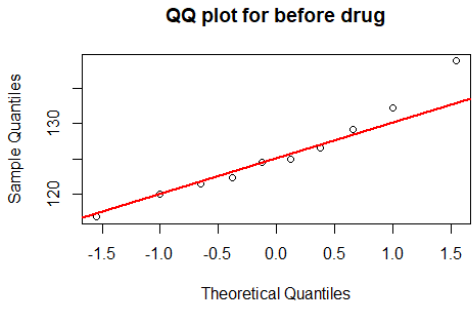
- H_0: observations of before/after are normally distributed
- H_1: observations of before/after are not normally distributed

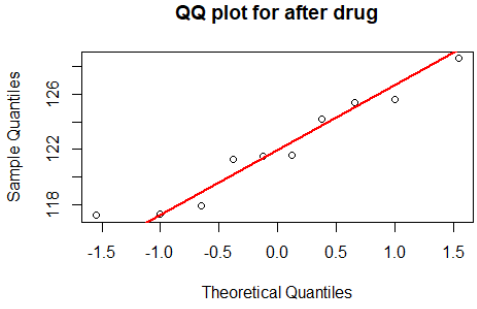
```
> shapiro.test(df$before)

        Shapiro-Wilk normality test

data:  df$before
W = 0.95863, p-value = 0.7701
```

P-value is 0.7701. For a significance level of $\alpha = 5\%$, H_0 is not rejected. Thus, observations are considered normally distributed.

```
> shapiro.test(df$after)

        Shapiro-Wilk normality test

data:  df$after
W = 0.9306, p-value = 0.4538
```

P-value is 0.4538. For a significance level of $\alpha = 5\%$, H_0 is not rejected. Thus, observations are considered normally distributed.





Anderson-Darling p-values were also in favor of null hypothesis.

```
> ad.test(df$before)

        Anderson-Darling normality test

data:  df$before
A = 0.23205, p-value = 0.7292

> ad.test(df$after)

        Anderson-Darling normality test

data:  df$after
A = 0.31356, p-value = 0.4876
```

Perform F test to compare the variances of before and after.

- H_0: variances of after, before subsets are equal
- H_1: variances of after, before subsets are not equal

```
        F test to compare two variances

data:  df$after and df$before
F = 0.36828, num df = 9, denom df = 9, p-value = 0.1529
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.09147525 1.48268834
sample estimates:
ratio of variances
         0.3682788
```

P-value is 0.1529. For a significance level of $\alpha = 5\%$, H_0 is not rejected. Thus, variances can be considered equal. Bartlett test p-value is also in favor of null hypothesis.

```
        Bartlett test of homogeneity of variances

data:  list(df$after, df$before)
Bartlett's K-squared = 2.044, df = 1, p-value = 0.1528
```

Perform t-test to check whether means of two subsets are different.

- H_0: mean of after is lower (or equal) than mean of before

- H_1: mean of after is greater than mean of before

```
        Paired t-test

data:  df$after and df$before
t = -2.1557, df = 9, p-value = 0.9703
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -6.698296       Inf
sample estimates:
mean of the differences
                  -3.62
```

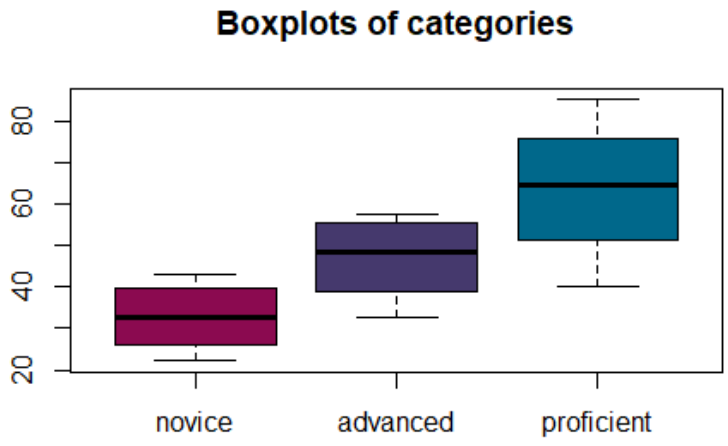
P-value is 0.9703. For a significance level of $\alpha = 5\%$, H_0 is not rejected. Thus, after is considered to have a lower mean than before, which means that the claim that the drug lowers blood pressure is true.

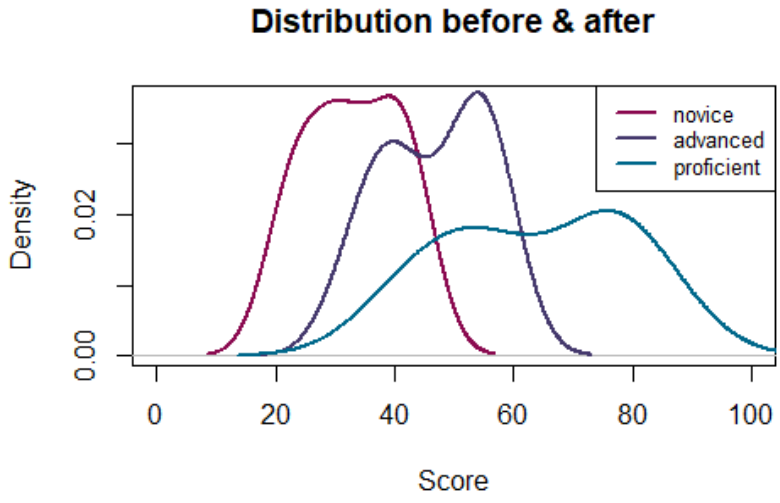
Exercise 4

1. Import, inspect and visualize the data into R.

```
     novice          advanced       proficient   
 Min.   :22.10   Min.   :32.50   Min.   :40.10  
 1st Qu.:27.05   1st Qu.:39.45   1st Qu.:52.50  
 Median :32.60   Median :48.40   Median :64.60  
 Mean   :33.04   Mean   :46.79   Mean   :63.89  
 3rd Qu.:39.50   3rd Qu.:54.92   3rd Qu.:75.65  
 Max.   :43.20   Max.   :57.70   Max.   :85.30  
```

We can already observe differences in the means, minimum and maximum values of each level. Let's observe the differences by visualizing the data.





2. Is there any difference among the groups in the ability of recalling previous cards?

We have three independent samples; therefore, we will test for any differences between the groups using ANOVA.

- H_0: means among groups are equal
- H_1: means among groups are not equal (at least one is different)

```
Call:
   aov(formula = score ~ factor(level), data = df1)

Terms:
                factor(level) Residuals
Sum of Squares       4777.317  3511.042
Deg. of Freedom             2        27

Residual standard error: 11.40345
Estimated effects may be unbalanced
> summary(fit)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(level)  2   4777    2389   18.37 9.21e-06 ***
Residuals     27   3511     130
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

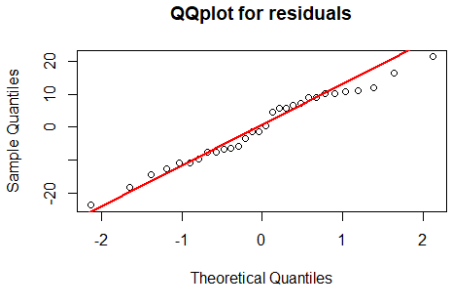
P-value is 9.21e-06. For a significance level of $\alpha = 5\%$, H_0 is rejected. Thus, there are differences among the groups. Now, let's check that the assumptions for performing ANOVA were not violated.

i. Normally distributed residuals

Perform Shapiro Wilk test to check the normality.

- H_0: residuals are normally distributed
- H_1: residuals are not normally distributed

```
        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.97185, p-value = 0.5909
```



P-value is 0.5909. For a significance level of $\alpha = 5\%$, H_0 is not rejected. Thus, we can say that residuals are normally distributed.

ii. Homogeneity of variances

Perform Bartlett, Fligner and Levene test to check homogeneity of variances among the three levels.

- H_0: variances of three levels are homogeneous
- H_1: variances are not homogeneous (at least one is different)

```
> bartlett.test(score ~ factor(level), data = df1)

        Bartlett test of homogeneity of variances

data:  score by factor(level)
Bartlett's K-squared = 4.6122, df = 2, p-value = 0.09965
```

P-value of Bartlett is in favor of H_0.

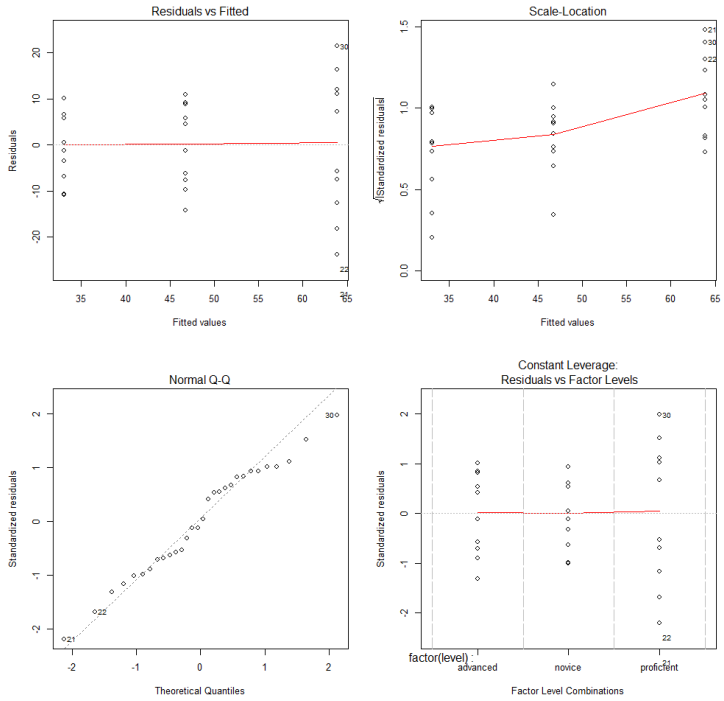
```
> fligner.test(score ~ factor(level), data = df1)

        Fligner-Killeen test of homogeneity of variances

data:  score by factor(level)
Fligner-Killeen:med chi-squared = 8.341, df = 2, p-value = 0.01544
```

P-values of Fligner is in favor of H_1.

```
> leveneTest(score ~ factor(level), data = df1)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value   Pr(>F)
group  2  5.9039 0.007464 **
      27
```

F-value of Levene is in favor of H_0.

iii. Diagnostic plots



Assumptions of ANOVA do not seem to be violated, meaning that the conclusion about the differences among the three levels is quite likely correct.

Nonetheless, we could run the non-parametric test of Kruskal-Wallis.

- H_0: means among groups are equal
- H_1: means among groups are not equal (at least one is different)

```
        Kruskal-Wallis rank sum test

data:  score by factor(level)
Kruskal-Wallis chi-squared = 17.387, df = 2, p-value = 0.0001677
```

P-value of Kruskal-Wallis test is also in favor of rejecting H_0 for a significance level of $\alpha = 5\%$.

In conclusion, both tests lead us into believing that there are significant differences among the three groups/levels.

3. Perform all pairwise t-tests. Is it valid to decide which groups are different by performing all pairwise comparisons via a t-test with significance level 5%?

Perform pairwise t-tests.

- H_0: mean of n+1 level is equal to mean of n level
- H_1: mean of n+1 level is not equal to mean of n level

```
        Pairwise comparisons using t tests with pooled SD

data:  score and factor(level)

           advanced novice
novice     0.0119   -
proficient 0.0024   1.9e-06

P value adjustment method: none
```

For a significance level 5% we conclude that all pairs have statistically significant differences in their means.

This method would create a Type I error 3 times the significance level we set (5%). In order to avoid that we could choose to transform the significance level. One technique is the Bonferroni one, which divides the significance level by the number of pairs compared, in this case 3.

4. Properly identify which groups are different at significance level 5%

Perform pairwise t-tests with Bonferroni adjustment.

- H_0: mean of n+1 level is equal to mean of n level
- H_1: mean of n+1 level is not equal to mean of n level

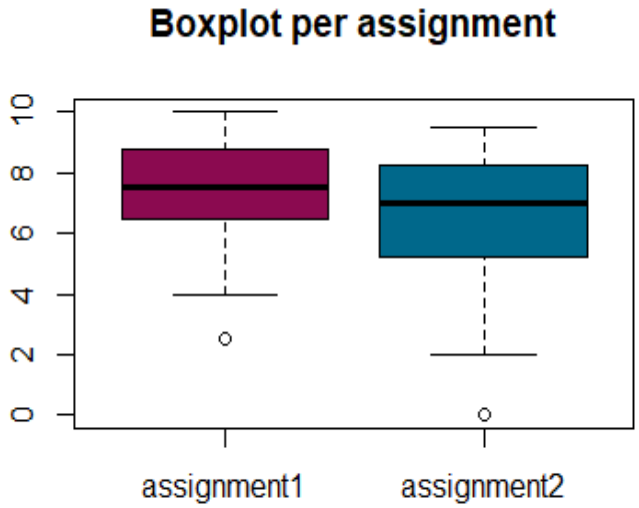
```
        Pairwise comparisons using t tests with pooled SD

data:  score and factor(level)

           advanced novice
novice     0.0358   -
proficient 0.0071   5.6e-06

P value adjustment method: bonferroni
```

For a significance level 5%/3 we conclude that advanced-proficient and novice-proficient pairs have statistically significant differences in their means, whereas novice-advanced pair does not.

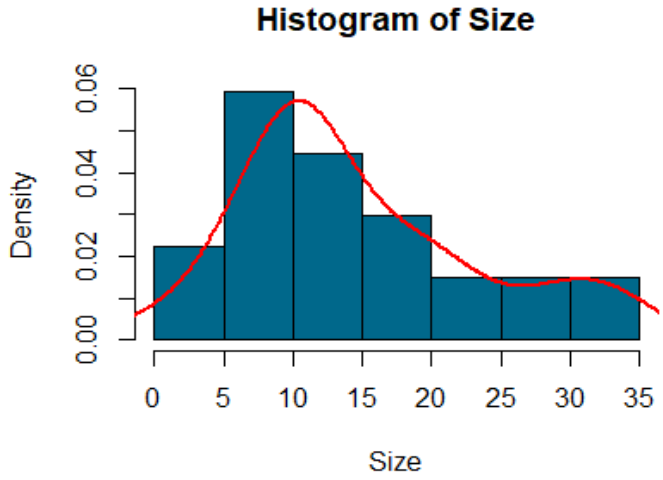
Exercise 5

1. Inspect and visualize the data.

```
   assignment1      assignment2         size
Min.   : 2.50   Min.   :0.000   Min.   : 0.00
1st Qu.: 6.50   1st Qu.:5.250   1st Qu.: 9.50
Median : 7.50   Median :7.000   Median :12.00
Mean   : 7.37   Mean   :6.537   Mean   :14.74
3rd Qu.: 8.75   3rd Qu.:8.250   3rd Qu.:19.00
Max.   :10.00   Max.   :9.500   Max.   :34.00
```



For assignment grades visualization a box plot is used. We can inspect a drop in the mean moving forward from assignment 1 to 2. Also, we see that there are outliers. In both assignments there are grades that are too low compared to the others.

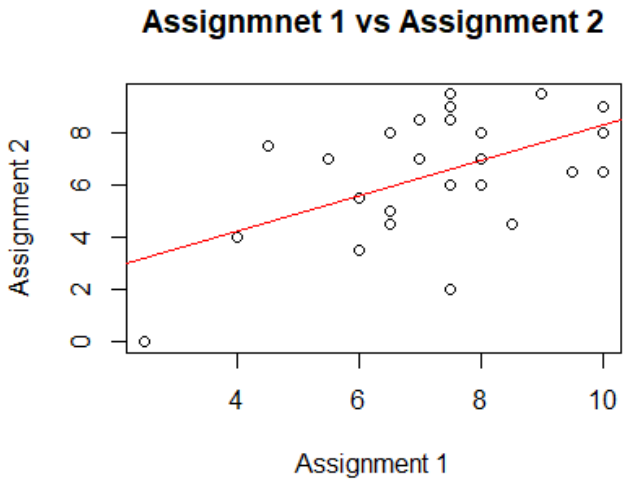


Size column is visualized through a histogram and a density line.

2. Fit a normal regression model in order to infer the grades of 2nd assignment using the 1st assignment grade as explanatory variable. Is the 2nd assignment grade affected by the 1st assignment grade? Plot the estimated regression line on top of the scatter-plot of the observed data. Superimpose the 90% confidence and prediction intervals.

```
> model <- lm(assignment2 ~ assignment1, data=grades)
> print(model)

Call:
lm(formula = assignment2 ~ assignment1, data = grades)

Coefficients:
(Intercept)  assignment1
     1.5140       0.6815
```

The linear regression model gives us the line assignment2 = 1.5140 + 0.6815assignmnet 1, which is shown as the red line on the scatterplot bellow.



Perform Pearson's product-moment correlation test.

- H_0: true correlation is equal to 0
- H_1: true correlation is not equal to 0

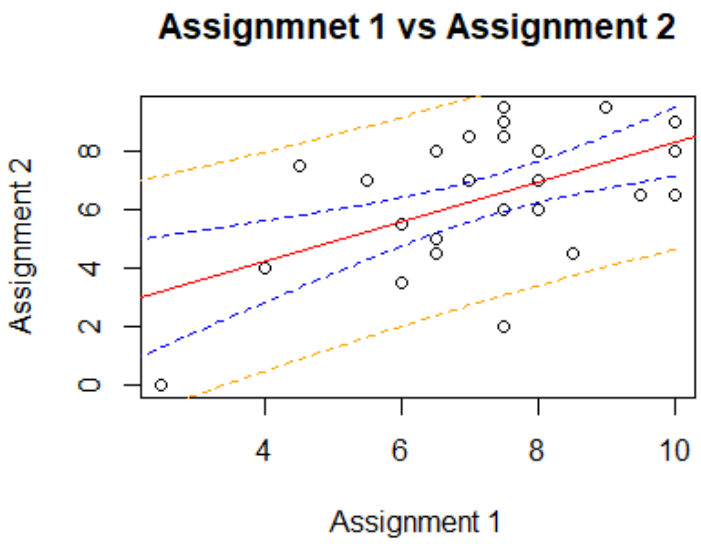
```
        Pearson's product-moment correlation

data:  grades$assignment1 and grades$assignment2
t = 3.1949, df = 25, p-value = 0.003763
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1991817 0.7624463
sample estimates:
      cor
0.5384401
```

P-value is 0.003763. For a significance level of $\alpha = 5\%$, H_0 is rejected. Thus, Pearson's test leads us to believe that there is indeed a correlation between the two assignment grades.

Spearman's rank correlation test with the same hypotheses, also, leads us to believe that there is correlation between the two assignment grades.

```
        Spearman's rank correlation rho

data:  grades$assignment1 and grades$assignment2
S = 1922.9, p-value = 0.03226
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4130247
```

Bellow, there are added the confidence and prediction interval lines.



Test modeling assumptions:

i. Linearity of the data:

Residuals vs Fitted diagram bellow shows if Assignment 1 and 2 are linearly related. A perfect linear relationship would give a red straight horizontal line. In our case it is somewhat horizontal. Also, we cannot observe a specific pattern in the residuals.

ii. Normality of residuals:

From the Normal Q-Q diagram bellow we can say that the residuals are normally distributed. Ideally all observations should lay on the line. In order to further check this assumption, we can perform Shapiro-Wilk test.

Perform Shapiro Wilk test to check the

- H_0: residuals normally distributed
- H_1: residuals are not normally distributed

```
> shapiro.test(model$residuals)

        Shapiro-Wilk normality test

data:  model$residuals
W = 0.95956, p-value = 0.3611
```

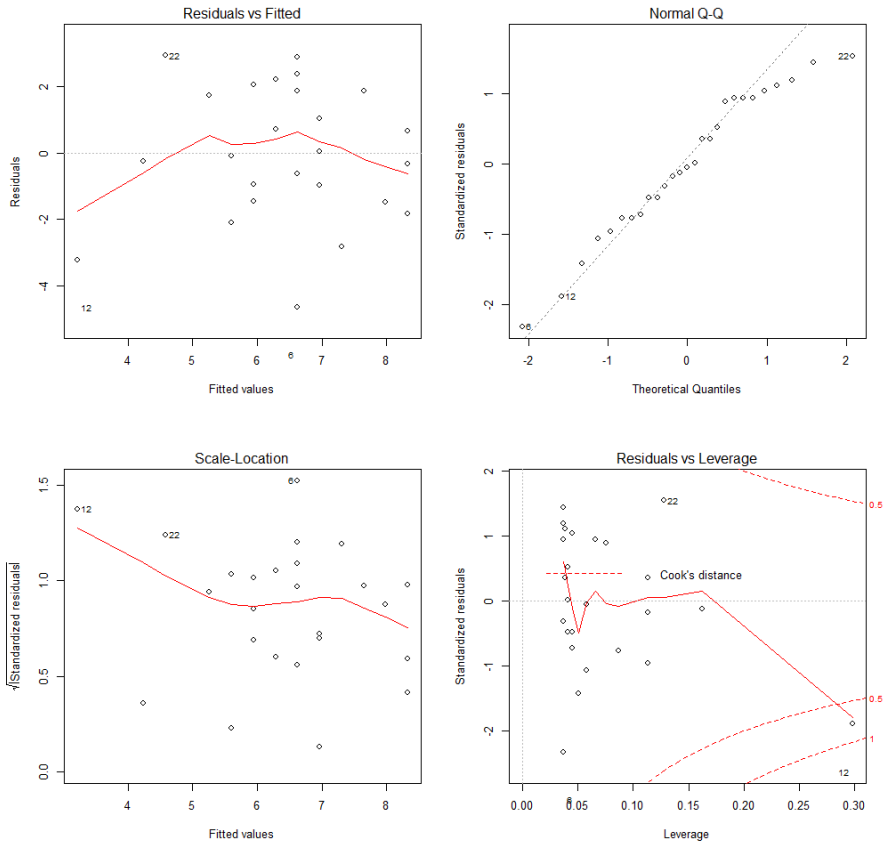
P-value is 0.3611. For a significance level of α = 5%, H_0 cannot be rejected. Thus, residuals can be considered normally distributed.

iii. Homogeneity of residuals variance:

In order to have homogeneous variances the Scale – Location diagram bellow should be a horizontal line with equally spread points. Again, although the assumption is not 100% violated it is not clear that we can procced with assuming that our linear regression model is perfect.

iv. Outliers

From the Residuals vs Leverage diagram bellow, we can see that there is one outlier that is influencing the regression results.



3. Fit a normal regression model in order to infer the grades of 2nd assignment using the 1st assignment grade and the number of report pages in 2nd assignment as explanatory variables. Describe the parameter estimates of the fitted model. Test whether this model fits better than the previous one.

```
Call:
lm(formula = assignment2 ~ assignment1 + size, data = grades)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0438 -0.9884 -0.2771  0.9359  3.1614

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.05590    1.24915   0.845 0.406295
assignment1  0.42575    0.17439   2.441 0.022383 *
size         0.15896    0.03709   4.286 0.000255 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.561 on 24 degrees of freedom
Multiple R-squared:  0.5977,    Adjusted R-squared:  0.5642
F-statistic: 17.83 on 2 and 24 DF,  p-value: 1.795e-05
```

The linear regression model gives us the line assignment2 = 1.05590 + 0.42575assignmnet 1 + 0.15896size.

This means that an increase of 1 grade in assignment 1 would give a 0.42575 increase in assignment 2 grade. Also, that a 1-page increase in the size of assignment 2 would give an increase of 0.15896 in assignment 2 grade. Finally, if someone would have a 0 grade in assignment 1 and would have a 0-paged assignment 2, would get 1.05590 as a grade. Obviously, this is not possible, but the prediction is very close to grade 0, that is the real grade that person would get.

For the p-values of each one of assignment 1 and size:

- H_0: true factor of assignment 1/size is equal to 0 ($\beta = 0$)
- H_1: true factor of assignment 1/size is not equal to 0

For a significance level of $\alpha = 5\%$, for both assignment 1 and size the null hypothesis is rejected. Meaning that their β values are not equal to 0. Note that when we check if $a\beta = 0$ for a variable we assume that all other variables are included in the model, thus assignment 1 and size hypotheses are checked independently.

Test modeling assumptions:

i. Linearity of the data:

Residuals vs Fitted diagram shows that assignment 2 and assignment 1 + size do not have a perfect linear relation. This assumption is not 100% violated but it is not perfect either.

ii. Normality of residuals:

From the Normal Q-Q diagram bellow we can say that the residuals are normally distributed. Ideally all observations should lay on the line. In order to further check this assumption, we can perform Shapiro-Wilk test.

Perform Shapiro Wilk test to check the

- H_0: residuals normally distributed
- H_1: residuals are not normally distributed

```
        Shapiro-Wilk normality test

data:  model1$residuals
W = 0.96323, p-value = 0.4364
```

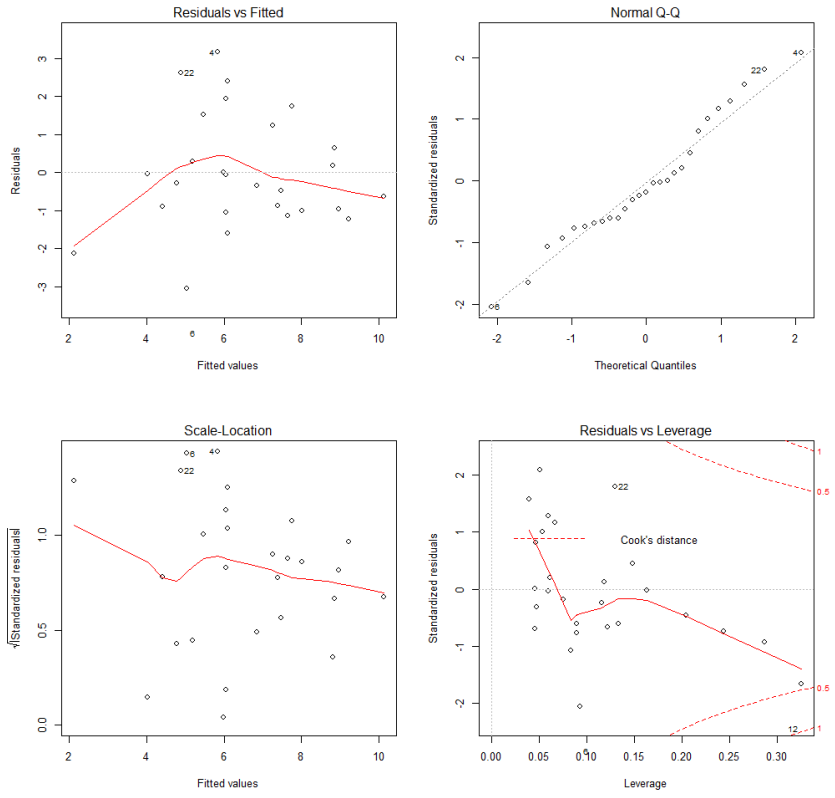
P-value is 0.4364. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected. Thus, residuals can be considered normally distributed.

iii. Homogeneity of residuals variance:

In order to have homogeneous variances the Scale – Location diagram bellow should be a horizontal line with equally spread points. This assumption stands better than the previous model. We cannot say that is violated either.

iv. Outliers

From the Residuals vs Leverage diagram bellow, we can see that there are no outliers.



In order to decide which model is better we will use ANOVA, since the two models are nested.

```
Analysis of Variance Table

Model 1: assignment2 ~ assignment1
Model 2: assignment2 ~ assignment1 + size
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     25 103.291
2     24  58.513  1    44.778 18.366 0.0002555 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F-value is very small. This indicates that by adding the size variable to the model there was statistically significant improvement. Null hypothesis would be that by adding the size variable the model was not improved, and alternative hypothesis that by adding the size variable the model was improved. As stated, we reject the null hypothesis in favor of the alternative.

5. Estimate the mean grade for a student with 1st assignment grade equal to 6 and 2nd assignment consisting of 10 pages. Give a 90% confidence interval.

```
> #estimate
> predict(model1, data.frame(assignment1 = 6, size = 10), interval = 'confidence', level=0.9)
       fit      lwr      upr
1 5.200013 4.538343 5.861684
```

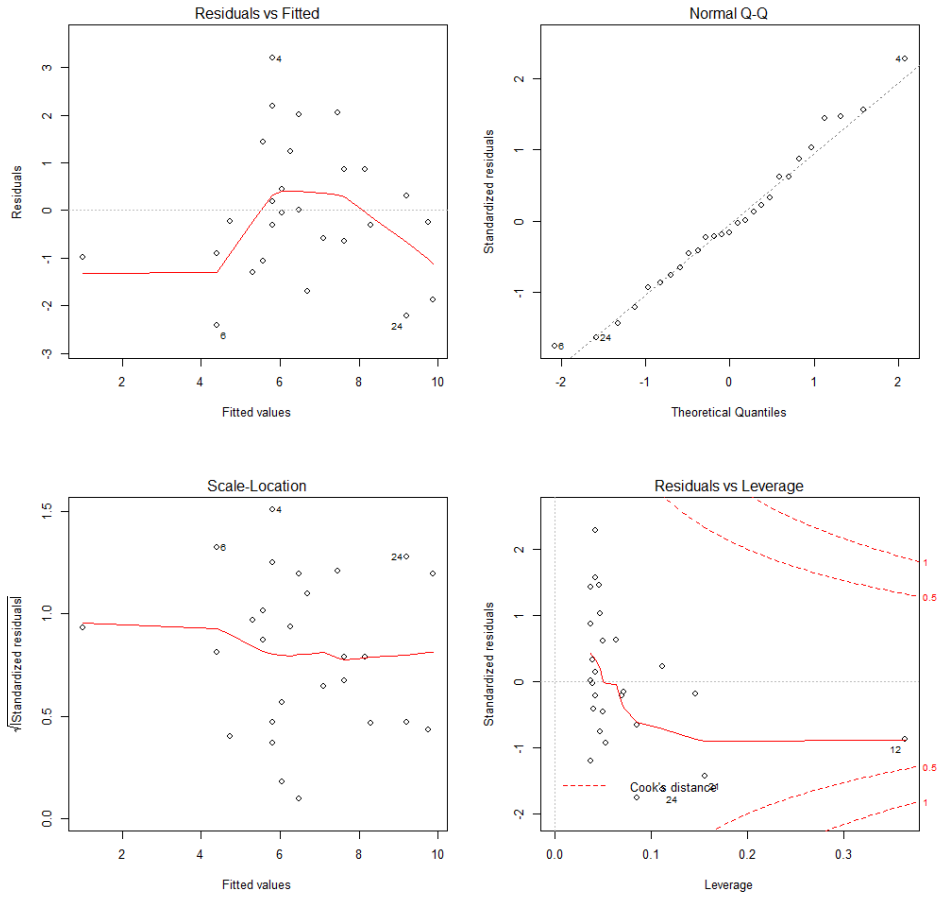
6. Predict the grade for a student with 1st assignment grade equal to 6 and 2nd assignment consisting of 10 pages. Give a 90% prediction interval.

```
> #predict
> predict(model1, data.frame(assignment1 = 6, size = 10), interval = 'predict', level=0.9)
       fit      lwr      upr
1 5.200013 2.447881 7.952146
```

7. Can you detect a better model? Present your findings in such a case.

```
Call:
lm(formula = assignment2 ~ sqrt(size), data = grades)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3988 -0.9443 -0.2242  0.8621  3.1891

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9898     0.8633   1.147    0.262
sqrt(size)    1.5246     0.2248   6.780 4.17e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.432 on 25 degrees of freedom
Multiple R-squared:  0.6478,    Adjusted R-squared:  0.6337
F-statistic: 45.97 on 1 and 25 DF,  p-value: 4.17e-07
```

A better model is the one that predicts based on the square root of the size of the assignment 2. P- value is too low indicating that β value could not be 0. Bellow are the diagrams that show no violation of assumtions.




```
> BIC(model)
[1] 122.7364
> BIC(model1)
[1] 110.6881
> BIC(model2)
[1] 103.8079
```

By calculating the BIC for each one of the three models we can see that the last one has the smallest value.

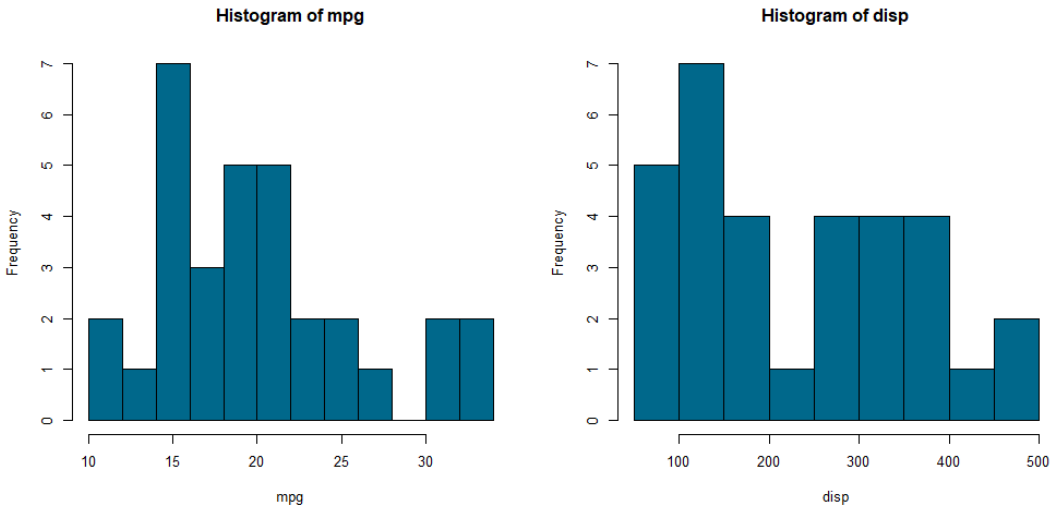
This means that it is the best among the three.

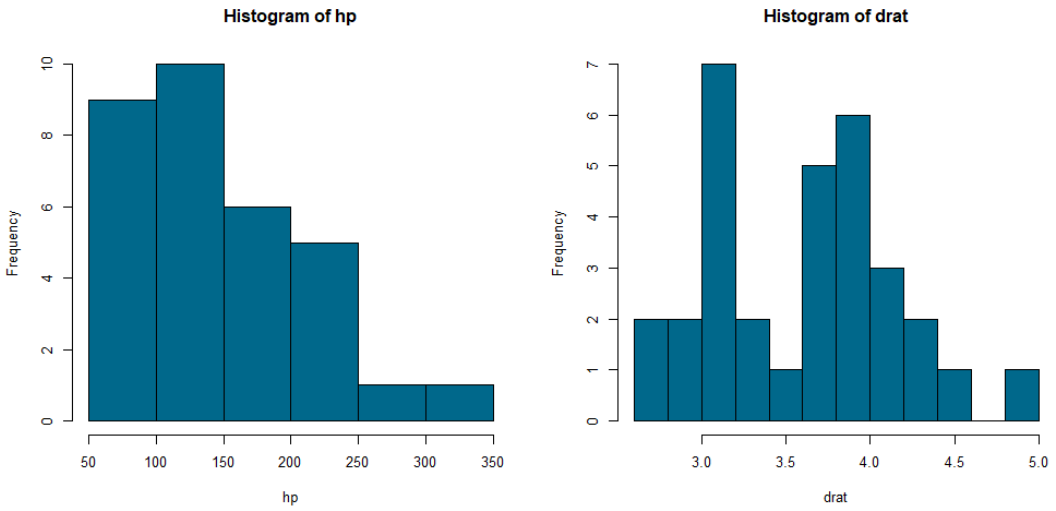
Exercise 6

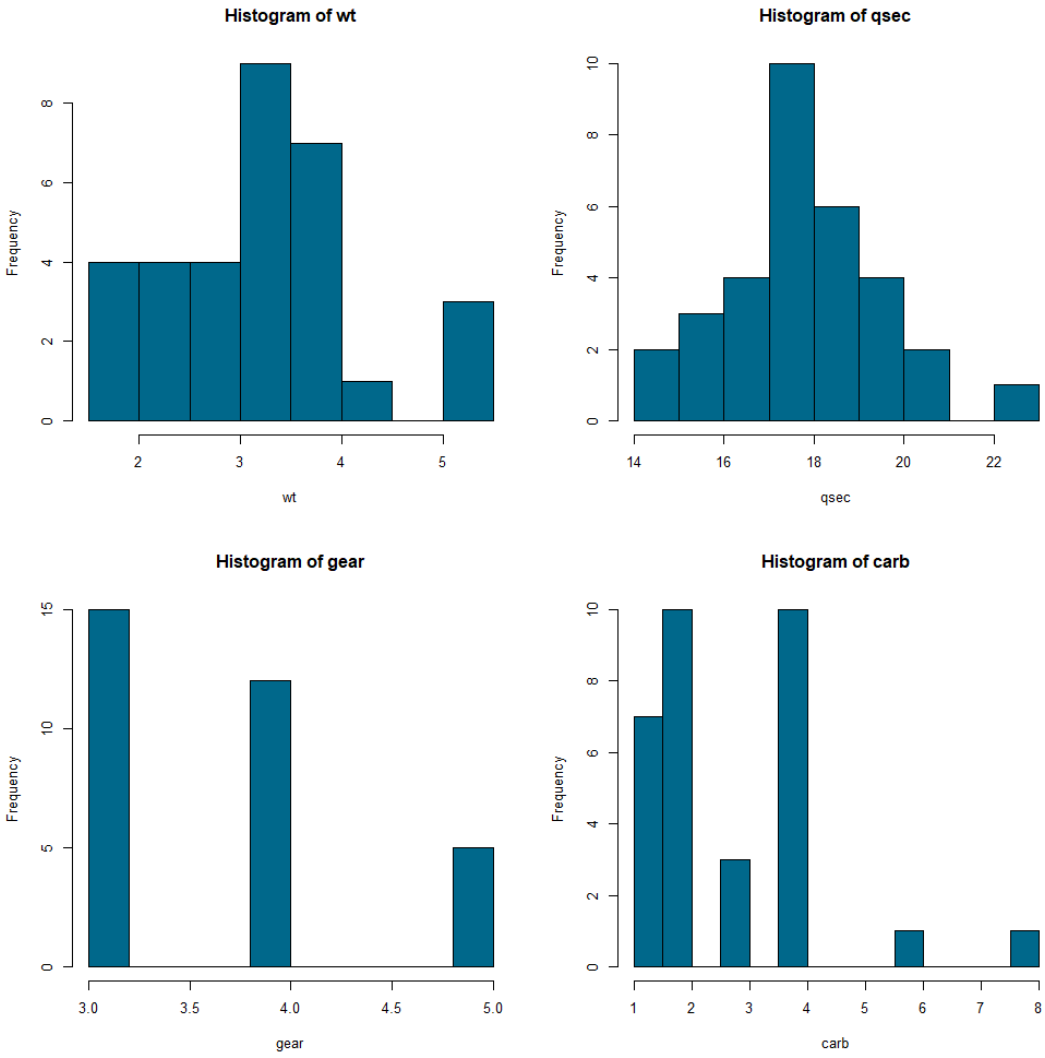
1. Report basic descriptive statistics for each variable and illustrate them on suitable diagrams.

```
mtcars

 11  Variables      32  Observations
---------------------------------------------------------------------------------------------------------------
mpg
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
      32        0       25    0.999    20.09    6.796    12.00    14.34    15.43    19.20    22.80    30.09    31.30

lowest : 10.4 13.3 14.3 14.7 15.0, highest: 26.0 27.3 30.4 32.4 33.9
---------------------------------------------------------------------------------------------------------------
cyl
       n  missing distinct
      32        0        3

Value          4     6     8
Frequency     11     7    14
Proportion 0.344 0.219 0.438
---------------------------------------------------------------------------------------------------------------
disp
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
      32        0       27    0.999    230.7    142.5    77.35    80.61   120.83   196.30   326.00   396.00   449.00

lowest :  71.1  75.7  78.7  79.0  95.1, highest: 360.0 400.0 440.0 460.0 472.0
---------------------------------------------------------------------------------------------------------------
hp
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
      32        0       22    0.997    146.7    77.04    63.65    66.00    96.50   123.00   180.00   243.50   253.55

lowest :  52  62  65  66  91, highest: 215 230 245 264 335
---------------------------------------------------------------------------------------------------------------
drat
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
      32        0       22    0.997    3.597   0.6099    2.853    3.007    3.080    3.695    3.920    4.209    4.314

lowest : 2.76 2.93 3.00 3.07 3.08, highest: 4.08 4.11 4.22 4.43 4.93
---------------------------------------------------------------------------------------------------------------
wt
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
      32        0       29    0.999    3.217    1.089    1.736    1.956    2.581    3.325    3.610    4.048    5.293

lowest : 1.513 1.615 1.835 1.935 2.140, highest: 3.845 4.070 5.250 5.345 5.424
---------------------------------------------------------------------------------------------------------------
qsec
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
      32        0       30        1    17.85    2.009    15.05    15.53    16.89    17.71    18.90    19.99    20.10

lowest : 14.50 14.60 15.41 15.50 15.84, highest: 19.90 20.00 20.01 20.22 22.90
---------------------------------------------------------------------------------------------------------------
vs
       n  missing distinct
      32        0        2

Value          0     1
Frequency     18    14
Proportion 0.562 0.438
---------------------------------------------------------------------------------------------------------------
am
       n  missing distinct
      32        0        2

Value          0     1
Frequency     19    13
Proportion 0.594 0.406
---------------------------------------------------------------------------------------------------------------
```

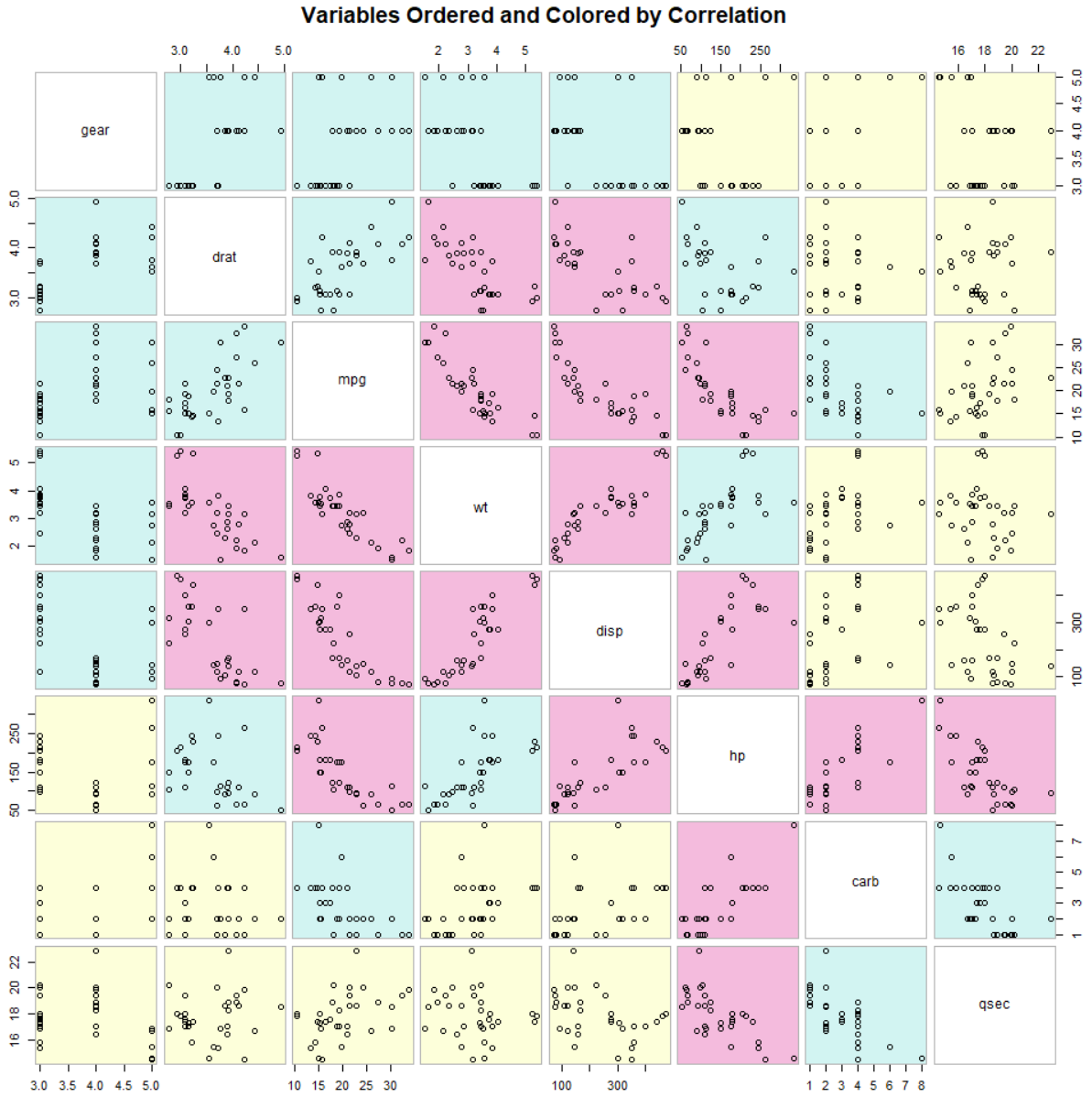
```
---------------------------------------------------------------------------
gear
       n  missing distinct     Info     Mean      Gmd
      32        0        3    0.841    3.688   0.7863

Value          3     4     5
Frequency     15    12     5
Proportion 0.469 0.375 0.156
---------------------------------------------------------------------------
carb
       n  missing distinct     Info     Mean      Gmd
      32        0        6    0.929    2.812    1.718

lowest : 1 2 3 4 6, highest: 2 3 4 6 8

Value          1     2     3     4     6     8
Frequency      7    10     3    10     1     1
Proportion 0.219 0.312 0.094 0.312 0.031 0.031
---------------------------------------------------------------------------
```






Histogram of wt
Frequency
0
2
4
6
8
2
3
4
5
wt
Histogram of qsec
Frequency
0
2
4
6
8
10
14
16
18
20
22
qsec
Histogram of gear
Frequency
0
5
10
15
3.0
3.5
4.0
4.5
5.0
gear
Histogram of carb
Frequency
0
2
4
6
8
10
1
2
3
4
5
6
7
8
carb

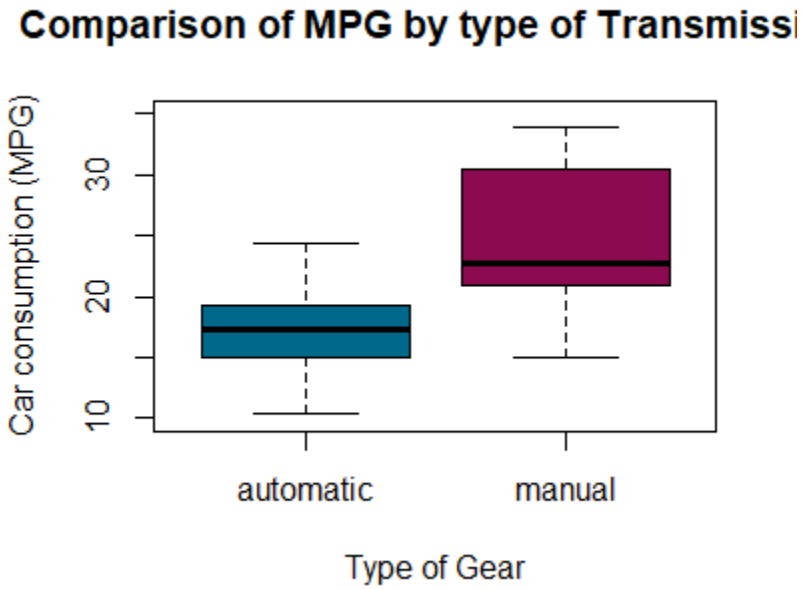
2. Produce all pairwise scatterplots for the numeric variables and compute the corresponding correlation coefficients.



Correlation coefficients

	mpg	disp	hp	drat	wt	qsec	gear	carb
mpg	1.0000000	0.8475514	0.7761684	0.68117191	0.8676594	0.41868403	0.4802848	0.5509251
disp	0.8475514	1.0000000	0.7909486	0.71021393	0.8879799	0.43369788	0.5555692	0.3949769
hp	0.7761684	0.7909486	1.0000000	0.44875912	0.6587479	0.70822339	0.1257043	0.7498125
drat	0.6811719	0.7102139	0.4487591	1.00000000	0.7124406	0.09120476	0.6996101	0.0907898
wt	0.8676594	0.8879799	0.6587479	0.71244065	1.0000000	0.17471588	0.5832870	0.4276059
qsec	0.4186840	0.4336979	0.7082234	0.09120476	0.1747159	1.00000000	0.2126822	0.6562492
gear	0.4802848	0.5555692	0.1257043	0.69961013	0.5832870	0.21268223	1.0000000	0.2740728
carb	0.5509251	0.3949769	0.7498125	0.09078980	0.4276059	0.65624923	0.2740728	1.0000000

3. Is there any difference in consumption between automatic and manual cars?



In order to check whether there is difference in consumption we will perform a t- test. But fist, we need to check that observations are normally distributed within each subset (automatic and manual) and if their variances are equal or not.

Perform Anderson Darling test to check whether observations are normally distributed.

- H_0: observations in automatic/manual are normally distributed
- H_1: observations in automatic/manual are not normally distributed

```
> #normality for manual
> ad.test(mtcars[mtcars$am == '0', 'mpg'])

        Anderson-Darling normality test

data:  mtcars[mtcars$am == "0", "mpg"]
A = 0.17192, p-value = 0.9166
```

P-value is 0.9166. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected. Thus, observations can be considered normally distributed.

```
> #normality for automatic
> ad.test(mtcars[mtcars$am == '1', 'mpg'])

        Anderson-Darling normality test

data:  mtcars[mtcars$am == "1", "mpg"]
A = 0.30016, p-value = 0.5298
```

P-value is 0.5298. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected. Thus, observations can be considered normally distributed.

Perform F test for variances.

- H_0: ratio of automatic, manual variances is 1

- H_1: ratio of automatic, manual variances is not 1

```
> #equality of variances
> res.ftest <- var.test(mtcars[mtcars$am == '0', 'mpg'], mtcars[mtcars$am == '1', 'mpg'])
> res.ftest

        F test to compare two variances

data:  mtcars[mtcars$am == "0", "mpg"] and mtcars[mtcars$am == "1", "mpg"]
F = 0.38656, num df = 18, denom df = 12, p-value = 0.06691
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1243721 1.0703429
sample estimates:
ratio of variances
         0.3865615
```

F-value is 0.06691. For a significance level of $\alpha = 5\%$, H_0 cannot be rejected. Thus, variances can be considered equal.

At last let's perform the t-test.

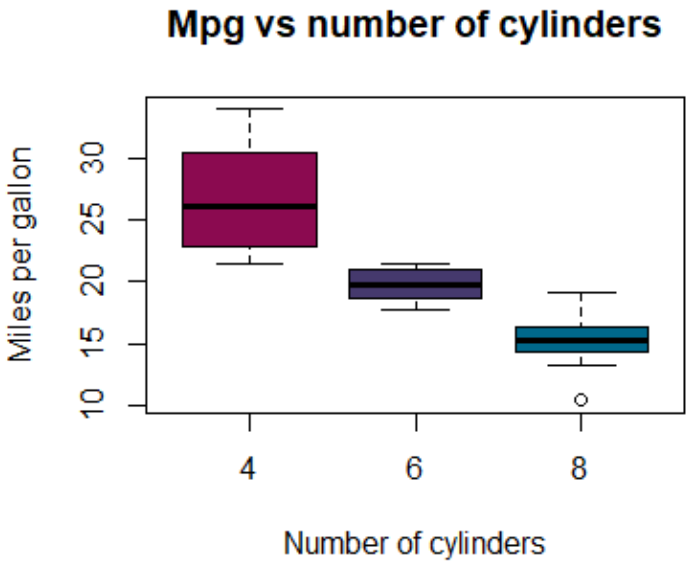
Perform F test for variances.

- H_0: difference in means of manual and automatic sets is equal to 0
- H_1: difference in means of manual and automatic sets is not equal to 0

```
        Two Sample t-test

data:  mtcars[mtcars$am == "0", "mpg"] and mtcars[mtcars$am == "1", "mpg"]
t = -4.1061, df = 30, p-value = 0.000285
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.84837  -3.64151
sample estimates:
mean of x mean of y
 17.14737  24.39231
```

F-value is 0.000285. For a significance level of $\alpha = 5\%$, H_0 can be rejected. Thus, there is indeed statistically significant difference in consumption between automatic and manual.

4. Is there any difference in consumption among cars with different number of cylinders?



We have three independent samples; therefore, we will test for any differences between the groups using ANOVA.

- H_0: means among groups are equal
- H_1: means among groups are not equal (at least one is different)

```
> fit <- aov(mpg~cyl,data=mtcars)
> fit
Call:
   aov(formula = mpg ~ cyl, data = mtcars)

Terms:
                     cyl Residuals
Sum of Squares  824.7846  301.2626
Deg. of Freedom        2        29

Residual standard error: 3.223099
Estimated effects may be unbalanced
> summary(fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
cyl          2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-value is too small. For a significance level of $\alpha = 5\%$, H_0 is rejected. Thus, there are differences among the groups. Now, let's check that the assumptions for performing ANOVA were not violated.

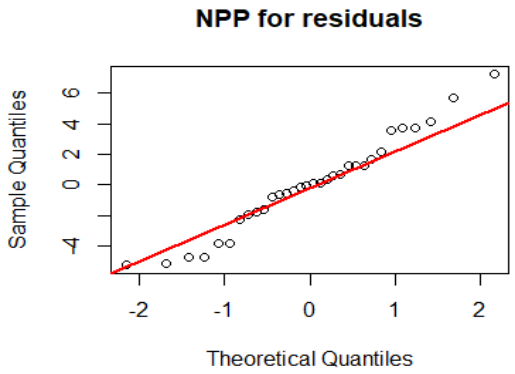
i. Normally distributed residuals

Perform Shapiro Wilk test to check the normality.

- H_0: residuals are normally distributed

- H_1: residuals are not normally distributed

```
        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.97065, p-value = 0.5177
```



P-value is 0.5177. For a significance level of $\alpha = 5\%$, H_0 is not rejected. Thus, we can say that residuals are normally distributed.

ii. Homogeneity of variances

Perform Bartlett and Fligner test to check homogeneity of variances among the three groups.

- H_0: variances of three groups are homogeneous
- H_1: variances are not homogeneous (at least one is different)

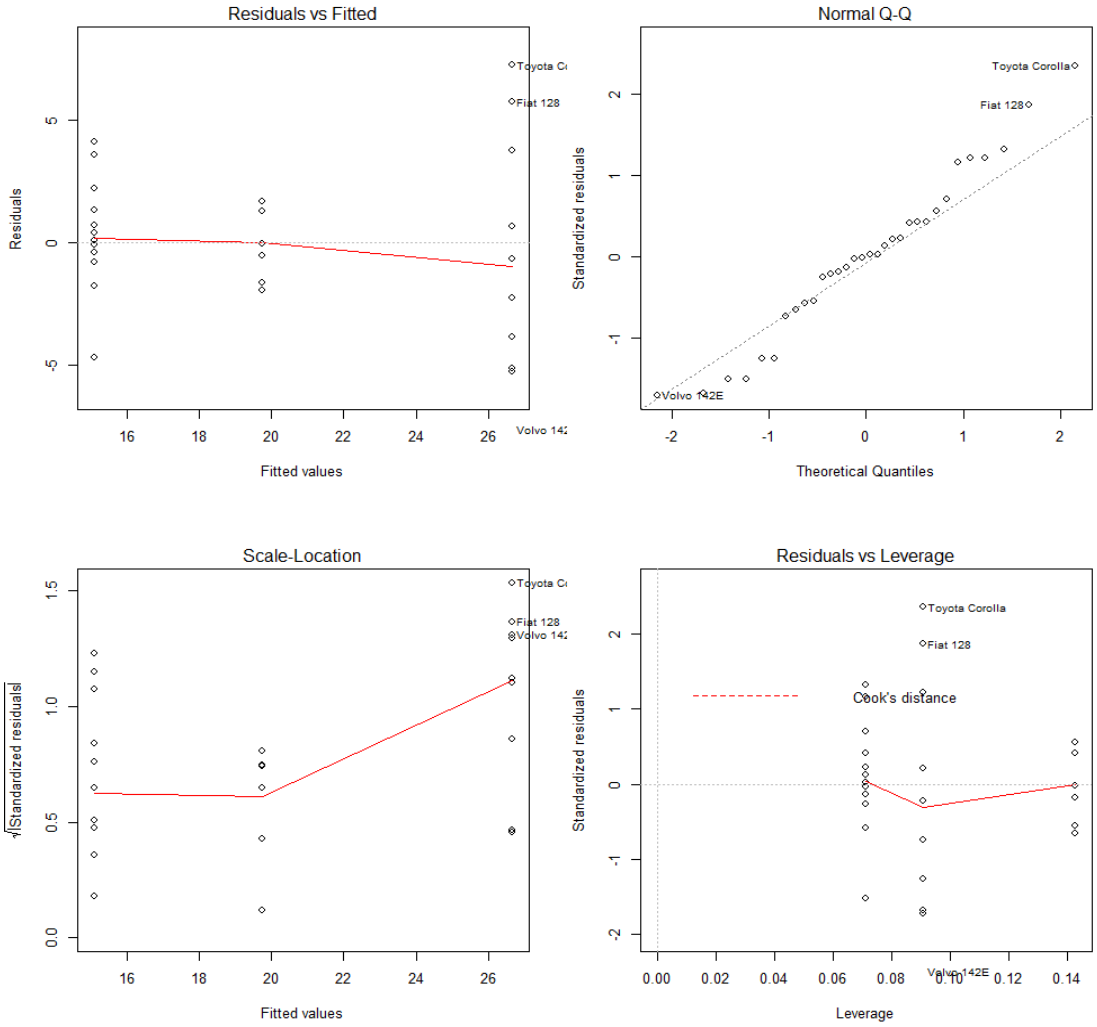
```
> bartlett.test(mpg~cyl,data=mtcars)

        Bartlett test of homogeneity of variances

data:  mpg by cyl
Bartlett's K-squared = 8.3934, df = 2, p-value = 0.01505

> fligner.test(mpg~cyl,data=mtcars)

        Fligner-Killeen test of homogeneity of variances

data:  mpg by cyl
Fligner-Killeen:med chi-squared = 6.8113, df = 2, p-value = 0.03319
```

P-value of Bartlett is in favor of H_1.

P-values of Fligner is in favor of H_1.

For a significance level of $\alpha = 5\%$, H_0 is rejected in both tests. Meaning that the assumption is violated. This is also shown in the Scale-Location diagram bellow.

iii. Diagnostic plots



Assumption of ANOVA on variance homogeneity is violated, meaning that we cannot draw any conclusion.

We should run a non-parametric test, as Kruskal-Wallis.

- H_0: means among groups are equal
- H_1: means among groups are not equal (at least one is different)

```
        Kruskal-Wallis rank sum test

data:  mpg by cyl
Kruskal-Wallis chi-squared = 25.746, df = 2, p-value = 2.566e-06
```

P-value of Kruskal-Wallis test is too small, thus in favor of rejecting H_0, for a significance level of $\alpha = 5\%$.

That means that there are statistically significant differences among the three groups.

5. Fit and interpret the full regression model using all available explanatory variables. According to the Bayesian Information Criterion, which variables mostly effect consumption? Check all modelling assumptions and interpret the selected model.

```
Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4734 -1.3794 -0.0655  1.0510  4.3906

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.81984   16.30602   1.093   0.2875
cyl6        -1.66031    2.26230  -0.734   0.4715
cyl8         1.63744    4.31573   0.379   0.7084
disp         0.01391    0.01740   0.799   0.4334
hp          -0.04613    0.02712  -1.701   0.1045
drat         0.02635    1.67649   0.016   0.9876
wt          -3.80625    1.84664  -2.061   0.0525 .
qsec         0.64696    0.72195   0.896   0.3808
vs1          1.74739    2.27267   0.769   0.4510
am1          2.61727    2.00475   1.306   0.2065
gear         0.76403    1.45668   0.525   0.6057
carb         0.50935    0.94244   0.540   0.5948
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.582 on 20 degrees of freedom
Multiple R-squared:  0.8816,    Adjusted R-squared:  0.8165
F-statistic: 13.54 on 11 and 20 DF,  p-value: 5.722e-07
```

Getting the p-value of one variable at a time, we cannot reject the null hypothesis that the estimate could be equal to zero (if all other variables are included in the model).

Using the Bayesian Information Criterion, we will try to find a better model. We will start with all variables and gradually remove until BIC is no longer decreasing.

Best model is:

```
Step:  AIC=67.17
mpg ~ wt + qsec + am
```

According to BIC wt, qsec and am are the variables that affect consumption the most.

In order to decide which model is better we will use ANOVA, since the two models are nested.

```
Analysis of Variance Table

Model 1: mpg ~ wt + qsec + am
Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 169.29
2     20 133.32  8    35.962 0.6743 0.7083
```

F-value is greater than the significance level α = 5%. Null hypothesis would be that by adding all the extra variables the model was not improved, and alternative hypothesis that by adding the model was improved. Null hypothesis cannot be rejected. The mpg~wt+qsec+am model is better.

```
Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4811 -1.5555 -0.7257  1.4110  4.6610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6178     6.9596   1.382 0.177915
wt           -3.9165     0.7112  -5.507 6.95e-06 ***
qsec          1.2259     0.2887   4.247 0.000216 ***
am1           2.9358     1.4109   2.081 0.046716 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497,    Adjusted R-squared:  0.8336
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

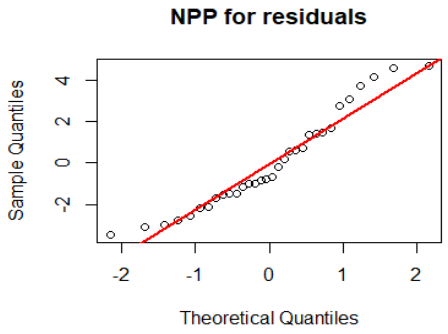
The line would be mpg = 9.6178 - 3.9165wt + 1.2259qsec + 2.9358am
P-values ae too small indicating that variables are significant for the model.

Check assumptions:

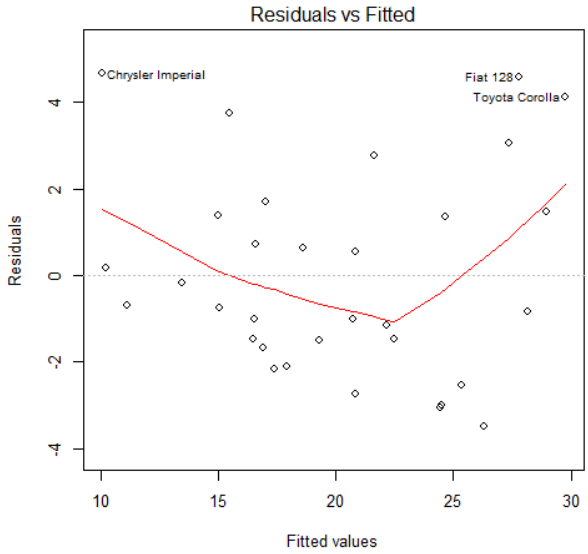
Perform Shapiro Wilk test to check whether residuals are normally distributed.

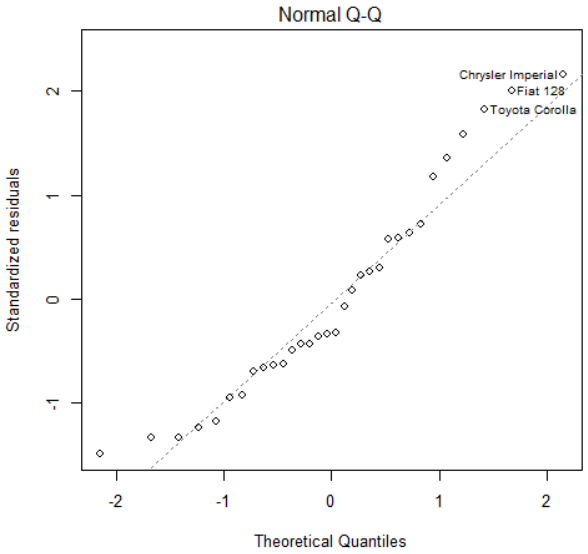
- H_0: residuals are normally distributed
- H_1: residualsare not normally distributed

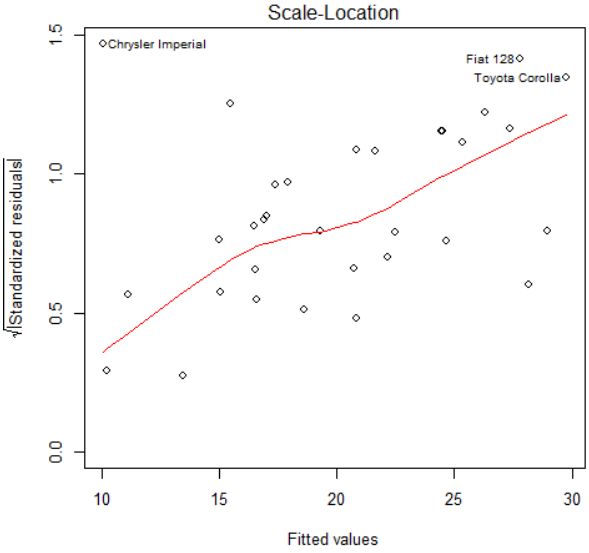
```
        Shapiro-Wilk normality test

data:  fit_bic$residuals
W = 0.9411, p-value = 0.08043
```

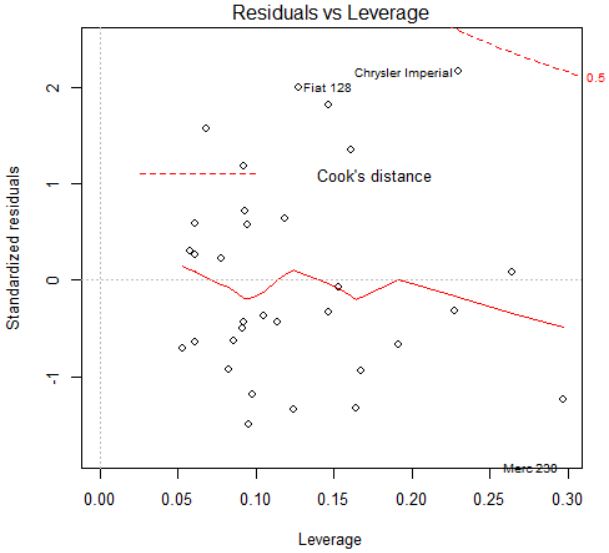


P-value is 0 .08043. For a significance level of α = 5%, H_0 cannot be rejected. Thus, residuals can be considered normally distributed.









Sofia Baltzi
10/12/2019