Probability and Statistics for Data Analysis

Assignment 3

Vasileios Galanos
MSc in Data Science(PT)
p3351902

December 11, 2019

Exercise 1.

1.

We will import the data into R using the following code:

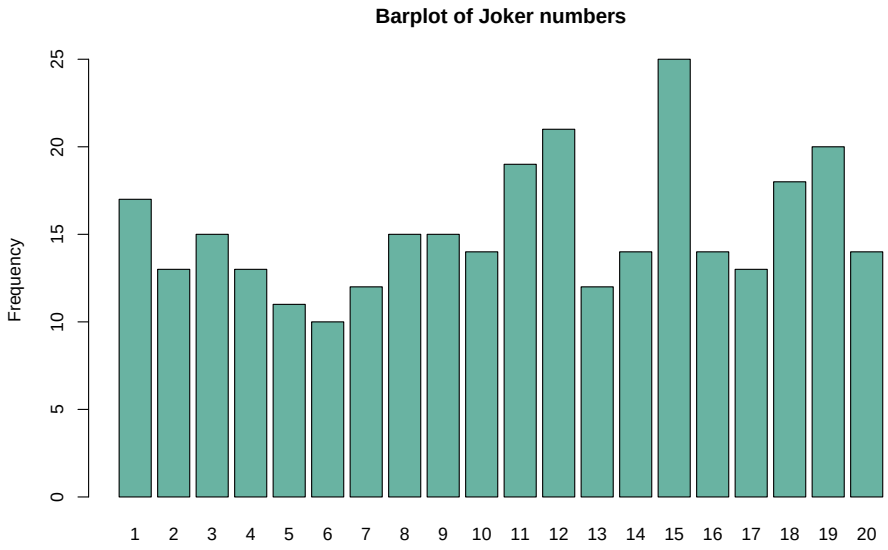
```
library("xlsx")
joker_numbers <- vector()
for (y in c(2017,2018,2019)){
  file <- read.xlsx(paste('Joker_',y,'.xlsx',sep=''), sheetIndex = 1
                    ,startRow=4, colIndex = 8,header = 'FALSE'
                    ,colClasses = c('numeric'))
  names(file)[1] <- 'joker'
  file = na.omit(file)
  joker_numbers <- c(joker_numbers,as.numeric(file$joker))
}
```

So for the period 2017-2019 we get the following table that represents the frequency of the winning joker numbers

Joker	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Freq	17	13	15	13	11	10	12	15	15	14	19	21	12	14	25	14	13	18	20	14

We can visualize the data using a **barplot**

```
barplot(joker_freqs
     , col = "#69b3a2"
     , main = "Barplot of Joker numbers"
     , ylab = 'Frequency'
)
```



2.

To test whether or not this is a fair lottery, we will simply test whether the theoretical model of the uniform distribution fits the data that we observed appropriately. We choose the uniform distribution because all the intervals of the same length on the distribution's support set are equally probable:

For $x \in [1, 2, \ldots, 20]$

$$f(x) = \frac{1}{20}$$

We will using Pearson's chi-squared test as a goodness of fit test that examines whether our observed frequency distribution differs from the theoretical or not.

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = n \sum_{i=1}^{k} \frac{(\frac{O_i}{n} - p_i)^2}{p_i} \sim \mathcal{X}^2_{k-l-1}$$

In our case:
$n = 305$
$k = 20$
$l = 0$ since we don't estimate any parameters in the uniform distribution
$p_i = \frac{1}{20} = 0.05$, which is the theoretical expected fraction of data for bin i
O_i : the observed data points in bin i, which we got from the frequency table earlier

Thus, using the following code we get:

```
sum <- 0
n <- length(joker_numbers)
for (k in (1:20)){
  o <- sum(joker_numbers == k)
  sum <- sum + (o/n - 0.05)^2/0.05
}
sprintf('%.5f',n*sum)
[1] "17.29508"
```

Now we use the chi-squared statistic that we got, to calculate a p-value by comparing the value of the statistic to a chi-squared distribution. The number of degrees of freedom is equal to the number of bins (20), minus the reduction in degrees of freedom (1).

```
> pchisq(17.29508,df = 19, lower.tail = FALSE)
[1] 0.5698844
```

Alternatively, we could have used R to calculate the p-value as follows:

```
m <- chisq.test(joker_freqs)
cat("The p-value is ", m$p.value, "\n")
The p-value is  0.5698842
```

Our hypothesis H_0 is that the observed data indeed came from a uniform distribution, against the alternative hypothesis H_1 that the data did not came from the uniform distribution. We got a p-value of 0.56, which is quite large, so we fail to reject the NULL hypothesis.

In conclusion, indeed this is a fair lottery.

Exercise 2.

1.

First, we import the data in R and create a separate vector for each type of drug treatment:

```
drug_response <- read.table('drug_response_time.txt',header = TRUE)

drug_A <- drug_response[drug_response$drug == 'A','time']
drug_B <- drug_response[drug_response$drug == 'B', 'time']
```

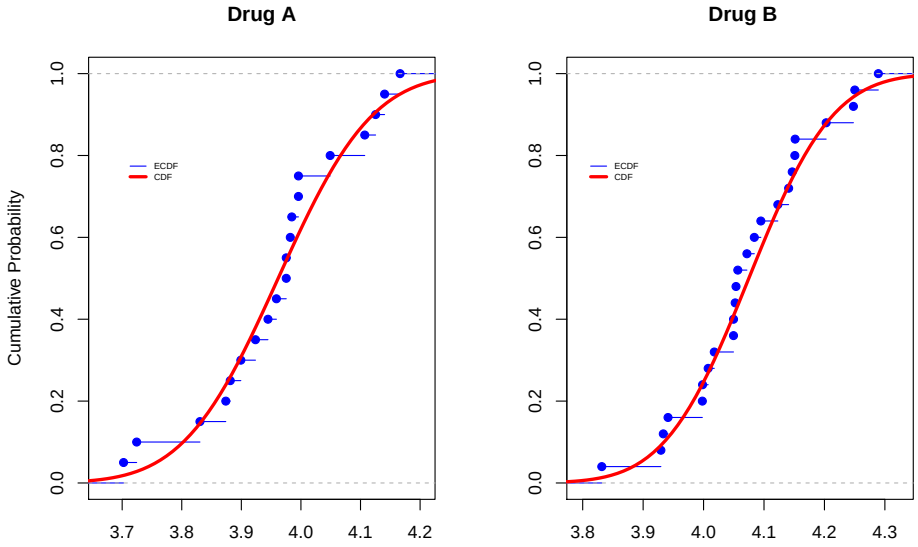
To test the normality assumption of response time within each treatment we will perform a Kolmogorov-Smirnov (KS) goodness of fit test. We will examine for each treatment, whether or not the sample of the observed data that we got, came from a specific continous distribution. In our case, we will examine whether it came from the normal distribution $\mathcal{N}(\bar{x}, s^2)$, where $\bar{x}$ is the sample mean of our data, and s^2 is the sample variance.
So we have the following hypothesis for treatment with $drug_k$:

$$H_0(\textit{Null hypothesis}) \text{ :the sample } drug_k \sim \mathcal{N}(\bar{x_k}, s_k^2)$$
$$H_1(\textit{alternative}) \text{ :Not } H_0$$

So we proceed to perfom the tests:

```
> ks.test(drug_A,"pnorm",mean(drug_A),sd(drug_A))

One-sample Kolmogorov-Smirnov test

data:  drug_A
D = 0.14218, p-value = 0.7623
alternative hypothesis: two-sided

> ks.test(drug_B,"pnorm",mean(drug_B),sd(drug_B))

One-sample Kolmogorov-Smirnov test

data:  drug_B
D = 0.088333, p-value = 0.9802
alternative hypothesis: two-sided
```

We get large p-values (0.76, 0.98), thus we fail to reject the Null hypothesis. So, we have statistical evidence that the data came from a normal distribution. To better visualize the KS test, we will visualize the Cumulative Density Function of the Normal Distribution vs the Empirical:

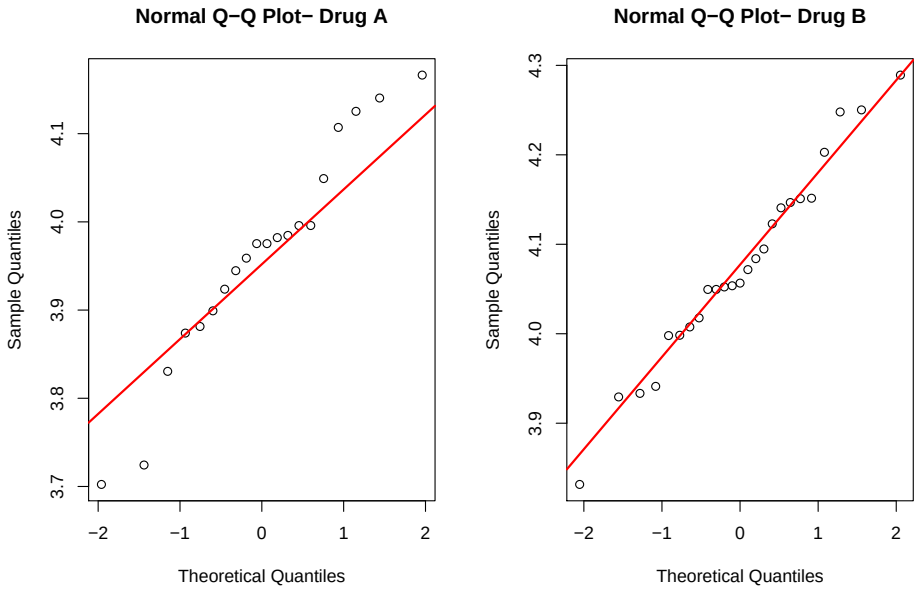


The results that we got from the KS tests were sufficient, but we will perfom some more tests as well:

```
> ad.test(drug_A)

Anderson-Darling normality test

data:  drug_A
A = 0.35519, p-value = 0.4244

> ad.test(drug_B)

Anderson-Darling normality test

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

```
> shapiro.test(drug_A)

Shapiro-Wilk normality test

data:  drug_A
W = 0.95593, p-value = 0.4662

> shapiro.test(drug_B)

Shapiro-Wilk normality test

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

Lastly, we will perform a QQ plot for the two treatments to visually check the normality of our data



2.

To test the homogeneity of the variances between treatments, we will perform a series of tests where all have they all same hypothesis:

$$H_0(Null\ hypothesis) : \sigma^2_{drug_A} = \sigma^2_{drug_B}$$
$$H_1(alternative) : \sigma^2_{drug_A} \neq \sigma^2_{drug_B}$$

```
> bartlett.test(time~drug,data = drug_response)

Bartlett test of homogeneity of variances

data:  time by drug
Bartlett's K-squared = 0.3405, df = 1, p-value = 0.5595

Fligner-Killeen test of homogeneity of variances

data:  time by drug
Fligner-Killeen:med chi-squared = 0.13374, df = 1, p-value = 0.7146

> library(car)
Loading required package: carData
> leveneTest(time~drug,data = drug_response)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.0855 0.7714
      43
```

So, in all the tests we fail the reject the Null hypothesis of the homogeneity of the variances between treatments. Thus, we have statistical evidence to assume an equal variance of response time between treatments.

3.

To test whether there is a difference in mean response time between the two drugs, we must test the following hypothesis:

$$H_0(Null\ hypothesis) : \mu_{drug_A} = \mu_{drug_B}$$
$$H_1(alternative) : \mu_{drug_A} \neq \mu_{drug_B}$$

We have two independent normal samples with uknown equal variances, so we can perform a t.test with the following test statistic:

$$T = \frac{\bar{X_1} - \bar{X_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-1}$$

In R,

```
> t.test(time~drug,data = drug_response, var.equal=T)
Two Sample t-test

data:  time by drug
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 in group A mean in group B
       3.961835        4.074892
```

We got a $p-value = 2P(t_{n-1} \geq |T|) = 0.002296 < 0.05$, thus it is highly unlikely to observe this statistic under the assumption that the means are equal. We reject the Null hypothesis.
We are 95% confident that there is a difference in mean response time between the two drugs.

Exercise 3.

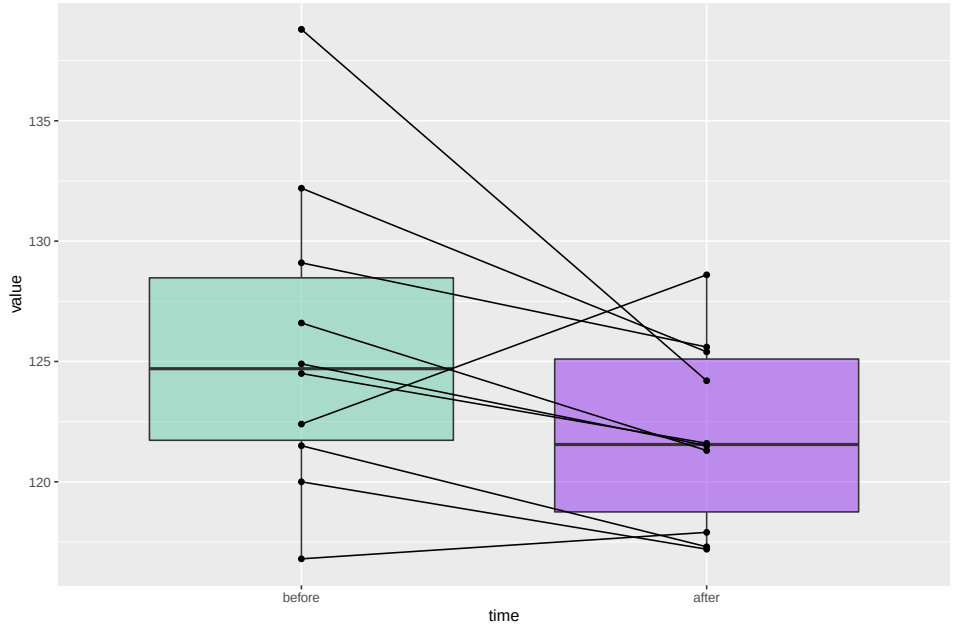
1.

We first import and visualize the data as follows:

```
before <- c(121.5, 122.4, 126.6, 120.0, 129.1, 124.9, 138.8, 124.5, 116.8, 132.2)
after <- c(117.3, 128.6, 121.3, 117.2, 125.6, 121.5, 124.2, 121.6, 117.9, 125.4)

#Visualization of the data
library(ggplot2)
d <- data.frame(before = before, after = after)
d$obs <- 1:nrow(d)
d2 <- tidyr::gather(d, time, value, -obs)
```

```
ggplot(d2, aes(time, value)) +
  geom_boxplot(fill = c('mediumaquamarine','purple2'), alpha = 0.5) +
  geom_point() +
  geom_line(aes(group = obs)) +
  scale_x_discrete(limits = c('before', 'after'))
```

In the next figure we see two boxplots that represent the two groups of data and the connection between the measurement of the systolic blood pressure before receiving the new drug and after receiving it. Visually, the median systolic blood pressure appears to be lower after receiving the drug.



We also report the summary statistics:

```
> summary(d[,c(1,2)])
     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
```

2.

Now, to test the claim of the pharmaceutical company, that their treatment reduces systolic blood pressure levels, we first state our hypothesis:

Let μ_d be the mean difference between between the paired sample before the treatment and after the treatment

$$H_0(\textit{Null hypothesis}) : \mu_d = \mu_d$$
$$H_1(\textit{alternative}) : \mu_d > 0 \quad \text{(upper-tailed)}$$

We run the test statistic for our paired sample and we get:

```
 > t.test(before,after, paired = TRUE, alternative="greater", conf.level = 0.99)

Paired t-test

data:  before and after
t = 2.1557, df = 9, p-value = 0.02974
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 -1.117963       Inf
sample estimates:
mean of the differences
                   3.62
```

We notice that 0 is in the 99% confidence interval.

The p-value $= P(T > t) \approx 0.02974$, so we reject the Null hypothesis. Thus, we have statistical evidence to believe that the drug reduces the systolic blood pressure, so we confirm the claim.

Exercise 4.

1.

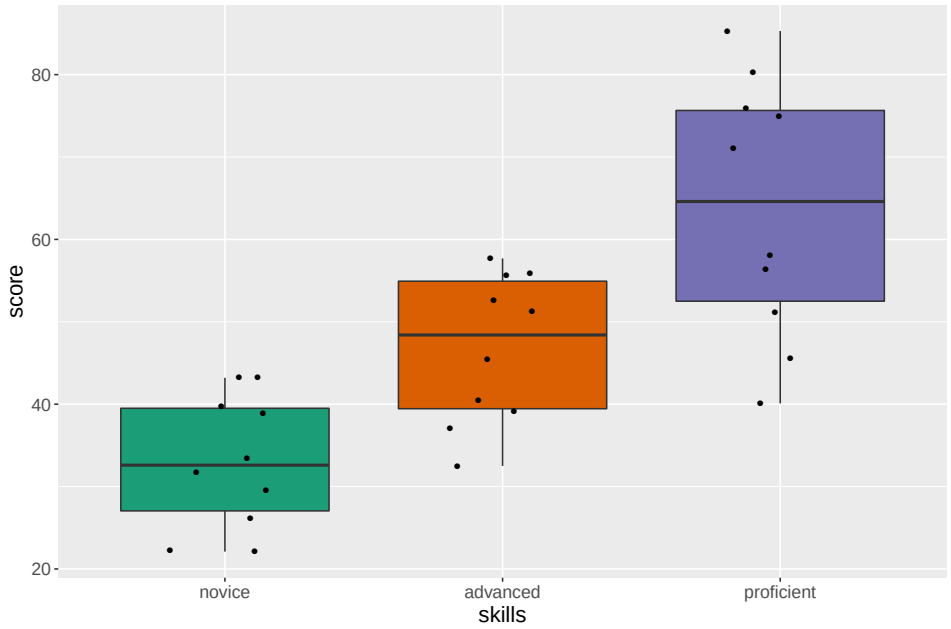
We first import the data in R,

```
a <- data.frame(skills = 'novice', score = c(22.10, 22.30 , 26.20
  , 29.60 , 31.70 , 33.50 , 38.90 , 39.70 , 43.20 , 43.20))
b <- data.frame(skills = 'advanced',score = c(32.50, 37.10, 39.10
, 40.50, 45.50, 51.30, 52.60, 55.70, 55.90, 57.70))
c <- data.frame(skills = 'proficient',score = c(40.10, 45.60, 51.20
, 56.40, 58.10, 71.10, 74.90, 75.90, 80.30, 85.30))

poker_players <- rbind(a, b, c)
```

Using the **psych** packages we provide the following summaires per group:

```
library(psych)
describeBy(poker_players$score, poker_players$skills, mat = TRUE)

> describeBy(poker_players$score, poker_players$skills, mat = F)

 Descriptive statistics by group
group: novice
   vars  n  mean   sd median trimmed   mad  min  max range  skew
X1    1 10 33.04 8.03   32.6   33.14 10.01 22.1 43.2  21.1 -0.06
------------------------------------------------------------------
group: advanced
   vars  n  mean   sd median trimmed   mad  min  max range skew
X1    1 10 46.79 9.03   48.4   47.21 11.42 32.5 57.7  25.2 -0.2
------------------------------------------------------------------
group: proficient
   vars  n  mean    sd median trimmed   mad  min  max range  skew
X1    1 10 63.89 15.62   64.6   64.19 18.31 40.1 85.3  45.2 -0.12
```

Next, we can visualize our data using **boxplots**:



We can clearly see the difference between the categories of our data. It appears that the more skilled a poker player is, the higher their score on in their ability to recall previous cards.

2.

To test whether there is a difference among the groups in the ability of recalling previous cards, we will use the ANOVA (Analysis of Variance)

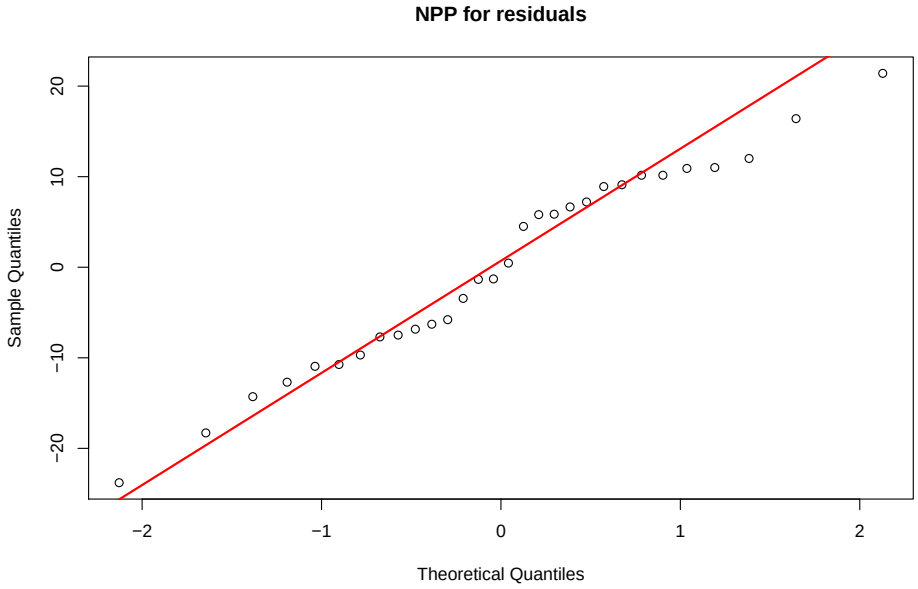
```
fit<-aov(score~factor(skills), data = poker_players)
fit
summary(fit)
>   summary(fit)
              Df Sum Sq Mean Sq F value   Pr(>F)
factor(skills)  2   4777    2389   18.37 9.21e-06 ***
Residuals      27   3511     130
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

. We have strong statistical evidence that there is a difference between the means. However, we must now, test the assumptions of the ANOVA.

Firstly, we will test the normality assumptions in the residuals using the hypothesis:

$$H_0(Null\ hypothesis): \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$
$$H_1(alternative): \quad Not \quad H_0$$

```
Shapiro-Wilk normality test

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



We fail to reject the Null hypothesis, so the assumption of normality of the residuals stands.

Next we check for the homogeneity of variance between the groups:

$$H_0(Null\ hypothesis): \quad \sigma^2_{novice} = \sigma^2_{advanced} = \sigma^2_{proficient}$$
$$H_1(alternative): \quad Not \quad H_0$$

```
>   bartlett.test(score~factor(skills), data = poker_players)

Bartlett test of homogeneity of variances

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

>   fligner.test(score~factor(skills), data = poker_players)

Fligner-Killeen test of homogeneity of variances

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

>   library(car)
>   leveneTest(score~factor(skills), data = poker_players)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value   Pr(>F)
group  2  5.9039 0.007464 **
      27
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The majority of the tests (2 out of 3) reject the Null hypothesis, so we cannot accept the assumption of the homogeneity of variances between the groups and the assumptions of ANOVA are violated. We proceed with a non-parametric approach: The Kruskal-Wallis one way ANOVA.

```
>   kruskal.test(score~factor(skills), data = poker_players)

Kruskal-Wallis rank sum test

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

We get a p-value = 0.0001677, thus we reject the Null Hypothesis of mean equality across the groups. At least one of the samples dominates the rest.

3.

We will now perform all pairwise t.tests to test the following hypotheses:

$H_0(\textit{Null hypothesis}) : \mu_k = \mu_j$

$H_1(\textit{alternative}) : \mu_k \neq \mu_j$ where $k, j \in \{\text{novice}, \text{intermediate}, \text{proficient}\}$

Novice vs Intermediate:

```
>   t.test(a$score,b$score, var.equal = F)

Welch Two Sample t-test

data:  a$score and b$score
t = -3.5975, df = 17.759, p-value = 0.002096
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.787791  -5.712209
sample estimates:
mean of x mean of y
    33.04     46.79
```

We reject the Null Hypothesis, so we have statistical evidence that there is a differenc in means between these two groups.

Novice vs Proficient:

```
>   t.test(a$score,c$score)

Welch Two Sample t-test

data:  a$score and c$score
t = -5.5537, df = 13.449, p-value = 8.235e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -42.80993 -18.89007
sample estimates:
mean of x mean of y
    33.04     63.89
```

We reject the Null Hypothesis, so we have statistical evidence that there is a differenc in means between these two groups.

Intermediate vs Proficient:

```
>   t.test(b$score,c$score)

Welch Two Sample t-test

data:  b$score and c$score
t = -2.9969, df = 14.411, p-value = 0.009361
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -29.305438  -4.894562
sample estimates:
mean of x mean of y
    46.79     63.89
```

We reject the Null Hypothesis, so we have statistical evidence that there is a difference in means between these two groups.

Let $A_i = \{$ type I error at the i-th test $\}$, $i \in 1, 2, 3$
Then, P(at least one type I error) $= P(A_1 \cup A_2 \cup A_3)$. But A_i are not independent from each other, thus we cannot calculate the probability exactly. We can find an upper limit, using Boole's inequality, though:

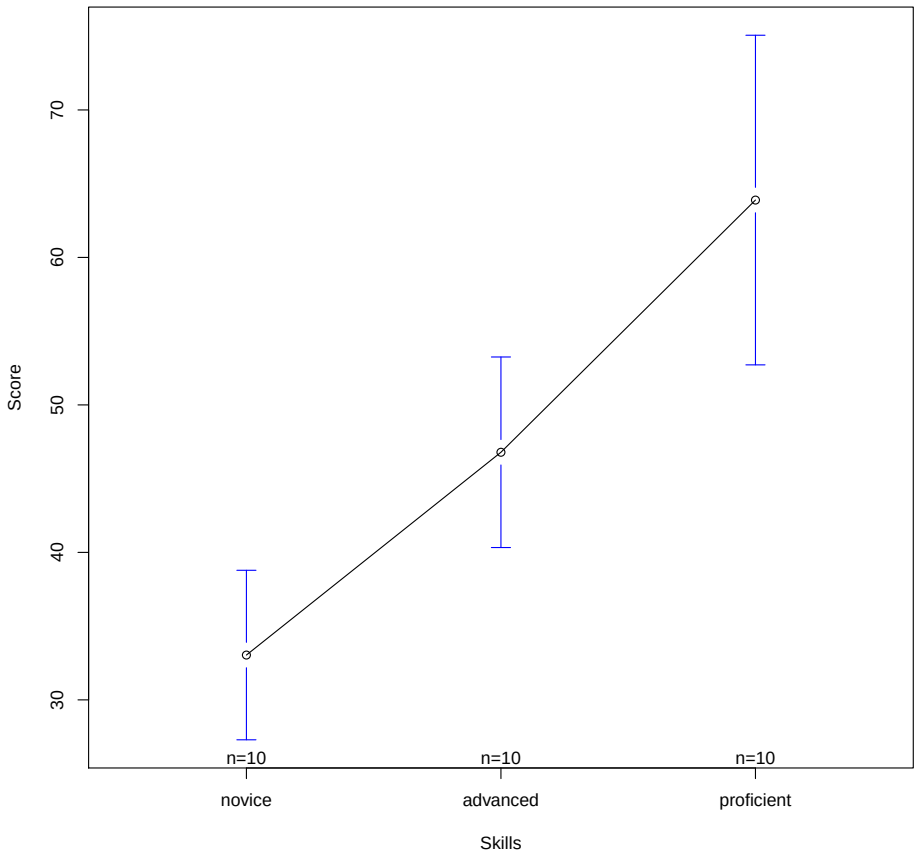
$$P(\text{at least one type I error}) \leq \sum_{i=1}^{3} P(A_i) = \sum_{i=1}^{3} \alpha = 3\alpha$$

So for 3 t-tests with significance level 5% the probability of at least one type I error can be up to 15%

4.

First of all, let's visualize the difference between the three groups:

```
library(gplots)
plotmeans(score~factor(skills), data = poker_players
          ,xlab="Skills", ylab="Score")
```



We can see that there is certainly a difference between all the pairs. Now to properly identify which groups are different at significance level 5%, we can use the Tukey HSD method on the ANOVA model that we created earlier:

```
>   TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = score ~ factor(skills), data = poker_players)

$`factor(skills)`
                      diff       lwr      upr     p adj
advanced-novice     13.75  1.105524 26.39448 0.0310169
proficient-novice   30.85 18.205524 43.49448 0.0000054
proficient-advanced 17.10  4.455524 29.74448 0.0065079
```

Clearly there is significant difference between all the pairs, especially between proficient and novice players.

Furthermore, we can perform p-value adjusted pairwise tests, using criteria like Bonferroni and Holm.

```
>   pairwise.t.test(poker_players$score, factor(poker_players$skills), p.adjust.metho

Pairwise comparisons using t tests with pooled SD
data:  poker_players$score and factor(poker_players$skills)
           novice  advanced
advanced   0.0358  -
proficient 5.6e-06 0.0071

P value adjustment method: bonferroni
>   pairwise.t.test(poker_players$score, factor(poker_players$skills), p.adjust.metho

Pairwise comparisons using t tests with pooled SD
data:  poker_players$score and factor(poker_players$skills)
           novice  advanced
advanced   0.0119  -
proficient 5.6e-06 0.0048
P value adjustment method: holm
```

We confirm the differences between all the groups, as before.

Next, using the Least Significant Differences (LSD) method,

```
>   DFE<-fit$df.residual
>   MSE<-deviance(fit)/DFE
>   library(agricolae)
>   print(LSD.test(poker_players$score, factor(poker_players$skills)
                     ,DFerror=DFE,MSerror=MSE,
+                   p.adj="bonferroni"))
$statistics
   MSerror Df     Mean       CV  t.value      MSD
  130.0386 27 47.90667 23.80346 2.552459 13.01697

$parameters
        test  p.ajusted                        name.t ntr alpha
  Fisher-LSD bonferroni factor(poker_players$skills)   3  0.05

$groups
           poker_players$score groups
proficient               63.89      a
advanced                 46.79      b
novice                   33.04      c
```

Similarly, performing the test with Scheffe method:'

```
>   print(scheffe.test(poker_players$score, factor(poker_players$skills)
                              ,DFerror=DFE,MSerror=MSE))
$statistics
   MSerror Df        F     Mean       CV  Scheffe CriticalDifference
  130.0386 27 3.354131 47.90667 23.80346 2.590031           13.20858

$parameters
     test                        name.t ntr alpha
  Scheffe factor(poker_players$skills)   3  0.05

$means
           poker_players$score       std  r  Min  Max   Q25  Q50    Q75
advanced                 46.79  9.030621 10 32.5 57.7 39.45 48.4 54.925
novice                   33.04  8.033292 10 22.1 43.2 27.05 32.6 39.500
proficient               63.89 15.621456 10 40.1 85.3 52.50 64.6 75.650
```

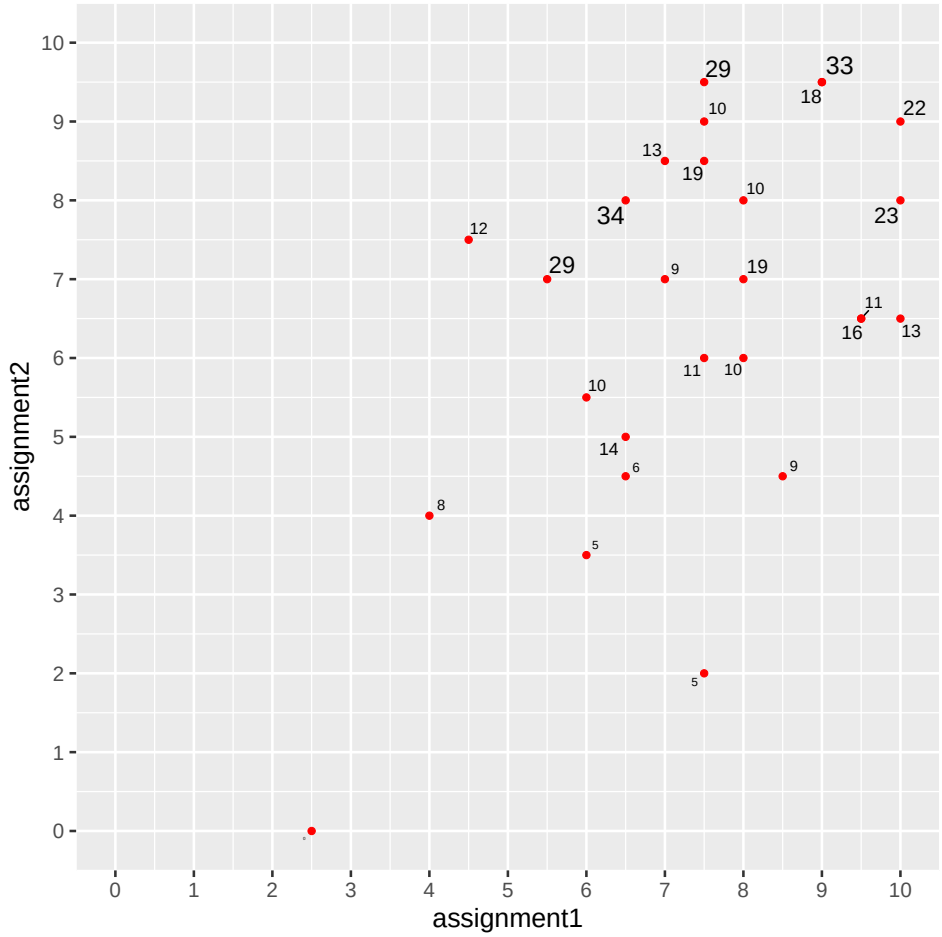
```
$groups
           poker_players$score groups
proficient                63.89      a
advanced                  46.79      b
novice                    33.04      c
```

In conclusion, we can safely state that all three groups are different at significance level 5%

Exercise 5.

1.

We visualize the data in a scatterplot, where the x axis corresponds to the assigment 1 grade, and the y axis to the assigment 2 grade. Also, above every point there is the number of report pages in assignment 2 scaled accordingly.



We can see that there is a correlation between assignment 1 grade and assignment 2 and the size of the report. For that reason we will try to fit linear regression models in order to infer the grades of 2nd assignment. We can also calculate the Pearson's product-moment correlation between our variables as follows:

```
cor.test(grades$assignment1, grades$assignment2)
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

Pearson's product-moment correlation

data:  grades$size and grades$assignment2
t = 4.9786, df = 25, p-value = 3.94e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4448524 0.8560624
sample estimates:
      cor
0.7055868
```

2.

We will now fit, a normal regression model in order to infer the grades of 2nd assignment using only the 1st assignment grade as explanatory variable

```
> fit<-lm(assignment2 ~ assignment1, data=grades)
> summary(fit)

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.6254 -1.4661 -0.1031  1.8523  2.9192
```

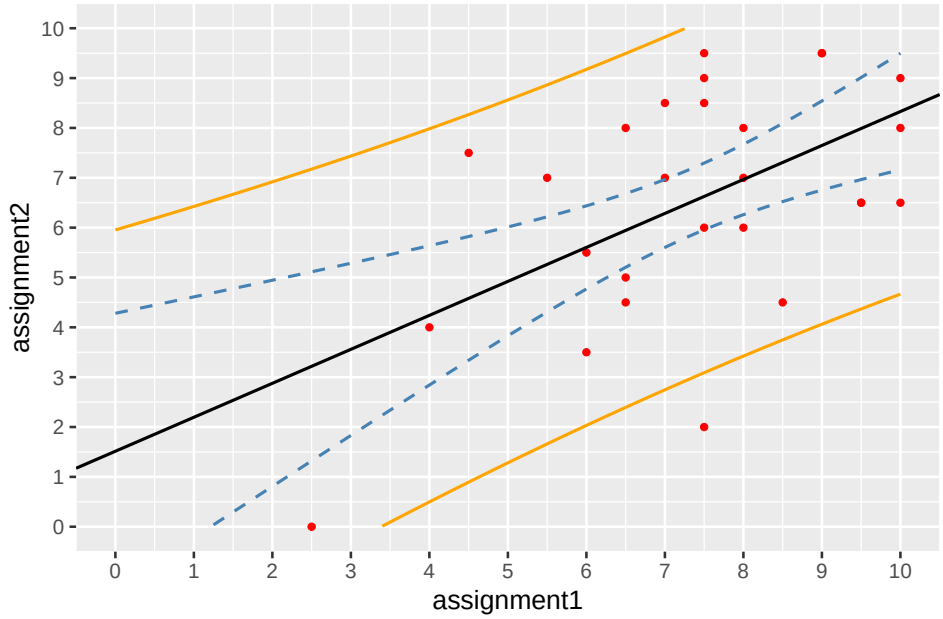
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5140     1.6202   0.934  0.35901
assignment1   0.6815     0.2133   3.195  0.00376 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.033 on 25 degrees of freedom
Multiple R-squared:  0.2899,Adjusted R-squared:  0.2615
F-statistic: 10.21 on 1 and 25 DF,  p-value: 0.003763
```

Using the coefficients that we found, our linear regression model is

$$\hat{\text{assignment2}}_i = 1.5140 + 0.6815 \times assignment1_i + \epsilon_i$$

Generally, p-value expresses the probability of observing the test statistic assuming the null hypothesis that the coefficient of assignment1 is zero ,which means that it has no effect in our response variable.($H_0 \beta_{assignment1} = 0$). As we can see in the summary of our model, the coefficient of our explanatory variable has a p-value < 0.05, so it appears that we have a statistical significance for the assignment1 variable. Indeed the 2nd assignment grade is affected by the 1st assignment grade.
We also got an R-squared (coefficient of determination) $= 0.2899$, so we explained $\approx 28\%$ of the total variation of the response variable assignment2 using this model.

We can now plot the estimated regression line on top of the scatter-plot of the observed data. We superimpose the 90% confidence intervals (blue dashed lines) and the prediction intervals (orange lines).



3.

We will now add a second explanatory variable in our previous mode, the number of report pages in the second assignment. We fit the new model as follows:

```
> fit2<-lm(assignment2 ~ assignment1 + size, data=grades)
```

Using the coefficients that we found, our linear regression model is

$$\hat{\text{assignment2}}_i = 1.0559 + 0.42575 \times assignment1_i + 0.15896 \times size_i + \epsilon_i$$

```
> summary(fit2)

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
```

Both coefficients are statistically significant, but it appears that size is the most significant of the two in our current model. Both the R^2 and Adjusted-R^2 where increased substantialy, so we conclude that our new model better explains tha variability of the response variable. To test that, since we have nested models, we can use the ANOVA, as follows:

```
> anova(fit,fit2)
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
```

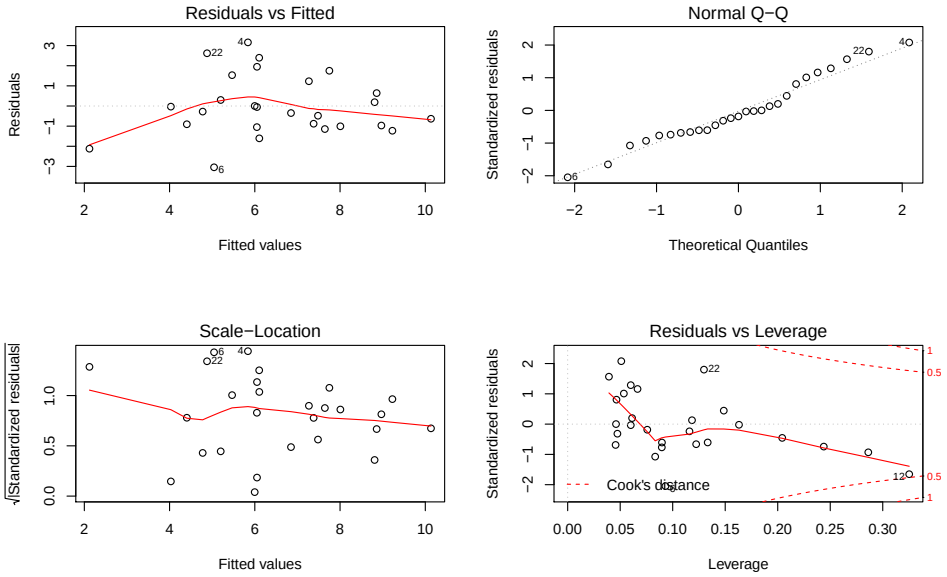
So the F-test to the hypothesis for the new variable:

$$H_0(\textit{Null hypothesis}) : \beta_{size} = 0$$
$$H_1(\textit{alternative}) : \beta_{size} \neq 0\}$$

gives us a p-value that is very small, thus we reject the Null hypothesis, and we conclude that the model fits better than the previous one.

4.

To test the modelling assumptions, we will first examine the diagnostic plots of our fitted model:



In the QQ-plot we can see that almost all the points appear to fall in the reference line. So the normality stands. We can confirm that assumption using the known normality tests:

```
> library(nortest)
> ad.test(fit2$residuals)

Anderson-Darling normality test

data:  fit2$residuals
A = 0.498, p-value = 0.1935

> shapiro.test(fit2$residuals)

Shapiro-Wilk normality test

data:  fit2$residuals
```

```
W = 0.96323, p-value = 0.4364
```

In the Residuals vs Fitted plot, the red line is almost horizontal at zero. Also there appears to be no pattern in the residual plot. Thus we can assume a linear relationship between the predictors and the outcome variables.

In the Residuals vs Leverage plot, we can see that there are high leverage points (#12, #22) and that is not good, because it can alter the results of our regression analysis.

In the Scale-Location plot, the residuals are spreaded equally along the ranges of the predictors. Thus, the homoscedasticity assumption stands.

```
> t.test(fit2$residuals)

One Sample t-test

data:  fit2$residuals
t = 4.528e-17, df = 26, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.5934458  0.5934458
sample estimates:
   mean of x
1.307267e-17
```

Thus, the mean value of the residuals is zero. All the modelling assumptions have been validated.

5.

We estimate the mean grade for a student with 1st assignment grade equal to 6 and 2nd assignment consisting of 10 pages. We also give the 90% confidence interval for our prediction:

```
 > predict(fit2, newdata = list(assignment1 = 6, size = 10), interval = 'confidence',
       fit      lwr      upr
1 5.200013 4.538343 5.861684
```

6.

We predict the grade for a student with 1st assignment grade equal to 6 and 2nd assignment consisting of 10 pages. We also give the 90% prediction interval :

```
> predict(fit2, newdata = list(assignment1 = 6, size = 10), interval = 'prediction',
       fit      lwr      upr
1 5.200013 2.447881 7.952146
```

7.

After a lot of trials for different variables and transformations of our explanatory variables we found the following model that best fits our data:

$$\hat{\text{assignment2}}_i = 4 + 0.08 \times \log(assignment1_i) \times size_i + 0.71 \times size_i - 5.21 \times size_i^2 + 2.22 \times size_i^3 + \epsilon_i$$

```
> summary(fit3)

Call:
lm(formula = assignment2 ~ I(log(assignment1) * size) + poly(size,
    3), data = grades)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1353 -0.8007 -0.2396  0.8577  2.8152

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.00225    2.18970   1.828 0.081181 .
I(log(assignment1) * size)  0.08500    0.07296   1.165 0.256491
poly(size, 3)1              0.71080    6.81606   0.104 0.917889
poly(size, 3)2             -5.21615    1.32096  -3.949 0.000683 ***
poly(size, 3)3              2.22224    1.31981   1.684 0.106366
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.282 on 22 degrees of freedom
Multiple R-squared:  0.7516,Adjusted R-squared:  0.7064
F-statistic: 16.64 on 4 and 22 DF,  p-value: 2.062e-06
```

Exercise 6.

1.

First we load the data into R using:

```
> require(stats)
> data(mtcars)
>
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

2.

We define **am, vs, cyl** as factor variables, and leave the rest as numeric:

```
mtcars$am   <- as.factor(mtcars$am)
mtcars$vs   <- as.factor(mtcars$vs)
mtcars$cyl  <- as.factor(mtcars$cyl)
```

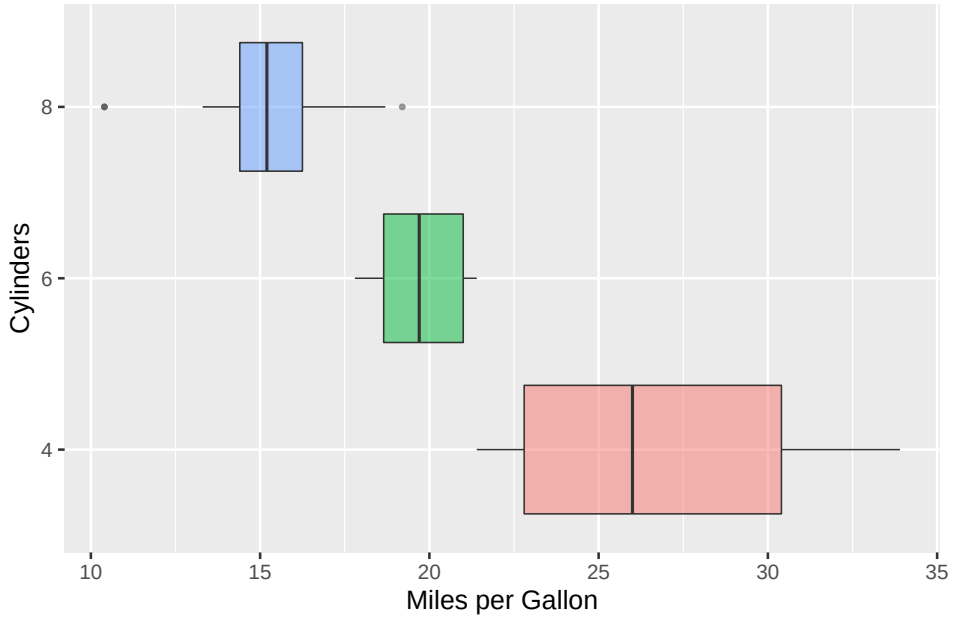
3.

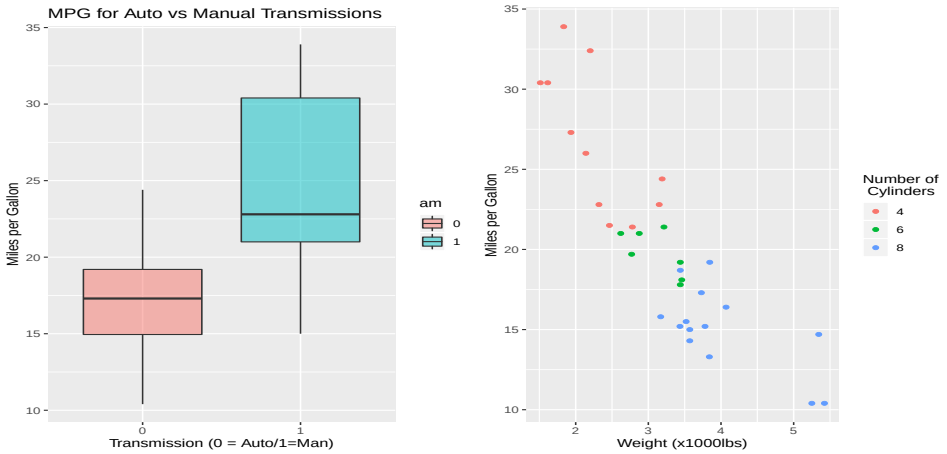
We report some basic descriptive statistics for each variable in our dataset as follows:

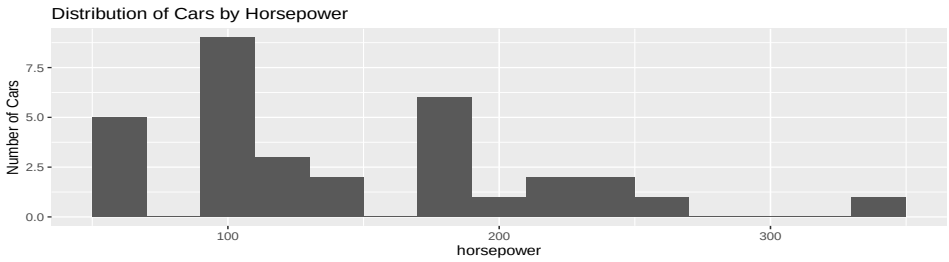
```
> describe(mtcars)[c('n','mean','median','sd','min','max','range')]
      n   mean median     sd   min    max  range
mpg  32  20.09  19.20   6.03 10.40  33.90  23.50
cyl* 32   2.09   2.00   0.89  1.00   3.00   2.00
disp 32 230.72 196.30 123.94 71.10 472.00 400.90
hp   32 146.69 123.00  68.56 52.00 335.00 283.00
drat 32   3.60   3.70   0.53  2.76   4.93   2.17
wt   32   3.22   3.33   0.98  1.51   5.42   3.91
qsec 32  17.85  17.71   1.79 14.50  22.90   8.40
```

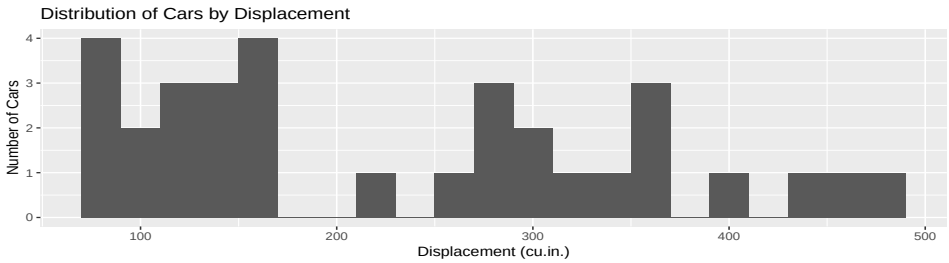
```
vs*  32   1.44   1.00   0.50  1.00   2.00   1.00
am*  32   1.41   1.00   0.50  1.00   2.00   1.00
gear 32   3.69   4.00   0.74  3.00   5.00   2.00
carb 32   2.81   2.00   1.62  1.00   8.00   7.00
```

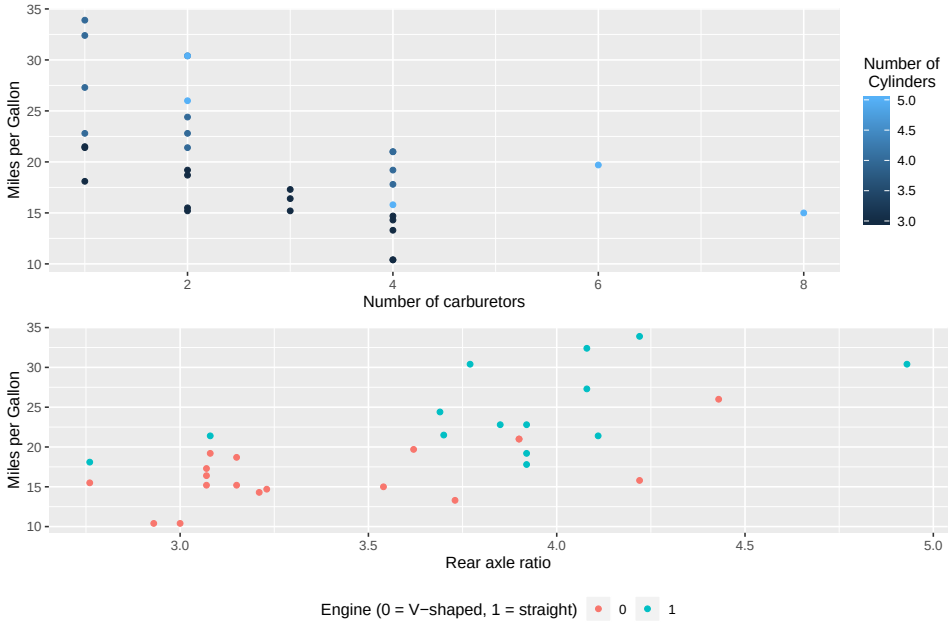
Now we will provide some illustrations to better understand our data and how:




MPG for Auto vs Manual Transmissions
Miles per Gallon
Transmission (0 = Auto/1=Man)
am
0
1
Weight (x1000lbs)
Number of Cylinders
4
6
8

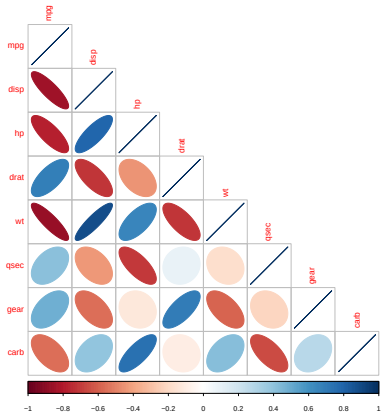

Distribution of Cars by Horsepower
Number of Cars
horsepower

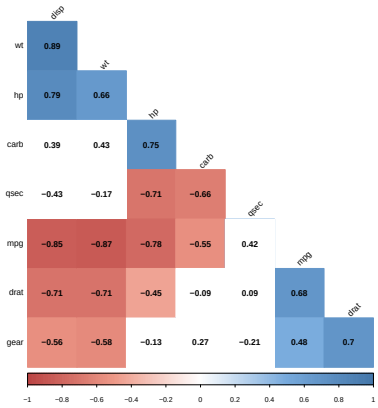

Distribution of Cars by Displacement
Number of Cars
Displacement (cu.in.)


Miles per Gallon
35
30
25
20
15
10
2
4
6
8
Number of carburetors
Number of Cylinders
5.0
4.5
4.0
3.5
3.0
Miles per Gallon
35
30
25
20
15
10
3.0
3.5
4.0
4.5
5.0
Rear axle ratio
Engine (0 = V−shaped, 1 = straight)
0
1

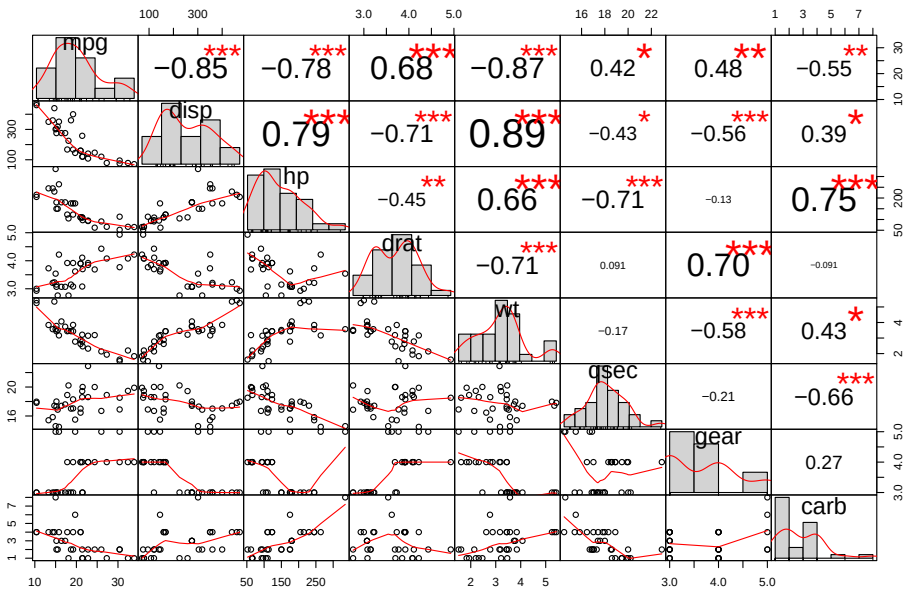
4.

In this section we will produce all pairwise scatterplots for the numeric variables, and we will also compute the corresponding correlation coefficients. In the two following figures we can visualize the correlation between the variables and also see the coefficients.





We will now produce the pairwise scatterplots, with the histograms of the variables at the diagonal, and the calculated coefficients at the upper triangle of the graph



5.

Earlier, we visualized the difference in means between cars with automatic transmission and cars with manual transmission using two boxplots. It appeared that there is a difference, but we will properly test that assumption as follows:

$$H_0(\textit{Null hypothesis}) : \mu_{auto} = \mu_{manual}$$
$$H_1(\textit{alternative}) : \mu_{auto} \neq \mu_{manual}\}$$

Before we perform the test statistic, we will first check the assumptions of normality and homogeneity of variances between the two groups.

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

Anderson-Darling normality test
```

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

>   ad.test(mtcars[mtcars$am == 1,'mpg'])

Anderson-Darling normality test

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

So the normality assumptions stands.

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

Bartlett test of homogeneity of variances

data:  mpg by am
Bartlett's K-squared = 3.2259, df = 1, p-value = 0.07248

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

Fligner-Killeen test of homogeneity of variances

data:  mpg by am
Fligner-Killeen:med chi-squared = 4.4929, df = 1, p-value = 0.03404

>   library(car)
>   leveneTest(mpg~am,data = mtcars)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  1  4.1876 0.04957 *
      30
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The majority of the tests rejects the null hypothesis that the variance is equal between the groups, so we assume that we have unequal variances. We will now perfom the t-test:

```
>   t.test(mpg~am,data = mtcars, var.equal = FALSE)

Welch Two Sample t-test

data:  mpg by am
t = -3.7671, df = 18.332, p-value = 0.001374
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.280194  -3.209684
sample estimates:
mean in group 0 mean in group 1
       17.14737        24.39231
```

We get a small p-value, thus we reject the null hypothesis. We conclude with 95% confidence that the true mean between the two groups differs.

6.

Now, we have $k = 3$ different groups of data, so we will perfom the ANOVA model to test if there is a difference in consumption among cards with different number of cylinders.

$$H_0(\textit{Null hypothesis}) : \mu_{cyl=4} = \mu_{cyl=6} = \mu_{cyl=8}$$
$$H_1(\textit{alternative}) : \text{Not } H_0\}$$

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

We get a really small value for the F-test, thus we have statistical evidence against the null hypothesis We will now test the assumptions of ANOVA:

```
>   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

>   library(car)
>   leveneTest(mpg~cyl,data=mtcars)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  2  5.5071 0.00939 **
      29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, the variance between the groups is not equal, thus the assumption of homogeneity is violated.

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

Shapiro-Wilk normality test

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

We got a large p-value, thus we fail to reject the null hypothesis and we assume the the residuals of our model are normally distributed. Furthermore, because one of the ANOVA assumptions is violated, we will perform a non-parametric test to test our hypothesis as follows:

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

```
Kruskal-Wallis rank sum test

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

In conclusion we reject the null hypothesis, because we have statistical evidence that the means between the three groups of data are not common.

7.

We fit the full regression model and print the summary in R as follows:

```
> fitall <- lm(mpg ~  cyl + disp + hp + drat
                   + wt + qsec + vs + am + gear + carb ,data = mtcars)
> summary(fitall)

Call:
lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs +
    am + gear + carb, 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
```

Our model, explains the 88% of the variability of the response variable.

Now we will perform a stepwise selection of the best model, using the Bayesian Information Criterion. We will try the forward method, where we will add an explanatory variable according to the lowest BIC, a backwards elimination method where we will remove the variable with the highest BIC, and lastly a stepwise selection method, where in each step we will add or remove a variable.

Using the following code in R:

```
> fitnull<-lm(mpg ~ 1, data = mtcars)
> stepFS<-step(fitnull, scope=list(lower = ~ 1,
+                                  upper= ~ cyl + disp + hp + drat + wt + qsec + vs +
+              direction="forward",k = log(n), criterion = "BIC", data=mtcars)
```

we found that the best model using the forward method is:

```
Step:  AIC=68.24
mpg ~ wt + hp
```

and using

```
> stepBE<-step(fitall, scope=list(lower = ~ 1,
+                                 upper= ~ cyl + disp + hp + drat + wt + qsec + vs +
+              direction="backward",k = log(n), criterion = "BIC", data=mtcars)
```

we found that the best model using the backwards elimination method is:

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

and using

```
> stepSR<-step(fitall, scope=list(lower = ~ 1,
+                                  upper= ~ cyl + disp + hp
+ drat + wt + qsec + vs + am + gear + carb ),
+              direction="both",k = log(n), criterion = "BIC", data=mtcars)
```

we found that the best model using the backwards elimination method is:

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

Now we examine which variables mostly effect the consumption in our previous model using:

Thus, according to the best model that we found, the variables that mostly affect consumption are :

1. Weight (1000 lbs)
2. 1/4 mile time
3. Transmission (0 = automatic, 1 = manual)

We interpret the model as follows:

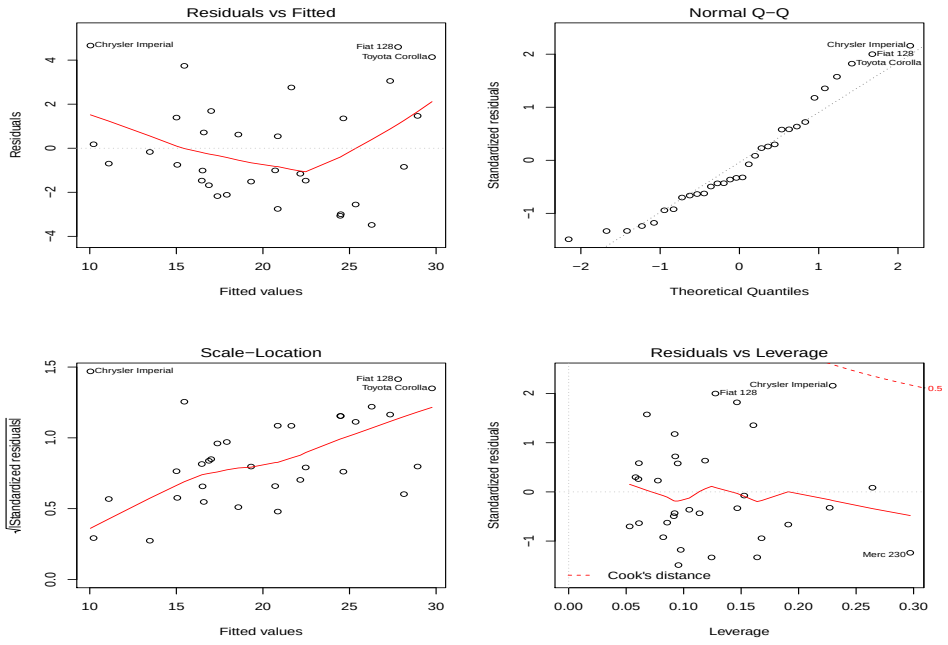
```
> fitbest <- lm(mpg ~  wt + qsec + am ,data = mtcars)
> summary(fitbest)

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
```

Iterestingly. we got a bigger adjusted R-squared metric from that of the full model, although we used far less variables to predict. That is mainly, due to the fact the this particular metric punishes complex models over simpler ones.

All of our coefficients appear to be statistically significant, with the first two more than the third. Using the estimated coefficients we have:

$$\hat{\text{mpg}}_i = 9.61 - 3.91 \times wt_i + 1.22 \times qsec_i + 2.93 \times am_i + \epsilon_i$$



It appears that are the modelling assumptions stand as well.