Statistical Modelling and Design of Experiments
1st Assignment
Beata Baczyńska

November 19, 2021

1 FIRST QUESTION: VISUALISATION, CHI SQUARE AND T-TEST

The dataset shows results for decathlon competition. This competition consists of 10 different disciplines.

- 100m (unit: seconds)

- Long jump (unit: metres)

- Shot put (unit: metres)

- High jump (unit: metres)

- 400 m (unit: seconds)

- 100 m hurdles (unit: seconds)

- Discus (unit: metres)

- Pole vault (unit: metres)

- Javelin (unit: metres)

- 1500 m (unit: seconds)

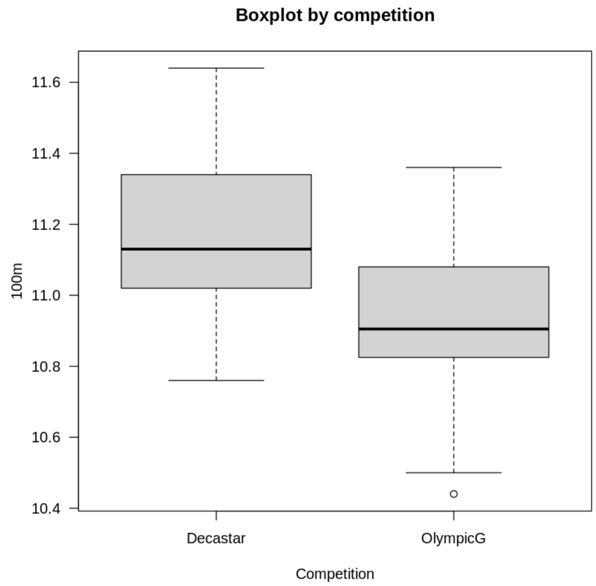
All columns with results for these disciplines consist of quantitative variables.

Also ‘Points’ are quantitative variables.

‘Rank’ and ‘Competition’ are both qualitative variables.

For ‘Competition’ we are having two values: ‘Decastar’ and ‘OlympicG’. The Decastar is an annual event and the Olympic is every 4 years, so we expect players in the OlympicG competition to have slightly better results (we are only guessing now).

1.1 Analyze the distribution of "X100m" according to the type of competition by using boxplot. Write your conclusion.



The boxplot method shows us the groups of numerical data through their quarterlies. The bold line indicates the median, so the value above which we find 50% of data and below which we find another 50% of data. The upper bound of the box indicates the 3rd quartile and below the bound of the box the 1st quartile. Outliers are marked as dots.The median during the OlympicG competition is lower than for Decastar competition. The distance between 75% and 25% is lower for OlympicG competition.Whole range for Decastar competition is 11.65 - 10.75 = 0.9 seconds and for OlympicG: 11.4 - 10.5 = 0.9 seconds, however for OlympicG the range between second and third quarter is tighter. For OlypicG we can also notice one outlier with value ~ 10.4.

1.2 Create a new categorical variable with two categories from the variable "X100m" by using 11 seconds as the cut-off point. Make a cross table from the new categorical variable and the "Competition". Are these two variables independent? Write your conclusion by checking marginal probabilities and test the independency of two variables by using Chi-Square test.

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
```

```
|-------------------------|


Total Observations in Table:  41


               | mydata$Competition
mydata$X100m11 |  Decastar |  OlympicG | Row Total |
---------------|-----------|-----------|-----------|
        faster |         2 |        19 |        21 |
               |     3.259 |     1.513 |           |
               |     0.095 |     0.905 |     0.512 |
               |     0.154 |     0.679 |           |
               |     0.049 |     0.463 |           |
---------------|-----------|-----------|-----------|
        slower |        11 |         9 |        20 |
               |     3.422 |     1.589 |           |
               |     0.550 |     0.450 |     0.488 |
               |     0.846 |     0.321 |           |
               |     0.268 |     0.220 |           |
---------------|-----------|-----------|-----------|
  Column Total |        13 |        28 |        41 |
               |     0.317 |     0.683 |           |
---------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  9.783628     d.f. =  1     p =  0.001760726

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  7.796182     d.f. =  1     p =  0.005235676
```

The marginal probabilities are giving us an idea about the proportion between columns (number of people in Decastar/OlympicG competition) and between rows (number of people with result better/worse than 11 seconds on 100 meters). We can see that there are more people in OlympicG competition (68.3%) than in Decastar (31.7%). The proportion between results better than 11 seconds and worse is more or less equal (51.2% and 48.8%). However, when we check cross table values we can see that in the Decastar competition there are only 2 people marked as 'faster' and 11 marked as 'slower'. The Conditional probability P('faster'|Decastar) = 0.154 and P('slower'|Decastar) = 0.846. For the OlympicG competition there are more people marked as 'faster' (19 people) than those marked as 'slower' (9 people). The Conditional probability here is:

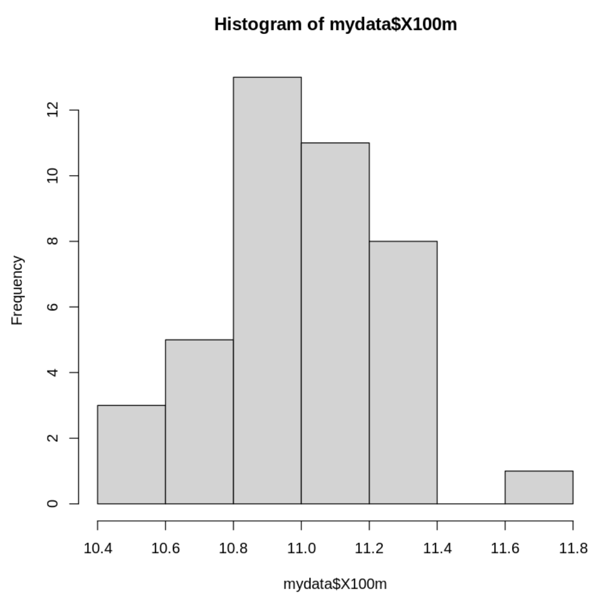
P('faster' | OlympicG) = 0.679 and P('slower' | OlympicG) = 0.321. Also when we read conditional probabilities P(Olympic | 'faster') = 0.905 and P(Decastar | 'faster') = 0.095 We can notice that there are faster people in OlympicG competition.

In my opinion this disproportion suggests that these two variables: Competition and categorical result on 100 meters are dependent.

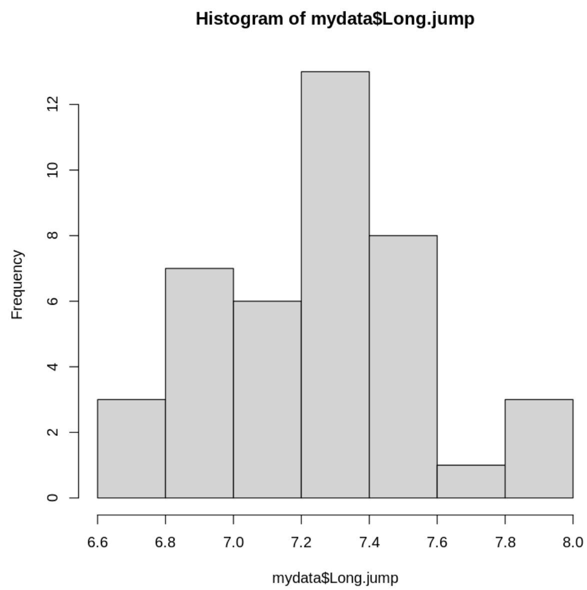
The p-values for both Chi-squared tests (with and without Yates' correction) are very small. We should reject our Ho hypothesis. That means the variables "Competition" and "X100m11" are dependent.

1.3 Visualize the distribution of quantitative variables by using proper graph. Which of these variables follow a Normal distribution?

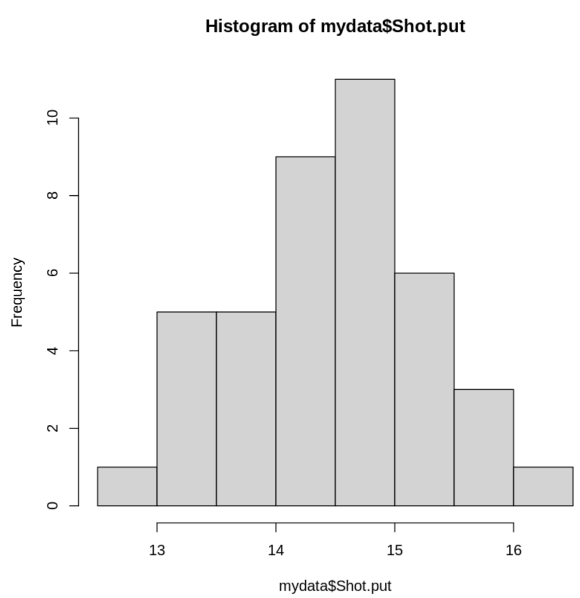
```
        Shapiro-Wilk normality test

data:  mydata$X100m
W = 0.9818, p-value = 0.7435
```



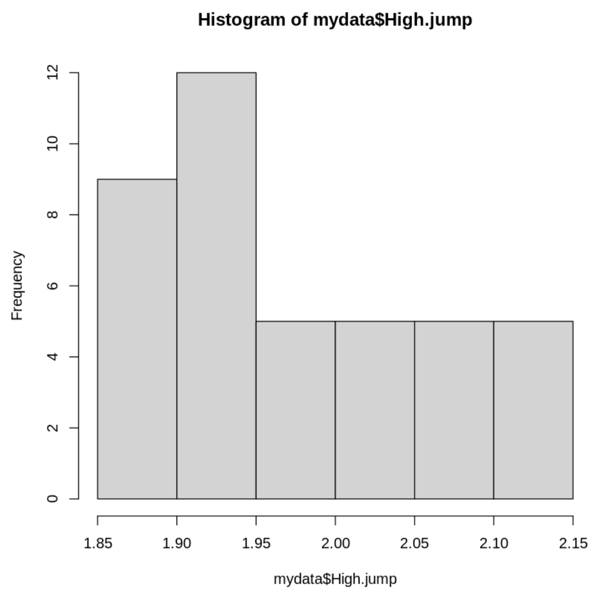

```
        Shapiro-Wilk normality test

data:  mydata$Long.jump
W = 0.98763, p-value = 0.9289
```



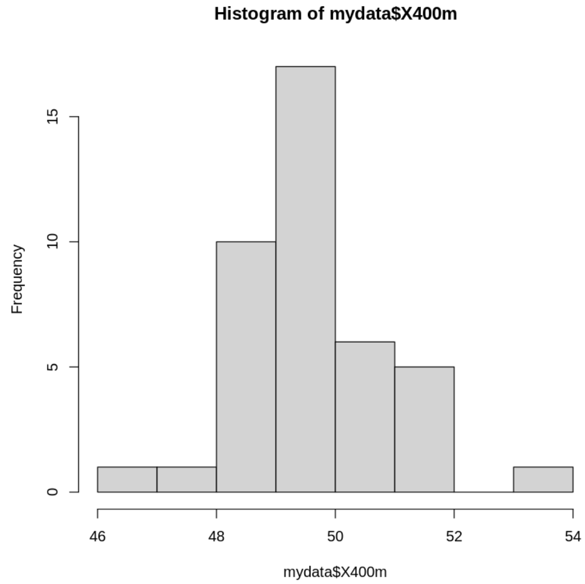

```
        Shapiro-Wilk normality test

data:  mydata$Shot.put
W = 0.9884, p-value = 0.9456
```



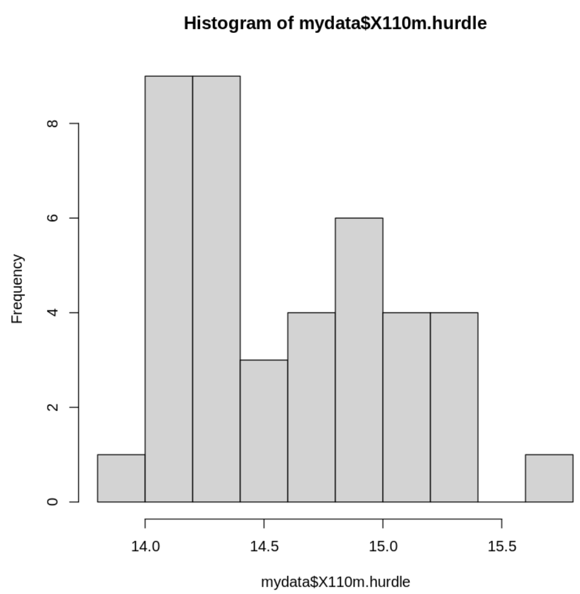
```
        Shapiro-Wilk normality test

data:  mydata$High.jump
W = 0.93734, p-value = 0.0255
```



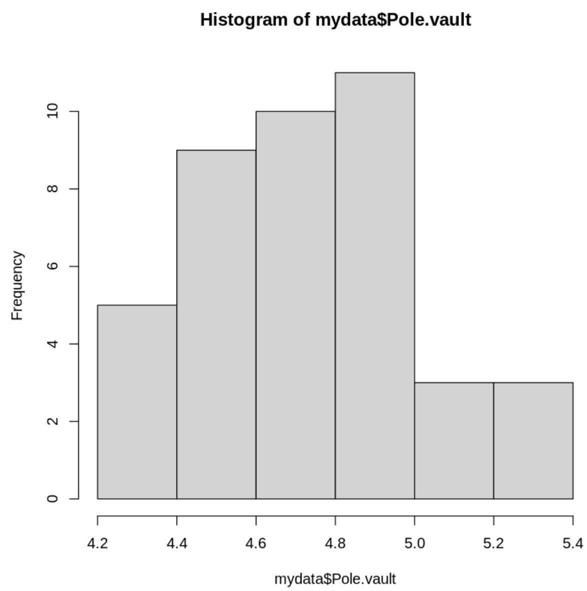

```
        Shapiro-Wilk normality test

data:  mydata$X400m
W = 0.95714, p-value = 0.1248
```



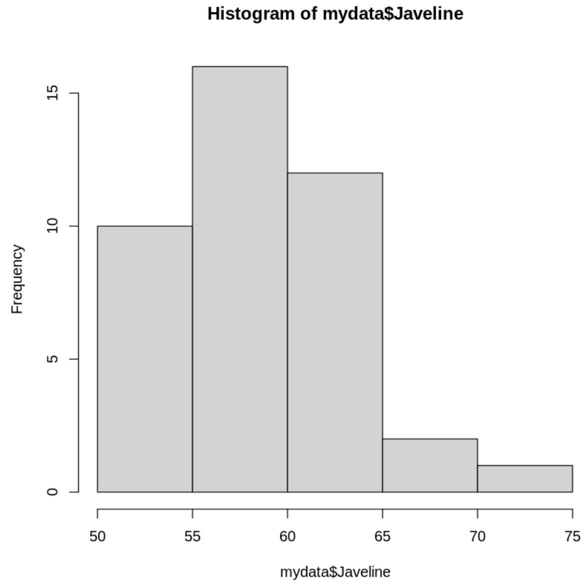

```
        Shapiro-Wilk normality test

data:  mydata$X110m.hurdle
W = 0.93087, p-value = 0.01544
```



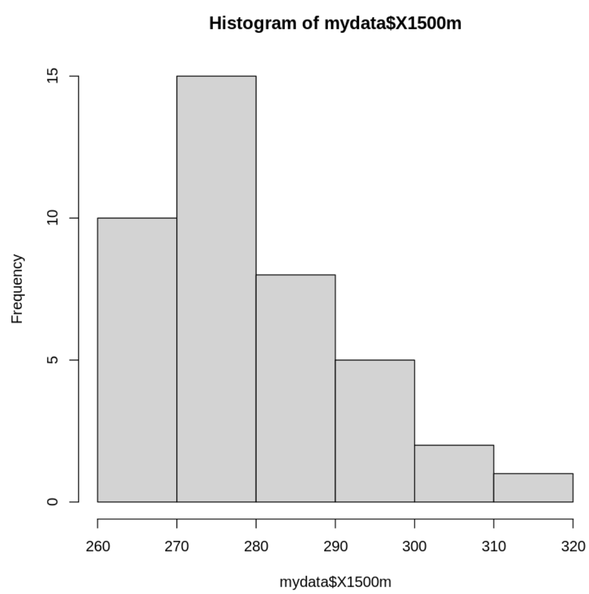
```
        Shapiro-Wilk normality test

data:  mydata$Pole.vault
W = 0.97003, p-value = 0.3456
```



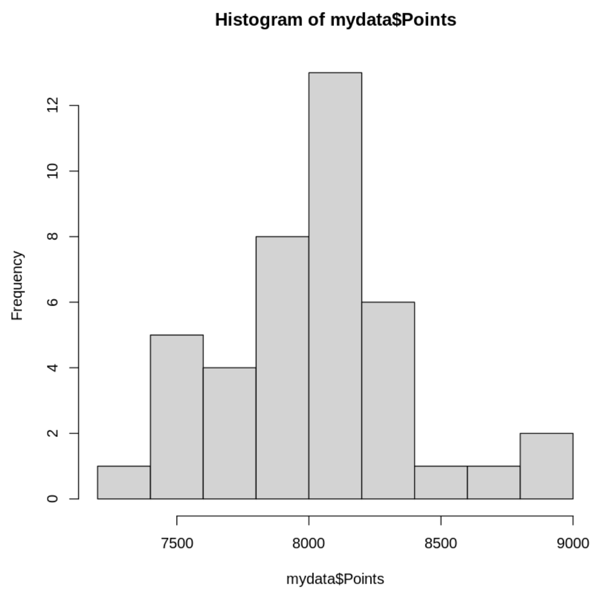

```
        Shapiro-Wilk normality test

data:  mydata$Javeline
W = 0.97106, p-value = 0.3732
```




```
        Shapiro-Wilk normality test

data:  mydata$X1500m
W = 0.93652, p-value = 0.02391
```



```
        Shapiro-Wilk normality test

data:  mydata$Points
W = 0.95584, p-value = 0.1123
```



By looking on the graph and checking if p-value in shapiro test is higher than 0.05, we can say that:

Variables:

- X100m

- Long.jump

- Shot.put

- X400m

- Pole.vault

- Javeline

- Points

follow Normal distribution.

Variables:

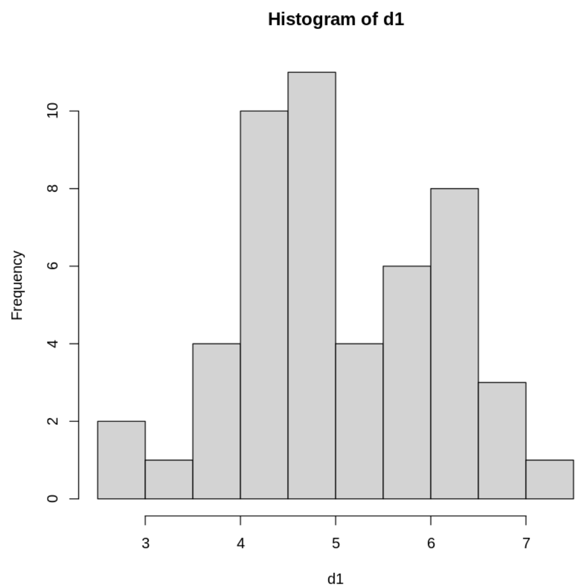
- High.jump

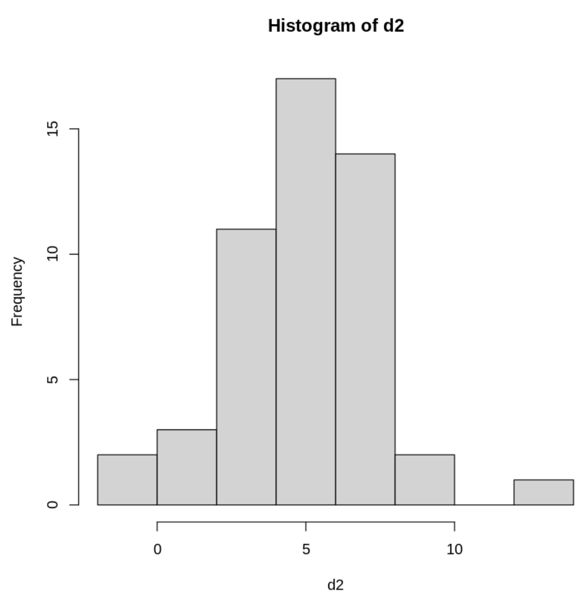
- X110m.hurdle

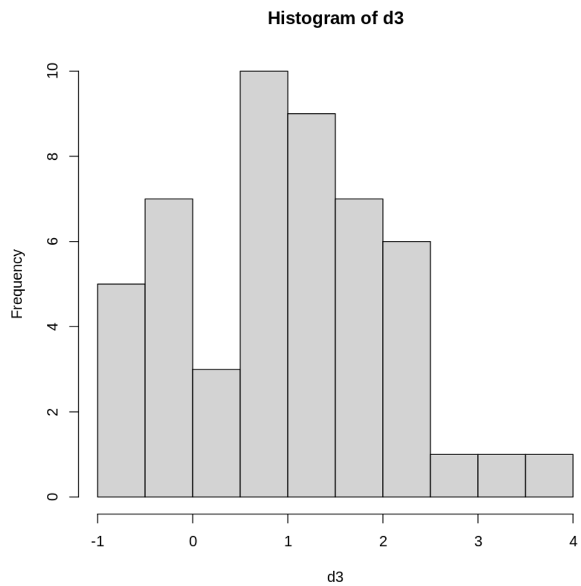
- X1500m

don’t follow Normal distribution.

1.4 Generate three Normally distributed random variables of length 50. Two of them should have the same mean, different standard deviations while the third one has a different mean but the same standard deviation with the first distribution. Use t test to compare mean differences between three variables.







H0 - There is no statistically significant difference between the samples

```
        Welch Two Sample t-test

data:  d1 and d2
t = 0.18568, df = 67.145, p-value = 0.8533
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6817320  0.8215868
sample estimates:
mean of x mean of y
 5.064934  4.995007
```

We do not reject H0 hypothesis as p-value is greater than 0.05

```
        Two Sample t-test

data:  d1 and d3
t = 18.83, df = 98, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.666683 4.530578
sample estimates:
mean of x mean of y
```

```
5.0649344 0.9663037
```

We reject the H0 hypothesis as the p-value is smaller than 0.05. There is significant difference

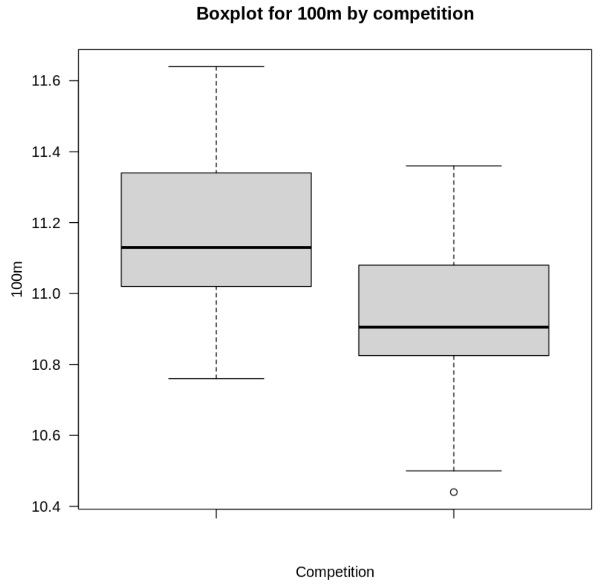
```
        Welch Two Sample t-test

data:  d2 and d3
t = 10.634, df = 68.385, p-value = 3.884e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.272819 4.784588
sample estimates:
mean of x mean of y
4.9950070 0.9663037
```

We reject the H0 hypothesis as the p-value is smaller than 0.05. There is a significant difference.

We can see that for tests where distributions have different means we always had to reject the H0 hypothesis and accept that there is significant difference between distributions.

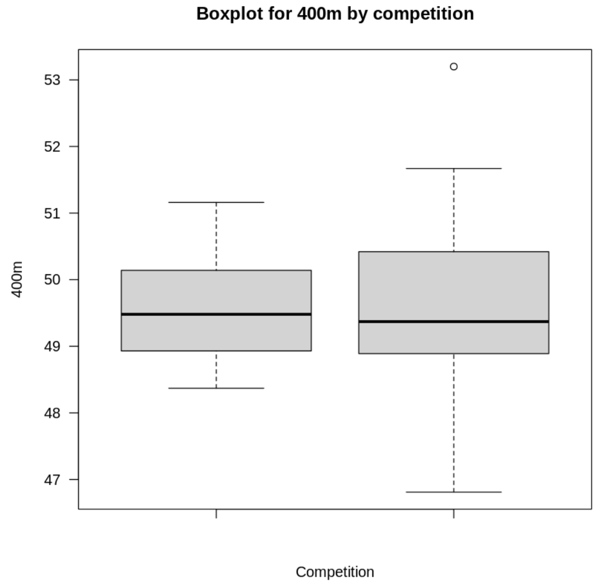
1.5 Test if there is a difference between two type of competitions according to the variables "X100m" and "X400m" by using t test.

```
        Welch Two Sample t-test

data:  mydata$X100m[mydata$Competition == "Decastar"] and
 ↪mydata$X100m[mydata$Competition == "OlympicG"]
t = 3.2037, df = 22.168, p-value = 0.00407
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.09164794 0.42769272
sample estimates:
mean of x mean of y
 11.17538  10.91571
```



We should reject the H0 hypothesis as the p-value is smaller than 0.05. There is a significant difference.

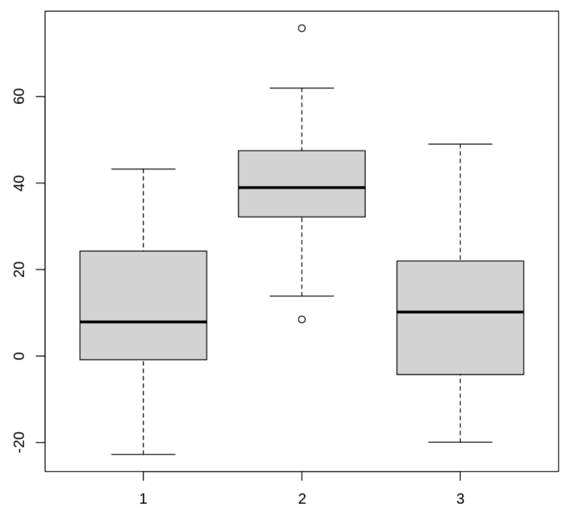
```
        Welch Two Sample t-test

data:  mydata$X400m[mydata$Competition == "Decastar"] and
 ↪mydata$X400m[mydata$Competition == "OlympicG"]
t = 0.05771, df = 32.106, p-value = 0.9543
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6858299  0.7258299
sample estimates:
mean of x mean of y
    49.63     49.61
```

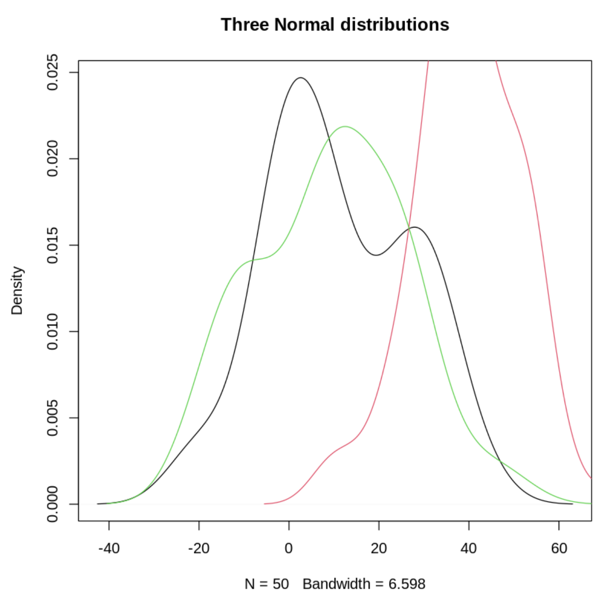


We accept the H0 hypothesis as p-value is higher than 0.05. There is no significant difference.

2 SECOND QUESTION : ANOVA

2.1 Generate three populations that follow a normal distribution, using your own algorithm. As an example, the first is a population that follows a normal distribution with a parameter mean=10, the second with mean=40, and the third with mean=10. Select the SAME variance for the three distributions at your convenience (a value >0).

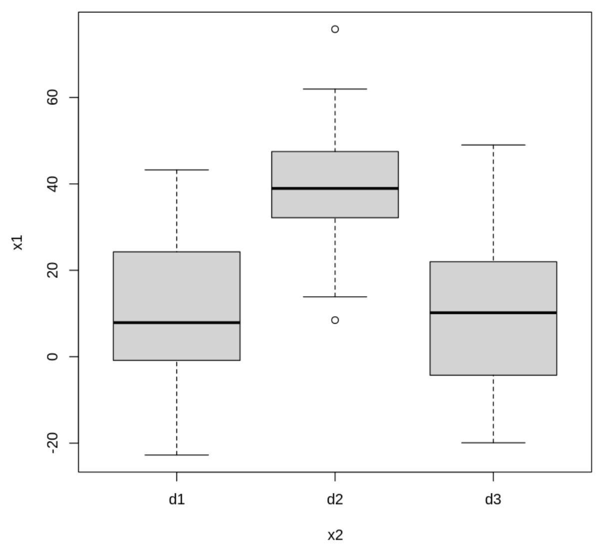




2.2 We want to analyze using an ANOVA if these three populations are different (or not) depending on the parameter selected. Analyze and explain the results obtained. Justify your answers. Remember to test the ANOVA assumptions. What do you expect on the assumptions?

```
             Df Sum Sq Mean Sq F value Pr(>F)
x2            2  29538   14769   64.99 <2e-16 ***
Residuals   147  33405     227
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

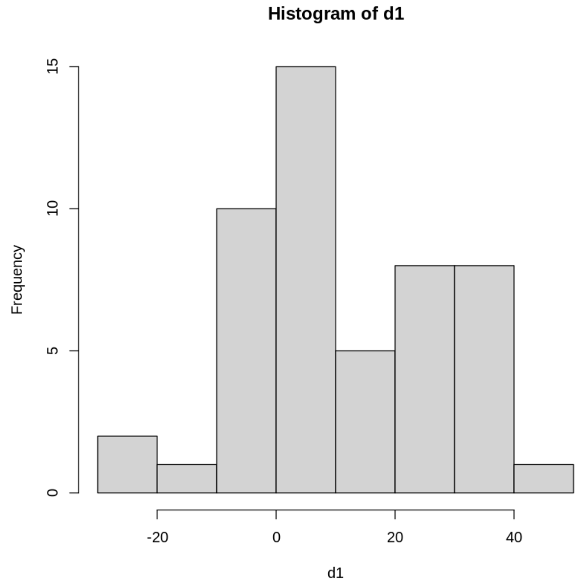
```
Warning message in Boxplot.default(mf[[response]], x, id = list(method =
id.method, :
“NAs introduced by coercion”
```

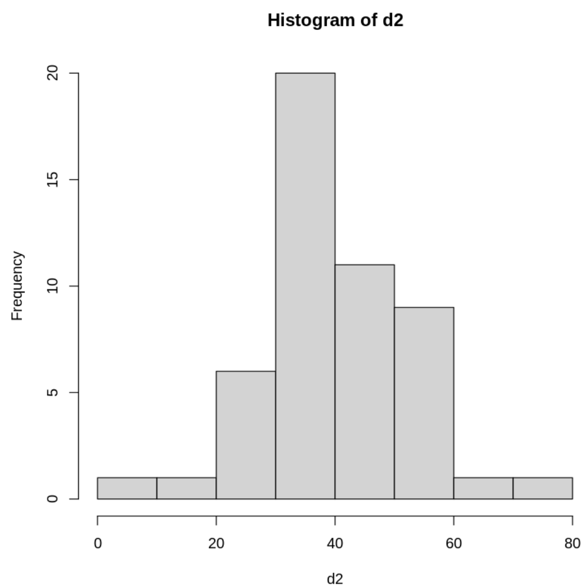


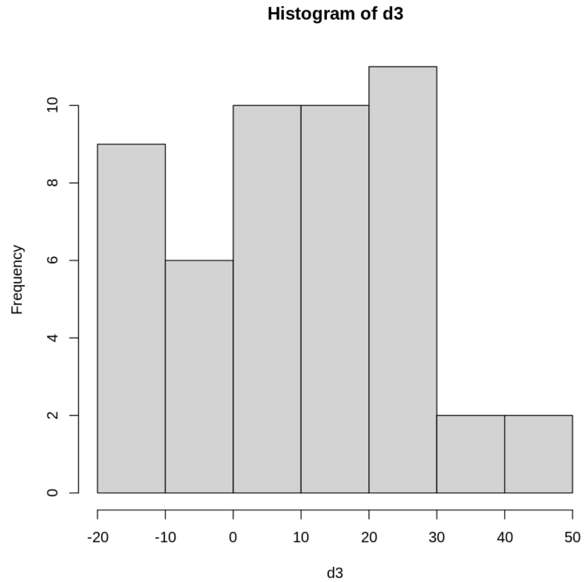
As the p-value is less than the significance level 0.05, we can conclude that at least one distribution is different

2.2.1 ANOVA assumptions:

- For each population, the response variable is normally distributed (checked by Histogram, Q-Q plot or Shapiro-Wilk hypothesis test).


Histogram of d1
Frequency
0
5
10
15
-20
0
20
40
d1


Histogram of d2
Frequency
0
5
10
15
20
0
20
40
60
80
d2




```
        Shapiro-Wilk normality test

data:  residuals(AnovaModel.1)
W = 0.99309, p-value = 0.69
```

The p-value is greater than 0.05, the assumption is valid.

- The variance of the response variable is the same for all of the populations (checked by Levene's Test or Breusch Pagan Test).

```
        studentized Breusch-Pagan test

data:  AnovaModel.1
BP = 5.4666, df = 2, p-value = 0.065
```

The p-value is greater than 0.05, the assumption is valid.

- The observations must be independent (checked by Durbin Watson test).

```
        Durbin-Watson test

data:  AnovaModel.1
DW = 1.9372, p-value = 0.5811
alternative hypothesis: true autocorrelation is not 0
```

The p-value is greater than 0.05, the assumption is valid.

2.3 We want to analyze if both Age and diabetes affects the risk factors. First categorize age in three groups: <=30 (young), 31-50 (middle age) and 50+ (old).

2.3.1 Questions:

- How does age influence on the risk factors associated with diabetes?

We were testing the hypothesis that there is no difference in means between groups. For the tests where this hypothesis was rejected (p-value smaller than 0.05) we can say that tested variable influences Age.

Variables related with age are:

- Pregnancies

- Glucose

- BloodPressure

- SkinThickness

- BMI

Insulin and DiabetesPedigreeFunction variables have no influence on age.

- Which of the risk factors are related with diabetes?

We were testing the hypothesis that there is no difference in means between groups. For the tests where this hypothesis was rejected we can say that the tested variable is related with diabetes.

Variables related with diabetes are:

- Pregnancies

- Glucose

- SkinThickness

- Insulin

- BMI

- DiabetesPedigreeFunction

Only the BloodPressure variable is not related.

- Detail the results of Two-Way ANOVA considering "Blood Pressure" as dependent variable, and the age groups and the indicator of diabetes as independent variables. Analyze the interaction term of two factors.

2.3.2 ANOVA assumptions

The observations within each sample must be independent (Durbin Watson).

```
Durbin-Watson test
```

```
data:  model20
DW = 1.9684, p-value = 0.6611
alternative hypothesis: true autocorrelation is not 0
```

The populations from which the samples are selected must be normal (Shapiro test).

```
        Shapiro-Wilk normality test

data:  residuals(model20)
W = 0.81264, p-value < 2.2e-16
```

The populations from which the samples are selected must have equal variances (Breusch Pagan test)

```
        studentized Breusch-Pagan test

data:  model20
BP = 9.7546, df = 5, p-value = 0.0825
```

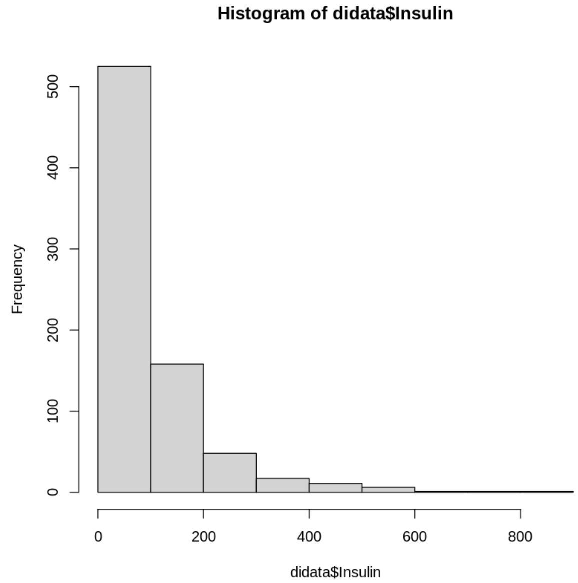
The normality assumption is not valid but it's the least important assumption.

There is no interaction.

Age:Outcome score is not significant.

We received only significant score for Age variable what means that Age is related with BloodPressure

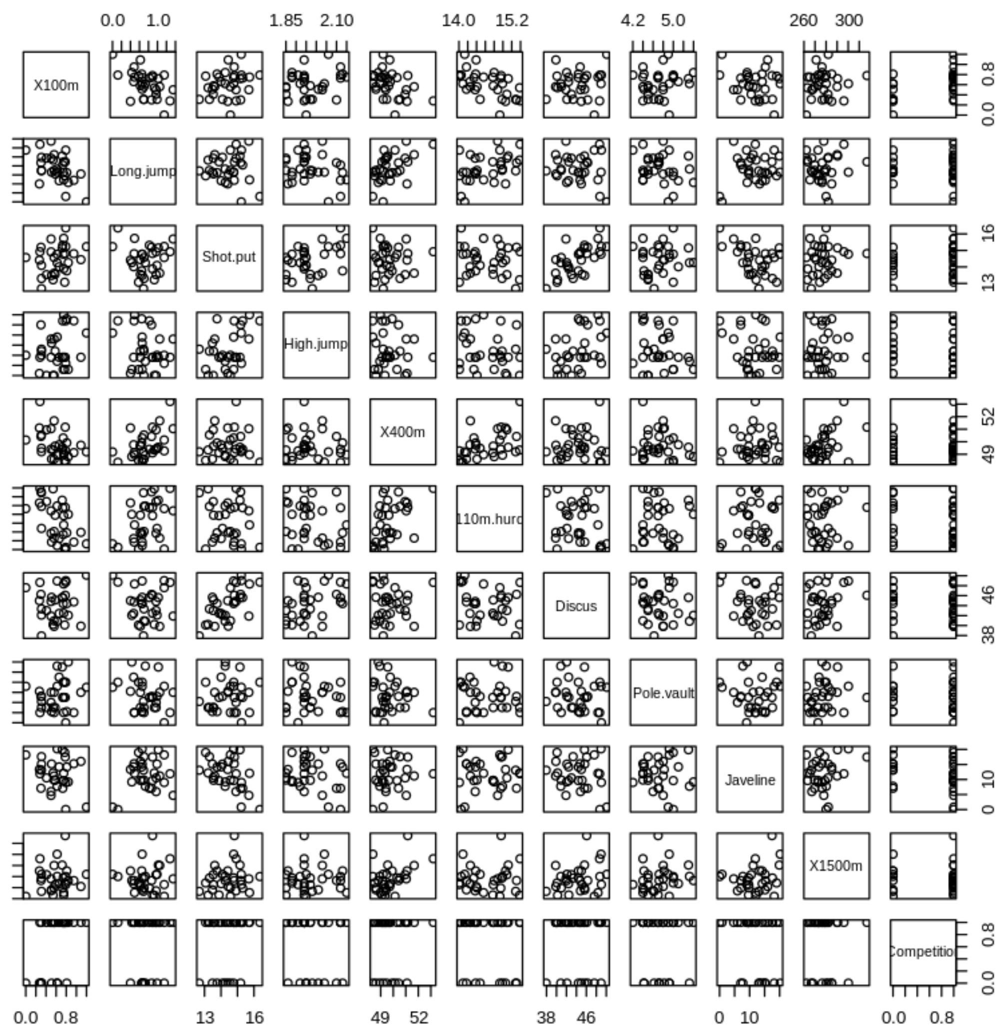
- Analyze the distribution of Insulin variable. What would you recommend to fit an ANOVA model on Insulin levels?



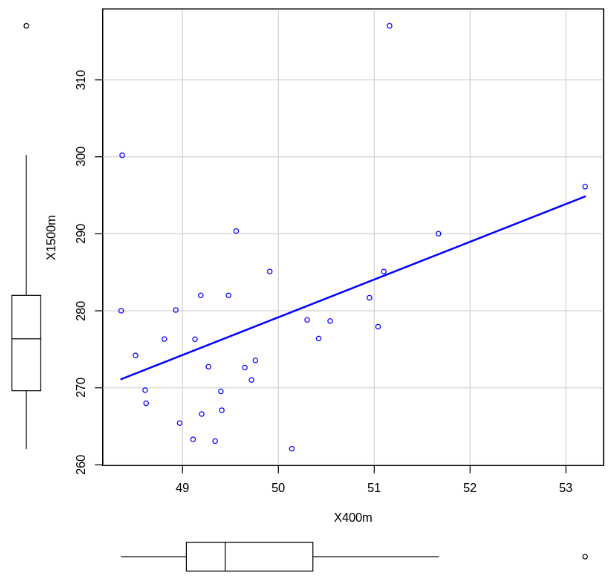
I would recommend checking the dataset and removing errors and outliers. The distribution should be more similar to Normal distribution.

3 THIRD QUESTION: DEFINE A LINEAR MODEL FOR AN ATHLETE IN THE 1500 M

3.1 Split data into train and test set



The highest correlation with 1500m is for 400m, so we will create simple linear regression with this feature.




```
Call:
lm(formula = X1500m ~ X400m, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-17.749  -6.411  -2.385   3.596  32.152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   34.105     87.982   0.388  0.70102
X400m          4.901      1.768   2.772  0.00949 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.66 on 30 degrees of freedom
Multiple R-squared:  0.2039,	Adjusted R-squared:  0.1773
F-statistic: 7.683 on 1 and 30 DF,  p-value: 0.009485
```

We can see that the decathlon$“400m” was marked by two stars. That means that variable is significant but not very strong (we didn’t receive three stars, which is a maximum).

The Residual standard error is equal 10.66, comparing it to the mean of the decathlon“1500m” (the variable we want to predict) which is equal to 277.9, we can say that the model is already relatively (quite) good. The Multiple R-squared is equal to 0.2039 which means that 20% of the variation of the response variable can be explained by using the decathlon$“400m” variable as the independent variable.

3.2 Multiple Linear Regression Model

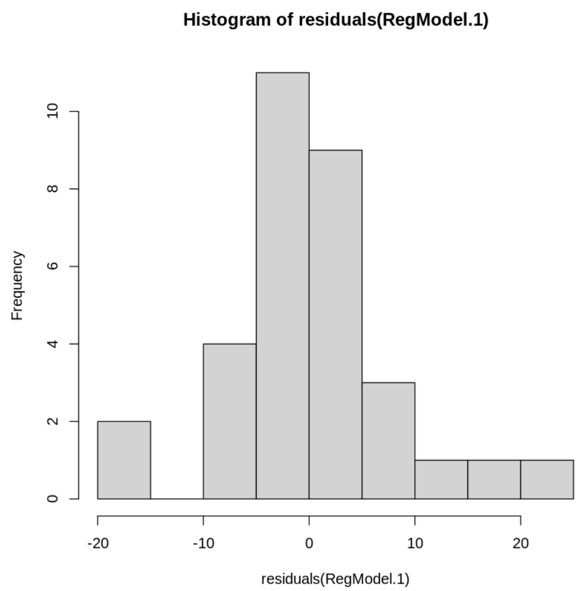
```
Call:
lm(formula = X1500m ~ X400m + X100m + Long.jump + Shot.put +
    High.jump + X110m.hurdle + Discus + Pole.vault + Javeline +
    Competition, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-19.3823  -3.4460  -0.2813   3.6758  21.2280

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -174.3305   121.2106  -1.438  0.16510
X400m           7.7429     2.1107   3.668  0.00143 **
X100m          20.9493    11.2484   1.862  0.07660 .
Long.jump      -0.9703     8.2885  -0.117  0.90792
Shot.put        1.2087     3.6820   0.328  0.74594
High.jump      -9.8716    24.8265  -0.398  0.69492
X110m.hurdle   -1.7192     5.1623  -0.333  0.74241
Discus          0.8561     0.8466   1.011  0.32341
Pole.vault      8.7085     7.1380   1.220  0.23598
Javeline        0.6821     0.4153   1.642  0.11543
Competition    -6.8212     5.3146  -1.283  0.21330
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.627 on 21 degrees of freedom
Multiple R-squared:  0.5454,	Adjusted R-squared:  0.3289
F-statistic: 2.519 on 10 and 21 DF,  p-value: 0.03575
```

Only 400m and (a little bit) 100m results seem to be significant.

We can check if we meet assumptions

3.3 Assumptions

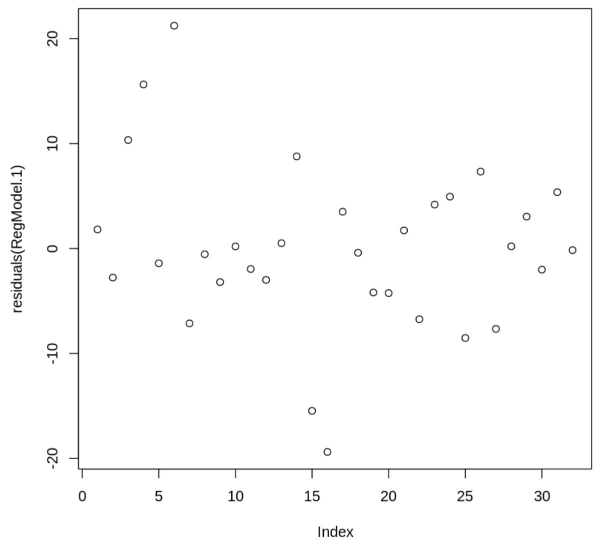
3.3.1 Normality




```
        Shapiro-Wilk normality test

data:  residuals(RegModel.1)
W = 0.96362, p-value = 0.3439
```

We accept the H0 hypothesis (p-value higher than 0.05). The error term follows a Normal distribution. The normality assumption is valid.

3.3.2 Homogenity of Variance




```
        studentized Breusch-Pagan test

data:  RegModel.1
BP = 10.496, df = 10, p-value = 0.3981
```

We accept the H0 hypothesis as the p-value is higher than 0.05. The hypothesis is about homogeneity of variance.

3.3.3 The independence of errors

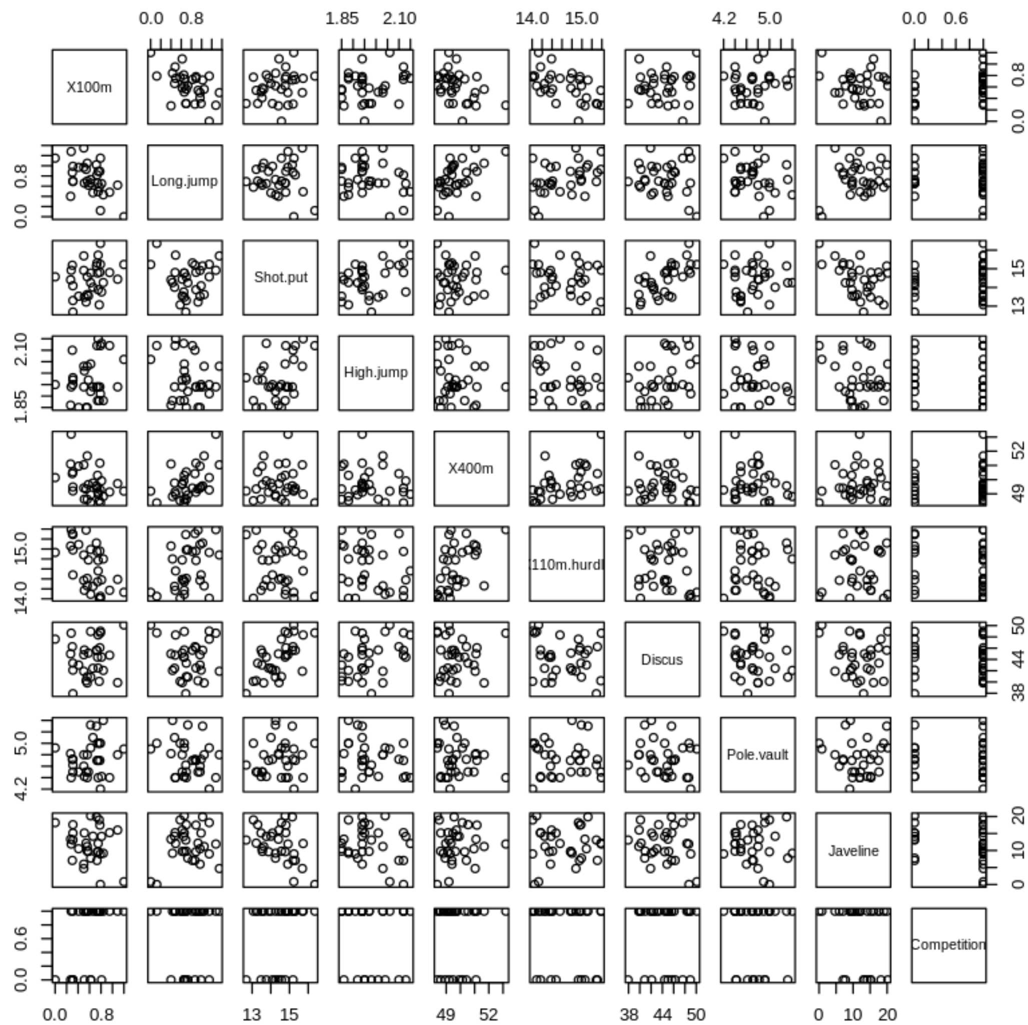
```
        Durbin-Watson test

data:  RegModel.1
DW = 2.1561, p-value = 0.6798
alternative hypothesis: true autocorrelation is not 0
```

There are no autocorrelations in the dataset.

H0 is accepted as p-value is higher than 0.05.

The errors are independent

3.3.4 Multicollinearity



We can see a correlation between Shot.put and Discus.

The highly correlated variables can affect results. So we need to delete one of the correlated variables.

Shot.put variable has the highest vif score, so that is the variable we should delete at the beginning.

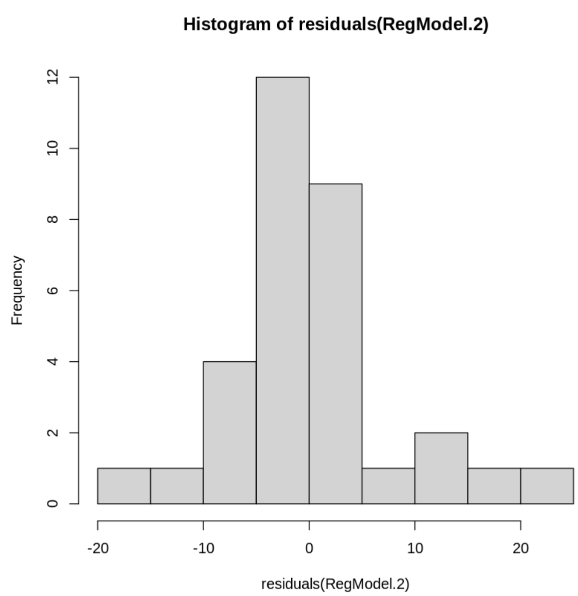
3.4 New Multiple Linear Regression

```
Call:
lm(formula = X1500m ~ X400m + X100m + Long.jump + High.jump +
    X110m.hurdle + Discus + Pole.vault + Javeline + Competition,
```

```
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.4759  -3.3426  -0.4669   3.3364  21.5818 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -173.23376  118.68218  -1.460  0.15852   
X400m           7.69448    2.06238   3.731  0.00116 **
X100m          21.36551   10.94777   1.952  0.06383 . 
Long.jump       0.07842    7.49140   0.010  0.99174   
High.jump      -7.22619   23.00127  -0.314  0.75636   
X110m.hurdle   -1.75596    5.05535  -0.347  0.73163   
Discus          1.04860    0.59831   1.753  0.09360 . 
Pole.vault      9.75740    6.25223   1.561  0.13288   
Javeline        0.62156    0.36452   1.705  0.10225   
Competition    -6.47193    5.10030  -1.269  0.21773   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.43 on 22 degrees of freedom
Multiple R-squared:  0.543,	Adjusted R-squared:  0.3561 
F-statistic: 2.905 on 9 and 22 DF,  p-value: 0.0199
```

3.5 Checking assumptions for the new model

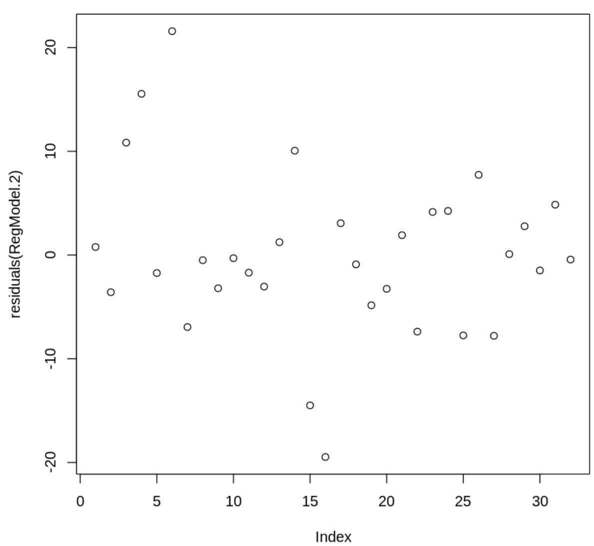
3.5.1 Normality



```
        Shapiro-Wilk normality test

data:  residuals(RegModel.2)
W = 0.95948, p-value = 0.2659
```

We accept H0 hypothesis (p-value higher than 0.05). The error term follows a Normal distribution. The normality assumption is valid.

3.5.2 Homogenity of Variance




```
        studentized Breusch-Pagan test

data:  RegModel.2
BP = 8.0252, df = 9, p-value = 0.5316
```

We accept the H0 hypothesis as the p-value is higher than 0.05. The hypothesis is about homogeneity of variance.

3.5.3 The independence of errors

```
        Durbin-Watson test
```

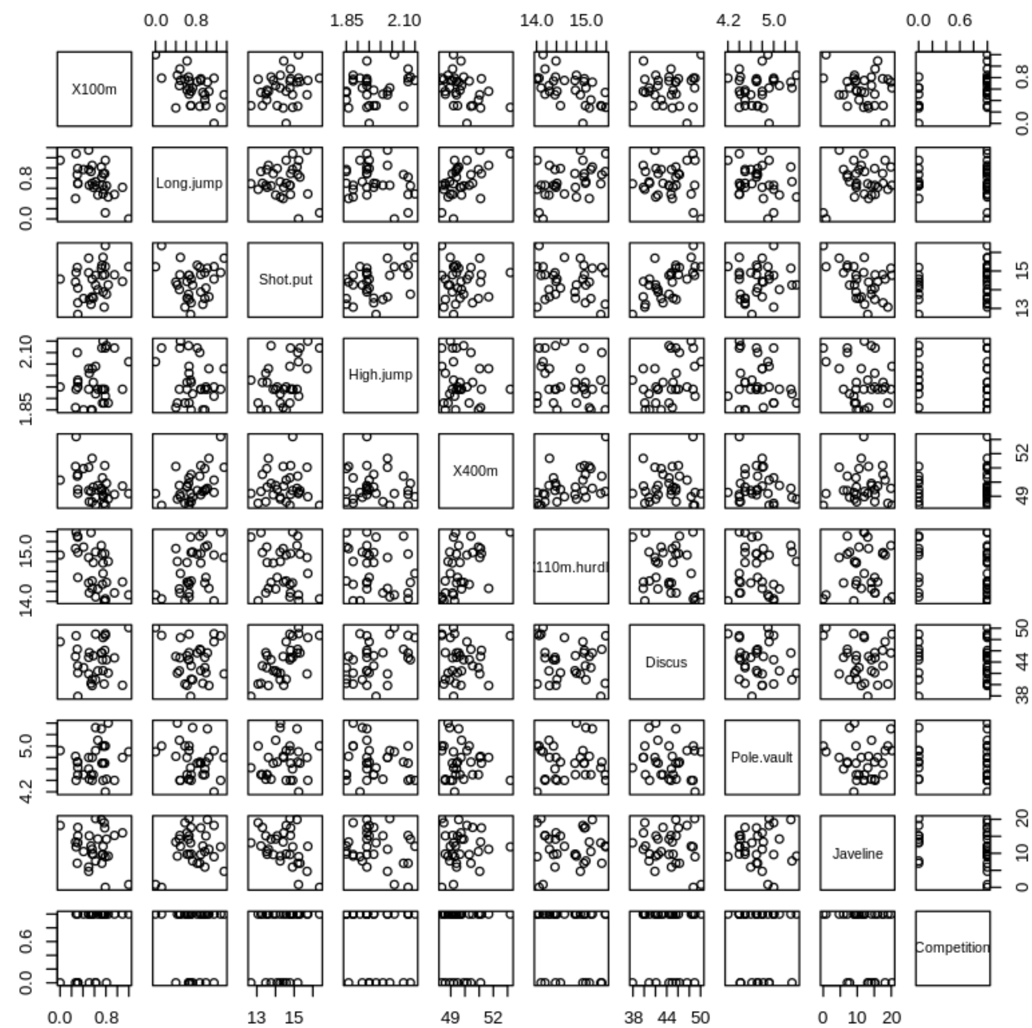
```
data:  RegModel.2
DW = 2.1766, p-value = 0.5921
alternative hypothesis: true autocorrelation is not 0
```

There are no autocorrelations in the dataset.

H0 is accepted as p-value is higher than 0.05.

The errors are independent

3.5.4 Multicollinearity




```
Start:  AIC=151.62
```

```
X1500m ~ X400m + X100m + Long.jump + High.jump + X110m.hurdle +
    Discus + Pole.vault + Javeline + Competition

               Df Sum of Sq    RSS    AIC
- Long.jump     1      0.01 1956.4 149.62
- High.jump     1      8.78 1965.2 149.76
- X110m.hurdle  1     10.73 1967.2 149.79
<none>                      1956.4 151.62
- Competition   1    143.19 2099.6 151.88
- Pole.vault    1    216.59 2173.0 152.98
- Javeline      1    258.56 2215.0 153.59
- Discus        1    273.15 2229.6 153.80
- X100m         1    338.70 2295.1 154.73
- X400m         1   1237.83 3194.3 165.31

Step:  AIC=149.62
X1500m ~ X400m + X100m + High.jump + X110m.hurdle + Discus +
    Pole.vault + Javeline + Competition

               Df Sum of Sq    RSS    AIC
- High.jump     1      8.83 1965.3 147.76
- X110m.hurdle  1     10.74 1967.2 147.80
<none>                      1956.4 149.62
- Competition   1    147.00 2103.4 149.94
- Pole.vault    1    223.04 2179.5 151.07
- Javeline      1    264.74 2221.2 151.68
- Discus        1    275.14 2231.6 151.83
- X100m         1    375.64 2332.1 153.24
- X400m         1   1341.20 3297.6 164.33

Step:  AIC=147.76
X1500m ~ X400m + X100m + X110m.hurdle + Discus + Pole.vault +
    Javeline + Competition

               Df Sum of Sq    RSS    AIC
- X110m.hurdle  1     10.33 1975.6 145.93
<none>                      1965.3 147.76
- Competition   1    139.87 2105.1 147.97
- Pole.vault    1    256.86 2222.1 149.69
- Discus        1    274.34 2239.6 149.95
- Javeline      1    297.63 2262.9 150.28
- X100m         1    376.28 2341.5 151.37
- X400m         1   1334.38 3299.6 162.35

Step:  AIC=145.93
X1500m ~ X400m + X100m + Discus + Pole.vault + Javeline + Competition

              Df Sum of Sq    RSS    AIC
```

```
<none>                        1975.6 145.93
- Competition  1    142.38 2118.0 146.16
- Pole.vault   1    247.59 2223.2 147.71
- Discus       1    305.12 2280.7 148.53
- Javeline     1    310.33 2285.9 148.60
- X100m        1    502.94 2478.5 151.19
- X400m        1   1421.08 3396.7 161.27


Call:
lm(formula = X1500m ~ X400m + X100m + Discus + Pole.vault + Javeline +
    Competition, data = train)

Coefficients:
(Intercept)        X400m        X100m       Discus   Pole.vault     Javeline
  -201.0539       7.4434      21.6470       1.0198       9.9059       0.6564
Competition
    -5.9655
```

The lowest AIC value the better. From the output of step() function we can see actual AIC score and AIC score when each variable will be deleted. Step by step we are deleting the variables to receive a model with as low AIC score as possible. When we see that deleting more variables will increase the AIC score we know that we already found the best multiple linear model.

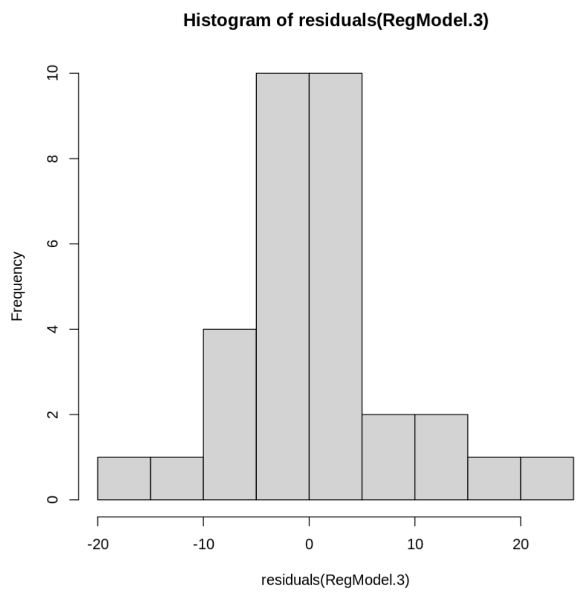
3.6 Third - final model

```
Call:
lm(formula = X1500m ~ X400m + X100m + Discus + Pole.vault + Javeline +
    Competition, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-19.2804  -3.5023  -0.2449   2.8593  21.3204

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -201.0539    95.4272  -2.107 0.045331 *
X400m          7.4434     1.7553   4.241 0.000267 ***
X100m         21.6470     8.5806   2.523 0.018377 *
Discus         1.0198     0.5190   1.965 0.060621 .
Pole.vault     9.9059     5.5964   1.770 0.088914 .
Javeline       0.6564     0.3312   1.982 0.058612 .
Competition   -5.9655     4.4443  -1.342 0.191573
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.89 on 25 degrees of freedom
```

```
Multiple R-squared:  0.5386,	Adjusted R-squared:  0.4278
F-statistic: 4.863 on 6 and 25 DF,  p-value: 0.002038
```

3.7 Checking assumptions for the final model

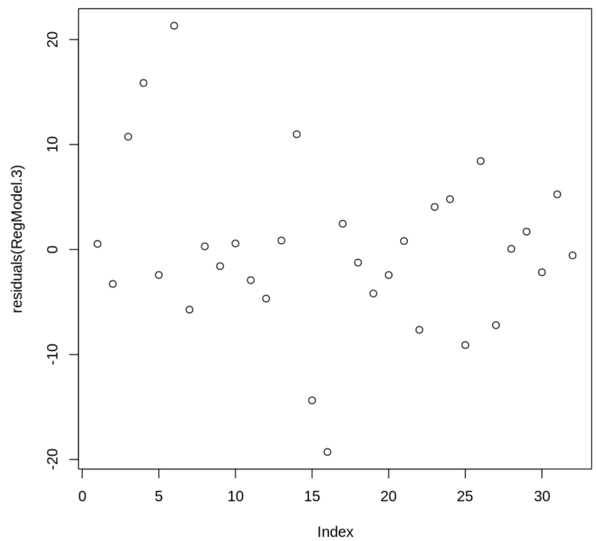
3.7.1 Normality




```
        Shapiro-Wilk normality test

data:  residuals(RegModel.3)
W = 0.95868, p-value = 0.2526
```

We accept the H0 hypothesis (p-value higher than 0.05). The error term follows a Normal distribution. The normality assumption is valid.

3.7.2 Homogenity of Variance




```
        studentized Breusch-Pagan test

data:  RegModel.3
BP = 7.3562, df = 6, p-value = 0.2892
```

We accept the H0 hypothesis as the p-value is higher than 0.05. The hypothesis is about homogeneity of variance.

3.7.3 The independence of errors

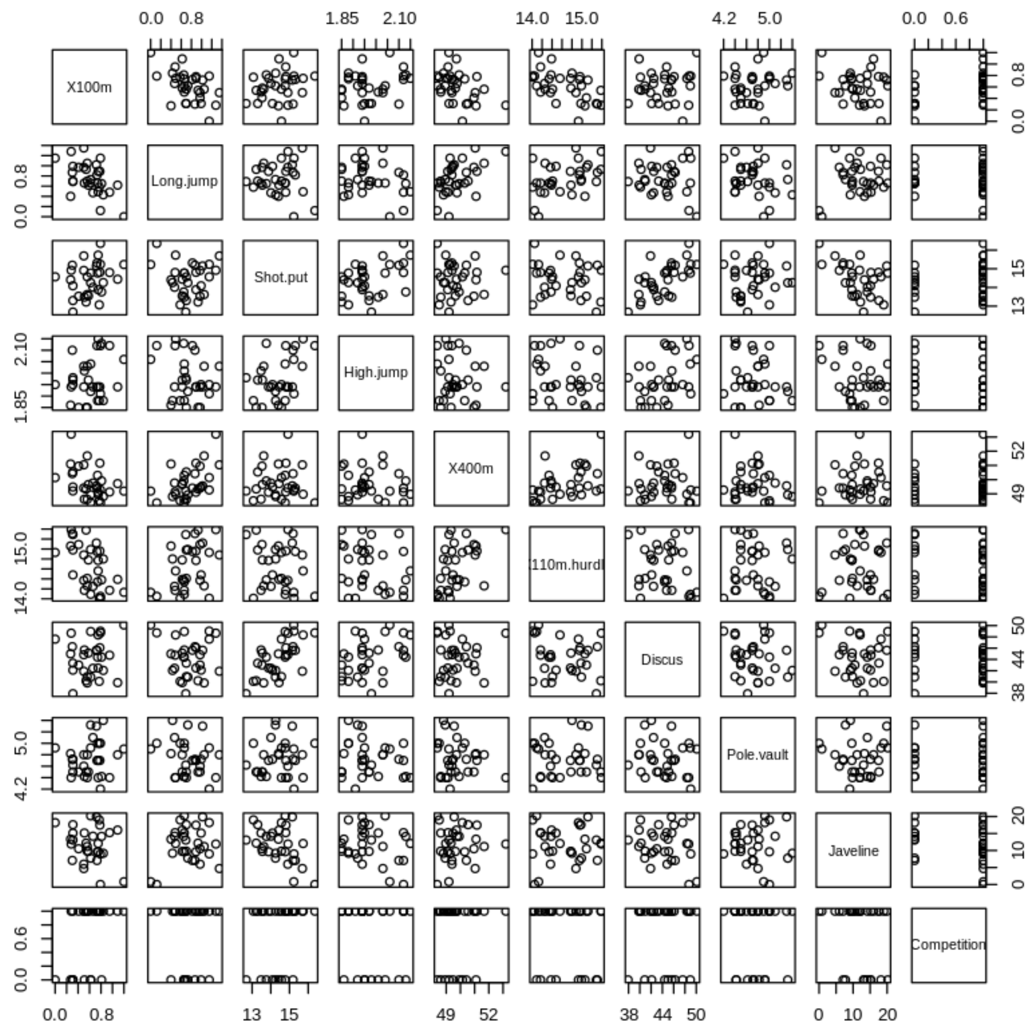
```
        Durbin-Watson test

data:  RegModel.3
DW = 2.2015, p-value = 0.4967
alternative hypothesis: true autocorrelation is not 0
```

There are no autocorrelations in the dataset.

H0 is accepted as p-value is higher than 0.05.

The errors are independent

3.7.4 Multicollinearity



3.8 Prediction on the test set

	fit	lwr	upr
SEBRLE	270.9752	249.1711	292.7793
CLAY	281.8743	260.4247	303.3239
WARNERS	264.8520	244.5093	285.1946
BOURGUIGNON	276.3583	254.8415	297.8751
Karpov	263.2864	240.6217	285.9512
Warners	264.3165	244.4791	284.1538
Hernu	260.8604	240.0874	281.6333
Pogorelov	281.3147	261.6349	300.9945
Ojaniemi	262.6185	242.4955	282.7416

[512]:
```
y <-test$'X1500m'
y
```

291.7

301.5

278.1

291.7

278.11

278.05

264.35

287.63

275.71

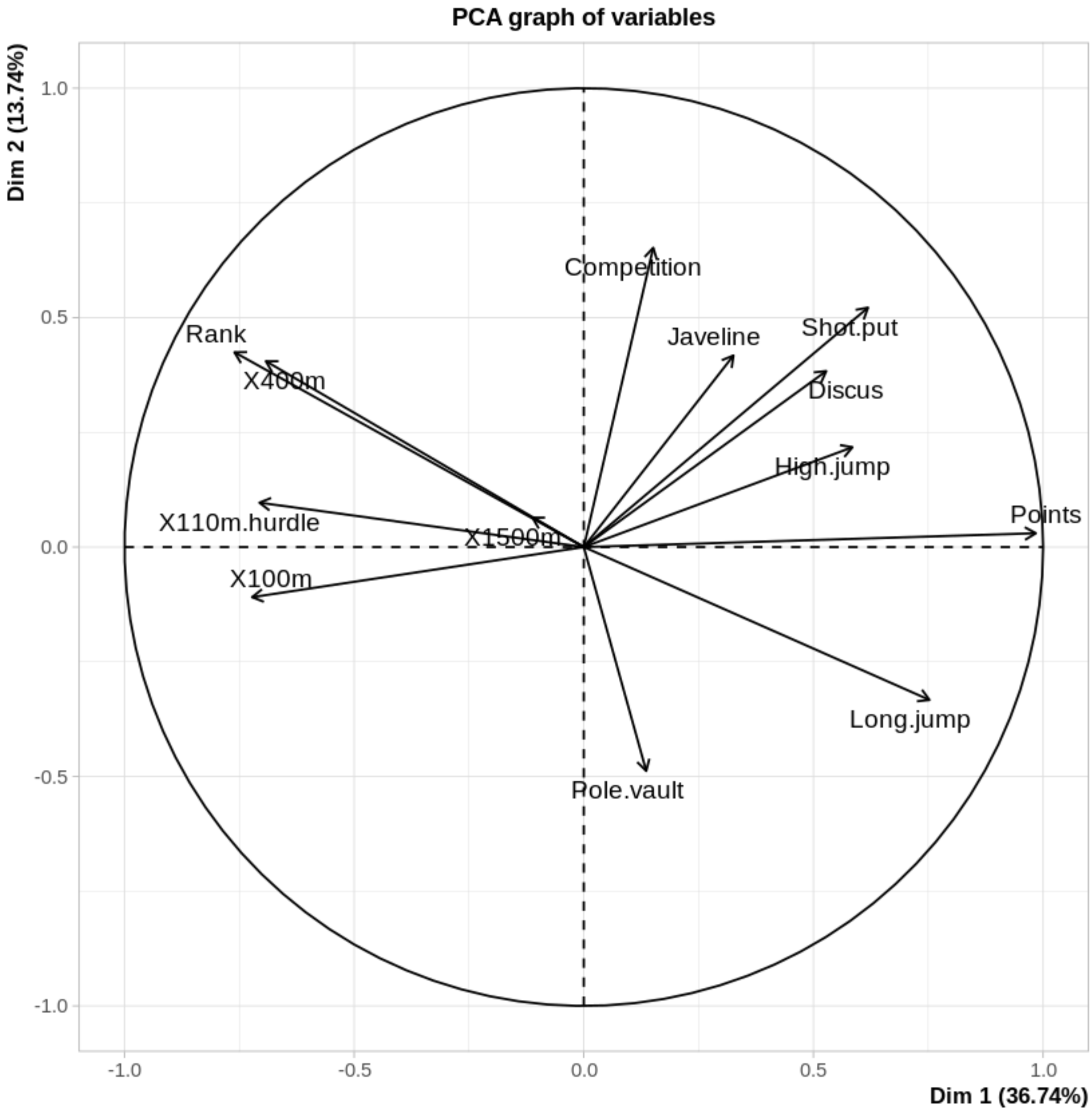
Now we can compare fitted values with real values. We can also check that all fitted values are in intervals (between lwr and upr value). That means our model works well.

RMSE value is smaller for the training set, which is correct as our model always should better fit for data it was training on than for a new one.

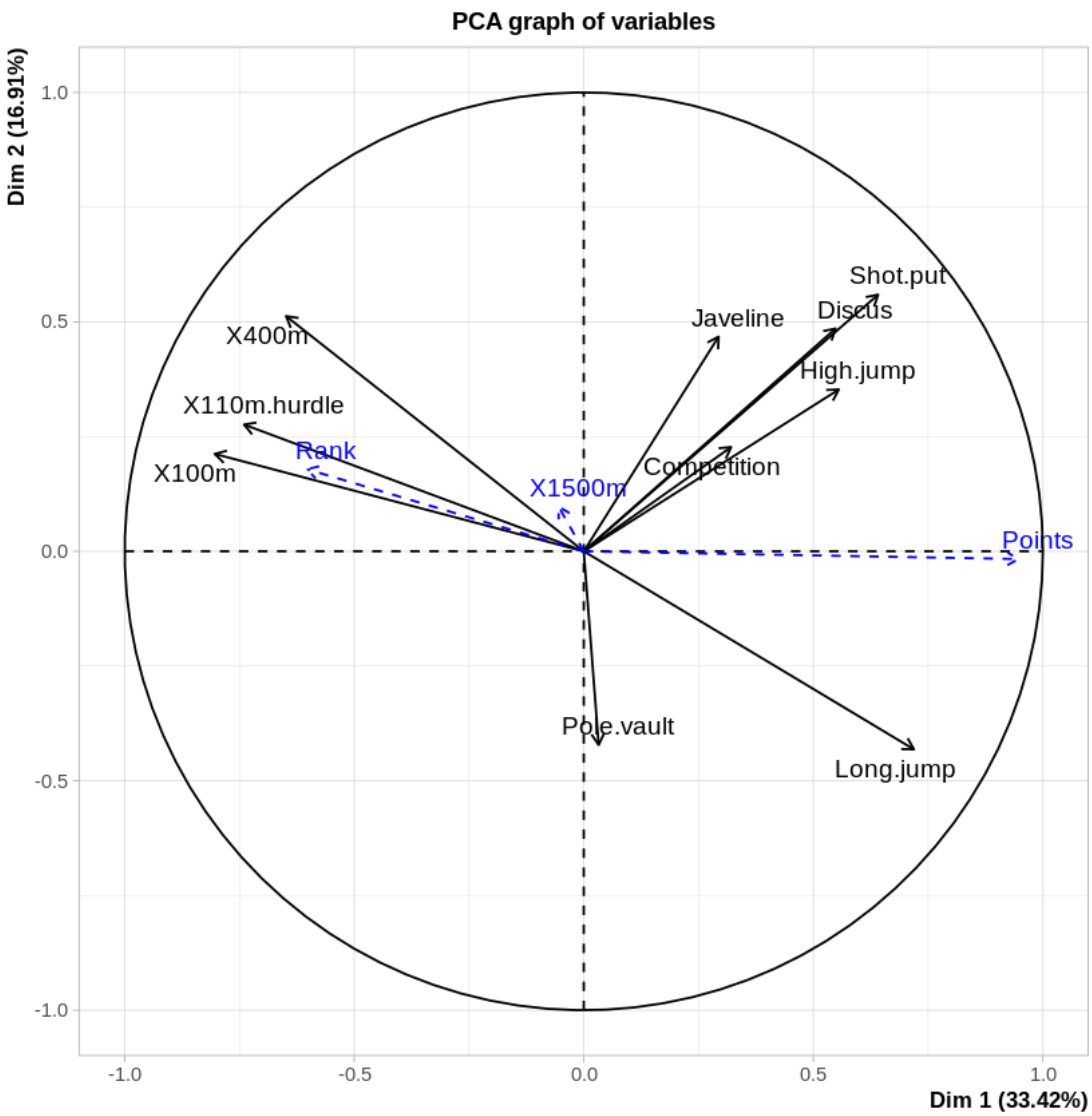
RMSE for train set -> 7.85730086630928

RMSE for test set -> 14.3656946422001

4 FOURTH QUESTION: WORKING WITH REAL DATA



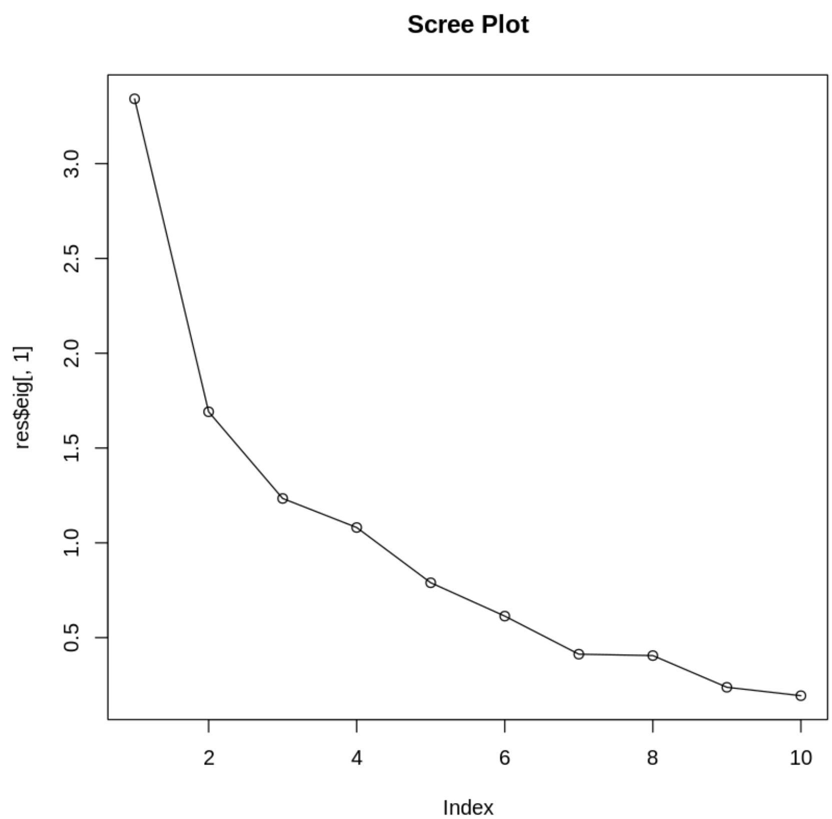
In PCA we should not consider the X1500m variable, as we want to predict this variable. Also from the previous task we know that we also should not include Rank and Points variables in our model.



After eliminating X1500m, Points and Rank variables from PCA the variation in the first dimension decreased and the variation in the second dimension increased.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.3416485	33.416485	33.41649
comp 2	1.6910124	16.910124	50.32661
comp 3	1.2338861	12.338861	62.66547
comp 4	1.0807109	10.807109	73.47258
comp 5	0.7894027	7.894027	81.36661
comp 6	0.6134908	6.134908	87.50151
comp 7	0.4126831	4.126831	91.62834
comp 8	0.4052996	4.052996	95.68134
comp 9	0.2380949	2.380949	98.06229
comp 10	0.1937711	1.937711	100.00000

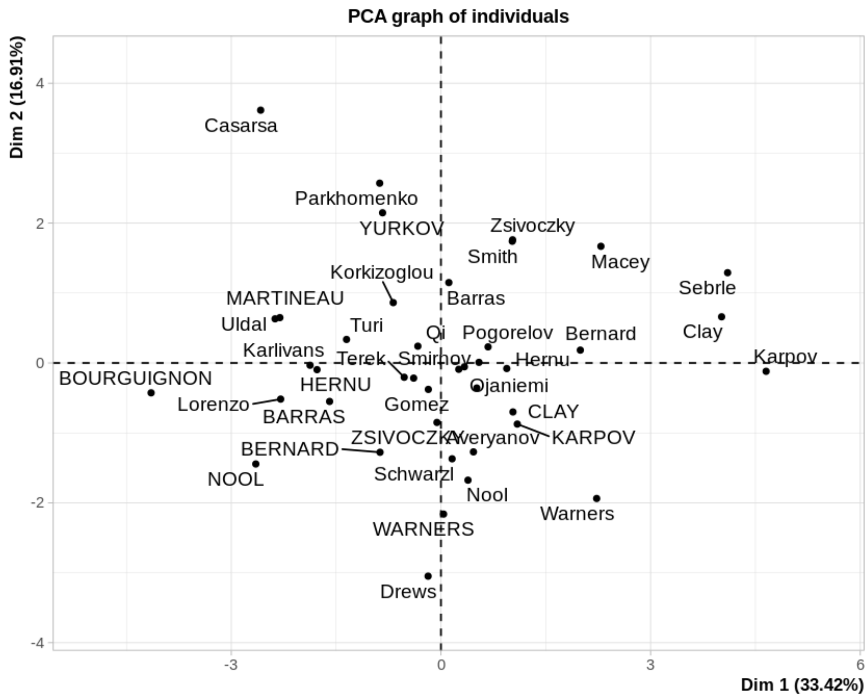
We can check the variation in each principal component, also the cumulative percentage of variance, which means we know how many principal components we should leave to remain a specific percentage of variance.



We should find the elbow, so the number of principal components where eigenvalues are leveling off. On this graph it is 3 principal components, this allows us to keep almost 63% of variation.

	Dim.1	Dim.2	Dim.3
X100m	-0.80375792	0.2121185	0.36644639
Long.jump	0.71902293	-0.4315127	0.10507862
Shot.put	0.64115334	0.5592746	0.09576916
High.jump	0.55546011	0.3526622	0.39219320
X400m	-0.64840502	0.5123973	-0.08437960
X110m.hurdle	-0.73996769	0.2760499	-0.02235663
Discus	0.54841299	0.4854139	0.36783035
Pole.vault	0.03238867	-0.4217449	0.13236792
Javeline	0.29365964	0.4678044	-0.22591777
Competition	0.32092230	0.2270800	-0.84504163

We can check loading of each variable in each dimension. The absolute value of the variable is the value of loadings. The first dimension is an indicator mainly of 'X100m', 'X110m.hurdle' and 'Long.jump' variables. The second dimension is an indicator mainly of 'Shot.put' and 'X400m' variables. The third dimension is an indicator mainly of the ‘Competition’ variable.



4.0.1 Principal Component Regression

We can notice that between PCs there is almost no correlation.

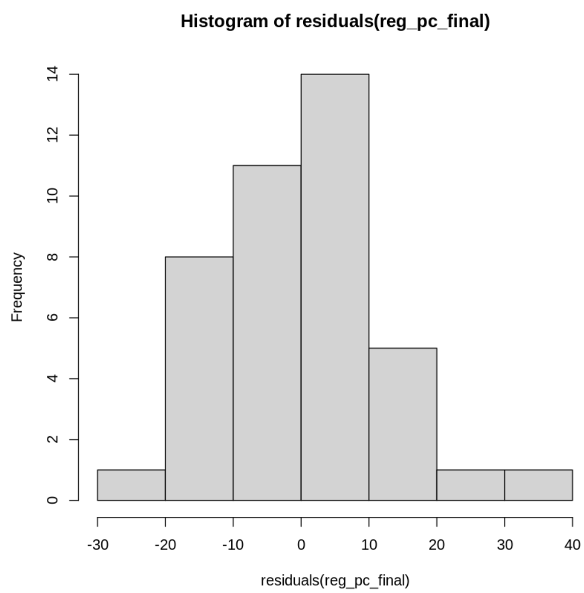
```
Call:
lm(formula = X1500m ~ PC1 + PC2 + PC3, data = decathlon)

Residuals:
    Min      1Q  Median      3Q     Max
-23.291  -5.933   0.166   4.245  38.333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 279.0249     1.8411 151.555   <2e-16 ***
PC1          -0.3440     1.0071  -0.342    0.735
PC2           0.8602     1.4158   0.608    0.547
PC3           2.1831     1.6574   1.317    0.196
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.79 on 37 degrees of freedom
Multiple R-squared:  0.05662,	Adjusted R-squared:  -0.01987
F-statistic: 0.7403 on 3 and 37 DF,  p-value: 0.5348
```

There is no multicollinearity as we are using principals components (no correlation between PCs)

Assumptions

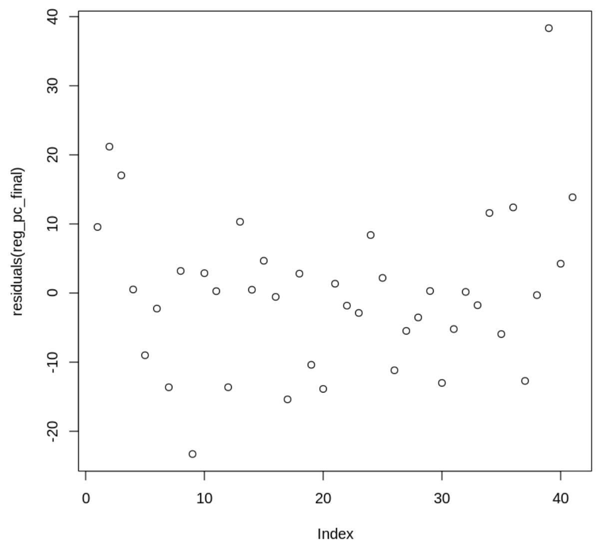
Normality




```
	Shapiro-Wilk normality test

data:  residuals(reg_pc_final)
W = 0.94901, p-value = 0.06459
```

p-value > 0.05, test is valid

Homogenity of Variance




```
        studentized Breusch-Pagan test

data:  reg_pc_final
BP = 2.6049, df = 3, p-value = 0.4566
```

p-value > 0.05, test is valid

The independence of errors

```
        Durbin-Watson test

data:  reg_pc_final
DW = 1.8034, p-value = 0.3867
alternative hypothesis: true autocorrelation is not 0
```

p-value > 0.05, test is valid

All assumptions are valid.

RMSE = 11.1988259897779