



**UNIVERSITAT POLITÈCNICA  
DE CATALUNYA  
BARCELONATECH**

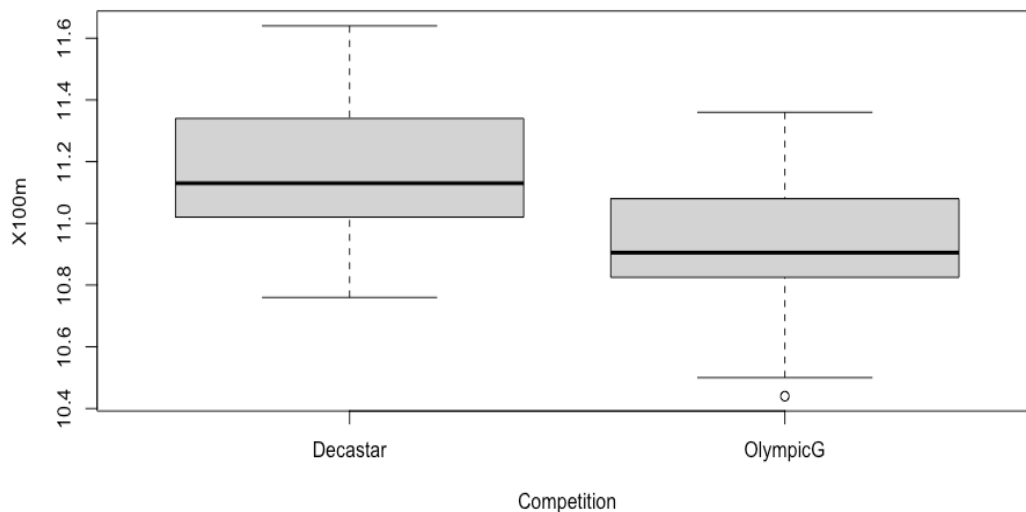
Report:  
SMDE - Assignment 1

Submitted by  
Mohana Fathollahi

April 21, 2020

## 1. First Question

a) Analyze the distribution of “X100m” according to the type of competition by using boxplot. Write your conclusion.



	Decastar	OlympicG	Description
IQR	0.35	0.3	Decastar has more dispersed data (variation of Decastar is higher than OlympicG)
Meidan	11.1	10.9	
Outlier	Nothing	One	
Skewness	Left skewed	Left skewed	

In order to have a robust idea it is better to have a data with less variation.

b) 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?

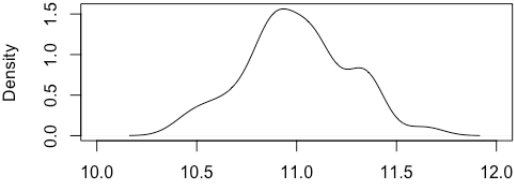
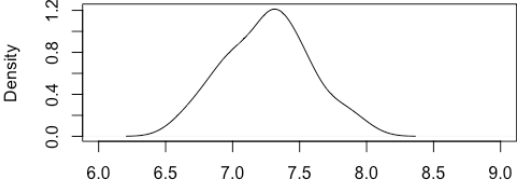
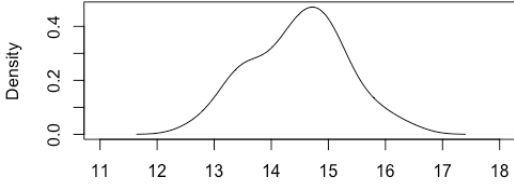
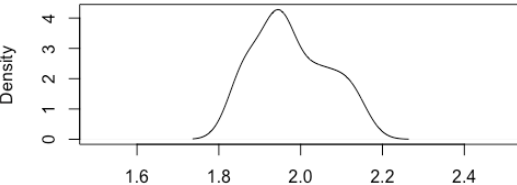
Write your conclusion by checking marginal probabilities and test the independence of two variables by using the Chi-Square test.

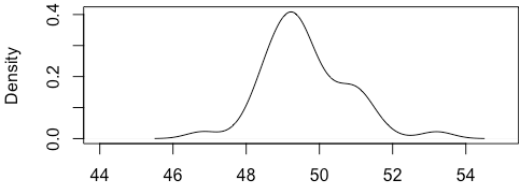
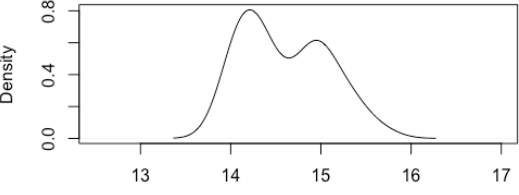
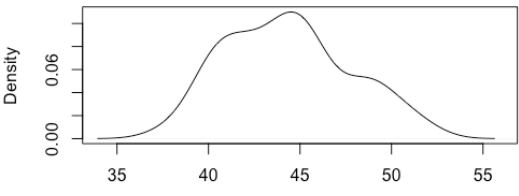
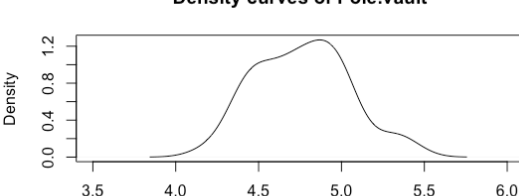
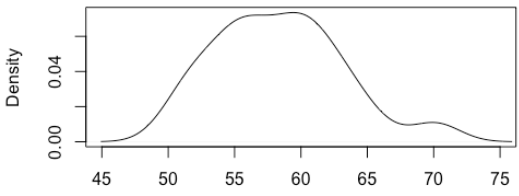
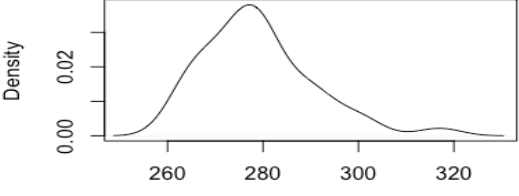
When we look at conditional probability results, we can find that the probability of the variable “x100m” being less than 11 is so different for two groups of competitions. We can find the same conclusion when x100m is larger than 11.

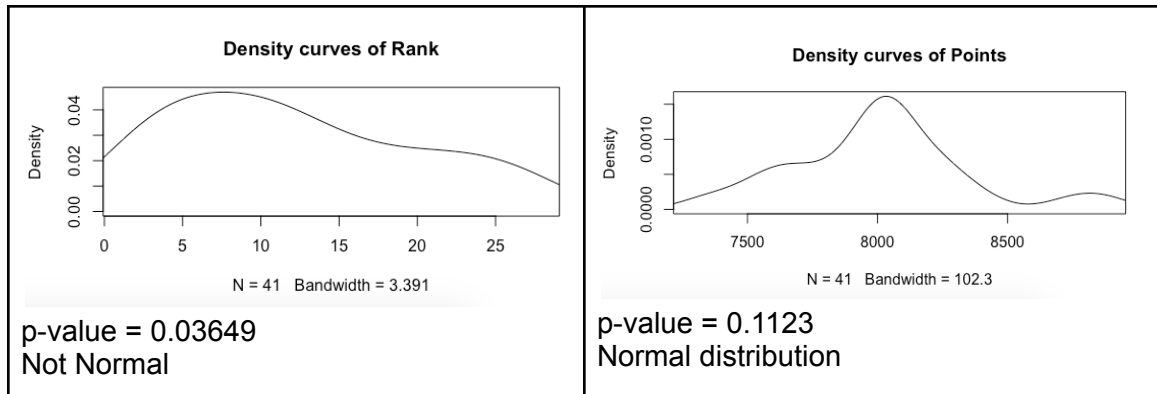
Therefore, we can conclude that these two variables are not independent.

The result of p-value is 0.005236 that is less than 0.05, so we reject the null hypothesis. Therefore two variables are dependent.

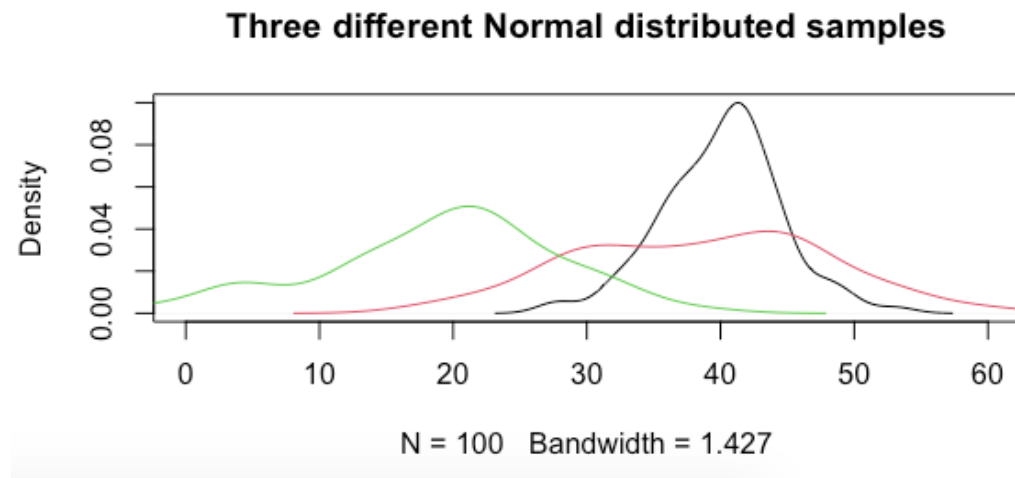
c) Visualize the distribution of quantitative variables by using proper graphs. Which of these variables follow a Normal distribution?

<p><b>Variable: X100m</b></p> <p><b>Density curves of X100m</b></p>  <p>N = 41 Bandwidth = 0.09268</p> <p>p-value = 0.7435 Accept (p-value&gt;0.05) normal distribution</p>	<p><b>Variable: Long.jump</b></p> <p><b>Density curves of Long.jump</b></p>  <p>N = 41 Bandwidth = 0.1355</p> <p>p-value = 0.9289 Accept (p-value&gt;0.05) normal distribution</p>
<p><b>Variable: Shot.put</b></p> <p><b>Density curves of Shot.put</b></p>  <p>N = 41 Bandwidth = 0.3483</p> <p>p-value = 0.9456 Normal distribution</p>	<p><b>Variable: High.jump</b></p> <p><b>Density curves of High.jump</b></p>  <p>N = 41 Bandwidth = 0.03809</p> <p>p-value = 0.0255 Not normal</p>
<p><b>Variable: X400m</b></p>	<p><b>Variable: X110m.hurdle</b></p>

<p style="text-align: center;"><b>Density curves of X400m</b></p>  <p style="text-align: center;">N = 41 Bandwidth = 0.4378</p> <p>p-value = 0.1248 Normal distribution</p>	<p style="text-align: center;"><b>Density curves of X110m.hurdle</b></p>  <p style="text-align: center;">N = 41 Bandwidth = 0.202</p> <p>p-value = 0.01544 Not normal</p>
<p style="text-align: center;"><b>Density curves of Discus</b></p>  <p style="text-align: center;">N = 41 Bandwidth = 1.333</p> <p>p-value = 0.3385 Normal distribution</p>	<p style="text-align: center;"><b>Density curves of Pole.vault</b></p>  <p style="text-align: center;">N = 41 Bandwidth = 0.1191</p> <p>p-value = 0.3456 Normal distribution</p>
<p style="text-align: center;"><b>Density curves of Javeline</b></p>  <p style="text-align: center;">N = 41 Bandwidth = 1.796</p> <p>p-value = 0.3732 Normal distribution</p>	<p style="text-align: center;"><b>Density curves of X1500m</b></p>  <p style="text-align: center;">N = 41 Bandwidth = 4.5</p> <p>p-value = 0.02391 Not normal</p>
<p><b>Variable: Rank</b></p>	<p><b>Variable: Points</b></p>

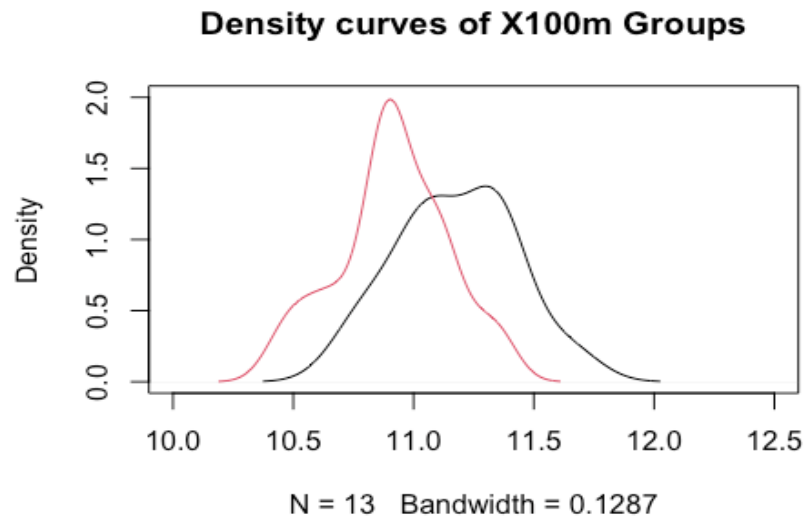


d) Generate three Normally distributed random variables of length 50. Two of them should have the same mean, different standard deviations while the third one has a different mean but the same standard deviation with the first distribution. Use t test to compare mean differences between three variables.



Dist_first= N(mean=40, sd=4)	Dist_second= N(mean=40, sd=10)	Dist_third= N(mean=20, sd=10)
data: Dist_first and Dist_second t = 1.5315, df = 198, p-value = 0.1272 > 0.05 null hypothesis accepted.		
data: Dist_third and Dist_second t = -15.035, df = 198, p-value < 2.2e-16 < 0.05 null hypothesis rejected.		
data: Dist_first and Dist_third t = 20.683, df = 198, p-value < 2.2e-16 < 0.05 null hypothesis rejected.		

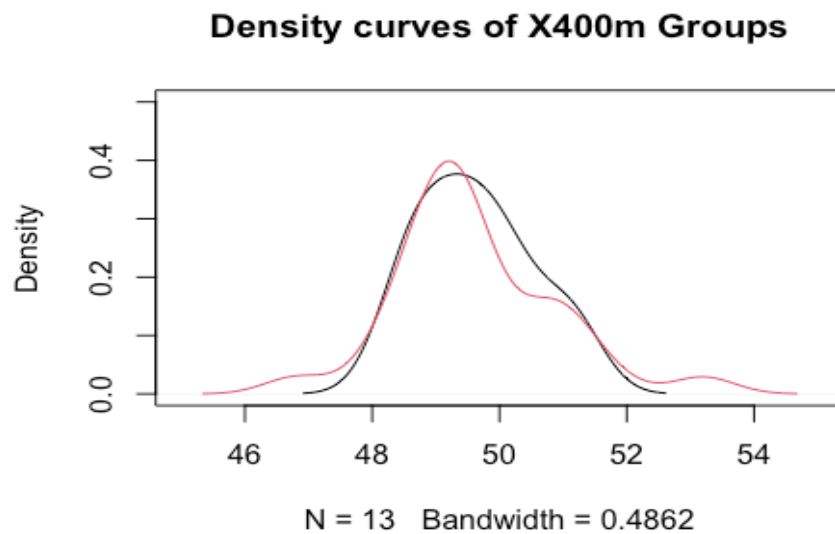
e) Test if there is a difference between two types of competitions according to the variables “X100m” and “X400m” by using t test.



```
# data: decathlon$X100m[g1] and decathlon$X100m[g2]
```

```
# t = 3.2811, df = 39, p-value = 0.002184
```

Based on this plot and p-value of t-test for two groups of competition we can find that the mean of x100m for two groups is different. P-value = 0.0021 < 0.05 that means null hypothesis is rejected.



```
# data: decathlon$X400m[g1] and decathlon$X400m[g2]
```

# t = 0.051016, df = 39, p-value = 0.9596

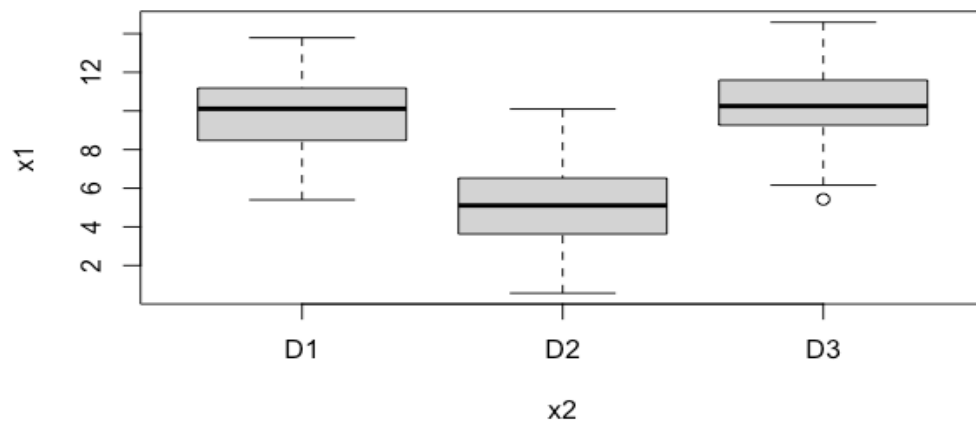
Based on this plot and p-value of t-test for two groups of competition we can find that the mean of x400m for two groups is not different. P-value = 0.96 > 0.05 that means null hypothesis is accepted.

## 2. Second Question

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=0, the second with mean=10, and the third with mean=0. Select the SAME variance for the three distributions at your convenience (a value >0).

We want to analyze using an ANOVA if these three populations are different (or not) depending on the parameter selected.

As expected from the code, the ANOVA results show that p-value is near zero, therefore there is enough evidence to reject the null-hypothesis of similarity. We can confirm that based on boxplot the second distribution has a different mean.



Assumptions:

1. Independency: based on Durbin-watson result; p-value = 0.6422 > 0.05. Therefore our data are independent.
2. Normality: based on shapiro test result; p-value = 0.7193 > 0.05. Therefore our data are normal.
3. Homogeneity of variance: based on Breusch-Pagan test, p-value = 0.7629. So our data have equal variances.

Once you finish the analysis and you are familiar with ANOVA test: on the dataset contained on Kaggle, named “Red and White Wine Quality”, we want to analyze if in both (type or quality) affects some properties of the wine. After combining the two datasets (one for red wines and one for white wines), you should create two variables. First, “type” that identifies if the wine is red or white, and second, wine quality categorized in three groups: <5 (low), 5-6 (medium) and >6 (high). Once you complete preprocessing steps, please answer to the following questions applying appropriate statistical techniques:

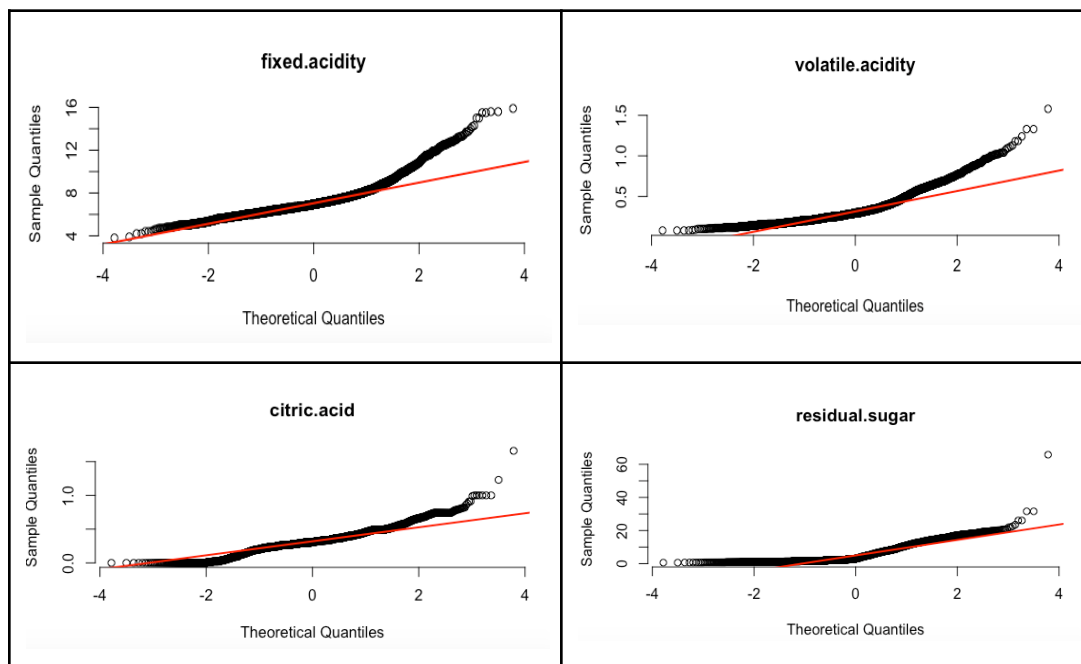
1) Which of the chemical properties influence the quality of the wines?

After applying the ANOVA model, we can find that all p-values are less than 0.05. Therefore, chemical properties are influenced by the quality of wine.

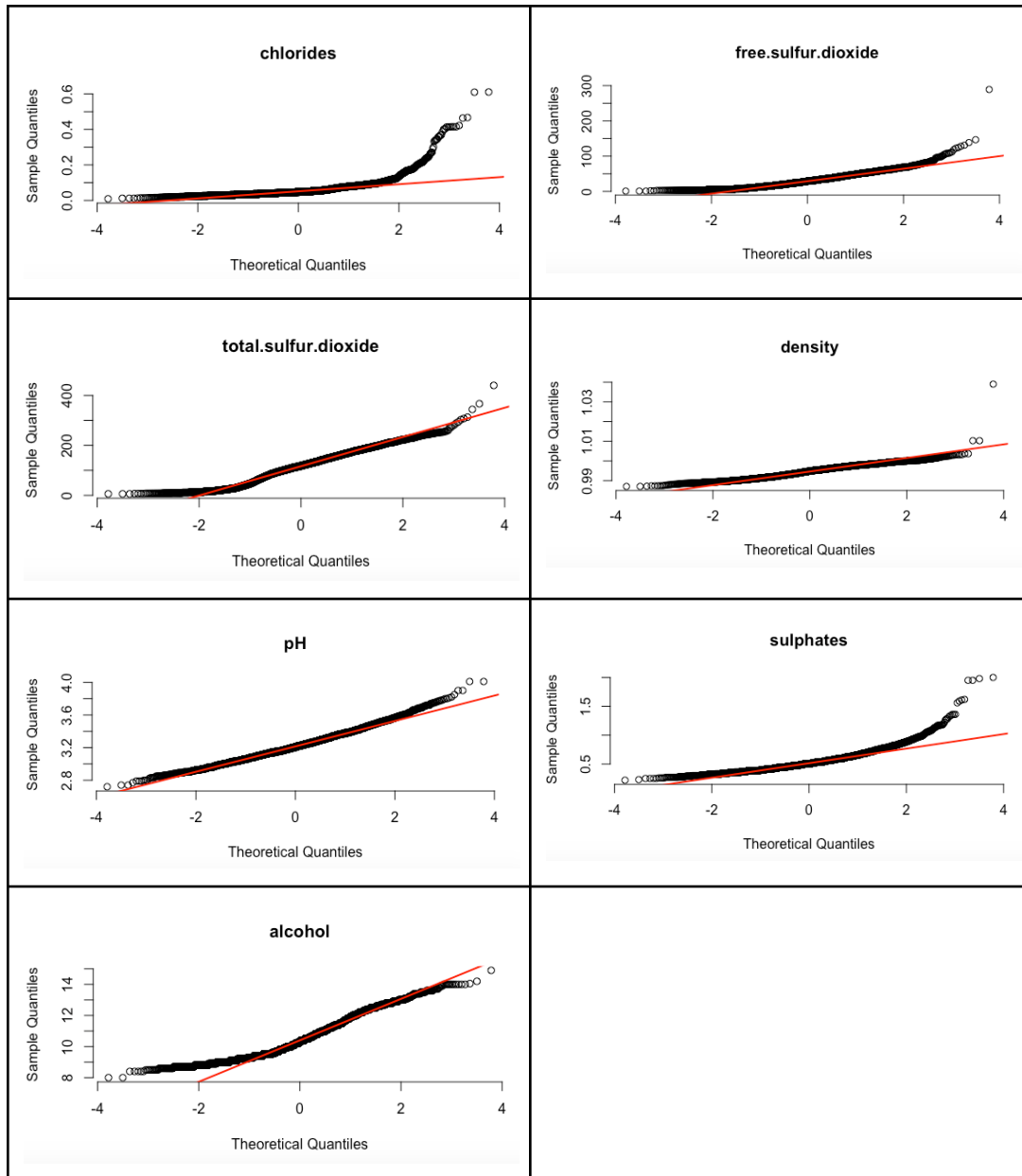
### Assumptions:

1. Independency: based on Durbin-watson result, all p-values are less than 0.05. So this assumption is not valid for the wine dataset.

2. Normality: based on the result of Q-Q plots just three of the variables are normal (Density, total.sulfur.dioxide, ph).







3. Homogeneity of variance: based on Breusch-Pagan test, p-values for just these three variables are larger than 0.05. For other variables, p-values are less than 0.05

Fixed.acidity: p-value = 0.4633, density: p-value = 0.2982, sulphates: p-value = 0.05731

Based on results of ANOVA assumptions, our ANOVA model is not reliable.

2) Which of the chemical properties are related with type of the wines?

After applying the ANOVA model, we can find that all p-values are less than 0.05. Therefore , chemical properties are influenced by the type of wine.

1. Independency: based on Durbin-watson result, all p-values are less than 0.05. So this assumption is not valid for the wine dataset.

2. We have same result as question number one

3. Homogeneity of variance: based on Breusch-Pagan test, p-values for just one variable is larger than 0.05. For other variables, p-values are less than 0.05

pH: p-value = 0.339

Based on results of ANOVA assumptions, our ANOVA model is not reliable.

### 3) How does type and quality of wines affect (separately and together) percentage of alcohol present in the wine?

Based on the ANOVA model; type and quality of wine have significant effects on the percentage of alcohol, separately and together. Percentage of alcohol will change between groups according to the type of wine and quality of the wine, because the type of wine and the quality of the wine have a significant effect on the alcohol.

In the second and first part all assumptions about one way ANOVA had been rejected. So we just need to test assumptions for two way ANOVA:

1. Independency: based on Durbin-watson result, p-values is less than 0.05. So this assumption is not valid.
2. Normality: based on Q-Q plot alcohol does not have a normal distribution
3. Homogeneity of variance: based on Leventest, null hypothesis for type and wine-quality are rejected. Therefore, variance of percentage of alcohol is not equal across types of wines and across quality of the wine.

Based on results of ANOVA assumptions, this model is not reliable.

### Post Hoc testing

Based on pairwise t-test with bonferroni correction:

1. Alcohol ~ type

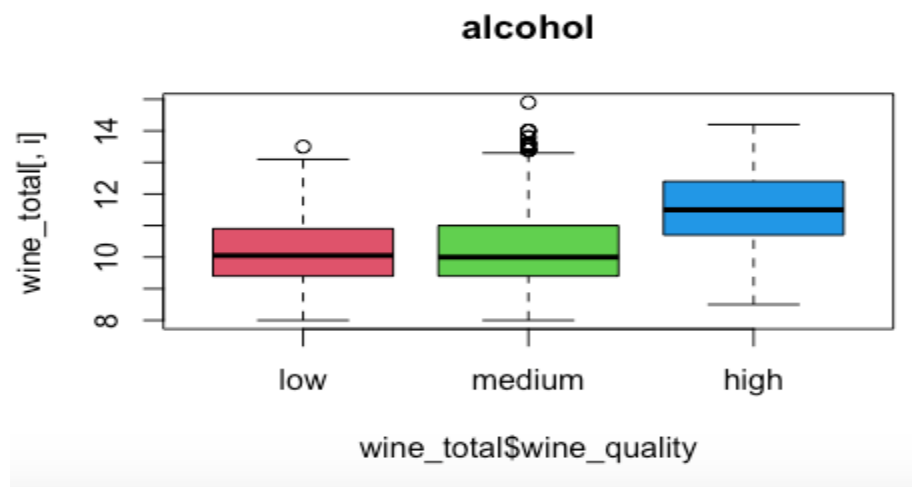
*Red*

*White 0.0079*

rejectH0: so, the difference between percentage of alcohol and type of the wine is coming from the difference between white and red wine.

2. Alcohol ~ wine\_quality

	low	medium
medium	0.78	-
high	<2e-16	<2e-16



The difference between percentage of alcohol and wine quality is coming from the difference between high quality and low quality, high quality and medium quality because their p-values are less than 0.05.

4. [Detail the results of Two-Way ANOVA considering as dependent variable “fixed acidity”, and independent variable “type” and “quality”.](#)

Based on the result of ANOVA test: the type of the wine has a significant effect on the fixed.acidity (reject H0) and the quality categorized of the wine does not have a significant effect on the fixed acidity(accept H0).

1. Independency: based on Durbin-watson result, p-values is less than 0.05. So this assumption is not valid.
2. Normality: based on Q-Q plot alcohol does not have a normal distribution
3. Homogeneity of variance: based on Leventest, variance for type is not equal but for quality is equal.

### Post Hoc testing

1. Based on pairwise t-test with bonferroni correction:

Red

White <2e-16

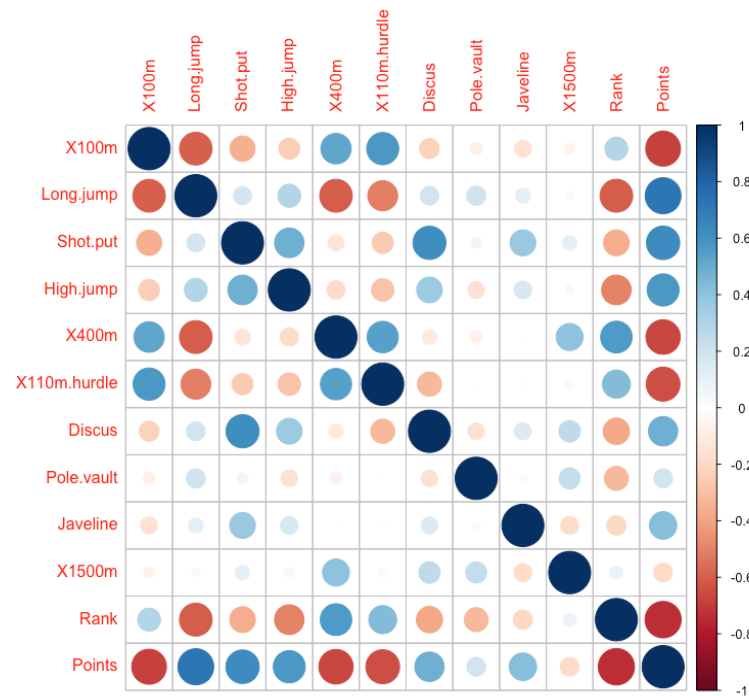
rejectH0: so, the difference between fixed acidity and type of the wine is coming from the difference between white and red wine.

### 3. Third Question

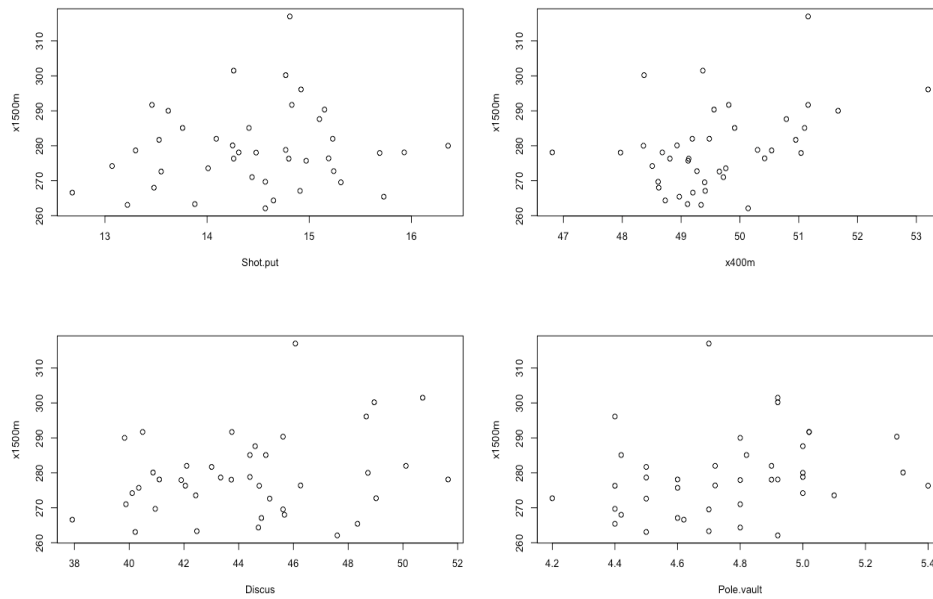
What is the linear expression that better predicts the behavior of an athlete for 1500m?

Creating a correlation matrix will help us to decide which values may be related with the 1500m value.

In this plot, positive correlations are displayed in a blue scale while negative correlations are displayed in a red scale.



As we can see from the above table, X400m, Discus, Pole.vault and shot.put have positive correlation with X1500m. In the plots below we can find how changes in these four variables can affect X1500m.



Variable	P-value (intercept)	P-value (slope)	MSE (Residual standard error)	Multiple R-squared	Adjusted R-squared
"x400m"	0.31909	0.00808	10.79	0.1666	0.1452
<p>Assumptions: # Model 1</p> <ol style="list-style-type: none"> <li>1. Normality test rejected (<math>p\text{-value} = 0.01742 &lt; 0.05</math>), <b>the error term does not follow a Normal distribution.</b></li> <li>2. Homogeneity of variance accepted (<math>p\text{-value} = 0.9739 &gt; 0.05</math>), the homogeneity of variances is provided.</li> <li>3. The independence of errors (<math>p\text{-value} = 0.3458 &gt; 0.05</math>), the errors/observations are independent.</li> </ol> <p>This model only explains <b>16.66%</b> of the variance</p>					
Variable	P-value (intercept)	P-value (slope)	MSE (Residual standard error)	Multiple R-squared	Adjusted R-squared
"Discus"	2.06e-12	0.103	11.42	0.06665	0.04272
Discus+x400m	0.92	x400m(0.003) Discus(0.033)	10.29	0.26	0.22

Assumptions: # Model 2

1. Normality test accepted (p-value = 0.1327 > 0.05), the error term follow a Normal distribution.
2. Homogeneity of variance accepted (p-value = 0.1491 > 0.05), the homogeneity of variances is provided.
3. The independence of errors (p-value = 0.3434 > 0.05), the errors/observations are independent.

This model only explains **26%** of the variance.

Variables	P-value (intercept)	P-value (slope)	MSE (Residual standard error)	Multiple R-squared	Adjusted R-squared
"Pole.vault"	6.36e-09	0.119	11.45	0.06123	0.03716
Discus+x400m+Pole.vault	0.29	x400m(0.0008) Discus(0.009) pole.vault(0.01)	9.605	0.37	0.32

Assumptions: # Model 3

1. Normality test accepted (p-value = 0.07727 > 0.05), the error term follows a Normal distribution.
2. Homogeneity of variance accepted (p-value = 0.2152 > 0.05), the homogeneity of variances is provided.
3. The independence of errors (p-value = 0.7275 > 0.05), the errors/observations are independent.

This model only explains **37%** of the variance

"Shot.put"	1.66e-09	0.471	11.74	0.01341	-0.01189
Discus+x400m+ Pole.vault + Shot.put	0.37	x400m(0.001) Discus(0.016) pole.vault(0.012) Shot.put(0.53)	9.684	0.38	0.31

# Model 4: In the case of this model, "Shot.put" is not significant. Other variables that we use in the previous model are significant.

Based on this table, the third model can predict most variations (37%) in X1500m and has least residual error(9.605) in comparison with other models. Furthermore, this model does not violate any of the assumptions required for the validity of the model.

Now use the expression to **predict** the behavior for a specific athlete. **Analyze and explain the results obtained.**

**Coefficients of best model have provided in below:**

(Intercept)	X400m	Discus	Pole.vault
-85.126139	4.839327	1.263457	14.286330

For example for Casarsa, linear regression model gives us 296.6658 as x1500m with a confidence interval of [285.1383, 308.1932]. Real value for this athlete is 296.12. For this case the model behaves well, but there are other cases where the difference between predicted and real value is large. For example, Clay's real value is 301.50 and predicted value is 288.1628 with a confidence interval of [281.0726, 295.2529] that has a higher difference compared to

**Is the model accurate? What do you expect?**

One way to test accuracy is comparing RMSE between training data and test data. If there is no high difference between these two values, we can say that our model is accurate.

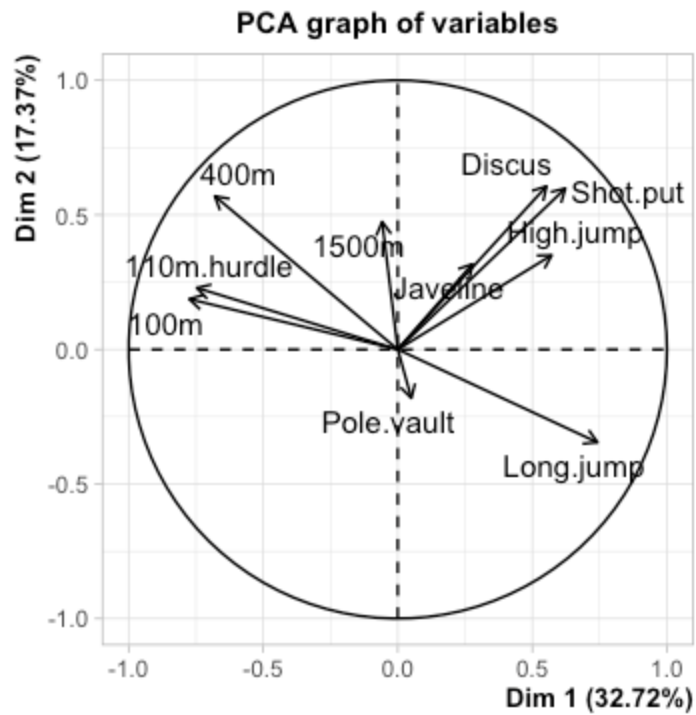
RMSE\_test = 7.464218

RMSE\_train = 9.984343

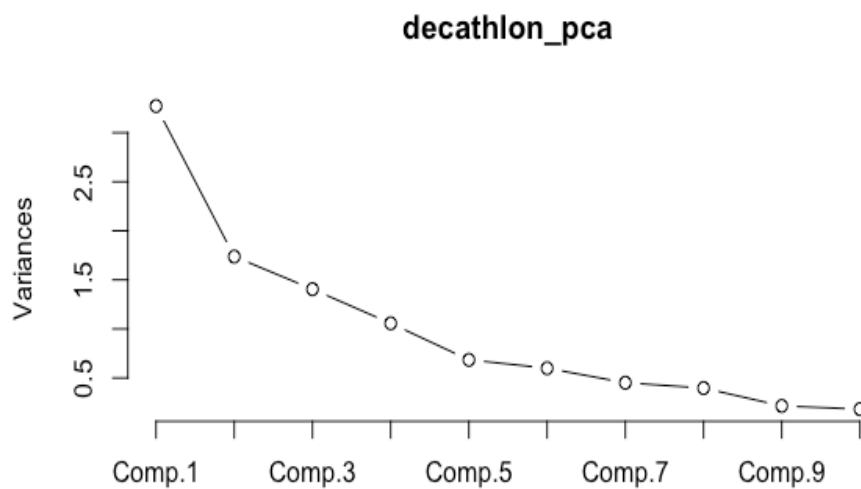
Difference between these two values is 1.16 that is not high, and we can say that our model is accurate.

#### **4. Fourth Question**

Define a PCA for the decathlon dataset and discuss why the model you own works well (or not) looking at the variables chart. We recommend using the FactoMineR package to do this analysis.



From the scree plot we can detect that two dimensions are enough to explain the variance of the data, which corresponds to the 50.09% cumulative percent of variability.



We want to construct a linear regression model to predict the points of each athlete. To do so you must first decide the number of principal components to be included in the regression as independent variables. Justify your answer.



Check the assumptions of the regression model.

Is the prediction accurate enough?

	Model_1	Model_2	Model_3
Number of PC	PC1+ PC2	PC1+ PC2 + PC3	PC1+ PC2 + PC3 +PC4
Adjusted R-squared	0.9145	0.9123	0.9746
Cumulative percentage of variance	50.09%	64.13%	74.71%
RMSE_test	96.0975	98.66786	48.8125
RMSE_train	103.9868	99.31829	57.84941
Assumptions/Normality-Shapiro test	p-value=0.365 >0.05	p-value=0.848 >0.05	p-value=0.005536 < 0.05
Homogeneity of Variance/ residual plot	Homogeneity of variances is provided	Homogeneity of variances is provided	Homogeneity of variances is provided
Independence of errors/Durbin watson test	p-value = 0.719 > 0.05	P-value = 0.5625 > 0.05	p-value = 0.722 >0.05

Based on this plot, the lowest RMSE belongs to Model-3 that has a 48.81 value, and it has highest R-adjusted that means this model can predict 97% of variation.