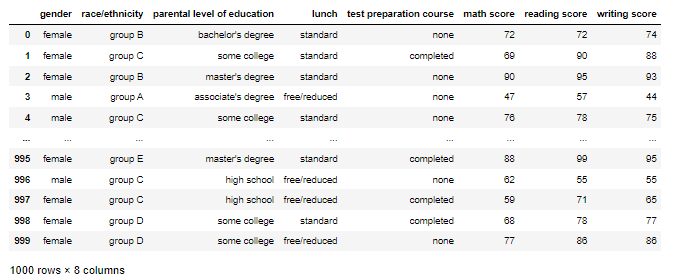
Cuculescu Andrei

Gr. 511

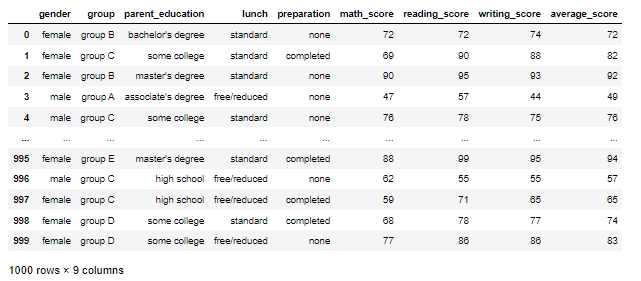
Project 2 SDS

1. For the first task I chose a popular dataset from Kaggle called “Students performances in exams” which purpose is to better understand the influence of different factors on students’ performance.

<https://www.kaggle.com/datasets/spscientist/students-performance-in-exams>

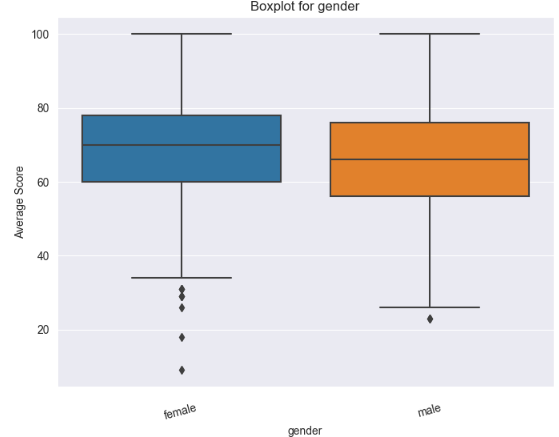


This is how the dataset looks like, the first thing we will do is to rename some columns and add a new column named ‘Average\_score’ which is the mean of the 3 exams grades.

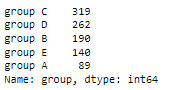


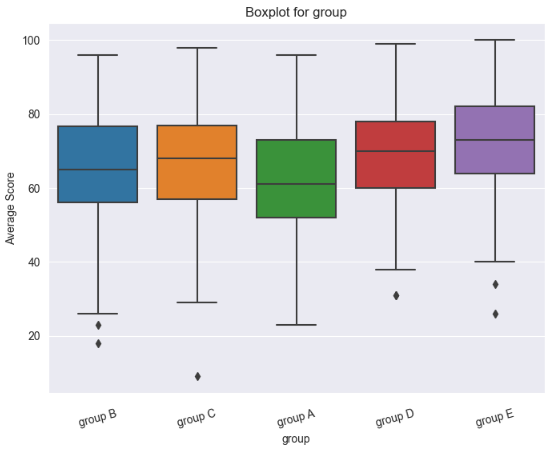
This is how our dataset looks right now, further I will do some plots to better understand it.



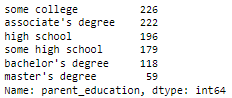


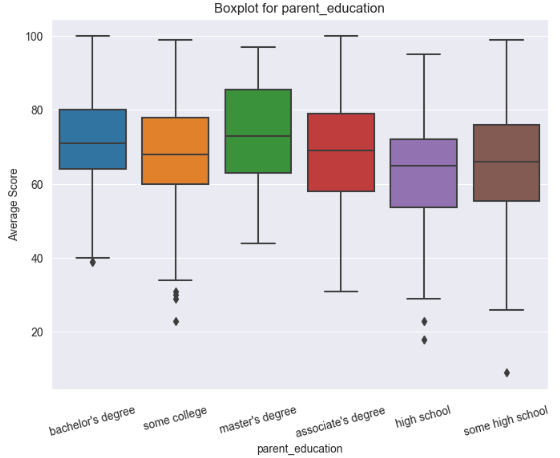
As we can see the distribution of male and female is close 52% with 48% but we can observe that female a slightly better performance in exams.





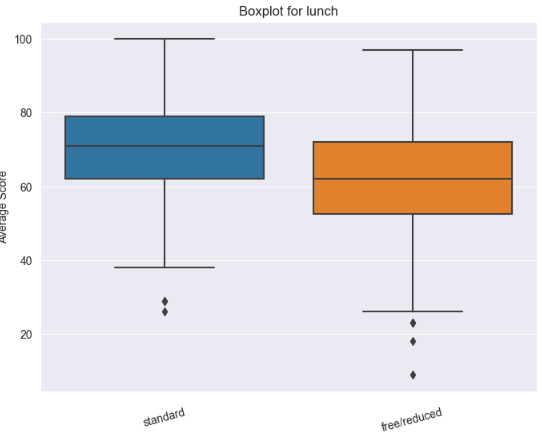
For the race groups we can see that group C represent 31% of our dataset and Group be only 8.9% but seems like students from the group E have the best performances.





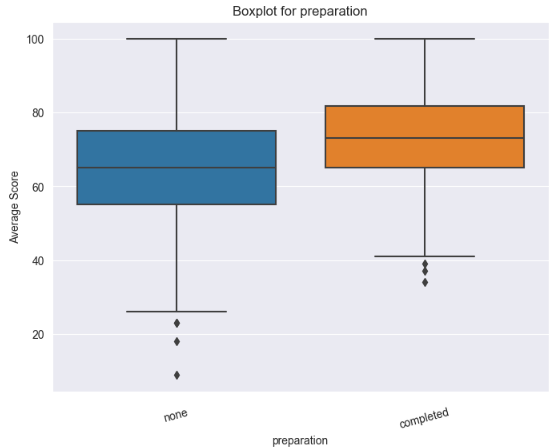
We can see that the education of the parent has a influence on students’ performance but also our sample for parents with master’s degree is small only 6% from our dataset.





Another slightly better performance for the students who have standard lunch, also represent 65% of the dataset.

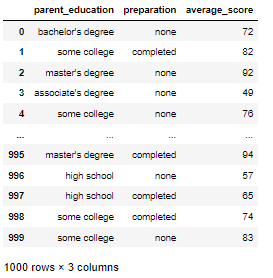


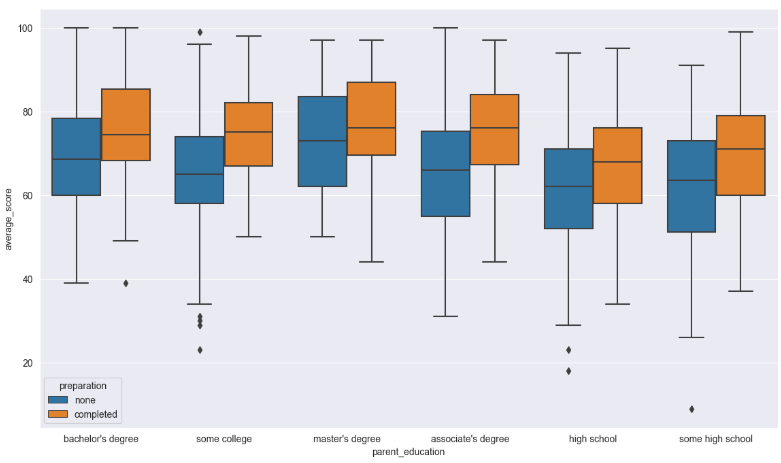


We can observe that the preparation course for exam had a benefic impact on the students how completed it, have a better performance than those who didn’t completed it.

Two-way ANOVAs (Analysis of Variance, also known as factor analyses) are statistical tests that examine the effects of two independent categorical variables (sometimes referred to as factors) on a continuous dependent variable. It is a development of the one-way ANOVA, which contrasts the means of one factor across many groups. [1]

For our task I only selected the data on parental education level and preparation course for the two-way anova task, and I looked at how these factors affected the average score.





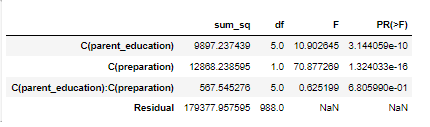
With Anova we will compare the means of more than two groups, our dependent variable is the average score of the student performance from the 3 exams (math, reading, writing) and the independent variables are the preparation course and the parent education. [3]

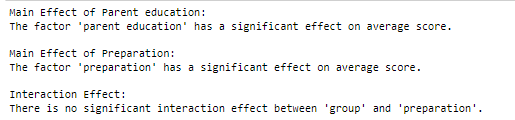
For Anova we will have two hypotheses:

* Null Hypothesis (H0): Groups means are equal which can be not be rejected if p-value > significance level (0.05)
* Alternate Hypothesis (H1): At least, one group has the mean different from the other groups and the null hypothesis is rejected p-value < significance level(0.05)

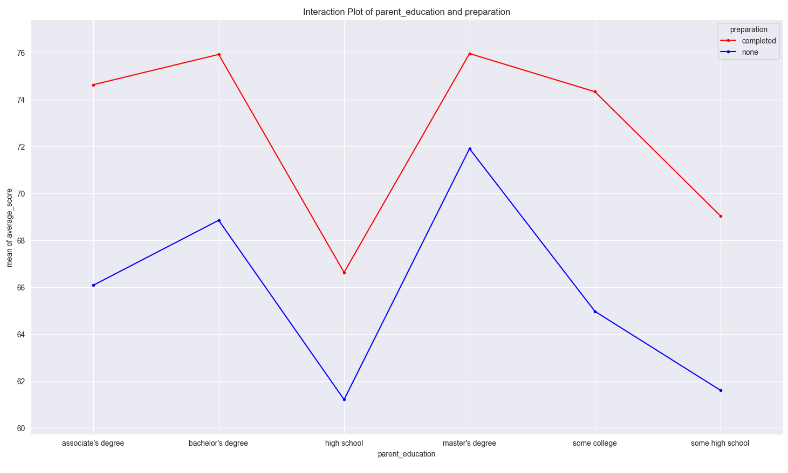
We will have 3 Hypothesis based on the two-way Anova analysis:

* The effect of the preparation course on the performance of the student in exams.
* The effect of the parent education on the performance of the student in exams.
* The effect of the preparation course and parent education on the performance of the student in exams.



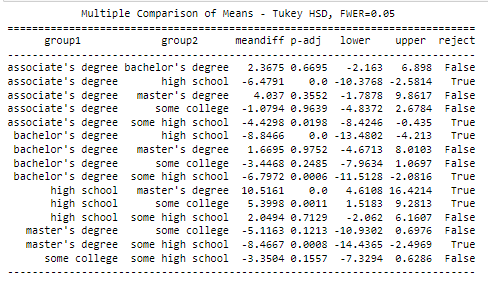


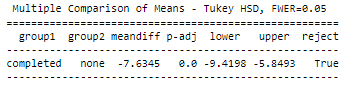
After we check the results, we can conclude that independently parent education and preparation course have a influence of the average score because the p value is lower that 0.05 so we reject the null hypothesis but together the two factors don’t have a significant effect on the average score and the p value is not lower than 0.05 so we fail to reject the null hypothesis.



From the interaction plot we can observe that the lines are quite parallel and from this and Anova results we can say that interaction is not very significant.

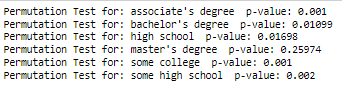
We will use the Tukey HSD test to do a multiple pairwise comparison (Post-hoc comparison) study to determine the pairings of significant difference groups because ANOVA does not reveal which groups differ substantially from one another. The test compares all possible pairs of means. [2]



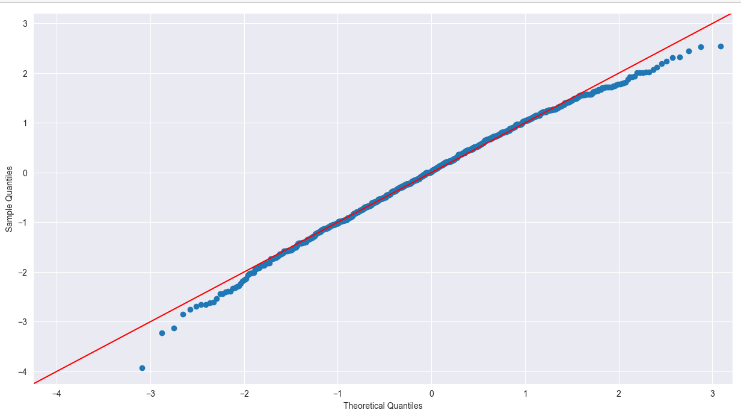


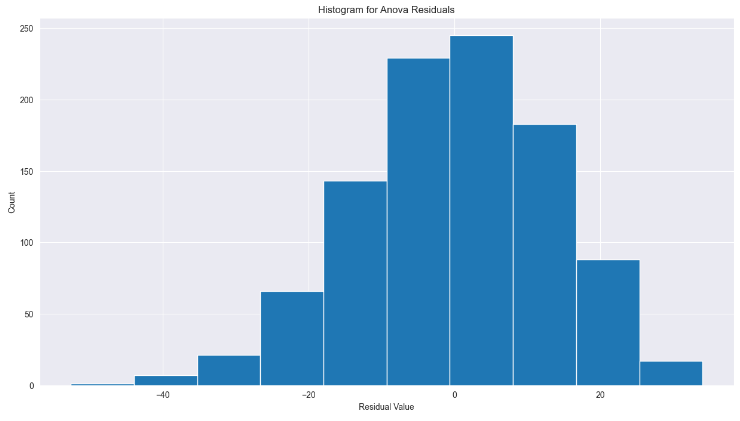
We can observe that for preparation course the statistical significance so we can reject the null hypothesis but for parent education there are some combinations that are not all rejecting null hypothesis.

The permutation test implies that two distinct groups came from the same distribution under the null hypothesis.



These findings imply that depending on the education level of the parents, the effect of taking the preparatory course on average scores may differ. Students who have completed the preparatory course tend to profit greatly, whether they have an associate's degree, bachelor's degree, high school education, some college education, or some high school education. The data, however, do not strongly support the notion that completion of the preparatory course makes a substantial difference in average scores for students with a master's degree.





Shapiro-Wilk Test for Normality

Null hypothesis: data is drawn from normal distribution.



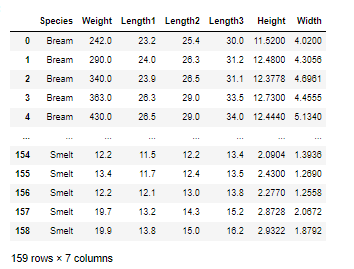
W, is 0.9928. Since the number is almost 1, this suggests that the data is quite close to a normal distribution.

The p-value of 0.8773 that was found is rather high. This implies that there is insufficient data to reject the null hypothesis of normality. As a consequence, the data is assumed to be normally distributed or to follow a distribution that is sufficiently near to a normal distribution based on the findings of the Shapiro-Wilk test.

2. For the Linear regression model task I took a fish market dataset.

<https://www.kaggle.com/datasets/aungpyaeap/fish-market?datasetId=229906&sortBy=voteCount>

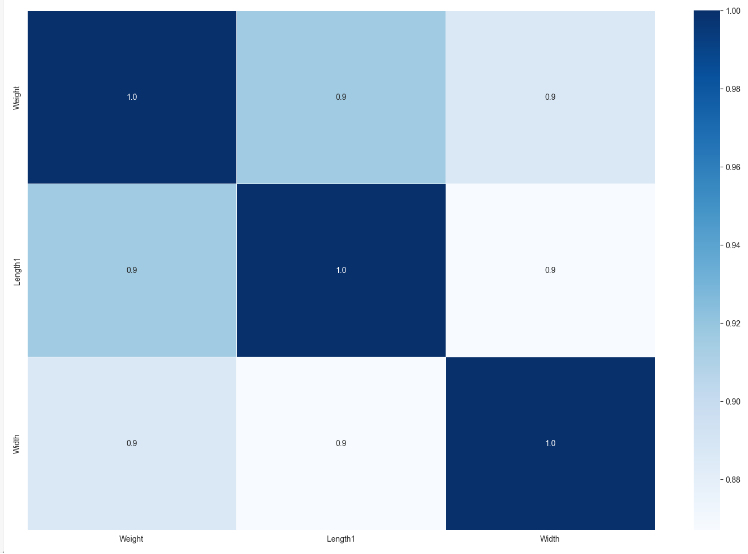
The dataset looks like this:



But the dataset I will work with is like this:

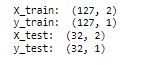


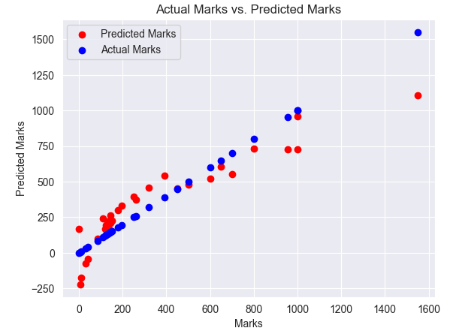
Using the length and width we will try to predict weight.



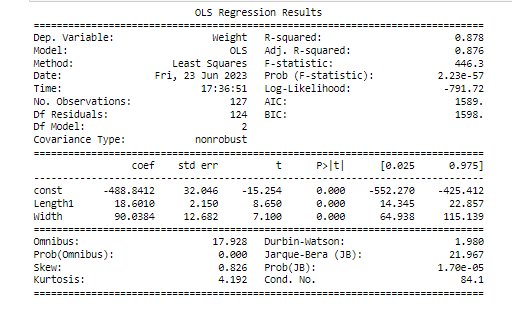
With this heat map we see the correlation between them.

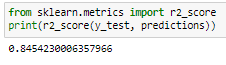
We split the dataset in 80% training and 20% test with the shape:



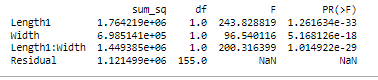


Here he have the prediction on the test samples.





R2 around 0,84 that is a good value.



We can deduct from the ANOVA table that all three of the H0 hypotheses are rejected and that the independent variables have an impact on weight both singly and collectively.



The extremely low (near 0) p-value indicates strong evidence against the null hypothesis of normality. The assumption of normality for the residuals is thus broken according to the Shapiro-Wilk test. In other words, the residuals' distribution does not fit neatly into the normal distribution.

Refernces

1. <https://www.investopedia.com/terms/a/anova.asp>

2.<https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/post-hoc/tukey-test-honest-significant-difference/>

3. <https://www.reneshbedre.com/blog/anova.html>