MARC VIOLIDES

Report Brainnest final project

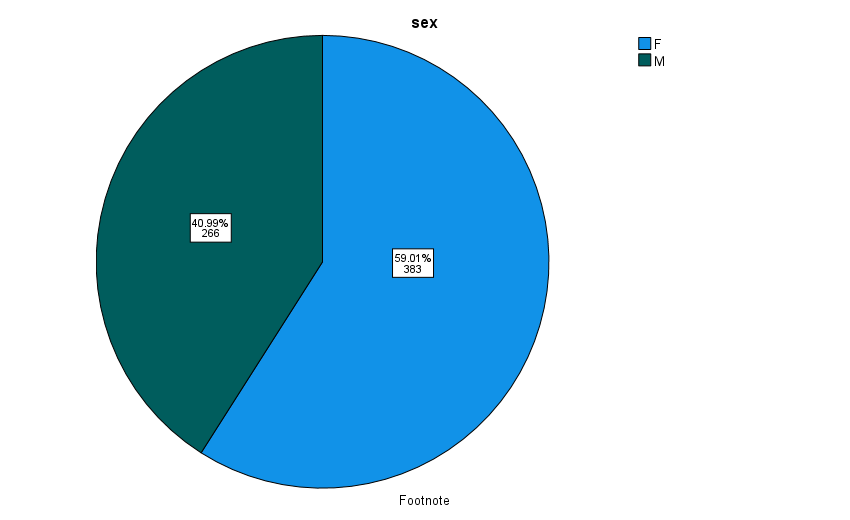
The chosen dataset taken from Kaggle (<https://www.kaggle.com/datasets/larsen0966/student-performance-data-sets>) shows how students perform on their exam and contains the following variables: Sex, Age, Parental status (whether they are married or divorced), study time, number of failures, and finally the grade they got on the exam. Note that there are originally 33 variables in the dataset but I decided to cut the number of variables and limit them to the most interesting ones that can influence grades in my opinion.

**UNIVARIATE ANALYSIS**

First step is to explore the data in a univariate way (each variable alone):

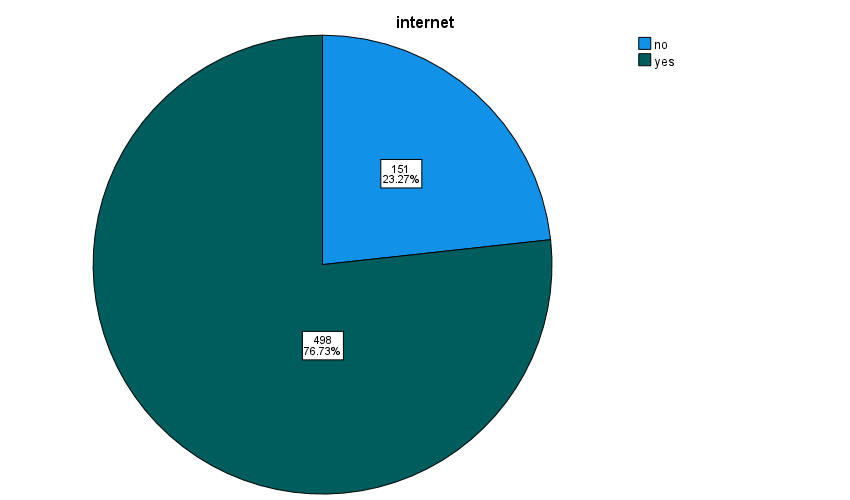
1. Categorical Variables

*Sex*



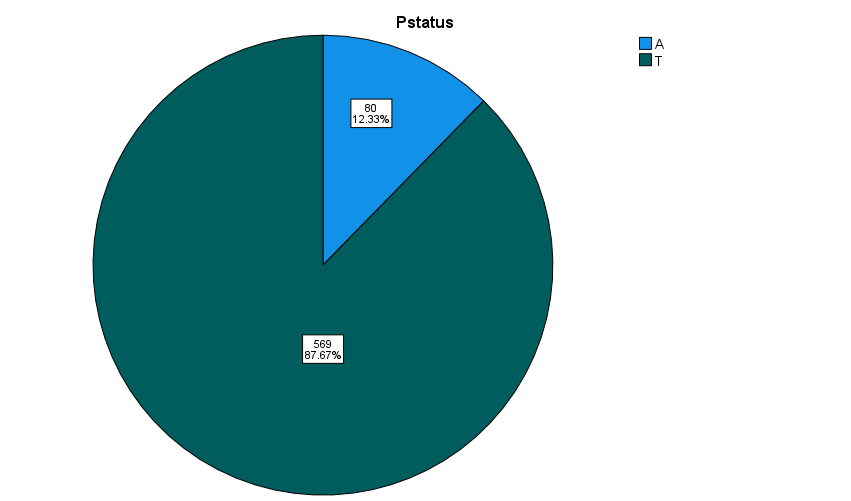
We notice that there are more Females than males in our dataset with 59% of individuals being women. This has no particular importance: in my opinion performance does not depend on the sex.

*Internet Use*



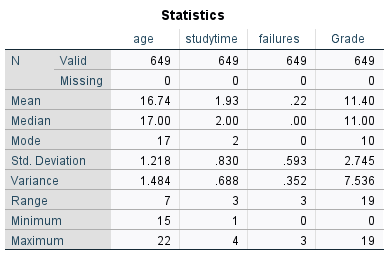
Moving forward, we get that there are more people using the internet (76.73%) than people who are not (23.27%), which is logical in our day and age. We shall see whether this impacts their grades or not.

*Parental Status*

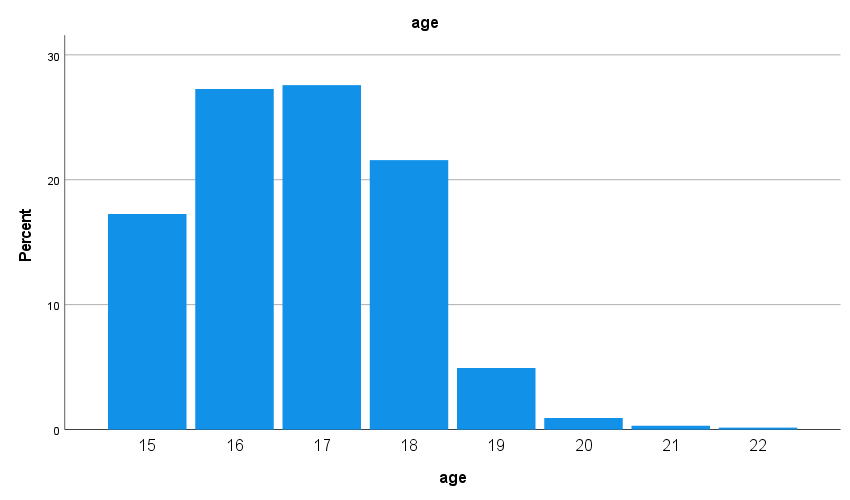


As for parental status, we observe that there are only a few divorced parents (12.33%) in our datasets which represents reality in a pretty accurate way.

1. Numerical variables

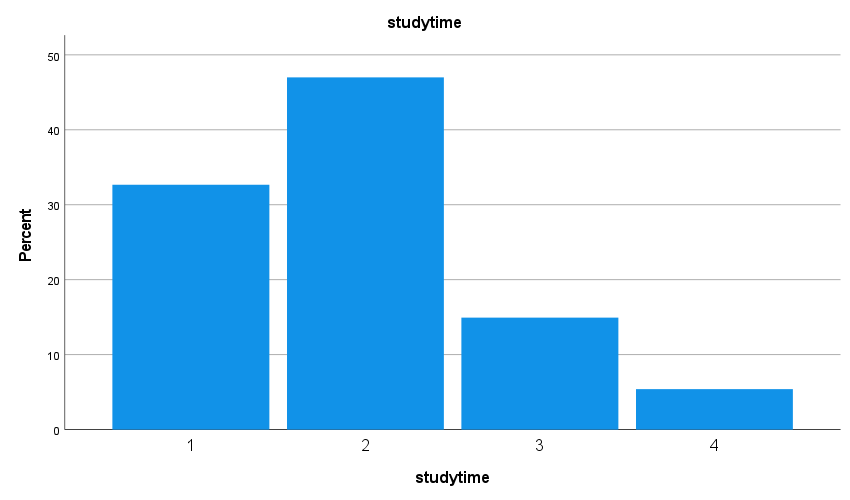


*Age*



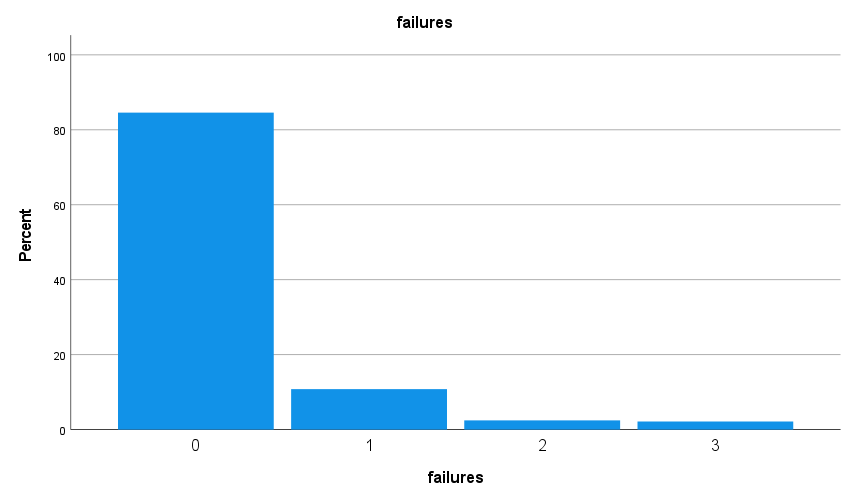
We see that age ranges from 15 to 22, with most of them being younger than 18, with a mean of 16.74 and a mode of 17.

*Study time*



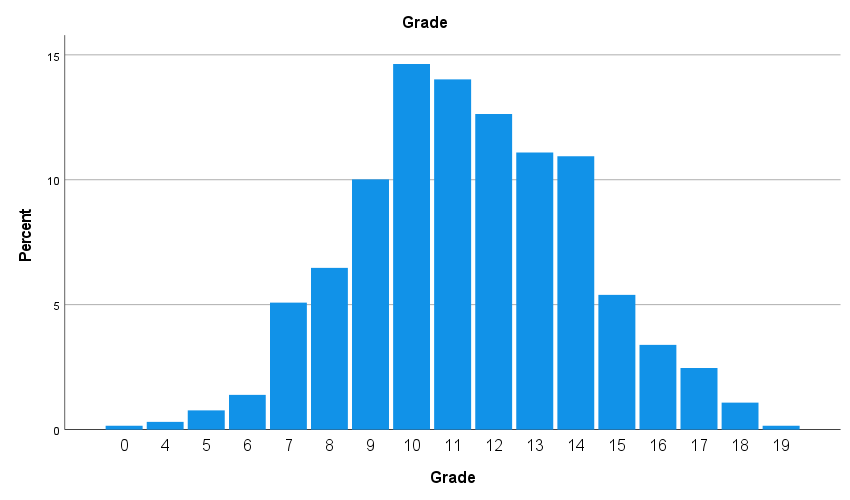
For the study time, most individuals spend on average 2 hours studying, with the values ranging from 1 to 4 hours.

*Number of failures*



Most students are good ones as the vast majority has 0 failures and the worst ones have 3.

*Grades*

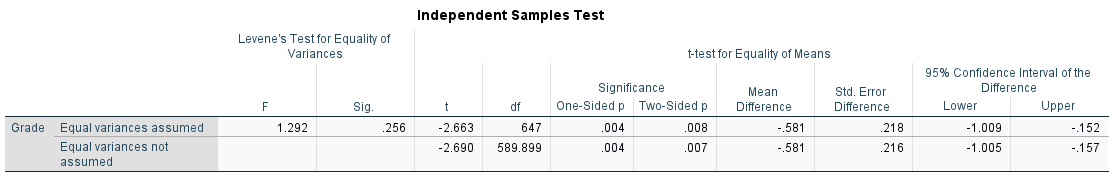


As for grades, data seems approximately normally distributed and we have that the mean and median are at around 11 and the mode at 10.

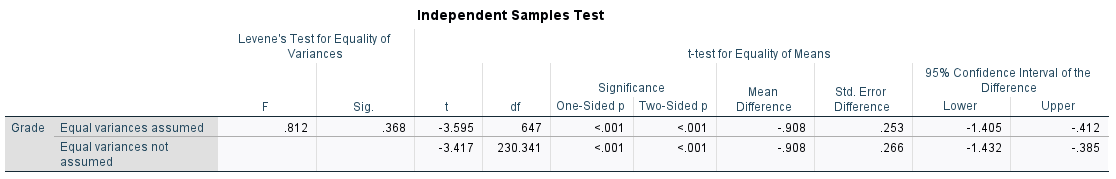
**BIVARIATE ANALYSIS (t-test, ANOVA, Pearson correlation, Cohen’s d)**

There are no significant outliers in our data, looking at each boxplot, so we can proceed to next step of analysis

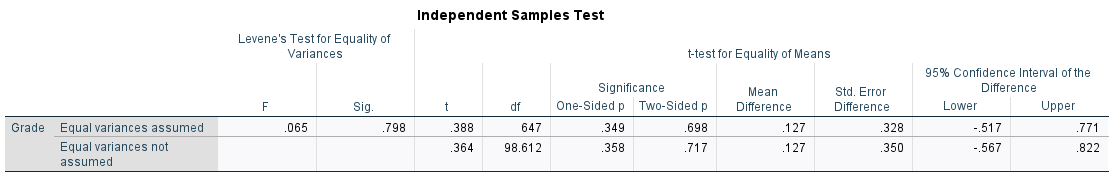
Having independence of observations, no significant outliers and normality, we can start with the t-tests to check what factors affect grades:



Starting with sex, as presumed, Levene’s test tell us with significance>0.05 so at 95% confidence level, there are no differences between groups (variances are equal across males and females with regards to grades). t value is smaller than -1.96 so there is a significant difference between mean of grades when comparing males to females. The mean grade for women is 11.64 and the mean grade for men is 11.06. However, Cohen’s d gives us a coefficient of -0.213 which means that this difference between means can be considered as negligeable.



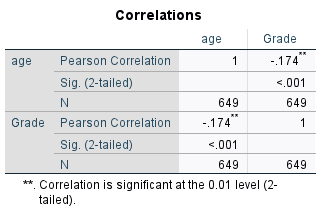
Again, for internet use, we notice that equal variances are assumed and that the means are significantly different (p-value < 0.01 < 0.05) so there is a difference in means between students who use the internet compared to people who do not. Cohen’s d in this case is -0.334 which is negligeable-to-average significance.



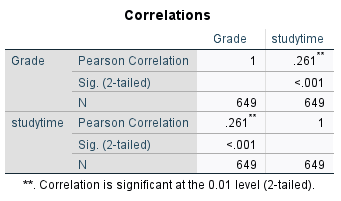
Weirdly enough, we notice that parental status has no significant effect on grades (p-value = .698 which is >0.05).

Here, we don’t need to use ANOVA test because we don’t have more than 2 groups in a category.

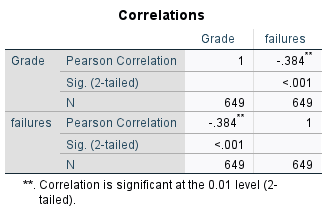
Now to check whether numerical data is associated with our dependent variable (here grades), we will see whether this is the case using Pearson’s correlation:



Starting with age, we find that there is a small significant negative correlation (because -0.174>-0.29) between the aforementioned variable and grade at 99% confidence level. This suggests that there is a negative linear association between these 2 variables.



As for study time, it has small-to-moderate correlation with grades at 99% confidence level. So there is slight linear correlation between the 2.



Lastly, we observe that failures seem to be the most correlated variable out of all the previous independent variables we just saw, with a correlation of -0.384 which means that there is a moderate linear association between these 2 variables.

To conclude, we can say that amongst the categorical variables, internet use seems to have the most prominent effect on grades, which is logical: the more you spend time on internet, the worse your grades are. Sex has a significant but extremely small effect on grades, as we would expect; but what is most surprising is that parental status has no effect on grades (you would expect kids to be affected by what is happening around them, but maybe it is because they are old enough to deal with it while focusing on themselves). As for numerical variables, failures have the highest correlation with grades which is logical (since the more you fail the worse your grades get); meanwhile study time and age both have significant correlation with grades, but as you would expect study time’s correlation is larger. However, this doesn’t mean that age doesn’t have an effect on grades because correlation does not mean causation, and Pearson’s correlation only measures linear relationship.