Students Perfomance in written tests

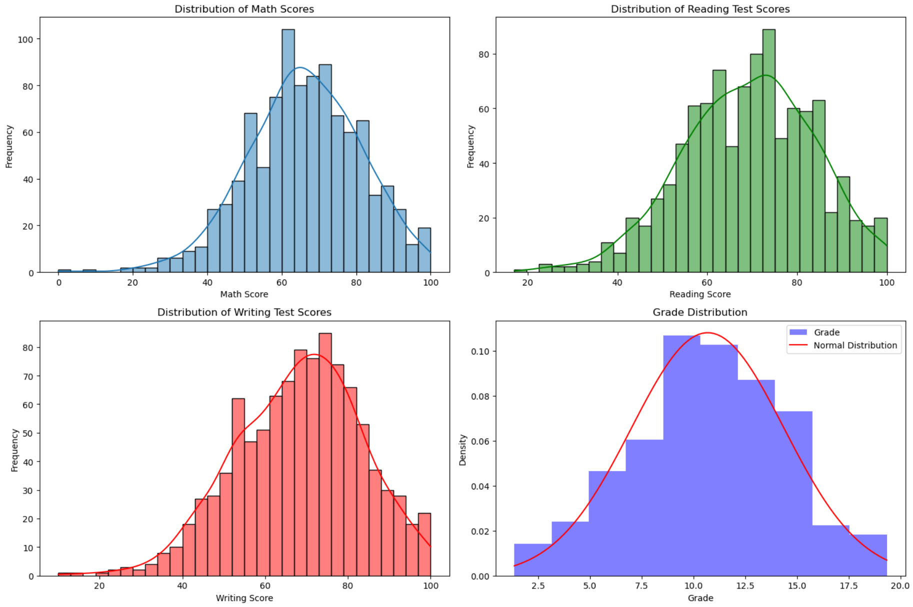
Mykyta Panchenko, Alen Sahinpasic

February 2024

1 Introduction

This part of the project advances our exploration of the Student's performance in tests dataset, which was introduced in our earlier work. Moreover, we will introduce a new dataset in order to make the regression more credible. Here, we focus on the correlation between a different variables that describe student's status and their performance. By analyzing this relationship, our objective is to ultimately develop a predictive model capable of estimating a student's performance based on some of the variables we will introduce later into the report.

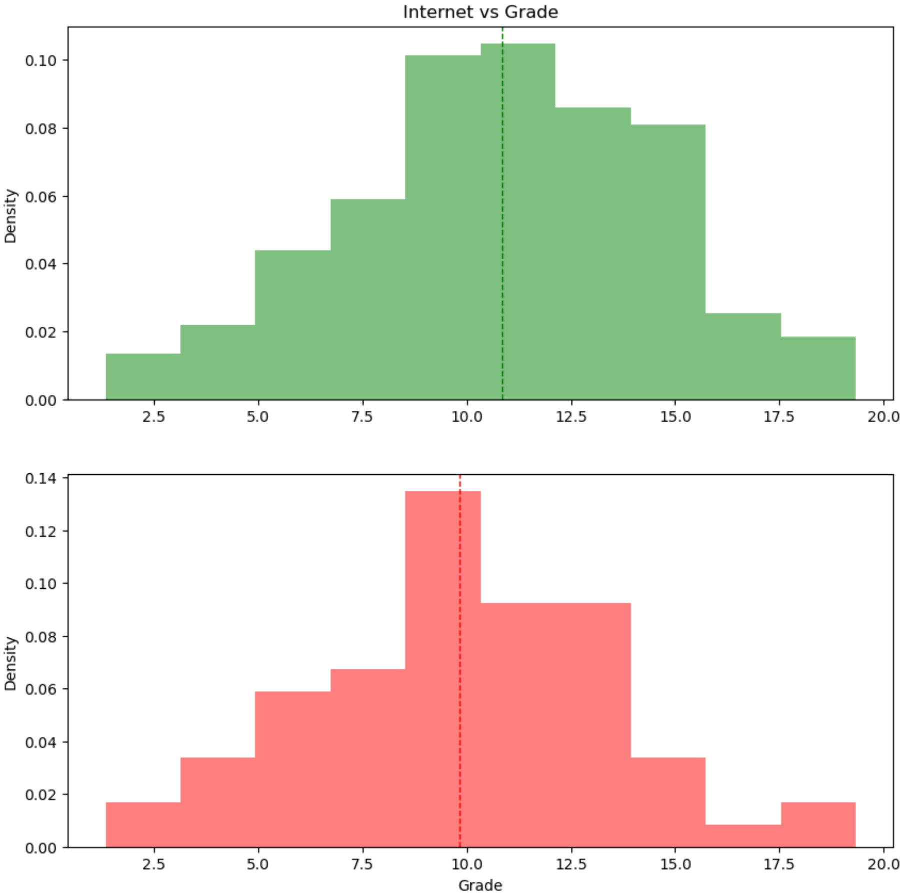
2 Distribution and Fitting

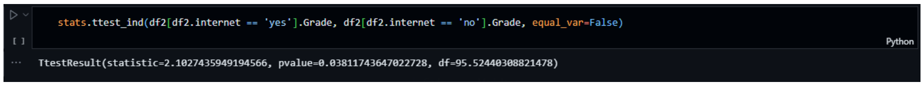


In the last report, we assumed that the distribution of scores for each test is approximately normal. In this part, we fitted normal distributions and performed Shapiro-Wilk tests. For the first 3 distributions which are a part of the first dataset, we got a 0.99 statistic with p≈0. On the other hand, with a second dataset, which is not artificial, we got p=0.0505, which means we can't reject the null hypothesis at 5 percent significance level.

3 Hypothesis testing

Hypothesis one: Performance of people who have access to the internet is the same as the performance of those who don't on 95 percent confidence level.



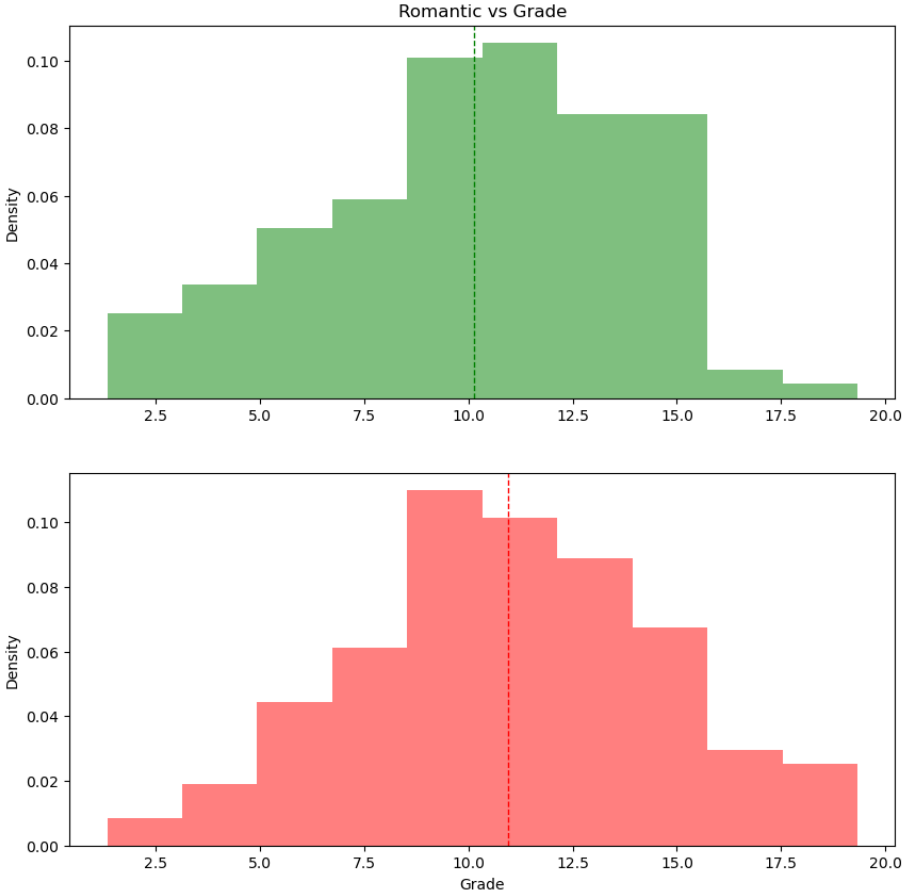


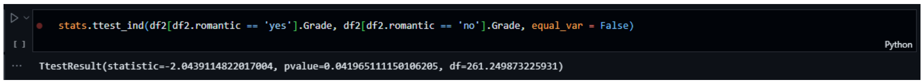
```
stats.ttest_ind(df2[df2.internet == 'yes'].Grade, df2[df2.internet == 'no'].Grade, equal_var=False)
```

```
TtestResult(statistic=2.1027435949194566, pvalue=0.03811743647022728, df=95.52440308821478)
```

The P-value is 0.04 with equal variance and 0.038 with Welch's t-test, therefore we can reject the null hypothesis on a 5 percent confidence level and say that the performance is not the same.

Hypothesis two: Performance of people who are in a romantic relationship is the same as the performance of those who aren't on 95 percent confidence level.





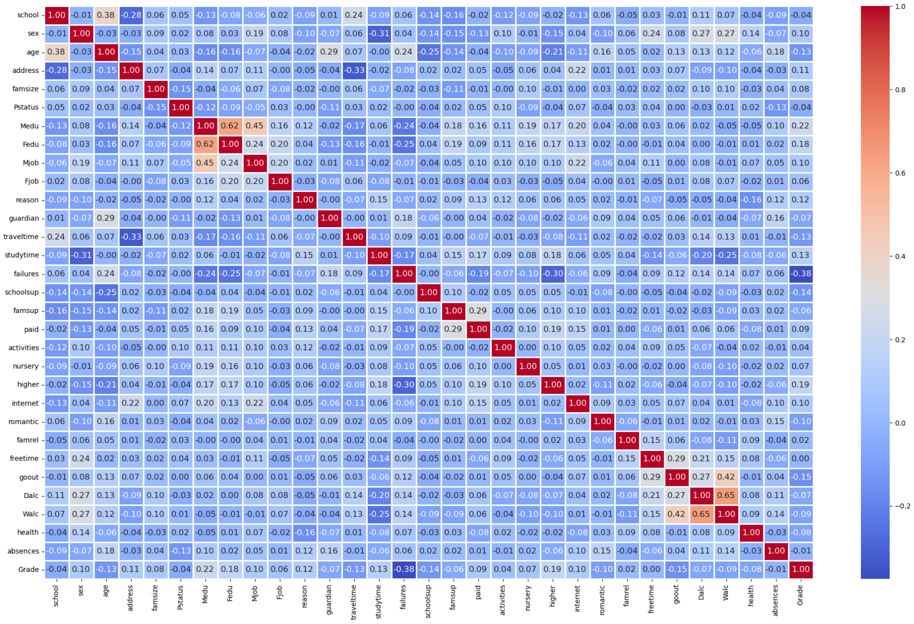
```
stats.ttest_ind(df2[df2.romantic == 'yes'].Grade, df2[df2.romantic == 'no'].Grade, equal_var = False)
```

```
TtestResult(statistic=-2.0439114822017004, pvalue=0.041965111150106205, df=261.249873225931)
```

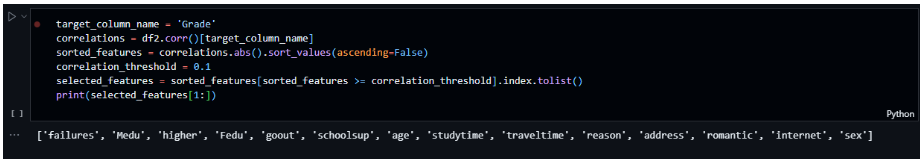
The P-value is 0.042 both with the Student's and Welch's t-test, therefore we can reject the null hypothesis on a 5 percent confidence level and say that the performance is not the same.

4 Linear Regression

First of all, we plotted a correlation matrix to see which variables are mostly correlated to the Grade. We selected features that had the correlation coefficient of >0.1.



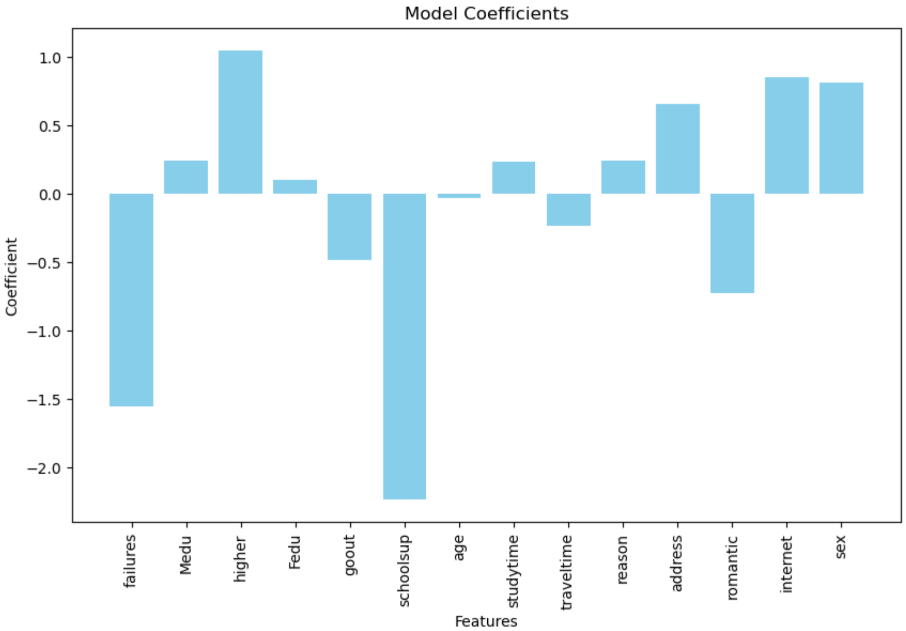
	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	Grade
school	1.00	-0.01	0.38	-0.28	0.06	0.05	-0.13	-0.08	-0.06	0.02	-0.09	0.01	0.24	-0.09	0.06	-0.14	-0.16	-0.02	-0.12	-0.09	-0.02	-0.13	0.06	-0.05	0.03	-0.01	0.11	0.07	-0.04	-0.09	-0.04
sex	-0.01	1.00	-0.03	-0.03	0.09	0.02	0.08	0.03	0.19	0.08	-0.10	-0.07	0.06	-0.31	0.04	-0.14	-0.15	-0.13	0.10	-0.01	-0.15	0.04	-0.10	0.06	0.24	0.08	0.27	0.27	0.14	-0.07	0.10
age	0.38	-0.03	1.00	-0.15	0.04	0.03	-0.16	-0.16	-0.07	-0.04	-0.02	0.29	0.07	-0.00	0.24	-0.25	-0.14	-0.04	-0.10	-0.09	-0.21	-0.11	0.16	0.05	0.02	0.13	0.13	0.12	-0.06	0.18	-0.13
address	-0.28	-0.03	-0.15	1.00	0.07	-0.04	0.14	0.07	0.11	-0.00	-0.05	-0.04	-0.33	-0.02	-0.08	0.02	0.02	0.05	-0.05	0.06	0.04	0.22	0.01	0.01	0.03	0.07	-0.09	-0.10	-0.04	-0.03	0.11
famsize	0.06	0.09	0.04	0.07	1.00	-0.15	-0.04	-0.06	0.07	-0.08	-0.02	-0.00	0.06	-0.07	-0.02	-0.03	-0.11	-0.01	-0.00	0.10	-0.01	0.00	0.03	-0.02	0.02	0.02	0.10	0.10	-0.03	0.04	0.08
Pstatus	0.05	0.02	0.03	-0.04	-0.15	1.00	-0.12	-0.09	-0.05	0.03	-0.00	-0.11	0.03	0.02	-0.00	-0.04	0.02	0.05	0.10	-0.09	-0.04	0.07	-0.04	0.03	0.04	0.00	-0.03	0.01	0.02	-0.13	-0.04
Medu	-0.13	0.08	-0.16	0.14	-0.04	-0.12	1.00	0.62	0.45	0.16	0.12	-0.02	-0.17	0.06	-0.24	-0.04	0.18	0.16	0.11	0.19	0.17	0.20	0.04	-0.00	0.03	0.06	0.02	-0.05	-0.05	0.10	0.22
Fedu	-0.08	0.03	-0.16	0.07	-0.06	-0.09	0.62	1.00	0.24	0.20	0.04	-0.13	-0.16	-0.01	-0.25	0.04	0.19	0.09	0.11	0.16	0.17	0.13	0.02	-0.00	-0.01	0.04	0.00	-0.01	0.01	0.02	0.18
Mjob	-0.06	0.19	-0.07	0.11	0.07	-0.05	0.45	0.24	1.00	0.20	0.02	0.01	-0.11	-0.02	-0.07	-0.04	0.05	0.10	0.10	0.10	0.10	0.22	-0.06	0.04	0.11	0.00	0.08	-0.01	0.07	0.05	0.10
Fjob	0.02	0.08	-0.04	-0.00	-0.08	0.03	0.16	0.20	0.20	1.00	-0.03	-0.08	0.06	-0.08	-0.01	-0.01	-0.03	-0.04	0.03	-0.03	-0.05	0.04	-0.00	0.01	-0.05	0.01	0.08	0.07	-0.02	0.01	0.06
reason	-0.09	-0.10	-0.02	-0.05	-0.02	-0.00	0.12	0.04	0.02	-0.03	1.00	-0.00	-0.07	0.15	-0.07	0.02	0.09	0.13	0.12	0.06	0.06	0.05	0.02	-0.01	-0.07	-0.05	-0.05	-0.04	-0.16	0.12	0.12
guardian	0.01	-0.07	0.29	-0.04	-0.00	-0.11	-0.02	-0.13	0.01	-0.08	-0.00	1.00	-0.00	0.01	0.18	-0.06	-0.00	0.04	-0.02	-0.08	-0.02	-0.06	0.09	0.04	0.05	0.06	-0.01	-0.04	-0.07	0.16	-0.07
traveltime	0.24	0.06	0.07	-0.33	0.06	0.03	-0.17	-0.16	-0.11	0.06	-0.07	-0.00	1.00	-0.10	0.09	-0.01	-0.00	-0.07	-0.01	-0.03	-0.08	-0.11	0.02	-0.02	-0.02	0.03	0.14	0.13	0.01	-0.01	-0.13
studytime	-0.09	-0.31	-0.00	-0.02	-0.07	0.02	0.06	-0.01	-0.02	-0.08	0.15	0.01	-0.10	1.00	-0.17	0.04	0.15	0.17	0.09	0.08	0.18	0.06	0.05	0.04	-0.14	-0.06	-0.20	-0.25	-0.08	-0.06	0.13
failures	0.06	0.04	0.24	-0.08	-0.02	-0.00	-0.24	-0.25	-0.07	-0.01	-0.07	0.18	0.09	-0.17	1.00	-0.00	-0.06	-0.19	-0.07	-0.10	-0.30	-0.06	0.09	-0.04	0.09	0.12	0.14	0.14	0.07	0.06	-0.38
schoolsup	-0.14	-0.14	-0.25	0.02	-0.03	-0.04	-0.04	0.04	-0.04	-0.01	0.02	-0.06	-0.01	0.04	-0.00	1.00	0.10	-0.02	0.05	0.05	0.05	-0.01	-0.08	-0.00	-0.05	-0.04	-0.02	-0.09	-0.03	0.02	-0.14
famsup	-0.16	-0.15	-0.14	0.02	-0.11	0.02	0.18	0.19	0.05	-0.03	0.09	-0.00	-0.00	0.15	-0.06	0.10	1.00	0.29	-0.00	0.06	0.10	0.10	0.01	-0.02	0.01	-0.02	-0.03	-0.09	0.03	0.02	-0.06
paid	-0.02	-0.13	-0.04	0.05	-0.01	0.05	0.16	0.09	0.10	-0.04	0.13	0.04	-0.07	0.17	-0.19	-0.02	0.29	1.00	-0.02	0.10	0.19	0.15	0.01	0.00	-0.06	0.01	0.06	0.06	-0.08	0.01	0.09
activities	-0.12	0.10	-0.10	-0.05	-0.00	0.10	0.11	0.11	0.10	0.03	0.12	-0.02	-0.01	0.09	-0.07	0.05	-0.00	-0.02	1.00	0.00	0.10	0.05	0.02	0.04	0.09	0.05	-0.07	-0.04	0.02	-0.01	0.04
nursery	-0.09	-0.01	-0.09	0.06	0.10	-0.09	0.19	0.16	0.10	-0.03	0.06	-0.08	-0.03	0.08	-0.10	0.05	0.06	0.10	0.00	1.00	0.05	0.01	0.03	-0.00	-0.02	0.00	-0.08	-0.10	-0.02	0.02	0.07
higher	-0.02	-0.15	-0.21	0.04	-0.01	-0.04	0.17	0.17	0.10	-0.05	0.06	-0.02	-0.08	0.18	-0.30	0.05	0.10	0.19	0.10	0.05	1.00	0.02	-0.11	0.02	-0.06	-0.04	-0.07	-0.10	-0.02	-0.06	0.19
internet	-0.13	0.04	-0.11	0.22	0.00	0.07	0.20	0.13	0.22	0.04	0.05	-0.06	-0.11	0.06	-0.06	-0.01	0.10	0.15	0.05	0.01	0.02	1.00	0.09	0.03	0.05	0.07	0.04	0.01	-0.08	0.10	0.10
romantic	0.06	-0.10	0.16	0.01	0.03	-0.04	0.04	0.02	-0.06	-0.00	0.02	0.09	0.02	0.05	0.09	-0.08	0.01	0.01	0.02	0.03	-0.11	0.09	1.00	-0.06	-0.01	0.01	0.02	-0.01	0.03	0.15	-0.10
famrel	-0.05	0.06	0.05	0.01	-0.02	0.03	-0.00	-0.00	0.04	0.01	-0.01	0.04	-0.02	0.04	-0.04	-0.00	-0.02	0.00	0.04	-0.00	0.02	0.03	-0.06	1.00	0.15	0.06	-0.08	-0.11	0.09	-0.04	0.02
freetime	0.03	0.24	0.02	0.03	0.02	0.04	0.03	-0.01	0.11	-0.05	-0.07	0.05	-0.02	-0.14	0.09	-0.05	0.01	-0.06	0.09	-0.02	-0.06	0.05	-0.01	0.15	1.00	0.29	0.21	0.15	0.08	-0.06	0.00
goout	-0.01	0.08	0.13	0.07	0.02	0.00	0.06	0.04	0.00	0.01	-0.05	0.06	0.03	-0.06	0.12	-0.04	-0.02	0.01	0.05	0.00	-0.04	0.07	0.01	0.06	0.29	1.00	0.27	0.42	-0.01	0.04	-0.15
Dalc	0.11	0.27	0.13	-0.09	0.10	-0.03	0.02	0.00	0.08	0.08	-0.05	-0.01	0.14	-0.20	0.14	-0.02	-0.03	0.06	-0.07	-0.08	-0.07	0.04	0.02	-0.08	0.21	0.27	1.00	0.65	0.08	0.11	-0.07
Walc	0.07	0.27	0.12	-0.10	0.10	0.01	-0.05	-0.01	-0.01	0.07	-0.04	-0.04	0.13	-0.25	0.14	-0.09	-0.09	0.06	-0.04	-0.10	-0.10	0.01	-0.01	-0.11	0.15	0.42	0.65	1.00	0.09	0.14	-0.09
health	-0.04	0.14	-0.06	-0.04	-0.03	0.02	-0.05	0.01	0.07	-0.02	-0.16	-0.07	0.01	-0.08	0.07	-0.03	0.03	-0.08	0.02	-0.02	-0.02	-0.08	0.03	0.09	0.08	-0.01	0.08	0.09	1.00	-0.03	-0.08
absences	-0.09	-0.07	0.18	-0.03	0.04	-0.13	0.10	0.02	0.05	0.01	0.12	0.16	-0.01	-0.06	0.06	0.02	0.02	0.01	-0.01	0.02	-0.06	0.10	0.15	-0.04	-0.06	0.04	0.11	0.14	-0.03	1.00	-0.01
Grade	-0.04	0.10	-0.13	0.11	0.08	-0.04	0.22	0.18	0.10	0.06	0.12	-0.07	-0.13	0.13	-0.38	-0.14	-0.06	0.09	0.04	0.07	0.19	0.10	-0.10	0.02	0.00	-0.15	-0.07	-0.09	-0.08	-0.01	1.00



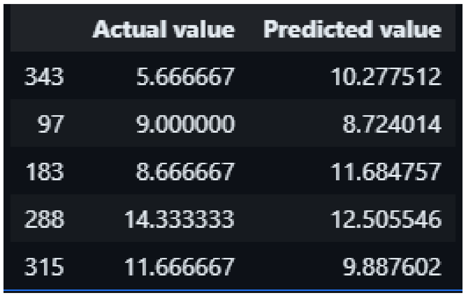
```
target_column_name = 'Grade'
correlations = df2.corr()[target_column_name]
sorted_features = correlations.abs().sort_values(ascending=False)
correlation_threshold = 0.1
selected_features = sorted_features[sorted_features >= correlation_threshold].index.tolist()
print(selected_features[1:])
```

```
['failures', 'Medu', 'higher', 'Fedu', 'goout', 'schoolsup', 'age', 'studytime', 'traveltime', 'reason', 'address', 'romantic', 'internet', 'sex']
```

We proceeded to build a linear regression model by dividing the dataset into train and test sets and obtained the coefficients of the regression model with an Intercept of 9.885774365061723 Then, we proceeded to test each feature if it's significant or not and build a model.We got an R^2 of 23.44 with Mean Absolute error of 2.74.



We would also like to present a comparison between predicted and actual values in the dataset. As we can see, R^2 of 23.44 clearly displays the perfomance of the model.



	Actual value	Predicted value
343	5.666667	10.277512
97	9.000000	8.724014
183	8.666667	11.684757
288	14.333333	12.505546
315	11.666667	9.887602

5 Conclusions

In conclusion, our exploration of the dataset has provided valuable insights. Through a combination of distribution fitting, statistical tests, and regression analysis, we've established a nuanced understanding of how variables interact within the confines of the dataset.

According to the linear regression, we have found out that the best performance would show someone who has parents with higher education, no failed classes in the past, who is aiming to continue studying in the university after school, who lives less than 15 minutes to school, has access to the internet and studies more than 10 hours a week. Variables that contributed the most were: Number of classes student failed in the past, Whether or not student receives extra educational support, and if the student wants to resume their studies.