Zewail City of Science and Technology

Machine learning course

Instructor: Mohamed Elshenawy

Ahmed Adel 201901464

Abdulla Sabry 201701484

Omar El-Sakka 201900773

University of Science and Technology March,

Problem definition, and Motivation.

We are working on the student performance dataset provided by UCI. The dataset deals with students who are in secondary school; two Portuguese schools. And it states some attributes for some students like father and mother education in order to predict final exam grades.

In our schooling systems nowadays, it's really important to create some models that can predict student performance, to make some conclusions based on them. Now schools can predict if students will pass a course or not, and based on it, they can define how much extra support they have to give to students. And help them create a unique and custom plan for each student

Basically, the website states that the aim of this dataset is to predict the final exam grade(G3) given first-period exam(G1) and second-period exam (G2). and in addition to that, we could make a pretty good model predicting G3 without G1, nor G2; to help schools predict G3, even before the start of the semester.

We created models that might help schools to know their students better, based on their grades; schools will be able to determine mother and father education, in addition to parents' status; if they are apart or not, that might help them give special support for some students. We also provided a model to predict how many days students will be absent given all grades in different exams. That might be useful, if the sheet of attendance is somehow lost, and they might need to estimate how many days each student came to school.

Moreover, we did great exploratory data analysis, to examine some interesting questions; for example, we did an EDA on interesting features related to students who failed at least once, and many many more.

And at the end of our project, we aim to give some tips to parents, for how to take care of their children, tips for getting high grades, all of these tips will be based on detailed analysis and hypothesis testing, they can examine themselves.

Dataset

1. Description
   The data set is in tabular form with many features regarding some students in secondary education of two Portuguese schools. And it records their performance in two different subjects; Mathematics (mat) and Portuguese language (por).

   The data has no missing values at all. In addition, it was correctly encoded by an ordinal encoder. Although, we made more cleaning which we will explain in the cleaning section.

   Our data set has 33 columns and here is their description.

School	Student school. GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira But that column was dropped to generalize our result to all schools
Sex	Student's sex (binary: 'F' - female or 'M' - male)
Age	Student's age (numeric: from 15 to 22)
Address	student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal: 'mother', 'father' or 'other')
traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)

nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
Internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first-period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20, output target)

2. Challenges.
    a. The date set was so small for models we used, about 650 for Portuguese language classroom and 350 for Mathematics classroom.

But we could overcome this problem, by combining two datasets to form 1050 rows which increased our accuracy by a very large difference.

We also dropped the school column, and hence, our models are aimed to be generalized for any school and any subject.

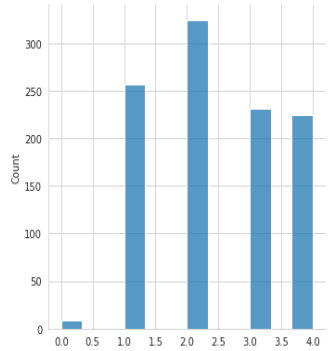
b. columns were so sufficient to describe student behavior, they were meaningful and repressive. But the column was encoded in ordinal. We hope not to do a more detailed analysis. And we might want to encode them in a different way, we might have wanted to put different scales other than them.

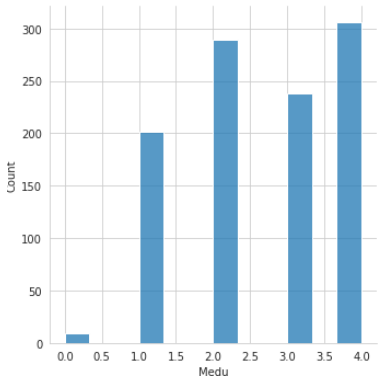
c. Website didn't provide which student they collected their data. And wanted to make sure randomly students were selected. And what criterion did they use, was it stratified or what? Does the percentage of each category of students in our dataset is really representative of the real distribution? There was not enough information about that.

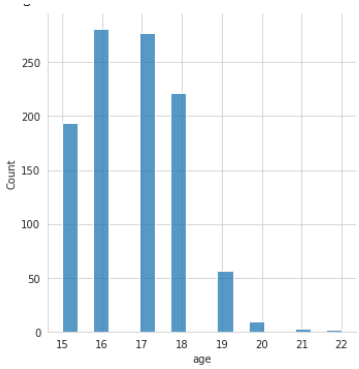
Exploratory data analysis.

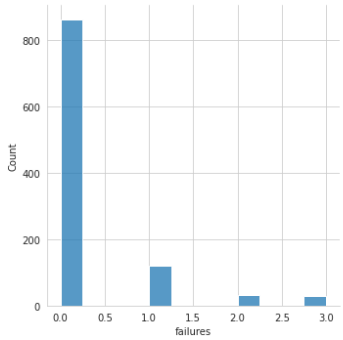
We had many questions, inspired by the dataset columns, and by our life as students, we tried to answer them via a detailed analysis. And here we will present our questions and how we approved the answer.

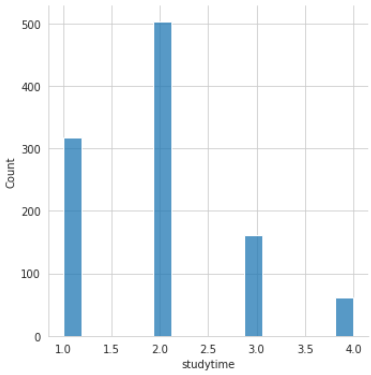
1. But before let's start with univariate analysis

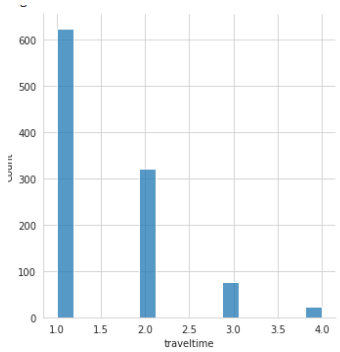


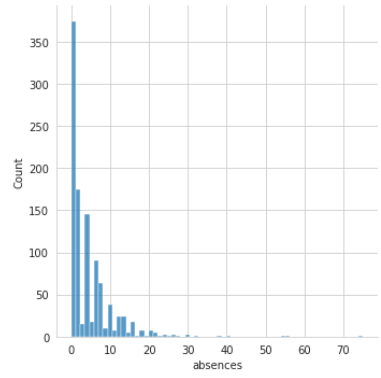


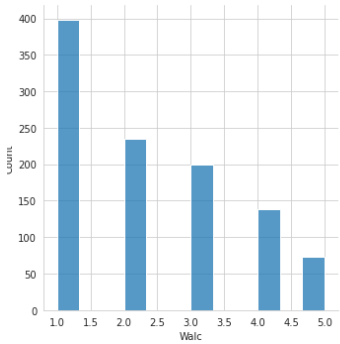


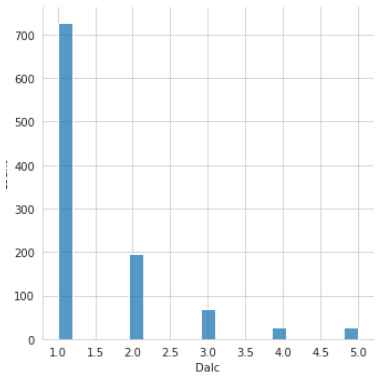


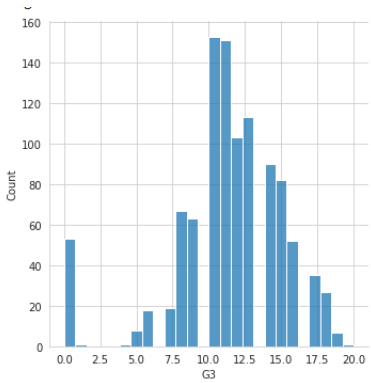


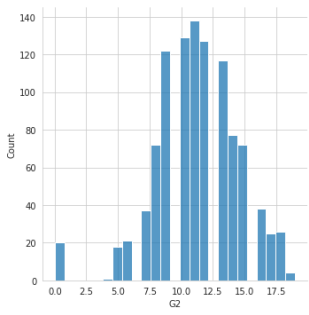


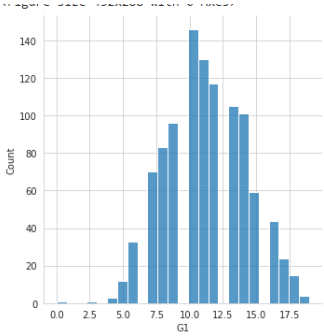










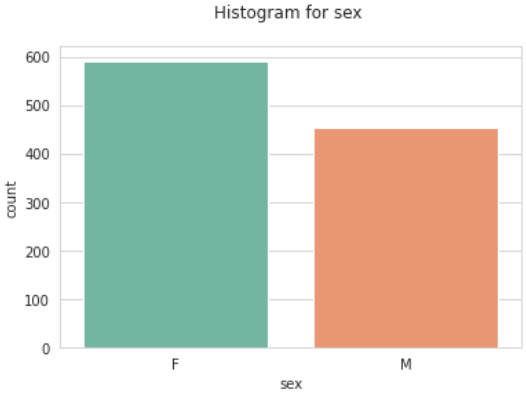


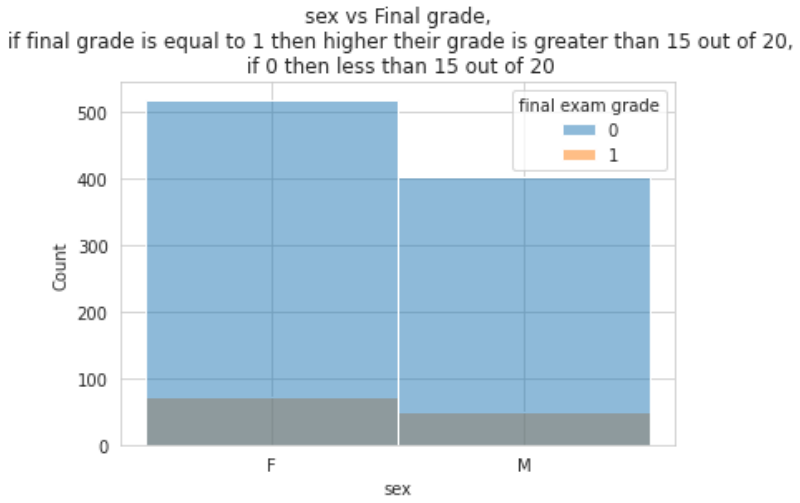
- Starting with age, we see most of our data set is concentrated between 15 and 19 years old.
- Most mothers and fathers are having a good education.
- Most students are having short travel time.
- And for alcohol consumption, their histogram is right-skewed, most of the students consume a very small amount of it
- For exam grades, they are perfectly normally distributed. Which fits any Gaussian distribution needed by any model.

2. which sex has higher grades in each exam?

We find in our data, males, and females are nearly balanced, and hence we can draw concise conclusions between them.
Males represent 44% and females represent 66%

And we plotted their grade distribution as below



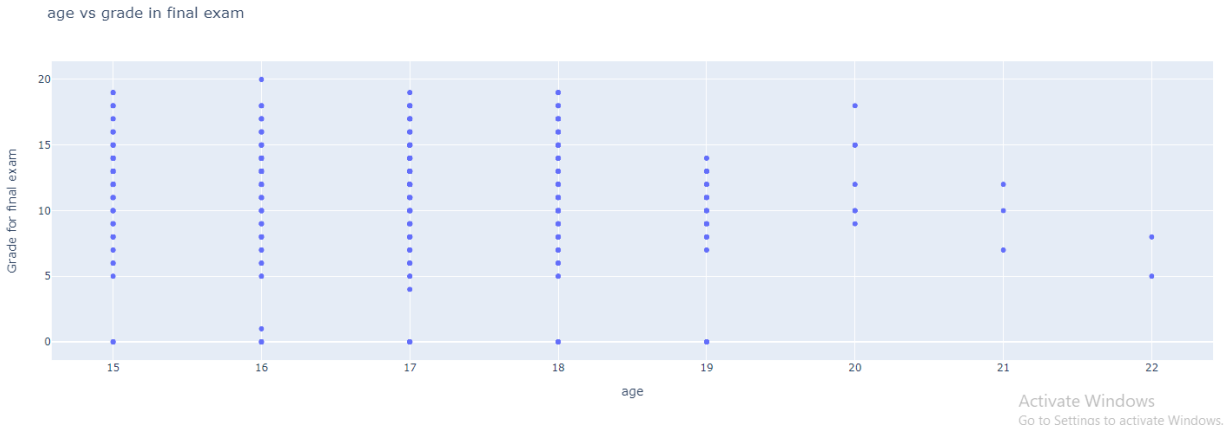


- And we found, there were 72 females(6.8%), and only 50 miles(4.8%) having grades greater than 15.
- And there were 519 females(49.7%) and 403 males(38.6%) having grades less than 15
- this is not analogous to that distribution of males and females $\frac{56.6}{43.9} \neq \frac{6.8}{4.8}$
- But their average grades in the final exam are very close (about 11.5)

To be even more sure, we used p-value to test our hypotheses, for the Null hypothesis, that males and females are having the same mean values for final exam grade.

And we found the p-value is equal to **0.68.** Meaning we can accept our null hypothesis, that they have the same average grade, not just due to some randomness.

3. Do students who are older have more chances to get higher grades? (because they should have more awareness) or they just have failed and don't care about the class?



secondary education: the second stage traditionally found in formal education, beginning about age 11 to 13 and ending usually at age 15 to 18. According to Britannica

So we say that max-age is 18 years old, we find starting from age 19, the performance drops severely down. The max grade for people of age greater than 18 is 18 out of 20.

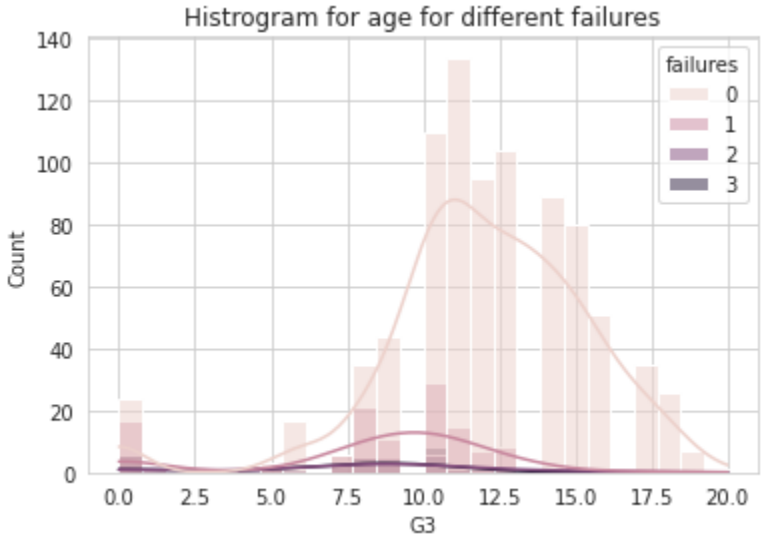
If a student is in secondary school at an age greater than 18, this might be due to he\she failed in school before, and that might be an indication that they are not interested in education at all.

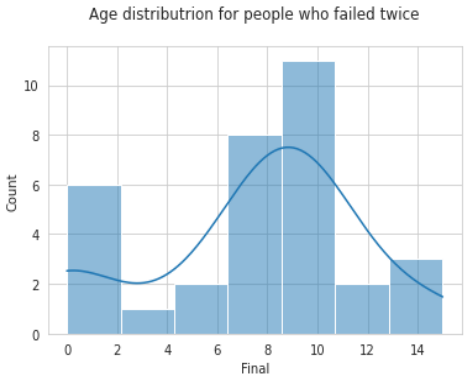
But that doesn't mean they are bad in all fields, they might be doing great extra activities.

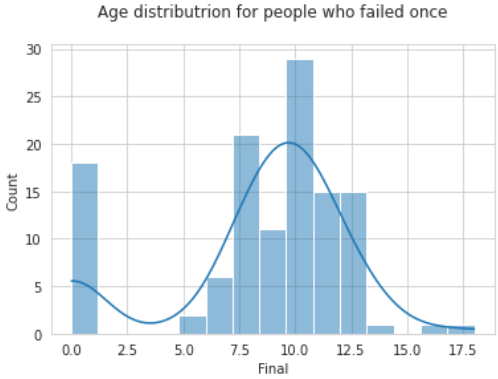
Let's explore them in the next section!!

4. What are the most interesting characteristics for students who have failed at least once?

- We started by plotting grade distribution for different failures numbers





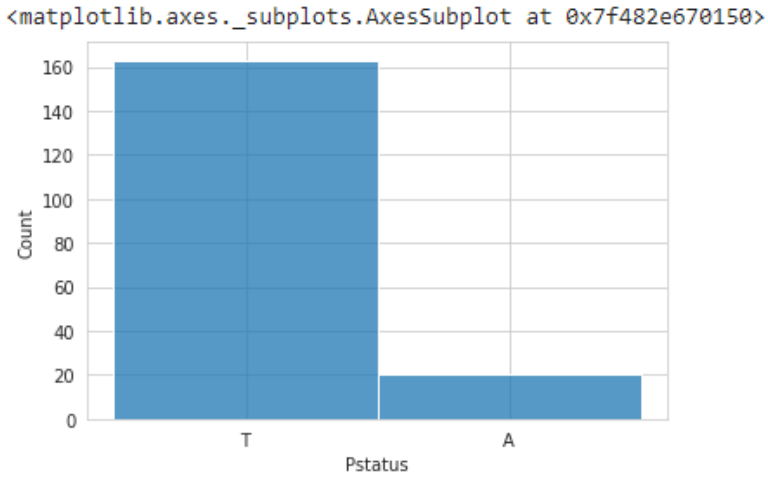


- Most people who have failed once, they have an average grade of 10, and a max of 17.5 out of 20
- Most people who have failed twice, they have an average grade of 10, and a max of 14 out of 20
- most people who have failed three times, they have an average grade of 10, and a max of 10 out of 20

We see max value decreases with increase in a number of failures.

- And then we started analyzing parent status effects

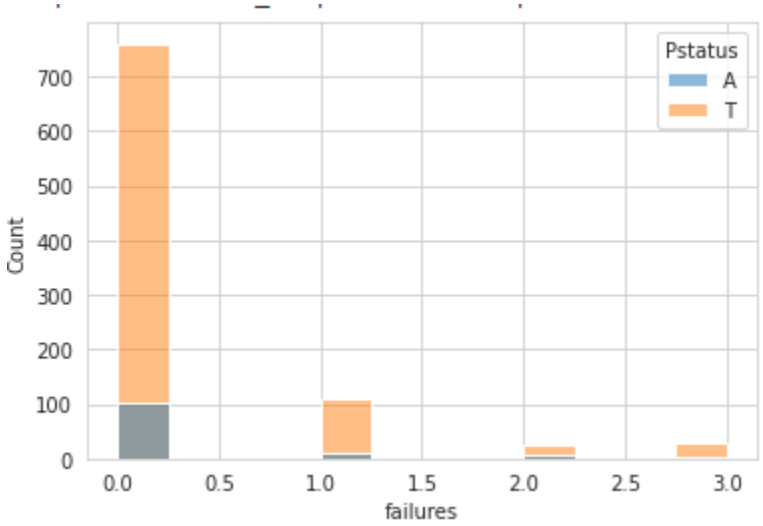
We used P-value to test if a student with a parent has the same average value as a student who has a normal family. And that was our null hypothesis. And we found out that the p-value is equal to 0.8. So we accepted our null hypothesis. And found out that divorce nearly has no effects on students.



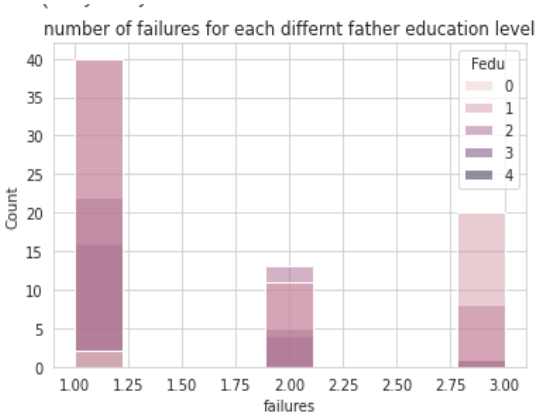
We actually read a paper confirming our hypothesis.

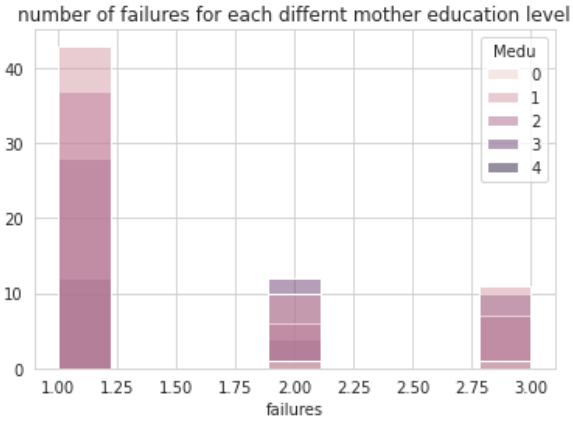
Noting there are more people who have their parents living together failed more.

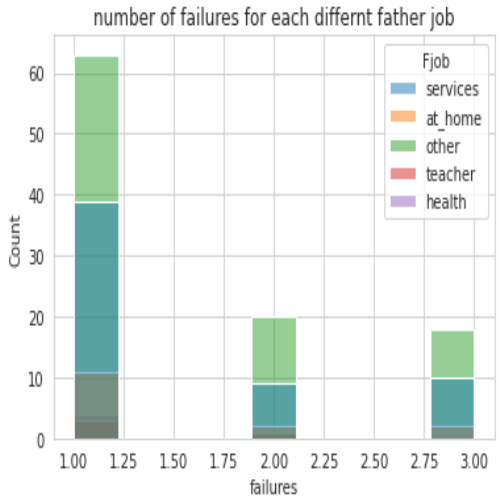
parents' breakups don't seem to affect students' performance. We read a lot that ensures that.

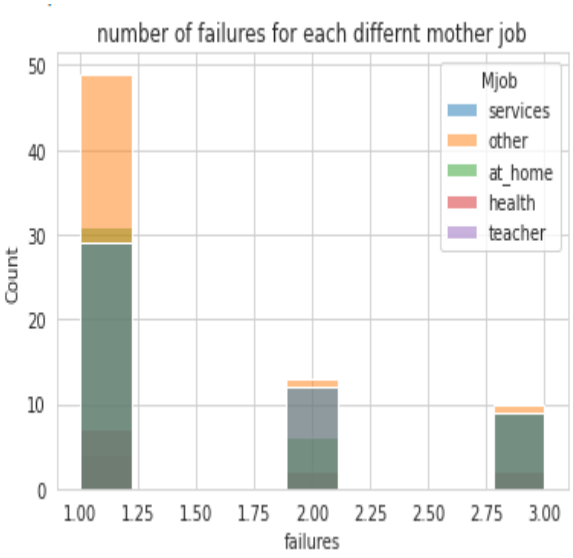


And then we examined the effects of mother and father education.









- It's observed as parents' education is better, the number of failures is less,

- Max number of students who fail at least once decreases with the increase of parents' education level

- Although some parents are highly educated, they fail 3 times. and for students who have parents with primary education only; mostly fail once.

- it's noted when father\mother is a teacher, students at max fail once not more. and same for mother being at home, it seems when father\mother is at home students fail at most once and the number is already so small. while when one is working; the number of fails increases (More analysis for this part is in a notebook.)

And here is the detailed number for different education levels and working statuses. that confirmed our hypothesis.

When mother education is 1

Number of students failed once	43
Number of students failed twice	10
Number of students failed three-time	11

When mother education is 2

Number of students failed once	37
Number of students failed twice	6
Number of students failed three-time	7

When mother education is 3

Number of students failed once	28
Number of students failed twice	12
Number of students failed three-time	10

When mother education is 4

Number of students failed once	12
Number of students failed twice	4
Number of students failed three-time	1

When father education is 1

Number of students failed once	40
Number of students failed twice	11
Number of students failed three-time	20

When father education is 2

Number of students failed once	40
Number of students failed twice	13
Number of students failed three-time	8

When father education is 3

Number of students failed once	22
Number of students failed twice	5

Number of students failed three-time 1

When father education is 4

Number of students failed once 12

Number of students failed twice 4

Number of students failed three-time 1

When a mother is working in services-

Number of students failed once 29

Number of students failed twice 12

Number of students failed three-time 9

When a mother is teacher

Number of students failed once 1

Number of students failed twice 0

Number of students failed three-time 0

When a father is working in services

Number of students failed once 39

Number of students failed twice 10

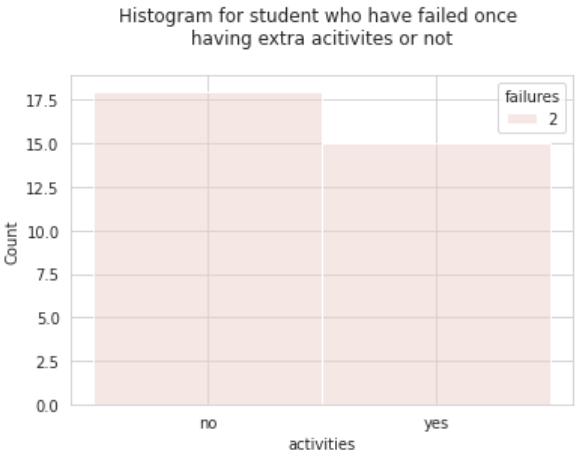
Number of students failed three-time 9

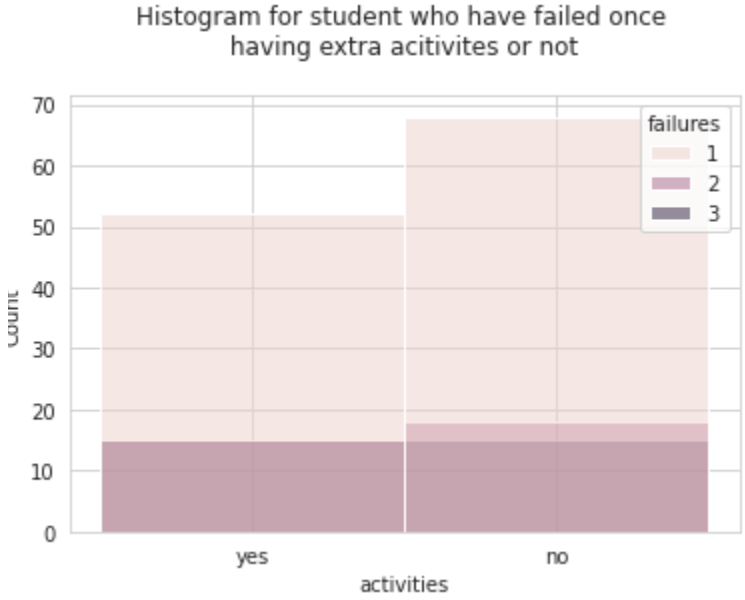
When a father is teacher

Number of students failed once 3

Number of students failed twice	1
Number of students failed three-time	0

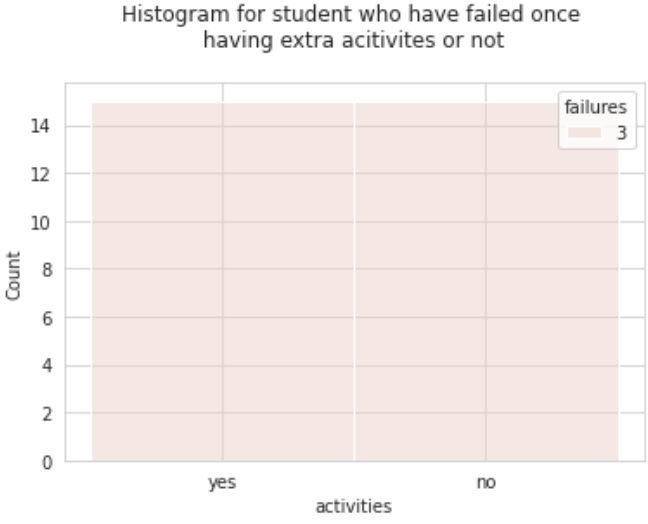
And then we examined the effects of activities.





We found that about 44% of students who have failed at least once are having activities.

That might let us decide that activities affect student performance badly, the majority of people who failed three times were having extra activities. **BUT** data size is so small that we can't generalize.

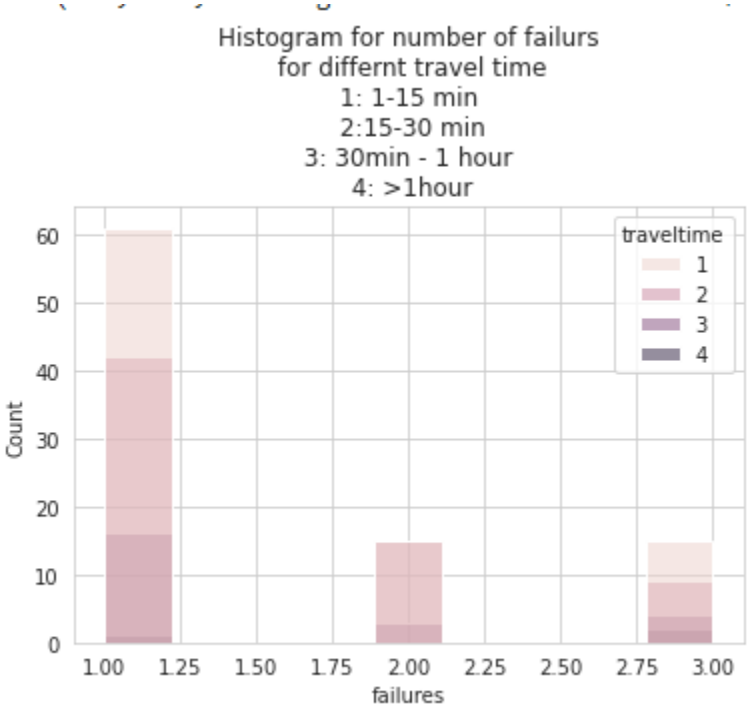


And we also looked at their travel time.

- Although people have travel time less than 15 minutes, some of them three up to 3 times.

  .

- but people failing twice, they have always travel time 15-30 minutes



We tested that students having travel time greater than 15 minutes have the same grade as less than 15 minutes (travel time =2).

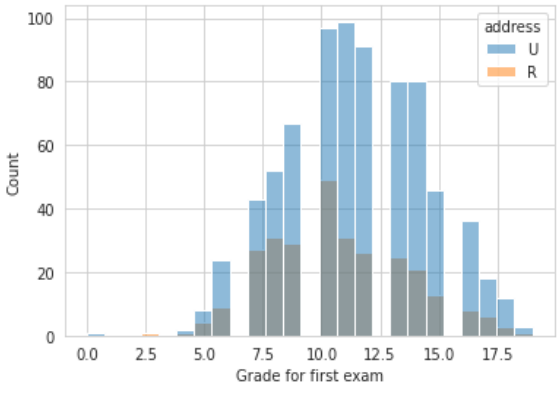
- And found a p-value equal to 0.01

so we reject our hypotheses. and we can say it's not due to randomness at all. it's better to get to a school that's near to your home.

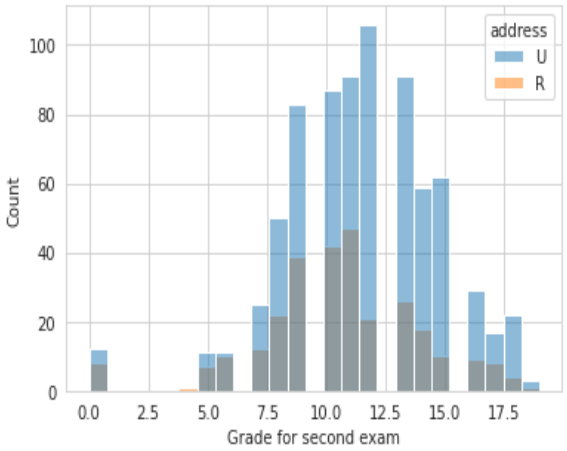
5. Does address affect student performance? (in each exam!!)

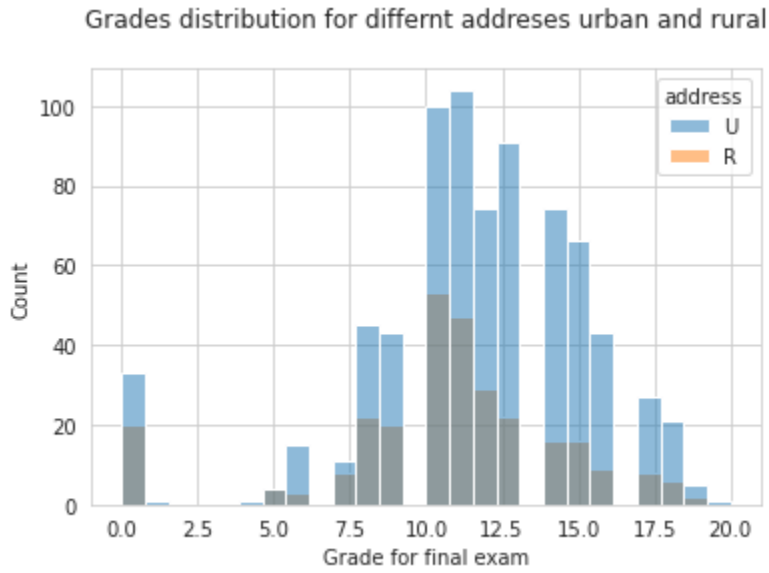
We find that our data is not balanced in an address at all.

Grades distribution for first exam for differnt addreses urban and rural



Grades distribution for second exam differnt addreses urban and rural

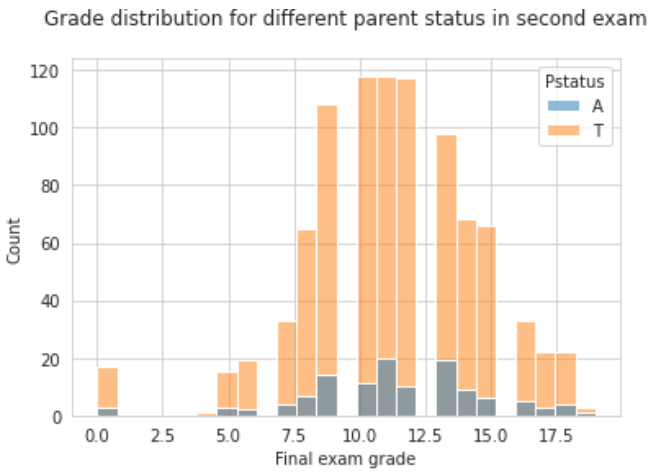


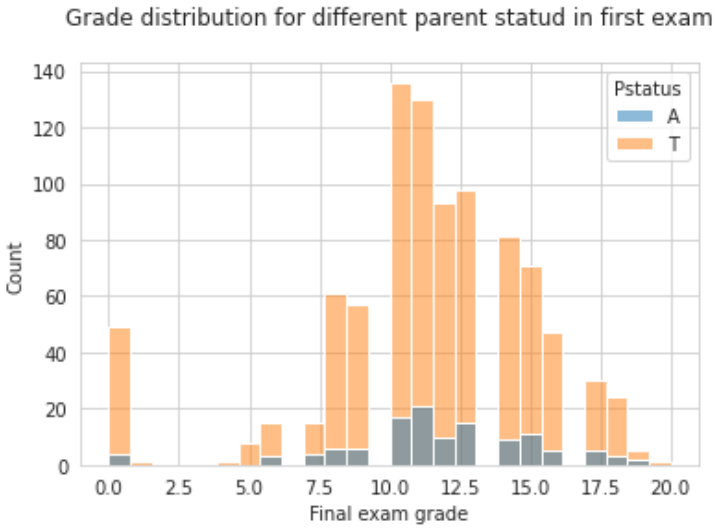


We tested a hypothesis that urban and rural average grades are equal. And we got a high p-value so that we can ensure no big difference between urban and rural at all.

6. Does parent status affect student performance overall?

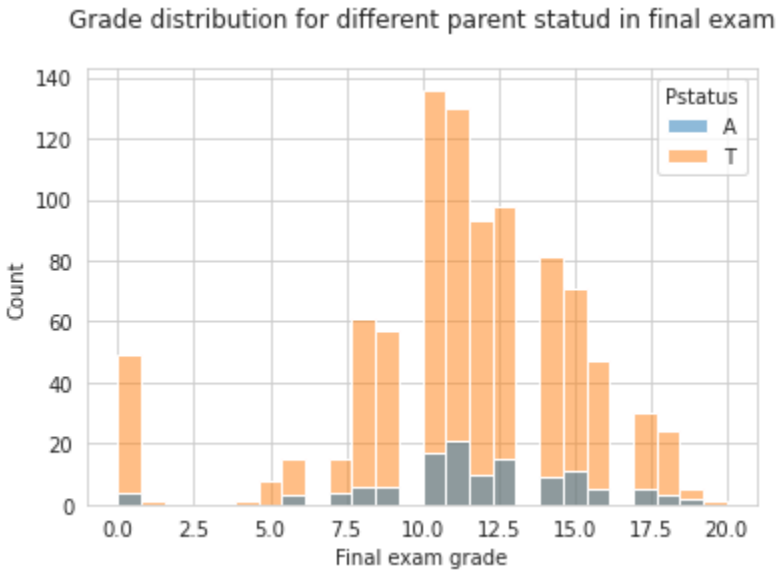
We plotted the final exam grades for different parent statuses.





We find that both categories are having the same distribution.

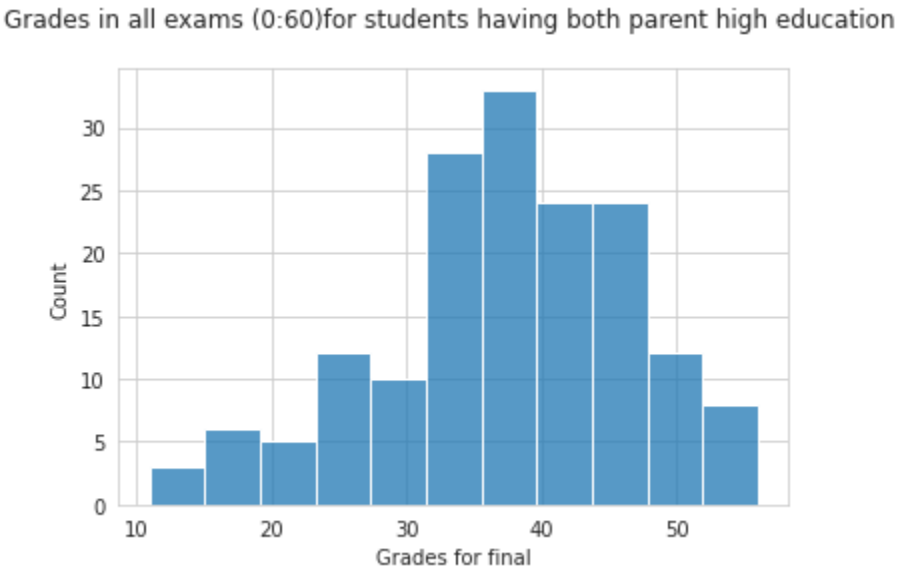
It's also found that for students having their parents apart, we couldn't find one with a grade greater than 19 in any exam. But it might be due to randomness. as not a big difference



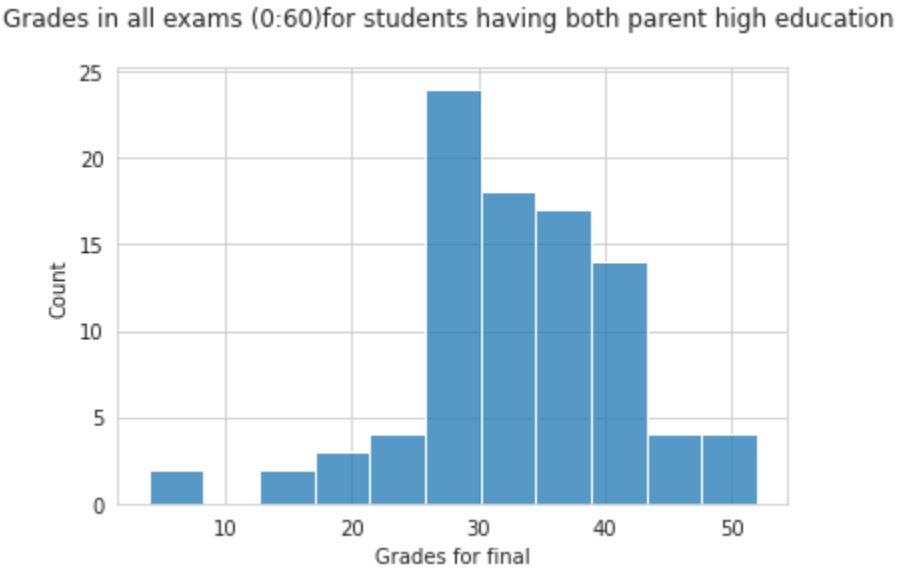
7. How does mother's and fathers' education affect student performance?

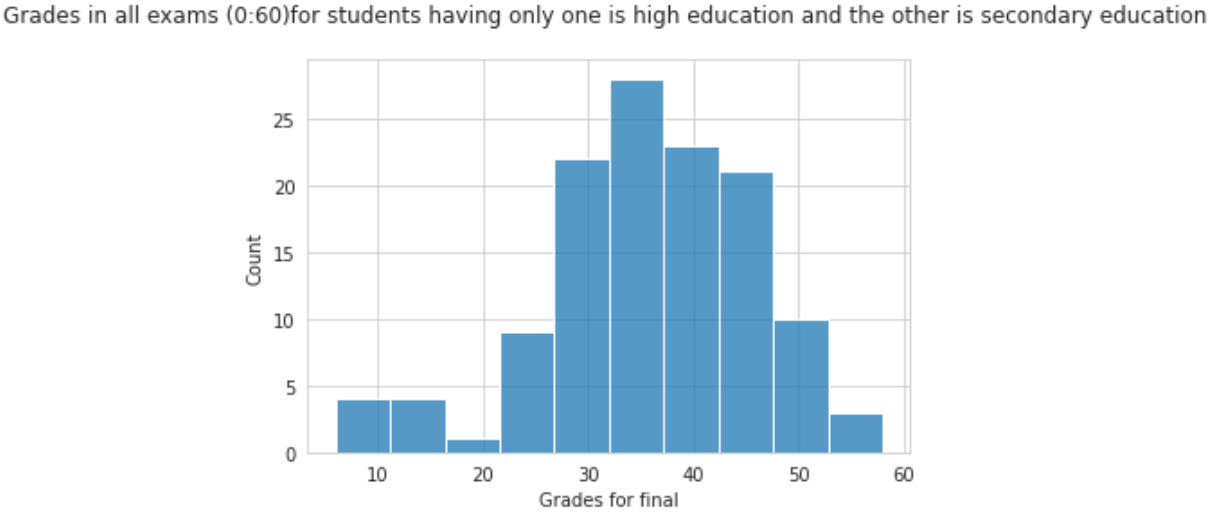
We summed grades in three exams, to be equal to 60 in our section here.

we find out that most of the students having their parents both of high education have high grades. their median value is 38 out of 60. And mean 37.



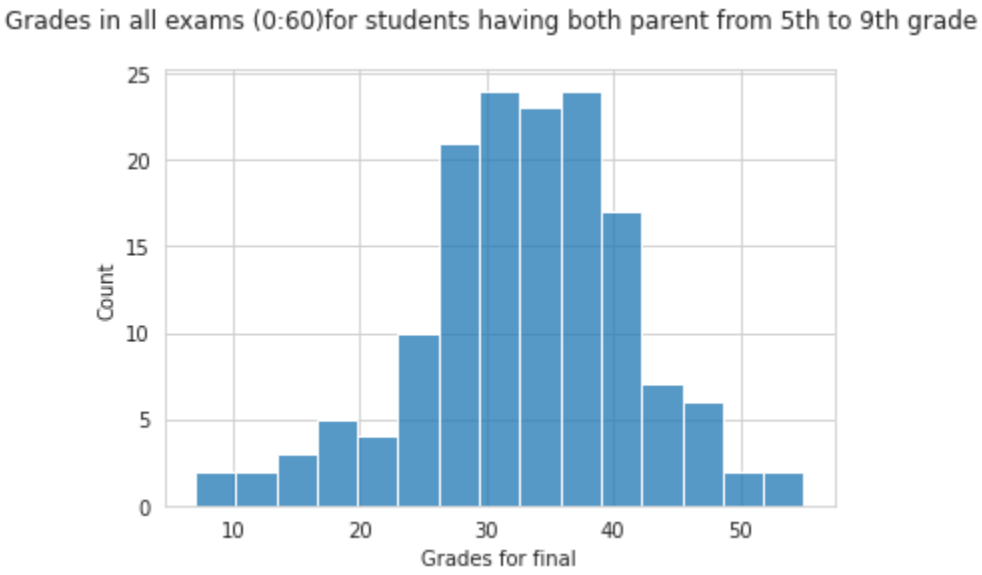
when both parents are educated from secondary school, students have a median of 34 out of 60 and mean 33





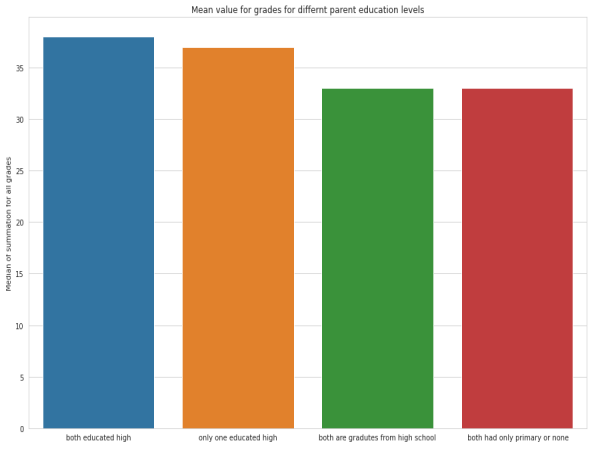
If one of the parents has high education, their average grade is 35 and the median is 37, better than the case when both are secondary education graduates.

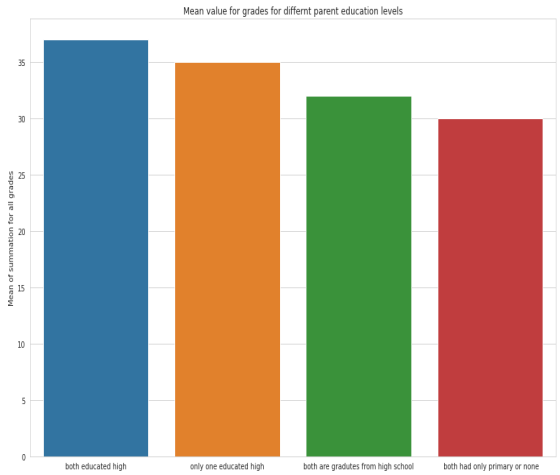
if both parents are from 5th to 9th grade, the average grade is 32, and the median is 33



And when a parent's had only primary education, or none, avg grade for their students is 31, and the median is 33

And here is a summary.

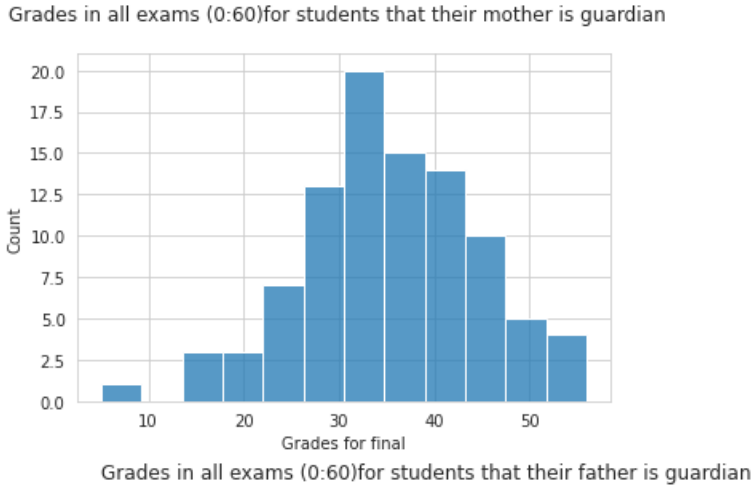




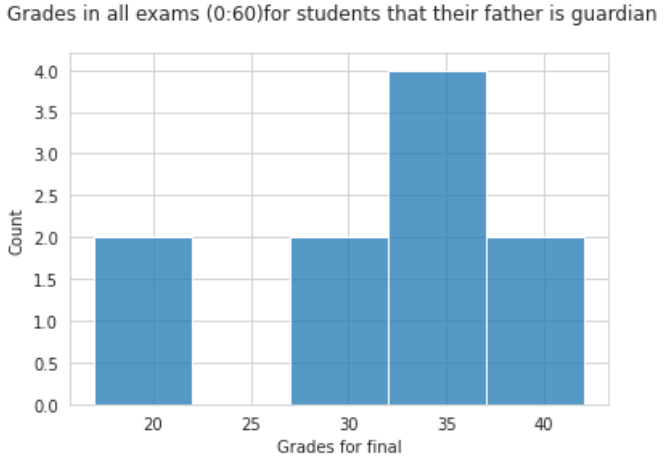
We can see a huge difference in student performance based on parents' education level. so we can conclude it's very important to have well-educated parents to motivate children to study.

8. If parents are apart; does it affect which guardian guards the student in his grades?

It seems when they are apart if the mother is guardian, the average grade is good it's 35



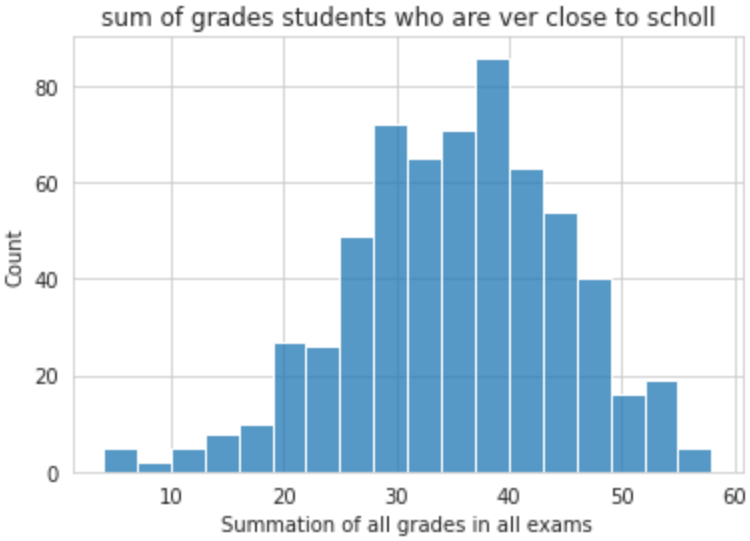
When the father is the guardian, they have a less average grade, mean =32.5



So we can conclude that mothers as guardians can do better than fathers. this might be due to fathers working more hours than women

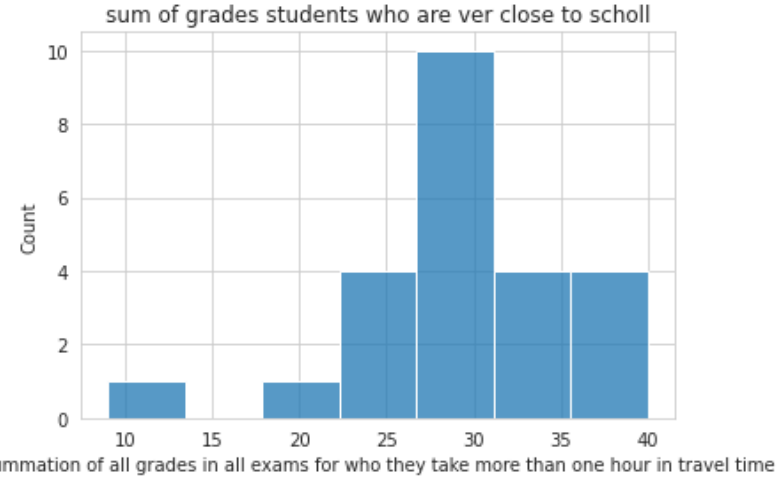
9. If a student's home is far away from school, does that affect the student's performance?

When they are close to school, they have an average grade of 36. but there are some weird outliers. there are some students with very low travel time but still, have 0 in exams

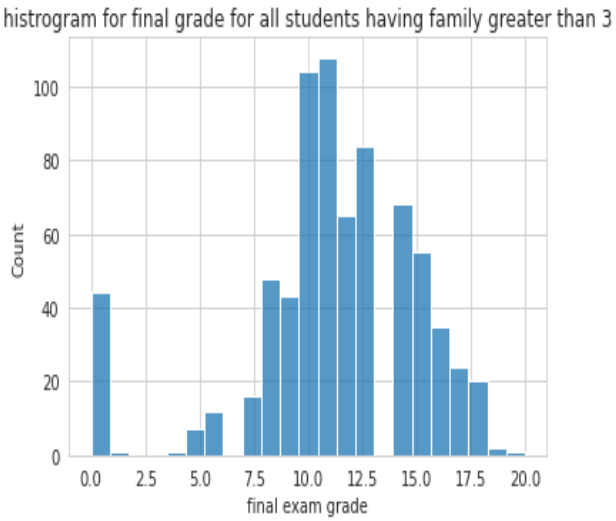


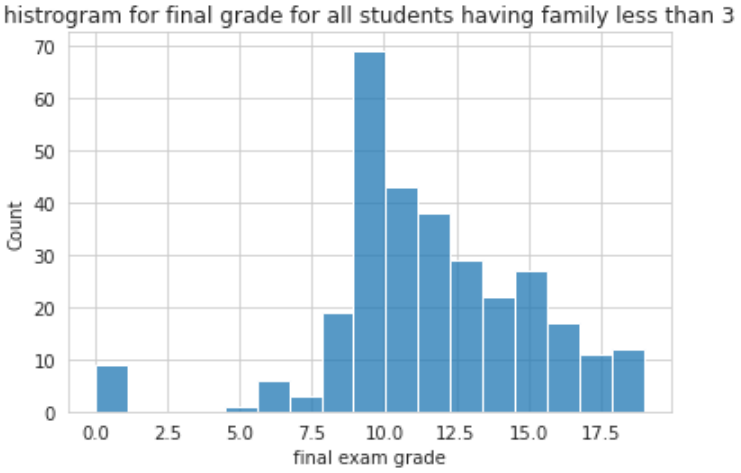
And for students having more than one hour, they get on average 29 out of 60.

So we can conclude that travel time affects student performance. and if they are close to the school they can perform much better



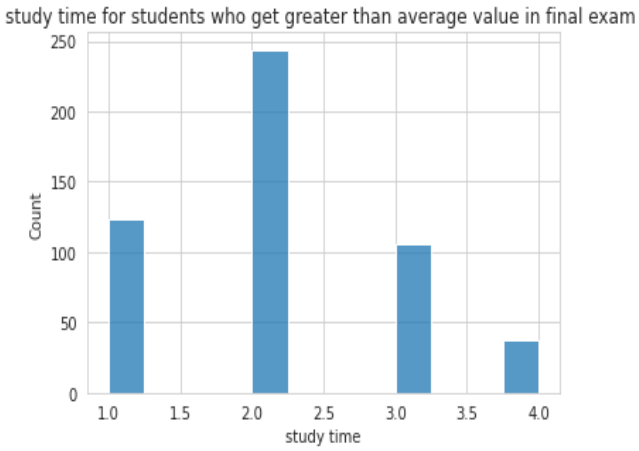
10. How does family size affect student performance?

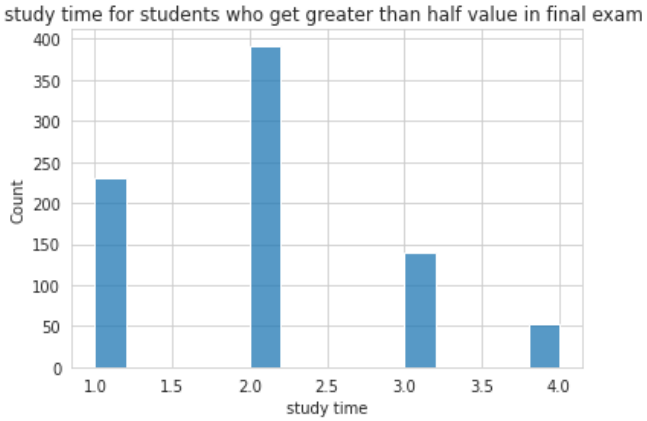


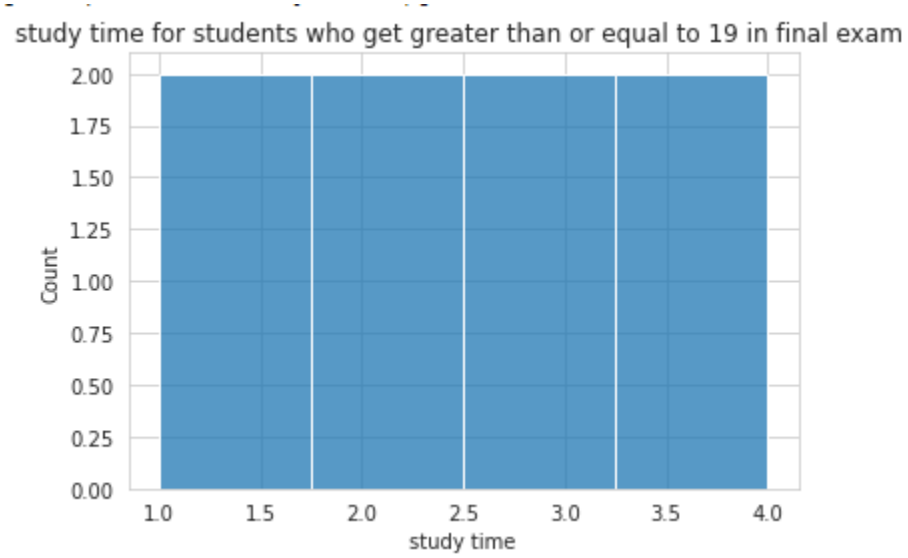


We found out if students have a family size greater than 3, or less than 3, they both have mean and median. And we concluded that there is no effect produced by student performance.

11. How many hours (in study time) should students have to get grades greater than half? and greater than average?





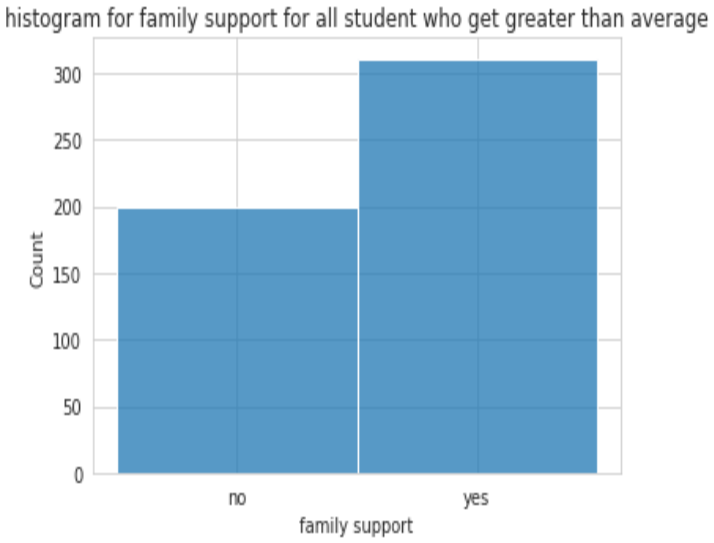


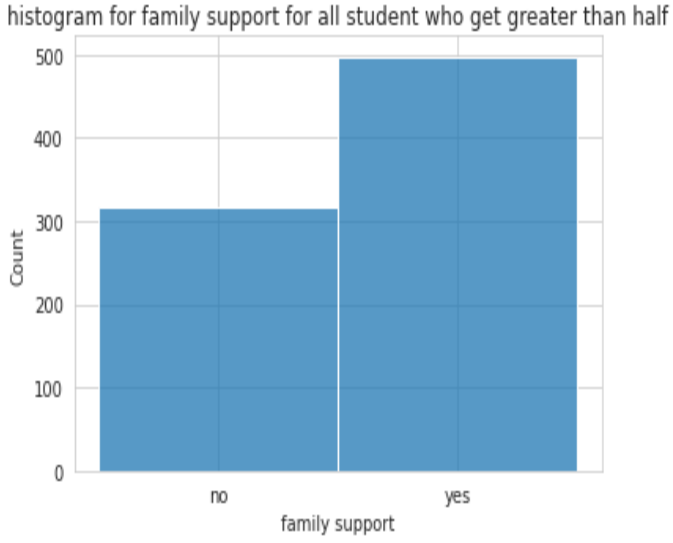
- We see here it takes about 2 hours to get grades greater than half, but from distribution above, we can say that are more people having study time up to 4.5 to just get greater than half or greater than average

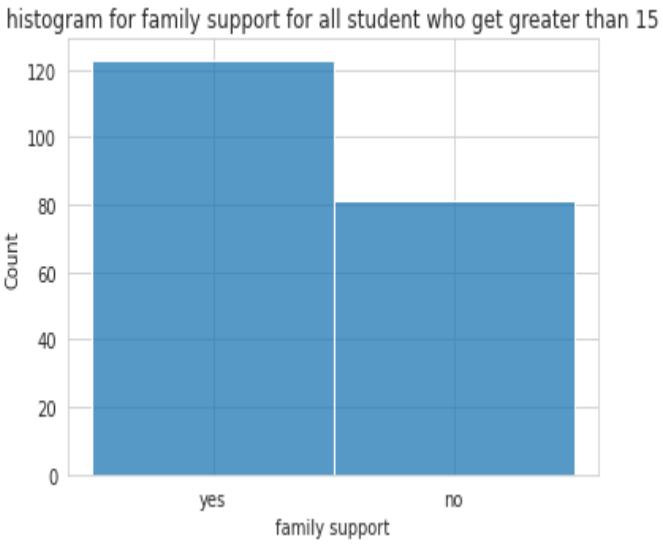
- But it's really clear that if they want to get high grades for like greater than or equal to 19, mean value of study time in 2.5 hours

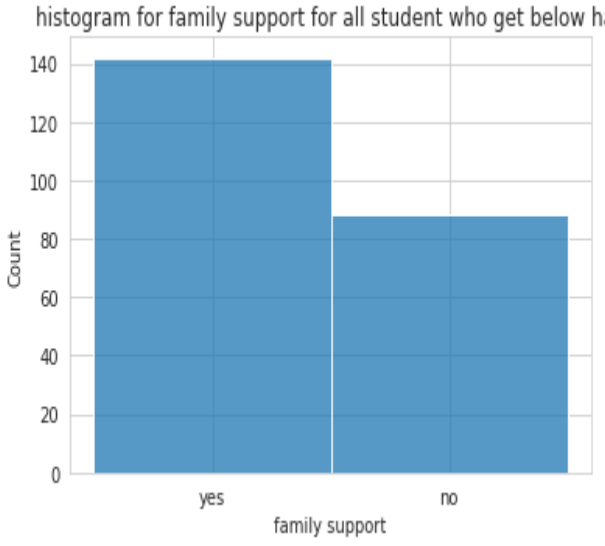
- But for having grades greater than 19, it might take some students more than 5 hours just to get the same grade

12. Is it necessary to have family support in education to get grades greater than average or greater than half?





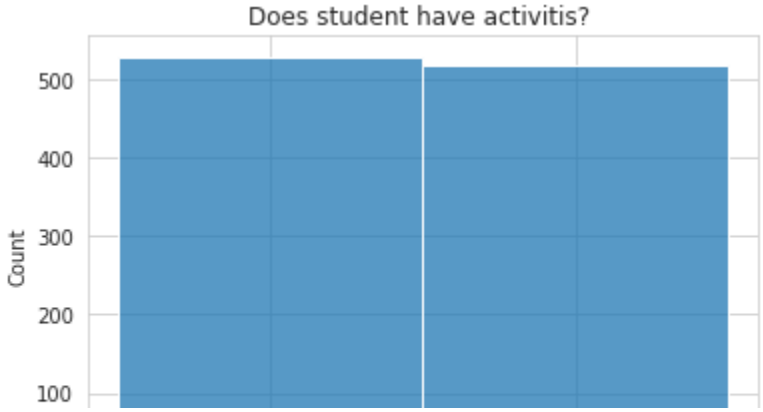


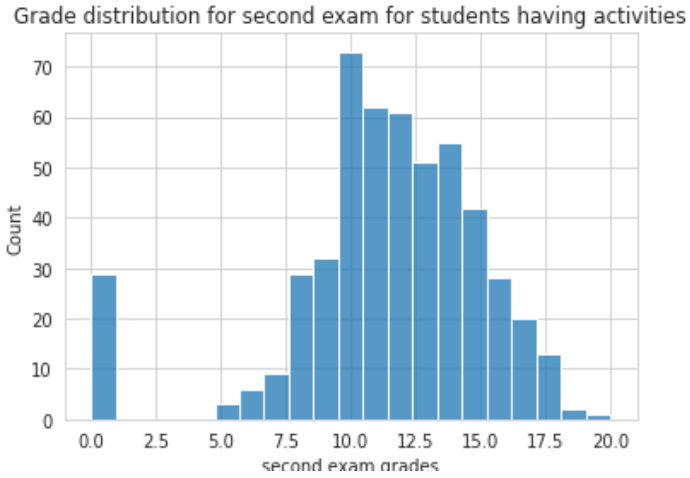


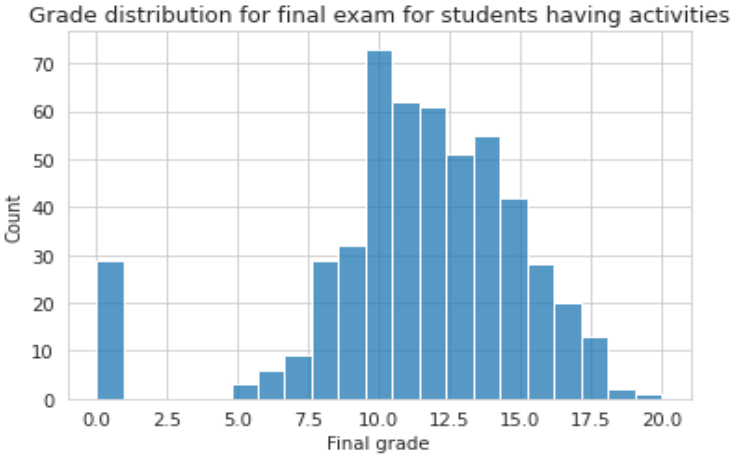
We see there is no serious indication that reflects the effect of family support on grades. But it would be a good insight to give family support to our students.

13. Do extra activities put more burden on students? can they also get a high grade while they do them?

We first started with checking the balance and found out, it's perfectly balanced.



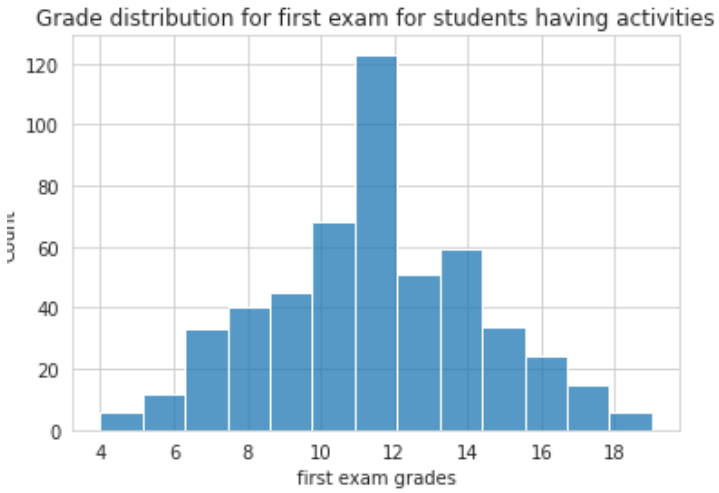




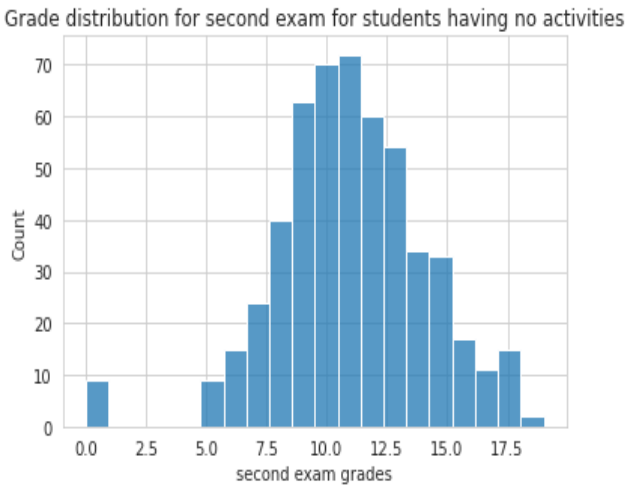
A first student having activities:

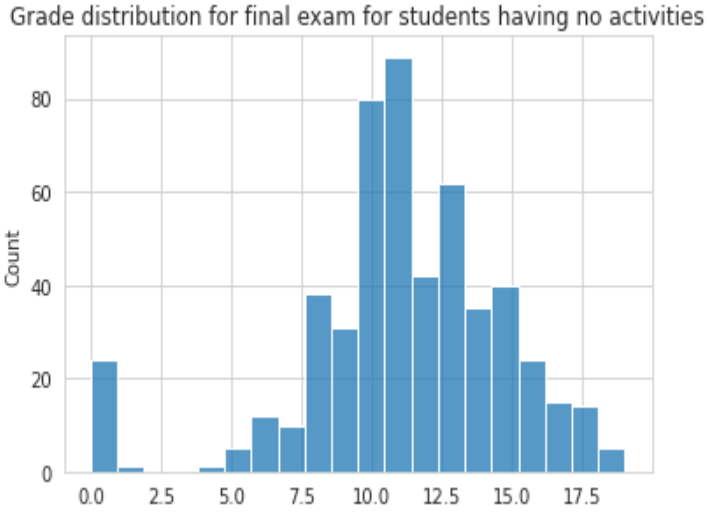
- Mean value for the first exam= 11

- Mean value for second exam= 11
- Mean value for final exam= 11

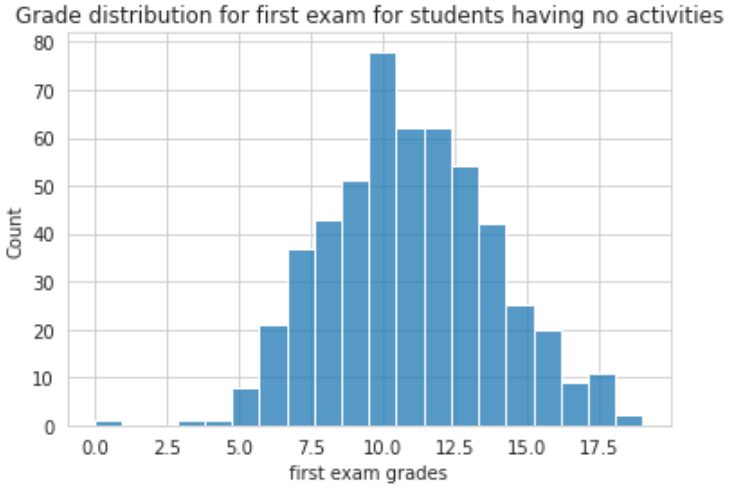


And for a student having no activities

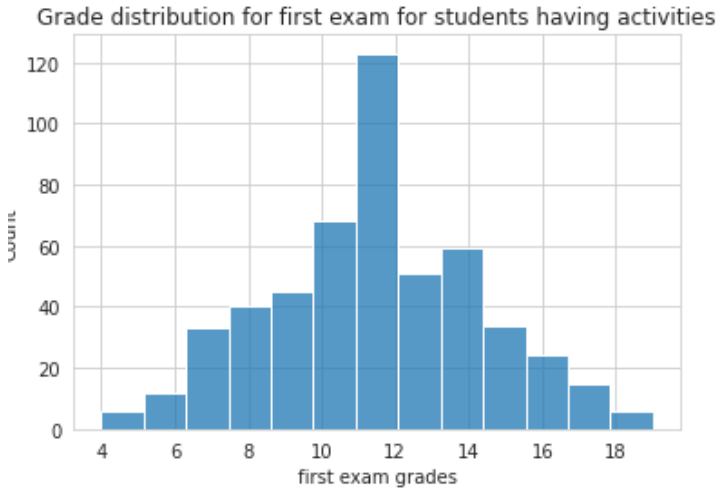




For a student having no activities:

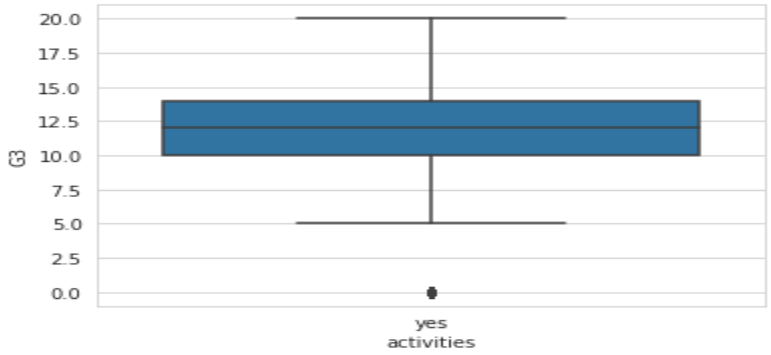


- Mean value for the first exam= 11
- Mean value for the second exam= 11
- Mean value for final exam= 11

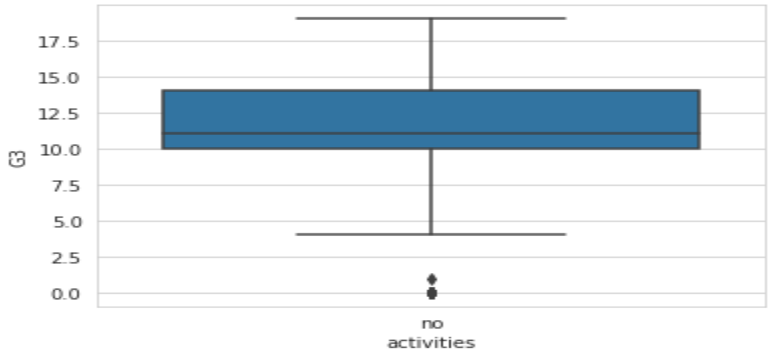


Both categories have nearly the same distribution.

```
ax = sns.boxplot(x="activities", y="G3", data=df_act)
```




```
ax = sns.boxplot(x="activities", y="G3", data=df_no_act)
```



We see here both kinds either doing activities or not have the same distribution, and the same mean value. We can infer that doing activities does not affect student performance very badly.

student performance cannot be determined by their activities

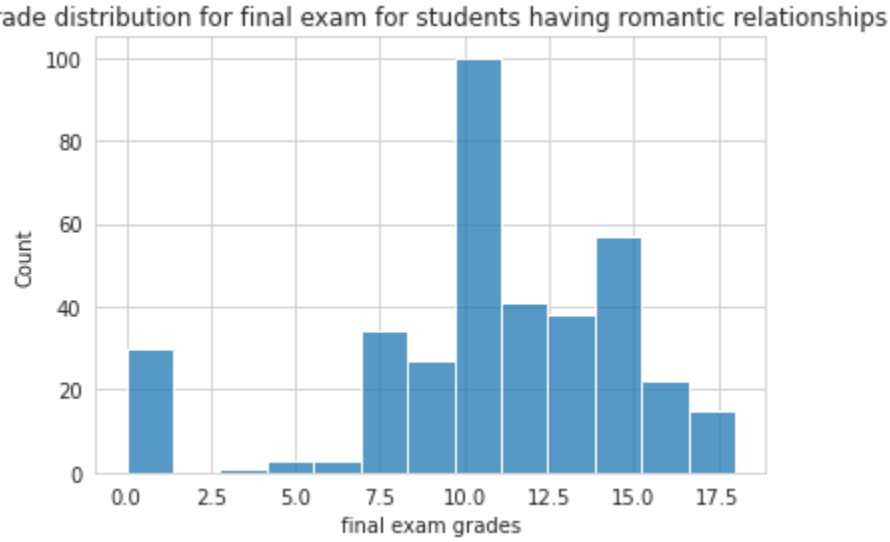
And we tested our null hypothesis, and it confirmed our analysis. For p-value 0.99

14. Are grades affected by romantic relationships?

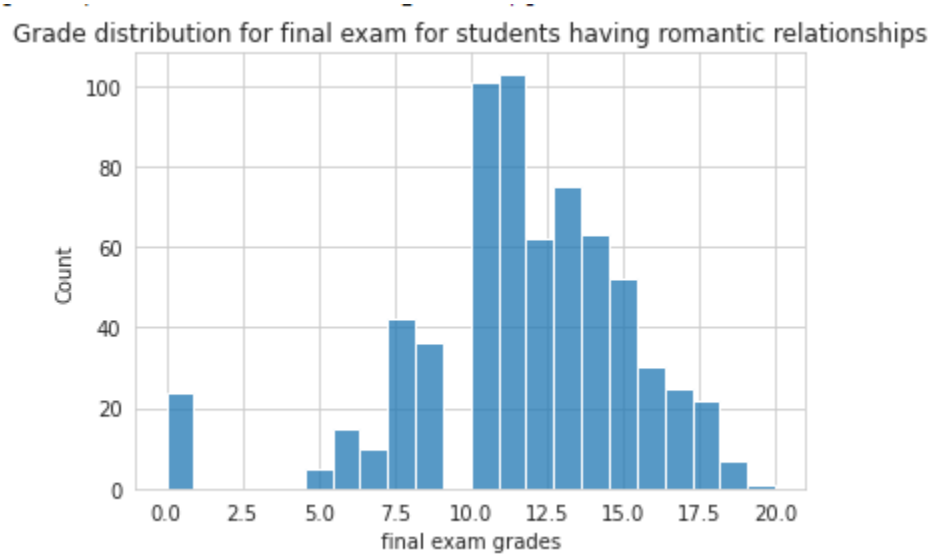
First, we started with a p-value, for hypotheses that both having and not having relationships have the same grade distribution, and found it's less than 5%, and we simply rejected it.

And here is the grade distribution for students having relationships.

We also found their mean for final exam grades is 10.4, and the median is 11.



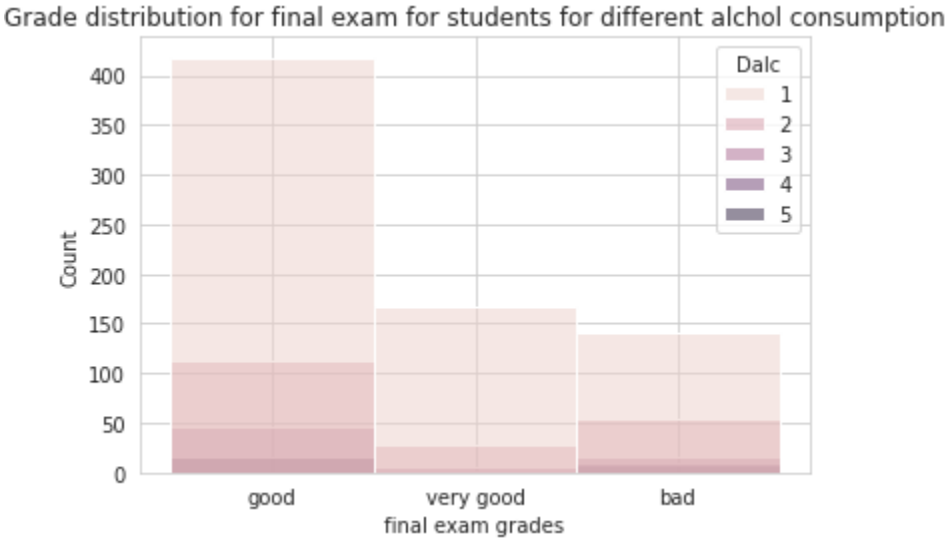
But for students having no relationships; they have a mean of 11.4 and a median of 12. Which is a bit higher than the other category.



We can infer that romantic relations and grades are negatively correlated.

15. Are grades affected by alcohol?

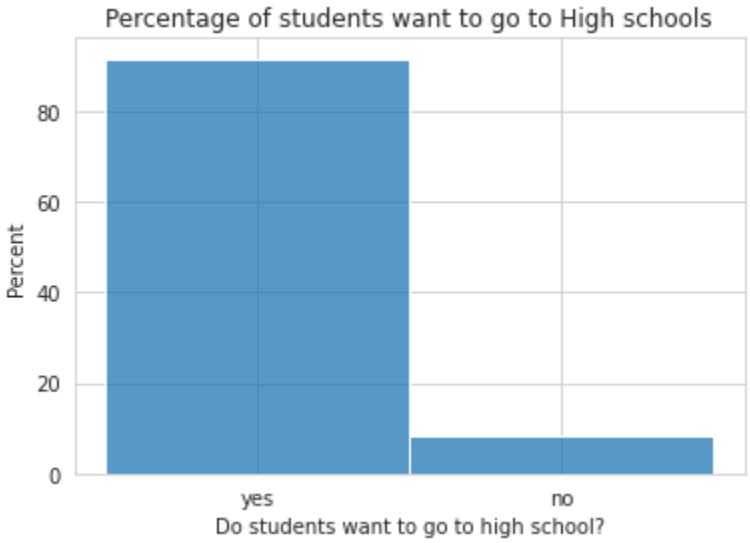
We started with a histogram of final exam grades



And we went directly to test hypotheses for using and non-using alcohol students, having the same average value. And we got a p-value very small, that we can infer it affects student performance

16. Is wanting to go to high school enough motivation to get high grades?

About 85 percent of our dataset want to go to high school



For students wanting to go to high school:

Mean in G3:11.6

Median in G3: 12

For students not wanting to go to high school:

Mean in G3: 8

Median in G3: 9

Yes they are different, but because of unbalanced data we cannot infer that their average value must be different

So we went to test our hypothesis using p-value for hypothesis stating both have the same average grade. And we found out that the p-value is less than 5%, so it's not due to randomness, so we can infer that they have different average grades.

It's pretty great to motivate your children to seek high school. It's pretty good motivation.

Cleaning and feature engineering

Although, data was so clean, we did an extra cleaning for modeling.

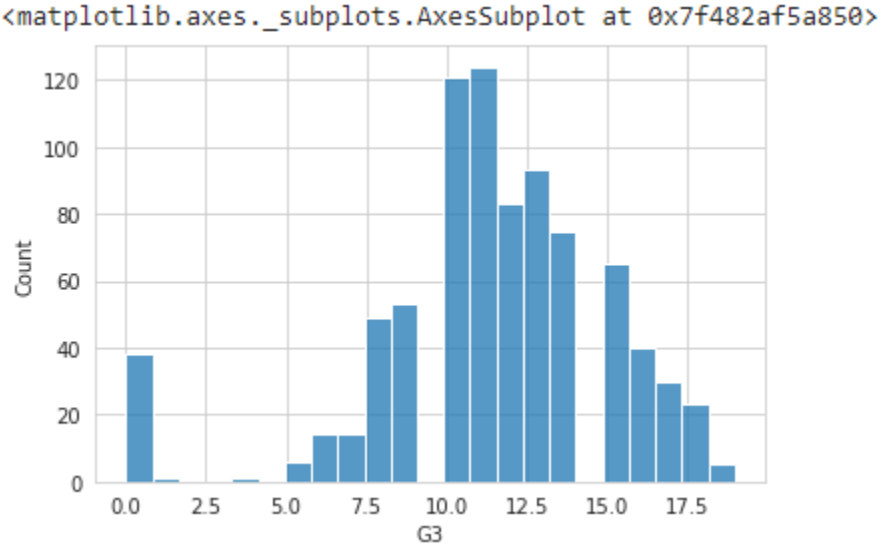
1. We dropped the school column to generalize our result to all schools

2. We did One hot encoding for
    a. Sex (male or female)
    b. Address (rural or urban)
    c. famsize(greater or less than 3)
    d. Pstatus(are parents apart or not)
    e. reason(the reason they got to that school?)
    f. Guardian (who is the guardian to the student?)
    g. schoolsup(does a student get support from school?)
    h. famsup(does a student get support from family?)
    i. Paid: (is student paying for the extra course)
    j. activities:(does student do extra activities?)
    k. nursery:(does the school have a nursery?)
    l. higher: (does the student want to go to high school)
    m. Internet: (does the student have internet access?)
    n. romantic:(is the student involved in a romantic relationship?)

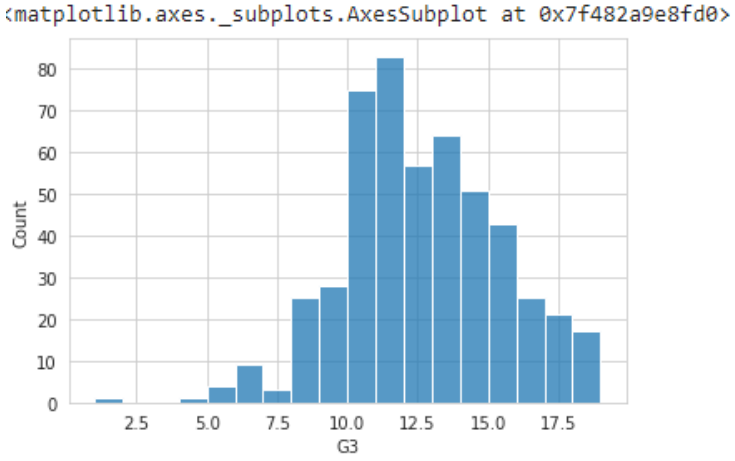
3. Medu and Fedu, we encoded them using ordinals but in a different way, we put different scales for different school levels, as if we didn't do that, the model will see the difference between primary and secondary as same as secondary and high school, which is not the case.

4. As noted from our EDA, we created some new meaningful features like
    a. both_at_home: are both parents at home?
    b. only_one_working: is there only one parent at home?
    c. both_working: are both parents working?
    d. And then create one hot encoder to check if any is working as a teacher or not.

- We found out that the target variable is perfectly normally distributed so we didn't need to do any transformation.

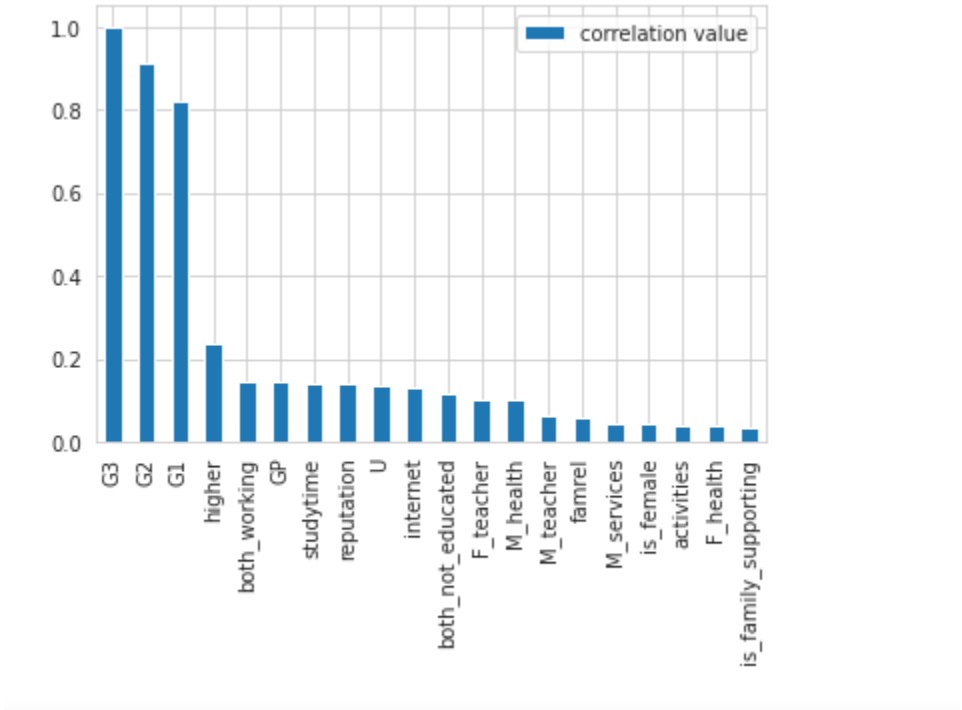


But there was some outlier on the left, so we deleted them using z-score. To be like that

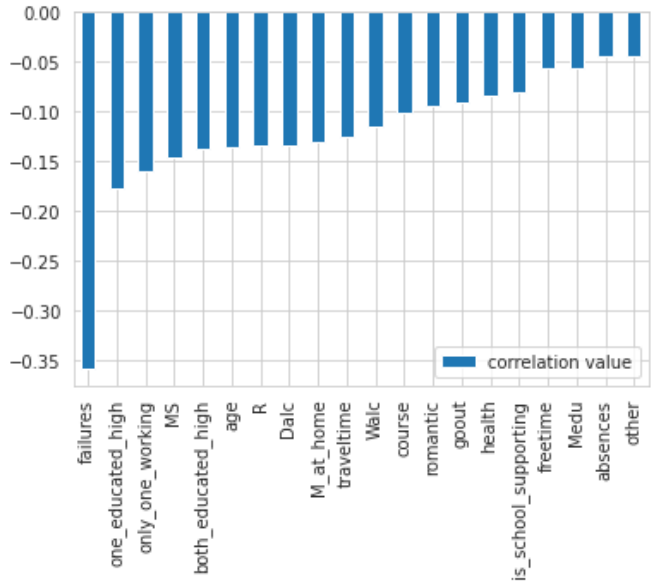


We looked at their correlation matrix, to drop the least correlated variables.

And found out the most important variables are:
1- positive correlation

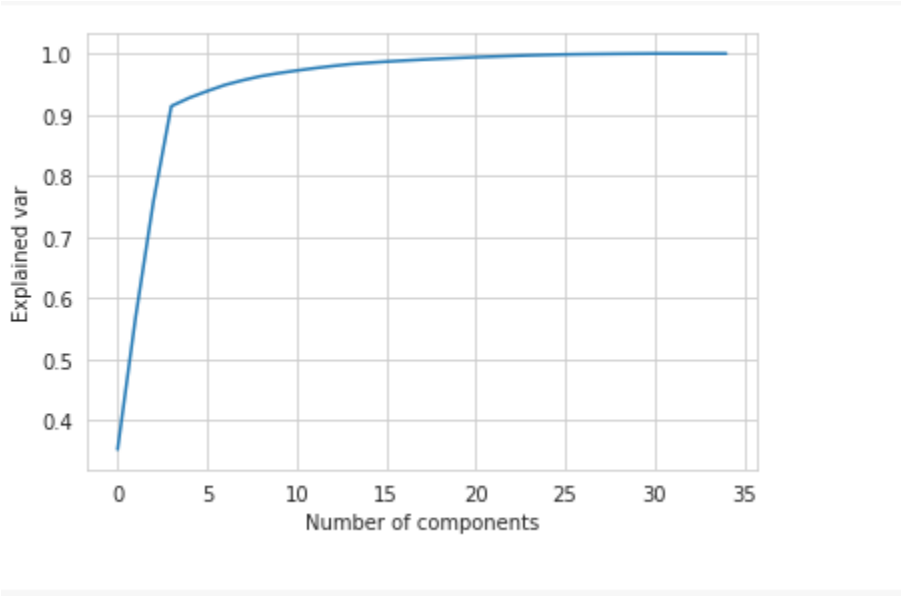


2- negative correlation



We dropped the uncorrelated columns, but we kept the original data frame also to try our model on it.

- We also tried to use PCA, to decrease the number of attributes. And we found out that about 35 attributes can describe the variance in our target variable



We let PCA choose the top 35 columns for us.

- Discretization

We found many notebooks online doing the same method, we and did the same fo compare our result to them

Grades between 0 and 9	0
Grades between 10 and 14	1
Grades between 15 and 20	2

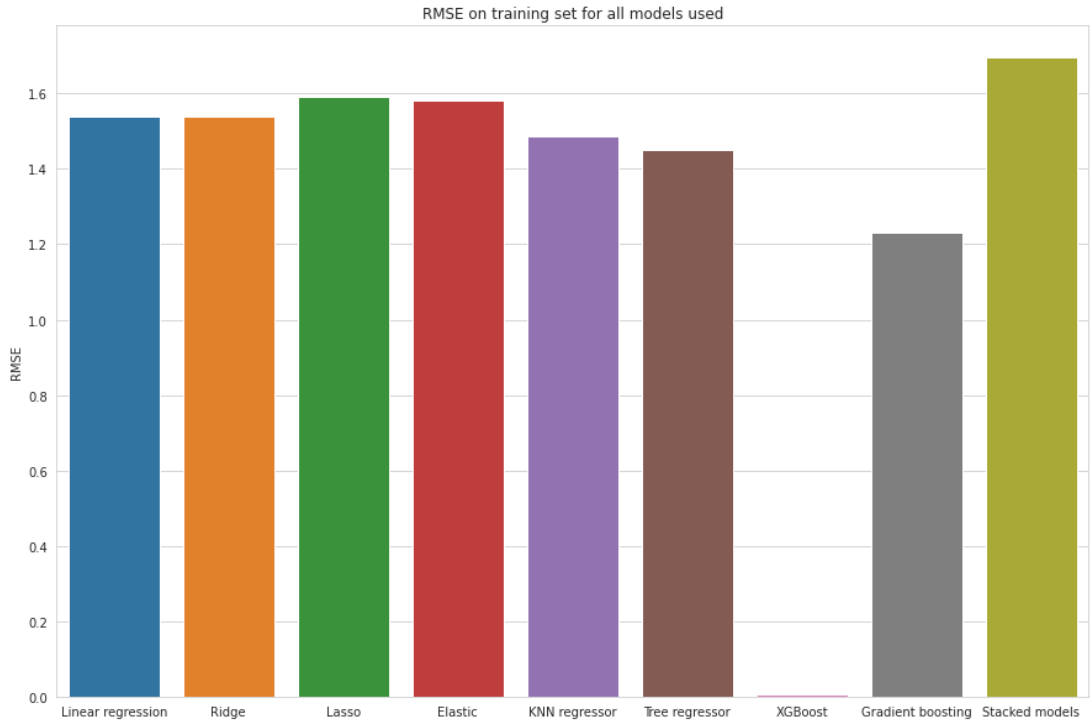
Modeling

We did many models to see which one will perform better. We firstly did our modeling on our raw data frame, then tried only columns produced by PCA, and tried data frame after dropping all uncorrelated columns.

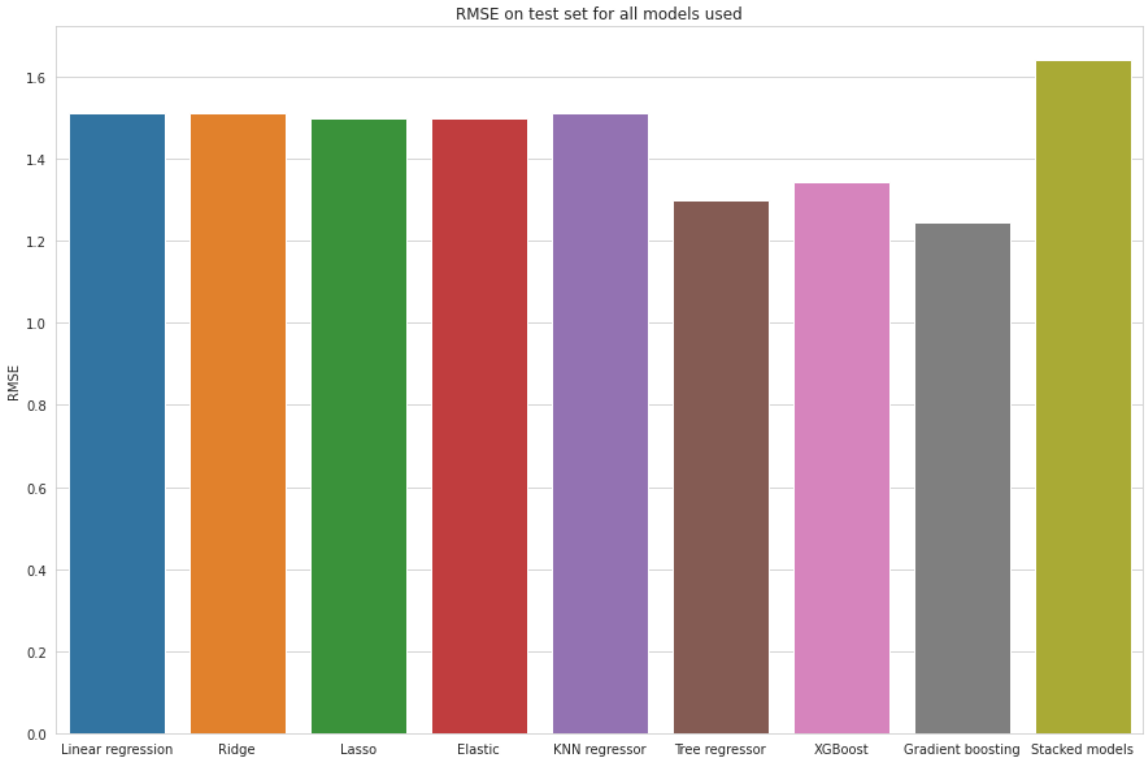
1. Models for predicting G3, given G1, and G2.

We used many models here

1. Linear regression
2. Ridge regression
3. Lasso regression
4. Elastic regression
5. KNN regressor
6. Tree regressor
7. XGBoost
8. Gradient boosting
9. Stacked models



And here is the root mean square error for all of our algorithms on the training set.



And here root means square error on the test set.

Notes.

- Linear, ridge, lasso, and elastic have the worst performance of all them
- XGBoost is probably overfitting the data as train error is much less than test error
- Gradient boosting is the most reasonable one, as they produce the least test error and not overfitting
- I was assuming that stacked models will outperform all of them, but it turns out it produces high RMSE in train and test.

Extra notes:

1. For ridge regression, it did its best for alpha =0.59
2. For Lasso, it did its best for alpha= 0.19 and fit_intercept is equal to True
3. For elastic it did its best for alpha equal to 0.16 and fit_intercept equal to True, and l1_ratio: equal to 0.29
4.

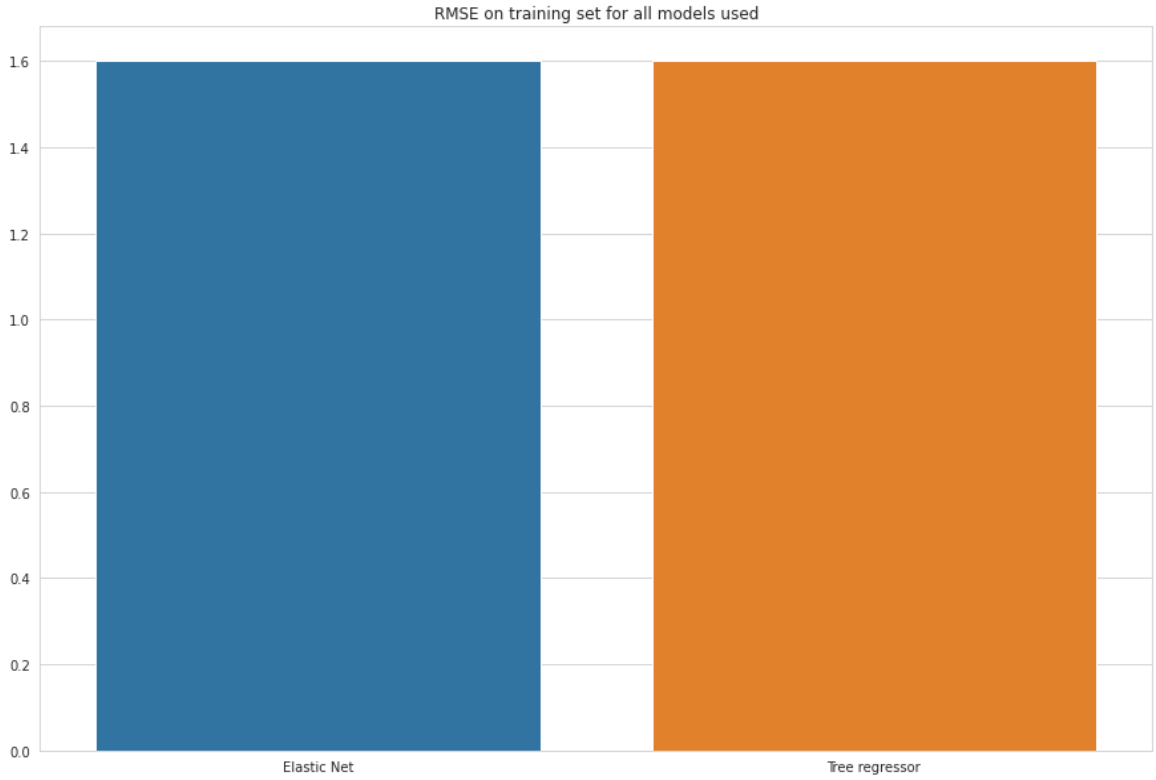
2. Predicting G3 given G1, and G2 on more clean data.

Here we dropped the most uncorrelated column and did our modeling again.

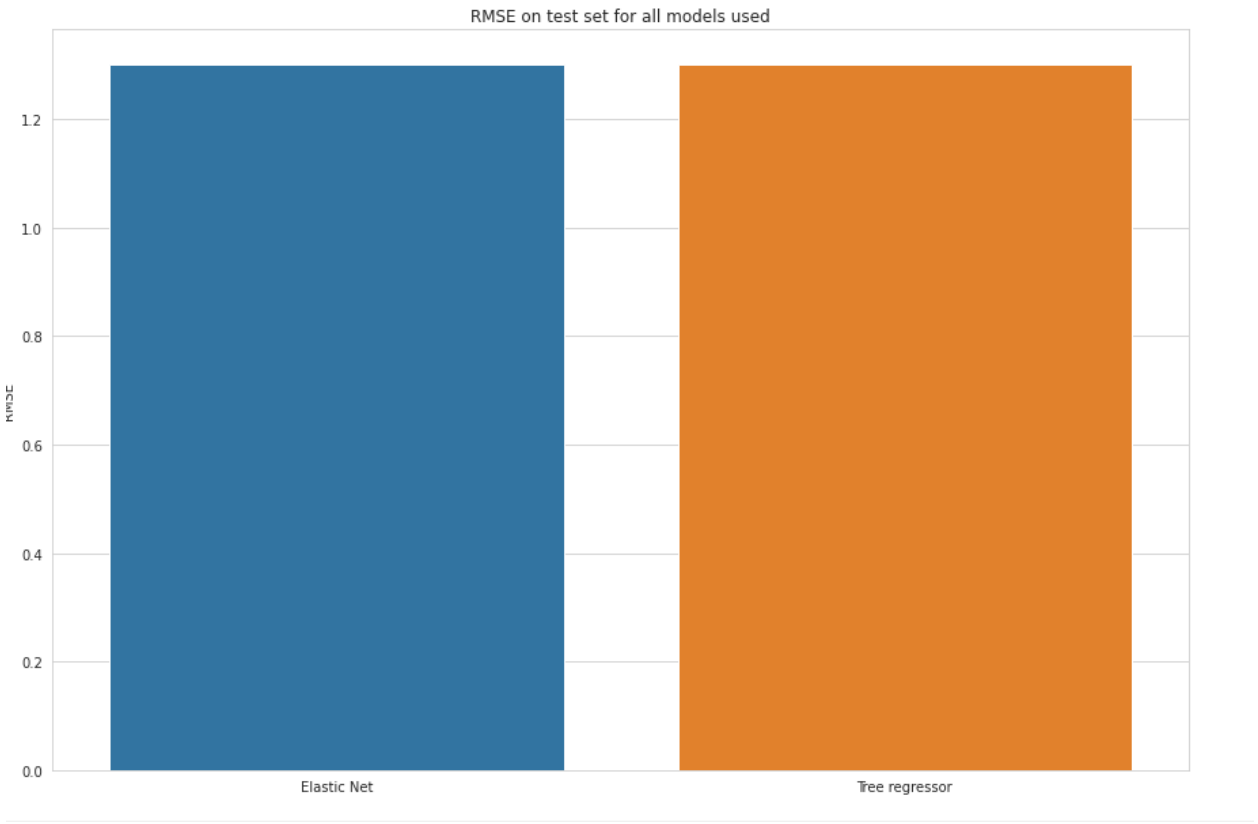
We tried

1. Tree regression
2. Elastic net

And here is the summary


RMSE on training set for all models used
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Elastic Net
Tree regressor

And on the test set



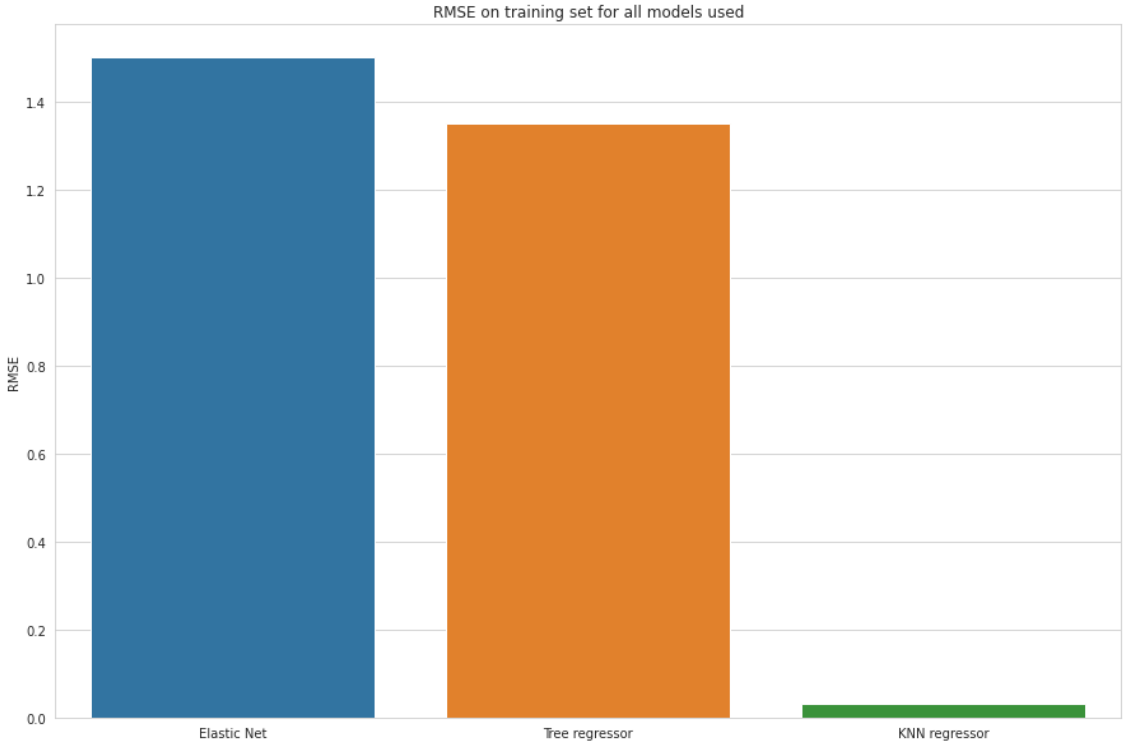
It performed worse than using the original dataset having all columns

3. Predicting G3 given G1 and G2 after choosing columns using PCA

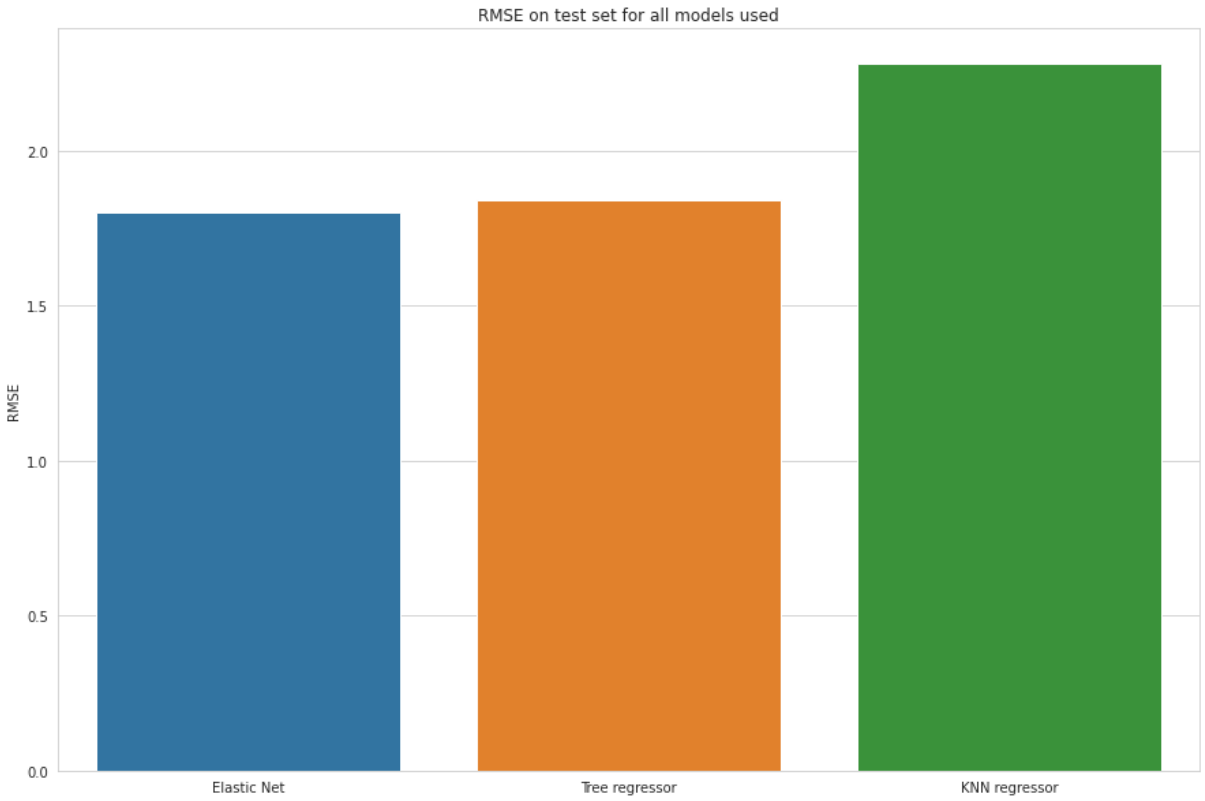
We tried

1. Elastic net
2. Decision tree
3. Knn regressor

And here is the summary



And on test set



We see here, that it was also worse than using raw data frame

Comparison

We will compare here our results to some notebooks found on Github

1. Student performance analysis
    - He got an RMSE of 1.54 in mathematics class
    - He got an RMSE of 1.26 in Portuguese class
    - Then the average in any class would be 1.4. And we outperformed him when using gradient boosting, as it was producing 1.34 in test error and 1.27 in train error.

    He also used a Neural network but the results were worse, about 4.4 in test error.

2. Comparison-of-Regression-Machine-Learning-Algorithms-of-Explaining-Students-Academic-Performance

I'm quoting from his notebook:

Results summary for Math.

	ML Technique	CV Score	Test Score	Parameters
1	Gradient Boosting Machines	0.26	0.26	max_depth=2, n_nestimators=100
2	Ridge Regression	0.10	0.20	alpha=100
3	SVR	0.22	0.18	kernel =rbf, gamma=0.1, C=10
4	Lasso Regression	0.10	0.18	alpha=0.1
5	Random Forest	0.30	0.17	max_depth=18
6	Decision Tree	0.22	0.15	max_depth=2
7	KNN	0.18	0.14	n_neighbors=5

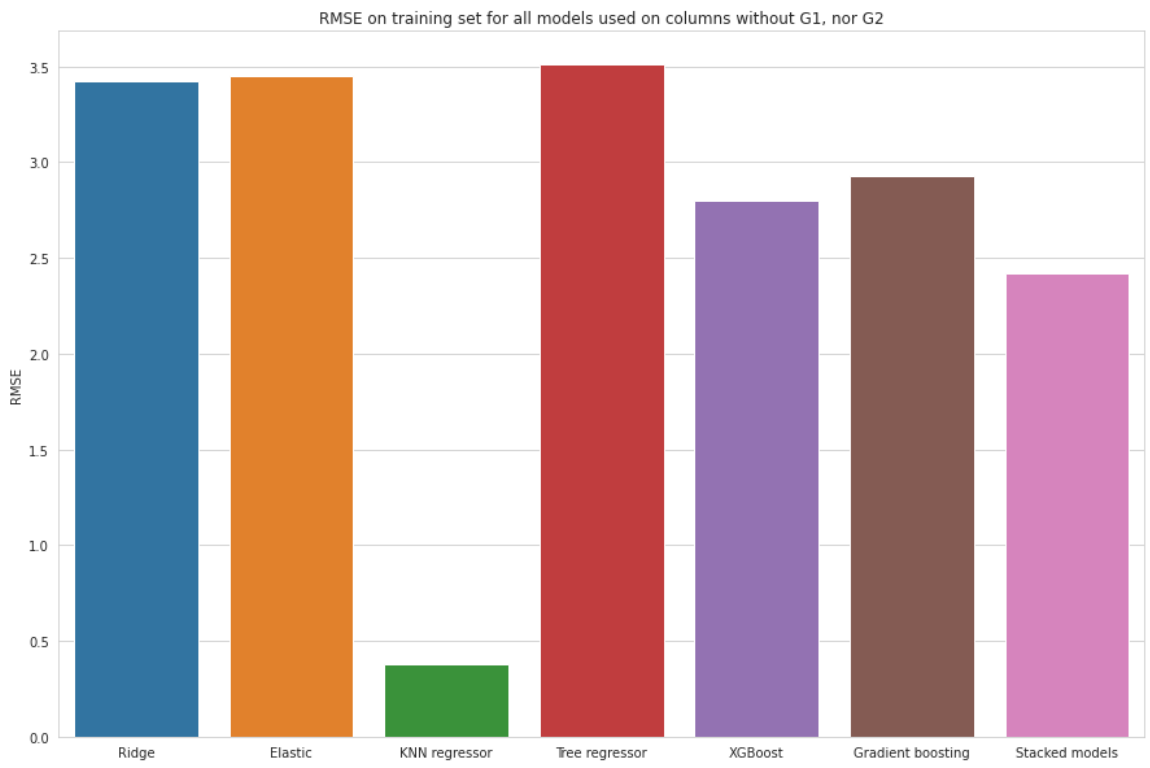
Results summary for Portuguese.

	ML Technique	CV Score	Test Score	Parameters
1	Random Forest	0.28	0.25	max_depth=13
2	Ridge Regression	0.27	0.25	alpha=0.1
3	SVR	0.21	0.23	kernel =rbf, gamma=0.001, C=10
4	Gradient Boosting Machines	0.31	0.21	max_depth=1, n_nestimators=100
5	KNN	0.15	0.20	n_neighbors=22
6	Lasso Regression	0.10	0.18	alpha=0.1
7	Decision Tree	0.19	0.15	max_depth=1

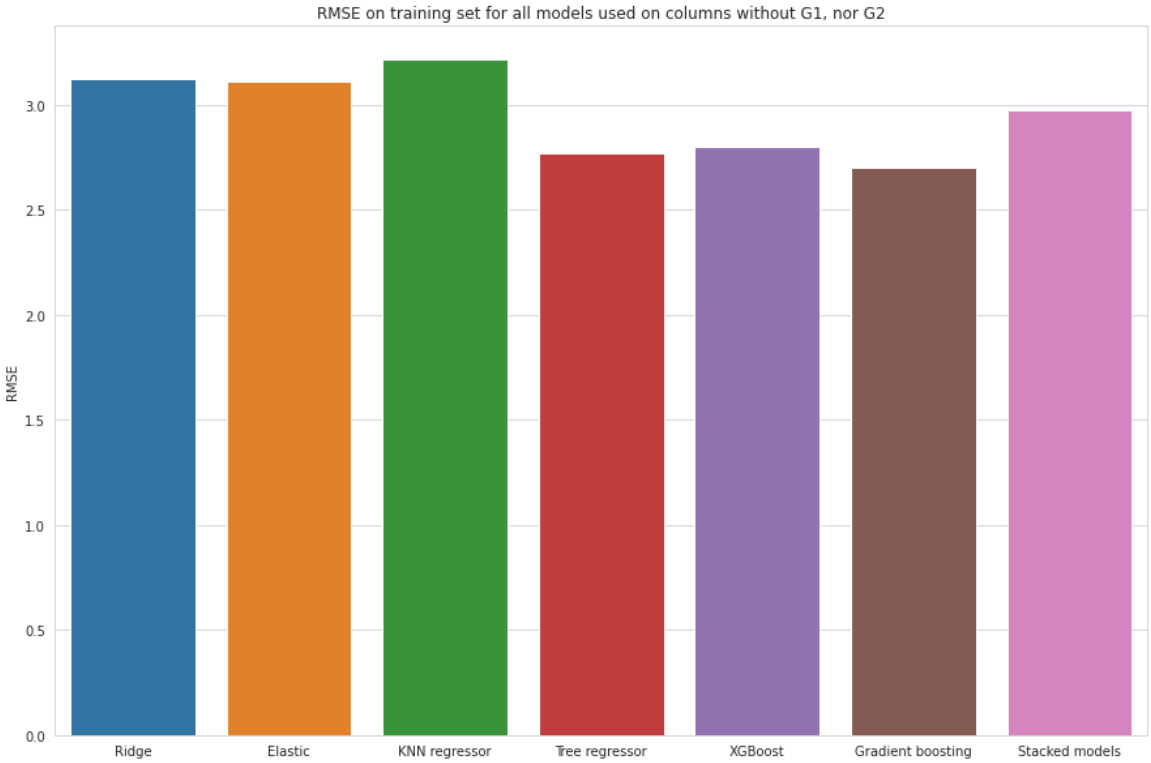
She has an average CV score 0.2, which we outperformed her

Models for predicting G3 not given G1, nor G2

Here we tried to predict the final grade, but we didn't include G1, not G2 column. We used multiple algorithms here. And here's the summary



And on test set:



We also find here that gradient boosting, and XGBoost are the best ones, with RMSE for training set 2.55 and test set 3.56 for gradient boosting, and for XGBoost, it got 3.49 for testing sample.

It’s a little worse than the last model. But sacrificed G1 and G2 columns which are the most important columns.

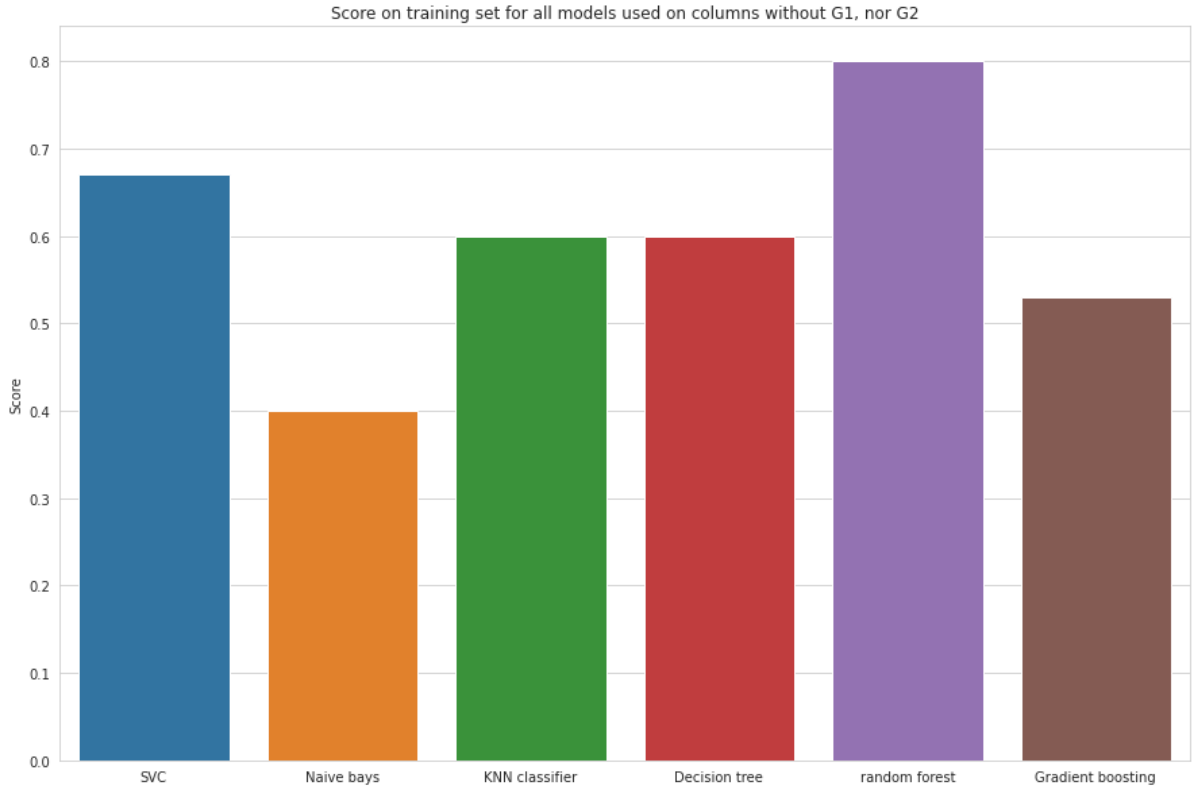
Classification

We tried to discretize our target variable and then classify it.
Final grade between 0-9 Bad
Final grade between 10-14 Good
Final grade between 15-20 Very good

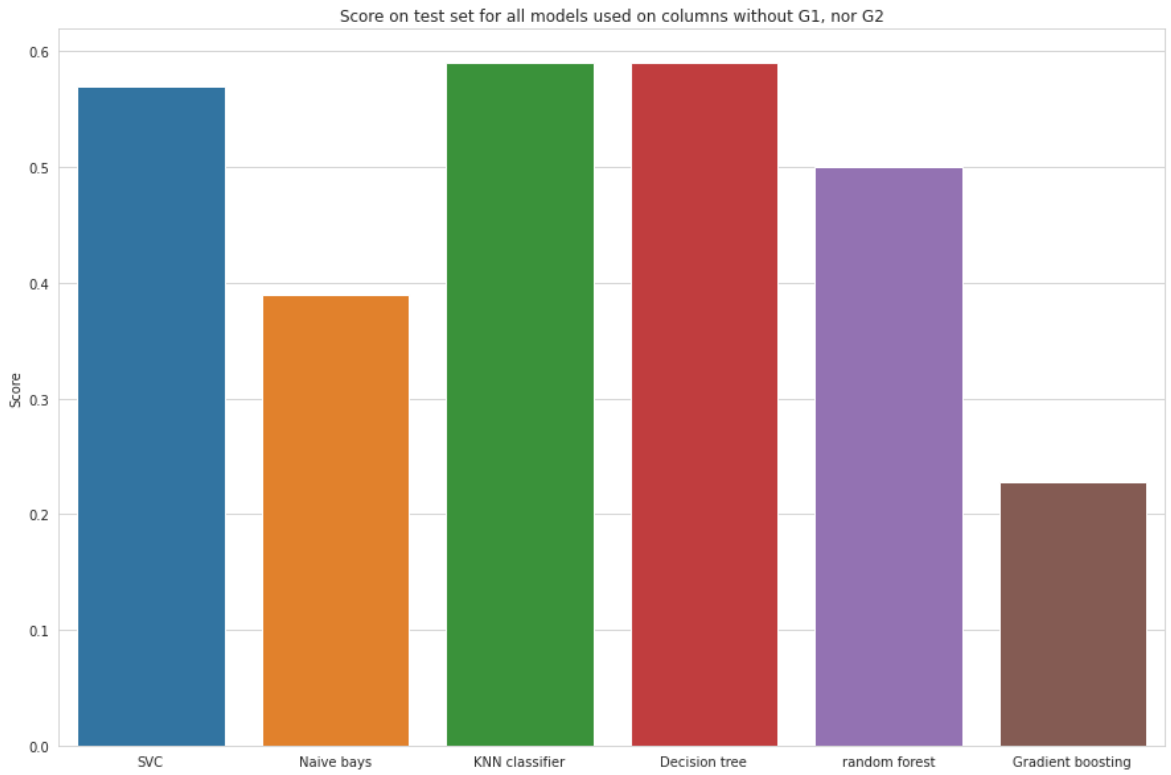
We tried many classification algorithms including

1. SVC
2. Naive bayes
3. Knn classifier
4. Decision tree classifier
5. Random forest
6. Gradient boosting classifier

And here is a summary for results



And on test set:



It’s found that out regression models were getting better score that classification models.

But the best score we got so far is Random forest

Other models

Also, we used this interesting data to predict extra information about parent state, mother education, father education, and absence of students

1) For parent state:
   1.1) we used decision tree that gives a score 86.6%
   1.2) Meanwhile KNN classifier that gives a score 88.5%
2) For mother education:
   2.1) decision tree bring a score of 48.8%
3) For father education:
   3.1) decision tree gives a score of 31.1%
4) For absence of student:
   4.1) using decision regressor, we got root mean square error 5.5
   4.2) using grid search (elastic net), we got root mean square error 5.45

Tools and libraries

1.) Numpy: is a fast and versatile library that enables a wide variety of mathematical operations on arrays.
2.) Scipy: is a library used for solving mathematical, scientific, engineering, and technical problems. We used it because it provides powerful statistical methods. As we are interested in calculating the p-value for null hypotheses.
3.) Math: is a module so useful for number theory. We installed this to be able to use functions (ex. Sqrt and log2)
4.) Collections: are containers that are used for storing a collection of data.
5.) Pandas: is a library that is used for data manipulation and analysis. Pandas is used for providing data frames for the data.
6.) Sklearn: is a very important library in machine learning generally that provides different machine learning models in regression, and classification. Also, it provides different preprocessing tools.
7.) Matplotlib: is a library that is used for visualizations and to present different types of plots
8.) Seaborn: is a library used for plotting statistical graphics.
9.) Plotly.express: this library is used for expressing graphs and plots.
10.) Statsmodels.api: is a module that has classes and functions associated with statistics.
11.) Xgboost: is a gradient boosting library that provides parallel boosting algorithms.

Conclusion

Helping schools to know what is better for their students, was a major challenge for us in that research. But we also focused on tips for parents that will help them raise a good student. But we strongly believe that grades are not everything in life; even a bad student at grades is certainly intelligent in some other field.

In conclusion, here are our final results found by our EDA.

1. There is no major difference between being a male and female in education. And it's all due to effort.
2. As the age increase, the performance of students decreases
3. For the student who failed at least once, their average grades is more inverse proportional to the number of failures.
4. We found out that students who have their parents apart have the same average grade same as a student who has their parents together
5. We found that students having one of their parents as a teacher, are having fewer failures, and they fail at most once.
6. If school is near to their hometown, it's much better for students, it helps them to get higher grades.
7. No major difference between living in urban or rural areas at all.
8. We found as parents get higher education, their parents get higher grades on average.
9. Women are better guardians than men regarding their child's grades.
10. Family size does not affect student performance
11. Students might need to study on aver for more than 2.5 hours to get grades greater than 19 out of 20, but it's also noted it's based on their abilities, some might need less than 2 hours, and up to five hours to get the same grade
12. It's pretty clear that family support is important for student spirit.
13. Half of the students who failed once are having extra activities. So it might put some extra burden on students.
14. Grades are affected by romantic relationships, in negative correlation.
15. Alcohol consumption decreases student performance.

16. Wanting to go to high school is a pretty good motivation for students to get high school.

Summary tips for parents

1. Don’t make your child have alcohol
2. Don’t make your child have romantic relationships
3. It’s better to have both parents as teachers, but the value is more important, so if parents are more familiar with their child's education system that shall help their child a lot.
4. It’s much better to register your child to a nearby school
5. If parents are having a great family issue, they don’t have to worry about their child's performance in school, as it’s not affected by their divorce.
6. It’s a pretty good motivation to motivate your child to go to high school.
7. To get your child to get high grades, you need to understand them, you need to make a custom number of hours to study to satisfy his/her needs
8.

Limitation

1. The data set was concentrated on Portuguese schools, we assume our models will have different values for test error when used with different cultures.

2. Columns were encoded in ordinal. We were hoping we did it ourselves, to see which is better for our models. And I can not know if any other encoders will behave differently or not.

3. Website didn’t provide which student they collected their data. And we wanted to make sure how randomly students were selected. And what criterion did they use, was it stratified or what? Does the percentage of each category of students in our dataset is really representative of the real distribution? There was not enough information about that. And that might affect our model in different regions

Thank You