

Main factors in predicting Student Performance.

A data Analysis Exploration by Ronald Beltrán

Table of Contents

Abstract

Keywords

Performance, Productivity, Correlation, Education, Causes.

Introduction

Data

For the purpose of this project, I researched on a variety of websites and found a medium-sized dataset from a highschool study in Portugal. I also found similar datasets from Turkey and India, however, given that I needed to compare my analysis to existing literature, I eventually decided to only use the dataset from Portugal.

Description

The dataset was named “Student Performance Dataset”, it can be found in diverse platforms of datasets like Kaggle or UCI ML Repository. It contain around 1044 observations in total, and it comes from a highschool in Portugal. It has a total of 32 features, between numerical and categorical.

This is the dataset's official description from the database website.

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

- 1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- 2 sex - student's sex (binary: 'F' - female or 'M' - male)
- 3 age - student's age (numeric: from 15 to 22)

- 4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
- 5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- 6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- 7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - "5th to 9th grade, 3 - "secondary education or 4 - "higher education)
- 8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - "5th to 9th grade, 3 - "secondary education or 4 - "higher education)
- 9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- 10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- 11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- 12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- 13 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)
- 14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- 15 failures - number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
- 16 schoolsup - extra educational support (binary: yes or no)
- 17 famsup - family educational support (binary: yes or no)
- 18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- 19 activities - extra-curricular activities (binary: yes or no)
- 20 nursery - attended nursery school (binary: yes or no)
- 21 higher - wants to take higher education (binary: yes or no)
- 22 internet - Internet access at home (binary: yes or no)
- 23 romantic - with a romantic relationship (binary: yes or no)
- 24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- 25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- 26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- 27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 29 health - current health status (numeric: from 1 - very bad to 5 - very good)
- 30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:

31 G1 - first period grade (numeric: from 0 to 20)

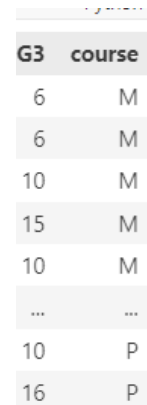
31 G2 - second period grade (numeric: from 0 to 20)

32 G3 - final grade (numeric: from 0 to 20, output target)

As we can see, the data contains different features related to students, including: Grades, Demographics, Social, Enviornmental, etc. As the aim of the study is to be a holistic analysis of factors affecting student performance, this is something that I desire in the dataset.

Cleaning

With regards to the data cleaning, I encountered, first of all, that the dataset from Portugal came in 2 separated .csv files -which is also explained in the official website-, one file contained the data from a math course and the other from a portuguese course. I merged them together, which in the process produced anew column that was named “_merge”, I eventually used this column as an indicator for the course, replacing each value according to which of the datasets did the row come from and renaming the column “course”, which is now a categorical variable.



G3	course
6	M
6	M
10	M
15	M
10	M
...	...
10	P
16	P

Image 1. The “course” column.

Additionally, when interacting with the data, I found that there was an outlier in the age column –there was a 22 year-old in a 17 year-olds data collection–, I decided to drop this observation (this individual’s data) from the study, as an outlier can affect the model and eventually the predictions of itself. It can also give very different information about the variance.

	school	sex	age	address	famsize	Pstatus	Medu	\
count	1044	1044	1044.000000	1044	1044	1044	1044.000000	
unique	2	2	NaN	2	2	2	NaN	
top	GP	F	NaN	U	GT3	T	NaN	
freq	772	591	NaN	759	738	923	NaN	
mean	NaN	NaN	16.726054	NaN	NaN	NaN	2.603448	
std	NaN	NaN	1.239975	NaN	NaN	NaN	1.124907	
min	NaN	NaN	15.000000	NaN	NaN	NaN	0.000000	
25%	NaN	NaN	16.000000	NaN	NaN	NaN	2.000000	
50%	NaN	NaN	17.000000	NaN	NaN	NaN	3.000000	
75%	NaN	NaN	18.000000	NaN	NaN	NaN	4.000000	
max	NaN	NaN	22.000000	NaN	NaN	NaN	4.000000	

Image 2. The outlier in age.

As a way to improve my efficiency, minimize the possibility of collinearity, and extracting the most out of the model, I dropped the G1 and G2 column (the columns from which G3 is computed from). As they are highly correlated with G3, which is the target variable of this study (the variable I am trying to predict). While doing this, I also double checked and dropped any possible Null-value rows that could be in the dataset.

```

editor, I can drop the rest of the grades, as they
nd can diminish the model value to extract importar

lDataframe.drop(columns = ['G1', 'G2']).dropna()

```

Image 3. Dropping G1, G2, and Nulls

After the data cleaning, I ended up with this preview of the dataset. I can't see all the features because it is truncated. But I find it appropriate to put the image of it in the report.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G3	course
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	no	4	3	4	1	1	3	6	6	M
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	no	5	3	3	1	1	3	4	6	M
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	no	4	3	2	2	3	3	10	10	M
3	GP	F	15	U	GT3	T	4	2	health	services	...	yes	3	2	2	1	1	5	2	15	M
4	GP	F	16	U	GT3	T	3	3	other	other	...	no	4	3	2	1	2	5	4	10	M
...
1039	MS	F	19	R	GT3	T	2	3	services	other	...	no	5	4	2	1	2	5	4	10	P
1040	MS	F	18	U	LE3	T	3	1	teacher	services	...	no	4	3	4	1	1	1	4	16	P
1041	MS	F	18	U	GT3	T	1	1	other	other	...	no	1	1	1	1	1	5	6	9	P
1042	MS	M	17	U	LE3	T	3	1	services	services	...	no	2	4	5	3	4	2	6	10	P
1043	MS	M	18	R	LE3	T	3	2	services	other	...	no	4	4	1	3	4	5	4	11	P

1044 rows × 32 columns

Image 4. Cleaned Dataset Preview.

Descriptive Statistics

To begin formally describing the data, I plotted every numerical value as a boxplot, to see what information can I extract from it.

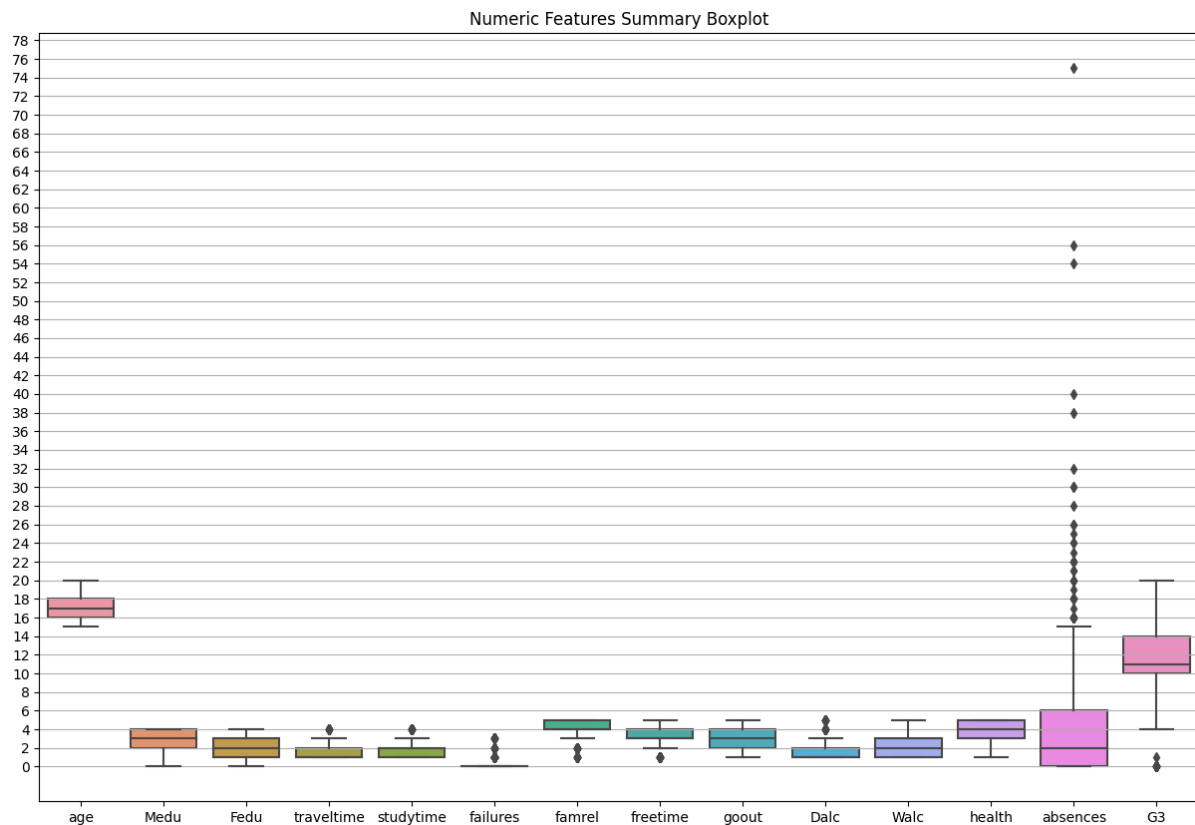


Image 5. Numeric Features Summary Boxplot

As a starting point, we can see that:

- This is a study of secondary educational level, the mean age is 17.
- On average, formal education of the mothers is higher than of fathers
- Travel and Study time are both, on average, low (these variables are ordinal ranging from 1 to 4)
- The number of failures is extremely low, on average the students have 0 failures, with a relatively few outliers that have, at most, 3 failures.
- Family relationships seem to be, on average, fairly good. (However I'm prone to believe this is self-reported information, or something equivalent). With some exceptions with a 1 at its lowest (this is an ordinal feature ranging from 1 to 5).
- Free time and going out with friends is also relatively good qualified, on average. (These are ordinal values ranging from 1 to 5).
- Alcohol consumption is higher on weekends, naturally enough, and there are a couple of outliers that consume a high amount of alcohol on weekdays as well (ordinal value from 1 to 5).

- Taking into account that “absences” is a numerical value and its maximum is 93, the number of absences is low on average, with some outliers that have at most 75
- The maximum grade is 20. The mean is above the threshold for passing the class, but there are some outliers that got very bad final grades.

Then I branched into describing the data more thoughtfully.

The “age” distribution is fairly normal, as we could infer from a uniform boxplot in the [Image 5], I used a histogram to see the “sex” distribution as well, and it seems to be also uniformly distributed accross all ages, however there are more women in this study.

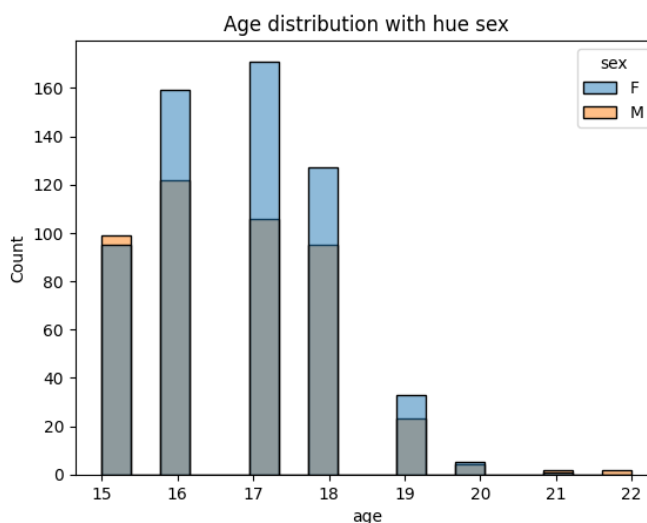


Image 6. Age distribution with hue sex

The distribution of grades is approximately normal, we can see the outliers that were visualized in the boxplot, apparently there are sound 53 of them who got almost a 0. This is a factor that increases the variance of the grades.

Mean 11.356111645813282
Std 3.8650450088046613
Variance 14.938572920085823
Skewness -0.9952493928298063

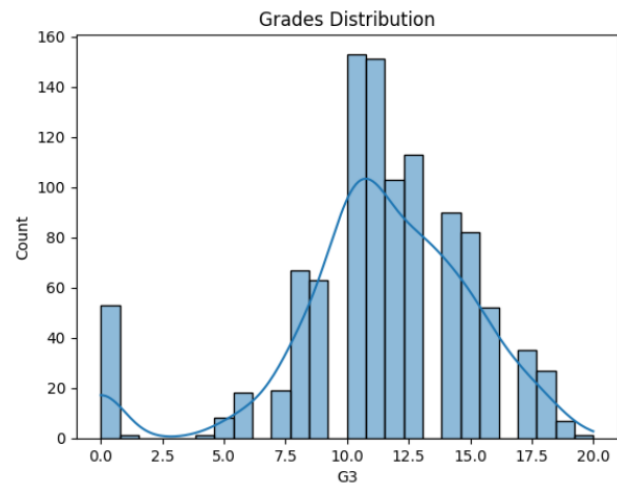


Image 7. Grades Distribution

I also used a heatmap to visualize the correlation between all the variables. This is one of the most important graphs of the notebook because it gives a guide on how to see the most possible most influential factors in determining "G3".



Image 8. Correlation Heatmap

The blue-ish the cell color, the more correlated the variables are. Negative correlation for red-ish cell colors. From this heatmap, we can say that:

- Parents education levels are correlated between each other.
- G3 is negatively correlated with the number of failures (which makes total sense)
- Weekend alcohol consumption is correlated with weekday consumption.
- “Free time” and “going out with friends” features are also somewhat correlated.

As a side note. When reading a heatmap, it could be easier just to read either side of the diagonal, as it is symmetric accross it.

I also visualized the relationship between several categorical variables and the final grades

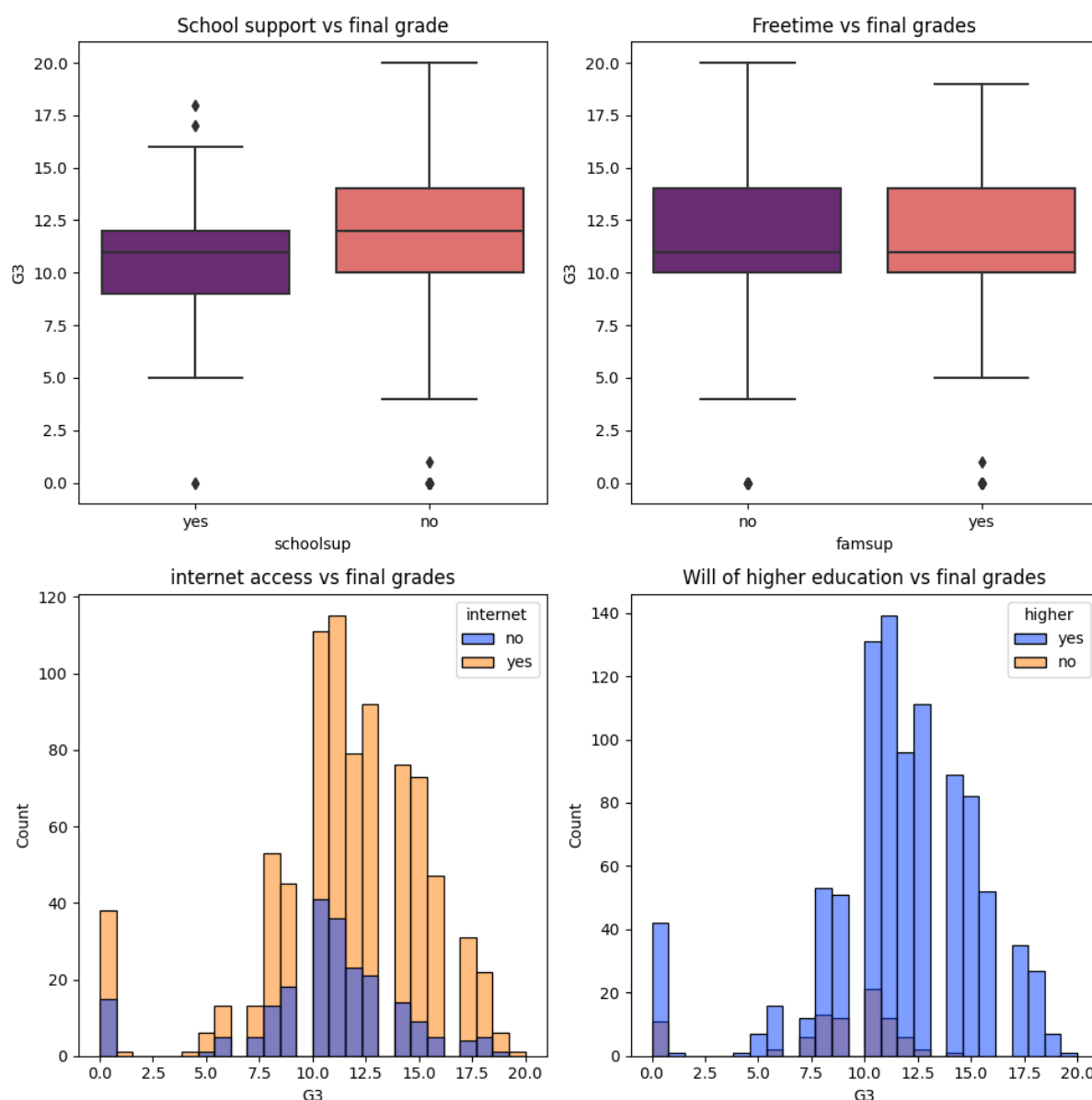


Image 9. Binary Variables Relationships with Final Grades

After plotting these graphs, there are some things to point out:

- Plot 1: School support somewhat negatively correlated with final grades, this makes sense because the more a student needs improvement, it its more likely to seek out school support. The grades of those that don't have educational support are more spread out.
- Plot 2: Surprisingly, family support doesn't correlate with positive grades, but the grades distribution for those who have family support are less spread out.
- Plot 3: Internet access means better grades, and there is more people with internet access than people without internet access.
- Plot 4: The students that want to continue to higher education have better grades, however they are very few. And around one student has relatively good grades but doesn't want to continue higher education.

Methodology

The method I am going to use to extract valuable insights from this dataset is a correlation analysis and a regression model. I already did the correlation analysis with the heatmap [Image 8]. Now I will do the preparation for the model with inferential statistics, to obtain information about the population taking my dataset as a sample of students.

I will obtain the confidence intervals of the following variables:

- The final Grade “G3”.
- The final grade of a subsample of students with each value of the “higher” variable. I will use these to compare the final grades averages of the students who want and who don’t want to continue to higher education.
- The final grade of a subsample of students with each value of the “going out” variable.

Then, I will perform statistical tests to know if the sample means of overall final grades of men, and grades of women, are different to their respective population means. In this case, as I’m dealing with a sample of 2 schools from Portugal, I will take as the population all of Portugal’s students.

Next I will perform hypothesis testing between 2 samples, for the “higher” variable, as I could see in the descriptive statistics section that there may be a significant difference between the groups, I will use this test to formalize that.

And to finish this part I will perform ANOVA testing between samples of the “goout” variable, to know if it is important to consider social interaction in the context of final grades performance.

With these tests and analyses, I expect to understand and formalize the effects of certain variables with student academic performance.

After the inferential statistics I will build the model. For this, I will use a multiple linear regression, with the predicted variable being “G3”. When done, I will interpret the results and draw conclusions.

Inferential Statistics

Confidence Intervals

I calculated the confidence intervals of several variables in the dataset:

- “G3”
- “G3” for each classification in the “higher” variable.
- “G3” for each classification in the “goout” variable.

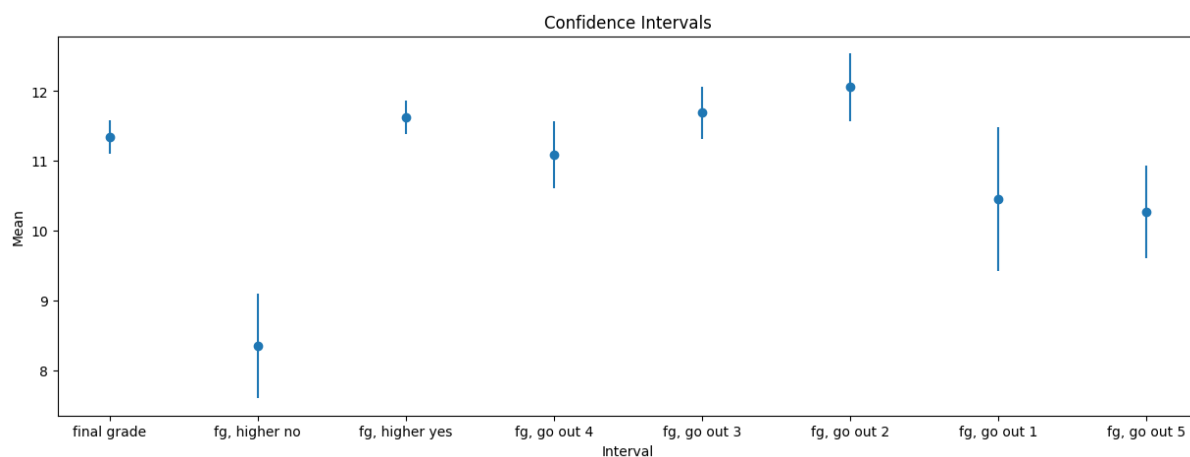


Image 10. Confidence intervals

From the plot, can already see some very different confidence intervals from the population with these characteristics. There seems to be a notable difference between the categories of “higher” variable.

- Inferential Statistics:
 - The students build confidence intervals for at least three variables **(15 points)**
 - Perform hypotheses testing for at least three variables (one sample tests) **(20 points)**.
 - In order to get the maximum score, students should accomplish the following:
 - The assumptions of the model are clearly stated.
 - Students check whether the main assumptions of the model are met
 - All the tests should be related to the main topic of the project.
 - Hypotheses are clearly stated
 - Results are satisfactorily compared with the literature.
 - All the results are presented either in a tabular or graphic form. All tables and graphs are properly numbered, labeled, and discussed. Students interpret results satisfactorily. Students support their findings with similar results from the literature. The contribution of the article must be clearly stated.

Model

As I mentioned in the previous section, I will be using a Random Forest Regressor model for predicting student outcome from this dataset.

Assumptions

Normality

The way I checked for normality in the variables that I obtained the confidence intervals of was by plotting a Q-Q Plot of the variables and doing a Shapiro test to formally say that the data is normal.

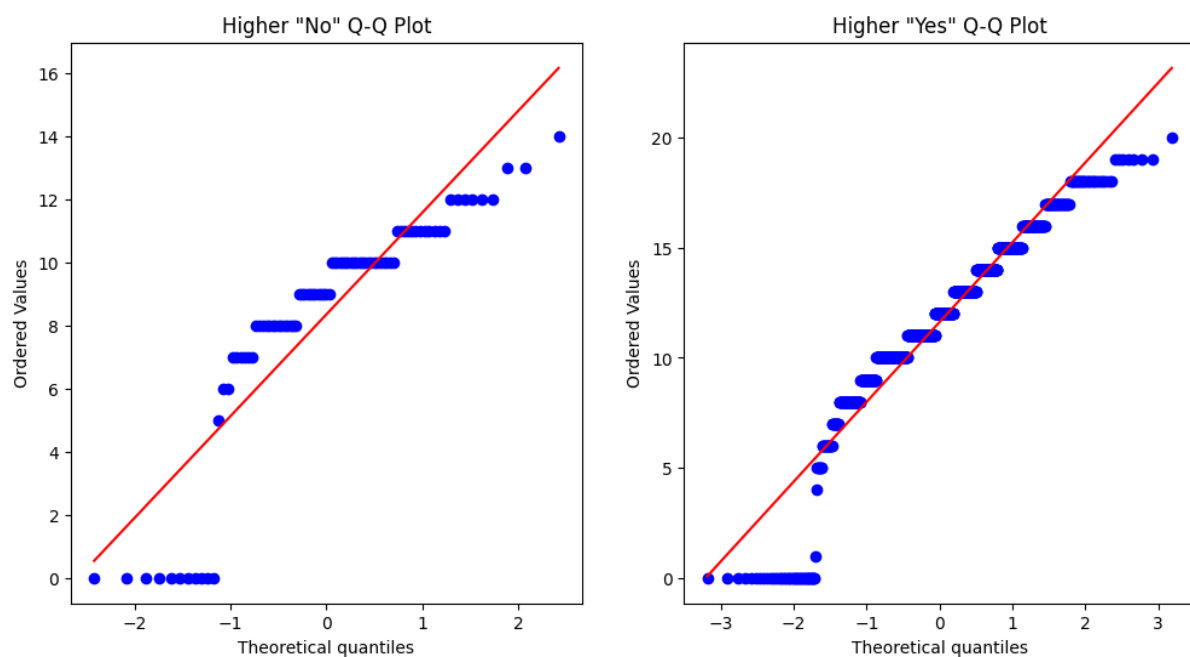


Image 11.1 The Q-Q plot to visualize the normality of the data samples of “higher”.

```
# Shapiro-Wilk Test is also used to check, but as a statistical test.
print(ss.shapiro(higher_no))
print(ss.shapiro(higher_yes))
```

✓ 0.3s

```
ShapiroResult(statistic=0.7977627515792847, pvalue=1.0373734093960252e-09)
ShapiroResult(statistic=0.9176795482635498, pvalue=2.2514292183399414e-22)
```

Image 11.2 The test of normality for “higher” variable with Shapiro-Wilk test.

As we can see in this plots and the Shapiro Tests, data from both samples is approximately normally distributed. Given that the statistic is very low, we can say that there was a very small difference between a normal distribution and the

sample. Even though the p-value rejected the null hypothesis of normality, the Q-Q Plots and the statistic proved that the data followed a normal distribution indeed. So I conclude this because I consider the small p-value a consequence of a large dataset.

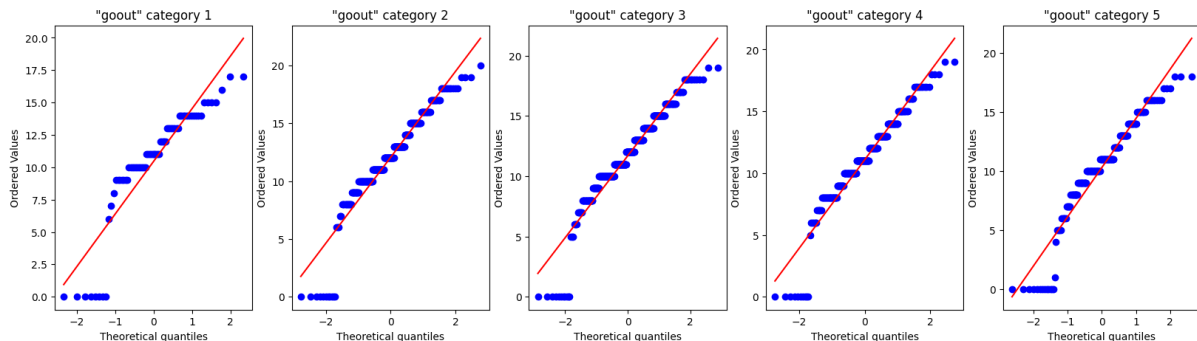


Image 12.1 The Q-Q Plots for the variable “goout”.

```
# Shapiro-Wilk Test is also used to check, but as a statistical test.
print(f'{ss.shapiro(goout_data)} \tfor # {category}')
```

✓ 0.7s

ShapiroResult(statistic=0.9306492209434509, pvalue=7.285465120077106e-09)	for # 4
ShapiroResult(statistic=0.9268427491188049, pvalue=9.595322800204364e-12)	for # 3
ShapiroResult(statistic=0.912743330001831, pvalue=7.45998471329834e-11)	for # 2
ShapiroResult(statistic=0.8306800127029419, pvalue=1.477290965112843e-07)	for # 1
ShapiroResult(statistic=0.920941174030304, pvalue=9.182016214026589e-08)	for # 5

Image 12.2 The Shapiro-Wilk test for normality.

The variable “goout” has similar results, the p-value is very small, however, taking into account that there is a lot of data, that the differences are small and that Q-Q Plots reflect a similar tendency to a normal distribution, then I conclude that all those samples are normally distributed.

Homogeneity of Variance

For this assumption I will do a Levene’s test for Homogeneity of Variance, which is a test that compares the mean variance across a set of samples.

```
LeveneResult(statistic=1.7086302918995935, pvalue=0.19145155131262367)
LeveneResult(statistic=1.1549038972216341, pvalue=0.3292684035563189)
```

Image 13. Results for the Lavene’s Test for Homogeneity of Variance.

In the [Image 12], the first result corresponds to the “higher” variable, while the second corresponds to the “goout” variable.

As we can see, the p-value indicates that there is not enough evidence to reject the null hypothesis that the data is homogeneous, that it is equal among the samples. Therefore, we conclude that homogeneity is met.

Hypotheses

One sample tests

Hypothesis Test 1

- Null Hypothesis: The average grade of the sample is the same as the population mean grade score (sample to portugal comparison).
- Alternative Hypothesis: The average grade of the sample is different than the population mean grade score (sample to portugal comparison).

T-statistic: -45.45406177653802, P-value: 3.765037059188488e-249

Image 14.1 Hypothesis Test 1 Result

We reject the null. This sample is different by -45.46 points. With a probability of this to happen by randomness of almost 0%

Hypothesis Test 2

- Null Hypothesis: The average grade for men in the sample is the same as the population mean grade score (sample to portugal comparison).
- Alternative Hypothesis: The average grade for men in the sample is different than the population mean grade score (sample to portugal comparison).

T-statistic: -30.661385504072605, P-value: 4.393463764292326e-112

Image 14.2 Hypothesis Test 2 Result

We reject the null. This sample is different by -30.66 points. With a probability of this to happen by randomness of almost 0%

Hypothesis Test 3

- Null Hypothesis: The average grade for women in the sample is the same as the population mean grade score (sample to portugal comparison).
- Alternative Hypothesis: The average grade for women in the sample is different than the population mean grade score (sample to portugal comparison).

T-statistic: -33.569671751947894, P-value: 7.03040735824691e-139

Image 14.3 Hypothesis Test 3 Result

We reject the null. This sample is different by -33.56 points. With a probability of this to happen by randomness of almost 0%

Hypothesis Test 4

- Null Hypothesis: The grades for students that want to take higher education are on average the same as of the students who don't want to.
- Alternative Hypothesis: The average grade of students that want to take higher education are on average different as of the students who don't want to.

```
T-statistic: 7.68185829714147
P-value: 3.619031073465375e-14
```

Image 14.4 Hypothesis Test 4 Result

We reject the Null Hypothesis, and as we found out in the boxplot from before, there is certainly a significant difference of the mean final grade for those who want to take higher education as opposed to those who don't want to. A difference of 7 more points.

Hypothesis Test 5

- Null Hypothesis: The grades for the students do not depend on how much do they go out with their friends.
- Alternative Hypothesis: The grades of the students depend on how much do they go out with their friends.

```
F-statistic: 6.993324259556865
P-value: 1.4834871220226462e-05
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.01
=====
group1 group2 meandiff p-adj lower upper reject
-----
1 2 1.606 0.0159 -0.0731 3.2851 False
1 3 1.258 0.0873 -0.372 2.888 False
1 4 0.6418 0.7307 -1.0537 2.3374 False
1 5 -0.1277 0.9993 -1.9041 1.6487 False
2 3 -0.348 0.8146 -1.3951 0.6991 False
2 4 -0.9642 0.0484 -2.1107 0.1823 False
2 5 -1.7337 0.0001 -2.9967 -0.4707 True
3 4 -0.6162 0.3323 -1.6895 0.4571 False
3 5 -1.3857 0.0016 -2.5827 -0.1888 True
4 5 -0.7695 0.2893 -2.0543 0.5153 False
-----
```

Image 14.5 Hypothesis Test 5 Result

After doing

Analysis of the Results

Interpretation of the most relevant Test

Conclusion

Limitations of the Analysis

Bibliography *Consistency in the format

Appendix *Not important enough material