Student Performance
Exploratory Analysis
TP1_NVR_TMM

Nick Romano, nromano@bellarmine.edu
Trevor McCormick, tmccormick2@bellarmine.edu

I. INTRODUCTION

Our data set came from the machine learning repository from UC Irvine. https://archive-beta.ics.uci.edu/ml/datasets/student+performance? It's a data set describing the performance of 662 students during their high school education. We were trying to analyze the grades of these students to find outliers, correlations, and differences in the various categories from the data set.

II. DATA SET DESCRIPTION

This data set contains 622 samples with 25 columns of various data types. A complete listing is shown in **Table 1**. In these tables we are indicating two things (nominal, ordinal, interval, or ratio) and the Pandas data type. In our dataset, we have 2 different Pandas data types, Object and Int64. We also have a few of each normal data type (nominal, ordinal, interval, and ratio), with ratio being the most common type. We managed to find a dataset that had exactly 0 pieces of missing data, which makes the whole set cleaner and easier to work with.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
School	Object, Nominal	0%
Sex	Object, Nominal	0%
Subject	Object, Nominal	0%
Age	Int64, Ratio	0%
Address	Object, Nominal	0%
Pstatus (Parental Marital Status)	Object, Ordinal	0%
Mother's education (Medu)	Int64, Ordinal	0%
Father's education (Medu)	Int64, Ordinal	0%
Travel time	Int64, Ordinal	0%
Study time	Int64, Ordinal	0%
Failures	Int64, Ratio	0%
Schoolup (School Support)	Object, Nominal	0%
Famsup (Family Support)	Object, Nominal	0%
Paid	Object, Nominal	0%
Activities	Object, Nominal	0%
Higher	Object, Nominal	0%
Internet	Object, Nominal	0%
Famrel (Relationship Quality)	Int64, Ordinal	0%
Free time	Int64, Ordinal	0%
Go out	Int64, Ordinal	0%
Dalc (Daily Alcohol)	Int64, Ordinal	0%
Walc (Weekend Alcohol)	Int64, Ordinal	0%
Health	Int64, Ordinal	0%
Absences	Int64, Ratio	0%
G1 (1st Period Grade)	Int64, Ratio	0%
G2 (2nd Period Grade)	Int64, Ratio	0%
G3 (Final Grade)	Int64, Ratio	0%

III. Data Set Summary Statistics

These tables show summaries of the student performance data. **Table 1** deals with statistics such as mean and standard deviation. **Table 2** shows the frequencies and proportions of the different categories in the dataset.

Table 2: Summary Statistics for Student Performance

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Age	662	16.8	1.26	15	16	17	18	22
Failures	662	0.33	.71	0	0	0	0	3
Absences	662	4.93	6.85	0	0	3	8	75
G1	662	10.72	3.08	3	8	10	13	19
G2	662	10.70	3.52	0	9	11	13	19
G3	662	10.72	4.10	0	9	11	13	20

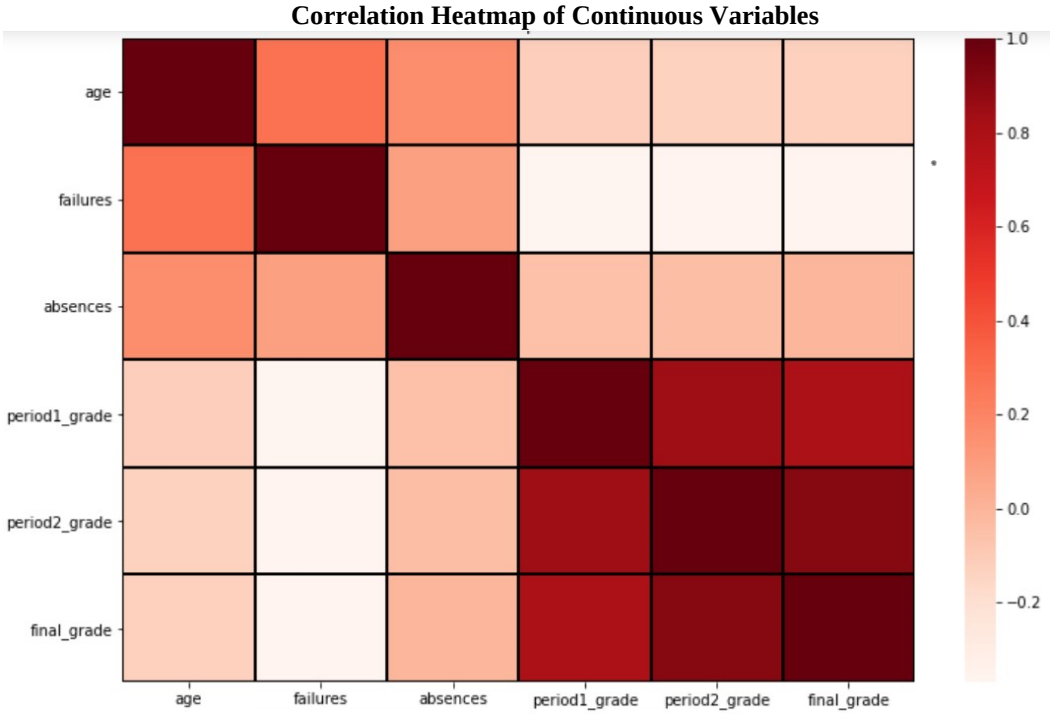
Table 3: Proportions for Student Performance

Category	*Frequency*	*Proportion (%)*
Subject:		
• Math	391	59%
• Portuguese	271	41%
School:		
• Gabriel Pereira	434	66%
• Mousinho da Silveira	228	44%
Sex:		
• Male	272	41%
• Female	390	59%
Address:		
• Rural	201	30%
• Urban	461	70%
Parent Status:		
• Living Apart	83	13%
• Living Together	579	87%
Mothers Education:		
• 0	6	1%
• 1	150	23%
• 2	190	29%
• 3	144	22%
• 4	172	26%
Fathers Education:		
• 0	7	1%
• 1	179	27%
• 2	216	33%
• 3	133	20%
• 4	127	19%
Travel Time:		
• 1	373	56%
• 2	220	33%
• 3	53	8%
• 4	16	2%

Study Time:		
• 1	217	33%
• 2	311	47%
• 3	99	15%
• 4	35	5%
School Support:		
• No	592	89%
• Yes	70	11%
Family Support:		
• No	262	40%
• Yes	400	60%
Paid:		
• No	469	71%
• Yes	193	29%
Activities:		
• No	343	52%
• Yes	319	48%
Higher:		
• No	71	11%
• Yes	591	89%
Internet:		
• No	158	24%
• Yes	504	76%
Family Relationship Quality:		
• 1	21	3%
• 2	29	4%
• 3	102	15%
• 4	328	50%
• 5	182	27%
Free Time:		
• 1	47	7%
• 2	110	17%
• 3	250	38%
• 4	184	28%
• 5	71	11%
Go Out:		
• 1	48	7%
• 2	149	23%
• 3	212	32%
• 4	147	22%
• 5	106	16%
Weekday Alcohol:		
• 1	460	69%
• 2	122	18%
• 3	45	7%
• 4	18	3%
• 5	17	3%
Weekend Alcohol:		
• 1	255	36%
• 2	147	22%
• 3	124	19%

• 4	90	14%
• 5	46	7%
Health:		
• 1	90	14%
• 2	78	12%
• 3	134	20%
• 4	110	17%
• 5	250	38%

Table 4: Correlation Table/Tables



IV. DATA SET GRAPHICAL EXPLORATION

With this dataset being the size that it is, there were many graphs we were able to make in order to show the data visually. The histogram (**Figure 1**) was very useful to see how the students did grade-wise over each period of the school year. You can begin to see slight decline in scores as the year goes on, which is expected. **Figure 2** shows the correlation between the different periods for the math classes. In these scatterplots, you can clearly see that there is a strong linear relationship between them. **Figure 3** was one that really stood out. The number of students who were planning on furthering their educations was far higher than I would have expected it to be. **Figure 4** and its results come as no surprise, as students consume more alcohol during the week, their academic performance starts to dwindle.

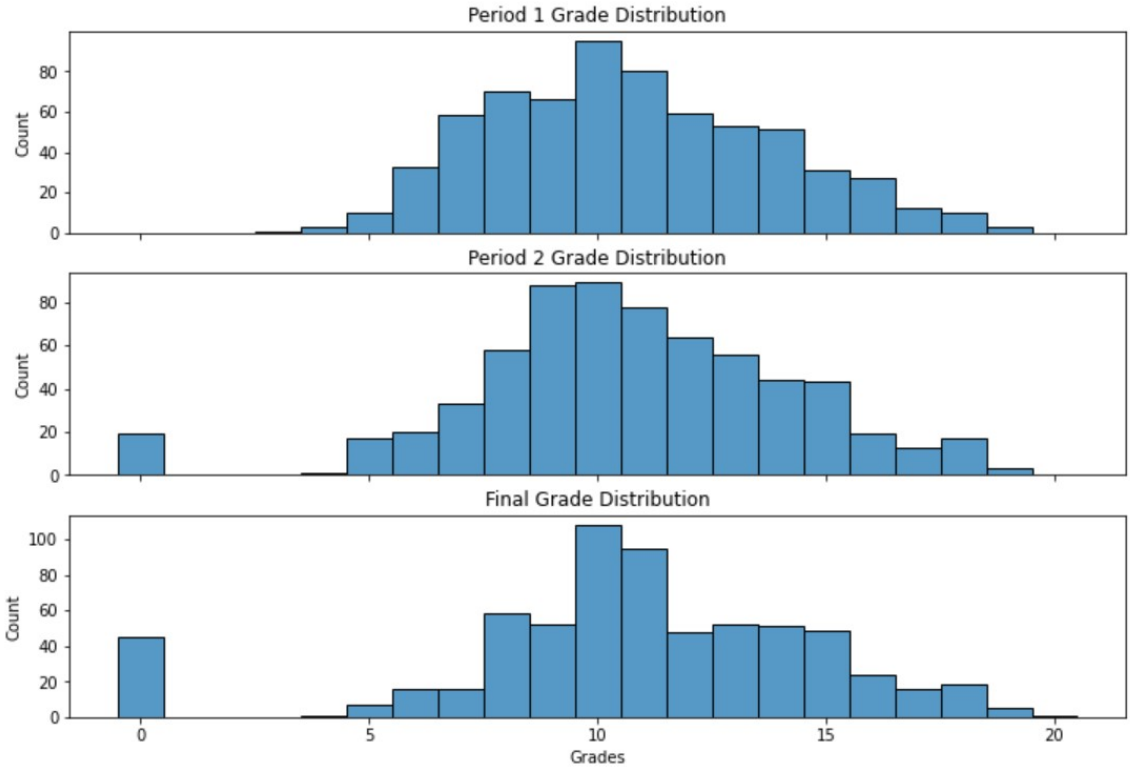


Figure 1: Histogram distributions of Grades from each grading period

As the periods go on, we noticed an increase in students receiving 0's for their performances in class.

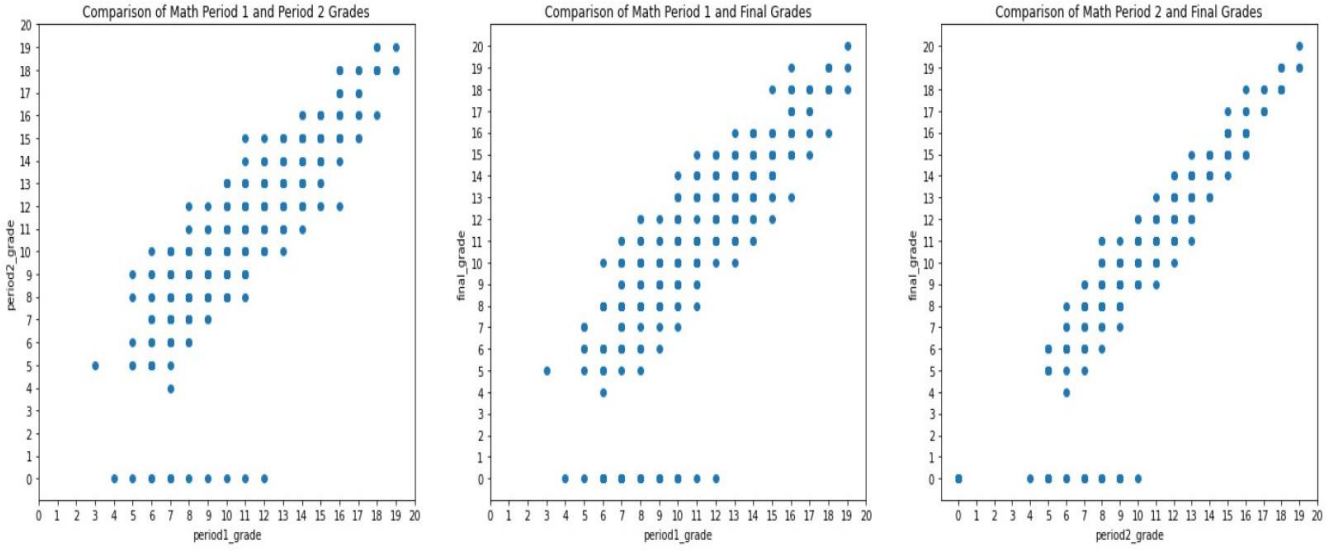


Figure 2: Scatterplots showing the relationship between the different grading periods of the math course

Judging from these plots, the distribution of grades is fairly linear.

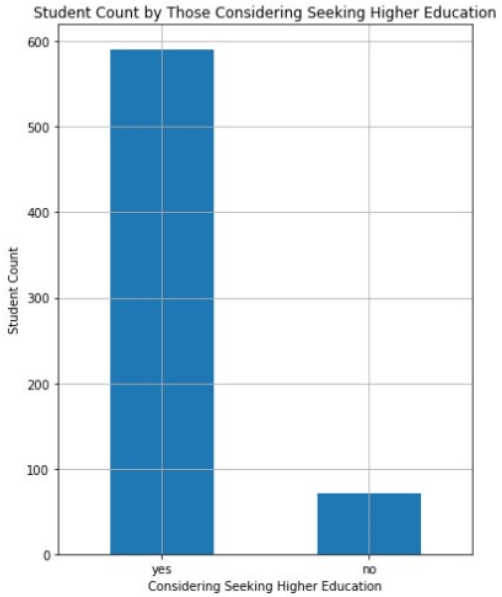


Figure 3: Number of students planning to seek higher education

We found this interesting that such a large number of students were planning on going to college to further their education. This is much higher than the percentage at the high school I (Trevor) went to.

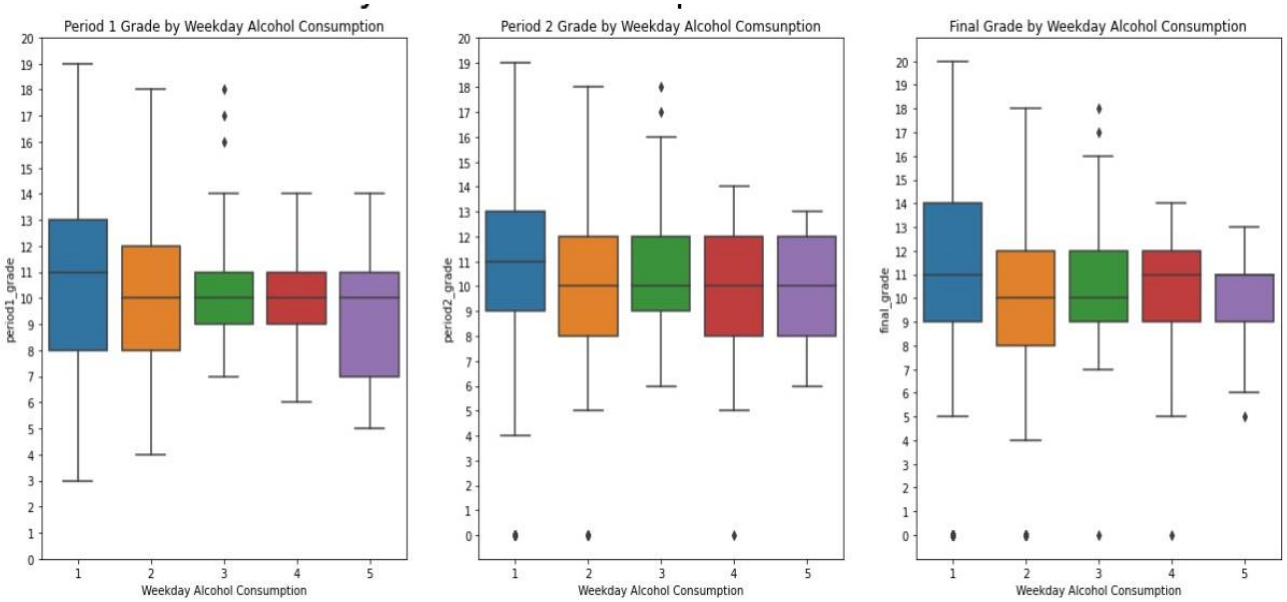


Figure 4: Weekday alcohol consumption by Period

As alcohol consumption grows higher, grades tent to become worse.

V. SUMMARY OF FINDINGS

This dataset had many different variables and categories of different sizes, types, and content. Correlations could be drawn from a very large combination of different categories. For example, the way that alcohol consumption effected the student performance was found in our data. The correlation being that alcohol overall did negatively affect the way a student was able to perform in class. Using various different graphs and charts, we were able to visualize almost every single category in the dataset. This was very helpful to understanding the data since it is much easier to grasp what's happening when you can look at the data in a clean and presentable way.