**The influence of demographic, social, and school related variables** **on student’s grades -** Milestone Report

**Table of Contents:**

1. Introduction
2. Data Acquisition/Cleaning
3. Data Exploration
4. Initial Findings
5. Next Steps

**1. Introduction**

This project studied student’s grades as compared to a range of variables. The data set was from Paulo Cortez, University of Minho, GuimarÃ£es, Portugal, <http://www3.dsi.uminho.pt/pcortez>. Student achievement in secondary education was gathered for two Portuguese schools. The attributes included the student’s grades, demographic, social and school related variables. The datasets that were provided included student information for two categories, mathematics and the Portuguese language. The data from both datasets were studied for this project, however only analysis of the Portuguese language is presented for this milestone report. The data was collected by using school reports and surveys. The student’s grades are provided for three periods, G1, G2, and G3. These correspond to three school periods with G3 being the final student grade. By analyzing these data sets, correlations between the attributes and the students grades can be explored with the goal of determining what impacts the students grades, and therefore what recommendations can be made to improve students school performance.

**Objectives**

The work performed for this project seeks to answer the following questions:

1. Which demographic, social, and/or school related variables have a positive effect on a student’s grades?
2. Do the demographic, social, and /or school related variables have the same impact on mathematics grades and Portuguese language grades?

**Data Set Information:**

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

**Data Set Features, Names, and Types**

The demographic, social, and school related variables that were investigated include:

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Set Feature** | **Name of Feature** | **Type** | **Value** |
| School of attendance | school | Binary | **GP** Gabriel Pereira or **MS** Mousinho da Silveira |
| Sex | sex | Binary | **F** or **M** |
| Age | age | Numeric | **15 to 22** |
| Address | address | Binary | **U** urban or **R** rural |
| Family size | famsize | Binary | **LE3** less or equal to 3 or **GT3** greater than 3 |
| Cohabitation status of parents | Pstatus | Binary | **T** living together or **A** apart |
| Mother’s education level | Medu | Numeric | **0** none, **1** 4th grade, **2** 5th to 9th grades, **3** secondary education, **4** higher education |
| Father’s education level | Fedu | Numeric | **0** none, **1** 4th grade, **2** 5th to 9th grades, **3** secondary education, **4** higher education |
| Mother’s job | Mjob | Categorical | **teacher**, **health** care, civil **services**, **at\_home**, or **other** |
| Father’s job | Fjob | Categorical | **teacher**, **health** care, civil **services**, **at\_home**, or **other** |
| Reason for choosing the school | reason | Categorical | Close to **home**, school **reputation**, **course** preference, or **other** |
| Students guardian (guardian) | guardian | Categorical | **mother**, **father**, **other** |
| Traveltime to the school | traveltime | Numeric | **1** < 15 min, **2** 15 to 30 min, **3** 30 min to 1 hour, **4** greater than 1 hour |
| Weekly study time | studytime | Numeric | **1** < 2 hours, **2** 2 to 5 hours, **3** 5 to 10 hours, **4** > 10 hours |
| Number of past failures | failures | Numeric | n if **1** <= n < **3**, else **4** |
| Extra educational support | schoolsup | Binary | **yes** or **no** |
| Family educational support | famsup | Binary | **yes** or **no** |
| Extra paid classes | paid | Binary | **yes** or **no** |
| Extra-curricular activities | activities | Binary | **yes** or **no** |
| Attended nursery school | nursery | Binary | **yes** or **no** |
| Wants to go on to higher education | higher | Binary | **yes** or **no** |
| In a romantic relationship | romantic | Binary | **yes** or **no** |
| Quality of family relationships | famrel | Numeric | From **1** – very bad to **5** - excellent |
| Free time after school | freetime | Numeric | From **1** – very low to **5** – very high |
| Going out with friends | goout | Numeric | From **1** – very low to **5** – very high |
| Workday alcohol consumption | Dalc | Numeric | From **1** – very low to **5** – very high |
| Weekend alcohol consumption | Walc | Numeric | From **1** – very low to **5** – very high |
| Current health status | health | Numeric | From **1** – very bad to **5** – very good |
| Number of school absences | absences | Numeric | **0** to **93** |
| First period grade (G1) | G1 | Numeric | **0** to **20** |
| Second period grade (G2) | G2 | Numeric | **0** to **20** |
| Final grade (G3) | G3 | Numeric | **0** to **20** |

**Client**

There are several potential clients that would benefit from the analyses of the data sets. These include the students, the parents of the students, the student’s teachers, and school administrators.

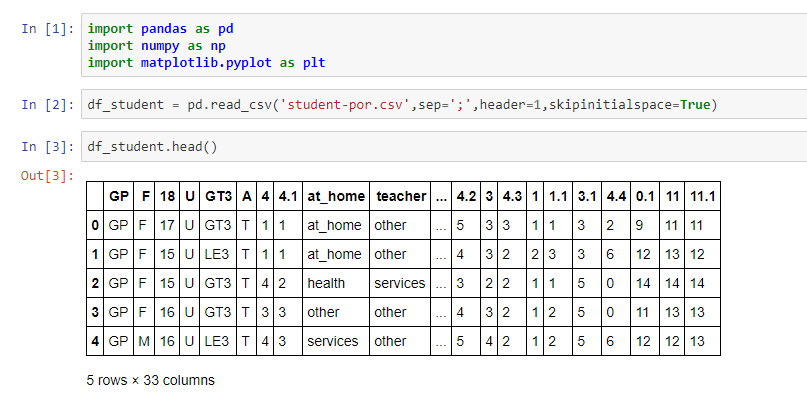
**2. Data Acquisition/Cleaning**

In this section I will explore the datasets and use preprocessing and data wrangling techniques to prepare the data. This will include the following steps:

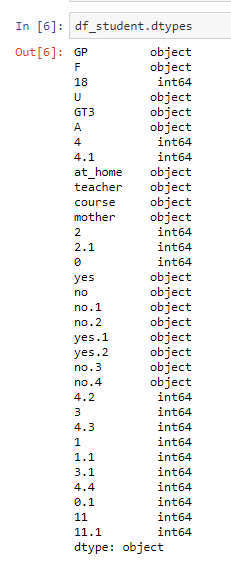
1. Loading the data and extracting general info and structure
2. Exploring data types
3. Identifying & dealing with missing values
4. Preprocessing techniques

**Data Information and Structure**

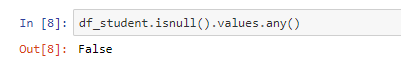
The data sets were fairly clean when I got them. There were no missing or NAN values. In order to utilize the data sets, I added column headers, and divided the data in to male and female groups.

****

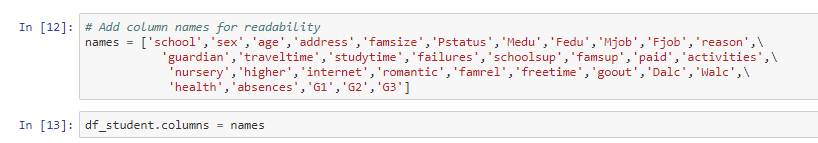
General data types:



Missing values?

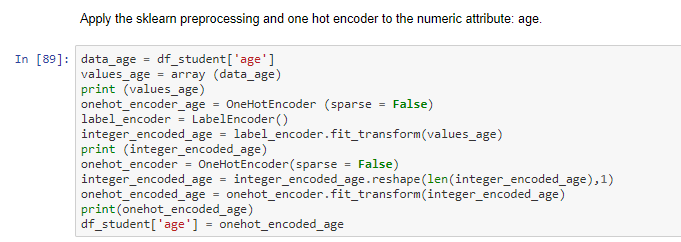


Add column names for readability

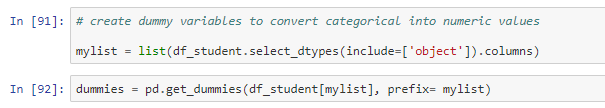


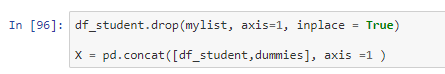
Preprocessing Techniques

Two preprocessing techniques were used on the data sets. sklearn preprocessing and one hot encoder were used to change the categorical attributes: age, Mjob, Fjob, reason, and guardian.



Pandas pd.get\_dummies was used to convert the Boolean categorical data into ones and zeros. These attributes included school, sex, address, famsize, Pstatus, schoolsup, famsup, paid, activities, nursery, higher, internet, and romantic. This resulted in the addition of 13 columns to the data set.





**3. Data Exploration**

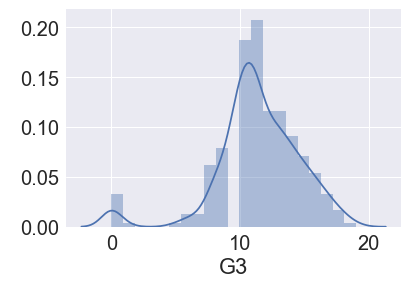
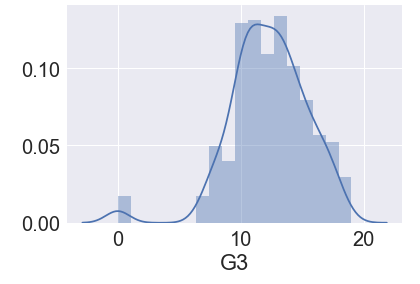
**Grades for Male vs. Female students**

The first comparison that I opted to explore was the difference between the grades of male and female students.



The data was separated in to two data frames, df\_student\_f and df\_student\_m. Presented below are grade distributions for male and female students for Portuguese language grades and math grades.

Grade Distribution for female students Grade Distribution for male students

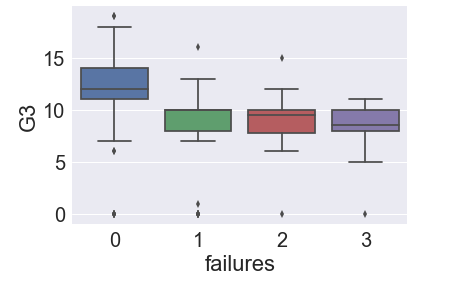
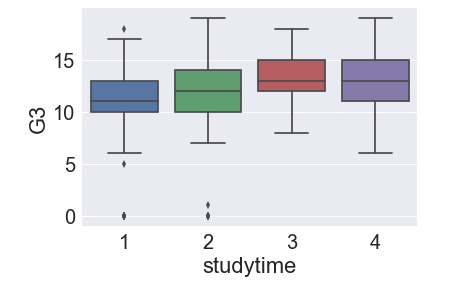


The distributions show a similar trend. The mean for the female grades in Portuguese is 12.3. The mean for the male grades in Portuguese is 11.4. A t-test was performed to determine if the means are indeed different. For the Portuguese language grades a t-value of -3.31 and a p-value of 0.00095 was calculated. These results indicate that groups are similar and the data didn’t occur by chance. So, there is statistically no difference between male and female grades.

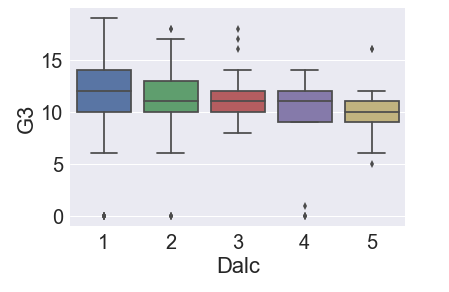
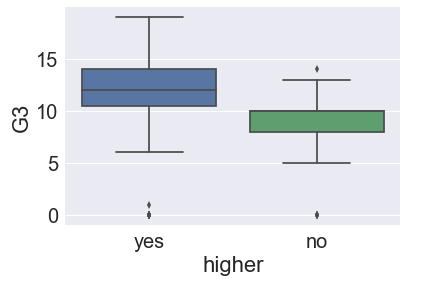
**Analysis of Grades vs. Demographic, Social and School Related Attributes**

By understanding which demographic, social and school related attributes affect grades, we can possibly improve grade outcomes. Box and whisker plots were created for each demographic, social, and school related attribute. The plots for amount of study time (studytime), number of failures (failures), interest in higher education (higher), quality of family relationships (famrel), and daily consumption of alcohol (dalc) showed more variation than the other variables. The box plots for these five attributes are plotted below.

Study time Failures

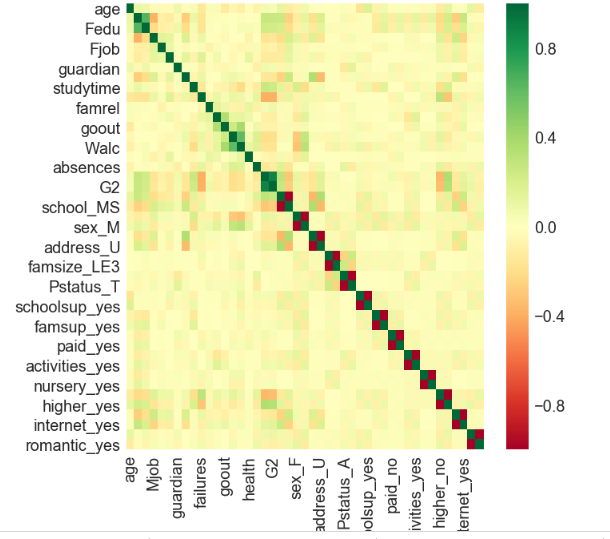


Interest in higher education Daily consumption of alcohol



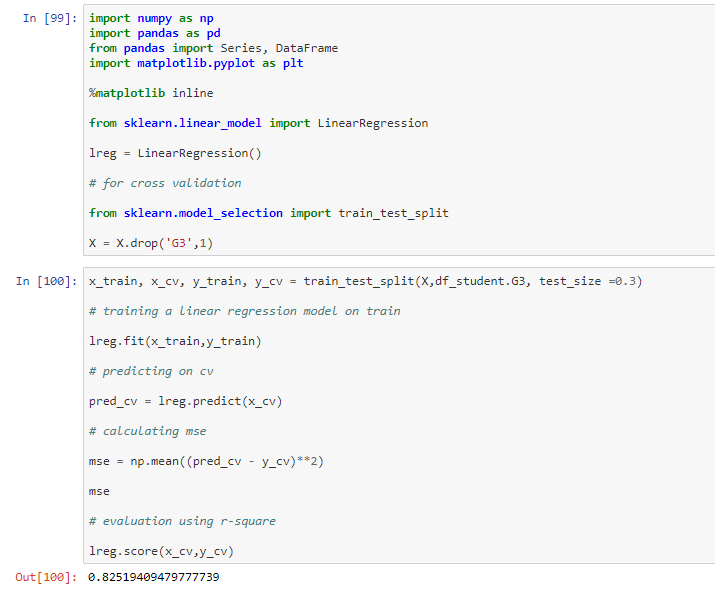
A Pearson correlation was used to test if there is a statistically significant linear relationship between the above attributes and student’s grades. The Pearson r-value for study time, failures, interest in higher education, and daily consumption of alcohol were 0.25, -0.39, 0.33, and -0.2, respectively. These correlation values indicate weak positive correlations between study time and interest in higher education with grades, and weak negative correlations between failures and daily consumption of alcohol with grades. I think that the weak correlations are a result of possible interdependence between some of the attributes. The Pearson r-value assumes that the variables are independent from one another.

Additional exploration between student’s grades and the attributes was conducted by using a correlation heatmap. The heatmap is displayed below. Not surprisingly, based on the above box plots, the correlation between attributes is weak at best.

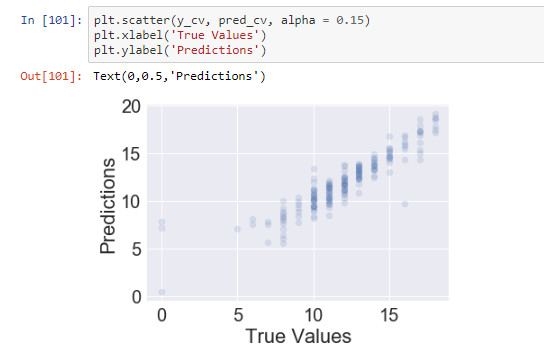


**4. Initial Findings**

After the data exploration a machine learning algorithm was utilized to create a predictive model. The train\_test\_split algorithm used a portion of the attribute data, in this case 30%, that was available to perform a linear regression on the selected portion of data, and tested the fit on the remaining data. The following code was used:



The score of the linear regression model was computed to be 0.87. A scatter plot shows the correlation between the true values and the predicted values.



**5. Next Steps**

The data has been explored and an initial model has been created to predict student’s grades based on the attributes. The next step involves optimizing the model. Optimizing the coefficient for each attribute will be explored using ridge regression and lasso regression. The attributes that maximize the accuracy of the model will be determined. Additionally, the math data that is provided within the same data set will be explored and the outcomes will be compared with the outcomes from the Portuguese language data.

**6. Resources**

* Dataset -  <http://www3.dsi.uminho.pt/pcortez>
* Project Proposal - [github link](https://github.com/emenriquez/Springboard-Coursework/blob/master/Capstone%20Project%201/Erik%20Enriquez%20-%20Capstone%20Project%201%20Proposal.pdf)
* Code (IPython Notebook) - https://github.com/hhtdata/SB2018/blob/master/Capstone1\_student-Por%20(10).ipynb