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).

Task 1: Introduction (200-400 words) [10 marks]

Provide a statement of the problem, outlining the problem that your chosen dataset addresses. The statement of the problem should briefly address the question:

What is the problem that you will investigate in this assignment?

Your introduction must describe:

• The aim of your work, what are you trying to achieve, and research questions you attempted to answer.

• All assumptions that your data must meet.

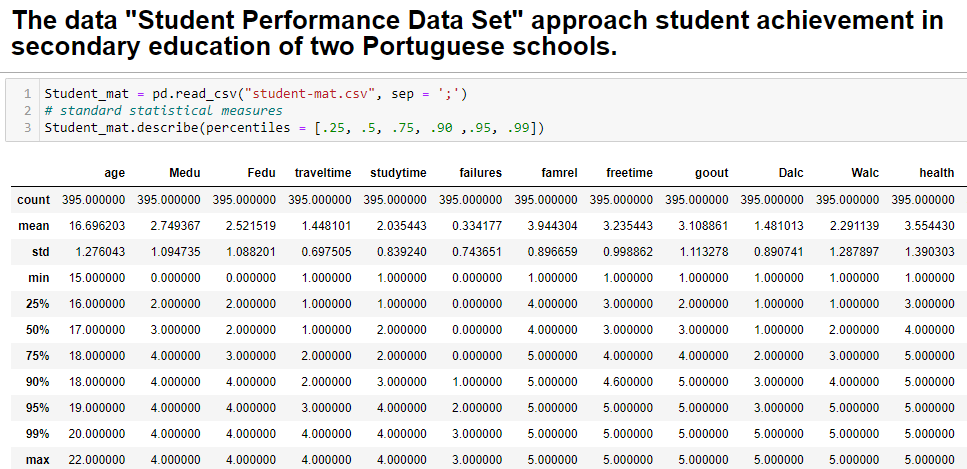
Answer:

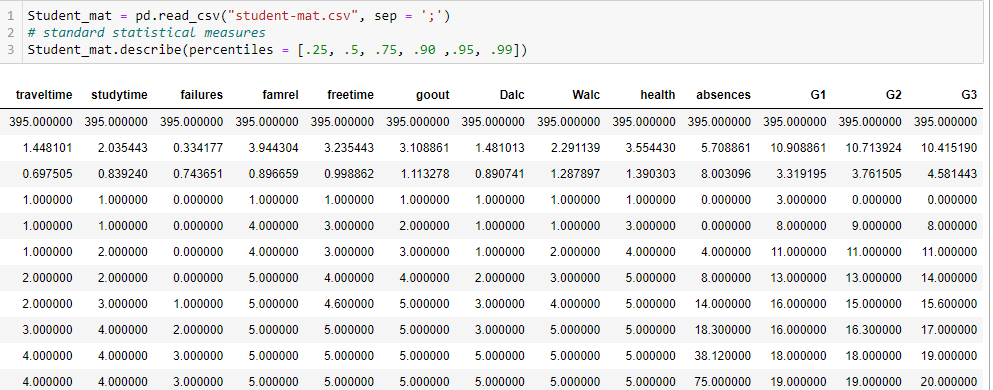
In the datasets provided from the UCI Machine Learning repository, I selected the data file information as "Student Performance Data Set". This student data set is to access the academic performance of students in two high schools in Portugal country. Statistical data attributes in the data are including student score, gender, age, social and school related features, and it was collected by using school reports and questionnaires. And especially in the student dataset, there are two distinct subjects provided related to the individual achievement of students is: Math (mat) and Portuguese language (por).

And to plot the data, analyse the data and make statistics of the data from the student data set, I will go step by step including “Data Exploration”, “Correlation Analysis”, “Linear Regression”, “Statistical Inference”.

Task 2: Data Exploration [20 marks]

The data "Student Performance Data Set" approach student achievement in secondary education of two Portuguese schools.





• How many features (attributes), instances and what data types are these?

The data Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires.

Attributes for both student-mat.csv (Math course):

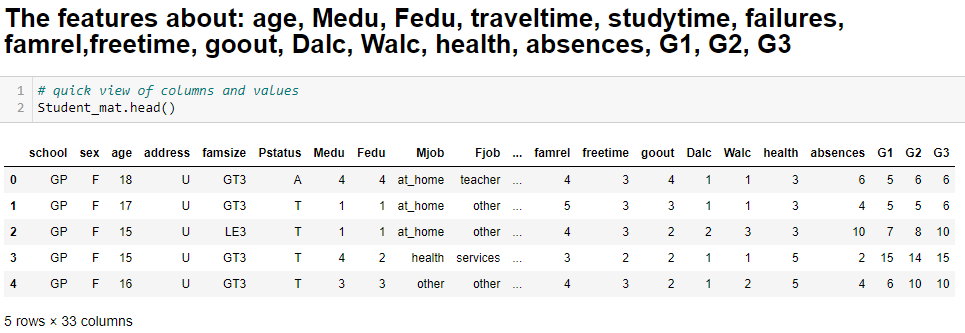
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)  
2 sex - student's sex (binary: 'F' - female or 'M' - male)  
3 age - student's age (numeric: from 15 to 22)  
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)  
5 famsize - family 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)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade)

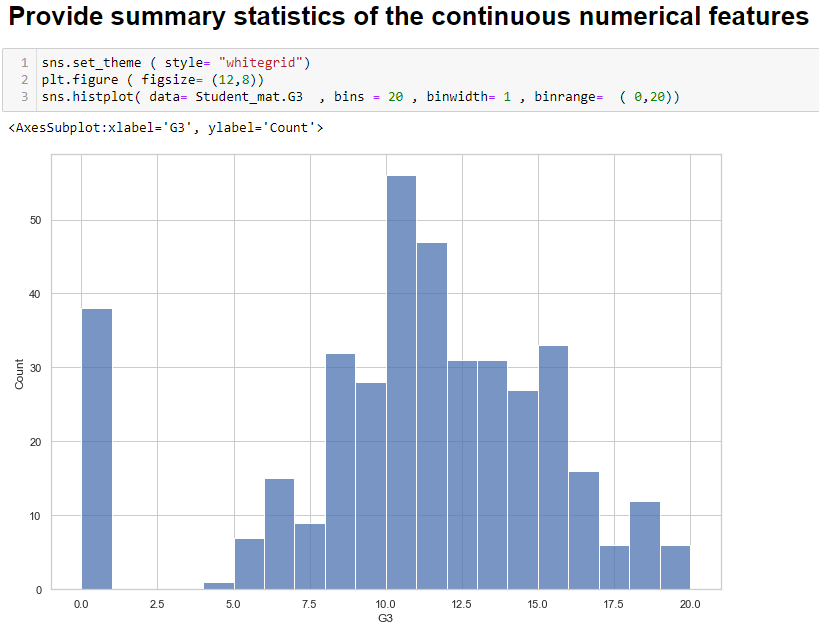
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services')   
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services')  
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')  
12 guardian - 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<=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)   
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)

Table

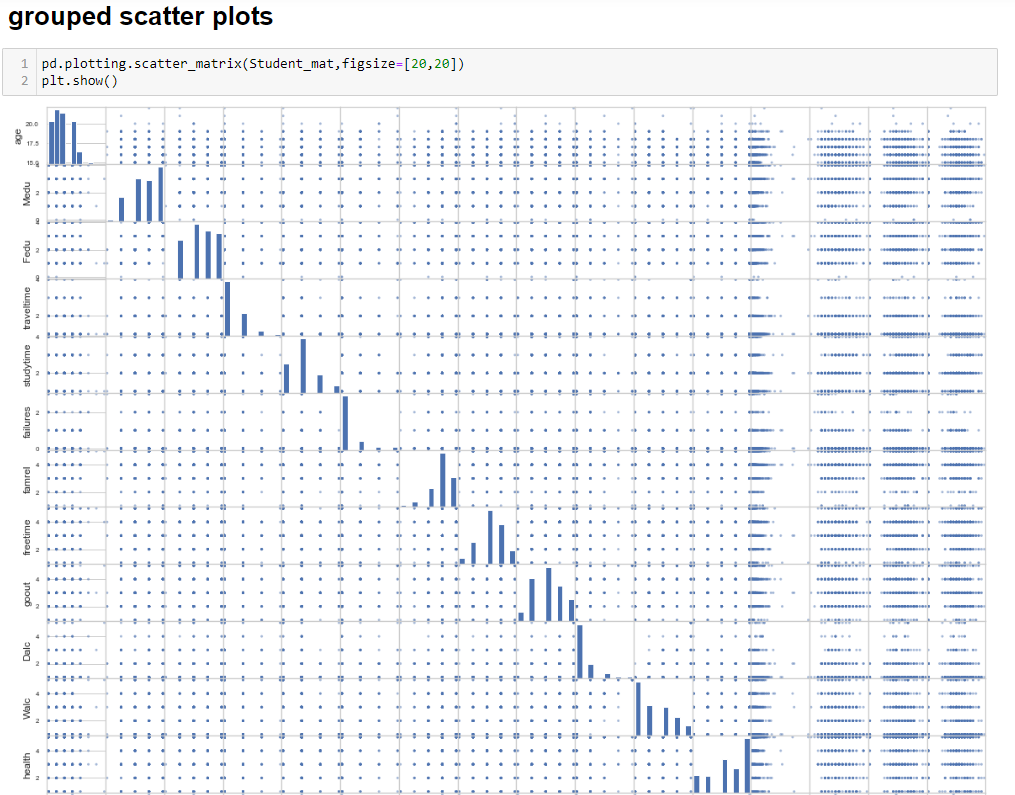
Description automatically generated

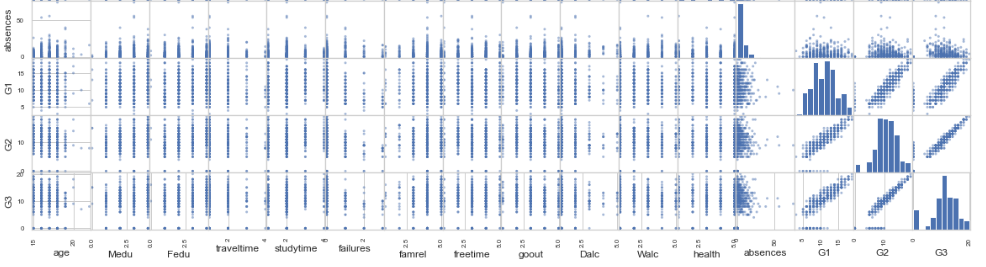


• Provide summary statistics of the continuous numerical features.



• Illustrate the features of your dataset using meaningful boxplots, histograms and grouped scatter plots (remember, these plots allow you to analyse the individual distribution of features, as well as the relationship between them).





• Explain what you can learn about the dataset from the diagrams.

Is there any sign of violation of assumption(s) (Có dấu hiệu vi phạm (các) giả định không)? If yes, explain your approach to handle it before moving to the next tasks? (Giải thích những gì bạn có thể học về tập dữ liệu từ các sơ đồ. Có bất kỳ dấu hiệu của vi phạm (các) giả định? Nếu có, hãy giải thích cách tiếp cận của bạn để xử lý nó trước khi chuyển sang các nhiệm vụ tiếp theo?)

các nhiệm vụ tiếp theo?)

Task 3: Correlation Analysis [15 marks]

Perform correlation analysis and provide correlation matrixes and plots. Discuss your findings in terms of

1. **correlation between the independent variables (Multi-Collinearity**) (mối tương quan giữa các biến độc lập (Multi-Collinearity)). Chọn sẵn luôn vài biến, hay là phải plot ra hết

Hướng giải quyết: Plot ra các giá trị independent variables(Multi-Collinearity), và giải thích.

1. between independent variables and dependent variables. Hỏi sao có cần plot ra 2 inde và depen hay không, hay chỉ giải thích.

Is there any sign of violation of assumption(s)? If yes, explain your approach to handling it before moving to the next task?

* Hướng giải quyết: Plot ra các giá trị không phụ thuộc và phụ thuộc rồi đưa ra hướng giải quyết với các giá trị đấy.

Task 4: Linear Regression [25 marks]

Taking the result of Task 3 into consideration, perform multivariate regression using Python. Provide the results including regression results, statistical significance metrics, and coefficients tables from this model. (thực hiện hồi quy đa biến bằng Python. Cung cấp các kết quả bao gồm kết quả hồi quy, số liệu ý nghĩa thống kê và bảng hệ số từ mô hình này.)

Task 5: Statistical Inference (300-400 words). [25 marks]

In this section, you are required to describe and analyse the results of your regression model. Are all the assumptions made in Task 1 satisfied? Provide evidence to support your answer. Compare your findings in Task 3’s with the coefficients table results generated in Task 4 and discuss your findings.

(Trong phần này, bạn được yêu cầu mô tả và phân tích kết quả của mô hình hồi quy của bạn. Tất cả các giả định được đưa ra trong Nhiệm vụ 1 có thỏa mãn không? Cung cấp bằng chứng để hỗ trợ câu trả lời của bạn. So sánh kết quả của bạn trong Nhiệm vụ 3 với kết quả bảng hệ số được tạo trong Nhiệm vụ 4 và thảo luận về những phát hiện của bạn.)

There will be 5 marks for the presentation of the assignment including spelling and grammar, layout, formatting, and readability of the figures.

* Có nên gộp 2 files ra không hay làm riêng từng files
* Plot hết mọi giá trị ra hay plot từng giá trị thôi.
* Tại sao lại chọn biến
* Có cần plot hết giữ liệu ra không
* Cô trả lời: Chọn 1 trong 2 files để làm, ví dụ chọn student mat thì thôi ko chọn student por, và ngược lại. Nếu combine cũng được nhưng sẽ khó và phải làm nhiều việc hơn là chỉ chọn 1 file.