Zachary DeNoto

DSC 550

Case Study: Analyze data to see what factors impact alcohol consumption in students

This analysis will look at high school students of various ages and other variables that apply to them such as sex, parents' education, wealth, etc. to see if any of the factors have an impact on their alcohol usage (both weekday and weekend consumption). There were two datasets for this analysis, one for students in a math class, and another dataset for students in a Portuguese language class. For the analysis below I have only used the dataset for the students in the Portuguese language class. Hopefully when this analysis is complete it will provide some insights on the impact of several factors on alcohol usage and to see if alcohol usage has any impact on school factors such as the number of absences.

Variables used from the dataset:

- school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- 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 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – 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 - 1 hour)
- studytime - weekly study time (numeric: 1 - 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)

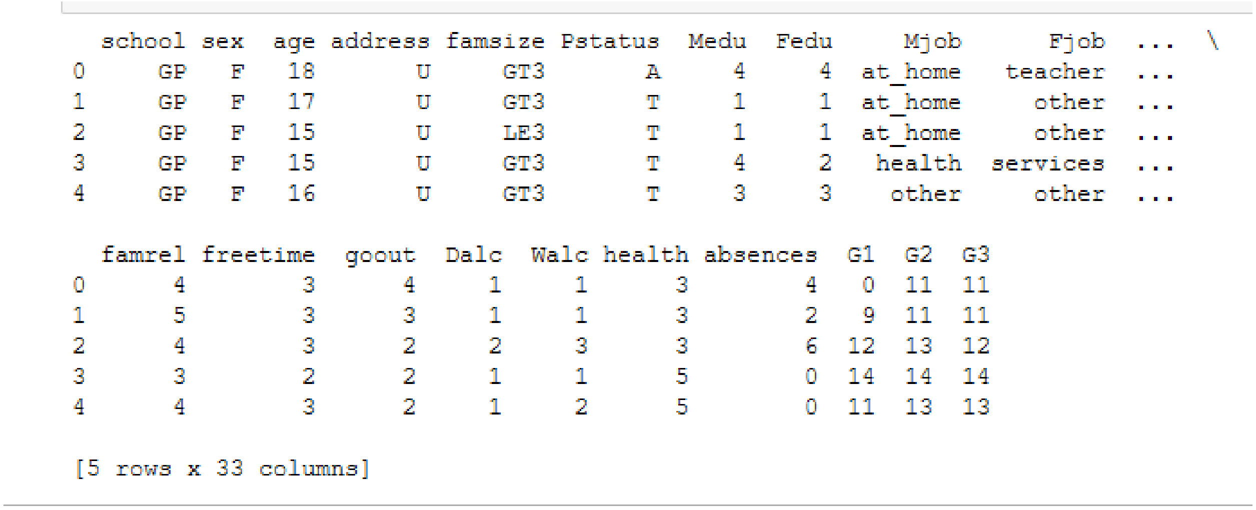
- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- 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)

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

- 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)

The step-by-step instructions to perform the graph analysis:

1. Load the data from the "student-por.csv" file into a DataFrame.
2. Display the dimensions of the file to see the number of columns and rows you are using.
3. Next, we display the first 5 rows of data so you can see the column headings and the type of data for each column. From this you can see we have 33 different columns.

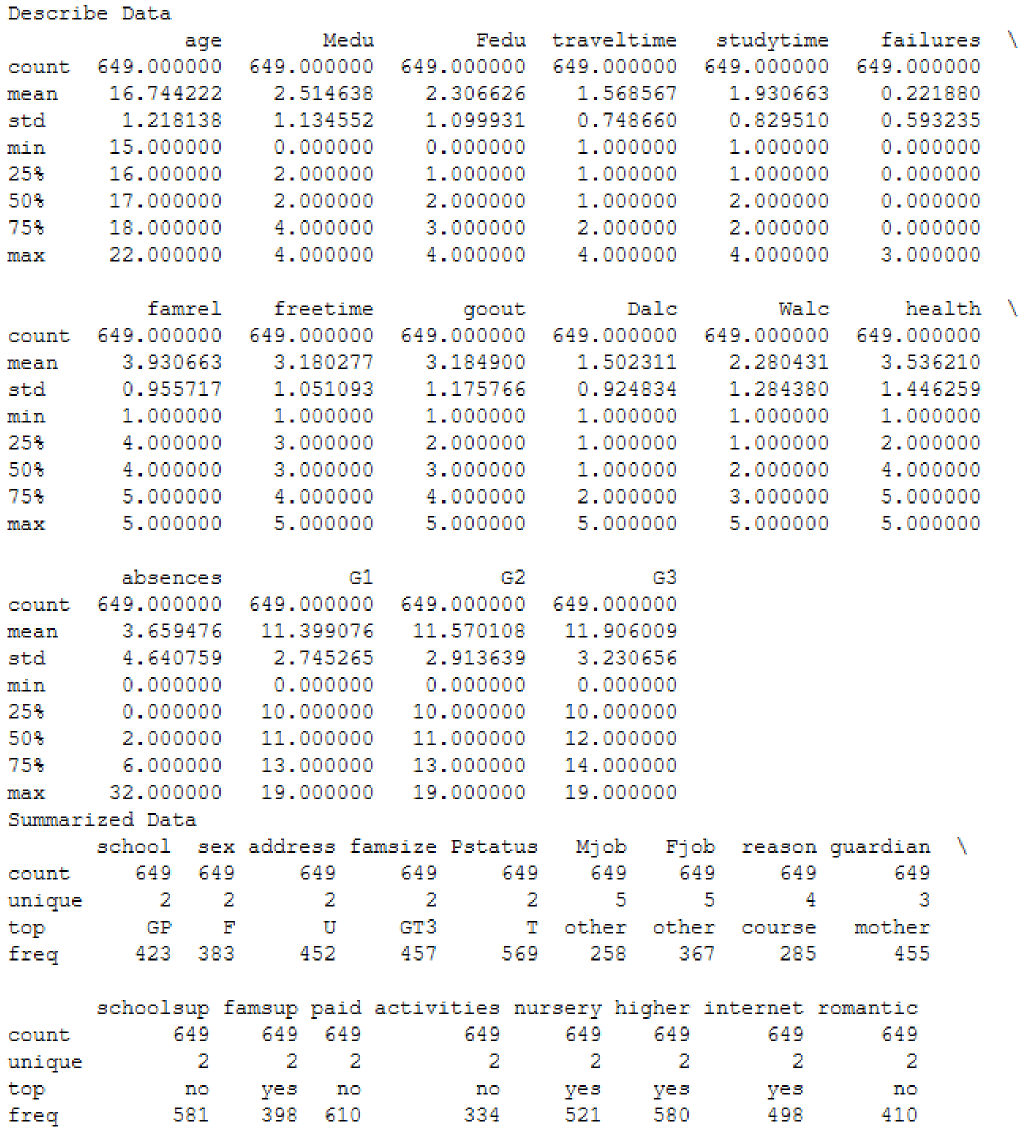


```
  school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  ...  \
0     GP   F   18       U     GT3       A     4     4   at_home   teacher  ...
1     GP   F   17       U     GT3       T     1     1   at_home     other  ...
2     GP   F   15       U     LE3       T     1     1   at_home     other  ...
3     GP   F   15       U     GT3       T     4     2    health  services  ...
4     GP   F   16       U     GT3       T     3     3     other     other  ...

  famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
0      4        3      4     1     1      3        4   0  11  11
1      5        3      3     1     1      3        2   9  11  11
2      4        3      2     2     3      3        6  12  13  12
3      3        2      2     1     1      5        0  14  14  14
4      4        3      2     1     2      5        0  11  13  13

[5 rows x 33 columns]
```

4. We have a few variables which we may not want to look at or change the values of. This includes possibly changing the sex from F to Female and M to Male. Other variables are

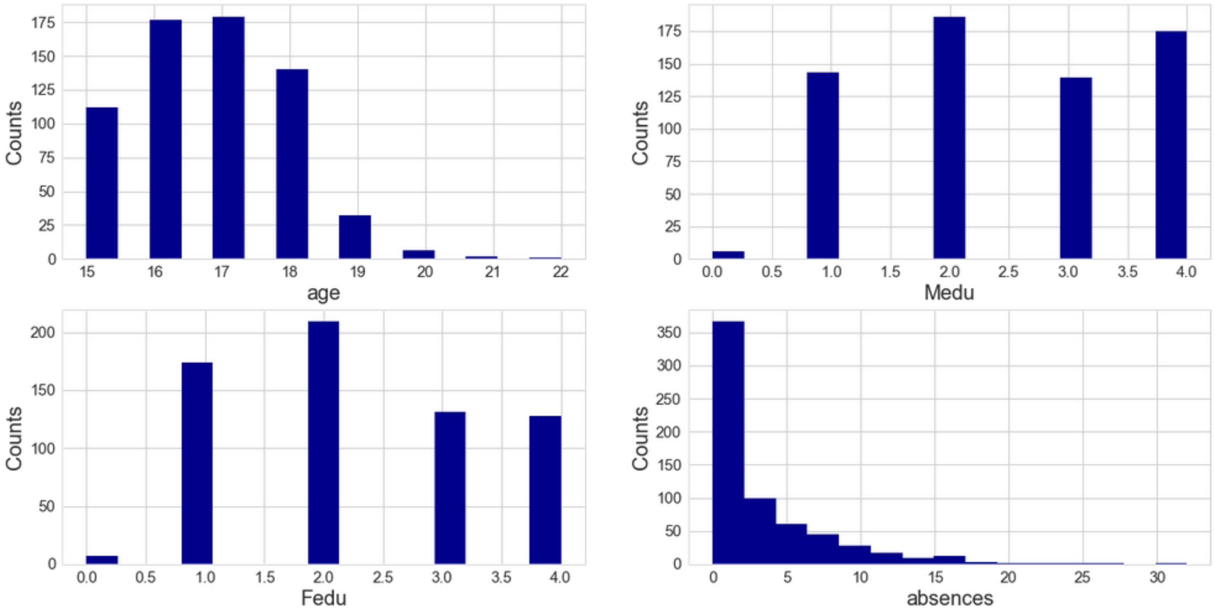
unclear what they are; for example, Medu and Fedu may not make sense to people, so the variable name could be changed to a more clear name such as Mothers_Education and Fathers_Education. Luckily there is no missing data so we do not need to drop any variables or rows.

5. Based on the summary information about the data, we can see that for all of the variables we have the same count, meaning no missing or blank data. Our data does not require any additional work and is ready to start a more in-depth analysis of our variables.

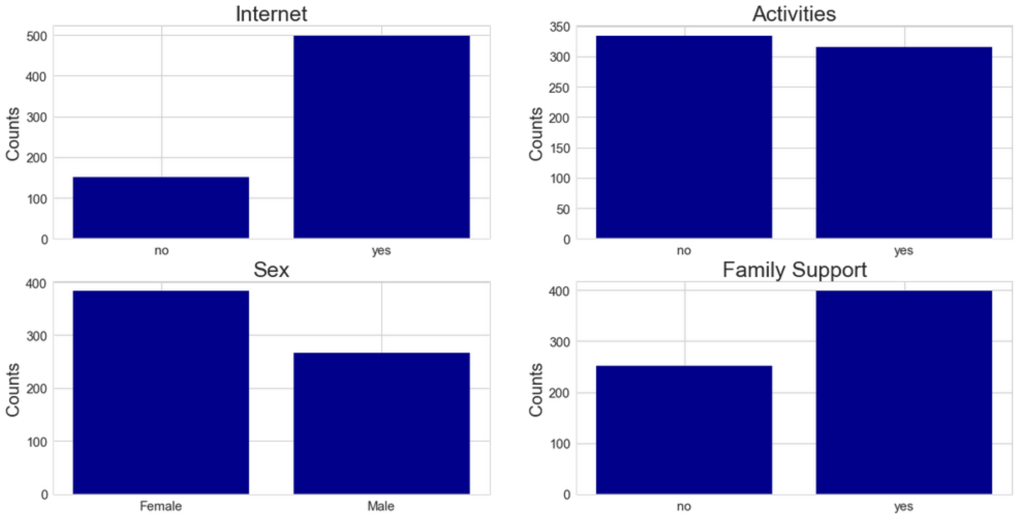


```
Describe Data
              age        Medu        Fedu  traveltime   studytime    failures  \
count  649.000000  649.000000  649.000000  649.000000  649.000000  649.000000
mean    16.744222    2.514638    2.306626    1.568567    1.930663    0.221880
std      1.218138    1.134552    1.099931    0.748660    0.829510    0.593235
min     15.000000    0.000000    0.000000    1.000000    1.000000    0.000000
25%     16.000000    2.000000    1.000000    1.000000    1.000000    0.000000
50%     17.000000    2.000000    2.000000    1.000000    2.000000    0.000000
75%     18.000000    4.000000    3.000000    2.000000    2.000000    0.000000
max     22.000000    4.000000    4.000000    4.000000    4.000000    3.000000

           famrel    freetime       goout        Dalc        Walc      health  \
count  649.000000  649.000000  649.000000  649.000000  649.000000  649.000000
mean     3.930663    3.180277    3.184900    1.502311    2.280431    3.536210
std      0.955717    1.051093    1.175766    0.924834    1.284380    1.446259
min      1.000000    1.000000    1.000000    1.000000    1.000000    1.000000
25%      4.000000    3.000000    2.000000    1.000000    1.000000    2.000000
50%      4.000000    3.000000    3.000000    1.000000    2.000000    4.000000
75%      5.000000    4.000000    4.000000    2.000000    3.000000    5.000000
max      5.000000    5.000000    5.000000    5.000000    5.000000    5.000000

         absences          G1          G2          G3
count  649.000000  649.000000  649.000000  649.000000
mean     3.659476   11.399076   11.570108   11.906009
std      4.640759    2.745265    2.913639    3.230656
min      0.000000    0.000000    0.000000    0.000000
25%      0.000000   10.000000   10.000000   10.000000
50%      2.000000   11.000000   11.000000   12.000000
75%      6.000000   13.000000   13.000000   14.000000
max     32.000000   19.000000   19.000000   19.000000
Summarized Data
       school  sex address famsize Pstatus   Mjob   Fjob  reason guardian  \
count     649  649     649     649     649    649    649     649      649
unique      2    2       2       2       2      5      5       4        3
top        GP    F       U     GT3       T  other  other  course   mother
freq      423  383     452     457     569    258    367     285      455

       schoolsup famsup paid activities nursery higher internet romantic
count        649    649  649        649     649    649      649      649
unique         2      2    2          2       2      2        2        2
top           no    yes   no         no     yes    yes      yes       no
freq         581    398  610        334     521    580      498      410
```

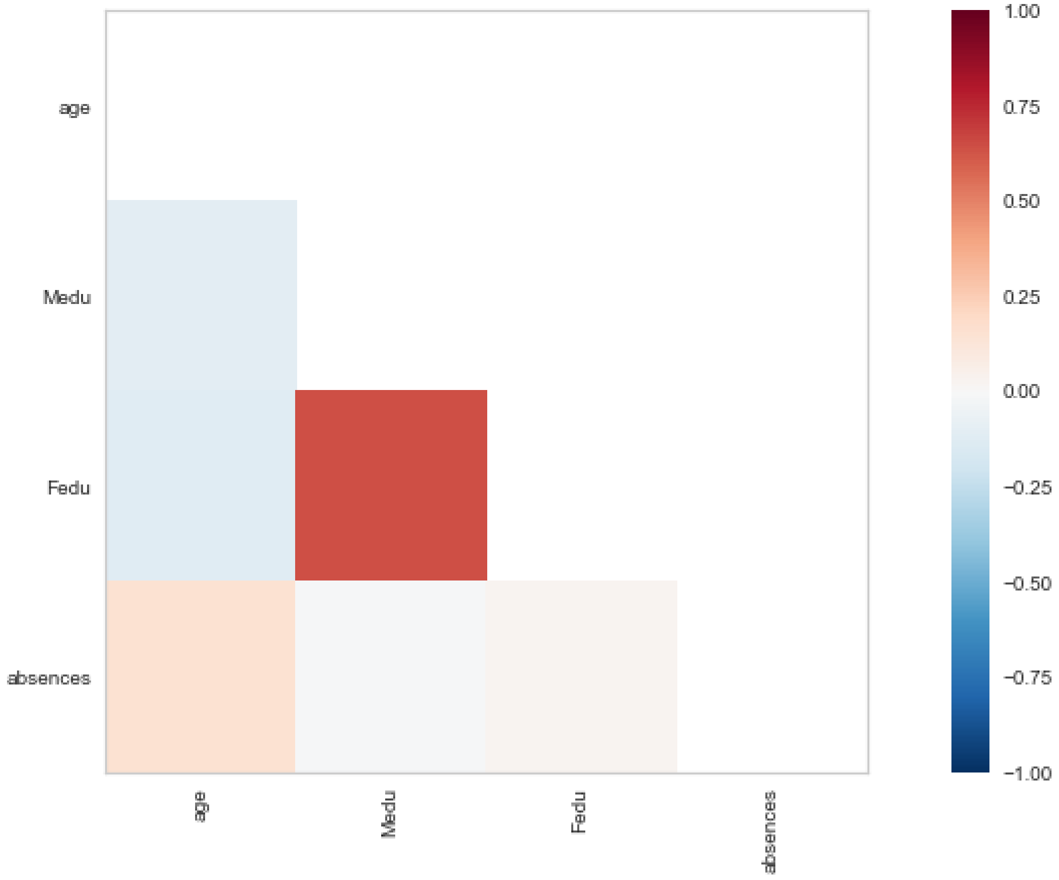
6. Next, make some histograms of several of the variables.
    a. Ages 17 and 16 were the most prevalent in the class.
    b. Most of the student had 0 to less than 5 absences.
    c. For the parents' educations, most of the fathers and mothers educations were between 5th and 9th grade education with a 2.0 being the most common.



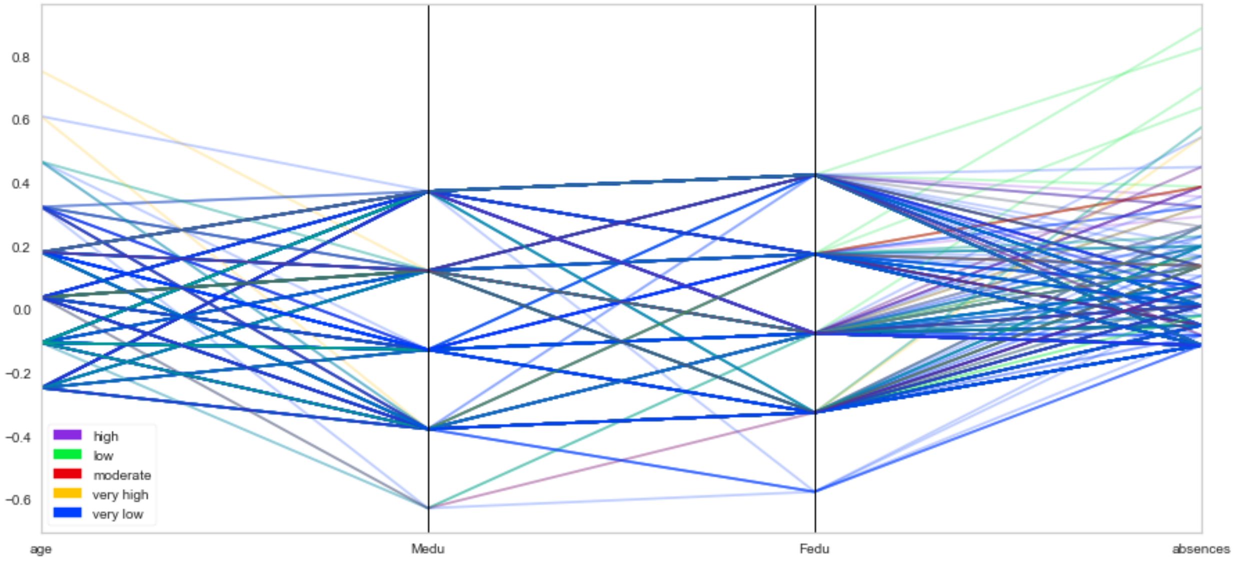
7. Make bar charts for variables with only a few options.
    a. From the bar charts we can see most students have Internet.
    b. You will notice there are more students with family support than students without family support.

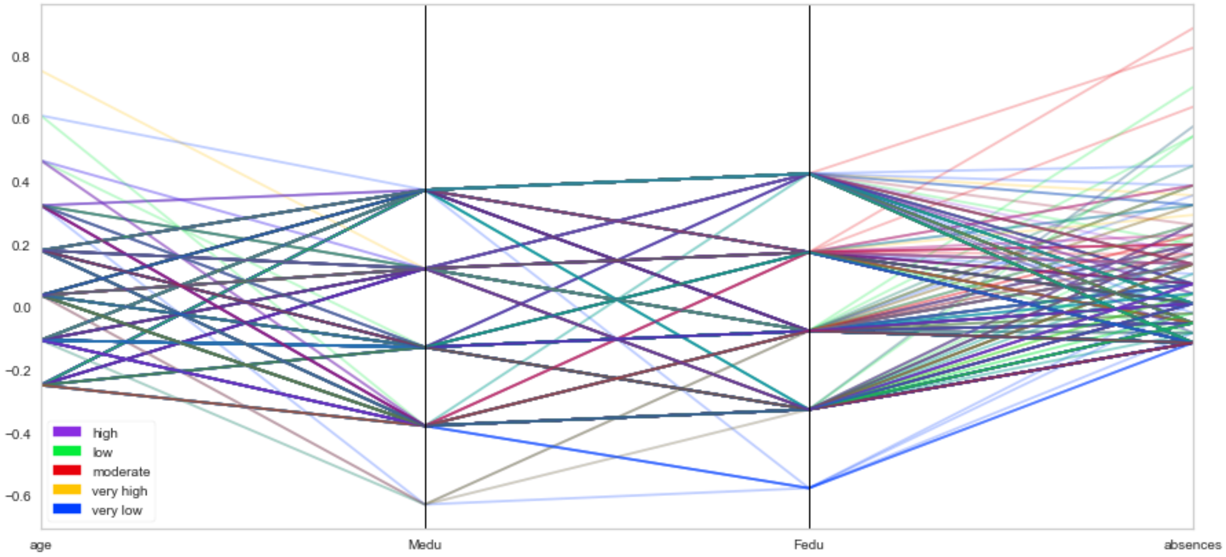


8. Next, look to see if the data is correlated by making Pearson Ranking charts.
    a. The correlation between the variables is low (1 or -1 is high positive or high negative, 0 is low or no correlation).
    b. These results show there is some correlation between the parents' education as well as a very small correlation between absences and age.

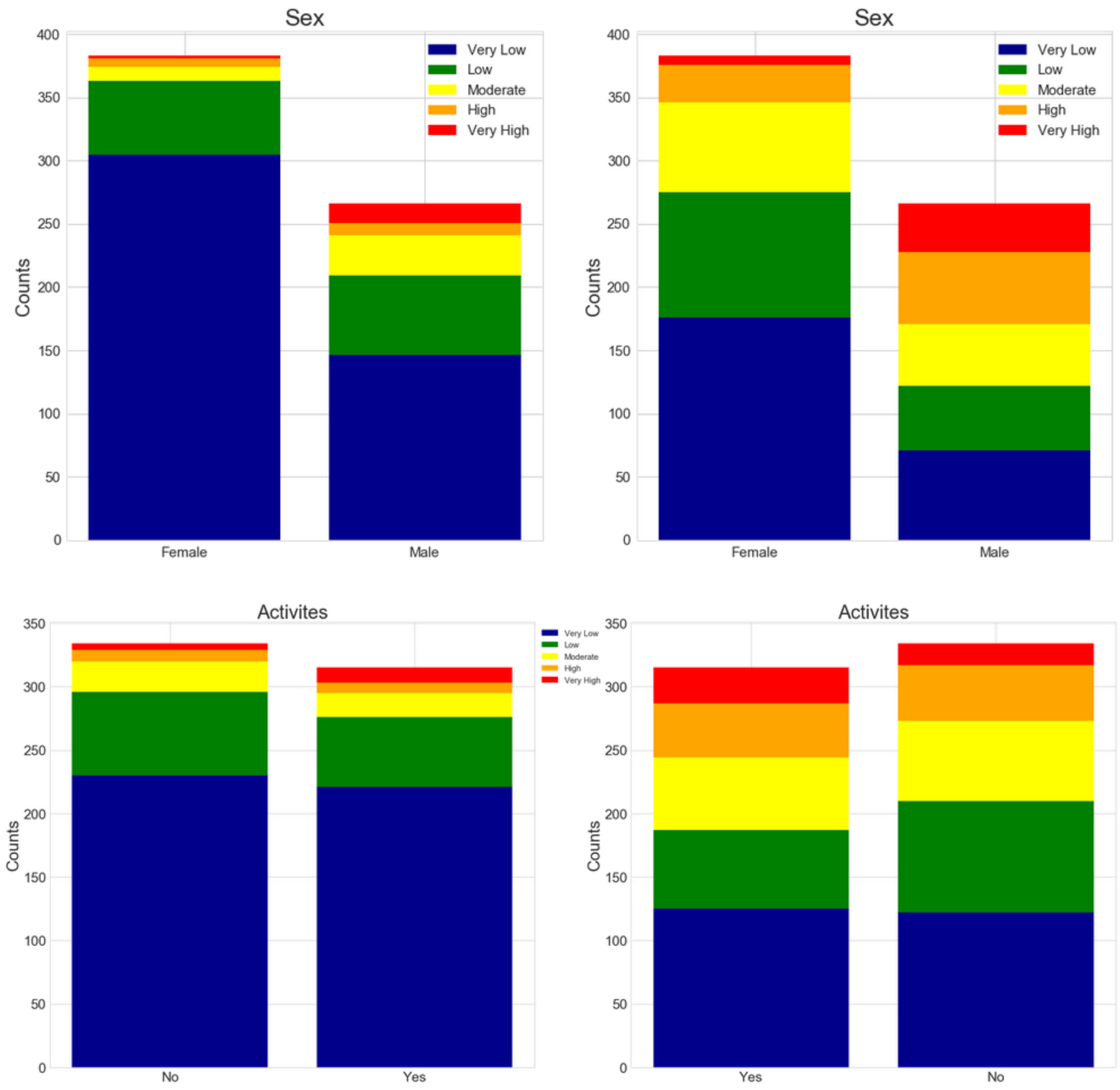


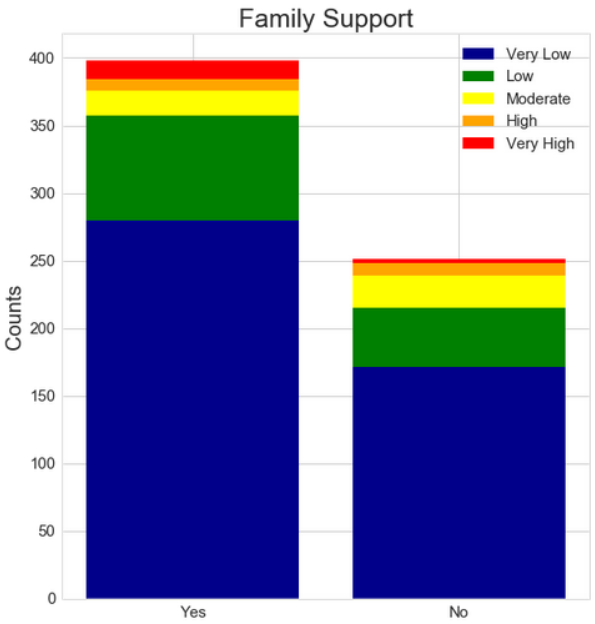
9. Next, use Parallel Coordinates visualization to compare the distributions of numerical variables between the different levels of alcohol consumption. The first visualization is for weekday alcohol consumption, while the second visual is for weekend alcohol consumption.
    a. Notice that the amount of alcohol consumption increases during the weekend as compared to the weekdays.

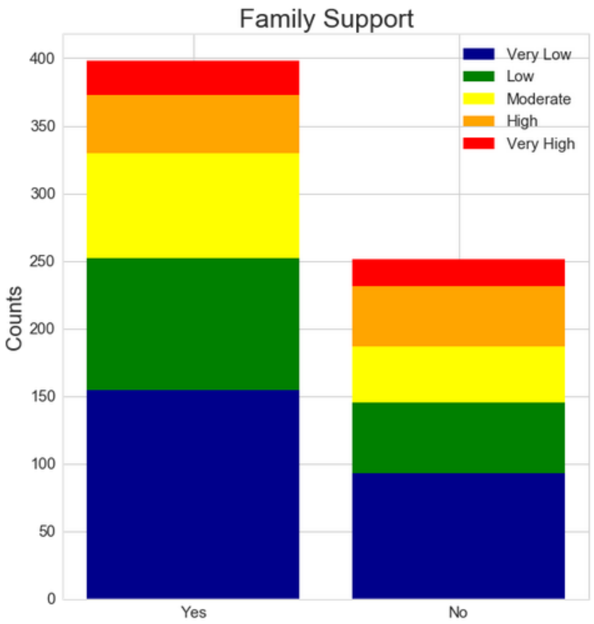




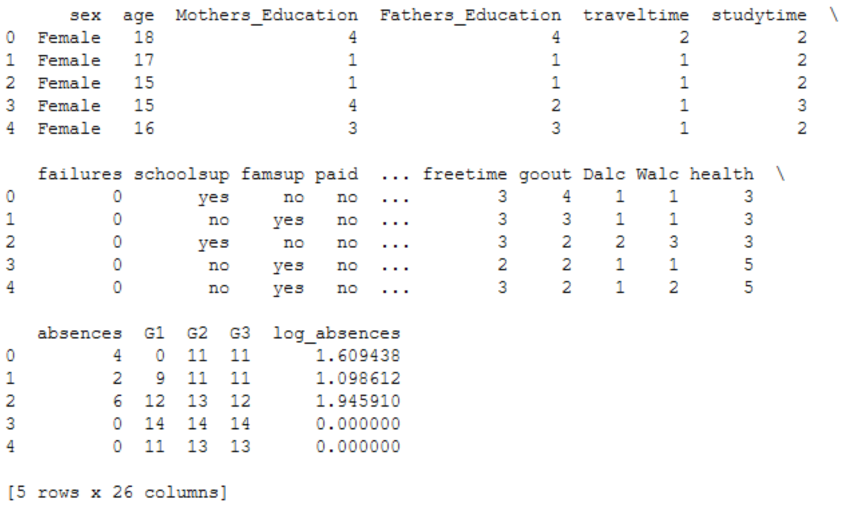
10. Next, use Stack Bar Charts to compare weekday and weekend alcohol consumption compared to the other variables. The left visualization is for weekday alcohol consumption, while the right visual is for weekend alcohol consumption.
    a. More students consume alcohol on the weekends compared to the weekdays.
    b. It appears that males consume more alcohol than females do during the weekdays as well as the weekends.





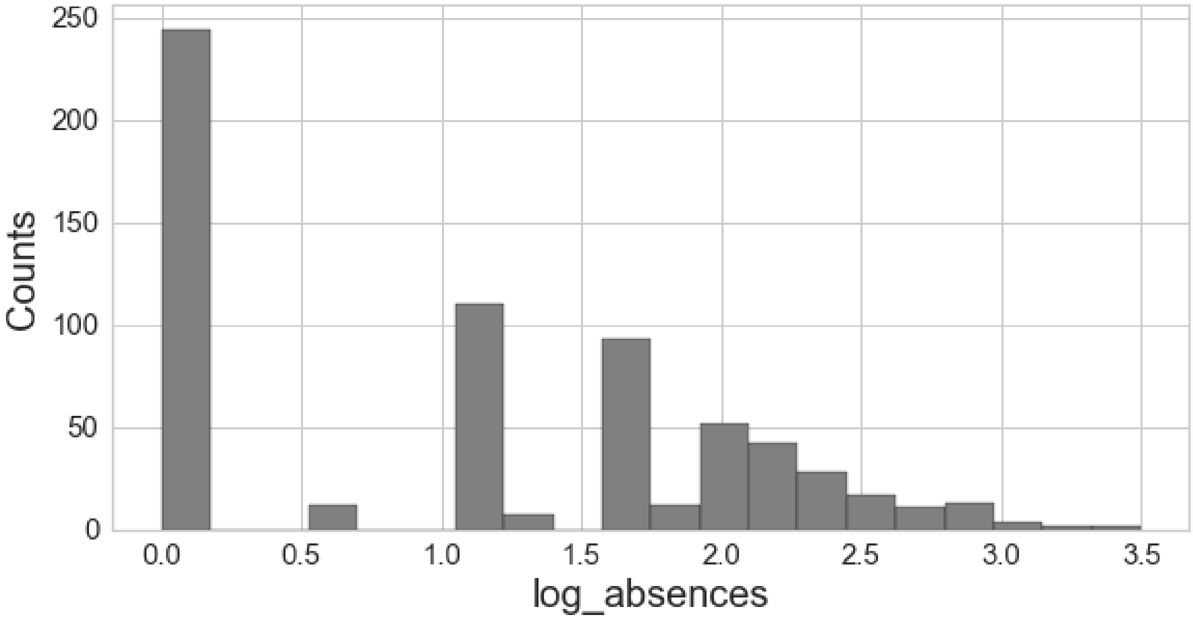


11. Now that some answers to the questions asked are coming to light, it is time to dig in deeper comparing weekday to weekend alcohol consumption with the features.
    a. First eliminate some unnecessary features and change some features names to be clearer.
    b. As previously noted, there are not any missing values in the dataset, allowing us to skip filling in missing values which can sometimes add bias to an analysis.

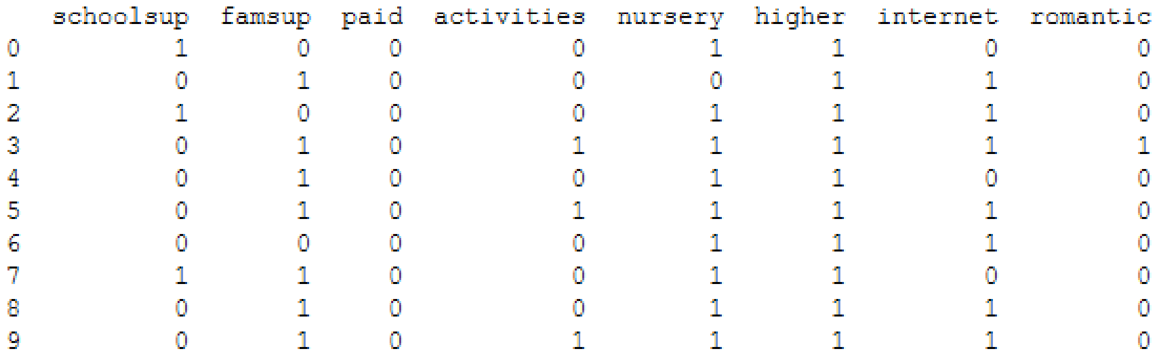


```
      sex  age  Mothers_Education  Fathers_Education  traveltime  studytime  \
0  Female   18                  4                  4           2          2
1  Female   17                  1                  1           1          2
2  Female   15                  1                  1           1          2
3  Female   15                  4                  2           1          3
4  Female   16                  3                  3           1          2

   failures schoolsup famsup paid  ... freetime goout Dalc Walc health  \
0         0       yes     no   no  ...        3     4    1    1      3
1         0        no    yes   no  ...        3     3    1    1      3
2         0       yes     no   no  ...        3     2    2    3      3
3         0        no    yes   no  ...        2     2    1    1      5
4         0        no    yes   no  ...        3     2    1    2      5

   absences  G1  G2  G3  log_absences
0         4   0  11  11      1.609438
1         2   9  11  11      1.098612
2         6  12  13  12      1.945910
3         0  14  14  14      0.000000
4         0  11  13  13      0.000000

[5 rows x 26 columns]
```

12. Next I plotted a new histogram using the log_absences feature, which are the logged values from the absences feature above as the previous histogram was very skewed. Though the new histogram is still skewed, is much easier to read and does not look nearly as bad as the previous histogram did.

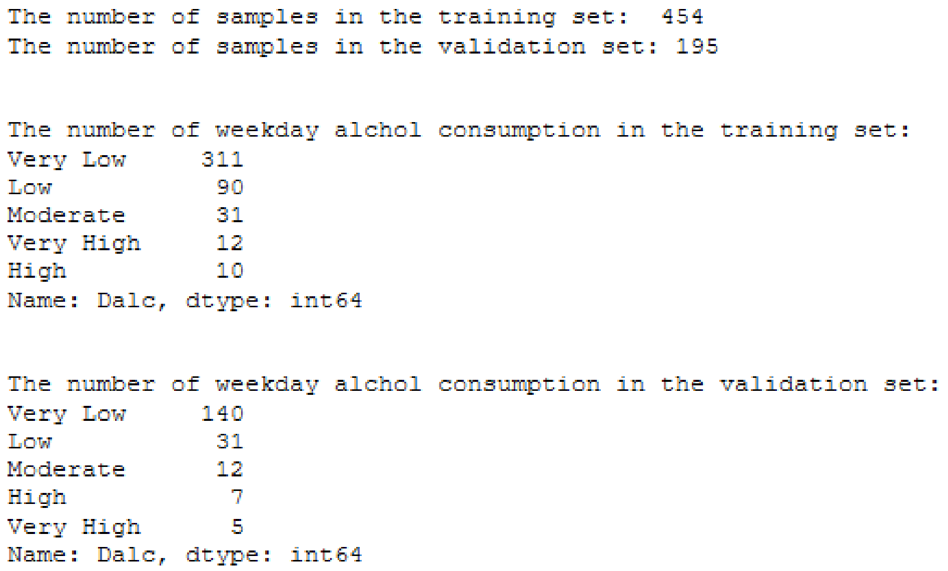


13. There were 8 features as seen below that was categorical data, which was difficult to do any analysis on. I changed the categorical features from yes and no answers to numbers (1’s and 0’s) to help run tests on these features.



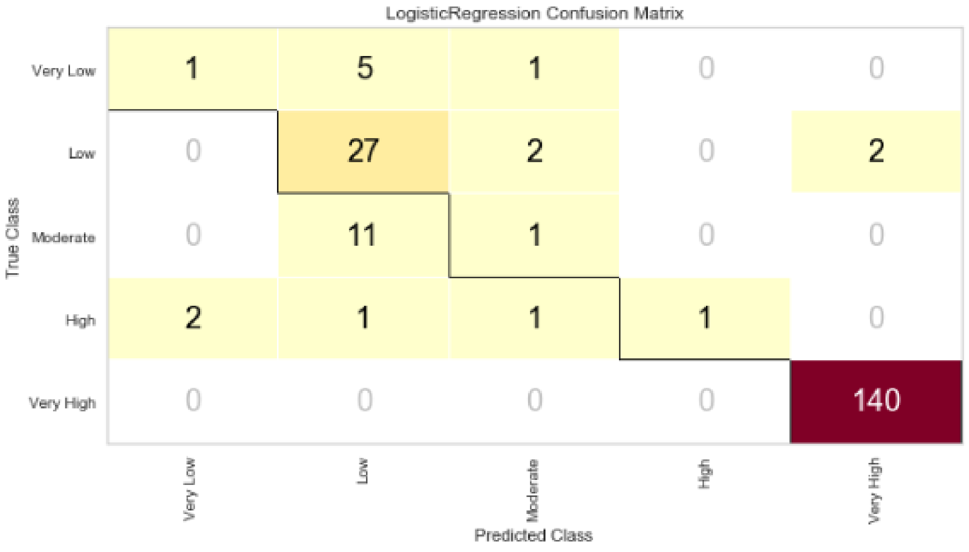
```
   schoolsup  famsup  paid  activities  nursery  higher  internet  romantic
0          1       0     0           0        1       1         0         0
1          0       1     0           0        0       1         1         0
2          1       0     0           0        1       1         1         0
3          0       1     0           1        1       1         1         1
4          0       1     0           0        1       1         0         0
5          0       1     0           1        1       1         1         0
6          0       0     0           0        1       1         1         0
7          1       1     0           0        1       1         0         0
8          0       1     0           0        1       1         1         0
9          0       1     0           1        1       1         1         0
```

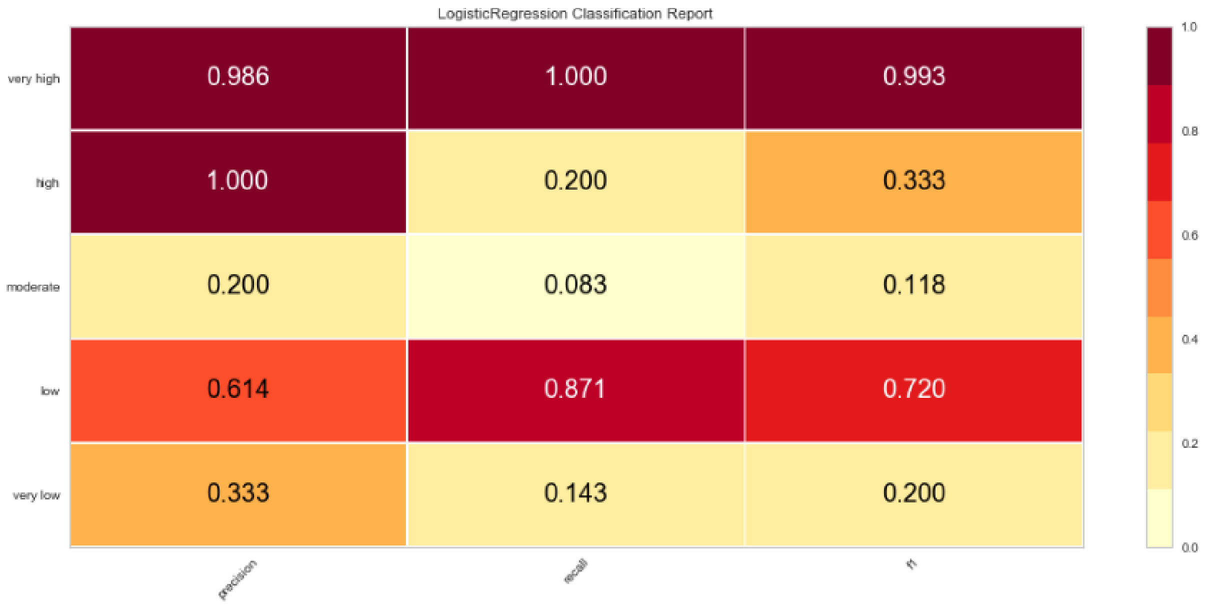
14. The next step is to split the data into groups, one for training and one for testing data. Afterwards I used the training data to predict the values in the test data.

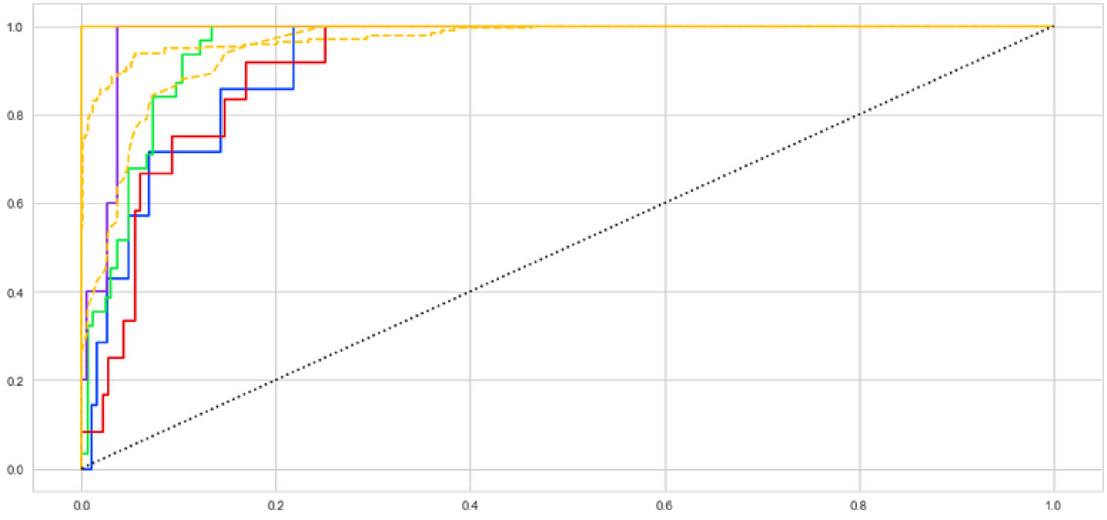


```
The number of samples in the training set:  454
The number of samples in the validation set: 195


The number of weekday alchol consumption in the training set:
Very Low     311
Low           90
Moderate      31
Very High     12
High          10
Name: Dalc, dtype: int64


The number of weekday alchol consumption in the validation set:
Very Low     140
Low           31
Moderate      12
High           7
Very High      5
Name: Dalc, dtype: int64
```

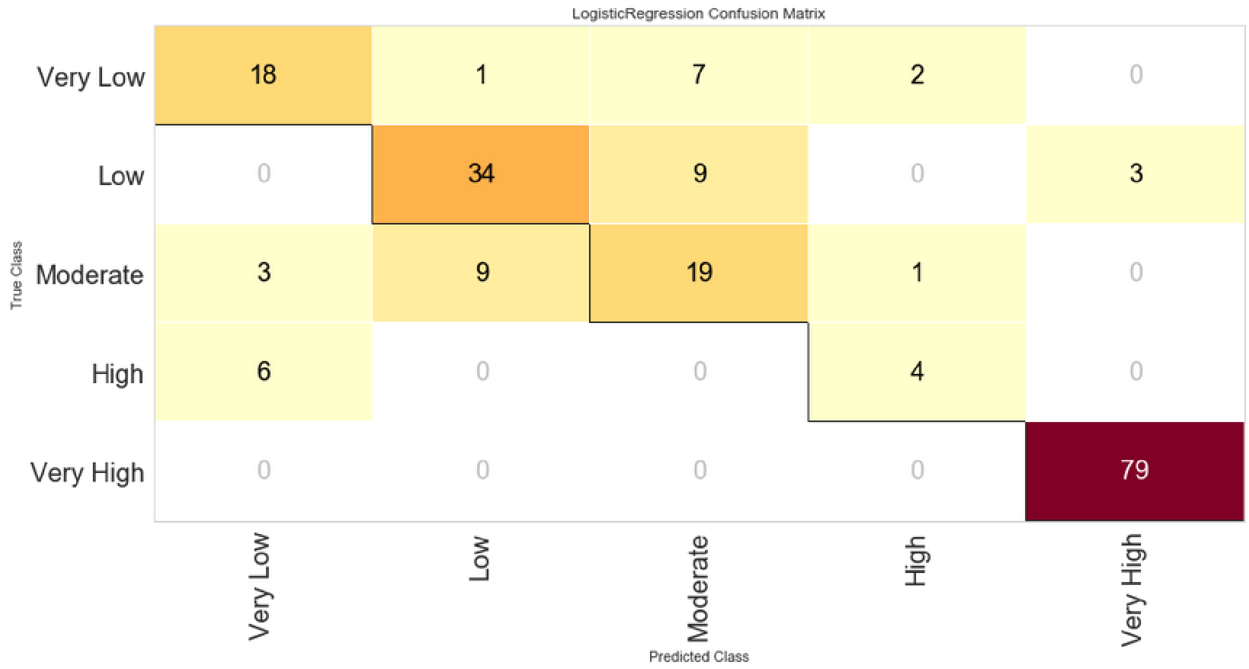
15. I tried to predict the weekday alcohol consumption of students. I used three different metrics for evaluation. Of the three tests, it appears that precision is the best slightly though all were good. It also appears like it is easiest to predict very high alcohol then low consumption for students compared to the other amount of alcohol consumption.
    a. Confusion matrix- 87%, which is a pretty good percentage.
    b. Precision, Recall, and F1 Score- All were good, with precision looking the best of the three.
    c. Roc Curve- Appears to be good as all the lines are above the dotted line meaning the probability is high for predicting all the different alcohol consumptions for students.


LogisticRegression Confusion Matrix
True Class
Very Low
Low
Moderate
High
Very High
1
5
1
0
0
0
27
2
0
2
0
11
1
0
0
2
1
1
1
0
0
0
0
0
140
Very Low
Low
Moderate
High
Very High
Predicted Class

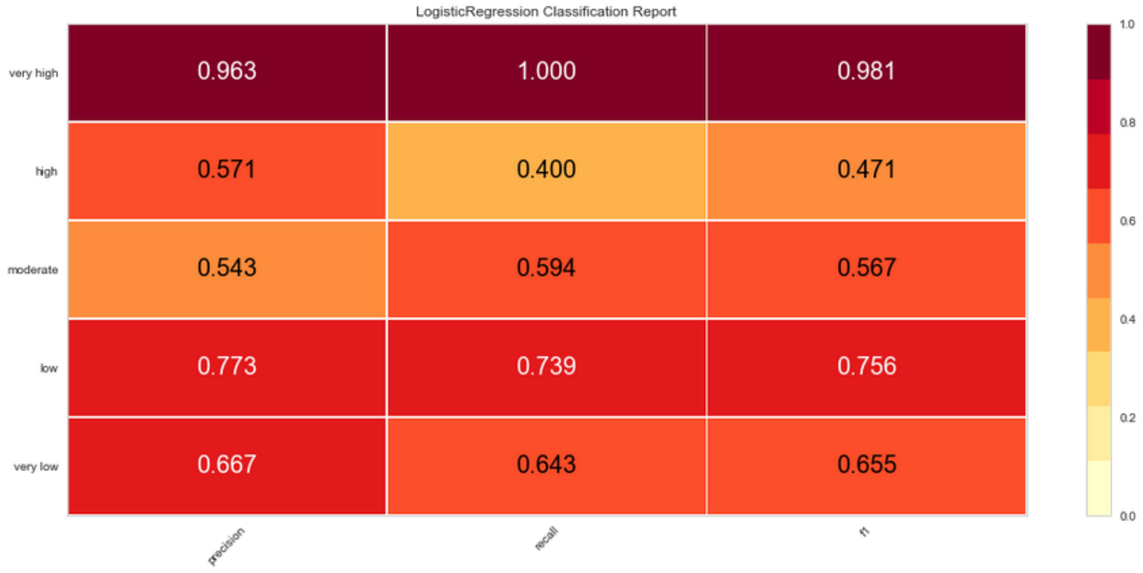

LogisticRegression Classification Report
very high
high
moderate
low
very low
0.986
1.000
0.993
1.000
0.200
0.333
0.200
0.083
0.118
0.614
0.871
0.720
0.333
0.143
0.200
precision
recall
f1
1.0
0.8
0.6
0.4
0.2
0.0


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

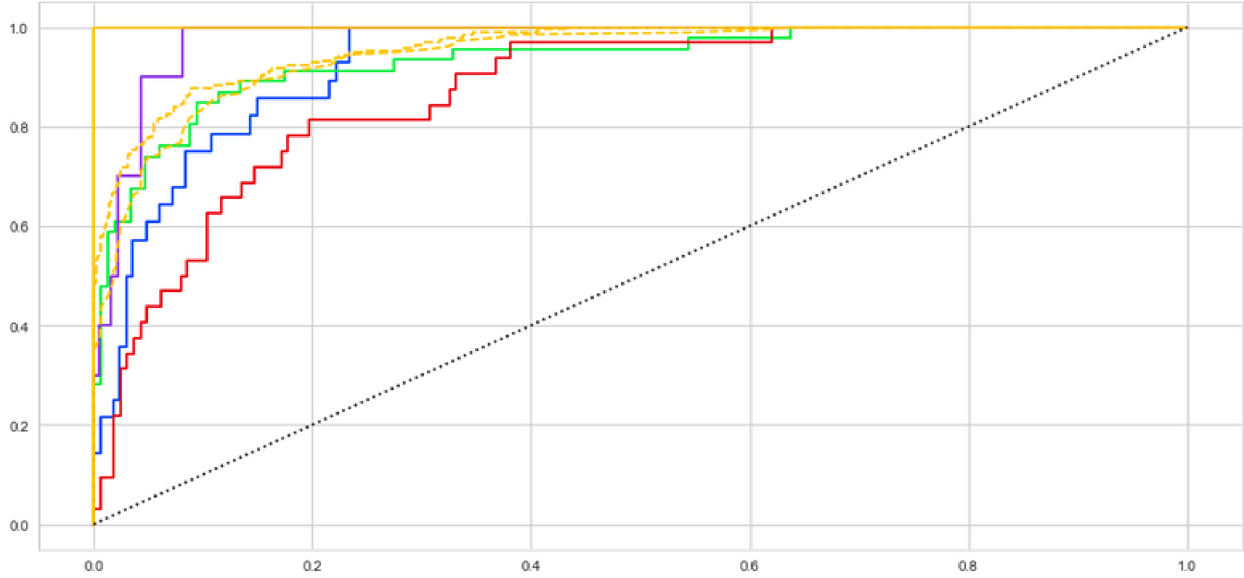
16. Next I tried to predict the weekend alcohol consumption of students. I used three different metrics for evaluation. Of the three tests, it appears that precision is the best slightly though all were good. It also appears like it is easiest to predict very high alcohol then low consumption for students compared to the other amount of alcohol consumption, just like results of the weekday alcohol consumption. It appears it is easier to predict, very low, moderate and high alcohol consumption compared to the weekday results.
    a. Confusion matrix- 79%, which is a pretty good percentage, but not as good as the weekday confusion matrix.
    b. Precision, Recall, and F1 Score- All were good, with precision looking the best of the three.
    c. Roc Curve- Appears to be good as all the lines are above the dotted line meaning the probability is high for predicting all the different alcohol consumptions for students. This Roc curve looks better than the Roc curve for weekday alcohol consumption.


LogisticRegression Confusion Matrix

True Class	Very Low	Low	Moderate	High	Very High
Very Low	18	1	7	2	0
Low	0	34	9	0	3
Moderate	3	9	19	1	0
High	6	0	0	4	0
Very High	0	0	0	0	79


LogisticRegression Classification Report

	precision	recall	f1
very high	0.963	1.000	0.981
high	0.571	0.400	0.471
moderate	0.543	0.594	0.567
low	0.773	0.739	0.756
very low	0.667	0.643	0.655



Conclusion:

Based on the analysis, there are several takeaways with the first being that there is a small correlation between age and absences. There was also a high correlation between the parent's education, which was a result originally not being looked at. Based on the factors of school support, family support, extra classes, activities, nursery, desire for higher education, if they have internet, and in a romantic relationship, it is easiest to predict if a student is a has a high alcohol consumption. After that, predicting students with low alcohol consumption are the next easiest to predict. For the rest of the alcohol consumption rates, it is more difficult to predict. It is easier to predict students alcohol consumption rate on the weekends compared to the weekdays, though it is not as accurate.