Name: Duong Nguyen

PSD: 1646288

Instructor: Christoph Eick

DATA MINING – PROJECT 1

Question 0:

Mean values of G1, G2, G3

G1 G2 G3

11.446178 11.625585 12.054602

Standard Deviations of G1, G2, G3

G1 G2 G3

2.723851 2.885887 2.962022

Quesion 1:

*Covariance matrix:*

G1 G2 G3

G1 7.419364 6.784504 6.852162

G2 6.784504 8.328344 8.084538

G3 6.852162 8.084538 8.773576

*Interpretations*: All G1, G2 and G3 have positive covariance, this means that for each pair, if an increase in one of the grades, there would be an increase for the other grade as well.

*Correlation matrix:*

G1 G2 G3

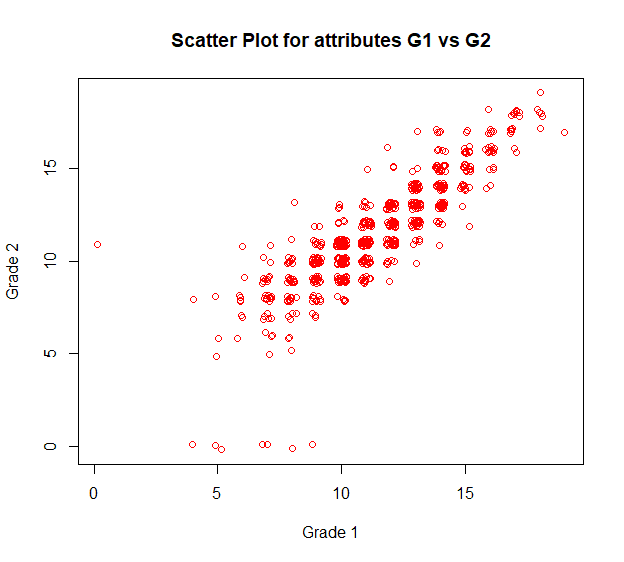
G1 1.0000000 0.8630887 0.8492899

G2 0.8630887 1.0000000 0.9457744

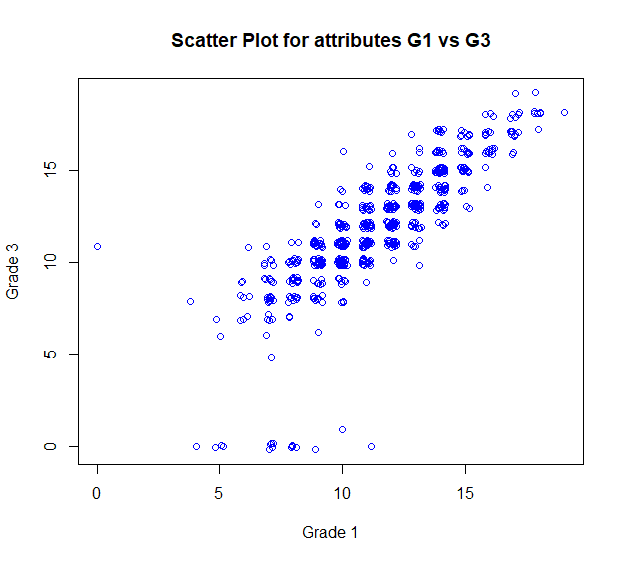
G3 0.8492899 0.9457744 1.0000000

*interpretation:* The diagonal values are always 1 because they represent the correlation between a variable and itself. From the table, the correlation of around 0.945 between G2 and G3 is almost perfect (meaning equal 1), this indicates a very strong positive linear relationship, we can say that if a student has a good grade 2, he or she is very likely going to get good grade 3 also. G1 and G2 also has a strong positive linear relationship as well, though not as strong as the one between G2 and G3. The linear relationship between G1 and G3 is around 0.849, which is also strong, but not as strong as the other 2 pairs' linear relationship.

QUESTION 2:



The scatter plot between G1 and G2 is what we expected, since we already calculated that the correlation between G1 vs G2 is a strong linearly. As grade 1 gets higher, Grade 2 tends to get higher also. We can also see some outliners, for example there is a student who missed/ made bad Grade 1 but he or she did okay on the the Exam 2. Besides there are some students who missed/got bad grade for Grade 2 also had done pretty badly on Grade 1.

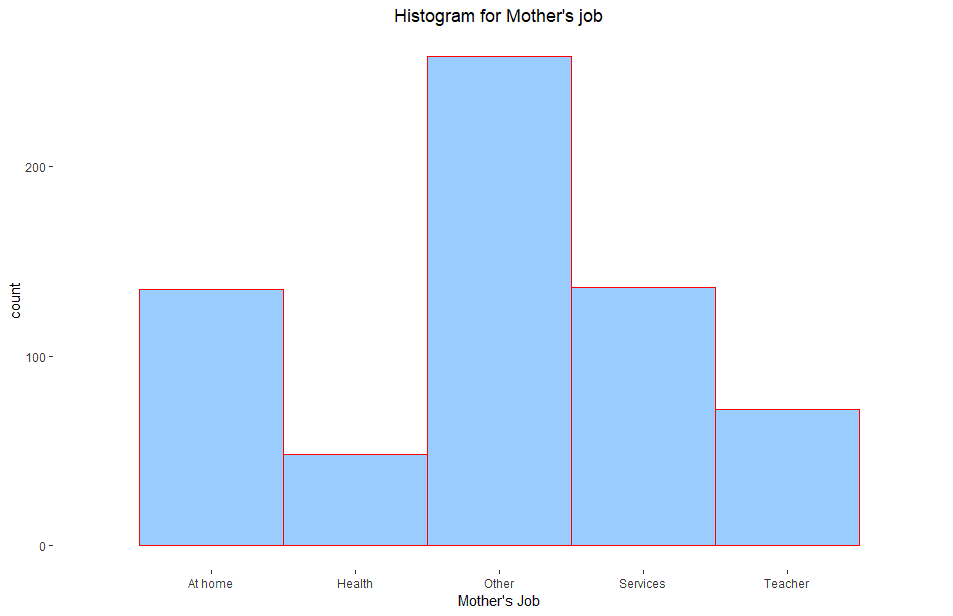


The scatter plot between G1 and G3 looks like the scatter plot of G1 and G2. It is what we expected also, since we already calculated that the correlation between them is also strong linearly. As G1 gets higher, G3 tends to get higher also. We can also see some outliners, for example there is a student who missed/ made bad Grade 1 but he or she did okay on the Exam 3, interesting thing is that happens to be the same student who had 0 for Grade 1 and 11 for Grade 2, he or she also got 11 for Grade 3. Furthermore, there are some students who missed/got bad grade for Grade 3 had previously done badly on Grade 1. And we also see there are quite a few more students that missed/did not do well on Grade 3 also.

QUESTION 3:

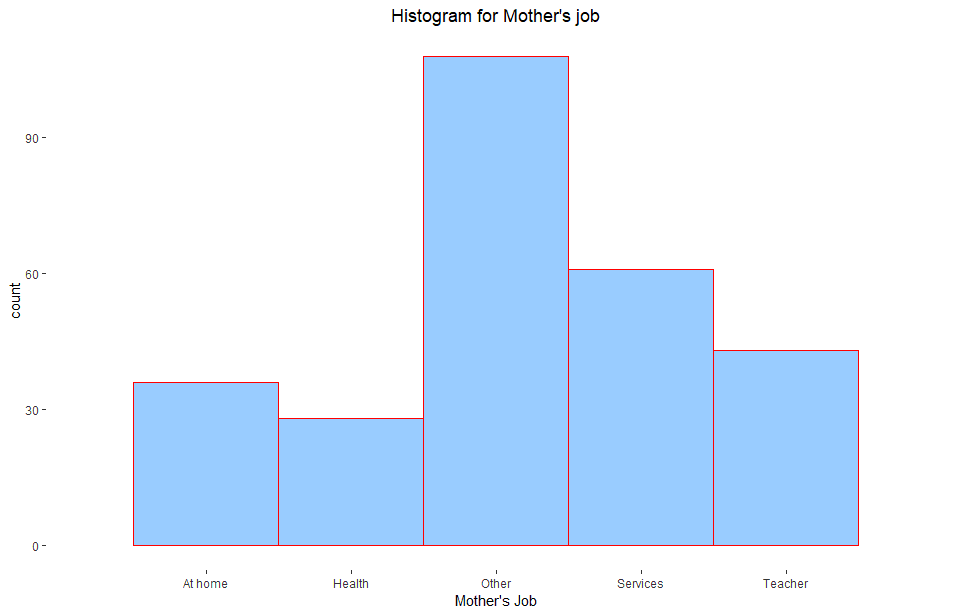
HISTOGRAMS FOR MOTHER'S JOB

***For All data points***



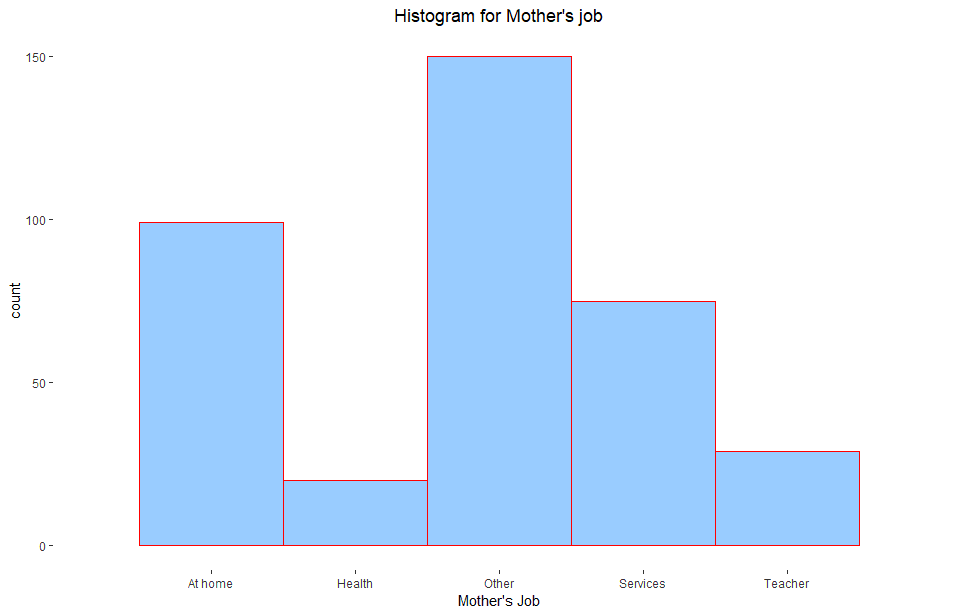
Among all the students, most of them have mother working in fields that are not specified. Not so many students have mother that works in Health. Around almost 100 students who have mother working as a teacher. While around 150 have mothers being housewife and 150 mothers working in services.

***For data points that G3 > 12***



For data points that G3 > 12, the proportions of mother’s jobs are pretty much the same with all data points proportion.

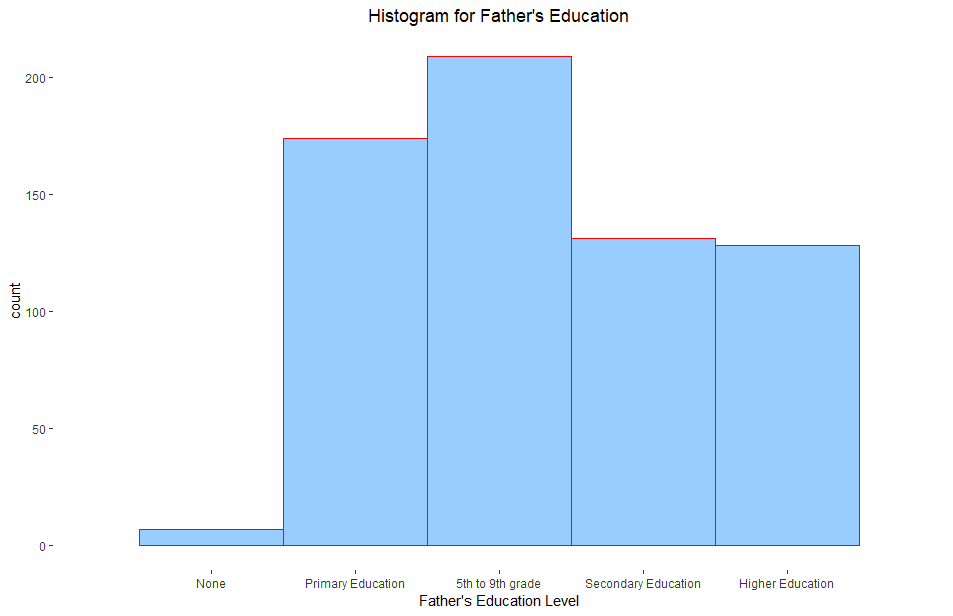
***For data points that G3 <= 12***



For data points that G3 <=13, there is a big number of mothers who are housewife, and not so many of them are teacher or work in health field.

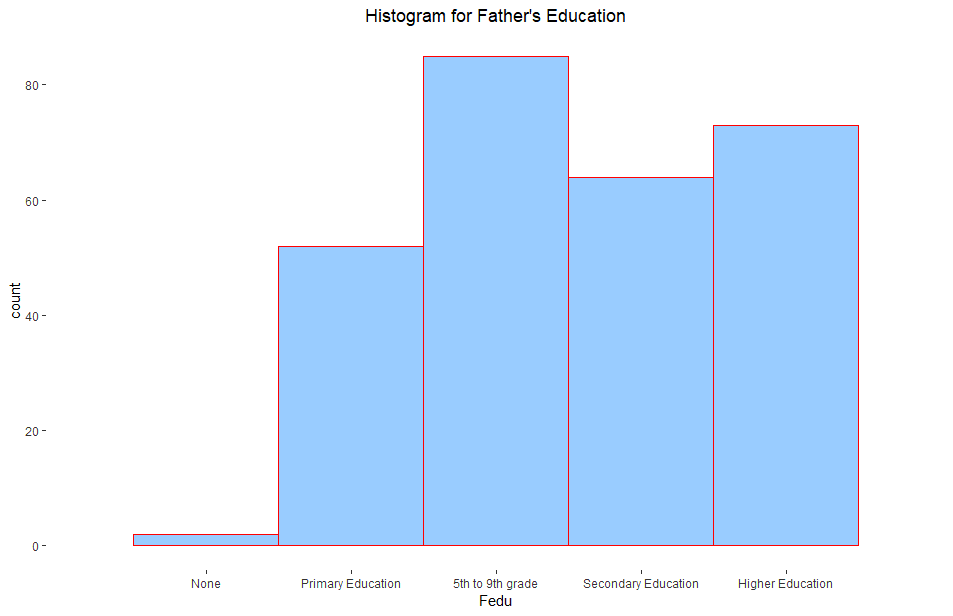
***HISTOGRAM FOR FATHER'S EDUCATION***

***For all data points***



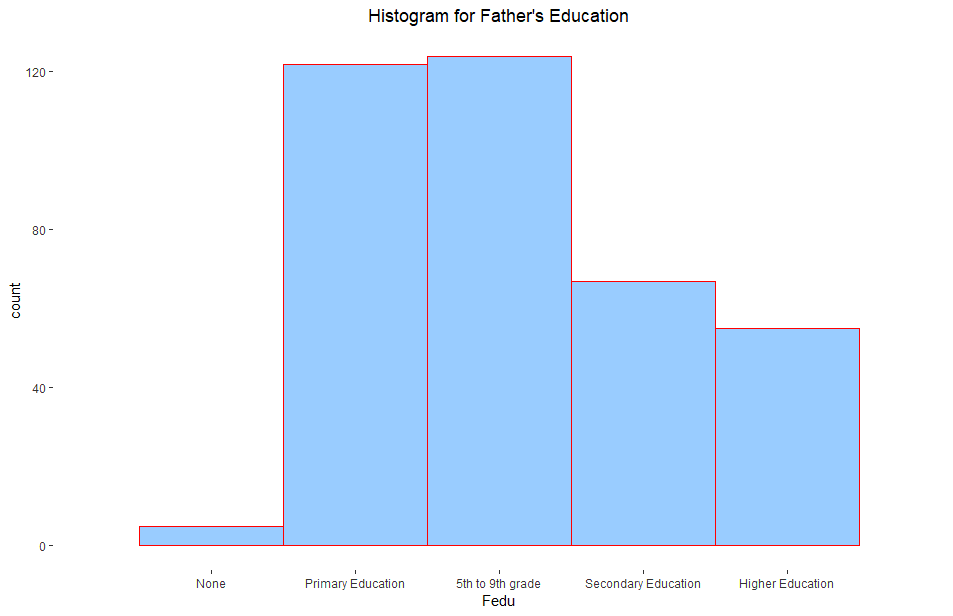
Most of the fathers were in or finished 5th-9th grade. And the ones who finished Secondary Education were likely to finish higher education as well. There are just some of them who did received none of the education.

***For data points that have G3 > 12***



In the group of student who have G3>12, a lot of their father finished higher education. And there are bigger proportion of fathers who finished secondary and higher education than ones who only finished Primary Education, opposite to the data showed in the diagram of all data points.

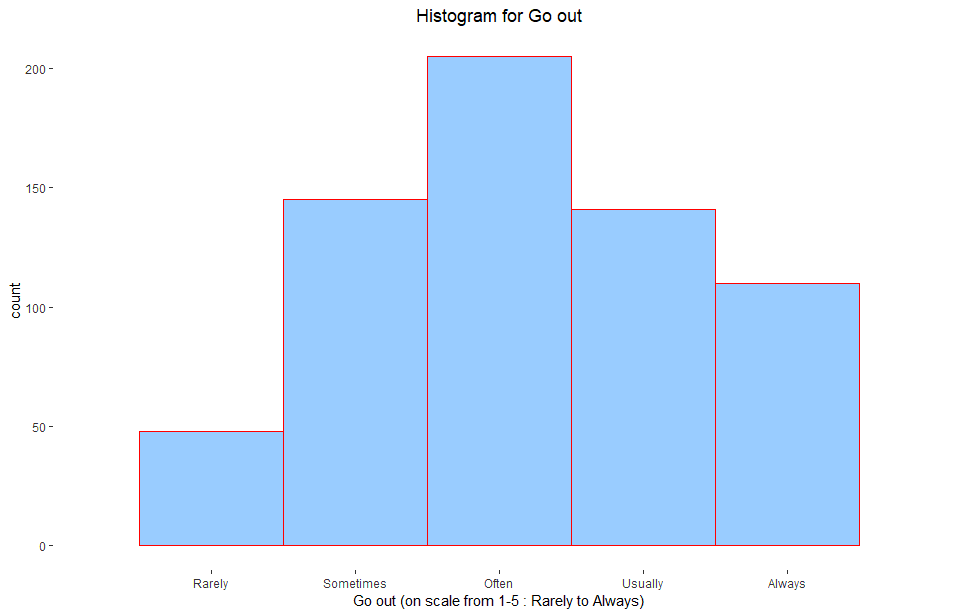
***For data points that have G3 <= 12***



For group of students who have G3 <= 12, there are less fathers who finished Secondary and or Higher education than ones who finished Primary and 5-9th grade.

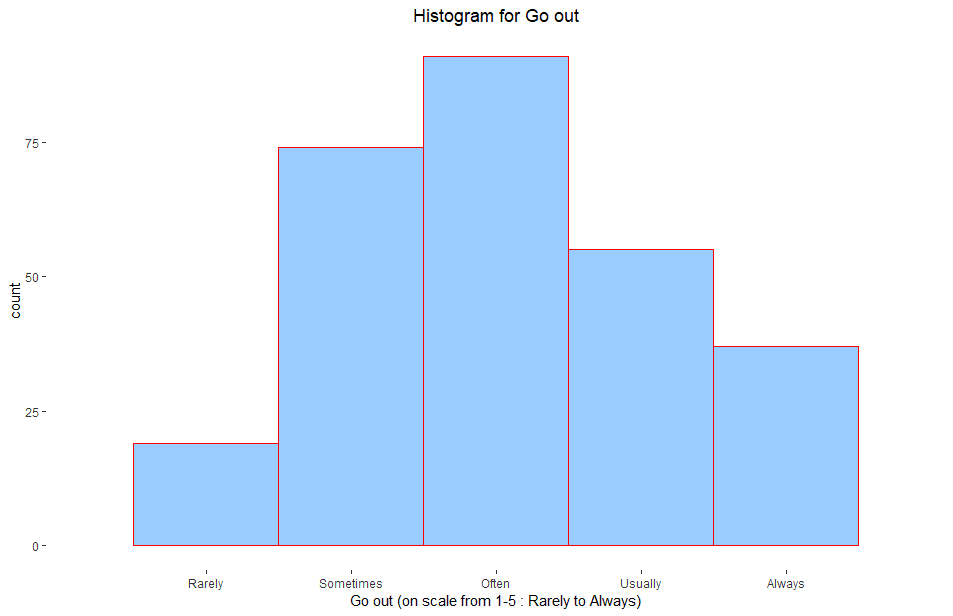
**HISTOGRAMS FOR GO OUT**

***For all data points***



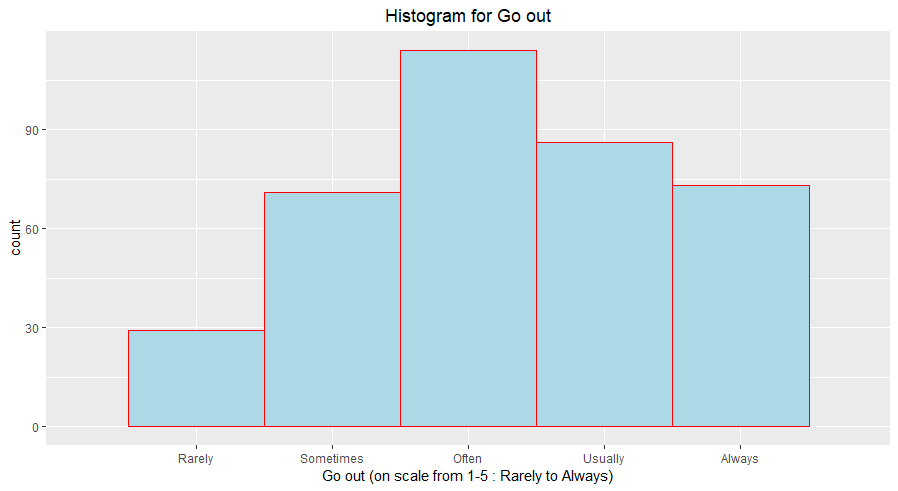
Most of the students do go out often. Around same number of them either go out sometimes or usually. There are more students always go out than rarely go out.

***For data points that have G3 > 12***



The distribution is pretty similar to distribution for all data points.

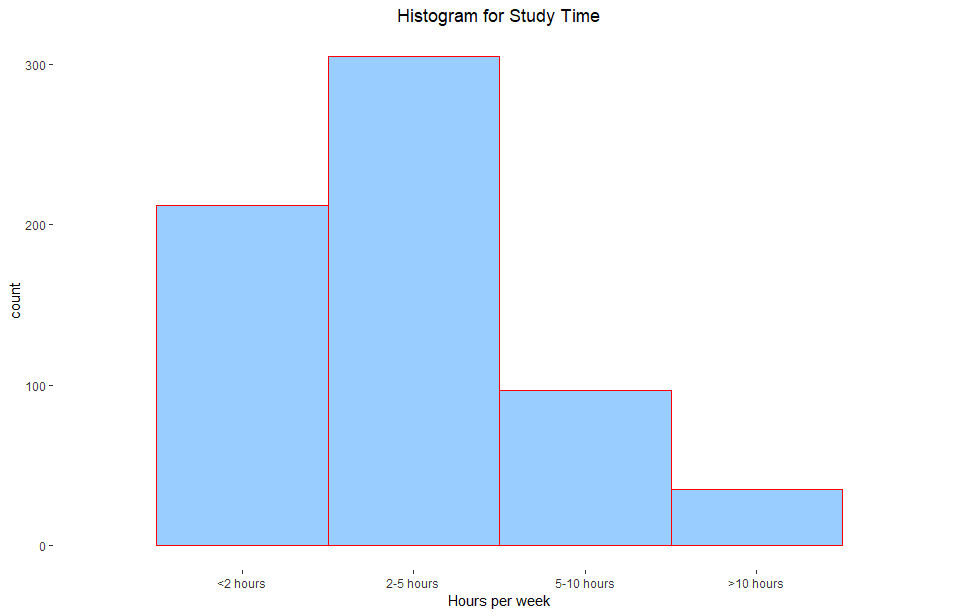
***For data points that have G3 <= 12***



Pretty similar to the diagram of all data points and of the group that have G3 > 12, except for there are more students in this group go out usually and always.

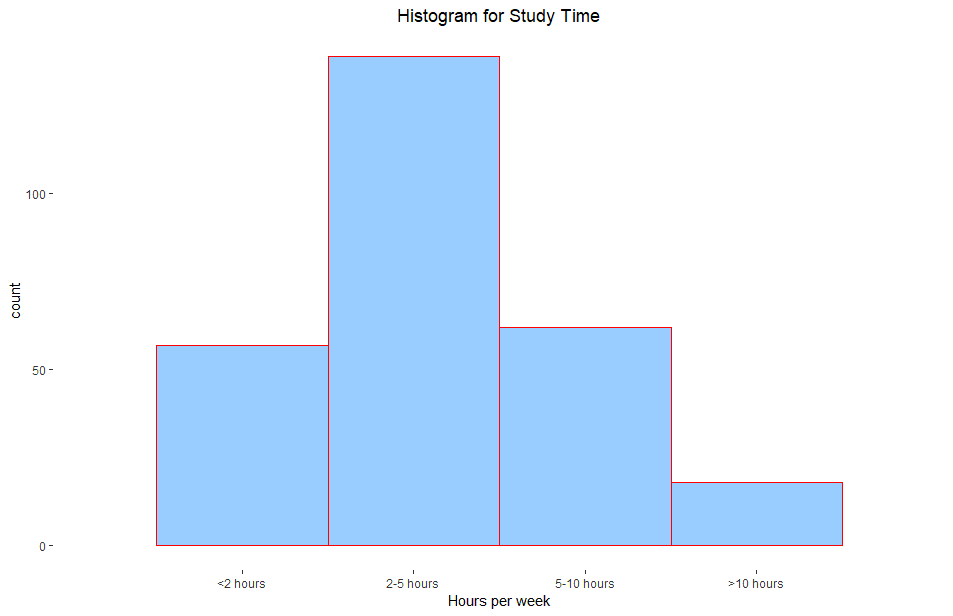
**HISTOGRAMS FOR STUDY TIME**

***For all data points***



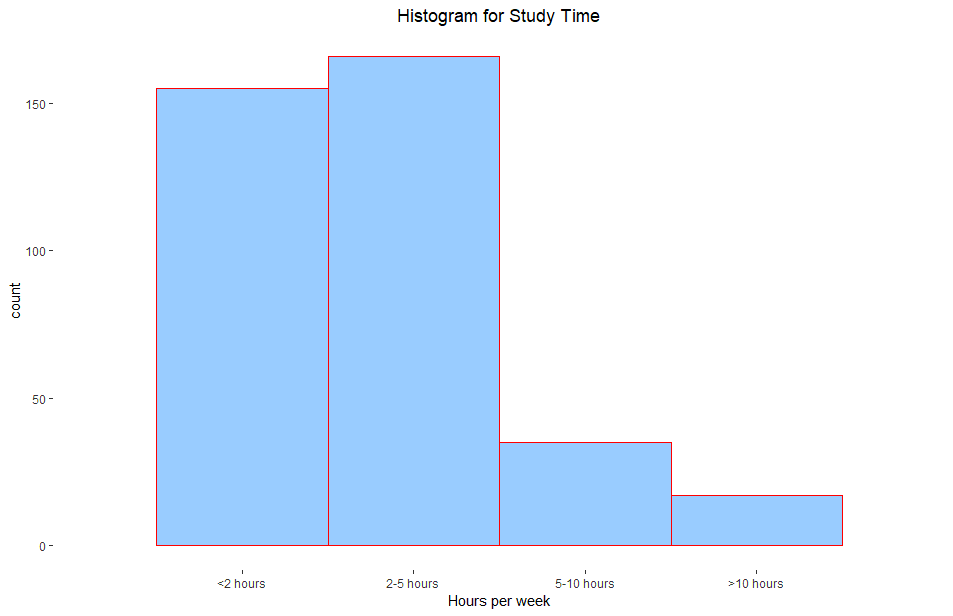
Many students spent less than 2 hours to study. Most of the students spent 2-5 hours to study per week. We can see not too many students spend more than 10 hours to study.

***For data points that have G3 > 12***



There are as many students who get G3 > 12 spent < 2 hours to study as students who spent 5-10 hours a week. However, most of them did spend 2-5 hours a week.

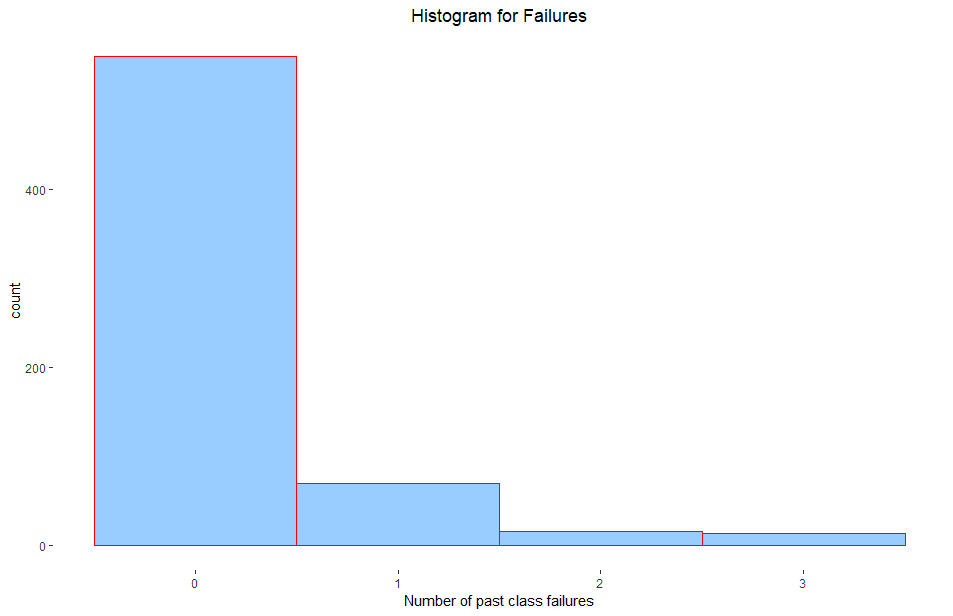
***For data points that have G3 <= 12***



A lot of students who have G3 <=12 spent less than 2 hours a week. Most of them spent 2-5 hours. Some of them spent > 10 hours a week, sadly their grade was still not desirable.

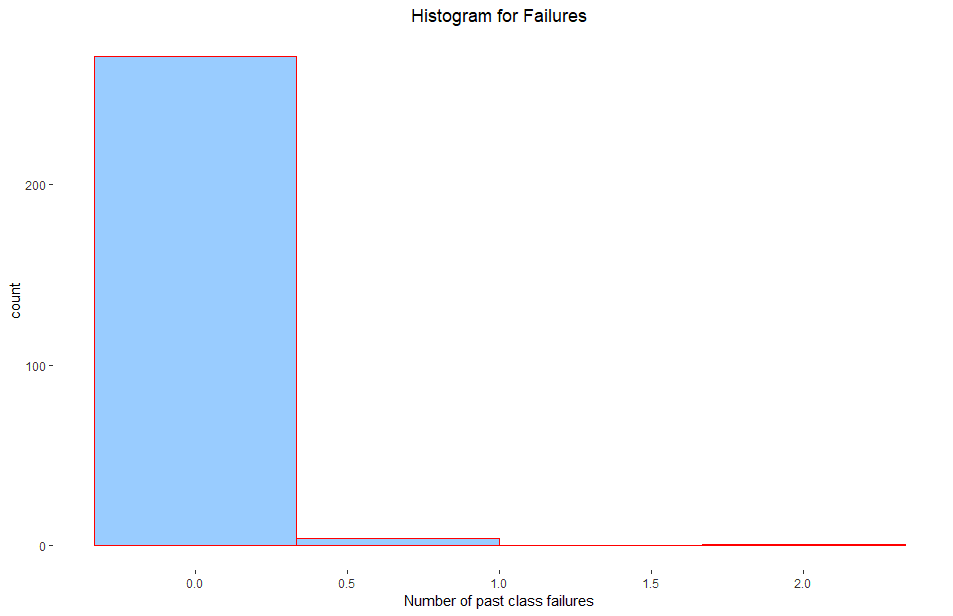
***HISTOGRAMS FOR FAILURES***

***For all data points***



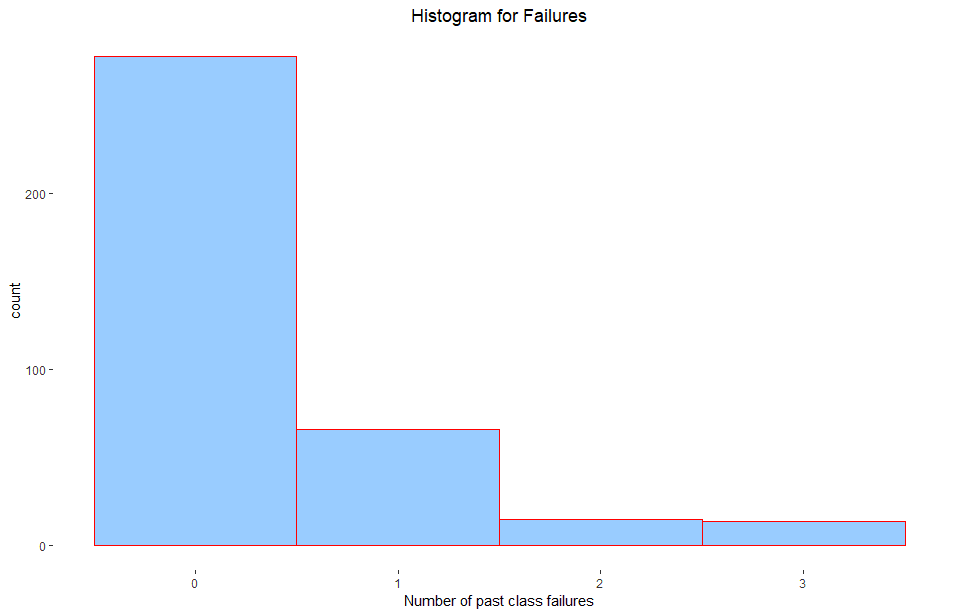
Most students have never failed a class before. Some of them failed more than 3, 4 times.

***For data points that have G3 > 12***



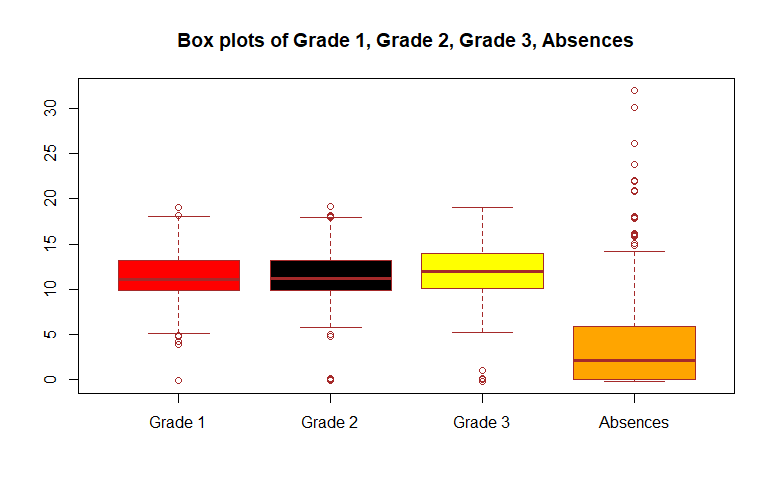
Most of students who have G3 > 12 has never failed a class before. There are barely any students of this group of higher grade have failed a class more than once.

***For data points that have G3 <= 12***



There are students who never failed before got G3 < 12. A few of them did fail more than three times. This distribution looks similar to general distribution for all data points.

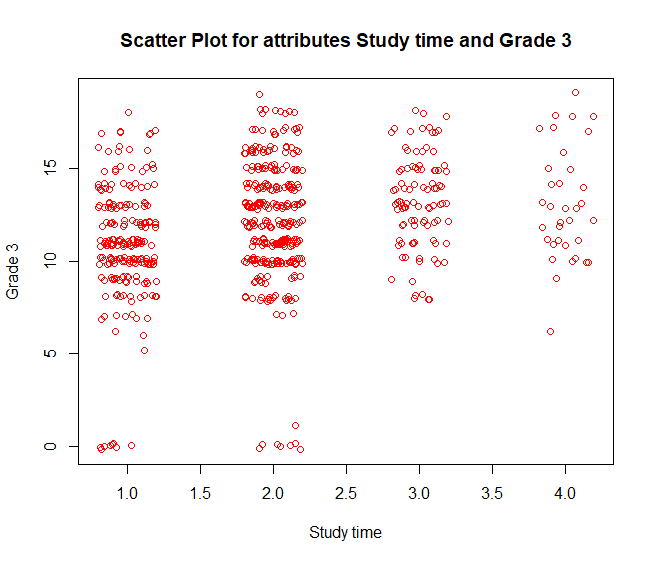
***QUESTION 4:***



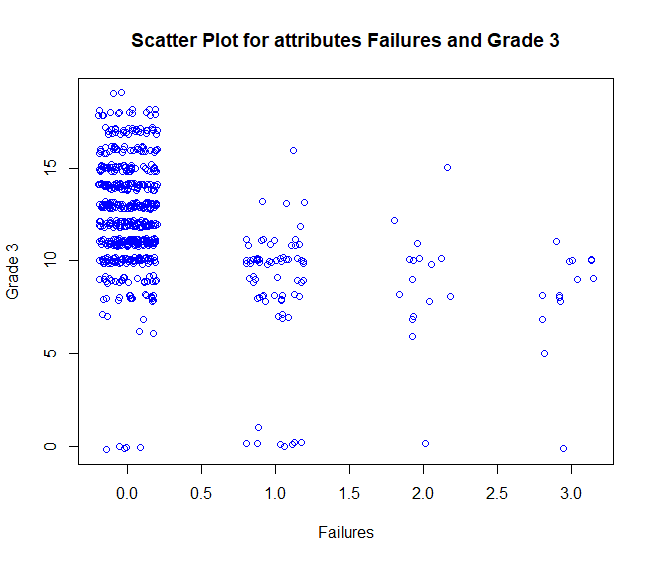
From the boxplots, we can see that Grade 1, Grade 2 and Grade 3 have pretty much the same mean, same range even though Grade 1's mean is slightly smaller and Grade 3 has a little bit wider range and interquartile range. Grade 1 and Grade 2 seem to have the same interquartile range and the outliers. Top 25% of Grade 1, Grade 2 and Grade 3 is around 13 or 14.

The mean of Absences is quite small (==2), which is a good thing, however, there also are a lot of outliers that means there are quite several students that take many absences. About 75% of students take 6 or less absences, and the rest 25% take somewhere between 6- and 32-days missing classes.

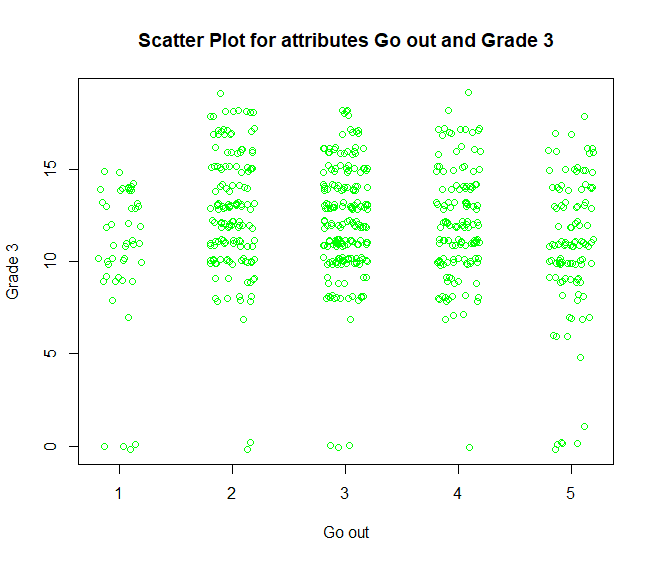
***QUESTION 5:***



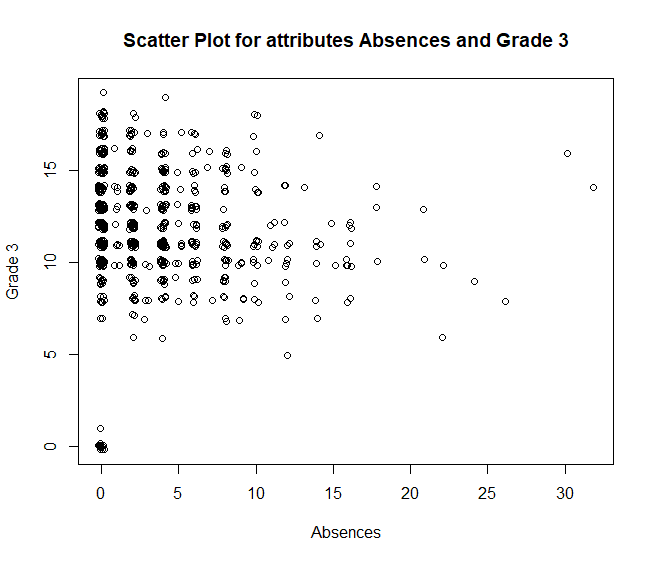
From the plot, there are a lot of students spend around less than 5 hours weekly to study (scale 1 represents < 2 hours and 2 represents <5 hours) and most of them got average Grade 3, students who spent less than 5 hours earned better Grade 3 than ones that spent less than 2 hours. Some of them made bad grades, especially there are quite a few students who spent between 2 and 5 hours a week make less than 3 for Grade 3. Students who spent more than 5 hours weekly studying made good grade generally. None of the students who studied more than 5 hours got below 5, however, there are still some made less than 10. Nonetheless, averagely, those students tend to earn higher than Grade 3's mean value.



Most of the students had never failed a class before, and the students that got best grade 3 are also in this group of students. Yet there are still a small number of them who never failed a class did get 0 or below 5 for Grade 3. The groups of students who failed classes more than once seem to make not a very good grade, as by looking at the plot, lots of them were making grade below the expected grade, even though there are some exceptions. For group of students who had failed class 4 times tend to make below average.

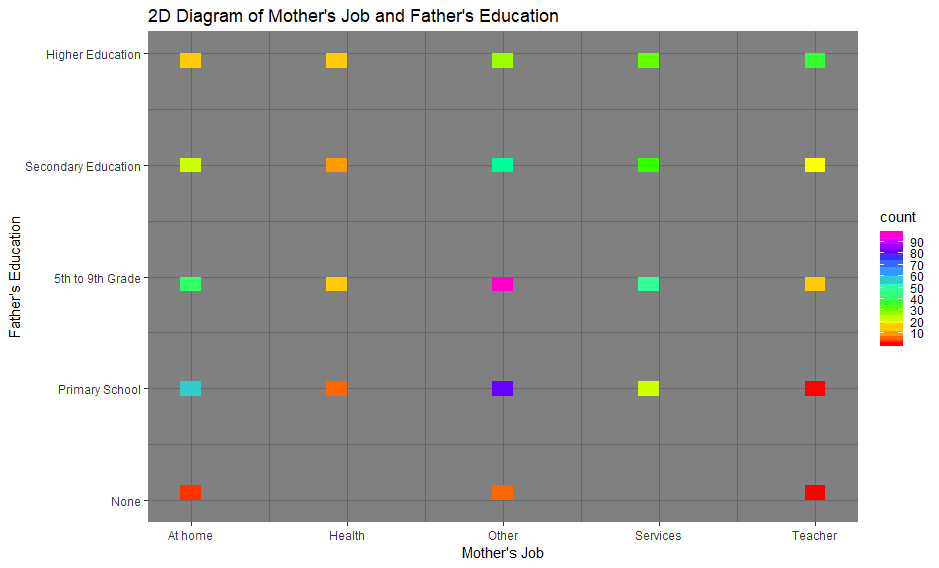


From the plot, there are students who barely go out that made bad grades and also some students who go out alot made bad grade. Without regard to how much time the students spending going out, it does not indicate the student would get a low grade if he or she goes out too much and vice versa. The group of students rarely go out made fair grades, but none of the highest grade was made by a student in this group. It does look like students who go out often, they make good grades. There are many students of the group for scale 2,3 and 4 made a very good Grade 3. It may not be obvious but more than half of the student in the group that always go out made less than 13.



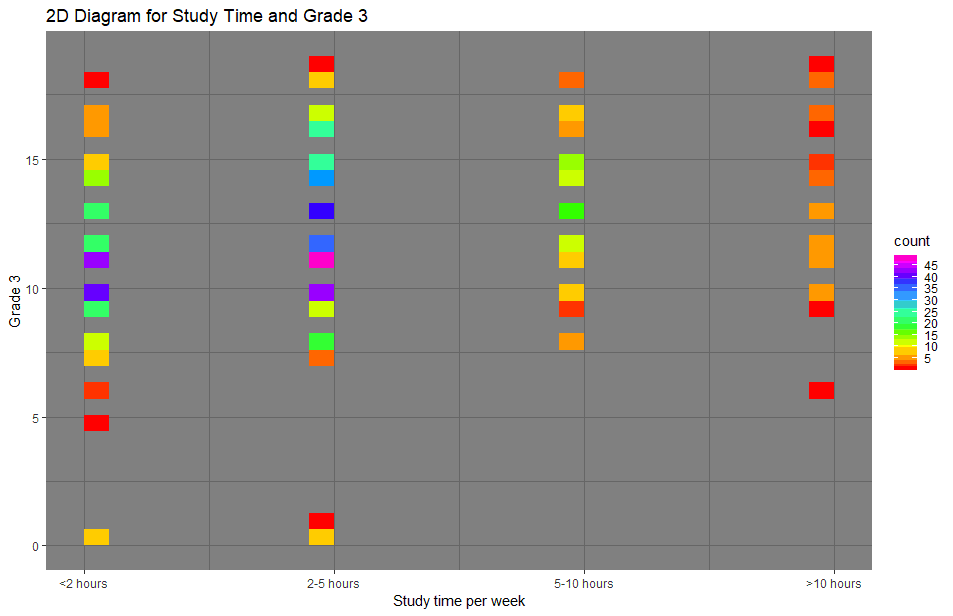
Many students took less than 6 absences. There are many students who took zero days off made better than average Grade 3. One would expect that student who took many absences would make bad grade, it is not a case here. All of them who made very low grade (below 5) had never missed a class. Students who take more than 20 days absent made below average, though none of them made grade that less than 5. Unusually, there are students who took more than 30 absences made above average grade.

***QUESTION 6:***



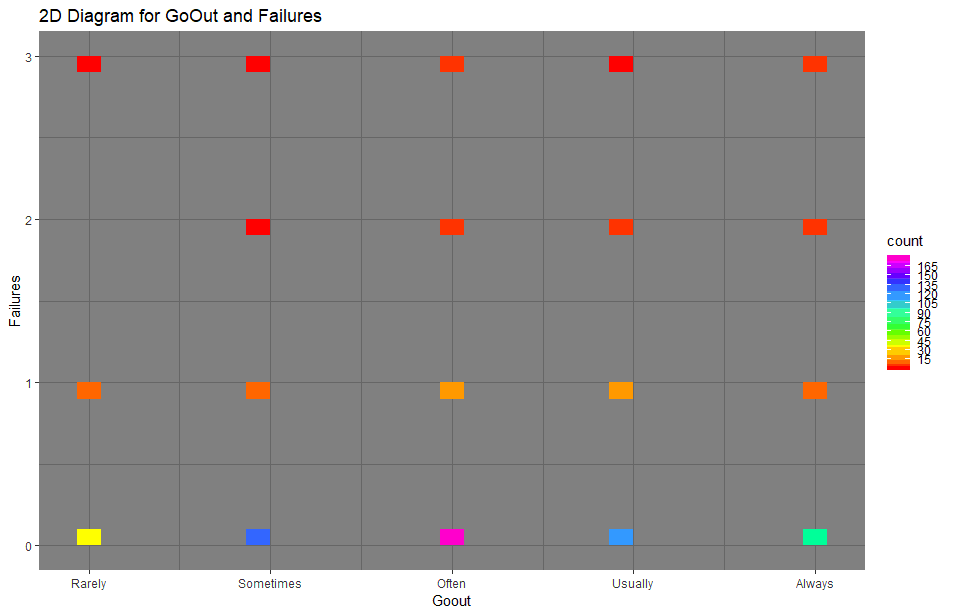
There are a lot of data points of Mother’s job other and Father’s education 5th to 9th grade. This makes sense if we look at the diagrams above showing that there are a big proportion of mothers reported their job as other and fathers who have done 5th to 9th grade. This may not be related but there are no couples that the wife work in Health or Services married a man with no education background.

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------



If we look at the purple squares, there are many counts of students who spend 2-5 hours a week, this confirms with the histogram in question 3. If we only look at the part containing students have grade more than 15, there are more students who spent 2-5 hours than student spent > 5 hours. That may indicate that study more hours do not always guarantee a better grade. However, there are some students who studied less than 5 hours got a bad G3 (below 5), while none of the students studying for more than 5 hours got G3 below 5.

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------



Most students going out often and had never failed a class. While surprisingly, there are quite many students who always go out had never failed a class. Therefore, going out often is not a good indicator of a bad performing students. Rather, students who spend their time outside with their friends and family tend to do well at school.

***QUESTION 7: FIT LINEAR MODEL***

lm(formula = G3 ~ ., data = grades)

Coefficients:

(Intercept) Fedu Studytime Failures Goout

0.06844 -0.02219 0.09130 -0.24836 -0.04151

Absences G1 G2 Mjob\_at\_home Mjob\_health

0.02400 0.13906 0.88234 0.01599 0.12872

Mjob\_other Mjob\_services Mjob\_teacher Reason\_course Reason\_home

-0.13646 0.04659 NA 0.16078 0.07250

Reason\_other Reason\_reputation

-0.33883 NA

After converting the two nominal attributes Mjob and reasons into binary, trying to fit a linear model to predict G3, we recieved NA for Mjob\_teacher and Reason\_reputation. This mean those two variables are highly correlated to some other variables that we won't get better prediction/ new information if we include them. I tried to drop those two columns and the model gives the exactly same coefficients:

Call:

lm(formula = G3 ~ ., data = grades\_test)

Coefficients:

(Intercept) Fedu Studytime Failures Goout Absences G1

0.06844 -0.02219 0.09130 -0.24836 -0.04151 0.02400 0.13906

G2 Mjob\_at\_home Mjob\_health Mjob\_other Mjob\_services Reason\_course Reason\_home

0.88234 0.01599 0.12872 -0.13646 0.04659 0.16078 0.07250

Reason\_other

-0.33883

After normalizing the dataset, I've gotten a new linear model:

Call:

lm(formula = normalized\_grades$G3 ~ ., data = normalized\_grades)

Coefficients:

(Intercept) Fedu Studytime Failures Goout

2.623e-17 -7.555e-03 2.344e-02 -4.561e-02 -1.511e-02

Absences G1 G2 Mjob\_at\_home Mjob\_health

3.448e-02 1.182e-01 7.958e-01 2.010e-03 1.043e-02

Mjob\_other Mjob\_services Mjob\_teacher Reason\_course Reason\_home

-2.069e-02 5.874e-03 NA 2.472e-02 9.445e-03

Reason\_other Reason\_reputation

-3.296e-02 NA

Again we see NA for Mjob\_teacher and Reason\_reputation.

Rsquare for the normalized dataset = 0.8535742.

After dropping G2 and 5 other insignificant attributes, we obtain a new linear model:

lm(formula = normalized\_grades$G3 ~ ., data = normalized\_grades)

Coefficients:

(Intercept) Studytime Failures Goout Absences

6.349e-17 2.910e-02 -9.427e-02 -2.669e-02 3.869e-02

G1 Mjob\_other Mjob\_teacher Reason\_course Reason\_other

7.809e-01 -3.472e-02 -5.568e-03 1.596e-02 -5.059e-02

Reason\_reputation

1.408e-02

Interpretation: After we drop 5 most insignificant attributes, the coefficients of some of the remaining attributes were getting bigger (G1, Failures...) , some got less significant (Absences, Reason\_course...). The positive coefficients tell us that the relationship between dependent and independent variables is a positive linear relationship, which mean if a positive-coefficient factor increases, the response variable also increases, while negative coefficients mean that if an independent variable is decreasing/increasing then the response variable will be increasing/ decreasing. For example in our model achieved above, G1 seems to have the strongest linear relationship with G3, so that if we know how the student did on Grade 1, we can pretty much predict how he or she would do on Grade 3, and this was confirmed from question 2. Failures has a negative coefficient, meaning that a student who failed more, tend to make worse grades, which makes sense. The coefficient of Goout is also negative, also for Mjob attributes and Reason\_other attributes; however, for these attributes, the coefficients are really small, so they do not play a very important role in predicting G3.

The Rsquared for second model = 0.6973308.

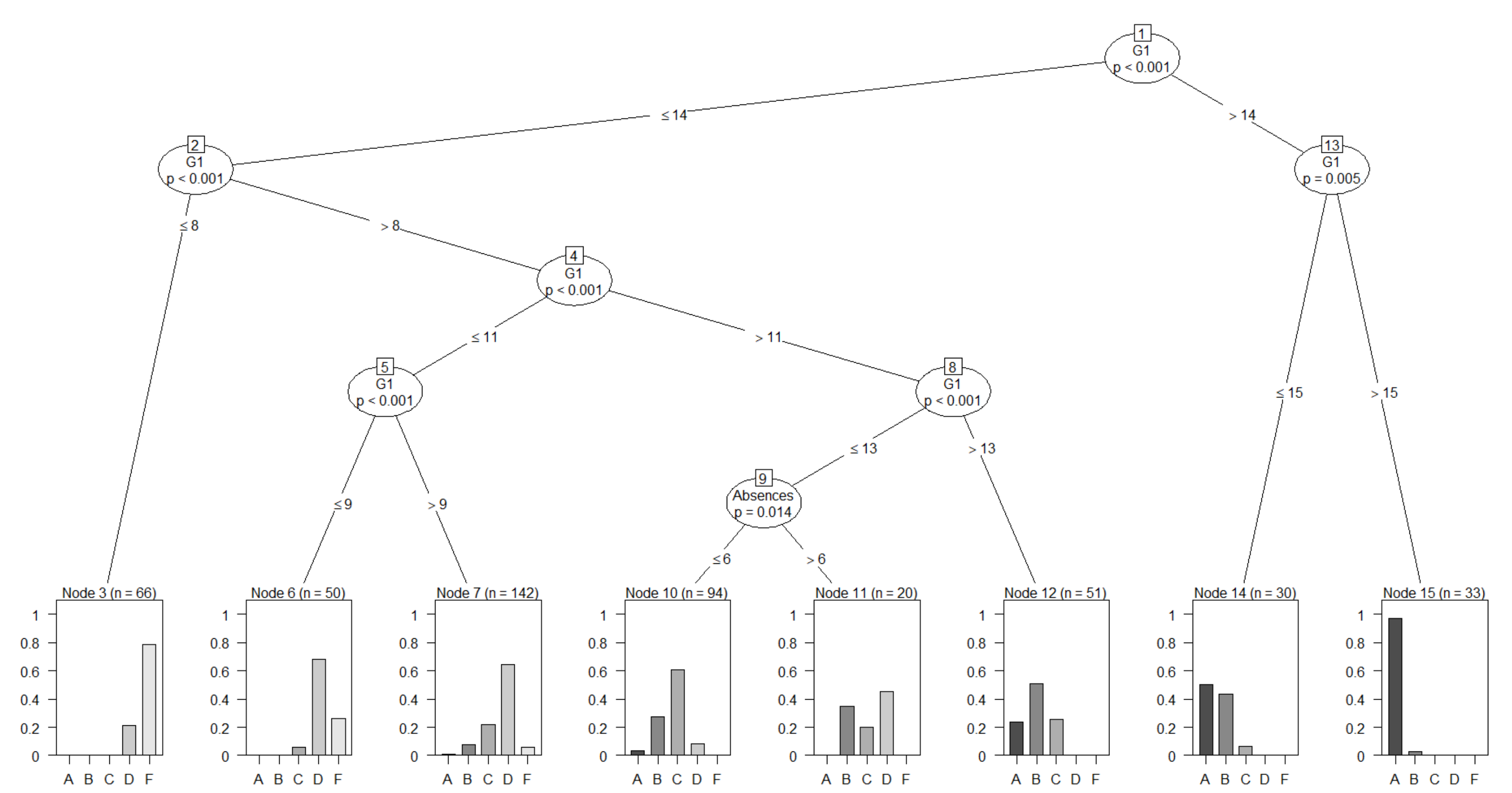
Comparing the 2 models: For the first model, the Rsquare is .853574, it is much higher than Rsquared from the second model. .853574 is high and close to 1, which indicates that the model we've just obtained is a pretty good linear fit to the data, also, one can say that the model accounts for 85% of the variance. For the second model, the independent variables only explain an estimated of around 69.7% of the variation in G3. However, having a high Rsquared does not always prove that the model is correctly specified. In order to get a more accurate result of how good a model is, we will need to explore further with other model diagnostic tools.

***QUESTION 8:***

***First model*** using ctree() from library package

I split the original dataset into 4 portions, 3 for training(train\_data) and 1 for testing (test\_data).

Below is the decision tree for train\_data:

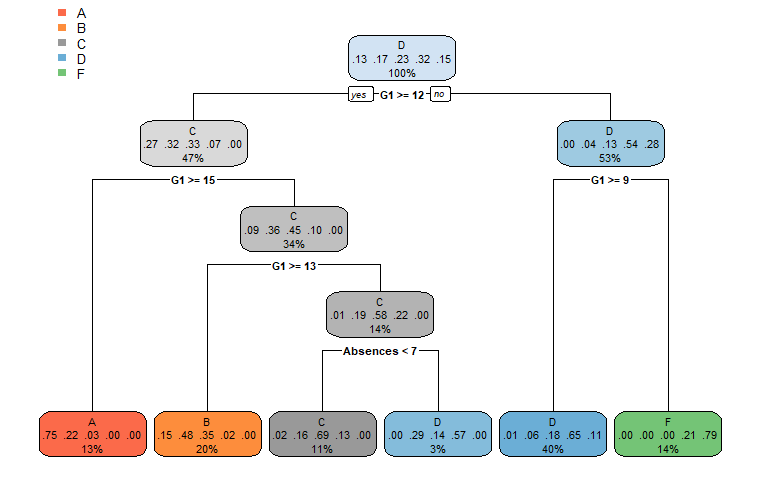


There are 15 nodes and it only used G1 and Absences among all 8 attributes to predict the Grade, and most of the results were predicted using G1 only, which means the other 6 attributes did not play an important part in predicting the Grade.

By looking at the tree, the student would have a very high chance (looks like more than 90%) to make an A. If he or she made between 14 and 15 in G1, he or she would likely make an A or B. It is probable for a student to make a C if he or she made in between 9 and 13. He or she would have around 70% chance to make a D if his/her G1 is between 8 and 9. If a student has G1 of less than 8, an F is very likely.

After training the dataset, the model obtained the accuracy rate of 65% for train\_data, however, when apply it to test\_data, the accuracy rate is only 59.5%. Thus there is 40% chance we will wrongly predict the Grade using this model.

***Second model*** using rpart() of library rpart(), and also using package rpart.plot to produce the tree picture



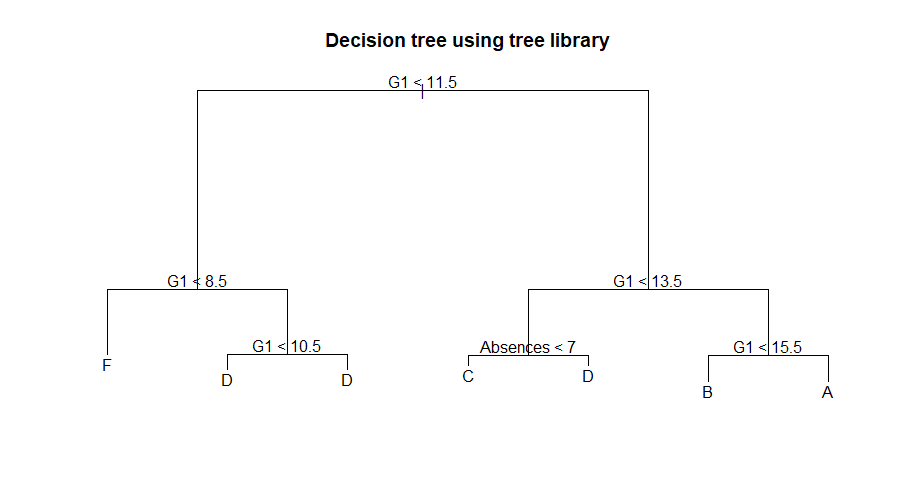
There are 8 nodes in the tree. And decision tree carried out using rpart() only used attributes G1 and Absences to predict G3, just like ctree above. And interesting thing is when it comes to the node where they use Absences attribute, they most likely are indicating the student would have grade C. So for rpart tree, first node where it has all the students, and most of them are having D. Then it started to branch the tree by asking the question of if Grade1 >= 12 or not, if yes, then it separates most students with Grade C and above to one group, and below C to another group, and the process keeps going till it reaches the acceptable results. By looking at the table, there would be 75% chance for a student to get an A if he or she made more than 15 on Grade 1. 48% chance to make a B if he or she make the grade in between 13 and 15. If he or she made less than 13 but more than 12 and have less than 7 absences, he or she would be likely making a C (69% chance). Otherwise, a student would make a D if he or she made less than 13 and have more than 7 absences or made more than 9 on Grade 1. If none of the cases above happened, then there is a 79% chance a student would make an F, 21% chance make a D, and barely a chance to make A, B or C.

Accuracy rate for rpart for train data = 0.6502058

Accuracy rate for rpart for test data = 0.5337423

Surprisingly that the accuracy rate for train data using rpart and ctree are almost the same, however, ctree did a better prediction job on test data when compare to rpart.

***Third model***



Again, this tree model also used G1 and Absences as classifiers. For this model, it predicts a student would get an A if his or her G1 >= 16, seemingly regardless of number of absences. The student will get B if he or she made either 14 or 15. If a student made 12 or 13, if he or she have more than 7 absences, there is a high chance of getting a D, otherwise a C is predicted. C and D are also possible if a student made 10 or 11. If he or she made less than 9, an F is very possible.

Compare to the other 2 models, the plot of this model looks much simpler. It does not show the probability for each outcome. But when we compare the accuracy rates, it did as a good job as other two models. However, it did a slightly better job than the other two models on test data.

Accuracy rate for train data using tree library = 0.6460905

Accuracy rate for test data using tree library = 0.601227

In summary, when it comes to predict G3, while we collected lots of data on different attributes, only previous grades as well as absences seem to matter, other attributes did not play a significant role.

***QUESTION 9:***

The dataset contains the information about student performance in Portugal during 2006-2007. The goal of the assignment as well as an original intention of collecting and exploring those data is to predict which factors would affect G3, which is a grade given upon completion of Secondary education in Portugal. Prior to analyzing the data, one would expect that G1 and G3 would strongly correlate with G3, and it did show after doing some statistics calculations: the correlation of around 0.934 between G2 and G3 is almost perfect, the linear relationship between G1 and G3 is around 0.866, which is also strong. However, besides G1 and G2, there were also other factors that affect G3. By applying some exploratory data analysis techniques (calculate correlation, covariance, plot diagrams, boxplots, etc), there have been some interesting findings. People would often think that if a student goes out often, he or she may not do very well, the data showed the opposite. The group of Portuguese students who often achieve good grades are group that go out often, and almost half of all students who hang out so frequently did have good grades. During doing the data analysis, we also saw that there is not a big difference between G1 G2 and G3 distribution. They tend to have the same mean and spread. Besides G1 and G2, G3 tends to have been affected by how often a student is absent as well. Most students that did not take too many absences tend to do well while students who took too many absences (>20 absences) tend to do poorly. Mother’s and Father’s job as well as their education level do not significantly affect the student’s performance at school. This was proved while fitting a linear model to predict G3 as well as while running decision trees to predict G3. In my opinion, to find out the true factors that affect student’s grades, we will need to continue collecting data on the existed attributes, as well as try to find other factors that would possibly influence the student’s study work and collect data on them. For example, because absences do affect the student’s performance, we can try to find out with he or she would take too many absences: is this because the classes are too boring? Is this because they must work too much and do not spend enough time to study? Is this because they were not able to get extra help outside of classrooms that made them fail? If we can address those problems, we can make changes to better the student’s success rate. Besides, for students that are successful, we can ask questions to find out what made they want to achieve good results, compare that with other struggling students so that we can influence them, to help them.