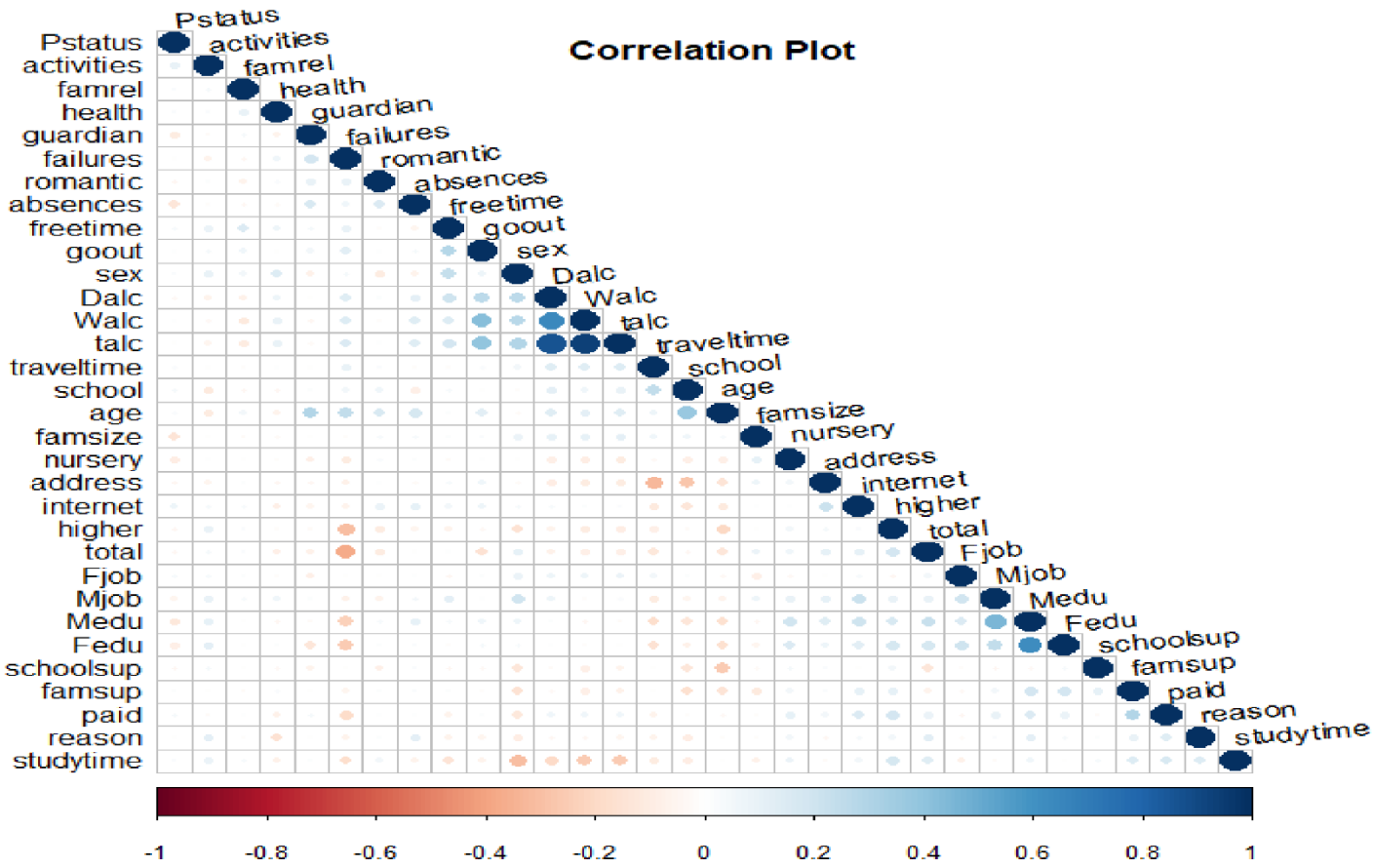
**STUDENT ALCOHOL CONSUMPTION**

- Team Dataholics (Himani Bhatt, Venkatesh Aditya, SriKrishna Sridhar, Mrunal Chaudhary)

## I) INTRODUCTION:

For this course project, we have decided to do exploratory data analysis for the 'Alcohol Consumption Data' which we collected from Kaggle website. The dataset was obtained after conducting a survey of 395 students' math courses in two secondary school from Portugal for the year 2005-06. It contains a lot of interesting social, gender and study information about students. Through this project we aim to see how alcohol consumption and other variables like mother's education, travel time, going out with friends, and past failures affect the overall grade of a student.




As can be seen from the correlation plot above, there are 33 variables in the dataset. When we plotted the correlation graph for it, it can be clearly seen that the failures, travel time, mother's education (Medu), and total alcohol consumption have a good correlation with the response variable 'total grades'. Also we verified and confirmed our results using the Boruta package which gave us the top 5 most correlated variables. A description of the variables in the dataset has been mentioned below on which we will be doing out analysis:

## II) DATA DESCRIPTION:

1. **total grades (totalgrades):** This is the response variable for our analysis. We defined this column as the sum of the grades of three exams conducted in the course of the semester and scaled it to 100 to get a percentage value. This variable has an almost normal distribution.

2. **Total Alcohol Consumption (Talc):** This is a predictor variable and our main point of interest. In the dataset we have two columns, 'weekly alcohol consumption' and 'daily alcohol consumption'. We tried to use these two as separate predictor variables and analyzed our results, only to conclude that both of them have an almost same effect. Hence we decided to combine these two variables, and make a new predictor variable 'Total alcohol consumption (Talc)' which we will be using in our analysis. The new variable is an ordinal variable with values from [1,5] 1 being 'very low consumption'

to 5 being 'very high consumption'. The new variable being the sum of 2 variables is in the order of 2-10. The variable so obtained has a highly skewed distribution.

3. **travel time (traveltime):** This is a predictor variable, which is again ordinal in nature. It depicts the amount of time taken to travel to school by a student. The values range from [1,4] and their explanations as given below:

1- less than 15 minutes, 2-> 15-30 minutes, 3-> 30 minutes to 1 hour, 4-> more than 1 hour. This variable has a highly skewed distribution.

4. **Mother's education (Medu):** This is a predictor variable, which is again ordinal in nature. It depicts the educational status of the student's mother. The values range from [1,4] and their explanations as given below:

0 -> No Education, 1 -> primary, 2 -> 4-9$^{th}$ grade, 3- secondary, 4 –> higher education. This variable has a highly skewed distribution.

5. **Going out (goout):** This is a predictor variable, which is again ordinal in nature. It depicts the how often a student goes out with his/her friends, family and others. The values range from [1,4] with one being doesn't go out at all, to 4 being frequent out goers.

6. **Failures (failures):** This is a predictor variable, which is ordinal in nature. It depicts past failures faced by a student. The values vary from [0,4], with the explanations given below:

0: no failures experienced, 1: student failed in 1 course, 2: student failed in 2 courses, 3: student failed in 3 courses, 4: student failed in 4 or more courses

Now that we have a brief data description available with us, we can proceed to the next section, the research questions which we aim to answer.

## III) RESEARCH QUESTIONS:

- How significantly does alcohol consumption affect total grade of students?
- What other factors impact total grade of students?
- What factors influence alcohol consumption?

In this report we present our analyses for the above questions. We have done bivariate, trivariate, categorical (since all our predictor variables are categorical) and hypervariate analysis. Let us have an in depth look at them.

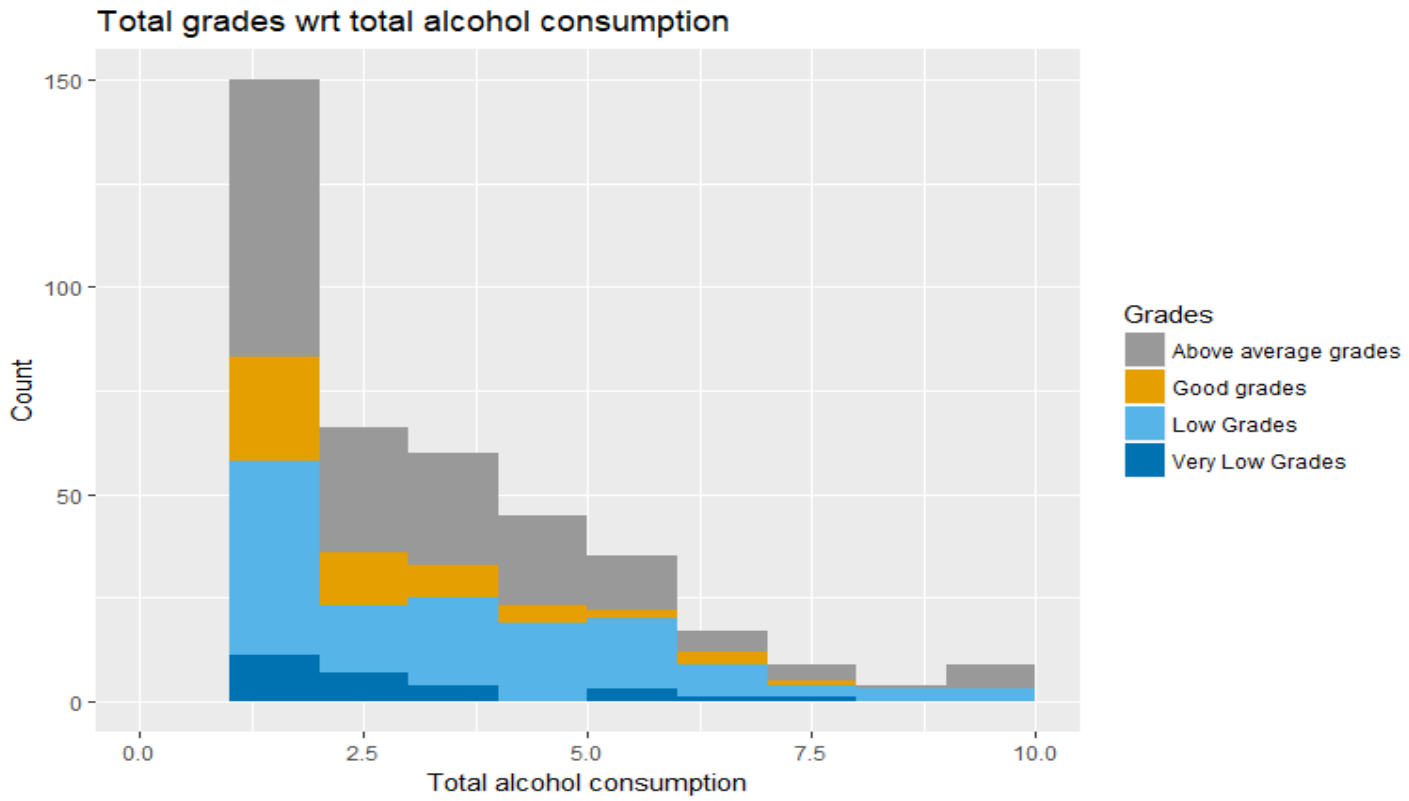
## IV) EXPLORATORY DATA ANALYSIS

### BIVARIATE ANALYSIS

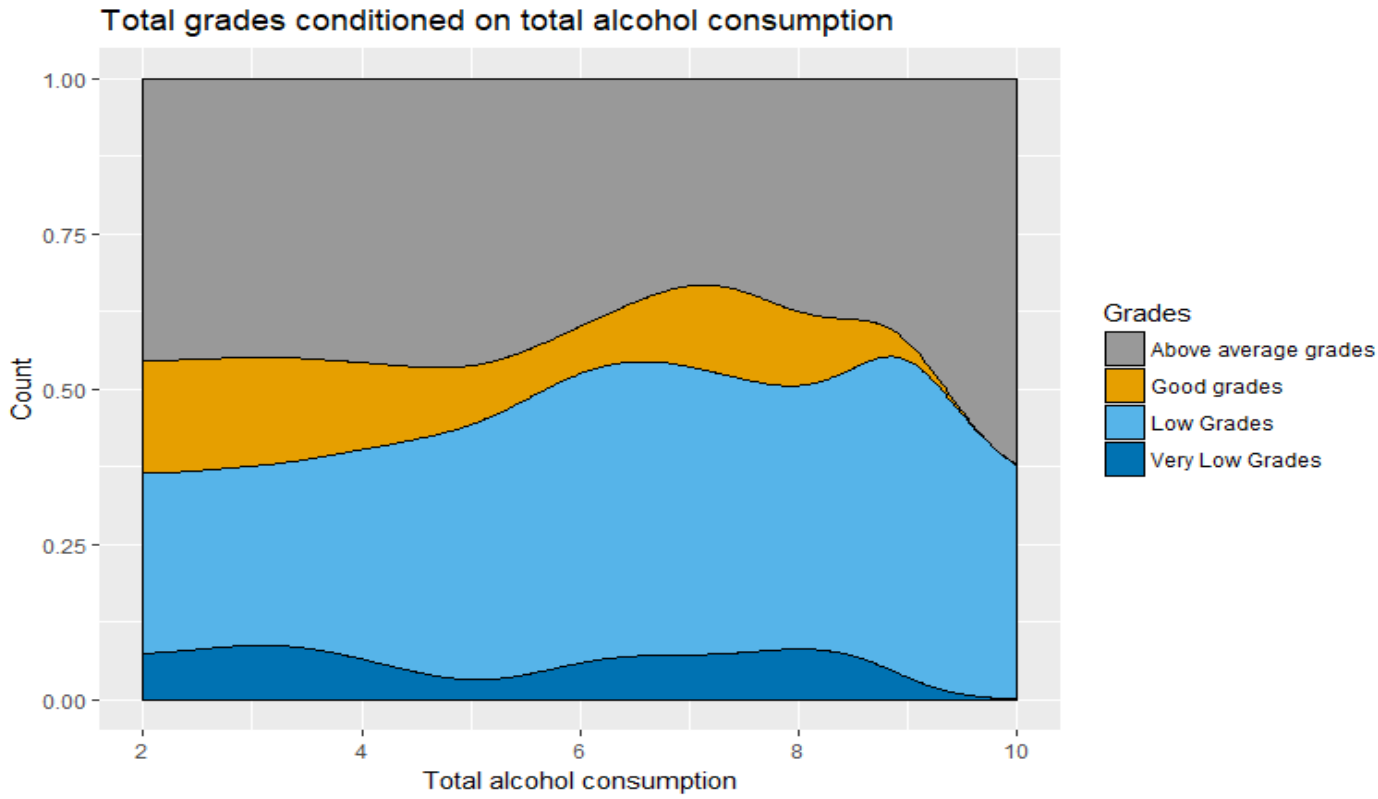
### Variation of Total Alcohol consumption with Total grades:

First we have divided our response variable, total grades into four categories – Very Low Grades, Low Grades, Above average grades and Good grades with respect to each quantile.

The below plot shows the conditional distribution of total alcohol consumption given the category of grades of the student. Here we have discretized total alcohol consumption. From the below plot we can see that the distribution of total alcohol consumption is highly right skewed.

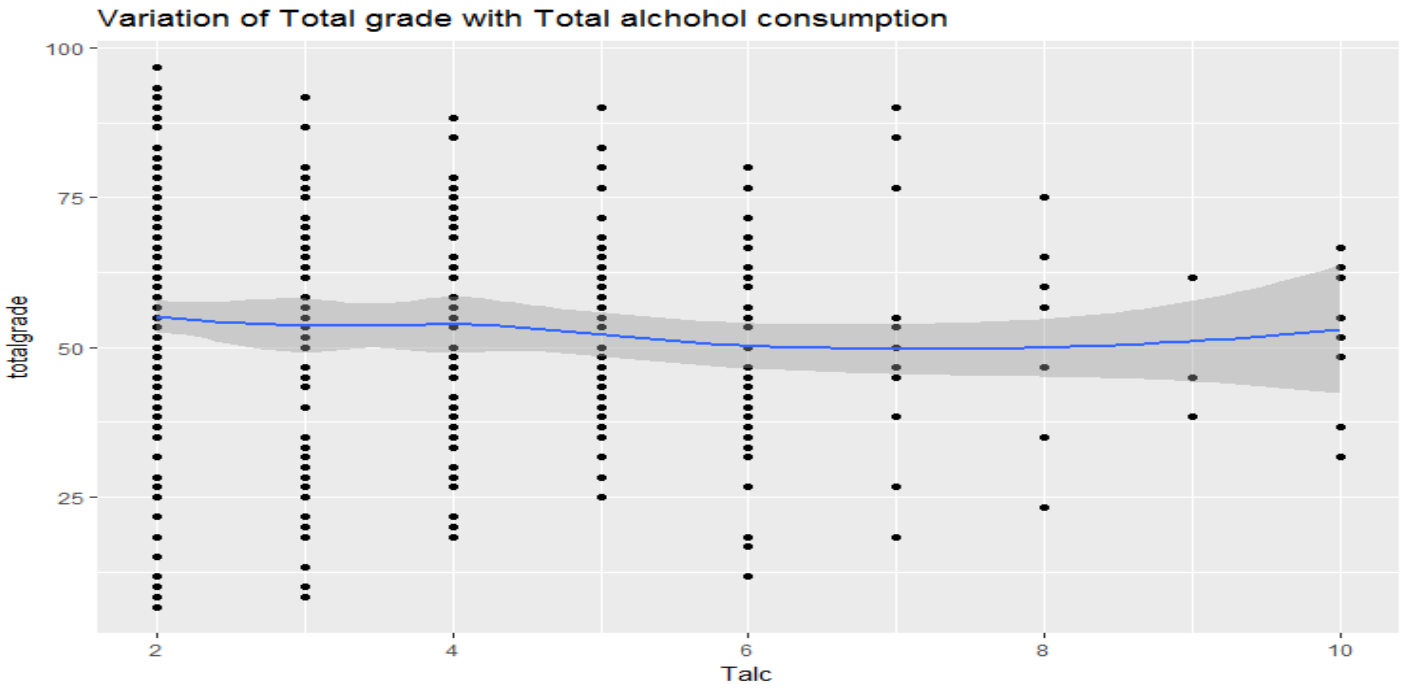



In the graph above, the total area is scaled to be equal to the number of observations. In order to answer the question, for a given total alcohol consumption, what proportion of students fall under categories of very low grades, low grades, good grades and above average grades, we plotted the below conditional density estimates.

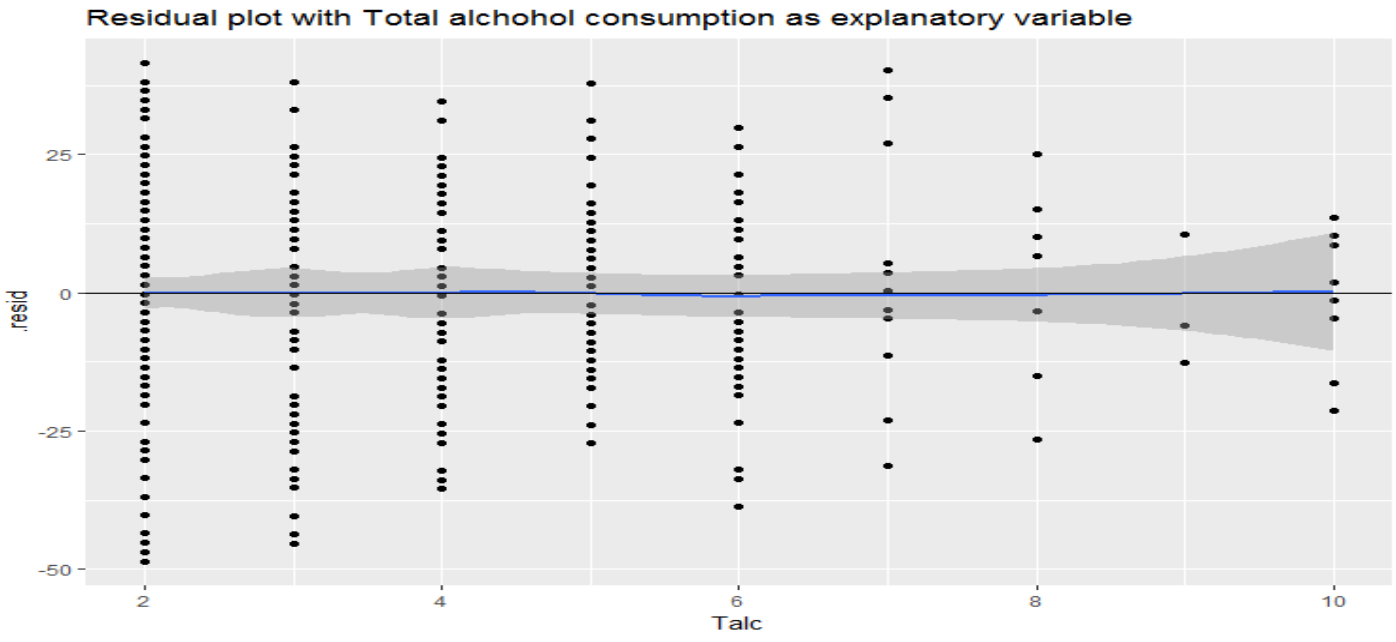


The proportion of students falling under each category does not have a uniform trend. The plot shows bumps in the distribution. For students with alcohol consumption between 8 – 10 , none of them get good grades. This can be because this region of exploratory variable has very less points (as the data is right skewed), so conditional distribution will not be well exhibited.

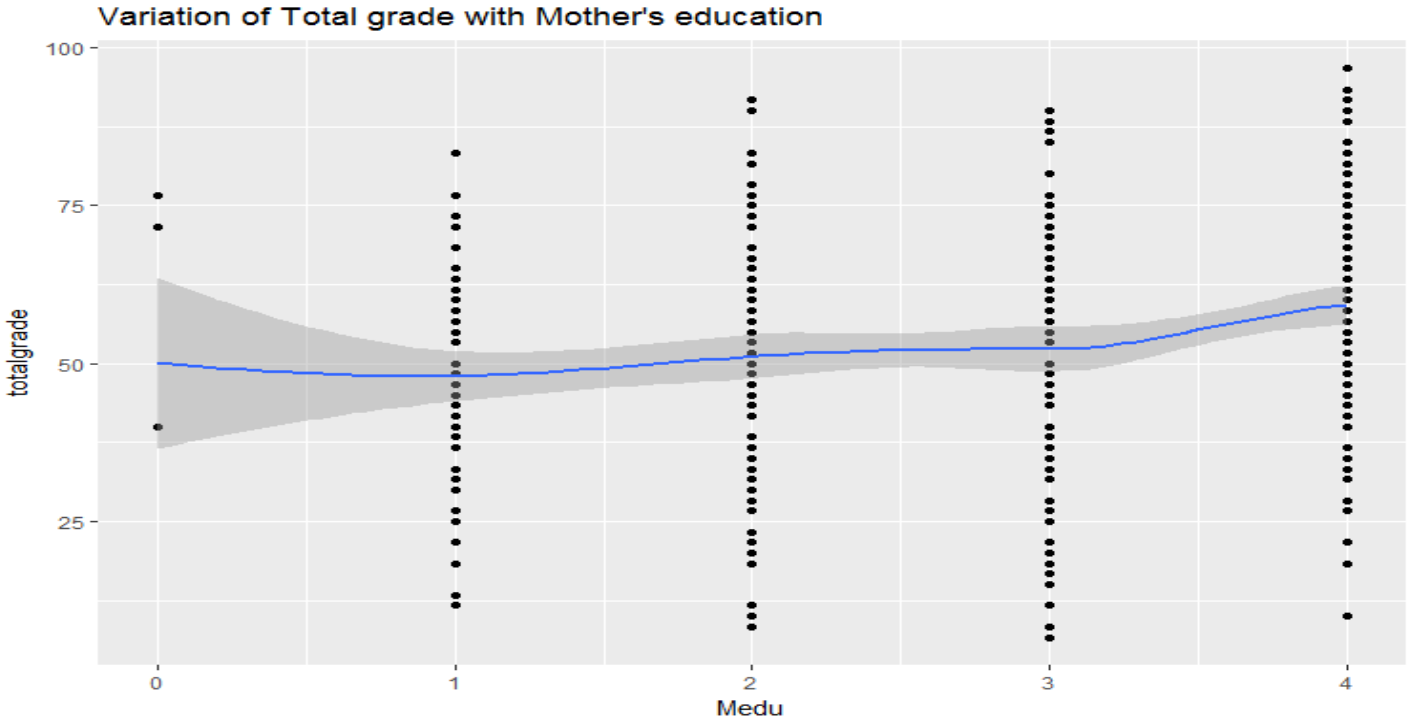
After looking at the distribution that we got in the above graphs, we have done model fitting using the numerical response variable 'totalgrades'. As the total alcohol consumption increases, the grades of students seems to follow a decreasing trend initially but there is no certain trend that we can see in the total grades or students with high alcohol consumption. Students with total alcohol consumption more than 6, tend to perform very poorly in exams.




LOESS model seems to be a good fit compared to others. It looks homoscedastic as well; except for a few outliers at the end. Compared to other predictors, the variance captured is less. The variance captured is around 8. Which suggests that the alcohol consumption is not a good predictor.

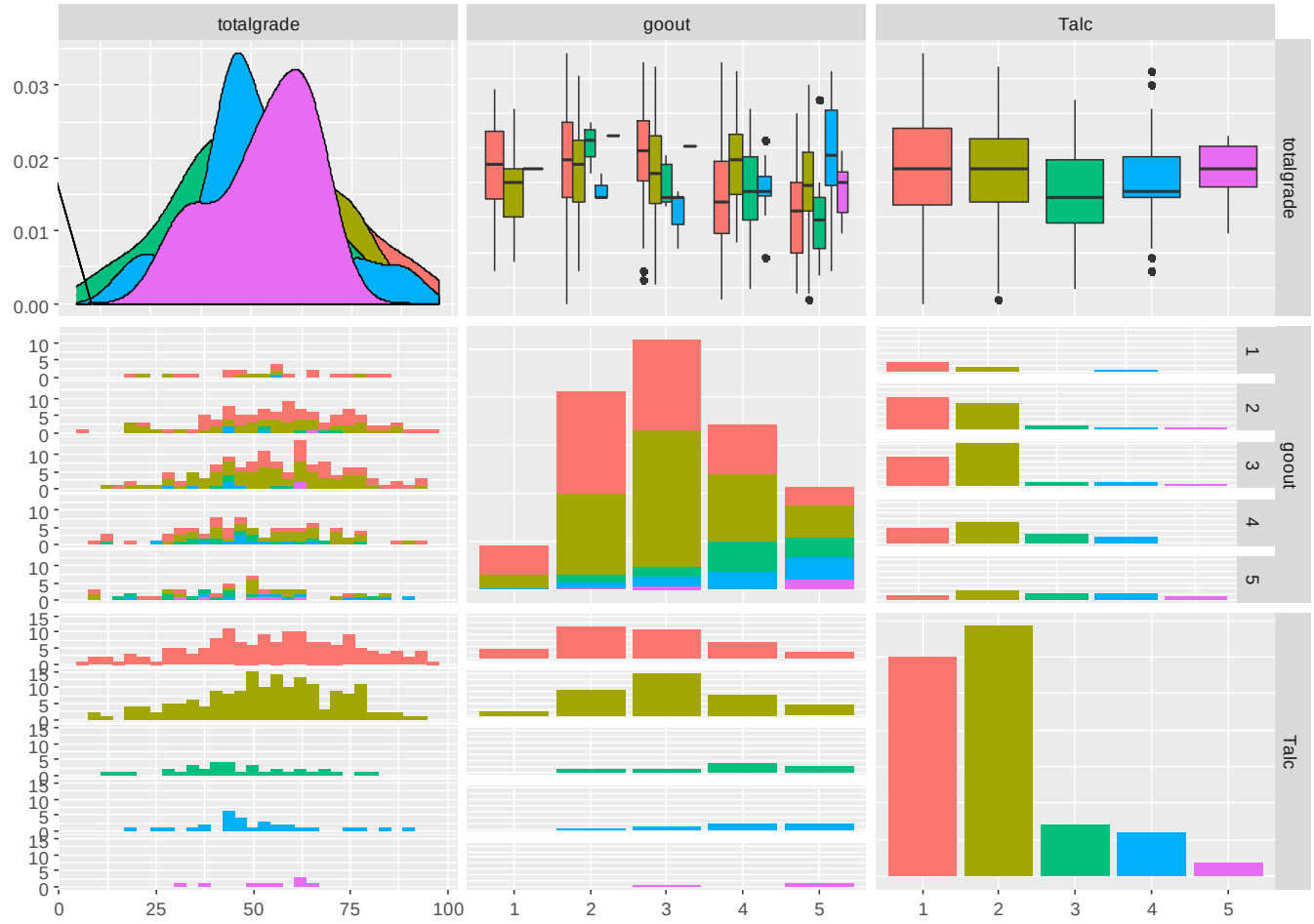


**Variation of Mother's education with total grade:** This is a plot to see how Mother's education varies with respect t o total grade. The overall trend is as a student's mother education increases, his/her grades also increase. This seems to be a logical assumption because mothers play an important part in a child's upbringing and it can be seen from the mode l fitting as well.

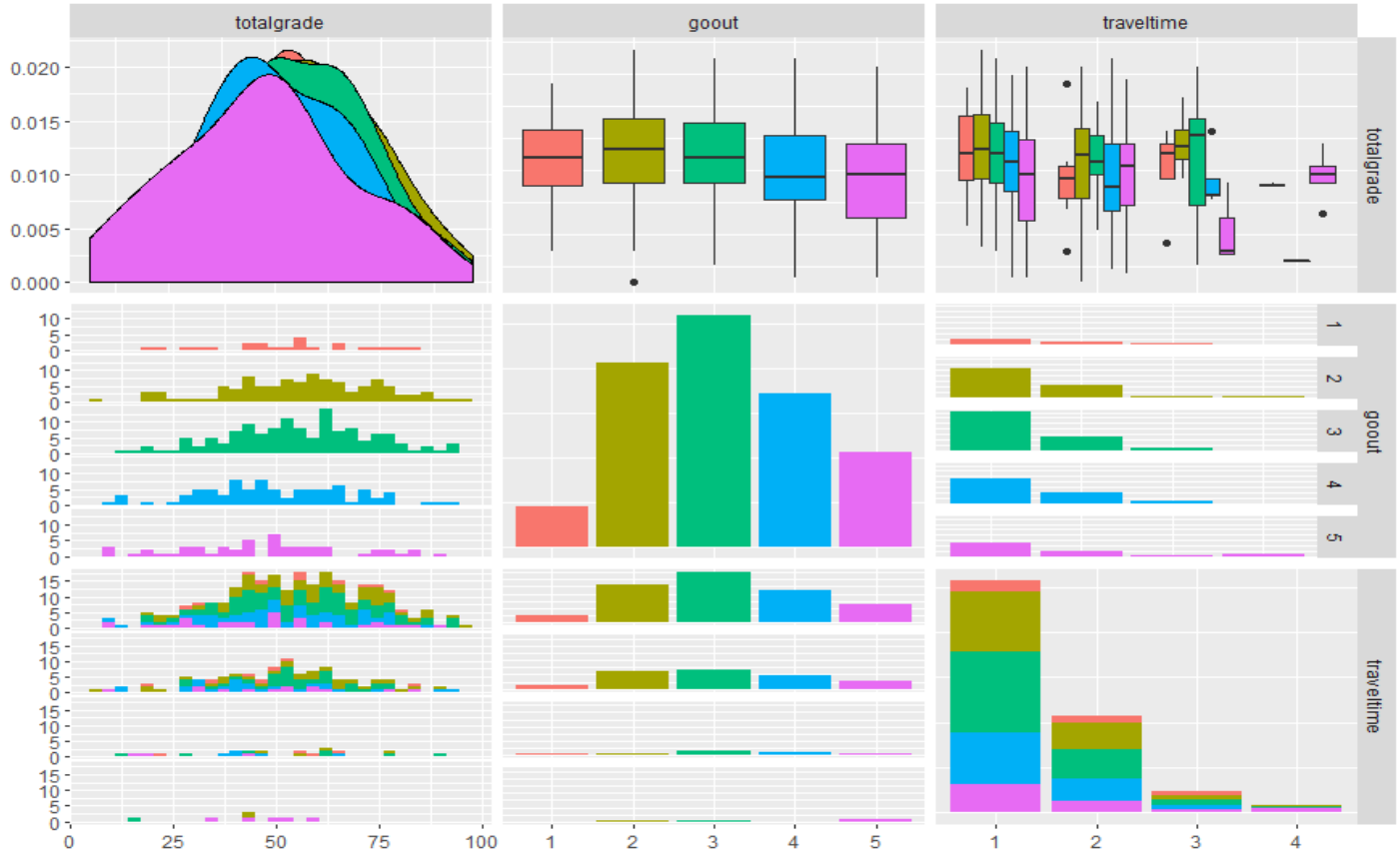



**TRIVARIATE ANALYSIS**

1. Using ggpairs, we looked at the relationship of total grades with respect to going out and total alcohol consumption, and how these two variables are in turn related with one another. As can be clearly seen the 'Talc' variable shares a negative relationship with the total grades, i.e. as alcohol consumption level increases, the total grades of students begin to decrease. This trend is almost true for all the alcohol consumptionlevels, but we can see that for heavy drinkers, the total grades showed an increasing ttrend. This can be attributed to the skewness of the data, since for heavy drinkers, we only have 17 out of the total 395 students. There is not a solid conclusion that we can draw as of now, about how going out affects this relation, but we can make a few guesses. The students who don't go out much have a lesser average grade value than those who go out once in a while. This looks like an interesting analysis to explore. The grades fluactuate for students who go out on a regular basis. The grades appear to be less for students who go out very frequently and are alcohol drinkers of level 2,3 and 4.

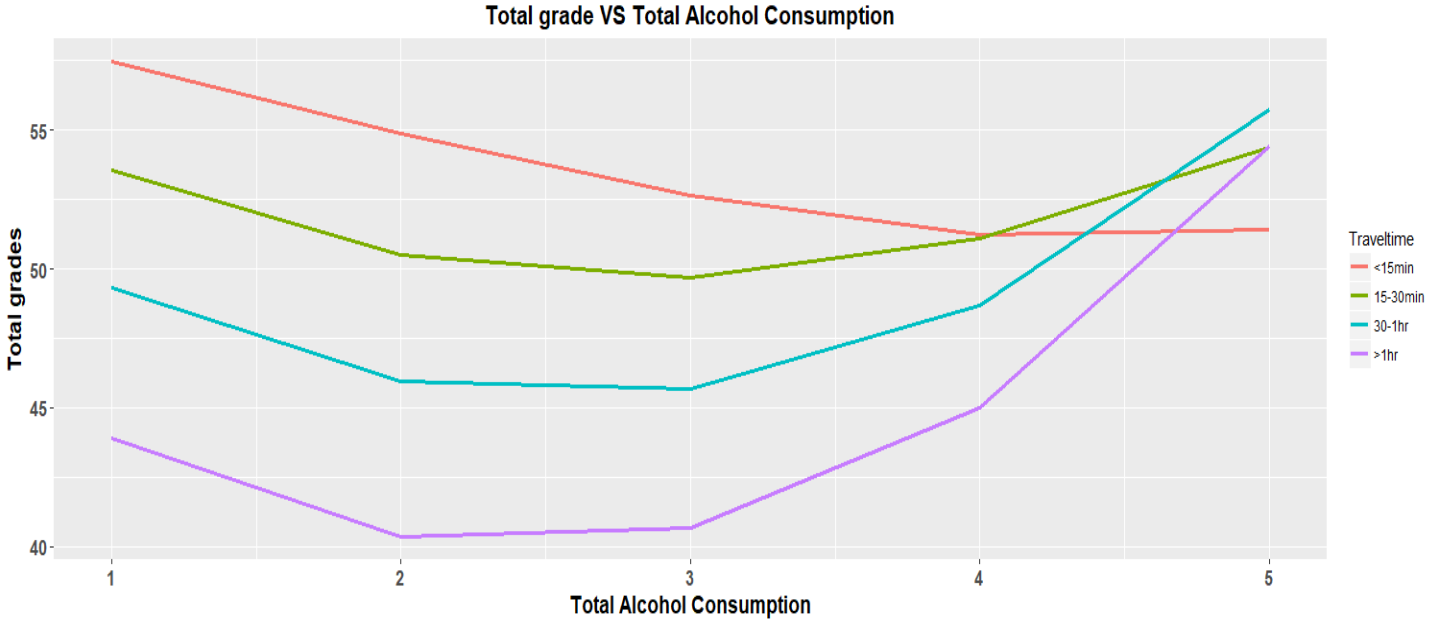



2. In the ggpairs plot below, we looked at the relationship of total grades with respect to going out and travel time, and how these two variables are in turn related with one another. We chose this combination since both mean a student is wasting time either in travelling or in going out, hence we aim to see how wastage of time can affect grades. Clearly going out is negatively correlated to the total grades, except for the first level, where the students do not go out at all. This is an interesting observation which we will be exploring later on. When we look at the how the interaction between going out and travel time to school affects, there is a negative relationship here as well. As can be seen from the graph below, the students who take more than an hour to reach school and go out frequently, their grades are lesser when compared to others.

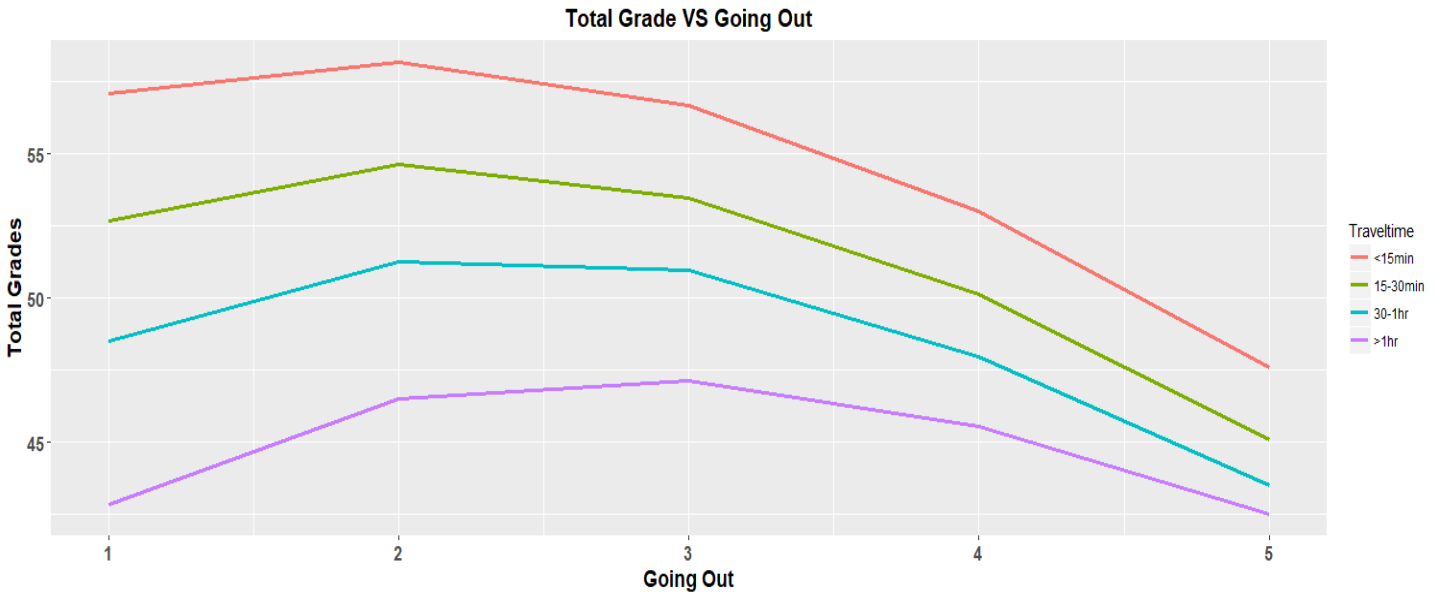



From the ggpairs, we made some interesting observations. Let us now have an in depth look into those.

1. **Total grades Vs total alcohol consumption**: As can be seen from the graph there is an initial negative trend observed in the total grades as the alcohol consumption of students increase, followed by a positive trend, suggesting that alcohol consumption increase the total grades of students, which is a rather an unsettling and counter intuitive observation. But if we look at the skewness, we realize that as mentioned before there are only 17 students who fall in the category of heavy drinkers (4,5). Hence, we will need more data to analyze and concretely base our hypothesis. Also, as expected, the total grades do on decreasing as travel time increases.



2. **Total grades Vs going out**: As can be clearly seen from the graph, the model we built using Loess follows a uniform trend across different factors of travel time. As can be seen from the graph, the students who take less than 15 minutes to travel, get higher grades. And the grades go on reducing as the travel time increases. This makes perfect sense, since travelling can get students exhausted and this may affect their studies. Also, if school is far away they might miss out on school more often. What appears interesting is that students who don't go out at all receive lesser grades than those students who go out occasionally. We guess this is because of: 'All work and no play made Jack a dull boy'. Some recreational activity is needed to refresh the mind. And hence going out once in a while is can contribute to bettering a students' performance.



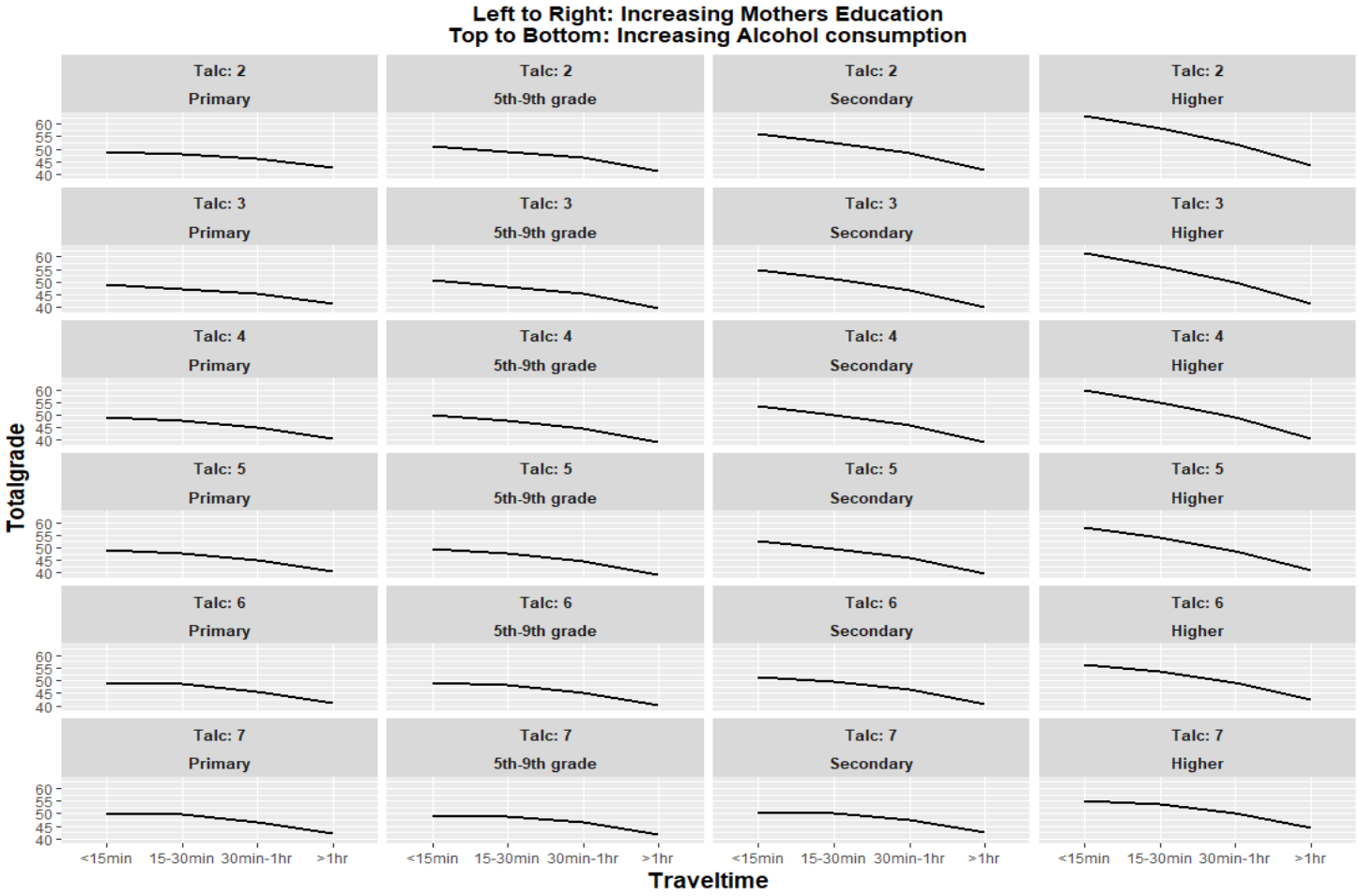

**MODEL FITTING:**

Since total grades VS going out gave a almost uniform trend, we decided to see how the spread location graph of the residuals and fitted values look like. The graph appears almost uniform and horizontal which is a good sign, but it lies far away from the y = 0, which suggests that the model is not that good a fit mostly due to the skewness in data. Thus, to make any substantial and concrete claims, we need more data, so that our distribution is normal and now skewed.

**HYPERVARIATE ANALYSIS**

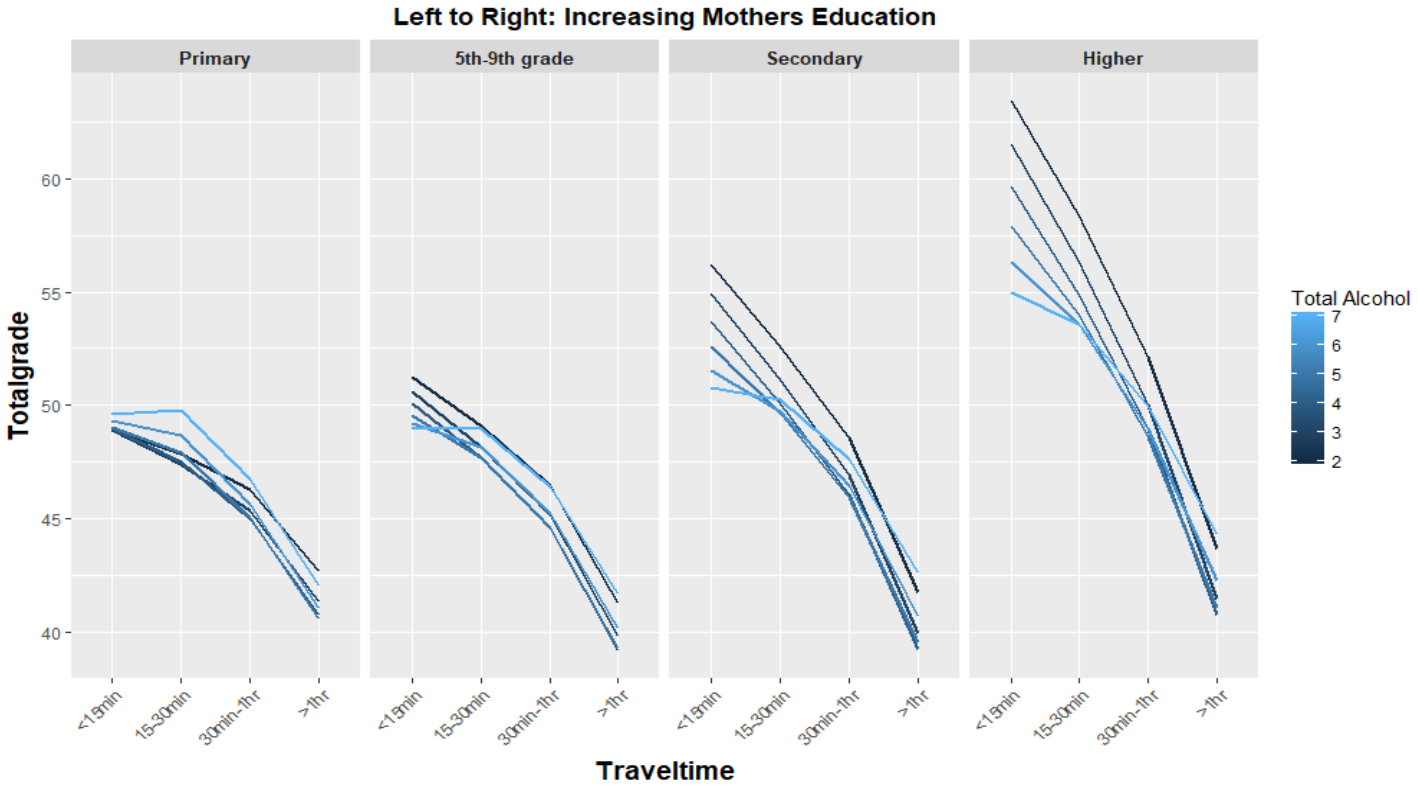
```
> table(d1$Talc)

  2   3   4   5   6   7   8   9  10
150  66  60  45  35  17   9   4   9
```

The graph beside shows the distribuition of data points which are used to build the loess model. As observable the dataset is very scarce, lacking records for lower education level, higher alcohol consumption and longer traver durations.

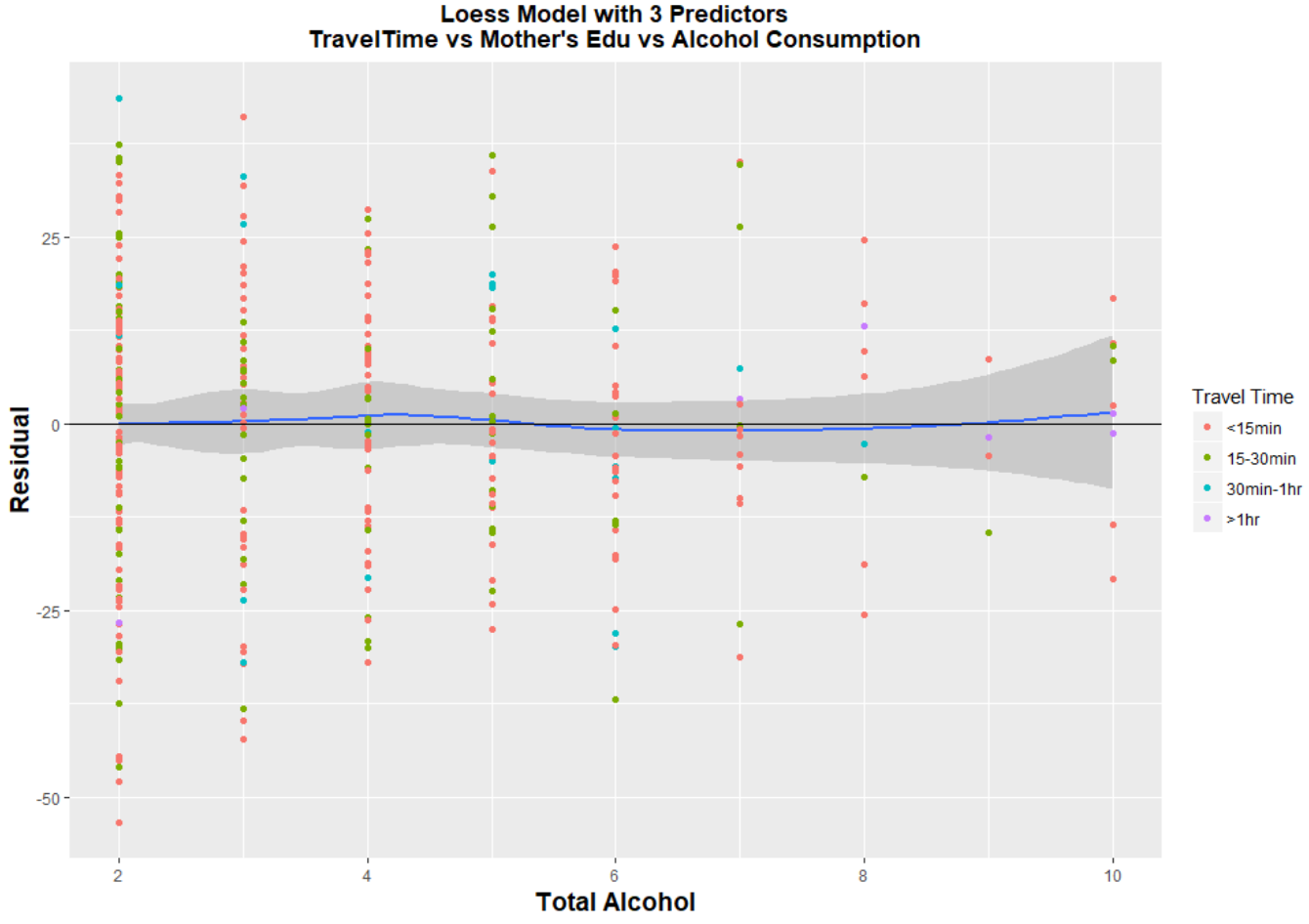


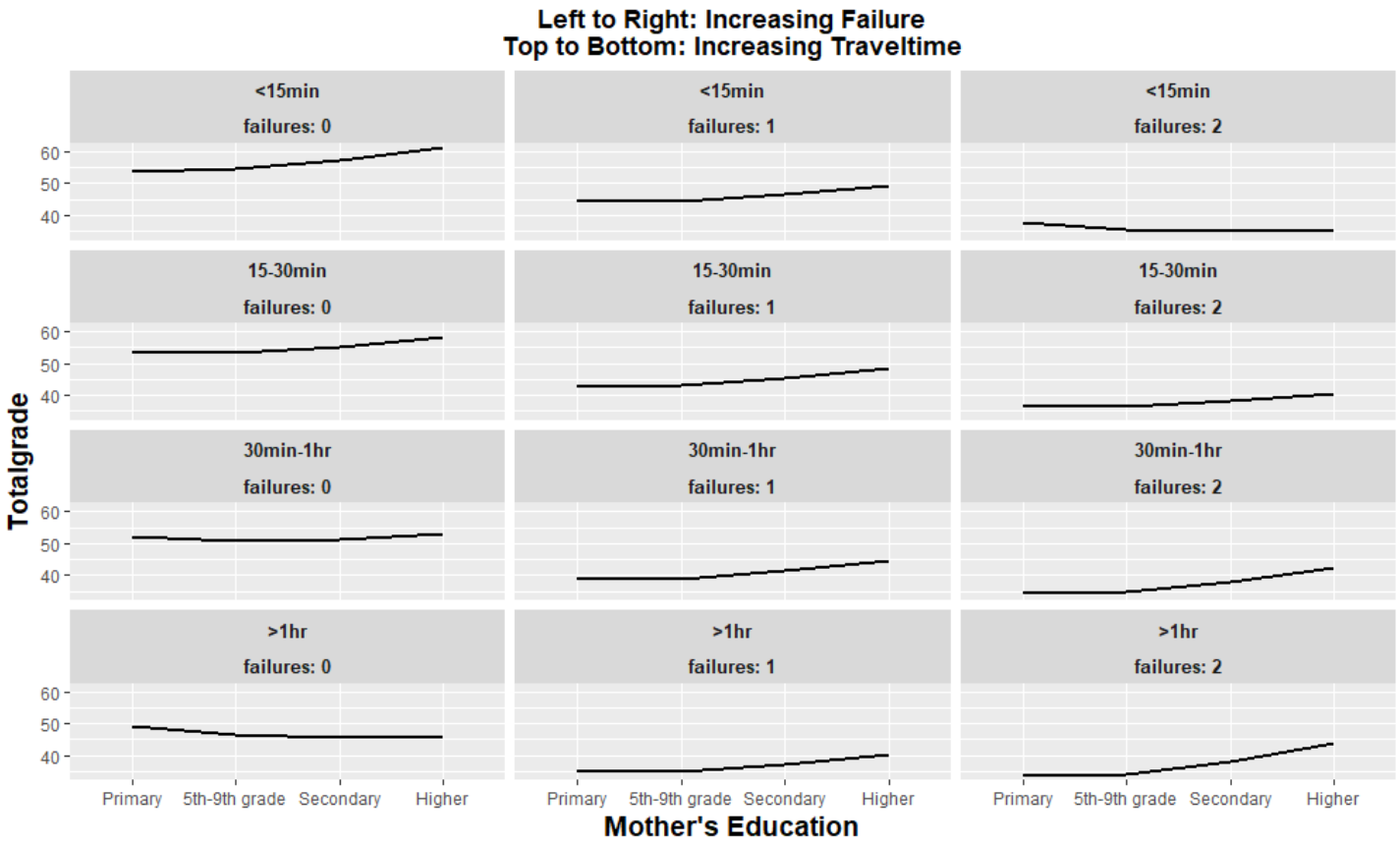

Testing different combinations of the predicator variables in Hypervariate analysis, we found that the model built with Mothers Education(Medu), Traveltime(traveltime) and Total Alcohol Consumption(Talc) gave the best results. This model when plotted with Traveltime on the x-axis and faceted by Mothers Education and Total alcohol consumption have a uniform pattern. The Total grade increases going upward and rightward across the facets and decreases rightward within the facet. Which makes sense as better educated mothers, lower alcohol consumption and shorter travel time account for better grades. The varying slope as we move downward indicate an interaction between alcohol consumption and travel time.

Summary or a neater version of the above graph is plotted below by removing the Talc facet dimension and using it as the colour aesthetic. In the below graph we see an oddity that highest alcohol consumption(index=7) seems to indicate better grades. This is precisely the unreliability aspect of the model built with scarce datapoints. The predictor is extremely biased by this that it even indicates that, “When the Mother has a primary education”- the trend is reversed; Students consuming more alcohol perform better. We cannot completely reject this anomaly without repeating the analysis with more data in these lacking categories.




The inference from the graphs is in line with our logical expectations, and the model accurately predicts grade as seen from the below residual graph; but the model is still unreliable because of the fact that the higher order alcohol consumption predictions have been fitted based on very few data points. This could cause the model to be highly biased.






The above predictor combination represents another good model.

## V) CONCLUSION

We set out to find interesting patterns between student grades and their alcohol consumption habits. Thorough our analysis we observed that alcohol is not a significant factor that affects the total grades of students although there was a correlation that affects the grade inversely. Travel Time and Mother's education are better predictors for total grades of students.

## VI) FUTURE WORK

A good future study on this relationship could include better curation of data from more schools and across different geographic locations. We would still expect the data to be skewed as secondary school students do not have adverse drinking habits. During data curation the factors such as what constitutes low or high drinking habits needs to be established as different subsections of people may have their own bench mark for low or high drinking habits.