Student Alcohol Consumption

Data Analysis

CONTENTS

Data Overview

Data Description

Context

The data were obtained in a survey of students math courses in secondary school. It contains a lot of interesting social, gender and study information about students. You can use it for some EDA or try to predict students final grade

Purpose of Analysis

After finding which variable makes the greatest impact on student alcohol consumption, I will segment students based on alcohol consumption level.

I will suggest marketing strategies for promoting alcohol consumption for each student segment.

I will find variables that have the greatest impact on the school record.

Data Description

Variable Description

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 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	- weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >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)

Data Description

Data Glimpse

No	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob
1	GP	F	18	U	GT3	A	4	4	at_home	teacher
2	GP	F	17	U	GT3	T	1	1	at_home	other
3	GP	F	15	U	LE3	T	1	1	at_home	other
4	GP	F	15	U	GT3	T	4	2	health	services
5	GP	F	16	U	GT3	T	3	3	other	other
6	GP	M	16	U	LE3	T	4	3	services	other
7	GP	M	16	U	LE3	T	2	2	other	other
8	GP	F	17	U	GT3	A	4	4	other	teacher
9	GP	M	15	U	LE3	A	3	2	services	other
10	GP	M	15	U	GT3	T	3	4	other	other

NO	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery
1	course	mother	2	2	0	yes	no	no	no	yes
2	course	father	1	2	0	no	yes	no	no	no
3	other	mother	1	2	3	yes	no	yes	no	yes
4	home	mother	1	3	0	no	yes	yes	yes	yes
5	home	father	1	2	0	no	yes	yes	no	yes
6	reputation	mother	1	2	0	no	yes	yes	yes	yes
7	home	mother	1	2	0	no	no	no	no	yes
8	home	mother	2	2	0	yes	yes	no	no	yes
9	home	mother	1	2	0	no	yes	yes	no	yes
10	home	mother	1	2	0	no	yes	yes	yes	yes

NO	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
1	yes	no	No	4	3	4	1	1	3	6	5	6	6
2	yes	yes	no	5	3	3	1	1	3	4	5	5	6
3	yes	yes	no	4	3	2	2	3	3	10	7	8	10
4	yes	yes	yes	3	2	2	1	1	5	2	15	14	15
5	yes	no	no	4	3	2	1	2	5	4	6	10	10
6	yes	yes	no	5	4	2	1	2	5	10	15	15	15
7	yes	yes	no	4	4	4	1	1	3	0	12	12	11
8	yes	no	no	4	1	4	1	1	1	6	6	5	6
9	yes	yes	no	4	2	2	1	1	1	0	16	18	19
10	yes	yes	no	5	5	1	1	1	5	0	14	15	15

Data Analysis

I will do EDA on Student Alcohol Consumption date set.

I will choose alcohol consumption level and math scores as Target variables.

I will develop two models in total.

1. The first model is for finding student segments based on alcohol consumption level.

2. The second model is for finding major variables that most affect math scores.

For the first model,
I will first segment students with the Decision Tree Modeling.

For the second model,
I will use the regression analysis to find variables that most affect math scores.

Lastly, I will suggest a marketing strategy for promoting student alcohol consumption based on analysis results.

Exploratory Data Analysis

Alcohol Consumption EDA

Data Structure

We have 395 Observations and 39 Variables.
Dalc and Walc variables are related to alcohol comsumption
Variables's scale is Likert scale. (1- Very Low ~ 5- Very High)
Dalc mean weekdays alcohol consumption.
Walc mean weekend alcohol consumption.

```
'data.frame':    395 obs. of  39 variables:
 $ school     : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
 $ sex        : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
 $ age        : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address    : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
 $ famsize    : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
 $ Pstatus    : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
 $ Medu       : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu       : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob       : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
 $ Fjob       : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
 $ reason     : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
 $ guardian   : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
 $ traveltime : int  2 1 1 1 1 1 1 2 1 1 ...
 $ studytime  : int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures   : int  0 0 3 0 0 0 0 0 0 0 ...
 $ schoolsup  : num  1 0 1 0 0 0 0 1 0 0 ...
 $ famsup     : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
 $ paid       : num  0 0 1 1 1 1 0 0 1 1 ...
 $ activities : Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
 $ nursery    : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ higher     : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ internet   : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
 $ romantic   : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ famrel     : int  4 5 4 3 4 5 4 4 4 5 ...
 $ freetime   : int  3 3 3 2 3 4 4 1 2 5 ...
 $ goout      : int  4 3 2 2 2 2 4 4 2 1 ...
 $ Dalc       : int  1 1 2 1 1 1 1 1 1 1 ...
 $ Walc       : int  1 1 3 1 2 2 1 1 1 1 ...
 $ health     : int  3 3 3 5 5 5 3 1 1 5 ...
 $ absences   : int  6 4 10 2 4 10 0 6 0 0 ...
 $ G1         : int  5 5 7 15 6 15 12 6 16 14 ...
 $ G2         : int  6 5 8 14 10 15 12 5 18 15 ...
 $ G3         : int  6 6 10 15 10 15 11 6 19 15 ...
```

Variables Manufacture

alcohol
- I created an alcohol variable based on Dalc and Walc variables.(Low, Middle, High, Very High)

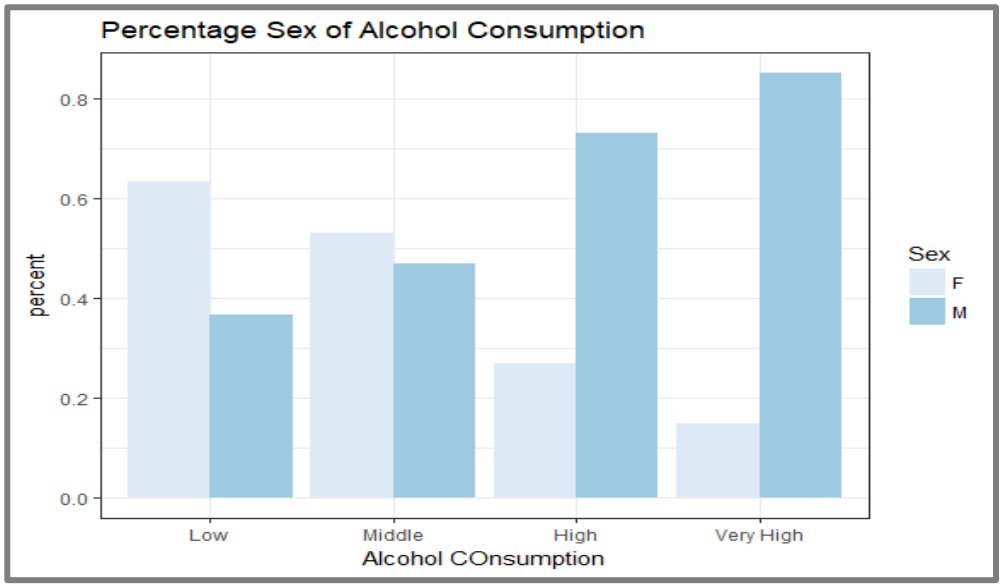
age_new
- I re-defined age variable because of obseravtion count.

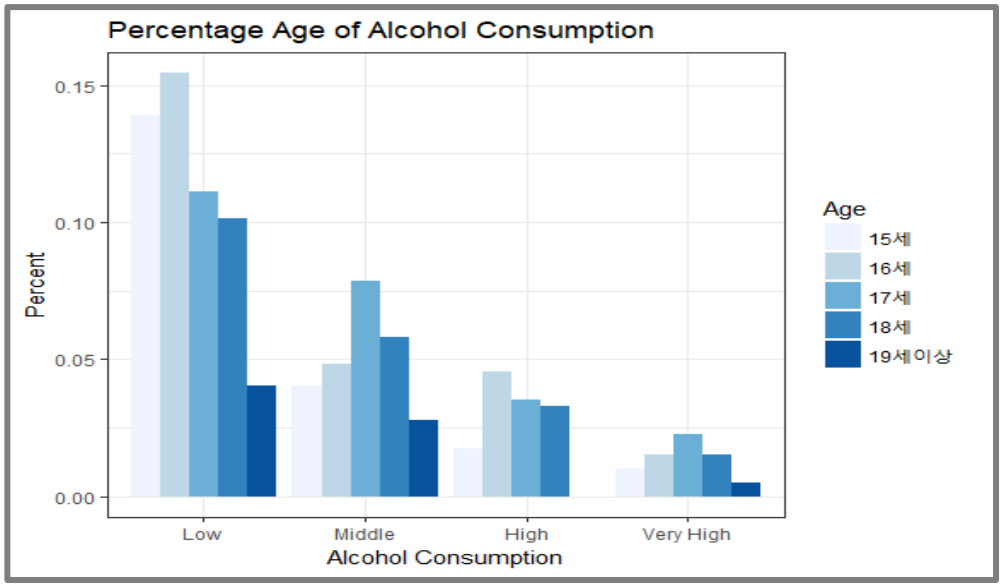
parents_edu_ct
- I created a parents_edu_ct variable based on Medu, Fedu variables.(Low, Middle, High)

extra_sup
- I creaded a extra_sup variable based on schoolsup and paid variables.(0,1,2)

Alcohol Consumption EDA

Alcohol consumption percentage based on sex or age



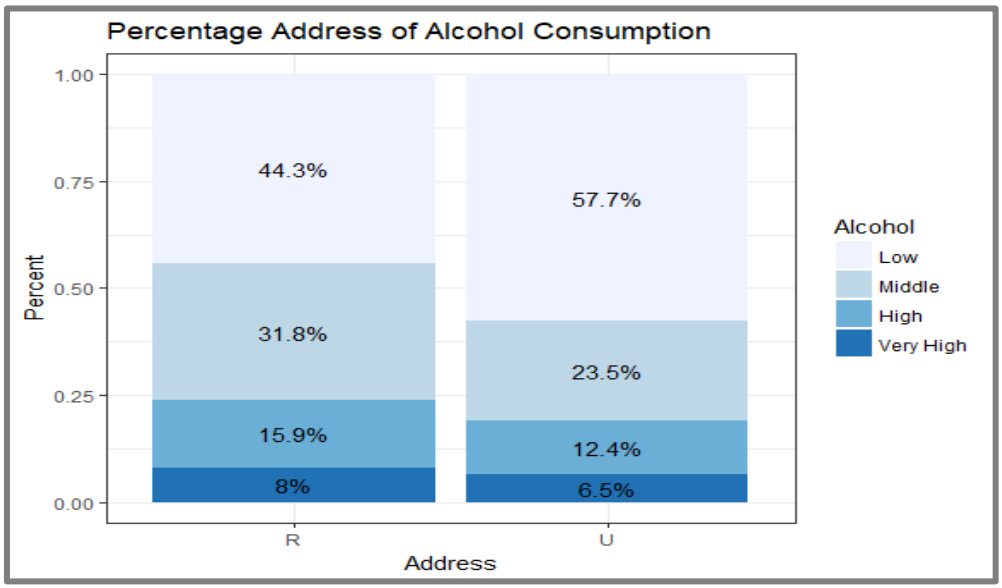


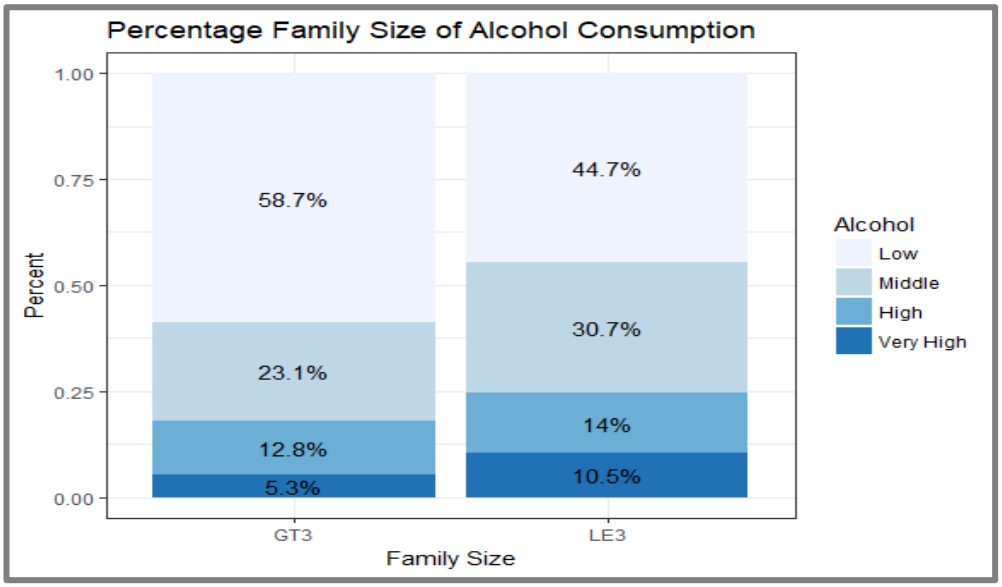
The first graph shows that male students in the high, and very high account for more percentage than female.

The second graph shows that older students in the high, and very high account for more percentage than younger students

Alcohol Consumption EDA

Alcohol consumption percentage based on address or family size



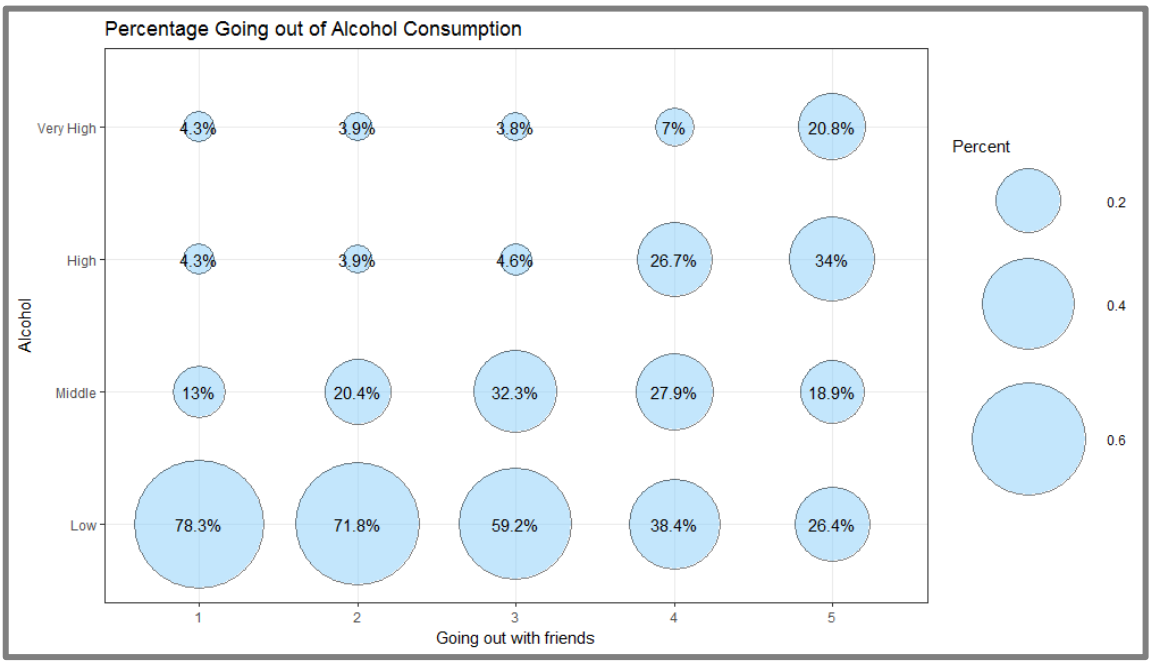


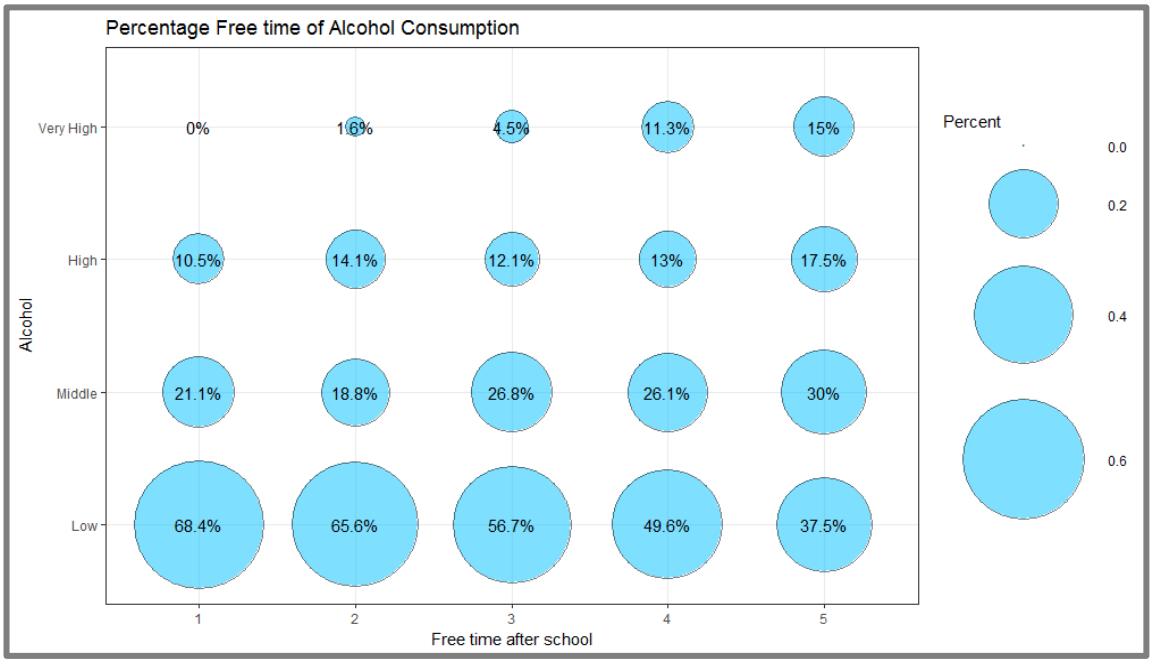
The first graph shows that students in the rural consume more alcohol than students in the urban

The second graph shows that students in the LE3 consume more alcohol than GT3

Alcohol Consumption EDA

Alcohol consumption based on Going out and Free time



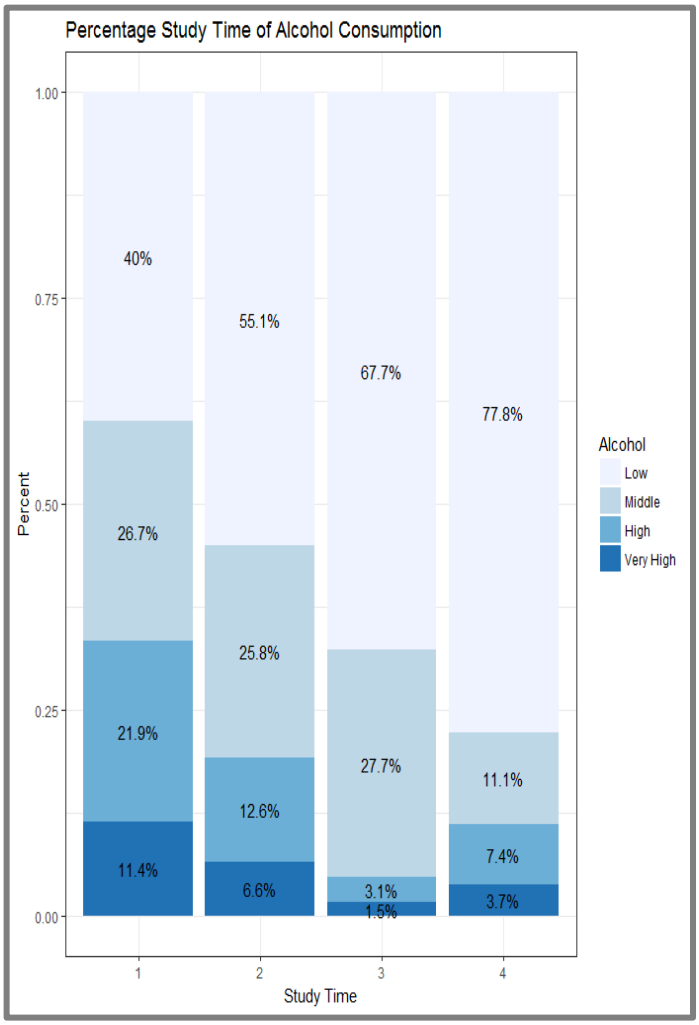


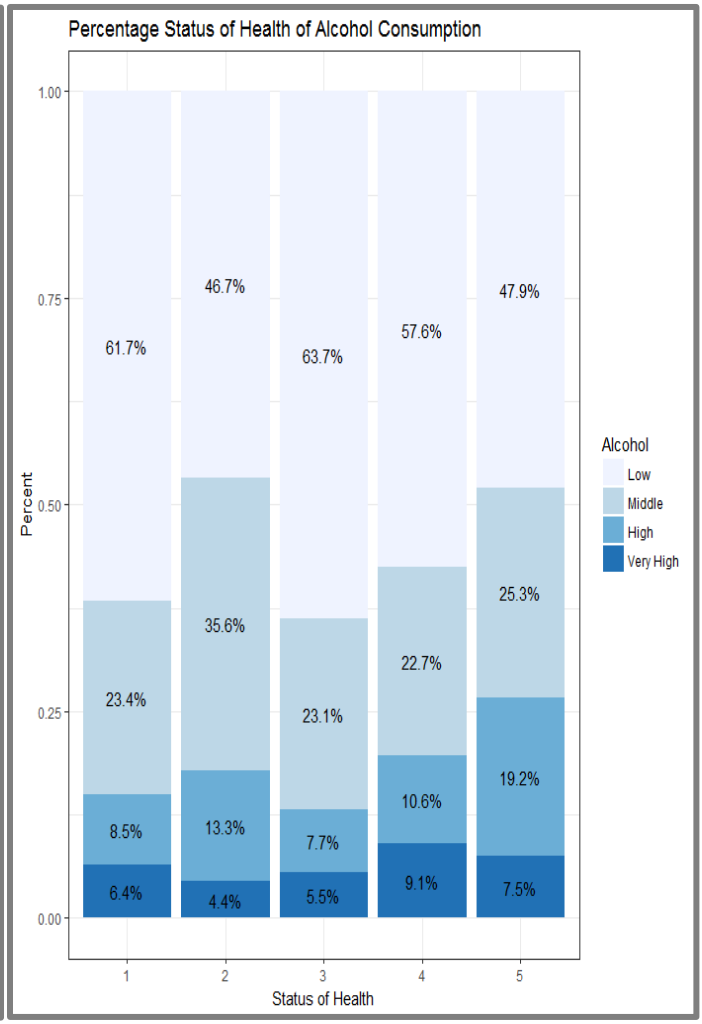
The more often students go out, the more they consume alcohol.

The more free time students have, the more they consume alcohol.

Alcohol Consumption EDA

Alcohol consumption based on Study Time and Health





The longer Study Time students have, the less they consume alcohol.

The healthier students are, the more they consume alcohol.

Data Structure

We have 395 Observations and 39 Variables.
G1, G2 and G3 variables are related to Math Grade

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)

```
'data.frame':    395 obs. of  39 variables:
 $ school     : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
 $ sex        : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
 $ age        : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address    : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
 $ famsize    : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
 $ Pstatus    : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
 $ Medu       : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu       : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob       : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
 $ Fjob       : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
 $ reason     : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
 $ guardian   : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
 $ traveltime : int  2 1 1 1 1 1 1 2 1 1 ...
 $ studytime  : int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures   : int  0 0 3 0 0 0 0 0 0 0 ...
 $ schoolsup  : num  1 0 1 0 0 0 0 1 0 0 ...
 $ famsup     : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
 $ paid       : num  0 0 1 1 1 1 0 0 1 1 ...
 $ activities : Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
 $ nursery    : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ higher     : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ internet   : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
 $ romantic   : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ famrel     : int  4 5 4 3 4 5 4 4 4 5 ...
 $ freetime   : int  3 3 3 2 3 4 4 1 2 5 ...
 $ goout      : int  4 3 2 2 2 2 4 4 2 1 ...
 $ Dalc       : int  1 1 2 1 1 1 1 1 1 1 ...
 $ Walc       : int  1 1 3 1 2 2 1 1 1 1 ...
 $ health     : int  3 3 3 5 5 5 3 1 1 5 ...
 $ absences   : int  6 4 10 2 4 10 0 6 0 0 ...
 $ G1         : int  5 5 7 15 6 15 12 6 16 14 ...
 $ G2         : int  6 5 8 14 10 15 12 5 18 15 ...
 $ G3         : int  6 6 10 15 10 15 11 6 19 15 ...
```

Variables Manufacture

G123
- I created a G123 as target bariable.(G123 = G1 + G2 + G3)

alcohol
- I created an alcohol variable based on Dalc and Walc variables.(Low, Middle, High, Very High)

age_new
- I re-defined age variable because of obseravtion count.

parents_edu_ct
- I created a parents_edu_ct variable based on Medu, Fedu variables.(Low, Middle, High)

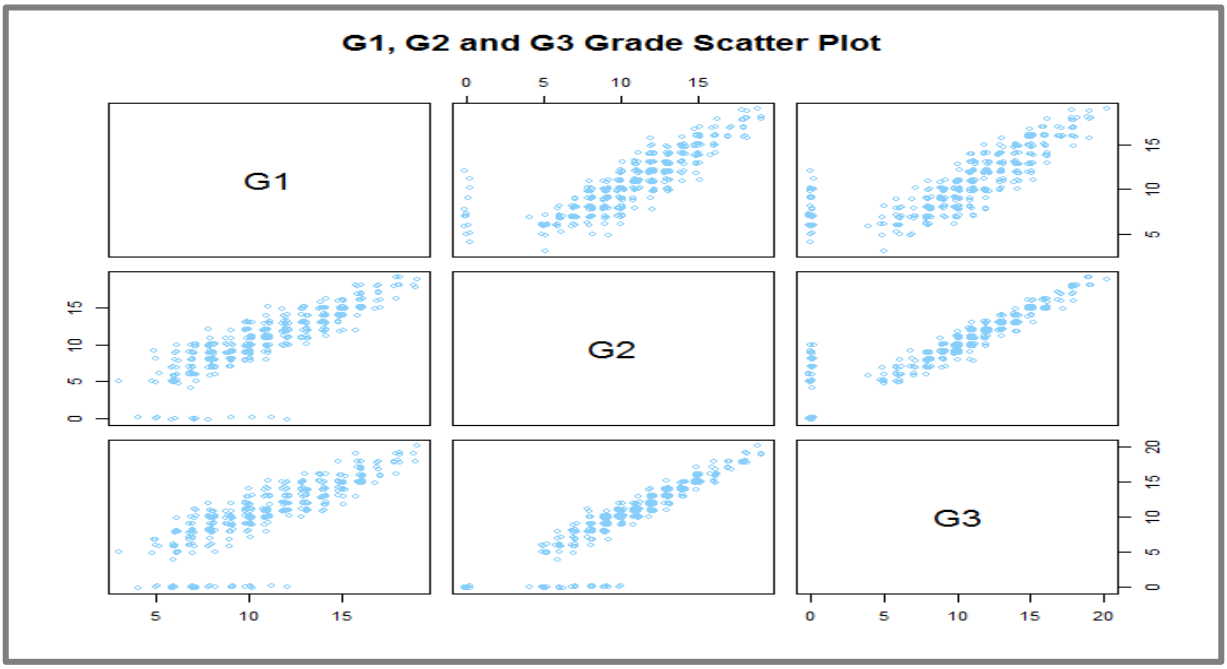
extra_sup
- I created a extra_sup variable based on schoolsup and paid variables.(0,1,2)

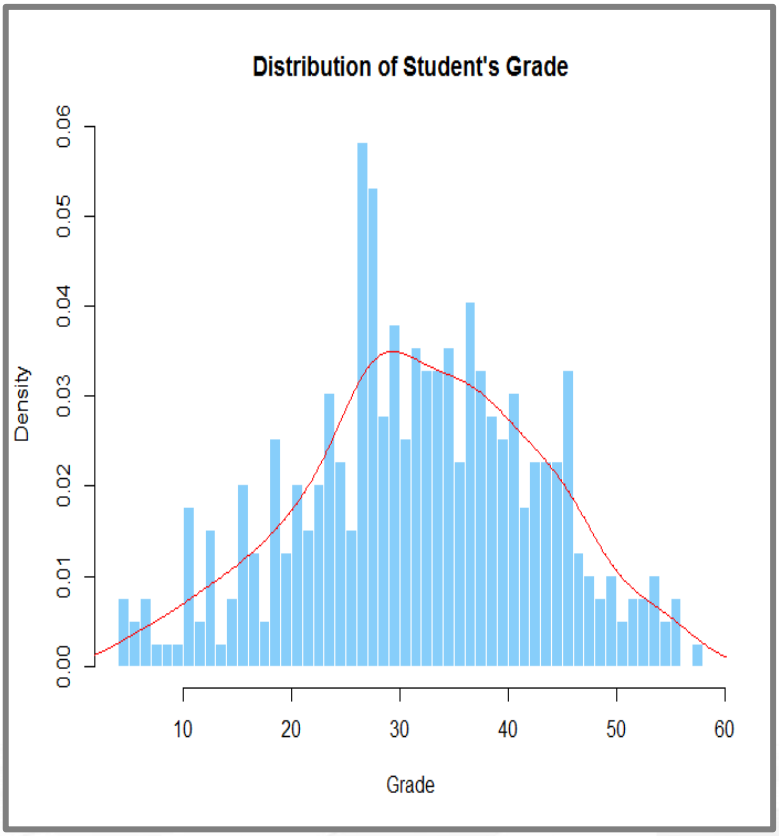
absences_check
- I created a absences_check variable. (no, yes)

freetime_goout
- I created a freetime_goout variable. (freetime – goout / -4 ~ +4)

Math Grade EDA

G1, G2 and G3 Math score variable



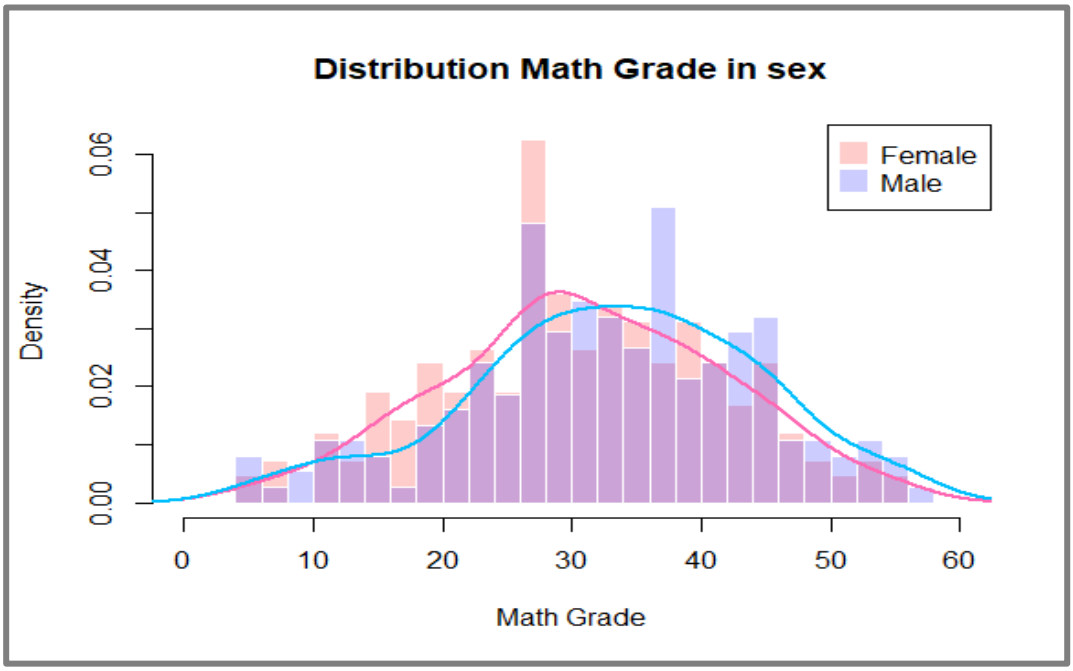


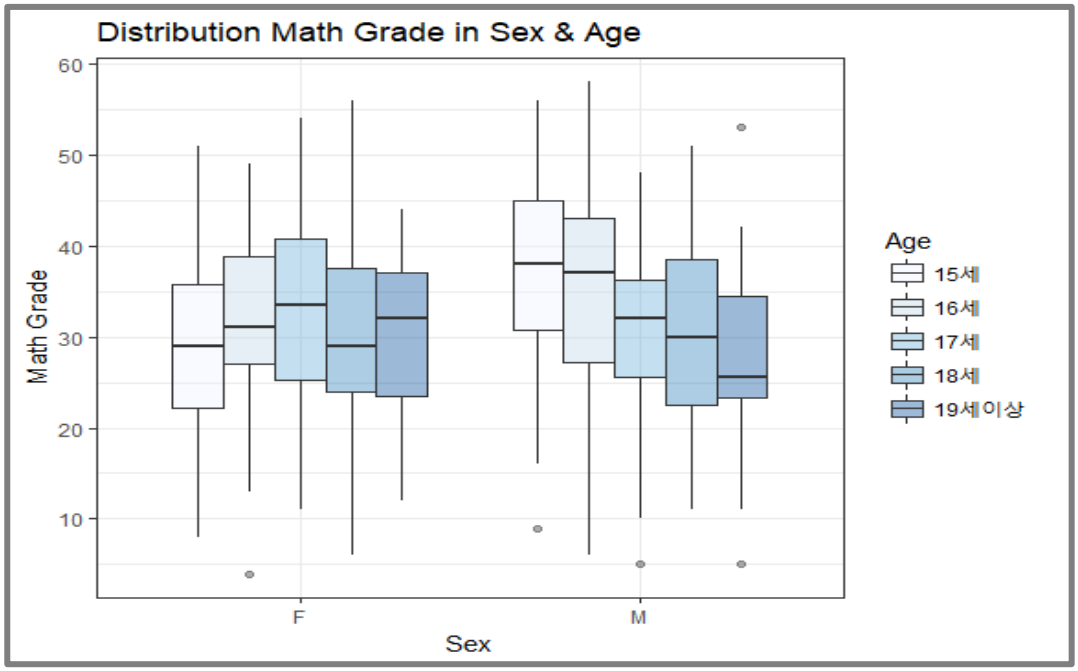
The first graph shows that G1, G2 and G3 variables have positive correlation. So I created just one variable G123 by calculating G1+G2+G3.

The second graph is histogram of G123.
This graph shows that G123 has normality with average score 32.04

Math Grade EDA

Distribution of G123 based on sex or age



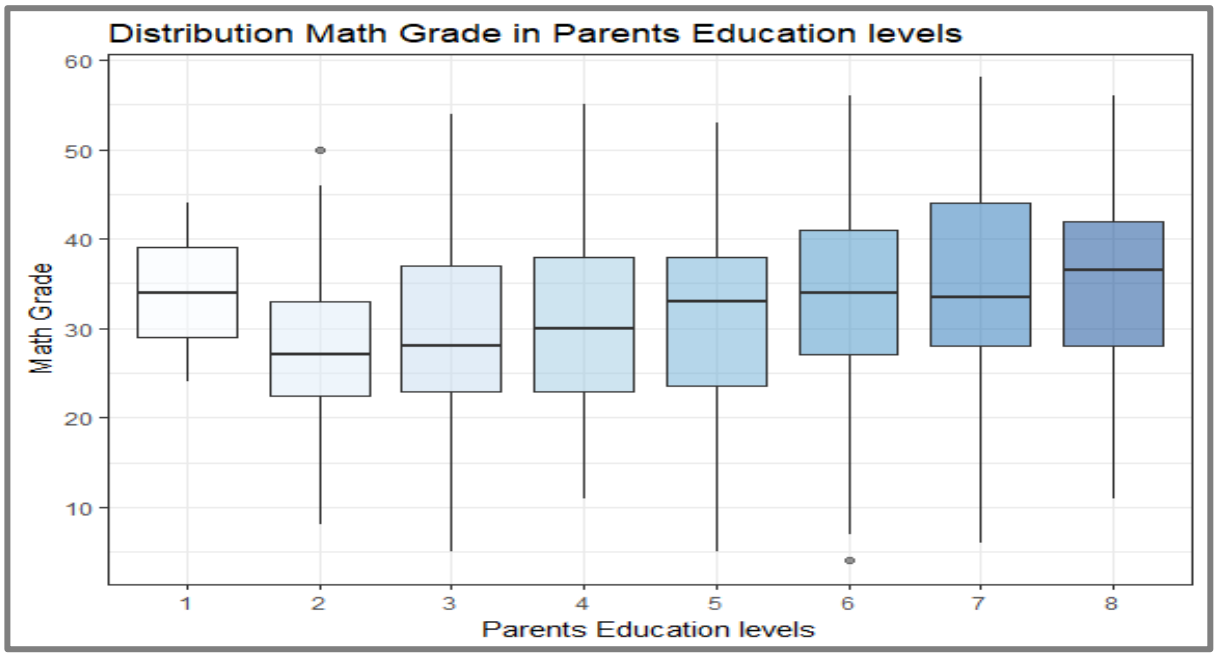


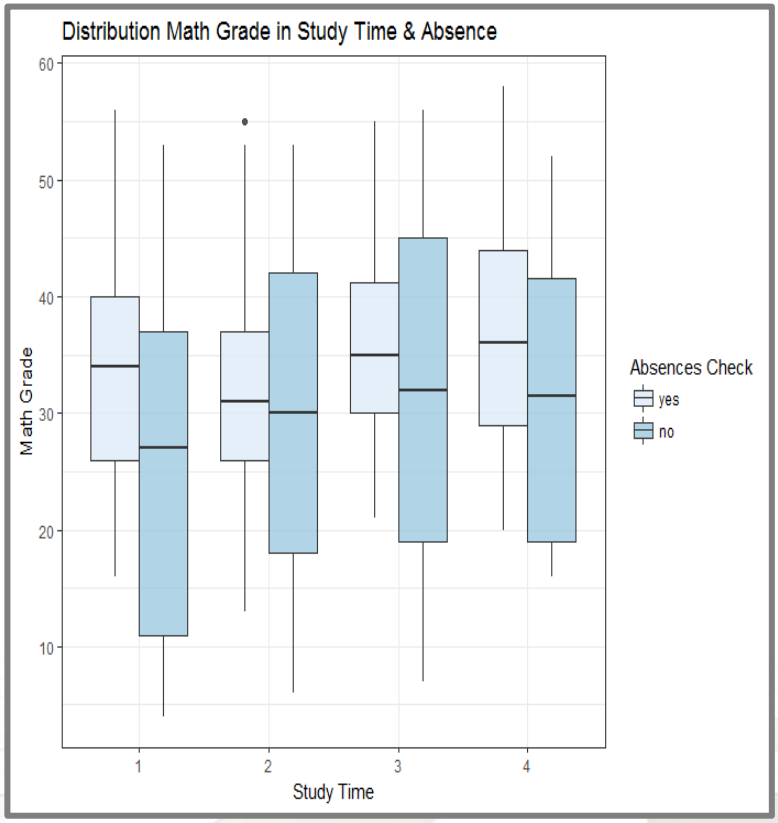
The longer Study Time students have, the less they consume alcohol.

The healthier students are, the more they consume alcohol.

Math Grade EDA

Distribution of G123 based on Parents-edu or study-time, absence

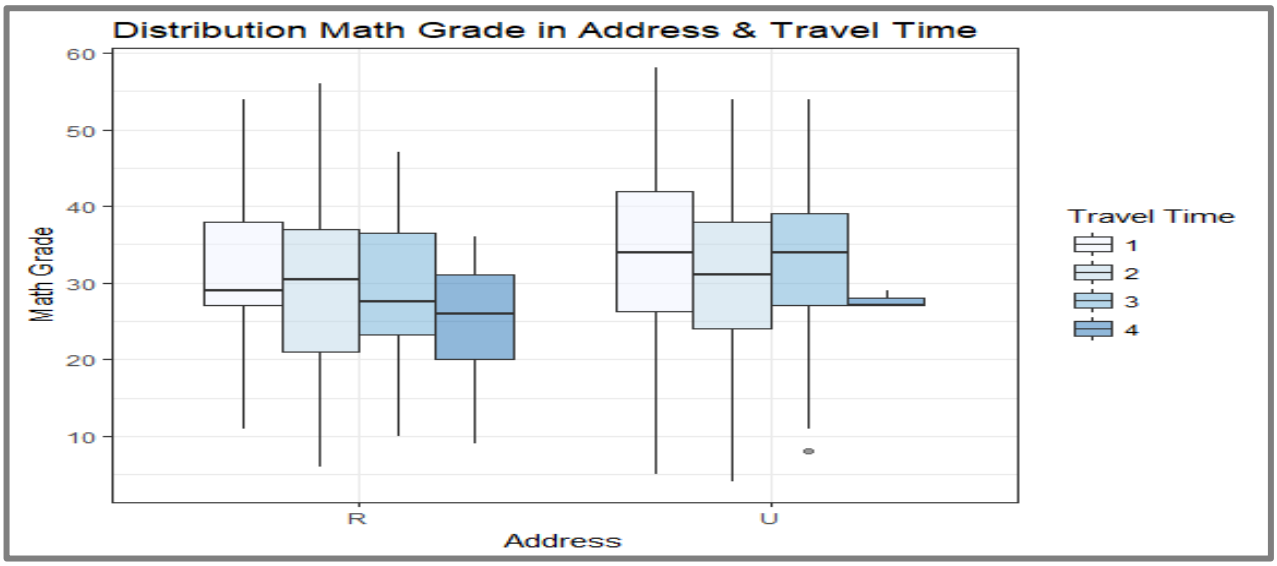


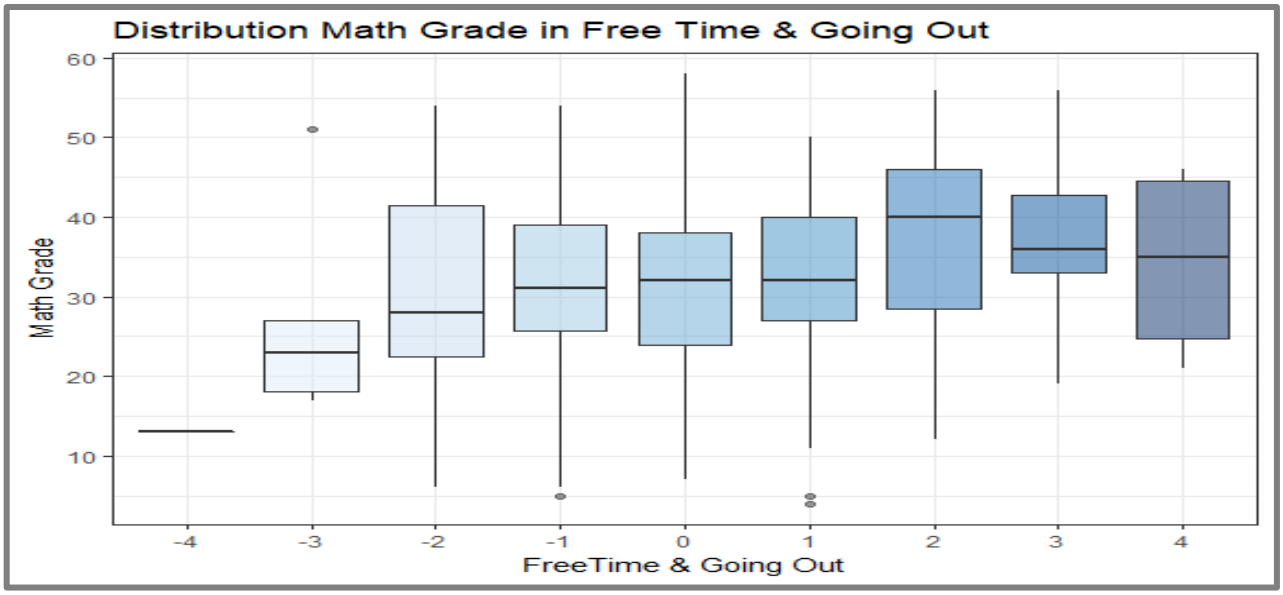


The better educated students' parents are the higher score students have

The more have study time the higher score students have and Absences "Yes" students have more score than "No" students

Distribution of G123 based on address,travel-time, free-time and goout



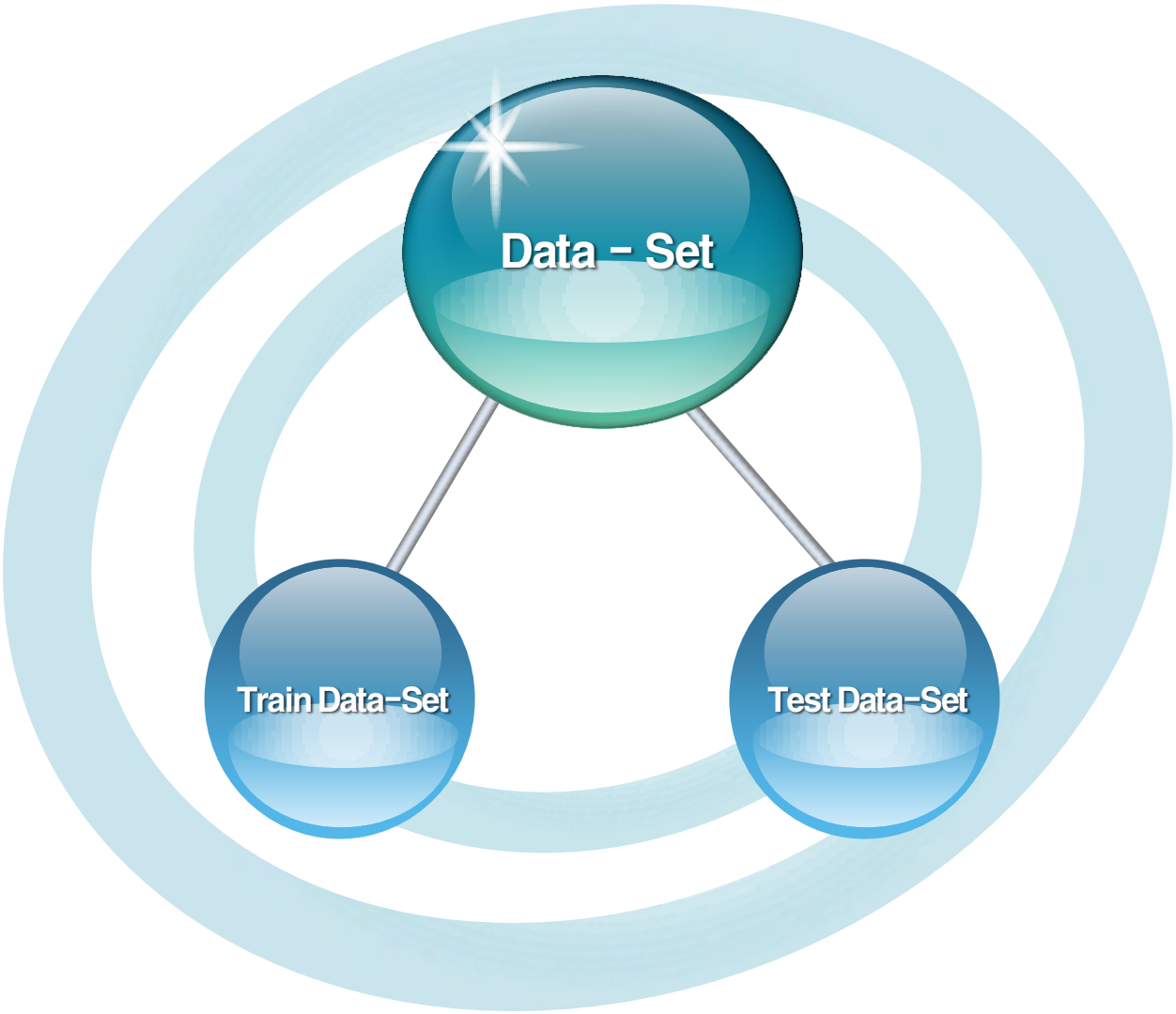


Students in urban have higher scores than in rural and the longer travel time students have the less scores they have.

Students that Free-time is shot and Going-out is high have low score
But, Free-time is long and Going-out is low have high score.

Modeling

Train and Test Data-set



I divided the whole data set at a ratio of 7 training data sets to 3 test data sets and used the data for creating models and performance test to produce an optimal model.

Train Data-Set : 276 obs
Objective : creating a predictive model
This Data-Set is train data to find student's feature

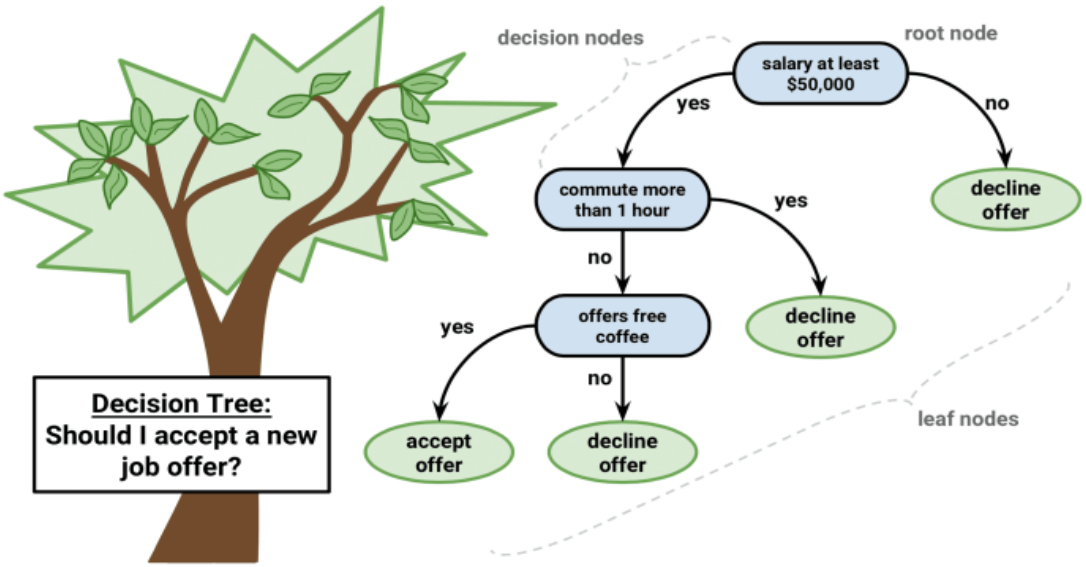
Test Data-Set : 119 obs
Objective : Test created model
This Data-Set is Test data that not used to create model to assess a predictive model

Decision Tree Modeling (1)

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Features of Decision Tree

- Simple to understand, interpret, visualize.
- Decision trees implicitly perform variable screening or feature selection.
- Can handle both numerical and categorical data. Can also handle multi-output problems.
- Decision trees require relatively little effort from users for data preparation.
- Nonlinear relationships between parameters do not affect tree performance.



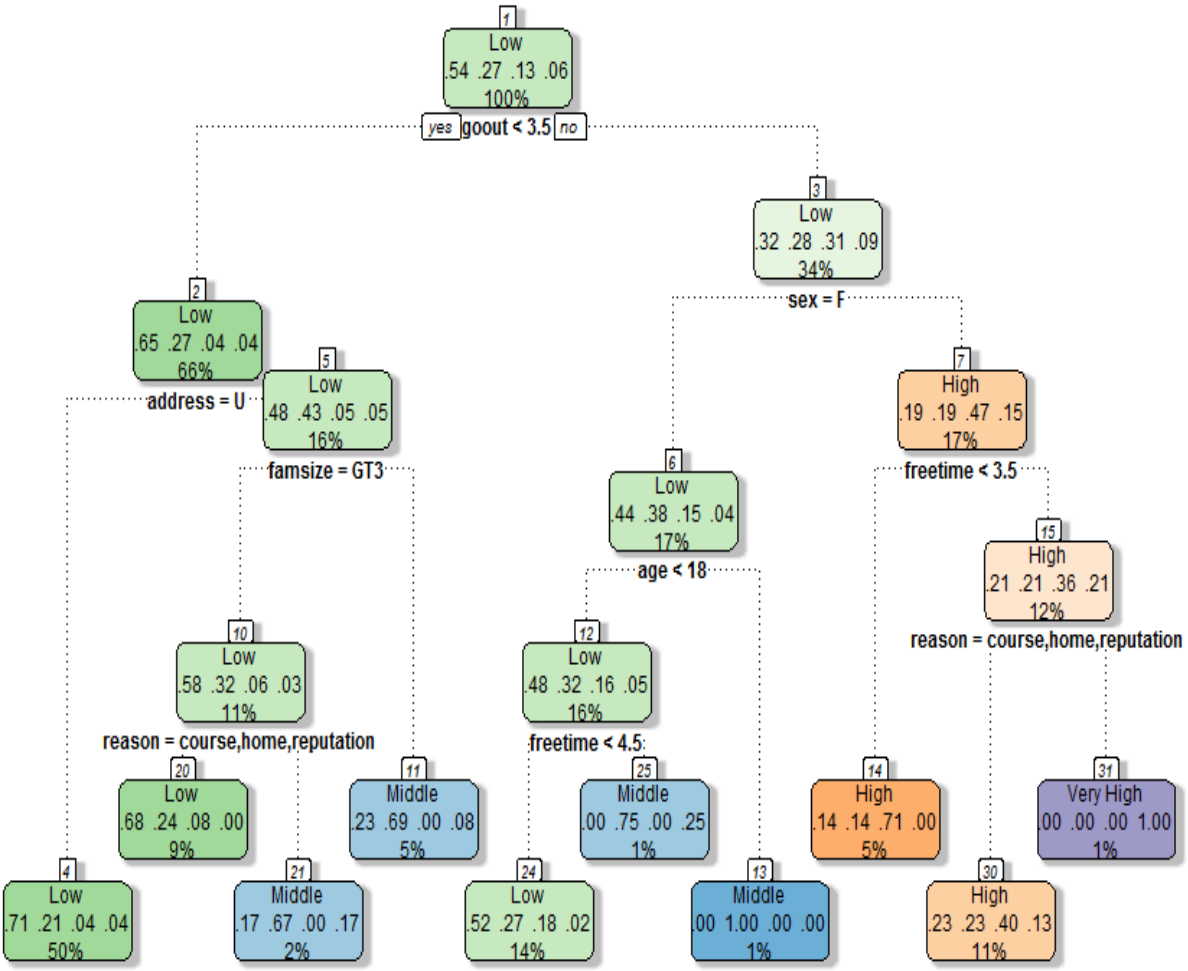
Decision Tree Modeling (2)

Finally I'll select variables to make Decision Tree model.

Variable importance
goout(33) > freetime(20) > sex(17) > reason(11) > famsize(8) > age(7) > address(5)

Decision Tree Modeling (3)

Decision Tree for Student Alcohol Consumption Prediction



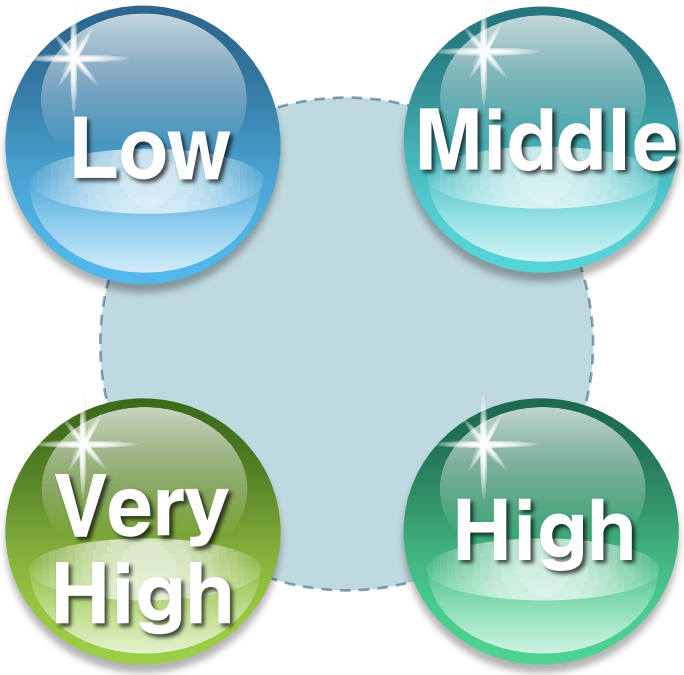
Rattle 2018-8-16 23:25:31 Administrator

Model Verification in Test Data-Set

	Reference			
Prediction	Low	Middle	High	Very High
Low	63	21	6	2
Middle	1	1	0	1
High	4	3	9	6
Very High	0	0	0	2

Decision Tree Modeling (4)

Alcohol consumption Student Segment



Low	Goout < 3.5 / Address = Urban Goout < 3.5 / Address = Rural / Family Size = GT3 / Reason = course,home,reputation Goout > 3.5 / Sex = F / age < 18 / freetime < 4.5
Middle	Goout < 3.5 / Address = Rural / Family Size = GT3 / Reason = other Goout < 3.5 / Address = Rural / Family Size = LE3 Goout > 3.5 / Sex = F / age < 18 / freetime > 4.5 Goout > 3.5 / Sex = F / age > 18
High	Goout > 3.5 / Sex = M / freetime < 3.5 Goout > 3.5 / Sex = M / freetime > 3.5 / Reason = course,home,reputation
Very High	Goout > 3.5 / Sex = M / freetime > 3.5 / Reason = other

Regression Model (1)

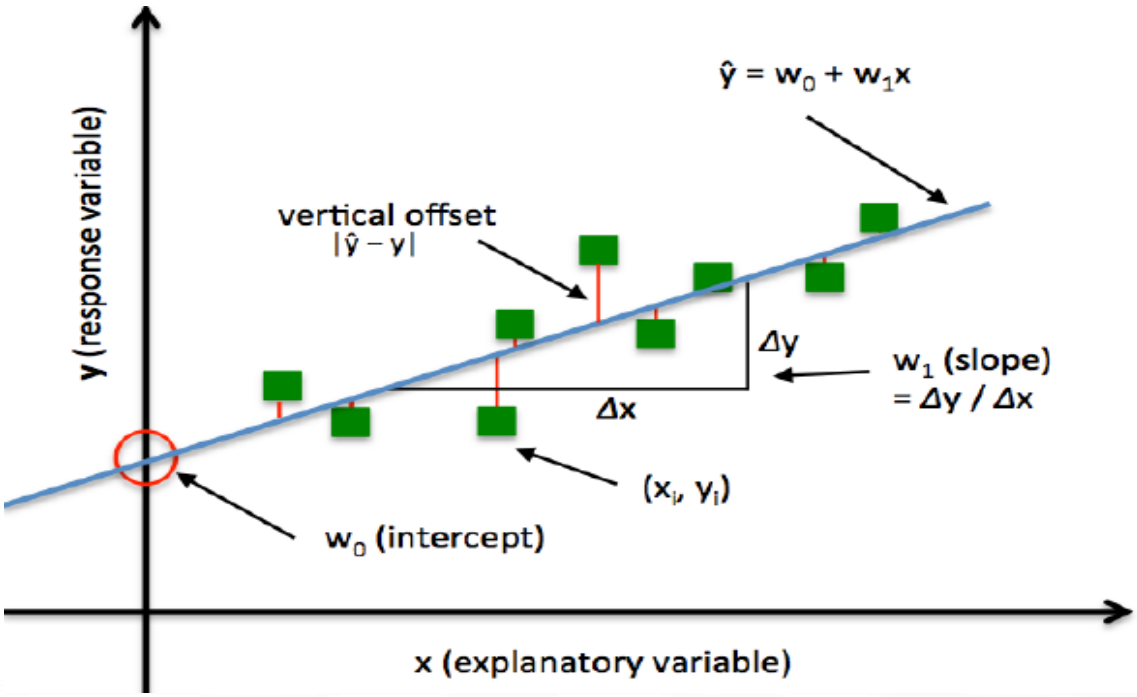
Regression is a statistical technique to determine the linear relationship between two or more variables. Regression is primarily used for prediction and causal inference.

Features of Regression Model

- Fast and easy to model and is particularly useful when the relationship to be modeled is not extremely complex and if you don't have a lot of data.

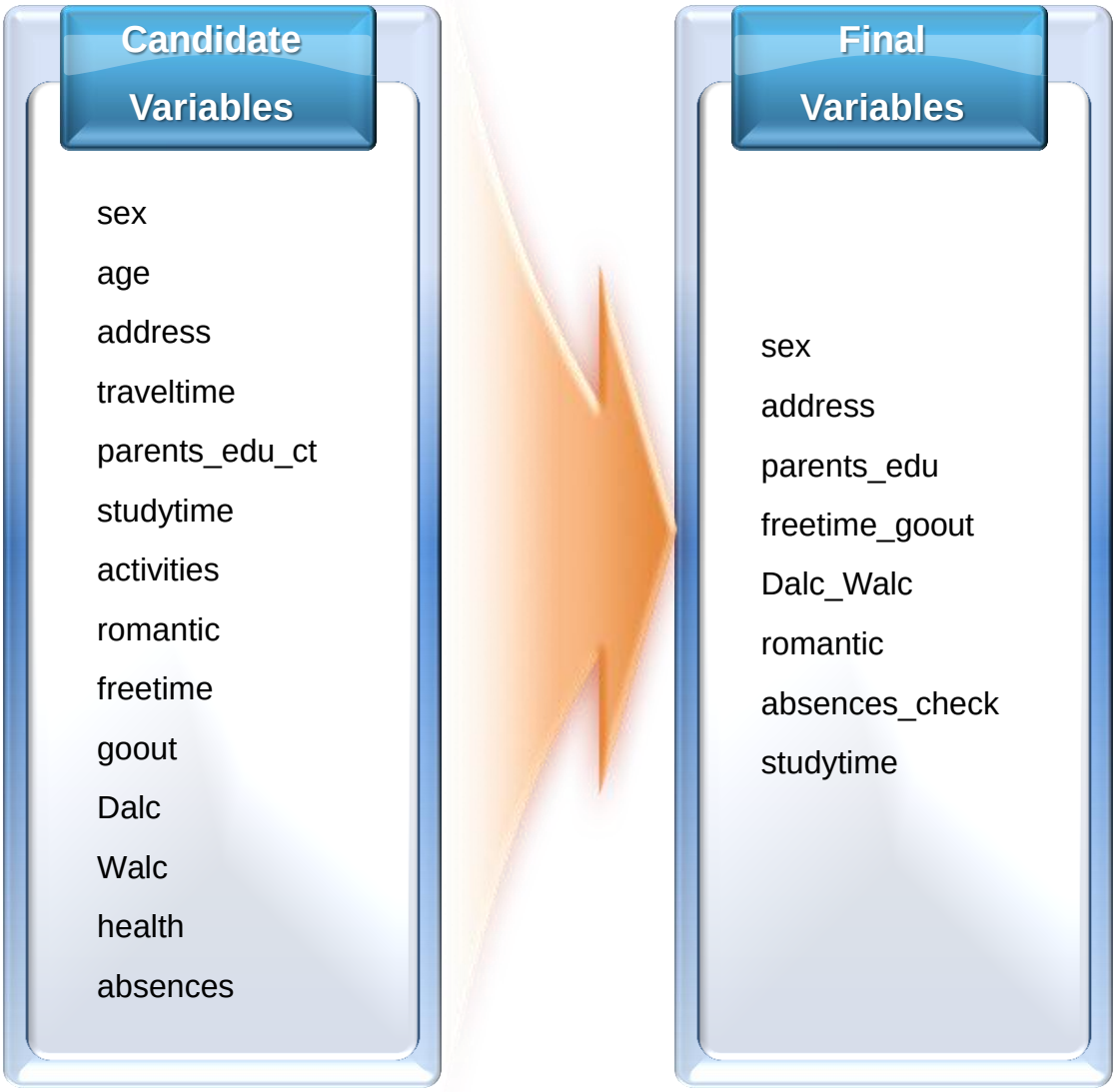
- Very intuitive to understand and interpret.

- Linear Regression is very sensitive to outliers.



Regression Model (2)

Finally I'll select variables to make Regression model.



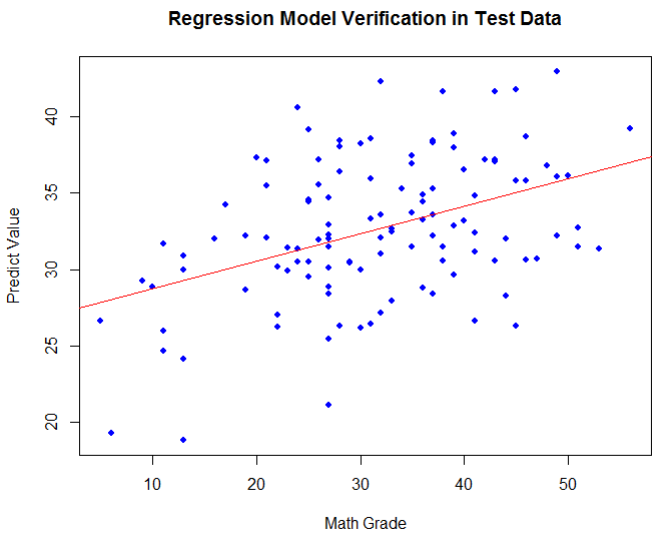
Regression Model (3)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.193	2.554	10.65	< 2e-16 ***
sexF	-2.912	1.142	-2.55	0.0111 *
addressR	-2.666	1.257	-2.12	0.0346 *
parents_edu	1.177	0.265	4.44	0.000012 ***
freetime_goout	1.322	0.426	3.10	0.0021 **
Dalc_Walc	-0.433	0.291	-1.49	0.1375
romanticyes	-2.026	1.102	-1.84	0.0667 .
absences_checkno	-5.138	1.156	-4.44	0.000012 ***
studytime	2.165	0.659	3.29	0.0011 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.2 on 386 degrees of freedom
Multiple R-squared: 0.168, Adjusted R-squared: 0.15
F-statistic: 9.72 on 8 and 386 DF, p-value: 0.00000000000259

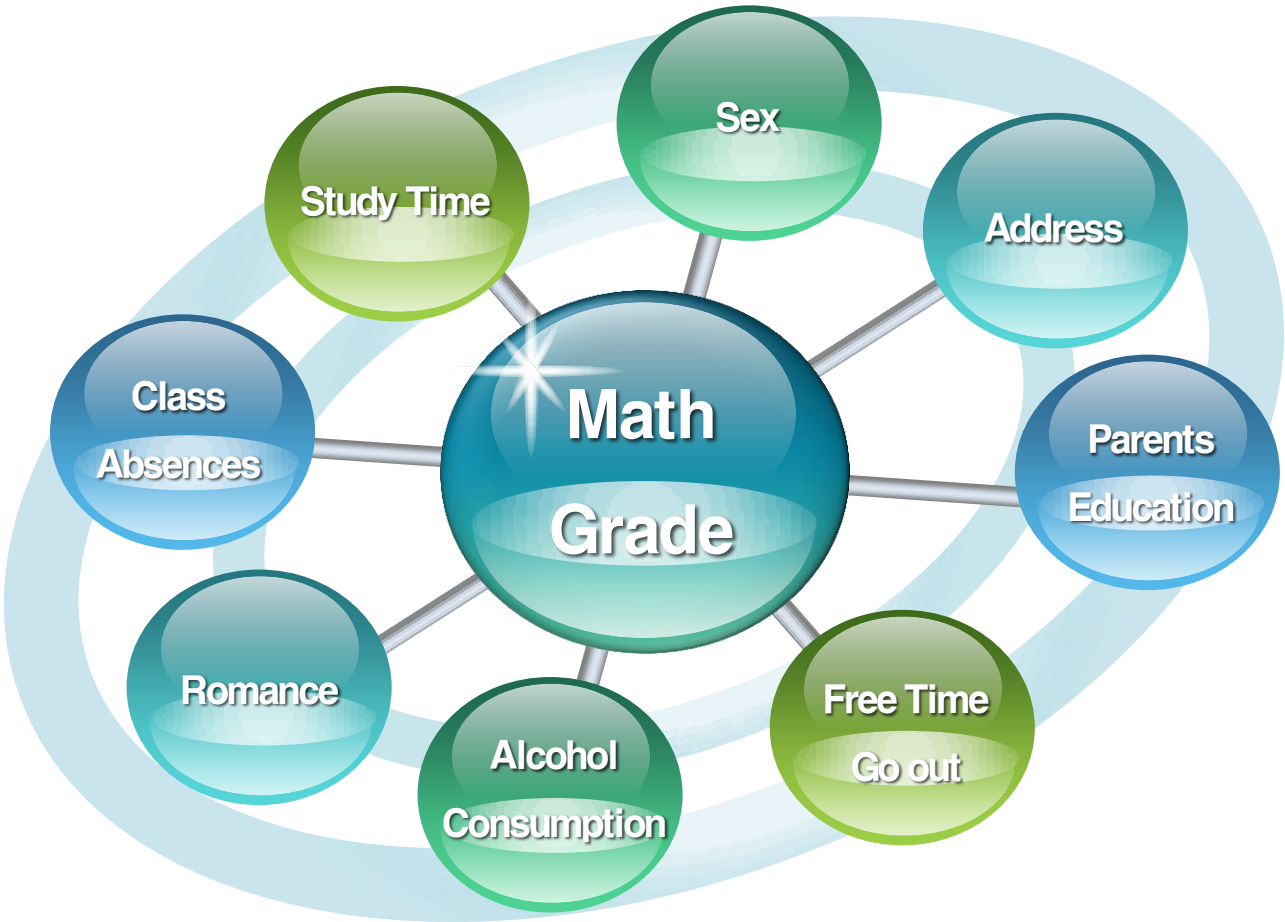


Mean Absolute Error : 8.113

Root Mean Square Error : 9.975

Mean Absolute Percentage Error : 37.51

Regression Model (4)



- Male students have higher math scores than female students.
- Students living in the urban have higher math scores than students in the rural.
- The better educated parents are the higher math scores their students have
- The more free time students have and the less often students go out, the higher math scores they have.
- The less students consume alcohol, the higher math scores they have.
- Students in a relationship have higher math scores than those not in a relationship.
- Students who were absent from school at least 1 time, have higher math scores than students who were never absent from school.
- The more hours students spend to study, the higher math scores they have.

Solutions to attracting students who are expecting graduation from high school

Students tend to be sensitive to the words and actions of their friends who have already experienced when trying new things.

Thus, by attracting students who have higher alcohol consumption level are likely to affect their friends who have zero alcohol consumption on brand choice.

Choose the students in the High / Very High groups classified by the Decision Tree Model, as the main target customers

Strategic solutions

First Strategic

As most of the students are male and go out more often with their friends, posting posters targeted for the male on the walls at busy streets are expected to attract targeted customers at lower costs.

Second Strategic

Most students live in the urban and have a large family.
Solution to making students recognize the alcohol brands of their parents' choice on family-gathering holidays such as Thanksgiving Day, and Christmas.
Conduct DM by giving out letters with discount or free tickets attached to students houses for the Thanksgiving/Christmas Season.