STUDENT MAT PROJECT

Student mat data is in the form of csv file. This data comes under classification and regression data. The data we are predicting is G3.

Data consisting of many variables like

Name of variable	Meaning of it
School	Student's School
Sex	Student's Sex
Age	Student's Age
Address	Student's Home Address Type
FamSize	Family Size
Pstatus	Parent's Cohabitation Status
Medu	Mother's Education
Fedu	Father's Education
Mjob	Mother's Job
Fjob	Father's Job
Reason	Reason To Choose This School
Guardian	Student's Guardian
TravelTime	Home To School Travel Time
StudyTime	Weekly Study Time
Failures	Number Of Past Class Failures
Schoolsup	Extra Educational Support
Famsup	Family Educational Support
Paid	Extra Paid Classes Within The Course Subject (Math Or Portuguese)
Activities	Extra-Curricular Activities
Nursery	Attended Nursery School
Higher	Wants To Take Higher Education
Internet	Internet Access At Home
Romantic	With A Romantic Relationship
Famrel	Quality Of Family Relationships
Freetime	Free Time After School
Goout	Going Out With Friends
Dalc	Workday Alcohol Consumption
Walc	Weekend Alcohol Consumption
Health	Current Health Status
Absences	Number Of School Absences
G1	First Period Grade
G2	Second Period Grade
G3	Final Grade

Now from this only few are considered for predicting the G3.

- Sex
- Age
- Address
- Pstatus
- Medu
- Fedu
- Mjob
- Fjob
- Studytime
- Traveltime
- Failures
- Higher
- Internet
- Gout
- Dalc
- Walc
- Health
- Absences
- G1
- G2
- G3

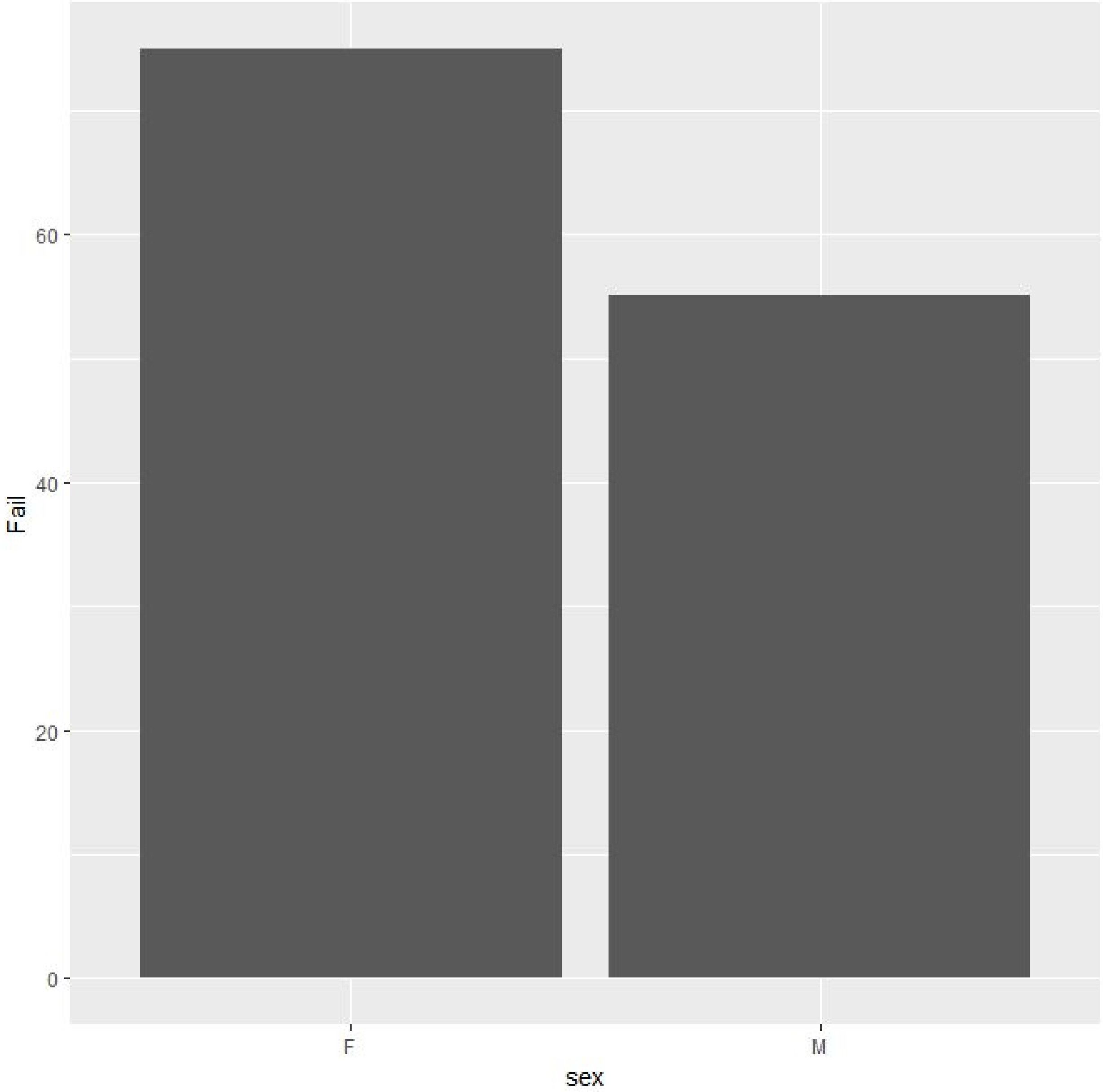
Analysing the data:

```
student_mat1<-student_mat %>%
+   mutate(pass=ifelse(G3>10,1,0),fail=ifelse(G3<10,1,0))%>%
+   filter(sex=="F"|sex=="M")%>%
+   group_by(sex)%>%
+   summarise(Pass=sum(pass),
+             Fail=sum(fail))
> student_mat1
# A tibble: 2 x 3
  sex    Pass  Fail
  <chr> <dbl> <dbl>
1 F      103.   75.
2 M      106.   55.
```

We have more male persons passed than female persons.

Plotting the fails of females and males by bar graph

```
student_mat1 %>%
+   ggplot(aes(x=sex,y=Fail))+geom_bar(stat = 'identity')
```



Females are more failed than males

Grouping the Week day alcohol consumption by G3 mean.

```
student_mat%>%
+   group_by(Walc)%>%
+   aggregate(G3~Walc,data=.,mean)%>%
+   arrange(desc(G3))
```

```
  Walc        G3
1    1 10.735099
2    3 10.725000
3    5 10.142857
4    2 10.082353
5    4  9.686275
```

Grouping the Working day alcohol consumption by G3 mean.

```
> student_mat%>%
+   group_by(Dalc)%>%
+   aggregate(G3~Dalc,data=.,mean)%>%
+   arrange(desc(G3))
  Dalc        G3
1    1 10.731884
2    5 10.666667
3    3 10.500000
4    4  9.888889
5    2  9.253333
```

So if we compare both week day and working day on week day its more consumption.

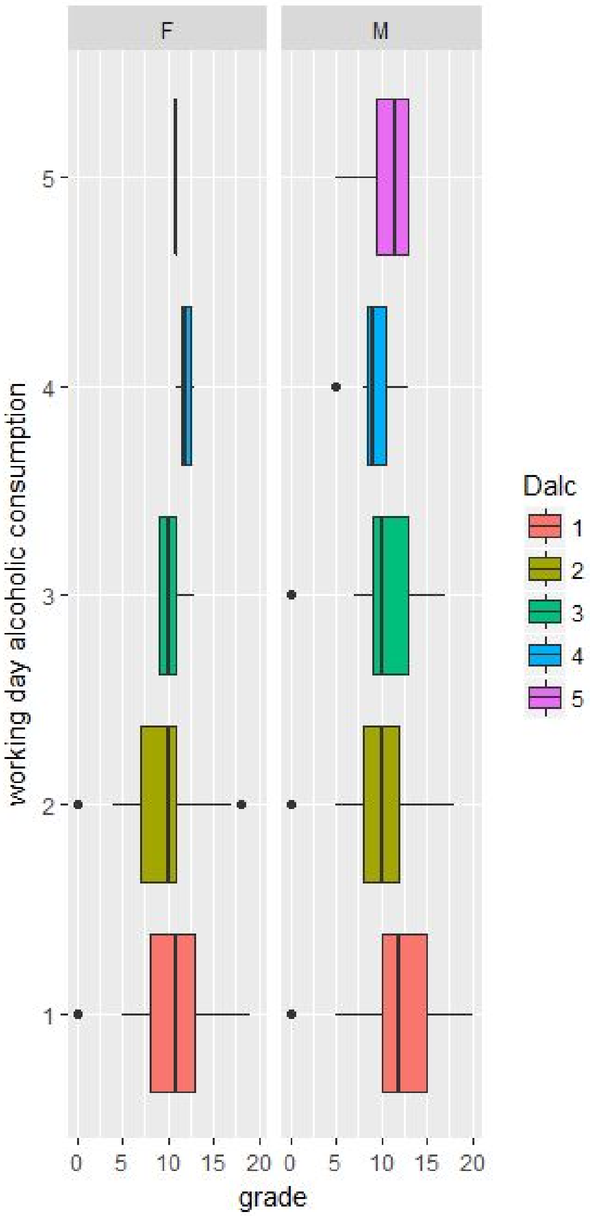
Grouping the goout by G3 mean.

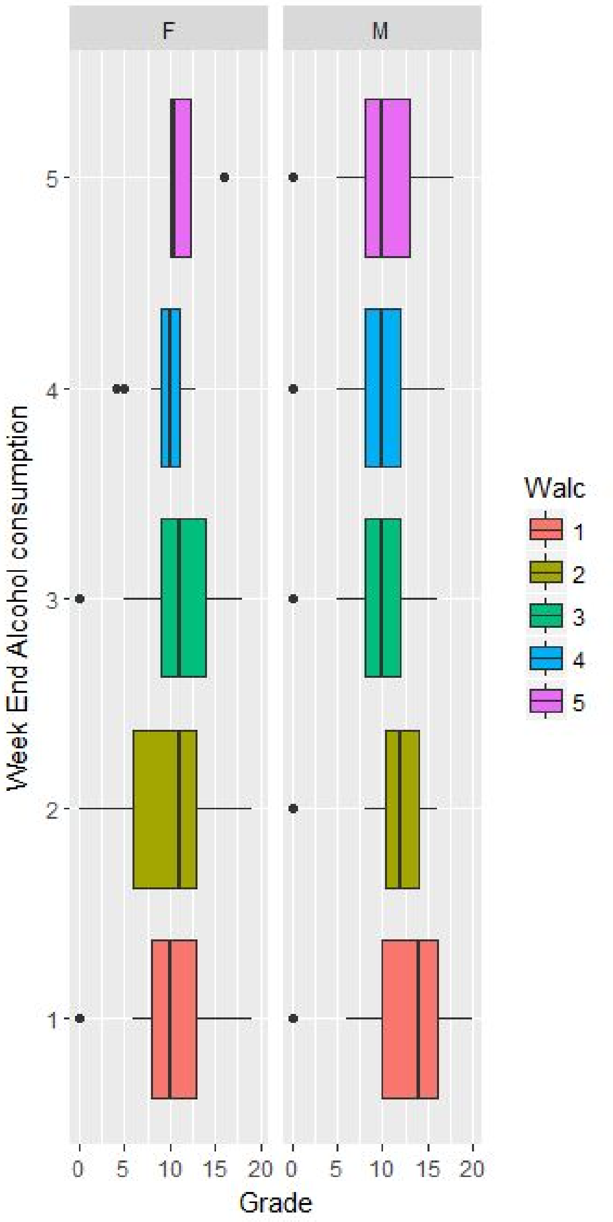
```
student_mat%>%
+   group_by(goout)%>%
+   summarise(averagescore=mean(G3,na.rm = TRUE))%>%
+   arrange(desc(averagescore))
# A tibble: 5 x 2
  goout averagescore
  <fct>        <dbl>
1 2             11.2
2 3             11.0
3 1              9.87
4 4              9.65
5 5              9.04
```

```
student_mat$Dalc<-as.factor(student_mat$Dalc)
> student_mat$Walc<-as.factor(student_mat$Walc)
```

Plotting the boxplot on alcoholic consumption by grade and sex

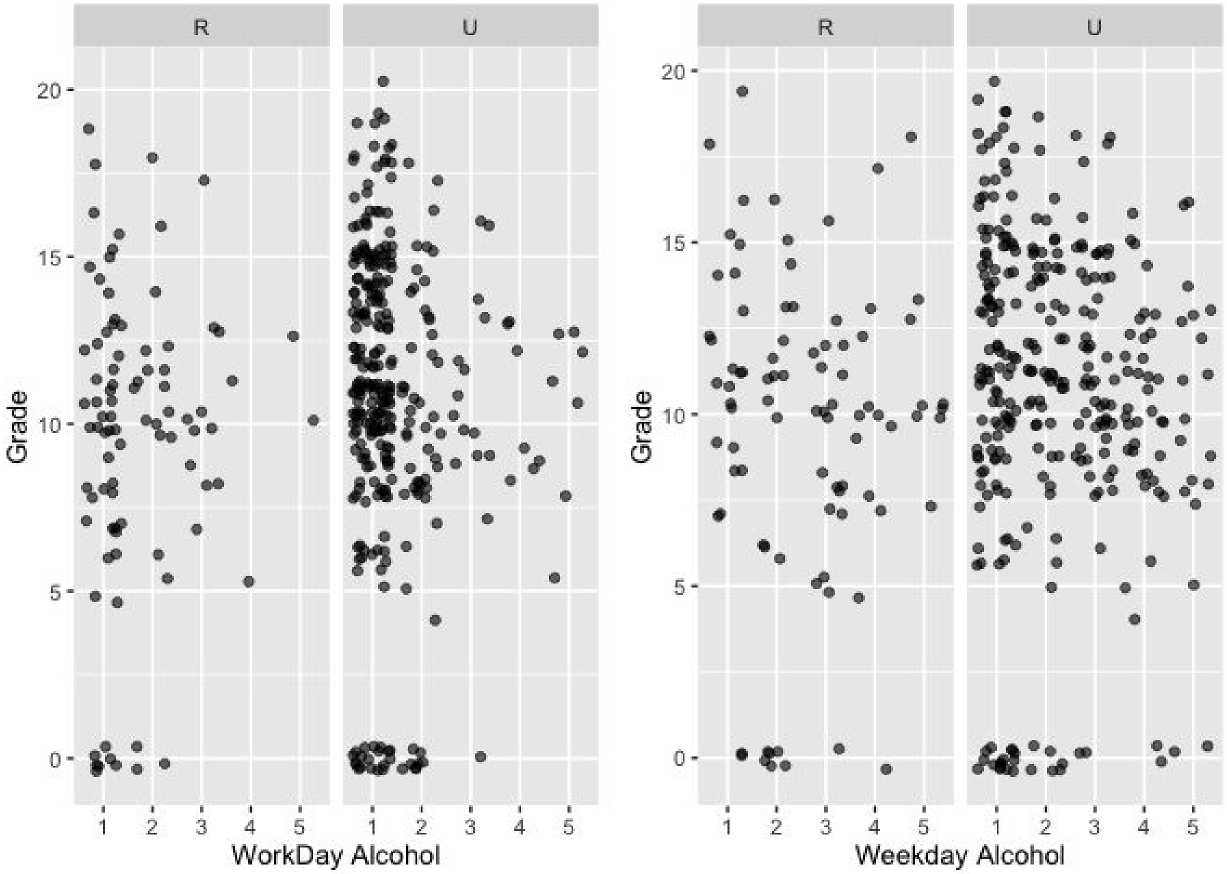
```
> g1a<-student_mat%>%
+   ggplot(aes(x=Dalc,y=G3,fill=Dalc))+
+   geom_boxplot()+
+   coord_flip()+
+   xlab("working day alcoholic consumption")+
+   ylab("grade")+
+   facet_grid(~sex)
> g1b<-student_mat %>%
+   ggplot(aes(x=Walc, y=G3, fill= Walc))+
+   geom_boxplot()+
+   coord_flip()+
+   xlab("Week End Alcohol consumption")+
+   ylab("Grade")+
+   facet_grid(~sex)
```





Plotting the boxplot on alcoholic consumption by grade and address

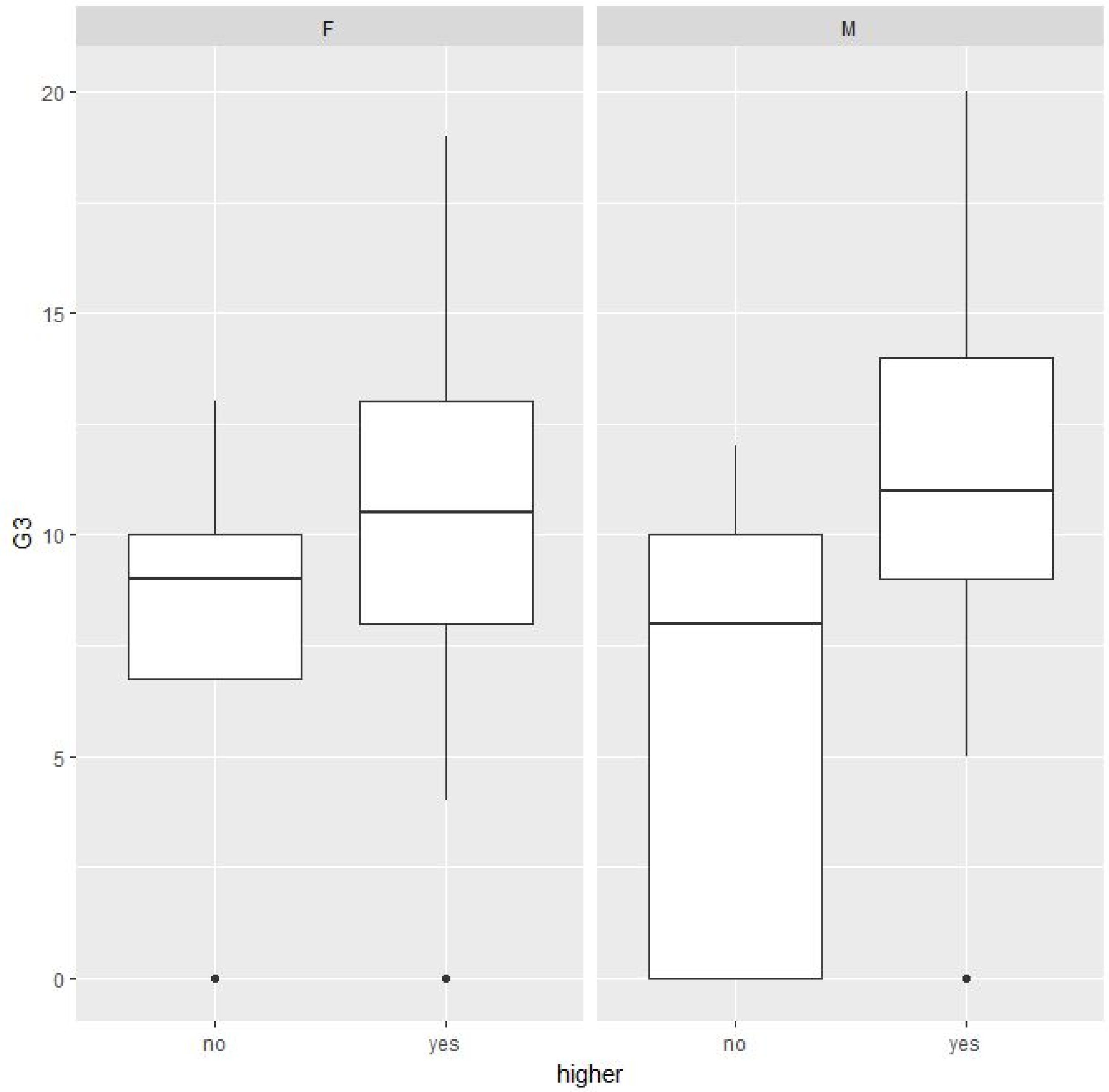
```
> g2a<-student_mat %>%
+   group_by(address)%>%
+   ggplot(aes(x=factor(Dalc), y= G3))+
+   geom_jitter(alpha=0.6)+
+   scale_x_discrete("WorkDay Alcohol")+
+   scale_y_continuous("Grade")+
+   facet_grid(~address)
> g2b<-student_mat %>%
+   group_by(address)%>%
+   ggplot(aes(x=factor(Dalc), y= G3))+
+   geom_jitter(alpha=0.6)+
+   scale_x_discrete("WeekDay Alcohol")+
+   scale_y_continuous("Grade")+
+   facet_grid(~address)
grid.arrange(g2a,g2b,ncol2)
```



In urban areas the alcoholic consumption is more both on work days and week days.

Plotting the boxplot on higher education by grade and sex

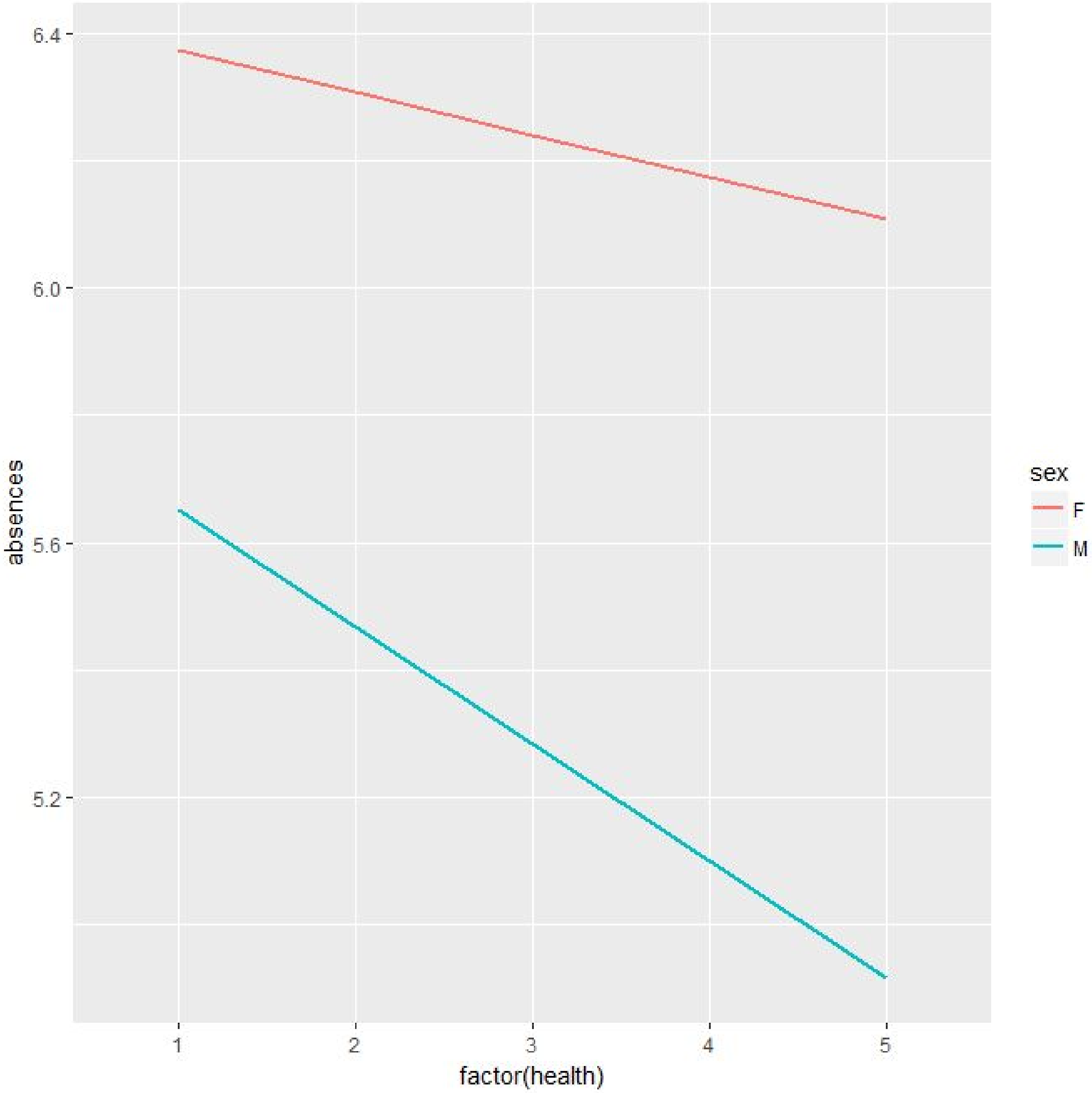
```
        student_mat%>%
+   ggplot(aes(x=higher, y=G3))+
+   geom_boxplot()+
+   facet_grid(~sex)
```



Male who opting for higher education having the more grade.

Plotting the line on health by absence and sex

```
student_mat%>%
+   group_by(sex)%>%
+   ggplot(aes(x=factor(health), y=absences, color=sex))+
+   geom_smooth(aes(group=sex), method="lm", se=FALSE)
```

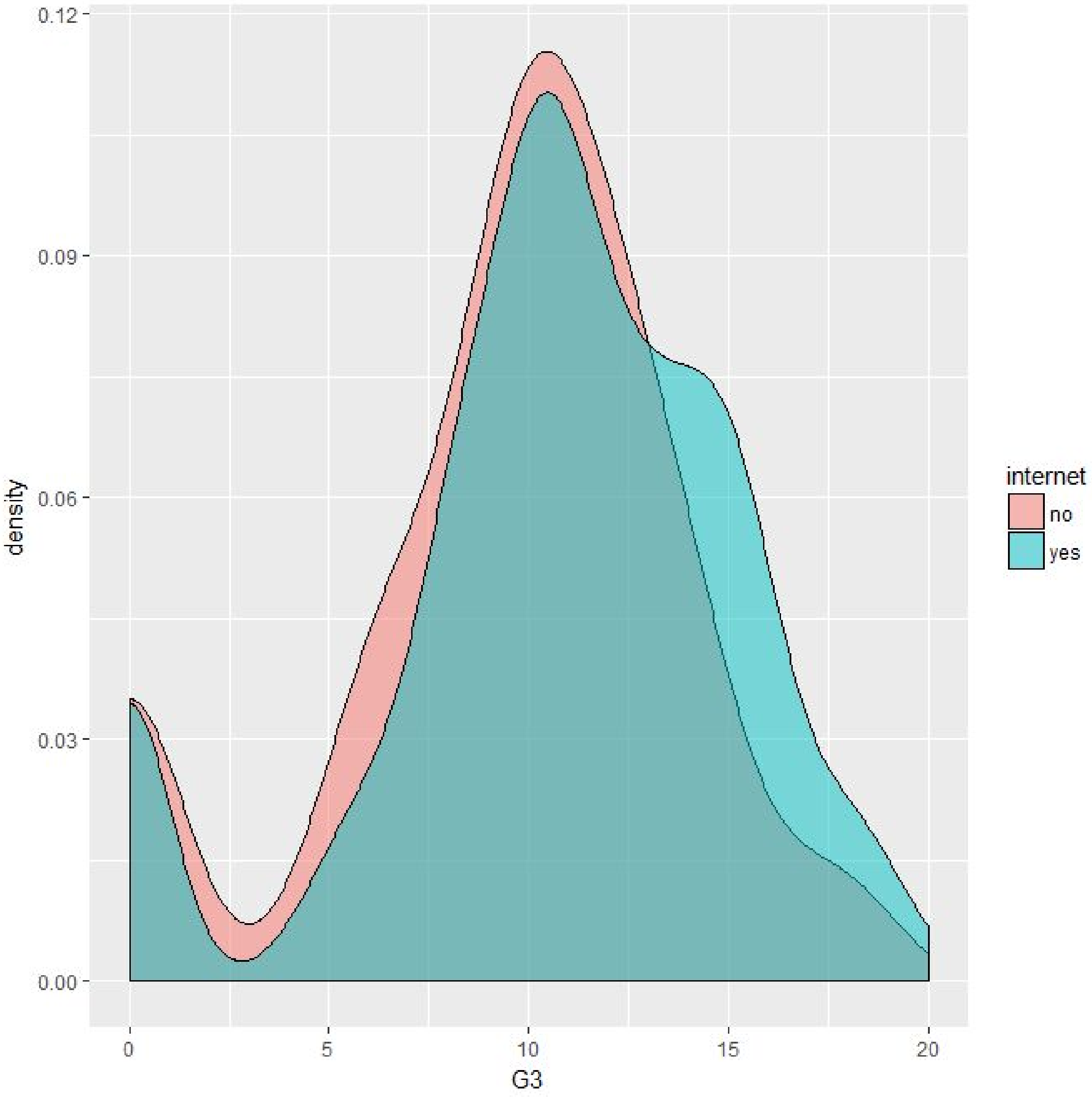


As the ill heath increasing absence rate for female increasing.

Plotting the area on internet by grade

```
> student_mat%>%
+   group_by(internet)%>%
+   ggplot(aes(x=G3, fill=internet))+
+   geom_density( alpha=0.5)
> group_by(internet)
```

A



Grade is more where there is no internet usage.

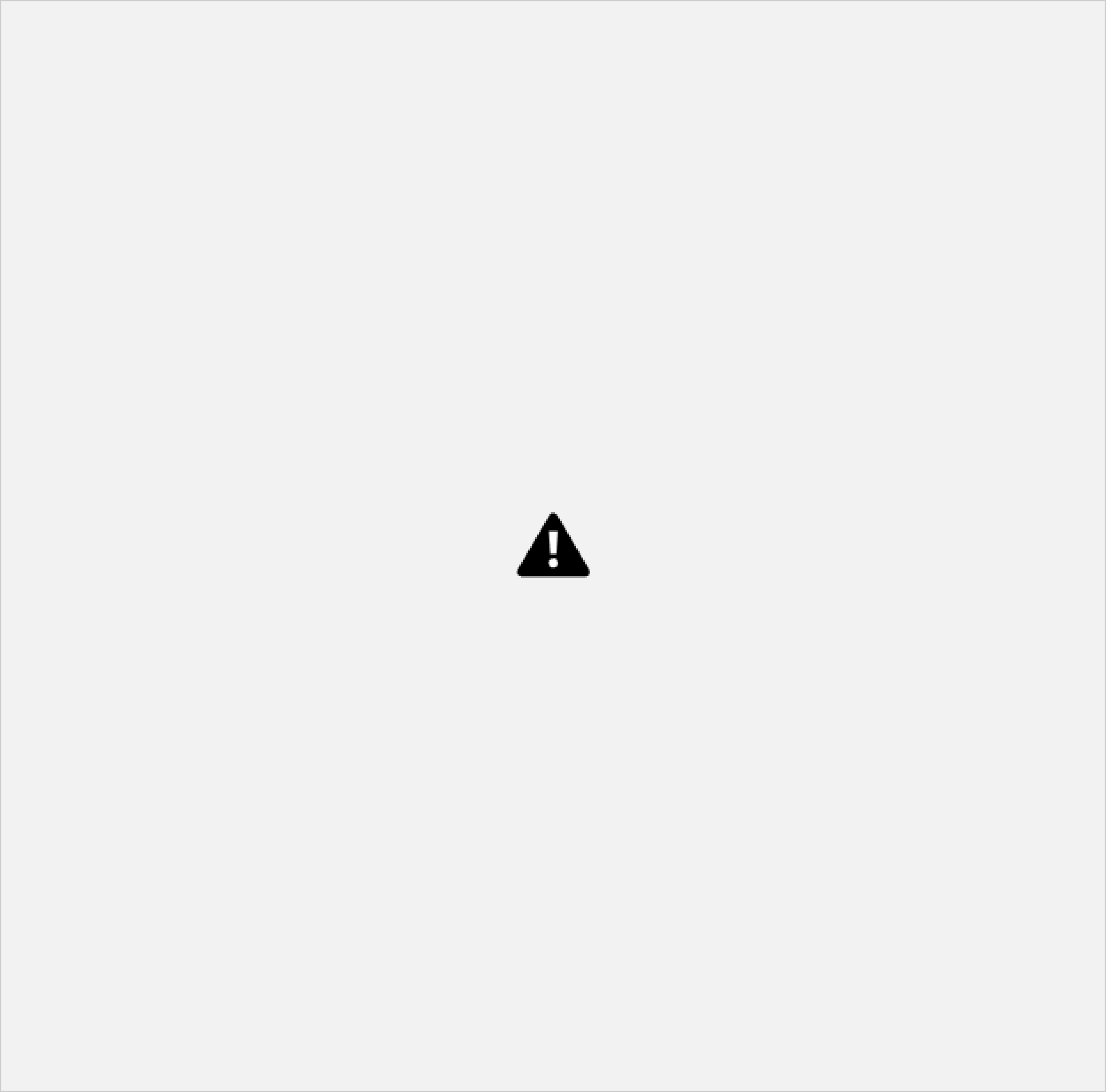
DECISION TREEE REGRESSION

```
> newdata<- student_mat%>%select(sex, age, address,Pstatus, Medu, Fedu, Mjob, 
Fjob,studytime,traveltime,failures,higher,internet,
+                          goout, Dalc,Walc,health, absences,G1, G2, G3)
> tree <- rpart(G3 ~ .,
+                data = newdata,
+                method = "class")
> imp <- varImp(tree)
> rownames(imp)[order(imp$Overall, decreasing=TRUE)]
 [1] "G2"         "G1"         "absences"   "Walc"       "failures"   "Fjob"
"Mjob"
 [8] "Dalc"       "Medu"       "health"     "studytime"  "goout"
"traveltime" "sex"
[15] "age"        "address"    "Pstatus"    "Fedu"       "higher"
"internet"
> printcp(tree)

Classification tree:
rpart(formula = G3 ~ ., data = newdata, method = "class")

Variables actually used in tree construction:
[1] absences Fjob     G1       G2

Root node error: 339/395 = 0.85823

n= 395


CP nsplit rel error  xerror     xstd
1 0.094395      0   1.00000 1.00000 0.020450
2 0.067847      1   0.90560 0.89381 0.024781
3 0.060472      2   0.83776 0.84366 0.026206
4 0.032448      4   0.71681 0.76696 0.027807
5 0.026549      7   0.61947 0.71386 0.028560
6 0.022124      9   0.56637 0.66667 0.029007
7 0.011799     11   0.52212 0.59292 0.029309
8 0.010000     13   0.49853 0.58702 0.029313
```

Plotting the cp tree

```
> plotcp(tree)
```

Here CP means complex parameters. Out of these many parameters it has taken only G1, G2, FJOB, absence as the independent variables.



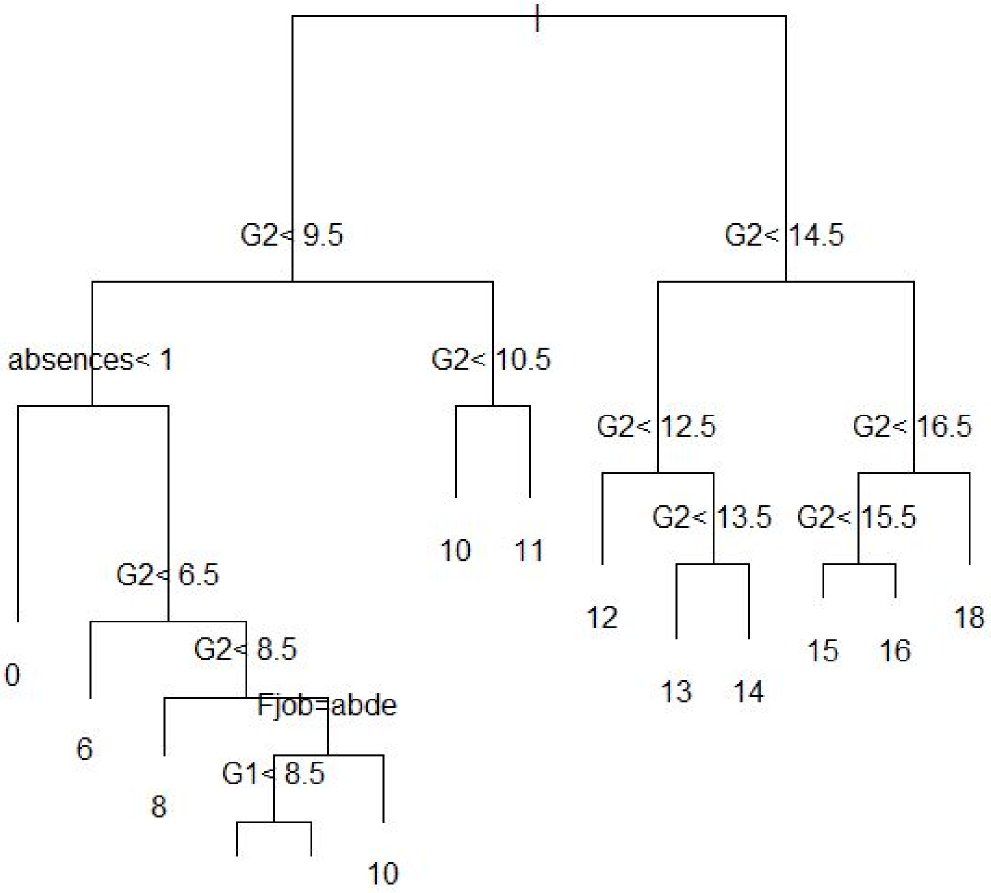
As the error decreases the size of the tree increases.

Plotting the decision tree

```
plot(tree)
> text(tree)
```

Ii

I



It is the trees of decision regression.