statistics-project

Jasleen Kaur

29/11/2019

```
##Importing some important libraries
library(devtools)
```

```
## Loading required package: usethis
```

```
library(easyGgplot2)
```

```
## Loading required package: ggplot2
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(reshape2)
library(ggplot2)
library(ggpubr)
```

```
## Loading required package: magrittr
```

```
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.
```

```
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and
```

```
##       if Arial Narrow is not on your system, please see http://bit.ly/arialnarrow
```

```
library(easyGgplot2)
library(faraway)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(glmnet)
```

```
## Loading required package: Matrix

## Loaded glmnet 3.0
```

```
library(ISLR)
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:faraway':
##
##     melanoma
```

Reading the file

```
##Reading the data file
data_ver_1 <- read.csv(file="student-por.csv", header=TRUE, sep=";")
#We have G3 as our response variables which is int column
head(data_ver_1, n=5)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob reason
## 1     GP   F  18       U     GT3       A    4    4 at_home  teacher course
## 2     GP   F  17       U     GT3       T    1    1 at_home    other course
## 3     GP   F  15       U     LE3       T    1    1 at_home    other  other
## 4     GP   F  15       U     GT3       T    4    2  health services   home
## 5     GP   F  16       U     GT3       T    3    3   other    other   home
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        0       yes     no   no         no
## 4   mother          1         3        0        no    yes   no        yes
## 5   father          1         2        0        no    yes   no         no
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
```

```
## 5     yes    yes       no       no      4        3     2    1    2      5
##   absences G1 G2 G3
## 1        4  0 11 11
## 2        2  9 11 11
## 3        6 12 13 12
## 4        0 14 14 14
## 5        0 11 13 13
```

EXPLORATORY ANALYSIS

1.1-Know the data

```
#Observe structure of the data
#We have 17 categorical variables and 16 int features including the response G3
str(data_ver_1)
```

```
## 'data.frame':    649 obs. of  33 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 0 0 0 0 0 0 0 0 ...
##  $ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
##  $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
##  $ paid      : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 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  4 2 6 0 0 6 0 2 0 0 ...
##  $ G1        : int  0 9 12 14 11 12 13 10 15 12 ...
##  $ G2        : int  11 11 13 14 13 12 12 13 16 12 ...
##  $ G3        : int  11 11 12 14 13 13 13 13 17 13 ...
```

```r
#check the sample size and number of feautres in dataset
#dimension of the data
dim(data_ver_1) #649 samples and 33 variables (32 predictors+1 response)
```

```
## [1] 649  33
```

```r
#summary of the data
summary(data_ver_1) #summary stats for all columns
```

```
##  school   sex          age        address famsize   Pstatus
##  GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80
##  MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569
##                   Median :17.00
##                   Mean   :16.74
##                   3rd Qu.:18.00
##                   Max.   :22.00
##       Medu            Fedu             Mjob           Fjob
##  Min.   :0.000   Min.   :0.000   at_home :135   at_home : 42
##  1st Qu.:2.000   1st Qu.:1.000   health  : 48   health  : 23
##  Median :2.000   Median :2.000   other   :258   other   :367
##  Mean   :2.515   Mean   :2.307   services:136   services:181
##  3rd Qu.:4.000   3rd Qu.:3.000   teacher : 72   teacher : 36
##  Max.   :4.000   Max.   :4.000
##         reason      guardian     traveltime      studytime
##  course     :285   father:153   Min.   :1.000   Min.   :1.000
##  home       :149   mother:455   1st Qu.:1.000   1st Qu.:1.000
##  other      : 72   other : 41   Median :1.000   Median :2.000
##  reputation:143                 Mean   :1.569   Mean   :1.931
##                                 3rd Qu.:2.000   3rd Qu.:2.000
##                                 Max.   :4.000   Max.   :4.000
##     failures      schoolsup famsup     paid     activities nursery
##  Min.   :0.0000   no :581   no :251   no :610   no :334    no :128
##  1st Qu.:0.0000   yes: 68   yes:398   yes: 39   yes:315    yes:521
##  Median :0.0000
##  Mean   :0.2219
##  3rd Qu.:0.0000
##  Max.   :3.0000
##  higher    internet  romantic      famrel         freetime
##  no : 69   no :151   no :410   Min.   :1.000   Min.   :1.00
##  yes:580   yes:498   yes:239   1st Qu.:4.000   1st Qu.:3.00
##                                Median :4.000   Median :3.00
##                                Mean   :3.931   Mean   :3.18
##                                3rd Qu.:5.000   3rd Qu.:4.00
##                                Max.   :5.000   Max.   :5.00
##      goout            Dalc            Walc          health
##  Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:2.000
##  Median :3.000   Median :1.000   Median :2.00   Median :4.000
##  Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean   :3.536
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:5.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000
##     absences            G1              G2              G3
##  Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00
```

```
##  1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00   1st Qu.:10.00
##  Median : 2.000   Median :11.0   Median :11.00   Median :12.00
##  Mean   : 3.659   Mean   :11.4   Mean   :11.57   Mean   :11.91
##  3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :32.000   Max.   :19.0   Max.   :19.00   Max.   :19.00
```

1.2- Cleaning and shaping data

```
#Drop the grade 1 and grade 2 columns and also fam_size ,address dosent seem to affect the final grade
data_ver_2 = subset(data_ver_1, select = -c(G1, G2,famsize,address))
## We can check for any missing values in dataset
cat("Number of missing values is ",sum(is.na(data_ver_2)),"\n\n")
```

```
## Number of missing values is  0
```

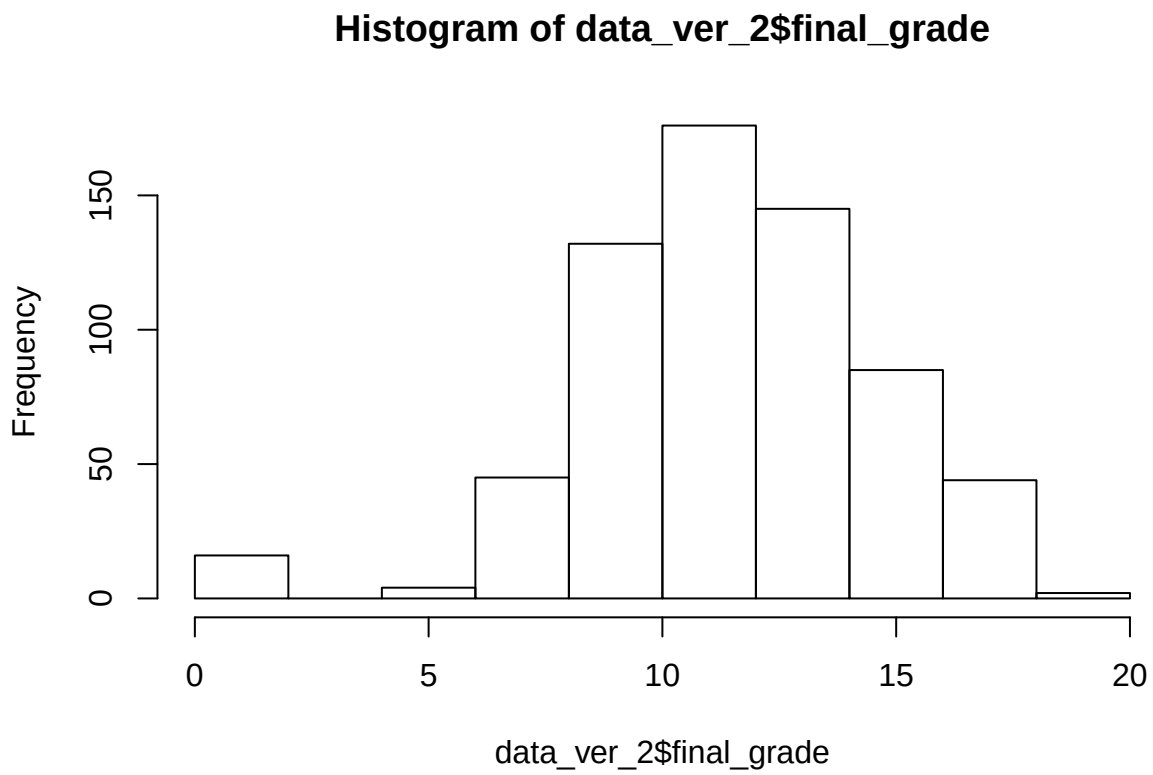
```
##Rename names of columns to some convinient names
#change the column name to more explicit name
#names(data_ver_2)-recheck the names
#change the column name to more explicit name
names(data_ver_2) <- c("school","sex","age","parents_cohab", "mom_edu","dad_edu","mom_job",
                       "family_relationship","free_time","social_time","workday_alch","weekend_alch",
                       "health","absences","final_grade")
##Lets check the structure of data again
##649 obs. of  29 variables (28 predictors and 1 response)
str(data_ver_2)
```

```
## 'data.frame':    649 obs. of  29 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 ...
##  $ parents_cohab      : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
##  $ mom_edu            : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ dad_edu            : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ mom_job            : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
##  $ dad_job            : 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 ...
##  $ travel_time        : int  2 1 1 1 1 1 1 2 1 1 ...
##  $ study_time         : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ edu_sup            : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
##  $ family_sup         : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
##  $ paid               : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 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 ...
##  $ family_relationship: int  4 5 4 3 4 5 4 4 4 5 ...
##  $ free_time          : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ social_time        : int  4 3 2 2 2 2 4 4 2 1 ...
```

```
##  $ workday_alch       : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ weekend_alch       : 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  4 2 6 0 0 6 0 2 0 0 ...
##  $ final_grade        : int  11 11 12 14 13 13 13 13 17 13 ...
```

1.3- Data Visualization

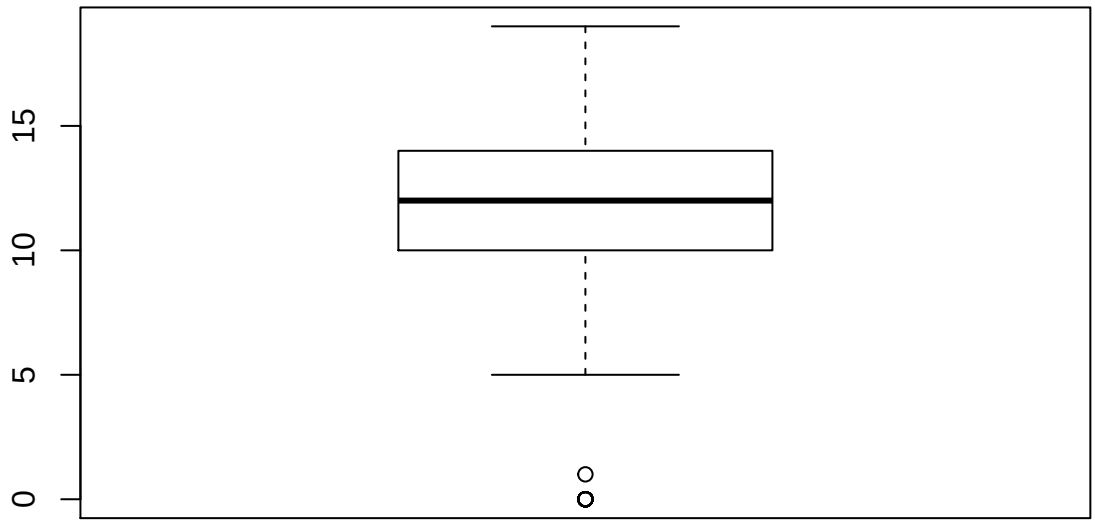
```
#histogram of grade,there seem some students with zero grade
hist(data_ver_2$final_grade)
```




```
#there 15 of 649 students got grade of 0
cat("Students with grade zero ",sum(data_ver_2$final_grade==0),"\n\n")
```

```
## Students with grade zero  15
```

```
#boxplot of the grade
boxplot <- boxplot(data_ver_2$final_grade)
```

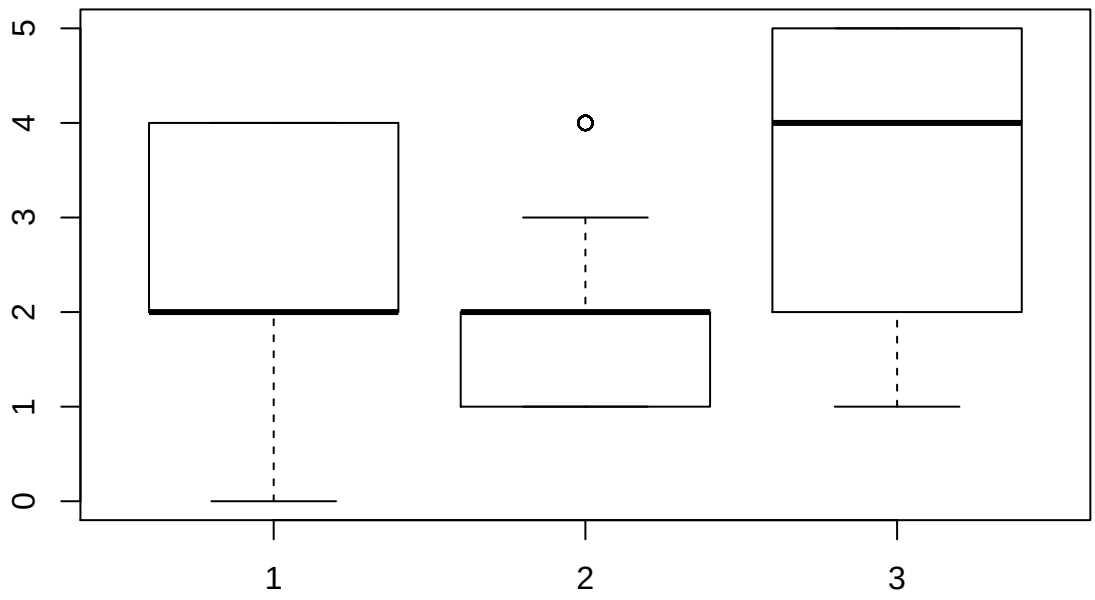



```
boxplot$out
```

```
##  [1] 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

1.4- Outliers

```
## We can check for outliers in numerical predictors from data_ver_2 like mom_edu,study_time,workday_al
##From the list of numerical variables absences seem more like response as it is inversely conveying th
boxplot(data_ver_2$mom_edu,data_ver_2$study_time,data_ver_2$health)
```

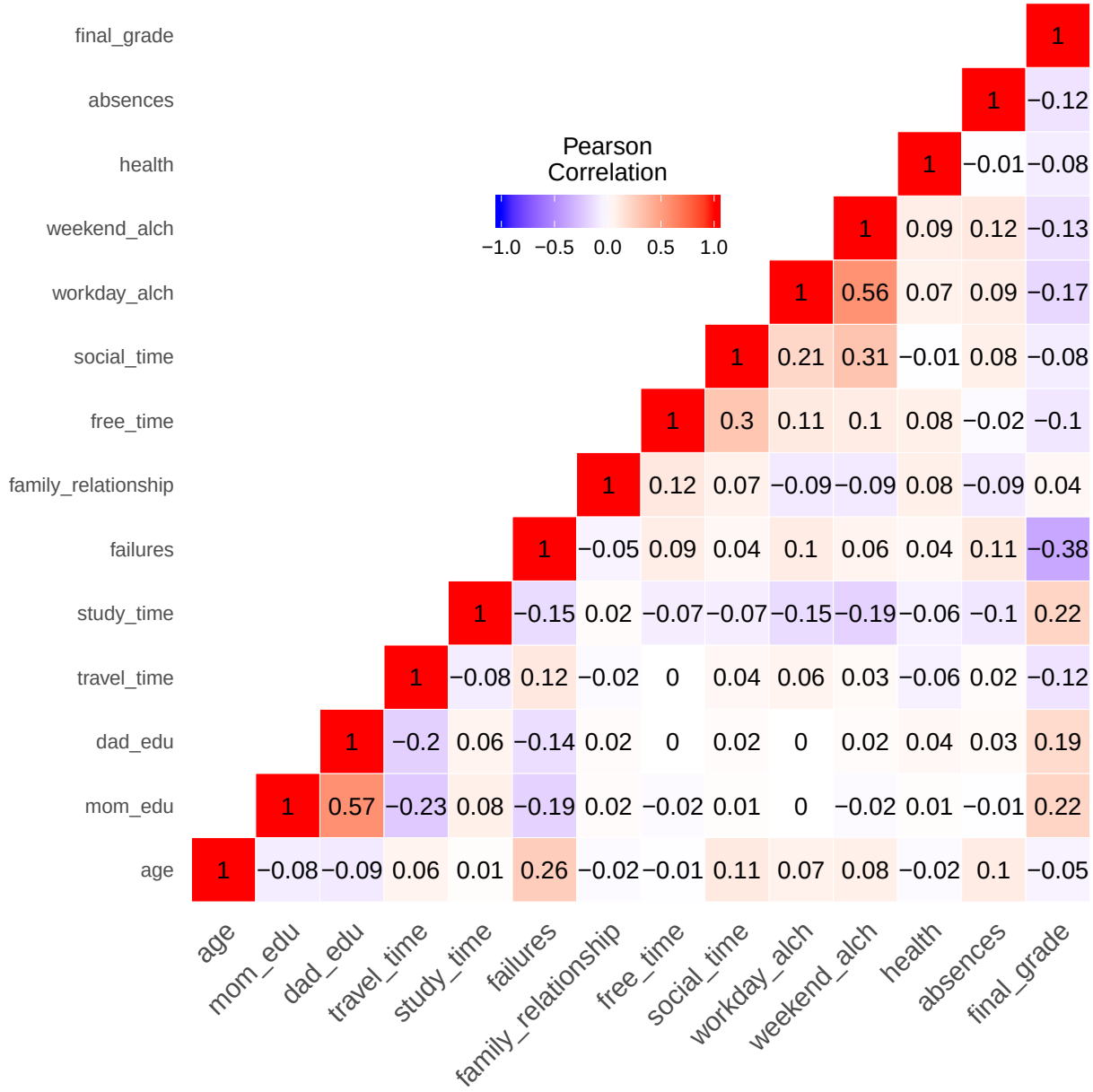



```
##No significant number of outliers found so we can skip rmeoving outliers from data and also most of t
```

1.4- FInd relationship between variables,if any

```
getcormatrix<- function (data_ver_2){
  #Filter  numerical column
numeric_col = dplyr::select_if(data_ver_2, is.numeric)

#Correlation matrix
cormat <- round(cor(numeric_col, method="kendall"),2)
cormat
get_upper_tri <- function(cormat){
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
  }
upper_tri <- get_upper_tri(cormat)
melted_cormat <- melt(upper_tri, na.rm = TRUE)
#Visualize the correltaion
ggheatmap <- ggplot(melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
  theme_minimal()+ # minimal theme
```

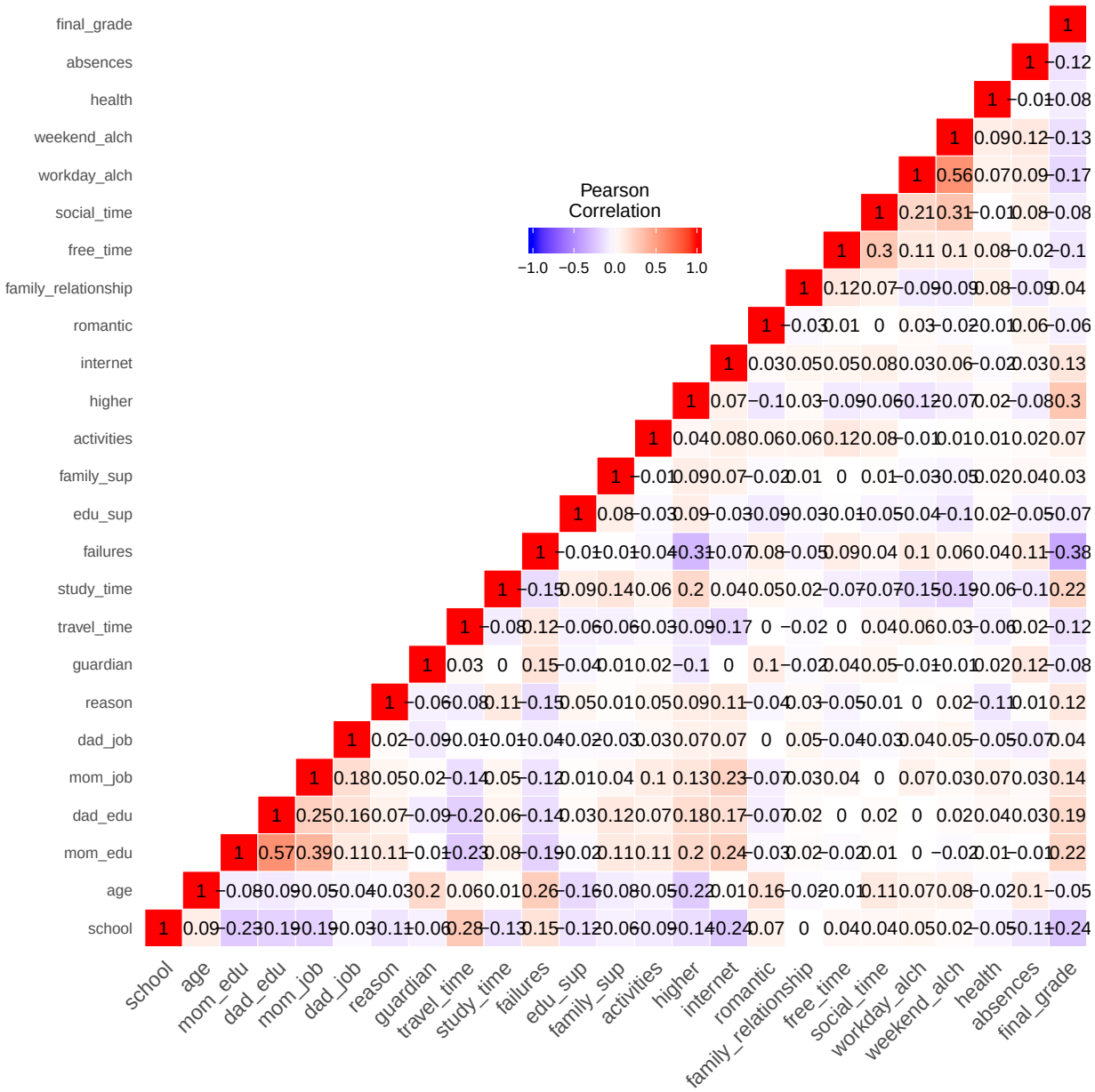
```
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1))+
 coord_fixed()

reorder_cormat <- function(cormat){
# Use correlation between variables as distance
dd <- as.dist((1-cormat)/2)
hc <- hclust(dd)
cormat <-cormat[hc$order, hc$order]
}

ggheatmap +
geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.grid.major = element_blank(),
  panel.border = element_blank(),
  panel.background = element_blank(),
  axis.ticks = element_blank(),
  legend.justification = c(1, 0),
  legend.position = c(0.6, 0.7),
  legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                title.position = "top", title.hjust = 0.5))
}
getcormatrix(data_ver_2)
```




```
## From the correlation matrix below we can see workday_alcoholic/weekend_alcoholic ,mom_edu/dad_edu ar
##Looks like study_time,dad_edu/mom_edu are explaining our response more positively related to response
## SOme predictors have negative correlation with final_grade mentioned in decreasing order like failur
##Some predcitos show least correlation with final_grade like family_relationship,age
```

```
## We got important predictors from above matrix.Lets transform few important categorical to numerical
data_ver_2$school<-as.numeric(data_ver_2$school)
data_ver_2$mom_job<-as.numeric(data_ver_2$mom_job)
data_ver_2$dad_job<-as.numeric(data_ver_2$dad_job)
data_ver_2$reason<-as.numeric(data_ver_2$reason)
data_ver_2$guardian<-as.numeric(data_ver_2$guardian)
data_ver_2$edu_sup<-as.numeric(data_ver_2$edu_sup) ##seem correlated to family_sup
data_ver_2$family_sup<-as.numeric(data_ver_2$family_sup)
```

```
data_ver_2$activities<-as.numeric(data_ver_2$activities)
data_ver_2$higher<-as.numeric(data_ver_2$higher)
data_ver_2$internet<-as.numeric(data_ver_2$internet)
data_ver_2$romantic<-as.numeric(data_ver_2$romantic)
##calling the correlation matrix function
getcormatrix(data_ver_2)
```




```
## From belwo matrix figure we can figure some more important correlations like:
## 1-mom_edu/mom_job are some positive correlations
## 2-Some relations like guardian/study_time,internet/guardian,romantic/dad_job,romantic/travel_time,fa
##3-FInally there seem negative correlation between failures and school in decreasin order.
## 4- There seem no exact collinearity between feautres.
```

1.5- Transforming int columns into categorical for modelling

For modelling, lets transform important variables to categorical predictors

```
##recheck str again using str(data_ver_2),choose some columns
col_to_factor <- c("school", "age", "mom_edu","dad_edu","mom_job","dad_job","reason","study_time","failu

data_ver_2[col_to_factor] <- lapply(data_ver_2[col_to_factor], factor)  ## as.factor() could also be us
##check the final structure before starting visual exploration
str(data_ver_2)
```

```
## 'data.frame':    649 obs. of  29 variables:
##  $ school             : Factor w/ 2 levels "1","2": 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                : Factor w/ 8 levels "15","16","17",..: 4 3 1 1 2 2 2 3 1 1 ...
##  $ parents_cohab      : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
##  $ mom_edu            : Factor w/ 5 levels "0","1","2","3",..: 5 2 2 5 4 5 3 5 4 4 ...
##  $ dad_edu            : Factor w/ 5 levels "0","1","2","3",..: 5 2 2 3 4 4 3 5 3 5 ...
##  $ mom_job            : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 2 3 4 3 3 4 3 ...
##  $ dad_job            : Factor w/ 5 levels "1","2","3","4",..: 5 3 3 4 3 3 3 5 3 3 ...
##  $ reason             : Factor w/ 4 levels "1","2","3","4": 1 1 3 2 2 4 2 2 2 2 ...
##  $ guardian           : num  2 1 2 2 1 2 2 2 2 2 ...
##  $ travel_time        : int  2 1 1 1 1 1 1 2 1 1 ...
##  $ study_time         : Factor w/ 4 levels "1","2","3","4": 2 2 2 3 2 2 2 2 2 2 ...
##  $ failures           : Factor w/ 4 levels "0","1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ edu_sup            : Factor w/ 2 levels "1","2": 2 1 2 1 1 1 1 2 1 1 ...
##  $ family_sup         : Factor w/ 2 levels "1","2": 1 2 1 2 2 2 1 2 2 2 ...
##  $ paid               : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ activities         : Factor w/ 2 levels "1","2": 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 "1","2": 2 2 2 2 2 2 2 2 2 2 ...
##  $ internet           : Factor w/ 2 levels "1","2": 1 2 2 2 1 2 2 1 2 2 ...
##  $ romantic           : Factor w/ 2 levels "1","2": 1 1 1 2 1 1 1 1 1 1 ...
##  $ family_relationship: Factor w/ 5 levels "1","2","3","4",..: 4 5 4 3 4 5 4 4 4 5 ...
##  $ free_time          : Factor w/ 5 levels "1","2","3","4",..: 3 3 3 2 3 4 4 1 2 5 ...
##  $ social_time        : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ workday_alch       : Factor w/ 5 levels "1","2","3","4",..: 1 1 2 1 1 1 1 1 1 1 ...
##  $ weekend_alch       : Factor w/ 5 levels "1","2","3","4",..: 1 1 3 1 2 2 1 1 1 1 ...
##  $ health             : Factor w/ 5 levels "1","2","3","4",..: 3 3 3 5 5 5 3 1 1 5 ...
##  $ absences           : int  4 2 6 0 0 6 0 2 0 0 ...
##  $ final_grade        : int  11 11 12 14 13 13 13 13 17 13 ...
```

We can divide our dataset into various categories to determine relations to various important predictors in one category with response variable 1-Student's information :school,age,sex,nursery,higher,romantic,reason,health 2-Sudent's family information :parents_cohab,mom_job,dad_job,guardian,family_sup,family_relationshp,mom_edu,mom_ 3-Student's study habits :study_time,travel_time,failures,edu_support,paid,absences 4-Student's leisure interests :activites,internet,free_time,social_time,social_time,workday_alch,weekend_alch

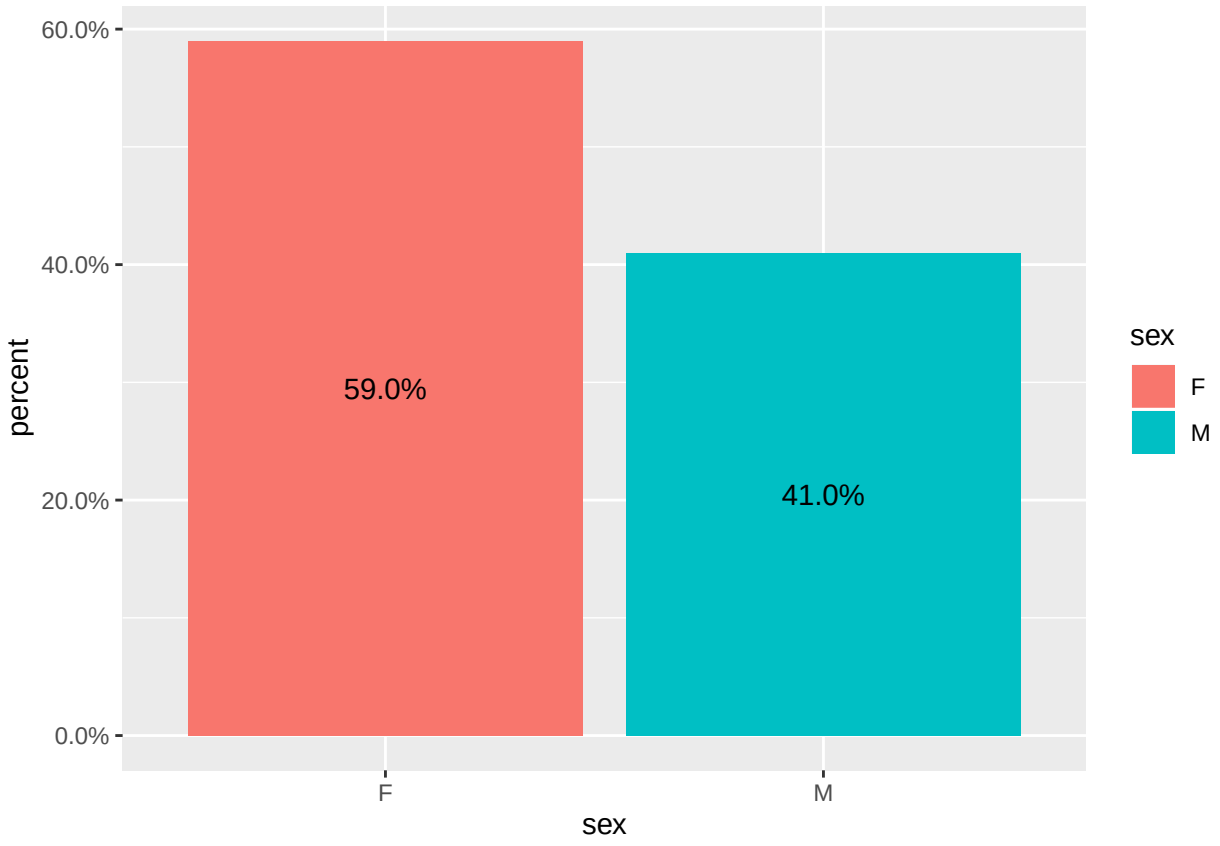
1.6- Visual Observations

1.6.1- final_grade vs sex('F' - female or 'M' - male)

```
#The distribution plot of test score vs. sex from first category

gender = data_ver_2 %>%
  group_by(sex) %>%
  summarize(count = n()) %>%
  mutate(percent = count/sum(count))
gender
```

```
## # A tibble: 2 x 3
##   sex   count percent
##   <fct> <int>   <dbl>
## 1 F       383   0.590
## 2 M       266   0.410
```

```
ggplot(gender, aes(sex, percent, fill = sex)) +
  geom_bar(stat='identity') +
  geom_text(aes(label=scales::percent(percent)), position = position_stack(vjust = .5))+
  scale_y_continuous(labels = scales::percent)
```



```
#(a) The gender constitution in the class: 59% of students are female in the class, and 41% are male.
#(b) Since the distribution is skewed, so we use the median as the center estimator.
#(c) From the grade distribution of female and male, female has higher median than male. It seems that
##code here
```

1.6.2- final_grade vs health (1 - very bad to 5 - very good)

```
##olivia-add code for health
```

1.6.3-final_grade vs romantic(binary-yes or no)

1.6.4-final_grade vs higher :wants to take higher education (binary: yes or no)

1.6.5-final_grade vs mom_edu:numeric : 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)

1.6.6-final_grade vs family_relationshp :quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

1.6.7-final_grade vs parents_cohab :parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

1.6.8-final_grade vs study_time :weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

1.6.9-final_grade vs edu_sup :extra educational support (binary: yes or no)

1.6.10-final_grade vs absences :number of school absences (numeric: from 0 to 93)

1.6.11-final_grade vs internet :Internet access at home (binary: yes or no)

1.6.12-final_grade vs workday_alch :workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

2- MODEL CONSTRUCTION

2.1- Feature Selection

2.2- BUilding some models

2.3- Comparing models

3- CHECKING MODEL ASSUMPTIONS

4- RESULT