EAS506LEC000 : STATISTICAL DATA MINING

HW ASSIGNMENT – 1

SUBMITTED BY : DITHYA SRIDHARAN

CLASS NUMBER : 48

QUESTION 1

PROBLEM STATEMENT

Consider the Student Performance Data Set on the UCI machine learning repository. Suppose that you are getting this data in order to build a predictive model for First Period Grades. Using the full dataset, investigate the data using exploratory data analysis such as scatterplots, and other tools we have discussed in class. Preprocess this data and justify your choices (elimination of outliers, elimination of variables, variable transformations, etc.)

1.1 LOADING THE DATA INTO R:

The data is available in the UCI repository and consists of two separate student subject details – Math, Portugese.

```
>df1=read.table("E:\\student-mat.csv",sep=";",header=TRUE)
>df2=read.table("E:\\student-por.csv",sep=";",header=TRUE)
```

1.2 DATA SUMMARY

We have 33 variables which act as independent variables which includes both categorical and numerical data. Among which, we will be using only 6-12 variable. There is one dependant variable for this particular problem(G1).

1.3 IDENTIFYING PERFORMANCE ANALYSIS CRITERIA

For the given problem, we add another column to the two dataframes (Pass.Fail) to calculate the performance of students in G1. If G1>10, the student is pass. Else he has failed G1.

```
>df1 <- mutate(df1, Pass.Fail = ifelse(G1 > 10, "Pass", "Fail"))
>df2 <- mutate(df2, Pass.Fail = ifelse(G1 > 10, "Pass", "Fail"))
>head(df1)
>head(df2)
```

1.4 EDA ANALYSIS

1.4.1 BAR PLOTS

Simple bar plots are used to identify the number of students who have failed in G1 and also to compare the number of males and females who have failed.

```
>attach(df1)
>b1 <- ggplot(data=df1, aes(x=Pass.Fail, fill=Pass.Fail)) +
             geom_bar(stat="count")
>attach(df2)
>b2 <- ggplot(data=df2, aes(x=Pass.Fail, fill=Pass.Fail)) +
             geom_bar(stat="count")
>grid.arrange(b1,b2,nrow=1)


>attach(df1)
>b3 <- ggplot(data=df1, aes(x=sex, fill=Pass.Fail)) +
              geom_bar(stat="count")
>attach(df2)
>b4 <- ggplot(data=df2, aes(x=sex, fill=Pass.Fail)) +
             geom_bar(stat="count")

>grid.arrange(b3,b4,nrow=1)
```

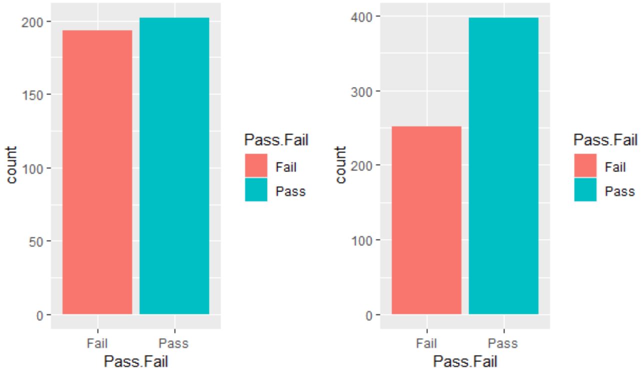
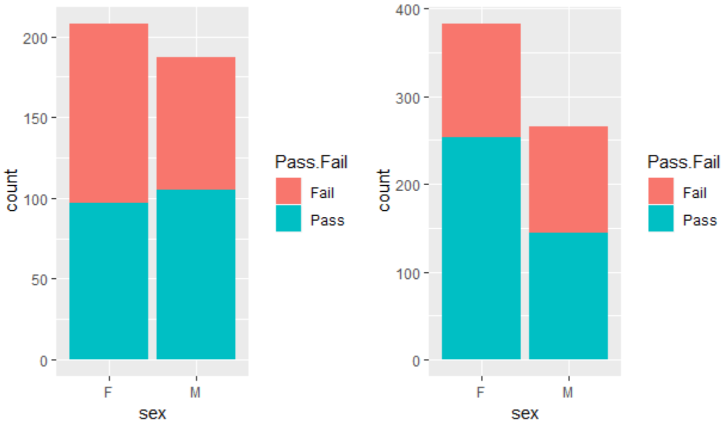


Fig 1.4.1 a) Bar plot for count of Pass or Fail among students. Left indicates math class and right indicates portugese class. It is seen that overall number of pass students are more than failed. Therefore, Pass% of G1 will be greater than Fail% of G1.

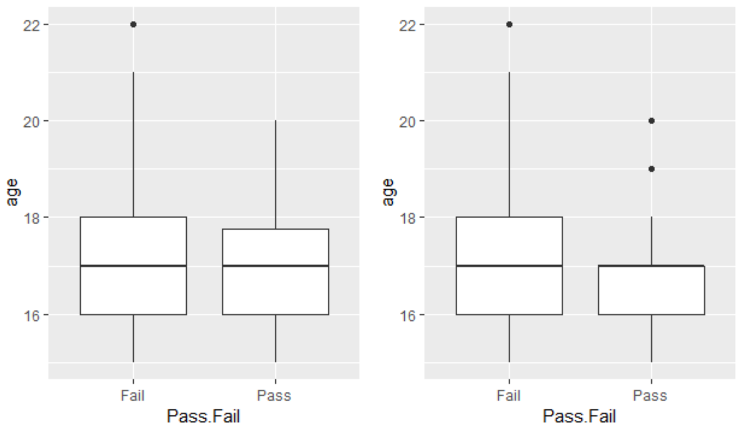
Fig 1.4.1 b)
It is seen that in math class, no of male pass students is more than no of female pass students. However, the degree of difference is not very much. In portugese class, no of female students who have passed is much higher than no of males who have passed.



Therefore, Gender variation is a significant variable affecting G1 performance.

1.4.2 BOX PLOTS

```
>c1 <- ggplot(data = df1, mapping = aes(x = Pass.Fail, y = age)) + geom_boxplot(
>c2 <- ggplot(data = df2, mapping = aes(x = Pass.Fail, y = age)) + geom_boxplot()
>grid.arrange(c1,c2,nrow=1)
```

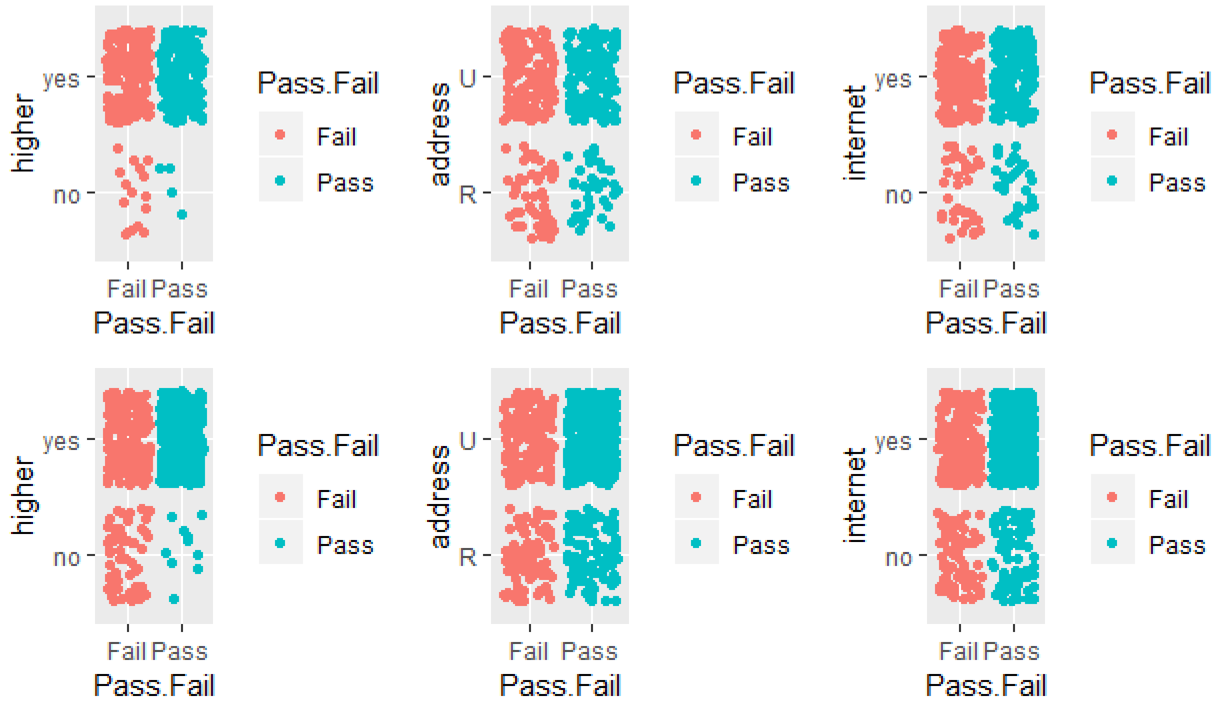


We can see that age is not a defining variable making significant changes to performance of students. So we can look for more significant variables.

1.4.3 JITTER PLOTS

)

```
>df1$sex <- as.factor(df1$sex)
>df2$sex <- as.factor(df2$sex)
>df1$address <- as.factor(df1$address)
>df2$address <- as.factor(df2$address)
>x1 <- ggplot(data = df1, mapping = aes(x = Pass.Fail, y = sex)) +
geom_jitter(aes(colour = Pass.Fail))
>x2 <- ggplot(data = df1, mapping = aes(x = Pass.Fail, y = address)) +
geom_jitter(aes(colour = Pass.Fail))
>x3 <- ggplot(data = df1, mapping = aes(x = Pass.Fail, y = internet)) +
geom_jitter(aes(colour = Pass.Fail))

>x4 <- ggplot(data = df2, mapping = aes(x = Pass.Fail, y = sex)) +
geom_jitter(aes(colour = Pass.Fail))
>x5 <- ggplot(data = df2, mapping = aes(x = Pass.Fail, y = address)) +
geom_jitter(aes(colour = Pass.Fail))
>x6 <- ggplot(data = df2, mapping = aes(x = Pass.Fail, y = internet)) +
geom_jitter(aes(colour = Pass.Fail))
>grid.arrange(x1,x2,x3,x4,x5,x6,nrow=2,ncol=3
```



Here, we study the effect of three different variables on the performance of G1. We find that on an average, students who are planning on a higher education perform better than those who do not.

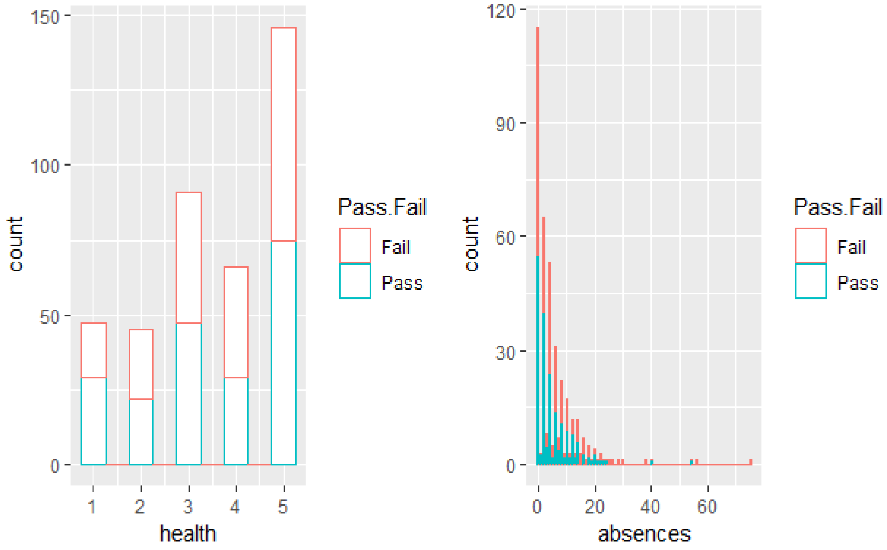
We find that the presence of internet is actually significant to the first period grades of the student. Since more students who have access to internet have passed.

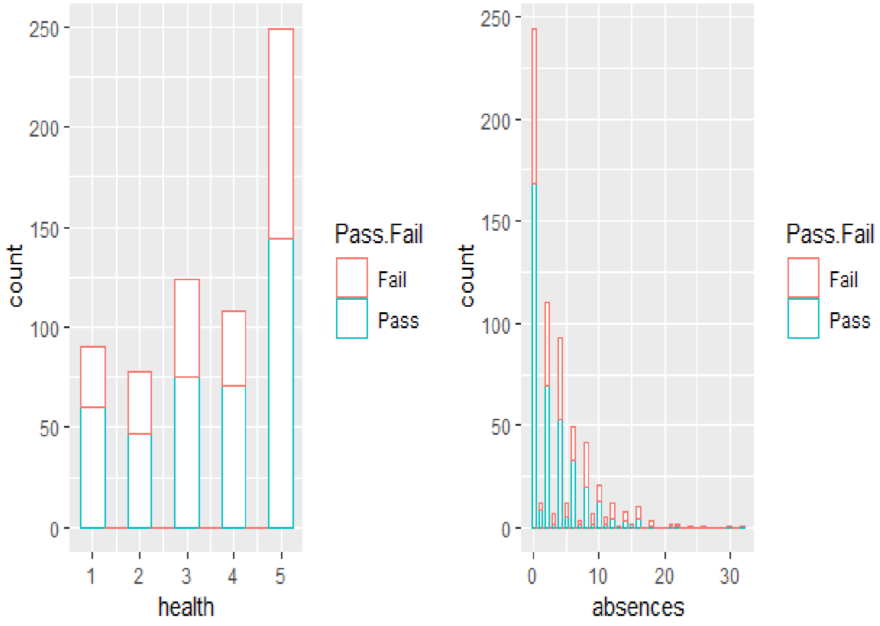
Not much can be inferred about Urban and rural based student performance. Hence we realise address is not a significant variable for modelling performance.

1.4.4 Histogram

```
>par(mfrow = c(2,2))
>p1 <- ggplot(df1, aes(x=health, color=Pass.Fail)) +
            geom_histogram(fill="white",binwidth = 0.5)
>p2 <- ggplot(df1, aes(x=absences, color=Pass.Fail)) +
            geom_histogram(fill="white",binwidth = 0.5)
>grid.arrange(p1,p2,nrow=1)
```

```
>p3 <- ggplot(df2, aes(x=health, color=Pass.Fail)) +
          geom_histogram(fill="white",binwidth = 0.5)
>p4 <- ggplot(df2, aes(x=absences, color=Pass.Fail)) +
          geom_histogram(fill="white",binwidth = 0.5)
>grid.arrange(p3,p4,nrow=1)
```

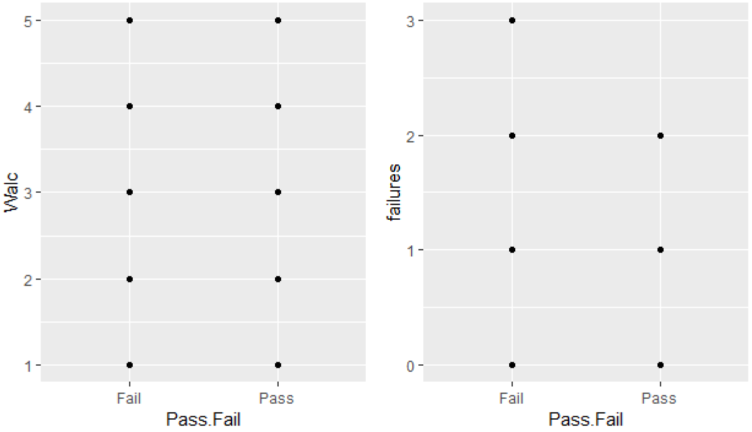


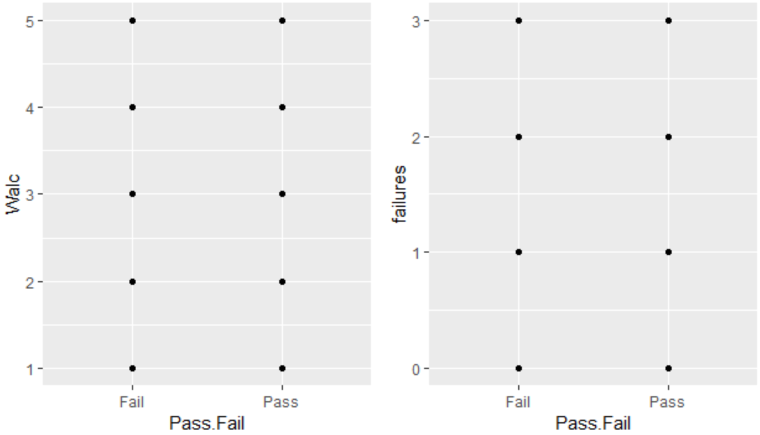


We can see that students who are more healthy perform better and students who have less number of absences perform better – as expected

1.4.4 SCATTER PLOTS

```
>y1 <- ggplot(df1, aes(x=Pass.Fail, y=Walc)) + geom_point()
>y2 <- ggplot(df1, aes(x=Pass.Fail, y=failures)) + geom_point()
>grid.arrange(y1,y2,nrow=1)

>y3 <- ggplot(df2, aes(x=Pass.Fail, y=Walc)) + geom_point()
>y4 <- ggplot(df2, aes(x=Pass.Fail, y=failures)) + geom_point()
>grid.arrange(y3,y4,nrow=1)
```





i. We can see that students with less previous failures perform better
ii. Alcohol consumption negatively affects the performance of students
iii. Therefore, they both are significant variables for modelling G1

1.5 OUTLIER ELIMINATION

Here, outliers that are found when we plot a boxplot are removed using the following commands.

```
age[!age %in% boxplot.stats(age)$out]
absences[!absences %in% boxplot.stats(absences)$out]
health[!health %in% boxplot.stats(health)$out]
```

1.6 DATA TRANSFORMATION

```
##Adding a column of data to see if student is a first generation student
df1$firstgen=ifelse(df1$Fedu>4&&df1$Medu>4,'1','0')

##Modifying age to be a range of values
d4=df1
brks=c(15,16,17,18,19,21)
d4$age=cut(d4$age,breaks=brks,include.lowest=TRUE)

##Modifying the type of data
df2$reason=gsub('C',"course",df2$reason)
df2$reason=gsub('R',"reputation",df2$reason)
```

1.7 VARIABLE ELIMINATION

```
vareli1 <- names(df1) %in% c("romantic", "guardian","firstgen","famrel","activities")
newdata1 <- df1[!vareli1]
vareli2 <- names(df2) %in% c("romantic", "guardian","firstgen","famrel","activities")
newdata2 <- df2[!vareli2]
```

Here, some elimination is performed to preprocess the data. Since the data is already tidy(without NA values), data cleaning refers to only selecting only the variables needed.

1.8 INSIGHTS ACQUIRED

i. Gender based – More females perform better in portugese G1, however this is less significant compared to other variables like attendance and health.
ii. Alcohol consumption – It has a detrimental effect on performance but lesser significant to attendance and health, especially weekday alcohol consumption.
iii. Higher – It is clearly seen that students who have plans of studying highers perform better than thos who do not.
iv. Health – As health deteriotes, performance is affected
v. Attendance – Students who miss less number of classes perform better as expected.
vi. Access to internet – Internet has a significant effect on the performance of G1

QUESTION 2

PROBLEM STATEMENT

Perform a multiple regression on the dataset you pre-processed in question one. The response are the first period grades. Use the lm() function in R.
a) Which predictors appear to have a significant relationship to the response.
b) What suggestions would you make to a first-year student trying to achieve good grades.
c) Use the * and : symbols to fit models with interactions. Are there any interactions that are significant?

2.2 UNDERSTANDING THE DATA

```
>head(newdata1)
>head(newdata2)
>summary(newdata1)
>summary(newdata2)
```

Dataframes 1 and 2 are both studied to interpret how to perform regression.

2.3 CHECKING CORRELATION BETWEEN G1 AND FINAL GRADE BY CORRELATION MATRIX AND USE OF lm()

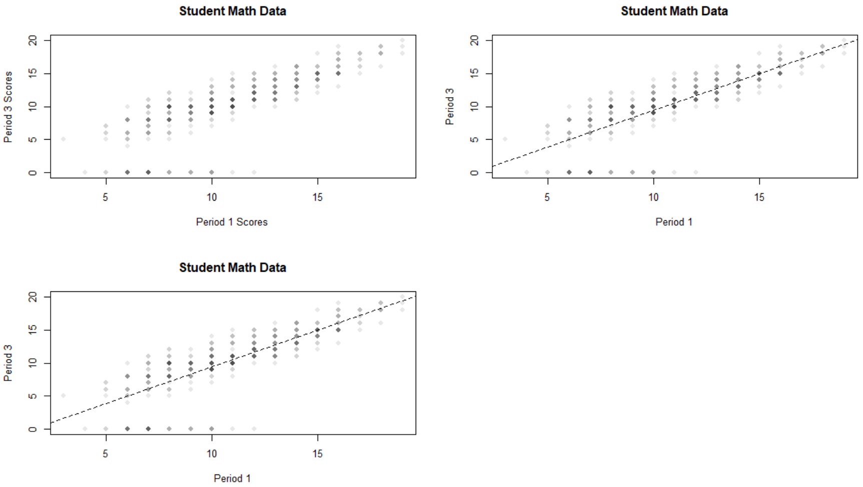
```
cor(newdata1[sapply(newdata1, is.numeric)])
cor(newdata2[sapply(newdata2, is.numeric)])
reg113 <- lm(G3 ~ G1, data = newdata1)
reg213 <- lm(G3 ~ G1, data = newdata2)
summary(reg113)
summary(reg213)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.65280    0.47475  -3.481 0.000555 ***
G1           1.10626    0.04164  26.568  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.743 on 393 degrees of freedom
Multiple R-squared:  0.6424,	Adjusted R-squared:  0.6414 
F-statistic: 705.8 on 1 and 393 DF,  p-value: < 2.2e-16
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.82040    0.30545   2.686  0.00742 ** 
G1           0.97250    0.02605  37.329  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.821 on 647 degrees of freedom
Multiple R-squared:  0.6829,	Adjusted R-squared:  0.6824 
F-statistic:  1393 on 1 and 647 DF,  p-value: < 2.2e-16
```

i. Yes, there is a significant relationship between G1 and G3 in math scores, b = 1.106, t = 26.57, p < .01

ii. The association between G1 and G3 are comparatively less significant for portugese.



The above figure describes the correlation between G1 and G3 with a regression line.

2.4 APPLYING STANDARD REGRESSION TO THE SIGNIFICANT DATA WE ANALYSED IN Q1

```
myvar <- c("internet","Walc","higher","absences","health")
nd <- newdata1[myvar]
head(nd)
plot(nd)

nd1 <- newdata2[myvar]
head(nd1)
plot(nd1)

results = lm(G1 ~ internet + Walc + absences + failures + higher, data=newdata1)
summary(results)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.3849312  0.8860062  11.721  < 2e-16 ***
internetyes  0.4525039  0.4212405   1.074    0.283
Walc        -0.1904148  0.1237028  -1.539    0.125
absences    -0.0006858  0.0198313  -0.035    0.972
failures    -1.4223122  0.2222761  -6.399  4.5e-10 ***
higheryes    1.1191879  0.7474279   1.497    0.135
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.099 on 389 degrees of freedom
Multiple R-squared:  0.1393,   Adjusted R-squared:  0.1282
F-statistic: 12.59 on 5 and 389 DF,  p-value: 2.418e-11
```

i. This shows that the variables we considered significant are indeed significant variables but could be better with respect to G1.
ii. Here, failures(-1.42) and highers(1.11) significantly affect performance of G1.
iii. Alcohol consumption(-0.19) is negatively correlated.
iv. Multiple regression is 13% which should be improved.

2.5 Finding regression and correlation wrt interactions

```
modx <- lm(G1 ~ failures*internet, data=newdata1)

mody <- lm(G1 ~ age:sex, data=newdata2)

summary(modx)

summary(mody)
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           10.8254     0.4336  24.967   <2e-16 ***
failures              -1.0164     0.4666  -2.179    0.030 *  
internetyes            0.7254     0.4720   1.537    0.125    
failures:internetyes  -0.6945     0.5229  -1.328    0.185    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.104 on 391 degrees of freedom
Multiple R-squared:  0.1322,	Adjusted R-squared:  0.1255 
F-statistic: 19.85 on 3 and 391 DF,  p-value: 5.387e-12
```

i. Students who have internet score 0.72 points higher in G1 than those who don’t.
ii. Students with past failures are negatively correlated.
iii. When both internet and failure interact there is a detrimental affect.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    11.7978     0.1064 110.896  < 2e-16 ***
failures:sexF  -1.9621     0.2455  -7.993 6.07e-15 ***
failures:sexM  -1.6361     0.2173  -7.528 1.74e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.536 on 646 degrees of freedom
Multiple R-squared:  0.149,	Adjusted R-squared:  0.1464 
F-statistic: 56.56 on 2 and 646 DF,  p-value: < 2.2e-16
```

It shows that interaction between failures and females is more. They are more negatively correlated(-1.96) to the performance of G1.

ANSWERS TO QUESTIONS ASKED IN PROBLEM STATEMENT

2a) The predictors G3(final grade), failures, alcohol consumption and internet have a significant effect on the response.

2b) A first year student should keep her health, alcohol consumption in check. Also previous failures affect the grades significantly so she should make sure she doesn't have any failure which may affect her. A student with internet scores marginally more than one who doesn't.

2c) Performed above

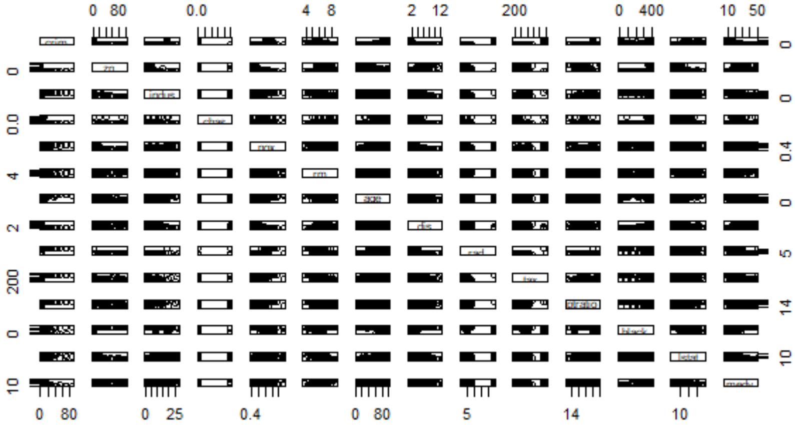
QUESTION 3

PROBLEM STATEMENT

This exercise concerns the
boston housing data in the MASS library (>library(MASS) >data(Boston)).
a) Make pairwise scatterplots of the predictors, and describe your findings.
b) Are any of the predictors associated with per capita crime rate?
c) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.
d) In this data set, how many of the suburbs average more than seen rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling

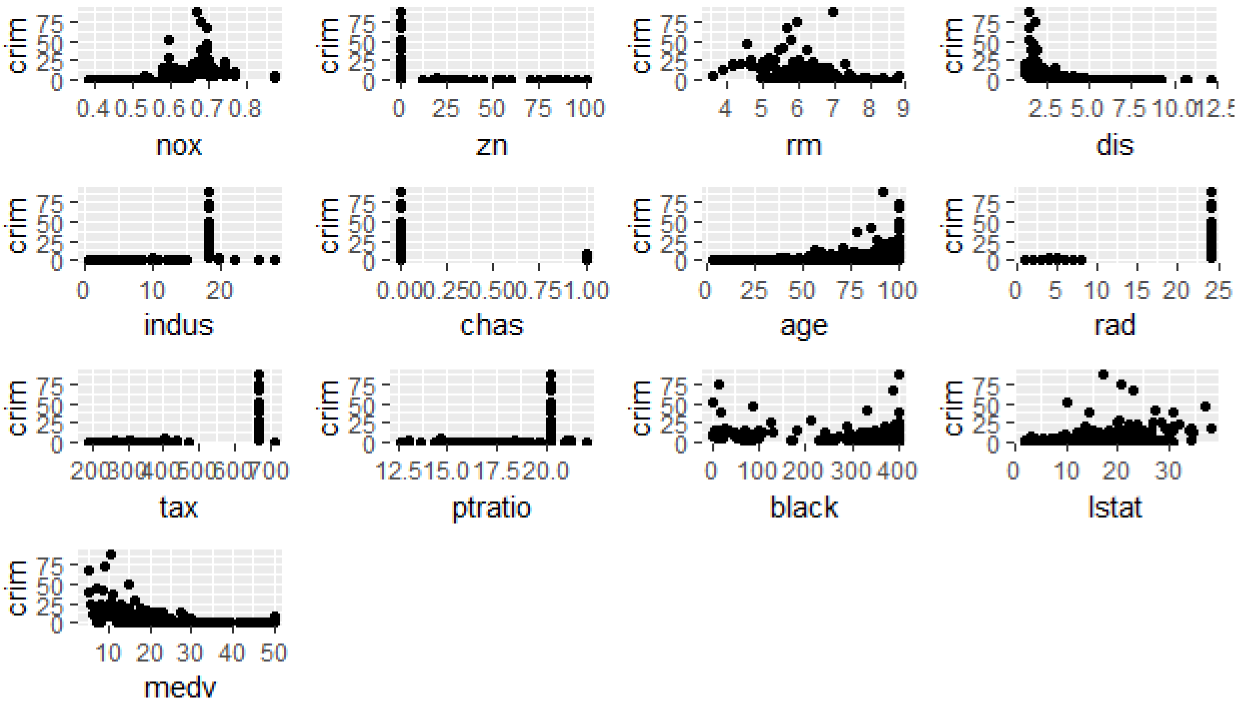
3.1UNDERSTANDING THE DATA

```
?Boston
nrow(Boston)
head(Boston)
pairs(Boston)
```



Boston data has 506 rows and 14 columns that contains information on the housing values in the suburbs of Boston.

3.2 Pair-wise scatterplots against crime data



```
>y1 <- ggplot(Boston, aes(x=nox, y=crim)) + geom_point()
>y2 <- ggplot(Boston, aes(x=zn, y=crim)) + geom_point()
>y3 <- ggplot(Boston, aes(x=rm, y=crim)) + geom_point()
>y4 <- ggplot(Boston, aes(x=dis, y=crim)) + geom_point()
>y5 <- ggplot(Boston, aes(x=indus, y=crim)) + geom_point()
>y6 <- ggplot(Boston, aes(x=chas, y=crim)) + geom_point()
>y7 <- ggplot(Boston, aes(x=age, y=crim)) + geom_point()
>y8 <- ggplot(Boston, aes(x=rad, y=crim)) + geom_point()
>y9 <- ggplot(Boston, aes(x=tax, y=crim)) + geom_point()
>y10 <- ggplot(Boston, aes(x=ptratio, y=crim)) + geom_point()
>y11 <- ggplot(Boston, aes(x=black, y=crim)) + geom_point()
>y12 <- ggplot(Boston, aes(x=lstat, y=crim)) + geom_point()
>y13 <- ggplot(Boston, aes(x=medv, y=crim)) + geom_point()

>grid.arrange(y1,y2,y3,y4,y5,y6,y7,y8,y9,y10,y11,y12,y13,nrow=4,ncol=4)
```

It’s hard to come to any conclusion about our predictors from the pair-wise scatter plots Hence we go for correlation matrix and histograms.

3.3 CORRELATION MATRIX WRT CRIME

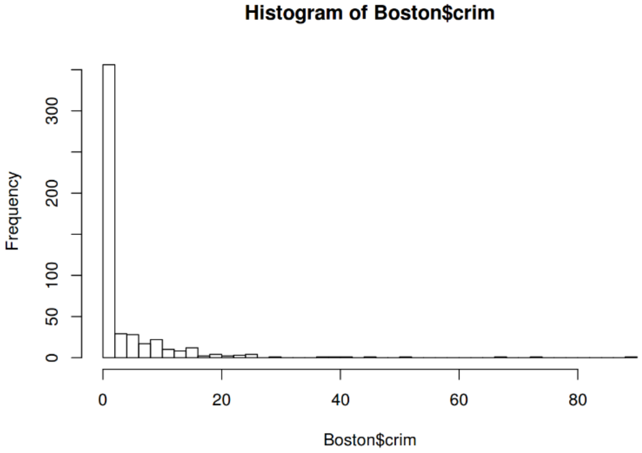
```
>attach(Boston)
>x <- (corrmatrix <- cor(Boston, use="complete.obs")[1,])
>x
```

```
crim          zn       indus        chas         nox          rm
 1.00000000 -0.20046922  0.40658341 -0.05589158  0.42097171 -0.21924670
        age         dis         rad         tax     ptratio       black
 0.35273425 -0.37967009  0.62550515  0.58276431  0.28994558 -0.38506394
      lstat        medv
 0.45562148 -0.38830461
```

i. Crime has some significant correlation with everything except chas.
ii. It has negative correlation with dis, medv, black.
iii. There might be relationship between crime and nox, rm, lstat, medv and black

3.5 HISTOGRAM PLOTTING FOR PER CAPITA CRIME

```
attach(Boston)
hist(Boston$crim,breaks = 50)
```

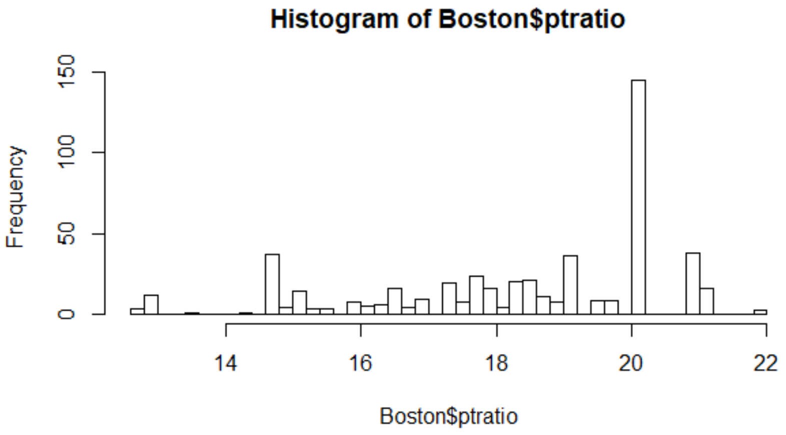


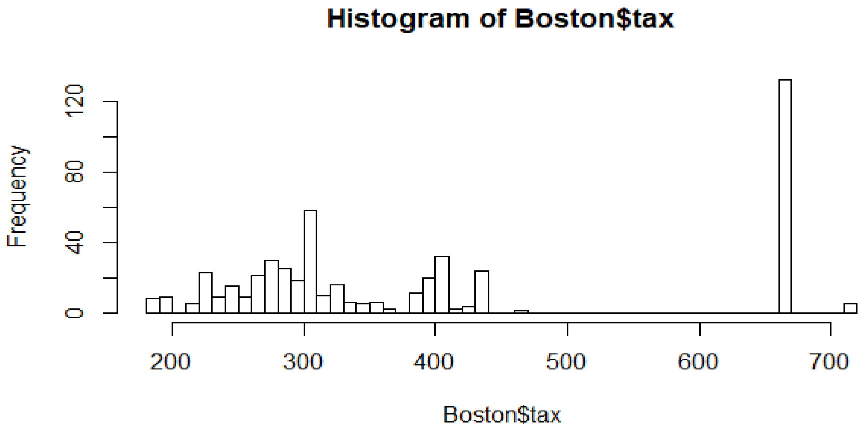
i. Most suburbs don't seem to have high crime rates

ii. Around 80% of the suburbs comes under crim<20

3.6 TAX-RATES AND PUPIL-TEACHER RATIO AS PREDICTORS

```
hist(Boston$tax, breaks = 50)
hist(Boston$ptratio, breaks = 50)
```





i. Maximum no of rows have ptratio>20
ii. Maximum no of rows have tax=666

Range of each predictor

```
range(Boston[crim>20,]$crim)
range(Boston[crim>20,]$nox)
range(Boston[crim>20,]$lstat)
range(Boston[crim>20,]$black)
```

```
> range(Boston[crim>20,]$crim)
[1] 20.0849 88.9762
> range(Boston[crim>20,]$nox)
[1] 0.597 0.740
> range(Boston[crim>20,]$lstat)
[1] 10.11 36.98
> range(Boston[crim>20,]$black)
[1]   2.6 396.9
```

i. The range of an observation variable is the difference of its largest and smallest data values. It is a measure of how far apart the entire data spreads in value.

ii. This shows that lstat is more spread than the other predictors.

ANSWERS TO QUESTIONS IN PROBLEM STATEMENT

3a) Peformed above
3b) Looks like rm, nox, lstat, medv and blacks are significantly associated to per capita crime
3c) Most of the suburbs don’t have high crime rates. Around 80-85% of the suburbs have per capita crime<20
3d) 64 suburbs have more than 7 rooms per dwelling and 13 suburbs have more than 8 rooms per dwelling. It is found that for suburbs with more than 8 rooms per dwelling, the average age indicates senior citizens and per capita crime seems to be less.

```
nrow(Boston[Boston$rm > 7, ])
nrow(Boston[Boston$rm > 8, ])
m <- (Boston[Boston$rm > 7, ])
m
```

QUESTION 4

PROBLEM STATEMENT

Compare the classification
performance of linear regression and k-nearest neighbor classification on the *zipcode* data. In particular, consider only the 2's and 3's for this problem, and k = 1,3,5,7,9,11, 13,15. Show both the training and the test error for each choice of k.
The *zipcode* data is available in the ElemStatLearn package

STUDYING THE TRAINING AND TEST DATA AND SUBSETTING ACCORDING TO QUESTION

```
ziptrain23 <- subset(zip.train, zip.train[,1]==2 | zip.train[,1]==3)
ziptest23 <- subset(zip.test, zip.test[,1]==2 | zip.test[,1]==3)
Traindf1 <- as.data.frame(ziptrain23)
Testdf1 <- as.data.frame(ziptest23)
head(Traindf1)
head(Testdf1)
```

REGRESSION CLASSIFICATION

```
trainreg <- lm(V1 ~ ., data=Traindf1)
prediction <- predict.lm(trainreg,Testdf1)
print(prediction)
```

We are using lm() function for standard regression calculation for V1 of the dataset.

ERROR ANALYSIS FOR REGRESSION

```
e1 <- mean((Traindf1$V1 - (prediction)) ^ 2)
print(e1)
```

The error of prediction is calculated using least squares method. The error comes to 0.540953

KNN PREDICTION AND ERROR ANALYSIS

```
knn.train <- Traindf1[,2:1389]
knn.test <- Testdf1[,2:364]

knn.train.V1 <- as.factor(Traindf1$V1)
knn.test.V1 <- as.factor(Testdf1$V1)


prediction_knn1 <- knn(train = Traindf1,test = Testdf1,cl = knn.train.V1,k = 1)
prediction_knn3 <- knn(train = Traindf1,test = Testdf1,cl = knn.train.V1,k = 3)
prediction_knn5 <- knn(train = Traindf1,test = Testdf1,cl = knn.train.V1,k = 5)
prediction_knn7 <- knn(train = Traindf1,test = Testdf1,cl = knn.train.V1,k = 7)
prediction_knn9 <- knn(train = Traindf1,test = Testdf1,cl = knn.train.V1,k = 9)
prediction_knn11 <- knn(train = Traindf1,test = Testdf1,cl = knn.train.V1,k = 11)
prediction_knn13 <- knn(train = Traindf1,test = Testdf1,cl = knn.train.V1,k = 13)
prediction_knn15<- knn(train = Traindf1,test = Testdf1,cl = knn.train.V1,k = 15)
#kne1 <- mean((Traindf1$V1 - (prediction_knn1)) ^ 2)

#Error analysis
err_rate1 <- round(mean(knn.train.V1!=prediction_knn1))
print(err_rate1)
err_rate3 <- round(mean(knn.train.V1!=prediction_knn3))
print(err_rate3)
err_rate5 <- round(mean(knn.train.V1!=prediction_knn5))
print(err_rate5)
err_rate7 <- round(mean(knn.train.V1!=prediction_knn7))
print(err_rate7)
err_rate9 <- round(mean(knn.train.V1!=prediction_knn9))
print(err_rate9)
err_rate11 <- round(mean(knn.train.V1!=prediction_knn11))
print(err_rate11)
err_rate13 <- round(mean(knn.train.V1!=prediction_knn13))
print(err_rate13)
err_rate15 <- round(mean(knn.train.V1!=prediction_knn15))
print(err_rate15)
```

The classification performance of both are generally comparable. KNN achieves the best results