# Report on SSLC Data Analysis

**Group No. 06**

Submitted by:

Ankit Mishra (MT201409)

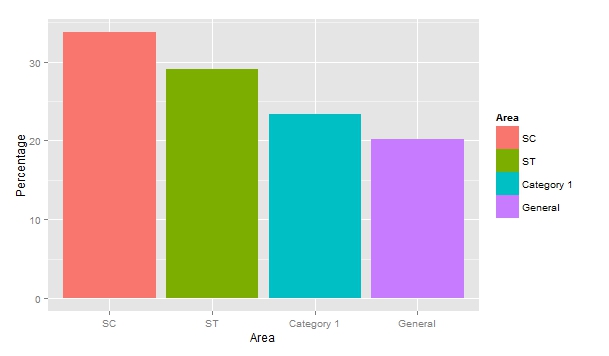
Shan Mehrotra (MT20140111)

Kumar Rahul (MT2014058)

## Negative observations

* 1. **Category wise performance comparison :**

Here number of students failed in different category is shown.



**R Code :**

#Measuring performance of students in different castes

caste1\_data <- filter(g6, NRC\_CASTE\_CODE == 1)#SC

count(caste1\_data)

mean(caste1\_data$TOTAL\_MARKS)

ggplot(caste1\_data, aes(x=caste1\_data$TOTAL\_MARKS)) + geom\_histogram(binwidth=5, color = "black", fill = "#FFFFE0")

caste2\_data <- filter(g6, NRC\_CASTE\_CODE == 2)#ST

count(caste2\_data)

mean(caste2\_data$TOTAL\_MARKS)

ggplot(caste2\_data, aes(x=caste2\_data$TOTAL\_MARKS)) + geom\_histogram(binwidth=5, color = "black", fill = "#FFFFE0")

caste3\_data <- filter(g6, NRC\_CASTE\_CODE == 3)#Cat1

count(caste3\_data)

mean(caste3\_data$TOTAL\_MARKS)

ggplot(caste3\_data, aes(x=caste3\_data$TOTAL\_MARKS)) + geom\_histogram(binwidth=5, color = "black", fill = "#FFFFE0")

caste4\_data <- filter(g6, NRC\_CASTE\_CODE == 4)#Gen

count(caste4\_data)

mean(caste4\_data$TOTAL\_MARKS)

ggplot(caste4\_data, aes(x=caste4\_data$TOTAL\_MARKS)) + geom\_histogram(binwidth=5, color = "black", fill = "#FFFFE0")

#Failed student percentage in all categories

#SC

#caste1\_fail\_data <- filter(caste1\_data, NRC\_RESULT == 'F')

caste1\_fail\_perc <- (sum(g6$NRC\_RESULT=='F' & g6$NRC\_CASTE\_CODE==1)/sum(g6$NRC\_CASTE\_CODE==1))\*100

caste1\_fail\_perc

#ST

#caste2\_fail\_data <- filter(caste2\_data, NRC\_RESULT == 'F')

caste2\_fail\_perc <- (sum(g6$NRC\_RESULT=='F' & g6$NRC\_CASTE\_CODE==2)/sum(g6$NRC\_CASTE\_CODE==2))\*100

caste2\_fail\_perc

#Cat1

#caste3\_fail\_data <- filter(caste3\_data, NRC\_RESULT == 'F')

caste3\_fail\_perc <- (sum(g6$NRC\_RESULT=='F' & g6$NRC\_CASTE\_CODE==3)/sum(g6$NRC\_CASTE\_CODE==3))\*100

caste3\_fail\_perc

#Gen

#caste4\_fail\_data <- filter(caste4\_data, NRC\_RESULT == 'F')

caste4\_fail\_perc <- (sum(g6$NRC\_RESULT=='F' & g6$NRC\_CASTE\_CODE==4)/sum(g6$NRC\_CASTE\_CODE==4))\*100

caste4\_fail\_perc

caste\_fail\_data <- data.frame(Area = factor(c("SC","ST","Category 1","General"), levels=c("SC","ST","Category 1","General")), Percentage = c(caste1\_fail\_perc, caste2\_fail\_perc,caste3\_fail\_perc,caste4\_fail\_perc))

ggplot(data=caste\_fail\_data, aes(x=Area, y=Percentage, fill=Area)) + geom\_bar(stat="identity")

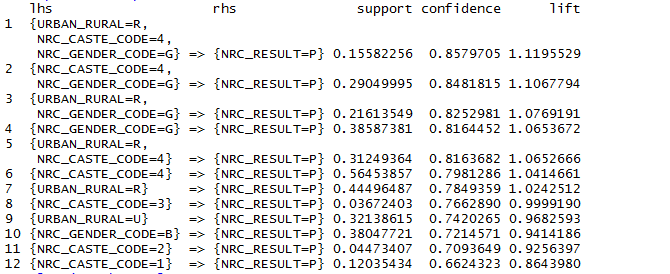
****

Fig. : Association Rule Mining

**R Code :**

#Performing Association Rule Mining on Caste Data

library(arules)

arule\_data <- g6[,c(6,12,13,30)]

arule\_data$NRC\_CASTE\_CODE <- as.factor(arule\_data$NRC\_CASTE\_CODE)

tbl\_df(arule\_data)

rules <- apriori(arule\_data)

inspect(rules)

rules <- apriori(arule\_data, parameter = list(minlen = 2, supp = 0.005, conf = 0.6), appearance = list(rhs=c("NRC\_RESULT=F","NRC\_RESULT=P"),default = "lhs"), control = list(verbose = T))

rules.sorted <- sort(rules, by="lift")

inspect(rules.sorted)

#pruning redundant rules

subset.matrix <- is.subset(rules.sorted, rules.sorted)

subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA

redundant <- colSums(subset.matrix, na.rm=T) >= 1

which(redundant)

# remove redundant rules

rules.pruned <- rules.sorted[!redundant]

inspect(rules.pruned)

#plotting the arules

library(arulesViz)

plot(rules.pruned, method="graph", control=list(type="items"))

plot(rules.pruned, method="paracoord", control=list(reorder=TRUE))

**Observation:**

* The number of student failed in SC/ST category is more.
* Maximum number of students passed lies in general category.
* From the association rules 1,2,5,6, we can observe that confidence of general category students are better than other categories.

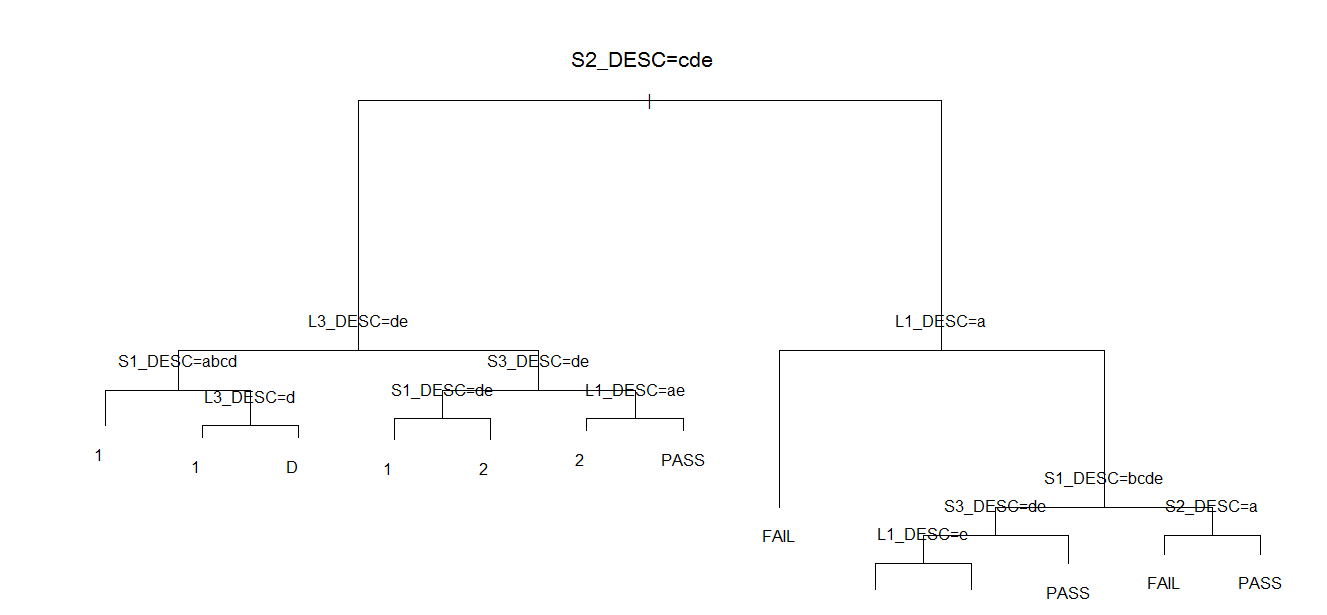
**Conclusion:**

* Students lying in SC/ST category have poor performance.

**b. Subject important for distinction :**

Performing Discretization + Classification

Here a=Fail, b=Pass, c=2, d=1, e=Distinction



**R Code :**

#Performing Discretization + Classification

library(rattle)

library(rpart.plot)

library(RColorBrewer)

g6$L1\_DESC<-cut(g6$L1\_MARKS,c(0,30,45,60,80,100),labels=c('F','P','2','1','D'))

g6$L2\_DESC<-cut(g6$L2\_MARKS,c(0,30,45,60,80,100),labels=c('F','P','2','1','D'))

g6$L3\_DESC<-cut(g6$L3\_MARKS,c(0,30,45,60,80,100),labels=c('F','P','2','1','D'))

g6$S1\_DESC<-cut(g6$S1\_MARKS,c(0,30,45,60,80,100),labels=c('F','P','2','1','D'))

g6$S2\_DESC<-cut(g6$S2\_MARKS,c(0,30,45,60,80,100),labels=c('F','P','2','1','D'))

g6$S3\_DESC<-cut(g6$S3\_MARKS,c(0,30,45,60,80,100),labels=c('F','P','2','1','D'))

ind<-sample(2,nrow(g6),replace = TRUE,prob = c(0.7,0.3))

train<-g6[ind==1,]

test<-g6[ind==2,]

myf<-NRC\_CLASS~L1\_DESC+L2\_DESC+L3\_DESC+S1\_DESC+S2\_DESC+S3\_DESC

tree1<-rpart(myf,data = train,control = rpart.control(minsplit = 10))

plot(tree1)

text(tree1)

print(tree1)

fancyRpartPlot(tree1)

#predicting the NRC\_CLASS

Prediction <- predict(tree1, test, type = "class")

Prediction

submit <- data.frame(original\_value= test$NRC\_CLASS,predicted\_value=Prediction)

submit

table(submit$original\_value==submit$predicted\_value)

xtab <- table(submit$original\_value,submit$predicted\_value)

library(caret)

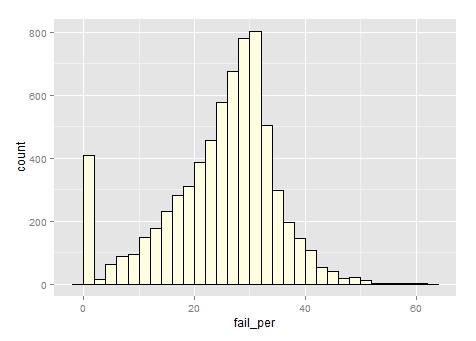
confusionMatrix(xtab) #all necessary parameters

**Conclusion:**

* Subjects L3 and S2 are important for getting distinction.

**c. Percentage distribution for failed students :**

Histogram showing percentage distribution of failed students.

****

**R Code :**

fail <- filter(g6, NRC\_RESULT == 'F')

fail\_per <- (fail$TOTAL\_MARKS/650)\*100

ggplot(fail, aes(x=fail\_per)) + geom\_histogram(binwidth=2, color = "black", fill = "#FFFFE0")

**Observation :**

* The data contain students whose result is shown failed even after getting more than 35 percent marks.

## Positive observations

* 1. **Comparing urban and rural students :**

Considering either L1=E and L2=k or L1=k and L2=E.

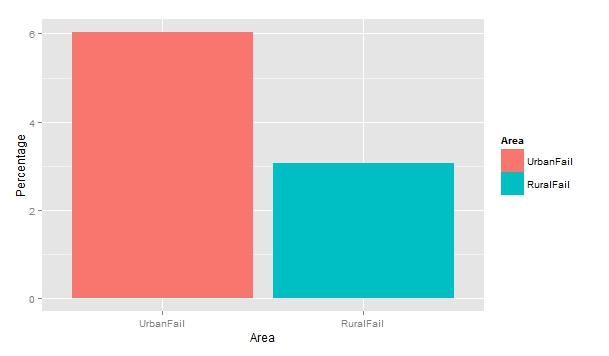


Fig.: L1=k and L2=E.

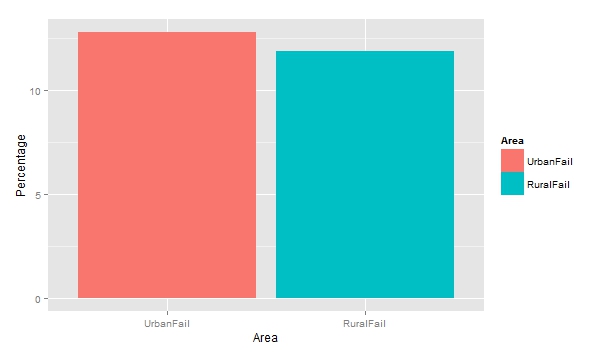


Fig.: L1=E and L2=K.

**R Code:**

#extracting L1 = K and L2 = E in urban

urb\_school\_L2\_data <- filter(g6, URBAN\_RURAL == "U", L1\_CODE == "01K",L2\_CODE == "31E")

str(urb\_school\_L2\_data)

urb\_l2\_mean <- mean(urb\_school\_L2\_data$L2\_MARKS)

table(urb\_school\_L2\_data$L2\_MARKS)

#extracting L1 = K and L2 = E in rural

rur\_school\_L2\_data <- filter(g6, URBAN\_RURAL == "R", L1\_CODE == "01K",L2\_CODE == "31E")

str(rur\_school\_L2\_data)

rur\_l2\_mean <- mean(rur\_school\_L2\_data$L2\_MARKS)

#drawing histogram for both data

library(ggplot2)

#ggplot(urb\_school\_L2\_data, aes(x=urb\_school\_L2\_data$L2\_MARKS)) + geom\_histogram(binwidth=2, color = "black", fill = "#FFFFE0")

#ggplot(rur\_school\_L2\_data, aes(x=rur\_school\_L2\_data$L2\_MARKS)) + geom\_histogram(binwidth=2, color = "black", fill = "#FFFFE0")

#calculating mean marks in L2 = E

urb\_l2\_mean <- mean(urb\_school\_L2\_data$L2\_MARKS)

rur\_l2\_mean <- mean(rur\_school\_L2\_data$L2\_MARKS)

#Percentage of students failed in L2 = E in urban

#urb\_l2\_fail <- filter( urb\_school\_L2\_data,L2\_RESULT == "F")

urb\_l2\_fail\_perc <- (sum( urb\_school\_L2\_data$L2\_RESULT == "F") / sum(urb\_school\_L2\_data$L2\_RESULT=='F' | urb\_school\_L2\_data$L2\_RESULT=='P')) \* 100

urb\_l2\_fail\_perc #12.46%

#Percentage of students failed in L2 = E in rural

#rur\_l2\_fail <- filter(rur\_school\_L2\_data, L2\_RESULT == "F")

rur\_l2\_fail\_perc <- (sum( rur\_school\_L2\_data$L2\_RESULT == "F") / sum(rur\_school\_L2\_data$L2\_RESULT=='F' | rur\_school\_L2\_data$L2\_RESULT=='P')) \* 100

rur\_l2\_fail\_perc #11.60%

gal\_fail\_data <- data.frame(Area = factor(c("UrbanFail","RuralFail"), levels=c("UrbanFail","RuralFail")), Percentage = c(urb\_l2\_fail\_perc, rur\_l2\_fail\_perc))

ggplot(data=gal\_fail\_data, aes(x=Area, y=Percentage, fill=Area)) + geom\_bar(stat="identity")

#extracting L1 = E and L2 = K in urban

urb\_l1Eng\_data <- filter(g6, URBAN\_RURAL == "U", L1\_CODE == "14E",L2\_CODE == "33K")

urb\_l1Eng\_mean <- mean(urb\_l1Eng\_data$L1\_MARKS)

urb\_l1Eng\_mean

#extracting L1 = E and L2 = K in rural

rur\_l1Eng\_data <- filter(g6, URBAN\_RURAL == "R", L1\_CODE == "14E",L2\_CODE == "33K")

rur\_l1Eng\_mean <- mean(urb\_l1Eng\_data$L1\_MARKS)

rur\_l1Eng\_mean

#Percentage of students failed in L1 = E in urban

#urb\_l1Eng\_fail <- filter(urb\_l1Eng\_data, L2\_RESULT == "F")

urb\_l1Eng\_fail\_perc <- (sum( urb\_l1Eng\_data$L2\_RESULT == "F") / sum(urb\_l1Eng\_data$L2\_RESULT=='F' | urb\_l1Eng\_data$L2\_RESULT=='P')) \* 100

urb\_l1Eng\_fail\_perc

#Percentage of students failed in L1 = E in rural

#rur\_l1Eng\_fail <- filter(rur\_l1Eng\_data, L2\_RESULT == "F")

rur\_l1Eng\_fail\_perc <- (sum( rur\_l1Eng\_data$L2\_RESULT == "F") / sum(rur\_l1Eng\_data$L2\_RESULT=='F' | rur\_l1Eng\_data$L2\_RESULT=='P')) \* 100

rur\_l1Eng\_fail\_perc

gal\_fail\_data <- data.frame(Area = factor(c("UrbanFail","RuralFail"), levels=c("UrbanFail","RuralFail")), Percentage = c(urb\_l1Eng\_fail\_perc, rur\_l1Eng\_fail\_perc))

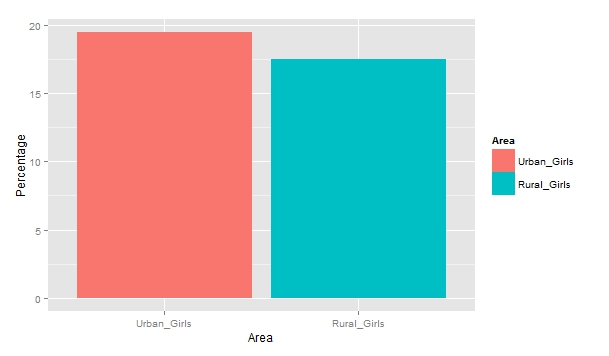
ggplot(data=gal\_fail\_data, aes(x=Area, y=Percentage, fill=Area)) + geom\_bar(stat="identity")

**Observation:**

* Rural students are performing better than urban student in English.

**Conclusion:**

* Rural students are performing better.
  1. **Analyzing performance of girls in rural and urban areas :**

Here number of girls failed in urban and rural areas are shown.

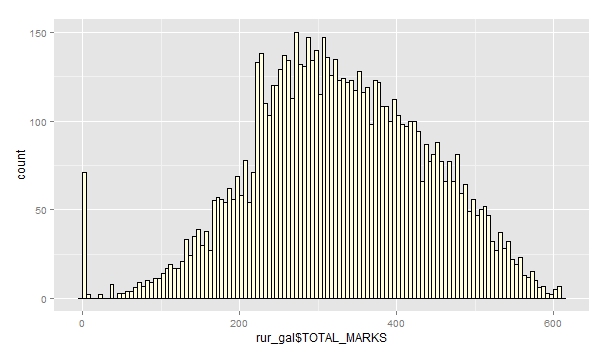


Fig. : Girls total marks in rural area.

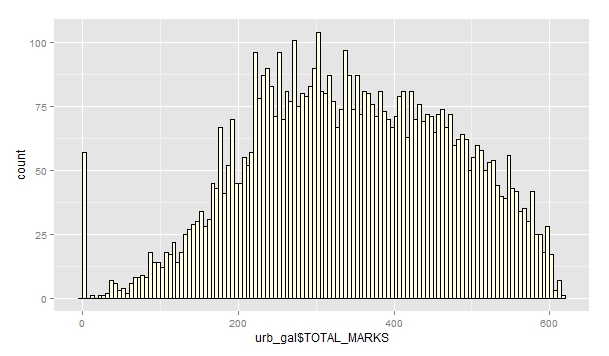


Fig. : Girls total marks in urban area.

**R Code :**

#Calculating No of girls

urb\_gal <- filter(g6, URBAN\_RURAL == "U", NRC\_GENDER\_CODE == "G")

rur\_gal <- filter(g6, URBAN\_RURAL == "R", NRC\_GENDER\_CODE == "G")

count(urb\_gal)

count(rur\_gal)

library(ggplot2)

#plotting histograms of total marks distribution

ggplot(urb\_gal, aes(x=urb\_gal$TOTAL\_MARKS)) + geom\_histogram(binwidth=5, color = "black", fill = "#FFFFE0")

mean(urb\_gal$TOTAL\_MARKS)

ggplot(rur\_gal, aes(x=rur\_gal$TOTAL\_MARKS)) + geom\_histogram(binwidth=5, color = "black", fill = "#FFFFE0")

mean(rur\_gal$TOTAL\_MARKS)

#No of girls failed in rural and urban areas

urb\_gal\_fail <- filter(urb\_gal, NRC\_RESULT == "F")

count(urb\_gal\_fail)

rur\_gal\_fail <- filter(rur\_gal, NRC\_RESULT == "F")

count(rur\_gal\_fail)

#Percentage of girls failed in Urban and Rural areas

#urb\_gal\_fail\_perc <- (sum(urb\_gal\_fail)/sum(urb\_gal))\*100

#urb\_gal\_fail\_perc #19.45%

#rur\_gal\_fail\_perc <- (sum(rur\_gal\_fail)/sum(rur\_gal))\*100

#rur\_gal\_fail\_perc #17.48

urb\_gal\_fail\_perc <- (sum(g6$URBAN\_RURAL == 'U' & g6$NRC\_RESULT == 'F' & g6$NRC\_GENDER\_CODE == "G")/sum(g6$URBAN\_RURAL == 'U' & g6$NRC\_GENDER\_CODE == "G"))\*100

urb\_gal\_fail\_perc

rur\_gal\_fail\_perc <- (sum(g6$URBAN\_RURAL == 'R' & g6$NRC\_RESULT == 'F' & g6$NRC\_GENDER\_CODE == "G")/sum(g6$URBAN\_RURAL == 'R' & g6$NRC\_GENDER\_CODE == "G"))\*100

#Bar Chart of girls failing in Urban vs Rural

gal\_fail\_data <- data.frame(Area = factor(c("Urban\_Girls","Rural\_Girls"), levels=c("Urban\_Girls","Rural\_Girls")), Percentage = c(urb\_gal\_fail\_perc, rur\_gal\_fail\_perc))

ggplot(data=gal\_fail\_data, aes(x=Area, y=Percentage, fill=Area)) + geom\_bar(stat="identity")

**Observation:**

* From the data, we have analyzed that girls in rural area are performing better than girls in urban area.

**Conclusion:**

* Girls in rural area are better than girls in urban area.