Report on SSLC Data Analysis

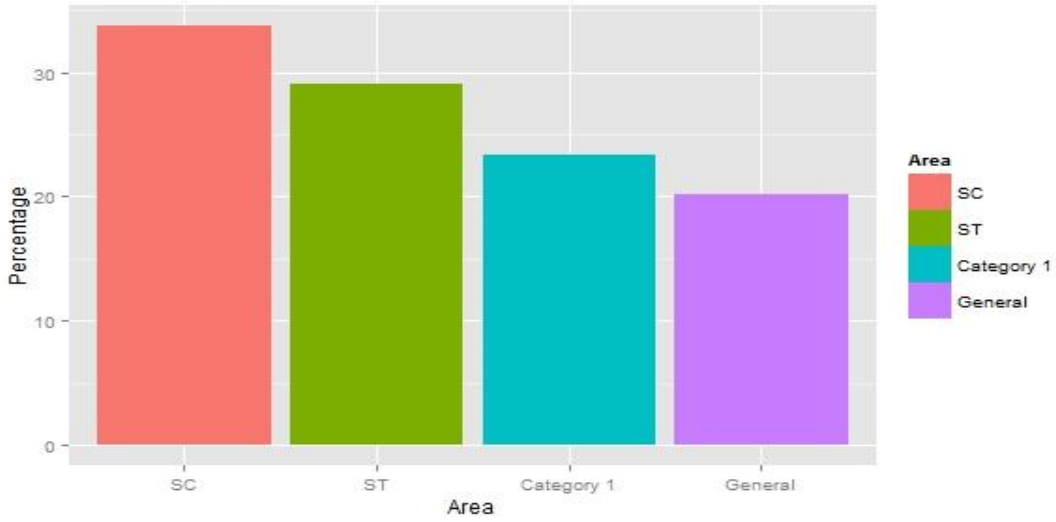
Group No. 06

Submitted by:
Ankit Mishra (MT201409)
Shan Mehrotra (MT20140111)
Kumar Rahul (MT2014058)

1. Negative observations

a. 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")
```

	lhs	rhs	support	confidence	lift
1	{URBAN_RURAL=R, NRC_CASTE_CODE=4, NRC_GENDER_CODE=G}	=> {NRC_RESULT=P}	0.15582256	0.8579705	1.1195529
2	{NRC_CASTE_CODE=4, NRC_GENDER_CODE=G}	=> {NRC_RESULT=P}	0.29049995	0.8481815	1.1067794
3	{URBAN_RURAL=R, NRC_GENDER_CODE=G}	=> {NRC_RESULT=P}	0.21613549	0.8252981	1.0769191
4	{NRC_GENDER_CODE=G}	=> {NRC_RESULT=P}	0.38587381	0.8164452	1.0653672
5	{URBAN_RURAL=R, NRC_CASTE_CODE=4}	=> {NRC_RESULT=P}	0.31249364	0.8163682	1.0652666
6	{NRC_CASTE_CODE=4}	=> {NRC_RESULT=P}	0.56453857	0.7981286	1.0414661
7	{URBAN_RURAL=R}	=> {NRC_RESULT=P}	0.44496487	0.7849359	1.0242512
8	{NRC_CASTE_CODE=3}	=> {NRC_RESULT=P}	0.03672403	0.7662890	0.9999190
9	{URBAN_RURAL=U}	=> {NRC_RESULT=P}	0.32138615	0.7420265	0.9682593
10	{NRC_GENDER_CODE=B}	=> {NRC_RESULT=P}	0.38047721	0.7214571	0.9414186
11	{NRC_CASTE_CODE=2}	=> {NRC_RESULT=P}	0.04473407	0.7093649	0.9256397
12	{NRC_CASTE_CODE=1}	=> {NRC_RESULT=P}	0.12035434	0.6624323	0.8643980

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.

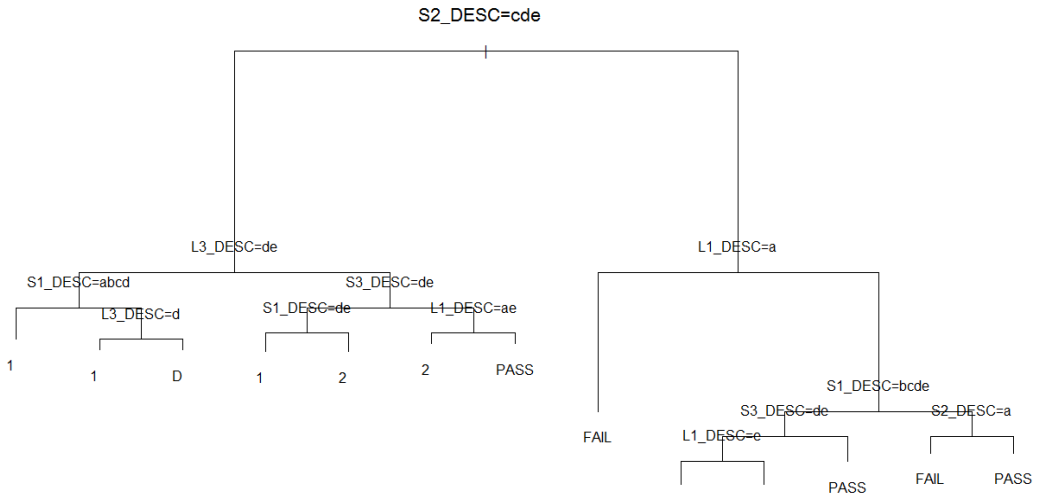
Conclusion:

- 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 :

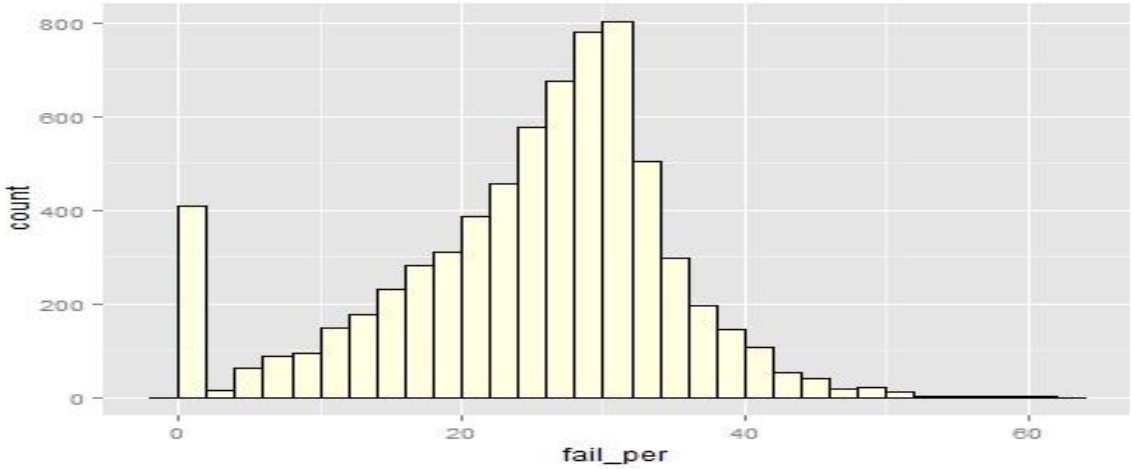
```
#Performing Discretization + Classification
library(rattle)
library(rpart.plot)
library(RColorBrewer)
g6$L1_DESC<-cut(g6$L1_MARKS,c(0,30,45,60,80,100),labels=c('F','P','2','1','D'))
g6$L2_DESC<-cut(g6$L2_MARKS,c(0,30,45,60,80,100),labels=c('F','P','2','1','D'))
g6$L3_DESC<-cut(g6$L3_MARKS,c(0,30,45,60,80,100),labels=c('F','P','2','1','D'))
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.

2. Positive observations

a. Comparing urban and rural students :

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

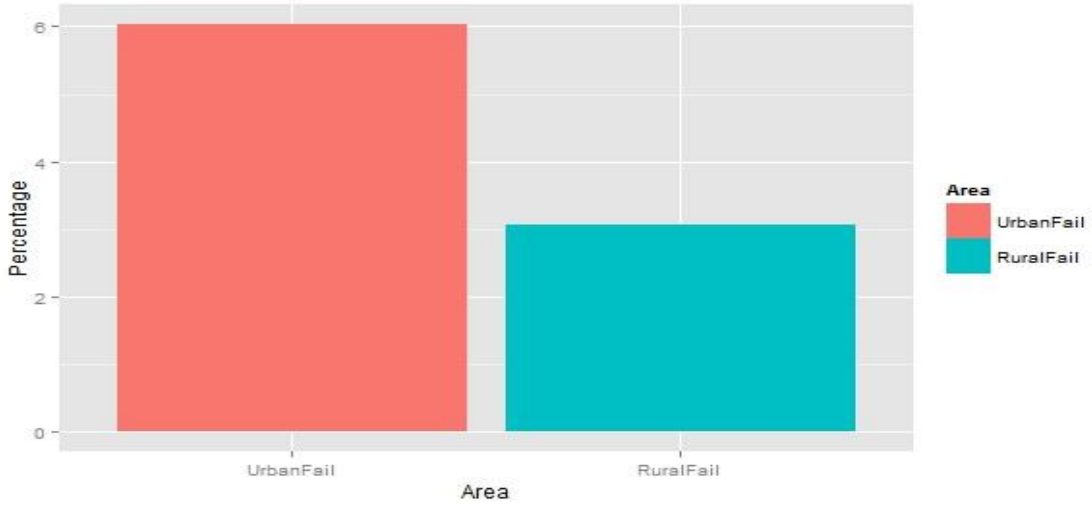


Fig.: L1=k and L2=E.

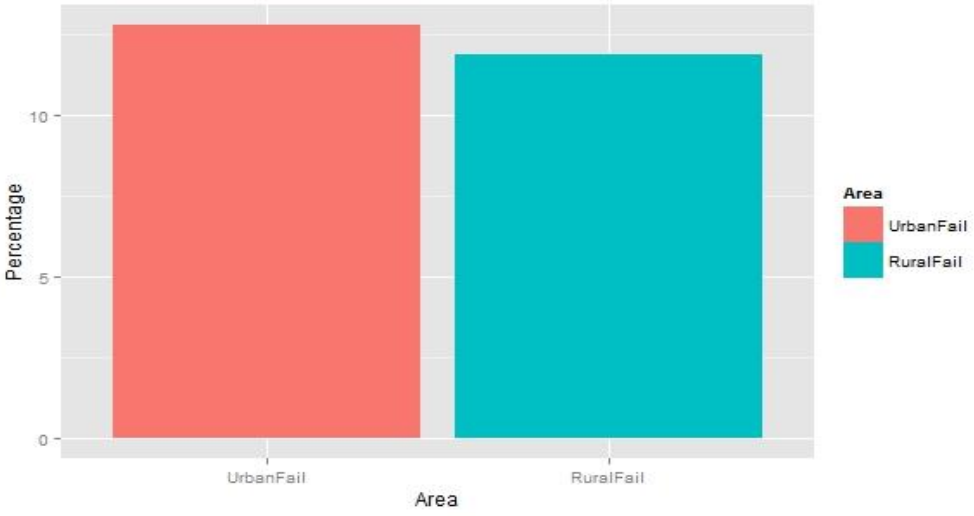


Fig.: L1=E and L2=K.

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:

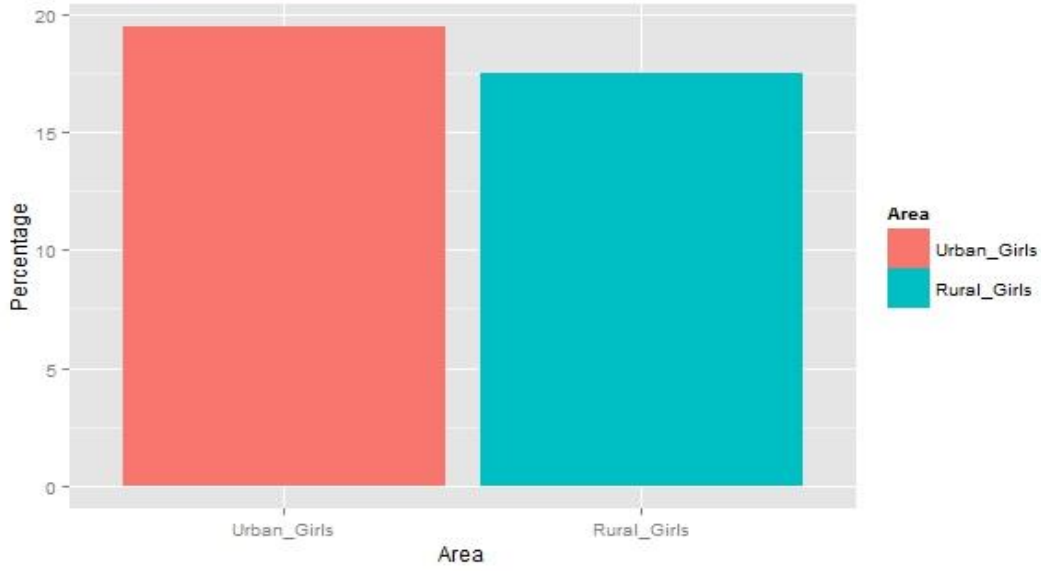
- Rural students are performing better than urban student in English.

Conclusion:

- ❖ Rural students are performing better.

b. Analyzing performance of girls in rural and urban areas :

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



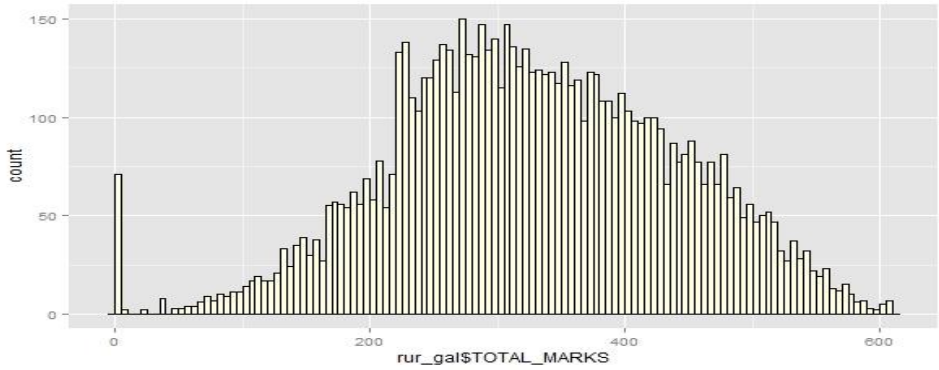


Fig. : Girls total marks in rural area.

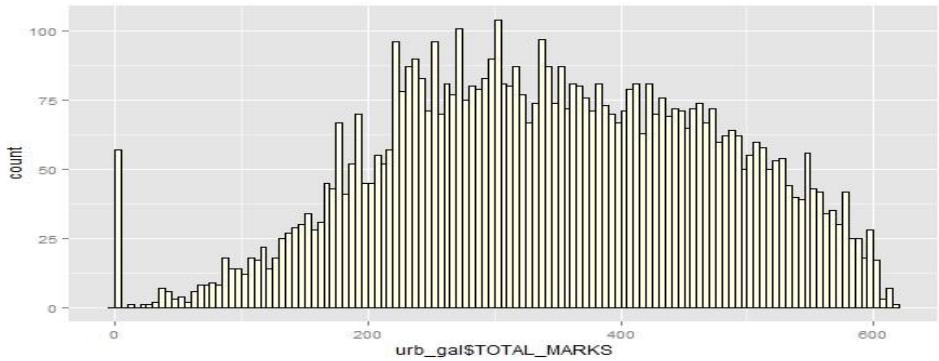


Fig. : Girls total marks in urban area.

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.