Data Analytics

Group 6

Ankit Mishra (MT2014009)

Kumar Rahul (MT2014058)

Shan Mehrotra (MT2014111)

Data Preparation

- Reviewed the data in Excel to gain insight.
- We found the following observations:
  1. All the students who were absent in the exams had '888' as their obtained marks.
  2. All the students whose marks were missing had result as NA.
  3. A student failing in any one of the exams is overall declared fail.
  4. Division of student is decided on the basis of available marks in exams excluding missing data.

Data Impurities

- We found following impurities in data:
  1. Marks columns had '*' in numeric data.
  2. DOB, Marks and many other columns had NA.
  3. Outliers such as '888' were present in the marks columns.
  4. Marks columns contained 'blanks' while their corresponding result columns had NA in them.
  5. Total Marks of many students were not equal to sum of marks of all exams.
  6. Division (NRC_Class) was wrongly specified.

Data Cleaning

- Removed the '*' from the data using 'gsub' and converted the vector to numeric data.

```
g6$L1_MARKS <- as.numeric(gsub('[^a-zA-Z0-9.]', '', g6$L1_MARKS))
```

- NA was present in many fields including non-numeric fields as well. We removed NA from numeric and date fields only.

```
g6$L1_MARKS[is.na(g6$L1_MARKS)] <- mean(g6$L1_MARKS, na.rm =T)
g6$DOB <- as.character(g6$DOB)
g6$DOB[is.na(g6$DOB)] <-"0/0/0000 0:00“
```

Contd...

- Blanks were present in Marks fields. Corresponding Result was NA. We imputed the blanks with mean and replaced NA with ‘P’.

```
blanks <- is.na(g6$L1_RESULT)
g6$L1_MARKS[blanks] <- mean(g6$L1_MARKS, na.rm =TRUE)
g6$L1_RESULT[blanks] <- “P”
```

- Replaced Outliers (888) with 0.

```
g6$L1_MARKS[g6$L1_MARKS == 888] <- 0
```

Contd...

- Since the Marks were changed due to replacement with mean, Total Marks and NRC_Class (division) was recomputed.

```
g6$TOTAL_MARKS = g6$L1_MARKS + g6$L2_MARKS +
g6$L3_MARKS + g6$S1_MARKS + g6$S2_MARKS + g6$S3_MARKS
```

- Computation of Division

```
g6$NRC_CLASS <- as.character(g6$NRC_CLASS)

g6$NRC_CLASS[(g6$TOTAL_MARKS/650)*100 >= 80 &
g6$NRC_RESULT == "P"] <- "D"
```

Exploratory Analysis

- Summary of the data.

	NRC_STUDENT_NAME	NRC_MOTHER_NAME	NRC_FATHER_NAME	NRC_CASTE_CODE	NRC_GENDER_CODE	NRC_MEDIUM	NRC_PHYSICAL_CONDITION	NRC_CENTER_CODE
POOJA	: 37	RATHNAMMA : 377	BASAPPA : 158	Min. :1.000	B:15538	E: 7373	B: 6	061SS : 31
ASHWINI	: 33	MANJULA : 331	BASAVARAJ : 146	1st Qu.:3.000	G:13925	H: 9	D: 19	002GG : 29
BASAVARAJ	: 23	RENUKA : 253	NAGARAJU : 133	Median :4.000		K:20575	H: 11	038QQ : 29
RENUKA	: 22	LAKSHMAMMA: 240	KRISHNAPPA: 120	Mean :3.281		L: 32	N:29374	050QQ : 29
SHILPA	: 21	GOWRAMMA : 232	MALLAPPA : 116	3rd Qu.:4.000		M: 562	P: 38	013OA : 28
MANJULA	: 19	(Other) :27907	(Other) :28759	Max. :4.000		T: 8	S: 13	023RA : 28
(Other)	:29308	NA's : 123	NA's : 31			U: 904	X: 2	(Other):29289

L1_MARKS	L1_RESULT	L2_MARKS	L2_RESULT	L3_MARKS	L3_RESULT	S1_MARKS	S1_RESULT	S2_MARKS	S2_RESULT	S3_MARKS	S3_RESULT
Min. : 0.00	A: 544	Min. : 0.00	A: 613	Min. : 0.00	A: 524	Min. : 0.00	A: 675	Min. : 0.00	A: 2	Min. : 0.00	A: 594
1st Qu.: 47.00	F: 3616	1st Qu.: 30.00	F: 3364	1st Qu.: 35.00	F: 2106	1st Qu.: 35.00	F: 3739	1st Qu.: 32.00	F: 5025	1st Qu.: 39.00	F: 2677
Median : 72.00	P:25303	Median : 40.00	P:25486	Median : 47.00	P:26833	Median : 47.00	P:25049	Median : 43.00	P:24436	Median : 56.00	P:26192
Mean : 70.22		Mean : 46.27		Mean : 51.04		Mean : 47.91		Mean : 44.12		Mean : 55.75	
3rd Qu.: 96.00		3rd Qu.: 63.00		3rd Qu.: 69.00		3rd Qu.: 62.00		3rd Qu.: 56.00		3rd Qu.: 74.00	
Max. :125.00		Max. :100.00		Max. :100.00		Max. :100.00		Max. :100.00		Max. :100.00	

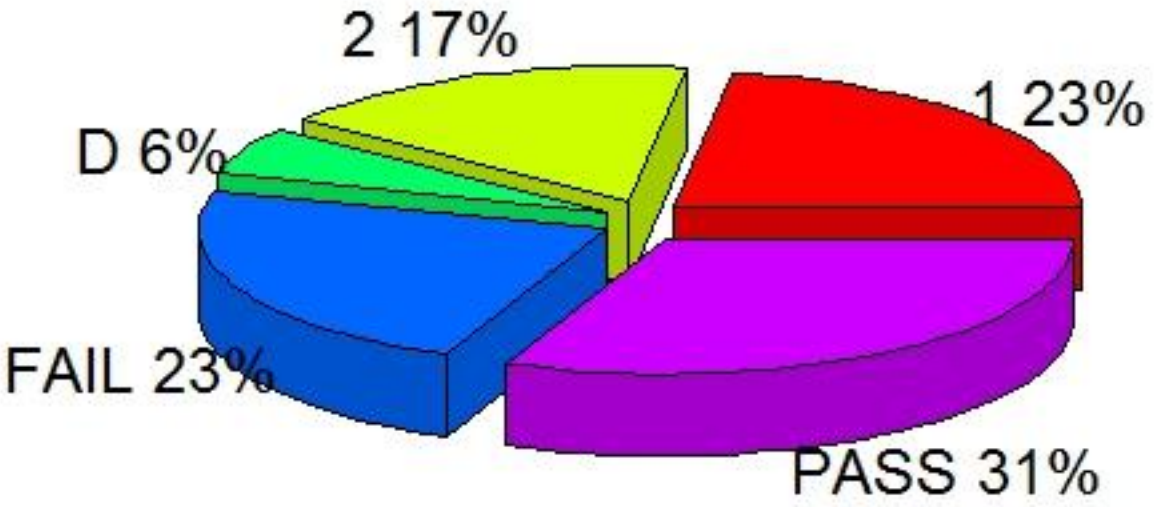
TOTAL_MARKS	NRC_RESULT	NRC_CLASS	CANDIDATE_TYPE	SCHOOL_NAME	L1_CODE	L2_CODE
Min. : 0.0	F: 6884	1 :6683	NSPR: 310	SIDDALINGESWARA RES. H S SIDDAGANGAMUTT, TUMKUR TUMKUR DISTRICT : 41	01K :24577	31E:26311
1st Qu.:231.0	P:22579	2 :5008	NSR : 2131	GOVERNMENT GIRLS JUNIOR COLLEGE TIPTUR, TUMKUR DISTRICT : 23	14E : 2216	33K: 3119
Median :309.0		D :1647	PF : 747	GPU COLLEGE GOWRIBIDNUR CHIKKABALLAPURA DISTRICT : 23	12U : 1237	X*0: 23
Mean :315.3		FAIL:6884	RF :26196	NATIONAL HIGH SCHOOL BASAVANAGUDI BANGALORE SOUTH : 22	16S : 640	X0 : 10
3rd Qu.:404.0		PASS:9241	RSPR: 9	GOVT JUNIOR COLLEGE FOR BOYS CHICKBALLAPUR CHIKKABALLAPURA DISTRICT : 21	08H : 561	
Max. :620.0			RSR : 70	GOVT. JUNIOR COLLEGE CHICKMAGALUR CHICKMAGALUR DISTRICT : 21	06H : 121	
				(Other) :29312	(Other): 111	

Contd..

- **Pie Chart to depict percentage distribution of division.**

```
w = table(g6$NRC_CLASS)


t = as.data.frame(w)


perc <- round(t$Freq/sum(t$Freq)*100)
lbls <- paste(t$Var1, perc) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels


library(plotrix)
pie3D(t$Freq, labels = lbls, explode = 0.1, main = "Pie chart of
Divisions")
```

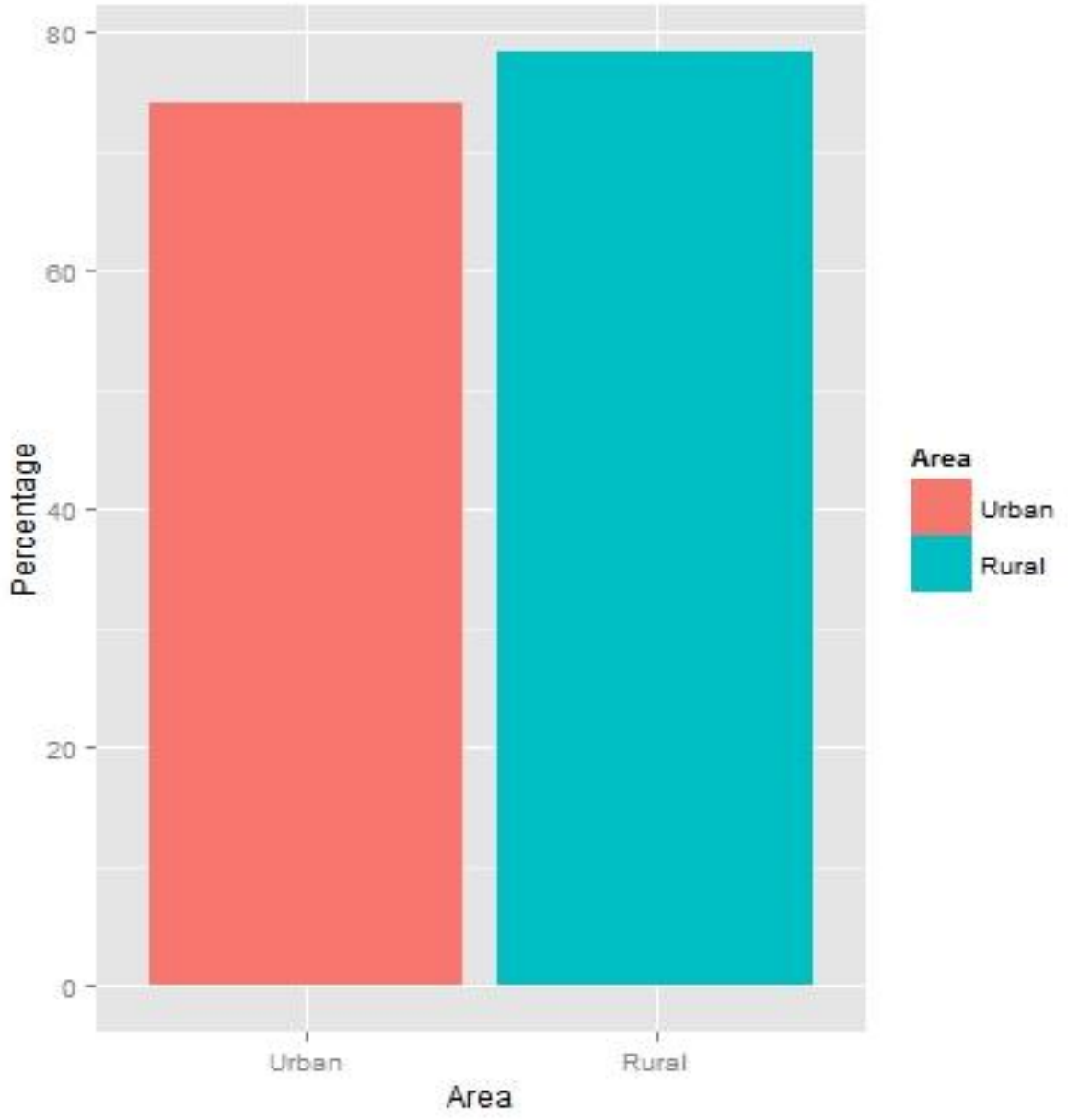
Pie chart of Divisions



Contd..

- **Bar Graph showing area-wise distribution of passed students.**

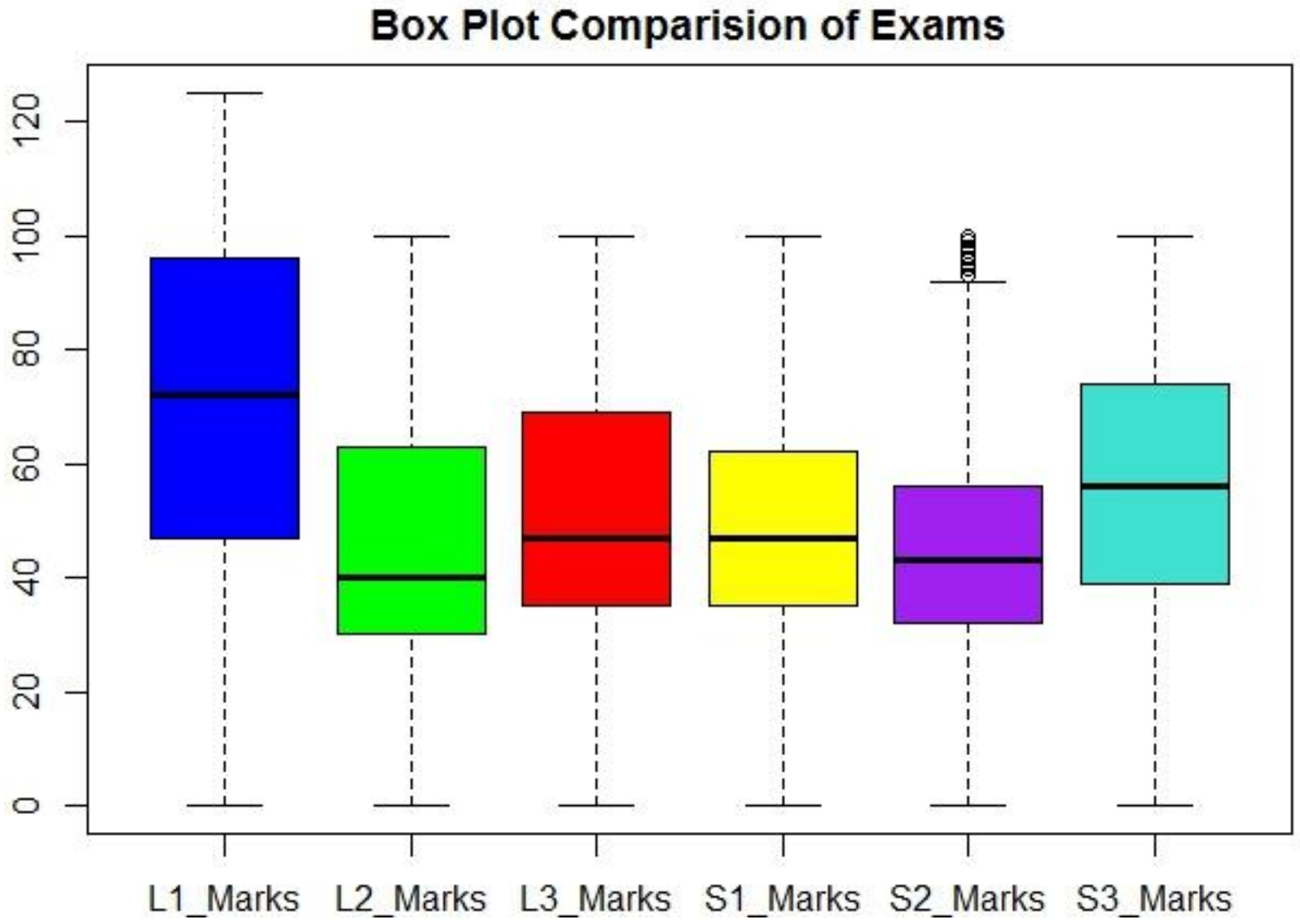
```
urb_pass_per <- (sum(g6$URBAN_RURAL == 'U' & g6$NRC_RESULT
   == 'P')/sum(g6$URBAN_RURAL == 'U'))*100


rur_pass_per <- (sum(g6$URBAN_RURAL == 'R' & g6$NRC_RESULT ==
   'P')/sum(g6$URBAN_RURAL == 'R'))*100


dat_pass <- data.frame(Area = factor(c("Urban","Rural"),
   levels=c("Urban","Rural")), Percentage = c(urb_pass_per, rur_pass_per))


ggplot(data=dat_pass, aes(x=Area, y=Percentage, fill=Area)) +
   geom_bar(stat="identity")
```


80
60
40
20
0
Percentage
Urban
Rural
Area
Area
Urban
Rural

Contd..

- **Box Plot showing the comparison of marks in different exams.**

```
color <- c("blue","green","red","yellow","purple","turquoise")

exam_name <-
c("L1_Marks","L2_Marks","L3_Marks","S1_Marks","S2_Marks","S3_Marks
")

boxplot(g6$L1_MARKS, g6$L2_MARKS, g6$L3_MARKS,
g6$S1_MARKS,g6$S2_MARKS, g6$S3_MARKS, data = g6, col = color,
names = exam_name, main = "Box Plot Comparision of Exams")
```

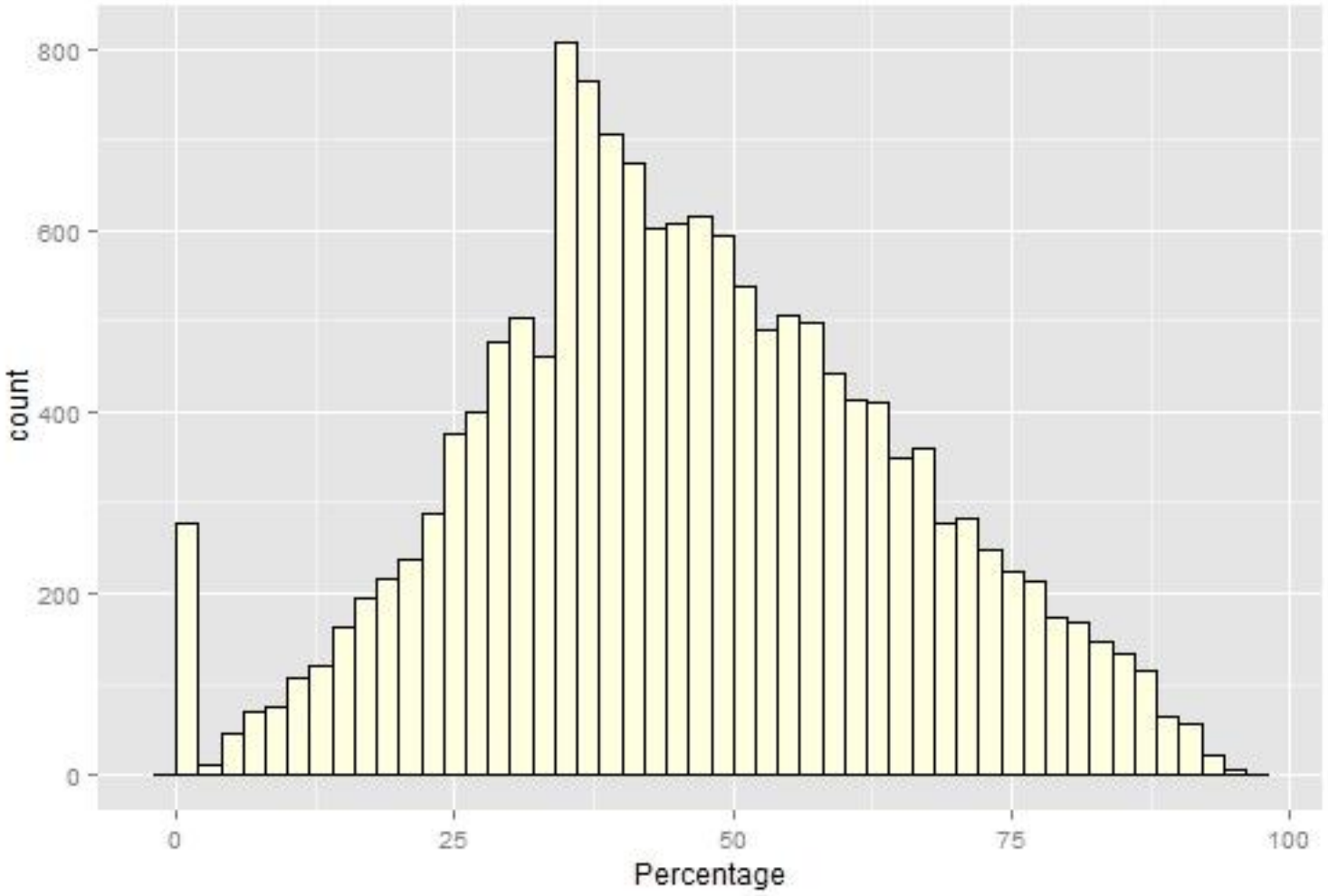

Box Plot Comparision of Exams
0
20
40
60
80
100
120
L1_Marks
L2_Marks
L3_Marks
S1_Marks
S2_Marks
S3_Marks

Contd..

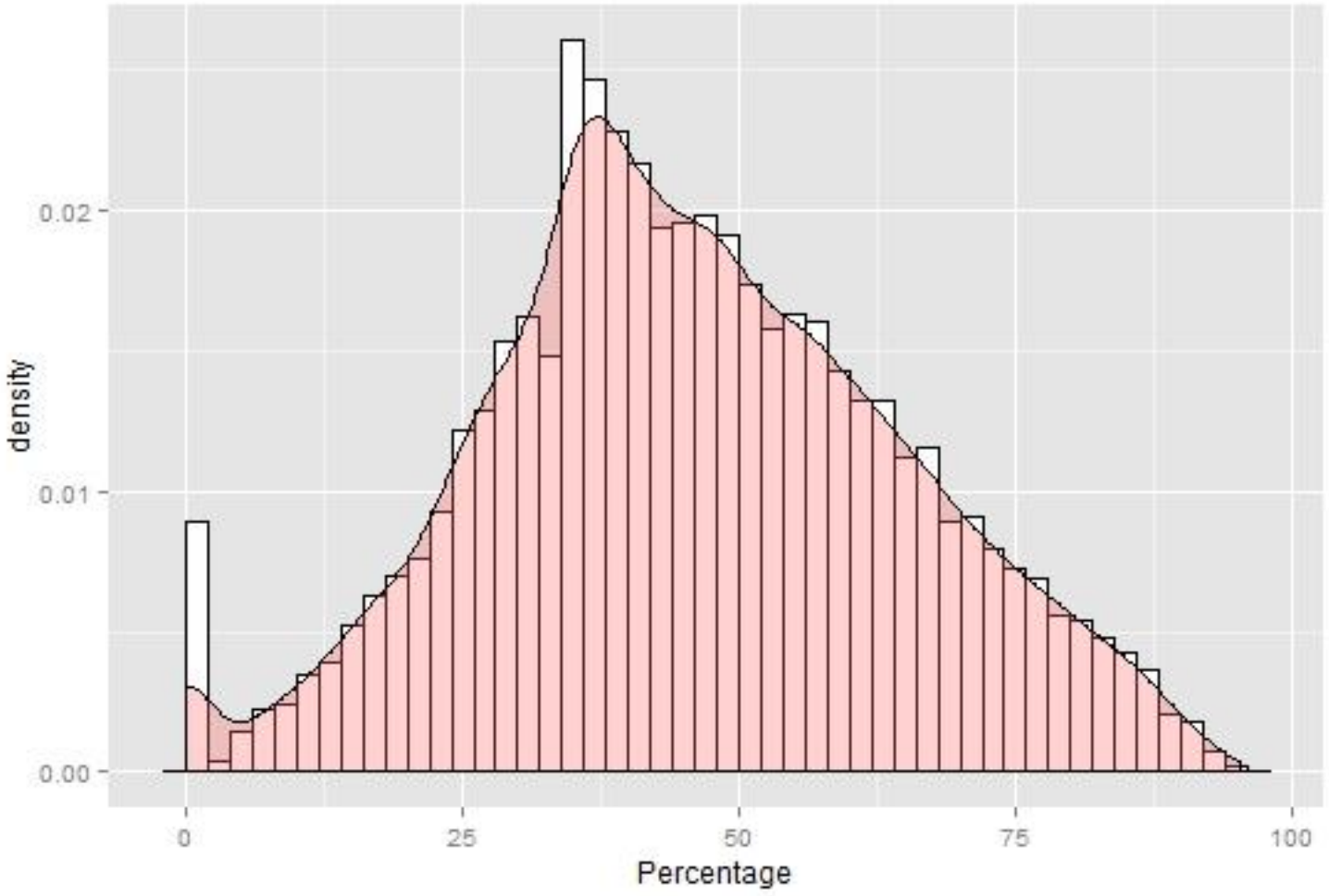
- **Histogram & Density Plot depicting the percentage distribution of Boys & Girls.**

```
boy_data <- filter(g6, NRC_GENDER_CODE == 'B')

Percentage <- (boy_data$TOTAL_MARKS/650)*100

ggplot(boy_data, aes(x=Percentage)) +
geom_histogram(binwidth=2, color = "black", fill = "#FFFFE0")
```

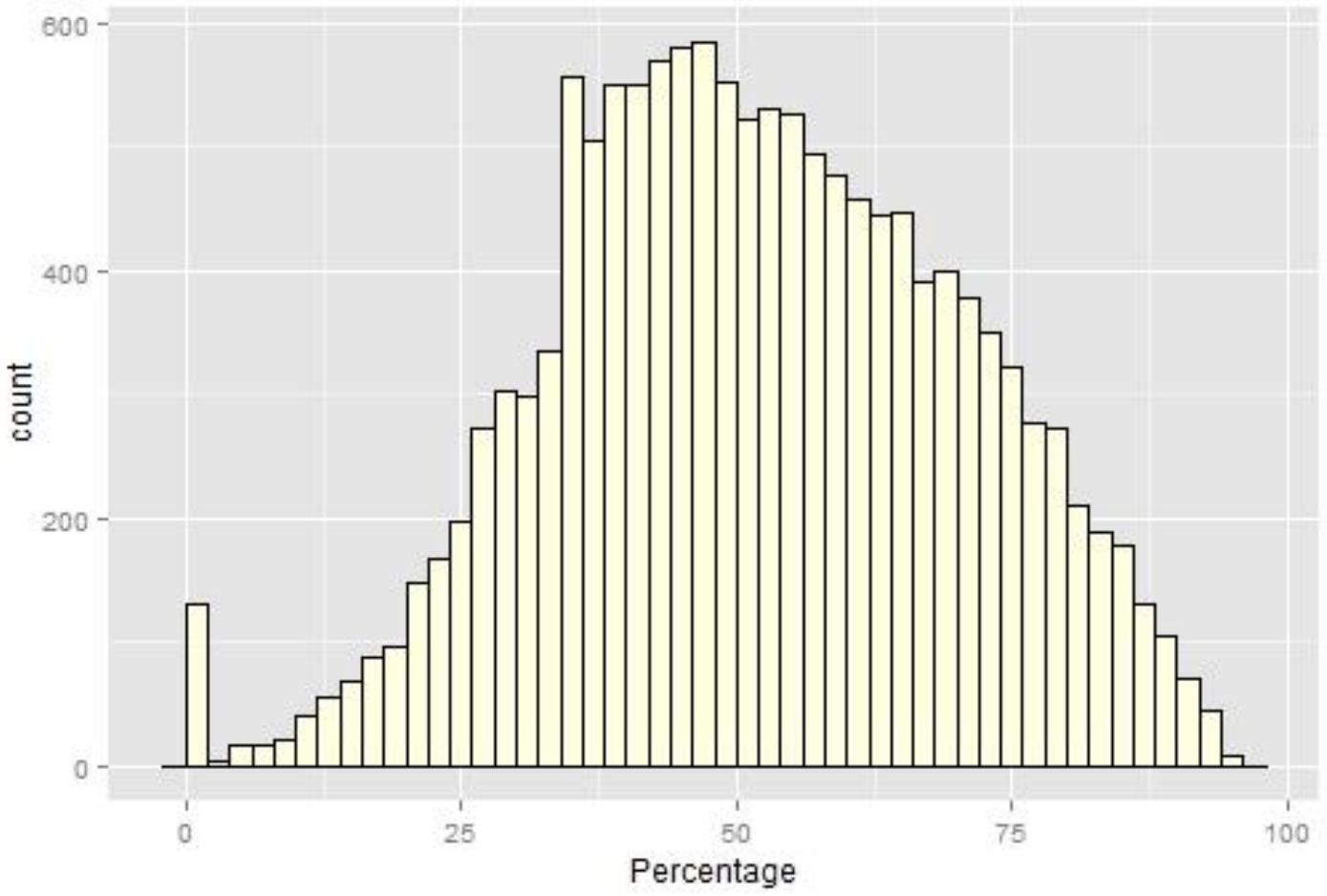
Histogram for Boys



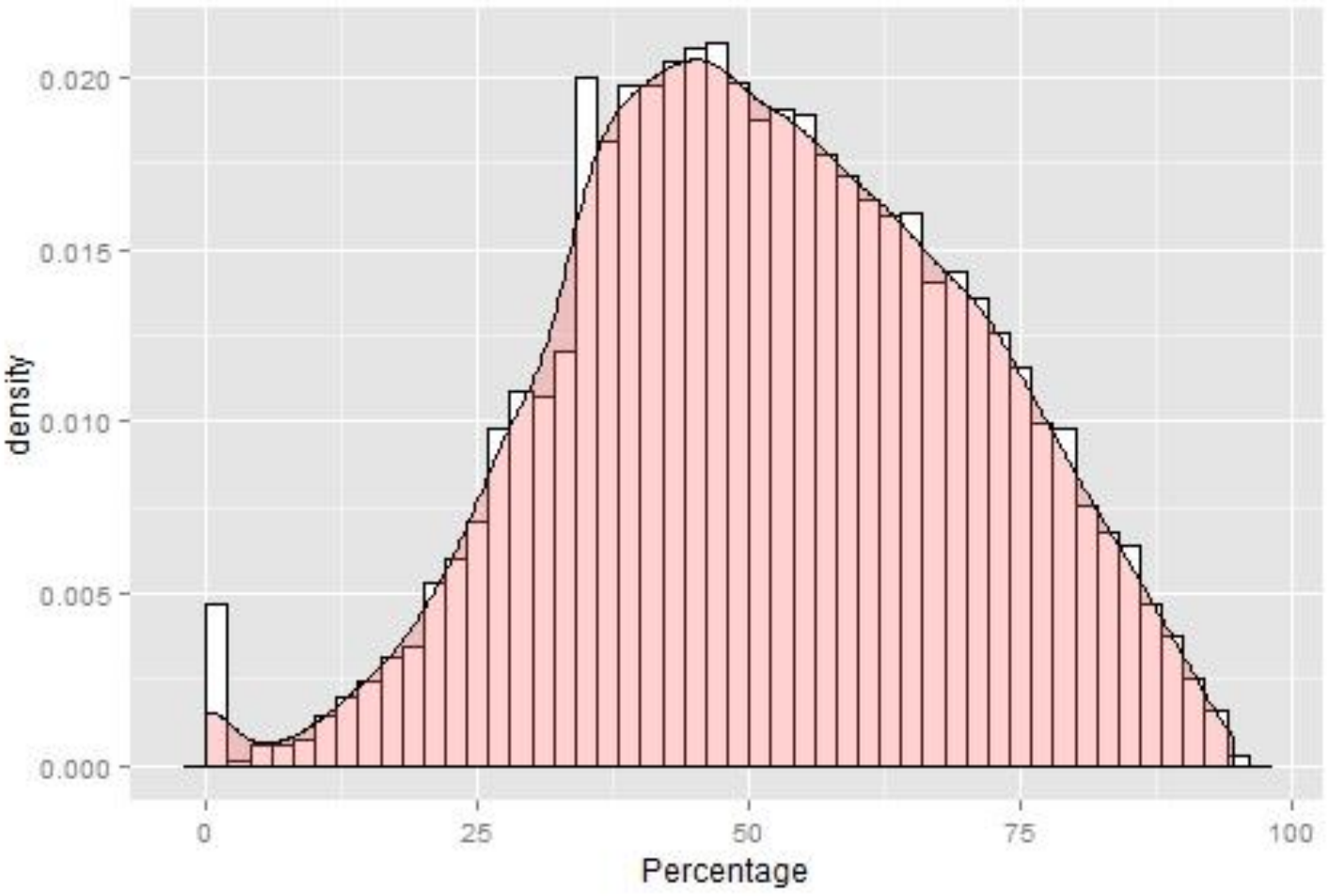
Density Plot for Boys



Histogram for Girls



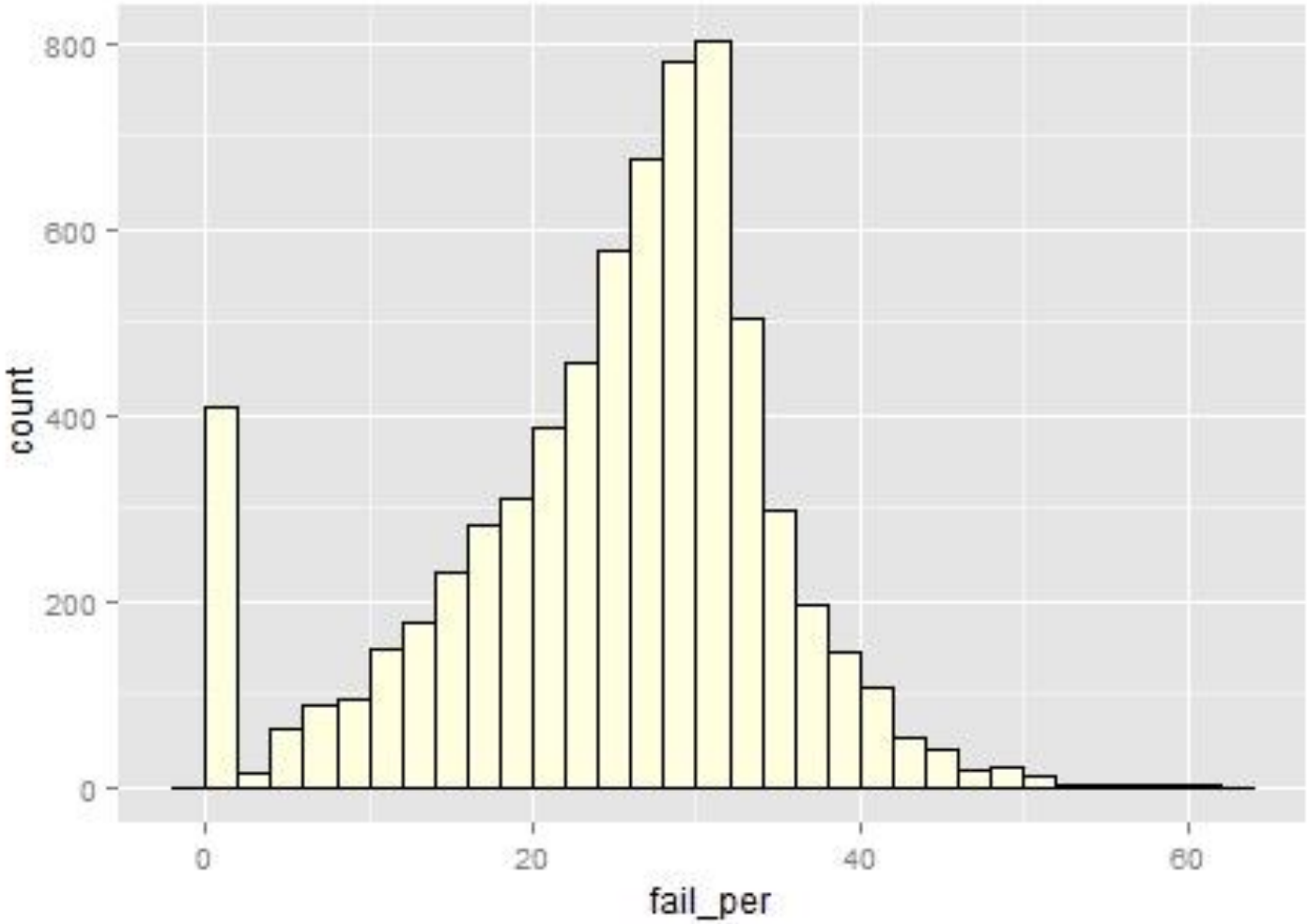
Density Plot for Girls



Contd..

- **Histogram showing percentage distribution of failed students.**

```
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")
```


800
600
400
200
0
count
0
20
40
60
fail_per

THANK YOU!!