Exercise 4 - Data Visualisation using R

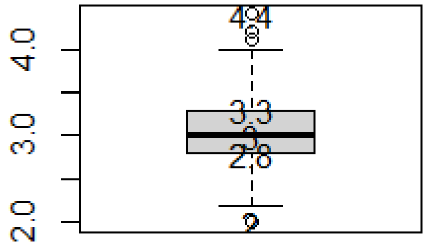
S.DEIVANAYKI (21BCS003)

06.02.2024

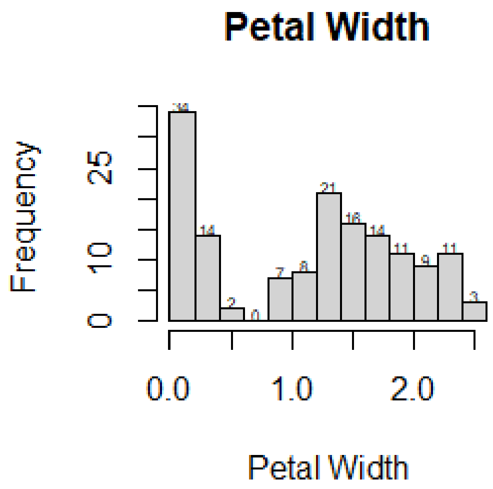
R Markdown

1. Extract the maximum information from a Histogram and Boxplot.

```
boxplot(iris$Sepal.Width)
text(y=fivenum(iris$Sepal.Width),x=1,labels = fivenum(iris$Sepal.Width))
```

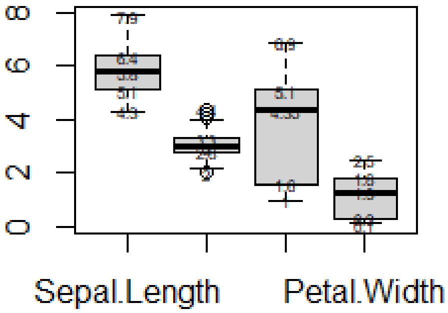


```
h <- hist(iris$Petal.Width,xlab="Petal Width",main="Petal Width")
text(h$mids,h$counts,labels=h$counts,cex=0.5, adj=c(0.5, 0))
```



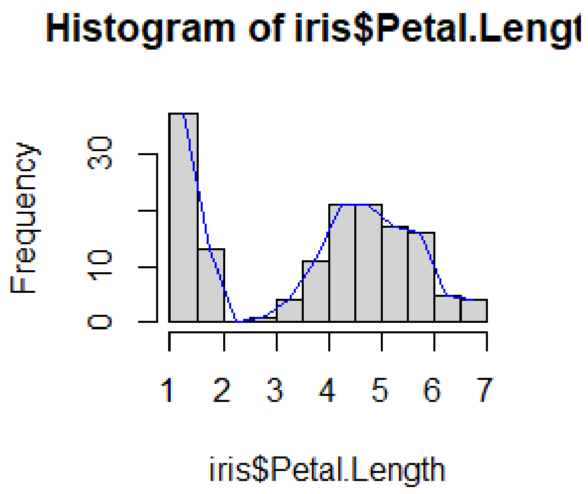
2. Consider the iris dataset, create new variable called boxplot_data that excludes the species column.

```
library(tidyverse)

boxplot_data <- iris %>% select(-Species)
boxplot(boxplot_data)
text(y=fivenum(boxplot_data$Sepal.Length),x=1,labels =
fivenum(iris$Sepal.Length),cex=0.5)
text(y=fivenum(boxplot_data$Sepal.Width),x=2,labels =
fivenum(iris$Sepal.Width),cex=0.5)
text(y=fivenum(boxplot_data$Petal.Length),x=3,labels =
fivenum(iris$Petal.Length),cex=0.5)
text(y=fivenum(boxplot_data$Petal.Width),x=4,labels =
fivenum(iris$Petal.Width),cex=0.5)
```



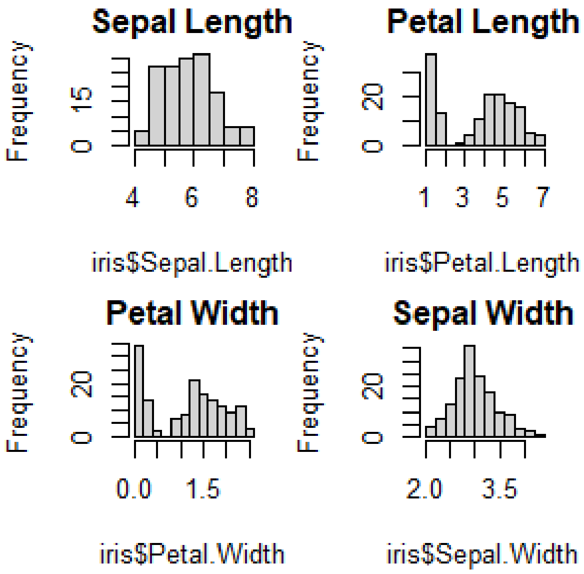
3. Create Histogram and illustrate the distribution look like within the petal length feature of the Iris data set.

```r
h<-hist(iris$Petal.Length)
mid <- (h$breaks[-1] + h$breaks[-length(h$breaks)]) / 2
lines(mid, h$counts, col = "blue")
```



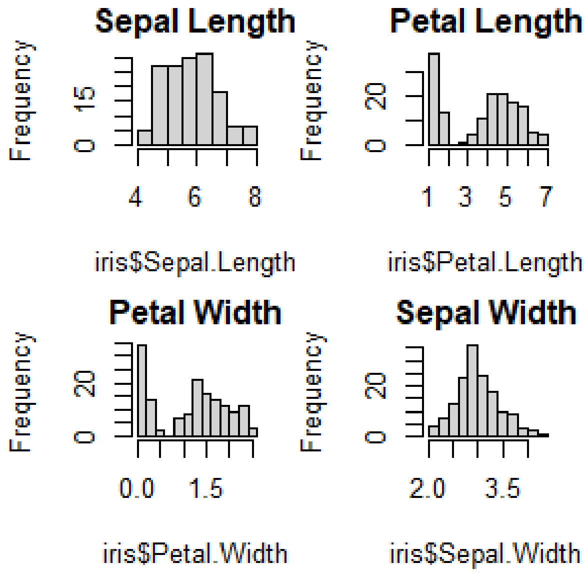
4. Create Histogram that plot multiple features at once for any dataset.

```r
par(mfrow=c(2, 2), mar=c(4, 4, 2, 1))
hist(iris$Sepal.Length, main="Sepal Length")
hist(iris$Petal.Length, main="Petal Length")
hist(iris$Petal.Width, main="Petal Width")
hist(iris$Sepal.Width, main="Sepal Width")
```



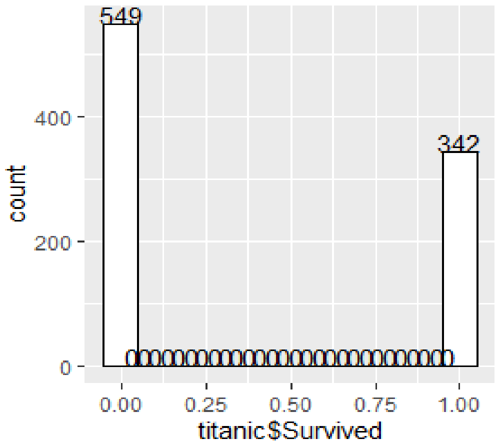
5. Plot every feature from the iris dataset in a histogram.

```r
par(mfrow=c(2, 2), mar=c(4, 4, 2, 1))
hist(iris$Sepal.Length, main="Sepal Length")
hist(iris$Petal.Length, main="Petal Length")
hist(iris$Petal.Width, main="Petal Width")
hist(iris$Sepal.Width, main="Sepal Width")
```



6. Consider a Titanic dataset and compare metric value across different subgroups of the data. Also assume you have a greater number of groups, which visualization method do you prefer over a column chart?

```r
titanic<-read.csv("C:/Users/DSL-A-B1/Downloads/Titanic.csv")
ggplot(titanic,aes(x=titanic$Survived))+geom_histogram(binwidth =
0.1,color="black",fill="white")+stat_bin(aes(label = ..count..), geom =
"text", vjust = 0)
```

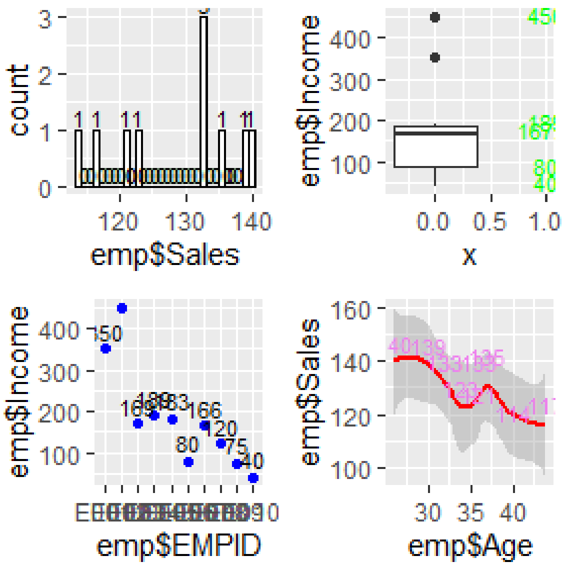


7. For the given dataset and plot different charts.

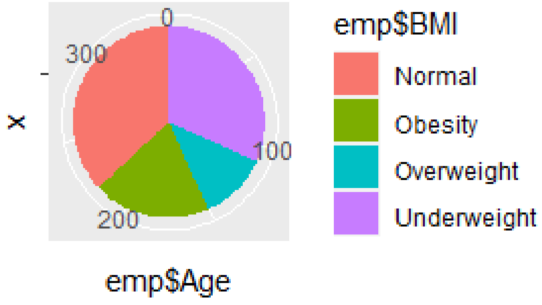
```
emp<-
data.frame(EMPID=c('E001','E002','E003','E004','E005','E006','E007','E008','E
009','E010'),Gender=c('M','F','F','M','F','M','M','F','M','M'),Age=c(34,40,37
,30,44,36,32,26,32,36),Sales=c(123,114,135,139,117,121,133,140,133,133),BMI=c
('Normal','Overweight','Obesity','Underweight','Underweight','Normal','Obesit
y','Normal','Normal','Underweight'),Income=c(350,450,169,189,183,80,166,120,7
5,40))
head(emp,3)
```

```
##   EMPID Gender Age Sales        BMI Income
## 1  E001      M  34   123     Normal    350
## 2  E002      F  40   114 Overweight    450
## 3  E003      F  37   135    Obesity    169
```

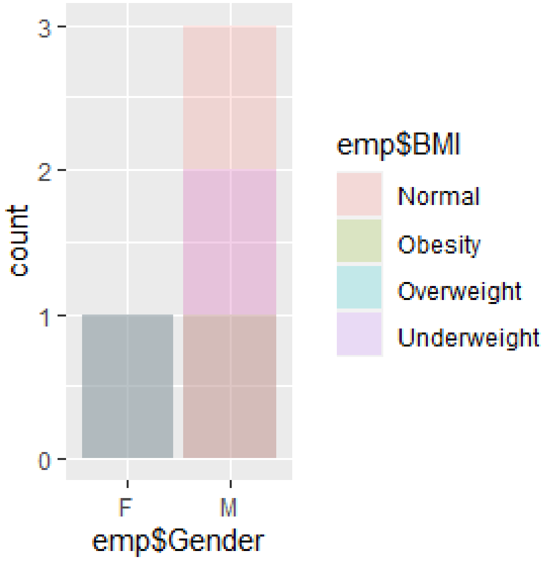
```
library(gridExtra)

f1 <-
ggplot(emp,aes(x=emp$Sales))+geom_histogram(color="black",fill="white")+stat_
bin(geom = "text", aes(label = ..count..),vjust=-0.2,size=3)

f2 <- ggplot(emp,aes(,y=emp$Income))+geom_boxplot()+annotate("text", x = 1, y
= fivenum(emp$Income), label =
fivenum(emp$Income),vjust=0.3,color="green",size=3)

f3 <- ggplot(emp,aes(x=emp$EMPID,y=emp$Income)) + geom_point(color="blue")
+geom_text(aes(label = Income), vjust = -0.5, hjust = 0.5, color =
"black",size=3)

f4 <- ggplot(emp,aes(x=emp$Age,y=emp$Sales)) +
geom_smooth(color="red")+geom_text(aes(label = Sales), color = "violet",
vjust = -0.5, hjust = 0.5,size=3)

grid.arrange(f1,f2,f3,f4,ncol=2)
```



```
ggplot(emp,aes(x="",y=emp$Age,fill=emp$BMI)) + geom_bar(stat="identity",
width=1) +  coord_polar("y", start=0)
```

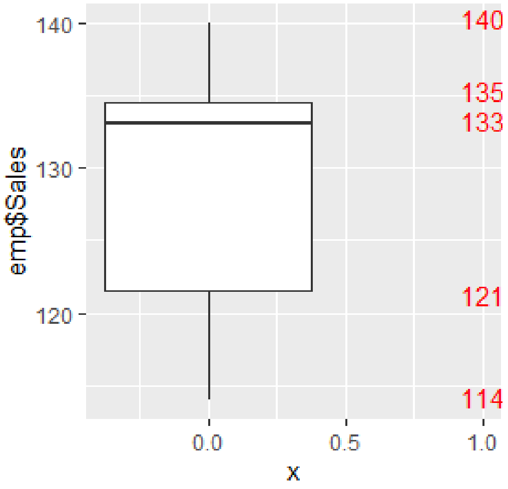


```
ggplot(emp,aes(x=emp$Gender,fill=emp$BMI)) + geom_bar(position="identity",
alpha=1/5)
```



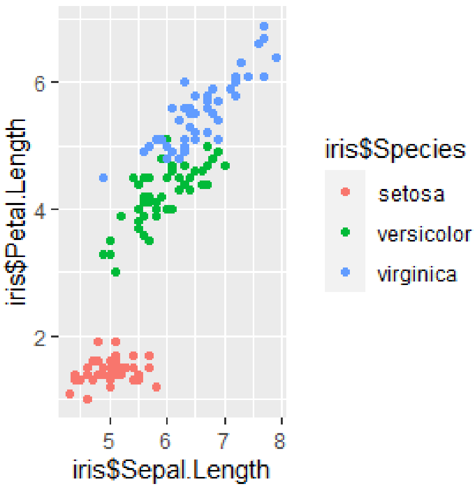
8. Consider the above dataset and draw the boxplot for the statistical data based on the minimum, first quartile, median, third quartile and maximum.

```
ggplot(emp,aes(y=emp$Sales))+geom_boxplot()+annotate("text", x = 1, y =
fivenum(emp$Sales), label = fivenum(emp$Sales),vjust=0.3,color="red")
```

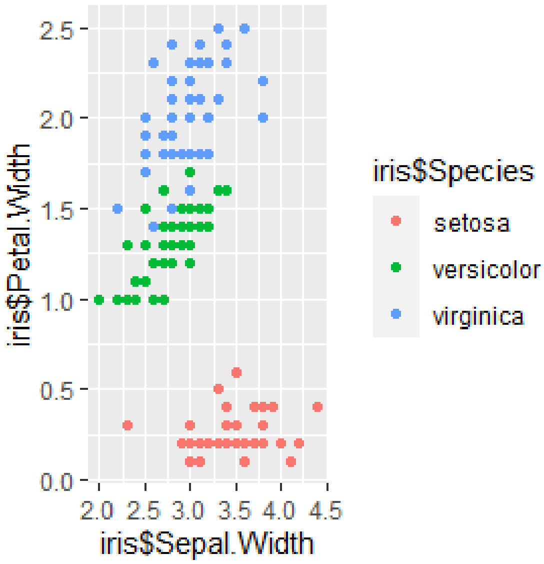


9. Make a scatterplot for the features in the Iris dataset.

```
ggplot(iris,aes(x=iris$Sepal.Length,y=iris$Petal.Length,color=iris$Species))
+ geom_point()
```

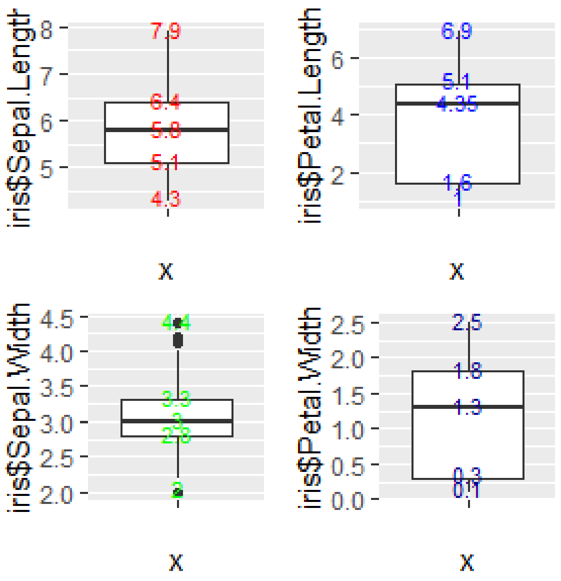


```
ggplot(iris,aes(x=iris$Sepal.Width,y=iris$Petal.Width,color=iris$Species)) +
geom_point()
```



10. Look at an individual feature through a boxplot.

```
p1 <- ggplot(iris,mapping = aes(x="",y=iris$Sepal.Length)) +
geom_boxplot()+annotate("text", x = 1, y = fivenum(iris$Sepal.Length), label
= fivenum(iris$Sepal.Length),vjust=0.3,color="red",size=3)
p2 <- ggplot(iris,mapping = aes(x="",y=iris$Petal.Length)) +
geom_boxplot()+annotate("text", x = 1, y = fivenum(iris$Petal.Length), label
= fivenum(iris$Petal.Length),vjust=0.3,color="blue",size=3)
p3 <- ggplot(iris,mapping = aes(x="",y=iris$Sepal.Width)) +
geom_boxplot()+annotate("text", x = 1, y = fivenum(iris$Sepal.Width), label =
fivenum(iris$Sepal.Width),vjust=0.3,color="green",size=3)
p4 <- ggplot(iris,mapping = aes(x="",y=iris$Petal.Width)) +
geom_boxplot()+annotate("text", x = 1, y = fivenum(iris$Petal.Width), label =
fivenum(iris$Petal.Width),vjust=0.3,color="darkblue",size=3)
grid.arrange(p1,p2,p3,p4,ncol = 2)
```



Exercise - 3

1. Create a synthetic dataset of 50 entries (randomly generated) with the following fields and store it in a file

```
stud <-
data.frame(StudentId=sample(1:50,50,replace=FALSE),Dept=sample(c('CSE','ECE',
'EEE','IT','Civil','Mech'),50,replace=TRUE),Year1=sample(125:130,50,replace=T
RUE),Year2=sample(135:140,50,replace=TRUE),Year3=sample(145:150,50,replace=TR
UE),Year4=sample(155:160,50,replace=TRUE))
head(stud)

##   StudentId Dept Year1 Year2 Year3 Year4
## 1        47  CSE   125   140   149   158
## 2        30  CSE   127   140   147   156
## 3        16  EEE   125   138   148   156


write_delim(stud,"Student.txt",delim = "\t")

studtab <- read_delim("C:/Users/DSL-A-
B1/Downloads/RLAB/Student.txt",delim="\t")

head(studtab)

## # A tibble: 6 × 6
##   StudentId Dept  Year1 Year2 Year3 Year4
##       <dbl> <chr> <dbl> <dbl> <dbl> <dbl>
## 1        46 Mech    129   137   148   160
## 2        21 ECE     126   137   146   159
## 3         1 IT      126   138   148   156
```

2. Create a new variable avg_height for each students and compute its value

```
studtab %>% group_by(StudentId) %>%
summarize(Avg_Height=(Year1+Year2+Year3+Year4)/4)
```

```
## # A tibble: 50 × 2
##    StudentId Avg_Height
##        <dbl>      <dbl>
##  1         1       142
##  2         2       142.
##  3         3       144.
##  4         4       142
```

3. Compute dept wise count of students

```
studtab %>% group_by(Dept) %>% summarize(No_Of_Student=n())
```

```
## # A tibble: 6 × 2
##   Dept  No_Of_Student
##   <chr>         <int>
## 1 CSE               4
## 2 Civil             7
## 3 ECE              14
```

Exercise - 4

1. Create a synthetic dataset of 100 entries (randomly generated) with the following fields and store it in a .csv file:

```
stud1 <-
data.frame(StudentId=sample(1:100,50,replace=FALSE),Dept=sample(factor(c('CSE
','ECE','EEE','IT','Civil','Mech')),100,replace=TRUE),Sem=sample(factor(c('I'
,'II','III','IV','V','VI','VII','VIII')),100,replace=TRUE),GPA=sample(seq(5,1
0,0.1),100,replace=TRUE))
head(stud1)
```

```
##   StudentId  Dept Sem GPA
## 1        44   EEE   I 5.7
## 2        29   EEE  VI 9.1
## 3        56   CSE   I 8.9
```

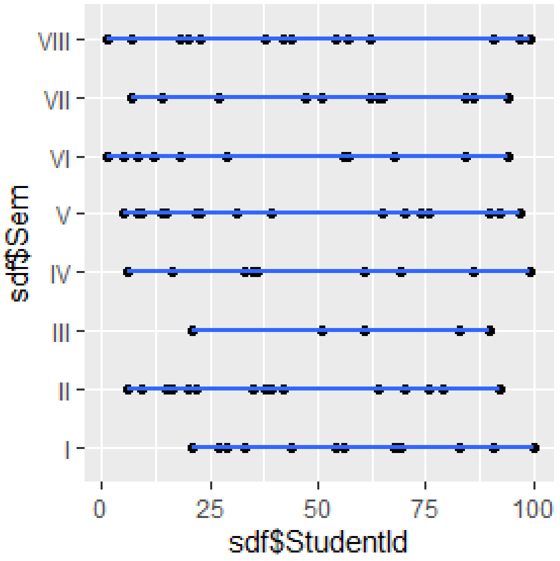
```
write.csv(stud1,"Student.csv")
```

```
sdf<-read.csv("C:/Users/DSL-A-B1/Downloads/Student.csv")
head(sdf)
```

```
##   X StudentId  Dept Sem GPA
## 1 1        44   EEE   I 5.7
## 2 2        29   EEE  VI 9.1
## 3 3        56    CSE    I 8.9
```

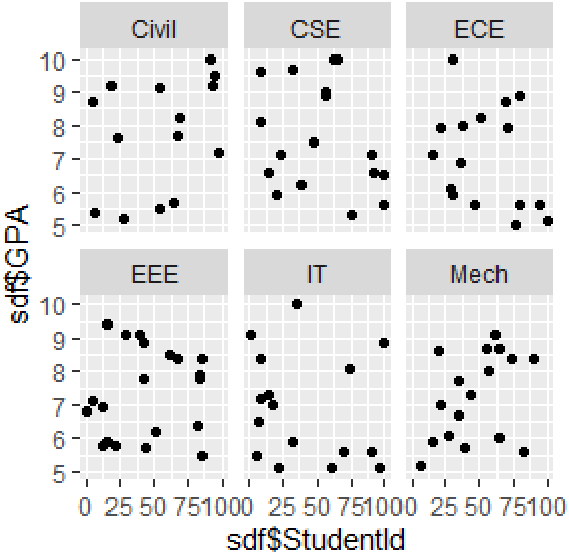
2. Scatterplot with smooth curve for every semester with different pattern in batch-wise mentioned

```
ggplot(sdf,aes(x=sdf$StudentId,y=sdf$Sem))+geom_point()+geom_smooth()
```



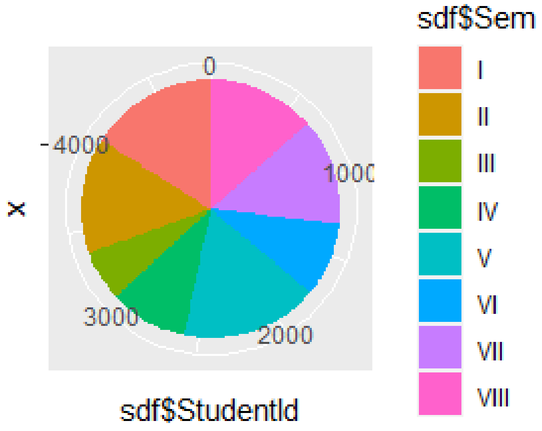
3. Draw subsets of scatterplot to plot GPA for each department.

```
ggplot(sdf)+geom_point(aes(x=sdf$StudentId,y=sdf$GPA))+facet_wrap(~
sdf$Dept,nrow = 2)
```



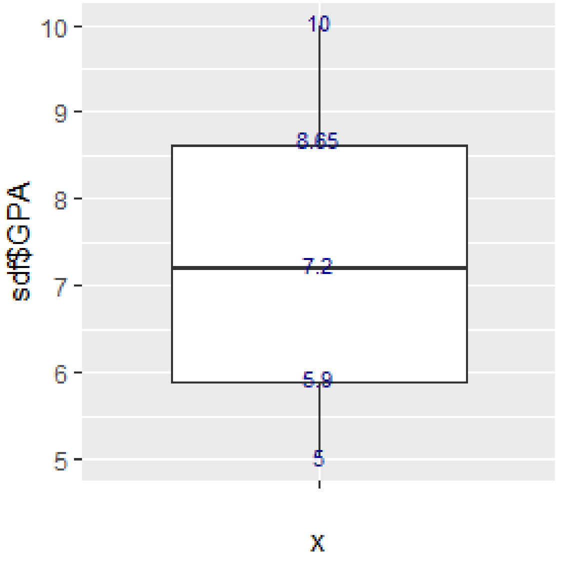
4. Bar chart specification mentioned in batch-wise

```
ggplot(sdf,aes(x="",y=sdf$StudentId,fill=sdf$Sem)) +
geom_bar(stat="identity", width=1) +  coord_polar("y", start=0)
```



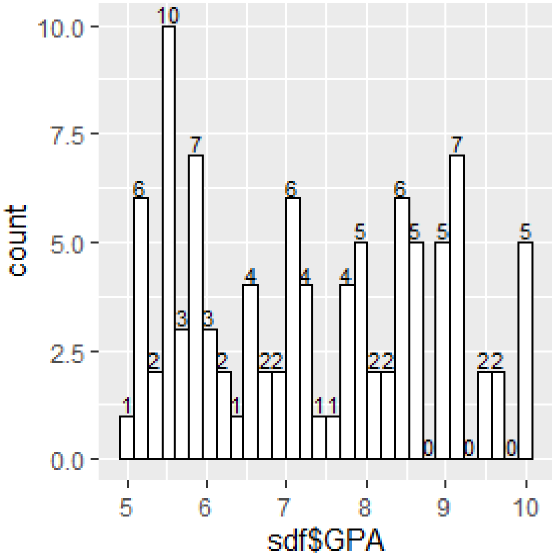
5. Identify the outlier’s using boxplot.

```
ggplot(sdf,mapping = aes(x="",y=sdf$GPA)) + geom_boxplot()+annotate("text", x
= 1, y = fivenum(sdf$GPA), label =
fivenum(sdf$GPA),vjust=0.3,color="darkblue",size=3)
```



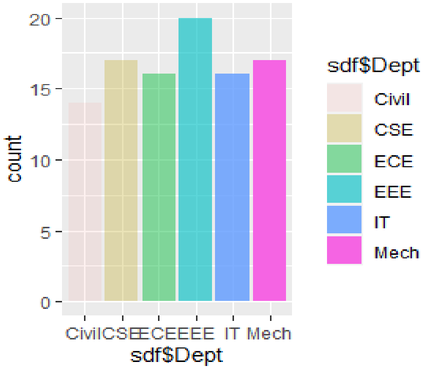
6. Draw histograms for count of GPA with different bin width & zoom to see any one region

```
ggplot(sdf,aes(x=sdf$GPA,bin_width=sdf$GPA))+geom_histogram(color="black",fil
l="white")+stat_bin(geom = "text", aes(label = ..count..),vjust=-0.2,size=3)
```



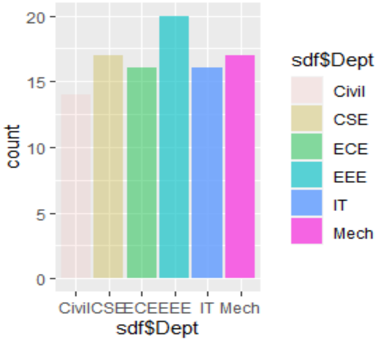
7. Transparency for every Department

```
ggplot(sdf,aes(x=sdf$Dept,fill=sdf$Dept)) + geom_bar(position="identity",
aes(alpha=sdf$Dept))
```



8. Number of students in each department with different colours each department

```
ggplot(sdf,aes(x=sdf$Dept,fill=sdf$Dept)) + geom_bar(position="identity",
aes(alpha=sdf$Dept))
```


count
Civil
CSE
ECE
EEE
IT
Mech
sdf$Dept
sdf$Dept
Civil
CSE
ECE
EEE
IT
Mech