Lab1

Priya Kurian(priku577) and Andreas christopoulos(andch552)

17 September 2018

Assignment 1

Made the required changes to the tree.pdf

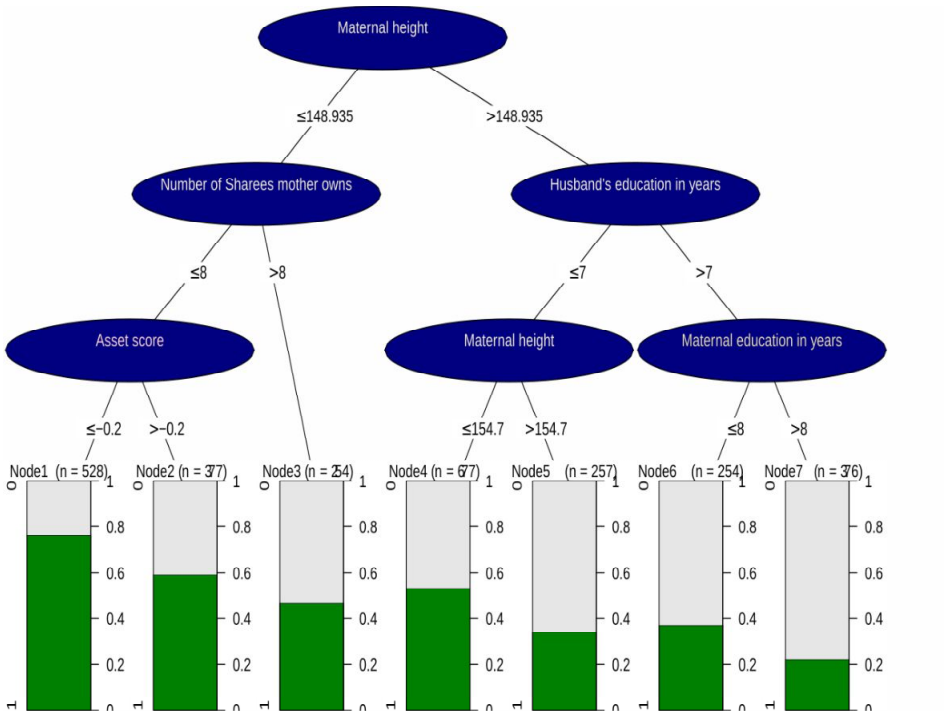


Figure 1. A tree from assignment 1.

Assignment 2

Read Data from file and load libraries

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────────────────────────────────────
──────────────────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 3.0.0     ✔ purrr   0.2.5
## ✔ tibble  1.4.2     ✔ dplyr   0.7.6
## ✔ tidyr   0.8.1     ✔ stringr 1.3.1
## ✔ readr   1.1.1     ✔ forcats 0.3.0
```

```
## ── Conflicts ────────────────────────────────────────────────────────────────────────────
──────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

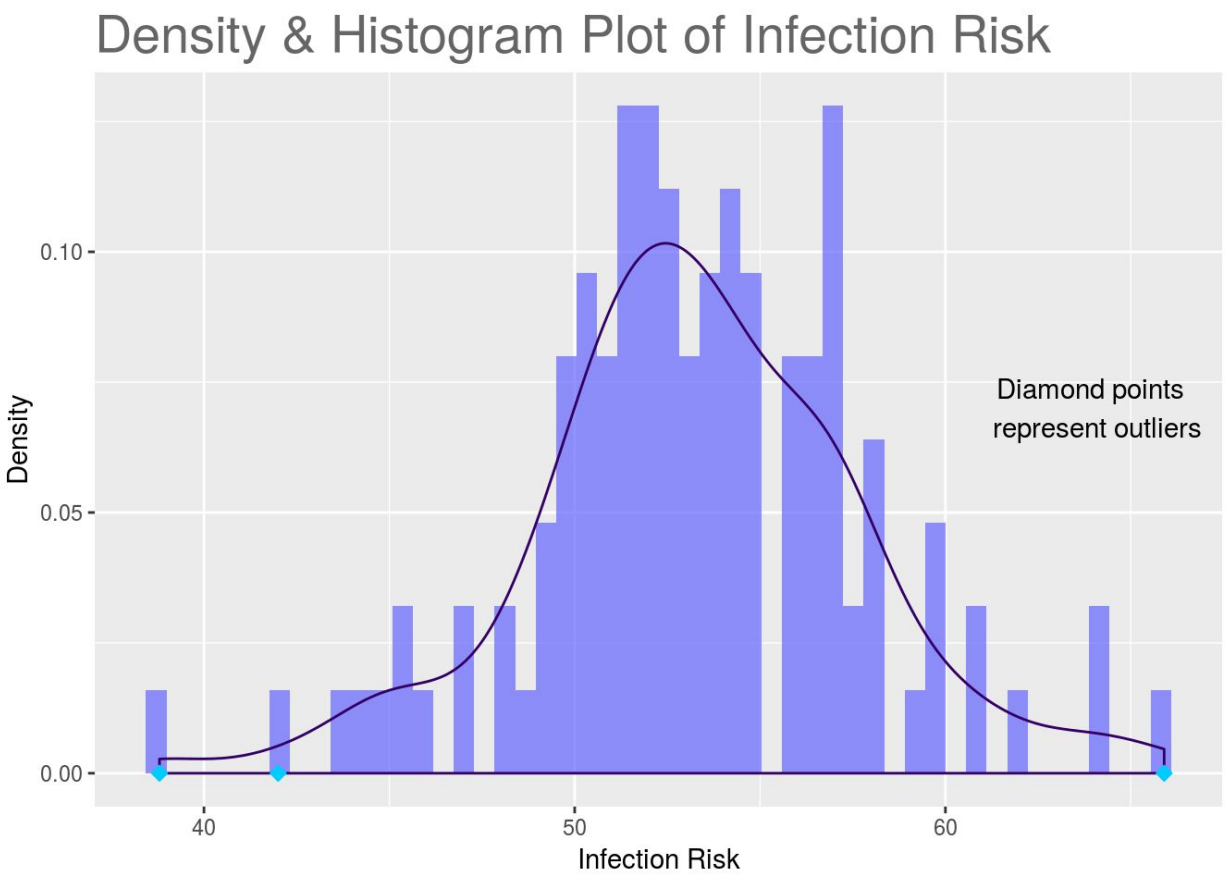
```
df<-read.table("SENIC.txt")
```

creating function to return indices

```
my_func<-function(x,name){
  col=x[[name]]
  quantile_1=quantile(col,0.25)
  quantile_3=quantile(col,0.75)
  l1=quantile_3+1.5*(quantile_3-quantile_1)
  l2=quantile_1-1.5*(quantile_3-quantile_1)
  indices=which(col>=l1 | col<=l2)

  return(indices)
}
```

Density Ploting of Infection Risk

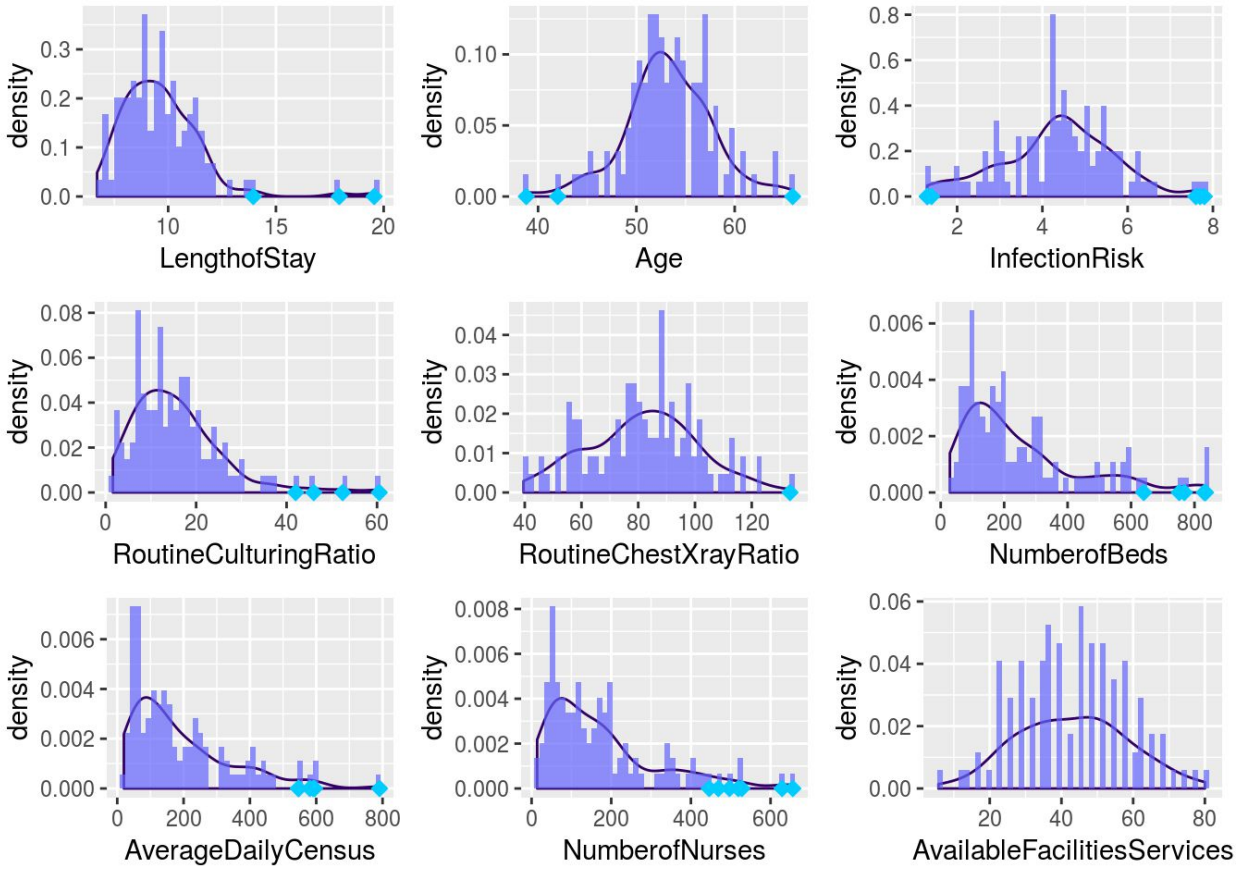
```
indices_infection<-my_func(df,"V3")
outliers_infection<-df$V3[indices_infection]
outliers <- tibble(x = outliers_infection, y = 0)

g<-ggplot(df,aes(x=V3))+geom_histogram(aes(y = ..density..),bins=50, alpha = 0.7,fill =
"#6666FF")+geom_density(col="#330066")
g<-g+geom_point(data = outliers, aes(x,y),size=2,colour="#00CCFF",pch=23,fill="#00CCFF")+
ggtitle("Density & Histogram Plot of Infection Risk")+labs(x="Infection Risk",y="Densit
y")
g<-g+theme(plot.title = element_text(color="#666666",size=21, hjust=0))
g+annotate(geom="text",x=64,y=0.07,label="Diamond points \n represent outliers")
```



outliers lie on either side of the graph probably to the ends of the graphs.

Producce graphs for all other variables

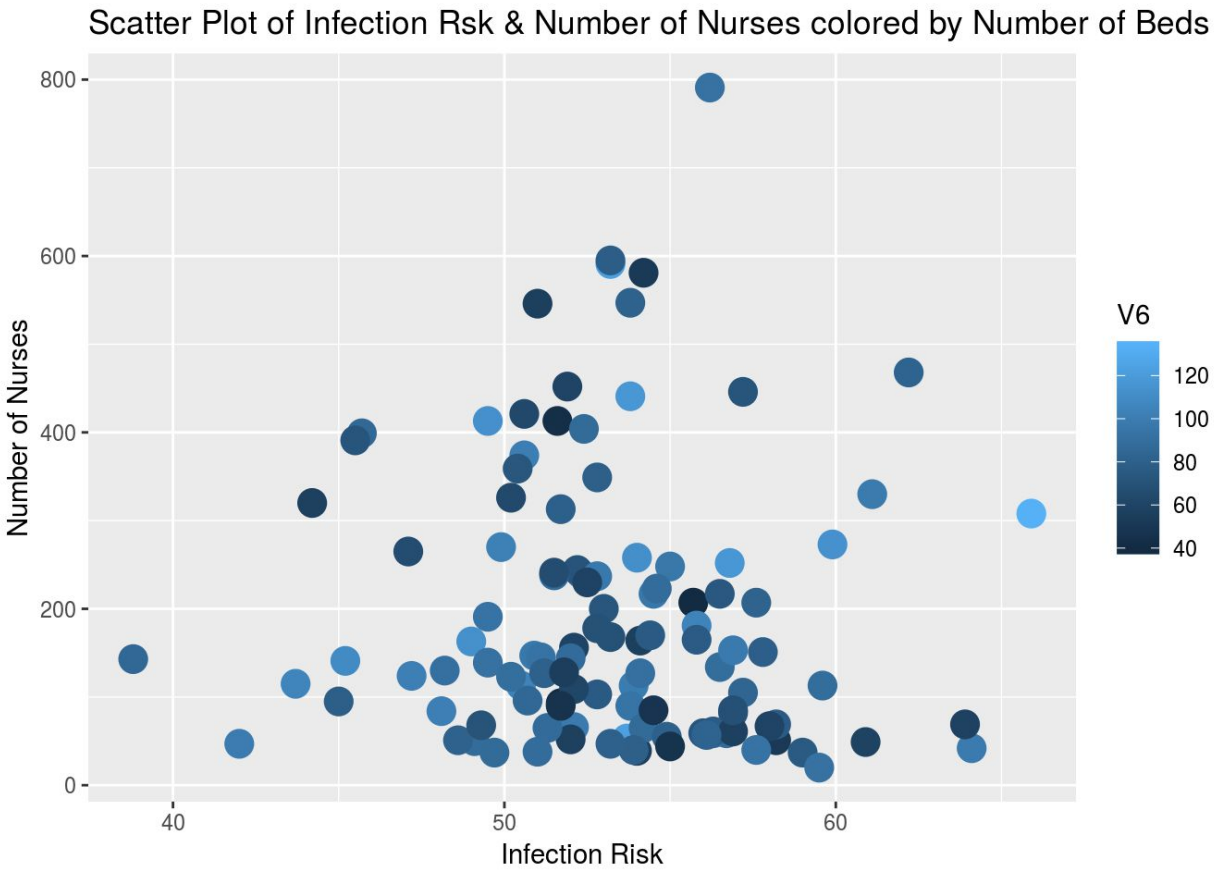
```
dt<-df[c(-1,-8,-9)]
names(dt)<-c('LengthofStay','Age','InfectionRisk','RoutineCulturingRatio',
                    'RoutineChestXrayRatio','NumberofBeds','AverageDailyCensus',
                    'NumberofNurses','AvailableFacilitiesServices')

myplots<-list()

for (name in names(dt))
{

  indices<-my_func(dt,name)
  outliers<-dt[[name]][indices]
  outliers_names<-tibble(x=outliers,y=0)
  myplots[[name]]<-ggplot(dt, aes_string(x =name)) + geom_density(col="#330066")+geom_his
togram(aes(y=..density..),bins=50, alpha = 0.7,fill = "#6666FF")+geom_point(data=outliers
_names,aes(x,y),size=2,color="#00CCFF",pch=23,fill="#00CCFF")
}
grid.arrange(grobs=myplots)
```



We could see that all graphs exept AvailableFacilitiesServices have outliers. We feel that the graphs of Age,InfectionRisk,RoutineChestXrayRatio,AvailableFacilitiesServices follow a normal distribution. Also the LengthofStay,RoutineCulturingRatio,NumberofBeds,AverageDailyCensus,NumberofNurses follow a chi-square distribution.

Scatter Plot

```
ggplot(df,aes(y=V10,x=V3,col=V6))+geom_point(size=5)+ggtitle("Scatter Plot of Infection R
sk & Number of Nurses colored by Number of Beds")+labs(x="Infection Risk",y="Number of Nu
rses")
```



From the graph we can see that the as the number of nurses increase the infection risk decreases.Also as the number of beds are less when compared to the people(overcrowded) the infection risk will be higher.

Plotly Graph

```
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

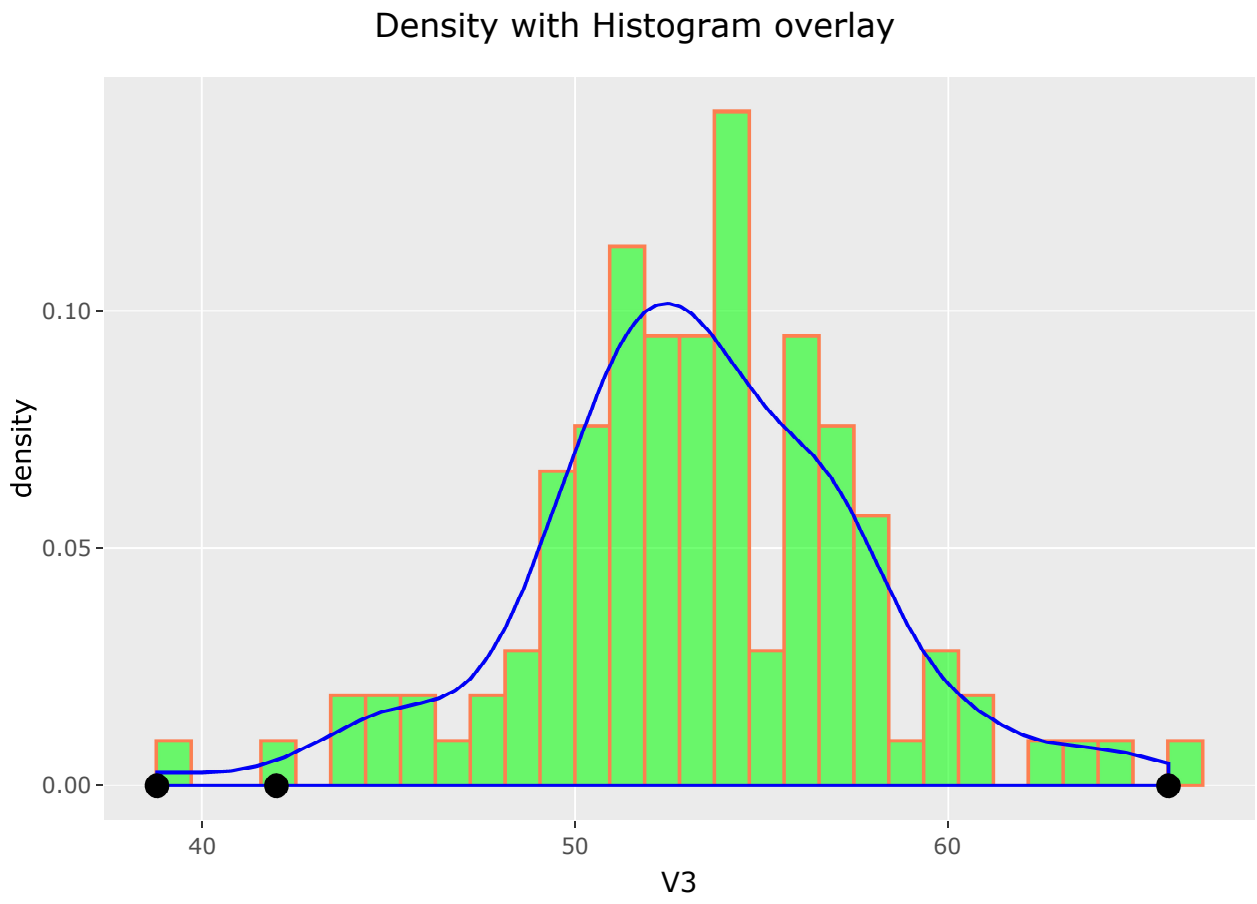
```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
indices_infection<-my_func(df,"V3")
outliers_infection<-df$V3[indices_infection]
outliers <- tibble(x = outliers_infection, y = 0)

plot<-ggplot(df,aes(V3))+geom_histogram(aes(y=..density..,alpha=0.7),col="coral",fill="gr
een")+geom_density(col="blue")+ggtitle("Density with Histogram overlay")+geom_point(data
= outliers, aes(x,y),size=3)
p<-ggplotly(plot)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

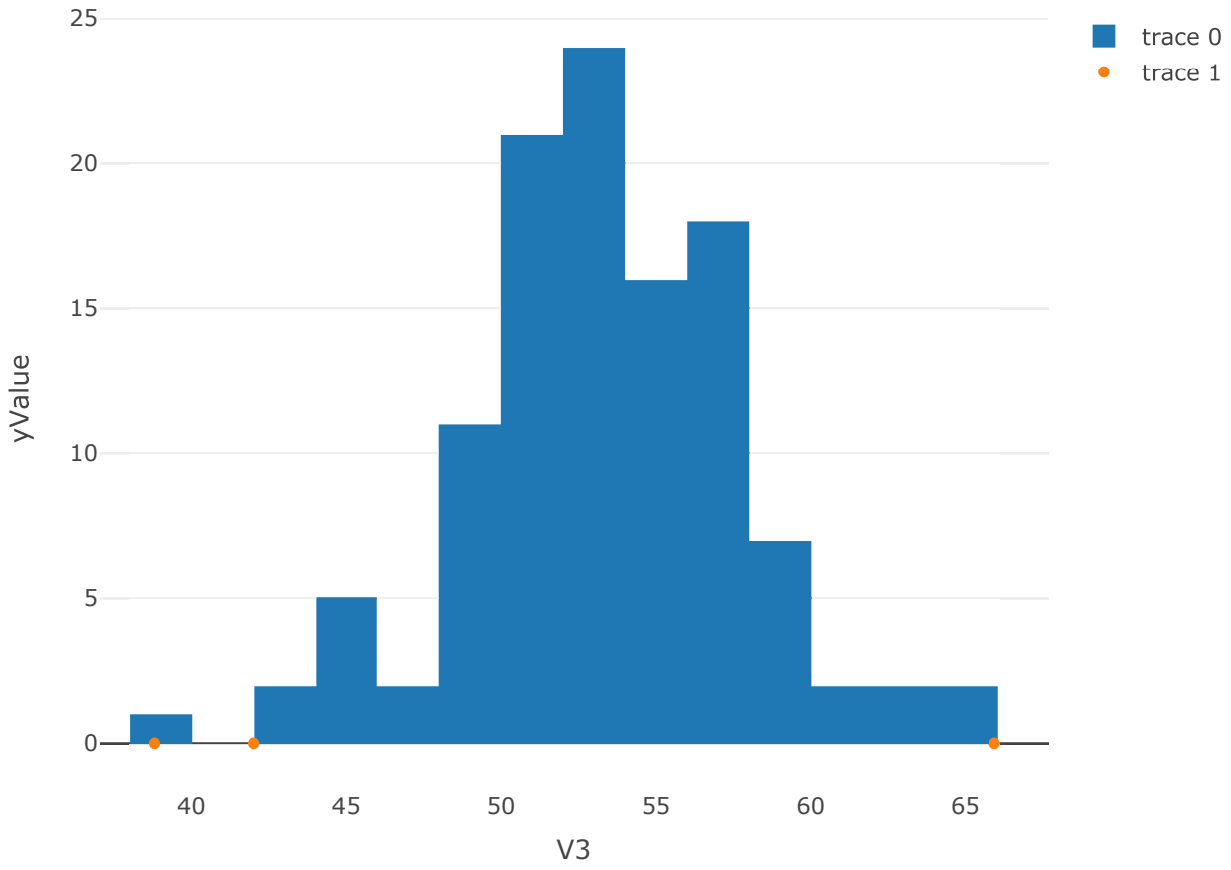
```
p
```



THe plotly has features like zoom in,zoom out,reset the axis,autoscaling,box slection,lassso selection etc.As a whole we feel like there are lot more that we can do with plotly graphs when compared to ggplot.

Plolty Plot Made with Pipe operator

```
Outlier_indices <- my_func(df,"V3")
Outlier_values<-df$V3[Outlier_indices]
yValue <- rep(0,length(Outlier_values))

hisPlot <- df %>% select(V3) %>% plot_ly(x=~V3,type="histogram")  %>%
  add_markers(x=~Outlier_values, y=~yValue)

hisPlot
```



Shiny App

```
library(shiny)
library(tidyverse)

df<-read.table("SENIC.txt")
dt<-df[c(-1,-8,-9)]
plot=list()


ui <- fluidPage(

  titlePanel("Density and Histogram Plot"),

  sidebarLayout(

    sidebarPanel(
      sliderInput(inputId = "bins",
                  label="Number of bins:",
                  min = 1,
                  max = 100,
                  value = 30),
      checkboxGroupInput("var",label= "Please Select Variable",
                  choices = c("Length of Stay"="V2", "Age"="V3", "Infection Risk"="V4","R
outine Culturing Ratio"="V5",
                              "Routine Chest X ray Ratio"="V6"," Number of Beds"="V7",
                              "Average Daily Censu"="V10","Number of Nurses"="V11","Avail
able Facilities Services"="V12")
    )),
    mainPanel(
      plotOutput("densPlot")
    )
  )
)

server <- function(input, output) {

  output$densPlot <- renderPlot({

      ggplot(dt, aes_string(x=input$var))+geom_density()+geom_histogram(aes(y = ..densit
y..),binwidth =input$bins, alpha = 0.7,fill = "#6666FF",col="red")

  })


}

# Run the application
shinyApp(ui = ui, server = server)
```

From the graphs we can see that the binwidth deoends on the scale of the X axis.If the x axis has higher range of value,then the binwidth should be more.

```
library(tidyverse)
library(ggplot2)
library(gridExtra)
df<-read.table("SENIC.txt")
my_func<-function(x,name){
  col=x[[name]]
  quantile_1=quantile(col,0.25)
  quantile_3=quantile(col,0.75)
  l1=quantile_3+1.5*(quantile_3-quantile_1)
  l2=quantile_1-1.5*(quantile_3-quantile_1)
  indices=which(col>=l1 | col<=l2)

  return(indices)
}
indices_infection<-my_func(df,"V3")
outliers_infection<-df$V3[indices_infection]
outliers <- tibble(x = outliers_infection, y = 0)

g<-ggplot(df,aes(x=V3))+geom_histogram(aes(y = ..density..),bins=50, alpha = 0.7,fill =
"#6666FF")+geom_density(col="#330066")
g<-g+geom_point(data = outliers, aes(x,y),size=2,colour="#00CCFF",pch=23,fill="#00CCFF")+
ggtitle("Density & Histogram Plot of Infection Risk")+labs(x="Infection Risk",y="Densit
y")
g<-g+theme(plot.title = element_text(color="#666666",size=21, hjust=0))
g+annotate(geom="text",x=64,y=0.07,label="Diamond points \n represent outliers")

dt<-df[c(-1,-8,-9)]
names(dt)<-c('LengthofStay','Age','InfectionRisk','RoutineCulturingRatio',
                    'RoutineChestXrayRatio','NumberofBeds','AverageDailyCensus',
                    'NumberofNurses','AvailableFacilitiesServices')

myplots<-list()

for (name in names(dt))
{

  indices<-my_func(dt,name)
  outliers<-dt[[name]][indices]
  outliers_names<-tibble(x=outliers,y=0)
  myplots[[name]]<-ggplot(dt, aes_string(x =name)) + geom_density(col="#330066")+geom_his
togram(aes(y=..density..),bins=50, alpha = 0.7,fill = "#6666FF")+geom_point(data=outliers
_names,aes(x,y),size=2,color="#00CCFF",pch=23,fill="#00CCFF")
}
grid.arrange(grobs=myplots)

ggplot(df,aes(y=V10,x=V3,col=V6))+geom_point(size=5)+ggtitle("Scatter Plot of Infection R
sk & Number of Nurses colored by Number of Beds")+labs(x="Infection Risk",y="Number of Nu
rses")



library(plotly)
indices_infection<-my_func(df,"V3")
outliers_infection<-df$V3[indices_infection]
outliers <- tibble(x = outliers_infection, y = 0)
```

```
plot<-ggplot(df,aes(V3))+geom_histogram(aes(y=..density..,alpha=0.7),col="coral",fill="gr
een")+geom_density(col="blue")+ggtitle("Density with Histogram overlay")+geom_point(data
= outliers, aes(x,y),size=3)
p<-ggplotly(plot)

p


Outlier_indices <- my_func(df,"V3")
Outlier_values<-df$V3[Outlier_indices]
yValue <- rep(0,length(Outlier_values))

hisPlot <- df %>% select(V3) %>% plot_ly(x=~V3,type="histogram")  %>%
  add_markers(x=~Outlier_values, y=~yValue)

hisPlot



library(shiny)
library(tidyverse)

df<-read.table("SENIC.txt")
dt<-df[c(-1,-8,-9)]
plot=list()


ui <- fluidPage(

  titlePanel("Density and Histogram Plot"),

  sidebarLayout(

    sidebarPanel(
      sliderInput(inputId = "bins",
                  label="Number of bins:",
                  min = 1,
                  max = 100,
                  value = 30),
      checkboxGroupInput("var",label= "Please Select Variable",
                  choices = c("Length of Stay"="V2", "Age"="V3", "Infection Risk"="V4","R
outine Culturing Ratio"="V5",
                              "Routine Chest X ray Ratio"="V6"," Number of Beds"="V7",
                              "Average Daily Censu"="V10","Number of Nurses"="V11","Avail
able Facilities Services"="V12")
    )),
    mainPanel(
      plotOutput("densPlot")
    )
  )
)

server <- function(input, output) {

  output$densPlot <- renderPlot({
```

```
      ggplot(dt, aes_string(x=input$var))+geom_density()+geom_histogram(aes(y = ..densit
y..),binwidth =input$bins, alpha = 0.7,fill = "#6666FF",col="red")

  })


}

# Run the application
shinyApp(ui = ui, server = server)
```