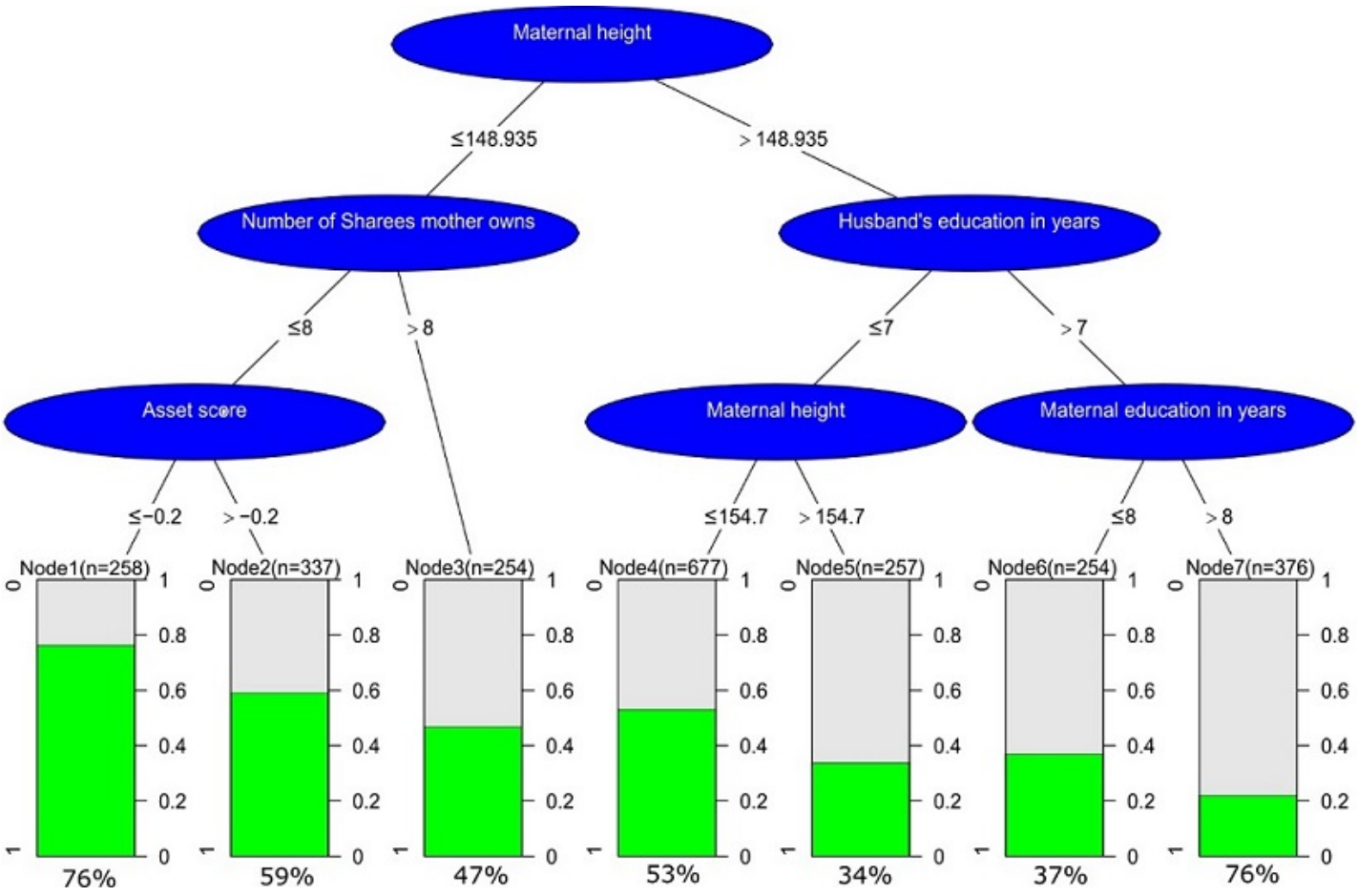
LAB 1

Group 22 - balra340 , tejma768

September 16, 2018

Assignment 1

Use Inkscape to produce a publication quality graph



Tree

Importing tree.pdf file using Inkscape resulted in the following Tree which is much better to view and is of publication quality.

Assignment 2

Question 1

Read data from SENIC.txt into R.

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

colnames(my_data) <- c("ID","STAY","AGE","RISK","CULTURE_RATIO","CHEST", "BEDS","AFFILIATION","REGION","CENSUS","NURSES","FS")
```

Question 2.1

Create a function that for a given column (vector) X computes first and third quantiles Q1 and Q3 with quantiles()

```
my_function <- function(x){

  Q1 <- quantile(x,0.25)
  Q3 <- quantile(x,0.75)
}
```

Question 2.2

Create a function that for a given column (vector) X returns indices of outlying observations, i.e. observation with X-values greater than Q3+1.5(Q3-Q1) or less than Q1-1.5(Q3-Q1).

```
my_function <- function(x){

  Q1 <- quantile(x,0.25)
  Q3 <- quantile(x,0.75)

  up <- Q3 + 1.5*(Q3 - Q1)
  down <- Q1 - 1.5*(Q3 - Q1)

  result1 <- which(x > up) #which function is used to check if x value is greater than up

  result2 <- which(x < down) #which function is used to check if x value is lesser than down

  final <- sort(c(result1,result2))

  return(final)

}
```

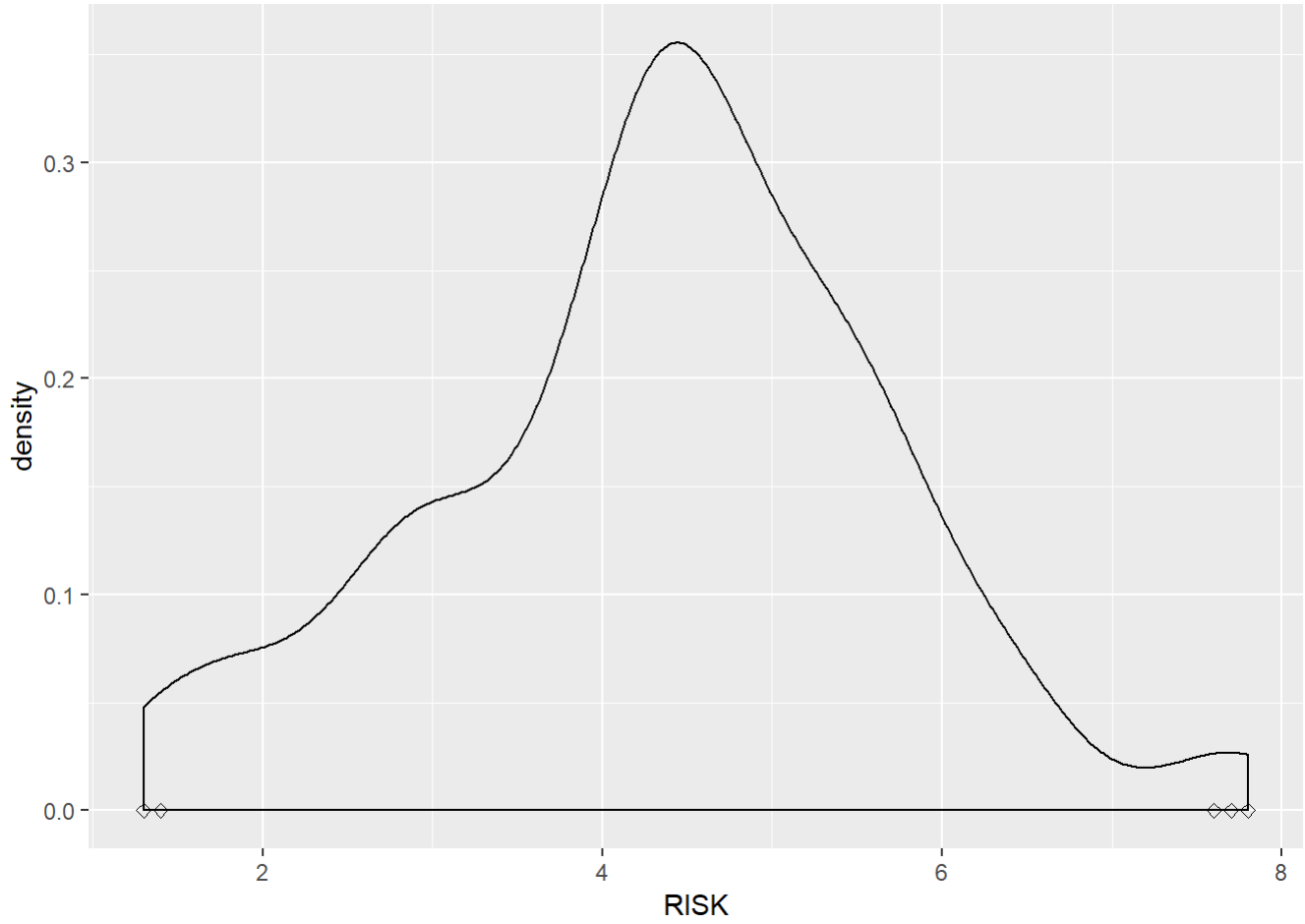
Question 3

Use ggplot2 and the function from step 2 to create a density plot of Infection risk in which outliers are plotted as a diamond symbol (???) . Make some analysis of this graph.

```
library(ggplot2)

output <- my_function(my_data$RISK)

output
```

```
## [1]  13  40  53  54  93 107
```

```
my_data2 <- my_data[output,] #creating a second data set


final <- ggplot() + geom_density(data = my_data, aes(x = RISK )) + geom_point(data = my_data2 ,a
es(x = RISK , y = 0 ,shape = 5)) +
         scale_shape_identity()

final
```



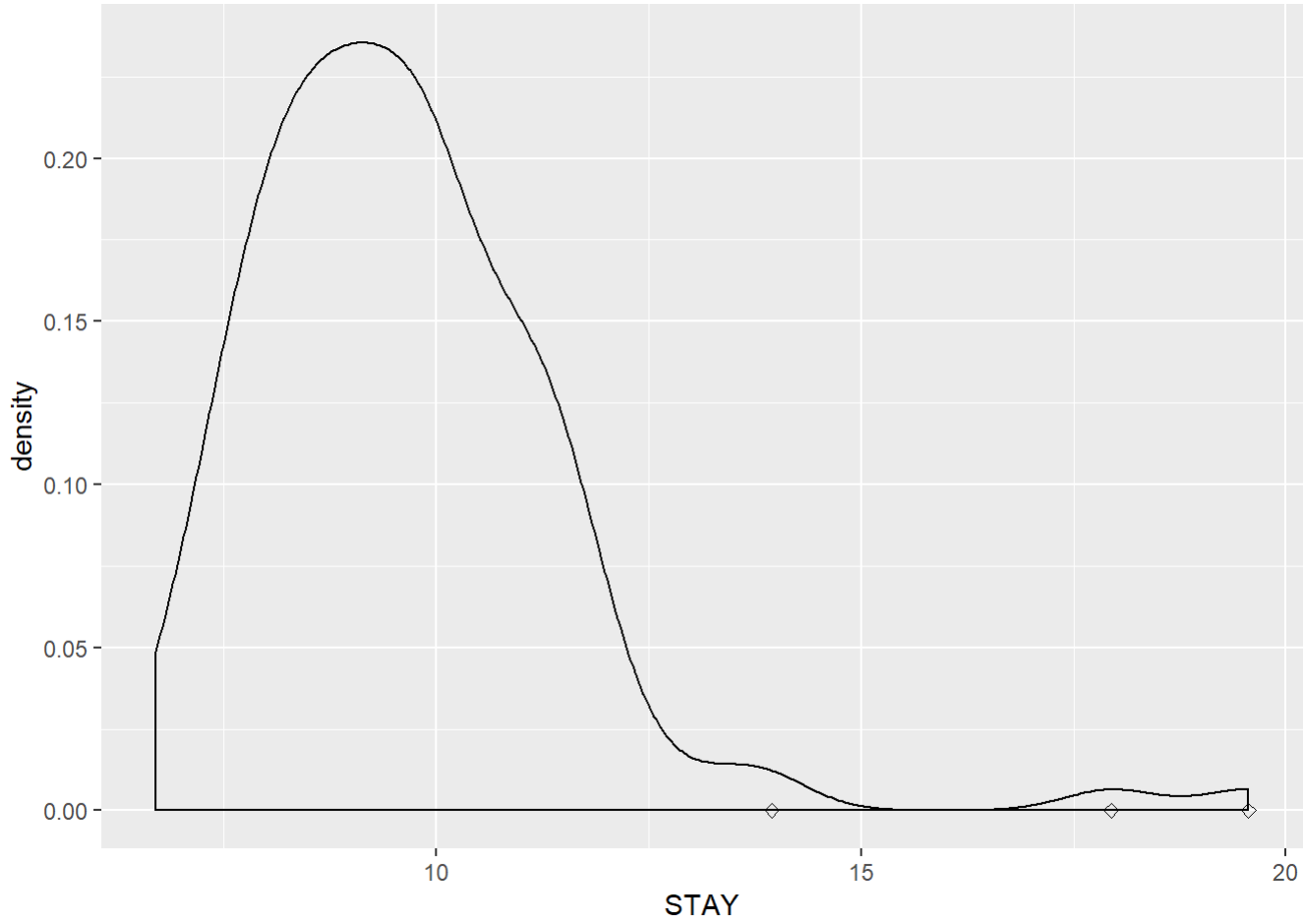
There are 5 outliers. We could infer from the graph the risk values of 2 to 7 is the most common in the data set.

Question 4

Produce graphs for all other quantitative variables in the data. Put these graphs into one (hint: arrangeGrob() in gridExtra package can be used) and make some analysis.

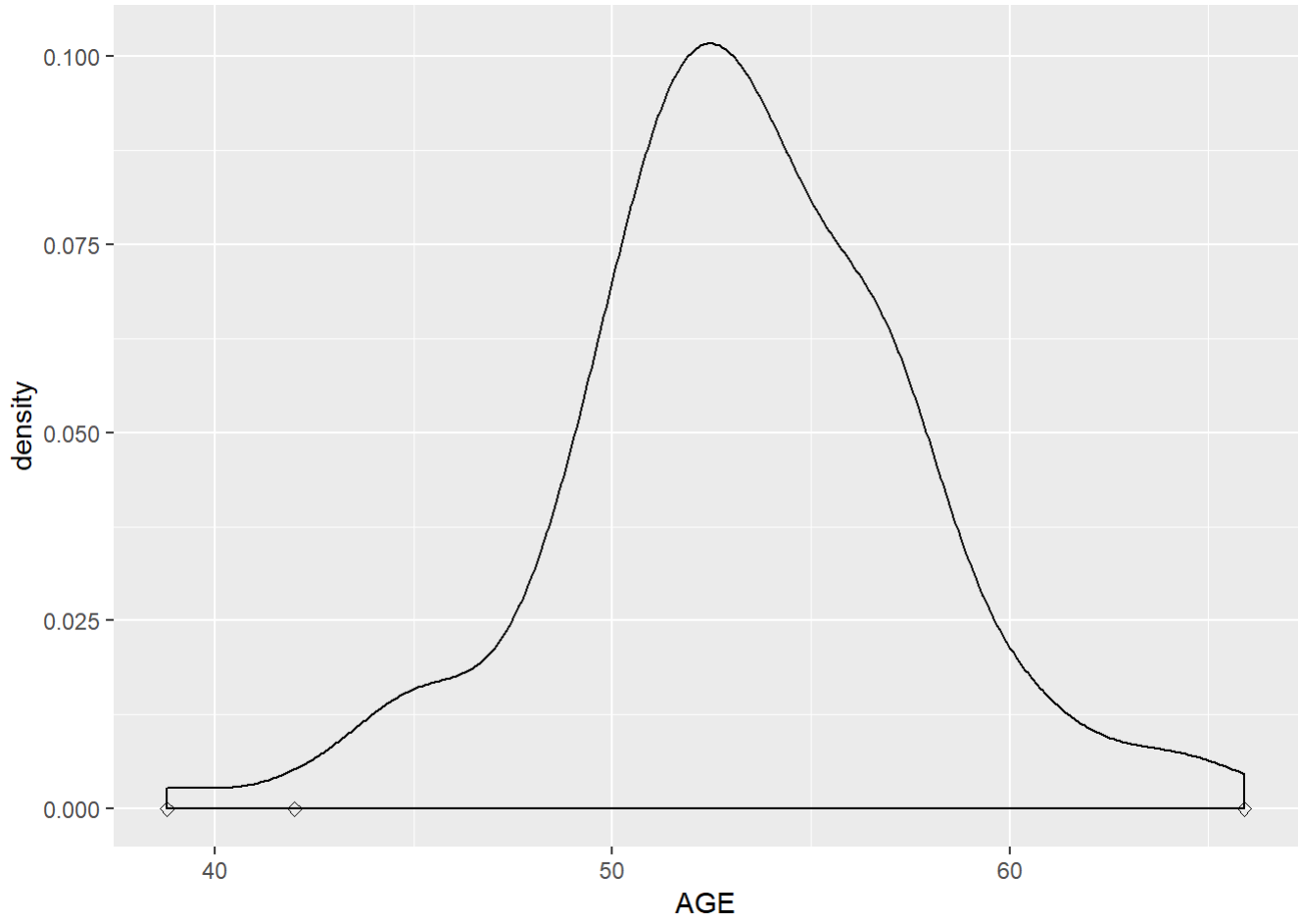
Stay

```
output_stay <- my_function(my_data$STAY)

my_data_stay <- my_data[output_stay,] #creating a second data set


stay <- ggplot() + geom_density(data = my_data, aes(x = STAY )) + geom_point(data = my_data_stay
 ,aes(x = STAY , y = 0 ,shape = 5))+
        scale_shape_identity()
stay
```



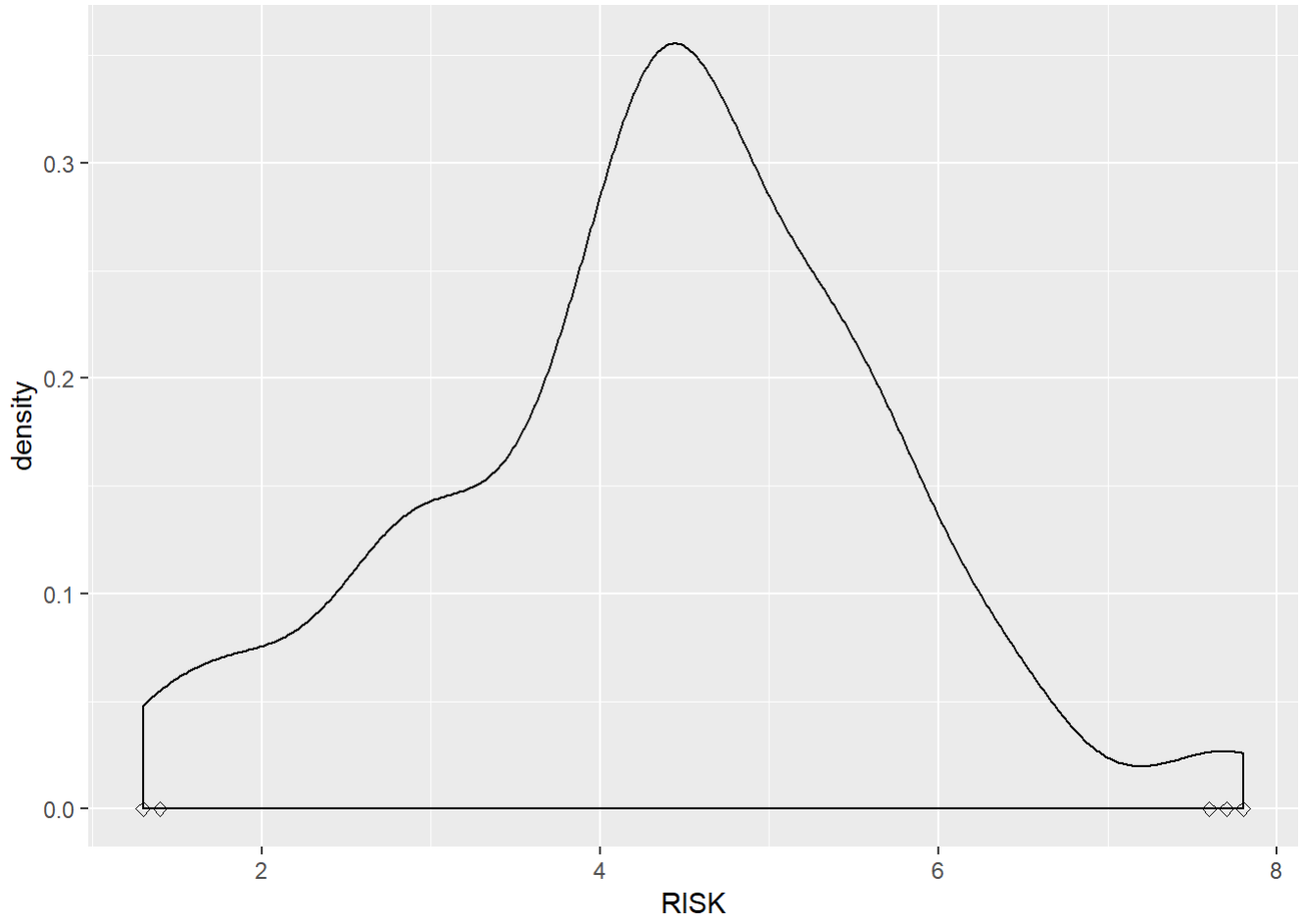
Age

```
output_age <- my_function(my_data$AGE)

my_data_age <- my_data[output_age,] #creating a second data set


age <- ggplot() + geom_density(data = my_data, aes(x = AGE )) + geom_point(data = my_data_age ,a
es(x = AGE , y = 0 ,shape = 5)) +
       scale_shape_identity()
age
```



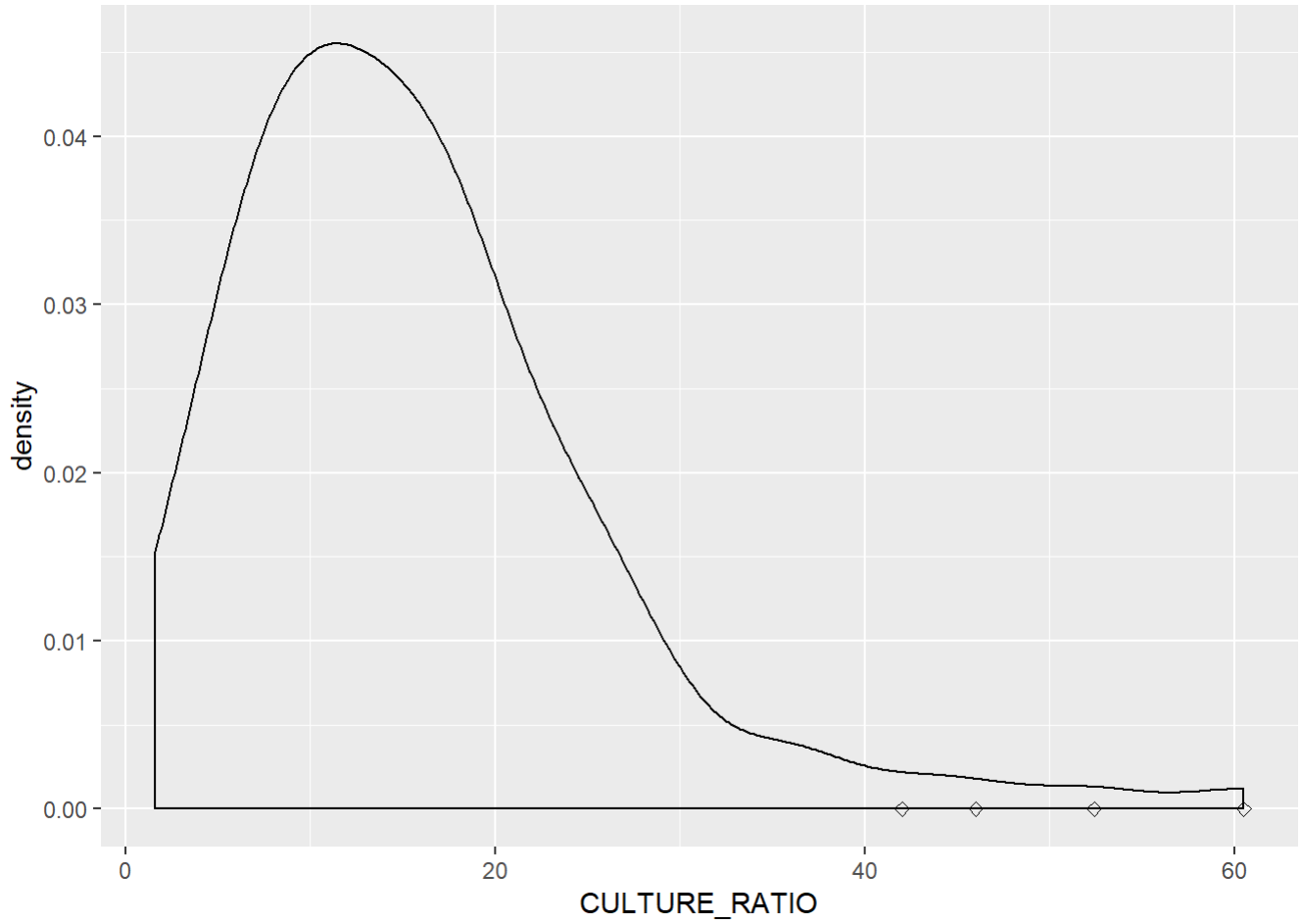
Risk

```
output_risk <- my_function(my_data$RISK)

my_data_risk <- my_data[output_risk,] #creating a second data set


risk <- ggplot() + geom_density(data = my_data, aes(x = RISK )) + geom_point(data = my_data_risk
 ,aes(x = RISK , y = 0 ,shape = 5))+         scale_shape_identity()
risk
```



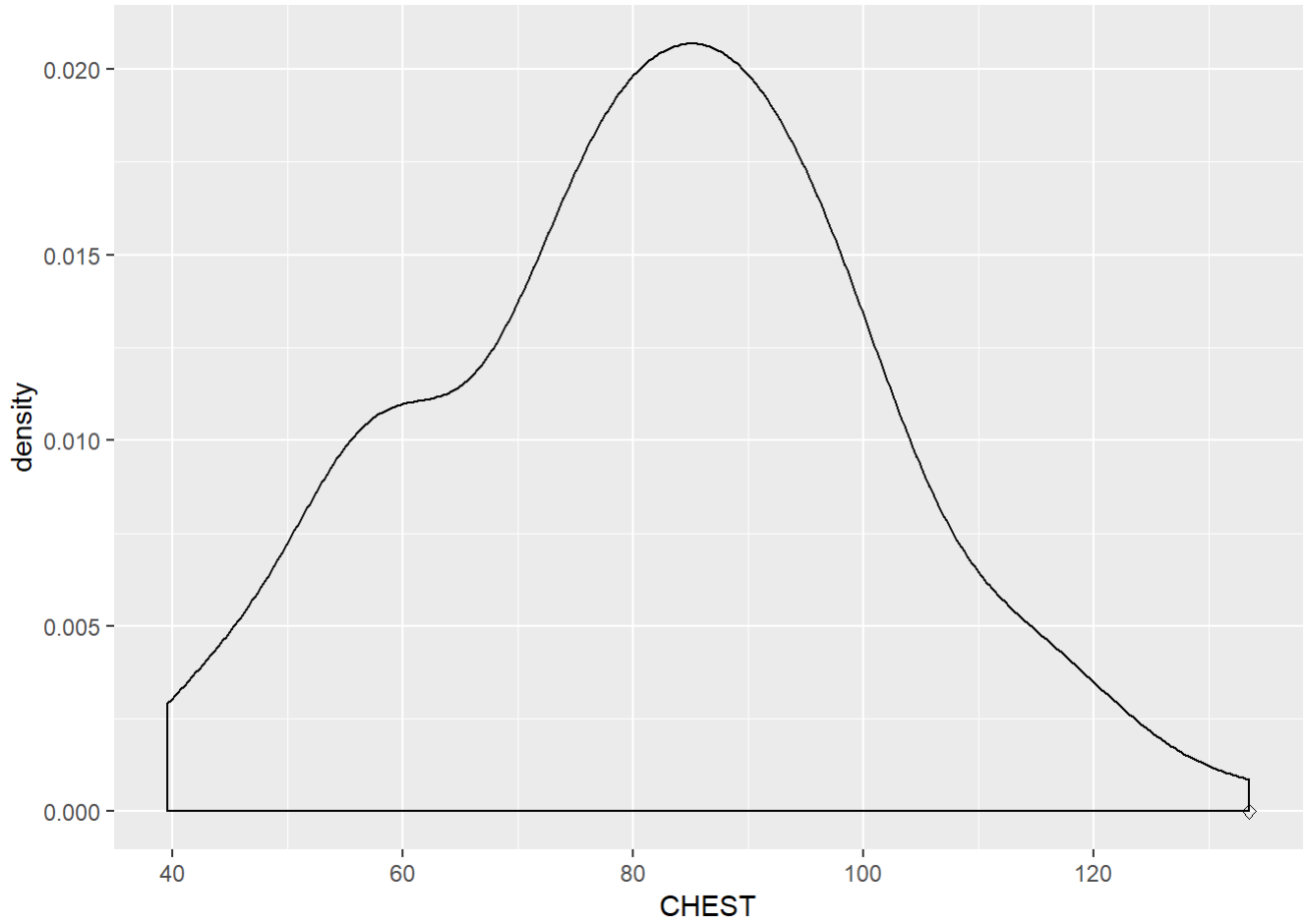
Culture Ratio

```
output_culture <- my_function(my_data$CULTURE_RATIO)

my_data_culture <- my_data[output_culture,] #creating a second data set


culture <- ggplot() + geom_density(data = my_data, aes(x = CULTURE_RATIO )) +
           geom_point(data = my_data_culture ,aes(x = CULTURE_RATIO , y = 0 ,shape = 5)) + scale
_shape_identity()

culture
```



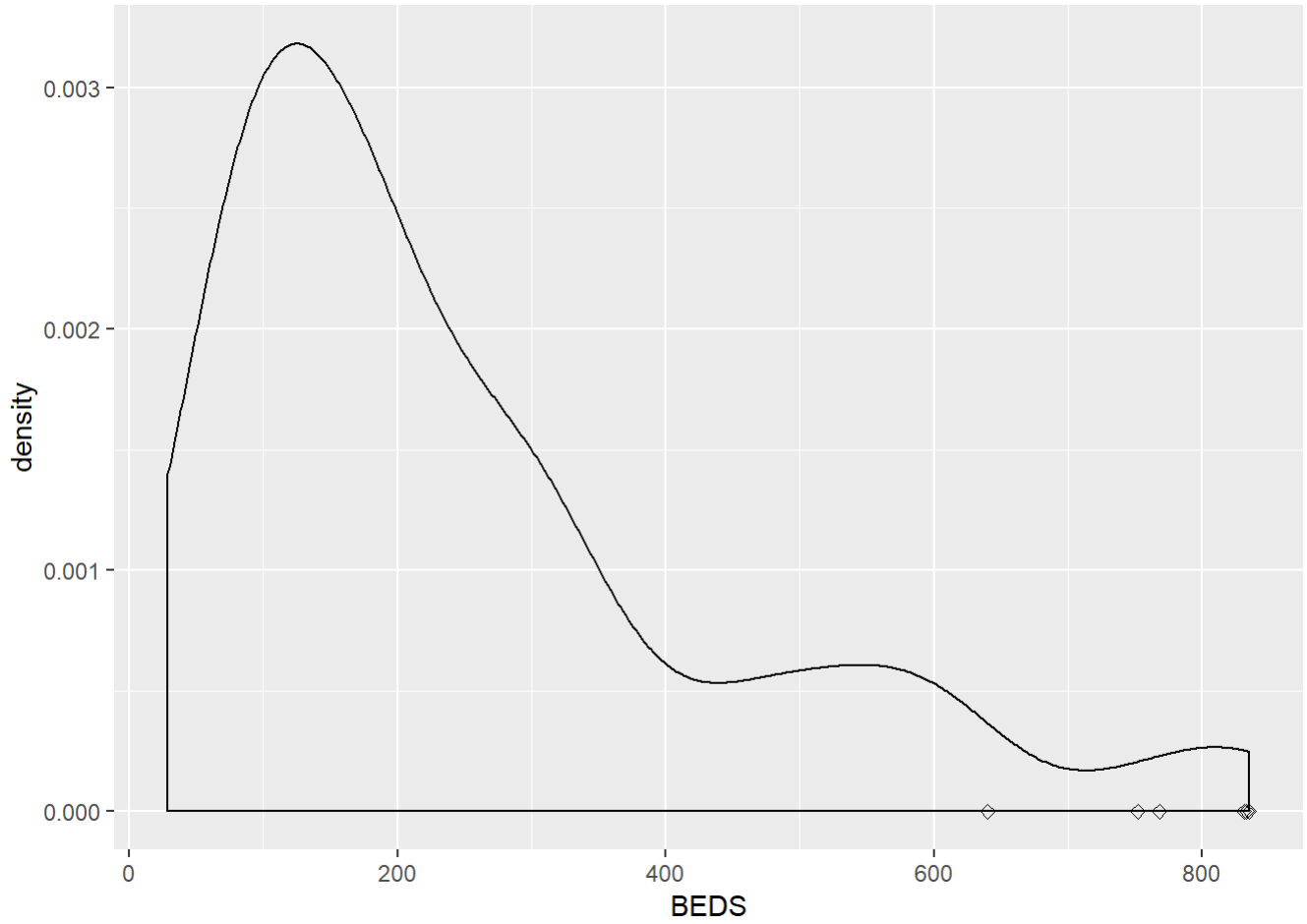
Chest

```
output_chest <- my_function(my_data$CHEST)

my_data_chest <- my_data[output_chest,] #creating a second data set


chest <- ggplot() + geom_density(data = my_data, aes(x = CHEST )) +
        geom_point(data = my_data_chest ,aes(x = CHEST , y = 0 ,shape = 5)) + scale_shape_ident
ity()

chest
```



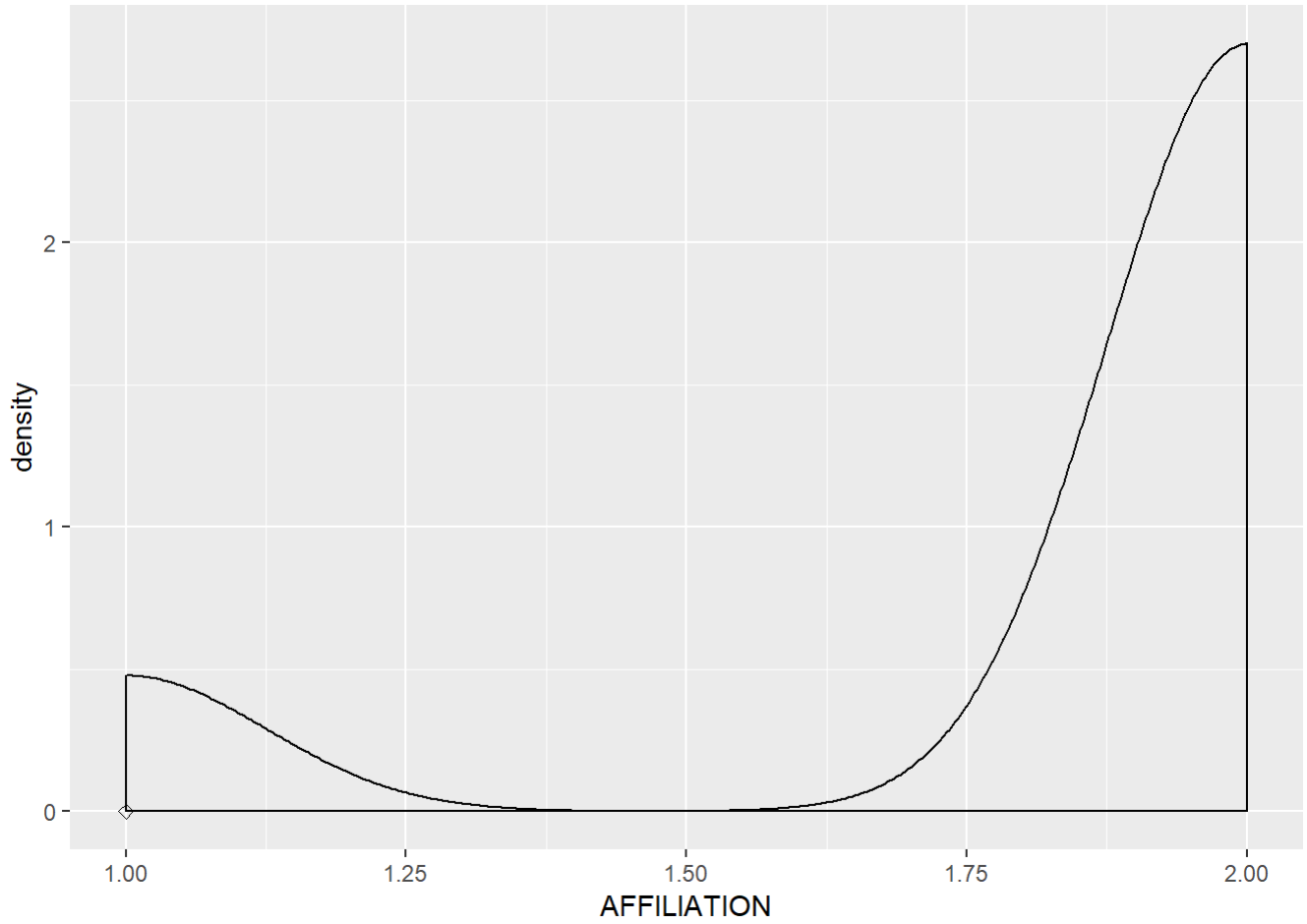
Beds

```
output_bed <- my_function(my_data$BEDS)

my_data_bed <- my_data[output_bed,] #creating a second data set


bed <- ggplot() + geom_density(data = my_data, aes(x = BEDS )) +
       geom_point(data = my_data_bed ,aes(x = BEDS , y = 0 ,shape = 5)) + scale_shape_identity()
bed
```



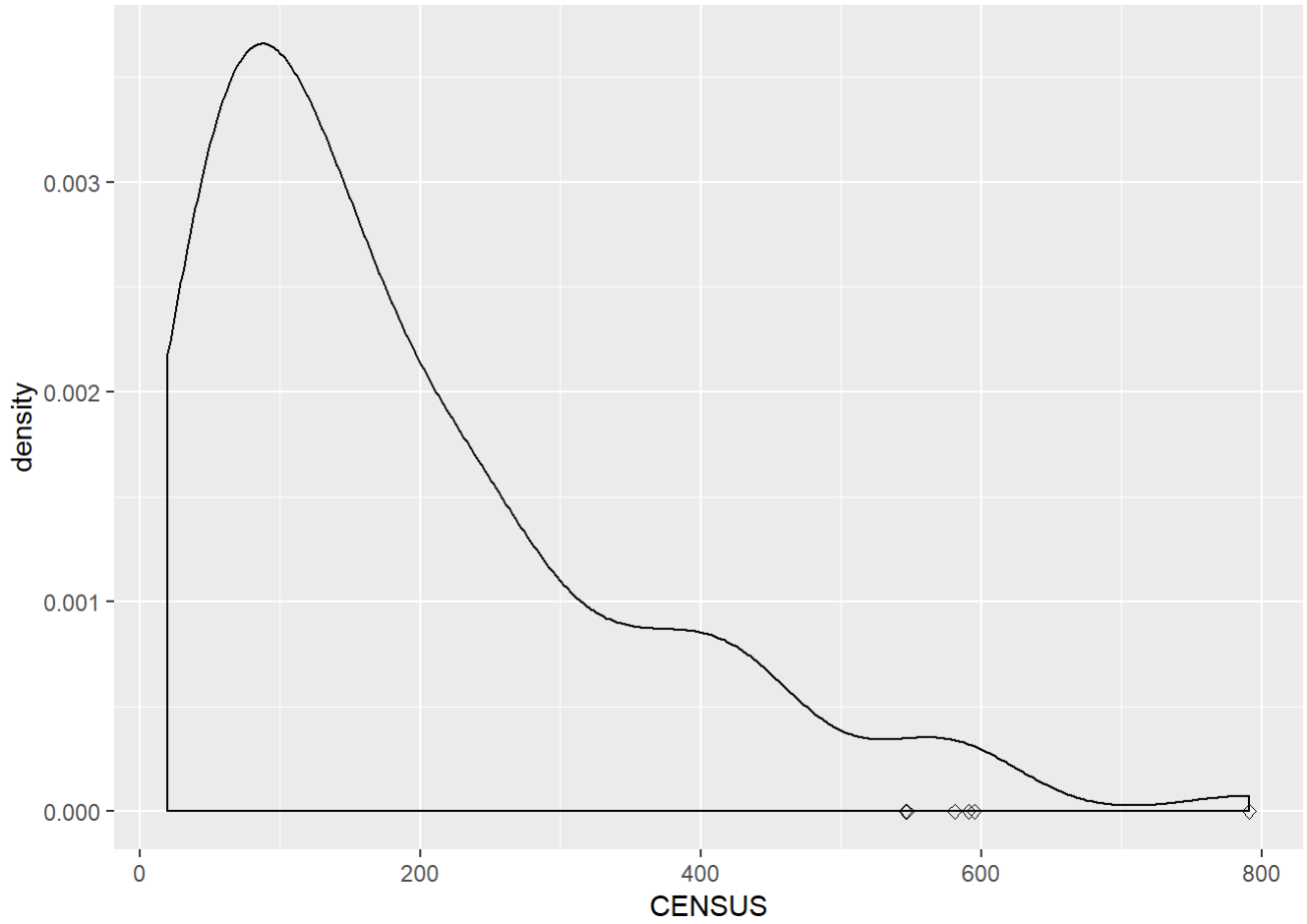
Affiliation

```
output_affliation<- my_function(my_data$AFFILIATION)

my_data_affiliation <- my_data[output_affliation,] #creating a second data set


affiliation <- ggplot() + geom_density(data = my_data, aes(x = AFFILIATION )) +
               geom_point(data = my_data_affiliation ,aes(x = AFFILIATION , y = 0 ,shape = 5)) +
 scale_shape_identity()
affiliation
```



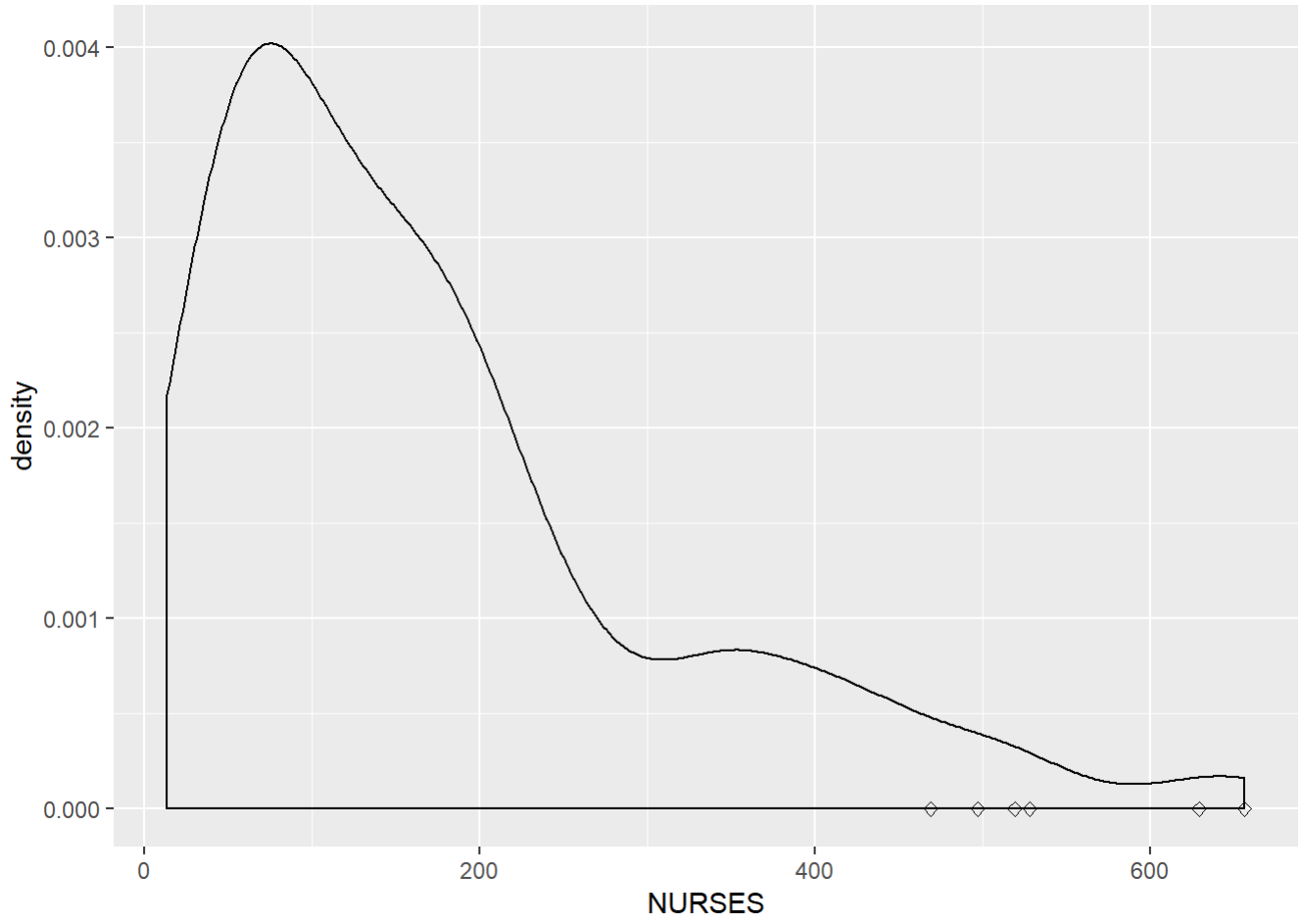
Census

```
output_census <- my_function(my_data$CENSUS)

my_data_census <- my_data[output_census,] #creating a second data set


census <- ggplot() + geom_density(data = my_data, aes(x = CENSUS )) +
          geom_point(data = my_data_census ,aes(x = CENSUS , y = 0 ,shape = 5)) + scale_shape_id
entity()
census
```

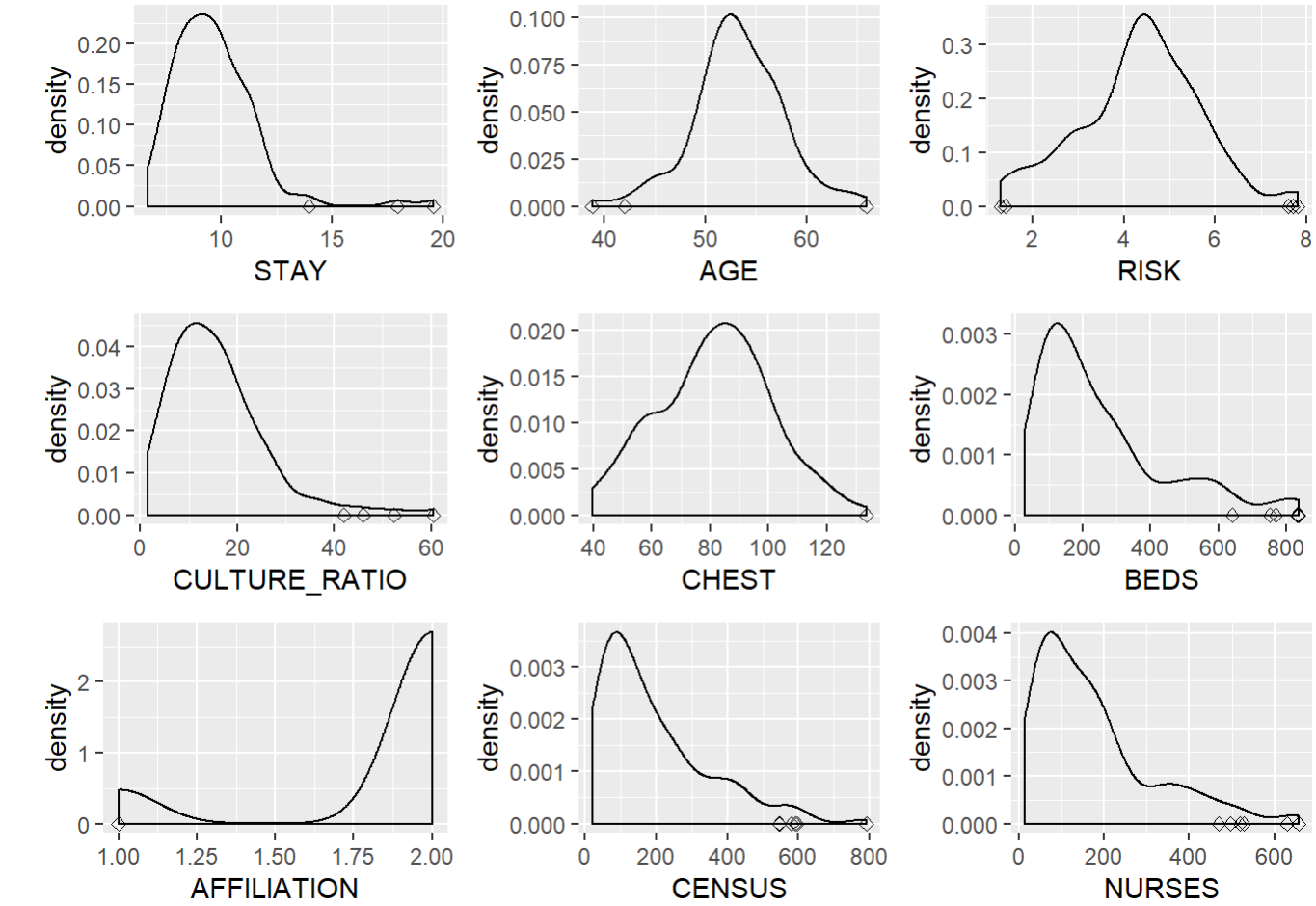


Nurses

```
output_nurse <- my_function(my_data$NURSES)

my_data_nurse <- my_data[output_nurse,] #creating a second data set

nurse <- ggplot() + geom_density(data = my_data, aes(x = NURSES )) +
         geom_point(data = my_data_nurse ,aes(x = NURSES , y = 0 ,shape = 5)) + scale_shape_iden
tity()
nurse
```




```
library(gridExtra)
combined <- arrangeGrob(stay,age,risk,culture,chest,bed,affiliation,census,nurse)

library(ggpubr)

as_ggplot(combined)
```



We can see most of the graphs are skewed right which means higher values are where outliers are present.

From the density plot of stay we get to know that very less people stay beyond 15 days. And that is shown in graph with two outliers beyond 15.

The age density plot helps us understand most observations were of age between 50 to 55.

The risk density plot shows that there is a 50 - 50 chance of getting an infection during the stay at the hospital.

Culture ratio density plot infers culture test wasn't performed for more than 50% of patients only in rare cases. And all the outliers are beyond 40 %.

Chest xray density plot shows almost everyone got a chest xray.

Density plot of no of beds show that most of the times the no of beds used were less than 400. However in rare cases there 600+ beds. This is clearly shown in thr graph with the outliers.

Next plot shows most of the hospitals didnot have a medical school affiliation.

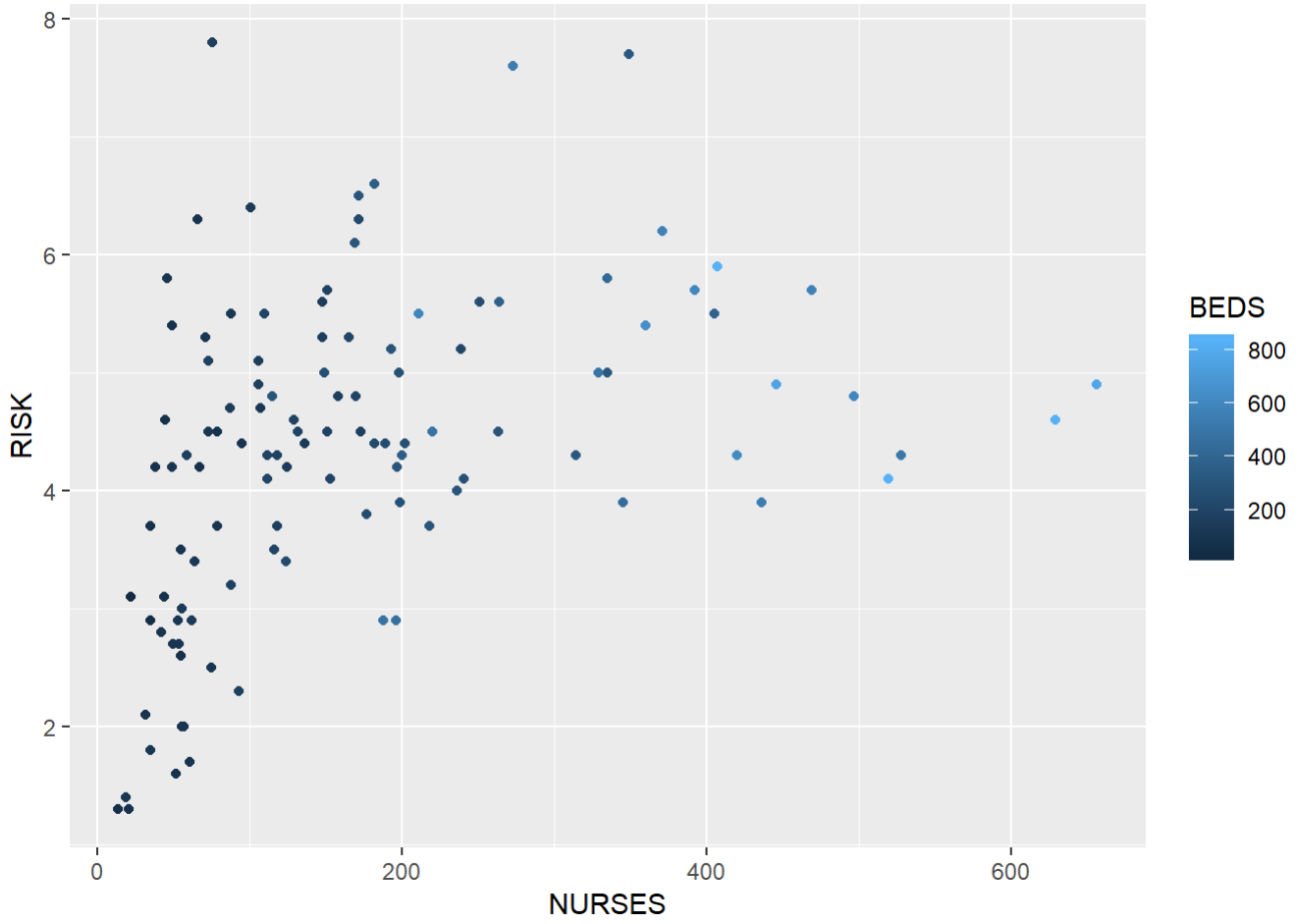
The census plot clearly supports the number of beds plots and they almost look similar.

From the last plot we can understand most of the times, each patient was given a nurse for personal attention.

Question 5

Create a ggplot2 scatter plot showing the dependence of Infection risk on the Number of Nurses where the points are colored by Number of Beds. Is there any interesting information in this plot that was not visible in the plots in step 4? What do you think is a possible danger of having such a color scale?

```
scatter <- ggplot(my_data,aes(x=NURSES,y=RISK)) + geom_point(aes(color = BEDS))
scatter
```



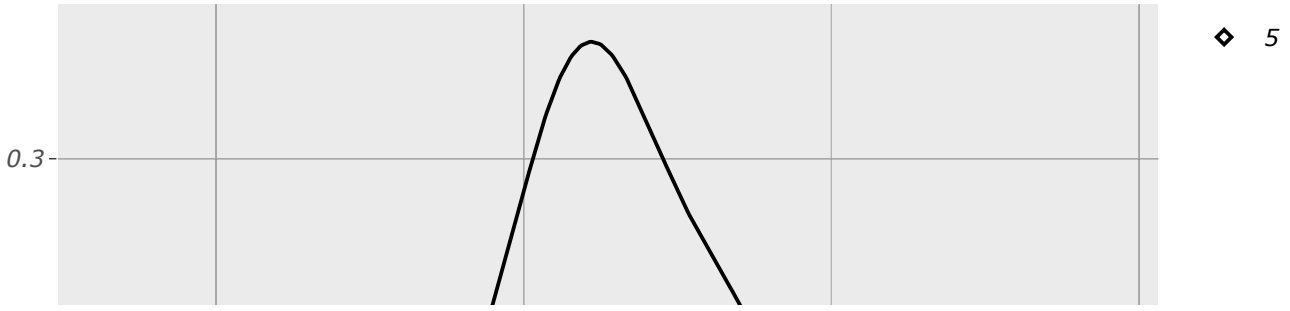
In the previous plot, we use different density which was giving information only about that variable. But when we use a scatter plot we can understand the dependence of one variable on the other. From the plot we can see that when there more number of beds we need more nurses to reduce the risk of infection.

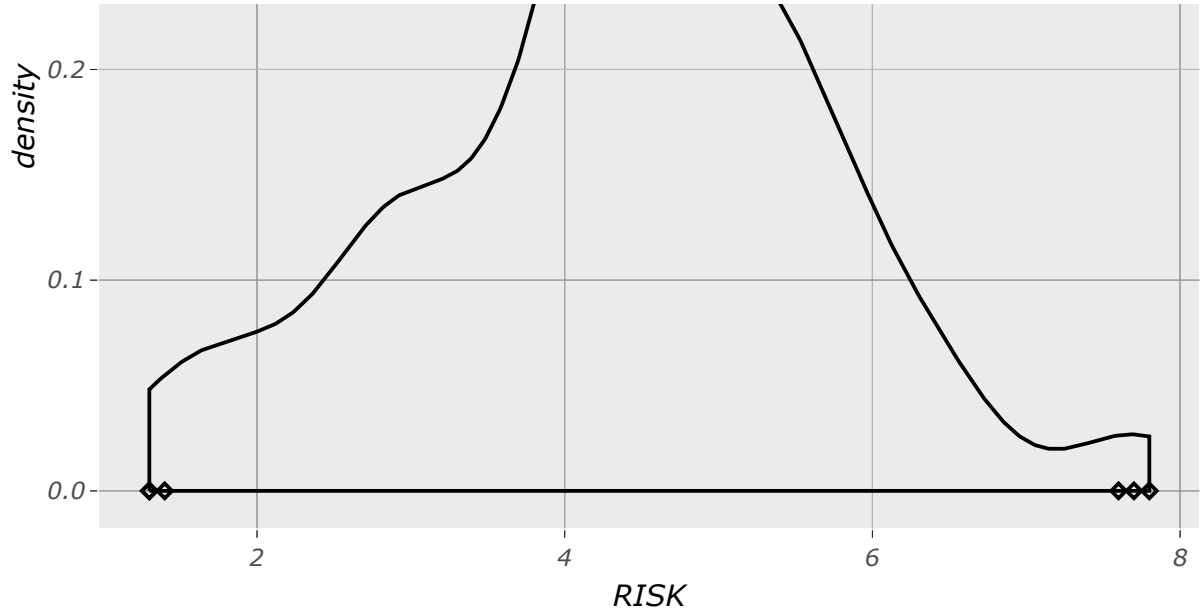
This color scale makes it difficult to distinguish because the colours are very similar.

Question 6

Convert graph from ggplot2 to Plotly with ggplotly function. What important new functionality have you obtained compared to the graph from step 3? Make some additional analysis of the new graph.

```
library(plotly)
ggplotly(final)
```



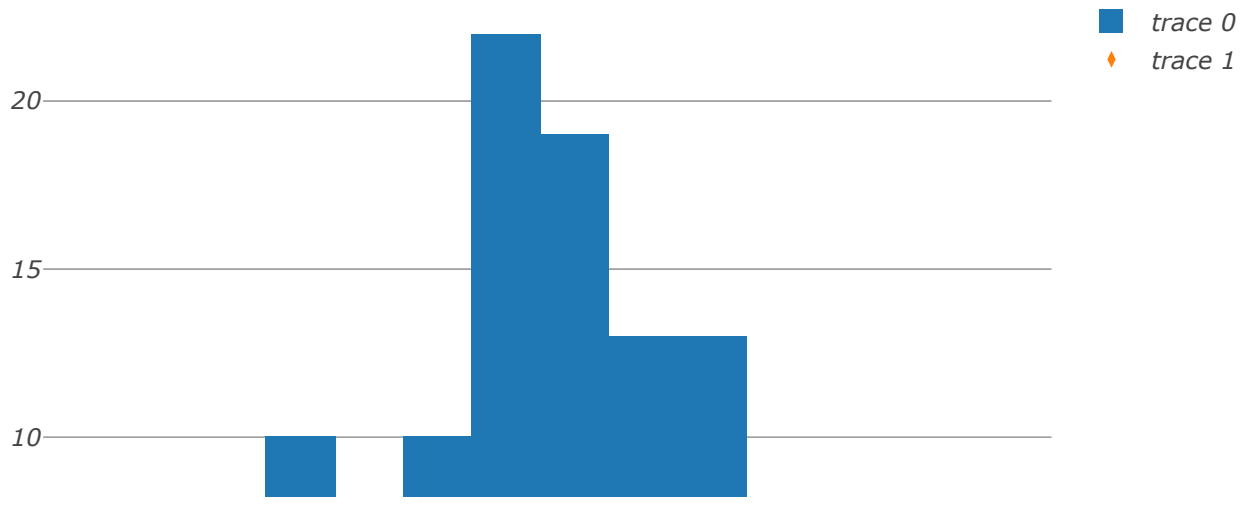


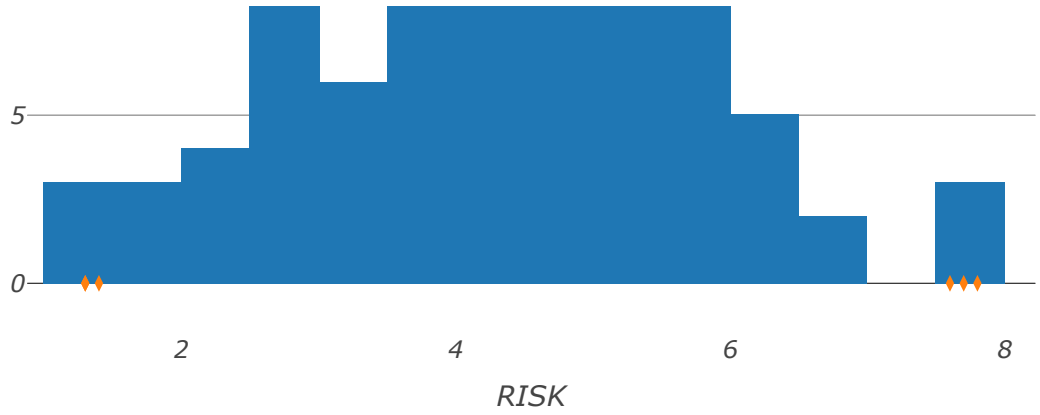
Using plotly helps us to make a more detailed analysis.Now its easy to find the maximum values, and the values of the outliers from just the graph alone.

Question 7

Use data plot-pipeline and the pipeline operator to make a histogram of Infection risk in which outliers are plotted as a diamond symbol (???) . Make this plot in the Plotly directly (i.e. without using ggplot2 functionality). Hint: select(), filter() and is.element() functions might be useful here.

```
library(dplyr)

output_risk <- my_function(my_data$RISK)

my_data_risk <- my_data[output_risk,] #creating a second data set


a <- my_data %>% select(RISK) %>% plot_ly(x = ~RISK , type = "histogram") %>% add_markers(x = my
_data_risk$RISK , y = 0,marker = list(symbol = 23) )

a
```





Question 8

Write a Shiny app that produces the same kind of plot as in step 4 but in addition includes: (a) Checkboxes indicating for which variables density plots should be produced. (b) A slider changing the bandwidth parameter in the density estimation ('bw' parameter).

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

reqd_vars <-
c("STAY","AGE","RISK","CULTURE_RATIO","CHEST","BEDS","AFFILIATION","CENSUS","NURSES","FS")

my_function <- function(x){

Q1 <- quantile(x,0.25) Q3 <- quantile(x,0.75)

up <- Q3 + 1.5(Q3 - Q1) down <- Q1 - 1.5(Q3 - Q1)

result1 <- which(x > up)

result2 <- which(x < down)

final <- sort(c(result1,result2))

return(final)

}

ui <- fluidPage(

sliderInput(inputId="bw_value", label="Choose bandwidth size", value=0.01, min=0.1, max=10),
checkboxGroupInput("var", "Variables", reqd_vars , inline=TRUE,selected = "STAY"), plotOutput("densPlot") )

server <- function(input, output) {

output$densPlot <- renderPlot({
```

```
chosen <- input$var

result <- vector("list" , length = length(chosen))




for (i  in 1:length(chosen)) {


  outliers <- my_function(my_data[,chosen[i]])

  my_data2 <- my_data[outliers,] #creating a second data set



  result[[i]] <- ggplot() + geom_density(data = my_data, aes_string(x = chosen[i] ),bw = input$b
w_value) + geom_point(data=my_data2, aes_string(x = chosen[i] , y = 0  )  , shape = 5) + scale_s
hape_identity()


}

final <- arrangeGrob(grobs = result)
grid.arrange(final)
```

}) }

shinyApp(ui = ui, server = server)

With increase in the value of bandwidth, the smoothness of the graph increases. A bandwidth value of 1.5 is ideal since it is decently smoothed and its easy to find the outliers value also