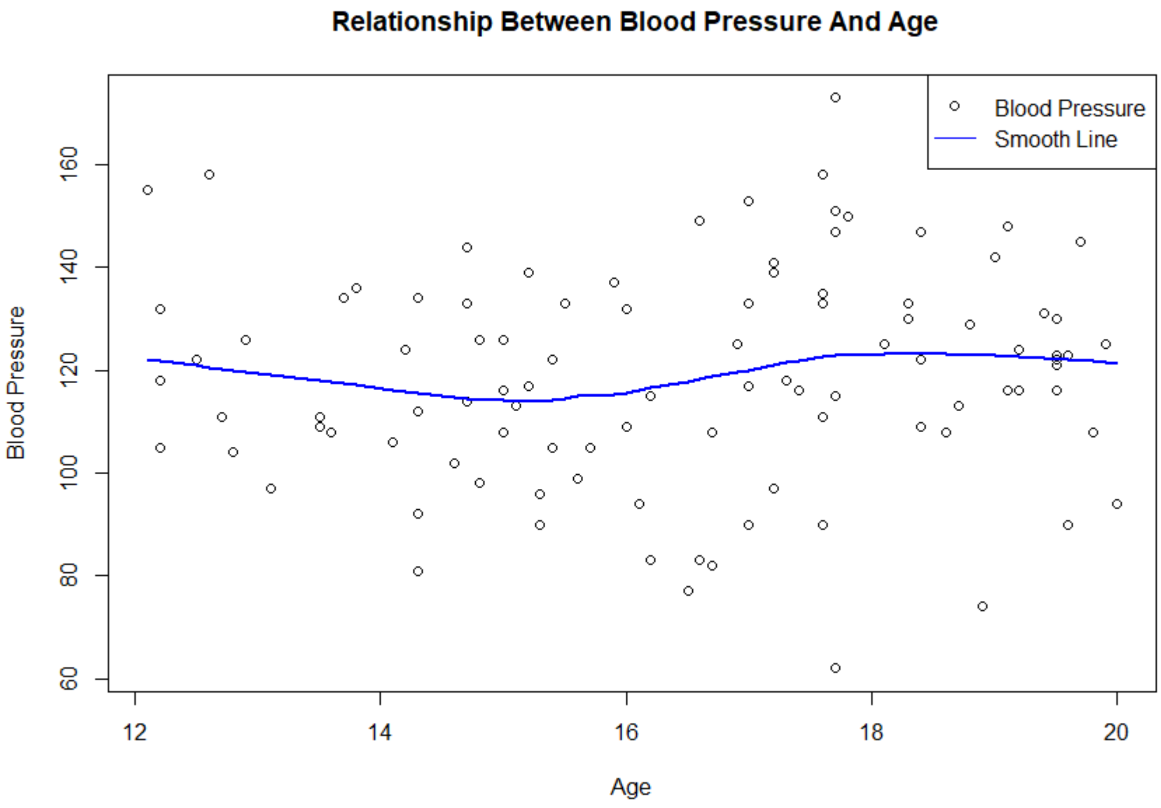
CS24200: Homework 4

Wei Huang

Fall 2019

1 Basic plots in R

1.1 Scatter Plot



Code:

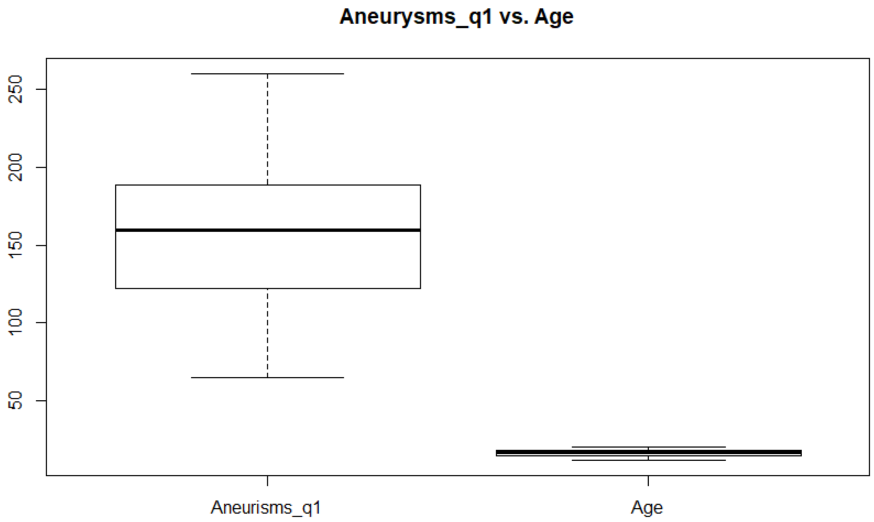
```
df = read.table(file = "../aneurysm_data.csv", sep = ',', header = TRUE)

# Q1.1
plot(df$Age, df$BloodPressure, main="Relationship␣Between␣Blood␣Pressure␣And␣Age",
      xlab = "Age", ylab = "Blood␣Pressure")
```

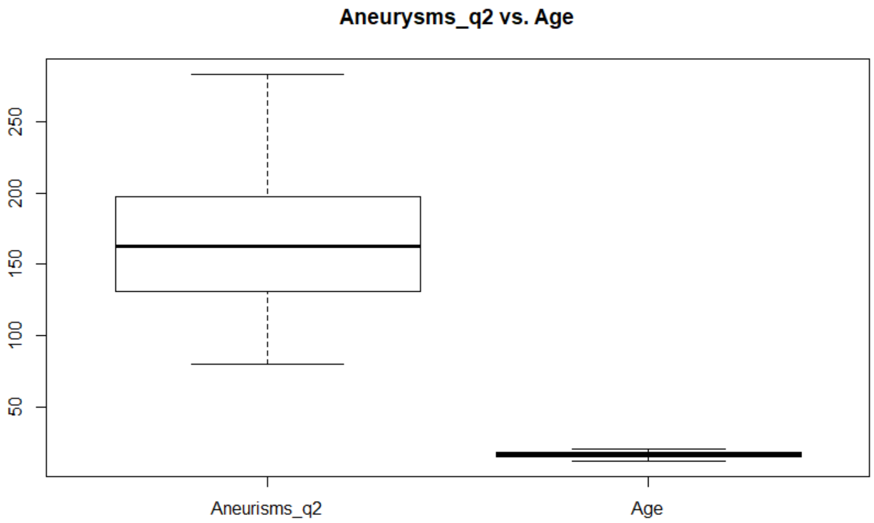
```
lines(lowess(df$BloodPressure~df$Age, f = 2/3), col = ’blue’, lwd = 2)
legend("topright", legend=c("Blood␣Pressure", "Smooth␣Line"),
       col=c("black", "blue"), pch = c(1,NA), lty= c(NA,1), cex=1)
```

1.2 Box Plot

1.2.1 (a) Aneurysms q1 vs. Age



1.2.2 (a) Aneurysms q2 vs. Age

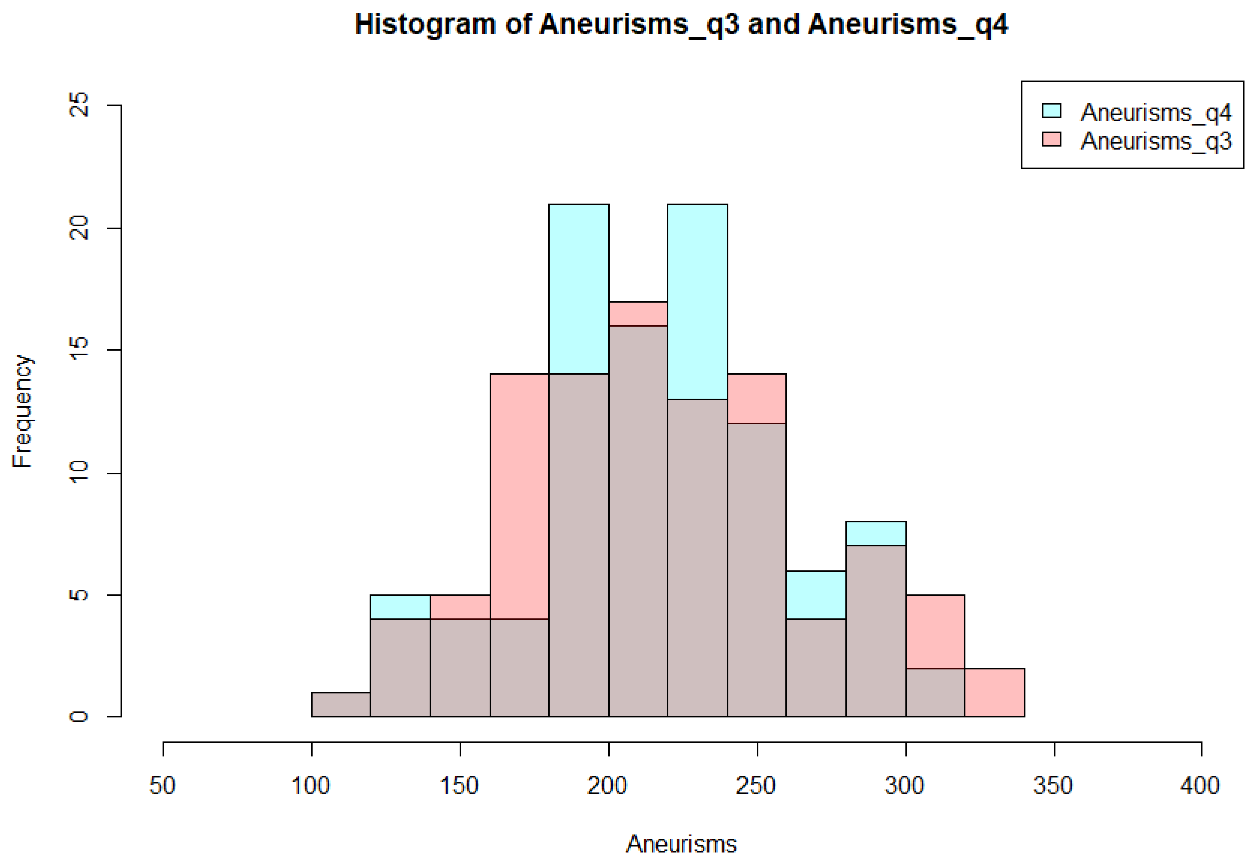


Code:

```
df = read.table(file = "../aneurysm_data.csv", sep = ',', header = TRUE)
```

```
# Q1.2
boxplot(df$Aneurisms_q1, df$Age, main="Aneurysms_q1␣vs.␣Age", names=c("Aneurisms_q1","Age"))
boxplot(df$Aneurisms_q2, df$Age, main="Aneurysms_q2␣vs.␣Age", names=c("Aneurisms_q2","Age"))
```

1.3 Histogram



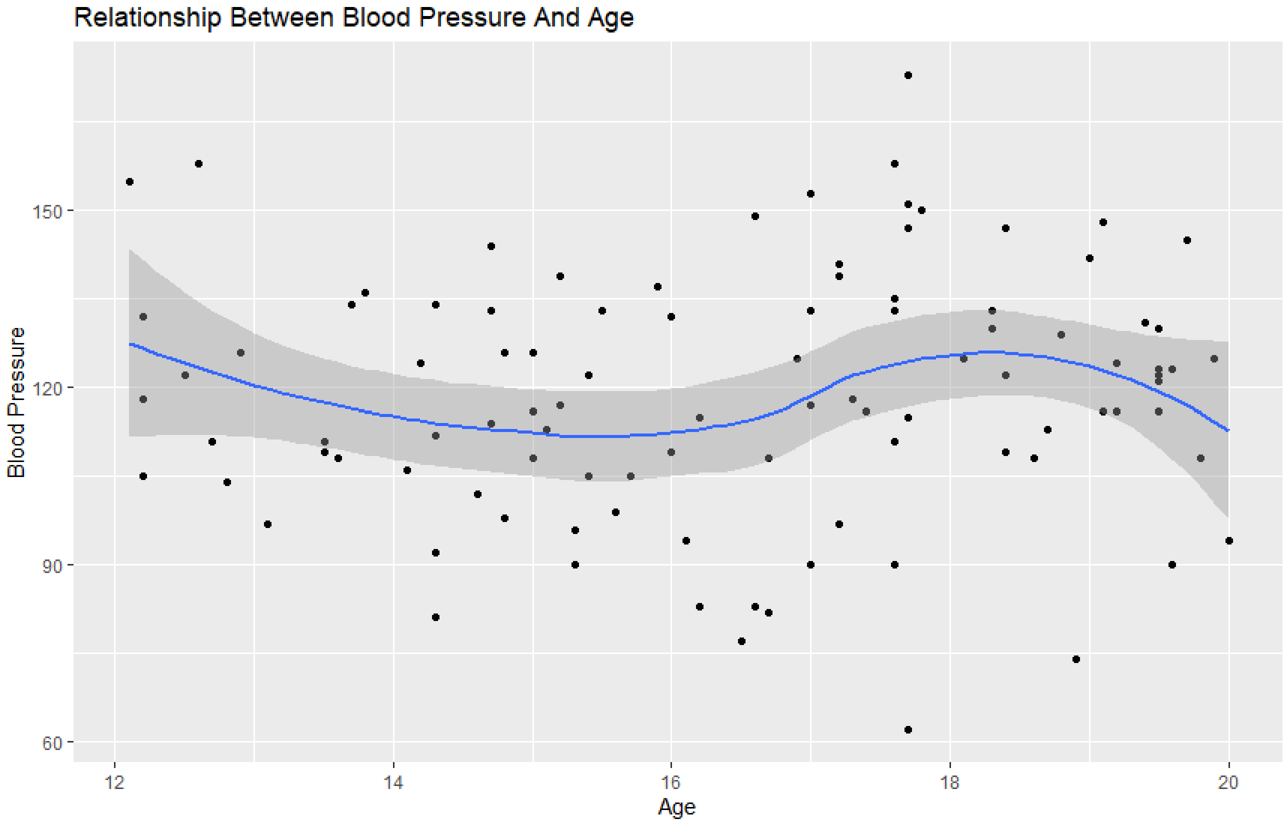
Code:

```
df = read.table(file = "../aneurysm_data.csv", sep = ’,’, header = TRUE)

# Q1.3
hist(df$Aneurisms_q4, col = rgb(0,1,1,0.25), xlim = c(50,400), ylim = c(0,25),
     main = "Histogram␣of␣Aneurisms_q3␣and␣Aneurisms_q4", xlab = "Aneurisms")
hist(df$Aneurisms_q3, add = T, col =rgb(1,0,0,0.25))
legend("topright", legend=c("Aneurisms_q4", "Aneurisms_q3"),
       fill=c(rgb(0,1,1,0.25), rgb(1,0,0,0.25)))
```

2 Using ggplot2

2.1 Scatter Plot

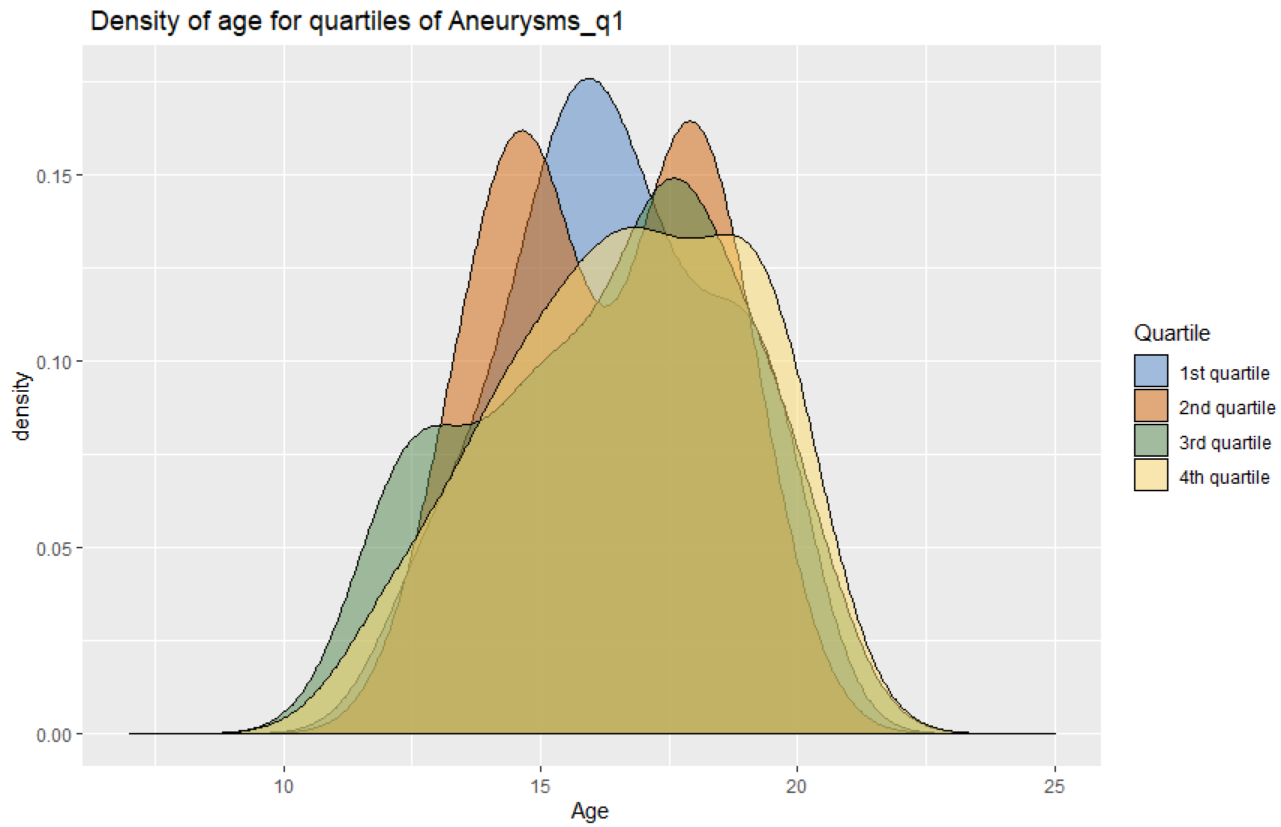


Code:

```
df = read.table(file = "../aneurysm_data.csv", sep = ',', header = TRUE)

# Q2.1
library("ggplot2")
p <- ggplot(df, aes(y=BloodPressure, x=Age))
p + geom_point() + geom_smooth(method = 'loess') +
  ggtitle("Relationship␣Between␣Blood␣Pressure␣And␣Age") + xlab("Age") + ylab("Blood␣Pressure")
```

2.2 Density Plot

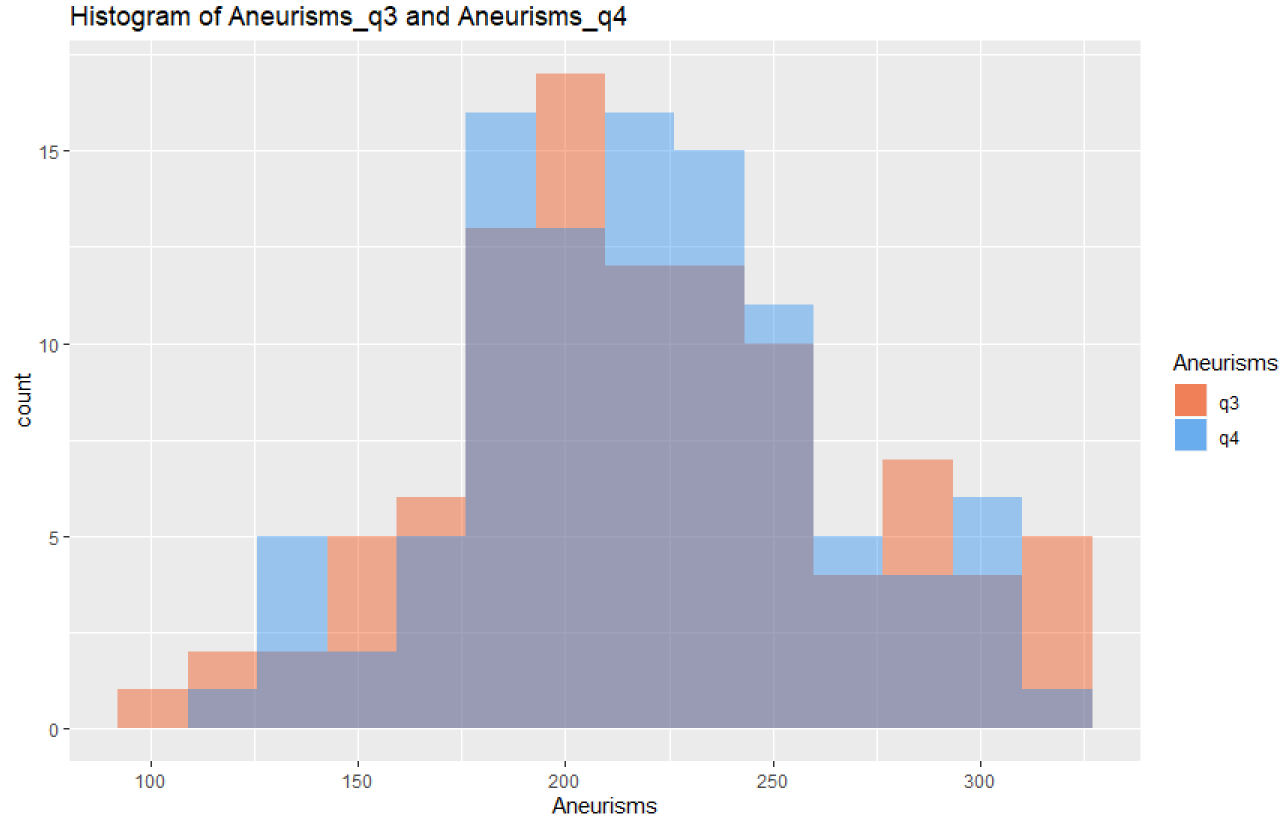


Code:

```
df = read.table(file = "../aneurysm_data.csv", sep = ',', header = TRUE)

# Q2.2
quart_1 <- subset(df, df$Aneurisms_q1 <= quantile(df$Aneurisms_q1,0.25))
quart_2 <- subset(df, df$Aneurisms_q1 > quantile(df$Aneurisms_q1,0.25) &
                    df$Aneurisms_q1 <= quantile(df$Aneurisms_q1, 0.5))
quart_3 <- subset(df, df$Aneurisms_q1 > quantile(df$Aneurisms_q1,0.5) &
                    df$Aneurisms_q1 <= quantile(df$Aneurisms_q1, 0.75))
quart_4 <- subset(df, df$Aneurisms_q1 > quantile(df$Aneurisms_q1,0.75))
quart_1$Quartile = "1st␣quartile"
quart_2$Quartile = "2nd␣quartile"
quart_3$Quartile = "3rd␣quartile"
quart_4$Quartile = "4th␣quartile"
quart = rbind(quart_1, quart_2)
quart = rbind(quart, quart_3)
quart = rbind(quart, quart_4)
ggplot(data = quart, aes(x=Age))+geom_density(aes(fill =Quartile, alpha=Quartile)) +
  scale_fill_manual(values = c("#4E84C4", "#D16103", "#52854C", "#FFDB6D")) +
  scale_alpha_manual(values = c(0.5, 0.5,0.5,0.5)) + xlim(c(7, 25)) +
  ggtitle("␣Density␣of␣age␣for␣quartiles␣of␣Aneurysms_q1") + xlab("Age")
```

2.3 Histogram



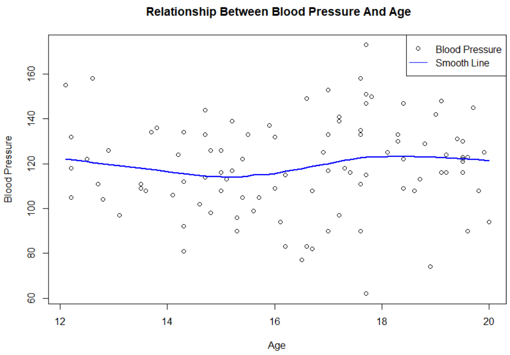
Code:

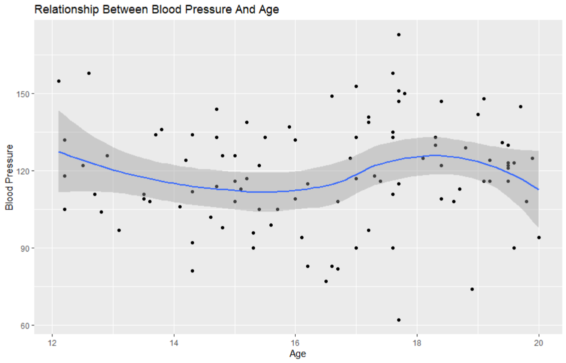
```
df = read.table(file = "../aneurysm_data.csv", sep = ',', header = TRUE)

# Q2.3
newdf = melt(df,id.vars=c("ID","Gender","Group","BloodPressure","Age",
                          "Aneurisms_q1", "Aneurisms_q2"),
             variable.name="q",value.name="Aneurisms")
ggplot(newdf,aes(x= Aneurisms) )+
  geom_histogram(data = subset(newdf, q == "Aneurisms_q3"), aes(fill = q),
                 bins = 14, alpha = 0.4) +
  geom_histogram(data = subset(newdf, q == "Aneurisms_q4"), aes(fill = q),
                 bins = 14, alpha = 0.4) +
  scale_fill_manual(name="Aneurisms", values=c("orangered2","dodgerblue2"),
                    labels=c("q3","q4")) +
  ggtitle("Histogram␣of␣Aneurisms_q3␣and␣Aneurisms_q4")
```

2.4 Discussion

I think ggplot produce more beautiful plots than basic R.

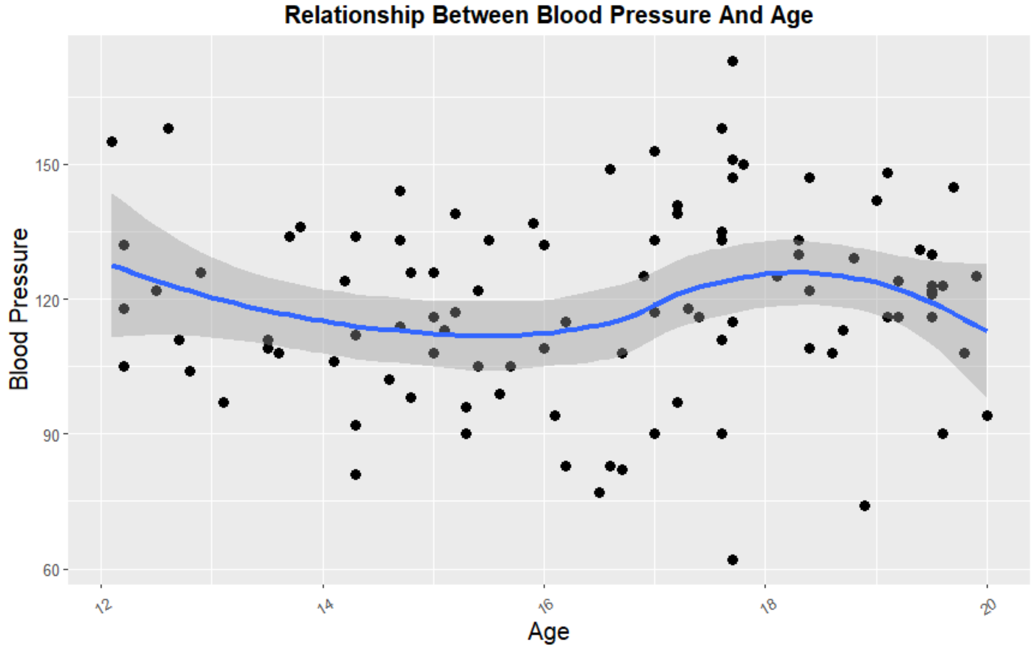


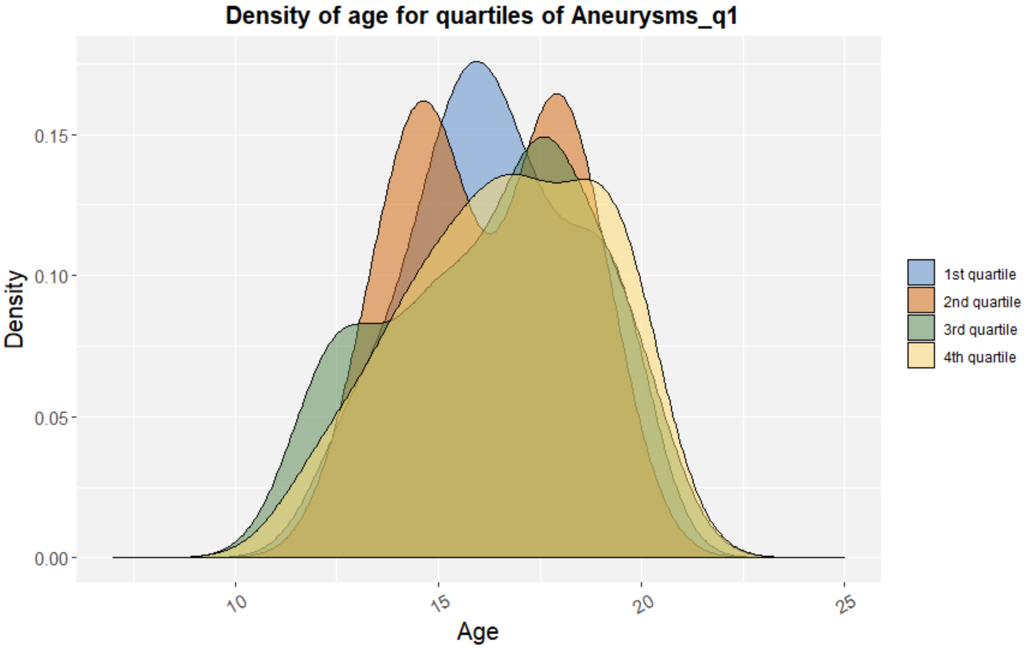


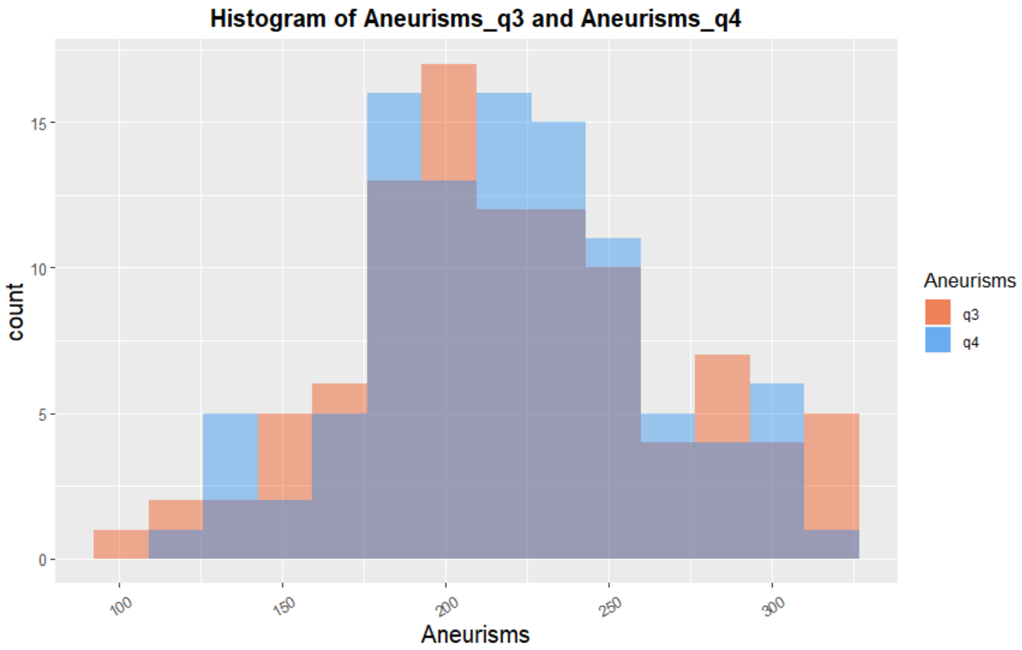
As figures show above, by comparing scatter plot and smooth line created by basic R (Q1.1) and ggplot (Q2.1), we can find out that graph created using ggplot is clearly than basic R. Plots are much easier to identify by using black points and gray panel than using circles by default in R. Grid lines on the graph helps a lot when trying to find the corresponding x,y coordinate for a specific point. Besides, ”geom_smooth” function not only provide a smooth curve that fit the line as ”lowess” does, but also it shows confidence interval around smooth which gives more infomation.

2.5 Customize Plots By Using Theme()

Theme() is a powerful function that can be used to make plots more pleasing and informative.
For example, the color, size and position of the title and subtitle can be modified by theme() function to make it clearer. It can change the size and rotate tick text or adjust style of legend. Grid and panel can be modified by theme() too.New plots are shown below.

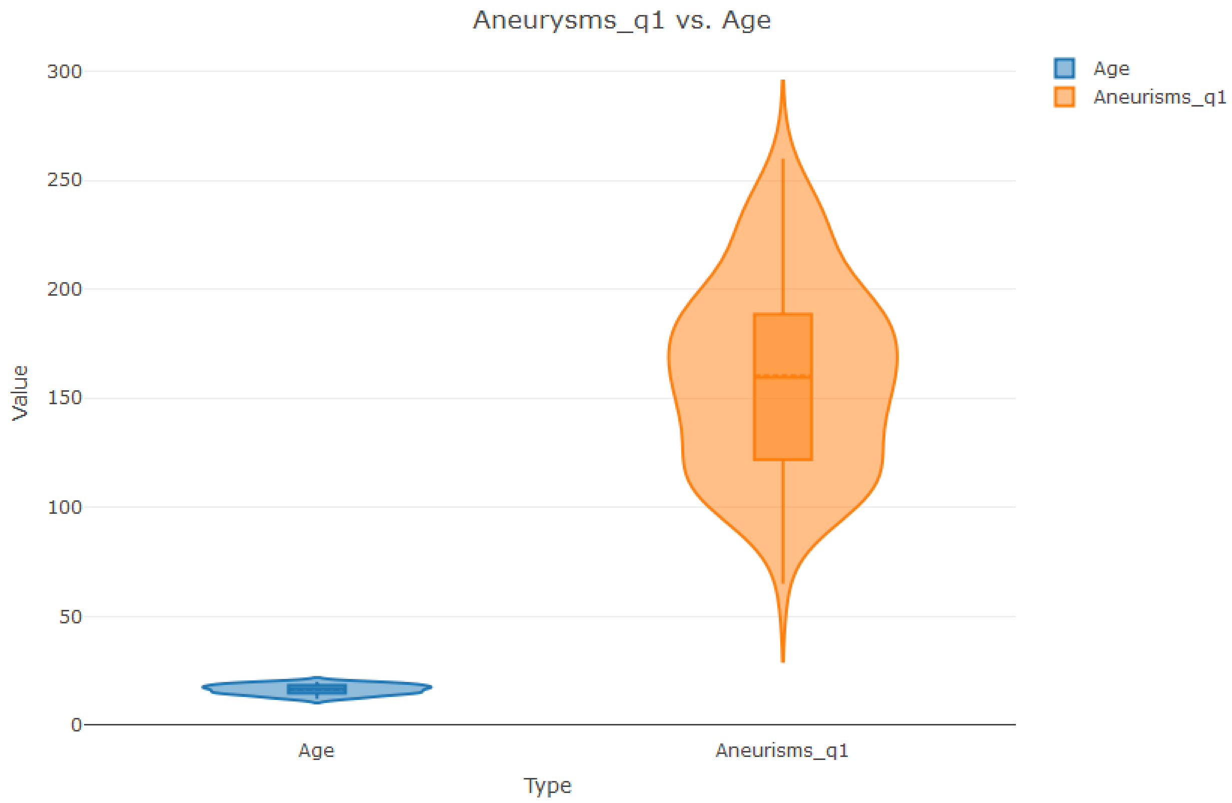





Histogram of Aneurisms_q3 and Aneurisms_q4
count
15
10
5
0
100
150
200
250
300
Aneurisms
Aneurisms
q3
q4

3 Using plotly

3.1 Violin Plot



Code:

```
df = read.table(file = "../aneurysm_data.csv", sep = ',', header = TRUE)

# Q3.1
library("plotly")
Sys.setenv("plotly_username"="WeiH")
Sys.setenv("plotly_api_key"="****************")
newdf2 = df[c(5,6)]
newdf2 = melt(newdf2, variable.name="Type",value.name="Value")
p <- plot_ly(newdf2, x = ~Type, y=~Value, split = ~Type,
             type = "violin", box = list(visible = T),meanline = list(visible = T)) %>%
  layout(title = "Aneurysms_q1␣vs.␣Age")
p
api_create(p, filename = "HW4_Q3")
# https://plot.ly/~WeiH/3/
```

3.2 Violin Plot vs. Box Plot

Violin plot not only displays information shown in box plot, such as mean, maximum, minimum values..., but also shows density plot of the data outside the box plot. In this case, people can get a preliminary understanding of this set of data by viewing the box plot and also generally know how the data is distributed.

3.3 Online Chart

Link: `https://plot.ly/~WeiH/3/`

4 Appendix

4.1 Question 1

```
df = read.table(file = "../aneurysm_data.csv", sep = ’,’, header = TRUE)

# Q1.1
plot(df$Age, df$BloodPressure, main="Relationship␣Between␣Blood␣Pressure␣And␣Age",
     xlab = "Age", ylab = "Blood␣Pressure")
lines(lowess(df$BloodPressure~df$Age, f = 2/3), col = ’blue’, lwd = 2)
legend("topright", legend=c("Blood␣Pressure", "Smooth␣Line"),
       col=c("black", "blue"), pch = c(1,NA), lty= c(NA,1), cex=1)

# Q1.2
boxplot(df$Aneurisms_q1, df$Age, main="Aneurysms_q1␣vs.␣Age", names=c("Aneurisms_q1","Age"))
boxplot(df$Aneurisms_q2, df$Age, main="Aneurysms_q2␣vs.␣Age", names=c("Aneurisms_q2","Age"))
?floor

# Q1.3
hist(df$Aneurisms_q4, col = rgb(0,1,1,0.25), xlim = c(50,400), ylim = c(0,25),
     main = "Histogram␣of␣Aneurisms_q3␣and␣Aneurisms_q4", xlab = "Aneurisms")
hist(df$Aneurisms_q3, add = T, col =rgb(1,0,0,0.25))
legend("topright", legend=c("Aneurisms_q4", "Aneurisms_q3"),
       fill=c(rgb(0,1,1,0.25), rgb(1,0,0,0.25)))
```

4.2 Question 2

```
# Q2.1
library("ggplot2")
p <- ggplot(df, aes(y=BloodPressure, x=Age))
p + geom_point() + geom_smooth(method = ’loess’) +
  ggtitle("Relationship␣Between␣Blood␣Pressure␣And␣Age") + xlab("Age") + ylab("Blood␣Pressure")

# Q2.2
quart_1 <- subset(df, df$Aneurisms_q1 <= quantile(df$Aneurisms_q1,0.25))
quart_2 <- subset(df, df$Aneurisms_q1 > quantile(df$Aneurisms_q1,0.25) &
                    df$Aneurisms_q1 <= quantile(df$Aneurisms_q1, 0.5))
quart_3 <- subset(df, df$Aneurisms_q1 > quantile(df$Aneurisms_q1,0.5) &
                    df$Aneurisms_q1 <= quantile(df$Aneurisms_q1, 0.75))
quart_4 <- subset(df, df$Aneurisms_q1 > quantile(df$Aneurisms_q1,0.75))
quart_1$Quartile = "1st␣quartile"
quart_2$Quartile = "2nd␣quartile"
quart_3$Quartile = "3rd␣quartile"
quart_4$Quartile = "4th␣quartile"
quart = rbind(quart_1, quart_2)
quart = rbind(quart, quart_3)
quart = rbind(quart, quart_4)
ggplot(data = quart, aes(x=Age))+geom_density(aes(fill =Quartile, alpha=Quartile)) +
  scale_fill_manual(values = c("#4E84C4", "#D16103", "#52854C", "#FFDB6D")) +
  scale_alpha_manual(values = c(0.5, 0.5,0.5,0.5)) + xlim(c(7, 25)) +
  ggtitle("␣Density␣of␣age␣for␣quartiles␣of␣Aneurysms_q1") + xlab("Age")
```

```
# Q2.3
library(reshape2)
library(knitr)
newdf = melt(df,id.vars=c("ID","Gender","Group","BloodPressure","Age",
                          "Aneurisms_q1", "Aneurisms_q2"),
             variable.name="q",value.name="Aneurisms")
ggplot(newdf,aes(x= Aneurisms) )+
  geom_histogram(data = subset(newdf, q == "Aneurisms_q3"), aes(fill = q),
                 bins = 14, alpha = 0.4) +
  geom_histogram(data = subset(newdf, q == "Aneurisms_q4"), aes(fill = q),
                 bins = 14, alpha = 0.4) +
  scale_fill_manual(name="Aneurisms", values=c("orangered2","dodgerblue2"),
                    labels=c("q3","q4")) +
  ggtitle("Histogram␣of␣Aneurisms_q3␣and␣Aneurisms_q4")
```

4.3 Question 3

```
# Q3.1
library("plotly")
Sys.setenv("plotly_username"="WeiH")
Sys.setenv("plotly_api_key"="****************")
newdf2 = df[c(5,6)]
newdf2 = melt(newdf2, variable.name="Type",value.name="Value")
p <- plot_ly(newdf2, x = ~Type, y=~Value, split = ~Type,
             type = "violin", box = list(visible = T),meanline = list(visible = T)) %>%
  layout(title = "Aneurysms_q1␣vs.␣Age")
p
api_create(p, filename = "HW4_Q3")
# https://plot.ly/~WeiH/3/
```