STA 138 Discussion 1 Solutions

Fall 2020

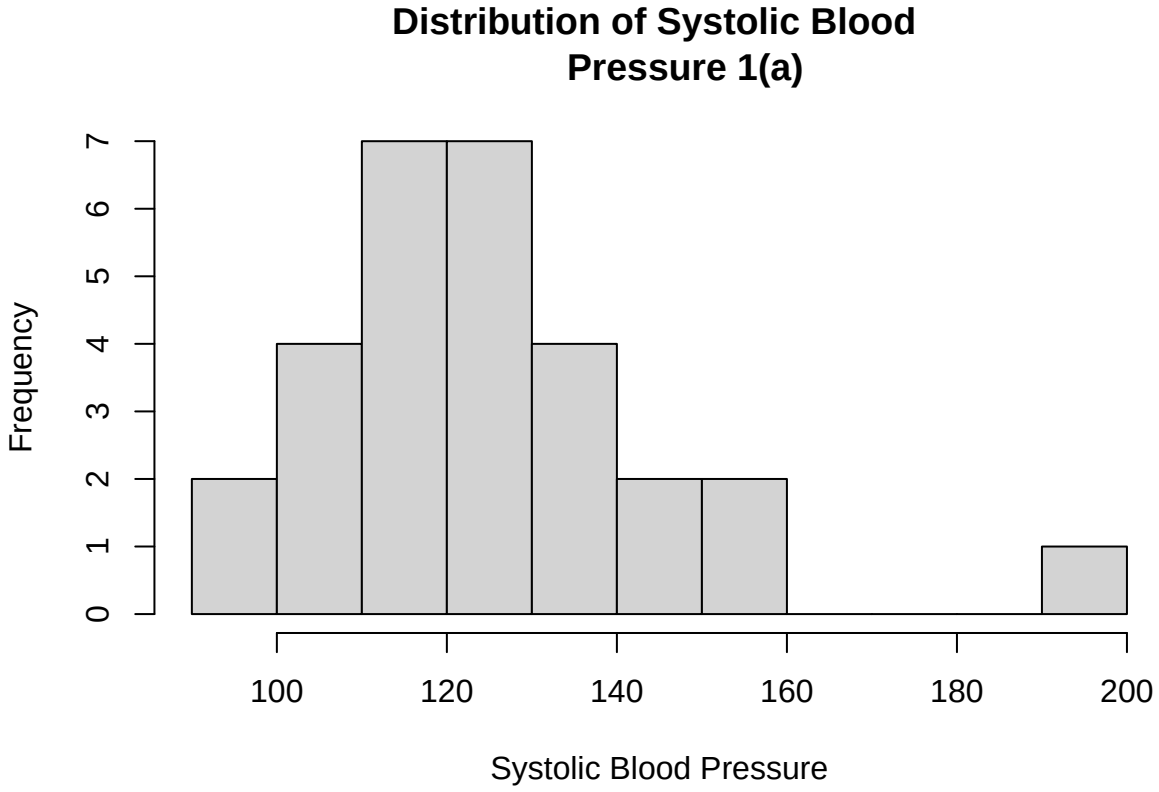
1. (Note that solutions here illustrate both base R and ggplot2; use of either or both is fine!)

`ggplot2` is a data visualization package for the statistical programming language R. Created by Hadley Wickham in 2005, ggplot2 is an implementation of Leland Wilkinson's Grammar of Graphics—a general scheme for data visualization which breaks up graphs into semantic components such as scales and layers.

For our discussions, we'll only use base R. However, you are highly encouraged to explore the usage of `ggplot2`.

(a)

- `break`: a single number giving the number of cells for the histogram
- `xlab`: x axis label
- `main`: main title



(b) With the exception of a single relative outlier, the data appears fairly symmetric.

(c)

- `horizontal`: logical indicating if the boxplots should be horizontal; default FALSE means vertical boxes.
- `main`: main title

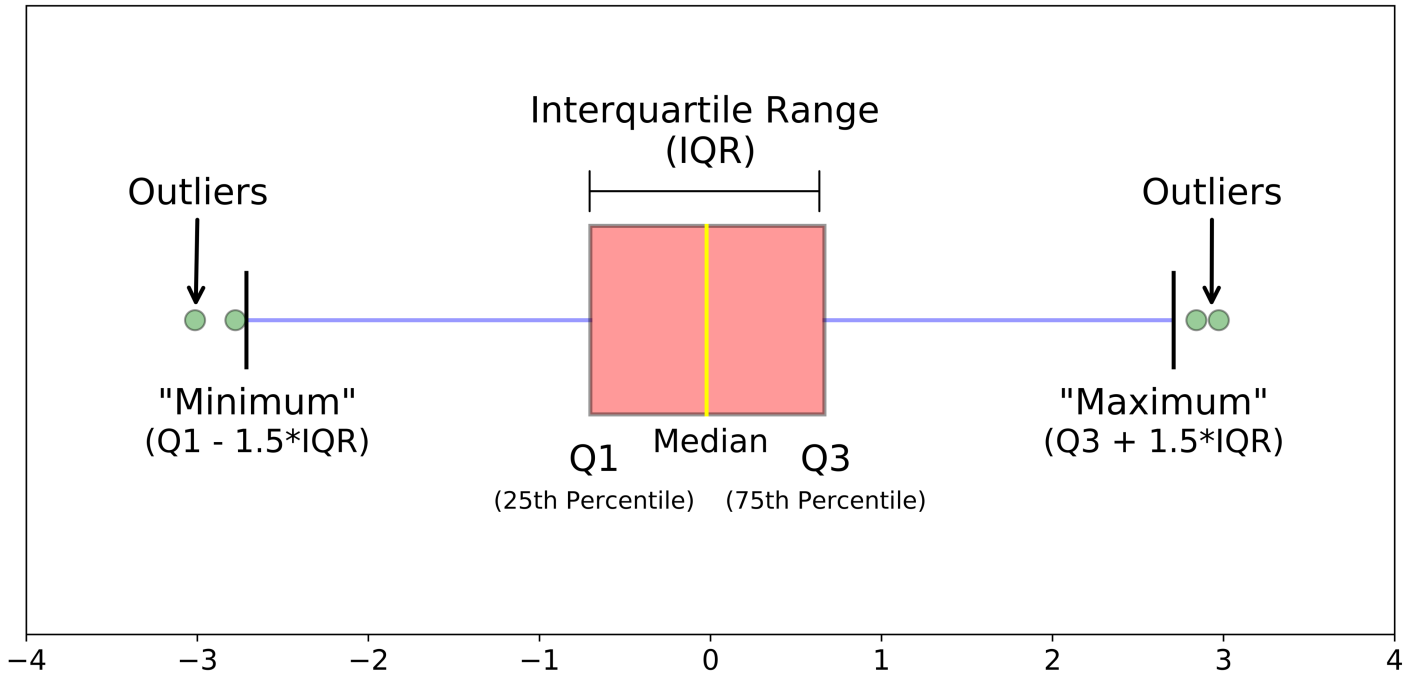
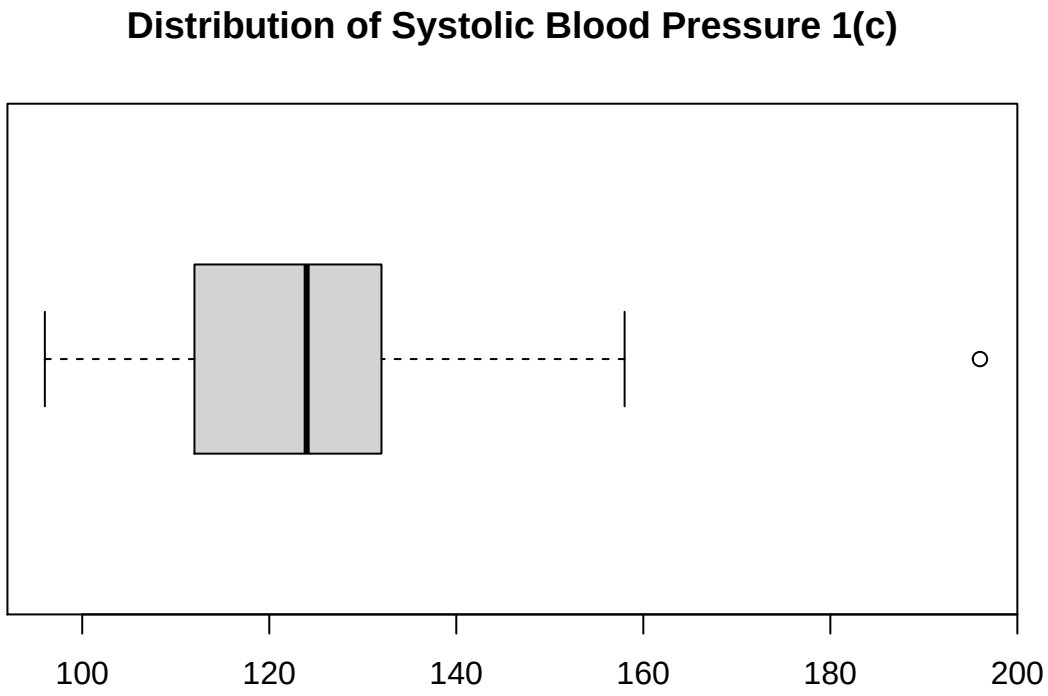


Figure 1: understanding boxplot

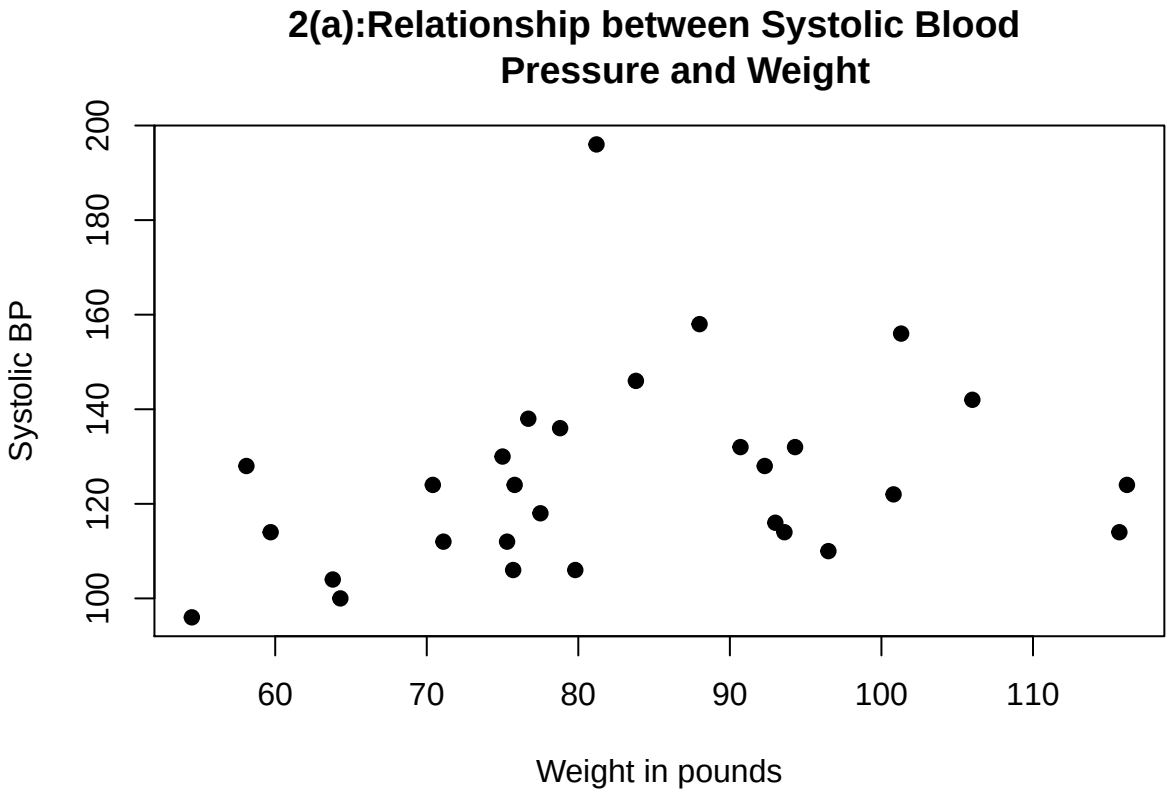


(d) There is one clear, unusually large observation apparent from the boxplot.

2.

(a)

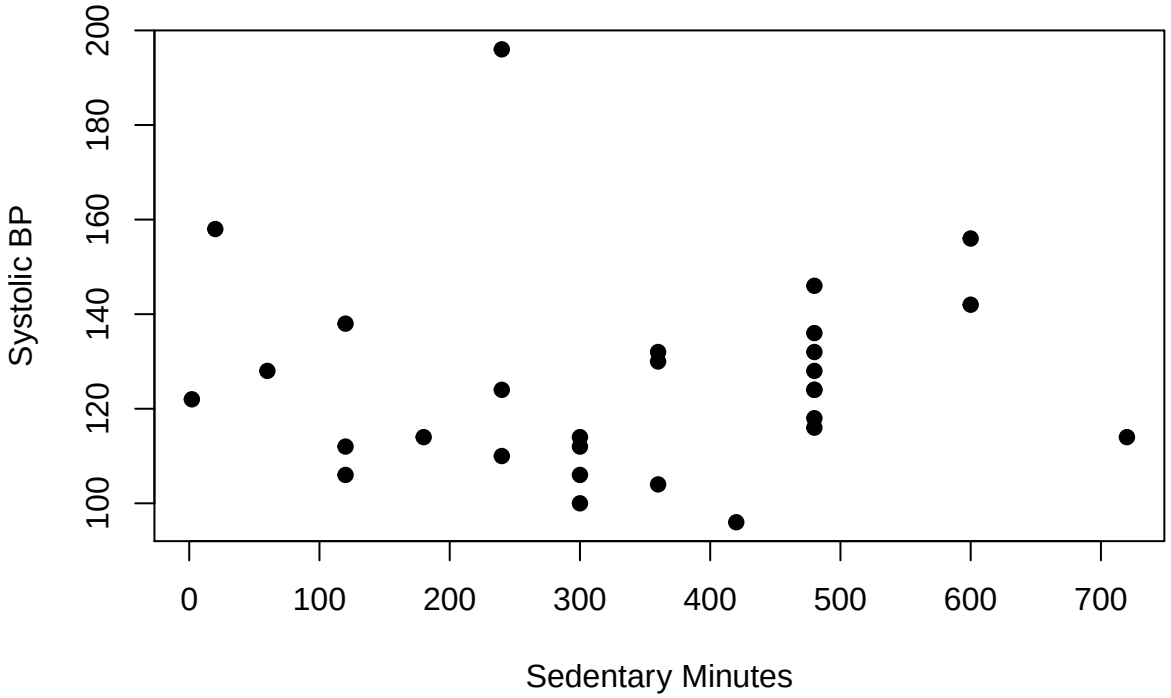
- `main`: main title
- `xlab`: x-axis label
- `ylab`: y-axis label
- `pch`: control points symbols (`pch = 19` solid circle)



(b) There seems to be a slight positive trend; as weight increases, the average systolic blood pressure appears to increase slightly as well.

(c)

2(c):Relationship between Systolic Blood Pressure and Sedentary Min

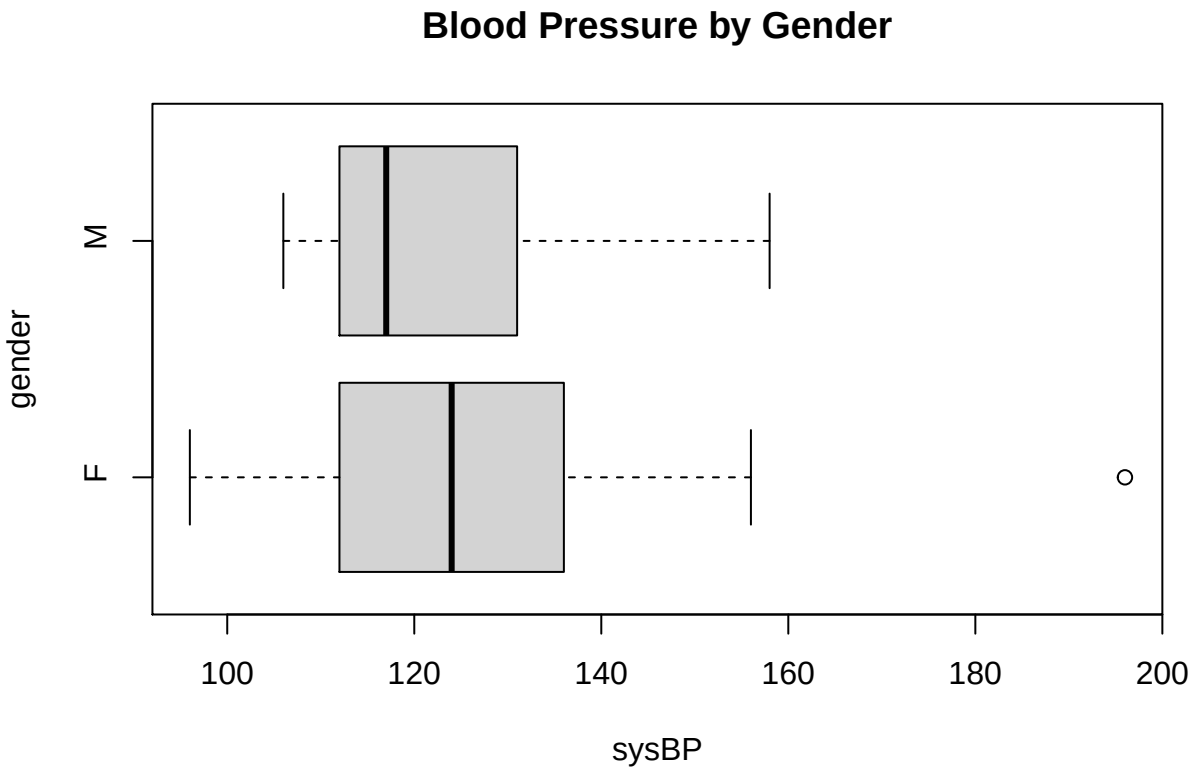


(d) There does not seem to be a notable trend here.

3.

(a)

- `formula`: a formula, such as `y ~ grp`, where `y` is a numeric vector of data values to be split into groups according to the grouping variable `grp` (usually a factor). Note that `~ g1 + g2` is equivalent to `g1:g2`.

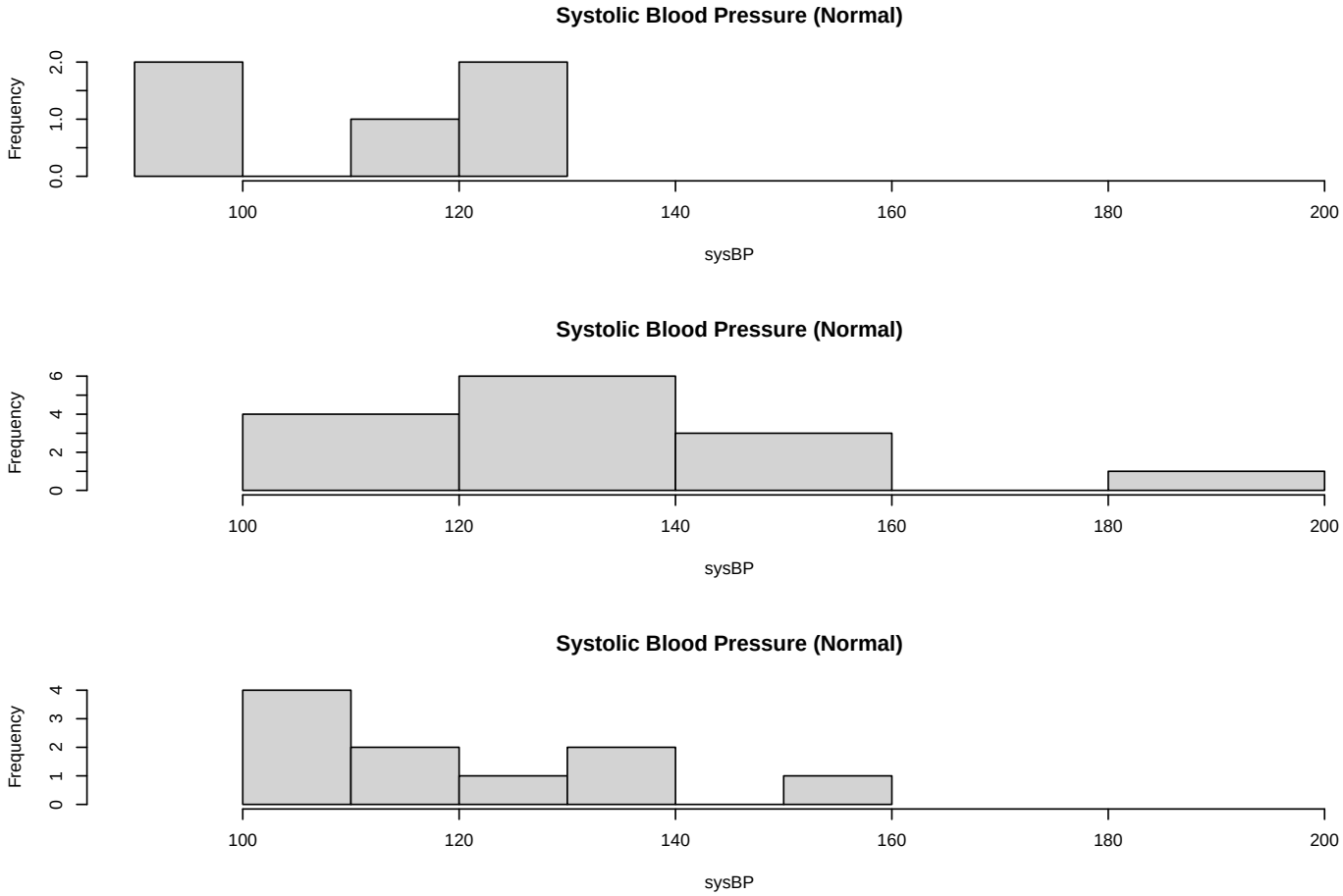


(b) The histogram for males and females appear to completely overlap, so they show no significant difference.

(c)

- `par()`: set or query graphical parameters

With the `par( )` function, you can include the option `mfrow=c(nrows, ncols)` to create a matrix of nrows by ncols plots that are filled in by row. `mfcol=c(nrows, ncols)` fills in the matrix by columns.



(d) The 'Normal' weight category appears to have lower systolic blood pressures on average than 'obese' and 'overweight,' but there does not appear to be much of a difference between 'overweight' and 'obese.'

4.

- `round`: rounds the values in its first argument to the specified number of decimal places (default 0)
- `aggregate`: Splits the data into subsets, computes summary statistics for each, and returns the result in a convenient form.

(a) The average systolic blood pressure is: 125.4483.

(b) A table of results follows:

	marriage	ave sysBP
1	divorced	133.50
2	married	124.77
3	nevermarried	115.50
4	other	126.00
5	widowed	132.75

(c) A table of results follows:

	marriage	sd sysBP
1	divorced	22.85
2	married	21.34
3	nevermarried	19.27
4	other	22.21
5	widowed	8.81

(d) A table of results follows:

	marriage	number
1	divorced	12
2	married	52
3	nevermarried	16
4	other	12
5	widowed	8

(e) The five number summary is: 54.5 (min), 75 (Q2), 79.800003 (Median), 93.599998 (Q3), 116.2 (Max)

Code Appendix

```
setwd(dirname(rstudioapi::getSourceEditorContext()$path))# set working directory to source file location
library(ggplot2)# Create Elegant Data Visualizations Using the Grammar of Graphics
patients101 = read.csv("patients101.csv")# import patients101.csv
#Problem 1 (a)
hist(patients101$sysBP, breaks=10, xlab = "Systolic Blood Pressure",main = "Distribution of Systolic Blood
     Pressure 1(a)")
#ggplot(patients101, aes(x = sysBP)) + geom_histogram(binwidth = 5,color = "black",fill = "white")+
#xlab("Systolic Blood Pressure")+ggtitle("1(a): Distribution of systolic blood pressure")
#Problem 1 (c)
boxplot(patients101$sysBP, horizontal = TRUE,main = "Distribution of Systolic Blood Pressure 1(c)")
# ggplot(patients101, aes(y=sysBP, x = factor(""")))+ geom_boxplot() + ylab("Systolic Blood Pressure") +
#xlab(" ")+ coord_flip() + ggtitle("1(a): Distribution of systolic blood pressure")
#Problem 2 (a)
plot(patients101$weight, patients101$sysBP, main = "2(a):Relationship between Systolic Blood
     Pressure and Weight",xlab = "Weight in pounds",ylab = "Systolic BP",pch = 19)
#qplot(weight, sysBP, data = patients101) +
#ggtitle("2(a):Relationship between Systolic Blood Pressure and Weight") +
#xlab("Weight in pounds") + ylab("Systolic BP")
#Problem 2 (c)
plot(patients101$sedmins, patients101$sysBP,
     main = "2(c):Relationship between Systolic Blood Pressure and Sedentary Minutes",
     xlab = "Sedentary Minutes",ylab = "Systolic BP",pch = 19)
#qplot(sedmins, sysBP, data = patients101) +
```

```r
#ggtitle("2(c):Relationship between Systolic Blood Pressure and Sedentary Minutes") +
#xlab("Sedentary Minutes") + ylab("Systolic BP")
#Problem 3 (a)
boxplot(sysBP ~ gender, data = patients101, main = "Blood Pressure by Gender",horizontal = TRUE)
#ggplot(patients101,aes(y = sysBP,x =gender)) + geom_boxplot() + ylab("Systolic Blood Pressure") +
# xlab("Gender") + ggtitle("Blood Pressure by Gender") + coord_flip()
#Problem 3 (c)
par(mfrow = c(3, 1))
hist(patients101$sysBP[patients101$obese==' normal'], main = "Systolic Blood Pressure (Normal)",
     xlab = "sysBP", xlim = c(90, 200))
hist(patients101$sysBP[patients101$obese==' obese'], main = "Systolic Blood Pressure (Normal)",
     xlab = "sysBP", xlim = c(90, 200))
hist(patients101$sysBP[patients101$obese==' overweight'], main = "Systolic Blood Pressure (Normal)",
     xlab = "sysBP", xlim = c(90, 200))
par(mfrow = c(1, 1))# reset to default
#ggplot(patients101, aes(x = sysBP)) + geom_histogram(binwidth = 10,color = "black",fill = "white") +
#facet_grid(obese ~.) +ggtitle("Systolic Blood Pressure by Obesity")
#Problem 4
#part(a)
aveBP = round(mean(patients101$sysBP), 4)
#part(b)
all.ave = aggregate(sysBP ~ marriage, data = patients101, mean)
#part(c)
all.sd = aggregate(sysBP ~ marriage, data = patients101, sd)
#part(d)
all.n= aggregate(sysBP ~ marriage, data = patients101, length)
#part(e)
fns.weight = fivenum(patients101$weight)
```