Assignment-3

Ojaas Hampiholi

2/3/2020

Loading Libraries:

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2
```

```
## -- Attaching packages ---------------------------------------------------------------
------------------ tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## Warning: package 'readr' was built under R version 3.6.2
```

```
## Warning: package 'purrr' was built under R version 3.6.2
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
## Warning: package 'forcats' was built under R version 3.6.2
```

```
## -- Conflicts ------------------------------------------------------------------------
------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(NHANES)
```

```
## Warning: package 'NHANES' was built under R version 3.6.2
```

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

```
## Warning: package 'gridExtra' was built under R version 3.6.2
```

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

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

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

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

```
library(dplyr)
```

Subsection 1:

The relation between Age and Average Systolic BP:

Since the data set is very large(10000 rows), I have taken mean of values of BP readings for common values of Age, Height and Weights depending on the comparisions to be made.

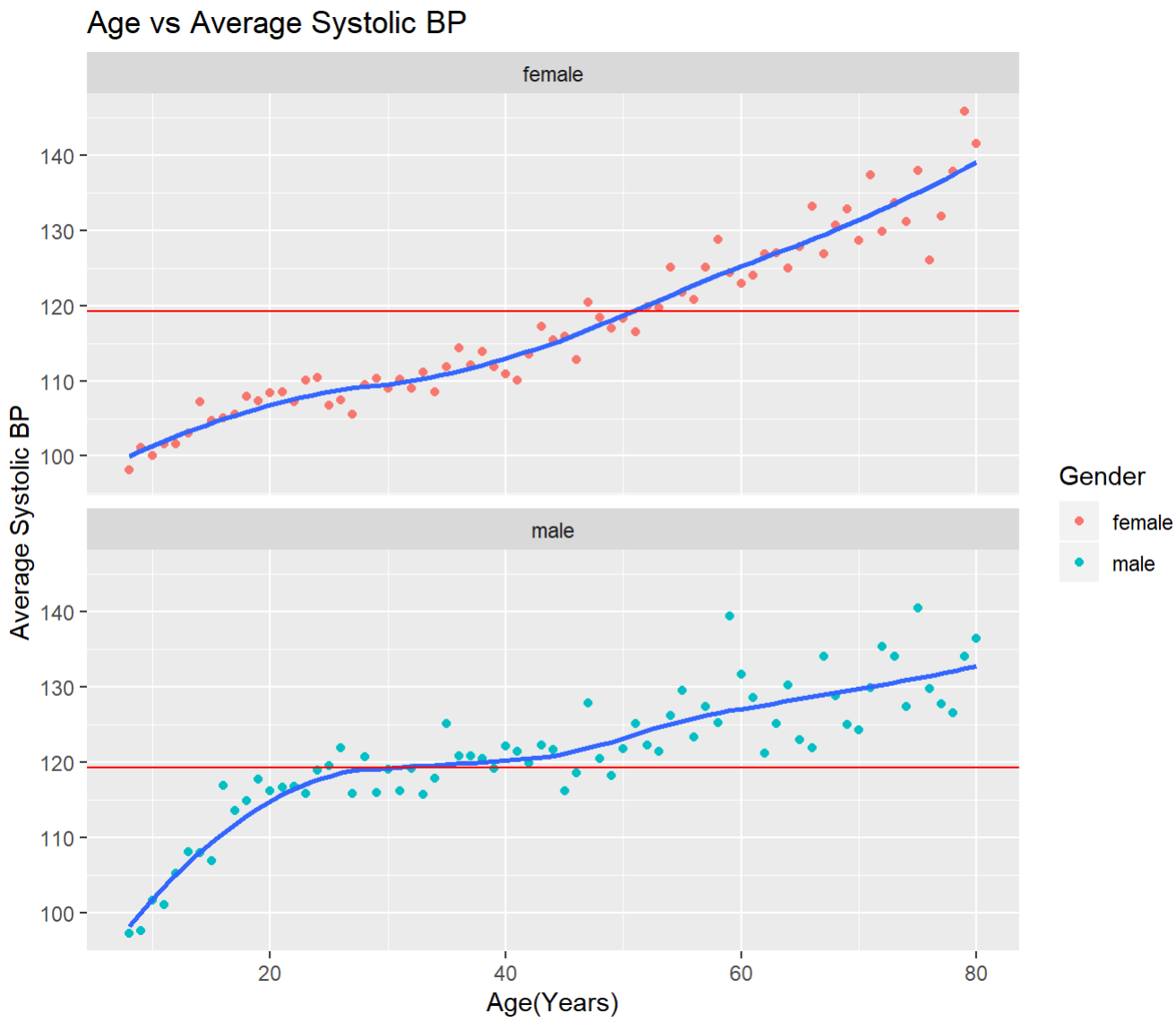
```
#View(NHANES)
workingData = subset(NHANES, select = c("BPSysAve", "Age", "Weight", "Height", "Gender"))
#View(workingData)

ageData = workingData %>% group_by(Gender, Age) %>% summarise(meanBP = mean(BPSysAve, na.rm = TR
UE))
#View(ageData)
ageData = na.omit(ageData)

ggplot(ageData, aes(x= Age, y= meanBP)) + geom_point(aes(color = Gender)) +
    geom_smooth(se = FALSE, span = 0.6) +
    geom_hline(yintercept = mean(ageData$meanBP), color = "red") +
    ggtitle("Age vs Average Systolic BP") +

    xlab("Age(Years)") + ylab("Average Systolic BP") +
    facet_wrap(~ Gender, nrow = 2) +
    xlim(8,80)
```

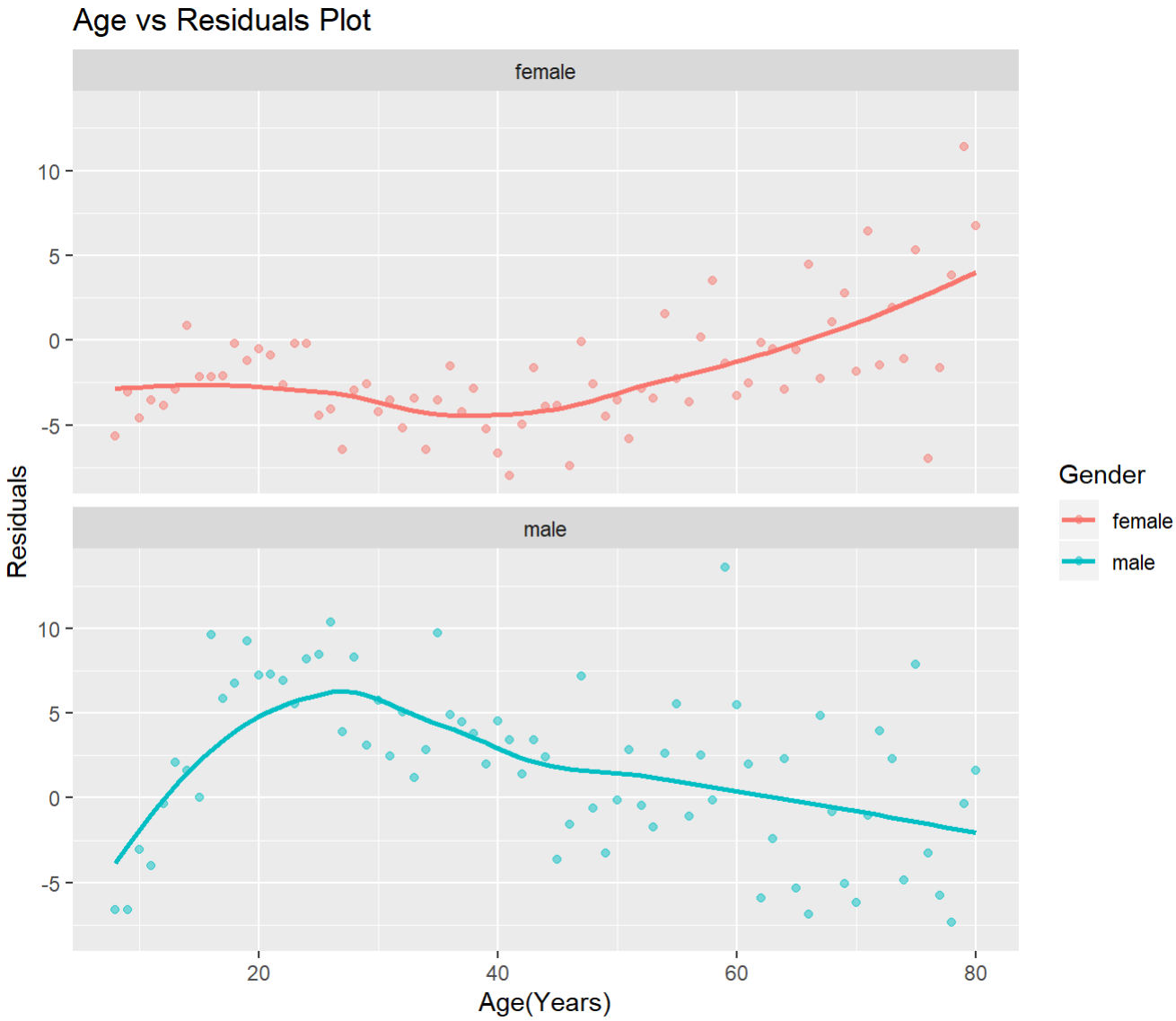
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The graph above shows the relationship between age and BP for women and men separately. From the trend that the data follows, it can be seen pretty clearly that the relation between Age and BP for women can be better displyaed by the loess curve, while for men, the graph has slight curve at younger ages and follows an approximately linear characteristic at older ages. Hence, loess model is the ideal fit for this comparision based on the dataset graph plotted above.

```
age = lm(meanBP ~ Age, data = ageData)
ageResiduals = data.frame(Age = ageData$Age, Gender = ageData$Gender, residual = residuals(age))
ggplot(ageResiduals, aes(x= Age, y= residual, color= Gender)) + geom_point(alpha = 0.5) +
  facet_wrap(~Gender,nrow = 2) + geom_smooth(se = FALSE) + xlab("Age(Years)") + ylab("Residuals"
)+
  ggtitle("Age vs Residuals Plot")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Here we can conclude that we can fit a loess model to the data by observing the residual plots. The residuals for women gender are distributed equally along the regression line, but the same is also valid for relationship for men. In case of men, the distribution of residuals around the line is more spread out as compared to women.

Subsection 2:

The relation between Height and Average Systolic BP:

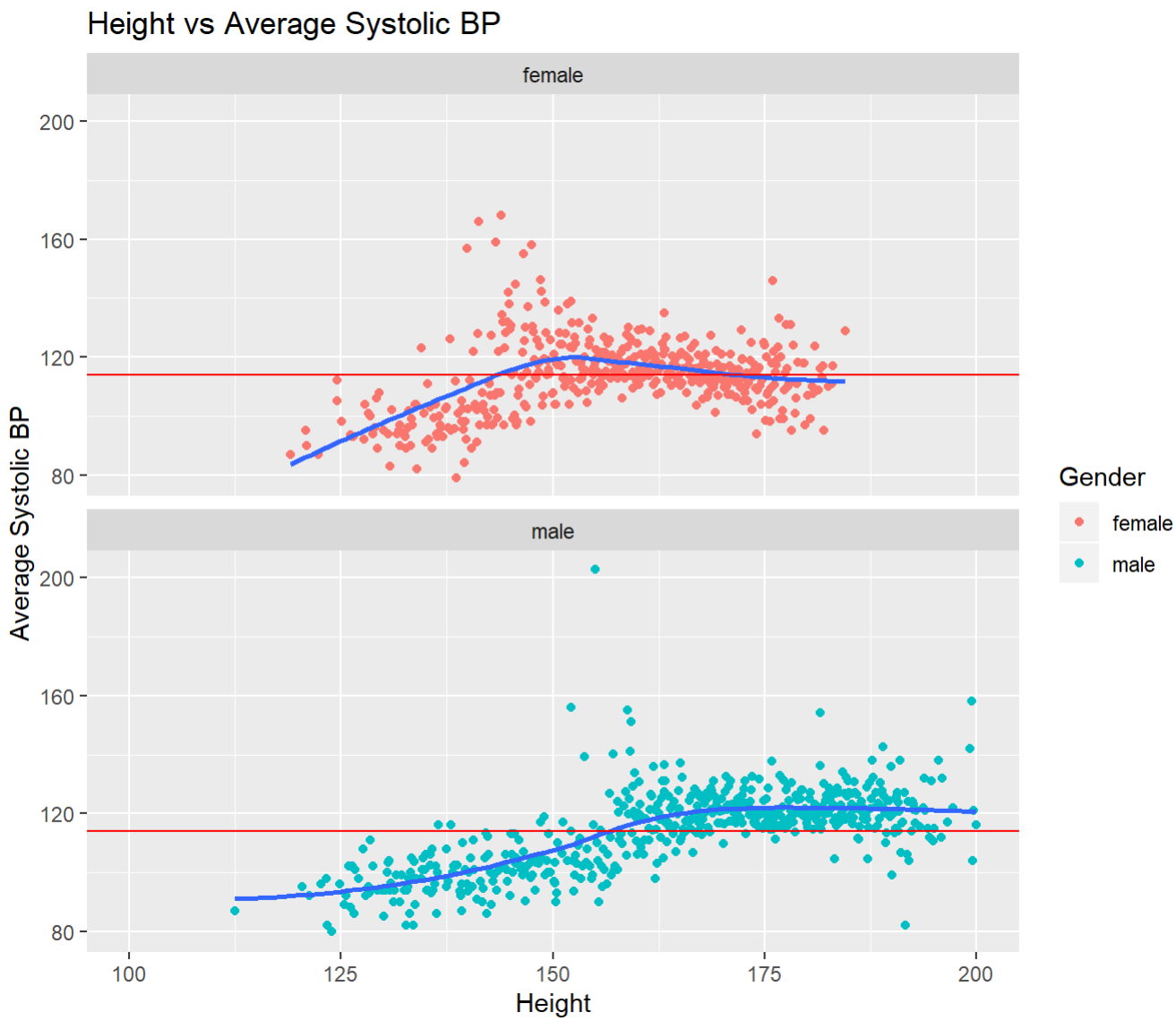
```
heightData = workingData %>%
            group_by(Gender, Height) %>%
              summarise(meanBP = mean(BPSysAve, na.rm = TRUE))
#View(ageData)
heightData = na.omit(heightData)

ggplot(heightData, aes(x= Height, y= meanBP)) + geom_point(aes(color = Gender)) +
    geom_smooth(se = FALSE) +
    geom_hline(yintercept = mean(heightData$meanBP), color = "red") +
    ggtitle("Height vs Average Systolic BP") +
    xlab("Height") + ylab("Average Systolic BP") +
    facet_wrap(~ Gender, nrow = 2) +
    xlim(100,200)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

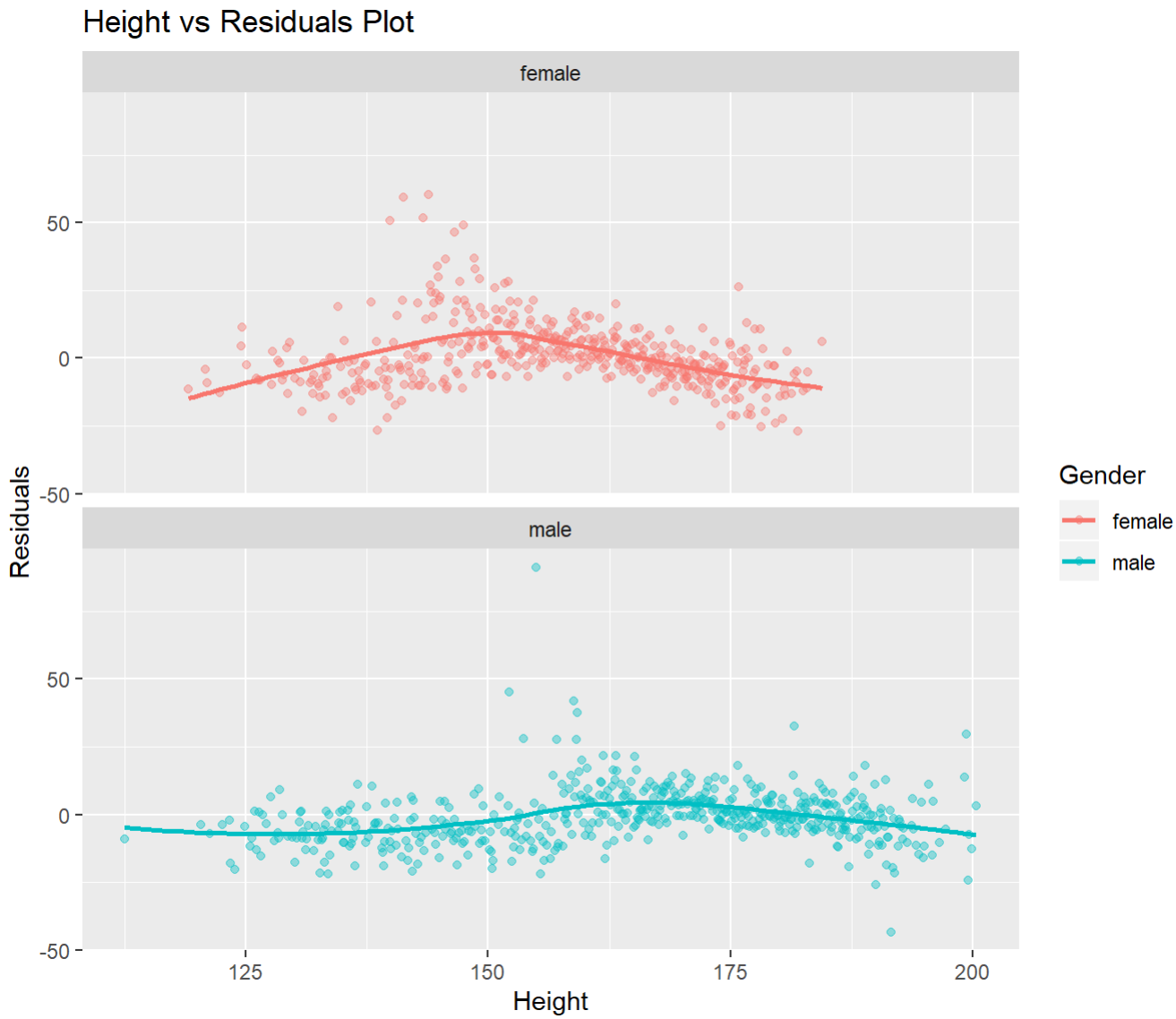
```
## Warning: Removed 1 rows containing missing values (geom_point).
```



A loess model can be fitted to the trend data for the relationship between Height and BP for both men and women. But, the graph for data shows outliers for both men and women for the comparision.The outliers for data provided for women are less extreme as compared to outliers for men.

```
height = lm(meanBP ~ Height, data = heightData)
heightResiduals = data.frame(Height = heightData$Height, Gender = heightData$Gender, residual =
residuals(height))
ggplot(heightResiduals, aes(x= Height, y= residual, color= Gender)) + geom_point(alpha = 0.4) +
  facet_wrap(~Gender,nrow = 2) + geom_smooth(se = FALSE) + xlab("Height") + ylab("Residuals")+
  ggtitle("Height vs Residuals Plot")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The data can be fit perfectly using a loess model. Applying log and square root transformation did not help the graph much. The spread of outliers across the data for women is less severe as compared to men.

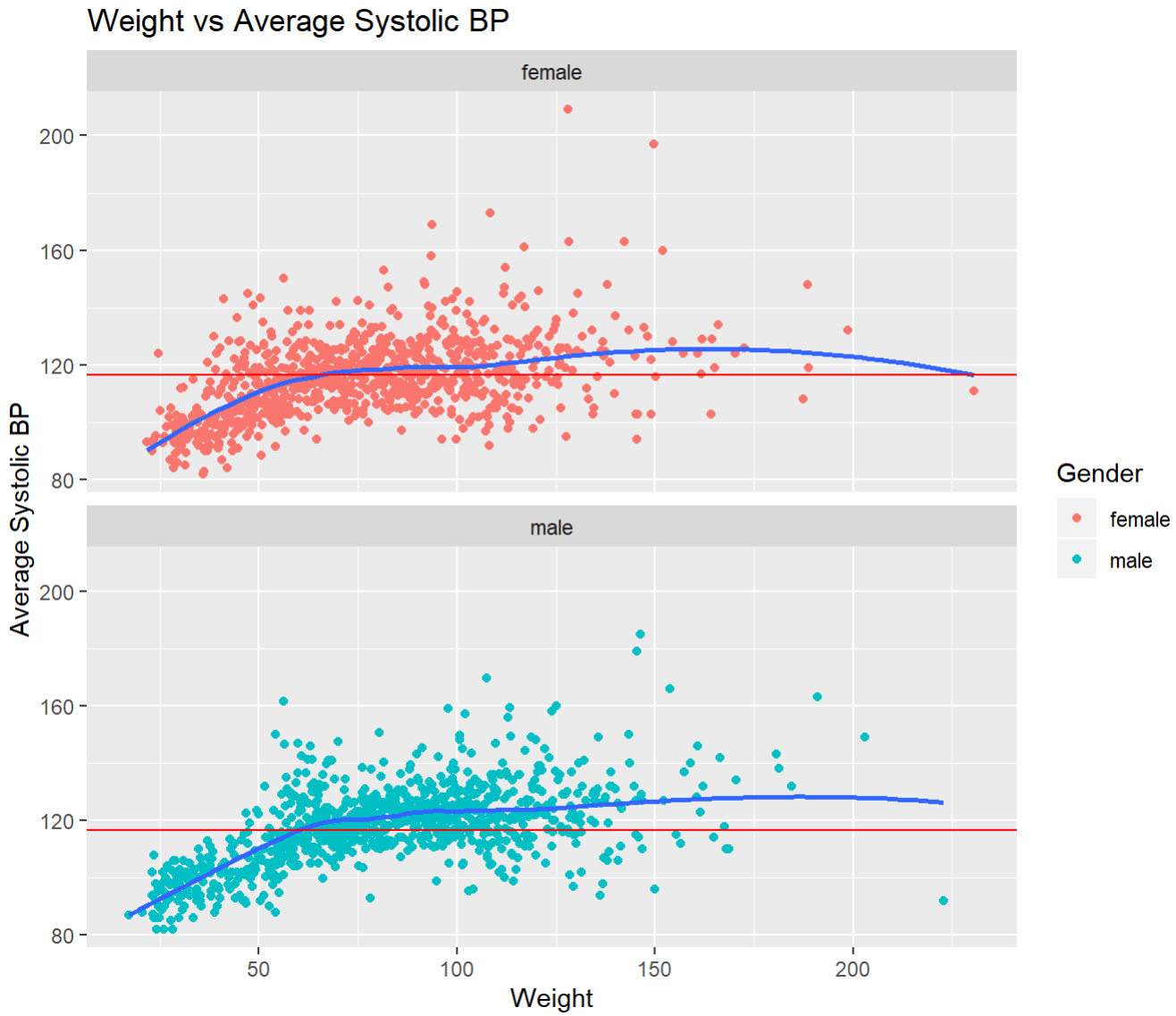
Subsection 3:

The relation between Weight and Average Systolic BP:

```
weightData = workingData %>%
            group_by(Gender, Weight) %>%
              summarise(meanBP = mean(BPSysAve, na.rm = TRUE))
#View(ageData)
weightData = na.omit(weightData)

ggplot(weightData, aes(x= Weight, y= meanBP)) + geom_point(aes(color = Gender)) +
    geom_smooth(se = FALSE, span = 0.6) +
    geom_hline(yintercept = mean(weightData$meanBP), color = "red") +
    ggtitle("Weight vs Average Systolic BP") +
    xlab("Weight") + ylab("Average Systolic BP") +
    facet_wrap(~ Gender, nrow = 2)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The graph above showws relationship between Weight and BP for men and women. In data for women, the outliers are more as compared to the data for men. Hence, a loess model is fitted here, which has a sharp curve at the smaller weights in both the genders, but the graph continues to have a curve even at higher weights.

```
weight = lm(meanBP ~ Weight, data = weightData)
weightResiduals = data.frame(Weight = weightData$Weight, Gender = weightData$Gender, residual =
residuals(weight))
ggplot(weightResiduals, aes(x= Weight, y= residual, color= Gender)) + geom_point(alpha = 0.4) +
  facet_wrap(~Gender,nrow = 2) + geom_smooth(se = FALSE) + xlab("Weight") + ylab("Residuals")+
  ggtitle("Weight vs Residuals Plot")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

The data cannot be fit perfectly using a loess model. The spread of outliers across the data for women is more severe as compared to men. Hence, linear model cannot be applied to the data without much problems and approximations. The log and square root transforms again provide no better results in this case as well.

Conclusions:

The graph for Age vs BP perfectly fits a loess model. Also, for Height and Weight the linear model is not the most optimum choice, this can be concluded from the residual graphs plotted above. But, a loess model can be applied in those cases as well, because it won't lead to a major problem, if and when a formal inference has to be made.