Assignment 3

Himani Anil Deshpande

Loading Data

```
health_data <-NHANES[,c("BPSysAve", "Age", "Weight", "Height", "Gender")]
#view(health_data)
total_samples = nrow(health_data)
total_samples
```

```
## [1] 10000
```

```
health_data %>%  summarize_all(funs(sum(is.na(.))))
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

```
## # A tibble: 1 x 5
##   BPSysAve   Age Weight Height Gender
##      <int> <int>  <int>  <int>  <int>
## 1     1449     0     78    353      0
```

We have a dataset of 10000 rows out of which we have 1449 missing values for BPSysAve and 78 missing values of Weight in Kilograms and 353 missing values for Height in centimeters.

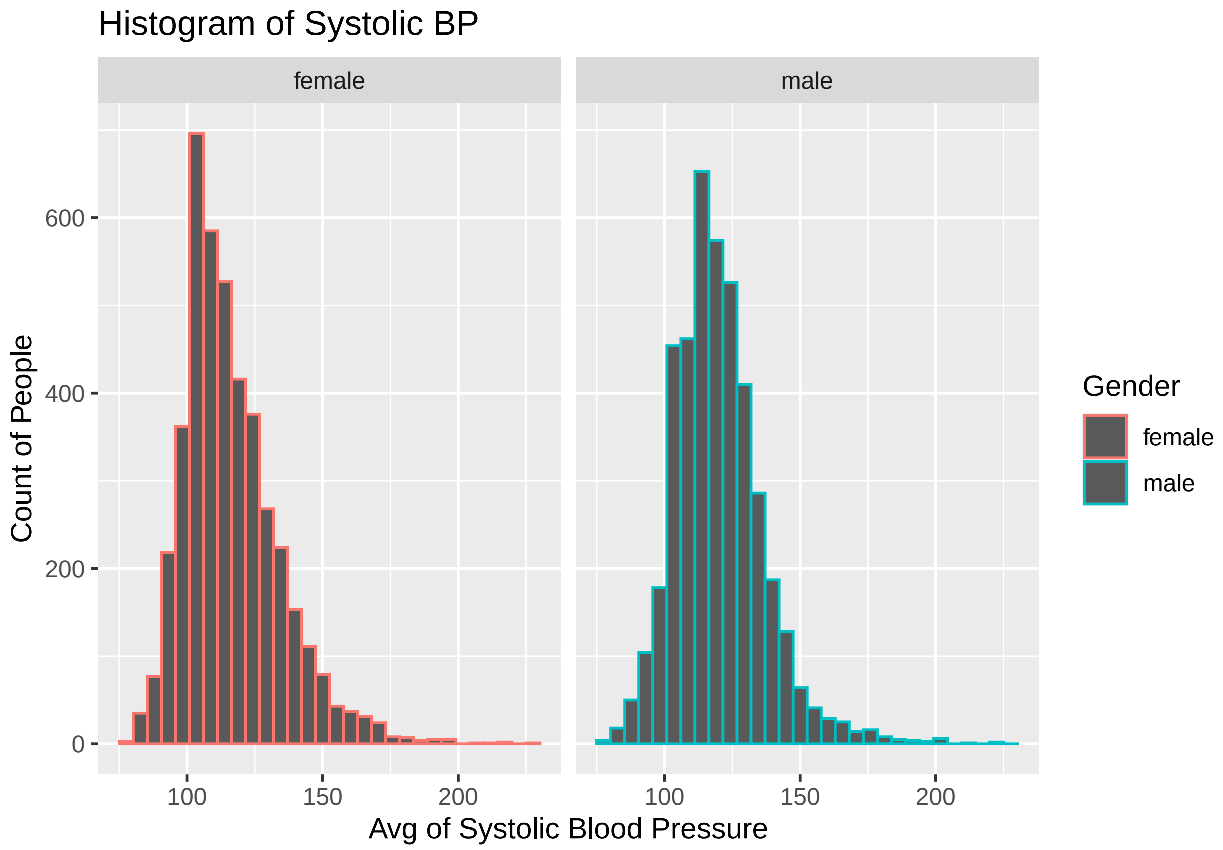
Distribution of Avg Systolic BP

Plotting the Distribution of response variable

```
ggplot(health_data, aes(x  = BPSysAve, color = Gender)) +
  ggtitle("Histogram of Systolic BP ")+
  geom_histogram() +
  xlab("Avg of Systolic Blood Pressure") +
  ylab("Count of People") +
 # scale_colour_manual(values = cbPalette) +
  facet_wrap(~Gender)
```

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

```
## Warning: Removed 1449 rows containing non-finite values (stat_bin).
```



Looks like the data is positively skewed(right skewed)

Plotting to check the Relationship between Avg Systolic Blood pressure and Age

BPSysAve is response variable and Age is the independent variable ## Scatter Plot

I am filtering the data to remove the rows which doesnt have BPSysAve and Age values. The lowest value for BPSysAve is 76 and Age is 0. We can afford removing the missing data as we have a huge dataset of 10000 rows.

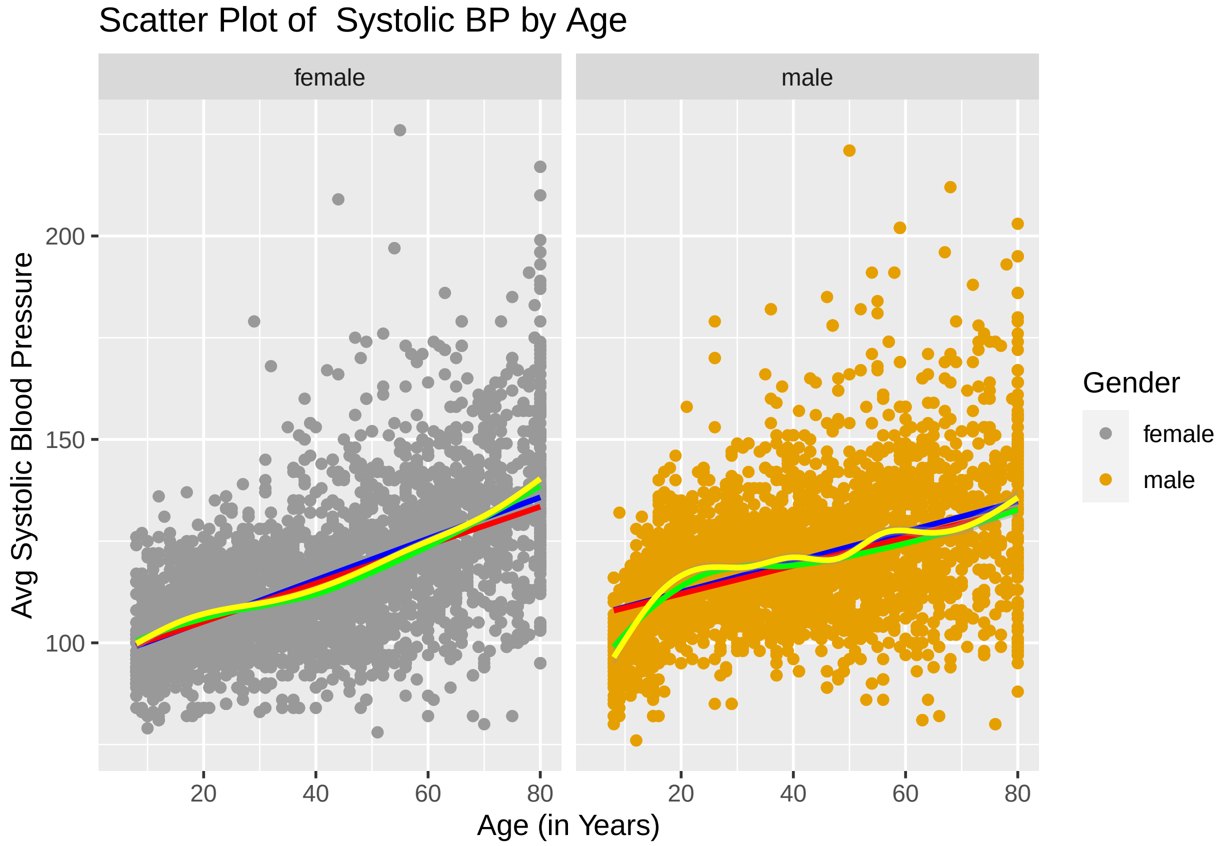
```
health_data_filter_age = filter(health_data, BPSysAve >= 0 , Age >= 0)



ggplot(health_data_filter_age, aes(x = Age, y = BPSysAve, color = Gender)) +
  ggtitle("Scatter Plot of  Systolic BP by Age")+
  xlim(5, 80)+
  ylab("Avg Systolic Blood Pressure") +
  xlab("Age (in Years)") +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  geom_smooth(method = "rlm",method.args = list(psi = psi.bisquare), color = "red") +
  geom_smooth(method = "loess", method.args = list(family = "symmetric"),color = "green") +
  geom_smooth(method = "gam", color = "yellow") +
  scale_colour_manual(values = cbPalette) +
   facet_wrap(~Gender)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'

## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```



From the above plot we can see that there seems to be equal number of samples above and below the lines. We can also say that the Systolic Blood Pressure increases as the age increase. Among the models (lm, rlm , gam and loess), the simplest model, i.e the linear model seems the best fit to describe the relationship between Systolic BP and Age.

Residuals

```
health_data.lm = lm(BPSysAve~Age , data = health_data_filter_age)


health_data.lm.df = data.frame(health_data_filter_age,
                                .resid = residuals(health_data.lm),
                                .fitted = fitted(health_data.lm))
```

Slope and Intercept of population

For Male's

```
coef(lm(BPSysAve~Age,data=health_data_filter_age %>% filter(Gender == "male")))
```

```
## (Intercept)         Age
##  105.097356    0.370387
```

For Female's

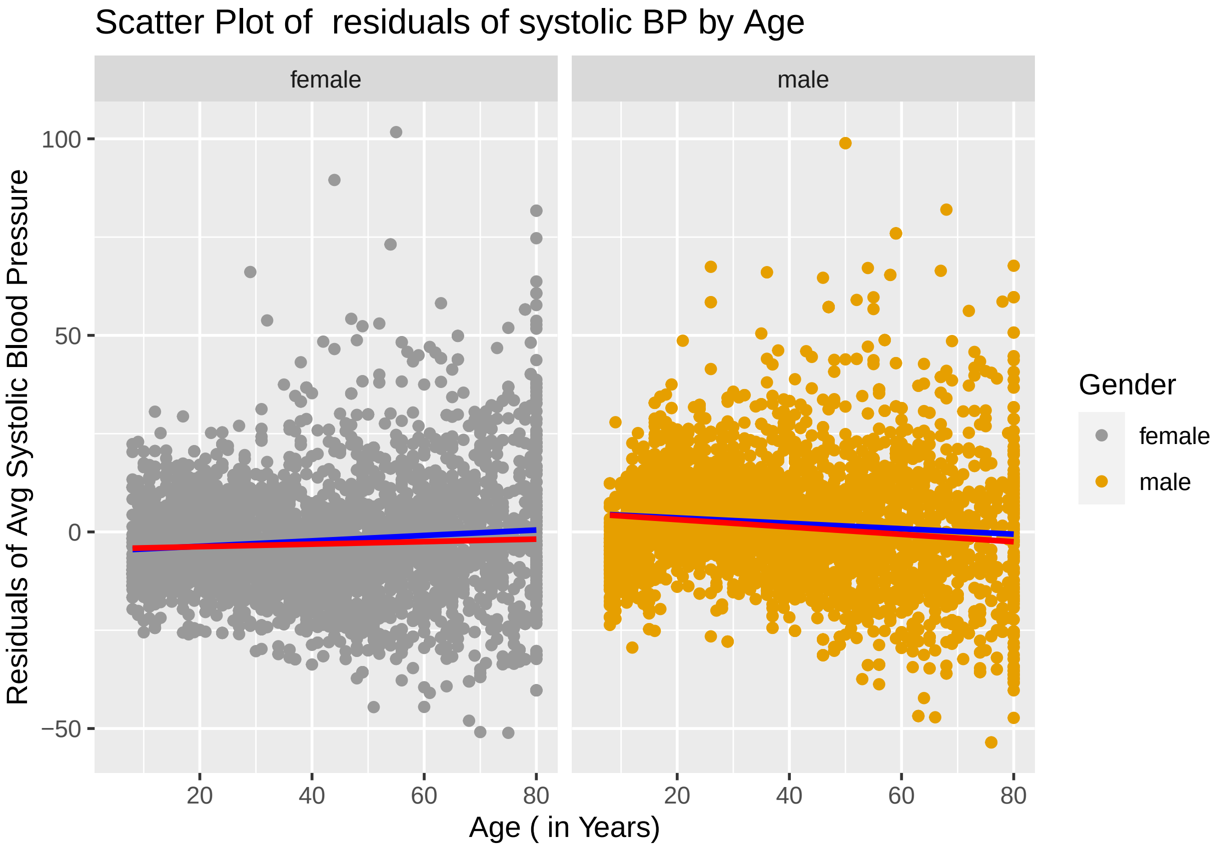
```
coef(lm(BPSysAve~Age,data=health_data_filter_age %>% filter(Gender == "female")))
```

```
## (Intercept)         Age
##  95.1194841   0.5082238
```

These slopes shows that the Female Population have higher increase in systolic Blood Pressure as the age increases compared to Male population

```
ggplot(health_data.lm.df, aes(x = Age, y = .resid, color = Gender)) +
  ggtitle("Scatter Plot of  residuals of systolic BP by Age")+
    xlim(5, 80)+
  ylab("Residuals of Avg Systolic Blood Pressure") +
  xlab("Age ( in Years)") +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  geom_smooth(method = "rlm",method.args = list(psi = psi.bisquare), color = "red", se = FALSE) +
  #geom_smooth(method = "gam", color = "yellow",se = FALSE) +
  #geom_smooth(method = "loess", color = "green",
   #          method.args = list(family = "symmetric"), se = FALSE) +
  scale_colour_manual(values = cbPalette) +
   facet_wrap(~Gender)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



The above plot help us in inferring that the relationship between the two variables is linear. The datapoints are spread equally above and below the line so there is no visible trend in the residual plot

Avg Systolic Blood pressure V/s Height

BPSysAve is response variable and Height is the independent variable ## Scatter Plot

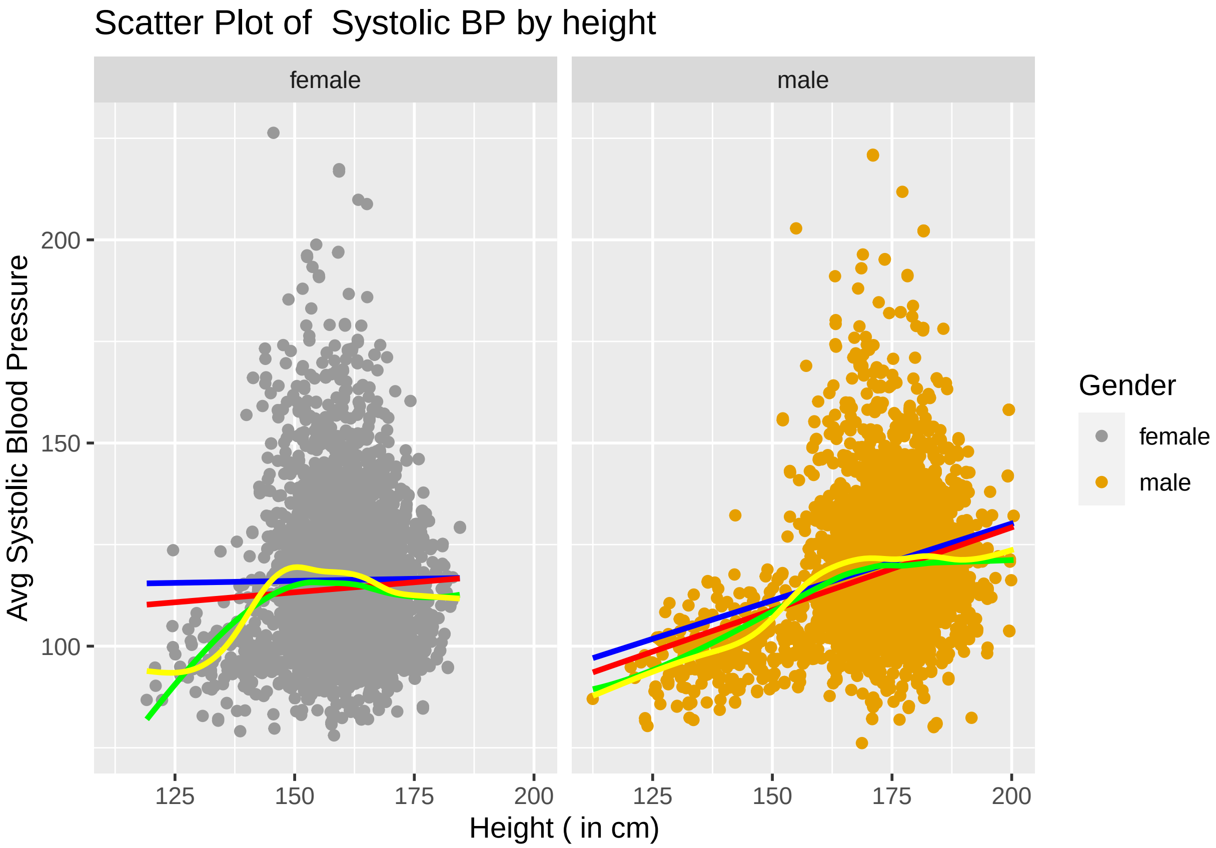
Filttering Data by removing missing values for BP and Height

```
health_data_filter_height = filter(health_data, BPSysAve >= 0 , Height >= 0)

ggplot(health_data_filter_height, aes(x = Height, y = BPSysAve, color = Gender)) +
  ggtitle("Scatter Plot of  Systolic BP by height")+
  ylab("Avg Systolic Blood Pressure") +
  xlab("Height ( in cm)") +
  geom_jitter() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  geom_smooth(method = "rlm",method.args = list(psi = psi.bisquare), color = "red", se = FALSE) +
  geom_smooth(method = "loess", method.args = list(family = "symmetric"),
            color = "green",  se = FALSE) +
  geom_smooth(method= "gam", se = FALSE, color = "yellow") +
  scale_colour_manual(values = cbPalette) +
   facet_wrap(~Gender)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'

## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```



From the above plots we can see that the line for gam's curve increases as we reach the height of 150-165 and then stays constant or at a plateau for Male population but slightly decrease for the female population. The linear model doesnt describe the trend as nicely as the gam model, hence we say that the gam model is better fit for the two variable relationship.

Residuals

```
health_data.height.gam = gam(BPSysAve~Height , data = health_data_filter_height)

health_data.height.gam
```

```
## Call:
## gam(formula = BPSysAve ~ Height, data = health_data_filter_height)
##
## Degrees of Freedom: 8498 total; 8497 Residual
## Residual Deviance: 2432946
```

```
health_data.height.gam.df = data.frame(health_data_filter_height,
                              .resid = residuals(health_data.height.gam),
                               .fitted = fitted(health_data.height.gam))
```

Slope and Intercept of population

For Male's

```
coef(gam(BPSysAve~Height,data=health_data_filter_height %>% filter(Gender == "male")))
```

```
## (Intercept)      Height
##  54.4963159   0.3784228
```

For Female's

```
coef(gam(BPSysAve~Height,data=health_data_filter_height %>% filter(Gender == "female")))
```

```
##  (Intercept)       Height
## 113.06813856   0.02008559
```

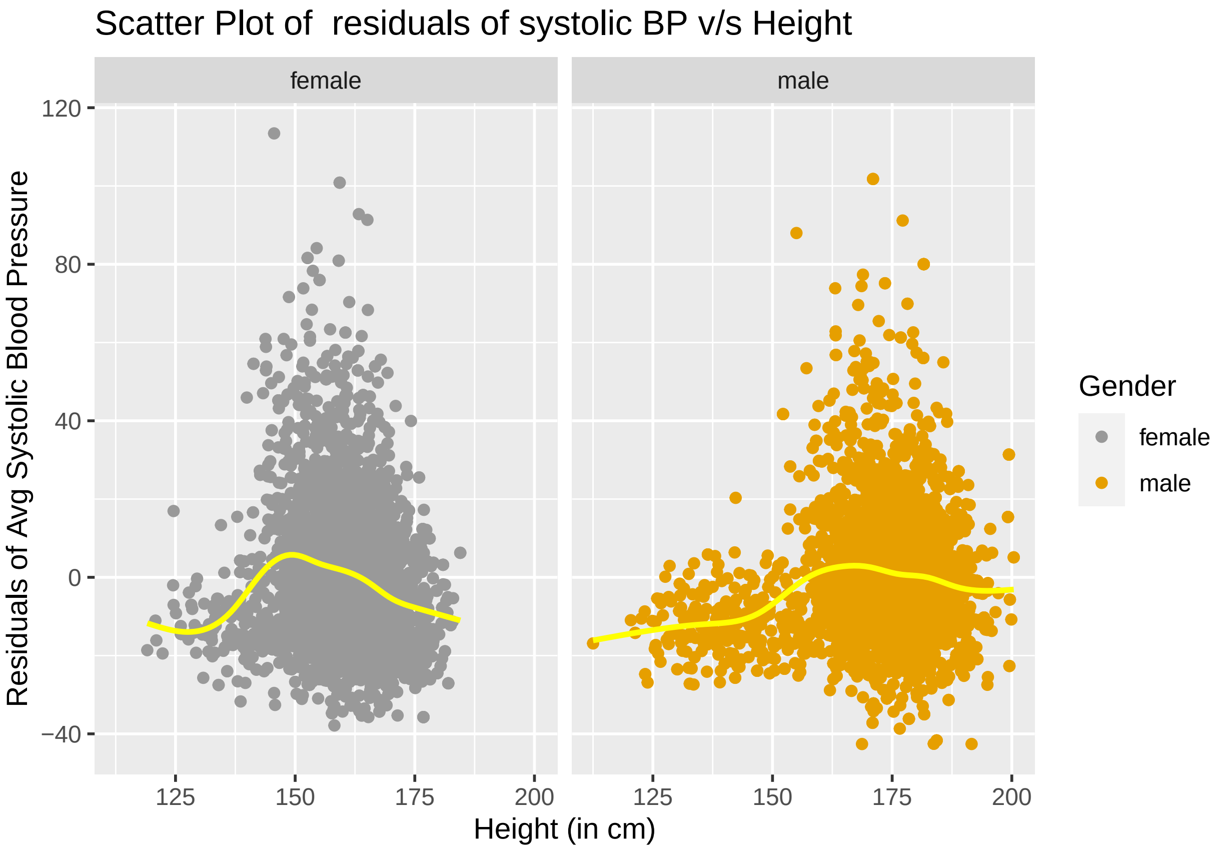
The Slope of the Male population is higher than the female population. It seems that the

```
ggplot(health_data.height.gam.df, aes(x =Height, y = .resid, color = Gender)) +
  ggtitle("Scatter Plot of  residuals of systolic BP v/s Height")+
  ylab("Residuals of Avg Systolic Blood Pressure") +
  xlab("Height (in cm)") +
  geom_point() +

  geom_smooth(method = "gam", color = "yellow", se= FALSE)+

  scale_colour_manual(values = cbPalette) +
   facet_wrap(~Gender)
```

```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```



From the plots, we can see that The data points in graph of male is more spread out then female. One main reason for that is that range of height in male is larger than range of height in female. The gam model fits the data population better than any other models

Avg Systolic Blood pressure V/s Weight

BPSysAve is response variable and Weight is the independent variable ## Scatter Plot

Filtering Data to remove missing BP and Weight values

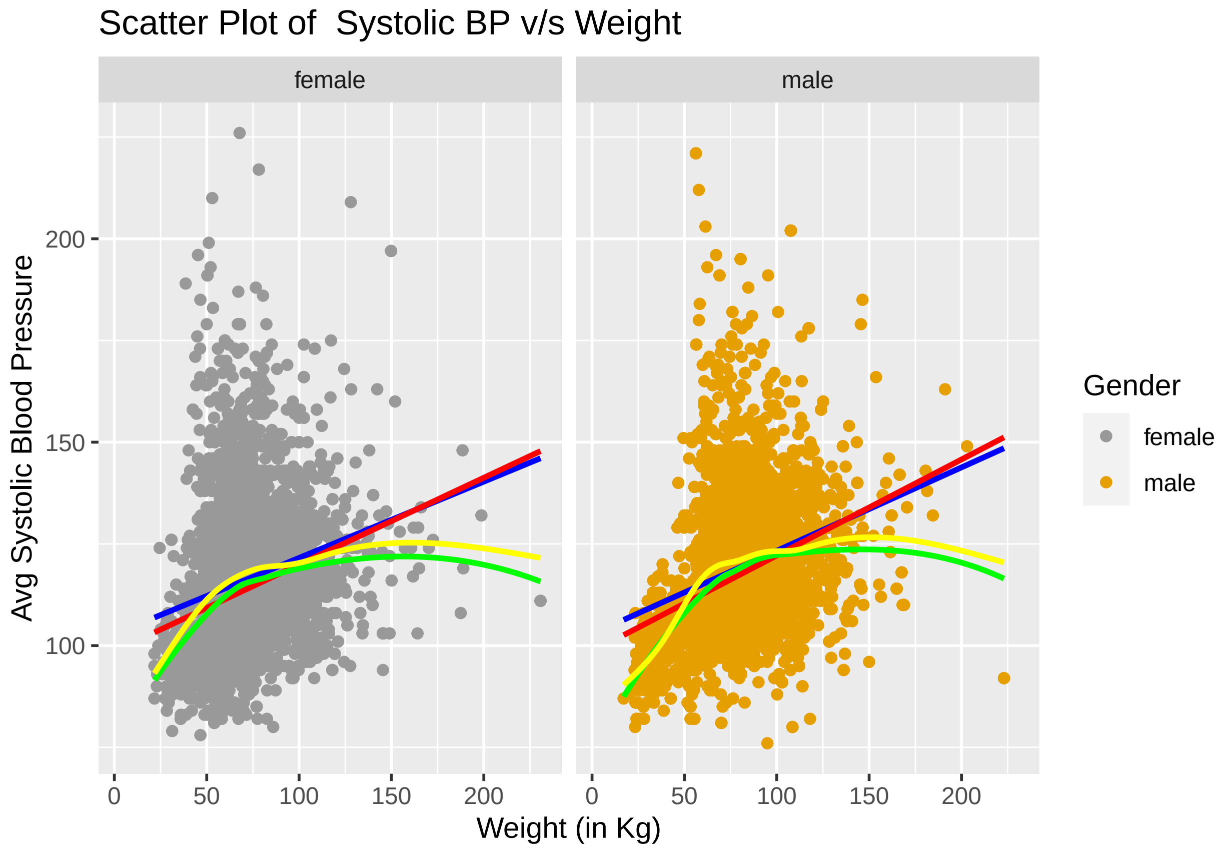
```
health_data_filter_weight = filter(health_data, BPSysAve >= 0 , Weight >= 0)

ggplot(health_data, aes(x = Weight, y = BPSysAve, color = Gender)) +
  ggtitle("Scatter Plot of  Systolic BP v/s Weight")+
  ylab("Avg Systolic Blood Pressure") +
  xlab("Weight (in Kg)") +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  geom_smooth(method = "rlm",method.args = list(psi = psi.bisquare), color = "red", se = FALSE) +
  geom_smooth(method = "loess", method.args = list(family = "symmetric"),
            color = "green",  se = FALSE) +
  geom_smooth(method= "gam", se = FALSE, color = "yellow") +
  scale_colour_manual(values = cbPalette) +
   facet_wrap(~Gender)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 1507 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 1507 rows containing non-finite values (stat_smooth).
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 1507 rows containing non-finite values (stat_smooth).
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
## Warning: Removed 1507 rows containing non-finite values (stat_smooth).
## Warning: Removed 1507 rows containing missing values (geom_point).
```



From the above graphs, we can see that loess fits the data much better than any other model. For both the Gender population, the loess's curve shows and increase in BP as the weight reaches in the range of 60 to 100 then after that it stays constant with a slight dip at the end, which could be dur to a few outliers.

Residuals

```
health_data.weight.lo = loess(BPSysAve~Weight , data = health_data_filter_weight)

health_data.weight.lo
```

```
## Call:
## loess(formula = BPSysAve ~ Weight, data = health_data_filter_weight)
##
## Number of Observations: 8493
## Equivalent Number of Parameters: 5.56
## Residual Standard Error: 16.29
```

```
health_data.weight.lo.df = data.frame(health_data_filter_weight,
                                 .resid = residuals(health_data.weight.lo),
                                .fitted = fitted(health_data.weight.lo))
```

Statistical Measure of population

For Male's

```
loess(BPSysAve~Weight,data=health_data_filter_weight %>% filter(Gender == "male"))
```

```
## Call:
## loess(formula = BPSysAve ~ Weight, data = health_data_filter_weight %>%
##     filter(Gender == "male"))
##
## Number of Observations: 4220
## Equivalent Number of Parameters: 5.59
## Residual Standard Error: 15.29
```

For Female's

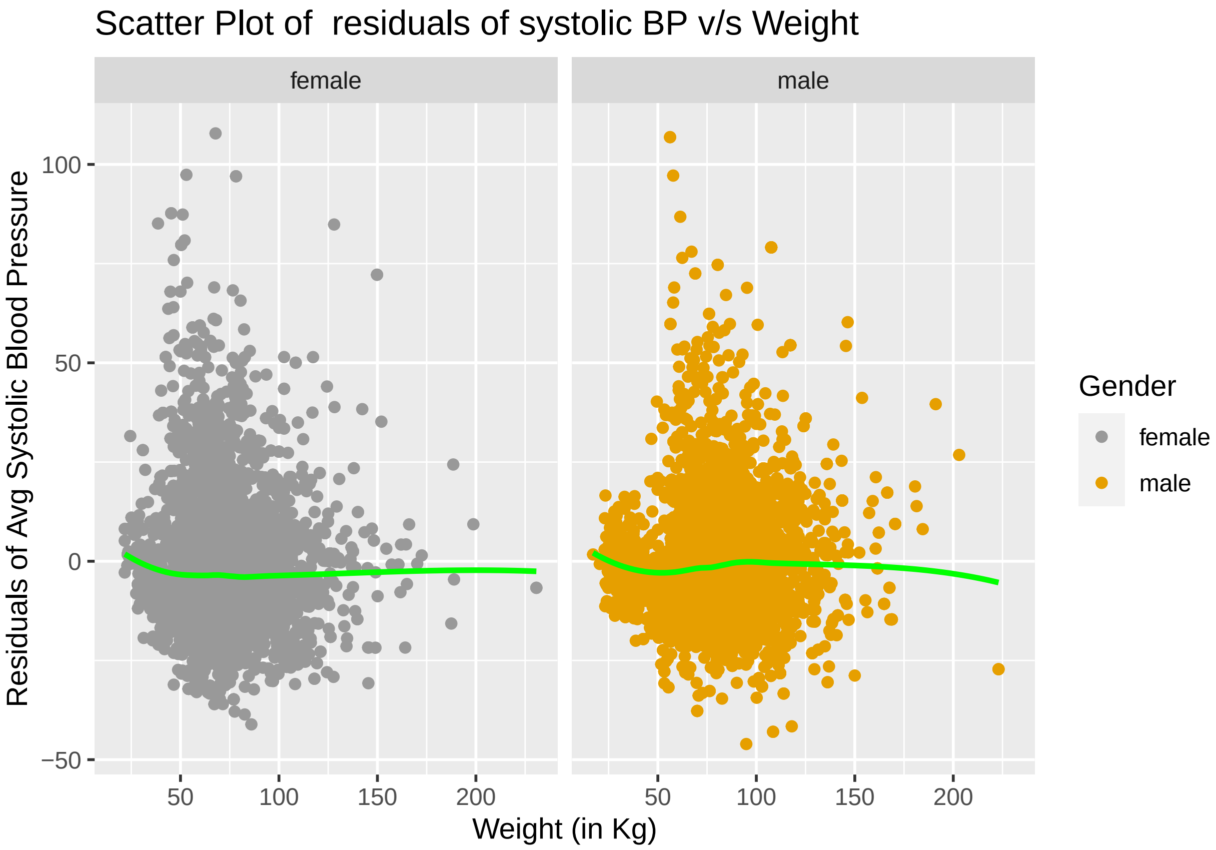
```
loess(BPSysAve~Weight,data=health_data_filter_weight %>% filter(Gender == "female"))
```

```
## Call:
## loess(formula = BPSysAve ~ Weight, data = health_data_filter_weight %>%
##     filter(Gender == "female"))
##
## Number of Observations: 4273
## Equivalent Number of Parameters: 5.62
## Residual Standard Error: 17.18
```

```
ggplot(health_data.weight.lo.df, aes(x = Weight, y = .resid, color = Gender)) +
  ggtitle("Scatter Plot of  residuals of systolic BP v/s Weight")+
  ylab("Residuals of Avg Systolic Blood Pressure") +
  xlab("Weight (in Kg)") +
  geom_point() +

  geom_smooth(method = "loess", color = "green",
              method.args = list(family = "symmetric"), se = FALSE) +
  scale_colour_manual(values = cbPalette) +
   facet_wrap(~Gender)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The residuals are equally scattered around the line. So the fit is reasonably good since there is no pattern in the residuals.