Assignment 1

Manning Smith

9/9/2024

Prompt

1. structure of the data set, including length, data type, names, and the components of the data set.
2. Average Measures for continuous variables, including Mean, Median, Standard deviation, Variance, Median Absolute Variance (deviation), maximum, Minimum, and Sum.
3. Frequency for categorical variables.
4. Plots of summary statistics (such as histograms or bar-plot) for one continuous variable and categorical variable respectively.

Data Set Information

- PatientID: unique patient identifier
- Age: age in years
- Sex: 1 = male; 0 = female
- ChestPain: chest pain type
- RestBP: resting blood pressure (in mm Hg on admission to the hospital)
- Chol: serum cholestoral in mg/dl
- Fbs: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- RestECG: resting electrocardiographic results. 0: normal ; 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) ; 2: showing probable or definite left ventricular hypertrophy by Estes' criteria.
- MaxHR: maximum heart rate achieved
- ExAng: exercise induced angina (1 = yes; 0 = no)
- Oldpeak: ST depression induced by exercise relative to rest
- Slope: the slope of the peak exercise ST segment. 1: upsloping; 2: flat; 3: downsloping
- Ca: number of major vessels (0-3) colored by flourosopy
- Thal: Thallium stress test, 3 =normal; 6 = fixed (defect); 7 = reversable (defect)
- AHD (the predicted attribute): angiographic heart disease

Import Data

```
Heart <- read.csv("Heart.csv")

#View(Heart)
```

Upon the initial look of the data set it contains information related to patient heart data. Both sex and age are provided. The sex of the patient is provided by (1,0), but information is given that 1 is male and 0 is female. There are some NA values for the 'Ca' variable, but this should not affect the data set. Overall, the data set looks pretty good to proceed with.

Question 1 | Structure

```r
# Structure Function
str(Heart)
```

```
## 'data.frame':    303 obs. of  15 variables:
##  $ PatientID: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Age      : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ Sex      : int  1 1 1 1 0 1 0 0 1 1 ...
##  $ ChestPain: chr  "typical" "asymptomatic" "asymptomatic" "nonanginal" ...
##  $ RestBP   : int  145 160 120 130 130 120 140 120 130 140 ...
##  $ Chol     : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ Fbs      : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ RestECG  : int  2 2 2 0 2 0 2 0 2 2 ...
##  $ MaxHR    : int  150 108 129 187 172 178 160 163 147 155 ...
##  $ ExAng    : int  0 1 1 0 0 0 0 1 0 1 ...
##  $ Oldpeak  : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ Slope    : int  3 2 2 3 1 1 3 1 2 3 ...
##  $ Ca       : int  0 3 2 0 0 0 2 0 1 0 ...
##  $ Thal     : chr  "fixed" "normal" "reversable" "normal" ...
##  $ AHD      : chr  "No" "Yes" "Yes" "No" ...
```

```r
# Column names
cols <- colnames(Heart)
#print(cols)
```

The structure of the data set can be seen above. There are 303 observations and 15 variables, most of the variables are integers, but there is one ID variable, 3 characters variable, and one numerical (meaning there are decimals).

Continuous Variables:
`Age`, `RestBP`, `Chol`, `MaxHR`, `OldPeak`

Categorical Variables:
`ChestPain`, `RestECG`, `Slope`, `Ca`, `Thal`, `AHD`

`ChestPain` is a categorical data point with the options of: 'typical', 'asymptomatic', 'nonanginal', and 'nontypical'.

`Thal` stands for Thallium stress test is also a categorical data point with the options of: 'fixed', 'normal', and 'reversible'.

`AHD` is a YES/NO binary data point.

Binary Variables:
`Sex`, `Fbs`, `ExAng`

Question 2 | Summary Statistics

```r
# Select all numerical data points
continuous_vars <- c("Age", "RestBP", "Chol", "MaxHR", "Oldpeak")

# Provide Summary
#summary(Heart)

# Group all results in a dataframe

summary_stats <- data.frame(
  Mean = sapply(Heart[continuous_vars], mean, na.rm = TRUE), # Mean
  Median = sapply(Heart[continuous_vars], median, na.rm = TRUE), # Median
  Std_Dev = sapply(Heart[continuous_vars], sd, na.rm = TRUE), # Standard deviation
  Variance = sapply(continuous_vars, function(x) var(Heart[[x]], na.rm = TRUE)), # Variance
  MAD = sapply(Heart[continuous_vars], mad, na.rm = TRUE), # Median
  Min = sapply(Heart[continuous_vars], min, na.rm = TRUE), # Minimum
  Max = sapply(Heart[continuous_vars], max, na.rm = TRUE), # Maximum
  Sum = sapply(Heart[continuous_vars], sum, na.rm = TRUE) # Sum
)

# Print Pretty Table
kable(summary_stats, caption = "Summary Statistics for Continuous Variables", digits = 5)
```

Table 1: Summary Statistics for Continuous Variables

	Mean	Median	Std_Dev	Variance	MAD	Min	Max	Sum
Age	54.43894	56.0	9.03866	81.69742	8.89560	29	77.0	16495
RestBP	131.68977	130.0	17.59975	309.75112	14.82600	94	200.0	39902
Chol	246.69307	241.0	51.77692	2680.84919	47.44320	126	564.0	74748
MaxHR	149.60726	153.0	22.87500	523.26577	22.23900	71	202.0	45331
Oldpeak	1.03960	0.8	1.16108	1.34810	1.18608	0	6.2	315

Table 2: Frequency Table for Sex

Var1	Freq
0	97
1	206

Table 3: Frequency Table for Chest Pain

Var1	Freq
asymptomatic	144
nonanginal	86
nontypical	50
typical	23

Question 3 | Frequency

```
# Select all categorical data points
categorical_vars <- c("Sex", "ChestPain", "Fbs", "RestECG", "ExAng", "Slope", "Ca", "Thal", "AHD")

freq_tables1 <- lapply(Heart[categorical_vars[1]], table)
kable(freq_tables1, caption = "Frequency Table for Sex", digits = 2)

freq_tables2 <- lapply(Heart[categorical_vars[2]], table)
kable(freq_tables2, caption = "Frequency Table for Chest Pain", digits = 2)

freq_tables3 <- lapply(Heart[categorical_vars[3]], table)
kable(freq_tables3, caption = "Frequency Table for Fasting Blood Sugar", digits = 2)

freq_tables4 <- lapply(Heart[categorical_vars[4]], table)
kable(freq_tables4, caption = "Frequency Table for Resting Electrocardiographic Result", digits = 2)

freq_tables5 <- lapply(Heart[categorical_vars[5]], table)
kable(freq_tables5, caption = "Frequency Table for Exercise Induced Angina", digits = 2)

freq_tables6 <- lapply(Heart[categorical_vars[6]], table)
kable(freq_tables6, caption = "Frequency Table for Slope", digits = 2)

freq_tables7 <- lapply(Heart[categorical_vars[7]], table)
kable(freq_tables7, caption = "Frequency Table for Number of Vessels", digits = 2)

freq_tables8 <- lapply(Heart[categorical_vars[8]], table)
kable(freq_tables8, caption = "Frequency Table for Thallium Stress Test", digits = 2)

freq_tables9 <- lapply(Heart[categorical_vars[9]], table)
kable(freq_tables9, caption = "Frequency Table for AHD", digits = 2)
```

Table 4: Frequency Table for Fasting Blood Sugar

Var1	Freq
0	258
1	45

Table 5: Frequency Table for Resting Electrocardiographic Result

Var1	Freq
0	151
1	4
2	148

Table 6: Frequency Table for Exercise Induced Angina

Var1	Freq
0	204
1	99

Table 7: Frequency Table for Slope

Var1	Freq
1	142
2	140
3	21

Table 8: Frequency Table for Number of Vessels

Var1	Freq
0	176
1	65
2	38
3	20

Table 9: Frequency Table for Thallium Stress Test

Var1	Freq
fixed	18
normal	166
reversable	117

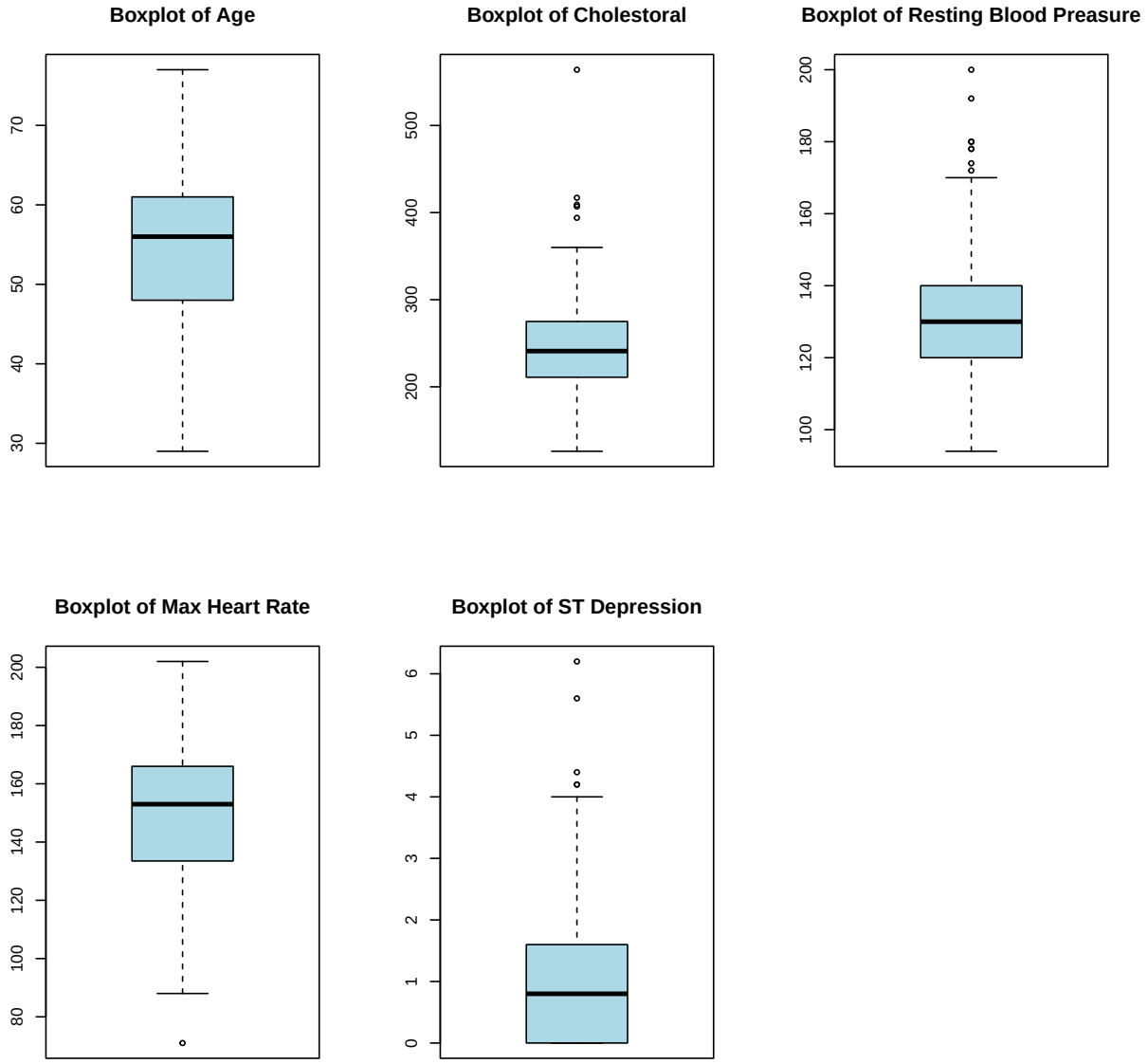
Table 10: Frequency Table for AHD

Var1	Freq
No	164
Yes	139

Question 4 | Plots

Continuous

```r
par(mfrow = c(2, 3))

# Age
#hist(Heart$Age, main = "Histogram of Age", xlab = "Age", col = "lightblue", border = "black")
boxplot(Heart$Age, main = "Boxplot of Age", col = "lightblue", border = "black", horizontal = FALSE)

# Chol
#hist(Heart$Chol, main = "Histogram of Cholestoral", xlab = "Cholestoral", col = "lightblue", border =
boxplot(Heart$Chol, main = "Boxplot of Cholestoral", col = "lightblue", border = "black", horizontal =

# RestBP
#hist(Heart$RestBP, main = "Histogram of Resting Blood Preasure", xlab = "RestBP", col = "lightblue", b
boxplot(Heart$RestBP, main = "Boxplot of Resting Blood Preasure", col = "lightblue", border = "black",

# MaxHR
#hist(Heart$MaxHR, main = "Histogram of Max Heart Rate", xlab = "MaxHR", col = "lightblue", border = "b
boxplot(Heart$MaxHR, main = "Boxplot of Max Heart Rate", col = "lightblue", border = "black", horizontal


# Oldpeak
#hist(Heart$Oldpeak, main = "Histogram of ST Depression", xlab = "Oldpeak", col = "lightblue", border =
boxplot(Heart$Oldpeak, main = "Boxplot of ST Depression", col = "lightblue", border = "black", horizonta
```

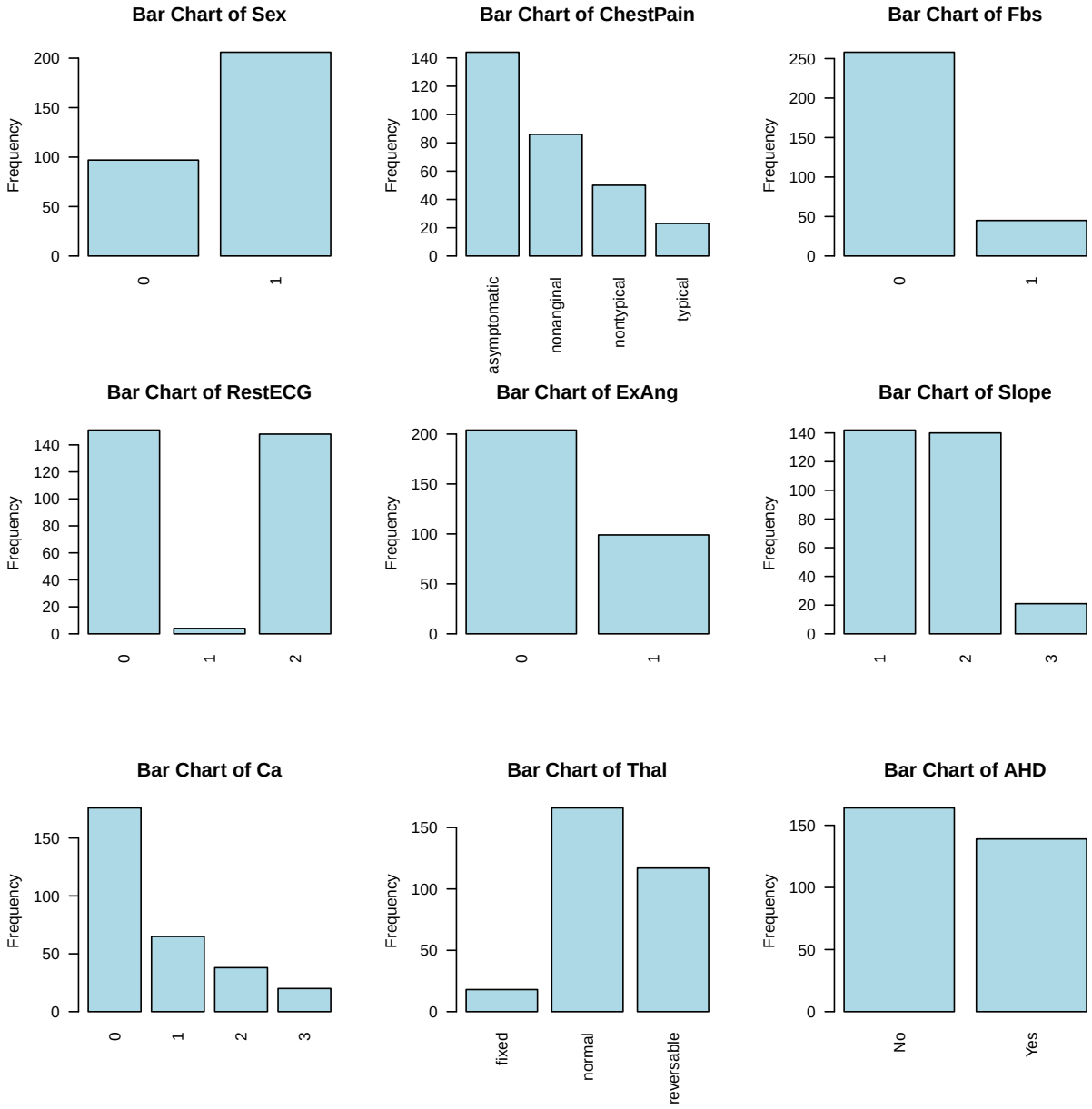


The `Age` variable is normally distributed between ages 29 and 77 with no apparent outliers.
The `Chol` variable is normally distributed with a few outliers on the higher end.
The `RestBP` variable is normally distributed with a few outliers on the higher end.
The `MaxHR` variable is normally distributed with one outlier on the low end.
The `Oldpeak` variable is right skewed with most of the data falling between 0 & 2 with data points as far as 6.

Categorical

```
par(mfrow = c(3, 3))

freq_tables <- lapply(Heart[categorical_vars], table)
```

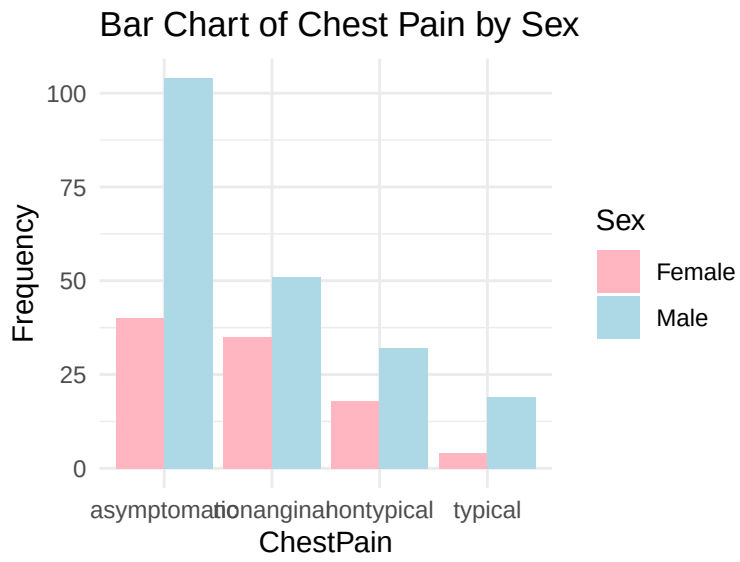
```r
plot_bar_chart <- function(variable) {
  freq_table <- table(Heart[[variable]])
  barplot(freq_table, main = paste("Bar Chart of", variable), col = "lightblue", border = "black",
          ylab = "Frequency", las = 2)
}

# Plot bar charts for each categorical variable
for (var in categorical_vars)
  plot_bar_chart(var)
```



The data set contains twice as many data points for males than females. This could be an issue when using all the data to predict female outcomes. You may only be able to generalize based on the subset gender populations.

Factor by Sex

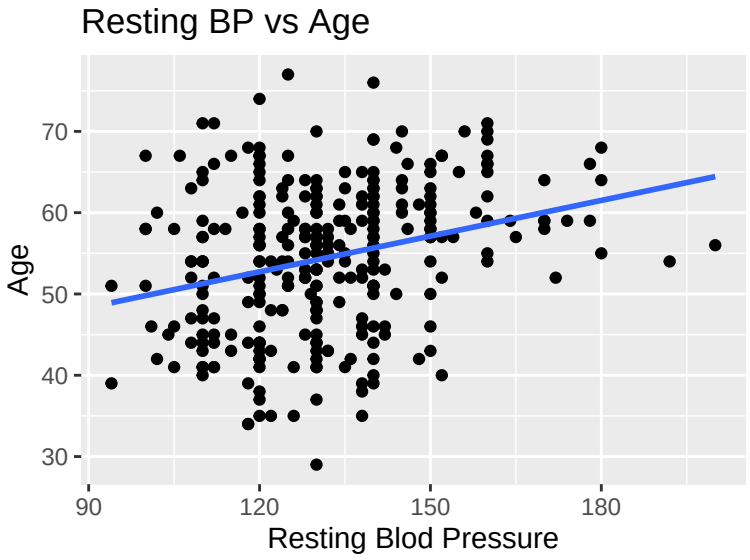
```
Heart$Sex <- factor(Heart$Sex, levels = c(0, 1), labels = c("Female", "Male"))

ggplot(Heart, aes(x = ChestPain, fill = Sex)) +
    geom_bar(position = "dodge") +
    labs(title = "Bar Chart of Chest Pain by Sex", x = "ChestPain", y = "Frequency") +
    scale_fill_manual(values = c("Female" = "lightpink", "Male" = "lightblue")) +
    theme_minimal()
```



Personal Exploration

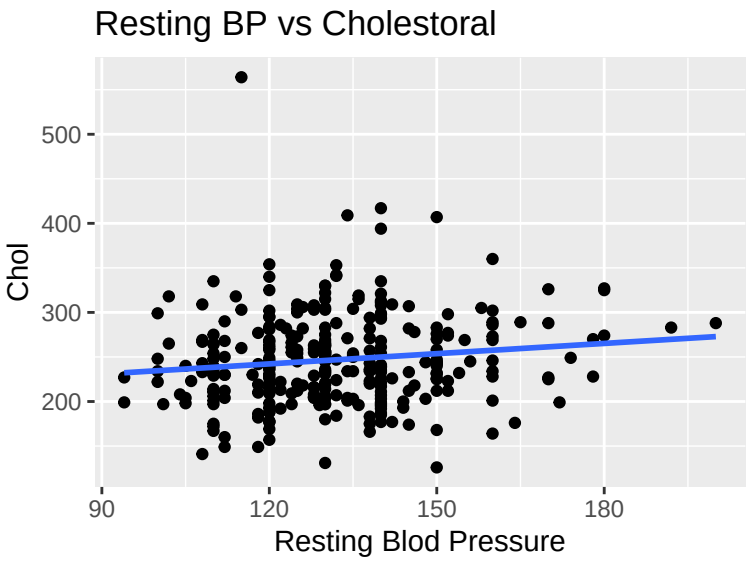
```
ggplot(Heart, aes(x = RestBP, y = Age)) +
  geom_point() +
  labs(x = "Resting Blod Pressure", y = "Age", title = "Resting BP vs Age") +
  geom_smooth(method = "lm", se = FALSE)
```

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



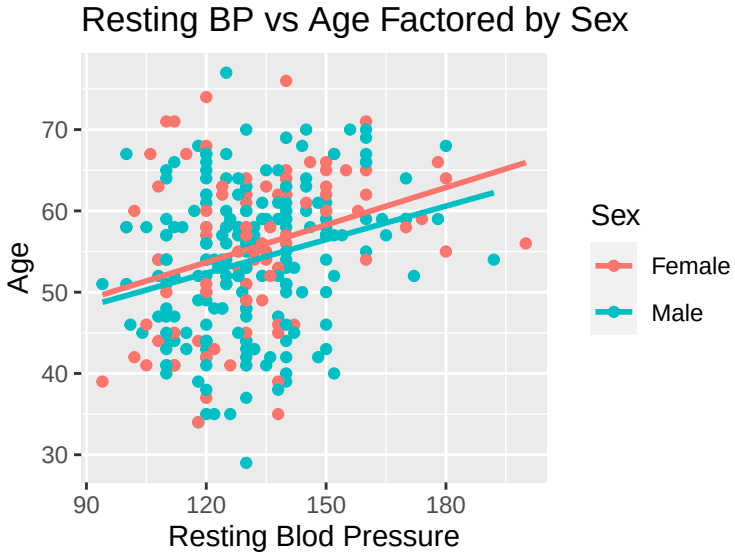

```
ggplot(Heart, aes(x = RestBP, y = Chol)) +
  geom_point() +
  labs(x = "Resting Blod Pressure", y = "Chol", title = "Resting BP vs Cholestoral") +
  geom_smooth(method = "lm", se = FALSE)
```

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



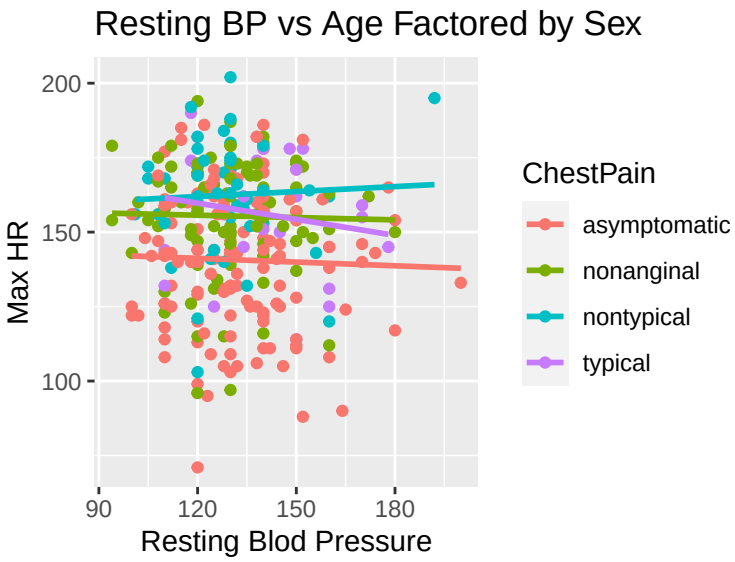

```
ggplot(Heart, aes(x = RestBP, y = Age, color = Sex)) +
  geom_point() +
  labs(x = "Resting Blod Pressure", y = "Age", colored = "Sex", title = "Resting BP vs Age Factored by S
  geom_smooth(method = "lm", se = FALSE)
```

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




```
ggplot(Heart, aes(x = RestBP, y = MaxHR, color = ChestPain)) +
  geom_point() +
  labs(x = "Resting Blod Pressure", y = "Max HR", colored = "Sex", title = "Resting BP vs Age Factored b
  geom_smooth(method = "lm", se = FALSE)
```

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




```
model1_Sex <- lm(RestBP ~ Age * Sex, data = Heart)
model2_Chol <- lm(RestBP ~ Age * Sex * Chol, data = Heart)
model3_ChestPain <- lm(RestBP ~ Age * ChestPain, data = Heart)
model4 <- glm(ExAng ~ RestBP + Sex + Chol, data = Heart, family = binomial)
model5 <- glm(Sex ~ RestBP + Chol, data = Heart, family = binomial)
model6 <- glm(Fbs ~ RestBP, data = Heart, family = binomial)

#model1_Sex
#model2_Chol
#model3_ChestPain
summary (model4)
```

```
##
## Call:
## glm(formula = ExAng ~ RestBP + Sex + Chol, family = binomial,
##     data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2387  -0.9427  -0.7452   1.3440   1.8779
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.338981   1.150529  -2.902  0.00371 **
## RestBP       0.008181   0.007132   1.147  0.25135
## SexMale      0.834724   0.294235   2.837  0.00455 **
## Chol         0.003804   0.002486   1.530  0.12591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 382.90  on 302  degrees of freedom
## Residual deviance: 372.12  on 299  degrees of freedom
```

```
## AIC: 380.12
##
## Number of Fisher Scoring iterations: 4
```

```
summary (model5)
```

```
##
## Call:
## glm(formula = Sex ~ RestBP + Chol, family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9590  -1.3092   0.7589   0.8852   1.2159
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.431584   1.077338   3.185  0.00145 **
## RestBP      -0.004789   0.007122  -0.672  0.50135
## Chol        -0.008186   0.002544  -3.218  0.00129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 379.94  on 302  degrees of freedom
## Residual deviance: 367.46  on 300  degrees of freedom
## AIC: 373.46
##
## Number of Fisher Scoring iterations: 4
```

```
summary (model6)
```

```
##
## Call:
## glm(formula = Fbs ~ RestBP, family = binomial, data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0741  -0.6056  -0.5053  -0.4198   2.3194
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.251826   1.212761  -4.330 1.49e-05 ***
## RestBP       0.026062   0.008759   2.975  0.00293 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 254.6  on 302  degrees of freedom
## Residual deviance: 245.8  on 301  degrees of freedom
## AIC: 249.8
##
## Number of Fisher Scoring iterations: 4
```

```
ggplot(Heart, aes(x = Chol, y = Age, color = Sex)) +
  geom_point() +
  labs(x = "Cholestoral", y = "Age", colored = "Sex", title = "Cholestoral vs Age Factored by Sex") +
  geom_smooth(method = "lm", se = FALSE)
```

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