Module 3

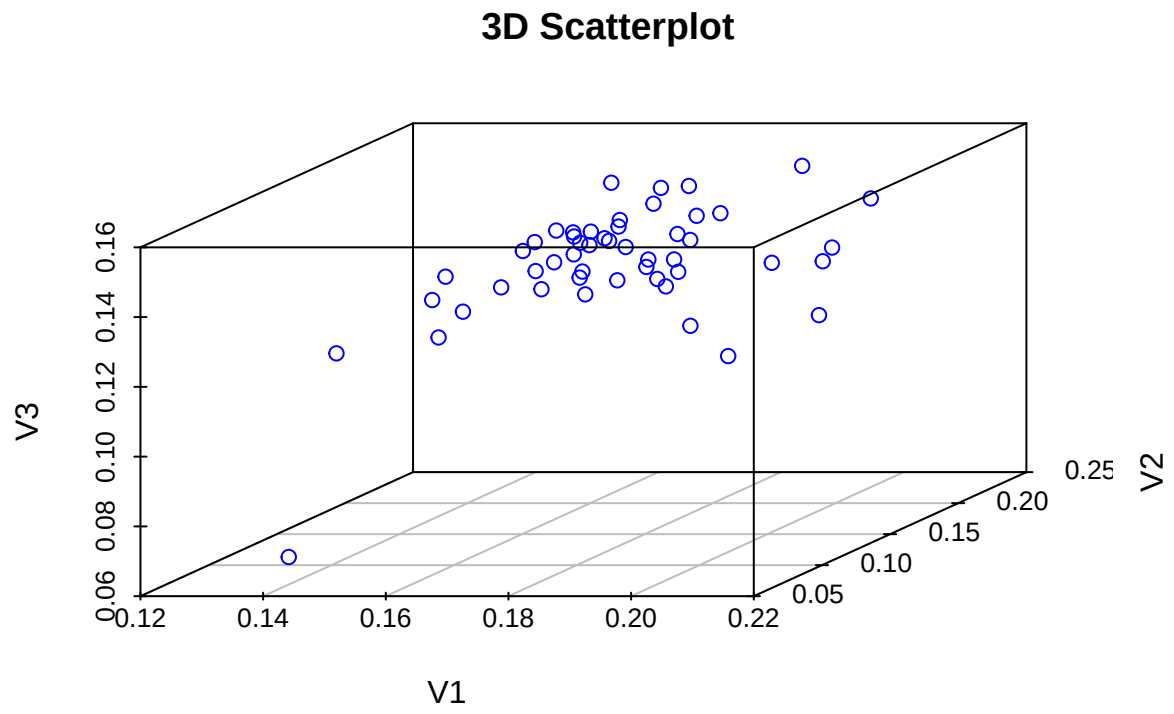
Terry Zhou

3/27/2022

1. Warm Up

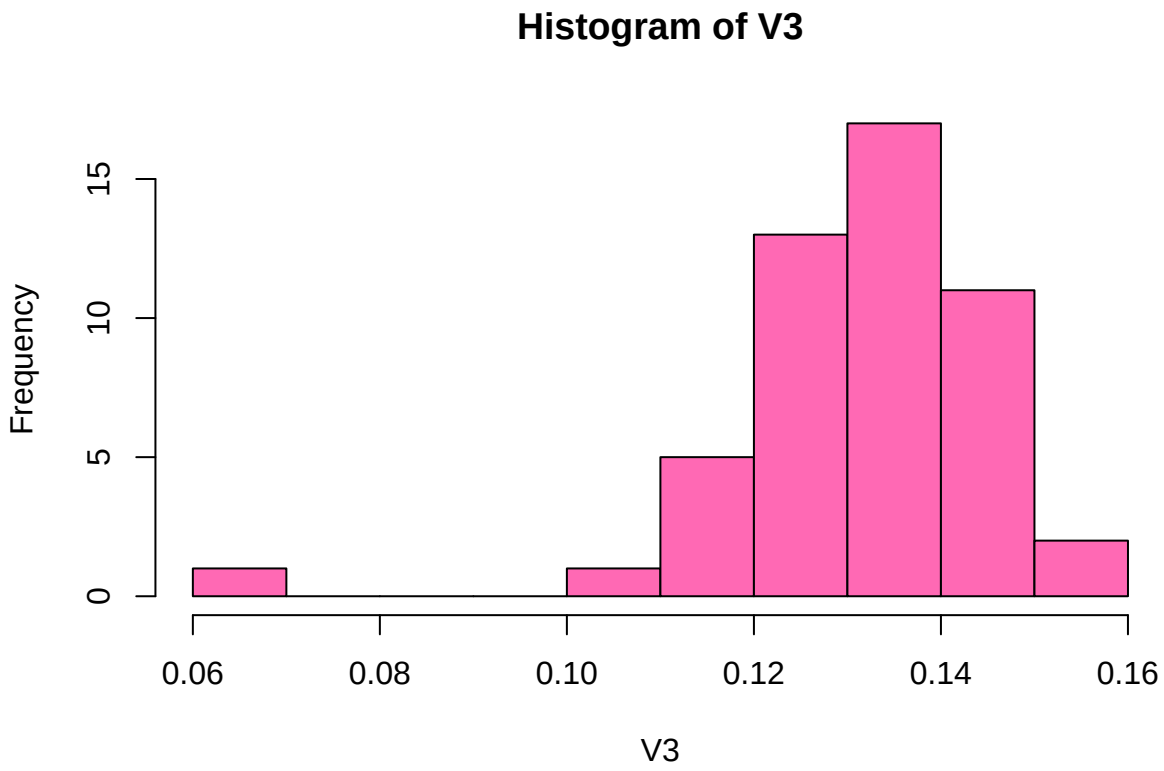
1. Scatterplot

```
library(scatterplot3d)
scatterplot3d(dat$V1, dat$V2, dat$V3,
              xlab = "V1", ylab = "V2", zlab = "V3",
              main = "3D Scatterplot", color = "blue")
```



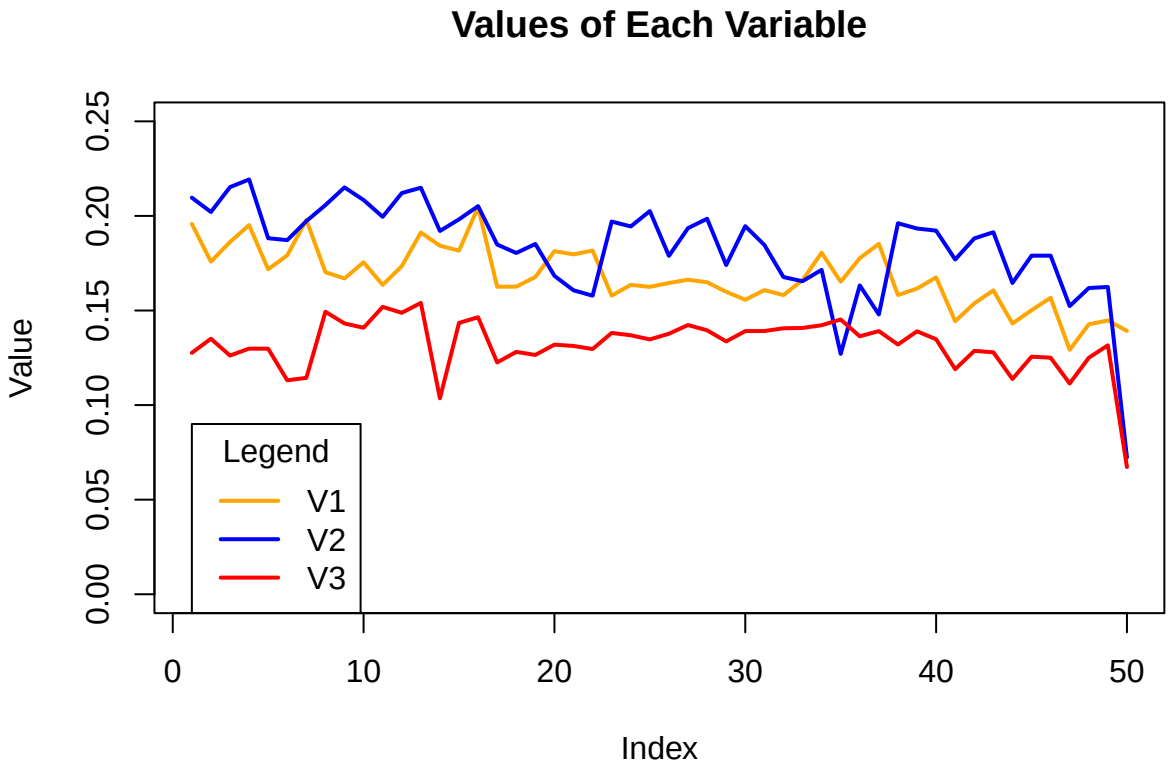
2. Histogram

```
hist(dat$V3, main = "Histogram of V3", xlab = "V3", col = "hotpink")
```



3. Line graph

```
plot(dat$V1, type = "l", col = "orange", ylim = c(0, 0.25), lwd = 2,
     xlab = "Index", ylab = "Value", main = "Values of Each Variable")
lines(dat$V2, type = "l", col = "blue", lwd = 2)
lines(dat$V3, type = "l", col = "red", lwd = 2)
legend(1, 0.09, legend = c("V1", "V2", "V3"),
       col = c("orange", "blue", "red"), lty = 1, lwd = 2,
       title = "Legend")
```


Values of Each Variable
Value
0.00
0.05
0.10
0.15
0.20
0.25
0
10
20
30
40
50
Index
Legend
V1
V2
V3

2. Data Visualization and Analysis on a Dataset 1

2.

```
data(iris)
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 ...
```

```
apply(is.na(iris), 2, which)
```

```
## integer(0)
```

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

The dataset contains 150 observations of 5 variables. The `Sepal.Length`, `Sepal.Width`, `Petal.Length`, and `Petal.Width` variables are all numerical. The `Species` variable is a factor with three levels. There are no missing values.

3.

```
summary(subset(iris, Species == "setosa"))
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100
##  1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400   1st Qu.:0.200
##  Median :5.000   Median :3.400   Median :1.500   Median :0.200
##  Mean   :5.006   Mean   :3.428   Mean   :1.462   Mean   :0.246
```

```
##  3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300
##  Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600
##        Species
##  setosa    :50
##  versicolor: 0
##  virginica : 0
##
##
##
```

```
summary(subset(iris, Species == "versicolor"))
```

```
##   Sepal.Length    Sepal.Width     Petal.Length   Petal.Width          Species
##  Min.   :4.900   Min.   :2.000   Min.   :3.00   Min.   :1.000   setosa    : 0
##  1st Qu.:5.600   1st Qu.:2.525   1st Qu.:4.00   1st Qu.:1.200   versicolor:50
##  Median :5.900   Median :2.800   Median :4.35   Median :1.300   virginica : 0
##  Mean   :5.936   Mean   :2.770   Mean   :4.26   Mean   :1.326
##  3rd Qu.:6.300   3rd Qu.:3.000   3rd Qu.:4.60   3rd Qu.:1.500
##  Max.   :7.000   Max.   :3.400   Max.   :5.10   Max.   :1.800
```

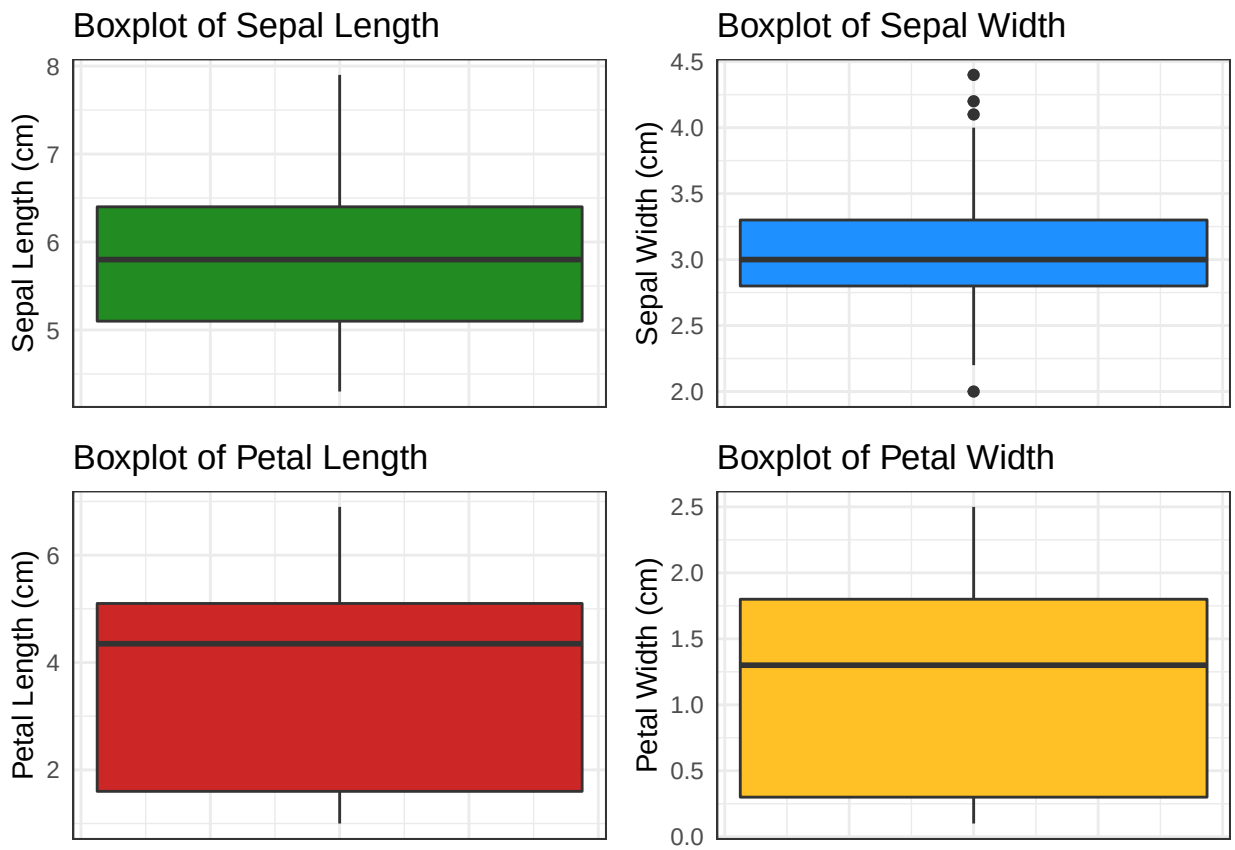
```
summary(subset(iris, Species == "virginica"))
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.900   Min.   :2.200   Min.   :4.500   Min.   :1.400
##  1st Qu.:6.225   1st Qu.:2.800   1st Qu.:5.100   1st Qu.:1.800
##  Median :6.500   Median :3.000   Median :5.550   Median :2.000
##  Mean   :6.588   Mean   :2.974   Mean   :5.552   Mean   :2.026
##  3rd Qu.:6.900   3rd Qu.:3.175   3rd Qu.:5.875   3rd Qu.:2.300
##  Max.   :7.900   Max.   :3.800   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    : 0
##  versicolor: 0
##  virginica :50
##
##
##
```

4.

```
library(ggplot2)
library(gridExtra)
plotlist <- list()
colors <- c("forestgreen", "dodgerblue", "firebrick3", "goldenrod1")
for (i in colnames(iris)[-5]){
  a <- (strsplit(i, split = "[.]"))[[1]][1]
  b <- (strsplit(i, split = "[.]"))[[1]][2]
  index <- which(colnames(iris) == i)
  plot <- ggplot(data = iris, aes_string(y = i)) +
    geom_boxplot(fill = colors[index]) +
    theme_bw() + ylab(paste(a, b, "(cm)", sep = " ")) +
    ggtitle(paste("Boxplot of", a, b, sep = " ")) +
```

```
    theme(axis.text.x = element_blank(), axis.ticks = element_blank())
    plotlist[[i]] <- plot
}
grobs <- arrangeGrob(grobs = plotlist)
grid.arrange(grobs)
```



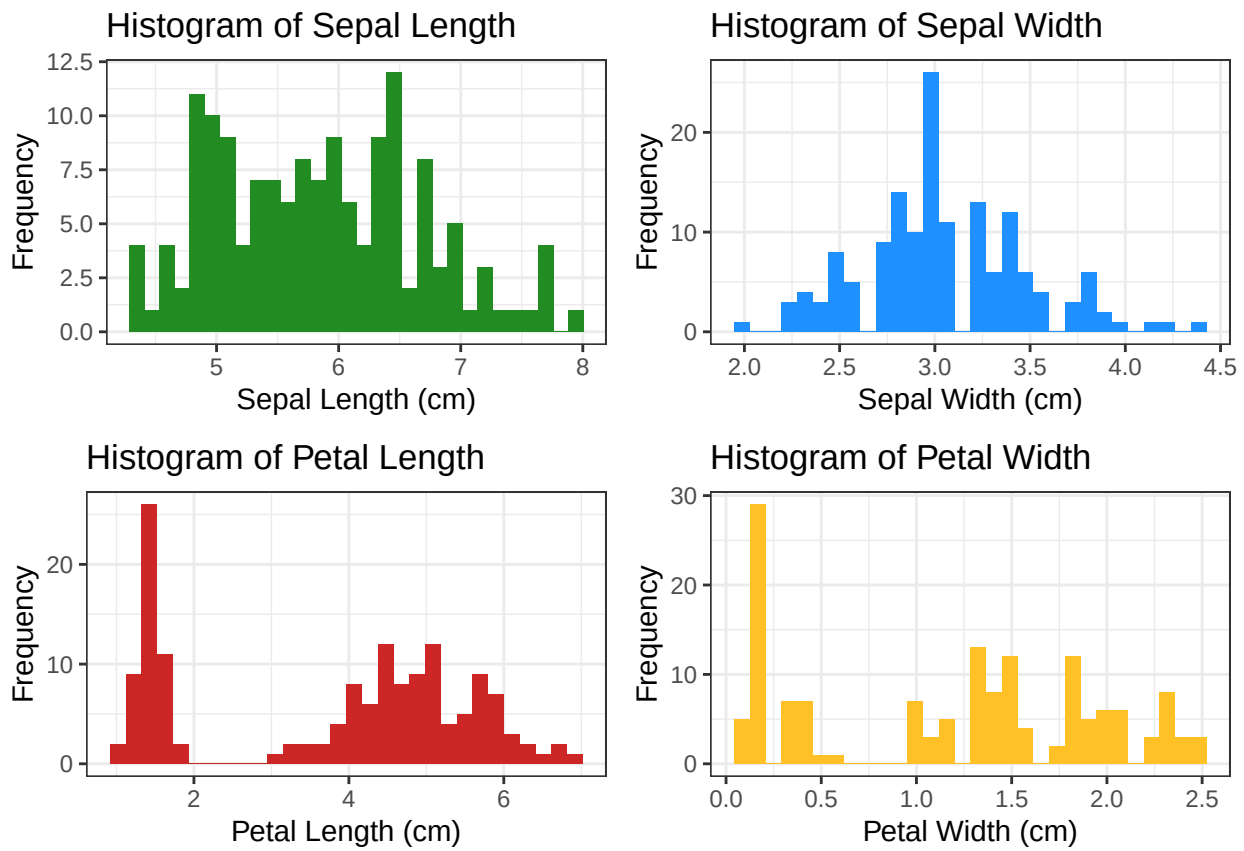
Only the sepal width plot shows outliers.

5.

```
histlist <- list()
colors <- c("forestgreen", "dodgerblue", "firebrick3", "goldenrod1")
for (i in colnames(iris)[-5]){
  a <- (strsplit(i, split = "[.]"))[[1]][1]
  b <- (strsplit(i, split = "[.]"))[[1]][2]
  index <- which(colnames(iris) == i)
  plot <- ggplot(data = iris, aes_string(x = i)) +
    geom_histogram(fill = colors[index]) +
    theme_bw() + xlab(paste(a, b, "(cm)", sep = " ")) + ylab("Frequency") +
    ggtitle(paste("Histogram of", a, b, sep = " "))
    histlist[[i]] <- plot
}
grobs.hist <- arrangeGrob(grobs = histlist)
```

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

```
grid.arrange(grobs.hist)
```



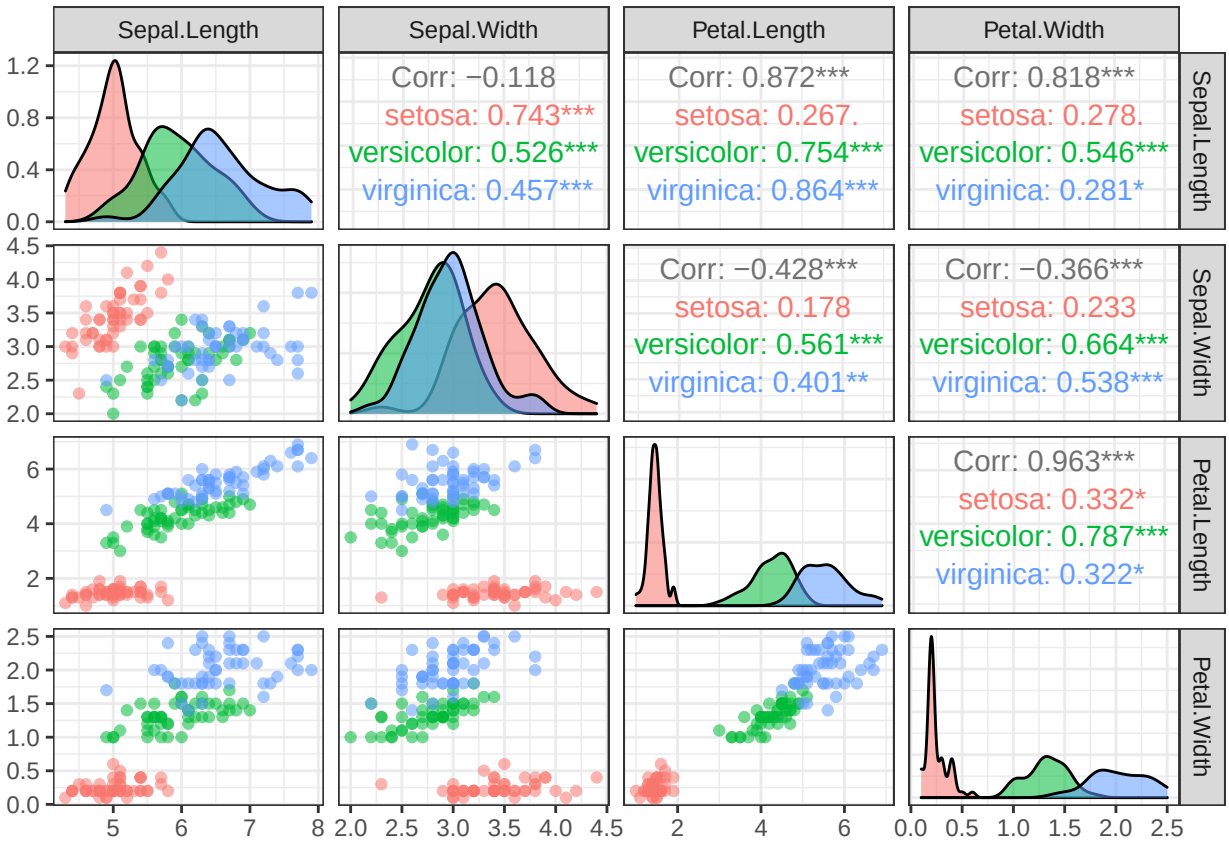
None of the distributions are normal, however the histogram of sepal width seems to be the most normal. There are gaps in the histogram for all four variables. The sepal width, petal length, and petal width histograms are unimodal, however the sepal length histogram seems multimodal. The sepal length, petal length, and petal width histograms all seem to be skew right.

6.

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(iris[, 1:4], aes(colour = iris$Species, alpha = 0.4)) + theme_bw()
```



There is a distinction between the three species for the petal width and petal length variables.

7.

```
cor.matrix <- cor(iris[, c(1:4)])
cor.matrix
```

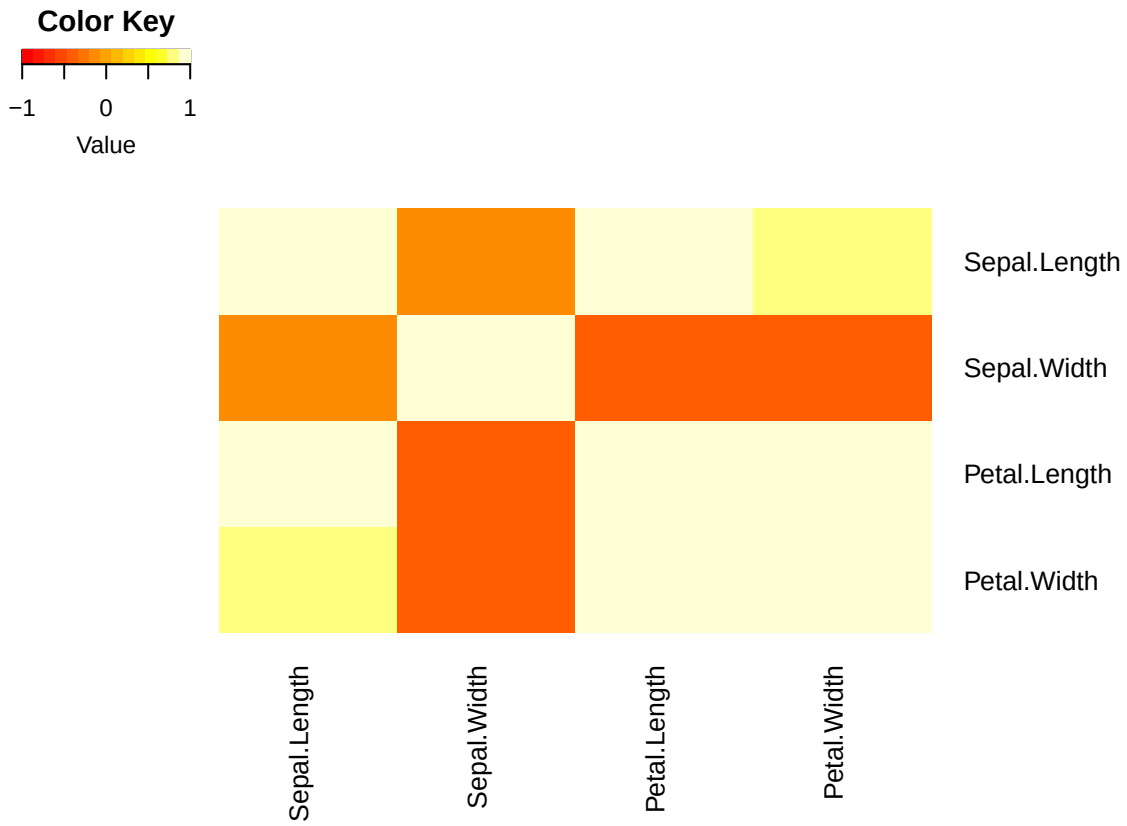
```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

```
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
heatmap.2(cor.matrix, cexRow = 1, cexCol = 1,
          Rowv = FALSE, Colv = FALSE, trace = c("none"), density.info = "none",
          lhei = c(1, 3), lwid = c(0.5, 1.5), margins = c(7, 7))
```



There are strong correlation between Sepal Length and Petal Length, Sepal Length and Petal Width, and Petal Length and Petal Width. This is because the correlation coefficients for each of these pairs is greater than 0.8.

8.

```
sepal.length.a <- t.test(iris$Sepal.Length, conf.level = 0.95)
sepal.length.a
```

```
##
##  One Sample t-test
##
## data:  iris$Sepal.Length
## t = 86.425, df = 149, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.709732 5.976934
## sample estimates:
## mean of x
##  5.843333
```

We are 95% confident that the mean sepal length of all observations is between 5.724154 cm and 5.962512 cm. The p-value of the one sample t-test is is less than 0.05, thus we can reject the null hypothesis and confirm that the mean sepal length is not equal to 0.

```
sepal.width.b <- t.test(iris$Sepal.Width, conf.level = 0.95,
                        alternative = "greater", mu = 4)
sepal.width.b
```

```
##
##  One Sample t-test
##
## data:  iris$Sepal.Width
## t = -26.488, df = 149, p-value = 1
## alternative hypothesis: true mean is greater than 4
## 95 percent confidence interval:
##  2.998429      Inf
## sample estimates:
## mean of x
##  3.057333
```

The null hypothesis is that the mean sepal width is equal to 4 cm. The alternative hypothesis is that the mean sepal width is greater than 4 cm. The one sample t-test gives a p-value of 1, which is greater than our alpha level of 0.05. Thus, we fail to reject the null hypothesis; there is not enough evidence to show that the mean sepal width is greater than 4 cm. The sample estimate for the mean sepal width is 3.057333. We are 95% confident that the mean sepal width is greater than 2.998429.

```
iris1 <- subset(iris, Species == "setosa" | Species == "versicolor")
var.test(Petal.Length ~ Species, iris1, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  Petal.Length by Species
## F = 0.13658, num df = 49, denom df = 49, p-value = 1.026e-10
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.07750613 0.24068043
## sample estimates:
## ratio of variances
##          0.1365804
```

The p-value for the F-test to compare the variances, which is less than our alpha value of 0.05. Thus, we can fail to reject the null hypothesis and conclude that there is a difference in the variance of the petal length in iris setosa and iris versicolor. Thus, we cannot assume that they have equal variances

```
t.test(Petal.Length ~ Species, iris1, alternative = "two.sided",
         conf.level = 0.99, var.diff = TRUE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Petal.Length by Species
```

```
## t = -39.493, df = 62.14, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group setosa and group versicolor is not eq
## 99 percent confidence interval:
##  -2.986265 -2.609735
## sample estimates:
##     mean in group setosa mean in group versicolor
##                    1.462                    4.260
```

The two sample t-test gives a p-value of less than 2.2e-16, which is less than the alpha level of 0.01. Thus, we can reject the null hypothesis and conclude that there is a difference in the mean petal length between the iris setosa and iris versicolor species.

```
petal.length.aov <- aov(Petal.Length ~ Species, data = iris)
summary(petal.length.aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  437.1  218.55    1180 <2e-16 ***
## Residuals   147   27.2    0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA gives a p-value of less than 2e-16 for the species variable. Thus, we can reject the null hypothesis and conclude, at a alpha = 0.05 significance level, that the average petal length is different between the three species categories.

9.

```
setosa <- subset(iris, Species == "setosa")
versicolor <- subset(iris, Species == "versicolor")
virginica <- subset(iris, Species == "virginica")
which(setosa$Sepal.Width == max(setosa$Sepal.Width))
```

```
## [1] 16
```

```
max(setosa$Sepal.Width)
```

```
## [1] 4.4
```

```
which(versicolor$Sepal.Width == max(versicolor$Sepal.Width)) + 50
```

```
## [1] 86
```

```
max(versicolor$Sepal.Width)
```

```
## [1] 3.4
```

```
which(virginica$Sepal.Width == max(virginica$Sepal.Width)) + 100
```

```
## [1] 118 132
```

```
max(virginica$Sepal.Width)
```

```
## [1] 3.8
```

Observation 16 has the highest septal width value (4.4 cm) for iris setosa, 86 has the highest value (3.4 cm) for iris versicolor, and 118 and 132 have the highest values (3.8 cm) for iris virginica.

10.

```
median(setosa$Sepal.Length)
```

```
## [1] 5
```

```
median(versicolor$Sepal.Length)
```

```
## [1] 5.9
```

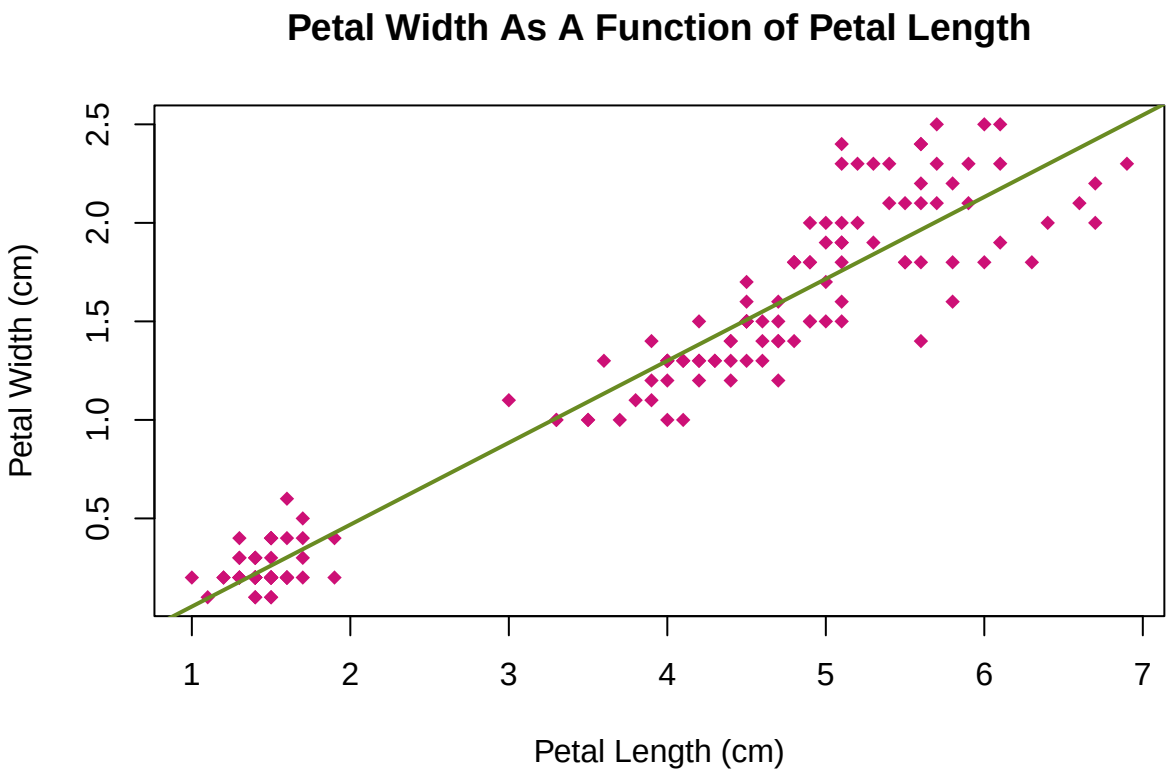
```
median(virginica$Sepal.Length)
```

```
## [1] 6.5
```

The median sepal lengths are 5 cm, 5.9 cm, and 6.5 cm, for iris setosa, iris versicolor, and iris virginica, respectivley.

11.

```
lm <- lm(iris$Petal.Width ~ iris$Petal.Length)
plot(x = iris$Petal.Length, y = iris$Petal.Width,
     xlab = "Petal Length (cm)", ylab = "Petal Width (cm)",
     main = "Petal Width As A Function of Petal Length", col = "deeppink3",
     pch = 18)
abline(lm, col = "olivedrab4", lwd = 2)
```




```
predict.2.5 <- as.numeric(lm$coefficients[1] + (lm$coefficients[2] * 2.5))
predict.2.5
```

```
## [1] 0.676313
```

There is a strong positive relationship between petal width and petal length; as petal length increases, so does petal width. At a petal length of 2.5 cm, we can expect the petal width to be 0.676313 cm.

2. Data Visualization and Analysis on a Dataset 2

1.

```
heart <- read.csv("processed.cleveland.data", header = FALSE)
colnames(heart) <- c("age", "sex", "cp", "trestbps", "chol", "fbs", "restecg",
                     "thalach", "exang", "oldpeak", "slope", "ca",
                     "thal", "num")
str(heart)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : num  1 1 1 1 0 1 0 0 1 1 ...
##  $ cp      : num  1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps: num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs     : num  1 0 0 0 0 0 0 0 0 1 ...
##  $ restecg : num  2 2 2 0 2 0 2 0 2 2 ...
##  $ thalach : num  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang   : num  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   : num  3 2 2 3 1 1 3 1 2 3 ...
##  $ ca      : chr  "0.0" "3.0" "2.0" "0.0" ...
##  $ thal    : chr  "6.0" "3.0" "7.0" "3.0" ...
##  $ num     : int  0 2 1 0 0 0 3 0 2 1 ...
```

```
apply(is.na(heart), 2, which)
```

```
## integer(0)
```

```
summary(heart)
```

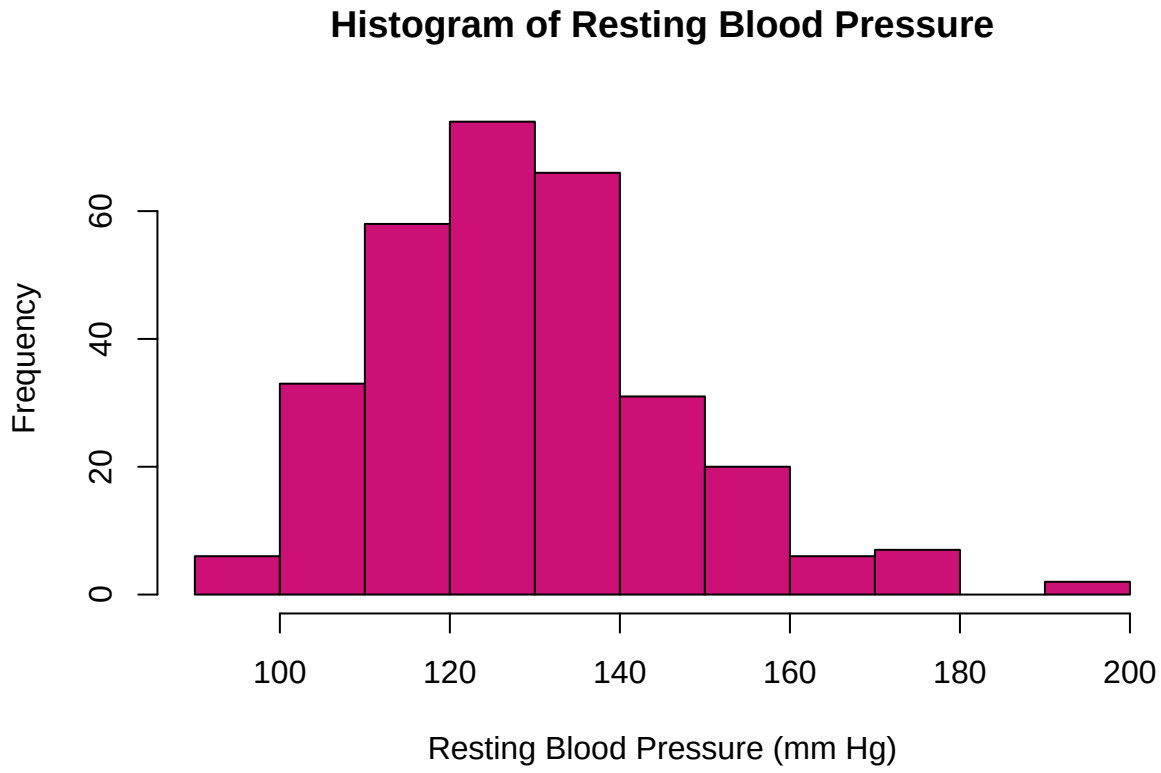
```
##       age             sex               cp           trestbps
##  Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0
##  1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0
##  Median :56.00   Median :1.0000   Median :3.000   Median :130.0
##  Mean   :54.44   Mean   :0.6799   Mean   :3.158   Mean   :131.7
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0
##       chol            fbs            restecg          thalach
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :241.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.7   Mean   :0.1485   Mean   :0.9901   Mean   :149.6
##  3rd Qu.:275.0   3rd Qu.:0.0000   3rd Qu.:2.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##      exang           oldpeak         slope            ca
##  Min.   :0.0000   Min.   :0.00   Min.   :1.000   Length:303
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   Class :character
##  Median :0.0000   Median :0.80   Median :2.000   Mode  :character
##  Mean   :0.3267   Mean   :1.04   Mean   :1.601
```

```
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000
##  Max.   :1.0000   Max.   :6.20   Max.   :3.000
##      thal                num
##  Length:303         Min.   :0.0000
##  Class :character   1st Qu.:0.0000
##  Mode  :character   Median :0.0000
##                     Mean   :0.9373
##                     3rd Qu.:2.0000
##                     Max.   :4.0000
```

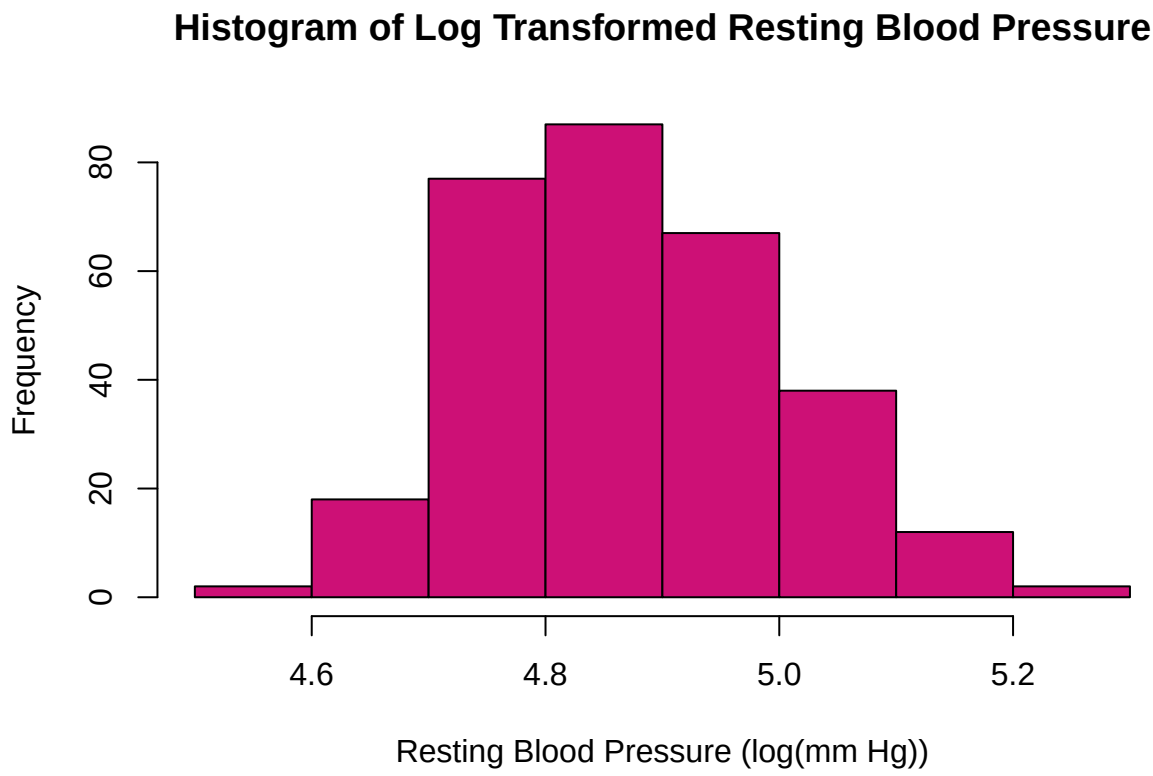
There are 14 variables. All of the variables except `ca` and `thal` are numeric; `ca` and `thal` are both characters.

3.

```
hist(heart$trestbps, main = "Histogram of Resting Blood Pressure",
     xlab = "Resting Blood Pressure (mm Hg)", col = "deeppink3")
```




```
hist(log(heart$trestbps),
     main = "Histogram of Log Transformed Resting Blood Pressure",
     xlab = "Resting Blood Pressure (log(mm Hg))", col = "deeppink3")
```



The resting blood pressure is slightly right skew. To reduce the skew, I would take the log transformation of the variable.

4.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
heart <- heart %>%
  mutate(age_group = case_when((age >= 20 & age <= 30) ~ "20-30",
                               (age >= 31 & age <= 40) ~ "31-40",
                               (age >= 41 & age <= 50) ~ "41-50",
                               (age >= 51 & age <= 60) ~ "51-60",
                               (age >= 61 & age <= 70) ~ "61-70",
                               (age >= 71 & age <= 80) ~ "71-80"))
age.tot <- heart %>%
  group_by(age_group) %>%
  summarize(tot = n())
cp.tot <- heart %>%
  group_by(cp) %>%
  summarize(tot = n())
tot <- heart %>%
  group_by(age_group, cp) %>%
  summarize(tot = n())
```

```
## `summarise()` has grouped output by 'age_group'. You can override using the `.groups` argument.
```

```
mat <- matrix(nrow = 7, ncol = 5)
mat[1, 4] <- 0
mat[1, 2] <- 0
mat[1, 1] <- 0
mat[1, 3] <- tot$tot[1]
mat[2, 4:1] <- tot$tot[2:5]
mat[3, 4:1] <- tot$tot[6:9]
mat[4, 4:1] <- tot$tot[10:13]
mat[5, 4:1] <- tot$tot[14:17]
mat[6, 4] <- 0
mat[6, 3:1] <- tot$tot[18:20]
mat[1:6, 5] <- age.tot$tot
mat[7, 4:1] <- cp.tot$tot
mat[7, 5] <- nrow(heart)
rownames(mat) <- c(age.tot$age_group, "Total")
colnames(mat) <- c("Asymptomatic", "Non-Anginal Pain", "Atypical Angina",
                   "Typical Angina", "Total")
mat
```

```
##       Asymptomatic Non-Anginal Pain Atypical Angina Typical Angina Total
## 20-30            0                0               1              0     1
## 31-40            6                6               2              3    17
## 41-50           29               25              20              2    76
## 51-60           66               34              20             10   130
## 61-70           41               19               5              8    73
## 71-80            2                2               2              0     6
## Total          144               86              50             23   303
```