Lab5Markdown

2023-03-07

DATA VISUALIZATION

1. Load the iris dataset into R and Explore its structure and summary statistics

```
# Loading the dataset

data(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa

```
# Exploring its structure

str(iris)
```

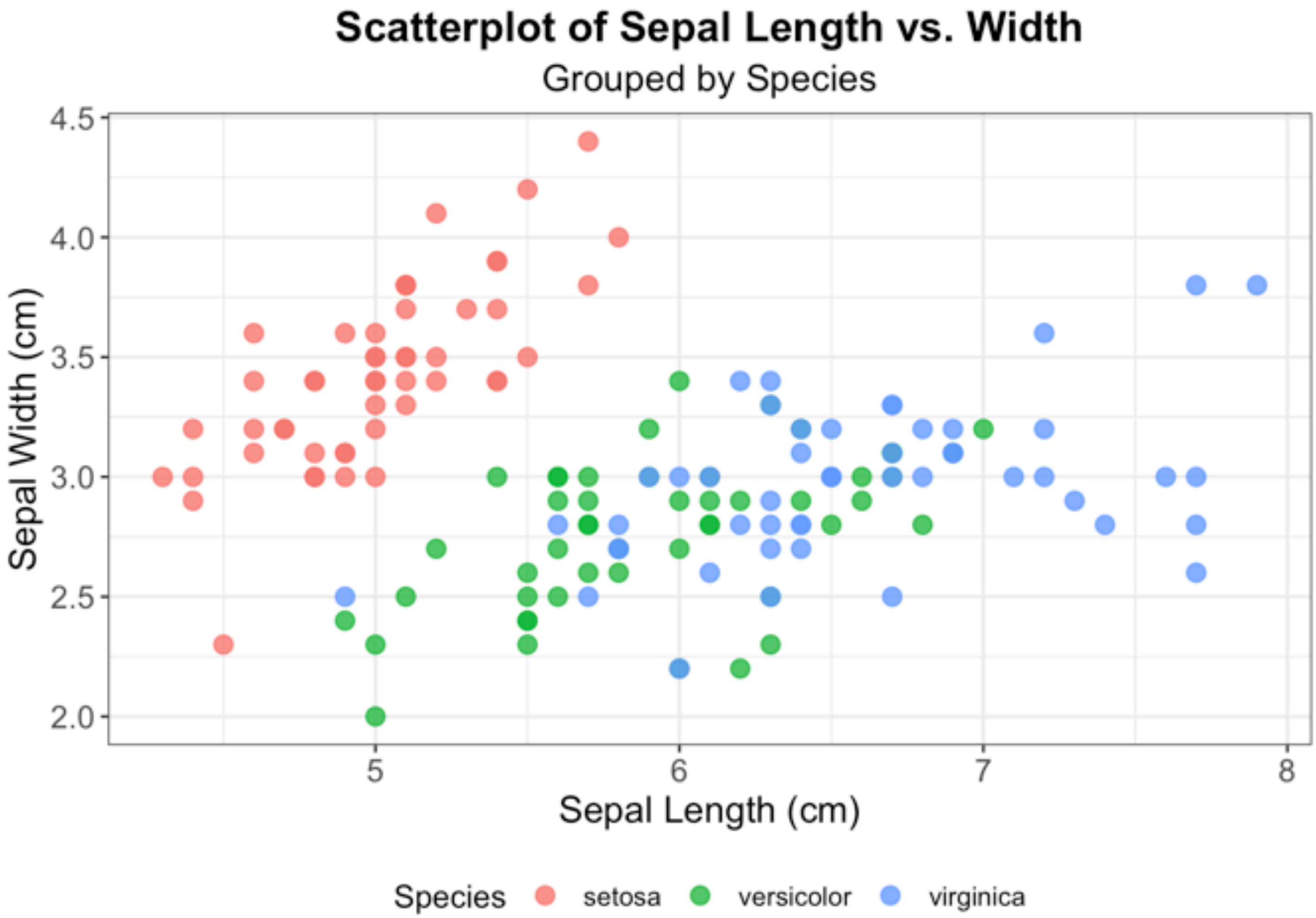
```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3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

```
# Exploring its summary statistic

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  
##                 
##                 
## 
```

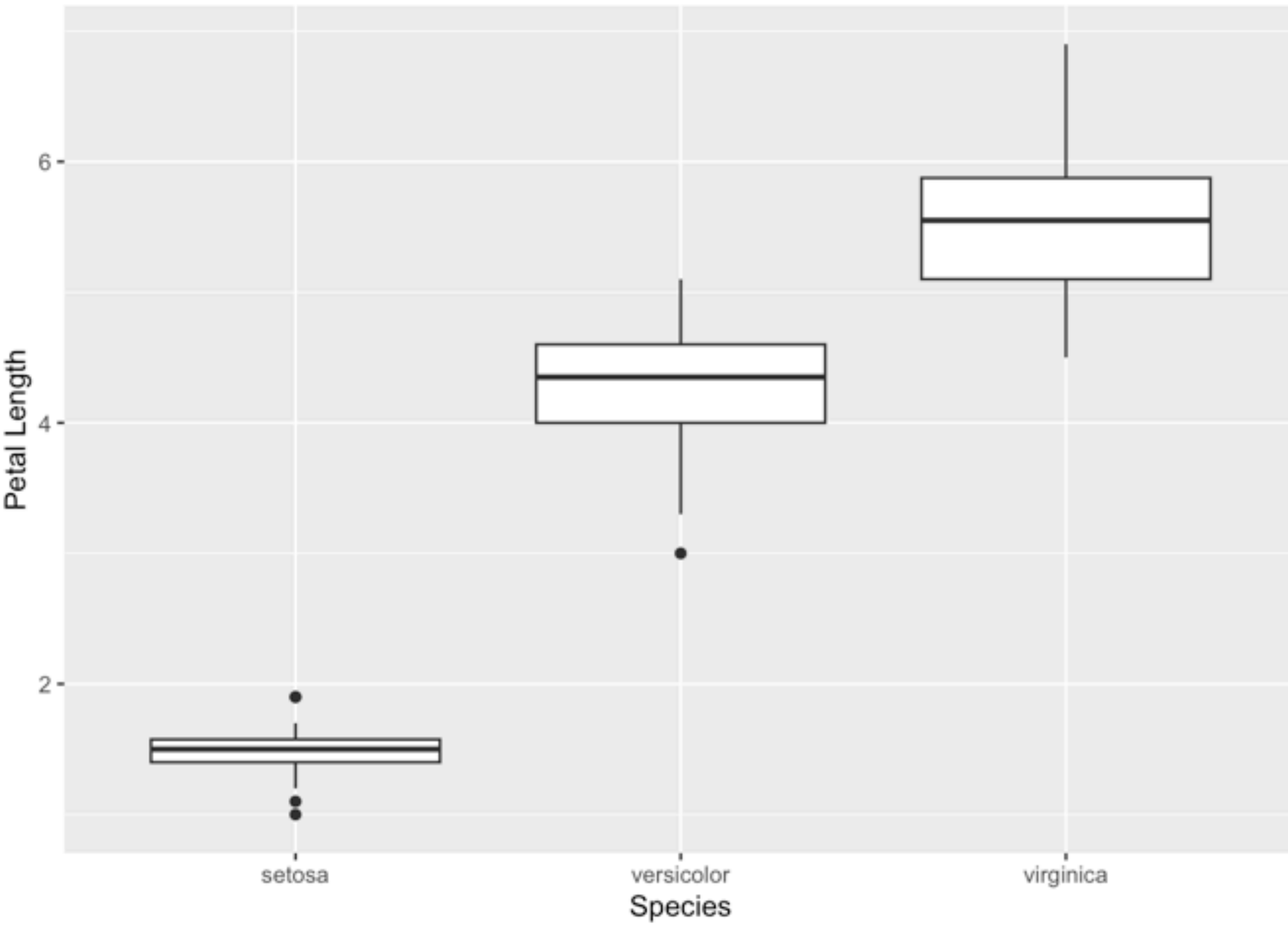
2. Create a scatterplot of the sepal length and width of the iris flowers. Use different colors to represent the different species of flowers.

```
# Create scatterplot with customized labels, colors, and design elements
library(ggplot2)
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(x = "Sepal Length (cm)", y = "Sepal Width (cm)",
       title = "Scatterplot of Sepal Length vs. Width",
       subtitle = "Grouped by Species",
       color = "Species") +
  theme_bw() +
  theme(plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
        plot.subtitle = element_text(size = 14, hjust = 0.5),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 14),
        legend.title = element_text(size = 12),
        legend.text = element_text(size = 10),
        legend.position = "bottom")
```



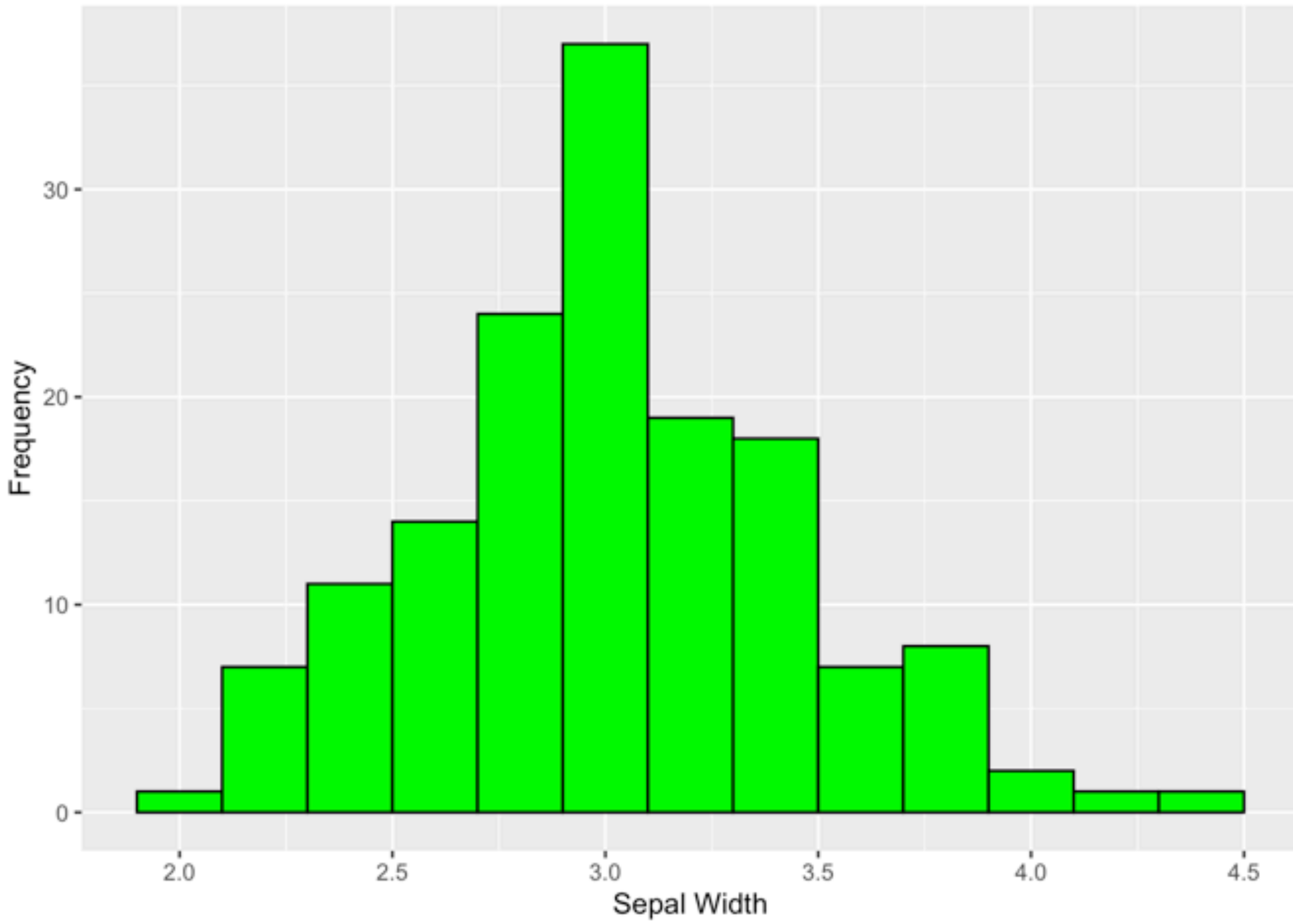
3. Create a boxplot of the petal length for each species of flower.

```
# Creating Boxplot
library(ggplot2)

ggplot(iris, aes(x = Species, y = Petal.Length)) +
  geom_boxplot() +
  labs(x = "Species", y = "Petal Length")
```



4. Create a histogram of the sepal width of the iris flowers.

```
# Creating Histogram
library(ggplot2)

ggplot(iris, aes(x = Sepal.Width)) +
  geom_histogram(binwidth = 0.2, color = "black", fill = "green") +
  labs(x = "Sepal Width", y = "Frequency")
```



LINEAR REGRESSION

1. Load the mtcars dataset into R and Explore its structure and summary statistics

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2

```
# Exploring its structure:
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
# Exploring its summary statistic:
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

2. Use linear regression to model the relationship between “mpg” (dependent variable) and “hp” (Independent variable). Interpret the regression coefficients and R-squared value

```r
# fit linear regression model
fit_LRM <- lm(mpg ~ hp, data = mtcars)

# display summary of the model
summary(fit_LRM)
```

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

3. Create a multiple linear regression model (using “hp” and “wt” as Independent variables, and mpg as a dependent variable).

```r
# fit multiple linear regression model
fit_MLRM <- lm(mpg ~ hp + wt, data = mtcars)

# display summary of the model
summary(fit_MLRM)
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## hp          -0.03177    0.00903  -3.519  0.00145 **
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

4. Compare the predictive power of the simple linear regression model (using “hp” as the predictor variable) and the multiple linear regression model (using “hp” and “wt” as predictor variables).

```
# loading mtcars dataset
data(mtcars)

# Fit simple linear regression model using hp as predictor variable
fit1 <- lm(mpg ~ hp, data = mtcars)

# Fit multiple linear regression model using hp and wt as predictor variables
fit2 <- lm(mpg ~ hp + wt, data = mtcars)

# Making predictions using simple and multiple regression models
pred1 <- predict(fit1, newdata = mtcars)
pred2 <- predict(fit2, newdata = mtcars)

# Calculating mean squared error (MSE) and root mean squared error (RMSE)
MSE1 <- mean((mtcars$mpg - pred1)^2)
MSE2 <- mean((mtcars$mpg - pred2)^2)
RMSE1 <- sqrt(MSE1)
RMSE2 <- sqrt(MSE2)

# Calculating R-squared values
Rsq1 <- summary(fit1)$r.squared
Rsq2 <- summary(fit2)$r.squared

# Display evaluation metrics
cat("Simple Linear Regression Model:\n")
```

```
## Simple Linear Regression Model:
```

```
cat("MSE:", MSE1, "\n")
```

```
## MSE: 13.98982
```

```
cat("RMSE:", RMSE1, "\n")
```

```
## RMSE: 3.740297
```

```
cat("R-squared:", Rsq1, "\n\n")
```

```
## R-squared: 0.6024373
```

```
cat("Multiple Linear Regression Model:\n")
```

```
## Multiple Linear Regression Model:
```

```
cat("MSE:", MSE2, "\n")
```

```
## MSE: 6.095242
```

```
cat("RMSE:", RMSE2, "\n")
```

```
## RMSE: 2.468854
```

```
cat("R-squared:", Rsq2, "\n")
```

```
## R-squared: 0.8267855
```