Data3421-Lab5-RyanBui

Ryan Bui

2023-03-10

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

##Problem A1: 1. Load the iris dataset into R and explore its structure and summary statistics.

```
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
##
##
##
```

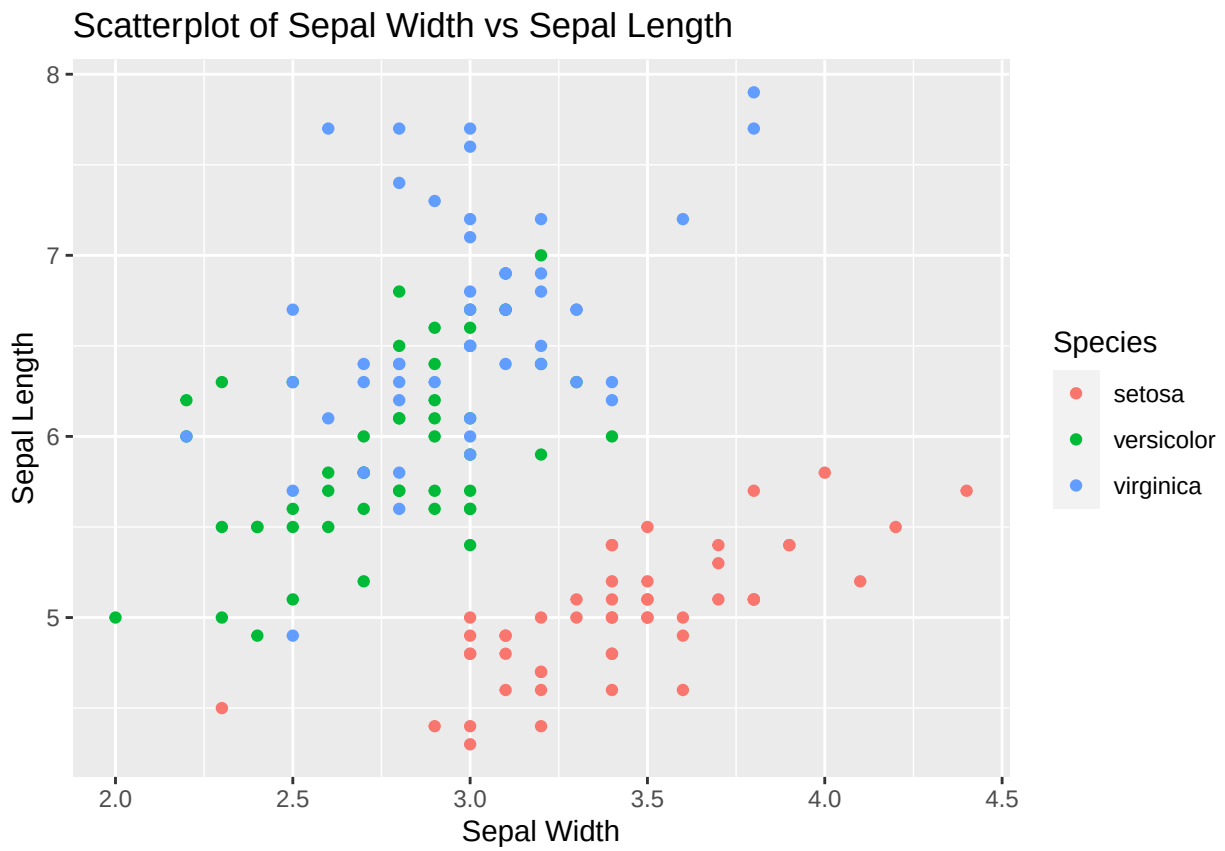
Problem A2:

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

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

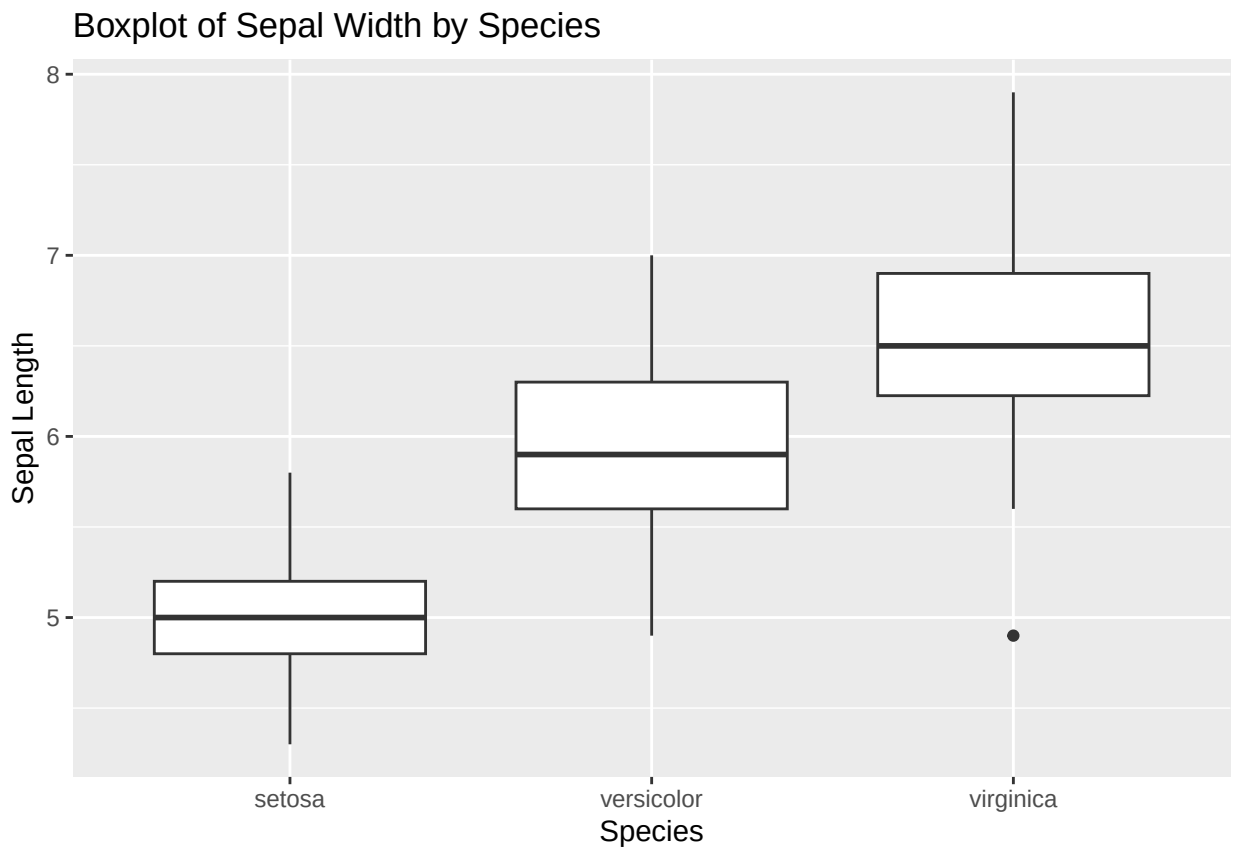
```
ggplot(iris, aes(x = Sepal.Width, y = Sepal.Length, color = Species)) +
  geom_point() +
  labs(title = "Scatterplot of Sepal Width vs Sepal Length",
       x = "Sepal Width", y = "Sepal Length")
```



Problem A3

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

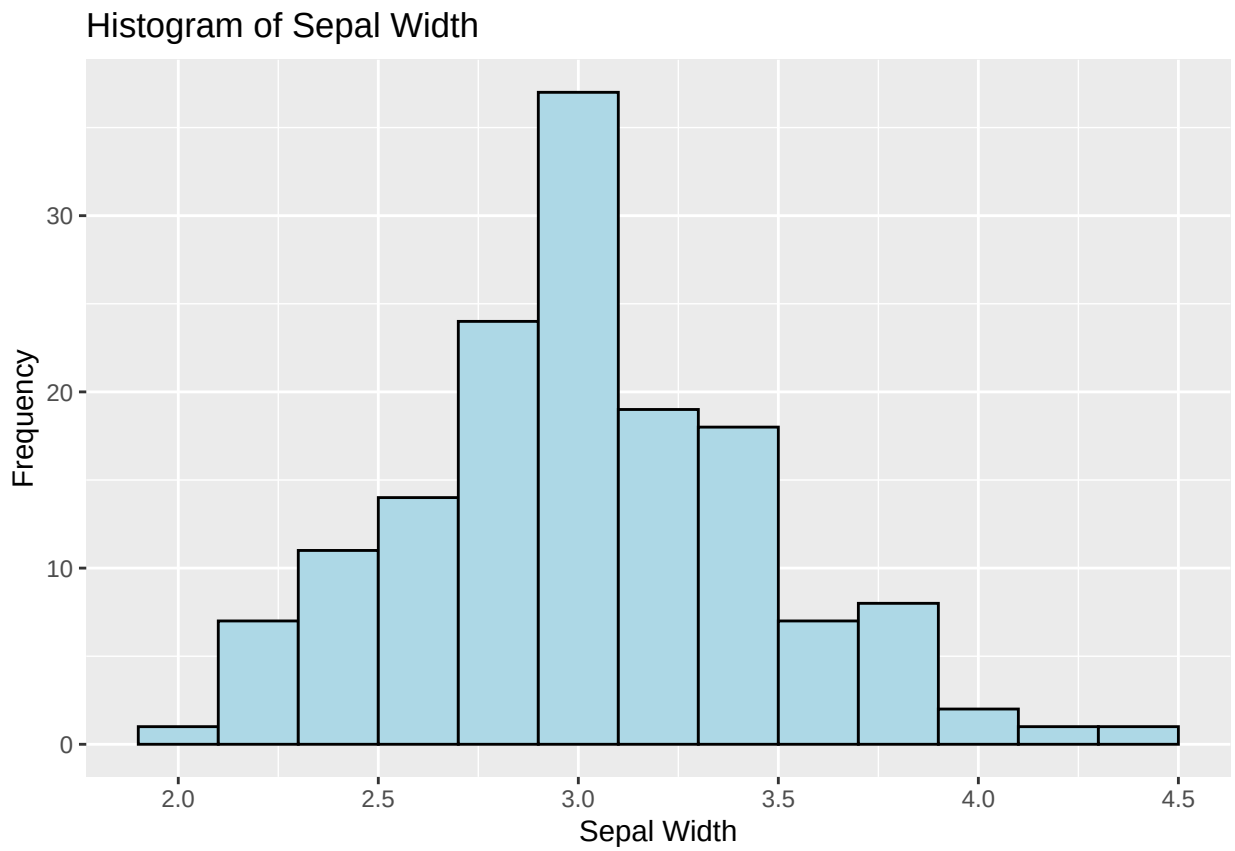
```
# Sub-setting the iris dataset into each species of flower
ggplot(iris, aes(x = Species, y = Sepal.Length)) +
  geom_boxplot() +
  labs(title = "Boxplot of Sepal Width by Species",
       x = "Species", y = "Sepal Length")
```



Problem A4

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

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



Problem B1

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

```
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
```

Problem B2

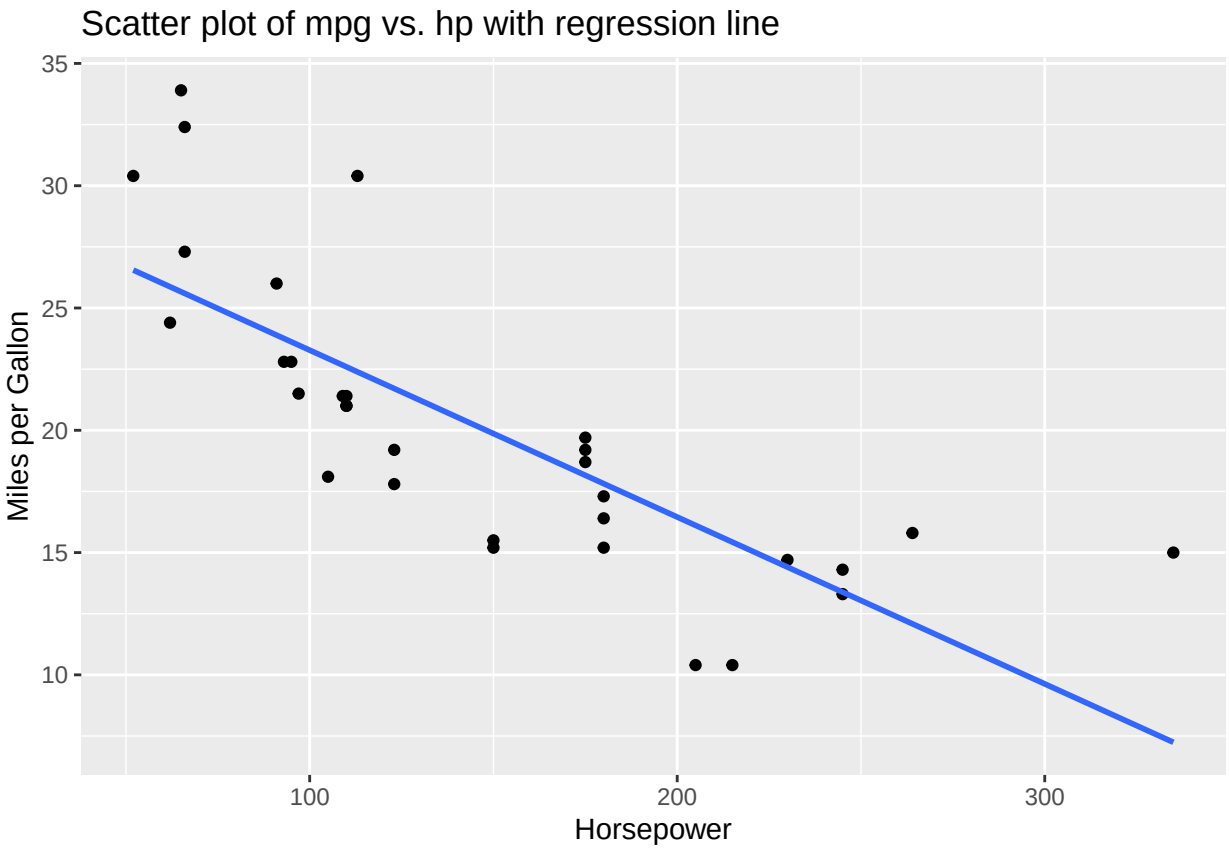
Use linear regression to model the relationship between "mpg" (dependent variable) and "hp" (Independent variable). Interpret the regression coefficients and R-squared value.

```
# Create a scatter plot of mpg vs. hp
model <- lm(mpg ~ hp, data = mtcars)
# Print the model summary
summary(model)
```

```
##
## 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
```

```
# Add the regression line to the scatter plot
ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point() +
  labs(x = "Horsepower", y = "Miles per Gallon") +
  ggtitle("Scatter plot of mpg vs. hp with regression line") +
  geom_smooth(method = "lm", se = FALSE)
```

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



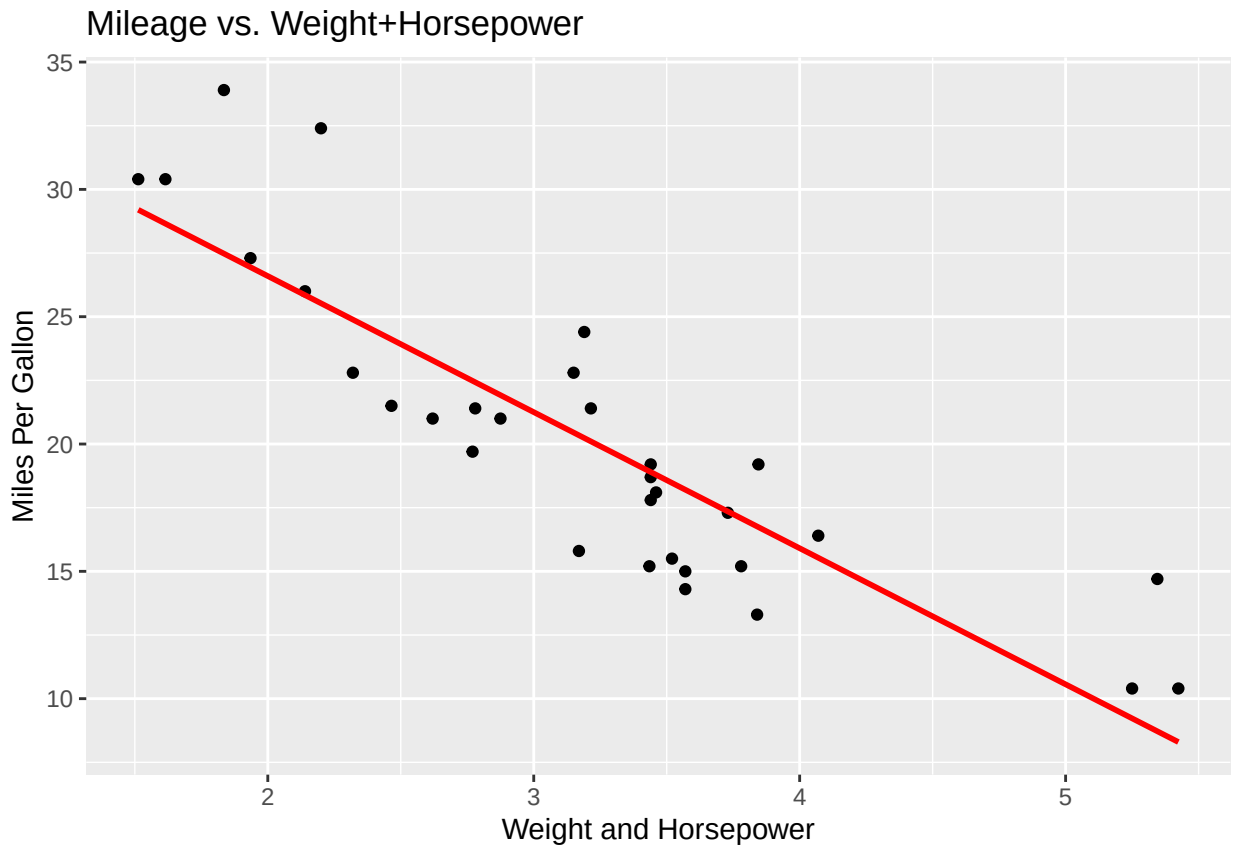
The adjusted R-Squared value is 0.5892 which means that about 59% of the variance in the response variable is explained by our independent variable which is horsepower. Looking at the coefficient of horsepower we can see that it looks like that for every point of horsepower you lose 0.0068 miles per gallon. ## Problem B3 Create a multiple linear regression model (using "hp" and "wt" as Independent variables, and mpg as a dependent variable).

```
model <- lm(mpg ~ hp + wt, data = mtcars)
summary(model)
```

```
##
## 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
```

```
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "red") +
  labs(title = "Mileage vs. Weight+Horsepower", x = "Weight and Horsepower", y = "Miles Per Gallon")
```



Problem B4

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).

Comparing the adjusted R-squared values the multiple linear regression model has the higher value with 0.8148 compared to the simple linear regression model which has 0.5892. So the multiple linear regression model can explain 81% of the variance in the dependent variable.

The horsepower coefficient seems to have gotten slightly larger from the original model it is not -0.03 where as before it was -0.006. It could be noted that the weight coefficient much larger than the other ones so one could infer that weight has a larger impact on the miles per gallon of the vehicle.