Iris Flower Analysis

Bin Han

2025-06-25

Project description

This project explores the well-known Iris dataset to understand relationships between flower measurements and species using statistical and visual analysis in R. The dataset contains measurements of sepal length, sepal width, petal length, and petal width for three species of iris flowers: setosa, versicolor, and virginica.

We begin by examining the structure of the dataset and interpreting what each variable represents in the context of a flower's anatomy. Through calculating correlations and creating scatterplots, we investigate how these features relate to each other numerically and visually. Linear regression models are then fitted to predict sepal length based on other variables, allowing us to assess the predictive power of sepal width and species.

By isolating specific species like setosa, we compare how relationships differ within a subgroup versus the full dataset. The project concludes with a prediction example to apply the learned model in practice.

Load the necessary libraries

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
```

```
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```r
data(iris)
```

Data exploration

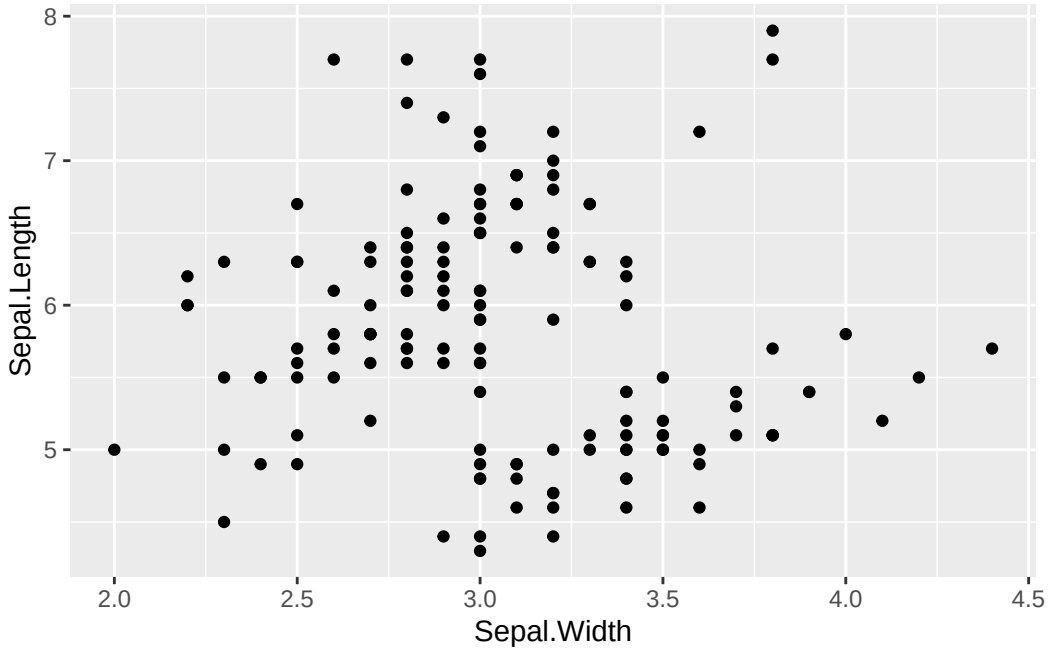
1. Compute the correlations between all numeric variables in the dataset. Save the result in a variable called correlations. Are the correlations positive or negative? Does this make sense to you? **There are positive correlation and negative correlation. This is chaotic.**

```r
correlations <- iris %>%
  select(,-Species) %>%
  cor()
correlations
```

```
             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
```

2. Use ggplot to create a scatterplot with the sepal width on the x axis against the sepal length on the y axis. Save the resulting plot in a variable called plot1. How does the plotted points match the correlations we computed before? **The plots points match the correlations we computed before, hard to find pattern.**

```r
plot1 <- iris %>%
  ggplot(aes(x=Sepal.Width, y=Sepal.Length))+
  geom_point()
plot1
```



3. Fit a linear regression model predicting the sepal length using sepal width. Save the model in a variable called model1. Does the estimated coefficient for sepal width match what have seen in tasks 1 and 2? ** The estimated coefficient for sepal width and lenght match what we have seen in Task 1 and Task 2". It seems inversely proportional.

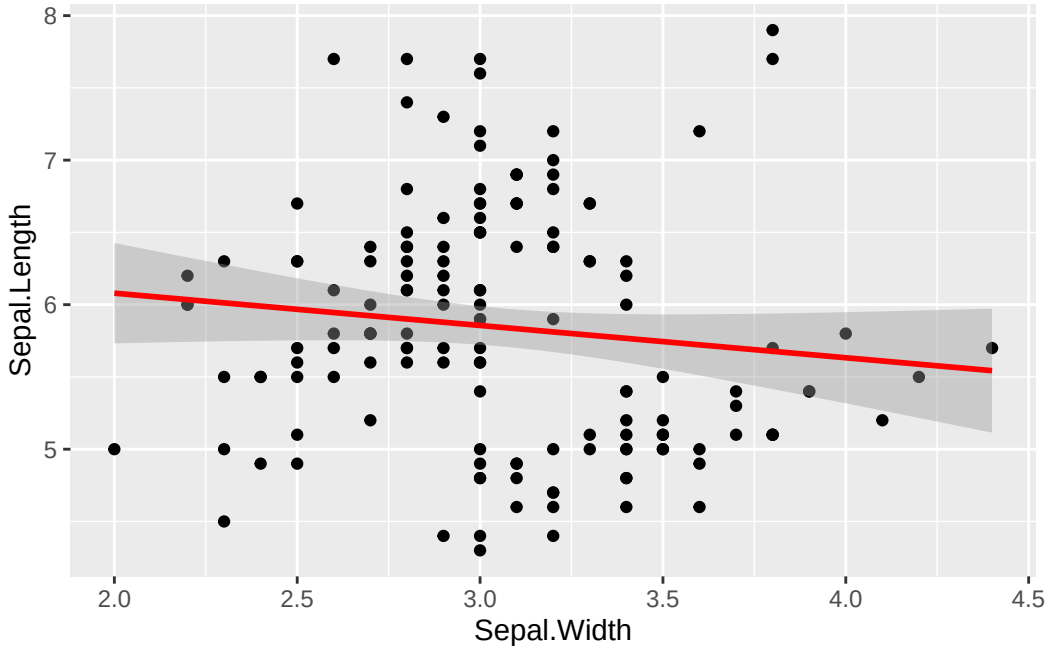
```
model1 <- iris %>%
  lm(Sepal.Length~ Sepal.Width, data=.)%>%
  summary()
model1
```

```
Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5561 -0.6333 -0.1120  0.5579  2.2226

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5262     0.4789   13.63   <2e-16 ***
Sepal.Width  -0.2234     0.1551   -1.44    0.152
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8251 on 148 degrees of freedom
Multiple R-squared:  0.01382,   Adjusted R-squared:  0.007159
F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519
```

```
ggplot(iris,aes(x=Sepal.Width, y=Sepal.Length))+
         geom_point()+
         stat_smooth(method="glm", color="red")
```

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



4. Compute the correlations between all numeric variables in the dataset using only the setosas. Save the result in a variable called correlations_setosa. Are these correlations positive or negative? How are they different from the overall correlations? Does this make sense **Now it make sense. since the length and Width becomes proportional.**

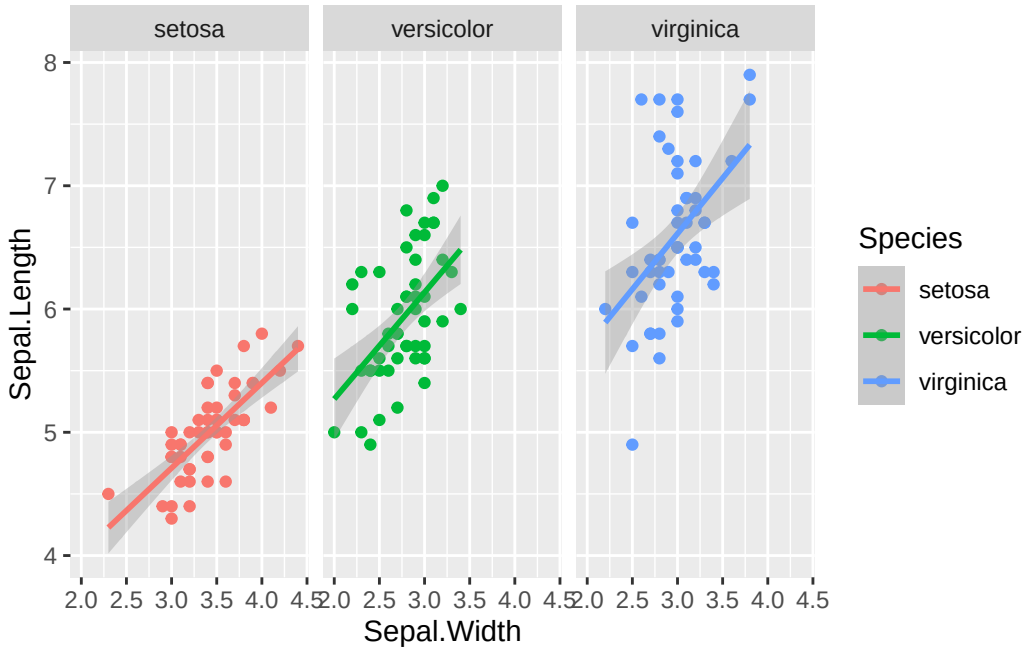
```
correlations_setosa<-iris%>%
  filter(Species=="setosa") %>%
  select(, -Species) %>%
  cor()
correlations_setosa
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000   0.7425467    0.2671758   0.2780984
Sepal.Width     0.7425467   1.0000000    0.1777000   0.2327520
Petal.Length    0.2671758   0.1777000    1.0000000   0.3316300
Petal.Width     0.2780984   0.2327520    0.3316300   1.0000000
```

5. Use ggplot to create a scatterplot with the sepal width on the x axis against the sepal length on the y axis. Color the points by flower species. Save the resulting plot in a variable called plot2 Does this match the correlations we computed for setosa flowers? **Setosa flowers width is proportional to the length.**

```
plot2<-iris%>%
  ggplot(aes(x=Sepal.Width, y=Sepal.Length,color=Species))+
  geom_point()+
  stat_smooth(method="glm")+
  facet_wrap(~Species)

plot2
```

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

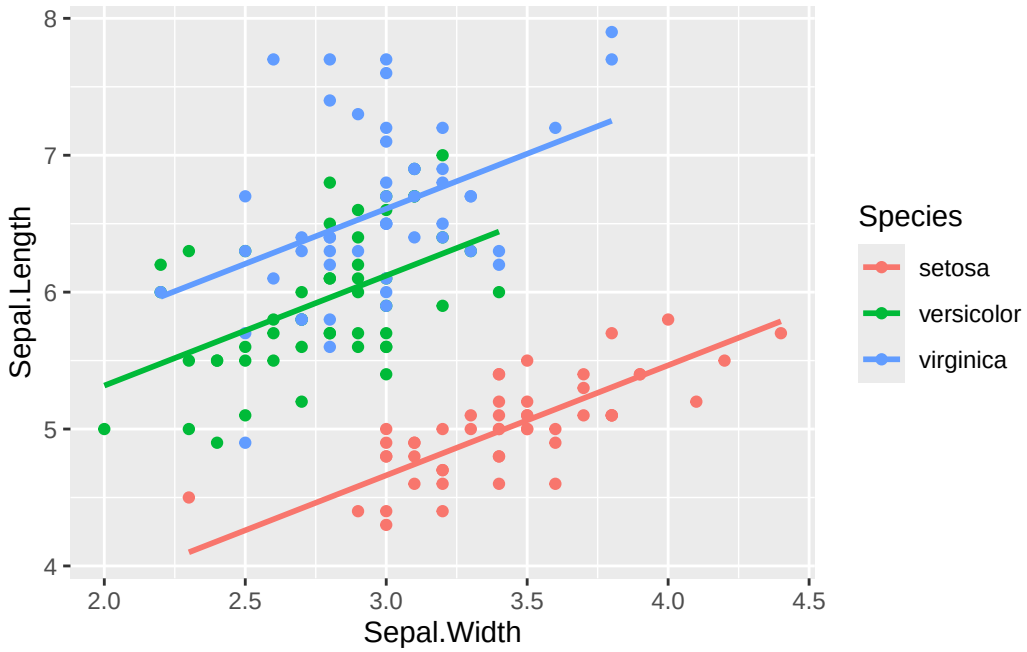


6.Fit a linear regression model predicting the sepal length using sepal width and the flower species. Save the model is a variable called model2. Does the estimated coefficient for

Sepal.Width make sense? What happened when we added species to our model? How does it change the interpretation? **The estimated coefficient means when the width increases one unit, the length increases 0.80 unit. The model gives different intercepts for different species. This model gives same influence of sepal width to sepal length. however with differenct initial point**

```
model2<-iris%>%
  lm(Sepal.Length~Sepal.Width+Species, data=.)
iris$pred<-predict(model2)
iris%>%
  ggplot(aes(x=Sepal.Width, y=Sepal.Length, color=Species))+
  geom_point()+
  geom_line(aes(y=pred), size=1)
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



7. Predict the sepal length of a setosa with a sepal width of 3.6 cm using your second model. Save the resulting prediction in a variable called prediction. Does the prediction appear reasonable? **For satosa, sepal length is 5.144cm for a sepal width of 3.6cm.**

```r
newdata<-data.frame(Sepal.Width=c(3.6), Species="setosa")
newdata
```

```
  Sepal.Width Species
1         3.6  setosa
```

```r
prediction <- predict(model2, newdata)
prediction
```

```
       1
5.144212
```