One-Way Anova: Iris Dataset

Bett Kipkemoi

2026-01-10

ONE-WAY ANOVA

The goal in this project is to analyze if test whether there are significant differences among the group means.

Understanding the Dataset

The Iris dataset is one of the most famous and widely used datasets in machine learning and statistics. It was introduced by the British statistician and biologist Ronald Fisher in 1936 as an example for discriminant analysis.

- Size: 150 samples (rows).
- Features: 4 numerical measurements (in centimeters):
  - Sepal length
  - Sepal width
  - Petal length
  - Petal width
- Target variable: Species of the iris flower (categorical, 3 classes with 50 samples each):
  - Iris setosa
  - Iris versicolor
  - Iris Virginica

The dataset is balanced, well-behaved, and often used for classification tasks, clustering, and statistical testing. The three species can be largely distinguished based on these measurements, especially petal length and width.

Data Cleaning and Exploration

To be able to load the dataset, clean, explore, visualize and perform all the analyses, we have to load the necessary libraries. For this project, ggplot2, car, dplyr and DescTools are used. After calling the libraries, we the load the data and explore accordingly. From the data exploration, there are no missing values and all 150 observations. No handling is required, as the data is clean and complete.

Exploring the Data

```
[1] 0
```

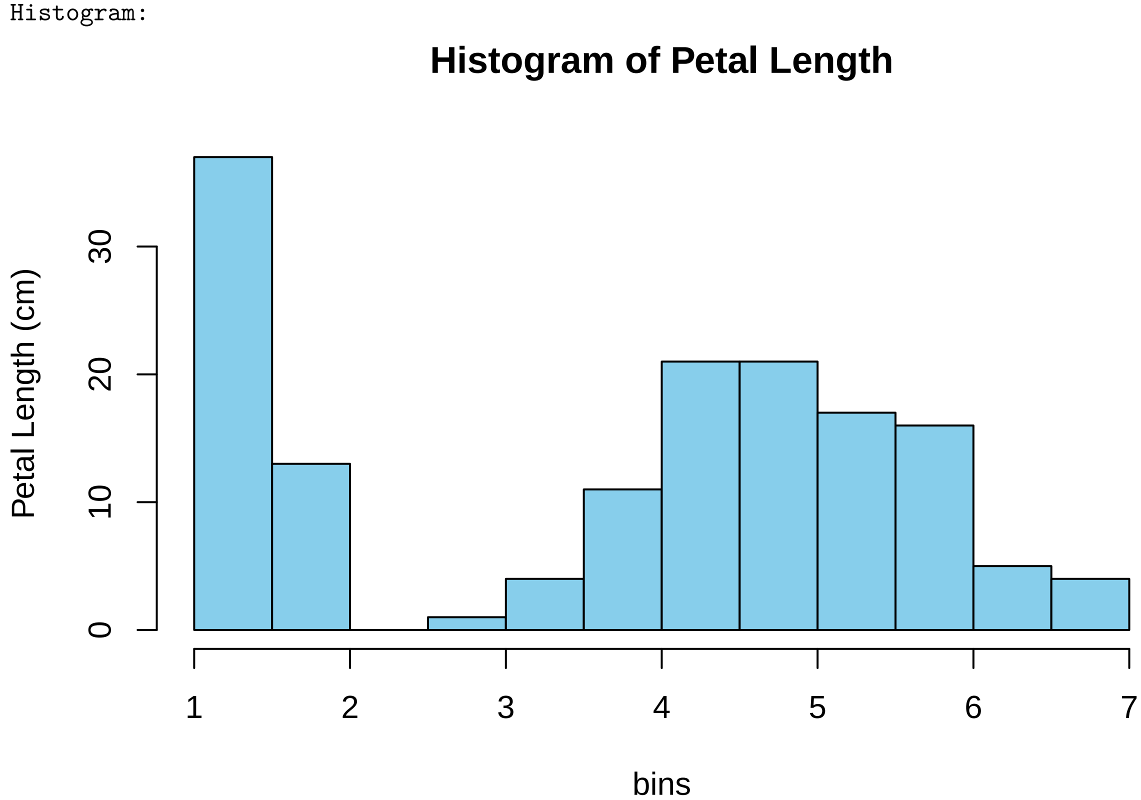
```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
           0            0            0            0            0
```

Summary Statistics and EDA

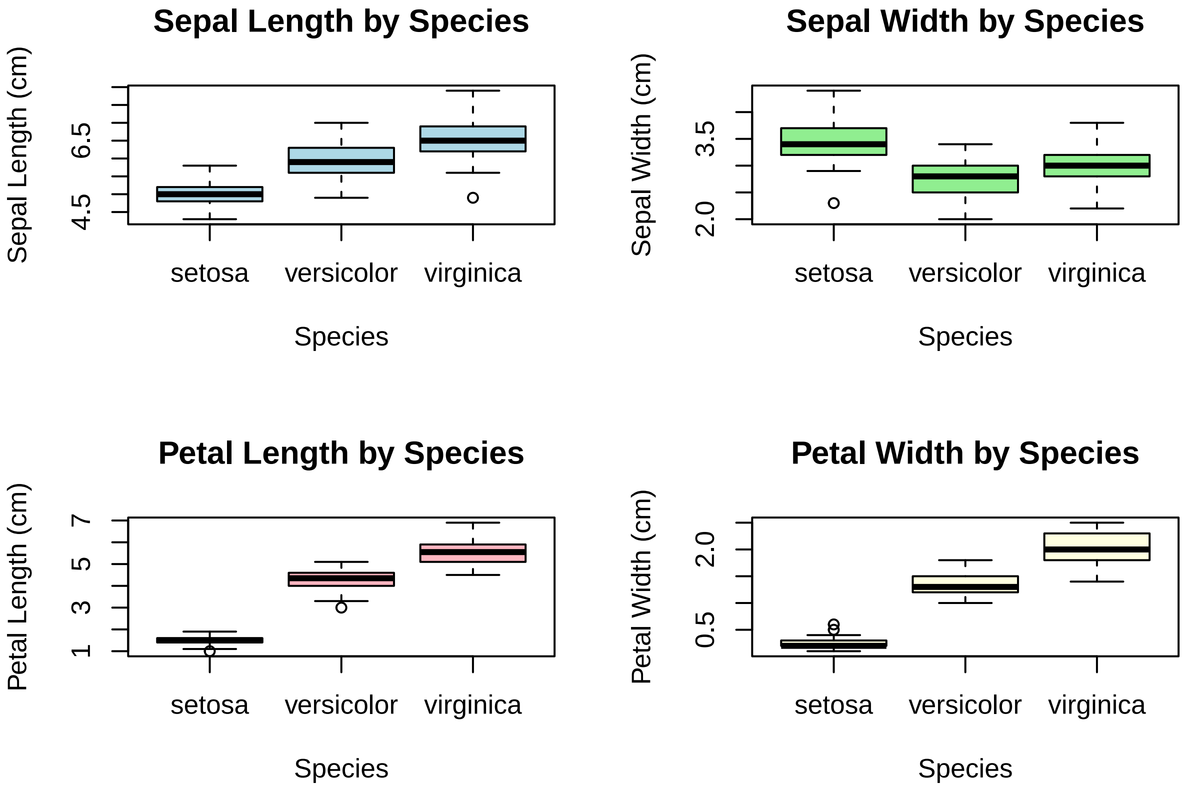
```
Summary Statistics:

  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
```

```
Histogram:
```



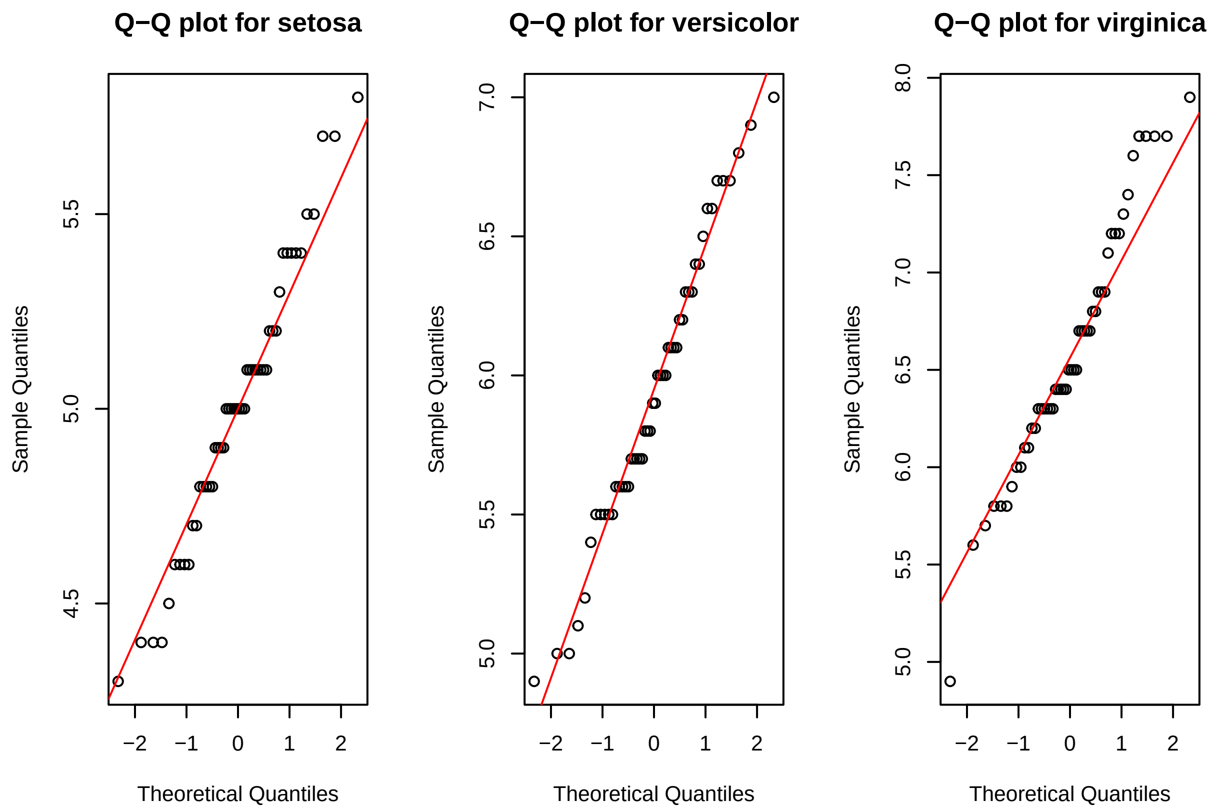
```
Box Plots:
```



Assumption Checking

Before conducting ANOVA, we have to check for the following assumptions:

- Normality: Are the response variable values within each group approximately normally distributed? We use visual methods (histograms, Q-Q plots)
- Homogeneity of Variances: Are the group variances approximately equal? Use Levene's test or Bartlett's test.




```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   2  6.3527 0.002259 **
      147
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the plots, distributions of sepal length and sepal width generally adhere to normality across all three species (setosa, versicolor, and virginica), although for setosa and virginica, there is a slight deviation at the 1-2 quantiles.

ANOVA Analysis

The null and alternative hypotheses are:

- H0 : The means of all groups are equal.
- H1 : At least one group mean is different.

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  63.21  31.606   119.3 <2e-16 ***
Residuals   147  38.96   0.265
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


ANOVA Results:

             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  63.21  31.606   119.3 <2e-16 ***
Residuals   147  38.96   0.265
---
```

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

The F-statistic of approximately 119.3 is very large, indicating substantial variation between group means (species) relative to within-group variation. The p-value (2×10ˆ16) is less than 0 '***', inferring there is strong evidence to reject the null hypothesis that the mean sepal length is equal across all three species. The between-group sum of squares (63.21) accounts for most of the total variation, underscoring the species effect. While ANOVA confirms at least one pairwise difference exists, it does not specify which pairs differ—follow-up post-hoc tests, such as the Tukey's HSD, would reveal whether all three pairwise comparisons are significant.

```
        Length Class  Mode
Species 12     -none- numeric


Post-hoc Analysis (Tukey's HSD):

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Sepal.Length ~ Species, data = iris)

$Species
                      diff       lwr       upr p adj
versicolor-setosa    0.930 0.6862273 1.1737727     0
virginica-setosa     1.582 1.3382273 1.8257727     0
virginica-versicolor 0.652 0.4082273 0.8957727     0
```

Post-Hoc Analysis and Interpretation

All three pairwise comparisons show highly significant differences (adjusted p = 0.0000). Iris virginica has the longest mean sepal length, significantly longer than both versicolor (by 0.652 cm) and setosa (by 1.582 cm). Iris versicolor has intermediate sepal length, significantly longer than setosa (by 0.930 cm). Iris setosa has the shortest mean sepal length. The 95% confidence intervals for all differences do not include zero, confirming that none of the pairwise differences could reasonably be due to chance. Tukey's HSD confirms that all three Iris species have statistically distinct mean sepal lengths, with a clear ordering: setosa < versicolor < virginica. This supports the strong discriminatory power of sepal length for distinguishing among the species, consistent with the extremely significant ANOVA result (F = 119.3, p < 2e-16). Although petal measurements typically separate the species even more clearly, sepal length alone is sufficient to detect highly significant differences between all pairs.

Code

```
knitr::opts_chunk$set(comment = NA)
library(ggplot2)
library(car)
library(dplyr)
library(DescTools)
# Step 1: Load the iris dataset
data(iris)
sum(is.na(iris))# checking for total missing values
colSums(is.na(iris)) # missing values in columns
knitr::opts_chunk$set(comment = "")
cat("\nSummary Statistics:\n")
summarystats <- summary(iris)
print(summarystats)
cat("\nHistogram:\n")
hist(iris$Petal.Length,
     main="Histogram of Petal Length",
     col="skyblue",
     xlab = "bins",
     ylab = "Petal Length (cm)")
# boxplots for each of the four features grouped by species in a 2x2 layout:
cat("\nBox Plots:\n")
# set up a 2x2 plotting layout
par(mfrow = c(2, 2))

# Boxplots for each feature by Species
boxplot(Sepal.Length ~ Species, data = iris,
        main = "Sepal Length by Species",
        col = "lightblue",
        ylab = "Sepal Length (cm)")

boxplot(Sepal.Width ~ Species, data = iris,
        main = "Sepal Width by Species",
        col = "lightgreen",
        ylab = "Sepal Width (cm)")

boxplot(Petal.Length ~ Species, data = iris,
        main = "Petal Length by Species",
        col = "lightpink",
        ylab = "Petal Length (cm)")

boxplot(Petal.Width ~ Species, data = iris,
        main = "Petal Width by Species",
        col = "lightyellow",
        ylab = "Petal Width (cm)")
## Normality - Q-Q plot
par(mfrow = c(1, 3))  # Arrange the plots side by side
for (species in levels(iris$Species)) {
  qqnorm(iris$Sepal.Length[iris$Species == species], main = paste("Q-Q plot for", species))
  qqline(iris$Sepal.Length[iris$Species == species], col = "red")
}

## Homogeneity of variance - Levene's Test
```

```r
leveneTest(Sepal.Length ~ Species, data = iris)
# Perform one-way ANOVA
anova_result <- aov(Sepal.Length ~ Species, data = iris)

# Check the summary of ANOVA
summary(anova_result)
## ANOVA Results
cat("\nANOVA Results:\n")
anova_summary <- summary(anova_result)
print(anova_summary)
# Post-hoc analysis using Tukey's HSD test
post_hoc_result <- TukeyHSD(anova_result)
summary(post_hoc_result)
# Post-hoc analysis (Tukey's HSD test)
cat("\nPost-hoc Analysis (Tukey's HSD):\n")
print(post_hoc_result)
```