**Iris Dataset Analysis**

**Project Overview**

This repository contains code and analysis related to the Iris dataset, a classic dataset used for introductory data analysis, hypothesis testing, and correlation. The goal of this project is to explore the Iris dataset, visualize its characteristics, and conduct hypothesis testing and correlation analysis to understand the relationships between features and species.

**Data Exploration**

* **Dataset:** The Iris dataset contains measurements of sepal length, sepal width, petal length, and petal width for 150 iris flowers from three species: setosa, versicolor, and virginica.
* **Exploratory Data Analysis (EDA):**
  + Histograms to visualize the distribution of sepal length, sepal width, petal length, and petal width.
  + Scatter plots to examine relationships between features.
  + Box plots to compare the distributions of petal length and petal width across species.

**Code:**

# Load necessary libraries

library(ggplot2)

# Load the Iris dataset

data(iris)

# Histogram of sepal length

ggplot(iris, aes(x = Sepal.Length)) +

geom\_histogram() +

labs(title = "Sepal Length Distribution")

A graph of a number of blue bars

Description automatically generated

# Scatter plot of sepal length vs. sepal width by species

ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +

geom\_point() +

labs(title = "Sepal Length vs. Sepal Width   by Species")

A graph with different colored dots

Description automatically generated

# Box plot of petal length by species

ggplot(iris, aes(x = Species, y = Petal.Length)) +

geom\_boxplot() +

labs(title = "Petal Length by Species")

A graph with different sizes and shapes

Description automatically generated with medium confidence

**Hypothesis Testing and Correlation Analysis**

* **Hypothesis Testing:** Conduct an ANOVA test to determine if there are significant differences in sepal length among the species.
* **Correlation Analysis:** Calculate correlation coefficients between sepal length, sepal width, petal length, and petal width.

**Code:**

# ANOVA test

model <- aov(Sepal.Length ~ Species, data = iris)

summary(model)

A screenshot of a computer

Description automatically generated

# Correlation matrix

correlation\_matrix <- cor(iris[, 1:4])

print(correlation\_matrix)

A screenshot of a computer code

Description automatically generated

**Results**

* **ANOVA:** The ANOVA test shows that there are significant differences in sepal length among the three species (p-value < 2e-16).
* **Correlation:** The correlation matrix reveals strong positive correlations between sepal length and petal length, as well as between sepal width and petal width. There are negative correlations between sepal length and sepal width, and between petal length and petal width.

**Conclusions**

Based on the analysis, we can conclude that sepal length is a significant factor in distinguishing the iris species. Additionally, the strong correlations between sepal length and petal length, as well as between sepal width and petal width, suggest that these features are interrelated.

**Note:** This repository provides a basic framework for analyzing the Iris dataset. You can extend it further by exploring other statistical techniques, machine learning algorithms, or visualization methods.