Iris Flower Dataset Analysis

A classic dataset in statistical classification

Deri Siswara

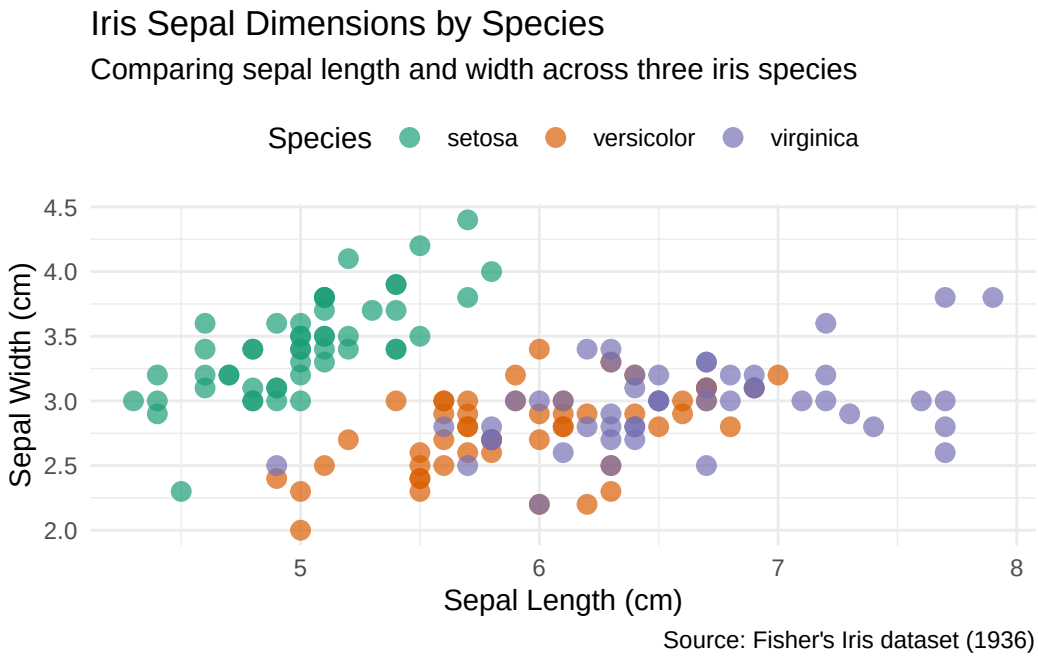
Species of Iris Flowers

```
# Summarize iris data by species
iris_summary <- iris %>%
  group_by(Species) %>%
  summarize(
    "Sepal Length" = mean(Sepal.Length),
    "Sepal Width" = mean(Sepal.Width),
    "Petal Length" = mean(Petal.Length),
    "Petal Width" = mean(Petal.Width)
  )
```

The dataset contains three species of Iris flowers: **setosa, versicolor, virginica**. The species with the largest average sepal length is **virginica**.

```
iris %>%
  ggplot(
    aes(x = Sepal.Length, y = Sepal.Width, color = Species)
  ) +
  geom_point(size = 3, alpha = 0.7) +
  labs(
    title = "Iris Sepal Dimensions by Species",
    subtitle = "Comparing sepal length and width across three iris species",
    color = "Species",
    caption = "Source: Fisher's Iris dataset (1936)"
  ) +
  xlab("Sepal Length (cm)") +
  ylab("Sepal Width (cm)") +
  scale_color_brewer(palette = "Dark2") +
```

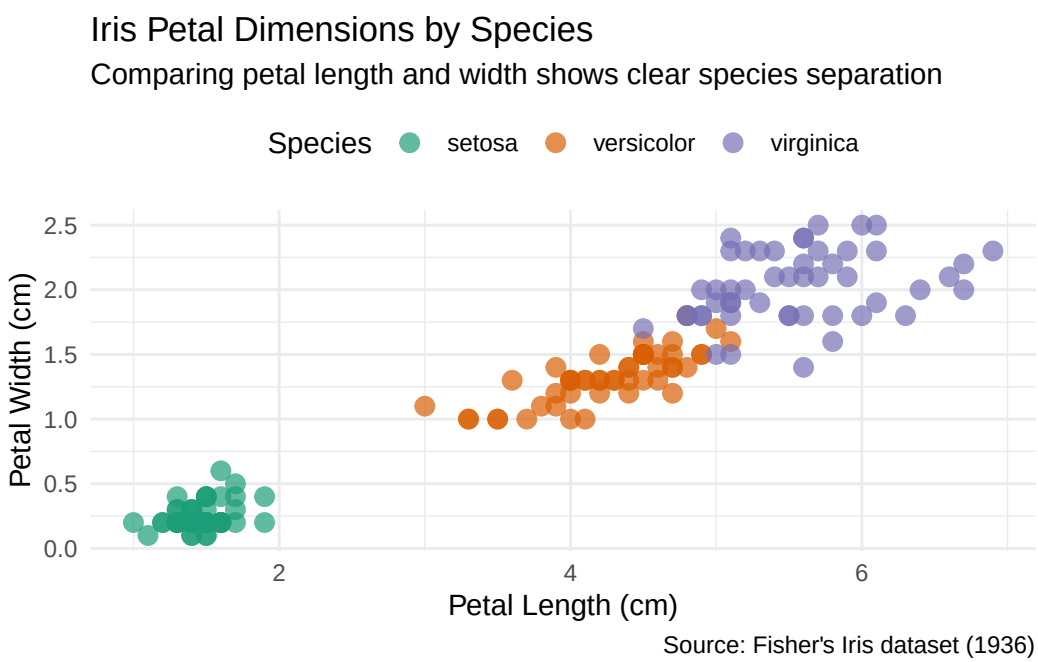
```
  theme_minimal() +
  theme(
    plot.caption.position = "plot",
    legend.position = "top"
  )
```



Petal Characteristics

```
iris %>%
  ggplot(
    aes(x = Petal.Length, y = Petal.Width, color = Species)
  ) +
  geom_point(size = 3, alpha = 0.7) +
  labs(
    title = "Iris Petal Dimensions by Species",
    subtitle = "Comparing petal length and width shows clear species separation",
    color = "Species",
    caption = "Source: Fisher's Iris dataset (1936)"
  ) +
  xlab("Petal Length (cm)") +
```

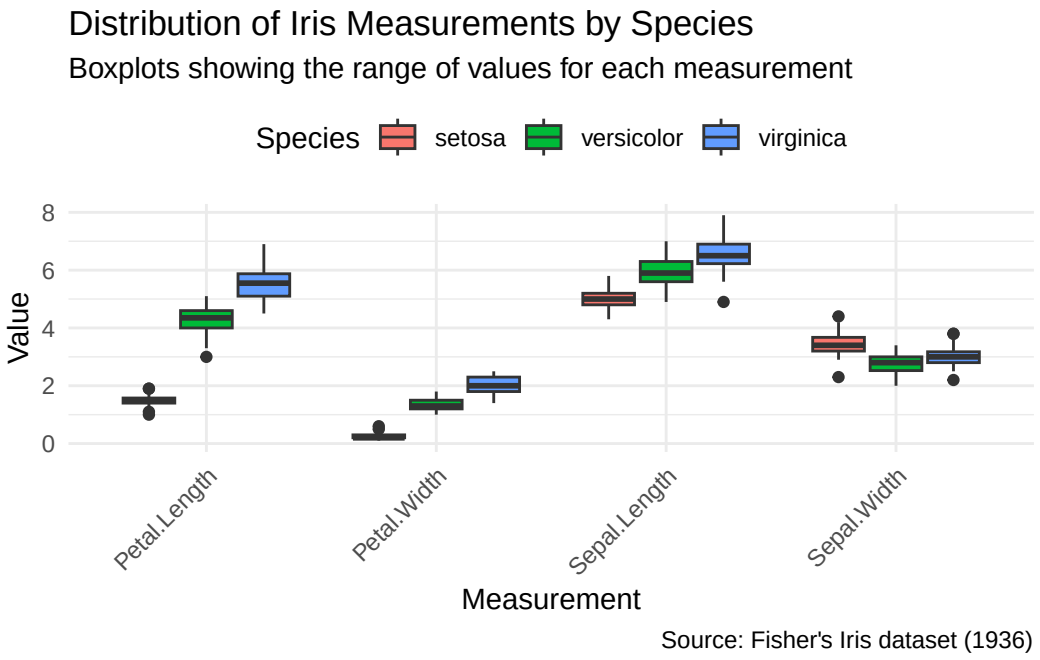
```
  ylab("Petal Width (cm)") +
  scale_color_brewer(palette = "Dark2") +
  theme_minimal() +
  theme(
    plot.caption.position = "plot",
    legend.position = "top"
  )
```



Correlation Between Measurements

```
iris %>%
  pivot_longer(cols = -Species,
               names_to = "Measurement",
               values_to = "Value") %>%
  ggplot(aes(x = Measurement, y = Value, fill = Species)) +
  geom_boxplot() +
  labs(
    title = "Distribution of Iris Measurements by Species",
    subtitle = "Boxplots showing the range of values for each measurement",
    fill = "Species",
```

```
    caption = "Source: Fisher's Iris dataset (1936)"
  ) +
  theme_minimal() +
  theme(
    plot.caption.position = "plot",
    legend.position = "top",
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



About the Dataset

The Iris flower dataset is a multivariate dataset introduced by the British statistician and biologist Ronald Fisher in his 1936 paper. It includes 50 samples from each of three species of Iris:

- Iris setosa
- Iris virginica
- Iris versicolor

Four features were measured from each sample: the length and width of the sepals and petals.

This dataset is often used for classification tasks in machine learning and statistics as a standard testing dataset.