Descriptive Analysis and Visualization of the Iris Dataset

Zemelak Goraga

2024-01-20

```
# Task 1: Load the iris dataset
library(datasets)
data(iris)
```

```
# Task 2: Inspect the iris dataset
head(iris)  # Display the first few rows of the dataset
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
summary(iris)  # Summary statistics of the dataset
```

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

```
# Task 3: Assign the iris dataset to a new variable(df)
df <- iris
```

```
# Task 4: Using dplyr to group by species and provide average sepal length
library(dplyr)
```

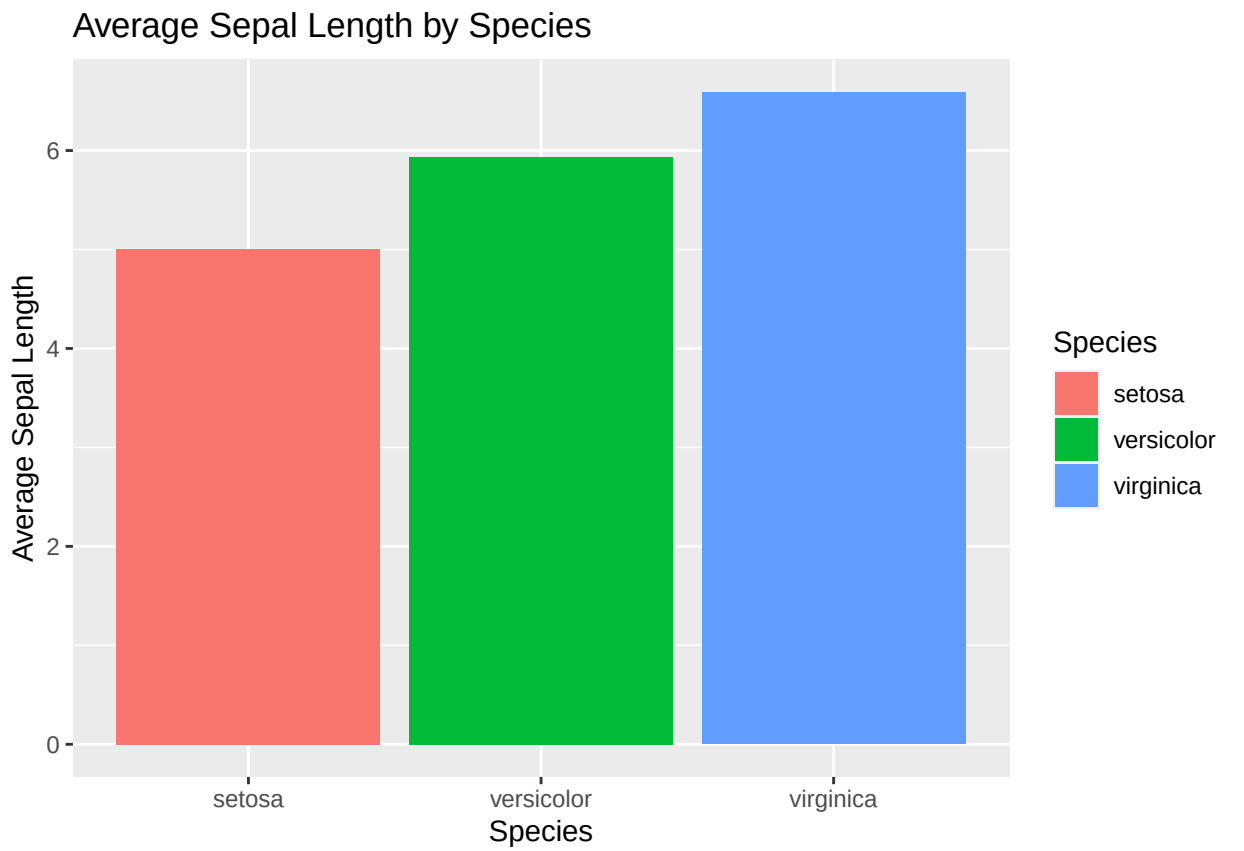
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
average_sepal_length <- df %>%
  group_by(Species) %>%
  summarise(avg_sepal_length = mean(Sepal.Length))

# Display the summary_data dataframe
print(average_sepal_length)
```

```
## # A tibble: 3 x 2
##   Species    avg_sepal_length
##   <fct>                 <dbl>
## 1 setosa                 5.01
## 2 versicolor             5.94
## 3 virginica              6.59
```

```r
# Task 4: Visualization using ggplot2
library(ggplot2)

# Create a bar plot
ggplot(average_sepal_length, aes(x = Species, y = avg_sepal_length, fill = Species)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Average Sepal Length by Species", x = "Species", y = "Average Sepal Length")
```




```
average_sepal_width <- iris %>%
  group_by(Species) %>%
  summarise(avg_sepal_width = mean(Sepal.Width))

# Display the summary_data dataframe
print(average_sepal_width)
```

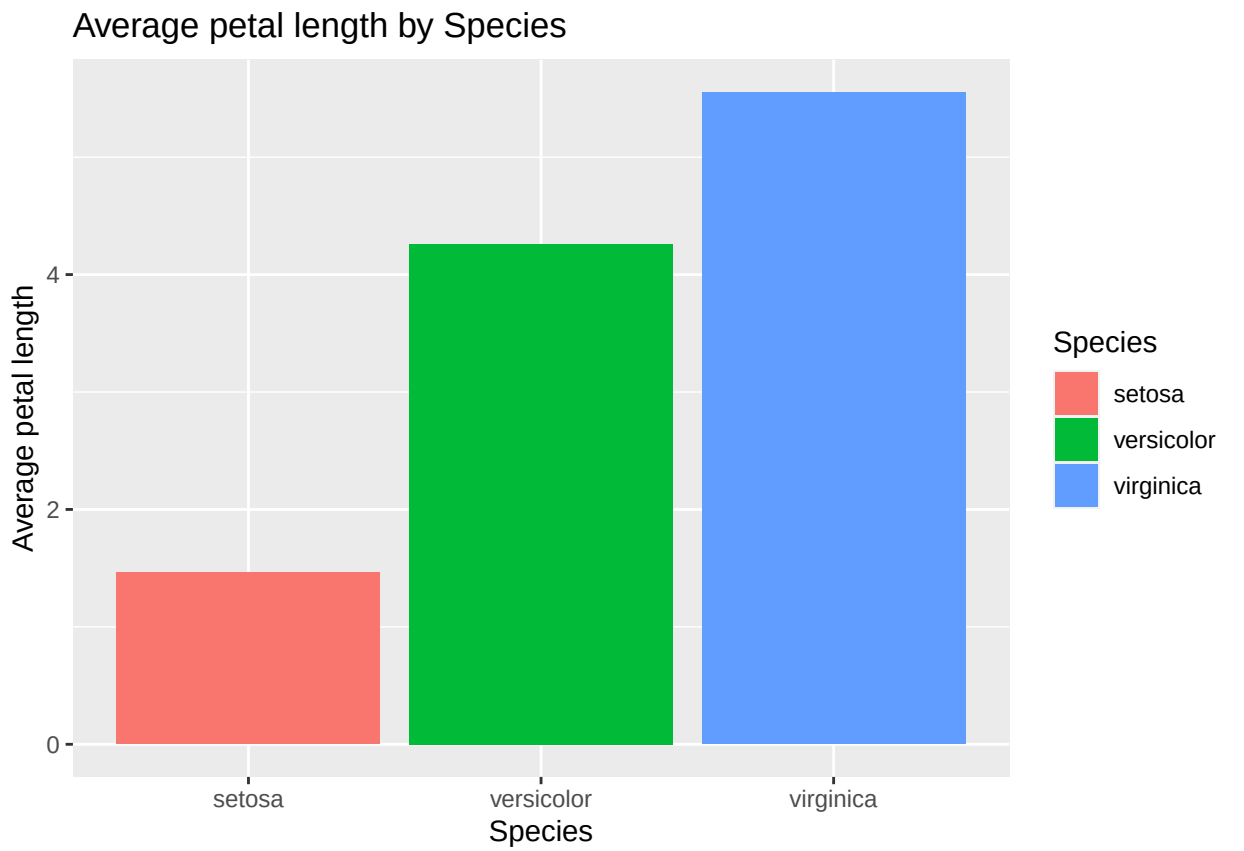
```
## # A tibble: 3 x 2
##   Species    avg_sepal_width
##   <fct>                <dbl>
## 1 setosa                3.43
## 2 versicolor            2.77
## 3 virginica             2.97
```

```
# Create a bar plot
ggplot(average_sepal_width, aes(x = Species, y = avg_sepal_width, fill = Species)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Average Sepal width by Species", x = "Species", y = "Average Sepal width")
```


```r
average_petal_length <- iris %>%
  group_by(Species) %>%
  summarise(avg_petal_length = mean(Petal.Length))

# Display the summary_data dataframe
print(average_petal_length)
```

```
## # A tibble: 3 x 2
##   Species    avg_petal_length
##   <fct>                 <dbl>
## 1 setosa                 1.46
## 2 versicolor             4.26
## 3 virginica              5.55
```

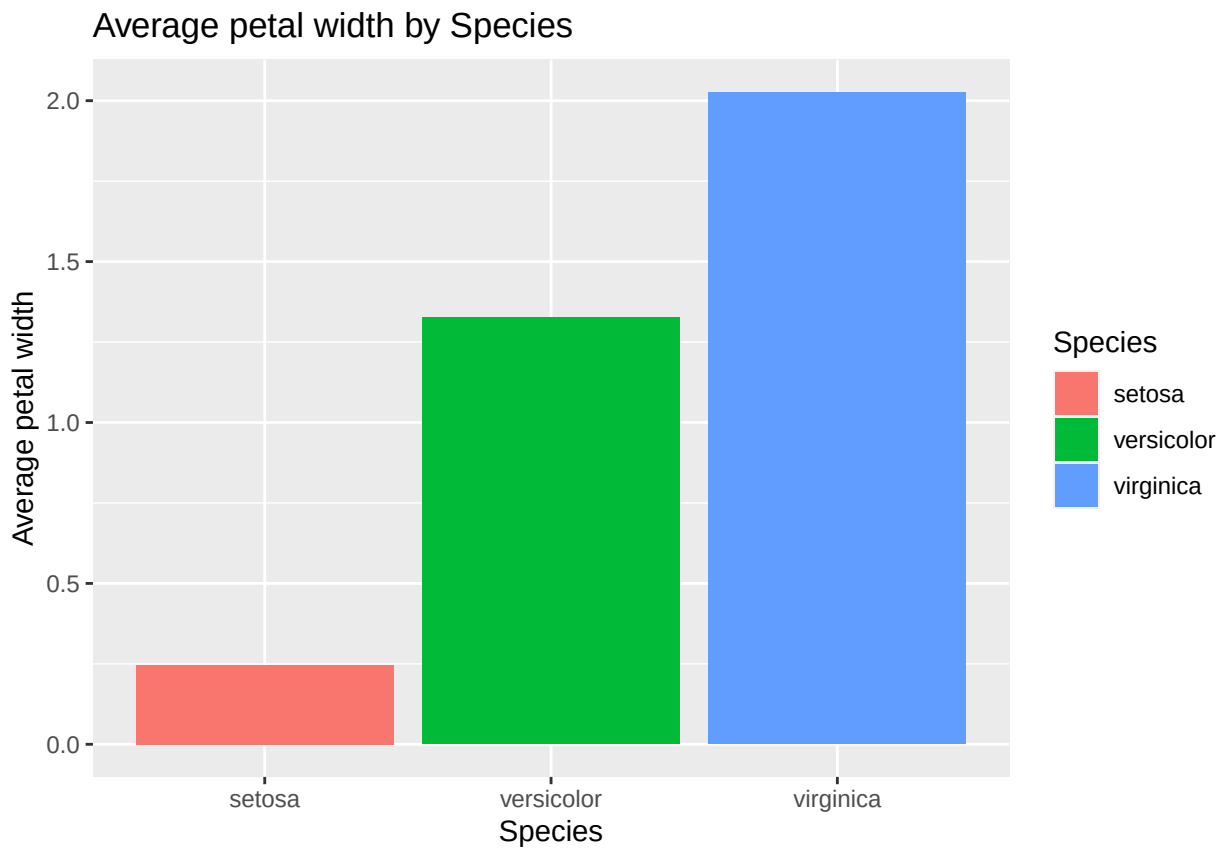
```r
# Create a bar plot
ggplot(average_petal_length, aes(x = Species, y = avg_petal_length, fill = Species)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Average petal length by Species", x = "Species", y = "Average petal length")
```



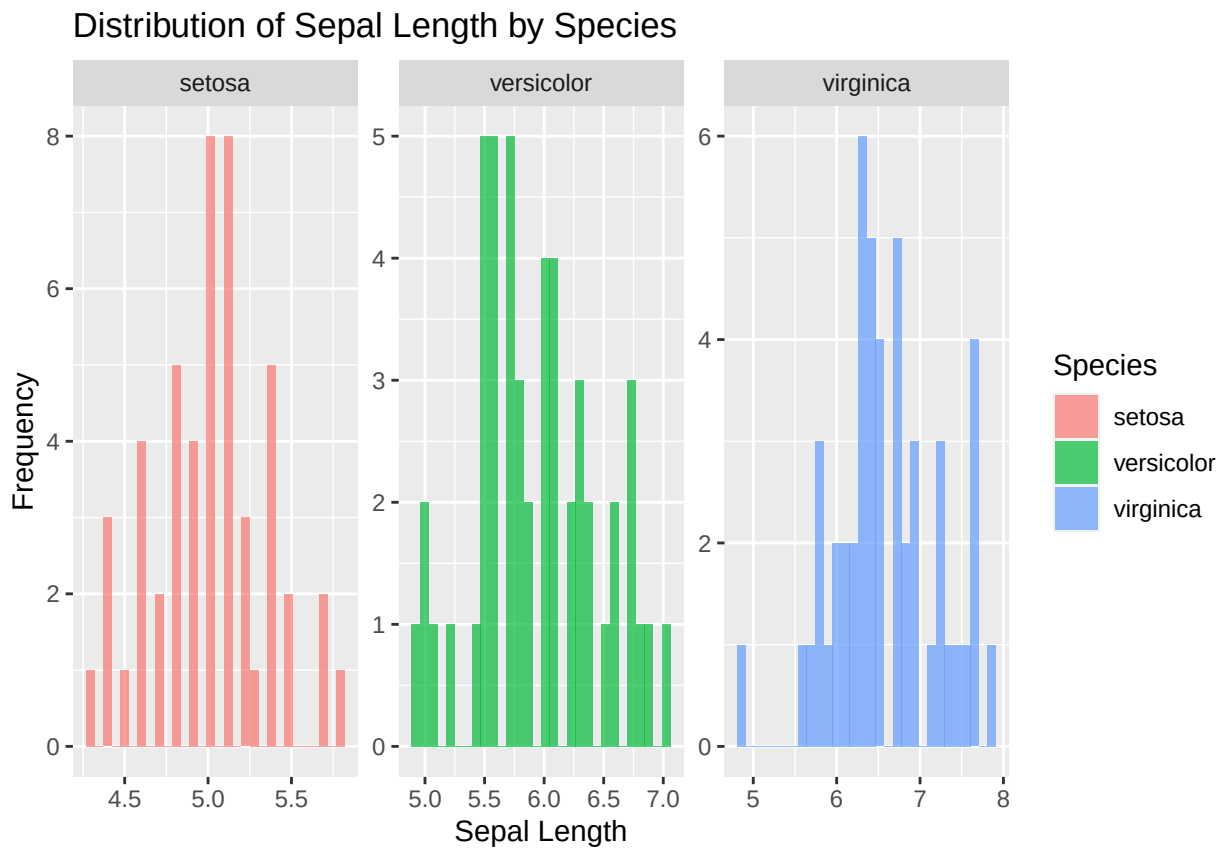

```
average_petal_width <- iris %>%
  group_by(Species) %>%
  summarise(avg_petal_width = mean(Petal.Width))

# Display the summary_data dataframe
print(average_petal_width)
```

```
## # A tibble: 3 x 2
##   Species    avg_petal_width
##   <fct>                <dbl>
## 1 setosa               0.246
## 2 versicolor           1.33
## 3 virginica            2.03
```

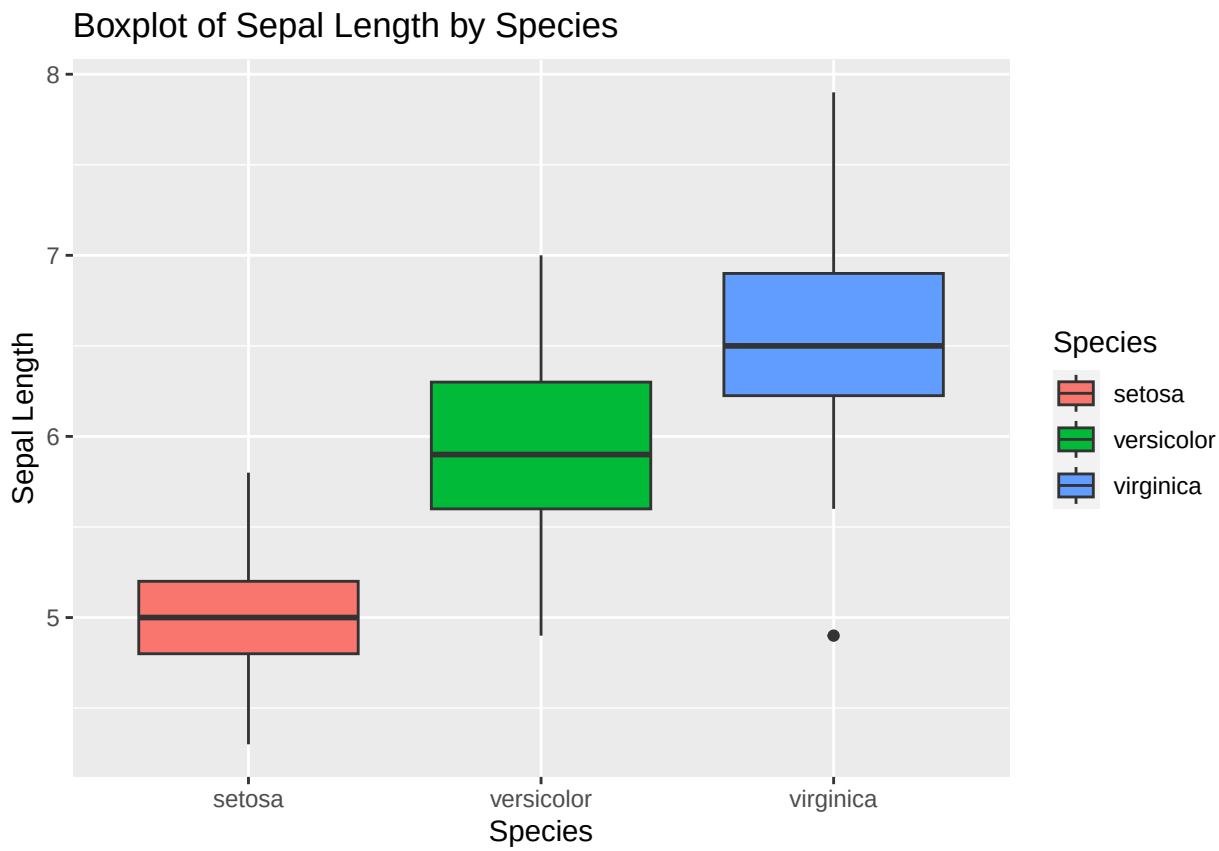
```
ggplot(average_petal_width, aes(x = Species, y = avg_petal_width, fill = Species)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Average petal width by Species", x = "Species", y = "Average petal width")
```



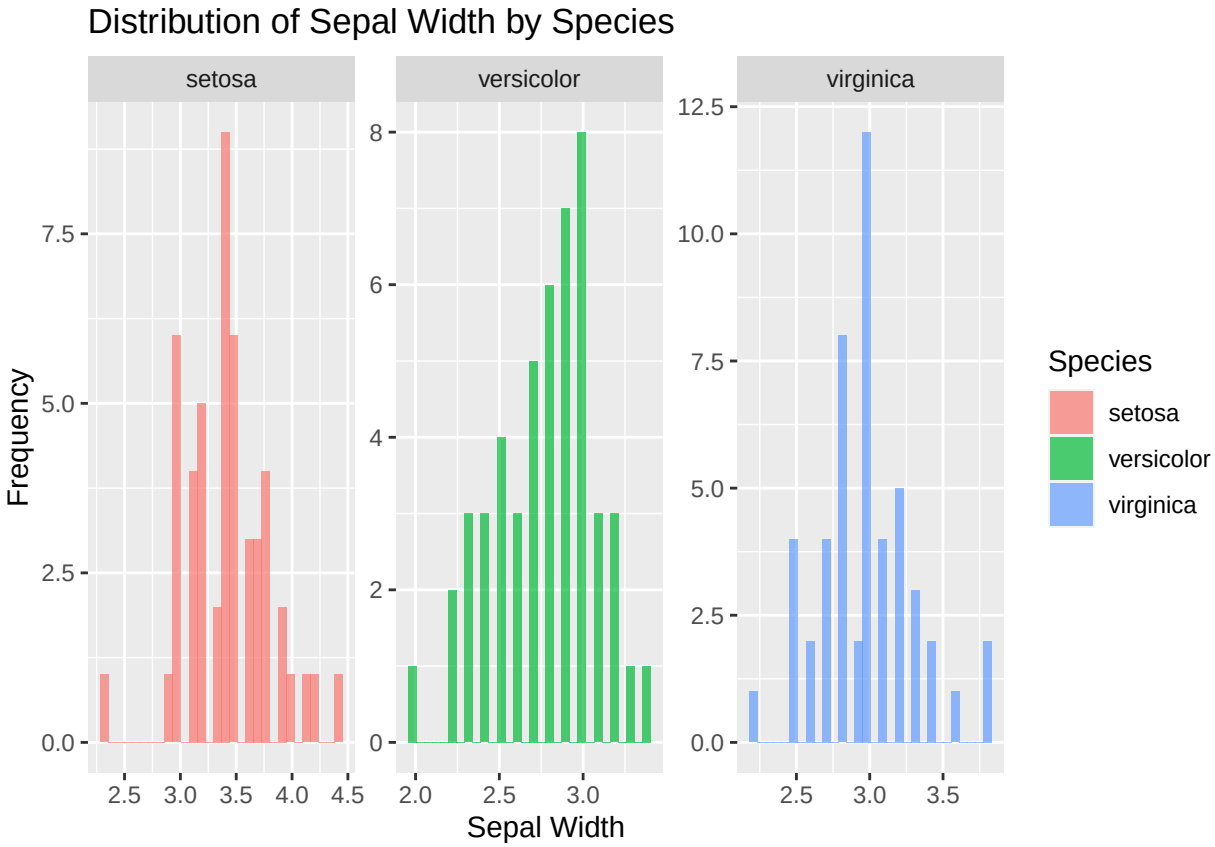

```
# Task 5: Visualizing Differences in Sepal Length, Sepal Width, Petal Length, and Petal Width using his
# 5.1. Sepal Length
ggplot(df, aes(x = Sepal.Length, fill = Species)) +
  geom_histogram(position = "identity", alpha = 0.7, bins = 30) +
  labs(title = "Distribution of Sepal Length by Species", x = "Sepal Length", y = "Frequency") +
  facet_wrap(~Species, scales = "free")
```



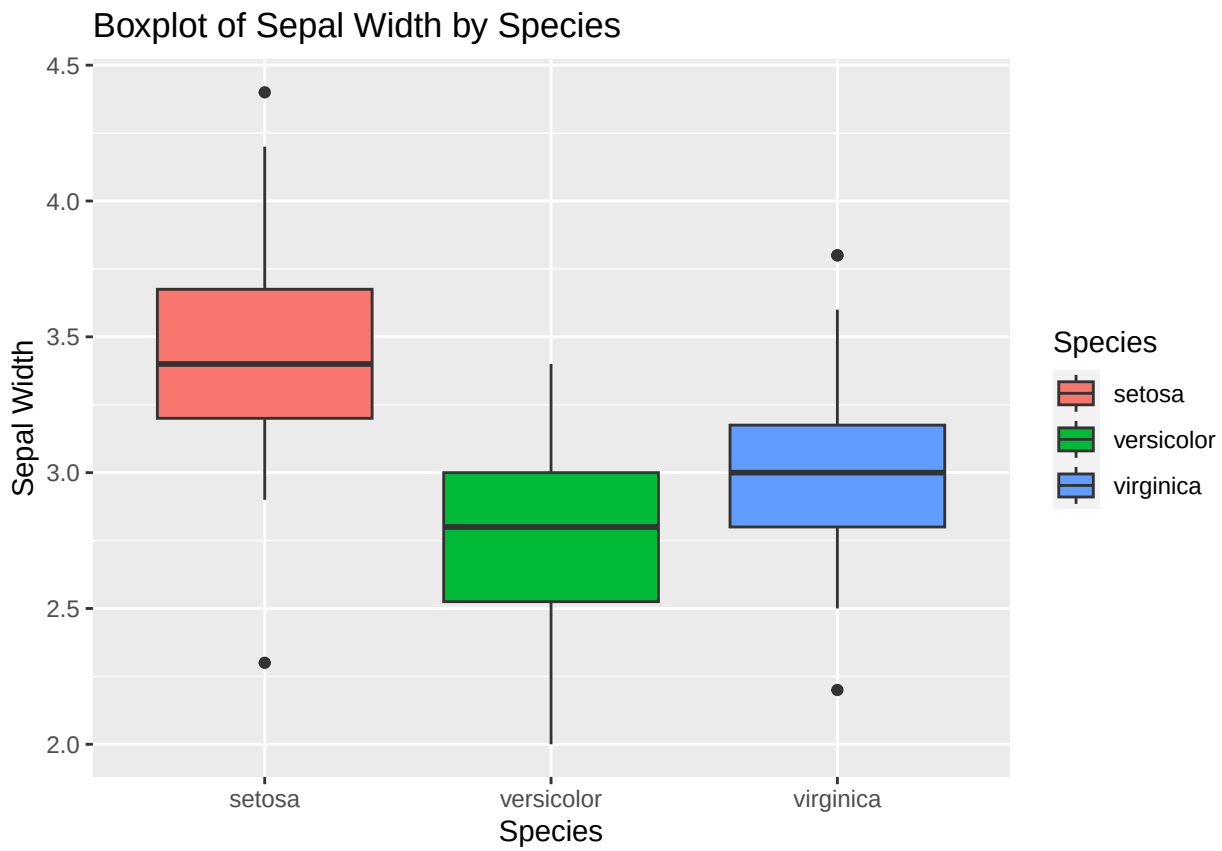

```
ggplot(df, aes(x = Species, y = Sepal.Length, fill = Species)) +
  geom_boxplot() +
  labs(title = "Boxplot of Sepal Length by Species", x = "Species", y = "Sepal Length")
```



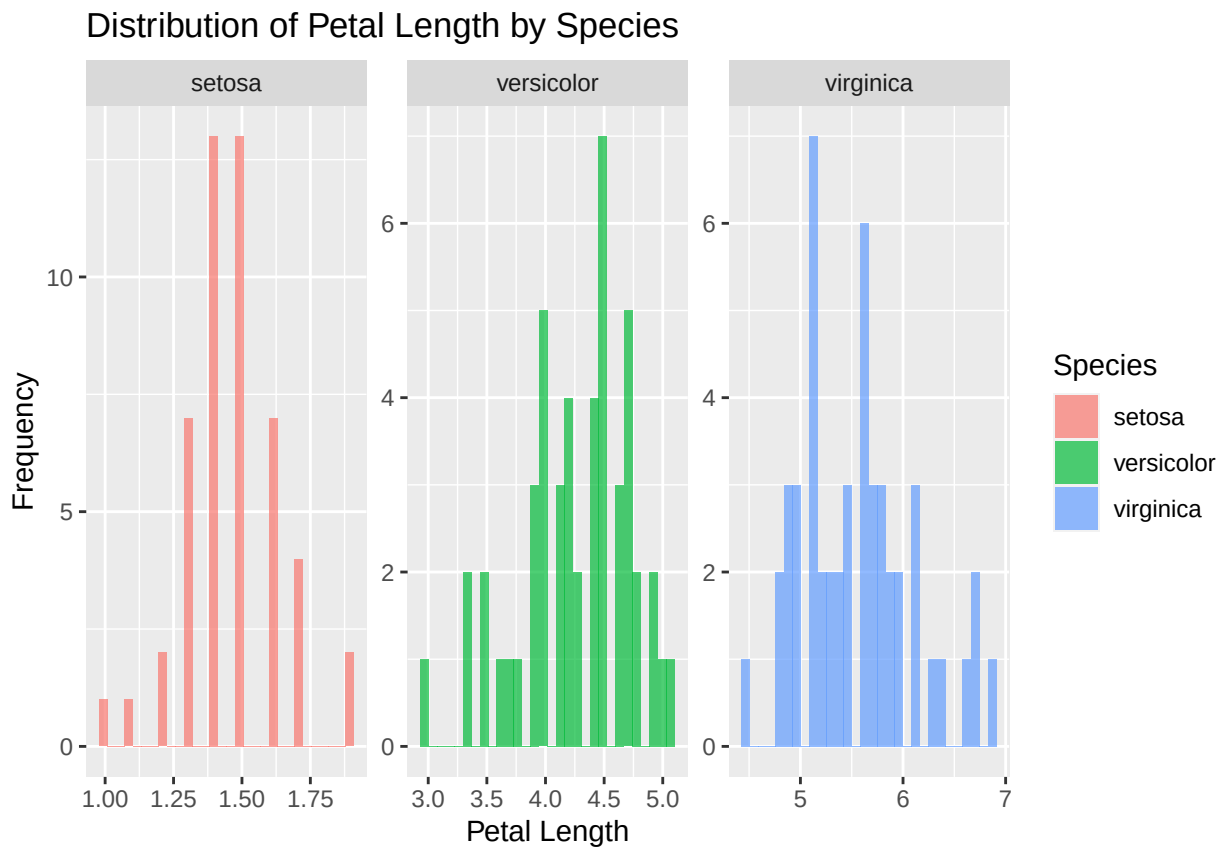

```
# 5.2. Sepal Width

ggplot(df, aes(x = Sepal.Width, fill = Species)) +
  geom_histogram(position = "identity", alpha = 0.7, bins = 30) +
  labs(title = "Distribution of Sepal Width by Species", x = "Sepal Width", y = "Frequency") +
  facet_wrap(~Species, scales = "free")
```



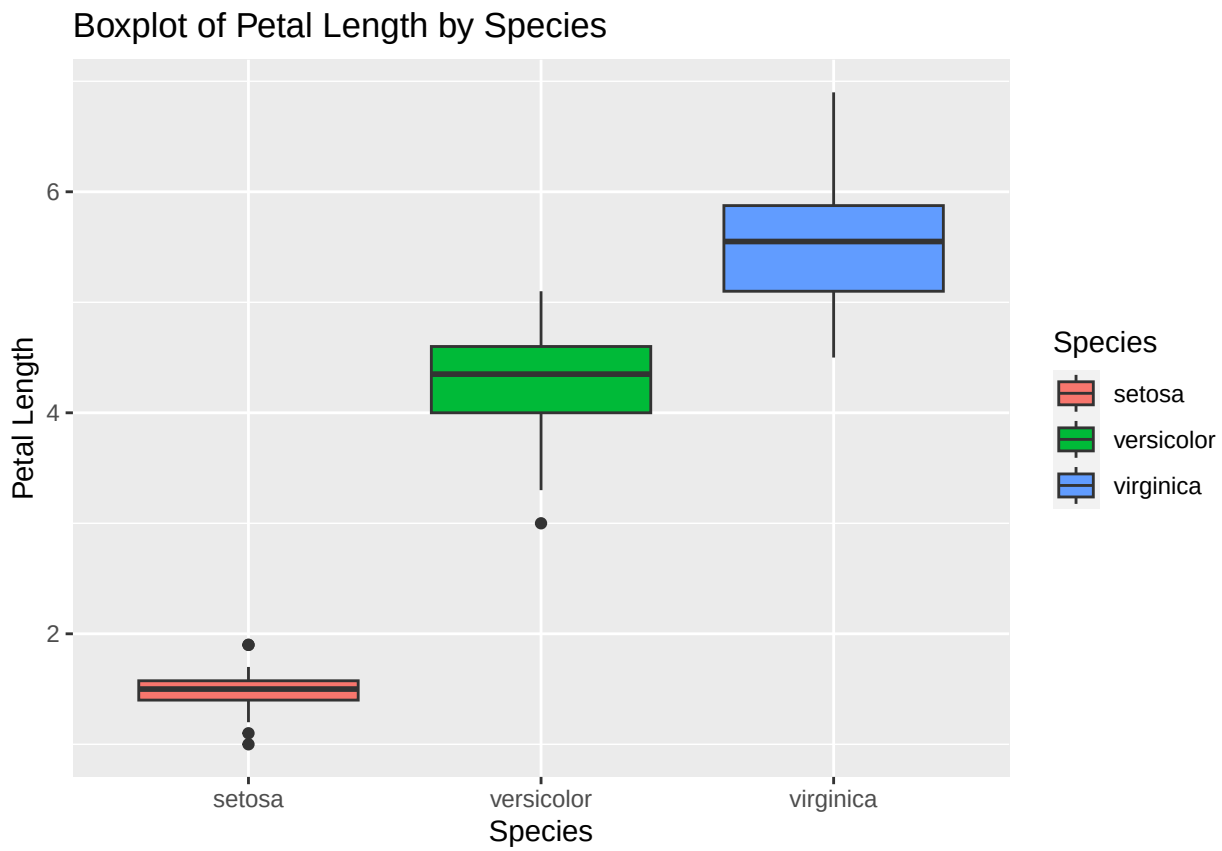

```r
ggplot(df, aes(x = Species, y = Sepal.Width, fill = Species)) +
  geom_boxplot() +
  labs(title = "Boxplot of Sepal Width by Species", x = "Species", y = "Sepal Width")
```



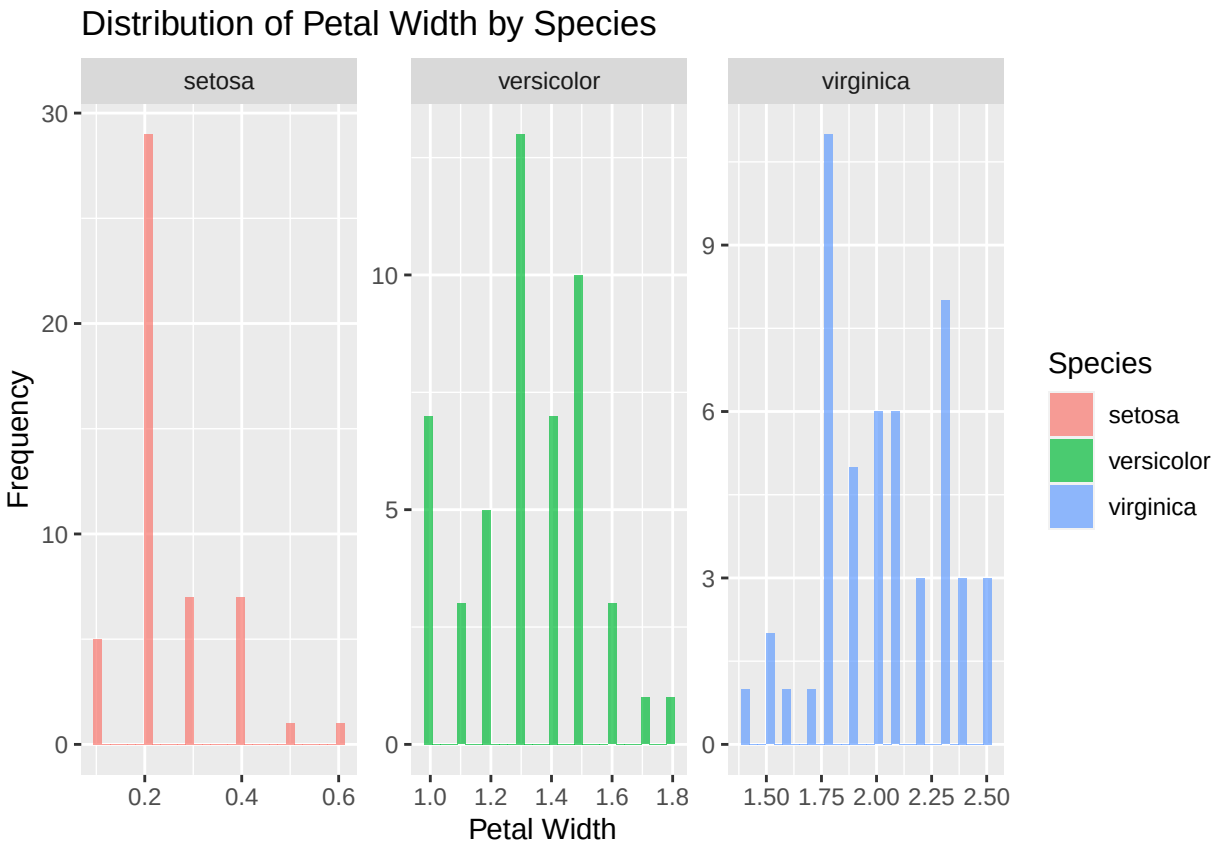

```r
# 5.3. Petal Length

ggplot(df, aes(x = Petal.Length, fill = Species)) +
  geom_histogram(position = "identity", alpha = 0.7, bins = 30) +
  labs(title = "Distribution of Petal Length by Species", x = "Petal Length", y = "Frequency") +
  facet_wrap(~Species, scales = "free")
```



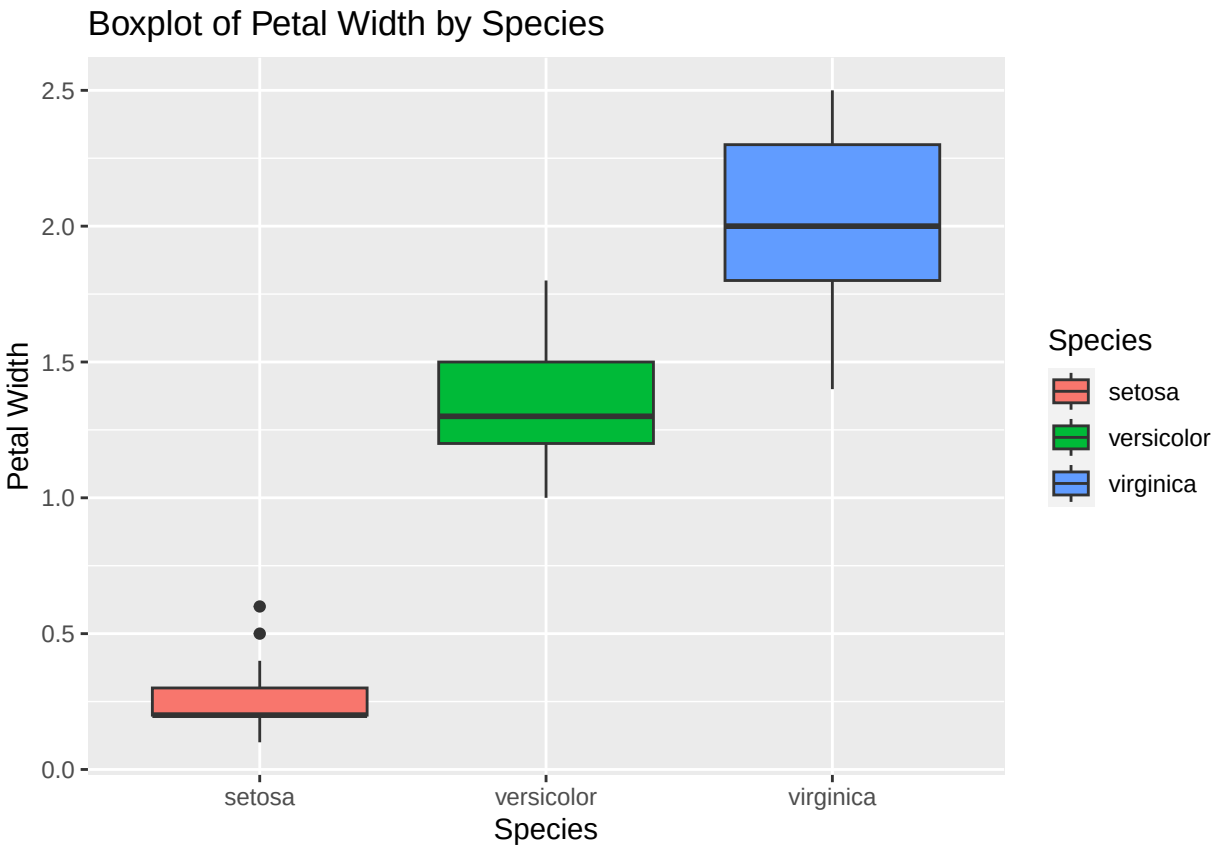

```r
ggplot(df, aes(x = Species, y = Petal.Length, fill = Species)) +
  geom_boxplot() +
  labs(title = "Boxplot of Petal Length by Species", x = "Species", y = "Petal Length")
```




```r
# 5.4. Petal Width

ggplot(df, aes(x = Petal.Width, fill = Species)) +
  geom_histogram(position = "identity", alpha = 0.7, bins = 30) +
  labs(title = "Distribution of Petal Width by Species", x = "Petal Width", y = "Frequency") +
  facet_wrap(~Species, scales = "free")
```




```
ggplot(df, aes(x = Species, y = Petal.Width, fill = Species)) +
  geom_boxplot() +
  labs(title = "Boxplot of Petal Width by Species", x = "Species", y = "Petal Width")
```



Task 6: Describing Differences Between Species

The dataset comprises three species of iris flowers: setosa, versicolor, and virginica. Descriptive statistics provide an overall view of the dataset, including the number of observations for each species. The average sepal length varies across species, with setosa having the shortest (5.006), followed by versicolor (5.936), and then virginica (6.588). In terms of sepal width, setosa has the highest average (3.428), while versicolor (2.770) and virginica (2.974) are relatively narrower. For petal length, setosa has the smallest average (1.462), followed by versicolor (4.260) and virginica (5.552), indicating an increasing trend. The average petal width also shows a similar pattern, with setosa having the smallest (0.246), followed by versicolor (1.326), and then virginica (2.026).

Visualizations with histograms and boxplots for each variable did further illustrate these differences. Sepal length tends to increase from setosa to virginica, while sepal width decreases. Petal length and width exhibit significant increases from setosa to virginica. These visualizations enhance our understanding of the variations in the selected variables among the three iris species.