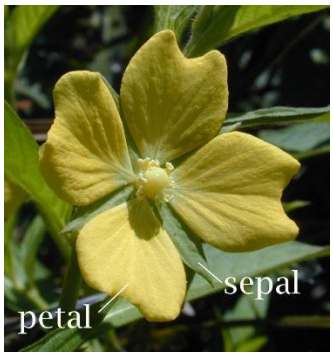
CYBR 520
Lab 1: Introduction to Exploratory Data Analysis (R version)



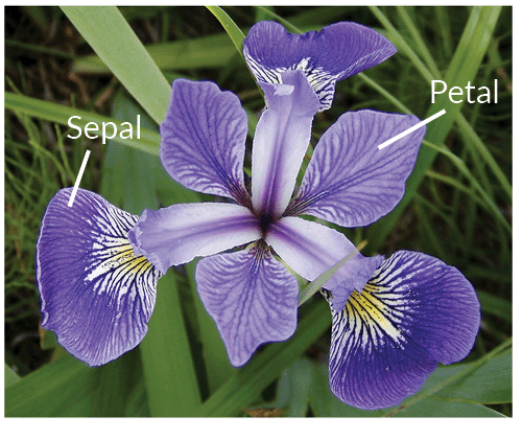
Dataset: Iris

Topics: Summary statistics, data visualization, variable relationships, feature distributions, and outlier detection

R Packages: tidyverse, tidymodels, GGally

Background: The iris dataset is a classic example dataset used for machine learning in R. Most, if not all flowers have a sepal and a petal. The sepal (Figure 1) [1] functions as a protector for the flower and also support the petals when the flower is in bloom:

The Iris dataset contains measurements (in cm) of petal length and width, and sepal length and width for 3 species of the iris flower: *iris versicolor, iris setosa,* and *iris virginica* (Figure 2)[2]. There are exactly 50 observations for each species type, bringing the total number of observations to 150. Each observation is also assigned membership to a specific species as well. The membership (aka "class label") is assigned based upon the name of the species "versicolor", "setosa", "virginica".



Iris Versicolor

Iris Setosa

Iris Virginica

[1] https://en.wikipedia.org/wiki/Petal#/media/File:Petal-sepal.jpg
[2] https://www.datacamp.com/community/tutorials/machine-learning-in-r

Lab Structure

The first part of this lab is a walkthrough with basic questions. Note that a companion walkthrough video is available on eCampus. Please answer each of the questions after the associated section of the walkthrough.

The second part of the lab, is where you will explore the data and test your analytics capabilities. Answer each question fully, and present all code and plots associated with answering each question.

Reproducibility of your code and your analysis to answer the questions are important, so include those as part of your companion walkthrough or README file with your lab submission.

REMEMBER: Take your time with this. Don't just rush through to complete the lab. Understand what you are doing, and how to leverage analytical techniques and software to solve cybersecurity problems. This lab is meant to expand your knowledge and skillset, not just for a grade…

1. Setup and Data Loading

Launch R Studio

Install the following packages:

```
library(tidyverse)
library(tidymodels)
library(GGally)
```

Now we need to load the data. Type:

```
data(iris)
```

Now set iris data to a variable

```
iris_tbl <- tibble::as_tibble(iris)
```

Notice that the iris data should now show up in your global environment. Lets now take a look at this dataset by typing:

```
glimpse(iris_tbl)
```

Question 1. What are the variable names and their data types? How many rows and columns does the dataset have?

There are 150 rows and 5 columns. The variable names are: Sepal.Length (double), Sepal.Width (double), Petal.Length (double), Petal.Width (double), Species (factor).

```
> glimpse(iris_tbl)
Rows: 150
Columns: 5
$ Sepal.Length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9,…
$ Sepal.Width  <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1,…
$ Petal.Length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5,…
$ Petal.Width  <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1,…
$ Species      <fct> setosa, setosa, setosa, setosa, setosa, setosa, s…
>
```

2. Exploratory Data Analysis – Iris Dataset

Now lets do some exploratory data analysis. Exploratory data analysis is a process of performing an initially investigation of the dataset using analysis and visualization techniques to discover patterns, spot outliers and anomalies, and relationships between variables in the data. This is an often skipped, but critical step in any analysis. It can also help you develop a more informed hypothesis for scientific inquiries. In essence, we're going to try to gain some insight on this dataset.

```
sum_wide <- iris_tbl %>%
  summarise(across(
    where(is.numeric),
    list(mean = ~mean(.x), sd = ~sd(.x), min = ~min(.x), max = ~max(.x)),
    .names = "{.col}_{.fn}"
  ))

print(sum_wide, n = Inf, width = Inf)
```

Question 2. What can you tell about the range and variability of each measurement?

Sepal.Length: Range = 7.9 - 4.3 =3.6, Variability = 0.828

Sepal.Length has a moderate variability which suggests that the sepal lengths vary moderately from the mean of 3.6.

Sepal.Width: Range = 4.4 - 2 = 2.4, Variability = 0.436

Sepal.Width has the smallest range and lowest standard deviation meaning that sepal width is usually close to the mean of 2.4.

Petal.Length: Range = 6.9 - 1 = 5.9, Variability = 1.77

Petal.Length has the largest range and highest standard deviation meaning that petal lengths vary greatly among the iris species.

Petal.Width: Range = 2.5 - 0.1 = 2.4, Variability = 0.762

Petal.Width has a moderate variability which suggests that the petal widths vary moderately from the mean of 2.4.

```
# A tibble: 1 × 16
  Sepal.Length_mean Sepal.Length_sd Sepal.Length_min Sepal.Length_max
              <dbl>           <dbl>             <dbl>             <dbl>
1              5.84           0.828               4.3               7.9
  Sepal.Width_mean Sepal.Width_sd Sepal.Width_min Sepal.Width_max
             <dbl>          <dbl>            <dbl>           <dbl>
1             3.06          0.436                2             4.4
  Petal.Length_mean Petal.Length_sd Petal.Length_min Petal.Length_max
              <dbl>           <dbl>             <dbl>             <dbl>
1              3.76            1.77                 1               6.9
  Petal.Width_mean Petal.Width_sd Petal.Width_min Petal.Width_max
             <dbl>          <dbl>            <dbl>           <dbl>
1             1.20          0.762              0.1             2.5
> |
```

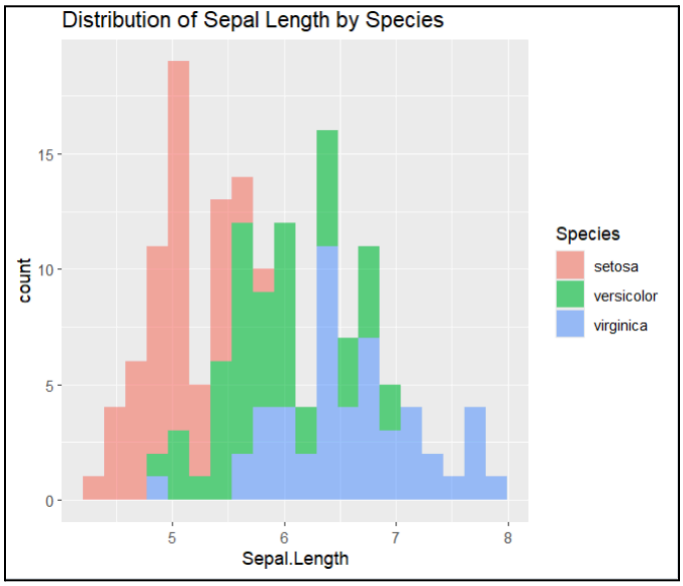
Lets now look at some univariate (single variable) visualizations

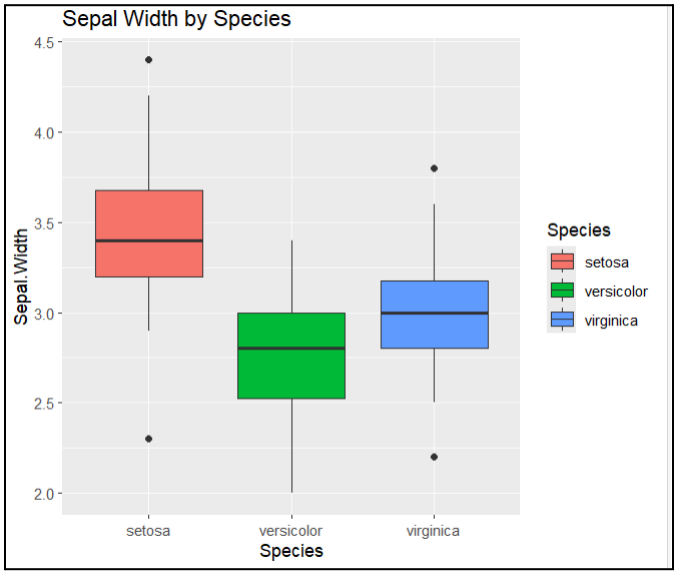
```
#Univariate Visualizations
# Histogram
ggplot(iris_tbl, aes(x = Sepal.Length, fill = Species)) +
  geom_histogram(alpha = 0.6, bins = 20) +
  labs(title = "Distribution of Sepal Length by Species")

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

Question 3. Which features show overlapping distributions between species? Which are more distinct?

The Sepal.Length features show overlapping distributions. Each species slightly overlaps with each other, but there is a distinct difference in which the Setosa has the shortest sepal length and the Virginica has the longest sepal length. The Sepal.Width feature is little more distinct due to the fact there is only really an overlap between the Versicolor and Virginica species. The Setosa width differs more from the other two iris species.





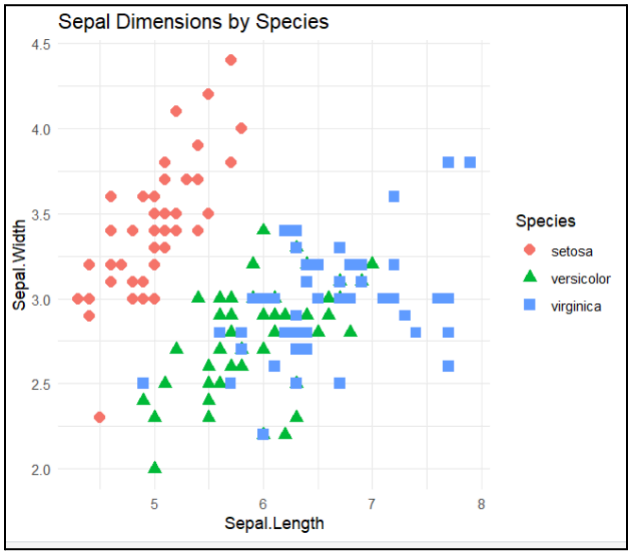
Now lets try some bivariate relationships!

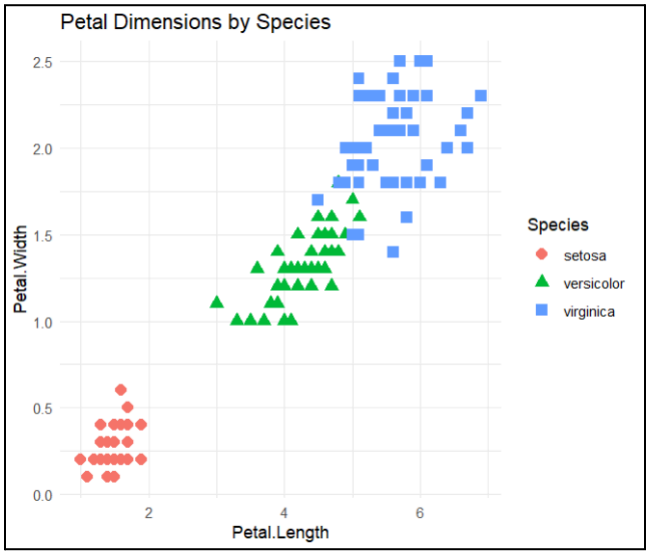
```
#Bivariate Relationships
ggplot(iris_tbl, aes(x = Sepal.Length, y = Sepal.Width,
                     color = Species, shape = Species)) +
  geom_point(size = 3) +
  theme_minimal() +
  labs(title = "Sepal Dimensions by Species")

ggplot(iris_tbl, aes(x = Petal.Length, y = Petal.Width,
                     color = Species, shape = Species)) +
  geom_point(size = 3) +
  theme_minimal() +
  labs(title = "Petal Dimensions by Species")
```

Question 4. Which pair (sepal or petal) shows clearer separation between species?

There is a clearer separation between species with the Petal Length and Width. You can see this because the data points for each species have virtually no overlap on the plot.

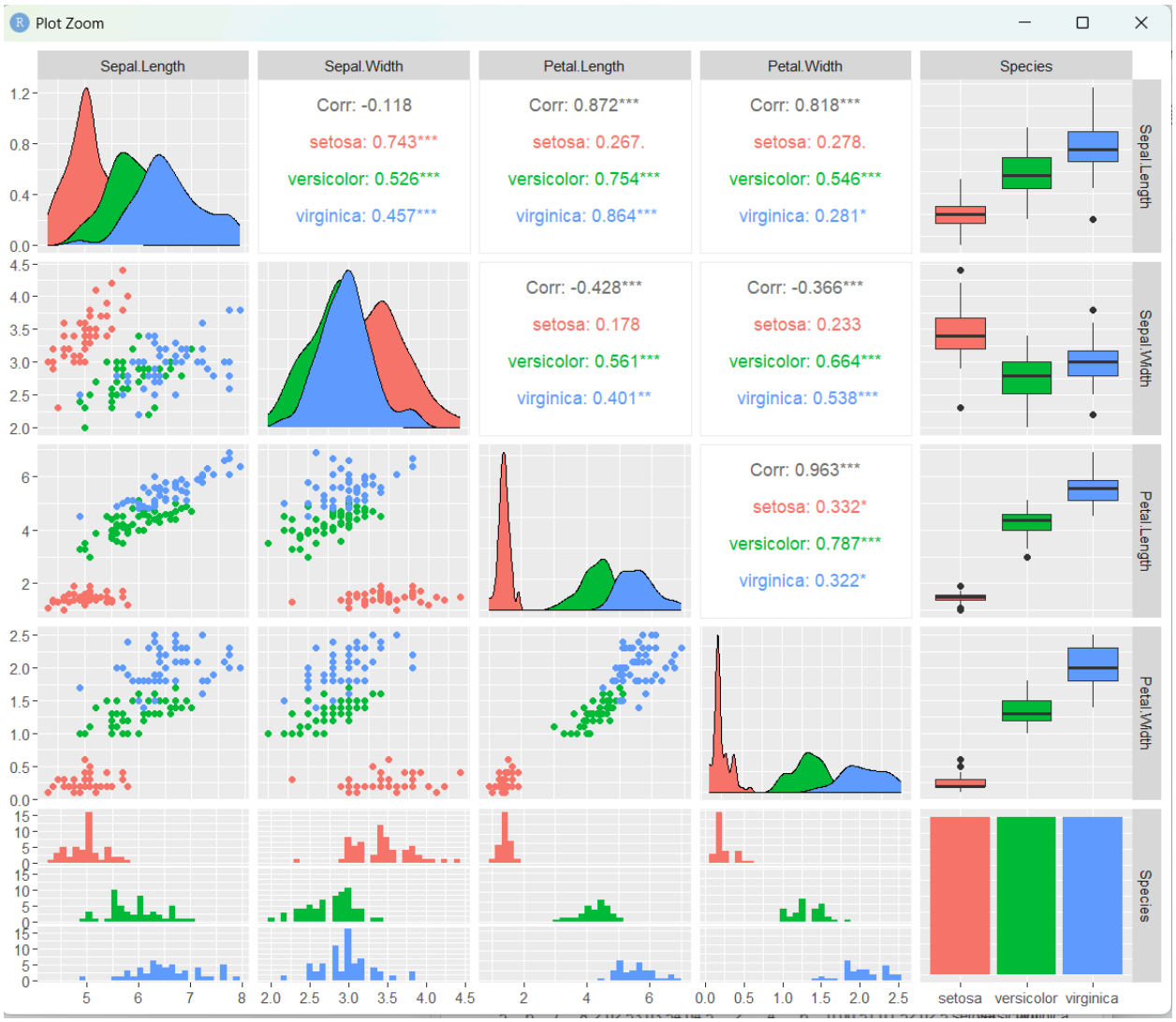




Finally, let's explore some correlations. If you have too many plots, use the left and right arrows in the visualization pane to go back and forth. These plots may also be very small, so use the zoom button as well which will bring up a pop up that you can stretch out!

```
cor(iris_tbl %>% select(where(is.numeric)))
ggpairs(iris_tbl, aes(color = Species))
```

```
> cor(iris_tbl %>% select(where(is.numeric)))
             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
```



3. Exploratory Section (Independent EDA)

Use tidymodel functions, or any other R package to answer these questions using the Iris dataset. Include your code and explanations within a readme file!

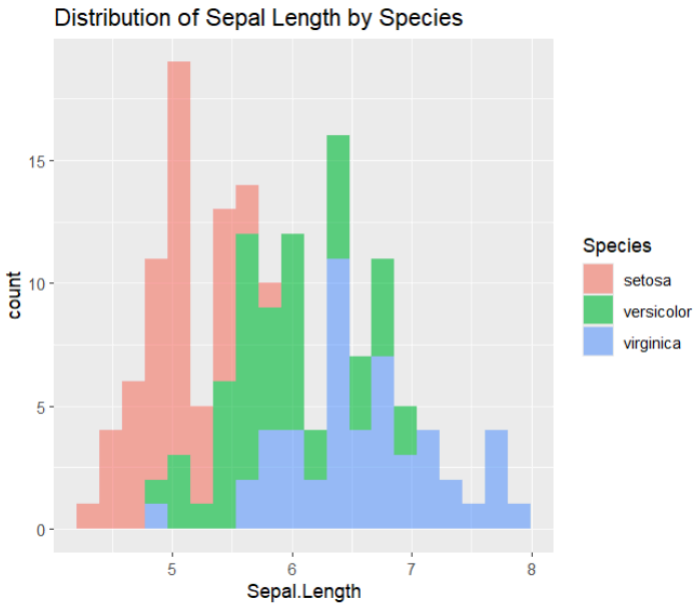
Question 5. Compute mean, median, and standard deviation for each numeric variable by species.

```
sum_by_species <- iris_tbl %>%
 group_by(Species) %>%
 summarise(across(
  where(is.numeric),
  list(
   mean = ~mean(.x, na.rm = TRUE),
   median = ~median(.x, na.rm = TRUE),
   sd = ~sd(.x, na.rm = TRUE)
  ),
  .names = "{.col}_{.fn}"
 ))
print(sum_by_species, n = Inf, width = Inf)
```

```
# A tibble: 3 x 13
  Species    Sepal.Length_mean Sepal.Length_median Sepal.Length_sd
  <fct>                  <dbl>               <dbl>           <dbl>
1 setosa                  5.01                   5           0.352
2 versicolor              5.94                 5.9           0.516
3 virginica               6.59                 6.5           0.636
  Sepal.Width_mean Sepal.Width_median Sepal.Width_sd Petal.Length_mean
             <dbl>              <dbl>          <dbl>             <dbl>
1             3.43                3.4          0.379              1.46
2             2.77                2.8          0.314              4.26
3             2.97                3            0.322              5.55
  Petal.Length_median Petal.Length_sd Petal.Width_mean
                <dbl>           <dbl>            <dbl>
1                 1.5           0.174            0.246
2                 4.35          0.470            1.33
3                 5.55          0.552            2.03
  Petal.Width_median Petal.Width_sd
               <dbl>          <dbl>
1                0.2          0.105
2                1.3          0.198
3                2            0.275
> |
```

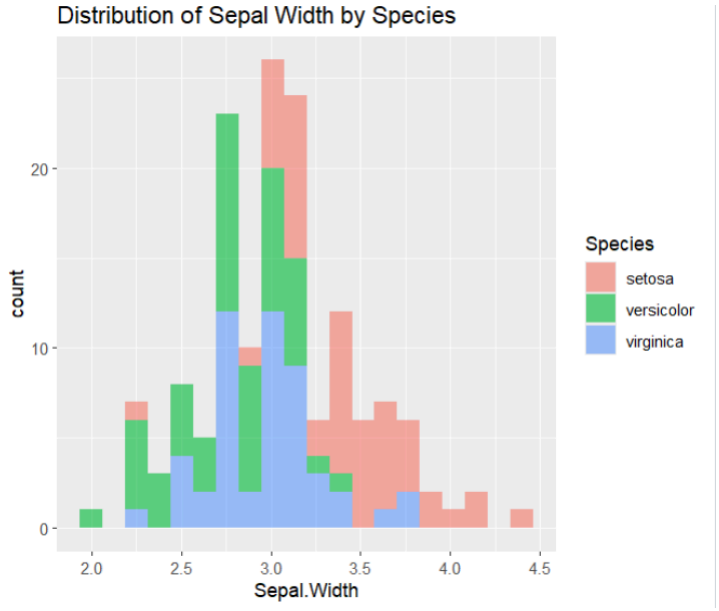
Question 6: Plot a histogram for each numeric variable. Comment on shape (skew, symmetry).

```
ggplot(iris_tbl, aes(x=Sepal.Length, fill = Species)) +
  geom_histogram(alpha = 0.6, bins = 20) +
  labs(title = "Distribution of Sepal Length by Species")
```



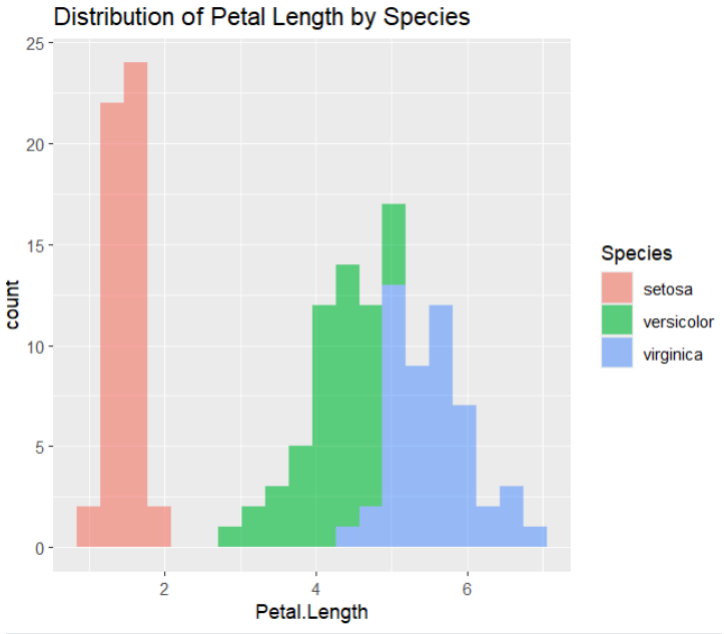
The Sepal.Length histogram has a mixed skewness. Versicolor and Virginica are more symmetrical while Setosa has a skewed right distribution.

```
ggplot(iris_tbl, aes(x = Sepal.Width, fill = Species)) +
  geom_histogram(alpha = 0.6, bins = 20) +
  labs(title = "Distribution of Sepal Width by Species")
```



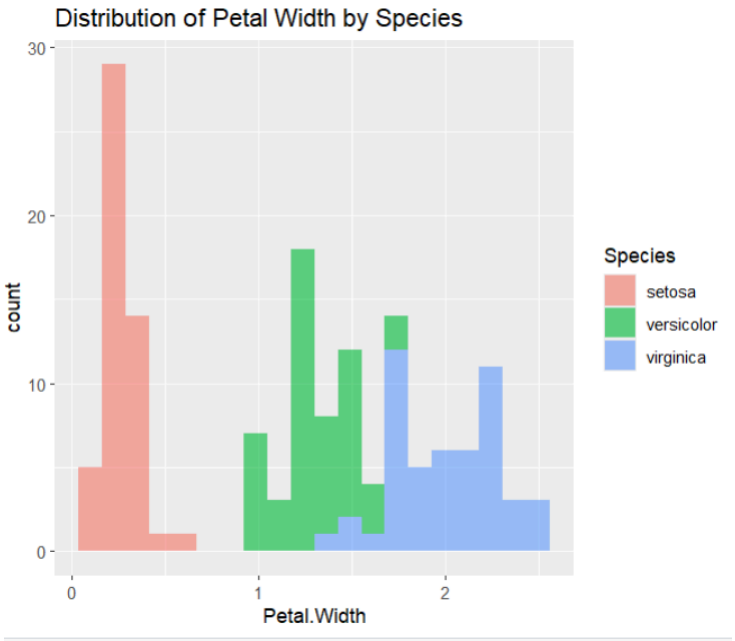
The Sepal.Width histogram has a more symmetrical distribution overall between the three species.

```
ggplot(iris_tbl, aes(x = Petal.Length, fill = Species)) +
  geom_histogram(alpha = 0.6, bins = 20) +
  labs(title = "Distribution of Petal Length by Species")
```



The Petal.Length histogram has an overall skewed left distribution. Both Versicolor and Virginica are skewed left. Setosa, however, is somewhat symmetrical individually, but not in relation to the histogram as a whole.

```
ggplot(iris_tbl, aes(x = Petal.Width, fill = Species)) +
  geom_histogram(alpha = 0.6, bins = 20) +
  labs(title = "Distribution of Petal Width by Species")
```

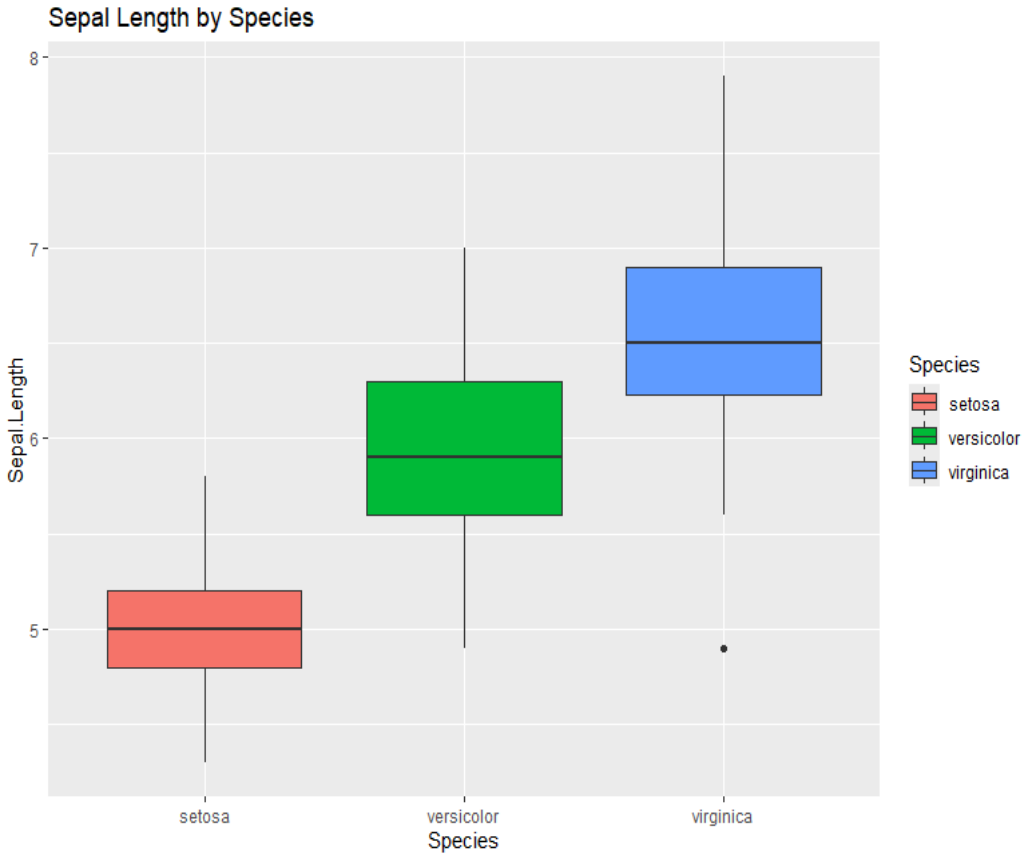


The skewness for the histogram of Petal.Width could be considered skewed right if you look at it as a whole. However when you start to look at each species individually, it gets more symmetrical.

Question 7: Create boxplots for each variable grouped by species. Identify any outliers.

Sepal Length by Species

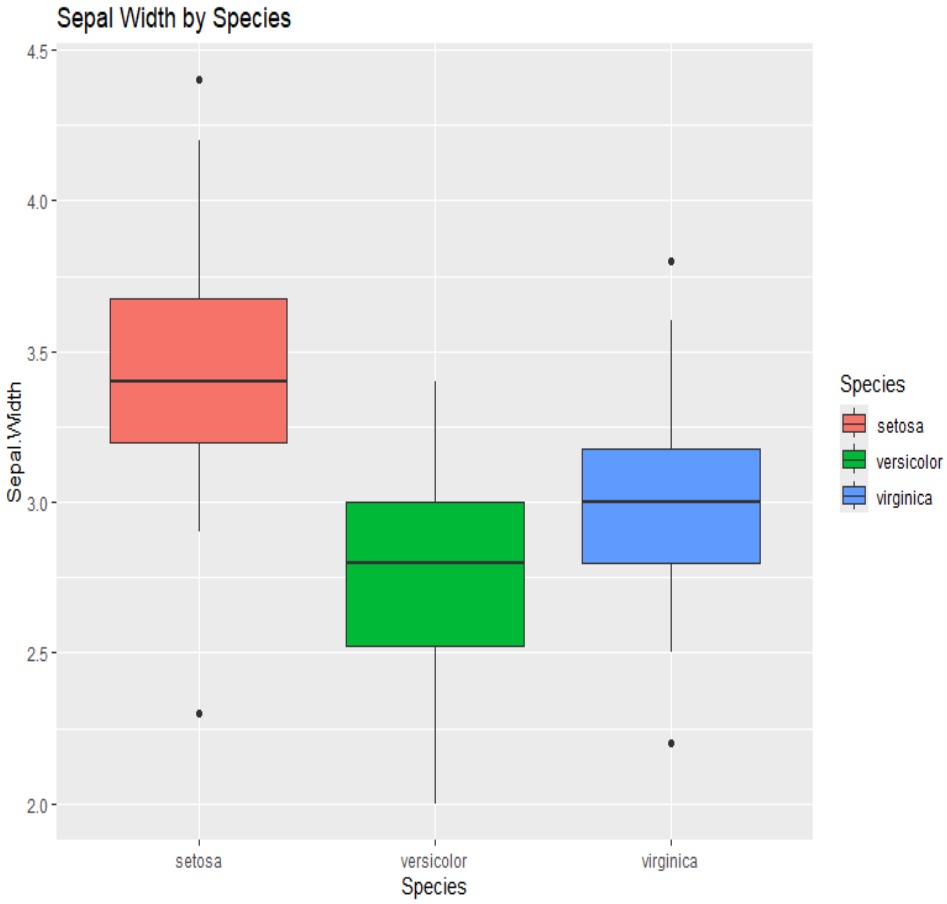
```
ggplot(iris_tbl, aes(x = Species, y = Sepal.Length, fill = Species)) +
geom_boxplot() +
        labs(title = "Sepal Length by Species")
```



There's one low outlier in Virgincia near the 5.0 mark

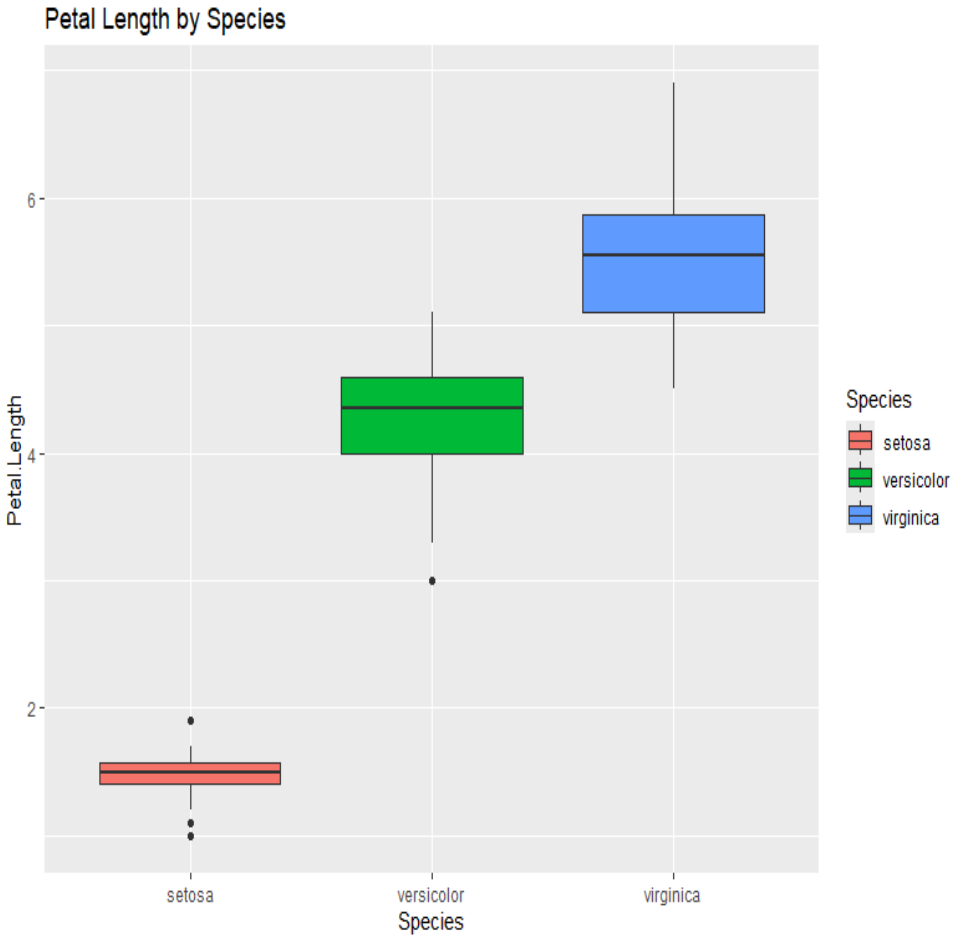
Sepal Width by Species

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



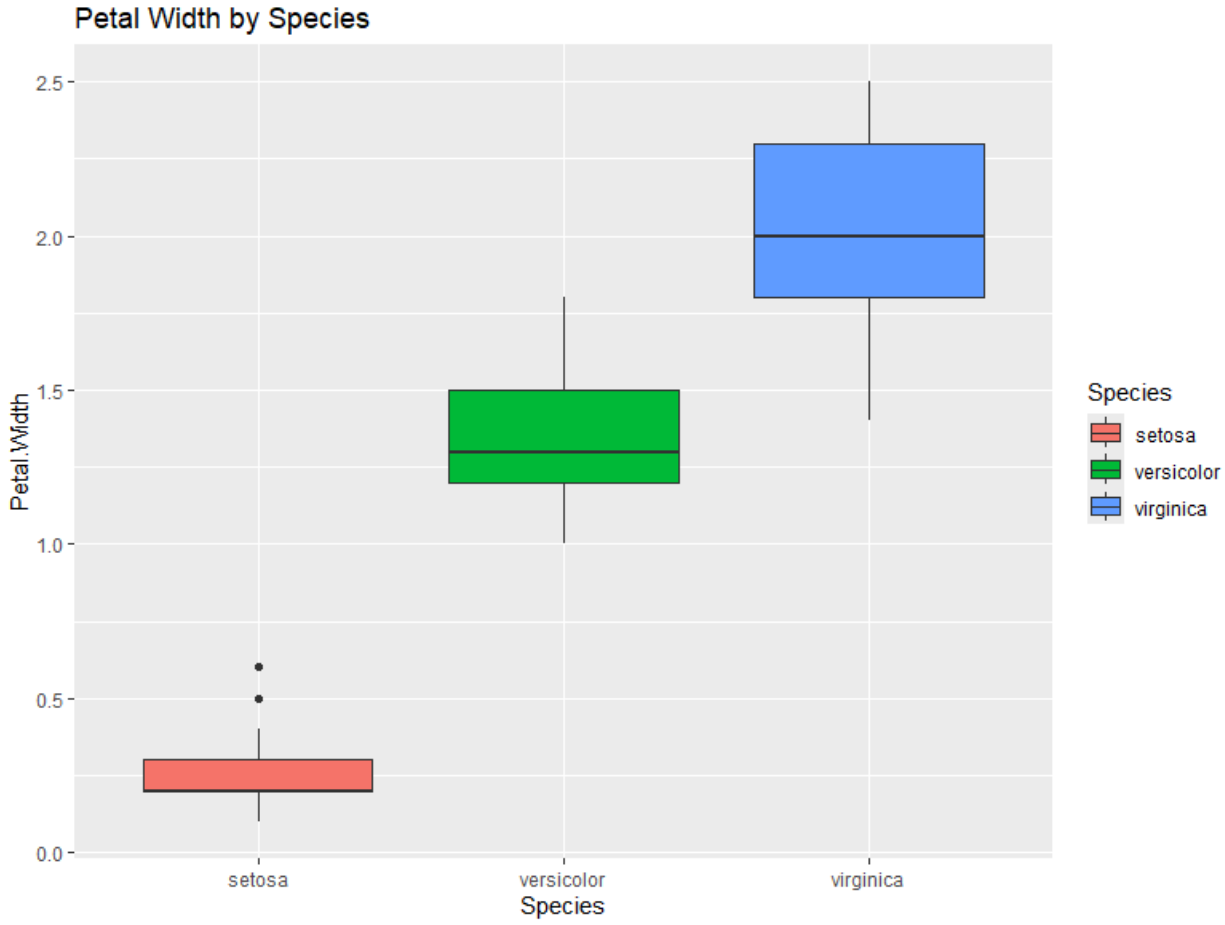
Petal Length by Species

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



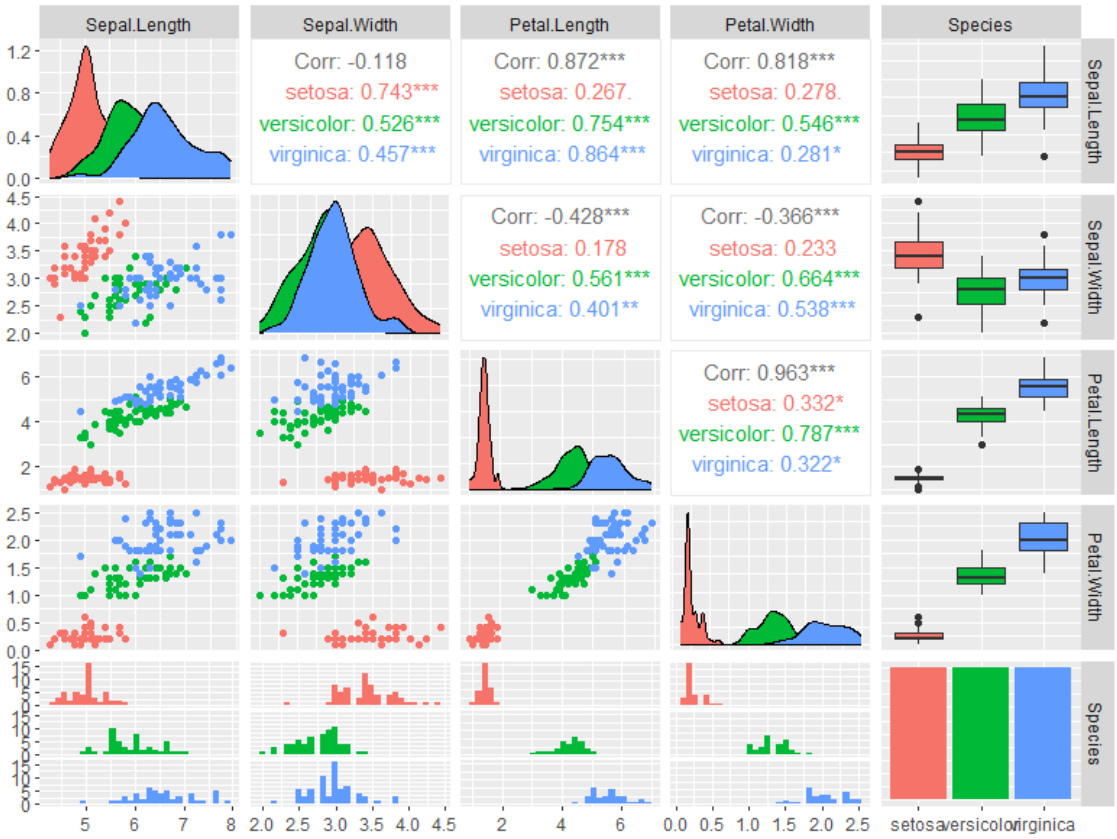
Petal Width by Species

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



Question 8: Create a scatterplot matrix (or ggpairs). Which variables are most related?

ggpairs(iris_tbl, aes(color = Species))



<u>The scatterplot matrix shows that Petal Length and Petal Width are the most related variables.</u>

<u>Correlation value 0.963. This means as petals get longer, they also get wider.</u>

Question 9: Compute and visualize correlations as a heatmap.

 # Compute correlation for numeric variables

cor_matrix <- cor(iris_tbl %>% select(where(is.numeric)))

 # Convert to long (tidy) format for ggplot

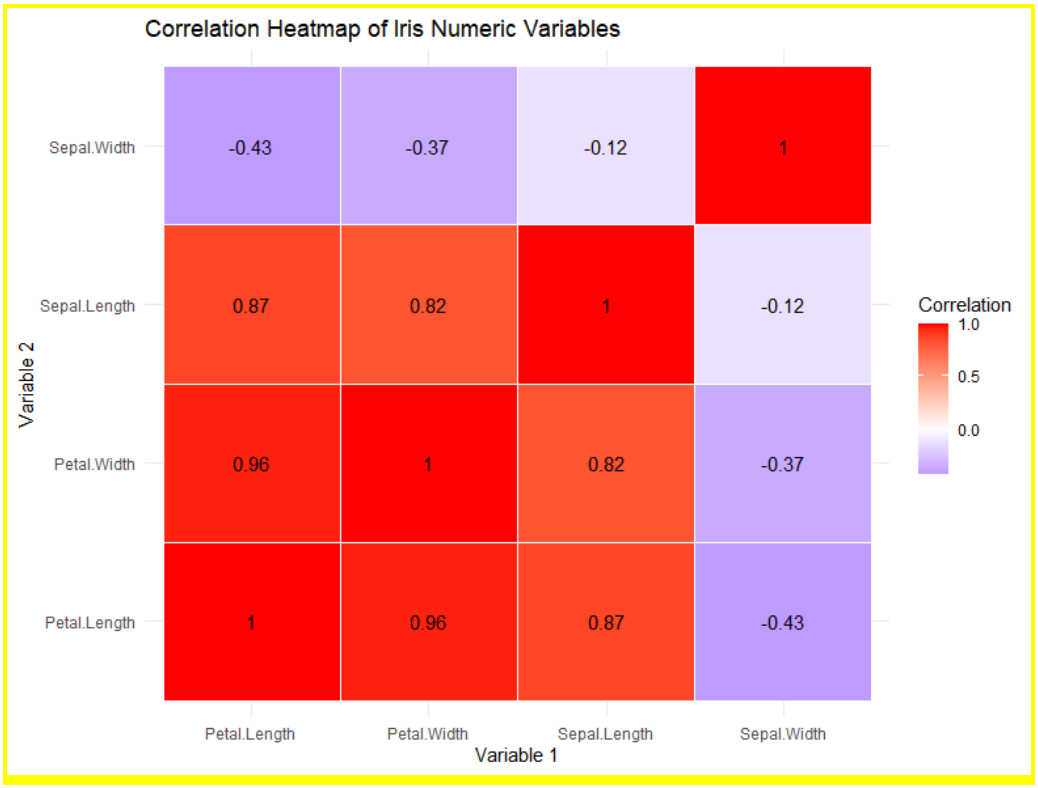
cor_data <- as.data.frame(cor_matrix) %>%

  rownames_to_column("Var1") %>%

  pivot_longer(cols = -Var1, names_to = "Var2", values_to = "value")

 # Plot the heatmap

```
ggplot(cor_data, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(value, 2)), size = 4) +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0) +
  labs(title = "Correlation Heatmap of Iris Numeric Variables",
       x = "Variable 1", y = "Variable 2", fill = "Correlation") +
  theme_minimal()
```



The heatmap uses the numeric variables for each pair of measurements. The red tiles indicate strong positive relationships, while purple tiles indicate weak or negative relationships.

The strongest correlation is between Petal Length and Petal Width (≈ 0.96), confirming that petal size dimensions increase together.

Question 10: Which species shows the greatest variability in petal dimensions?

 # Calculate standard deviation of petal dimensions by species

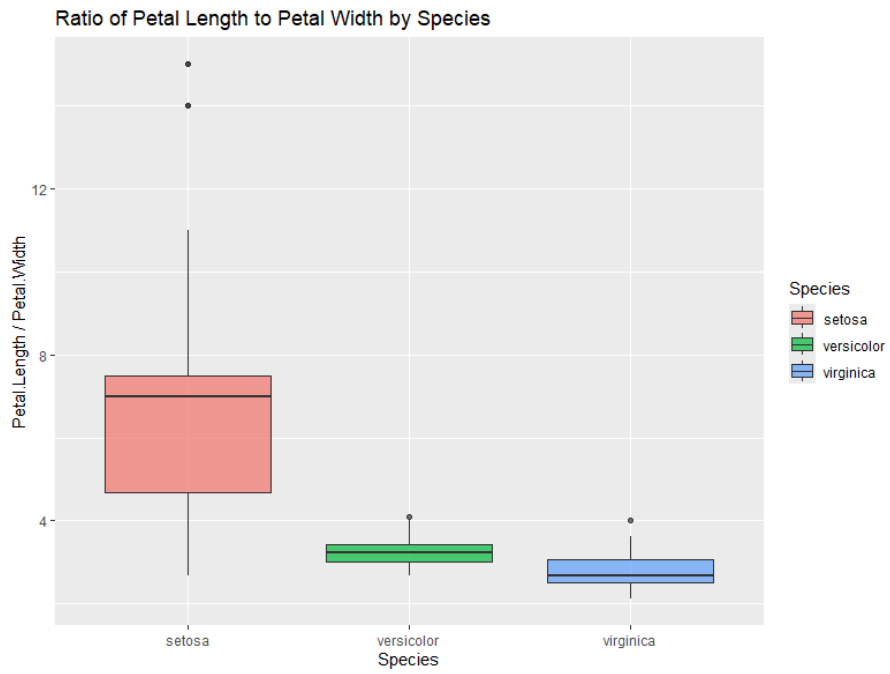
```
iris_tbl %>%
  group_by(Species) %>%
  summarise(
    Petal.Length_sd = sd(Petal.Length),
    Petal.Width_sd = sd(Petal.Width)
  )
```

```
# A tibble: 3 x 3
  Species    Petal.Length_sd Petal.Width_sd
  <fct>                <dbl>          <dbl>
1 setosa               0.174          0.105
2 versicolor           0.470          0.198
3 virginica            0.552          0.275
>
```

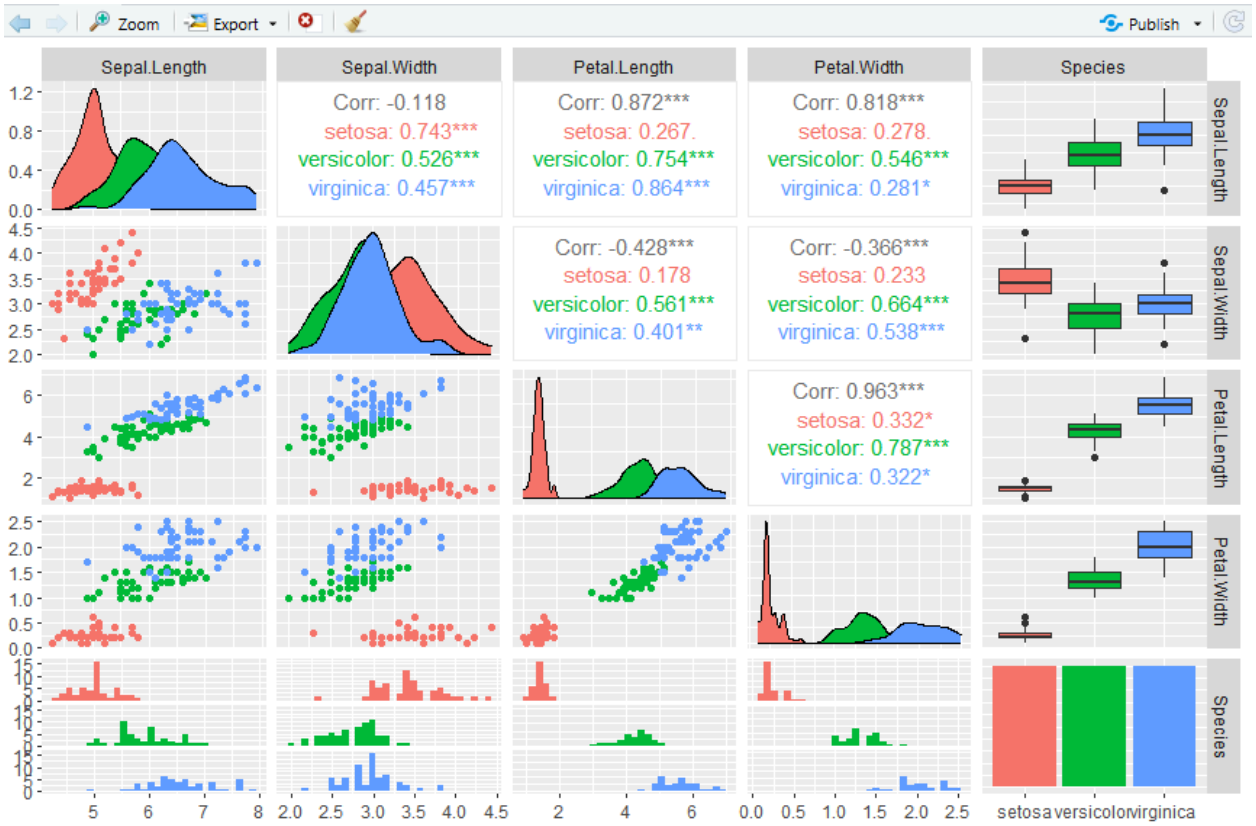
Virginica shows the greatest variability in petal dimensions. Based on its sd in both petal length and width, petal size among Virginica flowers differs a lot more than the other species.

Question 11: Compute and plot the ratio Petal.Length / Petal.Width by species.

```
iris_tbl %>%
  mutate(Ratio = Petal.Length / Petal.Width) %>%
  ggplot(aes(x = Species, y = Ratio, fill = Species)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Ratio of Petal Length to Petal Width by Species", y = "Petal.Length / Petal.Width")
```



Question 12: Identify any relationships between sepal and petal measurements. (Cierra)




```
# Load libraries

library(tidyverse)

library(GGally)

# Load dataset

data("iris")

# View relationships between Sepal and Petal features

ggpairs(iris, aes(color = Species))
```

The scatterplot matrix visualizes relationships among sepal and petal dimensions for each iris species.

- There is a strong positive correlation ($r \approx 0.87$) between Sepal.Length and Petal.Length, indicating that flowers with longer sepals also tend to have longer petals.

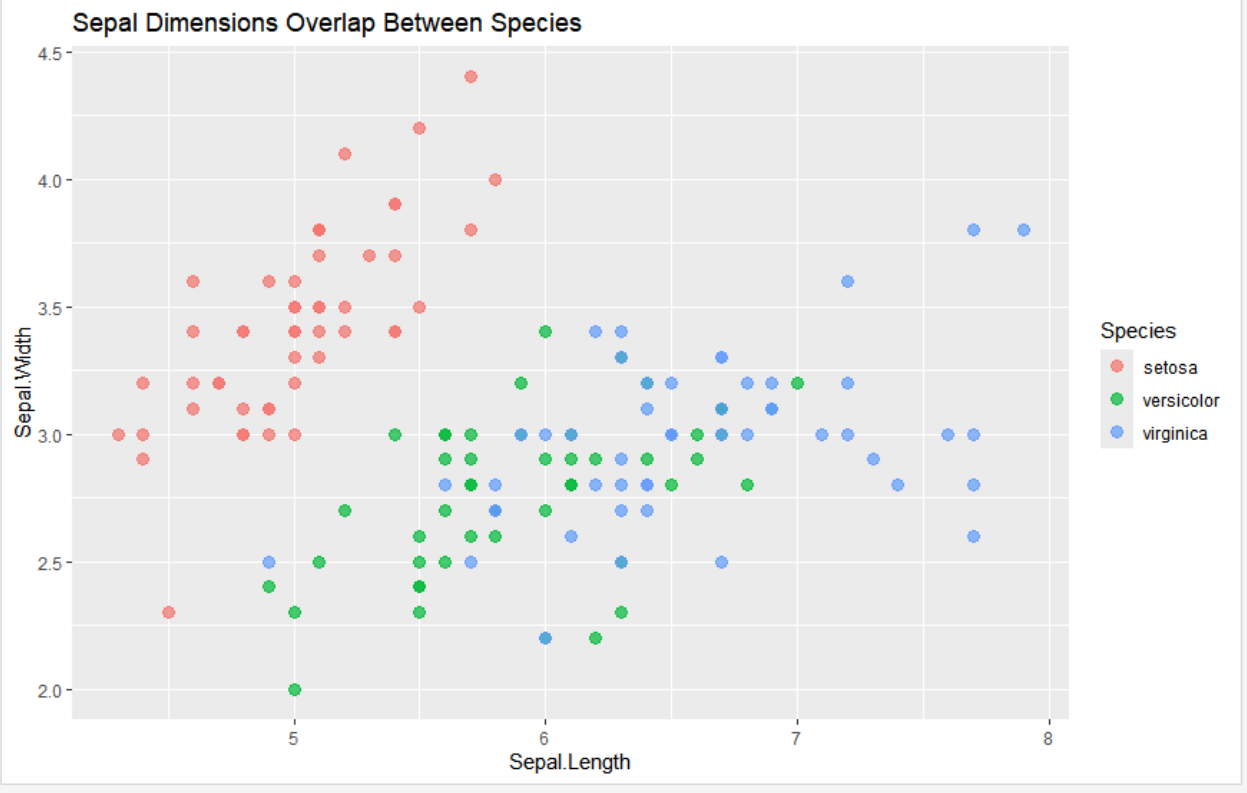
- Similarly, Petal.Length and Petal.Width have the strongest correlation ($r \approx 0.96$), showing that as petals become longer, they also become wider.

- Sepal.Width has weaker correlations with the other measurements and shows more variation within species.

- These relationships are strongest in *versicolor* and *virginica*, which exhibit more variability, while *setosa* remains smaller and more consistent in size.

This suggests that overall flower size increases proportionally across both petal and sepal features, especially among *versicolor* and *virginica* species.

Question 13: Discuss which features seem most useful for classifying species (qualitatively). (Cierra)




```
# Sepal comparison plot
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point(size = 3, alpha = 0.7) +
  labs(title = "Sepal Dimensions Overlap Between Species",
       x = "Sepal Length (cm)",
```

```
    y = "Sepal Width (cm)") +
  theme_minimal()
```

The scatterplot above illustrates how Sepal Length and Sepal Width vary across *Iris setosa*, *Iris versicolor*, and *Iris virginica*.
While there is some visible grouping:

- *Iris setosa* (red) generally clusters toward shorter and wider sepals,
- *Iris versicolor* (green) and *Iris virginica* (blue) show significant overlap in the middle and upper ranges.

Because of this overlap, sepal dimensions alone are not reliable for distinguishing between *versicolor* and *virginica*.

In contrast, when comparing petal measurements, the species are well separated, with *setosa* having small petals, *versicolor* mid-sized, and *virginica* the largest.

Question 14: Summarize your main insights in at least 1–2 paragraphs.

In the Iris dataset, we saw that there is a very strong correlation between floral characteristics and species. From observation, Petal length and width are by far the key difference in characteristics. We can use scatterplots to show that this is a good separation between Setosa, Versicolor, and Virginica, confirmed in a heatmap showing that these two features are highly correlated ($r \approx 0.96$) as the petals get wider as they get longer.

Sepals are a less discriminative feature compared to petals. We saw a moderate correlation between Sepal.Length and Petal.Size, and a wider range of Sepal Width for all species. So, from our dataset, we can say that petal features are a very strong feature and will definitely greatly impact future model and its performance

Question 15: What did you learn most with this lab (1-2 paragraphs maximum)