STA 326 2.0 Programming and Data Analysis with R [*]

Exploring `iris` dataset with `qplot`

31 March 2020

Contents

Stage 1: Planning your analysis

Step 1: Dataset overview and description

Before we get started let's look at the data and plan a analysis.

Load iris dataset

```
data(iris)
```

Here is a glimpse of the dataset.

```
head(iris)
```

```
  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
```

We have four quantitative variables: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width and one qualitative variable: Species

Step 2: One-way analysis

Let's look at the graphs we could use to explore variables one by one.

Plots that could be used to to summarize qualitative variables

- Bar chart
- Pie chart

Plots that could be used to to summarize quantitative variables

- Box and whisker plot
- Histograms
- Dot plots
- Density plot
- Stem and leaf displays

Note: Stem and leaf displays are best-suited for small to moderate datasets, whereas others such as histograms and Box and whisker plots are best-suited for large datasets. Box and whisker plots and histograms are also good at depicting differences between distributions and identifying outliers.

Step 3: Two-way analysis

Next, we will look at two variables at a time.

- Quantitative vs Quantitative: Scatter plots
- Quantitative vs Qualitative: Box plots/ Histograms/ Dot plots/ Density plots with groups allow us to compare across different levels of the qualitative variable. **Faceting** can be used to generate the same plot for different levels of the qualitative variable.

Step 4: Three-way analysis

Now, let's look at three variables at a time.

- Two quantitative variables and one qualitative variable: Scatter plot with different markers (eg: size, shapes, colours) for different levels of the qualitative variable.

Stage 2: Getting started with `qplot()` in the ggplot2 package.

Now we are going to use the `qplot` function to make some quick plots. This section demonstrates how different graphs can be plotted for various purposes using the `qplot`.

Recap: some important arguments in `qplot`

```
qplot(
  x, # X variable
  y, # Y variable
  data, # name of the dataframe
  facets = NULL,
  margins = FALSE,
  geom = "auto",
  xlim = c(NA, NA), # numeric vector of length 2 giving the x coordinates
  ylim = c(NA, NA), # numeric vector of length 2 giving the 7 coordinates
  log = "",
  main = NULL, # Figure title
  xlab = NULL, # X-axis title
  ylab = NULL, # Y-axis title
  asp = NA,
  stat = NULL,
  position = NULL,
)
```

One-way analysis

Load packages

```
library(ggplot2)
```

1. Summarizing qualitative variables

```
qplot(x = Species, data = iris, geom = "bar", ylab = "Count",
      main = "Composition of Species")
```

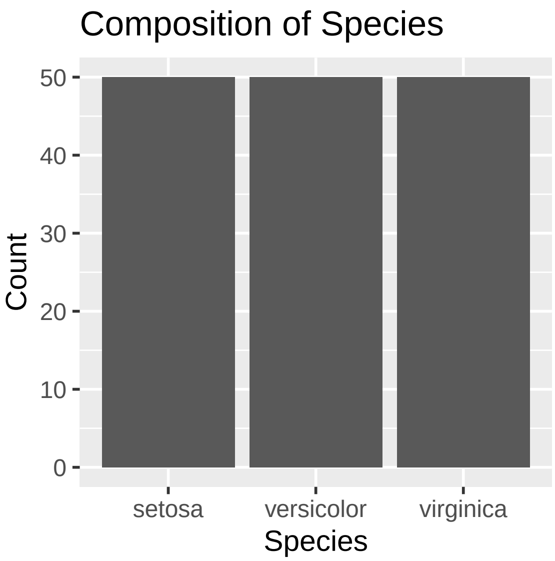


Figure 1: Composition of the sample

2. Summarizing quantitative variables

Here, I have drawn plots only for `Sepal.Width`. Please take suitable graphs for other quantitative variables in the iris dataset.

```
qplot(x = Sepal.Width, data = iris, geom = "histogram")
```

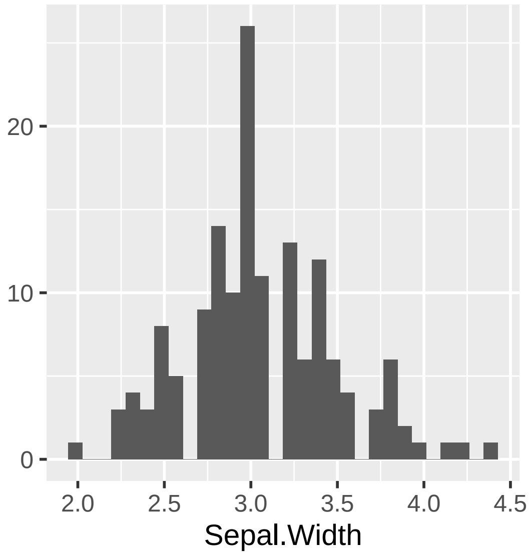


Figure 2: Histogram of sepal width

Histogram

```
qplot(x = Sepal.Width, data = iris, geom = "density")
```

Density plot

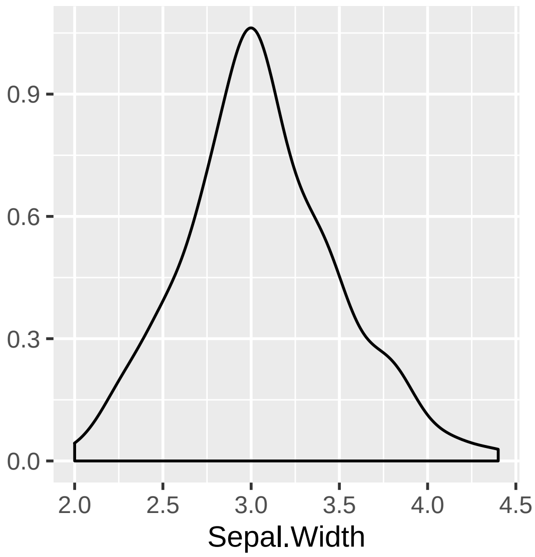


Figure 3: Density plot of sepal width

```
qplot(y = Sepal.Width, data = iris, geom = "boxplot")
```

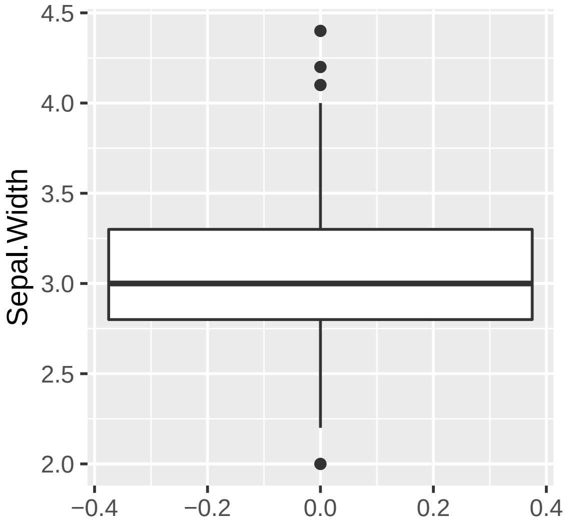


Figure 4: Boxplot of sepal width

Box and whisker plot

Two-way analysis

1. Visualizing two qualitative variables at a time

```r
qplot(x = Sepal.Length, y = Sepal.Width, data = iris, geom = "point")
```

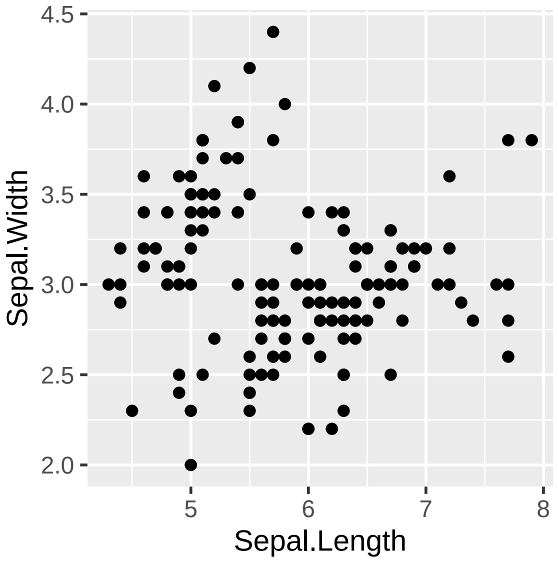


Figure 5: Scatterplot of sepal length and sepal width

2. Visualizing qualitative and quantitative variables

```r
qplot(x = Species, y = Sepal.Width, data = iris, geom = "boxplot")
```

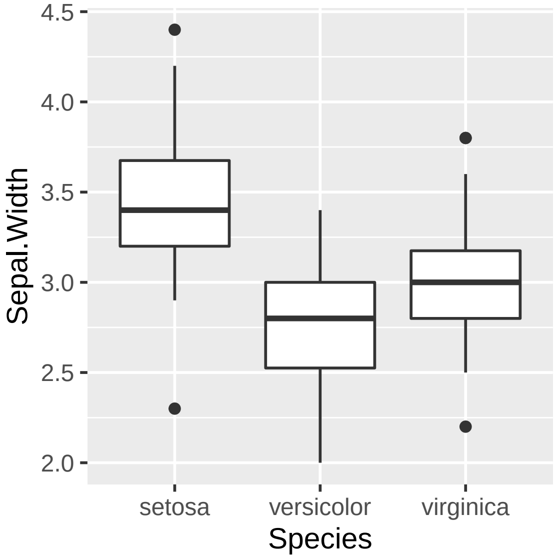


Figure 6: Boxplot of sepal width by species

```r
qplot(x = Species, y = Sepal.Width, data = iris, geom = "boxplot", fill = Species)
```

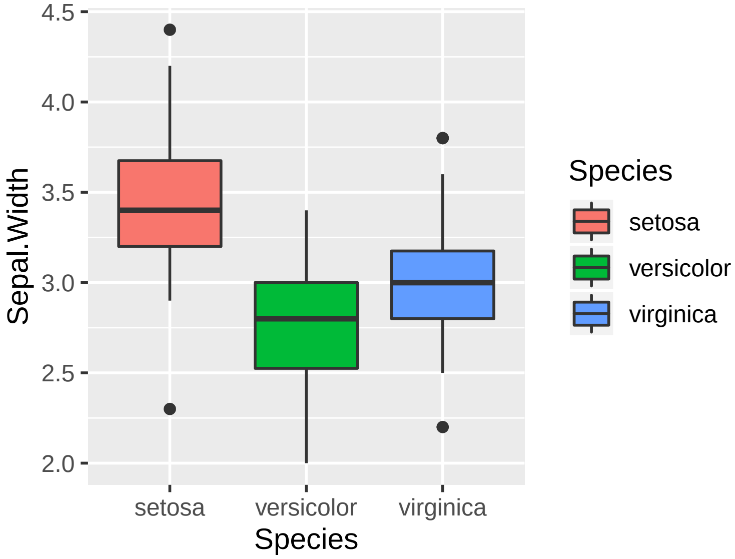


Figure 7: Boxplot of sepal width by species

Different ways to modify your graph

```
qplot(x = Species, y = Sepal.Width, data = iris, geom = c("point","jitter","boxplot"),
      alpha = 0.5, colour = Species)
```

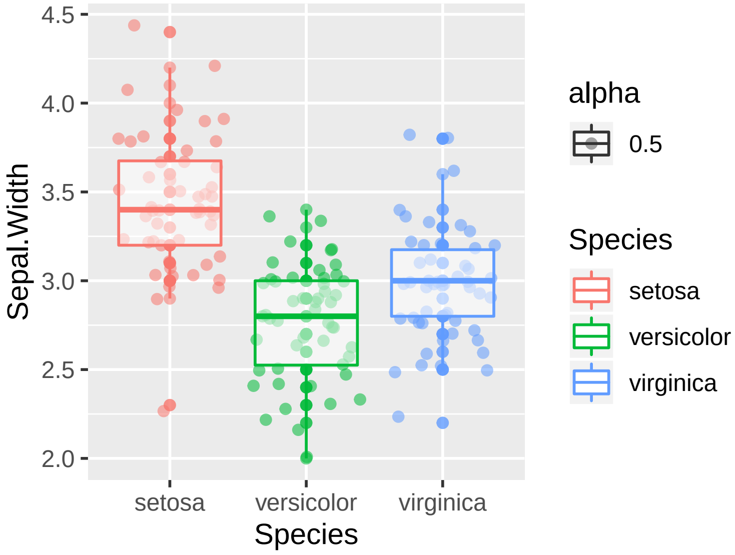


Figure 8: Boxplot of sepal width by species

```
qplot(x = Sepal.Width, data = iris, geom = c("histogram"),
     colour = Species)
```

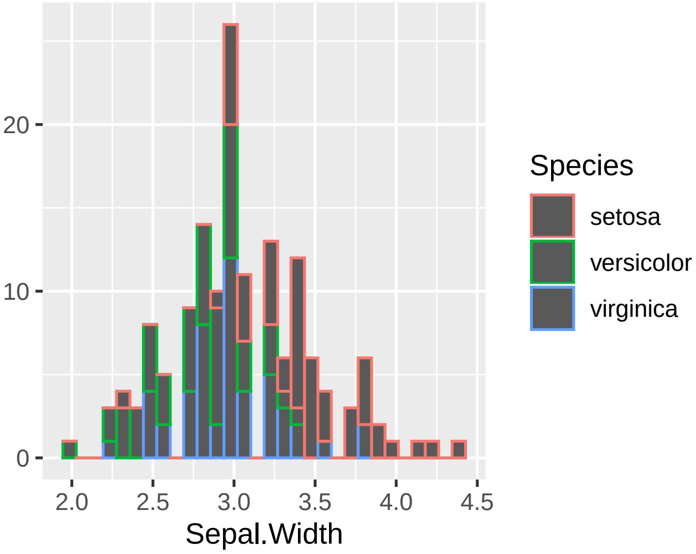


Figure 9: Histogram of sepal width

```
qplot(x = Sepal.Width, data = iris, geom = c("histogram"),
     fill = Species)
```

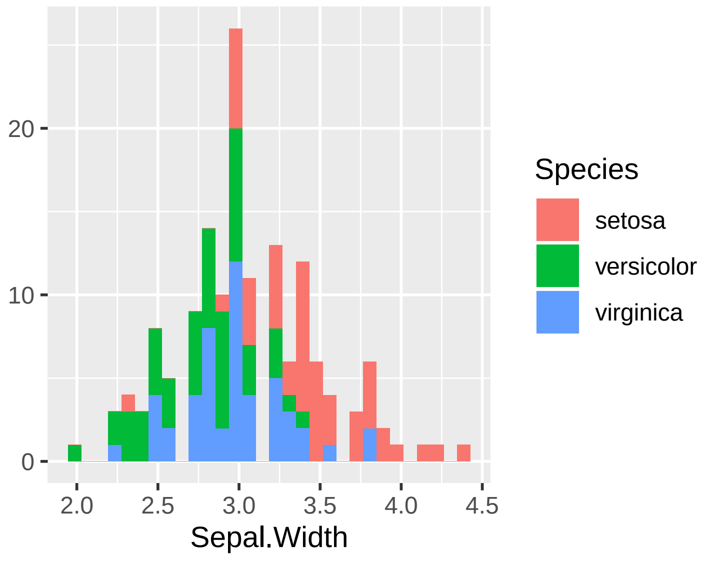


Figure 10: Histogram of sepal width

```
qplot(x = Sepal.Width, data = iris, geom = c("density"),
     colour = Species)
```

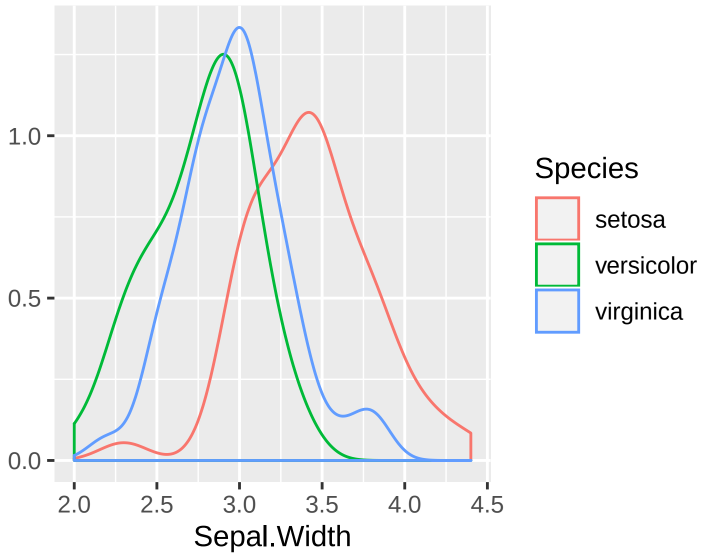


Figure 11: Density plot of sepal width

Three-way analysis

Everything on a single panel

```
qplot(x = Sepal.Length, y = Sepal.Width, data = iris,
      geom = "point", color = Species)
```

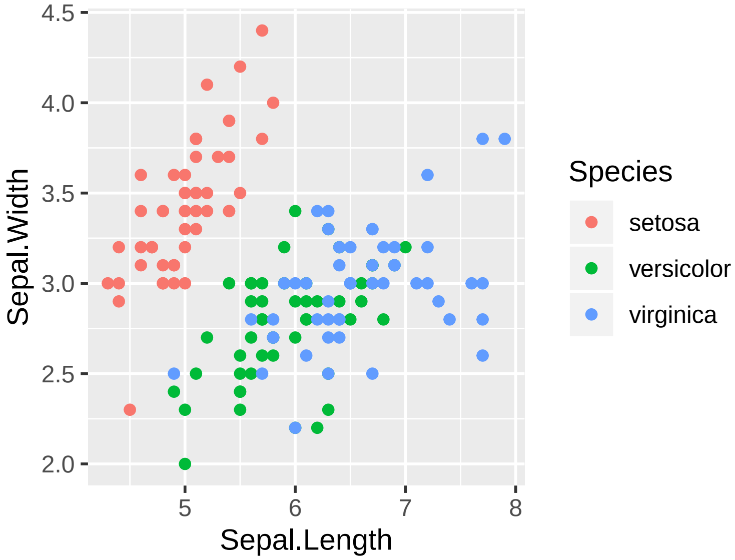


Figure 12: Scatterplot of sepal length and sepal width by species

Separate panels for each species: column-wise

```
qplot(x = Sepal.Length, y = Sepal.Width,
      facets = .~Species, data = iris, geom = "point")
```

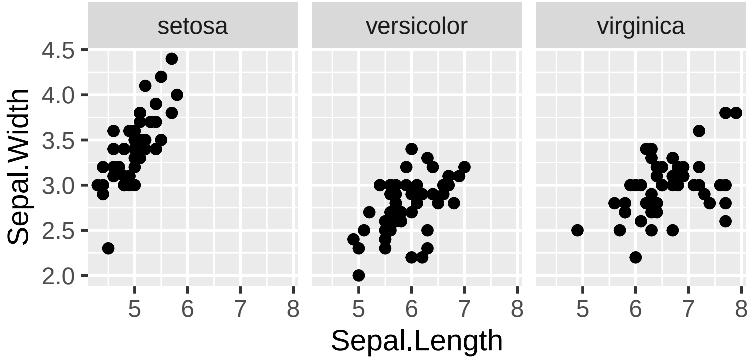


Figure 13: Scatterplot of sepal length and sepal width by species

Separate panels for each species: row-wise

```
qplot(x = Sepal.Length, y = Sepal.Width,
      facets = Species~., data = iris, geom = "point")
```

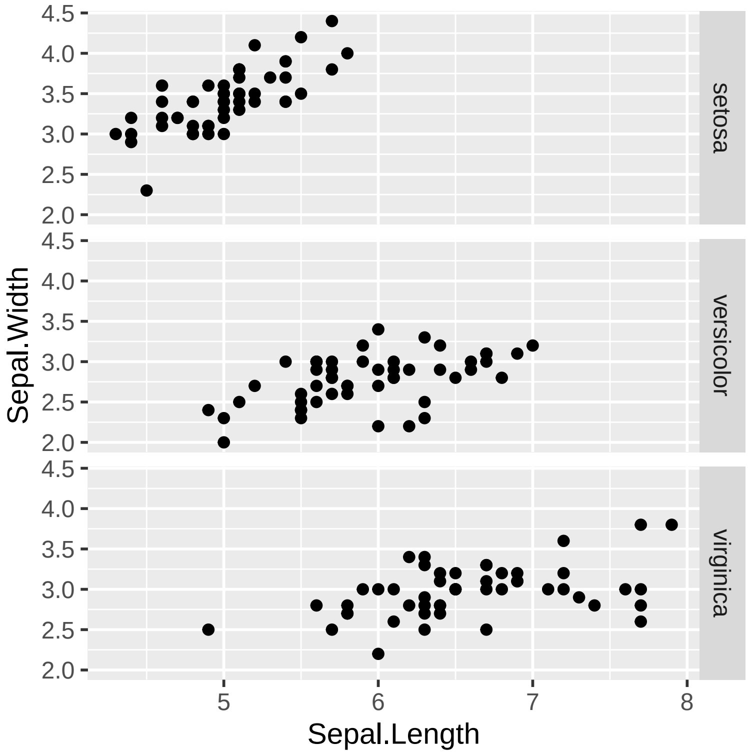


Figure 14: Scatterplot of sepal length and sepal width by species

`patchwork` package in R

```
library(patchwork)
```

First you need to assign a name for each graph. Here, I use `q1` and `q2`.

```
q1 <- qplot(x = Species, y = Sepal.Width, data = iris, geom = c("jitter","boxplot"),
      alpha = 0.5, colour = Species, main = "Distribution of Sepal.Width") + geom_boxplot(outlier.siz
q2 <- qplot(x = Species, y = Sepal.Length, data = iris, geom = c("jitter","boxplot"),
      alpha = 0.5, colour = Species, main = "Distribution of Sepal.Length") + geom_boxplot(outlier.si
```

Arrange multiple graphs row-wise use "/"

```
q1/q2
```

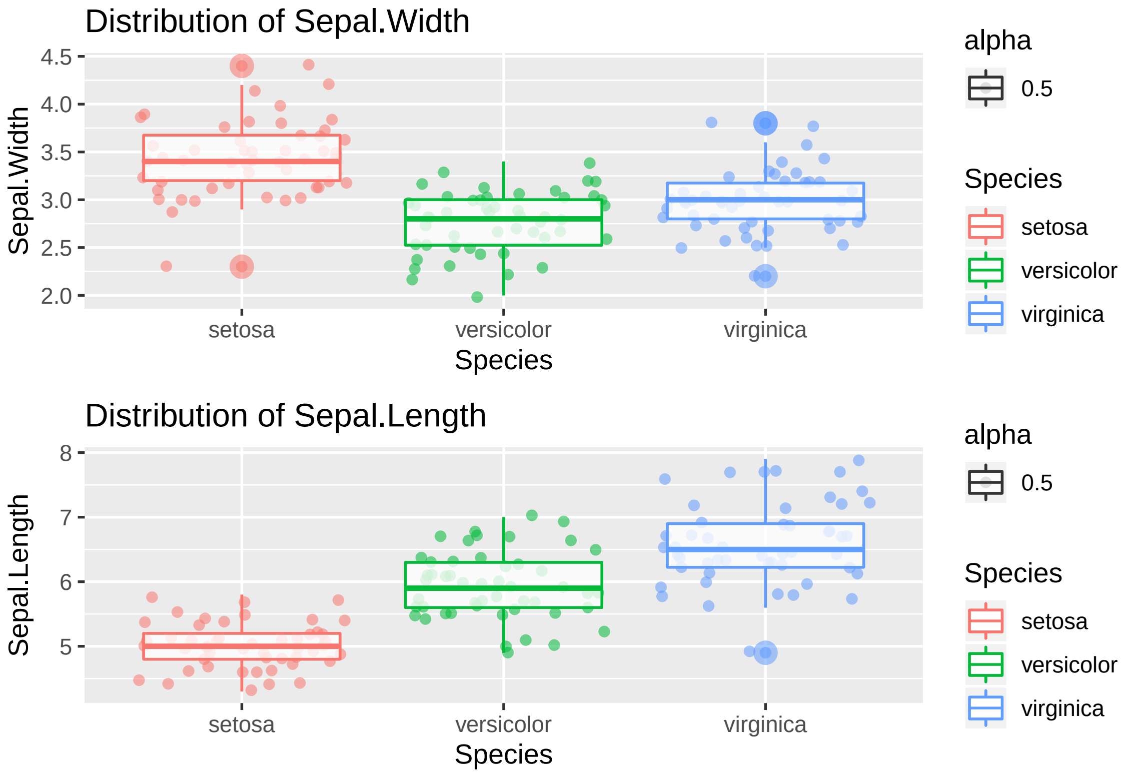


Figure 15: Arrange multiple graphs row-wise

Arrange multiple graphs column-wise: use “|”

```
q1|q2
```

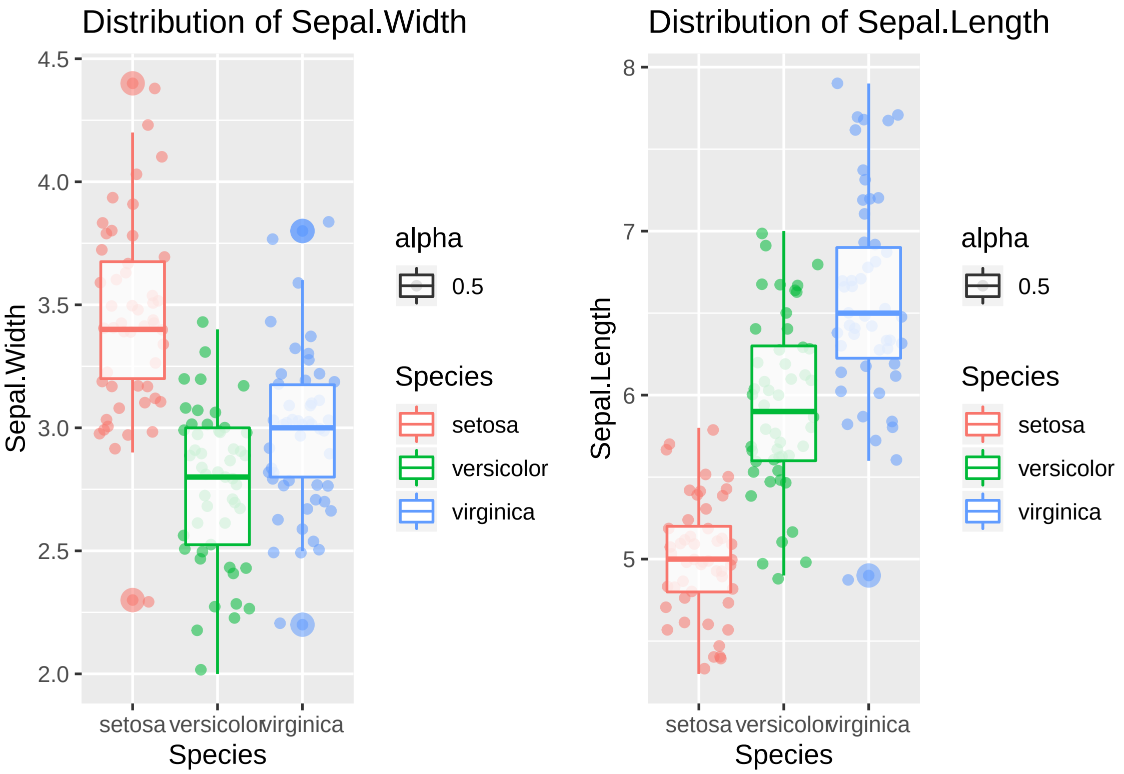


Figure 16: Arrange multiple graphs column-wise

Arrange multiple graphs both row-wise and column-wise

```
(q1|q2)/(q1|q2)
```

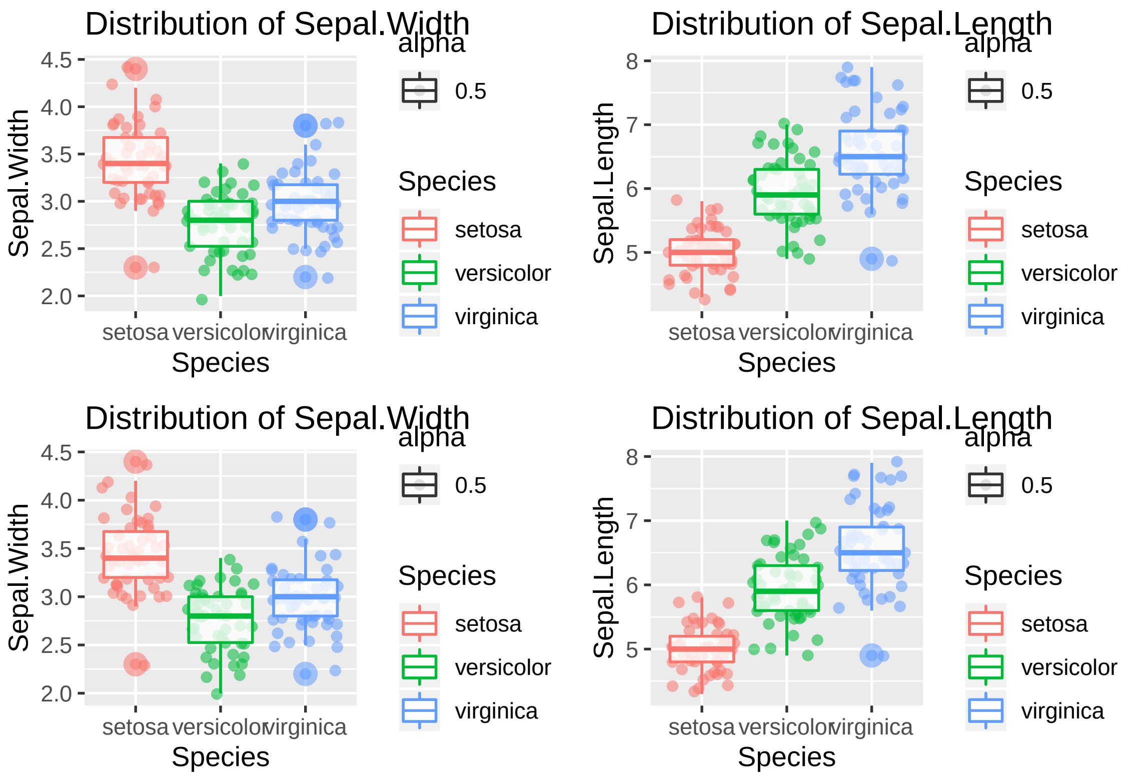


Figure 17: Arrange multiple graphs both row-wise and column-wise

Stage 3: Final analysis

You do not need to include all the graphs to your final analysis. Please select only the useful graphs which help you to tell the story in your dataset. Here is mine.

1. Composition of the sample

```
qplot(x = Species, data = iris, geom = "bar", ylab = "Count",
      colour = Species, fill = Species,
      main = "Composition of Species")
```

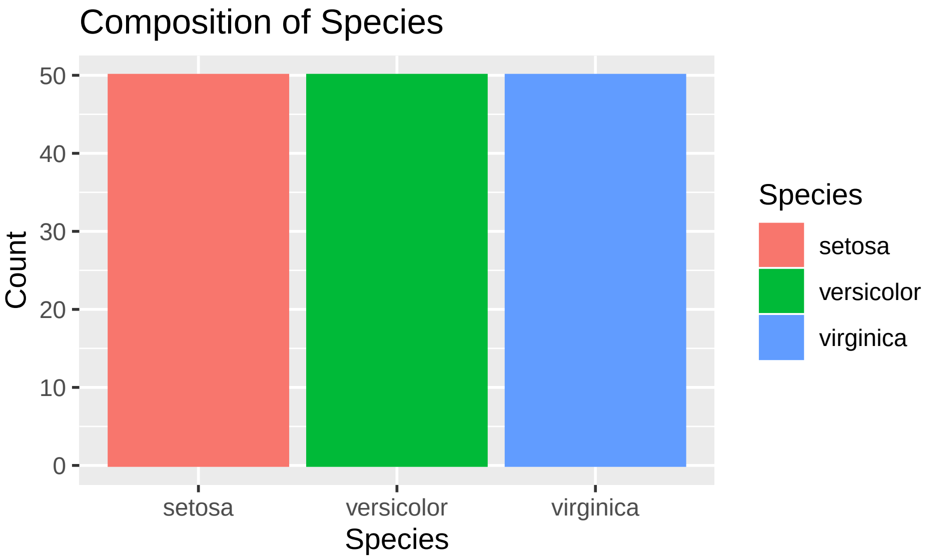


Figure 18: Composition of the sample

2. Distribution of the features of sepal and petal by species

```
q1 <- qplot(x = Species, y = Sepal.Width, data = iris, geom = c("jitter","boxplot"),
      alpha = 0.5, colour = Species, main = "(a) Distribution of Sepal Width") + geom_boxplot(outlier
q2 <- qplot(x = Species, y = Sepal.Length, data = iris, geom = c("jitter","boxplot"),
      alpha = 0.5, colour = Species, main = "(b) Distribution of Sepal Length") + geom_boxplot(outlie
q3 <- qplot(x = Species, y = Petal.Width, data = iris, geom = c("jitter","boxplot"),
      alpha = 0.5, colour = Species, main = "(c) Distribution of Petal Width") + geom_boxplot(outlier
q4 <- qplot(x = Species, y = Petal.Length, data = iris, geom = c("jitter","boxplot"),
      alpha = 0.5, colour = Species, main = "(d) Distribution of Petal Length") + geom_boxplot(outlie
(q1|q2)/(q3|q4)
```

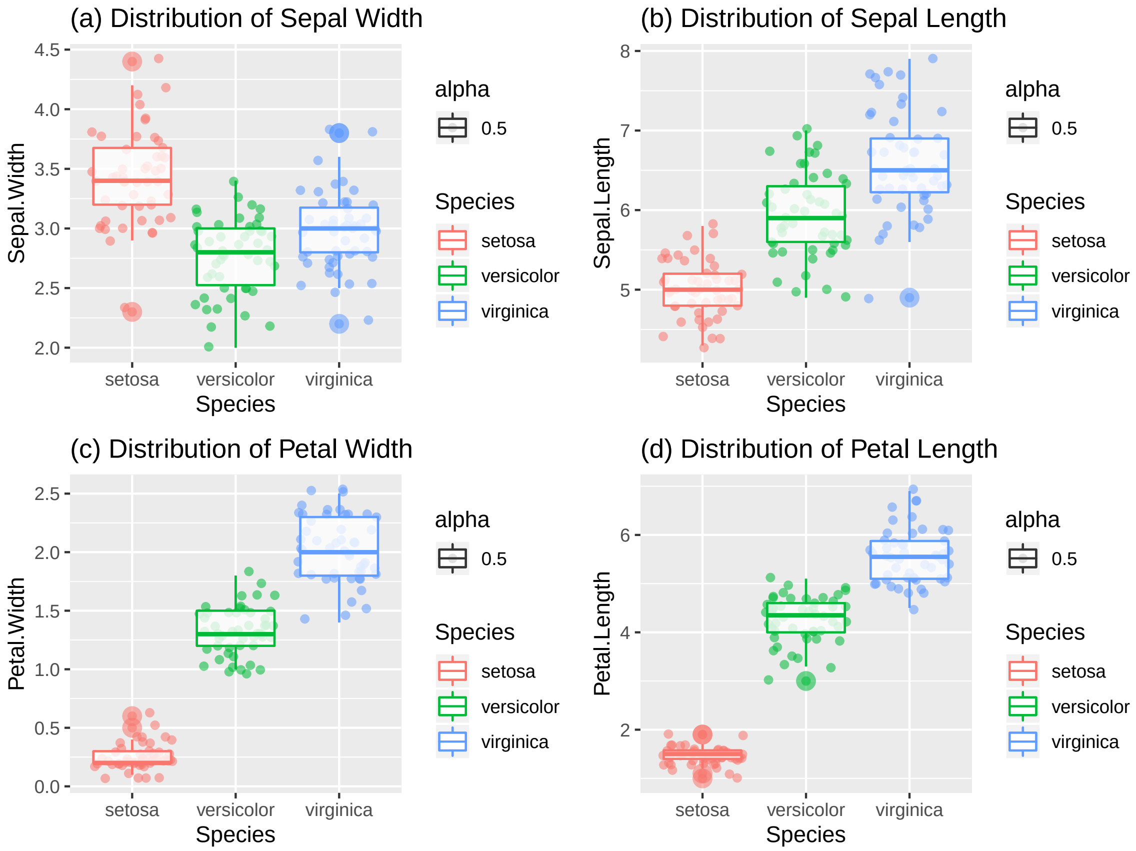
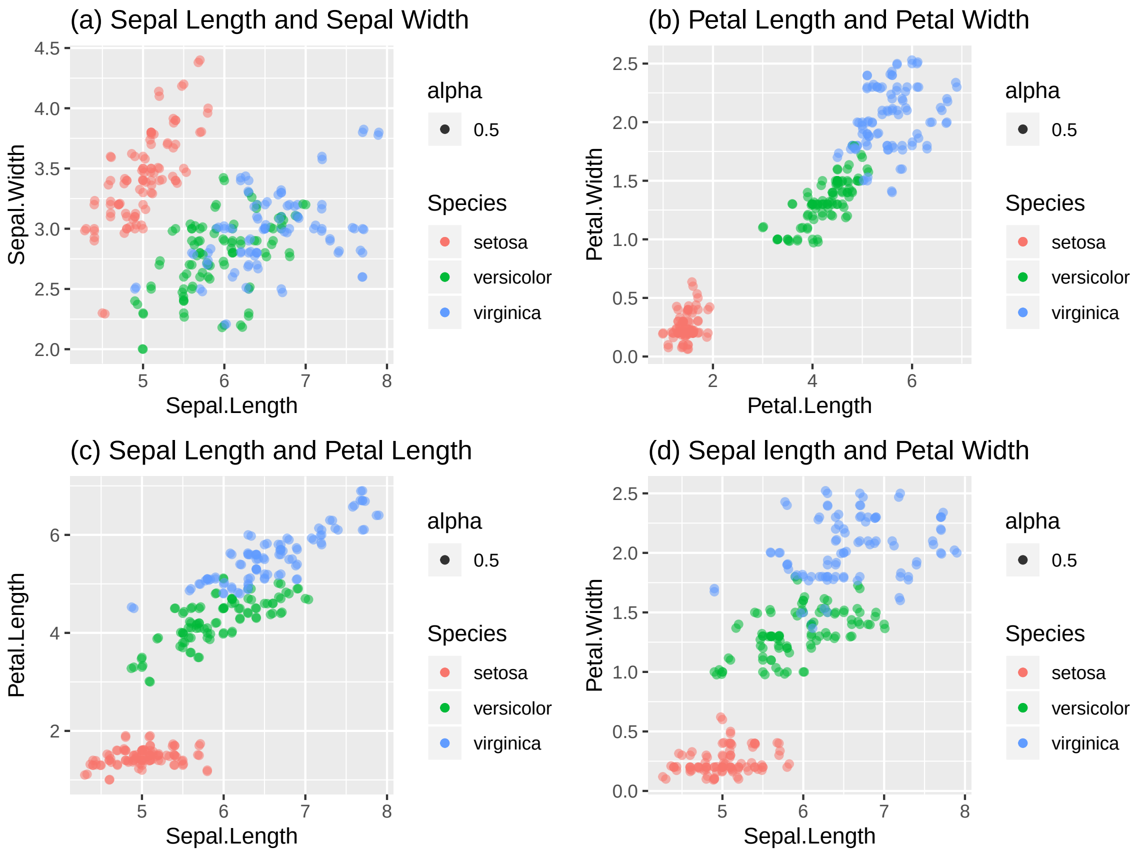


Figure 19: Distribution of features related to sepal and petal by species

3. Relationship between features of sepal and petal by species

```r
p1 <- qplot(x = Sepal.Length, y = Sepal.Width, data = iris, geom = c("point","jitter"),
      alpha = 0.5, colour = Species,
      main="(a) Sepal Length and Sepal Width")
p2 <- qplot(x = Petal.Length, y = Petal.Width, data = iris, geom = c("point","jitter"),
      alpha = 0.5, colour = Species,
      main = "(b) Petal Length and Petal Width")
p3 <- qplot(x = Sepal.Length, y = Petal.Length, data = iris, geom = c("point","jitter"),
      alpha = 0.5, colour = Species,
      main = "(c) Sepal Length and Petal Length")
p4 <- qplot(x = Sepal.Length, y = Petal.Width, data = iris, geom = c("point","jitter"),
      alpha = 0.5, colour = Species,
      main = "(d) Sepal length and Petal Width")
(p1|p2)/(p3|p4)
```


(a) Sepal Length and Sepal Width

(b) Petal Length and Petal Width

(c) Sepal Length and Petal Length

(d) Sepal length and Petal Width

Figure 20: Relationship between features of sepal and petal by species

Note: Interpret all figures in your final analysis.