Visualization by example

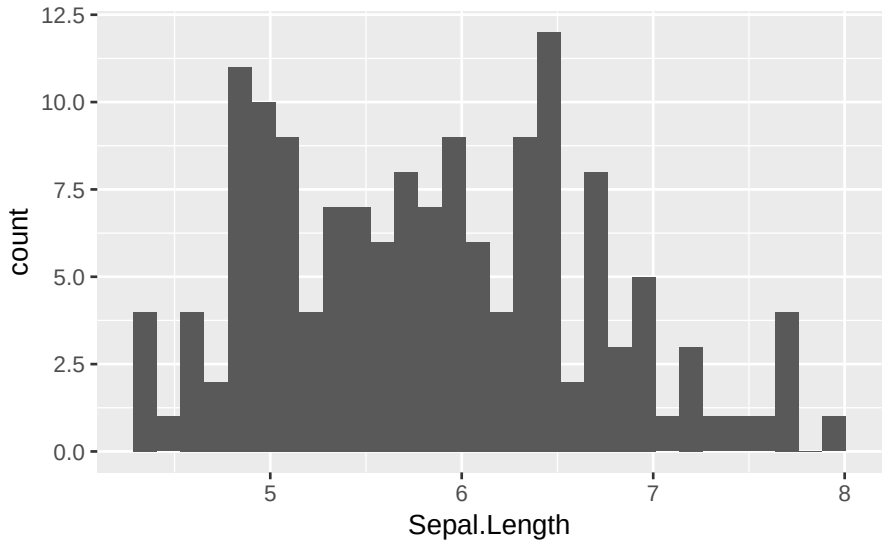
Gowri Prathap

The dataset we are working with is the iris, a R. A. Fisher's famous dataset that catalogs measurements of the Sepal and Petal of three species of iris flowers.

1. How many rows and columns does this dataset have?
   There are 4 columns and 150 rows.
2. Are the sepal and petal measurements stored as rows or as columns? The sepal and petal measurements are store as columns.
3. What are the three kinds of species of iris flower in the dataset? The three species of iris flower in the dataset are Setosa, Versicolor, virginica.
4. Why does the Species column repeat the same words multiple times (put another way, what does each row represent)? Flowers of the species Setosa, Versicolor, and Virginica are examined for sepal and petal measurements. Many flowers in these species are examined, that is why the name is repeated these many times.

Creating a histogram

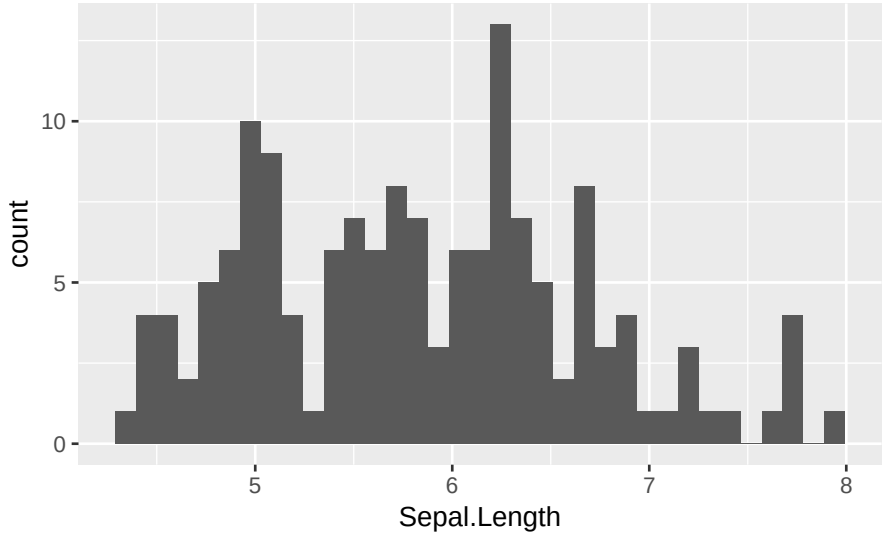
```
ggplot(data=iris) +
  geom_histogram(
    mapping = aes(x = Sepal.Length),
    bins = 30
  )
```



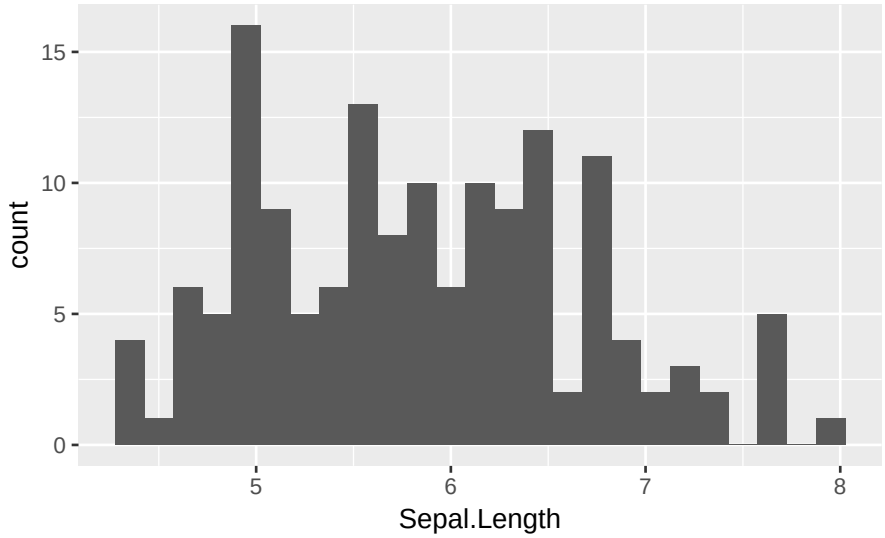
The values in the column 'Sepal.Length' from all rows is used to build the histogram. The height of each bar of the histogram gives the number of samples which have sepal length in the given range.

Changing the bin width

```
ggplot(data=iris) +
  geom_histogram(
    mapping = aes(x = Sepal.Length),
    bins = 35
  )
```



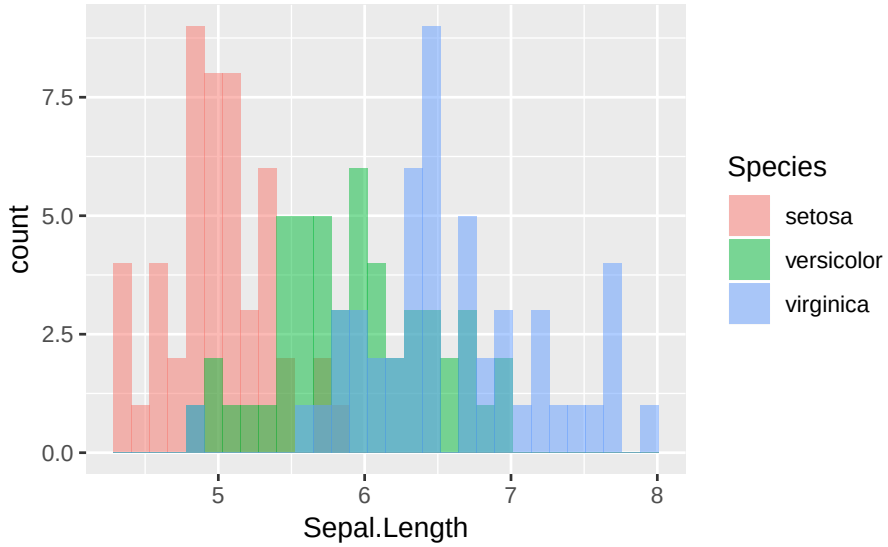

```
ggplot(data=iris) +
  geom_histogram(
    mapping = aes(x = Sepal.Length),
    bins = 25
  )
```



The first plot has more bins and shows more peaks and depressions, while the second plot shows lesser peaks. This shows that when the number of bins increases, the graph shows more detail, while when the number of bins decreases, the graph hides some detail.

Use of fill = “”

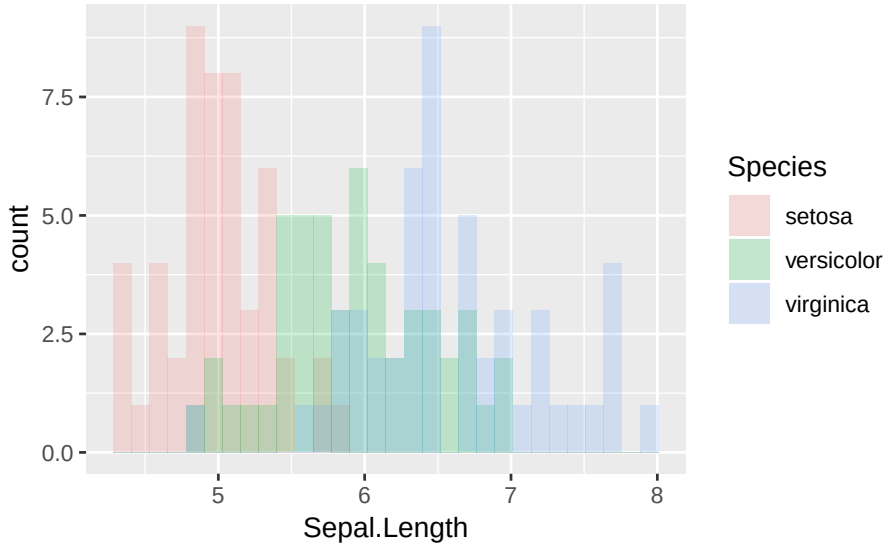
```
ggplot(data = iris) +
  geom_histogram(
    mapping = aes(x = Sepal.Length, fill = Species),
    bins = 30,
    alpha = 0.5,
    position = "identity"
  )
```



Adding fill = species divided the histogram into 3 separate ones (separated by colors), where there is one histogram each for each species. This changes our interpretation of our visualization, because it gives more detail about each species separately.

Use of alpha and position = “”

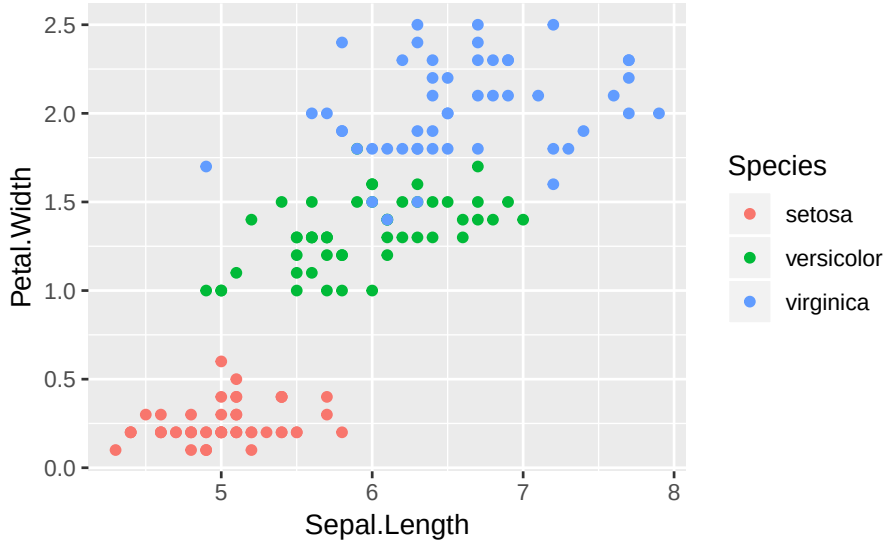
```
ggplot(data = iris) +
  geom_histogram(
    mapping = aes(x = Sepal.Length, fill = Species),
    bins = 30,
    alpha = 0.2,
    position = "identity"
  )
```



Changing alpha changes how transparent the bars of the histogram. Reducing alpha makes them more transparent, and increasing it makes it opaque. Removing position = "identity" causes the bars to not overlap anymore. Changing "identity" to "dodge" makes the histogram to bar charts, because the bars no longer touch each other. It puts overlapping objects besides one another.

Creating a scatterplot

```
ggplot(data = iris) +
  geom_point(
    mapping = aes(
      x = Sepal.Length,
      y = Petal.Width,
      color = Species
    ),
  )
```



There is a positive relationship between petal width and sepal length in case of the species. In case of virginica, this positive relationship is not as obvious as the other species. This scatterplot shows the relationship between sepal length and petal width. According to this plot, iris flowers with most sepal length have more petal width.