Homework 2

Sookja Kang, sk26949

This homework is due on Feb. 1, 2021 at 11:00pm. Please submit as a pdf file on Canvas.

In this homework you will be working with the `iris` dataset built into R. This data set contains measurements of flowers (sepal length, sepal width, petal length, petal width) for three different *Iris* species (*I. setosa*, *I. versicolor*, *I. virginica*).

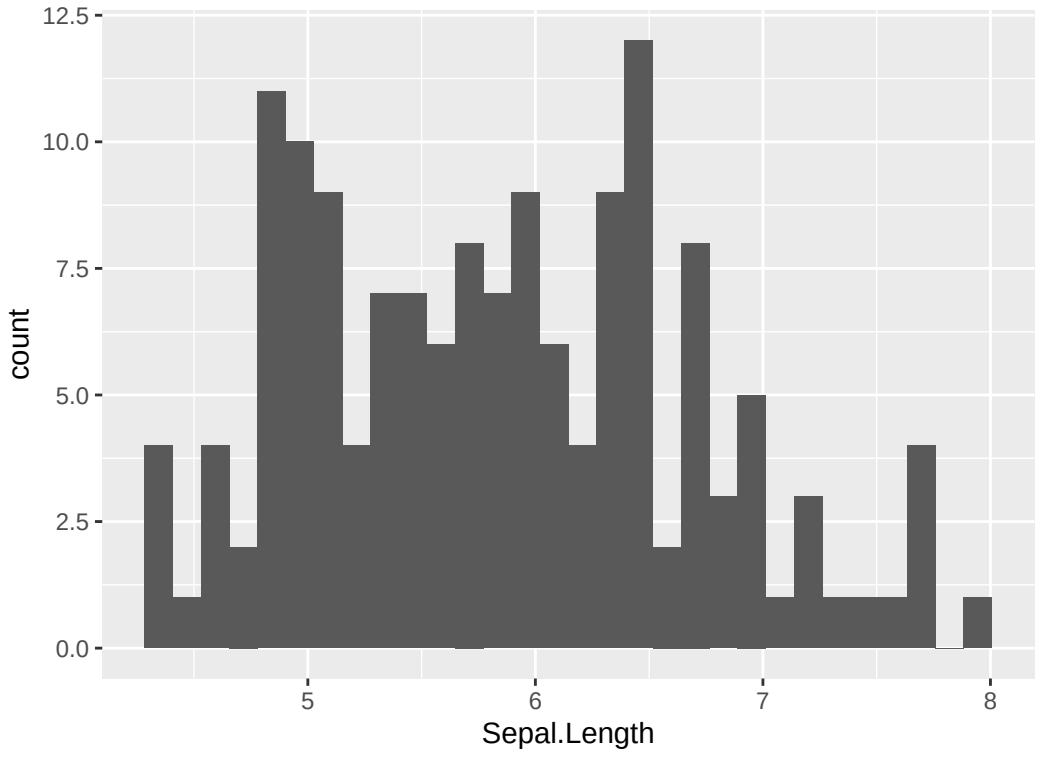
```
head(iris)
```

```
##   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
```

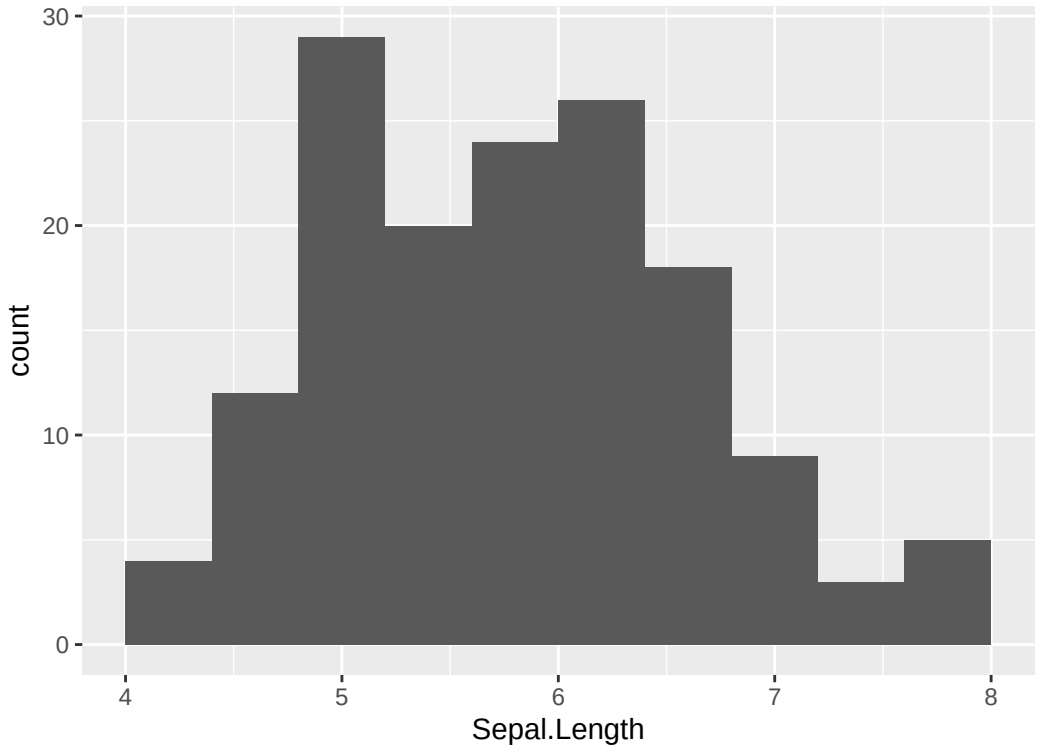
Problem 1: (6 pts) Use ggplot to make a histogram of the `Sepal.Length` column. Manually choose appropriate values for `binwidth` and `center`. Explain your choice of values in 2-3 sentences.

```
ggplot(iris, aes(Sepal.Length)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



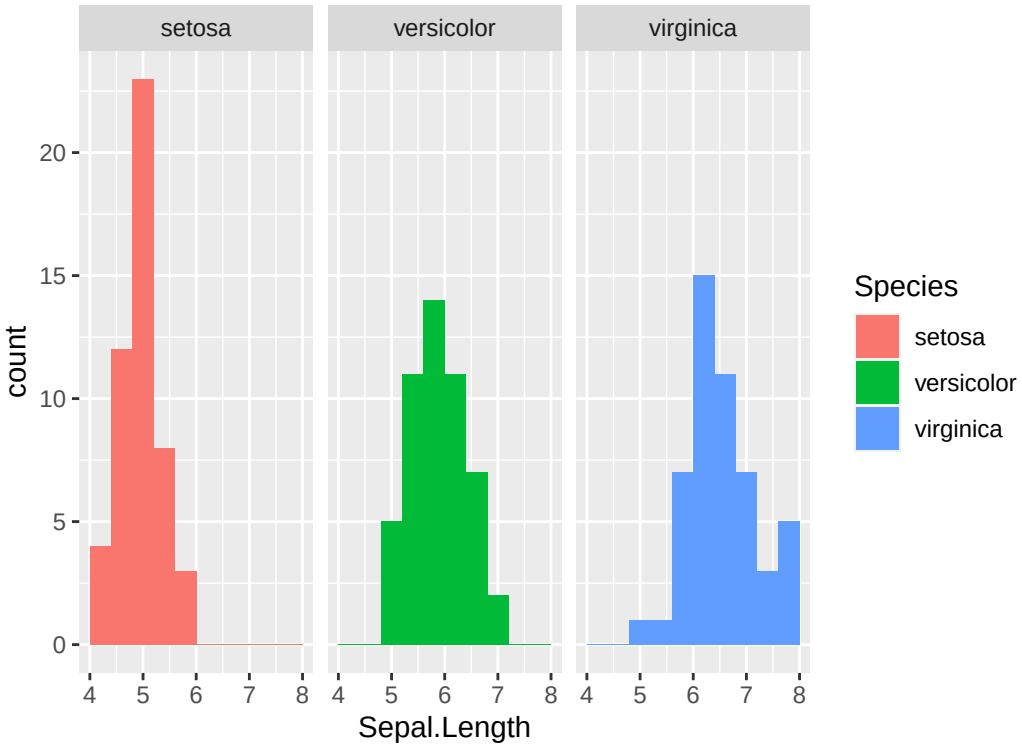

```
ggplot(iris, aes(Sepal.Length)) +
  geom_histogram(binwidth = .4,
                 center = .2)
```



When the bins = 30, the histogram is too peaky and difficult to visualize the main trend. Based on multiple trial runs with different bin width values (.2, .3, .4, .5, 1), the bin widths of .3 (center = .15) or .4 (center = .2) is appropriate to present this data trend that shows the dips and peaks.

Problem 2: (4 pts) Modify the plot from Problem 1 to show one panel per species. Hint: Use `facet_wrap()`. See Slide 14 from Class 2.

```
ggplot(iris, aes(Sepal.Length, fill = Species)) +
  geom_histogram(binwidth = .4,
                 center = .2) +
    facet_wrap(vars(Species))
```


setosa
versicolor
virginica
count
20
15
10
5
0
4
5
6
7
8
Sepal.Length
Species
setosa
versicolor
virginica