# Homework 3

*Sookja Kang, sk26949*

**This homework is due on Feb. 8, 2021 at 11:00pm. Please submit as a pdf file on Canvas.**

**Problem 1: (5 pts)** We will work again with the `iris` dataset built into R. It was previously introduced in Homework 2.

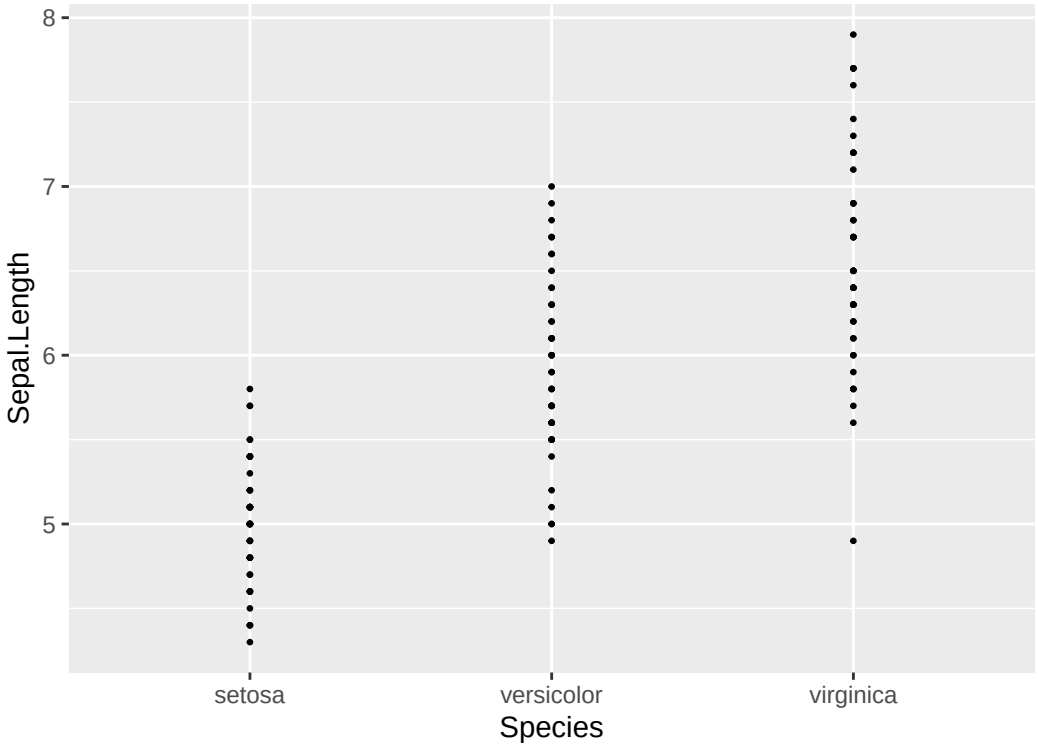
Make two different strip charts of sepal length versus species, the first one without horizontal jitter and second one with horizontal jitter. Explain in 1-2 sentences why the plot without jitter is highly misleading.

Hint: Make sure you do not accidentally apply vertical jitter. This is a common mistake many people make.

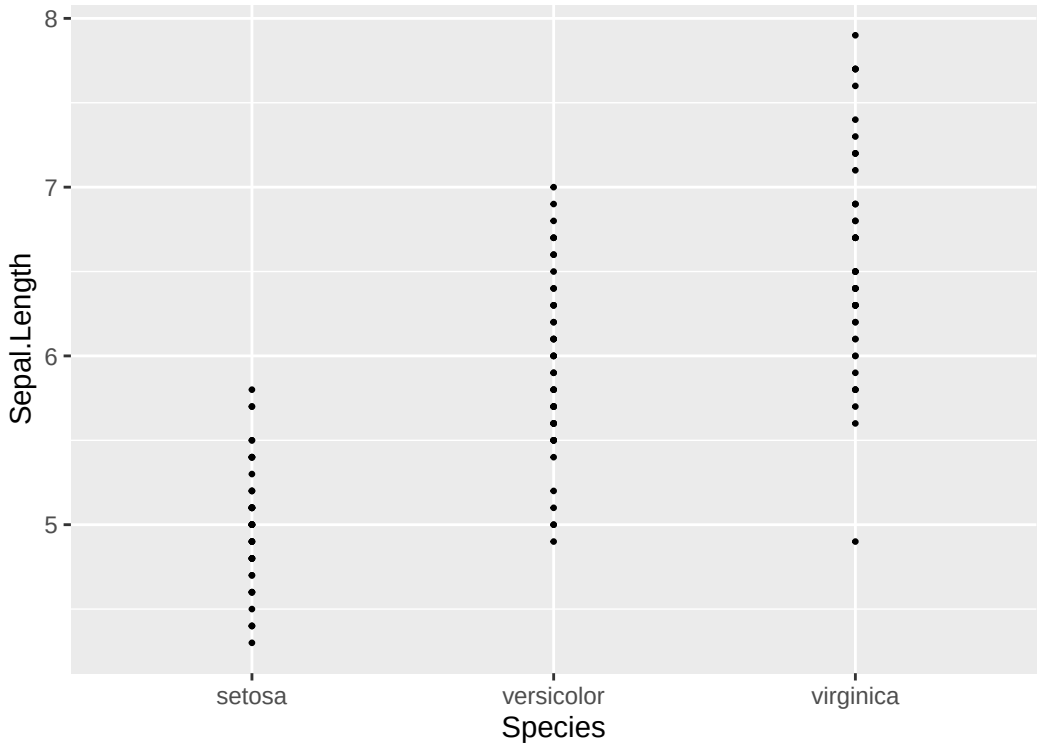
```
head(iris)
```

```
##   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
```

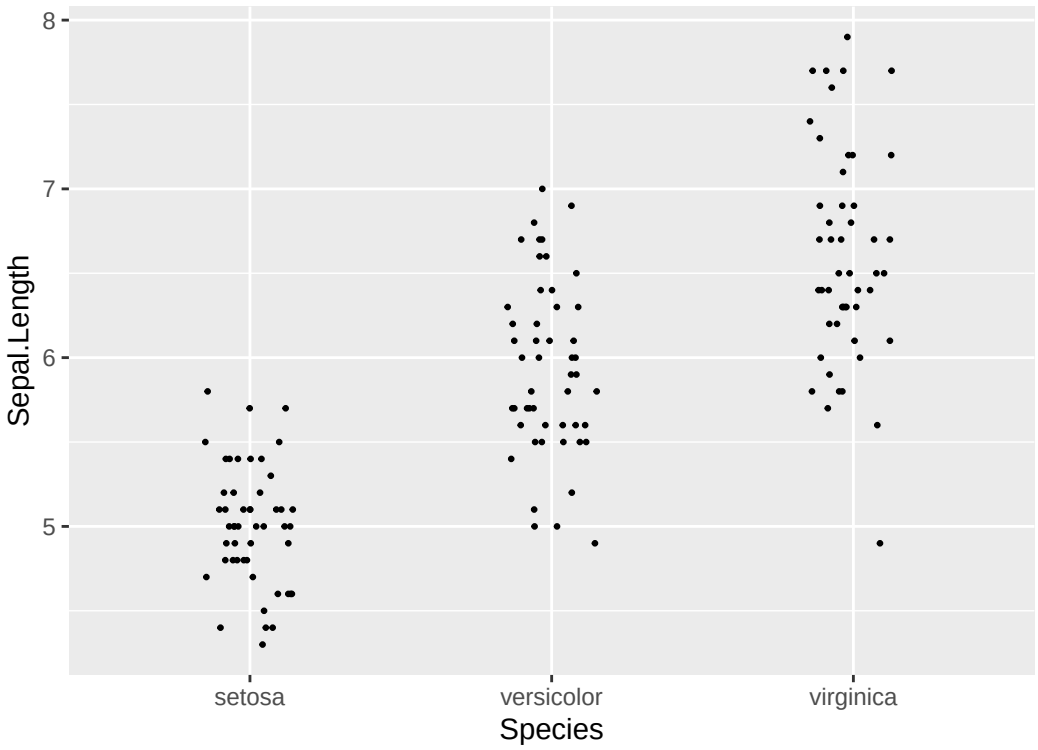
```
ggplot(iris, aes(Species, Sepal.Length)) +
  geom_point(size = 0.5)
```



```
ggplot(iris, aes(Species, Sepal.Length)) +
  geom_point(size = 0.5,
             position = position_jitter(width = 0, height = 0))
```




```
ggplot(iris, aes(Species, Sepal.Length)) +
  geom_point(size = 0.5,
             position = position_jitter(width = .15, height = 0))
```




*Each data points are plotted on top of each other on the first and second charts since these doesn't have horizontal jitter. Thus, it is difficult to see how sepal.lengths by species are distributed on the strip chart without jitter.*

**Problem 2: (5 pts)** For this problem, we will be working with the `Aus_athletes` dataset that comes with the **ggridges** package:

```
head(Aus_athletes)
```

```
##    rcc wcc   hc   hg ferr   bmi   ssf pcBfat   lbm height weight sex      sport
## 1 3.96 7.5 37.5 12.3   60 20.56 109.1  19.75 63.32  195.9   78.9   f basketball
## 2 4.41 8.3 38.2 12.7   68 20.67 102.8  21.30 58.55  189.7   74.4   f basketball
## 3 4.14 5.0 36.4 11.6   21 21.86 104.6  19.88 55.36  177.8   69.1   f basketball
## 4 4.11 5.3 37.3 12.6   69 21.88 126.4  23.66 57.18  185.0   74.9   f basketball
## 5 4.45 6.8 41.5 14.0   29 18.96  80.3  17.64 53.20  184.6   64.6   f basketball
## 6 4.10 4.4 37.4 12.5   42 21.04  75.2  15.58 53.77  174.0   63.7   f basketball
```

This dataset contains various physiological measurements made on athletes competing in different sports. Here, we are only interested in the columns `height`, indicating the athleete's height in cm, `sex`, indicating whether an athlete is male or female, and `sport`, indicating the sport the athlete competes in.

Visualize the distribution of athletes' heights by sex and sport with (i) boxplots and (ii) ridgelines. Make one plot per geom and do not use faceting. In both cases, put height on the x axis and sport on the y axis. Use color to indicate the athlete's sex.

The boxplot ggplot generates will have a problem. Explain what the problem is. (You do not have to solve it.)

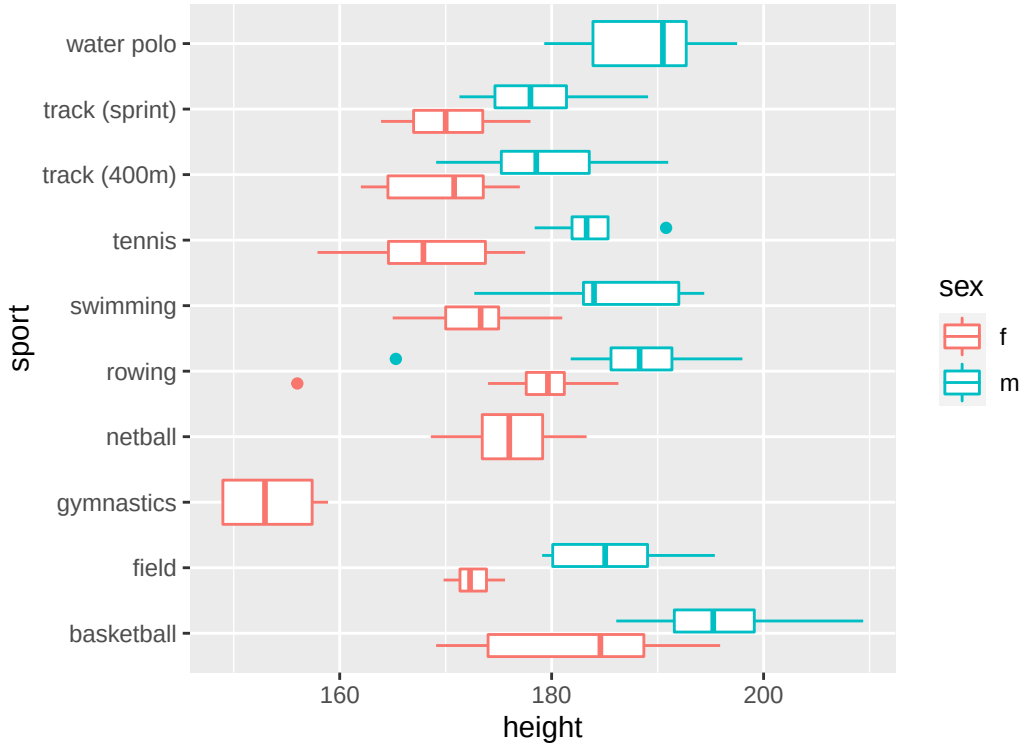
```
range(Aus_athletes$height)
```

```
## [1] 148.9 209.4
```

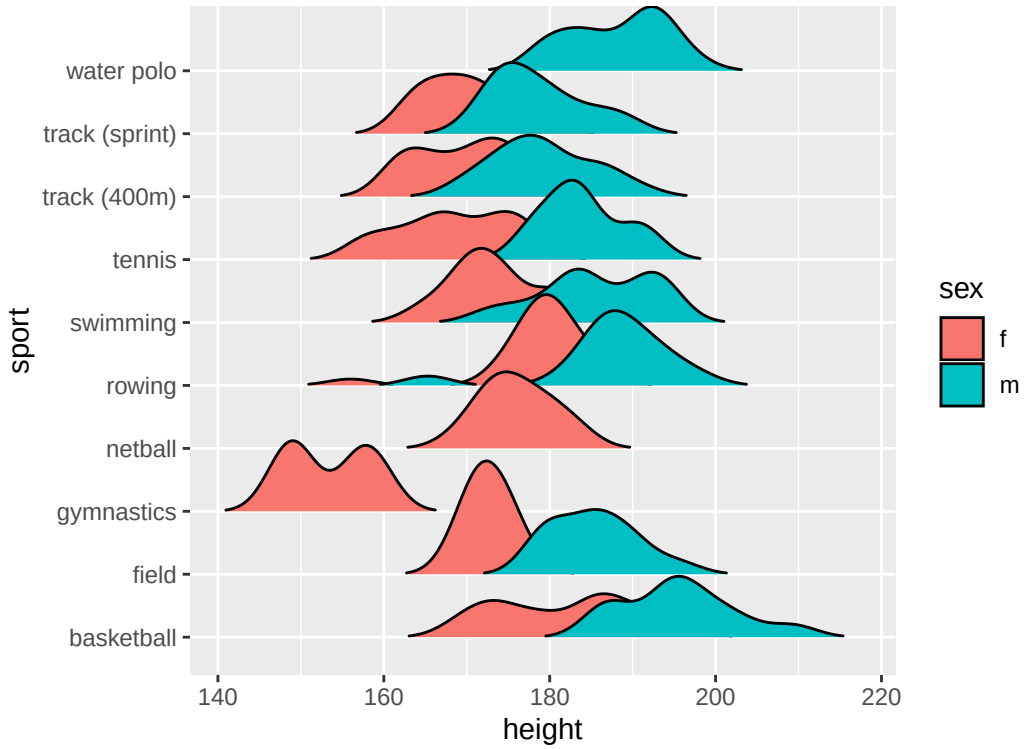
```
#view(Aus_athletes)
```

```
#heights (x) by sex and sports (y) with boxplot()
#ridgelines

ggplot(Aus_athletes, aes(height, sport, color = sex)) +
  geom_boxplot()
```




```
ggplot(Aus_athletes, aes(height, sport, fill = sex)) +
  geom_density_ridges(rel_min_height = 0.01,
                      bandwidth = 2.8)
```




*This boxplot chart provides a rough idea that male athletes are taller than female athletes. Gymnastics and netball have only female athletes; water polo has only male athletes. These affected the size of the boxplots (inconsistent sizes of boxplots). It needs to be corrected to show consistency of boxplots on this chart.*