Homework 2

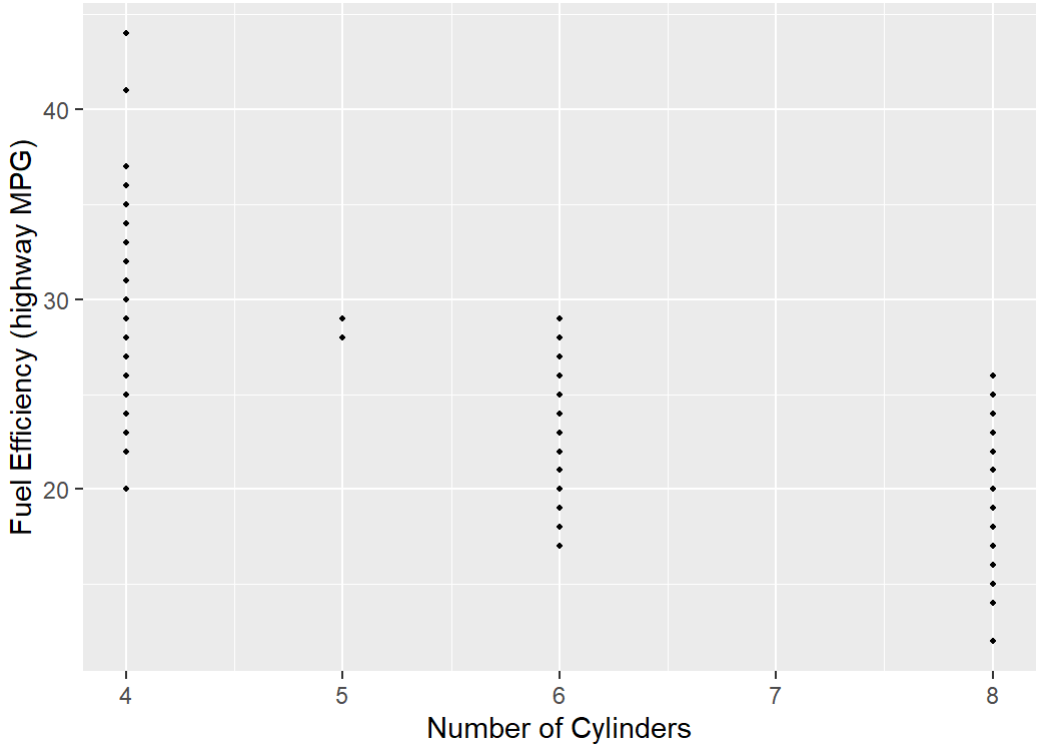
This homework is due on the deadline posted on edX. Please submit a .pdf file of your output and upload a .zip file with your .Rmd file.

Problem 1: We will work with the `mpg` dataset provided by **ggplot2**. See here for details: https://ggplot2.tidyverse.org/reference/mpg.html (https://ggplot2.tidyverse.org/reference/mpg.html)

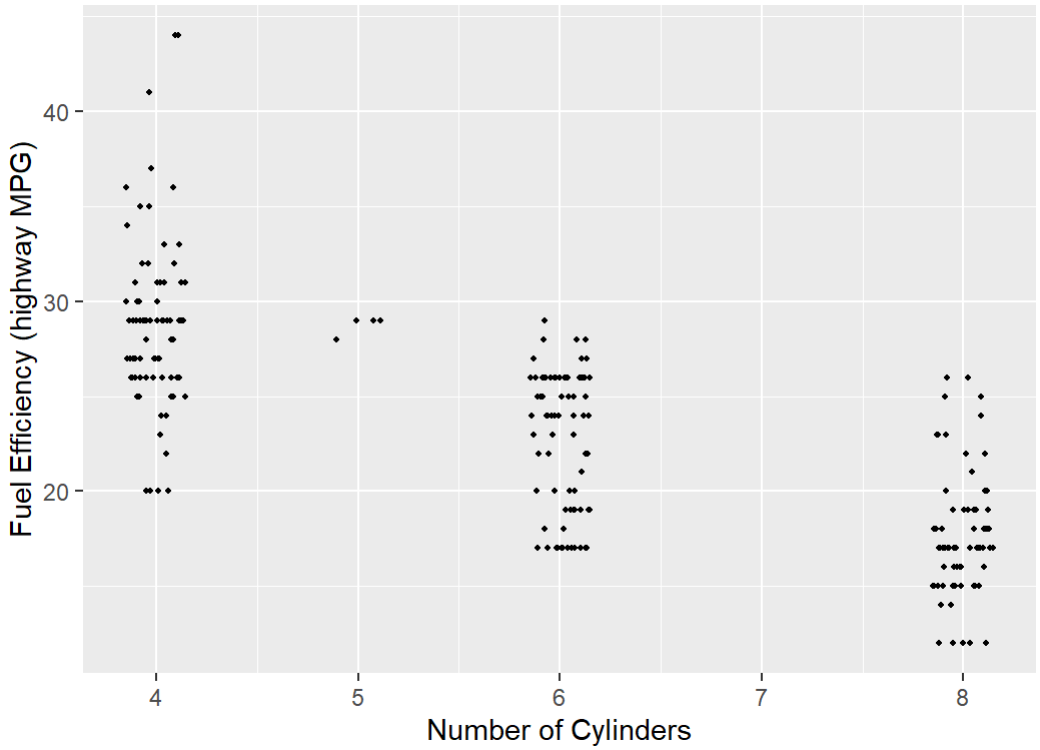
Make two different strip charts of highway fuel economy versus number of cylinders, the first one without horizontal jitter and second one with horizontal jitter. Explain in 1-2 sentences why the plot without jitter is highly misleading.

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

```
ggplot(mpg, aes(cyl, hwy)) +
  geom_point(size = 0.75) +
  ylab("Fuel Efficiency (highway MPG)") +
  xlab("Number of Cylinders")
```




```
ggplot(mpg, aes(cyl, hwy)) +
  geom_point(
    size = 0.75,
    position = position_jitter(
      width = 0.15,
      height = 0)) +
  ylab("Fuel Efficiency (highway MPG)") +
  xlab("Number of Cylinders")
```

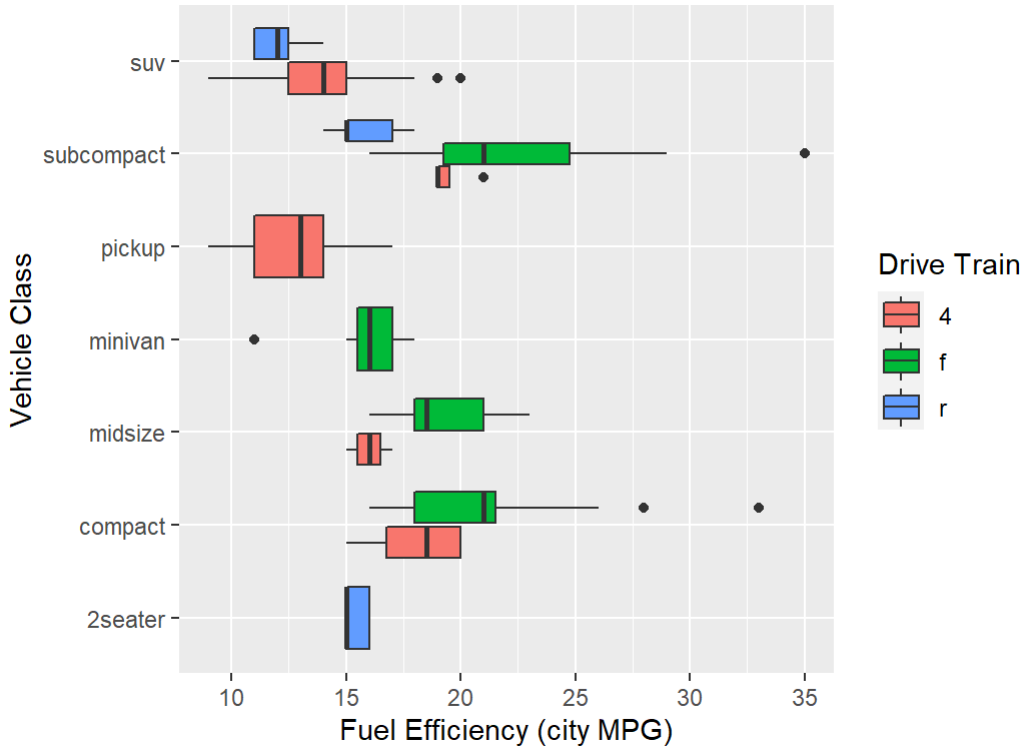


Notice that the points in the first plot are not very numerous in appearance; this is because there are actually many points overlapping each other. This is misleading because it does not accurately represent the number of data points in the dataset, while the lower plot with jitter reveals many more data points.

Problem 2: For this problem, we will continue working with the `mpg` dataset. Visualize the distribution of each car's city fuel economy by class and type of drive train with (i) boxplots and (ii) ridgelines. Make one plot per geom and do not use faceting. In both cases, put city mpg on the x axis and class on the y axis. Use color to indicate the car's drive train.

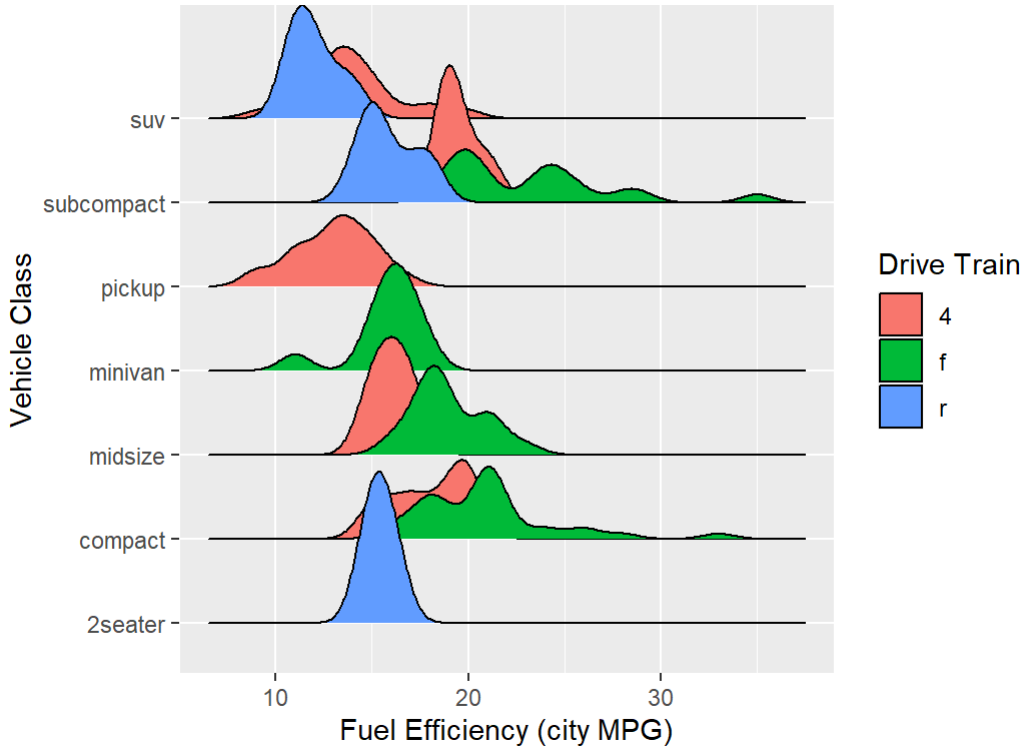
The boxplot ggplot generates will have a problem. Explain what the problem is. (You do not have to solve it.)

```
ggplot(mpg, aes(cty, class, fill = drv)) +
  geom_boxplot() +
  labs(
    y = "Vehicle Class",
    x = "Fuel Efficiency (city MPG)",
    fill = "Drive Train")
```




```
ggplot(mpg, aes(cty, class, fill = drv)) +
  geom_density_ridges() +
  labs(
    y = "Vehicle Class",
    x = "Fuel Efficiency (city MPG)",
    fill = "Drive Train")
```

```
## Picking joint bandwidth of 0.828
```



The issue with the default boxplots generated is that the width of the boxplots is dependent on the availability of data for each drive train in each class of vehicle. Pickup trucks, for example, only has data with 4 wheel drive, so it is particularly wide. This is ugly and even disorienting to the viewer.