Homework 2

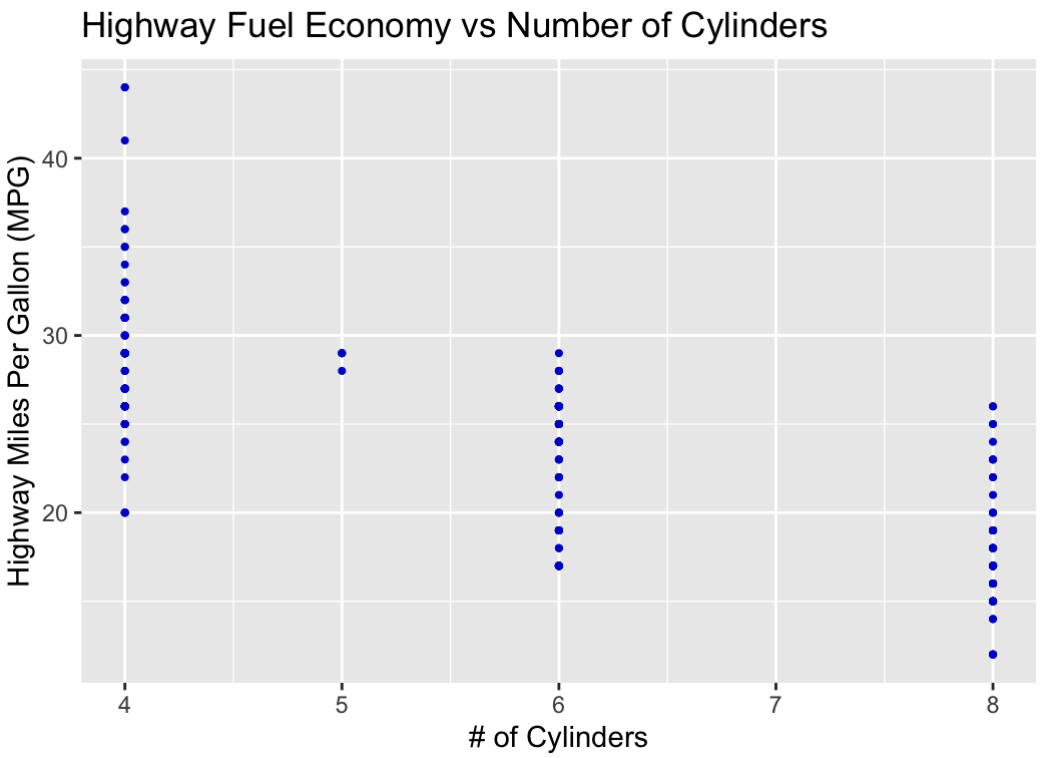
This homework is due on the deadline posted on edX. Please submit a .pdf file of your output and upload a .zip file containing your .Rmd file. Do NOT include your name or EID in your filenames.

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)

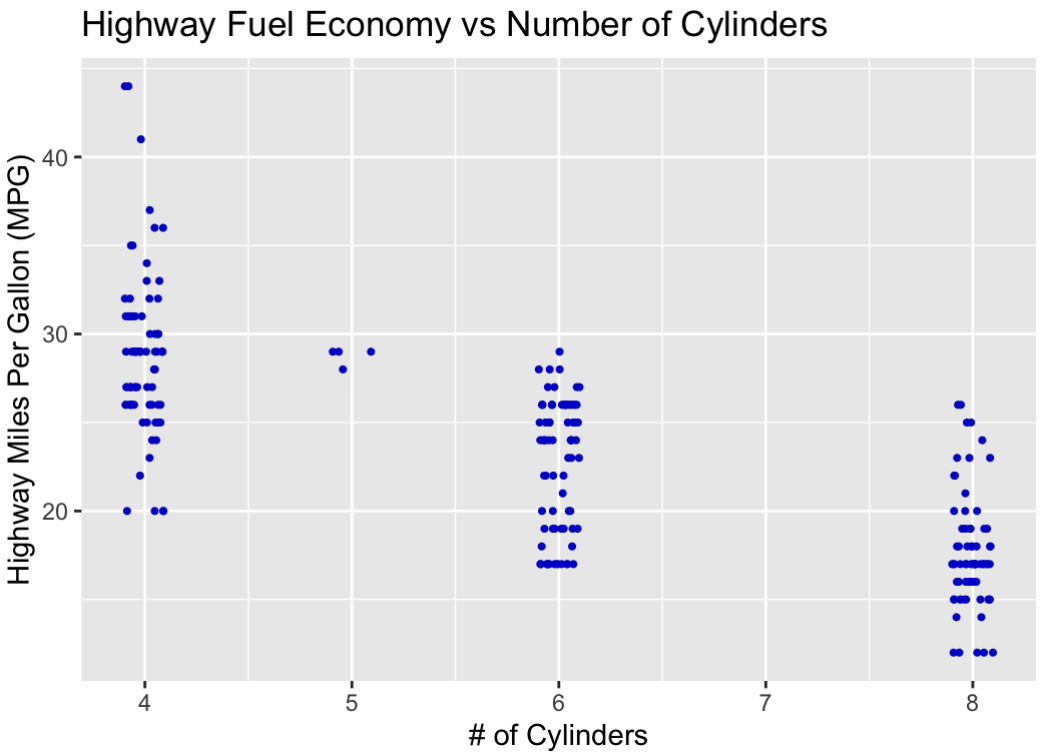
Make two different strip charts of highway fuel economy versus number of cylinders, the first one without horizontal jitter and second one with horizontal jitter. Explain in 1-2 sentences why the plot without jitter is highly misleading.

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, color="blue3",
             position = position_jitter(
               width = 0,
               height = 0
             )) +
  xlab('# of Cylinders') +
  ylab('Highway Miles Per Gallon (MPG)') +
  ggtitle('Highway Fuel Economy vs Number of Cylinders')
```



```
ggplot(mpg, aes(cyl, hwy)) +
  geom_point(size = 0.75, color="blue3",
             position = position_jitter(
               width = 0.1,
               height = 0
             )) +
  xlab('# of Cylinders') +
  ylab('Highway Miles Per Gallon (MPG)') +
  ggtitle('Highway Fuel Economy vs Number of Cylinders')
```



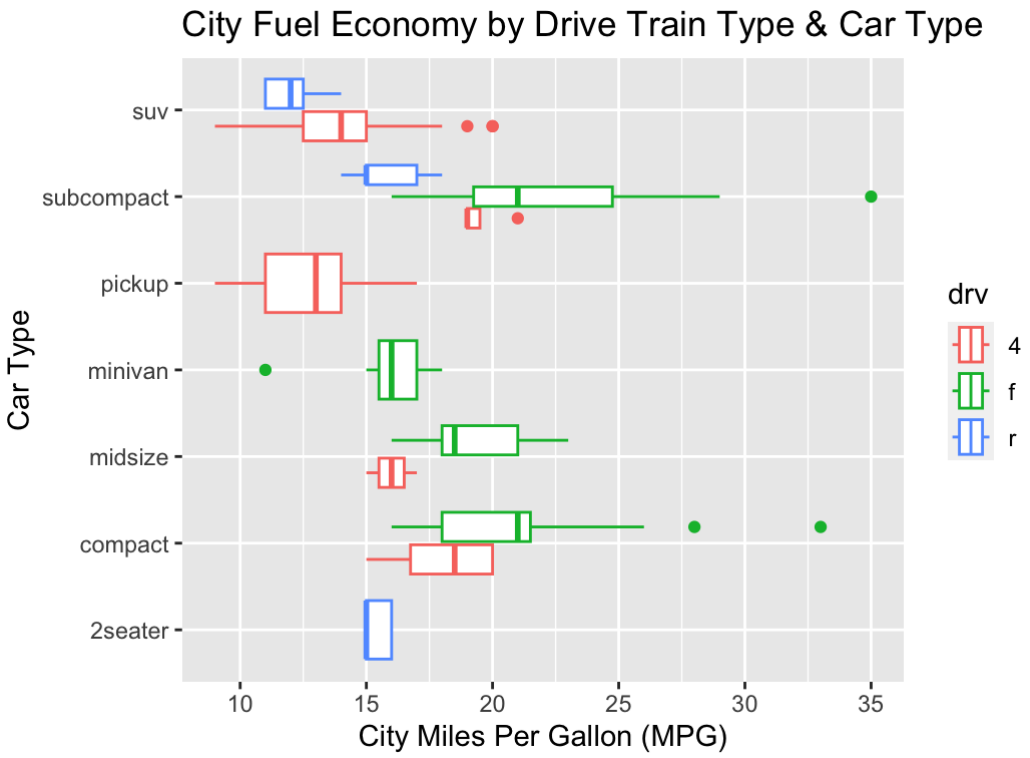
Your explanation goes here.

The reason the plot without jitter is misleading is because it underestimates the true size of the population within the dataset. For instance, there may be many 4 cylinder vehicles with a 30 Highway MPG, which would only display as one value in a plot without jitter.

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.

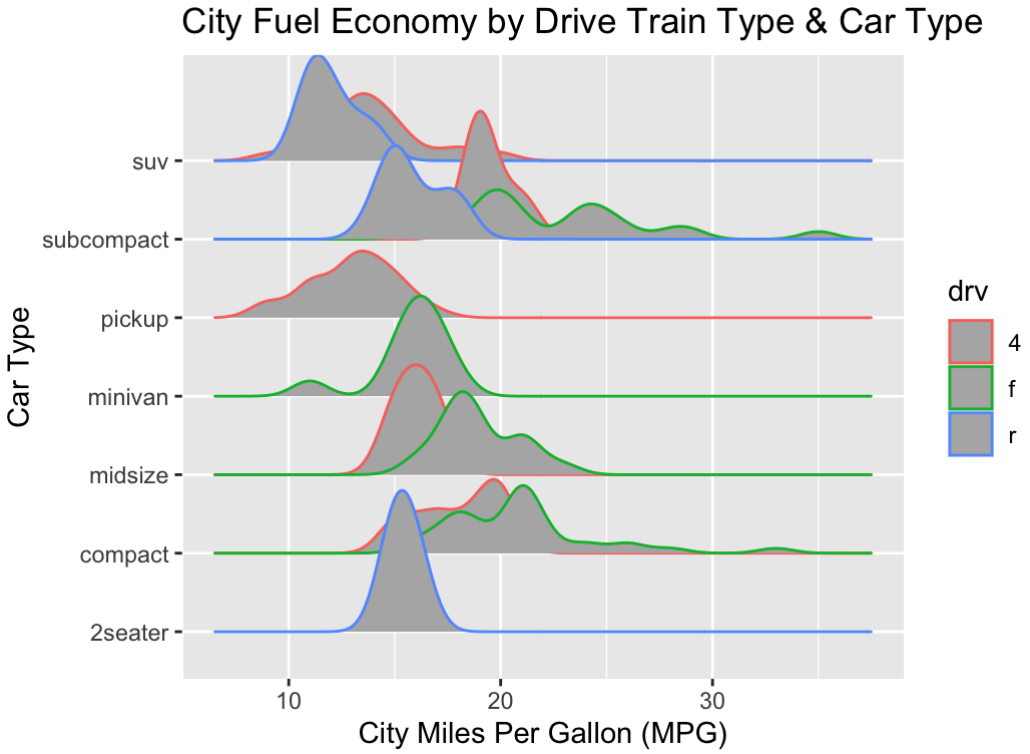
The boxplot ggplot generates will have a problem. Explain what the problem is. (You do not have to solve it.)

```
ggplot(mpg, aes(cty, class, color=drv)) +
  geom_boxplot() +
  xlab('City Miles Per Gallon (MPG)') +
  ylab('Car Type') +
  ggtitle('City Fuel Economy by Drive Train Type & Car Type')
```




```
ggplot(mpg, aes(cty, class, color=drv)) +
  geom_density_ridges() +
  xlab('City Miles Per Gallon (MPG)') +
  ylab('Car Type') +
  ggtitle('City Fuel Economy by Drive Train Type & Car Type')
```

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



Your explanation goes here.

The boxplot may be misleading in the fact that it could be difficult to interpret which boxplot is associated with what car type, as some boxplots are situated between each category. It is possible that this issue could be compounded if the field for which the color is based off of has many categories.