# Homework 4

*Ethan Chang - ehc586*

**This homework is due on Feb. 14, 2023 at 11:00pm. Please submit as a pdf file on Canvas.**

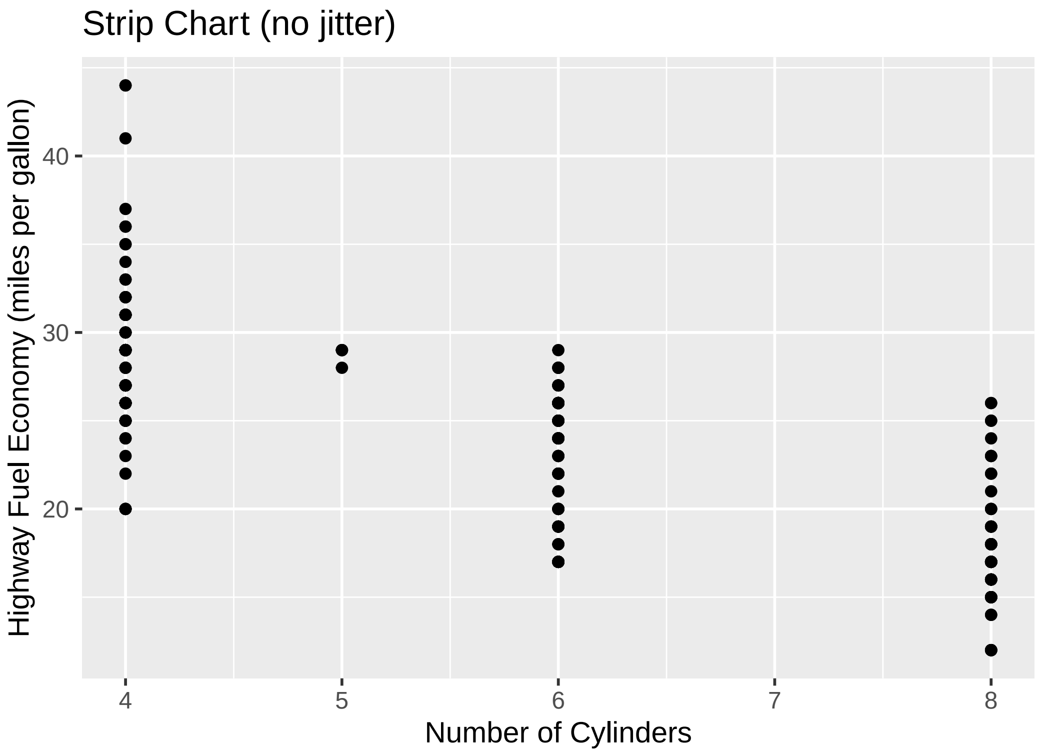
**Problem 1: (4 pts)** We will work with the `mpg` dataset provided by **ggplot2**. See here for details: https://ggplot2.tidyverse.org/reference/mpg.html

Make two different strip charts of highway fuel economy (`hwy`) versus number of cylinders (`cyl`), the first one without horizontal jitter and second one with horizontal jitter. In both plots, please replace names of the data columns (`hwy`, `cyl`) along the axes with nice, easily readable lables.

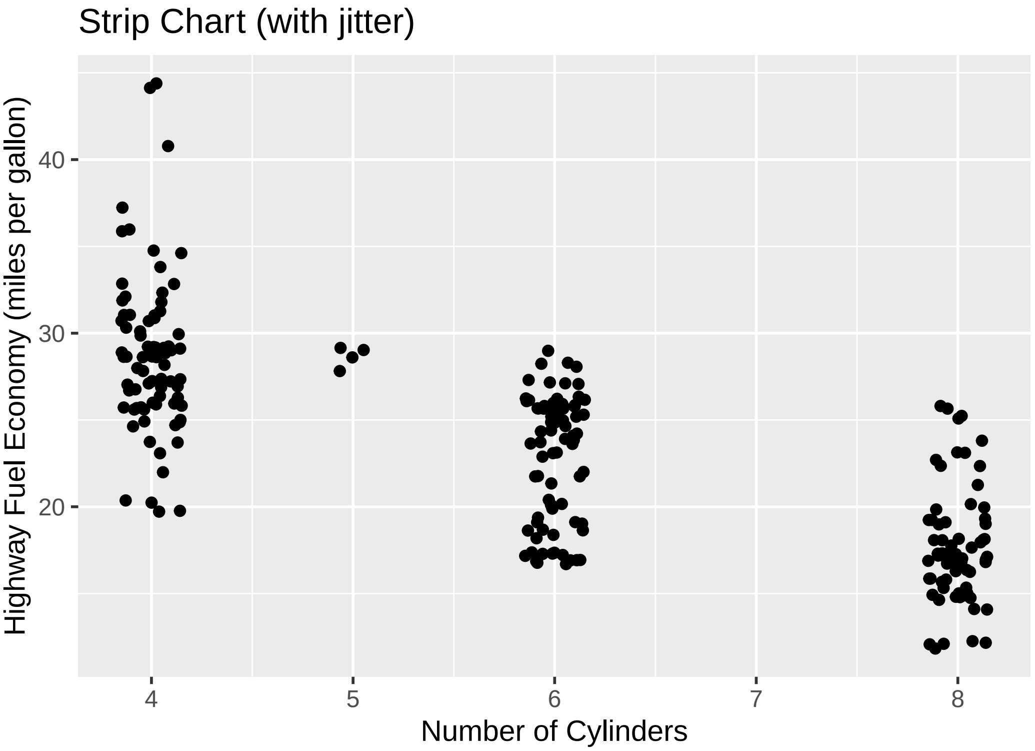
Explain in 1-2 sentences why the plot without jitter is misleading.

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() +
  xlab("Number of Cylinders") +
  ylab("Highway Fuel Economy (miles per gallon)") +
  ggtitle("Strip Chart (no jitter)")
```



```
ggplot(mpg, aes(cyl, hwy)) +
  geom_point(position = position_jitter(width = 0.15)) +
  xlab("Number of Cylinders") +
  ylab("Highway Fuel Economy (miles per gallon)") +
  ggtitle("Strip Chart (with jitter)")
```



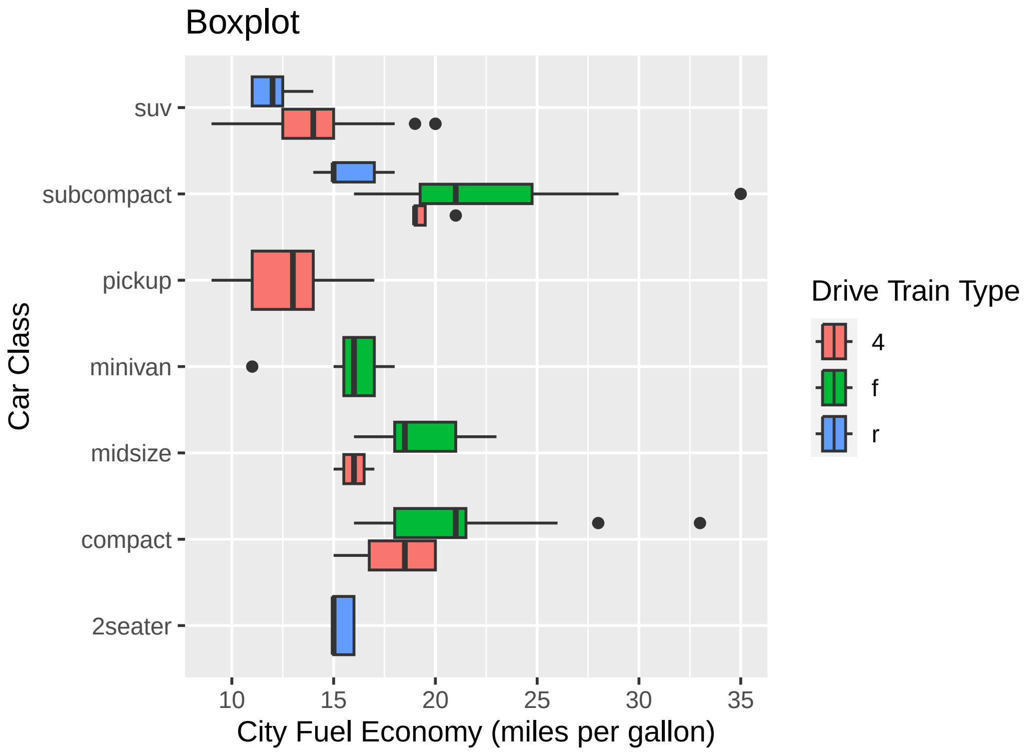
*From the graphs above, it can be seen that the plot without jitter is misleading as it can be seen that it has a lot less visible points compared to the plot with jitter. This is due to the overlapping of points with the same highway fuel economy and number of cylinders, making it seem as if there are a lot less points than there actually are and fails to show the true distribution of highway fuel economy for each number of cylinders.*

**Problem 2: (6 pts)** For this problem, we will continue working with the `mpg` dataset. Visualize the distribution of each car's city fuel economy by class (`class`) and type of drive train (`drv`) 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. As in Problem 1, rename the axis labels.

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

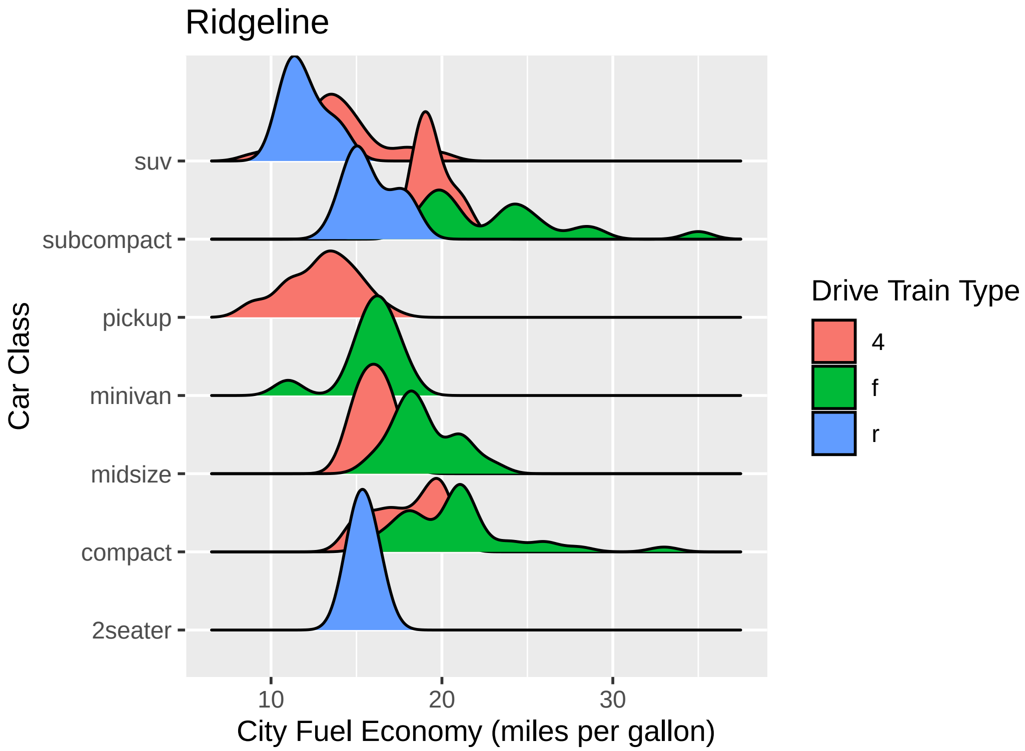
Hint: To change the name of the legend, use `+ labs(fill = "legend name")`

```
ggplot(mpg, aes(cty, class, fill = drv)) +
  geom_boxplot() +
  labs(title = "Boxplot", x = "City Fuel Economy (miles per gallon)",
       y = "Car Class", fill = "Drive Train Type")
```



```
ggplot(mpg, aes(cty, class, fill = drv)) +
  geom_density_ridges() +
  labs(title = "Ridgeline", x = "City Fuel Economy (miles per gallon)",
       y = "Car Class", fill = "Drive Train Type")
```

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



*From the boxplot above, it can be seen that ggplot's boxplots are all of different heights for each car class depending on the number of drive train types available. Car classes with more drive train types tend to have boxplots with shorter heights, while car classes with less drive train types tend to have boxplots with taller heights. This is a problem visually and can make the plot more difficult to interpret and compare.*