Homework 4

Enter your name and EID here

Connor Hanna cdh3663

This homework is due on Feb. 14, 2022 at 11:00am. 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

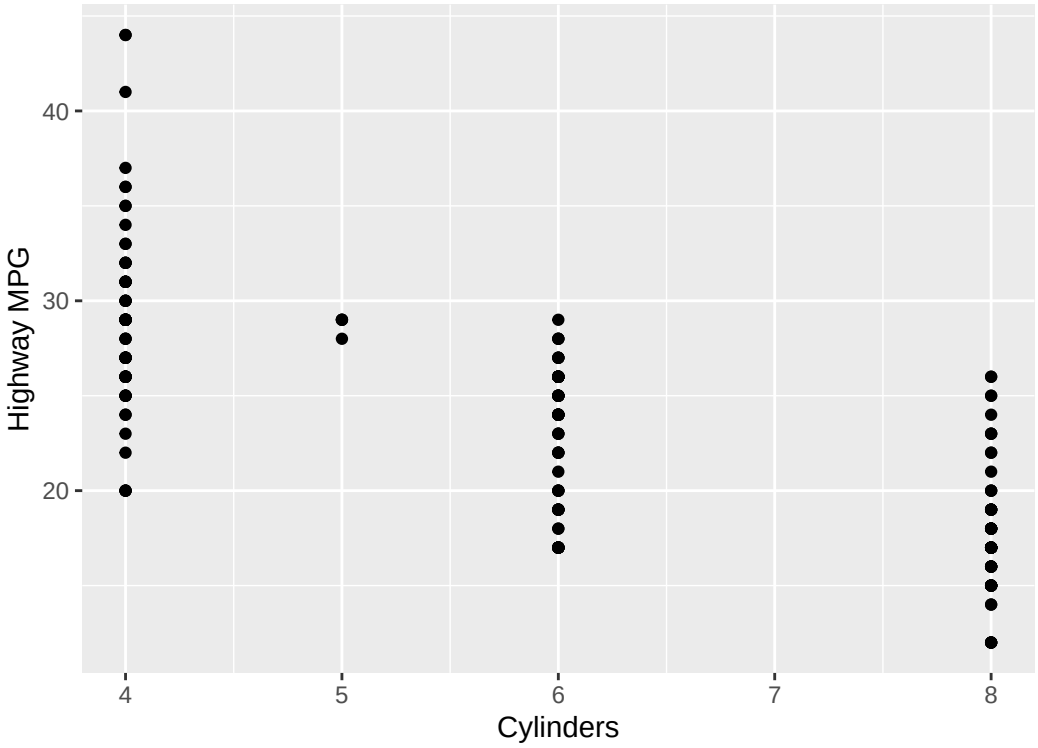
Make two different strip charts of highway fuel economy versus number of cylinders, the first one without horizontal jitter and second one with horizontal jitter. Explain in 1-2 sentences why the plot without jitter is highly misleading. Don't forget to rename axes labels.

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

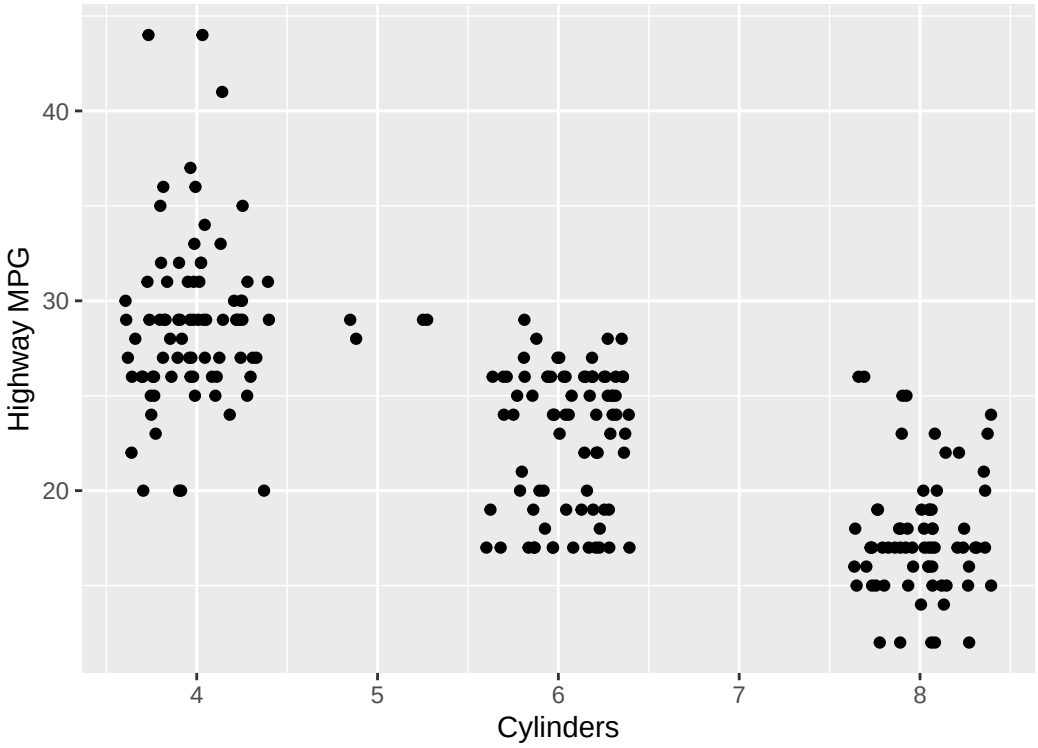
```
mpg
```

```
## # A tibble: 234 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4           1.8  1999     4 auto~ f        18    29 p     comp~
##  2 audi         a4           1.8  1999     4 manu~ f        21    29 p     comp~
##  3 audi         a4           2    2008     4 manu~ f        20    31 p     comp~
##  4 audi         a4           2    2008     4 auto~ f        21    30 p     comp~
##  5 audi         a4           2.8  1999     6 auto~ f        16    26 p     comp~
##  6 audi         a4           2.8  1999     6 manu~ f        18    26 p     comp~
##  7 audi         a4           3.1  2008     6 auto~ f        18    27 p     comp~
##  8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p     comp~
##  9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p     comp~
## 10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p     comp~
## # ... with 224 more rows
```

```
ggplot(mpg, aes(cyl, hwy)) +
  geom_point() +
  scale_x_continuous(
    name = "Cylinders"
  ) +
  scale_y_continuous(
    name = "Highway MPG"
  )
```




```
ggplot(mpg, aes(cyl, hwy)) +
  geom_point(position = position_jitter(height = 0.0)) +
  scale_x_continuous(
    name = "Cylinders"
  ) +
  scale_y_continuous(
    name = "Highway MPG"
  )
```



Your explanation goes here.

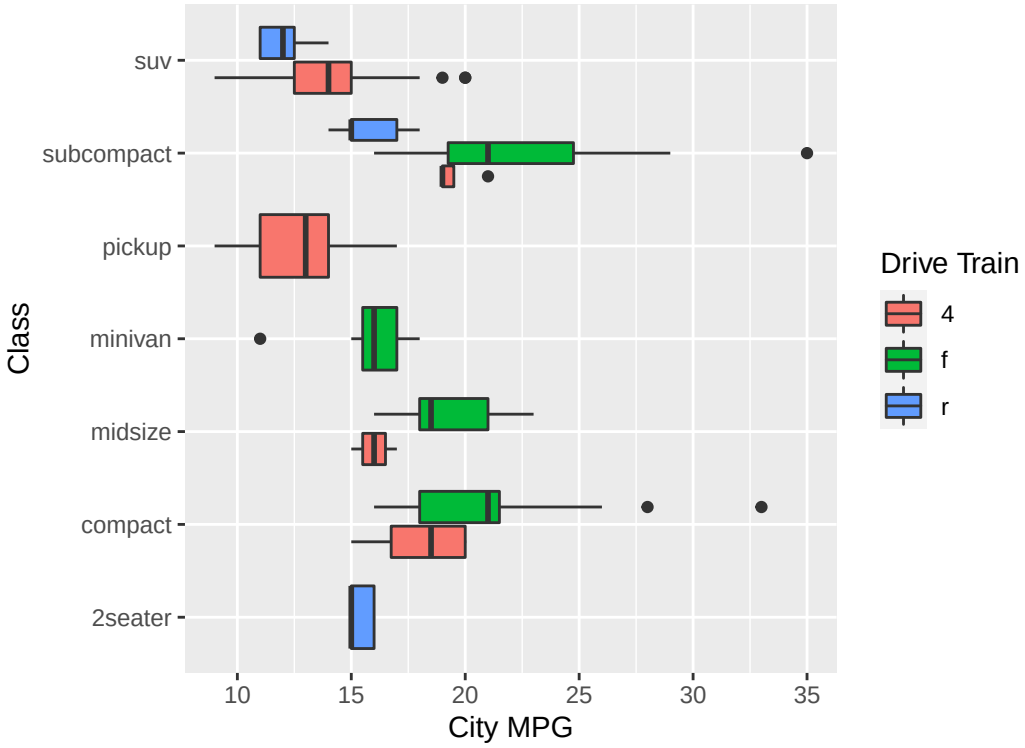
The plot without jitter doesn't show the density of observations along the y axis - making it look like the relationship between cylinders and highway MPG is much stronger than it actually is. It's also much more difficult to read, and it is ugly.

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 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. Don't forget to rename axes labels.

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

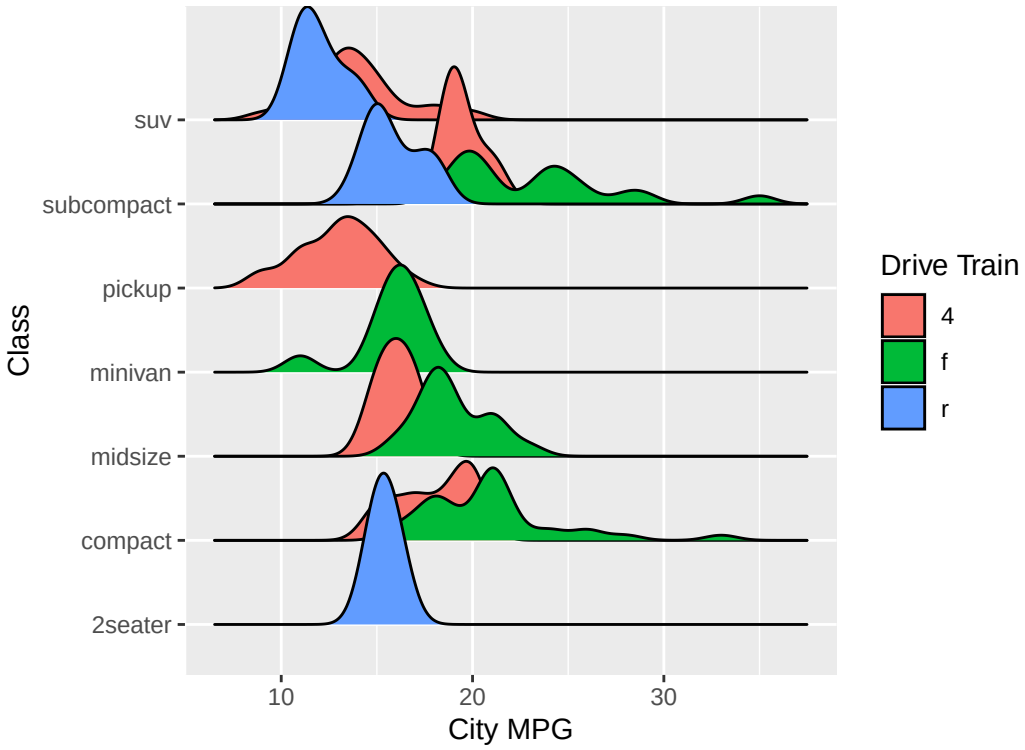
Hint: To change the name of the legend, use `+ labs(fill = "legend name")`

```
ggplot(mpg, aes(cty, class, fill = drv)) +
  geom_boxplot() +
  labs(fill = "Drive Train") +
  scale_x_continuous(
    name = "City MPG"
  ) +
  scale_y_discrete(
    name = "Class"
  )
```




```
ggplot(mpg, aes(cty, class, fill = drv)) +
  geom_density_ridges() +
  labs(fill = "Drive Train") +
  scale_x_continuous(
    name = "City MPG"
  ) +
  scale_y_discrete(
    name = "Class"
  )
```

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



Your explanation goes here.

The way the fill has rendered drivetrain information makes the boxplots difficult to read, since the spacing/width of drivetrain boxplots varies between vehicle classes. It also makes it difficult to tell which boxplots belong to which classes of vehicle.