Homework 2

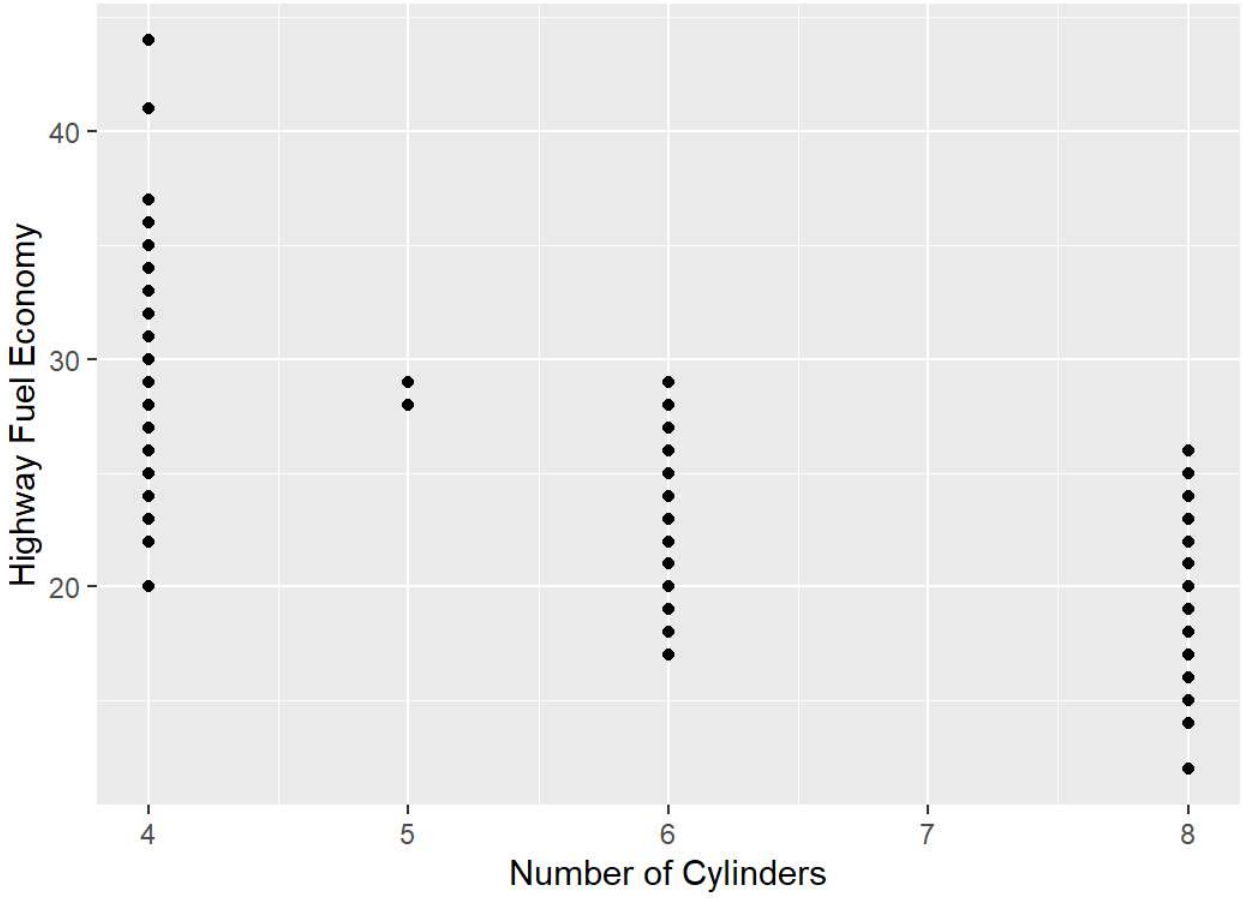
This homework is due on the deadline posted on edX. Please submit a .pdf file of your output and upload a .zip file with your .Rmd file.

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)

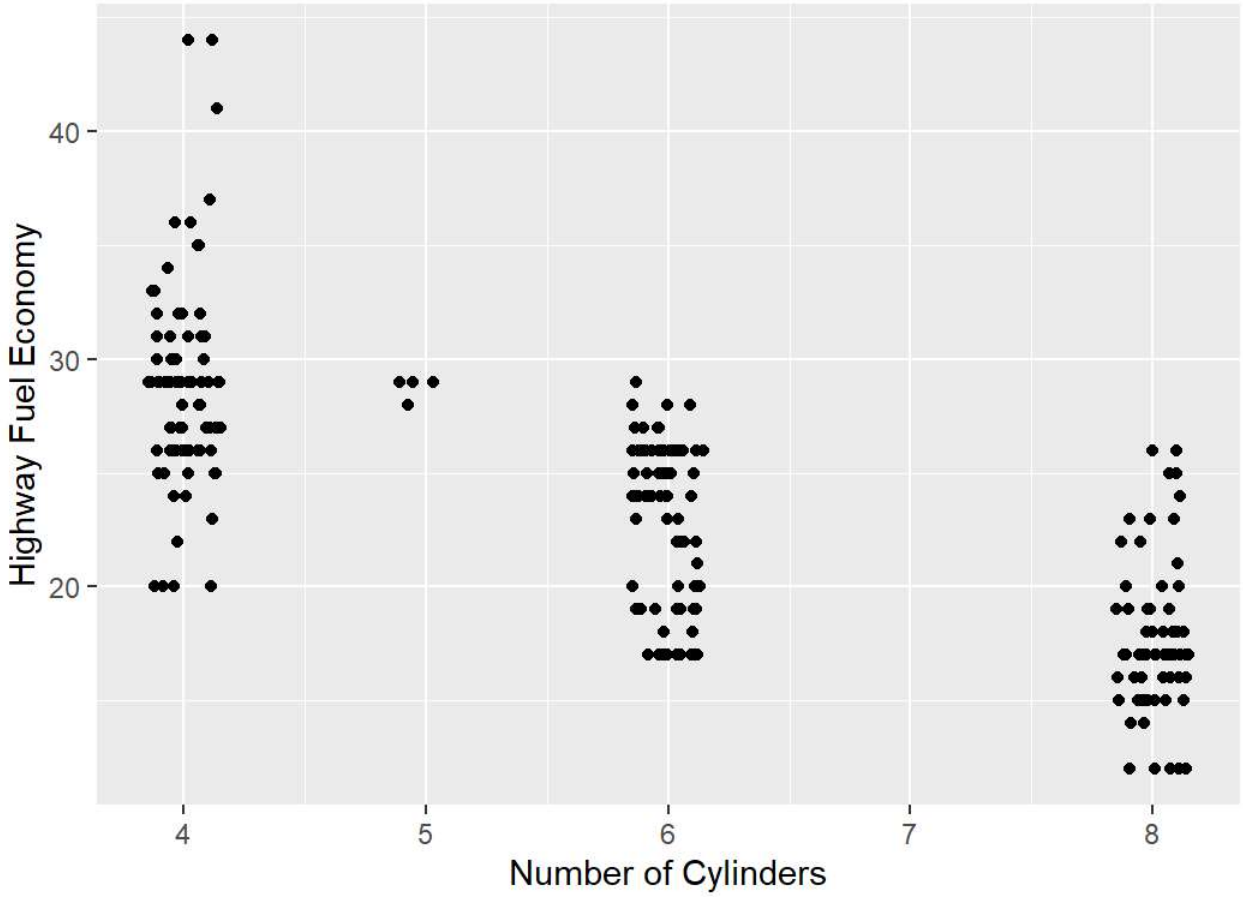
Make two different strip charts of highway fuel economy versus number of cylinders, the first one without horizontal jitter and second one with horizontal jitter. Explain in 1-2 sentences why the plot without jitter is highly misleading.

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()+ylab("Highway Fuel Economy")+xlab("Number of Cylinders")
```



```
ggplot(mpg,aes(cyl,hwy))+geom_point(position=position_jitter(height=0,width=0.15))+ylab("Highway Fuel Economy")+xlab("Number of Cylinders")
```



We could only see fewer points without jittering because lots of data points lie on top of each other and fully overlap, appear much less than they actually are, which is highly misleading.

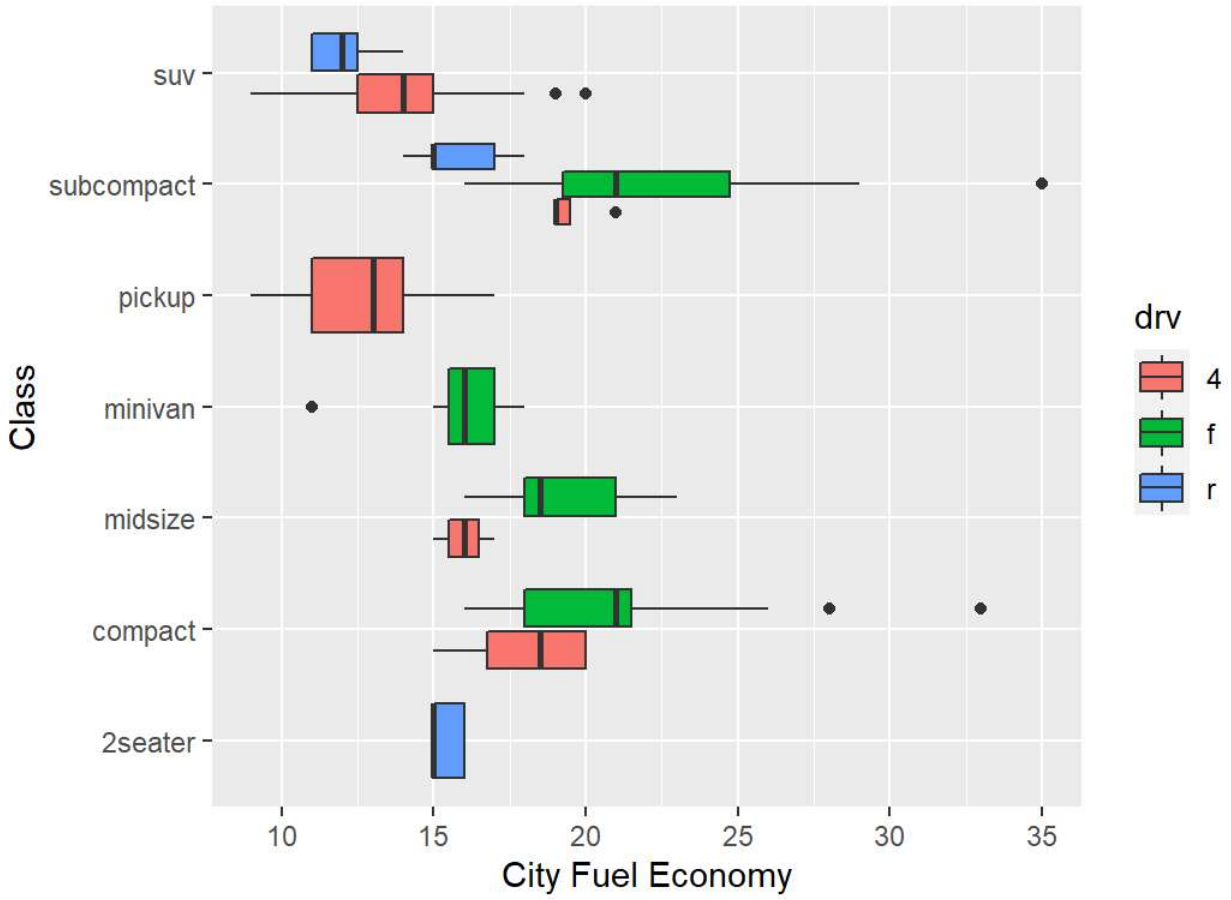
Problem 2: For this problem, we will continue working with the `mpg` dataset. Visualize the distribution of each car's city fuel economy by class and type of drive train with (i) boxplots and (ii) ridgelines. Make one plot per geom and do not use faceting. In both cases, put city mpg on the x axis and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

```
mpg
```

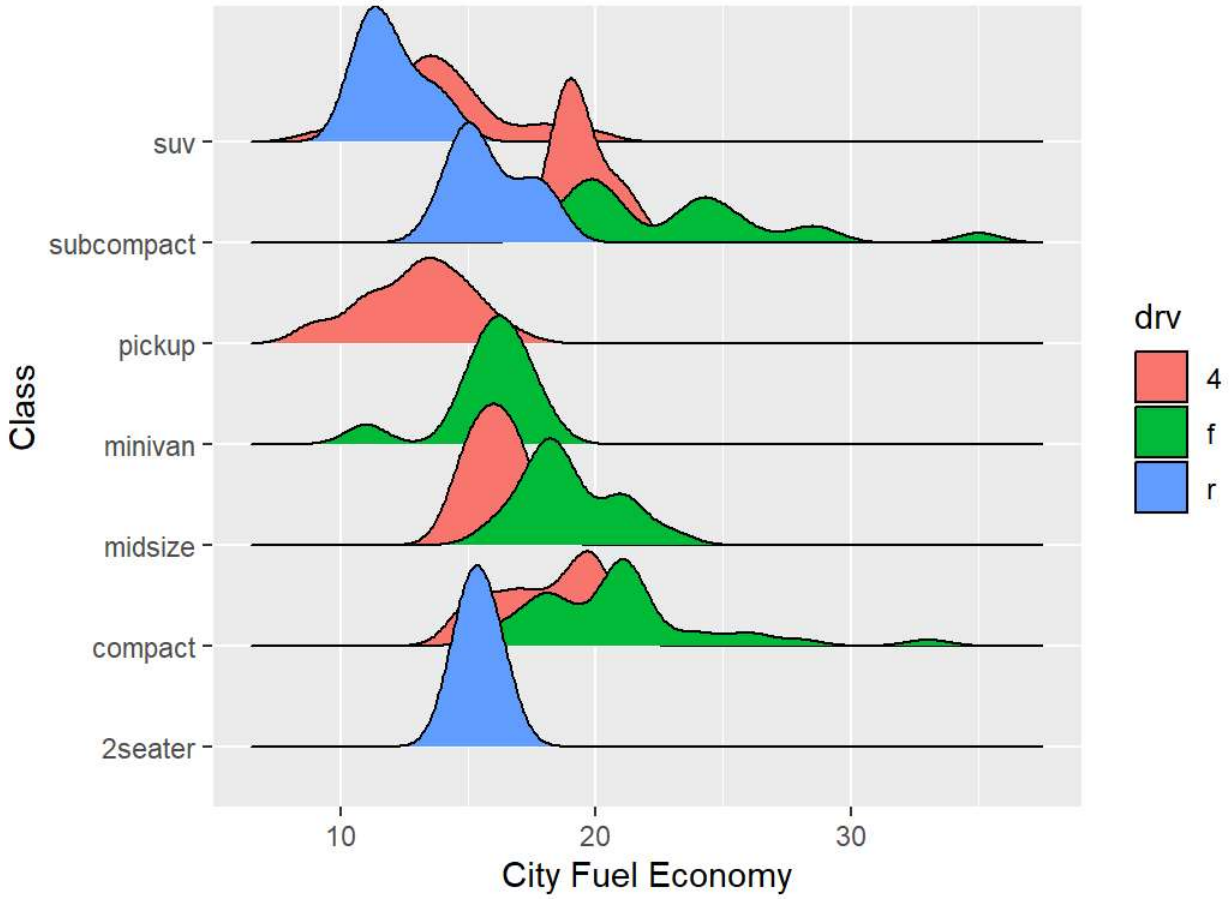
```
## # A tibble: 234 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    c
lass
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <
chr>
##  1 audi         a4           1.8  1999     4 auto~ f        18    29 p     c
omp~
##  2 audi       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

```
ggplot(mpg,aes(cty,class,fill=drv))+geom_boxplot()+ylab("Class")+xlab("City Fue
l Economy")
```




```
ggplot(mpg,aes(cty,class,fill=drv))+geom_density_ridges()+ylab("Class")+xlab("C
ity Fuel Economy")
```

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



Problem of the boxplot

1.The boxplot hides the multimodality and other features of distributions. For example, we only have 5 points for 2seater but we couldn't get the information from the boxplot.

2.The width of the boxplot for each class are different.For example, the boxplot for pickup is wider than that of SUV which is wider than that of subcompact.

3.Additionally, the boxplot is chaotic without reordering.