131-hw1

Ruoyu Li(9522913)

2022-10-01

Machine Learning Main Ideas

Question 1:

Define supervised and unsupervised learning. What are the difference(s) between them?

Supervised learning: Supervised learning is a type of machine learning that we have data about the actual observed outcome as well as the predictors. That is, we need to give the model both observed output and input, and we use the observed output to 'supervise' the performance of our models.

Unsupervised learning: Unsupervised learning uses machine learning algorithms that the models are only given predictors but not the actual observed response. Hence, the machine learning algorithms is without a 'supervisor' and we don't know what the answer key is. We can find hidden clustered pattern of the data using unsupervised learning.

Differences: The main difference between supervised and unsupervised learning is whether we have both the predictors and observed responses (observed input and output). In supervised learning we have both,and in unsupervised learning we only have the observed input.

Question 2:

Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

In the context of machine learning, the key difference between a regression model and a classification model is that for a regression model, the predicted outcome is quantitative (continuous numerical values) but for a classification model, the outcome is qualitative (categorical values).

Question 3:

Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

Commonly used metrics for Regression ML problems: (Training/Test) mean square error; mean absolute error, R squared.

Commonly used metrics for Classification ML problems:(Training/Test) error rate; Area Under the Curve (AUC) of The Receiver Operator Characteristic (ROC).

Question 4:

As discussed, statistical models can be used for different purposes. These purposes can generally be classified into the following three categories. Provide a brief description of each.

Descriptive models: Choose model to best visually emphasize a trend in data(i.e., using a line on a scatterplot). (From lecture Day1 pg39)

Inferential models: What features are significant? Aim is to test theories (Possibly) causal claims. State relationship between outcome & predictor(s). (From lecture Day1 pg39)

Predictive models: What combo of features fits best? Aim is to predict Y with minimum reducible error. Not focused on hypothesis tests. (From lecture Day1 pg39)

Question 5:

Predictive models are frequently used in machine learning, and they can usually be described as either mechanistic or empirically-driven. Answer the following questions.

Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

Mechanistic models are models that assume a parametric form for function f and it usually won't match true unknown f exactly. Mechanistic models aim to fit the observed inputs into the assumed parametric function f. Empirically-driven models don't give such assumptions about f; they require a large number of observations and are much more flexible by default. Empirically-driven models aim to find the best function f given the observed input.(Day1 pg38) The models are similar since they both want to fit a function f of the data in order to predict outcomes. The models differ that Mechanistic models make assumptions about f while Empirically-driven models don't.

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.

A mechanistic model is generally easier to understand because when we give an assumption of parametric function f, we can usually understand what f means since we're familiar with what we give. Also, the parametric function includes information about the relationship between our predictors. So, we don't need too many observations to determine the model. But empirically-driven models can develop a function f that is more flexible and complicated. It's harder to understand what f means and it takes a larger number of observations.

Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.

Mechanistic models should have a higher bias but lower variance because the parametric function won't match the unknown f exactly, which also means that it won't fit too exactly to the given data, leads to its lower variance. Empirically-driven models should have a lower bias but higher variance because the function f should fit quite well to the data, which results in the lower bias. But, this also means that f may not fit other data as well, leads to its higher variance.

Question 6:

A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:

Classify each question as either predictive or inferential. Explain your reasoning for each.

Given a voter's profile/data, how likely is it that they will vote in favor of the candidate?

This question is predictive, because in this question, we aim to predict the outcome (whethervote in favor of the candidate) given observed data (a voter's profile/data), we want to know what combo of features fits best.

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?

This question is inferential, because we want to know what features are significant here. We want to find the relationship between the outcome (a voter's likelihood of support for the candidate) and predictor (personal contact with the candidate).

Exploratory Data Analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(corrplot)
```

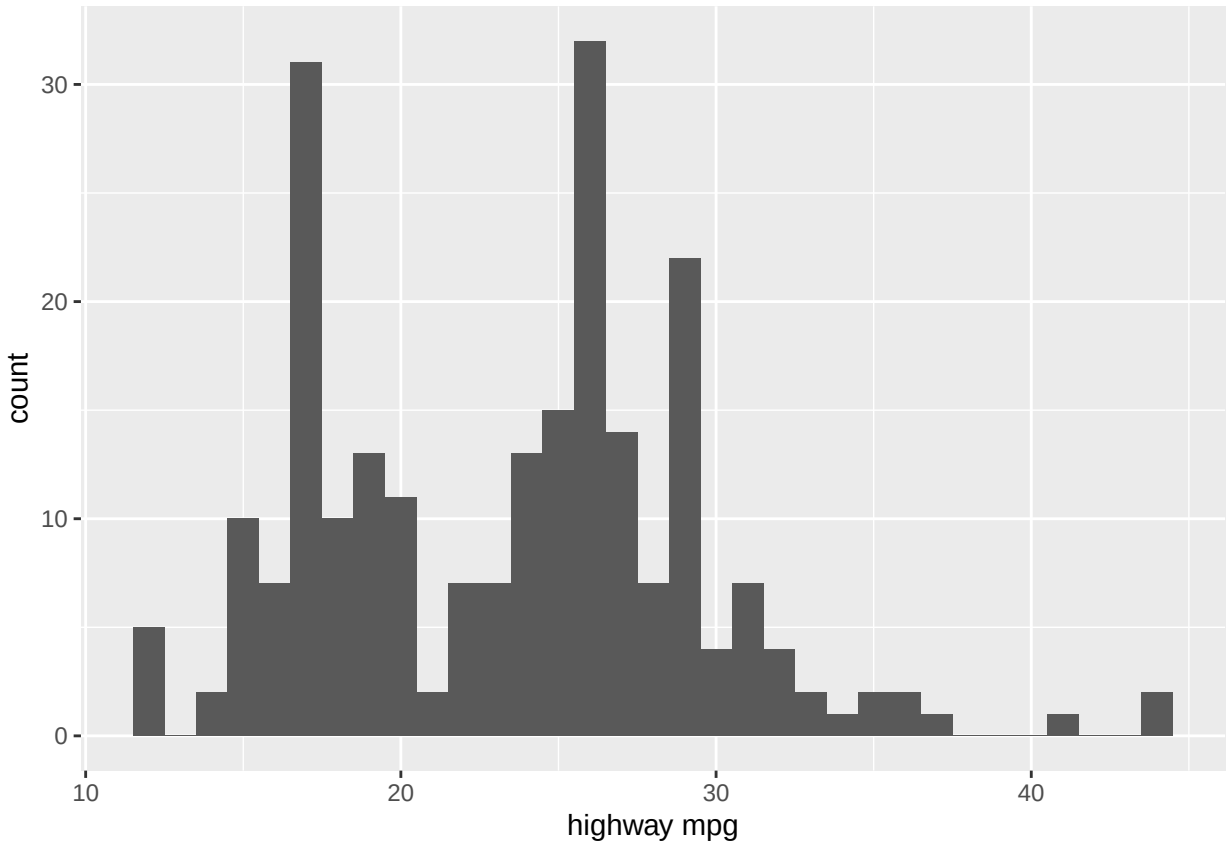
```
## corrplot 0.92 loaded
```

```
library(ggthemes)
```

Exercise 1:

We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.

```
ggplot(mpg, aes(x=hwy)) + geom_histogram(binwidth=1)+
  xlab('highway mpg')
```



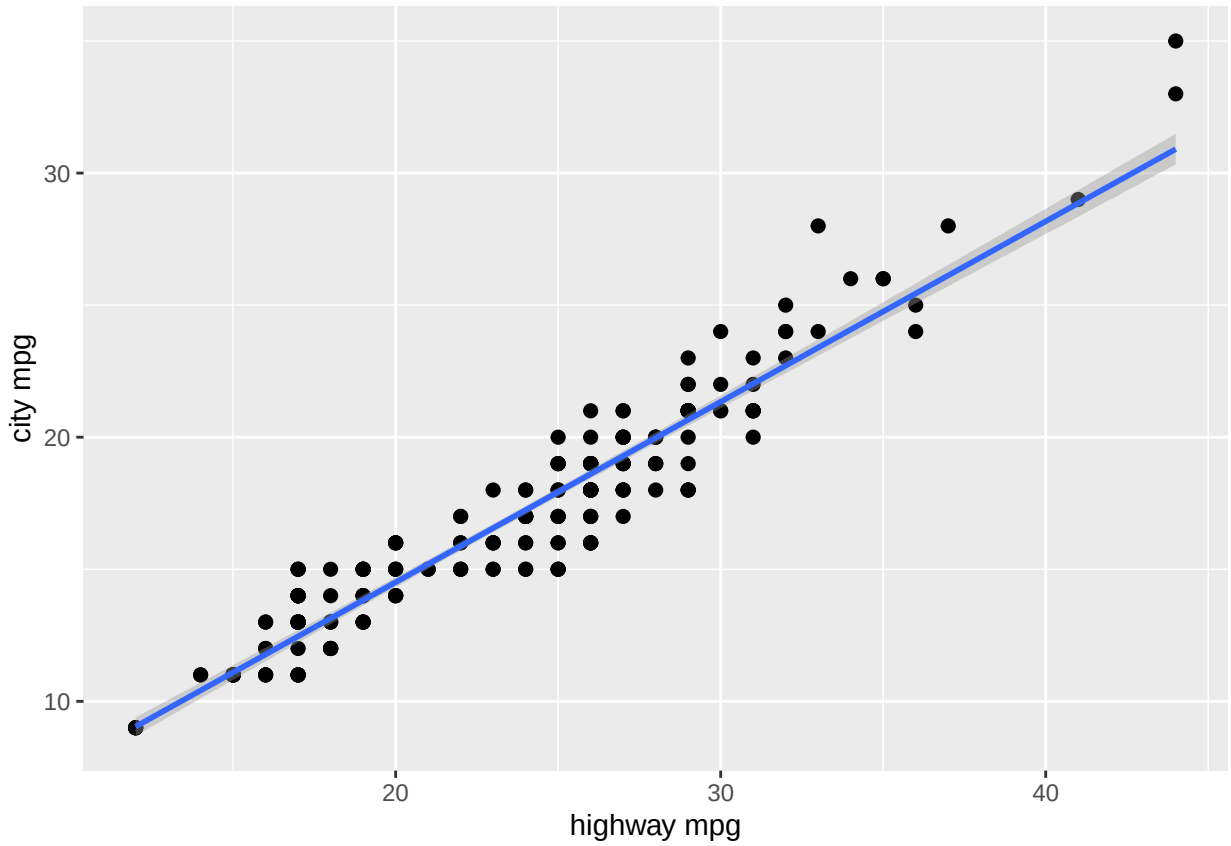
From this histogram, we can see that the distribution of variable hwy seems to be right skewed, with only very few observations over 40mpg but more oberservations on the left. Most of the observations fall in the range 15-35 highway mpg. Also, there are to notable peaks in this distribution, one around 17 mpg and one around 26 mpg.

Exercise 2:

Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?

```
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point(size=2) + geom_smooth(method=lm) +
  xlab('highway mpg') +
  ylab('city mpg')
```

```
## `geom_smooth()` using formula 'y ~ x'
```

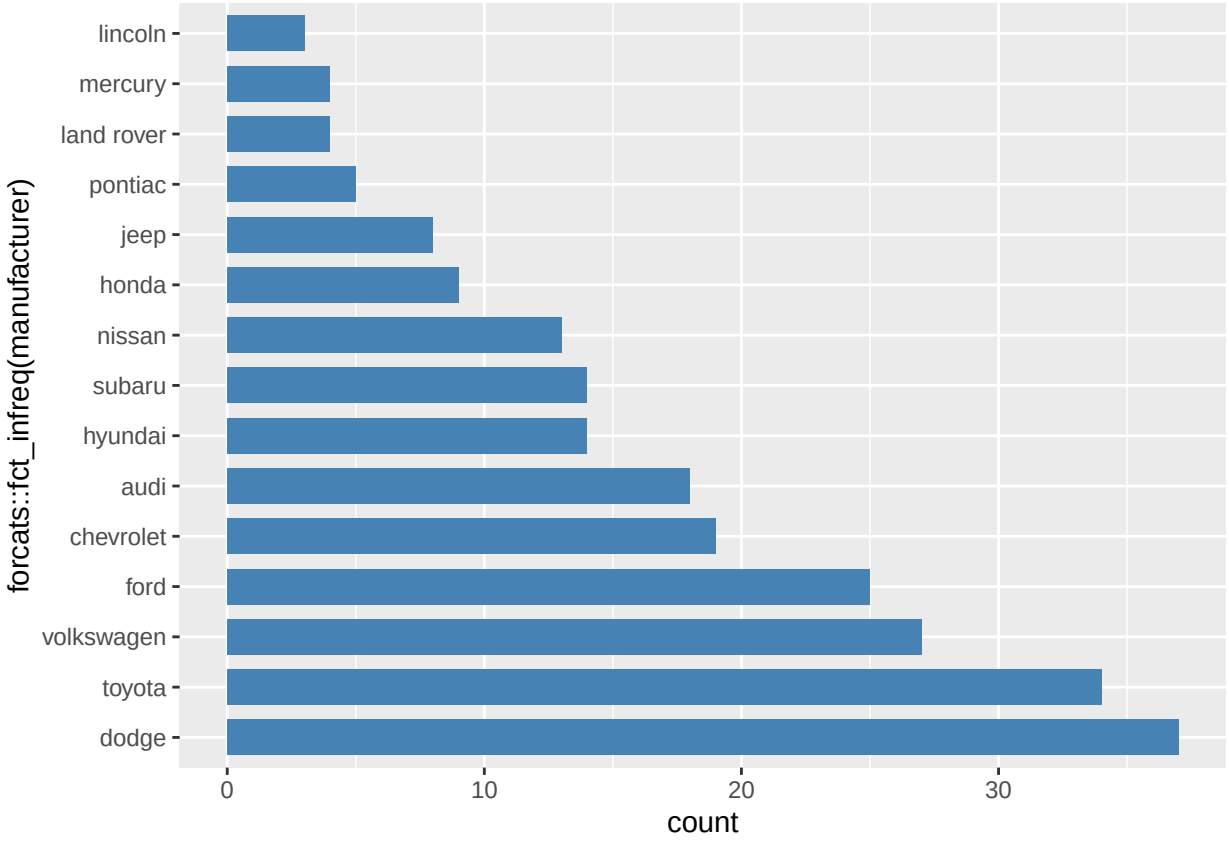


From the scatter plot, we find a positive correlation between the variable cty and the variable hwy. It's clear on the graph that there's a positive linear relationship between cty and hwy. This means that for each observation in this dataset, the higher hwy is, the higher cty would be, and vise versa. Also, all the observations seem to fall on a pretty uniform grid, and this makes sense since in the data set, cty and hwy are given by whole numbers only.

Exercise 3:

Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least?

```
ggplot(mpg, aes(x=forcats::fct_infreq(manufacturer))) +
  geom_bar(stat="count", width=0.7, fill="steelblue") +
  coord_flip()
```

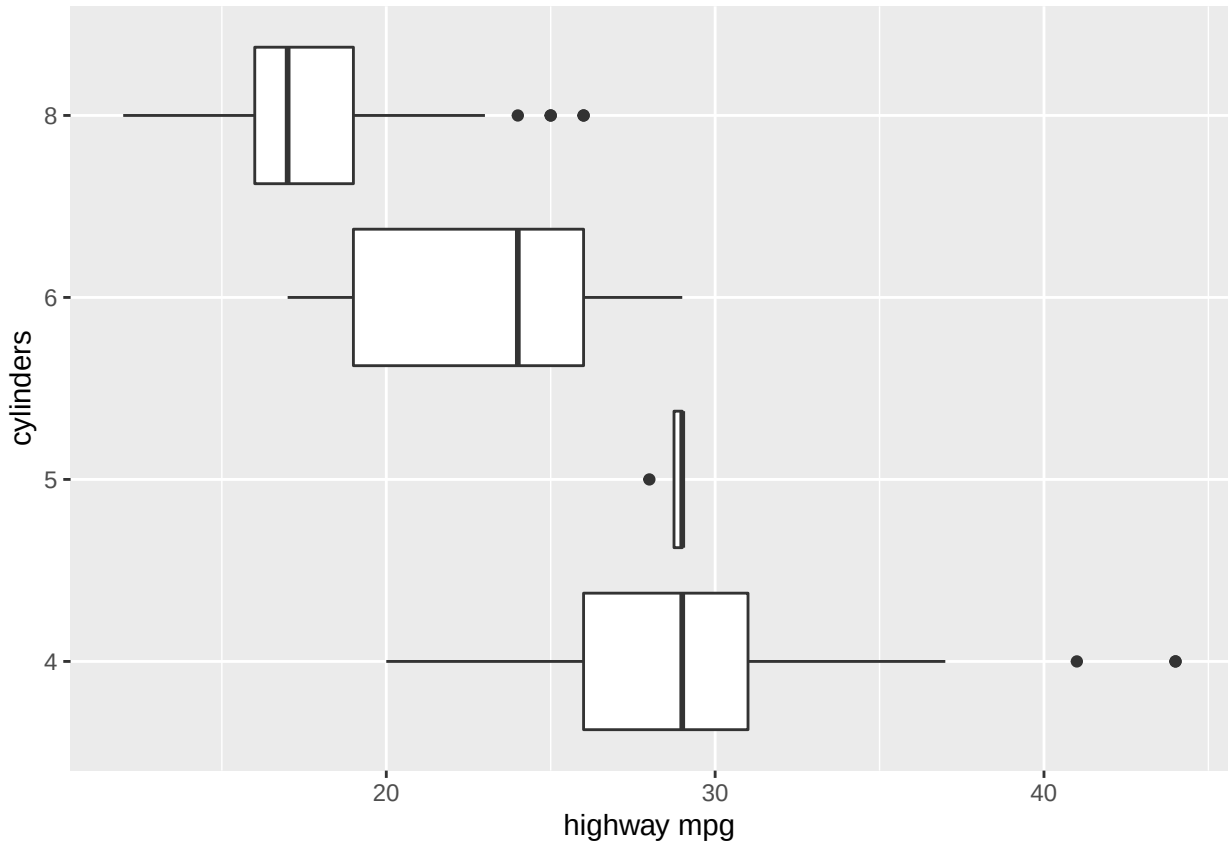


From this flipped bar plot, we can see that Dodge produces the most cars and Lincoln produces the least cars.

Exercise 4:

Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

```
ggplot(mpg,aes(x = hwy, y = factor(cyl))) + geom_boxplot() +
  xlab("highway mpg") +
  ylab("cylinders")
```

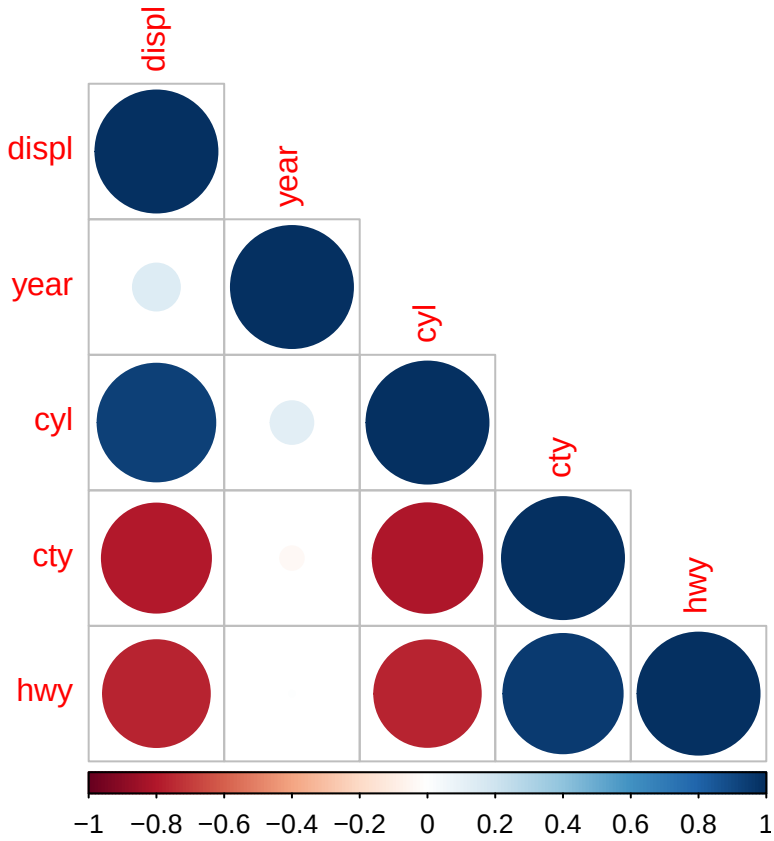


From this box plot, we can see a pattern that as the number of cylinders increases, the highway mpg tends to decrease.For cars with 4 cylinders and 8 cylinders, there seem to be some outliers with extraordinarily high highway mpg. Also, the data for cars with 5 cylinders seems to be quite concentrated, this might be result of the fact that there were only very few cars with 5 cylinders out there.

Exercise 5:

Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset. Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

```
M <-mpg[ , purrr::map_lgl(mpg, is.numeric)]
M1 = cor(M)
corrplot(M1, type = 'lower')
```



From the correlation matrix, we can find that cylinders and displacement, highway mpg and city mpg are strongly POSITIVELY correlated. These relationships makes sense, since usually cars with more cylinders also have larger displacement. Also, a car's fuel economy performance in city and highway should be positively correlated since a fuel efficient car in city should also perform great efficiency on highways.

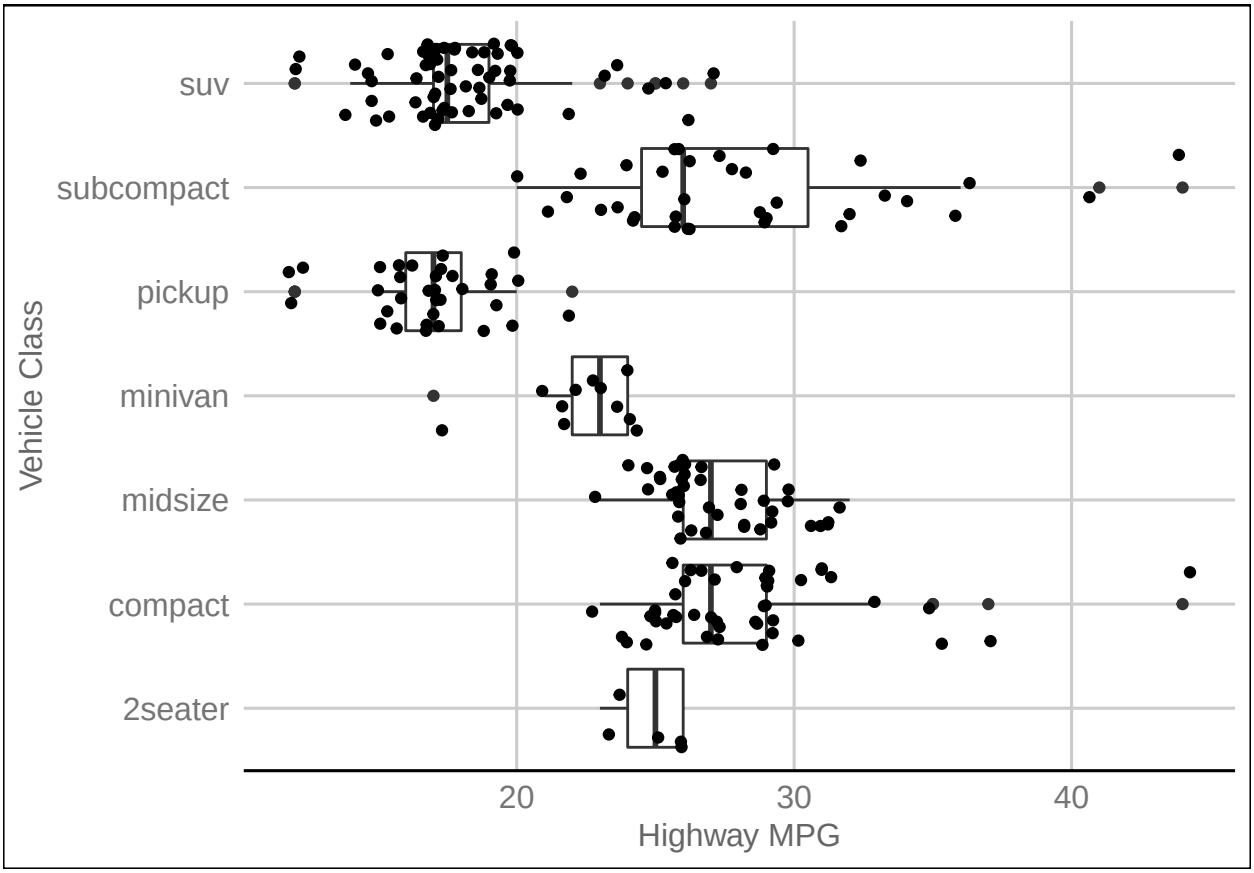
Then, we can see that city/highway mpg and displacement, city/highway mpg and cylinders are strongly NEGATIVELY correlated. This makes sense since firstly, we already found out that cylinders and displacement are positively correlated, and it's a common sense that cars with more cylinders and larger displacement usually have a lower fuel efficiency, which is also a lower city and highway mpg.

I am kind of surprise that there's not a strong correlation between city/highway mpg and year since I thought that older cars have worse fuel efficiencies. But when I look back at the data, all the cars in this dataset were manufactured in either 1999 or 2008, it could be possible that the car industry didn't improve their fuel efficiencies between these years.

231 Only:

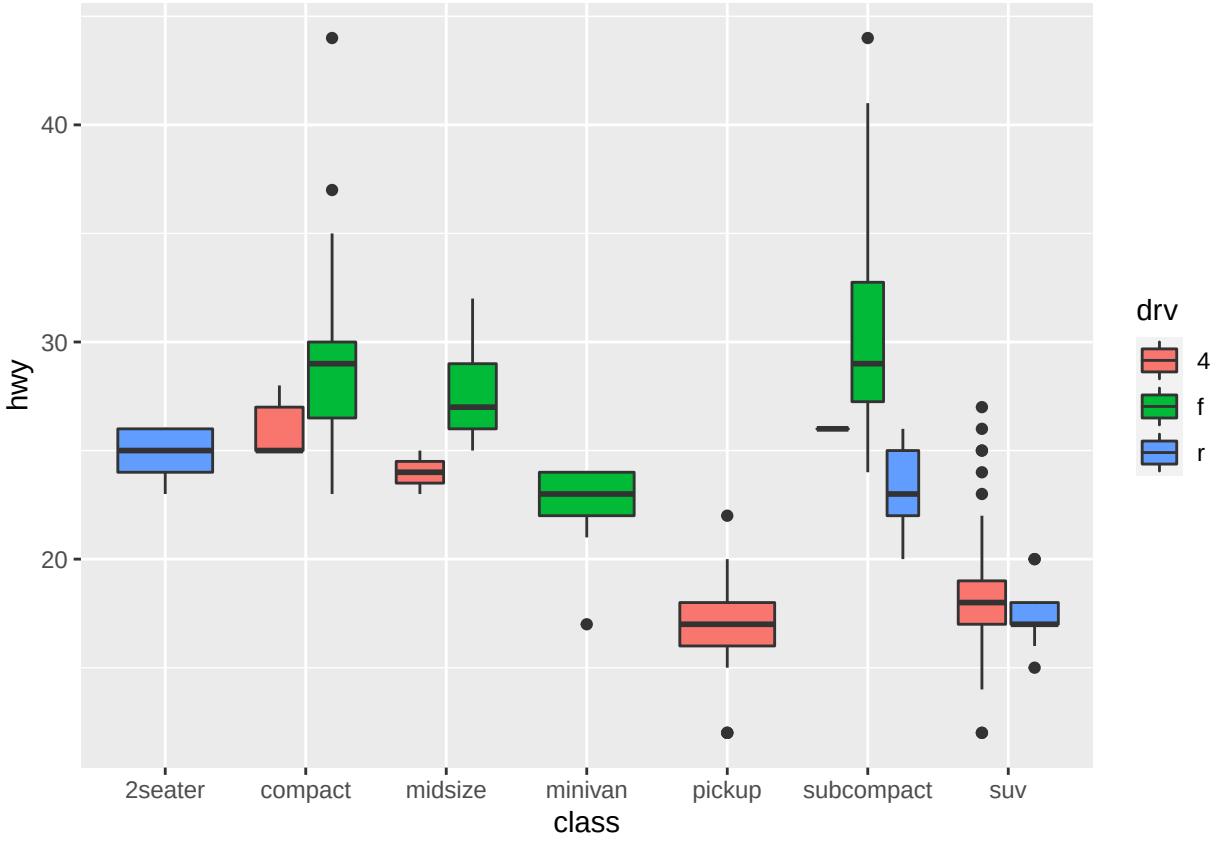
Exercise 6

```
ggplot(mpg,aes(x = hwy, y = class))+
  geom_boxplot() +
  geom_jitter() +
  theme_gdocs() +
  xlab('Highway MPG') +
  ylab('Vehicle Class')
```



Exercise 7

```
ggplot(mpg,aes(x = class, y = hwy, fill = drv)) +
  geom_boxplot()
```

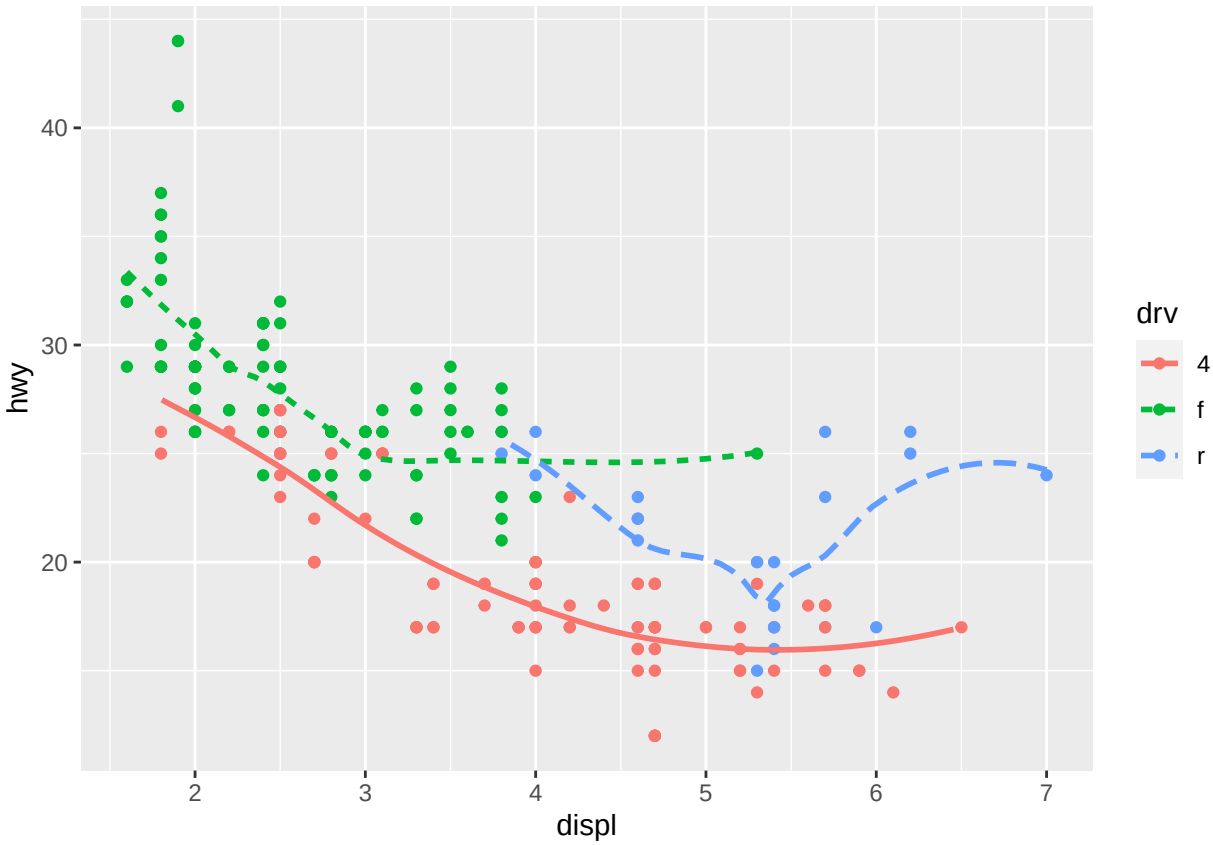


Exercise 8

```
ggplot(mpg, aes(x=displ, y=hwy, color=drv, linetype=drv)) +
  geom_point() +
  geom_smooth(method="auto", se=FALSE, fullrange=TRUE) #idk how to make all the lines blue here
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 69 rows containing missing values (geom_smooth).
```


hwy
40
30
20
2
3
4
5
6
7
displ
drv
4
f
r