HW1

shuhei kaneko

2022-09-24

Machine Learning Basic Ideas

Question 1:

In supervised learning, we have the information of outcome variable (Y_i) and the predictor ($\mathbf{X_i}$). The goal is to learn the function f such that $Y_i = f(\mathbf{X_i}) + \epsilon_i$. On the other hand, in unsupervised learning, we do **not** have the information of outcome (Y_i). Therefore, our goal in unsupervised learning is to generate plausible tendency of the dataset ($\mathbf{X_i}$) by Principal component analysis, clustering or some other technique.

Question 2:

In regression model, the outcome is numerical value (e.g. price, height, weight, blood pressure). On the other hand, in classification model, our outcome is categorical variable (e.g. Live/Dead, Red/Blue/Green).

Question 3:

For regression ML: Linear regression, Kernel regression

For classification ML: Logistic regression, Decision tree

Question 4:

Descriptive models: Describe the trends of the existing data. Choose model to best emphasize a trend in data.

Inferential models: Test which predictor (explanatory variable) is significantly related to the outcome possibly using asymptotic statistical theory. It might be possible to argue the causality depending on the assumption and the data structure.

Predictive models: Build the model that predict the future or unobserved outcome with minimum reducible error. Significance of each predictor is not important if our objective is to purely predict the outcome.

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?

Answer: Mechanistic model: We assume the parametric form of explanatory variable beforehand/ Empirically-driven model: we do not impose any assumption on f. They are similar in that both models aim to predict Y with minimum reducible error. They are different in interpretability and flexibility. In general, mechanistic model is easier to interpret (see the next question for detail) but less flexible because predicted Y is based on our assumption on the functional form of $\mathbf{X}$.

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

Answer: Mechanistic model is easier to understand. Suppose we would like to predict wage based on years of education. If we assume that these two are related linearly, it is easy to argue that "1 year increase in education will increase wage by XX dollars on average". On the other hand, in nonparametric model, it is impossible to summarize the relationship by the simple words.

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

Answer:In general, in mechanistic model, we have low variance and high bias. On the other hand, in empirically-driven model, we have high variance but low bias. The performance of the model is measured the sum of bias and variance. Therefore, too simple/flexible model is not a good idea.

Question 6:

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

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

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

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

Answer: First one: predictive, Second one: inferential. In the first one, our objective is to **predict** whether the voter will vote in favor of the candidate based on the characteristics ($\mathbf{X_i}$). In the second one, our objective is to test the significance of the personal contact in terms of the voter's likelihood of support for the candidate.

Exploratory Data Analysis

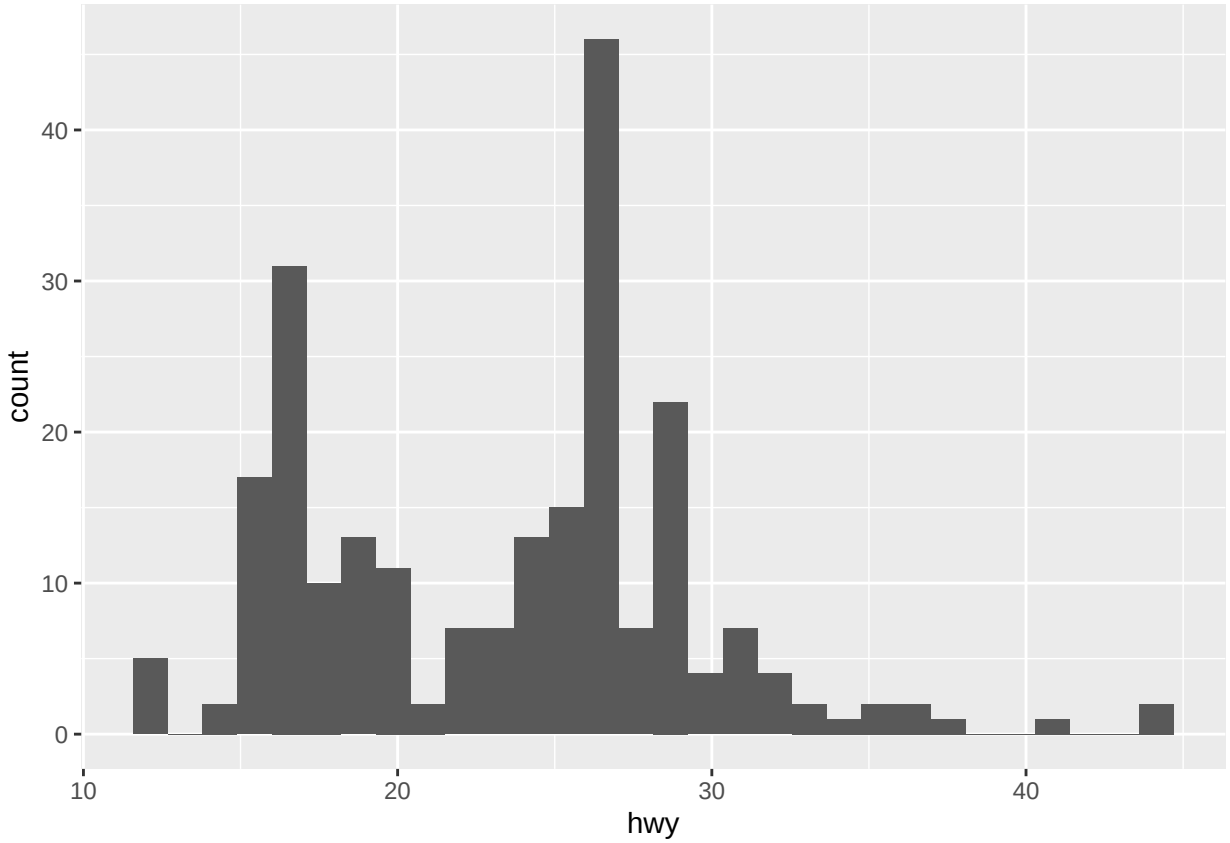
Exercise 1

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## 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
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
ex1 <- mpg %>%
  ggplot() +
  geom_histogram(aes(x = hwy))
plot(ex1)
```

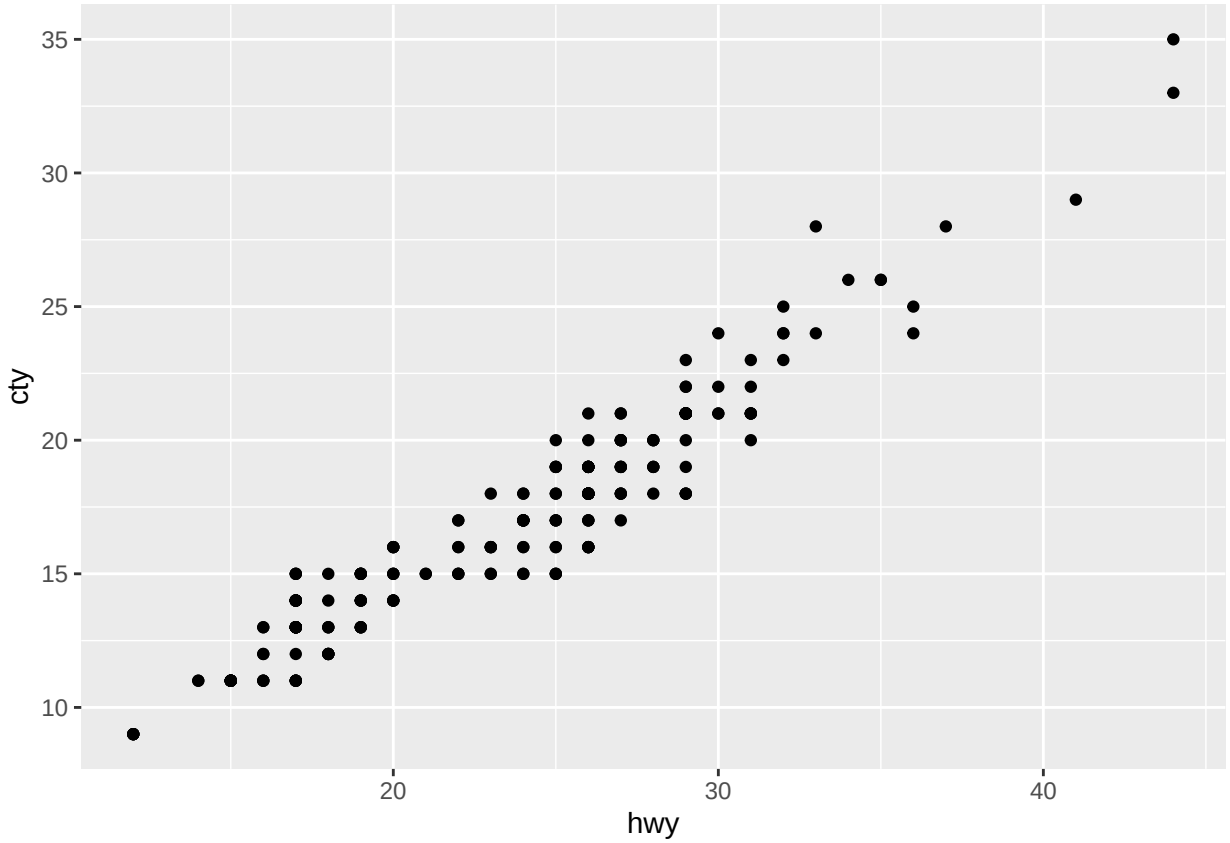
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



hwy variable has bimodal distribution: we have many observations around 15 and 25. Range is about 33 (max: about 45, min: about 12).

Exercise 2

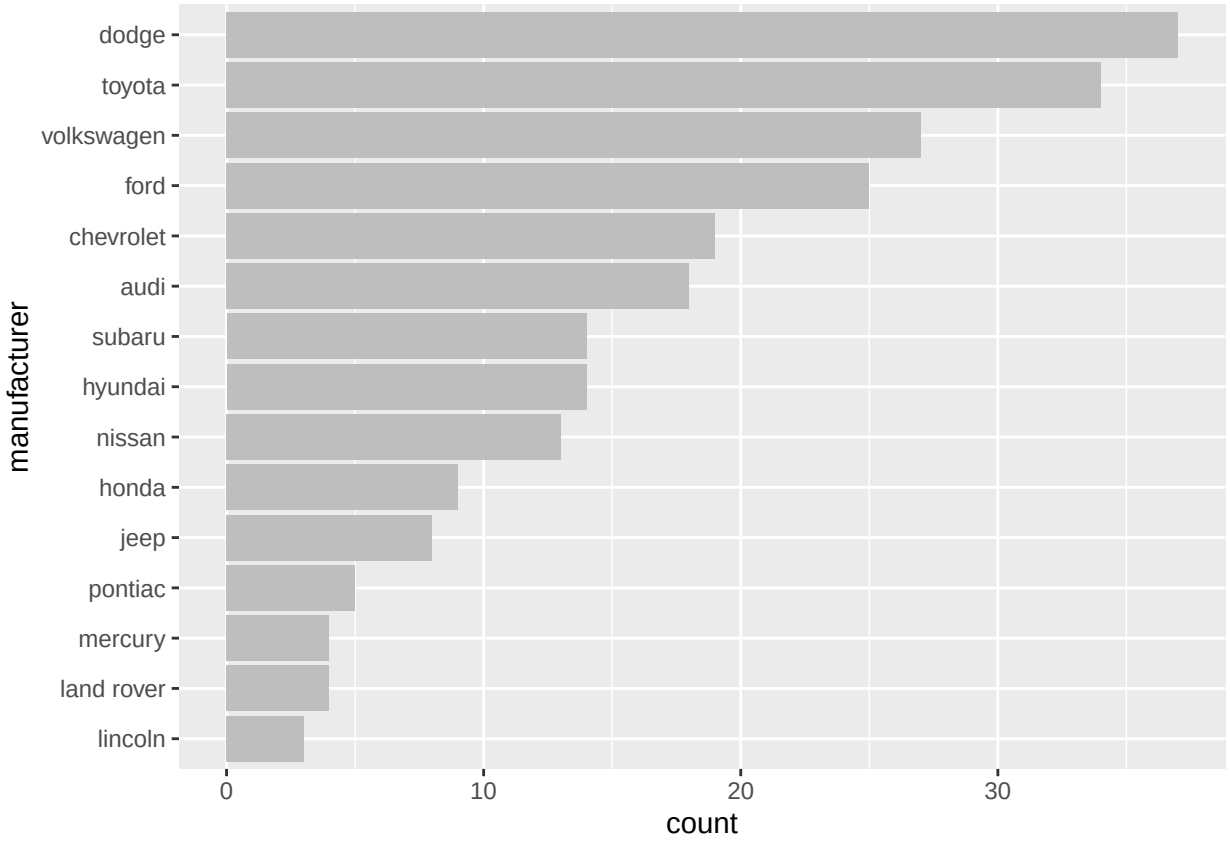
```
ex2 <- mpg %>%
  ggplot() +
  geom_point(aes(x= hwy, y=cty))
plot(ex2)
```



We can see strong positive relationship between hwy and cty. The relationship seems to be almost linear.

Exercise 3

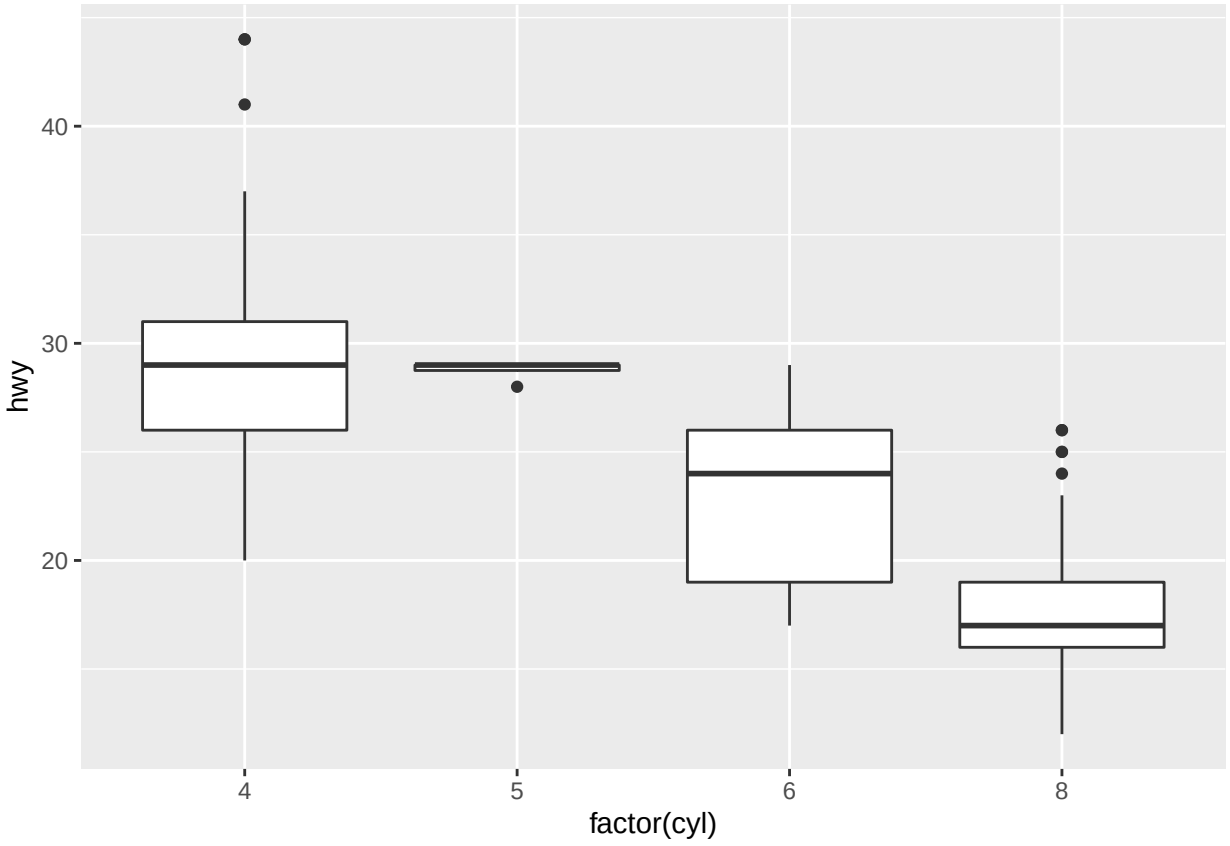
```
# The table() command aggregates the data by the indicated variable
tab_manu <- table(mpg$manufacturer)
df_manu <- as.data.frame(tab_manu)
# Rename the columns for readability.
colnames(df_manu) <- c("manufacturer", "count")

df_manu <- transform(df_manu, manufacturer=reorder(manufacturer, count))

ex3 <- ggplot(df_manu) +
       geom_bar(aes(x=manufacturer,y=count),stat="identity",fill="grey") +
       coord_flip()
plot(ex3)
```



Most: Dodge, Least: Lincoln

Exercise 4

```
ex4 <- mpg %>%
  ggplot(aes(x=factor(cyl), y=hwy)) +
    geom_boxplot()
plot(ex4)
```



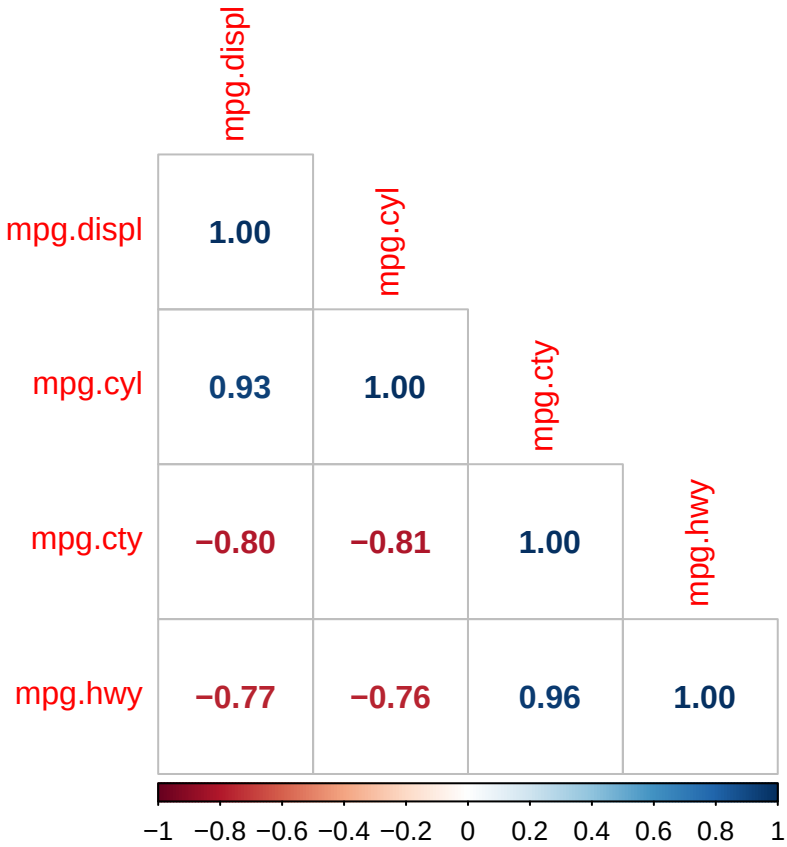
More number of cyl is negatively associated with hwy.

Exercise 5

```
#install.package("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

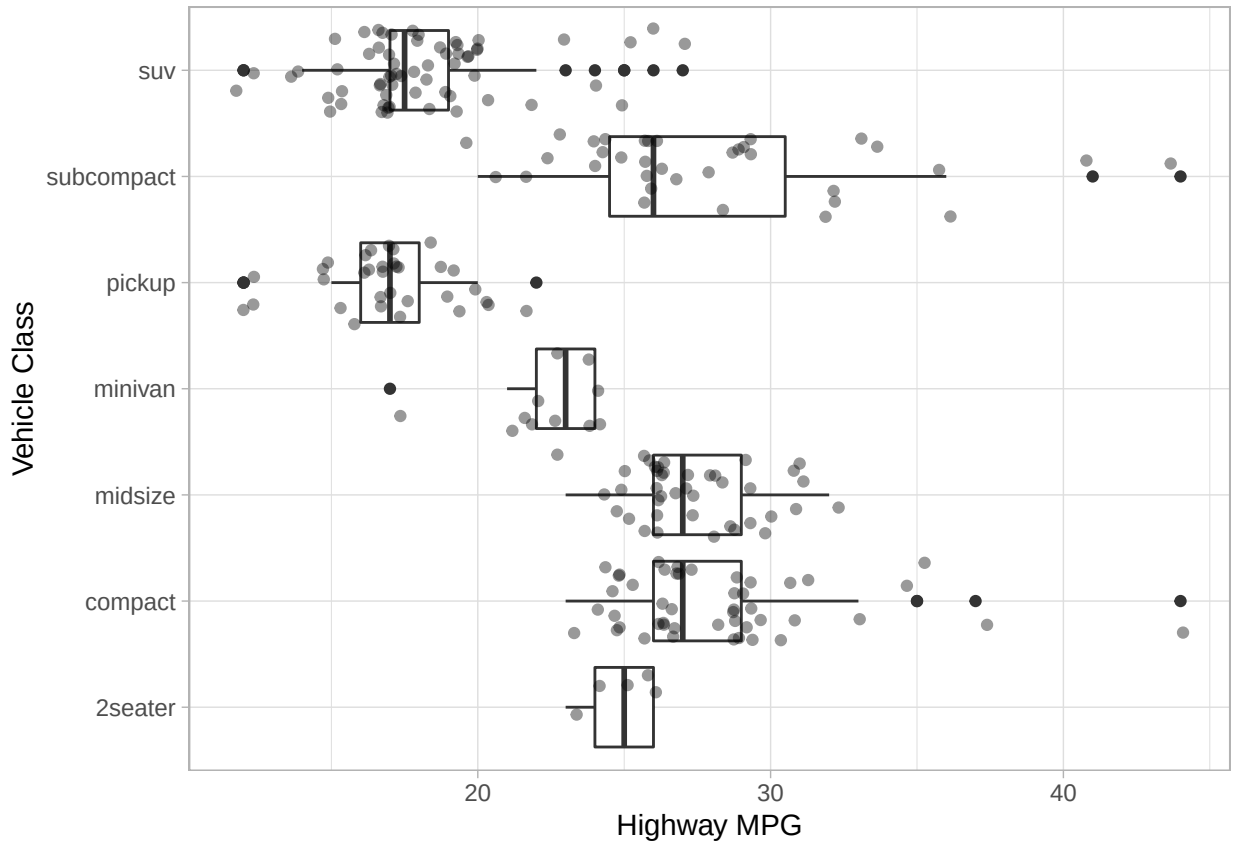
```
mpg_2 <- data.frame(mpg$displ, mpg$cyl, mpg$cty, mpg$hwy)
M <- cor(mpg_2)
corrplot(M, method = 'number', type = "lower")
```



City miles per gallon and highway miles per gallon have very high positive correlation, which seems to be very natural. Number of cylinder has negative correlation with cty and hty. This is because cars with many number of cylinder are more likely to be luxury car or super car. These types of car do not have a good fuel efficiency. The similar can be said about the displacement level. It is very natural that the cars with high displacement does not have good fuel efficiency, and also should be positively correlated with the number of cylinders.

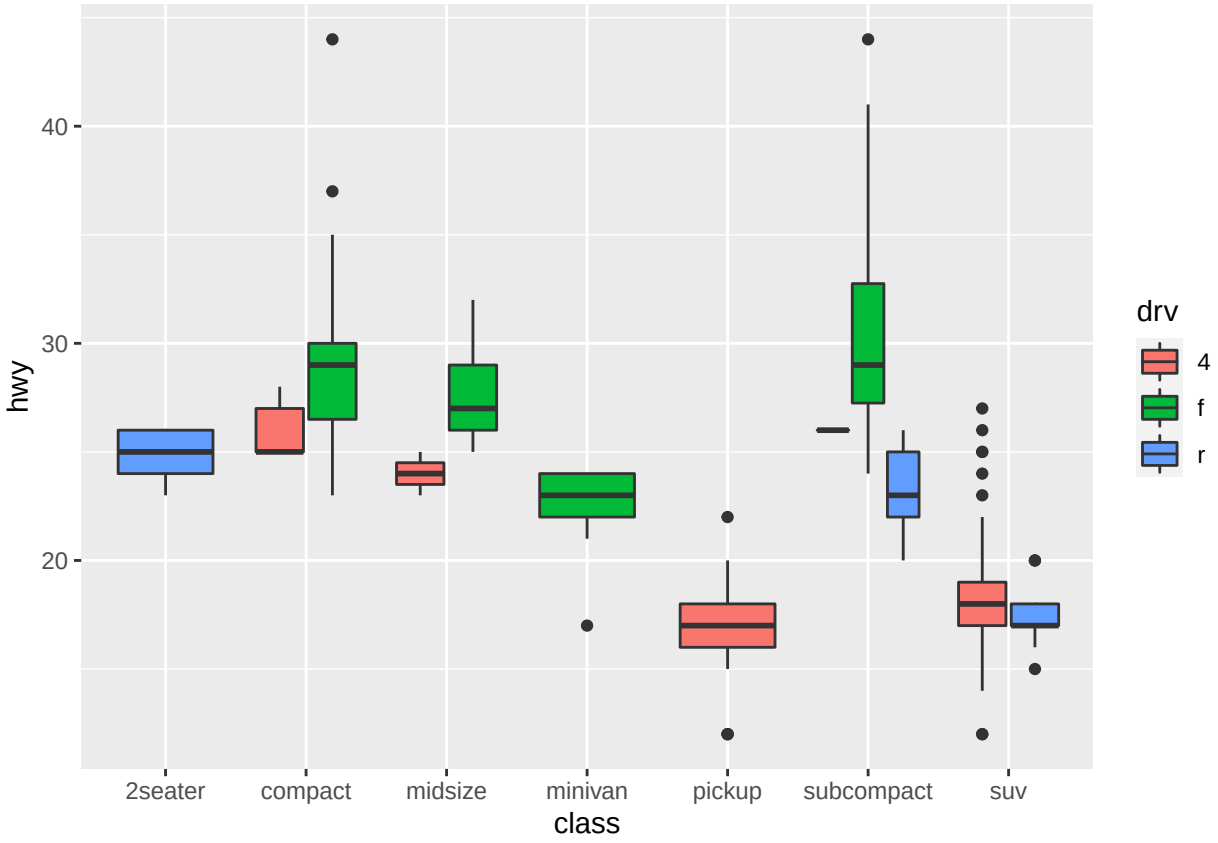
Exercise 6

```
#install.packages("ggthemes")
library(ggthemes)
ex6 <- mpg %>%
  ggplot(aes(x=hwy,y=class))+
  theme_light() +
  geom_boxplot()+
  geom_jitter(alpha = 0.4) +
  xlab("Highway MPG") +
  ylab("Vehicle Class")

plot(ex6)
```



Exercise 7

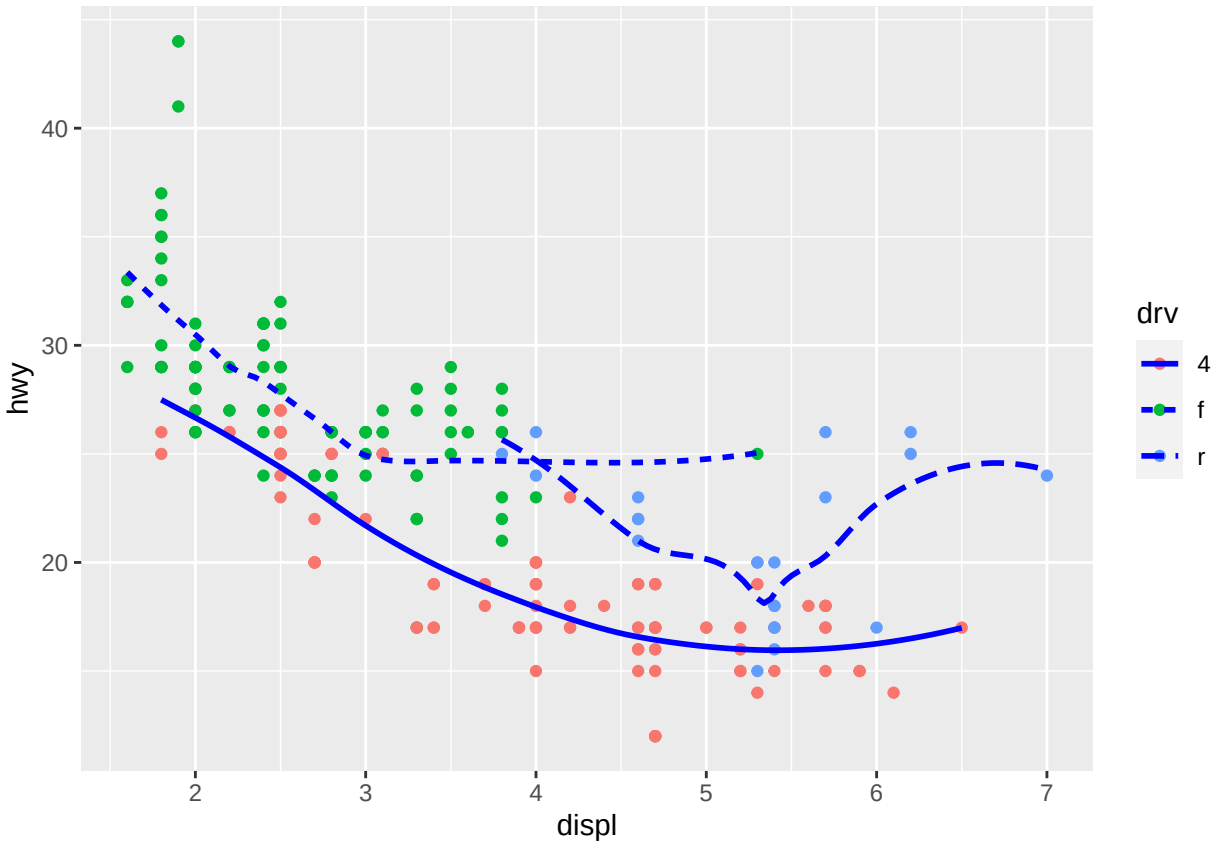
```
ex7 <- mpg %>%
  ggplot(aes(x=class, y=hwy, fill = drv)) +
    geom_boxplot()
plot(ex7)
```



Exercise 8

```
ex8 <- mpg %>%
  ggplot(aes(x = displ, y = hwy, group = drv, color = drv)) +
  geom_point() +
  geom_smooth(aes(linetype = drv), se=FALSE, method="loess", color = "Blue")

plot(ex8)
```

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


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