Pstat131HW1

Wentao Yu

2022-09-23

Machine Learning Main Ideas Question1:

The supervised learning is using the labeled data sets, and the unsupervised learning is using algorithm to clustering and analyze the unlabeled data sets. The main difference is the use of labeled data sets.

Question2:

According to the lecture slides no.34, 'day_1_131_231', the regression is the method where the response Y is quantitative. The classification is the method where the response Y is qualitative. That means the outputs of regression are numerical values, and the outputs of classification are categorical values.

Question3:

- Regression: Mean Square Error(MSE), Root Mean Square Error(RMSE), Mean Absolute Error(MAE)
- Classification: Accuracy, confusion matrix, AUC-ROC

Question4:

According to lecture sides no.39, 'day_1_131_231',

- Descriptive models:Choose model to best visually emphasize a trend in data.
- Inferential models:
    1. Aim is to test theories.
    2. (Possibly) casual claims.
    3. State relationship between outcomes & predictors.
- Predictive models:
    1. Aim is to predict Y with minimum reducible error.
    2. Not focused on hypothesis tests.

Question5:

- 1. Mechanistic: use a theory to predict what will happen in the real world.
  2. Empirically-driven: study real-world events to develop a theory.
  3. Differences: mechanistic assumes a parametric form for f, and it won't match true unknown f. However, empirical-driven has no assumptions aboout f, and it requires a large number of observations.
  4. Similarities: mechanistic can add more parameters to become more flexible, and empirical-driven is flexible by default. Also, both of them are overfitting.
- I think mechanistic seems to be easier to understand. As the mechanistic is given the theory to explore the real world, then we could try to find out the differences between real world events and theory. However, empirically-driven asks us to develop a theory by our own through real-world events, which somehow requires creativity.
- bias-variance tradeoff is a property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters. As both of mechanistic and

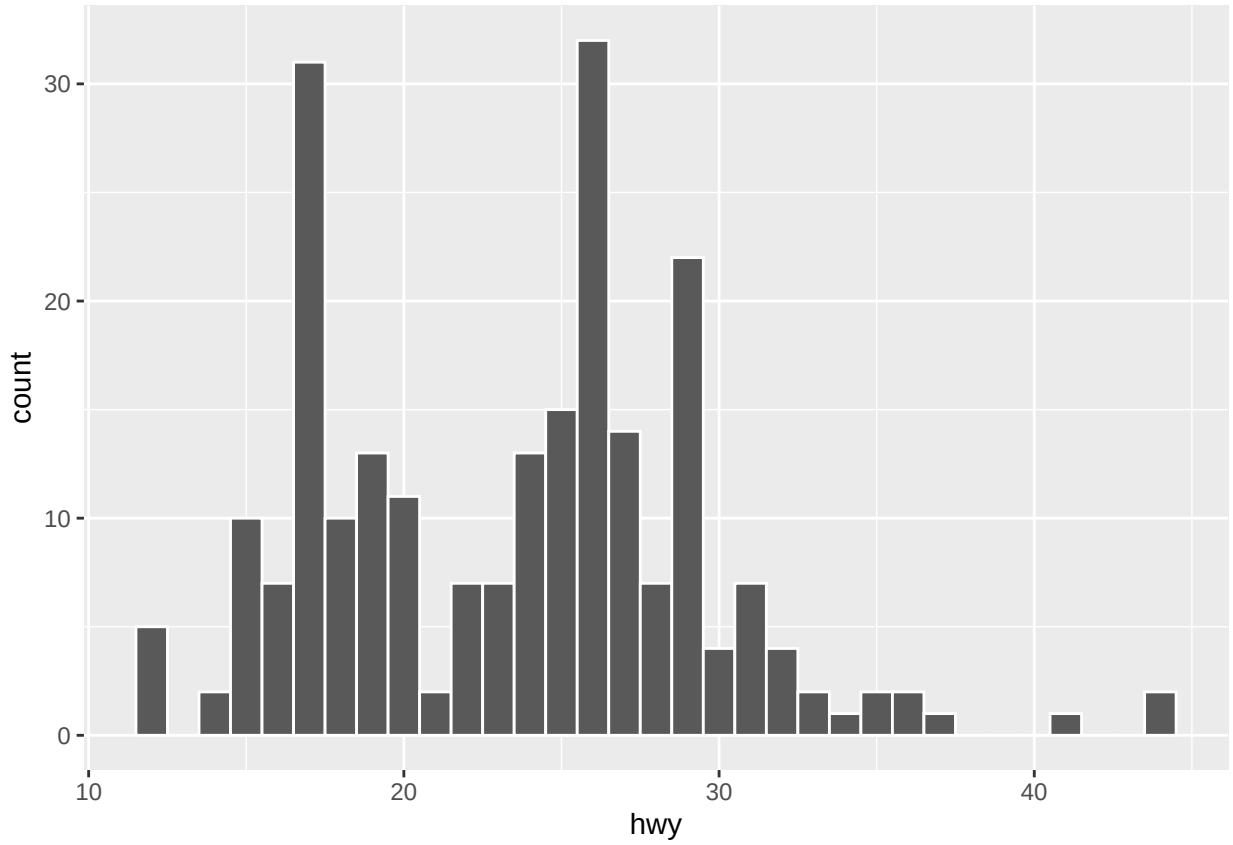
empirical-driven are overfitting which means they have high variance, then bias-variance tradeoff can help them to reduce the variance of parameter by increasing the bias.

Question6:

- This is a predictive model. As voters' files are given, we need to predict the voter's favor of candidates. This is a predictive model.
- This is an inferential model. Since it asks to find the change of voters' support if they have personal contact with the candidates, then it is trying to find the relationship between outcomes and predictors. Thus, this is an inferential model.

Exploratory Data Analysis Exercise1:

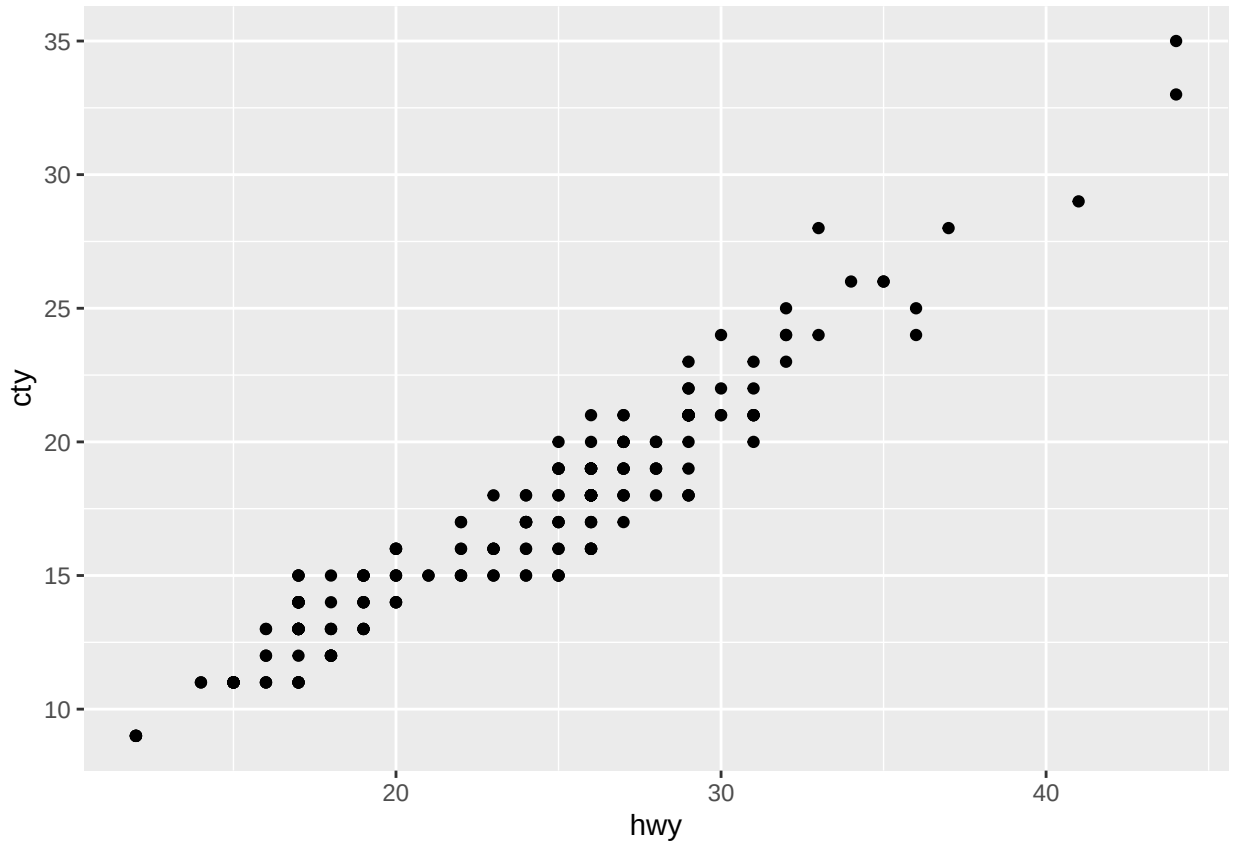
```
ggplot(mpg,aes(hwy))+geom_histogram(color = "white", binwidth = 1)
```



Explanation: I can see that there are two columns that are extremely higher than other. In addition, the main part of the data is between 14 - 37.

Exercise2:

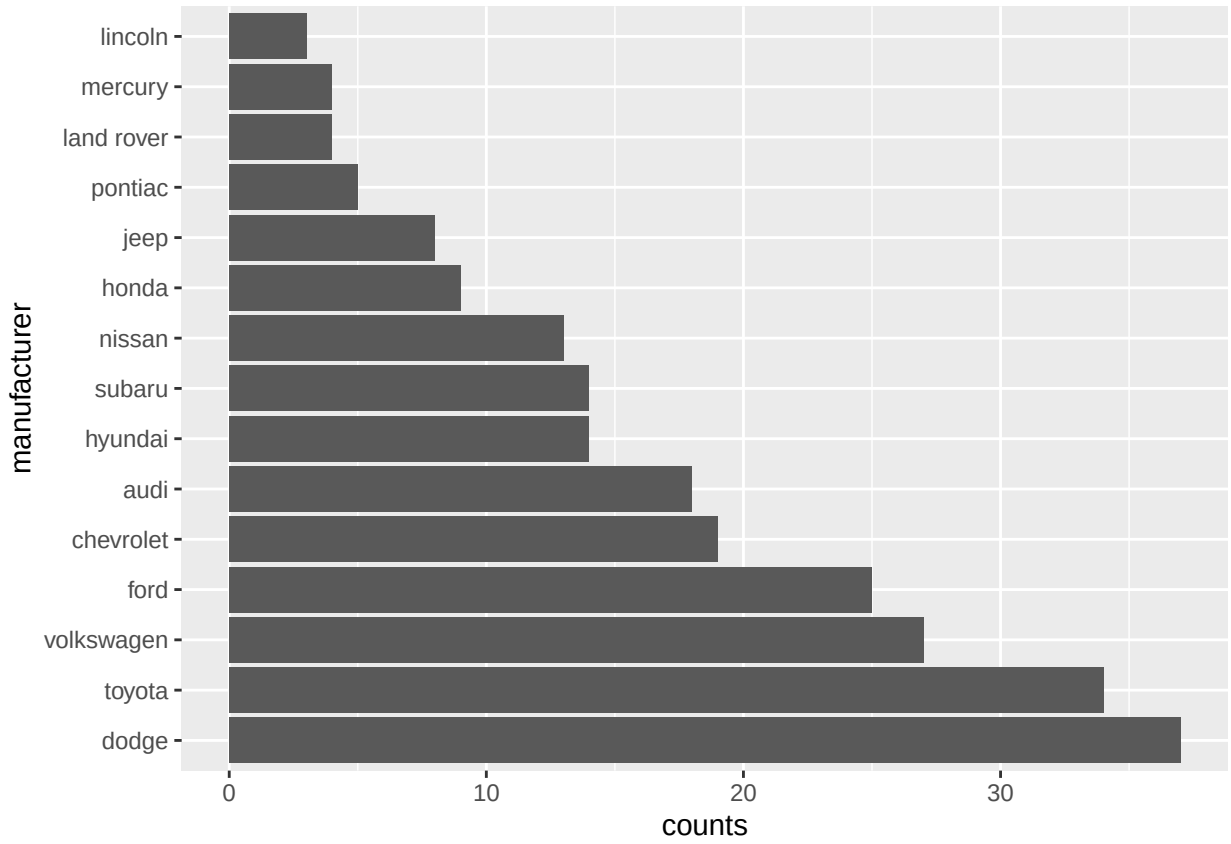
```
ggplot(mpg,aes(hwy,cty))+geom_point()
```



Explanation: I can see that there seems a positive linear relationship between hwy and cty. That represents the higher average highway MPG of the car, the higher average city MPG of the car also.

Exercise3:

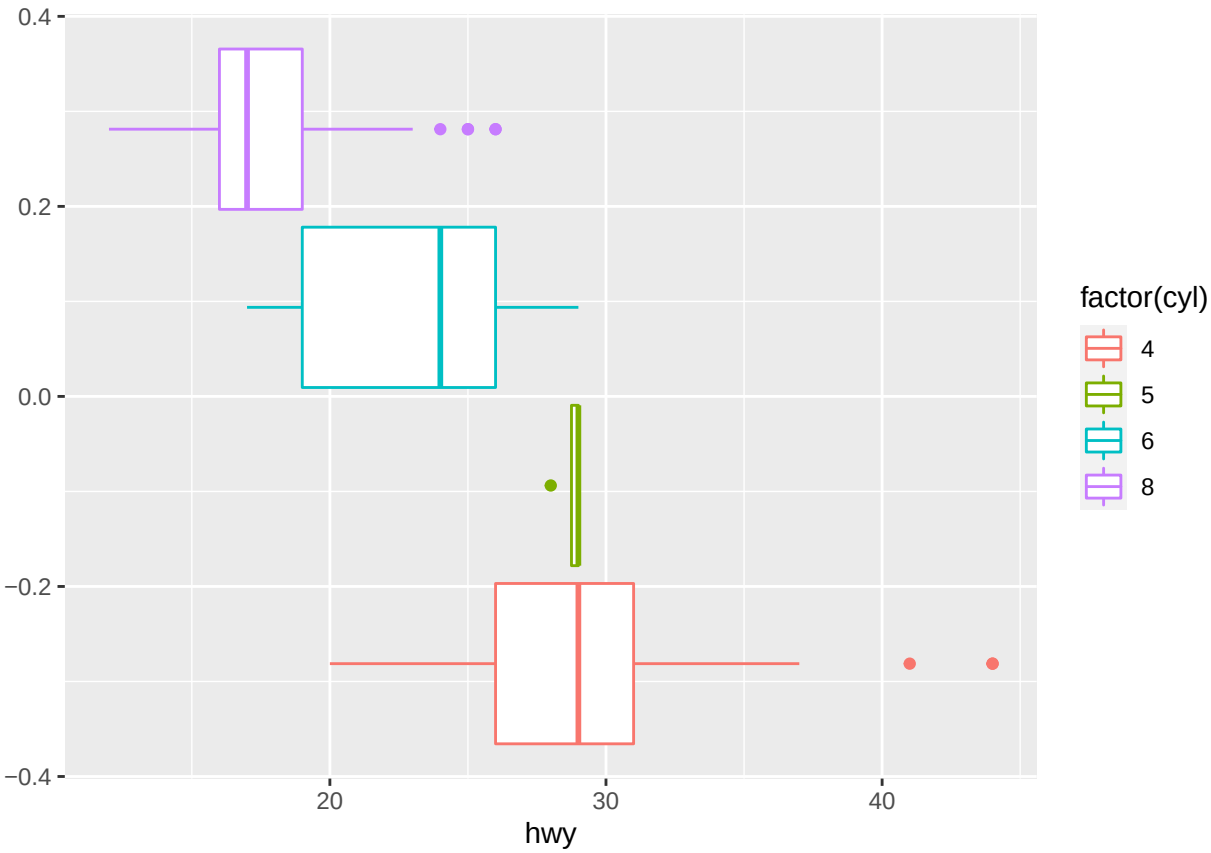
```
ggplot(mpg,aes(x = reorder(manufacturer,manufacturer, function(x)-length(x))))+geom_bar()+coord_flip()+
  labs(x = 'manufacturer', y = 'counts')
```



Explanation: Dodge produces the most cars, and Lincoln produces the least cars.

Exercise4:

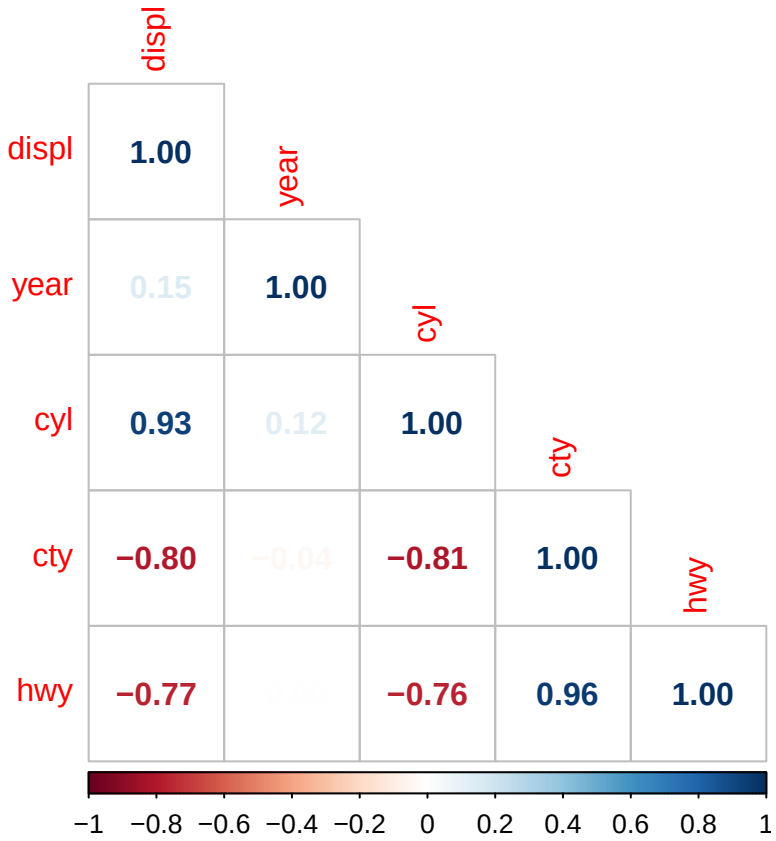
```
ggplot(mpg,aes(hwy, color= factor(cyl))) + geom_boxplot()
```



Explanation: I can see that cars with less cylinders(cyl) tend to have higher highway mpg(hwy).

Exercise5:

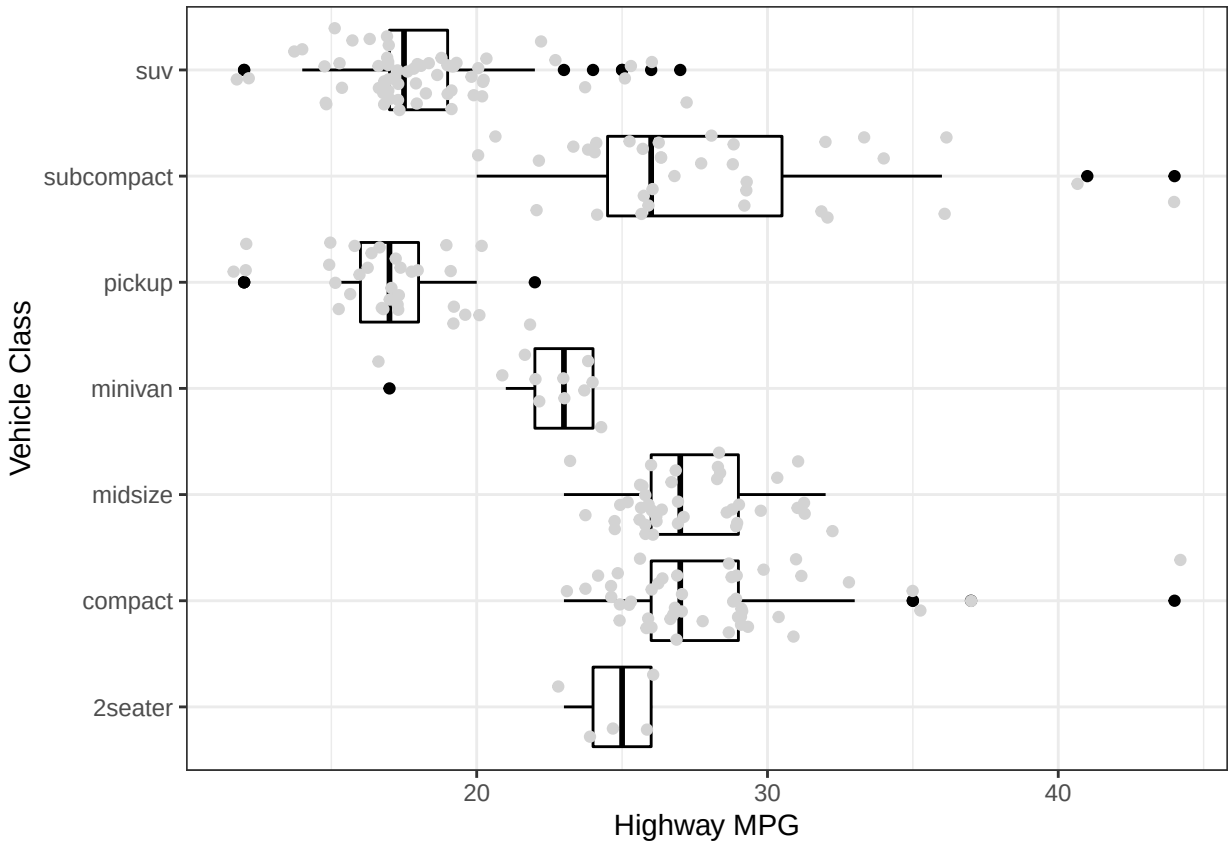
```
numeric_data <- select_if(mpg, is.numeric)
corrplot(cor(numeric_data),method = 'number', type = 'lower')
```



Explanation: hwy and cty have positive correlation. cyl and displ have positive correlation. cty has negative correlations with both displ and cyl. hwy has negative correlations with both displ and cyl.

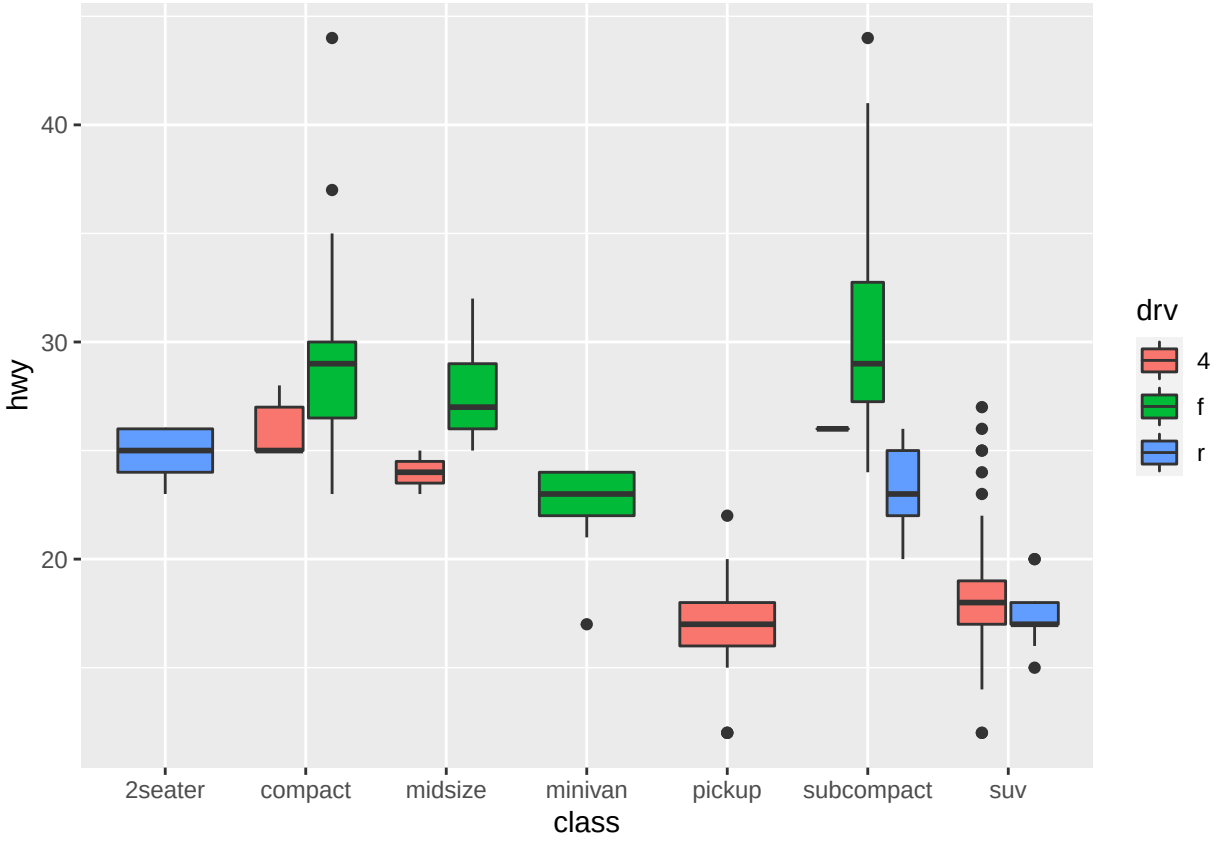
Exercise 6:

```
ggplot(mpg, aes(x = hwy, y= class))+
  geom_boxplot(color = 'black')+
  geom_point(position = 'jitter', color = 'light gray')+
  theme_bw()+
  labs(x='Highway MPG', y='Vehicle Class')
```



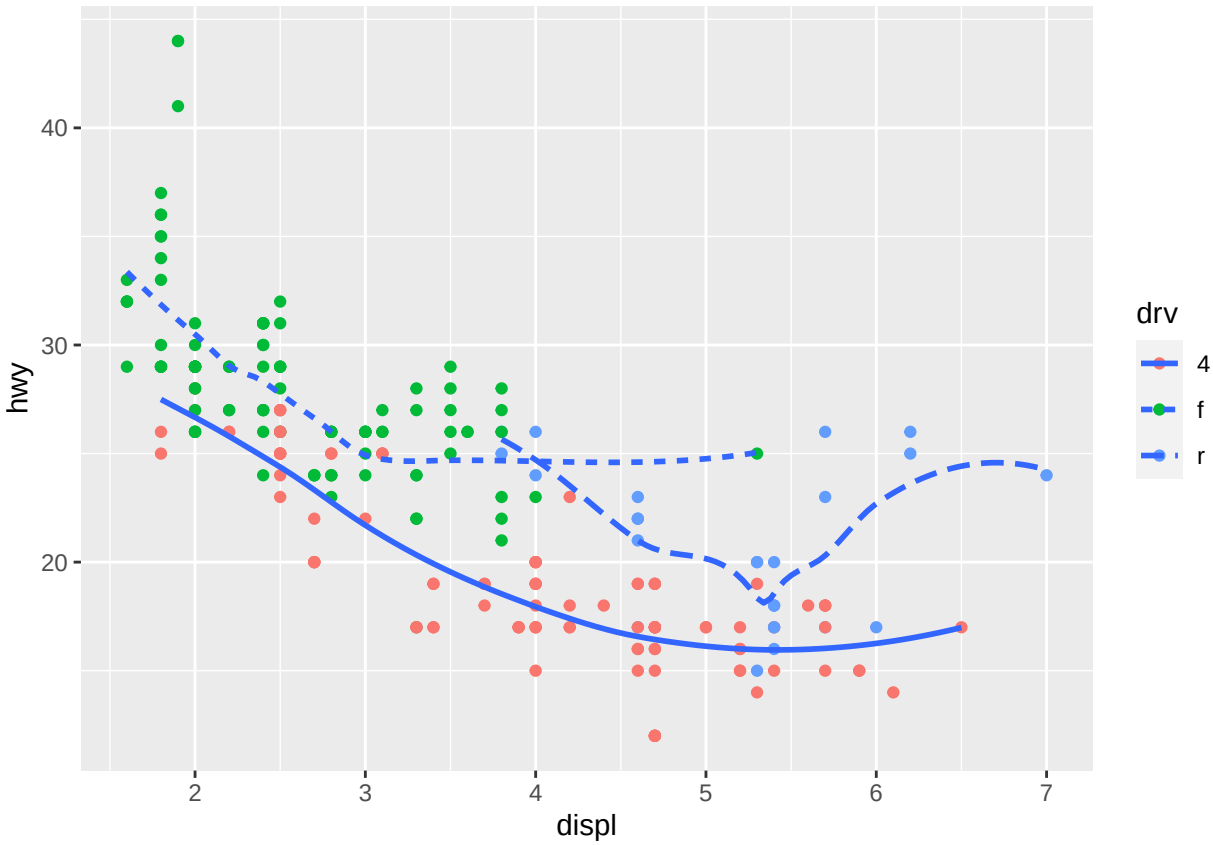
Exercise7:

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



Exercise8:

```
ggplot(mpg, aes(x = displ, y = hwy))+
  geom_point(aes(color = drv))+
  geom_smooth(formula = y ~x, method = 'loess', se = FALSE, aes(linetype=drv))
```


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