Pstat131HW1

Wentao Yu

2022-09-23

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: ???

Question4: According to lecture sides no.39, 'day_1_131_2321',

- 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.

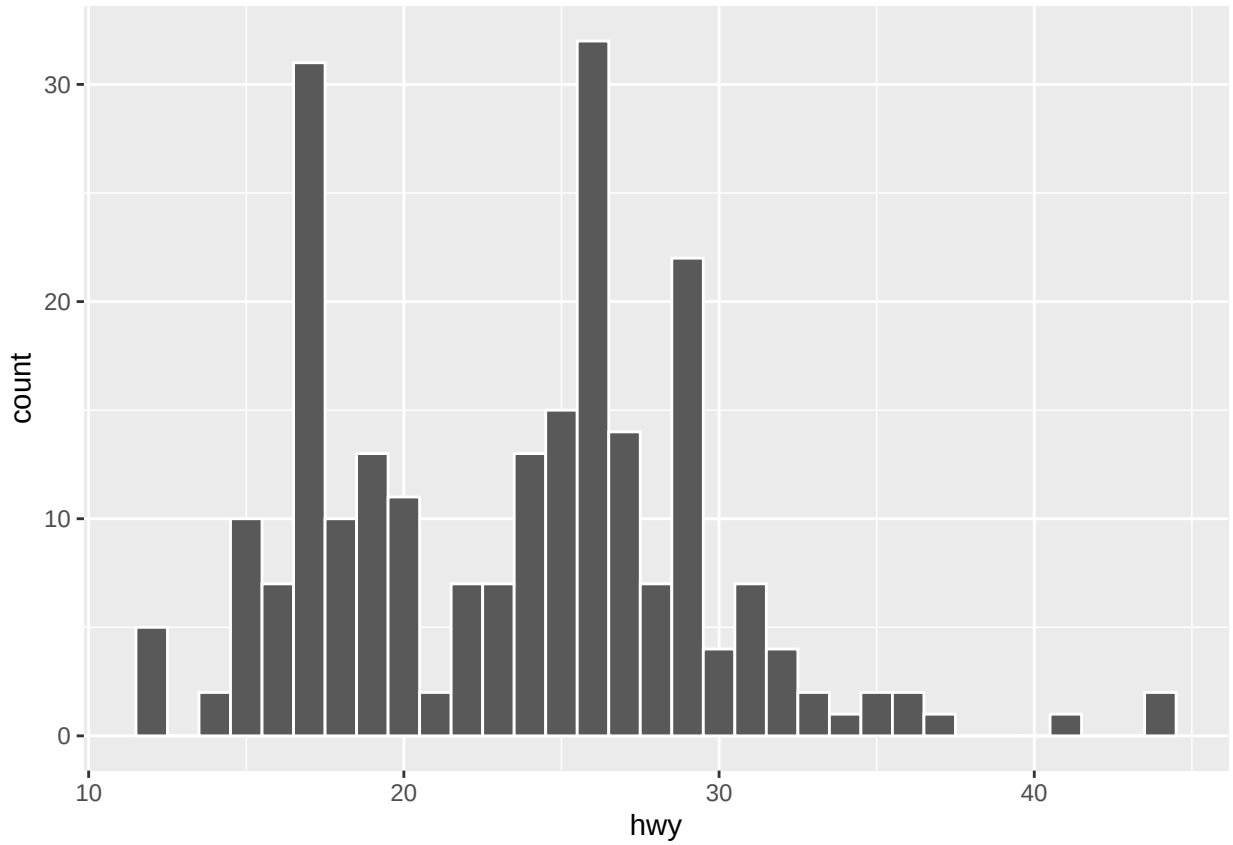
Question5:

- 1. Mechanics :
  2. Empirically-driven:
  3.
-
-

Question6:

Exercise1:

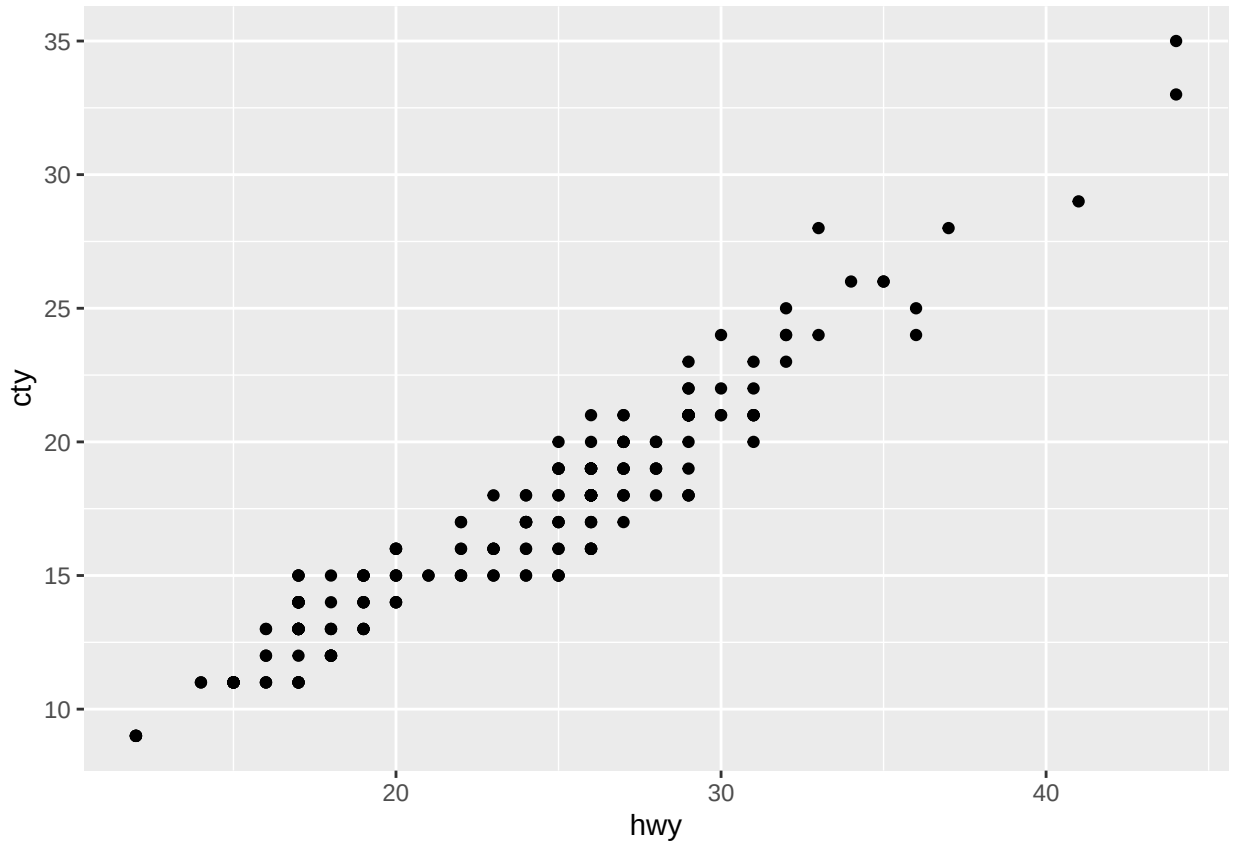
```
ggplot(mpg,aes(hwy))+geom_histogram(color = "white", binwidth = 1)
```



Explanation:

Exercise2:

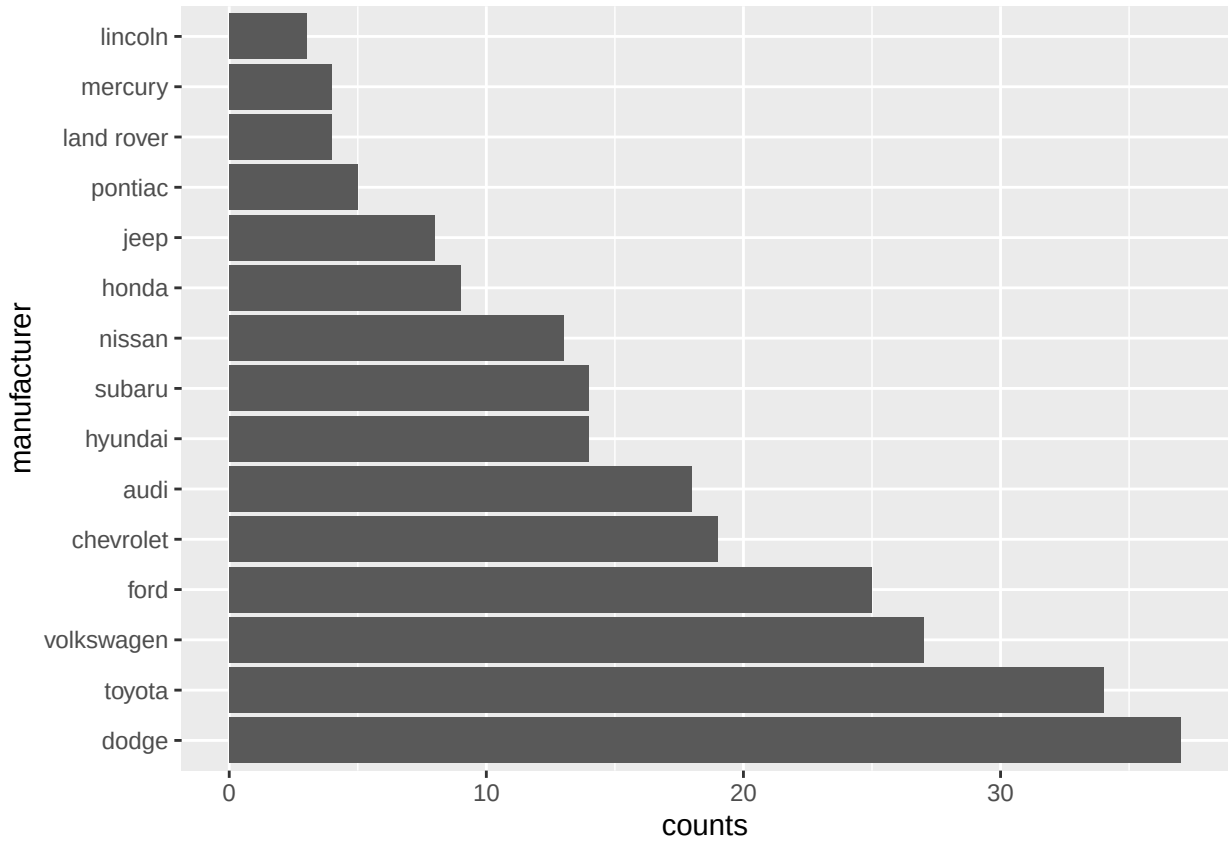
```
ggplot(mpg,aes(hwy,cty))+geom_point()
```



Explanation:

Exercise3:

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

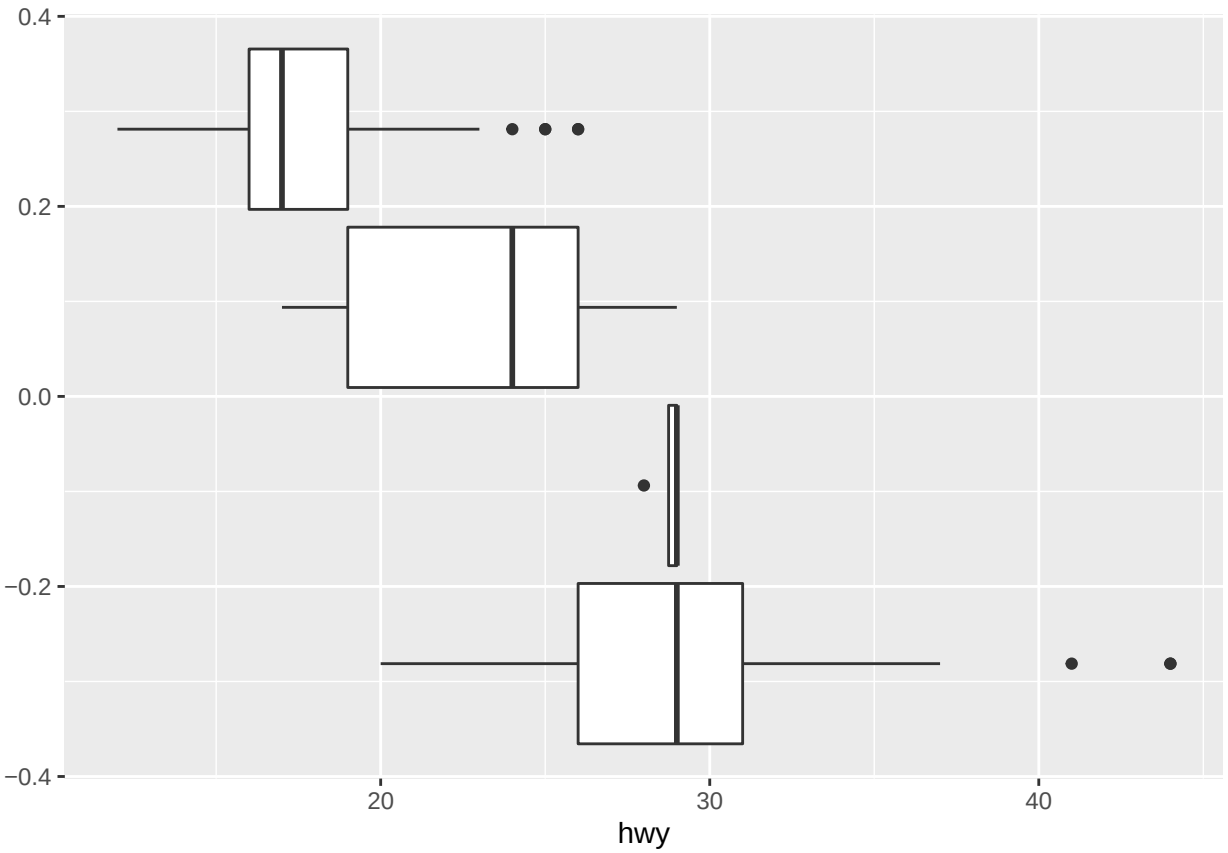



```
# how to change it to from highest to lowest
```

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

Exercise4:

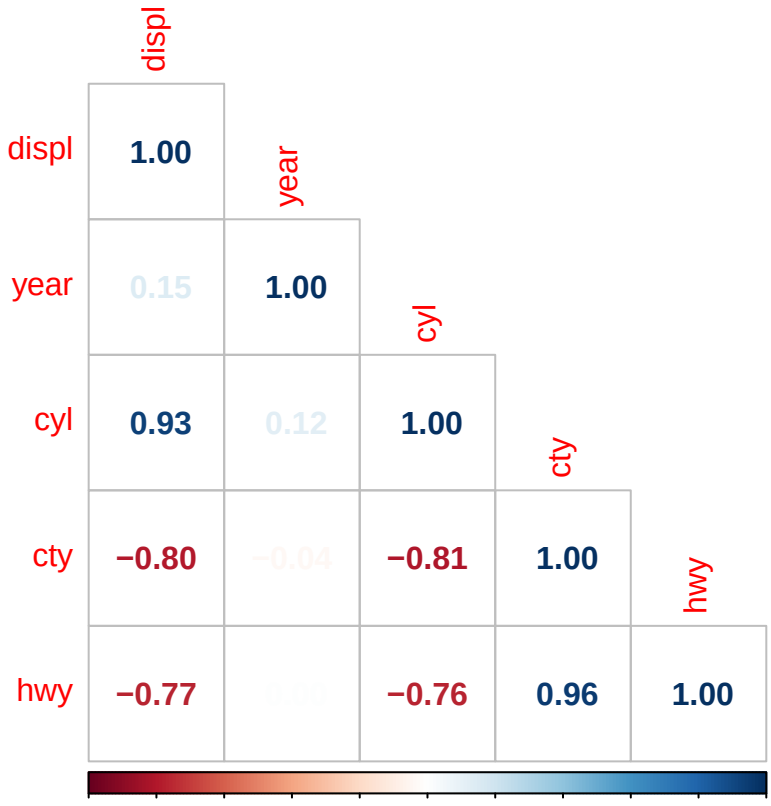
```
ggplot(mpg,aes(hwy)) + geom_boxplot(aes(group=cyl))
```



Explanation:

Exercise5:

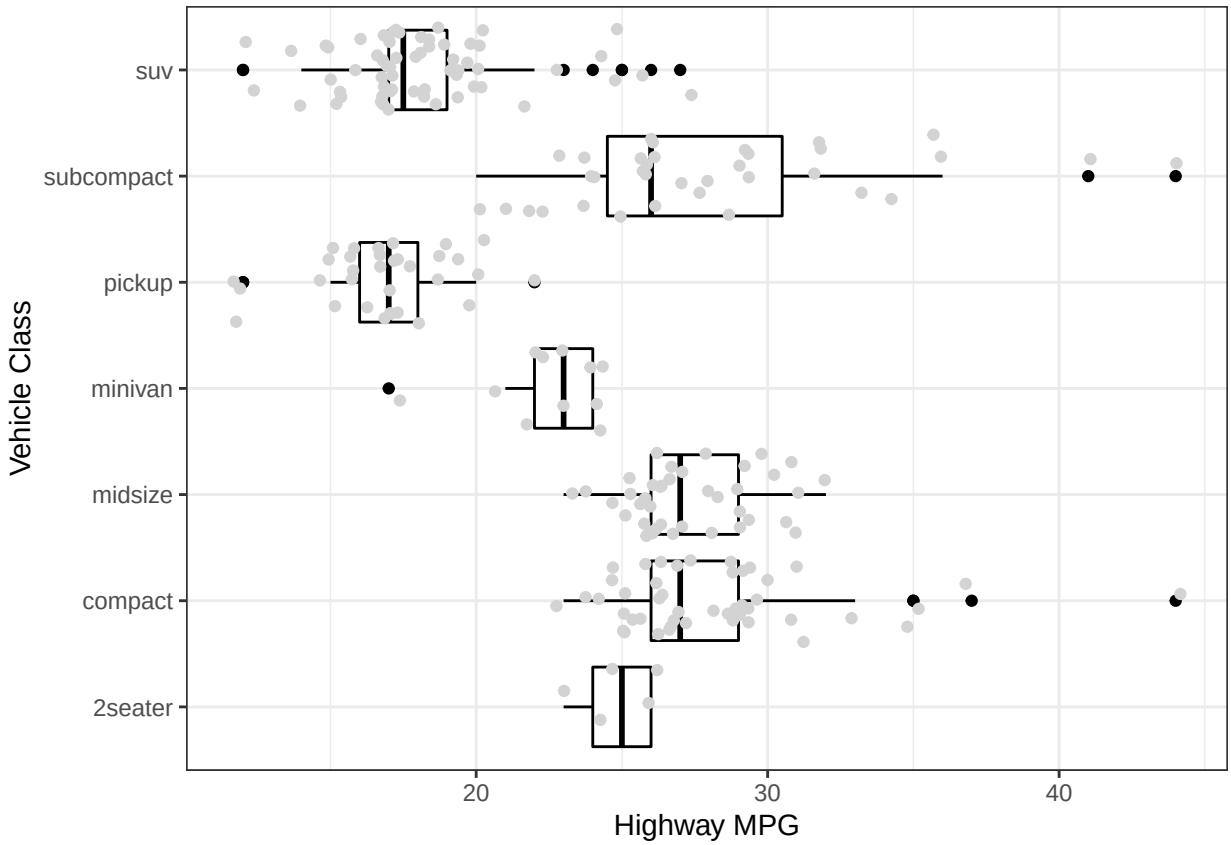
```
numeric_data <- select_if(mpg, is.numeric)
corrplot(cor(numeric_data),method = 'number', type = 'lower')
```



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.

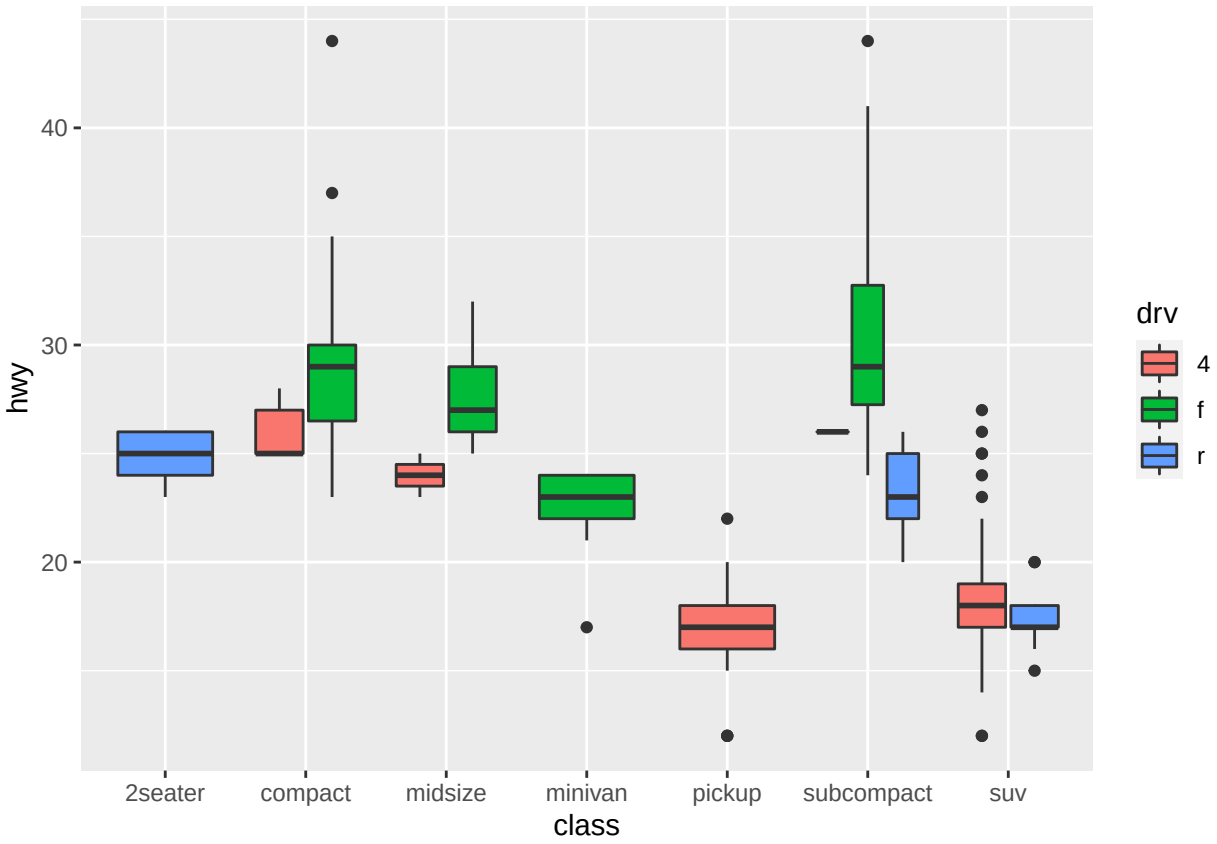
Exercise 6:

```
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')
```



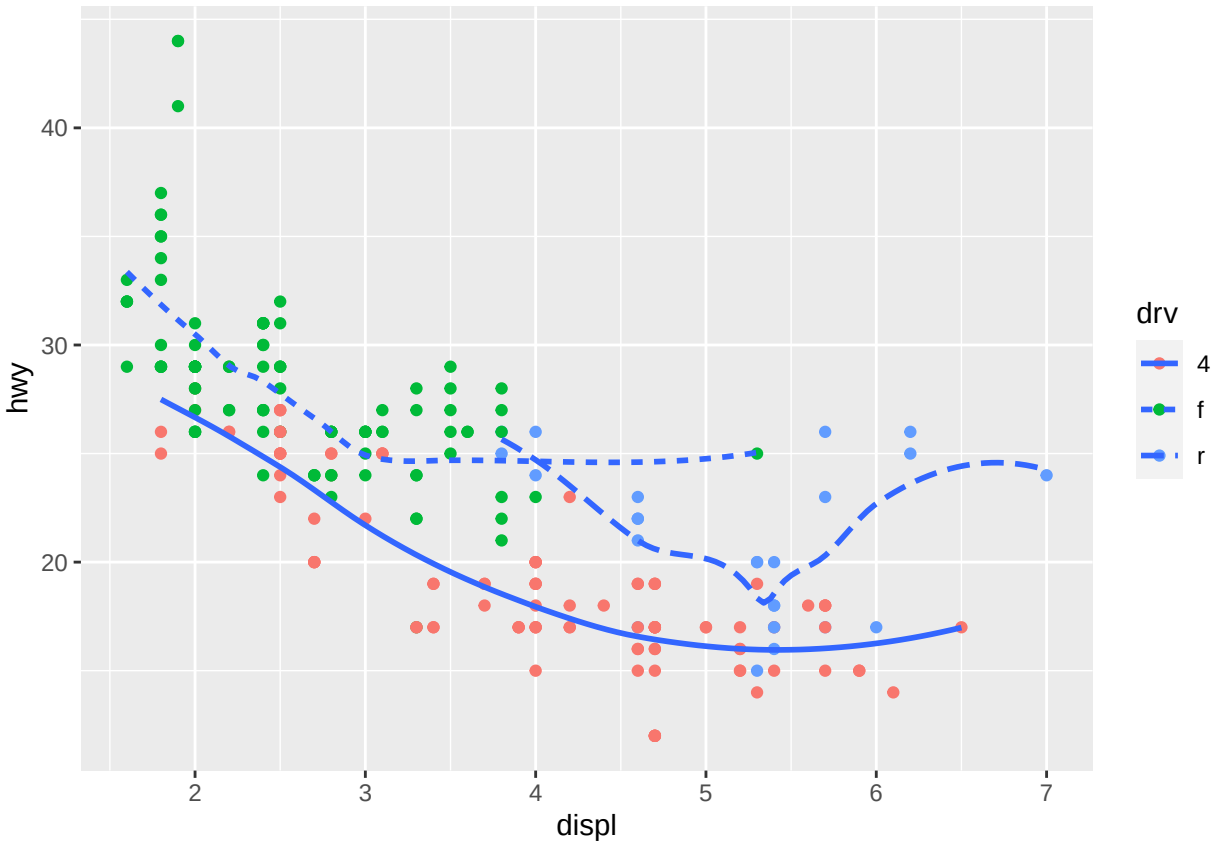
Exercise7:

```
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