PSTAT 131 HW1

Jiashu Huang

2022-10-02

Machine Learning Main Ideas
Question 1.

The main difference between supervised and unsupervised learning is the corresponding datasets. Supervised learning is only suitable for datasets that have predictor variables x and response variables y, while unsupervised learning fits data that only have predictor variables. Tasks such as prediction and estimation are common. The main goals of prediction, for example, is to derive a solution that predicts y using x most efficiently. Since unsupervised learning data doesn't have response variables, the main goal is mostly to learn and analyze the distribution of the datasets. Information about the dataset's distribution in a multi-dimension data space allows for tasks such as clustering.

Question 2.

The main difference between a regression model and a classification model is their corresponding type of response variables. Regression models predict quantitative reponse variables, which are numerical. Classification model deals with categorical data. Even if the response varaibles are represented by numbers, the values don't have any meanings and they are only labels that encode different categories.

Question 3.

Regression metrics: mean square error, rooted mean square error, and mean absolute error. Classification metrics: accuracy, confusion matrix(recall and precision), F1 Score, and the AUC-ROC curve.

Question 4.

Descriptive model aims to record and visualize a pattern or trend in a data. Predictive model aims to predict the response variables most efficiently (optimizing the evaluation metric chosen for the model). Inferential model aims to find the relationship between the predictors and the responses. The tasks could be, for example, finding the significant variables or interpret the model into arguments that have real-world meanings.

Question 5.

A mechanistic model is a mathematical equation (a fixed theory) describing the relationship between the predictors and the responses. A empirically-driven model is based on empirical observations rather than a fixed theory. Mechanistic models assumes a parametric form for the relationship between the predictors and the responses, whereas the empirically-driven model does not. Empirically-driven models are more flexible. As for similarities, they both could suffer issues of over fitting, and they both use predictors to predict responses.

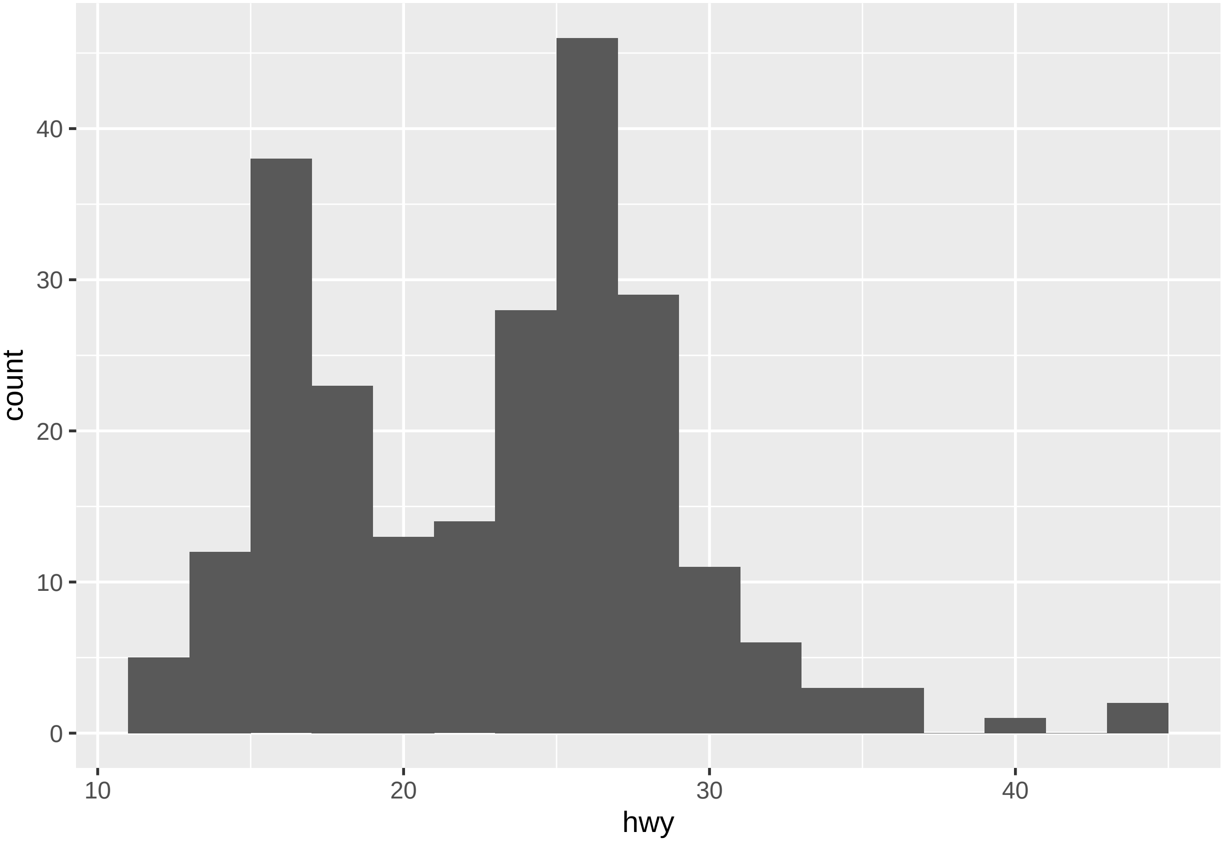
Question 6.

The first one is predictive, because the purposes is to estimate the likelihood of voting for a candidate. Only the result rather than the relationship is focused. The second one is inferential, since it tries to find out the influence of the change of one varibale on another. In this case, the relationship is focused much more than the result, which makes the question inferential.

Exploratory Data Analysis

Question 1.

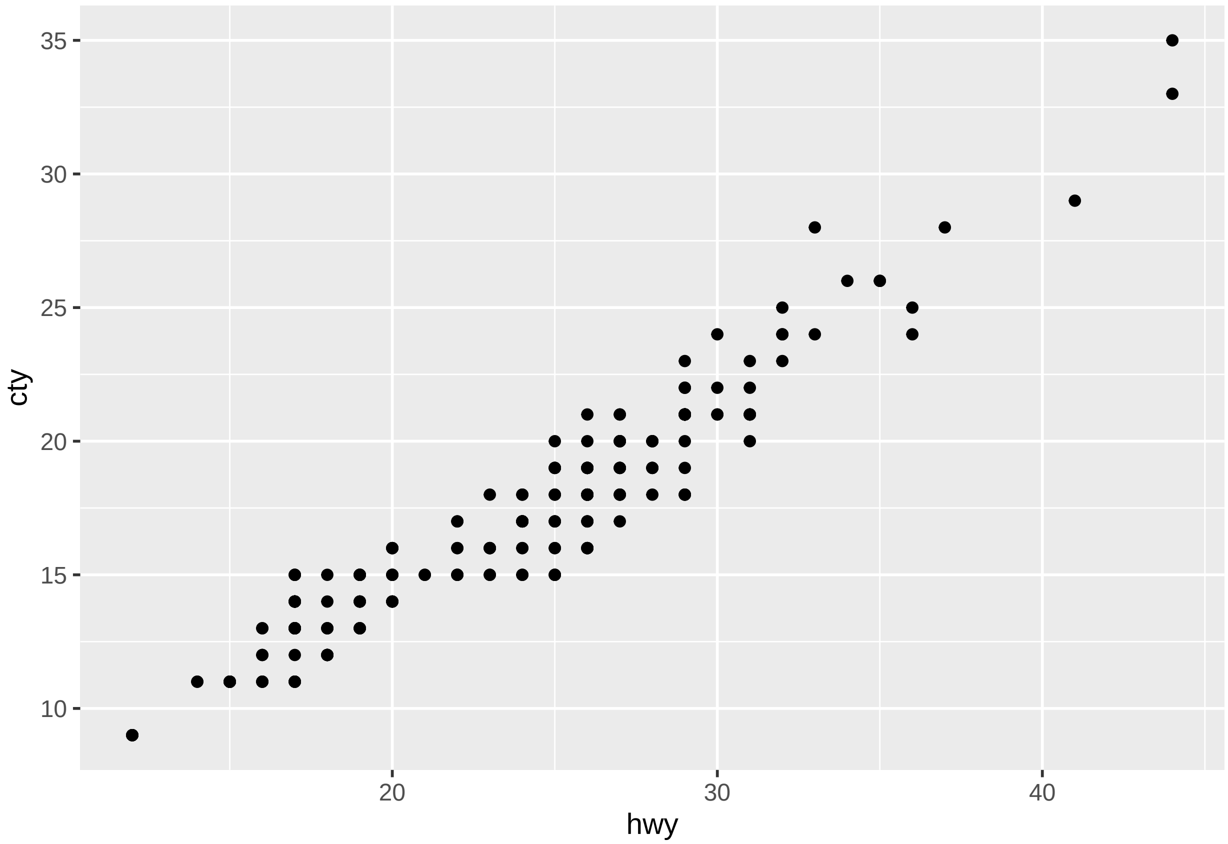
```
data('mpg')
ggplot(mpg)+geom_histogram(binwidth=2,aes(hwy))
```



The graph has two peaks at hwy = 17 and 27. It means that most cars have highway miles per gallon around 17 or 27 for some reason.

Question 2.

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



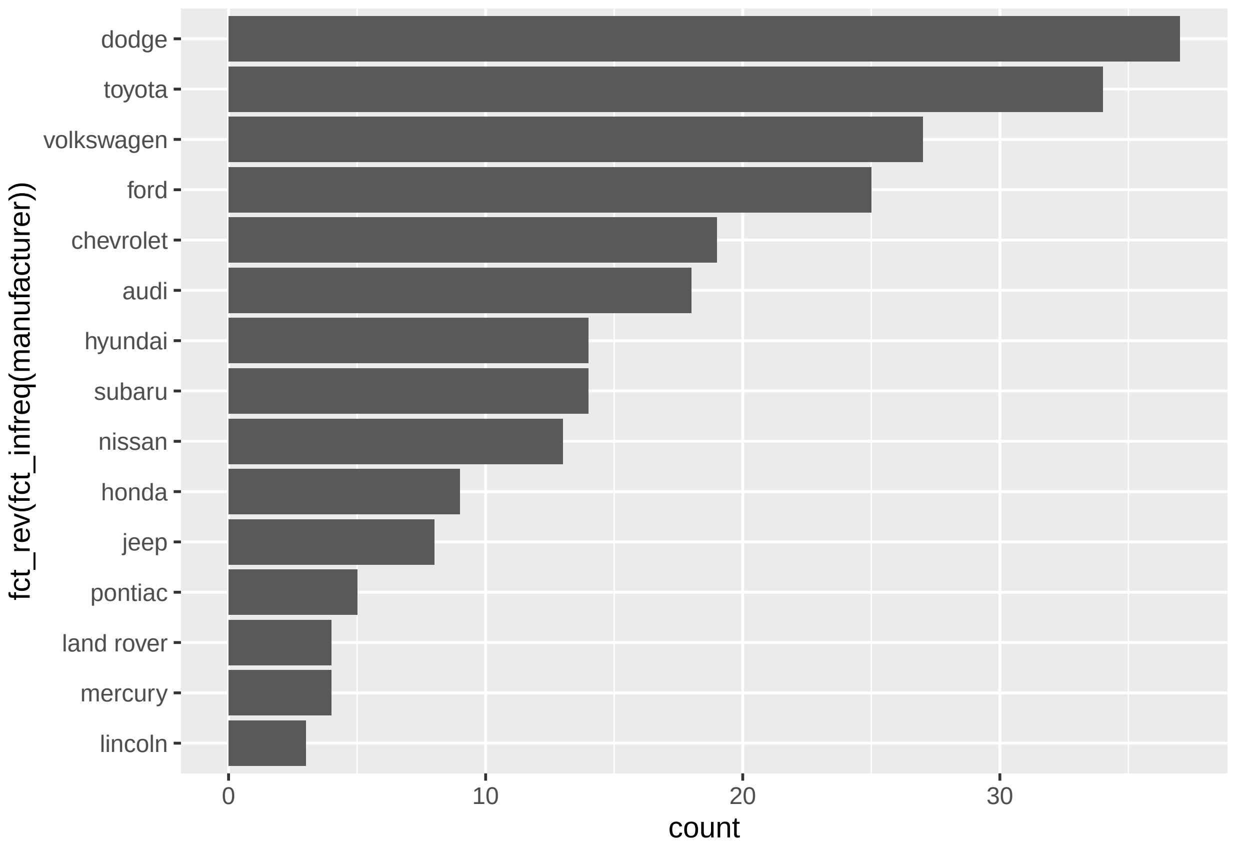
From the graph, it is reasonable to say that there is a positive linear relationship between hwy and cty. The more highway miles per gallon, the more city miles per gallon.

Question 3.

```
mpg[order()]
```

```
## # A tibble: 234 x 0
## # i Use `print(n = ...)` to see more rows
```

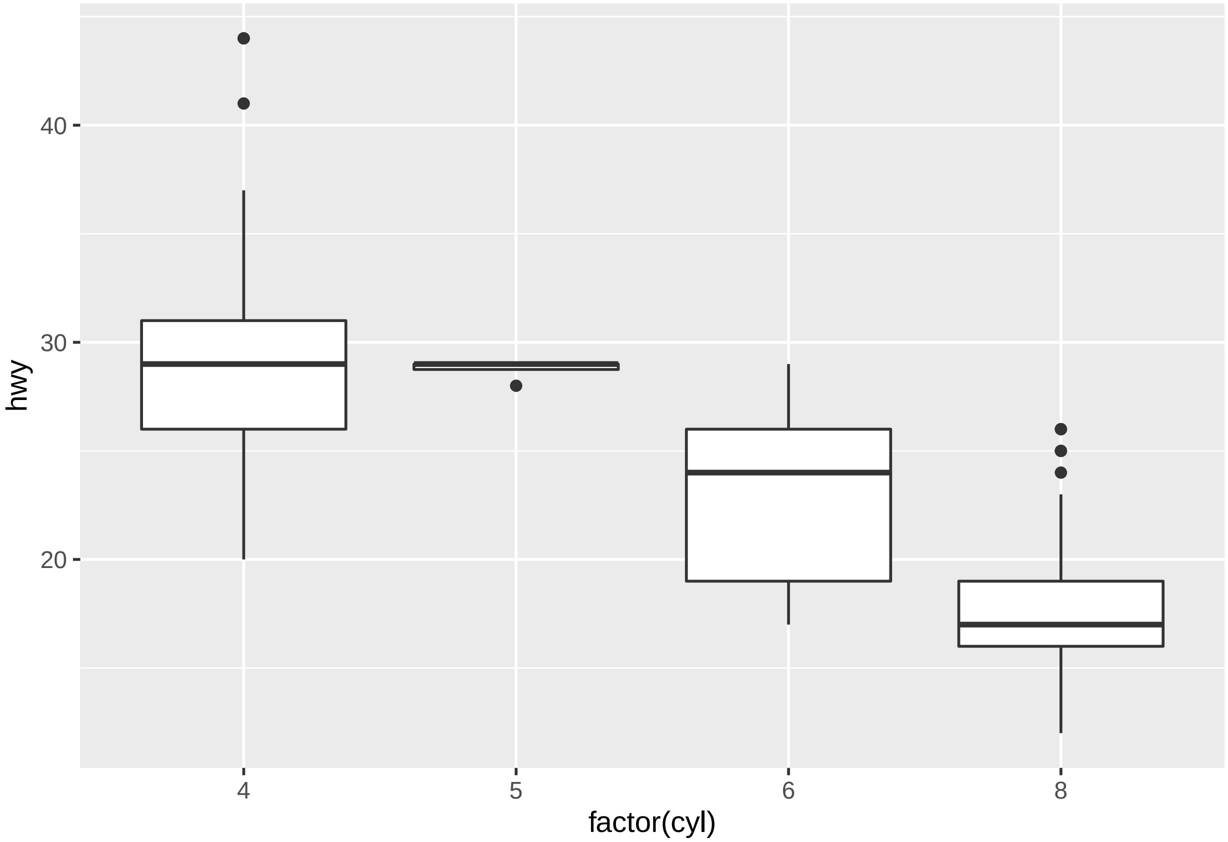
```
ggplot(mpg)+geom_bar(aes(fct_rev(fct_infreq(manufacturer))))+coord_flip()
```



Dodge produces the most cars. Lincoln produces the least.

Question 4.

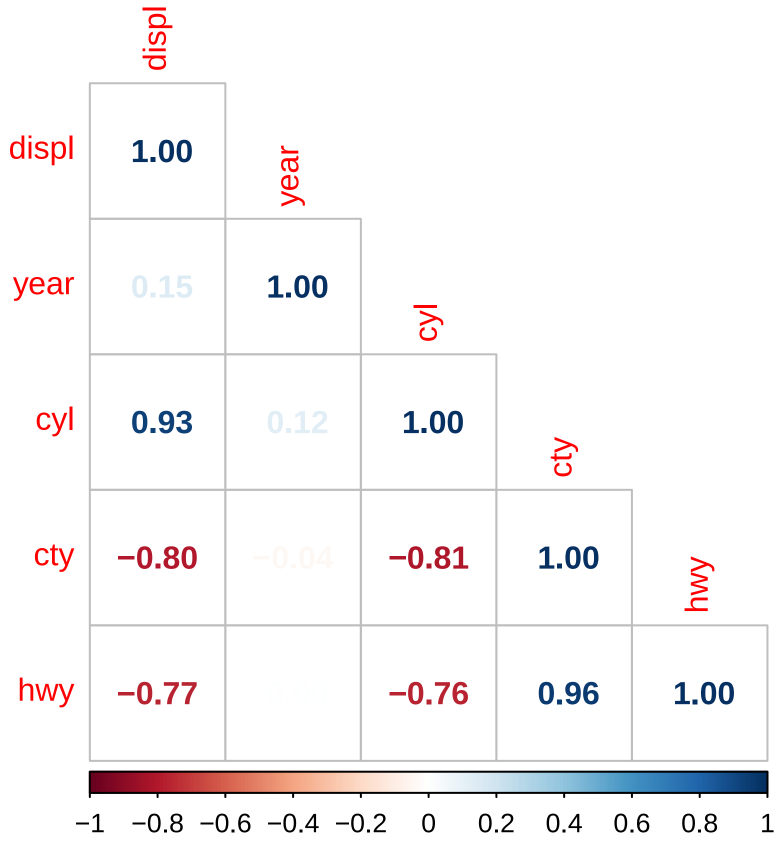
```
ggplot(mpg,aes(x=factor(cyl),y=hwy))+geom_boxplot()
```



There seems to be a negative linear relationship between hwy and cyl. The more cylinders a car model has, the less highway miles per gallon it is likely to have.

Question 5.

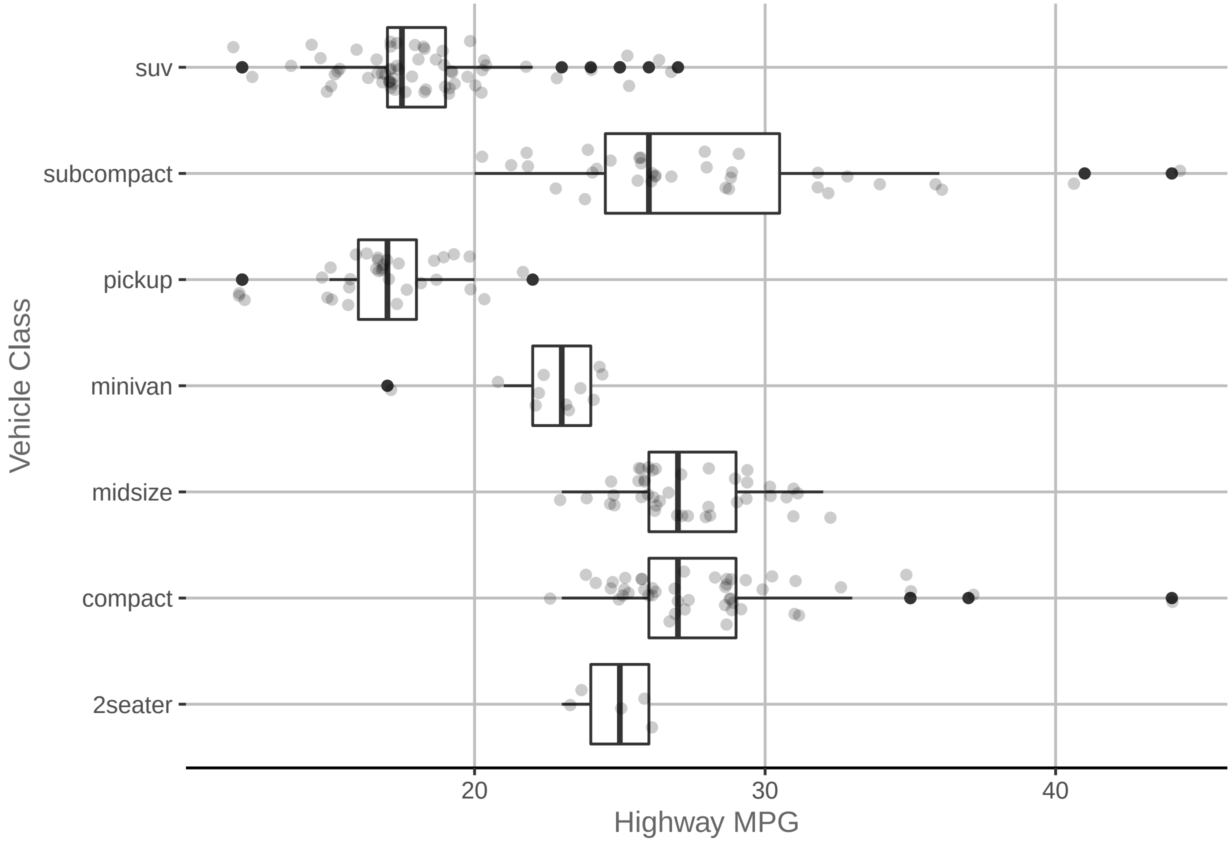
```
M = cor(mpg[,sapply(mpg, is.numeric)])
corrplot(M, method = 'number',type='lower')
```



disp is positively related to cyl, negatively related to cty and hwy. cty and hwy both have a negative relationship with cyl, while there is a positive relationship between them. I am surprised that year has little correlation with any other variables, indicating that the year of manufacture doesn't matter. The relationships between cty, cyl, and hwy are pretty intuitive. I don't really know what engine displacement means, so its corresponding relationships don't make sense to me.

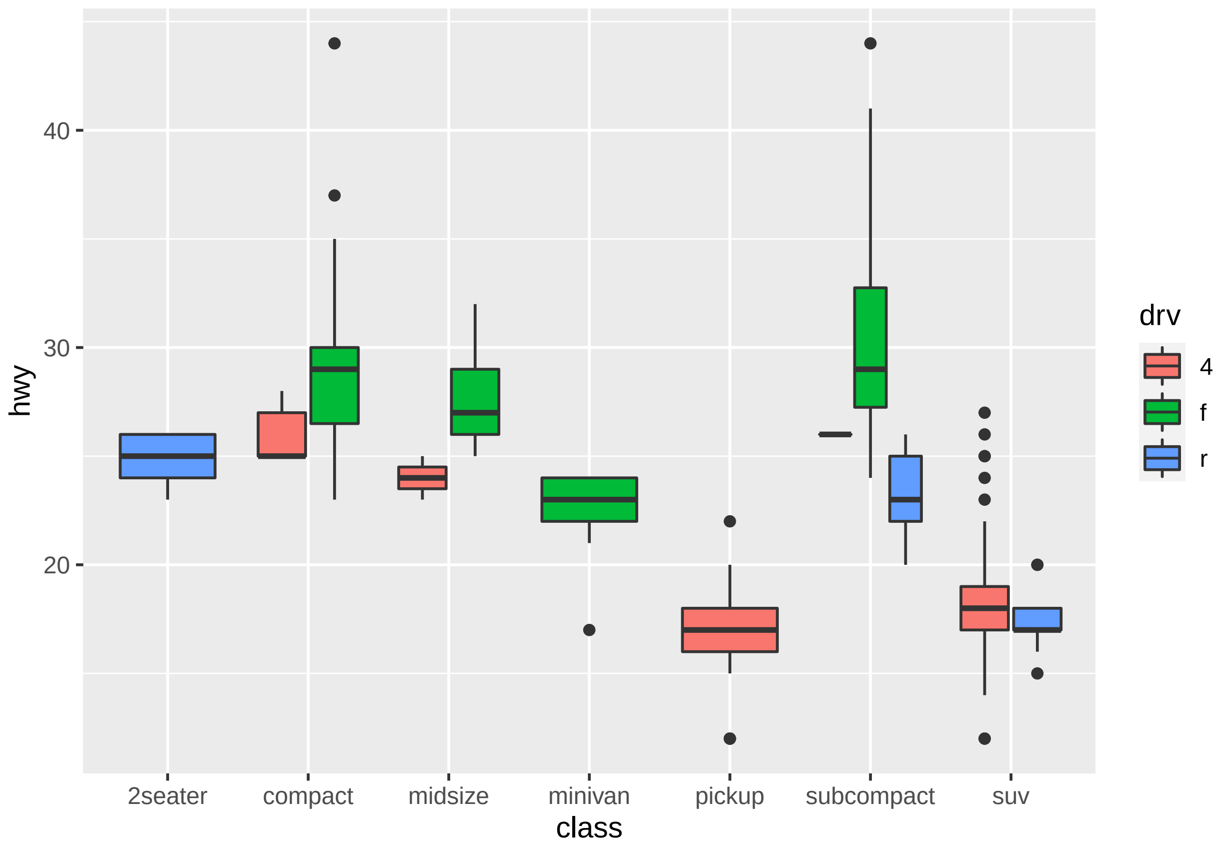
Question 6.

```
ggplot(mpg,aes(x=factor(class),y=hwy))+
  geom_boxplot()+
  geom_jitter(alpha=0.2,width=0.25)+
  coord_flip()+
  theme(
  text=element_text(color="#666666"),
  panel.border = element_blank(),
  panel.background = element_rect(fill = "white",
                                  colour = "white",
                                  size = 0.5, linetype = "solid"),
  panel.grid.major = element_line(size = 0.5, linetype = 'solid',
                                  colour = "grey"),
  panel.grid.minor = element_blank(),
  axis.line.x = element_line(size = 0.5, linetype = "solid",
                                     colour = "black"))+
  labs(y = "Highway MPG", x = "Vehicle Class",color="grey")
```



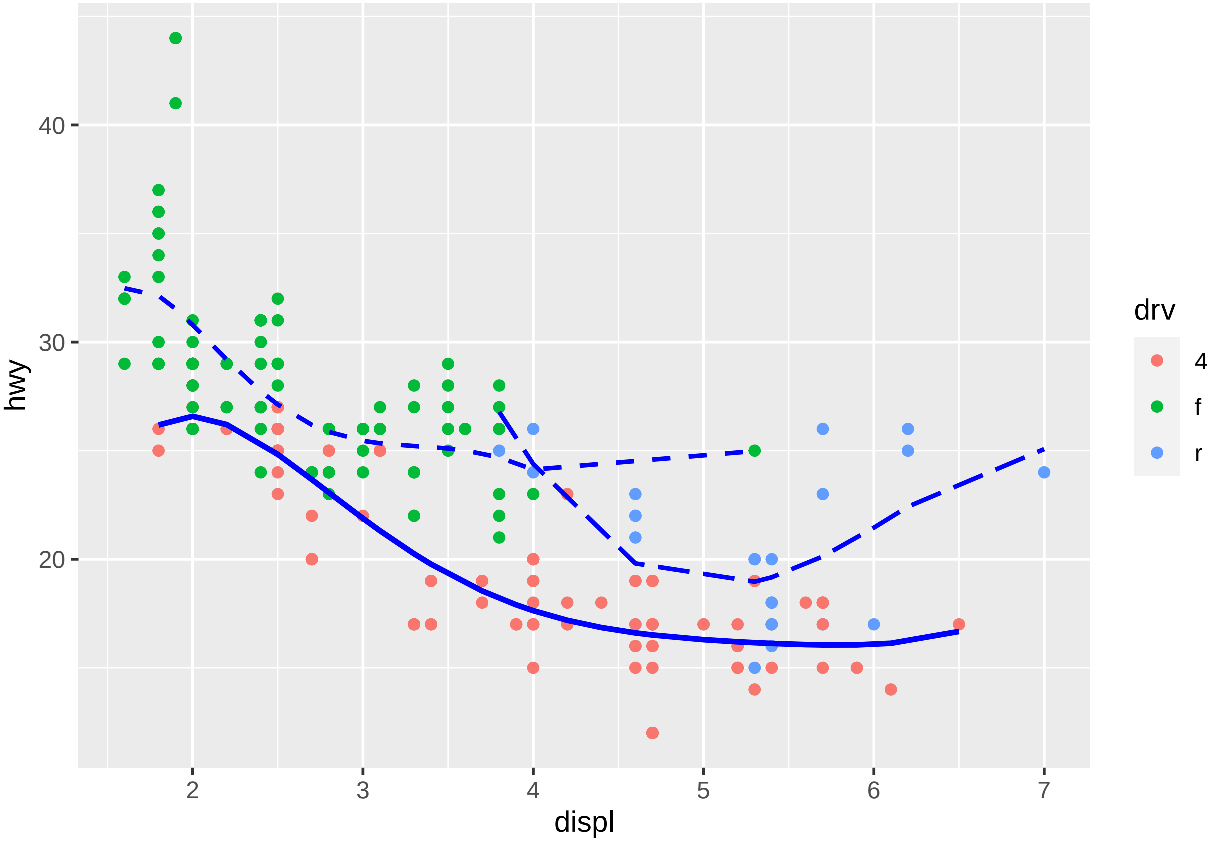
Question 7.

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



Question 8.

```
nls_fit4 <- nls(hwy ~ a + b * displ^(2) + c * displ^(3) + d * displ^(4) + e * displ^(5) + f * displ^(-1
predicted_4 <- data.frame(mpg_pred = predict(nls_fit4,mpg[mpg$drv==4,]), displ=mpg[mpg$drv==4,]$displ)
nls_fitf <- nls(hwy ~ a + b * displ^(2) + c * displ^(3) + d * displ^(4) + e * displ^(5) + f * displ^(-1
predicted_f <- data.frame(mpg_pred = predict(nls_fitf,mpg[mpg$drv=='f',]), displ=mpg[mpg$drv=='f',]$dis
nls_fitr <- nls(hwy ~ a + b * displ^(2) + c * displ^(3) + d * displ^(4), mpg[mpg$drv=='r',], start = li
predicted_r <- data.frame(mpg_pred = predict(nls_fitr,mpg[mpg$drv=='r',]), displ=mpg[mpg$drv=='r',]$dis
ggplot(mpg,aes(x=displ,y=hwy,colour=drv))+
  geom_point()+
  geom_line(color='blue',data = predicted_4, aes(y=mpg_pred, x=displ),size=1)+
  geom_line(color='blue',data = predicted_r, aes(y=mpg_pred, x=displ),linetype="longdash",size=0.8)+
  geom_line(color='blue',data = predicted_f, aes(y=mpg_pred, x=displ),size=0.8,linetype="dashed")
```


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