PSTAT231 HW1

muxi

2022-10-10

PSTAT 231 Homework 1

Machine Learning Main Ideas

Question 1

Supervised learning:

Machine learn from observed output and input to fit the model with labelled data: Regression/Classification

Unsupervised learning:

Machine only learn from the predictor data without labelled outcome. We try to capture the significant features of data: Clustering/ Dimensionality reduction/ Density estimation

Difference between supervised and unsupervised learning:

Availability of outcome data; Different targets; Supervised learning is a simple method for machine learning, while in unsupervised learning, we need powerful tools. Also unsupervised learning models are computationally complex.

Question 2

Difference between a regression model and a classification model:

Both regression model and classification model are supervised learning. But the labelled data Ys are different. In regression model, Y is quantitative or we can say that Y is continuous which represents numerical values. In classification model, Y is qualitative or we can say Y is discrete which represents categorical value.

Question 3

Commonly used metrics for regression ML problems:

Training MSE; Testing MSE; MRSE

Commonly used metrics for classification ML problems:

Training error rate; Testing error rate; Confusion Matrix

Question 4

Descriptive models: Choose model to best visually emphasize a trend in data.

Inferential models: Aim is to test theories which state relationship between outcome and predictor(s).

Predictive models: Aim is to predict Y with minimum reducible error.(from lecture)

Question 5

Mechanistic: Assume a parametric form or we can say that we have a prior distribution of f –target.

Empirically-driven: No assumptions about f. We only use the big data to explore f.

In Mechanistic, we can use the current information about the f to restrict it. In this case, we can add parameters when we receive more information. In Empirically-driven, we use large number of observed data to train a non-parametric model such like KNN to fit our data. All these methods may cause overfitting or we can say : the curese of dimensionality.

Question 6

The first one is predictive. Aim is to predict how likely they will vote in favor of the candidate.

The latter one is inferential. Aim is to test theories which state relationship voter's likelihood of support for the candidate and whether they had personal contact with the candidate.

Exploratory Data Analysis

Exercise 1

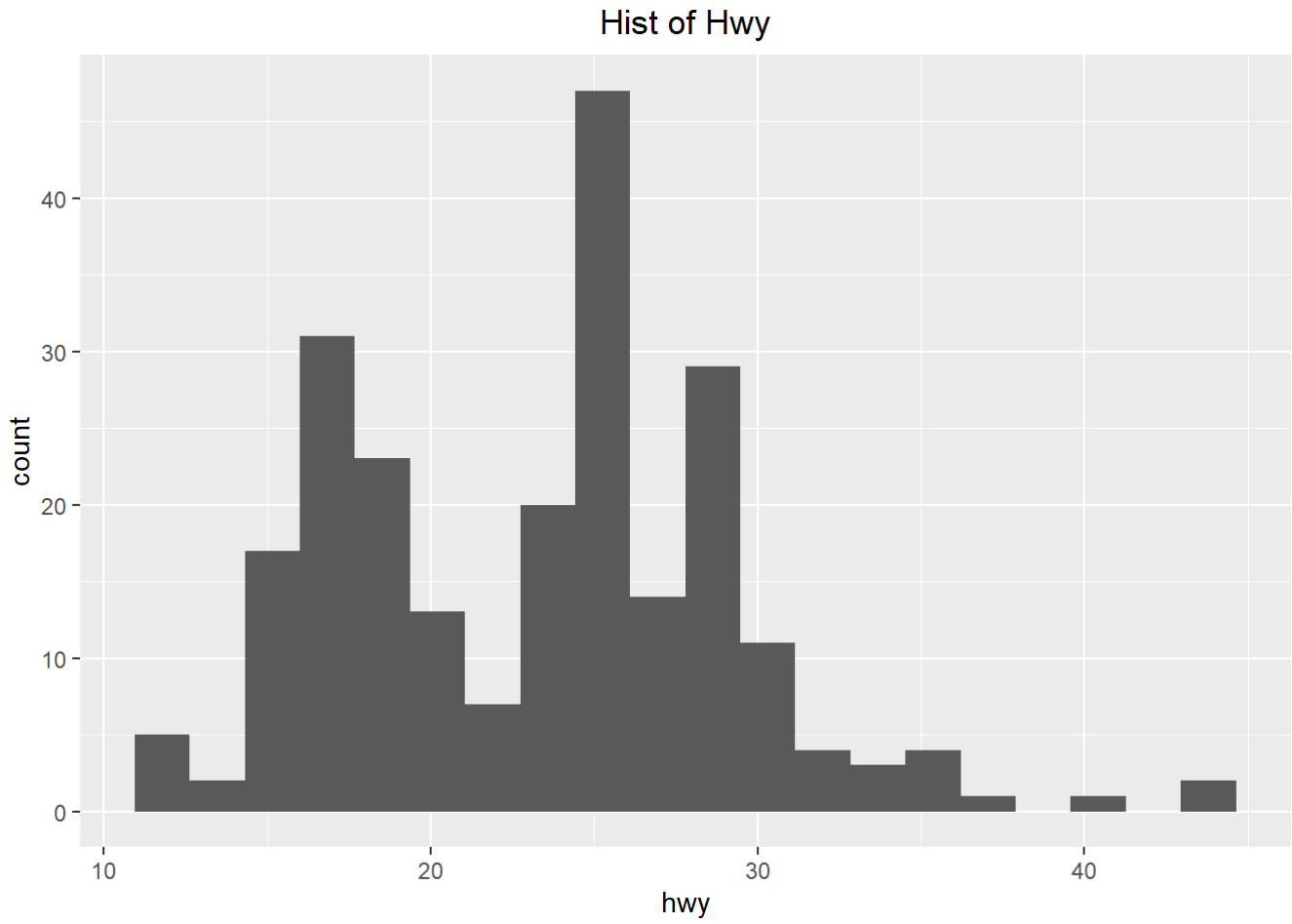
```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.8     ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1     ✓ stringr 1.4.1
## ✓ readr   2.1.3     ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
head(mpg)
```

```
## # A tibble: 6 × 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa…
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa…
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa…
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa…
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa…
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa…
```

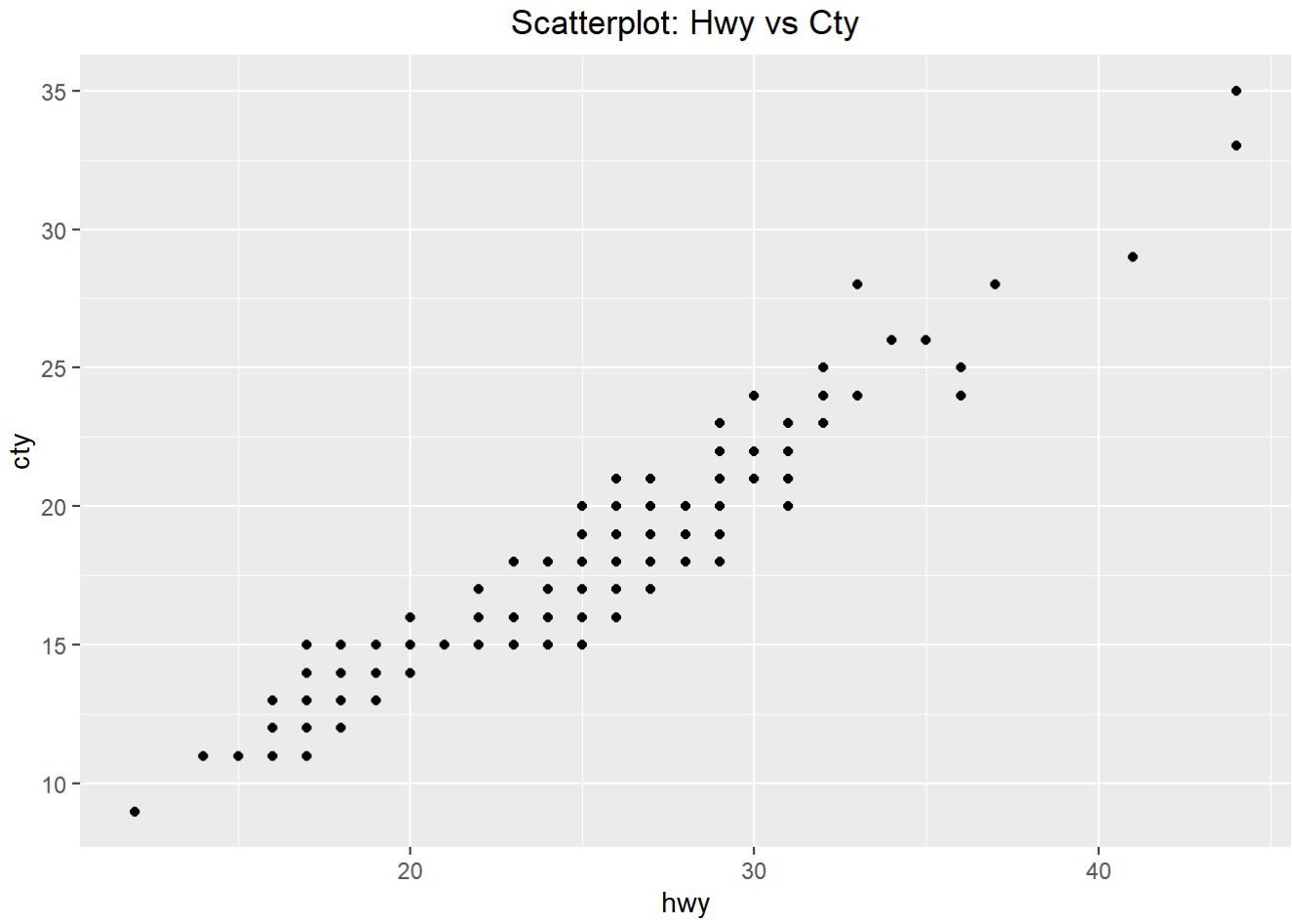
```
ggplot(data=mpg,aes(hwy))+geom_histogram(bins = 20)+ggtitle('Hist of Hwy')+
  theme(plot.title = element_text(hjust = 0.5))
```



It is clear that there exists several outliers larger than 37. The distribution looks like right-skewed and we could not ensure the accurate distribution of it. Also, it looks like a Multimodal distribution. The current informaiton are not efficient to see more characteristics of it. We may change the basic for exploration.

Exercise 2

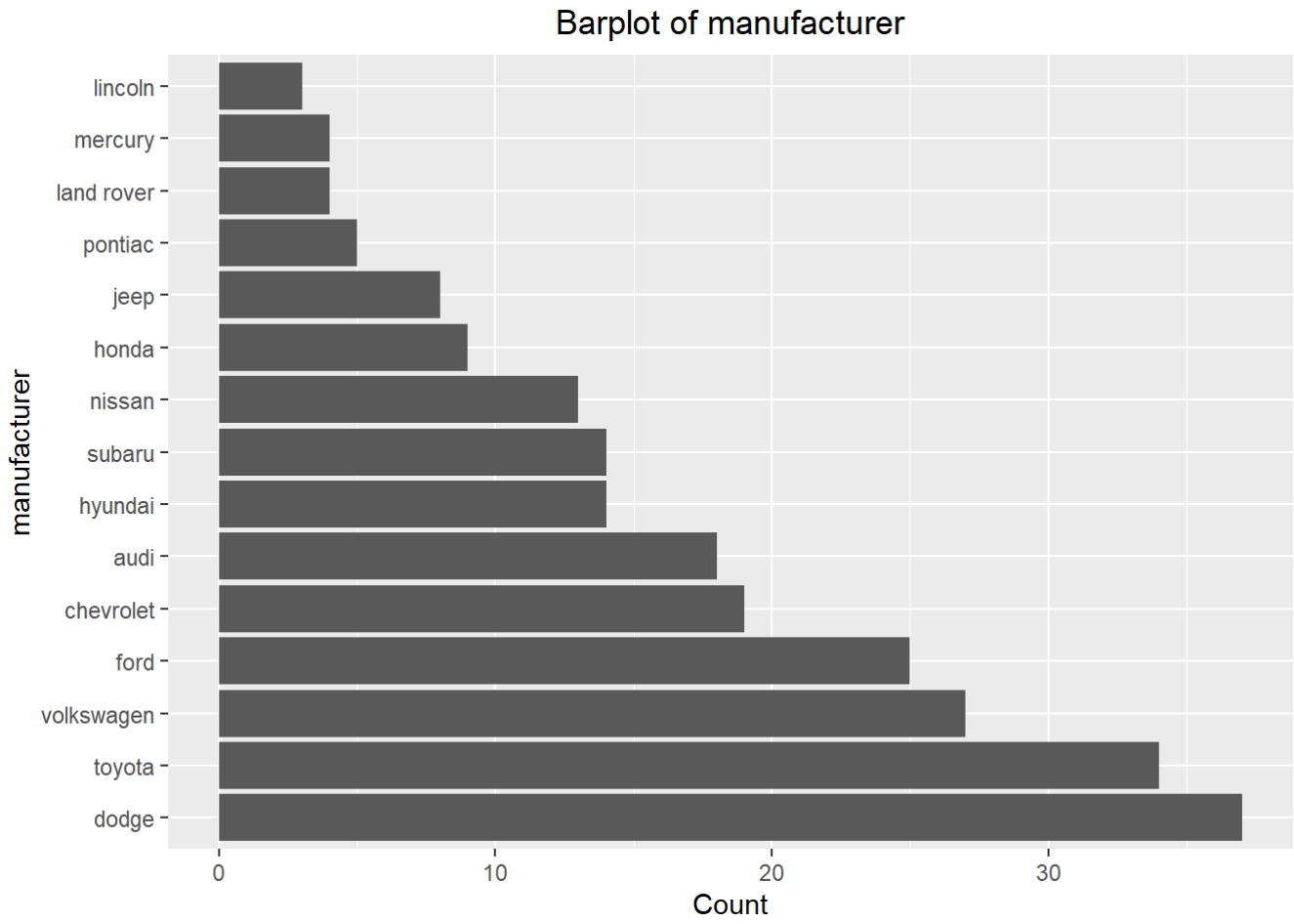
```
ggplot(data=mpg,aes(x=hwy,y=cty))+geom_point()+ggtitle('Scatterplot: Hwy vs Cty')+
  theme(plot.title = element_text(hjust = 0.5))
```



It looks like there is a strong positive correlation between hwy and cty, which means they move in the same direction. To be more specific, cty decreases as hwy decreases, or cty increases while hwy increases.

Exercise 3

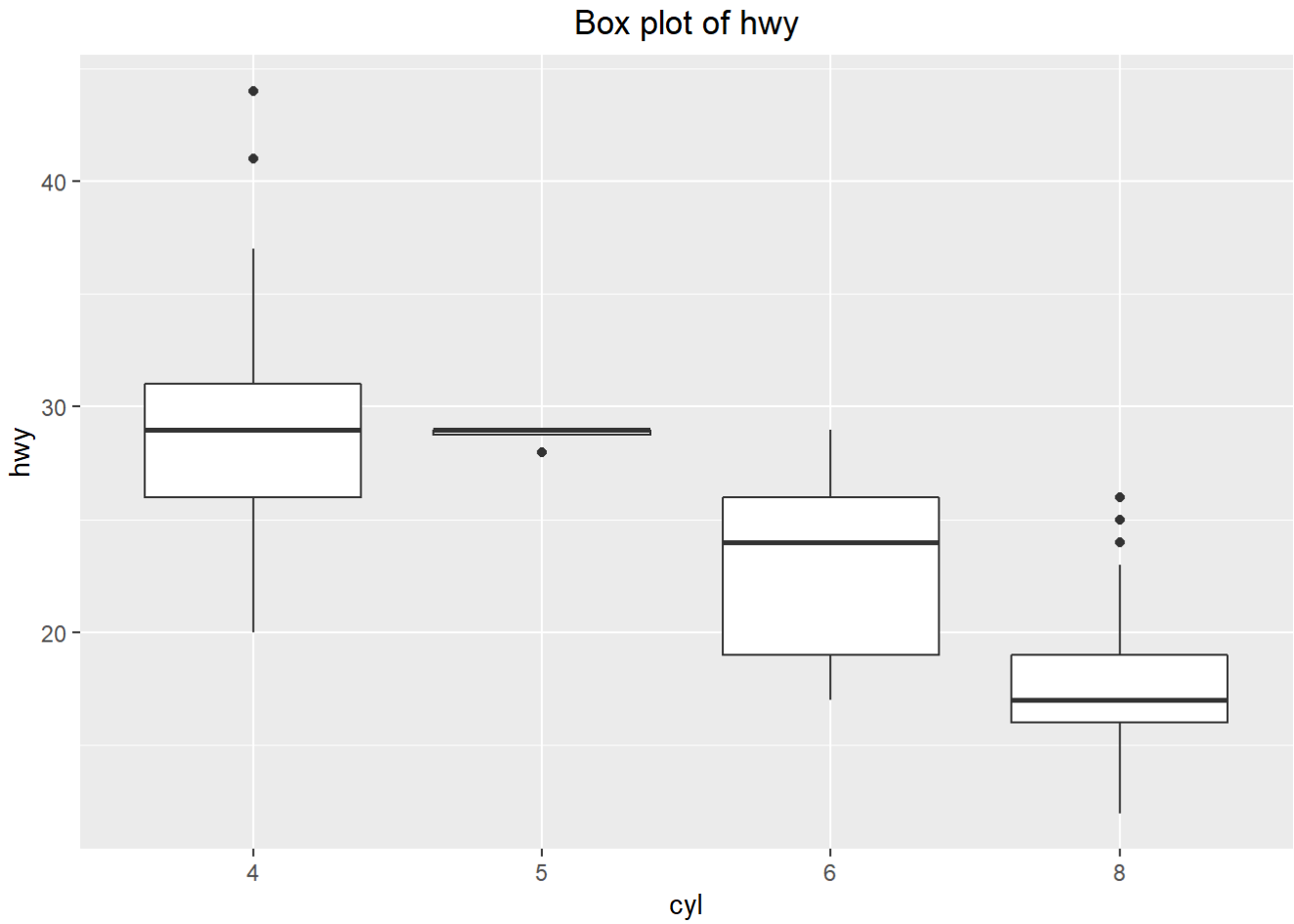
```
library(forcats)
k=c(t(mpg[,"manufacturer"]))
data1= data.frame(Count = k,value = sample(1:nrow(mpg)))
data1 %>%
  ggplot(aes(y = fct_infreq(Count))) +
  geom_bar() +
  labs(x = "Count",y="manufacturer")+ggtitle('Barplot of manufacturer')+
  theme(plot.title = element_text(hjust = 0.5))
```



Dodge produced the most cars and Lincoln produced the least.

Exercise 4

```
data2=mpg
data2$cyl[data2$cyl == 4]=1
data2$cyl[data2$cyl == 5]=2
data2$cyl[data2$cyl == 6]=3
data2$cyl[data2$cyl == 7]=4
data2$cyl[data2$cyl == 8]=5
data2$cyl = factor(data2$cyl,levels=c(1,2,3,4,5),labels=c("4","5","6","7","8"))
ggplot(data = data2,aes(x = cyl,y =hwy)) + geom_boxplot()+ggtitle('Box plot of hwy')+
  theme(plot.title = element_text(hjust = 0.5))
```



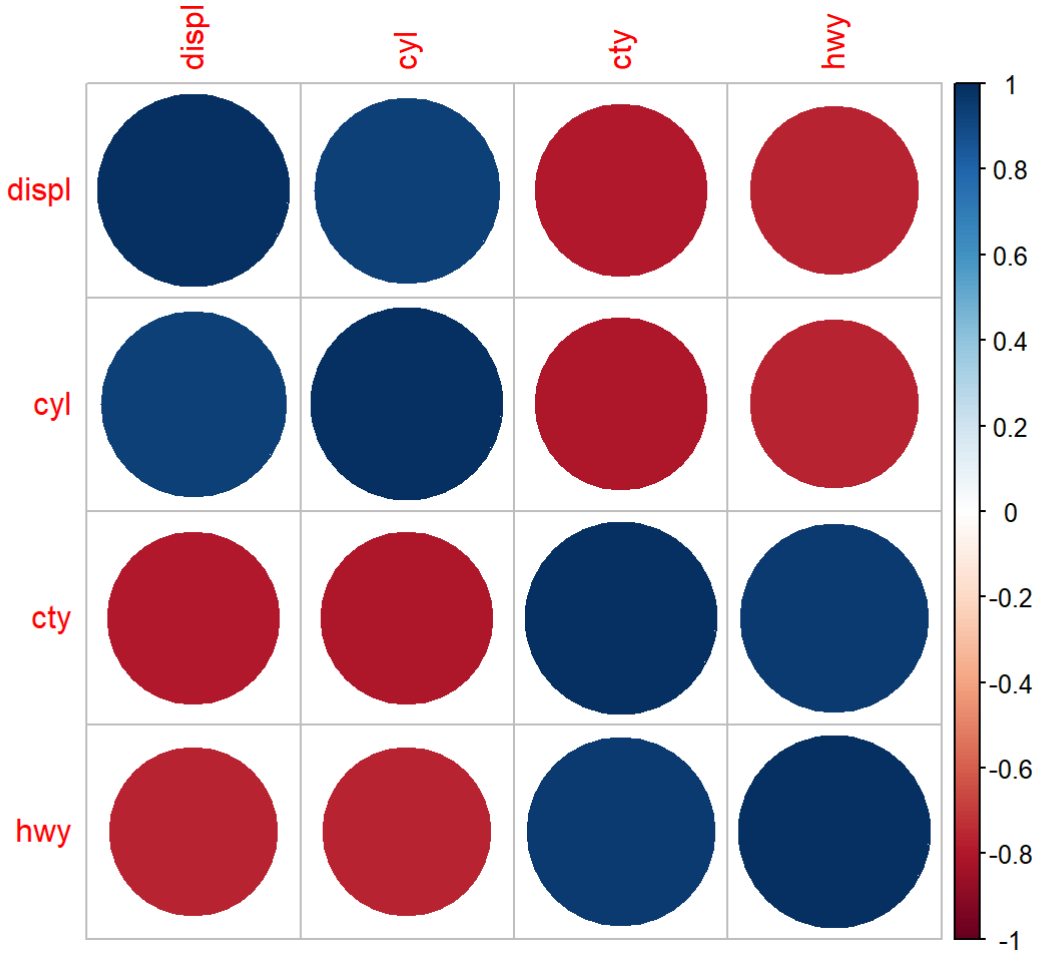
We could see that there is decreasing trend of hwy according to cyl. The mean keeps decreasing while the distribution between each level of cyl seems similar. I guess the cyl controls the average value of hwy. With higher cyl, the hwy becomes less.

Exercise 5

```
#Exercise 5
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
data3=mpg[,c(3,5,8,9)]
M = cor(data3)
corrplot(M)
```



From graph, we could see that displ is positively correlated with cyl and negatively correlated with cty and hwy.

Cyl is positively correlated with displ and negatively correlated with cty and hwy.

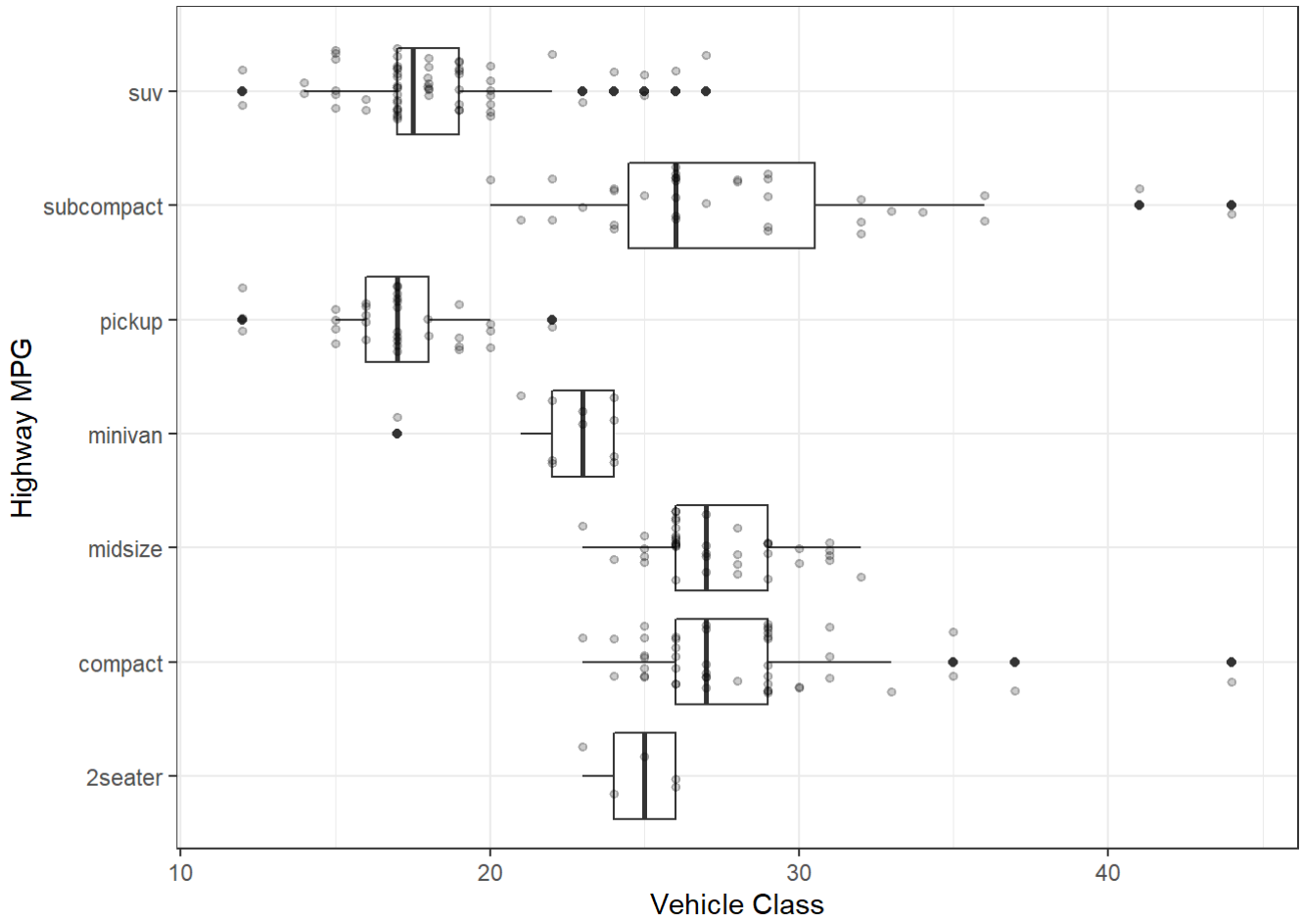
Cty is negatively correlated with displ and cyl and positively correlated with hwy.

Hwy is negatively correlated with displ and cyl and positively correlated with cty.

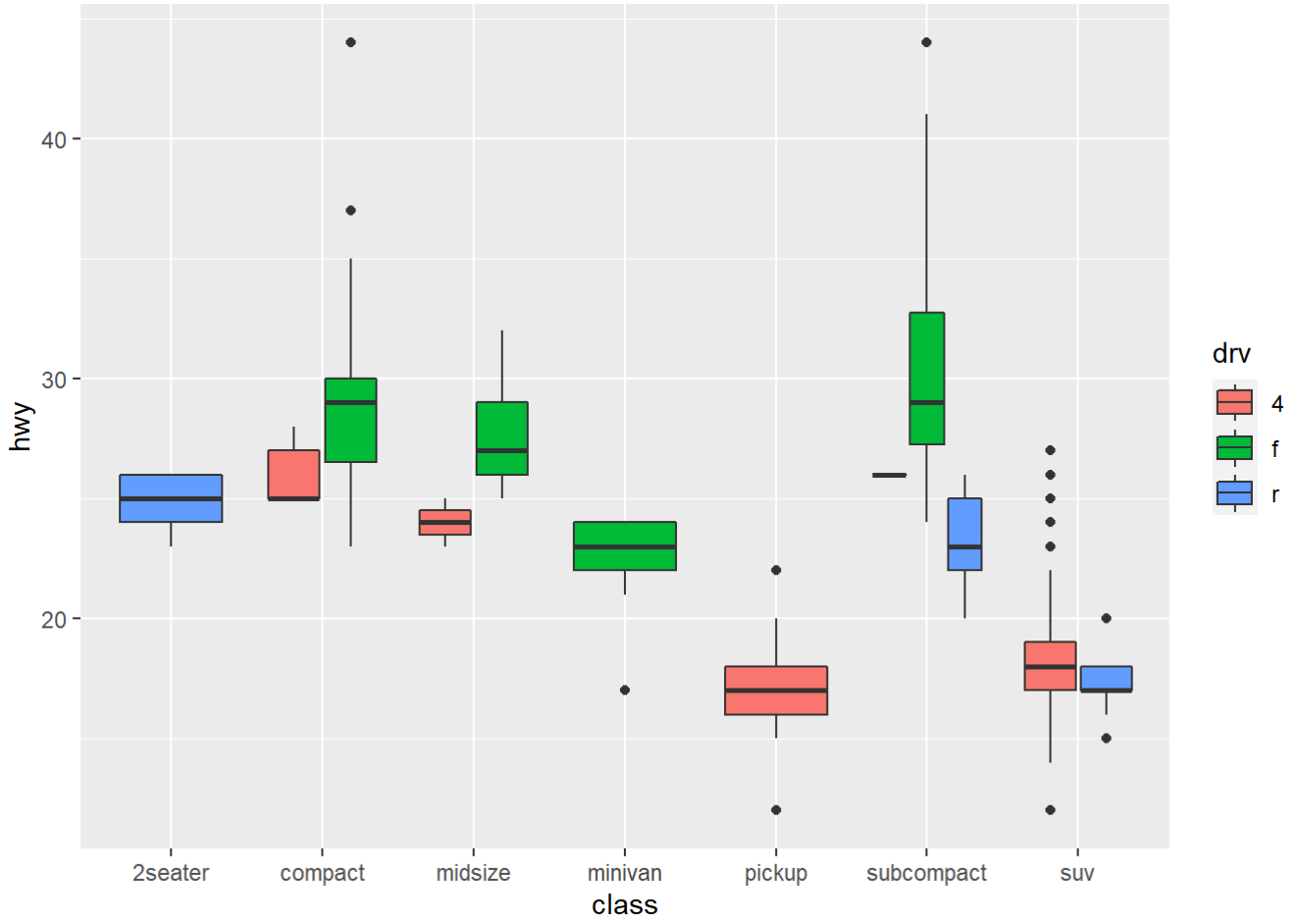
We notice that the four predictor could be divided into two groups which contains hwy and cyl, cty and hwy respectively.

One thing surprises me is that displ is negatively correlated with cty, cty is negatively correlated with cyl but hwy is positively correlated with cyl. To some degree, I believe this reveals some algorithm like two negatives make a positive.

```
## Graphic recreation
#Exercise 6
library(ggthemes)
p = ggplot(mpg, aes(class,hwy))
p + geom_boxplot()+geom_dotplot(binwidth = 1,alpha=0.2,stackratio=0.1,method="dotdensity",binax
is = "y",fill="black",stackdir = "centerwhole",dotsize=0.3,position = position_jitter(width=0.3
,height=0.001))+coord_flip()+theme_bw()+
  labs(x = "Highway MPG",y="Vehicle Class")
```

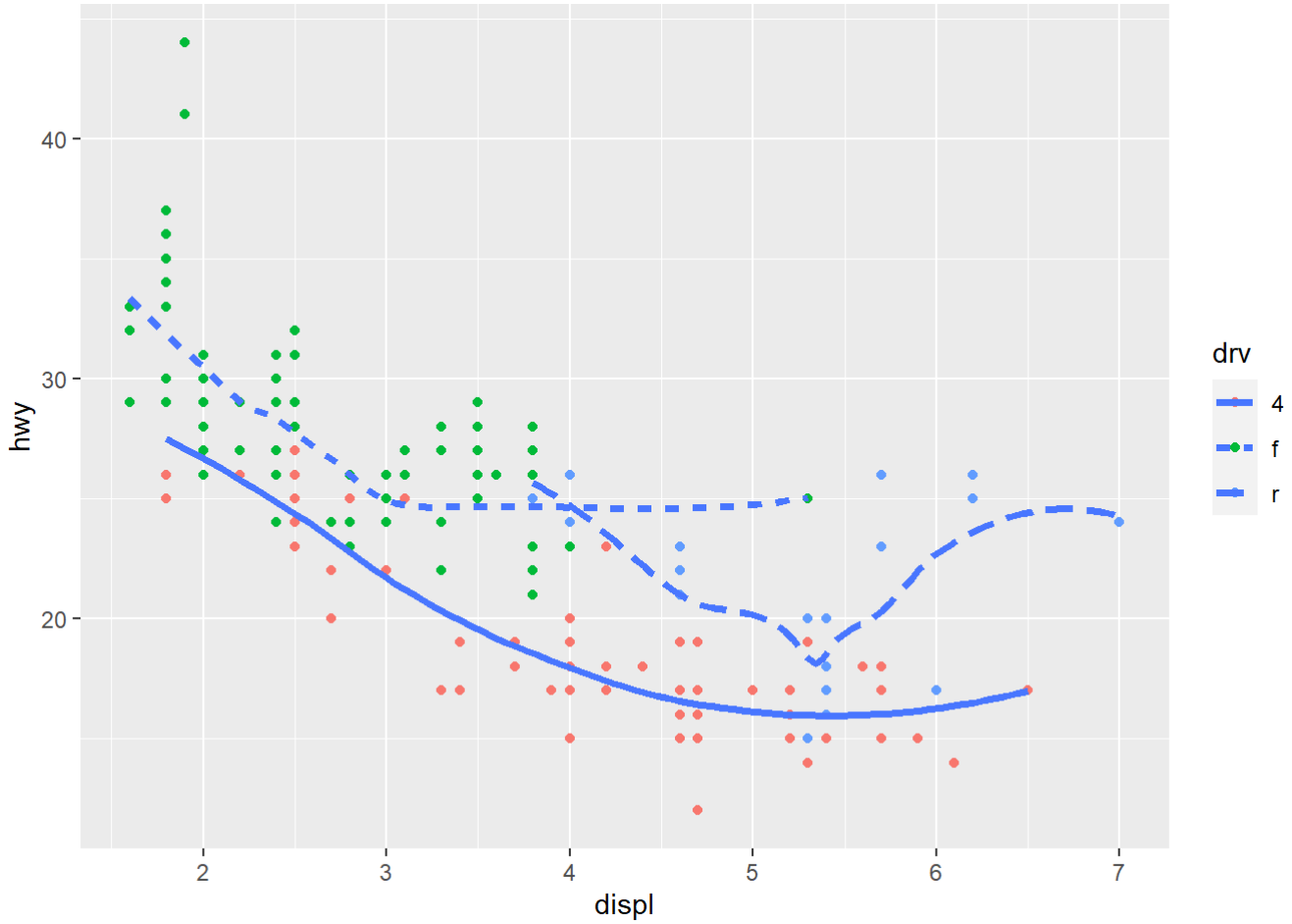



```
## Exercise 7
p = ggplot(mpg, aes(class,hwy))
p+geom_boxplot((aes(fill=drv)))
```




```
##Exercise 8
ggplot(mpg, aes(displ, hwy, colour = drv,linetype=drv)) +
  geom_point() +
  geom_smooth(size=1.3,se = FALSE,colour="royalblue1")
```

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


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