PSTAT 231 HW1

AO XU

2022-09-27

Machine Learning Main Ideas

Problem 1:

(a) Supervised learning is a machine learning approach that's defined by its use of labeled datasets.

(b) Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets.

(c) The main distinction between the two approaches is the use of labeled datasets. To put it simply, supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not.

Problem 2:

(a) While regression helps predict a continuous quantity, classification predicts discrete class labels.

(b) For Regression, Y is quantitative and numerical values; for Classification, Y is qualitative (From lecture slides).

Problem 3:

(a) For regression ML problems: Mean Squared Error, Mean Absolute Error

(b) For classification ML problem: F1 Score, AUC-ROC

Problem 4:

(a) Descriptive models: Choose model to best visually emphasize a trend in data;

(b) Predictive models: Aim is to predict Y with minimum reducible error, and not focused on hypothesis tests;

Inferential models: Aim is to test theories, (Possibly) causal claims and State relationship between outcome & predictor(s) (From lecture slides)

Problem 5:

(a) Mechanistic predictive models: uses a theory to predict what will happen in the real world.

(b) Empirically-driven predictive models: studies real-world events to develop a theory.

(c) Difference: Mechanistic predictive models assume a parametric form for f, won't match true unknown f, and could have more flexibility by adding parameters while empirically-driven predictive models have no assumptions about f, requires a large number of observations and is much more flexible by default.

(d) Sameness: Both of them are overfitting.

(e) I think a mechanistic predictive model is easier to understand since there is a parametric form for f to model. It looks like a more direct way.

(f) The bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters. (From Google)

Problem 6:

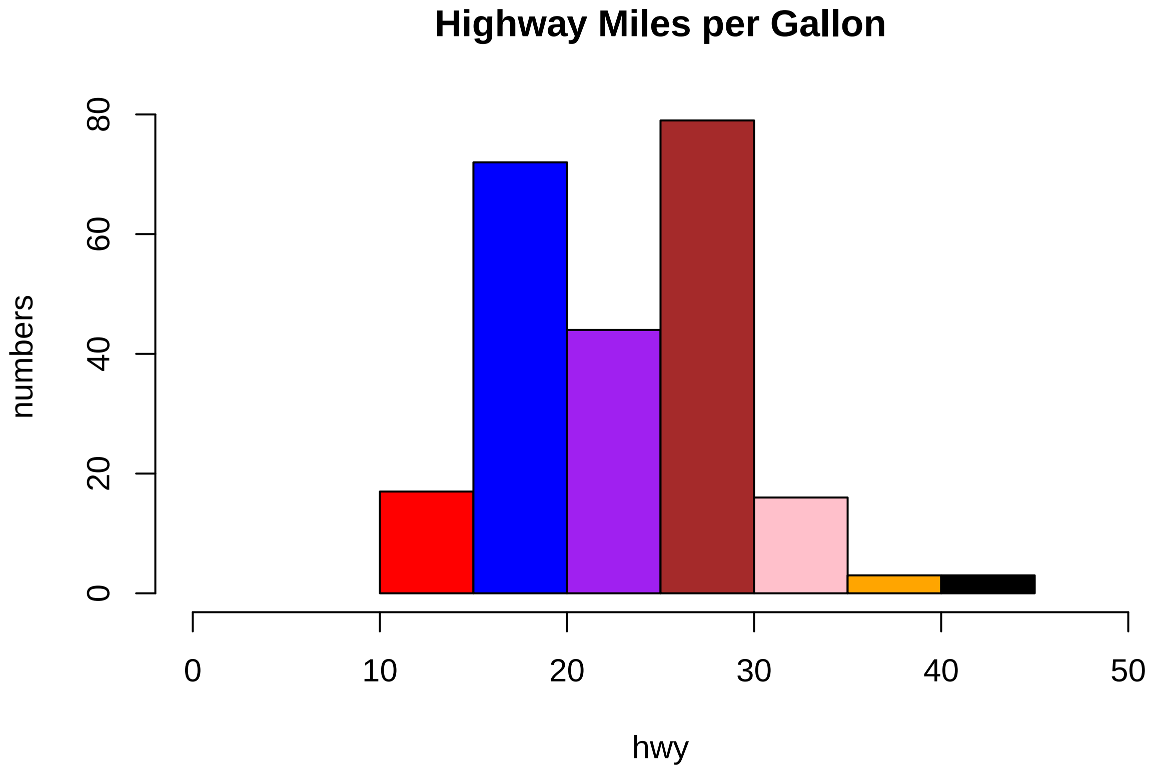
(a) For Question1, it's a predictive question since it is to predict Y;

(b) For Question2, it's an inferential question since it tests whether having the contact with candidate is important in voting choice.

Exploratory Data Analysis:

Exercise 1:

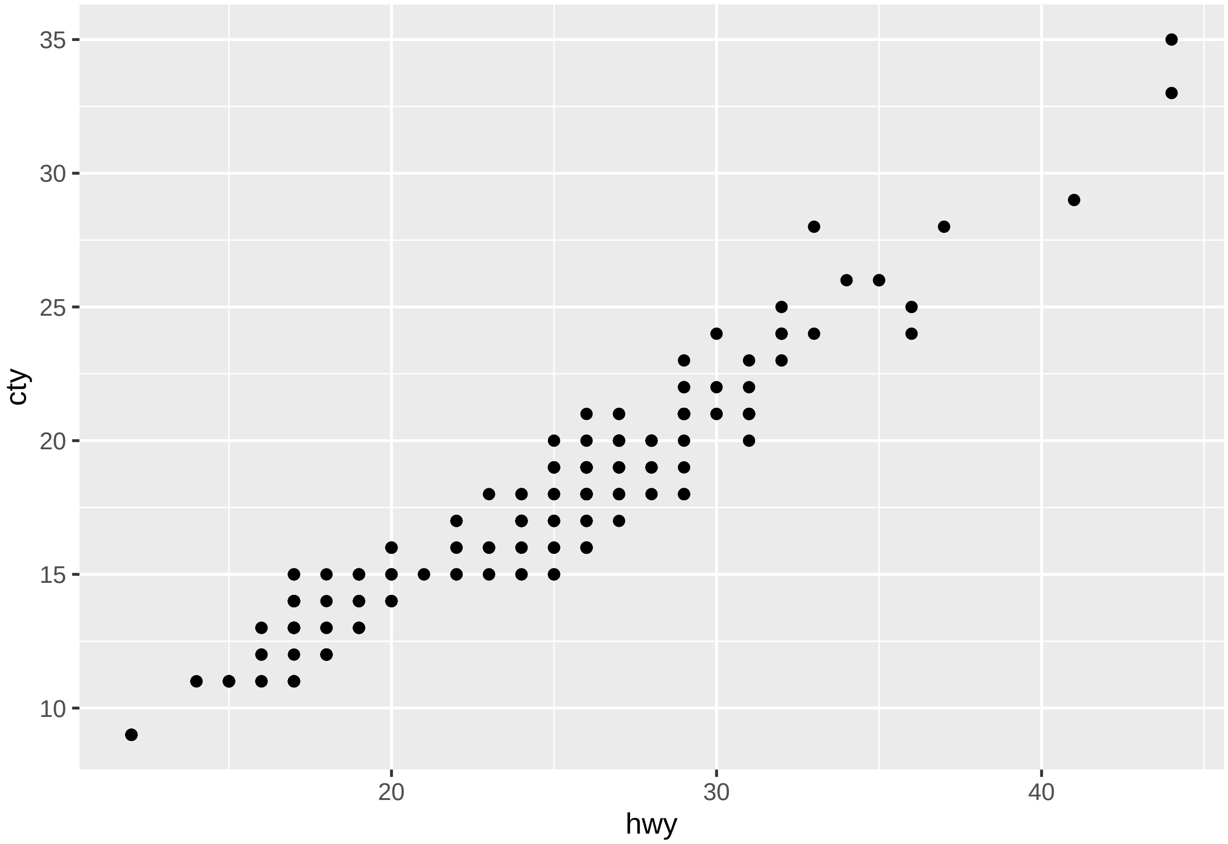
```
# histogram
library(ggplot2)
data("mpg")
colors<-c("red","blue","purple","brown","pink","orange","black")
hist(mpg$hwy, col=colors, main="Highway Miles per Gallon", breaks=7, xlim = range(0:50), xlab="hwy",ylab
```



```
# More cars are between 15-30 highway miles per gallon, fewer cars are in 10-15 and 30-45 highway miles
```

Exercise 2:

```
# scatterplot
library(ggplot2)
data("mpg")
ggplot(mpg, aes(x = hwy, y = cty)) +
  geom_point()
```



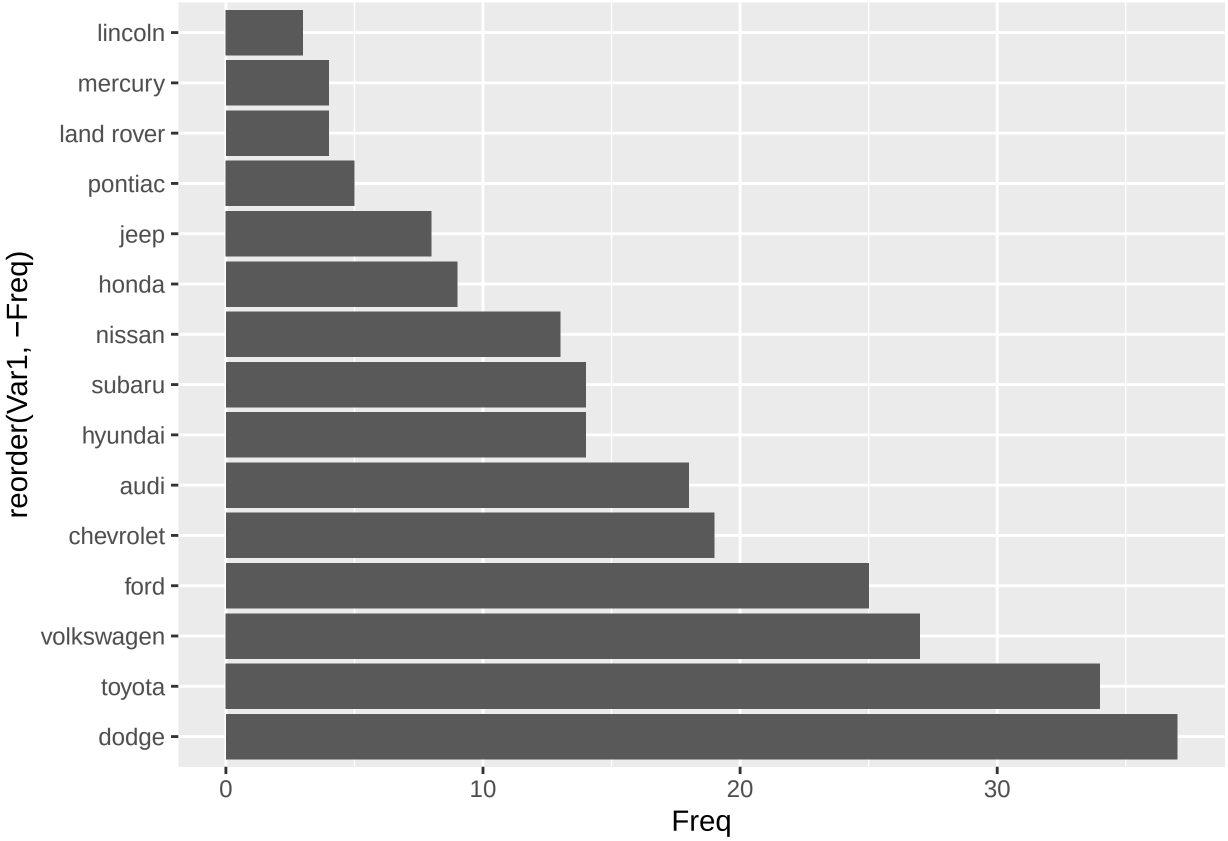
```
#I notice that as hwy increases, cty increases.
#Yes, there is a relationship between hwy and cty.
#It means that if cty is high then hwy is high and vice verca.
```

Exercise 3:

```
# bar plot
library(ggplot2)
data("mpg")
x <- ggplot2::mpg
x <- as.data.frame(table(x$manufacturer))
x$Var1 = as.character(x$Var1)
x
```

```
##          Var1 Freq
## 1        audi   18
## 2   chevrolet   19
## 3       dodge   37
## 4        ford   25
## 5       honda    9
## 6     hyundai   14
## 7        jeep    8
## 8  land rover    4
## 9     lincoln    3
## 10    mercury    4
## 11     nissan   13
## 12    pontiac    5
## 13     subaru   14
## 14     toyota   34
## 15 volkswagen   27
```

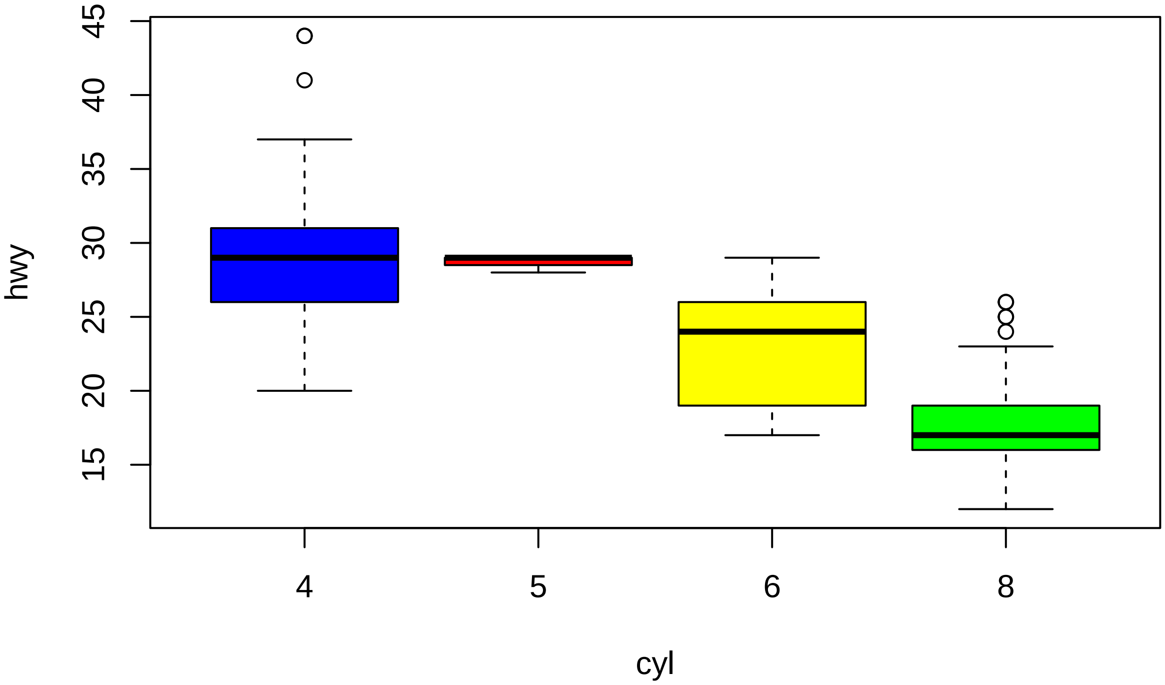
```
p2 <- ggplot(x,aes(x=reorder(Var1, -Freq),y=Freq))+geom_bar(stat='identity')+coord_flip()
p2
```



```
# Dodge produced most cars;
# Lincoin produced the least cars.
```

Exercise 4:

```
# box plot
library(ggplot2)
data("mpg")
boxplot(hwy ~ cyl, data = mpg,col = c("blue", "red", "yellow","green"))
```



```
# Yes, I see a pattern that generally when cyl is higher, then hwy is lower.
```

Exercise 5:

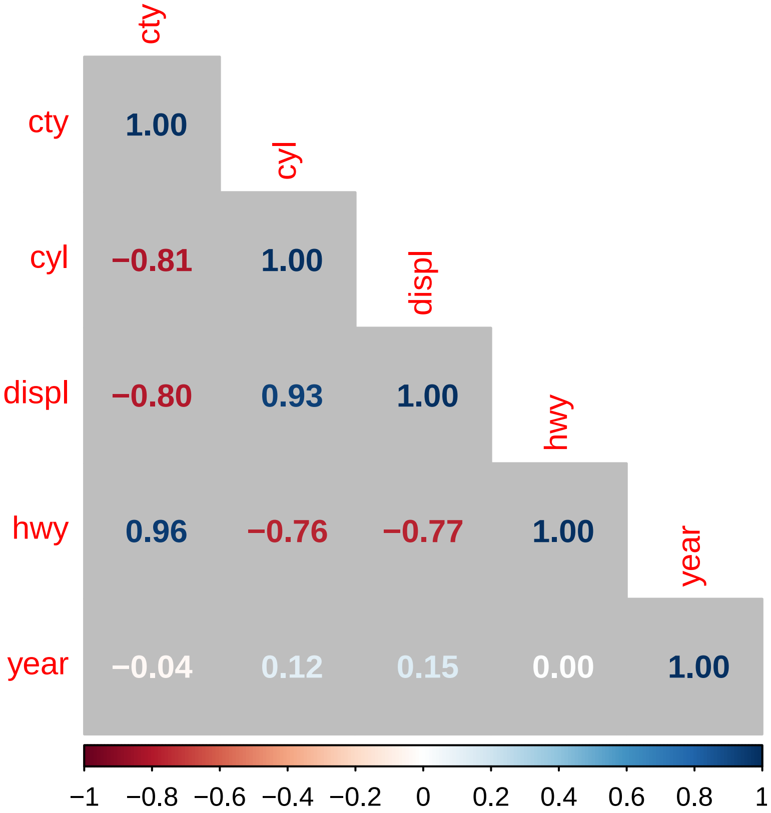
```
#install.packages('plyr', repos = "http://cran.us.r-project.org")
#install.packages("corrplot")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

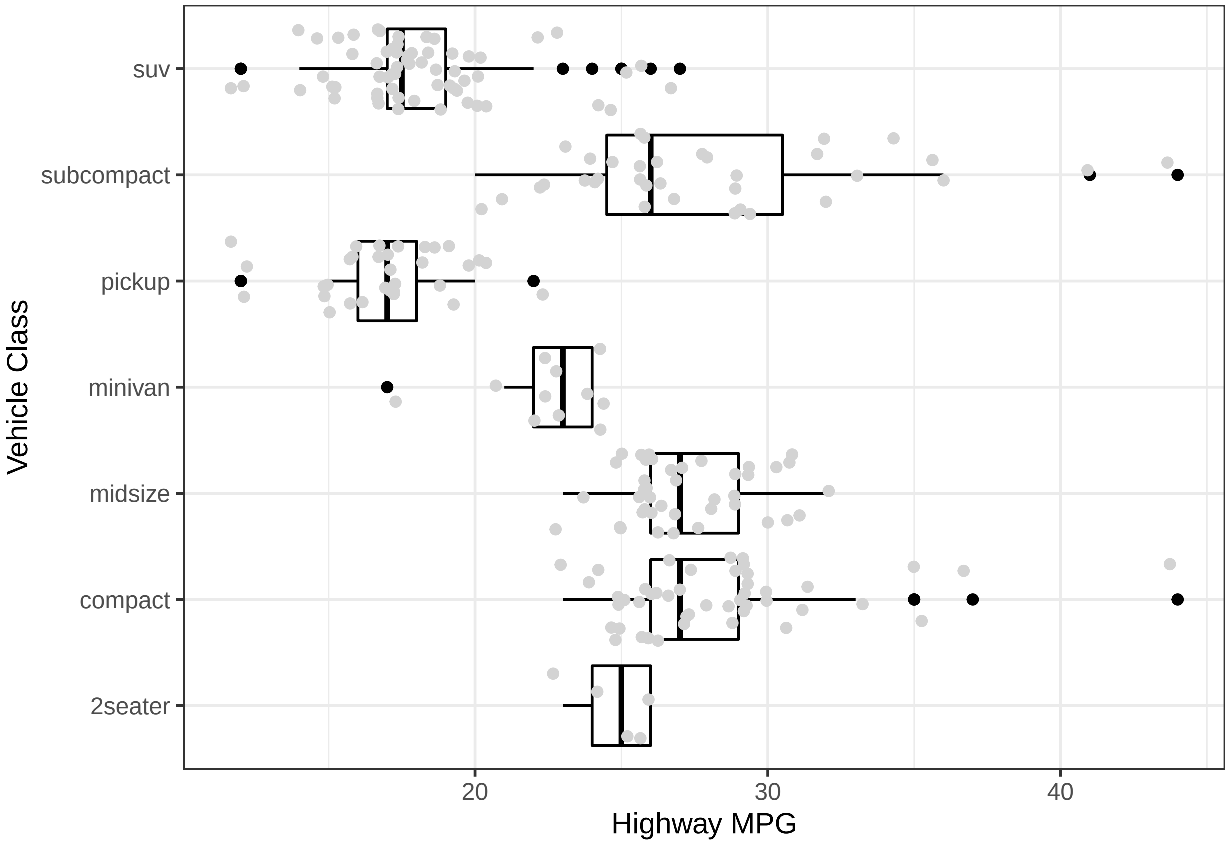
```
Matrix <- ggplot2::mpg %>% select_if(is.numeric) %>% cor(.)
corrplot(Matrix, method='number',type="lower",order='alphabet',bg="grey")
```



```
# (1)displ and cyl, (2)hwy and cty, (3)year and cyl, and (4)year and displ have postive relationships.
# (1)cyl and cty, (2)displ and cty, (3)hwy and cyl, and (4)hwy and displ have negative relationships.
# Yes, there relationship make sense to me since I think hwy and cty should be highly correlated.
# One thing surprises me is that year and hwy have no relationship.
```

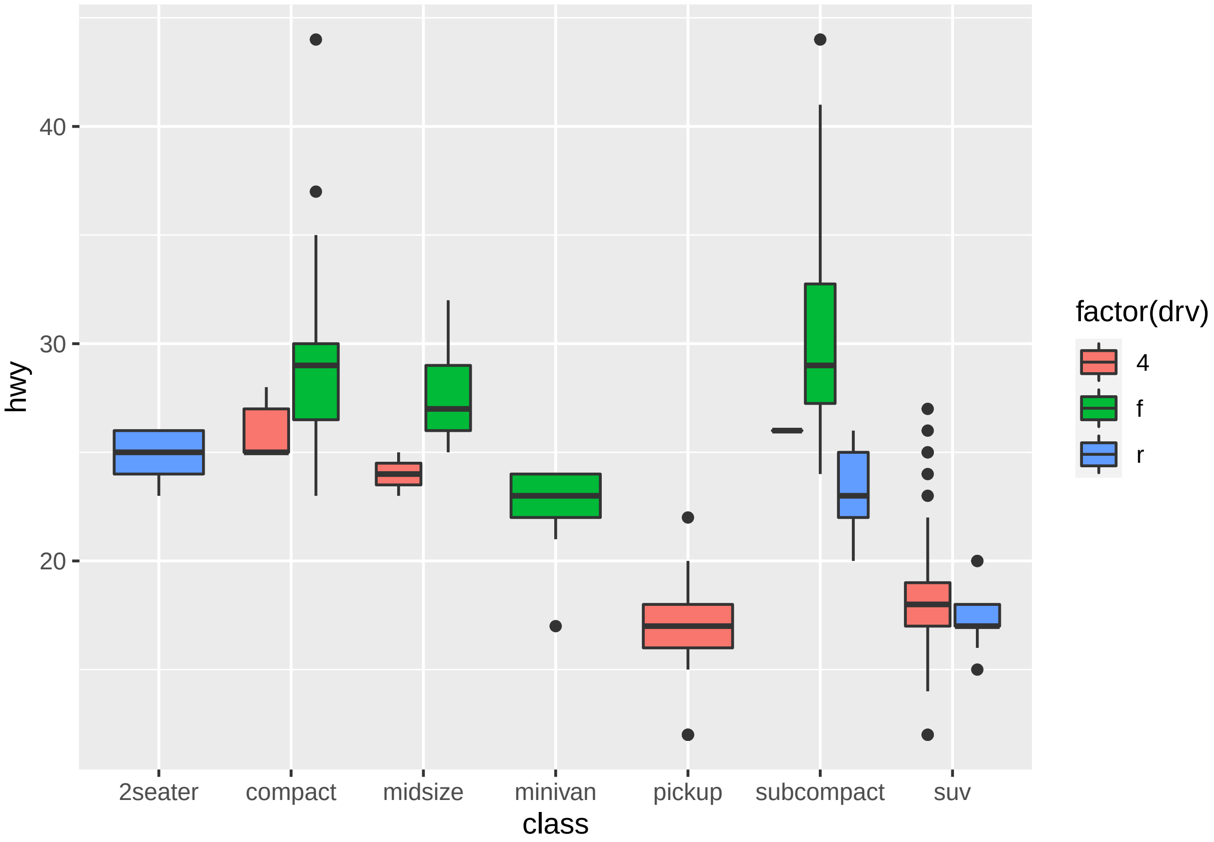
Exercise 6:

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



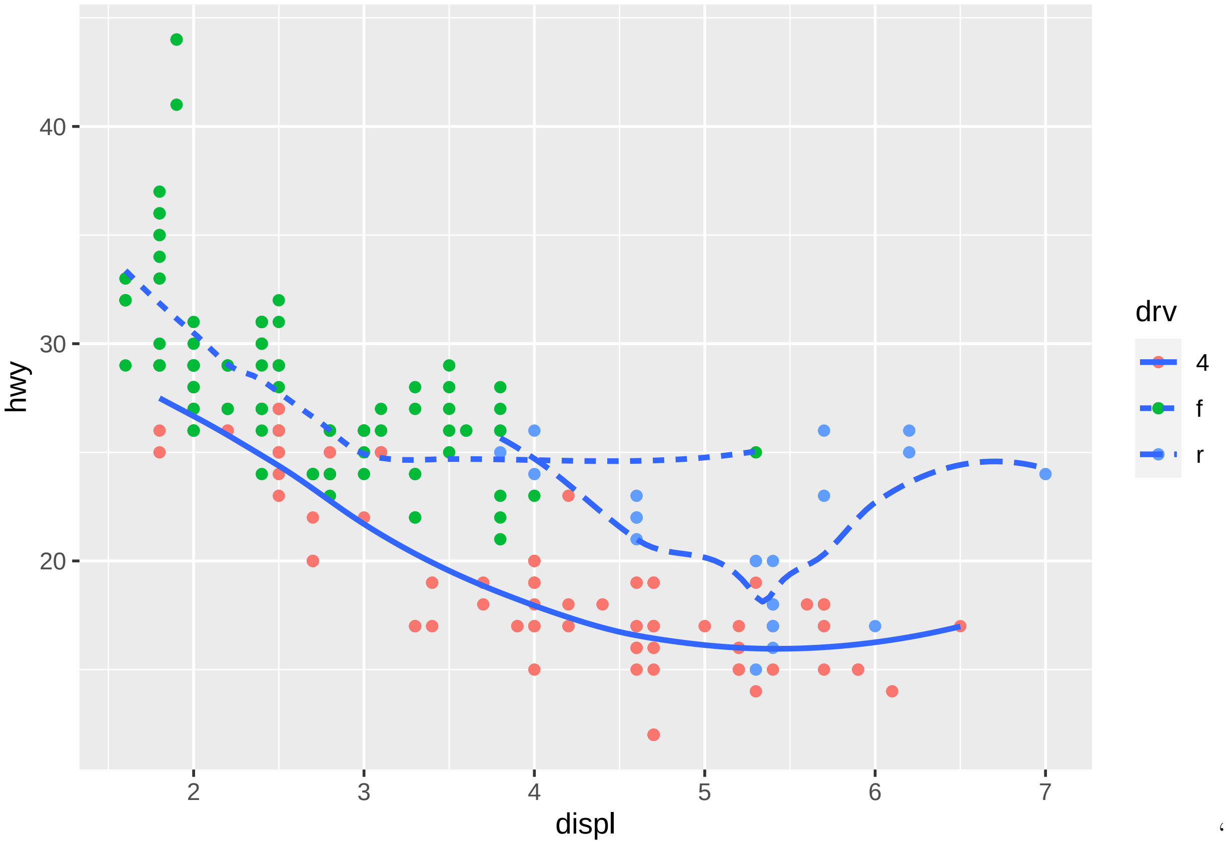
Exercise 7:

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



Exercise 8:

```
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

‘‘‘