PSTAT 231 HW1

Zongyi Han

2022-09-24

Answer1.

Supervised learning is one-to-one maps 1 inputs to 1 output
Unsupervised learning can discover patterns in data sets without human intervention(labeling)
The difference between them is that supervised learning needs labeling but unsupervised learning *doesn't* need labeling.

Answer2.

Regression takes Quantitative data
Classification takes qualitative data

Answer3.

For Regression ML, the metrics are MSE&RMSE
For Classification ML, the metrics are F-1 score and AUC-ROC

Answer4.

Descriptive model: Chose model best emphasize trend visually
Inferentialmodel: To test theories, state relationship between outcome and predictor
Predicative model: Predict Y with minimal error

Answer5.

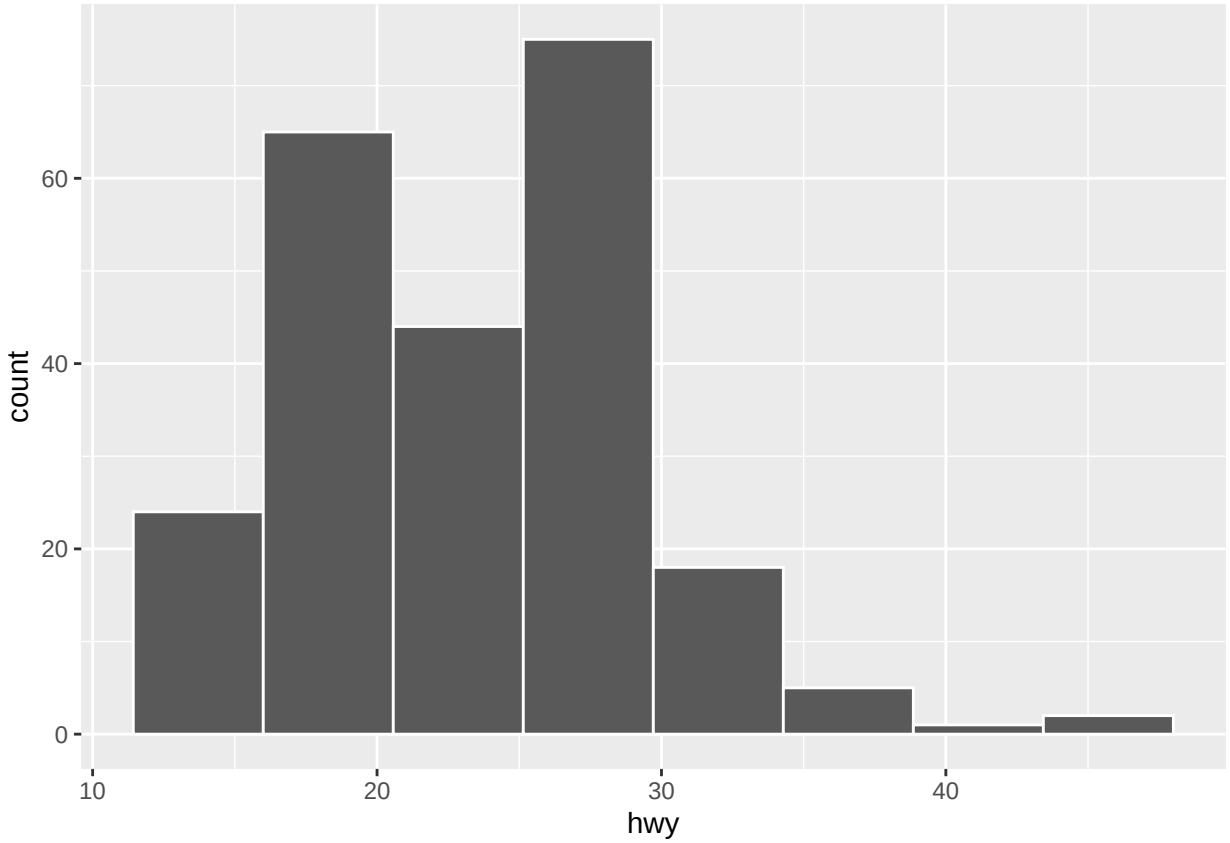
-Mechanistic is parametric
Empirically-driven is non-parametric.
-Mechanistic has less flexibility and needs assumptions. Latter does not need those things
both of them can be over fitting.
-Mechanistic can be easier to be understood b/c it has less flexibility.
-Bias-Variance trade off depends on flexibility of the methods, higher flexibility means low bias-variance trade off. One can expect mechanistic to have higher bias compare to empirically-driven model.

Answer6.

-Inferential. Assume voters in favor of candidate then use informational method to test if we accept H_0
-Predictive, b/c there is no assumption to be made.

Exercise 1

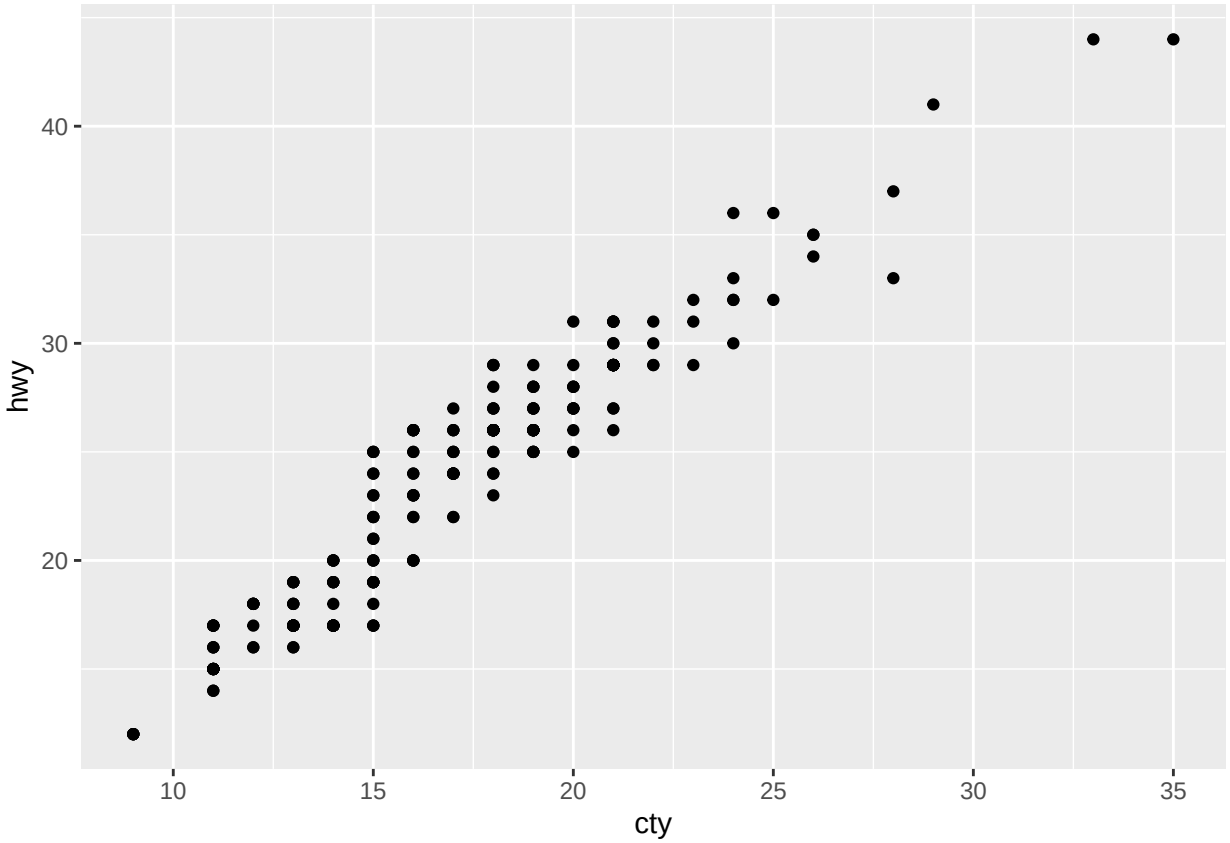
```
library(ggplot2)
ggplot(mpg, aes(hwy)) +
  geom_histogram(bins=8,color="white")
```



Most cars have highway fuel efficiency around 20~30mpg

Exercise 2

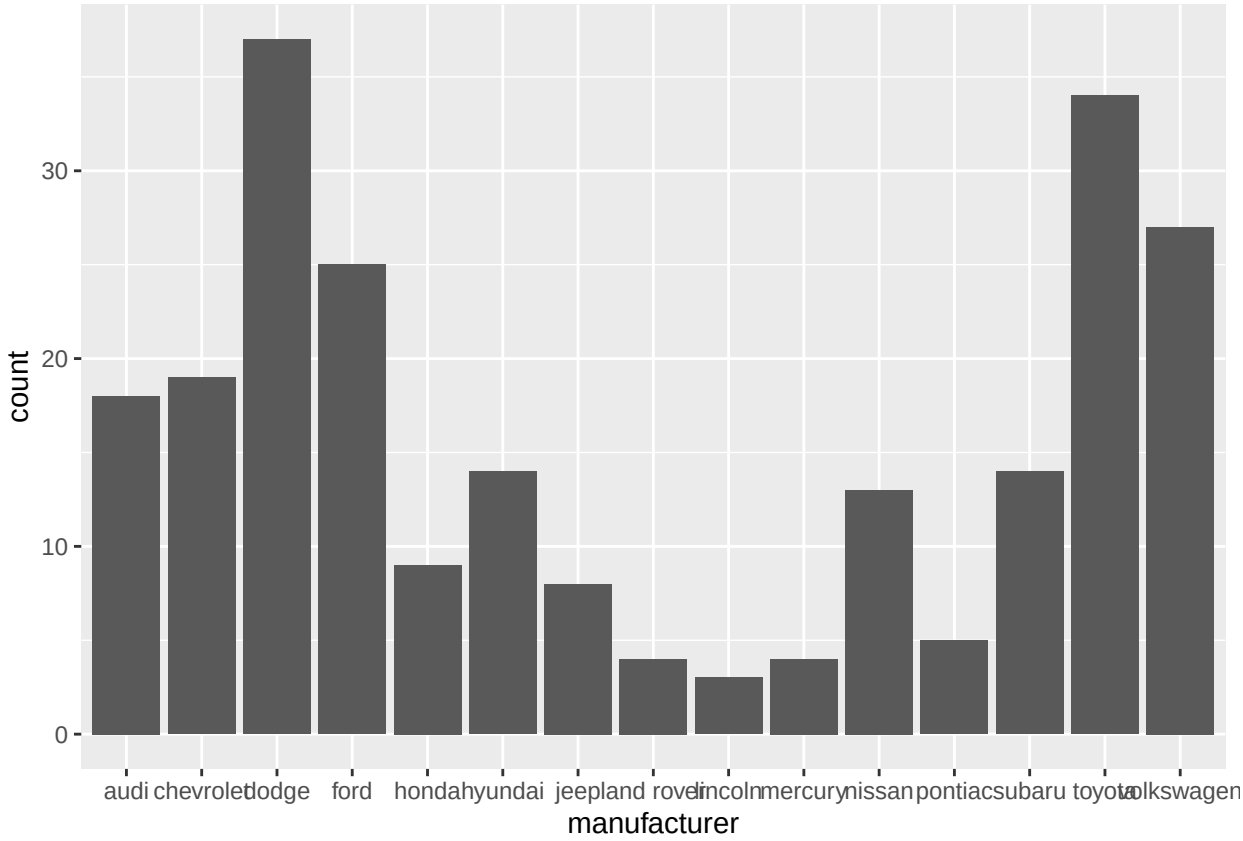
```
library(ggplot2)
ggplot(data = mpg) +
  geom_point(mapping = aes(x = cty, y = hwy))
```



higher city fuel efficiency higher highway fuel efficiency

Exercise 3

```
ggplot(data = mpg) +
  stat_count(mapping = aes(x = manufacturer))
```



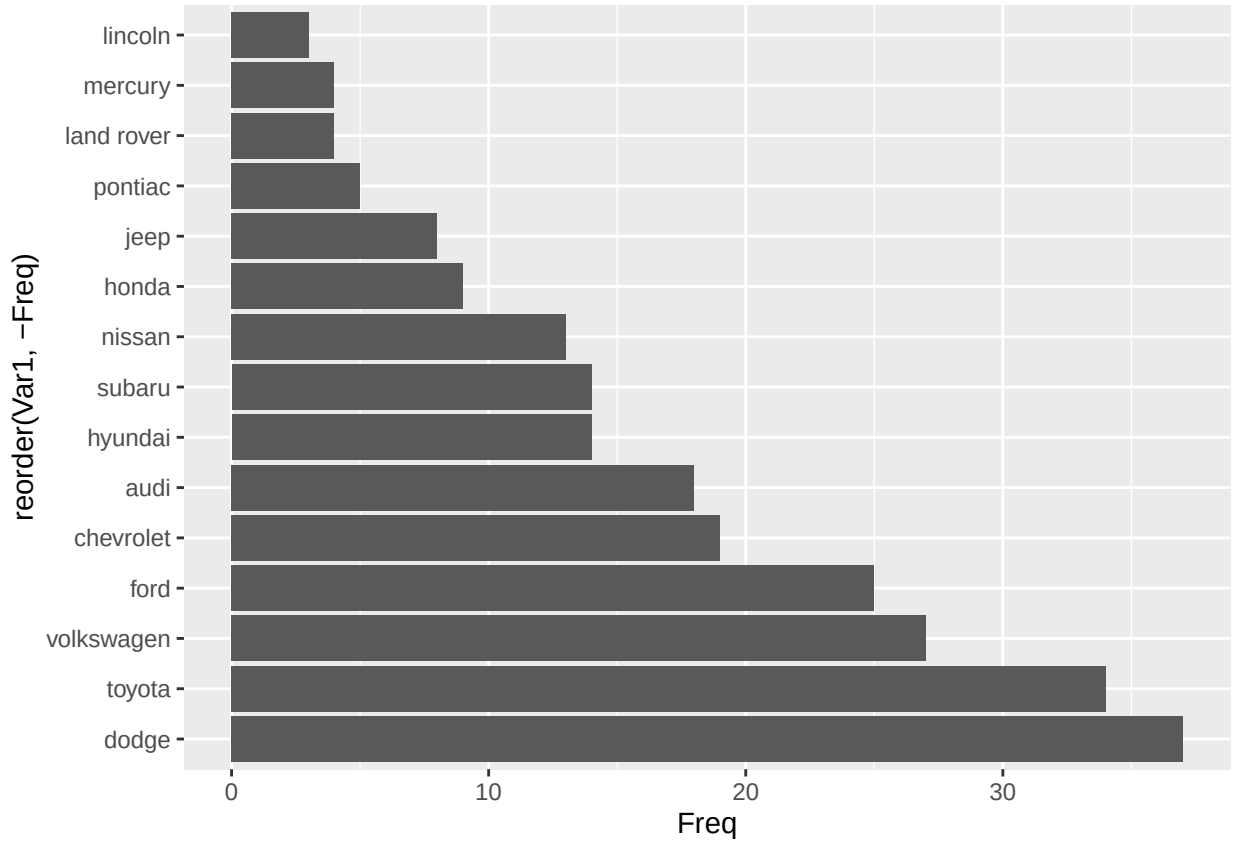
Exercise 4

```
a <- ggplot2::mpg
a <- as.data.frame(table(a$manufacturer))

a$Var1 = as.character(a$Var1)
a
```

```
##          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
```

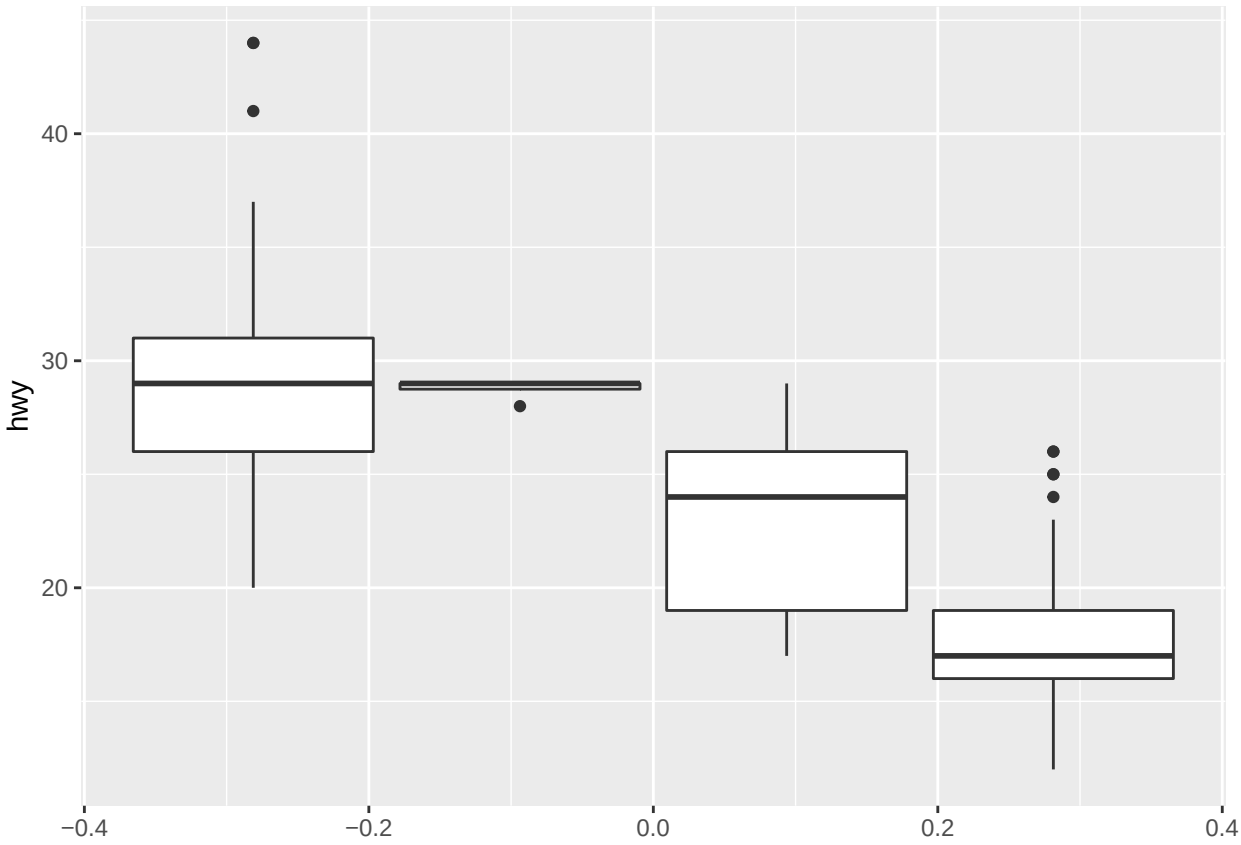
```
p2 <- ggplot(a, aes(x = reorder(Var1, -Freq), y = Freq)) +
        geom_bar(stat = 'identity')+coord_flip()
p2
```



Dodge makes most car

Exercise 5

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



less number of cyl , higher mpg on highway

Exercise 6

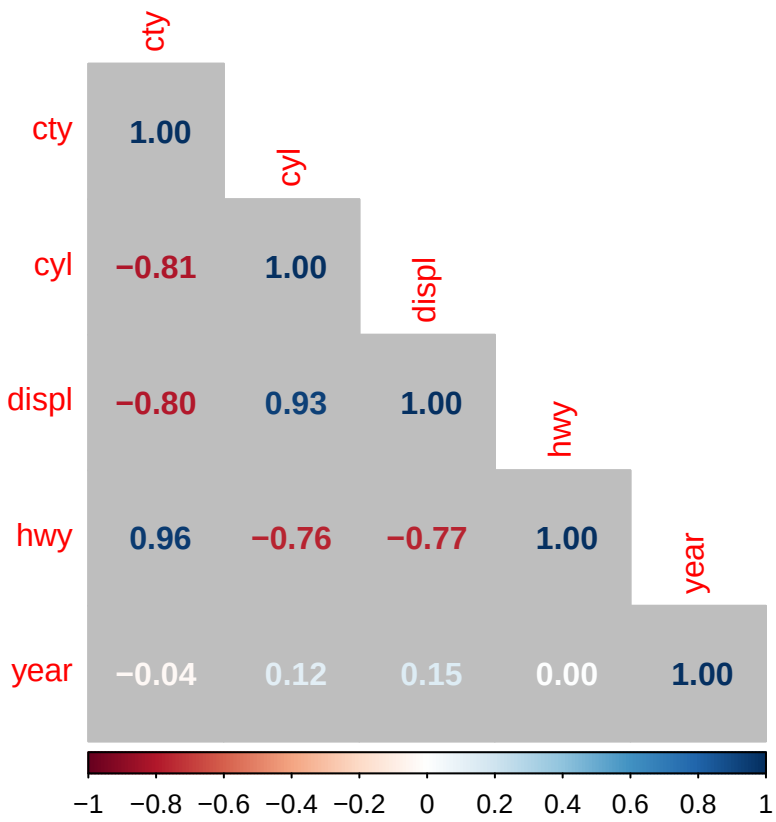
```
library(tidyverse)
```

```
## -- 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()
```

```
Matrx <- ggplot2::mpg %>%
    select_if(is.numeric) %>%
    cor(.)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(Matrx, method = 'number',type="lower", order = 'alphabet',bg = "grey")
```



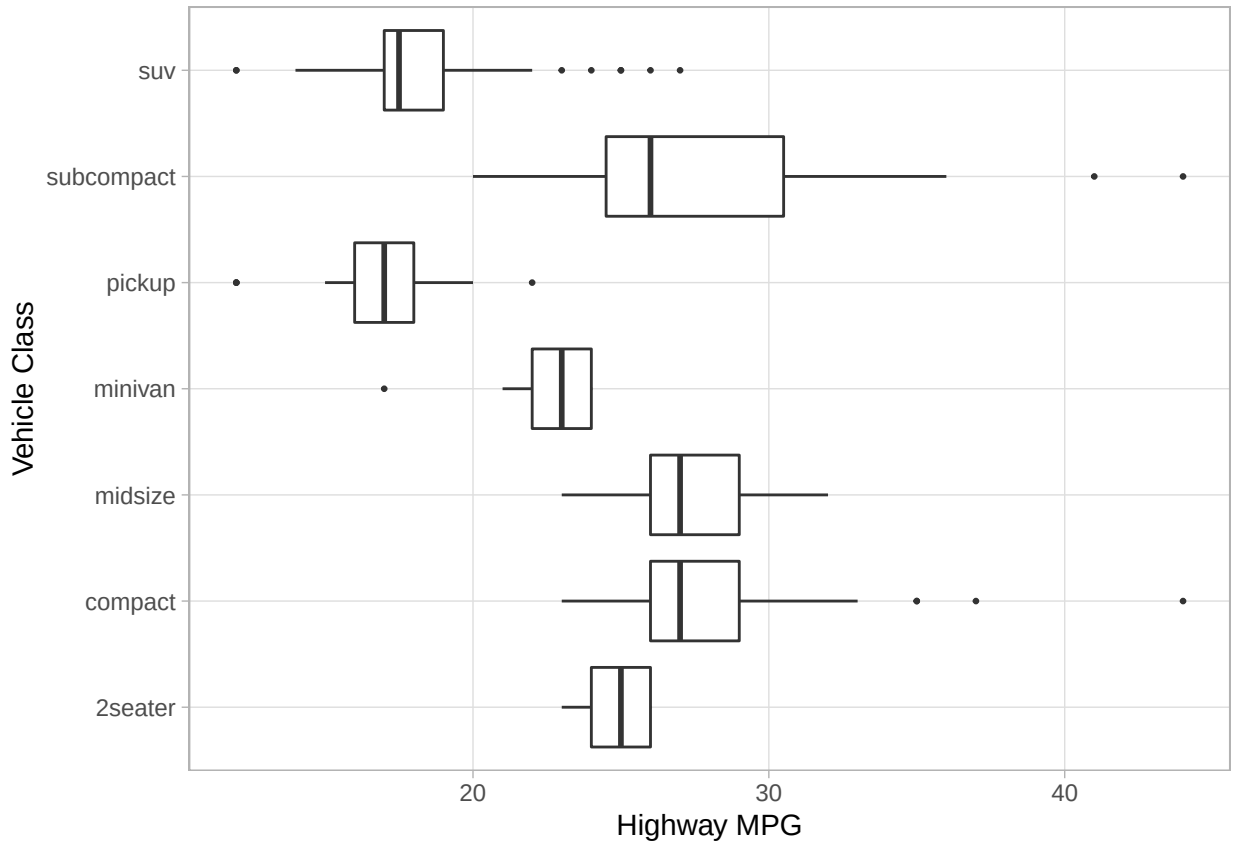
hwy are positively correlated with cty displ is positively correlated with cyl year is positively correlated with cyl and displ

cyl is negatively correlated with cty displ is negatively correlated with cty hwy is negatively correlated with cyl and displ year is negatively correlated with cty

These relationship make sense to me as it follows law of physics. No superise here.

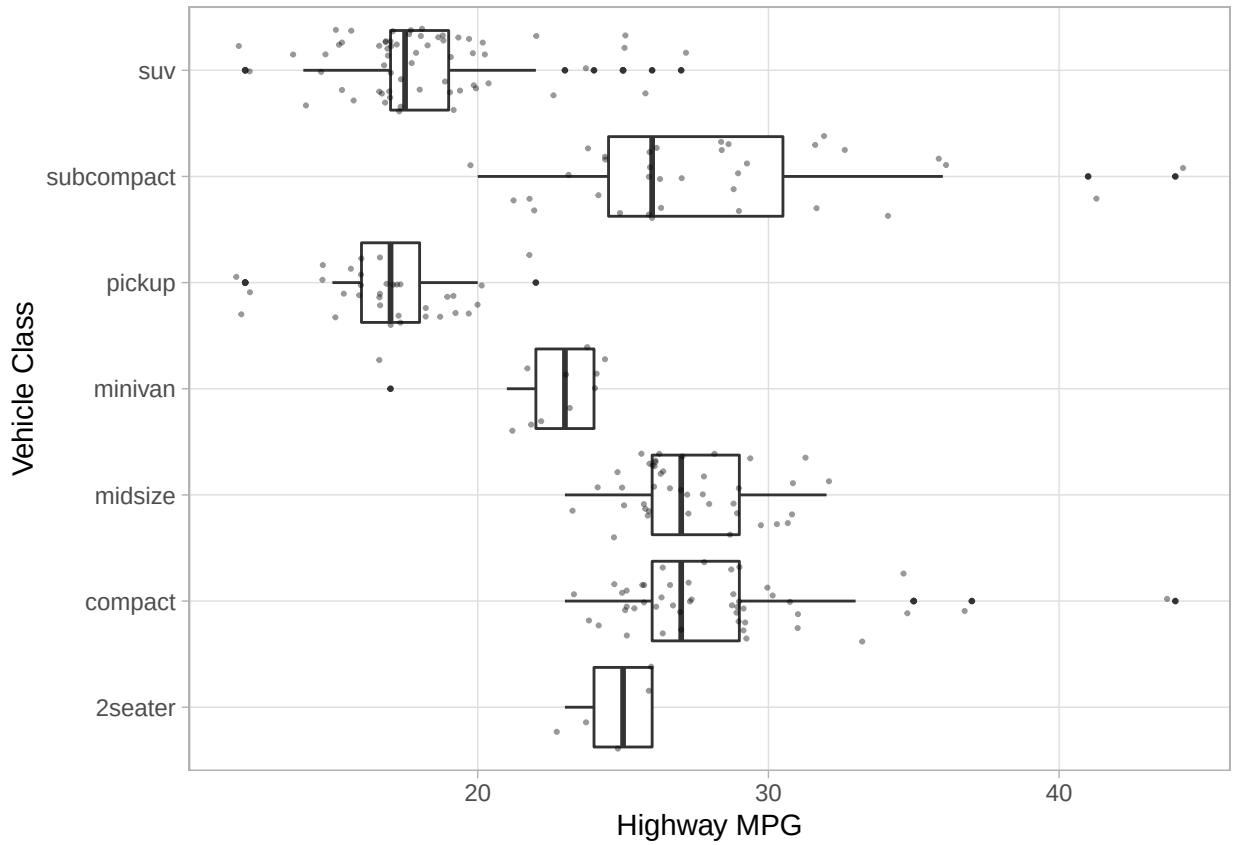
Exercise 7

```
ggplot(data = mpg, mapping = aes(x=hwy, y = class)) +
  geom_boxplot(outlier.size = 0.5)+
  #geom_point( position = position_jitterdodge(self))+
  theme_light()+
  xlab("Highway MPG")+
  ylab("Vehicle Class")+
  theme(panel.grid.minor = element_blank())
```



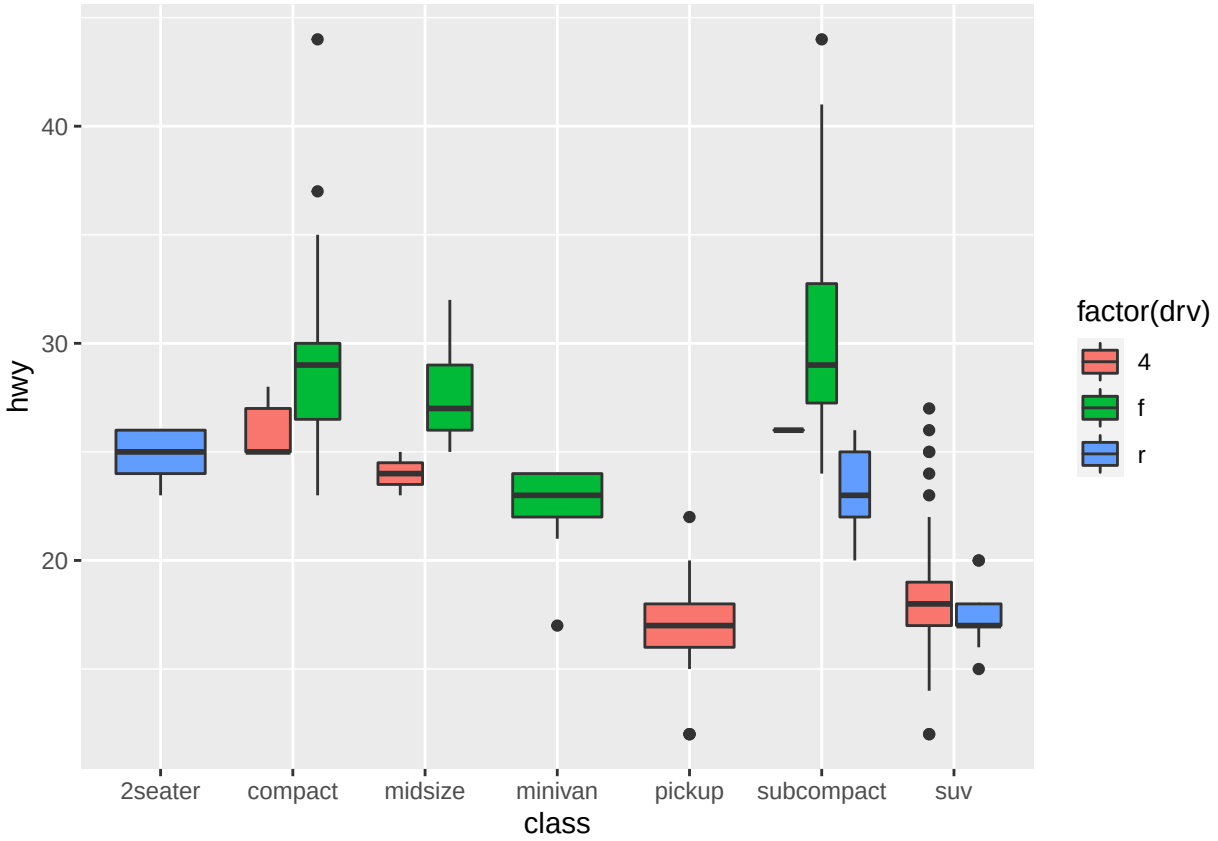

```
ggplot(data = mpg, mapping = aes(x=hwy, y = class)) +
  geom_boxplot(outlier.size = 0.5)+
  theme_light()+
  geom_jitter(color="black", size=0.4, alpha=0.4,stackdir = 'center')+
  xlab("Highway MPG")+
  ylab("Vehicle Class")+
  theme(panel.grid.minor = element_blank())
```

```
## Warning: Ignoring unknown parameters: stackdir
```



Exercise 8

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

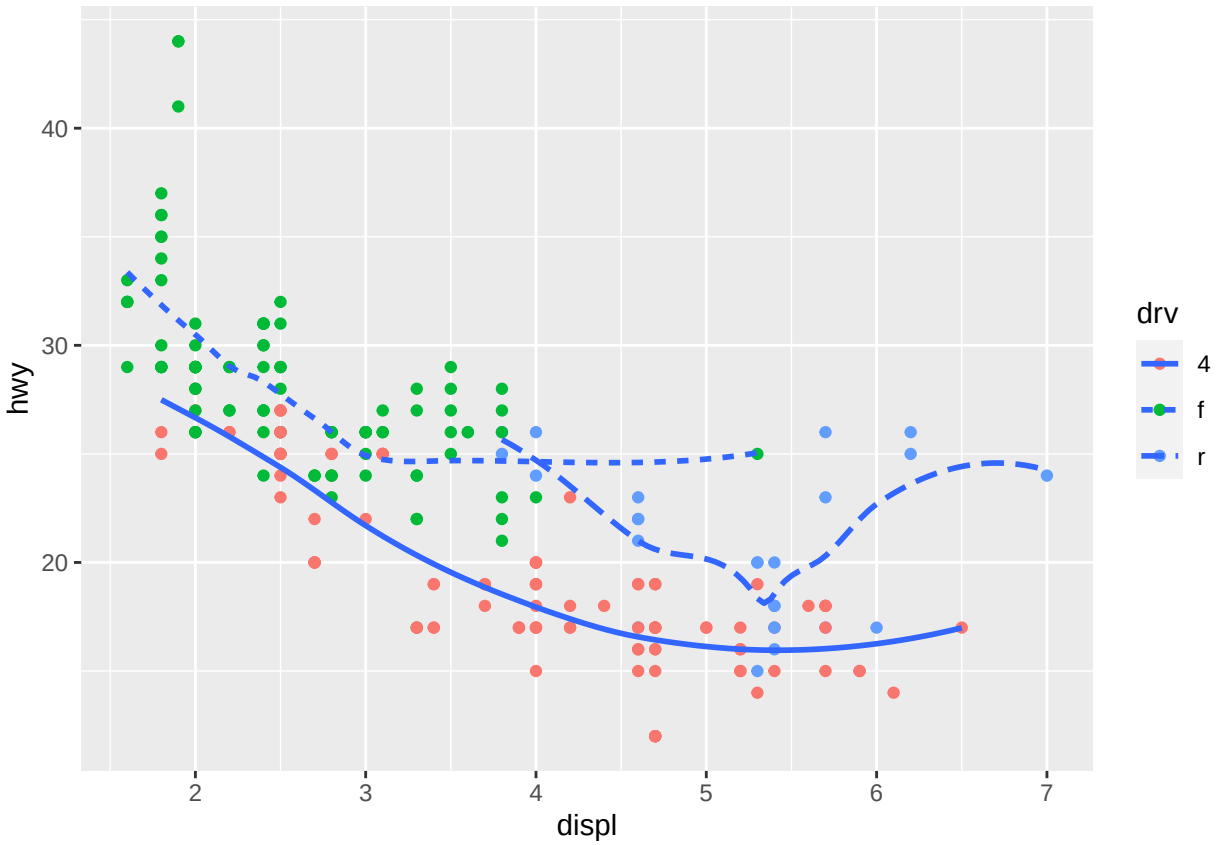



```
#cite from https://ggplot2.tidyverse.org/reference/position_dodge.html with modification
```

Exercise 9

```
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point(aes(colour = drv)) +
  geom_smooth(aes(linetype = drv), se = FALSE)
```

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


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