Question 1

- Supervised learning is the learning to predict a response given a range of predictor
- Unsupervised learning is the learning of the patterns or features of the response given a range of predictors
- The difference is that unlike, supervised learning, Unsupervised learning has no measure of the response, or the output (ESL #xi, ISLR p#26)

Question 2

Regression aims to find out a quantitative outcome with to a continuous measure of the data.

Classficiation gives a qualitative outcome with a qualitative measure of the data (ISLR p#28)

Question 3

Regression: 1. Least Square Linear Regression 2. Polynomial Regression

Classification: 1. K-clustering 2. Hiercharchical clustering (Lecture)

Question 4

- Descriptive models: models that emphasize a visual trend in the data such as a line on a scatterplot
- Predictive models: models that select the best combination of parameters to predict a response with the least minimum reducible error. Models that isn't focusing on hypothesis testing
- Inferential models: models that analyze the weights of selected features in a theory and reveal the relationship between predictor and outcome (can be causal) (Lecture)

Question 5

a

Mechanistic is to hypotheize a relationship betweeen the variables (parameters) and the outcome.

Empiricially-driven is to find out the relationship that best fit the true model solely based on past collected data (no prior assumption)

Differences:

1. Mechanistic has assumptions. Empiricially-driven has no assumptions
2. Mechanistic can be restricted so that it doesn't accurately reflect the true model as it is predefined on assumption, empirically-driven is more flexible if the data is large enough
3. Mechanistic requires less data as the parameters are usually simplified than the true model. Empiricially-driven requires more data (Lecture)

b

Mechanistic is easier to understand.

Since the assumption used by Mechanistic models are usually based on the known-property of the dataset. For example, we can predict the if a person is health or not by giving a range of known variables related to the health such as age, medical record, family health history These assumed factors have very natural biophysical relationship to the outcome health, thus easier to understand (Lecture)

c

Bias-variance tradeoff is that lowering the variance maynot always reduce the bias between the prediction and the true fit. In the Mechanistic model with selected parameters, the model usually will not overfit as much as a empirically-driven model. However, it doesn't necessary means that the model will not underfit, which means it doesn't include enough parameters to describe the relationship In the Empiricially-driven model with more flexibility, the model may be overfitting. Even though the training MSE is reduced, the testing MSE is not, since the model is only based on training data, unlimited testing data may produce a less accurate result. (Lecture)

Question 6

1. Predictive, the first question aims to predict the voting probability without focusing on what are the incentives to vote:want
2. Inferential, the team want to testing the hyothesis that if personal contact has anything to do with the voting probability

---

EDA

```
library(tidyverse)
```

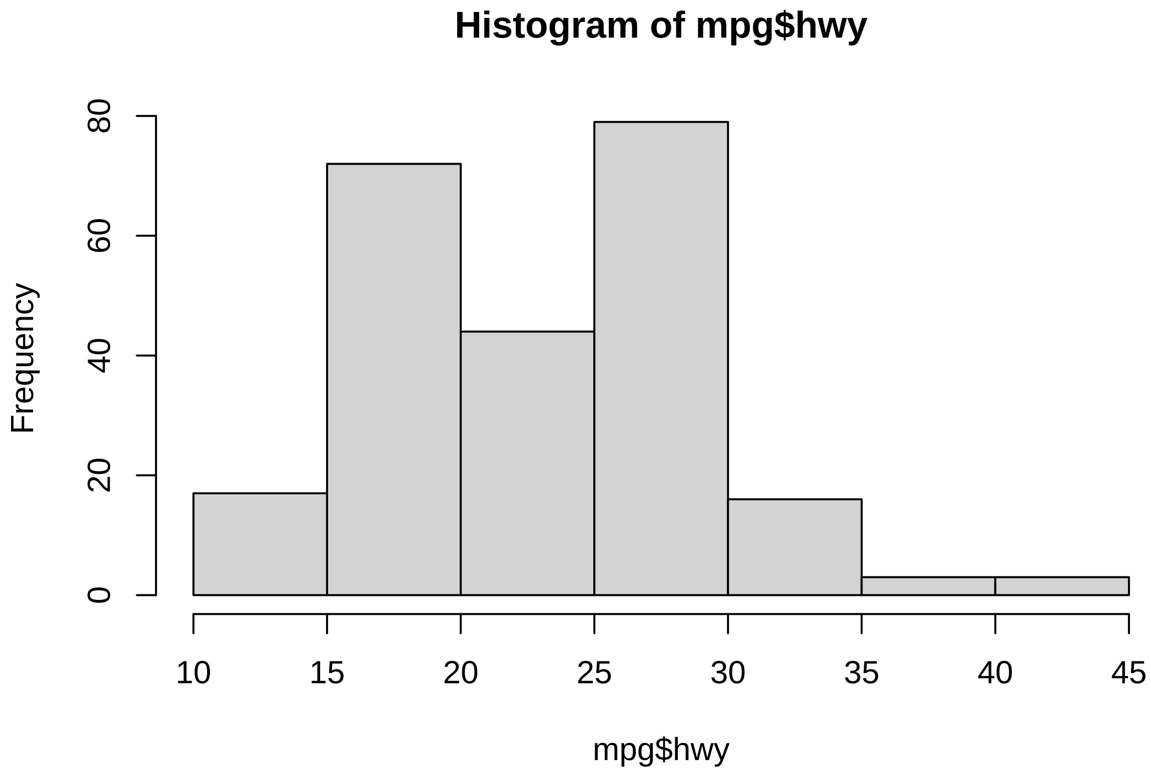
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
head(mpg)
```

```
## # A tibble: 6 x 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~
```

Ex 1

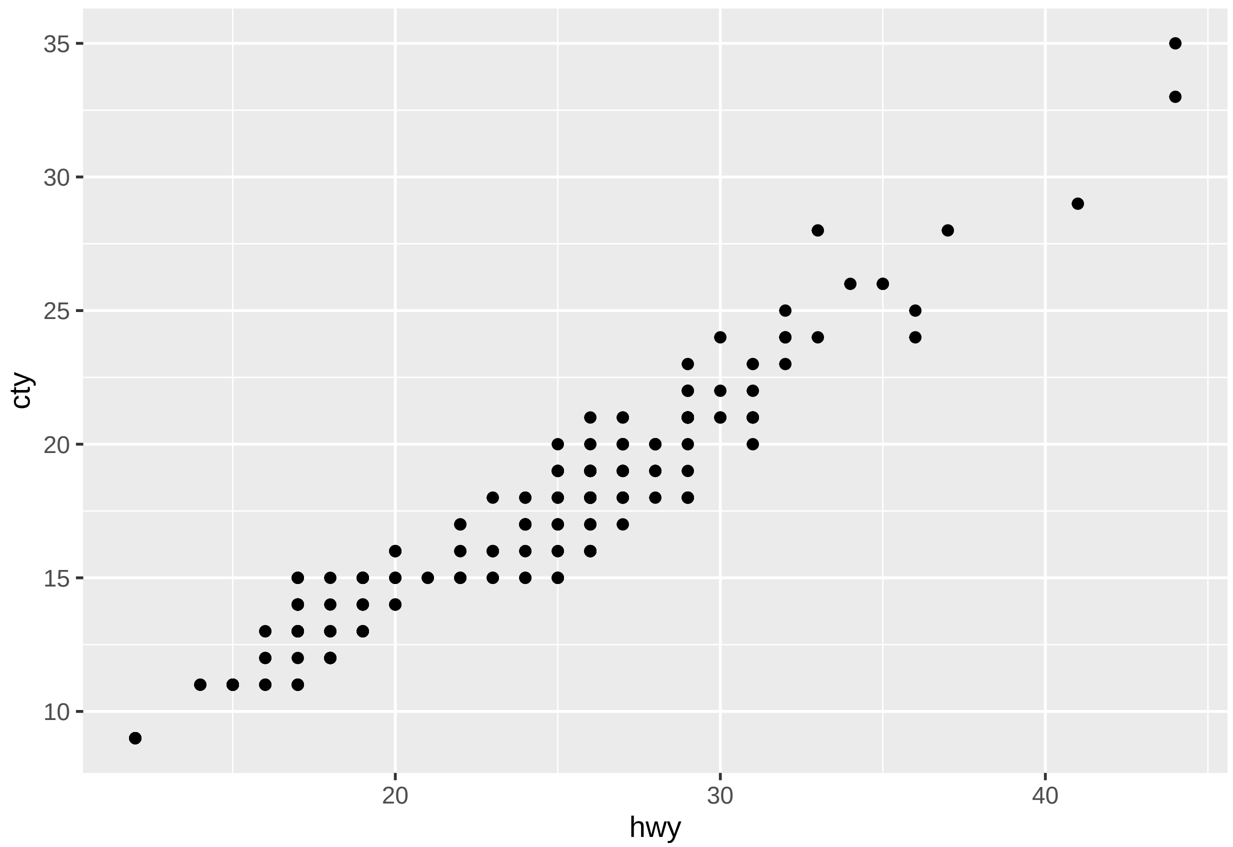
```
hist(mpg$hwy)
```



The histogram shows a right skewed distribution where most of the fuel economy is spread at the same as or lower than the medium suggesting that most cars have less than 30 mpg on hwy

Ex 2

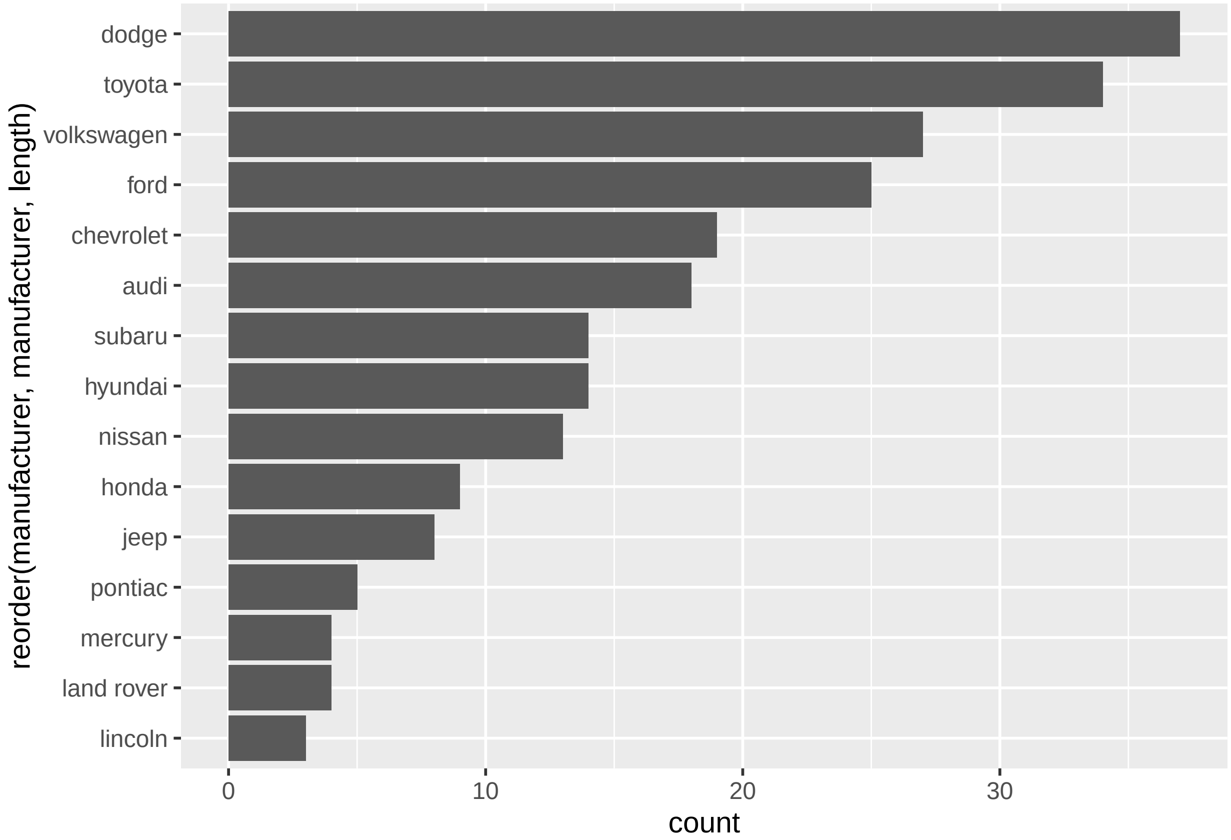
```
ggplot(mpg, aes(x = hwy, y = cty)) +
    geom_point()
```



The plot shows a clear trend that increase in hwy mpg will increase cty mpg. This suggests that hwy and cty mpg have a positive linear relationship, and is usually true is real-life.

Ex 3

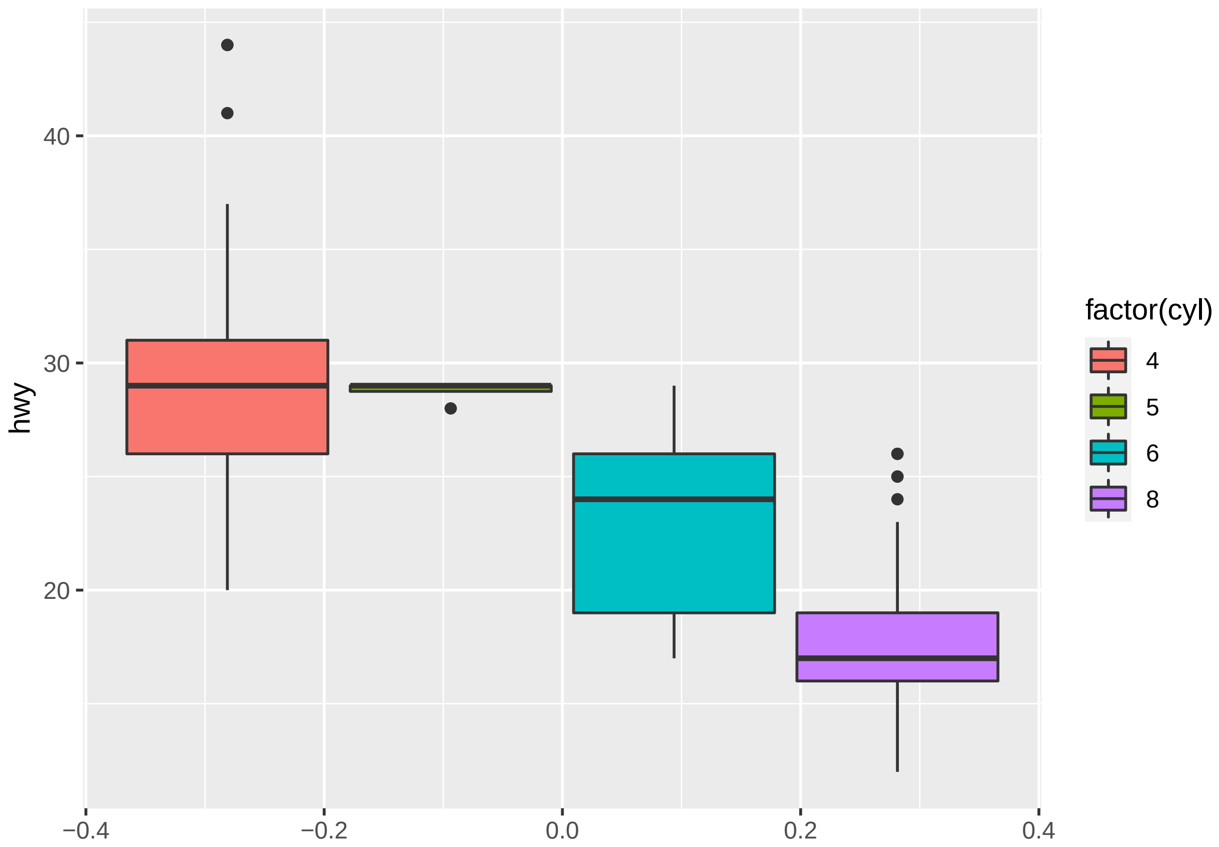
```
ggplot(mpg, aes(x = reorder(manufacturer, manufacturer, length))) +
    geom_bar() +
    coord_flip()
```



Dodge produced the most cars Lincoln produced the least cars

Ex 4

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



Pattern: negative linear relationship - more cylinders a car has, less hwy mpg it has

Ex 5

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

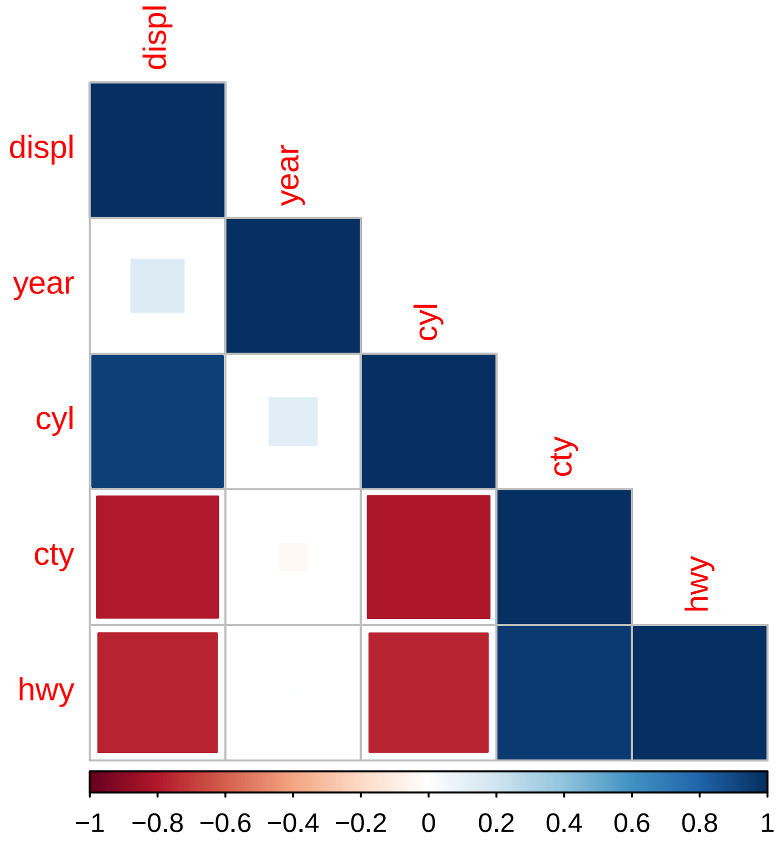
```
mpg_modified <- mpg %>% select(-c(manufacturer, model, trans, fl, class, drv))
head(mpg_modified)
```

```
## # A tibble: 6 x 5
##   displ  year   cyl   cty   hwy
##   <dbl> <int> <int> <int> <int>
## 1   1.8  1999     4    18    29
## 2   1.8  1999     4    21    29
## 3   2    2008     4    20    31
## 4   2    2008     4    21    30
## 5   2.8  1999     6    16    26
## 6   2.8  1999     6    18    26
```

```
corr_matrix <- cor(mpg_modified)
corr_matrix
```

```
##           displ        year       cyl        cty          hwy
## displ 1.0000000 0.147842816 0.9302271 -0.79852397 -0.766020021
## year  0.1478428 1.000000000 0.1222453 -0.03723229  0.002157643
## cyl    0.9302271  0.122245347  1.0000000 -0.80577141 -0.761912354
## cty   -0.7985240 -0.037232291 -0.8057714  1.00000000  0.955915914
## hwy   -0.7660200  0.002157643 -0.7619124  0.95591591  1.000000000
```

```
corrplot(corr_matrix, method = "square", type = "lower")
```

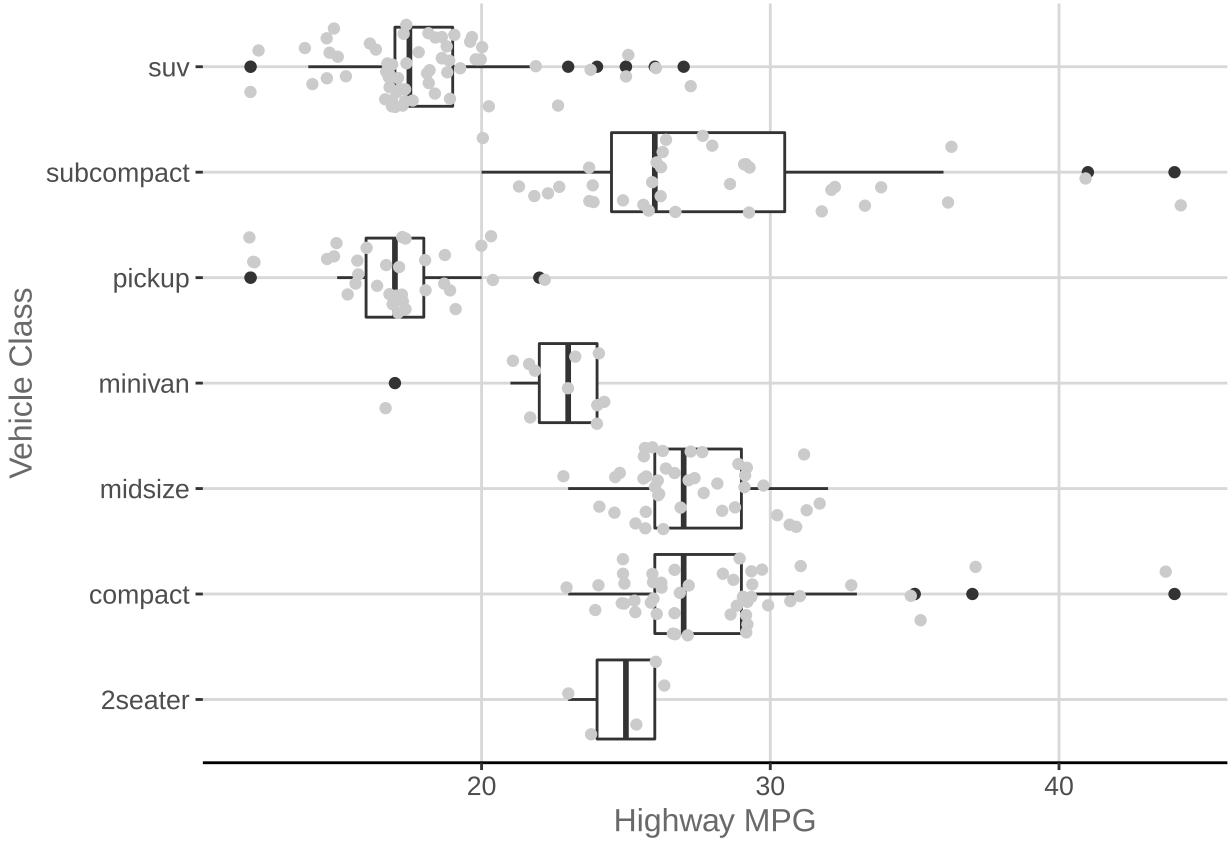


- displacement is positively correlates with cylinder number
- displacement is negatively correlates with cty mpg
- displacement is negatively correlates with hwy mpg
- year is marginally positively correlates with cylinder number
- cylinder is negative correlates with cty mpg
- cylinder is negative correlates with hwy mpg
- cty mpg is positively correlates with hwy mpg

Mostly relationships are easy to understand I'm suprised that year has no significant correlation with cty or hwy mpg since newer cars should be more efificency with better engine design

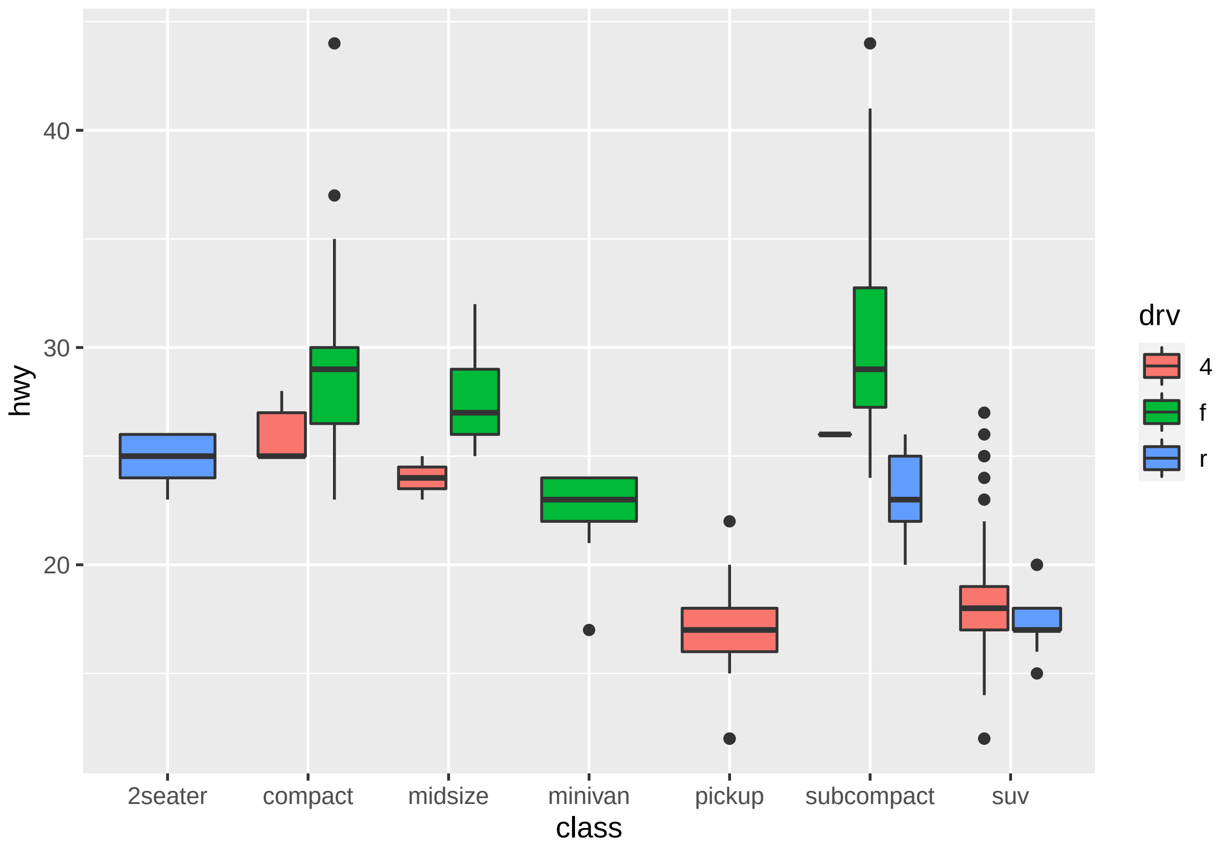
Ex 6

```
library(ggthemes)
ggplot(mpg, aes(x = hwy, y = class)) +
    geom_boxplot() +
    geom_point(position =  "jitter", color=c("#cbcbcb")) +
    xlab("Highway MPG") +
    ylab("Vehicle Class") +
    theme_hc() +
    theme(panel.grid.major.x = element_line(colour = "#D8D8D8"),
        axis.line = element_line(),
        axis.line.y = element_blank(),
        axis.title = element_text(color = "#696969")
        )
```



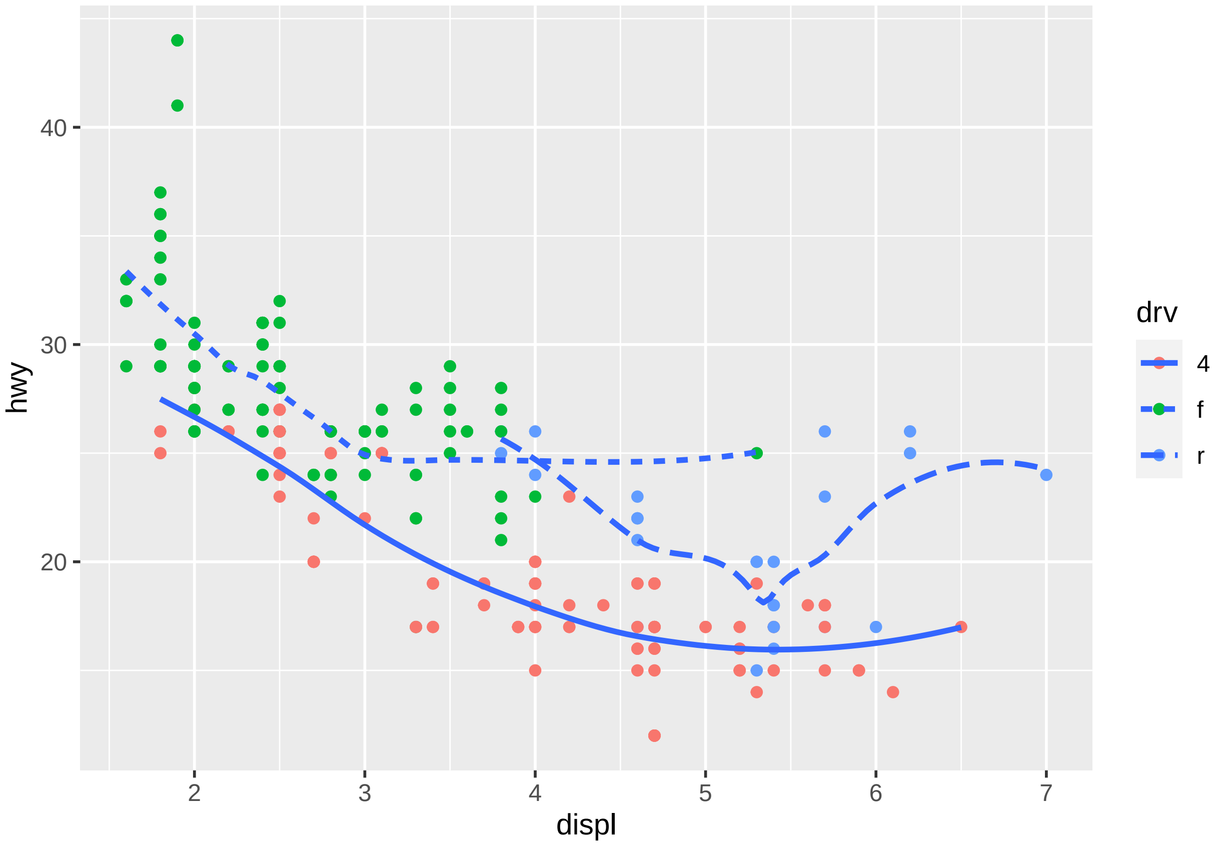
Ex 7

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



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

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


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