

Homework 1

R Markdown

```
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.1 --
```

```
## 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
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(tidymodels)
```

```
## Warning: package 'tidymodels' was built under R version 4.0.5
```

```
## -- Attaching packages ----- tidymodels 0.2.0 --
```

```
## v broom      0.7.12    v rsample      0.1.1
## v dials      0.1.0     v tune         0.2.0
## v infer      1.0.0     v workflows    0.2.6
## v modeldata  0.1.1     v workflowsets 0.2.1
## v parsnip    0.2.1     v yardstick    0.0.9
## v recipes    0.2.0
```

```
## Warning: package 'broom' was built under R version 4.0.5
```

```
## Warning: package 'dials' was built under R version 4.0.5
```

```
## Warning: package 'parsnip' was built under R version 4.0.5
```

```
## Warning: package 'recipes' was built under R version 4.0.5

## Warning: package 'tune' was built under R version 4.0.5

## Warning: package 'workflows' was built under R version 4.0.5

## Warning: package 'workflowsets' was built under R version 4.0.5

## -- Conflicts ----- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter() masks stats::filter()
## x recipes::fixed() masks stringr::fixed()
## x dplyr::lag() masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step() masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(ISLR)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(usethis)
```

Q1: In supervised learning there is a result Y (supervisor) where as in unsupervised learning there is not supervising result.

Q2: In a regression model Y is quantitative and has numerical values. In a classification model Y is qualitative and has categorical values.

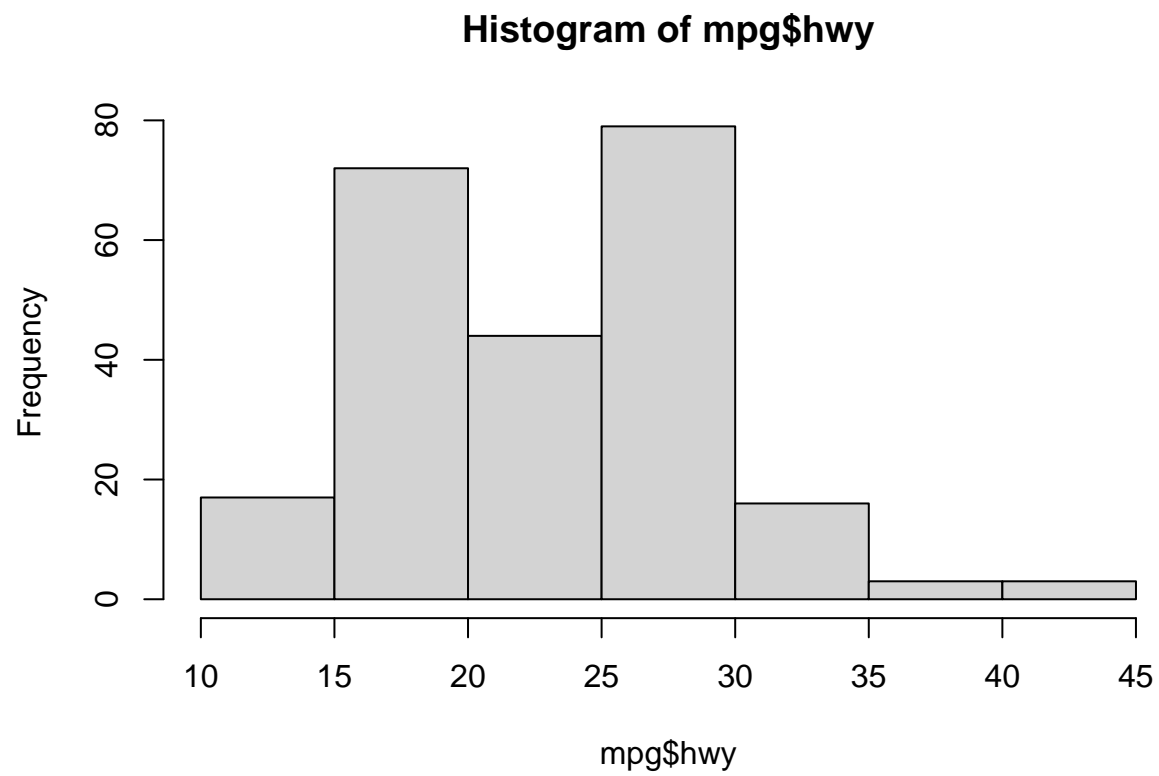
Q3:

Q4: Descriptive: Used to best visually emphasize trends in data (Lecture 2) Predictive: Use to try and predict the response with minimum reducible error (Lecture 2) Inferential: Used to test theories on the relationship between the outcome and predictors (Lecture 2)

Q5: - A mechanistic model assumes a parametric form for f . Where as the empirically driven model makes no assumptions about f (thus making it much more flexible by result) - I'd the mechanistic model would be easier to understand because f and the parametric form has constraints that make it easier to understand. The empirically driven model makes no assumptions and there for doesn't have these constraints. - The bias-variance tradeoff involves giving up flexibility for interpretability. Often times we will increase bias to shrink the variance in the model making it more useful.

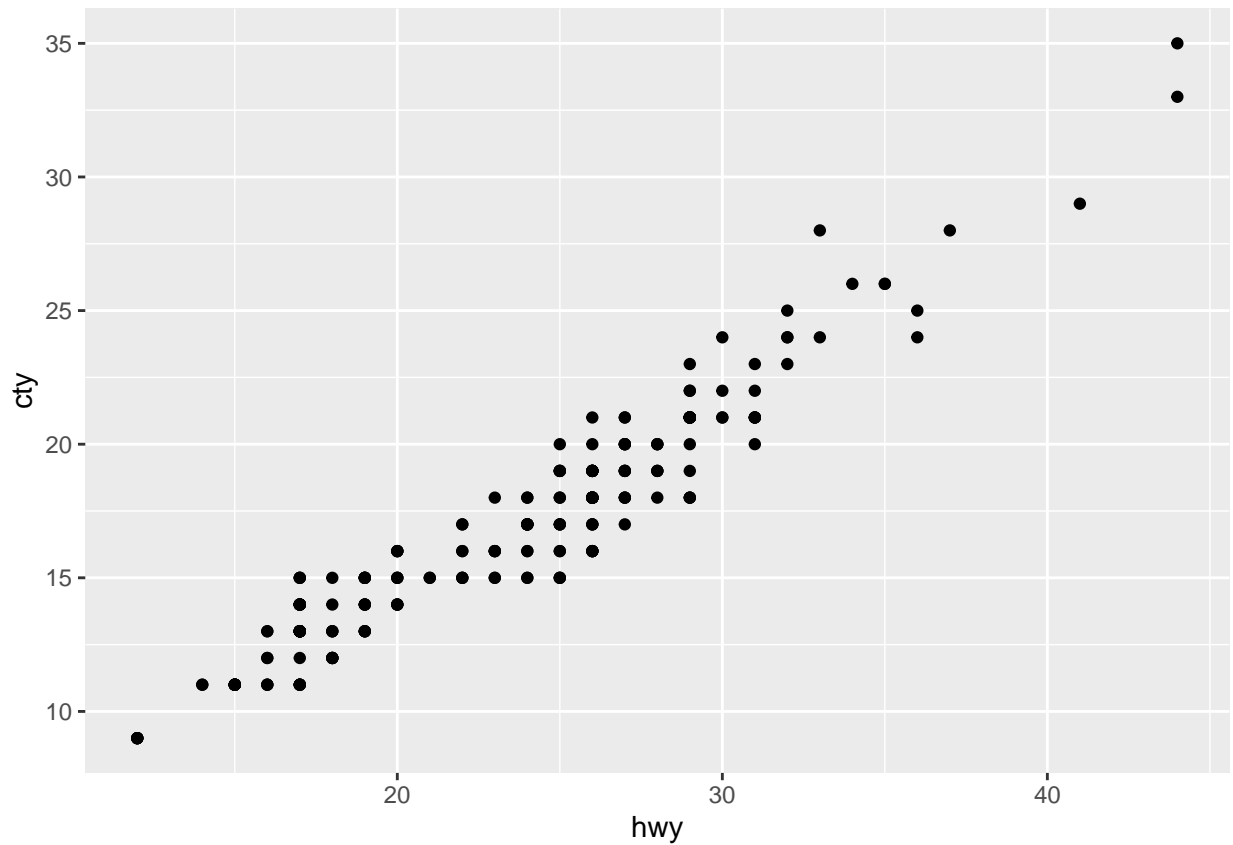
Q6: - The first question is predictive because the aim is to predict the response (in this case the vote) based on the inputs (voter profile) - The second one is inferential because we are trying to test a theory on the relationship between a voters relationship/contact to the candidate and their vote.

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



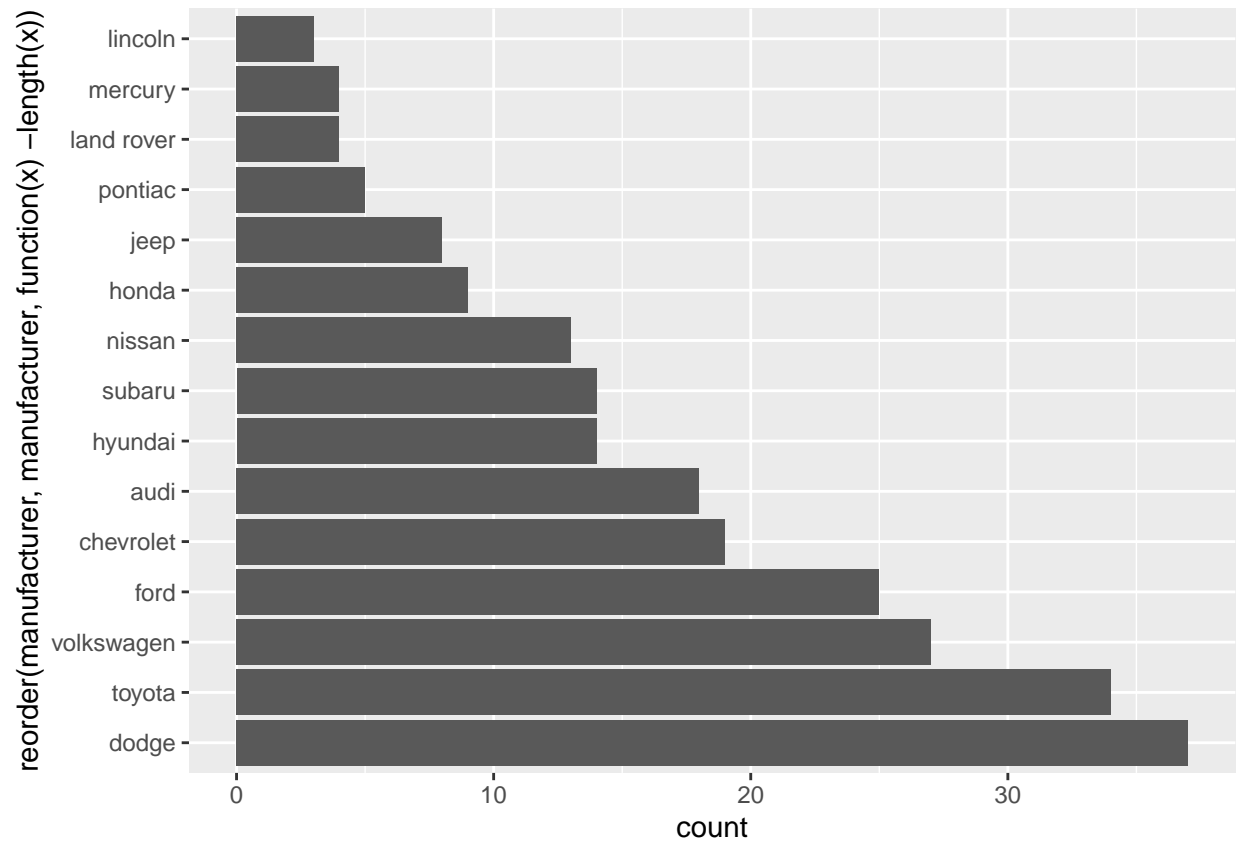
Most cars have between 25 and 30 mpg

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

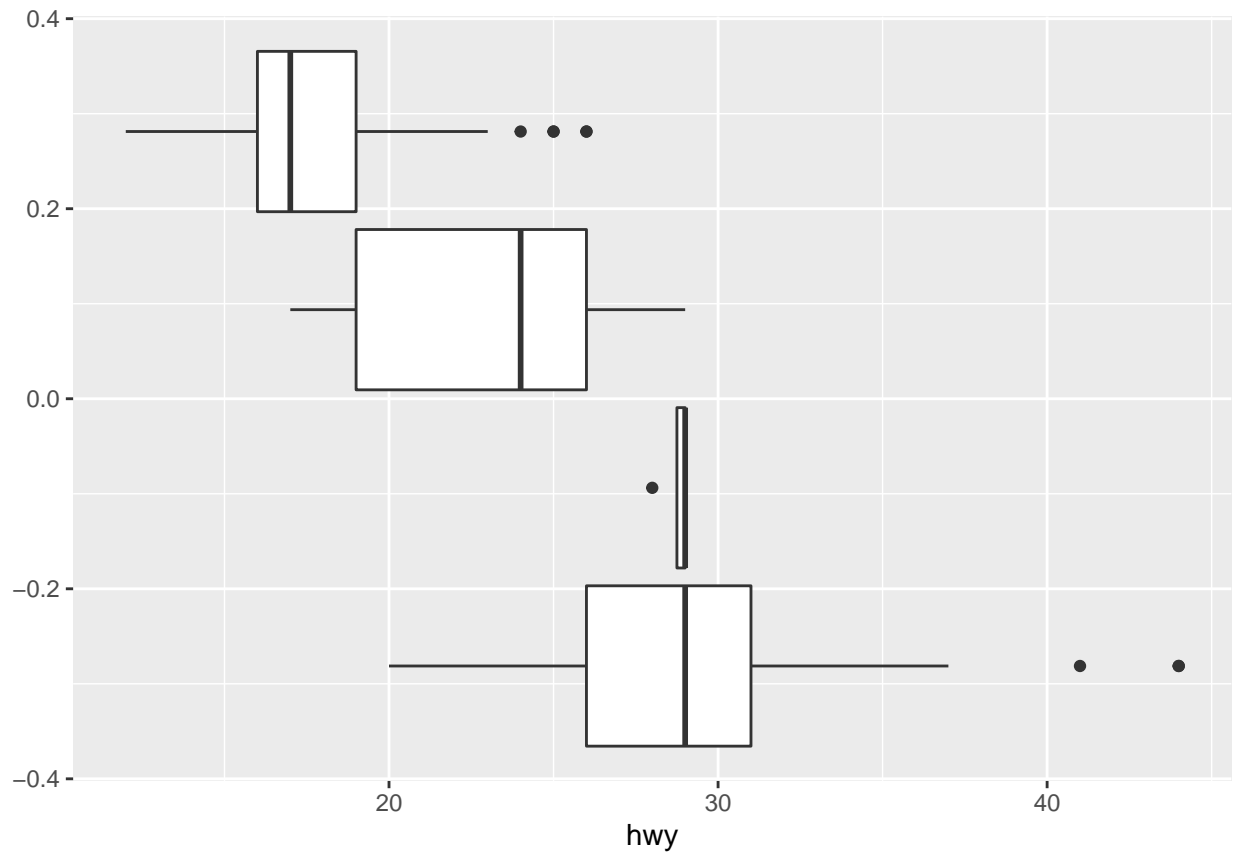


Highway mpg has a strong positive relationship with city mpg, they tend to increase together.

```
par(mar=c(2,8,2,2))  
ggplot(mpg, aes(y=reorder(manufacturer,manufacturer,function(x)-length(x)))) + geom_bar()
```

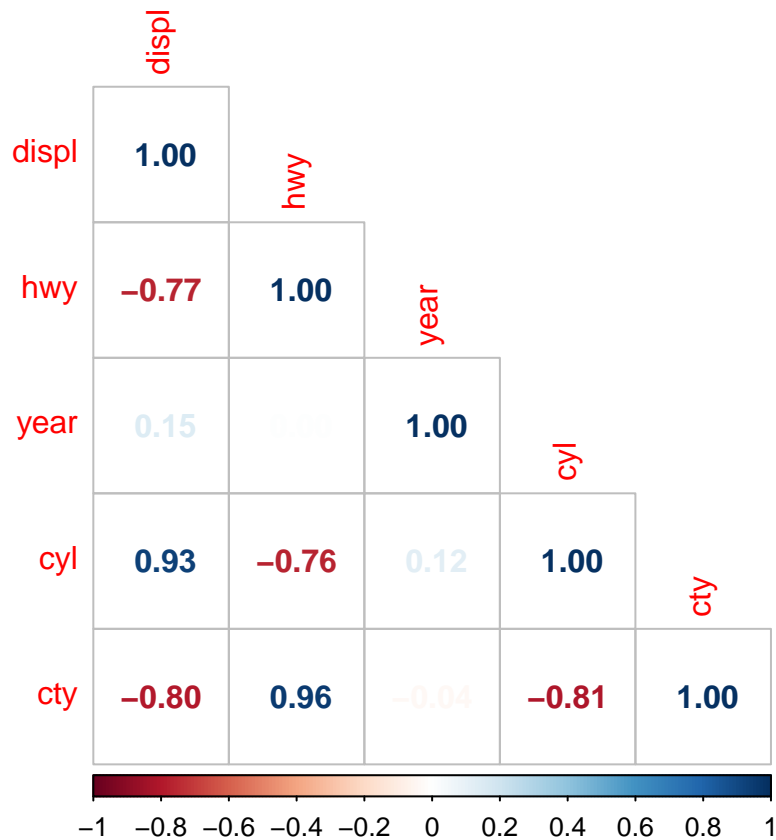


```
ggplot(mpg, aes(x=hwy, group=cyl)) + geom_boxplot()
```



Highway mpg tends to decrease as the number of cylinders increases.

```
mpg_num <- mpg %>% select(displ, hwy, year, cyl, cty)
corrplot(cor(mpg_num), method = 'number', type = 'lower')
```



Displacement is negatively correlated with hwy and city, but positively correlated with cylinders and year. Highway is negatively correlated with cylinder and positively correlated with city. year is positively correlated with city and cylinders. Cylinders is negatively correlated with city.

The only thing that surprised me was the positive correlation between city mpg and cylinders. I expected as cylinders increase.