Homework1

Question1: For each observation of the predictor measurement(s) xi, i = 1, . . . , n there is an associated response measurement yi. We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). Supervised learning can be separated into two types of problem: regression and classification. By contrast, unsupervised learning describes the somewhat more challenging situation in which for every observation i = 1,. . . ,n, we observe a vector of measurements xi but no associated response yi. It is not possible to fit a linear regression model, since there is no response variable to predict. The main distinction between the two approaches is whether there is an associated response measurement yi (from page26 of book)

Question2:

In regression model, the response variable Y is quantitative. Regression model help us to understand the relationship between dependent and independent variables and predict numerical values like price and blood pressure. In classification model, the response variable is qualitative. Classification model help us to classify spam/not spam email or survived/died. Logistic regression, decision tree and naive bayes are some examples of classification algorithms (from lecture1 31 33)

Question3: Commonly used metrics for regression ML problems: Mean Squared Error(MSE) Root Mean Squared Error(RMSE) Commonly used metrics for classification ML problems: Log Loss ROC AUC

Question4: Descripitve models: choose model to best visually emphasize a trend in data ex. using a line on a scatterplot Inferential models: what features are significant? Aim is to test theories Possibly causal claims State relationship between outcome and predictors Predictive models: What combo of features fits best? Aim is to predict Y with minimum reducible error Not focused on hypothesis tests (from lecture2 7)

Question5: Mechanistic: Based on Assume a parametric form for f and won't match true unknown f Can add more parameters to become more flexibility Have the problem of the overfitting Empirically driven: No assumption about f and required a larger number of observations Much more flexible by default Have the problem of the overfitting Empirically driven is easier to understand.

Question6: Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? This is a predictive question. We want to predict the possibility of voters will vote in favor of the candidate given their data.

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate? This is a inferential question. We want to know how personal contact with the candidate may influence voters

```
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)
```

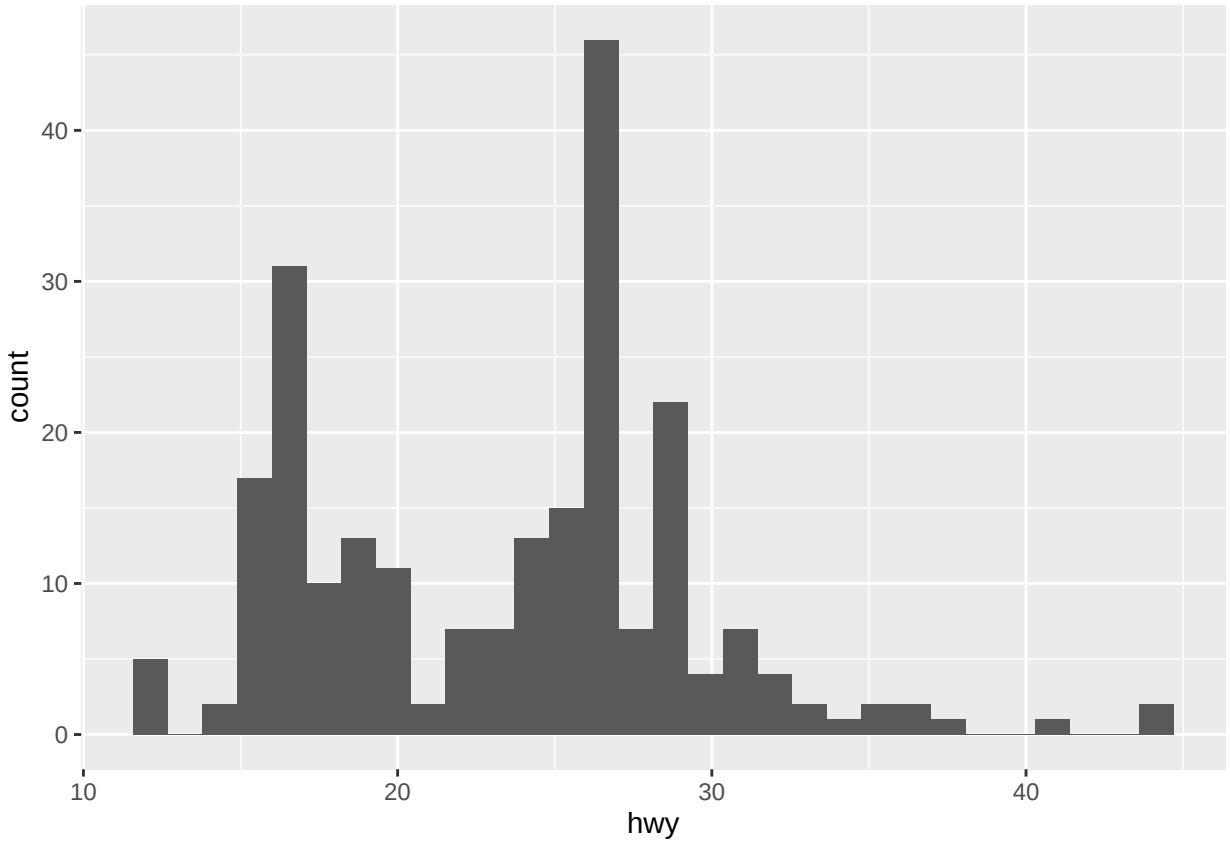
```
## Warning: package 'tidymodels' was built under R version 4.0.5

## -- 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 tidymodels_prefer() to resolve common conflicts.
```

```
library(ISLR)
```

Exercise1:

```
ggplot(data = mpg)+geom_histogram(mapping = aes(x = hwy))
```

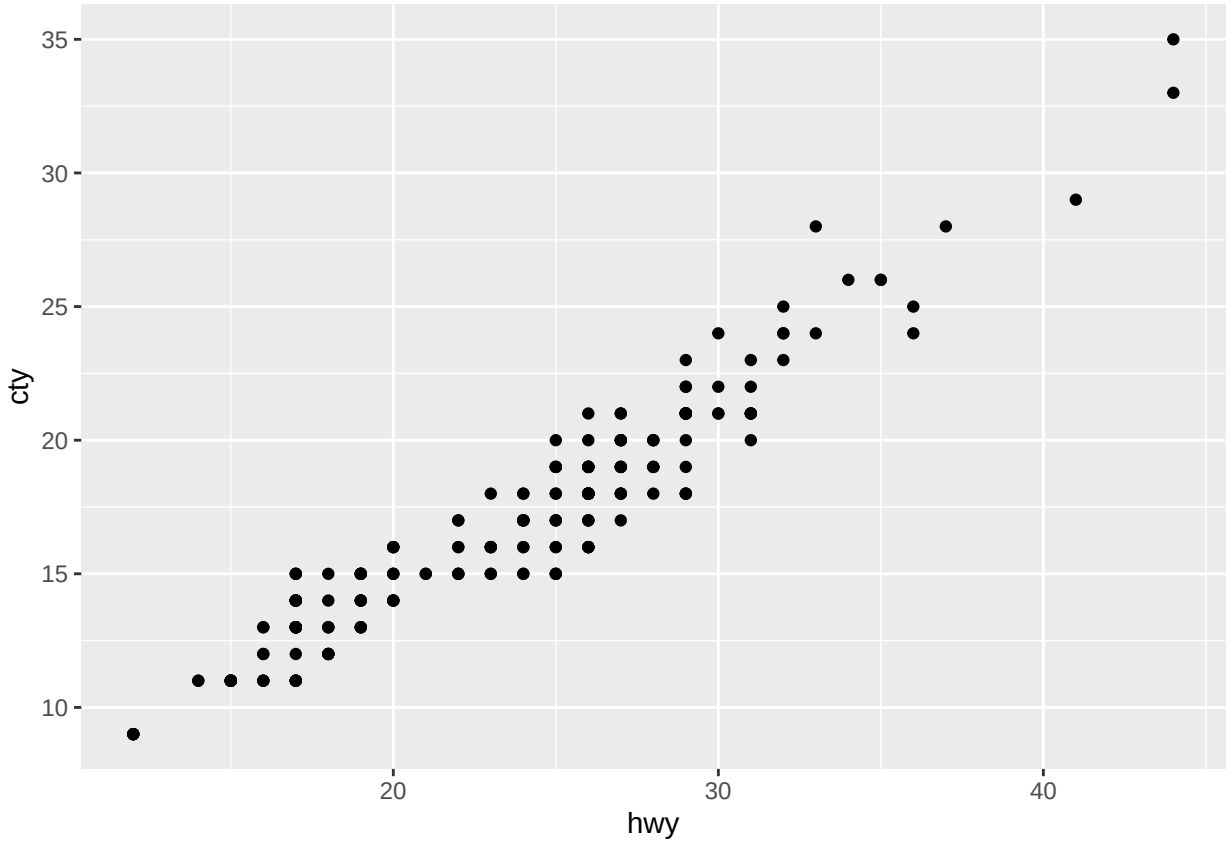
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



26 highway miles per gallon are the most common value

Exercise2:

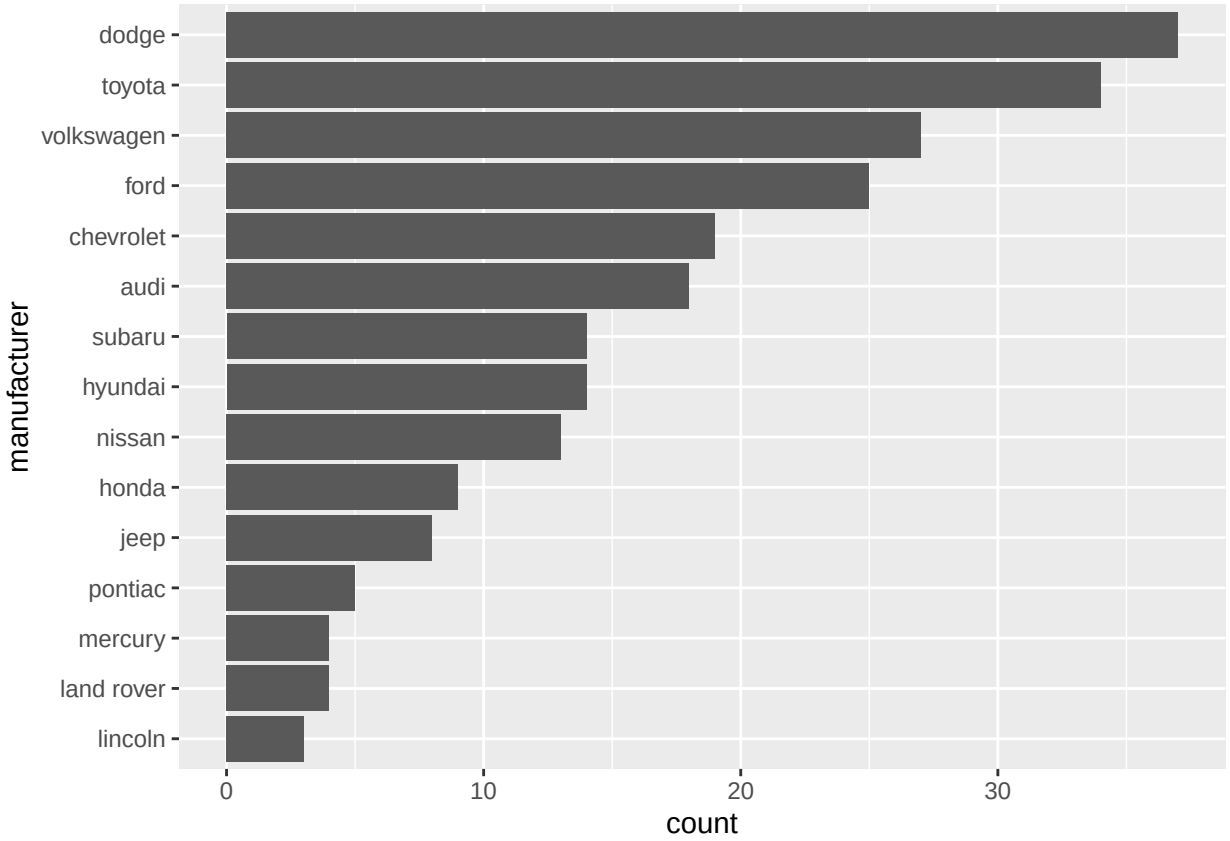
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = hwy, y = cty))
```



The trend of points seem upward slopping. There is relationship between cty and hwy, which means if we increase 1 highway miles per gallon, our city miles per gallon will increase.

Exercise3

```
ggplot(data = mpg) +
  geom_bar(mapping = aes(y = reorder(manufacturer, manufacturer, function(x) length(x)))) + ylab("manufa
```

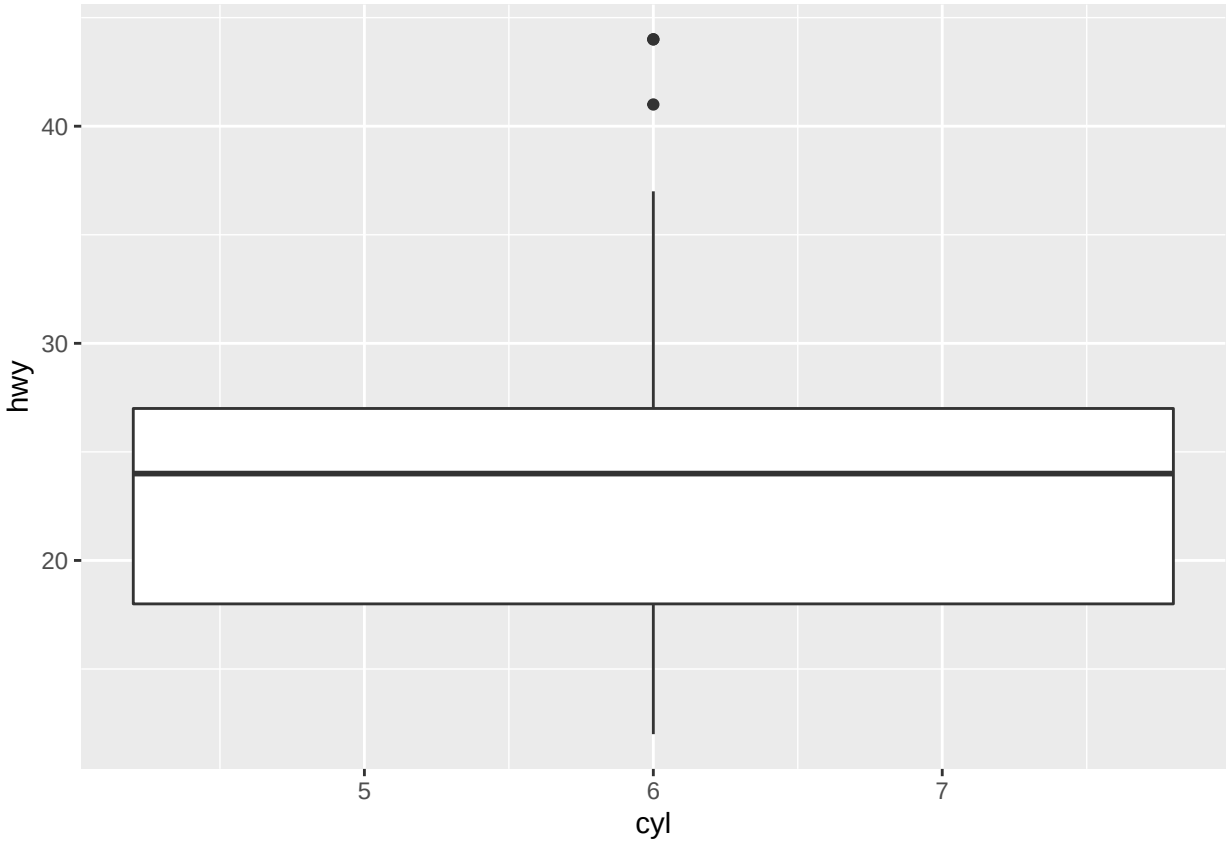


Dodge produced the most car and Lincoln produced the least

Exercise4

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

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```

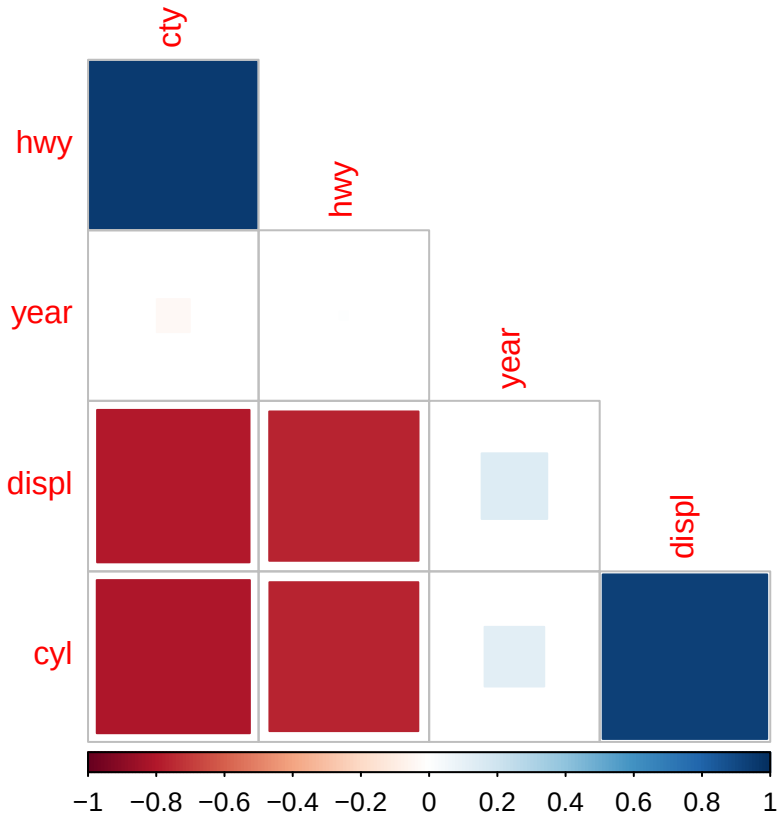


Exercise 5

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
M = cor(mpg[, -c(1,2,6,7,10,11)])
corrplot(M, method = 'square', order = 'FPC', type = 'lower', diag = FALSE)
```



cyl: number of cylinders cty: city miles per gallon hwy: highway miles per gallon displ: engine displacement, in litres

cyl is negative correlated with cty and hwy and positive correlated with displ displ is negative correlated with cty and hwy hwy is positively correlated with cty These relationship makes sense to me. The vehicle have more number of cylinders, which consume more gasoline