131 hw1

Tianyi Li

Contents

Question 1 Supervised learning: a machine learning technique that for each observation of the predictor measurements, there is an associated response to the predictors. (from page 26 of book An Introduction to Statistical Learning) Unsupervised learning: a machine learning technique that for every observation, we observe a vector of measurements but no associated response. (from page 26 of book An Introduction to Statistical Learning) Difference: In supervised learning, the actual data Y is the supervisor, while in unsupervised learning, we learn without a supervisor. (from the lecture1 slides)

Question 2 (from the lecture1 slides) Regression model: Y is quantitative, which are numerical values. Classification model: Y is qualitative, which are categorical values.

Question 3 for regression ML problems: mean absolute error, R-squared for classification ML problems: accuracy, F1-score

Question 4 (from the lecture2 slides) Descriptive models: choose model to best visually emphasize a trend in data. Inferential models: aim is to test theories,(possibly) causal claims, state relationship between outcome & predictor(s). Predictive models: aim is to predict Y with minimum reducible error, not focused on hypothesis tests.

Question 5 Mechanistic: make prediction using a theory. Empirically-driven: learn by experimenting. Difference: mechanistic models use a theory to predict while empirically-driven models develop a theory through experiments. Similarity: they are both predictive models which are used for predict Y. I think the empirically-driven model is easier to understand, since the conclusions of these models are driven from real-world events, which makes them more direct to see. The bias-variance tradeoff determines the performance of predictive models' algorithm by breaking down its prediction error.

Question 6 The first question: predictive, because it uses a voter's data to predict future outcome. The second question: inferential, because it aims to test theories and state relationship between outcome and predictor.

```
install.packages("tidyverse",repos="https://cran.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q_/b_fx39td6gb_d36sjcxswdnm0000gn/T//Rtmpyjepus/downloaded_packages
```

```
install.packages("tidymodels",repos="https://cran.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q_/b_fx39td6gb_d36sjcxswdnm0000gn/T//Rtmpyjepus/downloaded_packages
```

```
install.packages("ISLR",repos="https://cran.r-project.org")
```

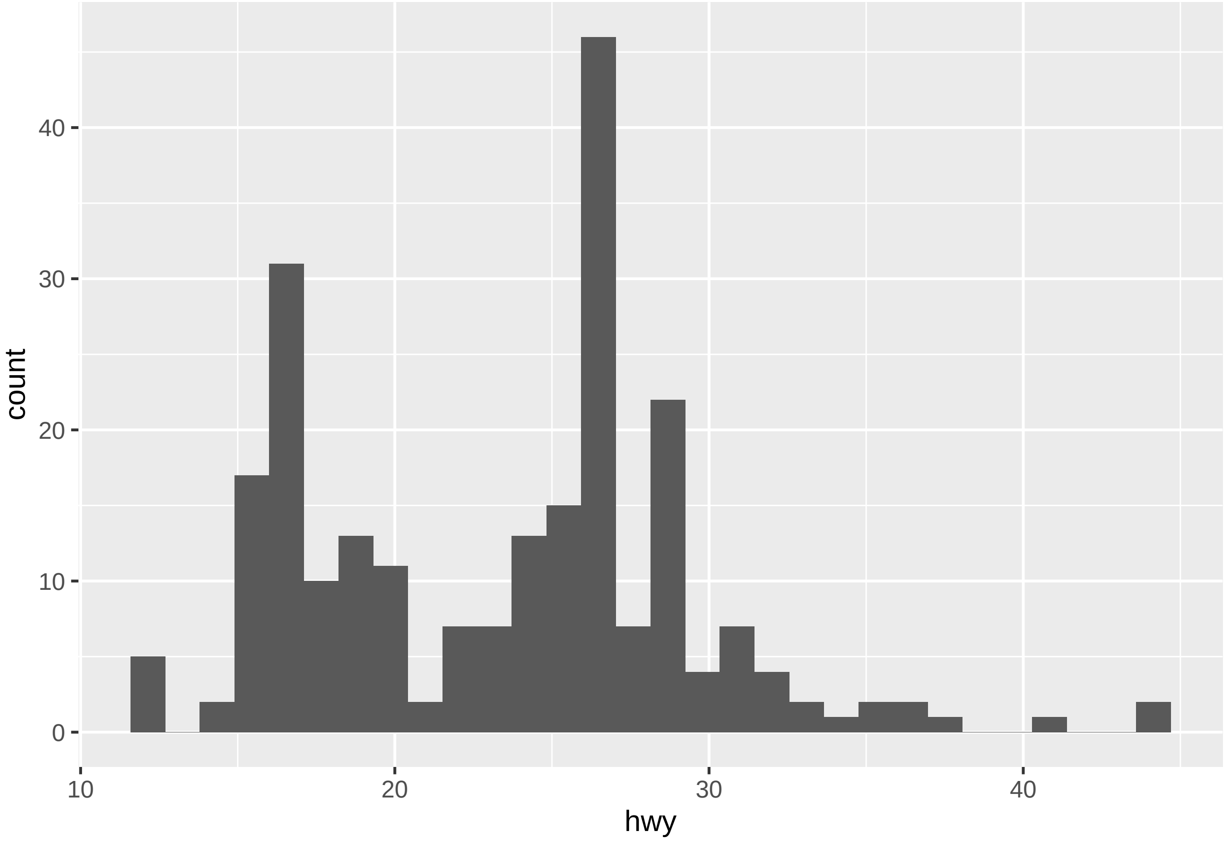
```
##
## The downloaded binary packages are in
##  /var/folders/q_/b_fx39td6gb_d36sjcxswdnm0000gn/T//Rtmpyjepus/downloaded_packages
```

```
library(tidyverse)
library(tidymodels)
library(ISLR)
```

```
mpg
```

Exercise 1

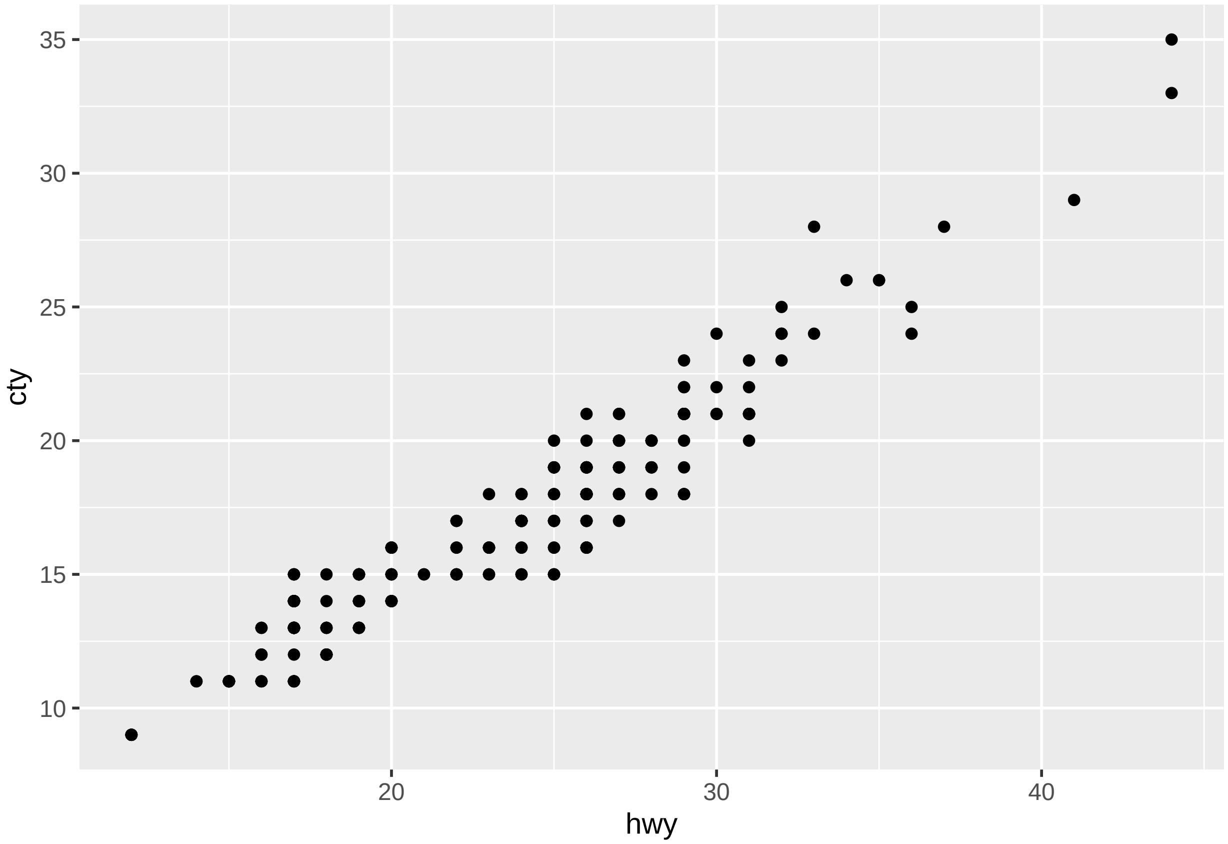
```
ggplot(data=mpg,mapping=aes(hwy))+geom_histogram()
```



There are two distinct peaks, but there is no clear pattern overall, so the distribution is random.

Exercise 2

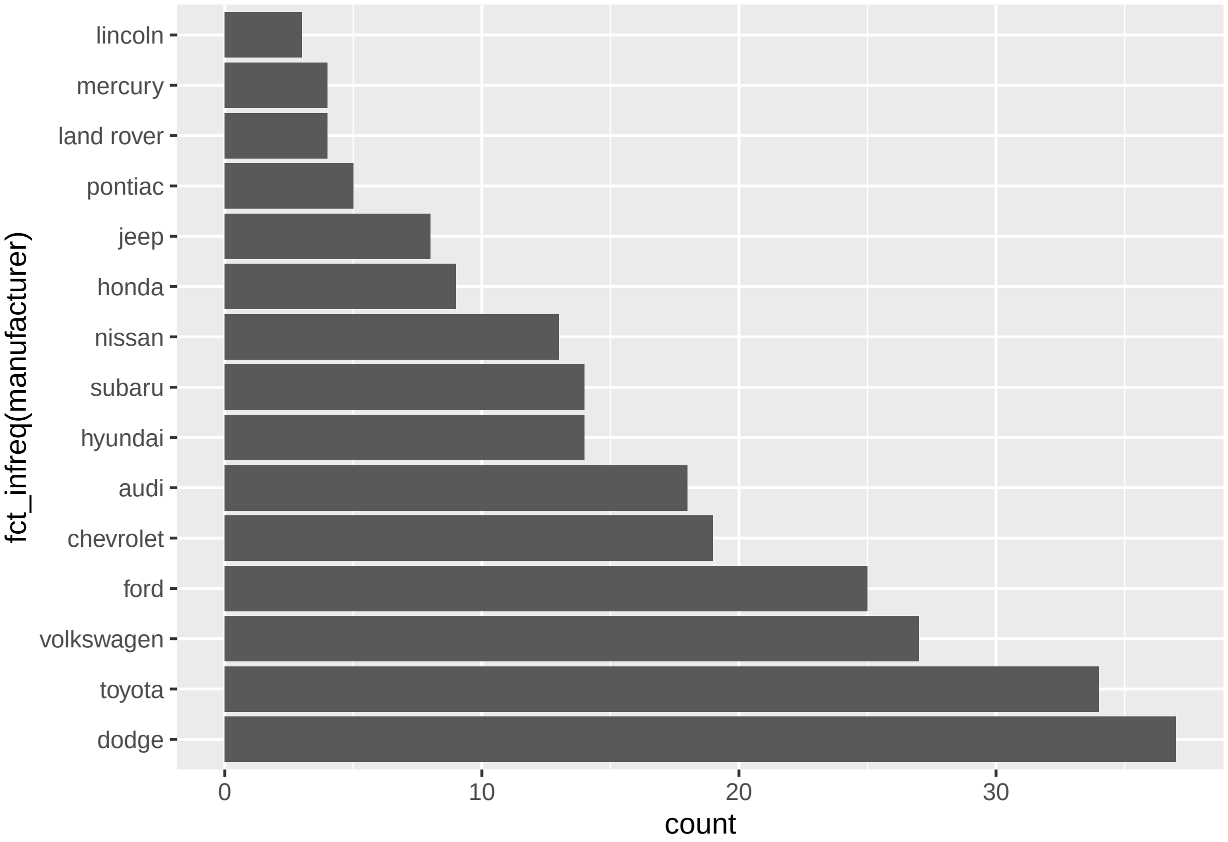
```
ggplot(data=mpg,mapping=aes(hwy,cty))+geom_point()
```



There is a positive correlation between hwy and cty, which means that as hwy increases, cty also increases.

Exercise 3

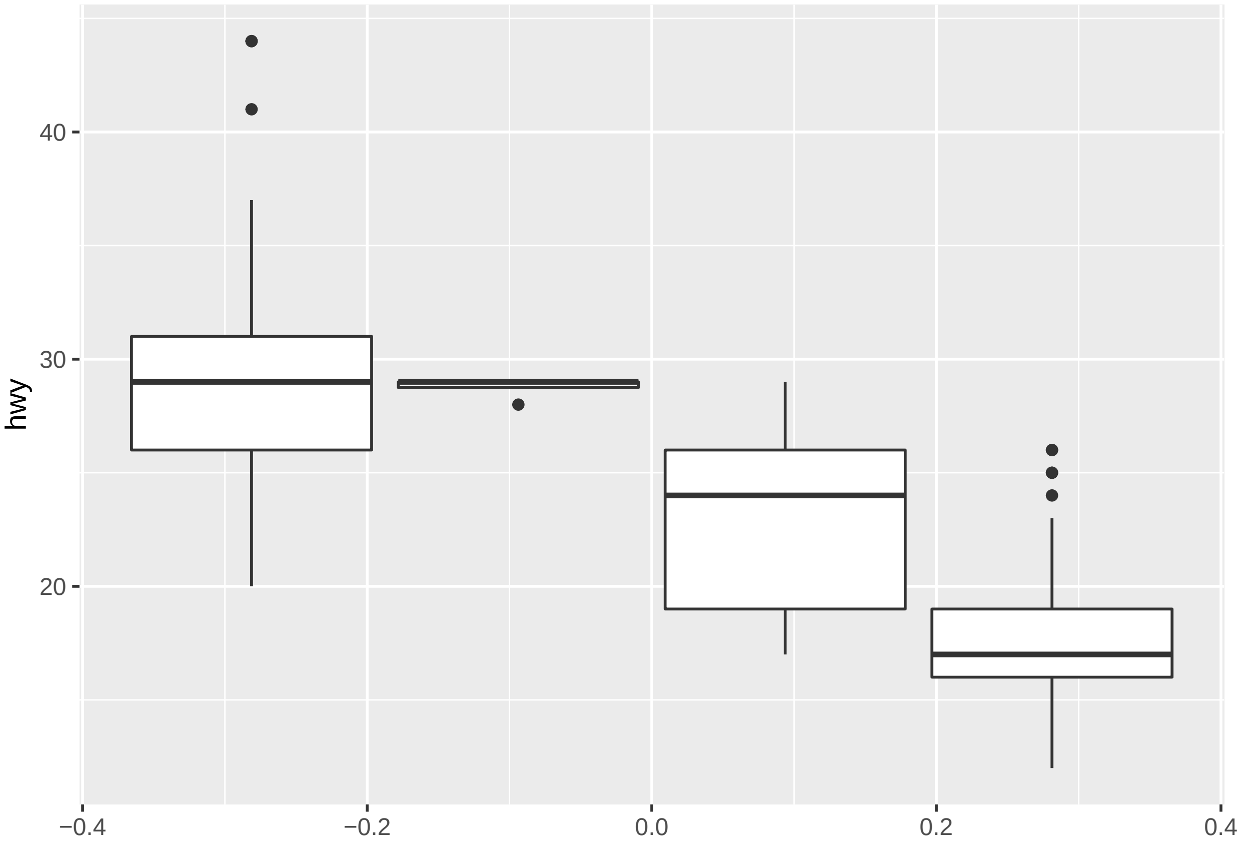
```
library(ggplot2)
library(tidyverse)
mpg %>%
  ggplot(aes(fct_infreq(manufacturer)))+
  geom_bar()+coord_flip()
```



Dodge produced the most cars, and Lincoln produced the least.

Exercise 4

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



Each of the four sections of the box plot is lower than another as cyl increases.

Exercise 5

```
install.packages("magrittr",repos="https://cran.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q_/b_fx39td6gb_d36sjcxswdnm0000gn/T//Rtmpyjepus/downloaded_packages
```

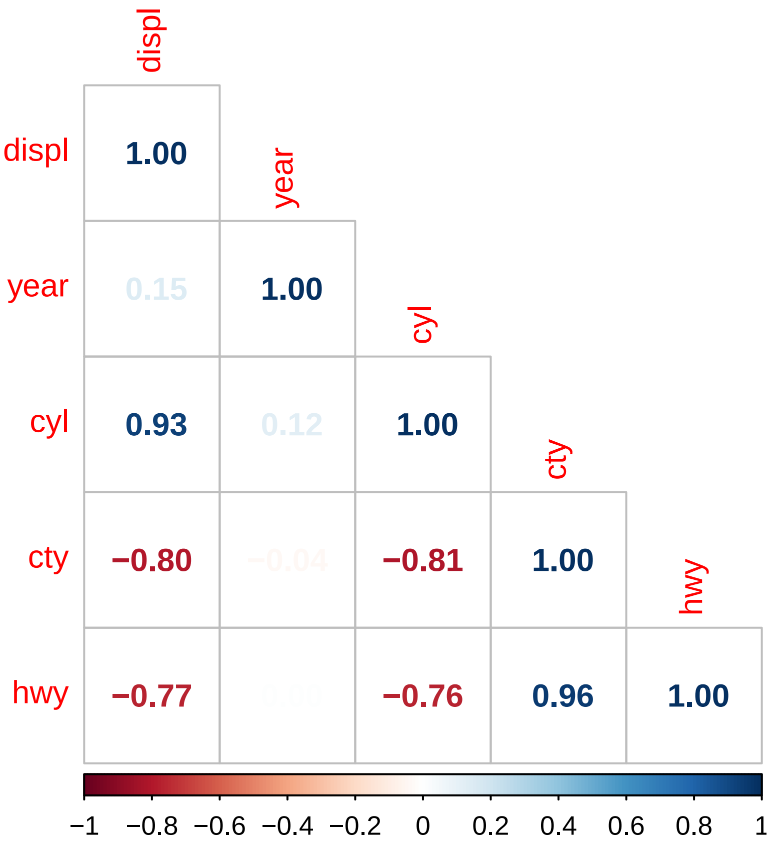
```
install.packages("dplyr",repos="https://cran.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q_/b_fx39td6gb_d36sjcxswdnm0000gn/T//Rtmpyjepus/downloaded_packages
```

```
library(magrittr)
library(dplyr)
install.packages('corrplot',repos="https://cran.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/q_/b_fx39td6gb_d36sjcxswdnm0000gn/T//Rtmpyjepus/downloaded_packages
```

```
library(corrplot)
corr=cor(mpg %>% dplyr::select(where(is.numeric)))
corrplot(corr,method=c("number"),type=c('lower'))
```



dispel is positively correlated with cyl, and negatively correlated with cty and hwy. cyl is negatively correlated with cty and hwy. cty is positively correlated with hwy. Most of these relationships make sense to me, but I'm surprised that the year has nearly no relationship with other variables.