131-hw1

Rita Han

2022-09-30

Q1:supervised vs unsupervised: Supervised is a track of machine learning that map input to output based on the observed mapping. Unsupervised is a track of machine learning that is without an observed mapping. The difference between them is like the former is learning with an answer key, the latter is without it.

Q2:regression model vs classification model: In machine learning, regression model is to predict a continuous quantity, and classification is to predict a discrete class. In order words, regression is for numerical value(e.g.price), and classification is for categorical values(e.g.survived) (Coburn, day_1_131_231.pdf, pg33).

Q3:two commonly used metrics for regression ML problems and classification ML problems: Regression: R-square, Root Mean Squared Error (RMSE). Classification: Accurary and F1 score.

Q4: Descriptive models: used to visualize the trend in data(e.g.using a line on a scatterplot) Inferential models: find the important features and the relationships between outcome and predictors. Predictive models: what groups of features can be the best combo to predict Y with minimum reducible error (Coburn, day_1_131_231.pdf, pg39).

Q5: mechanistic vs empirically-driven: -Mechanistic model is a parametric form that uses a theory to predict the real world outcome. -Empirically-driven is a non-parametric form that studies real-world events to predict and develop a theory, which are more flexible. They are both used to predict future, but based on different things, and are different in bias and variance(stated more below). Mechanistic models are easier to understand because they basically fit easy parametric forms. As Professor Coburn stated in the first week slides, no single method is the best choice for all data sets. The Mechanistic model tend to have higher bias and lower variance, but empiriclally-driven models have higher variance and lower bias. This also shows how bias-variance tradeoff is related to the two types of models.

Q6: The first situation is predictive, because it is predicting the outcome of how the voter will behave. The second is inferential, because inferential model is to look for the relationship between outcome and predictors, and in this situation, it is looking for the relationship between the predictor, the personal contact with the candidate and the votor behavior.

EDA

```
#install.packages("tidyverse")
#install.packages("tidymodels")
#install.packages("ISLR")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

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

```
library(tidymodels)
```

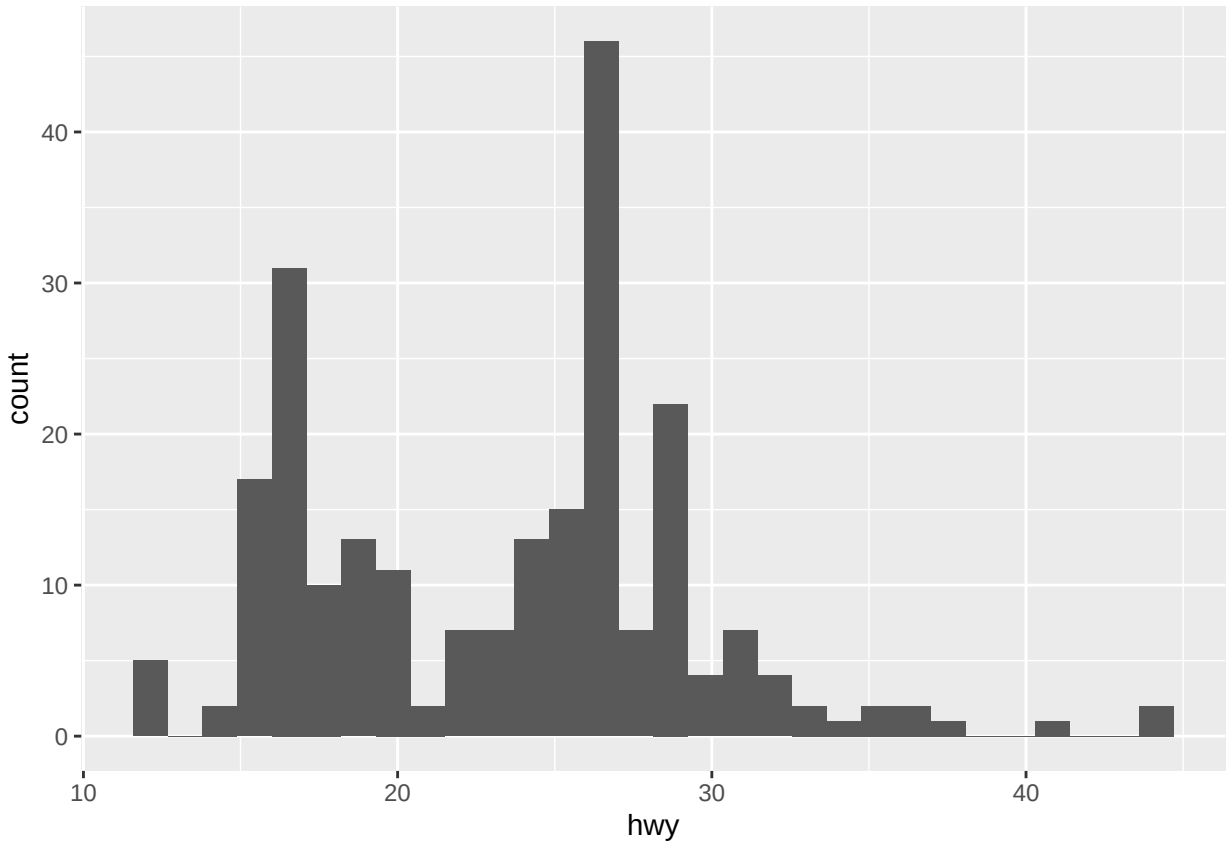
```
## -- Attaching packages --------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.0
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.2     v yardstick    1.1.0
## v recipes      1.0.1
## -- 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)
```

```
?mpg
```

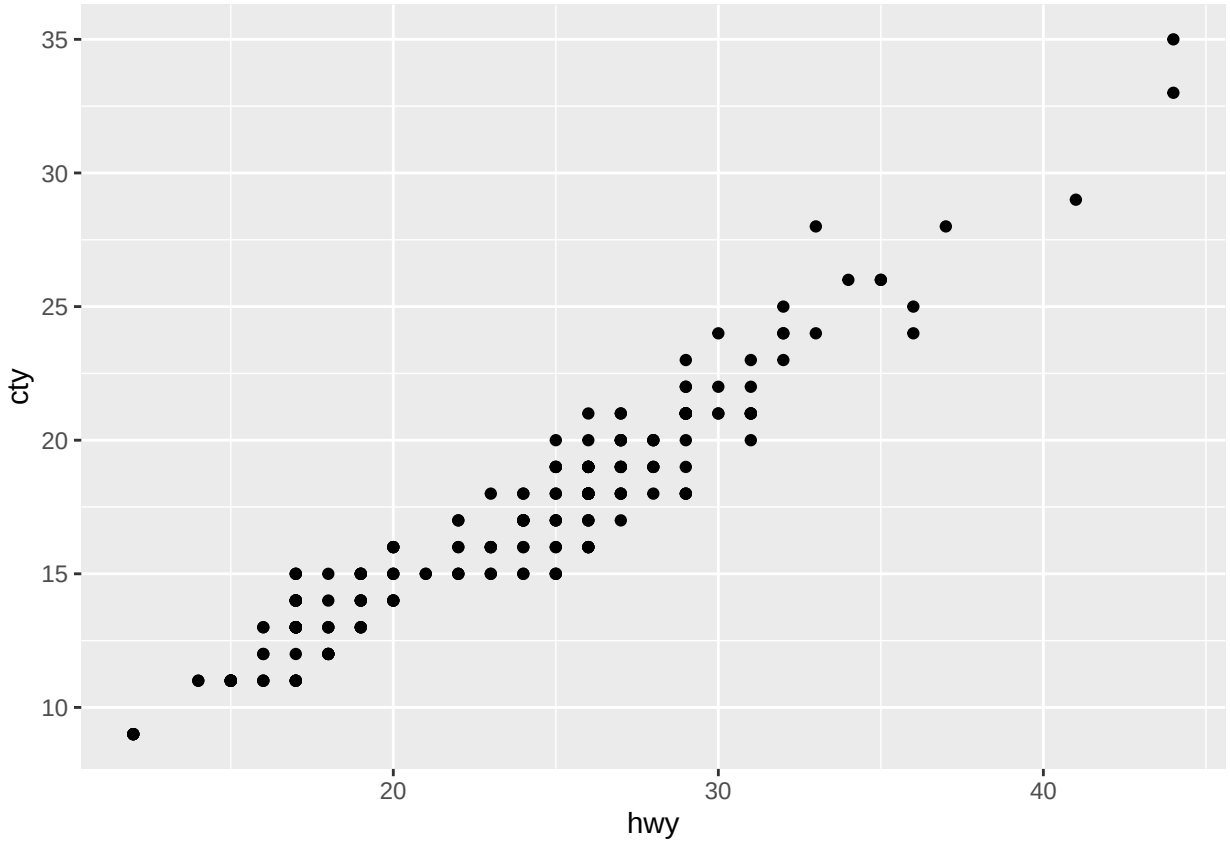
```
#Exc1:
ggplot(data=mpg, aes(x=hwy))+geom_histogram()
```

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



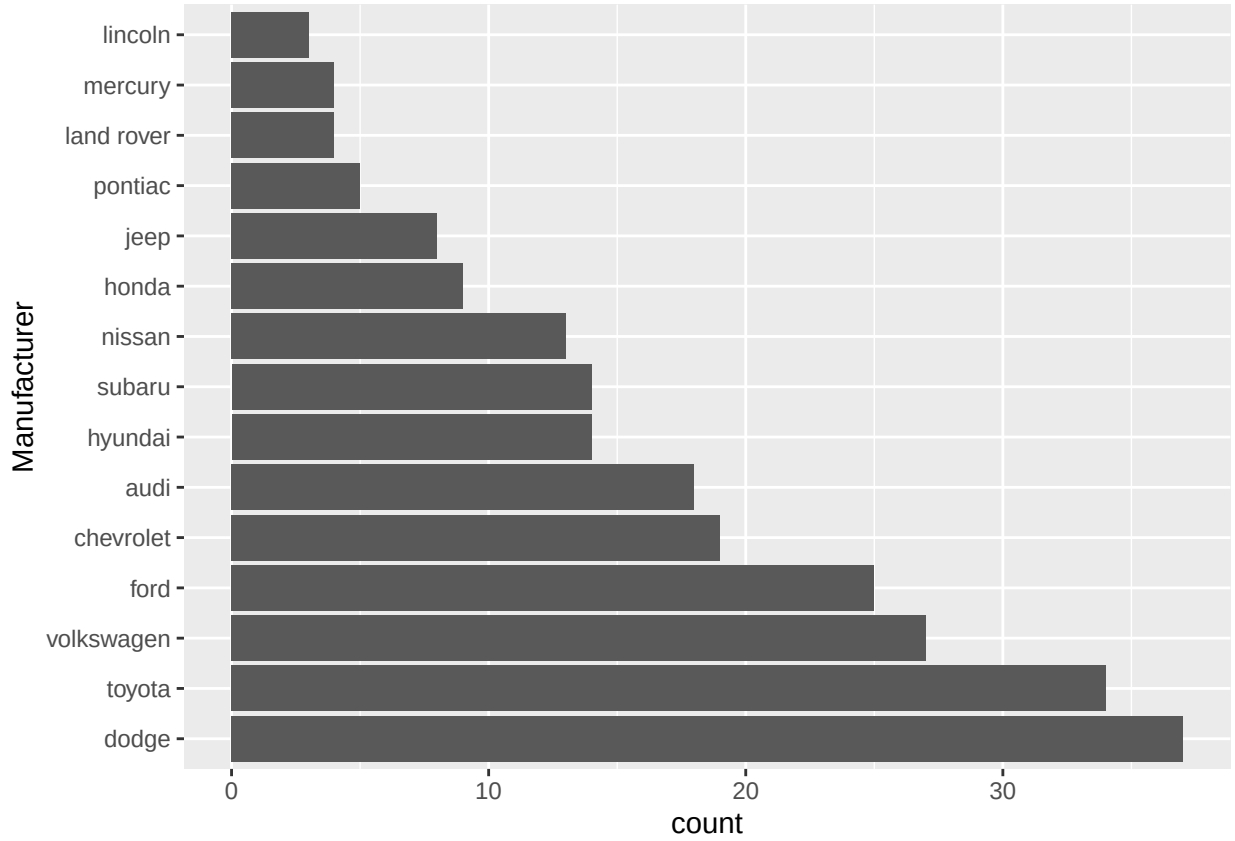
There is a peak between 16-17, and 25-28. The graph seems to be positively skewed. A few cars have 'hwy' higher than 40.

```
#Exc2
ggplot(data=mpg, aes(x=hwy,y=cty))+geom_point()
```



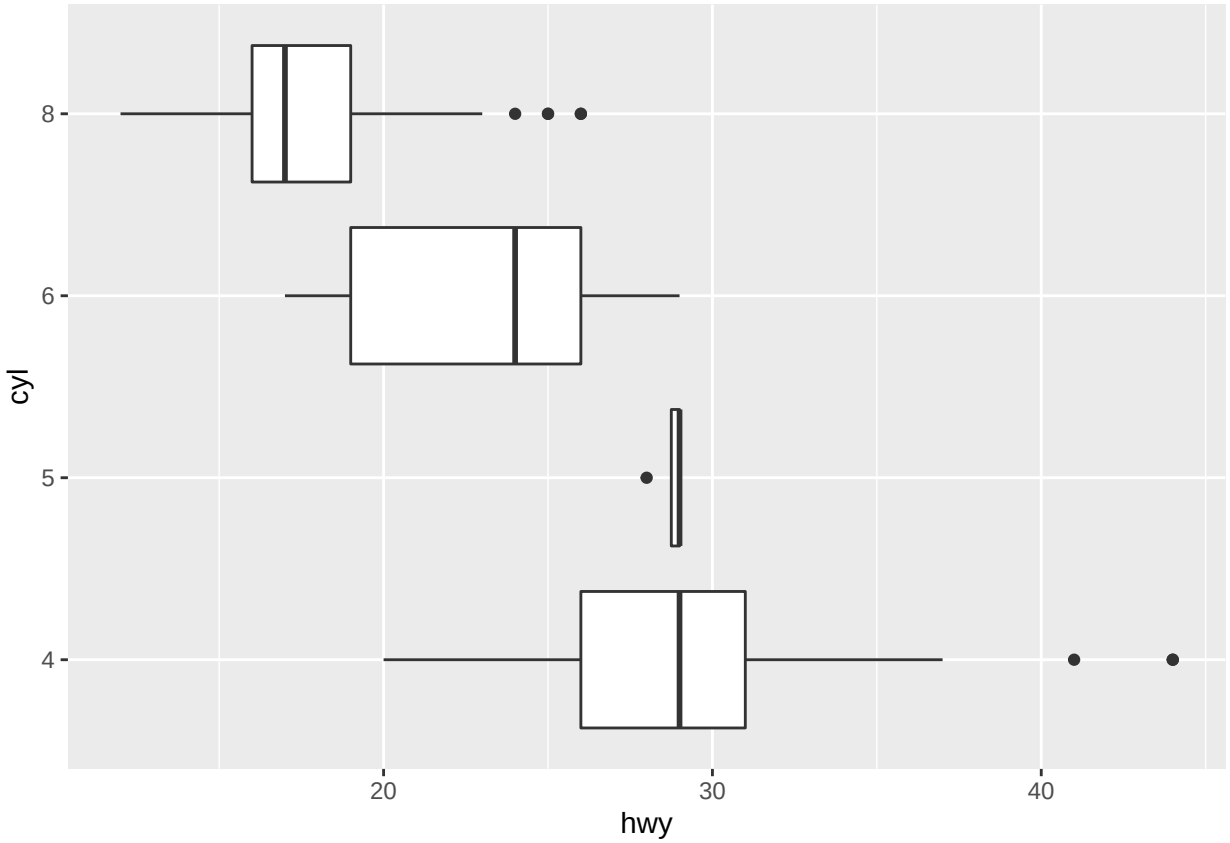
There is a obvious positive linear relationship between the two variables. This means if highway mpg increases, then the city mpg increases. The dots appears a pattern which implies the numbers might be rounded.

```
#Exc3
ggplot(data=mpg, aes(x=fct_infreq(manufacturer)))+geom_bar()+coord_flip()+
  xlab('Manufacturer')
```



Lincoin produced the least cars and Dodge produced the most.

```
#Exc4
ggplot(data=mpg, aes(x=hwy,y=factor(cyl)))+geom_boxplot()+xlab('hwy')+ylab('cyl')
```

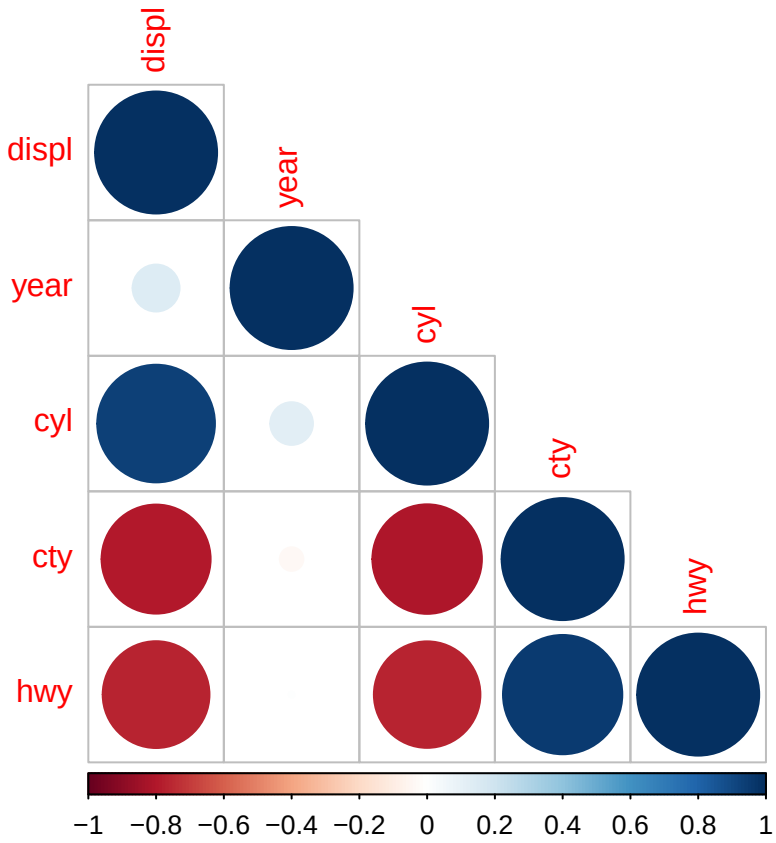


There is a pattern between the two variables. If the number of cylinder increase, the highway mpg decrease.

```
#Exc5
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
correlation=cor(mpg[, unlist(lapply(mpg, is.numeric))])
corrplot(correlation, method='circle', type='lower')
```



The number of cylinder is positively correlated to the engine displacement; they city mpg and the highway mpg is negatively correlated to the engine displacement and also to the number of cylinder; which are common senses in real life.