PSTAT131-HW1

Eric Liu

Machine Learning Main ideas

Question 1

Supervised learning: supervised learning is fitting models that relates the response to the predictors, with the aim of accurately predicting the response for future observations or better understanding the relationship between the response and the predictors (from page 26 of the book).
Unsupervised learning: unsupervised learning involves no response variable to predict, with the aim of understanding the relationships between the variables or between the observations.
The difference between supervised and unsupervised learning is that one uses both the response and the predictors, and the other one only uses the predictors or the inputs without the response or the output.

Question 2

In regression ML problems, the response or the output is quantitative and is numerical values.
In classification ML problems, the response or the output is qualitative and is categorical values.

Question 3

Metrics for regression models: MSE(mean square error), R^2(R square)
Metrics for classification models: error rate, Bayes classifier

Question 4

Descriptive models: descriptive models are used to best visually emphasize a trend in data, such as using a line on a scatterplot(from lecture).
Inferential models: inferential models are used to understand the relationship between the response and the predictors, and we need to know the exact form of the function f.
Predictive models: predictive models are used to best predict the future response with minimum reducible error with the choice of combo of features that fit the best, and we may not need to know the exact form of the function f.

Question 5

Mechanistic methods: first assume a parametric form for the function f, and use the training data to fit parameters.
Empirically-driven methods: make no assumptions about the function f, and seek an estimate of f that gets as close to the data points as possible.
The difference between them is that, mechanistic methods assumes the function f before the training, but empirically-driven methods doesn't need such assumption. Because of this, empirically-driven methods are more flexible by default, although mechanistic methods can be more flexible by adding parameters.

In addition, mechanistic methods may not match the true unknown f. Lastly, compared to mechanistic methods, empirically-driven methods require larger numbers of observations.
The similarity between them is that they can both result in overfitting.

In general, mechanistic model is easier to understand. Since the form of the function f is assumed before the training, we have more information about the model.

Generally, more flexible methods have higher variance and lower bias. Since empirically-driven methods are more flexible than mechanistic methods, empirically-driven methods will have higher variance and lower bias in general, compared to mechanistic methods, and mechanistic methods will have lower variance and higher bias. Because it is nearly impossible to find a method with both low variance and low bias, there is a bias-variance tradeoff, so there is also a tradeoff between the use of mechanistic methods and the use of empirically-driven methods.

Question 6

The first question is predictive, because we are asked to predict the response, which is the likelihood of voting in favor of the candidate, given the predictors, which are a voter's profile. Therefore, we focus on the prediction for this question.
The second question is inferential, because we are interested in the relationship between the response and the predictors, and in this question we are interested in the relationship between the likelihood of support between whether voters have personal contact with the candidate.

Exploratory Data Analysis

```
library(tidyverse)
```

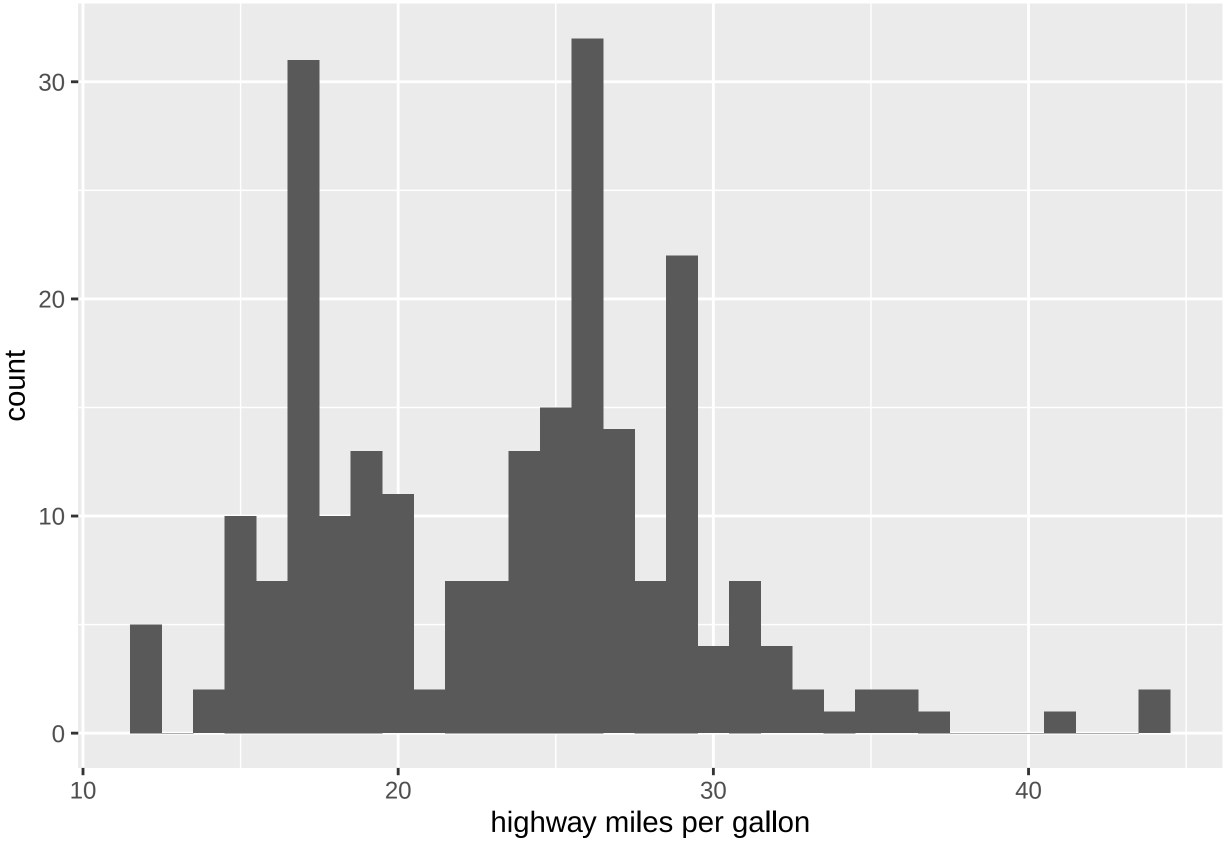
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
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~
```

Exercise 1

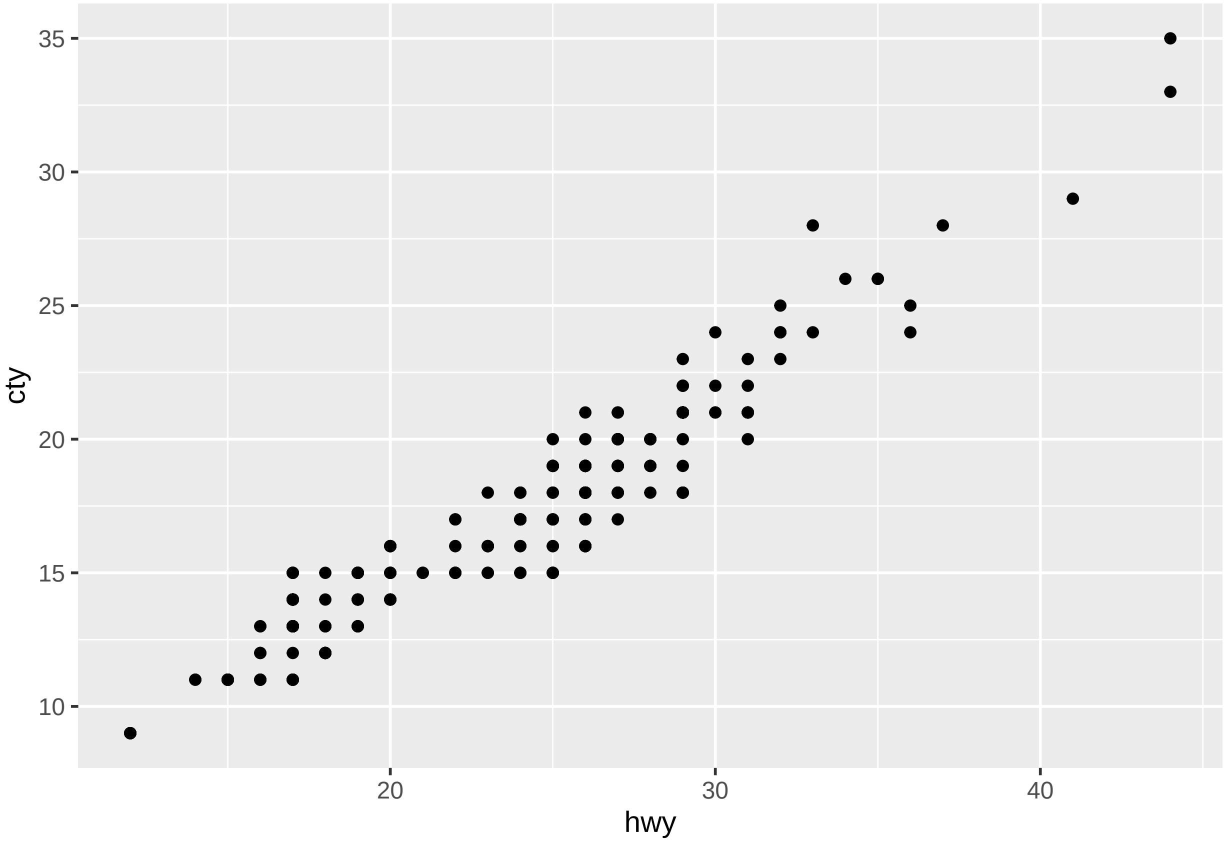
```
ggplot(mpg) + geom_histogram(mapping = aes(x=hwy), binwidth = 1) + labs(x="highway miles per gallon")
```



From the histogram, we can see that most cars have hwy in the range 10 to 40. There are 3 hwy values that have outstandingly high counts, which are around 17, 26, and 29.

Exercise 2

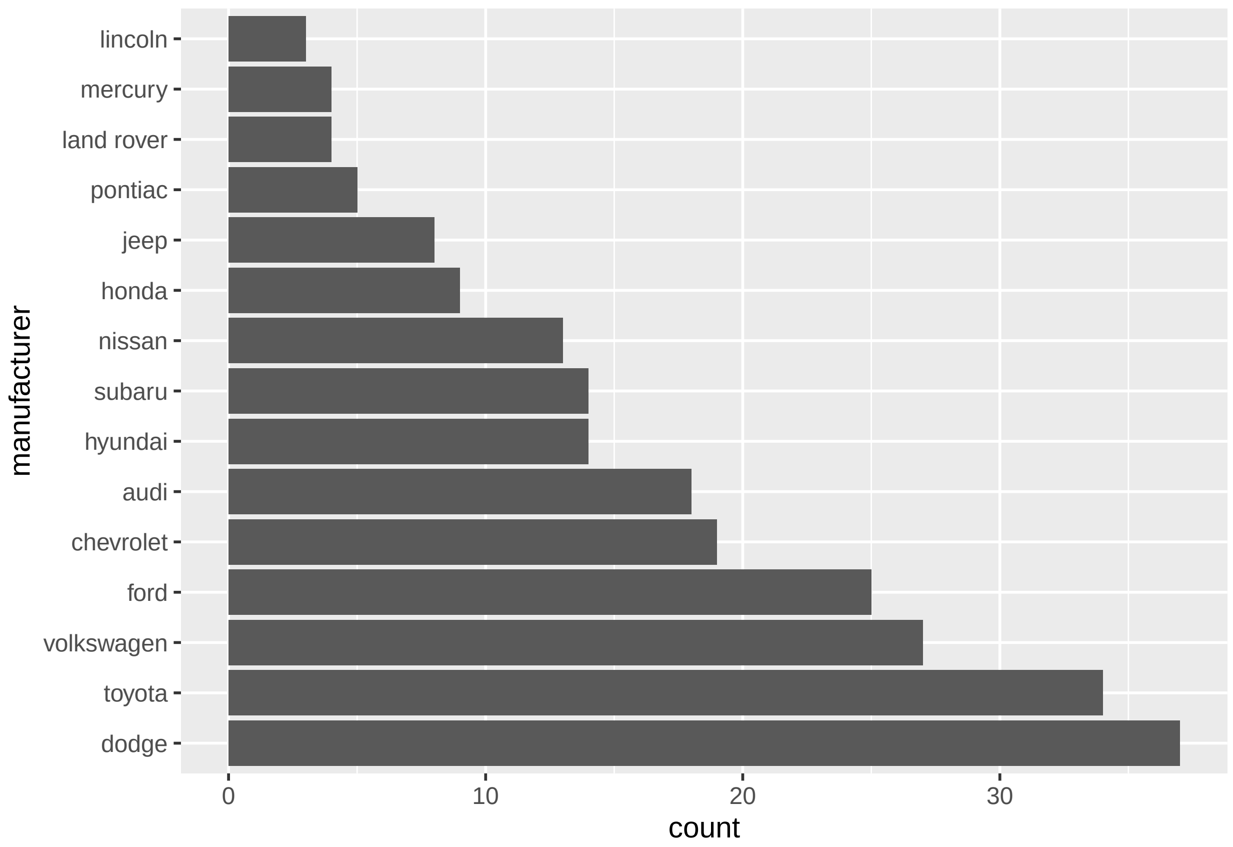
```
ggplot(mpg) + geom_point(mapping = aes(x=hwy, y=cty))
```



From the scatterplot, we can see that there is a almost positive linear relationship between hwy and cty, which means that cty increases as hwy increases and the increments of hwy and cty are nearly constantly proportional. In the context of this dataset, cars that have higher highway miles per gallon also have higher city miles per gallon, and cars that have lower highway miles per gallon also have lower city miles per gallon

Exercise 3

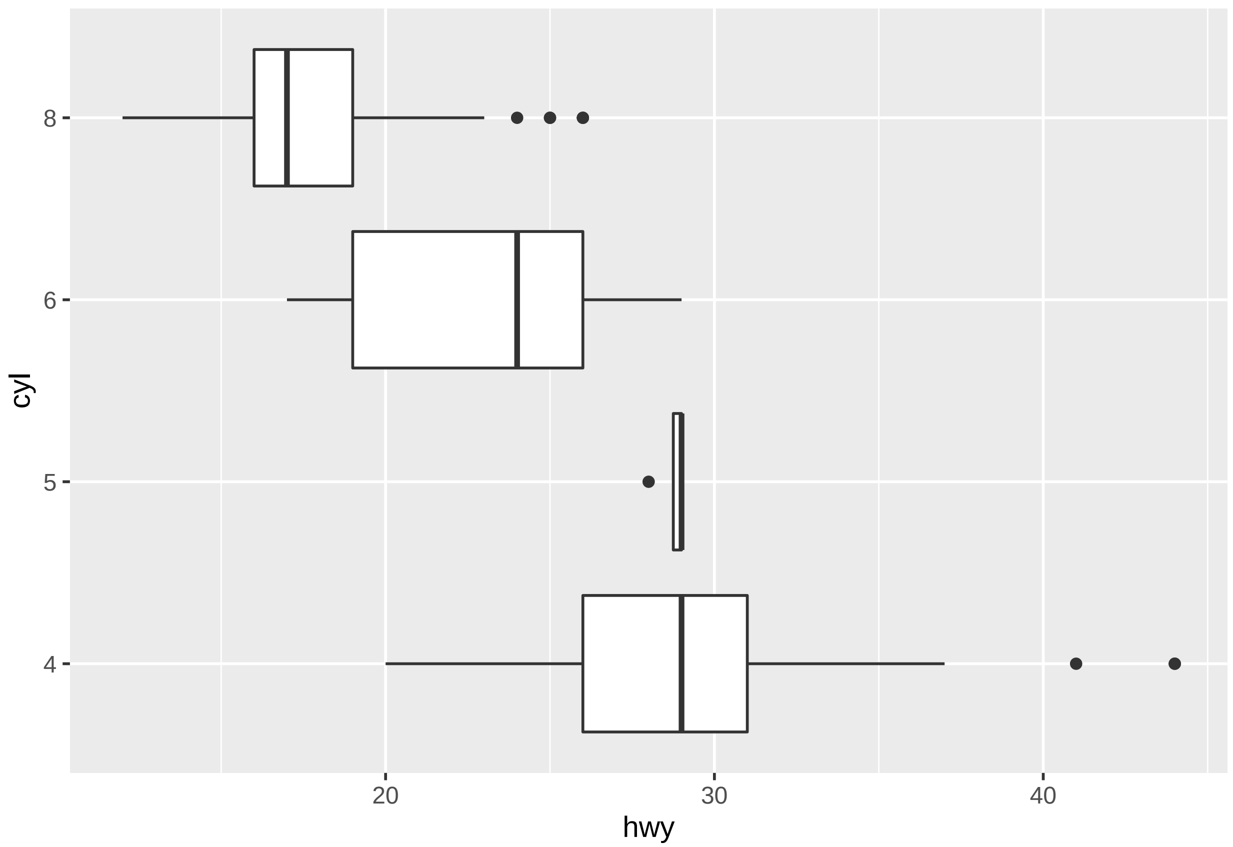
```r
ggplot(mpg) +
  geom_bar(mapping = aes(x=reorder(manufacturer,manufacturer,function(x)-length(x)))) +
  labs(x="manufacturer") +
  coord_flip()
```



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

Exercise 4

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



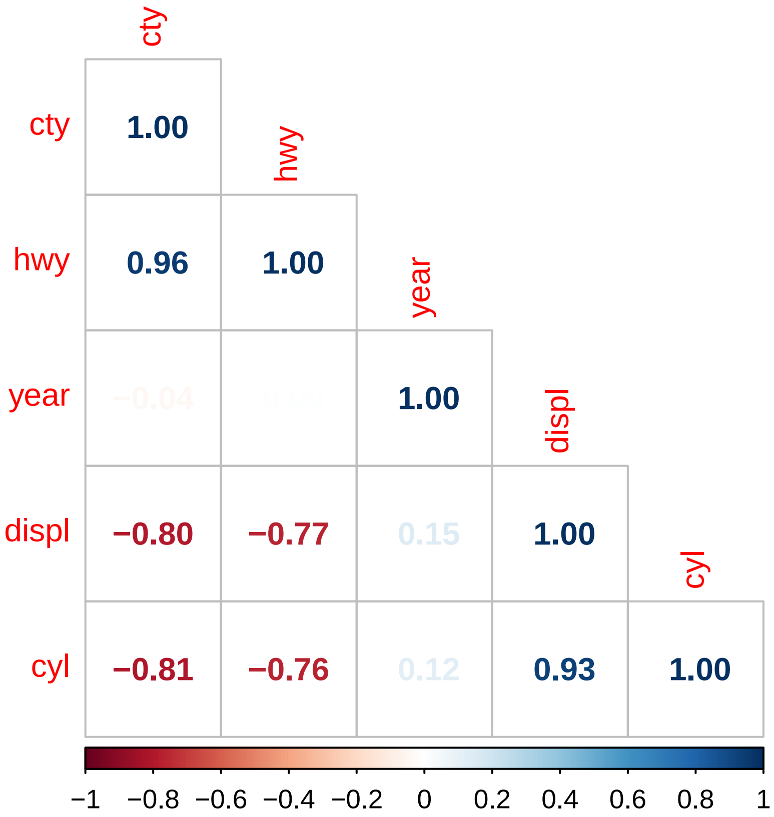
There is a pattern in the box plot. Cars with fewer cylinders have higher hwy on average.

Exercise 5

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

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



hwy and cty, cyl and displ are positively correlated.
Year are nearly not correlated with other variables.
displ and cty, displ and hwy, cyl and cty, cyl and hwy are negatively correlated.
All of these relationships make sense to me.