HW1

Xilong Li (3467966)

2022-04-02

Machine Learning Main Ideas

Qestion 1

1) Supervised learning means that there is an existing labeled dataset that can be learned from, as if there is an "answer key" such that we can "supervised" the machine on its accuracy of prediction and modeling.

2) Unsupervised learning means that there is no such labeled dataset that can be learned from. Thus, we cannot "supervise" its learning because we don't even know what the data should look like.

Another difference between supervised and unsupervised learning is that, while supervised learning involves regression and classification to establish models, unsupervised learning uses methods such as clustering to learn from unknown models.

Qestion 2

1) For the regression model, the response variable is quantitative, which means that it is numeric and can be directly calculated. Thus, the regression model allows us to predict trans based on existing labeled data. For example, we can use the existing data to establish model about the gas price for next month.
2) For the classification model, the response variable is qualitative, so it cannot be directly calculated as numbers. For example, classification model can be established on issues about whether a person is graduate or not, which is not numeric and is thus qualitative.

Qestion 3

1) Two commonly used metrics for regression ML problems: Mean Squared Error (MSE) and Root Mean Squared Error (RMSE);

2) Two commonly used metrics for classification ML problems: Accuracy and F1-score;

Qestion 4

1) Descriptive: Choose model to best visually emphasize a trend in data;
2) Inferential: what features are significant? — to test theories and state relationship between outcome & predictors;
3) Predictive: What combo of features fits best? — to predict Y with minimum reducible error
(Answers above are cited from Lecture 2)

Qestion 5

"A mechanistic model uses a theory to predict what will happen in the real world. The alternative approach, empirical modeling, studies real-world events to develop a theory." (cited from https://smallbusiness.chron.com/mechanistic-model-12706.html)
In addition, Mechanistic model assume a parametric form for f, while empirical model does not make any assumptions about f. Empirical model also requries a larger number of observations than mechanistic model. (cited from Lecture)

Usually, Empirical model has much more flexibility than Mechanistic model, since mechanistic model is parametric while the other is not parametric. Therefore, since more flexibility usually means less interpretability and sometimes more error, I think mechanistic-driven model is easier to understand.

According to the Bias-Variance trade-off, we always want to find a model that has low variance and low bias. However, we usually have a trade-off between the level of variance and bias. And also because Mechanistically-driven model has higher flexibility, which as a result will have low bias and high variance. On the opposite side, simpler model will have higher bias and lower variance. Therefore, the bias-variance trade-off would affect our selection between mechanistic and empirical model. But still, we want to find a model that has lowest variance and bias. And we also want to avoid overfitting in the model selection.

Qestion 6

1) The first question is predictive, because a data set about the voter is given in order to predict the likely choice that he/she will make in voting for the candidate.
2) The second question is inferential, because it aims to understand the relationship between "whether the vote has personal contact with the candidate" and "voter's likelihood of support for the candidate". This question strives to understand how these two factors are related.

Exploratory Data Analysis

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

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

```
library(ggplot2)
head(mpg)
```

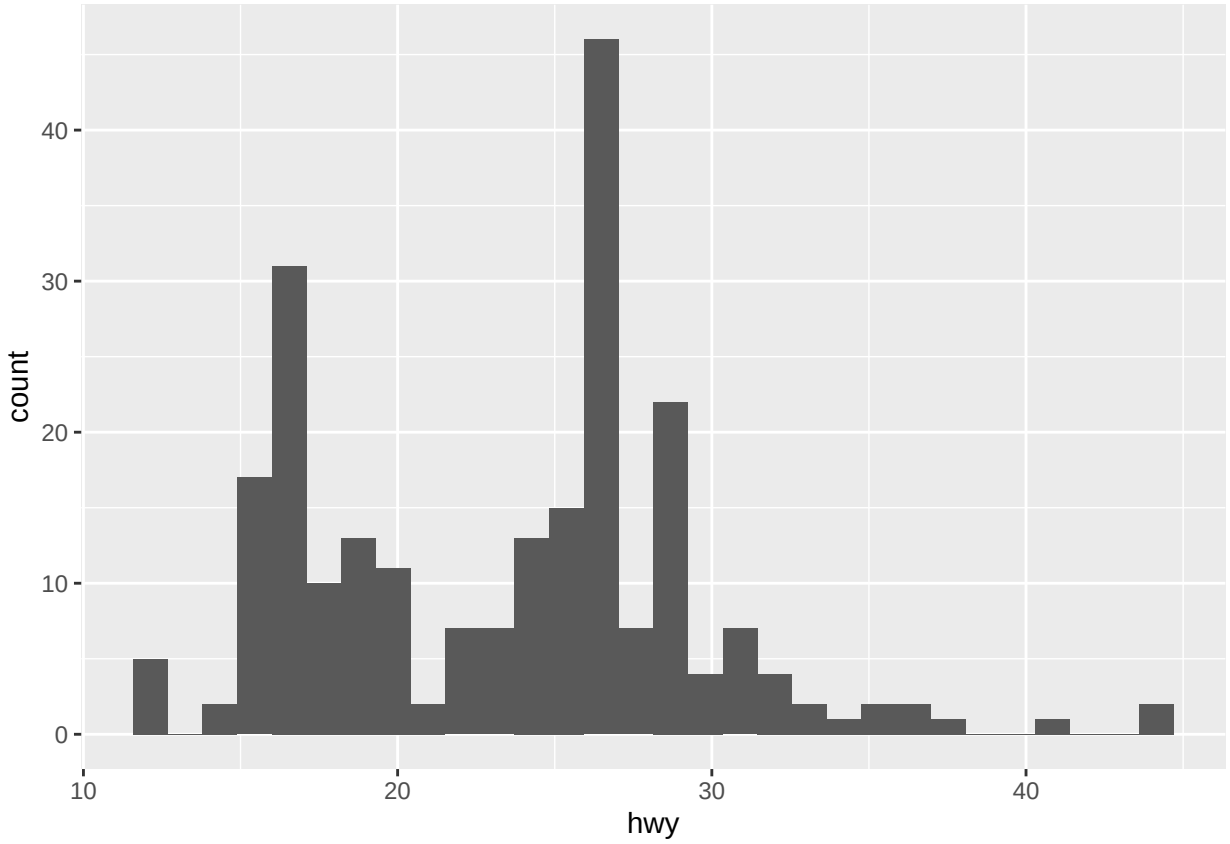
```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
```

```
## 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~
```

Qestion 1

```
ggplot(mpg, aes(hwy)) + geom_histogram()
```

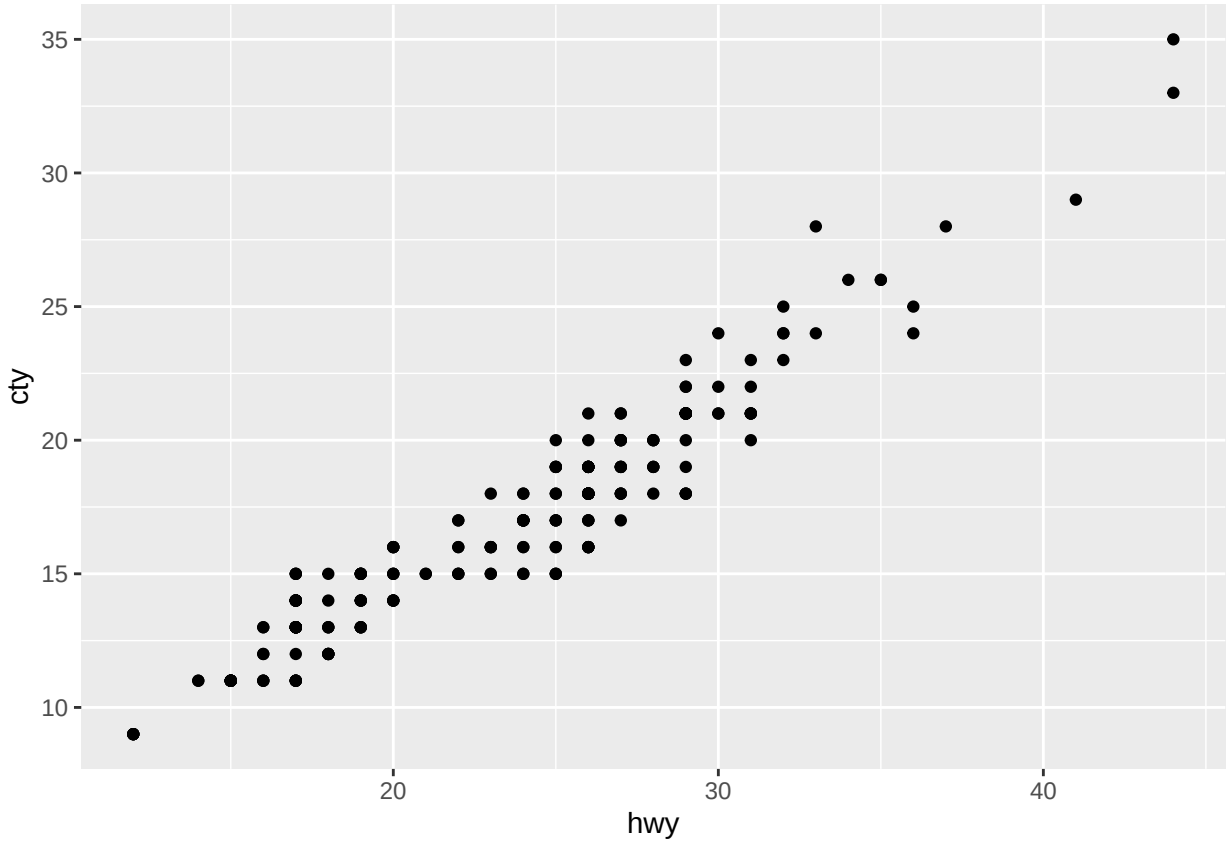
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As it is shown on the graph above, most highway miles per gallon data is less than 30. And there are two peaks on the graph: the first peak is between between 15 and 20, and the second peak is between 25 and 30.

Qestion 2

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



As the graph above has shown, there is a clear lineal relationship between hwy and cty, since all points are aligned to form an approximate line with positive slope.
This means that there is a positive linear relationship between hwy and cty: as hwy increases, cty would also increase.

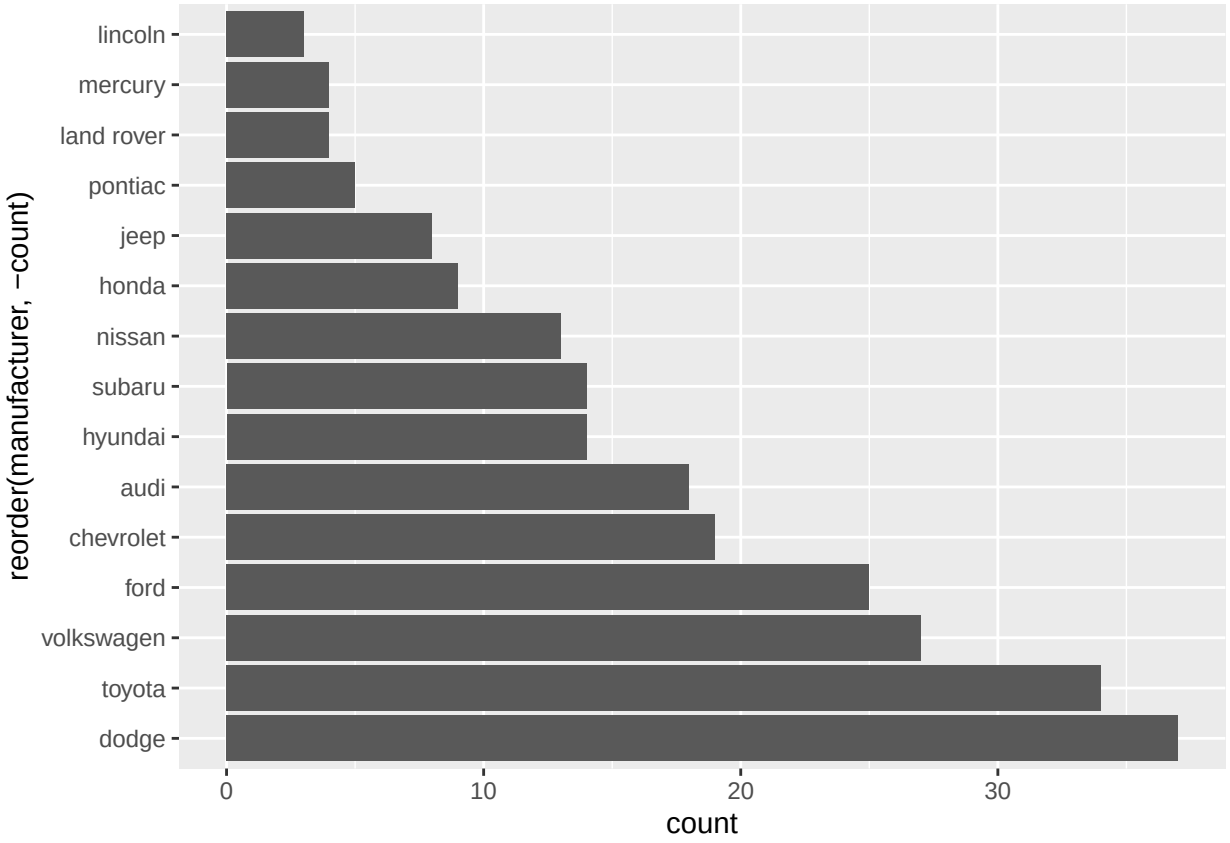
Qestion 3

```
ordered_data <- mpg %>%
  group_by(manufacturer) %>%
  summarise(count = n()) %>%
  arrange(count)
ordered_data
```

```
## # A tibble: 15 x 2
##    manufacturer count
##    <chr>        <int>
##  1 lincoln          3
##  2 land rover       4
##  3 mercury          4
##  4 pontiac          5
##  5 jeep             8
##  6 honda            9
##  7 nissan          13
##  8 hyundai         14
##  9 subaru          14
```

```
## 10 audi            18
## 11 chevrolet       19
## 12 ford            25
## 13 volkswagen      27
## 14 toyota          34
## 15 dodge           37
```

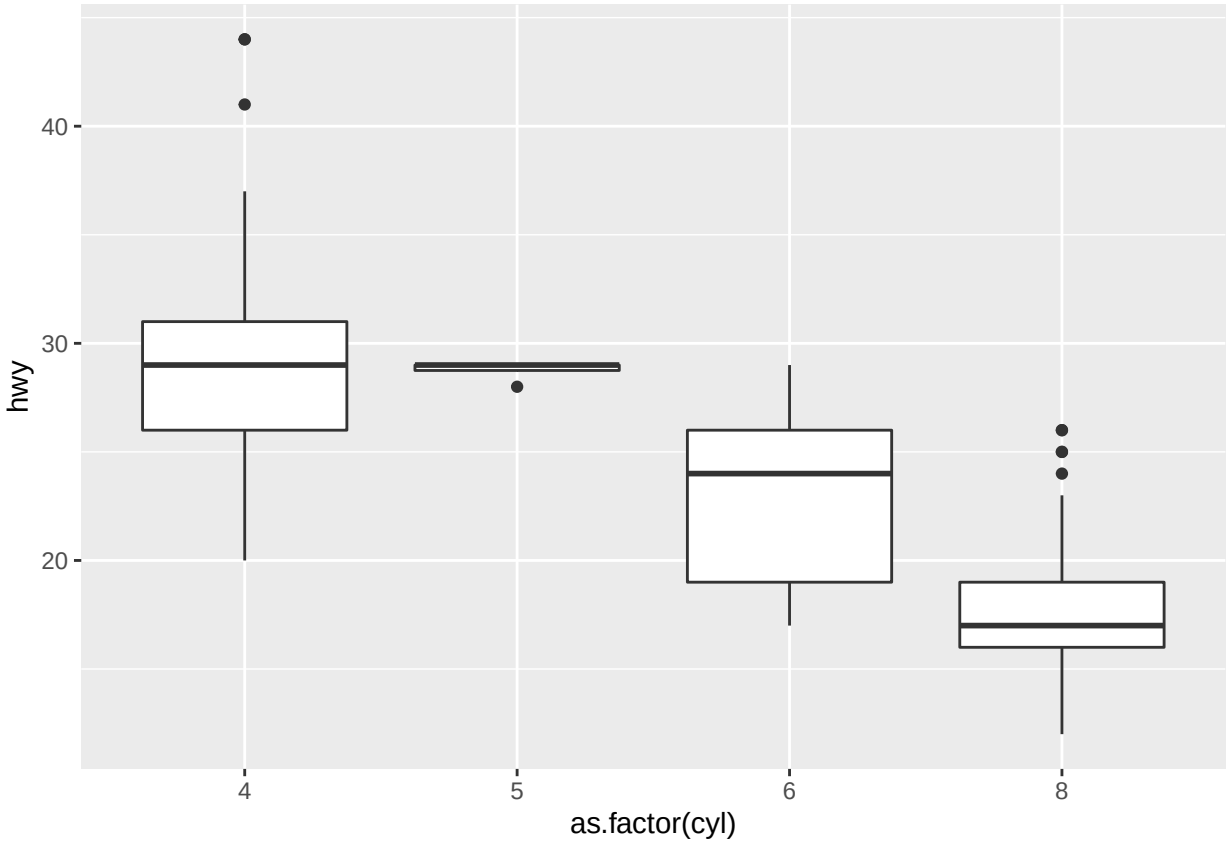
```
ggplot(ordered_data, aes(x = count, y = reorder(manufacturer, -count))) + geom_bar(stat = "identity")
```



Therefore, as the graph above has shown, dodge produced the most cars, and lincoln produced the least car.

Qestion 4

```
ggplot(data=mpg, aes(x= as.factor(cyl), y = hwy)) + geom_boxplot()
```



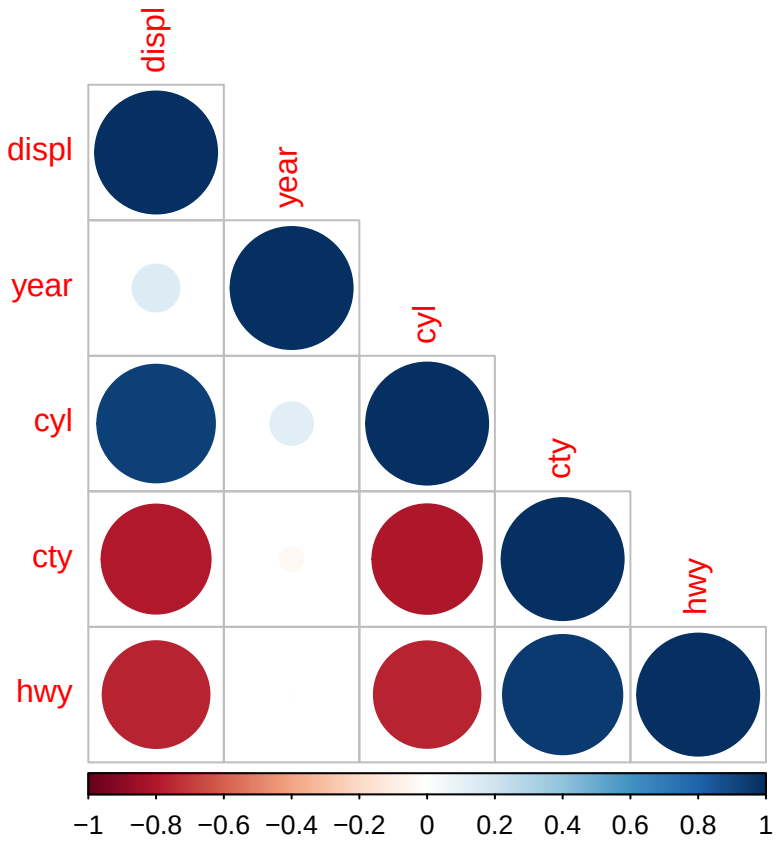
From the graph shown above, we can see that there is indeed a pattern. As cyl increases, hwy decreases. This means that there exists a negative relationship between cyl and hwy, which makes intuitive sense because as the car has more cyl, it would consume more gas and thus has lower miles per gallon.

Qestion 5

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

```
## corrplot 0.92 loaded
```

```
# select numeric parameters from mpg:
selected_mpg <- mpg %>%  select (-c(manufacturer, model, trans, drv, fl, class))
corrplot(cor(selected_mpg), type = 'lower')
```



As it is shown on the graph above: 1) displ has positive correlation with cyl; 2) displ has negative correlation with cty; 3) displ has negative correlation with hwy; 4) cyl has negative correlation with cty; 5) cyl has negative correlation with hwy; 6) cty has positive correlation with hwy;

Those correlation all make sense to me:
1) it is intuitive that the more cyl a car has, the more gas it would consume, and the less miles per gallon it would have, regardless whether it is on highway or city road. 2) it also makes sense there is a positive correlation between hwy and cty, since if a car has high miles per gallon on the high way road, it makes sense that it would also have high mpg on city road. 3) I don't understand what displ (engine displacement) is because I never heard of this term, but I guess it would make sense too :-)