pstat131_hw1

Cyril Wang

Machine Learning Main Ideas:

Question 1:

Supervised learning is when there is both input and output data, and that there is a related response y_i for each of the predictors x_i. We then find a model that relates the response to the predictors, in order to predict future responses or learn more about how the predictors relate to the response. Unsupervised learning has input data ($x_i's$), but has no response $y_i's$, or there is no output data, so we cannot fit a model. They differ in that unsupervised has no response that can "supervise" the learning.

Question 2:

The difference between a regression model and a classification model is the response variable (Y) of regression models are quantitative (Y takes on numerical values), while classification models responses are qualitative (Y takes on categorical values).

Question 3:

Two commonly used metrics for regression ML problems are the test and training mean squared error, which measures how close the predictions are to the actual responses for the test data and the training data. If the MSE is small, that means that the predictions and actual responses are fairly close. Two commonly used metrics for classification ML problems are the test and training error rate, which is the proportion of incorrect classification (from lecture slides day2) in the test and training data.

Question 4:

Descriptive models: statistical models that are used in order to visualize the patterns / trends in the data

Predictive models: statistical models that aim to predict the value of the response variable Y based on the predictors

Inferential models: statistical models that aim to understand the relationship between the response variable Y and the predictors X_i, and to see which predictors are significant.

Question 5:

Mechanistic models are ones where we assume that our function such that $Y = f(X_1 \dots X_p) + \epsilon$ has a parametric form, so rather than estimating f, we can estimate the coefficients of X instead. That being

said, the chosen model will usually be different from f, and can make our estimate poor. Although we can use flexible models, it requires a large number of parameters, which leads to overfitting.

Empirically-driven models are ones where we actually estimate f that matches or gets close to matching the actual data points, rather than assuming the form of f. Unlike mechanistic models, there is not the risk of having the model differ greatly from f, but does require a large number of observations. But similar to mechanistic models, they can result in overfitting and result in poor predictive estimates.

In general, a mechanistic model is easier to understand for me. Typically, there is a trade-off between model flexibility and interpretability, and such the restrictive mechanistic models are more interpretable. For inference, the coefficients in the mechanistic models make describing the relationship between a predictor and the response very simple.

The bias-variance tradeoff is related to the use of mechanistic or empirically-driven models, as while more flexible statistical models (like empirically-driven modesl)have higher variance, simpler and less flexible models (like mechanistic models) are likely to have greater bias due to differences in the real model and the estimated model resulting in poorer estimates. So, when choosing to use mechanistic models, we sacrifice bias in exchange for variance and vice versa for empirically-driven models.

Question 6:

The first question (how likely is it that they will vote in favor of the candidate?) would be predictive, as we are trying to estimate/predict the candidate's response using their profile as the predictors.

The second question (How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?) would be inferential, as we are trying to understand the association between a voter's support for the candidate (the response) and if they had contact with the candidate (a predictor).

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   2.1.2     v forcats 0.5.1

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

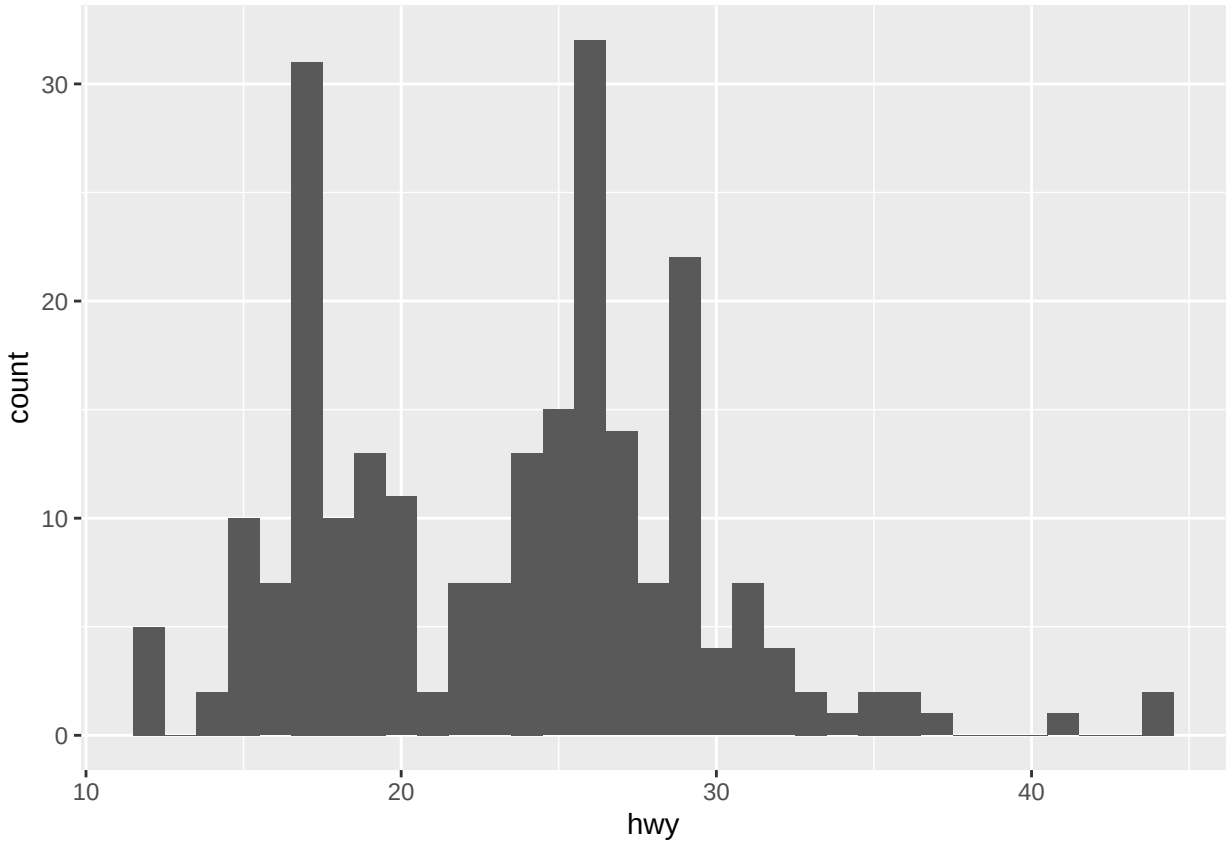
```
library(ggplot2)
summary(mpg)
```

```
##  manufacturer          model               displ            year
##  Length:234         Length:234         Min.   :1.600   Min.   :1999
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999
##  Mode  :character   Mode  :character   Median :3.300   Median :2004
##                                        Mean   :3.472   Mean   :2004
```

```
##                                        3rd Qu.:4.600   3rd Qu.:2008
##                                        Max.   :7.000   Max.   :2008
##       cyl           trans               drv                 cty
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00
##  Mean   :5.889                                         Mean   :16.86
##  3rd Qu.:8.000                                         3rd Qu.:19.00
##  Max.   :8.000                                         Max.   :35.00
##       hwy             fl                class
##  Min.   :12.00   Length:234         Length:234
##  1st Qu.:18.00   Class :character   Class :character
##  Median :24.00   Mode  :character   Mode  :character
##  Mean   :23.44
##  3rd Qu.:27.00
##  Max.   :44.00
```

Exercise 1

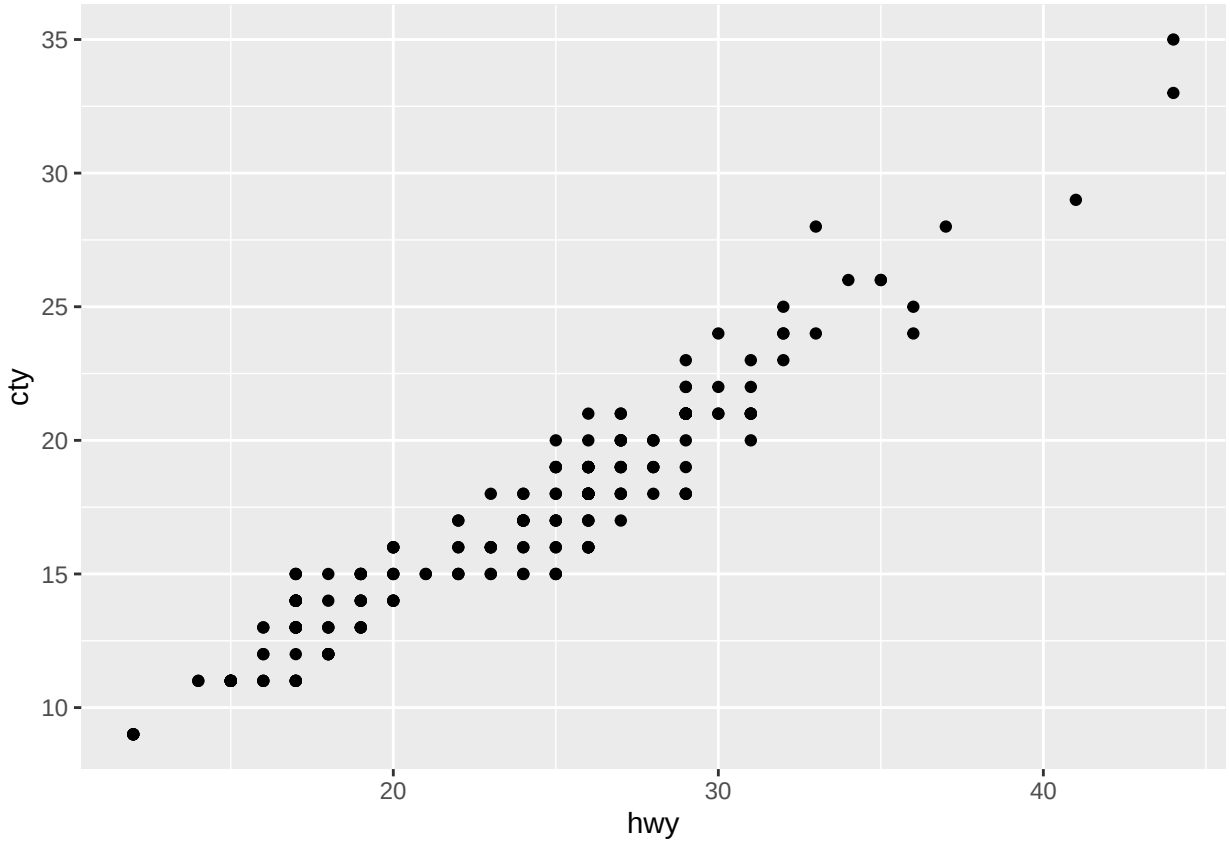
```
ggplot(mpg, aes(x=hwy)) + geom_histogram(binwidth = 1)
```



In the fuel economy data from 1999 to 2008 for 38 popular models of cars, the highway miles per gallon appears to be bi-modal, centered around 17 mpg and 26 mpg. For the most part, most of the values are between 15 and 30 mpg, but there are some points that may possibly be outliers, with values over 40 mpg.

Exercise 2

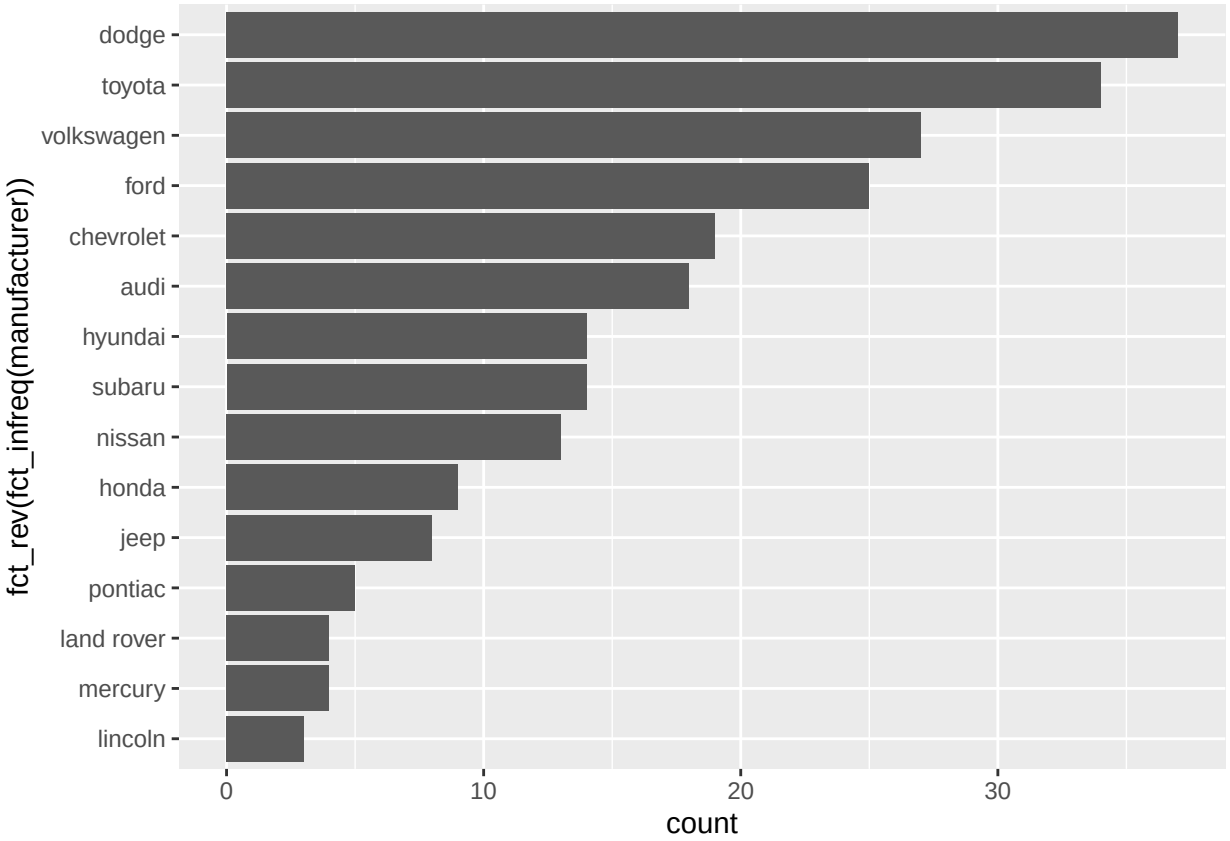
```
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point()
```



I notice that there is a linear relationship between hwy and cty, or that as the highway miles per gallon increases, city miles per gallon also increases.

Exercise 3:

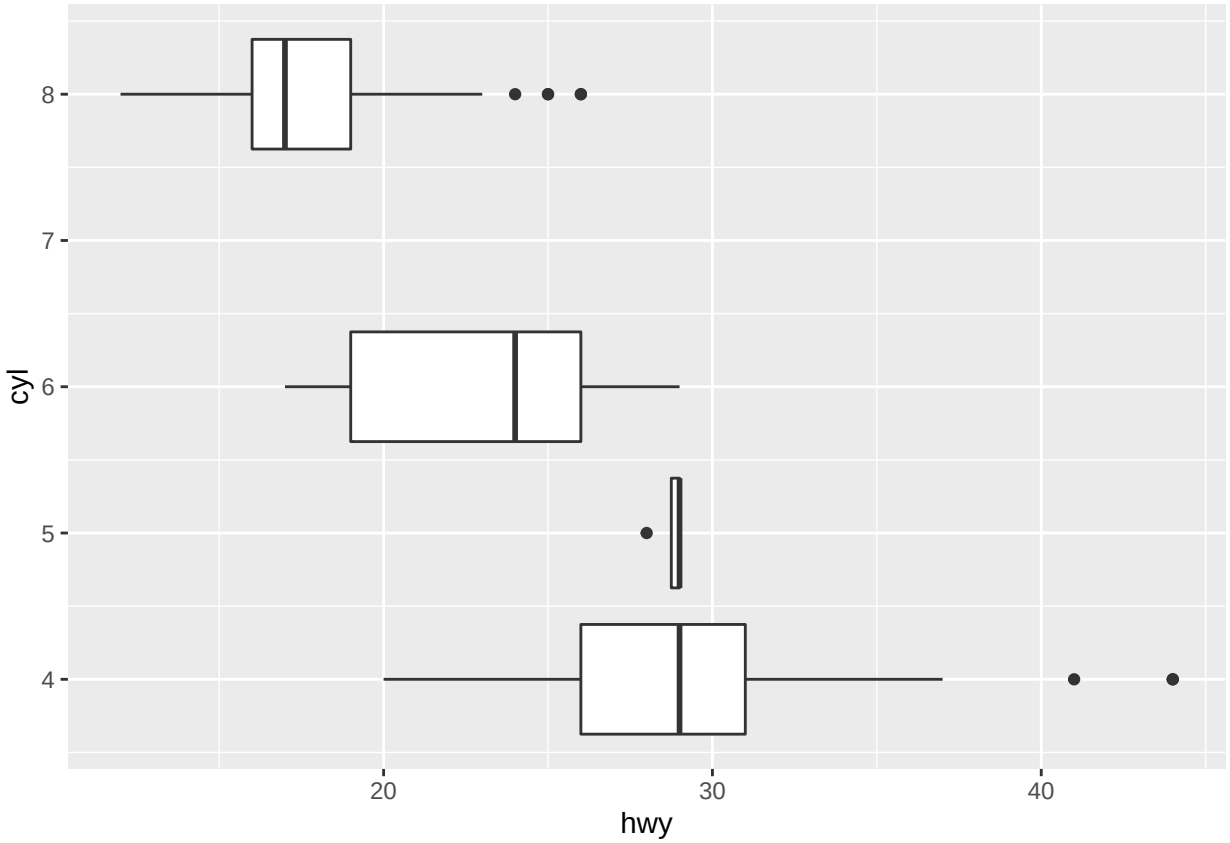
```
ggplot(mpg, aes(y=fct_rev(fct_infreq(manufacturer)))) + geom_bar()
```



The manufacturer that produced the most cars is Dodge, while the one that produced the least is Lincoln.

Exercise 4:

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



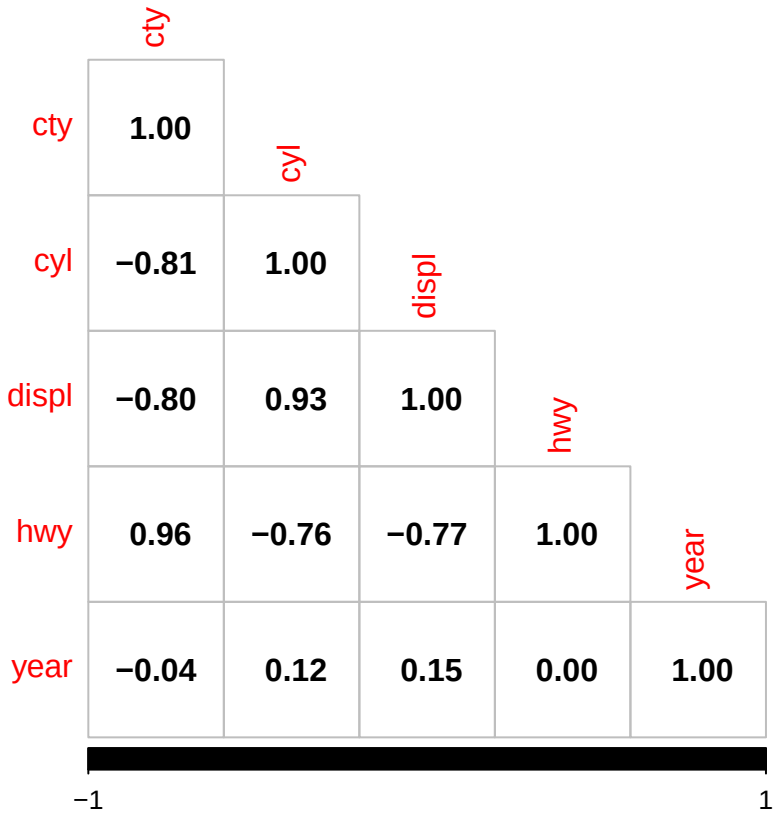
I notice that, on average, the highway miles per gallon increases as the number of cylinders decreases (i.e. that there is a negative correlation between hwy and cyl).

Exercise 5:

```
library("corrplot")
```

```
## corrplot 0.92 loaded
```

```
cor(select(mpg,where(is.numeric))) %>%
corrplot(method='number', order='alphabet', type= 'lower', col='black')
```



(I chose to make the text color black because it was hard to see some of the values) Of the numerical variables, cty is strongly, negatively correlated with cyl and displ, a strong, positive correlation with hwy. Cyl has a strong, positive correlation with displ and a strong, negative correlation with hwy. Displ has a strong, negative correlation with hwy. Year has a weak correlation with all the variables, with cty being negative and the others being positive. These make sense to me, that having more cylinders (where fuel is combusted) results in lower miles per gallon, as more fuel is consumed in exchange for more power. With a quick google search, it appears that the calculation for engine displacement involves the number of cylinders, explaining why those variables are correlated. The one that intrigues me is the correlation between year and displ. One would expect that as year increases, technology improves and fuel-efficiency increases, which would mean that year and displ should theoretically be positively-correlated. We do see that they are, but the correlation between the two is fairly weak, and I am a little suprised that it isn't higher.