PSTAT 131 HW 1

Eric

3/31/2022

1. Supervised learning is used when the data given to the model is labeled, which means that the model is able to see the observed input and output. Supervised learning predicts the outcome, Y, which can be categorical (classification) or numerical (regression). Supervised learning models include: linear regression, logistic regression, k-nearest neighbors, decision trees, random forests, support vector machines, and neural networks.
   Unsupervised learning is used when the data given to the model is unlabeled, which means that the model has no "answer key" or ground truths. This means that there is essentially no response variable. Unsupervised learning models include: PCA, k-means clustering, hierarchical clustering, and neural networks.
   The main difference is that supervised learning uses an "answer key", while unsupervised learning uses no "answer key". This leads to the fact that supervised learning is classification(or regression), while unsupervised learning has its own way to make sense of the data (eg. clustering). Due to this, different models are utilized depending on which type of learning method (supervised/unsupervised) is used.

2. The difference between a regression model and a classification model is that a regression model will predict numerical outputs, while a classification model will predict a categorical output.

3. Two commonly used metrics for regression ML problems are: Training MSE and Test MSE. Two commonly used metrics for classification ML problems are: Training Error Rate and Testing Error Rate.

4. Descriptive Models: We use this if we want to "best visually emphasize a trend in data" (lecture notes day 2). An example is using a line on a scatterplot.
   Inferential Models: We use this if we want to see "which features are significant" or "state the relationship between outcome and predictor(s)" (lecture notes day 2). The aim is to test theories and possibly make causal claims.
   Predictive Models: We use this if we want to "see which combo of features works best" and "predict Y with minimum reducible error" (lecture notes day 2). They are not focused on hypothesis tests, rather, we're usually focused on just making predictions about the future.

5.

i) In ML, mechanistic means that we assume that there is some parametric function that fits that data which we try to get as close as possible to. Generally, mechanistic describes something that relates to theories to explain phenomena.
   In ML, empirically-driven means that there is no assumption about the function that fits the data. Generally, empirically-driven describes something that is derived from or relates to experiment and observation (rather than theory).
   Differences: Mechanistic approach assumes a parametric function that fits the data. Empirically-driven approach assumes nothing about a function that fits the data and tends to require larger amounts of observations. The empirically driven approach also is more flexible by default.
   Similarities: Both are prone to overfitting the data.

ii) In general, a mechanistic model is easier to understand because you are assuming that there is a parametric function that exists that could explain the data. This means that we are probably well informed enough about the situation to form a theory. Additionally, we have parameters that we could interpret in a mechanistic model– whereas using an empirical model would not give us the luxury of analyzing the parameters.

iii) The bias-variance tradeoff is related to the use of mechanistic or empirically-driven models because it allows us to analyze when our model is overfitting or underfitting. A low complexity model has high bias and low variance, while a high complexity model has low bias and high variance. We want to find a model that has the smallest amount of bias and variance.

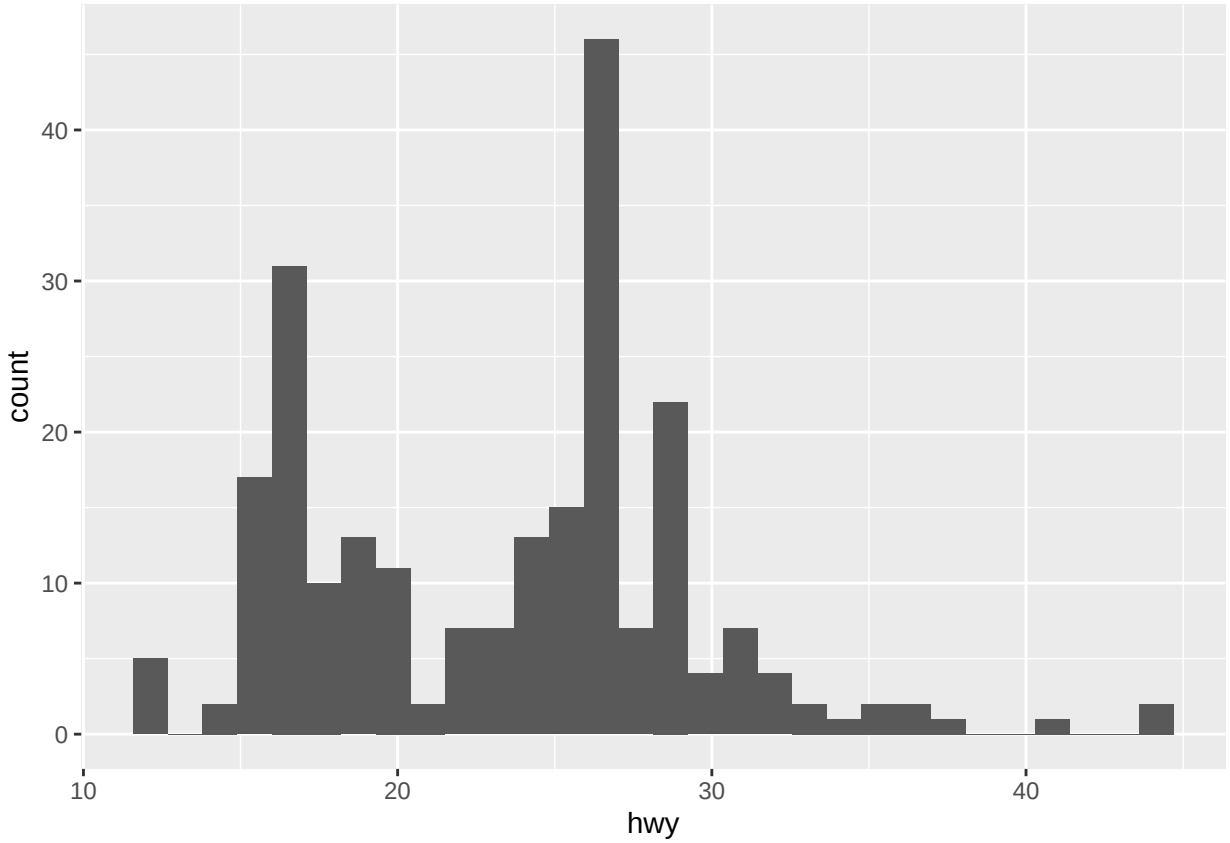
6. "Given a voter's profile/data, how likely is it that they will vote in favor of the candidate?"
   I would say that this is a predictive problem because we want to predict the outcome (likeiness of voting in favor of the candidate) based on a given set of attributes.
   "How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?"
   I would say that this is an inferential problem because we want to see if there is a relationship between the outcome and a particular predictor (personal contact w/ candidate and likelihood of support).

Exploratory Data Analysis

```
library(tidyverse)
library(ggplot2)
```

Exercise 1

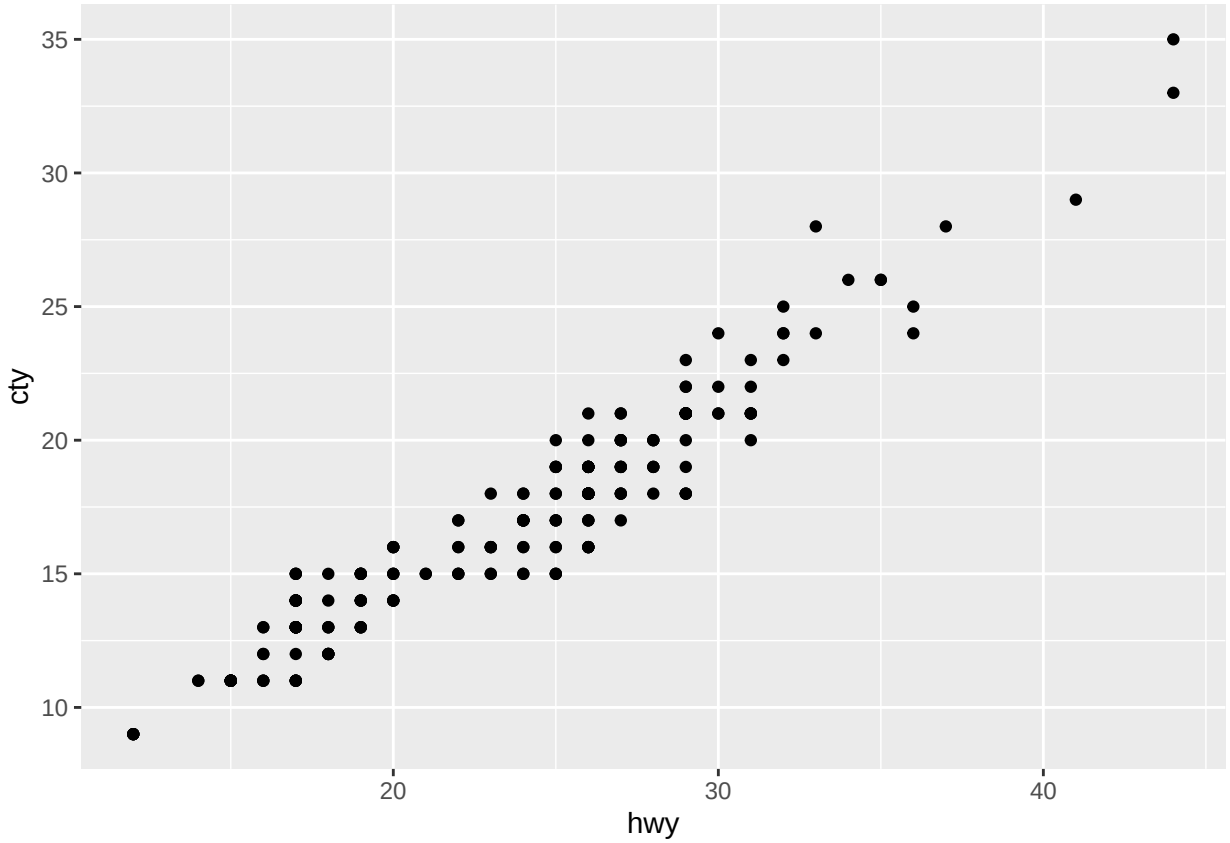
```
ggplot(mpg, aes(x = hwy)) + geom_histogram()
```



I see that the distribution of highway miles per gallon has two peaks. It seems that most cars seem to have highway miles per gallon in the mid 10's and mid 20's. Very few people have highway miles per gallon above 30.

Exercise 2

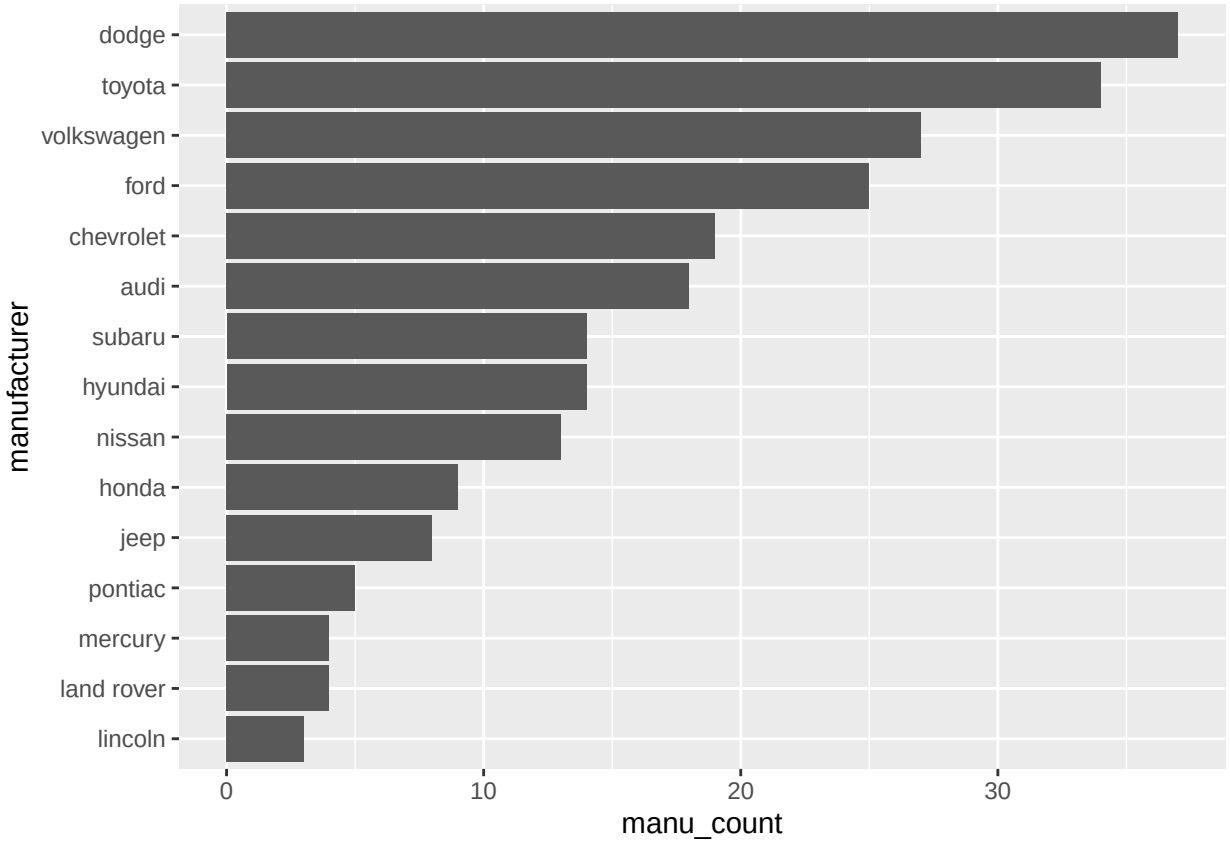
```
ggplot(mpg, aes(x = hwy, y = cty)) + geom_point()
```



I notice that there is a linear trend on the scatterplot. There seems to be a positive relationship between hwy and city. This means as hwy increases/decreases, cty increases/decreases as well.

Exercise 3

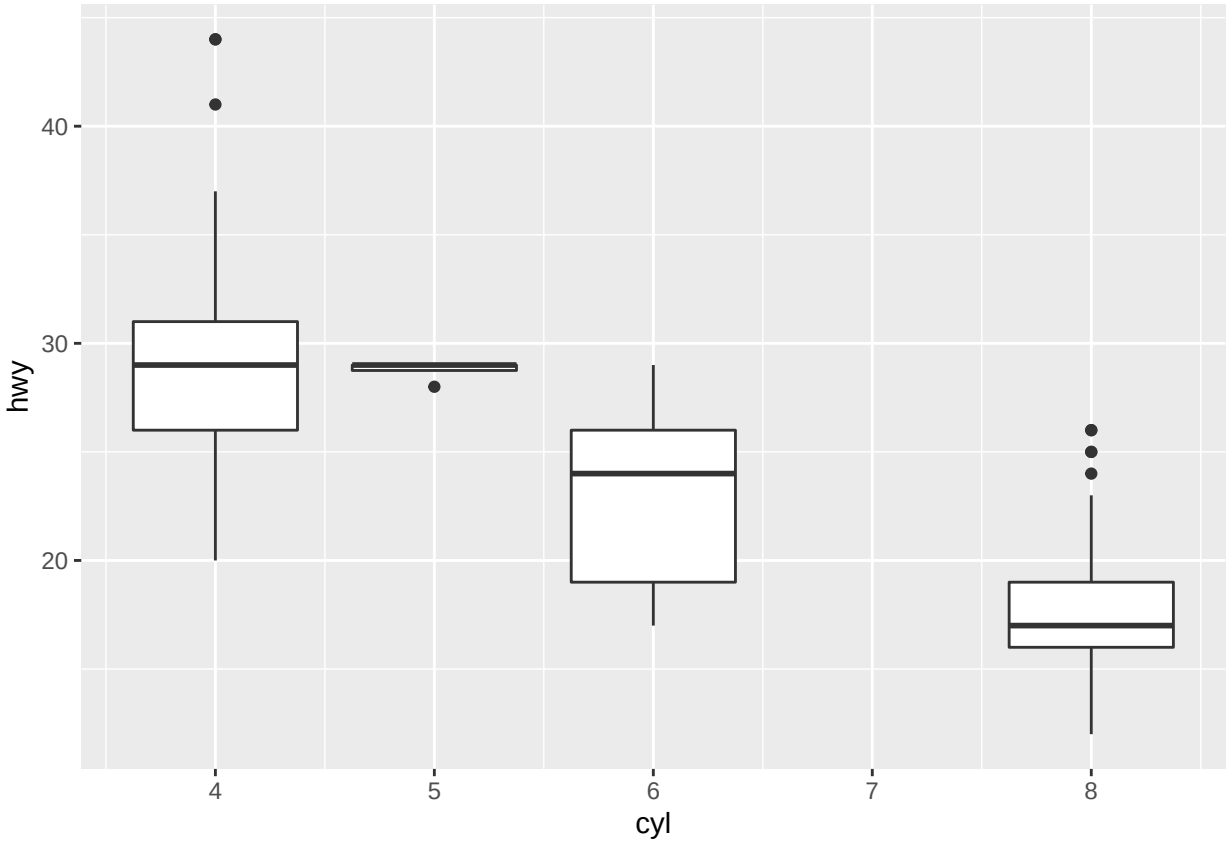
```
mpg_sorted <- mpg %>% group_by(manufacturer) %>% summarise(manu_count = n()) %>% arrange(manu_count)
mpg_sorted$manufacturer <- factor(mpg_sorted$manufacturer, levels = mpg_sorted$manufacturer)
x <- ggplot(mpg_sorted, aes(x = manufacturer, y = manu_count)) + geom_bar(stat = "identity")
x + coord_flip()
```



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

Exercise 4

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

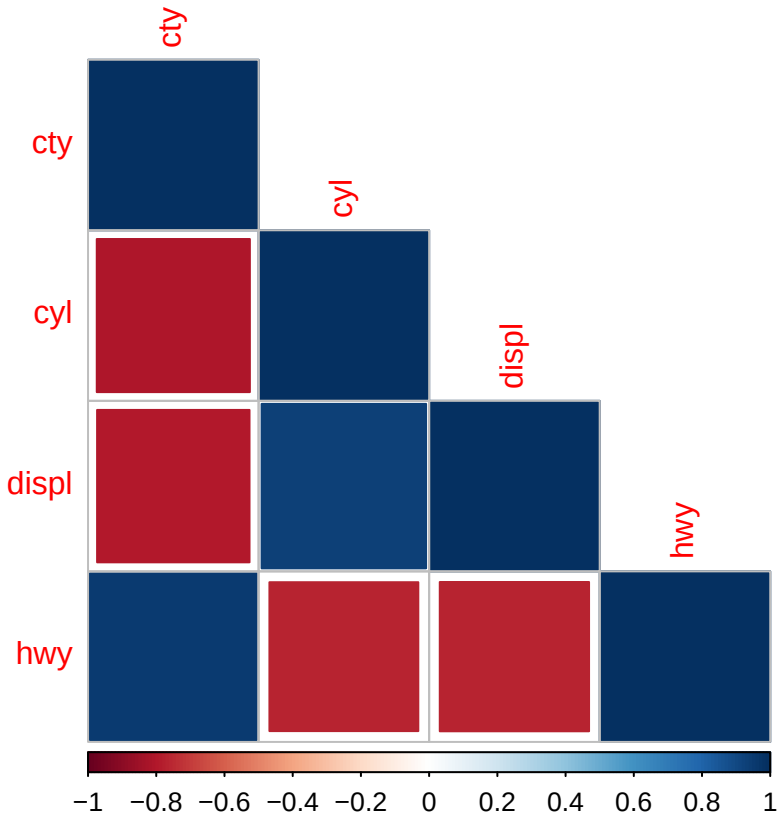


There does seem to be a pattern where a lower number of cylinders tends to have a larger number of highway miles per gallon.

Exercise 5

```
library(corrplot)
```

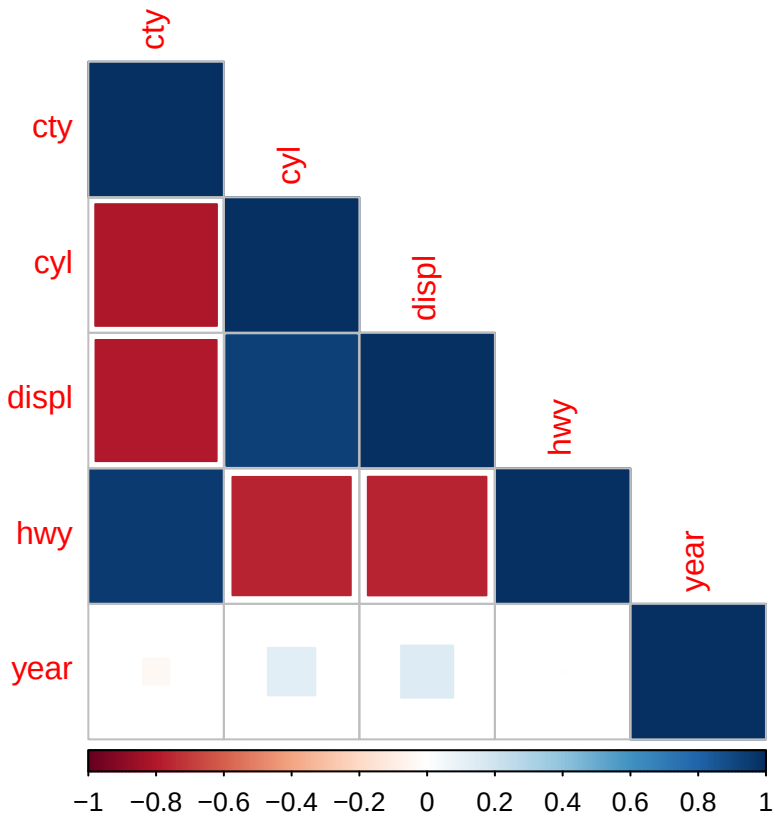
```
mpg_numerical <- mpg %>% select(displ, cyl, cty, hwy)
corr_mpg <- cor(mpg_numerical)
corrplot(corr_mpg, method = "square", order = "alphabet", type = "lower")
```



(**Note**: After talking to others and consulting the internet, I decided not to include year into this correlation chart. The reasoning is that the use of year here is a categorical variable (similar to a zip code), so it probably would not make sense to see if it has correlation with other variables. However, if it is required for the answer, year had no significant correlation with other variables– year was very slightly positively correlated with cyl and displ. I included an extra figure below)
The variables that are positively correlated are cty and hwy, and displ and cyl.
The variables that are negatively correlated are displ and cty, displ and hwy, cyl and cty, and cyl and hwy.
These relationships make sense because if two variables are strongly and positively correlated, then it would make sense that both of them would have similar correlations to another given variable in the dataset. This is why cty and hwy are both negatively correlated with displ and cyl.
Thinking about this problem in a non-mathematical way, these relationships do make sense. If a car has good city mpg, then it should almost certainly have the same/better highway mpg. Additionally, the engine displacement calculation depends on the number of cylinders. A higher number of cylinders means more engine displacement, so the positive correlation between displ and cyl makes sense as well.

Alternate Correlation Plot with Year Included

```
mpg_numerical <- mpg %>% select(displ, cyl, cty, hwy, year)
corr_mpg <- cor(mpg_numerical)
corrplot(corr_mpg, method = "square", order = "alphabet", type = "lower")
```


cty
cyl
displ
hwy
year
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1