PSTAT 131 Homework 1

Brett Goldman

9/28/2022

Maching Learning Main Ideas: Question 1

Define supervised and unsupervised learning. What are the difference(s) between them?

From the lecture, we know that supervised learning is when the response variable acts as a supervisor. The model is essentially given the correct answer to learn from in order to learn and improve itself. On the other hand, unsupervised learning is when the response variable is unknown to the model. The model has to learn from its mistakes without knowing what the truth looks like. The key difference between the two is the knowledge of the response variable in supervised learning, and the lack thereof in unsupervised learning.

Question 2

Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

From the lecture, simply put, a regression model is one where the response is quantitative. A classification model is one where the response is qualitative.

Question 3

Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

Two commonly used metrics for regression ML problems are Quality of Fit (MSE) and Bias-Variance tradeoff. Two commonly used metrics for classification ML problems are Bayes Classifier and K-Nearest Neighbors Classifier.

Question 4

As discussed, statistical models can be used for different purposes. These purposes can generally be classified into the following three categories. Provide a brief description of each.

Descriptive models:

Descriptive models choose a model to “best visually emphasize a trend in data” (from lecture). An example of this that we went over in lecture is a trend line on a scatterplot to show the direction that a response variable goes as the predictor increases or decreases.

Inferential models:

Inferential models decide which predictors are significant to the response. A lot of this is testing claims that Predictor A has a greater impact on the response than Predictor B.

Predictive models:

Predictive models try to predict a response variable with minimal error. They don't care about testing which predictor(s) is significant like an inferential model. They just want to get an accurate prediction of the response.

Question 5

Predictive models are frequently used in machine learning, and they can usually be described as either mechanistic or empirically-driven. Answer the following questions.

Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

Mechanistic models make assumptions to predict the outcome. These assumptions are based on some sort of theory. They use a function to assume some sort of relationship between the predictors and response.

On the other hand, empirically-driven models don't make assumptions or theories to predict the outcome. They rely solely on observations to create their model.

These model types differ in whether they make assumptions. Mechanistic models do make assumptions, which means that there is less of a need for as many observations. Empirically-driven models do not make assumptions, which means that they need a large number of observations to be accurate.

These models are similar in the sense that they can potentially both suffer from overfitting. They can both become too accurate to the random noise in the model, which is bad because that random noise varies from case to case.

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.

From my perspective, I would say that an empirically-driven model is easier to understand. If there are a large number of observation and a trend clearly develops, the model should be easy to create and understand. For a mechanistic model, you have to deal with some sort of theory in a field you may not be comfortable in.

Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.

Bias and variance have an inverse relationship, so if you increase bias in your model, you decrease variance and vice-versa. If you decrease variance too much, your model will not be flexible enough. However, if you increase variance too much, your model could be too flexible, where it becomes tough to interpret. There is a middle ground there between flexibility and interpretability. This tradeoff is more often seen in mechanistic models, as the theoretical function you are testing can be impacted by differing bias and variance.

Question 6:

A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:

1) Given a voter's profile/data, how likely is it that they will vote in favor of the candidate?

2) How would a voter’s likelihood of support for the candidate change if they had personal contact with the candidate?

Classify each question as either predictive or inferential. Explain your reasoning for each.

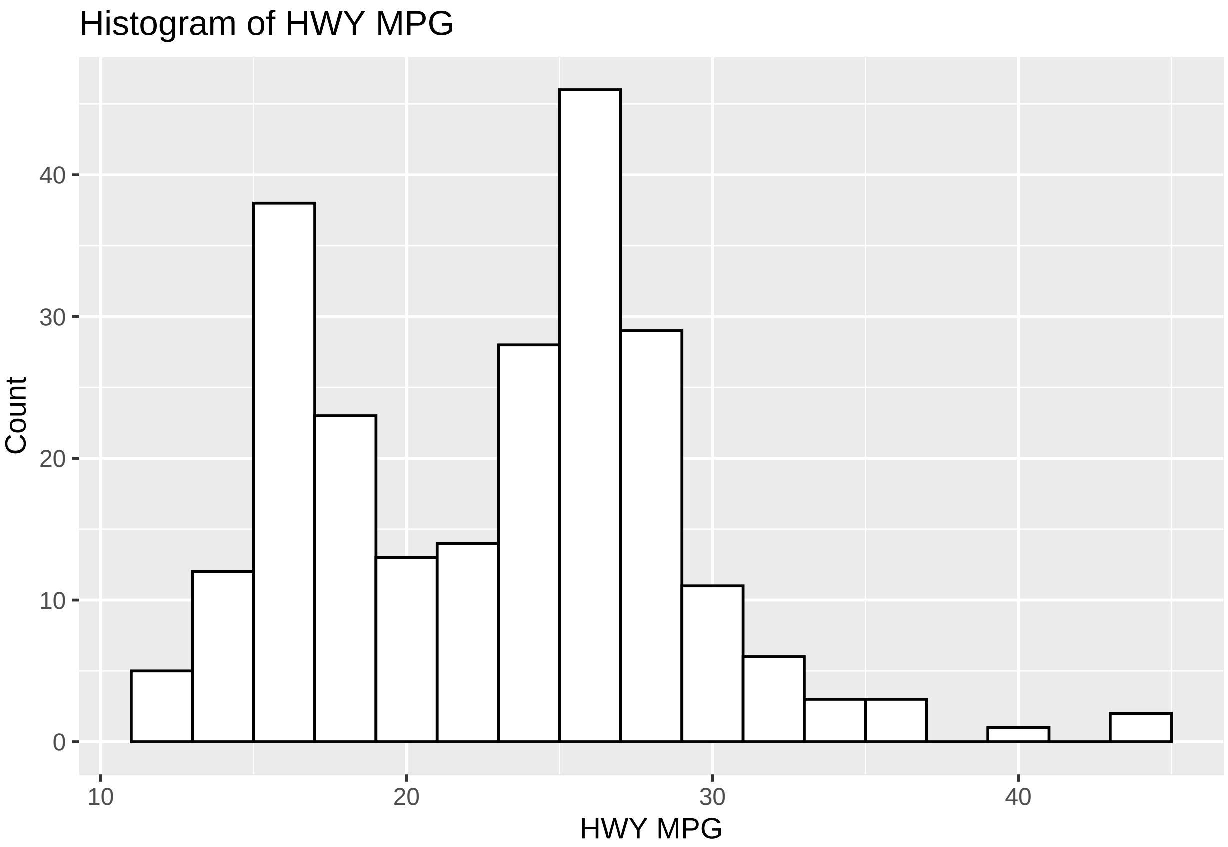
1) In this case, a predictive model is best. We are simply looking to predict the likelihood that they will vote for a candidate based on a variety of predictors. We don’t have any theories to test about what factors impact that probability more than others.
2) In this case, an inferential model is best. We are testing the theory that a voter’s support changes if they have had personal contact with the candidate. We’re not trying to predict a singular outcome like in the first question, instead we’re testing one variable against another.

Exploratory Data Analysis: Exercise 1:

```r
#loading in packages
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.7      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

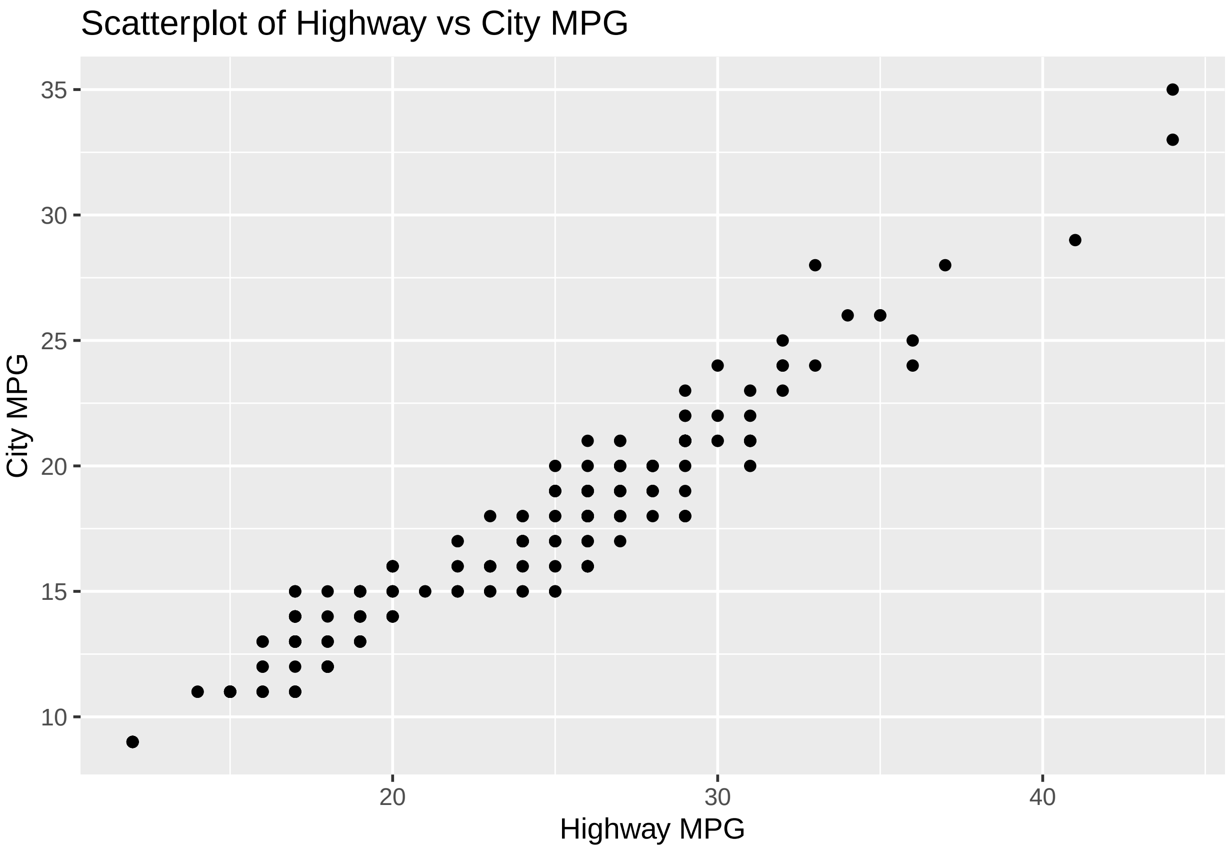
```r
ggplot(mpg, aes(x=hwy)) + geom_histogram(binwidth=2, color="black", fill="white") +
  labs(title='Histogram of HWY MPG', x='HWY MPG', y='Count')
```



I notice that there is a large peak at about 26 MPG and a smaller peak at about 16 MPG. From personal knowledge, this makes sense because 26 makes sense for most sedans and 16 makes sense for larger vehicles.There do appear to be sparse values over 40, which would be your more rarely seen fuel-efficient and hybrid vehicles.

Exercise 2:

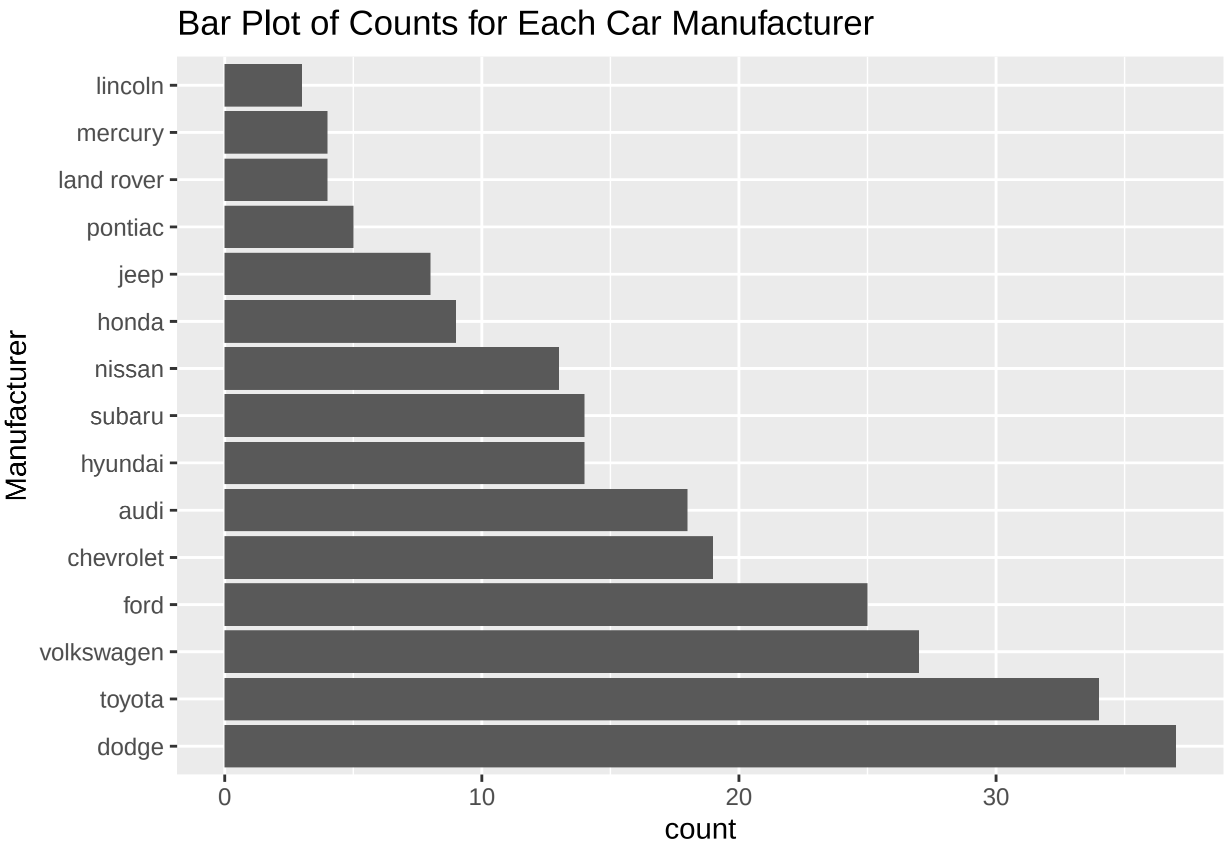
```
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point() +
  labs(title="Scatterplot of Highway vs City MPG", x="Highway MPG", y='City MPG')
```



There does seem to be a proportional relationship between hwy and cty. This means that as the highway MPG increases, the city MPG also increases. We can tell this relationship because there is a noticeable upward-pointing line from bottom left to top right.

Exercise 3:

```
ggplot(mpg, aes(x=reorder(manufacturer, manufacturer, function(x)-length(x)))) + geom_bar() +
  labs(title="Bar Plot of Counts for Each Car Manufacturer", x="Manufacturer", y="count") + coord_flip(
```

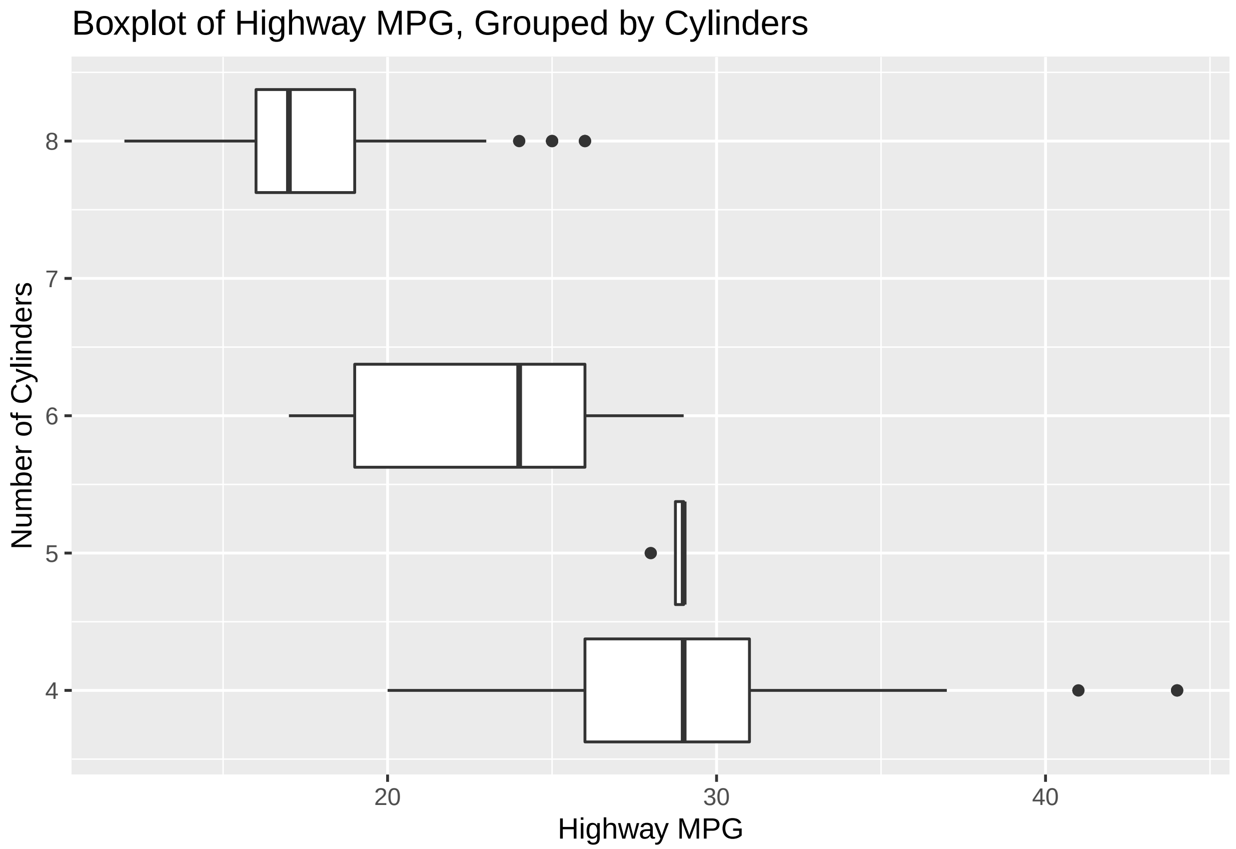


I used https://stackoverflow.com/questions/5208679/order-bars-in-ggplot2-bar-graph to help me with the ordering part of this problem, as I had never learned how to reorder a barplot in ggplot before now.

We can see that Dodge produced the most cars, and Lincoln produced the least cars.

Exercise 4:

```
ggplot(mpg, aes(x=hwy, group=cyl, y=cyl)) + geom_boxplot() +
  labs(title="Boxplot of Highway MPG, Grouped by Cylinders", x="Highway MPG", y="Number of Cylinders")
```



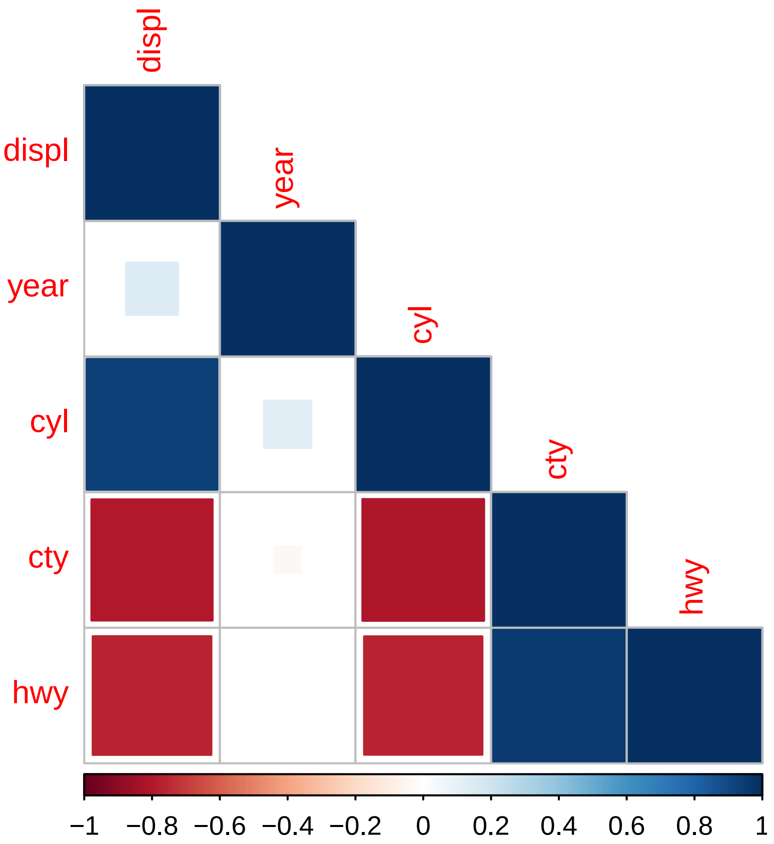
I do see a pattern, as it appears there is an inverse relationship between Highway MPG and number of Cylinders. As the number of cylinders decrease, the median HWY MPG tends to increase.

Exercise 5:

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mpgnumeric<-mpg[c('displ', 'year', 'cyl', 'cty', 'hwy')]
corrmpg<-corrplot(cor(mpgnumeric), method='square', type='lower')
```



```
corrmpg
```

```
## $corr
##             displ        year        cyl         cty          hwy
## displ  1.0000000  0.147842816  0.9302271 -0.79852397 -0.766020021
## year   0.1478428  1.000000000  0.1222453 -0.03723229  0.002157643
## cyl    0.9302271  0.122245347  1.0000000 -0.80577141 -0.761912354
## cty   -0.7985240 -0.037232291 -0.8057714  1.00000000  0.955915914
## hwy   -0.7660200  0.002157643 -0.7619124  0.95591591  1.000000000
##
## $corrPos
##    xName yName x y         corr
## 1  displ displ 1 5  1.000000000
## 2  displ  year 1 4  0.147842816
## 3  displ   cyl 1 3  0.930227102
## 4  displ   cty 1 2 -0.798523969
## 5  displ   hwy 1 1 -0.766020021
## 6   year  year 2 4  1.000000000
## 7   year   cyl 2 3  0.122245347
## 8   year   cty 2 2 -0.037232291
## 9   year   hwy 2 1  0.002157643
## 10   cyl   cyl 3 3  1.000000000
## 11   cyl   cty 3 2 -0.805771408
## 12   cyl   hwy 3 1 -0.761912354
## 13   cty   cty 4 2  1.000000000
## 14   cty   hwy 4 1  0.955915914
## 15   hwy   hwy 5 1  1.000000000
##
```

```
## $arg
## $arg$type
## [1] "lower"
```

The variables that appear to be significantly positively correlated with each other are: cylinders/displacement and city MPG/highway MPG.

The variables that appear to be significantly negatively correlated with each other are: city MPG/displacement, highway MPG/displacement, cylinders/city MPG, and cylinders/highway MPG.

These negative relationships make sense, as a higher number of cylinders and engine displacement typically indicate gas guzzling cars with worse MPG.

The cylinders and engine displacement being positively correlated makes sense too, as those are both indicators of bigger cars.

Finally, both MPGs being positively correlated makes sense. If a car is fuel efficient in one area, it's most likely fuel efficient in the other.

I'm very surprised that there's no correlation between year and city/highway MPG. You would think MPG would improve over the years, but that does not seem to be the case in this sample.