PSTAT 131 Homework 1

Logan Greenough

2022-09-28

Machine Learning Main Ideas

Question 1

Supervised learning is defined as learing in which you're given the input and the output, and the output acts as the "supervisor". Unsupervised learning is defined as learning in which you are given the input but not the output, and thus the learning lacks a "supervisor". The biggest difference between the two types of learning is in supervised learning you are able to see the output, while in unsupervised learning you are not able to see the output.(lecture)

Question 2

In the context of machine learning, the difference between a regression model and a classification model is that for the regression model the response to the predictors will be quantitative, also known as numerical values, while with the classification model the response to the predictors will be qualitative, also known as categorical values.(lecture)

Question 3

Two commonly used metrics for regression ML problems are price and blood pressure. Two commonly used metrics for classification ML problems are survived/died and spam/not spam.(lecture)

Question 4

A descriptive model is best used to visually emphasize a trend in the data, and an example of this could be a line on a scatterplot. A predictive model is most often utilized to predict a response variable, with minimum reducible error, given a set of input variables. An inferential model is best suited to test theories, potentially make causal claims, and to state relationships between outcomes and predictors.(lecture)

Question 5

Mechanistic models are models that assume a parametric form for f, which looks like $\beta_0 + \beta_1 + ... + \beta_n$. This model is good because you can add more parameters to make the model more flexible, but if you add to many parameters then the model will be overfit and will react in an unfavorable way to noise.

Empirically-driven models make no assumptions for f and tend to be more flexible by default, yet they fall short because they are require a larger number of observations and can also fall victim to overfitting.

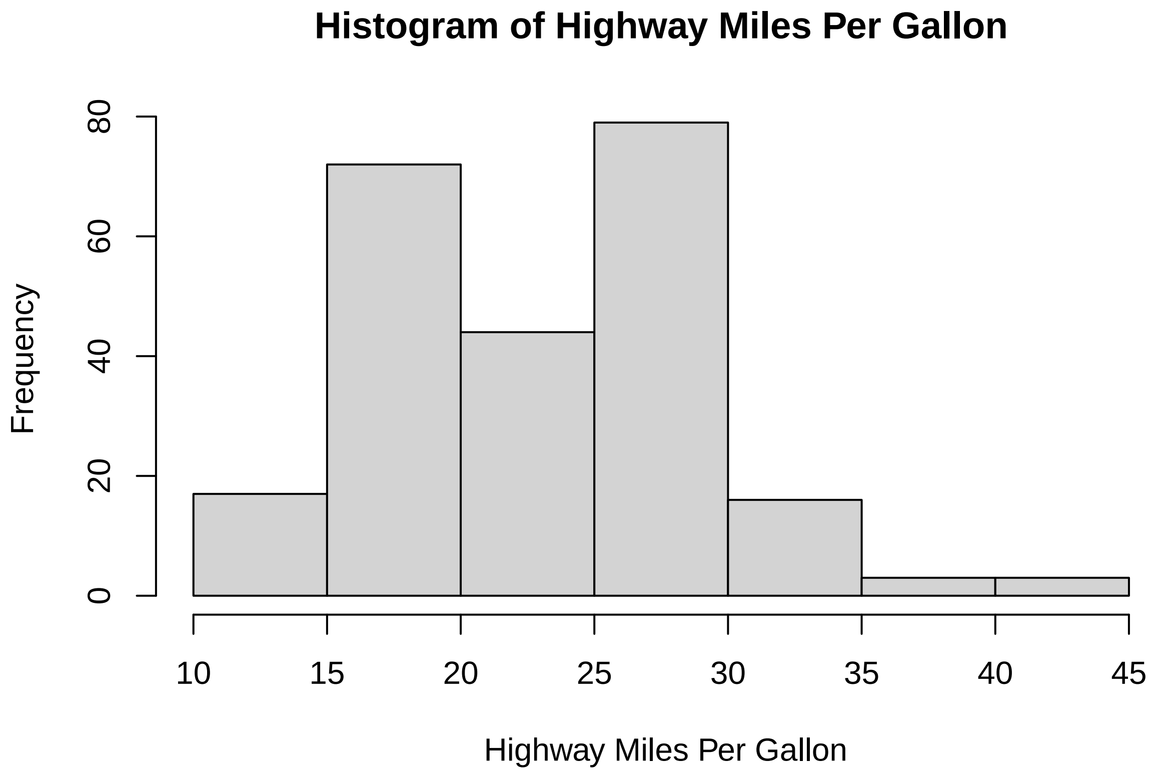
Therefore, these models are different in the sense that in mechanistic models you are making assumptions, while in empirically-driven you are not making assumptions. These models are similar in the sense that they have the same shortcomings and can fall victim to overfitting. In general, I would say that a mechanistic model is easier to understand. This is due to the fact that you are able to control the amount of parameters in the model, as well as understanding the effect that each parameter has on the model.

The bias-variance tradeoff principle states that if you have a model with high bias, then the variance will be low, and if the bias is low than the variance will be high. In relating this to the mechanistic model, since you are able to control the parameters(and make assumptions), these models have the potential to have more bias, and thus a lower variance. In relating this to the empirically driven models, no assumptions are made and thus the bias will tend to be lower with a higher variance.

Exploratory Analysis

Exercise 1

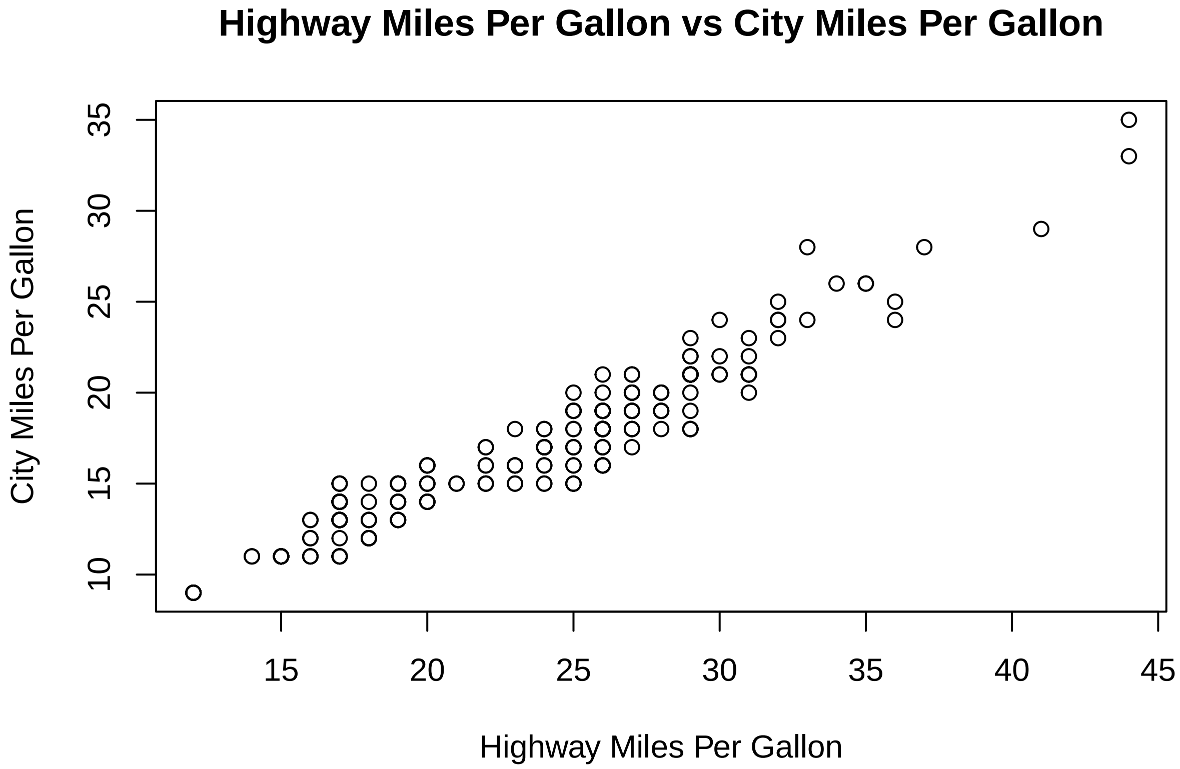
```r
hist(mpg$hwy, main = "Histogram of Highway Miles Per Gallon", xlab =" Highway Miles Per Gallon")
```



This histogram of the Highway variable from miles per gallon has a slight right skew to it. I think that if you were to change the way that this variable is grouped, you could probably make it look as though it follows a normal distribution.

Exercise 2

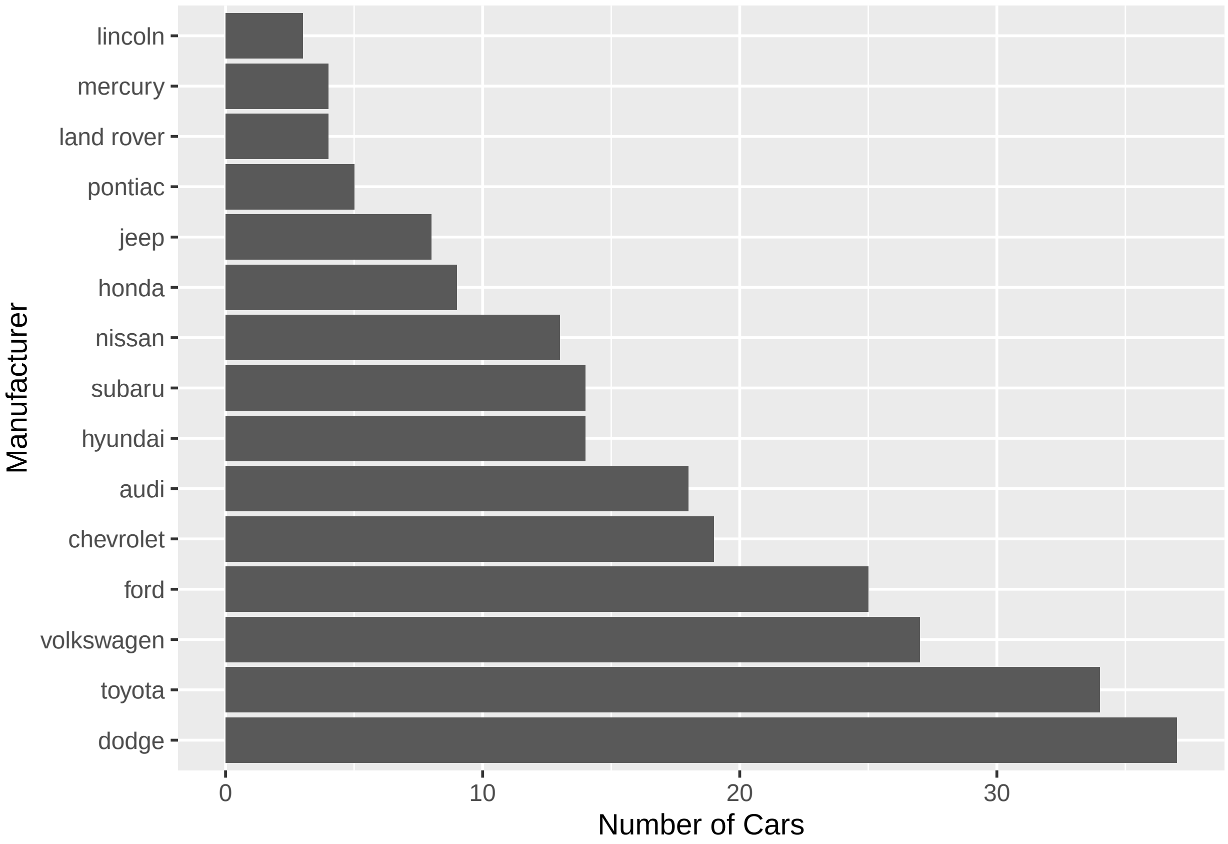
```
plot(mpg$hwy, mpg$cty, xlab = "Highway Miles Per Gallon", ylab = "City Miles Per Gallon", main = "Highwa
```



Based on this scatter plot, we can see that there is a clear positive trend line between the two. Therefore, the assumption can be made that as the highway miles per gallon increase then the city miles per gallon will also increase. Based on the scatter plot above, I would definitively say that there is a relationship between the highway miles per gallon and the city miles per gallon.

Exercise 3

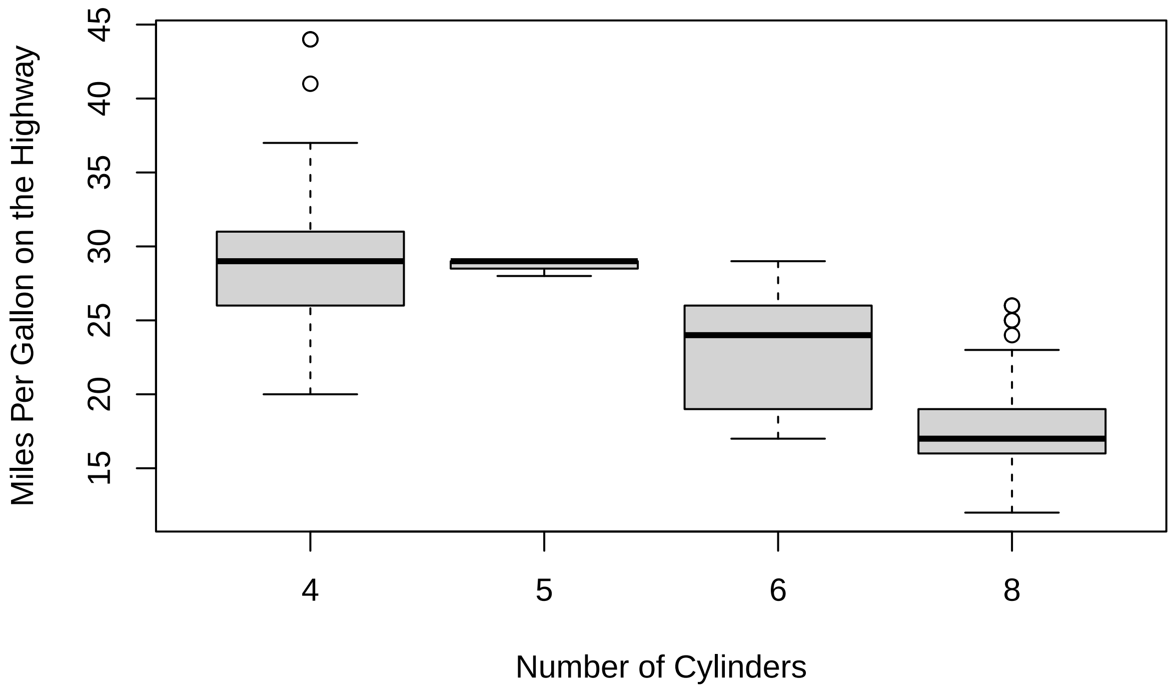
```
ggplot(data = mpg, aes(y = reorder(manufacturer, manufacturer, function(y)-length(y)))) + geom_bar(stat
```



To help with sorting the bar graph, I used statology.org. Based on this graph we can see that the manufacturer that made the most cars was Dodge, and the manufacturer that made the least was Lincoln.

Exercise 4

```
boxplot(formula=hwy ~ cyl, data=mpg, xlab= "Number of Cylinders", ylab = "Miles Per Gallon on the Highwa
```



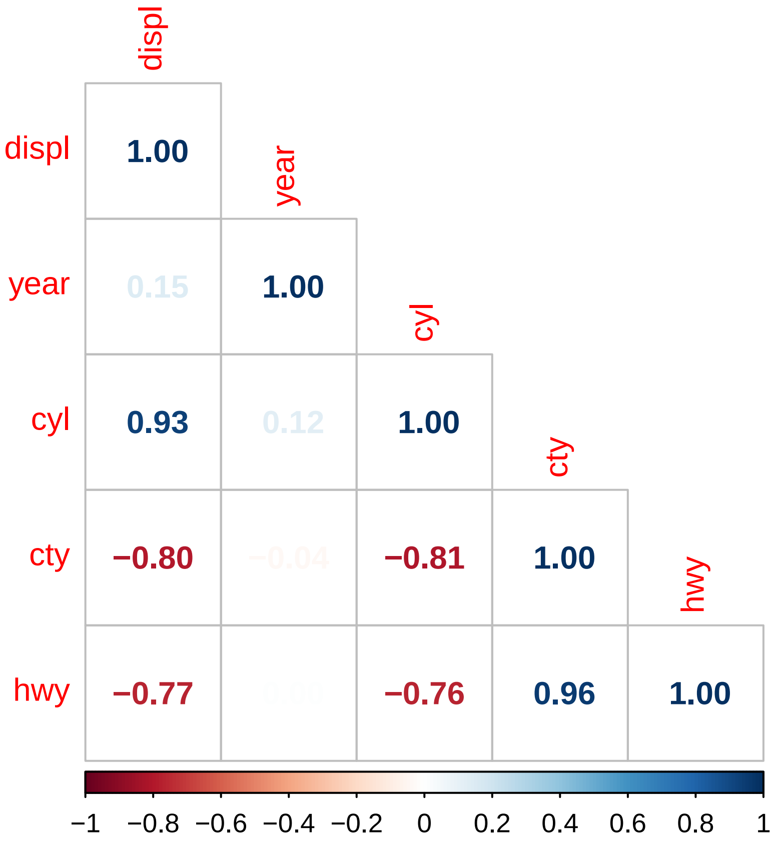
Based on this box plot, we are able to see that there is somewhat of a trend. This trend is that the more cylinders that a car has, then the less miles per gallon it will get on the highway.

Exercise 5

```
names(mpg)
```

```
##  [1] "manufacturer" "model"        "displ"        "year"         "cyl"
##  [6] "trans"        "drv"          "cty"          "hwy"          "fl"
## [11] "class"
```

```
fixed_mpg <- mpg[,c(-1,-2,-6,-7,-10,-11)]
M = cor(fixed_mpg)
corrplot(M, method = "number", type = "lower")
```



The variables that are positively correlated are number of cylinders and engine displacement, and city miles per gallon and highway miles per gallon. The variables that are negatively correlated are city miles per gallon and engine displacement, highway miles per gallon and engine displacement, city miles per gallon and number of gallons cylinders, and highway miles per gallon and number of cylinders. There is no strong correlation positive or negative for year and engine displacement, year and the number of cylinders, year and city miles per gallon, and year and highway miles per gallon.

These relationships all make sense to me. This is due to the fact that if an engine has more cylinders it will probably be less fuel efficient than a car that has fewer cylinders. Additionally, if a car has more cylinders, than it will displace more fuel as it runs, and vice versa. One thing that does surprise me is the fact that the year of the car doesn't have any relationships. This surprises me since newer cars tend to be more fuel efficient and get more miles per gallon. This could be attributed to when the data was collected, as this increase in fuel efficiency is a relatively new thing.