PSTAT 131 HW 1

Raymond Lee

2022-04-01

Machine Learning Main Ideas

1. In supervised learning, we know the output observed by the model. In unsupervised learning, we do not know the output observed by the model. Supervised learning is used to predict an output based on inputs. Unsupervised learning does not include an output, but we can still learn relationships and structure from the inputs.

2. Regression models involve quantitative responses while classification models involve qualitative responses.

3. For regression, we can use mean squared error and R squared. For classification, we can use Type 1 error and Type 2 error.

4. Descriptive models try to visually emphasize a trend in the data (lecture 2). Inferential models try to find which data are significant, test theories, possibly state causal claims, and describe the relationship between the outcome and predictors (lecture 2). Predictive models try to find which data best fits the model and predict the response with minimum reducible error (lecture 2).

5. Mechanistic means that a form is assumed for *f*, the relationship between the response and predictors. The data is then fit to the model. However, the model we choose may not match the true unknown form of *f*. We can try to address this by choosing flexible models that can fit many different possible forms for *f*. However, this requires estimating more parameters, which may lead to overfitting (pg. 21-22).

   Empirically-driven means that no assumptions about f are made, and an estimate of f that gets close to the data points as possible is sought. It can potentially accurately fit more possible forms of *f*. However, a larger number of observations is needed to accurately estimate f since f is estimated to a small number of parameters (pg. 23-24) Overfitting may also occur.

   A mechanistic model is easier to understand. Assuming a form for f allows us to estimate a set of parameters instead of fitting an arbitrary f (pg. 22).

   Bias is the error introduced by estimating a problem with a simpler model. Variance is the amount by which $\hat{f}$ would change if we obtained it with a different data set. As we use more flexible methods, the variance will increase, and the bias will decrease. However, at some point, increasing flexibility has little impact on bias but significantly increases the variance. We need to find a method/model for which the variance and squared bias are both low (pg. 34-36).

6. The first question is predictive because because we are trying to predict an outcome based on the data. The second question is inferential because describing the change involves the relationship between the outcome and predictors.

Exploratory Data Analysis

```
library(tidyverse)
```

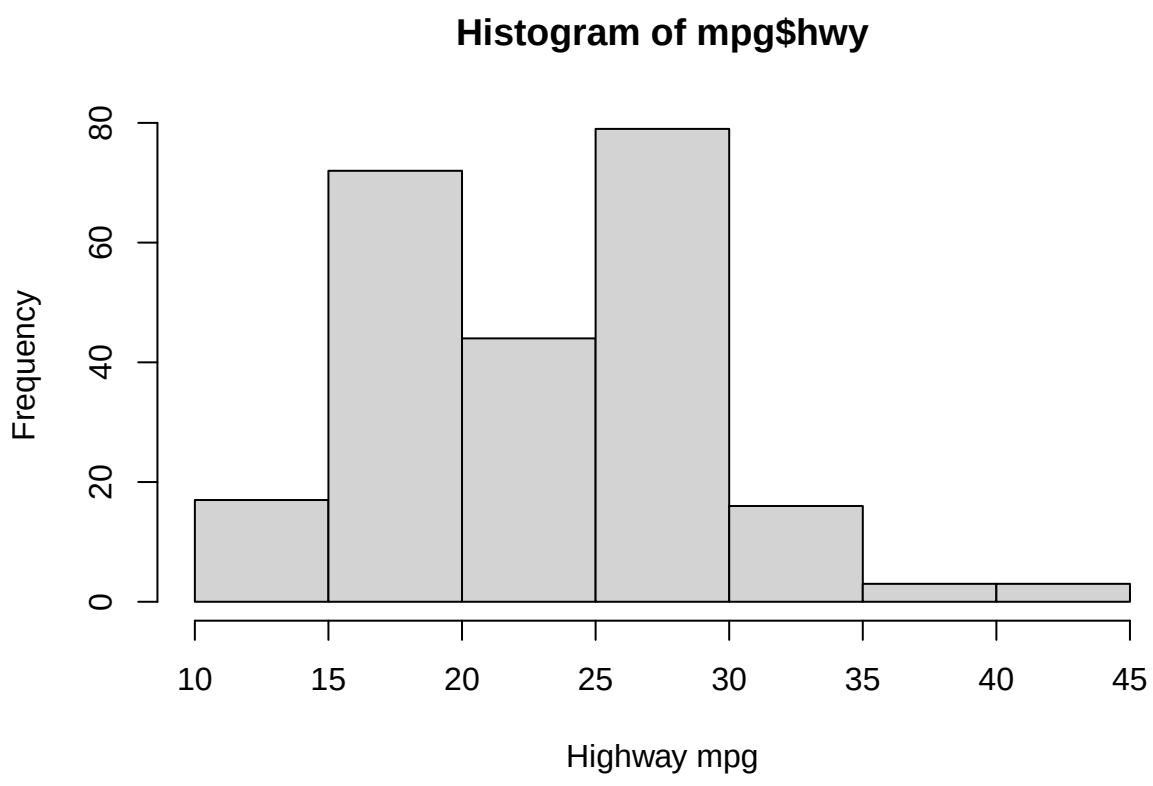
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## 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()
```

1.

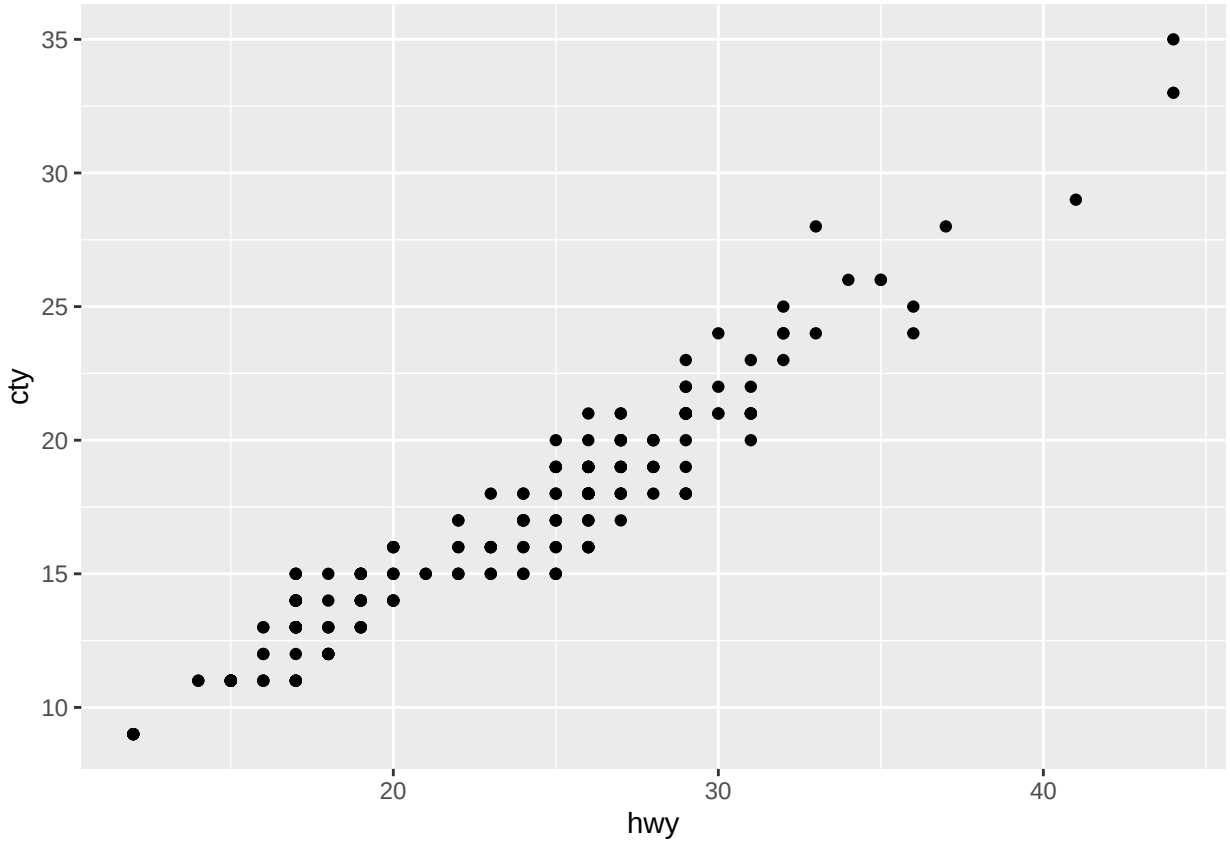
```
hist(mpg$hwy, xlab="Highway mpg")
```



The histogram seems to be skewed right. 25-30 highway mpg has the highest frequency.

2.

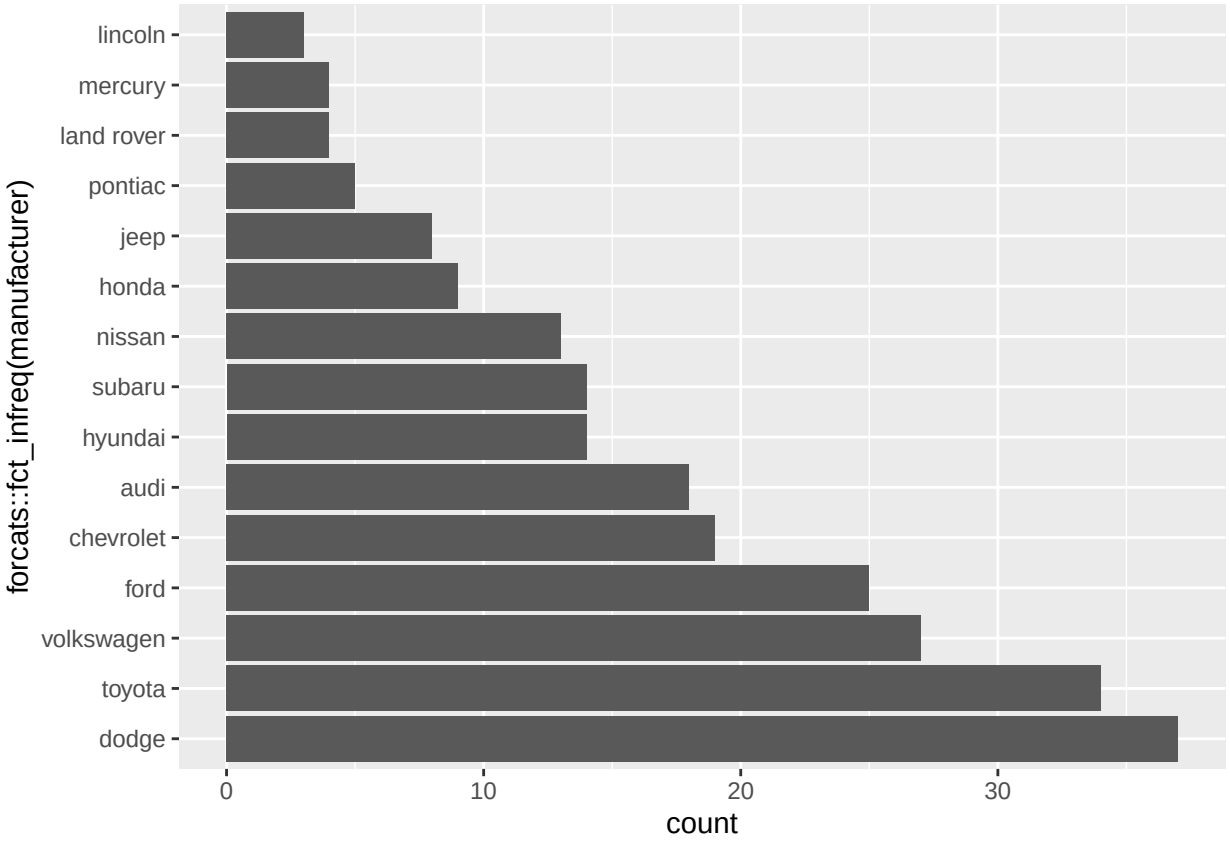
```
ggplot(mpg, aes(hwy, cty)) + geom_point()
```



Most points exist below hwy=35 and cty=30. hwy and cty seem to have a fairly strong, positive linear relationship. This means that as one variable increases, the other variable also increases.

3.

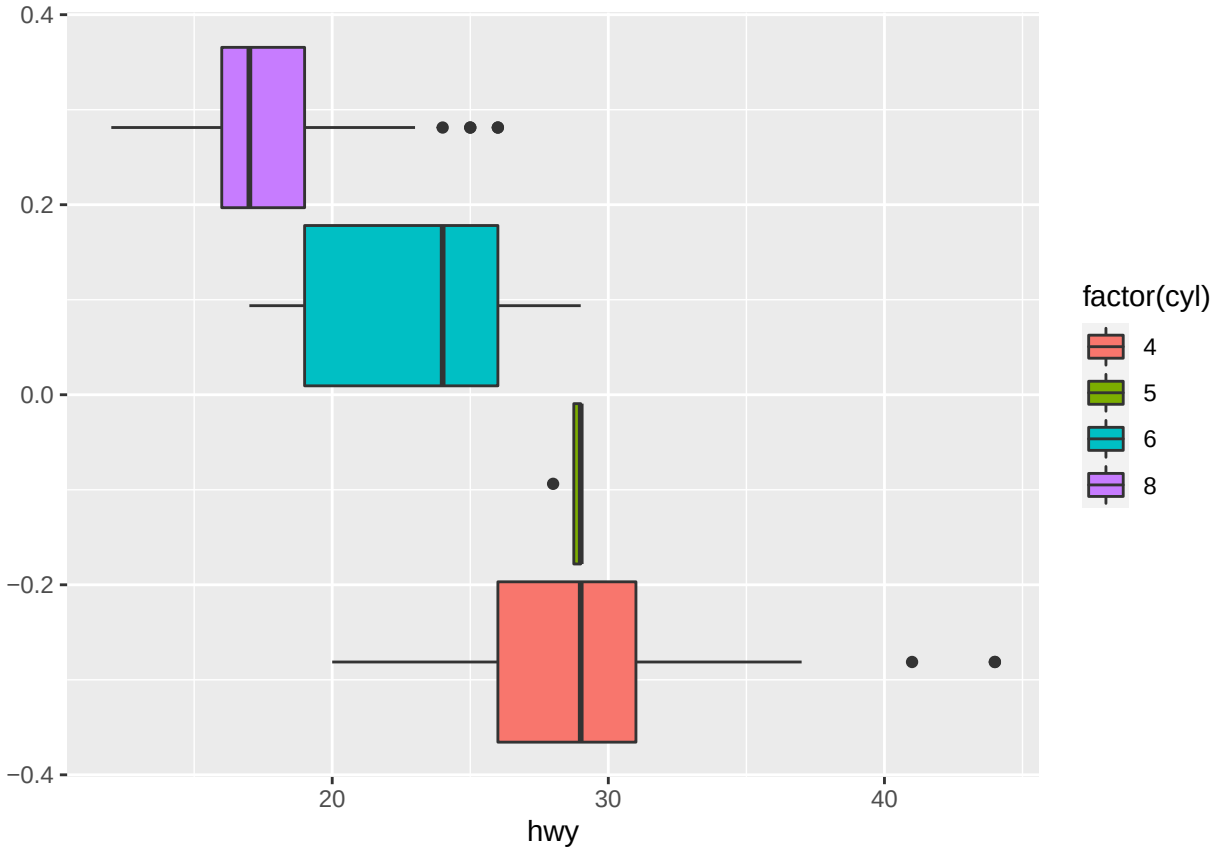
```
ggplot(mpg, aes(y=forcats::fct_infreq(manufacturer))) + geom_bar()
```



Dodge produced the most cars, and Lincoln produced the least cars.

4.

```
ggplot(mpg, aes(hwy, fill=factor(cyl))) + geom_boxplot()
```



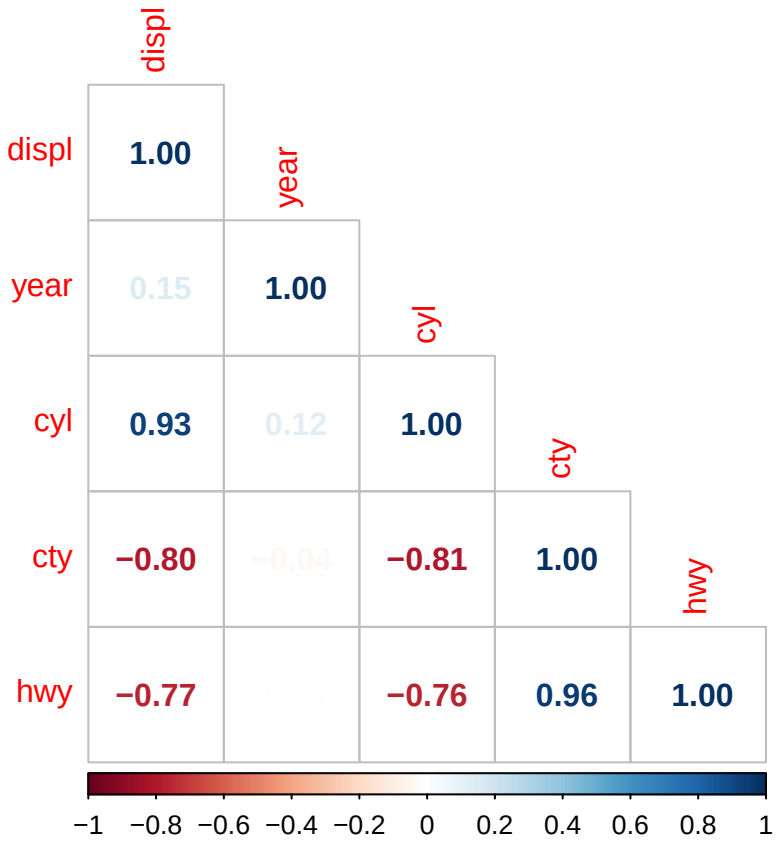
As the number of cylinders increases, there is a greater presence of higher hwy values.

5.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
df = subset(mpg, select=-c(1, 2, 6, 7, 10, 11))
cor_df = cor(df)
corrplot(cor_df, method="number", type="lower")
```



A lower triangle correlation matrix of the numerical variables was made. The categorical variables were omitted because we cannot calculate the correlation between categorical variables. Displ is positively correlated with year and cyl but negatively correlated with cty and hwy. Year is positively correlated with cyl. Cyl is negatively correlated with cty and hwy. Cty is positively correlated with hwy. Most of the relationships make sense, but it is interesting to me that year is positively correlated with cyl. That is, as the model year increases, the more cylinders there are. It is also interesting that displ is positively correlated with year. More recent models tend to have more engine displacement.