PSTAT131, Homework 1

Sofia Spasibenko

Machine Learning Main Ideas

Question 1:

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

Supervised learning is a statistical model that looks at predictors to predict future results. Unsupervised learning does not look at the response, and instead focuses on categorizing data. The key difference is whether or not there is a known output to the data (there is for supervised and there's not for unsupervised).

Question 2:

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

A regression model is one in which the result of the analysis is quantitative (e.g. a test score out of 100). This is not to be confused with a linear regression which may or may not be the best option for a regression model, but it can still be used to solve a regression problem. A classification model is one in which the results are qualitative (e.g. survived or died on the Titanic). Either model can be used for supervised learning.

Question 3:

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

For regression ML problems, mean squared error (MSE) and Root Mean Squared Error (RMSE) are two common metrics. For classification ML problems, accuracy and confusion matrix are two common metrics.

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:

Models that can help best visualize a trend in data.

Inferential models:

Models made to test theories, discover which feature are significant, and explore the relationships between predictors and the outcome.

Predictive models:

Models that predict the outcome with the minimum reducible error and which combination of features fit the model best.

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?

A mechanistic model type assumes a parametric form for f with certain standards the function will follow. An empirically-driven model type does not assume anything about f. Compared to the mechanistic model, the empirically-driven model requires more observations and is more flexible. However, both models can be affected by over-fitting.

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

Mechanistic models are often easier to understand because they have a higher interpretability which it trades off for less flexibility. There's a set way of analysis unlike the empirically-driven models where there's no standard way of approaching the problem.

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

Because mechanistic models are easier to interpret but less flexible, they contrast empirically-driven models which are more flexible, but harder to interpret. Thus, they favor different sides of the bias-variance tradeoff since mechanistic models err on the side of the variance and empirically-driven models err on the side of the bias.

Question 6:

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

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

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.

The first question is predictive because it is dealing with the outcome of the situation and how the predictor in the data profile affect the outcome. The second question is inferential because it's dealing with exploring a causal claim and if a feature is significant.

Exploratory Data Analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
mpg
```

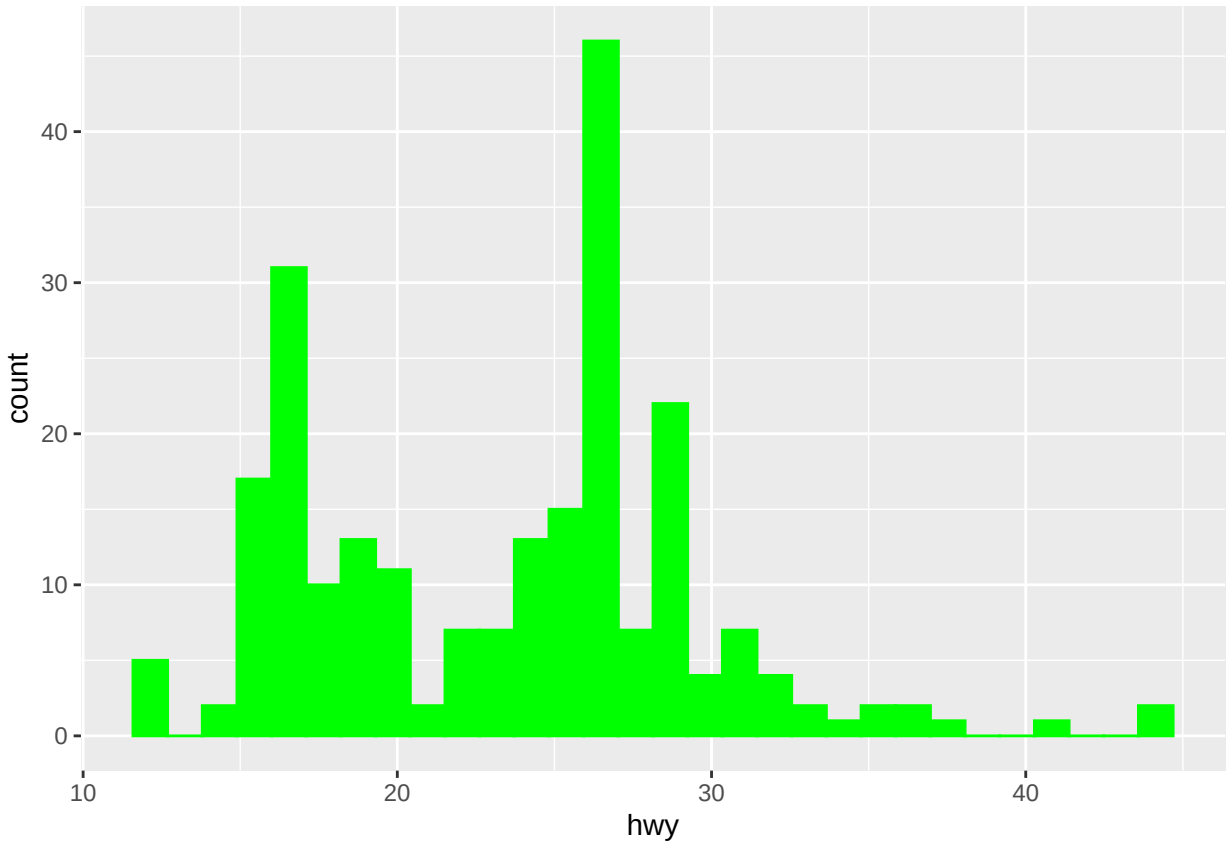
```
## # A tibble: 234 x 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4           1.8  1999     4 auto~ f        18    29 p     comp~
##  2 audi         a4           1.8  1999     4 manu~ f        21    29 p     comp~
##  3 audi         a4           2    2008     4 manu~ f        20    31 p     comp~
##  4 audi         a4           2    2008     4 auto~ f        21    30 p     comp~
##  5 audi         a4           2.8  1999     6 auto~ f        16    26 p     comp~
##  6 audi         a4           2.8  1999     6 manu~ f        18    26 p     comp~
##  7 audi         a4           3.1  2008     6 auto~ f        18    27 p     comp~
##  8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p     comp~
##  9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p     comp~
## 10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p     comp~
## # ... with 224 more rows
```

Exercise 1:

We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.

```
ggplot(mpg, aes(x=hwy)) + geom_histogram(color="green", fill="green")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

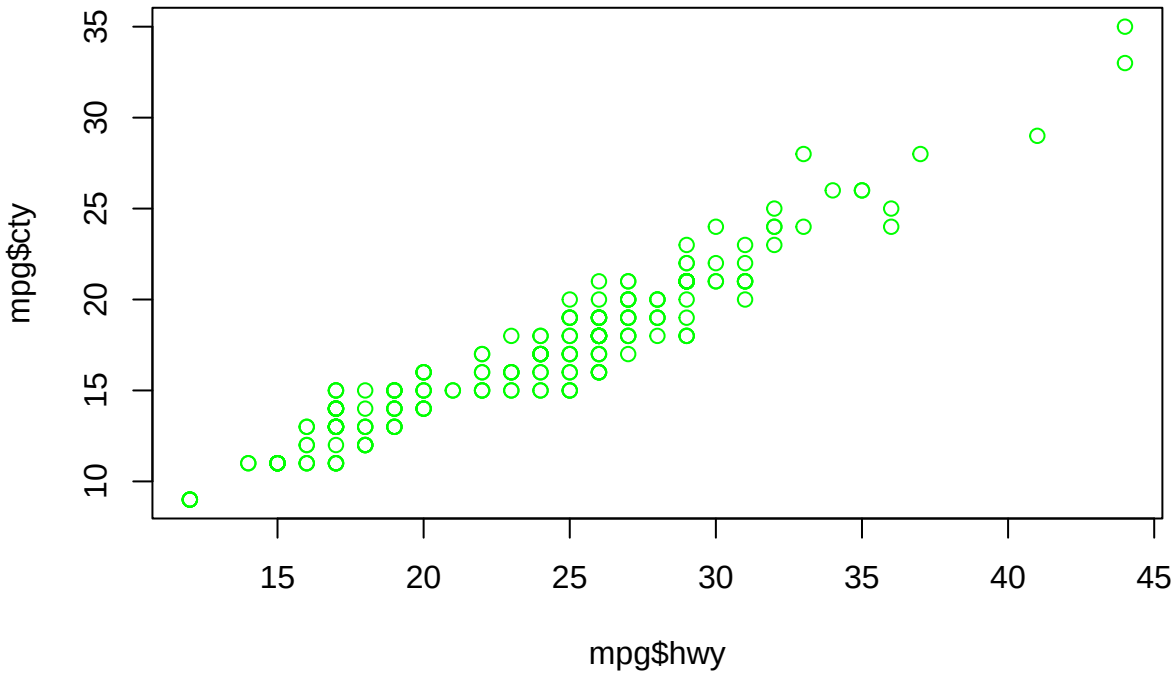


I see that the hwy variable seems to follow two normal curves side by side, with the second one being the more prominent one. It makes sense that there would be fewer cars with lower/higher highway mpg and would be mainly in the middle.

Exercise 2:

Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?

```
plot(mpg$hwy, mpg$cty, col="green")
```

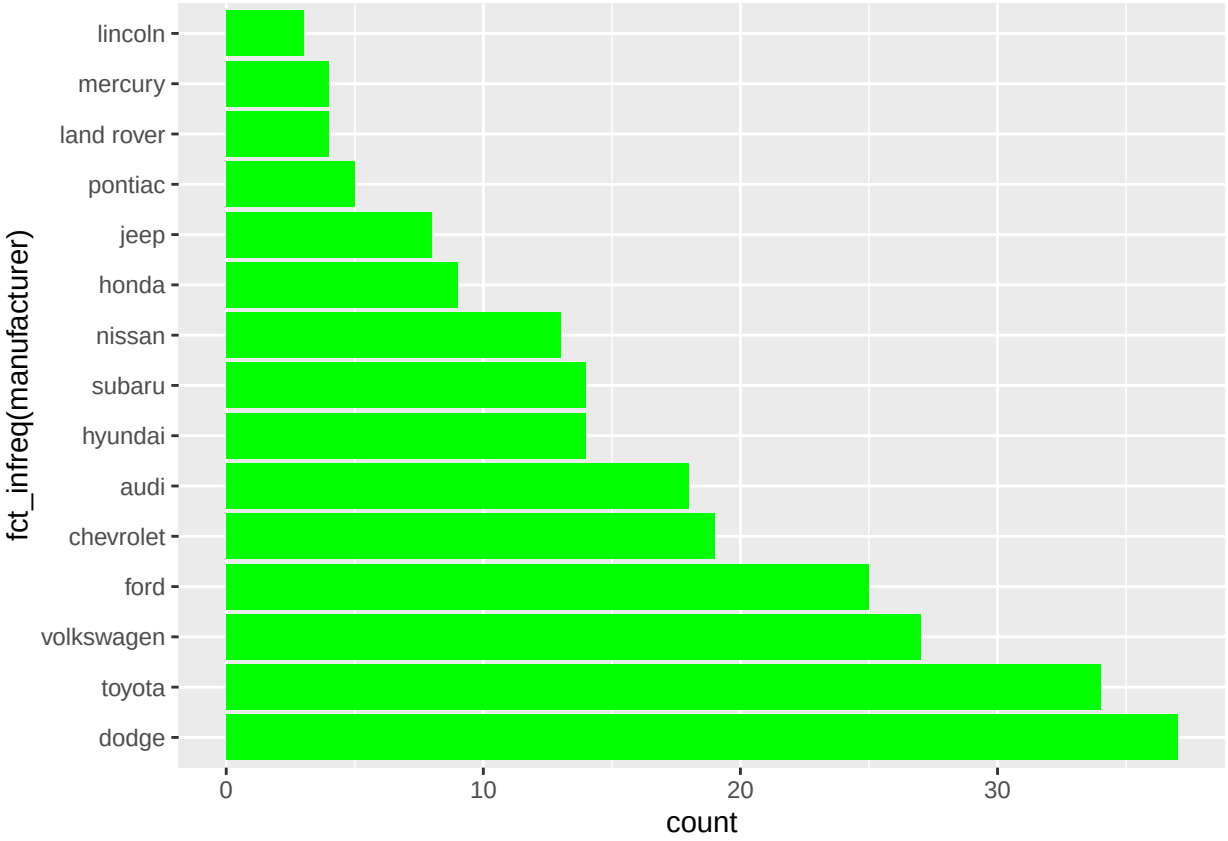


There seems to be a fairly strong positive linear correlation between cty and hwy which implies that if one variable increases, the other variable also increases. In general, if your highway mileage is high, your city mileage would also be high.

Exercise 3:

Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least?

```
library(forcats)
ggplot(mpg, aes(x=fct_infreq(manufacturer))) + geom_bar(fill = 'green') + coord_flip()
```

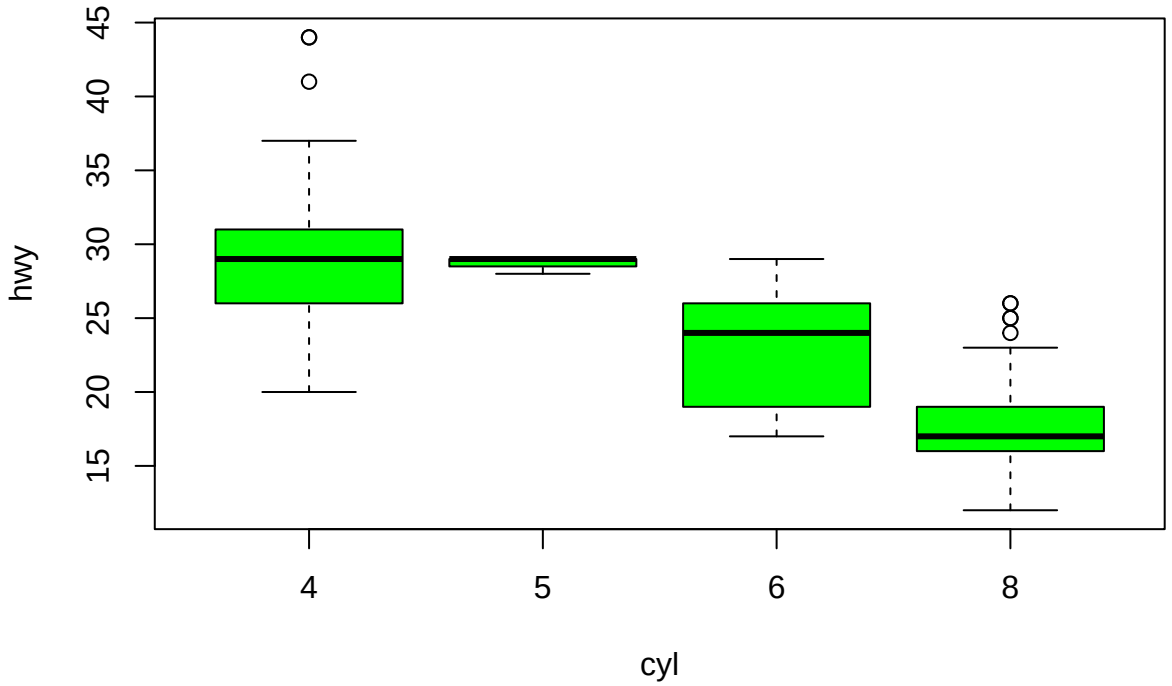


Dodge produced the most cars and Lincoln manufactured the least.

Exercise 4:

Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

```
boxplot(hwy~cyl, data=mpg, col='green')
```



It seems that the more cylinders a car has, the lower the highway miles per gallon are. However, it is also possible that they are distributed on a normal curve.

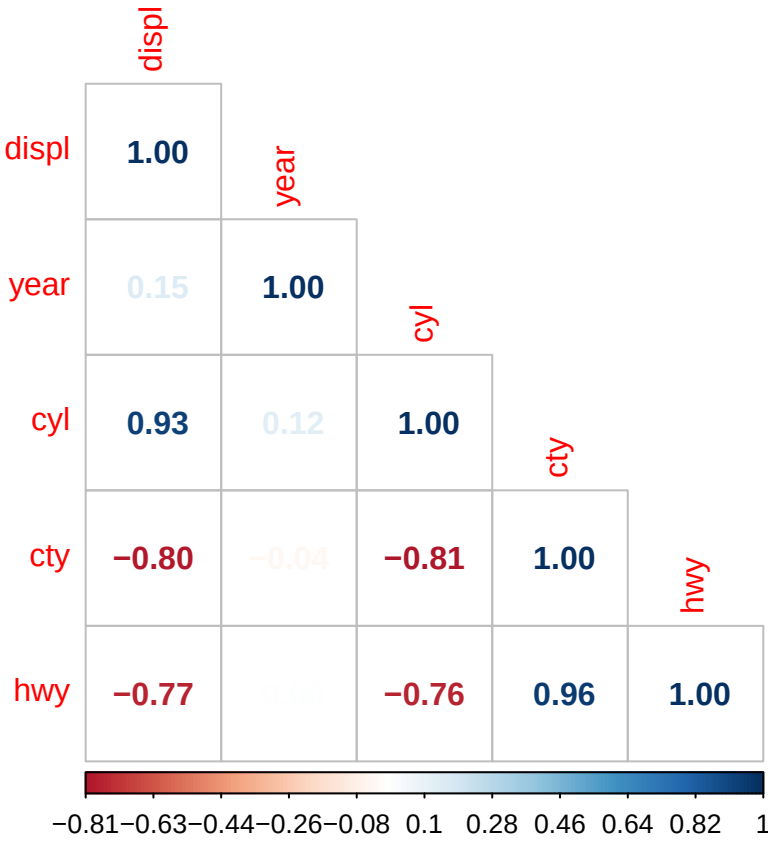
Exercise 5:

Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset. (Hint: You can find information on the package here.)

```
library('corrplot')
```

```
## corrplot 0.92 loaded
```

```
numMPG <- mpg
numMPG <- numMPG[, !sapply(numMPG, is.character)]
numMPG <- cor(numMPG)
corrplot(numMPG, is.corr = FALSE, type = "lower", method = 'number')
```



Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

Hwy is negatively correlated to displ and cyl, positively correlated to cty, and has no relation to year. Cty is positively correlated to displ and cyl, and slightly positively correlated to year. Cyl is negatively correlated to displ and slightly negatively correlated to year. Finally, year has a slight negative correlation to year. I was surprised to find out that year didn't seem to have a strong effect on the efficiency of city/highway mileage, but it makes sense that it wouldn't effect cylinders and consequently the engine displacement. The rest of the plot made sense: especially the city and highway mileage having such strong positive correlations with each other. Their negative correlations with the cylinders and engine displacement also made sense since more gas is needed to account for them.