PSTAT131 Homework 1

Shivani Kharva

2022-09-28

Homework 1

Machine Learning Main Ideas

Question 1

From the textbook page 26, supervised learning is when, for each observation of the predictor measurements x_i, i = 1,…,n, there is an associated response measurement y_i. In this case, the response is the supervisor. Models fitted to supervised learning tend to relate the response to the predictors. However, according to the textbook page 26, unsupervised learning is when, for each observation i,= 1,…,n, we observe a vector of measurements x_i, but no associated response y_i. We do not have a response variable that can supervise our analysis.

From Professor Coburn's office hours, supervised and unsupervised learning are mainly different due to the existence of a response variable (the supervisor) for supervised learning and the nonexistence of a response variable (no supervisor) for unsupervised learning.

Question 2

From the textbook page 28, regression models are those based on problems with a quantitative/numerical response whereas classification models are those based on problems involving a qualitative/categorical response (one with K different categories). Regression models are used for continuous outcomes, classification models are used for discrete outcomes. An example of a regression model is a linear regression model, and an example of a classification model is a logistic regression model.

According to textbook page 29, these two models are both used in machine learning and often, the type of model to be used may be decided based on whether the response is qualitative (classification) or quantitative (regression). However, as mentioned on textbook page 29, whether the predictors are qualitative or quantitative is less important because many statistical methods can be applied regardless of the predictor variable type with proper coding for qualitative predictors in advance.

Question 3

According to the textbook page 29 and Professor Coburn's office hours, two metrics for regression are the Mean Squared Error (MSE =

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}\,(x_i))^2$$

and the Root Mean Squared Error (RMSE = $\sqrt{MSE}$). According to Opengenus.org (https://iq.opengenus.org/performance-metrics-in-classification-regression/), two metrics for classification are accuracy (ratio of the number of correct predictions and the total number of predictions) and the F1-score (F1-

score = $2 * \frac{precision * recall}{precision + recall}$).

Question 4

Descriptive models: According to Lecture: Course Overview and Introduction slide 39, descriptive models visually illustrate trends in data. For example, using a line on a scatterplot is a descriptive model.

Inferential models: From Lecture: Course Overview and Introduction slide 39, inferential models aim to test theories. We are asking what features of the data are significant and attempting to learn whether there are any causal claims we can make. Inferential models aim to state a relationship between the outcome and the predictor(s).

Predictive models: Based on Lecture: Course Overview and Introduction slide 39, predictive models aim to predict Y (the response) with minimum reducible error. Predictive models aim to determine what combination of features (predictors) fits the best. These models are not focusing on hypothesis tests.

Question 5

Mechanistic (aka parametric) predictive models assume that there are some parameters involved. According to Lecture: Course Overview and Introduction slide 38, mechanistic predictive models assume a parametric form for f, such as assuming f is a linear function. However, this means that this type of predictive model almost never actually matched the true unknown f. According to Lecture: Course Overview and Introduction slide 38, empirically-driven (non-parametric) predictive models do not make any assumptions about f and do not assume any parameters. Because of this, these type of predictive models require a larger number of observations.

Based on Lecture: Course Overview and Introduction slide 38, The two model types are different because the mechanistic type assumes parameters and a parametric form of f whereas the empirically-driven type does not make any assumptions about parameters or f. Also, mechanistic model types require fewer observations than empirically-driven model types. Because of this difference, mechanistic model types require more parameters to be more flexible, while empirically-driven model types are much more flexible by default. However, according to Lecture: Course Overview and Introduction slide 38, the two model types can be similar in that these two model types may both lead to overfitting. In a mechanistic type model, if there are too many parameters, the model may become so flexible that it also fits random noise, which is overfitting. In an empirically-drive type model, since the model type is already much more flexible by default, it is more susceptible to overfitting by default as well.

From Professor Coburn’s office hours, mechanistic models tend to be easier to explain and understand because they tend to be simpler than empirically-driven models. For example, explaining a linear regression (mechanistic) would be much easier than explaining a decision tree with several nodes (empirically-driven). Thus, it may be harder to understand emprically-driven models simply because they are also harder to communicate and explain due to the complexity they can reach.

The bias-variance tradeoff is related to the use of mechanistic and empirically driven models in that the model type that you use may affect the bias-variance tradeoff. If a more empirically-driven model is used, such as drawing a curve that passes through every training observation, the method would have very low bias but high variance according to textbook page 36. If fitted to observations outside of the training set, the empirically-driven model would not be as accurate (because of overfitting). However, if a more mechanistic model is used, such as fitting a horizontal line to the data, the method might have very low variance but high bias according to textbook page 36. Using a more mechanistic model would allow for the training observations to be less fit to the chosen model, but using observations outside of the training set would not make the mechanistic model decrease in accuracy as with the empirically-driven model.

Question 6

“Given a voter’s profile/data, how likely is it that they will vote in favor of the candidate?”
This question is predictive because the campaign is attempting to predict the likelihood of the voter voting in favor of the candidate based on all the features (predictors) presented in their profile/data. The aim is to predict Y, the likelihood that the voter will vote in favor of the candidate, with minimum reducible error (which is what a predictive model is from Lecture: Course Overview and Introduction slide 39).

“How would a voter’s likelihood of support for the candidate change if they had personal contact with the candidate?”
This question is inferential because the campaign is attempting to see how changing a specific feature related to the voter would affect their likelihood to support the candidate. Rather than using all features to best predict the voter’s likelihood of supporting the candidate, the campaign is questioning how manipulating this one feature may affect the outcome. The aim is to test the relationship between the outcome (likelihood to vote in support of the voter) and a predictor (whether they had personal contact with the candidate) to see if any possible causal claims can be made (which is what an inferential model is from Lecture: Course Overview and Introduction slide 39).

Exploratory Data Analysis

Loading the tidyverse and other packages:

```
library(tidyverse)
```

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

```
library(ISLR)
library(ggplot2)
library(corrplot)
```

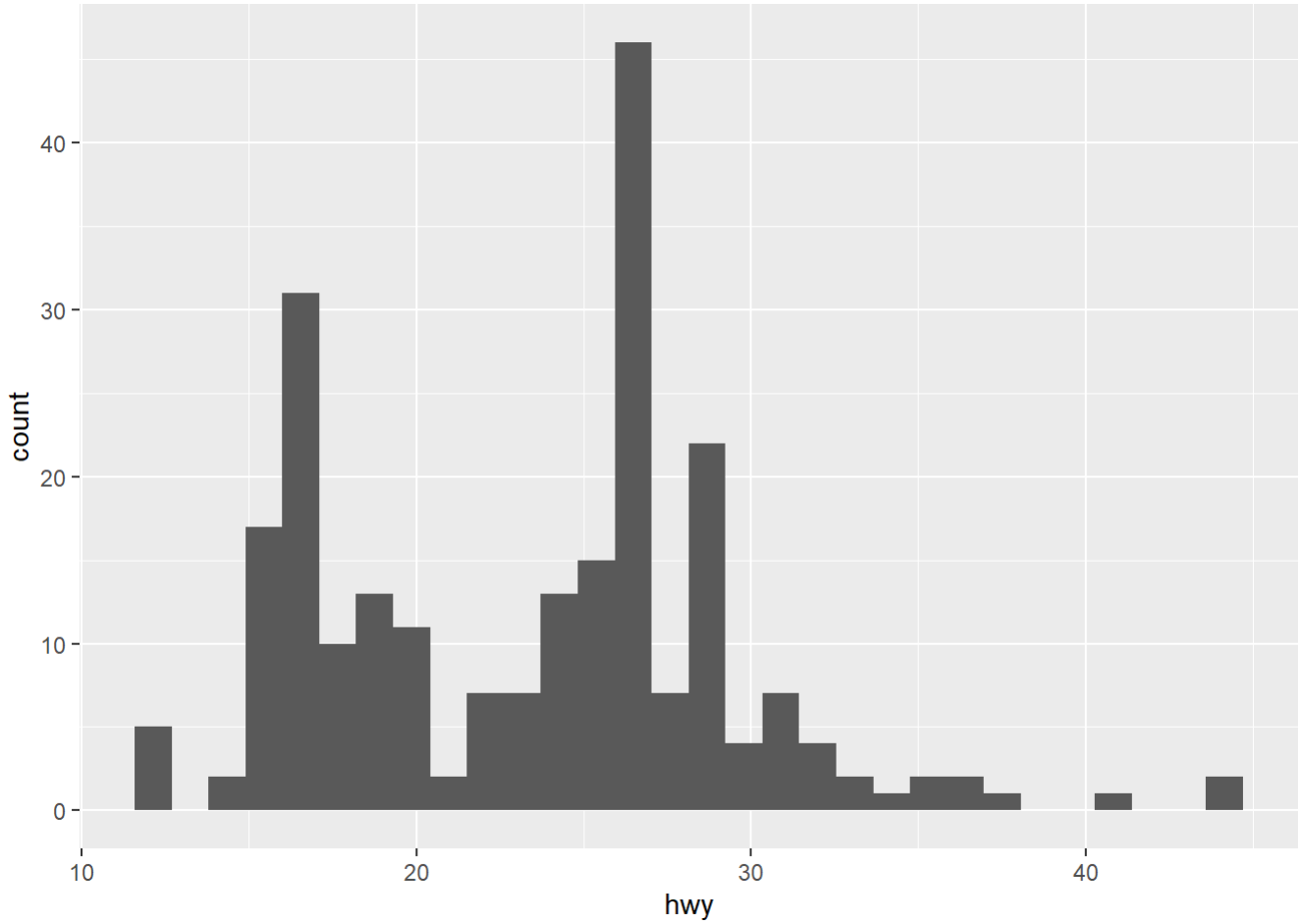
```
## corrplot 0.92 loaded
```

```
data(mpg)
```

Exercise 1

```
hwy_hist <- ggplot(mpg, aes(hwy)) +
  geom_histogram()

hwy_hist
```

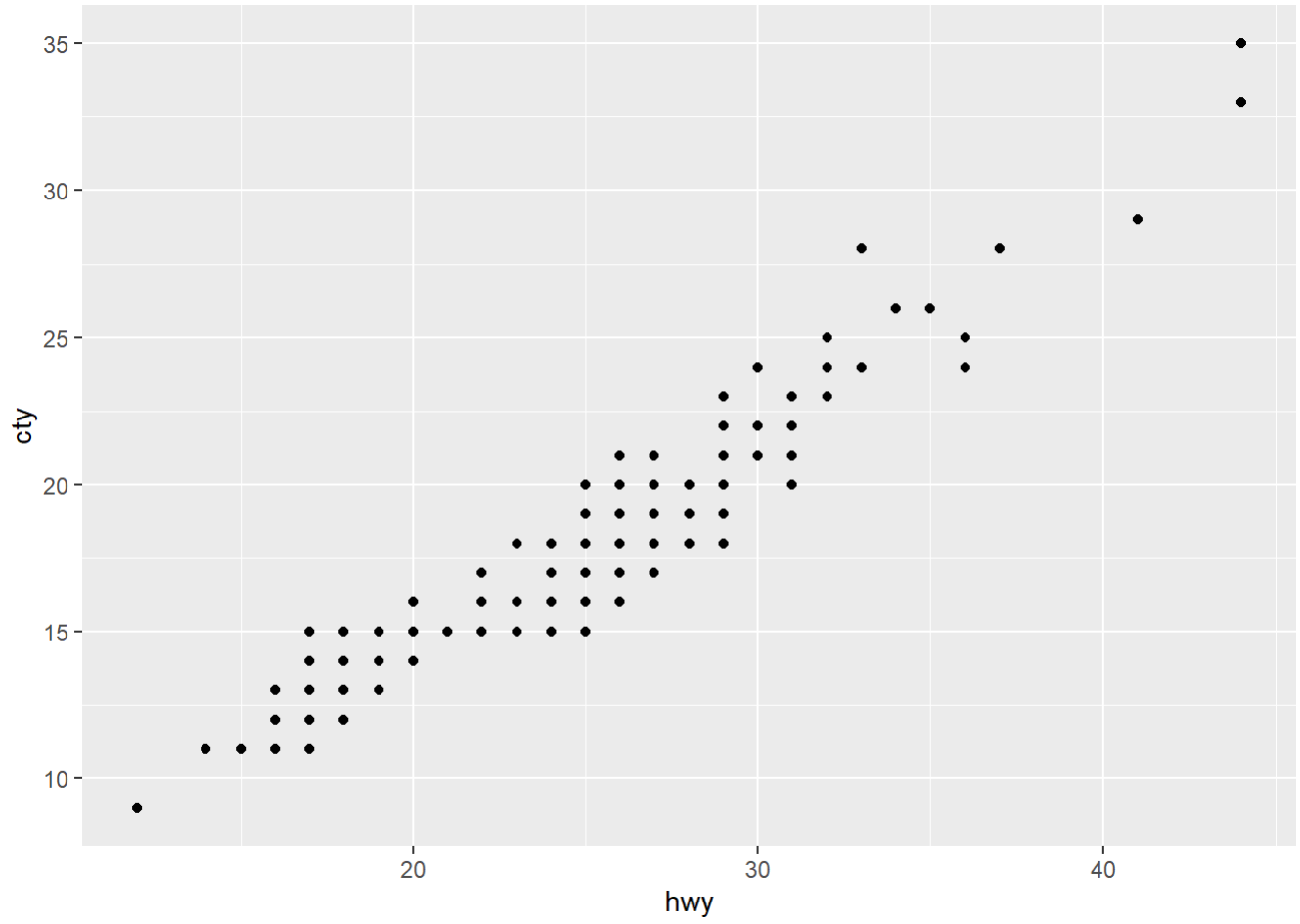
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram of highway miles per gallon is generally right skewed and bimodal. The two peaks are around 17 mpg and 27 mpg. It may also be important to note that the x-axis starts at 10 and ends at ~45.

Exercise 2

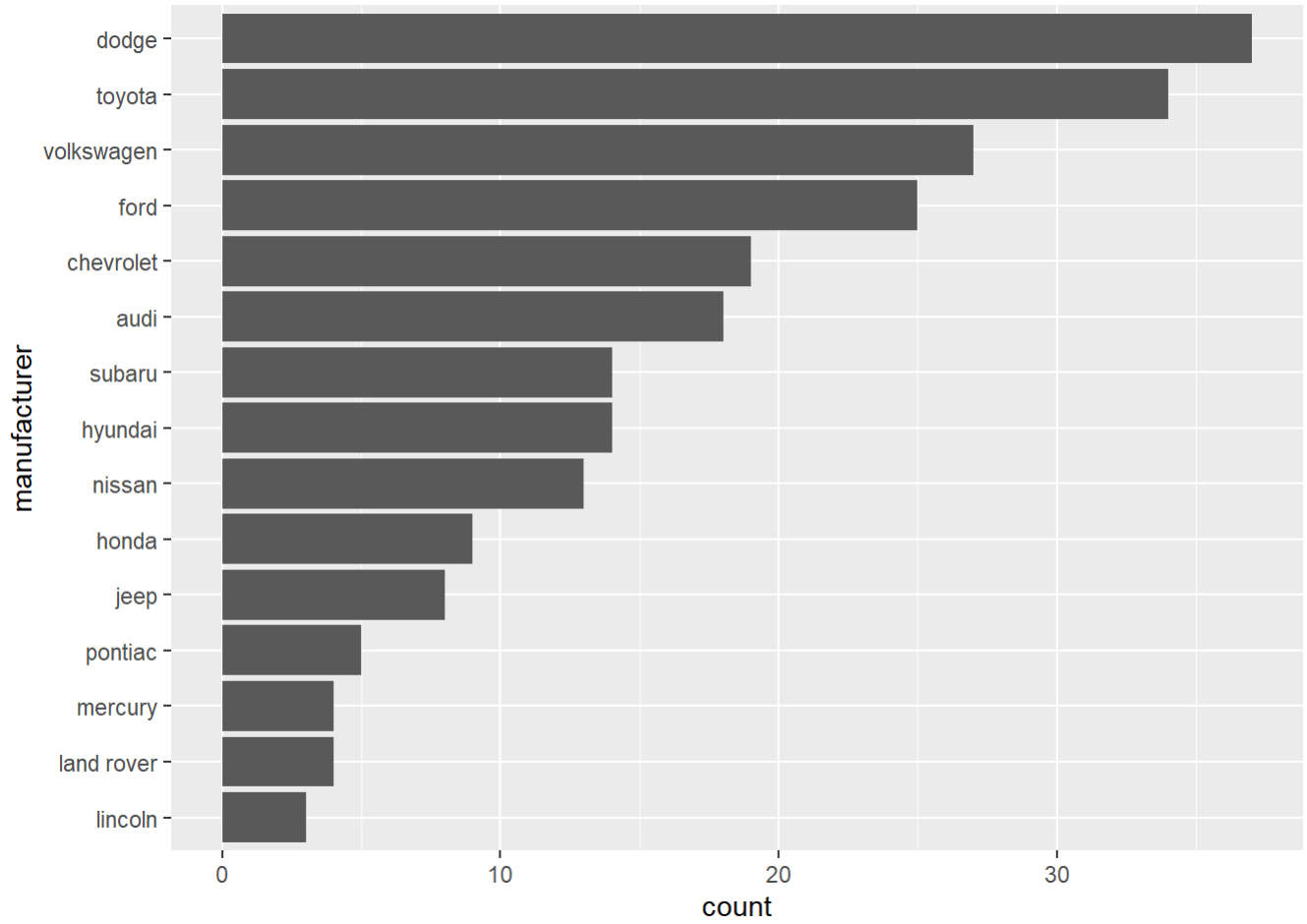
```
hwy_cty_scatter <- ggplot(mpg, aes(x=hwy, y=cty)) +
    geom_point()

hwy_cty_scatter
```



There is a strong positive relationship between `hwy` and `cty`. This means that, as the city miles per gallon increases, the highway miles per gallon also increases. Intuitively, this makes sense because those who get more mpg in any given city will also get more mpg on highways. Driving in the city tends to use a lot of fuel due to the constant stopping and starting of the car, so it makes sense that a car that gets a lot of mpg in the city will also get more mpg on the highway where there is much less exertion on the car. Although a given car does exert more energy in terms of having to use higher speeds on the highway, which is why highway miles per gallon are not extremely higher than city miles per gallon, but still higher.

Exercise 3

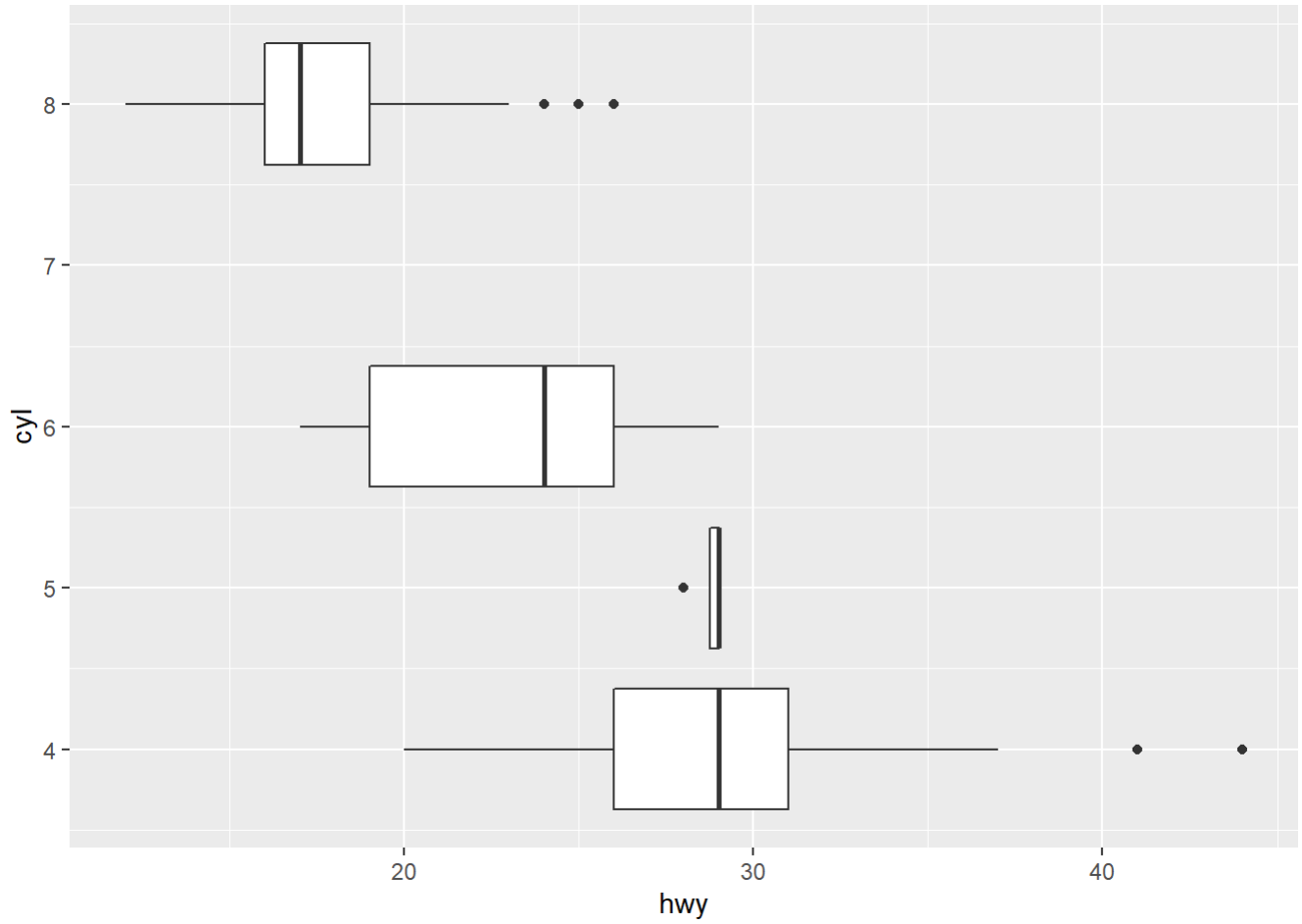
```
manufacturer_bar <- ggplot(mpg, aes(y=reorder(manufacturer, manufacturer, length))) +
  geom_bar() + labs(y="manufacturer")

manufacturer_bar
```



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

Exercise 4

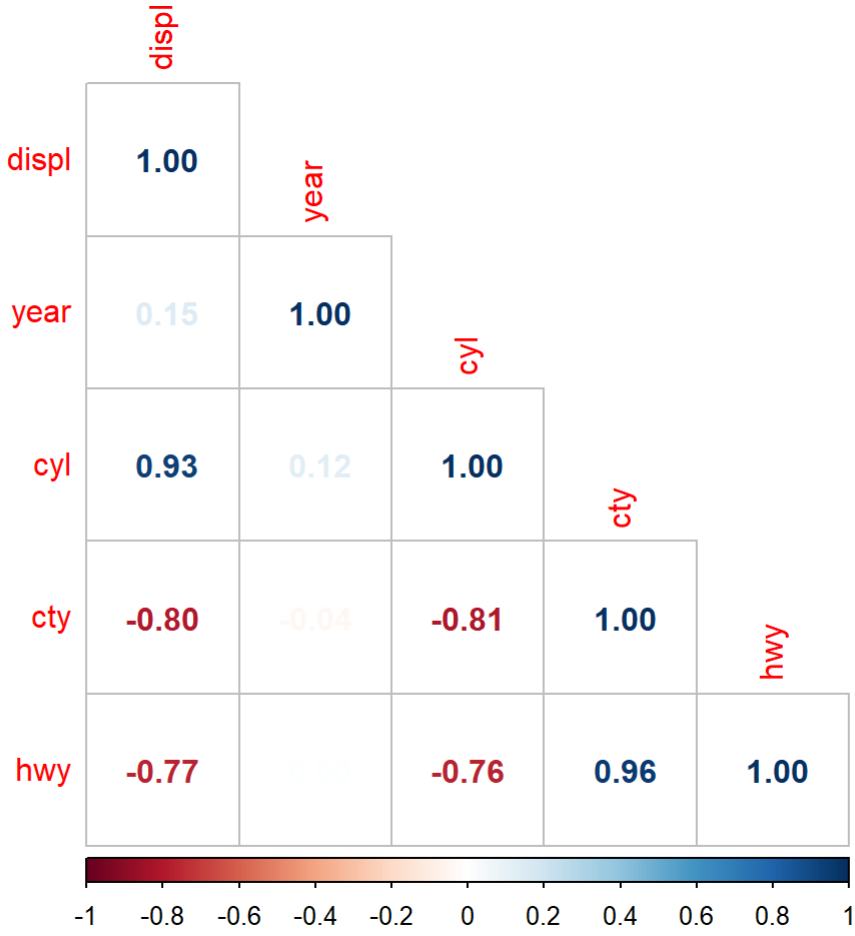
```
hwy_cyl_boxplot <- ggplot(mpg, aes(x=hwy, y = cyl, group=cyl)) +
  geom_boxplot()

hwy_cyl_boxplot
```



There is an apparent pattern shown with the boxplots. The general pattern appears to be that highway miles per gallon appear to generally decrease as the cylinder count increases.

Exercise 5

```
M <- cor(mpg[, !(names(mpg) %in% c("manufacturer", "model", "trans", "drv","fl", "class"))])
corrplot(M, method="number", type="lower")
```



Each of the variables has a perfect positive correlation with itself, which makes sense because all of the values are the same between a variable and itself.

`year` appears to have no correlation (with `cty` and `hwy`) or a very weak positive correlation (with `cyl` (0.12) and `displ` (0.15)) with all the variables. This indicates that each of the variables do not significantly change (increase or decrease) depending on the year.

`hwy` & `cty` have a strong positive correlation (0.96), which makes sense because those who get more mpg in any given city will also get more mpg on highways. Driving in the city tends to use a lot of fuel due to the constant stopping and starting of the car, so it makes sense that a car that gets a lot of mpg in the city will also get more mpg on the highway where there is much less exertion on the car. Although a given car does exert more energy in terms of having to use higher speeds on the highway, which is why highway miles per gallon are not extremely higher than city miles per gallon, but still higher.

Also, `cyl` & `displ` have a strong positive correlation (0.93). According to Wikipedia (https://en.wikipedia.org/wiki/Engine_displacement), “engine displacement is the measure of the cylinder volume swept by all of the pistons of a piston engine, excluding the combustion chambers.” So, it makes sense that the number of cylinders and engine displacement are positively correlated because, the more cylinders there are in a car, the more volume of the engine that is taken up by those cylinders (which is the engine displacement).

`cty` & `cyl` (-0.81) and `hwy` & `cyl` (-0.76) are both negatively correlated for the same reasons. According to Capital One (https://www.capitalone.com/cars/learn/finding-the-right-car/4-cylinder-vs-6-cylinder-which-is-more-fuel-efficient/1302), a car with more cylinders has more horsepower and also makes a car weigh more. Thus, it would make sense that a car with more cylinders in it would get less miles per gallon both on the highway and in the city because the car would be less fuel efficient and it would take more of the car’s energy to move the greater weight of the car (and vice versa for a car with fewer cylinders).

Furthermore, `cty` & `displ` (-0.80) and `hwy` & `displ` (-0.77) are both negatively correlated for the same reasons, which are related to the previous explaination (which might explain why the correlations are quite similar to the correlations mentioned in the previous explanation (-0.81 is close to -0.80; -0.76 is close to -0.77)). A car with more engine displacement is usually a car with more cylinders (as we saw from the strong positive correlation between `cyl` & `displ`). So, obviously, `displ` would have a similar correlation to `cty` and `hwy` as `cyl` because a higher value of `displ` would mean a higher value of `cyl` which, as I previously explained, means the car is less fuel efficient and is heavier. Therefore, it makes sense that cars get less miles per gallon both on the highway and in the city when there is greater engine displacement from a greater number of cylinders.