131 hw1 from Ruizhe Chen

2022-09-28

Question 1: Define supervised and unsupervised learning. What are the difference(s) between them? Supervised and unsupervised learning are both types of machine learning. In supervised learning, actual observed values of the outcome are known, and these values are used to "supervise" the performance of the models. In unsupervised learning, machine learning algorithms are used on data sets without actual labeled observations, to discover hidden patterns in the data.

Question 2: Explain the difference between a regression model and a classification model, specifically in the context of machine learning. The difference between regression and classification models, in the context of machine learning, is that regression models involve a continuous outcome, while classification models involve a categorical outcome.

Question 3: Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems. For classification, two common metrics are accuracy and the area under the ROC curve. For regression, two common metrics are RMSE, or root mean squared error, and R-squared.

Question4: 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: Used to visually emphasize a trend or trends in data, like fitting a linear regression to illustrate a relationship on a scatterplot. Inferential models: Used to make causal inferences about the relationship(s) between predictor(s) and the outcome. Interest is often in significance tests and implications for theories. Predictive models: Used with the goal of predicting the outcome variable as accurately as possible, with minimized reducible error.

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 assume a parametric form for the relationship between the predictor(s) and the outcome, which very likely will not match the true, unknown form. They tend to have higher bias and lower variance. Empirically-driven models make little to no assumptions about the form of the relationship, and are more flexible by default. They tend to have higher variance and lower bias. In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice. Mechanistic models tend to be easier to understand because they generally fit relatively simple parameteric forms. Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models. Mechanistic models have higher bias and lower variance; empirically-driven models have higher variance and lower 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; the campaign is focused on determining the probability of voter behavior. The second question is inferential; the campaign is more curious here about the specific relationship between a predictor, candidate interaction, and the outcome, voter behavior

exercise1:

```
#install.packages("magrittr") # package installations are only needed the first time you use it
#install.packages("dplyr")    # alternative installation of the %>%
library(magrittr) # needs to be run every time you start R and want to use %>%
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.0.0 --

## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tibble       3.1.8
## v infer        1.0.3     v tidyr        1.2.1
## v modeldata    1.0.1     v tune         1.0.0
## v parsnip      1.0.1     v workflows    1.1.0
## v purrr        0.3.4     v workflowsets 1.0.0
## v recipes      1.0.1     v yardstick    1.1.0

## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x purrr::discard()   masks scales::discard()
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
## x recipes::step()    masks stats::step()
## * Learn how to get started at https://www.tidymodels.org/start/
```

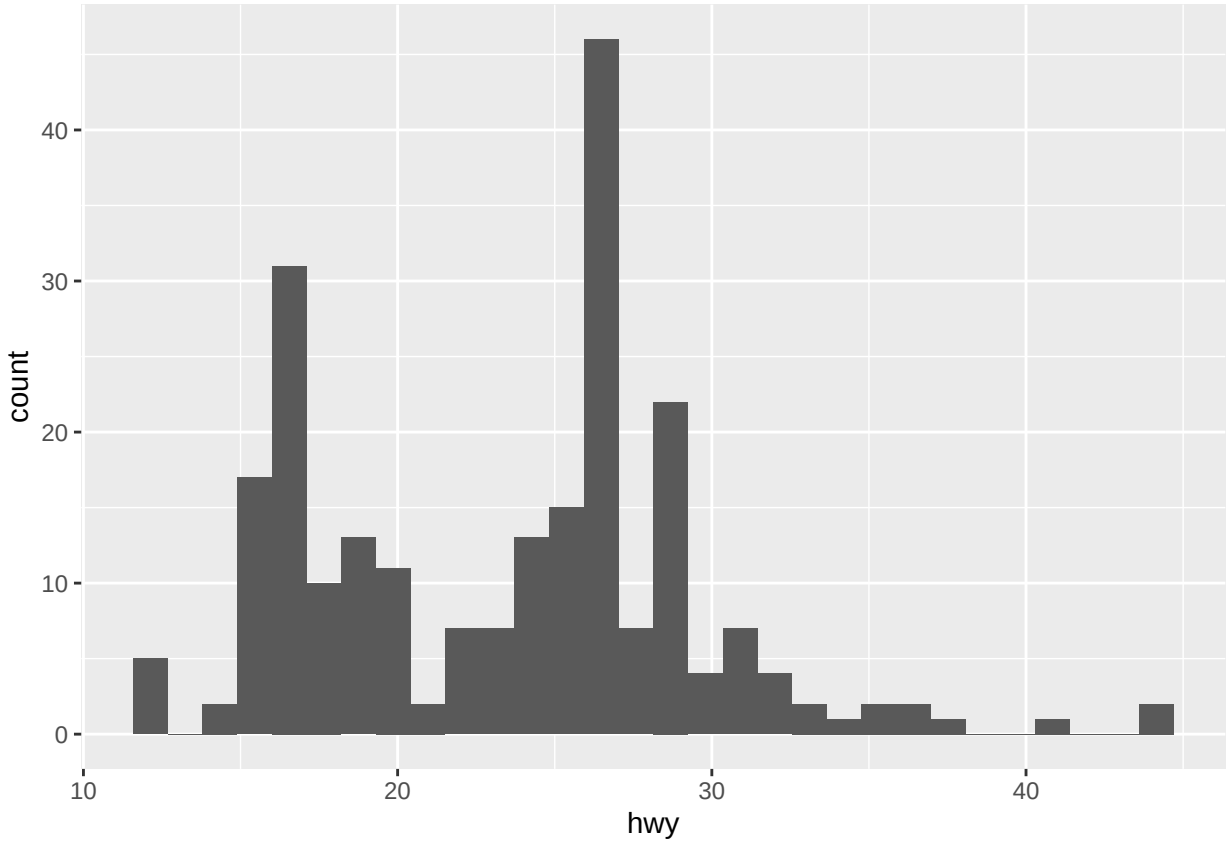
```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v readr   2.1.2     v forcats 0.5.2
## v stringr 1.4.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x tidyr::extract()    masks magrittr::extract()
## x dplyr::filter()     masks stats::filter()
## x stringr::fixed()    masks recipes::fixed()
## x dplyr::lag()        masks stats::lag()
## x purrr::set_names()  masks magrittr::set_names()
## x readr::spec()       masks yardstick::spec()
```

```
mpg %>%
 ggplot(aes(x = hwy)) +
 geom_histogram()
```

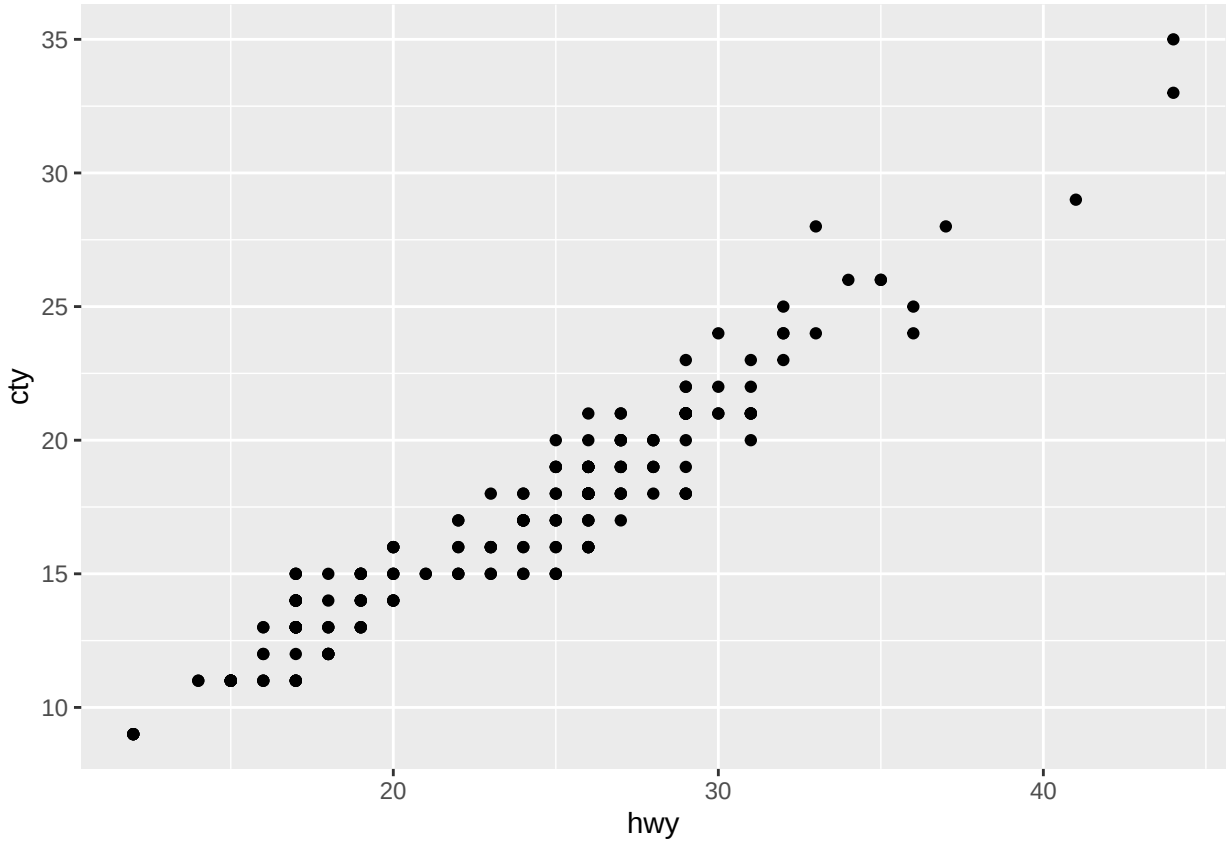
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The distribution of highway mileage definitely appears to be positively skewed, at least to a degree. It also almost looks bimodal, although those distributions are relatively rare; there's one peak around 26-27 mpg, and another around 16-17. Only a few cars have more than 40 highway mpg.

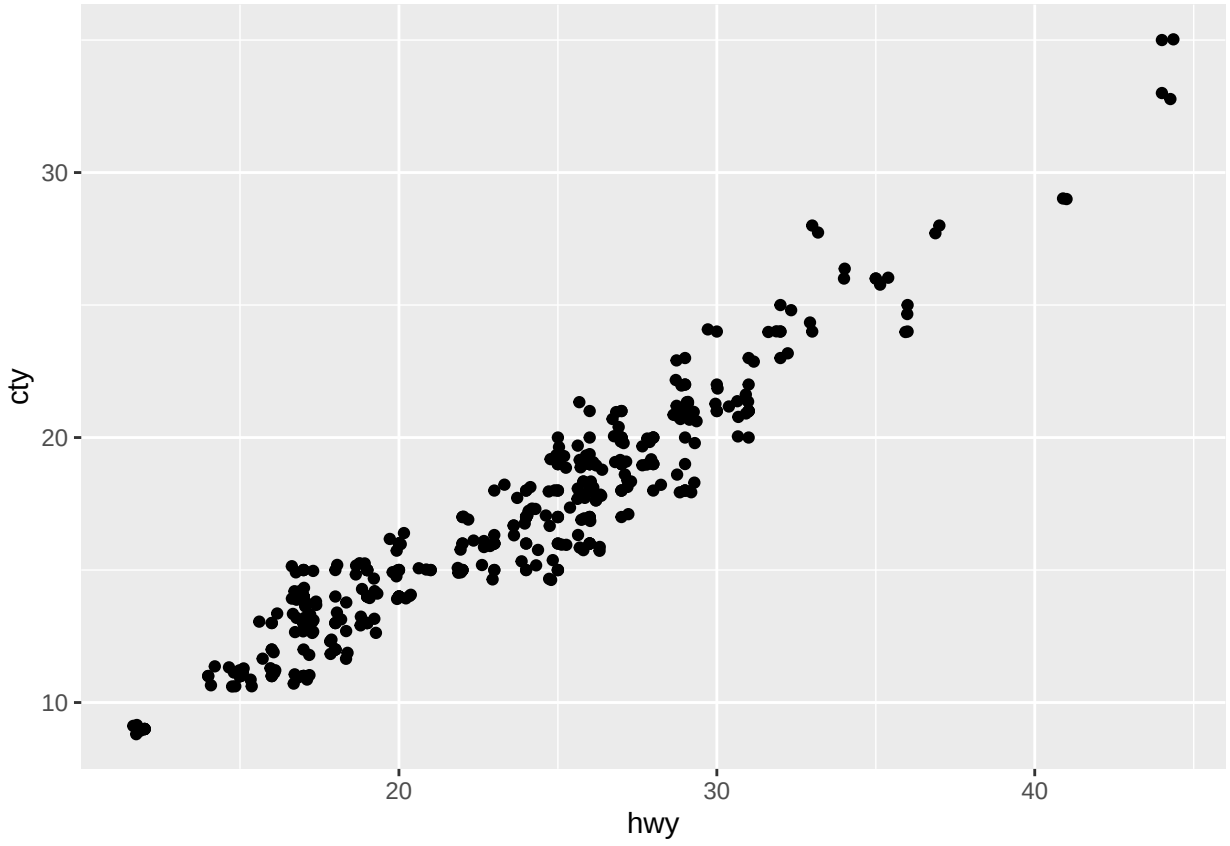
exercise2:

```
mpg %>%
 ggplot(aes(x = hwy, y = cty)) +
 geom_point()
```



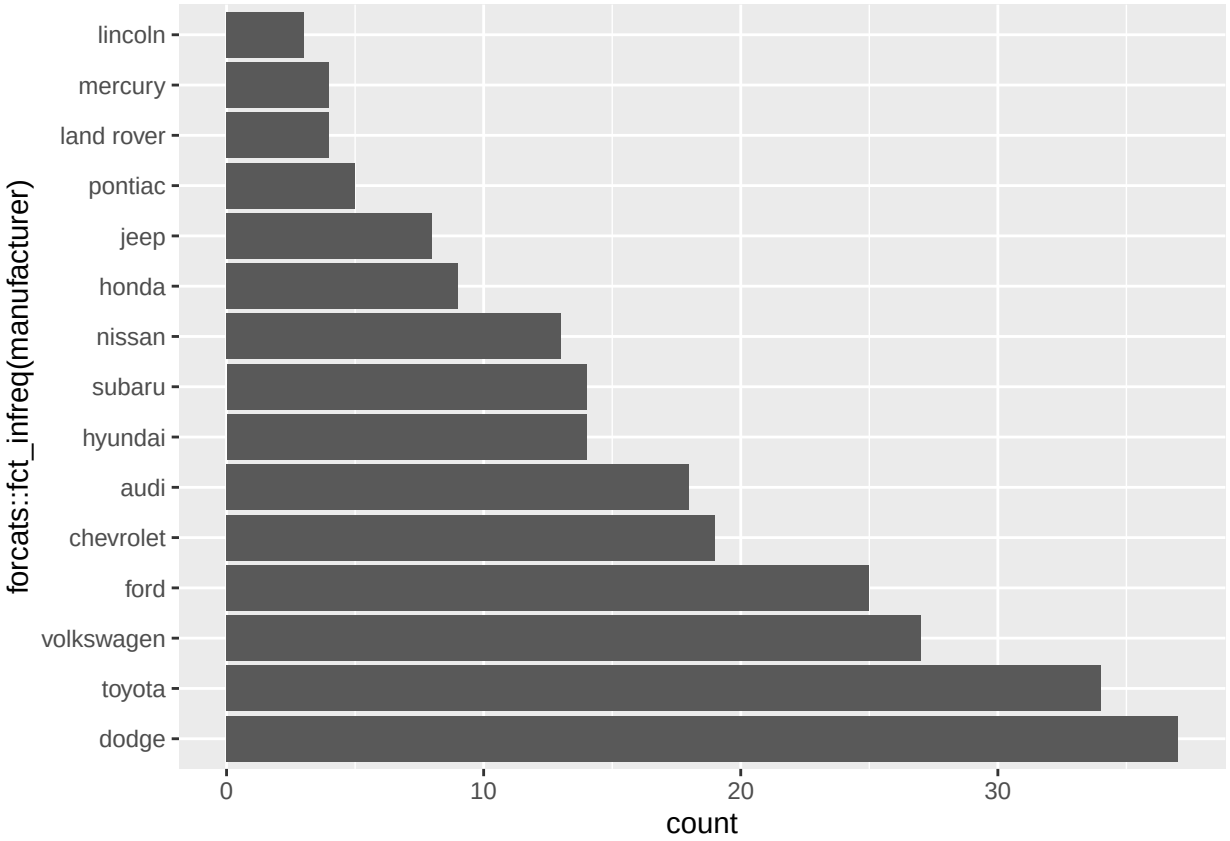
There certainly appears to be a positive linear correlation between highway mileage and city mileage. One thing to note is that the points seem to have a grid-like pattern. This is because the observations of both mileage variables are round numbers, so some of the points are appearing on top of each other. We could "fix" that with some jitter, or random noise, added:

```
mpg %>%
 ggplot(aes(x = hwy, y = cty)) +
 geom_point() +
 geom_jitter()
```



exercise3:

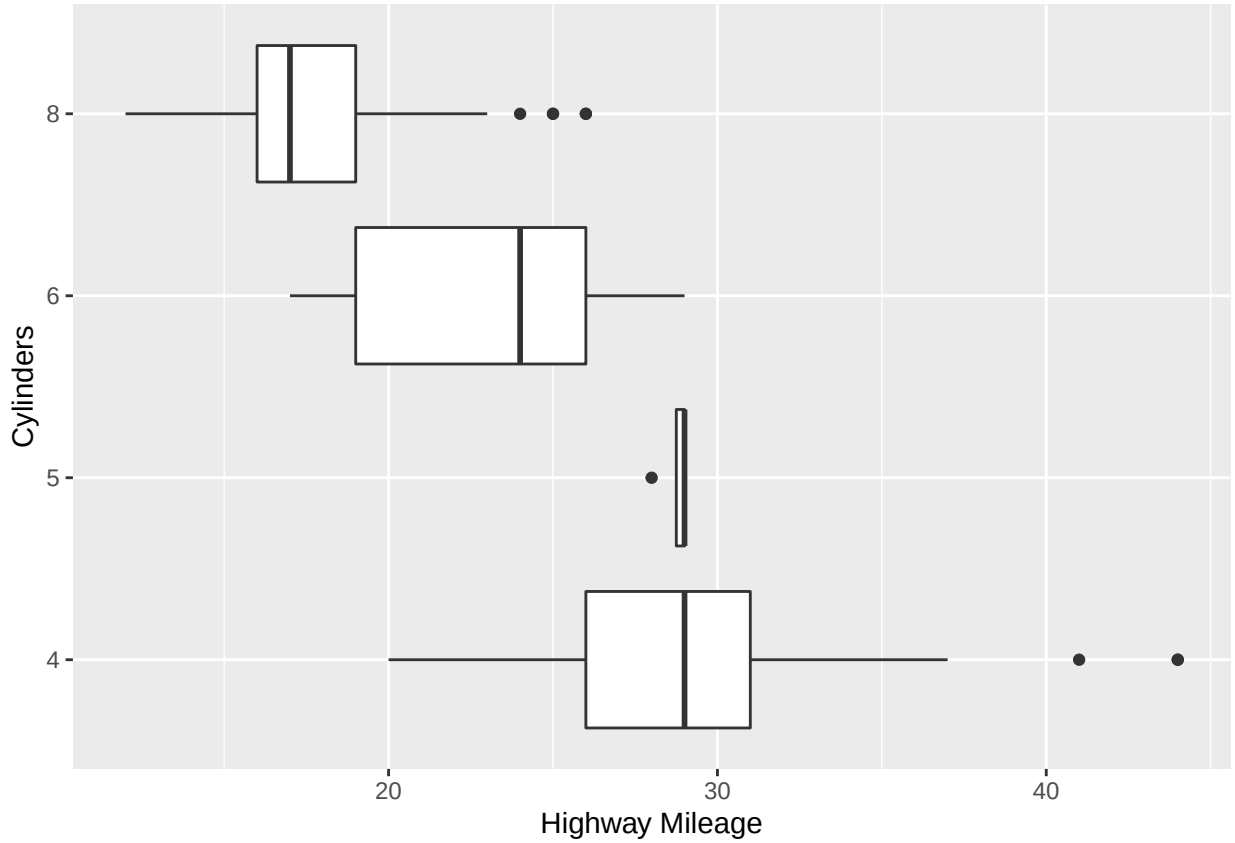
```
mpg %>%
 ggplot(aes(x = forcats::fct_infreq(manufacturer))) +
 geom_bar() +
 coord_flip()
```



Dodge produced the most cars, almost 40; Lincoln produced the least, below 5.

exercise4:

```
mpg %>%
 ggplot(aes(x = hwy, y = factor(cyl))) +
 geom_boxplot() +
 xlab("Highway Mileage") +
 ylab("Cylinders")
```



Yes; as the number of cylinders increases, the highway mileage tends to decrease. Cars with four cylinders have the highest mileage on average. Five-cylinder cars are close in terms of average, but there are far fewer five-cylinder cars in the data set.

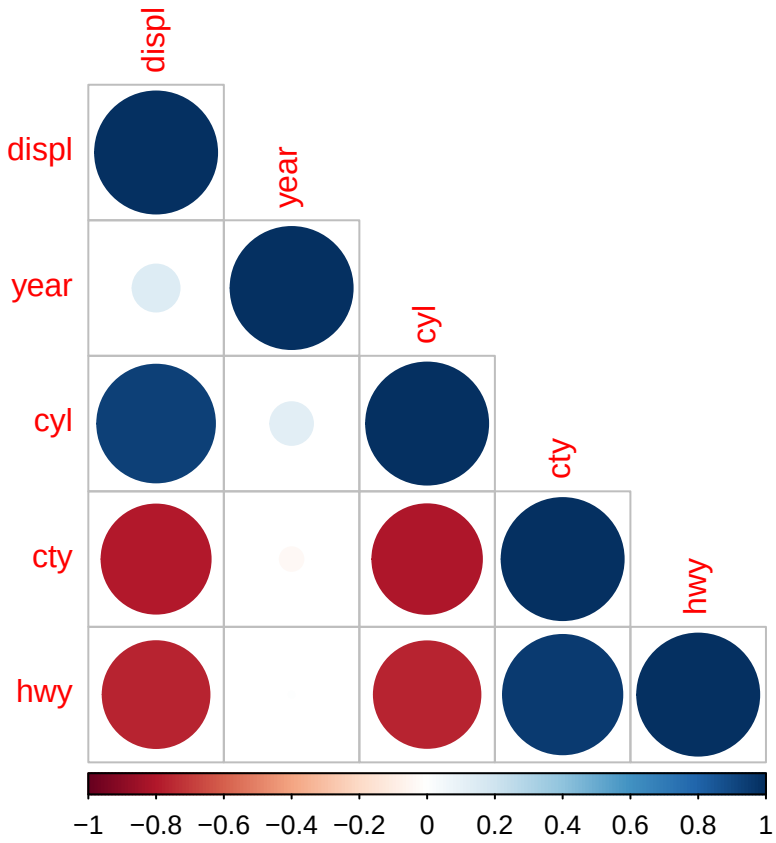
Exercise 5: Since there are some categorical variables in the data set, you can decide how to deal with these. One way is simply to exclude them from the plot, which is what is done here.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mpg %>%
 select(is.numeric) %>%
 cor() %>%
 corrplot(type = "lower")
```

```
## Warning: Predicate functions must be wrapped in `where()`.
##
##   # Bad
##   data %>% select(is.numeric)
##
##   # Good
##   data %>% select(where(is.numeric))
##
## i Please update your code.
## This message is displayed once per session.
```



Number of cylinders is positively correlated with displacement, and highway and city mileage are positively correlated with each other. Displacement is negatively correlated with both mileage variables, as is number of cylinders.