Homework_1

Vardan Martirosyan

2022-09-30

Machine Learning Main Ideas Questions

Question 1

We are asked to define supervised and unsupervised learning, and to state the differences between them. As stated in the lecture, supervised learning is when we use the “actual data ‘Y’ as the supervisor”. That is, we have to give the model the observed output and input, so that it can learn what is the correct output (and what is not), and adjust it’s learning process based on that information. Unsupervised learning is when the learning for the model occurs without giving the machine any indication on what answers are correct or not.

As can be inferred from above, the main difference between the two learning methods is that supervised learning uses input and output data that is classified, while unsupervised learning only uses input data that is classified. Another difference between the two learning methods is the types of algorithms that are used. As we learned in lecture, supervised learning consists of algorithms such as regression and trees, while unsupervised learning consists of clustering algorithms.

Question 2

We are asked to explain the difference between a regression model and a classification model, specifically in the context of machine learning. As stated in the lecture, a regression model has the response variables take on “quantiative, numerical values” (such as temperature, weight, etc). In a classification model, the response variables take on “qualatitive, categorical values” (such as color, car type, etc).

Question 3

We are asked to name two commonly used metrics for regression ML and classification ML problems each. For Regression Machine Learning, two commonly used metrics are the Mean Square Error and the Root Mean Square Error. For Classification Machine Learning, two commonly used metrics are the F1 score, and the accuracy rate.

Question 4

We are asked to give a brief description of descriptive, inferential, and predictive models. We do this as follows:

Descriptive models: In descriptive models, our goal is to choose the top model that can effectively visualize a pattern seen in our data. For example, as stated in lecture, we may consider “using a line on a scatterplot”.

Inferential models: Inferential models have a few goals in mind. (Most of the goals listed here are rephrased from Slide 39 of the Day 1 Slides). One goal is to determine which features in a model, if any, are the most important ones. Another goal for this type of model is to be able to determine the link between the predictors of a model, and it's outcomes. Yet another goal is to test theories about the model, and determine if these theories are indeed true or not. Finally, it could possibly be used to determine any causal claims.

Predictive models: As stated in lecture, hypothesis tests are not a main focus of these types of models. Additionally, the goal with predictive models is to determine which combination of features would work most effectively with the fit. Finally, as stated in the lecture, it's "aim is to predict Y with minimum reducible error".

Question 5.1

We are asked to define the mechanistic and empirically driven model types, and describe how they differ and how they are similar.

As stated in the lecture, mechanistic model types are parametric, and "assume a parametric form for f". In addition, they will not be equal to the actual value of f, which is not known. Additionally, by adding more parameters to the model, the model becomes "more flexible" (as stated in the lecture). Finally, we could overfit the model if we accidentally add a larger-than-necessary amount of parameters.

On the other hand, empirically driven (non-parametric) models have absolutely no prior assumptions regarding the function f. However, as stated in lecture, they do "require a larger number of observations" in order to work. By their initial construction, they are more flexible than parametric models, but also suffer from overfitting.

A similarity of both of the models is that they both have the ability to suffer from overfitting. The differences are as follows: in mechanistic models, there is an assumption made for the function f, but for the non-parametric model, there is no assumption made about f. Additionally, when looking at their default constructions, the non-parametric model is "much more flexible" (as stated in lecture) than the mechanistic model.

Question 5.2

We are asked to decide if mechanistic models or emprically driven models are easier to understand, and to explain our choice. I personally think that mechanistic models are easier to under, because we are making an implicit assumption about f and what it's form looks like. Even if it is not the exact true form of f, it still lets us perform different types of analysis that I (personally) find easier to understand than those of empirically driven models.

Question 5.3

We are asked to describe how bias-variance tradeoff is related to the use of mechanistic or empirically-driven models. As we learned in class, as the complexity/flexibility of the model increases, it has a lower bias and a higher variance. As the complexity/flexibility of a model decreases, it has a higher bias, but a lower variance. In the context of mechanistic and empirically driven models, we recall from our answer to 5.1 that empirically driven models were inherently more flexible than mechanistic ones. Thus, emprically driven models are more likely to have a higher variance, but a lower bias, than mechanisitic ones. So, if we were looking to choose a model that had a low bias, but a high variance, we might choose to pick an empirical model over a mechanistic one. This is how the bias-variance tradeoff is related to the use of mechanistic or empirically driven models. Depending on what types of tradeoffs we would be willing to give, in regards to the bias and variance, we can then choose to use a mechanistic or an empirically-driven model to accomplish our goals.

Question 6

I would classify the first question as predictive, and the second question as inferential.

The reason that I would classify the first question as predictive is because of the fact that they are trying to "predict" how likely a voter would vote in favor of the candidate based on the voter's profile/data. In other words, they want to use the voter's profile/data as predictors to determine how likely the physical action would be of the voter actually voting for the candidate.

The reason that I would classify the second question as inferential is because they are trying to infer the likliehood of the voter's change of support, based on an interaction with the candidate. Since the voter's change of support isn't something that can be easily measured (ie, compared to someone going and physically voting), I'd argue that this event is more likely to be inferential than predictive.

Exploratory Data Analysis Exercises

First, we call the library ‘ggplot2’ to be able to use it, along with the other libraries we need.

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## 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
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.0.0 --
## v broom        1.0.1     v rsample      1.1.0
## v dials        1.0.0     v tune         1.0.0
## v infer        1.0.3     v workflows    1.1.0
## v modeldata    1.0.1     v workflowsets 1.0.0
## v parsnip      1.0.1     v yardstick    1.1.0
## v recipes      1.0.1
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

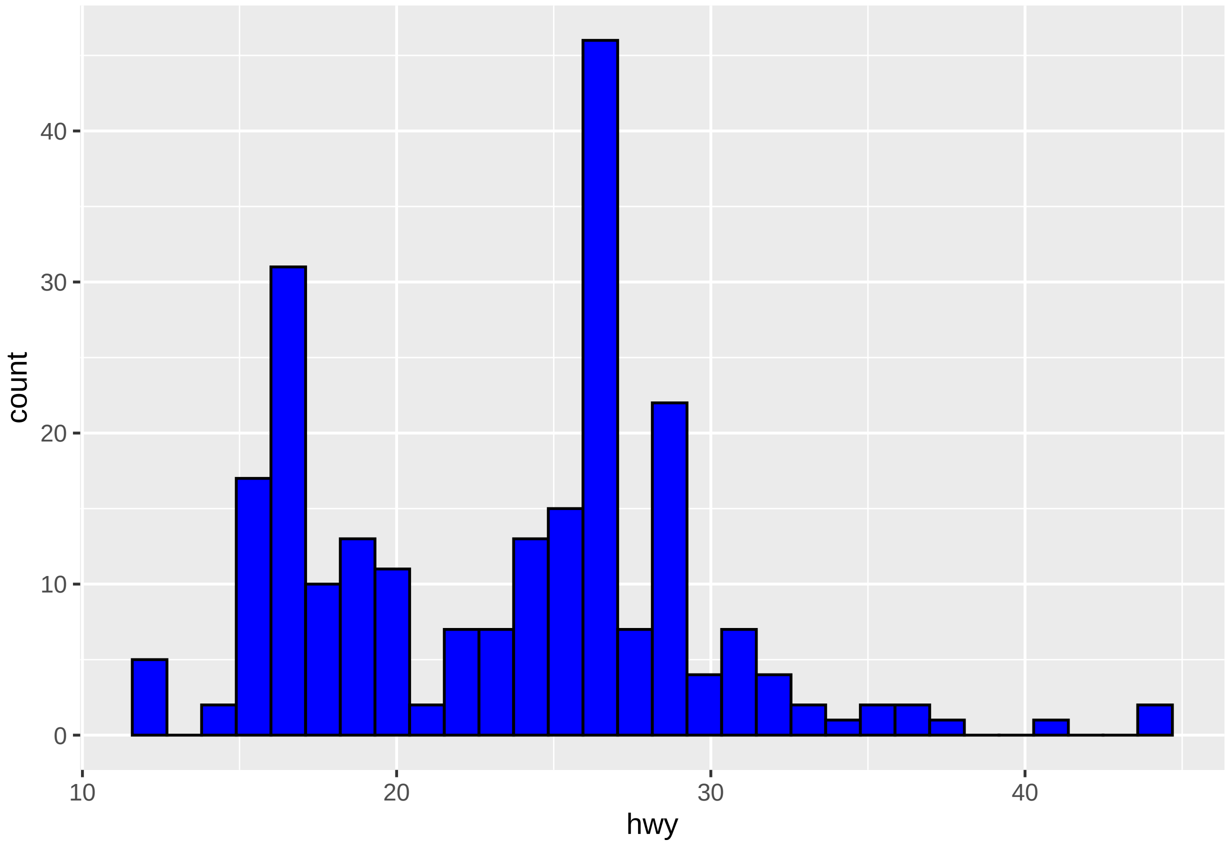
```
library(ISLR)
```

Question 1

We are asked to create a histogram of the variable ‘hwy’ of the ‘mpg’ dataset. Below is the histogram.

```
ggplot(data = mpg, aes(x = hwy)) + geom_histogram(color = 'black', fill = 'blue')
```

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

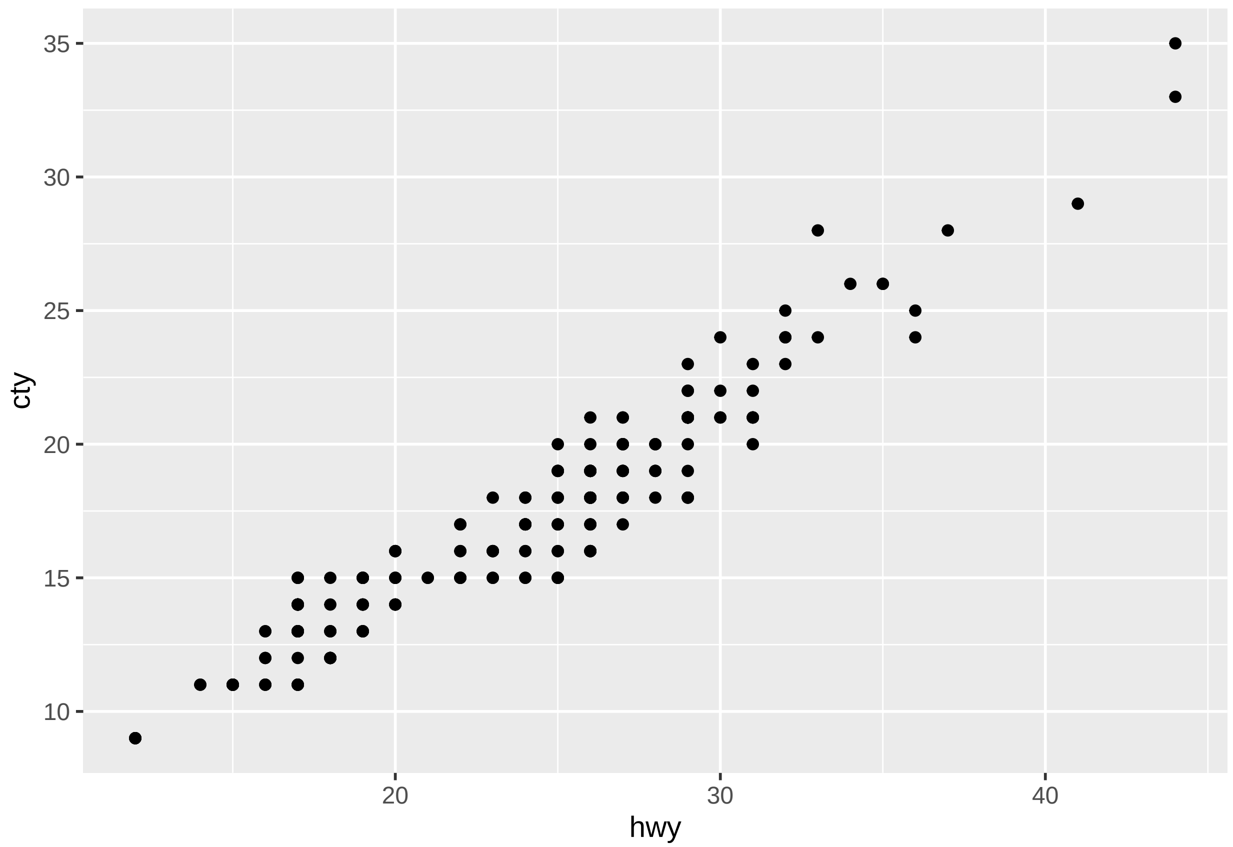


We are then asked to describe what we see/learn. Looking at the histogram, it seems to be bimodal, with two peaks happening at hwy = 17 and hwy = 27. In addition, most of the other values for this variable appear to be crowding around these two peaks, with some outliers on both ends (but most of the outliers primarily appearing on the higher end). From this, we can learn that most of the cars in the dataset seem to have an MPG of around 15-20 mpg or 25-30 mpg for their miles per gallon for their highway mileage.

Question 2

We are now asked to create a scatterplot, putting 'hwy' on the x-axis and 'cty' on the y-axis. Below is the scatterplot:

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



We are asked if there is a relationship between hwy and cty. Looking at the scatterplot, we state that there is indeed a relationship between hwy and cty: namely, a positive, linear relationship. This tells us that as the miles per gallon for the highway mileage increases, so does the miles per gallon for the city mileage.

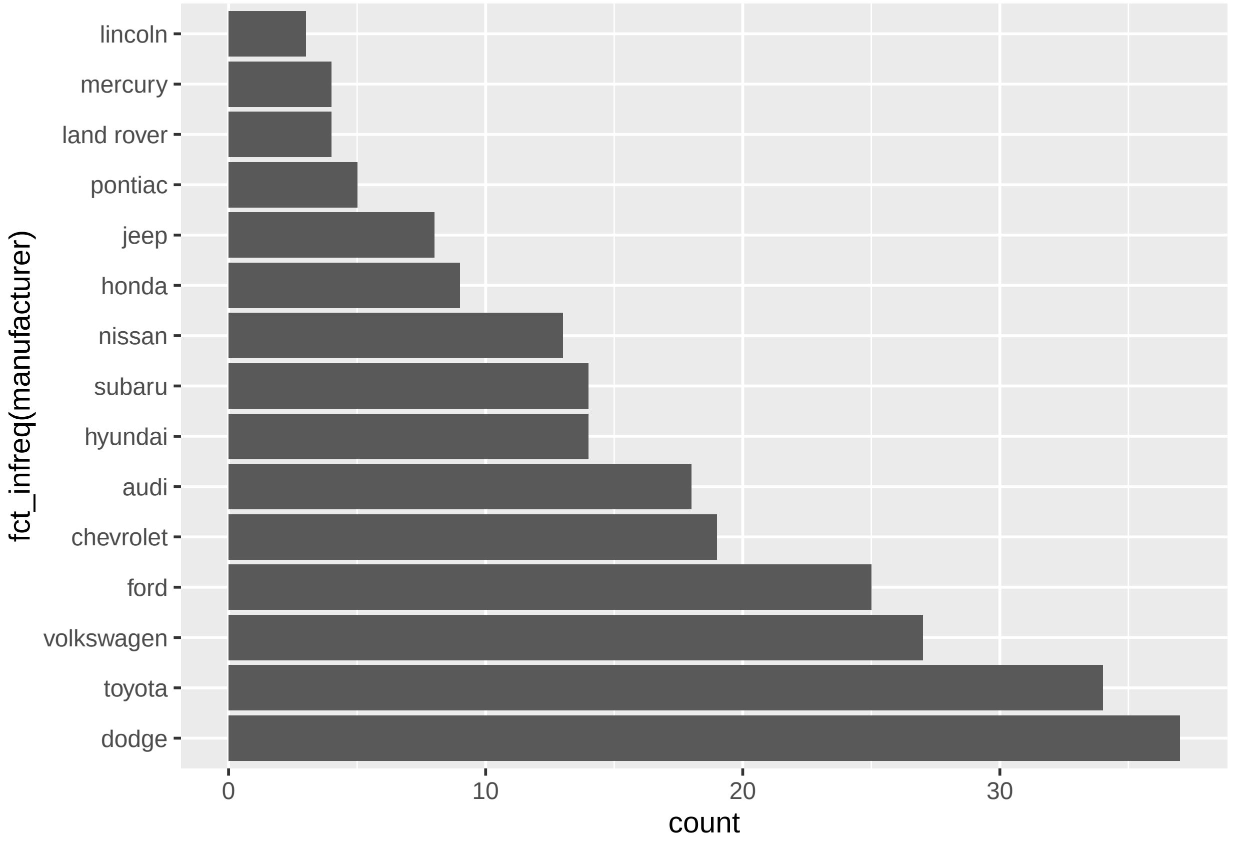
Question 3

We are asked to make a bar plot of manufacturer, flip it so that the manufacturers are on the y-axis, and to order the bars by height. To do this, we first need to call the library 'forcats' from tidyverse that will let us re-order the bars in the bar graph.

```
library(forcats)
```

Then, we can create the bar plot as requested, using the 'fct_infreq' function:

```
ggplot(mpg, aes(x = fct_infreq(manufacturer))) + geom_bar(stat = "count") + coord_flip()
```

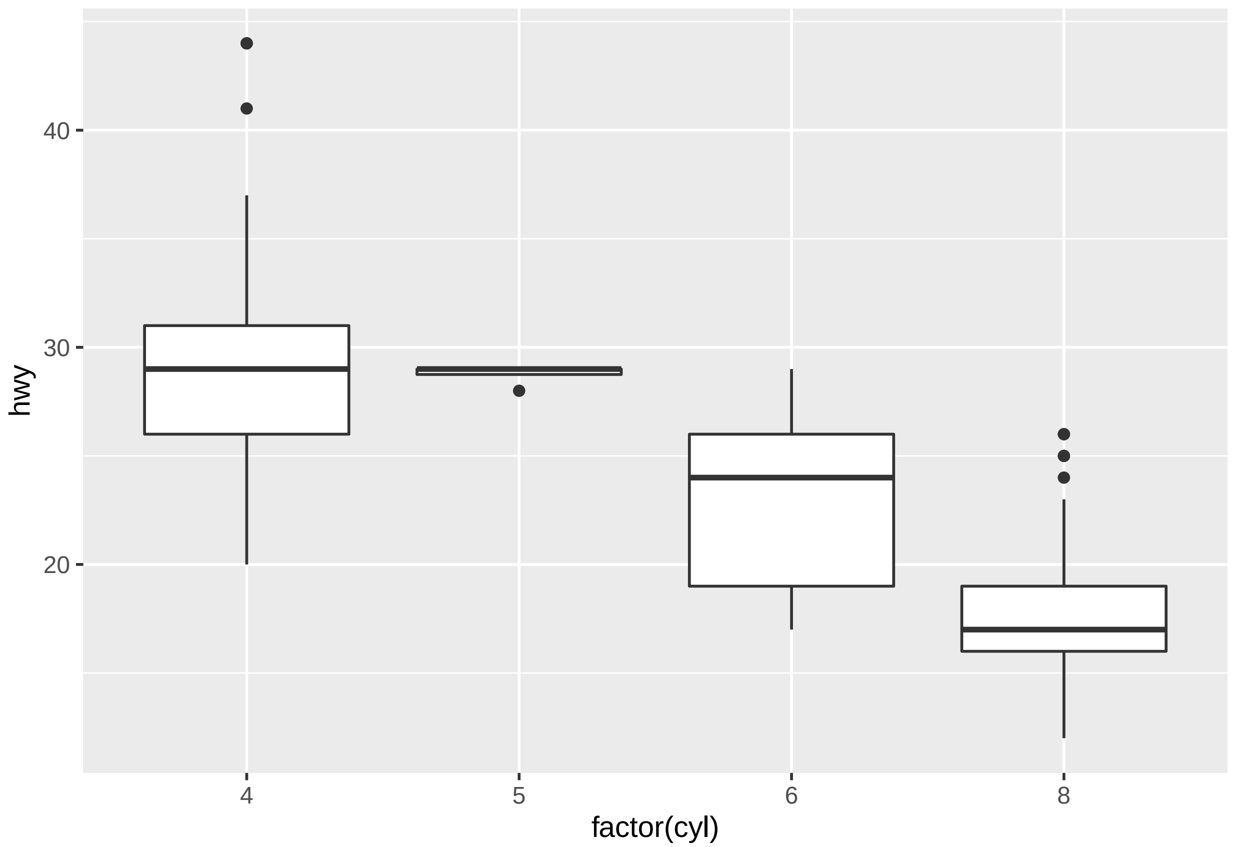


We are then asked which manufacturer produced the most cars, and which one produced the least. From the bar plot above, we see that the one that produced the most is Dodge, and the one that produced the least is Lincoln.

Question 4

We are asked to make a boxplot of hwy, grouped by cyl. We do this below as follows:

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



We are then asked if we see a pattern. Looking at this boxplot, I do indeed see a pattern. As the number of cylinder's increases, the miles per gallon for the highway mileage of cars decreases in a linear fashion.

Question 5

We are asked to use corrplot() to make a lower triangle correlation matrix of the mpg dataset. We first call the corrplot() package.

```
library(corrplot)
```

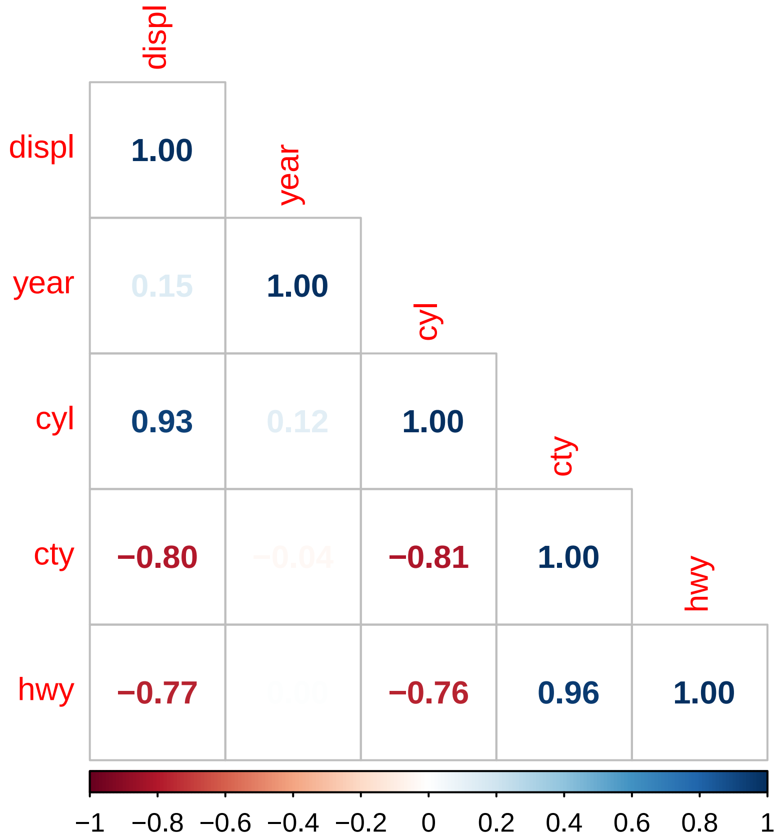
```
## corrplot 0.92 loaded
```

We then try to write the code to obtain a lower triangle correlation matrix of the mpg dataset as follows. However, when we try and run this line of code, we get thrown an error. In particular, the error is because 'manufacturer' is type 'character'. As suggested by Hanmo Li in office hours, to resolve this issue, we need to remove not only 'manufacturer' from the dataset, but all non-numeric variables. We do this as follows.

```
new_mpg <- mpg
new_mpg$manufacturer <- NULL
new_mpg$model <- NULL
new_mpg$trans <- NULL
new_mpg$drv <- NULL
#new_mpg$displ <- NULL
new_mpg$fl <- NULL
new_mpg$class <- NULL
```

Then, we can use corrplot() as desired:

```
corrplot(cor(new_mpg), method = 'number', type = 'lower')
```



We are asked which variables are positively or negatively correlated with each other. Additionally, we are asked to state if these relationships make sense to us, or surprise us. We first note that the 'cty' and 'hwy' variables are strongly positively correlated with one another. This one makes sense to me, since if the highway mileage for a car was high, then it should mean that the city mileage for the car is also similarly high. In other words, it doesn't really make sense for a car to have a high highway average mpg, but suddenly not have such a high city average mpg. Personally, I think it would be very strange if a car had a high highway miles per gallon (relative to other cars), then suddenly had a low city miles per gallon (relative to other cars). This is why it makes sense to me that the 'cty' and 'hwy' variables are strongly positively correlated with one another. Another strong positive relationship is between 'displ' and 'cyl'. This one makes sense, since the displacement variable represents the engine displacement (in liters), which is related to the number of cylinders a car has.

There are four strong negative relationships. The first two were the variables 'cty' and 'cyl', and 'hwy' and 'cyl'. This relationship was initially a bit surprising to me, since intuitively, I didn't expect the number of cylinder's in a car to drastically affect the performance of the car's average mpg on the highway or the city. However, after looking back at Exercise 4 in this section, this negative correlation makes sense, since we saw in that Exercise that as the number of cylinder's increased, the car's miles per gallon on the highway decreased. The other two strong negative relationships were 'displ' and 'cty', and 'displ' and 'hwy'. While I was initially suprised about this too, after looking at the previous two strong negative relationships, these last two relationships made sense to me, since 'displ' and 'cyl' are strongly related to one another.

Other than these relationships, there are no other positively/negatively correlated variables that made sense to me, or surprised me. (Based on what the question asked, I don't think we were meant to talk about the ‘Year’ correlations with the other variables, since it’s relationships with the other variables are not that strong or negative, and seem pretty insignificant).