Homework Assignment 1

Reynaldo Perez

October 2, 2022

Q1. In supervised learning, the actual data Y is the supervisor, and we need to give the model observed outputs and inputs. In unsupervised learning, the machine learns without a supervisor, meaning no response for our inputs. Key differences include different uses. For example, we may use supervised learning for linear regression and logistic regression, but not for unsupervised learning. On the contrary, unsupervised learning may be useful k-means clustering and neural networks.

Q2. In a regression model, Y is quantitative, whereas in a classification model, Y is qualitative. In other words, regression algorithms predict numerical values, whereas classification algorithms predict categorical values.

Q3. Two commonly used metrics for regression ML problems are mean squared error (MSE) and root mean squared error (RMSE). Two commonly used metrics for classification ML problems are the Bayes error rate, and precision.

Q4.

· Descriptive models: Choose a model that visually best emphasizes a trend in data.

· Inferential models: Aim to test theories and claims, as well as find relationships between outcome and predictors.

· Predictive models: Aim to predict Y with very little reducible error.

Q5.

· Mechanistic uses theory to predict real world phenomena. Empirically-driven uses observed data or past experiences. Differences include: mechanistic assumes a parametric form, whereas empirically-driven has no assumption about f. Furthermore, mechanistic won't match true unknown f, and empirically-driven requires a larger number of observations. Similarities include both of them being able to be flexible and overfitting.

· An empirically-driven model is easier to understand because we can gather information based on observation and past experiences, whereas a mechanistic model is all based on theory.

· Both models are able to be flexible. In bias-variance tradeoff, higher model flexibility leads to data points matching better. Both models are also overfitting, meaning the bias-variance tradeoff can be applied.

Q6.

· Predictive. We are aiming to predict a response Y.

· Inferential. We are aiming to test a theory and state the relationship between the outcome and predictor.

```r
library(tidyverse)  # Load tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.6      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.2
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)  # Load ggplot2
```

Exercise 1) Let us create a histogram of the *hwy* variable in the mpg data set.

```
nrow(mpg)  # Number of rows in the data set
```

```
## [1] 234
```

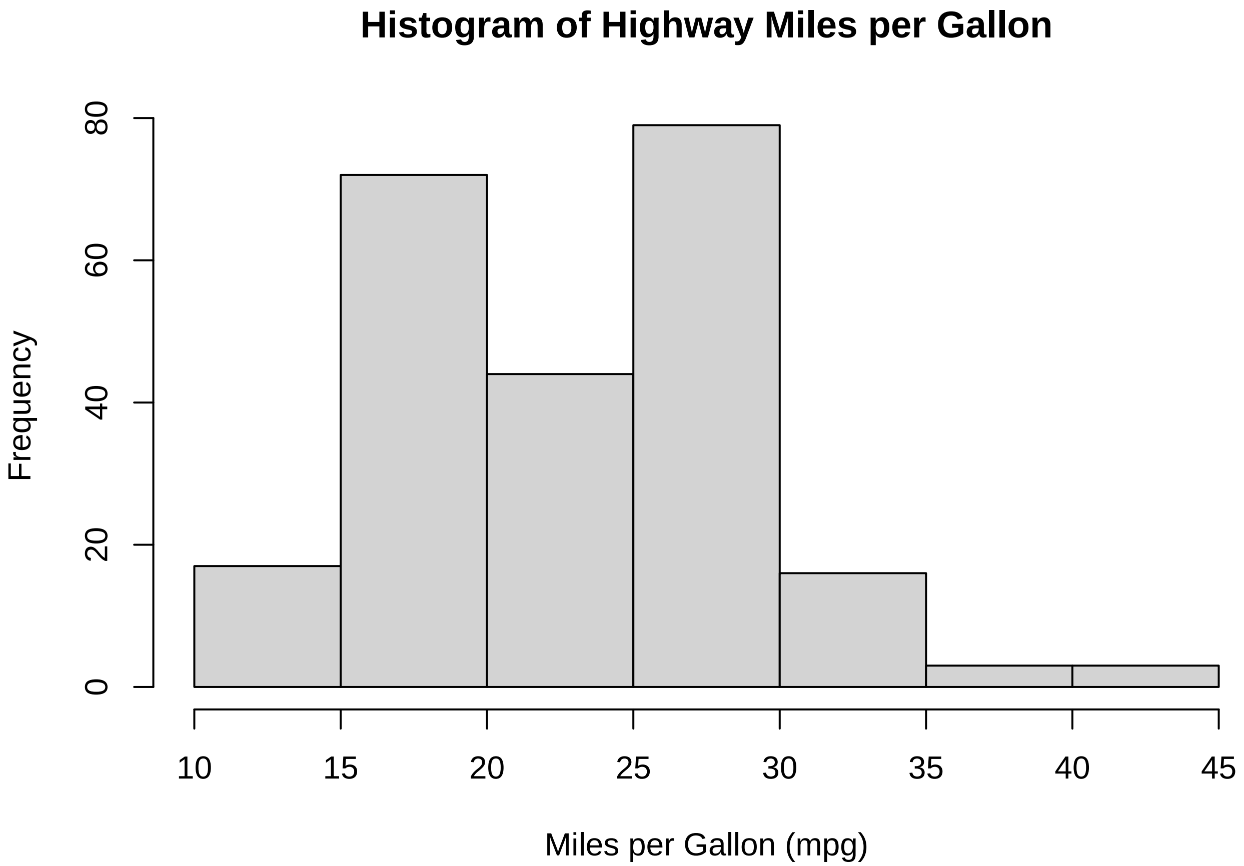
```
ncol(mpg)  # Number of columns in the data set
```

```
## [1] 11
```

```
colnames(mpg)  # Column names in the data set
```

```
##  [1] "manufacturer" "model"        "displ"        "year"         "cyl"
##  [6] "trans"        "drv"          "cty"          "hwy"          "fl"
## [11] "class"
```

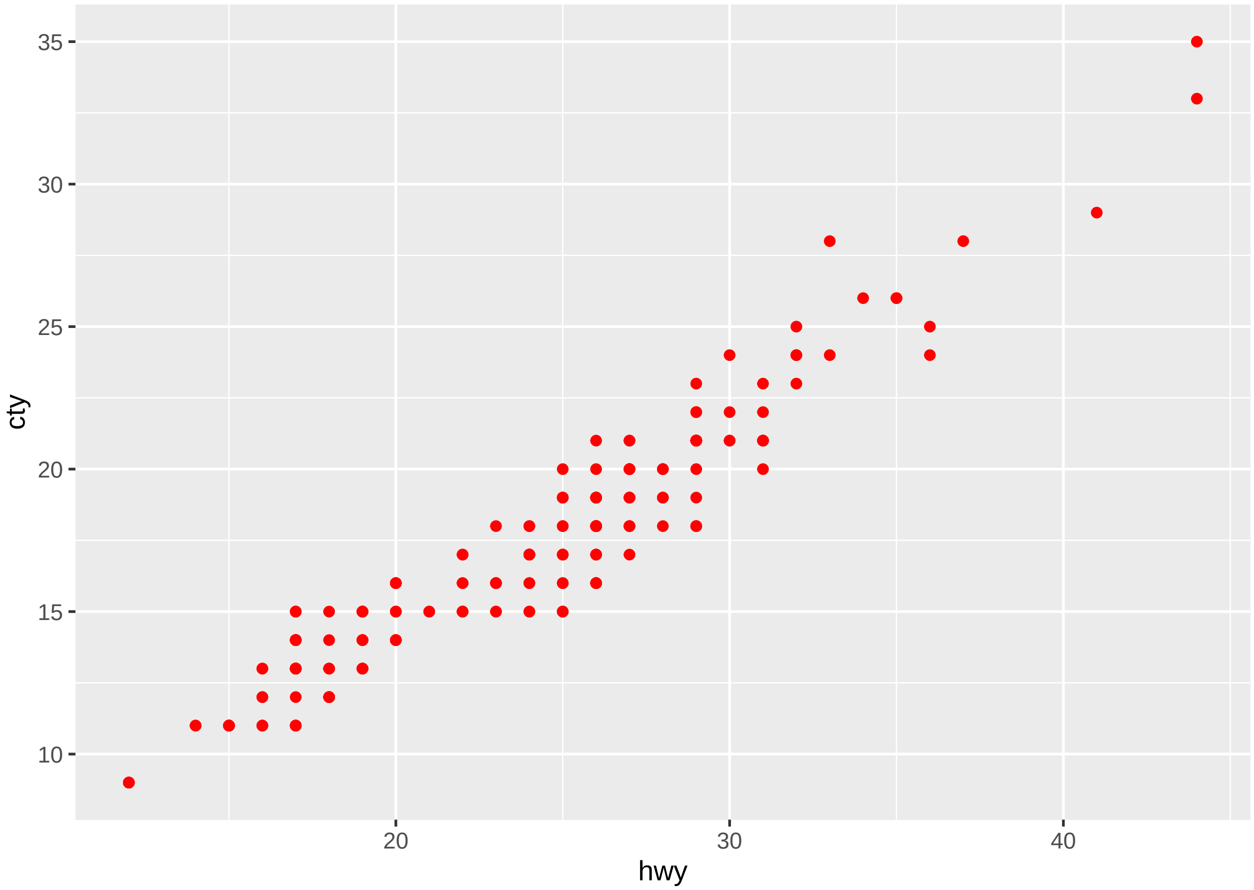
```
hist(mpg$hwy, main = paste("Histogram of", "Highway Miles per Gallon"), xlab = "Miles per Gallon (mpg)"
```



As one can see, this is a multimodal histogram with two peaks. This means most vehicles in this data set waste between 15 and 20 mpg and between 25 and 30 mpg. These two peaks mean that vehicles with lower mpg are fairly common, as well as with vehicles with higher mpg.

Exercise 2) Let us create a scatterplot with *hwy* on the x-axis and *cty* on the y-axis

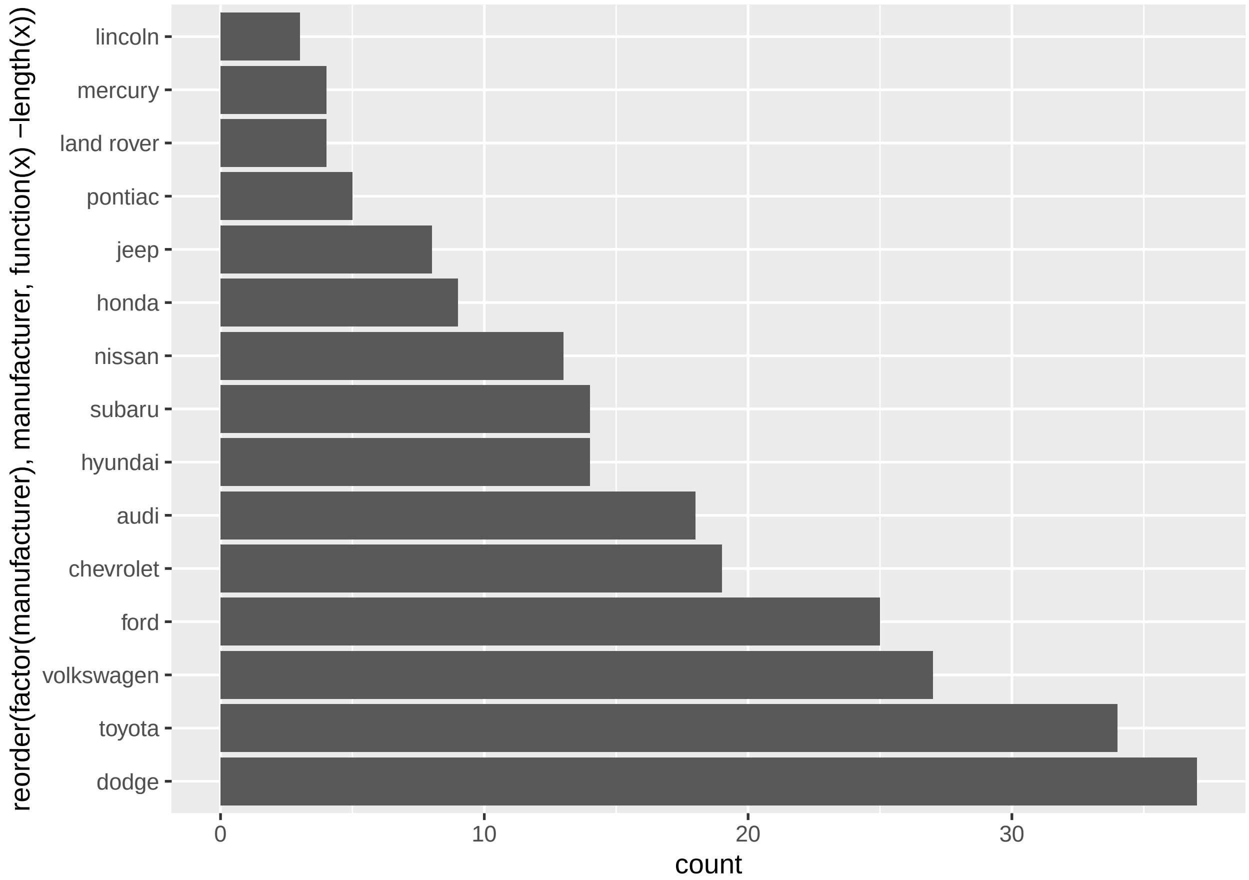
```
ggplot(data = mpg) + geom_point(mapping = aes(x = hwy, y = cty), color = "red") # Create scatterplot
```



Observe that the scatterplot has a positive correlation, meaning the *hwy* and *cty* variables have a strong relationship. This means that highway mpg tends to increase with city mpg.

Exercise 3) Let us make a bar plot.

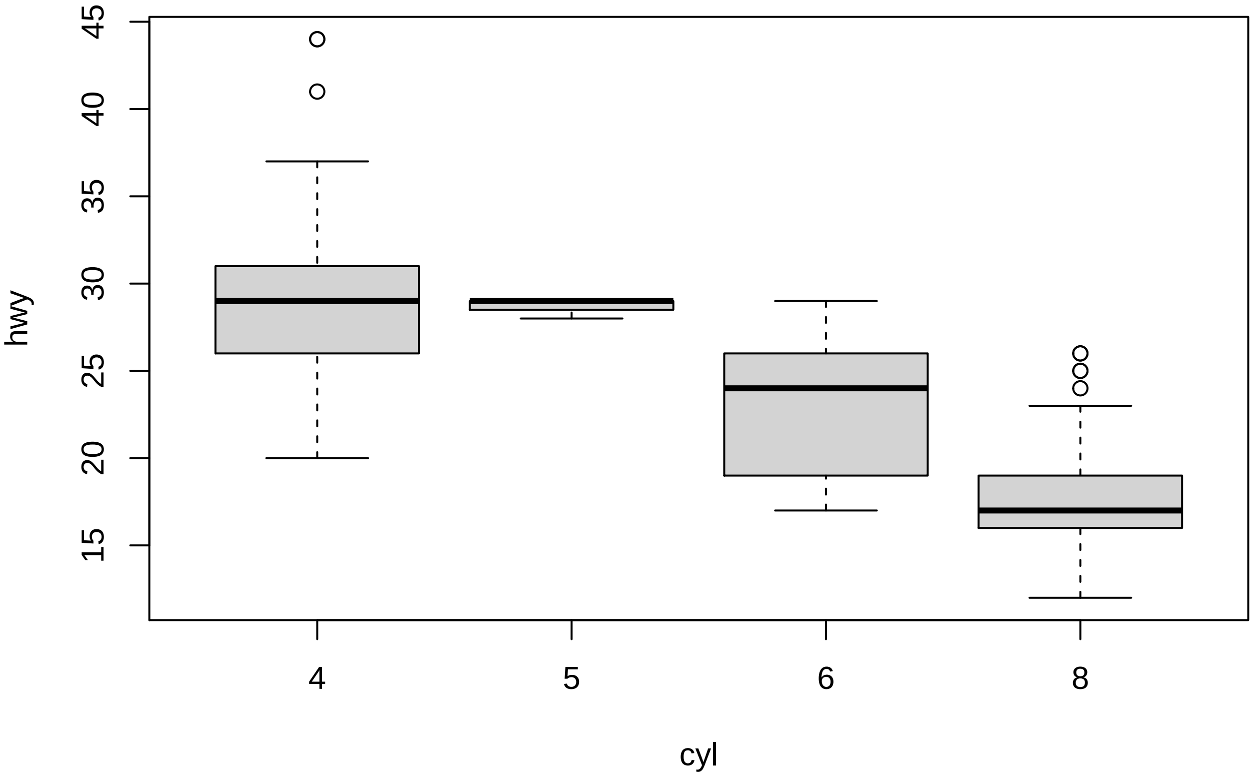
```
ggplot(data = mpg, mapping = aes(x = reorder(factor(manufacturer), manufacturer, function(x) -length(x))
```



The manufacturer Dodge produced the most number of cars. The manufacturer Lincoln produced the least number of cars. This is all based on the bar plot above.

Exercise 4) Let us make a boxplot of the *hwy* variable grouped by *cyl*.

```
boxplot(hwy ~ cyl, data = mpg)
```



Based on the boxplot, one can observe that a vehicle with more cylinders tend to have a smaller highway mileage. Furthermore, a vehicle with a smaller amount of cylinders tend to have a larger highway mileage.

Exercise 5) Let us make a correlation matrix.

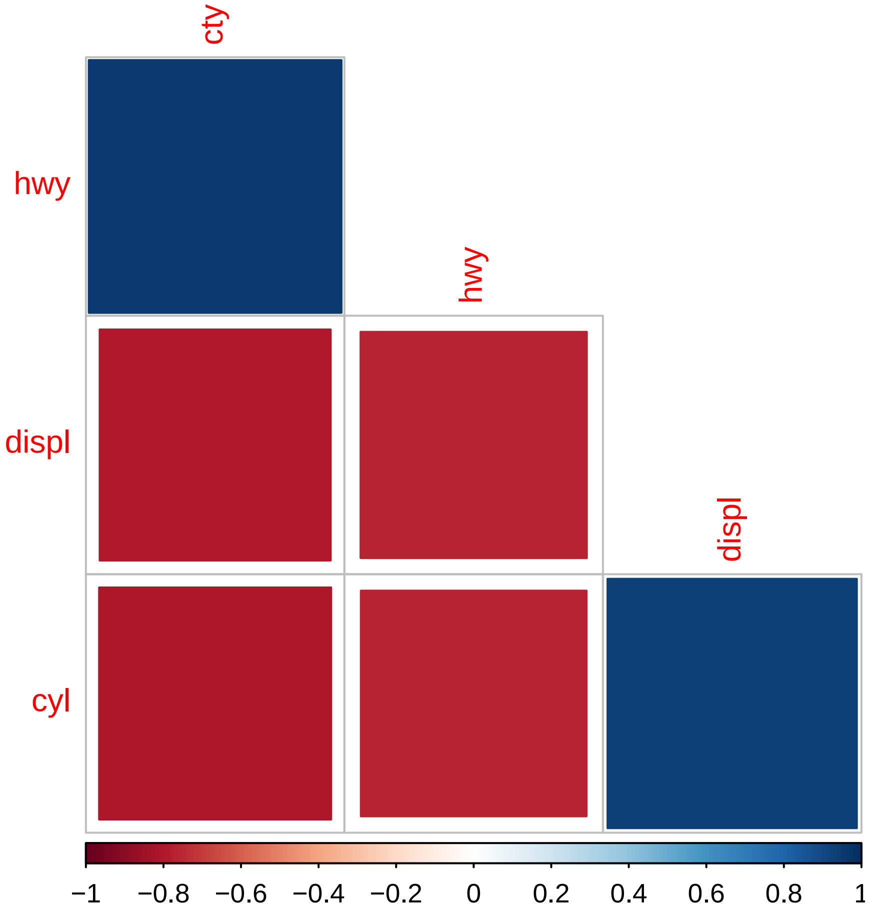
```
library(corrplot)  # Load corrplot
```

```
## corrplot 0.92 loaded
```

```
mpg2 <- mpg[, c("displ", "cyl", "cty", "hwy")]  # Only include numeric variables


M <- cor(mpg2)  # Compute variance

corrplot(M, method = 'square', order = 'FPC', type = 'lower', diag = FALSE) # Create correlation matrix
```



The blue squares indicate a positive correlation, whereas the red squares indicate a negative correlation between the variables. The *hwy* and *cty* variables have a strong positive correlation with each other, including the *cyl* and *displ* variables, which also have a strong positive correlation. On the other hand, the *displ* variable has a strong negative correlation with the *cty* and *hwy* variables, and the *cyl* variable also has a strong negative correlation with the *cty* and *hwy* variables.

The relationships do make sense. Highway and city mileage should have a positive relationship, as it wouldn't make sense otherwise. Additionally, it makes sense that that the number of cylinders has a negative correlation with highway and city mileage. The higher the amount of cylinders, the less mileage a car has (both highway and city mileage).