Lab01

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

```
# Supervised learning is a machine learning approach that's defined by its use of labele
d datasets. Supervised learning can be separated into two types of problems when data mi
ning: classification and regression.
# Unsupervised learning uses machine learning algorithms to analyze and cluster unlabele
d data sets. Unsupervised learning models are used for three main tasks: clustering, ass
ociation and dimensionality reduction.
# The difference between superviced learning and unsuperviced learning is that supervise
d learning uses labeled input and output data, while an unsupervised learning algorithm
 does not.
#Source: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning
```

```
# Explain the difference between a regression model and a classification model, specific
ally in the context of machine learning.
# The main difference between regression model and classification model is the data type
of Y.
# In a regression model, Y is quantitative and numerical values such as price and blood
 preesure.
# In a classification model. Y is a qualitative and categorical values such as married/
 not married.
```

```
# Name two commonly used metrics for regression ML problems. Name two commonly used metr
ics for classification ML problems.
# Not covered yet
```

```
# As discussed, statistical models can be used for different purposes. These purposes ca
n generally be classified into the following three categories. Provide a brief descripti
on of each.

# Descriptive models: Choose model to best visually emphasize a trend in data

# Inferential models: often used to compare the differences between the treatment groups
and aim is to testing theories, causal claims, stating relationship between outcome and
 predictor

# Predictive models: predict future behavior and aim is to predict Y with minimum reduci
ble error

# Source: lecture2 power point
```

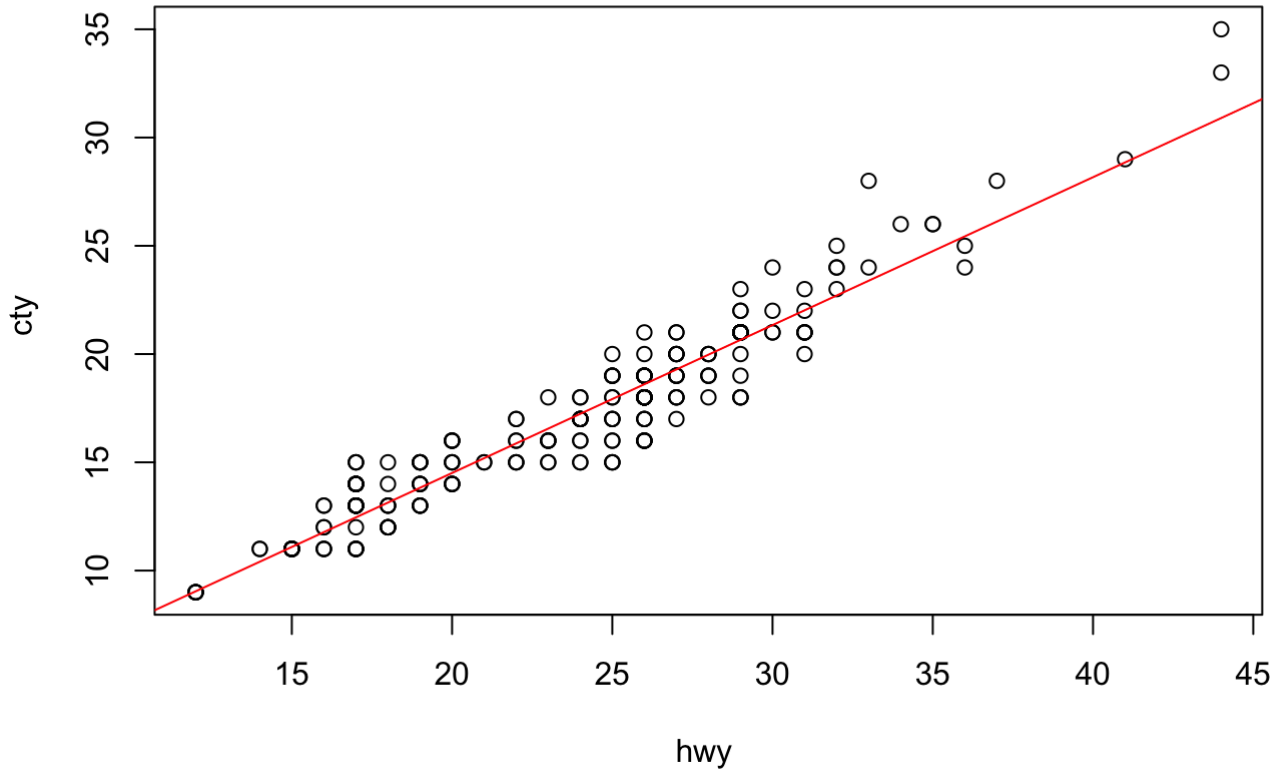
```
#Predictive models are frequently used in machine learning, and they can usually be desc
ribed as either mechanistic or empirically-driven. Answer the following questions.

#Define mechanistic. Define empirically-driven. How do these model types differ? How are
they similar?

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

#Describe how the bias-variance tradeoff is related to the use of mechanistic or empiric
ally-driven models.

#Not covered yet
```

```
# A political candidate’s campaign has collected some detailed voter history data from t
heir 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 can
didate?

# Ans: This is an inferential model becuase the campaign tries to state the relationship
between the predictor and outcomes.

# How would a voter’s likelihood of support for the candidate change if they had persona
l contact with the candidate?

# Ans: This is a predictive model because the campaign wants to know about the voter's f
uture behavior after they had personal contact with the candidate with a minimum reducib
le error.

# Classify each question as either predictive or inferential. Explain your reasoning for
each.
```

```
# We are interested in highway miles per gallon, or the hwy variable. Create a histogram
of this variable. Describe what you see/learn.\
library(ggplot2)
data("mpg")
hwy <- mpg$hwy
hist(hwy)
```

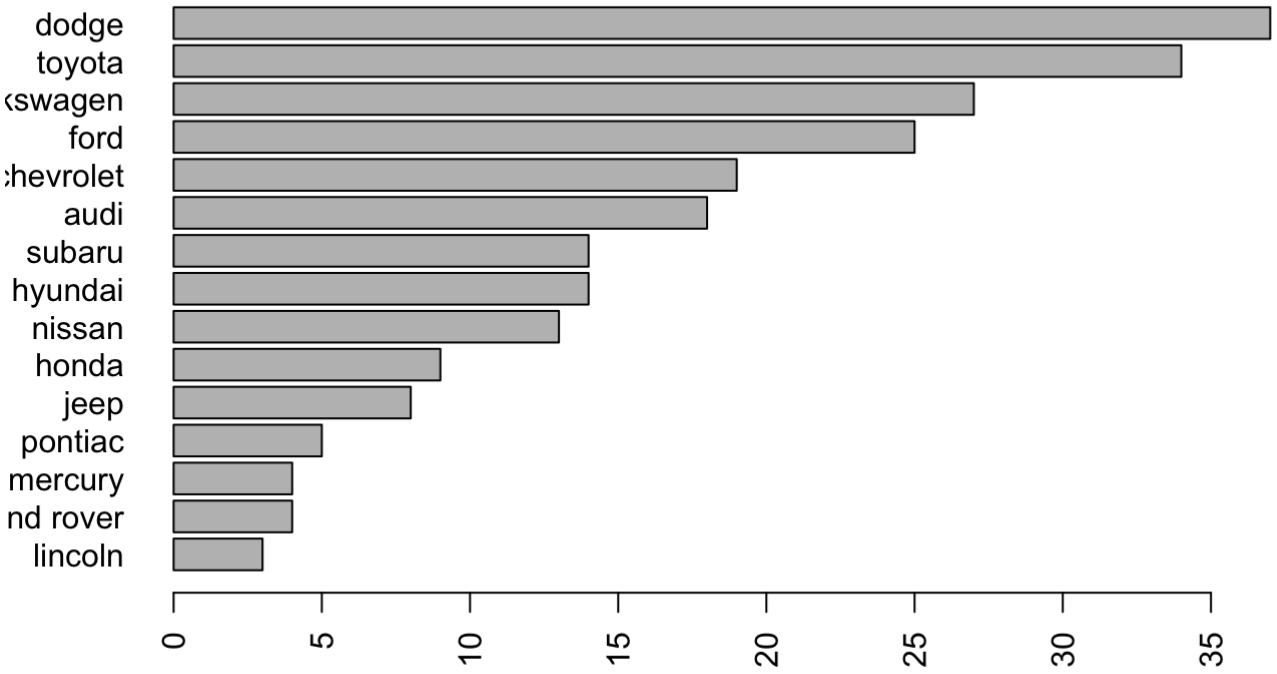

```
# I learned that the histogram is right-skwed and the maximum is 45. The median number i
s between 20-25.
```

```
# Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you n
otice. Is there a relationship between hwy and cty? What does this mean?
plot(mpg$hwy, mpg$cty,
   xlab="hwy ", ylab="cty ")
abline(lm(mpg$cty~mpg$hwy), col="red")
```




```
# I noticed that there is a positive linear relationship between hwy and cty.
# This means that cty increases linearly while hwy increases.
```

```
# 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?
counts <- table(mpg$manufacturer)
barplot(sort(counts), main="Car manufacture Distribution", horiz=TRUE, las =2)
```

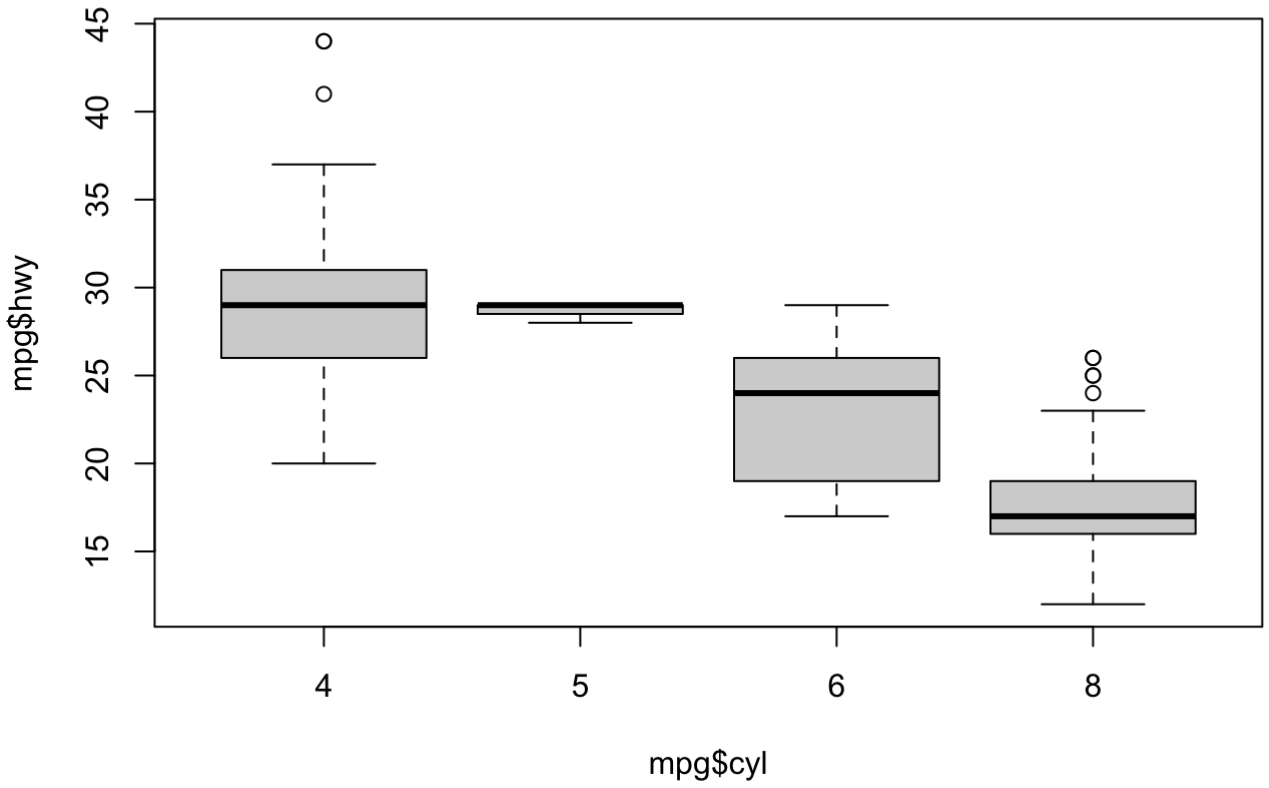



```
sort(counts)
```

```
##
##    lincoln land rover    mercury    pontiac       jeep      honda     nissan
##             3          4          4          5          8          9         13
##    hyundai     subaru       audi  chevrolet       ford volkswagen     toyota
##         14         14         18         19         25         27         34
##      dodge
##         37
```

```
# Based on the graph, lincoln produce the least and dodge produced the most.
```

```
# Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?
boxplot(mpg$hwy ~ mpg$cyl)
```




```
#I can see the pattern: with the increase of cyl, the mean of hwy is decreasing and the
 range of each cyl group is also moving donward on the y-axis. There is a negative relat
ionship between hwy and cyl.
```

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

# Which variables are positively or negatively correlated with which others? Do these re
lationships make sense to you? Are there any that surprise you?
library(corrplot)
```

```
## corrplot 0.92 loaded
```

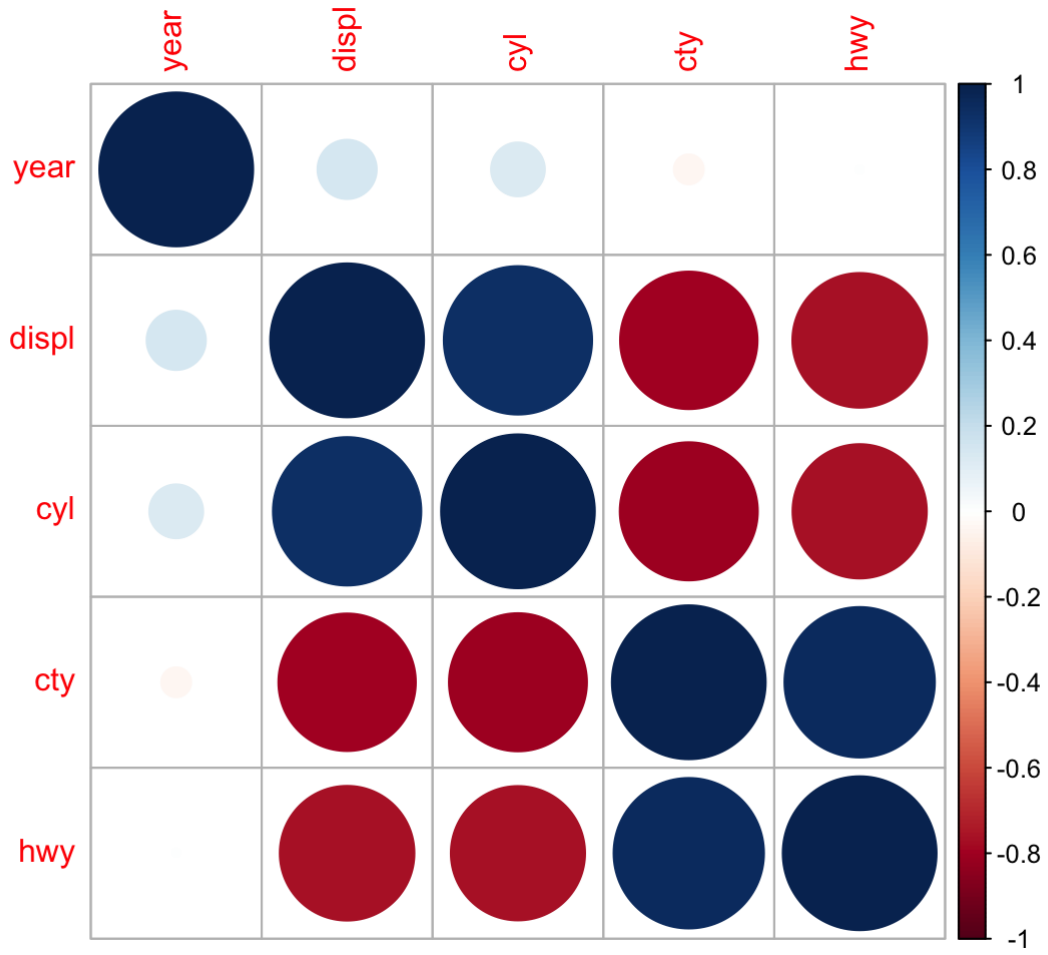
```
library(magrittr)
library(dplyr)
```

```
##
## 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
```

```
mpg_modified <- mpg %>% select(year, displ, cyl, cty, cyl, hwy) #by modifing the table,
 I only take the numerical values to do a corrplot, we can also try to transform categor
ical data to numerical data like 1 and 0
M = cor(mpg_modified)
corrplot(M)
```



#Based on the graph, I can see that displ and cty has a negative correlation, cty and cy l has a negative correlation. cyl and year has a weak position correlation. displ and ye ar has a weak positive correlation. cty and year does not have a mentionable correlatio n. hwy has a negative correlation with displ. cyl has a negative correlation with hwy. c ty has a positive relationship with hwy. These correlations make sense to me. If a car h as a low cty, then it is reasonable that it has a low hwy as well. Their relationship is positive for sure. I do not have any surprise since car pollution is not a new topic and we covered it in previous pstat class.