HW1

Rohan Dalal

April 2, 2022

Question 1:

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

Answer1:

Supervised learning: machine learning method in which models are trained using labeled input data Unsupervised learning: machinelearning method in which patterns inferred from the unlabeled input data Main difference between SUpervised and Unsupervides learning are: Supervised learning used to train the model so that it can predict the output when it is given new labeled data whereas The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown/unlabeled dataset.

Question 2:

Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

Answer2

Main difference between regression and classification model is that regression models are quantitative helps predicting a continuous quantity, where as classification models are qualitative predicts discrete class labels.

Question3:

Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

Answer3 2 main metrics for Regression problems are R Square/Adjusted R Square and Mean Square Error(MSE)/Root Mean Square Error(RMSE). 2 main metrics for classification problems are Accuracy and confusion matrix

Question4:

Describe : Descriptive models, Inferential models and Predictive models.

1) Descriptive models: models that describes real-world events and the relationships between factors responsible for them. Groups are clustered according to the descriptive factors.These models visually emphasize a trend in data using a line on a scatterplot
2) Inferential models: model that processes inferring properties on a population based on the properties of a sample of a population. Model allows to test a hypothesis or assess whether given data is generalizable to the broader population.
3) Predictive models: modles that predicts and forecast likely future outcomes with the help of historical and existing data. Model is built up by the number of predictors that are highly favorable to determine future decisions.Goal is to predict with minimum reducible error.

Question5

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

Mechanisitc: Parametric Model uses a theory to predict what will happen in the real world. Mechanistic models are useful if you have good data for making predictions. Empirically-driven: Non Parametric model does prediction by experimenting.Gather data, learn and train to predict.Empirical modeling is valuable when you're trying something new ##In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice. Empirical models are focused on describing the data with the specification of very few assumptions about the data being analyzed whereas Mechanistic models specify assumptions and attempt to incorporate known factors about the systems surrounding the data into the model, while describing the available data

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

Bias-variance of the parameter across models can be diminished by expanding the bias in the estimated parameters. By adding more info components/factors will assist with working on the data to fit better.

Question5

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?

a) Predictive: We are trying to predict/forecast based on histroical data of voter's profile hence it is a predictive model.
b) Inferential: Since we are trying ot process inferring properties if voter had personal contact with candidate.

```
data(mpg)
```

```
## Warning in data(mpg): data set 'mpg' not found
```

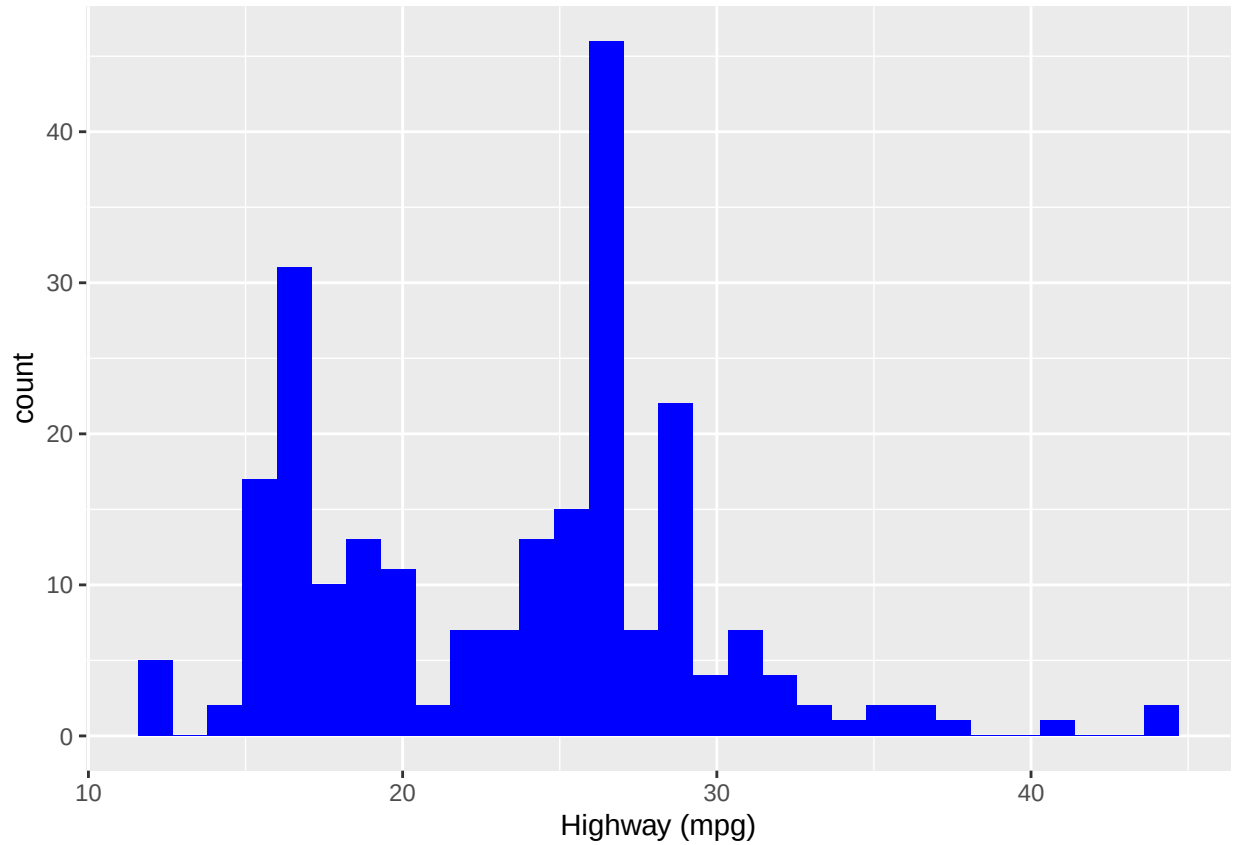
Exercise 1:

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)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
ggplot(data = mpg, aes(x = hwy)) + geom_histogram( fill = "blue", bins = 30) +  xlab("Highway (mpg)")
```

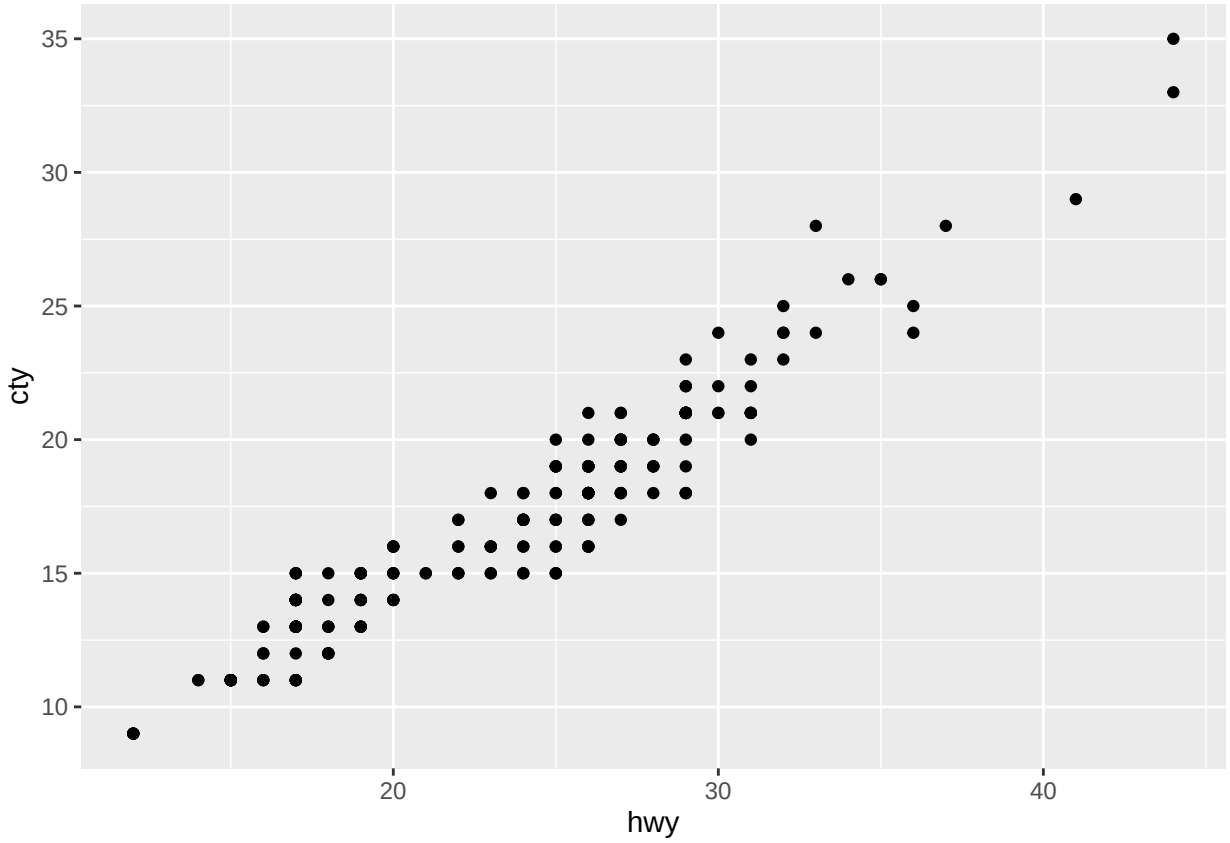


From the above histograms, we can see that "highway miles per gallon" data has a roughly normal distribution but a few observations "outliers" lie far from the other observations in the graph.

Exercise2:

Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?

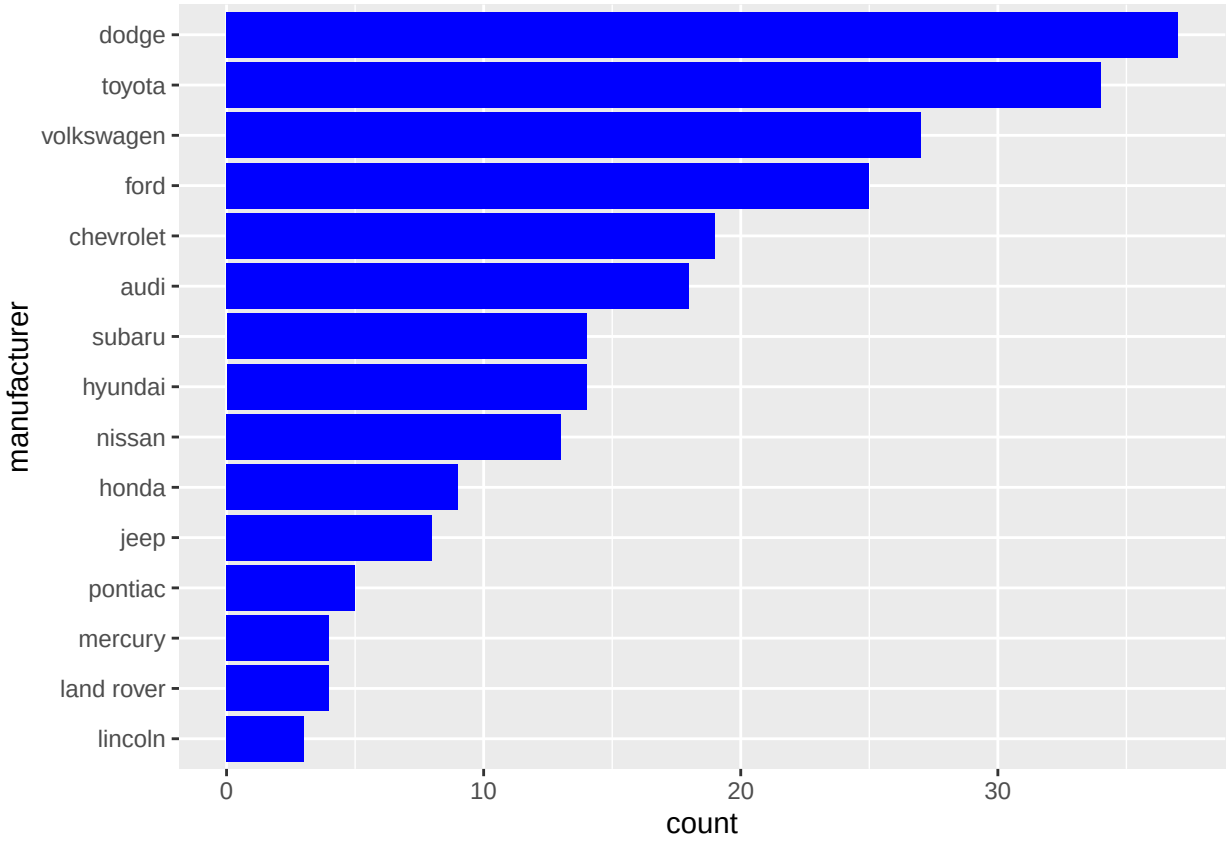
```
ggplot(mpg, aes(hwy, cty)) +
  geom_point()
```



Exercise 3:

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?

```
mpg <- within(mpg, manufacturer <-factor(manufacturer,levels=names(sort(table(manufacturer),
                                                    decreasing=FALSE))))
ggplot(mpg, aes(y=manufacturer)) + geom_bar(fill="blue")
```

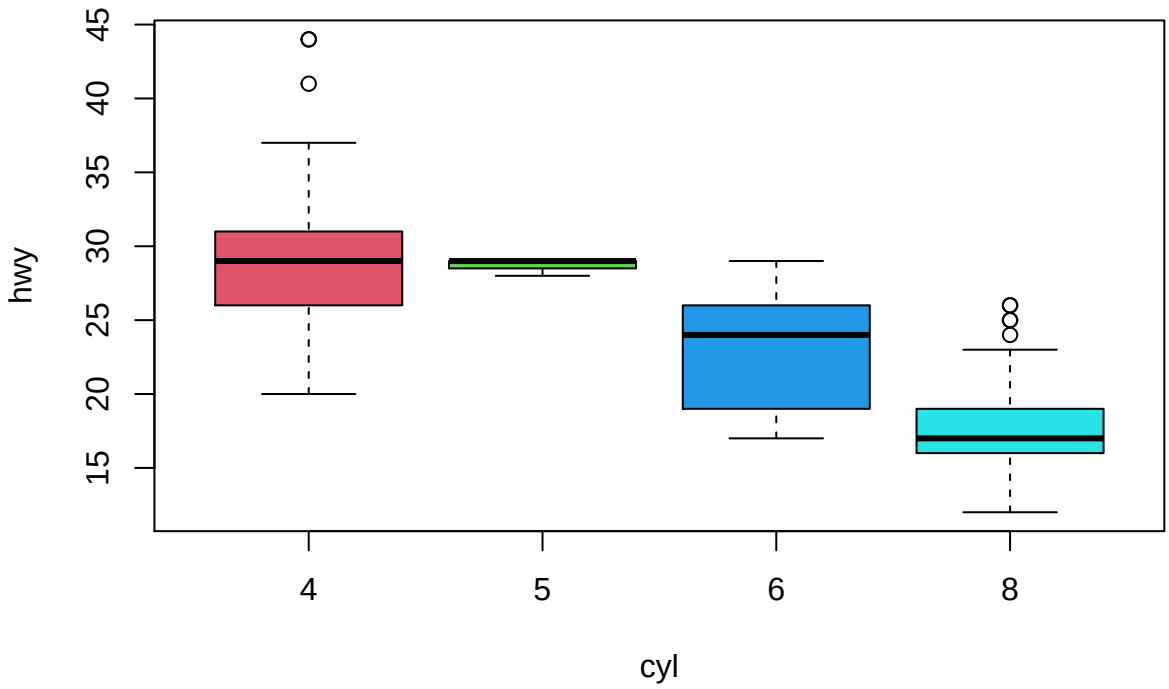


manufacturer produced the most cars: DODGE manufacturer produced the least cars: LINCOLN

Exercise 4:

Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

```
boxplot(data = mpg,hwy ~ cyl, col = c(2,3,4,5))
```



Pattern: High number of cylinder gives LOW mileage on HWY where as Low number of cylinder gives HIGH mileage on HWY

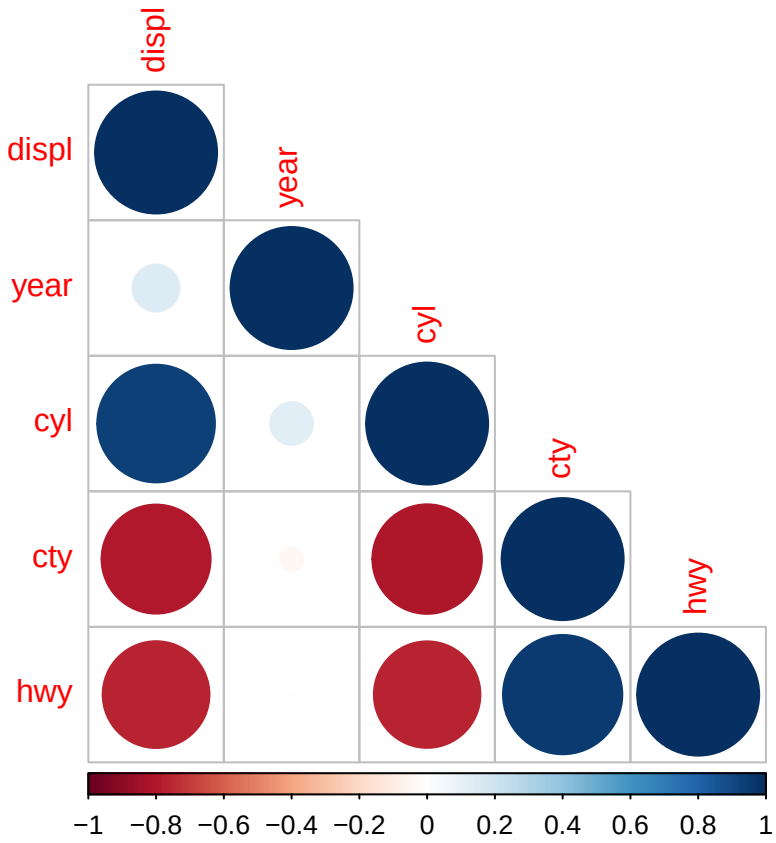
Exercise 5:

Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset.

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
data<-cor(mpg[sapply(mpg,is.numeric)])
corrplot(data, type="lower")
```



If the correlation coefficient is greater than zero, it is a positive relationship. Conversely, if the value is less than zero, it is a negative relationship and a value of zero indicates no relationship.Referring to the above diagram x axis is showing the correlation coefficient.