**Figure 1**  
*Load as Data Frame and Head Function*

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.2 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(rpart)  
library(rpart.plot)  
mod3cars.df <- read.csv("ToyotaCorolla2.csv", header = TRUE)  
head(mod3cars.df)

## Id Model Price Age\_08\_04 Mfg\_Month  
## 1 1 TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13500 23 10  
## 2 2 TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13750 23 10  
## 3 3 TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 13950 24 9  
## 4 4 TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors 14950 26 7  
## 5 5 TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors 13750 30 3  
## 6 6 TOYOTA Corolla 2.0 D4D HATCHB SOL 2/3-Doors 12950 32 1  
## Mfg\_Year KM Fuel\_Type HP Met\_Color Color Automatic CC Doors Cylinders  
## 1 2002 46986 Diesel 90 1 Blue 0 2000 3 4  
## 2 2002 72937 Diesel 90 1 Silver 0 2000 3 4  
## 3 2002 41711 Diesel 90 1 Blue 0 2000 3 4  
## 4 2002 48000 Diesel 90 0 Black 0 2000 3 4  
## 5 2002 38500 Diesel 90 0 Black 0 2000 3 4  
## 6 2002 61000 Diesel 90 0 White 0 2000 3 4  
## Gears Quarterly\_Tax Weight Mfr\_Guarantee BOVAG\_Guarantee Guarantee\_Period ABS  
## 1 5 210 1165 0 1 3 1  
## 2 5 210 1165 0 1 3 1  
## 3 5 210 1165 1 1 3 1  
## 4 5 210 1165 1 1 3 1  
## 5 5 210 1170 1 1 3 1  
## 6 5 210 1170 0 1 3 1  
## Airbag\_1 Airbag\_2 Airco Automatic\_airco Boardcomputer CD\_Player Central\_Lock  
## 1 1 1 0 0 1 0 1  
## 2 1 1 1 0 1 1 1  
## 3 1 1 0 0 1 0 0  
## 4 1 1 0 0 1 0 0  
## 5 1 1 1 0 1 0 1  
## 6 1 1 1 0 1 0 1  
## Powered\_Windows Power\_Steering Radio Mistlamps Sport\_Model Backseat\_Divider  
## 1 1 1 0 0 0 1  
## 2 0 1 0 0 0 1  
## 3 0 1 0 0 0 1  
## 4 0 1 0 0 0 1  
## 5 1 1 0 1 0 1  
## 6 1 1 0 1 0 1  
## Metallic\_Rim Radio\_cassette Parking\_Assistant Tow\_Bar  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 3 0 0 0 0  
## 4 0 0 0 0  
## 5 0 0 0 0  
## 6 0 0 0 0

**Figure 2**  
*Data Frame Dimensions*

dim(mod3cars.df)

## [1] 1436 39

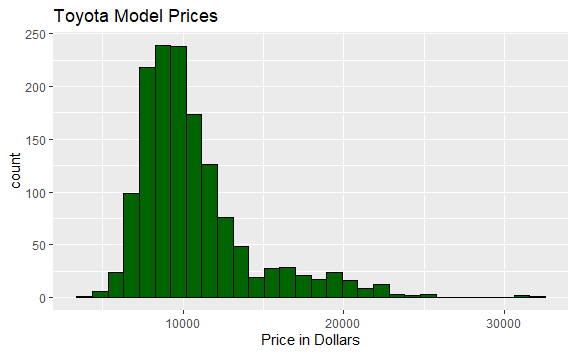
**Figure 3**  
*Attibute List with Column Numbers*

t(t(names(mod3cars.df)))

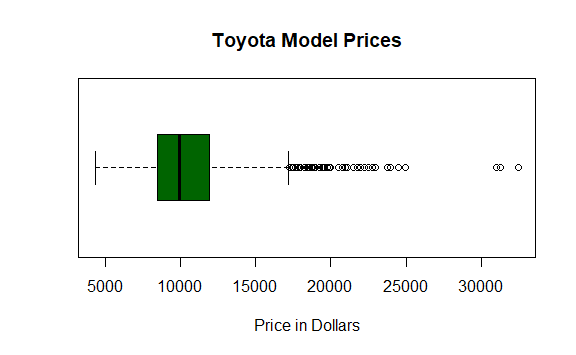
## [,1]   
## [1,] "Id"   
## [2,] "Model"   
## [3,] "Price"   
## [4,] "Age\_08\_04"   
## [5,] "Mfg\_Month"   
## [6,] "Mfg\_Year"   
## [7,] "KM"   
## [8,] "Fuel\_Type"   
## [9,] "HP"   
## [10,] "Met\_Color"   
## [11,] "Color"   
## [12,] "Automatic"   
## [13,] "CC"   
## [14,] "Doors"   
## [15,] "Cylinders"   
## [16,] "Gears"   
## [17,] "Quarterly\_Tax"   
## [18,] "Weight"   
## [19,] "Mfr\_Guarantee"   
## [20,] "BOVAG\_Guarantee"   
## [21,] "Guarantee\_Period"   
## [22,] "ABS"   
## [23,] "Airbag\_1"   
## [24,] "Airbag\_2"   
## [25,] "Airco"   
## [26,] "Automatic\_airco"   
## [27,] "Boardcomputer"   
## [28,] "CD\_Player"   
## [29,] "Central\_Lock"   
## [30,] "Powered\_Windows"   
## [31,] "Power\_Steering"   
## [32,] "Radio"   
## [33,] "Mistlamps"   
## [34,] "Sport\_Model"   
## [35,] "Backseat\_Divider"   
## [36,] "Metallic\_Rim"   
## [37,] "Radio\_cassette"   
## [38,] "Parking\_Assistant"  
## [39,] "Tow\_Bar"

**Figure 4**  
*Histogram and Boxplot of Prices*

ggplot(mod3cars.df) + geom\_histogram(mapping = aes(x = Price), bins = 30,  
 col = "black", fill = "dark green") +  
 xlab("Price in Dollars") +  
 ggtitle("Toyota Model Prices")



boxplot(mod3cars.df$Price, main = "Toyota Model Prices",  
 horizontal = TRUE, col = "dark green", xlab = "Price in Dollars")



**Figure 5**  
*Dimension Reduction*

mod3cars2.df <- select(mod3cars.df, Price, Age\_08\_04, KM, Fuel\_Type, HP,   
 Automatic, Doors, Quarterly\_Tax, Mfr\_Guarantee,   
 Guarantee\_Period, Airco, Automatic\_airco, CD\_Player,   
 Powered\_Windows, Sport\_Model, Tow\_Bar)  
head(mod3cars2.df)

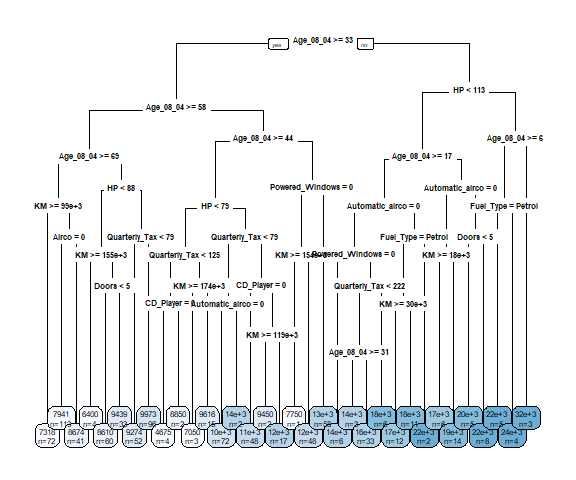
## Price Age\_08\_04 KM Fuel\_Type HP Automatic Doors Quarterly\_Tax  
## 1 13500 23 46986 Diesel 90 0 3 210  
## 2 13750 23 72937 Diesel 90 0 3 210  
## 3 13950 24 41711 Diesel 90 0 3 210  
## 4 14950 26 48000 Diesel 90 0 3 210  
## 5 13750 30 38500 Diesel 90 0 3 210  
## 6 12950 32 61000 Diesel 90 0 3 210  
## Mfr\_Guarantee Guarantee\_Period Airco Automatic\_airco CD\_Player  
## 1 0 3 0 0 0  
## 2 0 3 1 0 1  
## 3 1 3 0 0 0  
## 4 1 3 0 0 0  
## 5 1 3 1 0 0  
## 6 0 3 1 0 0  
## Powered\_Windows Sport\_Model Tow\_Bar  
## 1 1 0 0  
## 2 0 0 0  
## 3 0 0 0  
## 4 0 0 0  
## 5 1 0 0  
## 6 1 0 0

**Figure 6**  
*Regression Decision Tree*

set.seed(1)  
train.index <- sample(c(1:dim(mod3cars2.df)[1]), dim(mod3cars2.df)[1]\*0.6)  
mod3train.df <- mod3cars2.df[train.index, ]  
mod3valid.df <- mod3cars2.df[-train.index, ]  
mod3cars.dt <- rpart(Price ~., data = mod3train.df, method = "anova",   
 cp = 0.001, minbucket = 1, maxdepth = 30)

**Figure 7**  
*Decision Tree Plot*

rpart.plot(mod3cars.dt, type = 0, extra = 1, varlen = 0, cex = 0.5)



## Description of Steps Taken and Lessons Learned

To begin, multiple data exploration functions will be performed. After loading the appropriate libraries, the *csv* file is read and saved to the *mod3cars* data frame and the head(mod3cars) function is run to display as a sample of the data frame (See Figure 1). The data frame attributes are displayed in a flat-file format with the first 6 rows of data. Next, Figure 2 displays the dimensions of the data frame using the function dim(mod3cars.df) which is 1,436 rows and 39 columns. In Figure 3, the code t(t(names(mod3cars))) will display the data frame attributes in a list along with their respective column numbers. Visualizations to examine the Price attribute are added (See Figure 4). The histogram and box plot for Price shows data is normally distributed and has a right or positive skew. Since the assignment already provides instructions to limit the attributes for the decision tree, dimension reduction will be done next (See Figure 5). The select function will be used to create a new data frame *mod3cars2* with the reduced attributes. The head(mod3cars2) function is added as a means to verify the list of attributes in the new data frame is correct.

After data exploration, a regression tree will be created using the parameters from the assignment and then the created tree will be plotted. Figure 6 displays the code for creating a training data frame which is 60% of the mod3cars2 data frame. The regression tree is created using the ‘rpart’ function. The method of *anova* specifies the decision tree is a regression and not a classification tree. Two functions, prp or rpart.plot may be used to plot the tree. Rpart.plot was chosen as the display as it is easier to read for this example (See Figure 7).

A significant lesson learned in this exercise was the actual code for the decision tree. Both classification and regression trees use the same function, however, it is the *method* parameter that determines the type of tree created. Multiple questions arise with the use of *RMarkdown*. Mainly, the R package has several formatting limitations which conflict with APA requirements. Future consideration for the assignment instructions may be to clarify which directive is a higher priority. The two plot functions for displaying decision trees are frustrating. It is hard to tweak either plot function to assure legibility in larger trees and to avoid overlap in node labels. Centering plot images are also a challenge. Using fig.align = "center" in the code chunk does not appear to be supported.

## References

RStudio Team. (2021). RStudio: Integrated development environment for R. RStudio, PBC, Boston, MA. <http://www.rstudio.com/>.