Final Project

Abhishek Soalnki

10/16/2020

Introduction

This dataset provides fuel economy data from 2010 to 2012, 2014 to 2016, and 2018 to 2020 for popular models of cars.

```
library(tidyverse)
library(Hmisc)
library(funModeling)

setwd("C:/Users/abhis/Documents/Final case")

df2010 <- read.csv(file = "C:/Users/abhis/Documents/Final case/data2010.csv",
header = TRUE)
df2011 <- read.csv(file = "C:/Users/abhis/Documents/Final case/data2011.csv",
header = TRUE)
df2012 <- read.csv(file = "C:/Users/abhis/Documents/Final case/data2012.csv",
header = TRUE)

df2014 <- read.csv(file = "C:/Users/abhis/Documents/Final case/data2014.csv",
header = TRUE)
df2015 <- read.csv(file = "C:/Users/abhis/Documents/Final case/data2015.csv",
header = TRUE)
df2016 <- read.csv(file = "C:/Users/abhis/Documents/Final case/data2016.csv",
header = TRUE)

df2018 <- read.csv(file = "C:/Users/abhis/Documents/Final case/data2018.csv",
header = TRUE)
df2019 <- read.csv(file = "C:/Users/abhis/Documents/Final case/data2019.csv",
header = TRUE)
df2020 <- read.csv(file = "C:/Users/abhis/Documents/Final case/data2020.csv",
header = TRUE)
```

Merging and cleaning data

#The data is merged into three large files. First, 2012 to 2012; Second, 2014 to 2016; Third, 2018 to 2020.

```
df2010to12 <- rbind(df2010,df2011,df2012)
df2014to16 <- rbind(df2014,df2015,df2016)
df2018to20 <- rbind(df2018,df2019,df2020)
```

Cleaning the merged data

Once the data is merged. It must be cleaned to avoid outliers and bad data. Data Cleaning is the process of transforming raw data into consistent data that can be analyzed. It is aimed at improving the content of statistical statements based on the data as well as their reliability. Data cleaning may profoundly influence the statistical statements based on the data.

Data cleaning process was accomplished in three steps i.e. 1. Initial exploratory analysis 2. Visualazation exploration 3. NA cleaning

Initial exploratory analysis

The first thing that I did is check the class of the data frame:

```
class(df2010to12)

## [1] "data.frame"

class(df2014to16)

## [1] "data.frame"

class(df2018to20)

## [1] "data.frame"
```

Next, the number of columns and rows were checked for each dataframe

```
dim(df2010to12)

## [1] 11183    42

dim(df2014to16)

## [1] 14246    42

dim(df2018to20)

## [1] 13641    42
```

Finally, the summary of the data was analyzed

```
summary(df2010to12)

##        X              Year       Veh.Mfr.Code  Represented.Test.Veh.Make
##  Min.   :   0.0   Min.   :2010   NSX    :1602   NISSAN       :1221
##  1st Qu.: 931.5   1st Qu.:2010   GMX    :1192   BMW          : 904
##  Median :1863.0   Median :2011   TYX    :1118   CHEVROLET    : 823
##  Mean   :1880.2   Mean   :2011   BMX    :1083   TOYOTA       : 758
##  3rd Qu.:2795.0   3rd Qu.:2012   FMX    : 941   AUDI         : 708
##  Max.   :4143.0   Max.   :2012   ADX    : 791   Mercedes-Benz: 539
```

```
##                                  (Other):4456   (Other)      :6230
##                    Represented.Test.Veh.Model Test.Veh.Displacement..L.
##  Jetta                            :  208        Min.   :  0.001
##  R8                               :  155        1st Qu.:  2.400
##  TITAN KING-5.6LE SWB             :  139        Median :  3.456
##  328i                             :  126        Mean   :  3.535
##  NISSAN FRONTIER KING CAB SE 4X4:  121        3rd Qu.:  4.293
##  A3                               :  118        Max.   : 99.999
##  (Other)                          :10316
##  Vehicle.Type Rated.Horsepower X..of.Cylinders.and.Rotors  Engine.Code
##  Both :1338   Min.   :   1.0   Min.   : 3.000             1      : 564
##  Car  :6234   1st Qu.: 177.0   1st Qu.: 4.000             2      : 338
##  Truck:3611   Median : 261.0   Median : 6.000             CCTA   : 194
##               Mean   : 267.9   Mean   : 5.909             3      : 147
##               3rd Qu.: 317.0   3rd Qu.: 8.000             CBFA   : 147
##               Max.   :1200.0   Max.   :16.000             CJAA   : 138
##                                NA's   :1635               (Other):9655
##           Tested.Transmission.Type  X..of.Gears    Transmission.Lockup.
##  Automatic             :5638        Min.   :1.000   N:3570
##  Manual                :2396        1st Qu.:5.000   Y:7613
##  Semi-Automatic        :1946        Median :6.000
##  Continuously Variable: 948         Mean   :5.391
##  Automated Manual      : 141        3rd Qu.:6.000
##  Other                 :  64        Max.   :8.000
##  (Other)               :  50
##              Drive.System.Description      Transmission.Overdrive.Desc
##  2-Wheel Drive, Front   :4904        No gear ratio < 1  :  174
##  2-Wheel Drive, Rear    :4326        Top gear ration < 1:11009
##  4-Wheel Drive          :1082
##  All Wheel Drive        : 580
##  Part-time 4-Wheel Drive: 291
##
##
##  Equivalent.Test.Weight..lbs..   Axle.Ratio      N.V.Ratio
##  Min.   :2125                  Min.   :1.000   Min.   :  0.00
##  1st Qu.:3625                  1st Qu.:3.160   1st Qu.: 27.90
##  Median :4000                  Median :3.500   Median : 31.10
##  Mean   :4295                  Mean   :3.594   Mean   : 32.54
##  3rd Qu.:4750                  3rd Qu.:3.910   3rd Qu.: 35.80
##  Max.   :8500                  Max.   :9.730   Max.   :999.90
##
##                       Shift.Indicator.Light.Use.Desc
##  Equipped, not shifted by SIL        :  197
##  Equipped, shifted by SIL            :   28
##  Equipped, shifted by survey schedule:  105
##  Not eqipped                         :10853
##
##
##
##                             Test.Procedure.Description Test.Fuel.Type.Cd
```

```
##  HWFE                                     :4932         Min.   : 6.00
##  Federal fuel 2-day exhaust (w/can load):3517         1st Qu.:61.00
##  Federal fuel 3-day exhaust               :1016         Median :61.00
##  US06                                     : 495         Mean   :54.73
##  SC03                                     : 394         3rd Qu.:61.00
##  Cold CO                                  : 367         Max.   :62.00
##  (Other)                                  : 462
##                                  Test.Fuel.Type.Description Test.Category
##  Tier 2 Cert Gasoline                          :9032         FTP :5322
##  E85 (85% Ethanol 15% EPA Unleaded Gasoline): 686         HWY :4932
##  CARB Phase II Gasoline                        : 624         SC03: 394
##  Federal Cert Diesel 7-15 PPM Sulfur           : 423         US06: 495
##  Cold CO Premium (Tier 2)                      : 234         CD  :  40
##  Cold CO Regular (Tier 2)                      :  52
##  (Other)                                       : 132
##    THC..g.mi.       CO..g.mi.        CO2..g.mi.       NOx..g.mi.
##  Min.   :0.0000   Min.   :0.0000   Min.   : 117.8   Min.   :0.0000
##  1st Qu.:0.0043   1st Qu.:0.0929   1st Qu.: 249.3   1st Qu.:0.0037
##  Median :0.0152   Median :0.2139   Median : 325.7   Median :0.0085
##  Mean   :0.0280   Mean   :0.3878   Mean   : 342.5   Mean   :0.0130
##  3rd Qu.:0.0297   3rd Qu.:0.4800   3rd Qu.: 412.0   3rd Qu.:0.0160
##  Max.   :0.9320   Max.   :7.4165   Max.   :1012.0   Max.   :0.9100
##  NA's   :865      NA's   :831      NA's   :832      NA's   :914
##    CH4..g.mi.      N2O..g.mi.     RND_ADJ_FE        FE.Bag.1
##  Min.   :0.000   Min.   :0.00   Min.   :  7.90   Min.   : 7.278
##  1st Qu.:0.002   1st Qu.:0.01   1st Qu.: 21.00   1st Qu.:16.805
##  Median :0.004   Median :0.01   Median : 26.70   Median :20.600
##  Mean   :0.007   Mean   :0.01   Mean   : 28.91   Mean   :21.960
##  3rd Qu.:0.007   3rd Qu.:0.01   3rd Qu.: 35.50   3rd Qu.:25.500
##  Max.   :0.180   Max.   :0.01   Max.   :268.40   Max.   :63.002
##  NA's   :5383    NA's   :8330   NA's   :73       NA's   :8076
##     FE.Bag.2         FE.Bag.3      Target.Coef.A..lbf.
Target.Coef.B..lbf.mph.
##  Min.   :  8.00   Min.   : 8.093   Min.   : 1.843      Min.   :-0.8473
##  1st Qu.: 18.70   1st Qu.:21.578   1st Qu.:30.360      1st Qu.: 0.1500
##  Median : 22.72   Median :25.600   Median :36.700      Median : 0.3300
##  Mean   : 25.12   Mean   :26.846   Mean   :37.531      Mean   : 0.3637
##  3rd Qu.: 28.20   3rd Qu.:30.588   3rd Qu.:43.388      3rd Qu.: 0.5644
##  Max.   :116.42   Max.   :60.207   Max.   :83.000      Max.   : 1.5531
##  NA's   :8155     NA's   :8645
##  Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf.  Set.Coef.B..lbf.mph.
##  Min.   :0.00248            Min.   :-31.798   Min.   :-1.0429
##  1st Qu.:0.01744            1st Qu.:  8.653   1st Qu.:-0.0590
##  Median :0.02030            Median : 14.000   Median : 0.0942
##  Mean   :0.02220            Mean   : 14.426   Mean   : 0.1104
##  3rd Qu.:0.02606            3rd Qu.: 19.580   3rd Qu.: 0.2750
##  Max.   :0.20000            Max.   : 96.000   Max.   : 1.3109
##
##  Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd
##  Min.   :-0.08680        TWC    :8591
```

```
##  1st Qu.: 0.01804                :2129
##  Median : 0.02119        OC      : 147
##  Mean   : 0.02412        DPF     : 139
##  3rd Qu.: 0.02745        SCR     :  93
##  Max.   : 1.00000        NOXAD   :  46
##                          (Other):  38
##                  Aftertreatment.Device.Desc Police...Emergency.Vehicle.
##  Three-way catalyst           :8591         N:11118
##                               :2129         Y:   65
##  Oxidation catalyst           : 147
##  Diesel Particulate Filter    : 139
##  Selective Catalytic Reduction:  93
##  NOx Adsorber                 :  46
##  (Other)                      :  38
```

The summary data above indicates that 73 data values in the column RND_ADJ_FE for miles per gallon are missing. The missing values were replaced with the median to make the data consistent.

```
# Number of missing miles per gallon values
df2010to12 %>%
summarise(count = sum(is.na(RND_ADJ_FE)))

##   count
## 1    73
```

NA cleaning for 2010 to 2012 dataframe

```
# Replace missing values with the median for all numerical values
# mutate missing values, and modify the dataframe
df2010to12 <- df2010to12 %>%
  mutate(RND_ADJ_FE = replace(RND_ADJ_FE,
                                  is.na(RND_ADJ_FE),
                                  median(RND_ADJ_FE, na.rm = TRUE)))
#  X..of.Cylinders.and.Rotors column
df2010to12 <- df2010to12 %>%
  mutate( X..of.Cylinders.and.Rotors = replace( X..of.Cylinders.and.Rotors,
                                  is.na( X..of.Cylinders.and.Rotors),
                                  median( X..of.Cylinders.and.Rotors, na.rm =
TRUE)))
# THC..g.mi. column
df2010to12 <- df2010to12 %>%
  mutate(THC..g.mi. = replace(THC..g.mi.,
                                  is.na(THC..g.mi.),
                                  median(THC..g.mi., na.rm = TRUE)))

# CO..g.mi. column
df2010to12 <- df2010to12 %>%
  mutate(CO..g.mi.  = replace(CO..g.mi. ,
                                  is.na(CO..g.mi. ),
```

```
                                  median(CO..g.mi. , na.rm = TRUE)))

# CO2..g.mi.  column
df2010to12 <- df2010to12 %>%
  mutate(CO2..g.mi.  = replace(CO2..g.mi. ,
                                  is.na(CO2..g.mi. ),
                                  median(CO2..g.mi. , na.rm = TRUE)))

# NOx..g.mi.  column
df2010to12 <- df2010to12 %>%
  mutate(NOx..g.mi.  = replace(NOx..g.mi. ,
                                  is.na(NOx..g.mi. ),
                                  median(NOx..g.mi. , na.rm = TRUE)))
# CH4..g.mi.  column
df2010to12 <- df2010to12 %>%
  mutate(CH4..g.mi.  = replace(CH4..g.mi. ,
                                  is.na(CH4..g.mi. ),
                                  median(CH4..g.mi. , na.rm = TRUE)))

# N2O..g.mi.  column
df2010to12 <- df2010to12 %>%
  mutate(N2O..g.mi.  = replace(N2O..g.mi. ,
                                  is.na(N2O..g.mi.),
                                  median(N2O..g.mi. , na.rm = TRUE)))

# FE.Bag.1   column
df2010to12 <- df2010to12 %>%
  mutate(FE.Bag.1   = replace(FE.Bag.1  ,
                                  is.na(FE.Bag.1  ),
                                  median(FE.Bag.1  , na.rm = TRUE)))

# FE.Bag.2  column
df2010to12 <- df2010to12 %>%
  mutate(FE.Bag.2  = replace(FE.Bag.2,
                                  is.na(FE.Bag.2 ),
                                  median(FE.Bag.2 , na.rm = TRUE)))

# FE.Bag.3 column
df2010to12 <- df2010to12 %>%
  mutate(FE.Bag.3  = replace(FE.Bag.3,
                                  is.na(FE.Bag.3 ),
                                  median(FE.Bag.3 , na.rm = TRUE)))
```

NA cleaning for 2014 to 2016 dataframe

```
# Replace missing values with the median for all numerical values
# mutate missing values, and modify the dataframe
df2014to16 <- df2014to16 %>%
  mutate(RND_ADJ_FE = replace(RND_ADJ_FE,
```

```
                                  is.na(RND_ADJ_FE),
                                  median(RND_ADJ_FE, na.rm = TRUE)))
#  X..of.Cylinders.and.Rotors column
df2014to16 <- df2014to16 %>%
  mutate( X..of.Cylinders.and.Rotors = replace( X..of.Cylinders.and.Rotors,
                                  is.na( X..of.Cylinders.and.Rotors),
                                  median( X..of.Cylinders.and.Rotors, na.rm =
TRUE)))
# THC..g.mi. column
df2014to16 <- df2014to16 %>%
  mutate(THC..g.mi. = replace(THC..g.mi.,
                                  is.na(THC..g.mi.),
                                  median(THC..g.mi., na.rm = TRUE)))

# CO..g.mi. column
df2014to16 <- df2014to16 %>%
  mutate(CO..g.mi.  = replace(CO..g.mi. ,
                                  is.na(CO..g.mi. ),
                                  median(CO..g.mi. , na.rm = TRUE)))

# CO2..g.mi.  column
df2014to16 <- df2014to16 %>%
  mutate(CO2..g.mi.  = replace(CO2..g.mi. ,
                                  is.na(CO2..g.mi. ),
                                  median(CO2..g.mi. , na.rm = TRUE)))

# NOx..g.mi.  column
df2014to16 <- df2014to16 %>%
  mutate(NOx..g.mi.  = replace(NOx..g.mi. ,
                                  is.na(NOx..g.mi. ),
                                  median(NOx..g.mi. , na.rm = TRUE)))
# CH4..g.mi.  column
df2014to16 <- df2014to16 %>%
  mutate(CH4..g.mi.  = replace(CH4..g.mi. ,
                                  is.na(CH4..g.mi. ),
                                  median(CH4..g.mi. , na.rm = TRUE)))

# N2O..g.mi.  column
df2014to16 <- df2014to16 %>%
  mutate(N2O..g.mi.  = replace(N2O..g.mi. ,
                                  is.na(N2O..g.mi.),
                                  median(N2O..g.mi. , na.rm = TRUE)))

# FE.Bag.1   column
df2014to16 <- df2014to16 %>%
  mutate(FE.Bag.1   = replace(FE.Bag.1  ,
                                  is.na(FE.Bag.1  ),
                                  median(FE.Bag.1  , na.rm = TRUE)))
```

```r
# FE.Bag.2  column
df2014to16 <- df2014to16 %>%
  mutate(FE.Bag.2  = replace(FE.Bag.2,
                                    is.na(FE.Bag.2 ),
                                    median(FE.Bag.2 , na.rm = TRUE)))

# FE.Bag.3 column
df2014to16 <- df2014to16 %>%
  mutate(FE.Bag.3  = replace(FE.Bag.3,
                                    is.na(FE.Bag.3 ),
                                    median(FE.Bag.3 , na.rm = TRUE)))
```

NA cleaning for 2014 to 2016 dataframe

```r
# Replace missing values with the median for all numerical values
# mutate missing values, and modify the dataframe
df2018to20 <- df2018to20 %>%
  mutate(RND_ADJ_FE = replace(RND_ADJ_FE,
                                    is.na(RND_ADJ_FE),
                                    median(RND_ADJ_FE, na.rm = TRUE)))
#  X..of.Cylinders.and.Rotors column
df2018to20 <- df2018to20 %>%
  mutate( X..of.Cylinders.and.Rotors = replace( X..of.Cylinders.and.Rotors,
                                    is.na( X..of.Cylinders.and.Rotors),
                                    median( X..of.Cylinders.and.Rotors, na.rm =
TRUE)))
# THC..g.mi. column
df2018to20 <- df2018to20 %>%
  mutate(THC..g.mi. = replace(THC..g.mi.,
                                    is.na(THC..g.mi.),
                                    median(THC..g.mi., na.rm = TRUE)))

# CO..g.mi. column
df2018to20 <- df2018to20 %>%
  mutate(CO..g.mi.  = replace(CO..g.mi. ,
                                    is.na(CO..g.mi. ),
                                    median(CO..g.mi. , na.rm = TRUE)))

# CO2..g.mi.  column
df2018to20 <- df2018to20 %>%
  mutate(CO2..g.mi.  = replace(CO2..g.mi. ,
                                    is.na(CO2..g.mi. ),
                                    median(CO2..g.mi. , na.rm = TRUE)))

# NOx..g.mi.  column
df2018to20 <- df2018to20 %>%
  mutate(NOx..g.mi.  = replace(NOx..g.mi. ,
                                    is.na(NOx..g.mi. ),
                                    median(NOx..g.mi. , na.rm = TRUE)))
```

```r
# CH4..g.mi.  column
df2018to20 <- df2018to20 %>%
  mutate(CH4..g.mi.  = replace(CH4..g.mi. ,
                               is.na(CH4..g.mi. ),
                               median(CH4..g.mi. , na.rm = TRUE)))

# N2O..g.mi.  column
df2018to20 <- df2018to20 %>%
  mutate(N2O..g.mi.  = replace(N2O..g.mi. ,
                               is.na(N2O..g.mi.),
                               median(N2O..g.mi. , na.rm = TRUE)))

# FE.Bag.1   column
df2018to20 <- df2018to20 %>%
  mutate(FE.Bag.1   = replace(FE.Bag.1  ,
                               is.na(FE.Bag.1  ),
                               median(FE.Bag.1  , na.rm = TRUE)))

# FE.Bag.2  column
df2018to20 <- df2018to20 %>%
  mutate(FE.Bag.2  = replace(FE.Bag.2,
                               is.na(FE.Bag.2 ),
                               median(FE.Bag.2 , na.rm = TRUE)))

# FE.Bag.3 column
df2018to20 <- df2018to20 %>%
  mutate(FE.Bag.3  = replace(FE.Bag.3,
                               is.na(FE.Bag.3 ),
                               median(FE.Bag.3 , na.rm = TRUE)))

# Confirming the missing values have been replaced with the median
# Miles per gallon column
df2010to12 %>%
summarise(count = sum(is.na(RND_ADJ_FE)))

##   count
## 1     0

# X..of.Cylinders.and.Rotors column
df2010to12 %>%
summarise(count = sum(is.na( X..of.Cylinders.and.Rotors )))

##   count
## 1     0

#here
# THC..g.mi. column
df2010to12 %>%
summarise(count = sum(is.na( THC..g.mi. )))
```

```
##   count
## 1     0
```

```
# CO..g.mi. column
df2010to12 %>%
summarise(count = sum(is.na( CO..g.mi. )))
```

```
##   count
## 1     0
```

```
# CO2..g.mi. column
df2010to12 %>%
summarise(count = sum(is.na( CO2..g.mi.  )))
```

```
##   count
## 1     0
```

```
# NOx..g.mi. column
df2010to12 %>%
summarise(count = sum(is.na( NOx..g.mi. )))
```

```
##   count
## 1     0
```

```
# CH4..g.mi. column
df2010to12 %>%
summarise(count = sum(is.na( CH4..g.mi. )))
```

```
##   count
## 1     0
```

```
# X..of.Cylinders.and.Rotors column
df2010to12 %>%
summarise(count = sum(is.na( N2O..g.mi. )))
```

```
##   count
## 1     0
```

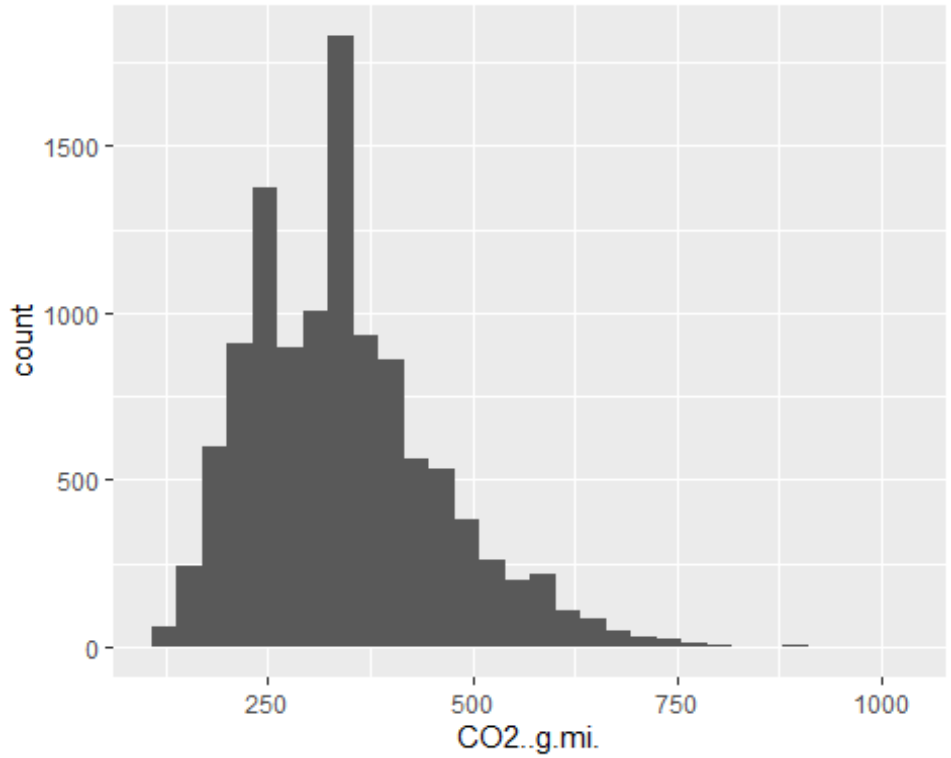
```
# X..of.Cylinders.and.Rotors column
df2010to12 %>%
summarise(count = sum(is.na( FE.Bag.1  )))
```

```
##   count
## 1     0
```

```
# X..of.Cylinders.and.Rotors column
df2010to12 %>%
summarise(count = sum(is.na( FE.Bag.2 )))
```

```
##   count
## 1     0
```

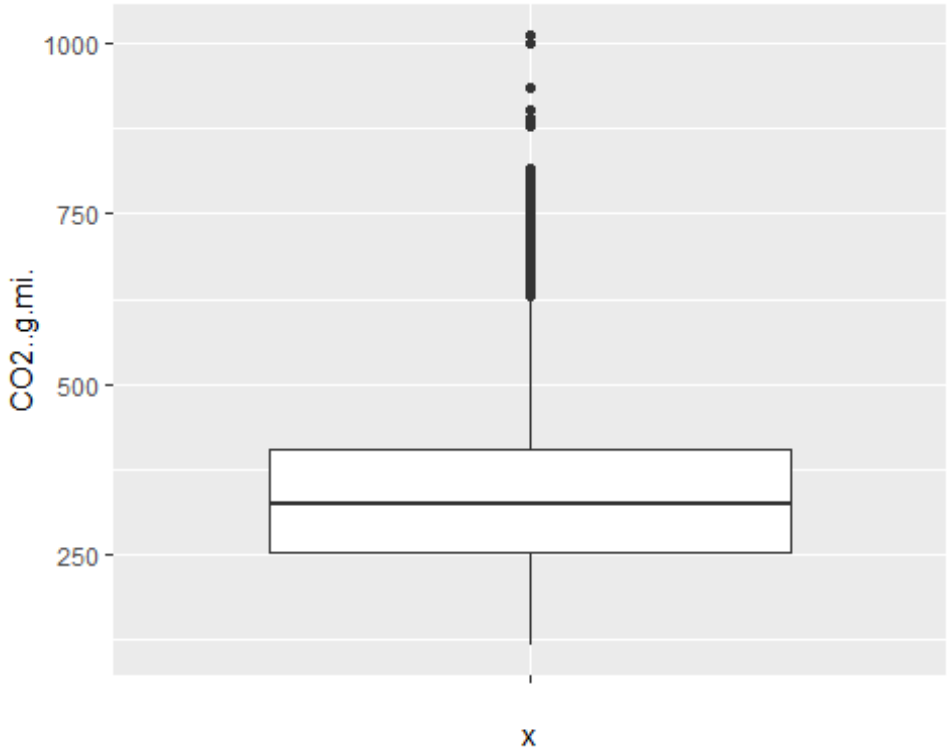
```
# X..of.Cylinders.and.Rotors column
df2010to12 %>%
summarise(count = sum(is.na( FE.Bag.3 )))

##   count
## 1     0

# Inspecting CO2 emissions
ggplot(data = df2010to12, aes(CO2..g.mi.))+geom_histogram()

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



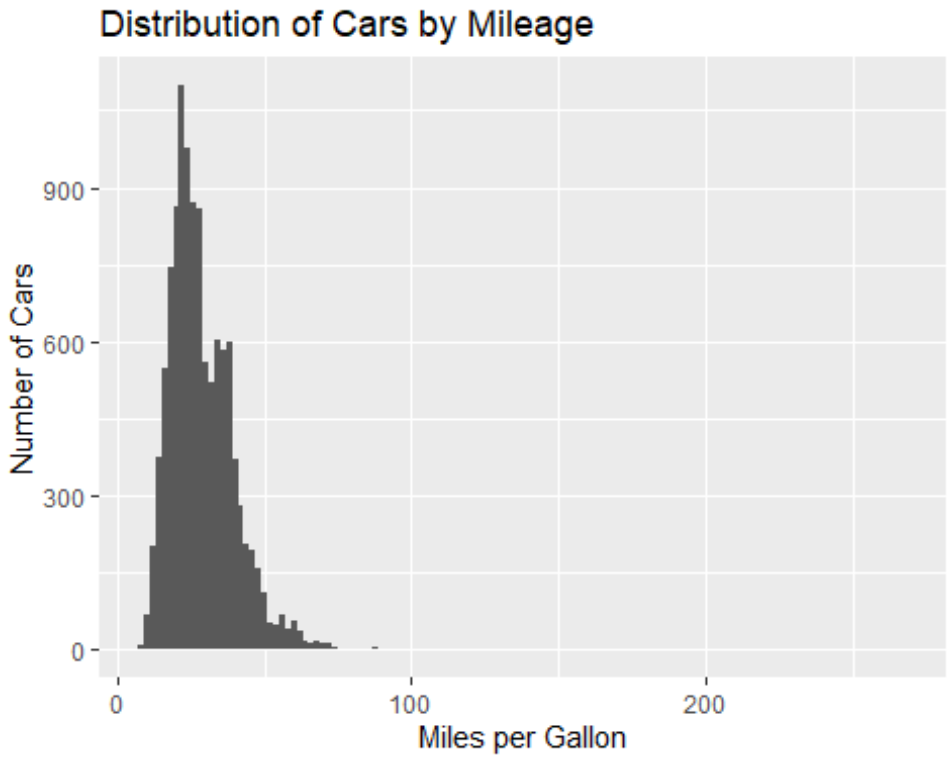
The distribution of CO2.g.mi is close to normal distribution, so it can be fit to a linear model.However, the CO2.g.mi data contains 831 missing values that need to be cleaned. For this case, the missing values were replaced with the median of the CO2.g.mi data.

Outlier detection on the CO2..g.mi using a box plot

```
ggplot(data = df2010to12, mapping = aes(x = "",y = CO2..g.mi.)) +
geom_boxplot()
```




```
ggplot(df2010to12, aes(RND_ADJ_FE)) +
  geom_histogram(binwidth = 2) + xlab('Miles per Gallon') + ylab('Number of
Cars') +
   ggtitle('Distribution of Cars by Mileage')
```




```
# Extracting the potential outliers

outliers <- boxplot.stats(df2010to12$CO2..g.mi.)$out
min(outliers)

## [1] 628.88

sum(outliers > 812.5)

## [1] 13
```

The extracted potential outliers are 217 in total. This is equal to 1.94% of the total observations.The minimum outlier value being 628.88. However, cross-checking with the CO2 emission histogram, the outliers lie beyond 812.5. These values are thirteen and equal 0.12% of the total observations. The dataframe was mutated to exclude these 13 values.

```
# Exclude CO2 emission values greater than 812.5 as outliers
# df2010to12 <- df2010to12[-c(df2010to12$CO2..g.mi. > 812.5)]

dim(df2010to12)

## [1] 11183    42

summary(df2014to16)

##        X             Year       Veh.Mfr.Code  Represented.Test.Veh.Make
##  Min.   :   0   Min.   :2014   BMX    :2492   BMW          :2240
##  1st Qu.:1187   1st Qu.:2014   TYX    :1527   Ford         : 971
```

```
##  Median :2374   Median :2015   FMX    :1150   AUDI         : 880
##  Mean   :2374   Mean   :2015   NSX    :1135   TOYOTA       : 847
##  3rd Qu.:3561   3rd Qu.:2016   GMX    :1065   CHEVROLET    : 749
##  Max.   :4831   Max.   :2016   CRX    : 952   Mercedes-Benz: 730
##                                (Other):5925   (Other)      :7829
##  Represented.Test.Veh.Model Test.Veh.Displacement..L. Vehicle.Type
##  Beetle     :  211          Min.   :  0.001           Both :1890
##  Jetta      :  150          1st Qu.:  2.000           Car  :8924
##  Passat     :  131          Median :  3.000           Truck:3432
##  ACCORD     :  129          Mean   :  3.329
##  Dart       :  113          3rd Qu.:  3.700
##  535d xDrive:  104          Max.   : 99.999
##  (Other)    :13408
##  Rated.Horsepower X..of.Cylinders.and.Rotors       Engine.Code
##  Min.   :   1.0   Min.   : 2.000             01          :  796
##  1st Qu.: 182.0   1st Qu.: 4.000             02          :  336
##  Median : 259.0   Median : 6.000             1           :  222
##  Mean   : 275.6   Mean   : 5.582             AA-100      :  207
##  3rd Qu.: 325.0   3rd Qu.: 6.000             2.0-N47-F30X:  136
##  Max.   :1200.0   Max.   :16.000             31A         :  118
##                                              (Other)     :12431
##
Tested.Transmission.Type
##  Automatic                                                          :4978
##  Semi-Automatic                                                     :4666
##  Manual                                                             :1862
##  Continuously Variable                                              :1392
##  Automated Manual- Selectable (e.g. Automated Manual with paddles): 639
##  Automated Manual                                                   : 396
##  (Other)                                                            : 313
##   X..of.Gears    Transmission.Lockup.            Drive.System.Description
##  Min.   :1.000   N: 3476              2-Wheel Drive, Front   :6010
##  1st Qu.:6.000   Y:10770              2-Wheel Drive, Rear    :6443
##  Median :6.000                       4-Wheel Drive          : 489
##  Mean   :6.051                       All Wheel Drive        :1158
##  3rd Qu.:8.000                       Part-time 4-Wheel Drive: 146
##  Max.   :9.000
##
##       Transmission.Overdrive.Desc Equivalent.Test.Weight..lbs..
Axle.Ratio
##  No gear ratio < 1  :  199        Min.   :2125                 Min.
:1.000
##  Top gear ration < 1:14047        1st Qu.:3625                 1st
Qu.:3.070
##                                   Median :4250                 Median
:3.330
##                                   Mean   :4270                 Mean
:3.453
##                                   3rd Qu.:4750                 3rd
Qu.:3.730
```

```
##                                  Max.   :8500                  Max.
:9.990
##
##    N.V.Ratio                            Shift.Indicator.Light.Use.Desc
##  Min.   :  0.00   Equipped, not shifted by SIL        :  372
##  1st Qu.: 25.80   Equipped, shifted by survey schedule:  158
##  Median : 28.40   Not eqipped                         :13716
##  Mean   : 31.07
##  3rd Qu.: 32.50
##  Max.   :999.00
##
##                            Test.Procedure.Description Test.Fuel.Type.Cd
##  HWFE                                    :6065         Min.   :19.00
##  Federal fuel 2-day exhaust (w/can load):4227         1st Qu.:61.00
##  Federal fuel 3-day exhaust              :1244         Median :61.00
##  US06                                    : 757         Mean   :53.99
##  CVS 75 and later (w/o can. load)        : 557         3rd Qu.:61.00
##  SC03                                    : 539         Max.   :62.00
##  (Other)                                 : 857
##                                Test.Fuel.Type.Description Test.Category
##  Tier 2 Cert Gasoline                       :11230        CD  : 218
##  Federal Cert Diesel 7-15 PPM Sulfur        : 1322        FTP :6667
##  E85 (85% Ethanol 15% EPA Unleaded Gasoline):  654        HWY :6065
##  CARB Phase II Gasoline                     :  364        SC03: 539
##  Cold CO Premium (Tier 2)                   :  294        US06: 757
##  Electricity                                :  202
##  (Other)                                    :  180
##    THC..g.mi.         CO..g.mi.          CO2..g.mi.      NOx..g.mi.
##  Min.   :0.000000   Min.   :  0.0000   Min.   :  0.0   Min.   :0.000000
##  1st Qu.:0.005052   1st Qu.:  0.0775   1st Qu.:225.7   1st Qu.:0.004299
##  Median :0.013500   Median :  0.1900   Median :289.6   Median :0.008900
##  Mean   :0.024235   Mean   :  0.5018   Mean   :309.1   Mean   :0.012881
##  3rd Qu.:0.025198   3rd Qu.:  0.3799   3rd Qu.:372.7   3rd Qu.:0.016000
##  Max.   :0.742710   Max.   :323.0000   Max.   :999.0   Max.   :0.483910
##
##    CH4..g.mi.         N2O..g.mi.         RND_ADJ_FE          FE.Bag.1
##  Min.   :  0.0000   Min.   :0.00000   Min.   :    0.00   Min.   : 0.00
##  1st Qu.:  0.0022   1st Qu.:0.01000   1st Qu.:   23.50   1st Qu.:23.21
##  Median :  0.0038   Median :0.01000   Median :   30.60   Median :23.21
##  Mean   :  0.0363   Mean   :0.00985   Mean   :   39.02   Mean   :23.56
##  3rd Qu.:  0.0064   3rd Qu.:0.01000   3rd Qu.:   40.10   3rd Qu.:23.21
##  Max.   :424.3200   Max.   :1.98000   Max.   :10000.00   Max.   :61.24
##
##     FE.Bag.2         FE.Bag.3     Target.Coef.A..lbf.
Target.Coef.B..lbf.mph.
##  Min.   :  0.00   Min.   : 0.00   Min.   :  0.00      Min.   :-0.84730
##  1st Qu.: 25.01   1st Qu.:28.03   1st Qu.: 30.47      1st Qu.: 0.04883
##  Median : 25.01   Median :28.03   Median : 37.59      Median : 0.26684
##  Mean   : 25.95   Mean   :28.39   Mean   : 38.28      Mean   : 0.22894
##  3rd Qu.: 25.01   3rd Qu.:28.03   3rd Qu.: 45.64      3rd Qu.: 0.42900
```

```
##  Max.   :126.29   Max.   :68.97   Max.   :150.00      Max.   : 1.72838
##
##  Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf. Set.Coef.B..lbf.mph.
##  Min.   :0.00000            Min.   :-30.01   Min.   :-0.9833
##  1st Qu.:0.01740            1st Qu.:  6.21   1st Qu.:-0.0419
##  Median :0.02061            Median : 12.20   Median : 0.0997
##  Mean   :0.02171            Mean   : 12.05   Mean   : 0.1064
##  3rd Qu.:0.02474            3rd Qu.: 18.79   3rd Qu.: 0.2437
##  Max.   :0.09442            Max.   :150.20   Max.   : 2.1980
##
##  Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd
##  Min.   :-0.08680              :  566
##  1st Qu.: 0.01728        DPF   :  389
##  Median : 0.01971        HCAD  :   15
##  Mean   : 0.02266        NOXAD :  173
##  3rd Qu.: 0.02412        OC    :  396
##  Max.   : 1.00000        SCR   :  369
##                          TWC   :12338
##                 Aftertreatment.Device.Desc Police...Emergency.Vehicle.
##                                    :  566        N:14107
##  Diesel Particulate Filter         :  389        Y:  139
##  HC-Adsorber                       :   15
##  NOx Adsorber                      :  173
##  Oxidation catalyst                :  396
##  Selective Catalytic Reduction     :  369
##  Three-way catalyst                :12338
```

```
# Number of missing miles per gallon values
df2014to16 %>%
summarise(count = sum(is.na(RND_ADJ_FE)))
```

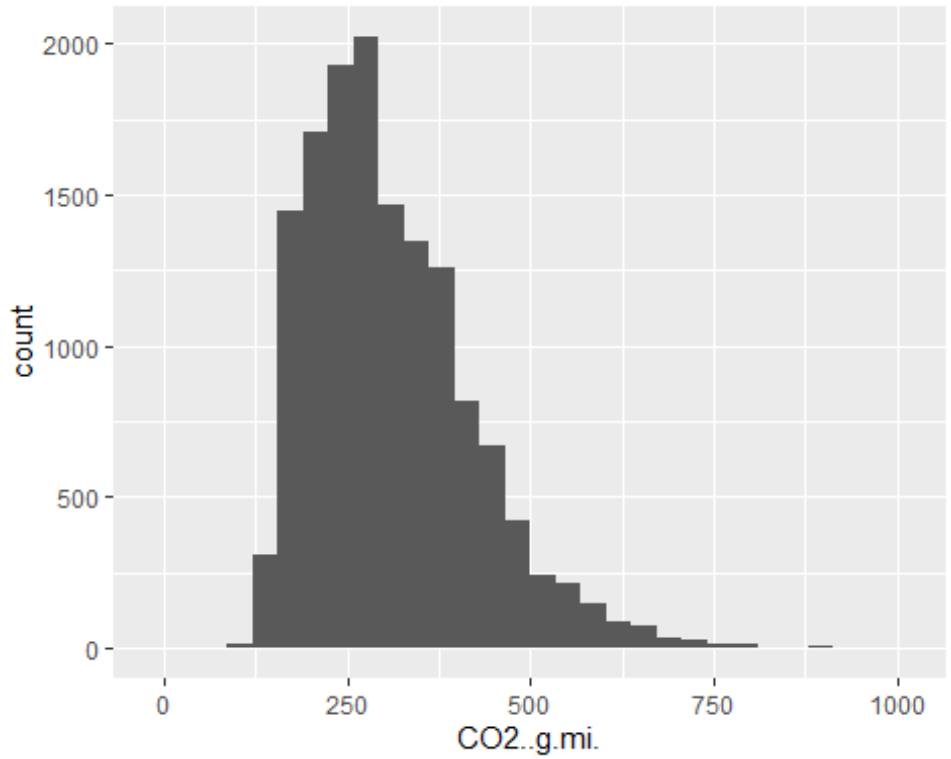
```
##   count
## 1     0
```

```
# Replace missing values with the median
# mutate missing values, and modify the dataframe
df2014to16 <- df2014to16 %>%
  mutate(RND_ADJ_FE = replace(RND_ADJ_FE,
                              is.na(RND_ADJ_FE),
                              median(RND_ADJ_FE, na.rm = TRUE)))

# Confirming the missing values have been replaced with the median
df2014to16 %>%
summarise(count = sum(is.na(RND_ADJ_FE)))
```

```
##   count
## 1     0
```

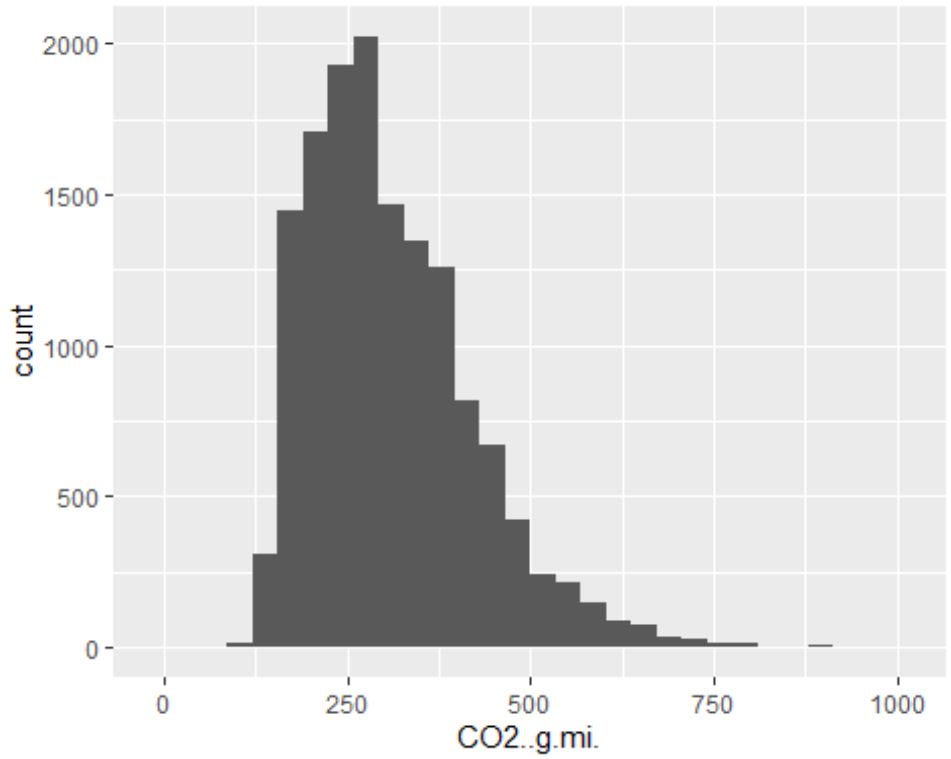
```
# Inspecting emissions
ggplot(data = df2014to16, aes(CO2..g.mi.))+geom_histogram()
```

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



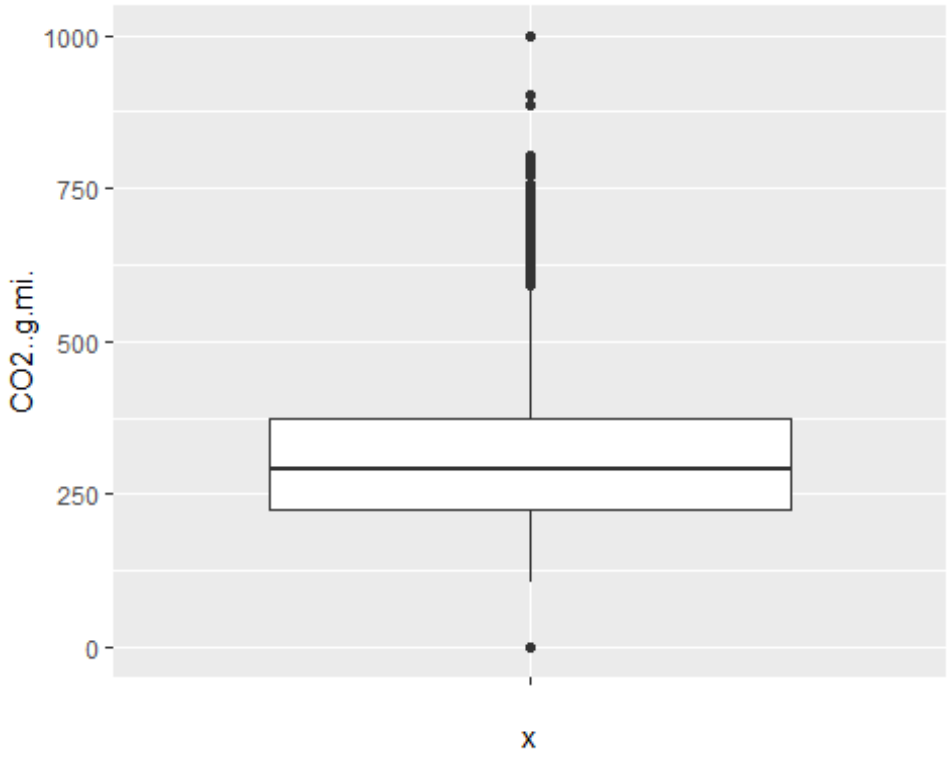
The distribution of CO2.g.mi is close to normal distribution, so it can be fit to a linear model.However, the CO2.g.mi data contains 831 missing values that need to be cleaned. For this case, the missing values were replaced with the median of the CO2.g.mi data.

```
# Number of missing CO2.g.mi values
df2014to16 %>%
summarise(count = sum(is.na(CO2..g.mi.)))

##   count
## 1     0

# mutate missing values, and modify the dataframe
df2014to16 <- df2014to16 %>%
  mutate(CO2..g.mi. = replace(CO2..g.mi.,
                              is.na(CO2..g.mi.),
                              median(CO2..g.mi., na.rm = TRUE)))

# The cleaned data histogram
ggplot(data = df2014to16, aes(CO2..g.mi.))+geom_histogram()

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



Outlier detection on the CO2..g.mi using a box plot

```
ggplot(data = df2014to16, mapping = aes(x = "",y = CO2..g.mi.)) +
geom_boxplot()
```



```r
# Extracting the potential outliers

outliers <- boxplot.stats(df2014to16$CO2..g.mi.)$out
min(outliers)
```

```
## [1] 0
```

```r
sum(outliers > 812.5)
```

```
## [1] 7
```

The extracted potential outliers are 217 in total. This is equal to 1.94% of the total observations.The minimum outlier value being 628.88. However, cross-checking with the CO2 emission histogram, the outliers lie beyond 812.5. These values are thirteen and equal 0.12% of the total observations. The dataframe was mutated to exclude these 13 values.

```r
# Exclude CO2 emission values greater than 812.5 as outliers
# df2010to12 <- df2010to12[-c(df2010to12$CO2..g.mi. > 812.5)]

dim(df2014to16)
```

```
## [1] 14246    42
```

```r
summary(df2018to20)
```

```
##    Model.Year        Year       Veh.Mfr.Code  Represented.Test.Veh.Make
##  Min.   :2018   Min.   :2018   BMX    :1610   BMW          :1319
##  1st Qu.:2018   1st Qu.:2018   TYX    :1512   Ford         :1055
##  Median :2019   Median :2019   GMX    :1504   CHEVROLET    : 957
##  Mean   :2019   Mean   :2019   FMX    :1230   HONDA        : 957
##  3rd Qu.:2020   3rd Qu.:2020   HNX    :1062   TOYOTA       : 936
##  Max.   :2020   Max.   :2020   VGA    :1036   Mercedes-Benz: 617
##                                (Other):5687   (Other)      :7800
##         Represented.Test.Veh.Model Test.Veh.Displacement..L. Vehicle.Type
##  CX-5                 :  135        Min.   :  0.001          Both :1562
##  CAMARO               :  133        1st Qu.:  2.000          Car  :8016
##  CIVIC 5DR            :  119        Median :  2.500          Truck:4063
##  F150 4x4             :  108        Mean   :  3.269
##  CIVIC 2DR COUPE 1.5L:   98        3rd Qu.:  3.500
##  Mustang              :   98        Max.   : 99.999
##  (Other)              :12950
##  Rated.Horsepower X..of.Cylinders.and.Rotors  Engine.Code
##  Min.   :   1.0   Min.   : 2.000             01     :  870
##  1st Qu.: 178.0   1st Qu.: 4.000             1      :  394
##  Median : 260.0   Median : 4.000             02     :  352
##  Mean   : 284.5   Mean   : 5.308             AA-100 :  188
##  3rd Qu.: 345.0   3rd Qu.: 6.000             03     :  186
##  Max.   :1500.0   Max.   :16.000             AA-200 :  175
##                                              (Other):11476
##
Tested.Transmission.Type
```

```
##  Semi-Automatic                                                  :4813
##  Automatic                                                       :3892
##  Continuously Variable                                           :1865
##  Manual                                                          :1138
##  Automated Manual- Selectable (e.g. Automated Manual with paddles): 866
##  Automated Manual                                                : 595
##  (Other)                                                         : 472
##   X..of.Gears     Transmission.Lockup.            Drive.System.Description
##  Min.   : 1.000   N: 3407              2-Wheel Drive, Front   :6069
##  1st Qu.: 6.000   Y:10234              2-Wheel Drive, Rear    :5518
##  Median : 7.000                        4-Wheel Drive          : 365
##  Mean   : 6.264                        All Wheel Drive        :1640
##  3rd Qu.: 8.000                        Part-time 4-Wheel Drive:  49
##  Max.   :10.000
##
##      Transmission.Overdrive.Desc Equivalent.Test.Weight..lbs..
Axle.Ratio
##  No gear ratio < 1 :  420        Min.   :2375                  Min.
:0.000
##  Top gear ratio < 1:13221        1st Qu.:3750                  1st
Qu.:3.130
##                                  Median :4250                  Median
:3.420
##                                  Mean   :4288                  Mean
:3.601
##                                  3rd Qu.:4750                  3rd
Qu.:3.800
##                                  Max.   :7000                  Max.
:9.700
##
##    N.V.Ratio                             Shift.Indicator.Light.Use.Desc
##  Min.   :  0.00   Equipped, not shifted by SIL        :  584
##  1st Qu.: 24.30   Equipped, shifted by survey schedule:  168
##  Median : 26.90   Not eqipped                         :12887
##  Mean   : 29.53   Equipped, shifted by SIL            :    2
##  3rd Qu.: 31.50
##  Max.   :155.10
##
##                             Test.Procedure.Description Test.Fuel.Type.Cd
##  HWFE                                      :5387         Min.   :19.00
##  Federal fuel 2-day exhaust (w/can load):3978         1st Qu.:61.00
##  Federal fuel 3-day exhaust                :1251         Median :61.00
##  US06                                      : 958         Mean   :56.79
##  SC03                                      : 692         3rd Qu.:61.00
##  Cold CO                                   : 608         Max.   :62.00
##  (Other)                                   : 767
##                                  Test.Fuel.Type.Description Test.Category
##  Tier 2 Cert Gasoline                        :11596        CD  : 419
##  Federal Cert Diesel 7-15 PPM Sulfur         :  716        FTP :6185
##  Electricity                                 :  398        HWY :5387
```

```
##  Cold CO Regular (Tier 2)                    :  345        SC03: 692
##  E85 (85% Ethanol 15% EPA Unleaded Gasoline):  323        US06: 958
##  Cold CO Premium (Tier 2)                    :  212
##  (Other)                                     :   51
##    THC..g.mi.         CO..g.mi.          CO2..g.mi.      NOx..g.mi.
##  Min.   :0.000000   Min.   :  0.0000   Min.   :  0.0   Min.   :0.000000
##  1st Qu.:0.004483   1st Qu.:  0.0864   1st Qu.:224.0   1st Qu.:0.003704
##  Median :0.010851   Median :  0.1810   Median :285.2   Median :0.007600
##  Mean   :0.020022   Mean   :  0.3472   Mean   :301.4   Mean   :0.011915
##  3rd Qu.:0.020300   3rd Qu.:  0.3487   3rd Qu.:357.6   3rd Qu.:0.013800
##  Max.   :1.205000   Max.   :323.0000   Max.   :971.0   Max.   :0.326003
##
##    CH4..g.mi.         N2O..g.mi.         RND_ADJ_FE          FE.Bag.1
##  Min.   :  0.0000   Min.   :0.000000   Min.   :    0.00   Min.   :  0.00
##  1st Qu.:  0.0018   1st Qu.:0.001000   1st Qu.:   24.70   1st Qu.: 24.12
##  Median :  0.0033   Median :0.001000   Median :   31.40   Median : 24.12
##  Mean   :  0.0704   Mean   :0.002556   Mean   :   47.03   Mean   : 24.59
##  3rd Qu.:  0.0057   3rd Qu.:0.001000   3rd Qu.:   40.80   3rd Qu.: 24.12
##  Max.   :424.3200   Max.   :1.980000   Max.   :10000.00   Max.   :999.00
##
##     FE.Bag.2         FE.Bag.3     Target.Coef.A..lbf.
Target.Coef.B..lbf.mph.
##  Min.   :  0.00   Min.   :  0.0   Min.   :15.43       Min.   :-0.8207
##  1st Qu.: 26.60   1st Qu.: 28.8   1st Qu.:30.30       1st Qu.: 0.0578
##  Median : 26.60   Median : 28.8   Median :37.80       Median : 0.2507
##  Mean   : 28.33   Mean   : 29.3   Mean   :39.06       Mean   : 0.2203
##  3rd Qu.: 26.60   3rd Qu.: 28.8   3rd Qu.:46.76       3rd Qu.: 0.4071
##  Max.   :999.00   Max.   :999.0   Max.   :86.80       Max.   : 2.4082
##
##  Target.Coef.C..lbf.mph..2. Set.Coef.A..lbf.  Set.Coef.B..lbf.mph.
##  Min.   :0.008707           Min.   :-99.900   Min.   :-1.9752
##  1st Qu.:0.018060           1st Qu.:  6.202   1st Qu.:-0.0066
##  Median :0.021668           Median : 12.072   Median : 0.1084
##  Mean   :0.022645           Mean   : 12.612   Mean   : 0.1232
##  3rd Qu.:0.025852           3rd Qu.: 18.542   3rd Qu.: 0.2286
##  Max.   :0.052210           Max.   : 64.520   Max.   : 8.5800
##
##  Set.Coef.C..lbf.mph..2. Aftertreatment.Device.Cd
##  Min.   :-0.03140              :  456
##  1st Qu.: 0.01721        DPF   :  209
##  Median : 0.02055        NOXAD:   50
##  Mean   : 0.02232        OC    :  203
##  3rd Qu.: 0.02538        OT    :   36
##  Max.   : 0.25994        SCR   :  224
##                          TWC   :12463
##                  Aftertreatment.Device.Desc Police...Emergency.Vehicle.
##                                     :  456        N:13507
##  Diesel Particulate Filter          :  209        Y:  134
##  NOx Adsorber                       :   50
##  Other                              :   36
```

```
##  Oxidation catalyst           :  203
##  Selective Catalytic Reduction:  224
##  Three-way catalyst           :12463
```

```
# Number of missing miles per gallon values
df2018to20 %>%
summarise(count = sum(is.na(RND_ADJ_FE)))
```

```
##   count
## 1     0
```

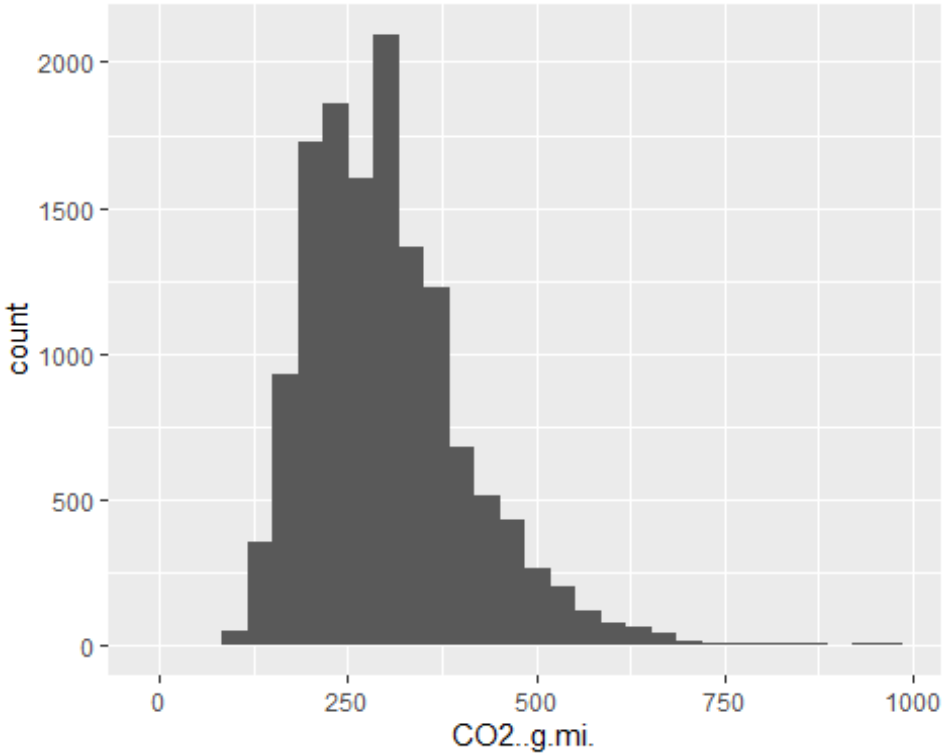
```
# Replace missing values with the median
# mutate missing values, and modify the dataframe
df2018to20 <- df2018to20 %>%
  mutate(RND_ADJ_FE = replace(RND_ADJ_FE,
                              is.na(RND_ADJ_FE),
                              median(RND_ADJ_FE, na.rm = TRUE)))
```

```
# Confirming the missing values have been replaced with the median
df2018to20 %>%
summarise(count = sum(is.na(RND_ADJ_FE)))
```

```
##   count
## 1     0
```

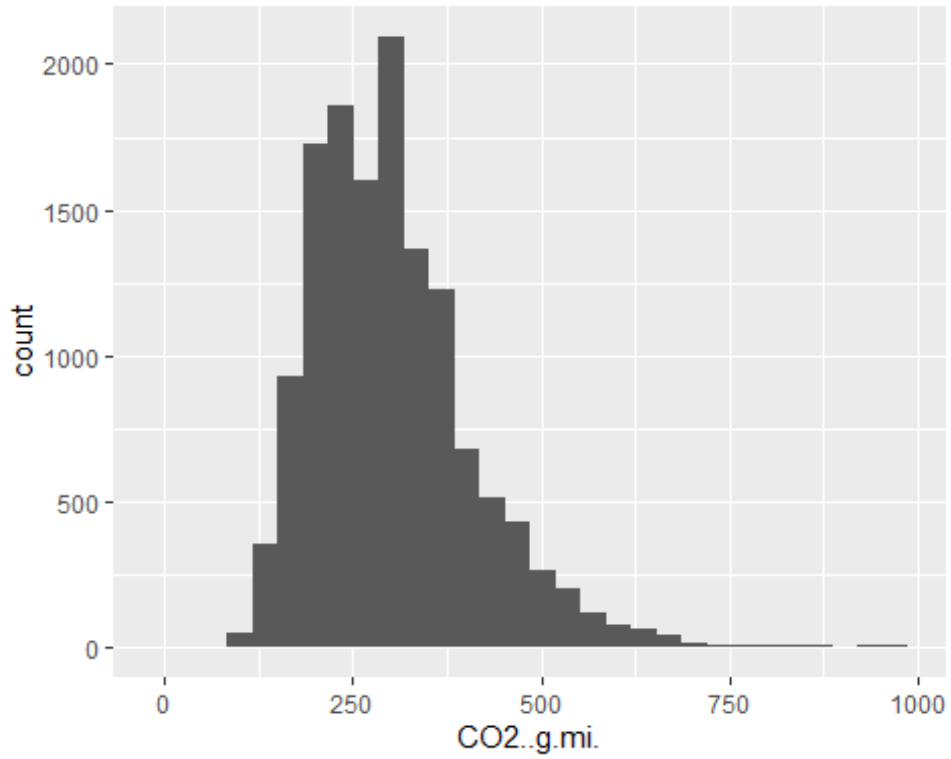
```
# Inspecting emissions
ggplot(data = df2018to20, aes(CO2..g.mi.))+geom_histogram()
```

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



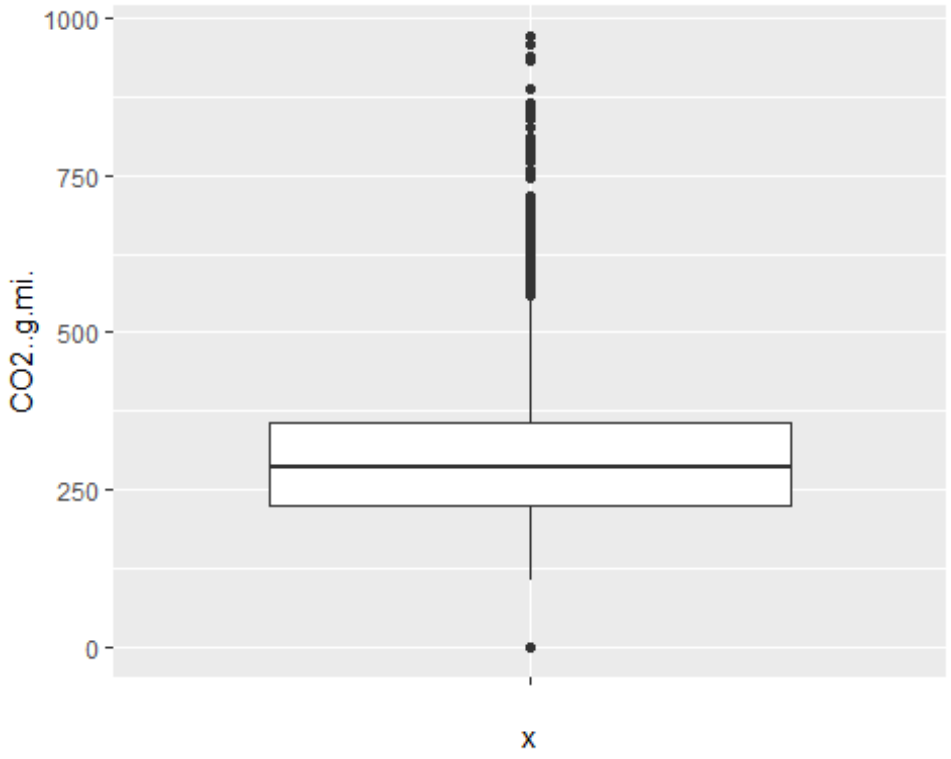
The distribution of CO2.g.mi is close to normal distribution, so it can be fit to a linear model.However, the CO2.g.mi data contains 831 missing values that need to be cleaned. For this case, the missing values were replaced with the median of the CO2.g.mi data.

```
# Number of missing CO2.g.mi values
df2018to20 %>%
summarise(count = sum(is.na(CO2..g.mi.)))

##   count
## 1     0

# mutate missing values, and modify the dataframe
df2018to20 <- df2018to20 %>%
  mutate(CO2..g.mi. = replace(CO2..g.mi.,
                              is.na(CO2..g.mi.),
                              median(CO2..g.mi., na.rm = TRUE)))

# The cleaned data histogram
ggplot(data = df2018to20, aes(CO2..g.mi.))+geom_histogram()

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



Outlier detection on the CO2..g.mi using a box plot

```
ggplot(data = df2018to20, mapping = aes(x = "",y = CO2..g.mi.)) +
geom_boxplot()
```



```
# Extracting the potential outliers

outliers <- boxplot.stats(df2018to20$CO2..g.mi.)$out
min(outliers)

## [1] 0

sum(outliers > 812.5)

## [1] 23
```

The extracted potential outliers are 217 in total. This is equal to 1.94% of the total observations.The minimum outlier value being 628.88. However, cross-checking with the CO2 emission histogram, the outliers lie beyond 812.5. These values are thirteen and equal 0.12% of the total observations. The dataframe was mutated to exclude these 13 values.

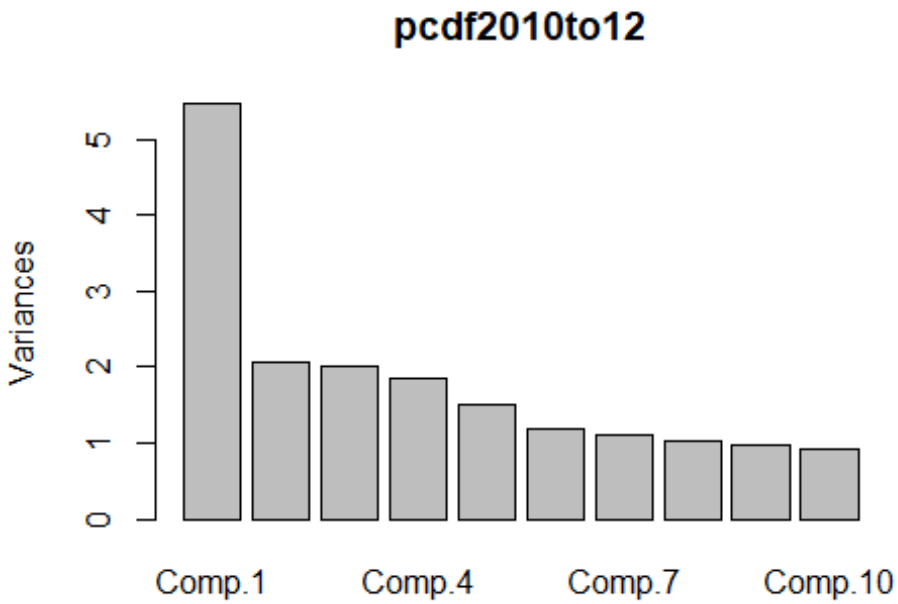
```
# Exclude CO2 emission values greater than 812.5 as outliers
# df2010to12 <- df2010to12[-c(df2010to12$CO2..g.mi. > 812.5)]

dim(df2018to20)

## [1] 13641    42
```

Principal components

2010 - 2012 dataframe

```
xdf2010to12 <-
cbind(df2010to12$Test.Veh.Displacement..L.,df2010to12$Rated.Horsepower,
df2010to12$X..of.Cylinders.and.Rotors, df2010to12$X..of.Gears,
df2010to12$Equivalent.Test.Weight..lbs..,df2010to12$Axle.Ratio,
df2010to12$N.V.Ratio, df2010to12$Test.Fuel.Type.Cd, df2010to12$THC..g.mi.,
df2010to12$CO..g.mi., df2010to12$CO2..g.mi.,df2010to12$NOx..g.mi.,
df2010to12$CH4..g.mi., df2010to12$N2O..g.mi., df2010to12$RND_ADJ_FE,
df2010to12$FE.Bag.1,df2010to12$FE.Bag.2, df2010to12$FE.Bag.3,
df2010to12$Target.Coef.A..lbf., df2010to12$Target.Coef.B..lbf.mph.,
df2010to12$Target.Coef.C..lbf.mph..2., df2010to12$Set.Coef.A..lbf.,
df2010to12$Set.Coef.B..lbf.mph., df2010to12$Set.Coef.C..lbf.mph..2.)

write.csv(xdf2010to12,'xdf2010to12.csv')
Num2010_12 <- read.csv("xdf2010to12.csv")
pcdf2010to12 <- princomp(xdf2010to12,cor = TRUE, scores = TRUE)
summary(pcdf2010to12)

## Importance of components:
##                           Comp.1     Comp.2     Comp.3     Comp.4
Comp.5
## Standard deviation     2.3390786 1.43473052 1.41848496 1.36166110
1.22688979
```

```
## Proportion of Variance 0.2279704 0.08576882 0.08383748 0.07725504
0.06271911
## Cumulative Proportion  0.2279704 0.31373918 0.39757666 0.47483170
0.53755080
##                            Comp.6     Comp.7     Comp.8     Comp.9
Comp.10
## Standard deviation     1.09192911 1.04939175 1.01046545 0.98589906
0.96588348
## Proportion of Variance 0.04967955 0.04588429 0.04254335 0.04049987
0.03887212
## Cumulative Proportion  0.58723035 0.63311465 0.67565800 0.71615787
0.75502999
##                           Comp.11    Comp.12   Comp.13    Comp.14
Comp.15
## Standard deviation     0.94431147 0.86951548 0.8252883 0.81042222
0.74518051
## Proportion of Variance 0.03715517 0.03150238 0.0283792 0.02736601
0.02313725
## Cumulative Proportion  0.79218516 0.82368755 0.8520667 0.87943275
0.90257000
##                           Comp.16    Comp.17   Comp.18   Comp.19
Comp.20
## Standard deviation     0.71691573 0.63525472 0.6066709 0.5249639
0.459849521
## Proportion of Variance 0.02141534 0.01681452 0.0153354 0.0114828
0.008810899
## Cumulative Proportion  0.92398534 0.94079986 0.9561353 0.9676181
0.976428959
##                            Comp.21     Comp.22     Comp.23     Comp.24
## Standard deviation     0.419927768 0.407885916 0.340648769 0.327036933
## Proportion of Variance 0.007347472 0.006932122 0.004835066 0.004456381
## Cumulative Proportion  0.983776431 0.990708553 0.995543619 1.000000000

plot(pcdf2010to12)
```




```
attributes(pcdf2010to12)

## $names
## [1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"
"call"
##
## $class
## [1] "princomp"
```

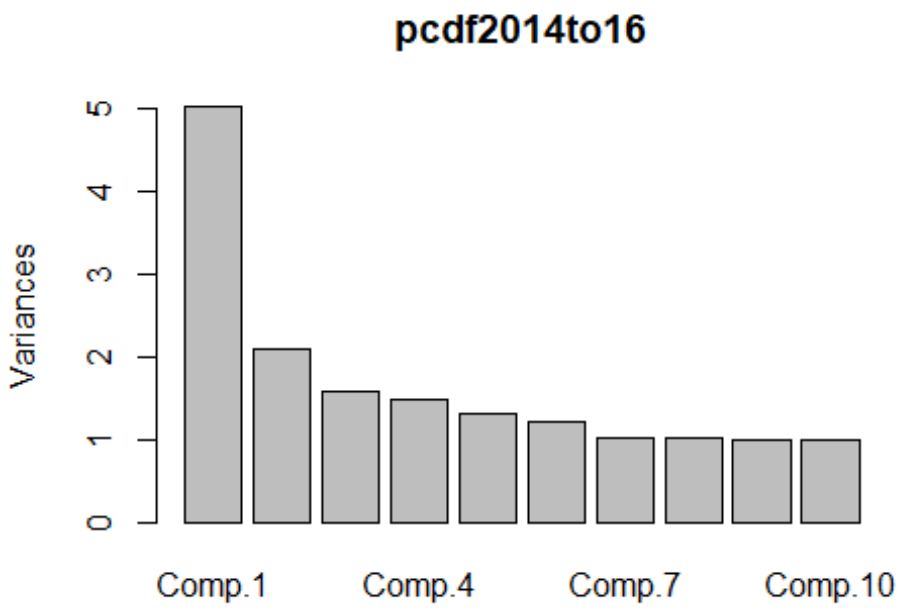
2014 - 2016 dataframe

```
xdf2014to16 <-
cbind(df2014to16$Test.Veh.Displacement..L.,df2014to16$Rated.Horsepower,
df2014to16$X..of.Cylinders.and.Rotors, df2014to16$X..of.Gears,
df2014to16$Equivalent.Test.Weight..lbs..,df2014to16$Axle.Ratio,
df2014to16$N.V.Ratio, df2014to16$Test.Fuel.Type.Cd, df2014to16$THC..g.mi.,
df2014to16$CO..g.mi., df2014to16$CO2..g.mi.,df2014to16$NOx..g.mi.,
df2014to16$CH4..g.mi., df2014to16$N2O..g.mi., df2014to16$RND_ADJ_FE,
df2014to16$FE.Bag.1,df2014to16$FE.Bag.2, df2014to16$FE.Bag.3,
df2014to16$Target.Coef.A..lbf., df2014to16$Target.Coef.B..lbf.mph.,
df2014to16$Target.Coef.C..lbf.mph..2., df2014to16$Set.Coef.A..lbf.,
df2014to16$Set.Coef.B..lbf.mph., df2014to16$Set.Coef.C..lbf.mph..2.)

write.csv(xdf2014to16,'xdf2014to16.csv')
Num2014_16 <- read.csv("xdf2014to16.csv")
```

```r
pcdf2014to16 <- princomp(xdf2014to16,cor = TRUE, scores = TRUE)
summary(pcdf2014to16)
```

```
## Importance of components:
##                           Comp.1     Comp.2     Comp.3     Comp.4
Comp.5
## Standard deviation     2.241019 1.44673308 1.25862585 1.21826163
1.14472262
## Proportion of Variance 0.209257 0.08720986 0.06600579 0.06184006
0.05459958
## Cumulative Proportion  0.209257 0.29646682 0.36247261 0.42431267
0.47891225
##                            Comp.6     Comp.7     Comp.8     Comp.9
Comp.10
## Standard deviation     1.10562781 1.00520508 1.00183032 1.00061712
0.99982106
## Proportion of Variance 0.05093387 0.04210155 0.04181933 0.04171811
0.04165176
## Cumulative Proportion  0.52984612 0.57194767 0.61376700 0.65548511
0.69713687
##                           Comp.11    Comp.12    Comp.13    Comp.14
Comp.15
## Standard deviation     0.99025716 0.97251750 0.95995947 0.90724934
0.89104757
## Proportion of Variance 0.04085872 0.03940793 0.03839676 0.03429589
0.03308191
## Cumulative Proportion  0.73799559 0.77740351 0.81580027 0.85009616
0.88317807
##                           Comp.16    Comp.17    Comp.18   Comp.19
Comp.20
## Standard deviation     0.77661946 0.75784771 0.68845156 0.5757687
0.52369817
## Proportion of Variance 0.02513074 0.02393055 0.01974856 0.0138129
0.01142749
## Cumulative Proportion  0.90830881 0.93223936 0.95198792 0.9658008
0.97722831
##                            Comp.21     Comp.22     Comp.23     Comp.24
## Standard deviation     0.477146953 0.371012800 0.321535503 0.278954642
## Proportion of Variance 0.009486217 0.005735437 0.004307712 0.003242321
## Cumulative Proportion  0.986714530 0.992449968 0.996757679 1.000000000
```

```r
plot(pcdf2014to16)
```




```
attributes(pcdf2014to16)

## $names
## [1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"
"call"
##
## $class
## [1] "princomp"
```

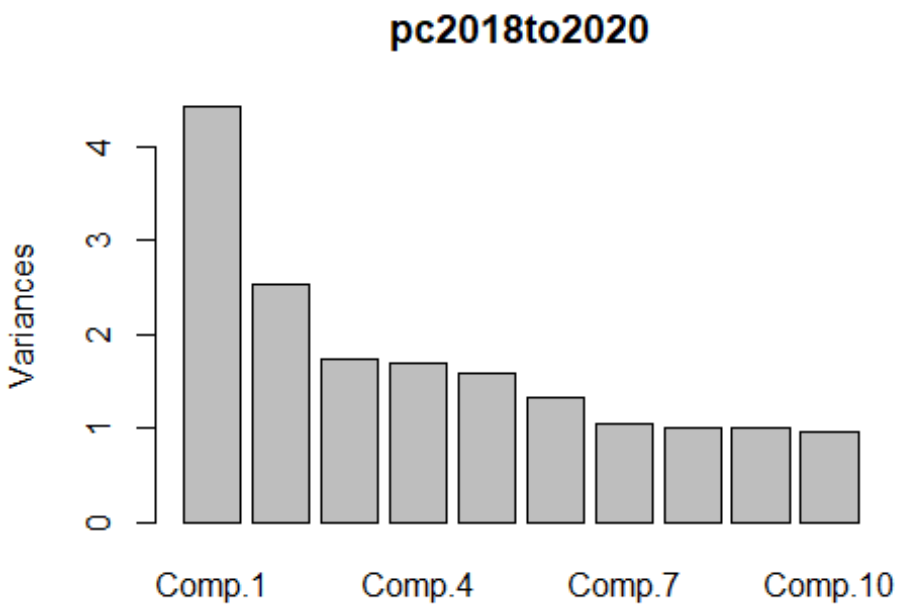
2018 - 2020 dataframe

```
xpc2018to2020 <-
cbind(df2018to20$Test.Veh.Displacement..L.,df2018to20$Rated.Horsepower,
df2018to20$X..of.Cylinders.and.Rotors, df2018to20$X..of.Gears,
df2018to20$Equivalent.Test.Weight..lbs..,df2018to20$Axle.Ratio,
df2018to20$N.V.Ratio, df2018to20$Test.Fuel.Type.Cd, df2018to20$THC..g.mi.,
df2018to20$CO..g.mi., df2018to20$CO2..g.mi.,df2018to20$NOx..g.mi.,
df2018to20$CH4..g.mi., df2018to20$N2O..g.mi., df2018to20$RND_ADJ_FE,
df2018to20$FE.Bag.1,df2018to20$FE.Bag.2, df2018to20$FE.Bag.3,
df2018to20$Target.Coef.A..lbf., df2018to20$Target.Coef.B..lbf.mph.,
df2018to20$Target.Coef.C..lbf.mph..2., df2018to20$Set.Coef.A..lbf.,
df2018to20$Set.Coef.B..lbf.mph., df2018to20$Set.Coef.C..lbf.mph..2.)

write.csv(xpc2018to2020,'xpc2018to2020.csv')
Num2018_20 <- read.csv("xpc2018to2020.csv")
```

```r
pc2018to2020 <- princomp(xpc2018to2020,cor = TRUE, scores = TRUE)
summary(pc2018to2020)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
Comp.5
## Standard deviation     2.1028312 1.5895017 1.31654112 1.30055803
1.26105999
## Proportion of Variance 0.1842458 0.1052715 0.07222002 0.07047713
0.06626135
## Cumulative Proportion  0.1842458 0.2895173 0.36173731 0.43221445
0.49847579
##                            Comp.6     Comp.7     Comp.8     Comp.9
Comp.10
## Standard deviation     1.15201693 1.02690636 1.00287405 0.99703851
0.98195858
## Proportion of Variance 0.05529763 0.04393903 0.04190652 0.04142024
0.04017678
## Cumulative Proportion  0.55377342 0.59771244 0.63961896 0.68103920
0.72121598
##                          Comp.11   Comp.12    Comp.13   Comp.14   Comp.15
## Standard deviation     0.9384839 0.9226899 0.89963755 0.8750968 0.8231598
## Proportion of Variance 0.0366980 0.0354732 0.03372282 0.0319081 0.0282330
## Cumulative Proportion  0.7579140 0.7933872 0.82711000 0.8590181 0.8872511
##                           Comp.16    Comp.17   Comp.18    Comp.19
Comp.20
## Standard deviation     0.81267432 0.66424127 0.6177670 0.58149420
0.55791855
## Proportion of Variance 0.02751831 0.01838402 0.0159015 0.01408898
0.01296971
## Cumulative Proportion  0.91476942 0.93315344 0.9490549 0.96314392
0.97611363
##                           Comp.21     Comp.22     Comp.23     Comp.24
## Standard deviation     0.476897870 0.380566973 0.344886989 0.286466365
## Proportion of Variance 0.009476316 0.006034634 0.004956126 0.003419291
## Cumulative Proportion  0.985589949 0.991624583 0.996580709 1.000000000
```

```r
plot(pc2018to2020)
```




```
attributes(pc2018to2020)

## $names
## [1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"
"call"
##
## $class
## [1] "princomp"
```

Selection of the principle components was based on cumulative proportion >= 85%. Based on this rule, 2010-12 dataframe had ten principle components, 2014-16 had fourteen principle components, and 2018-20 dataframe had fourteen principal components. #Once the principle components were diagnosed for all the three dataframes, the dimension reduction techniques are performed # Grouping similar vehicles using the reduced dimension #2010 - 2012 dataframe

```
principalcomps10_12 <- pcdf2010to12$scores[, 1:13]
groups <- kmeans(principalcomps10_12, 6)
attributes(groups)

## $names
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
##
## $class
## [1] "kmeans"
```

```r
write.csv(principalcomps10_12,'principalcomps10_12.csv')
pc2010_12 <- read.csv("principalcomps10_12.csv")

write.csv(groups$cluster,'clusters10_12.csv')
clusters10_12 <- read.csv("clusters10_12.csv")
```

2014 - 2016 dataframe

```r
principalcomps14_16 <- pcdf2014to16$scores[, 1:14]
groups2 <- kmeans(principalcomps14_16, 6)
attributes(groups2)

## $names
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
##
## $class
## [1] "kmeans"

write.csv(principalcomps14_16,'principalcomps14_16.csv')
pc2014_16 <- read.csv("principalcomps14_16.csv")

write.csv(groups2$cluster,'clusters14_16.csv')
clusters14_16 <- read.csv("clusters14_16.csv")
```

2018 - 2020 dataframe

```r
principalcomps18_20 <- pc2018to2020$scores[, 1:14]
groups3 <- kmeans(principalcomps18_20, 6)
attributes(groups3)

## $names
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
##
## $class
## [1] "kmeans"

write.csv(principalcomps18_20,'principalcomps18_20.csv')
pc2018_20 <- read.csv("principalcomps18_20.csv")

write.csv(groups3$cluster,'clusters18_20.csv')
clusters2018_20 <- read.csv("clusters18_20.csv")
```

Categorical data + principal components + clusters

```
# 2010 - 2012 non-numerical dataframe
xyzdf2010to12 = subset(df2010to12, select=-
c(Test.Veh.Displacement..L.,Rated.Horsepower,X..of.Cylinders.and.Rotors,X..of
.Gears,Equivalent.Test.Weight..lbs..,Axle.Ratio, N.V.Ratio,
Test.Fuel.Type.Cd, THC..g.mi., CO..g.mi., CO2..g.mi.,NOx..g.mi., CH4..g.mi.,
N2O..g.mi., RND_ADJ_FE, FE.Bag.1,FE.Bag.2, FE.Bag.3,
Target.Coef.A..lbf.,Target.Coef.B..lbf.mph.,Target.Coef.C..lbf.mph..2.,Set.Co
ef.A..lbf.,Set.Coef.B..lbf.mph.,Set.Coef.C..lbf.mph..2.))

# merging 2010 - 2012 non-numerical dataframe with the principal components
and clusters data
final2010_12 <- cbind(xyzdf2010to12,pc2010_12,clusters10_12)
# deleting the X columns

final2010_12 <- subset(final2010_12, select = -c(X))
final2010_12 <- subset(final2010_12, select = -c(X))
# Viewing a subset of the final data

head(final2010_12)
```

```
##   Year Veh.Mfr.Code Represented.Test.Veh.Make Represented.Test.Veh.Model
## 1 2010          ASX              Aston Martin                         DB9
## 2 2010          ASX              Aston Martin                         DB9
## 3 2010          ASX              Aston Martin                         DB9
## 4 2010          ASX              Aston Martin                         DB9
## 5 2010          ASX              Aston Martin                         DBS
## 6 2010          ASX              Aston Martin                         DBS
##   Vehicle.Type Engine.Code Tested.Transmission.Type Transmission.Lockup.
## 1          Car       AM09/                  Manual                    N
## 2          Car       AM09/                  Manual                    N
## 3          Car       AM09/           Semi-Automatic                   Y
## 4          Car       AM09/           Semi-Automatic                   Y
## 5          Car       AM08/           Semi-Automatic                   Y
## 6          Car       AM08/           Semi-Automatic                   Y
##   Drive.System.Description Transmission.Overdrive.Desc
## 1      2-Wheel Drive, Rear         Top gear ration < 1
## 2      2-Wheel Drive, Rear         Top gear ration < 1
## 3      2-Wheel Drive, Rear         Top gear ration < 1
## 4      2-Wheel Drive, Rear         Top gear ration < 1
## 5      2-Wheel Drive, Rear         Top gear ration < 1
## 6      2-Wheel Drive, Rear         Top gear ration < 1
##   Shift.Indicator.Light.Use.Desc              Test.Procedure.Description
## 1                    Not eqipped Federal fuel 2-day exhaust (w/can load)
## 2                    Not eqipped                                   HWFE
## 3                    Not eqipped Federal fuel 2-day exhaust (w/can load)
## 4                    Not eqipped                                   HWFE
## 5                    Not eqipped Federal fuel 2-day exhaust (w/can load)
## 6                    Not eqipped                                   HWFE
```

```
##   Test.Fuel.Type.Description Test.Category Aftertreatment.Device.Cd
## 1       Tier 2 Cert Gasoline           FTP
## 2       Tier 2 Cert Gasoline           HWY
## 3       Tier 2 Cert Gasoline           FTP                      TWC
## 4       Tier 2 Cert Gasoline           HWY                      TWC
## 5       Tier 2 Cert Gasoline           FTP
## 6       Tier 2 Cert Gasoline           HWY
##   Aftertreatment.Device.Desc Police...Emergency.Vehicle.   Comp.1
Comp.2
## 1                                                      N 4.033146
0.52597185
## 2                                                      N 1.238329
1.80105791
## 3         Three-way catalyst                           N 4.458835
1.41747109
## 4         Three-way catalyst                           N 1.984360
3.18751257
## 5                                                      N 4.064603 -
0.06193001
## 6                                                      N 1.345677
1.43828574
##       Comp.3     Comp.4     Comp.5    Comp.6      Comp.7     Comp.8
Comp.9
## 1  1.1700584 -2.3749371 -0.5759450 0.2566294  0.10587507 0.06775406
0.48211380
## 2  0.2060857 -0.4301523 -0.8608252 0.9170123  0.36524961 0.14836579
0.33115220
## 3 -0.4961915 -2.7173800 -0.2373641 1.4285829 -0.02122240 0.85942092
0.45378394
## 4 -0.1547307 -0.9286134 -0.7237715 1.4391482  0.57763590 0.90801433
0.19191120
## 5  0.8576252 -1.8406626 -0.5974745 0.3493939 -0.04775346 0.03983292
0.13772076
## 6  0.3820259  0.1143713 -0.9833862 0.7395540  0.28480646 0.13797974
0.08043463
##       Comp.10     Comp.11    Comp.12    Comp.13 X.1 x
## 1 -0.49340229 -0.27089327  0.0272832 -1.7285056   1 6
## 2 -0.21496499 -0.27613007 -0.4610970 -1.3960077   2 4
## 3 -0.81378425  1.01559892  0.6353505 -1.3239594   3 6
## 4 -0.45128180  0.64411776  0.4475580 -1.2002508   4 4
## 5 -0.23429232  0.14641256  0.1915930 -1.1069166   5 6
## 6  0.02031231  0.04814919 -0.1686152 -0.8932259   6 3

#2014-2016 non-numerical dataframe
xyzdf2014to16 = subset(df2014to16, select=-
c(Test.Veh.Displacement..L.,Rated.Horsepower,X..of.Cylinders.and.Rotors,X..of
.Gears,Equivalent.Test.Weight..lbs..,Axle.Ratio, N.V.Ratio,
Test.Fuel.Type.Cd, THC..g.mi., CO..g.mi., CO2..g.mi.,NOx..g.mi., CH4..g.mi.,
N2O..g.mi., RND_ADJ_FE, FE.Bag.1,FE.Bag.2, FE.Bag.3,
Target.Coef.A..lbf.,Target.Coef.B..lbf.mph.,Target.Coef.C..lbf.mph..2.,Set.Co
```

```
ef.A..lbf.,Set.Coef.B..lbf.mph.,Set.Coef.C..lbf.mph..2.))
final2014_16 <- cbind(xyzdf2014to16,pc2014_16,clusters14_16)

# deleting the X columns
final2014_16 <- subset(final2014_16, select = -c(X))

final2014_16 <- subset(final2014_16, select = -c(X))
# Viewing a subset of the final data
head(final2014_16)
```

```
##   Year Veh.Mfr.Code Represented.Test.Veh.Make Represented.Test.Veh.Model
## 1 2014          ASX              Aston Martin                         DB9
## 2 2014          ASX              Aston Martin                         DB9
## 3 2014          ASX              Aston Martin                   V8 VANTAGE
## 4 2014          ASX              Aston Martin                   V8 VANTAGE
## 5 2014          ASX              Aston Martin                 V8 VANTAGE S
## 6 2014          ASX              Aston Martin                 V8 VANTAGE S
##   Vehicle.Type Engine.Code Tested.Transmission.Type Transmission.Lockup.
## 1          Car       AM11/           Semi-Automatic                    Y
## 2          Car       AM11/           Semi-Automatic                    Y
## 3          Car       AM14/                   Manual                    N
## 4          Car       AM14/                   Manual                    N
## 5          Car       AM15/         Automated Manual                    Y
## 6          Car       AM15/         Automated Manual                    Y
##   Drive.System.Description Transmission.Overdrive.Desc
## 1      2-Wheel Drive, Rear         Top gear ration < 1
## 2      2-Wheel Drive, Rear         Top gear ration < 1
## 3      2-Wheel Drive, Rear         Top gear ration < 1
## 4      2-Wheel Drive, Rear         Top gear ration < 1
## 5      2-Wheel Drive, Rear         Top gear ration < 1
## 6      2-Wheel Drive, Rear         Top gear ration < 1
##   Shift.Indicator.Light.Use.Desc              Test.Procedure.Description
## 1                    Not eqipped Federal fuel 2-day exhaust (w/can load)
## 2                    Not eqipped                                    HWFE
## 3                    Not eqipped Federal fuel 2-day exhaust (w/can load)
## 4                    Not eqipped                                    HWFE
## 5                    Not eqipped Federal fuel 2-day exhaust (w/can load)
## 6                    Not eqipped                                    HWFE
##   Test.Fuel.Type.Description Test.Category Aftertreatment.Device.Cd
## 1       Tier 2 Cert Gasoline           FTP                      TWC
## 2       Tier 2 Cert Gasoline           HWY                      TWC
## 3       Tier 2 Cert Gasoline           FTP                      TWC
## 4       Tier 2 Cert Gasoline           HWY                      TWC
## 5       Tier 2 Cert Gasoline           FTP                      TWC
## 6       Tier 2 Cert Gasoline           HWY                      TWC
##   Aftertreatment.Device.Desc Police...Emergency.Vehicle.   Comp.1
Comp.2
## 1         Three-way catalyst                           N 4.552370
2.333922542
## 2         Three-way catalyst                           N 2.311290
```

```
1.246805538
## 3         Three-way catalyst                            N 3.356351
1.052301271
## 4         Three-way catalyst                            N 1.348085 -
0.085966526
## 5         Three-way catalyst                            N 3.170849
0.979820581
## 6         Three-way catalyst                            N 1.061324
0.003050348
##       Comp.3     Comp.4     Comp.5      Comp.6      Comp.7      Comp.8
## 1 -0.8997183 -0.3389642  1.5791618  0.62637229  0.03638080  0.02191681
## 2 -0.4418487 -0.1090729  3.0365107  0.06520599  0.14957520 -0.12672777
## 3  0.2173873 -0.5107763 -0.5190728  0.72555816 -0.10801233 -0.21737262
## 4  0.1924083  0.3096426  0.6508395  0.61593726  0.06305181 -0.41733349
## 5  0.4853718 -0.9102004 -0.3139763  0.30083265 -0.09255195 -0.03705374
## 6  0.8957075 -0.7350050  0.9706789 -0.28790072  0.07662031 -0.18458953
##          Comp.9     Comp.10     Comp.11   Comp.12     Comp.13    Comp.14
X.1 x
## 1 -0.2447297122  0.13581415 -0.14992024 0.4464896 -0.39502992  0.5575063
1 3
## 2  0.0159382456 -0.04817822 -0.37796279 0.4042980 -0.23905975  1.0244442
2 3
## 3 -0.0732604113 -0.05690131  0.18503856 0.5916335 -0.22276339  0.1400246
3 3
## 4 -0.1384509070 -0.09364818 -0.19127067 1.3628142 -0.33313203 -0.6286958
4 3
## 5 -0.1832832813  0.12792699  0.12487832 0.4038190 -0.14493270  0.7553137
5 3
## 6  0.0006317764 -0.04269855 -0.09138531 0.4469027  0.01436145  1.0045443
6 5
```

```
#2018-2020 non-numerical dataframe
xyzdf2018to20 = subset(df2018to20, select=-
c(Test.Veh.Displacement..L.,Rated.Horsepower,X..of.Cylinders.and.Rotors,X..of
.Gears,Equivalent.Test.Weight..lbs..,Axle.Ratio, N.V.Ratio,
Test.Fuel.Type.Cd, THC..g.mi., CO..g.mi., CO2..g.mi.,NOx..g.mi., CH4..g.mi.,
N2O..g.mi., RND_ADJ_FE, FE.Bag.1,FE.Bag.2, FE.Bag.3,
Target.Coef.A..lbf.,Target.Coef.B..lbf.mph.,Target.Coef.C..lbf.mph..2.,Set.Co
ef.A..lbf.,Set.Coef.B..lbf.mph.,Set.Coef.C..lbf.mph..2.))
final2018_20 <- cbind(xyzdf2018to20,pc2018_20,clusters2018_20)

# deleting the X columns
final2018_20 <- subset(final2018_20, select = -c(X))
final2018_20 <- subset(final2018_20, select = -c(X))

# Viewing a subset of the final data
head(final2018_20)
```

```
##   Model.Year Year Veh.Mfr.Code Represented.Test.Veh.Make
## 1       2018 2018          ASX               Aston Martin
```

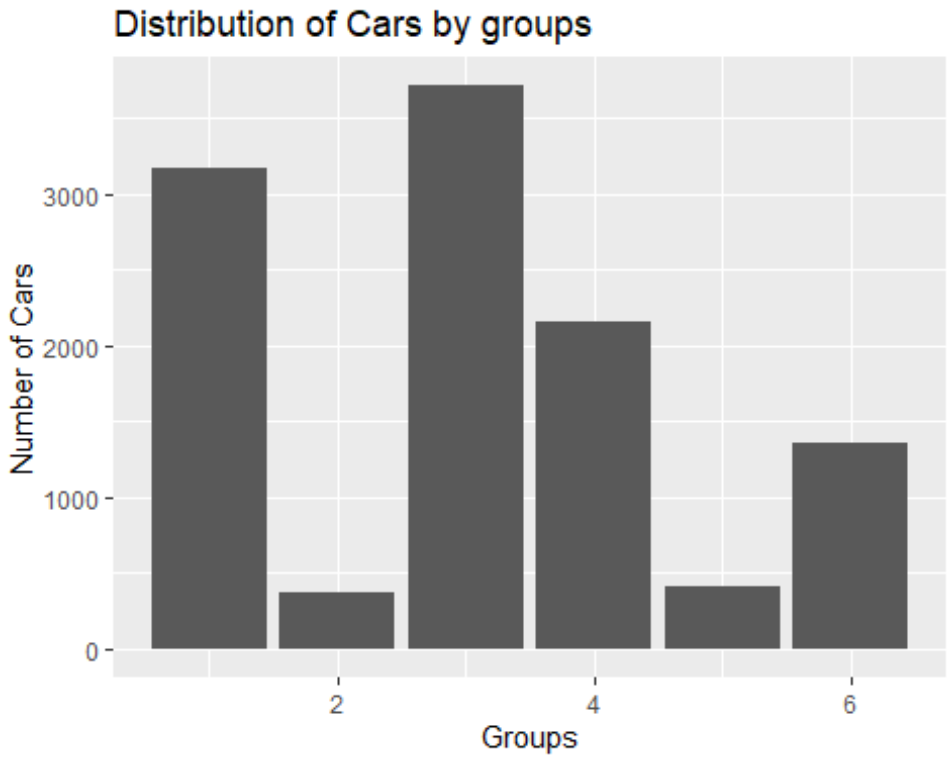
```
## 2       2018 2018          ASX              Aston Martin
## 3       2018 2018          ASX              Aston Martin
## 4       2018 2018          ASX              Aston Martin
## 5       2018 2018          ASX              Aston Martin
## 6       2018 2018          ASX              Aston Martin
##   Represented.Test.Veh.Model Vehicle.Type Engine.Code
Tested.Transmission.Type
## 1                      DB11          Car       AE31/           Semi-
Automatic
## 2                      DB11          Car       AE31/           Semi-
Automatic
## 3                   DB11 V8          Car      177950           Semi-
Automatic
## 4                   DB11 V8          Car      177950           Semi-
Automatic
## 5                  Rapide S          Car       AM29/           Semi-
Automatic
## 6                  Rapide S          Car       AM29/           Semi-
Automatic
##   Transmission.Lockup. Drive.System.Description
Transmission.Overdrive.Desc
## 1                    Y      2-Wheel Drive, Rear          Top gear ratio <
1
## 2                    Y      2-Wheel Drive, Rear          Top gear ratio <
1
## 3                    Y      2-Wheel Drive, Rear          Top gear ratio <
1
## 4                    Y      2-Wheel Drive, Rear          Top gear ratio <
1
## 5                    Y      2-Wheel Drive, Rear          Top gear ratio <
1
## 6                    Y      2-Wheel Drive, Rear          Top gear ratio <
1
##   Shift.Indicator.Light.Use.Desc              Test.Procedure.Description
## 1                    Not eqipped Federal fuel 2-day exhaust (w/can load)
## 2                    Not eqipped                                     HWFE
## 3                    Not eqipped Federal fuel 2-day exhaust (w/can load)
## 4                    Not eqipped                                     HWFE
## 5                    Not eqipped Federal fuel 2-day exhaust (w/can load)
## 6                    Not eqipped                                     HWFE
##   Test.Fuel.Type.Description Test.Category Aftertreatment.Device.Cd
## 1       Tier 2 Cert Gasoline           FTP                      TWC
## 2       Tier 2 Cert Gasoline           HWY                      TWC
## 3       Tier 2 Cert Gasoline           FTP                      TWC
## 4       Tier 2 Cert Gasoline           HWY                      TWC
## 5       Tier 2 Cert Gasoline           FTP                      TWC
## 6       Tier 2 Cert Gasoline           HWY                      TWC
##   Aftertreatment.Device.Desc Police...Emergency.Vehicle.   Comp.1
Comp.2
## 1         Three-way catalyst                           N 3.704372 -
```

```
0.49915515
## 2         Three-way catalyst                            N 2.603093 -
0.91346951
## 3         Three-way catalyst                            N 2.291884 -
0.49578040
## 4         Three-way catalyst                            N 1.464083 -
0.50666043
## 5         Three-way catalyst                            N 3.315479
0.09206583
## 6         Three-way catalyst                            N 2.081654 -
0.42916813
##        Comp.3     Comp.4     Comp.5      Comp.6     Comp.7    Comp.8
Comp.9
## 1 -0.31107524  0.5849613 -0.8845798 -0.64049421 -1.0196591 0.2172535 -
0.2818463
## 2 -1.21059652  0.8870872 -0.5147095 -0.22096302 -1.0221872 0.2335173 -
0.3339871
## 3 -0.39199819 -0.1454178 -0.2293182 -0.46060882 -0.3507771 0.2086055 -
0.1445008
## 4 -1.43300825  0.3269933  0.1606948  0.05218672 -0.5699829 0.1178649 -
0.1840725
## 5  0.07463202  0.5284699 -3.9560573  2.93352488 -2.9044878 0.4089148 -
0.8613364
## 6 -0.80331163  0.7424655 -3.5595517  3.34176741 -2.7398640 0.5115114 -
0.8718272
##      Comp.10      Comp.11     Comp.12    Comp.13    Comp.14 x
## 1 -0.8067005 -0.473867158  0.44785743  0.8339543  1.0031463 5
## 2 -0.7657143 -0.189709551  0.70908323  1.0320090  0.7100682 5
## 3 -0.4088161 -0.190715593  0.08125029  0.3085998  0.9139942 5
## 4 -0.3688699  0.007782901  0.31006237  0.7703169  0.3185616 3
## 5 -2.9783115  1.908106479 -1.29350321 -0.8161684 -0.0662340 5
## 6 -3.0146426  2.215951393 -1.02818146 -0.7270795 -0.2268597 5
```

Descriptive analytics for each group

Time Period 2010-2012

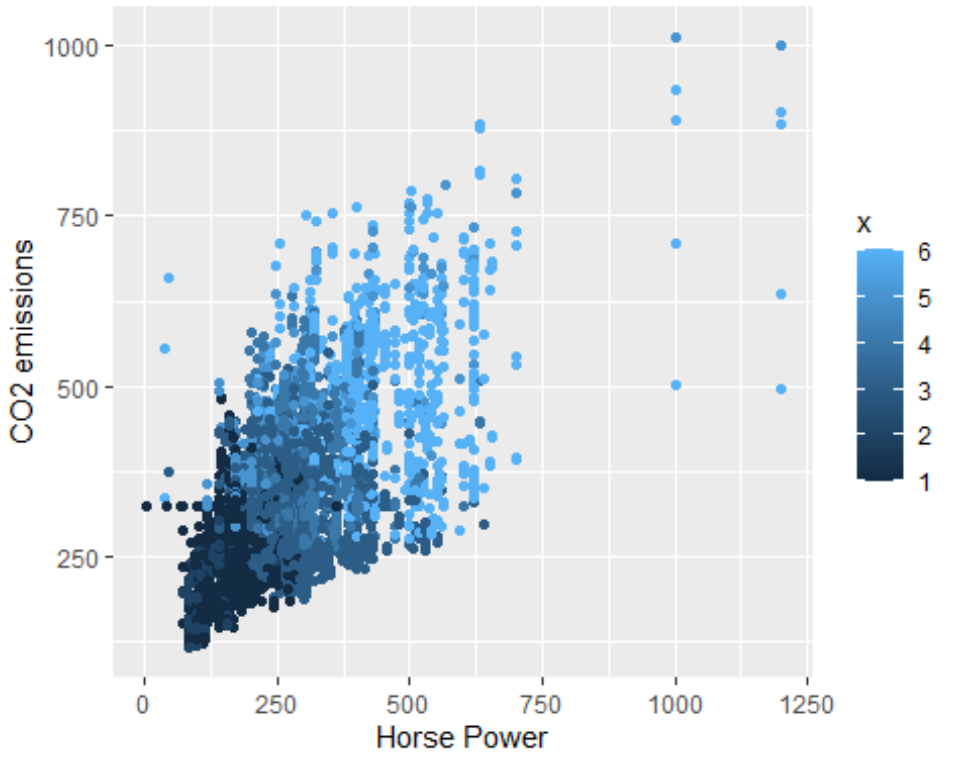
```
grouping2010to12 <- cbind(df2010to12,clusters10_12)
grouping2010to12 <- subset(grouping2010to12, select = -c(X))
grouping010to12 <- subset(grouping2010to12, select = -c(X))

ggplot(grouping010to12, aes(x)) +
  geom_bar() + xlab('Groups') + ylab('Number of Cars') +
   ggtitle('Distribution of Cars by groups')
```




```
ggplot(data=grouping010to12, aes(x=grouping010to12$Rated.Horsepower,
y=grouping010to12$CO2..g.mi., color=x)) +geom_point()+labs(x= "Horse Power",
y="CO2 emissions")
```

```
## Warning: Use of `grouping010to12$Rated.Horsepower` is discouraged. Use
## `Rated.Horsepower` instead.
```

```
## Warning: Use of `grouping010to12$CO2..g.mi.` is discouraged. Use
`CO2..g.mi.`
## instead.
```

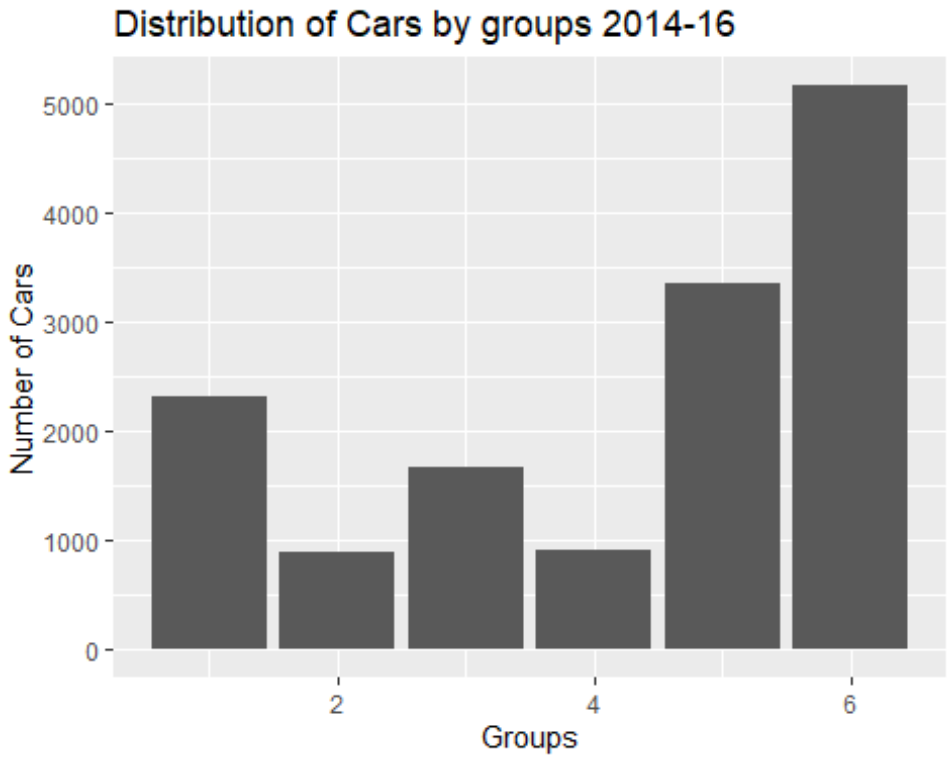


The above histogram for the years 2010 to 2012 shows that the highest number of cars are in group 2 and overall, group 1,2, and 3 has the highest number of cars.

The trend of the scatter plot displays that higher the horse power results in higher CO2 emissions which in real world would make sense since higher horse power means a more powerful engine which would lead to a higher overall emission of gases leading to higher emission of CO2 and for the years 2010 to 2012 the maximum data point are between around 125 horsepower to almost closer to 500 horsepowerand the highest being close to the range of 1250 horsepower with a alarmingly high CO2 emission.

Time Period 2014-2016

```
grouping2014to16 <- cbind(df2014to16,clusters14_16)
grouping2014to16 <- subset(grouping2014to16, select = -c(X))
grouping014to16 <- subset(grouping2014to16, select = -c(X))

ggplot(grouping014to16, aes(x)) +
  geom_bar() + xlab('Groups') + ylab('Number of Cars') +
   ggtitle('Distribution of Cars by groups 2014-16')
```




```
ggplot(data=grouping014to16, aes(x=grouping014to16$Rated.Horsepower,
y=grouping014to16$CO2..g.mi., color=x)) +geom_point()+labs(x= "Horse Power",
y="CO2 emissions")
```

```
## Warning: Use of `grouping014to16$Rated.Horsepower` is discouraged. Use
## `Rated.Horsepower` instead.
```

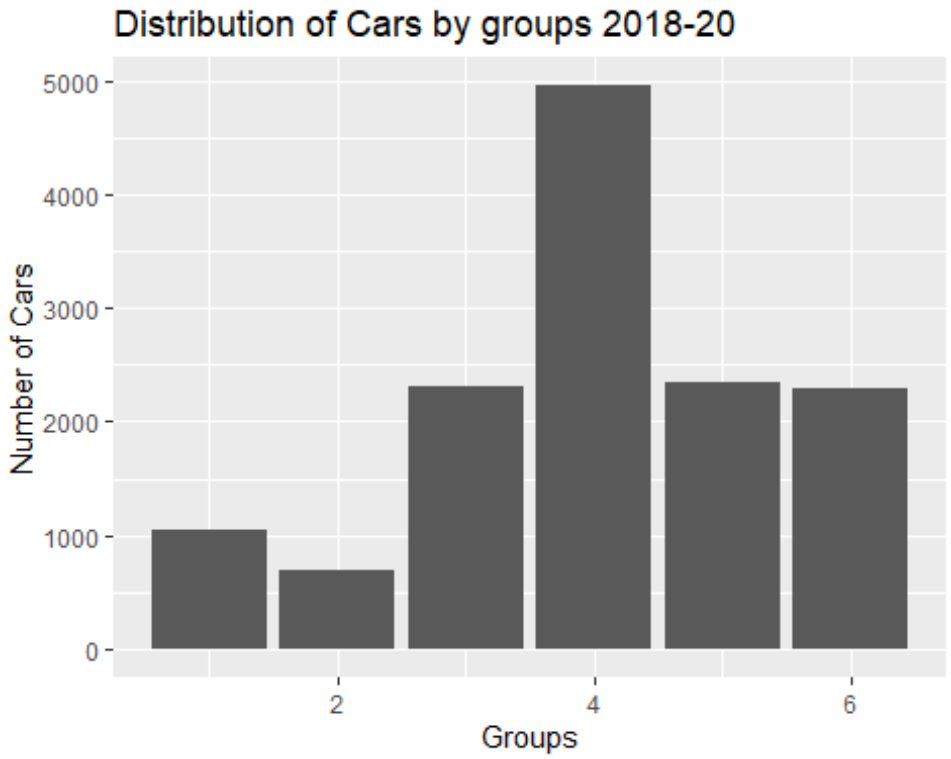
```
## Warning: Use of `grouping014to16$CO2..g.mi.` is discouraged. Use
`CO2..g.mi.`
## instead.
```

In this case, for the years 2014 to 2016 the histogram displays the highest number of vehicles in group 3. This histogram almost looks like a bell curve and the higher number of cars in group 3 and 4.

Here, if you take a look at the scatterplot, the trend is logically sound wherein, the CO2 emissions increases if the horsepower of the vehicle is high.The maximum amount of data points are in the range of 125 horsepower to close to 600 horsepower with the highest being at closer to 1250 with a very high CO2 emission rate.

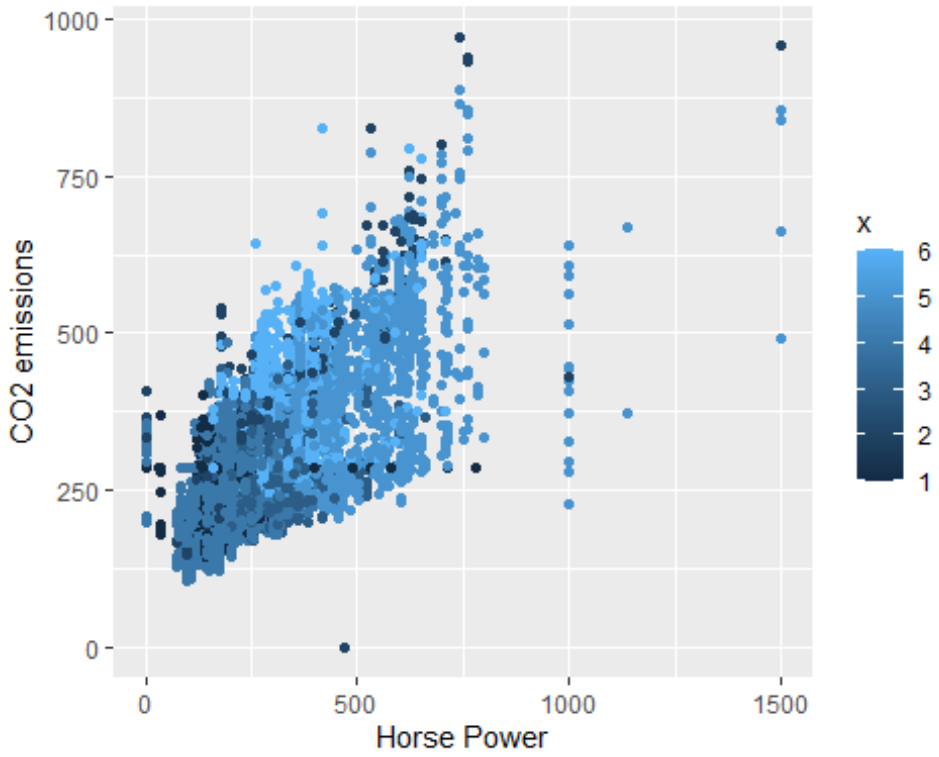
Time Period 2018-2020

```
grouping2018to20 <- cbind(df2018to20,clusters2018_20)

grouping018to20 <- subset(grouping2018to20, select = -c(X))

ggplot(grouping018to20, aes(x)) +
  geom_bar() + xlab('Groups') + ylab('Number of Cars') +
   ggtitle('Distribution of Cars by groups 2018-20')
```



The histogram here displays the highest number of cars in group 4 with a uniform level for group 5 and 6 and very low number of cars for group 3. Where as, group 2 and 1 are relatively high.

```
ggplot(data=grouping018to20, aes(x=grouping018to20$Rated.Horsepower,
y=grouping018to20$CO2..g.mi., color=x)) +geom_point()+labs(x= "Horse Power",
y="CO2 emissions")

## Warning: Use of `grouping018to20$Rated.Horsepower` is discouraged. Use
## `Rated.Horsepower` instead.

## Warning: Use of `grouping018to20$CO2..g.mi.` is discouraged. Use
`CO2..g.mi.`
## instead.
```



For the dataset for years 2018 to 2020, if you take a look at the scatterplot, the trend depicts that as the horsepower of the cars increase the CO2 emissions increase too.The maximum amount of data points are in the range of 110 horsepower to close to 550 horsepower with the highest being at at 1500, highest among all the three datasets with a very high CO2 emission rate.This leads to the conclusion that in the most recent years the CO2 emssions have alarmingly increased.

Predictive modelling

2010-2012 Period

```
modelA <- lm(df2010to12$RND_ADJ_FE~Tested.Transmission.Type, data =
df2010to12)
modelA

##
## Call:
## lm(formula = df2010to12$RND_ADJ_FE ~ Tested.Transmission.Type,
##     data = df2010to12)
##
## Coefficients:
##
(Intercept)
##
22.645
```

```
##
Tested.Transmission.TypeAutomatic
##
3.870
##
Tested.Transmission.TypeContinuously Variable
##
16.625
##
Tested.Transmission.TypeManual
##
8.431
##
Tested.Transmission.TypeOther
##
27.617
##
Tested.Transmission.TypeSemi-Automatic
##
4.893
##          Tested.Transmission.TypeSelectable Continuously Variable (e.g.
CVT with paddles)
##
17.638
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated
Manual with paddles)
##
9.155
```

2014-2016 Period

```
modelB <- lm(df2014to16$RND_ADJ_FE~Tested.Transmission.Type, data =
df2014to16)
modelB

##
## Call:
## lm(formula = df2014to16$RND_ADJ_FE ~ Tested.Transmission.Type,
##     data = df2014to16)
##
## Coefficients:
##
(Intercept)
##
31.8828
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated
Manual with paddles)
##
-0.9866
##
Tested.Transmission.TypeAutomatic
```

```
##
-1.5655
##
Tested.Transmission.TypeContinuously Variable
##
12.6882
##
Tested.Transmission.TypeManual
##
2.5235
##
Tested.Transmission.TypeOther
##
34.5386
##          Tested.Transmission.TypeSelectable Continuously Variable (e.g.
CVT with paddles)
##
8.4595
##
Tested.Transmission.TypeSemi-Automatic
##
17.7721
```

2018-2020 Period

```
modelC <- lm(df2014to16$RND_ADJ_FE~Tested.Transmission.Type, data =
df2014to16)
modelC

##
## Call:
## lm(formula = df2014to16$RND_ADJ_FE ~ Tested.Transmission.Type,
##     data = df2014to16)
##
## Coefficients:
##
(Intercept)
##
31.8828
## Tested.Transmission.TypeAutomated Manual- Selectable (e.g. Automated
Manual with paddles)
##
-0.9866
##
Tested.Transmission.TypeAutomatic
##
-1.5655
##
Tested.Transmission.TypeContinuously Variable
##
12.6882
```

```
##
Tested.Transmission.TypeManual
##
2.5235
##
Tested.Transmission.TypeOther
##
34.5386
##          Tested.Transmission.TypeSelectable Continuously Variable (e.g.
CVT with paddles)
##
8.4595
##
Tested.Transmission.TypeSemi-Automatic
##
17.7721
```