Exploratory data analysis on MPG dataset

Meghna Bajoria

2022-09-12

Loading the data

```
if(!require(ggplot2))
  install.packages("ggplot2",repos = "http://cran.us.r-project.org")

## Loading required package: ggplot2

mpg <- ggplot2::mpg
```

Checks to determine quality of data

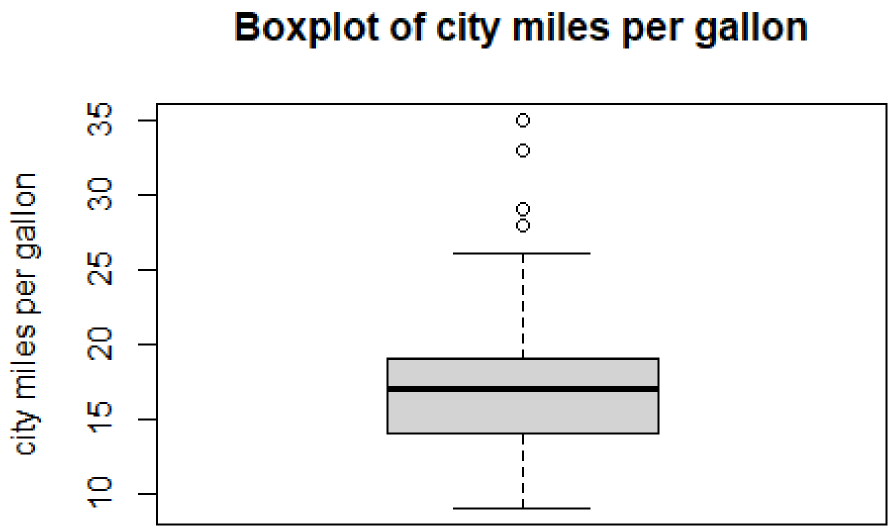
1. Missing data

```
sum(is.na(mpg))

## [1] 0
```

There are no missing values in the dataset.

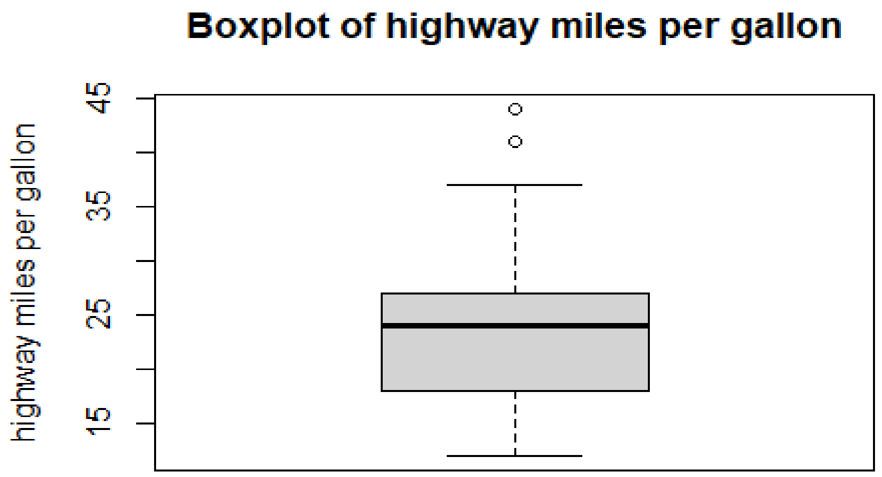
2. Finding outliers using boxplot

```
boxplot(mpg$cty,
  ylab = "city miles per gallon",
  main = "Boxplot of city miles per gallon")
```



We can see 4 outliers in city miles per gallon.

```
boxplot(mpg$hwy,
  ylab = "highway miles per gallon",
  main = "Boxplot of highway miles per gallon")
```



We can see 2 outliers in highway miles per gallon.

```
sum(duplicated(mpg))

## [1] 9
```

There are 9 duplicate rows.

3. Description of the data

```
dimension <- dim(mpg)
dimension

## [1] 234  11
```

Mpg dataset has 11 variables and 234 observations.

```
names(mpg)

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

The variable names in mpg dataset are shown above.

There are 3 numeric variables in the mpg dataset:
1. cty
2. hwy
3. displ

There are 8 categorical variables:
1. manufacturer
2. model
3. year
4. cyl
5. trans
6. drv
7. fl
8. class

Description of variables

1. manufacturer - name of car manufacturer

2. model - model name

3. year - year of manufacturing

4. cyl - number of cylinders

5. trans - type of transmission

6. drv - drive type

7. fl - fuel type

8. class - vehicle class

9. cty - city miles per gallon

10. hwy - highway miles per gallon

11. displ - engine displacement in litres

3. Summary statistics

```
summary(mpg)
```

```
##  manufacturer          model               displ            year     
##  Length:234         Length:234         Min.   :1.600   Min.   :1999  
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999  
##  Mode  :character   Mode  :character   Median :3.300   Median :2004  
##                                        Mean   :3.472   Mean   :2004  
##                                        3rd Qu.:4.600   3rd Qu.:2008  
##                                        Max.   :7.000   Max.   :2008  
##       cyl           trans               drv                 cty       
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00  
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00  
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00  
##  Mean   :5.889                                         Mean   :16.86  
##  3rd Qu.:8.000                                         3rd Qu.:19.00  
##  Max.   :8.000                                         Max.   :35.00  
##       hwy             fl               class          
##  Min.   :12.00   Length:234         Length:234        
##  1st Qu.:18.00   Class :character   Class :character  
##  Median :24.00   Mode  :character   Mode  :character  
##  Mean   :23.44                                        
##  3rd Qu.:27.00                                        
##  Max.   :44.00                                        
```

4. Summary statistics by grouping on categorical variable

We will be calculating summary statistics by grouping on categorical variable class.

```
tapply(mpg$cty, mpg$class, summary)
```

```
## $`2seater`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    15.0    15.0    15.0    15.4    16.0    16.0 
## 
```

```
## $compact
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   15.00   18.00   20.00   20.13   21.00   33.00 
## 
## $midsize
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   15.00   18.00   18.00   18.76   21.00   23.00 
## 
## $minivan
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   11.00   15.50   16.00   15.82   17.00   18.00 
## 
## $pickup
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       9      11      13      13      14      17 
## 
## $subcompact
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   14.00   17.00   19.00   20.37   23.50   35.00 
## 
## $suv
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.00   12.00   13.00   13.50   14.75   20.00
```

```
var(mpg$cty, y=NULL, na.rm = TRUE)
```

```
## [1] 18.11307
```

```
sd(mpg$cty, na.rm = TRUE)
```

```
## [1] 4.255946
```

```
range(mpg$cty,na.rm = TRUE)
```

```
## [1]  9 35
```

```
diff(range(mpg$cty,na.rm = TRUE))
```

```
## [1] 26
```

Here, pickup and SUV cars have the lowest city miles per gallon i.e 9 miles per gallon and subCompact car has the highest city miles per gallon.

```
tapply(mpg$hwy, mpg$class, summary)
```

```
## $`2seater`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    23.0    24.0    25.0    24.8    26.0    26.0 
## 
## $compact
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    23.0    26.0    27.0    28.3    29.0    44.0
```

```
## 
## $midsize
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   23.00   26.00   27.00   27.29   29.00   32.00 
## 
## $minivan
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   17.00   22.00   23.00   22.36   24.00   24.00 
## 
## $pickup
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   12.00   16.00   17.00   16.88   18.00   22.00 
## 
## $subcompact
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   20.00   24.50   26.00   28.14   30.50   44.00 
## 
## $suv
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   12.00   17.00   17.50   18.13   19.00   27.00
```

```
var(mpg$hwy, y=NULL, na.rm = TRUE)
```

```
## [1] 35.45778
```

```
sd(mpg$hwy, na.rm = TRUE)
```

```
## [1] 5.954643
```

```
range(mpg$hwy,na.rm = TRUE)
```

```
## [1] 12 44
```

```
diff(range(mpg$hwy,na.rm = TRUE))
```

```
## [1] 32
```

Here, pickup and SUV cars have the lowest highway miles per gallon i.e 12 miles per gallon. Compact and subCompact car has the highest highway miles per gallon.

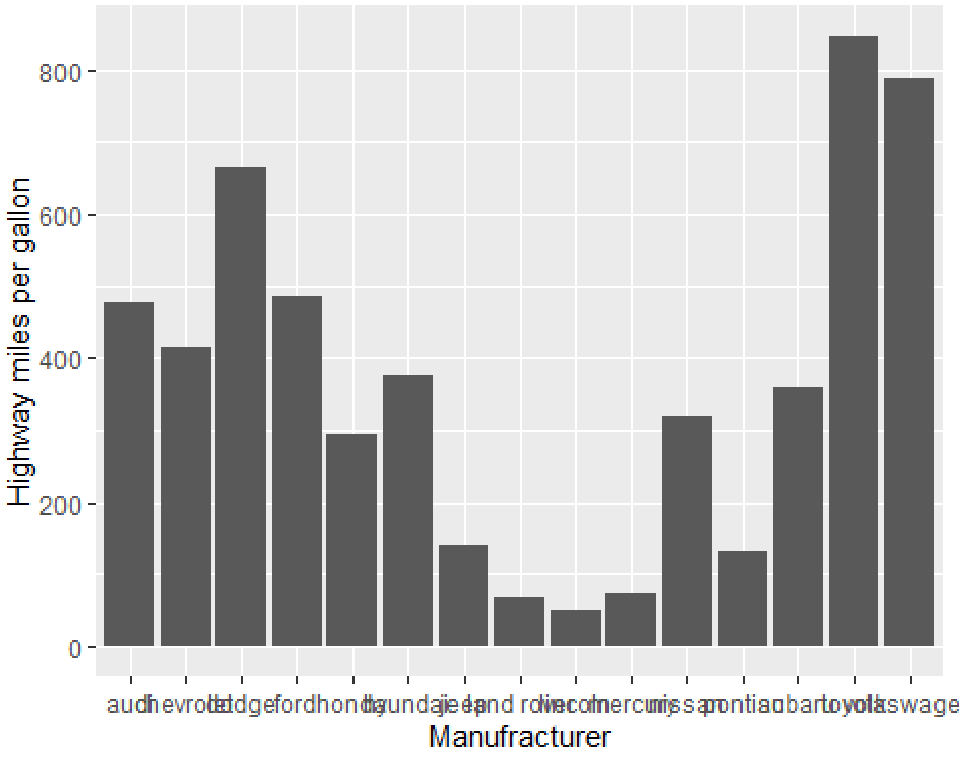
```
tapply(mpg$displ, mpg$class, summary)
```

```
## $`2seater`
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    5.70    5.70    6.20    6.16    6.20    7.00 
## 
## $compact
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.800   2.000   2.200   2.326   2.800   3.300 
## 
## $midsize
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.800   2.400   2.800   2.922   3.500   5.300
## 
## $minivan
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.400   3.300   3.300   3.391   3.800   4.000 
## 
## $pickup
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.700   3.900   4.700   4.418   4.700   5.900 
## 
## $subcompact
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.60    1.90    2.20    2.66    3.25    5.40 
## 
## $suv
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.500   4.000   4.650   4.456   5.300   6.500

var(mpg$displ, y=NULL, na.rm = TRUE)

## [1] 1.669158

sd(mpg$displ, na.rm = TRUE)

## [1] 1.291959

range(mpg$displ,na.rm = TRUE)

## [1] 1.6 7.0

diff(range(mpg$displ,na.rm = TRUE))

## [1] 5.4
```

Here, subCompact car has the lowest displacement i.e 1.6 and 2seater car has the highest displacement.

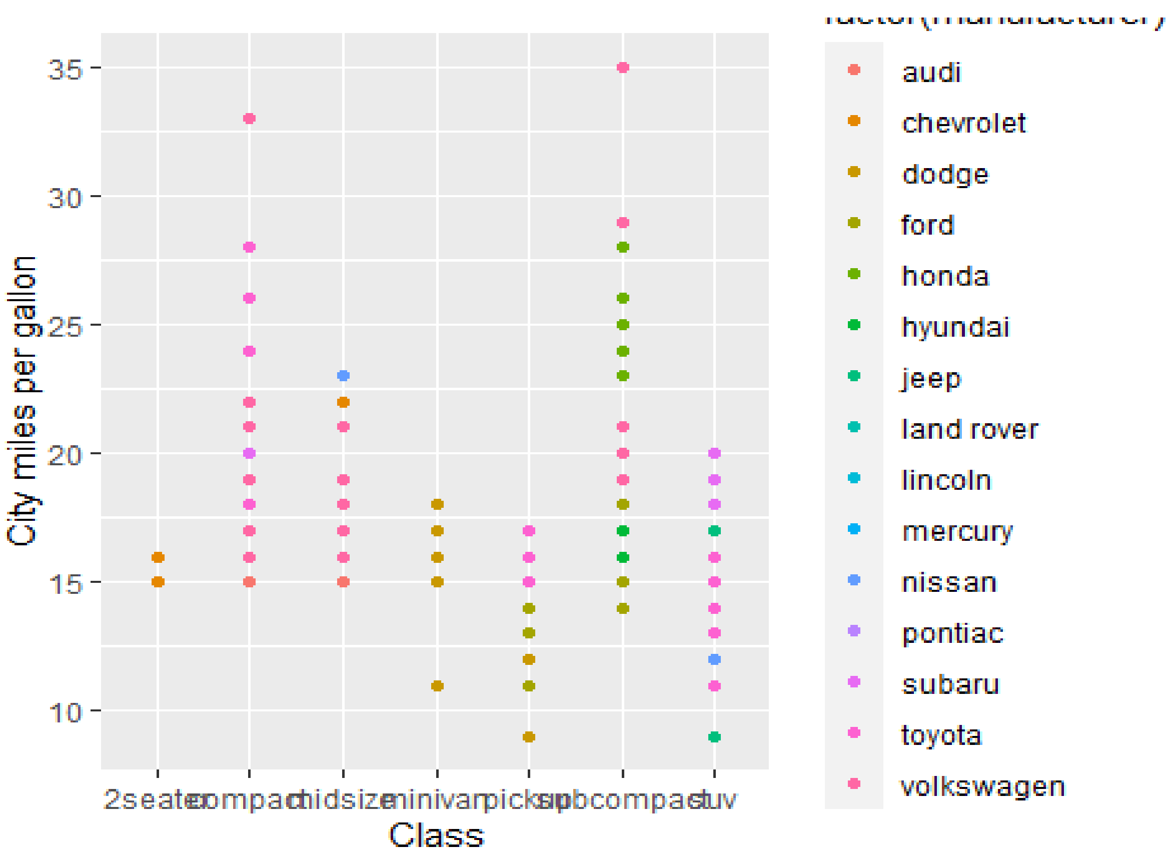
5. Visualizing relationship between variables

```
plot1=ggplot(mpg, aes(x = manufacturer, y = hwy)) +
  geom_bar(stat = "identity") +
  xlab("Manufracturer") +
  ylab("Highway miles per gallon")
plot1
```



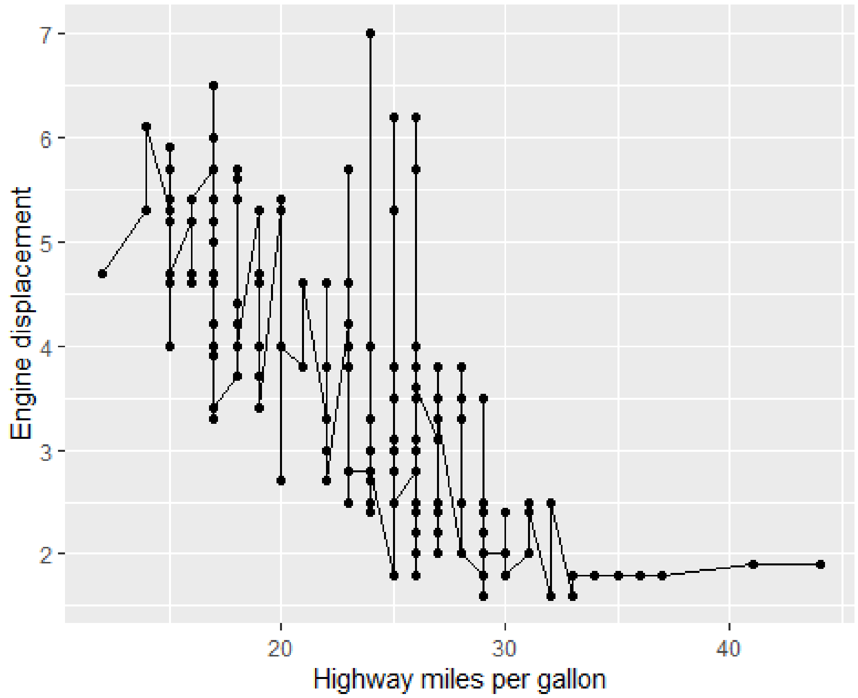
The first plot(above) shows the highway miles per gallon for each manufacturer. These variables show a general relationship between highway miles per gallon across all cars on the basis of the manifacturer. It can be concluded that Toyota has the best highway miles per gallon across all car models while Lincoln has the worst.

```
plot2 = ggplot(mpg, aes(class, cty, color = factor(manufacturer))) +
     geom_point() +
     xlab("Class") +
     ylab("City miles per gallon")
plot2
```



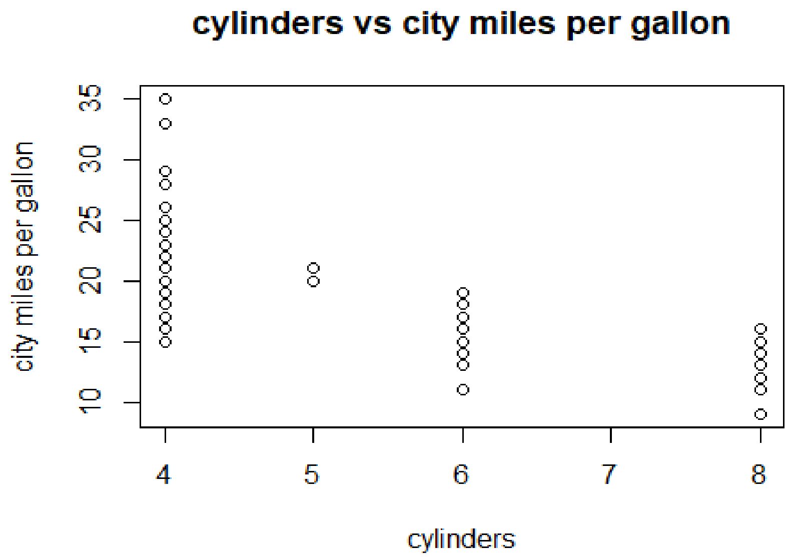
The second plot (above) shows every car's city miles per gallon vs its class. Also, each dot has been color coded to match its manufacturer.These variables were chosen to show that smaller cars like compact and subcompact give better miles per gallon than large cars like SUV and pickups.

```r
plot3 = ggplot(mpg, aes(hwy, displ)) +
  geom_point() +
  geom_line() +
  xlab("Highway miles per gallon") +
  ylab("Engine displacement")
plot3
```



The third graph (above) shows a decreasing trend between highway miles per gallon and engine displacement. It can be concluded that engine power decreases with increase in miles per gallon.

```
plot(x = mpg$cyl, y = mpg$cty,
    xlab = "cylinders",
    ylab = "city miles per gallon",
    main = "cylinders vs city miles per gallon"
)
```



The graph (above) shows that cars with 4 cylinders achieve the highest city miles per gallon.

Summary

To conclude we can say that the highway miles per gallon is highest with manufacturer as Toyota. Also, compact and subcompact have the best city miles per gallon. This helps us in inferring that the fuel economy and efficiency of the engine is good in these cars. We can also conclude that number of cylinders also play a crucial role in providing high city miles per gallon. Analysis shows that cars with 4 cylinders have the highest city miles per gallon.