EDA

Kusum

2022-10-14

## EDA asignment

This is my EDA assignment

library(tidyverse)  
library(dplyr)  
glimpse(mpg)

## Rows: 234  
## Columns: 11  
## $ manufacturer <chr> "audi", "audi", "audi", "audi", "audi", "audi", "audi", "…  
## $ model <chr> "a4", "a4", "a4", "a4", "a4", "a4", "a4", "a4 quattro", "…  
## $ displ <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1.8, 1.8, 2.0, 2.0, 2.…  
## $ year <int> 1999, 1999, 2008, 2008, 1999, 1999, 2008, 1999, 1999, 200…  
## $ cyl <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 8, 8, …  
## $ trans <chr> "auto(l5)", "manual(m5)", "manual(m6)", "auto(av)", "auto…  
## $ drv <chr> "f", "f", "f", "f", "f", "f", "f", "4", "4", "4", "4", "4…  
## $ cty <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 20, 19, 15, 17, 17, 1…  
## $ hwy <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 28, 27, 25, 25, 25, 2…  
## $ fl <chr> "p", "p", "p", "p", "p", "p", "p", "p", "p", "p", "p", "p…  
## $ class <chr> "compact", "compact", "compact", "compact", "compact", "c…

### A) Quality of the data

1. Missing values?

z1 <- mpg  
sum(is.na(z1))

## [1] 0

There are no missing values in the dataset

1. Outliers

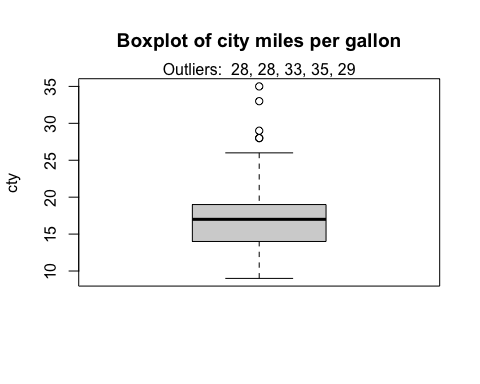
z1 <- mpg  
boxplot(z1$cty, ylab="cty",  
 main = "Boxplot of city miles per gallon")  
boxplot.stats(z1$cty)$out

## [1] 28 28 33 35 29

out <- boxplot.stats(z1$cty)$out  
out\_ind <- which(z1$cty %in% c(out))  
out\_ind

## [1] 100 197 213 222 223

mtext(paste("Outliers: ", paste(out, collapse = ", ")))



The outliers for cty were considered and there are as follows

a)Five points are considered as outliers:- 28,28, 33, 35 & 29

b)They are present in rows 100, 197, 213, 222, 223

### B) Description of the data

1. How big is it?

nrow(mpg)

## [1] 234

ncol(mpg)

## [1] 11

There are 234 observations and 11 variables in the dataset

1. How many numeric variables?

mpg %>%  
 count(displ, cty, hwy)

## # A tibble: 159 × 4  
## displ cty hwy n  
## <dbl> <int> <int> <int>  
## 1 1.6 23 29 1  
## 2 1.6 24 32 2  
## 3 1.6 25 32 1  
## 4 1.6 28 33 1  
## 5 1.8 16 25 1  
## 6 1.8 18 26 1  
## 7 1.8 18 29 2  
## 8 1.8 21 29 2  
## 9 1.8 24 30 1  
## 10 1.8 24 33 1  
## # … with 149 more rows

There are 3 numeric variables: displ, cty and hwy

1. How many categorical variables?

mpg %>%  
 count(manufacturer,model,year,cyl,trans,drv,fl,class)

## # A tibble: 205 × 9  
## manufacturer model year cyl trans drv fl class n  
## <chr> <chr> <int> <int> <chr> <chr> <chr> <chr> <int>  
## 1 audi a4 1999 4 auto(l5) f p compact 1  
## 2 audi a4 1999 4 manual(m5) f p compact 1  
## 3 audi a4 1999 6 auto(l5) f p compact 1  
## 4 audi a4 1999 6 manual(m5) f p compact 1  
## 5 audi a4 2008 4 auto(av) f p compact 1  
## 6 audi a4 2008 4 manual(m6) f p compact 1  
## 7 audi a4 2008 6 auto(av) f p compact 1  
## 8 audi a4 quattro 1999 4 auto(l5) 4 p compact 1  
## 9 audi a4 quattro 1999 4 manual(m5) 4 p compact 1  
## 10 audi a4 quattro 1999 6 auto(l5) 4 p compact 1  
## # … with 195 more rows

The dataset has 8 categorical variables: manufacturer, model, year, cyl, trans, drv, fl and class

1. Description of the variables

manufacturer - manufacturer name

model - model name

displ - engine displacement, in litres

year - year of manufacture

cyl - number of cylinders

trans - type of transmission

drv - the type of drive train, where f = front-wheel drive, r = rear wheel drive, 4 = 4wd

cty - city miles per gallon

hwy - highway miles per gallon

fl - fuel type

class - “type” of car

1. Are there any missing values?

sum(is.na(z1))

## [1] 0

There are 0 missing values in the dataset

1. Any duplicate rows?

sum(duplicated(z1))

## [1] 9

There are 9 duplicate rows

### C) Compute summary statistics :

1. Statistics for engine displacement (displ) :

r\_mode <- function(x) {  
 which.max(tabulate(x))  
}  
mpg %>%  
 summarise(average = mean(displ),  
 median = median(displ),  
 mode = r\_mode(displ),  
 stddev = sd(displ),  
 variance = var(displ),  
 range = max(displ) - min(displ))

## # A tibble: 1 × 6  
## average median mode stddev variance range  
## <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 3.47 3.3 2 1.29 1.67 5.4

1. Statistics for city miles per gallon (cty)

mpg %>%  
 summarise(average = mean(cty),  
 median = median(cty),  
 mode = r\_mode(cty),  
 stddev = sd(cty),  
 variance = var(cty),  
 range = max(cty) - min(cty))

## # A tibble: 1 × 6  
## average median mode stddev variance range  
## <dbl> <dbl> <int> <dbl> <dbl> <int>  
## 1 16.9 17 18 4.26 18.1 26

1. Statistics for highway miles per gallon (hwy)

mpg %>%  
 summarise(average = mean(hwy),  
 median = median(hwy),  
 mode = r\_mode(hwy),  
 stddev = sd(hwy),  
 variance = var(hwy),  
 range = max(hwy) - min(hwy))

## # A tibble: 1 × 6  
## average median mode stddev variance range  
## <dbl> <dbl> <int> <dbl> <dbl> <int>  
## 1 23.4 24 26 5.95 35.5 32

### D) Compute statistics on a numeric variable by grouping on a categorical variable:

1. Summary statistics for engine displacement (displ) :

r\_mode <- function(x) {  
 which.max(tabulate(x))  
}  
mpg %>%  
 group\_by(cyl) %>%  
 summarise(average = mean(displ),  
 median = median(displ),  
 mode = r\_mode(displ),  
 stddev = sd(displ),  
 variance = var(displ),  
 range = max(displ) - min(displ))

## # A tibble: 4 × 7  
## cyl average median mode stddev variance range  
## <int> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 4 2.15 2 2 0.315 0.0993 1.1  
## 2 5 2.5 2.5 2 0 0 0   
## 3 6 3.41 3.4 3 0.472 0.222 1.7  
## 4 8 5.13 5.2 4 0.589 0.347 3

1. Summary statistics for city miles per gallon (cty)

mpg %>%  
 group\_by(cyl) %>%  
 summarise(average = mean(cty),  
 median = median(cty),  
 mode = r\_mode(cty),  
 stddev = sd(cty),  
 variance = var(cty),  
 range = max(cty) - min(cty))

## # A tibble: 4 × 7  
## cyl average median mode stddev variance range  
## <int> <dbl> <dbl> <int> <dbl> <dbl> <int>  
## 1 4 21.0 21 21 3.50 12.2 20  
## 2 5 20.5 20.5 20 0.577 0.333 1  
## 3 6 16.2 16 18 1.77 3.15 8  
## 4 8 12.6 13 11 1.81 3.26 7

1. Summary statistics for highway miles per gallon (hwy)

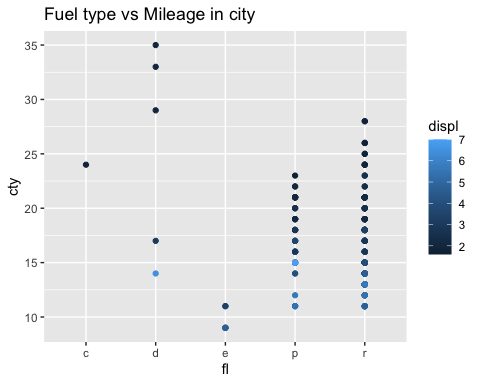
mpg %>%  
 group\_by(cyl) %>%  
 summarise(average = mean(hwy),  
 median = median(hwy),  
 mode = r\_mode(hwy),  
 stddev = sd(hwy),  
 variance = var(hwy),  
 range = max(hwy) - min(hwy))

## # A tibble: 4 × 7  
## cyl average median mode stddev variance range  
## <int> <dbl> <dbl> <int> <dbl> <dbl> <int>  
## 1 4 28.8 29 29 4.52 20.4 24  
## 2 5 28.8 29 29 0.5 0.25 1  
## 3 6 22.8 24 26 3.69 13.6 12  
## 4 8 17.6 17 17 3.26 10.6 14

### E) Visualize and transform to answer the questions asked. Visualizations to illustrate:

1. Relationship between variables

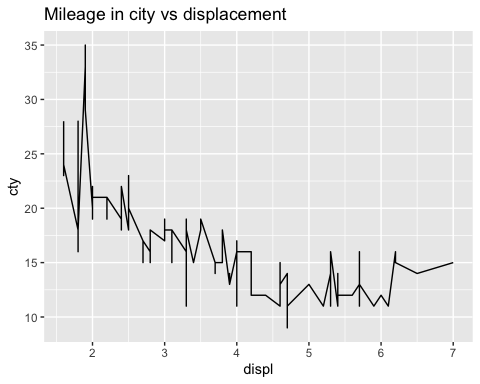
ggplot(data = mpg) +  
 ggtitle("Fuel type vs Mileage in city") +  
 geom\_point(mapping = aes(x = fl, y = cty, color = displ))



This graph shows that the mileage of a car in a city can be higher if d-type, p-type, or r-type fuel is used, which is dependent on engine displacement.

1. Trend

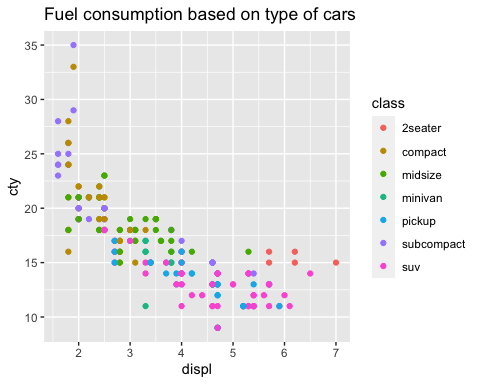
ggplot(data = mpg) +  
 ggtitle("Mileage in city vs displacement") +  
 geom\_line(mapping = aes(x = displ, y = cty))



From the above graph we can determine that the mileage is inversely proportional to the size of engines in a car i.e., the fuel consumption increases with the size of engines

1. Distribution of the variable(s):

ggplot(data = mpg) +  
 ggtitle("Fuel consumption based on type of cars") +  
 geom\_point(mapping = aes(x = displ, y = cty, color = class))



Fuel consumption based on different types of cars - from the graph we can conclude that SUVs consume the highest amount of fuel as compared to the others, irrespective of the engine size.

### F) Summarize your insights from the analysis

1. From the above visualizations, we can conclude that cars with bigger engines consume more fuel.
2. Cars that use p-type fuel and have a big engine have higher mileage compared to other types of fuel.