Analysis Project for mtcars dataset

Alaa Al-Abdali

Acknowledgment to: Dr. Yasir Al-Harthy

Version: 2

2023-07-21

# Introduction:

#### The mtcars dataset is a built-in dataset in R that contains measurements on 11 different attributes for 32 different cars.

#### This tutorial paper explains how to explore, summarize, and visualize the mtcars dataset in R.

# The aim of the project:

#### Explore the relationship between a set of variables and miles per gallon (MPG) to solve the following two questions:

#### 1- Is automatic or manual transmission better for “MPG”.

#### 2- Does the automatic or manual transmission have higher horsepower depending on “MPG”.

# Load necessary libraries:

library(ggplot2)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# Load the mtcars Data set:

# Since the mtcars data set is a built-in data set in R, we can load it by using the following command:  
  
data("mtcars")

# view first six rows of mtcars data set (By default is 6):  
  
head(mtcars)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

# Variables Description:

#### mpg = Miles/(US) gallon

#### cyl = Number of cylinders

#### disp = Displacement (cu.in.)

#### hp = Gross horsepower

#### drat = Rear axle ratio

#### wt = Weight (1000 lbs)

#### qsec = 1/4-mile time

#### vs = Engine (0 = V-shaped, 1 = straight)

#### am = Transmission (0 = automatic, 1 = manual)

#### gear = Number of forward gears

#### carb = Number of carburetors

# Summarize mtcars data set (summarize each variable in the data set):  
  
summary(mtcars)

## mpg cyl disp hp   
## Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0   
## 1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5   
## Median :19.20 Median :6.000 Median :196.3 Median :123.0   
## Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7   
## 3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0   
## Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0   
## drat wt qsec vs   
## Min. :2.760 Min. :1.513 Min. :14.50 Min. :0.0000   
## 1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000   
## Median :3.695 Median :3.325 Median :17.71 Median :0.0000   
## Mean :3.597 Mean :3.217 Mean :17.85 Mean :0.4375   
## 3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000   
## Max. :4.930 Max. :5.424 Max. :22.90 Max. :1.0000   
## am gear carb   
## Min. :0.0000 Min. :3.000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000   
## Median :0.0000 Median :4.000 Median :2.000   
## Mean :0.4062 Mean :3.688 Mean :2.812   
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :1.0000 Max. :5.000 Max. :8.000

#### Here for each of the 11 variables we can see the following information:

#### Min: The minimum value.

#### 1st Qu: The value of the first quartile (25th percentile).

#### Median: The median value.

#### Mean: The mean value.

#### 3rd Qu: The value of the third quartile (75th percentile).

#### Max: The maximum value.

# display rows and columns:  
  
dim(mtcars) # get the number of rows and number of columns

## [1] 32 11

#### There are 32 cars with descriptions on 11 different categorical variables.

# display column names:  
  
names(mtcars)

## [1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear"  
## [11] "carb"

# Check the structure of mtcars data:  
  
str(mtcars)

## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

# Check for missing data:  
  
colSums(is.na(mtcars))

## mpg cyl disp hp drat wt qsec vs am gear carb   
## 0 0 0 0 0 0 0 0 0 0 0

#### There are no missing values, so we are good.

# Converting the variables with discrete values to factor variables:  
  
mtcars$cyl <- as.factor(mtcars$cyl)  
mtcars$vs <- as.factor(mtcars$vs)  
mtcars$am <- as.factor(mtcars$am)  
mtcars$gear <- as.factor(mtcars$gear)  
mtcars$carb <- as.factor(mtcars$carb)

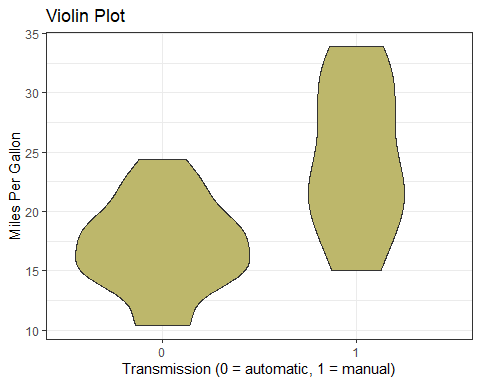
# Explore the statistics of our response variable mpg (miles per gallon):  
  
summary(mtcars$mpg)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.40 15.43 19.20 20.09 22.80 33.90

# Exploratory Data Analysis

# To determine the most effective transmission type for achieving higher miles per gallon (mpg), we will examine the correlation between mpg and the type of transmission (am):

g <- ggplot(data = mtcars, aes(x=am, y=mpg), draw\_quantiles=TRUE)  
g+theme\_bw() + geom\_violin(fill = "darkkhaki") + labs(title="Violin Plot", x="Transmission (0 = automatic, 1 = manual)",y="Miles Per Gallon")



# Observation:

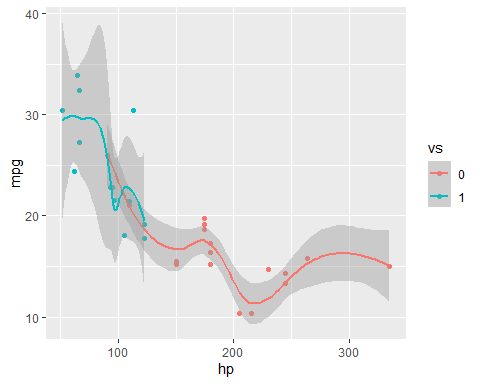
#### The provided exploratory fiddle chart compares the fuel efficiency (MPG) of vehicles with automatic and manual transmissions. The graph indicates a significant increase in MPG for cars equipped with a manual transmission compared to those with an automatic transmission. The distribution figures shown in the graph indicate that the MPG values for automatic transmission cars are tightly clustered around the median and concentrated around the first quartile.

# recommendation:

#### According to the predictions indicated by the graph, I recommend buying cars with an automatic engine because of the lower cost in terms of (mpg), given that cars with a manual engine have more consumption.

# check the relationship between horsepower (hP) and (mpg) differentiating between   
# the type of engine:  
  
ggplot(mtcars, aes(x = hp, y = mpg, color = vs)) + geom\_point() + geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

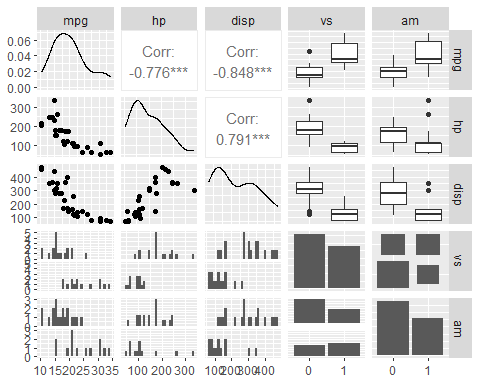


# Observation:

#### We note here that the relationship is slightly linear with only one outlier. The relationship indicates that straight engines with lower horsepower produced higher mpg among the middle values, while for V-shaped engines, despite higher horsepower, mpg values fell below average in the first quantile range.

# check the pairwise correlation between our desired variables:  
  
ggpairs(data = mtcars %>% select(mpg,hp,disp,vs,am))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# Observation:

We can see that the correlation coefficient is higher for the displacement (disp) of the engine, though the relationship is inversely related. The response variable mpg has a linear distribution with right skew.

# Deduction

#### Hypothesis Test and Confidence Interval:

#### Ho: The null hypothesis states that there is no difference between the transmission types.

#### Ha: The alternative hypothesis states that there is a significant difference between the mpg for different transmission types.

# T-Test transmission type and MPG:  
  
test\_result <- t.test(mpg ~ am, data=mtcars)  
test\_result$p.value

## [1] 0.001373638

#### With a p-value as low as 0.0013, we can reject the null hypothesis that the difference between transmission types is 0.

test\_result$estimate

## mean in group 0 mean in group 1   
## 17.14737 24.39231

#### The difference estimate between the 2 transmissions is 7.24494 MPG in favor of the manual.

# Modelling

#### # Since there are more than two explanatory variables, we can go for a multilinear regression model and fit the full model for data.

model <- lm(mpg ~ ., data = mtcars)  
summary(model)

##   
## Call:  
## lm(formula = mpg ~ ., data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5087 -1.3584 -0.0948 0.7745 4.6251   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.87913 20.06582 1.190 0.2525   
## cyl6 -2.64870 3.04089 -0.871 0.3975   
## cyl8 -0.33616 7.15954 -0.047 0.9632   
## disp 0.03555 0.03190 1.114 0.2827   
## hp -0.07051 0.03943 -1.788 0.0939 .  
## drat 1.18283 2.48348 0.476 0.6407   
## wt -4.52978 2.53875 -1.784 0.0946 .  
## qsec 0.36784 0.93540 0.393 0.6997   
## vs1 1.93085 2.87126 0.672 0.5115   
## am1 1.21212 3.21355 0.377 0.7113   
## gear4 1.11435 3.79952 0.293 0.7733   
## gear5 2.52840 3.73636 0.677 0.5089   
## carb2 -0.97935 2.31797 -0.423 0.6787   
## carb3 2.99964 4.29355 0.699 0.4955   
## carb4 1.09142 4.44962 0.245 0.8096   
## carb6 4.47757 6.38406 0.701 0.4938   
## carb8 7.25041 8.36057 0.867 0.3995   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.833 on 15 degrees of freedom  
## Multiple R-squared: 0.8931, Adjusted R-squared: 0.779   
## F-statistic: 7.83 on 16 and 15 DF, p-value: 0.000124

#### Since none of the coefficients have a p-value less than 0.05 I cannot conclude which variables are more statistically significant.

# Backward Elimination method:  
  
red\_model <- step(model, direction = "backward", trace = FALSE)  
summary(red\_model)

##   
## Call:  
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9387 -1.2560 -0.4013 1.1253 5.0513   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.70832 2.60489 12.940 7.73e-13 \*\*\*  
## cyl6 -3.03134 1.40728 -2.154 0.04068 \*   
## cyl8 -2.16368 2.28425 -0.947 0.35225   
## hp -0.03211 0.01369 -2.345 0.02693 \*   
## wt -2.49683 0.88559 -2.819 0.00908 \*\*   
## am1 1.80921 1.39630 1.296 0.20646   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.41 on 26 degrees of freedom  
## Multiple R-squared: 0.8659, Adjusted R-squared: 0.8401   
## F-statistic: 33.57 on 5 and 26 DF, p-value: 1.506e-10

#### The new model has 4 variables (cylinders, horsepower, weight, and transmission).

#### The R-squared value of 0.8659 confirms that this model explains about 87% of the variance in MPG. The p-values also are statistically significant because they have a p-value less than 0.05.

# Conclusion

#### The coefficients conclude that increasing the number of cylinders from 4 to 6 with decrease the MPG by 3.03. Further increasing the cylinders to 8 with decreasing the MPG by 2.16. Increasing the horsepower decreases MPG by 3.21 for every 100 horsepower. Weight decreases MPG by 2.5 for each 1000 lbs increase. A Manual transmission improves the MPG by 1.81.