# Executive Summary

This study intends to explore more about manual and automatic transmission cars. At the end of this study we will be able to answer wich transmission, manual or automatic, is better for gas consumption as well as the quantified difference in miles per galoon between these two types of transmission.

To reach this goal we will use the ***mtcars*** dataset, extracted from 1974 Motor Trend US magazine, and perform linear regressions and technics to improve the model.

The summary of the results is:

1. Manual transmission cars are more economic in gas consumption than automatic transmission cars
2. Manual transmission cars have the mpg ***increased by 1.8*** if compared to automatic transmission cars
3. The medians of automatic and manual transmission cars, that could be observed on the boxplot, are considerable different

# Exploratory Data Analises

To begin the study it is important to check the dataset and its variables:
The help page of the ***mtcars*** dataset has the format:

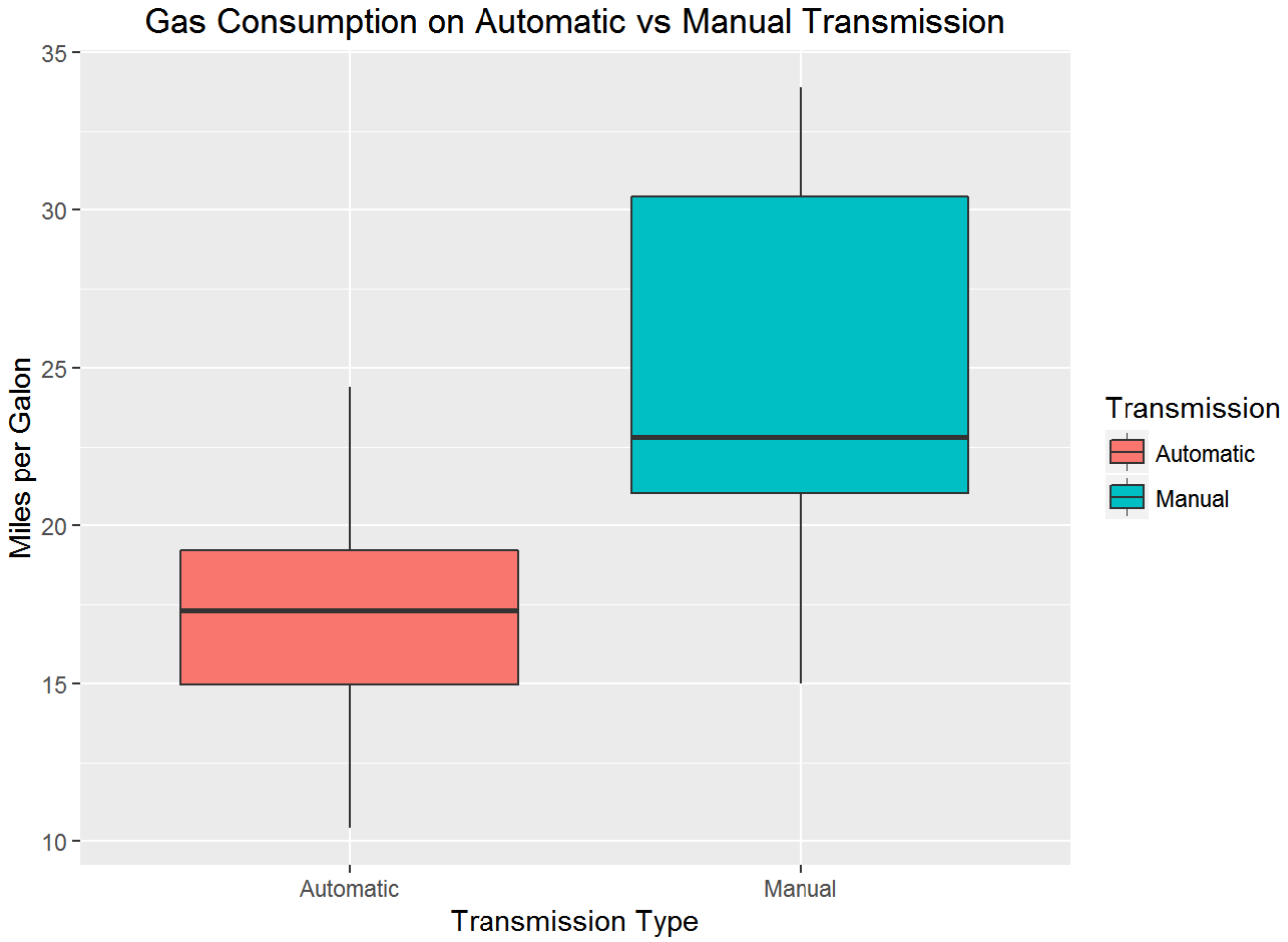
[, 1] ***mpg*** Miles/(US) gallon
[, 2] ***cyl*** Number of cylinders
[, 3] ***disp*** Displacement (cu.in.)
[, 4] ***hp*** Gross horsepower
[, 5] ***drat*** Rear axle ratio
[, 6] ***wt*** Weight (1000 lbs)
[, 7] ***qsec*** 1/4 mile time
[, 8] ***vs*** V/S
[, 9] ***am*** Transmission (0 = automatic, 1 = manual)
[,10] ***gear*** Number of forward gears
[,11] ***carb*** Number of carburetors

```
library(dplyr, quietly = TRUE, warn.conflicts = FALSE)
library(ggplot2, quietly = TRUE, warn.conflicts = FALSE)
data("mtcars")
head(mtcars, n = 5)
```

```
##                   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
```

We can plot a preliminary graph comparing the ***mpg*** in an automatic transmission vs manual transmission in order to guide our study:

```
ggplot(mtcars, aes(x = factor(am, labels = c("Automatic","Manual")), y = mpg)) +
geom_boxplot(aes(fill=factor(am, labels = c("Automatic","Manual")))) + ggtitle(label = "Gas C
onsumption on Automatic vs Manual Transmission") + xlab("Transmission Type") + ylab("Miles pe
r Galon") + labs(fill = "Transmission")
```




This graph showed us that the gas consumption on automatic cars, on avarage, tend to be higher than the manual cars.

We can perform the t-test to reject, or don't, the null hyphothesis:

```
t.test(as.numeric(mtcars$mpg) ~ as.factor(mtcars$am))$p.value
```

```
## [1] 0.001373638
```

The p-value < 0.05 let us to reject the null hypothesis that the mpg for automatic and manual cars are the same.

In order to discover wich variables have more impact on MPG we can study the Pearson correlation table:

```
cor(mtcars)[1,]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

The correlation table showed us that the variables ***cyl, disp, hp, drat, wt, vs and am*** are strong correlated with the outcome ***mpg***

# Data processing and transformation

As we could see in the preview of the dataset, we can transform some variables in factors:

```
mtcars <- mutate(mtcars, cyl = as.factor(cyl))
mtcars <- mutate(mtcars, vs = as.factor(vs))
mtcars <- mutate(mtcars, am = factor(am, labels = c("Automatic","Manual")))
mtcars <- mutate(mtcars, gear = as.factor(gear))
mtcars <- mutate(mtcars, carb = as.factor(carb))
```

# Model Construction

One approach to the linear model selection variables is the backward elimination, encapsulated in the ***step*** function in R. To use this technique we first fit a linear regression model with the outcome, in this case the ***mpg*** variable, and all other variables as predictors and, after that, we can use the ***step*** function as follows:

```
fit_initial <- lm(mpg ~ ., data = mtcars)
fit_step <- step(fit_initial, direction = "both", trace = FALSE)
summary(fit_step)
```

```
##
## 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 **
## amManual     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
```

If we compare the Adjusted R^2 we can see a good improvement in it with less variables than the first model:

```
data.frame(initial_model = summary(fit_initial)$adj.r.squared, step_model =
summary(fit_step)$adj.r.squared)
```

```
##   initial_model step_model
## 1     0.7790215  0.8400875
```

We can also check for anova between the model only considering the transmission type as predictor and the model recomended from step function:

```
anova(lm(mpg ~ am, data = mtcars), fit_step)
```

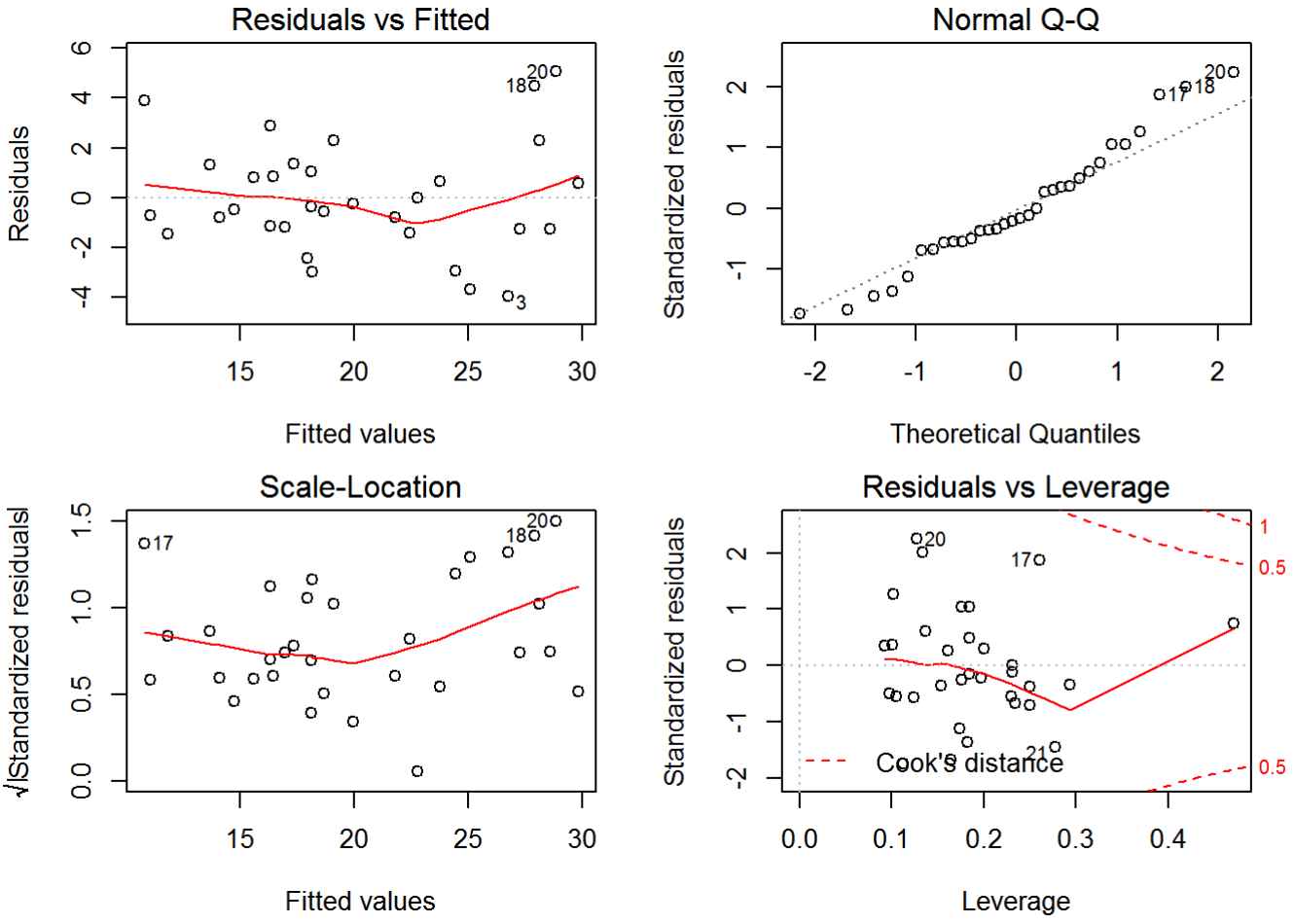
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This anova test, considering the p-value, let us to reject the null hypothesis that the variables included after the step function don't help the model to be more precise.

# Residuals

Let's study the residuals, and others diagnostics, in this model:

```
par(mfrow = c(2,2), mar = c(4,4,2,2))
plot(fit_step)
```




1. This residual plot show us an unbiased model (no pattern found)
2. The Q-Q Plot show us some normality distribution on the residuals
3. The scale-location plot is almost the same as the residual plot, no pattern could be observed so the model in unbiased
4. The residual vs leverage plot show us some outliers that, if removed, will change the linear regression model

# Inferences

There are differences between cars equiped with automatic transmission and manual transmission. As we could observe in the t-test and other tests performed in this study.

# Conclusions

The summary of the linear model lead us to the following conclusions:

1. If all the other variables, ***cyl, hp and wt***, hold the same, manual cars will have an ***increase*** on mpg of ***1.8***
2. If all the other variables, ***cyl, hp and am***, hold the same, for each unit increased in ***wt*** the ***mpg*** will ***decrease*** by ***2.5***
3. If all the other variables, ***cyl, am and wt***, hold the same, for each unit increased in ***hp*** the ***mpg*** will ***decrease*** by ***0.03***
4. If all the other variables, ***am, hp and wt***, hold the same, for cyl change from 4 to 6 it will ***decrease*** the ***mpg*** in ***3.03***
5. If all the other variables, ***am, hp and wt***, hold the same, for cyl change from 4 to 8 it will ***decrease*** the ***mpg*** in ***2.16***