EXECUTIVE SUMMARY

This report aims at investigating the following questions on the base of the database mtcars:

1. Is an automatic or manual transmission better for MPG (miles per gallons)

2. Quantify the MPG difference between automatic and manual transmissions

The database contains 32 car models of which 19 have automatic transmission and 13 have manual.

First we check the null hypothesis that manual and automatic transmissions cars are from the same population (the variable "am" is 0 for automatic transmission and 1 for manual.). For the sake of this projects we assume "mpg" to have a normal distribution, although it looks a little skewed.

A simple t-test confirm that the two transmissions are not from the same population as the interval does not contain 0 and the p-value 0.0013736 is sufficiently low.

```
## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95
```

We consider a linear model with "am" as a predictor and "mpg" as a response.

The relatevly high F-statistic and the very low p-value confirm that there might be a correlation, but if we look at how well the model fit, R-squared 0.36 is very low (1 being best fit, 0 being no fit at all). This, together with a Residual Standard Error (RSE) of 4.822, suggests that the model is not very accurate and that there might be other model to better explain the correlation. We look therefore to other variables that might be confounder. Based on a brief literature review, the variables that can be relevant to our analysis are weight, the number of cylinders and the rear axle ratio (drat in dataset). We can call primary this variables as this are the engine and car specifications rather than performance. Other variables as "qsec" (which measures the acceleration) and "hp" (which measure the power of the engine) can be good indicators but they are performance variables and are consequences of the primary specs. The number of cylinders can tell a lot about a car because it is related to the size and power of the engine and therefore the size and the type of the car. Displacement (the amount of fuel burnt per stroke of the engine) gives similar information about the engine but it is correlated with cylinders(see APPENDIX). The weight combined with the number of cylinders and the displacement, can tell whether we are talking about a small car (small cyl, wt and disp therefore high mpg), a big car (high cyl, wt and disp and low mpg) or a sport car (high cyl and disp, low weight, low mpg).

The predictor "cyl" explains 72.6 % of "mpg" variation. ADD DESCRIPTION OF THE SUMMARY

Once we factor in the weight "wt", the model explains 83 % of "mpg" variation with a RSE of 3 and very low p-values. In the final part we add the variable "am". From the graph in the appendix with "am" as a predictor, it seems that there is an interaction between "am" and "wt", as automatic cars in the dataset tend to be heavier than the manuals one (a t.test over the two populations confirm that, see APPENDIX).

The third model includes cyinders, weight, manual/automatic transmission and the interaction between the latter and the weight

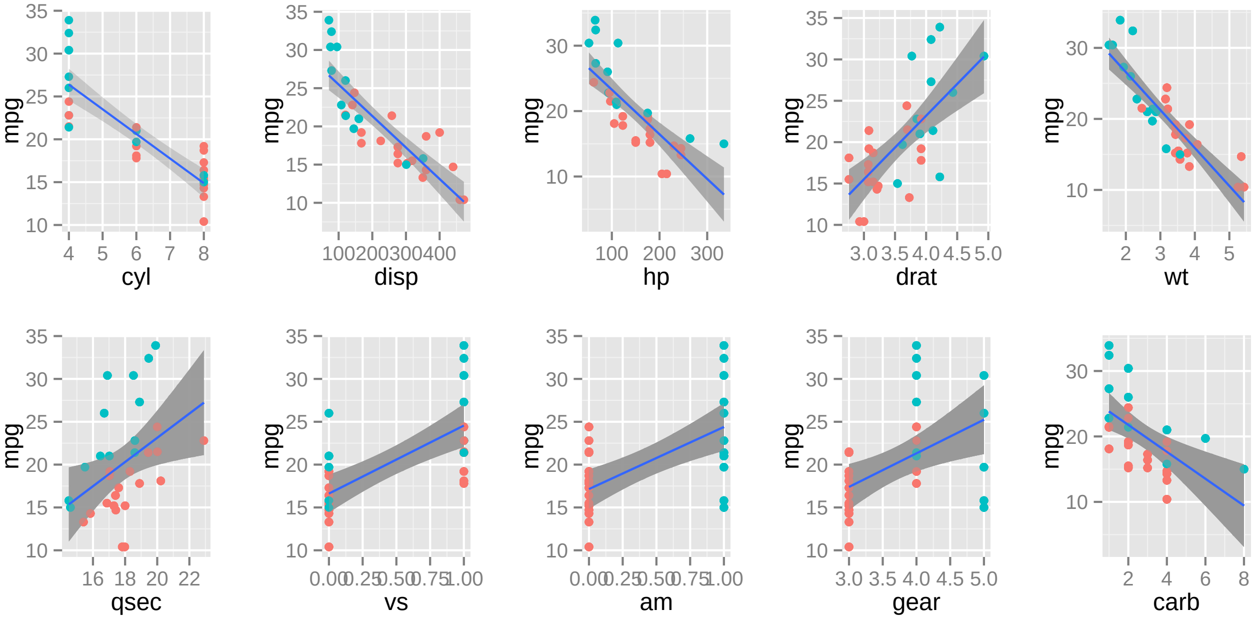
Last, we add the variables "drat" and the interaction between "disp" and "cyl".

We can now run an anova test to verify which of these variale can be dropped without loss of relevant information.

```
## Analysis of Variance Table
##
```

```
## Model 1: mpg ~ cyl
## Model 2: mpg ~ cyl + wt
## Model 3: mpg ~ cyl + wt + am + wt:am
## Model 4: mpg ~ cyl + wt + am + drat + wt:am
## Model 5: mpg ~ cyl + wt + am + drat + wt:am + cyl:disp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 308.33                                   
## 2     29 191.17  1   117.162 21.1492 0.0001054 ***
## 3     27 138.51  2    52.666  4.7535 0.0178001 *  
## 4     26 138.50  1     0.010  0.0019 0.9657499    
## 5     25 138.50  1     0.000  0.0001 0.9925400    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

APPENDIX Exploratory graphs



```
## [1] -0.6924953
```

```
## 
##  Welch Two Sample t-test
## 
## data:  wt[am == 0] and wt[am == 1]
## t = 5.4939, df = 29.234, p-value = 6.272e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8525632 1.8632262
## sample estimates:
## mean of x mean of y 
##  3.768895  2.411000
```

```
## 
## Call:
## lm(formula = wt ~ am)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3039 -0.3694 -0.2049  0.3156  1.6551
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7689     0.1646  22.895  < 2e-16 ***
## am           -1.3579     0.2583  -5.258 1.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7175 on 30 degrees of freedom
## Multiple R-squared:  0.4795, Adjusted R-squared:  0.4622
## F-statistic: 27.64 on 1 and 30 DF,  p-value: 1.125e-05
```

```
## [1] 0.9020329
```

```
##
##  Welch Two Sample t-test
##
## data:  disp[cyl == 4] and disp[cyl == 6]
## t = -4.423, df = 9.225, p-value = 0.001567
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -118.01437  -38.34147
## sample estimates:
## mean of x mean of y
##  105.1364  183.3143
```

```
##
##  Welch Two Sample t-test
##
## data:  disp[cyl == 6] and disp[cyl == 8]
## t = -7.0815, df = 17.931, p-value = 1.359e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -220.1712 -119.4002
## sample estimates:
## mean of x mean of y
##  183.3143  353.1000
```

```
##
## Call:
## lm(formula = disp ~ cyl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.985 -45.233   3.565  26.688 127.818
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -156.609     35.181  -4.452 0.000109 ***
## cyl           62.599      5.469  11.445  1.8e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.38 on 30 degrees of freedom
## Multiple R-squared:  0.8137, Adjusted R-squared:  0.8075
## F-statistic:   131 on 1 and 30 DF,  p-value: 1.803e-12
```