What car-specifics are efficient on a environmental and economic perspective for the year 1974?

Kevin Nilsson

2024-12-25

Introduction

In this report the dataset mtcars (motor trend car road tests) has been used with focus on what factors that influences fuel efficiency. This dataset descend from 1974 "Motor Trend" US Magazine and consists of a total of 32 car models. Different questions are answered:

- Are there any significant associations between number of cylinders and number of gears?
- How does transmission types, cylinders, horsepower affect the miles per gallon?
- What car-specifics are efficient on a environmental and economic perspective?

Different visualizations has been made, using scatterplots, histogram, bar charts and box plots. Correlation analysis is used to see the strength of associations and relationships between variables. Finally, a best model was developed and visualized which gave information about the most significant factors impacting fuel efficiency.

```
#loading libraries
library(ggplot2)
library(tidyverse)
library(dplyr)
library(readr)
library(purrr)
library(vcd)
library(car)

#Loading the dataset mtcars
data(mtcars)

#Taking a glimpse of the data
glimpse(mtcars)
```

```
## Rows: 32
## Columns: 11
## $ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8,~
## $ cyl  <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4, 4, 8,~
## $ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8, 16~
## $ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180, 180~
## $ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3.92,~
```

```
## $ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150, 3.~
## $ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90, 18~
## $ vs   <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0,~
## $ am   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,~
## $ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3,~
## $ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1, 2,~
```

Description of table above:

Dataset has 32 rows, 11 columns, and all variables are of numeric kind. Description of the variables below:

mpg = Miles per gallon

cyl = Number of cylinders

disp = Displacement

hp = Horsepower

drat = Rear axle ratio

wt = Weight

qsec = Quarter mile time

vs = Engine type (V/S)

am = Transmission type (0=Automatic/1=Manual)

gear = Number of forward gears

carb = Number of carburetors

Some variables will be transformed into factors for easier analysis purpose.

```
#Factor transformation
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs, labels=c("V-shaped", "Straight"))
mtcars$am <- factor(mtcars$am, labels=c("Automatic", "Manual"))
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)

#Converting car names to a column
mtcarscleaned <- tibble::rownames_to_column(mtcars,"car_name")

#Checking for missing values
sum(is.na(mtcarscleaned))
```

```
## [1] 0
```

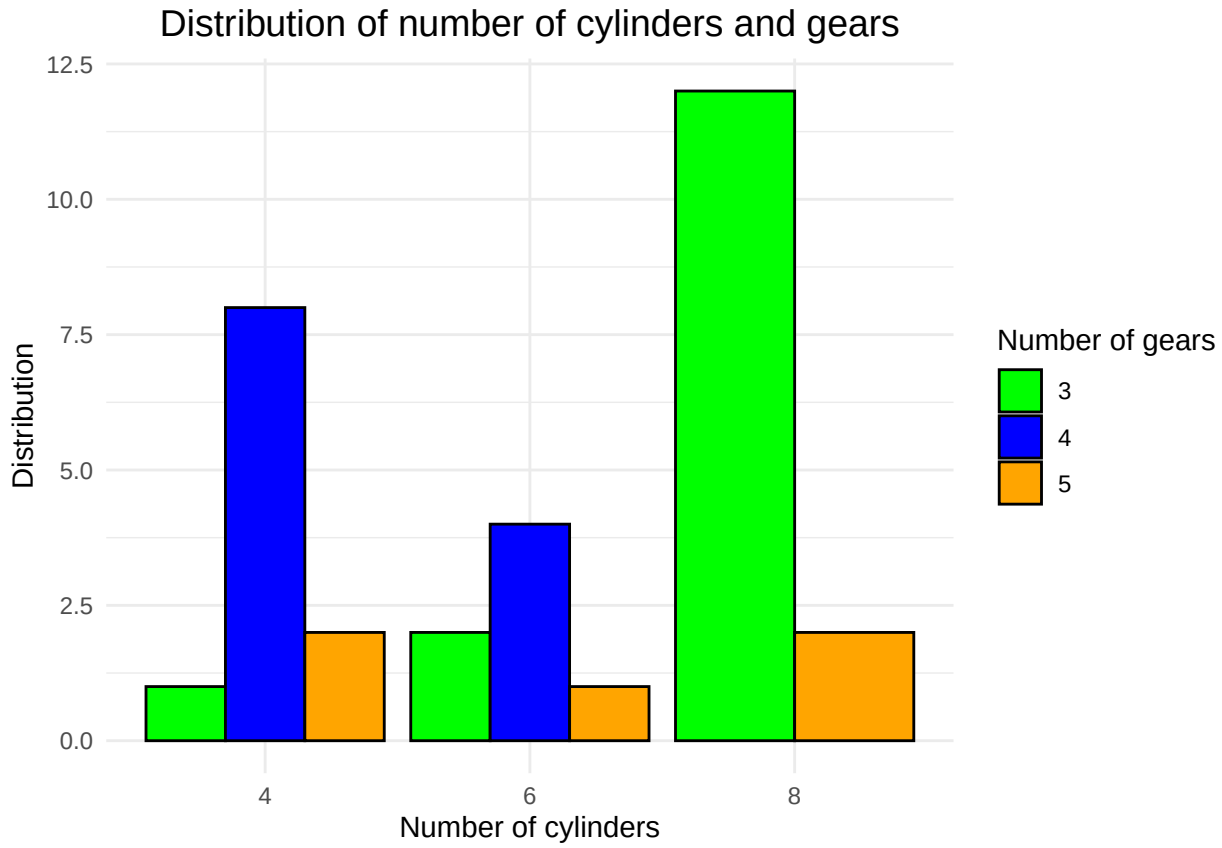
```
#checking if the factor transformation was successful
glimpse(mtcarscleaned)
```

```
## Rows: 32
## Columns: 12
## $ car_name <chr> "Mazda RX4", "Mazda RX4 Wag", "Datsun 710", "Hornet 4 Drive",~
## $ mpg      <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 1~
## $ cyl      <fct> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, 8, 8, 4, 4, 4, 4~
## $ disp     <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360.0, 146.7, 140.8~
```

```
## $ hp       <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 180,~
## $ drat     <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.69, 3.92, 3.92, 3~
## $ wt       <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3.150~
## $ qsec     <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15.84, 20.00, 22.90~
## $ vs       <fct> V-shaped, V-shaped, Straight, Straight, V-shaped, Straight, V~
## $ am       <fct> Manual, Manual, Manual, Automatic, Automatic, Automatic, Auto~
## $ gear     <fct> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3~
## $ carb     <fct> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, 4, 4, 1, 2, 1, 1~
```

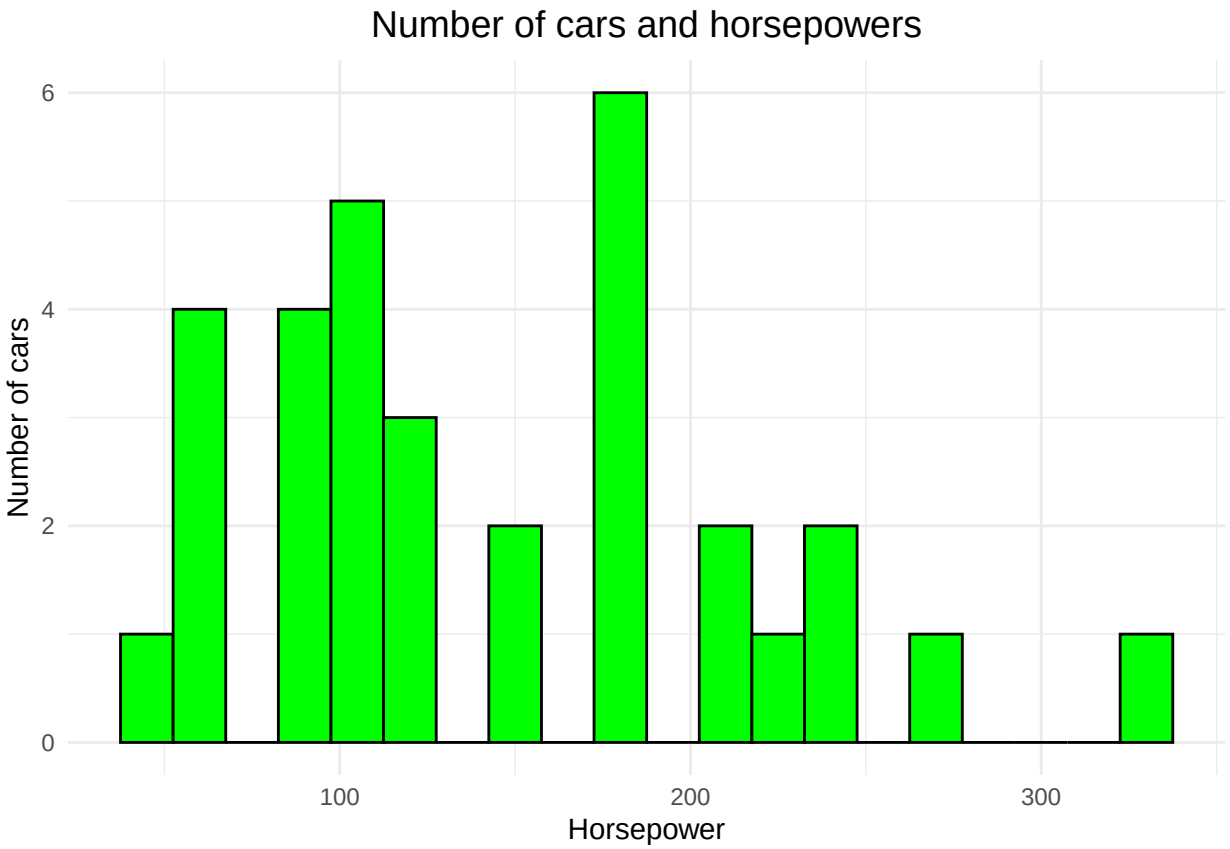
Description of table above:

Table above provides information that the factor transformations was successful. Now the dataset consist of 5 factor variables, 1 character, and 6 double numeric variables. There are zero missing values. Now some basic visualizations will be made.



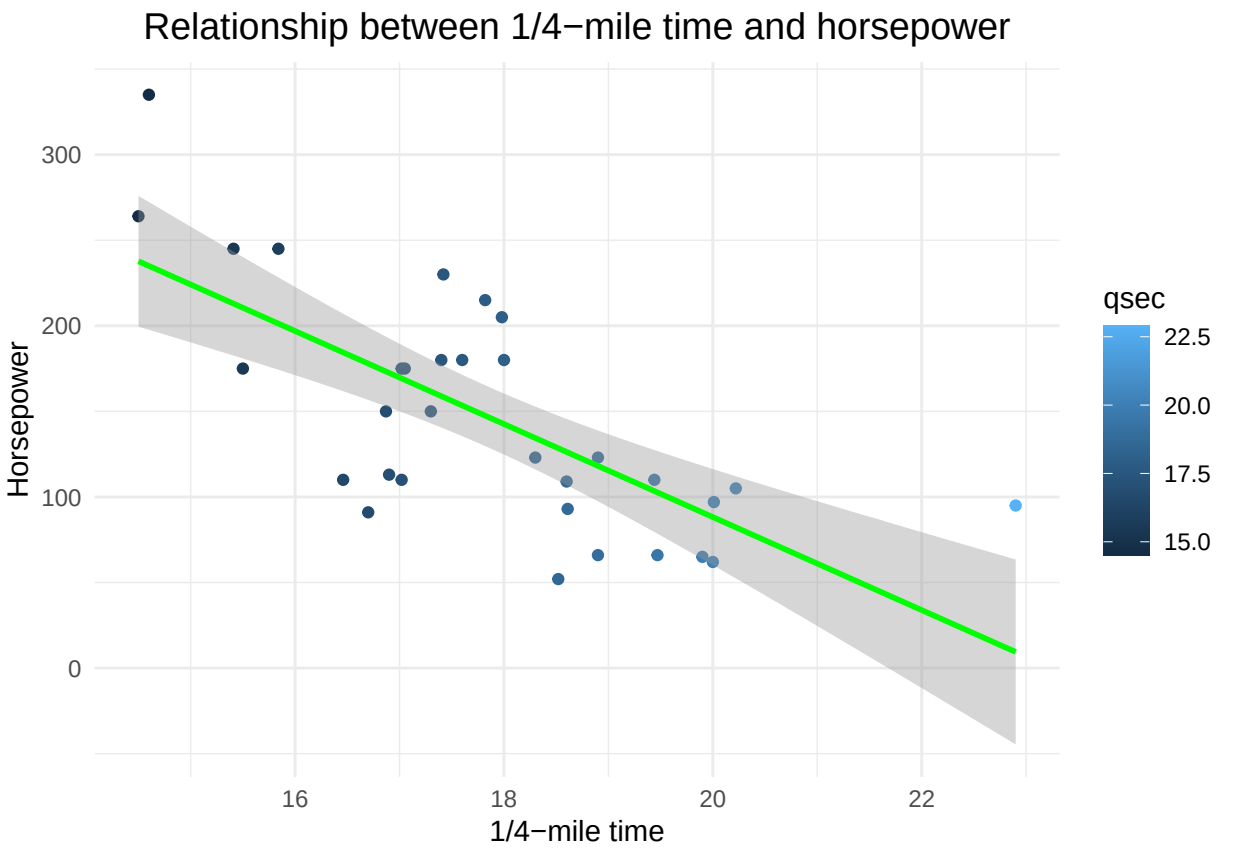
Bar chart interpretation:

The bar chart from above provides information about the relationship between number of cylinders and number of gears. Here we can see clearly that when increasing the number of cylinders, cars with 4 gears decreases significantly (to zero for the number of 8 cylinders). Cars with 3 gears increases significantly from approx. 1 car to approx. 12 cars when increasing cylinders from 4 to 8. Cars with 5 gears are almost similarly distributed through all cylinder categories. Conclusion can be made that cars with more cylinders tend to have less gears, and cars with less cylinders tends to have more gears. This is probably due to performance efficiency.



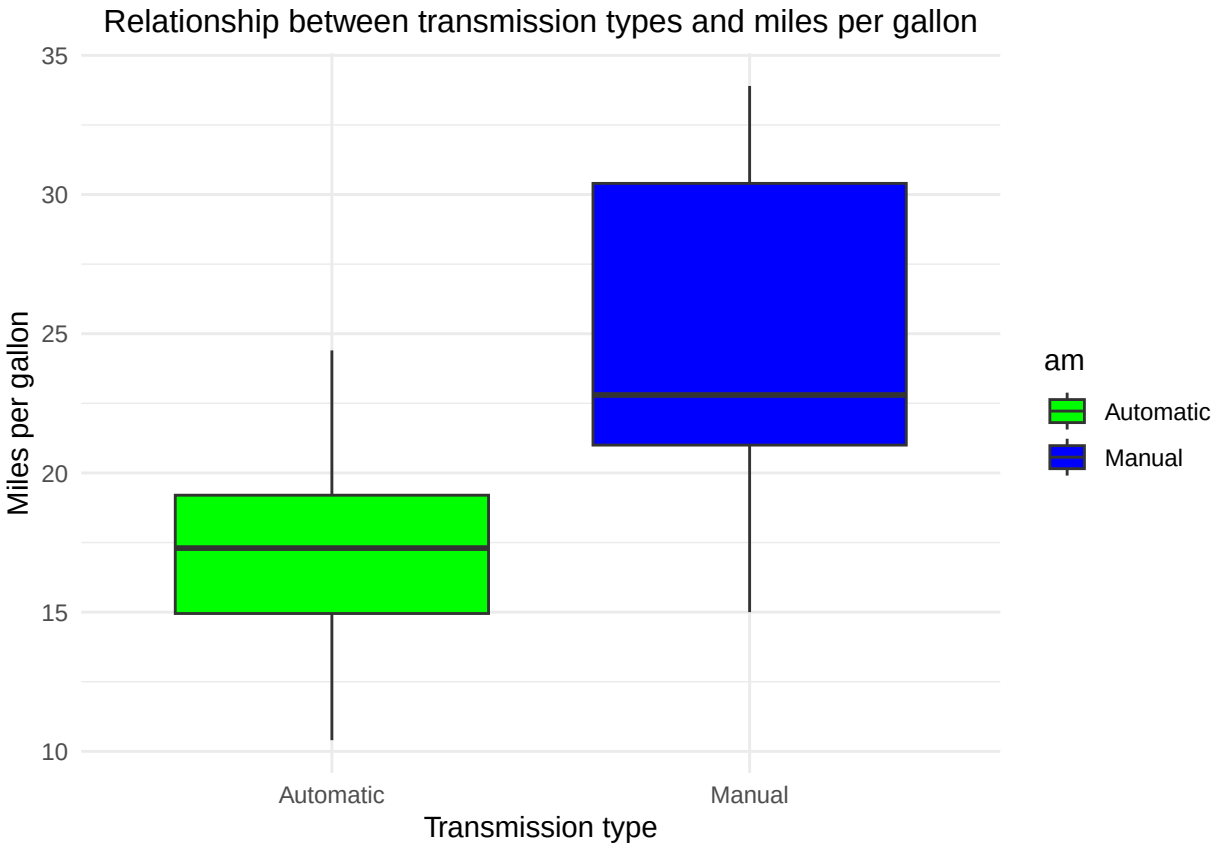
Interpretation of histogram above:

The histogram above are rightly skewed, which tells us the information that there are more cars of the lower horsepower than that of the higher horsepower. Very few cars have a horsepower above 250 hp. The most common are cars between 50 and 150 horsepowers.



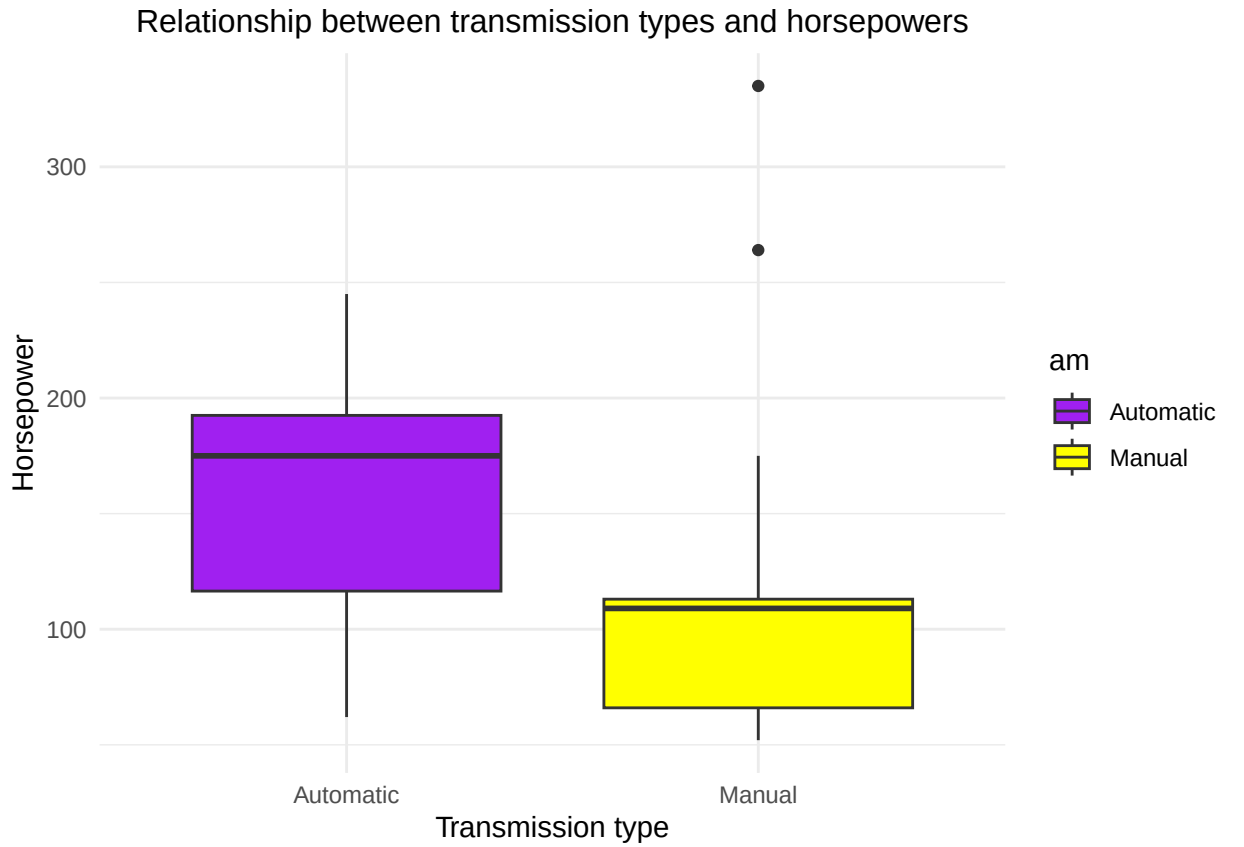
Interpretation of the scatterplot:

The scatterplot shows the relationship between 1/4-mile time and horsepower, telling us that the higher the horsepower the lesser the time in 1/4 miles and vice versa. The datapoints are also randomly distributed pretty close to the best fitted line, in no specific shape which tells us that the assumption of homoscedasticity holds and that there is a strong relation between the variables. The 1/4-mile time ranges from approx. 14.5 to 23 seconds and the data points are shown in darker blue when having a faster 1/4-mile time, and lighter blue with a slower 1/4-mile time.



Box plot interpretation:

The green box shows cars with automatic transmission, and it looks normally distributed, while the cars of manual transmission have a positive skew, which can be seen by the low median line in the box. The median value of manual cars are approx. 22.5 miles per gallon, whilst that of the automatic cars have a miles per gallon median equal to approx. 17.5. Conclusion can be made just by looking at the two boxes that cars with automatic transmission type tends to have a higher fuel consumption than cars with manual transmission, hence cars with manual transmission are better in an environmental and economical perspective.



Another relationship that is highly relevant is between transmission types and horsepower, to see this we simply take a look at the box plot from above. Here, its visible that manual cars tends to significantly have a lower horsepower than that of the automatic transmission types. Which also explains the conclusion in the last plot, where manual cars are more kind on the environment and reaches more miles per gallon. Also, there are two extreme values in cars with manual transmission, telling us that some cars with manual transmission also has really high horsepowers.

Decsrpitives and model comparisons

Adj. R-squared:

Model 1 = 0.5892

Model 2 = 0.7275

Model 3 = 0.7989

Model 4 = 0.7915

Here we can see that the adj. R-squared increases from model 1 to 3, and then starts to decrease a bit to model 4. So, model 3 is the model with the highest adj. R-squared, telling us that after adjusting for the number of independent variables in the model, this is the model that explains most of the variability in miles per gallon (79.89% of the variability in the dependent variable are explained by horsepowers, number of cylinders and transmission types).

Testing independent variables p-values on a 5%-level of significance below:

Model 1: mpg = 30.1 - 0.068hp

In model 1, only one independent variable (hp) which is statistically significant.

Model 2: mpg = 28.65 - 0.024hp - 5.97cyl6 - 8.521cyl8

In model 2, there are 3 independent variables. Here hp is no longer significant, and we can draw the conclusion that in this model horsepower is not significantly explaining the variance in miles per gallon anymore.

Model 3: mpg = 27.296 - 0.044hp - 3.925cyl6 - 3.533cyl8 + 4.158am

In model 3, there are 4 independent variables, now hp is back at significant again, and cyl8 is no longer significant.

Model 4: mpg = 28.141 - 0.054hp - 3.623cyl6 - 2.429cyl8 + 3.4am 0 0.216gear4 + 2.207gear5

The last model (model 4), with 6 independent variables, there are 3 variables that is not statistically significant, gear5, gear4, cyl8. We can also see that transmission manual, and cars with 4 gears, 5 gears has a positive impact on fuel consumption (an increase in these independent variables also increases the dependent).

This far, we have indications that the model 3 is the best one, but I will also use the AIC value approach to reduce the chance of overfitting before I draw the best model conclusions.

```
#Calculating the AIC values for each model
AIC(model1, model2, model3, model4)
```

```
##        df      AIC
## model1  3 181.2386
## model2  5 169.8964
## model3  6 161.0033
## model4  8 163.6950
```

Model 3 is the model with the lowest AIC value (161), hence conclusion can be made that this is the best model.

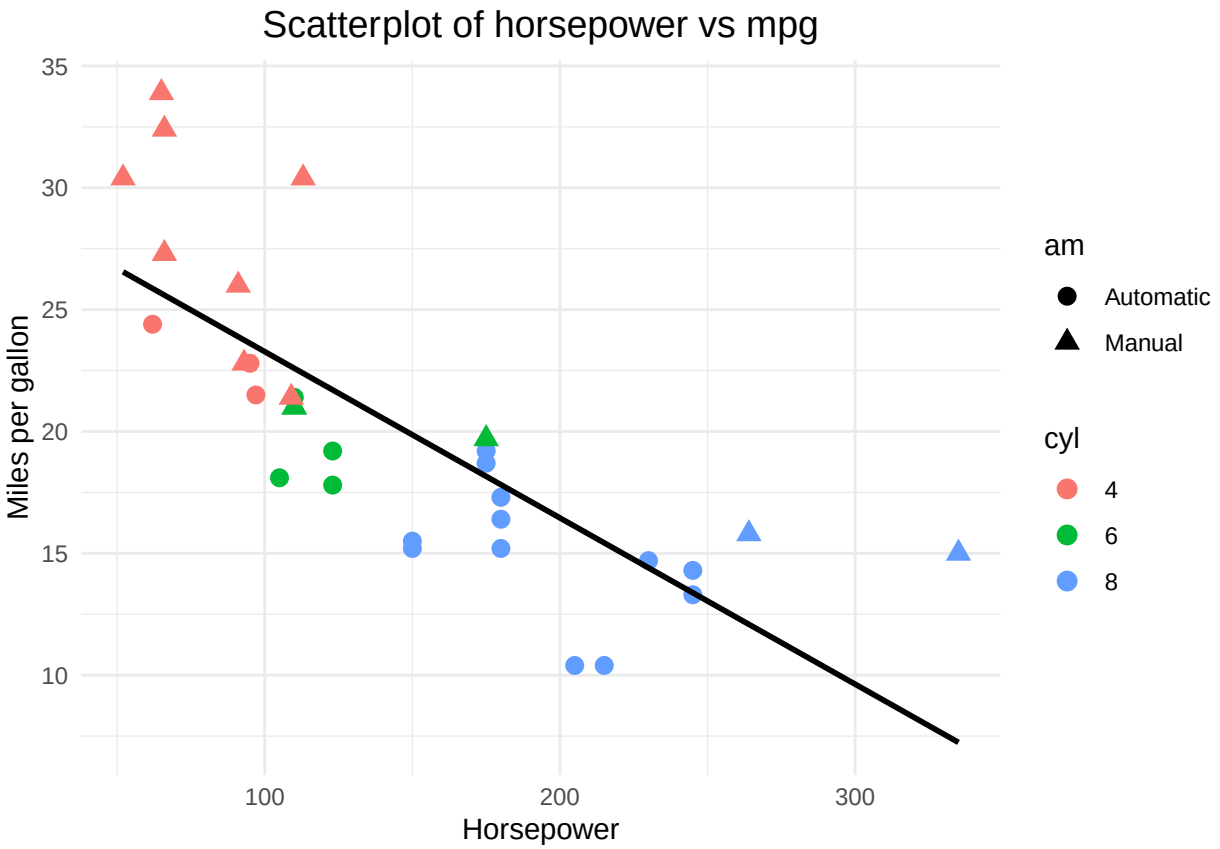
Model 3 -> mpg = 27.29 -0.044hp - 3.925cyl6 - 3.533cyl8 + 4.158am

Next step in the analysis I will check for multicollinearity.

```
#Computing VIF to evaluate multicollinearity
vif_results <- vif(model3)
print(vif_results)
```

```
##         GVIF Df GVIF^(1/(2*Df))
## hp  4.238903  1        2.058860
## cyl 5.486920  2        1.530496
## am  1.668652  1        1.291763
```

By analyzing the table above that includes VIF (Variance inflation factor) of the different independent variables of the best model we get the following results: hp has a VIF = 2.05, cyl = 1.53 and am = 1.29, hence, all are pretty close to 1, and I have no concerns of multicollinearity in this model. So, the independent variables are not highly correlated with each other which is a good sign.



By analyzing the scatter plot of the best model above, its clear that there is a negative relationship between miles per gallon and horsepower, higher horsepower results in less miles per gallon (as a cause of an increase in fuel consumption). Seen is also that there are mode cars with automatic transmissions with a higher horsepower, and manual transmissions tend to have a lower horsepower. Also, more cylinders seems to be significantly related to higher horsepower.

Summary

To summarize the results of this report, different conclusions has been made. Cars with more cylinders tend to have less gears, and cars with less cylinders tends to have more gears. This is probably due to performance efficiency. A higher horsepower is associated with a lower 1/4-mile time. Another conclusion that was made is that cars with automatic transmission type tends to have a higher fuel consumption than cars with manual transmission, hence cars with manual transmission are more kind on an environmental perspective but also in the economic aspect. This was also explained, as cars with manual transmissions tends to have lower horsepower than cars with automatic transmission.

After testing different models and finalizing the best model, it was visualized in a scatter plot. This told us that more cylinders seems to be significantly related to higher horsepower. It was also clear that cars with higher horsepower are related to a higher fuel consumption. So, to reach an efficient level of fuel consumption the recommendations of this report is to have a car with lower horsepower (up to 150hp), with manual transmission and with few cylinders (4).

Also, note once again that the data is collected from 1974, so the car specifics and efficiency on environment/economical perspective might have changed a bit since.