```
library(RColorBrewer)
library(ggExtra)
library(ggplot2)
library(ggthemes)
library(gridExtra)
library(hrbrthemes)
library(patchwork)
library(PerformanceAnalytics)
library(tidyverse)
library(vioplot)
```

Car information dataset analysis.

The dataset contains 399 rows of 9 features, which contains some general properties of cars. These 9 features are the following:

1. Name: Unique identifier for each automobile.
2. MPG: Fuel efficiency measured in miles per gallon.
3. Cylinders: Number of cylinders in the engine.
4. Displacement: Engine displacement, indicating its size or capacity.
5. Horsepower: Power output of the engine.
6. Weight: Weight of the automobile in pounds.
7. Acceleration: Capability to increase speed, measured in seconds to 60 miles/hour.
8. Model Year: Year of manufacture for the automobile model.
9. Origin: Country or region of origin for each automobile.

The dataset can be found via this **link**

Data exploration

In the output below, the first 5 lines of the data can be seen, and the overall structure.

```
# setwd("/media/sf_SF/Fedora/R_course/Assignment/")
car_data <- read.csv("Automobile.csv")
head(car_data)
```

```
##                       name mpg cylinders displacement horsepower weight
## 1 chevrolet chevelle malibu  18         8          307        130   3504
## 2         buick skylark 320  15         8          350        165   3693
## 3        plymouth satellite  18         8          318        150   3436
## 4             amc rebel sst  16         8          304        150   3433
## 5               ford torino  17         8          302        140   3449
## 6          ford galaxie 500  15         8          429        198   4341
##   acceleration model_year origin
## 1         12.0         70    usa
## 2         11.5         70    usa
## 3         11.0         70    usa
## 4         12.0         70    usa
## 5         10.5         70    usa
## 6         10.0         70    usa
```

```
str(car_data)
```

```
## 'data.frame':    398 obs. of  9 variables:
##  $ name        : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : int  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ model_year  : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : chr  "usa" "usa" "usa" "usa" ...
```

Let's check if there are empty values, while omitting these directly.

```
missing_values <- which(is.na(car_data), arr.ind = TRUE)
print(missing_values)
```

```
##      row col
## [1,]  33   5
## [2,] 127   5
## [3,] 331   5
## [4,] 337   5
## [5,] 355   5
## [6,] 375   5
```

```
print(paste("Column 5 is",colnames(car_data)[5]))
```

```
## [1] "Column 5 is horsepower"
```

```
car_data <- na.omit(car_data)
```

There are six missing values in the "horsepower" column.

The dataset contains four categorical variables (name, model_year, cylinders, and origin) and six numerical variables (mpg, displacement, horsepower, weight, and acceleration).

IMPORTANT NOTE: The number of cylinders falls under categorical variables, as it divides the cars into categorical groups based on their engine. Additionally, the miles per gallon will be disregarded, as the mpg values will be converted into liters per 100 kilometers (L/100km) and placed within a new column.

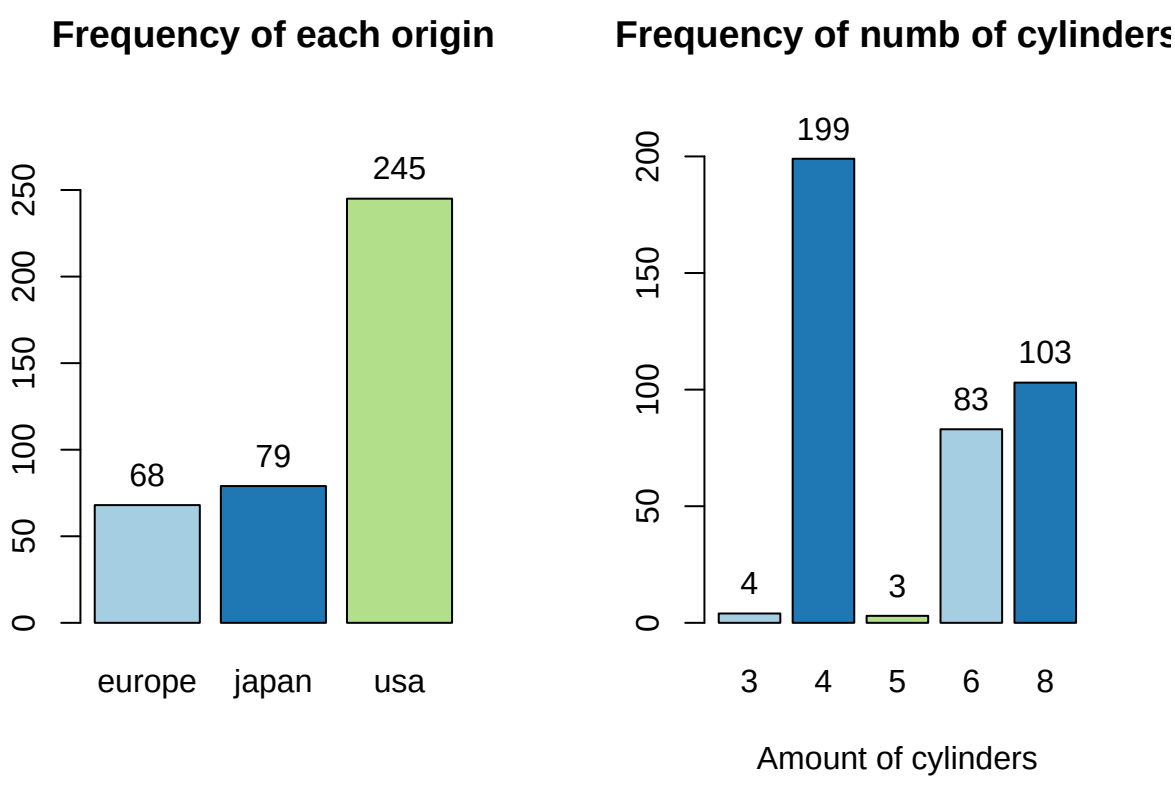
```
car_data$model_year <- as.character(car_data$model_year)
car_data$model_year <- paste0("19",car_data$model_year)
car_data$cylinders <- as.character(car_data$cylinders)
car_data$L_100km <- 235.215 / car_data$mpg
cat_var <- c("name", "brand", "model_year", "origin", "cylinders")
num_var <- c("L_100km", "displacement", "horsepower", "weight", "acceleration")
```

Let's take a look at the frequencies of each categorical variable. Because of the huge amount of unique car models, no representative barplot can be generated. The complete dataset contains 37 car brands

```
car_data$brand <- sapply(strsplit(car_data$name, " "), `[`, 1)

paste0("Unique car models: ",length(unique(car_data$name)),', Unique car brands: ', length(unique(car_da
```

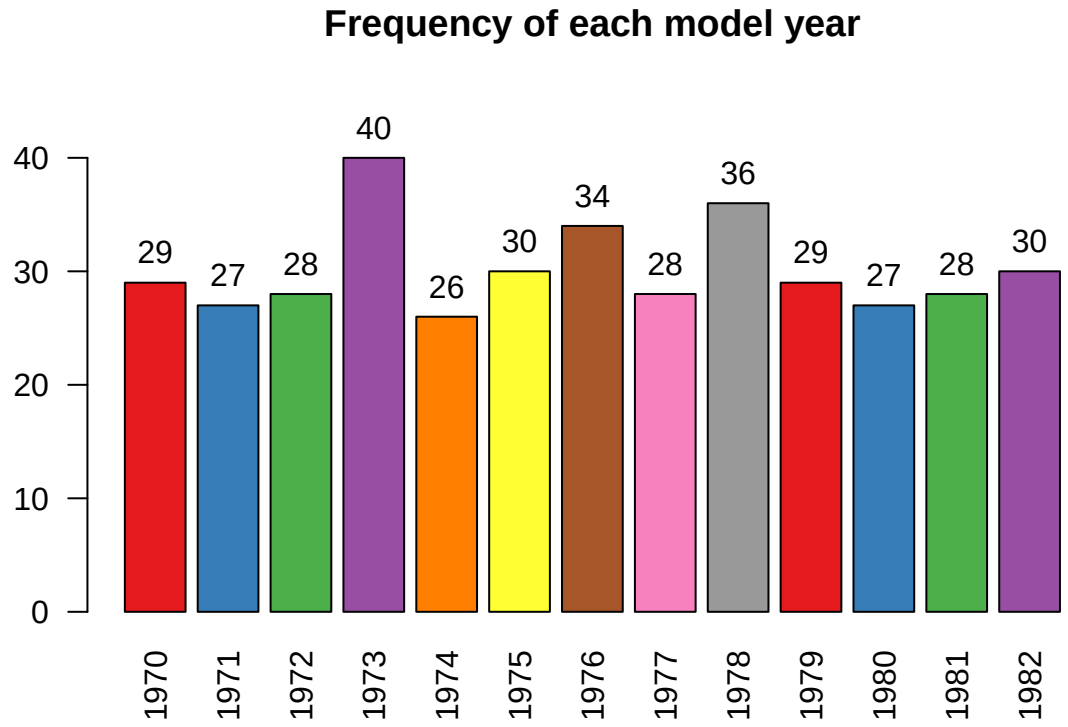
```
## [1] "Unique car models: 301, Unique car brands: 37"
```

```
par(mfrow = c(1,2))
bp <- barplot(table(car_data$origin),
        main = "Frequency of each origin",
        ylim = c(0,max(table(car_data$origin))+50),
        col = brewer.pal(3, "Paired"))
text(x=bp, y=table(car_data$origin),label=table(car_data$origin),pos=3)
bp <- barplot(table(car_data$cylinders),
        main = "Frequency of numb of cylinders",
        xlab = "Amount of cylinders",
        ylim = c(0,max(table(car_data$cylinders)+20)),
        col = brewer.pal(3, "Paired"))
text(x=bp, y=table(car_data$cylinders),label=table(car_data$cylinders),pos=3)
```



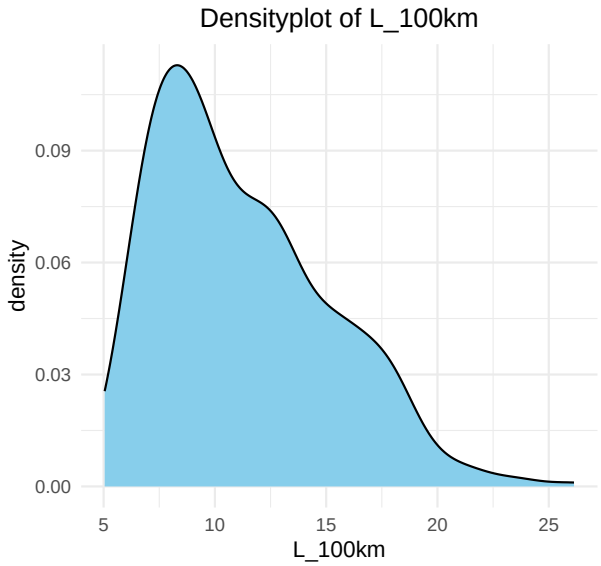

```
par(mfrow = c(1,1))
bp <- barplot(table(car_data$model_year),
        main = "Frequency of each model year",
        ylim = c(0,45),
        col = brewer.pal(12, "Set1"),
```

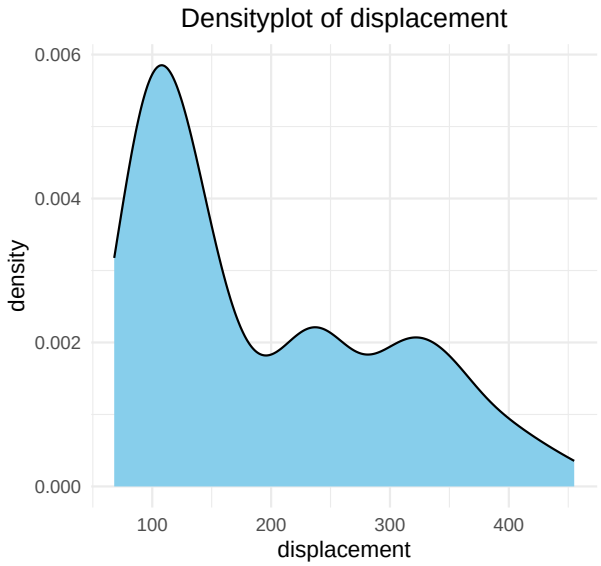
```
        las = 2)
text(x=bp, y=table(car_data$model_year),label=table(car_data$model_year),pos=3)
```

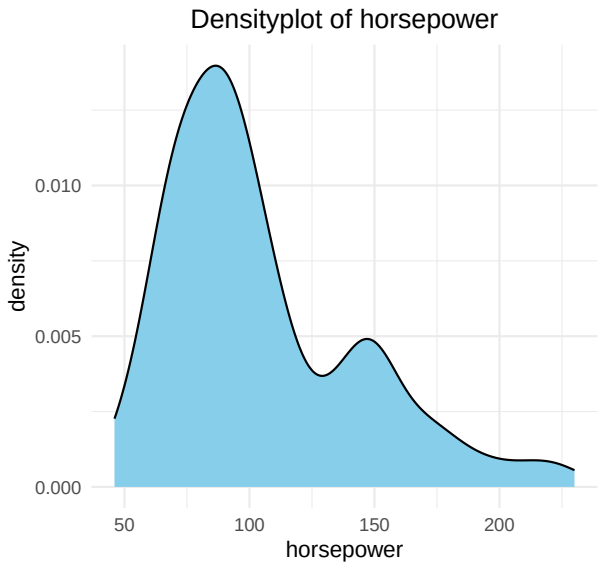


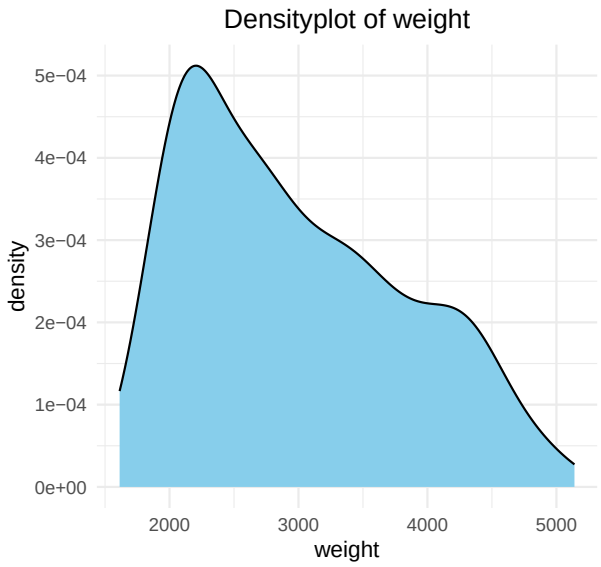
The plots above show that the model years are fairly normaly distributed. But on the other hand, the origin and amount of cylinders of the cars are not equally distributed. Let's take a look at the density and distributions of the numerical variables now.

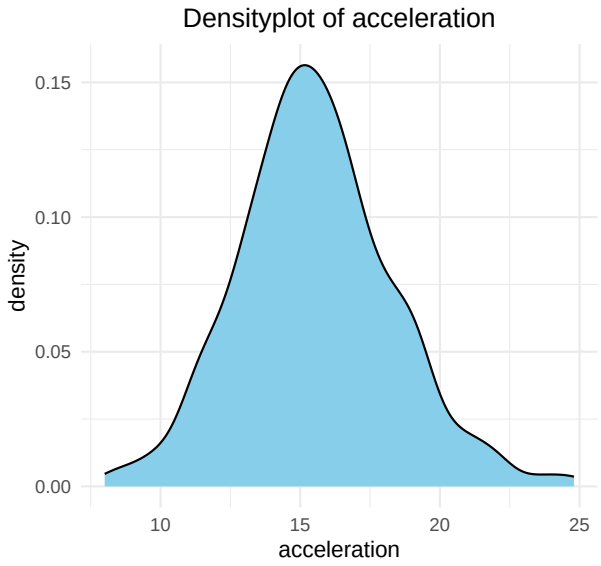
```
plot_list <- list() # Making a list to arrange the plots nicely later
for (var in num_var) {
  p <- ggplot(car_data, aes(x = !!sym(var))) +
    geom_density(fill = "skyblue", color = "black") +
    labs(title = paste("Densityplot of", var)) +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5))
  plot_list[[length(plot_list) + 1]] <- p
}
grid.arrange(grobs = plot_list, ncol = 2, nrow = 3) # arranging the plots in one figure
```


Densityplot of L_100km
density
0.09
0.06
0.03
0.00
5
10
15
20
25
L_100km


Densityplot of displacement
density
0.006
0.004
0.002
0.000
100
200
300
400
displacement


Densityplot of horsepower
density
0.010
0.005
0.000
50
100
150
200
horsepower


Densityplot of weight
density
5e−04
4e−04
3e−04
2e−04
1e−04
0e+00
2000
3000
4000
5000
weight

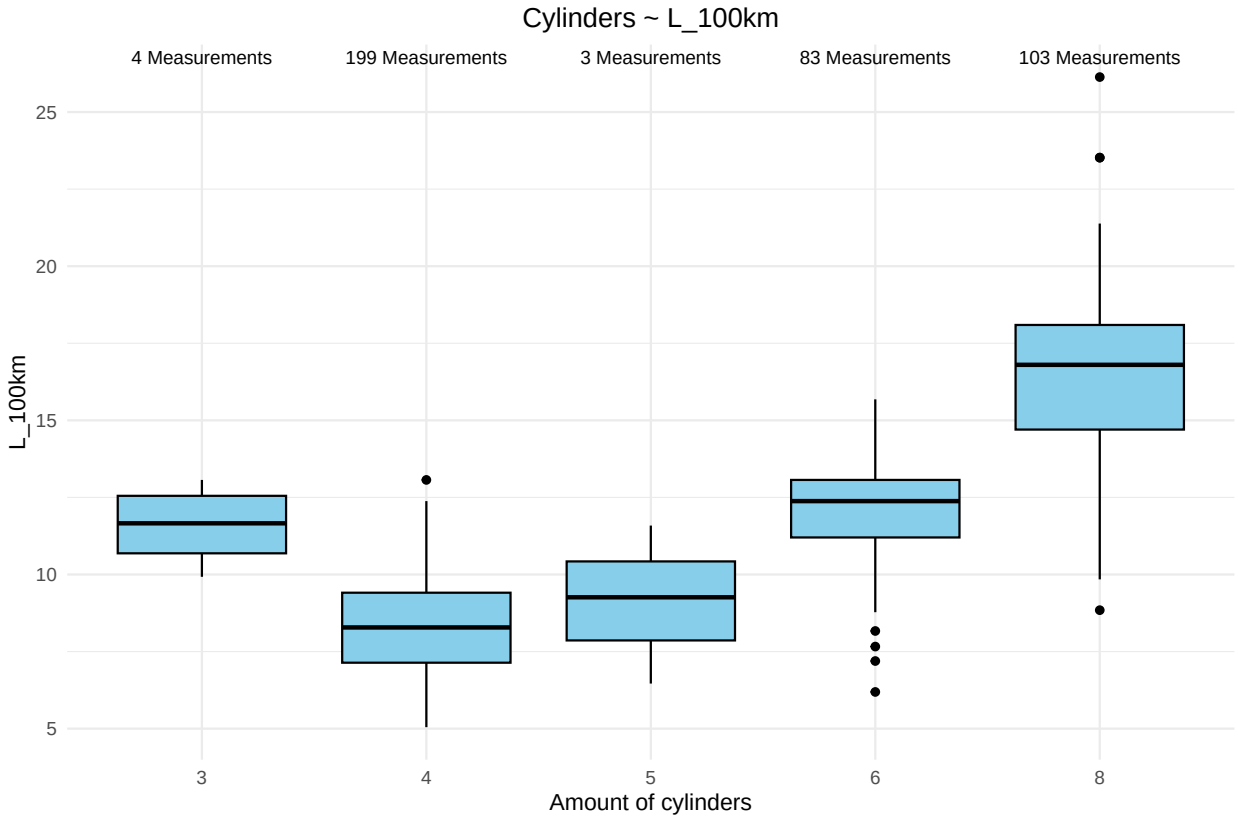

Densityplot of acceleration
density
0.15
0.10
0.05
0.00
10
15
20
25
acceleration

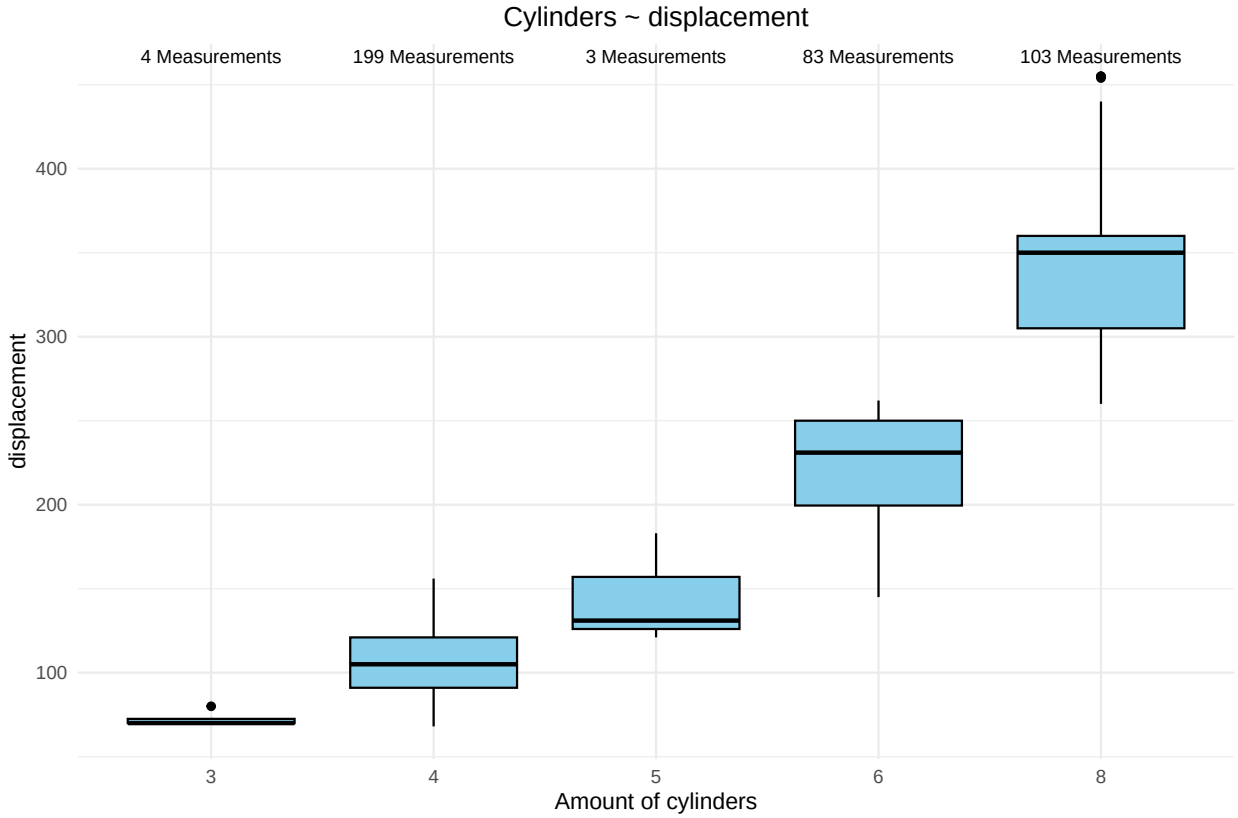
The acceleration variable has a normal distribution. The 4 other numerical variables all have a right skewness.

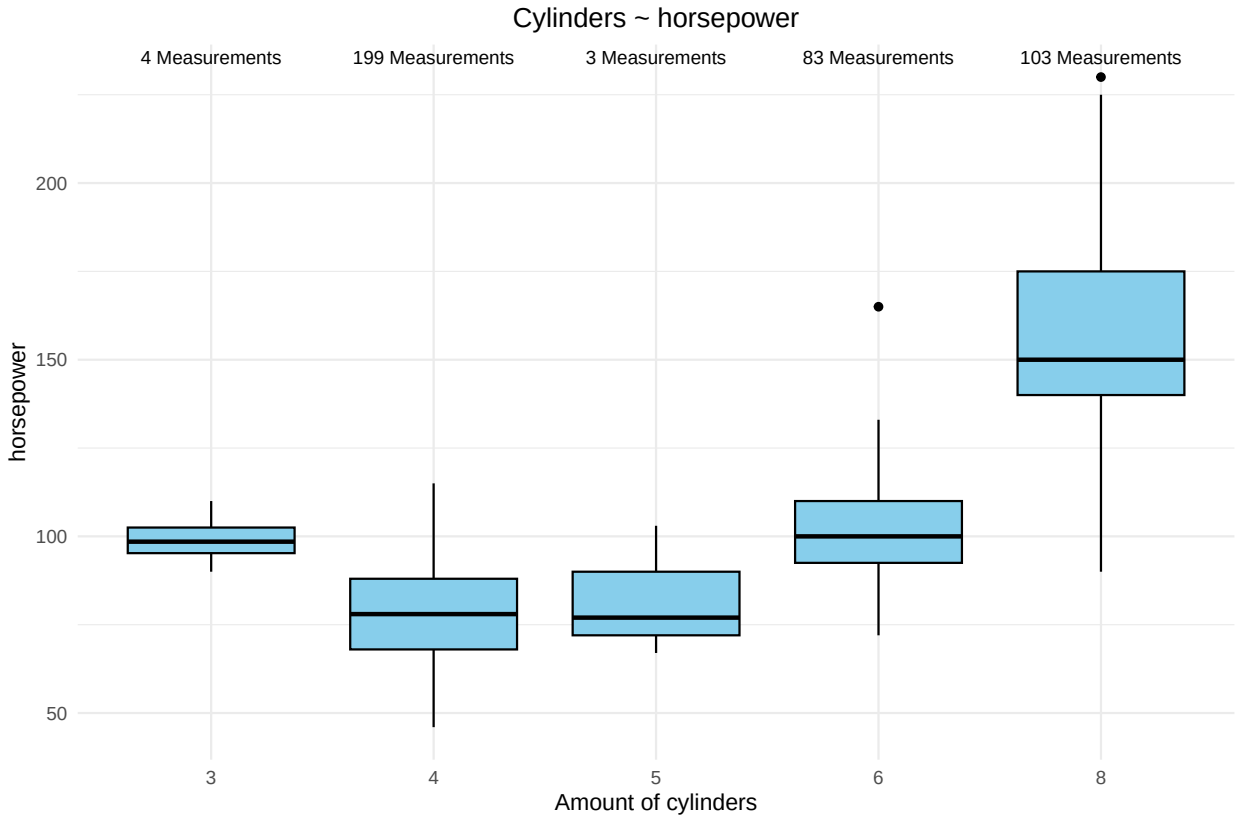
Question 1

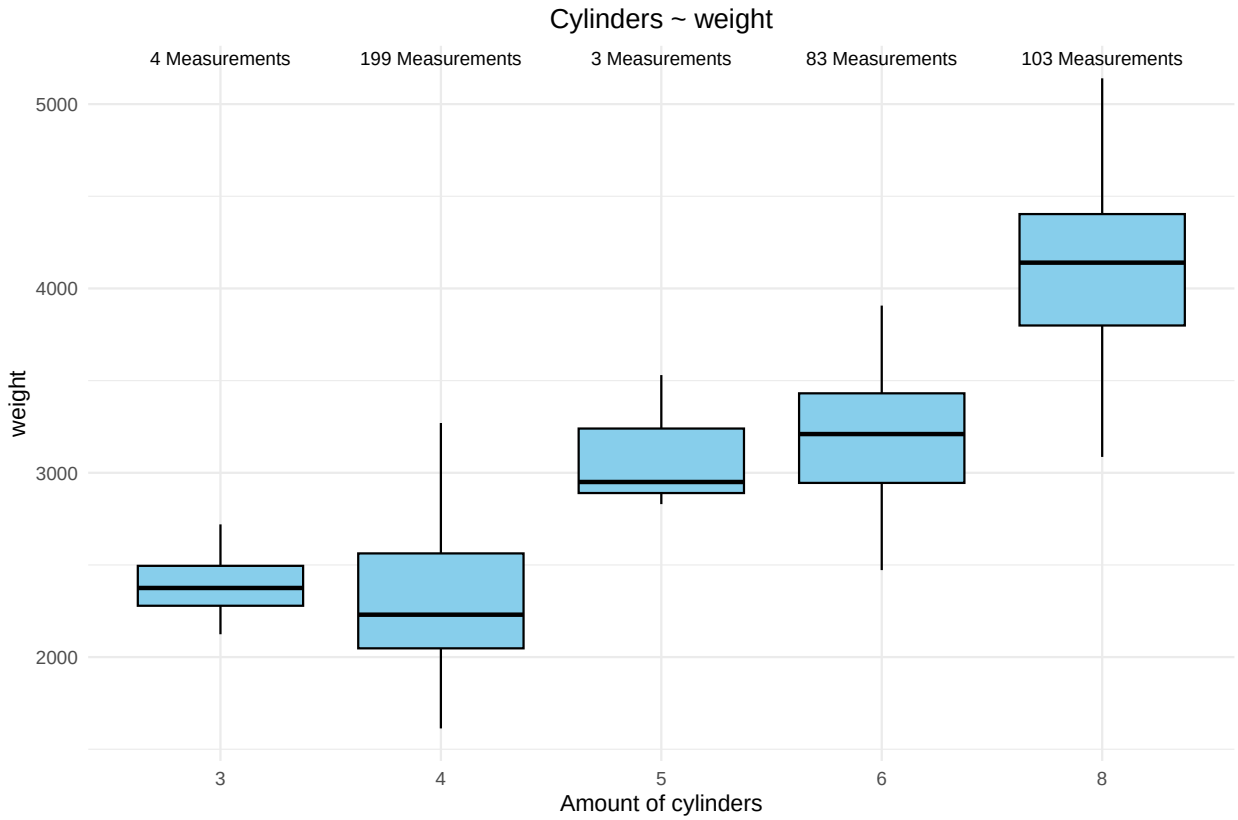
With this dataset, some genereal research questions can be asked. Firstly, do the amount of cylinders have a correlation with the numerical variables? To start, let's make some boxplots of the numerical variables, grouped by the amount of cylinders of the engine.

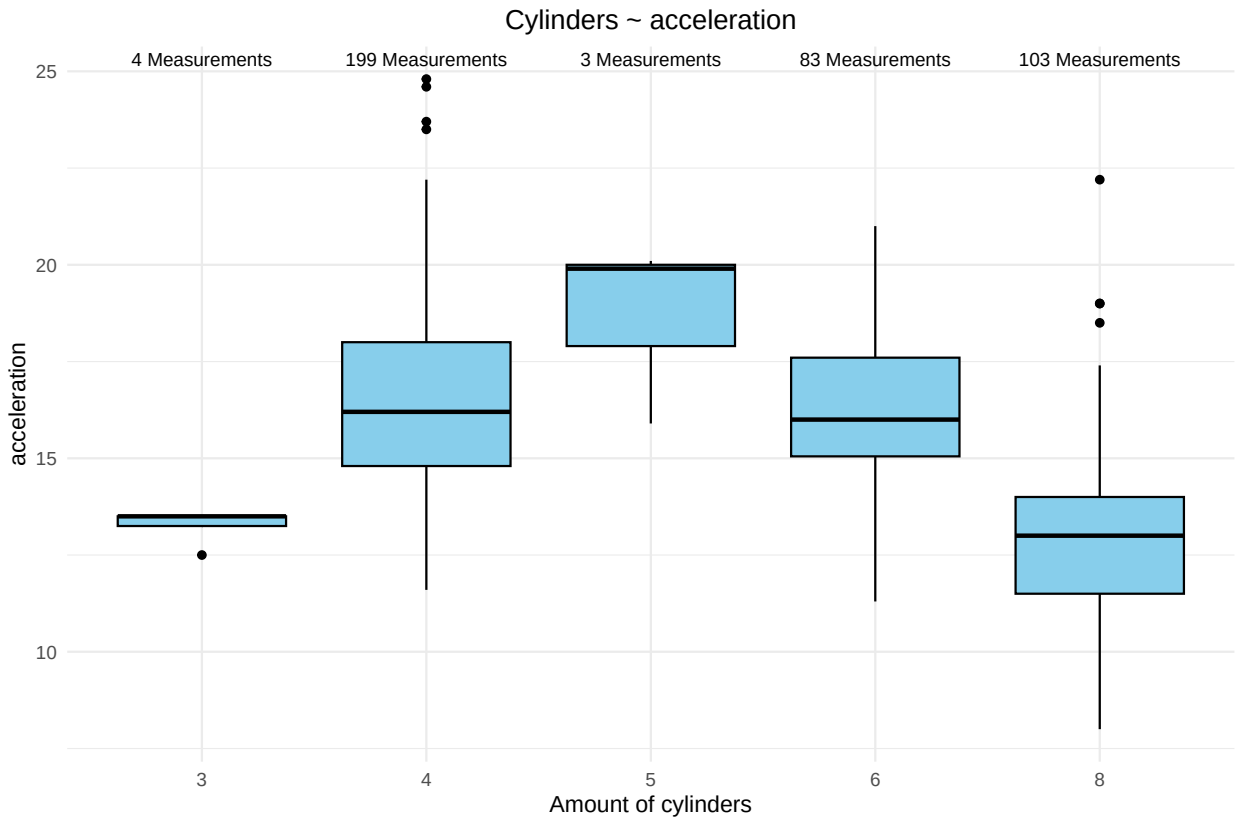
```
for (i in num_var) {
  # Making a table with the frequencies of each number of cylinders
  measurements <- car_data %>%
  group_by(cylinders) %>%
  summarise(count = sum(!is.na(!!sym(i))))

  p <- ggplot(car_data, aes(x = as.factor(cylinders), y = !!sym(i))) +
    geom_boxplot(fill = "skyblue", color = "black") +
    # Adding the frequency for the amount of cylinders
    geom_text(data = measurements, aes(label = paste(count, "Measurements"),
                                       x = as.factor(cylinders),
                                       y = max(car_data[[i]], na.rm = TRUE)),
              vjust = -1, size = 3) +
    labs(title = paste("Cylinders ~", i), x = "Amount of cylinders") +
    theme_minimal() +
     theme(plot.title = element_text(hjust = 0.5))
  plot_list[[length(plot_list) + 1]] <- p
  print(p)
}
```


Cylinders ~ L_100km
4 Measurements
199 Measurements
3 Measurements
83 Measurements
103 Measurements
L_100km
25
20
15
10
5
3
4
5
6
8
Amount of cylinders


Cylinders ~ displacement
4 Measurements
199 Measurements
3 Measurements
83 Measurements
103 Measurements
displacement
400
300
200
100
3
4
5
6
8
Amount of cylinders


Cylinders ~ horsepower
4 Measurements
199 Measurements
3 Measurements
83 Measurements
103 Measurements
200
150
100
50
horsepower
3
4
5
6
8
Amount of cylinders

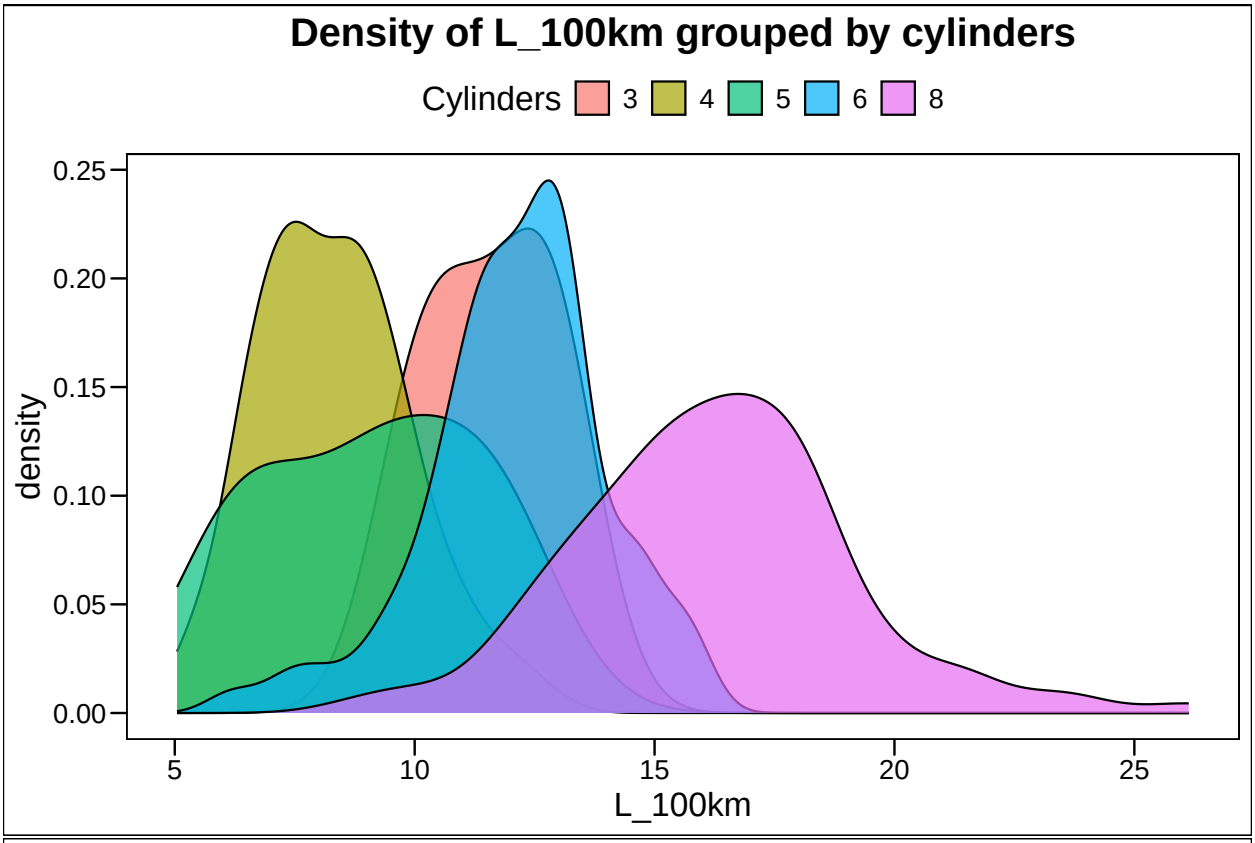

Cylinders ~ weight
4 Measurements
199 Measurements
3 Measurements
83 Measurements
103 Measurements
5000
4000
3000
2000
weight
3
4
5
6
8
Amount of cylinders

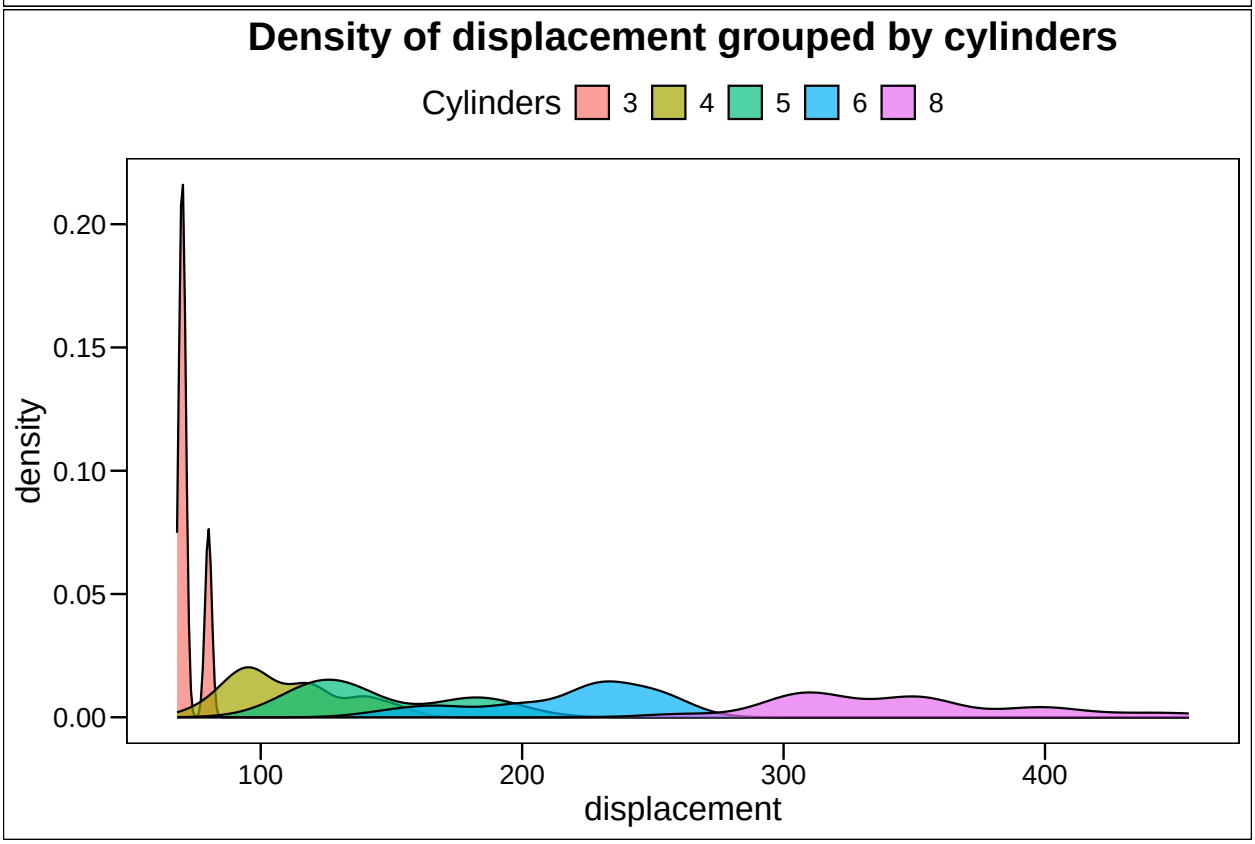


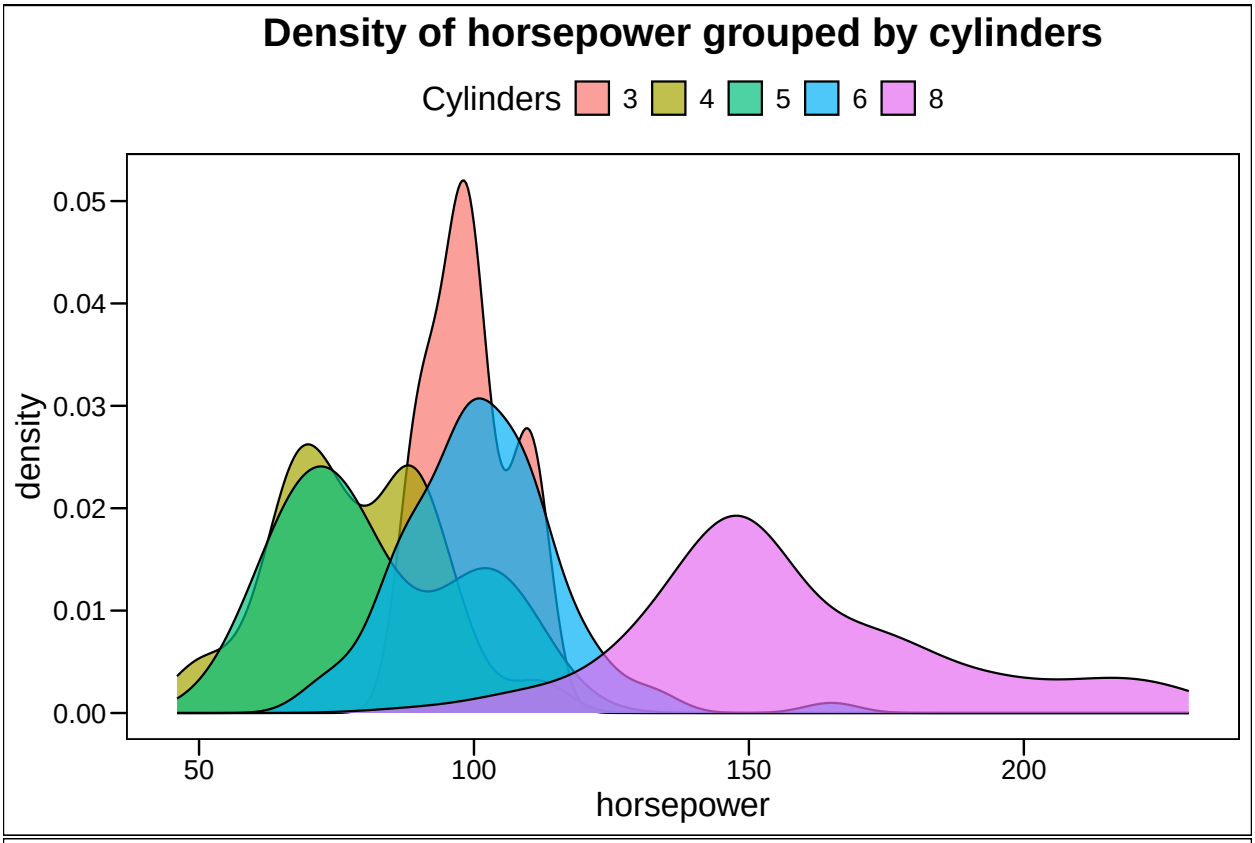
As the plots above depict, the median of fuel consumption (excluding the three cylinders), displacement, horsepower and weight increase with the number of cylinders for all variables, except for the acceleration of the cars.

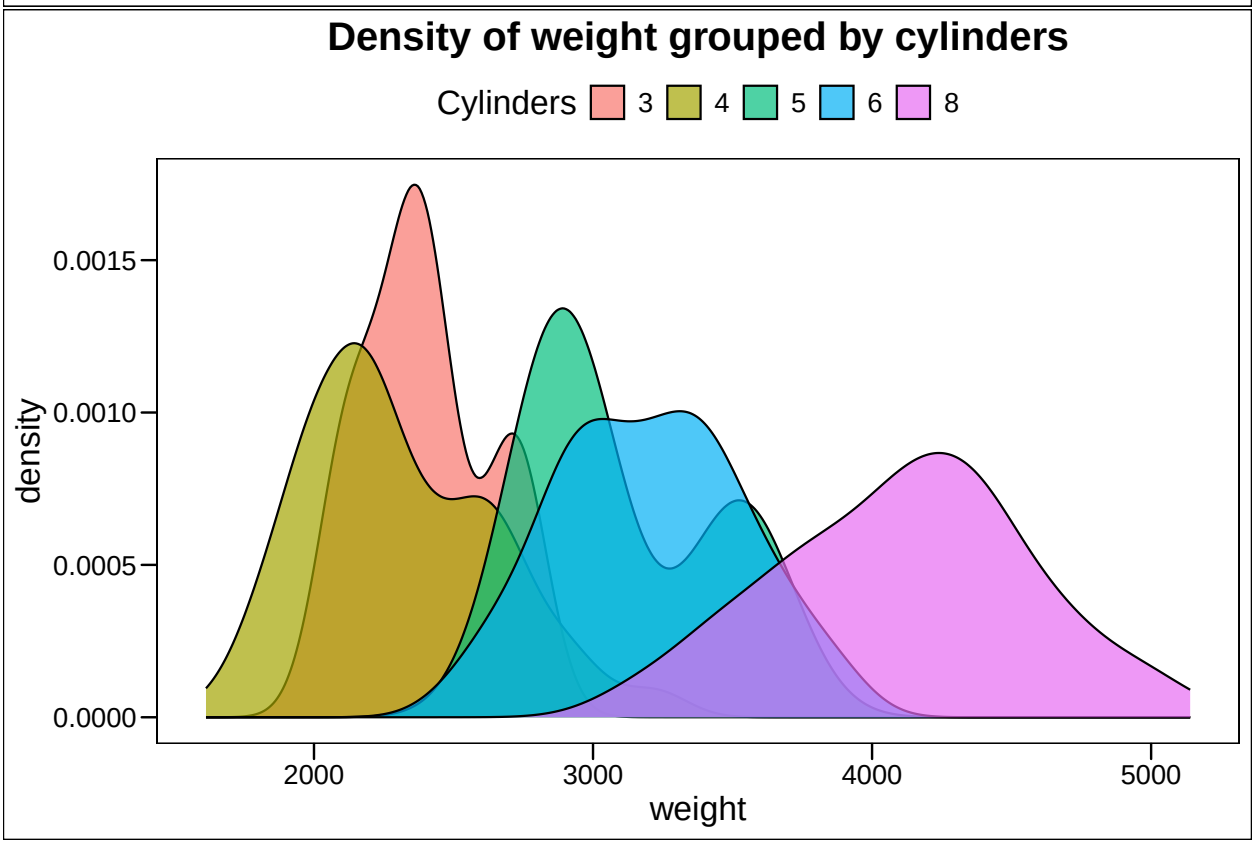
Now, what about the density of each variable plotted over each other? Can we see the same pattern as clearly here?

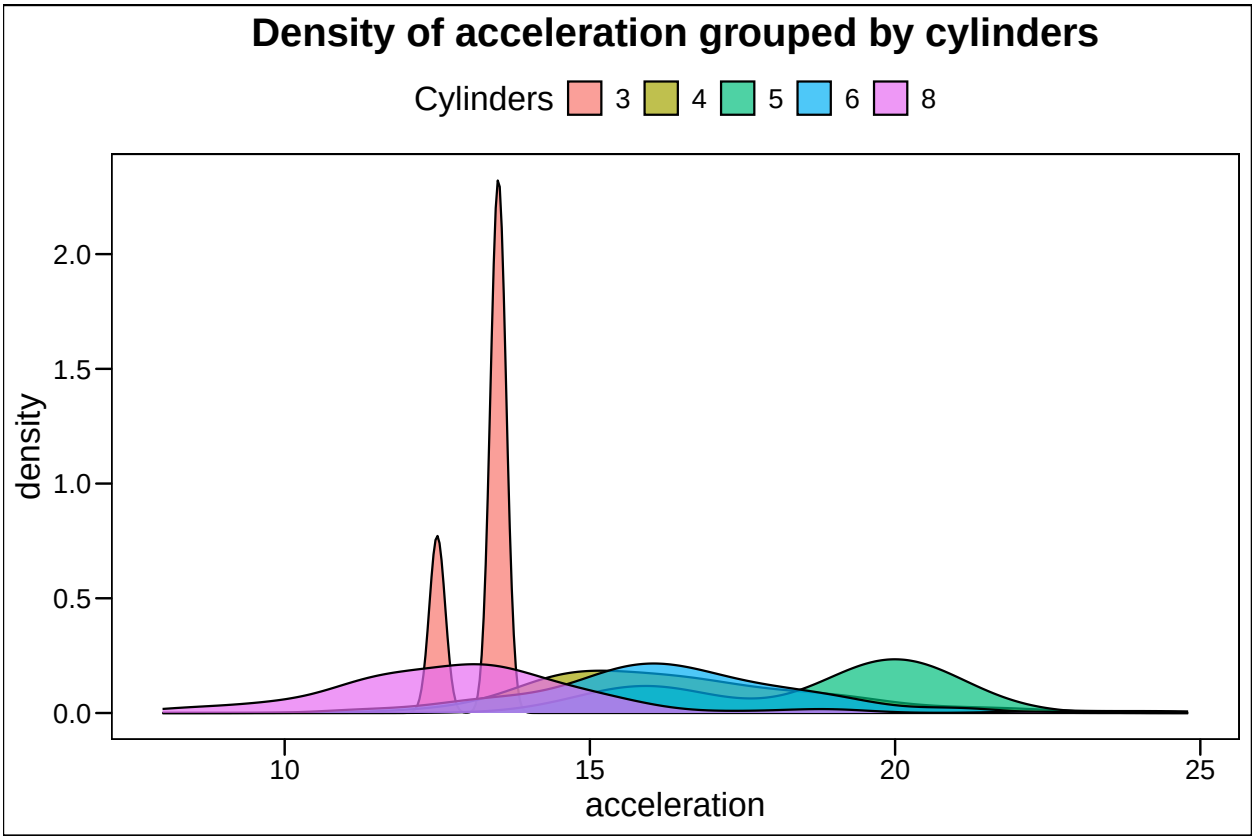
```
for (i in num_var){
  p <- ggplot(data = car_data, aes(x = car_data[,i], fill = as.factor(cylinders))) +
    geom_density(alpha = 0.7) +
    labs(title = paste("Density of",i,"grouped by cylinders"),
         x = i,
         fill = "Cylinders") +
    theme_base() +
    theme(plot.title = element_text(hjust = 0.5)) +
    theme(legend.position = "top")
  print(p)
}
```


Density of L_100km grouped by cylinders
Cylinders 3 4 5 6 8
0.25
0.20
0.15
0.10
0.05
0.00
density
5
10
15
20
25
L_100km


Density of displacement grouped by cylinders
Cylinders 3 4 5 6 8
0.20
0.15
0.10
0.05
0.00
density
100
200
300
400
displacement

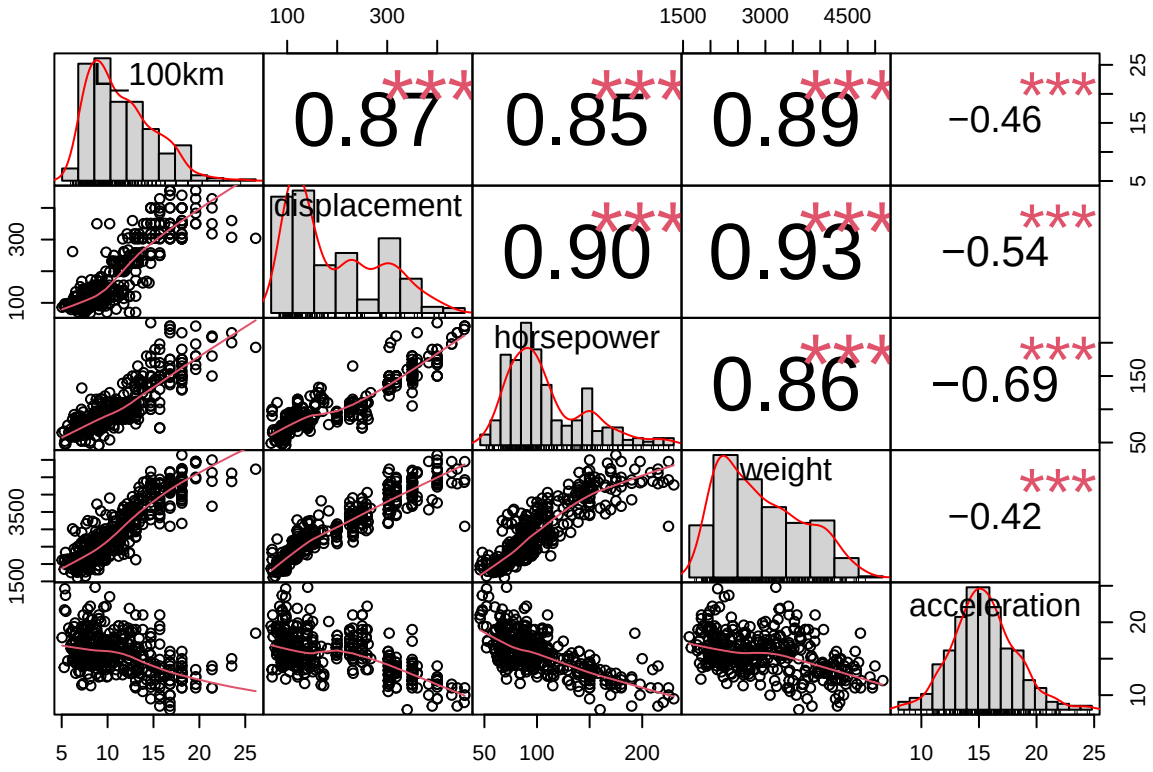

Density of horsepower grouped by cylinders
Cylinders 3 4 5 6 8
density
0.05
0.04
0.03
0.02
0.01
0.00
50
100
150
200
horsepower


Density of weight grouped by cylinders
Cylinders 3 4 5 6 8
density
0.0015
0.0010
0.0005
0.0000
2000
3000
4000
5000
weight



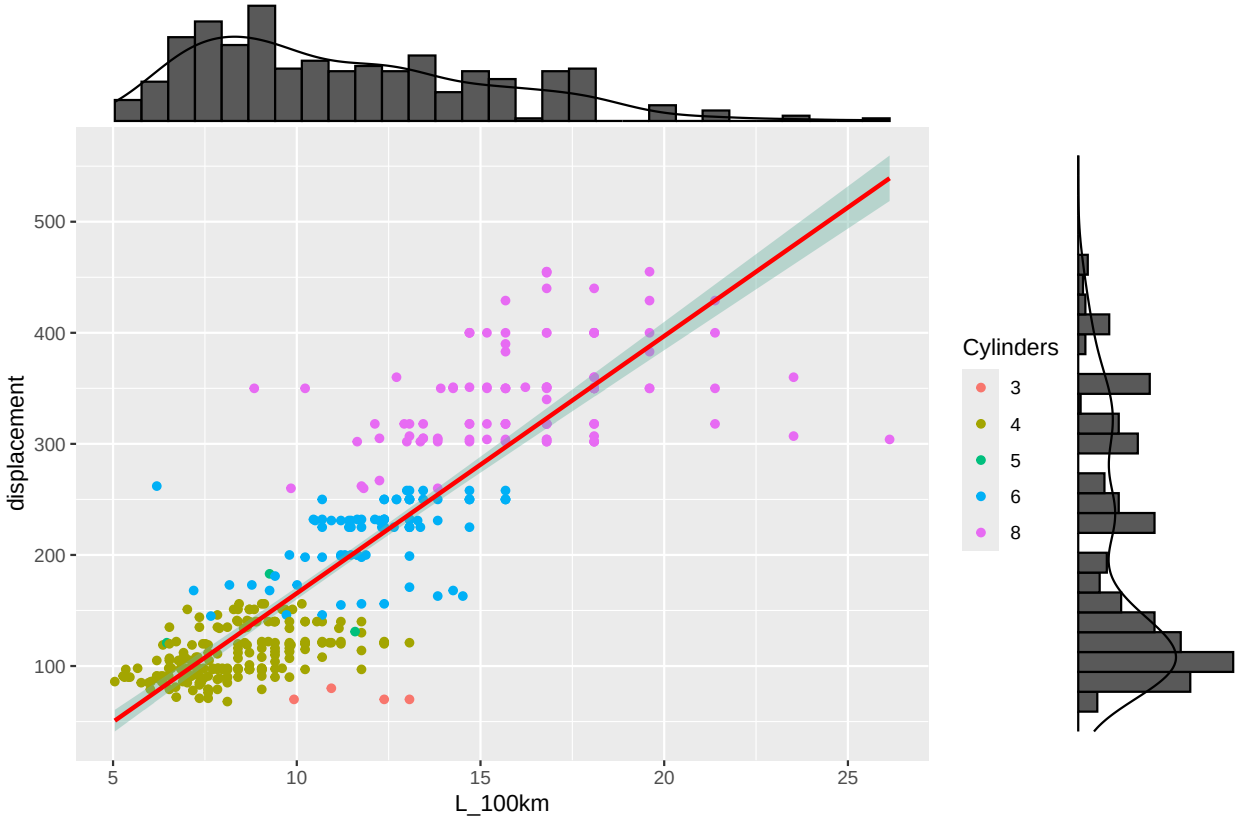
Here again, we observe a similar trend, with the exception of the acceleration variable. The density plots shift to the right as the number of cylinders increases. However, not every density plot is visually pleasing due to the limited sample size for 3 and 5 cylinders. Next, let's further investigate some correlatations, but now via a correlationplot of the numerical variables only.

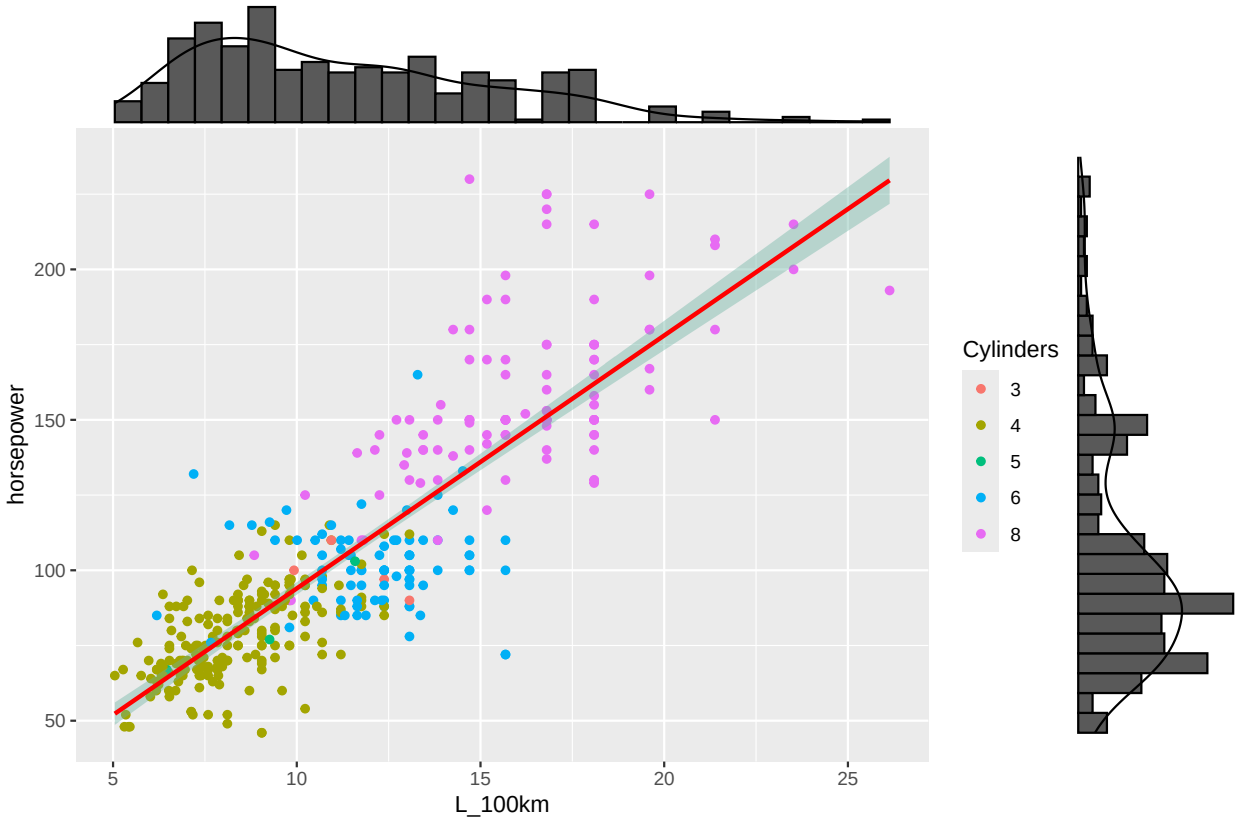
```
chart.Correlation(car_data[num_var],
                  method = "pearson",
                  histogram = TRUE,
                  pch = 16)
```

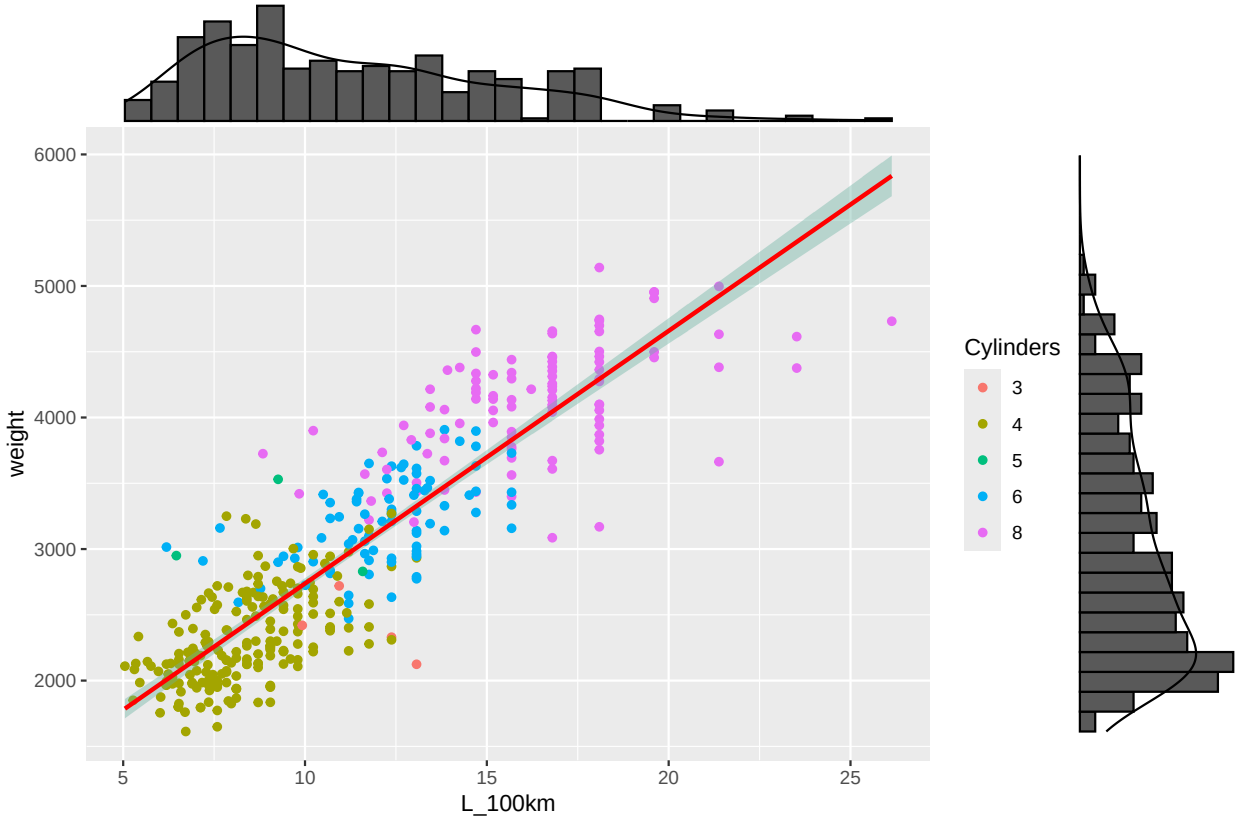


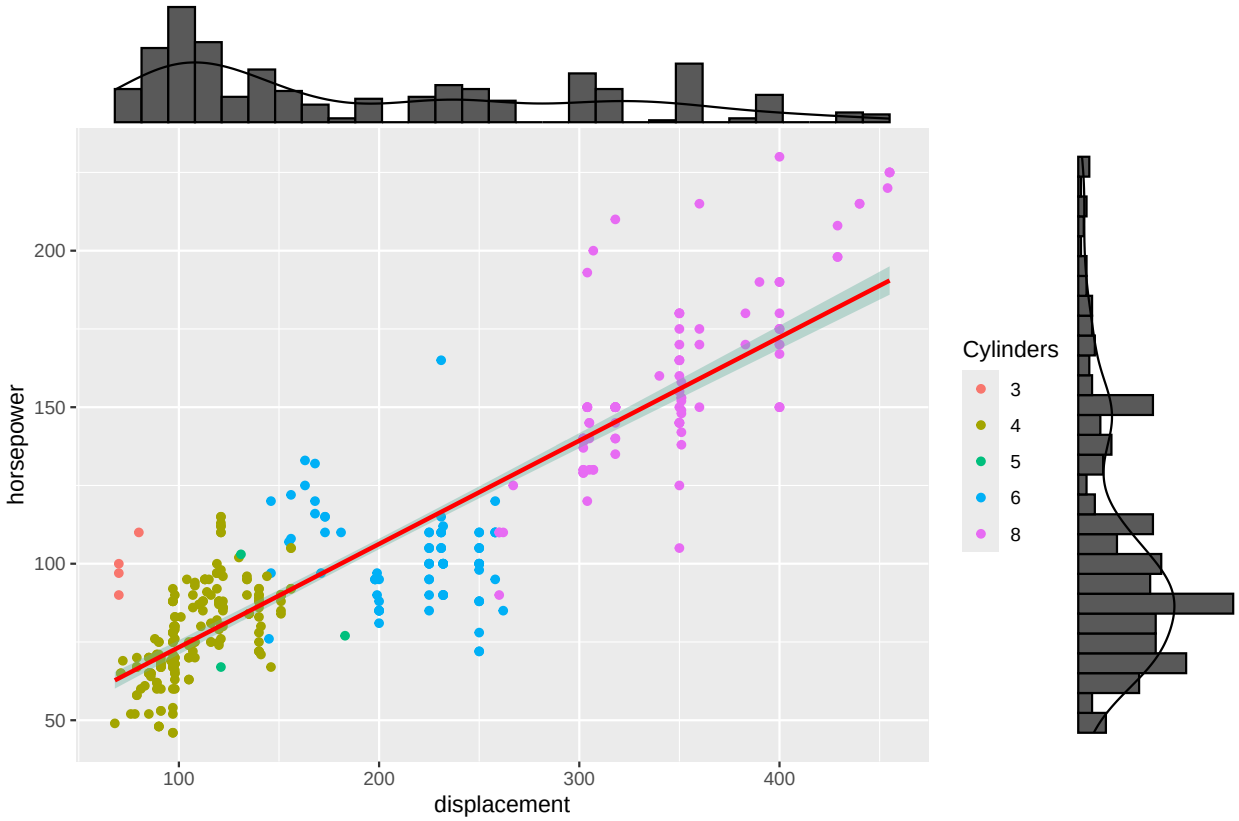
The correlation plot provides a comprehensive overview of the numerical variables that are strongly related through scatterplots, densigrams, and correlation coefficients. Now, let's narrow down our focus to the numerical variables that have an absolute correlation coefficient greater than 0.80.

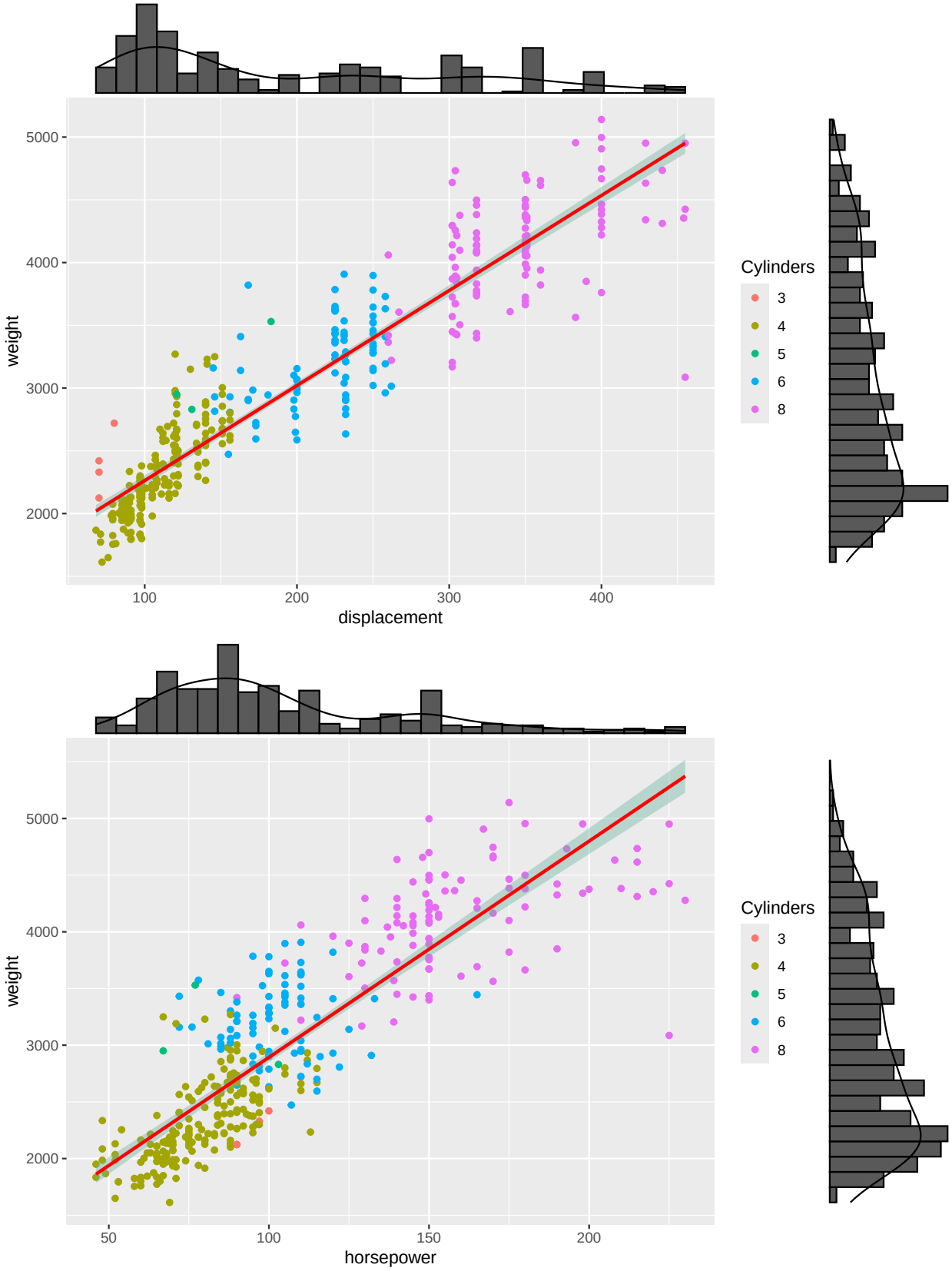
```
# Note that the code for only one plot is shown.
p <- ggplot(car_data ,aes_string(x = "L_100km",
                                 y = "displacement",
                                 color = "cylinders")) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE) +
  theme(legend.position = "right") +
  labs(color = "Cylinders")

p2 <- ggMarginal(p, type = "densigram")
print(p2)
```


displacement
500
400
300
200
100
5
10
15
20
25
L_100km
Cylinders
3
4
5
6
8


horsepower
200
150
100
50
5
10
15
20
25
L_100km
Cylinders
3
4
5
6
8


6000
5000
4000
3000
2000
weight
5
10
15
20
25
L_100km
Cylinders
3
4
5
6
8


200
150
100
50
horsepower
100
200
300
400
displacement
Cylinders
3
4
5
6
8



The red line in the scatter plot represents the fitting of a linear model to the data. The width of the red

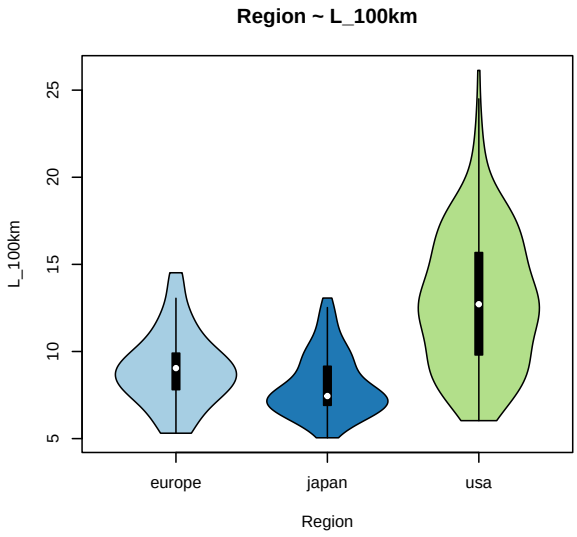
area surrounding the red line indicates the variance in the data for that particular point. A wider red area suggests higher variability in the data at that point.

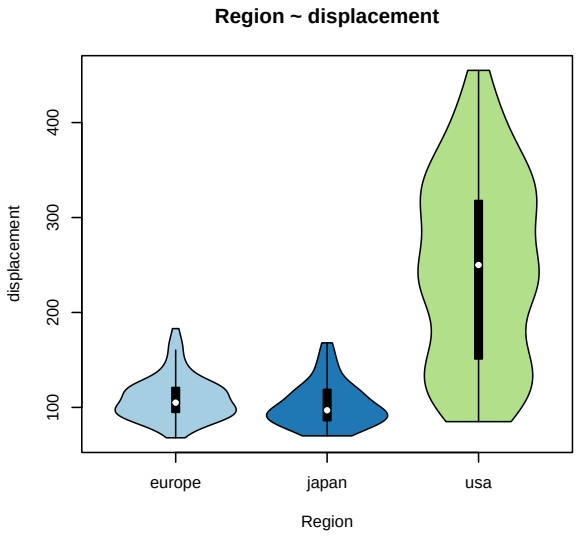
Based on the analysis of the previous plots, it can be concluded that the number of cylinders has a positive impact on the numerical variables. Additionally, the strong correlations observed in the last six plots further support this conclusion.

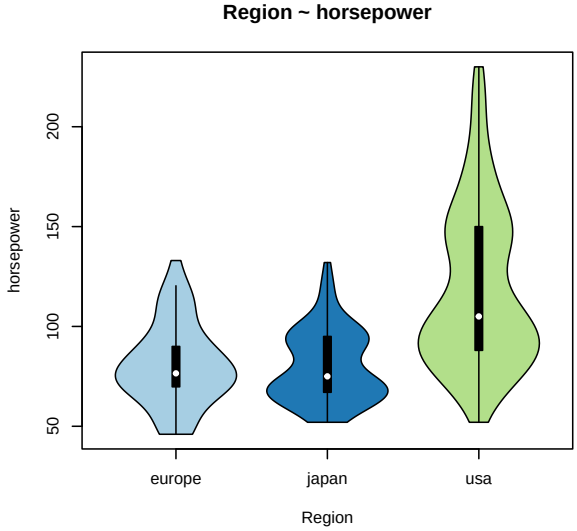
Question 2

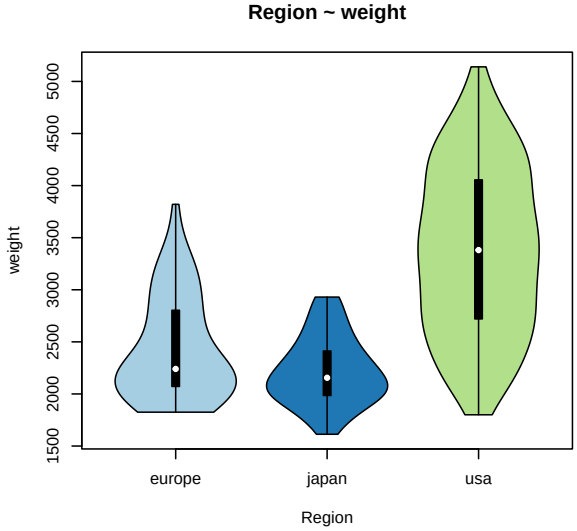
Another research question to explore is whether there are any differences in the numerical variables among the origins of the cars. To investigate this, a violin plot was created.

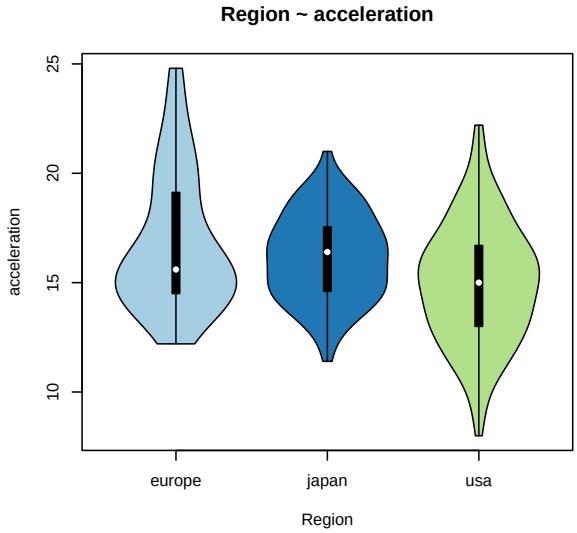
```
par(mfrow = c(3,2))
for (i in num_var) {
  p <- vioplot(car_data[,i] ~ car_data$origin,
               col = brewer.pal(3, "Paired"),
               xlab = "Region",
               ylab = paste(i),
               main = paste("Region ~",i))
  p
}
```


Region ~ L_100km
L_100km
5
10
15
20
25
europe
japan
usa
Region


Region ~ displacement
displacement
100
200
300
400
europe
japan
usa
Region


Region ~ horsepower
horsepower
50
100
150
200
europe
japan
usa
Region

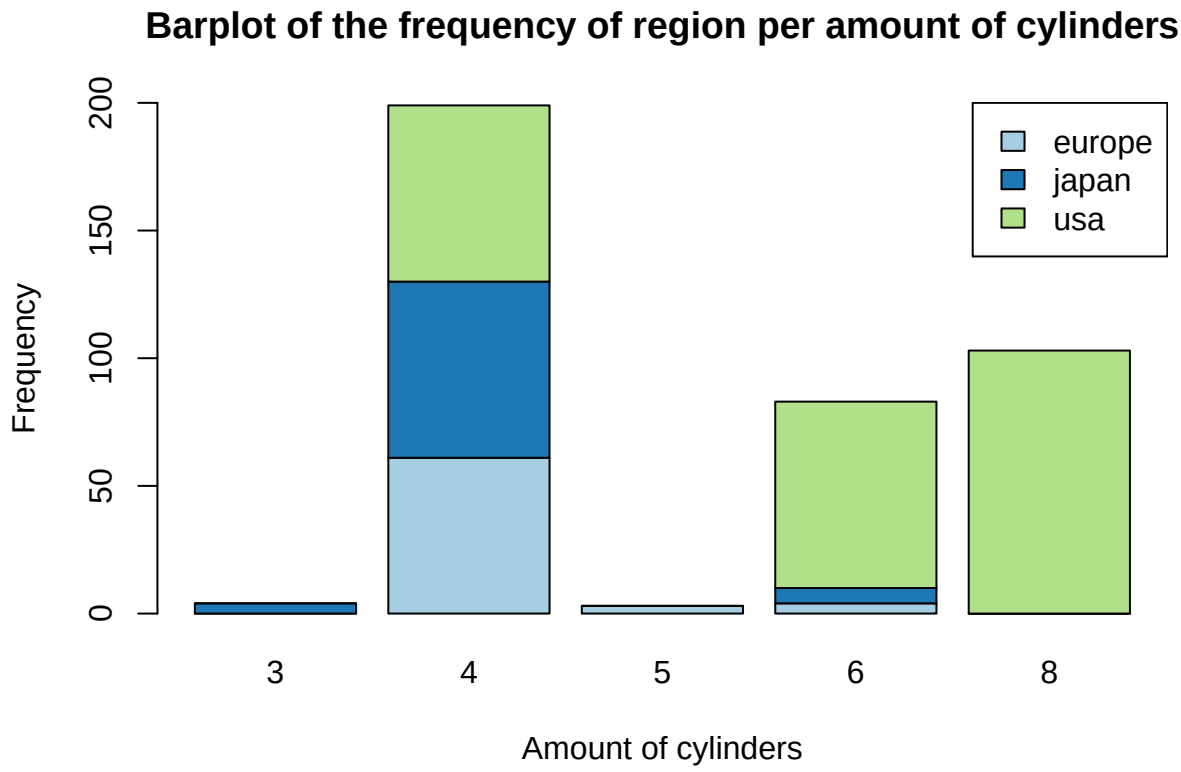

Region ~ weight
weight
1500
2000
2500
3000
3500
4000
4500
5000
europe
japan
usa
Region


Region ~ acceleration
acceleration
10
15
20
25
europe
japan
usa
Region

The violin plots above illustrate that cars in America tend to have higher fuel consumption, larger engines, more horsepower, and greater weight compared to those in Europe and Japan. However, they have similar acceleration characteristics. It's important to note that the dataset contains a higher proportion of cars from the USA, resulting in a larger variance for each numerical variable. This observation is also evident in the violin plots, where the plots for the USA have longer tails compared to the other two regions.

Considering our previous findings, it's plausible that there are more 8-cylinder cars in the USA. Let's explore this hypothesis further with a stacked barplot.

```
barplot(table(car_data[,c("origin", "cylinders")]),
        main = "Barplot of the frequency of region per amount of cylinders",
        xlab = "Amount of cylinders",
        ylim = c(0,200),
        ylab = "Frequency",
        col = brewer.pal(3, "Paired"))
legend("topright", legend = rownames(table(car_data$origin)), fill = brewer.pal(3, "Paired"))
```

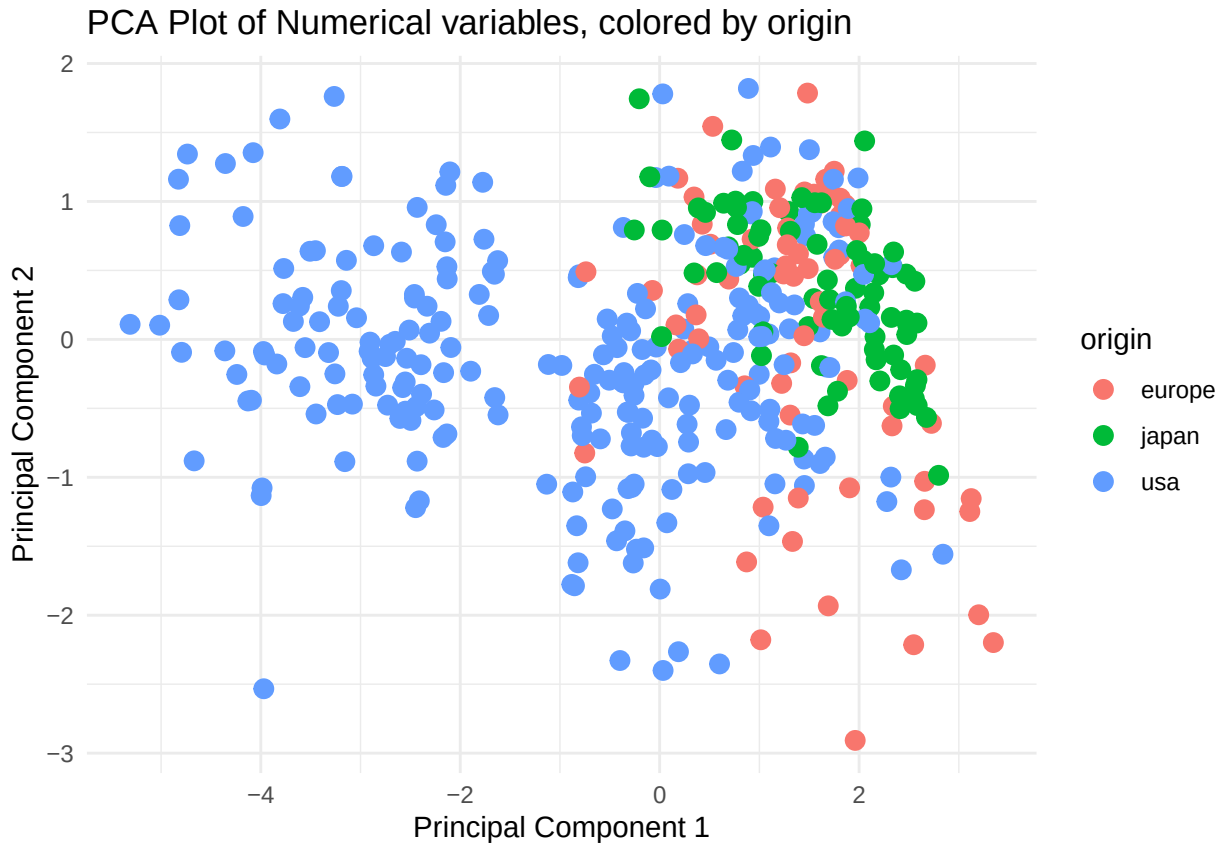


As depicted above, all 8-cylinder and almost all 6-cylinder cars originate from America, indicating heavier and more fuel-consuming engines.

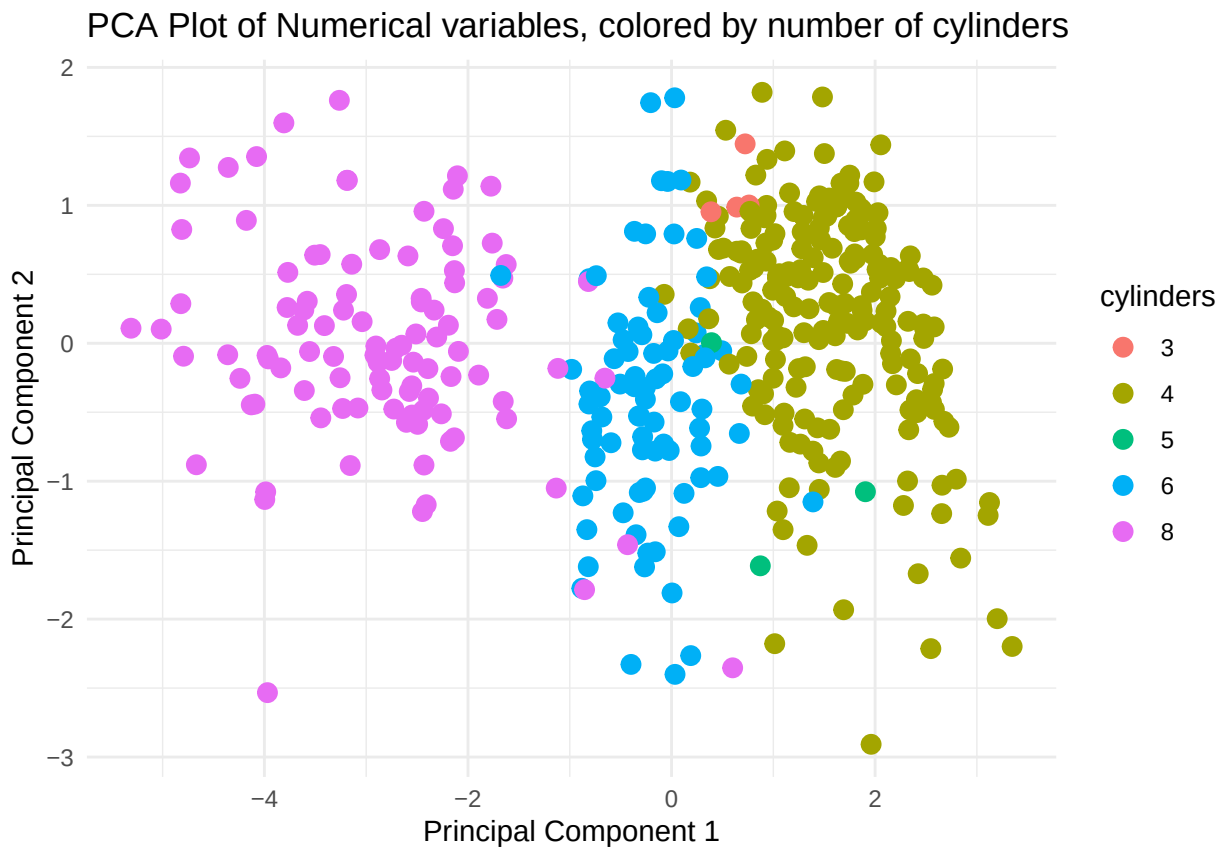
Conclusion via PCA

Lastly, let's investigate whether there is clustering based on the origin of the car or/and based on the number of cylinders, confirming the findings from research questions 1 and 2. To conduct this analysis, a principal component analysis (PCA) was performed.

```r
pca_result <- prcomp(na.omit(car_data[num_var]), scale. = TRUE, center = TRUE)
pca_df <- as.data.frame(pca_result$x)
pca_df <- cbind(pca_df, origin = car_data$origin)

# Create the PCA plot with origin coloring
ggplot(pca_df, aes(x = PC1, y = PC2, color = origin)) +
  geom_point(size = 3) +
  labs(title = "PCA Plot of Numerical variables, colored by origin",
       x = "Principal Component 1",
       y = "Principal Component 2") +
  theme_minimal()
```




```r
pca_df <- cbind(pca_df, cylinders = car_data$cylinders)
# Create the PCA plot with cylinder coloring
ggplot(pca_df, aes(x = PC1, y = PC2, color = cylinders)) +
  geom_point(size = 3) +
  labs(title = "PCA Plot of Numerical variables, colored by number of cylinders",
       x = "Principal Component 1",
       y = "Principal Component 2") +
  theme_minimal()
```



In the PCA plot, two distinct clusters are noticeable—one on the left and one on the right. Upon examining the plot where the number of cylinders is color-coded, a clear separation can be observed on the left side, corresponding to the 8-cylinder category, which is distinct from the rest of the dataset. It's noteworthy that all data points in this cluster belong to the USA region, as evident from the very first PCA plot. Overall, it appears that the number of cylinders serves as a clearer clustering factor compared to the region from which the car originated.