

The Procedure:

First, I went to the website where the data was stored. I obtained the data and from one of the links and put them all in a csv file where there are 9 variables: car name, mpg, cylinders, displacement, horsepower, weight, acceleration, model year and origin. Information of the csv file is given below.

```
> Cars.data<-read.csv("Cars.csv", header=TRUE)
> str(Cars.data)
'data.frame':   406 obs. of  9 variables:
 $ make       : Factor w/ 310 levels "amc.ambassador.brougham",...: 51 38 234 15 163 142 56 226 245 2 ...
 $ mpg        : Factor w/ 130 levels "10","11","12",...: 17 7 17 9 13 7 5 5 7 ...
 $ 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  : Factor w/ 94 levels "100","102","103",...: 16 34 28 28 23 41 46 45 47 39 ...
 $ 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      : int  1 1 1 1 1 1 1 1 1 1 ...
```

Findings:

The first factor I used to obtain information was origin because it was the least distinct with only three different countries (US, Europe and Japan). Organizing the data using origin as a factor allows us to see the different specs of cars of different origin.

```
> summary(cars.USA)
      mpg      cylinders      displacement      horsepower      weight      acceleration      model.year
Min.   : 9.00   Min.   :4.000   Min.   : 85.0   Min.   : 52.0   Min.   :1800   Min.   : 8.00   Min.   :70.0
1st Qu.:15.00   1st Qu.:4.000   1st Qu.:151.0   1st Qu.: 88.0   1st Qu.:2721   1st Qu.:13.00   1st Qu.:72.0
Median :18.50   Median :6.000   Median :250.0   Median :106.0   Median :3380   Median :15.00   Median :75.0
Mean   :20.08   Mean   :6.283   Mean   :247.9   Mean   :119.9   Mean   :3373   Mean   :14.94   Mean   :75.5
3rd Qu.:24.00   3rd Qu.:8.000   3rd Qu.:318.0   3rd Qu.:150.0   3rd Qu.:4055   3rd Qu.:16.70   3rd Qu.:78.0
Max.   :39.00   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140   Max.   :22.20   Max.   :82.0
NA's    :5
NA's     :4

> summary(cars.JPN)
      mpg      cylinders      displacement      horsepower      weight      acceleration      model.year
Min.   :18.00   Min.   :3.000   Min.   : 70.0   Min.   : 52.00   Min.   :1613   Min.   :11.40   Min.   :70.00
1st Qu.:25.70   1st Qu.:4.000   1st Qu.: 86.0   1st Qu.: 67.00   1st Qu.:1985   1st Qu.:14.60   1st Qu.:74.00
Median :31.60   Median :4.000   Median : 97.0   Median : 75.00   Median :2155   Median :16.40   Median :78.00
Mean   :30.45   Mean   :4.101   Mean   :102.7   Mean   : 79.84   Mean   :2221   Mean   :16.17   Mean   :77.44
3rd Qu.:34.05   3rd Qu.:4.000   3rd Qu.:119.0   3rd Qu.: 95.00   3rd Qu.:2412   3rd Qu.:17.55   3rd Qu.:81.00
Max.   :46.60   Max.   :6.000   Max.   :168.0   Max.   :132.00   Max.   :2930   Max.   :21.00   Max.   :82.00

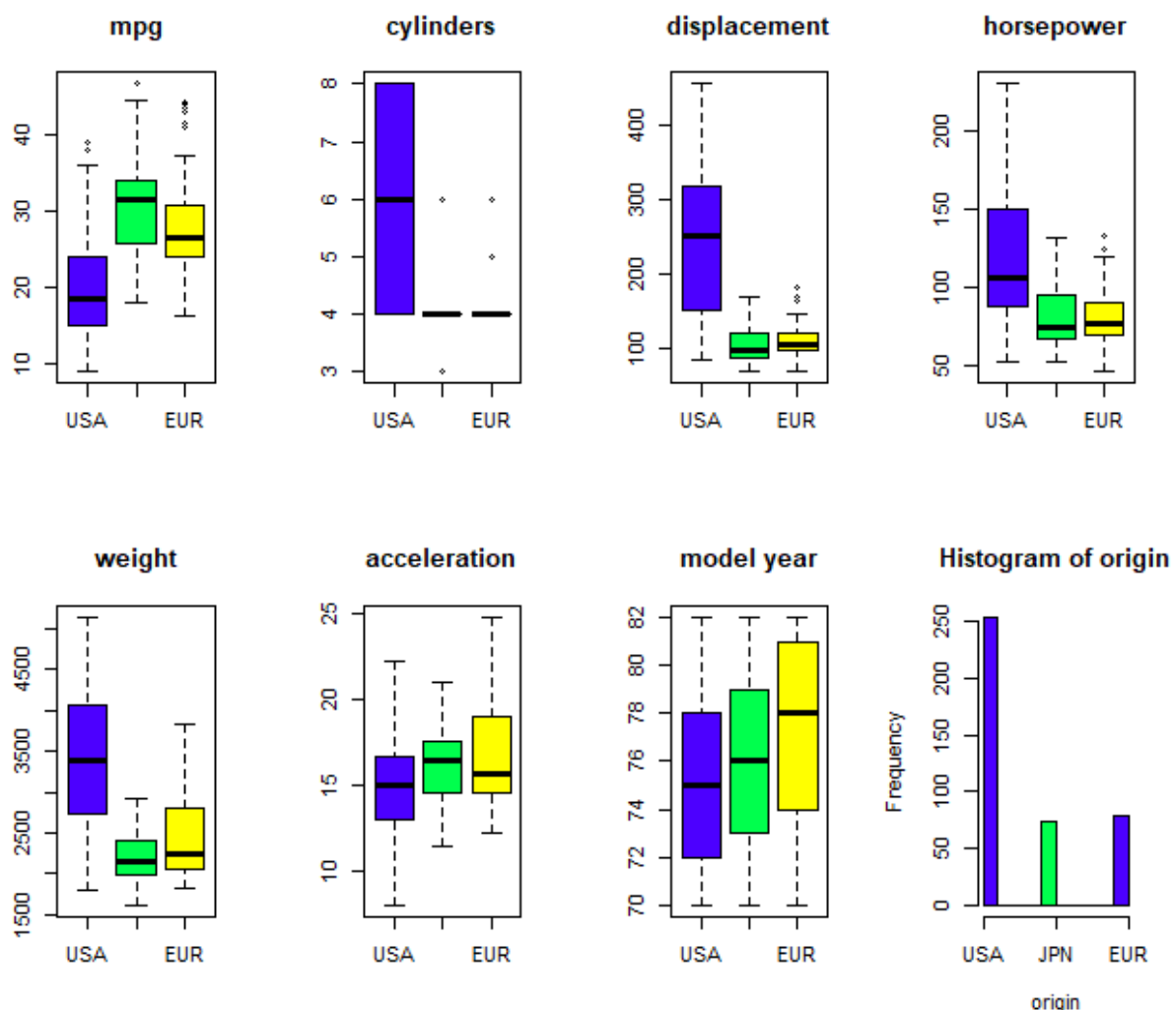
> summary(cars.EUR)
      mpg      cylinders      displacement      horsepower      weight      acceleration      model.year
Min.   :16.20   Min.   :4.000   Min.   : 68.0   Min.   : 46.0   Min.   :1825   Min.   :12.20   Min.   :70.00
1st Qu.:24.00   1st Qu.:4.000   1st Qu.: 96.0   1st Qu.: 69.5   1st Qu.:2065   1st Qu.:14.50   1st Qu.:73.00
Median :26.50   Median :4.000   Median :105.0   Median : 77.0   Median :2246   Median :15.70   Median :76.00
Mean   :27.89   Mean   :4.151   Mean   :109.5   Mean   : 81.0   Mean   :2431   Mean   :16.82   Mean   :75.74
3rd Qu.:30.65   3rd Qu.:4.000   3rd Qu.:121.0   3rd Qu.: 90.5   3rd Qu.:2800   3rd Qu.:19.00   3rd Qu.:79.00
Max.   :44.30   Max.   :6.000   Max.   :183.0   Max.   :133.0   Max.   :3820   Max.   :24.80   Max.   :82.00
NA's    :3
NA's     :2
```

From these summaries of US cars, European cars and Japan cars, we can see a lot of different things.

The average mpg for USA cars is 20.08 whereas for Japan and European cars it is 30.45, and 27.89 respectively. Cars with a higher mpg are more gas efficient so on average we can see that Japanese cars are more gas efficient than the USA and European cars. Next, we can see the information about cylinders by factor of origin. The average of cylinders for American, Japanese and European cars are 6.283, 4.101 and 4.151 respectively. Since the cylinder of a car is a power unit of an engine we can say that American cars have more powerful cars, which we can see further through the horsepower variable where the American cars have a much higher average horsepower. Moreover, engines are measured by displacement, usually expressed in liters (L) or cubic centimeters (cc). Displacement is the total volume of all the cylinders in an engine. The displacement variable further shows that the American cars are

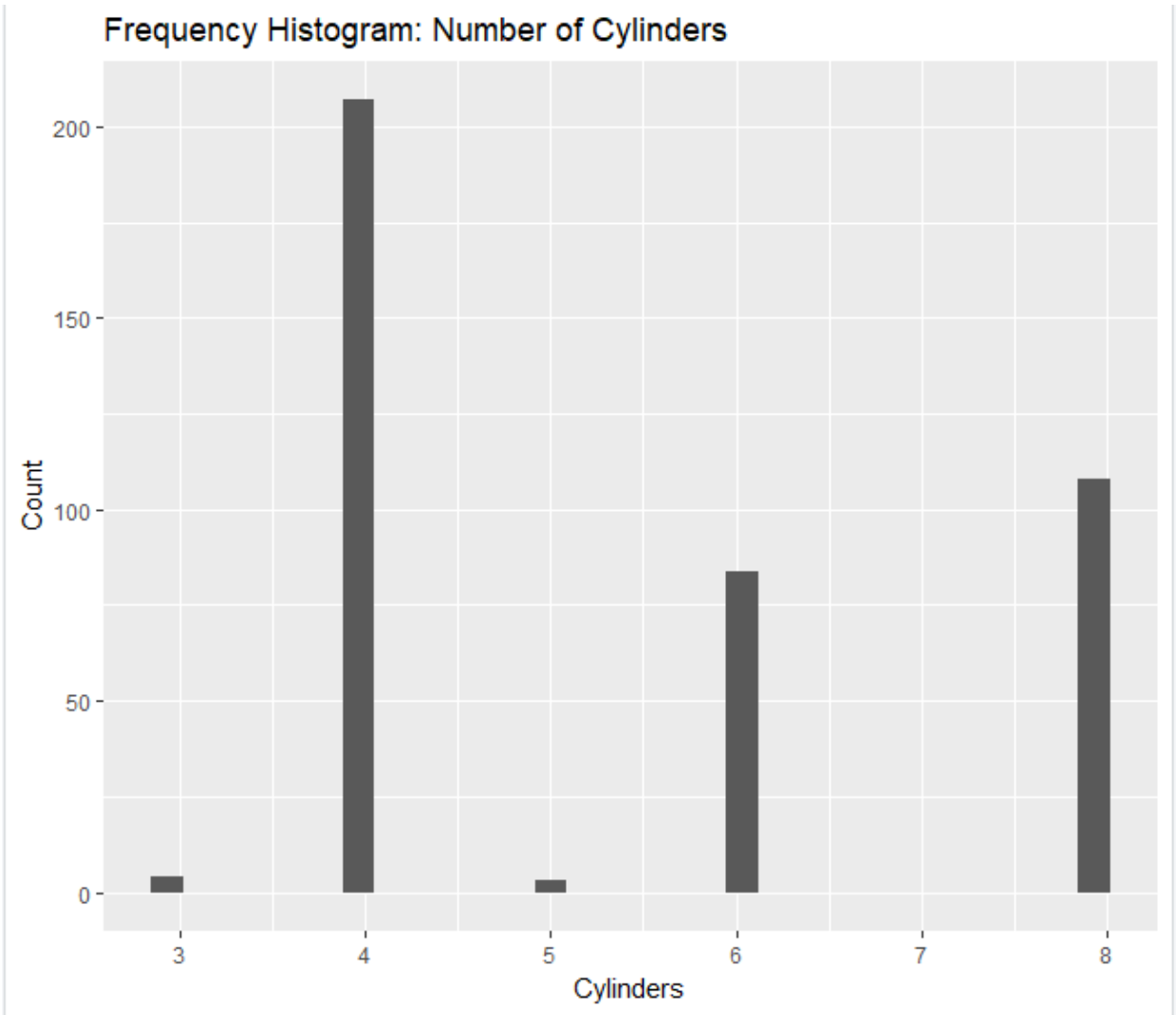
more powerful, and we can also say the more cylinders there are, the higher the displacement and therefore horsepower. In addition, we can see the relationships between origin and weight where we can see that average weight for American cars are heavier than the others. Then, the acceleration we can see that American cars have a lower acceleration than the other two. This could mean that there is a correlation between weight and acceleration but not necessarily horsepower. Lastly, we have the relationship between the origin and the model years.

This can all be further visualized through boxplots and histograms. We can see all the relationships with origin as a factor which helps us obtain data about American, European and Japanese cars. The last histogram shows us the frequency of the origins and this shows us that the majority of the data retrieved is from USA cars;



With the data so far, we can answer a good amount of questions, such as: Which origin of cars has the most powerful ones, which one has the newest ones, which one is more gas efficient, which has a better acceleration, and which are more abundant in this data.

If we want to obtain data without factoring by origin, then we can use univariate plots to make general findings. For example, we can obtain a frequency histogram using the cylinders variable to see the number of cylinders that is most prevalent in this data.



We can do this on all variables but in my opinion, I think it would be better to find correlations between variables. Which is what we will see next

	weight	cylinders	displacement	horsepower	acceleration
weight	1.000000	0.896632	0.932628	0.866586	-0.429479
cylinders	0.896632	1.000000	0.951882	0.844158	-0.521793
displacement	0.932628	0.951882	1.000000	0.898326	-0.558107
horsepower	0.866586	0.844158	0.898326	1.000000	-0.697124
acceleration	-0.429479	-0.521793	-0.558107	-0.697124	1.000000

Using the correlation function we can see the pairwise correlations between these variables. We can see that weight & cylinders, weight & displacement, and weight & horsepower are all correlated with more than 80% correlation. We can do the same with all the variables to find the pairwise correlations and we the lowest of them comes from acceleration. Therefore, we can say all these variables are highly correlated with each other except for acceleration, which further proves what we were able to see in the boxplots and summaries.

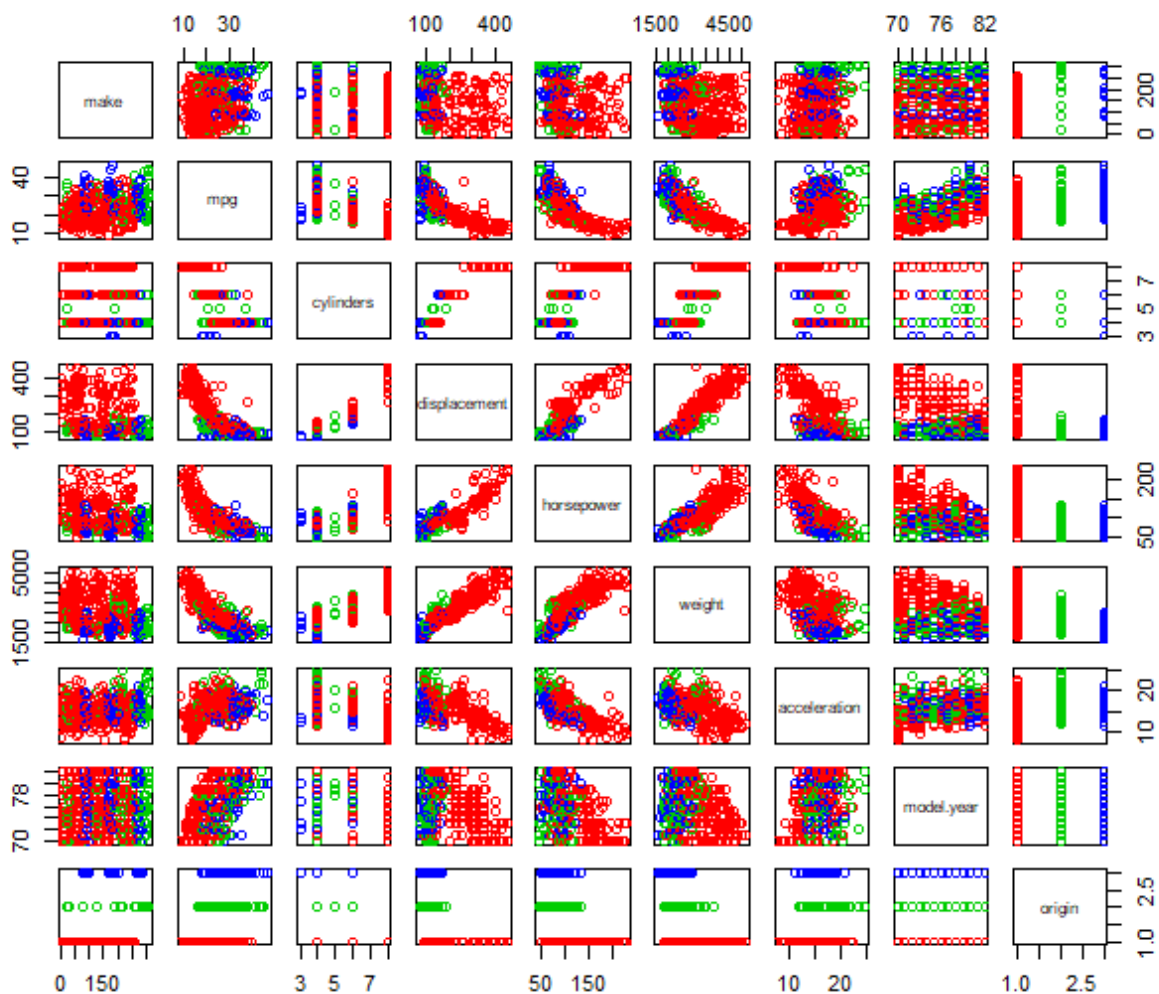
Next, we can see the correlation with the remaining variables: weight and mpg. From the boxplots we were able to see that they were correlated. Using the correlation function we see that

```

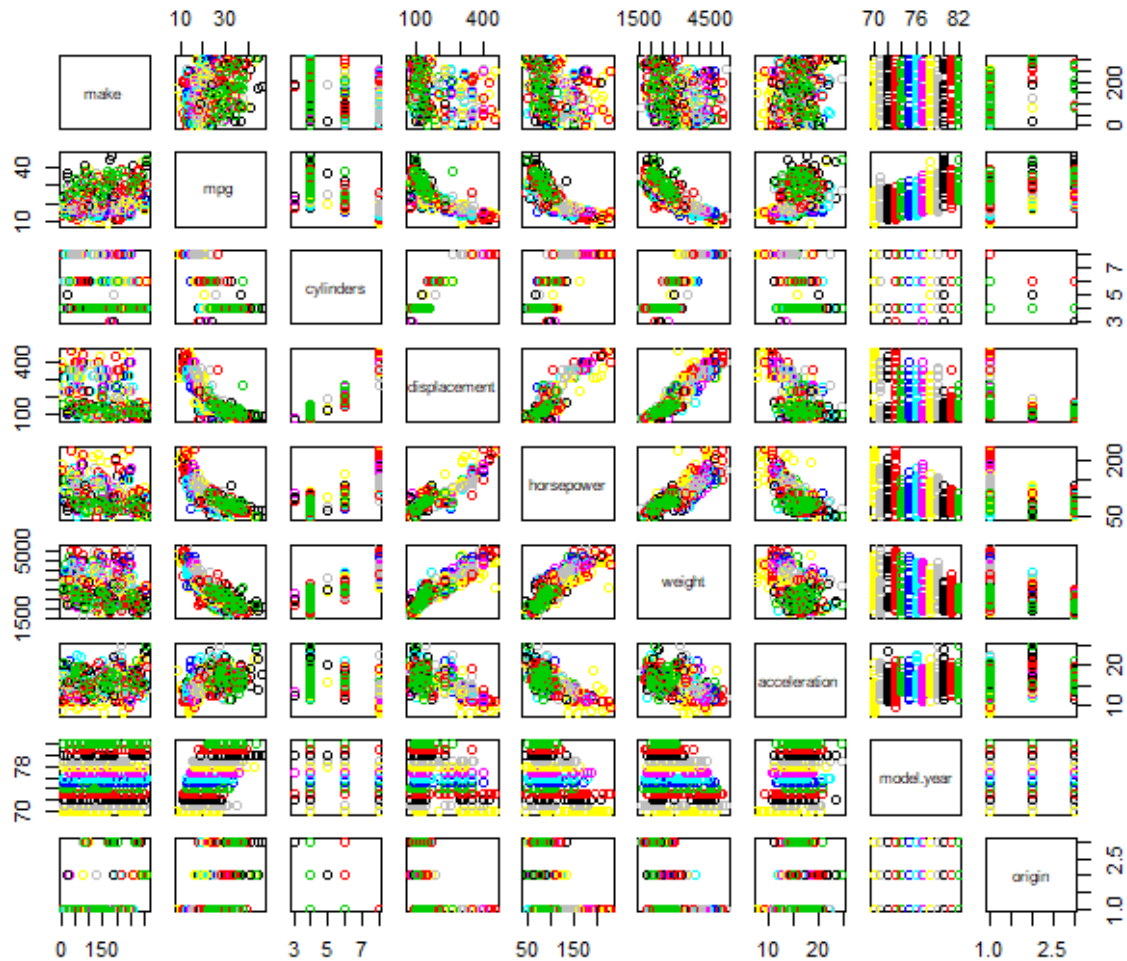
      weight      mpg
weight 1.000000 -0.8317409
mpg    -0.8317409 1.0000000

```

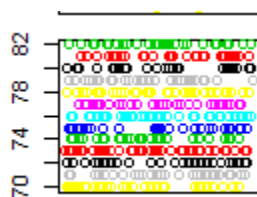
Weight and mpg are inversely correlated which makes sense since heavier cars would need more fuel, so the higher the weight, the less the mpg.



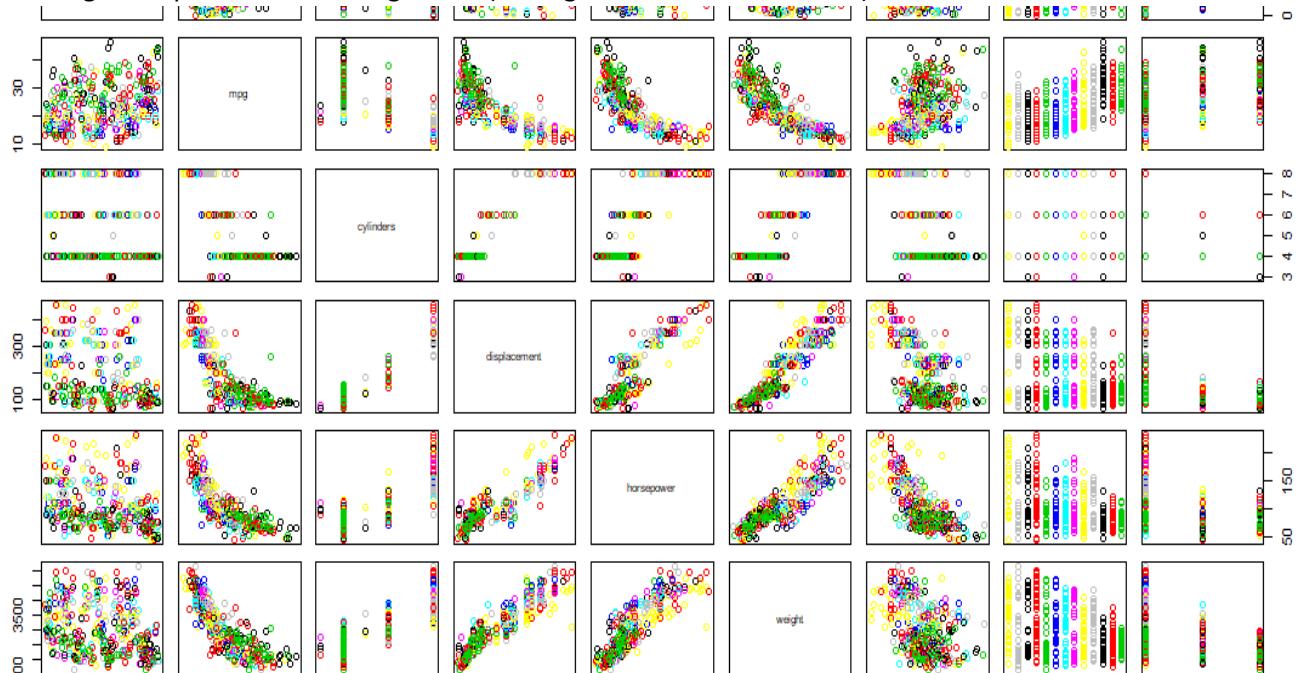
We can also see relationships between origins and other variables by using scatterplots where red is USA, green is European, and blue is Japanese. Moreover, we can use the same visual techniques and this time factor by model year, so we can see the evolution of specs as time increases.



In the second last row we can distinguish the colours with the model years by using the legend. The newest cars are green.



In the mpg row and weight column you can see that smaller/lighter cars with a higher mpg are becoming more prevalent as time goes on (since green are the newer cars)



Also, since the car names are available to us, we can find the top 3 cars for certain categories. The ones that seem important to me are mpg, horsepower, and acceleration.

```
> mpg_sorted <- Cars.data[order(Cars.data$mpg),]
> view(mpg_sorted)
> view(mpg_sorted)
> mpg_sorted <- Cars.data[order(-Cars.data$mpg),]
> head(mpg_sorted,3)
```

	make	mpg	cylinders	displacement	horsepower	weight	acceleration	model.year	origin
330	mazda.glc	46.6	4	86	65	2110	17.9	80	3
337	honda.civic.1500.gl	44.6	4	91	67	1850	13.8	80	3
333	vw.rabbit.c.(diesel)	44.3	4	90	48	2085	21.7	80	2

```
> horsepower_sorted<-Cars.data[order(-Cars.data$horsepower),]
> horsepower_sorted<-Cars.data[order(-Cars.data$horsepower),]
> head(horsepower_sorted,3)
```

	make	mpg	cylinders	displacement	horsepower	weight	acceleration	model.year	origin
124	pontiac.grand.prix	16	8	400	230	4278	9.5	73	1
9	pontiac.catalina	14	8	455	225	4425	10.0	70	1
20	buick.estate.wagon.(sw)	14	8	455	225	3086	10.0	70	1

```
> acceleration_sorted<-Cars.data[order(-Cars.data$acceleration),]
> head(acceleration_sorted,3)
```

	make	mpg	cylinders	displacement	horsepower	weight	acceleration	model.year	origin
307	peugeot.504	27.2	4	141	71	3190	24.8	79	2
403	vw.pickup	44.0	4	97	52	2130	24.6	82	2
334	vw.dasher.(diesel)	43.4	4	90	48	2335	23.7	80	2

These are the top 3 for each of those categories. It's funny to see how each origin dominates a certain one of those categories as if each country focuses on a certain aspect of a car. As we can see Japanese cars excel in gas efficiency, with the Mazda GLC having the highest mpg at 46.6. The Americans excel at making the most powerful cars with the Pontiac Grand Prix having the highest horsepower at 230. And the Europeans excel at making the car with the highest acceleration with the Peugeot 504 having the highest acceleration at 24.8 (m/s?)

Lastly, we can see the summary data of the entire data set.

To try and obtain more generalized information.

```
> summary(Cars.data)
```

	make	mpg	cylinders	displacement	horsepower	weight	acceleration	model.year	origin
ford.pinto	: 6	Min. : 9.00	Min. :3.000	Min. : 68.0	Min. : 46.00	Min. :1613	Min. : 8.00	Min. :70.00	1:254
amc.matador	: 5	1st Qu.:17.50	1st Qu.:4.000	1st Qu.:105.0	1st Qu.: 75.75	1st Qu.:2226	1st Qu.:13.70	1st Qu.:73.00	2: 73
ford.maverick	: 5	Median :23.00	Median :4.000	Median :151.0	Median : 95.00	Median :2822	Median :15.50	Median :76.00	3: 79
toyota.corolla	: 5	Mean :23.51	Mean :5.475	Mean :194.8	Mean :105.08	Mean :2979	Mean :15.52	Mean :75.92	
amc.gremlin	: 4	3rd Qu.:29.00	3rd Qu.:8.000	3rd Qu.:302.0	3rd Qu.:130.00	3rd Qu.:3618	3rd Qu.:17.18	3rd Qu.:79.00	
amc.hornet	: 4	Max. :46.60	Max. :8.000	Max. :455.0	Max. :230.00	Max. :5140	Max. :24.80	Max. :82.00	
(other)	:377	NA's :8			NA's :6				

For example the car that was in the dataset the most was Ford Pinto.