Liang,Xiyu 101086285 STAT5703 Assignment#1

Output of the code and interpretation.

1. Visualize data

Using R

By using summary and table command, we know that:

(1) We have 7 dimensions in the sample.

(2) Cars come from 3 regions (“USA”, “JAPAN”, “EUROPE”).

(3) The time period is from 1970 to 1982.

(4) The number of cylinders of cars’ engine is 3,4,5,6 and 8.

We can see these results clearly by using histograms.

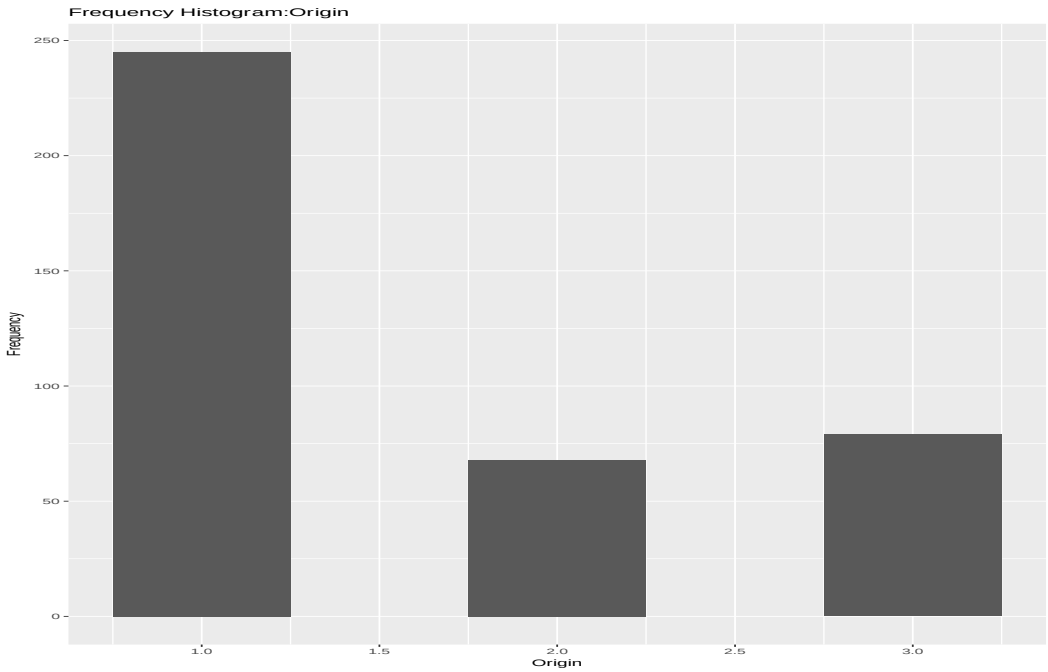

Figure 1.

There are three car manufactures in our sample, “1” represents “USA”, “2” represents “Japan” and “3” represents “Europe”.

The number of cars from USA, which is 245, is almost more than three times as many as cars either from Japan or Europe (68 and 79 respectively).

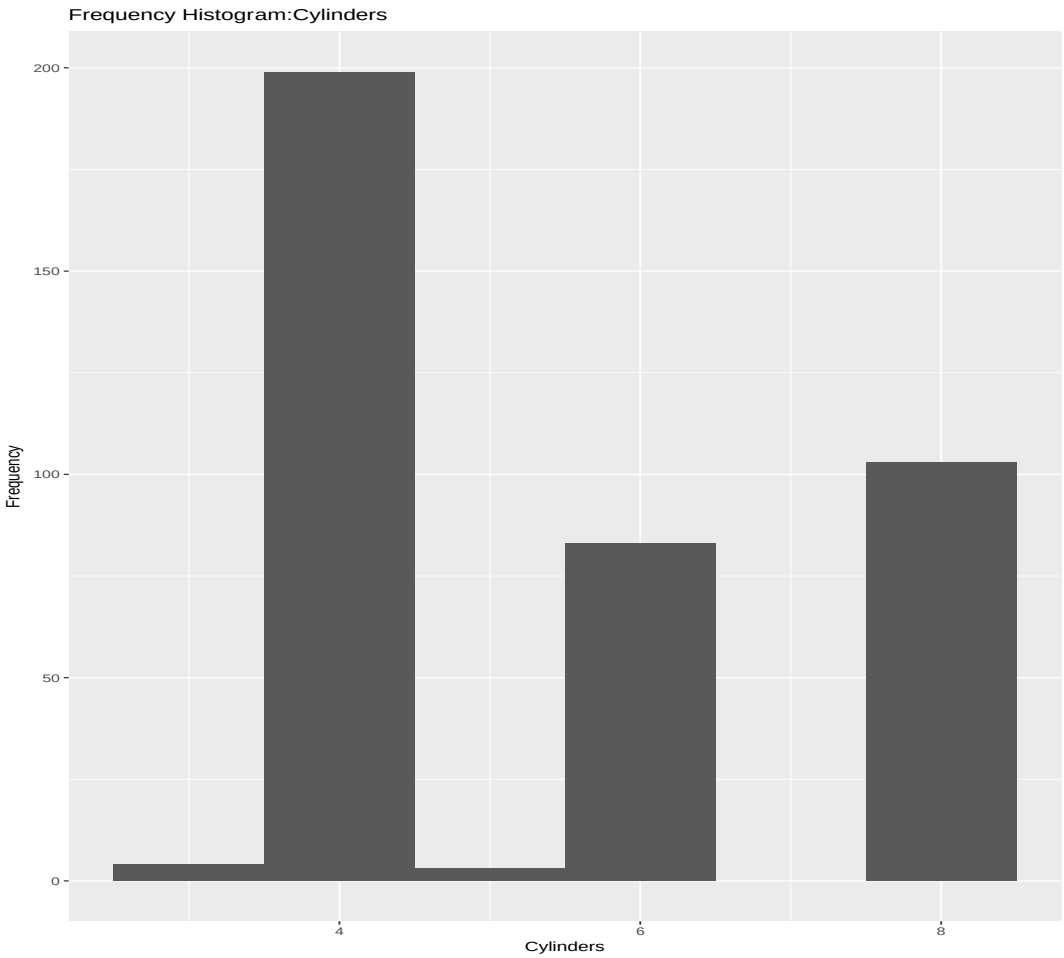

Figure 2.

In our sample data, compared with the number of four, six and eight-cylinder cars (199, 83, and 103 respectively), the counts of three and five-cylinder cars are relatively small, so we can remove them from our data in case they will be distractions then influence the pattern of later plots.

Scatterplot matrix plots the variable pairwise and can display variable relationships clearly, so we can use it to visualize our sample data.

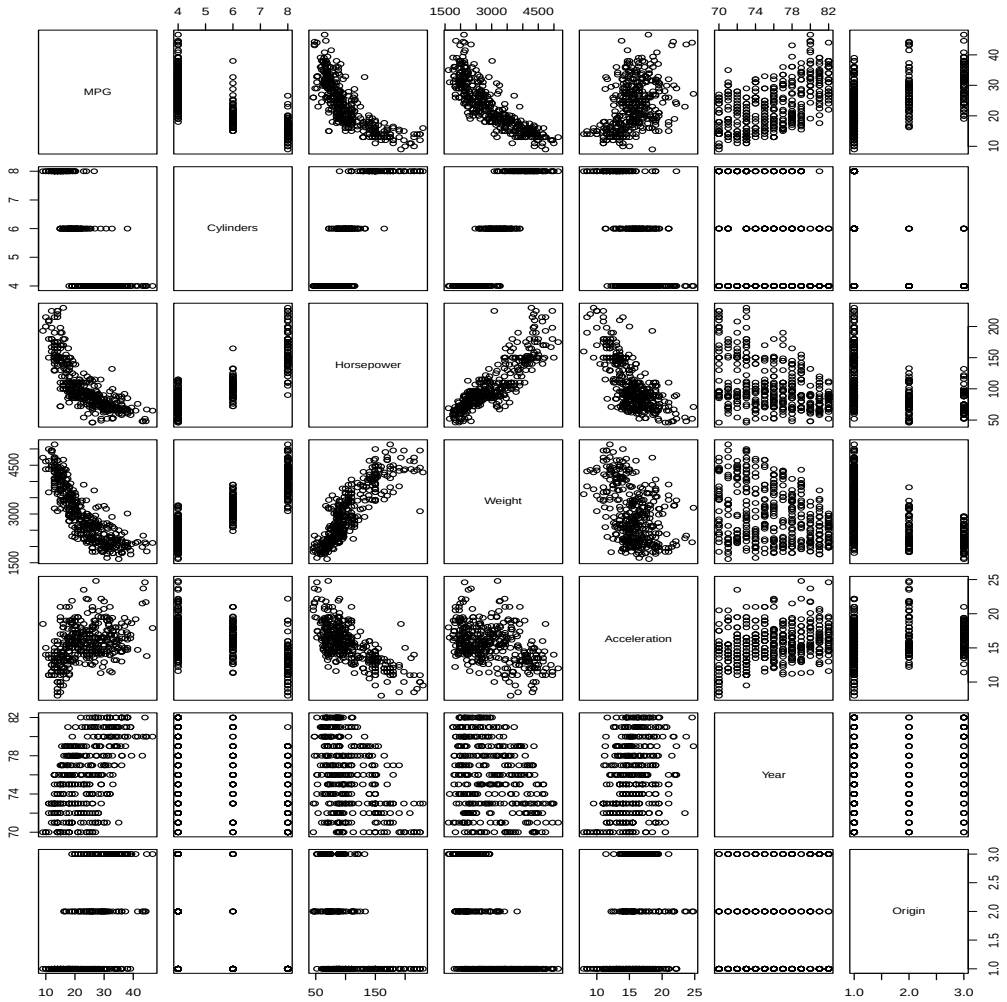

Figure 3.

By plotting the variables pairwise, we can get this standard pairs plot. It shows relationships between variables and the type of relationships.

We are able to see that there are high correlations between Horsepower and Weight, MPG and Horsepower, as well as MPG and Weight.

As the Weight increases, the Horsepower also increases, which means variables Horsepower and Weight have a generally positive relationship.

For the MPG and Horsepower pair of variables, when we read the graph from left to right, our points are decreasing so we could describe the trend as a decreasing trend. The relationship is in a negative way.

And by looking at the pairs plot of MPG and Weight, the weight of the vehicle is the explanatory variable and the MPG is the response variable, it shows that there is a negative association between these two variables, as the weight increases, the MPG decreases.

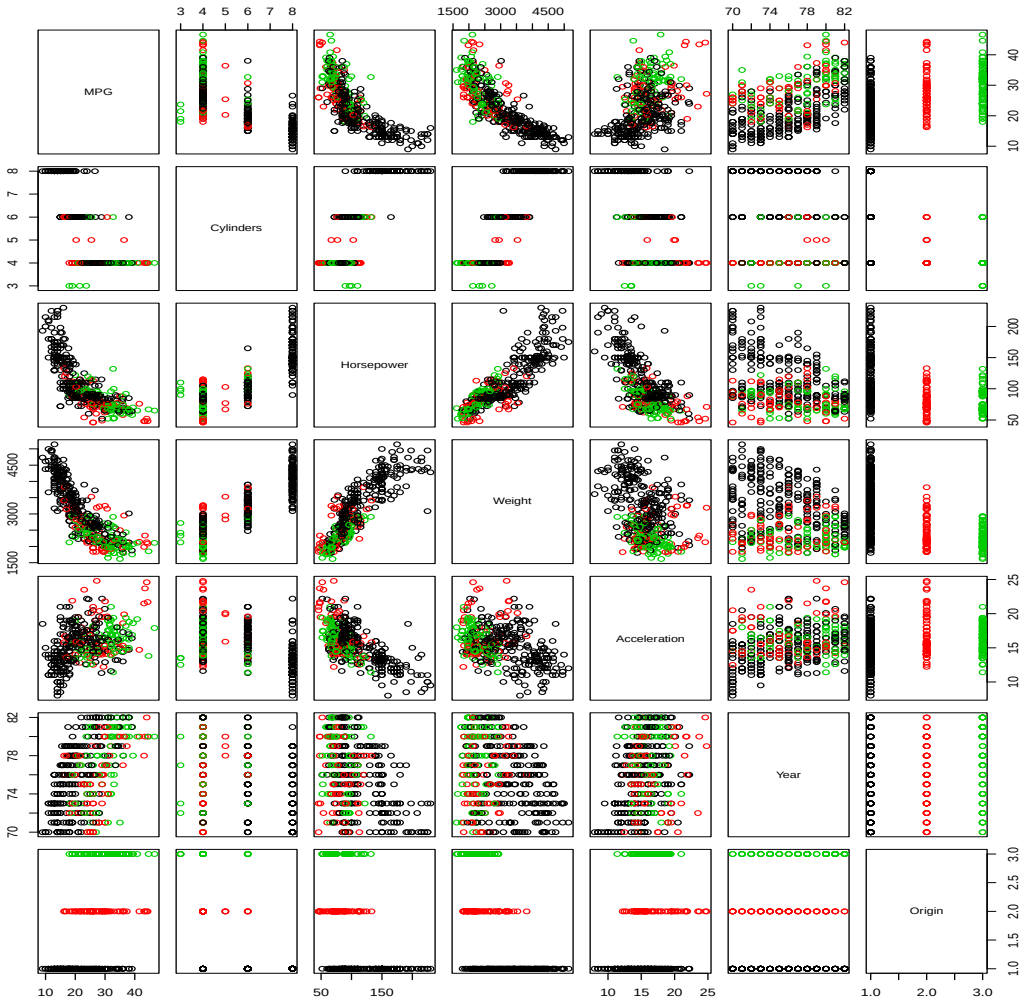

Figure 4.

Since we have 3 regions, we use color “green” to represent “Europe”, “red” for “Japan” and “black” for the “USA”.

Looking at the graph of MPG and Horsepower, and the graph of MPG and Weight (the first of the third row, the first of the fourth row, respectively), we can see that black points clustered in the

upper left corner, red and green points clustered in the lower right corner.

This shows that US cars have high horsepower but low MPG. And they weigh more than Japan and Europe cars.

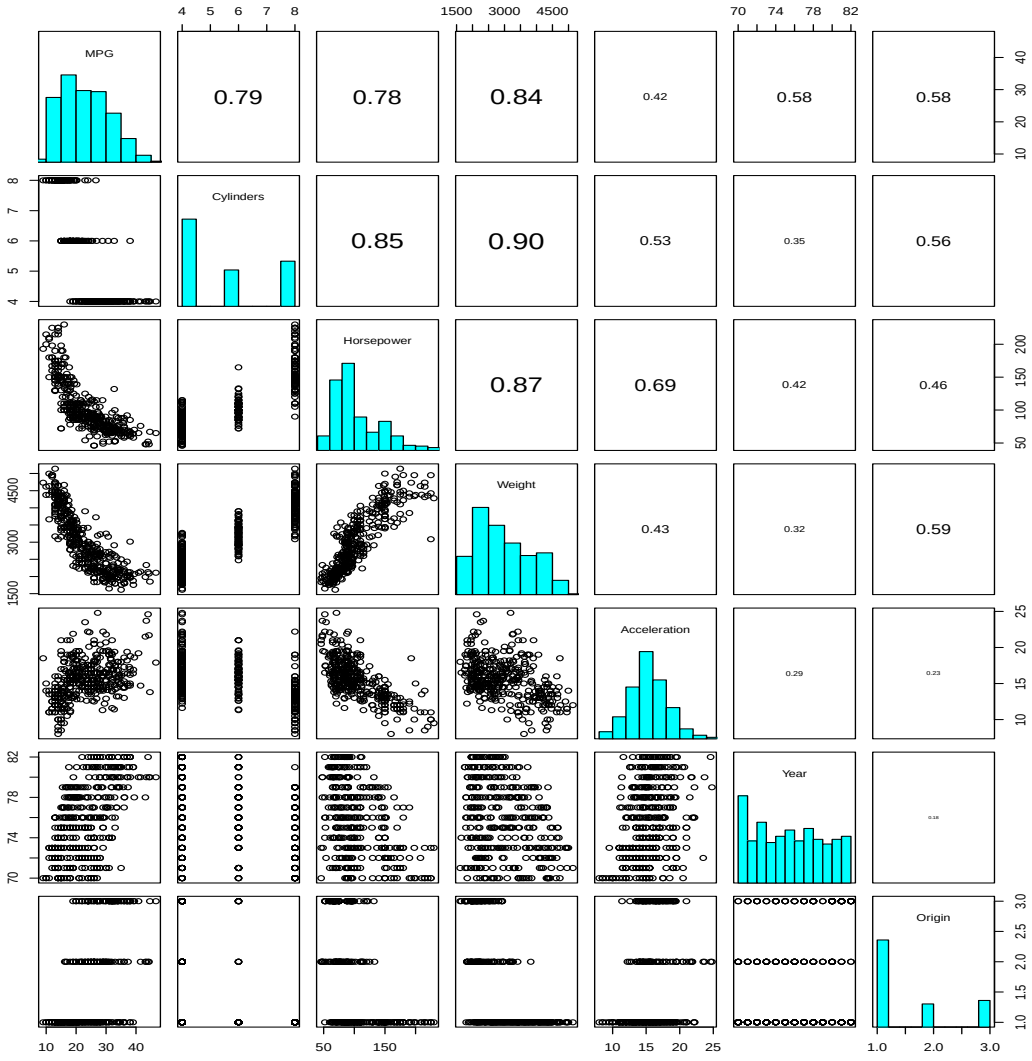

Figure 5.

The second pairs command shows the scatter plot of matrices above, with bivariate scatter plots below the diagonal, histograms on the diagonal, and the Pearson correlation above the diagonal.

By looking at histograms on the diagonal, we notice that MPG, Horsepower and Weight have similar right-skewed distribution, this indicates that there are relationships between these three variables. And there are some numbers in the upper triangle part, the larger the number is, the higher the correlation between variables.

As we can see, MPG and Cylinders, MPG and Horsepower, MPG and Weight; Cylinders and Horsepower, Cylinders and Weight; and Horsepower and Weight are highly correlated, which means the sample contains overlapped information and we can remove one or several of those related variables from our model.
Therefore, we can just use MPG, Cylinders and Weight variable to visualize how one of them will affect the others. Since cars in the sample come from three regions, we can also do comparisons by origin in order to find differences of cars from different origins.

We can also use Boxplot to discover relationships between variables.

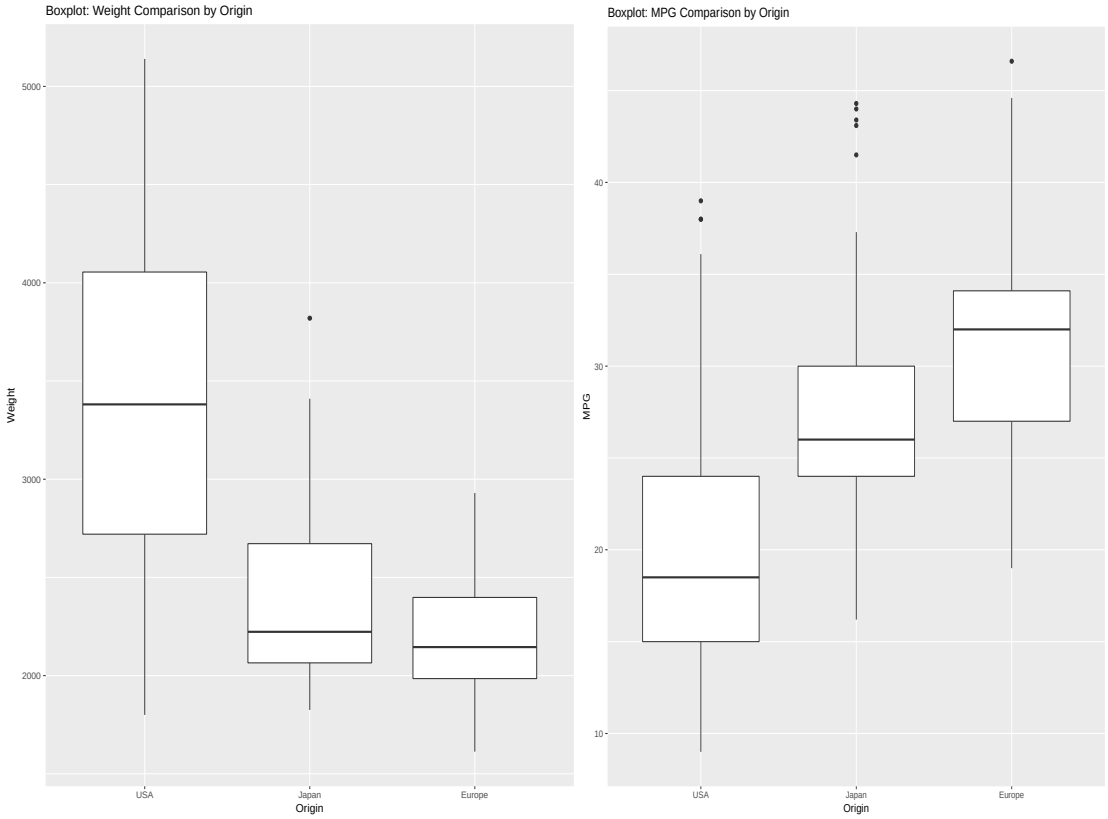

Figure 6.

American cars weigh more on average than Japan and Europe cars, and the average MPG of American cars is lower than that of either Japan or Europe. This verifies the result above that MPG and Weight have a negative relationship.

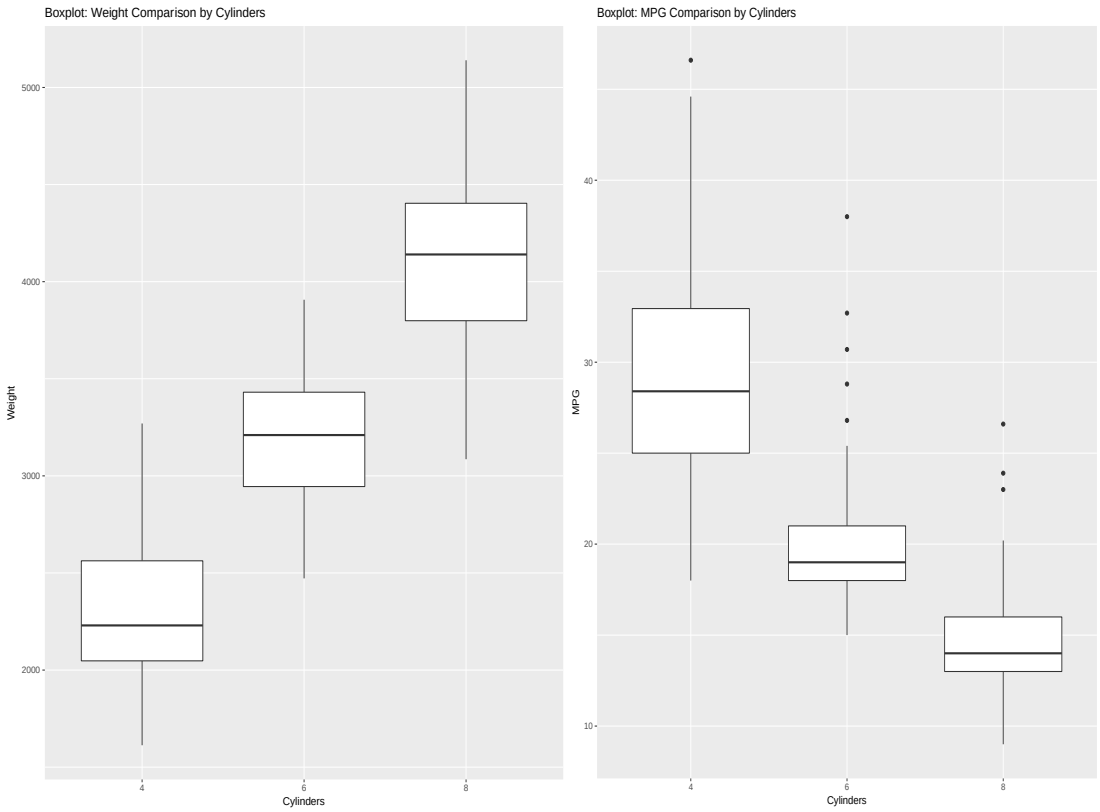

Figure 7.

The average MPG values for four-cylinder cars are higher than six-cylinder and eight-cylinder cars. So the number of cylinders has an effect on the MPG, the more cylinders a car has, the worse its fuel efficiency.

And by comparing the average Weight of cars with different cylinders, we know that in the sample, four-cylinder cars are lighter than six-cylinder and eight-cylinder cars.

Now we can compare the cars from each region by the number of cylinders.

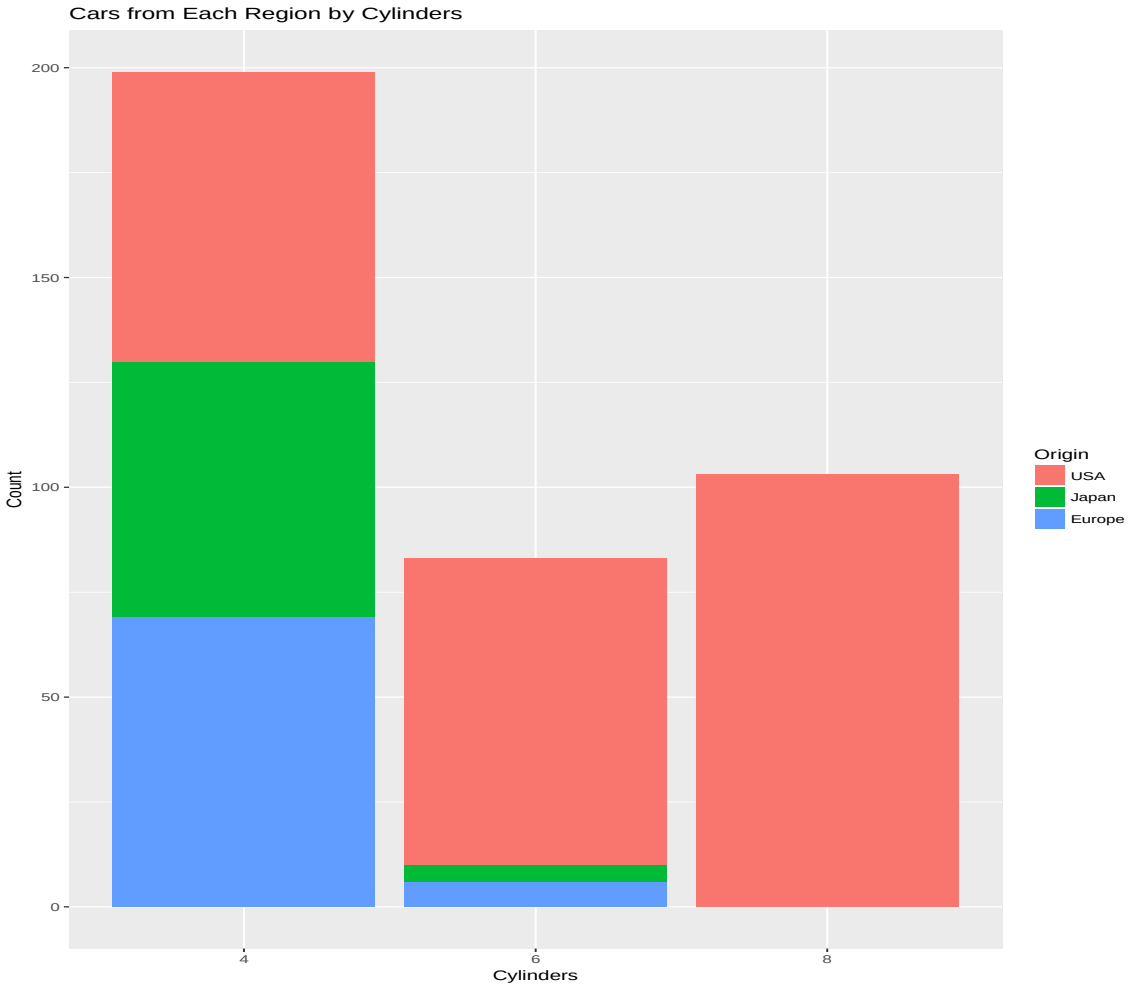

Figure 8.

All eight-cylinder cars and a majority of six-cylinder cars in the sample come from the US, six-cylinder cars seldom come from Japan or Europe. The number of four-cylinder cars is roughly equal across regions.

Then we are able to see how each region’s type of product changes over time.

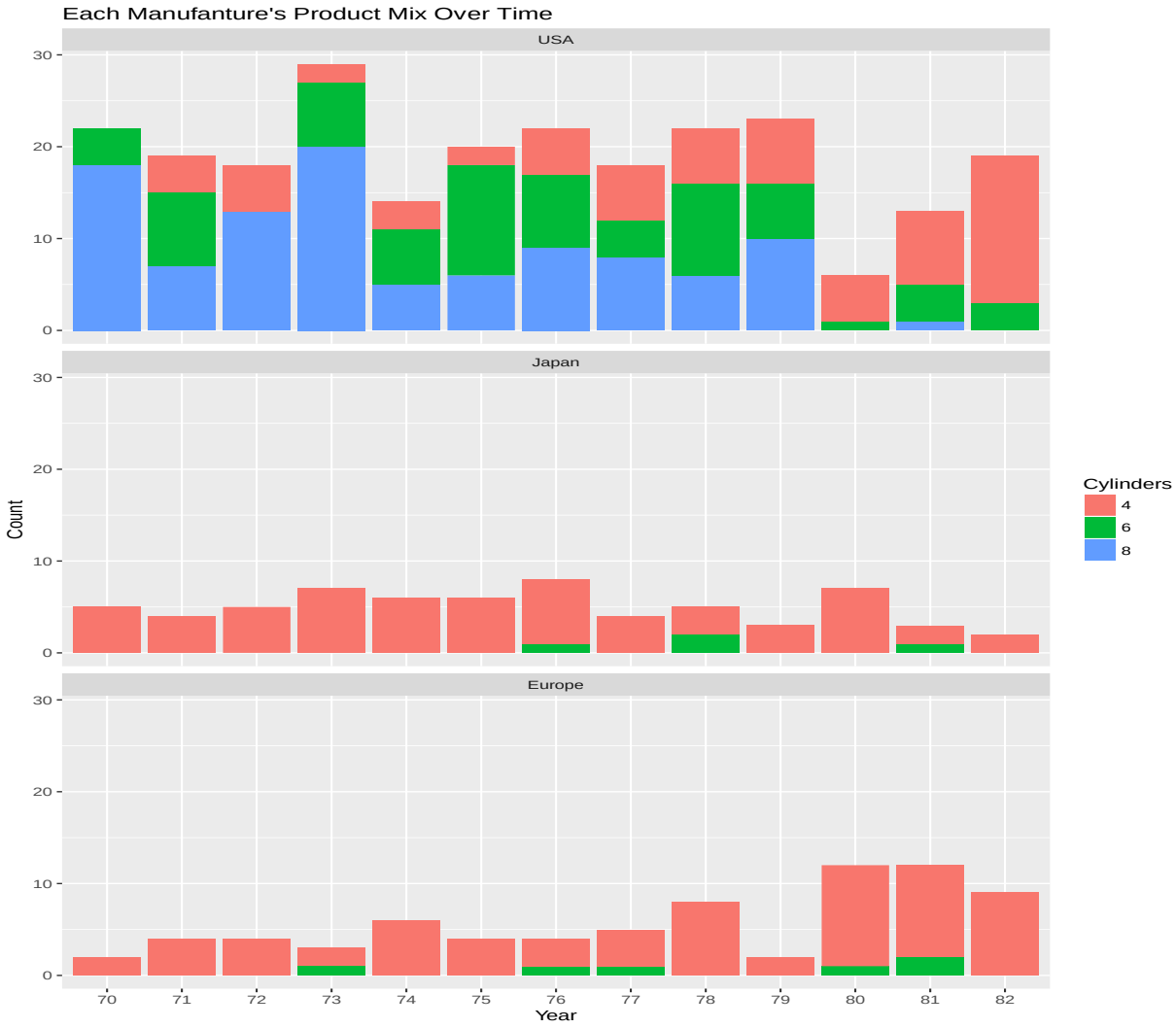

Figure 9.

In the USA section, although the number of four-cylinder cars increases over time, the majority of American cars’ engines are still six-cylinder or eight-cylinder until 1980.

While Japan and Europe almost only produce four-cylinder cars, and a small number of six-cylinder cars.

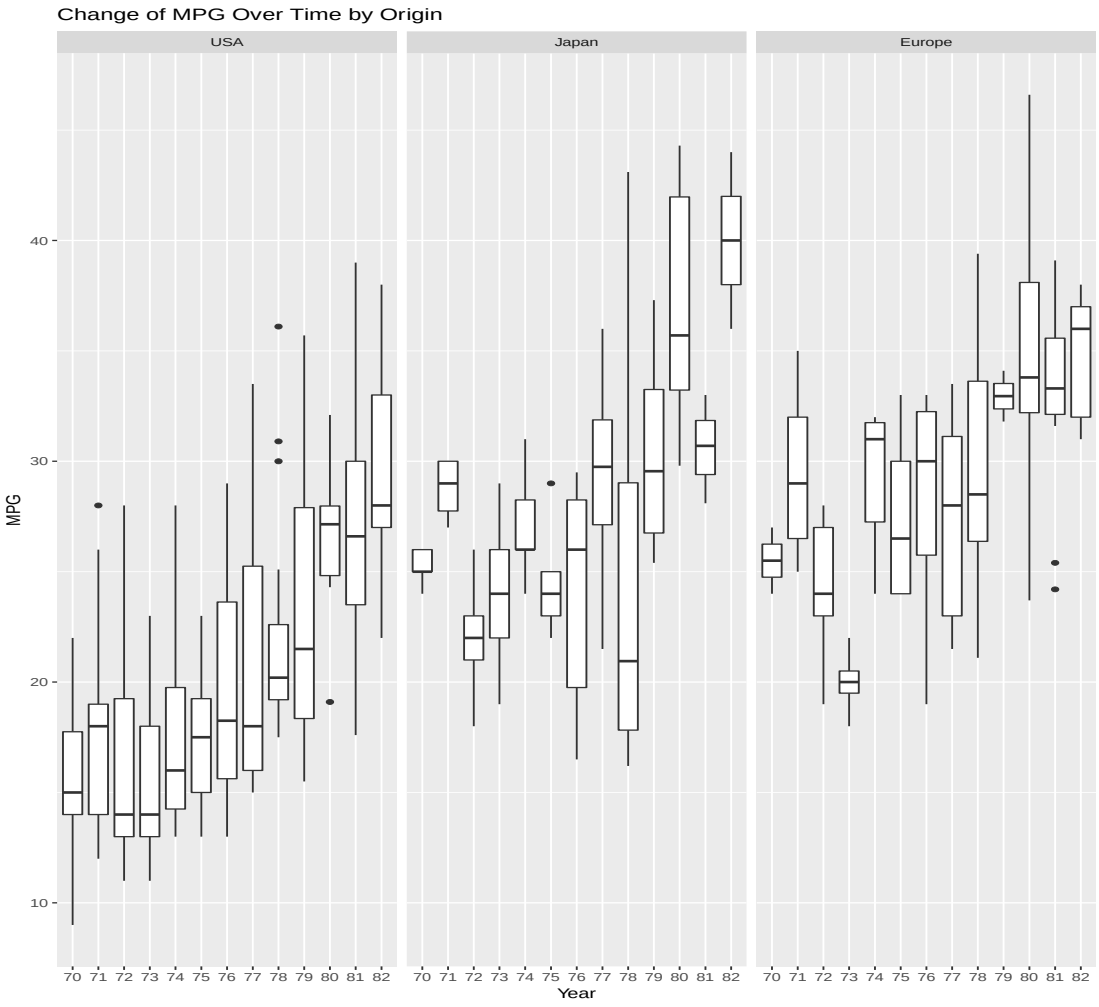

Figure 10.

We can see that the US, Japan and Europe tried to increase the average MPG of cars during the 13-year period.

US cars show much lower average MPG than Japan and Europe cars until 1980. From then, the fuel efficiency of US cars improved remarkably. And from Figure 9 we know that since 1980, US manufactures started shifting their product mix to include more four-cylinder cars.

Until the early 80’s, the average MPG of Japan and Europe cars was still much higher than that of US.

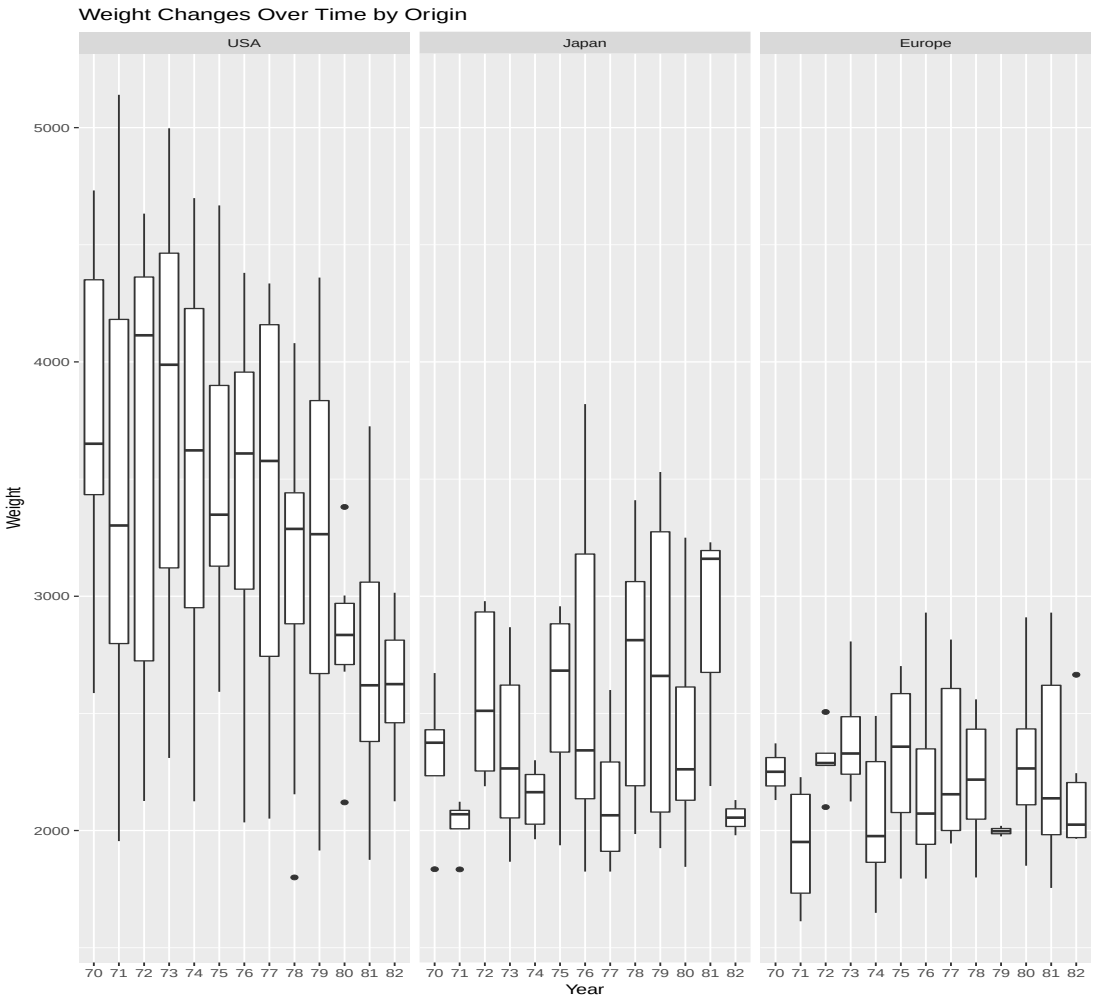

Figure 11.

We can notice that the average weight of US cars is way higher than Japan cars and Europe cars until 1980. This situation changed distinctly from the year of 1980.

The average weight of Europe cars shows the most constant trend.

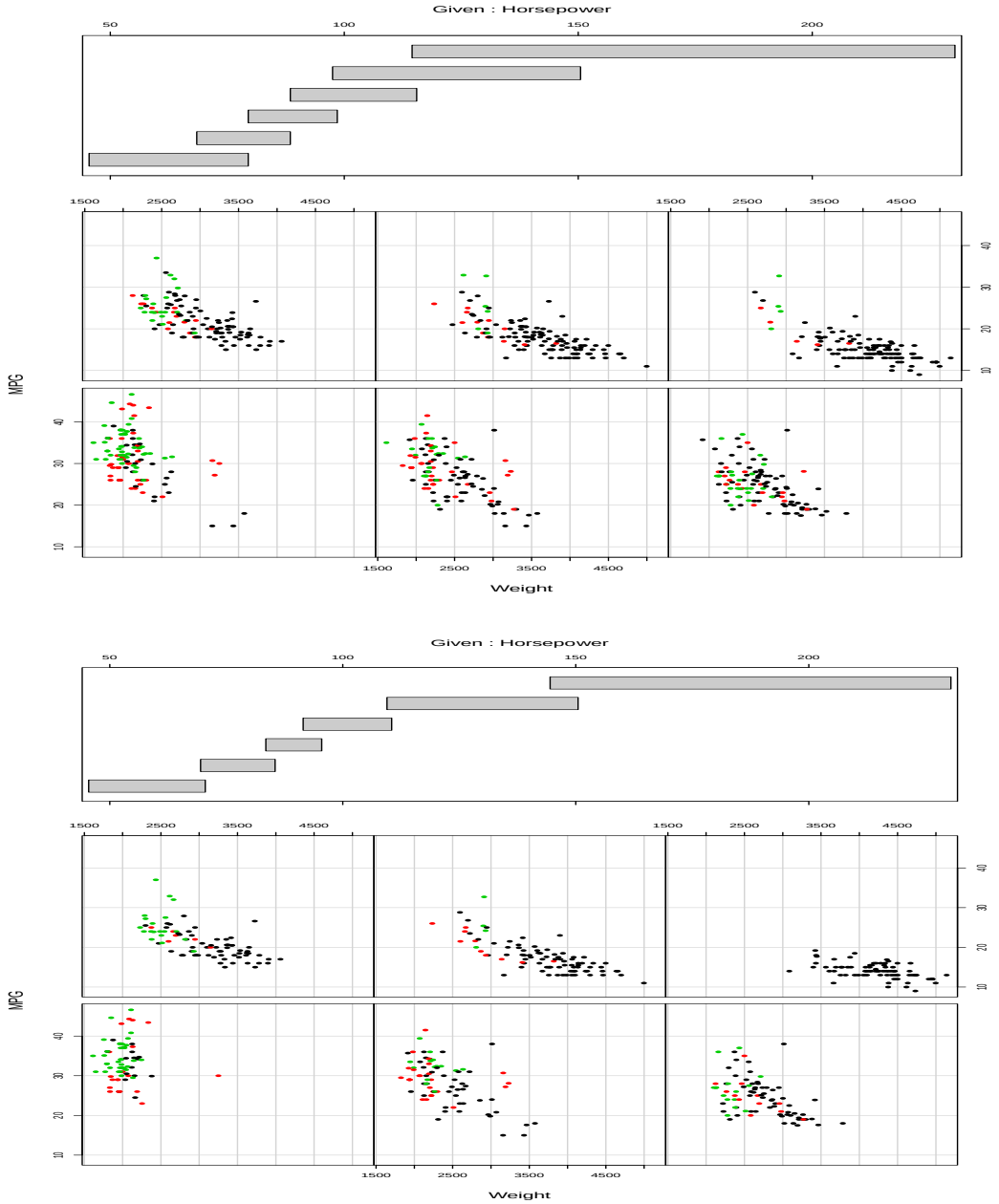

Figure 12.

The lower six panels show the pairwise plots for MPG against Weight for different ranges of Horsepower as shown in the upper panel. Each grouping has an equal number of cases, the overlap for the first figure is 0.5, the second one has reduced the overlap to 0.1. The greater the horsepower an engine of a car produces, the poorer miles per gallon, the heavier the car it is.

Within different ranges of horsepower, MPG is always in negative correlation to Weight.

Conclusions

1.The weight of a car will influence its MPG and Horsepower. A heavy car usually has worse fuel efficiency compared with a light car, while the horsepower of it may be better. Four-cylinder cars are the lightest, with the best gas mileage, and eight-cylinder cars are the heaviest.
If a customer wants to buy a car with low weight and good fuel efficiency, it is better to choose a four-cylinder car.

2.From 1970 to 1980, although the number of four-cylinder cars kept increasing, six and eight-cylinder cars still made up the large majority of the US product mix. Europe and Japan almost exclusively produced four-cylinder cars with just a few exceptions over the entire 13-year period.
So for customers who prefer six-cylinder and eight-cylinder cars, the US automakers are good choices.
In addition, this phenomenon can illustrate that why American cars weigh more, on average, and the fuel economy of American cars is not as good as Japan and Europe cars.
In 1975, US manufactures started producing more four-cylinder cars, in order to increase the average MPG, which is beneficial for them to compete with the other Europe and Japan automakers.

3.During the 13-year period, the US, Japan and Europe automakers were keep increasing the fuel economy of their products. The way that the US adopted is product transformation. With the increase in the number of four-cylinder cars, the average weight of US cars began to decline.

Using Ggobi

1. Scatterplot Matrix

Ggobi starts with simple scatterplot, since we have 7 dimensions so we can get 7 xy plots. Several plots are displayed below.

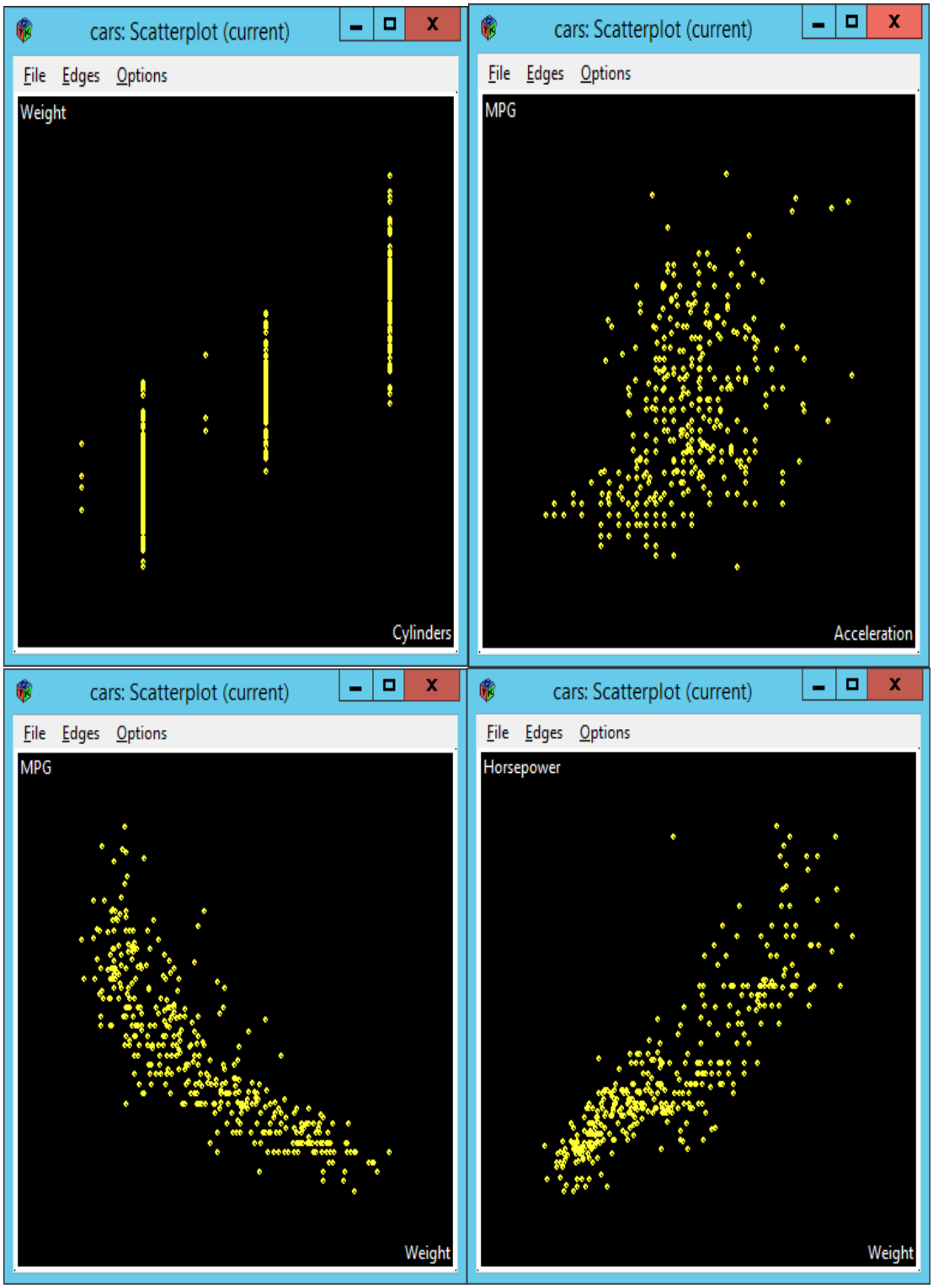

Figure 13.

Although we can also have a scatterplot matrix display.
Click [**Tools**] [**Automatic Brushing**], we select variable “Origin”, then use different color to represent different region of origin, “purple” represents “Europe”, “orange” for “Japan” and “yellow” for “USA”.

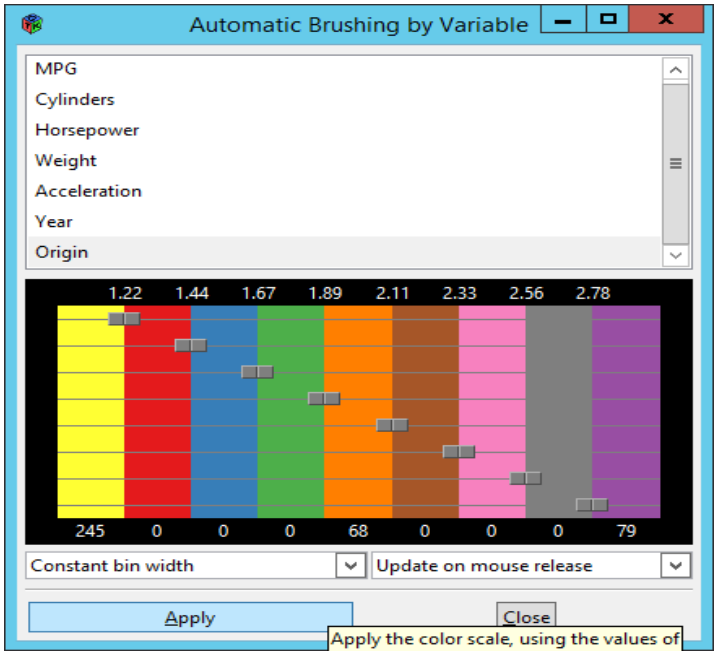

Figure 14.

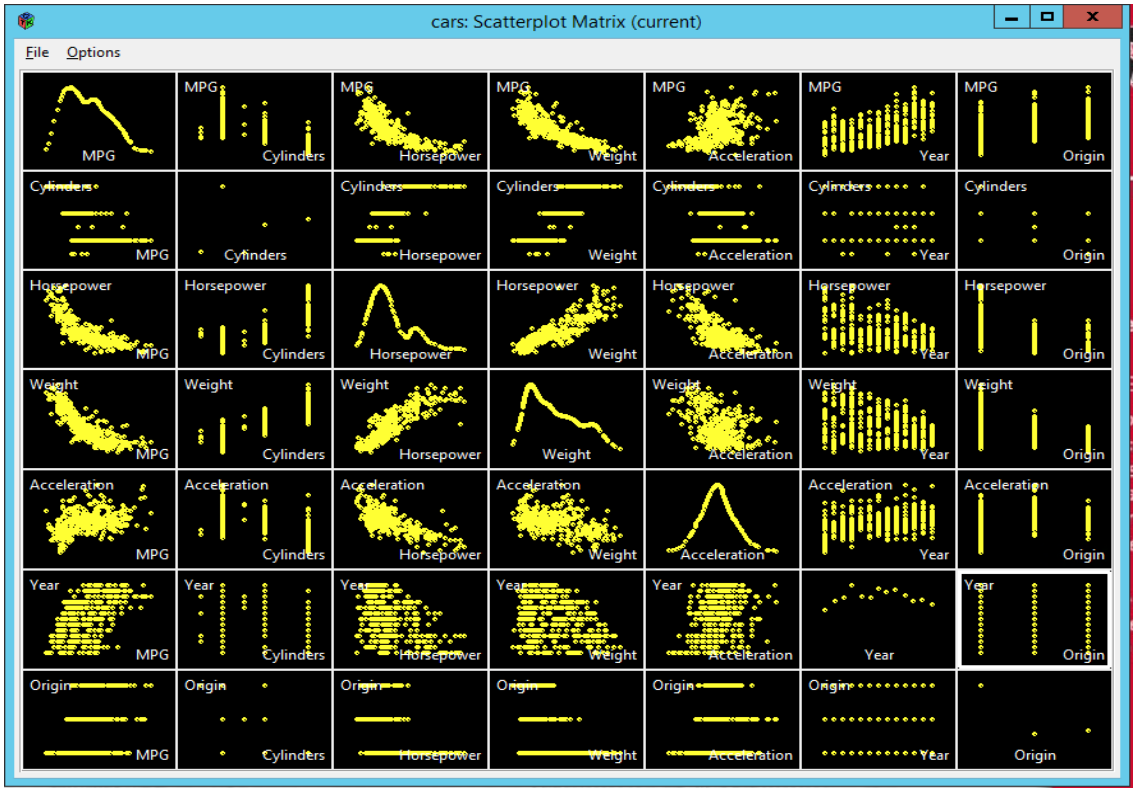

Figure 15.

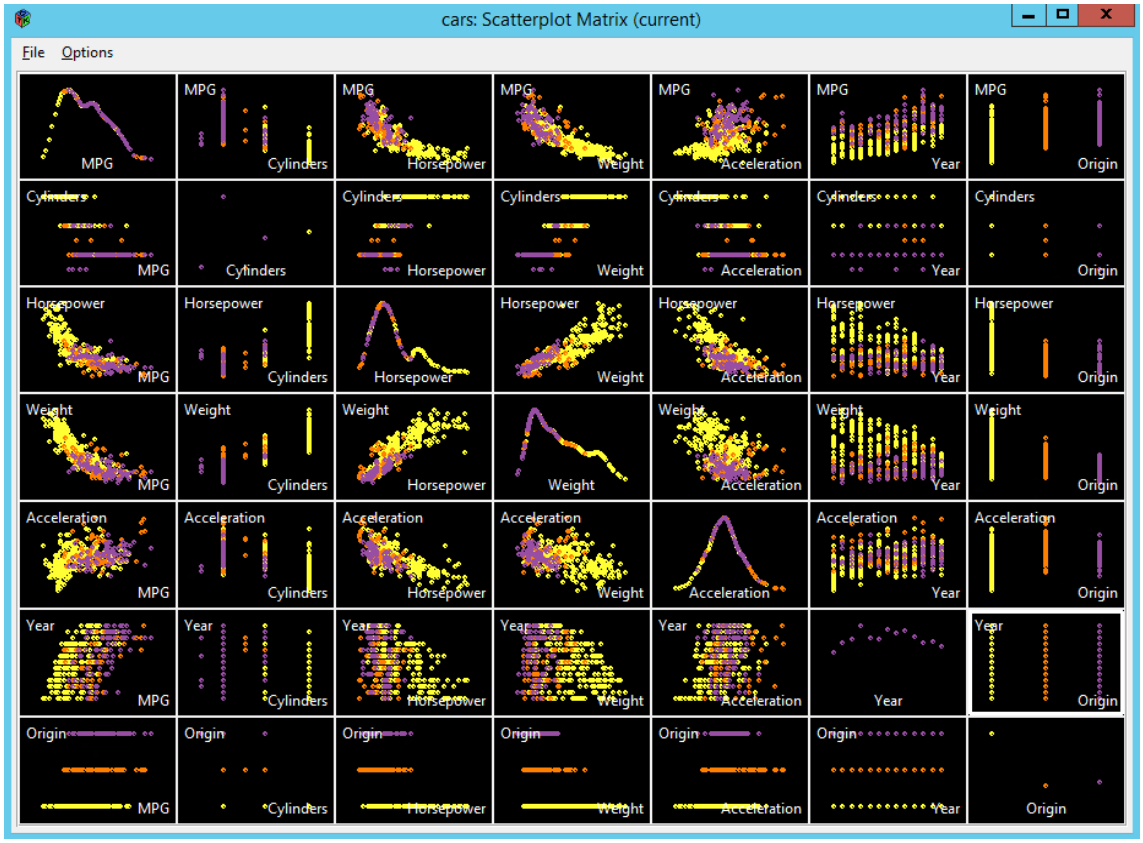

Figure 16.

The scatterplot matrix we get from Ggobi is almost the same as the scatterplot matrix we obtained by using R. Looking at the histograms on the diagonal, we have

(1) The distribution of MPG, Horsepower and Weight are all right-skewed, indicates that they may have associations.

(2) The distribution of Acceleration is mound-shaped, indicates that Acceleration's pairwise correlations with MPG, Weight, and Horsepower are less.

Looking at the lower triangle part of the scatterplot matrix, we know

(1) MPG and Weight, MPG and Horsepower are related negatively.

(2) The relationship between Weight and Horsepower is positive.

The information above we get from Ggobi is the same as the information we obtained from R.

2. 2D Tour

Different from R, Ggobi is able to help us to have a grand tour for our high dimensional data.

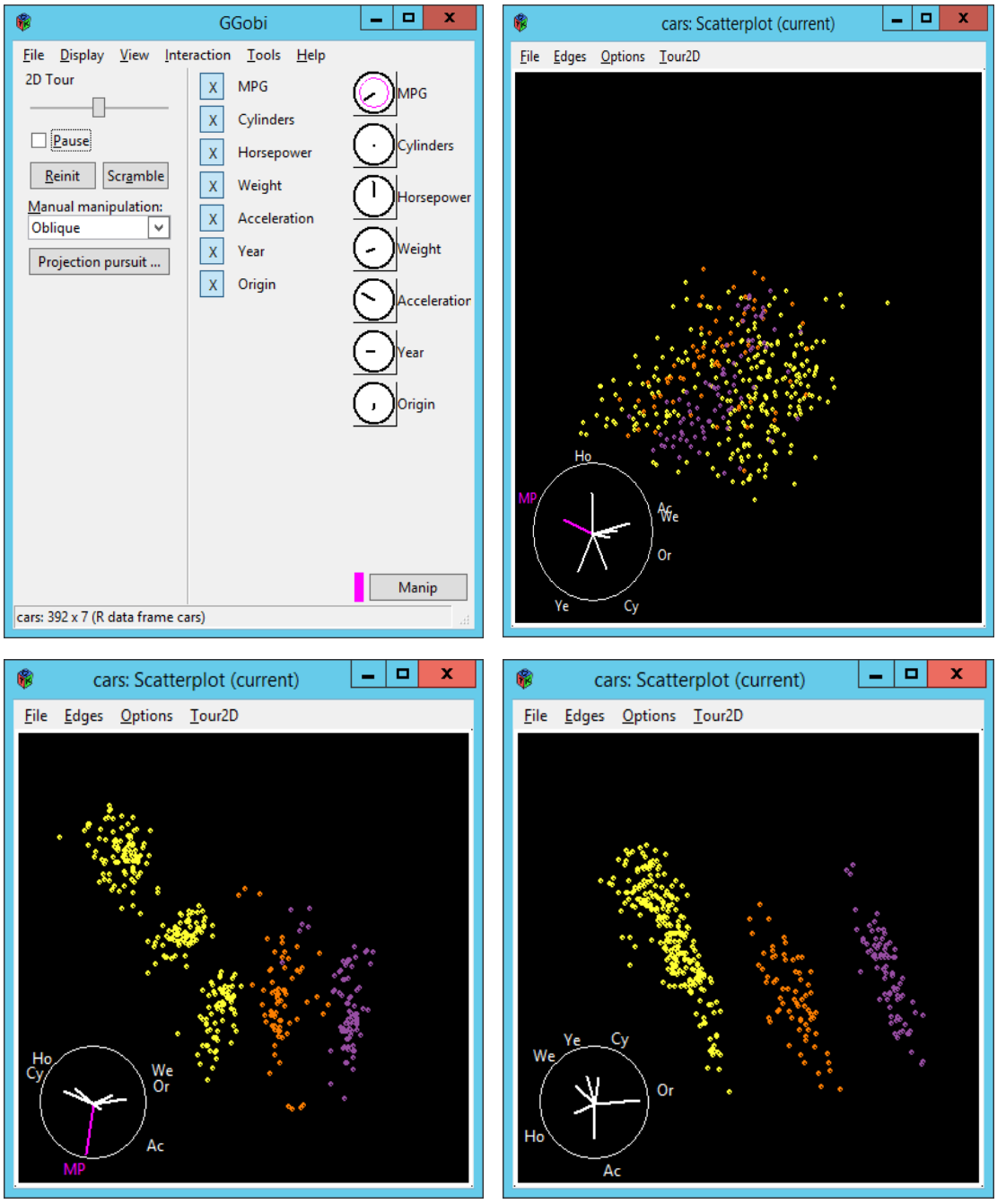

Figure 17.

Note the circle with the lines, it represents the projection of the seven axes on the two-dimensional display. We can see our data from all directions. Sometimes the points of the same color move together, but for much of the time, the projection of the clusters are mixed together.

3. Parallel Coordinates Plot

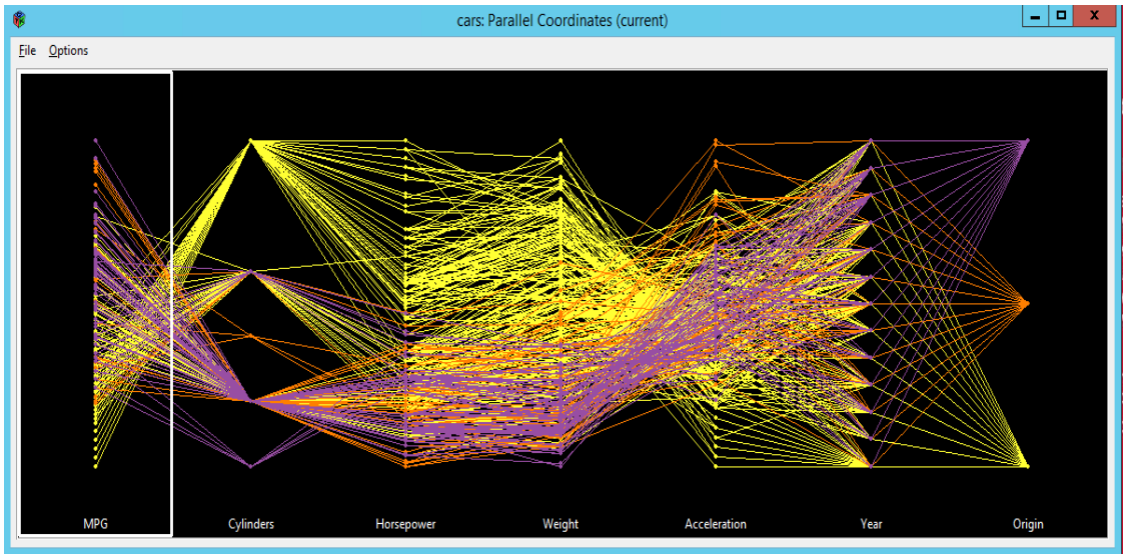

Figure 18.

In the Parallel Coordinates Plot, each data point has a value on each of the axes which are plotted vertically rather than at right angles to each other.

After moving the axes, we get Figure 19.

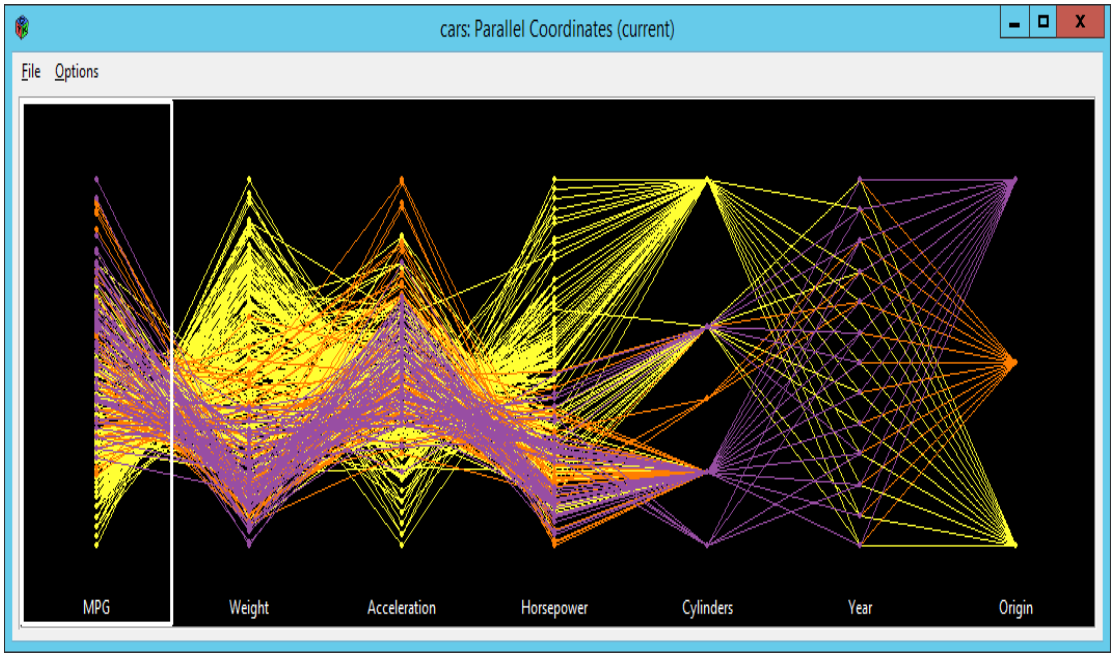

Figure 19.

Since “Cylinders” and “Origin” are discrete variables, and we can see that the variable “Year” cannot provide very useful information when we make the Parallel Coordinates Plot, so we remove them from our plot.

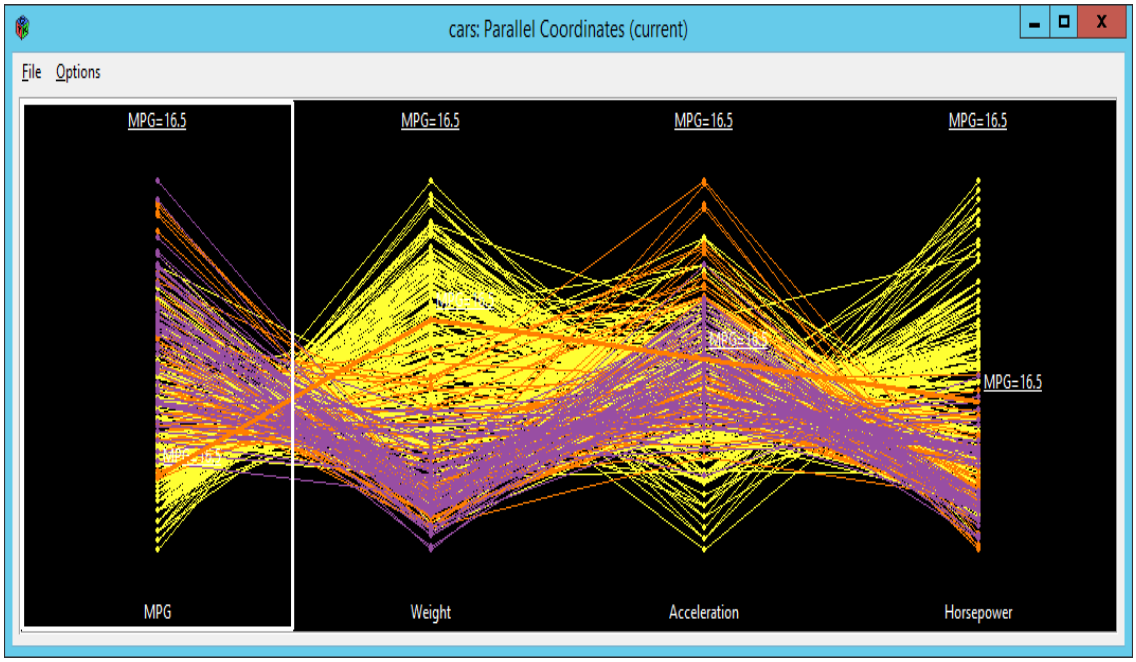

Figure 20.

When MPG<16.5, the yellow group is split from the other two groups, so the average MPG of US cars may be lower than Europe and Japan cars.

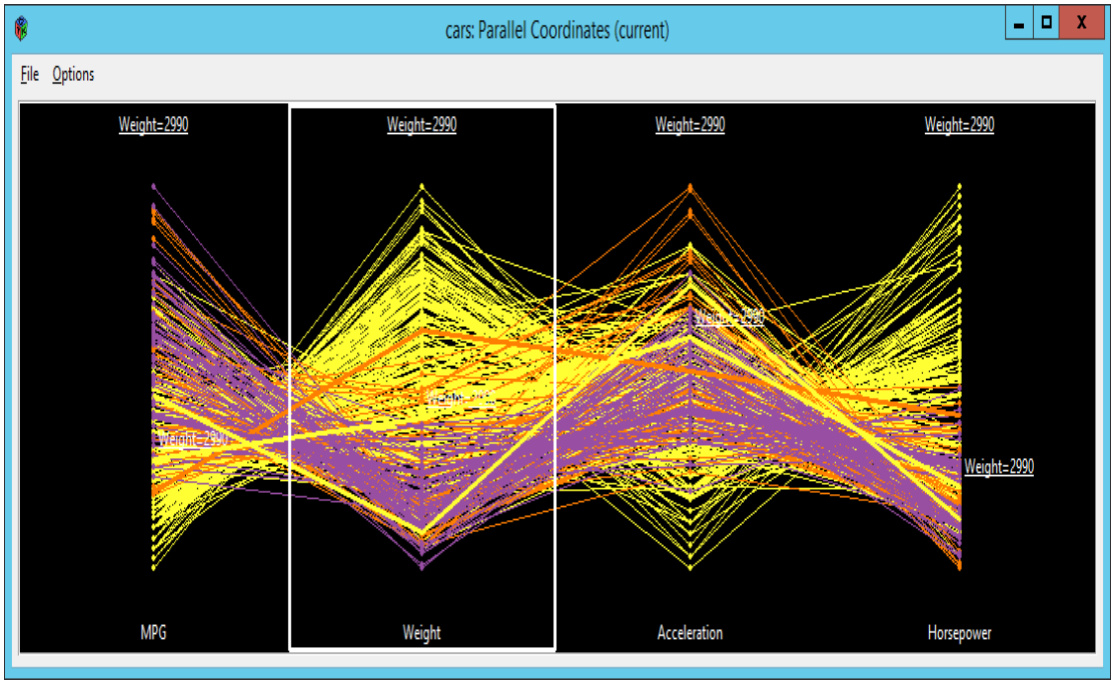

Figure 21.

In the weight section, yellow points clustered in the upper part while orange and purple points clustered below, so in average, US cars may weigh more.

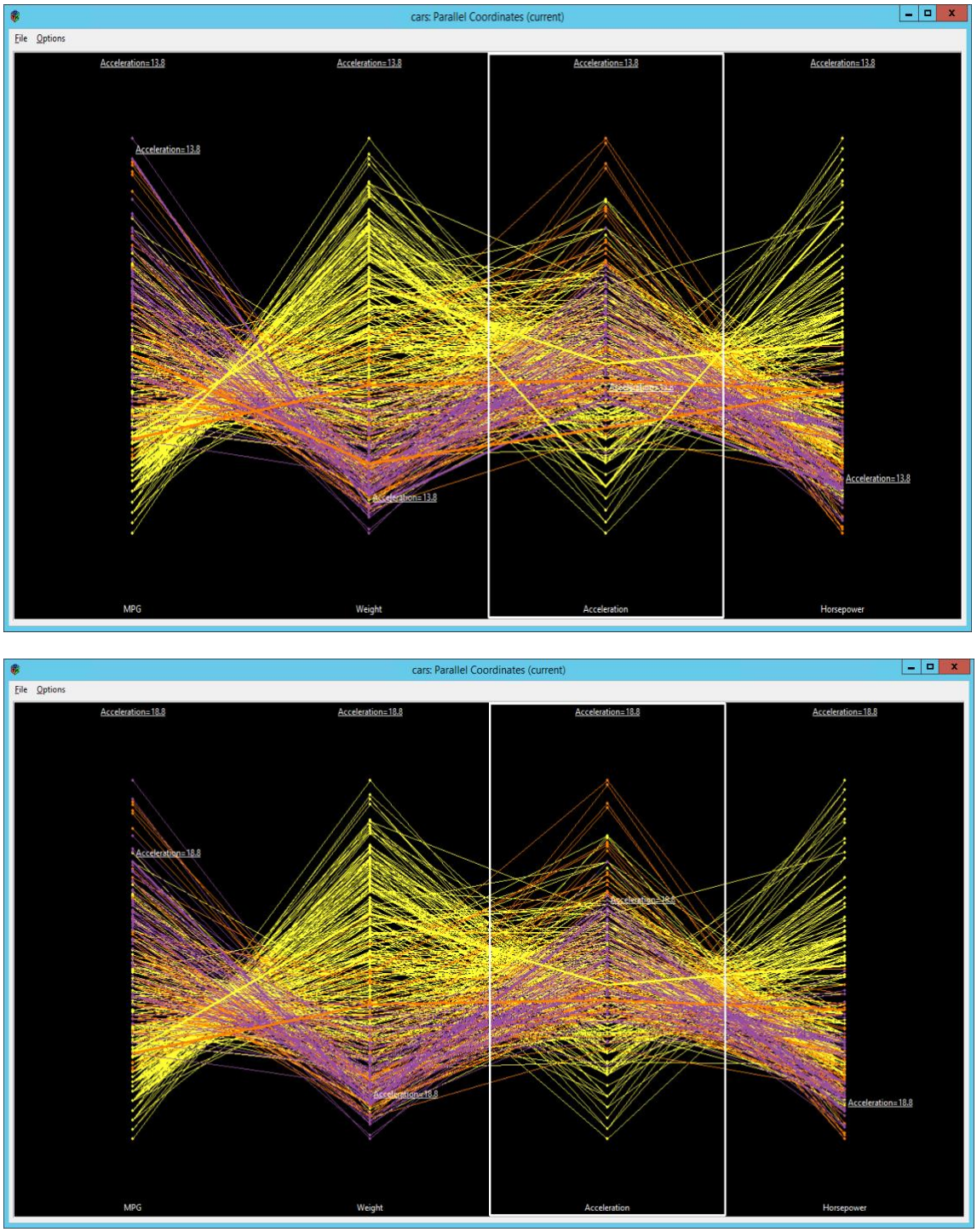

Figure 22.

In the above we see that if 13.8< Acceleration <18.8 we have one group (purple) clustered in this range, which indicates that in our sample, all the Acceleration values of Europe cars are in this area. Some US cars and only a few Japan cars' acceleration values are below 13.8.

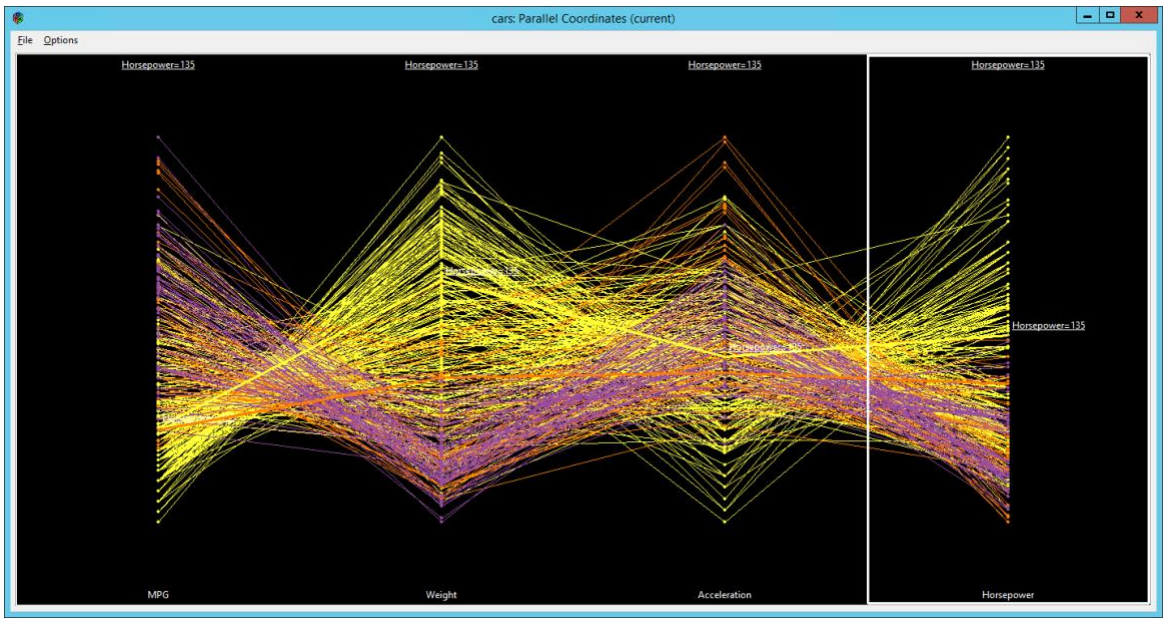

Figure 23.

When Horsepower> 135, only yellow points appear and both purple group and the orange group clustered in the lower part.
This indicates that in our sample, the average horsepower value of US cars may be larger than that of either Japan or Europe cars.

In summary, compared with Japan and Europe cars, US cars weigh more, with worse fuel efficiency but large horsepower.

2. Association Rule Mining

After transforming our data from the data frame format into transactions, we get a big matrix of items being bought together, ie, each row is a transaction, and the items in one row means they were bought together.

Then we can have a look about this transaction dataset.

```
> summary(itemlist)
transactions as itemMatrix in sparse format with
 19855 rows (elements/itemsets/transactions) and
 4034 columns (items) and a density of 0.006512985

most frequent items:
WHITE HANGING HEART T-LIGHT HOLDER           JUMBO BAG RED RETROSPOT
                              2260                              2092
          REGENCY CAKESTAND 3 TIER                     PARTY BUNTING
                              1989                              1686
           LUNCH BAG RED RETROSPOT                           (Other)
                              1564                            512067

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17
1532  830  686  671  688  613  589  600  621  534  558  502  481  540  544  548  463
  18   19   20   21   22   23   24   25   26   27   28   29   30   31   32   33   34
 454  462  429  433  341  337  304  253  262  262  215  270  239  191  191  161  172
  35   36   37   38   39   40   41   42   43   44   45   46   47   48   49   50   51
 144  134  131  127  127  125  113  117  100   95   97   96   88   85   94   73   71
  52   53   54   55   56   57   58   59   60   61   62   63   64   65   66   67   68
  62   77   76   68   58   53   51   58   43   47   42   36   36   37   41   41   39
  69   70   71   72   73   74   75   76   77   78   79   80   81   82   83   84   85
  30   29   37   20   31   42   22   23   17   25   14   12   24   20   20   19   16
  86   87   88   89   90   91   92   93   94   95   96   97   98   99  100  101  102
  13   15   17    6   14   10   16   14    7    9   10   10    7    6   13   14    3
 103  104  105  106  107  108  109  110  111  112  113  114  115  116  117  118  119
   7    6    5    1    4    5    6    4    3    6    6    7    3   10    3    5    8
 120  121  122  123  124  125  126  127  128  129  130  131  132  133  135  136  137
   5    7   11    6    3    2    6    5    5    2    3    3    1    5    5    7    4
 138  139  140  141  142  143  144  145  146  147  148  149  150  151  152  153  154
   2    4    7    6    2    1    2    5    6    4    6    2    2    5    3    4    2
 156  157  158  159  161  162  163  164  165  166  167  168  169  170  171  172  173
   6    4    3    4    2    3    5    3    1    5    1    4    3    5    3    1    3
 174  175  176  177  178  179  180  181  182  183  184  185  186  187  188  189  190
   3    3    3    3    5    5    5    3    2    2    7    2    2    4    4    3    1
 191  192  193  194  195  196  197  198  199  200  202  203  204  205  206  207  208
   1    3    6    2    3    2    4    2    5    1    1    1    4    6    4    1    1
```

We have 19,855 rows in total, which is the number of transactions as well. And there are 4,034 columns (items).

We can use the density to calculate how many items were purchased like so: 19855*4034*0.0065=520,617

In this summary, we can also find the most frequent items, which is "WHITE HANGING HEART T-LIGHT HOLDER".

By looking at the transaction length distribution and the transaction sizes, we can see that 1532 transactions contain only 1 item, 830 transactions for 2 and all the way up to the biggest transaction: just 1 for 1112 items, which means that in each transaction, most customers prefer to buy a small number of items.

```
 439  440  445  451  457  459  462  467  468  474  483  489  490  491  500  503  507
   2    2    1    1    1    1    1    1    1    1    1    1    1    1    1    2    1
 511  516  519  521  522  526  527  529  530  537  539  541  542  543  557  562  570
   1    2    1    1    1    1    2    2    1    1    1    1    1    1    1    1    1
 572  579  585  590  595  598  599  613  624  632  634  644  648  649  662  674  677
   1    1    1    1    3    2    1    1    1    1    1    1    1    1    1    1    1
 688  707  719  735  750 1112
   1    1    1    1    1    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    6.00   15.00   26.27   29.00 1112.00

includes extended item information - examples:
                     labels
1                  1 HANGER
2     10 COLOUR SPACEBOY PEN
3 12 COLOURED PARTY BALLOONS
```

Looking at the first quartile, third quartile, mean and median of the data, we can notice that the distribution of our data is right-skewed.

Now we use “Arules” package in R to do the association rule mining. Support is an indication of how frequently the itemset appears in the dataset, confidence is an indication of how often the rule has been found to be true. The lower support and confidence level we pick, the more rules we will get.
Usually we want both support and lift thresholds to be high, but in our case, there are a large number of transactions and a lot of different items in our dataset.
Based on the sample dataset we have, first, I set the support and confidence thresholds to 0.02 and 0.8, respectively. The minimum support threshold is 0.02, which means itemsets contain in our rules occur in at least 2% of our transactions. I think this level is acceptable for our data.

```
> summary(rules)
set of 14 rules

rule length distribution (lhs + rhs):sizes
2 3
7 7

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     2.0     2.5     2.5     3.0     3.0

summary of quality measures:
    support         confidence          lift            count
 Min.   :0.02055   Min.   :0.8016   Min.   : 7.611   Min.   :408.0
 1st Qu.:0.02166   1st Qu.:0.8431   1st Qu.:16.549   1st Qu.:430.0
 Median :0.02166   Median :0.9524   Median :38.931   Median :430.0
 Mean   :0.02305   Mean   :0.9198   Mean   :32.265   Mean   :457.7
 3rd Qu.:0.02166   3rd Qu.:1.0000   3rd Qu.:44.364   3rd Qu.:430.0
 Max.   :0.03188   Max.   :1.0000   Max.   :46.174   Max.   :633.0

mining info:
     data ntransactions support confidence
 itemlist         19855    0.02        0.8
```

The summary of rules contains useful information:

(1) We have 14 rules in total.

(2) The summary of quartile measures includes the range of support, confidence and lift.

(3) According to the mining information, we can see the number of transactions as well as the minimum support and confidence level we set earlier.

```
> inspect(rules[1:10])
    lhs                                rhs                                support    confidence lift     count
[1] {SET 3 RETROSPOT TEA}           => {SUGAR}                            0.02165701 1.0000000  46.17442 430
[2] {SUGAR}                         => {SET 3 RETROSPOT TEA}              0.02165701 1.0000000  46.17442 430
[3] {SET 3 RETROSPOT TEA}           => {COFFEE}                           0.02165701 1.0000000  38.93137 430
[4] {COFFEE}                        => {SET 3 RETROSPOT TEA}              0.02165701 0.8431373  38.93137 430
[5] {SUGAR}                         => {COFFEE}                           0.02165701 1.0000000  38.93137 430
[6] {COFFEE}                        => {SUGAR}                            0.02165701 0.8431373  38.93137 430
[7] {PINK REGENCY TEACUP AND SAUCER} => {GREEN REGENCY TEACUP AND SAUCER} 0.03188114 0.8263708  16.16511 633
[8] {SET 3 RETROSPOT TEA,SUGAR}     => {COFFEE}                           0.02165701 1.0000000  38.93137 430
[9] {COFFEE,SET 3 RETROSPOT TEA}    => {SUGAR}                            0.02165701 1.0000000  46.17442 430
[10] {COFFEE,SUGAR}                 => {SET 3 RETROSPOT TEA}              0.02165701 1.0000000  46.17442 430
```

Top 10 rules suggest that there is a strong relationship exists between the sale of “SET 3 RETROSPOT TEA”, “SUGAR” and “COFFEE”:

(1) 100% of the times a customer bought “SET 3 RETROSPOT TEA”, “SUGAR” is bought as well. 100% customers who bought “SUGAR” also bought “SET 3 RETROSPOT TEA”.

(2) 100% customers who bought “SET 3 RETROSPOT TEA” also bought “COFFEE”, 84.3% customers who bought “COFFEE” also bought “SET 3 RETROSPOT TEA”.

(3) 100% customers who bought “SUGAR” also bought “COFFEE”, while not all customers who bought “COFFEE” also bought “SUGAR”.

(4) 82.63% customers who bought “PINK REGENCY TEACUP AND SUGAR” also bought “GREEN REGENCY TEACUP AND SUGAR”, we can also say that for 82.63% of the transactions containing “PINK REGENCY TEACUP AND SUGAR” the rule is correct.
(5) 100% customers who bought any two of these items in the itemset {SET 3 RETROSPOT TEA, SUGAR, COFFEE} also bought the another one.

Lift explains the strength of association between the items on the left and right-hand side of the rule; the larger the lift the greater the link between the two items. For example, the lift for the first rule

{SET 3 RETROSPOT TEA} ---------- {SUGAR}

is 46.17442, suggests that the purchase of “SET 3 RETROSPOT TEA” increases the probability that “SUGAR” will also occur in that transaction.
As we can see, all lift values of these 10 rules are larger than 1, makes these rules potentially useful for predicting the consequent in future data sets.
Also indicates that the occurrence of the rule body (right-hand-side of the rule) has a positive effect on the occurrence of the rule head (left-hand-side of the rule).

Then I tried to lift the minimum support level to 0.03, in order to find 10 rules, the minimum confidence threshold has to go down to 0.5.

```
> summary(rules)
set of 13 rules

rule length distribution (lhs + rhs):sizes
 2
13

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2

summary of quality measures:
    support         confidence          lift            count
 Min.   :0.03017   Min.   :0.5035   Min.   : 5.493   Min.   :599.0
 1st Qu.:0.03188   1st Qu.:0.5787   1st Qu.: 6.429   1st Qu.:633.0
 Median :0.03223   Median :0.6236   Median :12.337   Median :640.0
 Mean   :0.03392   Mean   :0.6508   Mean   :11.192   Mean   :673.5
 3rd Qu.:0.03646   3rd Qu.:0.7205   3rd Qu.:14.565   3rd Qu.:724.0
 Max.   :0.04155   Max.   :0.8264   Max.   :16.165   Max.   :825.0

mining info:
     data ntransactions support confidence
 itemlist          19855    0.03        0.5
```

As we can see, we get 13 rules.

```
> inspect(rules[1:10])
    lhs                                 rhs                                 support    confidence lift
[1]  {PINK REGENCY TEACUP AND SAUCER}    => {ROSES REGENCY TEACUP AND SAUCER} 0.03016872 0.7819843  14.565008
[2]  {ROSES REGENCY TEACUP AND SAUCER}   => {PINK REGENCY TEACUP AND SAUCER}  0.03016872 0.5619137  14.565008
[3]  {PINK REGENCY TEACUP AND SAUCER}    => {GREEN REGENCY TEACUP AND SAUCER} 0.03188114 0.8263708  16.165115
[4]  {GREEN REGENCY TEACUP AND SAUCER}   => {PINK REGENCY TEACUP AND SAUCER}  0.03188114 0.6236453  16.165115
[5]  {ALARM CLOCK BAKELIKE RED}          => {ALARM CLOCK BAKELIKE GREEN}      0.03223369 0.6089439  12.337327
[6]  {ALARM CLOCK BAKELIKE GREEN}        => {ALARM CLOCK BAKELIKE RED}        0.03223369 0.6530612  12.337327
[7]  {LUNCH BAG PINK POLKADOT}           => {LUNCH BAG RED RETROSPOT}         0.03052128 0.5559633   7.057961
[8]  {JUMBO BAG PINK POLKADOT}           => {JUMBO BAG RED RETROSPOT}         0.04155125 0.6773399   6.428577
[9]  {JUMBO SHOPPER VINTAGE RED PAISLEY} => {JUMBO BAG RED RETROSPOT}         0.03424830 0.5787234   5.492616
[10] {LUNCH BAG  BLACK SKULL.}           => {LUNCH BAG RED RETROSPOT}         0.03228406 0.5035350   6.392383
    count
[1]  599
[2]  599
[3]  633
[4]  633
[5]  640
[6]  640
[7]  606
[8]  825
[9]  680
[10] 641
```

These items occurred in transactions more frequently.
(1)78% customers who bought “PINK REGENCY TEACUP AND SUGAR” also bought “ROSES REGENCY TEACUP AND SAUCER”, and 56% customers who bought “ROSES REGENCY TEACUP AND SAUCER” also bought “PINK REGENCY TEACUP AND SUGAR”.
(2) 82% of the times a customer bought “PINK REGENCY TEACUP AND SUGAR”, “GREEN REGENCY TEACUP AND SUGAR” is bought as well. 62% customers who bought “GREEN REGENCY TEACUP AND SUGAR” also bought “PINK REGENCY TEACUP AND SUGAR”.
(3)60% customers who bought “ALARM CLOCK BAKELIKE RED” also bought “ALARM CLOCK BAKELIKE GREEN”, 65% customers who bought “ALARM CLOCK BAKELIKE GREEN” also bought “ALARM CLOCK BAKELIKE RED”.
(4) We have almost 50% confidence to say that customers who bought “LUNCH BAG PINK POLKADOT” or “LUNCH BAG BLACK SKULL” would also buy “LUNCH BAG RED RETROSPOT”.
(5)67% customers who bought “JUMBO BAG PINK POLKADOT” and 57% customers who bought “JUMBO SHOPPER VINTAGE RED PAISLEY” also bought the “JUMBO BAG RED POLKADOT”.
And all the corresponding lift values are larger than 1, makes these rules we found are potentially useful.
And we can see that we will get different rules if we set different support and confidence thresholds.

3. Applications of association rule mining

(1) Beyond market basket analysis, association rule mining can be applied in various areas. First, association rule can be used for improving website effectiveness.

By using association rule mining on web usage log files, we can extract patterns of web user behavior, then we are able to conduct a web usage pattern analysis. The knowledge we get about website visitor behavior may be beneficial to increase client satisfaction, and the results are useful for a webmaster to increase their website effectiveness. In this case, we treat a web resource as an item, and a website visitor session can be considered as a transaction of items. For example, if we have the web usage log files of a website in January 2015 and January 2018 respectively, we can conduct an analysis to find out the changes of support, confidence and lift levels of association rules during these two different time periods, which can help webmasters to make an action to increase the website effectiveness.

The raw web usage log files will contain a lot of web requests and all related information. Since we do not need some irrelevant visits so we have to clean our data first. And we need to group all entries into visitor sessions. During a website visit, when the visitor browses the website and then returns, wen treat this visit as a session.

The final data we will use will contain the corresponding visitor sessions in January 2015 and January 2018, respectively.

Once we have cleaned up our data, we can conduct the association rule mining.

We use the package “Arules” and the apriori function in R to mine association rules. And we need to set the appropriate minimum support and confidence thresholds levels. Since we have data in two different time periods, ie, January 2015 and January 2018, we will get two rule sets. The main goal of our study is to investigate how those rules change over time, with these information, webmasters can make decisions to improve their website structure easily.

For instance, we want to compare the rule sets generated from the web usage log file of Carleton University in January 2015 and January 2018.

From January 2015 to January 2018, the website of Carleton University may be changed slightly. Some links and pages may have been removed and some pages may be added, but most of old pages still exist. Therefore, most rules found in the old rule sets still there which gives us the opportunity to compare changes of rules and their confidence levels. The changes in the confidence levels of the association rules may reflect the changes in web visitor behavior.

Suppose we get two rules set:

Carleton Central-----cuLearn confidence (2015): 0.68
Carleton Central-----cuLearn confidence (2017): 0.42

cuLearn-----Carleton Central confidence (2015):0.31
cuLearn-----Carleton Central confidence (2017):0.22

Based on the knowledge that in the January of 2015, 68% students who visited Carleton Central page also visited the cuLearn page, the webmaster decided to add a new link from the Carleton Central page to the cuLearn page. Then in January 2018, however, finding that the confidence dropped significantly (almost 20%) to 0.42, so the webmaster can consider to remove the cuLearn page link from the Carleton Central page. The webmaster may decide to watch the association between two pages in the future.

Besides, the opposite link from cuLearn to Carleton Central had not been added in 2015, and the decrease of the confidence level (from 0.31 to 0.22) confirmed that the earlier decision of the webmaster.

(2) Association rule mining can also be used for insurance claim analysis. This application will help insurance companies use insurance data to gain useful information in order to manage different claims properly as well as reduce claim costs. For business owners, with association rule mining, they can conduct effective safety program.

For example, in Workers Compensations insurance, understanding the cause of injury is crucial for preventing repeat injuries. In this case, association rule mining can enable insurers and self-insurers to understand the pattern of circumstances related to the injuries.

So if the data we have is worker compensation data, by using the association rule mining, the study can help prevent workplace injuries and it will provide useful information for employers who want to improve safety program then take prevent measures.

First, we need to prepare our data. The claim data should contain details of injuries and incidents, such as the place and time when incidents occurred, the information of injured people, cause of injuries and affected body locations, etc. Since we will gain a large number of rules, we also need to define our goals. If we want to find the major causes of head injuries in our data, we can find the most common circumstances and factors associated with head injuries.

Once we have appropriate data and clear goals, we can use R to conduct association rule mining. The lower the level of support and confidence picked, the more rules will emerge.

For instance, if an employer is seeing a lot of eye injuries in his company and those injury claims cost a lot annually. He wants to prevent these injuries as well as conduct an employee safety program in his workplace. By doing this, he needs to understand the circumstances associated with eye injuries. After running the association rule mining, we get a rule as below

{Foreign Object, Night Shift, Debris} ---- {Eye Injuries}

Support: 1.67% Confidence: 0.75 Lift: 12.18

This rule tells us the eye injuries are highly associated with night shift, and the cause of injury is likely being hit by foreign object such as debris. The confidence level is 0.74 which is highly significant. Therefore, an eye safety program is essential and this program can require employees wearing protective gear such as goggles to protect their eyes from being injured.