CARLTON UNIVERSITY

STAT 5703-Assignment 1

By- Hemant Gupta (101062246)-STAT5703

Index

SOLUTION-1

Part 1 Code: - Please refer file Assignment1_Q1.R

1. We unzip the cars.zip file from the link http://davis.wpi.edu/xmdv/datasets/cars.html and unzip it.
2. We check that there are cars.OKC file which contains the data. First row contain header and second row contain the garbage value.
3. Manually delete the garbage value and save the file in location.
4. In Code, first we set the path for cars data file i.e. cars.OKC
5. Now we read the data using read.table command with header equals to TRUE.
6. Factoring the Origin column of Cars.dat and assigning it to Cars.col.
7. On looking at the value of Cars in Year Column, we found there are few values which has year greater then 100, we clean that data by reducing their value by 100.
8. We have no NA value in my data
9. Plotting the complete Car.dat using Pair Plot.
10. Install ggobi library and using it plot Scatterplot and Parallel Coordinates Graphs
11. Now we draw co-plot of all five parameter in reference year with conditional on origin
12. Subset the Cars.dat based on Origin #Categorize data by origin
13. We draw the box plot of six parameter with respect to origin to compare.
14. Subsetting the Cars.USA data on the years.
15. Plotting Cars.USA data based on the Year.
16. Subsetting the Cars.EUR data on the years.
17. Plotting Cars.EUR data based on the Year.
18. Subsetting the Cars.JPN data on the years.
19. Plotting Cars.JPN data based on the Year.

Part 2: - Output and Deduction: -

Scatter Plot and Parallel Coordinate Plot:- This show the Overall picture of the data. We can predict few things but not with conformity.

Coplot: - Helps us to see the changes yearly on each parameter for each origin but it will not show the overall average of data clearly

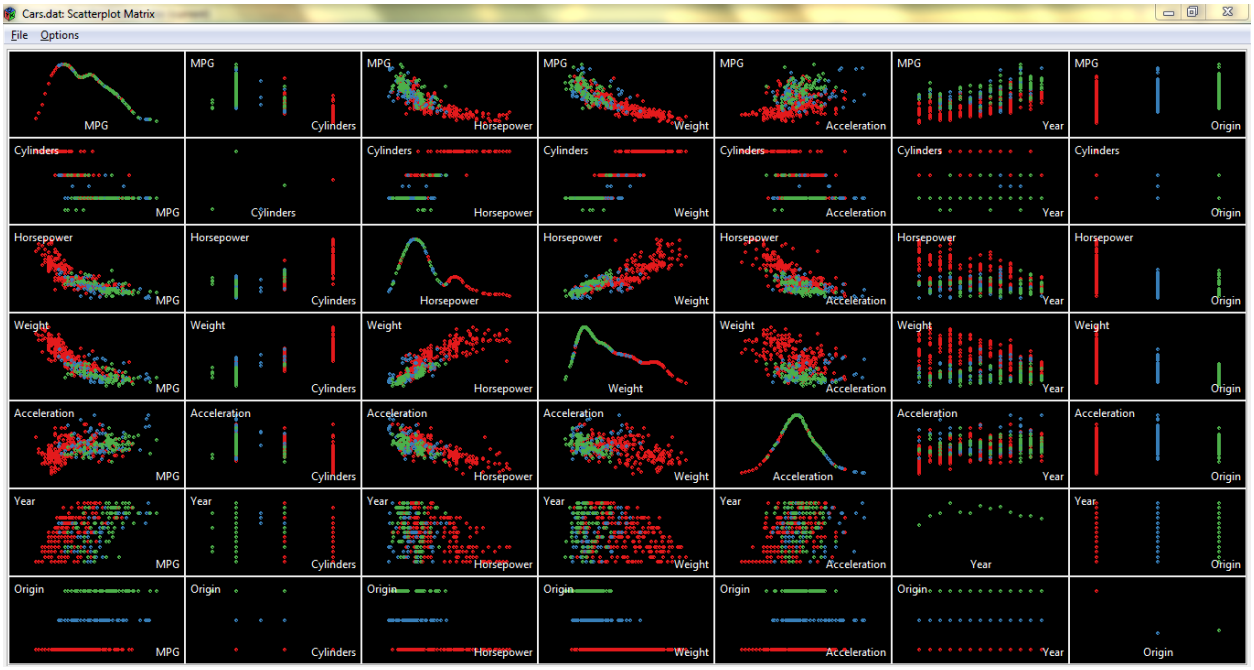
Boxplot: - Help to determine the average value in each column and deduction becomes easier.

We have use different plot like scatterplot, co-plot, boxplot to deduct something from the graphs. General Deduction of Overall Graph

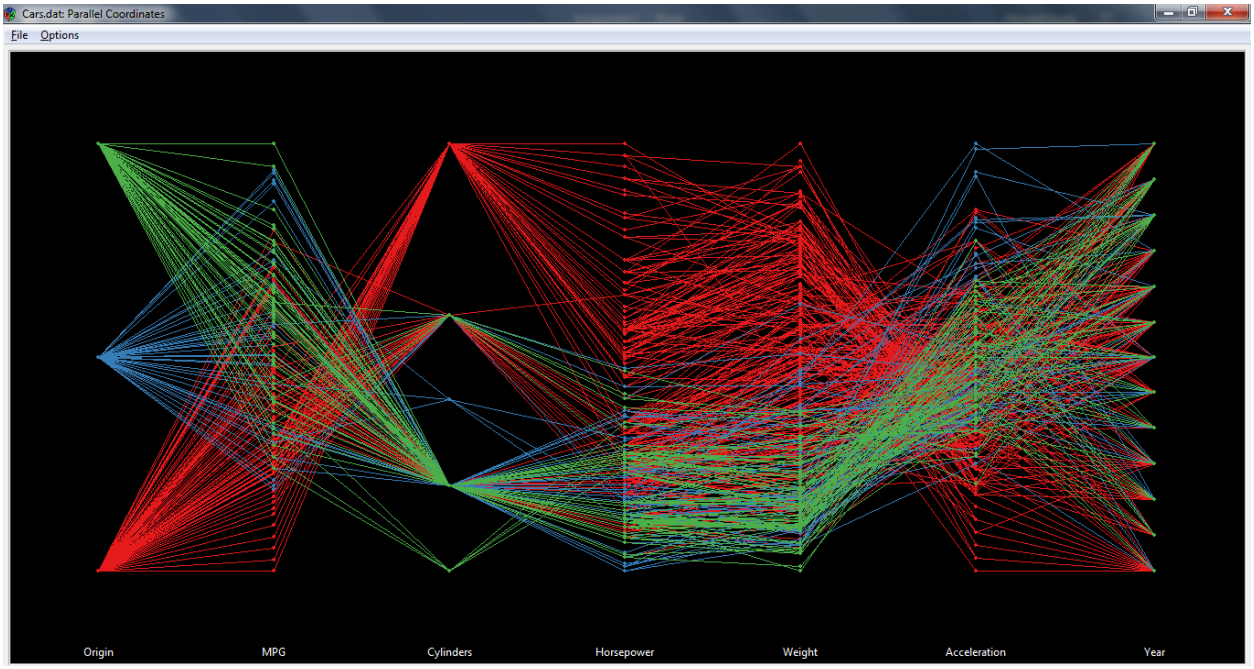
1. USA earlier in 1970's prefer cars with large no. of cylinders and large horsepower then Japan and Europe but as moving towards 1982, they also prefer 4 cylinder car as it gives them high MPG.
2. Each country trying to also reduce the weight of the cars.
3. Over the Year everyone trying to improve the acceleration.

Analysis OF the Graphs: -

1. **Scatter plot Graph**: - Red Represent USA, Japan Represent Europe and Green Represent Europe
    a. Cars in USA has high no. of cylinders but their MPG is low whereas Cars in Europe has less no. of cylinders but their MPG is high.
    b. Horsepower in USA Cars is very high as they are using large no. of cylinders in their cars in comparison to cars in Europe and Japan.
    c. Weight of the Cars is also directly proportional to number of cylinder in car. According to plot it seems USA cars has large weight with respect to no. of cylinders
    d. Acceleration in Japan Cars are high as compare to cars in USA and Europe.

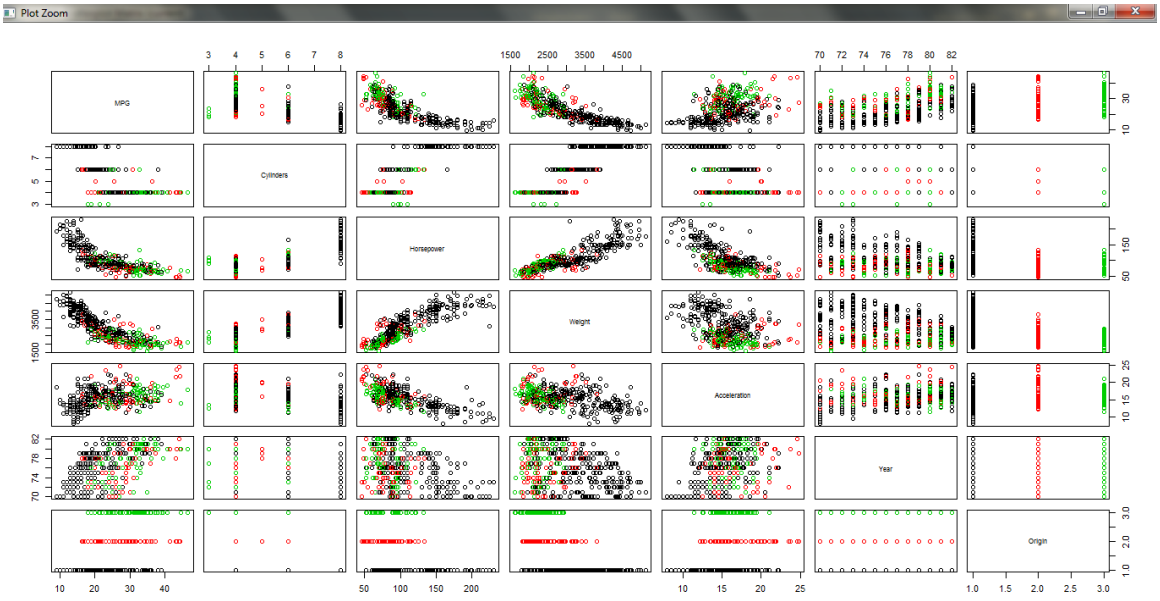


2. **Parallel Coordinates Graph:** - Red Represent USA, Blue Represent Europe and Green Represent Japan
    a. Using this graph we can confirm that cars made in USA has high number of cylinders, larger horsepower and larger weight as compare to Cars in Europe and Japan.



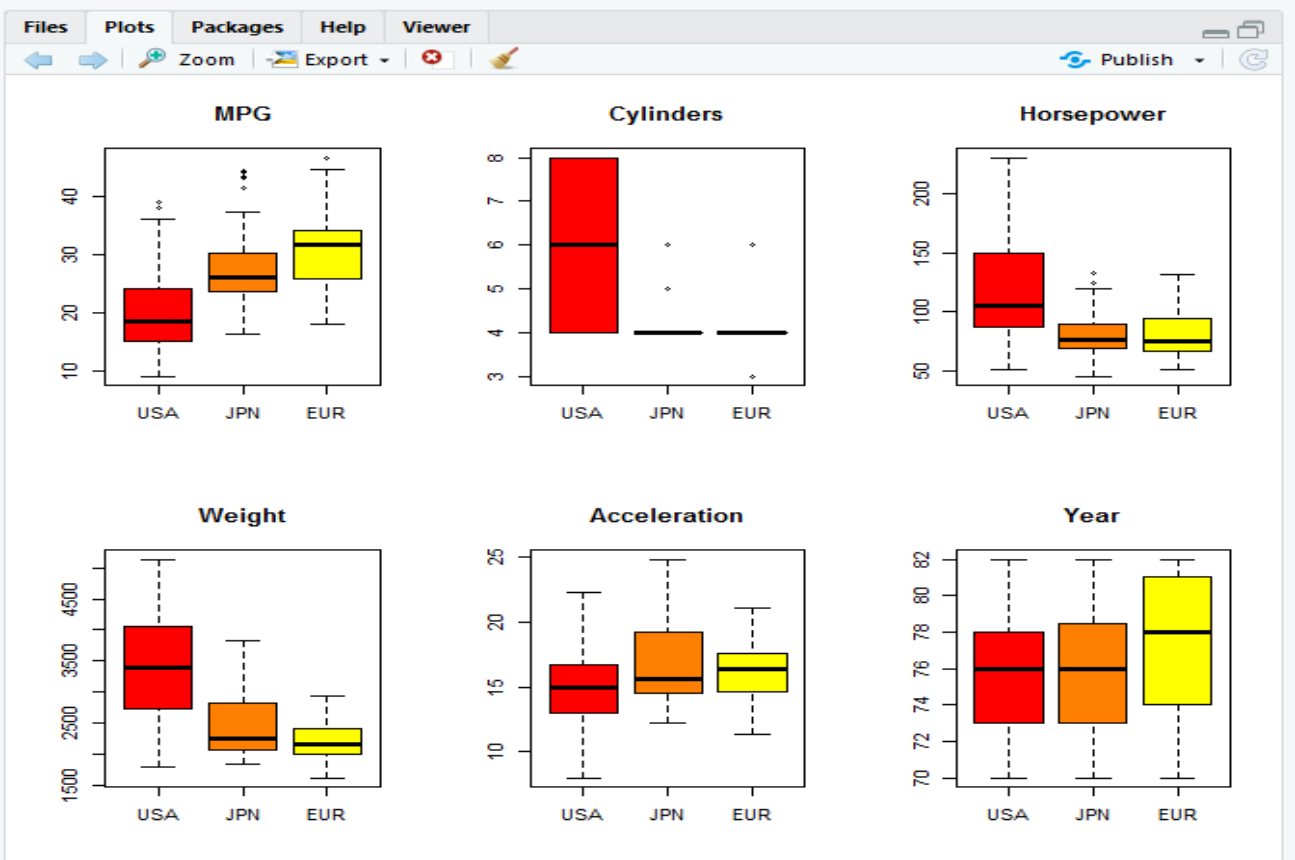
3. **Pair Plot**: - Black Represent USA, Red Represent Japan and Green Represent Europe. This plot also provides the scale for deduction.

    a. Cars in USA has high no. of cylinders up to 8 but their MPG is low between whereas Cars in Europe has less no. of cylinders mainly cars use 4 cylinders but their MPG is high.
    b. Horsepower in USA Cars is very high above 200 as they are using large no. of cylinders in their cars in comparison to cars in Europe and Japan.
    c. Weight of the Cars is also directly proportional to number of cylinder in car. According to plot it seems USA cars has large weight upto 4500 with respect to no. of cylinders
    d. Acceleration in Japan Cars are high up to 25 as compare to cars in USA and Europe.

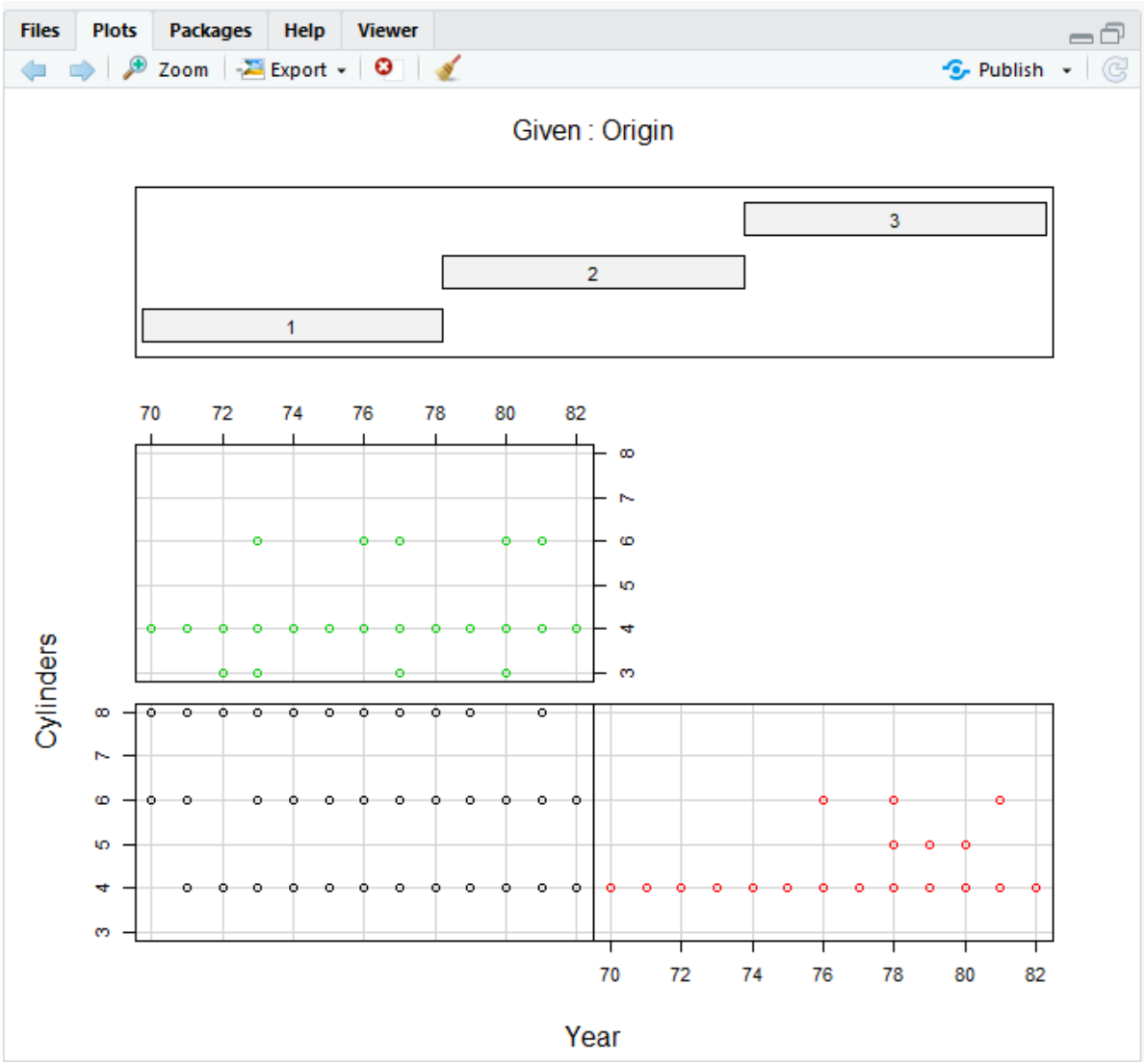


4. **BoxPlot:** - In this plot we have plotted all parameter with respect to Origin. For more proper deduction.

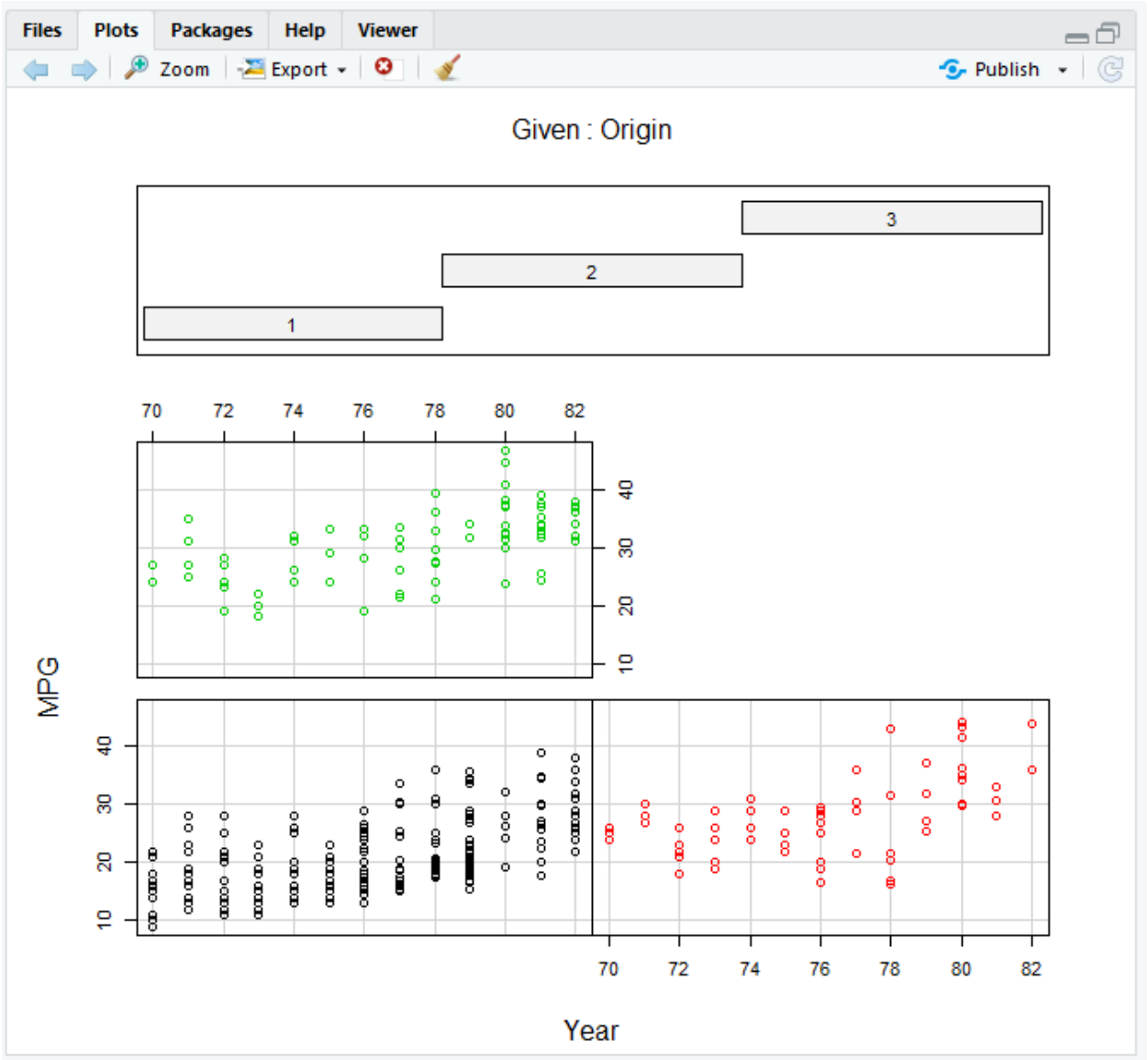
    a. MPG- Europe Cars has high Miles per gallon then USA and JAPAN. Average of 30 MPG and maximum 35.
    b. Cylinders- Average number of Cylinder in USA cars are 6 and maximum is 8 while cars in Europe and Japan mainly uses 4 cylinders.
    c. Horsepower- Average horse power in USA cars is 100 while for cars in Japan and Europe its nearly equal to 75.
    d. Weight in USA cars are very high as compare to cars in Japan and Europe.
    e. Average acceleration of cars in these three countries are same but Japanese cars has highest value in acceleration.



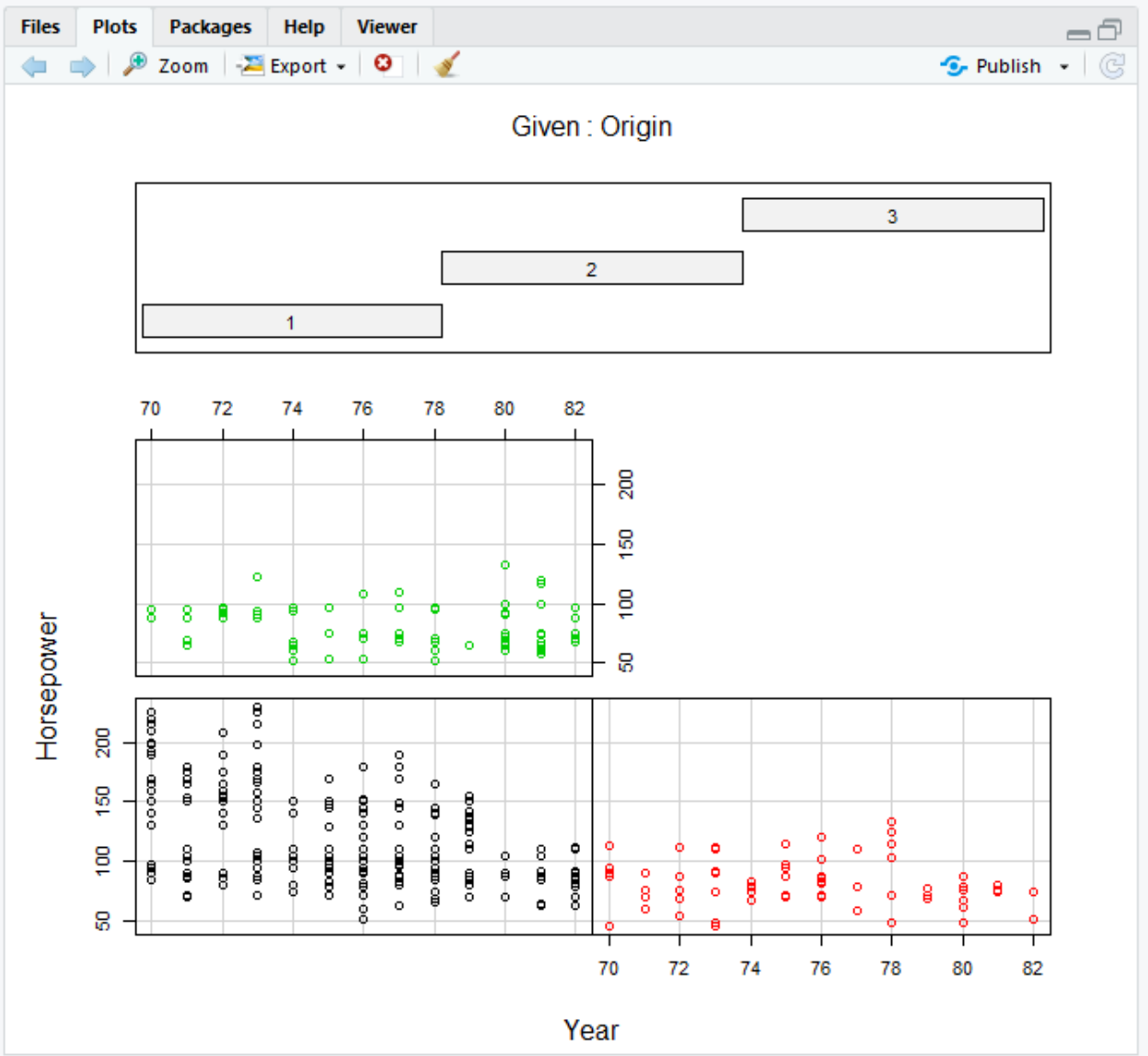
5. **CoPlot Cylinders**- Black Represent USA , Red Represent Japan and Green Represent Europe.
   Here we have plot variation in cylinders yearly for each origin.
   a. In USA there are commonly distribution of no. of cylinders yearly between 4, 6 and 8.
   b. In Japan, they start with 4 cylinder then in year 1976 they move to 6, then in year 1978, they started on 5 cylinder cars and in 1982 they again move to 4 cylinder cars.
   c. In Europe, maximum no. of cars contain 4 cylinders, they also tried with 3 and 6 cylinders between 1972 to 1981.



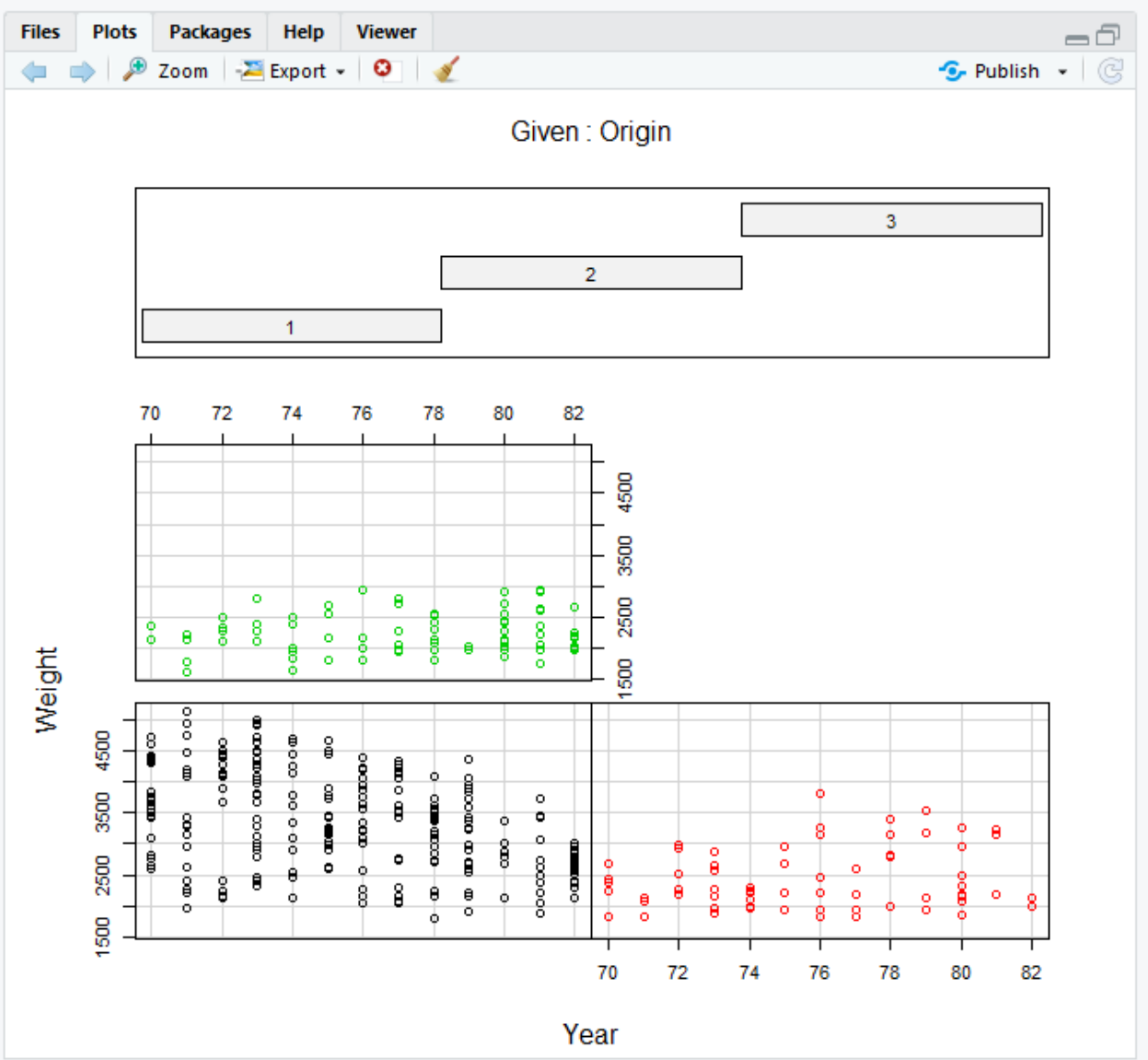
6. **CoPlot MPG-** Black Represent USA , Red Represent Japan and Green Represent Europe. Here we have plot variation in cylinders yearly for each origin.
   a. In USA there is improvement in increasing the MPG value yearly in 1982 they reach 40MPG as compared to nearly 20 MPG in 1970.
   b. In Japan also there is improvement in increasing the MPG value yearly in 1980 they reach 45 MPG as compared to nearly 25 MPG in 1970..
   c. In Europe, there is improvement in increasing the MPG value yearly in 1980 they reach nearly 50 MPG as compared to nearly 25 MPG in 1970. But in next two years it has decreased to 40 MPG.



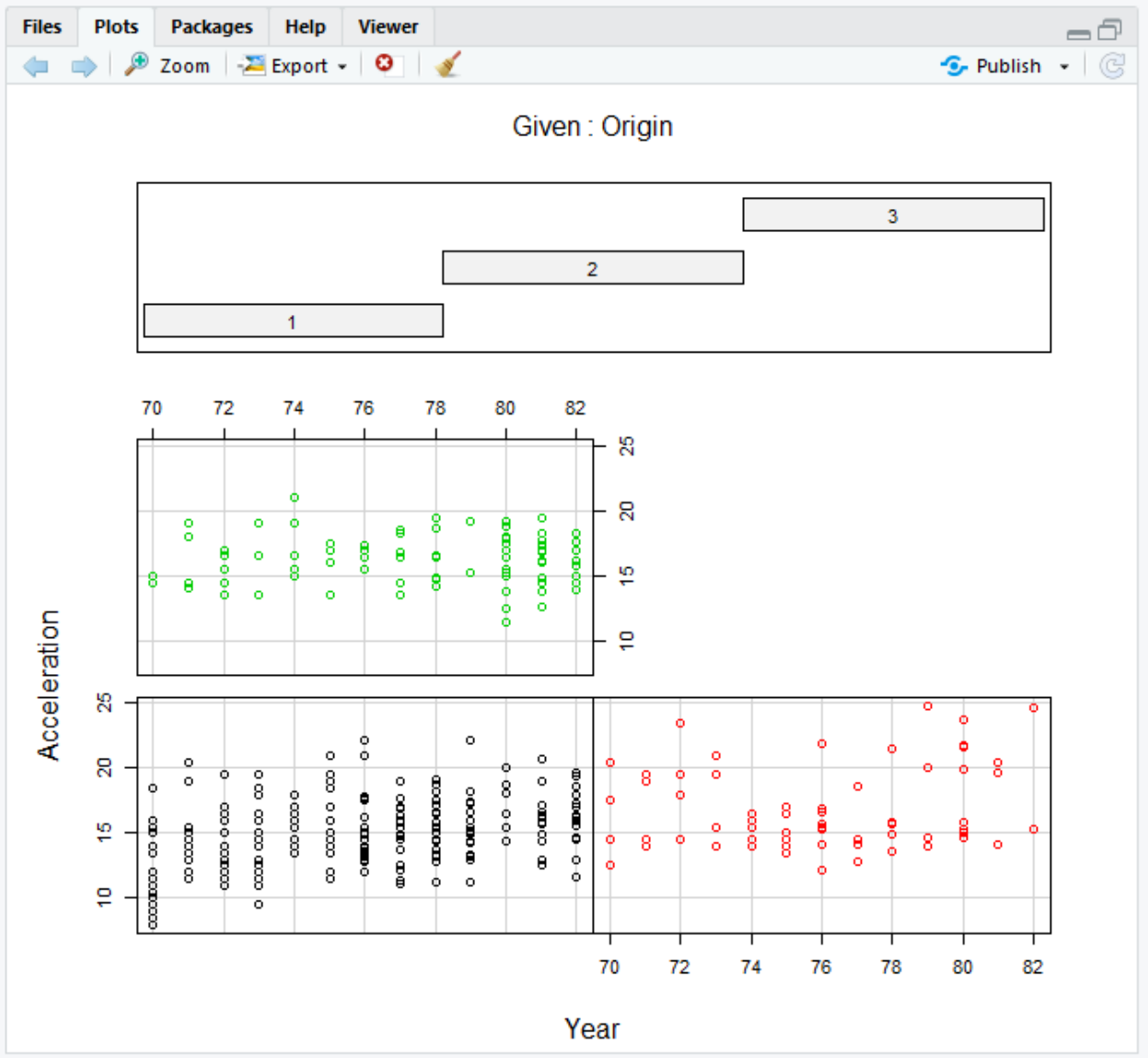
7. **CoPlot Horsepower**- Black Represent USA , Red Represent Japan and Green Represent Europe.
   Here we have plot variation in cylinders yearly for each origin.
   a. In USA, there is decrease in horsepower value yearly in 1982 they reach to 100 as compared to nearly 250 in 1970.
   b. In Japan , horsepower remain nearly on an average or below 100 each year.
   c. In Europe, horsepower remain nearly on an average or below 100 each year.



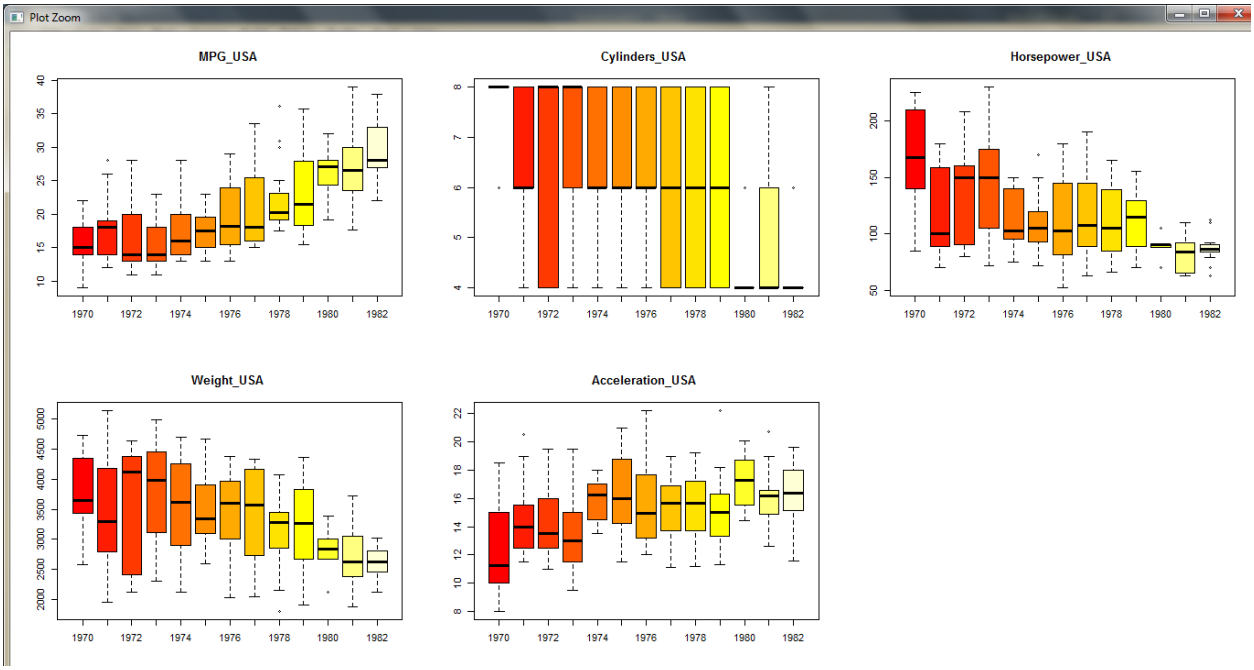
8. **CoPlot Weight**- Black Represent USA , Red Represent Japan and Green Represent Europe. Here we have plot variation in cylinders yearly for each origin.
   a. In USA, there is decrease in weight of cars value yearly in 1982 they reach to below 3000 as compared to nearly 4500 in 1970.
   b. In Japan, it's like a bell curve in 1970 they start with 2500 and move to maximum of nearly 4000 in 1976 and then decrease to 2000 in 1982.
   c. In Europe, weight of the cars mainly remain between 2000 and 2500 all years.



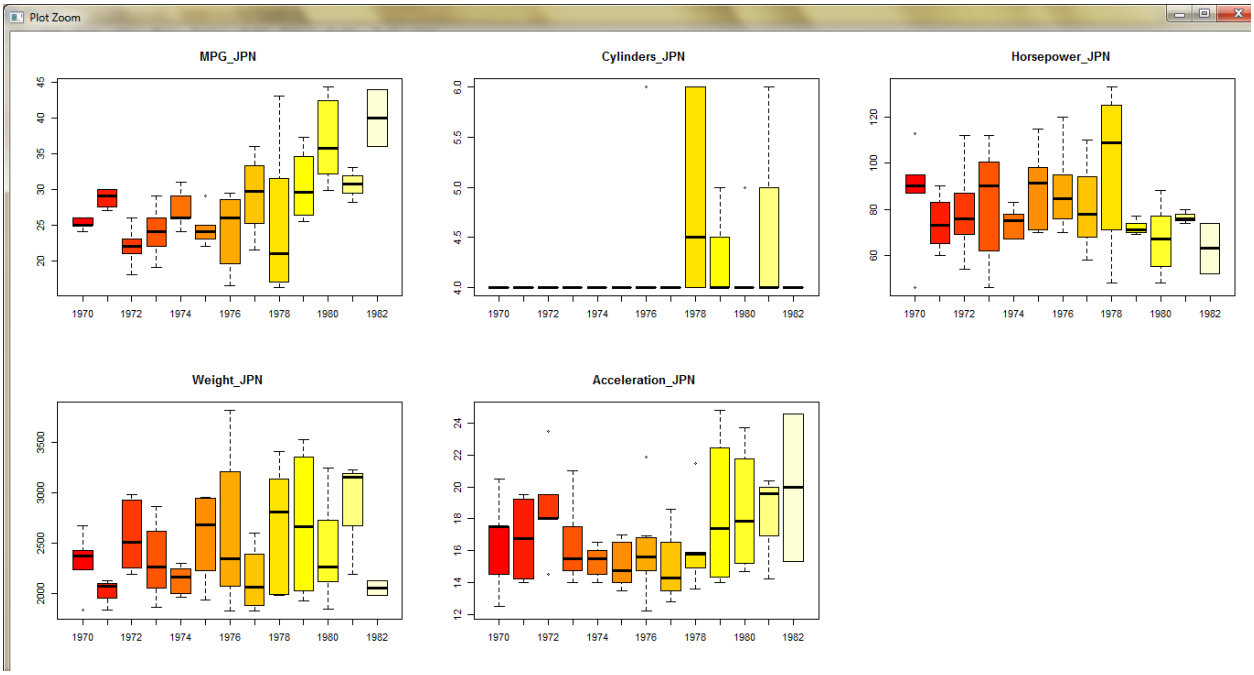
9. **CoPlot Acceleration**- Black Represent USA , Red Represent Japan and Green Represent Europe.
   Here we have plot variation in cylinders yearly for each origin.
   a. In USA, car acceleration moves from 10 to 20 in each year.
   b. In Japan, it's like scatter plot it moves to maximum of nearly 25 in 1979.
   c. In Europe, car acceleration moves from 12 to 20 in each year.



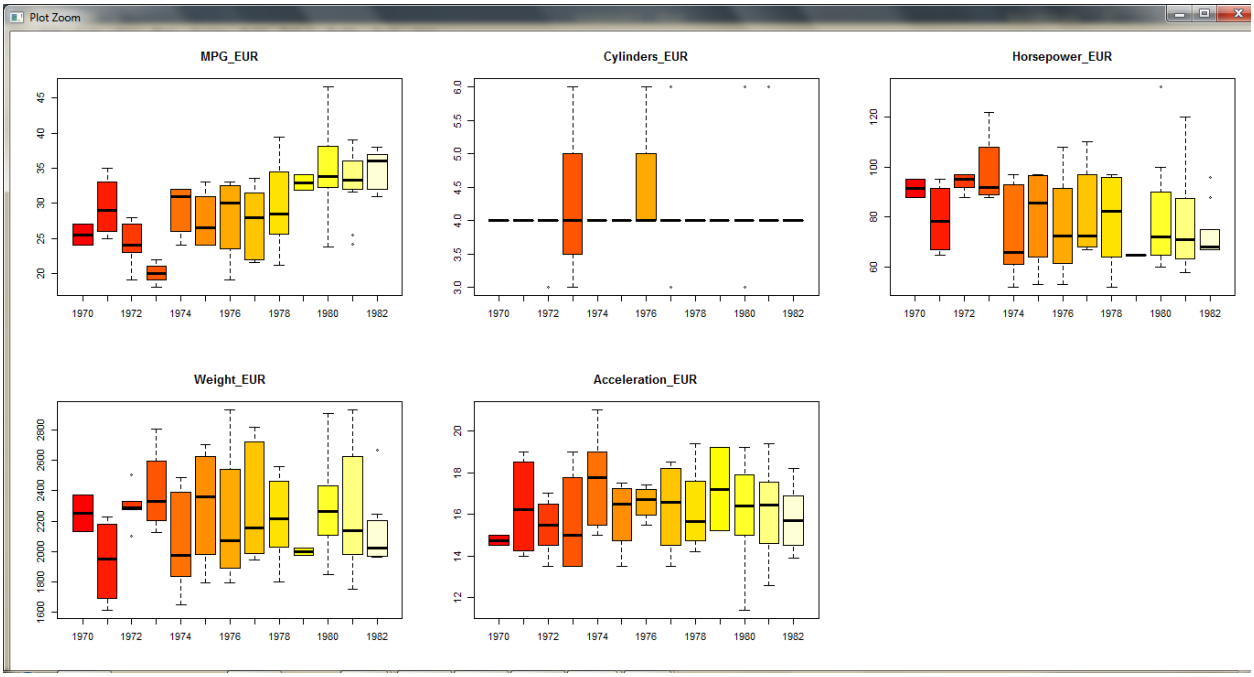
10. **BoxPlot USA-**
    - **a.** In plot, there is yearly increase in Miles per gallon values and acceleration of cars.
    - **b.** In plot, there is decrease in no. of cylinders used in cars and horsepower and weight.



11. **BoxPlot Japan-**
    a. In plot, there is yearly increase in Miles per gallon values and acceleration of cars.
    b. In plot, they normally use 4 cylinders cars but in 1978 they also tried with 6 cylinders and an 1982 move to the 4 cylinder again.
    c. There is not much variation in case of horse power but very high fluctuation in case of weight of the car..



12. **BoxPlot Europe**-
    a. In plot, there is yearly increase in Miles per gallon values and decrease in horsepower of cars.
    b. In plot, they normally use 4 cylinders cars but in 1973 they also tried with 5 cylinders and in 1977 move to the 4 cylinder again.
    c. There is not much variation in case of weight and acceleration.



SOLUTION -2

Code: - Please check the Assignment1_Q2.R

Algorithm:-

1. We set the Data directory in the code
2. We then load the data in the variable online_data.
3. Split the original_data on the basis of description and invoice no.
4. Apply arules on the output and using support and confidence plot scatterplot graph such that we get values of support and confidence for 10 best rules.
5. Plot the rules
6. Plot the rules for minimum 3 items and minimum 4 items
7. Now we remove the cancelled transaction from original data.
8. Apply arules on the output and using support and confidence plot scatterplot graph such that we get values of support and confidence for 10 best rules.
9. Plot the rules
10. Then we take subset of original data where country is equal to United Kingdom.
11. Apply arules on the output and using support and confidence plot scatterplot graph such that we get values of support and confidence for 10 best rules.
12. Plot the rules.

.

Output and Deduction: -

1. As we filter the data maximum value of support and confidence increases with respect to support on original data.
2. In case of rules applied 2 items nearly 900 people buy and then on 3 items nearly 550 people buy therefore maximum value of support decreases as many people do not buy three items together but confidence and lift increase due to less no. people by 3 items together.
3. In case of rules applied 2 items and then on 3 items maximum value of support decreases as many people do not buy three items together but confidence and lift increase due to less no. people by 3 items together.
4. Less than 300 people buy 4 items together and therefore support decrease but increase in confidence and lift.

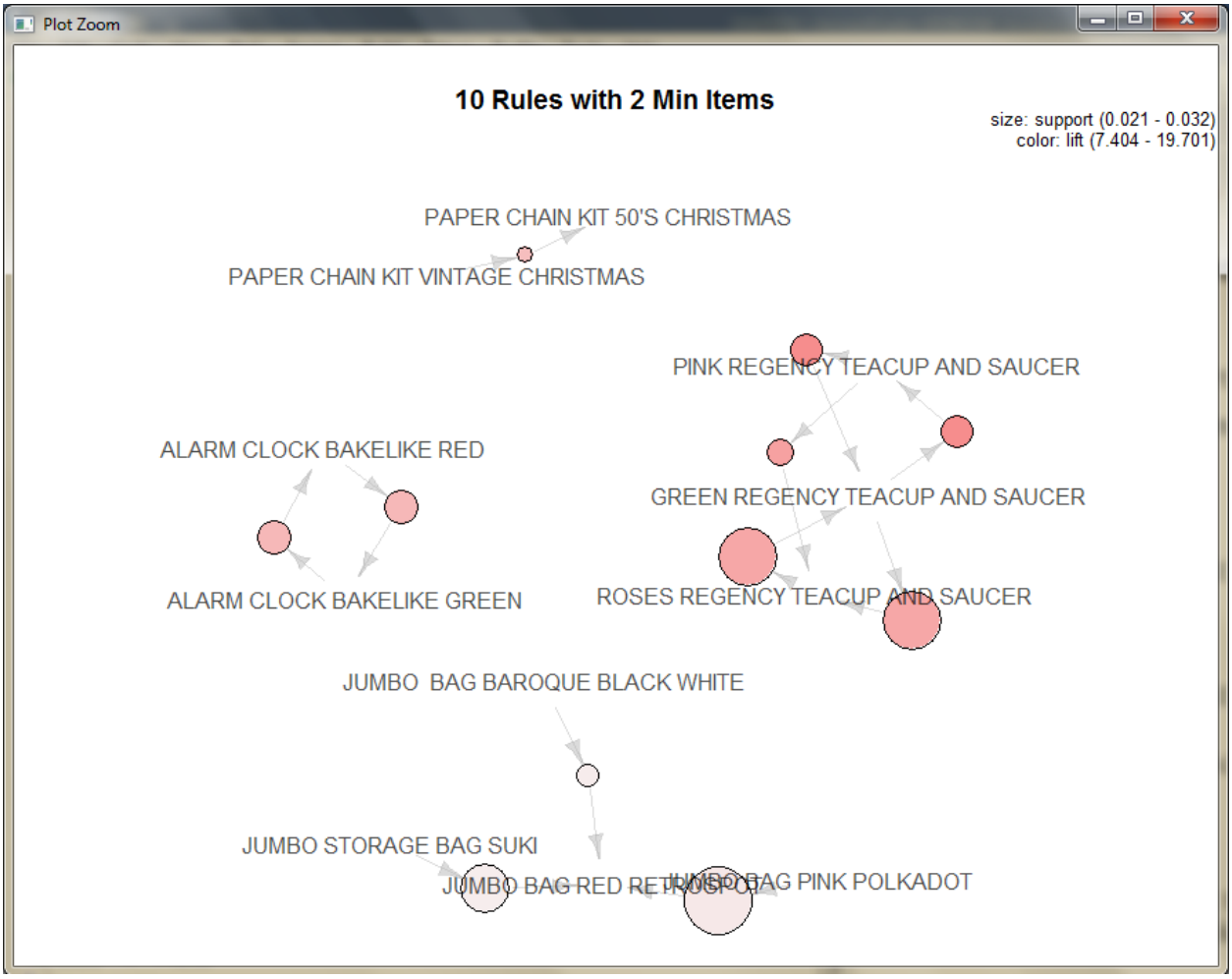
   And with the help of rules and no. of people participating we can check the which items people usually buy together and then in case someone select one item to purchase we can suggest him/her another item based on these rules.
   Generally,
   JUMBO BAG PINK POLKADOT with JUMBO BAG RED RETROSPOT, gives highest support in each case.
   In 2 items, with high confidence and lift, mostly people likely to buy PINK REGENCY TEACUP AND SAUCER with GREEN REGENCY TEACUP AND SAUCER.

   Below is few more deduction with respect to output.

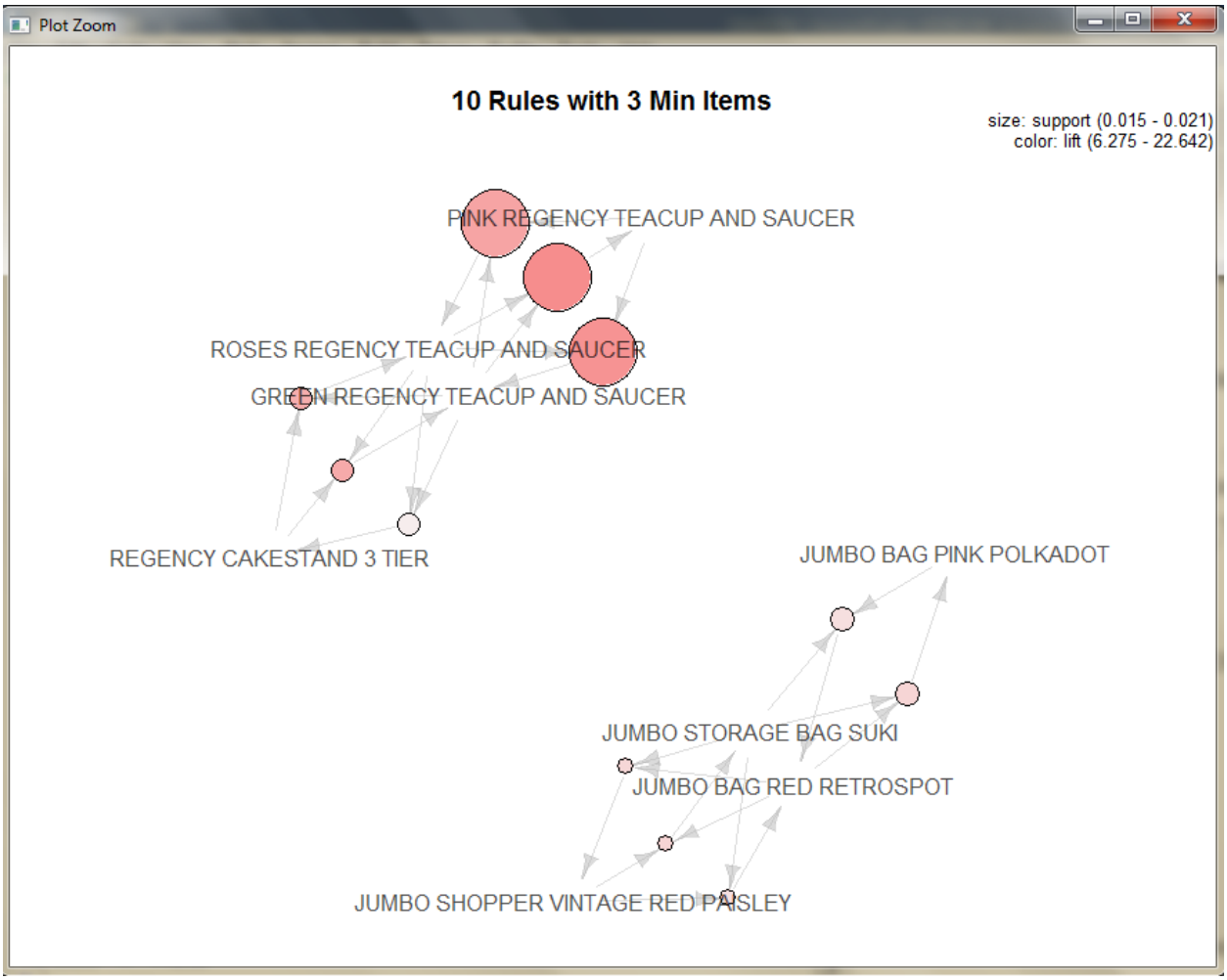
Output: -

1. We put the arules on original data and get output of rule. And then we apply arules to get relation of rules with minimum 3 items and minimum 4 items
   a. 873 people buy JUMBO BAG PINK POLKADOT with JUMBO BAG RED RETROSPOT, gives highest support.
   b. With high confidence and lift, 644 people likely to buy PINK REGENCY TEACUP AND SAUCER with GREEN REGENCY TEACUP AND SAUCER
   c. In case of 3 items, with high support, confidence and lift, 549 people likely to buy PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER with GREEN REGENCY TEACUP AND SAUCER
   d. In case of 4 items, with high support, confidence and lift, 303 people likely to buy PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER, REGENCY CAKESTAND 3 TIER with GREEN REGENCY TEACUP AND SAUCER

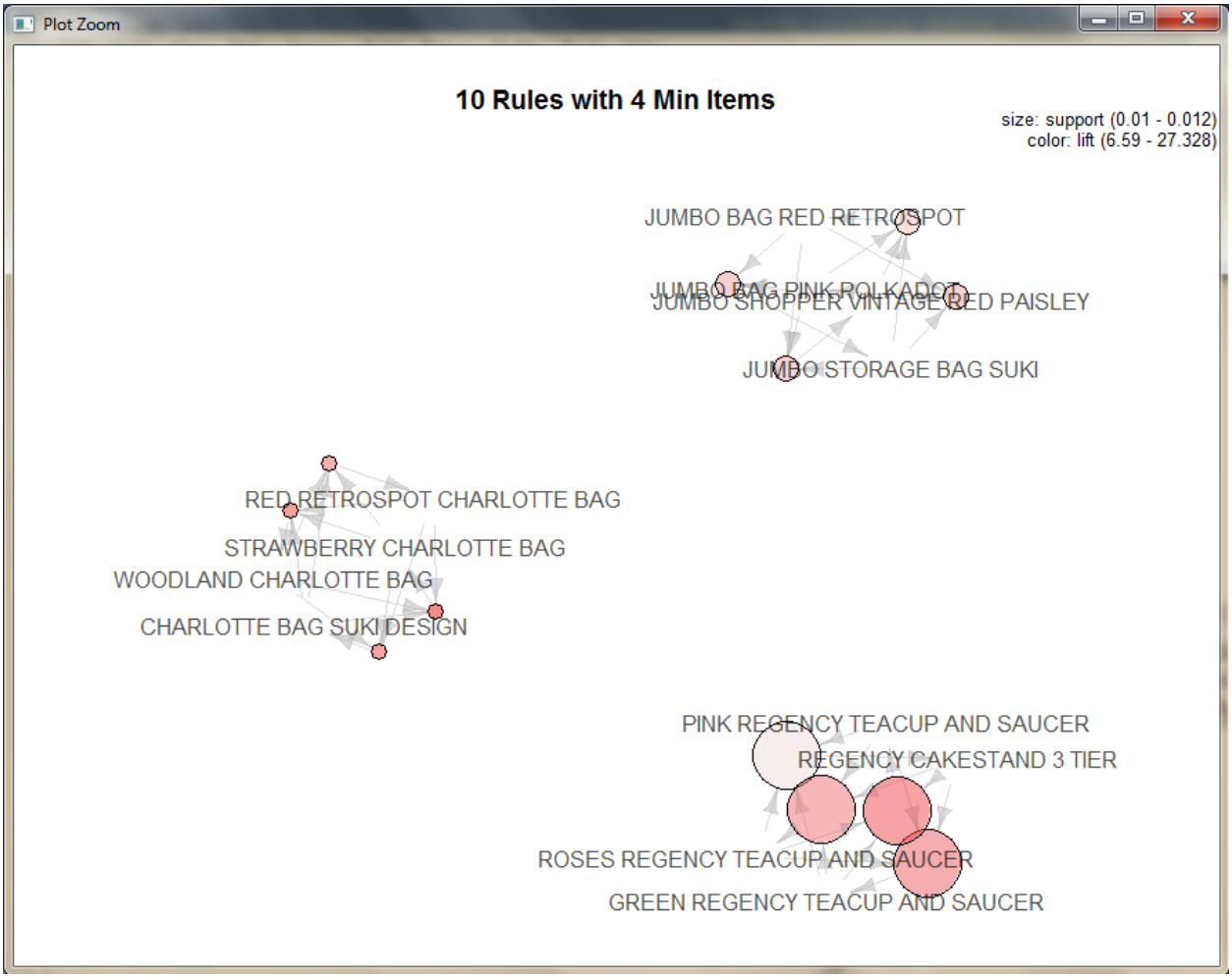
```
> inspect(head(rules,n=10))
     lhs                                  rhs                                  support    confidence lift      count
[1]  {PINK REGENCY TEACUP AND SAUCER}     => {ROSES REGENCY TEACUP AND SAUCER} 0.02370656 0.7665418  17.726280 614
[2]  {PINK REGENCY TEACUP AND SAUCER}     => {GREEN REGENCY TEACUP AND SAUCER} 0.02486486 0.8039950  19.700540 644
[3]  {GREEN REGENCY TEACUP AND SAUCER}    => {PINK REGENCY TEACUP AND SAUCER}  0.02486486 0.6092715  19.700540 644
[4]  {PAPER CHAIN KIT VINTAGE CHRISTMAS} => {PAPER CHAIN KIT 50'S CHRISTMAS}  0.02142857 0.6670673  14.766704 555
[5]  {ALARM CLOCK BAKELIKE RED}           => {ALARM CLOCK BAKELIKE GREEN}      0.02494208 0.5975948  15.462244 646
[6]  {ALARM CLOCK BAKELIKE GREEN}         => {ALARM CLOCK BAKELIKE RED}        0.02494208 0.6453546  15.462244 646
[7]  {JUMBO  BAG BAROQUE BLACK WHITE}     => {JUMBO BAG RED RETROSPOT}         0.02289575 0.6261880   7.596379 593
[8]  {JUMBO BAG PINK POLKADOT}            => {JUMBO BAG RED RETROSPOT}         0.03216216 0.6766856   8.208973 833
[9]  {ROSES REGENCY TEACUP AND SAUCER}    => {GREEN REGENCY TEACUP AND SAUCER} 0.03027027 0.7000000  17.152318 784
[10] {GREEN REGENCY TEACUP AND SAUCER}    => {ROSES REGENCY TEACUP AND SAUCER} 0.03027027 0.7417219  17.152318 784
>
```




```
> inspect(rules.3.items)
     lhs                                                rhs                                      support confidence      lift count
[1]  {PINK REGENCY TEACUP AND SAUCER,
      ROSES REGENCY TEACUP AND SAUCER}    => {GREEN REGENCY TEACUP AND SAUCER}    0.02119691  0.8941368 21.909313   549
[2]  {GREEN REGENCY TEACUP AND SAUCER,
      PINK REGENCY TEACUP AND SAUCER}     => {ROSES REGENCY TEACUP AND SAUCER}    0.02119691  0.8524845 19.713703   549
[3]  {GREEN REGENCY TEACUP AND SAUCER,
      ROSES REGENCY TEACUP AND SAUCER}    => {PINK REGENCY TEACUP AND SAUCER}     0.02119691  0.7002551 22.642456   549
[4]  {JUMBO BAG PINK POLKADOT,
      JUMBO STORAGE BAG SUKI}             => {JUMBO BAG RED RETROSPOT}            0.01606178  0.8030888  9.742389   416
[5]  {JUMBO BAG RED RETROSPOT,
      JUMBO STORAGE BAG SUKI}             => {JUMBO BAG PINK POLKADOT}            0.01606178  0.5675307 11.940735   416
[6]  {JUMBO SHOPPER VINTAGE RED PAISLEY,
      JUMBO STORAGE BAG SUKI}             => {JUMBO BAG RED RETROSPOT}            0.01509653  0.7447619  9.034817   391
[7]  {JUMBO BAG RED RETROSPOT,
      JUMBO SHOPPER VINTAGE RED PAISLEY}  => {JUMBO STORAGE BAG SUKI}             0.01509653  0.5724744 12.345617   391
[8]  {JUMBO BAG RED RETROSPOT,
      JUMBO STORAGE BAG SUKI}             => {JUMBO SHOPPER VINTAGE RED PAISLEY}  0.01509653  0.5334243 11.639165   391
[9]  {GREEN REGENCY TEACUP AND SAUCER,
      ROSES REGENCY TEACUP AND SAUCER}    => {REGENCY CAKESTAND 3 TIER}           0.01590734  0.5255102  6.275110   412
[10] {REGENCY CAKESTAND 3 TIER,
      ROSES REGENCY TEACUP AND SAUCER}    => {GREEN REGENCY TEACUP AND SAUCER}    0.01590734  0.7672253 18.799561   412
[11] {GREEN REGENCY TEACUP AND SAUCER,
      REGENCY CAKESTAND 3 TIER}           => {ROSES REGENCY TEACUP AND SAUCER}    0.01590734  0.7953668 18.392857   412
```




```
> inspect(rules.4.items)
     lhs                                                                rhs                                    support    confidence     lift count
[1]  {GREEN REGENCY TEACUP AND SAUCER,
      PINK REGENCY TEACUP AND SAUCER,
      ROSES REGENCY TEACUP AND SAUCER}                              => {REGENCY CAKESTAND 3 TIER}          0.01169884  0.5519126  6.590381   303
[2]  {PINK REGENCY TEACUP AND SAUCER,
      REGENCY CAKESTAND 3 TIER,
      ROSES REGENCY TEACUP AND SAUCER}                              => {GREEN REGENCY TEACUP AND SAUCER}   0.01169884  0.8991098 22.031167   303
[3]  {GREEN REGENCY TEACUP AND SAUCER,
      PINK REGENCY TEACUP AND SAUCER,
      REGENCY CAKESTAND 3 TIER}                                     => {ROSES REGENCY TEACUP AND SAUCER}   0.01169884  0.8757225 20.251084   303
[4]  {GREEN REGENCY TEACUP AND SAUCER,
      REGENCY CAKESTAND 3 TIER,
      ROSES REGENCY TEACUP AND SAUCER}                              => {PINK REGENCY TEACUP AND SAUCER}    0.01169884  0.7354369 23.780044   303
[5]  {CHARLOTTE BAG SUKI DESIGN,
      STRAWBERRY CHARLOTTE BAG,
      WOODLAND CHARLOTTE BAG}                                       => {RED RETROSPOT CHARLOTTE BAG}       0.01007722  0.8585526 21.177632   261
[6]  {CHARLOTTE BAG SUKI DESIGN,
      RED RETROSPOT CHARLOTTE BAG,
      STRAWBERRY CHARLOTTE BAG}                                     => {WOODLAND CHARLOTTE BAG}            0.01007722  0.7767857 23.865658   261
[7]  {RED RETROSPOT CHARLOTTE BAG,
      STRAWBERRY CHARLOTTE BAG,
      WOODLAND CHARLOTTE BAG}                                       => {CHARLOTTE BAG SUKI DESIGN}         0.01007722  0.7957317 23.053077   261
[8]  {CHARLOTTE BAG SUKI DESIGN,
      RED RETROSPOT CHARLOTTE BAG,
      WOODLAND CHARLOTTE BAG}                                       => {STRAWBERRY CHARLOTTE BAG}          0.01007722  0.7744807 27.328407   261
[9]  {JUMBO BAG PINK POLKADOT,
      JUMBO SHOPPER VINTAGE RED PAISLEY,
      JUMBO STORAGE BAG SUKI}                                       => {JUMBO BAG RED RETROSPOT}           0.01038610  0.8677419 10.526705   269
[10] {JUMBO BAG PINK POLKADOT,
      JUMBO BAG RED RETROSPOT,
      JUMBO SHOPPER VINTAGE RED PAISLEY}                            => {JUMBO STORAGE BAG SUKI}            0.01038610  0.7078947 15.266006   269
[11] {JUMBO BAG PINK POLKADOT,
      JUMBO BAG RED RETROSPOT,
      JUMBO STORAGE BAG SUKI}                                       => {JUMBO SHOPPER VINTAGE RED PAISLEY} 0.01038610  0.6466346 14.109382   269
```

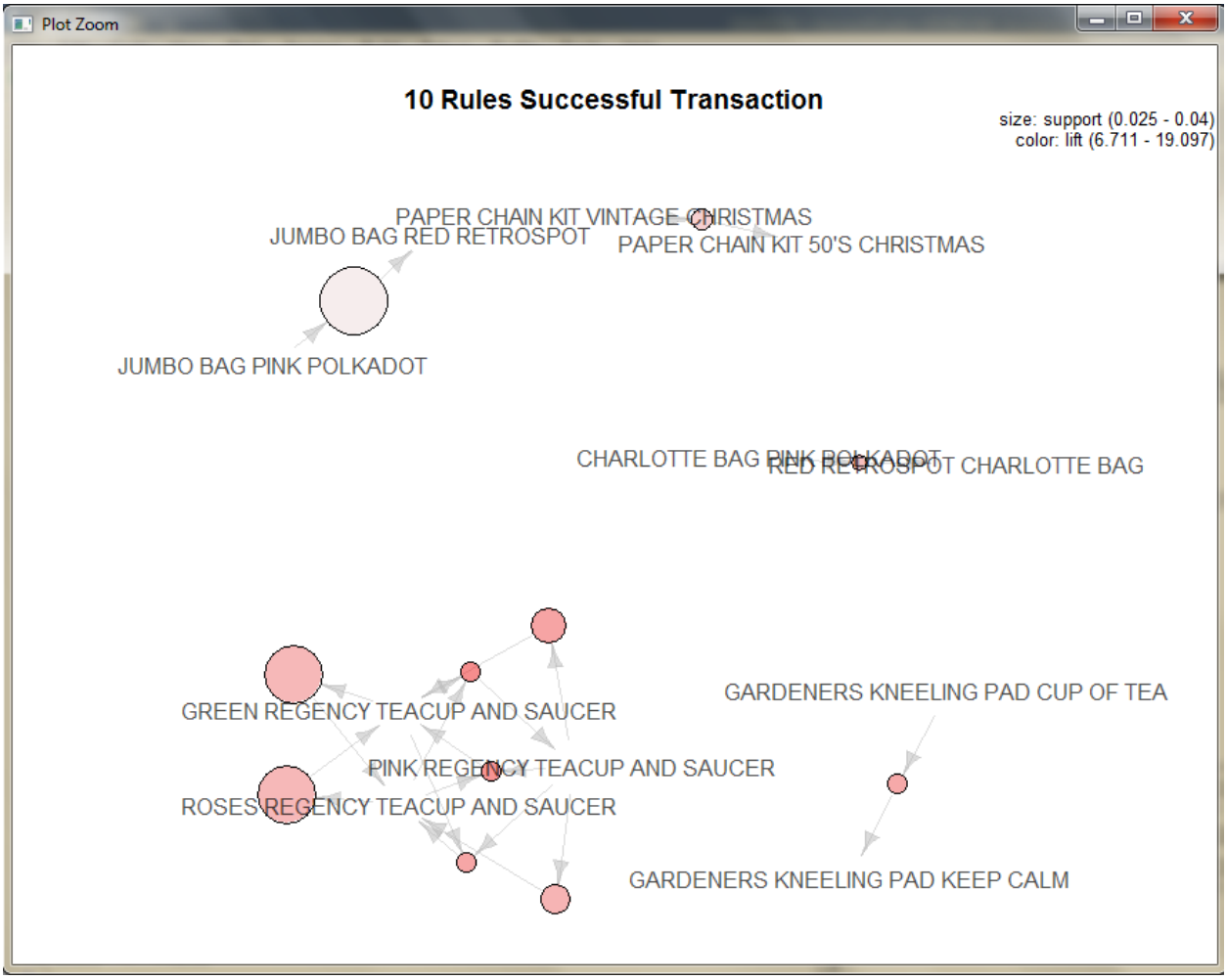


2. We put the arules on data and after removing the cancelled entries get output of rule.

a. 825 people buy JUMBO BAG PINK POLKADOT with JUMBO BAG RED RETROSPOT, gives highest support but value of support decreases as less no. of data.

b. In 3 items, with high confidence and lift, 542 people likely to buy PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER with GREEN REGENCY TEACUP AND SAUCER

```
> inspect(head(rules_success_trnx, n=10))
     lhs                                    rhs                                     support confidence      lift count
[1]  {PINK REGENCY TEACUP AND SAUCER}    => {ROSES REGENCY TEACUP AND SAUCER}  0.02889811  0.7819843 15.205414   599
[2]  {PINK REGENCY TEACUP AND SAUCER}    => {GREEN REGENCY TEACUP AND SAUCER}  0.03053840  0.8263708 16.875875   633
[3]  {GARDENERS KNEELING PAD CUP OF TEA} => {GARDENERS KNEELING PAD KEEP CALM} 0.02634118  0.7203166 16.353475   546
[4]  {PAPER CHAIN KIT VINTAGE CHRISTMAS} => {PAPER CHAIN KIT 50'S CHRISTMAS}   0.02663064  0.6731707 12.028865   552
[5]  {JUMBO BAG PINK POLKADOT}           => {JUMBO BAG RED RETROSPOT}          0.03980124  0.6773399  6.711234   825
[6]  {CHARLOTTE BAG PINK POLKADOT}       => {RED RETROSPOT CHARLOTTE BAG}      0.02518333  0.7025572 14.083758   522
[7]  {ROSES REGENCY TEACUP AND SAUCER}   => {GREEN REGENCY TEACUP AND SAUCER}  0.03705133  0.7204503 14.712801   768
[8]  {GREEN REGENCY TEACUP AND SAUCER}   => {ROSES REGENCY TEACUP AND SAUCER}  0.03705133  0.7566502 14.712801   768
[9]  {PINK REGENCY TEACUP AND SAUCER,
      ROSES REGENCY TEACUP AND SAUCER}   => {GREEN REGENCY TEACUP AND SAUCER}  0.02614821  0.9048414 18.478377   542
[10] {GREEN REGENCY TEACUP AND SAUCER,
      PINK REGENCY TEACUP AND SAUCER}    => {ROSES REGENCY TEACUP AND SAUCER}  0.02614821  0.8562401 16.649292   542
>
>
```

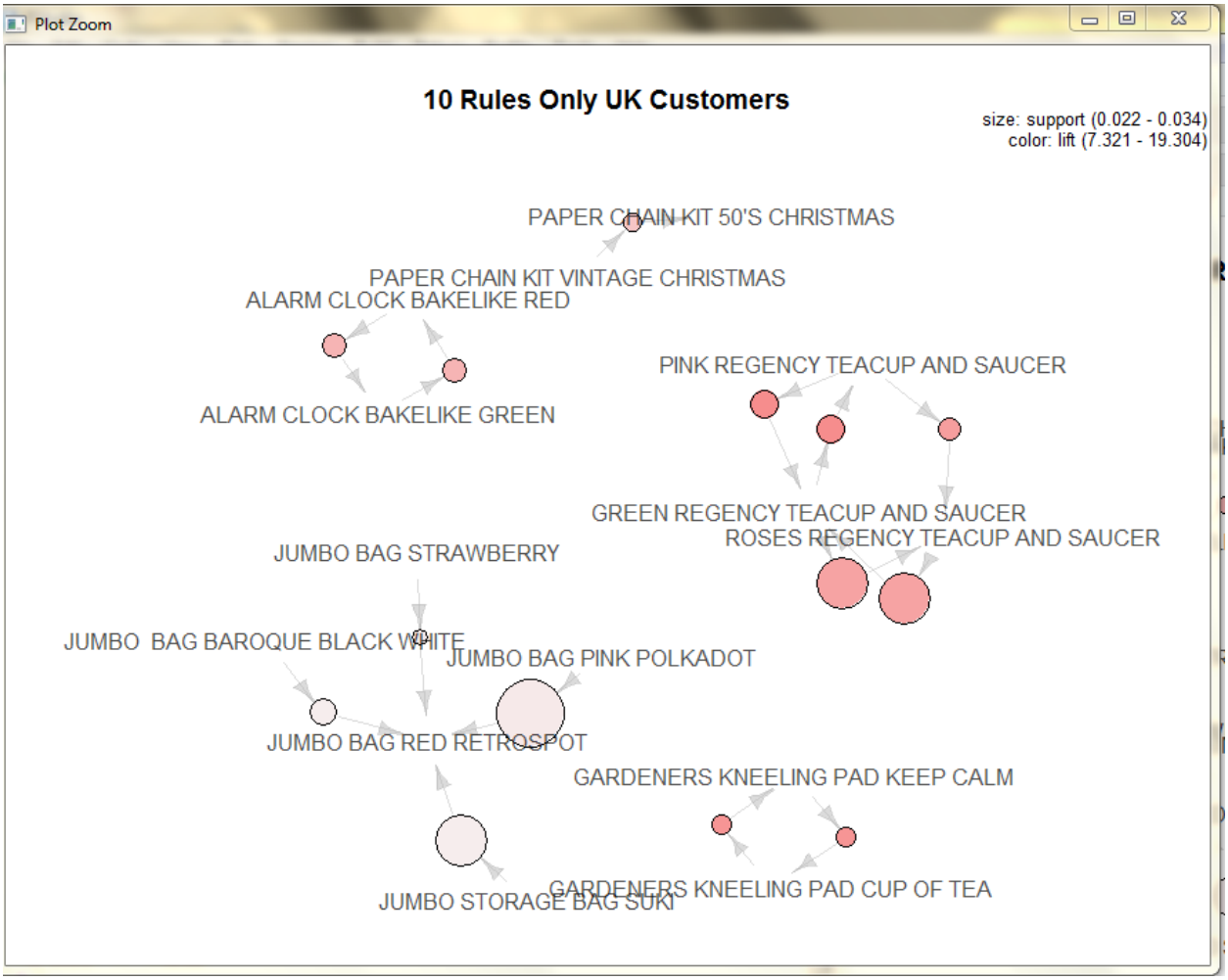


3. We put the arules on data which has country belong to United Kingdom as 495478 entries belong to UK and get output of rule.

a. 578 people buy JUMBO BAG PINK POLKADOT with JUMBO BAG RED RETROSPOT, gives highest support but value of support decreases as less no. of data.

b. In 2 items, with high confidence and lift, 585 people likely to buy PINK REGENCY TEACUP AND SAUCER with GREEN REGENCY TEACUP AND SAUCER.

```
> inspect(head(rules.UK,n=10))
     lhs                                  rhs                                    support confidence      lift count
[1]  {PINK REGENCY TEACUP AND SAUCER}  => {ROSES REGENCY TEACUP AND SAUCER}  0.02370818  0.7588556 17.775227   557
[2]  {PINK REGENCY TEACUP AND SAUCER}  => {GREEN REGENCY TEACUP AND SAUCER}  0.02489997  0.7970027 19.303899   585
[3]  {GREEN REGENCY TEACUP AND SAUCER} => {PINK REGENCY TEACUP AND SAUCER}   0.02489997  0.6030928 19.303899   585
[4]  {ALARM CLOCK BAKELIKE RED}        => {ALARM CLOCK BAKELIKE GREEN}      0.02421895  0.5933264 15.592405   569
[5]  {ALARM CLOCK BAKELIKE GREEN}      => {ALARM CLOCK BAKELIKE RED}        0.02421895  0.6364653 15.592405   569
[6]  {JUMBO BAG STRAWBERRY}            => {JUMBO BAG RED RETROSPOT}         0.02234613  0.6521739  7.746296   525
[7]  {PAPER CHAIN KIT VINTAGE CHRISTMAS} => {PAPER CHAIN KIT 50'S CHRISTMAS} 0.02298459  0.6658446 13.794845   540
[8]  {GARDENERS KNEELING PAD CUP OF TEA} => {GARDENERS KNEELING PAD KEEP CALM} 0.02315485 0.7186262 18.594056   544
[9]  {GARDENERS KNEELING PAD KEEP CALM}  => {GARDENERS KNEELING PAD CUP OF TEA} 0.02315485 0.5991189 18.594056   544
[10] {JUMBO  BAG BAROQUE BLACK WHITE}  => {JUMBO BAG RED RETROSPOT}         0.02460203  0.6262189  7.438011   578
>
```


Plot Zoom
10 Rules Only UK Customers
size: support (0.022 - 0.034)
color: lift (7.321 - 19.304)
PAPER CHAIN KIT 50'S CHRISTMAS
PAPER CHAIN KIT VINTAGE CHRISTMAS
ALARM CLOCK BAKELIKE RED
ALARM CLOCK BAKELIKE GREEN
PINK REGENCY TEACUP AND SAUCER
GREEN REGENCY TEACUP AND SAUCER
ROSES REGENCY TEACUP AND SAUCER
JUMBO BAG STRAWBERRY
JUMBO BAG BAROQUE BLACK WHITE
JUMBO BAG PINK POLKADOT
JUMBO BAG RED RETROSPOT
GARDENERS KNEELING PAD KEEP CALM
GARDENERS KNEELING PAD CUP OF TEA
JUMBO STORAGE BAG SUKI

SOLUTION-3.1

Example 1: - Survival on the Titanic

The sinking of the Titanic is a famous event, and new books are still being published about it. These data were originally collected by the British Board of Trade in their investigation of the sinking. Note that there is not complete agreement among primary sources as to the exact numbers on board, rescued, or lost. Very detailed data about the passengers is now available on the Internet, at sites "http://www.rmplc.co.uk/eduweb/sites/phind"

This data set provides information on the fate of passengers on the terrifying voyage of the ocean liner, summarized according to economic status (class), sex, age and survival.

With the help of this data set we can predict some rules and probability of people who survived and what is the percentage of people survived.

A 5-dimensional array resulting from cross-tabulating 1313 observations on 5 variables.

In this data set there are 4 parameters

1. Class: - Values = 1st class, 2nd Class, 3rd Class, Crew
2. Sex: - Value = Male or Female
3. Age: - Value = Child or Adult
4. Survived: - Value = Yes or No
5. Sex-Code: - Value = 1 (i.e. Female) or 0 (i.e. Male)

Dataset also contain First and Last name of the person on the ship.

With these 5 parameter we can predict some rules. Here Class also suggest which floor of the ship.

With the help of association rules we can get the support, confidence and lift and predict some rules about survival.

Code:- - Code is in file name Assignment1_Q3_Titanic.R.

Deductions: -

1. With the support of 53.9%, 708 number of male passengers did not survive.
2. 160 people in second class did not survive while 193 people in first class survived.
3. With Support of 23%, 308 number of females survived.
4. With a lift of 224%, 94 females in 2nd class survived on the other hand with a lift of 130% 147 men in 2nd class did not survived.
5. With a lift of 273%, 134 females in 1st class survived on the other hand with a lift of 143% 120 men in second class did not survived.
6. With a lift of 164%, 80 females in 3rd class survived on the other hand with a lift of 158% 120 female in 3rd class did not survived.
7. With a lift of 114%, 440 males in 3rd class did not survived
8. With a support of 0.29 females in first class survived with count 134 and with support of 0.20 females in second class survived.
9. With a support of 0.19 males in first class survived with count 59.

Mostly people die who belong to 3rd class and large number of men unable to survive as they belong to 3rd Class. Mostly women survived in each class but few unable to survive in 3rd Class.

Output: -

```
> inspect(rules)
     lhs                          rhs          support    confidence lift      count
[1]  {PClass=2nd}              => {Sex=male}   0.13099772 0.6164875  0.9522918 172
[2]  {PClass=2nd}              => {Survived=0} 0.12185834 0.5734767  0.8725086 160
[3]  {PClass=1st}              => {Survived=1} 0.14699162 0.5993789  1.7488544 193
[4]  {PClass=1st}              => {Sex=male}   0.13632902 0.5559006  0.8587030 179
[5]  {Survived=1}              => {Sex=female} 0.23457730 0.6844444  1.9451852 308
[6]  {Sex=female}              => {Survived=1} 0.23457730 0.6666667  1.9451852 308
[7]  {PClass=3rd}              => {Sex=male}   0.37928408 0.7014085  1.0834698 498
[8]  {Sex=male}                => {PClass=3rd} 0.37928408 0.5858824  1.0834698 498
[9]  {PClass=3rd}              => {Survived=0} 0.43564356 0.8056338  1.2257210 572
[10] {Survived=0}              => {PClass=3rd} 0.43564356 0.6628042  1.2257210 572
[11] {Sex=male}                => {Survived=0} 0.53922315 0.8329412  1.2672674 708
[12] {Survived=0}              => {Sex=male}   0.53922315 0.8203940  1.2672674 708
[13] {PClass=2nd,Survived=1}   => {Sex=female} 0.07159177 0.7899160  2.2449343  94
[14] {PClass=2nd,Sex=female}   => {Survived=1} 0.07159177 0.8785047  2.5632814  94
[15] {PClass=2nd,Sex=male}     => {Survived=0} 0.11195735 0.8546512  1.3002978 147
[16] {PClass=2nd,Survived=0}   => {Sex=male}   0.11195735 0.9187500  1.4191985 147
[17] {PClass=1st,Survived=1}   => {Sex=female} 0.10205636 0.6943005  1.9731961 134
[18] {PClass=1st,Sex=female}   => {Survived=1} 0.10205636 0.9370629  2.7341414 134
[19] {PClass=1st,Sex=male}     => {Survived=0} 0.09139375 0.6703911  1.0199577 120
[20] {PClass=1st,Survived=0}   => {Sex=male}   0.09139375 0.9302326  1.4369357 120
[21] {PClass=3rd,Survived=1}   => {Sex=female} 0.06092917 0.5797101  1.6475312  80
[22] {PClass=3rd,Sex=female}   => {Survived=0} 0.10053313 0.6226415  0.9473097 132
[23] {Sex=female,Survived=0}   => {PClass=3rd} 0.10053313 0.8571429  1.5851107 132
[24] {PClass=3rd,Sex=male}     => {Survived=0} 0.33511043 0.8835341  1.3442414 440
[25] {PClass=3rd,Survived=0}   => {Sex=male}   0.33511043 0.7692308  1.1882353 440
[26] {Sex=male,Survived=0}     => {PClass=3rd} 0.33511043 0.6214689  1.1492799 440
>
```

```
> inspect(rules.survived)
     lhs             rhs            support    confidence lift      count
[1]  {Age=18}     => {Sex=female} 0.02444444 1.0000000  1.4610390  11
[2]  {Age=45}     => {Sex=female} 0.02000000 0.8181818  1.1953955   9
[3]  {Age=19}     => {Sex=female} 0.02000000 0.7500000  1.0957792   9
[4]  {Age=22}     => {Sex=female} 0.02444444 0.8461538  1.2362637  11
[5]  {Age=36}     => {Sex=female} 0.02444444 0.6875000  1.0044643  11
[6]  {PClass=2nd} => {Sex=male}   0.05555556 0.2100840  0.6657593  25
[7]  {PClass=2nd} => {Sex=female} 0.20888889 0.7899160  1.1540980  94
[8]  {Sex=female} => {PClass=2nd} 0.20888889 0.3051948  1.1540980  94
[9]  {PClass=3rd} => {Sex=male}   0.12888889 0.4202899  1.3319045  58
[10] {Sex=male}   => {PClass=3rd} 0.12888889 0.4084507  1.3319045  58
[11] {PClass=3rd} => {Sex=female} 0.17777778 0.5797101  0.8469791  80
[12] {Sex=female} => {PClass=3rd} 0.17777778 0.2597403  0.8469791  80
[13] {Sex=male}   => {PClass=1st} 0.13111111 0.4154930  0.9687660  59
[14] {PClass=1st} => {Sex=male}   0.13111111 0.3056995  0.9687660  59
[15] {PClass=1st} => {Sex=female} 0.29777778 0.6943005  1.0144001 134
[16] {Sex=female} => {PClass=1st} 0.29777778 0.4350649  1.0144001 134
>
```

```
> inspect(rules.dead)
     lhs                      rhs          support    confidence lift      count
[1]  {Age=24}              => {Sex=male}   0.01506373 0.8125000  0.9903778  13
[2]  {Age=25}              => {Sex=male}   0.01506373 0.8125000  0.9903778  13
[3]  {Age=20}              => {PClass=3rd} 0.01622248 0.8235294  1.2424928  14
[4]  {Age=20}              => {Sex=male}   0.01622248 0.8235294  1.0038219  14
[5]  {Age=28}              => {Sex=male}   0.01738123 0.8823529  1.0755234  15
[6]  {Age=18}              => {Sex=male}   0.01506373 0.6842105  0.8340024  13
[7]  {Age=26}              => {PClass=3rd} 0.01622248 0.7368421  1.1117041  14
[8]  {Age=26}              => {Sex=male}   0.01969873 0.8947368  1.0906185  17
[9]  {Age=21}              => {PClass=3rd} 0.01622248 0.6666667  1.0058275  14
[10] {Age=21}              => {Sex=male}   0.01969873 0.8095238  0.9867501  17
[11] {Age=22}              => {PClass=3rd} 0.01738123 0.6818182  1.0286872  15
[12] {Age=22}              => {Sex=male}   0.02085747 0.8181818  0.9973035  18
[13] {Age=30}              => {Sex=male}   0.02085747 0.7826087  0.9539425  18
[14] {PClass=1st}          => {Sex=male}   0.13904983 0.9302326  1.1338852 120
[15] {Sex=male}            => {PClass=1st} 0.13904983 0.1694915  1.1338852 120
[16] {Sex=female}          => {PClass=3rd} 0.15295481 0.8571429  1.2932068 132
[17] {PClass=3rd}          => {Sex=female} 0.15295481 0.2307692  1.2932068 132
[18] {PClass=2nd}          => {Sex=male}   0.17033604 0.9187500  1.1198888 147
[19] {Sex=male}            => {PClass=2nd} 0.17033604 0.2076271  1.1198888 147
[20] {PClass=3rd}          => {Sex=male}   0.50984936 0.7692308  0.9376358 440
[21] {Sex=male}            => {PClass=3rd} 0.50984936 0.6214689  0.9376358 440
[22] {PClass=3rd,Age=26}   => {Sex=male}   0.01506373 0.9285714  1.1318604  13
[23] {Age=26,Sex=male}     => {PClass=3rd} 0.01506373 0.7647059  1.1537433  13
[24] {PClass=3rd,Age=22}   => {Sex=male}   0.01506373 0.8666667  1.0564030  13
[25] {Age=22,Sex=male}     => {PClass=3rd} 0.01506373 0.7222222  1.0896465  13
>
```

SOLUTION-3.2

Example 2: - School Science Survey Data

Description

Data frame has 1385 rows and 7 columns. The data are on attitudes to science, from a survey where there were results from 20 classes in private schools and 46 classes in public schools.

This data frame contains the following columns:

State: - Values as factor with levels ACT (Australian Capital Territory, NSW (New South Wales)
PrivPub: - Values as factor with levels Private and Public
School: - Values a factor, coded to identify the school
Class: - Values a factor, coded to identify the class
Sex: - Values a factor with levels f(female) and m (male)
Like: - Values a summary score based on two of the questions, on a scale From 1 (dislike) to 12 (like)

Class: - Values a factor with levels corresponding to each class

With the help of this dataset we are able to make some rules based on

1. how many public and Private schools are in ACT and NSW,
2. Class of Students like Science
3. Male or Female student like Science

And many more.
And with the help of this database we can try to improve structure of Science in particular school or for a parent if he wanted his child in School where Science is liked then he can use the rules.

Source: -
Francine Adams, Rosemary Martin and Murali Nayadu, Australian National University

Code: - Code is in file name Assignment1_Q3_Science.R

Deduction: -

1. With a support of 0.32 451 female student are in Class 1.
2. With a support of 0.33 female student in Public School with a count of 467.
3. With a highest support of 0.064, Male student of Class 1 in ACT public school like science half with a count of 89.
4. With a support of 0.061, female student of Class 1 in ACT public school like science half with a count of 85.
5. With a support of 0.031, male student of Class 1 in ACT public school like science with a count of 48.

Output 2: With minlen =2

```
> inspect(rules)
     lhs                                rhs                 support   confidence lift      count
[1]  {sex=f}                         => {class=1}           0.3256318 0.6526773  1.0260591 451
[2]  {class=1}                       => {sex=f}             0.3256318 0.5119183  1.0260591 451
[3]  {sex=f}                         => {PrivPub=public}    0.3371841 0.6758321  1.0032449 467
[4]  {PrivPub=public}                => {sex=f}             0.3371841 0.5005359  1.0032449 467
[5]  {sex=f}                         => {State=ACT}         0.4772563 0.9565847  0.9916690 661
[6]  {sex=m}                         => {class=1}           0.3090253 0.6184971  0.9723252 428
[7]  {sex=m}                         => {PrivPub=public}    0.3350181 0.6705202  0.9953596 464
[8]  {sex=m}                         => {State=ACT}         0.4859206 0.9725434  1.0082130 673
[9]  {State=ACT}                     => {sex=m}             0.4859206 0.5037425  1.0082130 673
[10] {class=1}                       => {PrivPub=public}    0.4555957 0.7162316  1.0632162 631
[11] {PrivPub=public}                => {class=1}           0.4555957 0.6763130  1.0632162 631
[12] {class=1}                       => {State=ACT}         0.6187726 0.9727582  1.0084357 857
[13] {State=ACT}                     => {class=1}           0.6187726 0.6414671  1.0084357 857
[14] {PrivPub=public}                => {State=ACT}         0.6736462 1.0000000  1.0366766 933
[15] {State=ACT}                     => {PrivPub=public}    0.6736462 0.6983533  1.0366766 933
[16] {class=1,sex=f}                 => {State=ACT}         0.3155235 0.9689579  1.0044960 437
[17] {State=ACT,sex=f}               => {class=1}           0.3155235 0.6611195  1.0393309 437
[18] {State=ACT,class=1}             => {sex=f}             0.3155235 0.5099183  1.0220505 437
[19] {PrivPub=public,sex=f}          => {State=ACT}         0.3371841 1.0000000  1.0366766 467
[20] {State=ACT,sex=f}               => {PrivPub=public}    0.3371841 0.7065053  1.0487780 467
[21] {State=ACT,PrivPub=public}      => {sex=f}             0.3371841 0.5005359  1.0032449 467
[22] {class=1,sex=m}                 => {State=ACT}         0.3018051 0.9766355  1.0124552 418
[23] {State=ACT,sex=m}               => {class=1}           0.3018051 0.6210996  0.9764164 418
[24] {PrivPub=public,sex=m}          => {State=ACT}         0.3350181 1.0000000  1.0366766 464
[25] {State=ACT,sex=m}               => {PrivPub=public}    0.3350181 0.6894502  1.0234604 464
[26] {PrivPub=public,class=1}        => {State=ACT}         0.4555957 1.0000000  1.0366766 631
[27] {State=ACT,class=1}             => {PrivPub=public}    0.4555957 0.7362894  1.0929912 631
[28] {State=ACT,PrivPub=public}      => {class=1}           0.4555957 0.6763130  1.0632162 631
>
```

Output 2: With minlen =5

```
> inspect(rules)
     lhs                                      rhs                 support    confidence lift      count
[1]  {PrivPub=public,class=1,sex=m,like=8}  => {State=ACT}         0.03465704 1.0000000  1.0366766 48
[2]  {State=ACT,class=1,sex=m,like=8}       => {PrivPub=public}    0.03465704 0.9056604  1.3444155 48
[3]  {State=ACT,PrivPub=public,sex=m,like=8} => {class=1}          0.03465704 0.7058824  1.1097015 48
[4]  {State=ACT,PrivPub=public,class=1,like=8} => {sex=m}          0.03465704 0.6075949  1.2160679 48
[5]  {PrivPub=public,class=1,sex=f,like=3}  => {State=ACT}         0.03104693 1.0000000  1.0366766 43
[6]  {State=ACT,class=1,sex=f,like=3}       => {PrivPub=public}    0.03104693 0.7543860  1.1198548 43
[7]  {State=ACT,PrivPub=public,sex=f,like=3} => {class=1}          0.03104693 0.6935484  1.0903116 43
[8]  {State=ACT,PrivPub=public,class=1,like=3} => {sex=f}          0.03104693 0.6056338  1.2138970 43
[9]  {PrivPub=public,class=1,sex=f,like=4}  => {State=ACT}         0.03465704 1.0000000  1.0366766 48
[10] {State=ACT,class=1,sex=f,like=4}       => {PrivPub=public}    0.03465704 0.5714286  0.8482621 48
[11] {State=ACT,PrivPub=public,sex=f,like=4} => {class=1}          0.03465704 0.5714286  0.8983298 48
[12] {State=ACT,PrivPub=public,class=1,like=4} => {sex=f}          0.03465704 0.5000000  1.0021708 48
[13] {PrivPub=public,class=1,sex=m,like=4}  => {State=ACT}         0.03465704 1.0000000  1.0366766 48
[14] {State=ACT,class=1,sex=m,like=4}       => {PrivPub=public}    0.03465704 0.7384615  1.0962157 48
[15] {State=ACT,PrivPub=public,sex=m,like=4} => {class=1}          0.03465704 0.7164179  1.1262643 48
[16] {State=ACT,PrivPub=public,class=1,like=4} => {sex=m}          0.03465704 0.5000000  1.0007225 48
[17] {PrivPub=public,class=2,sex=f,like=6}  => {State=ACT}         0.03393502 1.0000000  1.0366766 47
[18] {State=ACT,class=2,sex=f,like=6}       => {PrivPub=public}    0.03393502 0.7966102  1.1825349 47
[19] {State=ACT,PrivPub=public,class=2,like=6} => {sex=f}          0.03393502 0.5949367  1.1924564 47
[20] {PrivPub=public,class=1,sex=f,like=6}  => {State=ACT}         0.06137184 1.0000000  1.0366766 85
[21] {State=ACT,class=1,sex=f,like=6}       => {PrivPub=public}    0.06137184 0.7391304  1.0972086 85
[22] {State=ACT,PrivPub=public,sex=f,like=6} => {class=1}          0.06137184 0.6204380  0.9753764 85
[23] {State=ACT,PrivPub=public,class=1,like=6} => {sex=f}          0.06137184 0.4885057  0.9791324 85
[24] {PrivPub=public,class=1,sex=m,like=6}  => {State=ACT}         0.06425993 1.0000000  1.0366766 89
[25] {State=ACT,class=1,sex=m,like=6}       => {PrivPub=public}    0.06425993 0.8240741  1.2233040 89
[26] {State=ACT,PrivPub=public,sex=m,like=6} => {class=1}          0.06425993 0.6691729  1.0519915 89
[27] {State=ACT,PrivPub=public,class=1,like=6} => {sex=m}          0.06425993 0.5114943  1.0237277 89
>
```

APPENDIX-1: Code-Cars

```
############################################################
#
#               Question 1
############################################################
 #install rggobi package

## Setting the Path of Directory

drive="E:"
path.upto <- paste("STAT5703-HEMANT-101062246-Assignment1", sep="/" )
code.dir <- paste(drive, path.upto,"Code", sep="/")
data.dir <- paste(drive, path.upto,"Data","cars.OKC", sep="/")
work.dir <- paste(drive, path.upto,"Work", sep="/")
setwd(work.dir)

##Reading the CSV format of Cars Data

Cars.dat <- read.table(data.dir, header=TRUE)

## Factoring the Car Data ORigin Column
Cars.dat$Origin = as.factor(Cars.dat$Origin)

Cars.col <- Cars.dat$Origin


## Cleaning the data for Column Year
for (j in 1:length(Cars.dat$Year)){
  if(Cars.dat[j,6] > 100){
    Cars.dat[j,6] = Cars.dat[j,6]-100
  }
}


## Ggobi PLot : Scatterplot and Parallel COordinate gives overall
picture of
###              data but its hard to scale it.

library(rggobi)
g <- ggobi(Cars.dat)

display(g[1], "Scatterplot Matrix")

display(g[1], "Parallel Coordinates Display")

## Plotting the Cars.dat with respect to origin Pairs Plot as it also
contain scale for each parameter

pairs(Cars.dat, col=Cars.col)
```

```
#Categorize data by origin

Cars.USA=Cars.dat[Cars.dat$Origin==1,-7]
Cars.JPN=Cars.dat[Cars.dat$Origin==2,-7]
Cars.EUR=Cars.dat[Cars.dat$Origin==3,-7]


par(mfrow=c(2,3))

##Setting X axis name
x.names=c("USA","JPN","EUR")

## PLotting the Box Plot of All 6 parameter based on Origin and
keeping na.rm =TRUE to avoid NA values
for(i in 1:6){
  plot.title=colnames(Cars.dat)[i]

boxplot(Cars.USA[,i],Cars.JPN[,i],Cars.EUR[,i],main=plot.title,xaxt="n
", col= heat.colors(3), na.rm=TRUE)
  axis(1,at = 1:3,labels=x.names)
}

### CO-PLOT :-To get the concentation of points with yearly for each
origin

coplot(Cylinders ~ Year| Origin, data = Cars.dat, col=
Cars.dat$Origin)
coplot(MPG ~ Year| Origin, data= Cars.dat,col= Cars.dat$Origin)
coplot(Horsepower ~ Year| Origin, data= Cars.dat,col= Cars.dat$Origin)
coplot(Weight ~ Year| Origin, data= Cars.dat,col= Cars.dat$Origin)
coplot(Acceleration ~ Year | Origin, data= Cars.dat,col=
Cars.dat$Origin)

#####We use Box plot to get the Aveage value of the data per year
#####Coplot shows the distribition of data point but clearly didnot
mention
#####average avlue concentartion . This helps in deduction Accurately.

#Setting Data Of USA based on Year
Cars.USA.70=Cars.USA[Cars.USA$Year==70,-6]
Cars.USA.71=Cars.USA[Cars.USA$Year==71,-6]
Cars.USA.72=Cars.USA[Cars.USA$Year==72,-6]
Cars.USA.73=Cars.USA[Cars.USA$Year==73,-6]
Cars.USA.74=Cars.USA[Cars.USA$Year==74,-6]
Cars.USA.75=Cars.USA[Cars.USA$Year==75,-6]
Cars.USA.76=Cars.USA[Cars.USA$Year==76,-6]
Cars.USA.77=Cars.USA[Cars.USA$Year==77,-6]
Cars.USA.78=Cars.USA[Cars.USA$Year==78,-6]
Cars.USA.79=Cars.USA[Cars.USA$Year==79,-6]
Cars.USA.80=Cars.USA[Cars.USA$Year==80,-6]
Cars.USA.81=Cars.USA[Cars.USA$Year==81,-6]
```

```
Cars.USA.82=Cars.USA[Cars.USA$Year==82,-6]

##BoxPlot of Cars.USA data based on Yearly variation and keeping na.rm
=TRUE to avoid NA values
par(mfrow=c(2,3))
x.names=c("1970","1971","1972","1973","1974","1975","1976","1977","197
8","1979","1980","1981","1982")
for(i in 1:5){

  plot.title <- paste(colnames(Cars.USA)[i], "USA ",sep = '_')

boxplot(Cars.USA.70[,i],Cars.USA.71[,i],Cars.USA.72[,i],Cars.USA.73[,i
],Cars.USA.74[,i],Cars.USA.75[,i],Cars.USA.76[,i],Cars.USA.77[,i],Cars
.USA.78[,i],Cars.USA.79[,i],Cars.USA.80[,i],Cars.USA.81[,i],Cars.USA.8
2[,i],main=plot.title, xaxt="n", col= heat.colors(13),na.rm=TRUE)
  axis(1,at = 1:13,labels=x.names)
}

#Setting Data Of Japan based on Year

Cars.JPN.70=Cars.JPN[Cars.JPN$Year==70,-6]
Cars.JPN.71=Cars.JPN[Cars.JPN$Year==71,-6]
Cars.JPN.72=Cars.JPN[Cars.JPN$Year==72,-6]
Cars.JPN.73=Cars.JPN[Cars.JPN$Year==73,-6]
Cars.JPN.74=Cars.JPN[Cars.JPN$Year==74,-6]
Cars.JPN.75=Cars.JPN[Cars.JPN$Year==75,-6]
Cars.JPN.76=Cars.JPN[Cars.JPN$Year==76,-6]
Cars.JPN.77=Cars.JPN[Cars.JPN$Year==77,-6]
Cars.JPN.78=Cars.JPN[Cars.JPN$Year==78,-6]
Cars.JPN.79=Cars.JPN[Cars.JPN$Year==79,-6]
Cars.JPN.80=Cars.JPN[Cars.JPN$Year==80,-6]
Cars.JPN.81=Cars.JPN[Cars.JPN$Year==81,-6]
Cars.JPN.82=Cars.JPN[Cars.JPN$Year==82,-6]

##BoxPlot of Cars.JPN data based on Yearly variation and keeping na.rm
=TRUE to avoid NA values
par(mfrow=c(2,3))
x.names=c("1970","1971","1972","1973","1974","1975","1976","1977","197
8","1979","1980","1981","1982")
for(i in 1:5){
  plot.title <- paste(colnames(Cars.USA)[i], "JPN ",sep = '_')

boxplot(Cars.JPN.70[,i],Cars.JPN.71[,i],Cars.JPN.72[,i],Cars.JPN.73[,i
],Cars.JPN.74[,i],Cars.JPN.75[,i],Cars.JPN.76[,i],Cars.JPN.77[,i],Cars
.JPN.78[,i],Cars.JPN.79[,i],Cars.JPN.80[,i],Cars.JPN.81[,i],Cars.JPN.8
2[,i],main=plot.title,xaxt="n", col= heat.colors(13), na.rm=TRUE)
  axis(1,at = 1:13,labels=x.names)
}

#Setting Data Of Europe based on Year and keeping na.rm =TRUE to avoid
NA values
```

```
Cars.EUR.70=Cars.EUR[Cars.EUR$Year==70,-6]
Cars.EUR.71=Cars.EUR[Cars.EUR$Year==71,-6]
Cars.EUR.72=Cars.EUR[Cars.EUR$Year==72,-6]
Cars.EUR.73=Cars.EUR[Cars.EUR$Year==73,-6]
Cars.EUR.74=Cars.EUR[Cars.EUR$Year==74,-6]
Cars.EUR.75=Cars.EUR[Cars.EUR$Year==75,-6]
Cars.EUR.76=Cars.EUR[Cars.EUR$Year==76,-6]
Cars.EUR.77=Cars.EUR[Cars.EUR$Year==77,-6]
Cars.EUR.78=Cars.EUR[Cars.EUR$Year==78,-6]
Cars.EUR.79=Cars.EUR[Cars.EUR$Year==79,-6]
Cars.EUR.80=Cars.EUR[Cars.EUR$Year==80,-6]
Cars.EUR.81=Cars.EUR[Cars.EUR$Year==81,-6]
Cars.EUR.82=Cars.EUR[Cars.EUR$Year==82,-6]

##BoxPlot of Cars.EUR data based on Yearly variation
par(mfrow=c(2,3))
x.names=c("1970","1971","1972","1973","1974","1975","1976","1977","197
8","1979","1980","1981","1982")
for(i in 1:5){
  plot.title <- paste(colnames(Cars.USA)[i], "EUR",sep = '_')

boxplot(Cars.EUR.70[,i],Cars.EUR.71[,i],Cars.EUR.72[,i],Cars.EUR.73[,i
],Cars.EUR.74[,i],Cars.EUR.75[,i],Cars.EUR.76[,i],Cars.EUR.77[,i],Cars
.EUR.78[,i],Cars.EUR.79[,i],Cars.EUR.80[,i],Cars.EUR.81[,i],Cars.EUR.8
2[,i],main=plot.title,xaxt="n", col= heat.colors(13),na.rm = TRUE)
  axis(1,at = 1:13,labels=x.names)
}
```

APPENDIX-2: Code-Online Retail

```
############################################################

#                    Question 2

############################################################
## Install readxl package
## Install arulesViz package
## Install arules package


## Setting up the directory path
drive="E:"
path.upto <- paste("STAT5703-HEMANT-101062246-Assignment1", sep="/" )
code.dir <- paste(drive, path.upto,"Code", sep="/")
data.dir <- paste(drive, path.upto,"Data","Online Retail.xlsx",
sep="/")
work.dir <- paste(drive, path.upto,"Work", sep="/")
```

```
setwd(work.dir)

## Installing the library
library("readxl")

## reading the Data file
online_data <- read_excel(data.dir)

##Spliting the data on behalf of Description and Invoice No.

original_data <- split(online_data$Description,online_data$InvoiceNo)
head(original_data)

##Setting Up the arules
library(arules)
library(arulesViz)

##Applying Ariori Rules
rules <- apriori(original_data, parameter=list(support=0.01,
confidence=0.1,minlen=2))

###Plotting Value of Support and Confidence to get new value to get
best 10 rules
plot(rules, measure=c("support", "confidence"), shading="lift",main =
"All Rules Original_Dataset")

##Getting Rules with new value of Support and Confidence from Graph
rules <- apriori(original_data, parameter=list(support=0.02131,
confidence=0.5893))

summary(rules)
inspect(head(rules,n=10))

######Plotting rules for minmimum 2 items
plot(rules, method="graph", control=list(type="Description"),main =
"10 Rules with 2 Min Items")

## Applying arules for relation between 3 items on original data
rules.3.items <- apriori(original_data, parameter=list(support=0.015,
confidence=0.5, minlen=3))
inspect(rules.3.items)

######Plotting rules for minmimum 3 items
plot(rules.3.items, method="graph",
control=list(type="Description"),main = "10 Rules with 3 Min Items")



## Applying arules for relation between 4 items on original data
rules.4.items <- apriori(original_data, parameter=list(support=0.01,
confidence=0.1, minlen=4))
```

```
inspect(rules.4.items)

######Plotting rules for minmimum 3 items
plot(rules.4.items, method="graph",
control=list(type="Description"),main = "10 Rules with 4 Min Items")



################## Removing Cancelled Transactions############


new_data <- (subset(online_data, Quantity > 0))

data_success_trnx <- split(new_data$Description,new_data$InvoiceNo)

library(arules)
## Applying arules
rules_success_trnx <- apriori(data_success_trnx,
parameter=list(support=0.01, confidence=0.1))

###Plotting Value of Support and Confidence to get new value to get
best 10 rules
plot(rules_success_trnx, measure=c("support", "confidence"),
shading="lift", main = "All Rules Successful Transaction")

##Getting Rules with new value of Support and Confidence from Graph
rules_success_trnx <- apriori(data_success_trnx,
parameter=list(support=0.025, confidence=0.66))

summary(rules_success_trnx)
inspect(head(rules_success_trnx, n=10))

######Plotting rules
plot(rules_success_trnx, method="graph",
control=list(type="Description"),main = "10 Rules Successful
Transaction")



####################SUBSET of DATA WITH UK Customers############


table(online_data$Country == "United Kingdom")
## As as 495478 entries are from United Kingdom
## Therefore we take the subset of UK data to apply arules

OR.UK<- (subset(online_data, Country == "United Kingdom"))
data.UK <- split(OR.UK$Description,OR.UK$InvoiceNo)
head(data.UK)

#######Applying Arules##################
```

```
rules.UK <- apriori(data.UK, parameter=list(support=0.01,
confidence=0.1))

###Plotting Value of Support and Confidence to get new value to get
best 10 rules
plot(rules.UK, measure=c("support", "confidence"), shading="lift",
main = "All Rules Only UK Customers")

##Getting Rules with new value of Support and Confidence from Graph
rules.UK <- apriori(data.UK, parameter=list(support=0.02131,
confidence=0.5893, minlen = 2))

summary(rules.UK)
inspect(head(rules.UK,n=10))

######Plotting rules
plot(rules.UK, method="graph", control=list(type="Description"),main =
"10 Rules Only UK Customers")
```

APPENDIX-3: Code-Titanic

```
#################################################################3

# Question -3 : Bounus Implementation

#################################################################

#install arules package

##Setting Directory Path

drive="E:"
path.upto <- paste("STAT5703-HEMANT-101062246-Assignment1", sep="/" )
code.dir <- paste(drive, path.upto,"Code", sep="/")
data.dir <- paste(drive, path.upto,"Data", sep="/")
work.dir <- paste(drive, path.upto,"Work", sep="/")
setwd(work.dir)

## Reading Data from File
titanic.file <- paste(data.dir,"Titanic.csv", sep="/")
titanic.dat = read.csv(titanic.file,header=TRUE)

## Factorizing the data
titanic.dat$Index <- as.factor(titanic.dat$Index)
titanic.dat$SexCode <- as.factor(titanic.dat$SexCode)
titanic.dat$Survived <- as.factor(titanic.dat$Survived)
```

```
## Subsetting the data on the basis of people Survived and Not
Survived

titanic.dat.survived = (subset(titanic.dat, Survived == 1))
titanic.dat.dead = (subset(titanic.dat, Survived == 0))

## Removing Column Index, LastName and FirstName,SexCode, Survival
titanic.dat.survived <- titanic.dat.survived[,c(4,5,6)]
titanic.dat.dead = titanic.dat.dead[,c(4,5,6)]

## Removing Column Index, LastName and FirstName,SexCode
titanic.dat <- titanic.dat[,c(4,5,6,7)]

## Rules implementation on whole dataset
library(arules)
rules <- apriori(titanic.dat, parameter=list(support=0.05,
confidence=0.5, minlen=2))
summary(rules)
inspect(rules)

## Rules implementation on dataset of survived people
rules.survived <- apriori(titanic.dat.survived,
parameter=list(support=0.02, confidence=0.2, minlen=2))
summary(rules.survived)
inspect(rules.survived)

## Rules implementation on dataset of Dead people
rules.dead <- apriori(titanic.dat.dead, parameter=list(support=0.015,
confidence=0.1, minlen=2))
summary(rules.dead)
inspect(rules.dead)
```

APPENDIX 4: Code-Science Survey

```
############################################################3

# Question -3 : Bounus Implementation

#################################################################
##install arules package

##Setting Directory Path

drive="E:"
path.upto <- paste("STAT5703-HEMANT-101062246-Assignment1", sep="/" )
code.dir <- paste(drive, path.upto,"Code", sep="/")
```

```
data.dir <- paste(drive, path.upto,"Data", sep="/")
work.dir <- paste(drive, path.upto,"Work", sep="/")
setwd(work.dir)

## Reading Data from File
Science.file <- paste(data.dir,"Australian_School_Science_Survey.csv",
sep="/")
Science.dat = read.csv(Science.file,header=TRUE)

str(Science.dat)
Science.dat <- Science.dat[,c(2,3,4,5,6,7)]

## Factorizing the data
Science.dat$school <- as.factor(Science.dat$school)
Science.dat$class <- as.factor(Science.dat$class)
Science.dat$like <- as.factor(Science.dat$like)


## Rules implementation on whole dataset
library(arules)
## Minimum Variable must be 2
rules <- apriori(Science.dat, parameter=list(support=0.3,
confidence=0.5, minlen=2))
summary(rules)
inspect(rules)

## Minimum Variable must be 5
rules.min5 <- apriori(Science.dat, parameter=list(support=0.03,
confidence=0.4, minlen=5))
summary(rules.min5)
inspect(rules.min5)
```