【數據科學演算法】

期末報告

報告者：李健立

學號：407190080

使用資料：Cars93

研究主題：

什麼樣的特性 or 性能會使車子價格昂貴

此研究中的變數：

- Type(種類)：Compact , Large , Midsize , Small , Sporty , Van
- Price(平均價格)：low , medium , high , **supreme(研究重點)**
- Horsepower(馬力)：low , medium , high
- RPM(最大馬力時的每分鐘轉速)：low , medium , high
- Man.trans.avail(是否有手動變速箱)：Yes , No
- Fuel.tank.capacity(油箱容量)：low , medium , high
- Weight(重量)：light , normal , heavy
- Origin(來源)：USA , non-USA

執行：car = Cars93[, c("Type" , "Price" , "Horsepower" , "RPM" ,

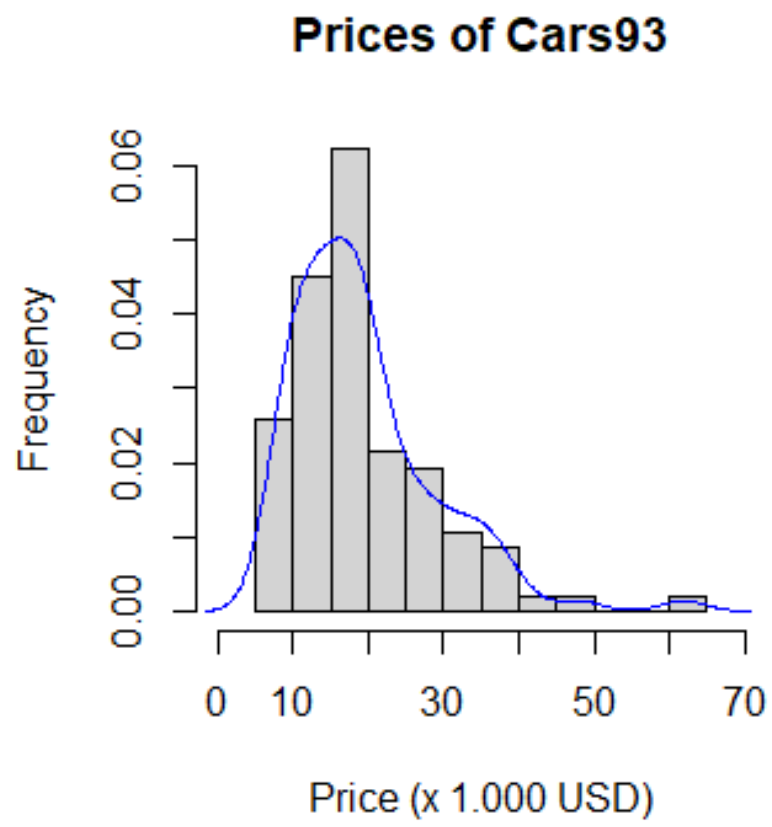
"Man.trans.avail" , "Fuel.tank.capacity" , "Weight" , "Origin")]

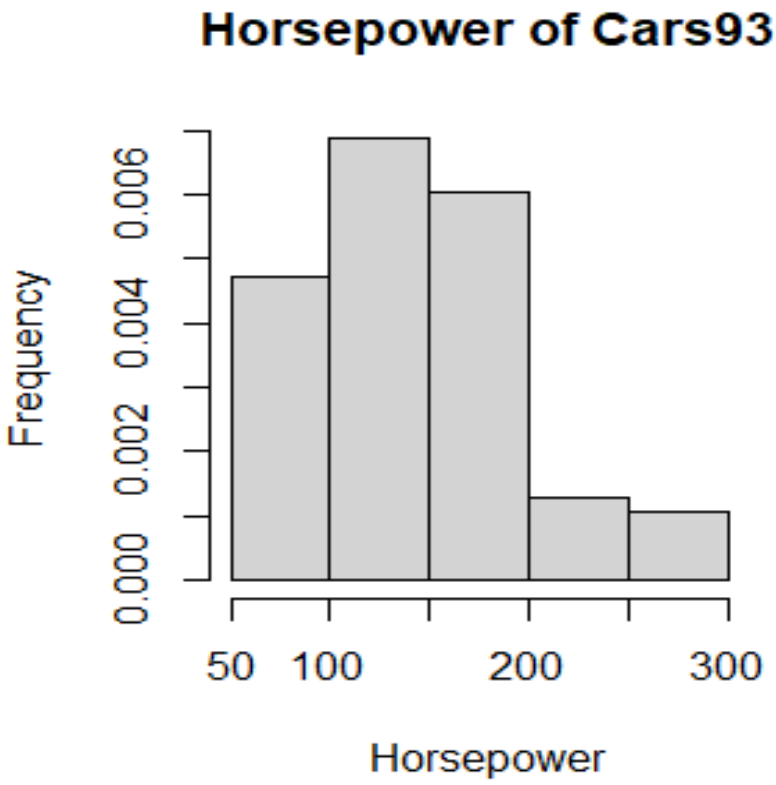
(在 Cars93 的眾多變數中取出我要調查的資料)

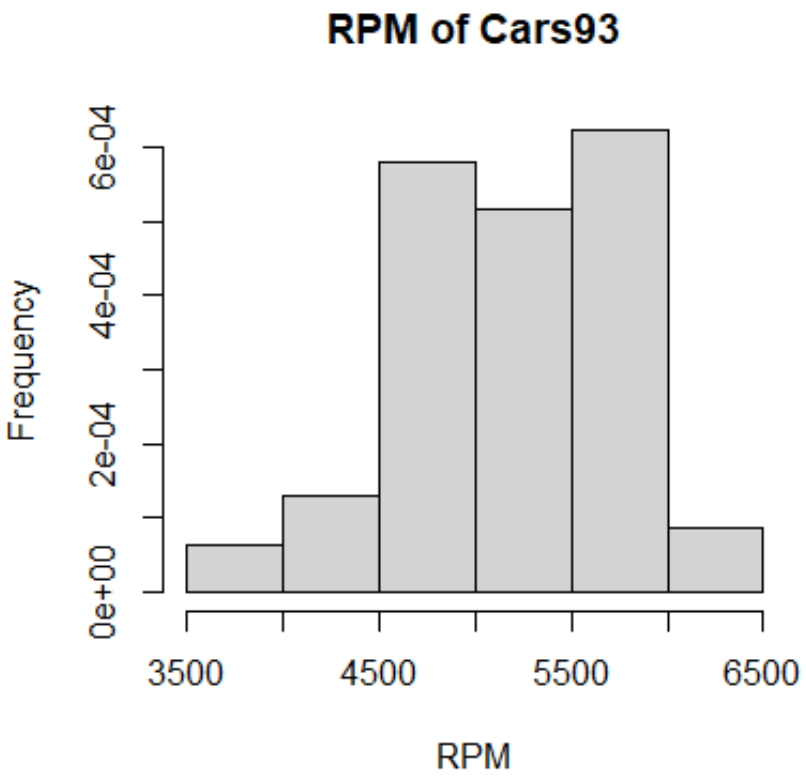
並把數據轉換成變數 (low , medium , high 等)

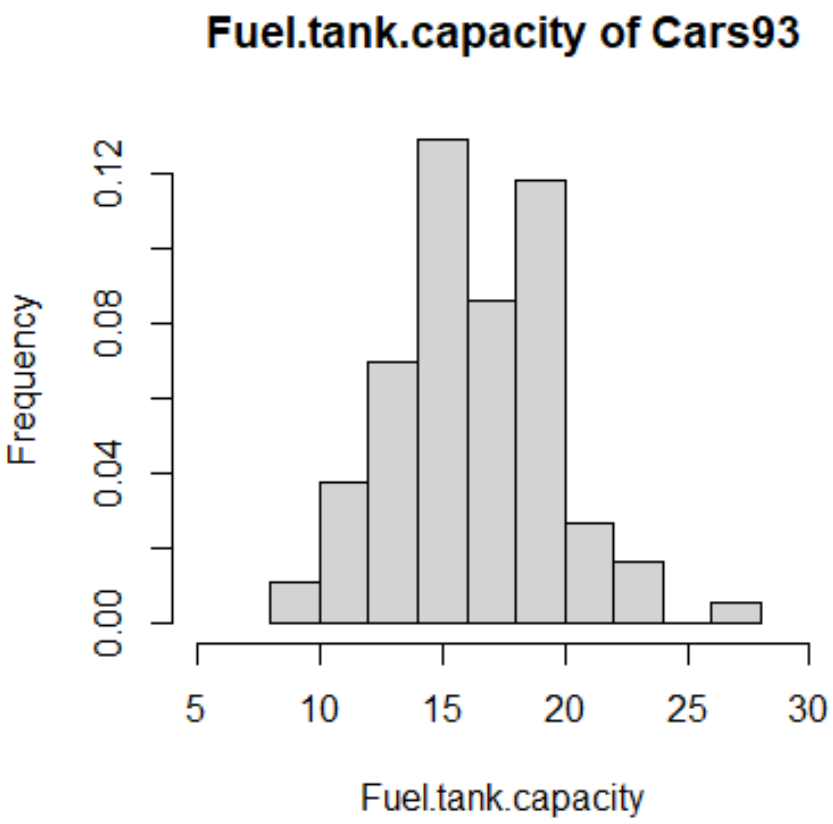
(↑此處執行 R)

以下展示一些變數的 histogram：


Prices of Cars93
Frequency
0.00
0.02
0.04
0.06
0
10
30
50
70
Price (x 1.000 USD)


Horsepower of Cars93
Frequency
0.000
0.002
0.004
0.006
50
100
200
300
Horsepower


RPM of Cars93
Frequency
0e+00
2e-04
4e-04
6e-04
3500
4500
5500
6500
RPM


Fuel.tank.capacity of Cars93
Frequency
0.00
0.04
0.08
0.12
5
10
15
20
25
30
Fuel.tank.capacity

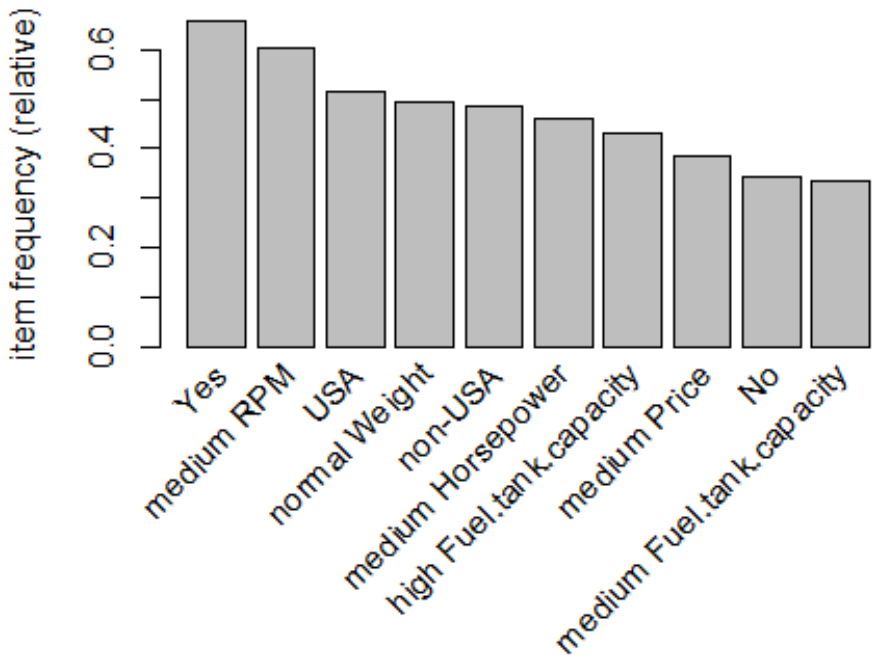
這邊把 data(car) 轉成 transactions 以便等等訓練模型：

```
#####把檔案寫入CSV#####
write.csv(car, "car.csv", row.names=F)
library(knitr)
kable(car[1:5,])

#####輸出CSV文件#####
install.packages('arules')
library(arules)
car_tran <- read.transactions("car.csv", sep = ",", skip = 1, rm.duplicates=TRUE)
summary(car_tran) ##transactions
```

執行：itemFrequencyPlot(car_tran, topN=10) 來看前 10 名的出現頻率

>>>



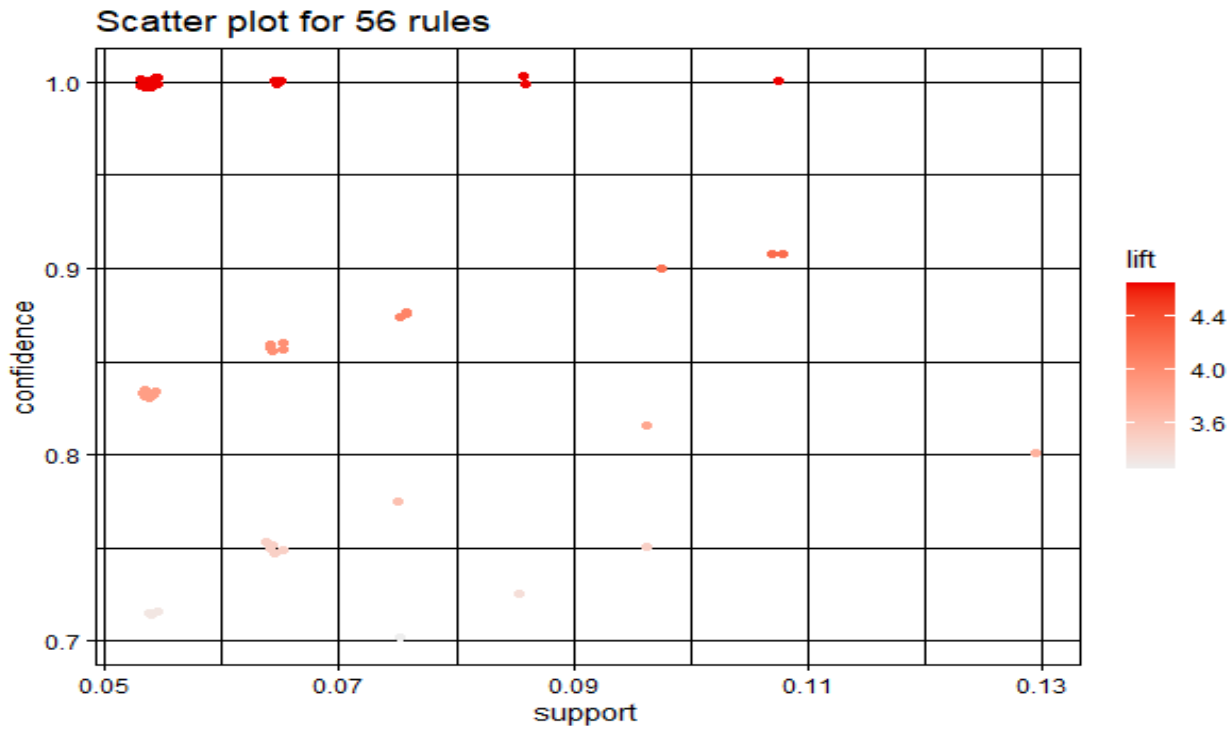
可發現大多的車都有手動變速箱(Yes),且差不多有一半是 non-USA

接著可以開始訓練模型了：

```
#####訓練模型#####
car_tran_rule <- apriori(car_tran, parameter=list(support = 0.1, confidence = 0.8, minlen = 2))
car_tran_rules <- apriori(car_tran,
                  parameter = list(minlen = 3, support = 0.05, confidence = 0.7),
                  appearance = list(rhs="supreme Price", default="lhs"), control = list(verbose=F))
##支援度support，亦即X和Y同時出現的次數 ÷ 所有交易數 ，指同時擁有lhr & rhs特性的車中占全部車的比例
##信賴度confidence，亦即X和Y同時出現的次數 ÷ X出現的次數 ，指擁有lhs特性的車中也有rhs特性的比例
```

此處我訓練了兩個模型，一個是初始設定的 s = 0.1 , c = 0.8，但多加了 minlen = 2 來刪除所有少於兩個項目的規則，這個模型訓練的目的是要看原始資料的規則，看能不能發現另一項顯性的研究(共 595 rules)；**另一個模型是本次主要的研究內容，方法是固定 right hand sides 的變數為”supreme Price” ，目的是要看 left hand sides 有什麼特性 or 性能會使車子價格變昂貴，並降低 confidence & support 來使資料得到適合的 rules(共 56 rules)。**

下圖為第二個模型的 confidence & support 的分布圖：



接著用”lift”排序一下 rules 的順序：

```
#####排序特定的值#####
inspect(sort(car_tran_rule, by="lift")[1:15])
inspect(sort(car_tran_rules, by="lift")[1:15])
```

lift = confidence/support(Y) (這裡的 support 是 Y 出現次數 ÷ 所有交易數，和上面的不同)

此處用”lift”來排序是因為 lift 在此研究中可解讀成是：在擁有 lhs 特性的情況下，價格是”supreme Price”的**可能性有多大**,所以 lift 越高越好。

執行：inspect(sort(car_tran_rule, by="lift")[1:15])

>>>

```
     lhs                                                rhs              support   confidence coverage  lift     count
[1]  {high Fuel.tank.capacity, high Horsepower, Midsize} => {supreme Price}  0.1075269 1.0000000  0.1075269 4.650000 10
[2]  {high Horsepower, Midsize}                         => {supreme Price}  0.1075269 0.9090909  0.1182796 4.227273 10
[3]  {high Horsepower, non-USA}                         => {supreme Price}  0.1075269 0.9090909  0.1182796 4.227273 10
[4]  {low Fuel.tank.capacity, medium RPM}               => {low Horsepower} 0.1290323 1.0000000  0.1290323 4.043478 12
[5]  {low Horsepower, non-USA}                          => {light Weight}   0.1290323 1.0000000  0.1290323 4.043478 12
[6]  {low Fuel.tank.capacity, medium RPM, Small}        => {low Horsepower} 0.1075269 1.0000000  0.1075269 4.043478 10
[7]  {low Horsepower, non-USA, Small}                   => {light Weight}   0.1075269 1.0000000  0.1075269 4.043478 10
[8]  {low Fuel.tank.capacity, low Horsepower, non-USA}  => {light Weight}   0.1182796 1.0000000  0.1182796 4.043478 11
[9]  {light Weight, low Fuel.tank.capacity, medium RPM} => {low Horsepower} 0.1182796 1.0000000  0.1182796 4.043478 11
[10] {low Fuel.tank.capacity, low Price, medium RPM}    => {low Horsepower} 0.1290323 1.0000000  0.1290323 4.043478 12
[11] {low Fuel.tank.capacity, medium RPM, Yes}          => {low Horsepower} 0.1290323 1.0000000  0.1290323 4.043478 12
[12] {low Horsepower, low Price, non-USA}               => {light Weight}   0.1290323 1.0000000  0.1290323 4.043478 12
[13] {low Horsepower, non-USA, Yes}                     => {light Weight}   0.1290323 1.0000000  0.1290323 4.043478 12
[14] {low Price, medium RPM, non-USA}                   => {light Weight}   0.1075269 1.0000000  0.1075269 4.043478 10
[15] {low Fuel.tank.capacity, low Price, medium RPM, Small} => {low Horsepower} 0.1075269 1.0000000  0.1075269 4.043478 10
```

先不看前三列(因為前三列的 rhs 為本次研究問題)，列 4 展現了當擁有{低油箱容量、中等 RPM}時，很大的可能性會是{低馬力}；列 5 則是當擁有{低馬力、non-USA}時，很大的可能性會是{輕重量}。

執行：inspect(sort(car_tran_rules, by="lift")[1:15])

>>>

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{high Fuel.tank.capacity, high RPM, Midsize}	=>	{supreme Price}	0.05376344	1	0.05376344	4.65	5
[2]	{heavy Weight, high Horsepower, Midsize}	=>	{supreme Price}	0.06451613	1	0.06451613	4.65	6
[3]	{heavy Weight, Midsize, non-USA}	=>	{supreme Price}	0.05376344	1	0.05376344	4.65	5
[4]	{high Horsepower, Midsize, No}	=>	{supreme Price}	0.05376344	1	0.05376344	4.65	5
[5]	{high Fuel.tank.capacity, high Horsepower, Midsize}	=>	{supreme Price}	0.10752688	1	0.10752688	4.65	10
[6]	{high Horsepower, Midsize, non-USA}	=>	{supreme Price}	0.08602151	1	0.08602151	4.65	8
[7]	{heavy Weight, high Horsepower, non-USA}	=>	{supreme Price}	0.05376344	1	0.05376344	4.65	5
[8]	{high Horsepower, medium RPM, non-USA}	=>	{supreme Price}	0.06451613	1	0.06451613	4.65	6
[9]	{heavy Weight, high Fuel.tank.capacity, high Horsepower, Midsize}	=>	{supreme Price}	0.06451613	1	0.06451613	4.65	6
[10]	{heavy Weight, high Horsepower, Midsize, non-USA}	=>	{supreme Price}	0.05376344	1	0.05376344	4.65	5
[11]	{heavy Weight, high Fuel.tank.capacity, Midsize, non-USA}	=>	{supreme Price}	0.05376344	1	0.05376344	4.65	5
[12]	{high Fuel.tank.capacity, high Horsepower, Midsize, No}	=>	{supreme Price}	0.05376344	1	0.05376344	4.65	5
[13]	{high Fuel.tank.capacity, high Horsepower, Midsize, non-USA}	=>	{supreme Price}	0.08602151	1	0.08602151	4.65	8
[14]	{high Fuel.tank.capacity, high Horsepower, medium RPM, Midsize}	=>	{supreme Price}	0.06451613	1	0.06451613	4.65	6
[15]	{high Fuel.tank.capacity, high Horsepower, Midsize, Yes}	=>	{supreme Price}	0.05376344	1	0.05376344	4.65	5

上圖為固定 rhs 為{supreme Price}後的規則，注意畫螢光筆的列，列 12 比列 5 多了一個{No}的特性；列 13 則是多了{non-USA}，但他們的 lift 值卻是一樣的，這代表多了{No}和{non-USA}後並沒有增加價格是{supreme Price}的可能性，因此列 12 與列 13 的規則在此研究中是多餘的。

所以我們把多餘的規則篩選掉，只留下重點來看就好：

```
#####刪除多餘的規則#####
rules_lift <- sort(car_tran_rules, by = 'lift')
rules_pruned <- rules_lift[!is.redundant(rules_lift, measure="lift")]
inspect(rules_pruned) ##從原本56個rules裁減到剩24個rules
```

is.redundant 的意思就是上面所講的，當 lhs 多了一些特性，但”lift”並沒有增加，就會判定是多餘的。

最後來看一下篩選過後的規則(用 lift 排序)：

執行：inspect(rules_pruned)

>>>

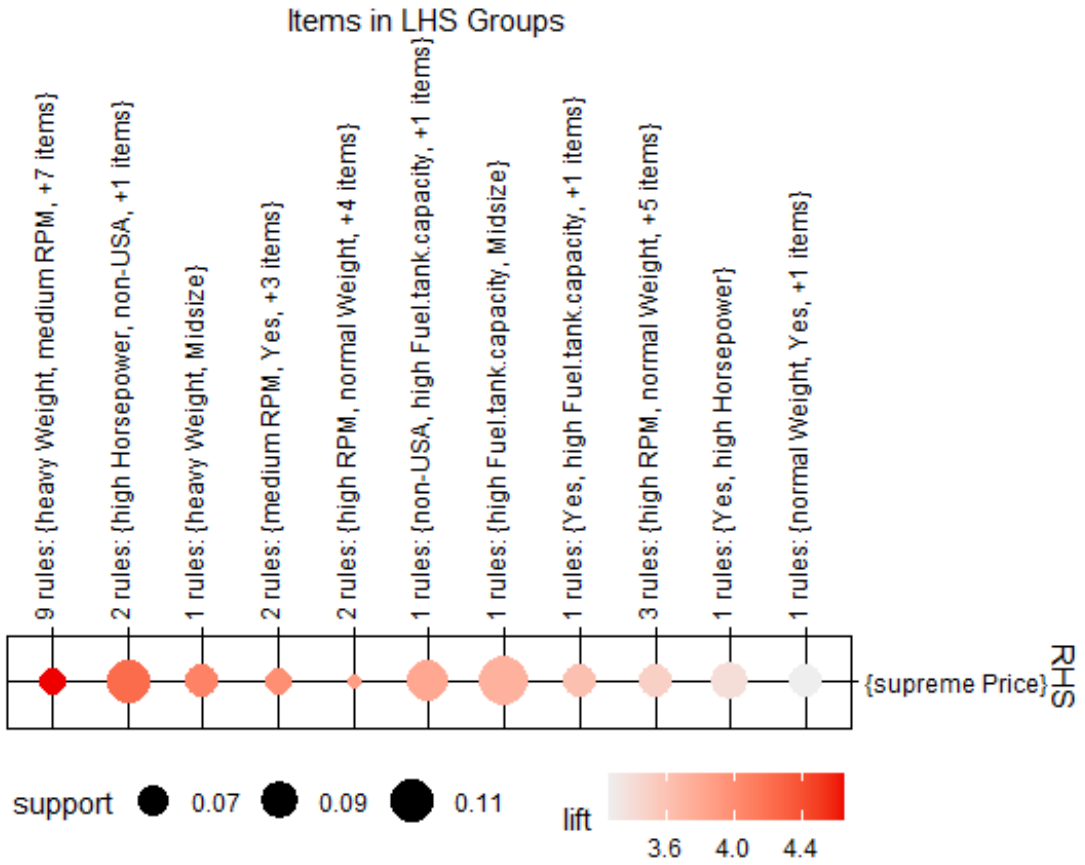
	lhs		rhs	support	confidence	coverage	lift	count
[1]	{high Fuel.tank.capacity, high RPM, Midsize}	=>	{supreme Price}	0.05376344	1.0000000	0.05376344	4.650000	5
[2]	{heavy Weight, high Horsepower, Midsize}	=>	{supreme Price}	0.06451613	1.0000000	0.06451613	4.650000	6
[3]	{heavy Weight, Midsize, non-USA}	=>	{supreme Price}	0.05376344	1.0000000	0.05376344	4.650000	5
[4]	{high Horsepower, Midsize, No}	=>	{supreme Price}	0.05376344	1.0000000	0.05376344	4.650000	5
[5]	{high Fuel.tank.capacity, high Horsepower, Midsize}	=>	{supreme Price}	0.10752688	1.0000000	0.10752688	4.650000	10
[6]	{high Horsepower, Midsize, non-USA}	=>	{supreme Price}	0.08602151	1.0000000	0.08602151	4.650000	8
[7]	{heavy Weight, high Horsepower, non-USA}	=>	{supreme Price}	0.05376344	1.0000000	0.05376344	4.650000	5
[8]	{high Horsepower, medium RPM, non-USA}	=>	{supreme Price}	0.06451613	1.0000000	0.06451613	4.650000	6
[9]	{high Fuel.tank.capacity, high Horsepower, medium RPM, Yes}	=>	{supreme Price}	0.05376344	1.0000000	0.05376344	4.650000	5
[10]	{high Horsepower, Midsize}	=>	{supreme Price}	0.10752688	0.9090909	0.11827957	4.227273	10
[11]	{high Horsepower, non-USA}	=>	{supreme Price}	0.10752688	0.9090909	0.11827957	4.227273	10
[12]	{heavy Weight, Midsize}	=>	{supreme Price}	0.07526882	0.8750000	0.08602151	4.068750	7
[13]	{high Fuel.tank.capacity, Midsize, Yes}	=>	{supreme Price}	0.06451613	0.8571429	0.07526882	3.985714	6
[14]	{high Horsepower, medium RPM, Yes}	=>	{supreme Price}	0.06451613	0.8571429	0.07526882	3.985714	6
[15]	{high RPM, Midsize}	=>	{supreme Price}	0.05376344	0.8333333	0.06451613	3.875000	5

注意上圖畫螢光筆的列，我的 support 最低限度設定為 0.05，但這四列的 support 值都接近 0.1，且 lift 一樣很高，代表他們的參考價值相較下會高一些。

結論：

通常**高的油箱容量、高馬力、類型為 Midsize 且非美國製造**的車子會較為昂貴，似乎符合封面的那台車呢！

下圖為規則的氣球圖：



最後可再將 rules 寫入 csv 檔以便後續查看：

```
#####將rule寫進CSV#####
write(rules_pruned, file = "carrule.csv", sep=",", row.names=F)
```

以上為我的報告，感謝聆聽。