Reglas de asociación: Trabajo Final

Juan Ramón Gómez Berzosa

31/1/2019

Introducción: Análisis y preprocesamiento dataset CAR

En primer lugar vamos a hablar sobre el dataset que vamos a utilizar para realizar nuestra extracción de reglas de asociación. El dataset elegido está en la base de datos de Keel (original de UCI) y se llama "Car". Se trata de un conjunto de datos de clasificación el cual recoge características asociadas a coches con el fin de que sirvan para clasificarlos en función de una etiqueta llamada "acceptability." Esta etiqueta indica cual es nivel de aceptabilidad por los clientes del coche atendiendo a dichas características.

En primer lugar leeremos el dataset y realizaremos un análisis previo de este, llevando a cabo un preprocesamiento adecuado para poder aplicar los algoritmos de reglas pertinentes para extraer la máxima información posible de este.

```
## 'data.frame':    1728 obs. of  7 variables:
##  $ Buying       : Factor w/ 4 levels "high","low","med",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Maint        : Factor w/ 4 levels "high","low","med",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Doors        : Factor w/ 4 levels "2","3","4","5more": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Persons      : Factor w/ 3 levels "2","4","more": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Lug-Boot     : Factor w/ 3 levels "big","med","small": 3 3 3 2 2 2 1 1 1 3 ...
##  $ Safety       : Factor w/ 3 levels "high","low","med": 2 3 1 2 3 1 2 3 1 2 ...
##  $ Acceptability: Factor w/ 4 levels "acc","good","unacc",..: 3 3 3 3 3 3 3 3 3 3 ...
```

El dataset contiene un total de 1728 observaciones, teniendo disponibles 7 predictores y una variable de clasificación. Todos los datos son factores, por lo que nos vendrá bien posteriormente para tratarlo con los algoritmos de reglas de asociación.

Las variables que intervienen en el proceso de decisión son:

- *Buying*: Precio de compra del coche.
- *Maint*: Precio de mantenimiento del coche.
- *Doors*: Número de puertas del coche.
- *Persons*: Número de personas que pueden viajar en el coche.
- *Lug-Boot*: Tamaño del maletero para equipaje.
- *Safety*: Nivel de seguridad estimado del coche.

La variable *Acceptability* es un factor de 4 niveles dónde se indica, como hemos dicho antes, el nivel de aceptación que tiene un coche por sus compradores. El nivel de aceptabilidad de un coche viene determinado por su precio, precio de compra y de mantenimiento, y sus características técnicas, las de confort como número de puertas, personas y el tamaño del maletero, y el nivel de seguridad del coche.

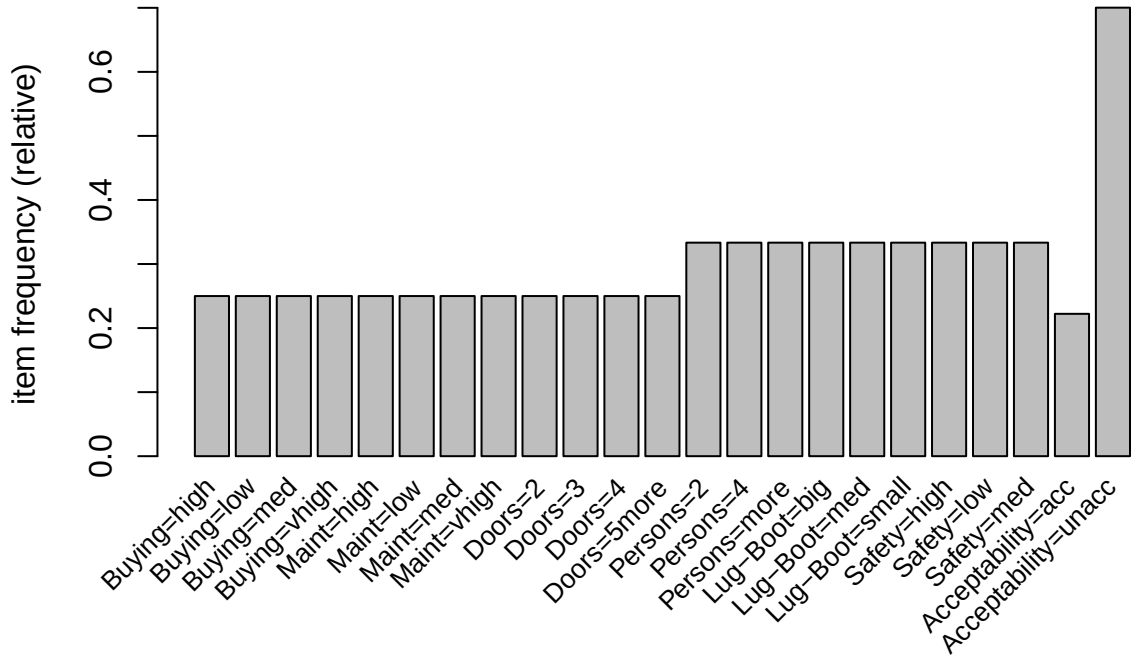
Cabe destacar que el conjunto de datos no tiene datos perdidos por lo que hemos visto en la página de keel, aunque también lo hemos comprobado. Una vez terminado esto vamos a mostrar un resumen del dataset.

```
##    Buying      Maint       Doors     Persons     Lug-Boot    Safety
##  high :432   high :432   2    :432   2   :576   big  :576   high:576
##  low  :432   low  :432   3    :432   4   :576   med  :576   low :576
##  med  :432   med  :432   4    :432   more:576   small:576   med :576
##  vhigh:432   vhigh:432   5more:432
##  Acceptability
##  acc  : 384
##  good :  69
```

```
##  unacc:1210
##  vgood:  65
```

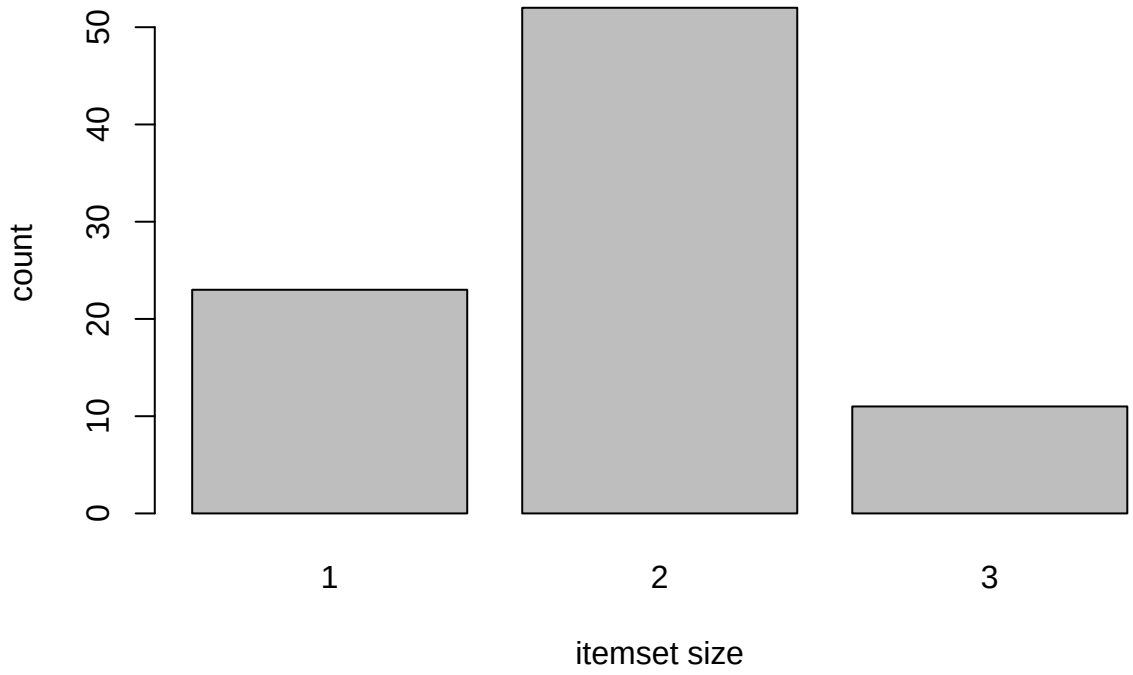
Podemos apreciar como el número de observaciones de cada tipo para cada variable está muy equilibrado, tanto que tenemos el mismo número de casos en cada clase para todos los atributos excepto para la aceptabilidad, donde si podemos apreciar un gran desbalanceo ya que el 70% de las observaciones tienen como variable de aceptabilidad la no aceptación. Por tanto en primer lugar lo analizaremos de forma genérica y después extraermos las observaciones correspondientes a aceptaciones de coches para intentar extraer más información.

Como análisis inicial podemos observar que el item que más se repite sería la no acceptabilidad del coche, mientras que vemos que el número de personas por coche tiene la misma frecuencia para 2, 4 o más personas. También podemos corroborar que no hay valores perdidos, pues hemos obtendio un total de 1728 transacciones las cuales tienen 7 atributos. Ahora mostraremos los items más frecuentes, aunque antendiendo a lo anterior y al resumen inicial, podemos darnos cuenta que inicialmente van a tener una frecuencia muy similar.



Como habíamos podido adelantar, vemos que todos los items pertenecientes a los distintos tipos de la misma variable tienen una distribución idéntica, teniendo como mínimo todos una frecuencia del 25%. Cabe destacar que el item más frecuente es la no aceptación del coche y el menos frecuente la aceptación. Por tanto, podemos afirmar que el conjunto de datos es bastante desbalanceado ya que practicamente 2/3 se corresponden con observaciones de la clase no aceptación y el tercio restante se reparte entre los otros 3 tipos de aceptación. También cabe destacar que es un conjunto de datos preparado para machine learning por lo que posiblemente no podamos obtener una información muy relevante con la búsqueda de reglas.

Atendiendo a lo que hemos comentado anteriormente y esa distribución de items frecuentes, no podemos sacar muchas conclusiones sólo fijándonos en esto, ya que lo único que podríamos decir es que tenemos más coches que no son aceptados a coches que si que lo son porque es el único atributo donde encontramos una variabilidad de frecuencia dentro de sus clases. A continuación, vamos a sacar los itemsets más frecuentes que tengan una frecuencia mínima de 0.1, utilizando el algoritmo apriori para intentar extraer información más relevante.



El número de itemsets frecuentes que hemos obtenido es 86, el número tan bajo se debe a que realmente tenemos pocas características y no muchas observaciones, con lo cuál tiene lógica. También podemos ver los itemsets más frecuentes son los formados por 2 elementos, un único item en el antecedente y otro en el consecuente, estando seguramente la mayoría formados por el item no aceptación ya que era el más frecuente, estando presente en 2/3 del dataset. Cabe destacar que también son muy frecuentes los de 1 único item, pero de estos no vamos a poder sacar a priori reglas de utilidad. Los que menos soporte tienen son los itemsets formados por 3 items y seguramente sean los más interesantes a la hora de extraer reglas de utilidad.

No merece la pena ponernos a analizar itemsets cerrados o máximales ya que hemos encontrado muy pocos itemsets frecuentes de por sí, así que pasaremos a extraer las reglas, siendo los criterios de soporte y confianza del 10% y 80% respectivamente. El 10% de soporte nos dice que la regla tendrá que estar soportada por aproximadamente 173 transacciones, lo cual es un número aceptable. Cabe destacar que a partir de ahora todas las reglas que se tienen en cuenta son teniendo en cuenta además de las medidas clásicas de soporte y confianza, la medida lift para evitar reglas con un alto valor de soporte en el consecuente y que por tanto sean engañosas a la hora de interpretar.

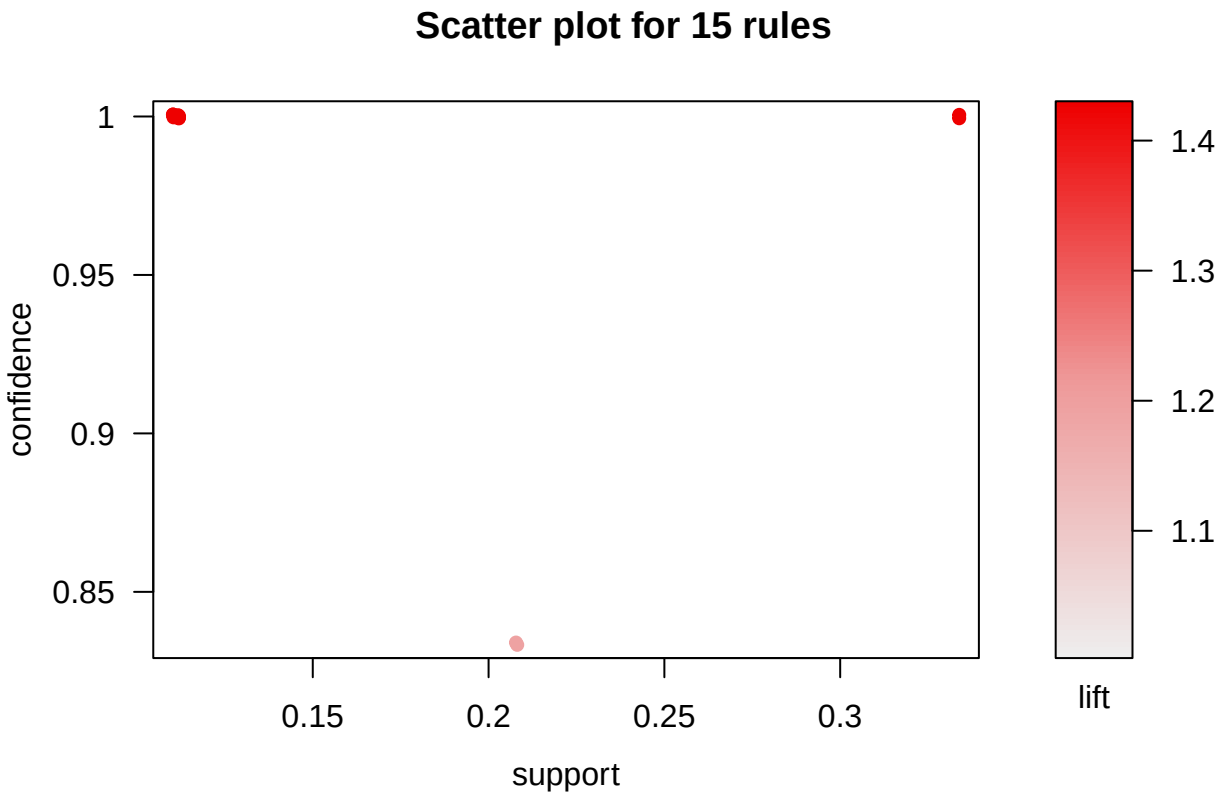
Como era de esperar atendiendo al número de itemsets frecuentes, el número de reglas que hemos obtenido ha sido muy bajo. Hemos obtenido 15 reglas, siendo la gran mayoría de 3 items, como habíamos anticipado antes, teniendo un soporte bajo siendo como máximo del 33%. En cuanto a confianza, las reglas tienen de media una confianza del 97% y unos valores de lift todos por encima de 1.

En general, todas las reglas tienen en su consecuente el item de no ser aceptado, lo cual tiene lógica ya que era el item que más se repetía el dataset. Encontramos reglas muy lógicas como que los coches que tienen una baja seguridad no tienen aceptación por parte de los clientes.

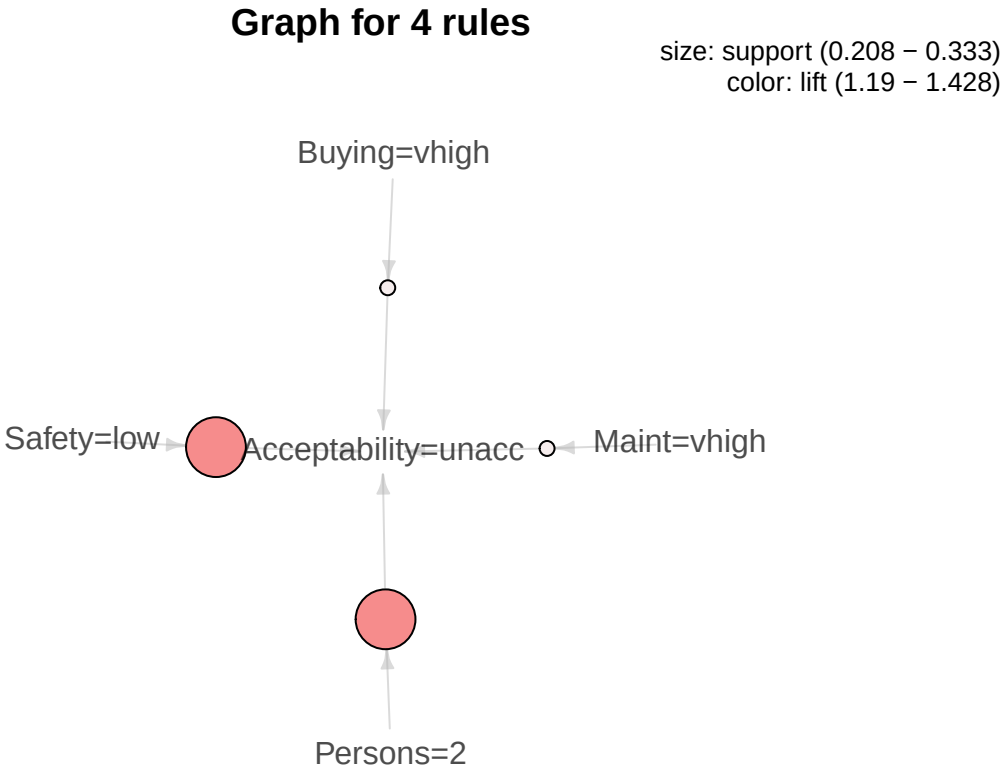
En principio no eliminaremos las reglas redudantes ya que tenemos muy pocas, así que procedremos a analizarlas.

```
## Loading required package: grid
```

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```



Por lógica, deberíamos de fijarnos en las reglas que están en el intervalo del 20-35%, ya que tienen un soporte y confianza aceptables, sin ser un soporte muy bajo para ser reglas inútiles ni un soporte muy alto para ser reglas poco interesantes a priori.



Podemos ver que los coches con seguridad baja nunca tienen una aceptación, en ninguno de los casos. Esto

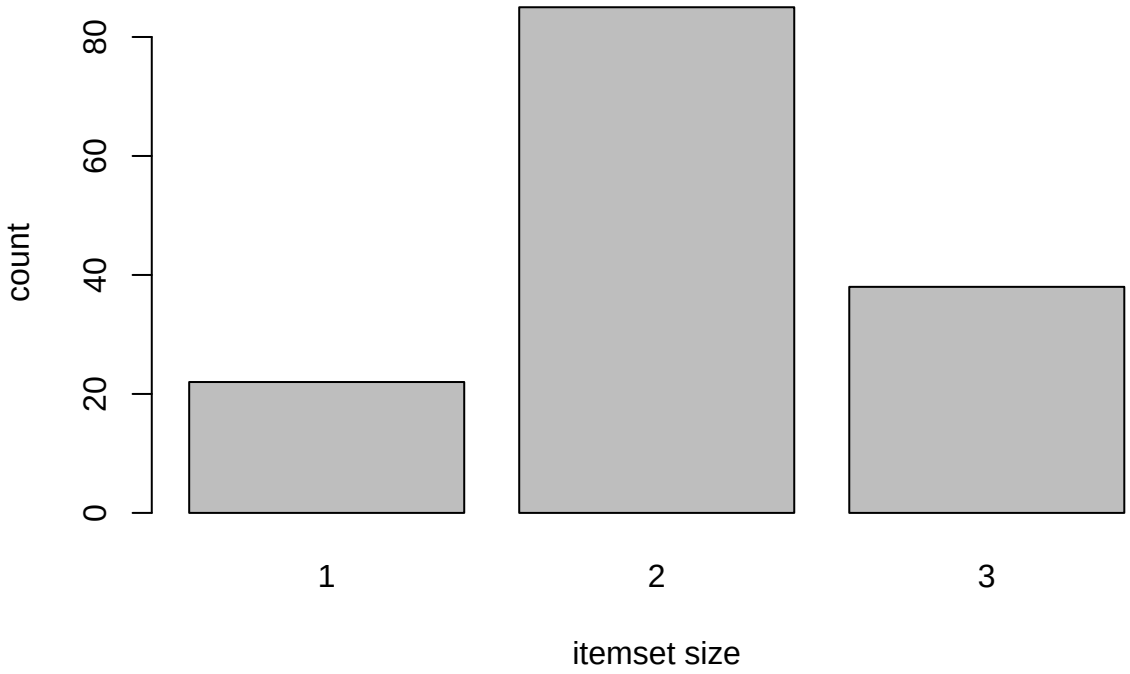
también sucede con los coches de dos plazas como es lógico desde mi punto de vista, ya que las personas suelen comprar sus coches para viajar, moverse por la ciudad o ir a trabajar y por norma general acaban formando una familia por lo que no pueden ir en un coche de 2 plazas. Desde mi punto de vista, estos coches son más pensados para temas específicos como la fórmula 1, deportivos de lujo, entre otros, pero estos datos no son tenidos en cuenta ya que se tratan de excepciones y por tanto no se recogen en este conjunto de datos.

Sin embargo, lo que si podemos ver es que los coches los cuales tienen un alto valor para ser adquiridos o mantenidos aunque tienen una alta confianza de que suelen ser coches no aceptados, hay casos en los que no se cumple. Esto puede deberse a las personas que están interesadas en coches de alta gama como son del tipo porsche, mercedes, audi... donde el precio de los coches es muy alto y normalmente suelen ser caros mantenerlos (debido a que son caras las piezas si se averían, la gasolina que consumen... etc) pero la gente los adquiere porque están interesados en este tipo de coches por lo que suelen tener una buena aceptacción.

También hemos visto en las reglas anteriores que factores como el tamaño del maletero influían a la hora de no aceptar un coche, sin embargo eran factores secundarios y normalmente todo era condicionado por el tema de la seguridad, número de personas y el precio de compra y mantenimiento, siendo los dos primeros los principales temas buscados a la hora de la aceptación de un coche.

Sin embargo, lo que más podemos destacar es que de todas las reglas que hemos obtenido hasta ahora, ninguna nos lleva a una situación en la que la aceptación del coche sea al menos baja. Por tanto, vamos a repetir este análisis pero escogiendo las observaciones del conjunto de datos referentes a coches que si son aceptados.

```
##    Buying       Maint        Doors     Persons     Lug-Boot    Safety
##  high :108   high :118   2    :106   2   :  0   big  :208   high:299
##  low  :174   low  :164   3    :132   4   :264   med  :184   low :  0
##  med  :164   med  :164   4    :140   more:254   small:126   med :219
##  vhigh: 72   vhigh: 72   5more:140
##  Acceptability
##  acc  :384
##  good : 69
##  unacc:  0
##  vgood: 65
```



Ahora el número de itemsets frecuentes es mayor, 145, a cuando teníamos en cuenta las observaciones de la

no aceptabilidad de los coches. Ahora veremos si podemos extraer reglas relevantes.

Hemos obtenido un total de 18 reglas, siendo la mayoría (11 en particular) de 3 items de tamaño y el resto de 2 items. El soporte máximo está en un 34% y tenemos valores de lift todos por encima de 1, mientras que la confianza de media no es muy alta pero tiene especial sentido al haber hecho esto sobre el conjunto de datos eliminando las observaciones referentes a la no aceptabilidad para centrarnos en las restantes. De estas 18 reglas nos vamos a centrar en aquellas que no son redundantes.

Al eliminar reglas redundantes hemos disminuido la cantidad a apenas unas 9 reglas. De estas reglas las cuales nos hemos centrado solo en aquellas transacciones donde las personas han aceptado los coches (ya sea con un nivel más alto o más bajo de aceptación), podemos ver que se da que sólo si la aceptación ha sido excelente tenemos una seguridad de coche alta, por lo tanto podemos ver que la seguridad es la prioridad.

Las demás reglas nos dan condiciones que se dan para la acepación de un coche, sin embargo es el nivel de aceptación más bajo y por tanto está en los umbrales de las reglas que hemos dado anteriormente para los coches no aceptados. Podemos ver por tanto que coches con una seguridad media son normalmente aceptados, aunque con bastante frecuencia. Podemos ver que los coches que son caros tanto de adquirir o mantener son normalmente aceptados, sin embargo los que lo son muy caros en ambos sentidos ya no se suele dar con tanta asiduidad, suponiendo que en este caso entrarían las personas con mayor poder adquisitivo o más dispuestas a gastarse tanto dinero en un vehículo. Esto último lo analizaremos posteriormente en el apartado de análisis por grupos de reglas.

También podemos destacar la regla que dice que normalmente si el maletero es pequeño, el coche es aceptado, aunque con un nivel de aceptación muy bajo. Al realizar una búsqueda de reglas bajando los niveles de soporte al 5% y de confianza al 70%, hemos encontrado una regla con alto cumplimiento en este grupo la cual trata de que, dentro de los coches que son aceptados, los coches con maletero pequeño son por norma general altamente seguros. Esta regla tiene bastante sentido y es muy interesante ya que los coches que tengan una menor capacicad en el maletero van a llevar menos equipaje y por tanto se reduce la probabilidad de que el coche tienda a colear en algunas curvas debido al exceso de equipaje en el maletero.

Análisis de ítems negados

El análisis que hemos hecho hasta ahora nos ha servido para reafirmar información que intuíamos aunque sobre todo para destacar las dos principales variables que influyen a la hora de la aceptación de un coche, por encima del precio del vehículo o del mantenimiento: - *Número de personas* - *Seguridad*

Sin embargo, aunque esta información es interesante desde nuestro punto de vista y atendiendo a los datos disponibles, ahora nos vamos a centrar en intentar profundizar un poco más en esta línea. Vamos a proceder a realizar un análisis de ítems negados para ver si podemos encontrar otras reglas que no habían sido detectadas anteriormente. Como anteriormente hemos visto que la variable de la capacidad del maletero estaba presente en las reglas tanto teniendo en cuenta los coches no aceptados como solamente teniendo en cuenta los aceptados, vamos a probar a negar dicho item.

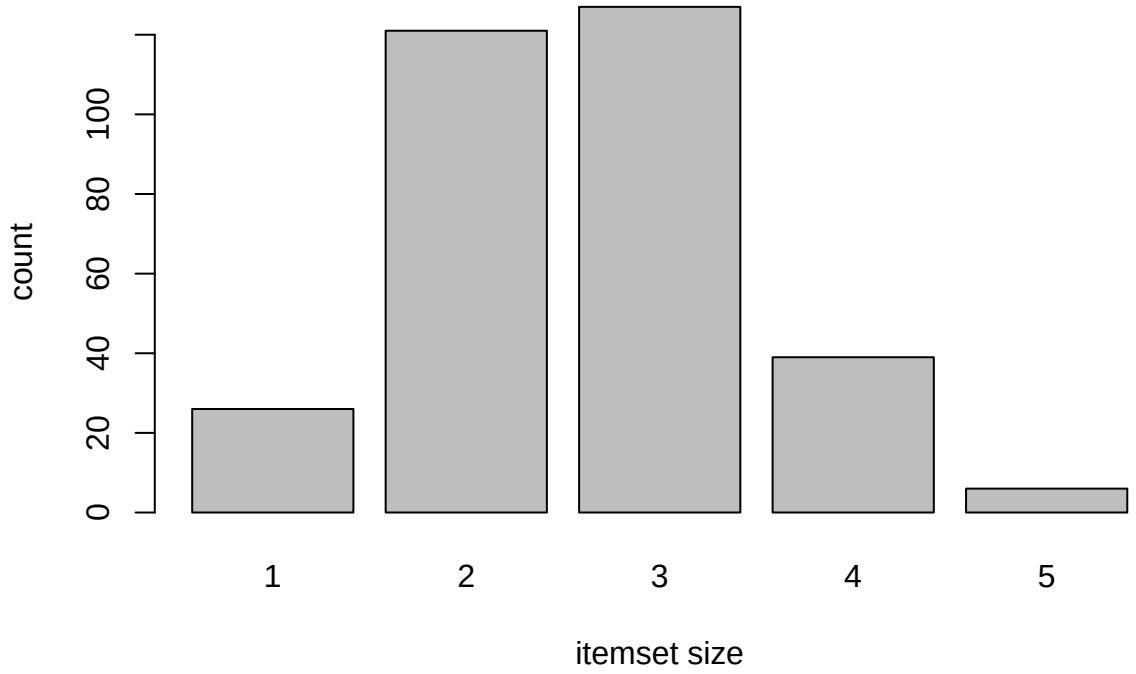
```
summary(carNegados)
```

```
##    Buying       Maint        Doors      Persons     luggBootSmall luggBootMed
##  high :432   high :432   2    :432   2   :576   FALSE:1152    FALSE:1152
##  low  :432   low  :432   3    :432   4   :576   TRUE : 576    TRUE : 576
##  med  :432   med  :432   4    :432   more:576
##  vhigh:432   vhigh:432   5more:432
##  luggBootBig    Safety    Acceptability
##  FALSE:1152   high:576   acc  : 384
##  TRUE : 576   low :576   good :  69
##               med :576   unacc:1210
##                          vgood:  65
```

Ahora vamos a pasar el dataset a transacciones para poder analizarlo mediante reglas de asociación.

Ahora tenemos otros items que son frecuentes debido a los negados que hemos obtenido. Vamos a analizar ahora los itemsets frecuentes.

```
barplot(table(size(iCarsNegados)), xlab = "itemset size", ylab = "count")
```

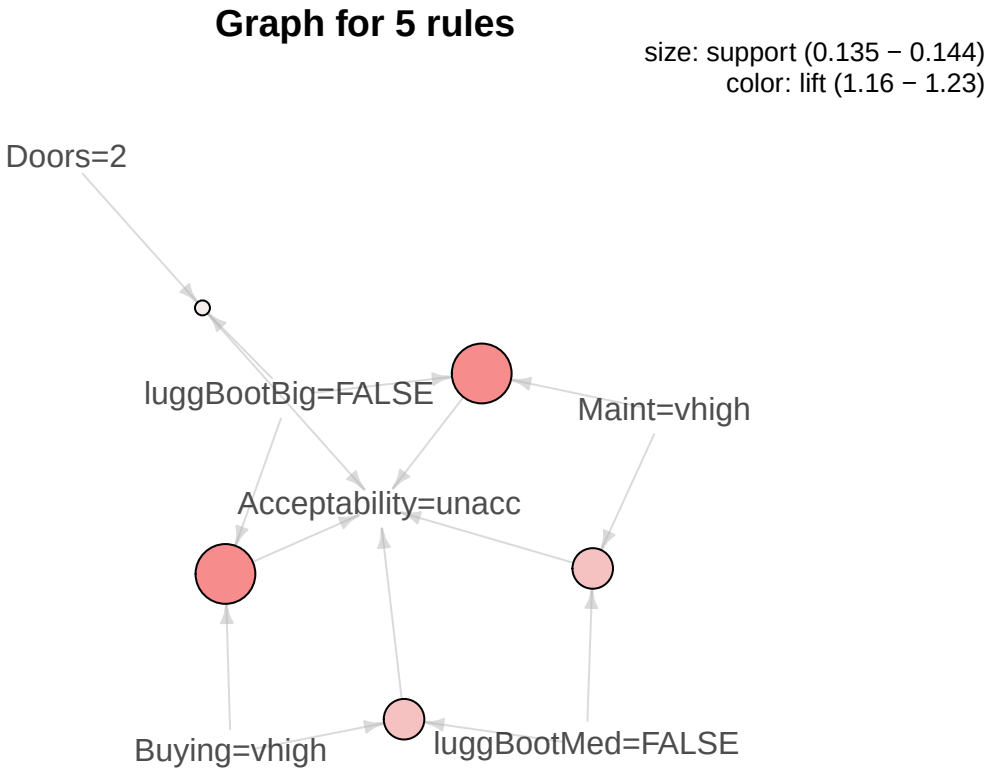


Como podemos apreciar la cantidad de itemsets frecuentes que se han generado el triple a la que teníamos inicialmente, 319, introduciéndose itemsets frecuentes de 4 y 5 items y siendo los más más frecuentes los de 2 y 3 items. Vamos a generar ahora las reglas.

Como es lógico también, el número de reglas ha crecido, teniendo un total de 200. Sin embargo, al introducir los items negados tenemos muchas reglas redundantes y carentes de utilidad, por lo que las eliminaremos.

Una vez eliminadas las reglas redundantes obtenemos un total de 15 reglas, de las cuales tenemos tenemos reglas que no son de utilidad como: si tenemos un tamaño de maletero grande no tenemos un tamaño de maletero pequeño, así que procederemos a eliminar a mano este tipo de reglas inútiles.

```
##     lhs                      rhs                      support confidence     lift count
## [1] {Persons=2}           => {Acceptability=unacc} 0.3333333  1.0000000 1.428099   576
## [2] {Safety=low}          => {Acceptability=unacc} 0.3333333  1.0000000 1.428099   576
## [3] {Buying=vhigh}        => {Acceptability=unacc} 0.2083333  0.8333333 1.190083   360
## [4] {Maint=vhigh}         => {Acceptability=unacc} 0.2083333  0.8333333 1.190083   360
## [5] {Buying=vhigh,
##      luggBootBig=FALSE}   => {Acceptability=unacc} 0.1435185  0.8611111 1.229752   248
## [6] {Maint=vhigh,
##      luggBootBig=FALSE}   => {Acceptability=unacc} 0.1435185  0.8611111 1.229752   248
## [7] {Buying=vhigh,
##      luggBootMed=FALSE}   => {Acceptability=unacc} 0.1400463  0.8402778 1.200000   242
## [8] {Maint=vhigh,
##      luggBootMed=FALSE}   => {Acceptability=unacc} 0.1400463  0.8402778 1.200000   242
## [9] {Doors=2,
##      luggBootBig=FALSE}   => {Acceptability=unacc} 0.1354167  0.8125000 1.160331   234
```



Esto nos deja un total de 9 reglas, de las cuales las 4 primeras son las mismas que obtuvimos en el primer análisis y por tanto no las mostramos en la gráfica. Como nuevas reglas podemos destacar con más soporte y confianza que normalmente los coches que son caros y no tienen el maletero grande no son aceptados, pasando esto también pero con los coches con un maletero mediano y con un poco menos de soporte y confianza. También podemos comentar la última regla la cual nos dice que normalmente los coches que tienen dos puertas y no tienen el maletero grande tampoco son aceptados. Desde mi punto de vista esto es interesante ya que aunque ya sabíamos que reglas nos llevaban principalmente a la no aceptación del coche, esto nos de una información extra que corrobore algunas características de los coches que no tienen aceptación, por lo que podría ser útil a la hora de comercialización.

Análisis por grupos de reglas

El análisis por grupos de reglas en este caso puede ser un poco complicado por lo que hemos visto, ya que hemos obtenido un número bajo de reglas y por tanto no hay mucho en lo que comparar. Sin embargo, hemos encontrado un caso claro que comentaremos a continuación y después comentaremos los resultados que hemos podido obtener bajando los límites de confianza y soporte para obtener más reglas.

De entre lo que hemos visto anteriormente podemos destacar las dos reglas lógicas que nos afirmaban que los coches que son muy caros de adquirir o mantener son no aceptados por normal general, estas reglas tenían una confianza cercana al 83%. Sin embago, cuando hemos obtenido las reglas correspondientes al conjunto de datos pero descartando los coches no aceptados y atendiendo únicamente a los coches aceptados en sus diferentes reglas, hemos obtenido otras dos reglas que se corresponderían con las excepciones de las dos comentadas anteriormente. Estas dos reglas nos indicaban que si los coches eran muy caros de adquirir o mantener entonces los coches eran aceptados, aunque con el nivel más bajo de aceptación, las cuales tenían una confianza del 100% respecto a los datos segmentados.

Por tanto, podemos afirmar que estos casos son los que restaban del 17% restante de confianza de las primeras dos reglas y desde nuestro punto de vista se pueden corresponder con los casos que hemos comentado antes

de aquellas personas que están interesadas en coches de alta gama bien por su poder de adquisición o bien por gusto propio.

En cuanto al resto, hemos intentado hacer un análisis de reglas reduciendo los parámetros de soporte mínimo y confianza, observando tanto el conjunto de datos original como el conjunto de datos reducido del que se habían extraído las observaciones de coches no aceptados y no hemos podido encontrar más grupos de reglas que nos llevaran a una anomalía clara o una excepción u otro tipo de reglas correspondientes a este campo, ya que la gran mayoría de las reglas tenían la no aceptación o la aceptación en su nivel más bajo en el consecuente (debido en mi opinión en gran parte al tipo de conjunto de datos y su desbalanceo), nunca se incluía en el antecedente, por lo que era más difícil aún de detectar grupos de reglas de este tipo.

Queda claro que el conjunto de datos está mas orientado a un tratamiento mediante técnicas de aprendizaje automático que utilizando métodos para extraer información de bases de datos como son las reglas de asociación, ya que aunque según mi criterio hemos podido obtener alguna regla de utilidad no queda del todo claro que estas reglas de cara a un experto en el dominio puedan ser descartadas.