Alessia Poli, 922096, EMA
a. a. 2019 2020

ESERCIZIO IN R
MARKET BASKET ANALYSIS: DATASET VenditeOnLine.csv
Problema da risolvere:
Il processo di identificazione delle associazioni tra articoli sul comportamento di acquisto dei clienti rappresenta un elemento cruciale per l'elaborazione di opportune strategie di marketing orientate a rendere più efficaci le politiche promozionali.

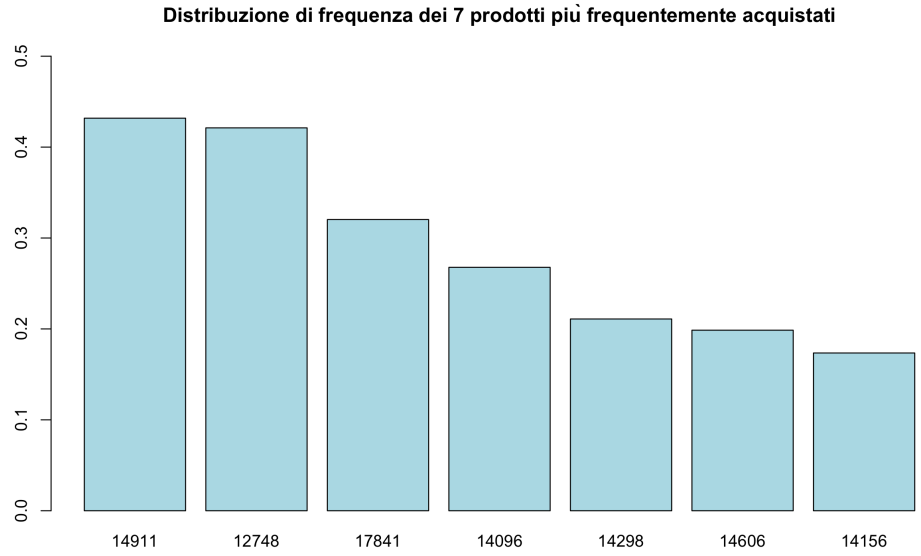
I dati a disposizione:
I dati a disposizione riguardano le transazioni avvenute nell'intervallo temporale compreso tra lo 01/12/2010 e il 09/12/2011 attraverso un portale inglese di vendite al dettaglio on line. Gli articoli in vendita appartengono principalmente alla categoria "regali per tutte le occasioni". Ogni riga identifica uno specifico acquirente (che può apparire più volte all'interno del dataset a seconda del numero di volte in cui effettua un acquisto) e il suo comportamento di acquisto viene esplicitato sia in termini di articolo, sia in termini di codice articolo. Le variabili presenti nel dataset sono le seguenti:

- CustomerID: identificativo dell'acquirente;
- Items: tipologia di articolo;
- StockCode: codice articolo.

Obiettivi:

2. INDIVIDUAZIONE DEI 7 PRODOTTI PIÙ FREQUENTEMENTE ACQUISTATI E RELATIVA RAPPRESENTAZIONE GRAFICA

```
barplot(head(VenditeOnLine_freq,7),ylim=c(0,0.5),col="light
blue",main="Distribuzione di frequenza dei 7 prodotti più frequentemente
acquistati") # grafico a barre della frequenza dei 7 prodotti più frequentemente
acquistati
```



Il grafico a barre rappresenta la distribuzione di frequenze dei 7 prodotti più frequentemente acquistati: è possibile osservare sull'asse delle ascisse il codice relativo al prodotto e sull'asse delle ordinate la frequenza relativa di ciascuno di essi.

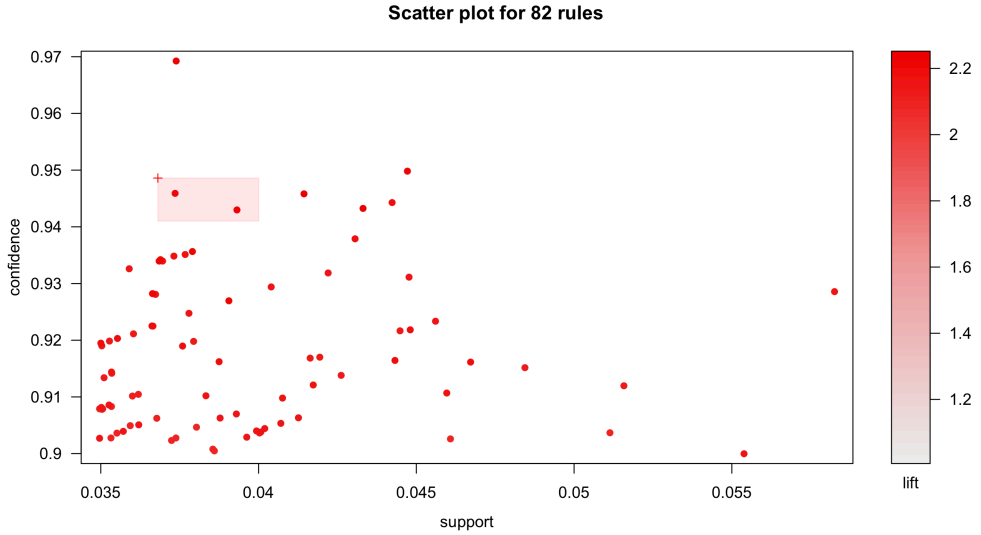
3. IDENTIFICAZIONE DI AL MASSIMO 100 REGOLE ASSOCIATIVE, FISSATI UN ADEGUATO LIVELLO DI SUPPORT E DI CONFIDENCE

```
>  VenditeOnLine.rules<-apriori(VenditeOnLine,parameter=list(supp=0.035,conf=0.9,target="rules"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.9    0.1    1 none FALSE            TRUE       5   0.035      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 147

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[4372 item(s), 4224 transaction(s)] done [0.10s].
sorting and recoding items ... [411 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 done [0.02s].
writing ... [82 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Con un livello di support pari al 35% e un livello di confidence al 90% è possibile identificare 82 regole associative tra gli items presenti nel dataset transazionale.

4. COMMENTO (IN TERMINI DI SUPPORT, CONFIDENCE E LIFT) DI DUE REGOLE ASSOCIATIVE A SCELTA

```
plot(VenditeOnLine.rules,engine="interactive") # scatter plot delle regole associative che permette la selezione arbitraria di regole associative
```



Lo scatterplot è il grafico a dispersione in cui vengono riportate combinazioni di variabili di un dataset. Ciascun punto rappresenta una delle 82 regole associative ("se una condizione, quindi il risultato"), di

cui vengono analizzate il livello di support (ascisse), il cui livello minimo è pari 3,5%, e il livello di confidence (ordinate), il cui livello minimo è pari a 90%. Inoltre l'intensità del colore di ciascun punto rappresenta il livello di lift della relativa regola (maggiore l'intensità del colore, maggiore il livello di lift).

Per una data regola A→B, dova A è corpo e B è testa, la **support** della regola A→B è la frequenza relativa che indica la proporzione di transazioni in cui è osservata la regola stessa.

La **confidence** della regola, invece, misura la forza della relazione tra due oggetti, ossia la probabilità condizionata della testa rispetto al corpo.

Infine la **lift** collega la confidence della regola con il support della testa (B) della stessa regola, misurando in che rapporto si verifica la regola associativa rispetto alla probabilità degli items che la costituiscono.

```
Number of rules selected: 2
    lhs               rhs     support    confidence lift     count order
[1] {14456,15529} => {12748} 0.03929924 0.9431818  2.239460 166   3
[2] {14088,14156} => {14911} 0.03740530 0.9461078  2.190986 158   3
```

Count rappresenta il numero delle transazioni che contengono il termine della regola e **order** il numero di elementi della regola.
Analisi della prima regola:

- Support: circa il 3.9% delle transazioni comprende sia i prodotti 14456 e 15529 che il prodotto 12748,
- Confidence: circa il 94,3% delle transazioni che comprendono i prodotti 14456 e 15529, allora comprendono anche 12748,
- Lift: il numero delle transazioni che contengono sia i prodotti 14456 e 15529 sia il prodotto 12748 è pari a 2,24 volte il numero delle transazioni che contengono o solo i prodotti 14456 e 15529 o solo il prodotto 15529,
- Count: 166 transazioni contengono la regola [1] per cui, se si verificano 14456 e 15529, allora anche 12748,
- Order: la regola contiene 3 elementi, 2 nel corpo (lhs), ossia gli items 14456 e 15529, e 1 nella testa (rhs), ossia l'item 12748.

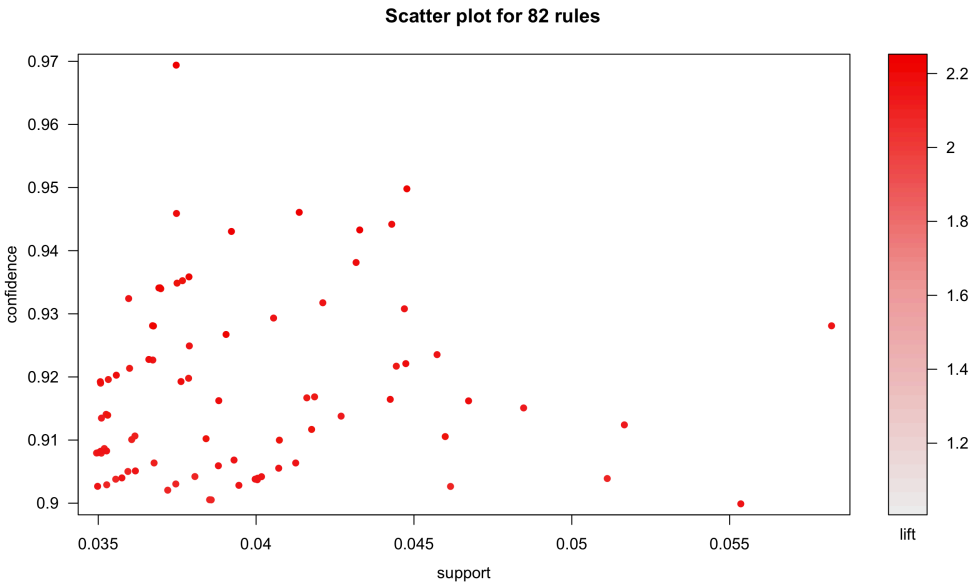
Analisi della seconda regola:

- Support: circa il 3.7% delle transazioni comprende sia i prodotti 14088 e 14156 che il prodotto 14911,
- Confidence: circa il 94,6% delle transazioni che comprendono i prodotti 14088 e 14156, allora comprendono anche 14911,
- Lift: il numero delle transazioni che contengono sia i prodotti 14088 e 14156 sia il prodotto 14911 è pari a 2,19 volte il numero delle transazioni che contengono o solo i prodotti 14088 e 14156 o solo il prodotto 14911,
- Count: 158 transazioni contengono la regola [2] per cui, se si verificano 14088 e 14156, allora anche 14911,

- Order: la regola contiene 3 elementi, 2 nel corpo (lhs), ossia gli items 14088 e 14156, e 1 nella testa (rhs), ossia l'item 14911.

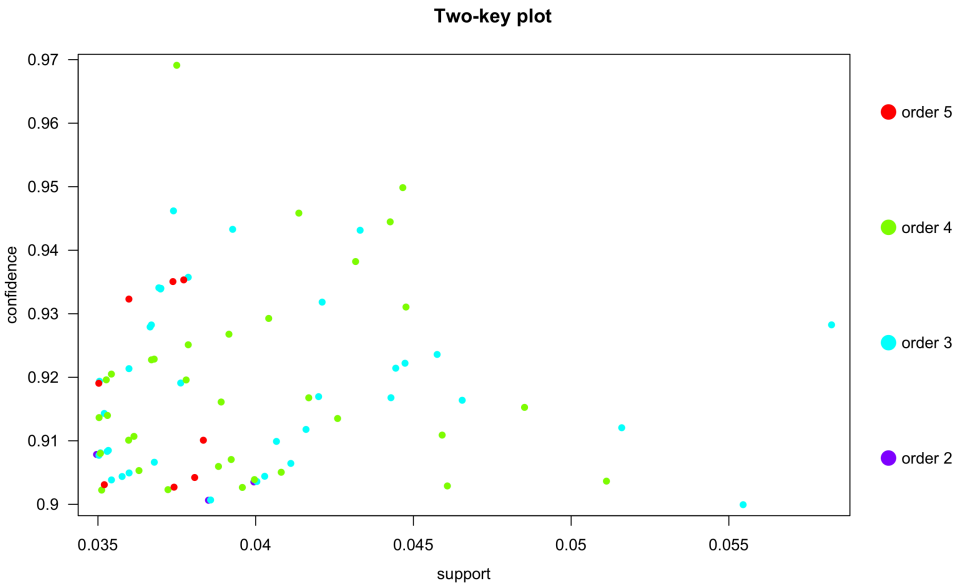
5. RAPPRESENTAZIONE DELLO SCATTER PLOT DELLE REGOLE OTTENUTE AL PUNTO 3, IN FUNZIONE DEGLI INDICI DI SUPPORT, CONFIDENCE E LIFTE E LO SCATTER PLOT DELLE REGOLE OTTENUTE AL PUNTO 3, RIPORTANTE L'ORDINE DI GRANDEZZZA DELLE REGOLE

```
plot(VenditeOnLine.rules) # scatter plot delle regole associative individuate
secondo i parametri di interesse
```



Scatterplot delle 82 regole associative identificate per un livello minimo di support del 3,5% (ascisse) e un livello minimo di confidence del 90% (ordinate). Il livello di lift è rappresentato dall'intensità del colore del punto nello scatterplot (maggiore l'intensità del colore, maggiore il livello di lift).

```
plot(VenditeOnLine.rules,method="two-key plot") # scatter plot delle regole
associative individuate secondo i parametri di interesse e riportante l'ordine
di grandezza delle regole
```

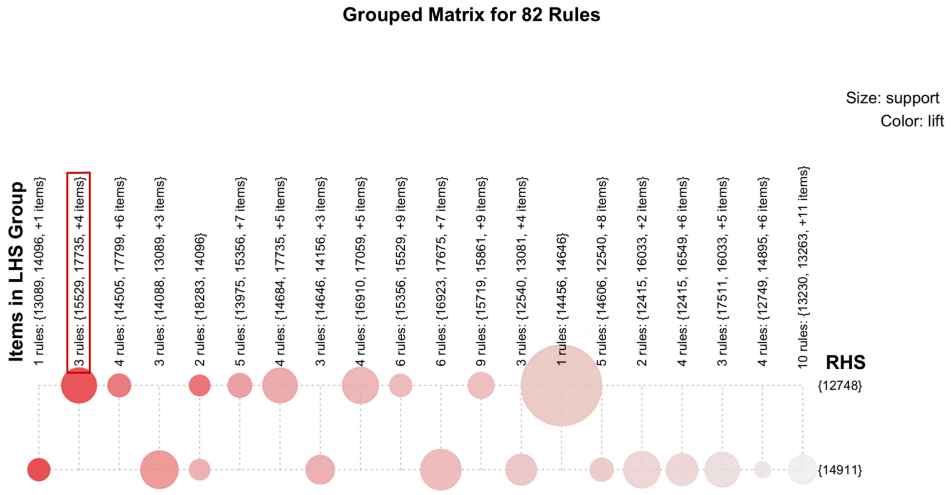


Scatterplot delle 82 regole associative identificate per un livello minimo di support del 3,5% (ascisse) e un livello minimo di confidence del 90% (odrinate, lato sinistro), di cui viene anche indicato l'order (ordinate lato destro). In questo caso viene sacrificata la rappresentazione grafica del livello di lift, a favore della visualizzazione dell'ordine di grandezza delle regole.
A seconda dell'ordine di grandezza, il punto rappresentativo di ciascuna delle regole associative assume un colore diverso, come indicato sulla destra del grafico: se l'ordine di grandezza è pari a 2 il punto sarà di colore viola, se l'ordine di grandezza è pari a 3 il punto sarà di colore azzurro, se l'ordine di grandezza è pari a 4 il punto sarà di colore verde e se l'ordine di grandezza è pari a 5 il punto sarà di colore rosso.

6. RAPPRESENTAZIONE DELLE REGOLE OTTENUTE AL PUNTO 3, CON UN GRAFICO A PALLONCINI, INDICANDO UNA DELLE REGOLE CARATTERIZATE DAL LIFT PIÙ ELEVATO

Nel grafico a palloncini ogni regola è rappresentata da un palloncino all'intersezione tra gli items del corpo (LHS, ordinate) e della testa (RHS, ascisse).
L'intensità del colore di ciascun cerchio rappresenta il livello di lift (maggiore è l'intensità del colore, maggiore è il livello di lift). Invece, la dimensione dei palloncini rappresenta il livello di support di ciascuna regola, ossia la frequenza con cui si verifica la regola associata al palloncino.

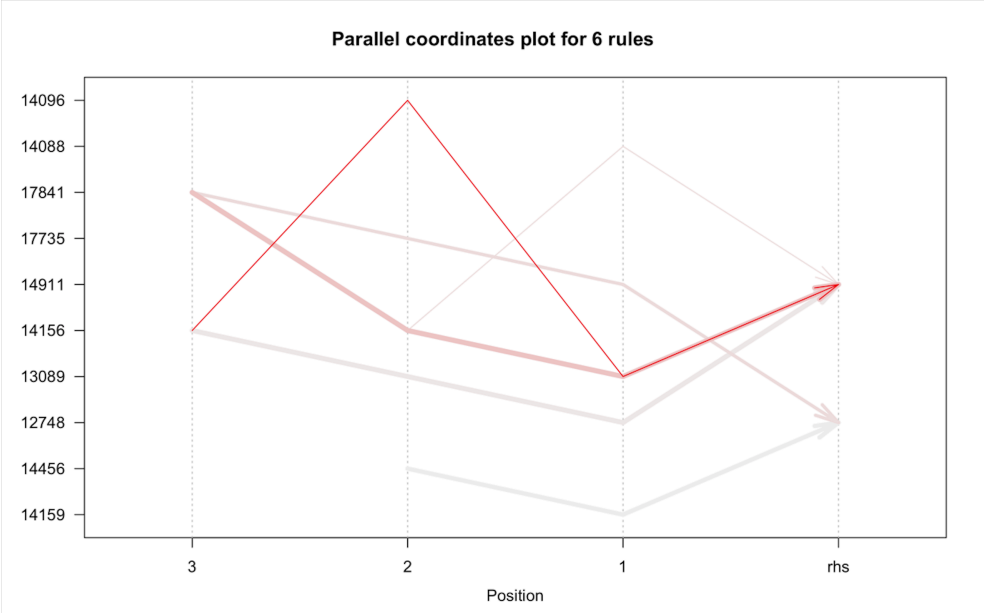
```
plot(VenditeOnLine.rules,method="grouped") # visualizzazione delle regole con
grafico a palloncini
```



Il gruppo con il maggior livello di lift consiste in 3 regole che contengono gli items identificati dai codici 15529 e 17735 e altri 4 items nel corpo (LHS) e tutte le regole hanno come testa (RHS) l'item identificato dal codice 12748.

7. RAPPRESENTAZIONE DELLE PRIME 6 REGOLE ASSOCIATIVE CARATTERIZZATA DA UN MAGGIOR VALORE DI CONFIDENCE ATTRAVERSO UN GRAFICO CHE RIPORTI SOLO LE MISURE DI SUPPORT E CONFIDENCE.

```
VenditeOnLine_confidence<-head(sort(VenditeOnLine.rules,by="confidence"),6) # estrazione delle 6 regole associative con maggior confidence
plot(VenditeOnLine_confidence,method="paracoord",shading="confidence") # grafico delle 6 regole associative con maggiore confidence
```



Alessia Poli, 922096, EMA
a. a. 2019 2020

Il grafico a frecce rappresenta le prime 6 regole associative, caratterizzate da un maggior valore di confidence (come specificato nel primo comando). Il livello di support è indicato dallo spessore delle frecce (maggiore lo spessore, maggiore il livello di support); il livello di confidence è indicato dal colore delle frecce (maggiore l'intensità del colore, maggiore il livello di confidence), perché specificato, all'interno del comando, dallo strumento "shading". Se questo non vi fosse stato, il colore avrebbe rappresentato il livello di lift delle regole associative.
Sull'asse delle ordinate sono elencati i codici relativi agli specifici items e sull'asse delle ascisse è indicata la posizione occupata da ciascun item all'interno del corpo; il punto in cui cade la freccia cade rappresenta la testa della regola associativa.