ANNO ACCADEMICO 2021/2022
DATA SCIENCE PER L'ECONOMIA E LE IMPRESE
DATA ANALYSIS FOR BUSINESS DECISIONS

Assignment 1

Cluster Analysis per segmentazione della clientela

A CURA DI: RICCARDO ROCCHI, BENITO DEL PESCHIO, CHIARA ANTOGNOZZI E FEDERICA VOLPI

Obiettivo

Lo scopo della seguente analisi è quello di effettuare una segmentazione della clientela, sulla base di dati fornitici inerenti al comportamento di consumo. Nello specifico, il dataset di riferimento contiene informazioni sulle singole transazioni dei clienti, includendo dettagli sul prodotto considerato (id prodotto, variante, quantità, ..).
Individuare gruppi omogenei di clienti rappresenta in ambito commerciale un punto di partenza cruciale per comprendere a pieno le dinamiche del proprio mercato di riferimento, nonché per realizzare successivamente un'efficace strategia di marketing.

Metodologia

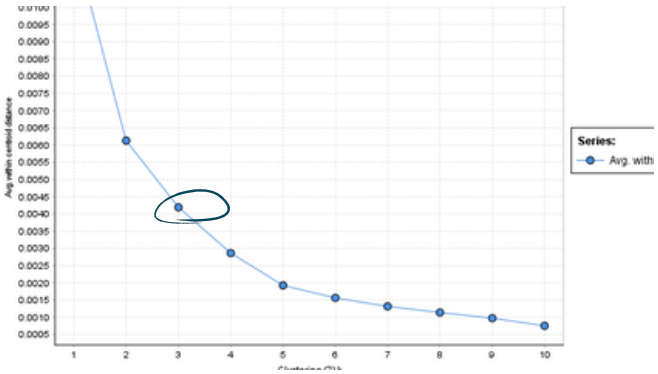
Prima di poter effettuare un'efficace segmentazione della clientela, sono stati implementati i seguenti step:

- aggregazione dei record riferiti allo stesso cliente, per ottenere ID univoci;
- creazione di una matrice di correlazione in grado di individuare variabili altamente correlate tra loro che, fornendo informazioni simili, sono da considerare ridondanti ai fini dell'analisi;
- eliminazione di quei record per cui il valore del 'discount' usufruito per l'acquisto supera il valore lordo di vendita. Pur non avendo assoluta certezza che quest'ultimo fosse un vincolo da rispettare, la scelta di eliminare tali record è stata effettuata vista la scarsa perdita di informazione (pari solamente allo *0.014% del totale);*
- eliminazione delle variabili "orderd item quantity" e "retrurns" in quanto, una volta aggregati i dati per cliente, risultavano solo 2 record in cui i prodotti erano stati restituiti;
- verifica di eventuale presenza di valori mancanti o significativamente discostanti dalla maggior parte, senza evidenti risultati;
- standardizzazione di ogni variabile (attraverso la tecnica min-max) per poter ottenere valori confrontabili tra loro;
- selezione delle variabili da utilizzare nell'analisi e nell'algoritmo scelto. In particolare sono state considerate solo features numeriche e eliminate quelle che presentavano un'elevata correlazione, ovvero ridondanti in quanto fornivano poca informazione aggiuntiva. A partire dalla seguente matrice, le variabili selezionate come utili sono state 'Gross Sales', 'Discount', 'Total Sales', 'Ordered Item Quantity'.

	Net Quantity	Gross Sales	Discounts	Returns	Net Sales	Taxes	Total Sales	Returned Item Quantity	Ordered Item Quantity
Net Quantity	1.000000	0.385819	-0.084897	0.157515	0.352742	0.338748	0.216322	0.211056	0.967661
Gross Sales	0.385819	1.000000	-0.722134	0.062968	0.892956	0.900862	-0.047786	0.086023	0.372157
Discounts	-0.084897	-0.722134	1.000000	0.270160	-0.707701	-0.709383	-0.024583	0.233671	-0.147078
Returns	0.157515	0.062968	0.270160	1.000000	0.073179	0.072014	0.027571	0.820231	-0.050674
Net Sales	0.352742	0.892956	-0.707701	0.073179	1.000000	0.975786	0.185032	0.098053	0.335244
Taxes	0.338748	0.900862	-0.709383	0.072014	0.975786	1.000000	0.067654	0.096492	0.321343
Total Sales	0.216322	-0.047786	-0.024583	0.027571	0.185032	0.067654	1.000000	0.036942	0.211575
Returned Item Quantity	0.211056	0.086023	0.233671	0.820231	0.098053	0.096492	0.036942	1.000000	-0.042341
Ordered Item Quantity	0.967661	0.372157	-0.147078	-0.050674	0.335244	0.321343	0.211575	-0.042341	1.000000

Analisi e presentazione dei risultati

L'algoritmo utilizzato è stato quello del K-means, che permette di individuare un numero finito di cluster omogenei all'interno e distinguibili tra loro. Per la scelta dei centroidi iniziali si è optato per l'utilizzo del K-means ++, che rappresenta un'ottimizzazione del tradizionale K-means.
Come riportato nel grafico a fianco, la scelta è ricaduta su una segmentazione della clientela basata su 3 gruppi. Sebbene, come sotto riportato, il numero di gruppi ottimale oscilli tra 3 e 6, la scelta di 3 è stata dettata dalla volontà di effettuare una segmentazione più chiara e di facile interpretazione.



Curva utilizzata per la scelta del numero di gruppi ottimale.

I clienti possono essere facilmente divisi sulla base degli ordini effettuati e delle vendite generate/fatturato generato:

- il cluster 0 contiene clienti che hanno acquistato in media 1-2 volte:
- il cluster 1 contiene clienti che in media hanno acquistato più di 2-3 volte;
- il cluster 2 contiene clienti che hanno acquistato in media più di 5 volte

Le caratteristiche fondamentali di ogni cluster sono visibili nelle tabelle sottostanti.

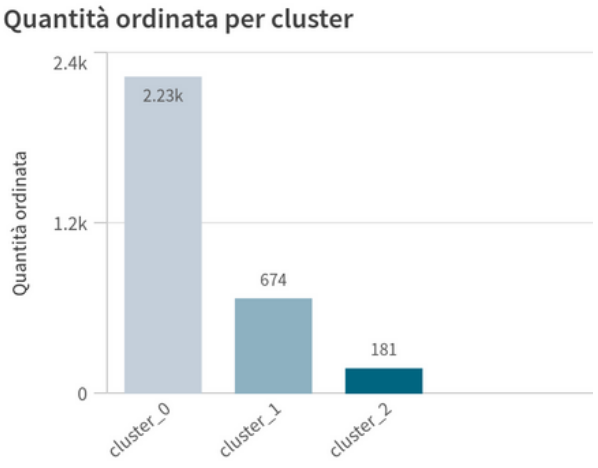
	Quantità ordinata in media	Discount medio usufruito	Vendite totali in media
cluster_0	1.0154475	-1776.5061	819.47297
cluster_1	2.1812298	-3819.6699	1811.521
cluster_2	5.1714286	-8182.8857	5313.4286

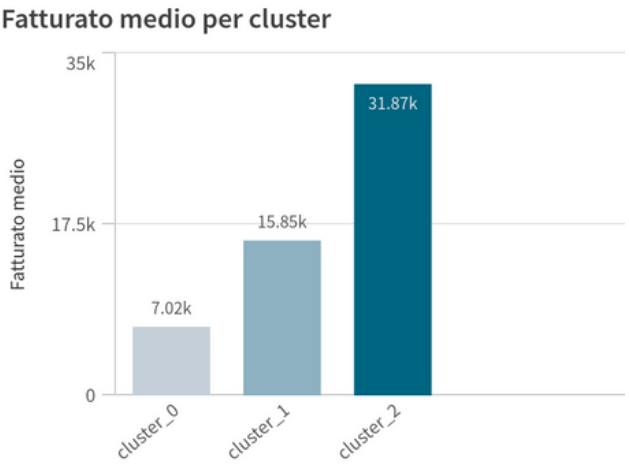
Numerosità dei gruppi

Gruppo	Numero dei clienti
cluster_0	2201
cluster_1	309
cluster_2	35
Totali	**2545**

Dal confronto delle tabelle di cui sopra, è interessante notare che il cluster 2, pur essendo il meno numeroso, è composto da clienti che hanno generato in media il fatturato più alto. Per questo motivo, si potrebbe interpretare questo gruppo come composto da clienti di nicchia, che acquistano a prescindere dal prezzo applicato al bene.

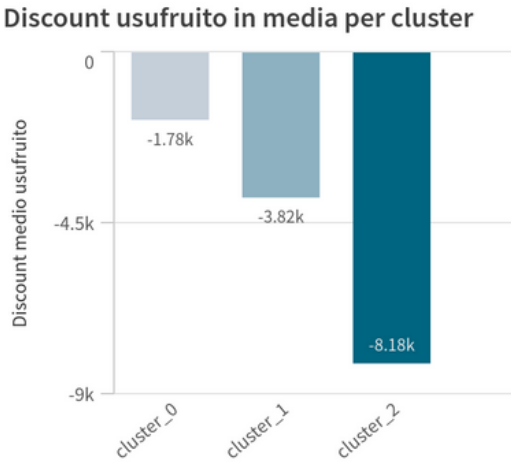
I grafici riportati di seguito, a sostegno di questa tesi, mostrano come in corrispondenza del cluster 2 la quantità ordinata in totale è significativamente più bassa rispetto ai restanti, ma capace di generare di per sé il 58.22% del fatturato totale.





Ciò nonostante, le vendite totali generate dal cluster 2 sono nettamente inferiori rispetto a quelle dei restanti gruppi e rappresentano circa il 7,3% del totale.

La differenza tra l'elevato fatturato e le vendite totali del cluster 2 è motivata dal grafico sottostante che rappresenta l'ammontare del discount in media usufruito dai clienti nell'acquisto, molto elevato per coloro che si trovano nel cluster 2.



TOTALE VENDITE PER CLUSTER IN PERCENTUALE