Unblind Consulting

# E-COMMERCE DATA DRIVEN OVERVIEW

## CUSTOMERS THROUGH CLUSTERING

Belardinelli Marco, Cicalè Juri,
Lepore Marco Antonio, Manzo Manuel

# GOAL SETTING

Individuazione di **gruppi omogenei (Cluster Analysis)** nella clientela dell'azienda sulla base di un dataset di riferimento contenente i dati relativi alle singole transazioni commerciali avvenute nell'e-commerce di proprietà.

La risultante segmentazione potrà poi essere base di lavoro ideale per l'implementazione di **strategie di marketing coerenti**, centrate e dall'impatto potenzialmente dirimente per quanto riguarda il livello della presenza sul mercato dell'azienda.

# ANALYSIS & FINAL RESULTS

Di seguito le principali risultanze dell'analisi effettuata, corredate da approfondimenti grafici volti a facilitarne e a veicolarne una comprensione quanto più immediata possibile; in appendice le specifiche tecniche relative al lavoro svolto, ove vi fosse la necessità di verificare, in maniera analitica, le logiche di elaborazione alla base di quanto sinteticamente esposto nel presente report.

L'analisi ha evidenziato la presenza di **tre gruppi omogenei** nella clientela con le seguenti caratteristiche:

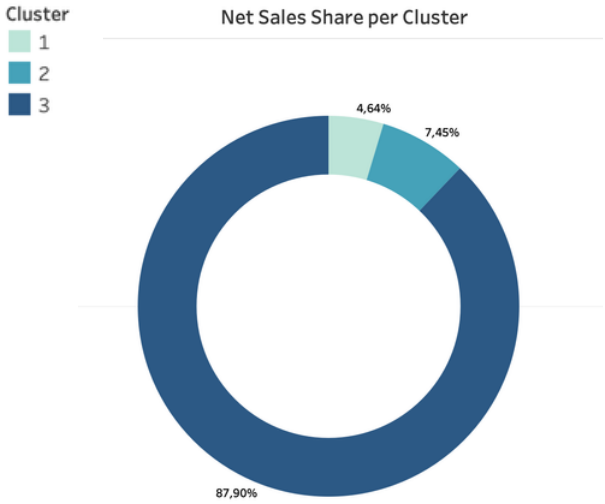
| | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| Cluster Size | 415 | 85 | 2030 |
| Total Revenue | 735.445,00 | 1.180.229,00 | 13.919.264,00 |
| Average Revenue | 1.772,16 | 13.885,05 | 6.856,78 |
| Total Net Quantity | 528 | 240 | 2275 |
| Average Net Quantity | 1,27 | 2,82 | 1,21 |
| Total Discount | 300.100,00 | 281.250,00 | 4.735.328,00 |
| Average Discount | 723,00 | 3.308,00 | 2.333,00 |

Il **cluster 3**, ovvero quello che raccoglie la maggior parte della clientela, evidenzia un acquisto con probabile focus su un prodotto (o famiglia di prodotti) specifico, non caratterizzato da una scontistica particolarmente vantaggiosa (rispetto a quelle generalmente applicate) ma di evidente ed importante richiamo commerciale poiché in grado di catalizzare buona parte del potenziale di vendita dell'azienda.

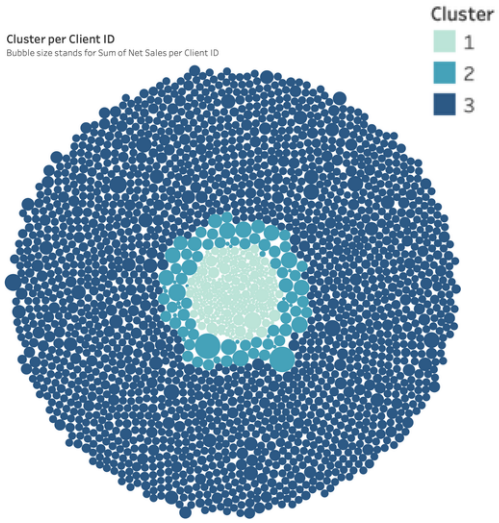
Il **cluster 2** è quello con meno clientela (circa il 3% del totale) ma con l'acquisto medio per cliente più elevato e lo sconto medio più basso. Viene rappresentato un segmento di clienti di estrema minoranza che compra più prodotti (circa 3 per cliente), spende mediamente di più ma usufruisce di una scontistica inferiore.

Il **cluster 1** rappresenta circa 1/5 dell'intero dataset ed è caratterizzato da un volume di acquisti per cliente decisamente più basso degli altri ma con un maggiore sconto applicato, è possibile immaginare di essere dinanzi a quel cliente che acquista un prodotto più inflazionato, in maniera occasionale, motivato prevalentemente dalla scontistica più rilevante rispetto alla media applicata.

Con alcuni grafici specifici possiamo dettagliare e visualizzare meglio quanto appena introdotto:

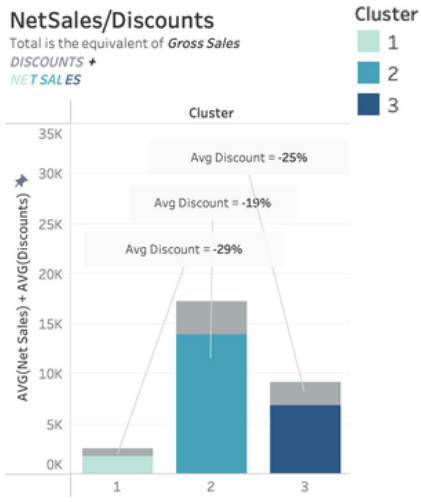



Questa rappresentazione ci da una visione subitanea ed indiscutibile del **ruolo dei gruppi** nel sostenere il fatturato globale dell'azienda.




Il grafico a bolle ci restituisce un'idea immediata della **rilevanza dimensionale dei cluster** anche al loro interno con il ruolo specifico del singolo cliente in relazione all'importo acquistato.


Qui viene evidenziata la prevalenza schiacciante di **clientela che acquista meno di due prodotti**, (tendenzialmente uno) ed il cluster 2 come eccezione con quasi 3 prodotti per cliente acquistati.




Interessante rimarcare quanto evidenziato in precedenza con questo grafico dedicato (dove possiamo vedere chiaramente lo sconto in percentuale in relazione al fatturato del singolo cluster), il livello di sconto non sembra essere il principale vettore di orientamento nell'acquisto poiché **il cluster che usufruisce dello sconto più rilevante è quello che produce il fatturato minore**.

# APPENDICE TECNICO METODOLOGICA

Possiamo suddividere il lavoro effettuato in due fasi principali:

**ANALISI ED ELABORAZIONE PRELIMINARE DEL DATASET**

**CLUSTERING**

Per quanto riguarda la presa in carico del dataset abbiamo operato un lavoro di **pulizia generale** e **valutazione preliminare dei dati,** nello specifico:

- Verificato la presenza di outliers, valori nulli e duplicati;
- Verificato ed eliminato errori ed incoerenze strutturali di alcune osservazioni quantitative:
  - le osservazioni per cui Discounts era maggiore di Gross Sales,
  - le osservazioni per cui la somma tra Net Sales e Discounts non era uguale a Gross Sales,
  - ricostruzione della colonna Total Sales come somma di Net Sales e Taxes (poiché presentava valori anomali per gran parte dei record).

Abbiamo poi implementato una **matrice di correlazione** ed eliminato tutte le variabili con indice superiore al 75%.

Per la fase due abbiamo valutato come soluzione ideale l'adozione dell'**algoritmo K-Means** e poi, di conseguenza, operato nel seguente modo:

- Dopo aver **raggruppato le osservazioni per ID cliente**, abbiamo utilizzato l'elbow method, la gap statistics e un ciclo for che misurasse la silhouette al variare del numero di cluster, dai risultati ottenuti.

- Abbiamo deciso di attuare una segmentazione della clientela in 3 cluster (le variabili che sono state utilizzate nell'algoritmo sono "net sales", "net quantity" e "returned items").

- Riguardo la scelta dei centroidi è stata ripetuta l'esecuzione dell'algoritmo numerose volte con seed differenti per definire quelli ottimali.