WHOLESALE CUSTOMERS DATA

In questo progetto mi occuperò di analizzare un dataset che contiene informazioni sulle spese annuali di vari clienti, suddivise per categorie di prodotti (freschi, latte, alimentari, surgelati, detersivi e prodotti di carta, gastronomia) e riportate in unità monetarie.

**OBIETTIVO**: comprendere le differenze tra i diversi tipi di clienti (di diverse regioni: Lisbona, Porto o altro) con cui collabora un distributore all’ingrosso (hotel, ristorante, bar oppure venditore al dettaglio).

**METODOLOGIA**:

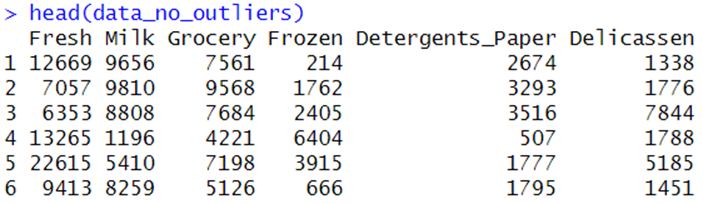
1. Metodo Z-score
2. Analisi Componenti Principali (PCA)
3. Clustering: metodo K-medie
4. **IDENTIFICARE ED ELIMINARE VALORI ANOMALI: METODO Z-SCORE**

Utilizzando questo approccio posso lavorare con più variabili per identificare outliers (valori anomali) in uno spazio multidimensionale per facilitare l’analisi.

Rimuovere i valori anomali mi aiuta a garantire, infatti, che i risultati delle analisi riflettano meglio le caratteristiche della popolazione principale, migliorando così le prestazioni e l'affidabilità delle tecniche utilizzate.

Uno Z-score indica quante deviazioni standard sono un valore lontano dalla media.

Si importa una soglia 3 o -3 in quanto valori con Z-score assoluti superiori sono considerati outliers.



**NUMERO DI RIGHE RIMASTE:** 414

**NUMERO DI RIGHE INIZIALI:** 440

1. **APPLICARE ANALISI COMPONENTI PRINCIPALI (PCA)**

Dopo aver normalizzato i dati e rimosso i valori anomali, possiamo applicare PCA a questi dati puliti per identificare le dimensioni che massimizzano la varianza delle variabili coinvolte.

PCA, infatti, riduce il numero di variabili nel dataset mantenendo la maggior parte dell'informazione, poiché il dataset con troppe variabili può diventare complesso e difficile da interpretare.

1. Osserviamo se vi è correlazione tra i dati per poter o meno utilizzare PCA:

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

Osservando la matrice di correlazione è possibile notare come:

-Milk abbia una correlazione positiva molto forte con Grocery;

-Grocery abbia una correlazione molto forte con Detergents\_Paper

-Detergents\_Paper abbia una correlazione molto forte con Grocery e Milk

Mentre le variabili Fresh, Frozen e Delicassen hanno correlazioni molto basse.

È quindi sensato eseguire la PCA sui dati standardizzati, cioè sulla matrice di correlazione piuttosto che sulla matrice di covarianza su Milk, Grocery e Detergents\_Paper mentre Fresh, Frozen e Delicassen sono debolmente correlate con le altre variabili: questo potrebbe indicare che queste due variabili non contribuiranno molto alla variabilità spiegata dalle prime componenti principali.

1. Selezionare il numero delle componenti principali:

Utilizzo la funzione prcomp() per effettuare la PCA sui miei dati, questa funzione mi centra le variabili in modo tale che abbia un valore della media zero.

Successivamente seleziono il numero delle componenti principali, risultato della funzione, attraverso la Regola di Kaiser. Essa suggerisce di mantenere tante componenti quanti sono gli autovalori maggiori di 1.

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamenteOsserviamo dalla tabella che le uniche due componenti con un valore proprio >1 sono PC1 e PC2.

È possibile confermare questo risultato utilizzando lo scree plot della varianza spiegata:

Immagine che contiene testo, linea, diagramma, schermata

Descrizione generata automaticamente

La prima componente spiega il 44.71% della varianza.

La seconda componente spiega il 22.72% della varianza

La terza componente spiega il 12.88% della varianza.

La quarta componente spiega il 12.11% della varianza.

La quinta componente spiega il 5.64% della varianza.

La sesta componente spiega il 1.93% della varianza.

Quindi la prima e seconda componente catturano gran parte dell’informazione dei dati originali.

1. Interpretazione delle componenti principali:

Adesso analizziamo la correlazione delle variabili alle componenti principali prescelte per capire quali contribuiscono maggiormente:

Immagine che contiene testo, diagramma, linea, Diagramma

Descrizione generata automaticamente

Le variabili Detergents\_Paper, Grocery e Milk contribuiscono significativamente alla prima componente (Dim1), suggerendo che queste categorie sono fortemente correlate tra loro e spiegano una buona parte della varianza in quella direzione.

D'altra parte, Fresh, Frozen, e Delicassen sono più importanti per la seconda componente (Dim2), suggerendo che queste variabili rappresentano una dimensione diversa della variabilità nei dati, separata da quella spiegata da Detergents\_Paper e le altre.

1. Riduzione dimensione matrice:

Dalla teoria sappiamo che l'esecuzione della PCA si riduce alla decomposizione propria della matrice di correlazione dei dati centrati. Gli autovettori corrispondenti agli autovalori ordinati (ordine non decrescente) sono rispettivamente le nostre componenti principali e le loro varianze.

Eseguirò una riduzione di dimensionalità creando un modello PCA usando solo 2 dimensioni, invece di 6.

Immagine che contiene testo, Carattere, schermata

Descrizione generata automaticamente

1. **APPLICARE CLUSTERING: METODO K-MEDIE**

Dopo aver ridotto la dimensionalità con PCA, eseguo il clustering per identificare gruppi o pattern nascosti nei dati. Il clustering, in questo caso, potrebbe aiutarmi a identificare gruppi di clienti con comportamenti di spesa simili, facilitando l'ottimizzazione dei servizi di consegna.

Utilizzo nello specifico il metodo del k-medie perché è uno degli algoritmi di clustering più semplici e veloci da eseguire, soprattutto su dataset già ridotti dimensionalmente con PCA. Dopo la riduzione, il numero di variabili è inferiore, quindi k-means può convergere più velocemente e trovare cluster in maniera più efficiente.

1. Individuazione numero dei cluster:

Utilizzo il Metodo del Gomito con la distanza euclidea in quanto di facile interpretazione.

L'obiettivo del metodo del gomito è identificare il punto in cui l'aggiunta di ulteriori cluster non riduce significativamente la somma delle distanze intra-cluster, suggerendo che si è raggiunto un buon compromesso tra il numero di cluster e la loro compattezza.

Nel grafico, si nota una riduzione significativa della WSS per i primi 3 o 4 cluster, mentre oltre il 4° cluster la riduzione diventa meno pronunciata. Questo punto di flessione viene chiamato "gomito" e indica il numero ottimale di cluster.

Secondo il grafico, il numero consigliabile sarebbe 3.

Immagine che contiene testo, linea, schermata, diagramma

Descrizione generata automaticamente

1. Analisi cluster:

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamenteDopo aver suddiviso le mie osservazioni in 3 cluster, analizzo quali variabili influenzano maggiormente i diversi gruppi:

-Cluster 1:

Questo gruppo spende molto di più in Grocery (16.545 unità) e Milk (10.481 unità) rispetto agli altri cluster.

Spendono anche una quantità moderata in Detergents\_Paper (7.296 unità).

È il gruppo che spende meno in Fresh e Frozen, indicando che i loro consumi sono focalizzati su beni di consumo e non su alimenti freschi o surgelati.

-Cluster 2 (Verde):

Si distingue per essere quello che spende di più in Fresh (22.146 unità) e in Frozen (5.336 unità).

Spendono molto meno in Grocery e Detergents\_Paper, il che suggerisce che questo gruppo potrebbe essere focalizzato sull'acquisto di prodotti freschi e surgelati piuttosto che su beni di consumo quotidiano.

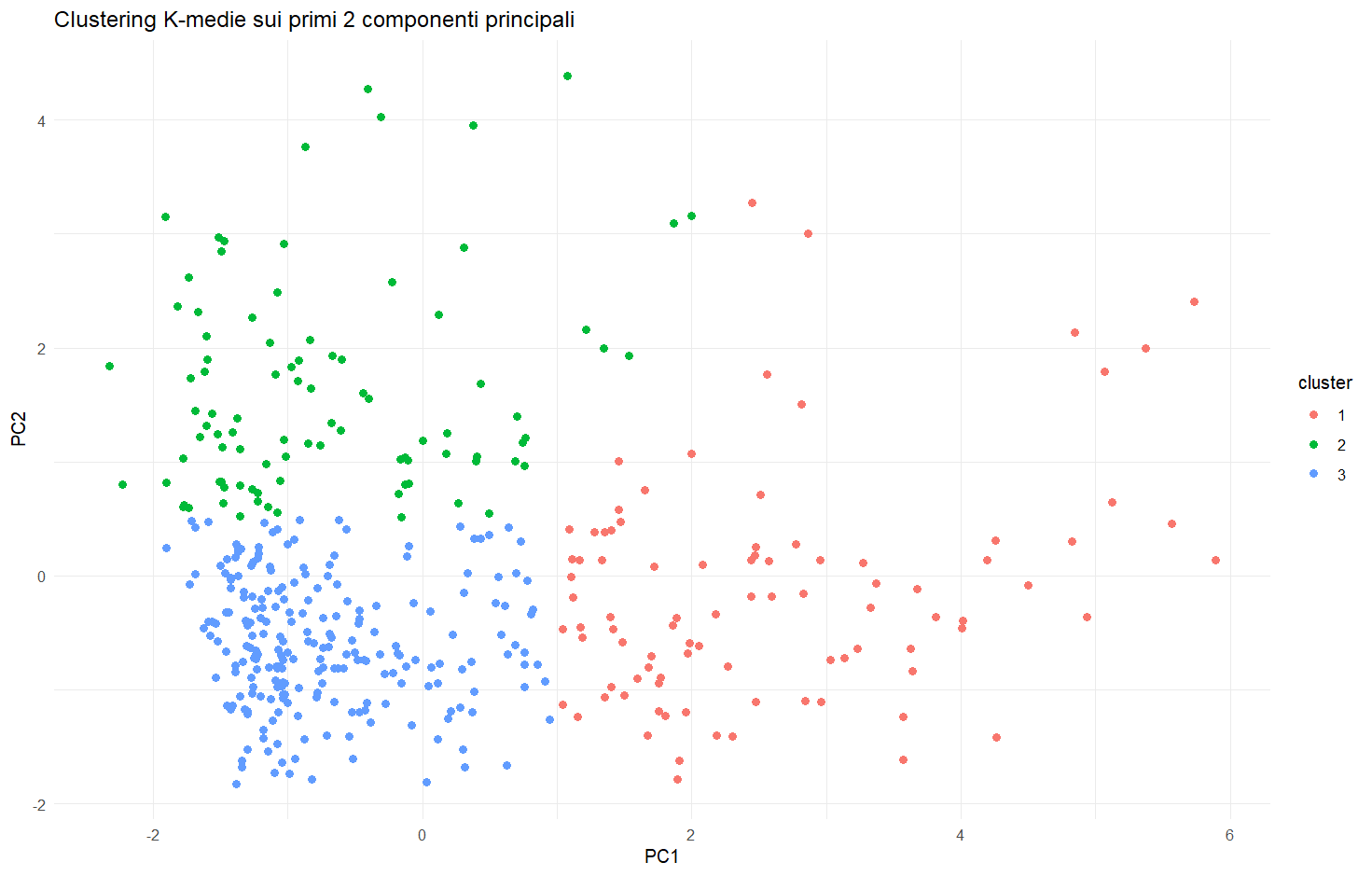
Questo cluster rappresenta probabilmente clienti che acquistano prodotti freschi in grandi quantità, come ristoranti o catering.

-Cluster 3 (Blu):

Spende una quantità moderata in Fresh (8.233 unità), ma molto meno rispetto al cluster 2.

Le spese in tutte le altre categorie (Milk, Grocery, Frozen, Detergents\_Paper, Delicatessen) sono le più basse rispetto agli altri cluster.

Questo cluster potrebbe rappresentare consumatori che acquistano quantità limitate di una vasta gamma di prodotti, senza focalizzarsi su una categoria specifica.



-Cluster 1:

è distribuito principalmente a destra lungo l'asse PC1, suggerendo che questo gruppo ha un pattern di spesa più concentrato su prodotti di consumo come Grocery e Milk, che influenzano fortemente il PC1. La loro posizione indica che sono distanti rispetto agli altri cluster lungo PC1, il che significa che il loro comportamento di spesa è fortemente influenzato da beni di consumo, come conferma anche la tabella (alta spesa in Grocery e Milk).

-Cluster 2 :

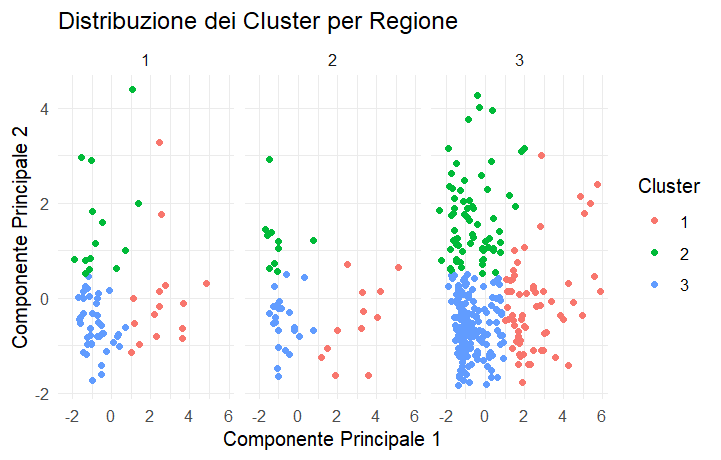
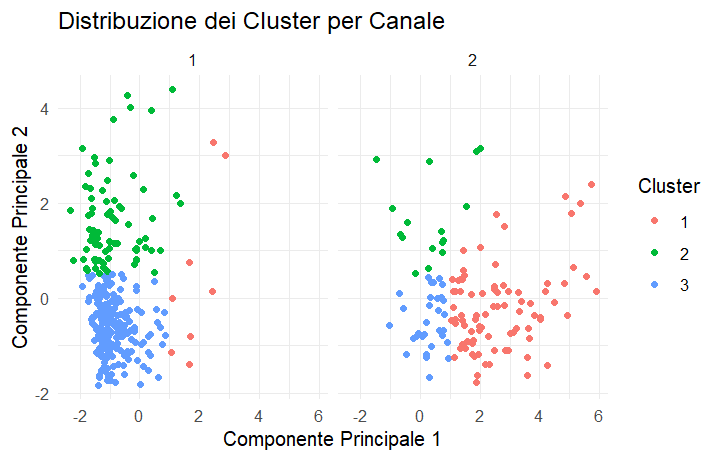
si concentra nella parte alta del grafico, lungo l'asse PC2. Questo indica che la loro spesa è influenzata fortemente da prodotti freschi e surgelati (Fresh, Frozen), che contribuiscono maggiormente al PC2. Sono separati verticalmente rispetto agli altri cluster proprio perché le variabili che incidono sul loro comportamento (Fresh e Frozen) dominano lungo PC2. Questo spiega perché, nella tabella, il cluster 2 ha la spesa più alta in Fresh e Frozen.

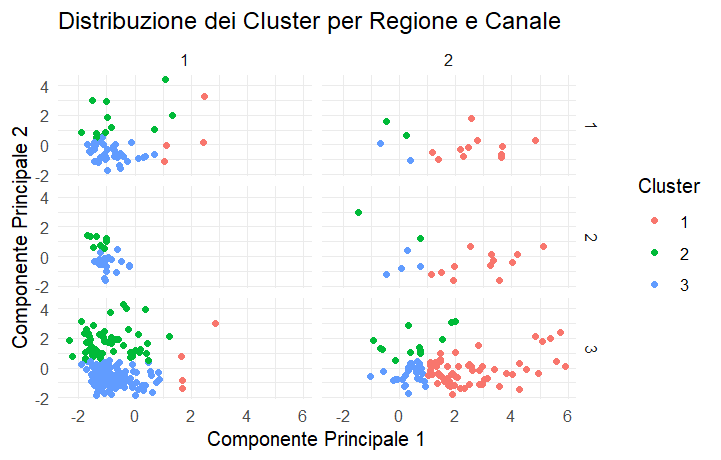
Cluster 3:

è distribuito principalmente nella parte sinistra del grafico, a valori negativi lungo PC1, il che suggerisce che spendono meno in Grocery e Milk rispetto al cluster 1, ma hanno una distribuzione più bilanciata su altre categorie. Questo si allinea con la tabella, dove il cluster 3 ha spese più basse in tutte le categorie rispetto agli altri gruppi.

Il grafico mostra che c’è poca sovrapposizione tra i cluster, suggerendo che i gruppi individuati sono ben distinti in termini di comportamento di spesa. Questo è un indicatore che il K-means ha trovato un buon numero di cluster che rappresentano gruppi di consumatori con pattern di acquisto chiaramente differenziati.

Adesso osserviamo come si distribuiscono i cluster per le variabili qualitative “Channel” e “Region”:





-Cluster 1 (Rosso):

-Distribuzione per Regione:

In Lisbona (Regione 1), Cluster 1 ha una buona distribuzione, ma è meno predominante rispetto agli altri cluster.

In Porto (Regione 2), Cluster 1 domina chiaramente, suggerendo che i clienti di Porto hanno un comportamento di spesa simile, con una concentrazione più alta nel cluster 1.

Nella regione Altro (Regione 3), Cluster 1 è ben rappresentato, ma non domina in maniera schiacciante come a Porto.

-Distribuzione per Canale:

Nei clienti hotel (Canale 1), Cluster 1 è distribuito, ma è meno predominante rispetto agli altri cluster.

Nei clienti della vendita al dettaglio (Canale 2), Cluster 1 è nettamente dominante, suggerendo che questo cluster rappresenta la maggior parte dei clienti di vendita al dettaglio, il che può indicare una similarità nei loro comportamenti di acquisto.

-Interpretazione:

Cluster 1 rappresenta un gruppo di clienti più omogeneo, soprattutto nella regione di Porto e nel canale della vendita al dettaglio. Questo cluster potrebbe rappresentare una clientela con comportamenti di spesa meno variegati, probabilmente con esigenze più standardizzate. I clienti potrebbero essere più orientati verso acquisti di prodotti di consumo rapido e quotidiano come Grocery e Milk.

-Cluster 2 (Verde):

-Distribuzione per Regione:

In Lisbona (Regione 1), Cluster 2 è presente in modo significativo, con una buona distribuzione dei clienti che si distinguono dagli altri cluster. Questo suggerisce una certa varietà nei comportamenti di spesa.

In Porto (Regione 2), Cluster 2 ha una presenza inferiore, indicando che i clienti di Porto nel Cluster 2 sono meno frequenti.

Nella regione Altro (Regione 3), Cluster 2 è presente ma non domina.

-Distribuzione per Canale:

Nei clienti hotel (Canale 1), Cluster 2 è ben rappresentato, suggerendo che i clienti di questo cluster tendono a spendere di più o ad acquistare prodotti più specializzati.

Nei clienti della vendita al dettaglio (Canale 2), Cluster 2 è meno dominante, ma ancora presente.

-Interpretazione:

Cluster 2 sembra rappresentare clienti con comportamenti di spesa più variati, in particolare a Lisbona e nel canale hotel. Questi clienti potrebbero essere più propensi a comprare una gamma di prodotti più ampia o con comportamenti d'acquisto più sofisticati, forse legati ad acquisti specializzati o occasionali come Fresh e Frozen.

-Cluster 3 (Blu):

-Distribuzione per Regione:

In Lisbona (Regione 1), Cluster 3 è ben distribuito, indicando una clientela con comportamenti di spesa molto differenziati.

In Porto (Regione 2), Cluster 3 ha una presenza ridotta rispetto agli altri cluster.

Nella regione Altro (Regione 3), Cluster 3 è piuttosto presente, suggerendo che in questa regione ci sono clienti con comportamenti di spesa che si distinguono da quelli di Porto e Lisbona.

-Distribuzione per Canale:

Nei clienti hotel (Canale 1), Cluster 3 è ben rappresentato, suggerendo che i clienti di questo cluster abbiano comportamenti d'acquisto specifici e diversificati.

Nei clienti della vendita al dettaglio (Canale 2), Cluster 3 è presente ma non dominante, indicando che questo cluster tende a essere più diversificato in termini di comportamenti d'acquisto.

-Interpretazione:

Cluster 3 rappresenta un gruppo di clienti con comportamenti di spesa molto variegati, in particolare nelle regioni Altro e nel canale hotel. Questo cluster potrebbe rappresentare clienti che fanno acquisti più occasionali di ogni genere.

Il mio progetto ha dimostrato l’efficacia dell’uso combinato di PCA e k-means per la segmentazione dei dati. Attraverso la riduzione della dimensionalità, sono riuscita ad identificare tre cluster distinti che rappresentano gruppi di dati con caratteristiche simili. Questo mi ha permesso di ottenere una visione chiara delle differenze e delle somiglianze tra le varie regioni e canali.

I risultati ottenuti suggeriscono che esistono segmenti ben definiti nel dataset, che possono essere utilizzati per sviluppare strategie mirate e ottimizzare le decisioni aziendali. La validità del modello di clustering è stata confermata dalla chiara separazione dei cluster, indicando che le tecniche utilizzate sono appropriate per questo tipo di analisi.

* I clienti di Lisbona hanno tendenze di spesa diversificate, con una buona rappresentanza dei tre cluster, suggerendo una clientela più variegata.
* A Porto, il cluster 1 sembra dominare, indicando che i clienti hanno comportamenti di spesa più simili tra loro.
* Vendita al dettaglio tende ad avere una maggiore concentrazione di clienti nel cluster 1, mentre gli hotel, ristoranti e bar presentano più variazione nei cluster.