



Walmart

Riccardo Baratto
Matricola: 2129899

Strumenti statistici per l'analisi di dati aziendali



Walmart

Walmart è una multinazionale statunitense che gestisce la più grande catena al mondo di grandi magazzini e negozi al dettaglio, nota per i prezzi competitivi e l'ampia varietà di prodotti disponibili.

Fondata nel 1962, è diventata nel tempo anche un gigante dell'e-commerce, capace di competere con Amazon nel mercato statunitense.

E' presente in 22 paesi.



Dataset iniziale

Il dataset è stato trovato su Kaggle, e riguarda un campione delle vendite dell'e-commerce dell'azienda.

Il dataset presenta 555.068 osservazioni e 10 variabili.

Ogni osservazione fa riferimento a una transazione e contiene delle informazioni riguardanti il cliente.



Dataset iniziale

User_ID -> ID utente (numerica)

Product_ID -> ID prodotto (fattore con 3631 livelli)

Gender-> Genere (fattore con 2 livelli)

Age -> Età (fattore con 7 livelli)

Occupation -> Professione (fattore con 21 livelli)

City_category -> Categoria della città (fattore con 3 livelli)

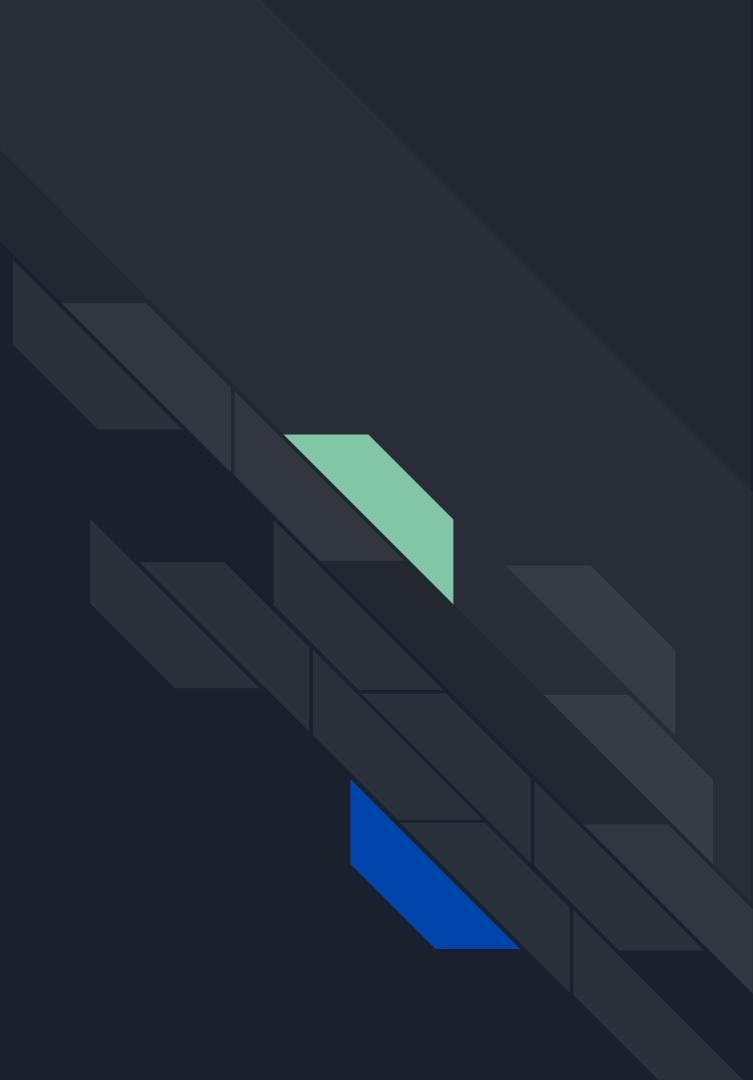
Stay_In_Current_City_Years-> Numero di anni di permanenza nella città attuale (fattore con 5 livelli)

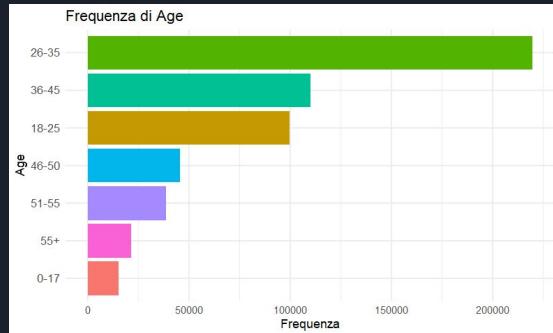
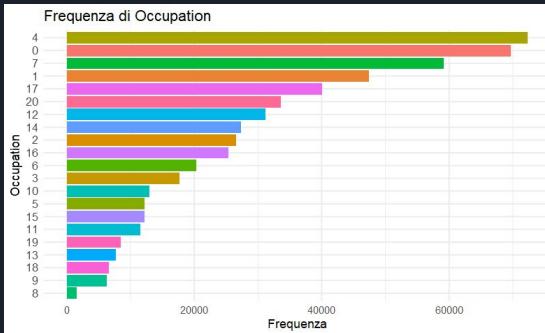
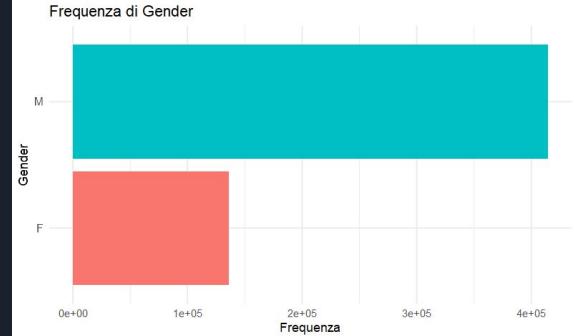
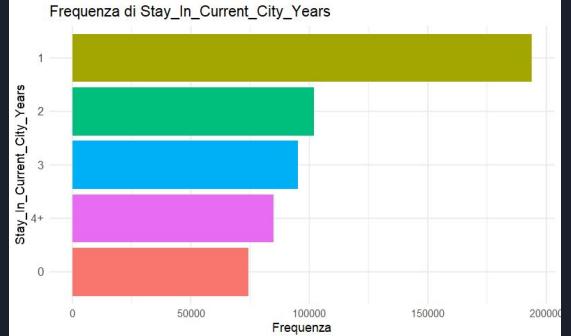
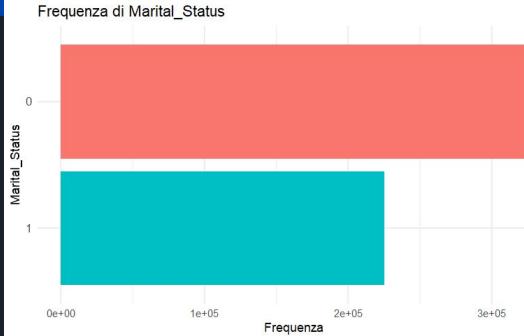
Marital Status-> Stato civile (fattore con 2 livelli)

Product_Category -> Categoria prodotto (fattore con 20 livelli)

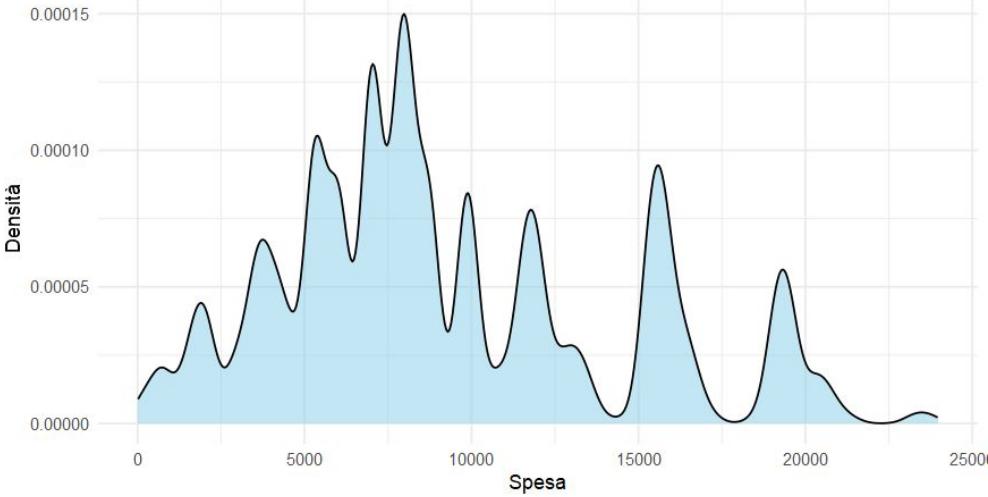
Purchase -> Costo del prodotto

Analisi esplorativa

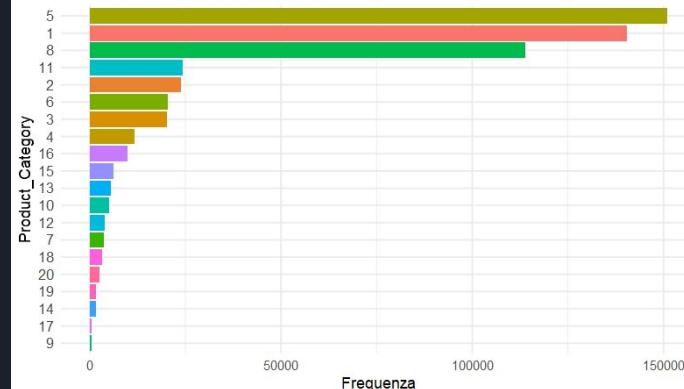




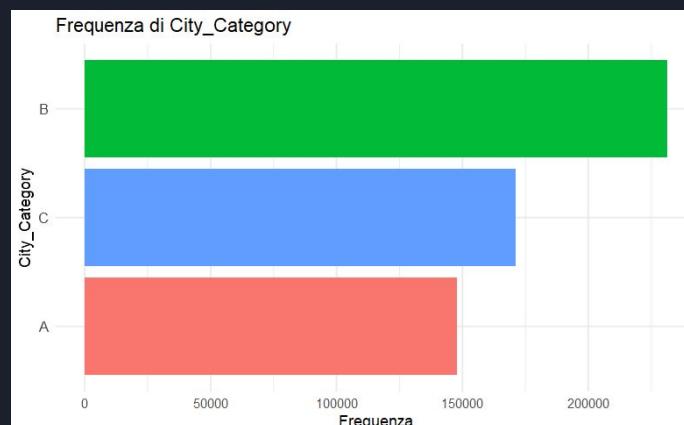
Distribuzione della spesa



Frequenza di Product_Category



Frequenza di City_Category





Domande di ricerca

- 01 Cosa influenza la spesa media di ogni utente?
- 02 Esistono gruppi di clienti che mostrano pattern simili di spesa?
- 03 Quali combinazioni di prodotti vengono acquistate più frequentemente all'interno di ciascun gruppo?



Domanda di ricerca 01

Si è interessati ad analizzare quali variabili influenzano la spesa media di ogni utente.

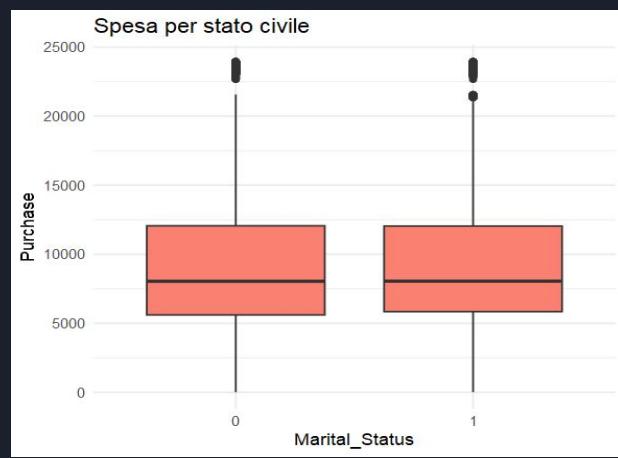
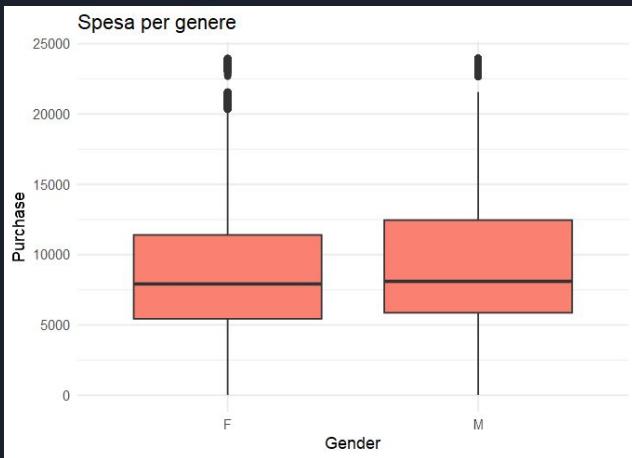
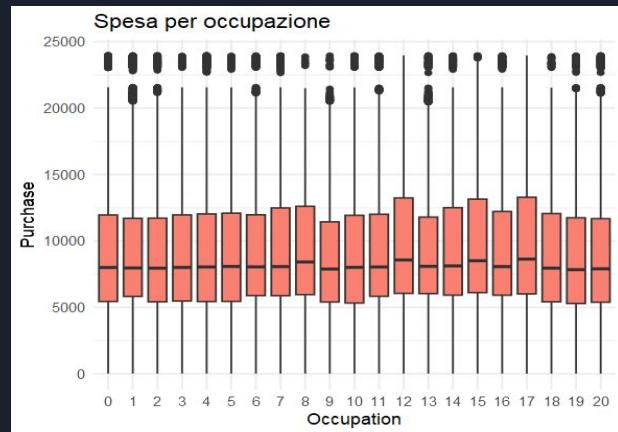
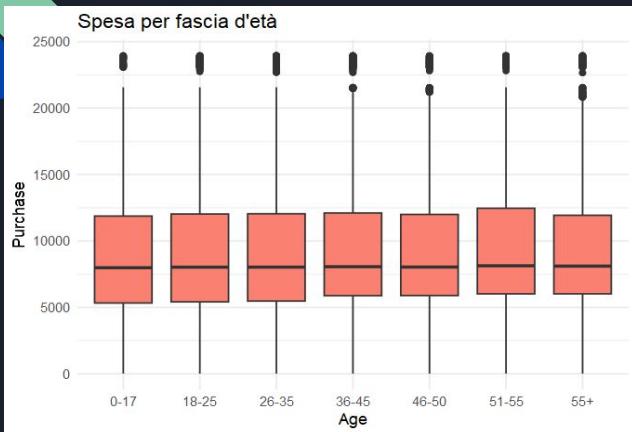
A tal fine, sono stati costruiti dei modelli con l'obiettivo di stimare la spesa media.

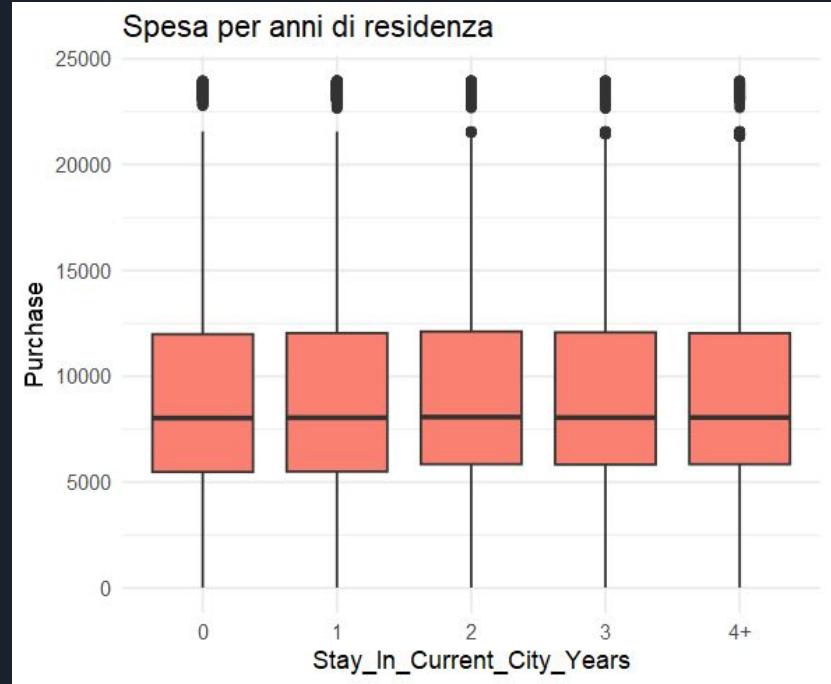
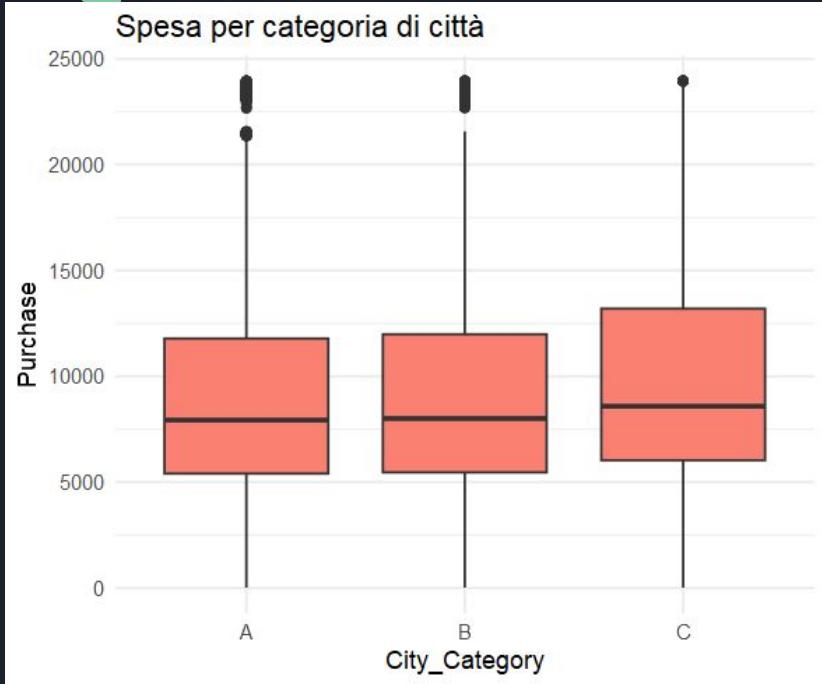
Dal dataset originale è stata eliminata la variabile *Product_ID* e dopo è stata effettuata un'aggregazione per *User_ID* e *Product_Category*, in modo da ottenere, per ciascuna combinazione, la spesa media.

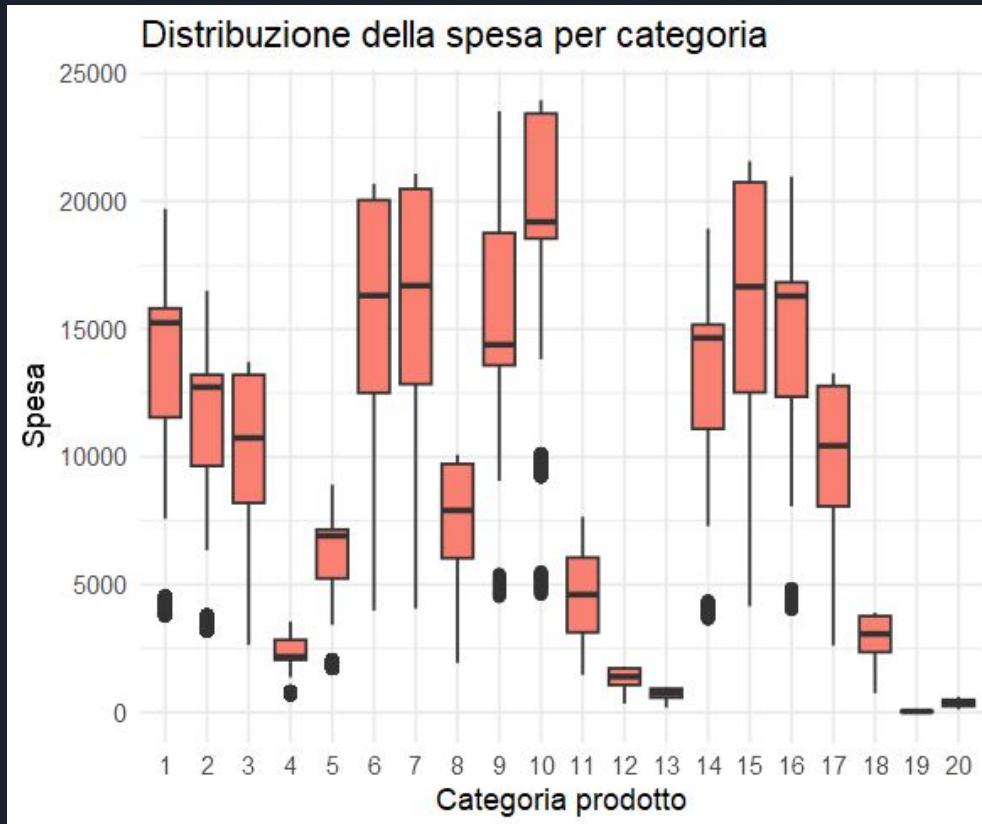
Questa quantità costituirà la variabile risposta dei modelli.

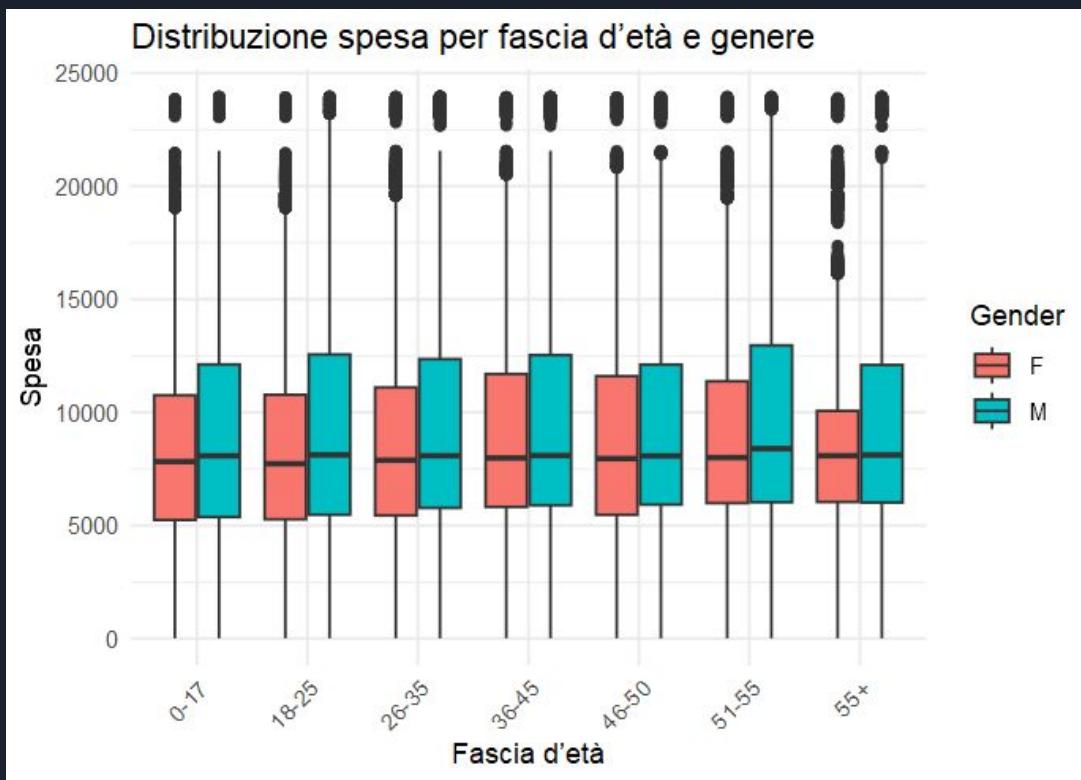
Dopo l'aggregazione, il dataset risultante comprende 56.782 osservazioni.



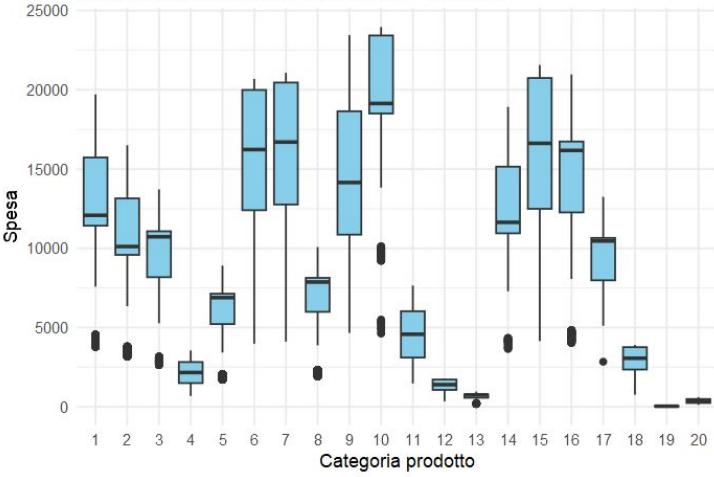




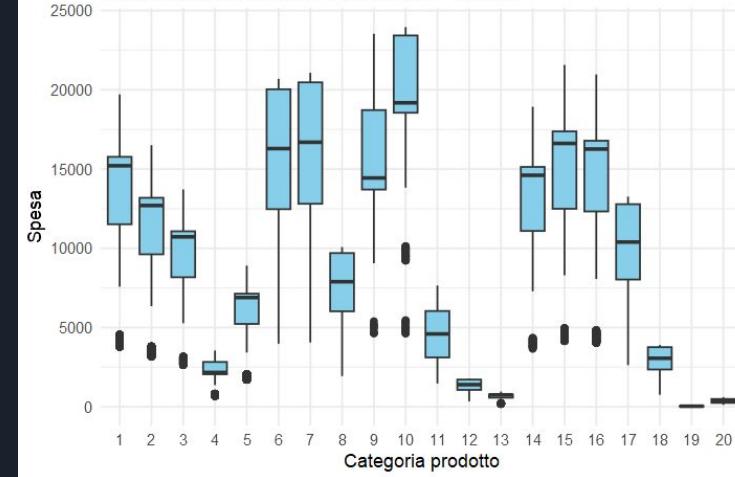




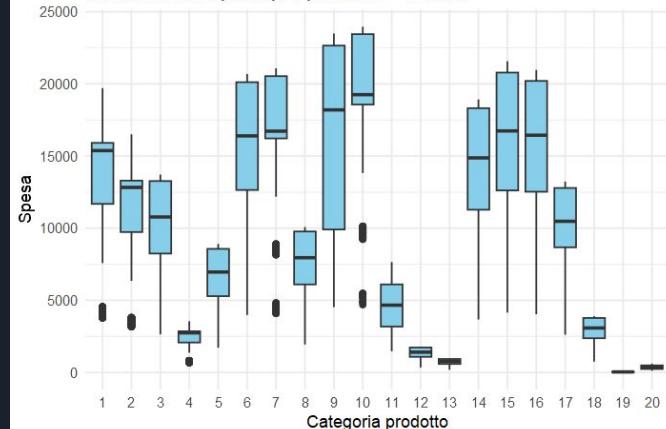
Distribuzione spesa per prodotto – Città A



Distribuzione spesa per prodotto – Città B



Distribuzione spesa per prodotto – Città C



Dataset domanda di ricerca 01

Viene creato un insieme di stima contenente l'80% delle osservazioni(45427) osservazioni e un insieme di verifica contenente il 20% delle osservazioni(11355)



Modelli domanda di ricerca 01

I modelli stimati sono:

- modello lineare con stepwise
- modello ridge
- MARS
- Random forest
- Boosting
- Bagging
- XGBoost

I modelli vengono confrontati attraverso l'RMSE



Confronto modelli

Modello	RMSE
Regressione lineare	2363
Regressione Ridge	2370
MARS	2382
Random Forest	2641
Boosting	4202
Bagging	2540
XGBoost	2387



Modello lineare

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	13988.589	108.133	129.365	< 2e-16	***
Product_Category2	-2442.789	52.442	-46.581	< 2e-16	***
Product_Category3	-3888.031	53.957	-72.059	< 2e-16	***
Product_Category4	-11650.192	56.452	-206.375	< 2e-16	***
Product_Category5	-7655.827	48.388	-158.219	< 2e-16	***
Product_Category6	1886.979	53.491	35.276	< 2e-16	***
Product_Category7	2145.511	75.674	28.352	< 2e-16	***
Product_Category8	-6480.182	48.575	-133.406	< 2e-16	***
Product_Category9	1253.010	134.135	9.341	< 2e-16	***
Product_Category10	5470.265	63.717	85.853	< 2e-16	***
Product_Category11	-9280.949	55.045	-168.606	< 2e-16	***
Product_Category12	-12735.885	74.147	-171.765	< 2e-16	***
Product_Category13	-13330.639	64.368	-207.101	< 2e-16	***
Product_Category14	-703.852	91.177	-7.720	1.19e-14	***
Product_Category15	901.045	62.629	14.387	< 2e-16	***
Product_Category16	1298.748	57.489	22.591	< 2e-16	***
Product_Category17	-4062.995	134.251	-30.264	< 2e-16	***
Product_Category18	-11067.356	81.371	-136.011	< 2e-16	***
Product_Category19	-14057.601	72.948	-192.708	< 2e-16	***
Product_Category20	-13710.306	61.535	-222.804	< 2e-16	***

Modello lineare

Occupation1	4.416	48.945	0.090	0.928108	
Occupation2	-9.240	60.378	-0.153	0.878373	
Occupation3	281.003	70.560	3.982	6.83e-05	***
Occupation4	160.773	48.177	3.337	0.000847	***
Occupation5	-34.572	85.980	-0.402	0.687621	
Occupation6	227.887	65.316	3.489	0.000485	***
Occupation7	12.268	46.145	0.266	0.790342	
Occupation8	-807.495	215.666	-3.744	0.000181	***
Occupation9	-16.850	101.582	-0.166	0.868252	
Occupation10	-198.460	106.332	-1.866	0.061990	.
Occupation11	206.190	78.964	2.611	0.009026	**
Occupation12	121.984	53.276	2.290	0.022046	*
Occupation13	29.723	90.262	0.329	0.741931	
Occupation14	118.839	58.989	2.015	0.043951	*
Occupation15	307.191	75.469	4.070	4.70e-05	***
Occupation16	-7.496	62.291	-0.120	0.904209	
Occupation17	139.150	49.775	2.796	0.005183	**
Occupation18	-17.374	108.567	-0.160	0.872859	
Occupation19	-379.173	102.231	-3.709	0.000208	***
Occupation20	7.757	58.022	0.134	0.893654	

Modello lineare

Age18-25	-251.646	99.690	-2.524	0.011596	*
Age26-35	-128.852	99.805	-1.291	0.196698	
Age36-45	45.613	101.345	0.450	0.652656	
Age46-50	12.325	105.124	0.117	0.906666	
Age51-55	198.295	105.998	1.871	0.061386	.
Age55+	138.996	110.856	1.254	0.209905	
City_CategoryB	-52.825	31.263	-1.690	0.091086	.
City_CategoryC	202.050	29.908	6.756	1.44e-11	***

Il modello è stato ottenuto utilizzando lo stepwise, il quale ha selezionato tutte le variabili tranne “Stay_in_current_city_years”, “Gender” e “Marital_Status”.

Come riferimento per le variabili fattoriali si sono presi:

City_Category -> A

Age -> 0-17

Product_Category -> 1

Occupation -> 0



Domanda di ricerca 02

Si è interessati a esplorare l'esistenza di gruppi omogenei di clienti che presentano comportamenti di acquisto simili.

A tal fine è stato effettuato un clustering, con l'obiettivo di segmentare gli utenti.

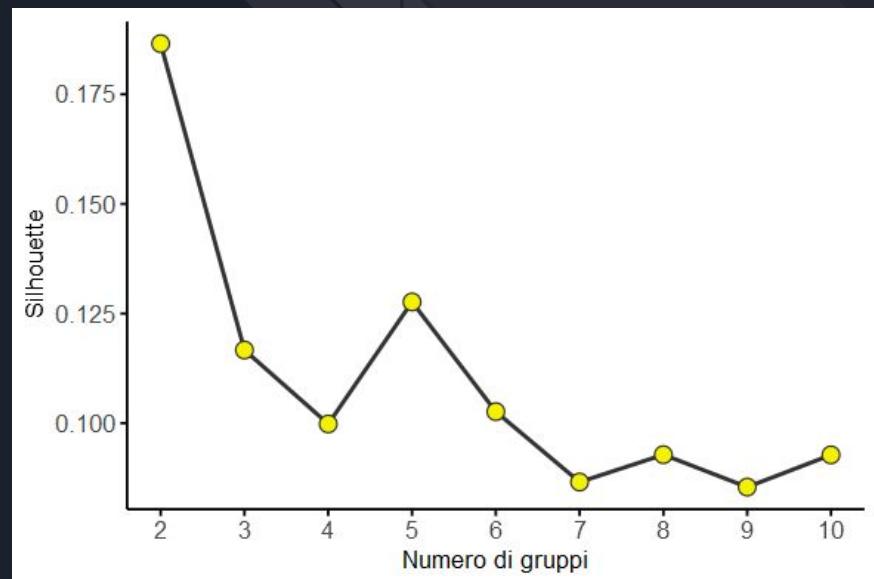
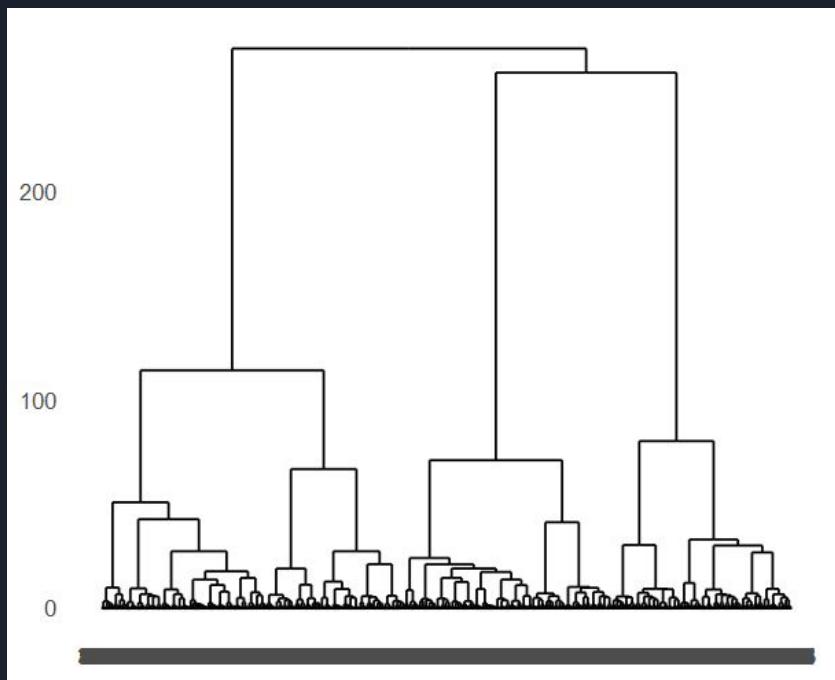
Dataset domanda di ricerca 02

Il dataset viene aggregato per User_id e viene calcolata la media degli acquisti per ogni utente.

Così facendo il dataset diventa composto da 5891 osservazioni.



Dendogramma e silhouette



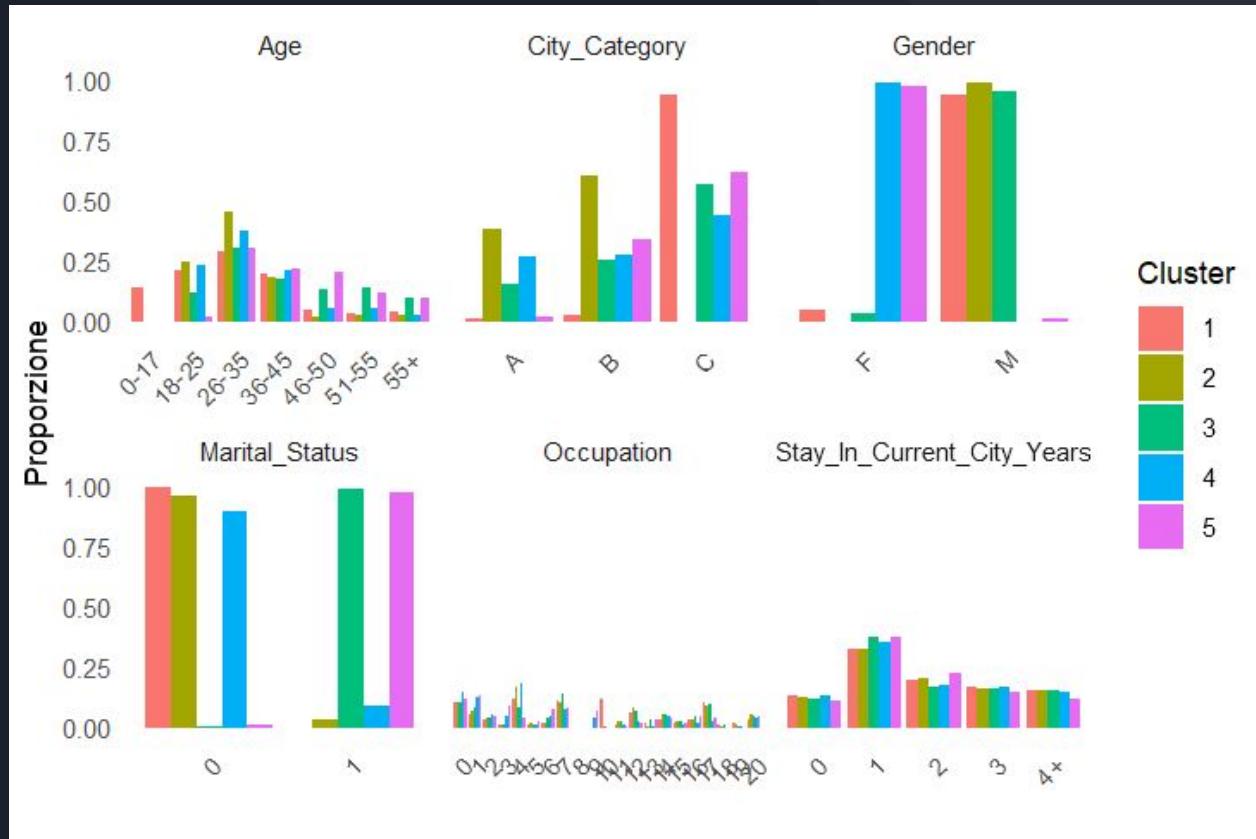
Numero di cluster

Il dendrogramma suggerisce una suddivisione compresa tra 3 e 5 gruppi, mentre l'indice di silhouette indica come ottimale una soluzione a 2 cluster. Tuttavia, si è scelto di adottare una segmentazione in 5 cluster, in quanto rappresenta un massimo locale della silhouette all'interno di una tendenza complessivamente decrescente, seguita da un netto calo. Questo per bilanciare la qualità della segmentazione e interpretabilità dei gruppi.

Caratteristiche dei gruppi - variabili quantitative

Cluster	Media	SD
1	9837	2057
2	9691	1743
3	9751	1877
4	8965	1631
5	9027	1735

Caratteristiche dei gruppi - variabili qualitative



Cluster 1

La fascia più rappresentata è 26-35 anni (30%), seguita da 18-25 (22%) e 36–45 (20%).

La maggior parte degli utenti vive nella città di categoria C (95%).

Fortemente dominato da maschi (95%), con una piccola percentuale di femmine (5%).

Quasi tutti sono non sposati (99%).

L'occupazione più frequente è la 10 (12%). Le altre sono progressivamente meno comuni.

Prevalentemente vivono in quella città da 1 anno (33%), seguita da 2 anni (20%), 3 anni (17%), 4+ anni (16%) e 0 anni (14%).

Questo cluster presenta una spesa media elevata, pari a 9837, indicando una spesa relativamente alta.

Cluster 2

La fascia d'età più rappresentata è quella tra i 26 e i 35 anni (46%), seguita da 18–25 anni (26%) e 36–45 anni (19%).

La maggior parte degli utenti risiede in città di categoria B (61%), seguita da quelle di categoria A (38%).

Il cluster è fortemente dominato da uomini (99.4%) e quasi totalmente composto da persone non sposate (99%).

L'occupazione più frequente è la numero 4, che rappresenta circa il 17% del totale.

La maggior parte degli utenti vive nella città attuale da un anno (32%).

La spesa è leggermente inferiore rispetto al cluster 1, con minore variabilità. Gli utenti in questo gruppo tendono a mantenere spese più stabili e leggermente più basse.

Cluster 3

La fascia d'età più rappresentata è quella tra i 26 e i 35 anni (30%), seguita da 36–45 anni (18%) e 51–55 anni (15%). Le fasce più giovani (18–25) e quelle oltre i 55 anni sono molto meno presenti.

La maggior parte degli utenti risiede in città di categoria C (57%), seguita da quelle di categoria B (26%).

Il cluster è quasi interamente composto da uomini (96%) e da persone sposate (99,5%).

L'occupazione più frequente è la numero 7 (14%).

La durata di permanenza più frequente è 1 anno (37%).

Simile ai cluster 1 e 2, con spesa media intorno a 9.750. Questo cluster mostra un comportamento di spesa medio-alto livello.

Cluster 4

La fascia d'età più rappresentata è 26–35 anni (38%), seguita da 18-25 (24%) e 36-45 (22%). Le fasce più anziane (oltre i 45 anni) sono presenti ma meno frequenti.

La maggioranza vive in città di categoria C (45%), seguita da categoria B (28%) e categoria A (27%).

Il cluster è quasi esclusivamente femminile (99,5%), con il 90% delle persone non sposate.

L'occupazione più frequente è la numero 4 (19%).

La permanenza nella città è più spesso di 1 anno (36%), seguita da 2 anni (18%) e 3 anni (17%).

Cluster con spesa media più bassa rispetto ai primi tre, con minore variabilità.

Cluster 5

La fascia d'età più rappresentata è 26–35 anni (31%), seguita da 36-45 (22%) e 46-50 (21%). Le età sono distribuite abbastanza uniformemente, con una leggera prevalenza delle fasce adulte e mature.

Quasi tutti vivono in città di categoria C (63%), con pochissimi utenti nelle categorie A (3%) .

Il cluster è quasi esclusivamente femminile(99%).

La stragrande maggioranza è sposata (98%).

Le occupazioni più comuni sono la 1 (13%)e la 0 (12%).

La permanenza in città è più spesso di 1 anno (38%), seguita da 2 anni (23%) e 3 anni (15%).

Cluster simile al 4, con spesa leggermente superiore ma sempre inferiore ai primi tre cluster.



Domanda di ricerca 03



Si è interessati a esplorare quali combinazioni di prodotti vengono acquistate più frequentemente all'interno di ciascun gruppo.

Queste informazioni risultano particolarmente utili in un'ottica di targeting mirato, poiché permettono di identificare pattern ricorrenti di co-acquisto all'interno dei cluster.

Regole associative - Cluster 1

lhs	rhs	support	confidence	coverage	lift	count
{11, 2}	{15}	0.24	0.54	0.45	1.50	344
{11, 6}	{15}	0.22	0.53	0.41	1.48	308
{11, 13}	{6}	0.20	0.90	0.23	1.47	287
{11, 3}	{15}	0.21	0.51	0.40	1.43	292
{13, 8}	{6}	0.25	0.87	0.29	1.41	356

Regole associative - Cluster 2

lhs	rhs	support	confidence	coverage	lift	count
{15, 18}	{13}	0.22	0.86	0.25	1.70	256
{13, 4}	{18}	0.24	0.60	0.42	1.70	283
{10, 18}	{13}	0.21	0.84	0.25	1.67	252
{10, 13}	{18}	0.21	0.56	0.38	1.64	252
{16, 18}	{13}	0.23	0.83	0.28	1.64	270

Regole associative - Cluster 3

lhs	rhs	support	confidence	coverage	lift	count
{10, 16}	{13}	0.20	0.71	0.30	1.80	362
{10, 15}	{13}	0.21	0.71	0.30	1.80	366
{13, 15}	{10}	0.21	0.77	0.27	1.76	366
{15, 4}	{13}	0.22	0.69	0.32	1.74	386
{15, 16}	{13}	0.20	0.68	0.30	1.74	363

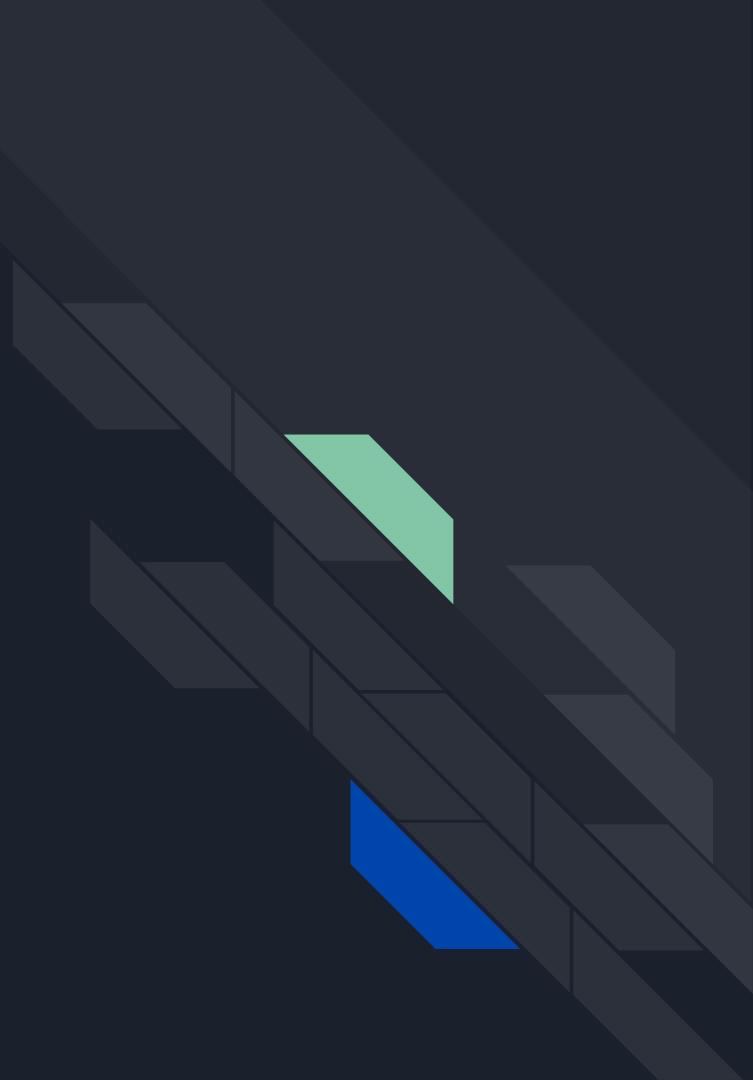
Regole associative - Cluster 4

lhs	rhs	support	confidence	coverage	lift	count
{15, 6}	{13}	0.21	0.75	0.28	1.95	196
{13, 2}	{15}	0.20	0.62	0.32	1.91	191
{13, 6}	{15}	0.21	0.61	0.34	1.88	196
{15, 2}	{13}	0.20	0.71	0.28	1.87	191
{1, 13}	{15}	0.22	0.60	0.37	1.81	210

Regole associative - Cluster 5

lhs	rhs	support	confidence	coverage	lift	count
{1, 10}	{13}	0.21	0.65	0.32	1.95	123
{13, 5}	{10}	0.21	0.64	0.33	1.93	122
{1, 13}	{10}	0.21	0.63	0.33	1.93	123
{13, 8}	{10}	0.21	0.63	0.33	1.93	123
{10, 5}	{13}	0.21	0.65	0.32	1.92	122

Conclusioni



Fattori che influenzano la spesa media

L'analisi ha evidenziato che il modello stepwise è il modello più performante, seguito da ridge e MARS.

Sono state selezionate tutte le variabili tranne
“Stay_in_current_city_years”,
“Gender” e “Marital_Status”.

L'analisi mostra che la categoria di prodotto è il fattore principale che determina la spesa media: alcune categorie sono associate a una spesa significativamente più alta, mentre altre riducono fortemente l'importo speso.

L'occupazione influisce moderatamente: alcune professioni tendono a spendere di più, altre meno, suggerendo differenze di potere d'acquisto o preferenze tra gruppi lavorativi.

L'età gioca un ruolo limitato, con i più giovani (18-25 anni) che spendono in media meno rispetto agli altri, mentre le differenze tra le altre fasce d'età non sono rilevanti.

Infine, la città di residenza ha un piccolo impatto: chi vive nella città di categoria “C” spende leggermente di più rispetto alla città di riferimento, mentre l'effetto per la categoria “B” è trascurabile.



Segmentazione clientela

Cluster 1: Uomini tra i 26-35 non sposati, residenti in città di categoria C, con una spesa media alta.

Cluster 2: Uomini tra i 26-35 non sposati, residenti in città di categoria B, con una spesa media medio-alta.

Cluster 3: Uomini tra i 26-35 sposati, residenti in città di categoria C, con una spesa media alta.

Cluster 4: Donne tra i 26-35, non sposate, residenti in città di categoria C, con una spesa media medio-bassa.

Cluster 5: Donne tra i 26-35 sposate, residenti in città di categoria C, con una spesa media media-bassa.



Regole associative

I cinque cluster mostrano chiaramente abitudini di acquisto correlate, con confidenze tra il 50% e il 90% e lift fino a 1.95. Questo indica che clienti in ogni gruppo tendono a comprare certi prodotti insieme con frequenza superiore al caso.

In particolare, i cluster 2 e 3 evidenziano forti pattern di cross-selling (confidenza > 70%, lift > 1.7), suggerendo ottime opportunità per offerte bundle mirate.

Anche i cluster 4 e 5 confermano questa tendenza con lift elevati (fino a 1.95), sottolineando il potenziale di campagne promozionali personalizzate per aumentare la spesa media e la fidelizzazione.



Strategia di Marketing

I cluster mostrano che non esiste un unico “cliente tipo”, ma piuttosto segmenti distinti con esigenze, capacità di spesa e preferenze di acquisto differenti.

I cluster 1 e 3 sono quelli che spendono di più in media, mentre il cluster 4 è quello che spende meno.

Le città di categoria B rappresentano il mercato più redditizio, però la spesa media nelle varie città varia di pochissimo.



Strategia di Marketing

Si potrebbe definire delle offerte o promozioni mirate, sviluppando dei bundle basati sui prodotti più associati.

Adattare il messaggio promozionale in base al profilo dell'utente, usando i canali di comunicazione più adatti ad ogni segmento.

Implementare dei programmi di fedeltà che premiano acquisti ripetuti e tramite l'adozione di bundle.

Offrire dei vantaggi esclusivi per incentivare il cross-selling.

Grazie per l'attenzione!

