UNIVERSITA' POLITECNICA DELLE MARCHE

FACOLTA' DI INGEGNERIA

Corso di Laurea triennale in Ingegneria Gestionale

Clustering di clienti e data visualization tramite PowerBI

Customer clustering and data visualization using PowerBI

Relatore:

Prof. **Primo Zingaretti**

Correlatore:

Dott. **Rocco Pietrini**

Tesi di Laurea di:

Luca Roganti

A.A. **2022** / **2023**

INDICE

Abstract

La tesi utilizza i concetti teorici del clustering e della data visualization per analizzare i clienti di un'azienda di moda.
Grazie ai due strumenti principali Power Bi e Python viene effettuata una cluster analysis in low coding al fine di tracciare le tendenze e i comportamenti dei vari segmenti.
Il risultato finale coincide con l'identificazione di cinque gruppi di clienti, ognuno dei quali con determinate abitudini e caratteristiche, che sono messe in risalto dall'utilizzo di visualizzazioni e grafici appartenenti al mondo della data visualization.

1.Introduzione

Nel contesto dell'attuale era digitale, caratterizzata da un'enorme mole di informazioni disponibili, la capacità di raccogliere e gestire i dati è diventata un elemento fondamentale per il successo di un'azienda. L'esplosione dei dati generati dalle interazioni degli utenti, dalle transazioni commerciali, dai dispositivi connessi e dalle piattaforme digitali ha aperto nuove opportunità, ma ha anche presentato sfide significative per le organizzazioni di ogni settore.

Le aziende di successo hanno compreso che i dati rappresentano una preziosa risorsa aziendale e che la loro corretta gestione può fornire un vantaggio competitivo significativo.

La raccolta e l'analisi dei dati consentono di ottenere una visione approfondita delle attività aziendali, dei comportamenti dei clienti, delle tendenze di mercato e delle prestazioni operative. Queste informazioni possono essere utilizzate per prendere decisioni informate, identificare nuove opportunità di crescita e ottimizzare le operazioni aziendali.

La raccolta dei dati non riguarda solo la quantità, ma anche la qualità e la pertinenza delle informazioni. Le aziende devono adottare strategie e strumenti adeguati a garantire l'integrità, l'accuratezza e la completezza dei dati raccolti. Inoltre, devono essere in grado di gestire efficacemente la crescente complessità dei dati, provenienti da diverse fonti e in vari formati.

La gestione dei dati richiede anche un'attenzione particolare alla sicurezza e alla privacy. Con l'aumento delle minacce cibernetiche e delle normative sulla protezione dei dati, è essenziale adottare misure adeguate a proteggere le informazioni sensibili e garantire la conformità alle leggi e ai regolamenti applicabili.

L'obiettivo di questa è tesi è stato quello di esplorare strumenti moderni di data analytics e visualization in ottica di Business Intelligence per un noto brand internazionale nel settore della moda. L'elaborazione dei dati forniti dall'azienda ha permesso di unire strumenti di data visualization moderni a tecniche di Machine Learning per analisi più avanzate come il clustering dei clienti.

Nello specifico si partirà introducendo l’argomento generale della business inteligence, per poi nel capitolo due parlare delle basi teoriche delle tecniche e

degli strumenti adoperati per la raccolta e l'analisi dei dati. Nel terzo capitolo si approfondiranno i materiali e metodi specifici che hanno segnato lo studio del caso in esame, dedicando attenzioni ai problemi e le soluzioni adottate. Nel capitolo 4 verranno poi presentati i risultati, sottoforma di grafici accompagnati da descrizioni esplicative, per poi chiudere nel quinto capitolo con le conclusioni ed un accenno ai possibili sviluppi futuri.
Attraverso questa analisi, sarà possibile comprendere appieno come i dati siano diventati una risorsa di valore inestimabile per le aziende moderne e come una gestione attenta e oculata di questa risorsa possa contribuire al successo e alla sostenibilità delle organizzazioni nel contesto digitale in continua evoluzione.
Nell'ambito della gestione dei dati ci concentreremo soprattutto sul metodo della cluster analysis e sull'utilizzo dello strumento Power Bi, due realtà appartenenti ed affini al mondo più ampio della business inteligence.

1.1 La business inteligence

La Business Intelligence (BI) è un insieme di processi, tecnologie e strumenti che consentono alle aziende di acquisire, gestire, analizzare e presentare dati aziendali in modo da supportare la presa di decisioni informate. L'obiettivo principale della business intelligence è quello di fornire informazioni significative e pertinenti alle diverse funzioni aziendali, come la gestione, il marketing, le vendite, le operazioni e le finanze.
La BI coinvolge l'estrazione dei dati da diverse fonti, come database aziendali, sistemi transazionali, applicazioni esterne e fonti di dati esterne. I dati vengono quindi trasformati e modellati in un formato che può essere facilmente analizzato e interpretato. Questo processo di trasformazione e modellazione dei dati è spesso denominato "data warehousing".
Una volta che i dati sono stati trasformati, vengono utilizzati strumenti e tecniche di analisi per scoprire modelli, tendenze e relazioni nei dati stessi. Ciò può comportare l'utilizzo di metodi statistici, algoritmi di apprendimento automatico e altre tecniche avanzate per estrarre informazioni significative.
Infine, i risultati dell'analisi vengono visualizzati attraverso dashboard, report e altre forme di rappresentazione visiva.

Questo consente agli utenti aziendali di comprendere facilmente le informazioni e prendere decisioni basate sui dati in modo tempestivo.
Rispondiamo alla domanda ‘perché un’azienda dovrebbe affidarsi alla business inteligence?’.
Un'azienda dovrebbe affidarsi alle analisi di business intelligence per cinque principali motivi:

1.Migliorare la presa di decisioni: l'analisi di business intelligence consente di raccogliere, elaborare e analizzare grandi quantità di dati aziendali provenienti da diverse fonti. Questi possono fornire informazioni preziose sulle tendenze del mercato, il comportamento dei clienti, le prestazioni delle vendite e molto altro ancora. Con queste informazioni a disposizione, i dirigenti possono prendere decisioni più informate e basate su dati concreti.

2.Identificare opportunità di mercato: l'analisi di business intelligence può aiutare a individuare nuove opportunità di mercato. Attraverso l'analisi dei dati dei clienti, ad esempio, un'azienda può identificare segmenti di clientela potenzialmente redditizi o individuare nuove esigenze dei clienti che possono essere soddisfatte con nuovi prodotti o servizi. Questo consente all'azienda di rimanere competitiva e di anticipare le esigenze del mercato.

3.Ottimizzare i processi aziendali: la business intelligence può contribuire a migliorare l'efficienza operativa e ridurre i costi. Attraverso l'analisi dei processi aziendali e dei dati relativi alle prestazioni, un'azienda può individuare aree di inefficienza o spreco e prendere provvedimenti per migliorare tali processi. Ad esempio, può ottimizzare la gestione dell'inventario, migliorare la catena di approvvigionamento o ridurre i tempi di produzione.

4.Monitorare le prestazioni aziendali: le tecniche del mondo della business intelligence consentono di monitorare le prestazioni aziendali in tempo reale. Attraverso l'uso di dashboard e report personalizzati, i dirigenti possono tenere traccia dei principali indicatori di performance (KPI) e valutare il raggiungimento degli obbiettivi. Questo consente loro di intervenire tempestivamente se si verificano deviazioni rispetto ai piani o se si evidenziano aree di miglioramento.

5. Migliorare la comprensione dei clienti: Attraverso l'analisi dei dati dei clienti, come ad esempio i modelli di acquisto, le preferenze e le interazioni sui canali digitali, un'azienda può ottenere una comprensione più approfondita del proprio pubblico di riferimento. Questo può aiutare a personalizzare le offerte di prodotti o servizi, migliorare l'esperienza del cliente e aumentare la fedeltà dei clienti.
I risultati riportati in questa tesi riguarderanno soprattutto quest'ultimo punto, in quanto si è basata l'analisi su informazioni (origini di dati) riguardanti i clienti già in contatto con l'azienda.

2. Stato dell'arte

Questa tesi è il risultato di uno studio basato sulle caratteristiche di clienti appartenenti ad una azienda dell'alta moda italiana.
Come già accennato in precedenza, lo studio ha riguardato due tematiche principali: l'analisi di clustering e la data visualization.
Il tema del clustering dei clienti è stato già affrontato da Zaramella Luca con la sua tesi dal titolo 'Cluster Analysis per la segmentazione della clientela utilizzando il Software SAS® ENTERPRISE MINER™ '.
Alcuni concetti teorici importanti che si possono estrarre sono: l'importanza del cliente, il suo ciclo di vita e la segmentazione.
L'azienda è un sistema aperto e dinamico fortemente influenzato da variabili esterne, per avere successo il vantaggio competitivo di conoscere i bisogni del cliente è fondamentale e per conoscerli bisogna saper analizzare i dati a disposizione.
Il ciclo di vita di ogni cliente sarà composto da quattro fasi: prospect (potenziale cliente), responder (coloro che si dimostrano interessati all'azienda), clienti effettivi (coloro che utilizzano il prodotto) ed ex clienti (tutti quelli che non usano più il prodotto o abbandonano il servizio) [1].
Questa tesi si occupa dei clienti effettivi e di ex clienti, in quanto sono le uniche categorie che sono effettivamente entrate in contatto con l'azienda e di cui quindi si hanno dati a disposizione, ma è importante conoscere tutte le fasi per capire a pieno il loro comportamento.
Per quanto riguarda invece la data visualization, una buona disamina è stata effettuata da Riccardo Terenzi in "Progettazione e realizzazione di una campagna di data analytics a supporto delle attività di vendita di un calzaturificio".
In essa vengono approfonditi i temi di data analytics e data visualization, nel caso specifico di un'azienda calzaturiera.
"È fondamentale che gli utenti aziendali siano capaci di comprendere i risultati dell'analisi e di fornire un feedback. Si cerca di spingere gli utenti a porsi domande che non si erano fatti, o a guardare i risultati da un punto di vista

differente. Per far ciò è fondamentale scegliere il tipo di visualizzazione più adatto, in quanto quest'ultimo può influenzare fortemente l'utente" [2].

2.1 Clustering

Il termine cluster indica, generalmente, un gruppo che può essere composto da una serie di elementi, di solito, molto omogenei tra loro o, in generale, accomunati da caratteristiche comuni (features).
La definizione di gruppi omogenei è utile in tutti quei casi in cui vi sia la necessità di:

- ridurre la complessità dei dati rispetto alle unità
- riunire i dati in maniera significativa e per mezzo di metodi quantitativi
- scoprire i legami esistenti tra casi
- costruire sistemi di classificazione automatica che consentono di immagazzinare informazioni, documenti, ecc.; nelle scienze biologiche ciò viene definito tassonomia
- esplorare i dati in una forma grafica che sia semplice, sintetica e intuitiva

Il clustering inteso come processo è un insieme di tecniche di analisi multivariata volte alla selezione e raggruppamento di elementi omogenei in un insieme di dati.
Quando si parla di analisi multivariata si intende una metodologia statistica che valuta l'effetto di un insieme di variabili indipendenti o predittive su una (o più) variabili dipendenti o di esito.
Più semplicemente il risultato dell'analisi che stiamo effettuando dipenderà da più dati o variabili.
Per esempio, se studiamo il meteo di una città, sappiamo che il risultato finale dipenderà dall'insieme delle varianti: precipitazioni, umidità, inquinamento, vento ecc.

Questo testo approfondisce la cluster analysis soprattutto riguardo al suo utilizzo in azienda e più nello specifico nel mondo del marketing, ma in realtà il

clustering è un processo che viene applicato in svariati ambiti, vengono di seguito elencati alcuni esempi significativi.
Un mondo nel quale il clustering è di fondamentale importanza è la medicina. Come spiegano molti medici, i pazienti molto spesso sono affetti da più patologie, la così detta comorbilità.

“I pazienti di medicina interna sono per lo più anziani, con multiple co-morbilità, solitamente croniche. L’alta prevalenza di co-morbilità e la multi-morbilità - secondo le rispettive definizioni - ha un impatto significativo sia sull’esito clinico che sul trattamento e la gestione di queste persone. Il clustering è il processo d’inquadramento nosografico che raggruppa in associazioni significative alcune condizioni morbose con una malattia indice, in modo da evidenziare le possibili congruenze (ma anche le incongruenze), con tutte le interconnessioni e le possibili interferenze esistenti. Le nostre decisioni dovrebbero essere assunte, oltre che sui problemi clinici immediati, anche in base ai fattori contestuali, che devono essere presi in considerazione, con una saggia selezione delle priorità” [3].

Da ciò capiamo come anche la nostra salute è studiata con la tecnica dell’analisi multivariata e come ogni risultato e scelta in questo campo siano frutto della concomitanza di moltissime variabili.
Successivamente anche gli istituti di statistica come Istat ed Eurostat utilizzano la cluster analysis per aiutare gli scienziati sociali a fotografare meglio le nostre società.
Ad esempio, è stato richiesto a 110 studenti universitari, provenienti da due atenei italiani, selezionati tra le Facoltà di Scienze Motorie e Medicina, di compilare un questionario riguardo il loro stile di vita.
I dati raccolti sono poi stati analizzati tramite clustering partizionale ed il risultato in output sono stati due gruppi.
Il gruppo 1 caratterizzato principalmente da soggetti maschili della Facoltà di Scienze Motorie con stile di vita attivo e rapporti affettivi familiari distaccati; il gruppo 2 prevalentemente composto da soggetti femminili della Facoltà di Medicina con stile di vita maggiormente sedentario e relazioni affettive stabili. Inoltre, è stata confermata la correlazione tra bassi valore di ‘AGEs’ (un marker

biomedico che indica l'assunzione di cibi dannosi) e stili di vita salutari [4](figura 1).

Tab. 5- Student Groups

Sporty Students (Group 1)	**Sedentary Students (Group 2)**
Men	Women
Report spending more time more time with the family	Report spending less time with family
Tend to be more satisfied with the time spent with the family	Tend to be more dissatisfied with family time and would like to spend more time at home
Coming from Foro Italico University	Coming from the Sapienza University
Less smokers	Slightly more smokers
Mixed composition of students and workers	Mostly just students
Have slightly fewer stable relationships	Tend to have stable relationships
Graduates	High school diploma
BMI above average	BMI below average
Age below average	Age mostly above average

Figura 1, Confronto tra Sportly Students e Sedentary Students

Un altro ambito applicativo del clustering è il mondo dello sport, un caso di applicazione molto interessante riguarda il calcio.

Un' azienda che si occupa principalmente di football analytics, ovvero la Soccerment, ha reinventato i classici ruoli (portiere, difensore, centrocampista, attaccante) mediante l'uso del clustering.

L'idea principe da cui si sviluppò tutto fu che 'il ruolo non è più una posizione, ma una funzione' [5].

Infine, il clustering caratterizza anche il settore moda, nel corso di questa tesi si è posta l'attenzione sui clienti, ma abbiamo esempi di come le moderne tecniche di machine learning possano migliorare servizi forniti dall'azienda, ad esempio velocizzare i resi, risolvere il problema della stagionalità dei capi e migliorare l'esperienza online del cliente.

Si possono approfondire questi temi nell'articolo 'Multi Clustering Recommendation System for Fashion Retail'.

"In this paper, a recommendation solution in the context of fashion retail is proposed. The aim has been to solve the above-mentioned problems of cold start, computational complexity, low number of returns in the shops of fashion retails and long period for returning, the needs of more mediated interactions in

the shops and more direct interactions online, and the effects of the seasonality of products" [6].
Un ultimo esempio relativo ad aziende del mondo fashion è descritto nella tesi Advanced Analytics per il Marketing: clustering dei clienti fidelizzati.
"Questo elaborato descrive un progetto di cluster analysis svolto per un'azienda di fashion retail. Lo scopo di questo progetto è clusterizzare i clienti fidelizzati dell'azienda di fashion retail e testare varie piattaforme che permettono analisi avanzate, in particolare il clustering. Le clusterizzazioni sono state effettuate tramite il linguaggio R, sono stati sperimentati diversi algoritmi di clustering, tenendo in considerazione efficienza e qualità del risultato" [7].
Questo esempio è quello più vicino al progetto svolto, sono stati scelti anche algoritmi di clustering simili, una notevole differenza sta invece nel linguaggio di programmazione adoperato.

2.2 Classificazioni del clustering

È possibile identificare ben 15 criteri di classificazione [8] dei metodi di raggruppamento, occorre però dire che, solo alcuni di tali criteri vengono considerati utili all'atto pratico, in particolare si considerano principalmente le seguenti classi:

1. clustering partizionale/clustering gerarchico
2. hard clustering/soft clustering
3. clustering agglomerativo/clustering divisivo
4. clustering sequenziale/clustering simultaneo
5. clustering incrementale/clustering non incrementale
6. clustering eterogeneo/clustering omogeneo

1.Clustering partizionale/clustering gerarchico

Il clustering partizionale prevede la suddivisione del set di dati in un numero prefissato di cluster, in cui ogni oggetto(pattern) viene assegnato a uno e un solo cluster. In questa tecnica, i cluster vengono creati in modo iterativo,

partendo da una configurazione iniziale casuale e regolando i centroidi dei cluster in modo che la somma delle distanze tra i pattern e il centroide sia minima. Il clustering partizionale è una tecnica veloce e adatta a set di dati di grandi dimensioni, ma presenta alcuni svantaggi come la dipendenza dal numero di cluster prefissato, la sensibilità alla posizione iniziale dei centroidi e la limitazione alla scoperta di forme di cluster non sferiche.

Il clustering gerarchico, invece, crea una gerarchia di cluster in cui i dati vengono suddivisi in insiemi sempre più piccoli e dettagliati. Questa tecnica permette di scoprire forme di cluster complesse e di visualizzare la struttura gerarchica dei dati, ma può essere computazionalmente costosa e può richiedere una grande quantità di memoria.

In sintesi, il clustering partizionale e il clustering gerarchico sono due tecniche di clustering che si differenziano per il modo in cui creano i cluster e la struttura gerarchica dei dati. La scelta tra le due tecniche dipende dalle caratteristiche del set di dati e dagli obiettivi della ricerca, ma entrambe possono essere utili per scoprire pattern nascosti nei dati e fornire insight utili per la comprensione dell'argomento di studio (figura 2).

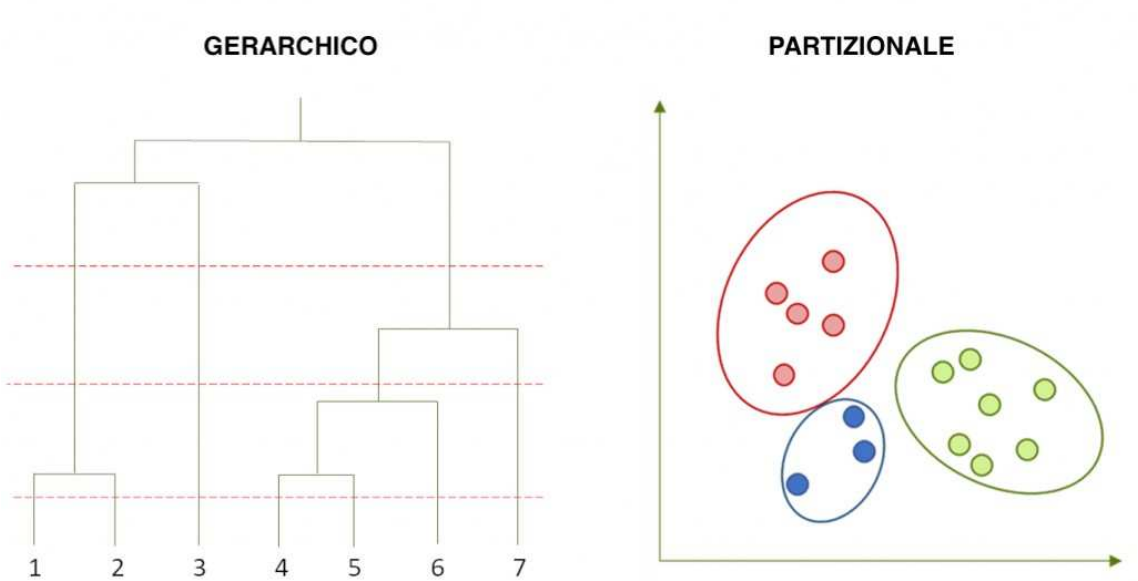


Figura 2,clustering gerarchico e partizionale, nel grafico del clustering gerarchico i cluster finali sono rappresentati dai numeri, mentre in quello partizionale sono contraddistinti dai diversi colori

2.Hard clustering/soft clustering

Nell'hard clustering un pattern può appartenere ad uno ed un solo cluster, viene anche chiamato cluster esclusivo. Al contrario nel soft clustering il pattern può essere in condivisione tra più insiemi ed è anche chiamato 'fuzzy clustering' o clustering non esclusivo. Per esempio, raggruppare le persone per data di nascita è esclusivo (ognuno ne avrà solo una), mentre raggruppare le persone per sport praticati non è esclusivo (ogni persona può appartenere a più insiemi che sono rappresentati degli sport). Inoltre, nel soft clustering ogni pattern può avere associata una 'membership', ovvero un punteggio che indica quanto appartenga ad un cluster, i pattern che si trovano graficamente ai bordi avranno un punteggio più basso di quelli in prossimità del centro.

3. Clustering agglomerativo/clustering divisivo

La differenza tra questi due gruppi sta nella modalità di formazione dei cluster. Nel primo caso si parte con tanti gruppi quanti sono gli elementi per poi unirli con le così dette operazioni di 'merge', ovvero la fusione. Nel secondo caso invece si inizia con un unico cluster contenente tutti i dati, e successivamente si divide i cluster fino al raggiungimento di una determinata condizione, in questo caso si effettueranno le operazioni di 'split', ovvero frazionamenti. Questa in realtà è una sottocategoria del clustering gerarchico in quanto come possiamo capire, in entrambi i casi il numero di clustering non viene scelto a priori.

4. Clustering sequenziale/clustering simultaneo

Nella modalità sequenziale i pattern vengono processati uno alla volta, mentre in quella simultanea tutti assieme. Quasi tutte le tecniche agglomerative sono sequenziali.

5. Clustering incrementale/clustering non incrementale

All' arrivo di nuovi dati il clustering può essere aggiornato in maniera incrementale oppure potrebbe essere necessario riesaminare l'intero dataset. La scelta è molto importante, in quanto negli ultimi anni abbiamo a che fare con dataset sempre più grandi e con moli di dati in espansione; perciò, non sempre è possibile procedere con la seconda opzione.

6.Clustering eterogeneo/clustering omogeneo

In un cluster eterogeneo i gruppi possono avere dimensioni, forme e densità molto diverse, mentre in uno omogeneo i diversi raggruppamenti saranno accomunati da medesima forma, densità e dimensioni.

2.3 Algoritmi di clustering

Definito il clustering, analizziamo la strada per arrivare all' output finale, ovvero i vari algoritmi che elencano i passi da seguire per poter associare ciascun punto dato ad un gruppo specifico.
Ci sono moltissimi algoritmi che possono essere utilizzati, è possibile classificarli sommariamente come di seguito specificato:

- metodi di partizionamento
- metodi gerarchici
- metodi basati sulla densità
- metodi basati sulla griglia
- metodi basati sul modello

Introduciamo un esempio appartenente ad ognuna delle categorie descritte, premettendo che i risultati di questa tesi sono stati ottenuti utilizzando l'algoritmo più diffuso, ovvero il K-means.

K-means

Il k-means [9] è un algoritmo di clustering partizionale ampiamente utilizzato in ambito di analisi dei dati. Questo algoritmo si basa sulla suddivisione del dataset in un numero predefinito di gruppi, dove ogni gruppo rappresenta una categoria di oggetti che sono il più simili possibile tra loro.
L'algoritmo funziona in modo iterativo. Inizialmente, viene selezionato un numero k di centroidi in maniera casuale all'interno del dataset, dove k è il numero di gruppi che si vogliono identificare.
Il centroide è un punto appartenente allo spazio degli attributi che media le distanze tra tutti i dati appartenenti al cluster, rappresenta quindi una sorta di baricentro del cluster ed in generale, non è uno dei punti del dataset.
Successivamente, ogni oggetto del dataset viene assegnato al gruppo il cui centroide è il più vicino in termini di distanza Euclidea.
Una volta che tutti gli oggetti sono stati assegnati a un gruppo, i centroidi di ogni gruppo vengono aggiornati calcolando la media aritmetica degli oggetti di ogni cluster. Quindi, il processo di assegnazione degli oggetti ai gruppi e di aggiornamento dei centroidi viene ripetuto in modo iterativo, fino a quando la posizione dei centroidi non si stabilizza ed i pattern vengono assegnati in modo stabile ai cluster.
Si tratta di un algoritmo molto efficiente per l'elaborazione di dataset grandi, perché la sua complessità computazionale è linearmente dipendente dalla dimensione del data set.
Il k-means è molto efficace per identificare cluster circolari o sferici, ma può presentare problemi se i cluster sono di forma irregolare. Inoltre, l'algoritmo è molto sensibile alla scelta iniziale dei centroidi e alla loro posizione all'interno del dataset.
Per mitigare questi problemi, esistono alcune varianti dell'algoritmo k-means, come il k-means++ e il k-medoids [10]. Il k-means++ migliora la scelta iniziale dei centroidi selezionandoli in modo intelligente, mentre il k-medoids utilizza un oggetto del dataset come centroide, invece della media aritmetica. Una 'modifica' che spesso viene apportata al k-means standard è quella di unire due

cluster se la distanza tra i due relativi centroidi è minore di una soglia prefissata, oppure se contengono pochi punti.

Gerarchico aggregato

Questo algoritmo, come intuibile dal nome appartiene alla famiglia dei clustering gerarchici [11], e funziona seguendo i passi successivamente descritti.
Inizialmente ogni punto è trattato come un singolo cluster, ovvero se ci sono X punti dati nel nostro set di dati, allora abbiamo X cluster. Poi viene selezionata una metrica di distanza che misuri la distanza tra due insiemi, ad esempio, utilizzeremo il collegamento medio.
Ad ogni iterazione, vengono combinati due cluster in uno. I due cluster da combinare sono selezionati come quelli con il legame medio più piccolo. Cioè, secondo la nostra metrica di distanza selezionata, questi due gruppi hanno la distanza più piccola tra loro; quindi, sono i più simili e dovrebbero essere combinati.
Quest'ultimo passaggio viene ripetuto finché non raggiungiamo la radice dell'albero, ovvero abbiamo un solo cluster che contiene tutti i punti dati. In questo modo possiamo selezionare quanti cluster vogliamo alla fine, semplicemente scegliendo quando smettere di combinare i cluster cioè quando smettiamo di costruire l'albero.
Il clustering gerarchico non richiede di specificare il numero di cluster e possiamo persino selezionare quale numero di cluster sembra migliore. Inoltre, l'algoritmo non è sensibile alla scelta della distanza metrica; tutti tendono a funzionare ugualmente bene mentre con altri algoritmi di clustering, la scelta della metrica della distanza è fondamentale.

Mean-shift

Un altro algoritmo (appartenente alla famiglia degli algoritmi basati sulla densità) utilizzato per la cluster analysis è il mean-shift [12].
L'algoritmo mean-shift funziona calcolando la densità di probabilità di ogni punto all'interno del dataset, in modo da identificare i picchi locali che rappresentano i centroidi dei cluster. Il processo si basa sulla traslazione

iterativa dei punti all'interno del dataset verso il picco locale di densità di probabilità più vicino, fino a quando la convergenza non viene raggiunta.
In pratica, l'algoritmo utilizza una finestra, che calcola la densità di probabilità di ogni punto all'interno del dataset. La finestra (detta finestra di parzen) viene centrata sul punto di riferimento e viene traslata iterativamente verso il picco locale più vicino, fino a quando il punto di riferimento non converge al centroide del cluster.
Una volta che tutti i punti del dataset convergono ai loro centroidi, l'algoritmo mean-shift restituisce i centroidi dei cluster e i punti del dataset associati a ciascun cluster.
L'algoritmo mean-shift è particolarmente utile in contesti in cui il numero di cluster e la loro forma sono sconosciuti, come nel riconoscimento dei pattern. Tuttavia, può presentare alcune limitazioni, come la sensibilità alla finestra utilizzata per calcolare la densità di probabilità, e la tendenza a produrre cluster di dimensioni disuguali.

Sting

I metodi di clustering basati sulla griglia quantizzano lo spazio in un numero finito di celle che formano una struttura su cui vengono effettuate tutte le operazioni.
Il principale vantaggio di questi approcci consiste nel loro tempo di elaborazione ridotto; tale tempo, infatti, è indipendente dal numero degli oggetti da clusterizzare essendo dipendente, soltanto, dal numero di celle in ciascuna dimensione dello spazio quantizzato.
La tecnica più diffusa appartenente a questa classe è lo Sting [13], acronimo di STatistical INformation Grid.
Vi sono, diversi livelli di celle rettangolari che corrispondono a diversi livelli di risoluzione, queste celle formano una struttura gerarchica; ciascuna cella ad un livello elevato viene partizionata per formare un certo numero di celle ad un livello immediatamente inferiore.

L’algoritmo pre-calcola e memorizza alcune informazioni statistiche (quali la media, il valore massimo e il valore minimo) relative agli attributi di ciascuna cella della griglia.
Quando i dati vengono caricati nel database, i parametri statistici delle celle a livello più basso vengono calcolati direttamente da essi. I valori della distribuzione possono essere specificati dall’utente, se sono noti a quest’ultimo, oppure possono essere frutto di test di ipotesi.
Il tipo di distribuzione di una cella a più alto livello può essere determinato sulla base della maggioranza dei tipi di distribuzione delle corrispondenti celle a più basso livello. Se le distribuzioni delle celle a più basso livello sono in disaccordo tra loro e nessuna di esse prevale nettamente sulle altre, il tipo di distribuzione della cella ad alto livello è posto ad un valore nullo.

Expectation Maximization (EM)

I metodi di clustering basati sul modello tentano di ottimizzare la corrispondenza tra i dati e un qualche modello matematico predefinito dall’utente.
L’ algoritmo model-based per eccellenza è l’Expectation Maximization di solito abbreviato con EM [14].
L’ EM può essere visto come un’estensione del paradigma k-means soltanto che, invece di assegnare ciascun oggetto ad un cluster in modo rigido, l’EM assegna ciascun oggetto ad un cluster secondo un peso che rappresenta la probabilità di appartenenza. L'algoritmo di EM funziona in modo iterativo e si basa sul principio di massimizzazione della likelihood, ovvero la produttoria di tutte le funzioni di probabilità.
Nella fase di Expectation, l'algoritmo di EM stima la probabilità che ogni punto del dataset appartenga a ciascun cluster. Questa stima viene effettuata utilizzando la distribuzione di probabilità attualmente stimata dei dati e la regola di Bayes.
la regola di Bayes permette di calcolare la probabilità di un'ipotesi o di un evento A, data l'osservazione di un'informazione B. Questo calcolo si basa sulla conoscenza della probabilità di B, la probabilità di A, e la probabilità di B dato A.
La formula della regola di Bayes è la seguente:

$$P(A \vee B) = P(B \vee A) * P(A) / P(B)$$

Nella fase di Maximization, l'algoritmo di EM aggiorna i parametri della distribuzione di probabilità in modo da massimizzare la likelihood del dataset. Questa stima viene effettuata utilizzando la stima della probabilità di appartenenza dei punti ai cluster ottenuta nella fase di Expectation.
Il processo di Expectation Maximization viene ripetuto iterativamente fino a quando la stima dei parametri della distribuzione di probabilità non converge.
L'algoritmo di EM è particolarmente utile in contesti in cui i dati seguono una distribuzione di probabilità nota, come la distribuzione normale o la distribuzione di Poisson. Tuttavia, l'algoritmo può presentare alcune limitazioni, come la sensibilità alla scelta dei parametri iniziali della distribuzione di probabilità e la tendenza a convergere ai minimi locali della likelihood.

2.4 Data visualization

La data visualization è il processo di rappresentazione visiva delle informazioni mediante grafici, diagrammi, mappe o altre forme visive. È una disciplina che combina elementi di design, statistica e storytelling per comunicare in modo efficace concetti complessi e dati numerici in modo chiaro e comprensibile.
“L’obiettivo principale della Data Visualization è permettere all’utente una comprensione qualitativa e semplice dei contenuti. Essa consiste nella trasformazione di oggetti, numeri e concetti in una forma che possa essere facilmente interpretata dall’occhio umano” [15].
La visualizzazione dei dati ha quattro scopi principali: generazione di idee, illustrazione di idee, scoperta visiva e visualizzazione quotidiana.
La visualizzazione dei dati stimola la creatività, attraverso la rappresentazione grafica dei dati, è possibile individuare modelli, tendenze o relazioni che potrebbero non essere evidenti nella forma grezza dei dati stessi.
La data visualization è uno strumento efficace per comunicare e illustrare idee complesse in modo chiaro e conciso. Trasformare i dati in grafici, diagrammi o mappe può semplificare concetti complessi e consentire una comprensione più rapida ed efficiente.

L'interattività delle visualizzazioni dei dati consente agli utenti di esplorare diverse prospettive e dettagli, spostandosi tra diversi livelli di aggregazione o filtrando i dati in base a determinati parametri. Questa scoperta visiva può portare a nuove intuizioni e approfondimenti sull'argomento in esame.
Infine, la visualizzazione dei dati è utile nel quotidiano in quanto semplifica la comprensione di numeri complessi e rende più facile il monitoraggio delle tendenze nel tempo o il confronto tra diverse variabili.
Uno degli strumenti più utilizzati per la visualizzazione dati è Power Bi [16].

2.5 Power Bi

Power BI è una piattaforma unificata e scalabile per la business intelligence aziendale e in modalità self-service. È possibile connettersi ai dati, visualizzarli ed inserire inoltre gli oggetti visivi nelle app usate giornalmente.
Power Bi è stato lo strumento fondamentale per la data visualization del caso di studio.
Power BI nasce nel 2006, come un progetto top secret con il nome di ‘Gemini’ da un'idea di Thierry D'hers e Amir Netz del team ‘SQL Server Reporting Services’ di Microsoft. Questo progetto aveva lo scopo di sfruttare la potenza di uno strumento chiamato ‘SQL Server Analysis Services’ (SSAS) per trasformarlo in un motore di archiviazione “in-memory”.
Gemini fu lanciato nel 2009 con il nome di ‘Power Pivot’ come estensione di Excel e non fu molto scaricato ed utilizzato fino a quando uno dei suoi creatori, Rob Collie, ne iniziò a parlare all’interno del suo blog online.
Successivamente fu introdotto quello che oggi è Power Query per la trasformazione dei dati (allora di chiamava Data Explorer), anche se ancora c’era un problema, ovvero quello di scambiarsi grandi moli di fogli Excel contenenti dati via e-mail.
Nel 2015 vennero aggiunti altri add-on per migliorare l’aggiornamento e la manipolazione dei dati ed il nome venne cambiato a quello odierno, queste due mosse fecero arrivare 500.000 nuovi iscritti in pochi giorni e la piattaforma divenne di dominio pubblico.

Esistono tre tipi di licenze di Power BI per gli utenti: Free, Pro e Premium Per User (PPU).
La scelta del tipo di licenza dipende da tre principali caratteristiche: la posizione in cui il contenuto viene archiviato, come si interagisce con quest'ultimo e se il contenuto usa la funzionalità (o sottoscrizione) Premium.
Free: gli utenti con licenze gratuite possono usare il servizio Power BI per connettersi ai dati e creare report e dashboard per il proprio uso.
Non possono usare le funzionalità di condivisione o collaborazione di Power BI con altri utenti o pubblicare contenuto nelle aree di lavoro di altre persone. Tuttavia, gli utenti pro e PPU possono condividere contenuto e collaborare con utenti Free se il contenuto viene salvato nelle aree di lavoro ospitate in capacità Premium.
Pro: Power BI Pro è una singola licenza per utente che consente agli utenti di creare contenuto e leggere e interagire con il contenuto pubblicato da altri utenti nel servizio Power BI.
Gli utenti con questo tipo di licenza possono condividere i risultati e collaborare con altri utenti che possiedono la medesima licenza.
Premium Per User (PPU): Una licenza PPU per utente fornisce al titolare della licenza tutte le funzionalità di Power BI Pro più l'accesso alla maggior parte delle funzionalità basate sulla capacità Premium.
Una licenza PPU di Power BI sblocca l'accesso a un'ampia gamma di funzionalità e tipi di contenuto disponibili solo tramite Premium. Questo accesso è limitato al titolare della licenza PPU e ad altri colleghi che hanno anche loro una licenza PPU.
Ad esempio, per collaborare e condividere il contenuto in un'area di lavoro PPU, tutti gli utenti devono avere una licenza PPU.
Infine, la 'sottoscrizione Premium' consente agli utenti Pro o PPU di creare e salvare il contenuto nelle aree di lavoro di capacità Premium. Possono quindi condividere l'area di lavoro con i colleghi che dispongono di qualsiasi tipo di licenza.

2.5.1 Report

Una delle principali funzionalità di PowerBI è la creazione di report.

Un report di Power BI è una visualizzazione multi-prospettiva in un set di dati, con oggetti visivi che rappresentano risultati e informazioni dettagliate da tale set di dati.
Un report può avere un singolo oggetto visivo o molte pagine piene di oggetti visivi. Si può notare un esempio di report nella figura 3.

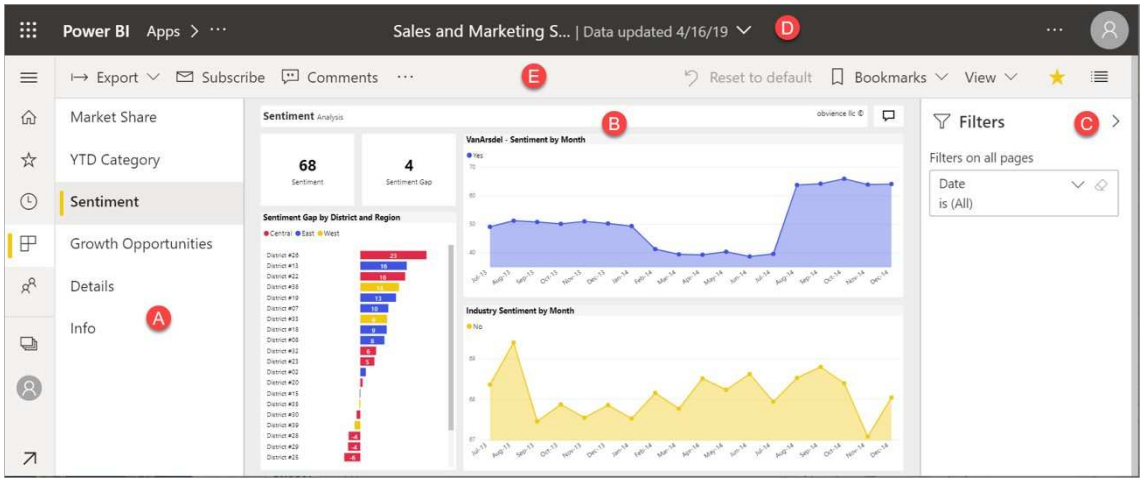


Figura 3, esempio di un report [17]

La creazione di un report in Power Bi avviene seguendo i successivi quattro passi:

1.connessione a origine dati
2.elaborazione dati
3.configurazione modello dati
4.creazione visualizzazioni

1.Connessione a origine dati

È possibile importare dati nell'applicazione per poi procedere con l'elaborazione, oppure avere una connessione diretta con l'origine, con il vantaggio di accedere ai dati in tempo reale.
Power Bi offre molteplici opzioni per la connessione a origini dati. La prima è l'utilizzo di fogli di calcolo Excel. Essi sono molto comuni e utilizzati tra gli utenti aziendali, ma allo stesso tempo hanno dimensioni limitate e sono soggetti a

potenziali errori umani nell'inserimento dati. La seconda opzione è la connessione a Database relazionali, in questo caso le dimensioni non sono un problema ed è possibile creare relazioni tra tabelle. L'ultima opzione è il collegamento a servizi cloud come SharePoint o Google Analytics. Qui i lati negativi sono rappresentati dai costi per l'utilizzo, mentre il loro vantaggio sta nell'aggiornamento automatico delle istanze.

2.Elaborazione dati

La fase di elaborazione dati comprende innanzitutto le operazioni di pulizia, quali la rimozione dei dati duplicati, la correzione di errori, la gestione dei valori mancanti e la conversione ad un formato uniforme.
Una volta 'pulito' il dataset si passa a trasformare i dati, questo include l'applicazione di filtri, la suddivisione di colonne, l'aggregazione dei dati, la creazione di colonne calcolate utilizzando formule e la combinazione di dati provenienti da diverse origini.
A questo punto se necessario è possibile definire relazioni tra le tabelle per creare un modello dati coerente e consentire analisi incrociate tra dati correlati. Le relazioni possono essere stabilite utilizzando chiavi primarie e chiavi esterne. Infine, l'ultima operazione dell'elaborazione dei dati è l'aggiunta di calcoli, si possono creare colonne calcolate sui dati esistenti (ad esempio l'aggiunta della colonna 'cluster').

3. Configurazione modello dati

Durante la configurazione del modello dati, vengono create nuove tabelle per organizzare i dati in base alle necessità specifiche dell'analisi. Le tabelle possono rappresentare entità o concetti distinti all'interno del dominio dei dati, come clienti, prodotti o vendite. Per ogni tabella, è possibile definire le colonne e i rispettivi tipi di dati e si possono definire gerarchie per organizzare le colonne in modo gerarchico, ad esempio anno, trimestre e mese. Infine, in questa fase c'è la possibilità di applicare regole di sicurezza per limitare l'accesso e la visibilità dei dati in base ai ruoli degli utenti.

4. Creazioni di visualizzazioni

Power BI offre una vasta gamma di visualizzazioni tra cui scegliere, come grafici a barre, grafici a torta, grafici a linee, mappe, tabelle e molti altri. La scelta della visualizzazione corretta dipende dal tipo di dati e dal messaggio che si desidera comunicare. Una volta selezionata una visualizzazione, è possibile configurarne le proprietà per adattarla alle esigenze specifiche. Ad esempio, si vanno a personalizzare i colori, le etichette, i titoli, gli assi e altre caratteristiche per migliorare l'aspetto visivo e l'usabilità.
Uno dei plus di Power Bi sta nell'interattività, perciò sono implementate funzioni per: filtrare i dati, selezionare punti specifici su un grafico per ottenere ulteriori dettagli e utilizzare funzionalità di drill-through per esplorare i dati a diversi livelli di dettaglio. Le varie rappresentazioni create possono essere racchiuse in dashboard, strumenti interattivi fondamentali per avere una panoramica immediata delle prestazioni aziendali.

2.6 Python e PyCaret

Power Bi tra le altre funzionalità permette l'integrazione del linguaggio di programmazione Python.
Attraverso l'uso di Power Query e il modulo "Python Script" all'interno di Power BI Desktop, è possibile scrivere codice Python per elaborare, filtrare, aggregare o trasformare i dati prima di importarli nel modello di dati di Power BI.
Questo consente agli utenti di sfruttare la potenza e la flessibilità di Python per manipolare i dati in modo avanzato prima di creare visualizzazioni interattive e dashboard.
Python [18] è un linguaggio di programmazione orientato agli oggetti noto per la sua chiarezza, potenza e flessibilità. Si tratta di un linguaggio interpretato, il che significa che un interprete legge ed esegue il codice direttamente, senza compilazione.

Python è inoltre facile da apprendere, comprendere e usare, con una sintassi pulita e uniforme.
La filosofia alla base della creazione di Python, infatti, si concentra principalmente sulla leggibilità e manutenibilità del codice.
Il linguaggio viene fornito con una vasta libreria standard per l'elaborazione di stringhe, protocolli Internet, unit testing, registrazione, profilazione e analisi del codice Python, e interfacce del sistema operativo. Inoltre, è possibile potenziare ed allargare le funzionalità di Python grazie alle estensioni di terze parti.
Un vantaggio di questo linguaggio è quello di godere di una comunità di sviluppatori molto attiva e collaborativa.
Ci sono numerosi forum di discussione, gruppi di utenti e risorse online che offrono supporto, consigli e risorse per aiutare gli sviluppatori a risolvere problemi e approfondire le loro competenze.
Una delle librerie più popolari in Python per l'analisi dei dati e la creazione di modelli è PyCaret [19].
PyCaret è una libreria open-source che semplifica notevolmente il processo di sviluppo e confronto di modelli di machine learning.
Offre un'interfaccia semplice per eseguire una vasta gamma di compiti, come la preparazione dei dati, la selezione delle caratteristiche, la creazione e la valutazione dei modelli, consentendo agli utenti di concentrarsi sulla logica del modello in modalità low coding.
Pycaret supporta algoritmi di machine learning per: compiti di regressione (es. regressione lineare, regressione logistica), classificazione (es. K-nearest neighbors (KNN)) rilevamento delle anomalie (es. One-Class Support Vector Machine (OCSVM)) e naturalmente clustering, nel corso di questa tesi è stato utilizzato proprio per il clustering dei clienti.

3.MATERIALI E METODI

3.1 Descrizione caso di studio

I dati utilizzati per lo studio in questione sono rappresentati da un dataset, sotto forma di foglio Excel denominato "ContactActive", messo a disposizione da un noto brand della moda italiana, attivo nella produzione di abbigliamento per uomo. Un approfondimento sul dataset è contenuto nella prossima sezione.
Lo strumento selezionato per le analisi e la data visualization è stato Power Bi, per tre principali motivi.

1. Robustezza e capacità di analizzare grande quantità di dati
2. Ampia gamma di visualizzazioni, con funzionalità interattive
3. Possibilità di effettuare la cluster analysis in Power Bi grazie all'implementazione di Python

Un aspetto negativo che si può citare di Power Bi è la disponibilità. Nello specifico è disponibile solo per sistemi operativi Windows e questo potrebbe essere un problema per chi dispone di altri sistemi operativi.
Ritornando alle motivazioni per cui ho scelto Power Bi, bisogna specificare che il processo di clustering sarebbe stato possibile anche seguendo altre strade, ma la libreria Pycaret ha il grande vantaggio di essere una libreria low coding e con poche righe di codice è stato possibile mantenere il giusto livello di astrazione, non entrare troppo nel dettaglio dei meccanismi della cluster analysis ed ottenere allo stesso modo il risultato desiderato.
La clusterizzazione in Power Bi sarebbe stata possibile anche senza l'utilizzo di Python, creando un grafico scatter plot e utilizzando la funzionalità 'trova clustering automaticamente' che colora ogni pattern del grafico a dispersione con il colore assegnato all'insieme di riferimento.
Il grande vantaggio dell'utilizzo di Python però sta nella possibilità di utilizzare molteplici features nell'algoritmo di clustering, mentre la funzionalità descritta, rappresenta un mero raggruppamento grafico bidimensionale, risultando quindi

banale e non adatto alle complessità richieste dal clustering dei consumatori, influenzati da molteplici variabili.

3.2 Dataset

Il dataset utilizzato è rappresentato da un export del database interno all'azienda, fornito dalla stessa sotto forma di file Excel (denominato ContactActive) con struttura tabellare, formato da 384353 righe (clienti) e 135 colonne (features), dati che sono stati raccolti in un arco temporale di 7 anni e che rappresentano i clienti dell'azienda sotto forma di caratteristiche demografiche, d'acquisto e comportamentali.
In realtà, delle 135 features, non tutte sono state rilevanti per il risultato finale. Alcune sono state scartate perché non ritenute importanti, altre invece perché ridondanti (ad esempio la data di nascita non ci dà nessuna informazione aggiuntiva rispetto all'attributo 'age' già esistente).
Le features scelte sono state le seguenti:

FEATURE	DESCRIZIONE
Age	età del cliente
Age group	età del cliente, raggruppata per decadi (21-30,31-40)
Average ticket	importo medio dello scontrino in euro
Best store	negozio preferito dal cliente in termini di importo totale speso nell'ultimo anno
Country of residence	paese di residenza
Loyalty	indice che classifica la fedeltà del cliente in 12 livelli (01. Subscriber, 02. Prospect Cs, 03. Prospect Store, 04. New, 05. New Loyal, 06. Occasional Retained, 07. Occasional Reactivated, 08. Loyal Reitaned, 09. Loyal Reactivated, 10. Sleeper, 11. Inactive, 12. Lost)
Macro Area	macroregione di appartenenza
RFM Frequency	indice sulla frequenza di acquisto

RFM Monetary	indice sul valore speso dal cliente
RFM Recency	indice basato sul tempo trascorso dall'ultimo acquisto del cliente
RFM Score	indice che combina i valori di RFM Frequency (F), RFM Monetary (M) e RFM Recency (R)
Total Amount	spesa totale effettuata dal cliente
Total Amount ly	spesa relativa all'anno precedente
Total Amount cy	spesa relativa all'anno corrente
Total return quantity	quantità di resi effettuata
Total transaction	numero transizioni per cliente
E-mail cliccate	numero di e-mail aperte dal cliente
E-mail ricevute	numero di e-mail ricevute dal cliente
E-mail spedite	numero di e-mail spedite dal cliente

Tabella 1, descrizione features

3.3 Metodi

In figura 4 è mostrata la flowchart seguita, ovvero la sequenza di operazioni effettuate per arrivare ai report finali.

I quattro blocchi rappresentati sono le quattro fasi principali (in ordine cronologico), che a loro volta comprenderanno altre 'sotto-operazioni'.

L'obbiettivo sarà quello di aggiungere una colonna al dataset di partenza, che identifica il gruppo di appartenenza di ogni cliente, da lì si procederà con la data visualization.

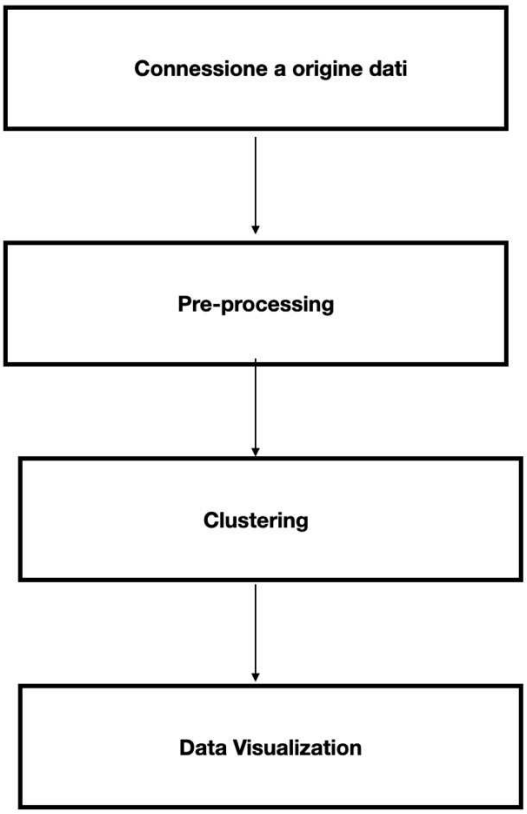


Figura 4, flowchart

Connessione a origine di dati

La connessione a origine di dati di PowerBI consente di importare direttamente il dataset utilizzato "ContactActive"(figura 5).

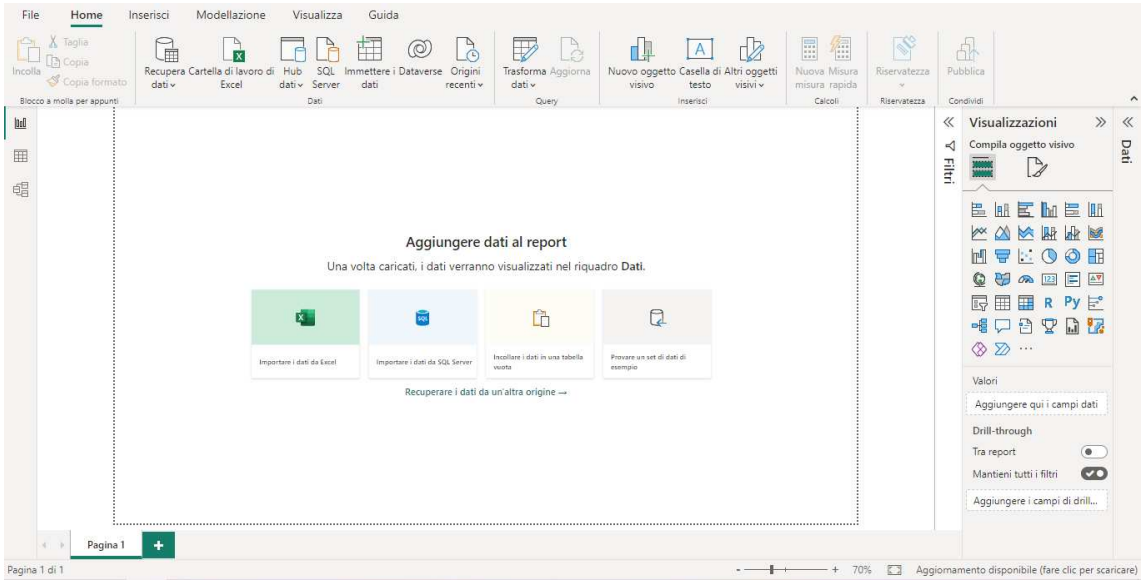


Figura 5, connessione a origine dati

Una volta selezionato il foglio Excel è possibile caricarlo in Power Bi Desktop o trasformare i dati. Nel caso di studio è stato necessario questo passaggio intermedio per effettuare le operazioni di pulizia del dataset (figura 6).

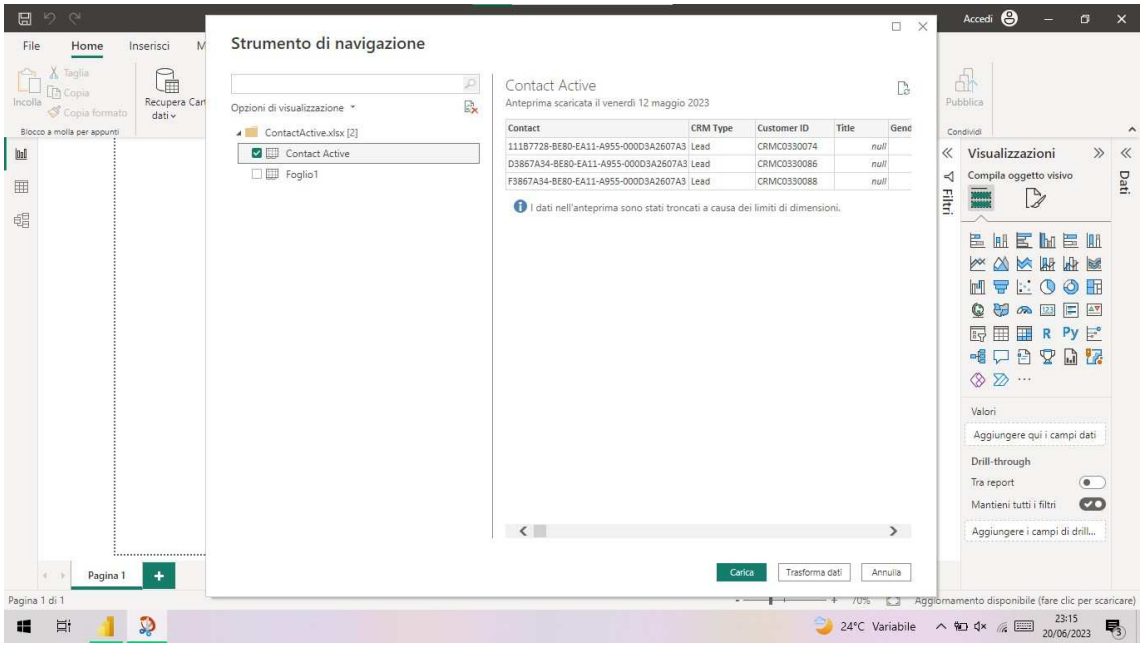


Figura 6, scelta dataset da importare

Quello che è stato ottenuto a valle della fase di importazione del dataset è la schermata mostrata in figura 7, ovvero l'intero dataset di partenza all'interno dell'Editor di Power Query, da dove inizia la fase del pre-processing.

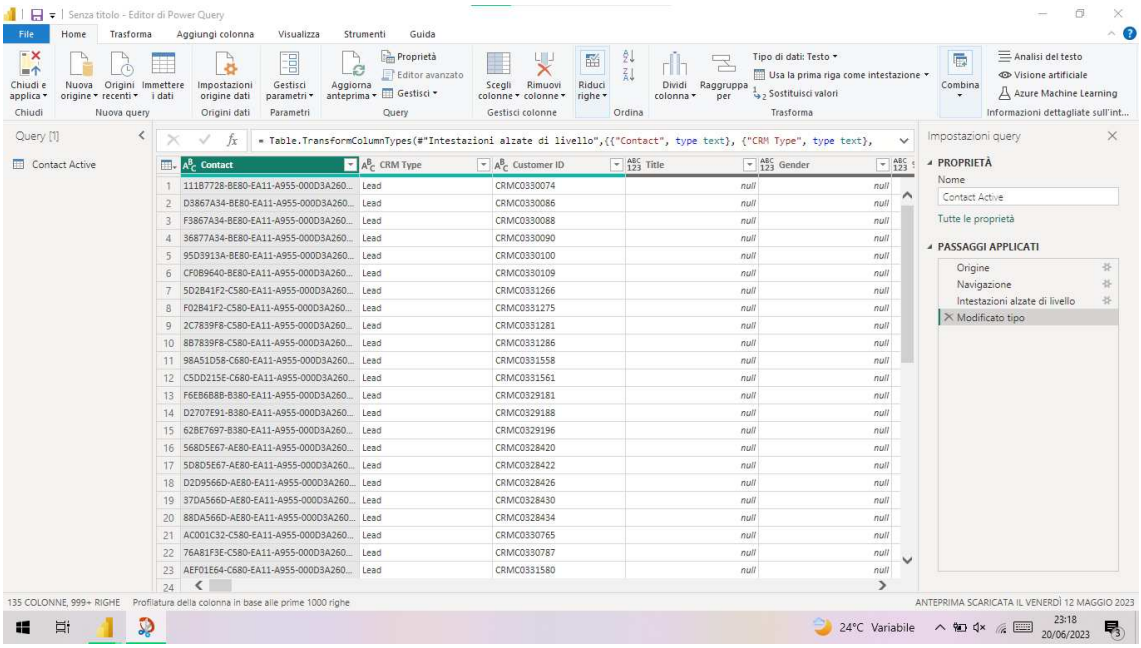


Figura 7, dataset nell'Editor Power Query

Pre-processing

Le operazioni che caratterizzano il pre-processing sono tre: la scelta delle colonne o features rilevanti per l'analisi, la scelta delle righe e l'eliminazione dei valori nulli.

Le colonne scelte sono stata già elencate nella sezione 3.2 riguardante il dataset, a livello pratico selezioneremo dal pannello visibile in figura 8 le colonne da mantenere.

Per quanto riguarda la scelta delle righe, sono state scelte solo quelle con valore 'customer' in corrispondenza dell'attributo CRM Type, in quanto nel dataset iniziale erano presenti anche soggetti come i dipendenti dell'azienda, ma l'analisi è stata concentrata solo sulla figura del cliente.

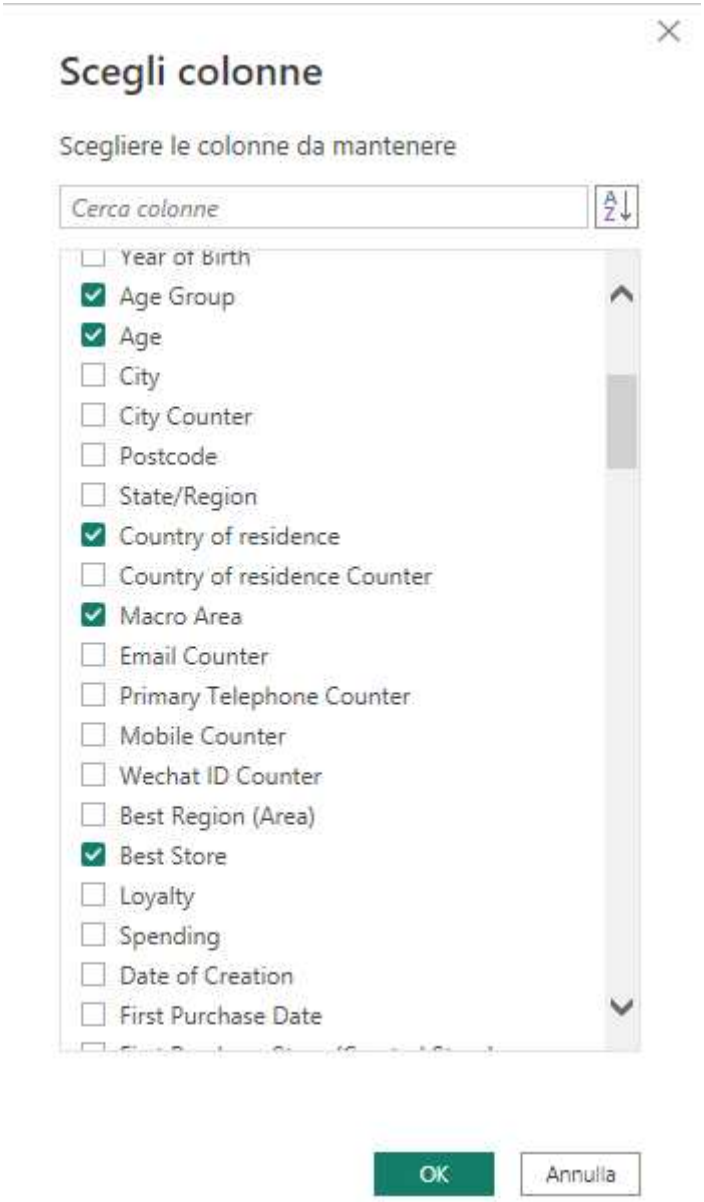


Figura 8, pannello scelta features

Per quanto riguarda i valori nulli, in un set di dati essi possono influenzare significativamente le prestazioni degli algoritmi di apprendimento automatico per diverse ragioni:

Precisione: gli algoritmi di apprendimento automatico sono modelli matematici che richiedono un input per produrre un output. Se non c'è input (il valore è nullo), l'output non può essere calcolato con precisione, portando a previsioni meno accurate.

Addestramento e Test: molti algoritmi di apprendimento automatico non supportano dati con valori mancanti e, se questi valori nulli non sono gestiti correttamente, gli algoritmi potrebbero fallire. Inoltre, i valori nulli nel set di dati di test potrebbero portare a risultati distorti.
Sbilanciamento dei Dati: i valori nulli possono causare uno sbilanciamento nel set di dati, in particolare se i valori nulli sono più presenti in una classe di dati rispetto ad un'altra. Questo può portare il modello ad essere parziale verso la classe con meno valori nulli.
Rappresentazione delle features: i valori mancanti possono distorcere lo spazio rappresentativo delle features, riducendo la qualità complessiva della comprensione del modello della distribuzione dei dati.
Potenza Statistica: meno dati equivale a meno potenza statistica. Questo è particolarmente vero nei test statistici in cui i valori nulli potrebbero influenzare i risultati.
Ecco perché nel caso studio in esame si è scelto di procedere con l'eliminazione (figura 9).
L'eliminazione non è stata un problema dal momento che il numero di righe rimanenti risultava comunque adeguato alle successive analisi ed inoltre in questo modo si ha una rappresentazione più veritiera del cliente.
Inoltre, in seguito all'importazione sono stati selezionati i giusti tipi di dato per le colonne laddove PowerBI falliva nel riconoscerli automaticamente, questo per poi facilitare le fasi di data visualization e clustering.
Naturalmente ci sono delle colonne in cui i valori nulli sono stati ammessi, per esempio sarà possibile avere un cliente che non ha mai ricevuto una e-mail e perciò la sua istanza relativa all'attributo 'E-mail spedite' avrà valore nullo.
Inoltre, in alcune colonne i valori nulli sono stati sostituiti con il valore zero.

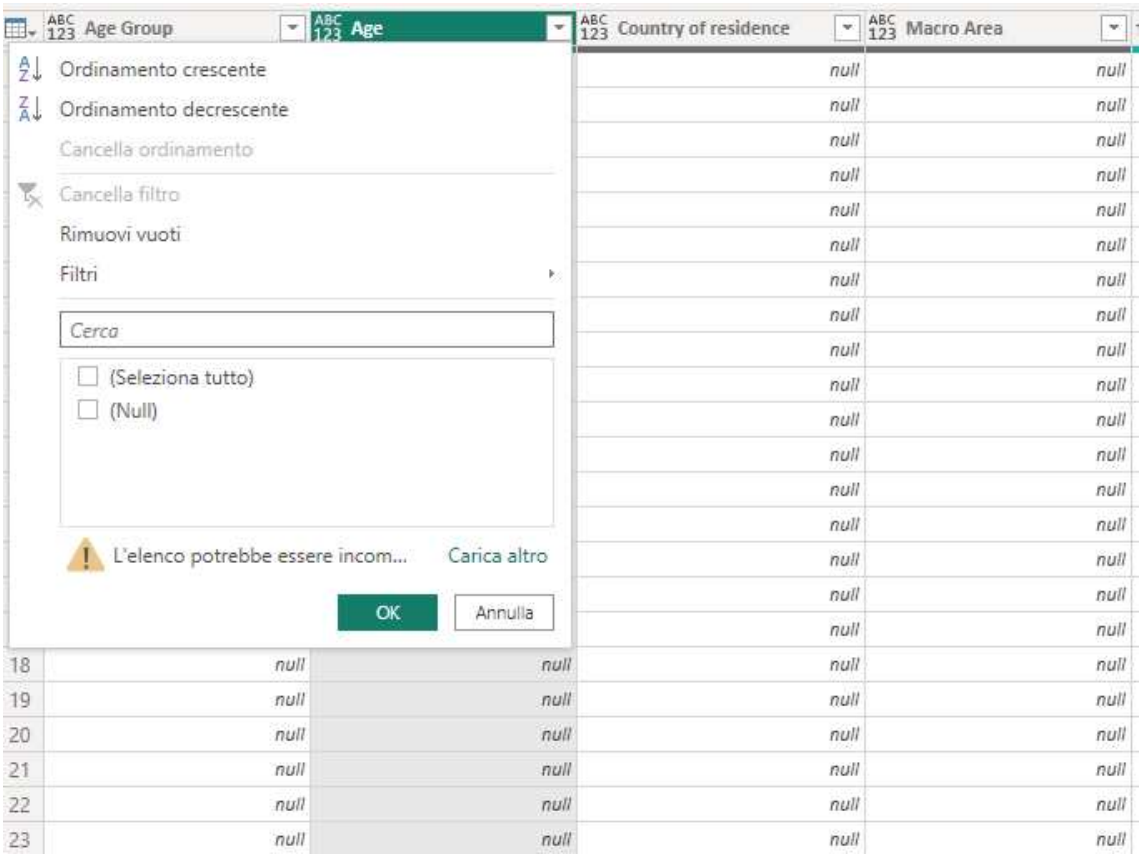


Figura 9, eliminazione valori nulli

Clustering

Una volta terminato il procedimento di pre-processing, bisogna assegnare ogni cliente al relativo cluster.

Per la cluster analysis si è fatto ricorso alla possibilità di utilizzare all'interno di PowerBI degli script Python e richiamare quindi funzionalità esterne avanzate fornite da PyCaret.

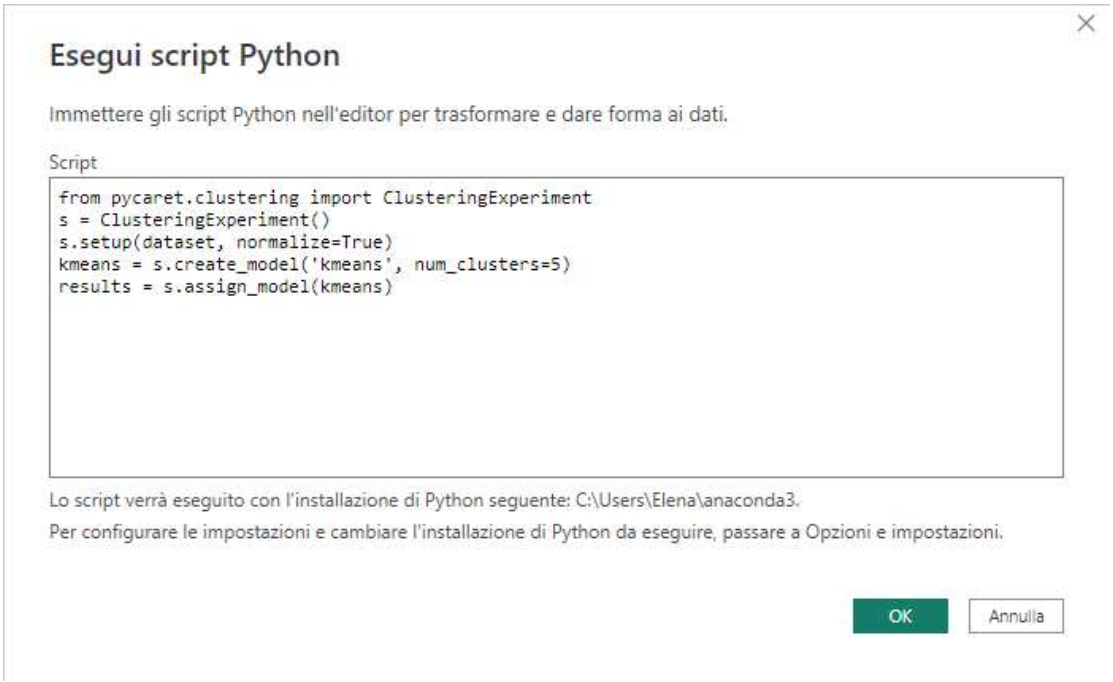


Figura 10,codice Python

A questo punto è stato eseguito il codice Python, che è commentato di seguito (e mostrato in figura 10):

```
from pycaret.clustering import ClusteringExperiment:
```

Una volta scaricata la libreria Pycaret viene importata la classe 'ClusteringExperiment' dal modulo 'Pycaret. clustering', questo modulo e questa classe sono responsabili della creazione e della validazione del modello di clustering, permettendo di mantenere un alto livello di astrazione.
Pycaret automatizza il pre-processing per l'algoritmo di Machine Learning utilizzato, cioè trasforma le features categoriche (ovvero features i cui valori sono discreti e appartenenti a categorie (es. Macro Area) in numeriche (questo perché gli algoritmi di ML prendono in input variabili numeriche).
Ciò viene effettuato con Il metodo 'One Hot Encoding'. Il One-Hot Encoding trasforma una variabile categorica con N categorie in N variabili binarie, ciascuna corrispondente a una categoria specifica.

Per esempio, la variabile categorica "sesso" con due valori assumibili: uomo e donna verrà trasformata in due variabili corrispondenti al genere, a cui verrà assegnato un valore binario vero o falso.
Un problema riscontrato in fase di progetto è stato scegliere tra le features il 'Customer Id' una matricola che identifica ogni cliente, ma essendo un valore univoco vi erano tanti valori quanti clienti e una volta che questo attributo veniva trasformato in formato numerico a causa della tecnica di 'One Hot Encoding', il dataset ottenuto saturava la memoria rendendo impossibile il clustering.
Dopo un'attenta revisione si è convenuto di non includere tale colonna nel dataset, in quanto effettivamente priva di contenuto informativo.
Pycaret è basato su Scikit-learn [20], una libreria molto utilizzata per il machine learning che fornisce una gamma di algoritmi di apprendimento, come K-means, DBSCAN, Agglomerative Clustering e molti altri.
Questi algoritmi sono pronti all'uso e possono essere facilmente applicati a set di dati per eseguire l'analisi di clustering.

```
s = ClusteringExperiment ()
```

Crea un'istanza per il modello di clustering denominata 's'.

```
s.setup(dataset, normalize=True)
```

il metodo 'setup' della classe 'ClusteringExperiment' prende come argomenti il dataset e l'opzione 'normalize'(=True). Quello che fa è configurare l'esperimento utilizzando il dataset fornito, normalizzando i dati, ovvero ridimensionandoli in modo che abbiano media uguale a zero e deviazione standard uguale a uno.
La normalizzazione è fondamentale per poter confrontare tutte le features sulla stessa scala e far sì che esse contribuiscano in maniera omogenea al clustering finale.

```
kmeans = s.create_model('kmeans', num_clusters=5)
```

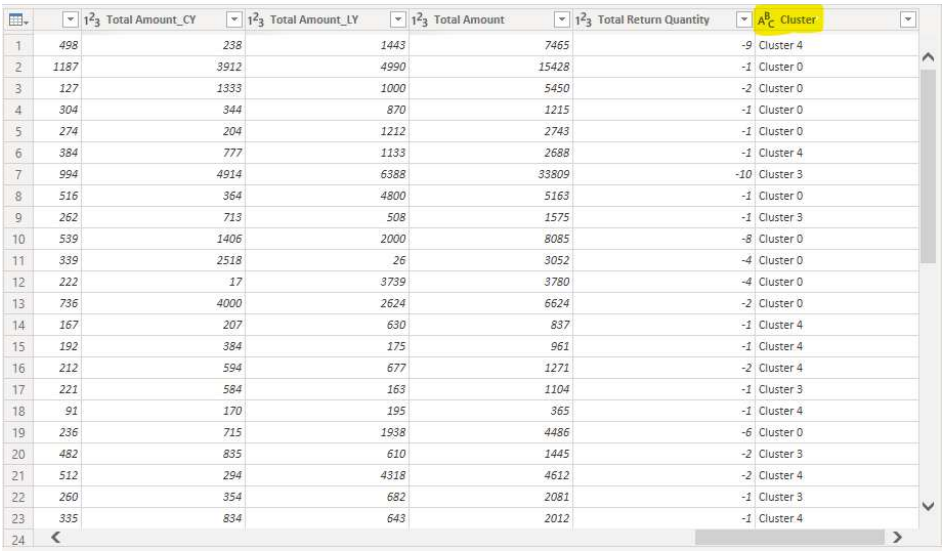
Con questa istruzione, vengono definiti due parametri del modello di clustering, ovvero l'utilizzo dell'algoritmo K-means ed il numero di cluster finali. Poniamo il numero di cluster uguale a cinque.

```
results = s.assign_model(kmeans)
```

L'ultima istruzione assegna i punti dati del dataset ai cluster utilizzando il modello K-means e memorizza i risultati nella variabile 'results'.
Il processo di clustering si è concluso in circa quattro ore e trascorso questo lasso di tempo, viene restituita la tabella (figura 11) contenente proprio la variabile 'results'.
Questa variabile non sarà nient'altro che una tabella rappresentante il dataset che avevamo ottenuto dopo il pre-processing con l'aggiunta della colonna 'cluster' che assegna ad ogni cliente un valore da zero a quattro (cinque gruppi), posta in evidenza nella figura 12.

	Name	Value
1	dataset	Table
2	results	Table

Figura 11, results



	Total Amount_CY	Total Amount_LY	Total Amount	Total Return Quantity	Cluster
1	498	238	1443	7465	-9 Cluster 4
2	1187	3912	4990	15428	-1 Cluster 0
3	127	1333	1000	5450	-2 Cluster 0
4	304	344	870	1215	-1 Cluster 0
5	274	204	1212	2743	-1 Cluster 0
6	384	777	1133	2688	-1 Cluster 4
7	994	4914	6388	33809	-10 Cluster 3
8	516	364	4800	5163	-1 Cluster 0
9	262	713	508	1575	-1 Cluster 3
10	539	1406	2000	8085	-8 Cluster 0
11	339	2518	26	3052	-4 Cluster 0
12	222	17	3739	3780	-4 Cluster 0
13	736	4000	2624	6624	-2 Cluster 0
14	167	207	630	837	-1 Cluster 4
15	192	384	175	961	-1 Cluster 4
16	212	594	677	1271	-2 Cluster 4
17	221	584	163	1104	-1 Cluster 3
18	91	170	195	365	-1 Cluster 4
19	236	715	1938	4486	-6 Cluster 0
20	482	835	610	1445	-2 Cluster 3
21	512	294	4318	4612	-2 Cluster 4
22	260	354	682	2081	-1 Cluster 3
23	335	834	643	2012	-1 Cluster 4
24					

Figura 12, colonna cluster

Data visualization

L'ultima fase della flowchart è la visualizzazione dei risultati.
Una volta a disposizione il dataset mostrato in figura 12, è bastato importarlo dal Editor Power Query all'interfaccia di Power Bi Desktop e utilizzare gli strumenti grafici messi a disposizione per mettere in risalto le informazioni ritenute più opportune.

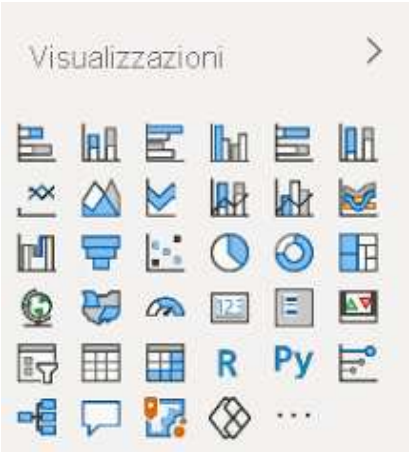


Figura 13, strumenti grafici Power Bi

In figura 13 sono presenti i principali oggetti visivi, ma in Power Bi è possibile importarne altri o anche crearne degli appositi.
I report ottenuti successivamente a queste operazioni sono mostrati nel capitolo 4.

4.Risultati

Il presente capitolo costituisce un punto focale della tesi, in cui verranno presentati in dettaglio i risultati ottenuti dall'analisi di clustering condotta utilizzando il software Power BI.

Questo capitolo sarà suddiviso in due sezioni principali: "Report generali" e "Report clustering", al fine di fornire un'esaustiva panoramica delle informazioni emerse dal processo di analisi.

La sezione dei "Report generali" metterà in luce i principali risultati emersi dall'intero studio, fornendo una visione generale dell'insieme dei dati esaminati. Verranno presentati grafici, tabelle e diagrammi rappresentativi, che offriranno una panoramica intuitiva delle caratteristiche dei dati e delle relazioni rilevanti individuate. Inoltre, saranno forniti commenti e osservazioni sulle tendenze identificate.

La seconda sezione, "Report clustering", sarà incentrata sull'approfondimento dei risultati ottenuti tramite l'applicazione delle tecniche di clustering. Saranno presentati grafici specifici, rappresentazioni visive e statistiche che illustreranno i cluster identificati e le loro caratteristiche distintive. Ogni cluster verrà analizzato individualmente, con particolare attenzione alle sue peculiarità, ai pattern di comportamento che lo contraddistinguono e alle possibili implicazioni che emergono dai dati.

Al fine di supportare le argomentazioni, verranno forniti commenti dettagliati per ogni risultato presentato. Inoltre, saranno evidenziati i punti di forza e le limitazioni del processo di clustering, fornendo una valutazione critica delle scelte adottate.

Importante sottolineare che i grafici sono in due dimensioni, per motivi di complessità della rappresentazione, ma i risultati prodotti e l’appartenenza dei pattern ai cluster non dipende dalle due dimensioni visibili, ma dal totale delle venti features ritenute importanti per l’analisi.

4.1 Report Generali

Prima di arrivare a definire e caratterizzare i cluster di clienti, ho effettuato un’analisi generale su di loro.

Quest' analisi utilizza semplicemente alcuni dei più noti strumenti di data visualization di Power Bi per mettere in risalto le caratteristiche generali degli elementi appartenenti al nostro dataset.
Esempi delle caratteristiche citate saranno informazioni riguardo l'età, la provenienza e le abitudini di acquisto dei clienti.
Il primo report estratto è mostrato in figura 14, si tratta di un grafico a torta relativo ai consensi concessi dai clienti.
I clienti possono dare o meno l'autorizzazione ad essere raggiunti da offerte e promozioni dell'azienda via sms e via e-mail.
Inoltre, l'utente autorizza l'azienda ad utilizzare i propri dati per analisi e studi di marketing.
Il grafico a torta è ottimale per le rappresentazioni in forma percentuale, nell'esempio a primo impatto risaltano in verde coloro che hanno fornito il loro consenso ed in blu invece saranno rappresentati coloro che non l'hanno fatto.
Possiamo notare che gli individui preferiscono essere raggiunti via e-mail piuttosto che per messaggio (81.33% vs 75.65%) e che gli stessi sono più disposti a svelare i propri dati per finalità di marketing rispetto a quelle di Analysis (81.3% vs 77.8%), anche se bisogna dire che in entrambi i casi i valori sono molto vicini tra loro.

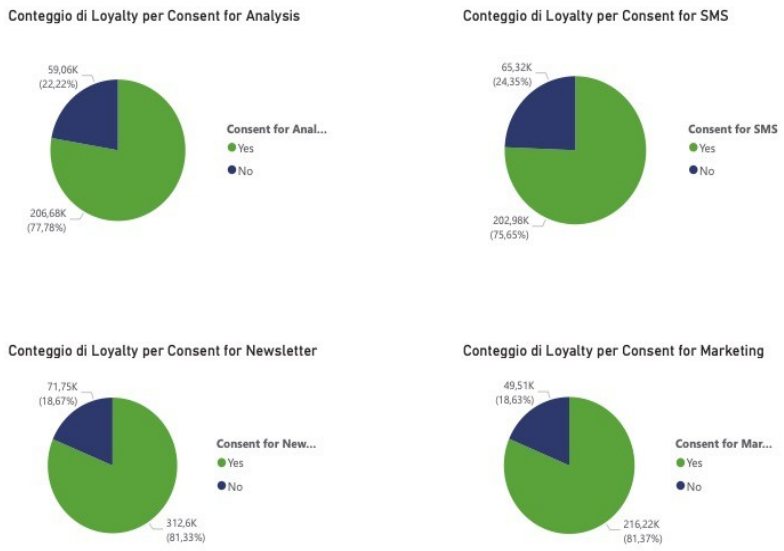


Figura 14, consensi clienti

Uno dei dati che più interessa alle aziende naturalmente è la spesa effettuata dai clienti, anche nei prossimi report e in quelli relativi al clustering sarà uno dei dati su cui verrà posta la lente d'ingrandimento.

Come spiegato nel capitolo 3.2 la spesa totale nel nostro dataset si trova sotto la voce 'Total Amount', nel report (figura 15) è mostrato come essa sia collegata all'area in cui viene effettuata ed al valore della Loyalty del cliente.

Per esplicare le considerazioni accennate sono stati utilizzati un grafico a barre raggruppate ed un grafico ad aree, che sono funzionali ad evidenziare le differenze tra i valori (in un caso la loyalty posta sulle ordinate, nell'altro la Best Region posta sulle ascisse).

Si evince che i clienti che spendono di più sono quelli appartenenti alla categoria 'New' (ovvero quelli che acquistano per la prima volta) ed i clienti che acquistano soprattutto negli Stati Uniti, quest'ultima informazione era prevedibile, dato che la maggior parte dei punti vendita dell'azienda si trova proprio negli Stati Uniti.

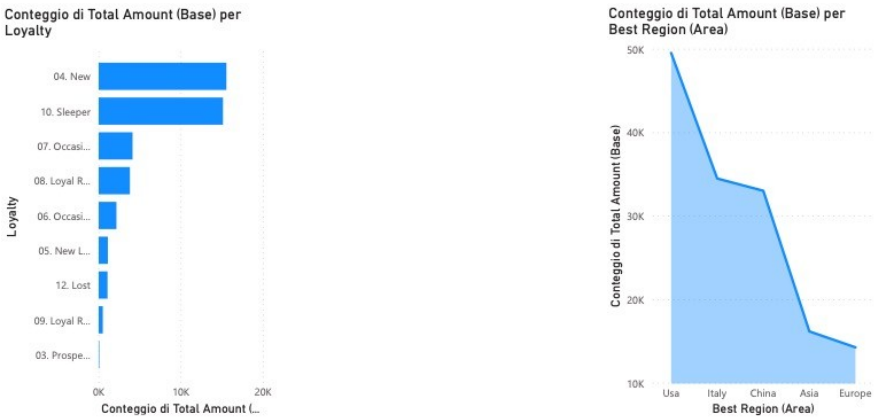


Figura 15, Total Amount / Loyalty /Best Region

Continuando ad analizzare la variabile del Total Amount in figura 16 sono stati utilizzati un oggetto visivo denominato ‘misuratore’ ed un grafico a barre in pila.
Il misuratore ci dà un’indicazione grafica molto intuitiva del valore del Total Amount dell’anno scorso rispetto al Total Amount complessivo (relativo ai sette anni di raccolta dati).
Si può notare che l’azienda è in netta crescita, in quanto ben otto milioni sui sedici complessivi derivino dall’attività aziendale dell’anno appena trascorso.
Il grafico a barre in pila invece è utile per ripartire il valore del Total Amount dell’anno scorso tra i vari gruppi d’età dei clienti, come si evincerà anche in alcuni dei grafici successivi la maggior parte degli incassi avvengono grazie al contributo della clientela appartenente al gruppo 51-60 anni ed al gruppo 41-50.

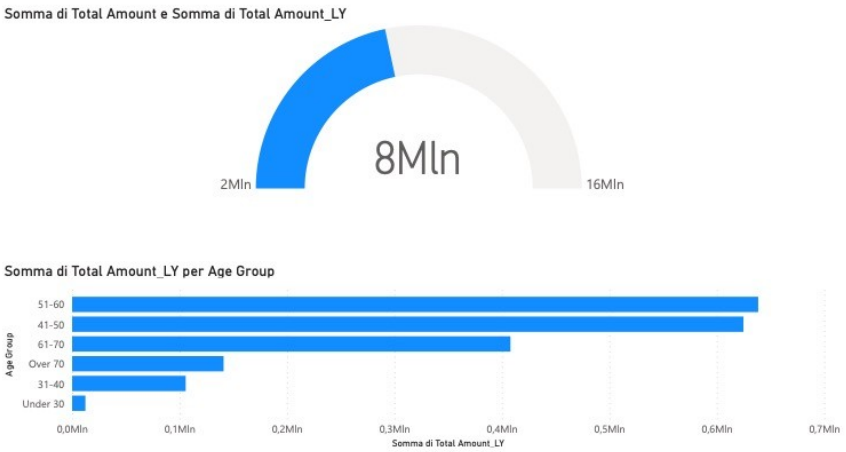


Figura 16, Confonto tra Total Amount last year e Age Group

Uno dei punti di forza di Power Bi, come già accennato, è l'interattività e nella figura 17 si può osservare questa funzionalità.

Sono rappresentate parallelamente una tabella con i cinque livelli dell'indice RFM Frequency ed un istogramma a colonne raggruppate.

Cliccando a piacere in uno dei cinque livelli, viene evidenziata sull'istogramma la partizione di clienti con quel valore di RFM Frequency appartenente ad ogni fascia d'età.

Nell'immagine è stata catturata un'istantanea dei pattern di livello 1 e si nota che essi hanno tendenzialmente tra i 41 ed i 50 anni, mentre nessuno sotto ai 30.

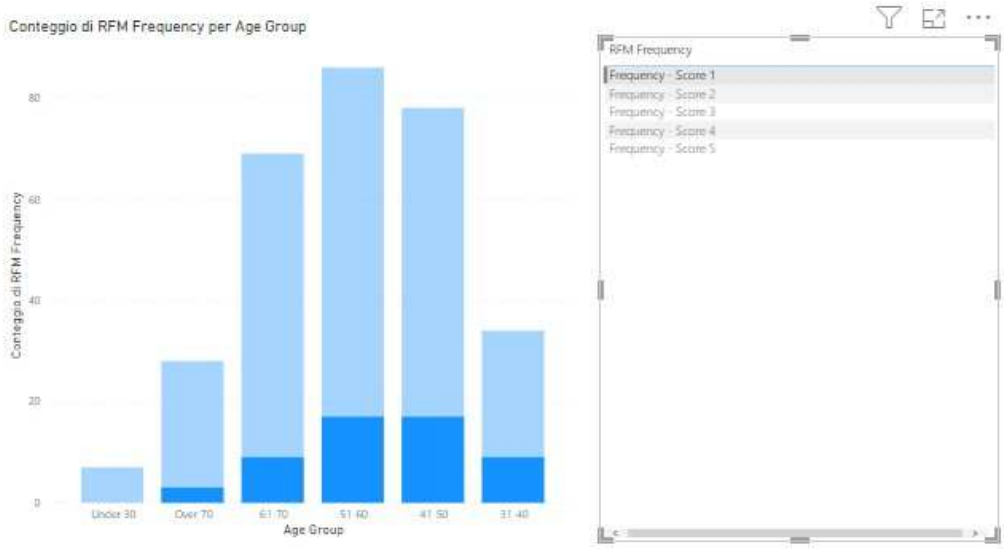


Figura 17, Interattività tra RFM Frequency e Age Group

4.2 Report Clustering

Dopo aver visto i report generali si scende nel dettaglio con i report derivanti dal processo di clustering.

L'oggetto visivo più utilizzato è lo scatter plot o grafico a dispersione, dove ogni pattern è rappresentato con un punto e i vari cluster sono differenziati dai colori.

Una premessa da fare è che nonostante il pre-processing, il dataset potrebbe contenere degli outlier rappresentati ad esempio da un erroneo inserimento dei dati in fase di popolamento del database.

Per ovviare a questo problema si possono utilizzare i filtri che Power Bi mette a disposizione.

Nello specifico alla rappresentazione in figura 18 che mostra il rapporto Age/Total Amount è stata impostata la limitazione dell'età tra i valori di 15 e 90 anni (si può notare un pattern sul valore di 2 anni e un altro sul valore di 110, che naturalmente non possono essere clienti reali).

Poi è stato posto un filtro dinamico ai fini di rappresentare solo i valori di Total Amount desiderati dall'utente grazie al dispositivo di scalatura sulle ordinate (impostato in figura 18 a 20k), questo ha uno scopo prettamente grafico, ovvero

quello di poter visualizzare i pattern in maniera meno ‘schiacciata’ e porre l’attenzione su qualsiasi elemento raffigurato.

Il report successivo all’applicazione dei filtri è visibile nella figura 19.

Quello che si evince dai grafici è che i clienti più redditizi per l’azienda sono quelli del cluster 0 ed hanno principalmente tra i 55 ed i 65 anni, nessuno è un under 30, mentre il cluster meno redditizio è il numero 4.

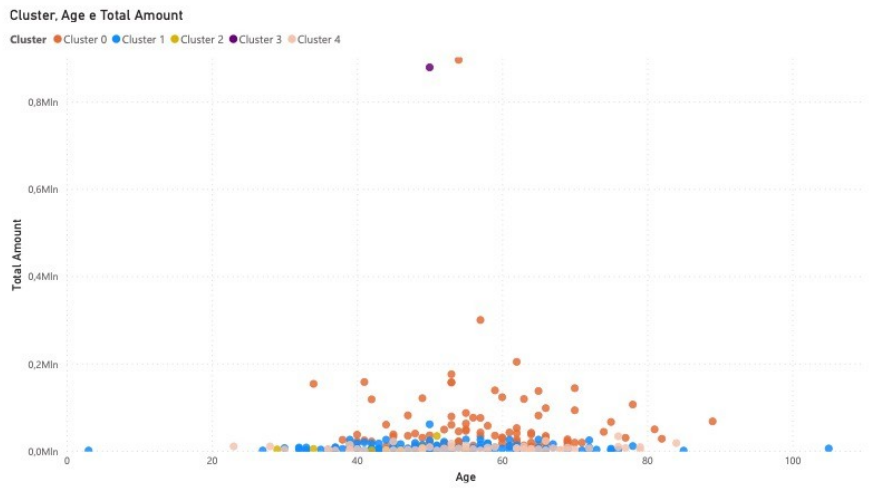


Figura 18, Clustering Age Total Amount pre filtraggio

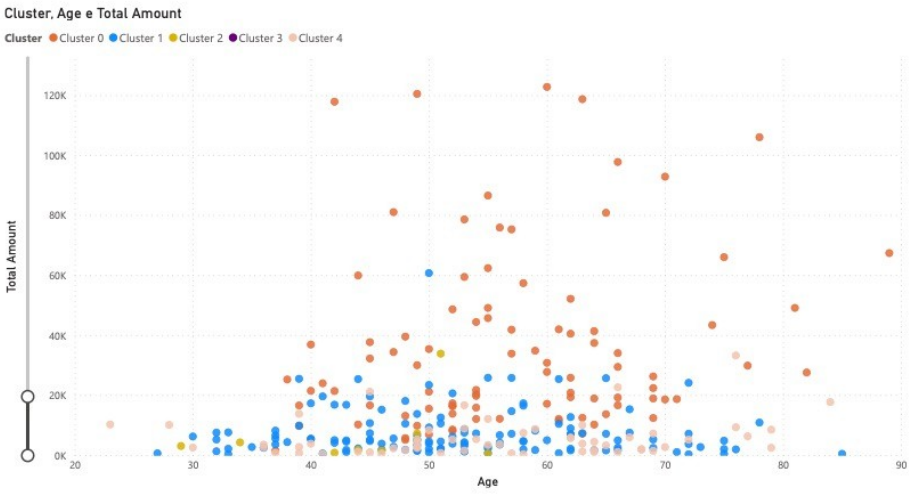


Figura 19, Clustering Age Total Amount post filtraggio

L’analisi mostrata negli ultimi due grafici prosegue in figura 20, dove nelle ordinate è presente l’Average Ticket Base al posto del Total Amount.
Le informazioni generali riprendono quelle già argomentate, anche in questo caso gli scontrini medi più alti sono emessi ai clienti del cluster 0 ed i picchi corrispondono a individui intorno ai 55 anni.
In questa trattazione però si può scendere nel dettaglio della differenza tra cluster, grazie all’interattività delle rappresentazioni, cliccando o su un pattern specifico o su un insieme.
In figura 20 si è posto il focus su un elemento del cluster 0 (l’elemento a cui è associata l’etichetta), mentre in figura 21 invece è stato messa in risalto la somma di Average Ticket per ogni gruppo utilizzando un oggetto visivo denominato ‘scheda con più righe’.
L’informazione interessante che deriva da questa operazione è che la somma degli scontrini medi dei cluster 2 e 3 è minore dello scontrino medio del singolo pattern in figura 20.
Perciò ad esempio in fase di progettazione di una campagna pubblicitaria lato marketing, per l’azienda idealmente sarebbe più proficuo rivolgersi a quel singolo utente che all’intero gruppo due.

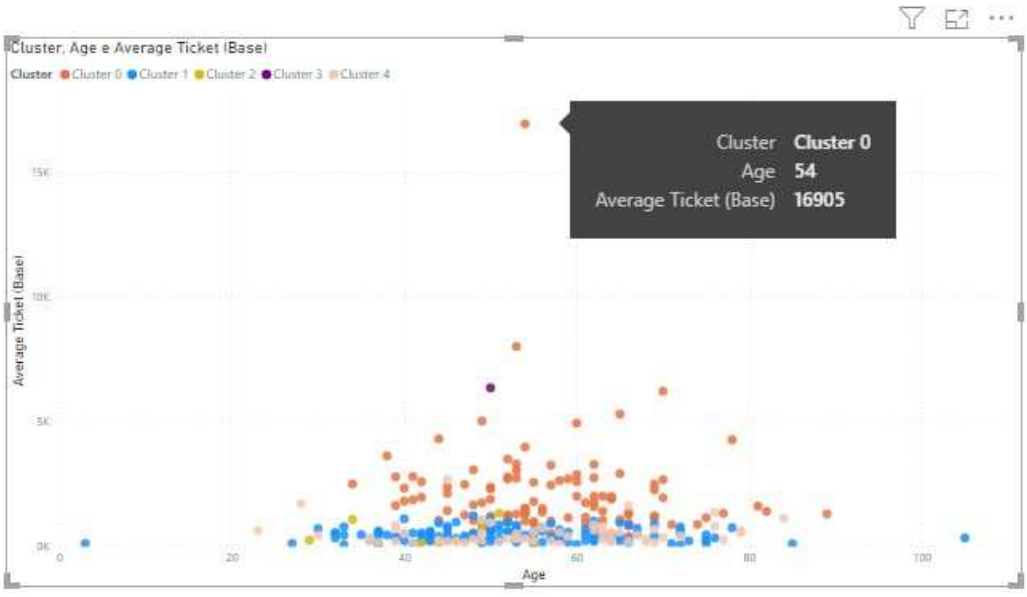


Figura 20, Clustering Average Ticket Age

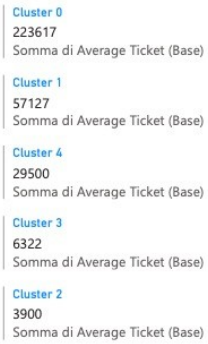


Figura 21, Somma di Average Ticket di ogni cluster

Per terminare l'analisi riguardante l'età, in figura 22 sono mostrate le features Age e Total Transactions.

Questa volta l'età è stata posta nelle ordinate, questo semplicemente per individuare in maniera più intuitiva i pattern e avere una visione più chiara delle tendenze.

In questo caso non c'è una netta prevalenza di un cluster rispetto agli altri come invece accadeva nei report precedenti.

Il pattern con il maggior numero di transazioni (ben 139) sarà il tre.

Escludendo il gruppo tre la media più alta di transazioni apparterrà al cluster zero, con un valore di 27,49, mentre il cluster numero due avrà con una media di 10,88 transazioni il valore più basso in questo campo.

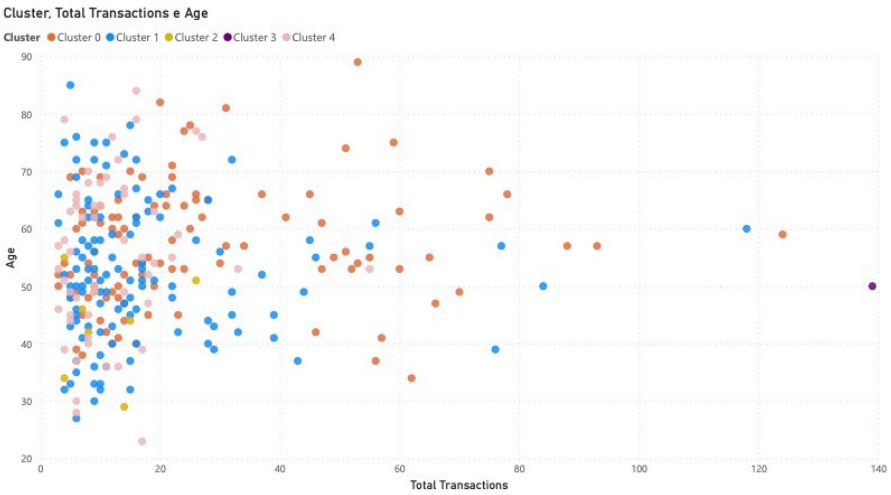


Figura 22, clustering age Total Transactions

Nella figura 23 è stata sfruttata una potenzialità degli oggetti visivi di Power Bi, ovvero quello di associare la grandezza dei pattern all'interno dello scatter plot ad una feature.

In questo caso per questioni di chiarezza grafica non sono stati identificati i 5 cluster con i rispettivi colori per poter porre in risalto la correlazione tra attributi.

Nello specifico nell'asse delle X abbiamo l'età dei clienti e in quello delle Y l'RFM Score.

Si può notare che i punti hanno una distribuzione gaussiana, ovvero i valori più alti di RFM Score sono per le età 'centrali' (dai 50 ai 60 anni).

In aggiunta a ciò, a valori alti di RFM Score corrispondono valori elevati di Total Amount e questo si può notare come già spiegato dalle dimensioni delle circonferenze dei pattern.

Questo risultato è importante perché rivela la bontà dell'indice RFM, ovvero clienti a cui è stato assegnato un valore elevato sono poi effettivamente quelli che fanno crescere l'azienda in termini di ricavi.

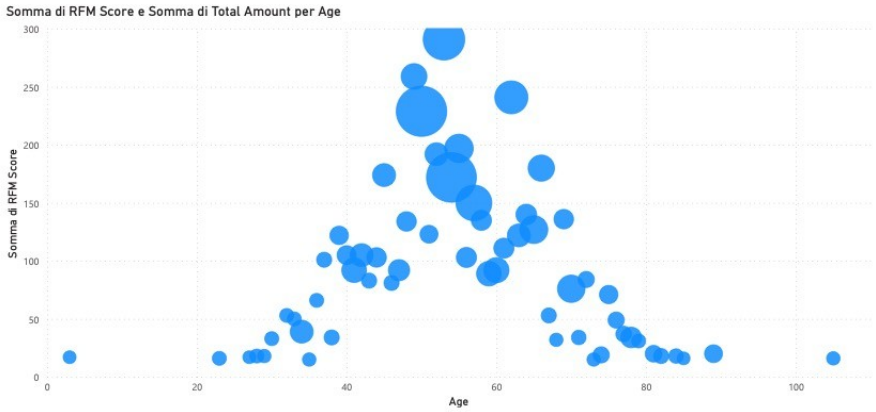


Figura 23, Clustering RFM Score e Age, con Total Amount come dimensione dei pattern

Un aspetto interessante da studiare delle attività appartenenti al mondo della moda è quello dei resi.

In figura 24 si è utilizzato lo scatter plot per ricavare alcune informazioni relative a questo fenomeno.

Nel dettaglio, nelle ascisse è presente la quantità di merce restituita (che cresce verso sinistra essendo i valori che indicano un capo riportato dal cliente registrati in segno negativo), nelle ordinate troviamo invece l'RFM Score.

La dimensione dei pattern rappresenta il Total Amount ed il colore degli stessi la categoria di Loyalty di appartenenza leggibile nella legenda.

In sunto i clienti che effettuano più resi hanno un RFM Score basso e fanno parte della categoria al livello otto dei 'Loyal Retained', perciò sono clienti non molto legati all'azienda ed effettuano resi di quantità generalmente non costose.

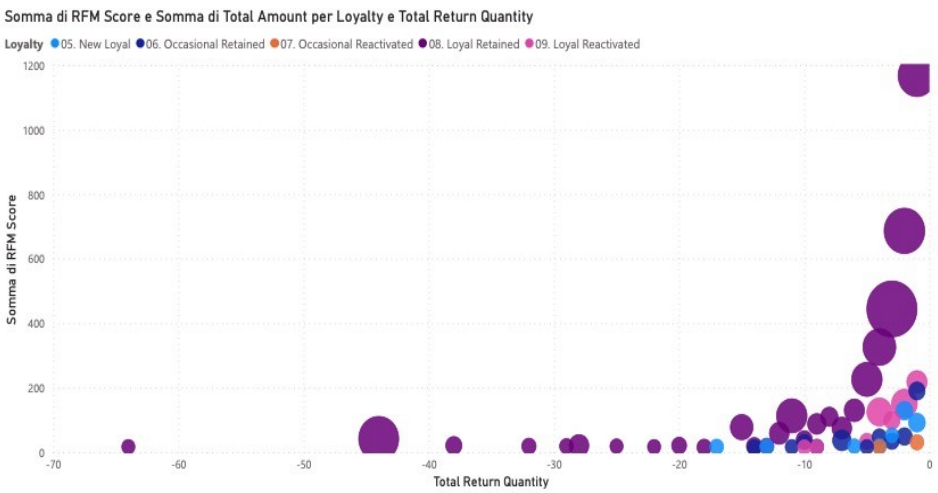


Figura 24, Grafico relativo ai resi

In figura 14 si erano analizzati i consensi dei clienti e la loro volontà ad essere raggiunti per promozioni e offerte da parte dell'azienda. In figura 25 si va ad approfondire quanto queste comunicazioni (sotto forma di e-mail) interessano realmente al cliente.

Questa analisi è basata sul rapporto tra e-mail cliccate ed e-mail ricevute dal consumatore.

Dal grafico a dispersione si nota che i cluster che cliccano più e-mail sono il numero zero e quattro, ma questo è dovuto alla numerosità dei due insiemi.

Approfondendo lo studio con l'aggiunta dell'istogramma a colonne in pila in forma percentuale si nota che il cluster più interessato alle e-mail o comunque più propenso ad aprirle è il cluster due con il 15,7% di e-mail cliccate sul totale di quelle ricevute.

Seguono il gruppo uno con il 14,81%, il cluster quattro con il 14,44%, lo zero con il 10,09% ed infine il tre con solo il 2,54%.

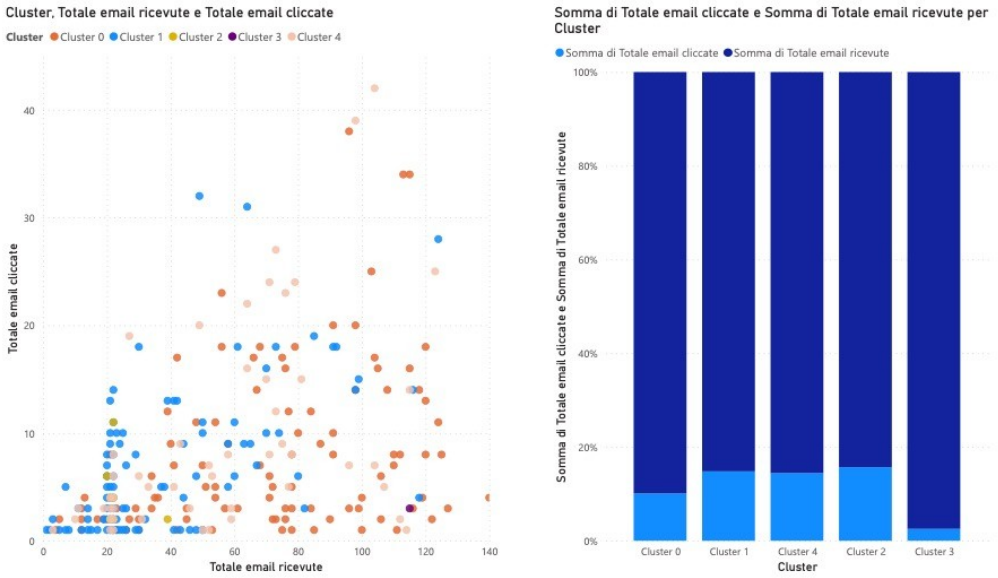


Figura 25, Rapporto tra le e-mail ricevute e quelle cliccate

L'esempio nel report appena discusso svela che nonostante risulti il più utilizzato lo scatter plot non è sempre lo strumento più adatto per analisi di clustering, a tale proposito in figura 26 è mostrato come per alcuni tipi di cluster analysis si possano utilizzare oggetti visivi alternativi, come ad esempio le mappe.

La mappa ci mostra i punti vendita dell'azienda sparsi in tutto il mondo, e per ognuno è presente un pattern tanto più grande quanti più individui hanno dichiarato quel negozio come punto acquisto di fiducia.

Inoltre, ogni pattern sarà un grafico a torta che indica il frazionamento di cluster al suo interno, ovvero quali clienti acquistano maggiormente nel dato negozio.

A partire dalla mappa, si può affermare che la boutique dichiarata come preferita dai clienti è quella localizzata a Londra in 'New Bond Street'.

I clienti del cluster 1 sono dispersi in tutto il mondo, mentre quelli del cluster 4 acquistano solo in Inghilterra ed Irlanda.

I clienti meno numerosi sono quelli del cluster 3 e preferiscono acquistare in Italia.

Il paese con più punti vendita sono gli Stati Uniti, dove i negozi preferiti sono quelli situati al Madison di New York e nel Kansas.

Nelle tre boutique della Cina presenti a Shangai i consumatori appartengono quasi esclusivamente al cluster 0.

I punti vendita meno ‘amati’ dai clienti sono quelli in Estonia, Cipro e Bulgaria.

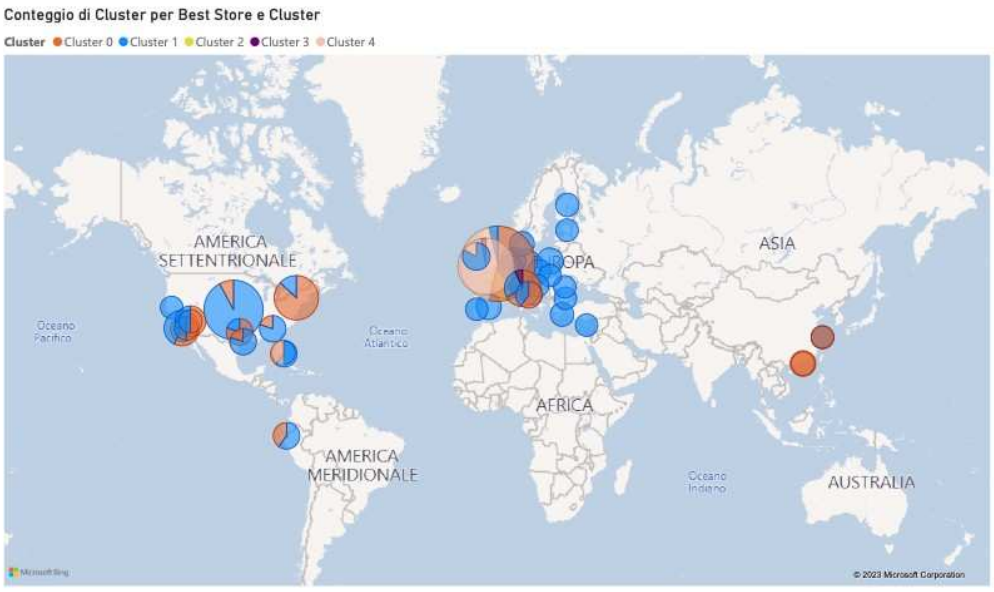


Figura 26, Mappa della collocazione dei cluster nel mondo

5.Conclusioni e sviluppi futuri

Questa tesi ha introdotto inizialmente il campo interdisciplinare della business inteligence, chiarendone i punti di forza ed i suoi ambiti applicativi, grazie alla business inteligence le organizzazioni possono estrarre valore dai dati per ottenere una visione chiara e approfondita delle proprie attività.
Successivamente si sono argomentati i concetti teorici di clustering e data visualization, l'approccio alla cluster analysis ha consentito di identificare pattern, relazioni o segmenti nascosti nei dati che possono essere utilizzati per prendere decisioni più informate, la data visualization è fondamentale invece per esplorare e comunicare i risultati in modo efficace ed intuitivo.
Poi si è passati alla descrizione dei materiali e metodi adoperati come Power Bi, Python e Pycaret, identificando le motivazioni per le quali sono stati funzionali al progetto.
Infine, si è giunti alla presentazione e discussione dei risultati, rappresentati da report seguiti da commenti.
Una volta conclusa l'analisi si possono fornire all'azienda i risultati (i report stessi) ed un riassunto dei vari cluster identificati che potranno essere utilizzati per le finalità discusse.

Il cluster zero rappresenta il gruppo con l'RFM Score più elevato. Gli appartenenti al gruppo fanno parte principalmente della categoria 'New Loyal'. Questi clienti sono particolarmente proficui per l'azienda e ciò è evidente guardando i valori di Total Amount e Average Ticket.

Il cluster uno contiene il maggior numero di rappresentanti, nonostante ciò, il loro Total Amount complessivo è minore di quello del cluster 0. Una loro caratteristica è quella di effettuare tanti acquisti, ma di prodotti piuttosto economici.

Il cluster due è uno dei cluster più legati e interessati all'azienda, è stato mostrato nel report delle e-mail cliccate. Questo legame è visibile anche analizzando i valori dell'RFM Frequency, dove la maggior parte dei pattern

appartengono al livello uno, indice di come gli acquisti avvengano con alta frequenza.

Il cluster tre è il meno numeroso, pochissimi clienti ne fanno parte, allo stesso tempo però è il gruppo che in media spende di più. Geograficamente gli appartenenti a questo gruppo si trovano in Italia e acquistano principalmente nel punto vendita di Milano.

Infine, il cluster quattro è composto da clienti principalmente del Regno Unito. I valori di RFM, Total Amount e Total Transactions sono tra i più bassi. Si tratta del cluster meno proficuo per l'azienda.

Per quanto riguarda gli sviluppi futuri si potrebbe partire dal progetto impostato in questa tesi e approfondirlo sotto diversi aspetti.
Per esempio, sarebbe interessante svolgere un'analisi simile a quella condotta ma incrociando un altro dataset a disposizione dell'azienda, come quello sulle transazioni specifiche o comunque un'origine dati contenente informazioni riguardanti i clienti.
Un'altra via percorribile è quella di partire dallo stesso dataset, ma utilizzare tecniche di machine learning per il clustering più specifiche.
Il numero ottimale di cluster è stato scelto in modo empirico seguendo le indicazioni del brand, ma andrebbe selezionato attraverso un apposito criterio, ad esempio quello del gomito, che prevede di iterare il K-means per diversi valori di K calcolando ogni volta la somma delle distanze al quadrato tra ogni centroide ed i punti del proprio cluster.
Graficando i valori di K (asse orizzontale) e i valori della somma delle distanze al quadrato (asse verticale), si ottiene un grafico. Questo grafico deve essere letto da destra verso sinistra e si deve trovare il punto in cui la curva tende a salire in modo più consistente, selezionando il valore di K corrispondente come numero dei cluster.
Altrettanto interessante sarebbe mettere a confronto algoritmi diversi di clustering al fine di trovare quello che meglio performa relativamente ai dati in oggetto.

Infine, a livello generale si potrebbero utilizzare i concetti e le funzionalità descritte nella tesi per studiare altri ambiti dell'azienda come la produzione, le risorse umane e molto altro ancora.

Bibliografia

[1] L. Zaramella, *Cluster Analysis per la segmentazione della clientela utilizzando il Software SAS® ENTERPRISE MINERTM.*

[2] R. Terenzi, *Progettazione e realizzazione di una campagna di data analytics a supporto delle attività di vendita di un calzaturificio.*

[3] F. Tangianu and R. Nardi, "[Il clustering diseases in medicina interna come strumento di approccio alla complessità dei pazienti".

[4] Gianmpaoli, Lenzi, Parente, Frajese and Tessitore, "Abitudini, stili di vita e socialità: una cluster analysis di studenti universitari italiani," *Igiene e Sanità pubblica,* 2021.

[5] A. Gagliardi, The Clustering Project.

[6] Bellini, Palesi and Nesi, "Multi Clustering Recommendation System for Fashion Retail.".

[7] G. Di Marzo, *Advanced Analytics per il Marketing: clustering dei clienti fidelizzati,* 2018.

[8] Bailey and Kenneth, "Typologies and taxonomies: An introduction to classification techniques".

[9] Lloyd and Stuart, "Least squares quantization in PCM," *Information Theory,* 1982.

[10] Kaur, Kamal, Kaur and Singh, "K-Medoid clustering algorithm-a review," *Int. J. Comput. Appl. Technol.*

[11] J. A. Hartigan, "Clustering algorithms," *New York: Wiley,* 1975.

[12] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *EEE Transactions on Pattern Analysis and Machine Intelligence,* 2002.

[13] Wei Wang, Jiong Yang and Richard R. Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining," 1997.

[14] Dempster, Laird and Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society,* 1977.

[15] Kumar Sm, "Meena Belwal," 2017.

[16] Microsoft, "Microsoft Power BI: Visualizzazione dei dati," [Online]. Available: https://powerbi.microsoft.com/it-it/.

[17] "Report in Power BI," [Online]. Available: https://learn.microsoft.com/it-it/power-bi/consumer/end-user-reports.

[18] "Python.org," [Online]. Available: https://www.python.org/.

[19] "PyCaret: Home," [Online]. Available: https://pycaret.org/.

[20] "scikit-learn Machine Learning in Python," [Online]. Available: https://scikit-learn.org/stable/.

Ringraziamenti

Alla fine di questo elaborato, mi sembra doveroso dedicare uno spazio per ringraziare brevemente tutte le persone che mi hanno aiutato in questo percorso.
Vorrei ringraziare tutta la mia famiglia, in particolar modo i miei genitori che mi hanno sempre supportato e spronato appoggiando ogni mia decisione, mia sorella, che tra le altre cose è stata fonte di curiosità per avvicinarmi agli argomenti affrontati in questo percorso, la mia seconda famiglia di Pollenza
Scalo che mi ha sfamato e accompagnato dall'asilo alla laurea e tutti i miei amici che hanno alleggerito i momenti più pesanti (grazie di esistere come direbbe qualcuno).