A McDonald's Data Analysis - Progetto B 2

Introduzione

L'analisi della segmentazione del mercato è uno strumento potente che aiuta le aziende a indirizzare i loro sforzi di marketing in modo più efficacemente ed efficientemente. Dividendo un mercato grazie all'analisi dati in gruppi distinti (o clusters) con esigenze, preferenze e caratteristiche simili, le aziende possono adattare i loro prodotti, servizi e strategie di marketing per soddisfare meglio le esigenze dei clienti.

Questo progetto scritto in Python in ambiente Jupyter ha l'obiettivo di costruire un'analisi approfondita sui gusti della clientela di uno dei famosi ristoranti della catena di McDonald's.

Abbiamo utilizzato le seguenti tecniche di analisi e modellazione dati:

1. **PCA** (*Principal Component Analysis*)
2. **Clustering K-Means**
3. **Clustering gerarchico**

🎯 *Obiettivi*

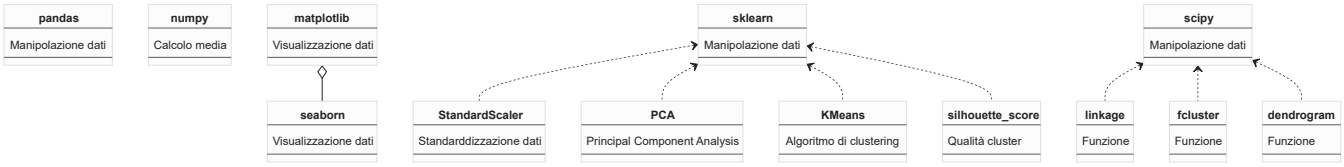
- **Segmentazione dei clienti:** L'obiettivo è suddividere i clienti di McDonald's in gruppi distinti in base alle loro preferenze e caratteristiche.
  - Questo viene fatto utilizzando tecniche di cluster analysis, come *K-Means* e *Clustering gerarchico*.
- **Rilevazione di correlazioni tra le features:** L'analisi cerca di scoprire se esistono relazioni significative tra le diverse caratteristiche dei clienti e le loro preferenze riguardo ai prodotti McDonald's.
  - Questo viene fatto principalmente attraverso *l'analisi delle componenti principali* (PCA) e *l'analisi della matrice di correlazione*.

✓ **Data science per costruire una strategia di marketing**

L'idea è che identificando gruppi di clienti con preferenze simili, McDonald's possa adattare le sue strategie di marketing e i suoi prodotti in modo più efficace per soddisfare le esigenze di ciascun segmento.

Workflow

Librerie utilizzate



Primo sguardo ai dati

1. **Lettura del dataset ("mcdonalds.csv"):** Il codice utilizza la libreria pandas (`pd.read_csv`) per leggere il file CSV "mcdonalds.csv" e memorizzarlo in un DataFrame chiamato `data` .
   - Il DataFrame è una struttura dati tabulare che consente di manipolare e analizzare i dati in modo efficiente.

	yummy	convenient	spicy	fattening	greasy	fast	cheap	tasty	expensive	healthy	disgusting	Like	Age	VisitFrequency	Gender
0	No	Yes	No	Yes	No	Yes	Yes	No	Yes	No	No	-3	61	Every three months	Female
1	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	+2	51	Every three months	Female
2	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	+1	62	Every three months	Female
3	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	+4	69	Once a week	Female
4	No	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes	No	+2	49	Once a month	Male
5	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No	No	No	+2	55	Every three months	Male
6	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No	+2	56	Every three months	Female
7	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	No	I love it!+5	23	Once a week	Female
8	No	No	No	Yes	Yes	No	No	No	Yes	No	Yes	I hate it!-5	58	Once a year	Male
9	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No	No	+1	32	Every three months	Female
10	No	Yes	No	Yes	No	Yes	Yes	No	No	No	Yes	-2	53	Every three months	Female

2. **Ispezione iniziale dei dati:**
   - `data.info()` : Fornisce un riepilogo delle colonne presenti nel DataFrame, inclusi i loro tipi di dati (ad esempio, oggetto, int64) e il numero di valori non nulli. Questa funzione è utile per identificare rapidamente eventuali valori mancanti e per comprendere la struttura generale dei dati.

```
>>>data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1453 entries, 0 to 1452
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   yummy           1453 non-null   object
 1   convenient      1453 non-null   object
 2   spicy           1453 non-null   object
 3   fattening       1453 non-null   object
 4   greasy          1453 non-null   object
 5   fast            1453 non-null   object
 6   cheap           1453 non-null   object
 7   tasty           1453 non-null   object
 8   expensive       1453 non-null   object
 9   healthy         1453 non-null   object
 10  disgusting      1453 non-null   object
 11  Like            1453 non-null   object
 12  Age             1453 non-null   int64 
 13  VisitFrequency  1453 non-null   object
 14  Gender          1453 non-null   object
dtypes: int64(1), object(14)
memory usage: 170.4+ KB
```

Descrizione del dataset

Il dataset analizzato contiene i risultati di un sondaggio condotto sui clienti di McDonald's. Include sia **dati demografici** (sesso ed età) sia **preferenze** relative a diversi aspetti dell'esperienza McDonald's.

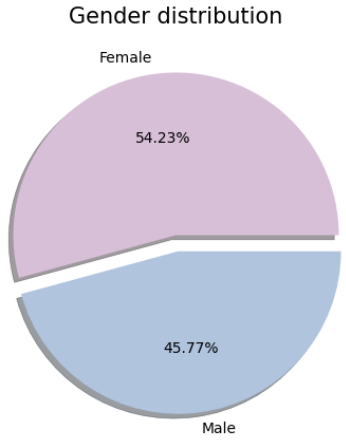
	yummy	convenient	spicy	fattening	greasy	fast	cheap	tasty	expensive	healthy	disgusting	Like	Age	VisitFrequency	Gender
0	No	Yes	No	Yes	No	Yes	Yes	No	Yes	No	No	-3	61	Every three months	Female

Panoramica delle features

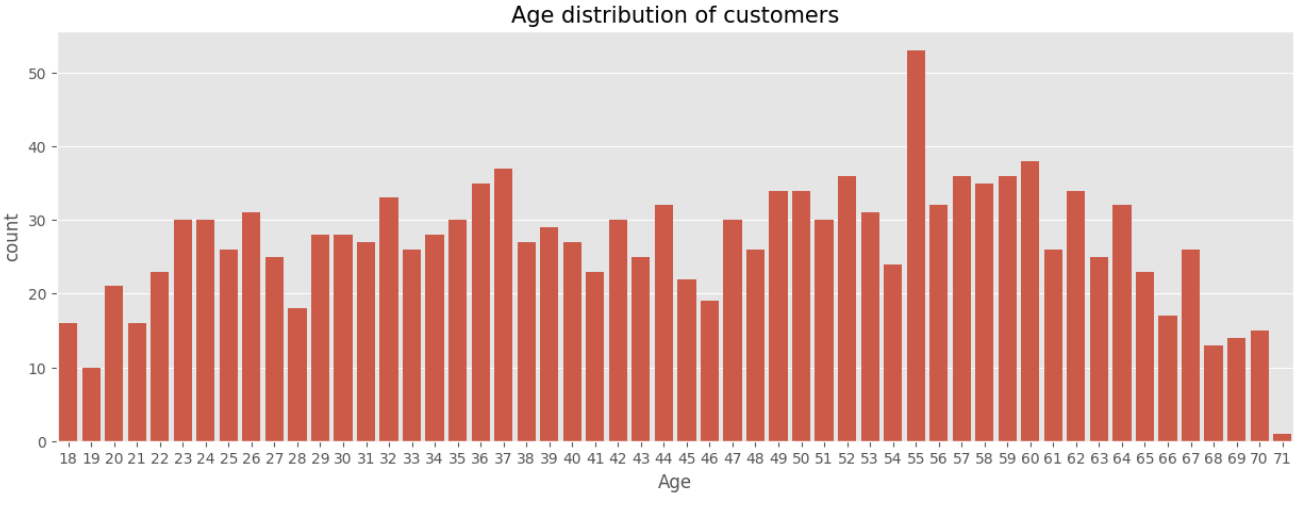
Le variabili possono essere classificate in due categorie principali:

1. **Dati Demografici:**
   - **Sesso:** Maschio o Femmina.
   - **Età:** L'età dei partecipanti al sondaggio è stata raccolta e successivamente categorizzata in fasce d'età per l'analisi.
2. **Preferenze e Opinioni:**
   - **Valutazioni delle caratteristiche del cibo McDonald's:** Ai clienti è stato chiesto di valutare aspetti specifici del cibo come "gustoso", "conveniente", "salutare", "grasso", ecc.
   - **Frequenza delle visite:** Questa variabile indica quanto spesso un cliente visita McDonald's (ad esempio, più volte alla settimana, una volta alla settimana, una volta al mese, ecc.).
   - **Gradimento generale:** Una valutazione complessiva dell'esperienza McDonald's espressa su una scala da -5 a +5.

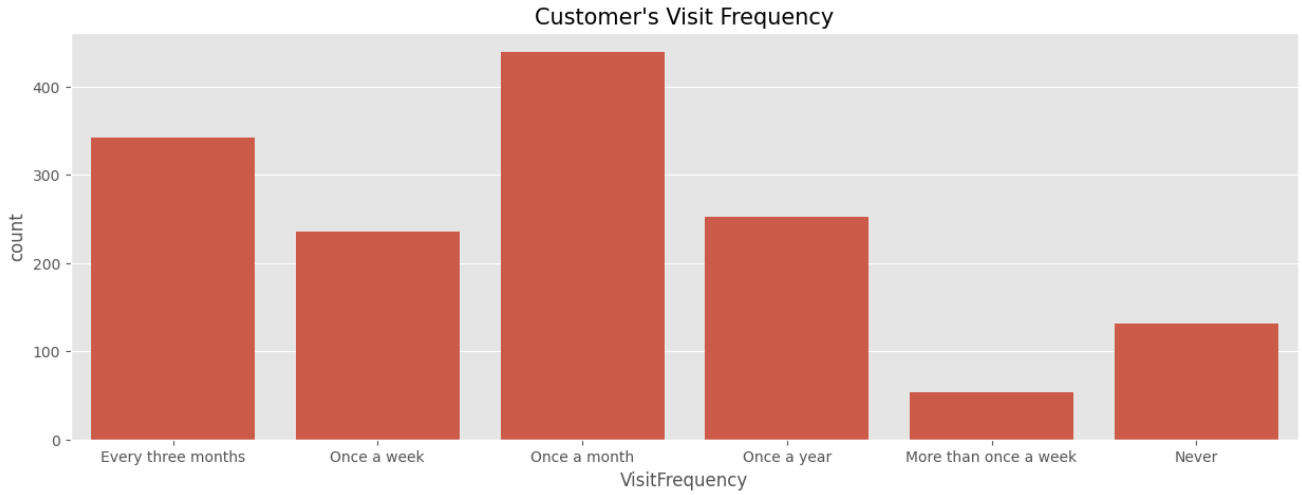
Visualizzazioni grafiche con osservazioni



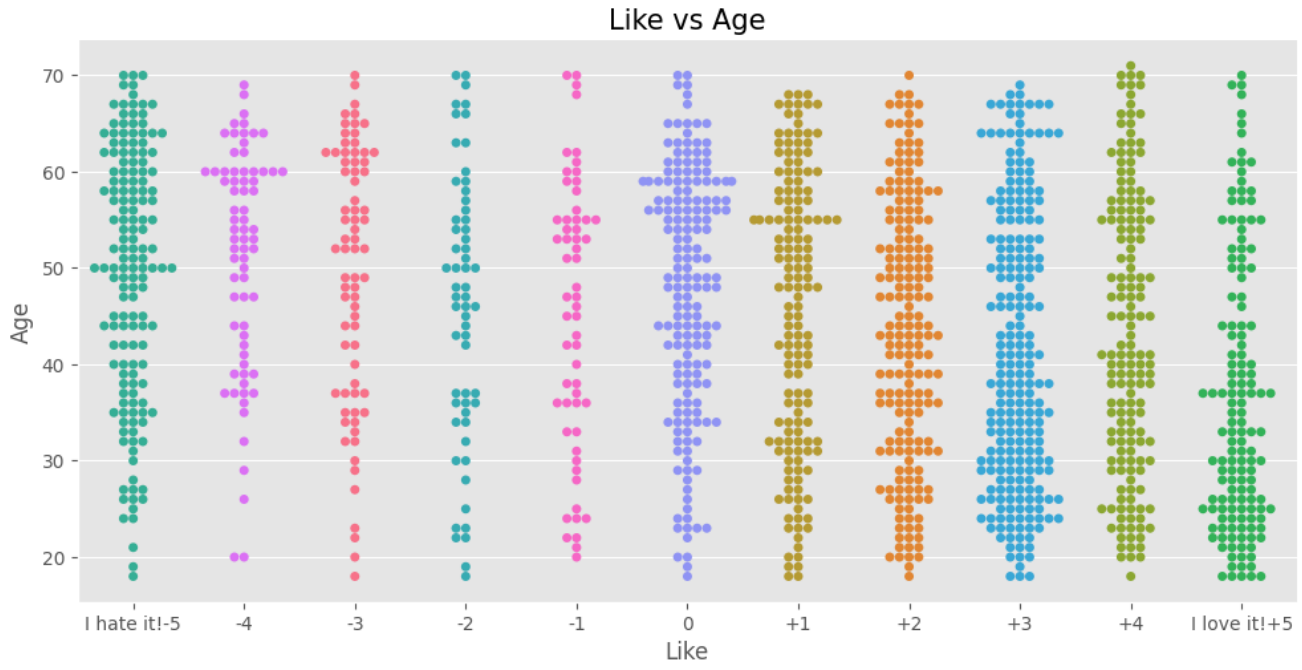
- Il grafico a torta sulla distribuzione di genere mostra una leggera **prevalenza di clienti di sesso femminile (54.23%)** rispetto a quelli di sesso maschile (45.77%).
  - *La differenza non è molto marcata, suggerendo che McDonalds attrae in modo abbastanza equilibrato entrambi i sessi.*



- **Maggioranza tra i 30 e i 60 anni**: La maggior parte dei clienti ha un'età compresa tra i 30 e i 60 anni, con picchi intorno ai 35-40 anni e ai 55-60 anni.
  - Il grafico suggerisce che McDonald's potrebbe voler concentrare le proprie campagne di marketing sulle fasce d'età tra i 30 e i 60 anni.
- **Pochi clienti giovani e anziani**: Ci sono relativamente pochi clienti sotto i 25 anni e sopra i 65 anni.
  - L'azienda potrebbe anche voler sviluppare prodotti o promozioni che attraggano maggiormente i clienti più giovani o più anziani.



- **"Once a month" è la frequenza più alta**: Il grafico mostra chiaramente che la maggior parte dei clienti tende a frequentare il McDonald's una volta al mese e pochissimi invece lo frequentano più volte alla settimana.
  - L'azienda potrebbe modellare le campagne affinché i clienti siano spinti ad accedere al servizio più frequentemente dividendo questo divario tra le 2 frequenze di visita



- **Apprezzamento generale**: I clienti tendono a dare più votazioni positive rispetto a quelle negative.
- **Fasce d'età più grandi**: La fascia d'età dei clienti nel range (50, 70) è quella che non ha apprezzato particolarmente il McDonald's, nonostante ci siano comunque clienti che lo apprezzino.
- **Fasce d'età più giovani**: Quasi tutti i clienti più giovani apprezzano il servizio.

Problemi o limitazioni dei dati

1. **No dati mancanti:** Non sono presenti valori mancanti nel dataset, il che semplifica l'analisi e non richiede tecniche di imputazione dei dati mancanti.
   - `data.isna().sum()` : Calcola e stampa il numero di valori mancanti (NaN) per ciascuna colonna.

2. **Mancanza di contesto per alcune variabili:** Alcune variabili, come la valutazione generale dell'esperienza McDonald's ("Like"), sono state raccolte su una scala numerica senza fornire un contesto chiaro ai partecipanti. Ciò potrebbe introdurre un certo grado di soggettività nelle risposte.
3. **Assenza di informazioni temporali:** Il dataset non include informazioni sul periodo in cui è stato condotto il sondaggio. Questo potrebbe essere rilevante se le preferenze dei clienti o le caratteristiche del cibo McDonald's fossero cambiate nel tempo.

Discretizzazione dati

La discretizzazione dei dati è un passaggio fondamentale in questo progetto per adattare le variabili al tipo di analisi da svolgere.

- Sono stati utilizzati dizionari Python per sostituire valori categorici con valori numerici.

	yummy	convenient	spicy	fattening	greasy	fast	cheap	tasty	expensive	healthy	disgusting	Like	Age	VisitFrequency	Gender
0	0	1	0	1	0	1	1	0	1	0	0	-3	4	2	0
1	1	1	0	1	1	1	1	1	1	0	0	2	3	2	0
2	0	1	1	1	1	1	0	1	1	1	0	1	4	2	0
3	1	1	0	1	1	1	1	1	0	0	1	4	4	4	0
4	0	1	0	1	1	1	1	0	0	1	0	2	3	3	1
5	1	1	0	1	0	1	1	1	0	0	0	2	3	2	1
6	1	1	1	1	0	1	0	1	1	1	0	2	3	2	0
7	1	1	0	1	1	1	1	1	0	0	0	5	1	4	0
8	0	0	0	1	1	0	0	0	1	0	1	-5	4	1	1
9	1	1	0	1	1	1	0	1	1	0	0	1	1	2	0
10	0	1	0	1	0	1	1	0	0	0	1	-2	3	2	0

Questo processo è stato applicato a:

1. **Colonne binarie (Sì/No, Maschio/Femmina):**
   - Colonne come `yummy`, `convenient`, `spicy`, ecc., che originariamente contenevano valori "*Si*" o "*No*", sono state convertite in variabili binarie 0/1.
     - Il dizionario `replacement_dict` mappa "*Si*" a 1 e "*No*" a 0. Analogamente, la colonna `Gender` è stata trasformata mappando "*Maschio*" a 1 e "*Femmina*" a 0.
2. **Colonna "Like":**
   - La colonna `Like`, che rappresentava un gradimento su una scala da "*I hate it! -5*" a "*I love it! +5*", è stata convertita in valori numerici da -5 a 5 utilizzando lo stesso dizionario `replacement_dict`.
3. **Colonna "VisitFrequency":**
   - La colonna `VisitFrequency`, che indicava la frequenza delle visite con valori testuali (ad esempio, "*Più di una volta alla settimana*"), è stata trasformata in una scala ordinale da 0 a 5.
     - Il dizionario `VisitSostitution` mappa ogni frequenza di visita a un valore numerico corrispondente, dove 0 rappresenta "*Mai*" e 5 rappresenta "*Più di una volta alla settimana*".
4. **Colonna "Age":**
   - La colonna `Age`, che conteneva l'età dei clienti, è stata discretizzata in fasce d'età. Sono stati definiti dei bin (`age_bins`) che rappresentano gli intervalli di età e delle etichette (`age_labels`) corrispondenti a ciascun intervallo.
     - La funzione `pd.cut` è stata utilizzata per assegnare a ciascun cliente l'etichetta della fascia d'età appropriata.

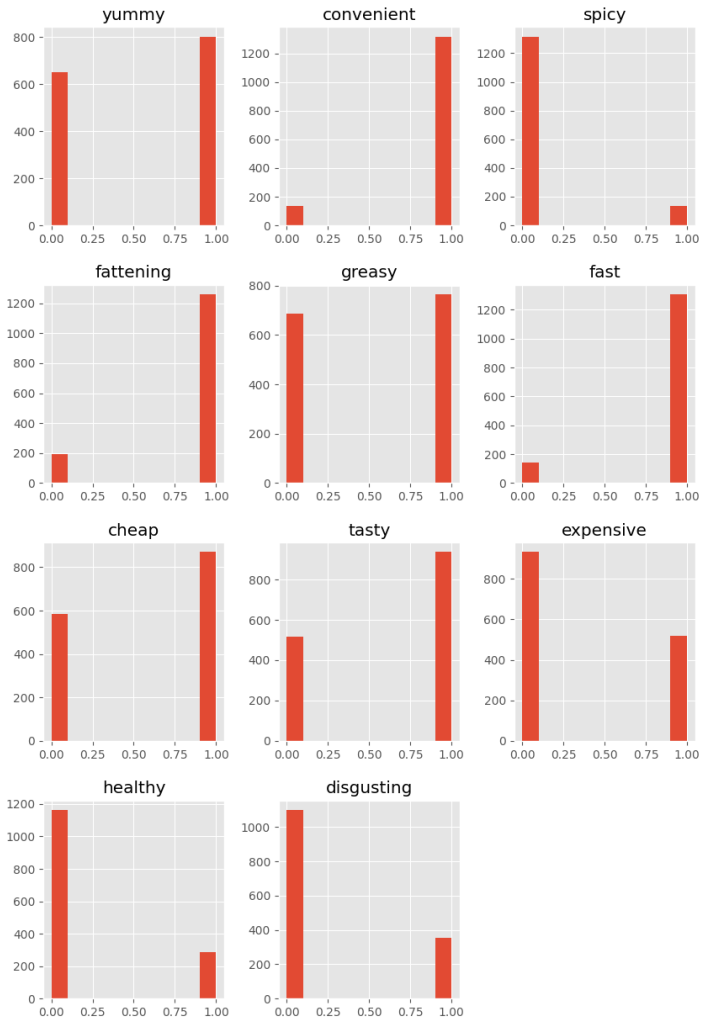
Limitazioni

- **Perdita di informazioni:** La discretizzazione comporta una perdita di granularità nei dati. Ad esempio, nella colonna `Age`, non si distingue più tra un cliente di 25 anni e uno di 30 anni, entrambi appartenenti alla stessa fascia d'età.
- **Scelta dei bin:** La definizione dei bin nella discretizzazione dell'età è arbitraria e potrebbe influenzare i risultati dell'analisi.

Analisi esplorativa (EDA)

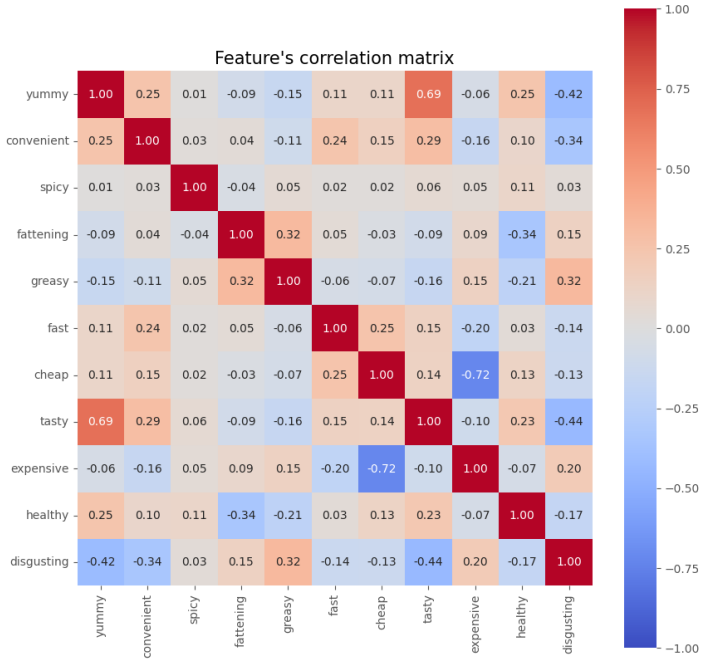
L'analisi esplorativa dei dati (EDA) è stata fase cruciale nel progetto di analisi dati. Si tratta di un processo di indagine iniziale in cui si utilizzano tecniche statistiche e visualizzazioni per:

- **Comprendere la struttura dei dati:** Quali sono le variabili presenti? Quali sono i loro tipi e distribuzioni?
- **Identificare relazioni e pattern:** Esistono correlazioni tra le variabili? Ci sono valori anomali o tendenze interessanti?
- **Verificare la qualità dei dati:** Ci sono valori mancanti o errori che devono essere affrontati?



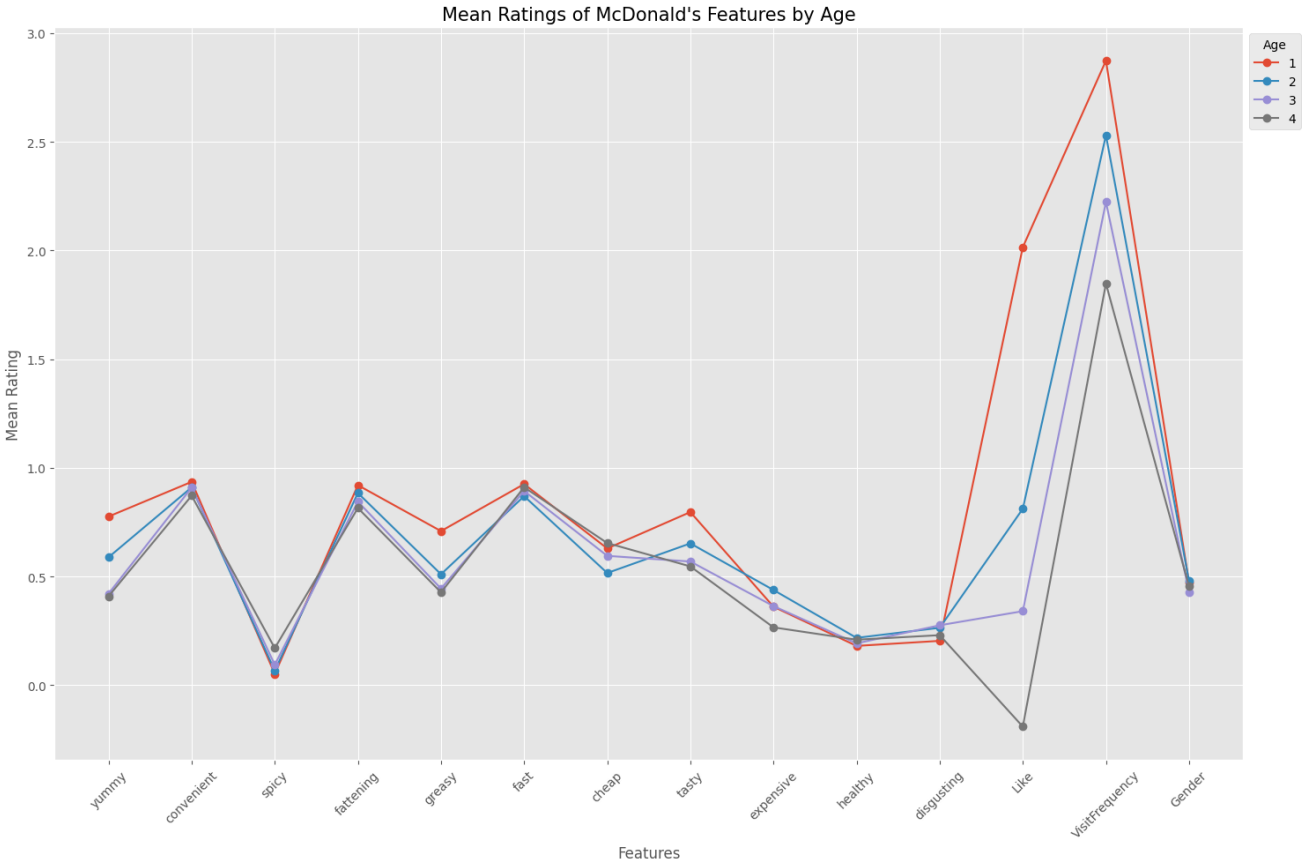
Il cibo è percepito principalmente come delizioso, conveniente, piccante, ingrassante, grasso, veloce, economico, gustoso, ma non sano. E' anche considerati costoso, ma non disgustoso.

- **Yummy, Convenient, Spicy, Fattening, Greasy, Fast, Cheap, Tasty**: Per questi attributi, la maggioranza dei giudizi è 1.
- **Expensive**: Anche per questo attributo, la maggioranza dei giudizi è 1, suggerendo che molti considerano il cibo della catena di fast food costoso.
- **Healthy**: La maggioranza dei giudizi è 0, indicando che il cibo non è considerato sano dalla maggior parte della popolazione campione.
- **Disgusting**: La maggioranza dei giudizi è 0, indicando che il cibo non è considerato disgustoso dalla maggior parte degli intervistati.



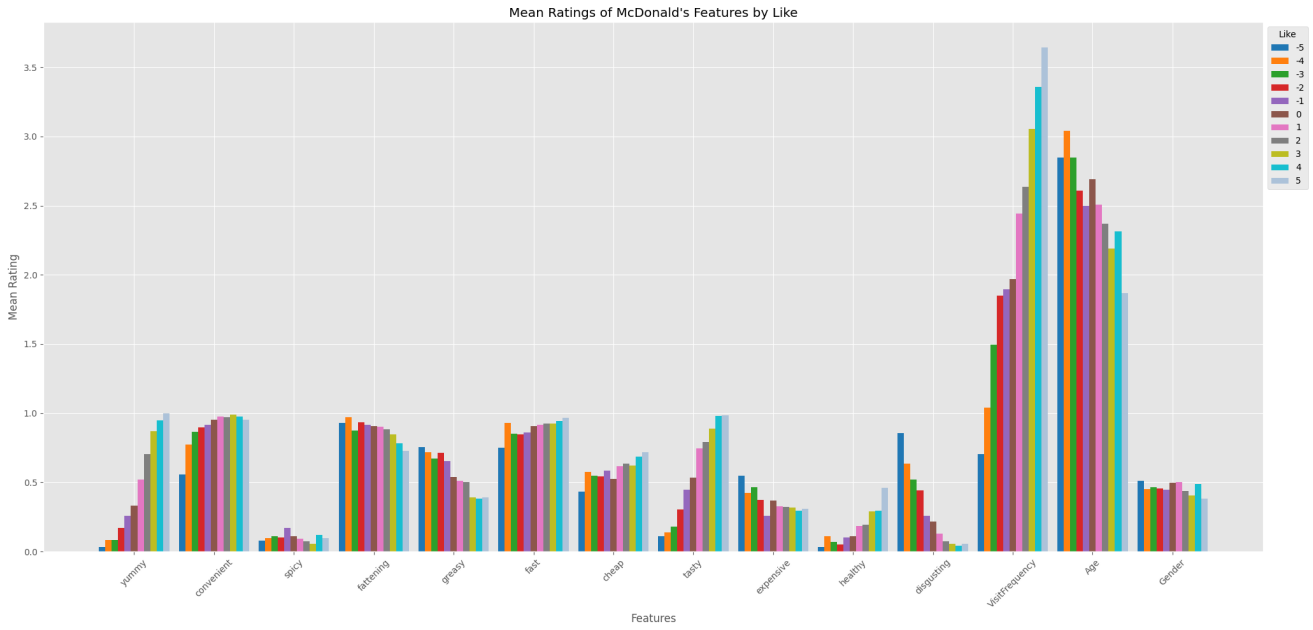
Dal grafico si nota che la maggior parte delle features non è correlata tra di loro, nonostante alcune eccezioni.

- **Correlazioni positive (rosso):** Tendenzialmente "*yummy*" è la feature con più correlazioni positive, in particolare "*yummy*" e "*tasty*" hanno il valore positivo più alto, ciò sta ad indicare che i clienti tendono a considerare un prodotto buono anche gustoso.
- **Correlazioni negative (blu):** In generale "*disgusting*" è correlata negativamente a molte features, tra cui "*yummy*", "*convenient*" e "*tasty*", come è anche intuibile. Le feature, tuttavia, con il valore più negativo sono "*expensive*" e "*cheap*", dunque i clienti che trovano il McDonalds costoso certamente non lo definiscono anche economico.



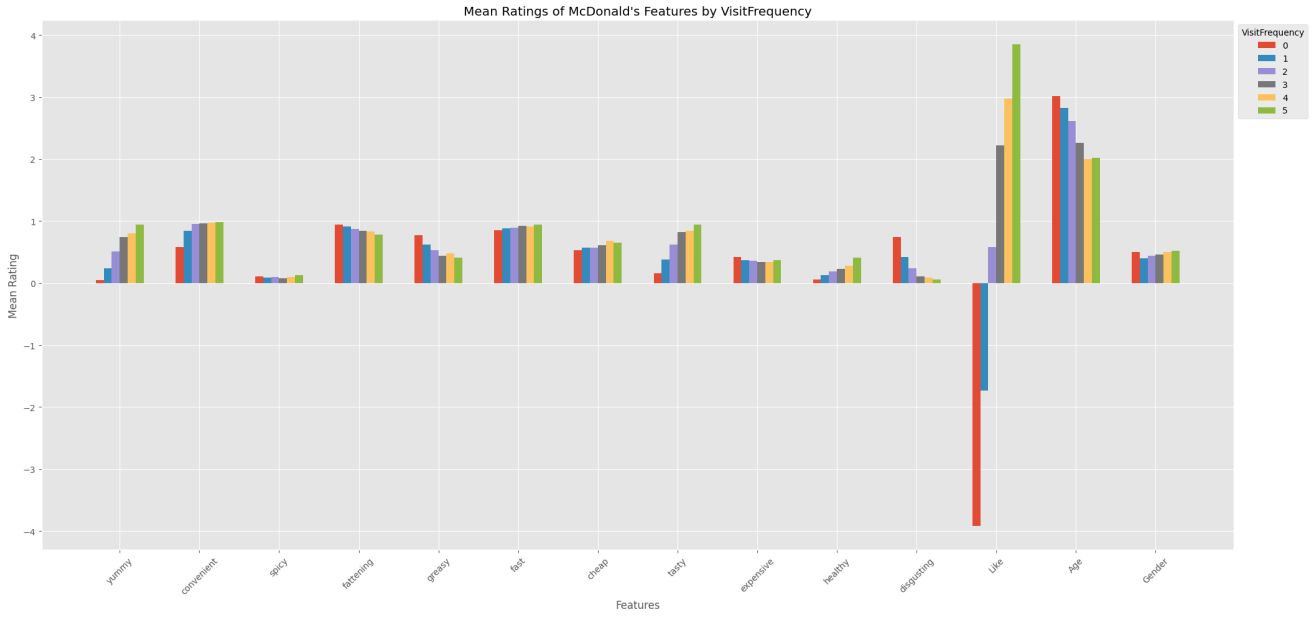
Nel grafico vengono rappresentate 4 linee che rappresentano i bins in cui abbiamo suddiviso le fasce d'età dei clienti. *Nonostante alcune features siano condivise da tutte le fasce d'età o, nel caso di gender, rappresentino un'equa distribuzione*, si possono notare alcuni andamenti specifici tra le varie fasce.

- **Fasce d'età più giovani (rosso)**: I clienti più giovani sono quelli che apprezzano maggiormente il servizio, motivo per cui tendono a frequentarlo di più e considerarlo gustoso nonostante siano quelli che lo considerino più "*greasy*".
- **Fasce d'età dai 33 ai 45 (blu)**: Questa fascia di clienti è quella che si trova nella media, non ha opinioni particolari riguardo al servizio se non per il costo: questa fascia di età considera il McDonald's costoso.
- **Fasce d'età dai 45 ai 71 (viola e grigio)**: Queste due fasce d'età tendono ad avere opinioni condivise quasi per tutto. Il grafico infatti mostra che dai 45 anni in poi si tende ad apprezzare sempre meno il cibo McDonald's dunque evitano di frequentarlo nonostante lo reputino conveniente. Perciò si evidenza come i clienti in questa fascia di età preferiscano la qualità al risparmio, al contrario dei più giovani.



Il grafico mostra per ogni features l'andamento dell'apprezzamento: Per alcune di queste l'opinione non crea divari particolarmente evidenti nel livello di gradimento, come per "*spicy*" ad esempio. L'ultima feature mostra come non cambi l'apprezzamento in base al genere ma che abbia invece un impatto importante in base al livello di frequenza del McDondald's.

- **Apprezzamento generale**: I clienti tendono a dare votazioni positive/negative coerentemente all'opinione delle features che esprimono gradimento/sgradimento del servizio. Lo si può notare benissimo con "*yummy*" e "*disgusting*", dato che l'andamento di una è opposto all'altra.



Il grafico mostra per ogni features l'andamento della frequenza: I valori sono organizzati affinché quelli più piccoli indichino minor frequenza (0 = Never) e quelli più grandi invece una maggior frequenza (5 = More than once a week)

- **Andamento generale**: Si può notare infatti che i clienti più abituali sono quelli che apprezzano di più il McDonald's al contrario di chi lo frequenta di meno.
- **Persone che non sono mai andate al McDonald's(rosso)**: Dal grafico si mette in evidenzia che coloro che non hanno mai mangiato al McDonald non hanno un opinione positiva di esso.
  - L'azienda dovrebbe attenzionare l'opinione di queste persone affinché siano più invogliate a provare il servizio o comunque non influenzino negativamente l'opinione generale.

PCA

- **Il problema:** Ci si trova di fronte a un gran numero di variabili. Queste variabili possono essere correlate tra loro, rendendo l'analisi molto complessa e di difficile interpretazione.
- **La soluzione:** La PCA è una tecnica statistica che mira a ridurre la dimensionalità dei dati: cerca di combinare le variabili originali in un numero inferiore di nuove variabili, chiamate componenti principali (PC).
  - Queste PC sono costruite in modo da catturare la maggior parte della varianza presente nei dati originali, eliminando ridondanze e correlazioni.

La riduzione della dimensionalità è stata una fase cruciale per semplificare i nostri dati e prepararli per il clustering. *Abbiamo utilizzato l'analisi delle componenti principali (PCA) per raggiungere questi obiettivi:*

1. **Standardizzazione delle features:** Abbiamo standardizzato le variabili numeriche per garantire che avessero media zero e deviazione standard unitaria. Questo passaggio è importante per evitare che variabili con scale diverse dominino l'analisi PCA.
2. **Applicazione della PCA:** Abbiamo applicato la PCA per ridurre le dimensioni del nostro dataset, mantenendo le componenti principali che spiegano la maggior parte della varianza nei dati (PC1 e PC2).
3. **Analisi dei pesi:** Ogni PC è una combinazione lineare delle variabili originali. I coefficienti di questa combinazione, chiamati pesi, indicano l'importanza di ciascuna variabile originale nella definizione della PC.
4. **Visualizzazione dei risultati:** Abbiamo creato una matrice semplice, uno *scree plot* per valutare la varianza spiegata da ciascuna componente principale e un biplot per visualizzare la relazione tra le variabili originali e le componenti principali.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
0	-0.704334	0.437016	-0.268698	-0.872074	1.521184	-0.470160	0.030969	0.687116	-0.367598	-0.321161	1.701170
1	0.467820	-0.364277	1.596835	-0.004835	-0.462385	0.449321	-0.087351	-0.446003	-0.221855	-0.191268	1.467681
2	-0.191986	-1.712949	0.339413	3.368168	1.266802	-0.148058	0.606634	0.668576	-1.377226	1.259300	-0.128530
3	0.116991	1.155122	1.003913	0.469589	-1.141750	0.857182	-0.015843	-0.390275	1.578539	-0.092189	-0.233201
4	0.034724	1.390267	-0.792275	0.473031	0.270488	0.847963	1.804085	0.700019	-1.630339	-0.092449	-0.033144
5	1.611088	0.546626	0.434235	-0.741550	-0.206557	-0.465167	-0.671111	0.401152	0.451211	-0.159923	0.082956
6	1.162820	-2.554323	0.081259	2.758296	1.257602	-0.454271	-0.330374	1.611138	-0.291112	-0.161685	-0.251076
7	1.074743	0.819610	1.214468	-0.078911	-0.894950	0.092274	-0.055646	-0.574769	-0.242037	-0.129561	0.020773
8	-4.566454	-1.227881	-1.245089	-0.163407	-2.298765	-0.363442	-0.309906	0.469046	-0.198857	0.029066	0.080847
9	-0.122070	-1.537454	1.852180	-0.238391	0.061907	0.491465	-0.105498	-0.577937	-0.348401	-0.068196	0.056023

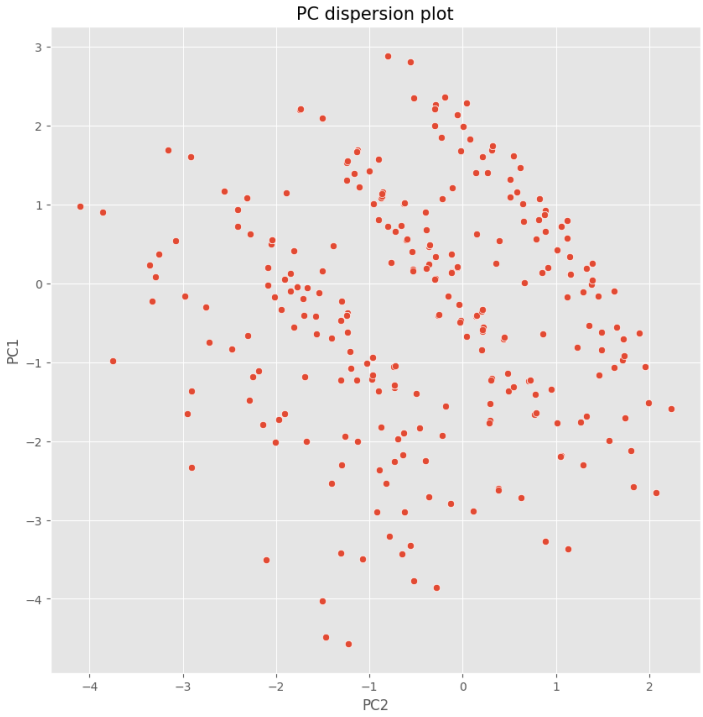
Dopo aver applicato la PCA, si ottiene una tabella che mostra la percentuale di varianza spiegata da ciascuna componente principale.

- *In genere, le prime PC spiegano la maggior parte della varianza totale, mentre le ultime ne spiegano una porzione molto piccola.*

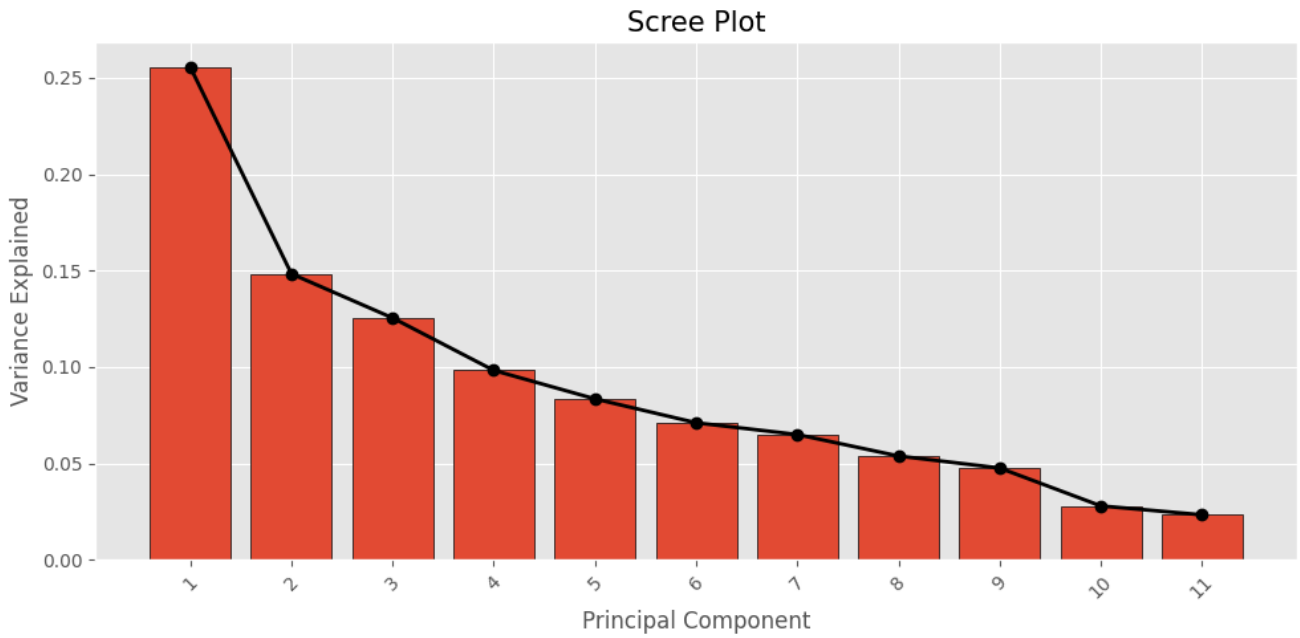
In sintesi..

- La PCA ci ha permesso di ridurre la complessità dei dati, facilitando l'identificazione di pattern e gruppi di clienti nelle fasi successive.
- L'analisi ha rilevato che la maggior parte della varianza nei dati può essere spiegata da due componenti principali. La prima componente principale è correlata positivamente con caratteristiche come "gustoso", "conveniente" e "veloce", e negativamente con "malsano" e "disgustoso". Questo suggerisce che la maggior parte dei clienti apprezza la velocità e la convenienza, ma è anche attenta alla salute.

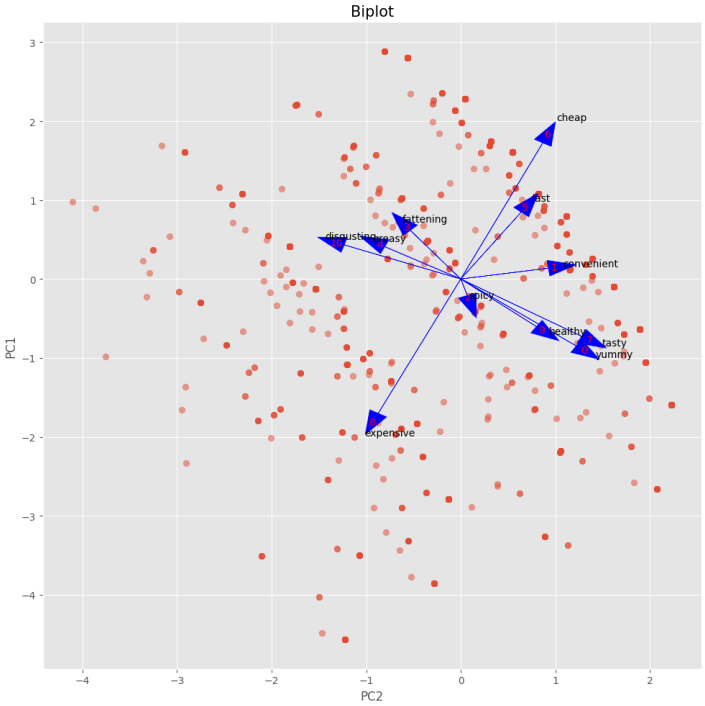
Visualizzazioni grafiche con osservazioni



Il grafico mostra una distribuzione relativamente omogenea dei campioni: come si nota dalla disposizione dei punti intorno al centro, senza cluster distinti. La maggiore varianza di PC1 suggerisce che le differenze maggiori nei dati sono catturate dalla prima componente principale.

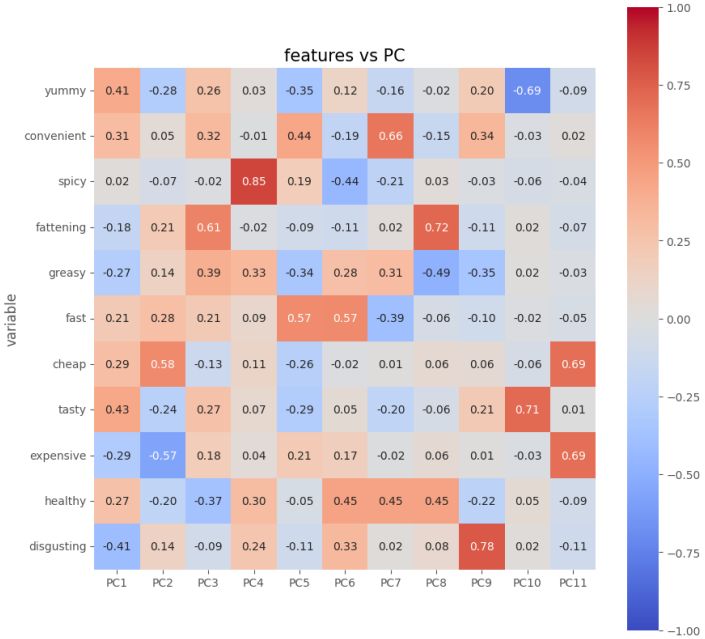


Il grafico mostra per ogni componente principale la proporzione della varianza totale: Nello specifico la prima componente è quella che rappresenta al meglio la varianza dei dati, con un distacco netto rispetto alla seconda componente, in cui possiamo trovare il così detto ginocchio, ovvero il punto di curva che descrive il numero di componenti principali ottimale.



Il grafico è una sovrapposizione di score e loadings e nello specifico si può notare il modo in cui ogni variabile contribuisce alle due componenti principali e con le altre variabili.

- **Correlazioni tra variabili**: Le variabili con correlazioni negative sono quelle che saltano subito all'occhio poiché sono quelle che costruiscono un angolo di circa 180° come cheap ed expensive, così come le features che costituiscono un'opinione negativa dell'azienda ("*fattening*", "*greasy*" e "*disgusting*"), che infatti hanno una correlazione positiva data dalla vicinanza delle frecce e della direzione, e quelle che invece ne creano una positiva ("*tasty*", "*yummy*", "*healthy*") anch'esse con correlazione positiva.
- **Correlazioni tra variabili e PC**: In generale il livello di parallelismo delle features rispetto agli assi delle PC ne determinano il contributo in quella componente.
  - Si nota infatti che "*convenient*" essendo parallela all'asse x ha un contributo maggiore nella seconda componente principale, mentre "*spicy*" al contrario è più influente nella prima componente principale anche se il suo contributo è minimo per via della lunghezza della freccia.



La heatmap tra features e PC permette di vedere il contributo di ogni singola feature per tutte le componenti principali. Si conferma il contributo minimo di "*spicy*" per quasi tutte le componenti tranne che per la PC4, mentre alcune delle feature mostrano il loro contributo più elevato nelle PC finali anche se proprio in quelle la percentuale di varianza, come ha mostrato lo Scree Plot è minima.

- **La scelta dunque delle prime due PC è confermata essere la più esatta** rispetto all'heatmap poiché mostra chiaramente che sono proprio quelle le componenti in cui le features contribuiscono più o meno tutte.

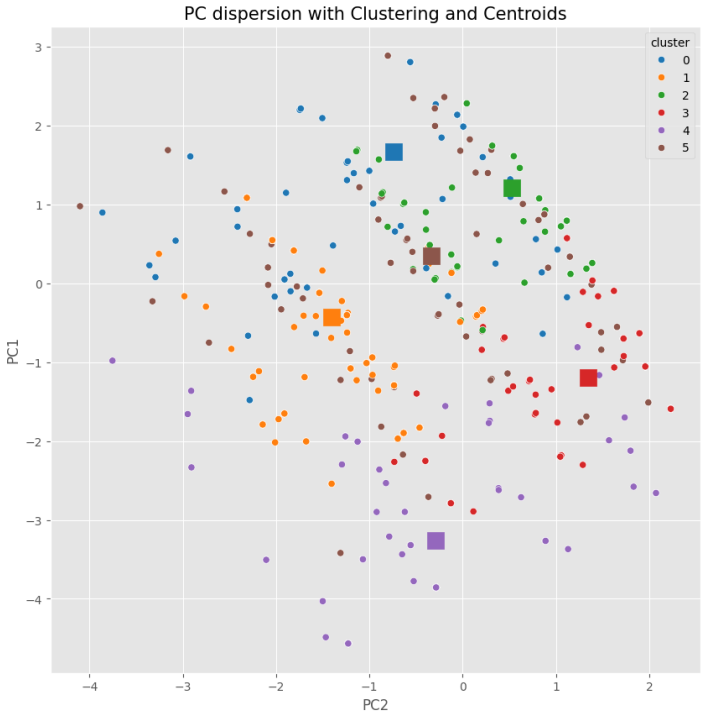
Clustering

Il clustering è stato il cuore del nostro progetto, consentendoci di identificare gruppi di clienti con preferenze simili. Abbiamo utilizzato due tecniche principali:

K-Means Clustering

Il clustering K-Means è un algoritmo di clustering partizionale. L'obiettivo è dividere i clienti in un numero predefinito di gruppi (K), in modo che i clienti all'interno di ciascun gruppo siano il più simili possibile tra loro e il più diversi possibile dai clienti degli altri gruppi.

- K-means è efficiente dal punto di vista computazionale, ma richiede di specificare in anticipo il numero di cluster.
- Utilizzo del **metodo Elbow** e **Silhouette Score** per determinare il numero ottimale di cluster.

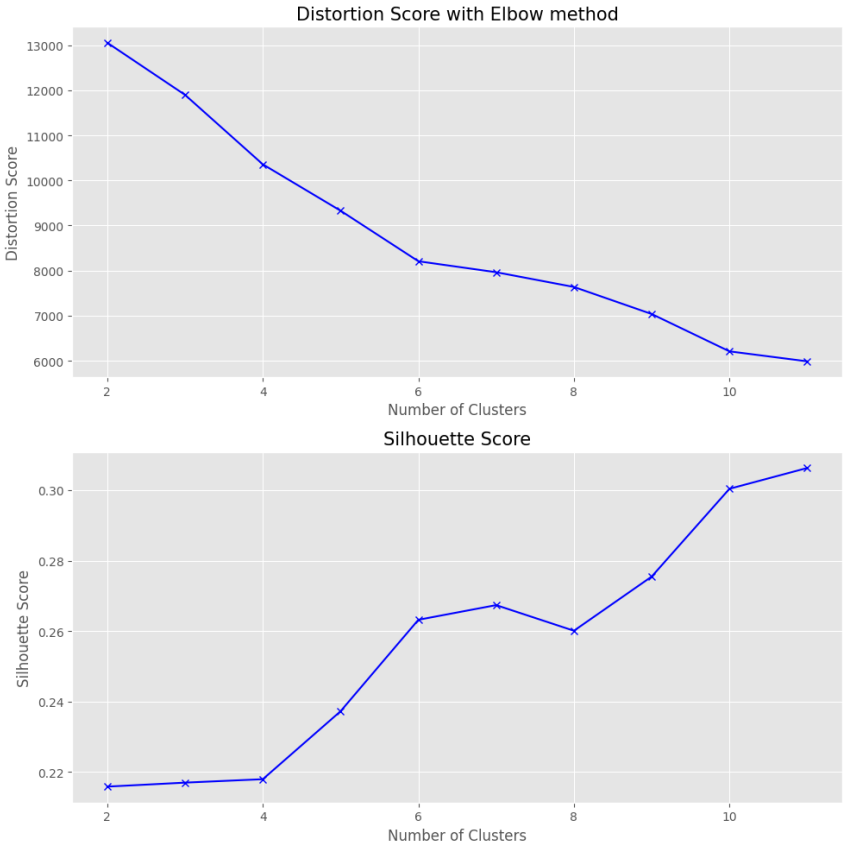


I grafico mostrano il modo in cui i punti vengono proiettati nelle 2 componenti e vengano suddivisi in cluster.

- Dal modo in cui sono distribuiti i cluster possiamo notare che **nonostante i centroidi siano abbastanza distanziati tra di loro, la forma dei cluster non è molto coesa** e i punti dati "invadono" altri cluster, in alcuni casi si sovrappongono. Probabilmente questo è dato dalla somiglianza dei punti dati che non permette dunque una separazione più netta tra i cluster e i punti dati stessi.

Determinare il numero ottimale di clusters

- **Metodi basati sulla varianza:** Questi metodi cercano di trovare il numero di cluster che minimizza la varianza intra-cluster e massimizza la varianza inter-cluster. Un esempio è il metodo del gomito (elbow method), che consiste nel tracciare un grafico della varianza spiegata in funzione del numero di cluster e cercare il punto in cui la curva si "piega" (come un gomito).
- **Metodi basati sulla silhouette:** Questi metodi valutano la qualità dei cluster in base alla silhouette, una misura di quanto bene un cliente si adatta al proprio cluster rispetto agli altri cluster. Si cerca il numero di cluster che massimizza la silhouette media.

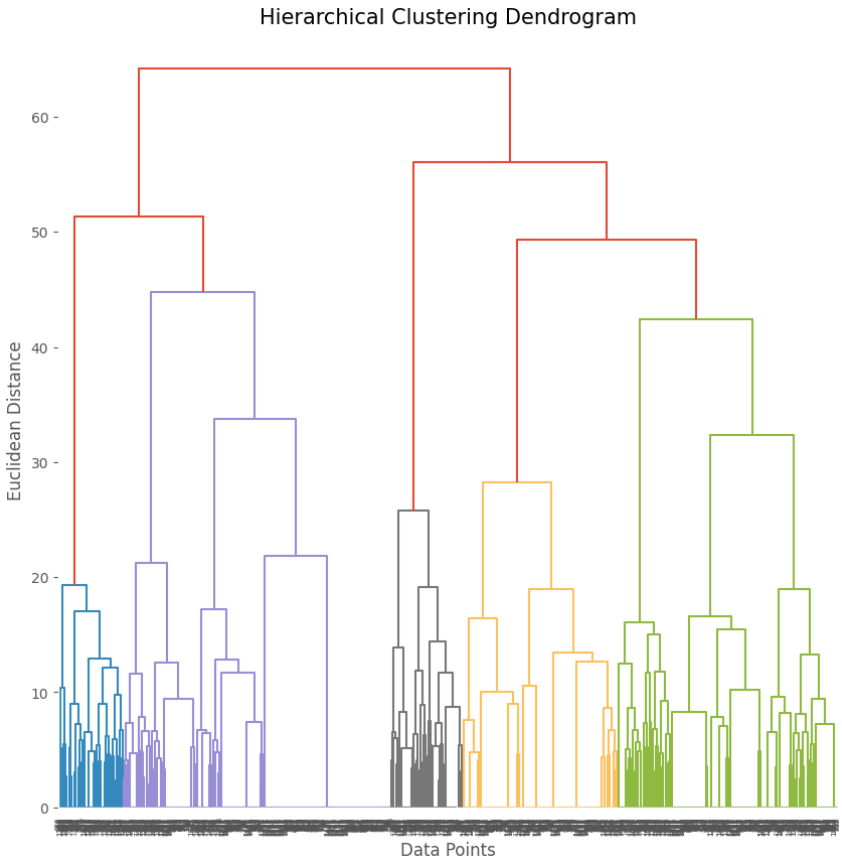


- **Elbow method**: Nel grafico risulta chiaro come il punto di gomito in cui la somma delle distanze al quadrato diventa marginale è dato da 6 clusters
- **Silhouette Score**: Il coefficiente di silhouette ad esempio è una metrica utilizzata per valutare la qualità di un clustering, misurando la coesione all'interno dei cluster e la separazione tra cluster.
  - A differenza del *metodo del "gomito"*, che si concentra sulla valutazione globale del clustering basandosi sulla somma totale entro il cluster di square (WSS), il coefficiente di silhouette fornisce informazioni sulla qualità del clustering a livello di singolo punto e di cluster.
  - In questo caso, nonostante il livello massimo di Silhouette Score è dato da 11 clusters, abbiamo optato per un numero di cluster che rappresentasse bene il silhouette score senza tuttavia ignorare ciò che avevamo visto attraverso l'Elbow Method.

- **Sono stati identificati 6 cluster di clienti**, ciascuno con preferenze e caratteristiche demografiche distinte.
  - Ad esempio, un cluster è composto principalmente da giovani che apprezzano il cibo gustoso e conveniente, mentre un altro è formato da clienti più anziani e attenti alla salute.

Clustering gerachico

- *Il Clustering gerarchico* è un algoritmo di clustering agglomerativo. Inizia considerando ogni cliente come un cluster a sé stante e poi unisce progressivamente i cluster più simili tra loro fino a ottenere un unico grande cluster utilizzando il metodo Ward.
  - Il clustering gerarchico non richiede di specificare in anticipo il numero di cluster, ma può essere meno efficiente dal punto di vista computazionale rispetto a K-means.



Il dendrogramma mostra come i dati sono stati raggruppati in 6 cluster principali con diverse strutture gerarchiche e vari livelli di similarità. La distanza euclidea fornisce un'idea della somiglianza tra i cluster e all'interno dei cluster stessi.

- I cluster a livelli più bassi rappresentano sotto-cluster che si uniscono in cluster più grandi man mano che si sale nell'albero.

Approfondimento sui dendrogrammi

Si voglia notare che in un dendrogramma l'asse delle Y indica la distanza o la dissimilarità tra i cluster, nel nostro progetto l'asse è stato chiamato "distanza euclidea".

- Chiamiamo *rami* le linee verticali che collegano i cluster rappresentando le fusioni tra i cluster, queste si incontrano in dei punti chiamati nodi che rappresentano invece i momenti in cui i cluster si sono uniti.
  - L'altezza di un nodo rispetto all'asse Y indica la distanza tra i cluster uniti in quel punto. Nella parte finale abbiamo le foglie descriventi i rami terminali del dendrogramma e quindi le osservazioni originali.
- La creazione di un dendrogramma ci ha permesso di poter facilitare la comprensione della struttura sottostante ai dati pertanto la similarità che intercorre tra le persone oggetto del nostro studio.
  - Il dendrogramma ci fa inoltre notare la presenza di outlier e anomalie nei dati, che potrebbero essere indicativi di errori di misurazione o di fenomeni insoliti.

Conclusioni

In questo progetto di analisi dei dati, abbiamo esplorato le preferenze dei clienti di McDonald's utilizzando tecniche di analisi esplorativa dei dati (EDA), riduzione della dimensionalità (PCA) e clustering (K-means e gerarchico).

1. Attraverso l'**EDA**, abbiamo ottenuto una comprensione approfondita delle caratteristiche dei clienti e delle loro opinioni sul cibo e sul servizio McDonald's. Abbiamo identificato le caratteristiche più apprezzate e quelle meno apprezzate, nonché le correlazioni tra le diverse variabili.
2. **La PCA** ci ha permesso di ridurre la complessità dei dati, identificando le componenti principali che spiegano la maggior parte della varianza. Questo ci ha aiutato a visualizzare i dati in modo più semplice e a preparare il terreno per l'analisi dei cluster.
3. Il **clustering**, sia K-means che gerarchico, ci ha permesso di identificare gruppi distinti di clienti con preferenze e caratteristiche simili. Abbiamo descritto i profili di questi cluster, evidenziando le loro caratteristiche demografiche, i loro

comportamenti d'acquisto e le loro preferenze specifiche.

Nonostante la clientela del fast food abbia una buona opinione nei riguardi del cibo servito, l'azienda dovrebbe avere maggior riguardo verso le fasce più anziane, investendo sulla qualità e non solo sul prezzo competitivo

Questo progetto ci ha offerto un'opportunità unica di applicare le conoscenze teoriche e tecniche a un caso reale e tangibile. Abbiamo imparato come affrontare le diverse fasi di un progetto di data science, dalla pulizia e preparazione dei dati all'interpretazione dei risultati.

Lavorare su questo progetto ci ha aperto alle potenzialità che le tecniche di machine learning sono in grado di offrire, mostrandoci come questi metodi possono essere utilizzati per risolvere problemi reali e per ottenere informazioni preziose.