DO YOU WANT SOME

CINEMA

JUST ME AND YOU ?

Lab. Analisi dei dati

Professor Furio Camillo

Sommario

1. Introduzione 3

1.1 Descrizione del questionario 5

2. Analisi dei dati 9

3. Analisi delle componenti principali 14

3.1 Individuazione dell'effetto taglia ed eliminazione dello stesso 14

3.2 Individuazione del numero di componenti principali 16

4. Clustering 18

***4.1* Scelta del metodo di *clustering* 18**

***4.2* Come interpretare i *cluster* 20**

4.3 Test Chi-Quadro 21

***4.4* Analisi dei *cluster* 23**

4.4.1 *Cluster* 1 – "Gli Aristogatti" 23

4.4.2 Cluster 2 – "I Mercenari" 24

4.4.3 Cluster 3 – La "Dead poets society" 25

4.4.4 Cluster 4 – Gli "Inside out" 26

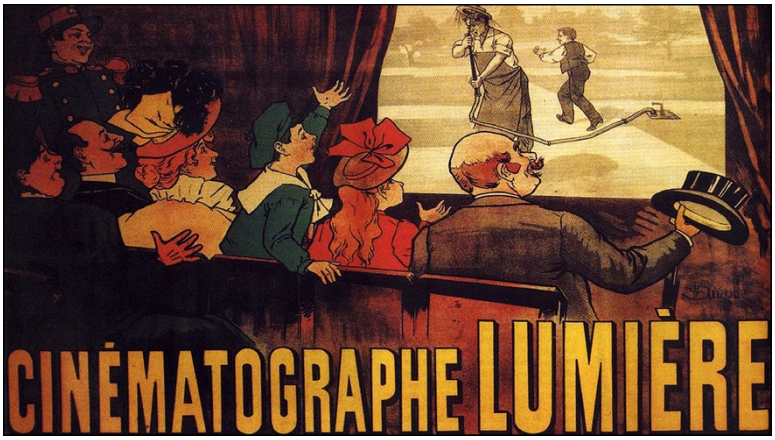
4.4.5 Cluster 5 – I drughi di "Arancia meccanica" 27

***4.5* Similarità tra *cluster* 28**

5. Conclusioni 29

1. Introduzione

Ad oggi andare al cinema o vedere un film a casa fa parte degli usi comuni, quasi della quotidianità di ogni individuo. Vedere un film sul grande schermo risulta essere scontato e alla portata di tutti. Gli individui, chi più chi meno, vedono nel cinema, in particolar modo nella vasta estensione che il termine può assumere, un'industria da cui la società non può più prescindere completamente. Ma non è sempre stato così.



L'esperienza cinematografica concettualmente vicina a quella che noi intendiamo, seppur lontana dalle moderne e avanzate tecnologie di cui disponiamo oggi e che contribuiscono a renderla unica, nasce nel 1895 grazie all'intuizione dei fratelli Louis e Auguste Lumière che decisero di proiettare, ad uso di più individui contemporaneamente, le immagini impresse su pellicola da una cinepresa[1].

Da quel momento iniziano ad essere proiettati i primi cortometraggi e film nelle sale, pubbliche e private, di gran parte dell'Europa, accogliendo diverse tipologie di pubblico, facendosi strada sia nelle fiere che nei salotti borghesi, finanche ad arrivare alla nobiltà. Il cinema divenne ben presto un affare di tutti.

Entrò fermamente a far parte della cultura popolare sia come forma di intrattenimento che come mezzo di comunicazione di massa, dando voce a chiunque volesse farsi ascoltare. Assunse, e ancora oggi interpreta, diversi ruoli in base al messaggio da far passare, che sia quello narrativo, propagandistico, pubblicitario, pedagogico o scientifico. Ben presto, molte sono le filosofie che hanno abbracciato il cinema, plasmandone i contenuti e gli attori, tuttavia per goderne appieno, prima che la televisione divenisse alla portata di tutti, era necessario recarsi in una sala appropriatamente predisposta alla visione. È così che, nel tempo, la sala

[1] L'invenzione della cinepresa, nonché della macchina che consentiva la visione dei fotogrammi impressi su pellicola, si deve a Thomas Edison che li realizzò nel 1889. I fratelli Lumière ebbero tuttavia l'intuizione di proiettare tali fotogrammi per permettere ad un pubblico, non più solo ad un individuo per volta, di goderne.

cinematografica, comunemente chiamata anch'essa "cinema", è passata dall'essere uno sterile luogo in cui vedere un film all'essere un'esperienza completa.

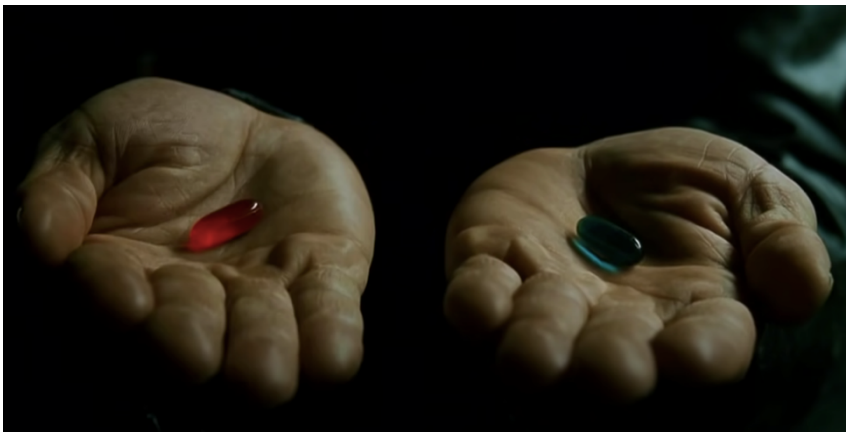
Andare al cinema, oggi, per i più, significa staccare dalla routine giornaliera, organizzarsi con la famiglia, gli affetti e gli amici più cari per trascorrere insieme una serata diversa di cui poi potersi ricordare, vivere l'emozione di vivere le avventure di qualcun altro concedendosi il lusso di essere solo un testimone e nel mentre, perché no, mangiare qualche pop-corn, normale, al burro o al caramello, e bere una coca-cola.

Apprezzare un film al cinema non è la stessa cosa di vederlo a casa, è diventato più un rito, ha assunto una sacralità propria. E forse è proprio questo che l'ha salvato, e malgrado con gran difficoltà, continua a salvarlo dalle recenti minacce contemporanee, come il covid, e dalla concorrenza sempre più spietata dello streaming da casa. Lo scoppio della pandemia ha avuto il duplice effetto di disabituare il pubblico ad andare in sala e di abituarlo a vedere i film in streaming. Cause di forza maggiore hanno obbligato l'individuo, che si fa pubblico e mente collettiva nel cinema, a rinunciare alla ritualità dell'esperienza in favore di una più sicura soluzione casalinga. Ad oggi, potendo guardare ormai alla pandemia come ad un qualcosa di passato, anche se non lontano, la speranza è quella di tornare a riempire le sale cinematografiche proprio in virtù di questa ritualità che rivestono, di spazio sociale ed esperienziale, di condivisione, in cui è possibile apprezzare al meglio quella che è definita come "la settima arte".

E proprio la passione e l'interesse comune, del nostro gruppo, per il cinema hanno alimentato la curiosità di scoprire se questi fossero altrettanto diffusi. Per farlo abbiamo deciso di elaborare un questionario finalizzato a determinare i motivi che portano le persone ad andare al cinema e quanto queste vivono fino in fondo l'esperienza dello stesso.

1.1 Descrizione del questionario

Il questionario è stato elaborato tramite l'uso di Google Forms e somministrato ad un pubblico eterogeneo al fine di ottenere un'analisi che fosse il più possibile veritiera. Successivamente, ci siamo avvalsi dell'uso del software SAS per gestire, analizzare e presentare i dati ottenuti.
Il questionario è composto da 25 domande in totale di cui le prime due volte ad indagare se piaccia o meno il cinema come esperienza, ed eventuali motivazioni a sostegno del no; cinque domande sociodemografiche; otto domande comportamentali, volte a scoprire eventuali abitudini dell'individuo relative al cinema; dieci domande finalizzate ad indagare proprio i motivi che spingono le persone ad andare al cinema, cui era possibile rispondere con una votazione da 1 a 7.



Il questionario è stato diffuso in italiano ed aveva ottenuto, al momento dell'analisi, 152 risposte.

Il questionario è stato articolato come di seguito:

1. ***Ti piace andare al cinema?***

   - Si
   - No

2. ***Se hai risposto "No", seleziona il motivo per cui non ti piace andare al cinema***

   - Non mi piace vedere i film al cinema
   - Il cinema costa troppo rispetto al servizio che propone
   - Preferisco aspettare che escano i film in streaming e vederli a casa in tranquillità

3. ***Sesso***

- Uomo
- Donna

4. ***Età***

- Meno di 18 anni
- Tra i 18 e i 30 anni
- Tra i 31 e i 50 anni
- Oltre i 50 anni

5. ***Occupazione***

- Studente
- Lavoratore dipendente
- Lavoratore autonomo
- Pensionato
- Non occupato

6. ***Titolo di studio***

- Licenza media
- Diploma superiore
- Laurea triennale
- Laurea magistrale
- Post-Laurea

7. ***Provenienza***

- Nord Italia
- Centro Italia
- Sud Italia
- Isole

8. Quante volte al mese vai al cinema?

- Nessuna
- Una volta
- Tra le due e le tre volte
- Più di tre volte

9. Quando vai al cinema, ti limiti ad acquistare solo il biglietto?

- Si
- No

10. Il biglietto del cinema preferisci acquistarlo:

- In cassa
- Online

11. Hai un abbonamento al cinema?

- Si
- No

12. In quali posti preferisci sederti quando vai al cinema?

- In basso
- Al centro
- Nelle postazioni VIP
- In alto

13. Quanto sei disposto a spendere quando vai al cinema?

- Meno di 5€
- Tra i 5 e i 10€
- Più di 10€

14. Al cinema preferisci andarci in:

- Estate
- Inverno
- Indifferente

15. Quanto tempo sei disposto ad impiegare per raggiungere il cinema?

- Meno di quindici minuti
- Tra i quindici e i quarantacinque minuti
- Più di quarantacinque minuti

Dopo aver letto attentamente le affermazioni, esprima il suo grado di accordo o disaccordo, contrassegnando il numero corrispondente alla sua reazione:
1 (*Completamente in disaccordo*) - **2** (*Per la maggior parte in disaccordo*) - **3** (*Leggermente in disaccordo*) - **4** (*Né d'accordo né in disaccordo*) - **5** (*Abbastanza d'accordo*) - **6** (*Molto d'accordo*) - **7** (*Completamente d'accordo*).

16. Il cinema rappresenta un momento di evasione dalla vita quotidiana.

17. Il cinema è il luogo in cui vado con gli amici per divertirmi.

18. Vado al cinema per avere un maggior coinvolgimento con il film.

19. Il cinema è il luogo in cui vado per rilassarmi.

20. Preferisco vedere un film al cinema piuttosto che in streaming.

21. Il cinema mi permette di apprezzare particolari caratteristiche sonore e visive del film.

22. Il cinema è il posto giusto per poter apprezzare l'arte dei film.

23. Il cinema è il luogo in cui vado per passare una serata romantica.

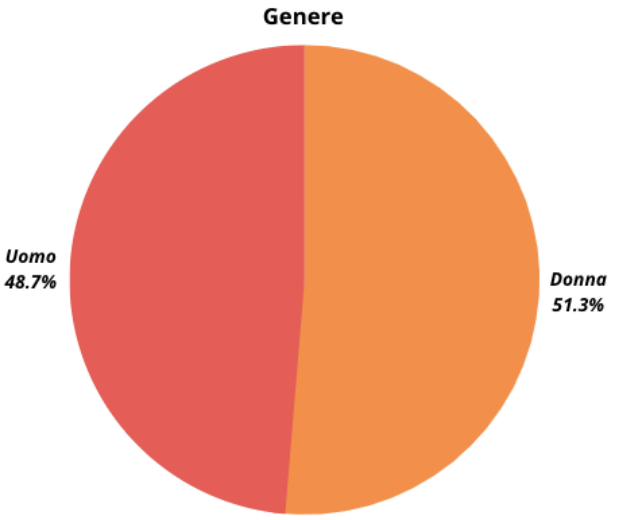
24. Il cinema è un momento di aggregazione con la famiglia.

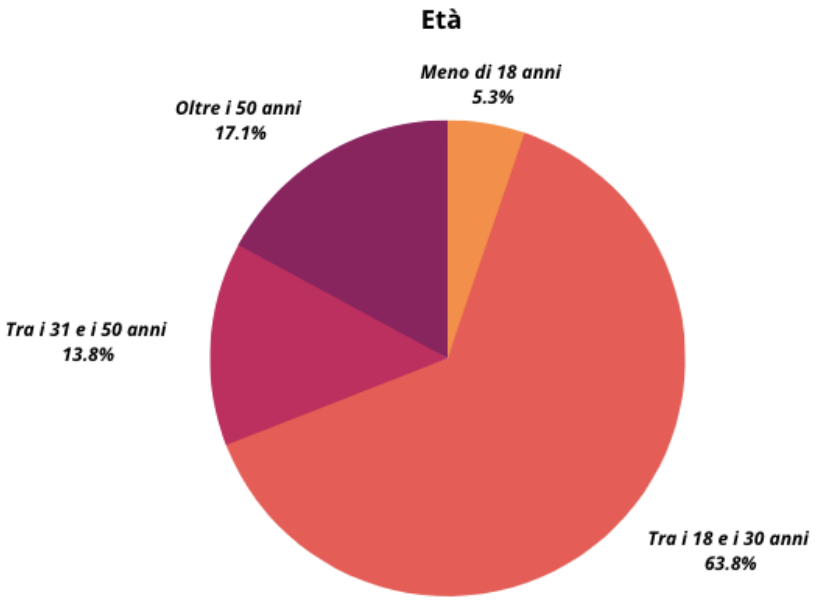
25. Il cinema è il luogo in cui posso dare libero sfogo alle emozioni.

2. Analisi dei dati

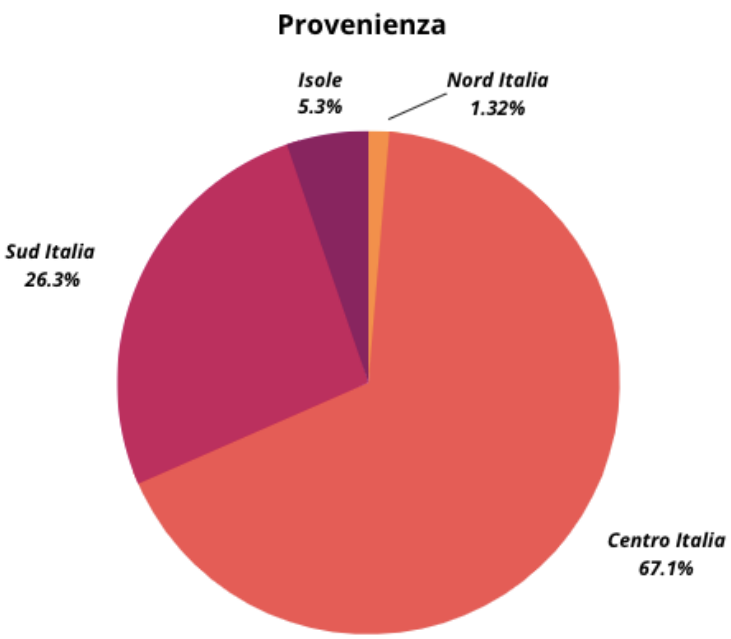
Prima di procedere con le dovute analisi delle risposte al questionario da noi somministrato è opportuno porre all'attenzione del lettore alcune considerazioni in merito alla natura del campione di individui a nostra disposizione. Ciò al fine di contestualizzare al meglio le risposte ottenute e di pensare a delle soluzioni che siano efficaci.

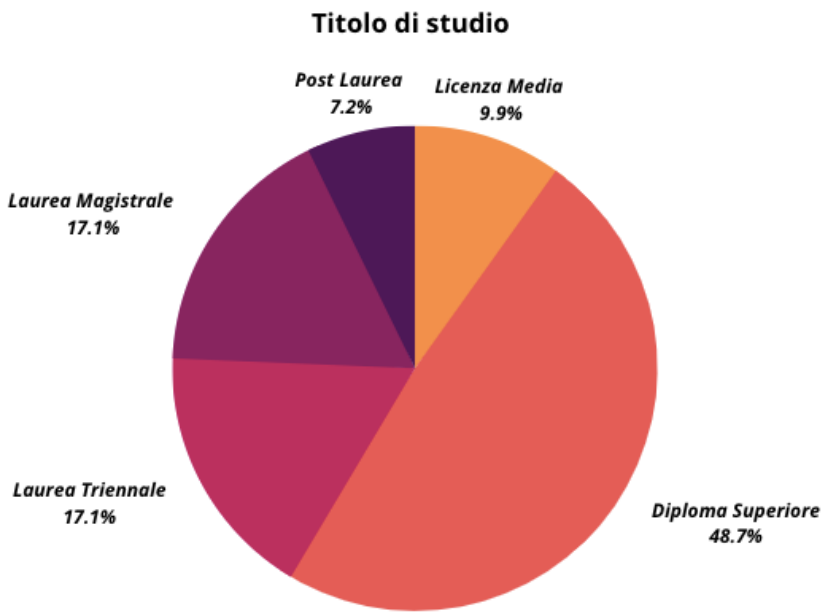
Il campione da noi esaminato è composto da 153 individui, quasi equamente distribuiti tra il genere femminile e quello maschile. È infatti possibile apprezzare, nel grafico a fianco, come, tra i rispondenti, il 48.7% risulta essere composto da uomini mentre il 51.3% risulta essere composto da donne. Tale distribuzione ci garantisce un'equa rappresentazione dei generi nelle risposte ottenute dal questionario. Ai fini di un'analisi descrittiva del campione stesso risulta utile andare a indagare altre caratteristiche della popolazione di riferimento quali possono essere l'età, la provenienza, il titolo di studio nonché l'occupazione. Per quanto riguarda la prima delle caratteristiche precedentemente citate, l'età, era possibile scegliere tra quattro modalità: "*Meno di 18 anni*", "*Tra i 18 e i 30 anni*", "*Tra i 31 e i 50 anni*" e "*Oltre i 50 anni*". La seconda è risultata essere quella estremamente più corposa, rappresentando il 63.8%; seguono gli individui con più di 50 anni, con il 17.1%; quelli tra i 31 e i 50 anni, con il 13.8%; chiudono i minorenni, che hanno raggiunto appena il 5.3%. Il campione risulta quindi essere composto, per più dei due terzi, da giovani e giovanissimi.

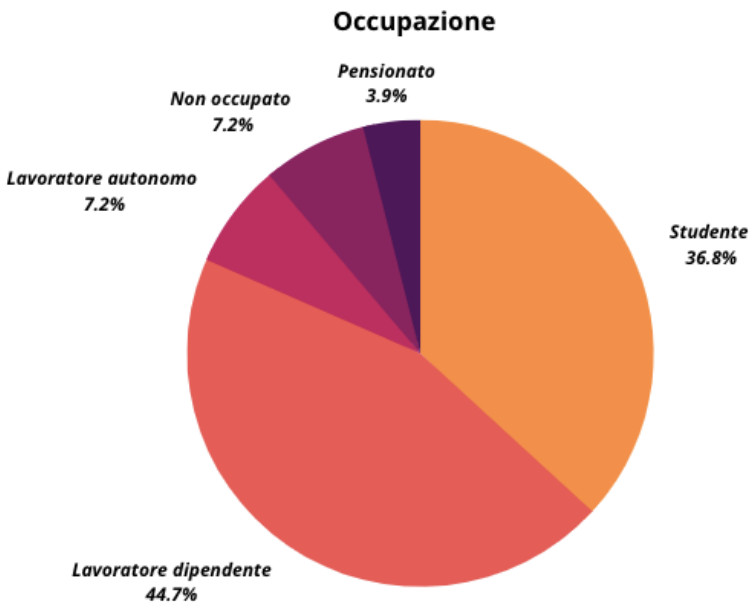




Per quanto riguarda, invece, la provenienza, il nostro campione di individui risulta essere un po' sbilanciato, in quanto quasi il 70% proviene dal centro Italia, mentre il restante 30% si distribuisce, sempre in modo disomogeneo, tra sud, nord e isole. È possibile, infatti, apprezzare come più del 25% delle risposte provengano dal sud, il 5% dalle isole e appena l'1% dal nord. Tale forte squilibrio è da imputarsi principalmente al fatto che noi redattori del questionario proveniamo dal centro e dal sud Italia quindi, avendo somministrato la survey ai nostri contatti, anche i rispondenti provengono essenzialmente dalla stessa parte del Paese. In riferimento al titolo di studio, invece, la distribuzione è risultata essere più o meno in linea con i dati nazionali. Le modalità più esigue sono emerse essere anche quelle più lontane, in termini di livello di istruzione, tra loro, ovvero la licenza media, che ha raggiunto il 9.9%, e i titoli di studio conseguibili dopo la laurea magistrale, che ha raggiunto, invece, il 7.2%. A farla da padrone, con quasi il 50%, spicca il diploma superiore, seguito da una percentuale pressoché uguale di laureati in triennale e di laureati in magistrale. Come poc'anzi detto, i dati risultano essere in linea con quelli nazionali.

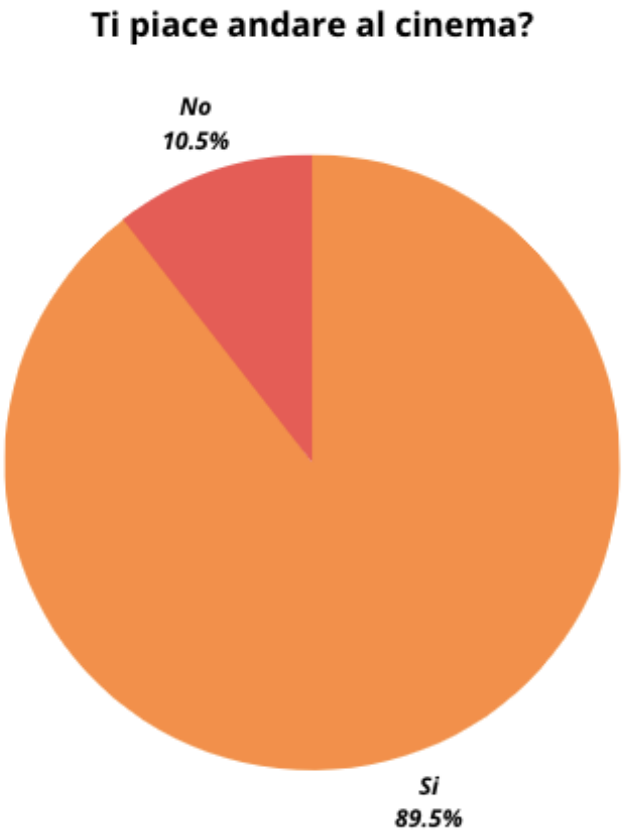




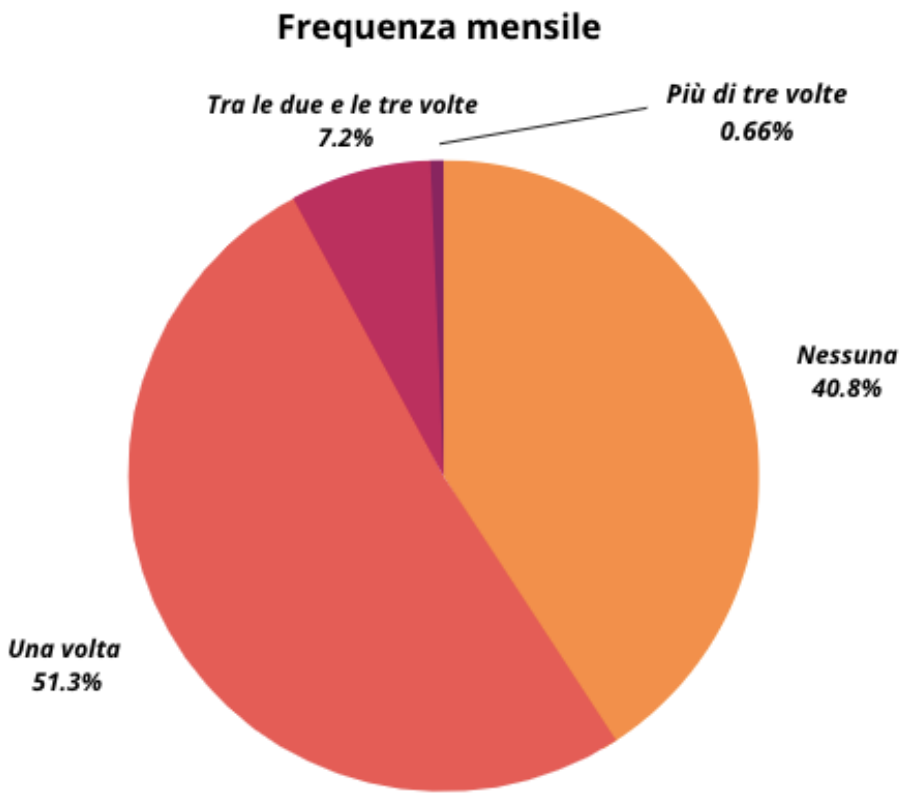


Per quanto riguarda, infine, lo stato di occupazione dei rispondenti, il 36.8% del campione è risultato essere composto da studenti, di ogni ordine e grado; il 44.7% è composto da lavoratori dipendenti; il 7.2% è rappresentato da lavoratori autonomi; anche i non occupati risultano essere il 7.2% mentre la restante parte del campione, corrispondente al 3.9%, risulta essere composto da pensionati. Come è possibile apprezzare nel grafico qui sopra, quindi, le modalità principalmente rappresentate sono quelle di studente e lavoratore dipendente.

Dopo aver analizzato il campione da un punto di vista sociodemografico è interessante andare ad esaminare anche le percentuali emerse dalle risposte ad alcune delle domande somministrate.

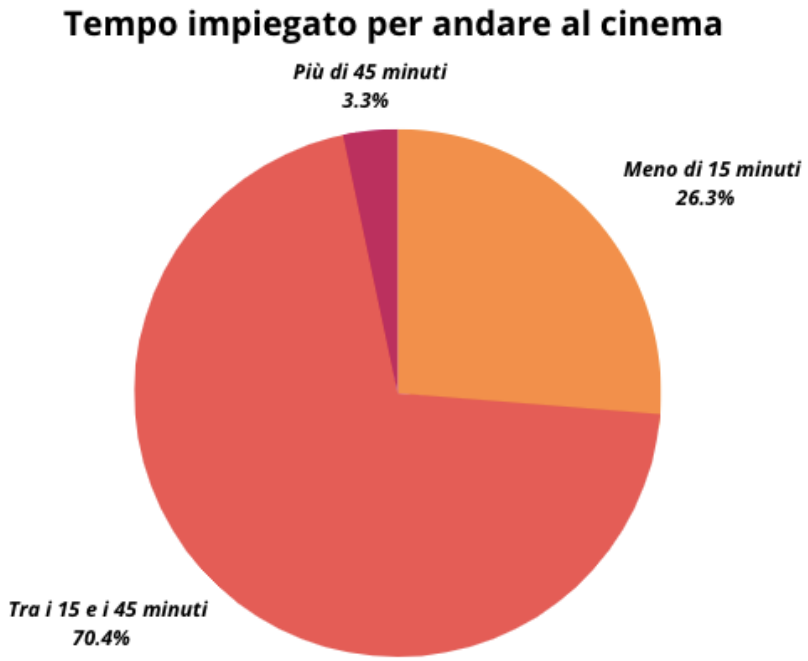


Come si può apprezzare dal grafico precedente, solo al 10% di tutti i rispondenti non piace andare al cinema. A nostro avviso è importante riportare tali percentuali, in quanto essere coscienti del fatto che al 90% della popolazione in questione piace andare al cinema, da una parte ci fa capire che "l'esperienza cinema" ancora gode di attrattiva nei confronti del pubblico e dall'altra che non è il cinema a non essere apprezzato, quanto ciò che gli ruota intorno e a esso viene associato. In sostanza, possiamo dire che alla gente piace, piacerebbe, andare al cinema, anche più spesso di quanto già vanno, ma una serie di fattori ad esso collegato portano l'individuo a desistere da tale volontà.



Ancora, importante è interpretare il grafico di cui sopra. Da esso emerge che nonostante al 90% delle persone totali del campione piaccia andare al cinema, il 40.8% non va praticamente mai al cinema con una frequenza mensile, quanto piuttosto, magari, con una frequenza più annuale. Emerge inoltre come il 51.3% dei rispondenti si reca al cinema al più una volta al mese, quasi sicuramente a godersi il Blockbuster mensile, mentre appena l'8% restante è un frequentatore che potremmo definire assiduo. Infatti, di questo 8%, il 90% si reca al cinema tra le due e le tre volte al mese, mentre il restante 10% ci va anche più di tre volte.

Infine, per puro scopo conoscitivo, è interessante sottolineare come più di un terzo del campione è disposto a spendere tra i 15 e i 45 minuti per raggiungere il cinema che ritiene essere il migliore per soddisfare le proprie aspettative dall'esperienza del cinema.



Dopo un'attenta analisi delle risposte possiamo quindi concludere che il campione è prevalentemente composto da individui del centro Italia, in possesso almeno del diploma superiore, di età medio-alta, cui piace andare al cinema anche se non sono assidui frequentatori.

3. Analisi delle componenti principali

3.1 Individuazione dell'effetto taglia ed eliminazione dello stesso

L'obiettivo dell'analisi è l'individuazione di diversi gruppi, i quali rappresenteranno le diverse tipologie di clienti e fruitori del cinema. Quindi, cercheremo di ottenere dei *cluster* sulla base dei 153 questionari e delle 10 domande poste.

Per fare ciò sarà necessario eseguire in prima battuta l'analisi delle componenti principali, la quale ci permetterà di eseguire una riduzione della dimensione del nostro *dataset* lungo le colonne (quindi lungo le 10 domande sottoposte agli intervistati). Così facendo, selezioneremo un numero di componenti principali minore rispetto al numero di domande proposte, ma che ridurranno il numero di variabili che descrivono un insieme di dati a un numero minore di variabili latenti, limitando il più possibile la perdita di informazioni.
La prima cosa da verificare è la presenza o meno dell'***effetto taglia*** - ***size effect*** - che misura la forza della relazione tra due variabili in una popolazione o la loro associazione. Infatti, la valutazione è spesso condizionata alla percezione della scala di misurazione suggerita agli intervistati. Il livello di coinvolgimento di questi ultimi, infatti, è spesso la vera causa della prima componente principale, ed è perciò necessario eliminare tale effetto.
La tabella qui di seguito contiene gli autovettori e ci indica la composizione della singola componente principale: i valori ivi presenti rappresentano i coefficienti di correlazione - tra la variabile e la componente principale - ponderati per il $\boldsymbol{\lambda}$, ovvero la lunghezza della componente principale, ossia la sua capacità di rappresentazione. Questi coefficienti di correlazione, in realtà, non sono altro che il coseno dell'angolo tra la variabile $\boldsymbol{x_i}$ e la componente principale stessa: più è alto il loro valore più tale variabile la caratterizza. Come è possibile notare dal riquadro rosso presente nella "***Tabella degli autovettori***"[2], la prima componente presenta tutti i coefficienti di correlazione positivi e ciò significa che non permette di discriminare al meglio gli individui.

[2] La "***Tabella degli Autovettori***" è stata ottenuta utilizzando la ***PROC PRINCOMP di SAS***

Autovettori											
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
d_1	d_1	0.308487	-.167358	0.559071	0.019445	0.220137	-.303073	0.438493	0.381545	0.160534	-.244849
d_2	d_2	0.283061	0.396685	0.199354	0.473459	-.353763	-.104580	-.158926	-.376503	0.373400	-.236826
d_3	d_3	0.326758	-.352192	-.169765	0.083287	0.348308	0.354386	0.373402	-.555420	0.183746	0.069578
d_4	d_4	0.326222	-.043939	0.558109	-.003502	0.170872	0.316883	-.507493	-.029034	-.269532	0.345686
d_5	d_5	0.344296	-.017567	0.066954	-.476075	-.615195	0.101131	0.264329	-.165062	-.377053	-.148475
d_6	d_6	0.325521	-.384290	-.243494	0.232443	-.390973	-.212368	-.027305	0.280678	0.195532	0.563312
d_7	d_7	0.343162	-.319122	-.338536	0.210216	0.137055	-.159090	-.376988	0.085924	-.347148	-.554523
d_8	d_8	0.263238	0.463032	-.236422	0.291286	0.051623	0.512197	0.265757	0.463871	-.157184	0.006611
d_9	d_9	0.297459	0.443525	-.174349	-.098851	0.324390	-.567281	0.105229	-.213239	-.296107	0.321520
d_10	d_10	0.333888	0.166501	-.206689	-.590782	0.133689	0.080094	-.301749	0.171107	0.561330	-.096664

Come è possibile apprezzare dalla tabella seguente, la presenza dell'***effetto taglia*** è spiegata anche dalla correlazione elevata tra la media dei voti e la prima componente principale. Si può notare, inoltre, un'alta correlazione tra la ***PC1*** e la media. Questo vuol dire che la prima componente principale discrimina gli individui semplicemente a partire dalla media che questi generano rispetto alle 10 variabili. In questo caso sarà necessario creare un *clustering* su tutto tranne che su ***PC1***, in quanto non siamo interessati a generare *cluster* su un costrutto che va a catturare un comportamento del tipo: "Mi piace il cinema e quindi assegno solo voti alti", o il contrario.

Coefficienti di correlazione di Pearson, N = 152 Prob > |r| sotto H0: Rho=0		
	avg_i	Prin1
avg_i	1.00000	0.99362 <.0001
Prin1	0.99362 <.0001	1.00000

Per risolvere questo problema abbiamo eseguito una trasformazione, dove abbiamo rimosso ad ogni individuo la propria media.

- Se il voto dell'individuo è maggiore rispetto alla sua media voti, allora il nuovo valore sarà:

$$New = \frac{Old_i - Avg_i}{Max_i - Avg_i}$$

- Se il voto dell'individuo è minore rispetto alla sua media voti, allora il nuovo valore sarà:

$$New = \frac{Old_i - Avg_i}{Avg_i - Min_i}$$

- Se il voto dell'individuo è uguale alla sua media voti - o il valore è mancante - allora il nuovo valore sarà pari a zero

Così facendo, abbiamo sfruttato una funzione lineare a pezzi, in quanto quest'ultima risulta essere lineare prima e dopo la media, dove il valore minimo viene spinto a -1 mentre il valore massimo a +1.

3.2 Individuazione del numero di componenti principali

Dopo aver risolto il problema legato all'***effetto taglia***, procediamo all'individuazione del numero di componenti principali che verranno utilizzate. Il numero di componenti da selezionare è pari a cinque in quanto queste, da sole, riescono già a spiegare il 67% della variazione dei dati. L'autovalore, infatti, non è altro che una lunghezza utile a considerare tutti gli individui all'interno della componente e ne misura la variabilità. Secondo la regola di ***Kaiser*** si dovrebbero includere nel modello finale tutte le componenti cui corrisponde un autovalore uguale o maggiore di 1. La selezione della quinta variabile è stata effettuata in quanto possiede comunque un autovalore molto prossimo a 1; in più ci garantisce un aumento della variazione spiegata del 10% circa.

Autovalori della matrice di correlazione

	Autovalore	Differenza	Proporzione	Cumulativa
1	2.00313376	0.60986260	0.2003	0.2003
2	1.39327116	0.18257305	0.1393	0.3396
3	1.21069811	0.09089520	0.1211	0.4607
4	1.11980291	0.17345111	0.1120	0.5727
5	0.94635180	0.12685951	0.0946	0.6673
6	0.81949229	0.09055959	0.0819	0.7493
7	0.72893270	0.07544427	0.0729	0.8222
8	0.65348843	0.02524933	0.0653	0.8875
9	0.62823910	0.13164936	0.0628	0.9503
10	0.49658974		0.0497	1.0000

Osservando la nuova "***Tabella degli autovettori***"

qui di sotto possiamo notare che l'***effetto taglia*** è stato risolto grazie alla trasformazione precedentemente effettuata. Tale tabella non fa altro che spiegarci quale aspetto viene catturato dalle componenti principali. La proiezione dei punti sulla componente principale avrà quindi degli errori che saranno ortogonali rispetto alla ***PC1***. Successivamente, troveremo la ***PC2*** che ne minimizzerà gli errori da un altro punto di vista, ortogonale rispetto alla ***PC1***. Così facendo abbiamo creato uno spazio dove i due vettori avranno un angolo retto, con $\boldsymbol{\lambda}\ \boldsymbol{PC1} > \boldsymbol{\lambda}\ \boldsymbol{PC2}$, e questo ci permette di vedere il grado di decremento esplicativo delle componenti principali: se il decremento è elevato allora stiamo "spiegando" rumore.

Grazie all'**Analisi delle componenti principali** riusciamo, quindi, a creare uno spazio ben segmentato che permetterà al nostro algoritmo di *clustering* di riuscire a trovare una migliore divisione dei gruppi.

Autovettori										
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
new_1	0.310397	-.451077	0.220122	0.239557	0.026787	0.221660	0.494063	0.276009	0.395483	-.260395
new_2	-.263174	-.291412	-.259972	0.493105	0.369565	0.001590	0.026044	-.615288	0.127349	0.054996
new_3	0.353096	0.158415	-.084332	0.232908	-.476940	-.624755	0.244356	-.231919	0.163884	0.175287
new_4	0.200397	-.460946	0.347994	0.337251	-.213148	0.028618	-.526708	0.082393	-.258669	0.339355
new_5	0.223122	0.251998	0.250306	0.249428	0.680521	-.453901	-.084489	0.251896	-.107632	-.102654
new_6	0.446256	0.106734	-.403389	0.001740	0.238750	0.330785	0.219373	0.107620	-.151077	0.614644
new_7	0.364283	0.258214	-.366680	0.370687	-.166162	0.280675	-.307079	-.015075	-.130792	-.557441
new_8	-.409933	0.100374	-.328017	0.349058	-.101264	-.074727	-.177246	0.580354	0.407695	0.205338
new_9	-.341770	0.249580	0.234687	0.439070	-.178343	0.155568	0.465826	0.084460	-.540418	0.029253
new_10	0.055456	0.510007	0.482955	0.131247	-.004526	0.368400	-.141851	-.255852	0.473103	0.202972

4. *Clustering*

4.1 Scelta del metodo di *clustering*

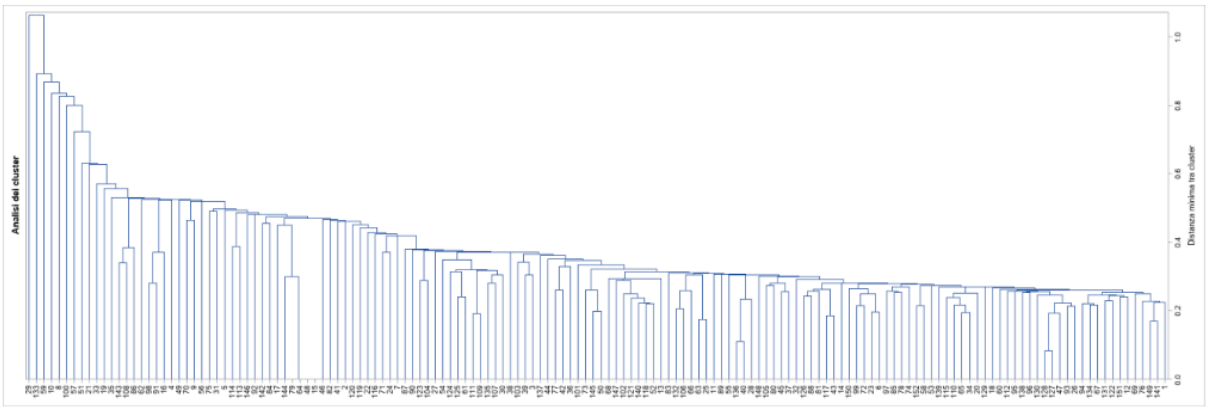
Il *clustering* si basa sui concetti di somiglianza e dissomiglianza, e questi concetti, a loro volta, si basano sulla distanza tra i vari punti: più un punto è vicino a un altro, più questi individui sono simili. La bontà delle analisi ottenute dagli algoritmi di *clustering* dipende molto dalla scelta della metrica, quindi da come è calcolata la distanza. Gli algoritmi di *clustering* raggruppano gli elementi sulla base della loro distanza reciproca e l'appartenenza o meno a un insieme dipende da quanto l'elemento preso in esame è distante dall'insieme stesso.
Con il *clustering* gerarchico viene costruita, appunto, una gerarchia di partizioni, caratterizzate da un numero crescente di gruppi, visualizzata da una rappresentazione ad albero – meglio noto come **dendrogramma** - in cui sono rappresentate le fasi di unione/partizionamento dei gruppi. Questi algoritmi sono a loro volta suddivisi in due classi:

- **Agglomerativo** → questi algoritmi assumono che inizialmente ogni *cluster* contenga un singolo punto; ad ogni passo, poi, vengono fusi i *cluster* più "vicini" fino ad ottenere un grande singolo *cluster*

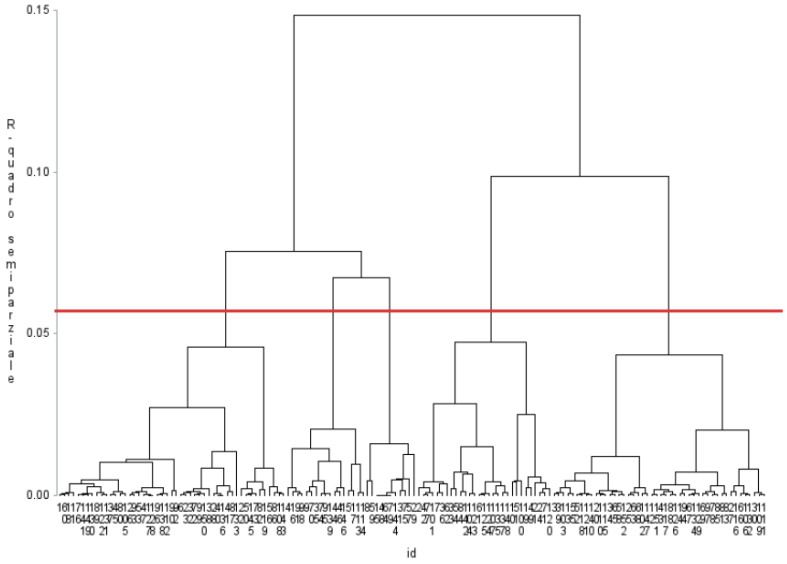
- **Divisivo** → questi algoritmi, invece, iniziano considerando lo spazio organizzato in un singolo grande *cluster* contenente tutti i punti, e di volta in volta lo dividono in due. A ogni passo viene selezionato un *cluster,* in base a una certa misura, e poi questo viene suddiviso in due *cluster* più piccoli

In entrambi i tipi di *clustering* gerarchico sono necessarie funzioni per selezionare la coppia di *cluster* da fondere ("**Agglomerativo**"), oppure il *cluster* da dividere ("**Divisivo**").
Ad esempio, il "***minimum linkage***" calcola la distanza tra due *cluster* come la minima distanza tra elementi appartenenti a *cluster* diversi. Il rischio di usare questo metodo è che spesso genera un ***clustering nested*** - cioè un annidamento, ovvero l'assenza di *cluster* naturali - vale a dire la presenza di punti equidistanti tra loro. Infatti, se nel grafico avessimo un punto ad elevata densità e se l'algoritmo del ***minimum linkage*** dovesse partire da quest'ultimo, si genererebbe il cosiddetto annidamento.

Come si può vedere dalla figura di seguito, è come se si costruisse un gruppo e a quel gruppo aggiungessimo, di volta in volta, un individuo e così facendo il gruppo diventerebbe più grande.



Per risolvere tale problema è possibile utilizzare il metodo di ***Ward***, il quale si basa sulla scomposizione della devianza di ***Fisher***. Tutto inizia dal concetto secondo cui la $DEV_{TOT} = DEV_{INT} + DEV_{FRA}$, quindi suggerisce di riunire, ad ogni tappa del processo, i due gruppi dalla cui fusione deriva il minimo incremento possibile della devianza "entro". Nel passare da ***k+1*** a ***k*** gruppi – aggregazione - DEV_{INT} aumenta, mentre ovviamente DEV_{FRA} diminuisce. Ad ogni passo effettuato seguendo il metodo di ***Ward*** si aggregano tra loro quei gruppi per cui vi è il minor incremento della devianza entro i gruppi stessi. Lì dove la devianza "esplode" si uniscono i punti più distanti, dunque la distanza di fusione aumenta. La figura sottostante mostra l'applicazione del metodo di ***Ward*** al nostro caso studio.



Sulla base del presente **dendrogramma** abbiamo individuato cinque *cluster*. La scelta è stata guidata dall'entità del "salto" necessario alla fusione dei gruppi che gli algoritmi di *clustering* gerarchici generano.

4.2 Come interpretare i *cluster*

Si potrebbe pensare che per interpretare i *cluster* generatisi basti confrontare la media del gruppo *j-esimo*, quindi se più alta o bassa, con quella dell'intero campione. Tuttavia, quando si effettuano tali ragionamenti è necessario considerare anche la variabilità. Poniamoci un quesito: se la prima variabile del *cluster j-esimo* ha una media di 8.5, mentre la media generale è di 7.06, la differenza che c'è, è da considerarsi come tanta o poca? Per saperlo occorre capire quale sia la probabilità che questa differenza sia generata dal caso oppure no. Per questo motivo si fa ricorso al ***t-value***, che non è altro che la differenza tra la media del gruppo e quella generale, il tutto standardizzato secondo un ***ϑ stimato***.

$$t_{value} = \frac{\bar{x}_{gruppo} - \bar{x}_{gen}}{\hat{\theta}}$$

Ovviamente per il ϑ stimato non è possibile calcolare la media di ϑ_{gruppo} e ϑ_{gen}, in quanto la prima è una distribuzione parziale di quella generale. Per questo motivo **Satterthwaite** suggerisce un aggiustamento in base ai gradi di libertà.

L'ipotesi nulla di questo *test* è che la media del gruppo e la media generale sono uguali; quindi, più la probabilità di osservare un valore diverso da zero tende a zero, allora tanto più la variabile caratterizza quel *cluster*.

Dopo aver calcolato tutti i ***t-value*** delle variabili sarà possibile ordinarli e quelli con i valori, in valore assoluto, più elevati mi permetteranno di identificare meglio il *cluster*. Tendenzialmente se il ***t-value*** sarà maggiore di due allora quella variabile sarà significativa nell'interpretazione del *cluster*. Questa stessa analisi può essere effettuata anche osservando i ***p-value***.

4.3 Test Chi-Quadro

Il *test* di indipendenza del ***Chi-quadrato*** è un *test* non parametrico che si effettua con lo scopo di verificare se esiste associazione, o dipendenza, tra due variabili categoriche.
Il sistema d'ipotesi del *test* è il seguente:

- H_0, ipotesi nulla: la *variabile uno* è indipendente dalla *variabile due*

- H_1, ipotesi alternativa: la *variabile uno* non è indipendente dalla *variabile due*

Questo *test* consiste nella costruzione di una tabella a doppia entrata, sulla base delle frequenze teoriche e di quelle osservate, da cui è possibile confrontare questi valori e calcolare il ***Chi-quadro***, che sarà un valore compreso tra 0 e $+\infty$. Grazie a questo *test* riusciamo a confrontare due variabili qualitative.
Dopo averlo calcolato tale grandezza è necessario conoscere qual è la probabilità che ci siano dei ***Chi-quadro*** più grandi di quello da noi calcolato e più il nostro ***Chi-quadro*** sarà grande, più sarà difficile trovare valori che siano più elevati. Oltre ad utilizzare tale metodo di confronto tra *test* e valore critico, è possibile procedere al confronto tra il ***p-value***[3] e il livello di significatività (α):

- Se il ***p-value*** è minore del livello di significatività prefissato, allora sarà necessario rifiutare H_0 affermando che le due variabili possono essere considerate dipendenti per un livello pari a 0.05

- Se il ***p-value*** è maggiore del livello di significatività prefissato, allora non si dovrà rifiutare H_0 ma si dovrà affermare che non è possibile concludere che le due variabili sono dipendenti

Più sarà basso il ***p-value*** più quella variabile sarà determinante nella generazione delle opinioni sugli individui osservati e questo ci permetterà di descrivere al meglio il *cluster*. Dalla tabella sottostante è possibile vedere i ***p-value*** delle variabili per ciascun *cluster*, dove quelli

[3] Il ***p-value***, o valore p, è la probabilità di ottenere uno specifico insieme di osservazioni nel caso in cui l'ipotesi nulla fosse vera

evidenziati in rosso rappresentano la dipendenza tra la variabile analizzata ed il *cluster,* con un livello di significatività dello 0.2.

Cluster	sesso	Piace_andare_al_cinema	Età	Occupazione	Titolo_di_studio	Provenienza	Quante_volte_al_cinema
CL1	0,14	0,062	0,2	0,3	0,03	0,16	0,19
CL2	0,46	0,93	0,65	0,22	0,78	0,48	0,12
CL3	0,36	0,29	0,18	0,36	0,44	0,19	0,35
CL4	0,19	0,48	0,31	0,108	0,26	0,75	0,21
CL5	0,79	0,31	0,91	0,83	0,82	0,19	0,2

Cluster	Acquisti_solo_biglietto	Online_o_cassa	Dove sederti	Disposto_pagare	Stagione_cinema	Tempo per andare al cinema
CL1	0,26	0,65	0,28	0,6	0,72	0,82
CL2	0,67	0,6	0,24	0,99	0,42	0,84
CL3	0,18	0,83	0,12	0,24	0,98	0,77
CL4	0,35	0,75	0,84	0,81	0,21	0,06
CL5	0,63	0,21	0,3	0,34	0,98	0,52

4.4 Analisi dei *cluster*

4.4.1 *Cluster* 1 – "Gli Aristogatti"

Il primo *cluster* è composto da 11 rispondenti - circa il 7% del totale - principalmente giovani, con età compresa tra 18 e 30 anni (54%) e provenienti dal Sud Italia. Questo è il *cluster* di coloro che vedono il cinema come momento di condivisione e di intrattenimento, infatti danno molto peso al fatto di andare al cinema con la famiglia o con il proprio *partner* per passare una serata romantica. Inoltre, tali individui non sono interessati all'aspetto artistico (***t-value*** -4,2), visivo e sonoro del cinema (***t-value*** più alto pari a -6,43) e preferiscono guardare i film in *streaming* piuttosto che al cinema (***t-value*** –2,47). Con questo *cluster* vogliamo identificare coloro che vanno al cinema per passare una serata con figure intime, indipendentemente dal film o dall'evento. In questo *cluster* troviamo principalmente persone che hanno come titolo di studio al massimo un diploma superiore (81%). Leggendo tra le righe, ci è sembrato opportuno associare i membri di questo *cluster* al cartone "**Gli Aristogatti**", in quanto abbiamo ritenuto che le loro caratteristiche si confacessero bene a quelle dei personaggi dello storico cartone *Disney*.

label	tvalue_1	prob_1
Particolari visivi e sonori	-6,43254	3,83E-05
Aggregazione con la famiglia	4,956512	0,000164144
Arte del film	-4,20157	0,001373137
Coinvolgimento	-3,4396	0,004689976
Serata romantica	3,086738	0,009069573
Cinema meglio che streaming	-2,47157	0,03009238
Evasione	-2,25369	0,044380728

4.4.2 Cluster 2 – "I Mercenari"

Il secondo *cluster* è composto da 49 rispondenti - circa il 32% del totale - principalmente giovani, con un'età compresa tra 18 e 30 anni (67%) e provenienti dal Centro Italia. Qui troviamo coloro che vanno al cinema per delle occasioni particolari. Possiamo affermare che sono persone che non sono appassionate né del cinema né della sua arte (***t-value*** -4,77) e che non vengono coinvolte molto dal film (***t-value*** -1,99). Per questi motivi possiamo concludere che coloro che appartengono a questo *cluster* sono i tipici clienti dei film ***Blockbuster***[4], cioè dei veri e propri mercenari che vanno al cinema solo per cavalcare il successo di questi film e avere qualcosa di cui parlare con amici e parenti. In più, coloro che appartengono a questo gruppo, si limitano ad andare al cinema "nessuna" (39%) o al massimo "una volta" (47%) al mese e questo ci permette di avvalorare la loro occasionalità nell'andare al cinema.

label	tvalue_2	prob_2
Arte del film	-4,76717	7,19E-06
Aggregazione con la famiglia	-2,56409	0,011749785
Serata romantica	-2,16648	0,032738099
Coinvolgimento	-1,98487	0,050499344
Sfogo le mie emozioni	-1,58244	0,116674429

[4] I ***Blockbuster*** sono film dal grande successo commerciale.

4.4.3 Cluster 3 – La “Dead poets society”

Il terzo *cluster* è composto da 46 rispondenti - circa il 30% del totale - principalmente giovani con un’età compresa tra 18 e 30 anni (31%) e provenienti dal Centro Italia. Qui troviamo gli appassionati del cinema, cioè coloro che apprezzano l’arte del film (***t-value*** 7,58), i particolari visivi e sonori della sala (***t-value*** 6,15), il coinvolgimento che questa regala (***t-value*** 4,29) e che vedono il cinema come un momento di evasione (***t-value*** 2,06). Questa passione si può notare anche dal fatto che tali individui preferiscono sedersi nelle zone centrali del cinema, cioè nei posti dove la visibilità del film e l’audio sono tendenzialmente migliori. Inoltre, questi ultimi tendono ad acquistare esclusivamente il biglietto, evitando l’acquisto di cibo e bevande, probabilmente per godere appieno del film. Tale appassionato fervore ci ha ricordato il discorso di Robin Williams ai propri studenti nel film "***Dead poets society***" (“**L’attimo fuggente**”) ed è per questo che abbiamo insignito i membri di tale *cluster* del titolo di membri del *club* esclusivo di cui fanno parte i personaggi del film.

label	tvalue_3	prob_3
Arte del film	7,584352	1,35E-12
Particolari visivi e sonori	6,1462	4,37E-09
Serata romantica	-4,44539	1,80E-05
Coinvolgimento	4,287631	3,70E-05
Aggregazione con la famiglia	-3,53287	0,000620082
Amici	-3,23155	0,001675229
Evasione	2,0599	0,042482994

4.4.4 Cluster 4 – Gli “Inside out”

Il quarto *cluster* è composto da 29 rispondenti - circa il 19% del totale - principalmente giovani con un’età compresa tra 18 e 30 anni (48%) e provenienti dal Centro Italia. A questo *cluster* appartengono coloro che empatizzano tanto con il film che vanno a vedere e per questo motivo vedono il cinema come un luogo per sfogare le proprie emozioni (***t-value*** 4,55) che vengono generate dall’arte insita nei film (***t-value*** 3,79). Oltretutto, tali individui tendono a preferire il cinema allo *streaming* (***t-value*** 2,4), questo a sottolineare l’intenzione di esprimere le proprie emozioni in sala. Per questi motivi siamo sicuri che i membri di questo *cluster* hanno riso, pianto e si sono emozionati con le vicende di uno dei cartoni più belli e commoventi della recente storia del cinema, “***Inside out***”. Per descrivere ancora meglio questo gruppo possiamo affermare che i suoi membri sono perlopiù lavoratori dipendenti disposti ad impiegare tra i 15 e i 45 minuti per andare al cinema.

label	tvalue_4	prob_4
Sfogo le mie emozioni	4,547251	5,53E-05
Arte del film	3,784987	0,000389025
Aggregazione con la famiglia	3,872964	0,000414353
Evasione	-2,5995	0,013316612
Cinema meglio che streaming	2,423214	0,020063405
Relax	-1,93633	0,060918351
Serata romantica	1,692881	0,098852219
Amici	-1,59703	0,118929348

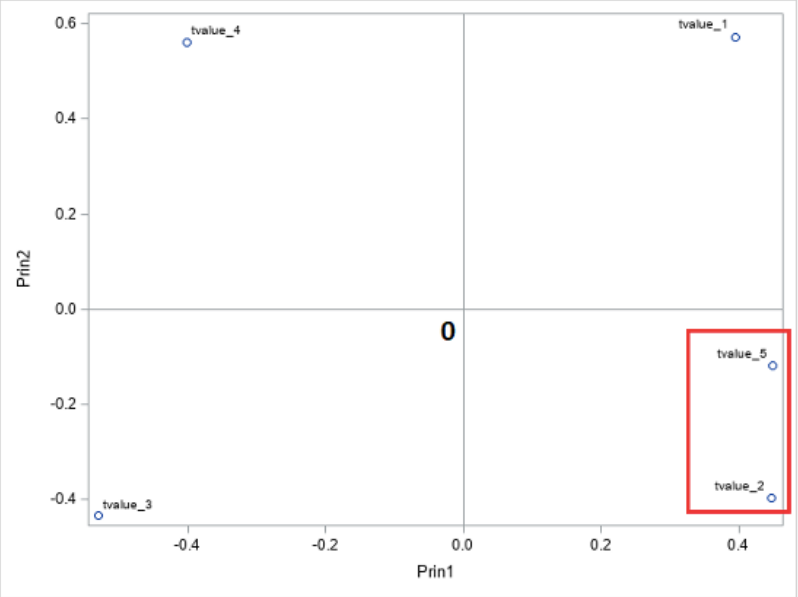
4.4.5 Cluster 5 – I drughi di “Arancia meccanica”

Il quinto *cluster* è composto da 17 rispondenti - circa il 11% del totale - principalmente giovani con un’età compresa tra 18 e 30 anni (65%) e provenienti dal Centro Italia. In questo *cluster* è possibile individuare coloro che vanno a divertirsi con figure più intime come amici (***t-value*** 11,84) o con il/la proprio/a compagno/a (***t-value*** 4,92). Queste persone non sono interessate a sfogare le proprie emozioni al cinema (***t-value*** -9,24), ma principalmente a divertirsi e passare una serata in allegria con gli amici e proprio per questo preferiscono lo *streaming* al cinema (***t-value*** -2,06). Non essendo degli appassionati, anche in questo cluster si ritrovano delle persone che non vanno spesso al cinema, al massimo vanno una volta al mese. Vista la natura dei motivi che spingono tali individui al cinema, abbiamo voluto paragonarli ai noti protagonisti di un film *cult* degli anni ‘70, “***Arancia meccanica***”.

label	tvalue_5	prob_5
Amici	11,8436	5,10E-17
Sfogo le mie emozioni	-9,24019	2,96E-16
Serata romantica	4,921587	7,51E-05
Cinema meglio che streaming	-2,06099	0,052778057

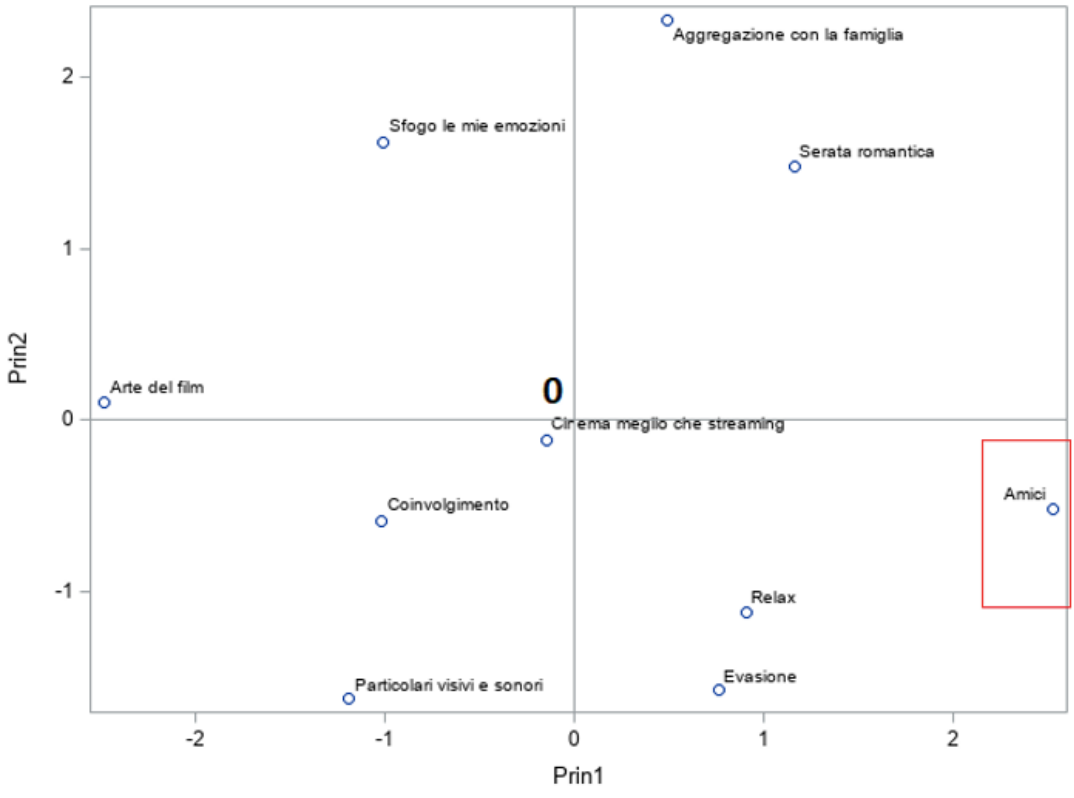
4.5 Similarità tra *cluster*

Osservando il grafico qui presente, che mostra il cerchio delle correlazioni, possiamo notare che il ***cluster 2*** e ***5*** sono molto simili. Tuttavia, per capire sotto quale punto di vista questi due



cluster sono simili sfruttiamo il secondo grafico sottostante, in cui sono posizionate le variabili e le loro rispettive correlazioni. Dalla lettura di tale grafico è possibile affermare che la similitudine tra i ***cluster 5*** e ***2*** è data dalla variabile "il cinema è il luogo in cui vado con gli amici per divertirmi".

Non è possibile sovrapporre i due grafici in quanto l'analisi delle componenti principali non ha la stessa metrica nella misurazione degli individui e nelle distanze delle colonne.



5. Conclusioni

Al giorno d'oggi quasi nessuno non è mai stato al cinema. Quest'ultimo fa parte della vita di tutti ma, abbiamo scoperto, in termini differenti da persona a persona. Sicuramente l'analisi da noi condotta ha evidenziato delle criticità che affliggono il mondo del cinema da tempo ma che, soprattutto negli ultimi due anni di pandemia, si sono notevolmente acutizzate. Sarebbe assurdo sfruttare tale spazio per affibbiare delle colpe, se ce ne sono, a dei motivi e a degli attori ben precisi; quindi, ci limiteremo a sottolineare brevemente quanto emerso dalla nostra analisi, cercando, ove possibile, anche di suggerire delle modeste iniziative volte a cercare di riportare il pubblico in sala.

A nostro avviso, uno dei problemi del cinema può essere identificato nella mancanza di un pubblico stabile che non vede il cinema come un evento speciale ma che lo vive come una realtà da frequentare periodicamente, in base a ciò che è in grado di dare al pubblico: un po' come può essere la chiesa per le famiglie la domenica mattina o lo stadio per i tifosi nel fine settimana. L'analisi ha infatti evidenziato una forte presenza di un pubblico polarizzato, caratterizzato o da individui fortemente appassionati all'arte del cinema e alle emozioni che è in grado di far scaturire, o da individui quasi completamente disinteressati al cinema stesso che si recano in sala solo per rimanere aggiornato sulle tendenze del momento. Andando infatti a leggere le due componenti principali individuate, una si può dire che rappresenta la dimensione dell'occasionalità, l'altra invece va a intercettare, in un certo modo, la dimensione sensoriale, sia da un punto di vista più intimo e interno, sia da un punto di vista più estetico ed esterno.

In virtù di tali osservazioni, nonché dei *cluster* emersi, abbiamo pensato a dei piccoli suggerimenti per cercare di motivare gli individui a frequentare maggiormente il cinema e di migliorare la loro esperienza in sala.

1. **Gli Aristogatti** → si è pensato a sconti per biglietti, *pop-corn's bucket* e bibite in formato famiglia o all'organizzazione di serate romantiche per le coppiette, tramite la proiezione di film *cult* romantici

2. **I Mercenari** → si è pensato all'aumento della fruizione di *merchandising* tematici e alla ri-proiezione di film culto per cercare di cavalcare l'effetto nostalgia

3. **La Dead poets society** → si è pensato al potenziamento delle sale ***ISense***, al miglioramento delle postazioni *vip* e a offerte personalizzate come abbonamenti e sconti

4. **Gli Inside-out** → si è pensato al potenziamento dei gadget a tema film, alla creazione di *QR codes* che rimandano a siti che contengono contenuti speciali – come anteprime, interviste, curiosità ed eventuali rimandi a fumetti e libri da cui derivano i film – o ancora a foto ricordo, scattate con le polaroid, con i cartonati dei protagonisti dei film

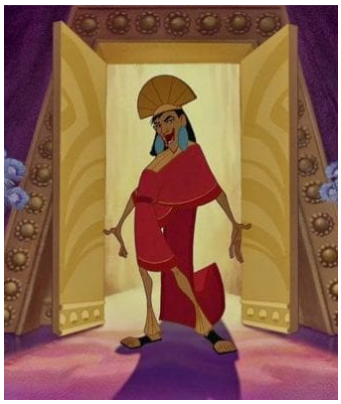
5. **I drughi di Arancia Meccanica** → si è pensato ad offerte di *cross-selling* per ristoranti, *fast-food*, gelaterie e sale giochi posti nelle vicinanze dei cinema per cercare di invogliarli non tanto tramite il film quanto attraverso il post serata

Non vogliamo, tramite questa analisi, avere la presunzione di risolvere i molteplici problemi del cinema ma il nostro deve essere interpretato come un semplice gesto di chi ha legato, e continua a legare, al cinema dei ricordi e che non riuscirebbe proprio a immaginare una società priva di esso.

"***Hasta la vista, baby***"[5]

Sebastiano Costa

Roberta Graziano



Luca Mastrocesare

[5] Arnold Schwarzenegger in "***Terminator 2: Judgment Day***"