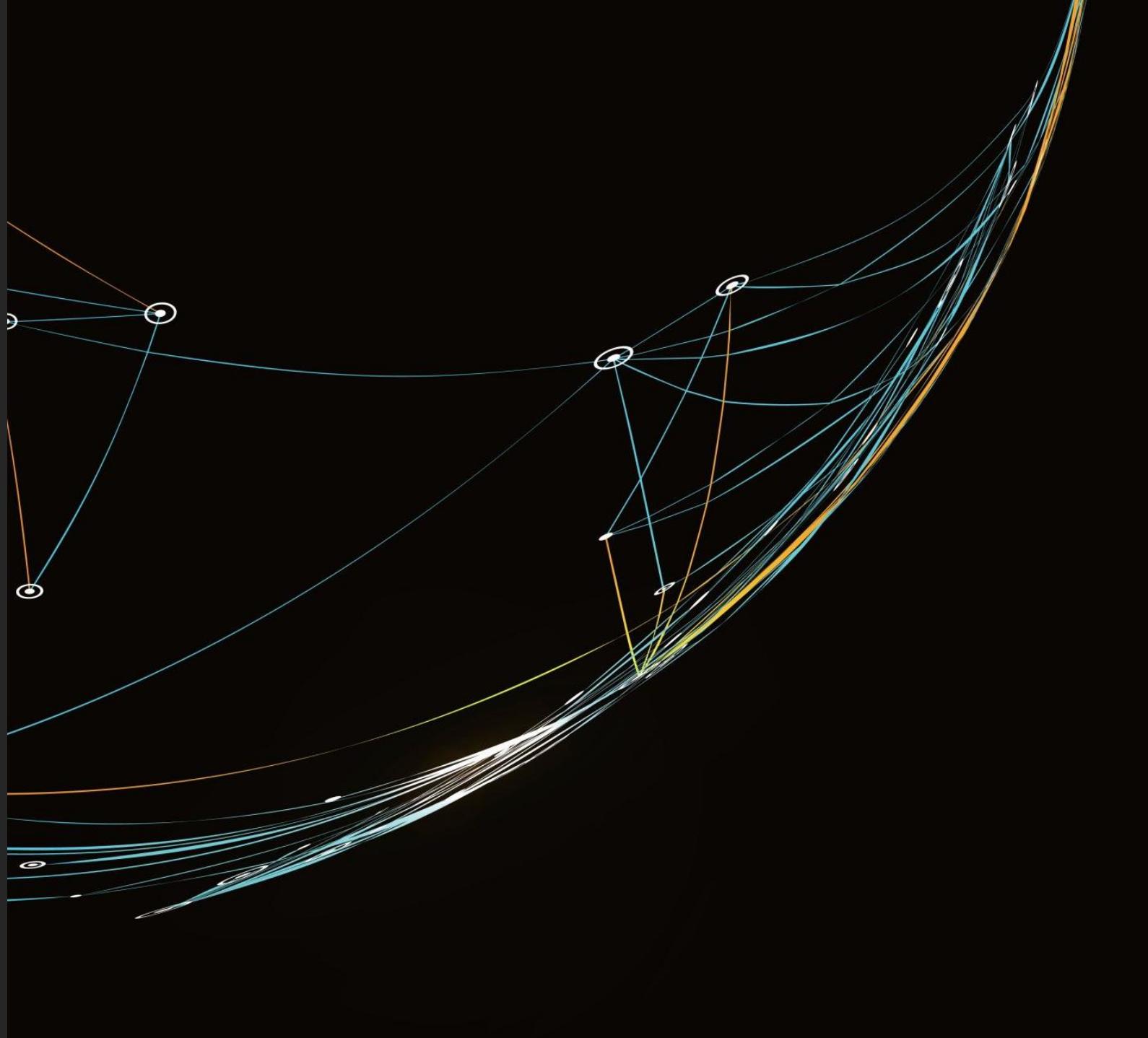


PROGETTO FINALE DI DATA ANALYSIS

DI MATTEO FRANCESCO
BIASIO



start2impact
UNIVERSITY



PREMESSE DELL'ANALISI

- Come è a noi ben noto, il perno su cui i contenuti della nostra editoria ruotano è quello di carattere **scientifico-tecnologico**. Con l'obiettivo di accedere da una **platea più estesa**, rendendo le nostre **notizie** più **accessibili e comprensibili**, questa analisi si è concentrata sullo sviluppo di una **STRATEGIA** atta allo scopo.
- Cercheremo *insieme* un modo per accrescere *l'appeal* delle nostre pubblicazioni e *strizzare l'occhio* anche ad un pubblico *non di nicchia* o *non addetto ai lavori*.
- *Come si stanno comportando i lettori? Da cosa sono più attratti? Riusciremo a stimolare il loro interesse proponendo articoli dal taglio innovativo?*

INTRODUZIONE AI DATI E PRIME MANIPOLAZIONI

- Nel nostro arsenale abbiamo a disposizione una *banca dati* inerente gli *articoli più letti nell'ultimo triennio*.
- Assicurati dell'*integrità dell'informazione a nostra disposizione* – non nullità su tutte le variabili – procediamo alla *conversione dei DATETIME* relative alle date di lettura dell'articolo ed iscrizione dell'utente (`read_date` e `subscription_date`)
- Come si può osservare in
 - «Progetto Finale Data Analysis di Matteo Francesco Biasio.ipynb» - Python
 - «Progetto Finale Data Analysis di Matteo Francesco Biasio» - Tableau Public,
- Link Tableau e Repository GitHub (Python):
 
- iniziamo a fare una prima *indagine esplorativa* circa la distribuzione degli **articoli letti**.

INDAGINE ESPLORATIVA

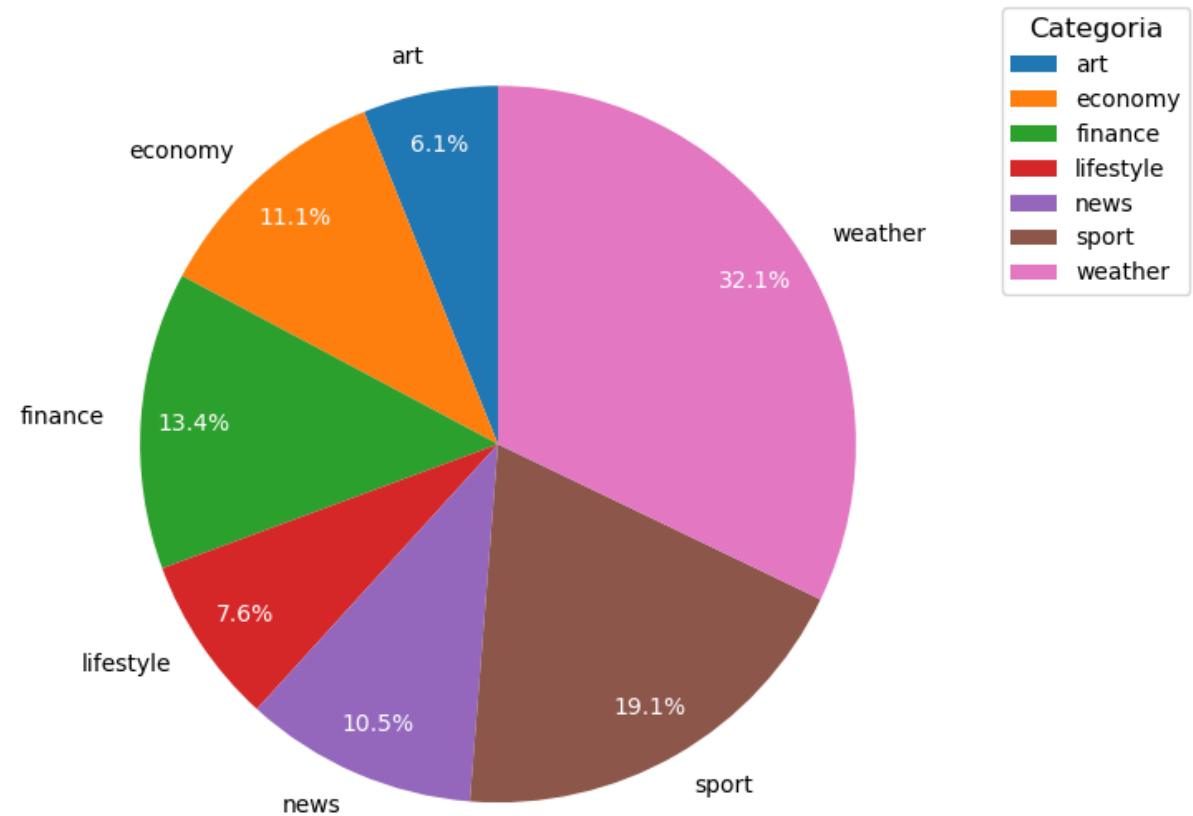
DISTRIBUZIONE DELLE LETTURE – ANALISI AGGREGATA

Come si può evincere dal **grafico a torta**, su TUTTO L'ARCO TEMPORALE a disposizione

- **weather**, con il 32,1% - quasi 1 articolo su 3 -, risulta la **categoria più letta MEDIANTE**
- a seguirla vi sono **sport** e **finance**, con il 19,1% e il 13,4%

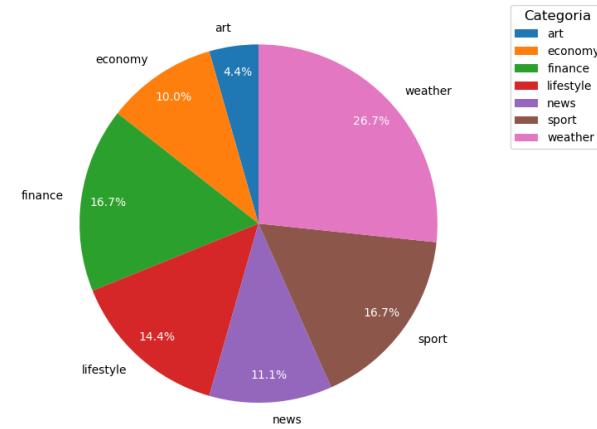
Questa analisi introduttiva viene condotta su un campione complessivo di 999 osservazioni.

Percentuale della distribuzione degli articoli

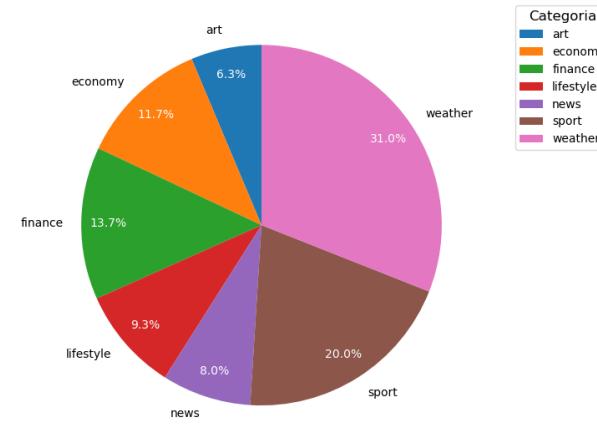


DISTRIBUZIONE DELLE LETTURE – ANALISI ANNUALE

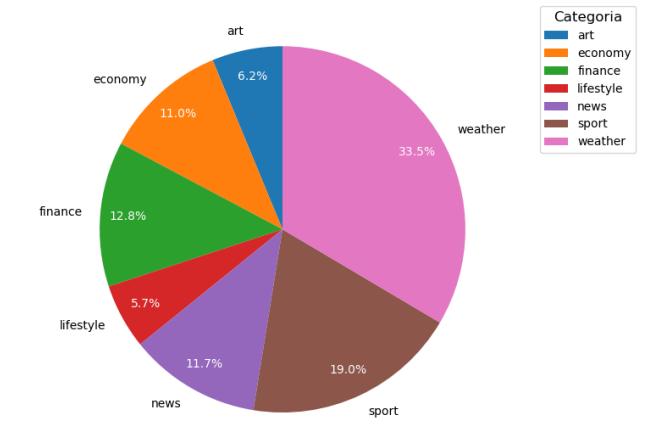
Percentuale della distribuzione degli articoli - 2021



Percentuale della distribuzione degli articoli - 2022



Percentuale della distribuzione degli articoli - 2023



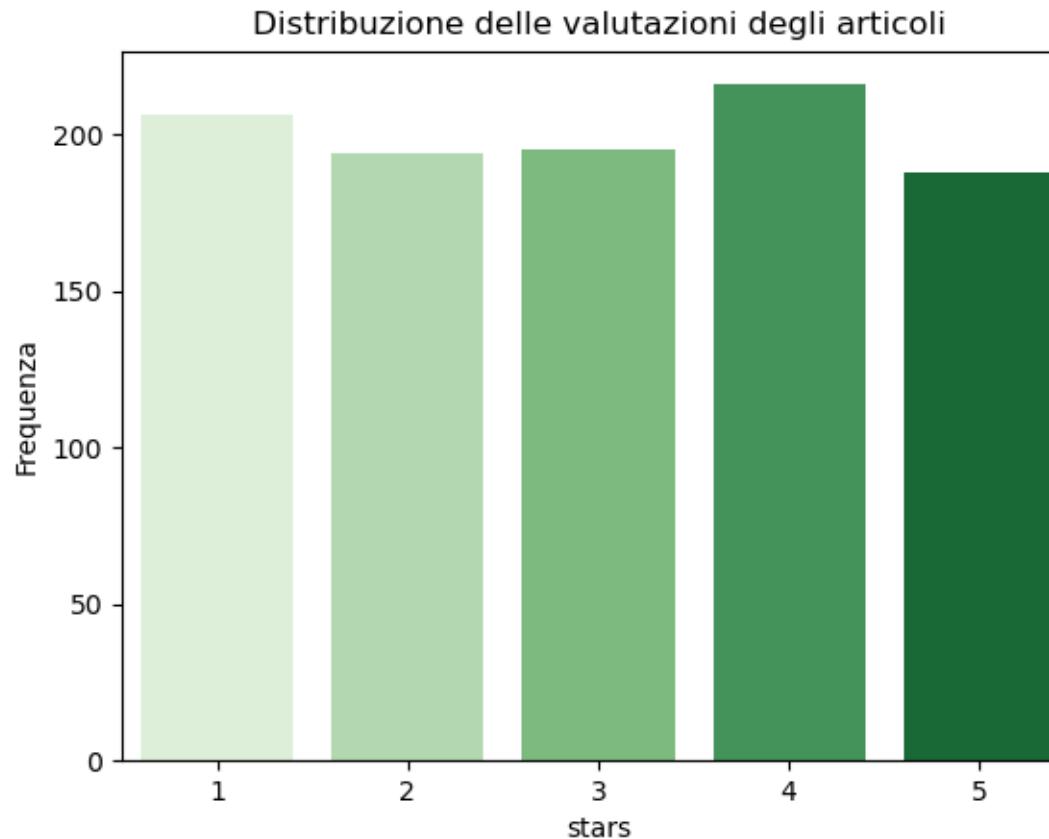
Conducendo un'indagine su **granularità annuale**, si consideri come, nel tempo, vi sia stato un *trend di crescita* del numero di letture: siamo passata da **90** articoli letti nel **2021**, passando per **300** nel **2022** fino ai **609** attestabili nel **2023**. Questo **fenomeno di crescita** ci permette di valutare, con una *discreta robustezza*, anche:

- quanto **weather** continui a mantenere una **posizione di leadership**, anzi accrescendo il suo bacino d'utenza, passando dal 26,7% del 2021 (24 letture sulle 90 annuali) al 31,0% del 2022 (93 su 300) al 33,5% del 2023 (204 su 609)
- Considerando i **pesi dei singoli anni** in termini di letture, le **restanti categorie** mostrerebbero una **distribuzione pressochè simile**, fatto salvo per **finance**, che mostra un **calo di interesse**, a livello annuale, passando dal 16,7% del 2021 al 12,8% al 2023. Stessa sorte tocca a **lifestyle** che passa dal 14,4% del 2021 al 5,7% del 2023. **news** si mostra altalenante, mentre **sport**, nel complesso, acquisisce qualche punto – pur avendo una **flessione** dal 2022 al 2023 – passando comunque dal 16,7% del 2021 al 19,0% del 2023.

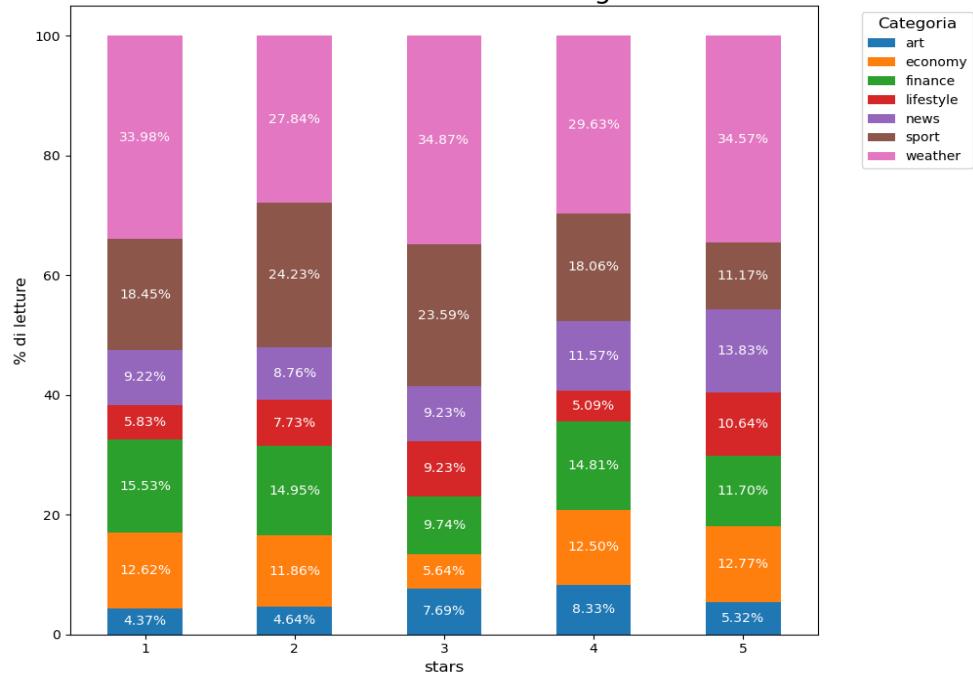
DISTRIBUZIONE DELLE VALUTAZIONI MEDIE – ANALISI AGGREGATA

Volendo poi valutare questa *distribuzione* – su tutto l’arco temporale a nostra disposizione – con un *focus di dettaglio maggiore*, si decide di sondare **come le letture si suddividono** sulla base della “*percezione individuale*” della qualità dell’articolo, attraverso la **VALUTAZIONE MEDIA**:

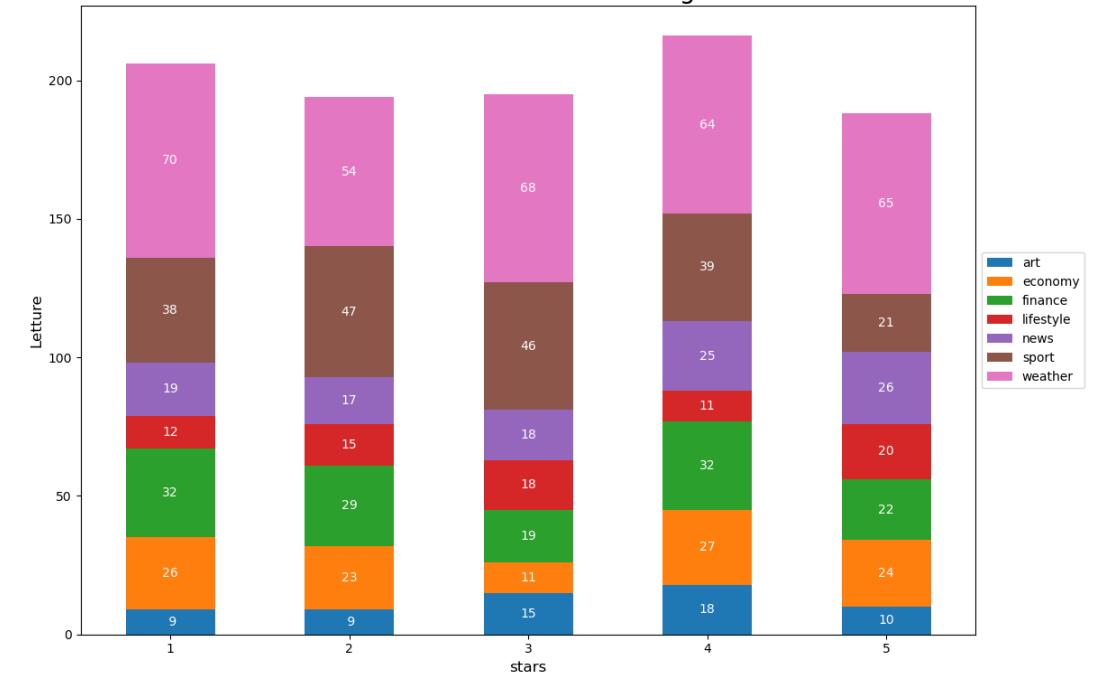
- Ad un primo sguardo, sembrerebbe **non percepirti ALCUNA ANOMALIA** nelle valutazioni; i 999 utenti si *distribuiscono quasi uniformemente* su **TUTTA LA SCALA DI VALORI** inerente la valutazione
- In media, gli utenti si ritiene giudichino gli articoli con *ampia discrezionalità* nella *valutazione*, motivo per il quale *non spicca un certo tipo di voto rispetto ad un altro.*



Percentuale delle valutazioni degli articoli



Distribuzione delle valutazione degli articoli



DISTRIBUZIONE DELLE VALUTAZIONI MEDIE – ANALISI AGGREGATA

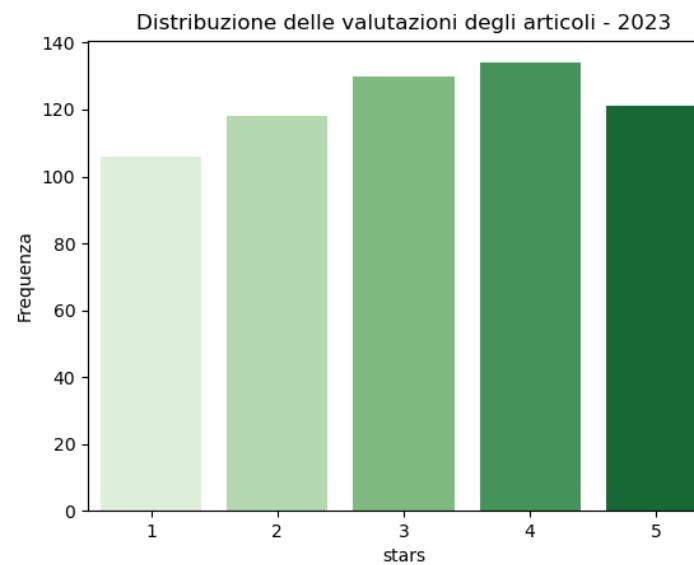
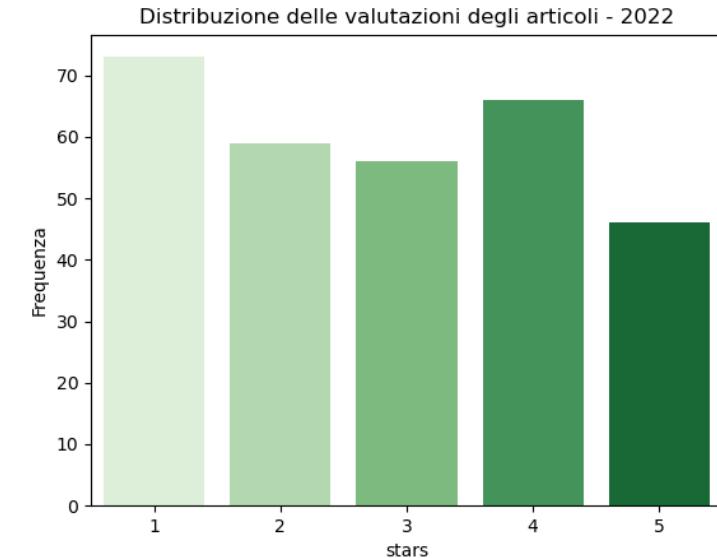
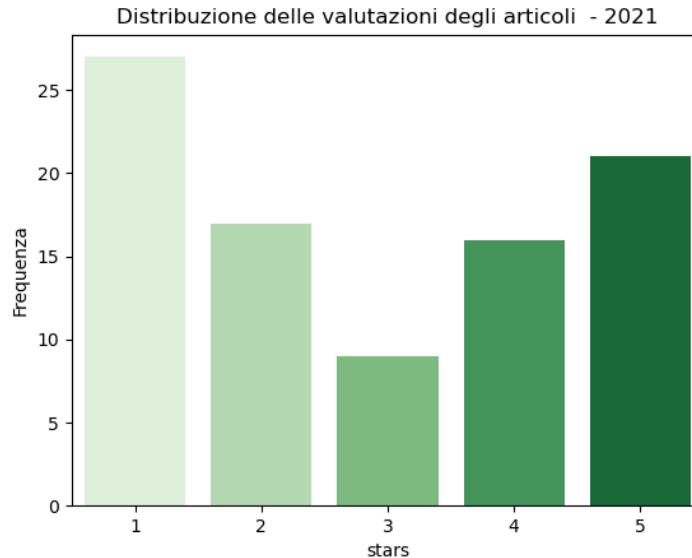
Anche analizzando la distribuzione aggregata per **valutazione stars** e **category**, non sembrano emergere anomalie in termini di “scelte individuali”:

- **Sport** mostra solo una **contrazione** della propria *percentuale* – sul totale degli articoli letti che hanno ottenuto la valutazione massima = 5 – arrivando ad essere pari all’11,17%
- Sorti simili spettano a **lifestyle** e ad **economy**, dove tale “*contrazione*” si manifesta per la categoria di voto pari a 4 e 3 rispettivamente (5,09% per la prima e 5,64% per la seconda)

DISTRIBUZIONE DELLE VALUTAZIONI MEDIE – ANALISI ANNUALE

MA verificando lo stesso fenomeno su **GRANULARITÀ ANNUALE**, viene evidenziata una sospetta anomalia circa le valutazioni **medie**. Differentemente da come ci si poteva aspettare con **una semplice analisi aggregata**, emergono comportamenti quasi agli antipodi:

- il 2022 e il 2023 sembrano essere quasi anni di **assestamento e irrobustimento** della *distribuzione delle valutazioni*
 - il 2022 mostra una sorta di irrobustimento delle **valutazioni intermedie**, prossime alla sufficienza o appena sufficienti (tra 2 e 4)
 - il 2023 “*consacra*” questo irrobustimento, mostrando la tendenza alla formazione di una **distribuzione NORMALE** (c.d. CAMPANA GAUSSIANA), che inizia a concentrare buona parte delle **valutazioni** attorno al **valor medio/sufficienza** (riducendo lentamente la deviazione standard)



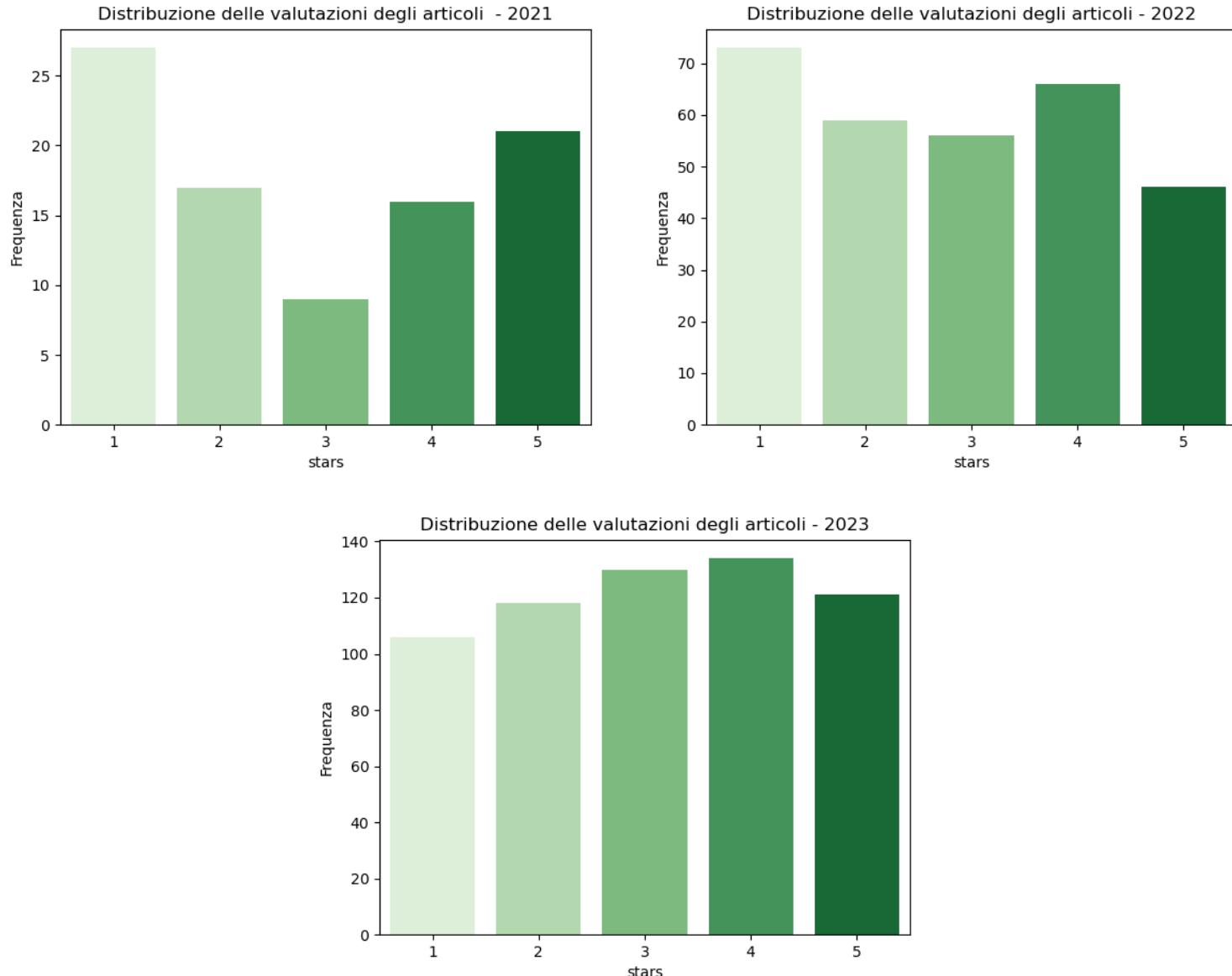
DISTRIBUZIONE DELLE VALUTAZIONI MEDIE – ANALISI ANNUALE

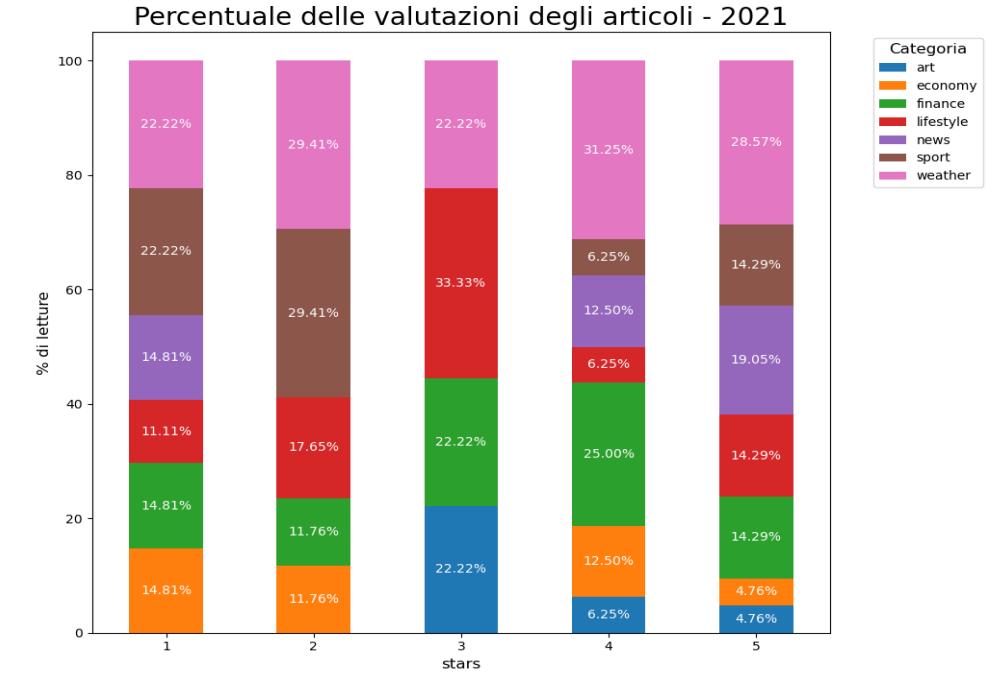
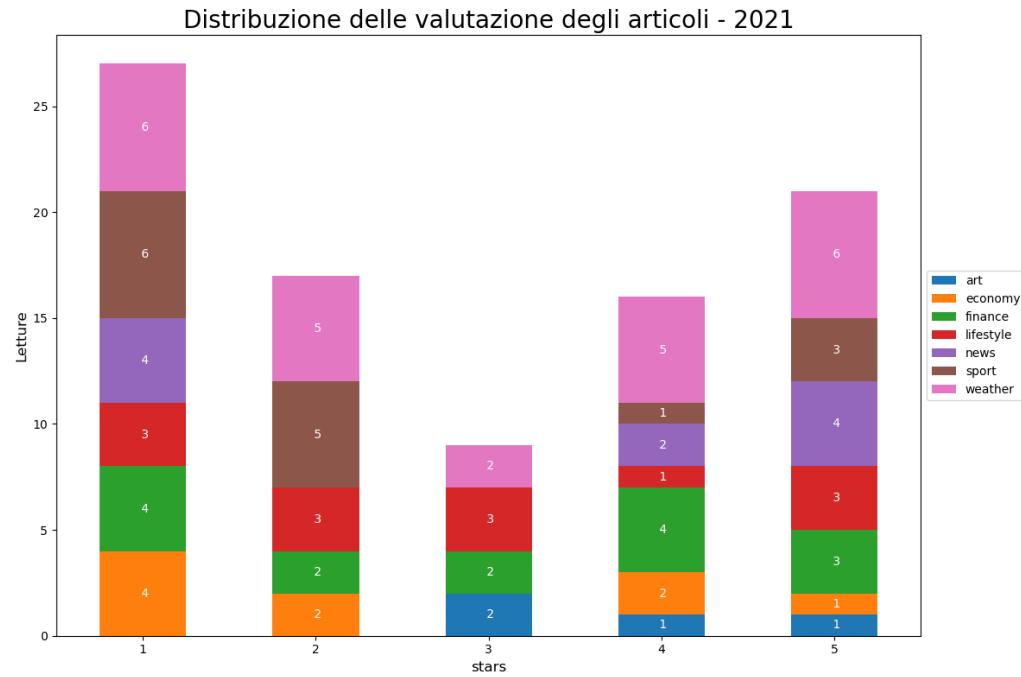
Come è possibile osservare con questi istogrammi, l'ANOMALIA sorge al 2021. La distribuzione mostra una sorta di *comportamento parabolico convesso*:

- dalla categoria di valutazione media 1, la *peggiore*, il numero di letture di articoli *scende*,
- fino al raggiungimento di un *punto di minimo coincidente* con la categoria di valutazione 3,
- per poi *risalire* la china *fino alla categoria 5*

Si presenterebbe, dunque, una DISTRIBUZIONE DELLE VALUTAZIONI DELLE LETTURE POLARIZZATA, per cui vengono espresse principalmente valutazioni medie agli estremi della scala di valori.

Si decide, dunque, di porre l'attenzione della nostra analisi al **solo 2021**.

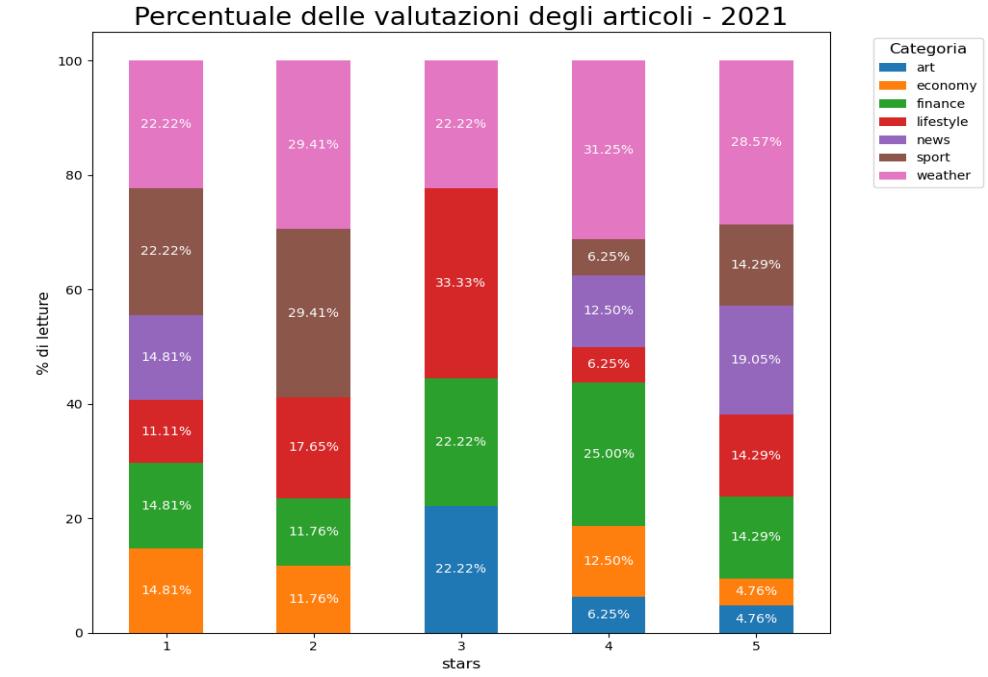
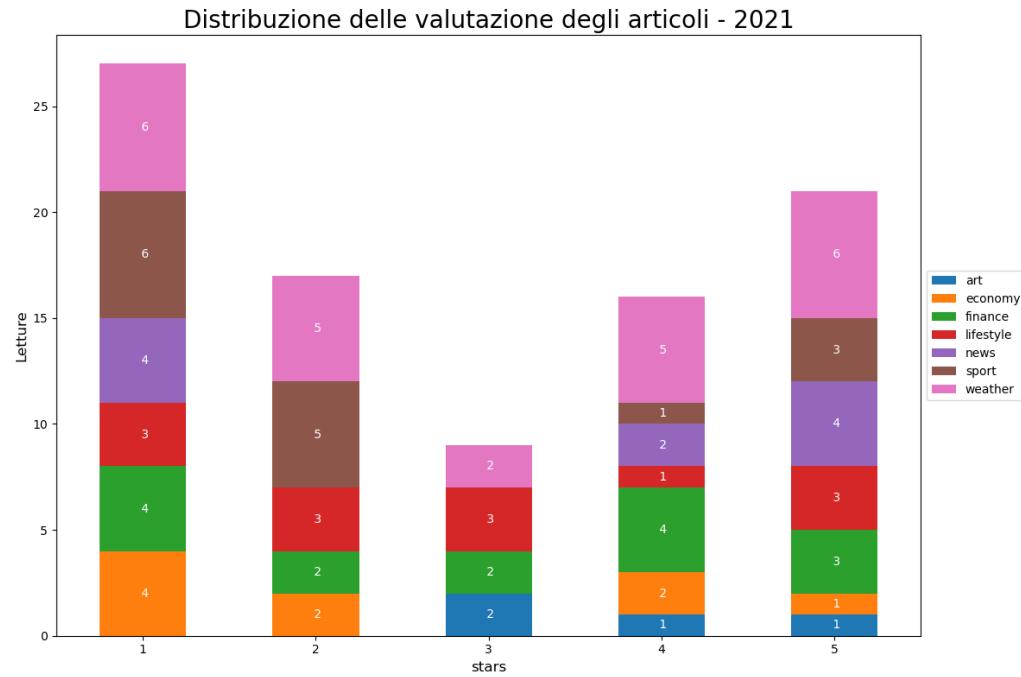




DISTRIBUZIONE DELLE VALUTAZIONI MEDIE: CATEGORIE

Ad una prima indagine sul dataset «anomalo» al 2021, valutando la distribuzione delle valutazioni in termini **assoluti e percentuali**, sembrerebbe che

- per ogni livello di valutazione, una media del 25-27% delle letture, *per ogni stars e complessivamente*, è relativo alla categoria **weather**
- seguono **finance** e **lifestyle**



DISTRIBUZIONE DELLE VALUTAZIONI MEDIE: CATEGORIE

Risulta possibile mostrare come alcune **categorie** mostrino una *maggior POLARIZZAZIONE* in termini di valutazione dando uno sguardo al grafico *percentuale*:

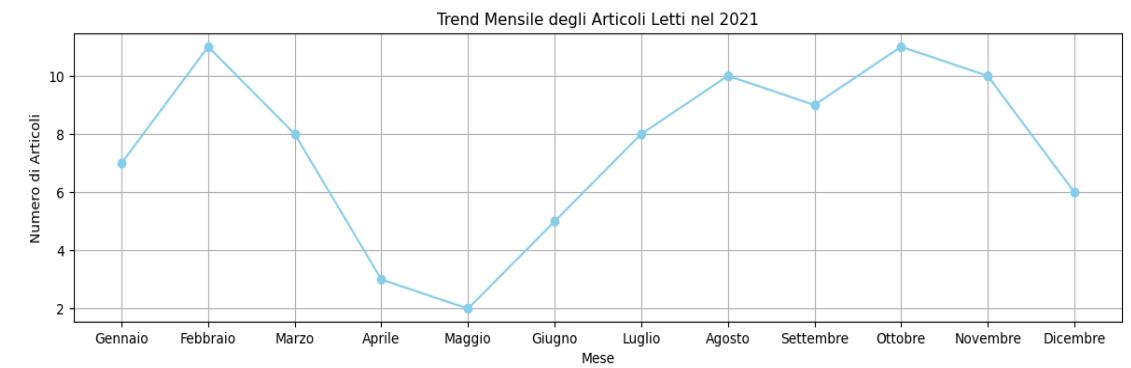
- **art** è mediamente concentrata su **valutazioni positive**, o comunque, sopra la soglia della sufficienza (3-5)
- **sport** ed **economy** non presentano una distribuzione equa, in quanto le valutazioni sono **poste agli estremi (1-2 o 4-5)**: sono però **fortemente sbilanciate**, anche in termini assoluti, verso *valutazioni negative*

Per il resto sono equamente distribuite, fatto salvo per **lifestyle**, per cui, considerando anche il livello complessivo di numerosità per la valutazione 3 (la più "debole"), si presenta la maggior concentrazione di valutazioni per la **categoria** (una sorta di **POLARIZZAZIONE "CONVERGENTE"**)

TREND MENSILE

ANNO 2021:

Come abbiamo potuto osservare, il comportamento «anomalo» si evidenzia nell'anno 2021. Data questa premessa, si decide di approfondire l'analisi in questo anno.



MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE

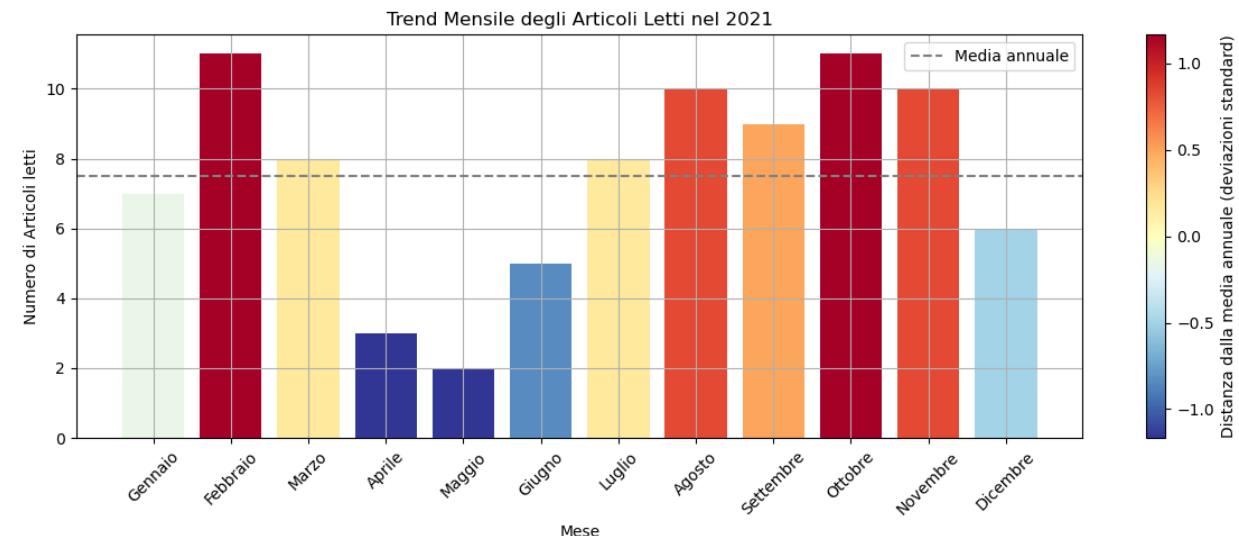
Estratto il mese dal campo `read_date` in `read_month`, monitoriamo ora il trend mensile all'anno 2021.

Da una prima analisi temporale,

- si evince come il secondo trimestre dell'anno mostri una flessione discendente dell'interesse alla lettura degli articoli.
- questo trend viene compensato con una ripresa della DOMANDA DI INFORMAZIONE nei mesi estivi, perdendo di efficacia nel periodo delle vacanze natalizie

Occorre valutar il motivo di questa flessione stagionale

MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE (DEVIAZIONI STANDARD)



A corredo di questa prima indagine, ora mostriamo come i *mesi del secondo trimestre* siano ampiamente sotto la media annuale, *pur non incidendo in maniera cospicua sul valore medio*

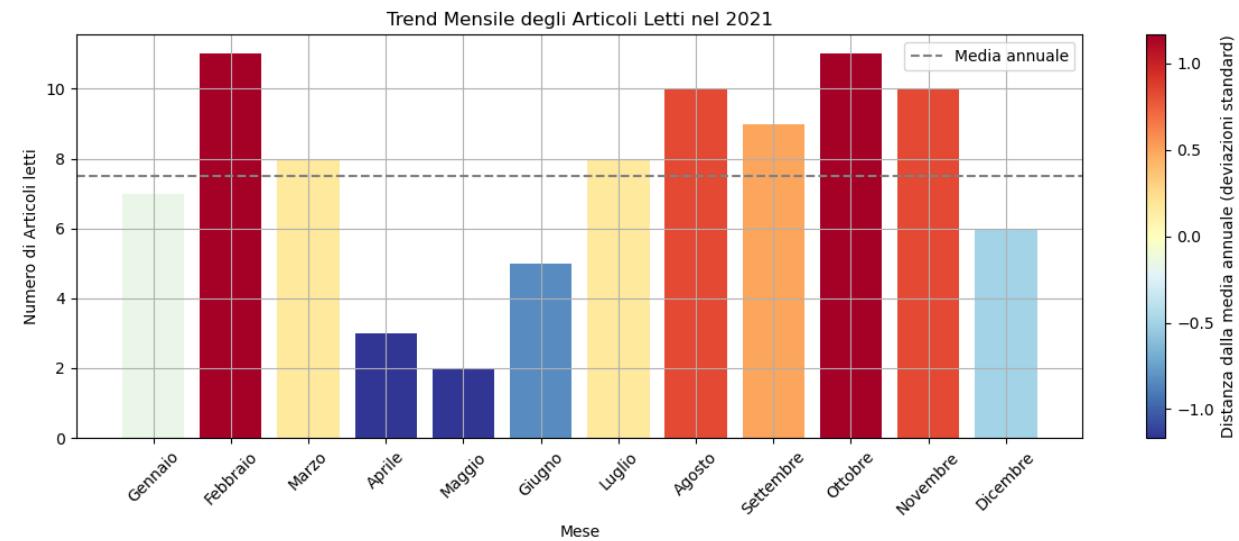
Utilizzando un *diagramma a barre* - con *mappatura cromatica graduata* - possiamo mostrare il PESO DEL PERIODO DI FLESSIONE in relazione alla DISTANZA DALLA MEDIA ANNUALE come

$$N \text{ art. mensili letti (norm.)} = \frac{\sum \text{Letture mensili} - \text{media mensile annuale}}{\text{deviazione standard}}$$

Sfruttando tale misura come *indice per la graduazione* del plot

- all'aumentare di tale *differenza*, si *intensifica la colorazione*;

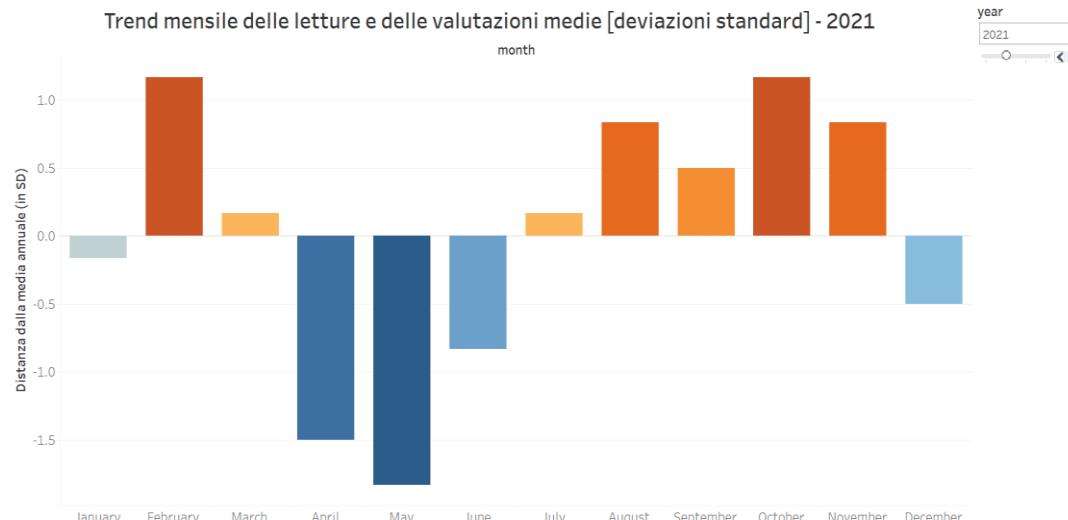
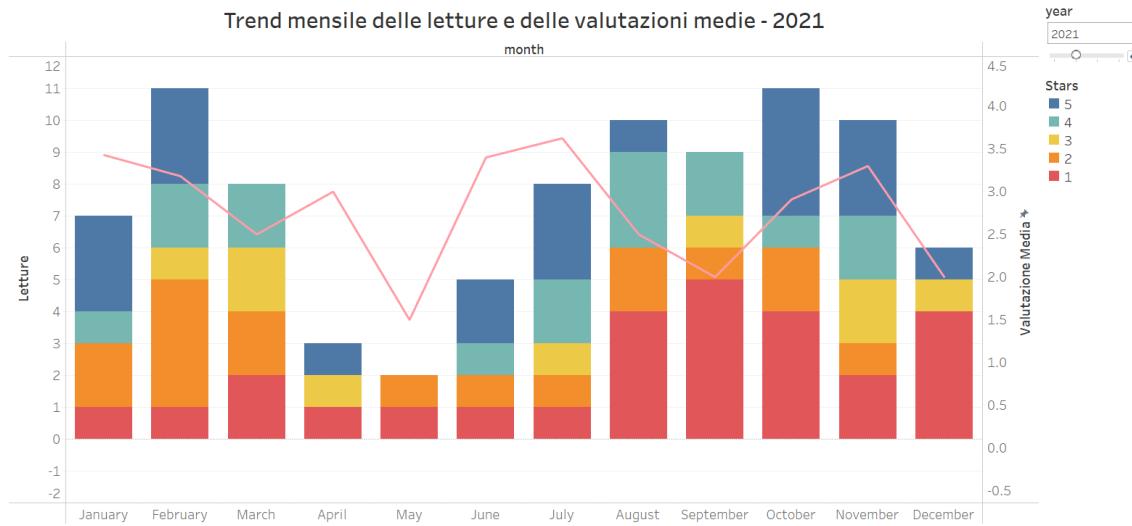
MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE (DEVIAZIONI STANDARD)



- al collasso verso il $\lim_{x \rightarrow t^\pm} (Letture\ mensili(x) - media(t)) = 0$
- la scala tende ad un **coloratura sbiadita di giallo/bianco**
 - se tale *misura* risulta $< \text{media annuale}$ la colorazione assume la scala di **colori freddi (BLU)**
 - se tale *misura* risulta $> \text{media annuale}$ la colorazione assume la scala di **colori caldi (ROSSO)**

N.B. A prescindere dal valore k della *stddev*, essendo considerabile come una **costante**, occorre valutare solo il comportamento del **numeratore** in termini di limite:

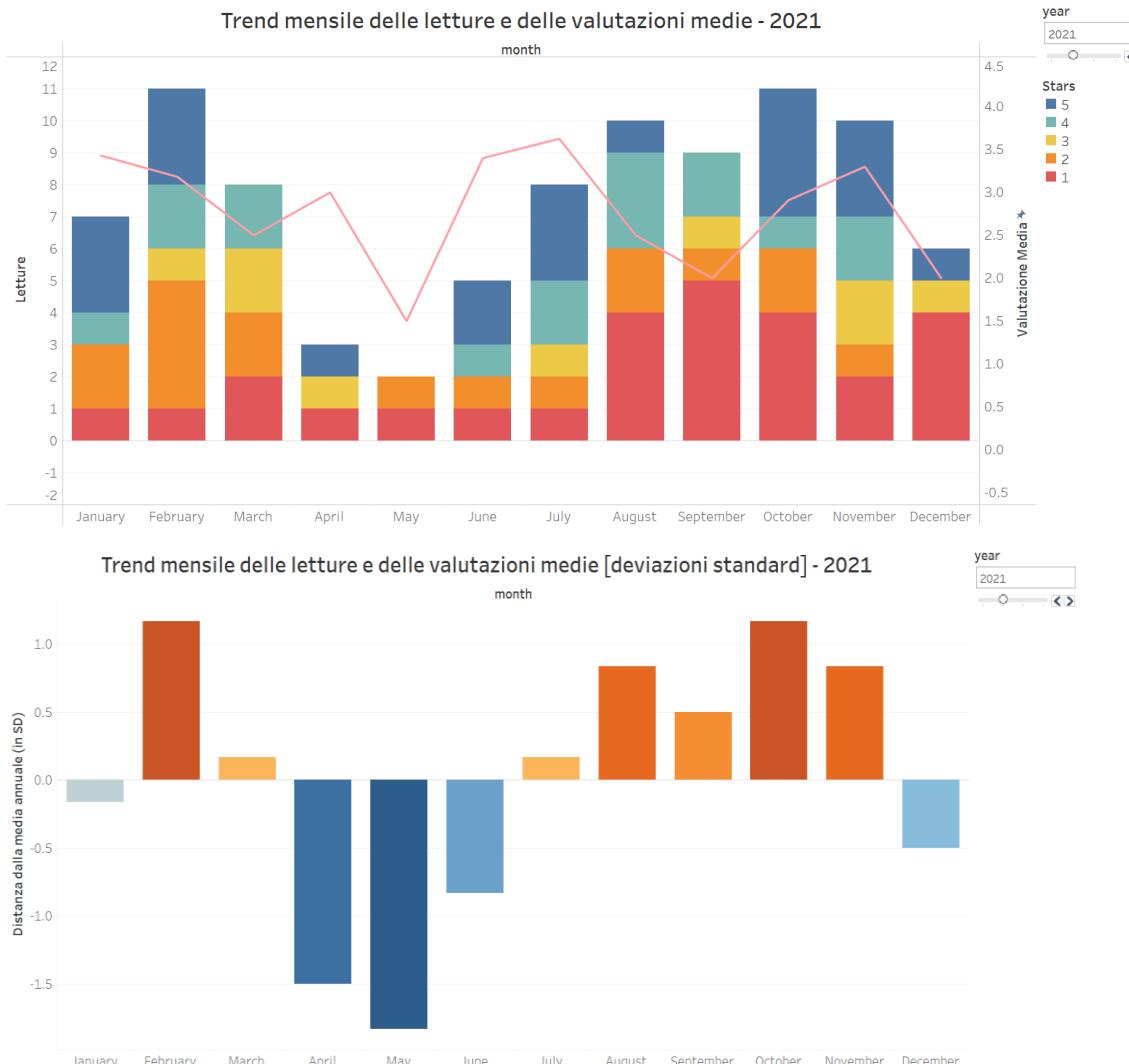
$$\left(\lim_{x \rightarrow t^\pm} \frac{(x-t) \sim 0}{k} \rightarrow 0 \right)$$



MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE (DEVIAZIONI STANDARD)

Quello che ne emergerebbe potrebbe essere un **fenomeno ciclico** della serie storica ascrivibile alla **STAGIONALITÀ**:

- si presenta un *tasso di lettura* ampiamente **sopra la media annuale** a
 - **febbraio** e a cavallo tra la fine del periodo estivo e quello autunnale
 - si mostra un *tasso di lettura* di articoli **ben sotto la media annuale** tra
 - **aprile** e **giugno**, in ripresa nei mesi successivi, evidenziando un tasso di **poco sopra la soglia media a luglio** - così come a **marzo**, che mostra un cambio di trend **repentino in descrescita** -.
 - il periodo natalizio **dicembre - gennaio** mostra un periodo di **stanca** inconsueto: il cambio di trend, rispetto a novembre, non è facilmente *predictable*.
 - Tuttavia, sarebbe interessante analizzare se questo comportamento **si ripeta anche negli anni adiacenti**, al fine di accettare la natura della serie storica
- N.B. su Tableau le normalizzazioni sono ricreate usando Level of Detail Expressions - LOD



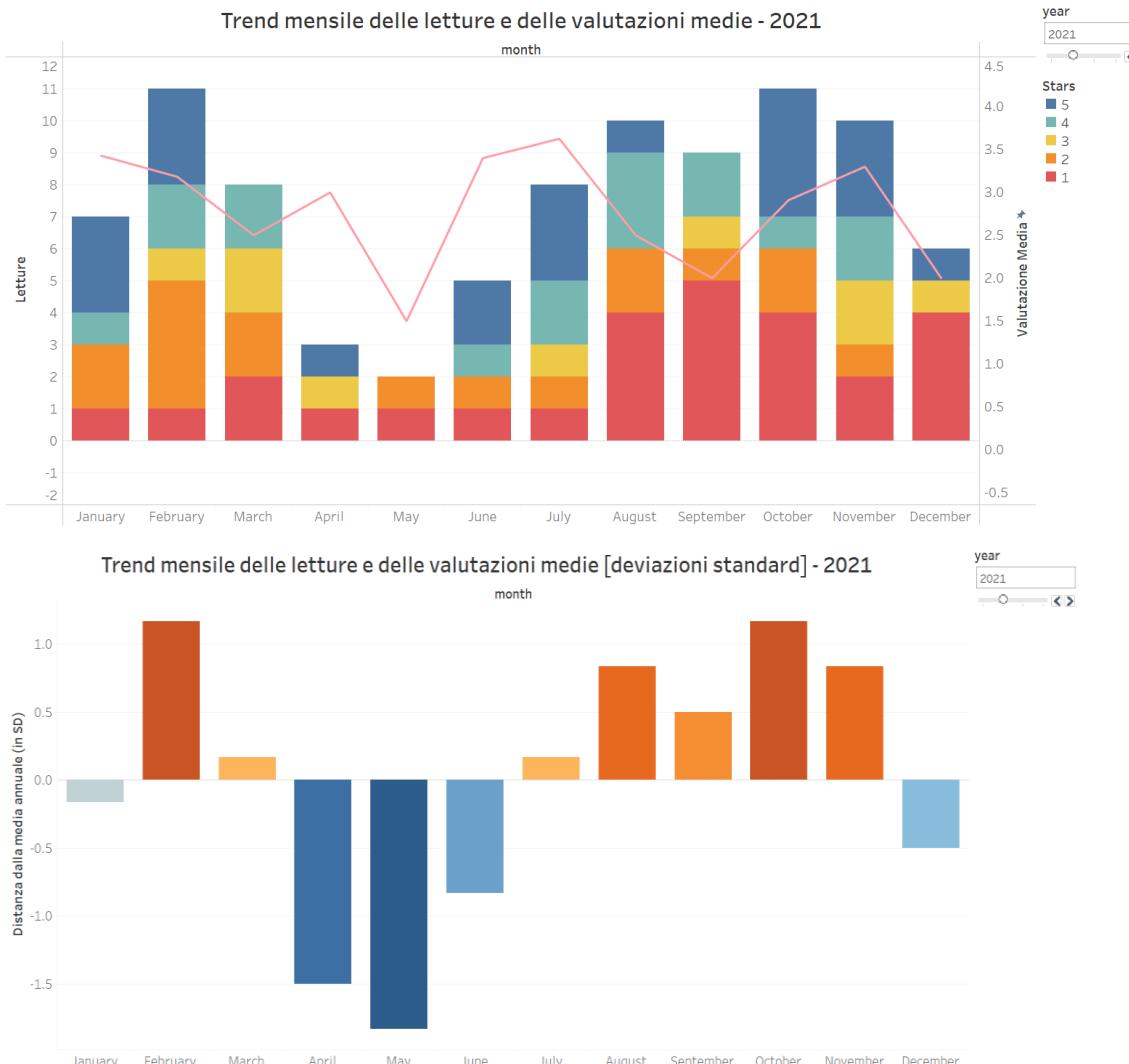
MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE (DEVIAZIONI STANDARD)

Da notare, inoltre, come la **valutazione media mensile** mostri una **convergenza** con il **numero di letture mensili**: quale potrebbe essere la motivazione dietro questo fenomeno?

Questo comportamento lo si evidenzia valutando il grafico relativo al **numero di letture mensili** qui a fianco e nella slide precedente: le barre sono, in questo caso, colorate in modo tale da osservare la **distribuzione delle valutazioni tra gli articoli letti nel mese**.

Valutando in parallelo letture, **valutazione media mensile** e la **distribuzione mensile delle valutazioni**, si nota come

- nella **seconda parte dell'anno**, che tendono ad essere mediamente **insufficienti** – dopo luglio – riflettono una tendenza, degli utenti, a **fornire valutazioni di molto sotto la sufficienza** (buona parte sono pari a 1). Ne è un'eccezione **novembre**, dove la distribuzione delle valutazioni permette, mediamente, di essere sopra la sufficienza, per quel mese.



MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE (DEVIAZIONI STANDARD)

Ne consegue, invece, un **comportamento opposto** nei mesi estivi (giugno-luglio), dove la distribuzione – **ESTREMAMENTE POLARIZZATA** nel caso di giugno – è **sbilanciata** verso valori positivi, il che conferisce alla media una **valutazione più che sufficiente**.

Questa **polarizzazione** mostra caratteri ancora più accentuati nel mese precedente, **maggio**, dove le valutazioni sono **TUTTE CONCENTRATE SU VALORI GRAVEMENTE INSUFFICIENTI**, condizione che motiva questa «*interruzione della serie*» inerente la crescita della valutazione media che il trend mostrava da dopo il mese di **marzo**

Aprile, infatti, mostra una **valutazione media sufficiente**, e la cosa è spiegata dal fatto che gli utenti conferiscono le 3 valutazioni disponibili agli **estremi della scala di valori (1,5) e sul valore della mediana (3)**

Infine, i mesi invernali **gennaio-febbraio**, mostrano ancora una distribuzione **semi-polarizzata** agli estremi positivi e negativi tale per cui, in media, le valutazioni risultano di poco sopra la soglia della sufficienza.

Marzo mostra poi come buona parte delle valutazioni siano **appena sufficienti o insufficienti**, il che spiega la flessione di questo mese.

TREND MENSILE

ANALISI
CATEGORICA –
PRIMA PARTE

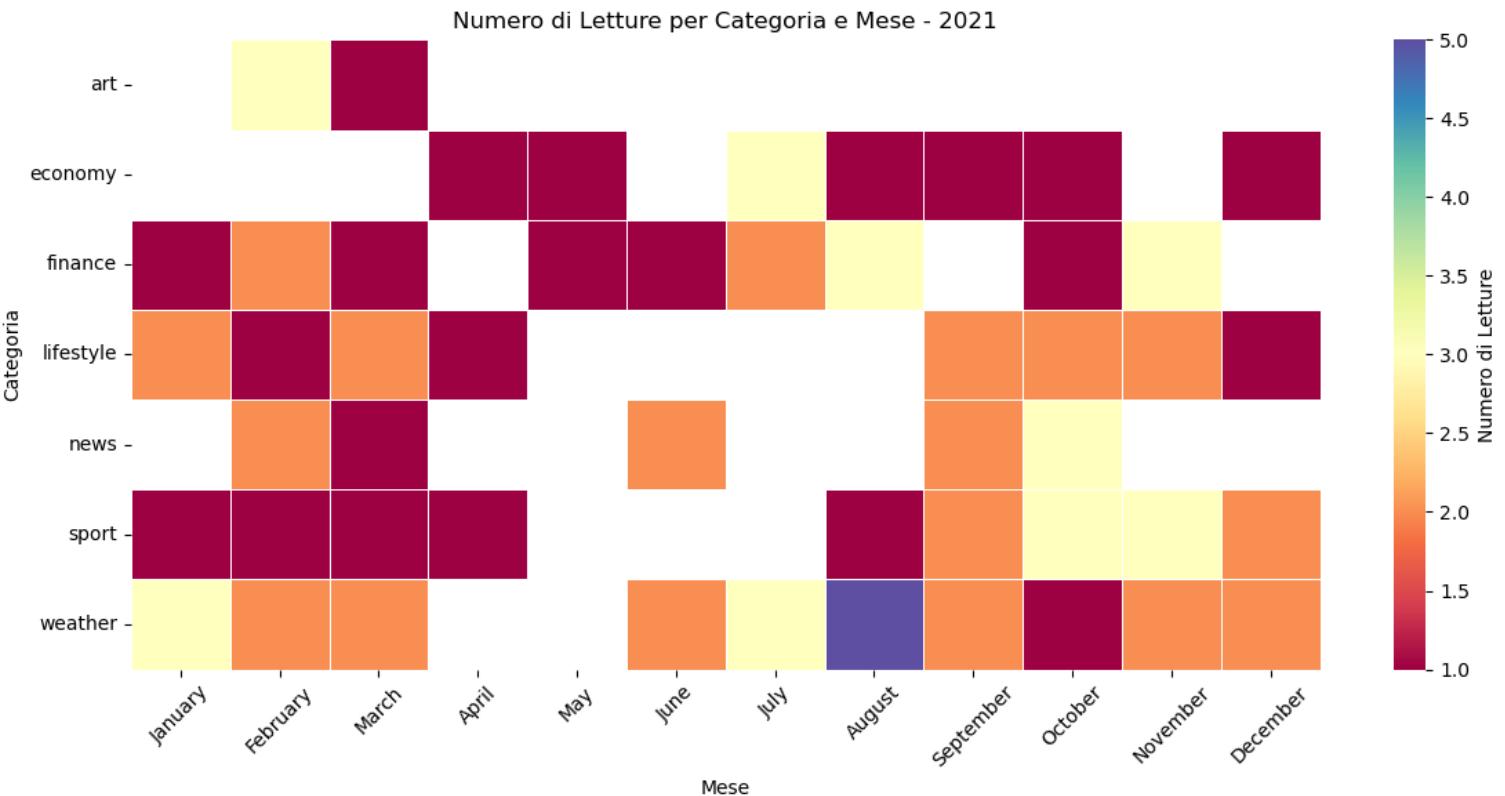
MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE, PER CATEGORIE

Sulla linea dei presupposti del punto precedente - su cui si basano le nostre fondamenta di analisi - ora possiamo porre la lente di ingrandimento sulla **distribuzione temporale delle letture**, in relazione alle **categorie** o **topic** degli articoli

Costruiamo una sorta di **Pivot** delle letture per **month** e **category** e ne leggiamo i potenziali *insights* attraverso una **Heat Map**

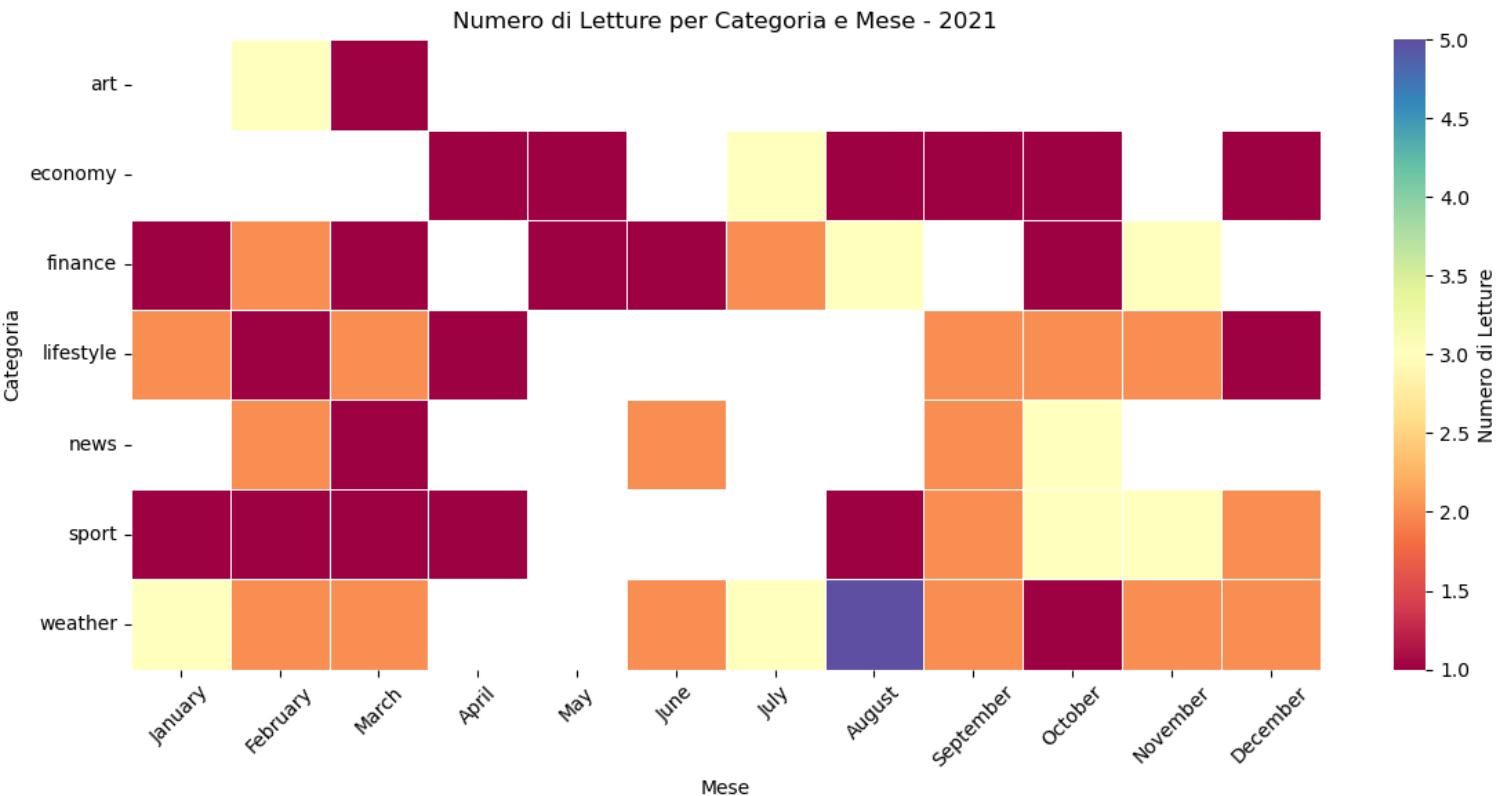
Com'era già stato sottolineato nell'analisi della **distribuzione delle letture per category e stars**,

- la categoria **weather** si mostra come la **tematica di maggior interesse** *su quasi tutto l'arco temporale*, evidenziando, tuttavia, una **flessione delle letture** nel periodo di secca di aprile - maggio e in ottobre.



MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE, PER CATEGORIE

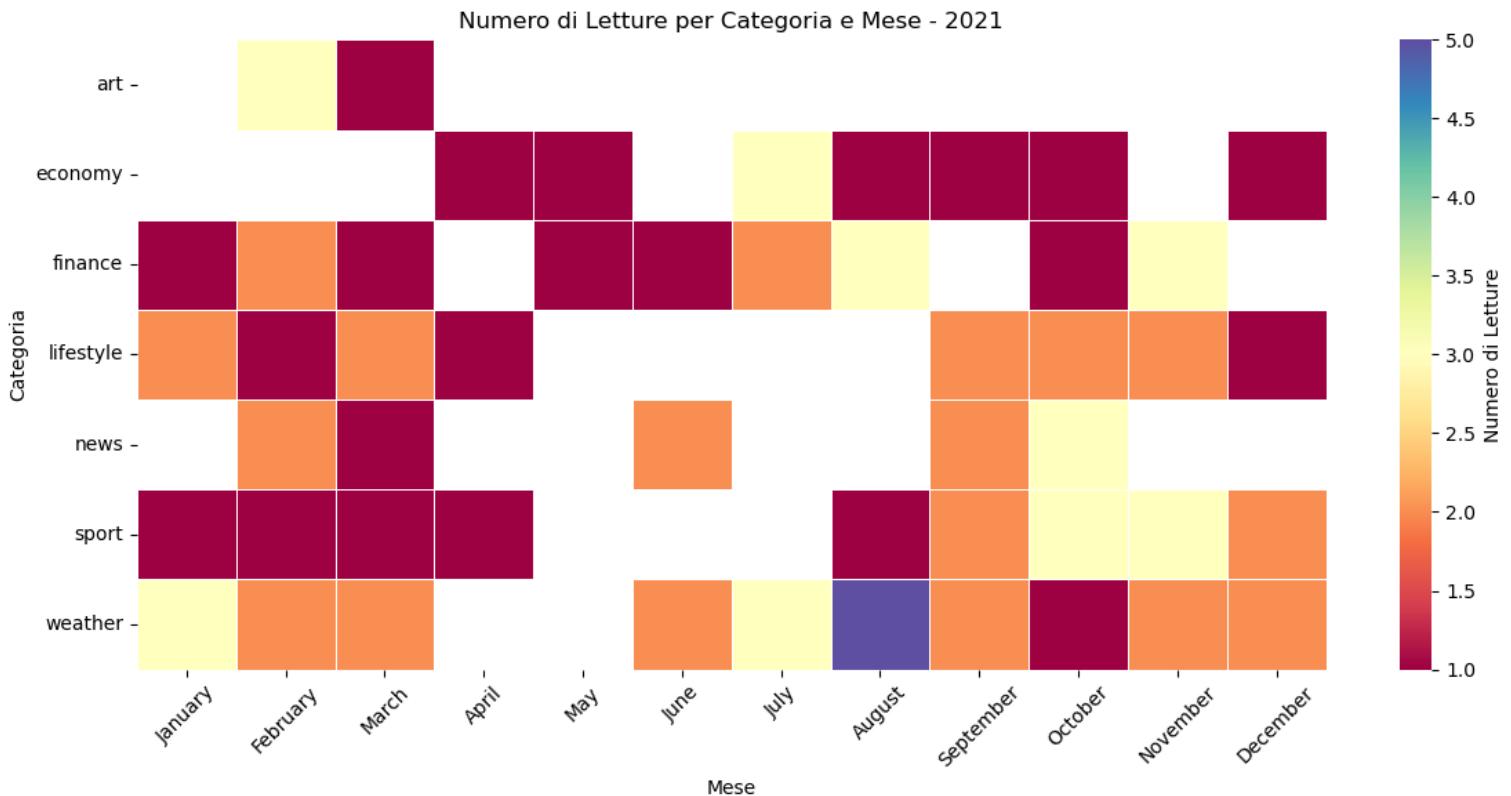
- Gennaio e Giugno sono periodi di discreto interesse, mentre Agosto rappresenta il periodo di **picco** per la categoria
- **economy** e **finance** si mostrano come categorie che **compensano**, seppur in modo marginale, il **periodo di secca primaverile**, suscitando grande interesse entrambe nel periodo estivo - **finance** anche a novembre
- **lifestyle**, **news** e **sport** si dichiarano come categorie di interesse nel **periodo autunnale** - **news** e **lifestyle** anche in **inverno**, in modo alternato



MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE, PER CATEGORIE

- nota di colore è rappresentata da **news**, che mostra un comportamento anomalo in giugno rispetto al trend di periodo
- **art** si conferma come la categoria di *minor interesse*: l'unico periodo in cui sono stati letti articoli al riguardo è quello di **febbraio - marzo**

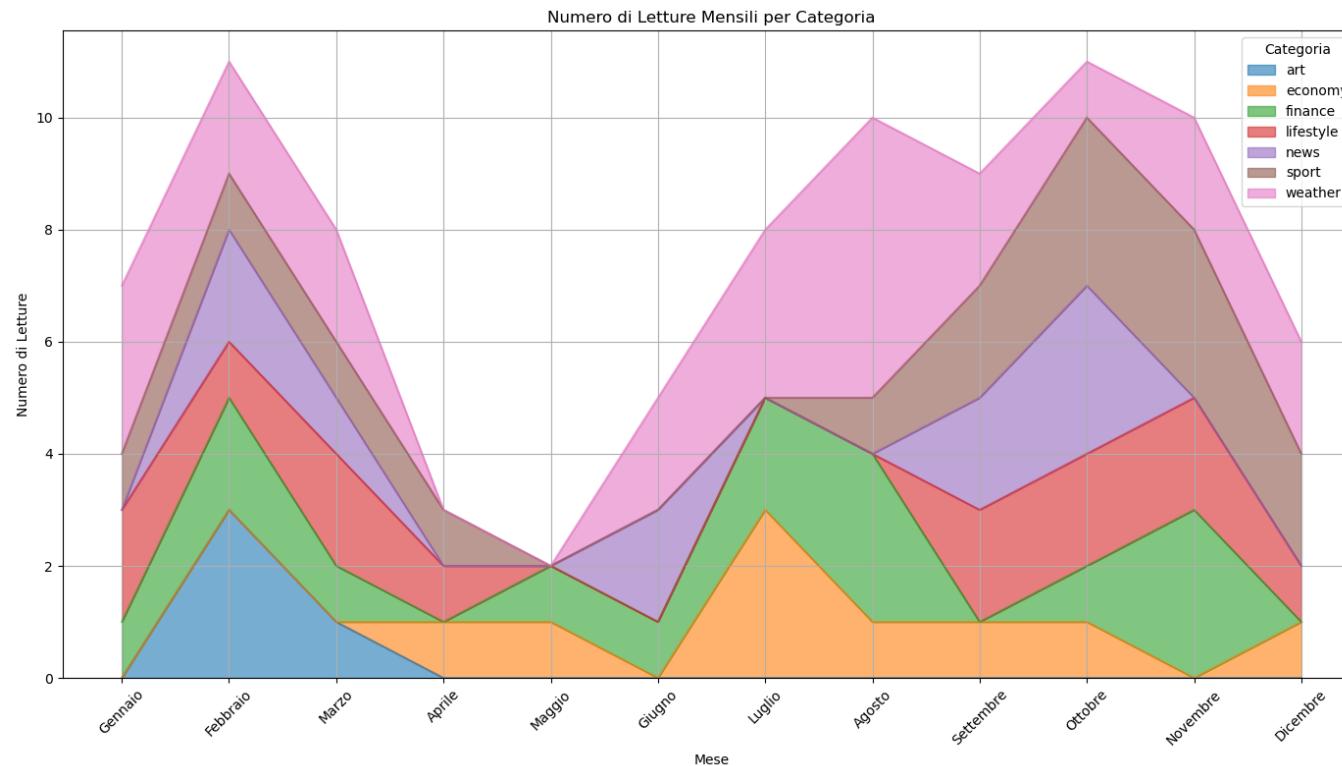
Per mostrare ulteriormente la dimensione di questo fenomeno e gli impatti o contributi delle varie categorie, mostriamo il tutto attraverso un **diagramma ad area CATEGORICO**, mostrando i *trend temporali* e i contributi delle categorie nei vari mesi



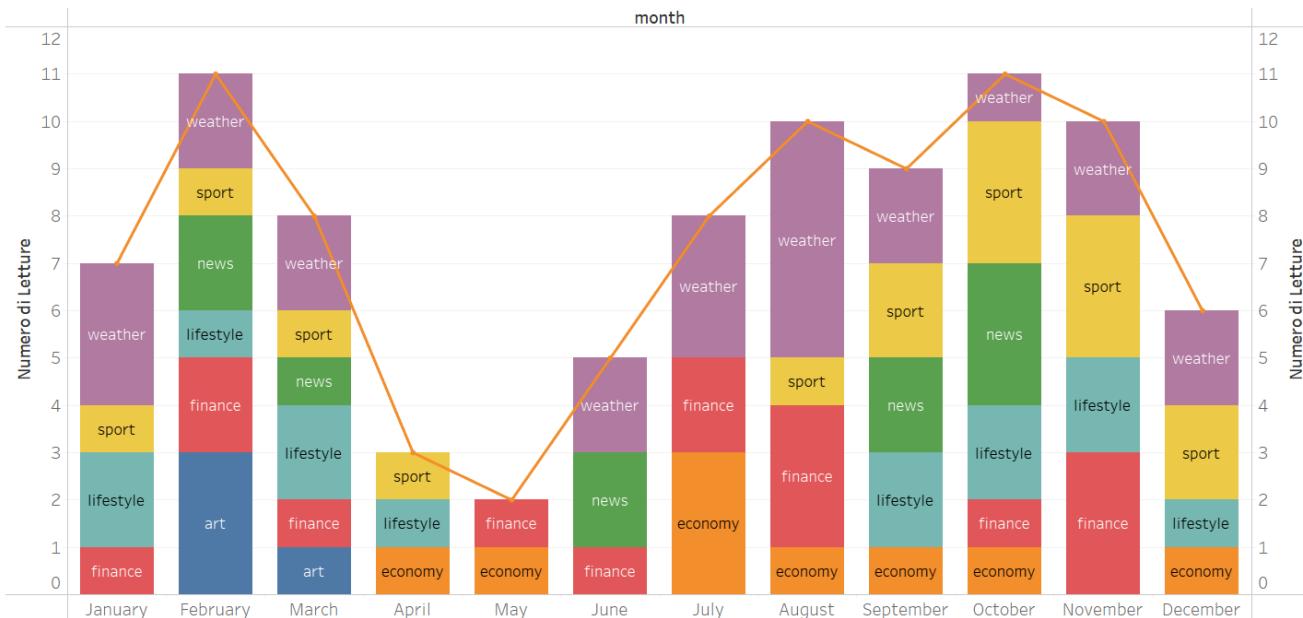
MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE, PER CATEGORIE

Come si è potuto parzialmente evidenziare già con la Heat Map, **weather** si mostra come la **categoria con più visibilità**, salvo nella flessione primaverile - ottobrina

- il contributo nella fase primaverile veste i panni di **sport, economy, finance, lifestyle**
- dopo il picco invernale di **art**, questa smette di essere di interesse
- **lifestyle** si mostra negli occhi dei lettori nel periodo di letargo: si potrebbe associare al tipico fenomeno della vita sana durante l'arco temporale autunnale - invernale - primaverile, sinonimo che gli utenti amano seguire ed informarsi su hint del vivere bene, per arrivare alla stagione estiva nel pieno della forma



Numero di Letture per CATEGORIA e Mese - 2021

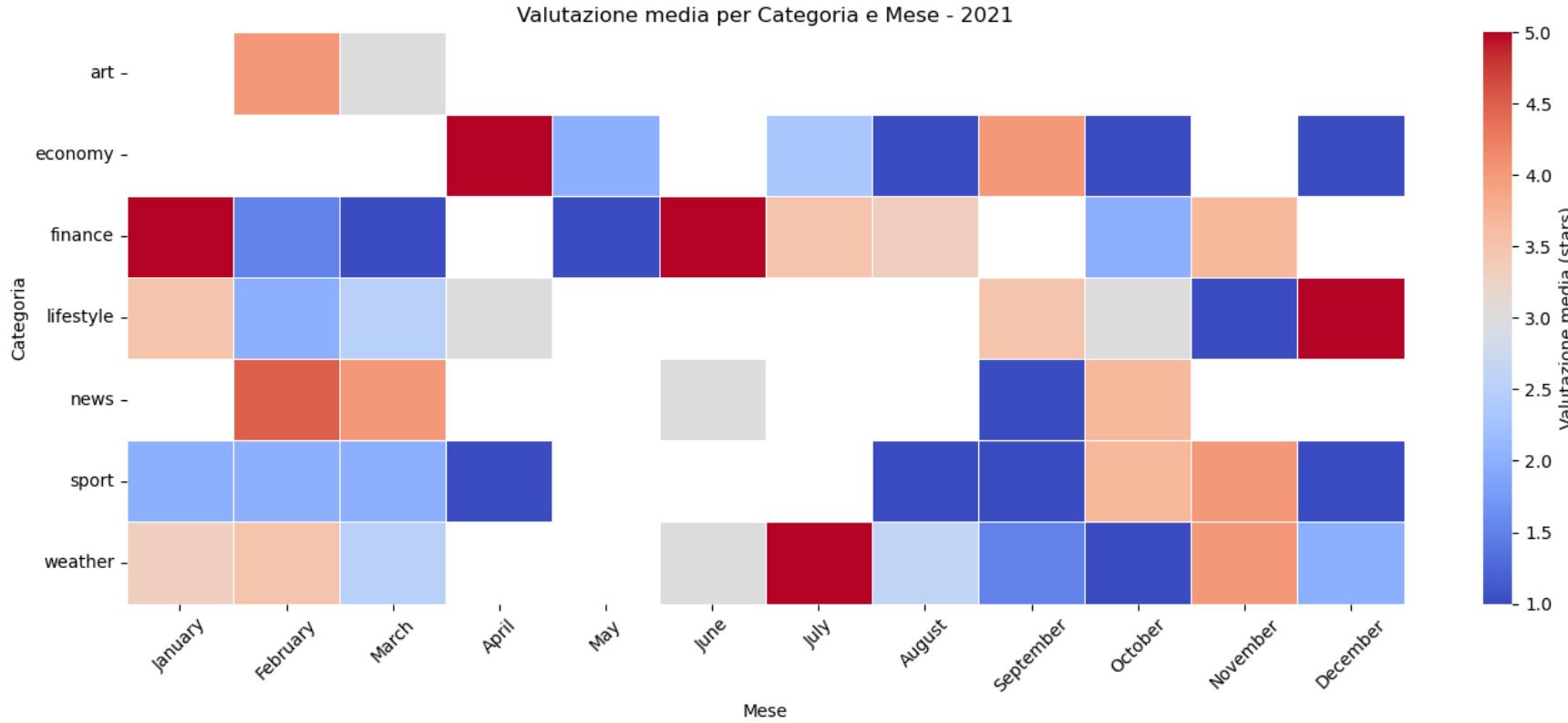


MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE, PER CATEGORIE- STACKED BAR

- **sport** mostra un trend simile, suscitando interesse nel periodo tra luglio-dicembre e gennaio-maggio: coincide con i termini tradizionali della maggior parte delle attività sportive (fatto salvo per meeting continentali ed intercontinentali o sport motoristici), sia in termini di **pratica agonistica**, sia in termini di **spettatori**
- **finance** mostra un **interesse medio** su tutto l'arco annuale, fatto salvo per aprile, settembre, dicembre
- **news** mostra un behaviour anomalo e discontinuo: i periodi di interesse sono **masimo bimestrali**, intervallati da un bimestre di **stanca**

TREND MENSILE

ANALISI
CATEGORICA –
SECONDA PARTE

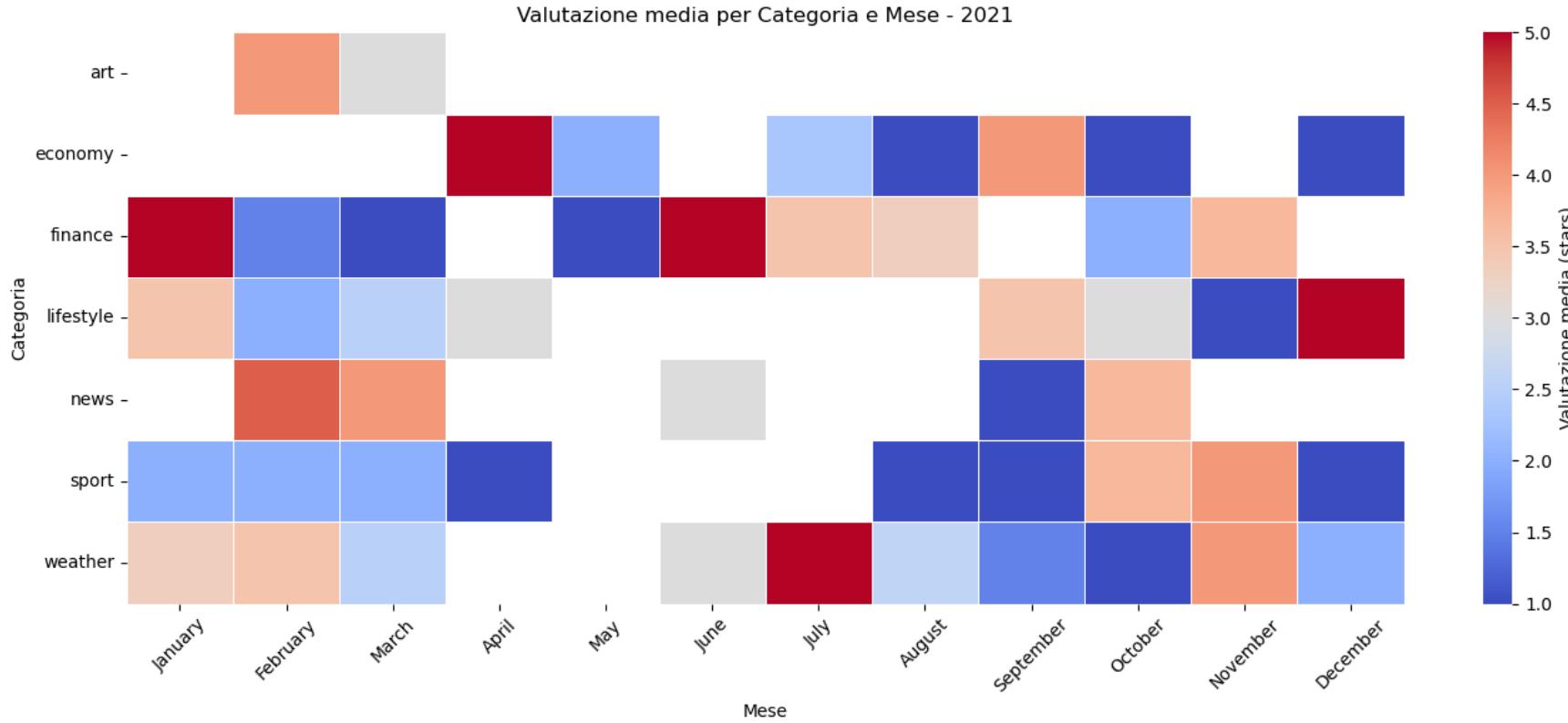


MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE, PER CATEGORIA E VALUTAZIONE

Seguendo la procedura per mappare i trend di interesse monitorati al punto precedente, ora valutiamo qual è stata la valutazione media stars mensile per category

Valutando in parallelo la **Heat Map** con quanto precedentemente analizzato,

- sono **SPORADICHE**, quasi ascrivibili ad *outliers*, le mensilità in cui la **valutazione media stars** supera la soglia della sufficienza;
 - solo **weather**, **economy**, **finance** e **lifestyle** riescono ad ottenere un **CALDO ACCOGLIMENTO** con valutazioni medie ampiamente sopra la sufficienza - per quei mesi in cui vengono letti articoli inerenti tali argomenti -;
 - questo fenomeno si manifesta **solo nel breve termine**, per singoli mesi (es. **weather** a luglio o **economy** ad aprile); non si riesce a mantenere una valutazione tale da giustificare un interesse continuato nel tempo.



MONITORAGGIO TREND MENSILE: NUMERO DI LETTURE, PER CATEGORIA E VALUTAZIONE

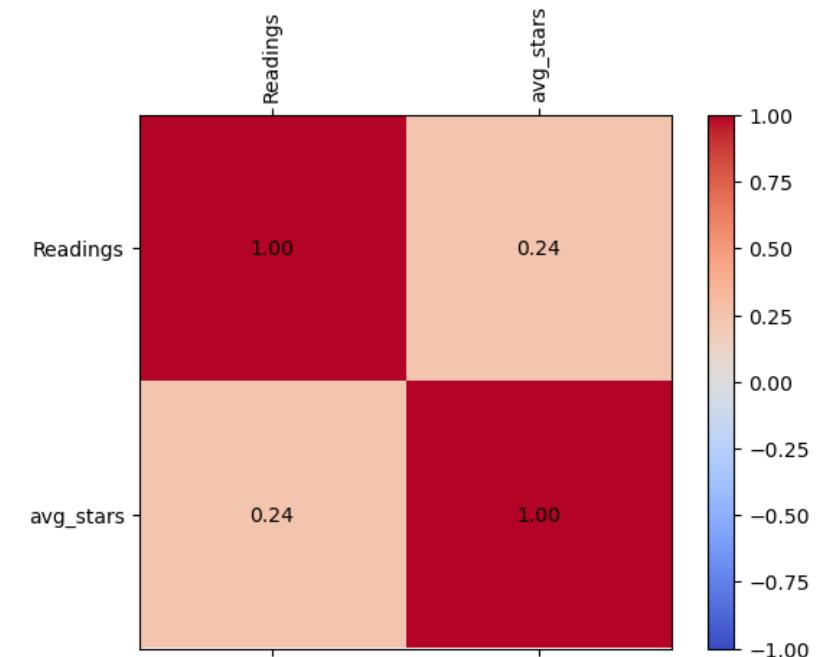
- **weather, finance, sports, news e lifestyle** riescono a mantenere interessi bimestrali degli utenti, ma *ottenendo valutazioni di poco sopra la linea di galleggiamento = 3*
 - **art**, per i mesi in cui si manifesta interesse, propone delle valutazioni medie di poco sopra la sufficienza
- Per il resto dei mesi in cui si manifesta interesse, l'accoglimento, in termini di valutazione è molto FREDDO
- A giudicare dalla scala cromatica della Heat Map, sembrano esserci delle convergenze in termini di valutazione e numero di letture. Sarebbe interessante monitorare se vi sia qualche CORRELAZIONE tra la valutazione media stars e gli EFFETTI TEMPORALI di essa sulle letture readings successive.

ANALISI DI CORRELAZIONE

CORRELAZIONE LETTURE – VALUTAZIONE MEDIA

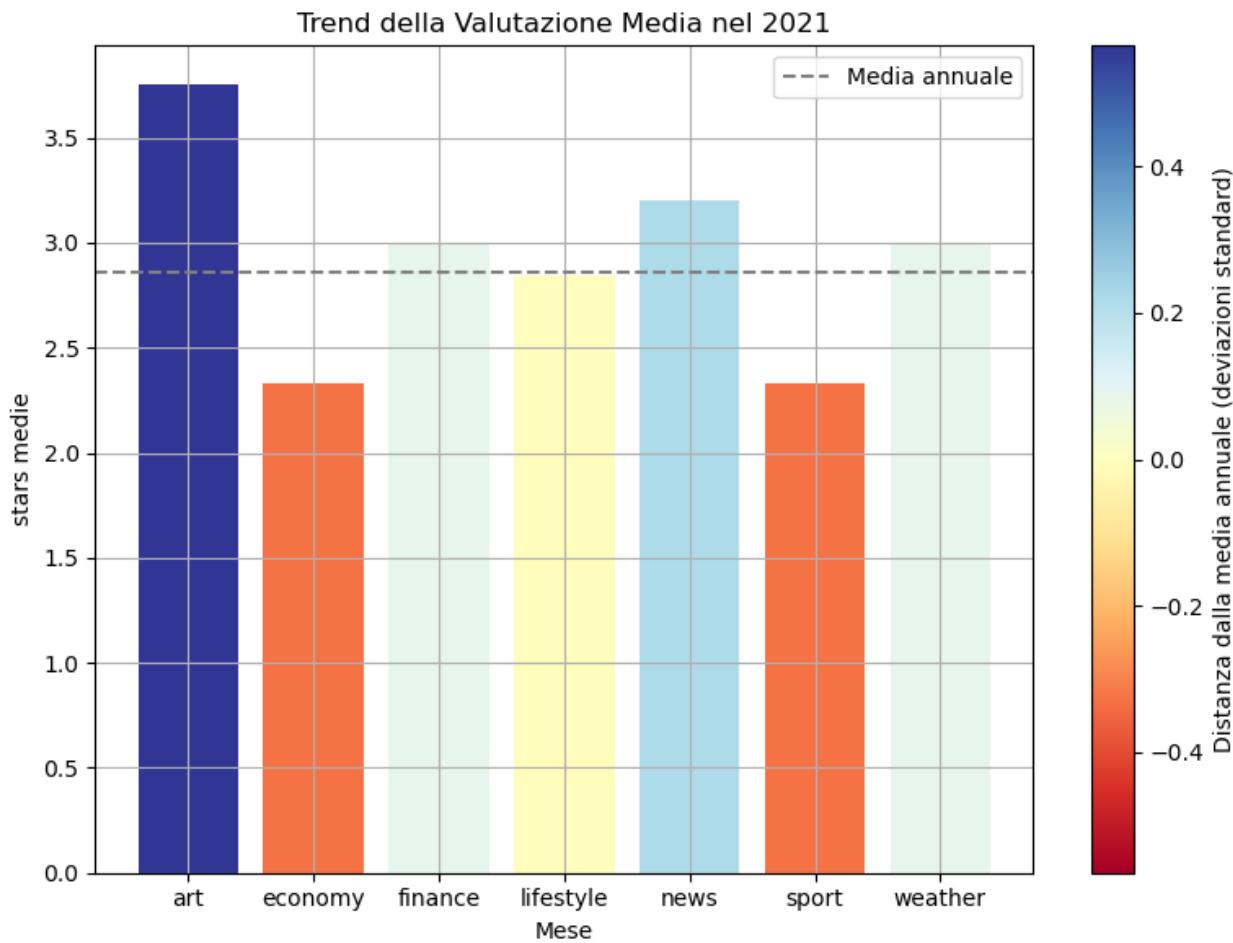
Creata un dataframe MERGED tra valutazioni **stars** e numero di letture mensili **Readings**, sembrerebbe non esserci una CORRELAZIONE tra le due variabili (0.241793)

La qualità delle notizie in termini di **redazione**, **attinenza**, **sensibilità**, **comunicazione** e - NON PER ULTIMO - **contenuto** sono leve che potrebbero incidere sulla **longevità** o **fedeltà** dei lettori nel lungo periodo.
Ma allo stato attuale, dato il coefficiente di correlazione stimato, sembrerebbe non esserci alcuna significatività al riguardo.





CATEGORY – VALUTAZIONE MEDIA



INTERAZIONE CATEGORIA – VALUTAZIONE MEDIA

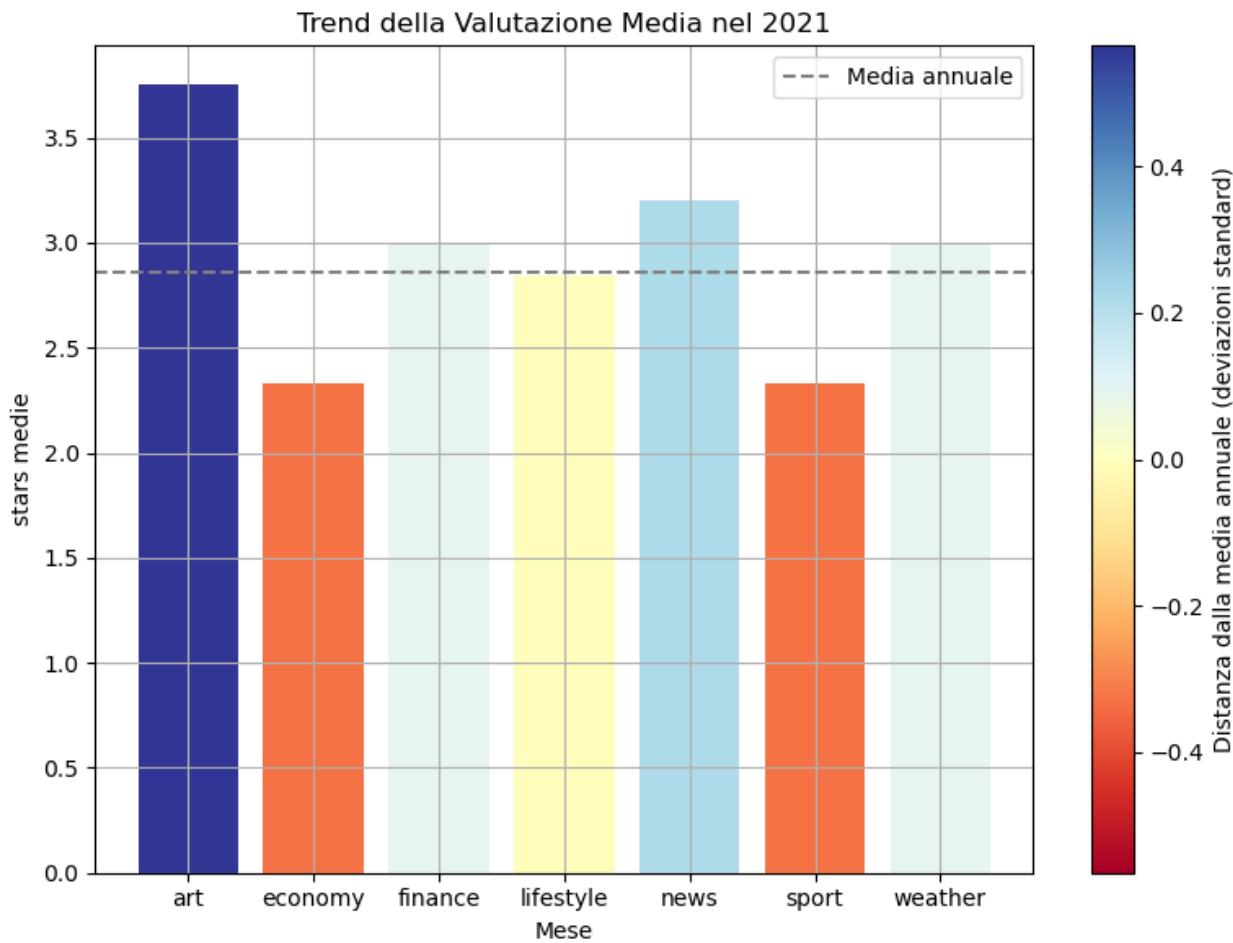
Come già visto per quanto riguarda l'andamento del numero di letture a livello mensile, utilizzo un *diagramma a barre* - con mappatura cromatica graduata - per mostrare il peso delle valutazioni **MIE stars**, per **category**, in relazione alla *distanza dalla media annuale* come

$$val.\text{media categoria (norm.)} = \frac{val.\text{media annua categoria} - valutazione media annua}{\text{deviazione standard}}$$

Sfruttando tale INDICATORE come *indice per la graduazione* del plot

- all'aumentare di tale differenza, si *intensifica* la *colorazione*;
- al **collasso** verso il

$$\lim_{x \rightarrow t^\pm} (val.\text{media annua cat.}(x) - val.\text{media annua}(t)) = 0$$

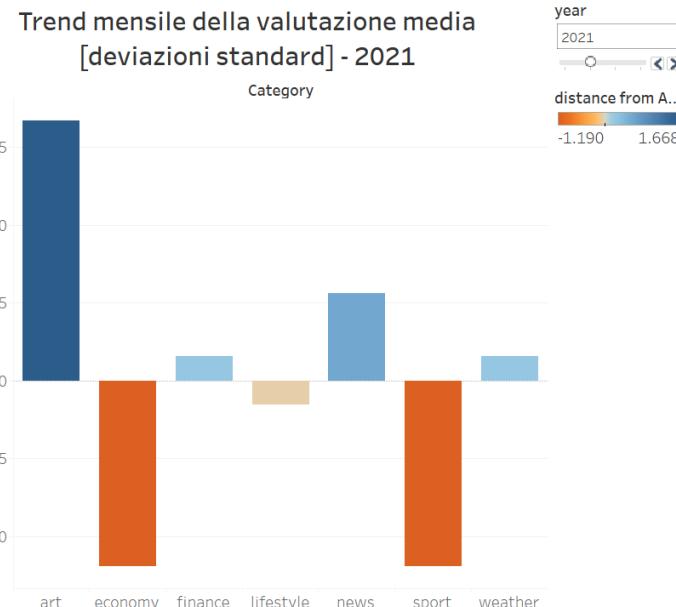
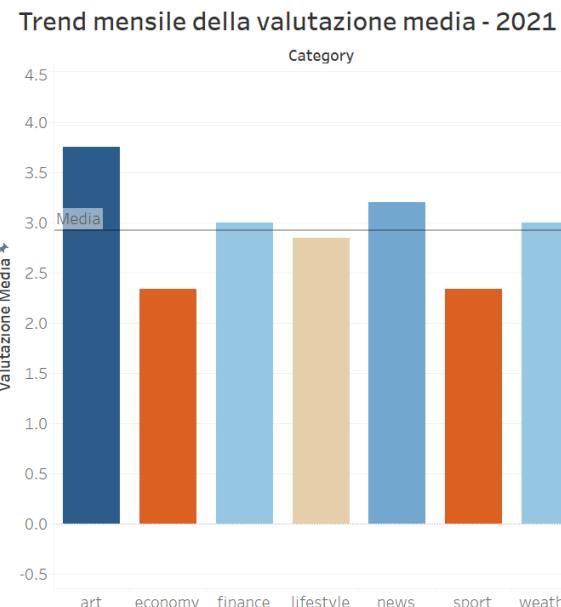


INTERAZIONE CATEGORIA – VALUTAZIONE MEDIA

- la scala tende ad un **coloratura sbiadita di giallo/bianco**
 - se tale **INDICATORE** risulta $< \text{media annuale}$ la colorazione assume la **scala di colori caldi (ROSSO)**
 - se tale **INDICATORE** risulta $> \text{media annuale}$ la colorazione assume la **scala di colori freddi (BLU)**

N.B. A prescindere dal valore k della $stddev$, essendo considerabile come una **costante**, occorre valutare **solo** il comportamento del **numeratore** in termini di limite:

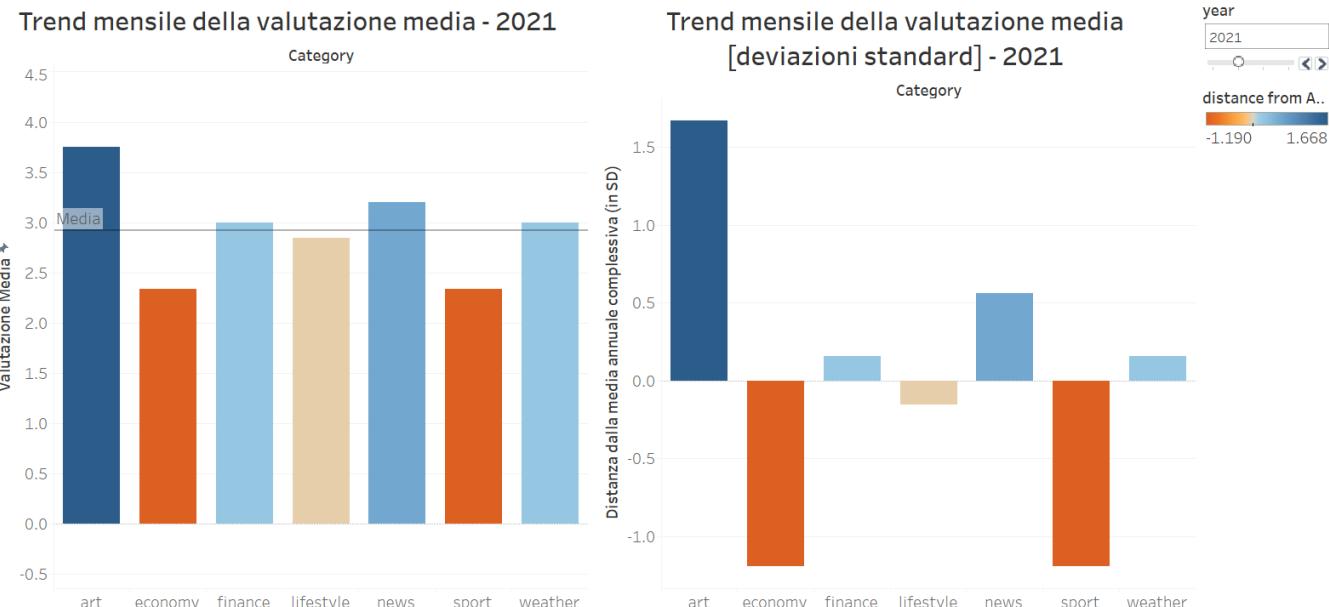
$$\left(\lim_{x \rightarrow t^\pm} \frac{(x-t) \sim 0}{k} \rightarrow 0 \right)$$



INTERAZIONE CATEGORIA – VALUTAZIONE MEDIA

Come si evince dal **bar graph** a fianco, solo una categoria ottiene una valutazione media annua **AMPIAMENTE SOPRA LA MEDIA**: art

- **finance, lifestyle, news, weather** sono le categorie per cui la valutazione media annua è almeno **UGUALE o SUPERIORE** alla **MEDIA ANNUA**, *che comunque risulta INSUFFICIENTE*
- solo **finance, news e weather** mostrano una valutazione media ANNUA di categoria almeno sufficiente



INTERAZIONE CATEGORIA – VALUTAZIONE MEDIA

- mentre **lifestyle**, **economy** e **sport** sono mediamente INSUFFICIENTI, dove le *ULTIME DUE si discostano negativamente in modo ampio rispetto alla media annua*.

Nel complesso, le **valutazioni MEDIE ANNUE** ottenute dagli articoli ci danno una indicazione di quanto il mercato dell'informazione offra una **proposta di valore mediamente SCARSA**. Vi è la possibilità, dunque, di poter penetrare nel mercato con un carattere distintivo tale da poter accumulare quote di mercato in modo agevole.

SEGMENTAZIONE DEMOGRAFICA



SEGMENTAZIONE DEMOGRAFICA – 2021

Decido di provare a sondare altri terreni, valutando quale sia la **distribuzione delle letture** in relazione alla **category** e alle **platform** utilizzate.

Effettuato il raggruppamento al fine di creare a una PIVOT per monitorare i **conteggi**, opto per l'analisi in termini di **conteggi normalizzati**, sulla base di questo criterio:

- $$\text{numero lettura categoria}_{ji} = \frac{(\text{numero di lettura}_{ji} - \text{media lettura}_j)}{\text{deviazione standard}_j}$$
$$j = \text{categoria di lettura}; i = 1 \rightarrow k : \text{tipologia di platform}$$
- Di fatto, verrà creato un indice, inteso come **numero di letture, per category e platform**, come differenza rispetto alla media di categoria, in deviazioni standard

SEGMENTAZIONE DEMOGRAFICA – 2021

Ai fini di ottimizzare il processo di valutazione, si decide di creare **4 funzioni – metodi**, sfruttando il dataset in formato *wide*:

- *Righe: le 3 platform*
- *Colonne: le variabili inerenti ogni **category**, raggruppate per **avg, sd, norm***

Queste funzioni sono state create al fine di poter *riutilizzarle* successivamente nell'analisi. Qui sotto ne citiamo la **definizione con i parametri**

- `def avg_calc_method(df)`
- `def std_calc_method(df)`
- `def norm_calc_method(df)`
- `def norm_extraction(original_df)`
- `def ren_var(df)`

Nella sostanza, ognuna delle prime tre crea la *media/deviazione standard//conteggio normalizzato* di **categoria**, definendo un numero di variabili pari al numero iniziale delle categorie (che in origine sono le colonne del raggruppamento).

Alla fine si avranno $N * 4$ colonne (dove N è il numero originario di categorie in colonne). La **quarta funzione** estraе solo le colonne dei conteggi normalizzati.

La **quinta** le *rinomina* ai fini di chiarezza.

SEGMENTAZIONE DEMOGRAFICA – 2021

Quello che ne risulterà è la costruzione di un *piccolo dataframe riassuntivo*, utile ad una Heat Map per **platform** e **category**, ma con VALORI NORMALIZZATI:



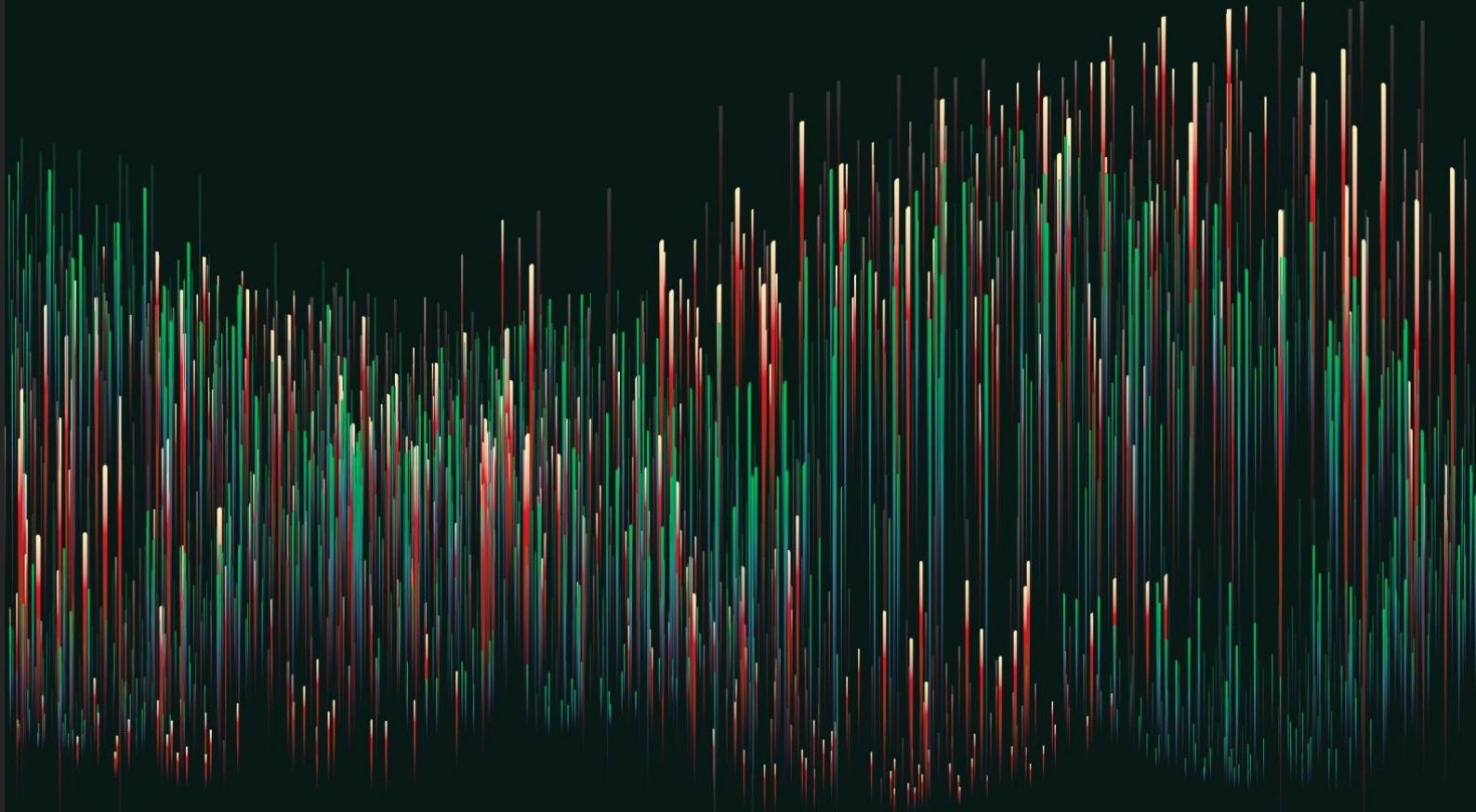
La scelta è data dal fatto che l'utilizzo di valori assoluti avrebbe inevitabilmente sovrastimato l'effetto DIMENSIONE per alcune categorie (es. **weather**, la categoria, come sappiamo, più interessante in base alle letture);



Decidendo di utilizzare VALORI NORMALIZZATI come *differenza* rispetto alla MEDIA di CATEGORIA, pesata per la DEVIAZIONE STANDARD di CATEGORIA, possiamo *ridurre gli EFFETTI DIMENSIONALI PER LE CATEGORIE* e apprendere con più chiarezza, per ogni categoria, quale sia la piattaforma PREFERITA. MA ORA VEDIAMO GLI SCRIPT NEL DETTAGLIO.

SEGMENTAZIONE DEMOGRAFICA

PYTHON SCRIPT



SEGMENTAZIONE DEMOGRAFICA – SCRIPT

CALCOLO DEI CONTEGGI

Come primo step, viene creato un raggruppamento *unstack* per **platform** e **category**, dove vengono monitorati i conteggi assoluti del numero di letture che rispettano tali condizioni

```
platform_category_2021 = Papers_2021.groupby(['platform','category']).size().unstack(fill_value=0)  
platform_category_2021
```

	category	art	economy	finance	lifestyle	news	sport	weather
platform								
mobile	1	2	7	3	1	3	8	
pc	0	4	4	5	4	8	7	
tablet	3	3	4	5	5	4	9	

SEGMENTAZIONE DEMOGRAFICA – SCRIPT

CALCOLO DELLE MEDIE

Attraverso il **method avg_calc_method()** si accede al *dataframe* dei conteggi appena creato.

Con un ciclo **for** che itera sulle colonne (**category**) creo, per ognuna di esse, la corrispondente **mean()** di **category** tra le row (**platform**) disponibili:

- Creo una *istanza* **avg_value** che concatena una stringa *prefisso* «**avg_**» e la stringa del nome della colonna (**'name_of_cat'**)
- Questa stringa verrà utilizzata come **nome** per la *variabile media* associata alla categoria
- Di seguito viene poi calcolata la **mean()** di categoria, per ognuna delle colonne disponibili

Il metodo ritorna il *dataframe aggiornato* in cui saranno ora presenti $N * 2$ colonne, con N che rappresenta il numero di **colonne originali**.

```
def avg_calc_method(df):  
  
    for cat in list(df.columns):  
        # iterando sulle colonne rappresentanti le categorie, creo una variabile "avg_name_of_cat":  
        # estrago la STRINGA del nome della VARIABILE contenuta momentaneamente in 'cat' e la  
        # uso come SUFFISSO in una nuova variabile - il cui PREFISSO sarà 'avg_' - che rappresenterà il VALOR MEDIO di categoria  
        # tra le platform disponibili. Questa Stringa verrà usata per inizializzare la nuova colonna nel dataframe IN USO  
  
        ### stringa usata per inizializzare la nuova variabile: `avg_` + `name_of_category`  
        avg_value = "avg_" + str(cat)  
  
        # inizializzazione della variabile `avg_name_of_category``  
        df[avg_value] = df[str(cat)].mean()  
  
    ## ESEMPIO LOCALE  
    ## ora abbiamo 7 X 2 = 14 colonne:  
    ##### 7 rappresentano i valori NOMINALI di lettura, per CATEGORIA, attraverso le diverse PLATFORM  
    ##### 7 rappresentano il valor MEDIO delle letture, per CATEGORIA, in base alla distribuzione NOMINALE tra le platform  
  
    ##### es. avg_economy = sum(economy)/N_economy (N_platform o N_rows) = (2+4+3)/3 = 9/3 = 3  
  
    return df  
  
platform_category_2021 = avg_calc_method(platform_category_2021)
```

	category	art	economy	finance	lifestyle	news	sport	weather	avg_art	avg_economy	avg_finance	avg_lifestyle	avg_news	avg_sport	avg_weather
platform	mobile	1	2	7	3	1	3	8	1.333333	3.0	5.0	4.333333	3.333333	5.0	8.0
	pc	0	4	4	5	4	8	7	1.333333	3.0	5.0	4.333333	3.333333	5.0	8.0
	tablet	3	3	4	5	5	4	9	1.333333	3.0	5.0	4.333333	3.333333	5.0	8.0

SEGMENTAZIONE DEMOGRAFICA – SCRIPT

CALCOLO DELLE DEVIAZIONI STANDARD

Attraverso il `method std_calc_method()` si accede al `dataframe` dei conteggi aggiunto delle `mean()` appena create.

Con un ciclo `for` che itera sulle colonne (`category`) creo, per ognuna di esse, la corrispondente `std()` di categoria tra le row (`platform`) disponibili:

- Creo una *istanza* `std_value` che concatena una stringa prefisso «`std_`» e la stringa del nome della colonna ('`name_of_cat`')
- Questa stringa verrà utilizzata come `nome` per la *variabile deviazione standard* associata alla categoria
- Di seguito viene poi calcolata la `std()` di categoria, per ognuna delle colonne disponibili
- N.B. viene valutato che le variabili a cui si accede NON SIANO RELATIVE alle `mean()` attraverso il `method startswith()`

Il metodo ritorna il `dataframe` aggiornato in cui saranno ora presenti $N * 3$ colonne, con N che rappresenta il numero di colonne originali.

```
def std_calc_method(df):  
  
    for cat in list(df.columns):  
        # iterando sulle colonne delle categorie, creo una variabile "std_`name_of_cat`":  
        # estraggo la STRINGA del nome della VARIABILE contenuta momentaneamente in 'cat'  
        # CONTROLLO CHE TALE VARIABILE NON SIA UNA MEDIA e la  
        # uso come SUFFISSO in una nuova variabile - il cui PREFISSO sarà `std_` - che rappresenterà la DEVIAZIONE STANDARD di categoria  
        # tra le platform disponibili. Questa Stringa verrà usata per inizializzare la nuova colonna nel dataframe IN USO  
  
        #controllo che la variabile nel ciclo NON SIA UNA MEDIA  
        if cat.startswith("avg_")==0:  
            ### stringa usata per inizializzare la nuova variabile: `std_` + `name_of_category`  
            std_value = "std_" + str(cat)  
  
            # inizializzazione della variabile `std_`name_of_category``  
            df[std_value] = df[str(cat)].std()  
  
    ##ESEMPIO LOCALE  
    ## ora abbiamo 7 X 3 = 21 colonne:  
    ##### 7 rappresentano i valori NOMINALI di lettura, per CATEGORIA, attraverso le diverse PLATFORM  
    ##### 7 rappresentano il valor MEDIO delle letture, per CATEGORIA, in base alla distribuzione NOMINALE tra le platform  
    ##### 7 rappresentano la DEVIAZIONE STANDARD delle letture, per CATEGORIA  
  
    return df  
  
platform_category_2021 = std_calc_method(platform_category_2021)
```

weather	avg_art	avg_economy	avg_finance	...	avg_news	avg_sport	avg_weather	std_art	std_economy	std_finance	std_lifestyle	std_news	std_sport	std_weather
8	1.333333	3.0	5.0	...	3.333333	5.0	8.0	1.527525	1.0	1.732051	1.154701	2.081666	2.645751	1.0
7	1.333333	3.0	5.0	...	3.333333	5.0	8.0	1.527525	1.0	1.732051	1.154701	2.081666	2.645751	1.0
9	1.333333	3.0	5.0	...	3.333333	5.0	8.0	1.527525	1.0	1.732051	1.154701	2.081666	2.645751	1.0

SEGMENTAZIONE DEMOGRAFICA – SCRIPT

CALCOLO DEI CONTEGGI ‘NORMALIZZATI’

Attraverso il **method norm_calc_method()** si accede al **dataframe** dei conteggi aggiunto delle **mean()** e delle **std()** appena create.

Con un ciclo **for** che itera sulle colonne (**category**) creo, per ognuna di esse, la corrispondente **norm()** di categoria tra le row (**platform**) disponibili:

- Creo una *istanza* **norm_value** che concatena una stringa **prefisso «norm_value_»** e la stringa del nome della colonna (**‘name_of_cat’**)
- Creo altresì due *istanze* atte a concatenare la **stringa** del **nome** delle colonne analizzata nel ciclo ai **prefissi ‘avg_’ e ‘std_’**
- La stringa **norm_value** verrà utilizzata come **nome** per la **variabile conteggio normalizzato** associata alla categoria, così come le altre due, usate, per accedere alla **mean()** e alla **std()** di categoria

```
def norm_calc_method(df):  
  
    for cat in list(df.columns):  
        # iterando sulle colonne delle categorie, creo una variabile "norm_value_name_of_cat":  
        # estraggo la STRINGA del nome della VARIABILE contenuta momentaneamente in 'cat'  
        # CONTROLLO CHE TALE VARIABILE NON SIA UNA MEDIA o UNA DEVIAZIONE STANDARD e la  
        # uso come suffisso in una nuova variabile - che inizia con 'norm_value_` - che rappresenterà LA DIFFERENZA DALLA MEDIA di categoria, basata  
        # sulla DEVIAZIONE STANDARD, tra le platform disponibili. Questa Stringa verrà usata per inizializzare la nuova colonna nel dataframe  
        # IN USO  
        if cat.startswith("avg_")==0 | cat.startswith("std_")==0:  
  
            #stringa che permette di individuare a quale colonna riferirsi in termini di MEDIA (es. se `cat` == `art`, questa sarà `avg_art`)  
            ## verrà usata per accedere alla variabile presente nel dataset, utile a fare il calcolo della variabile `norm_value_`  
            avg_var = "avg_"+str(cat)  
  
            #stringa che permette di individuare a quale colonna riferirsi in termini di DEVIAZIONE STANDARD (es. se `cat` == `art`, questa sarà `std_art`)  
            ## verrà usata per accedere alla variabile presente nel dataset, utile a fare il calcolo della variabile `norm_value_`  
            std_var = "std_"+str(cat)  
  
            ### stringa usata per inizializzare la nuova variabile: `norm_value` + `name_of_category`  
            norm_value = "norm_value_"+str(cat)  
            # inizializzazione della variabile `norm_value_name_of_category``  
            df[norm_value] = (df[cat] - df[avg_var])/df[std_var]  
  
            ####ESSEMPIO LOCALE  
            ## ora abbiamo 7 X 4 = 28 colonne:  
            ##### 7 rappresentano i valori NOMINALI di Letture, per CATEGORIA, attraverso le diverse PLATFORM  
            ##### 7 rappresentano il valor MEDIO delle letture, per CATEGORIA, in base alla distribuzione NOMINALE tra le platform  
            ##### 7 rappresentano la DEVIAZIONE STANDARD delle letture, per CATEGORIA  
            ##### 7 rappresentano i valori NORMALIZZATI come DIFFERENZA RISPETTO ALLA MEDIA DI CATEGORIA, pesata sulla DEVIAZIONE STANDARD, delle letture,  
            ##### per CATEGORIA  
  
            return df  
  
platform_category_2021 = norm_calc_method(platform_category_2021)
```

	_news	std_sport	std_weather	norm_value_art	norm_value_economy	norm_value_financial	norm_value_lifestyle	norm_value_news	norm_value_sport	norm_value_weather	
81666	2.645751	1.0	-0.218218	-1.0	1.154701	-1.154701	-1.120897	-0.755929	0.0		
81666	2.645751	1.0	-0.872872	1.0	-0.577350	0.577350	0.320256	1.133893	-1.0		
81666	2.645751	1.0	1.091089	0.0	-0.577350	0.577350	0.800641	-0.377964	1.0		

SEGMENTAZIONE DEMOGRAFICA – SCRIPT

CALCOLO DEI CONTEGGI ‘NORMALIZZATI’

Di seguito viene poi calcolato tale conteggio *normalizzato* di categoria, per ognuna delle colonne disponibili, come segue

$$\text{numero di letture normalizzato}_{ji} = \frac{(\text{numero di letture}_{ji} - \text{media lettura}_j)}{\text{deviazione standard } d_j}$$

j = categoria di lettura; i = 1 → k : tipologia di platform

- N.B. viene valutato che le variabili a cui si accede NON SIANO RELATIVE alle **mean()** o **std()** attraverso il **method startswith()**

Il metodo ritorna il *dataframe aggiornato* in cui saranno ora presenti $N * 4$ colonne, con N che rappresenta il numero di **colonne originali**.

```
def norm_calc_method(df):  
  
    for cat in list(df.columns):  
        # iterando sulle colonne delle categorie, creo una variabile "norm_value_name_of_cat":  
        # estraggo la STRINGA del nome della VARIABILE contenuta momentaneamente in 'cat'  
        # CONTROLLO CHE TALE VARIABILE NON SIA UNA MEDIA o UNA DEVIAZIONE STANDARD e la  
        # uso come suffisso in una nuova variabile - che inizia con 'norm_value_` - che rappresenterà LA DIFFERENZA DALLA MEDIA di categoria, basata  
        # sulla DEVIAZIONE STANDARD, tra le platform disponibili. Questa Stringa verrà usata per inizializzare la nuova colonna nel dataframe  
        # IN USO  
        if cat.startswith("avg_")==0 | cat.startswith("std_")==0:  
  
            #stringa che permette di individuare a quale colonna riferirsi in termini di MEDIA (es. se `cat` == `art`, questa sarà `avg_art`)  
            ## verrà usata per accedere alla variabile presente nel dataset, utile a fare il calcolo della variabile `norm_value_`  
            avg_var = "avg_" + str(cat)  
  
            #stringa che permette di individuare a quale colonna riferirsi in termini di DEVIAZIONE STANDARD (es. se `cat` == `art`, questa sarà `std_art`)  
            ## verrà usata per accedere alla variabile presente nel dataset, utile a fare il calcolo della variabile `norm_value_`  
            std_var = "std_" + str(cat)  
  
            ### stringa usata per inizializzare La nuova variabile: `norm_value` + `name_of_category`  
            norm_value = "norm_value_" + str(cat)  
            # inizializzazione della variabile `norm_value_name_of_category``  
            df[norm_value] = (df[cat] - df[avg_var]) / df[std_var]  
  
            ####ESSEMPIO LOCALE  
            ## ora abbiamo 7 X 4 = 28 colonne:  
            ##### 7 rappresentano i valori NOMINALI di Letture, per CATEGORIA, attraverso le diverse PLATFORM  
            ##### 7 rappresentano il valor MEDIO delle letture, per CATEGORIA, in base alla distribuzione NOMINALE tra le platform  
            ##### 7 rappresentano la DEVIAZIONE STANDARD delle letture, per CATEGORIA  
            ##### 7 rappresentano i valori NORMALIZZATI come DIFFERENZA RISPETTO ALLA MEDIA DI CATEGORIA, pesata sulla DEVIAZIONE STANDARD, delle letture,  
            ##### per CATEGORIA  
  
            return df  
  
platform_category_2021 = norm_calc_method(platform_category_2021)
```

	_news	std_sport	std_weather	norm_value_art	norm_value_economy	norm_value_financial	norm_value_lifestyle	norm_value_news	norm_value_sport	norm_value_weather	
81666	2.645751	1.0	-0.218218		-1.0	1.154701	-1.154701	-1.120897	-0.755929		0.0
81666	2.645751	1.0	-0.872872		1.0	-0.577350	0.577350	0.320256	1.133893		-1.0
81666	2.645751	1.0	1.091089		0.0	-0.577350	0.577350	0.800641	-0.377964		1.0

SEGMENTAZIONE DEMOGRAFICA – SCRIPT

ESTRAZIONE DEI CONTEGGI NORMALIZZATI

Attraverso poi il method **norm_extraction()** si accede al *dataframe* completo per estrarre i soli conteggi *normalizzati*.

Estratta una *lista* delle variabili del *dataframe*, itero e creo una *sublist* che contiene le SOLE variabili *normalizzate* utilizzando il metodo **startswith()**

Uso la **sublist new_list** appena creata per accedere alle sole variabili di interesse (le *normalizzate*) e ritornarle nel *dataframe* da utilizzare per la **Heat Map**

```
def norm_extraction(original_df):

    #estratto una lista delle colonne del dataframe IN USO
    varlist = list(original_df.columns)
    varlist

    #tengo le variabili normalizzate `norm_value_`
    newlist = [x for x in varlist if x.startswith('norm_value_')]

    ##utilizzo questa nuova lista per selezionare le variabili che mi serviranno per la HEAT MAP
    new_df = original_df[newlist]
    return new_df

norm_platform_category_2021 = norm_extraction(platform_category_2021)
```

category	norm_value_art	norm_value_economy	norm_value_finance	norm_value_lifestyle	norm_value_news	norm_value_sport	norm_value_weather	
platform								
mobile	-0.218218	-1.0	1.154701	-1.154701	-1.120897	-0.755929	0.0	
pc	-0.872872	1.0	-0.577350	0.577350	0.320256	1.133893	-1.0	
tablet	1.091089	0.0	-0.577350	0.577350	0.800641	-0.377964	1.0	

SEGMENTAZIONE DEMOGRAFICA – SCRIPT

RINOMINA DELLE VARIABILI

Infine, con il **method ren_var()** rinomino le variabili normalizzate ai fini estetici per la Heat Map.

Per ogni *variabile category*,

- con un ciclo **for** e
- valutando che la variabile sia una *normalizzata* attraverso il **method startswith()**

Creo una istanza **new_name**, che ha l'obiettivo di estrarre i caratteri della stringa del nome della variabile *successivi* al *suffisso* ‘**norm_value_**’, comune a **tutte** le variabili (**length** fissa a 11 caratteri)

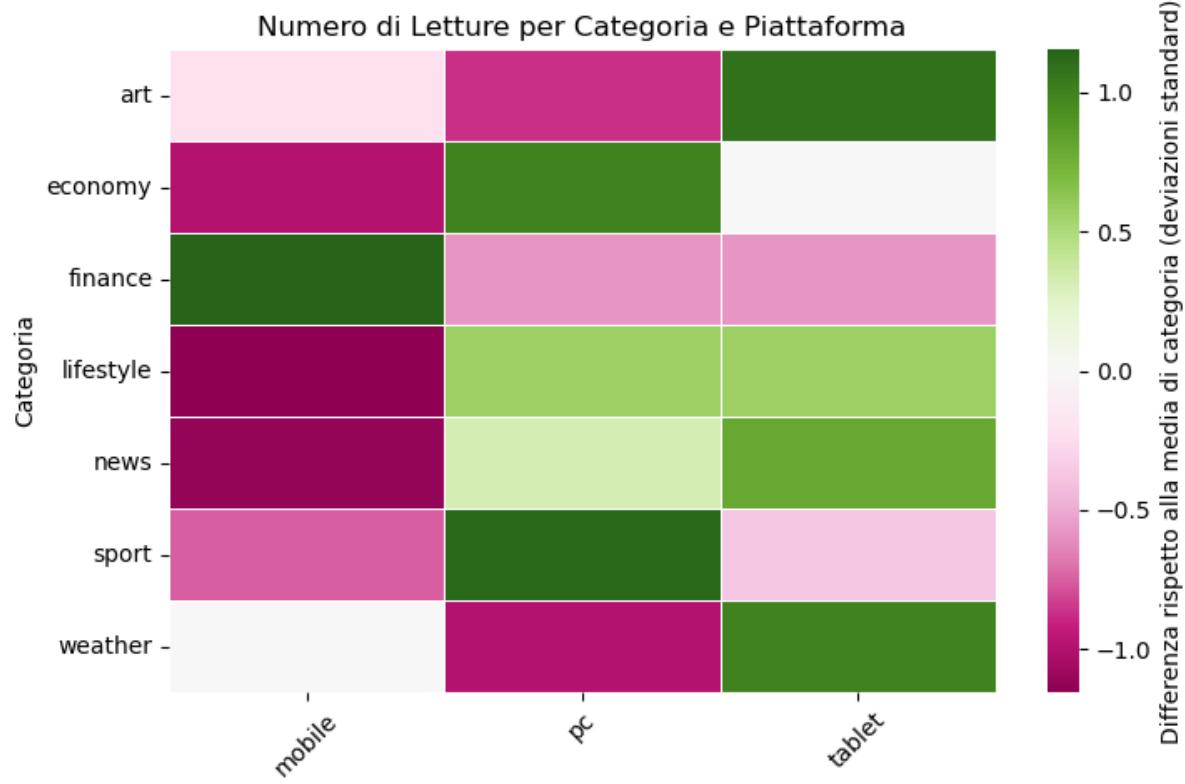
Rinomino, dunque, la relativa variabile **var**, attualmente analizzata, con il **new_name**, applicandolo definitivamente al *dataframe* (**inplace = True**) e viene ritornato aggiornato.

```
def ren_var(df):  
  
    ## RINOMINO LE VARIABILI rimuovendo `norm_value_`  
  
    for var in list(df):  
        if str(var).startswith('norm_value_'):  
            # valutata la natura della variabile che 'var' assume nel ciclo (inizi con `norm_value_`)  
            #estraiamo il nome originale, glissando i primi 11 caratteri (index 0-10), ovvero 'norm_value_'  
            new_name = str(var[11:])  
            #ora associamo l'estrazione, che equivale al nome ORIGINALE della variabile `art, economoy, weather etc.  
            df.rename(columns={var: new_name}, inplace=True)  
  
    return df  
  
norm_platform_category_2021 = ren_var(norm_platform_category_2021)
```

category	art	economy	finance	lifestyle	news	sport	weather
platform							
mobile	-0.218218	-1.0	1.154701	-1.154701	-1.120897	-0.755929	0.0
pc	-0.872872	1.0	-0.577350	0.577350	0.320256	1.133893	-1.0
tablet	1.091089	0.0	-0.577350	0.577350	0.800641	-0.377964	1.0

SEGMENTAZIONE DEMOGRAFICA

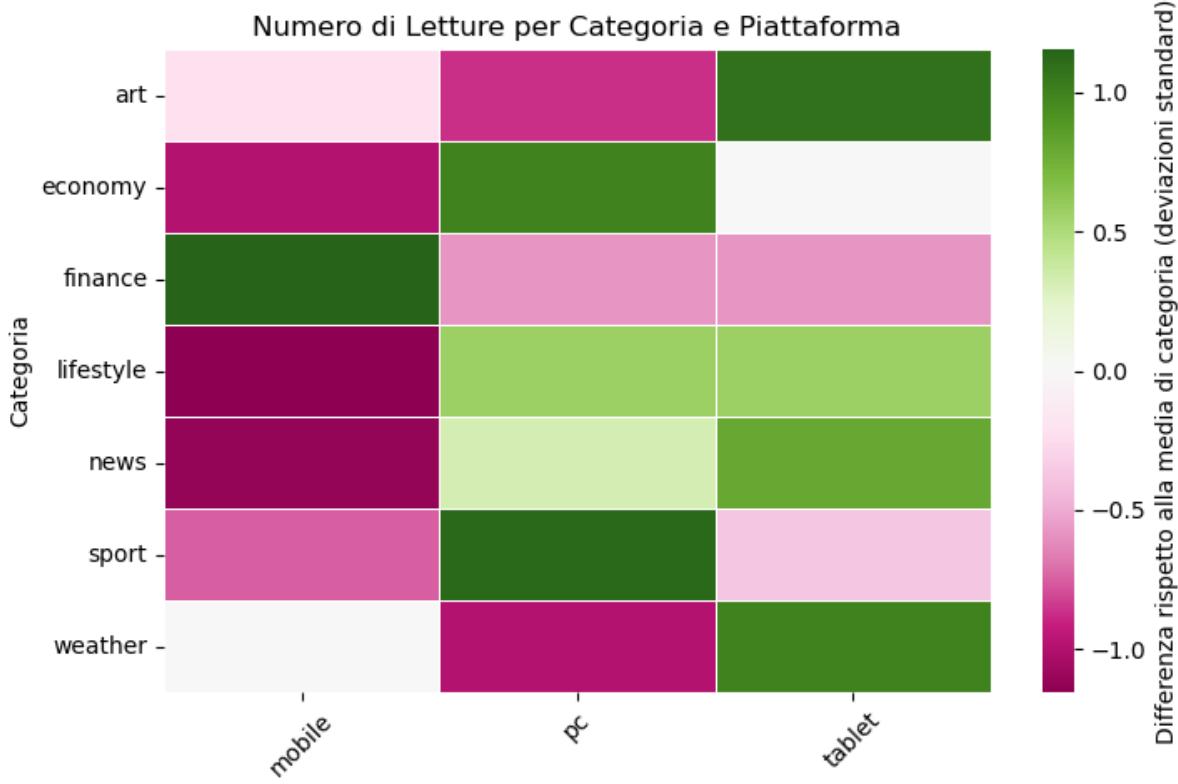
ANALISI
QUANTITATIVA E
CONSIDERAZIONI



SEGMENTAZIONE DEMOGRAFICA – HEAT MAP

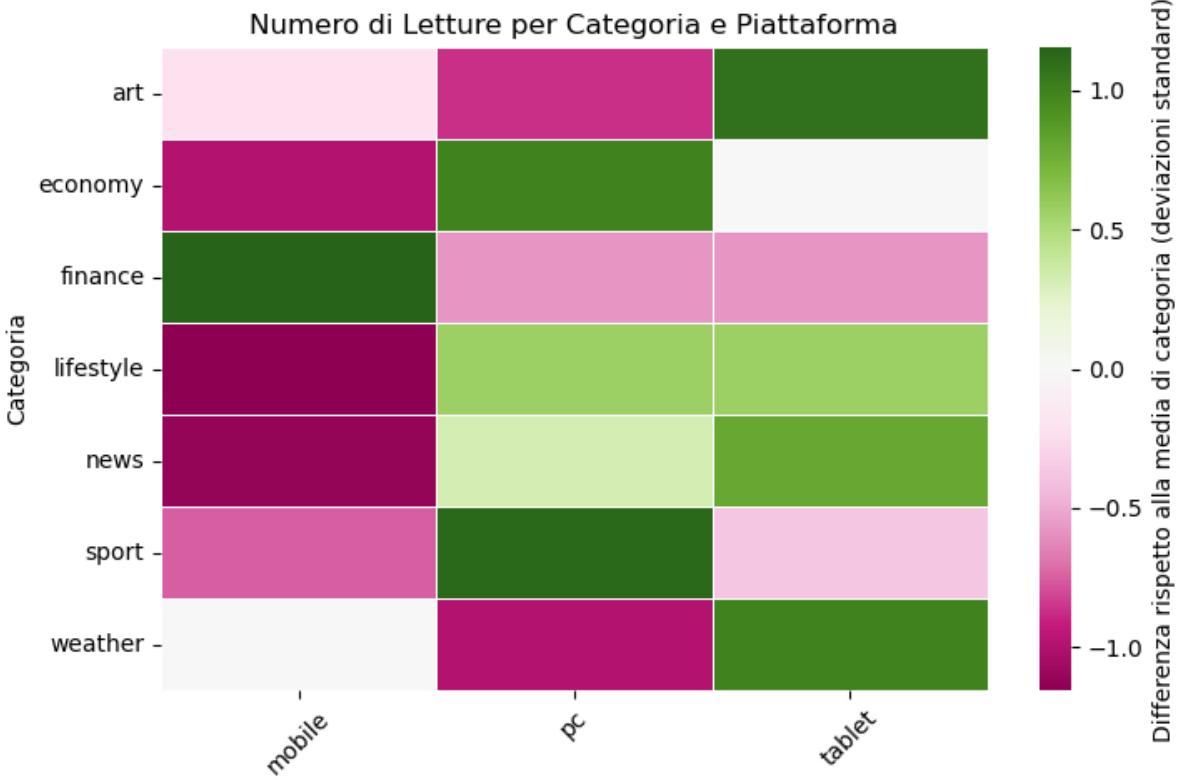
Ora possiamo evidenziare in modo abbastanza chiaro, quali potrebbero essere le **PIATTAFORME PREFERITE** in funzione della **CATEGORIA ESAMINATA**

- per **weather**, insieme ad **art** si mostra come i lettori siano indirizzati, mediamente, alla lettura dei relativi articoli attraverso l'utilizzo del **tablet**:
 - la differenza, PESATA per STDEV, positiva rispetto alla media di categoria (sebbene siano impercettibili a livello nominale (es. valori 8-7-9 per **weather**)) ci fa supporre che la scelta sia dovuta al fatto che gli articoli inerenti a queste due categorie *richiedano un device dalle dimensioni comparabili a quelle di un PC*:



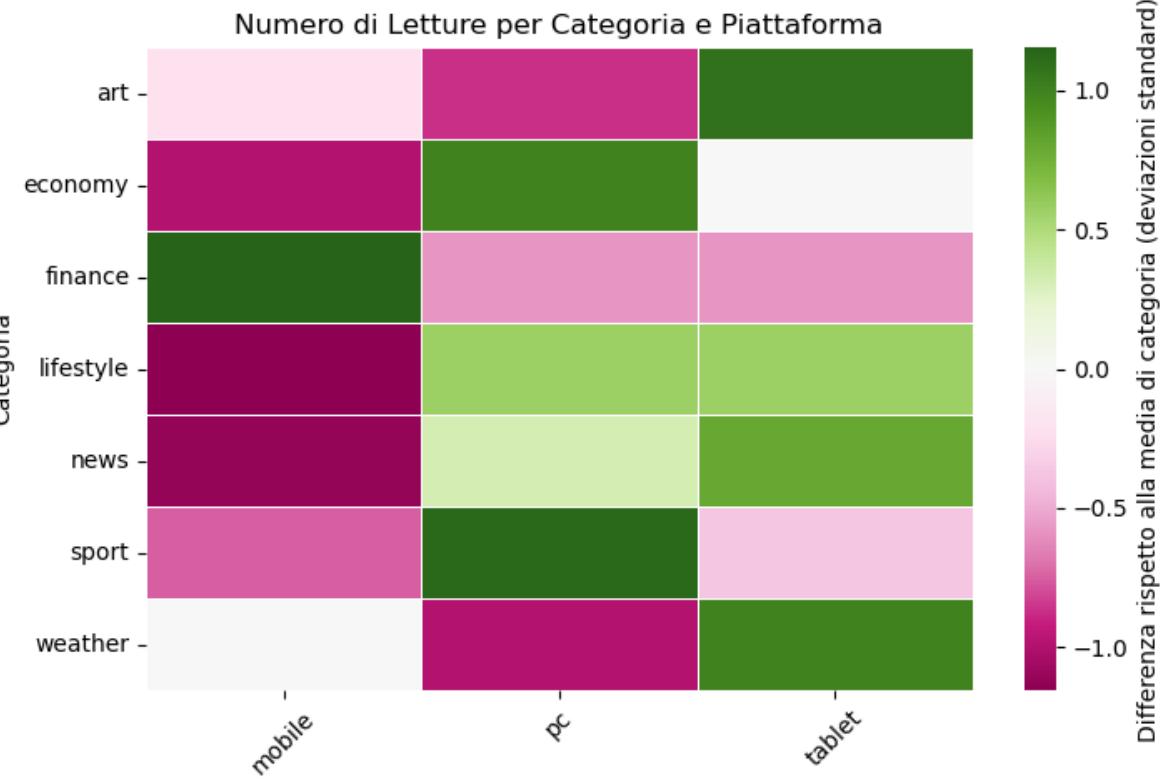
SEGMENTAZIONE DEMOGRAFICA – HEAT MAP

- un **tablet** ha il vantaggio funzionale di poter essere *facilmente accessibile anche durante trasferte*, ed è possibile tenerlo in borsa;
- in più trattasi di categorie il cui *impatto grafico* ha una sua **rilevanza**: **art** potrebbe richiedere di visualizzare immagini ad *alta risoluzione di opere d'arte* o di *installazioni avveniristiche di qualche designer*. Risulta il giusto **compromesso** tra **mobile** (portatile in modo agevole, ma con uno schermo, la cui dimensione potrebbe non rendere giustizia - così come non rendere agevole la consultazione - alle opere illustrate negli articoli) e **pc** (schermo di grandi dimensioni ma con un handling e una fruibilità non sempre immediata)
- allo stesso modo **weather** richiede, per gli utenti più consapevoli e informati, di **consultare dati e mappe cartografiche** con differenti *layer*, ad esempio sui carotaggi che aiutano a monitorare il meteo tra le diverse ere geologiche, la cui complessità e dimensione richiede l'ausilio di uno strumento il cui display sia di ratio appropriata, con lo stesso compromesso visto per **art**



SEGMENTAZIONE DEMOGRAFICA – HEAT MAP

- **economy** e **sport** mostrano una *differenza positiva*, rispetto alla media di categoria, per quanto riguarda la consultazione su **pc**:
 - le *tematiche* di **economy**, soprattutto quando si tratta di consultazione di, ad esempio, articoli inerenti la **valutazione di politiche pubbliche** con **strumenti econometrici**, richiedono l'ausilio di un device ad alta risoluzione (per visualizzare meglio tabelle riassuntive di *test di convergenza* o *scatter plot di regressioni* e *modelli probabilistici* come *probit* o *multiprobit*), ma anche un contesto relativamente adeguato, silenzioso, come studi e/o uffici, trattandosi di tematiche altamente complesse. L'ausilio del PC risulta relativamente motivato.
 - articoli di **sport** potrebbero avere una più agevole lettura su grandi schermi, come il **pc**, a riprodurre le dimensioni di una *testata giornalistica cartacea*, riconducendosi ad una sorta di *effetto nostalgia* per la carta stampata.



SEGMENTAZIONE DEMOGRAFICA – HEAT MAP

- **lifestyle** e **news** mostrano un comportamento simile come predilezione alla consultazione attraverso **pc** e **tablet**:
 - la **differenza** in termini **positivi**, rispetto alla media di categoria, potrebbe essere motivata dal fatto che articoli di questo genere potrebbero essere **consultati in fasi giornaliere non troppo stressanti** - magari durante i pasti o nel tempo libero - dunque non c'è la necessità di una preferenza specifica in termini di device
- **finance** si mostra come una sorta di *anticonformista* nell'economia di questa **Heat Map**: si mostrerebbe una predilezione per la consultazione attraverso **mobile**: la motivazione è abbastanza comprensibile, in quanto potrebbe trattarsi di articoli per cui si richiede una **consultazione immediata**, soprattutto se si ha a che fare, come mestiere, con i **mercati finanziari e mobiliari**, le **quotazioni in borsa**, **titoli azionari e/o trading di criptovalute**

Possiamo ora approfondire lo studio, concentrandosi sul tasso di attenzione dei lettori, in termini di preferenze di **LUNGHEZZA** dell'articolo, **USANDO LO STESSO PRINCIPIO DI RAPPRESENTAZIONE DI QUESTA ANALISI**



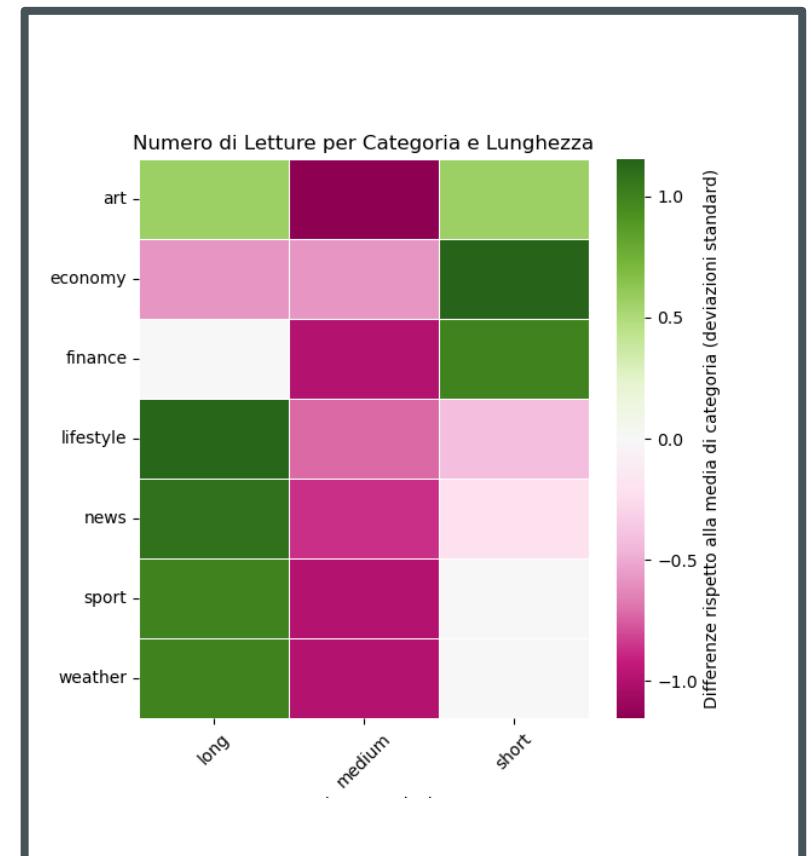
LUNGHEZZA MEDIA

LUNGHEZZA MEDIA DEGLI ARTICOLI, PER CATEGORIA

Sfruttando i *methods* appena creati, ora il nostro focus si concentra sull'individuare *hint* tra la lunghezza media degli articoli **length** e **category**.

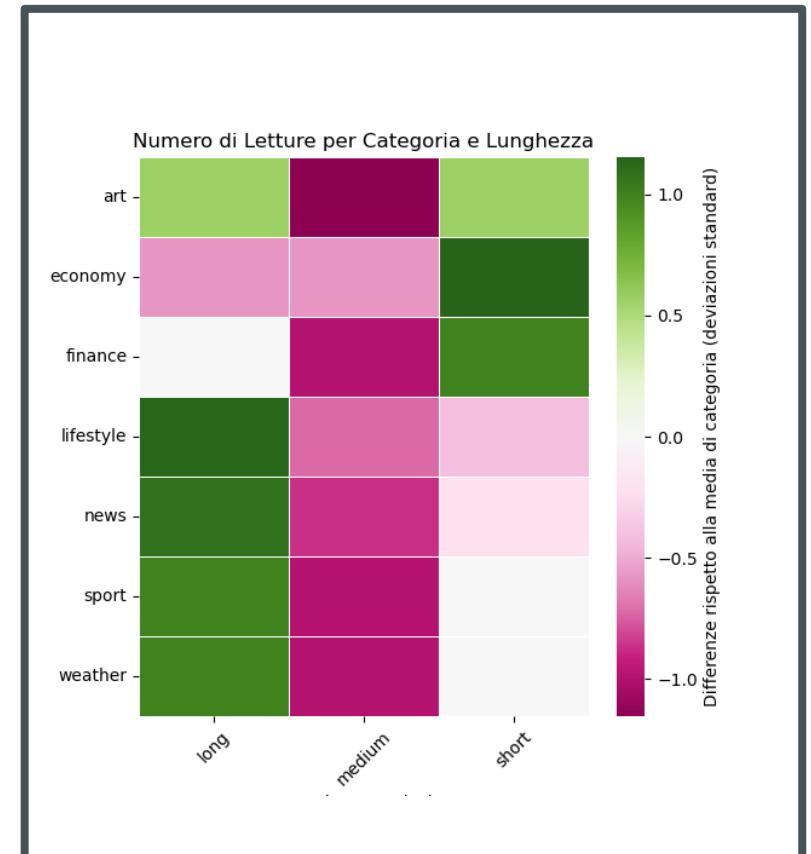
La Heat Map mostra dei comportamenti dei lettori discretamente evidenziabili: si presenta, per la grande maggioranza, un fenomeno di *POLARISMO* nelle scelte della lunghezza **length** degli articoli (o **short** o **long**).

- **lifestyle, news, sport, weather** mostrano come vi sia una preferenza, in termini di *differenza positiva rispetto alla media di categoria*, pesata per SD di categoria, alla lettura di articoli di lunga redazione **long**; vi è invece una *media preferenza* per le letture con un numero ridotto di caratteri **short**.
 - il fenomeno potrebbe essere motivato dal fatto che, per queste categorie, *si voglia andare più nel dettaglio*, mostrando una volontà ad approfondire specifiche tematiche. Queste possono *influenzare anche le decisioni del quotidiano* che possono avere *risvolti nel medio-lungo termine*, come nel caso di **lifestyle, news o weather**



LUNGHEZZA MEDIA DEGLI ARTICOLI, PER CATEGORIA

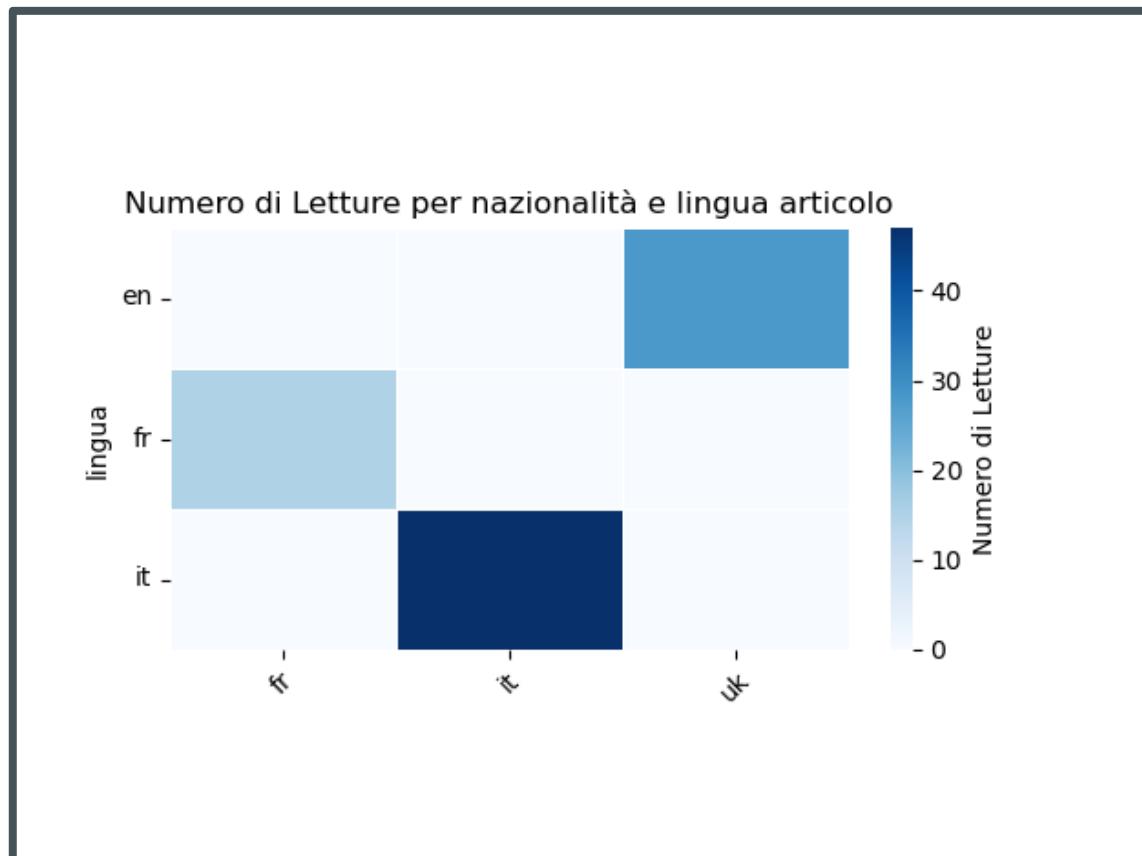
- nel caso di **economy** e **finance** il fenomeno è *l'opposto*: si mostra una predilezione sopra la media per le lettture **short**, mentre nella media per le long nel solo caso di **finance**:
 - la scelta potrebbe essere motivata dal fatto che, *benchè si tratti di argomenti che hanno risvolti sulla collettività*, vi sia una **predilezione alla sintesi**, in modo tale da poter prendere *decisioni in maniera si consapevole*, ma anche **rapida**, attraverso **NUMERI**, come avviene spesso nei **mercati mobiliari** o in borsa (cosa leggermente diversa nelle scelte politiche-economiche collettive).





LANGUAGE DIVERGENCE E GUSTI NAZIONALI

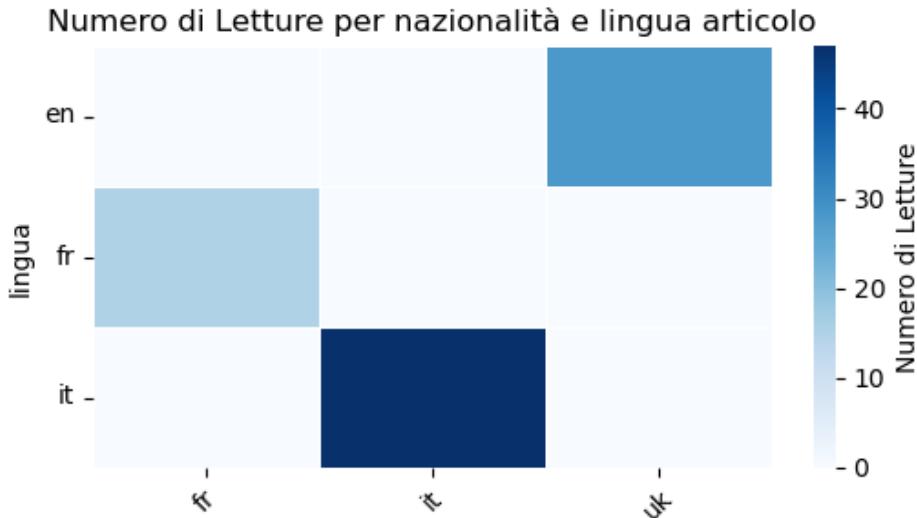
DISTRIBUZIONE LANGUAGE - COUNTRY



Come è chiaramente evidenziabile, *non esiste* alcuna **DEVIAZIONE LINGUISTICA** tra la *lingua in cui* è redatto l'articolo **language** e la *nazionalità del lettore country*: sembrerebbe dunque che vi sia una tendenza alla **soddisfazione informativa** attraverso l'ausilio di stampa nazionale, che non richieda invece di assumere informazioni attraverso stampa estera, condizione abbastanza limitante.

A questo proposito vorrei muovere una **CRITICA** nei confronti di un contesto come questo: è vero che tendenzialmente siamo propensi a guardare il nostro orticello, ma un occhio al giardino del vicino sicuramente accrescerebbe il nostro **bagaglio culturale**. Andrebbe dunque *valutata la conoscenza di altre lingue dei lettori* (che, nel caso italiano, mostra dei limiti nella conoscenza dell'inglese).

DISTRIBUZIONE LANGUAGE - COUNTRY



La conoscenza di altre lingue inciderebbe sul progresso culturale delle popolazioni nazionali: sarebbe un forte driver a motivare la scelta di altri autori, che redigono in altre lingue, considerando che la valutazione media degli articoli attualmente è piuttosto scarsa.

Sembrerebbe quasi che i lettori siano contenti di questa situazione, che non fa altro che mettere la controparte, gli editori, nelle condizioni di non "evolversi". Ma, in realtà, la scelta di altre lingue e, potenzialmente, di articoli più interessanti, potrebbe avere conseguenze a doppia valenza:

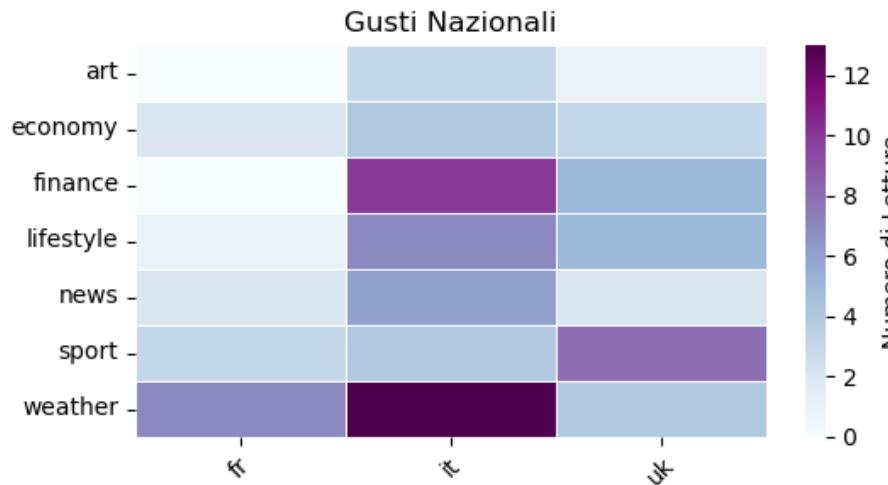
- potrebbe accrescere il bagaglio conoscitivo dei lettori, leggendo anche in altre lingue;
- motiverebbe i giornalisti a competere in termini di qualità e a produrre contenuti di livello, innalzando l'asticella del settore in termini di contenuti offerti

Come sempre, le responsabilità di un fenomeno stanno nel mezzo. Occorre un compromesso delle parti.

GUSTI PER NAZIONE: CATEGORY

Si evince come

- gli italiani prediligano letture inerenti le categorie **weather**, **finance** e **lifestyle**
 - sono tematiche strettamente *legate al quotidiano*, ma che possono avere risvolti anche nel medio-lungo termine:
 - l'Italia negli ultimi anni, come nel resto del mondo, ha subito *movimenti repentini* in termini di **cambiamento climatico**; si potrebbe dunque pensare ad una scelta collettiva di assumere **maggior consapevolezza** riguardo a quale sia la *situazione attuale del nostro ambiente* e a come *il nostro agire abbia influito su tale cambiamento*.
 - la scelta **lifestyle** si potrebbe motivare attraverso i **fenomeni sociali** occorsi durante il periodo di clausura forzata del Covid-19: le persone, abituate ad *intensi turni di lavoro*, ad una vita frenetica, sembra abbiano scoperto e riscoperto il **piacere di dedicarsi al proprio IO**, prediligendo scelte indirizzate al **benessere individuale**, dalle scelte di acquisti al supermercato, al seguire una vita sana, alla scelta di affidarsi ad un terapeuta, come uno **psicologo**. Sotto questo ultimo punto i decisori collettivi nazionali si sono movimentati, fornendo **"Bonus" per le famiglie** con reddito inferiore ad una specifica soglia, ma non è ancora abbastanza.



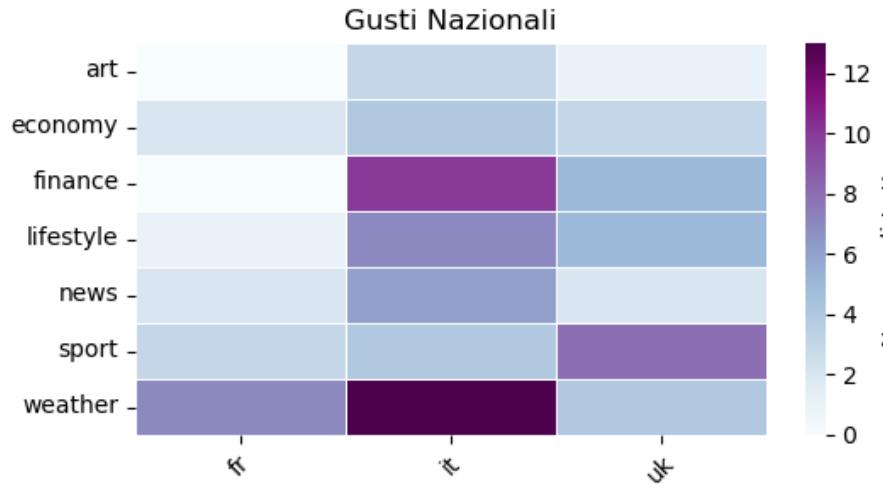
GUSTI PER NAZIONE: CATEGORY

- **finance** si motiva con una crescente consapevolezza, del popolo italiano, dello stato dell'arte del **sistema finanziario nazionale**, sempre più al **collasso**. Questo potrebbe incidere sulle scelte individuali in termini di accesso ad informazioni che rendano scelte del quotidiano più ponderate e supportate da dati.

Gli inglesi sono principalmente interessati a tematiche di **sport**, **finance** e **lifestyle**

- lo **sport** lo si potrebbe motivare con un crescente interesse durante la continua lotta, in diversi sport, con la rappresentanza italiana, durante le Olimpiadi di Tokyo 2021, gli Europei di calcio e durante altri meeting occorsi in quel periodo
- **finance** è un tema, nel mondo anglosassone, ormai ricorrente; per **lifestyle** potrebbe essere fatto un discorso simile a quello italiano, anche se le procedure di *isolamento forzato* erano ben **diverse** dal caso nostrano

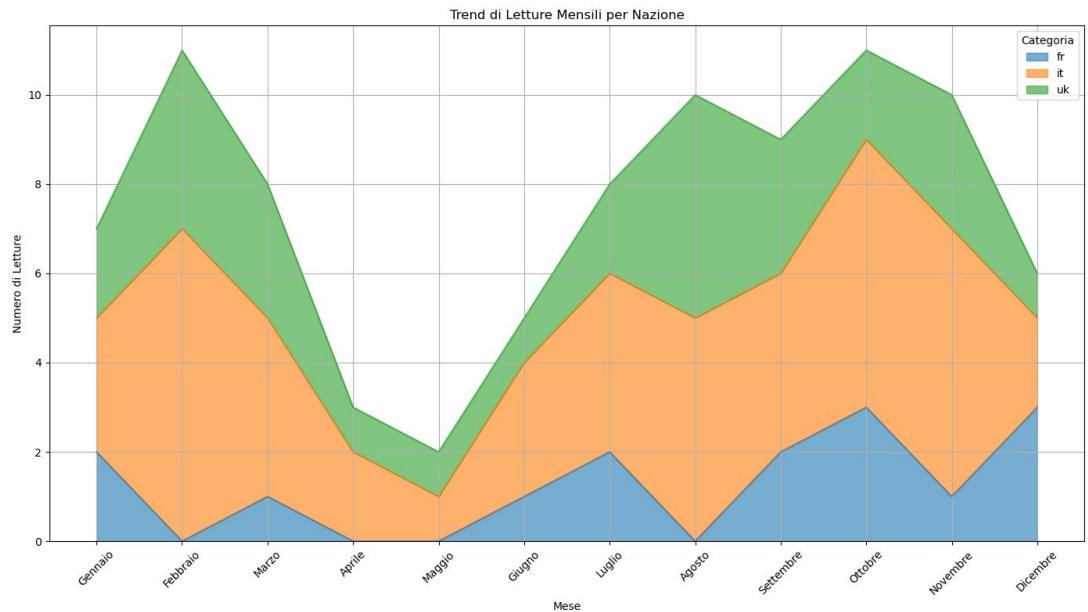
I francesi sono tendenzialmente *convergenti* verso articoli inerenti **weather**



GUSTI PER NAZIONE: TREND TEMPORALI

L'Area Plot mostra come la grande maggioranza delle letture siano riferite a lettori inglesi o italiani

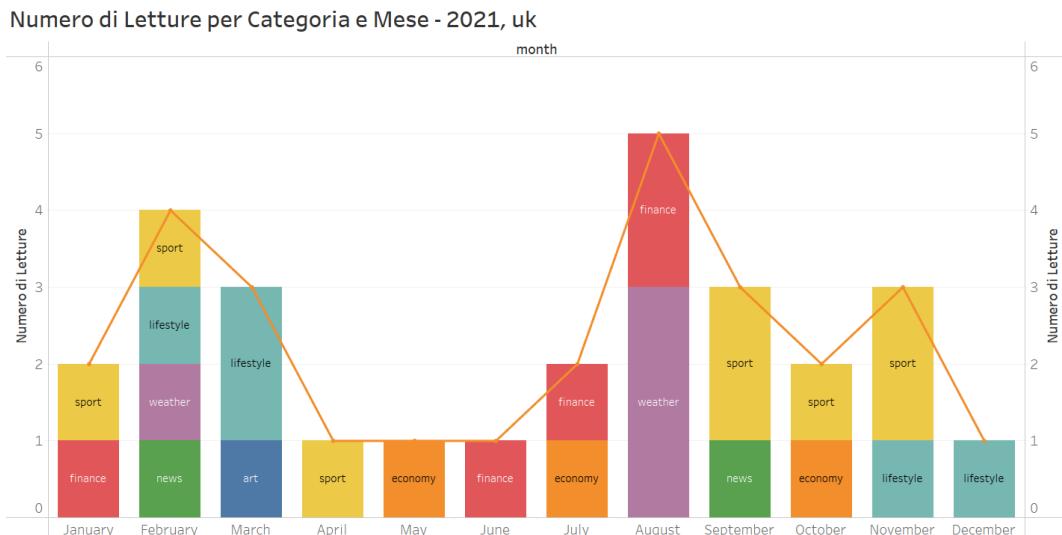
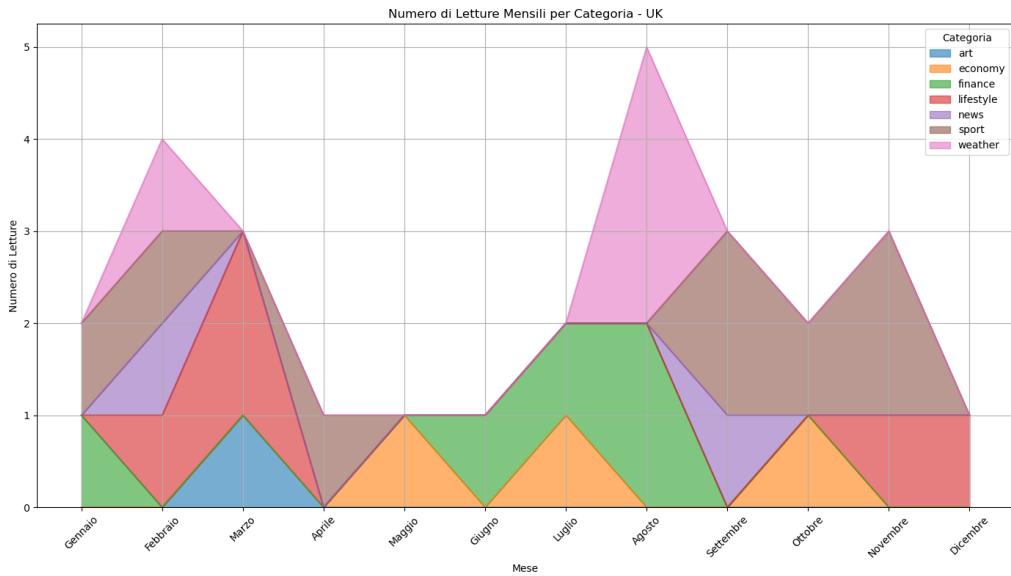
- gli italiani si mostrano i *più interessati alle letture*, evidenziando sempre *tassi superiori* rispetto agli inglesi, che mostrano un periodo di *flessione superiore a quello italiano*, nel caso primaverile
- i francesi si mostrano i *meno interessati alla lettura*, avendo periodi, come Aprile, Maggio e Agosto, in cui *non vengono letti articoli*



GUSTI PER NAZIONE: TREND TEMPORALI – CASO INGLESE

Gli Inglesi mostrano degli *interessi non continui*:

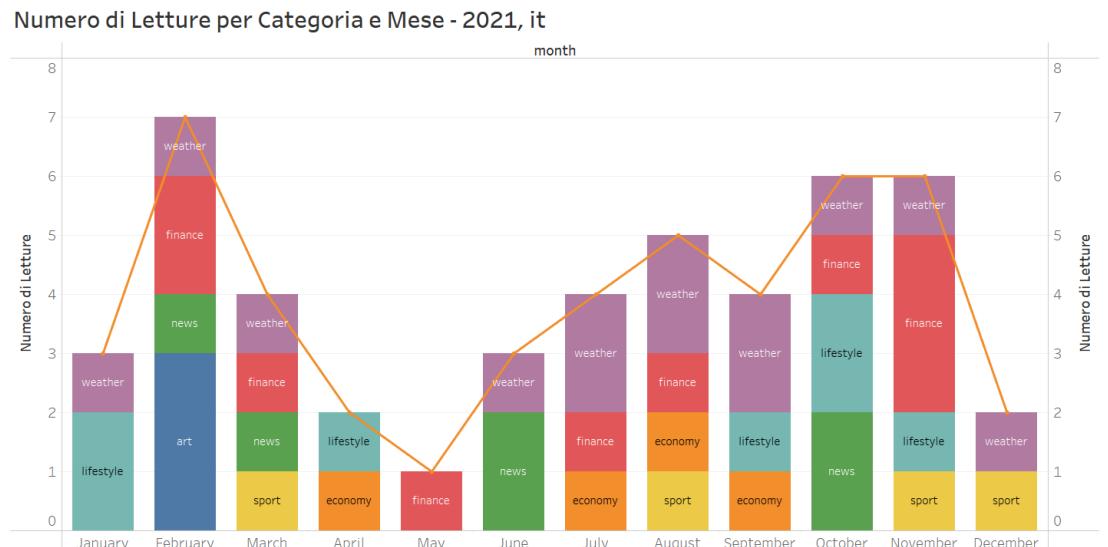
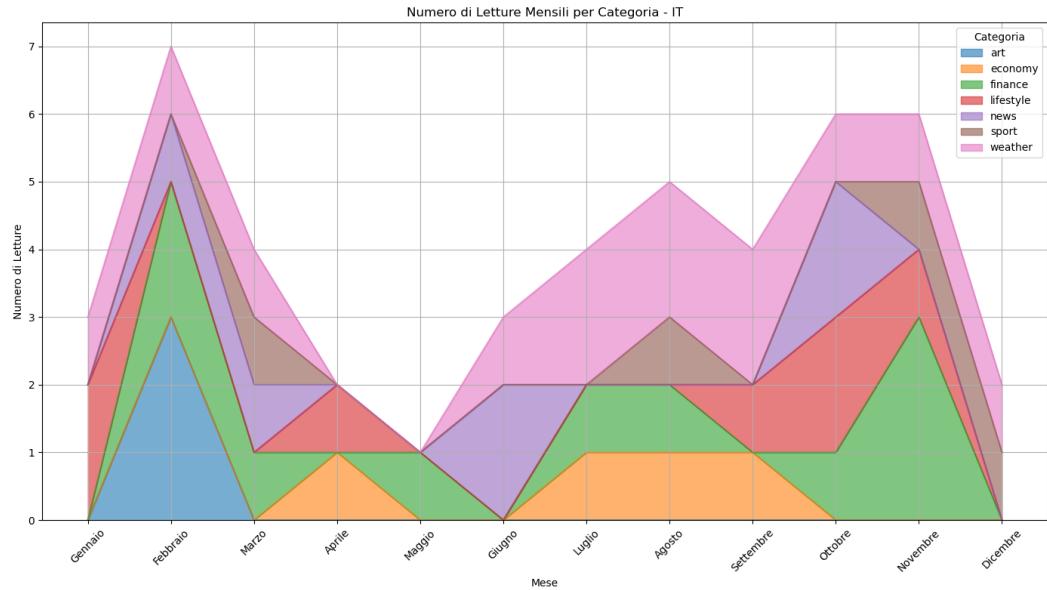
- leggono di **economy** con *cadenza continua massimo mensile*
- sono interessati all'**art** solo a Marzo
- sono più interessati allo **sport**, per quanto si mostri assenza di interesse da maggio ad agosto. Il periodo di crescita dell'interesse coincide con la **fine delle manifestazioni estive**, in cui il *Leitmotiv* è stato la continua lotta con le compagini Italiane in ogni manifestazione sportiva.
- leggono di **finance** principalmente nel periodo estivo
- si informano su **lifestyle** e **weather** sporadicamente



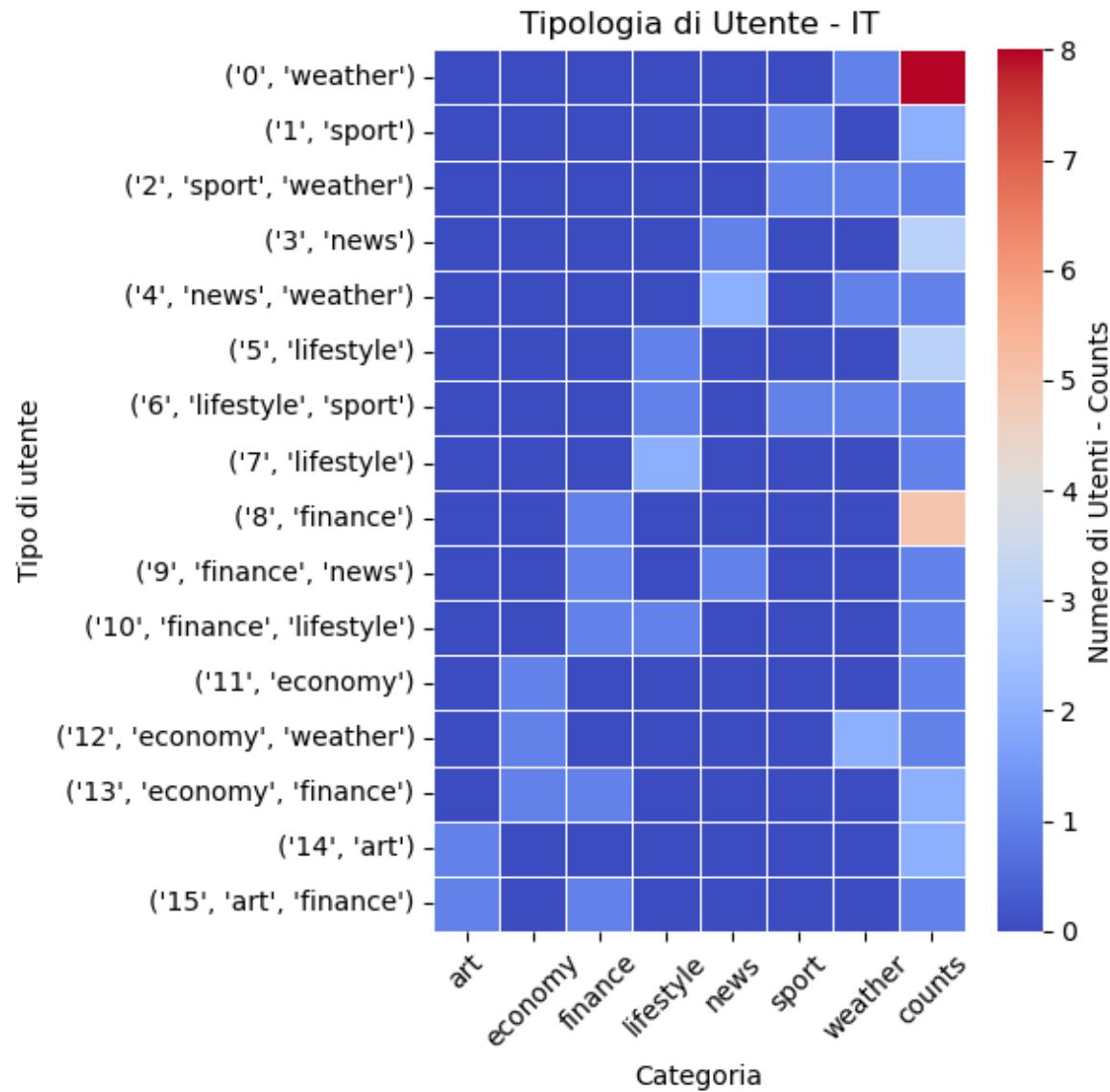
GUSTI PER NAZIONE: TREND TEMPORALI – CASO ITALIANO

Si mostri, come nel caso italiano, l'interesse maggiore sia ascrivibile a tematiche **weather**, che ricopre tassi importanti su tutto l'anno, tranne l'assenza in Aprile - Maggio

- si legge di **finance** con *abbastanza continuità*, salvo qualche assenza in alcuni mesi dell'anno;
- vi è discontinuità per **news**, **sport**, e **lifestyle**
- vi è una cadenza mono-trimestrale nel caso di **economy**
- **art** è relegata al solo *periodo di febbraio*



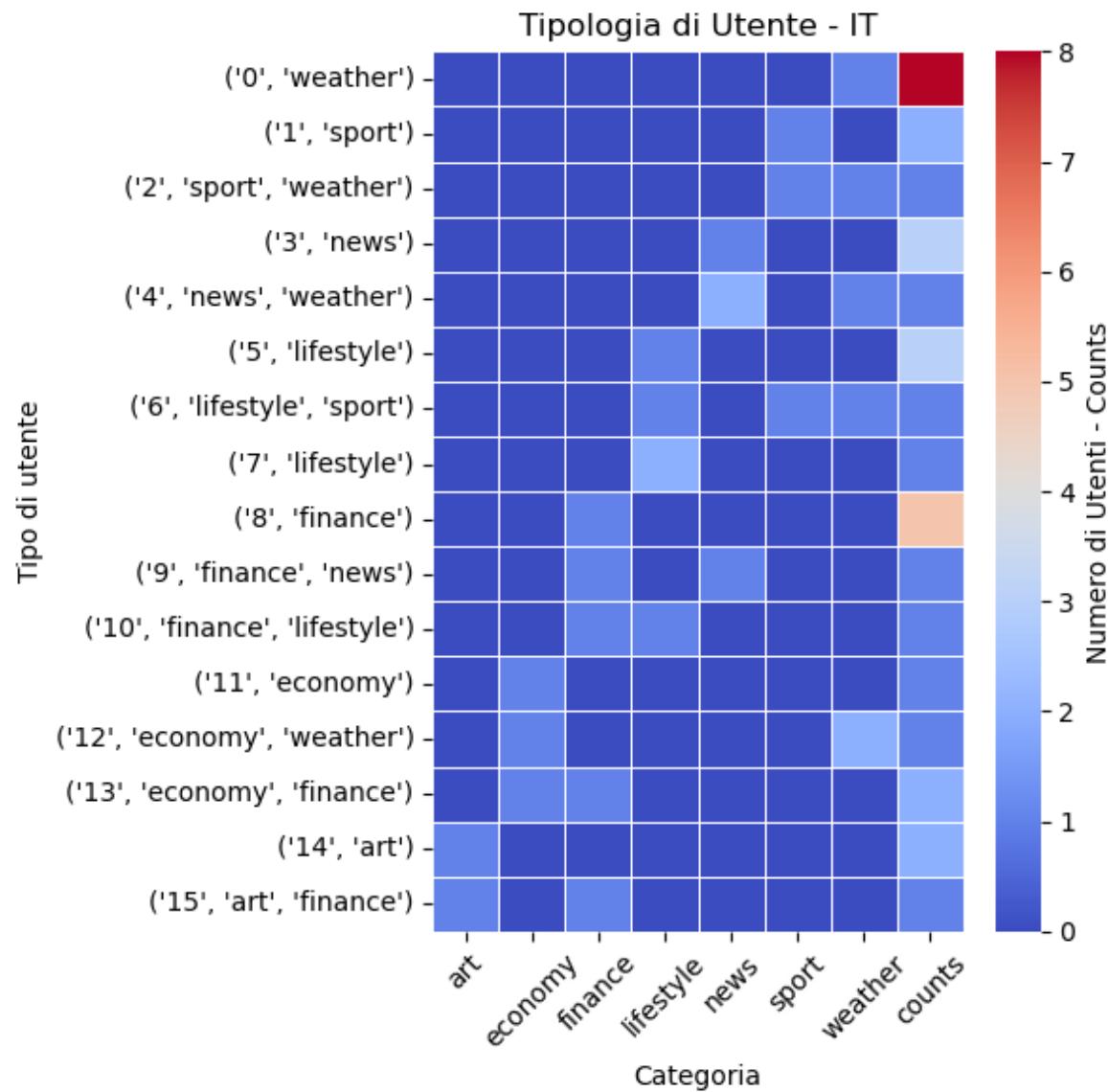
PERSONAS ANALYSIS



TIPOLOGIA DI UTENTE - IT

Al fine di indirizzare l'indagine sotto forma di **strategia** e di **PERSONAS** di utenti tipo, usando TABLEAU, creo questa **Heat map** su Python che evidenzi la frequenza del *numero di utenti* con interessi specifici in termini di **numero di letture per categoria**:

- ci sono principalmente utenti che leggono *solo* di **weather** o **finance** seguiti da chi legge riguardo *solo* **lifestyle** o **news**
- infine vi sono una parte che leggono di **sport**, **art** *separatamente* e **economy** e **finance** *insieme*



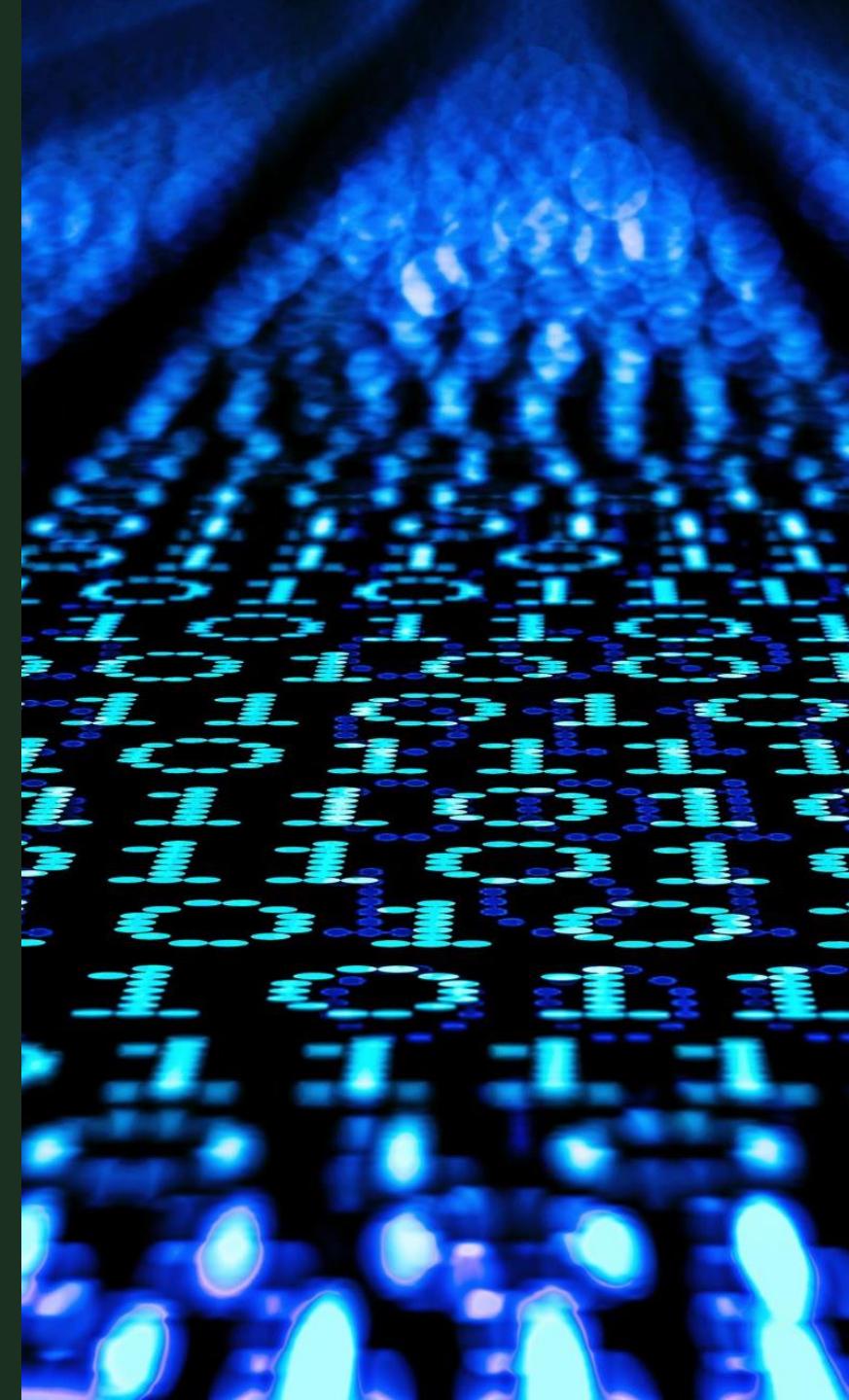
TIPOLOGIA DI UTENTE - IT

Le tipologie di utente ITALIANI qui proposte riflettono quella che è la **distribuzione degli interessi nel dataset al 2021**, considerando le categorie disponibili.

Ma come si è potuto raggiungere questo livello di dettaglio? Mostriamo i ora i passaggi necessari.

TIPOLOGIA DI UTENTE – IT

PYTHON SCRIPT



```

IT = Papers_2021[Papers_2021['country']=='it']
IT

user_cat = IT.groupby(['user_uuid','category']).size().unstack(fill_value=0)
user_cat

```

category	art	economy	finance	lifestyle	news	sport	weather
user_uuid							
19	0	0	0	0	0	0	1
24	0	0	0	1	0	0	0
34	0	0	1	0	0	0	0
39	0	0	0	2	0	0	0
48	0	0	0	0	0	1	1
53	0	0	0	1	0	0	0
57	0	0	0	0	0	1	0
60	0	0	1	0	0	0	0
69	0	0	1	0	0	0	0
76	0	0	1	0	0	0	0
77	0	0	0	0	0	0	1

TIPOLOGIA DI UTENTE - IT

Questa struttura viene ricavata attraverso un

- `groupby(['user_uuid', 'category'])`

valutando la `size()` – numero di letture, per quell'utente, per quella `category`, e strutturando l'estrazione `unstack(fill_value = 0)`.

Avremo così un dataframe *wide*, come una PIVOT, ove le righe sono rappresentate dagli utenti `user_uuid` e le colonne dalle `category`.

N.B. sono inizializzate a “0” le celle per cui non si trova corrispondenza della coppia `user_uuid` - `category`

```

#lista di variabili da usare come riferimento per il raggruppamento
varlist = ['art','economy','finance','lifestyle','news','sport','weather']

counts = user_cat.groupby(varlist).size().reset_index(name='counts')

#estratto i nomi delle colonne 'category' per cui, per ogni riga vale la
#condizione di "non nullità" (lettura = 0) e le salvo come campo, per ogni
#riga. Le userò come parametri per gli indici di riga
new_index = counts.apply(lambda row: row[(row > 0)
                                         & (row.index != 'counts')].index,
                         axis=1)

#converto in un DF la lista di questa estrazione
new_index = pd.DataFrame(new_index.tolist(), index=counts.index).iloc[:, :2]
new_index.columns = ['cat1','cat2']

#conversione a stringa degli indici originali
new_index.index = new_index.index.astype(str)
counts.index = counts.index.astype(str)

#salvo l'indice originale come variabile
counts['Original_Index'] = counts.index
#definisco un metodo che mi permette di valutare,
#attraverso i campi appena creati, quante e quali categorie
#sono lette dal singolo utente. Verranno usate per riclassificare
#L'indice e darne più chiarezza per la Heat Map
#def build_index(row):
#    index = [row['Original_Index']]
#
#    #se il campo delle variabili che contengono le categorie
#    #lette dall'utente è non nullo, allora aggiunge
#    #tale categoria alla tupla, da usare come indice
#
#    if pd.notna(row['cat1']):
#        index.append(row['cat1'])
#    if pd.notna(row['cat2']):
#        index.append(row['cat2'])
#    return tuple(index)
#
#join per indici originali
counts = counts.join(new_index)

#creazione dell'istanza che contiene INDICE + LISTA CATEGORIE LETTE
new_id = counts.apply(build_index, axis=1)

#definizione del nuovo indice del DF PIVOT
counts.index = new_id
#aggiunta alla lista originale del raggruppamento di 'counts'
#per salvarlo come variabile del sotto dataset 'counts'
varlist.append('counts')

counts = counts[list(varlist)]
counts

```

TIPOLOGIA DI UTENTE - IT

A partire dalla prima estrazione, viene mappata la **frequenza** del **numero di utenti** che hanno **interessi simili** e **numero di lettura simili**, attraverso un successivo **groupby(varlist)**.

varlist rappresenta l'**elenco delle categorie**: attraverso questo raggruppamento, possiamo individuare quanti utenti (variabile **counts**) hanno **stessi interessi e con stessa dimensione** (numero di letture per quella categoria)

Dopo il raggruppamento estraggo, attraverso un **apply(lambda)** sull'istanza **counts**, il nome delle colonne, che rappresentano le **category**, per cui quell'utente ha letto ALMENO un articolo [escludendo la variabile '**counts**'] e le salvo, ad una ad una, in un **dataframe new_index** da usare come nuova chiave per gli *indici*.

Converto poi gli indici di **new_index** e **counts** in **strings** (necessario per la successiva costruzione della **Heat map**)

```

#lista di variabili da usare come riferimento per il raggruppamento
varlist = ['art','economy','finance','lifestyle','news','sport','weather']

counts = user_cat.groupby(varlist).size().reset_index(name='counts')

#estratto i nomi delle colonne 'category' per cui, per ogni riga vale la
#condizione di "non nullità" (lettura = 0) e le salvo come campo, per ogni
#riga. Le uso come parametri per gli indici di riga
new_index = counts.apply(lambda row: row[(row > 0)
                                         & (row.index != 'counts')].index,
                          axis=1)

#converto in un DF la lista di questa estrazione
new_index = pd.DataFrame(new_index.tolist(), index=counts.index).iloc[:, :2]
new_index.columns = ['cat1','cat2']

#conversione a stringa degli indici originali
new_index.index = new_index.index.astype(str)
counts.index = counts.index.astype(str)

#salvo l'indice originale come variabile
counts['Original_Index'] = counts.index
#definisco un metodo che mi permette di valutare,
#attraverso i campi appena creati, quante e quali categorie
#sono lette dal singolo utente. Verrano usate per riclassificare
#L'indice e darne più chiarezza per la Heat Map
def build_index(row):
    index = [row['Original_Index']]

    #se il campo delle variabili che contengono le categorie
    #lette dall'utente è non nullo, allora aggiunge
    #tale categoria alla tupla, da usare come indice

    if pd.notna(row['cat1']):
        index.append(row['cat1'])
    if pd.notna(row['cat2']):
        index.append(row['cat2'])
    return tuple(index)

#join per indici originali
counts = counts.join(new_index)

#creazione dell'istanza che contiene INDICE + LISTA CATEGORIE LETTE
new_id = counts.apply(build_index, axis=1)

#definizione del nuovo indice del DF PIVOT
counts.index = new_id
#aggiunta alla lista originale del raggruppamento di 'counts'
#per salvarlo come variabile del sotto dataset 'counts'
varlist.append('counts')

counts = counts[list(varlist)]
counts

```

TIPOLOGIA DI UTENTE - IT

Salvo provvisoriamente gli indici di **counts** come colonna '**Original_Index**' del *dataframe*.

Definisco inoltre un nuovo metodo chiamato

- **build_index(row)**

Lo strumento serve per valutare la **natura** e la **numerosità** delle categorie (i cui campi sono appena stati popolati) lette dal singolo utente. Queste verranno poi salvate in una **tupla** e usate come nuova chiave per gli indici del *dataframe* **counts**

- Creo una istanza **index**, dove salvo **Original_Index**
- Se i campi delle due variabili **cat1** e **cat2** *non sono nulli* vengono aggiunti all'istanza **index**

Dopo il **join** tra **counts** e **new_index** viene creata questa istanza.

```

#lista di variabili da usare come riferimento per il raggruppamento
varlist = ['art','economy','finance','lifestyle','news','sport','weather']

counts = user_cat.groupby(varlist).size().reset_index(name='counts')

#estratto i nomi delle colonne 'category' per cui, per ogni riga vale la
#condizione di "non nullità" (lettura = 0) e le salvo come campo, per ogni
#riga. Le userò come parametri per gli indici di riga
new_index = counts.apply(lambda row: row[(row > 0)
                                         & (row.index != 'counts')].index,
                          axis=1)

#convertendo in un DF la lista di questa estrazione
new_index = pd.DataFrame(new_index.tolist(), index=counts.index).iloc[:, :2]
new_index.columns = ['cat1','cat2']

#conversione a stringa degli indici originali
new_index.index = new_index.index.astype(str)
counts.index = counts.index.astype(str)

#salvo l'indice originale come variabile
counts['Original_Index'] = counts.index
#definisco un metodo che mi permette di valutare,
#attraverso i campi appena creati, quante e quali categorie
#sono lette dal singolo utente. Verranno usate per riclassificare
#L'indice e darne più chiarezza per la Heat Map
def build_index(row):
    index = [row['Original_Index']]

    #se il campo delle variabili che contengono le categorie
    #lette dall'utente è non nullo, allora aggiunge
    #tale categoria alla tupla, da usare come indice

    if pd.notna(row['cat1']):
        index.append(row['cat1'])
    if pd.notna(row['cat2']):
        index.append(row['cat2'])
    return tuple(index)

#join per indici originali
counts = counts.join(new_index)

#creazione dell'istanza che contiene INDICE + LISTA CATEGORIE LETTE
new_id = counts.apply(build_index, axis=1)

#definizione del nuovo indice del DF PIVOT
counts.index = new_id
#aggiunta alla lista originale del raggruppamento di 'counts'
#per salvarlo come variabile del sotto dataset 'counts'
varlist.append('counts')

counts = counts[list(varlist)]
counts

```

TIPOLOGIA DI UTENTE - IT

Dopo di che, la **list** di **tuple** appena generata viene usata come nuovo **indice** nel *dataframe* **PIVOT** **counts**.

Aggiungo, infine, anche la colonna '**counts**' all'istanza **varlist** e estraggo definitivamente le variabili utili alla costruzione della **Heat Map**.

	art	economy	finance	lifestyle	news	sport	weather	counts
(0, weather)	0	0	0	0	0	0	1	8
(1, sport)	0	0	0	0	0	1	0	2
(2, sport, weather)	0	0	0	0	0	1	1	1
(3, news)	0	0	0	0	1	0	0	3
(4, news, weather)	0	0	0	0	2	0	1	1
(5, lifestyle)	0	0	0	1	0	0	0	3
(6, lifestyle, sport)	0	0	0	1	0	1	1	1
(7, lifestyle)	0	0	0	2	0	0	0	1
(8, finance)	0	0	1	0	0	0	0	5
(9, finance, news)	0	0	1	0	1	0	0	1
(10, finance, lifestyle)	0	0	1	1	0	0	0	1
(11, economy)	0	1	0	0	0	0	0	1
(12, economy, weather)	0	1	0	0	0	0	2	1
(13, economy, finance)	0	1	1	0	0	0	0	2
(14, art)	1	0	0	0	0	0	0	2

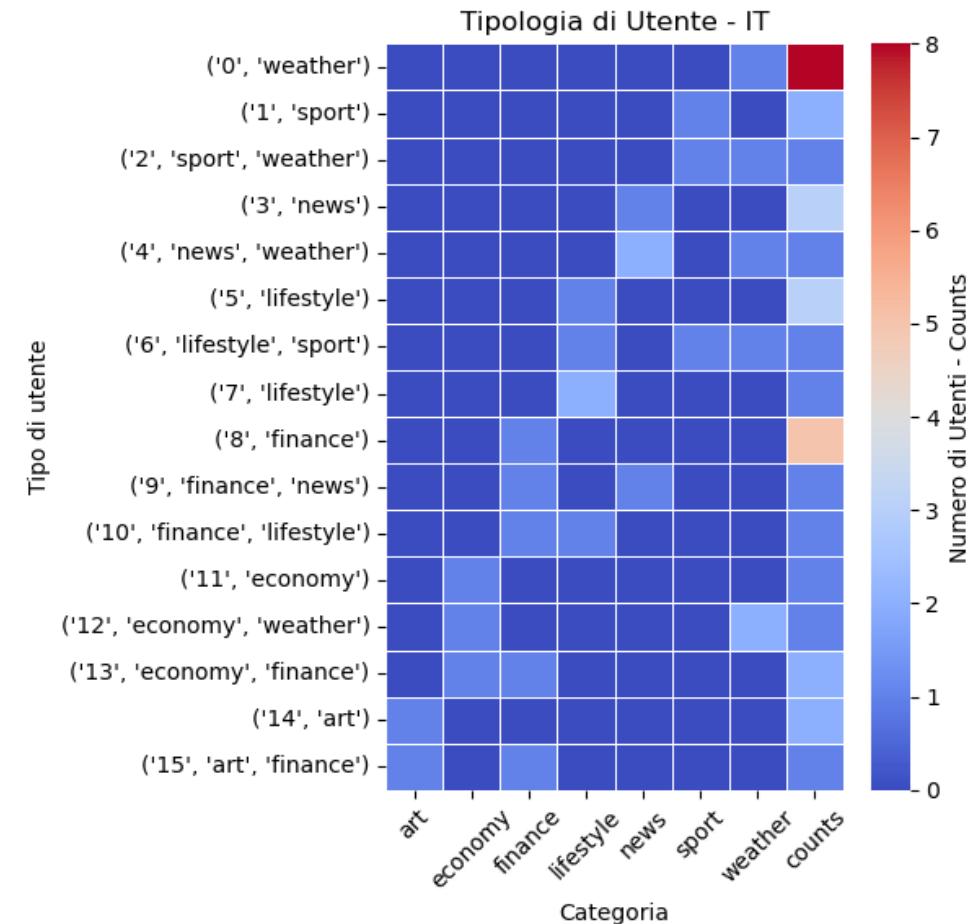
TIPOLOGIA DI UTENTE - IT

Questa è la tabella definitiva su cui si basa la **Heat Map** della *Tipologia di Utente – IT*.

Di fatto, vengono contati il numero di utenti con **stessa sequenza numerica** [art, economy, finance, lifestyle, news, sport, weather], dove si possono assumere valori ≥ 0

se $n_{i,j} = 0$: assenza del carattere (lettura per la categoria)

	art	economy	finance	lifestyle	news	sport	weather	counts
(0, weather)	0	0	0	0	0	0	1	8
(1, sport)	0	0	0	0	0	1	0	2
(2, sport, weather)	0	0	0	0	0	1	1	1
(3, news)	0	0	0	0	1	0	0	3
(4, news, weather)	0	0	0	0	2	0	1	1
(5, lifestyle)	0	0	0	1	0	0	0	3
(6, lifestyle, sport)	0	0	0	1	0	1	1	1
(7, lifestyle)	0	0	0	2	0	0	0	1
(8, finance)	0	0	1	0	0	0	0	5
(9, finance, news)	0	0	1	0	1	0	0	1
(10, finance, lifestyle)	0	0	1	1	0	0	0	1
(11, economy)	0	1	0	0	0	0	0	1
(12, economy, weather)	0	1	0	0	0	0	2	1
(13, economy, finance)	0	1	1	0	0	0	0	2
(14, art)	1	0	0	0	0	0	0	2



```
merge_IT = IT.merge(user_cat, how='inner', on='user_uuid')
merge_IT
```

journalist_id	language	length	country	subscription_date	platform	article_id	stars	read_month	art	economy	finance	lifestyle	news	sport	weather
117	it	short	it	2020-08-24	tablet	5128	3	February	1	0	0	0	0	0	0
111	it	long	it	2020-02-12	tablet	732766	5	July	0	1	0	0	0	0	2
110	it	short	it	2020-02-12	pc	212930	5	April	0	1	0	0	0	0	2
116	it	short	it	2020-02-12	pc	712540	1	October	0	1	0	0	0	0	2
111	it	short	it	2020-04-06	pc	612403	3	November	0	0	1	1	0	0	0
116	it	long	it	2020-04-06	mobile	132857	1	October	0	0	1	1	0	0	0

```
merge_IT = merge_IT.merge(counts, how='inner', on=['art', 'economy', 'finance', 'lifestyle', 'news', 'sport', 'weather'])
merge_IT
```

journalist_id	language	length	country	subscription_date	platform	article_id	stars	read_month	art	economy	finance	lifestyle	news	sport	weather	counts
117	it	short	it	2020-08-24	tablet	5128	3	February	1	0	0	0	0	0	0	2
112	it	long	it	2020-01-09	tablet	532353	4	February	1	0	0	0	0	0	0	2
111	it	long	it	2020-02-12	tablet	732766	5	July	0	1	0	0	0	0	2	1
110	it	short	it	2020-02-12	pc	212930	5	April	0	1	0	0	0	0	2	1
116	it	short	it	2020-02-12	pc	712540	1	October	0	1	0	0	0	0	2	1
111	it	short	it	2020-04-06	pc	612403	3	November	0	0	1	1	0	0	0	1

TIPOLOGIA DI UTENTE - IT

La doppia procedura di *merge* permette poi di associare la sequenza degli interessi e il conteggio del numero di utenti, relativa al *singolo lettore*, attraverso **user_uuid**, nel primo caso, e attraverso la sequenza stessa [**'varlist'** originale], nel secondo:

- In questo modo si può descrivere **quali potrebbero essere le preferenze *inerenti variabili individuali*** per gli utenti che hanno un indirizzo in termini di **preferenze di categoria** [es. I lettori di solo **art** che preferenza abbiano in termini di lunghezza degli articoli **length**].

Questa procedura è necessaria al fine di condurre un primo approccio di **PERSONAS ANALYSIS**.



PERSONAS ANALYSIS

PERSONAS ANALYSIS - PASSAGGI PREPARATORI

Per rendere fattibile tale studio sono state mappate, con *un doppio merge*, le osservazioni per cui vale la condizione di

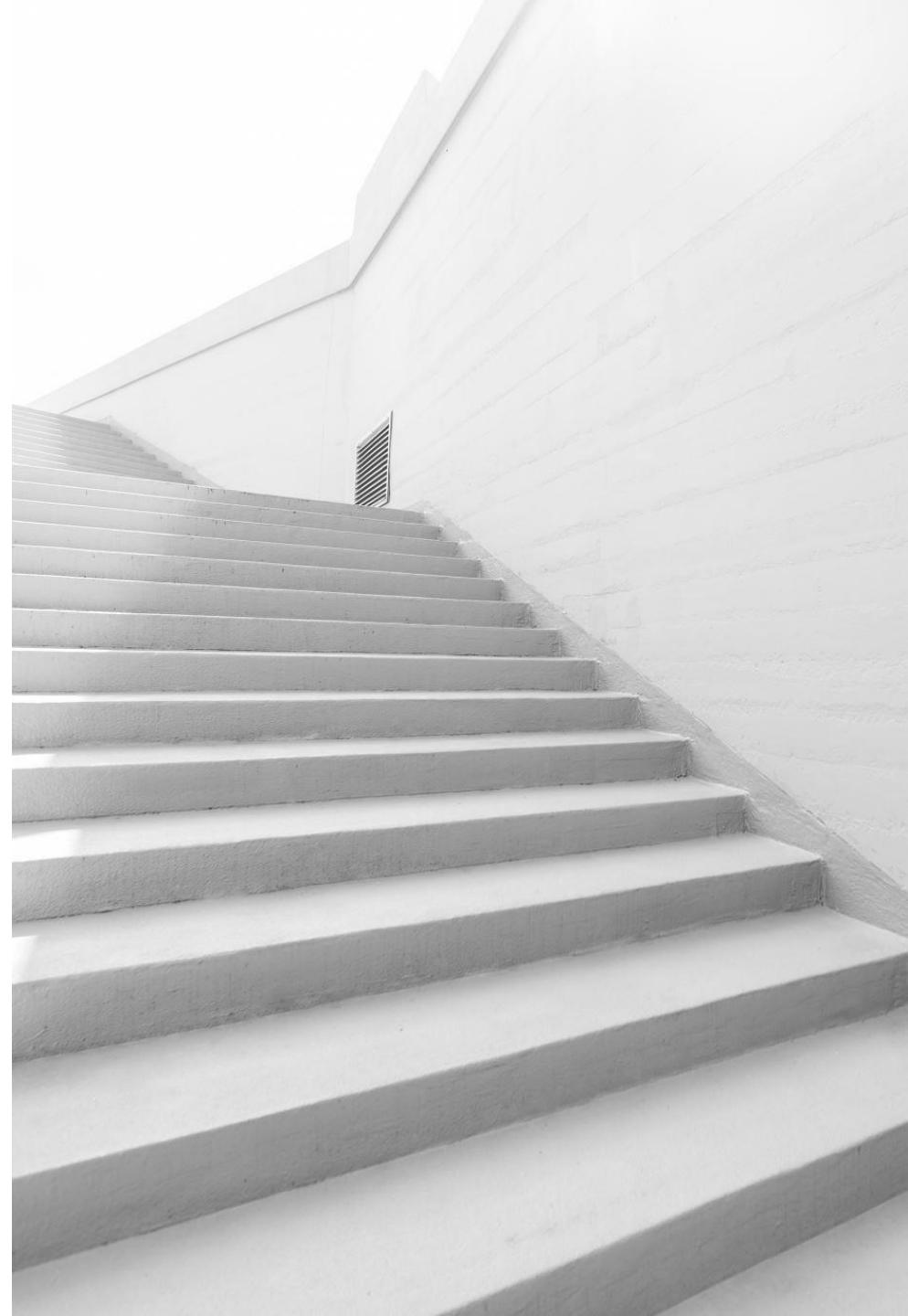
- *numerosità* (l'utente fa parte del gruppo di *counts elevati*) [usando come chiave `user_uuid`]
- *unicità* (il gruppo legge solo articoli di certe categorie) [usando come chiave la combinazione sequenziale es. `[0,0,0,1,0,0,2]` delle variabili conteggio delle *category*, che valuta *se un utente ha letto 0 o più articoli di una categoria.*]

Fatto questo, viene creato un `method`

- `def define_personas(row)`

che si occupa di mappare la *tipologia di utenti italiani* in base a queste *singolarità (solo una categoria o più di una)* *CREANDO LE PERSONAS*

Viene poi estratto il dataset risultante per definire le *caratteristiche individuali tipo* di chi rientra in questo gruppo di utenti, *da analizzare con Tableau.*





PERSONAS ANALYSIS

PYTHON SCRIPT

PERSONAS – SCRIPT

Come anticipato nella slide precedente, si è definita una **funzione – method** che ha l’obiettivo di mappare le *tipologie di utenti italiani*.

Lo script a fianco mostra come **define_personas()** riceva come parametro il *dataframe* in cui sono contenute le informazioni sui singoli utenti, associati alla **sequenza che riflette i loro interessi** in termini di lettura – che chiameremo, d’ora in poi, **list** – e il **conteggio del numero di utenti** che presentano la **STESSA SEQUENZA**.

Viene definita una nuova colonna – **persona** – che ha l’obiettivo di immagazzinare una **string** che identifichi a che *tipologia di persona il singolo utente sia riconducibile*, sulla base della sequenza di **list**.

Si accede alle variabili inerenti i **conteggi di lettura (values)**, per **categoria**, attraverso la **list** precedentemente inizializzata

Convertiti gli elenchi in liste, viene poi calcolato, per ogni utente,

- Il numero di zeri num_zeros
- Il numero di non zeri num_non_zeros, come differenza tra la lunghezza della sequenza e il numero di zeri (len(values) - num_zeros)

```
# Definizione delle personas
def define_personas(row):

    #inizializziamo il campo personas
    row['persona'] = ''

    #lista delle variabili di categoria da controllare sotto forma di COMBINAZIONE SEQUENZIALE
    list=['art', 'economy', 'finance', 'lifestyle', 'news', 'sport', 'weather']

    #conversione dell'elenco a Lista
    values = row[list].tolist()

    # Conta il numero di zeri e non zeri
    num_zeros = values.count(0)
    num_non_zeros = len(values) - num_zeros

    #Sapendo che La sequenza deve contenere 5 o 6 ZERI, viene fatto fatto il controllo
    #ANCHE che L'unico valore diverso da 0 possa essere MAGGIORE (in quanto vi sono TIPI
    #DI UTENTE che Leggono più di un articolo per categoria)

    ##controllo della correttezza della sequenza e della nazionalità
    ##E poi successiva classificazione della PERSONAS

    #sport
    if (num_zeros == 6 and values[5] > 0):
        row['persona'] = 'Sports Enthusiast - Gimbo Tamperi'

    #weather
    elif (num_zeros == 6 and values[6] > 0):
        row['persona'] = 'Weather Guru - Mario Giuliacci'

    #lifestyle
    elif (num_zeros == 6 and values[3] > 0):
        row['persona'] = 'Lifestyle Addicted - Chiara Ferragni'

    #news
    elif (num_zeros == 6 and values[4] > 0):
        row['persona'] = 'News Reader - Enrico Mentana'

    #finance
    elif (num_zeros == 6 and values[2] > 0):
        row['persona'] = 'Finance Professional - Mario Draghi'

    #art
    elif (num_zeros == 6 and values[0] > 0):
        row['persona'] = 'Art Expert - Alessandro Orlando'

    #finance and economy
    elif (num_zeros == 5 and values[1] > 0 and values[2] > 0):
        row['persona'] = 'Economy and Finance Researcher - Prof. Filippo Cossetti'

    else:
        row['persona'] = 'Casual Reader - Chiara'

return row
```

PERSONAS – SCRIPT

In base gli output generati dalla **Heat Map**, viene fatta una classificazione delle personas riferendoci alla **rilevanza del numero di utenti** che hanno una specifica combinazione di zeri e non zeri (specifici interessi) [es. solo **weather**, la più rilevante, era comune a ben 8 utenti]

- Sulla base di ciò, con multipli **if - elif**, si verifica che nella sequenza vi siano un numero di 5 o 6 zeri (valutando il conteggio **num_zeros**) **num_zeros == 6** o **num_zeros == 5** e quale sia/siano il/i valore/i, nella **sequenza – lista** dell'utente, maggiori/i di zero [in base al fatto che ci siano **interessi singoli** o **multipli**].
- Sapendo quale sia la disposizione della lista di variabili estratte [**category** in ordine alfabetico] verifichiamo quindi la condizione **values[i] > 0** sfruttando l'indice posizionale (es. $i = 0 \rightarrow art$; $i = 3 \rightarrow lifestyle$...). N.B. in caso di **interessi multipli** si valuta la validità concorrente delle multiple condizioni (es. **finance** and **economy**)

In base a quale sia il caso rispettato, viene inizializzata la stringa relativa, contenuta in **persona**, di conseguenza.

In casi non rientranti tra quelli di spicco (numero elevato di utenti con stessa combinazione) l'utente è classificato come «**Casual Reader - Chiara**»

Dopo di che il metodo ritorna il *dataframe*, aggiornandolo o creando una nuova istanza.

Es. di chiamata del metodo: **Papers_2021_pers_IT = merge_IT.apply(define_personas, axis = 1)**

```
# Definizione delle personas
def define_personas(row):

    #inizializziamo il campo personas
    row['persona']=''

    #lista delle variabili di categoria da controllare sotto forma di COMBINAZIONE SEQUENZIALE
    list=['art', 'economy', 'finance', 'lifestyle', 'news', 'sport', 'weather']

    #conversione dell'elenco a Lista
    values = row[list].tolist()

    # Conta il numero di zeri e non zeri
    num_zeros = values.count(0)
    num_non_zeros = len(values) - num_zeros

    #Sapendo che La sequenza deve contenere 5 o 6 ZERI, viene fatto fatto il controllo
    #ANCHE che L'unico valore diverso da 0 possa essere MAGGIORE (in quanto vi sono TIPI
    #DI UTENTE che Leggono più di un articolo per categoria)

    ##controllo della correttezza della sequenza e della nazionalità
    ##E poi successiva classificazione della PERSONAS

    #sport
    if (num_zeros == 6 and values[5] > 0):
        row['persona'] = 'Sports Enthusiast - Gimbo Tamperi'

    #weather
    elif (num_zeros == 6 and values[6] > 0):
        row['persona'] = 'Weather Guru - Mario Giuliacchi'

    #lifestyle
    elif (num_zeros == 6 and values[3] > 0):
        row['persona'] = 'Lifestyle Addicted - Chiara Ferragni'

    #news
    elif (num_zeros == 6 and values[4] > 0):
        row['persona'] = 'News Reader - Enrico Mentana'

    #finance
    elif (num_zeros == 6 and values[2] > 0):
        row['persona'] = 'Finance Professional - Mario Draghi'

    #art
    elif (num_zeros == 6 and values[0] > 0):
        row['persona'] = 'Art Expert - Alessandro Orlando'

    #finance and economy
    elif (num_zeros == 5 and values[1] > 0 and values[2] > 0):
        row['persona'] = 'Economy and Finance Researcher - Prof. Filippo Cossetti'

    else:
        row['persona'] = 'Casual Reader - Chiara'

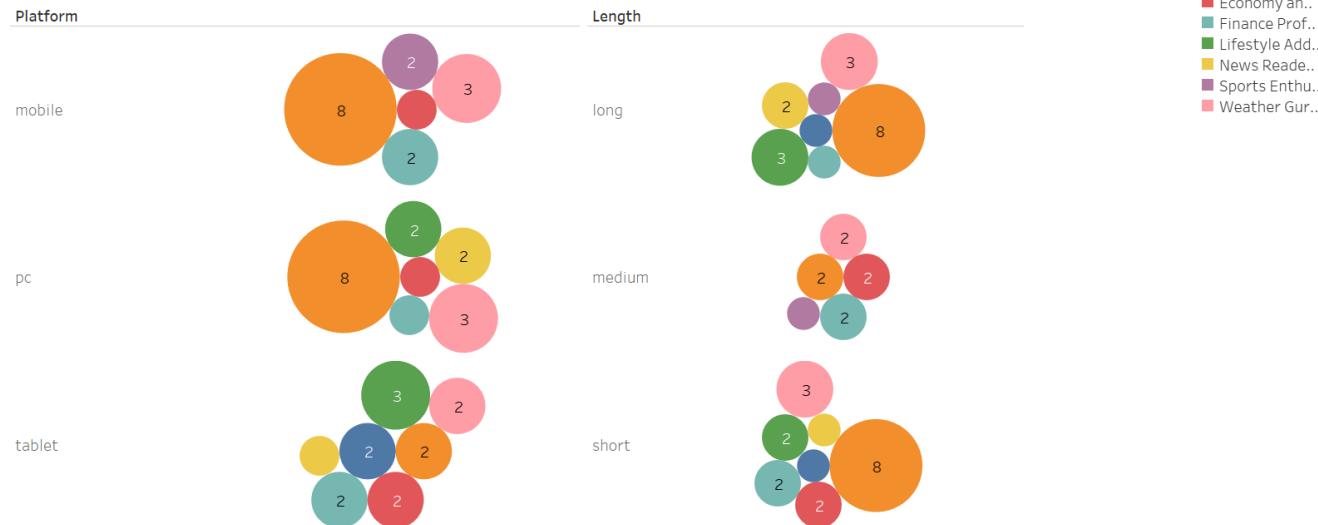
    return row
```



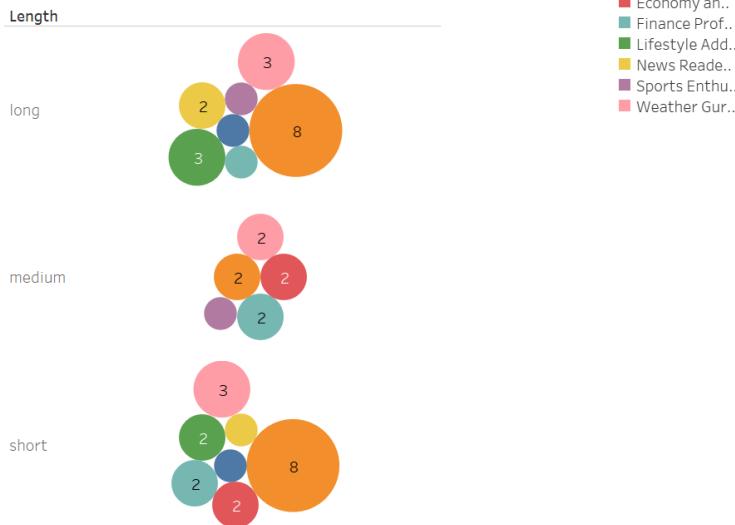
PERSONAS ANALYSIS

VALUTAZIONI
QUANTITATIVE, *INSIGHTS* E
CONSIDERAZIONI

Personas - Distribuzione Letture per Platform -2021



Personas - Distribuzione Letture per Lunghezza Articolo - 2021



ANALISI PERSONAS – LETTURE PER PLATFORM E LUNGHEZZA ARTICOLO

- Ciò che possiamo dedurre da questa classificazione è che, nel complesso, - escludendo gli *utenti casuali* -, coloro che leggono **esclusivamente weather** (**Weather Guru - Mario Giuliacci**) si dimostrano grandi lettori, accaniti esclusivamente su **tematiche climatiche** e che sono **equamente distribuiti** in termini di **preferenze di platform** (solo in **tablet** i **Lifestyle Addicted - Chiara Ferragni** hanno la meglio)
- Si dimostra altrettanta equità nelle distribuzioni anche in termini di **lunghezza** degli articoli **length**, dove queste due categorie spiccano tra le categorie **long** e **short** (nel caso di **short** solo **weather**)

Personas - Numero di Letture per Valutazione Media - 2021

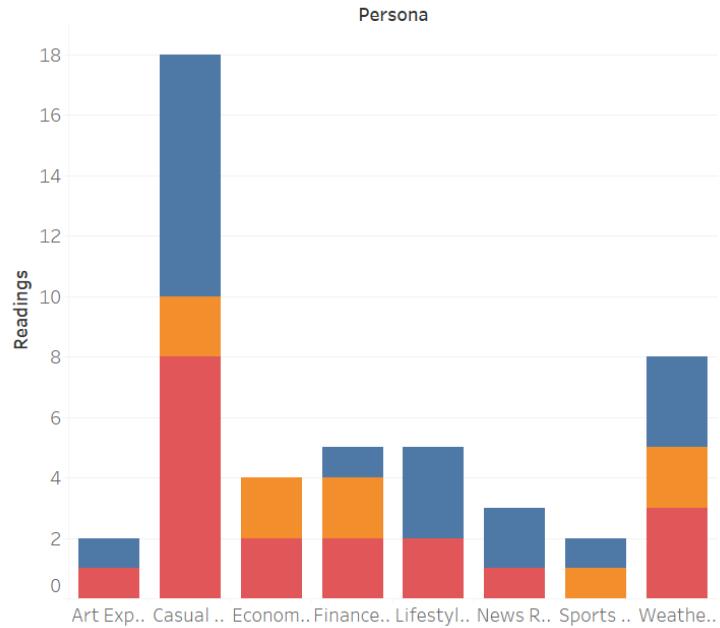


ANALISI PERSONAS – LETTURE PER VALUTAZIONE MEDIA

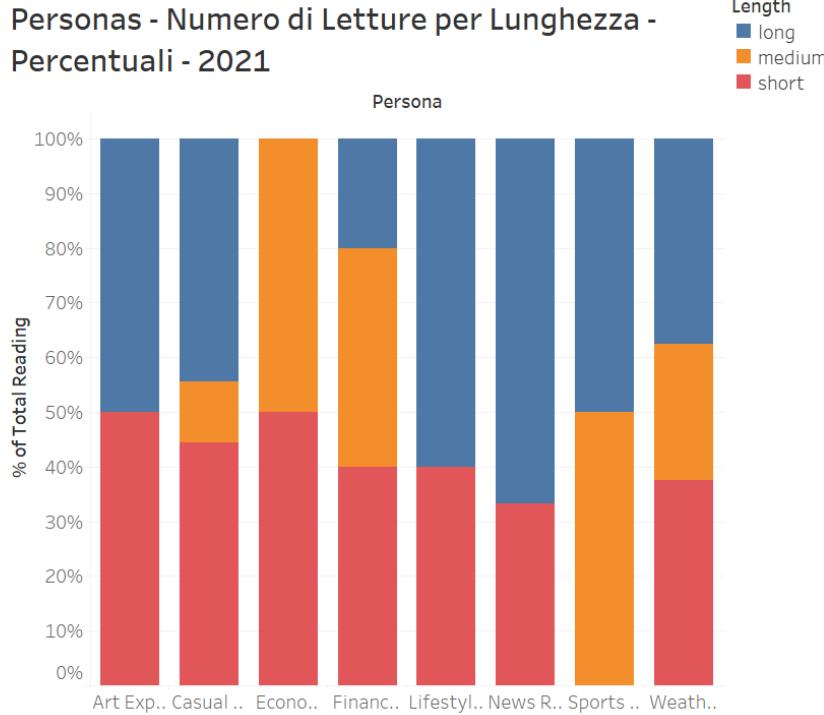
- Come si evince dal plot, per gli utenti che leggono solo di specifiche categorie, quelle più pesanti nel computo totale, (**weather e lifestyle**), si presenta una distribuzione *uniforme delle valutazioni*, manifestando ancora la tendenza a non avere una robusta qualità degli articoli offerti.
- Per **Sport Enthusiast - Gimbo Tamberi** questa stima è *al ribasso* (valutazioni scarse)
- Per i **News Reader - Enrico Mentana** la valutazione è addirittura *polarizzata* tra gli estremi *minimi* e *massimi* (anche per chi appartiene alla categoria **Economy and Finance Researcher - Prof. Filippo Cossetti**)

ANALISI PERSONAS – LETTURE PER LUNGHEZZA

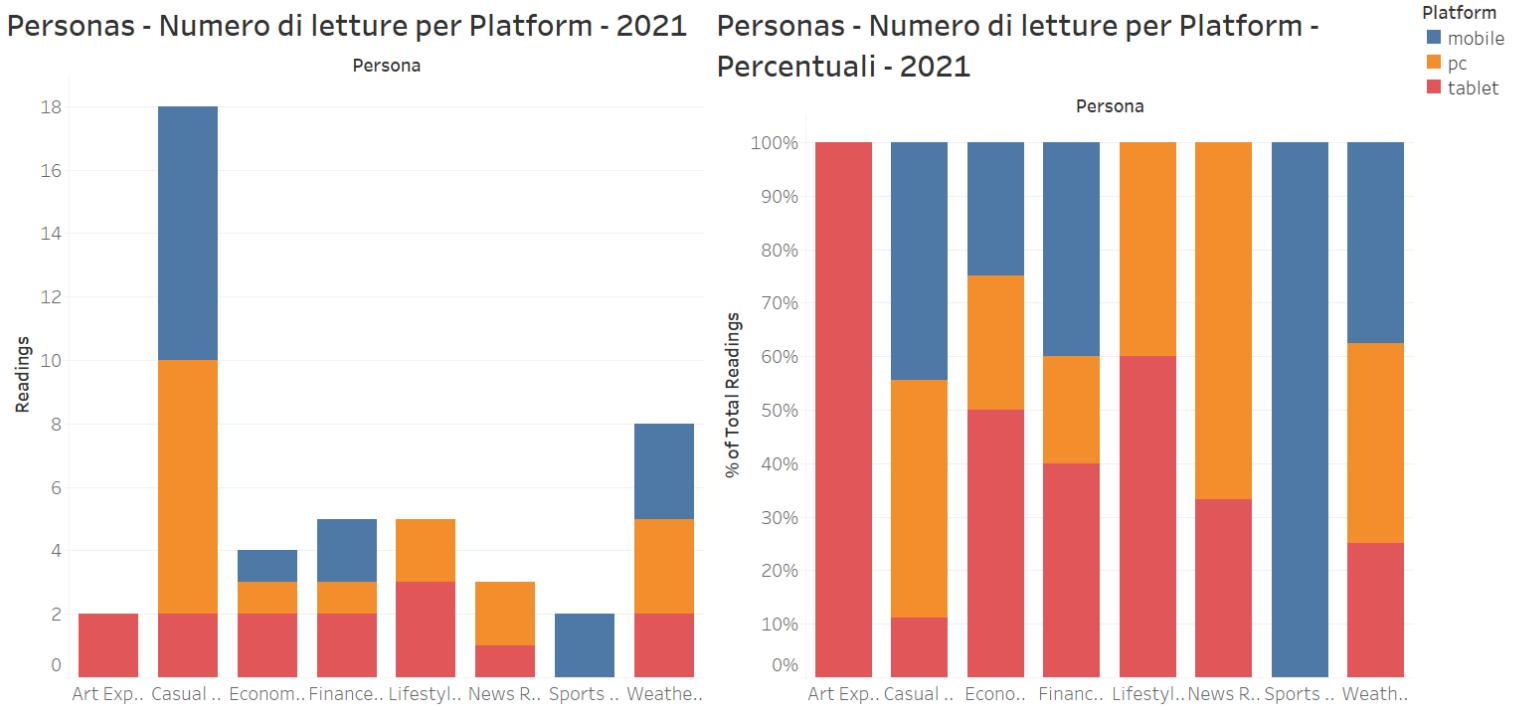
Personas - Numero di Letture per Lunghezza -
2021



Personas - Numero di Letture per Lunghezza -
Percentuali - 2021



- Nel complesso, sempre escludendo la question **Casual Reader - Chiara**, si stima come i lettori accaniti di solo **weather** siano i più rappresentativi della popolazione, seguiti dai lettori di sola **finance** o **lifestyle**.
- Valutando la distribuzione in termini **percentuali**, si presenta un fenomeno di **polarizzazione** della scelta tra **long** e **short**, con qualche presenza di lunghezza **medium**.
- Solo **Sports Enthusiast** ha una convergenza verso testi **medio-lunghi** (indice di come ci si voglia indirizzare su articoli che illustrino il fenomeno descritto nel dettaglio), mentre **Economy and Finance Researcher** predilige testi **medio-corti** (orientandosi all'immediatezza del contenuto, arrivando al nodo della questione)



ANALISI PERSONAS – LETTURE PER PLATFORM

Stesso discorso è valevole, escludendo **Casual Reader**, monitorando la distribuzione per **Platform**:

- A livello percentuale, per questioni “grafiche”, gli **Art Expert** prediligono la lettura da **tablet**
- Gli **Sports Enthusiast** sono orientati alla lettura da **mobile**
- **Lifestyle Addicted** e **News Reader** sono orientati tra **pc** e **tablet** (in un rapporto 40-60 e ~70-30 rispettivamente)
- **Economy Researcher**, **Weather Guru**, **Finance Professional** si distribuiscono, con diversi pesi, tra le tre tipologie di piattaforme disponibili.

CONSIDERAZIONI FINALI

CONSIDERAZIONI FINALI: PROPOSTA STRATEGICA

Tenuto conto dell'analisi appena svolta, gli *hint* che possiamo assumere sono molteplici:

- Sapendo che nel complesso, sia a livello aggregato, sia a livello solo *italiano*, la valutazione media è polarizzata agli estremi, non robusta o **mediamente** non sufficiente, questo ci dà ispirazione su come articoli a stampo **tecnologico-scientifico** possano rivelarsi strategici per accogliere utenti da diverse *sfere di interesse*.
- L'applicazione scientifico-tecnologica è *multi-tentacolare*: può estendere la propria rete a molteplici ambiti, in quanto ormai *tutto può essere filtrato* con un taglio **empirico-sperimentale-tech**.
- Data la potenziale insoddisfazione degli utenti, considerando queste valutazioni mutevoli, e questa presenza forsennata del **tech** in ogni ambito, possiamo fornire ai nuovi arrivati articoli «*introduttivi*» al mondo tech, con linguaggio largamente condivisibile e semplice, in cui vengono forniti i pilastri del metodo scientifico – tech, come svolgere ricerca in questi ambiti, e come questi siano inevitabilmente *connessi* con le loro sfere di interesse (**economy, finance, weather...**)

CONSIDERAZIONI FINALI: PROPOSTA STRATEGICA

- Riuscendo a **penetrare** facilmente nel **mercato**, possiamo compensare tale insoddisfazione, traendone vantaggio, ma *facendo il bene della collettività*:
 - Fornendo articoli di qualità, migliorando la valutazione media, riusciremo a garantire più **fedeltà** dei nostri lettori *nel tempo*
 - Introducendo i lettori alla *lettura GUIDATA non in lingua natia* (es. lettori italiani verso l'inglese [i paper di ricerca accademica sono prevalentemente redatti inglese]) si stimola una **crescita del livello di educazione media** che porterà benefici a cascata sul quotidiano; i nostri, come gli altri, **autori** saranno altresì stimolati a proporre **articoli sempre più di qualità**, per spostare le quote di mercato. **È UN BENEFICIO PER ENTRAMBE LE PARTI.**
- Sapendo che, mediamente, vi sia questa forma di **polarismo** in termini di **length**, possiamo provare ad *indirizzare l'interesse* verso letture più approfondite, ma **PER GRADI**. L'obiettivo è **stimolare l'individuo** ad una **lettura più immersiva** nuovamente, provando a ribaltare il fenomeno del *livello di attenzione da smartphone*, su cui si basano i contenuti dei social media (lunghezza media ~ 15 secondi). **ALTRO BENEFICO A FAVORE DELLA COLLETTIVITÀ.**

Trattandosi di realtà multipiattaforma, possiamo orientarci verso una **proposta customizzata**, come contenuti, in termini di visualizzazione grafica, in modo da poter *strizzare l'occhio* ad una platea di pubblico sempre più ampia.

Ricordiamoci: la Scienza (e il Tech) è Progresso. Individuale e, soprattutto, Collettivo.