Progetto finale di Data Analysis di Sebastiano Iaci

- ## Il dataset

In questo progetto ho analizzato il dataset 'information_data' in formato .csv utilizzando tre metodi differenti : **SQL** , **Python** e **Tableau**.

```
information_data.head()
```

	read_date	user_uuid	category	journalist_id	language	length	country	subscription_date	platform	article_id	stars
0	19-2-2022	190	sport	117	fr	long	fr	28-6-2021	tablet	331708	2
1	25-2-2021	243	art	117	it	short	it	24-8-2020	tablet	5128	3
2	19-12-2023	36	weather	115	en	long	uk	7-2-2021	tablet	733403	5
3	27-3-2023	162	finance	111	it	short	it	6-9-2022	tablet	612638	1
4	22-10-2023	181	economy	119	fr	short	fr	9-12-2020	tablet	211458	3

Questo è un estratto del dataframe in pandas e contiene dei dati su delle riviste digitali in cui sono elencati rispettivamente:

- La data di lettura dell'articolo
- L'Id che identifica il lettore
- La tipologia dell'articolo
- L'Id che identifica lo scrittore dell'articolo
- La lingua
- La lunghezza (o formato)
- Il paese
- La data di iscrizione del lettore nel sito di lettura
- La piattaforma su cui è stato letto
- L'Id dell'articolo
- Le valutazioni date dal lettore che vanno da un punteggio minimo di 1 ad un massimo di 5

Qui il dataset completo :

https://drive.google.com/file/d/1JlbeI2bPZWMU6zTjCaZ9nxklaCJU-TvW/view?usp=sharing

- # Lo scopo del progetto e le informazioni che ho estratto

Lo scopo del progetto simulato è quello di individuare le tendenze del comportamento degli utenti in merito agli articoli analizzando quindi metriche che mettono in relazione la tipologia di articolo con il comportamento dei lettori in modo da proporre una strategia che ottimizzi un prodotto che possa funzionare in base alle tendenze e ai comportamenti visualizzati. Il prodotto in questione sarebbe riferito ad una testata online fittizia di divulgazione scientifica che miri ad eliminare il divario tra il mondo scientifico e pubblico.

Essendo che alcune di queste informazioni sono ridondanti mi limito a riportarle per intero senza suddividerle dalla fonte di provenienza e indicherò in modo sommario i punti chiave, per gli approfondimenti si possono vedere i vari link.

- ## Articoli e lettori:

Il dataset conta **242** lettori unici. La categoria più seguita è il meteo , quella di meno l'arte.

Meteo: **180** lettori (74%) - incremento di letture nel 2022

Sport: **132** lettori (55%) - incremento di letture nel 2022

Finance: **102** lettori (42%) - incremento di letture nel 2022

News: **87** lettori (36%) - incremento di letture nel 2023

Economy: **87** lettori (36%) - incremento di letture nel 2022

Lifestyle: 66 lettori (25%) - incremento di letture nel 2022

Art: **54** lettori (22%) - incremento di letture nel 2022

- **Categorie col maggior numero di punti massimi (5) :**

weather: 65 ; news: 26 ; economy:24 ; finance: 22 ; sport: 21 ; lifestyle: 20 ; art:10

- **Categorie col maggior numero di punti minimi (1):**

weather: 70 ; sport: 38 ; finance:32 ; economy:26 ; news:19 ; lifestyle: 12 ; art:9

- **Massimo conteggio di voti dati dai lettori per categoria suddivisi per mesi e anni :**

Art : **24** (Ottobre 2023) Economy: **35** (Luglio 2023) Finance: **34** (Luglio 2023) Lifestyle: **17** (Luglio 2022)
News: **35** (Dicembre 2023) Sport: **48** (Settebre 2023) Weather: **76** (Luglio 2023)

- **Numero di lettori (Id univoci) che leggono soltanto una categoria:**

13 - Weather

1 - Economy

3 - News

3 - Art

3 - Finance

3 - Liferstyle

4 - Sport

- **ID lettori con la maggior frequenza di lettura per categoria (% sul totale) :**

ID 163 - art (34,72%)

ID 106 - economy (35,40%)

ID 170 - finance (11,98%)

ID 108 - lifestyle (63,55%)

ID 106 - news (43,15%)

ID 155 - sport (14,98%)

ID 88 - weather (9,94%)

- Dall'anno 2021 al 2024 , l'intervallo di anni che ricopre il periodo temporale del dataset, le categorie che hanno avuto un incremento del punteggio medio sono state : **finance** (60,49%) , **sport** (48,44%) e **lifestyle** (16,07%)

- Il formato preferito di articolo è quello **lungo** , la piattaforma digitale più usata in tutto è il **tablet** (per i dettagli di preferenza per paese e per anno vedere su Tableau) .

- **L’analisi**

Utilizzando **SQL** , **Python** e **Tableau** ho avuto modo di analizzare il dataset a **360 gradi.** Da questa analisi approfondita ho ricavato alcune informazioni ridondanti sotto una forma diversa ma è stato integrando grazie alla visualizzazione grafica dei dati con le librerie **matplotlib** , **seaborn** e il software **Tableau** che ho potuto vedere nella totalità la relazione profonda tra i dati che erano di mio interesse.

Per cominciare ho usato SQL dove mi sono dedicato ad una prima ‘scrematura’ delle informazioni e analizzandole sotto un aspetto più generale.

```sql
-- In questa query ho selezionato le categorie con il maggior numero numero di punti massimi (5) contandoli e ordinandoli in ordine discendente
SELECT category, COUNT(*) AS max_star_count
FROM information_data
WHERE stars = (SELECT MAX(stars) FROM information_data)
GROUP BY category
ORDER BY max_star_count DESC
-- Risultato= weather: 65 ; news: 26 ; economy:24 ; finance: 22 ; sport: 21 ; lifestyle: 20 ; art:10


--Qui ho fatto lo stesso ma selezionando per punteggi minimi e contandoli
SELECT category, COUNT(*) AS min_star_count
FROM information_data
WHERE stars = (SELECT MIN(stars) FROM information_data)
GROUP BY category
ORDER BY max_star_count DESC
-- Risultato= weather: 70 ; sport: 38 ; finance:32 ; economy:26  ; news:19 ; lifestyle: 12 ; art:9
```

Qui il link:

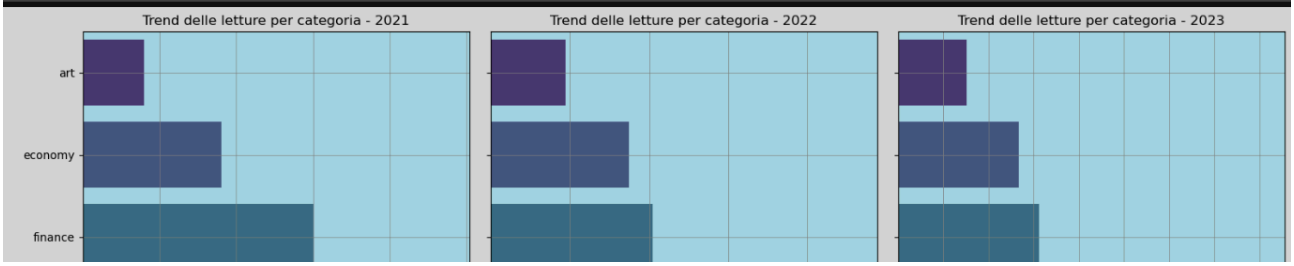
https://drive.google.com/file/d/1YOhJe2si05GEwaRNwHk4Ru8IIDSfJybg/view?usp=sharing

In aggiunta questo è il dataset modificato con la query di SQL in cui ho estratto le categorie preferite raggruppate per paesi , poi caricato a parte e visualizzato con matplotlib:

https://drive.google.com/file/d/1dIhWEBk-Ftoa4gWx6eAn6AixrRROattc/view?usp=sharing

Come anticipato, con pandas seaborn e matplotlib ho cominciato a scavare più a fondo soprattutto tramite la visualizzazione in grafici.

```python
for i, year in enumerate(sorted(years)):
    yearly_data = category_trends[category_trends['year'] == year]
    sns.barplot(ax=axes[i], data=yearly_data, x='counts', y='category', palette='viridis')
    axes[i].set_title(f"Trend delle letture per categoria - {year}")
    axes[i].set_xlabel("Numero di Letture")


    axes[i].grid(True, linestyle='-', linewidth=0.5, color='grey')
    axes[i].set_facecolor('#A4D6E1')

    if i == 0:
        axes[i].set_ylabel("Categoria")
    else:
        axes[i].set_ylabel("")

plt.tight_layout()
plt.show()



# con questo bar plot possiamo vedere meglio i trend di lettura raggruppate per anno
```



Qui il link:

https://drive.google.com/file/d/1MrhpSH-0Rw73rkezuUqhGhhERXslyyq0/view?usp=sharing

Ed infine grazie a Tableau ho potuto avere una panoramica ancora più chiara ed interattiva dell'insieme.

Link di Tableau public:

https://public.tableau.com/views/ProgettoFinaleDataAnalysisdiSebastianoIaci/Dashboard1?:language=it-IT&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

- **Conclusioni**

Analizzare uno o più dataset con più strumenti offre la possibilità di trarre delle conclusioni più solide e dettagliate. Nel caso di questo dataset gli elementi chiave che accomunano le preferenze dei lettori possono essere generici come la durata preferita degli articoli che è quella lunga ma in altri casi , soprattutto grazie alla visualizzazione grafica, si possono vedere delle tendenze che è opportuno esplorare nel dettaglio in modo da definire degli aspetti specifici che spesso dipendono dalle singole circostanze come la preferenza delle piattaforme di lettura per paese e anno ma anche le categorie preferite sempre per anno e nazione.