



DANIELA SURCHICEAN

Fashion

PROGETTO FINALE DATA ANALYSIS



DANIELA SURCHICEAN

Introduzione

Questo dataset offre un'opportunità completa per mettere alla prova le tue competenze in modo olistico, spaziando dall'utilizzo di SQL a Tableau.

Ti permetterà di affinare le tue capacità di analisi e di esplorare la tua creatività in un contesto ricco di dati.

Progetto

PYTHON

<https://drive.google.com/drive/u/0/folders/1WGvCHsybayGSt9vGu2nVynwjr9DgfSNm>

TABLEAU

https://public.tableau.com/app/profile/daniela.surchicean/viz/Fashion_Progetto_Finale_Data_Analisys_Daniela_Surchicean/StoriaGenerale



Fashion

PROGETTO FINALE DATA ANALYSIS



DANIELA SURCHICEAN

Progetto

Sei interessato a esaminare i dati relativi al 2021 di un mercato online di abbigliamento di seconda mano, dove gli utenti possono mettere in vendita i propri capi. All'interno del set di dati, avrai accesso ai dettagli di tutte le transazioni effettuate dagli utenti, consentendoti di condurre analisi approfondite sulle prestazioni dei prodotti e dei designer, nonché sui trend temporali. Potrai inoltre esaminare i cluster di utenti per comprendere meglio il comportamento degli acquirenti. Di seguito troverai il link al dataset su cui potrai basare le tue analisi.





DANIELA SURHICEAN

Descrizione Dataset

Il dataset contiene informazioni dettagliate relative alle transazioni all'interno di un marketplace di abbigliamento second hand.

Include dati come l'identificativo dell'utente che vende, la categoria dell'abbigliamento, l'identificativo del designer, la lingua dell'utente, la fascia di prezzo, la nazionalità dell'utente, la data di acquisto, la piattaforma di pagamento, l'identificativo del prodotto, la valutazione media e la data di iscrizione dell'utente.

Questi dati offrono un'ampia gamma di informazioni utili per analizzare le transazioni, i comportamenti degli utenti e le prestazioni dei prodotti all'interno del marketplace.

user_uuid	Identificativo dell'utente che vende
category	Categoria di abbigliamento
designer_id	Identificativo del designer
language	Lingua dell'utente che vende
level	Fascia di prezzo
country	Nazionalità dell'utente che vende
purchase_date	La data in cui l'utente ha venduto il prodotto
platform	Piattaforma da cui è stato fatto il pagamento
item_id	Identificativo del prodotto
stars	Stelle medie assegnate al prodotto (da 1 a 5)
subscription_date	Giorno in cui l'utente che vende si è iscritto

EDA con Python

Analisi esplorativa dei dati

L'Analisi Esplorativa dei Dati (EDA) è il processo fondamentale per comprendere e familiarizzare con un set di dati prima di intraprendere ulteriori analisi statistiche. Consiste nell'esplorare e visualizzare i dati per scoprire modelli, tendenze, anomalie e relazioni tra le variabili.

L'EDA con Python sfrutta la potenza e la flessibilità del linguaggio di programmazione Python per eseguire l'analisi in modo efficiente e intuitivo, la libreria Python utilizzata per l'EDA, tra cui è Pandas, specializzata per la gestione e l'analisi di dati tabulari.

Nell'approccio non grafico, utilizzo funzioni come shape, summary, describe, isnull, info, datatypes e altro.

Nell'approccio grafico, utilizzo i grafici come i grafici a dispersione, a scatola, a barre, densità e correlazione.



Esplorazione dei dati in Python

- Caricamento dei dati: carichiamo il dataset nel tuo ambiente di lavoro Python utilizzando Pandas. Successivamente leggiamo i dati da file CSV, Excel, database o altre fonti. (Nell'analisi in questione abbiamo a disposizione il file csv);
- Importazione delle librerie: importiamo le librerie fondamentali per l'analisi dei dati, tra cui: Pandas, NumPy e Matplotlib. Queste librerie offrono funzionalità avanzate per la manipolazione, l'elaborazione e la visualizzazione dei dati.
- Esplorazione iniziale dei dati: la prima analisi è l'analisi esplorativa dei dati che esamina le righe del dataset per comprendere la struttura e il formato dei dati. (Nell'analisi in questione abbiamo utilizzato i metodi: `head()`, `info()`, `describe()`, utili proprio per ottenere una visione generale sul dataset);



Esplorazione dei dati in Python

- Pulizia dei dati: identifichiamo e gestiamo i valori mancanti, i valori duplicati o eventuali errori nei dati, durante questa fase di analisi abbiamo utilizzato i metodi `isnull()`, `drop_duplicates()`, `fillna()`, utili per pulire il dataset.
- Analisi delle variabili: esploriamo le diverse variabili del dataset e calcoliamo statistiche come media, mediana e deviazione standard.
- Esplorazione delle relazioni tra le variabili : analizziamo le relazioni tra le variabili del dataset, utilizzando grafici come scatterplot, correlazione o grafici a barre.





DANIELA SURHICEAN

Andamento delle vendite

#DATA ANALYSIS AND VISUALISATION

#calcolo in numero di vendite per ogni anno

```
vendite_per_anno = dtfashion['yquarter_purchase'].value_counts()  
vendite_per_anno_asc = vendite_per_anno.sort_values(ascending=True)  
vendite_per_anno_asc
```

#creo il grafico a linee per mostrare l'andamento annuale delle vendite

```
line_plot = sns.lineplot(x=vendite_per_anno_asc.index, y=vendite_per_anno_asc.values, marker='s', markersize=8)
```

```
plt.xlabel('Anno')  
plt.ylabel('Numero Vendite/Acquisti')  
plt.title('Andamento Annuale delle Vendite/Acquisti')
```

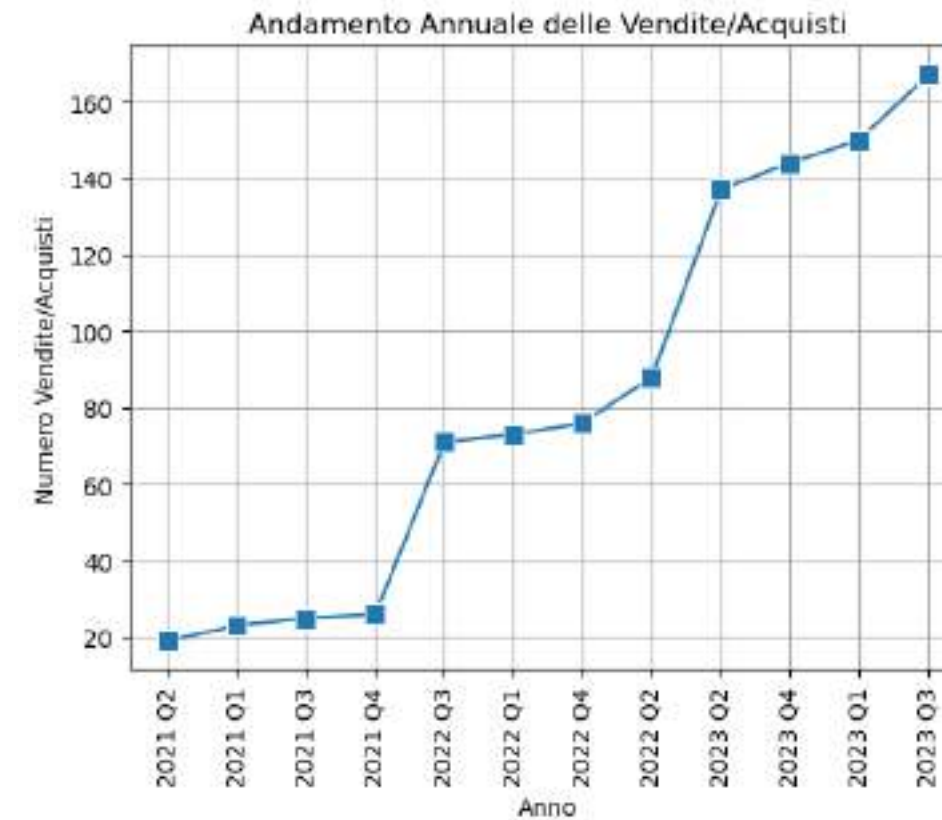
```
plt.xticks(rotation=90)  
plt.grid(True)  
plt.show()
```

#visualizzo le vendite raggruppate per anno utilizzando un grafico a barre

```
bar_plot = sns.barplot(x=vendite_per_anno_asc.index, y=vendite_per_anno_asc.values)  
plt.xlabel('Anno')  
plt.ylabel('Numero di vendite/acquisti')  
plt.title('Numero di Vendite/Acquisti per Anno')  
plt.grid(True)  
plt.xticks(rotation=90)  
for index, value in enumerate(vendite_per_anno_asc.values):  
    bar_plot.text(index, value, str(value), ha='center', va='bottom')  
plt.show()
```

#creo il grafico a torta sulle vendite raggruppate per anno

```
plt.figure(figsize=(10, 6))  
plt.pie(vendite_per_anno_asc.values, labels=vendite_per_anno_asc.index, autopct='%1.1f%%')  
plt.title('Distribuzione Vendite/Acquisti per Anno')  
plt.show()
```

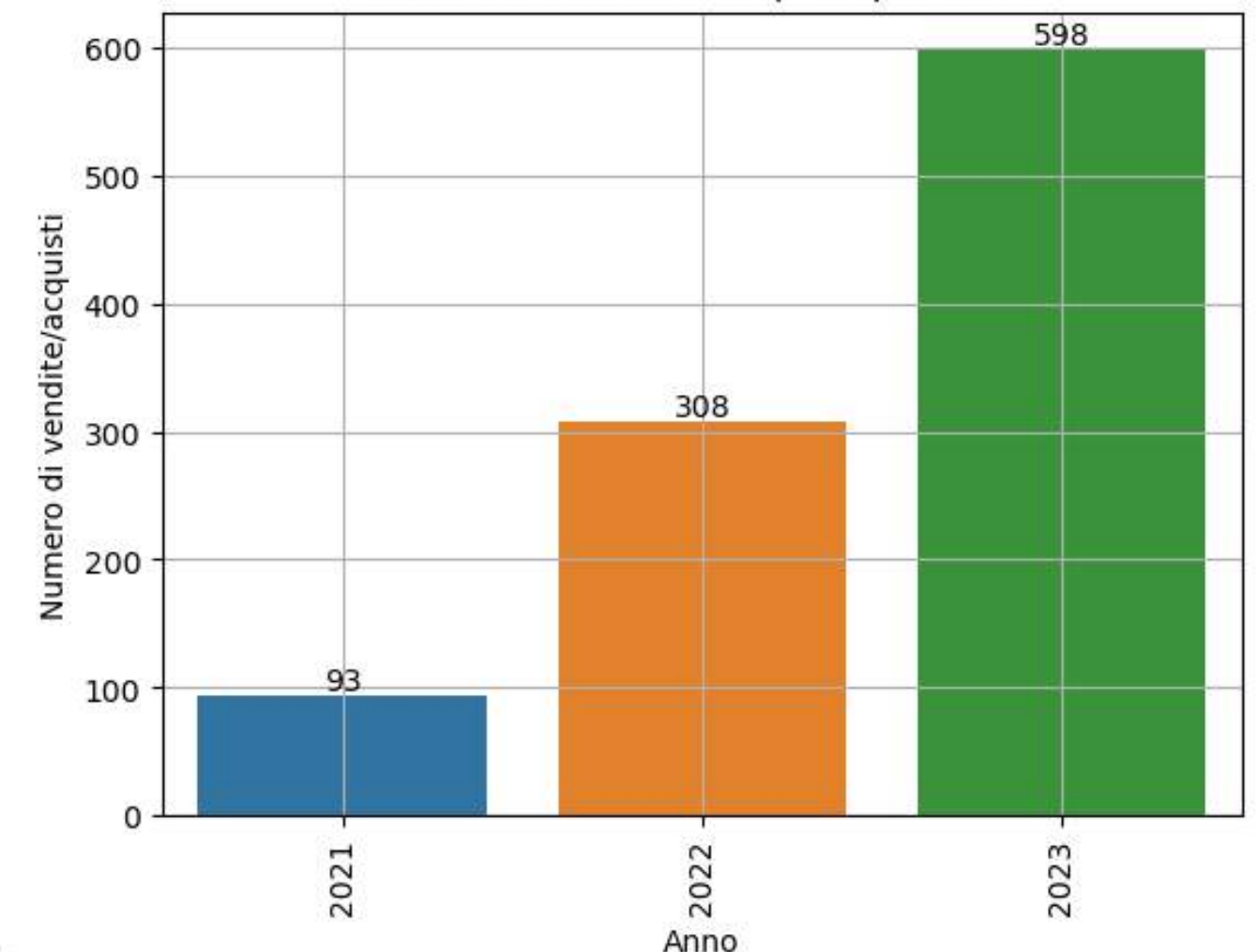


yquarter_purchase

2021 Q2	19
2021 Q1	23
2021 Q3	25
2021 Q4	26
2022 Q3	71
2022 Q1	73
2022 Q4	76
2022 Q2	88
2023 Q2	137
2023 Q4	144
2023 Q1	150
2023 Q3	167

Name: count, dtype: int64

Numero di Vendite/Acquisti per Anno



purchase_year

2023	598
2022	308
2021	93

Name: count, dtype: int64

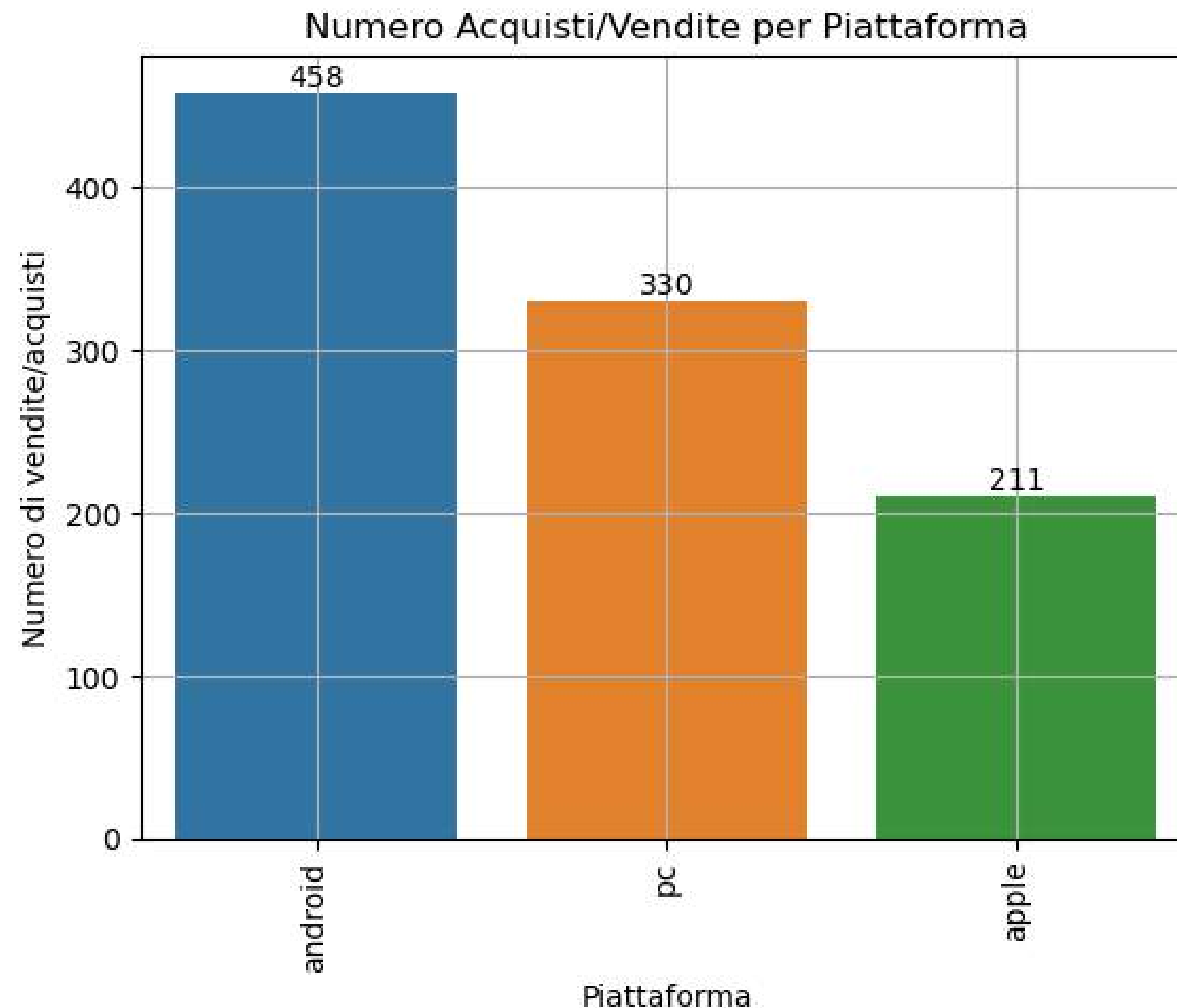


DANIELA SURHICEAN

Andamento delle vendite in base alla piattaforma

```
#calcolo il numero di acquisti per ogni piattaforma  
acquisti_per_piattaforma = dtfashion['platform'].value_counts()  
acquisti_per_piattaforma
```

```
#calcolo il numero di acquisti per ogni piattaforma  
bar_plot = sns.barplot(x=acquisti_per_piattaforma.index, y=acquisti_per_piattaforma.values)  
plt.xlabel('Piattaforma')  
plt.ylabel('Numero di vendite/acquisti')  
plt.title('Numero Acquisti/Vendite per Piattaforma')  
plt.grid(True)  
plt.xticks(rotation=90)  
for index, value in enumerate(acquisti_per_piattaforma.values):  
    bar_plot.text(index, value, str(value), va='bottom')  
plt.show()
```



```
platform  
android    458  
pc         330  
apple      211  
Name: count, dtype: int64
```

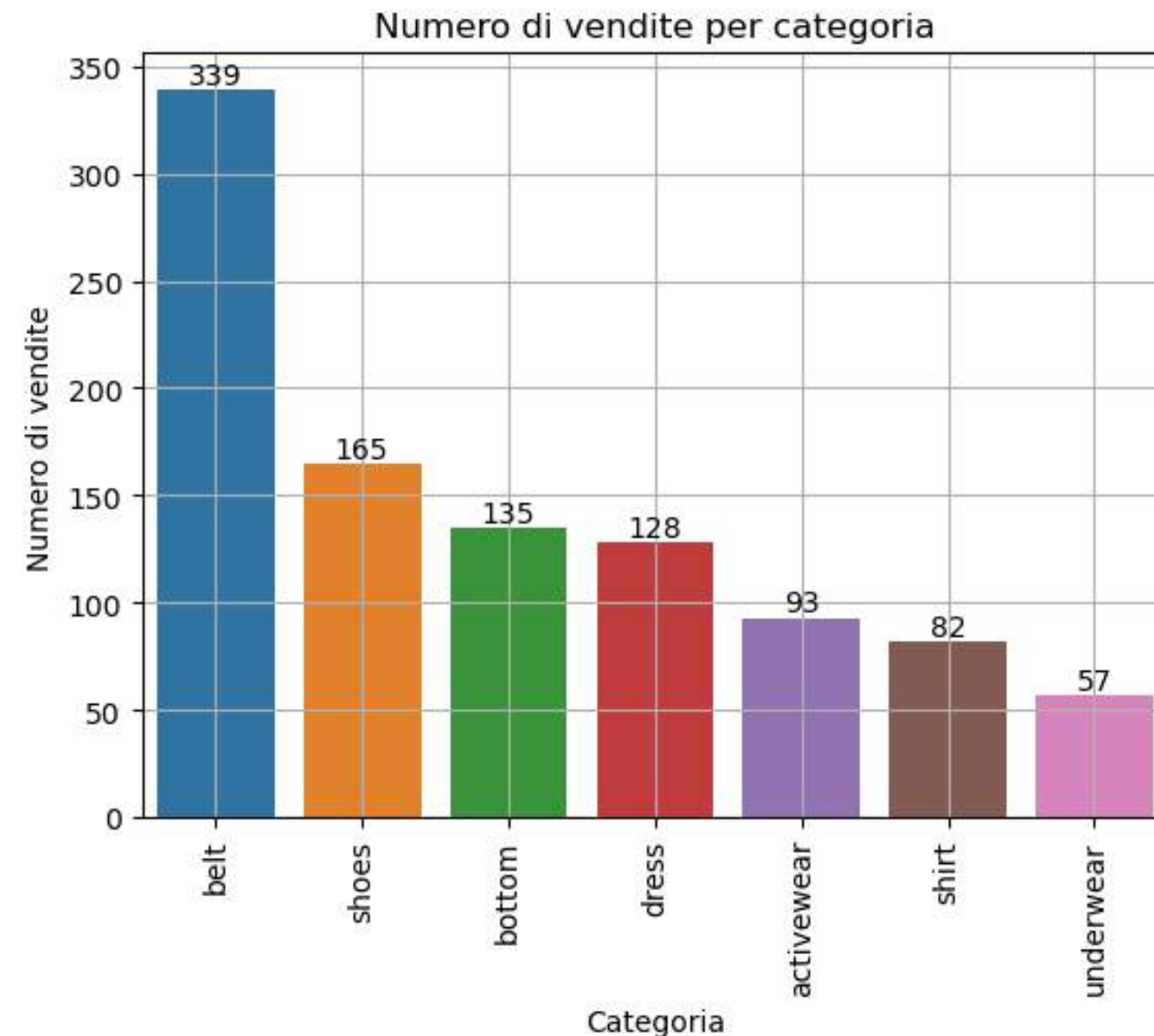



DANIELA SURCHICEAN

Andamento delle vendite in base alla categoria

```
#calcolo in numero di vendite per ogni categoria prodotto  
vendite_per_cate = dtfashion['category'].value_counts()  
vendite_per_cate
```

```
#visualizzo queste vendite utilizzando un grafico a barre  
bar_plot = sns.barplot(x=vendite_per_cate.index, y=vendite_per_cate.values)  
plt.xlabel('Categorie')  
plt.ylabel('Numero di vendite')  
plt.title('Numero di Vendite per Categoria')  
plt.grid(True)  
plt.xticks(rotation=90)  
for index, value in enumerate(vendite_per_cate.values):  
    bar_plot.text(index, value, str(value), ha='center', va='bottom')  
plt.show()
```



```
category  
belt      339  
shoes     165  
bottom    135  
dress     128  
activewear 93  
shirt     82  
underwear 57  
Name: count, dtype: int64
```

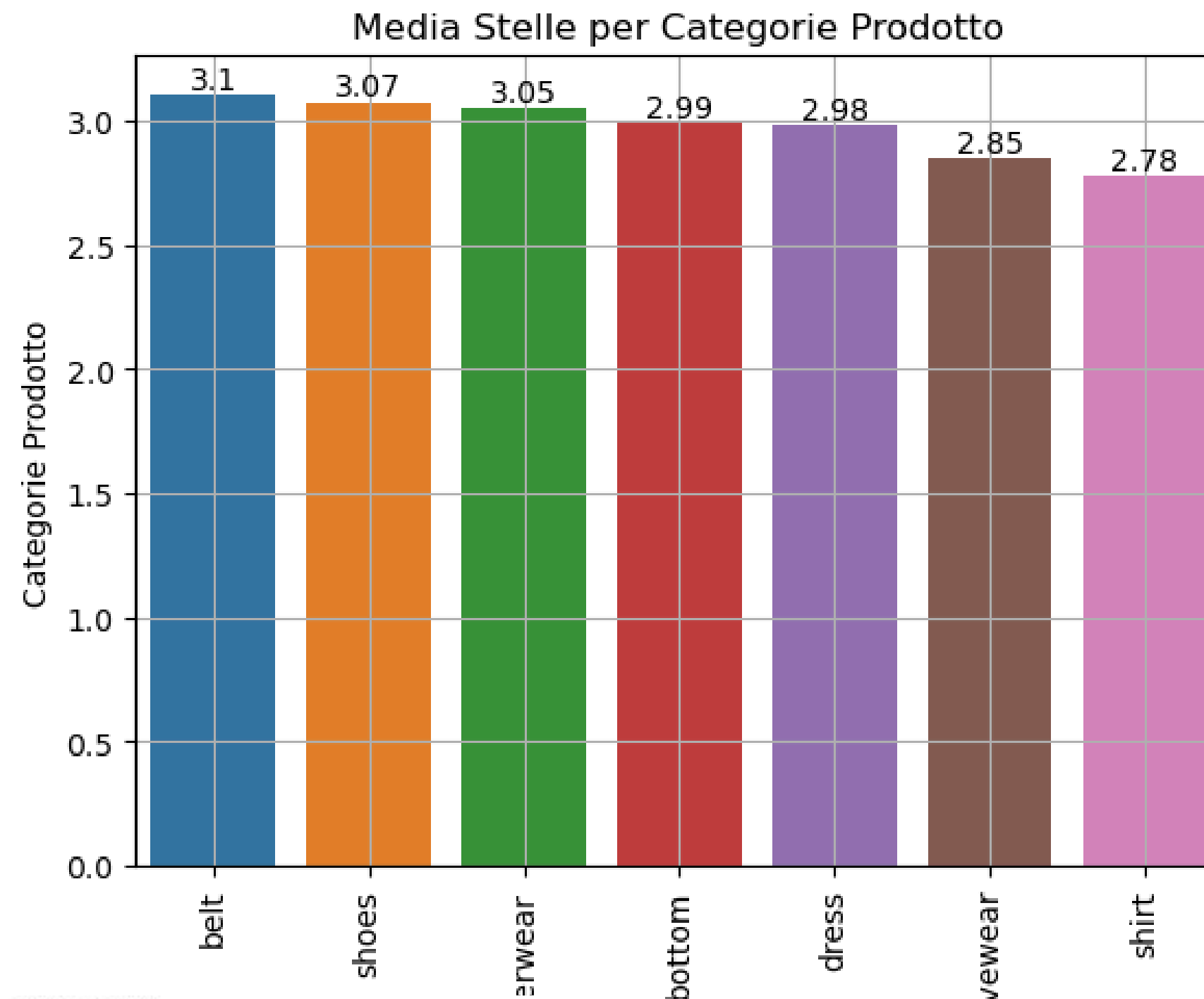



DANIELA SURHICEAN

Analisi del gradimento dei prodotti in base alla categoria

```
#calcolo la media delle stelle assegnate ai prodotti raggruppati per categoria prodotto
media_stelle_per_cate = dtfashion.groupby('category')['stars'].mean()
media_stelle_per_cate_desc = media_stelle_per_cate.sort_values(ascending=False)
media_stelle_per_cate_desc
```

```
per il grafico a barre per mostrare la media di stelle assegnata ad ogni categoria prodotto
plot = sns.barplot(x=media_stelle_per_cate_desc.index, y=media_stelle_per_cate_desc.values)
plot.xlabel('Stelle')
plot.ylabel('Categorie Prodotto')
plot.title('Media Stelle per Categorie Prodotto')
plot.xticks(rotation=90)
plot.grid(True)
for index, value in enumerate(media_stelle_per_cate_desc.values):
    bar_plot.text(index, value, s=round(value, 2), ha='center', va='bottom')
plot.show()
```



category	stars
belt	3.103245
shoes	3.072727
underwear	3.052632
bottom	2.985185
dress	2.976562
activewear	2.849462
shirt	2.780488

Name: stars, dtype: float64



DANIELA SURHICEAN

Analisi dei prodotti in base alla categoria di prezzo

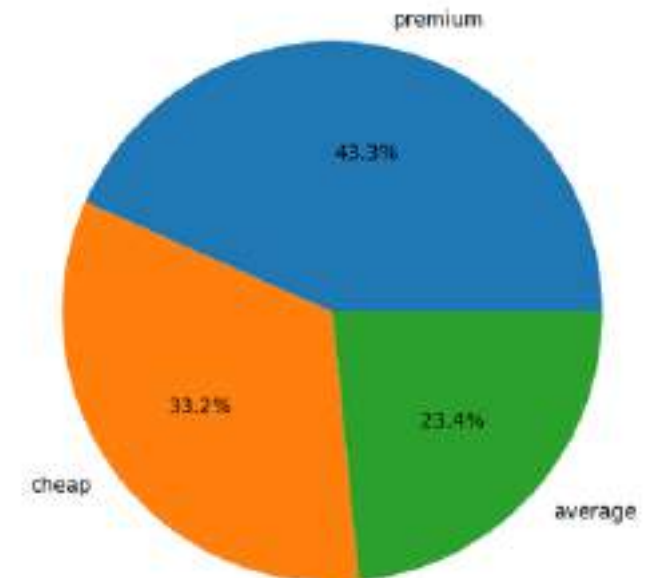
```
#calcolo in numero di prodotti raggruppandoli per la fascia di prezzo corrispondente  
prodotti_perlivello_prezzo = dfashion['level'].value_counts()  
prodotti_perlivello_prezzo
```

```
#creo il grafico a barre sul raggruppamento dei prodotti per fascia di prezzo di appartenenza  
bar_plot = sns.barplot(x=prodotti_perlivello_prezzo.values, y=prodotti_perlivello_prezzo.index)  
plt.xlabel('Numero di Prodotti')  
plt.ylabel('Categorie di Prezzo')  
plt.title('Numero Prodotti per Categorie di Prezzo')  
plt.xticks(rotation=90)  
plt.grid(True)  
plt.show()
```

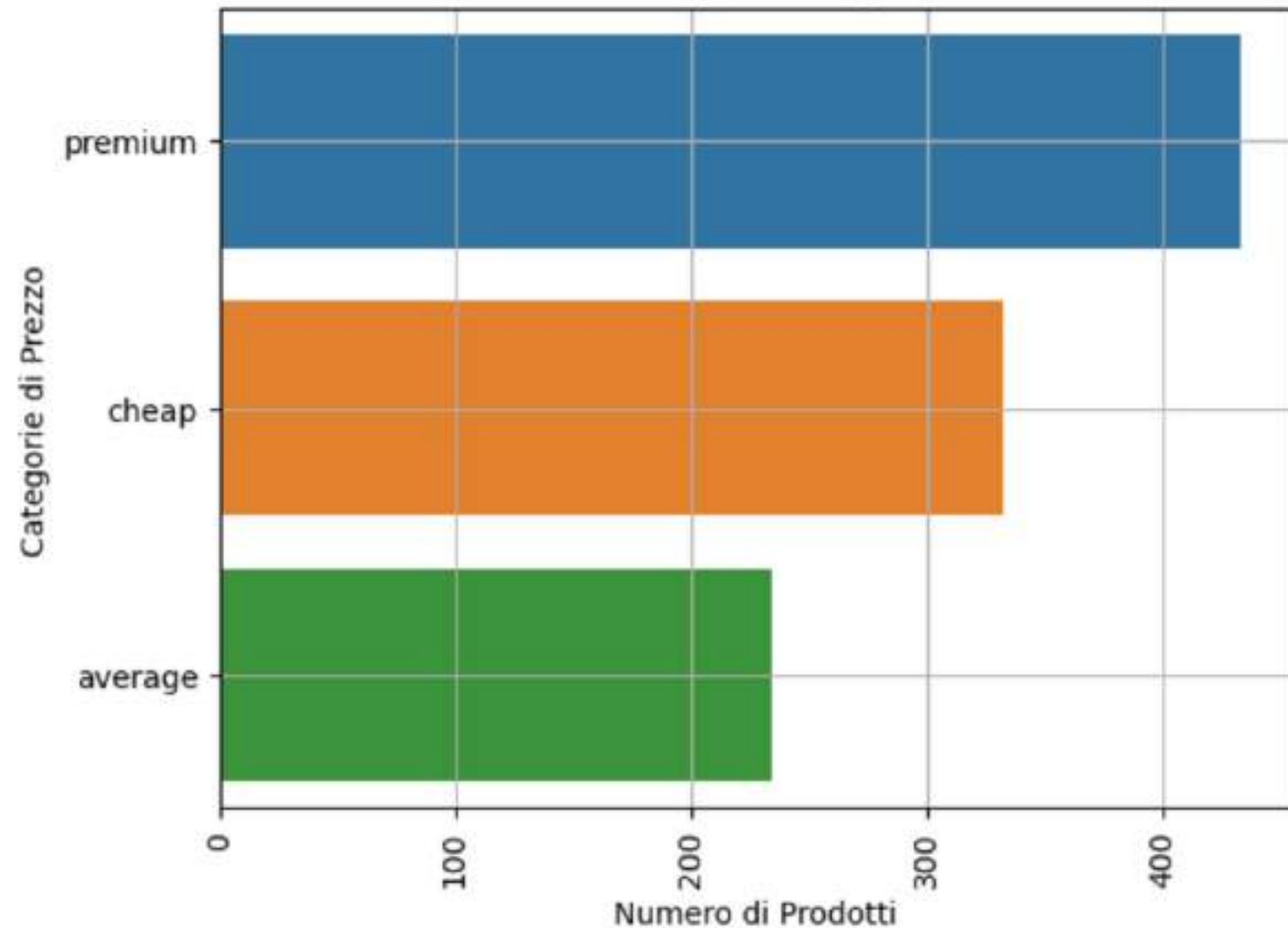
```
#creo il grafico a torta sul raggruppamento dei prodotti per fascia di prezzo di appartenenza  
plt.figure(figsize=(10, 6))  
plt.pie(prodotti_perlivello_prezzo.values, labels=prodotti_perlivello_prezzo.index, autopct='%1.1f%%')  
plt.title('Percentuale di Prodotti per Categorie di Prezzo')  
plt.show()
```

```
level  
premium    433  
cheap      332  
average    234  
Name: count, dtype: int64
```

Percentuale di Prodotti per Categorie di Prezzo



Numero Prodotti per Categorie di Prezzo



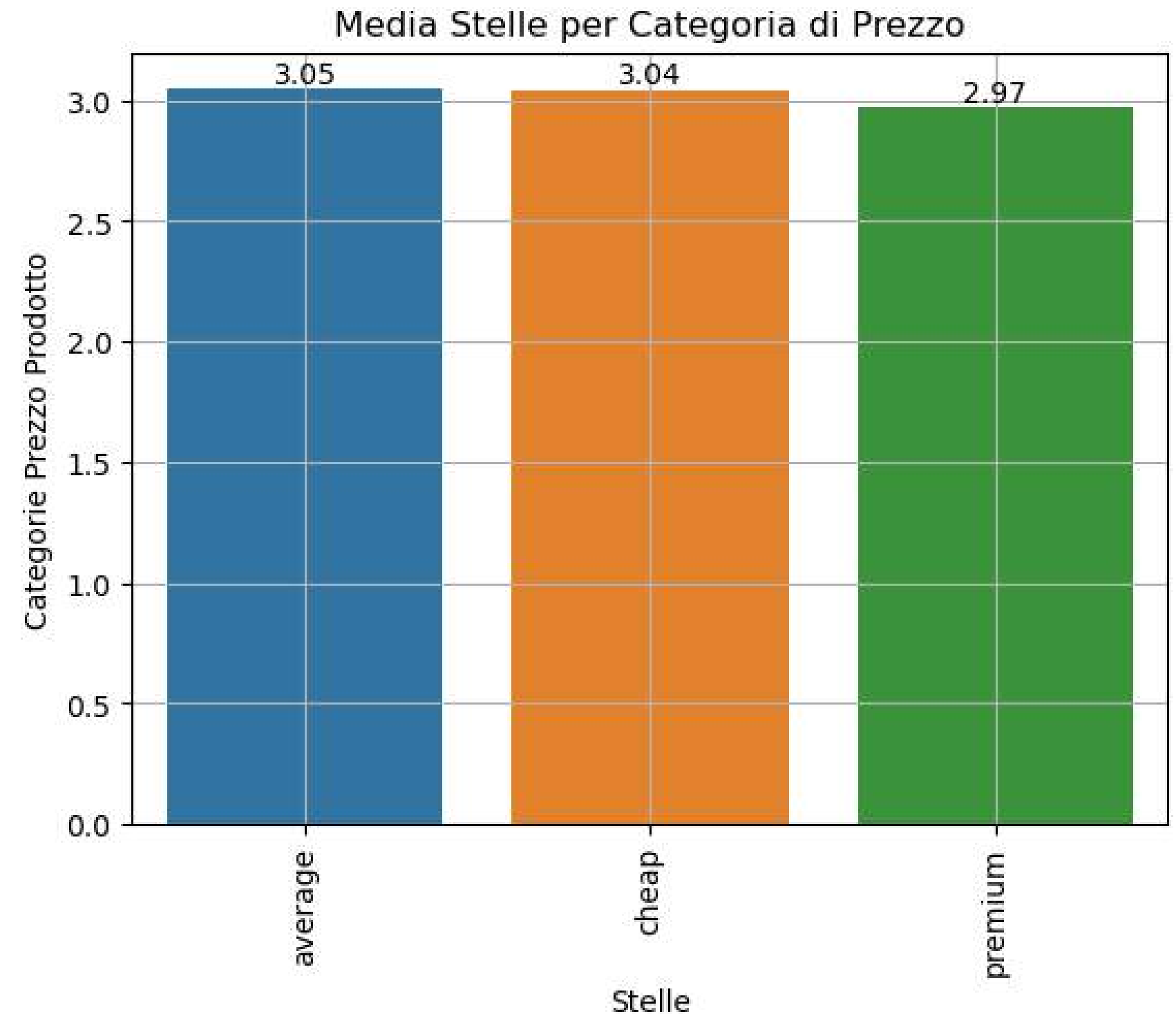


DANIELA SURHICEAN

Analisi del gradimento dei prodotti in base alla categoria di prezzo

```
#calcolo la media delle stelle assegnate ai prodotti raggruppati per categoria prezzo
media_stelle_per_cprezzo = dtfashion.groupby('level')['stars'].mean()
media_stelle_per_cprezzo_desc = media_stelle_per_cprezzo.sort_values(ascending=False)
media_stelle_per_cprezzo_desc
```

```
#creo il grafico a linee per mostrare la media di stelle assegnata ad ogni categoria prezzo
bar_plot = sns.barplot(x=media_stelle_per_cprezzo_desc.index, y=media_stelle_per_cprezzo_desc.values)
plt.xlabel('Stelle')
plt.ylabel('Categorie Prezzo Prodotto')
plt.title('Media Stelle per Categoria di Prezzo')
plt.xticks(rotation=90)
plt.grid(True)
for index, value in enumerate(media_stelle_per_cprezzo_desc.values):
    bar_plot.text(index, value, s=round(value, 2), ha='center', va='bottom')
plt.show()
```



```
level
average    3.047009
cheap      3.042169
premium    2.972286
Name: stars, dtype: float64
```

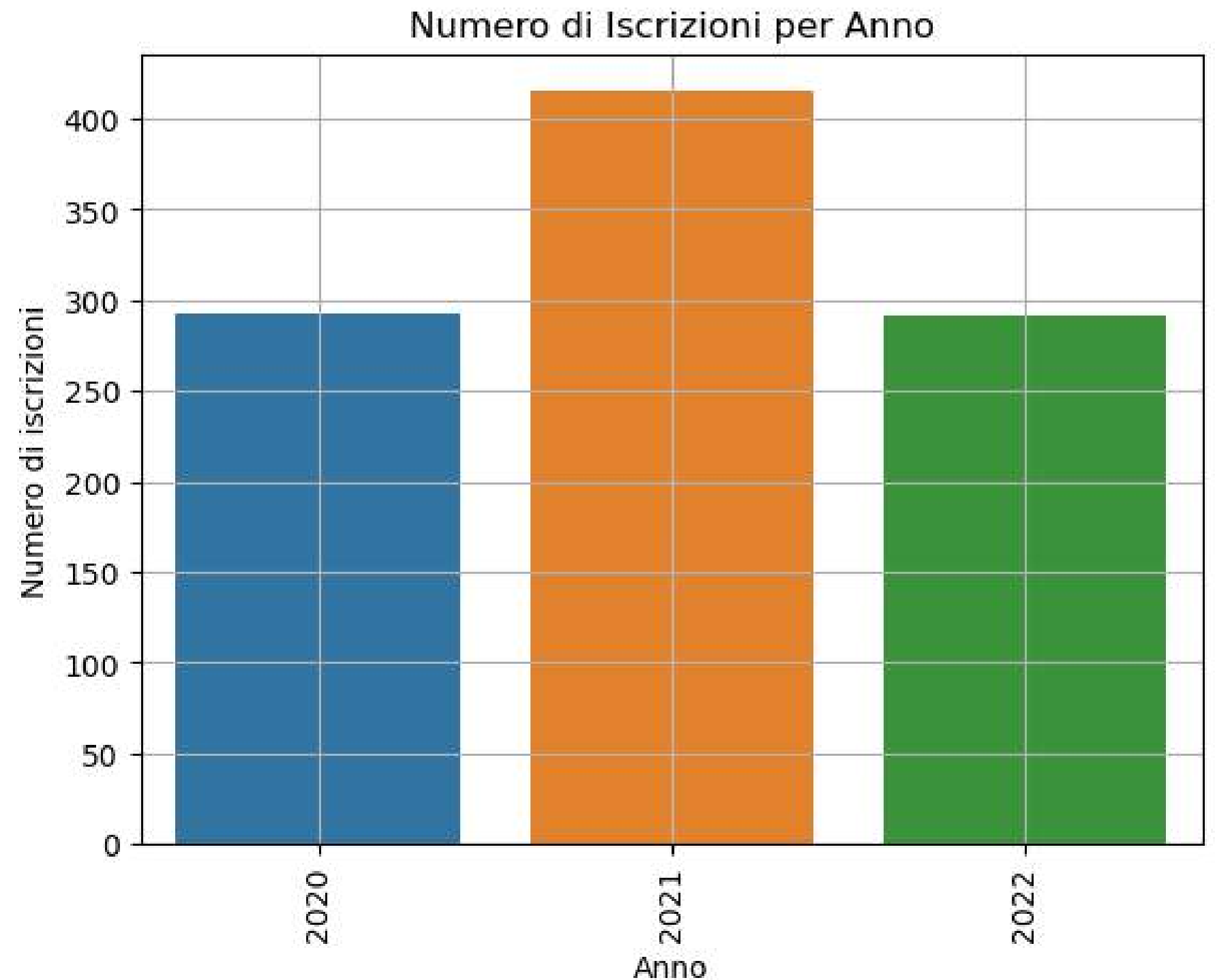



DANIELA SURHICEAN

Analisi delle iscrizioni annuali

```
#calcolo il numero di iscrizioni per ogni anno  
iscrizioni_per_anno = dtfashion['subscription_year'].value_counts()  
iscrizioni_per_anno
```

```
#visualizzo le iscrizioni raggruppate per anno utilizzando un grafico a barre  
bar_plot = sns.barplot(x=iscrizioni_per_anno.index, y=iscrizioni_per_anno.values)  
plt.xlabel('Anno')  
plt.ylabel('Numero di iscrizioni')  
plt.title('Numero di Iscrizioni per Anno')  
plt.grid(True)  
plt.xticks(rotation=90)  
plt.show()
```



```
subscription_year  
2021    415  
2020    293  
2022    291  
Name: count, dtype: int64
```




DANIELA SURCHICEAN

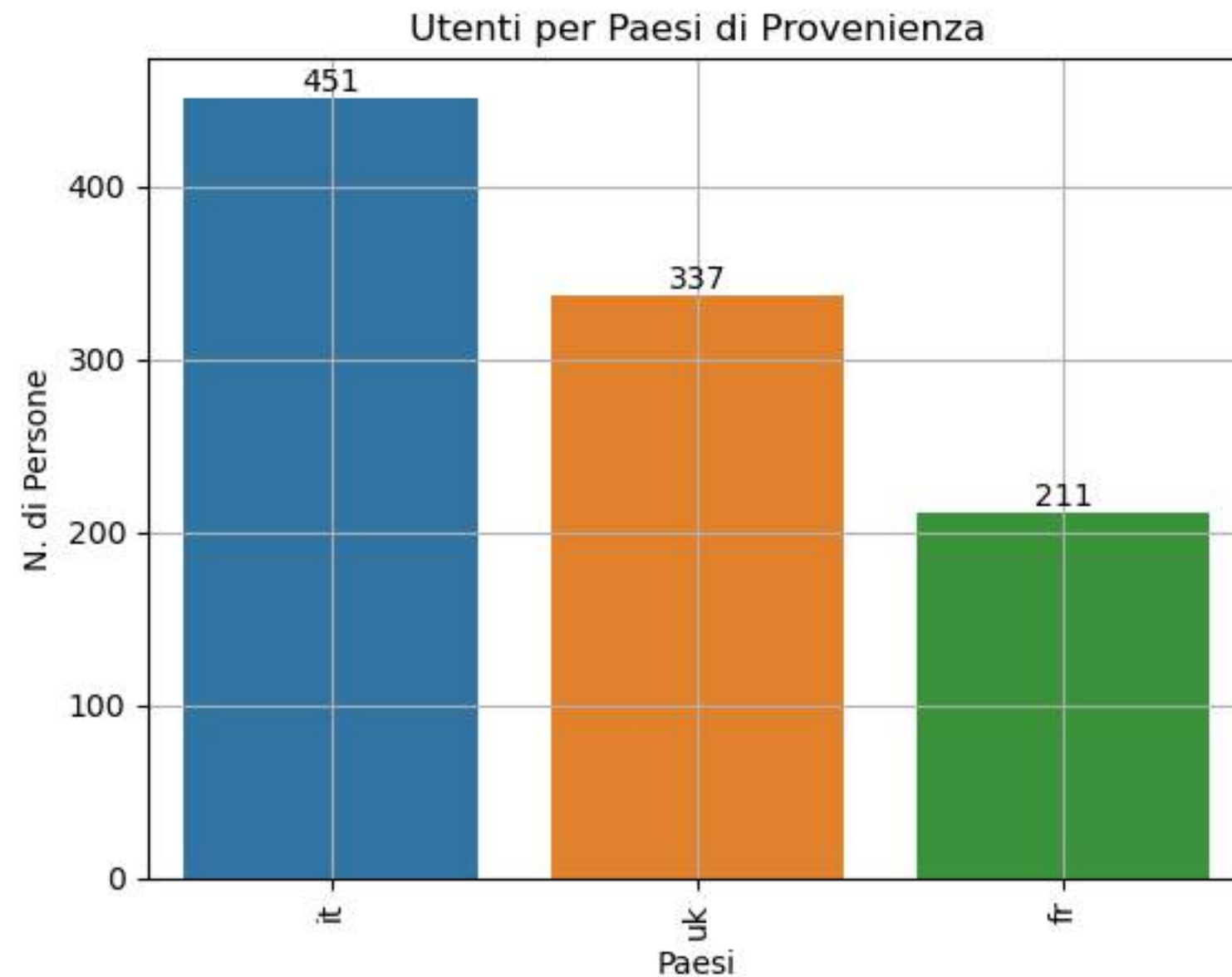
Analisi delle nazionalità degli iscritti

```
#calcolo in numero di persone raggruppandole per paese di provenienza
persone_per_paese = dtfashion['country'].value_counts()
persone_per_paese
```

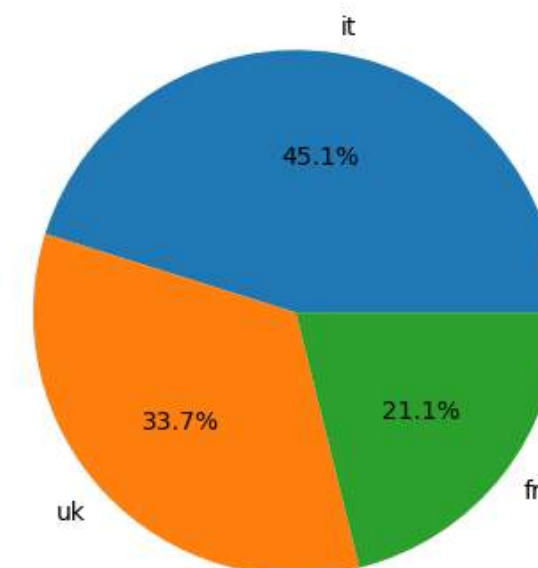
```
#creo il grafico a barre sul raggruppamento delle persone per paese di provenienza
bar_plot = sns.barplot(x=persone_per_paese.index, y=persone_per_paese.values)
plt.xlabel('Paesi')
plt.ylabel('N. di Persone')
plt.title('Utenti per Paesi di Provenienza')
plt.grid(True)
plt.xticks(rotation=90)
for index, value in enumerate(persone_per_paese.values):
    bar_plot.text(index, value, str(value), ha='center', va='bottom')

plt.show()
```

```
#creo il grafico a torta sul raggruppamento degli utenti per paese di provenienza
pie_chart = plt.pie(persone_per_paese.values, labels=persone_per_paese.index, autopct='%1.1f%%')
plt.title('Distribuzione delle Persone in base ai Paesi di Provenienza')
plt.show()
```



Distribuzione delle Persone in base ai Paesi di Provenienza



```
country
it      451
uk      337
fr      211
Name: count, dtype: int64
```

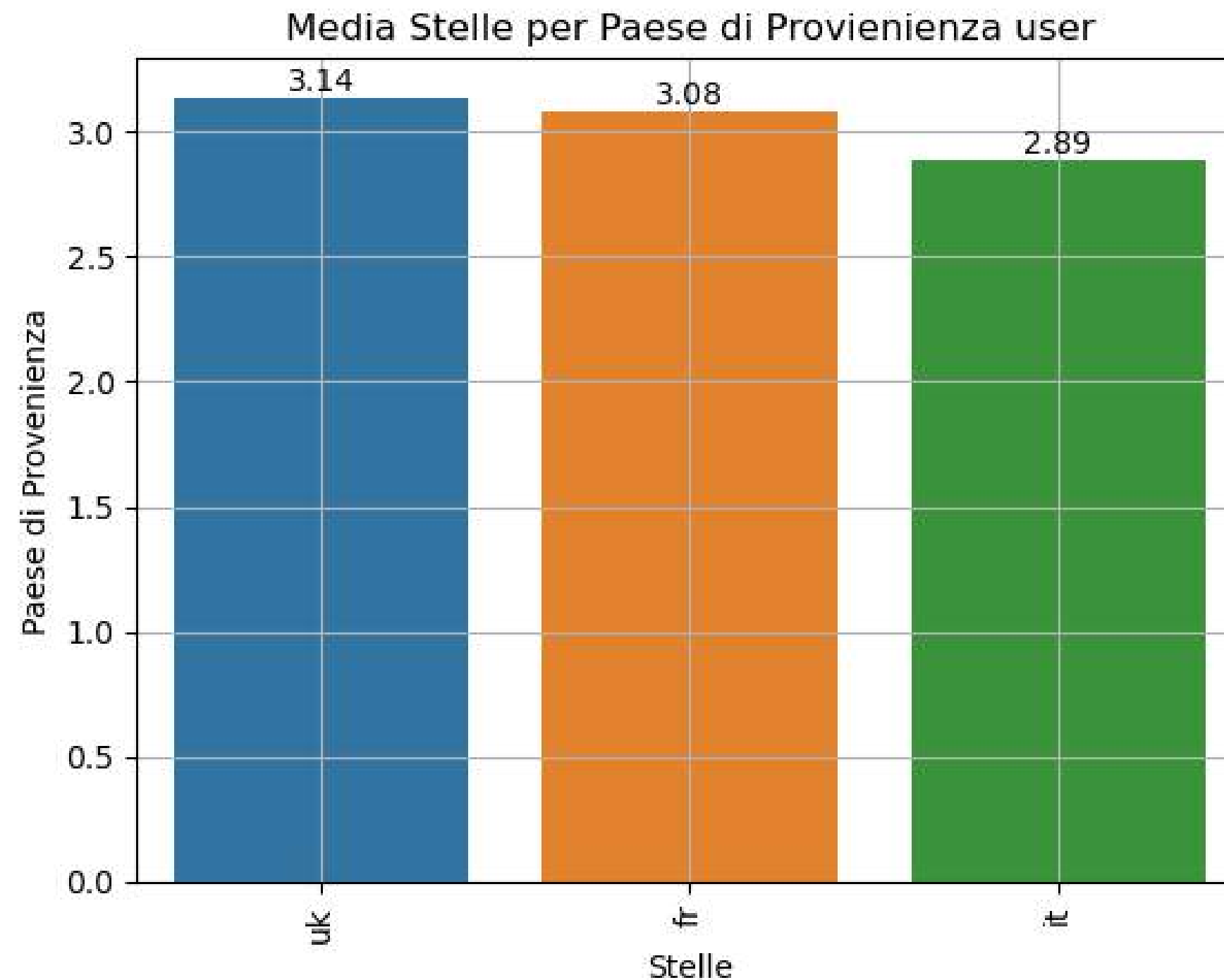



DANIELA SURHICEAN

Analisi del gradimento degli acquisti in base al paese di provenienza dell'utente

```
#calcolo la media delle stelle assegnate alle vendite raggruppate per paese di provenienza dello user
media_stelle_per_cprezzo = dtfashion.groupby('country')['stars'].mean()
media_stelle_per_cprezzo_desc = media_stelle_per_cprezzo.sort_values(ascending=False)
media_stelle_per_cprezzo_desc
```

```
#creo il grafico a barre per mostrare la media di stelle assegnata ad ogni categoria prezzo
bar_plot = sns.barplot(x=media_stelle_per_cprezzo_desc.index, y=media_stelle_per_cprezzo_desc.values)
plt.xlabel('Stelle')
plt.ylabel('Paese di Provenienza')
plt.title('Media Stelle per Paese di Provenienza user')
plt.xticks(rotation=90)
plt.grid(True)
for index, value in enumerate(media_stelle_per_cprezzo_desc.values):
    bar_plot.text(index, value, s=round(value, 2), ha='center', va='bottom')
plt.show()
```



```
country
uk      3.136499
fr      3.080569
it      2.889135
Name: stars, dtype: float64
```


Conclusioni EDA

Analisi esplorativa dei dati

Dopo aver effettuato EDA possiamo affermare che:

- la categoria dove si effettuano più vendite è quella riguardante le cinture (Belt);
- la maggior parte degli acquisti hanno una fascia di prezzo costosa ovvero appartengono ad utenti Premium;
- la categoria più apprezzata dagli utenti è quella delle cinture (Belt);
- gli utenti presenti nel dataset provengono da Italia, Inghilterra e Francia, e la maggioranza di loro proviene dall'Italia;
- la maggior parte degli acquisti viene effettuata dalla piattaforma Android;
- gli utenti unici presenti sono 241;
- le vendite per le tipologie utenti variano tra 2,4 fino a 2,9;
- i venditori più affidabili, a cui è assegnata una valutazione più alta fanno parte della categoria premium.



ESPLORAZIONE DATI 2021

Eseguiamo l'analisi dei dati relativi al solo anno 2021 avvalendoci della libreria Mito.

Mito è una libreria Python opensource che facilita l'analisi e l'automazione di dati in fogli di calcolo; offre un'interfaccia intuitiva e flessibile per lavorare con fogli di calcolo Excel, CSV e DataFrame di Pandas. Inoltre utilizza un'interfaccia a foglio di calcolo familiare, simile a Excel, per visualizzare e manipolare i dati. E' possibile importare e utilizzare tutte le funzionalità descritte della libreria Mito in Jupyter Notebook . Così che ogni modifica apportata al foglio di calcolo genererà il Python equivalente nella cella di codice sottostante. Ad esempio, se crei una tabella pivot in Mito, il codice pivot Pandas apparirà automaticamente, con la documentazione nella cella del codice seguente.



ESPLORAZIONE DATI 2021

HomeInsertDataFormulasCode

ImportExportCopy

Default

Conditional FormattingFormatInsertDeleteFilterFind & ReplaceType

PivotMergeGraphAI

HelpUpgrade to Mito Pro

purchase_date | purchase_date

	user_uid	category	designer_id	language
	int	str	int	
2	158	belt	106	it
12	207	bottom	112	it
16	110	dress	108	fr
17	145	underwear	117	en
44	89	shirt	110	en
57	67	activewear	110	it
76	243	belt	121	en
77	182	activewear	112	en
117	69	underwear	107	en
138	239	belt	115	it
142	115	shoes	110	fr
143	34	bottom	112	en

purchase_date

Dtypedatetime

SortAscendingDescending

Filter

Where>=01/01/2021

And<=31/12/2021

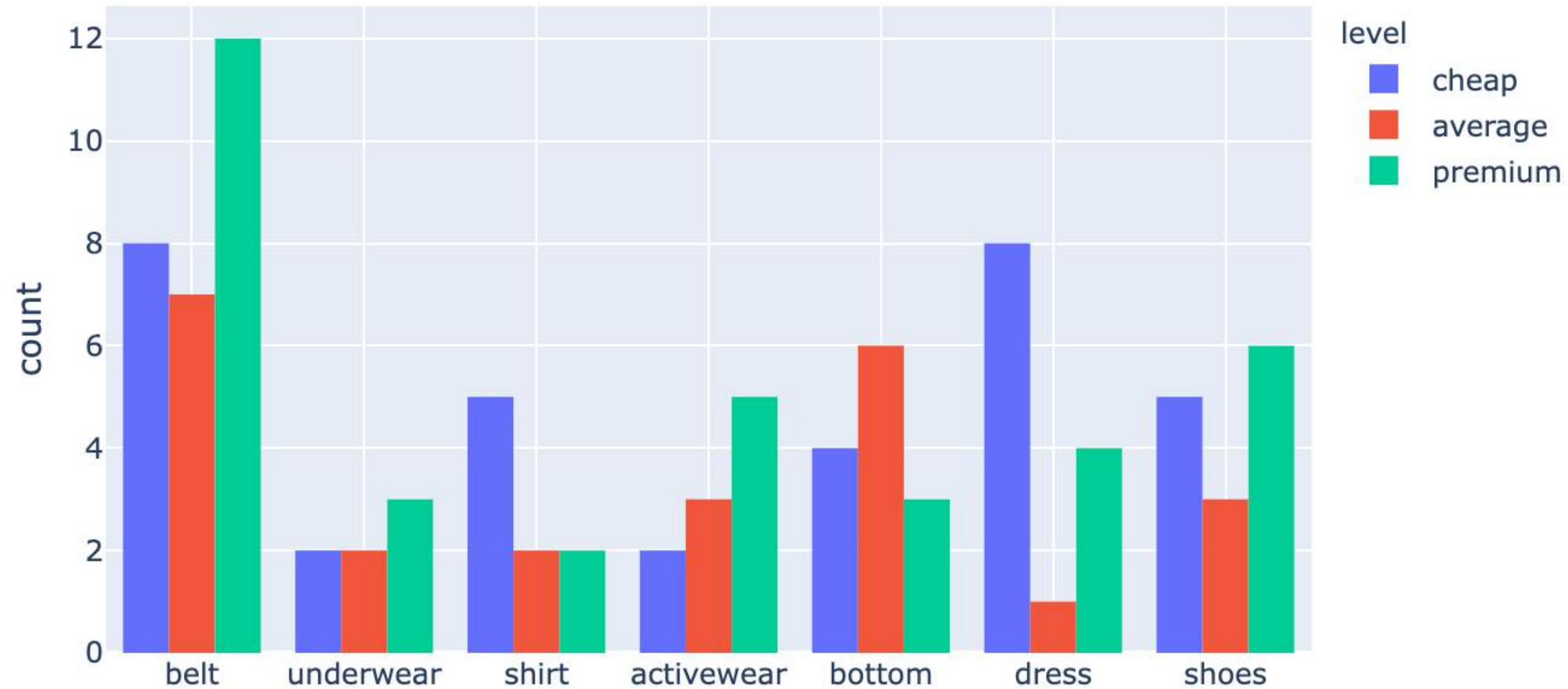
+ Add Filter

Removed an additional 906 rows

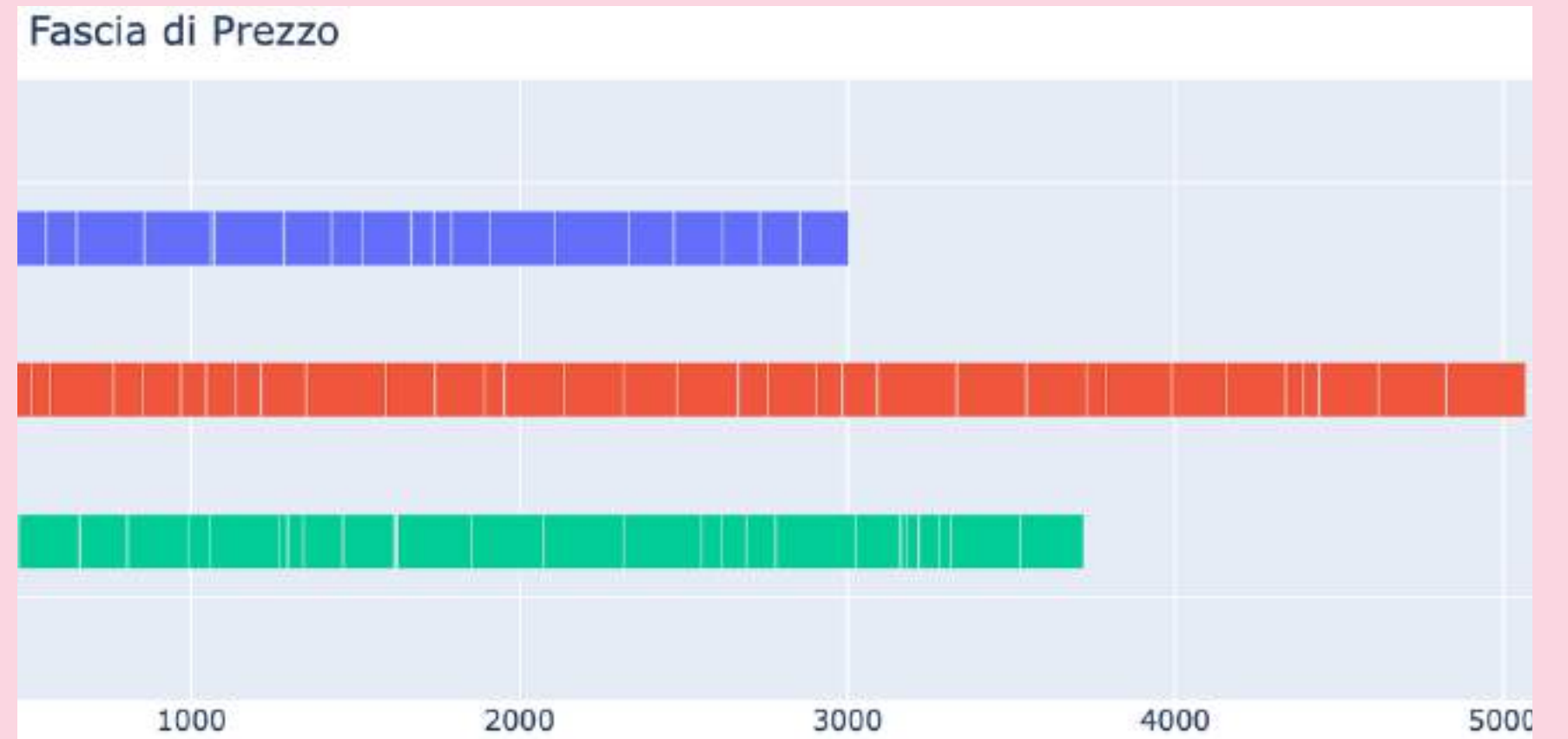
Filter/SortValuesStats

+fashionfashion_1fashion_2fashion_3fashion_493 rows, 11 cols

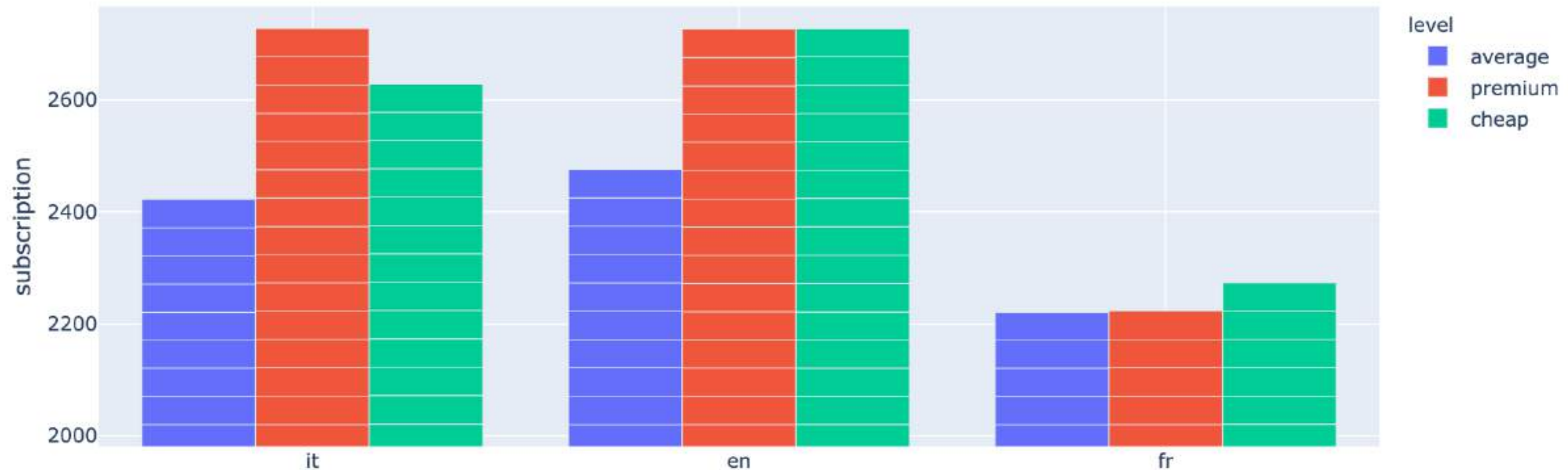
Vendite 2021



ESPLORAZIONE DATI 2021



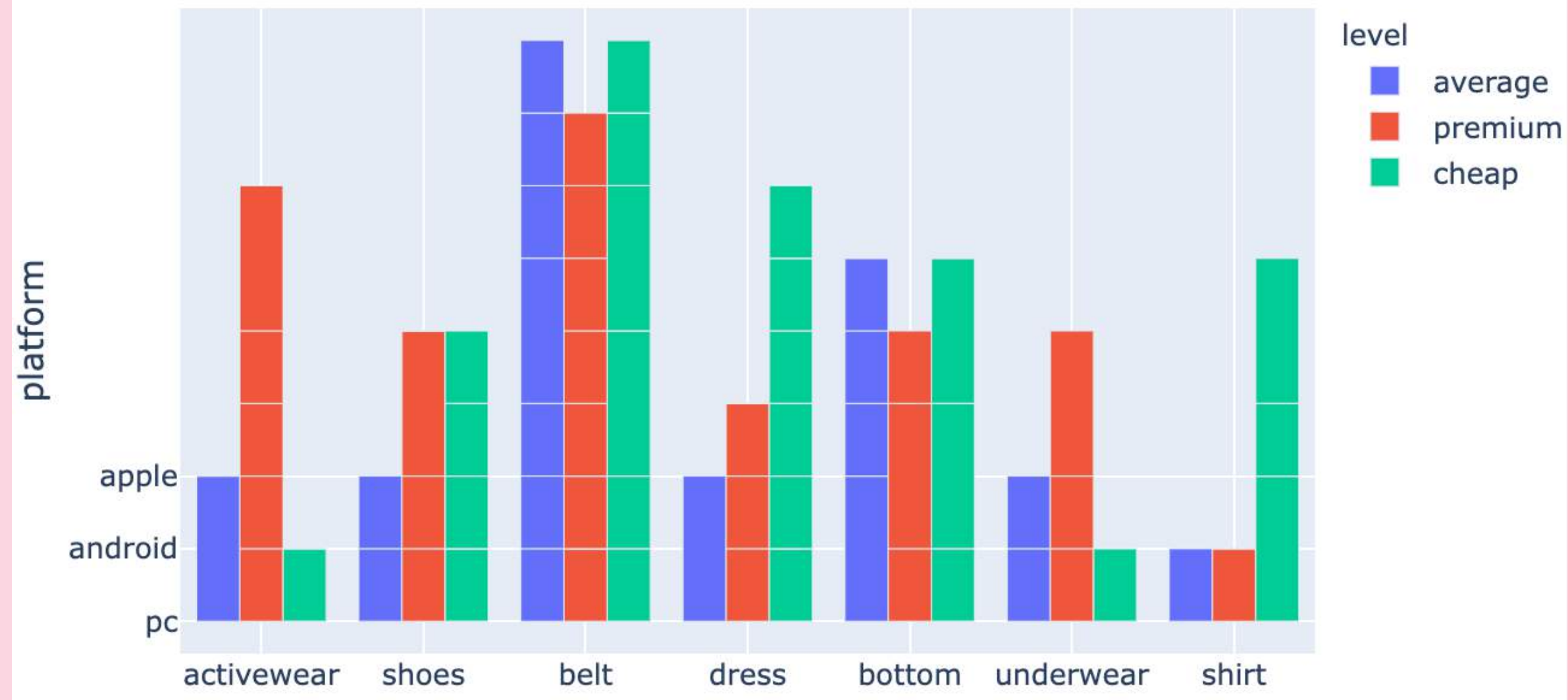
Paesi di Provenienza delle Tipologia di Utenti



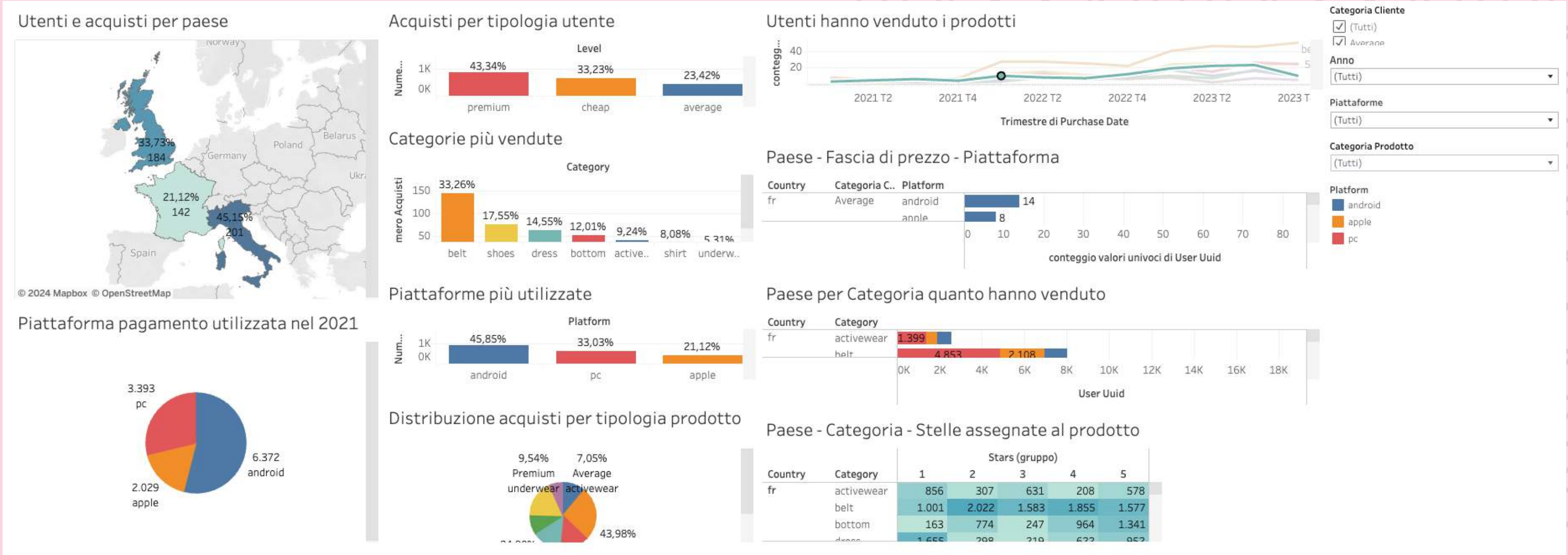
ESPLORAZIONE DATI 2021



Categorie Vendute da Piattaforma 2021



Creazione Dashboard Generale TABLEAU



Creazione Dashboard TABLEAU relativa al 2021

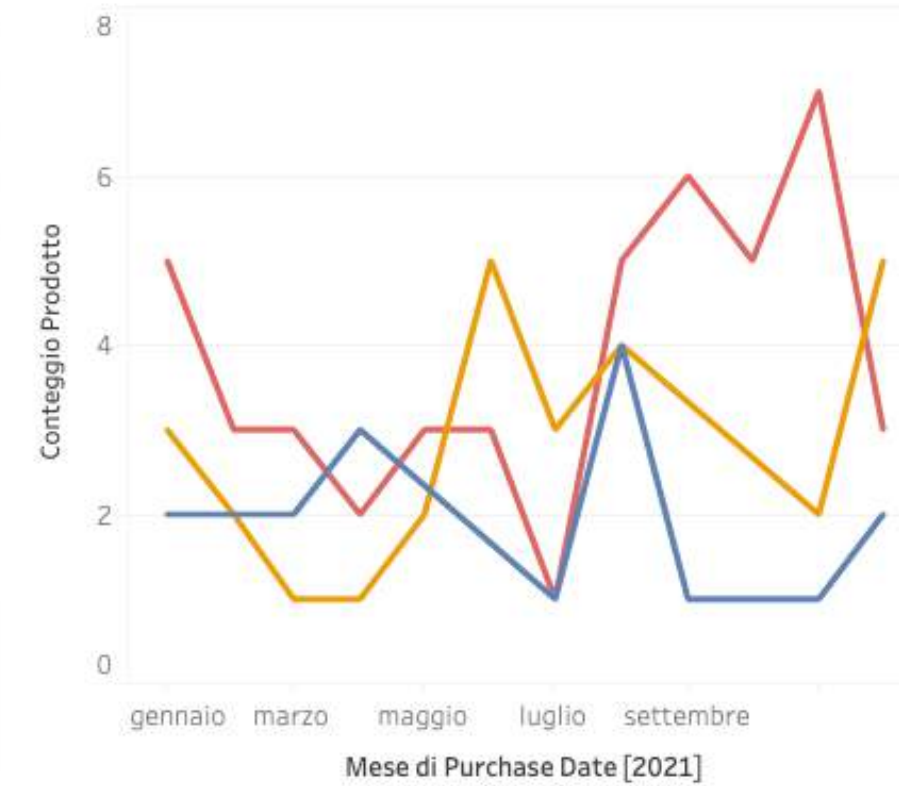


DANIELA SURCHICEAN

Utenti iscritti nei mesi del 2021



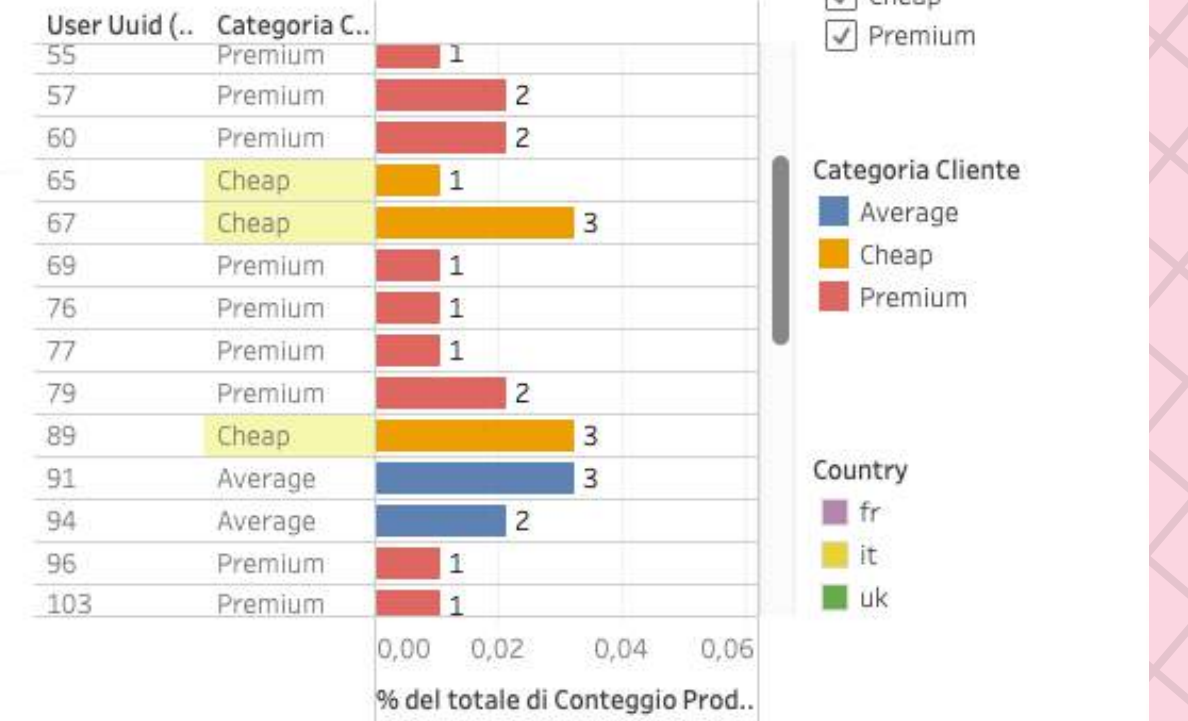
Venduti prodotti nei mesi del 2021



Fascia di prezzo e utente che vende nel 2021



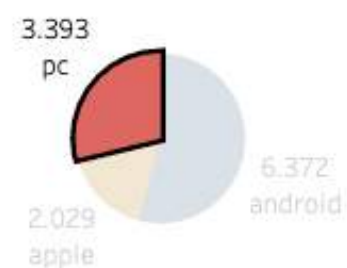
Gli utenti e le fasce di prezzo di vendita



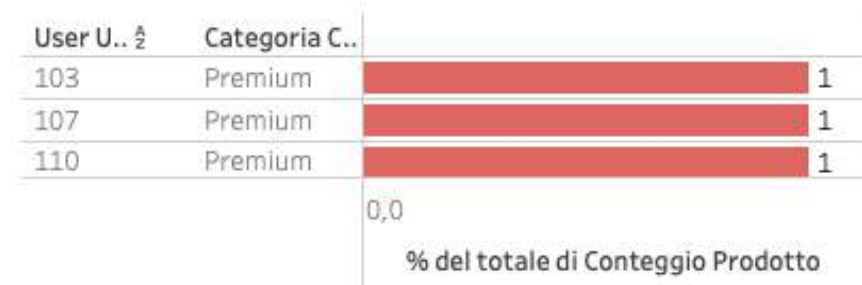
Paese - Categoria - Stelle assegnate al prodotto 2021

Country	Categ..	Stars (gruppo)				
		1	2	3	4	5
fr	Premium	697	1.140	946	1.064	
	Cheap	489	797	180	429	
	Average	335	78	262	768	
it	Premium	2.258	3.476	2.221	2.584	

Piattaforma pagamento utilizzata nel 2021



Utenti con fascia di prezzo cheap e con un solo prodotto nel sito



Categorie dell'utente che vende nel 2021



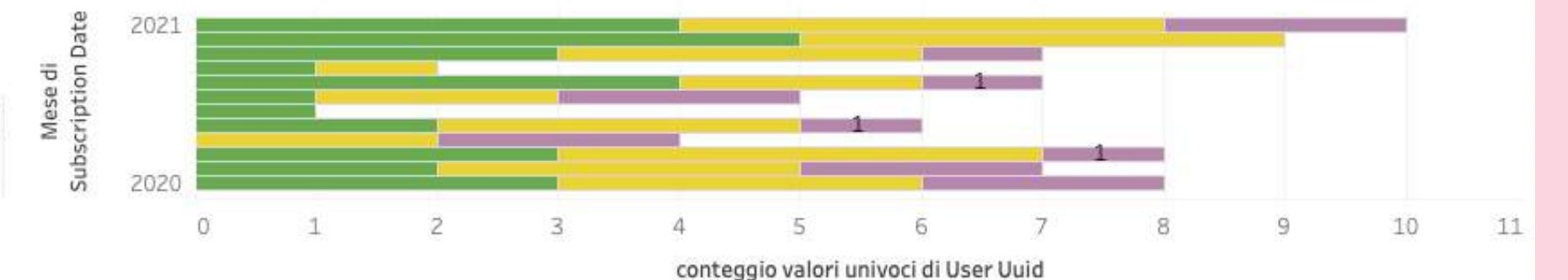
Andamento delle vendite 2021

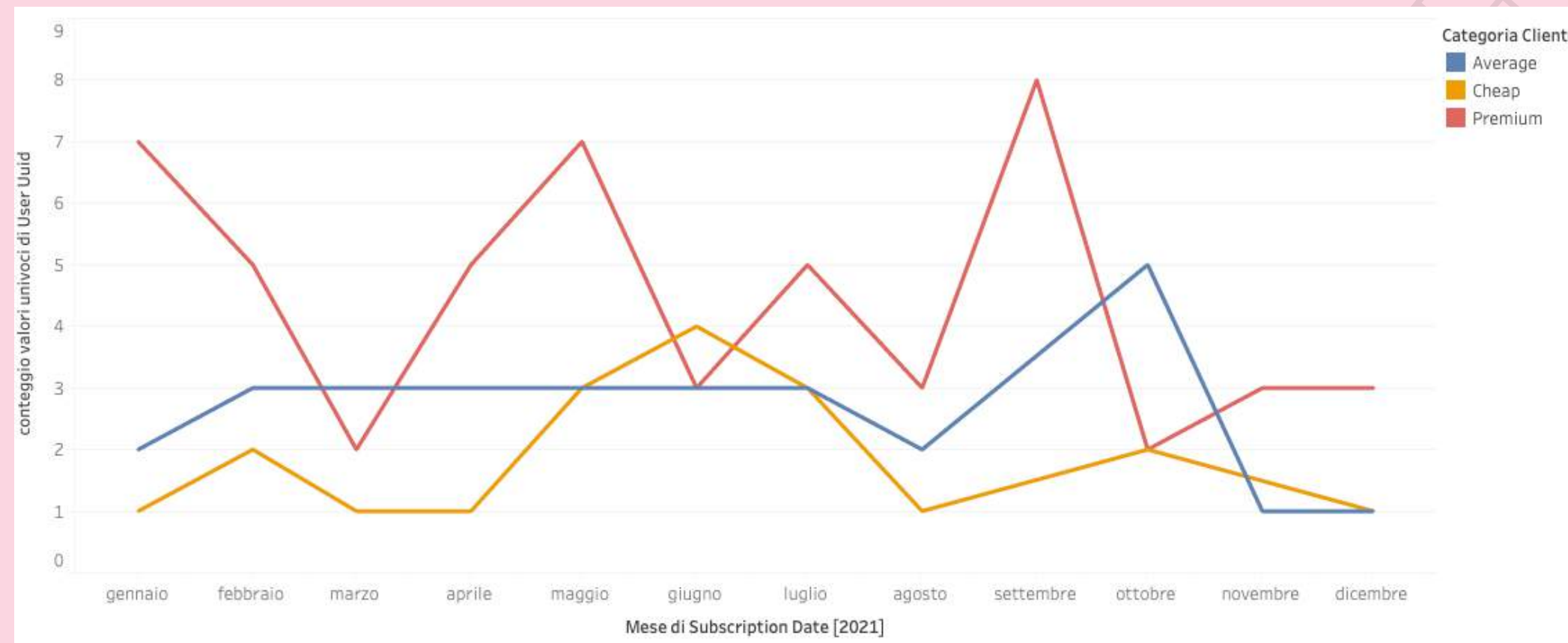


Utenti con fascia di prezzo premium e con più di 2 prodotti nel sito

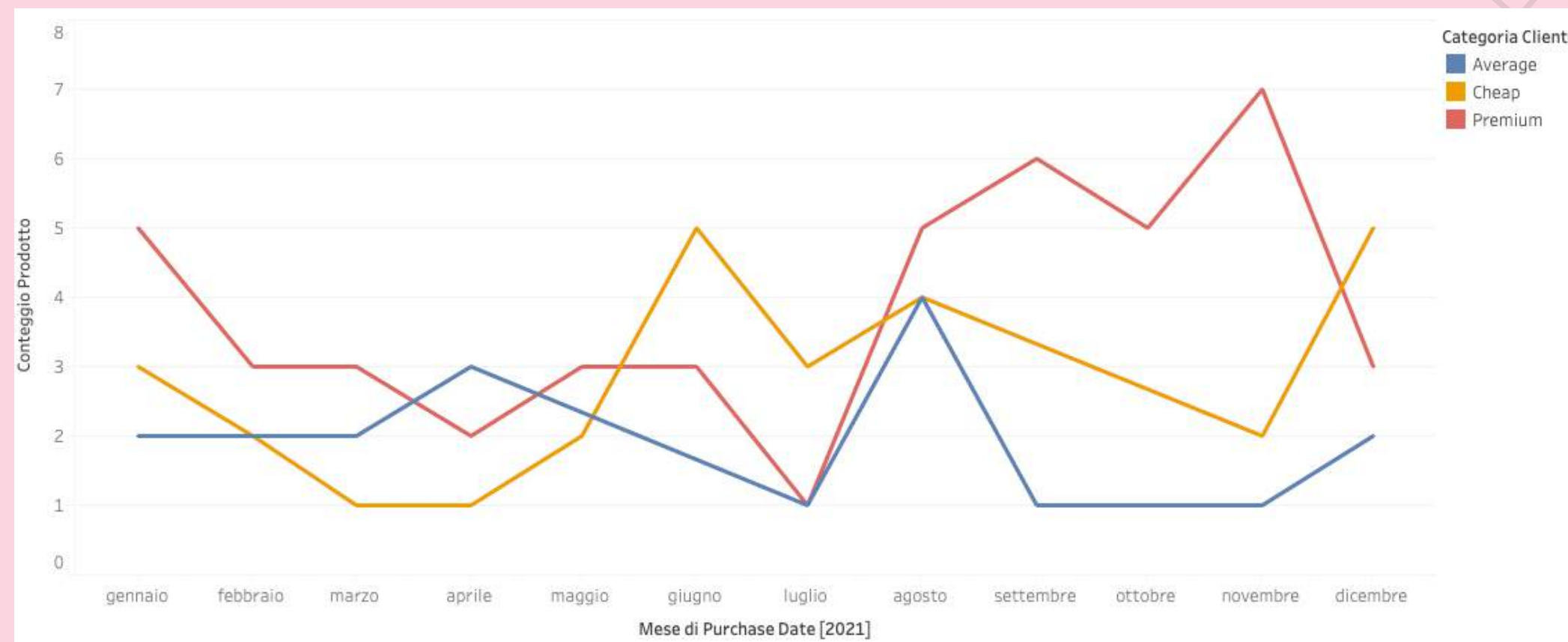


Mesi 2021 - Paesi - Utenti



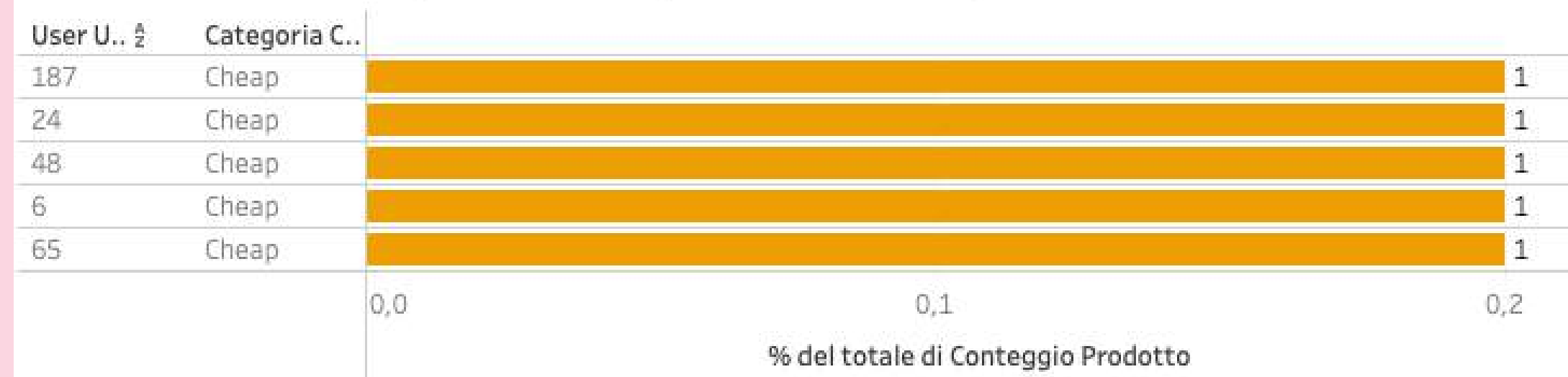


Maggio è stato il mese dove si sono iscritti più utenti all'interno del sito e con più valore.



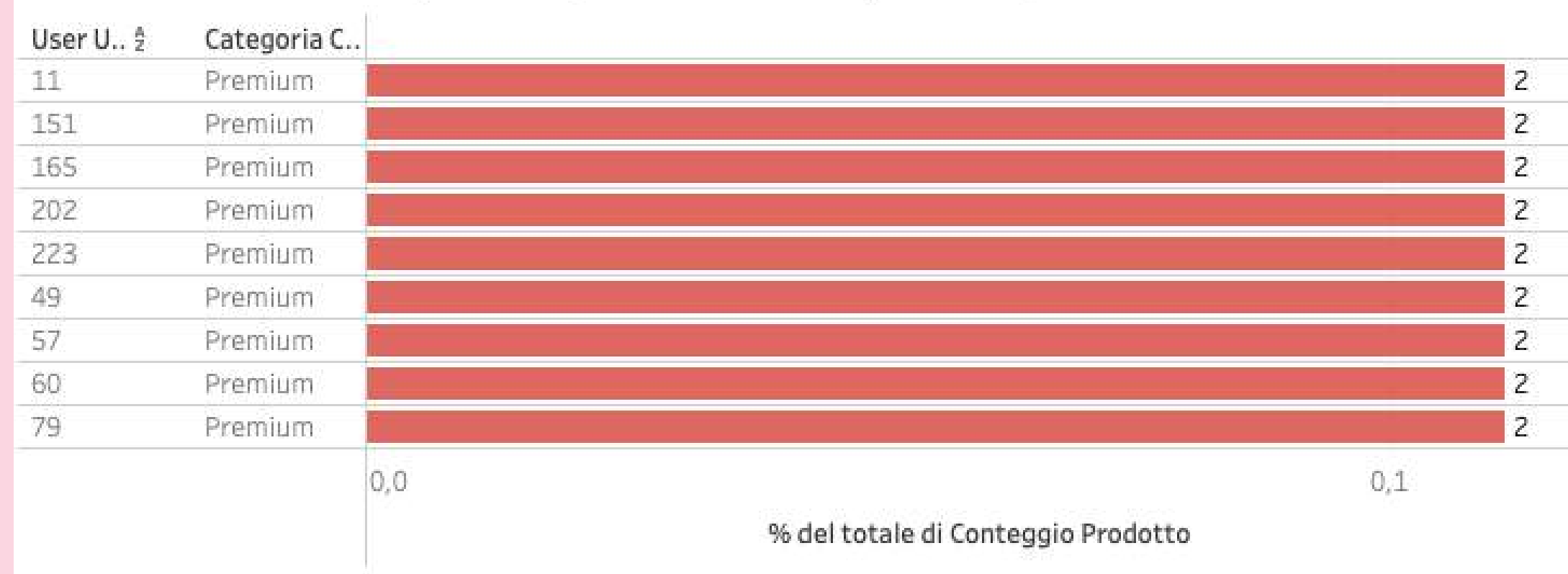
Nel mese di febbraio si sono venduti più prodotti con la fascia di prezzo premium.

Utenti con fascia di prezzo cheap e con un solo prodotto nel sito



Inserendo il filtro di un solo prodotto nel sito, (considerando solo il 2021) e con una fascia di prezzo cheap, possiamo risalire a questi utenti mostrati nel grafico. Inserendo il filtro di due o più prodotti nel sito, (considerando solo il 2021) e con una fascia di prezzo premium, possiamo risalire a questi utenti mostrati nel grafico.

Utenti con fascia di prezzo premium e con più di 2 prodotti nel sito



Per questi utenti si dovranno adottare delle strategie di fidelizzazione con l'obiettivo di mantenere questi suddetti utenti nel sito e invogliarli ad inserire prodotti nuovi.

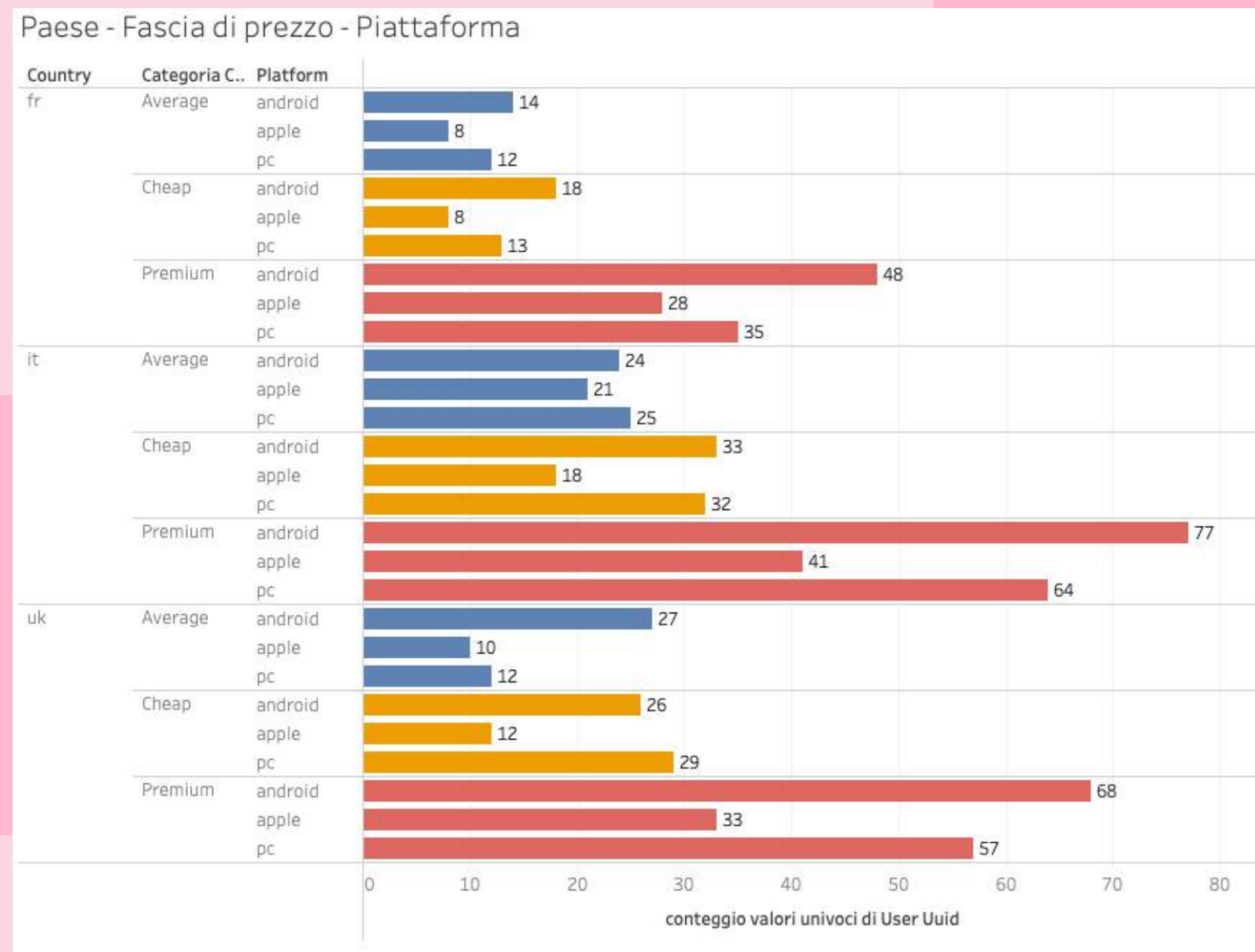


DANIELA SURHICEAN

CREAZIONE DI PERSONAS

Creiamo 3 personas:

- L'utente che fa per la maggior parte vendite costose (level=premium) lo chiameremo: Camille
- L'utente che fa principalmente vendite a prezzo nella media (level=average) lo chiameremo: Emily
- E infine, l'utente che soprattutto fa vendite convenienti (level=cheap) lo chiameremo: Gabriel





DANIELA SURHICEAN

Creazione di alcuni personaggi di utenti tipo



CAMILLE

PREMIUM

Utenti che hanno più del 70%
delle vendite sulla fascia
premium.



EMILY

AVERAGE

Utenti che hanno più del 70%
delle vendite sulla fascia
average.



GABRIEL

CHEAP

Utenti che hanno più del 70%
delle vendite sulla fascia
cheap.



DANIELA SURCHICEAN

Strategia

PROPOSTA

Utenti Average

Possiamo affermare che la categoria dove si effettuano più vendite è quella relativa alla cintura. La maggior parte degli utenti iscritti viene dall'Italia. I maggiori acquisti hanno una fascia di prezzo costosa e vengono effettuati per la maggior parte da piattaforme con sistema Android. Gli utenti iscritti unici nel dataset sono 241. La categoria che ha un punteggio di stelle più alto è quella relativa alla cintura. I punteggi medi con le stelle più alte sono riferiti a prodotti venduti da persone provenienti dal Regno Unito, seguito da Francia e Italia, nonostante che la maggior parte degli iscritti venga proprio dall'Italia.

Quindi la strategia che si potrebbe applicare è quella di concentrarsi sulla fidelizzazione degli utenti già presenti nel marketplace proponendo uno sconto sulle spedizioni degli oggetti. Inoltre, si potrebbe informare gli utenti sugli oggetti che si vendono di più come le cinture. Per le categorie che hanno poche vendite si potrebbe optare per uno sconto per chi acquista fornito dalla "piattaforma" stessa.

