DataSet Kaggle

1.0 Caricamento e Visualizzazione dei Dati dei Titoli Netflix

In [99]:
```
# Importa le librerie, carica il file csv nel dataframe e lo stampa
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
percorso_file_csv = "C:\\Users\\zetam\\Desktop\\2 Superiore\\Robotica\\netflix_titles.cs
df = pd.read_csv(percorso_file_csv)
print(df)
```

```
     show_id     type                  title         director  \
0         s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1         s2  TV Show          Blood & Water              NaN
2         s3  TV Show              Ganglands  Julien Leclercq
3         s4  TV Show  Jailbirds New Orleans              NaN
4         s5  TV Show           Kota Factory              NaN
...      ...      ...                    ...              ...
8802   s8803    Movie                 Zodiac    David Fincher
8803   s8804  TV Show            Zombie Dumb              NaN
8804   s8805    Movie             Zombieland  Ruben Fleischer
8805   s8806    Movie                  Zoom     Peter Hewitt
8806   s8807    Movie                Zubaan      Mozez Singh

                                                   cast        country  \
0                                                   NaN  United States
1     Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2     Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                   NaN            NaN
4     Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India
...                                                 ...            ...
8802  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...  United States
8803                                                NaN            NaN
8804  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...  United States
8805  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...  United States
8806  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...          India

              date_added  release_year rating   duration  \
0     September 25, 2021          2020  PG-13     90 min
1     September 24, 2021          2021  TV-MA  2 Seasons
2     September 24, 2021          2021  TV-MA   1 Season
3     September 24, 2021          2021  TV-MA   1 Season
4     September 24, 2021          2021  TV-MA  2 Seasons
...                  ...           ...    ...        ...
8802   November 20, 2019          2007      R    158 min
8803        July 1, 2019          2018  TV-Y7  2 Seasons
8804    November 1, 2019          2009      R     88 min
8805    January 11, 2020          2006     PG     88 min
8806       March 2, 2019          2015  TV-14    111 min

                                          listed_in  \
0                                     Documentaries
1       International TV Shows, TV Dramas, TV Mysteries
2     Crime TV Shows, International TV Shows, TV Act...
3                               Docuseries, Reality TV
4     International TV Shows, Romantic TV Shows, TV ...
...                                                 ...
8802                  Cult Movies, Dramas, Thrillers
8803            Kids' TV, Korean TV Shows, TV Comedies
8804                          Comedies, Horror Movies
8805              Children & Family Movies, Comedies
8806    Dramas, International Movies, Music & Musicals

                                            description
0     As her father nears the end of his life, filmm...
1     After crossing paths at a party, a Cape Town t...
2     To protect his family from a powerful drug lor...
3     Feuds, flirtations and toilet talk go down amo...
4     In a city of coaching centers known to train I...
...                                                 ...
8802  A political cartoonist, a crime reporter and a...
8803  While living alone in a spooky town, a young g...
8804  Looking to survive in a world taken over by zo...
8805  Dragged from civilian life, a former superhero...
8806  A scrappy but poor boy worms his way into a ty...
```

```
[8807 rows x 12 columns]
```

1.1 Identificazione del Tipo di Programma più Frequente nei Titoli Netflix

In [41]:
```
# Conta quante volte compare ogni tipo di programma e stampa quello con il numero maggio
# utilizzando il metodo idmax
import pandas as pd
percorso_file_csv = "C:\\Users\\zetam\\Desktop\\2 Superiore\\Robotica\\netflix_titles.cs
df = pd.read_csv(percorso_file_csv)
tipo_programma = df['type'].value_counts().idxmax()
print(tipo_programma)
```

Out[41]:
```
'Movie'
```

1.2 Conteggio dei Programmi Netflix per Anno di Rilascio

In [11]:
```
# Conta quanti programmi ci sono per ogni anno e stampa i numeri
import pandas as pd

anno_programma = df['release_year'].value_counts()
print(anno_programma)
```

```
release_year
2018    1147
2017    1032
2019    1030
2020     953
2016     902
        ...
1959       1
1925       1
1961       1
1947       1
1966       1
Name: count, Length: 74, dtype: int64
```

1.3 Identificazione dell'Anno con il Maggior Numero di Programmi Netflix

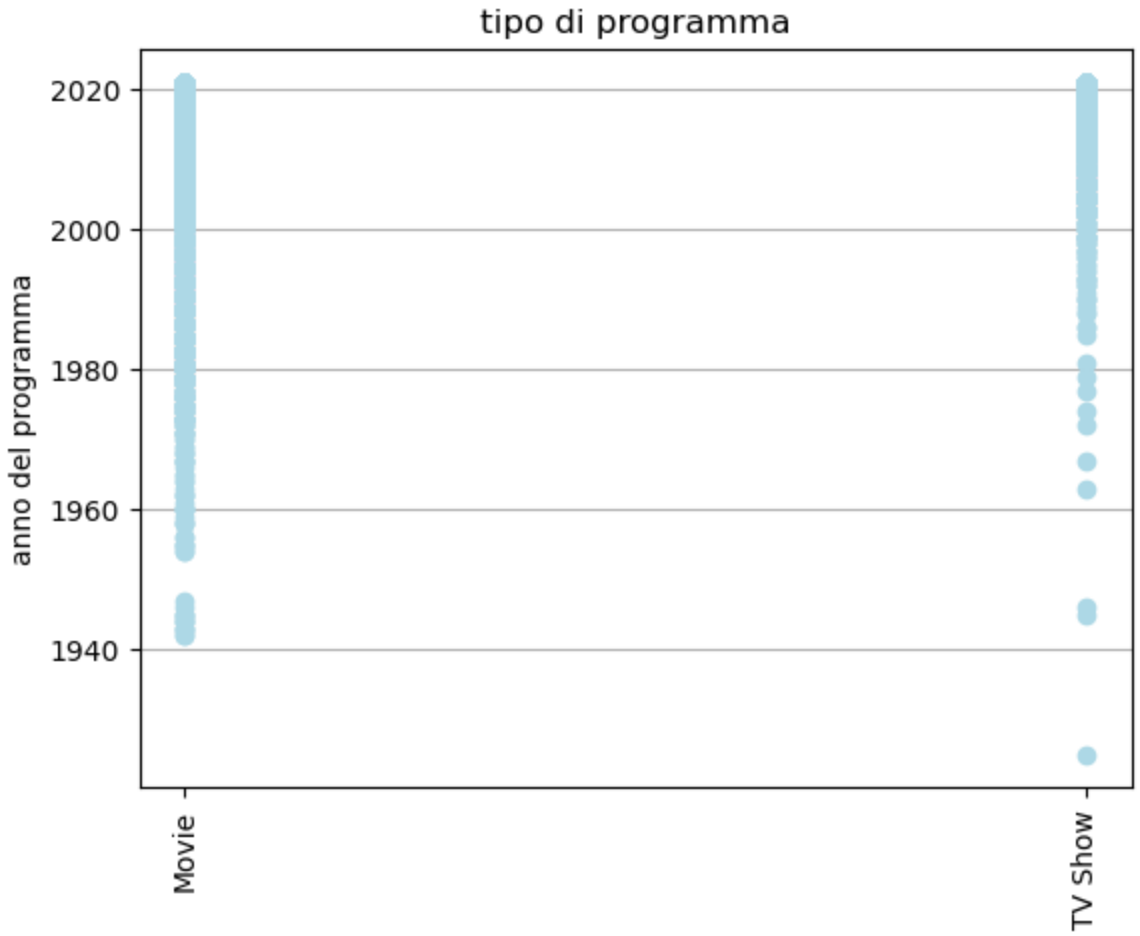
In [12]:
```
# Conta quanti programmi ci sono per ogni anno e stampa l'anno che ne ha di più
import pandas as pd
anno_prog = df['release_year'].value_counts().idxmax()
print(anno_prog)
```

```
2018
```

1.4 Visualizzazione della Distribuzione dei Tipi di Programmi Netflix nel Corso degli Anni

In [38]:
```
# Grafico a dispersione che mostra la distribuzione dei tipi di programmi negli anni
import matplotlib.pyplot as plt
plt.plot(df['type'],df['release_year'], marker='o', linestyle='', color='lightblue')
plt.title('tipo di programma')
plt.ylabel('anno del programma')
tation=90)
```

```
plt.grid(True, axis="y")
plt.show()
```



1.5 Identificazione e Stampa delle Righe con Valori Mancanti nel DataFrame

In [40]:
```
# identifica le righe con valori mancanti e lo stampa
righe_con_dati_mancanti = df[df.isnull().any(axis=1)]
righe_con_dati_mancanti
```

Out[40]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	D
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	T
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	T
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	
...	...	...	...	...	...	...	...	...	...	...	
8795	s8796	TV Show	Yu-Gi-Oh! Arc-V	NaN	Mike Liscio, Emily Bauer, Billy Bob Thompson, ...	Japan, Canada	May 1, 2018	2015	TV-Y7	2 Seasons	A
8796	s8797	TV Show	Yunus Emre	NaN	Gökhan Atalay, Payidar Tüfekçioglu, Baran Akbu...	Turkey	January 17, 2017	2016	TV-PG	2 Seasons	T
8797	s8798	TV Show	Zak Storm	NaN	Michael Johnston, Jessica Gee-George, Christin...	United States, France, South Korea, Indonesia	September 13, 2018	2016	TV-Y7	3 Seasons	
8800	s8801	TV Show	Zindagi Gulzar Hai	NaN	Sanam Saeed, Fawad Khan, Ayesha Omer, Mehreen ...	Pakistan	December 15, 2016	2012	TV-PG	1 Season	
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	

2475 rows × 12 columns

1.6 Calcolo e Stampa del Numero Totale di Righe con Dati Mancanti

In [42]:
```
# calcola il numero totale di righe con dati mancanti e lo assegna alla variabile tot_da
alla variabile tot_dati_mancanti
tot_dati_mancanti = righe_con_dati_mancanti.shape[0]
tot_dati_mancanti
```

Out[42]:
```
3475
```

1.7 Identificazione e Rimozione delle Righe con Valori Mancanti dal DataFrame

In [43]:
```
# identifica le righe con valori mancanti e le rimuove dal dataframe df1, poi lo stampa
df1=df.dropna(inplace=False)
df1
```

Out[43]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	TV-MA	125 min
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	September 24, 2021	2021	TV-14	9 Seasons
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-13	104 min
12	s13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel, ...	Germany, Czech Republic	September 23, 2021	2021	TV-MA	127 min
24	s25	Movie	Jeans	S. Shankar	Prashanth, Aishwarya Rai Bachchan, Sri Lakshmi...	India	September 21, 2021	1998	TV-14	166 min
...	...	...	...	...	...	...	...	...	...	...
8801	s8802	Movie	Zinzana	Majid Al Ansari	Ali Suliman, Saleh Bakri, Yasa, Ali Al-Jabri, ...	United Arab Emirates, Jordan	March 9, 2016	2015	TV-MA	96 min
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min

5332 rows × 12 columns

1.8 Creazione di una Matrice Booleana per Indicare Valori Mancanti nel DataFrame

In [44]:
```
# Utilizza il metodo isnull() sul DataFrame df  per creare una matrice booleana (vslori
# missing_matrix che indica se c'è un valore mancante (NaN) in ciascuna posizione del Da
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
missing_matrix = df.isnull()
missing_matrix
```

Out[44]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	desc
0	False	False	False	False	True	False	False	False	False	False	False	
1	False	False	False	True	False	False	False	False	False	False	False	
2	False	False	False	False	False	True	False	False	False	False	False	
3	False	False	False	True	True	True	False	False	False	False	False	
4	False	False	False	True	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	
8802	False	False	False	False	False	False	False	False	False	False	False	
8803	False	False	False	True	True	True	False	False	False	False	False	
8804	False	False	False	False	False	False	False	False	False	False	False	
8805	False	False	False	False	False	False	False	False	False	False	False	
8806	False	False	False	False	False	False	False	False	False	False	False	

8807 rows × 12 columns

1.9 Selezione e Stampa dei Nomi delle Colonne Numeriche del DataFrame

In [49]:
```
# selezioa le colonne del Df che contengono dati numerici e le mette nella variabile num
numeric_cols = df.select_dtypes(include=['number'])
numeric_cols.columns
```

Out[49]:
```
Index(['release_year'], dtype='object')
```

2.0 Calcolo del Numero di Valori Mancanti per Ogni Colonna in un DataFrame

In [51]:
```
# calcola il numero di valori mancanti per ogni colonna
df.isnull().sum()
```

Out[51]:
```
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
description        0
dtype: int64
```

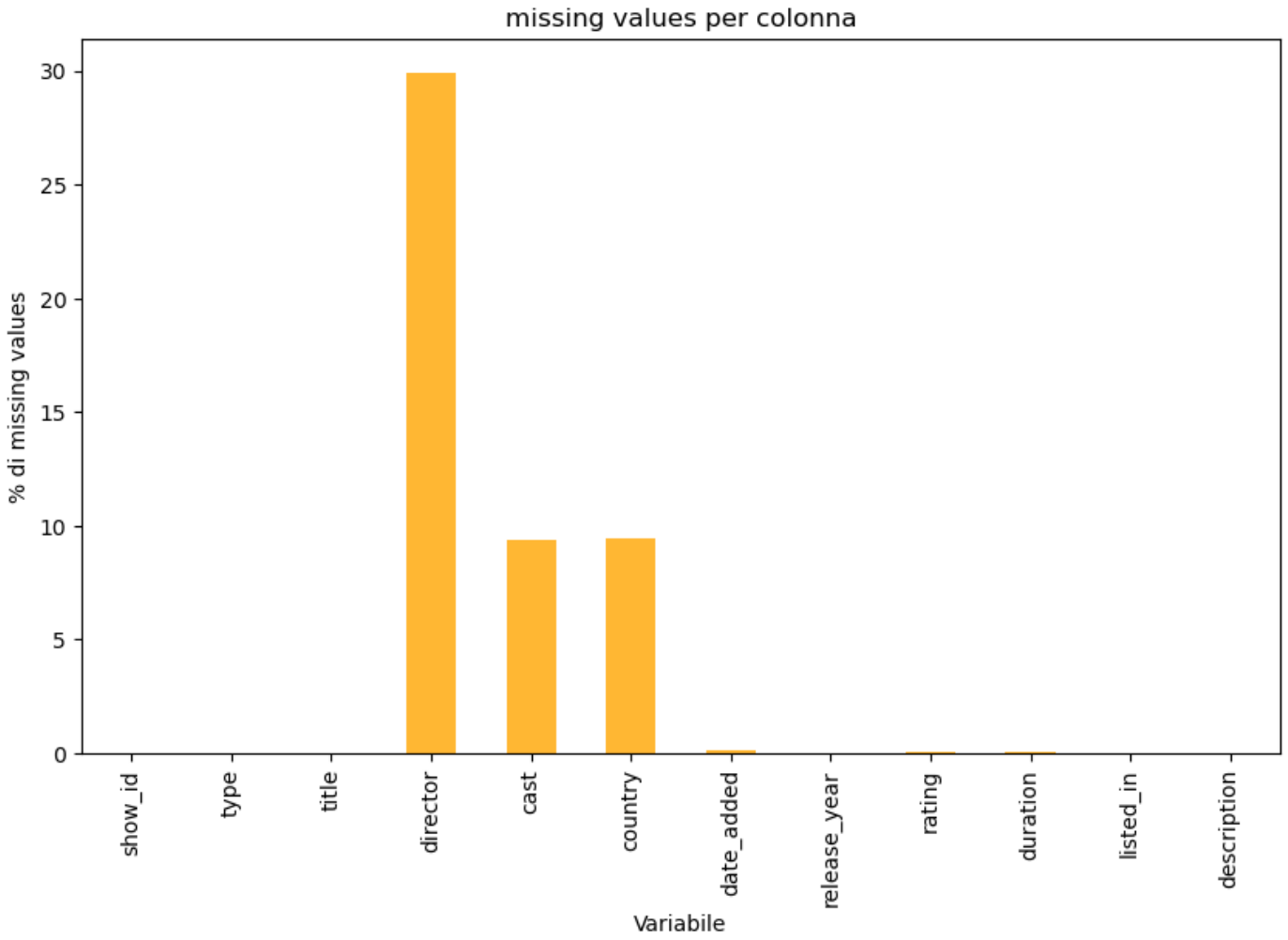
2.1 Calcolo della Percentuale di Valori Mancanti per Ogni Colonna in un DataFrame

In [53]:
```
# calcola per ogni colonna la percentuale di valori mancanti su tutte le righe del dataf
missing_percent = df.isnull().sum() / len(df) * 100
missing_percent
```

Out[53]:
```
show_id          0.000000
type             0.000000
title            0.000000
director        29.908028
cast             9.367549
country          9.435676
date_added       0.113546
release_year     0.000000
rating           0.045418
duration         0.034064
listed_in        0.000000
description      0.000000
dtype: float64
```

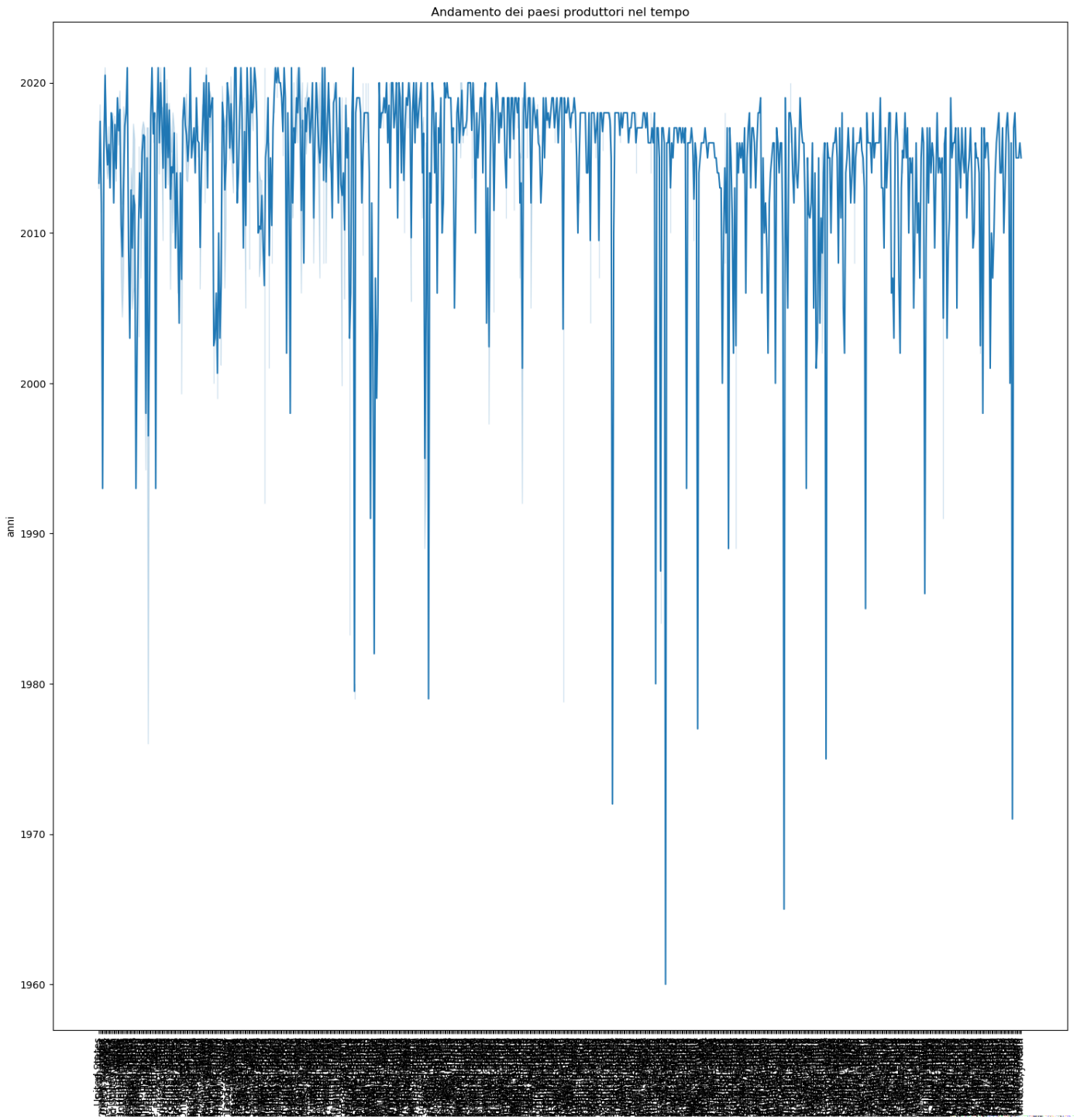
2.2 Calcolo della Percentuale di Valori Mancanti per Ogni Colonna in un DataFrame e Creazione del Grafico a Barre Corrispondente

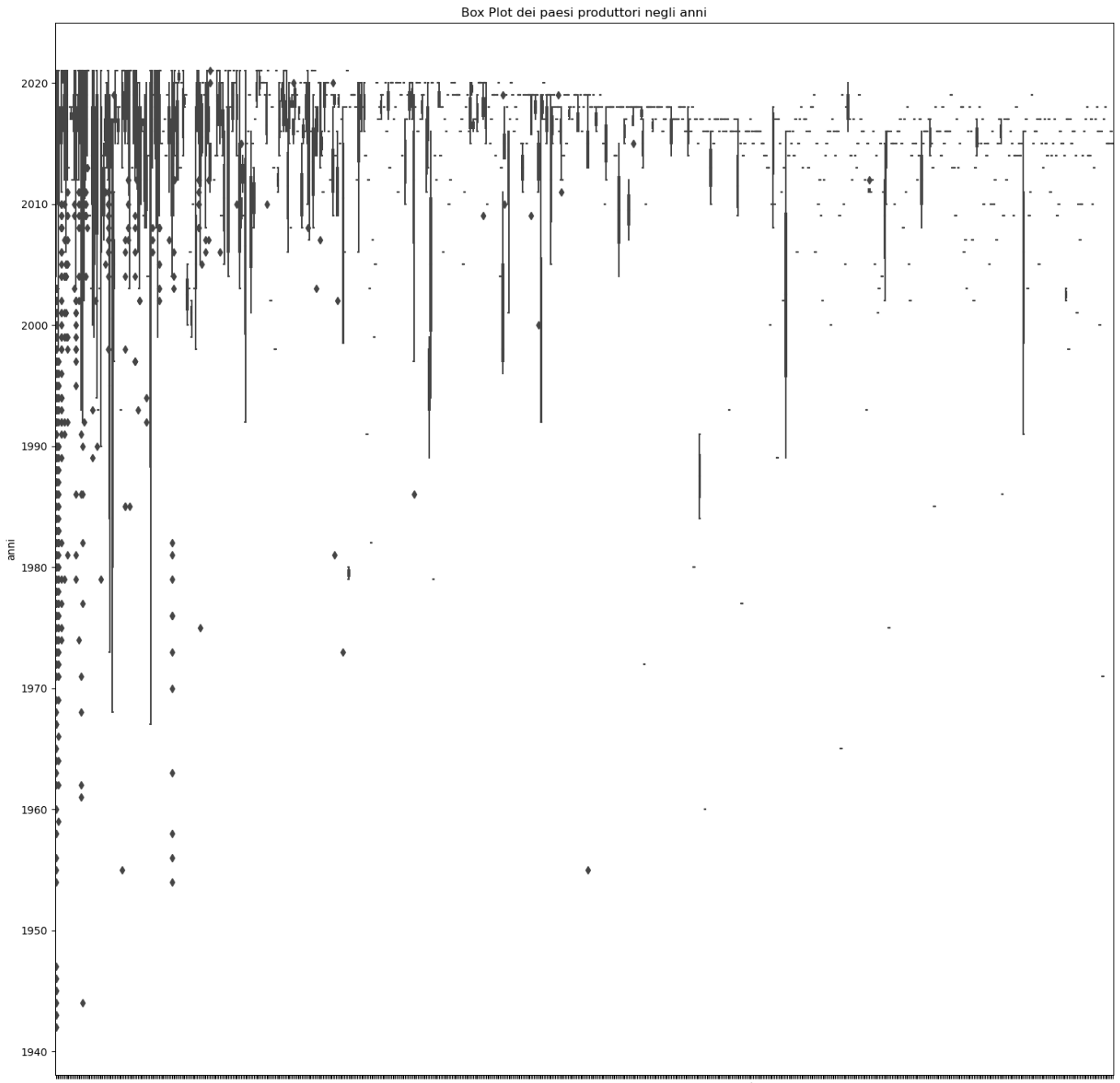
In [62]:
```
# calcola per ogni colonna la percentuale di valori mancanti su tutte le righe del dataf
missing_percent= (df.isnull().sum()) / len(df) * 100
plt.figure(figsize=(10,6))
missing_percent.plot(kind='bar', color='orange', alpha=0.8)
plt.xlabel('Variabile')
plt.ylabel('% di missing values')
plt.title('missing values per colonna')
plt.xticks(rotation=90)
plt.show()
```



2.3 Visualizzazione dell'Andamento dei Paesi Produttori nel Tempo tramite un Grafico Lineare e un Box Plot

In [66]:
```
# Visualizza un grafico dei paesi produttori nel tempo
plt.figure(figsize=(2^16, 2^16))
sns.lineplot(x='country', y='release_year', data=df)
plt.title('Andamento dei paesi produttori nel tempo')
plt.xlabel('country')
plt.ylabel('anni')
plt.xticks(rotation=90)
plt.show()
# Visualizza una box plot dei paesi produttori nel tempo
plt.figure(figsize=(2^16, 2^16))
sns.boxplot(x='country', y='release_year', data=df)
plt.title('Box Plot dei paesi produttori negli anni')
plt.xlabel('paesi')
plt.ylabel('anni')
plt.xticks(rotation=90)
plt.show()
```





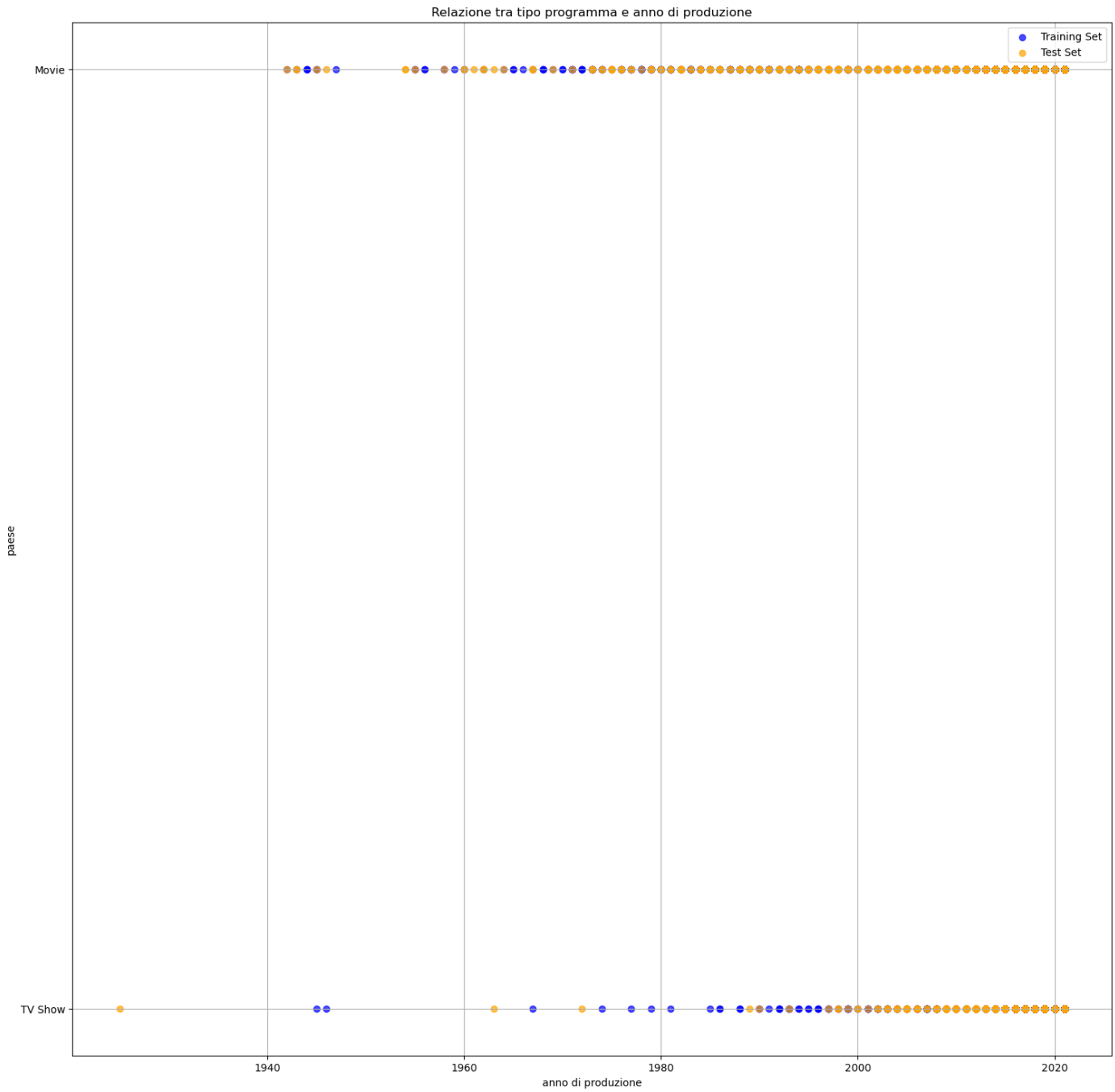
paesi

Austr

2.4 Suddivisione del Dataset in Training e Test Set, Creazione di un Grafico a Dispersione e Stampa delle Dimensioni dei Set

In [70]:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
# Suddivisione del dataset in training set (70%) e test set (30%)
X_train, X_test, y_train, y_test = train_test_split(df['release_year'],df['type'],test_s
# Creazione di un grafico a dispersione
plt.figure(figsize=(2^16, 2^16))
plt.scatter(X_train, y_train, label='Training Set', color='blue', alpha=0.7)
plt.scatter(X_test, y_test, label='Test Set', color='orange', alpha=0.7)
plt.xlabel('anno di produzione')
plt.ylabel('paese')
plt.title('Relazione tra tipo programma e anno di produzione')
plt.legend()
plt.grid(True)
plt.show()
# Stampare le dimensioni dei training set e test set
print("Dimensioni del Training Set (tipo programma e anno di produzione):",X_train.shape
print("Dimensioni del Test Set (tipo programma e anno di produzione):",X_test.shape, y_t
```




```
Dimensioni del Training Set (tipo programma e anno di produzione): (6164,) (6164,)
Dimensioni del Test Set (tipo programma e anno di produzione): (2643,) (2643,)
```

2.5 Creazione di Tre Subset Casuali da un DataFrame

In [100…

```
# Creare tre subset di dimensioni simili
# primo subset: campione causale di 1/3 delle righe del df di partenza
subset1 = df.sample(frac=1/3)
# stampa il numero di righe del subset1
l1=len(subset1)
print(l1)
df = df.drop(subset1.index)
# secondo subset: campione casuale con metà delle righe rimanenti (la metà dei 2/3 riman
subset2 = df.sample(frac=1/2)
# stampa il numero di righe del subset2
l2=len(subset2)
print(l2)
df = df.drop(subset2.index)
# terzo subset: le righe restanti
subset3 = df
umero di righe del subset3
```

```
l3=len(subset3)
print(l3)
```

```
2936
2936
2935
```

2.6 Calcolo delle Percentuali dei Valori Unici per il Paese nel Subset1

In [101…

```
percentuali_subset1 = subset1['country'].value_counts(normalize=True)
percentuali_subset1
```

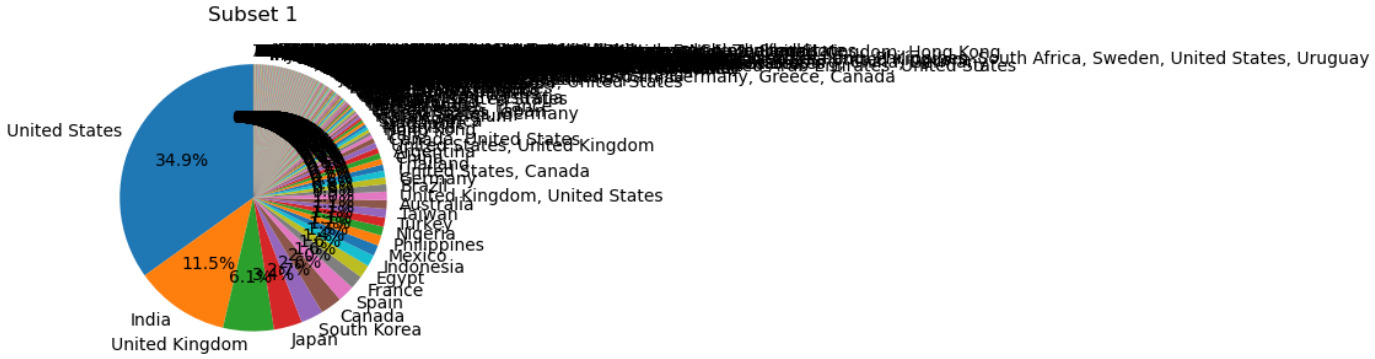
Out[101]:

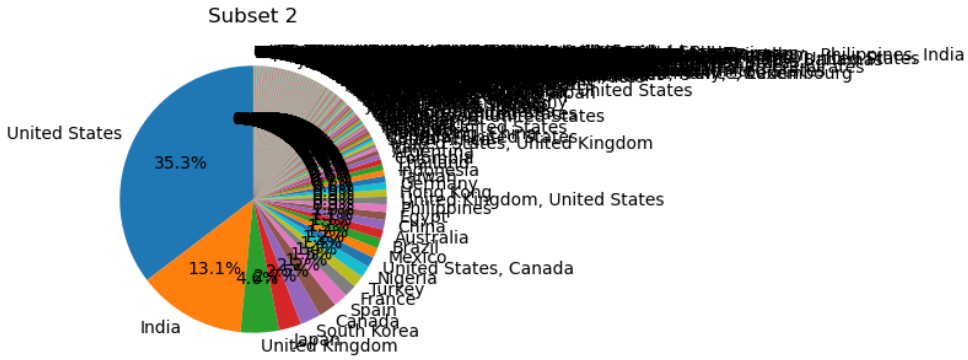
```
country
United States                                0.348907
India                                        0.114544
United Kingdom                               0.061417
Japan                                        0.034288
South Korea                                  0.027129
                                               ...
Canada, Australia                            0.000377
South Korea, Canada, United States, China    0.000377
Norway, United States                        0.000377
Australia, New Zealand, United States        0.000377
Taiwan, China, France, United States         0.000377
Name: proportion, Length: 304, dtype: float64
```

2.7 Visualizzazione delle Distribuzioni dei Valori ‘Country’ nei Tre Subset con Grafici a Torta

In [103…

```
percentuali_subset1 = subset1['country'].value_counts(normalize=True)
percentuali_subset2 = subset2['country'].value_counts(normalize=True)
percentuali_subset3 = subset3['country'].value_counts(normalize=True)
# Creare i grafici a torta
fig, axs = plt.subplots(3, 1, figsize=(6, 12))
# Subset 1
axs[0].pie(percentuali_subset1, labels=percentuali_subset1.index, autopct='%1.1f%%', sta
axs[0].set_title('Subset 1')
# Subset 2
axs[1].pie(percentuali_subset2, labels=percentuali_subset2.index, autopct='%1.1f%%', sta
axs[1].set_title('Subset 2')
# Subset 3
axs[2].pie(percentuali_subset3, labels=percentuali_subset3.index, autopct='%1.1f%%', sta
axs[2].set_title('Subset 3')
# Mostrare il grafico
plt.show()
```





2.8 Divisione dei Subset in Training e Test Set e Visualizzazione delle Distribuzioni dei Valori 'Country' con Grafici a Torta

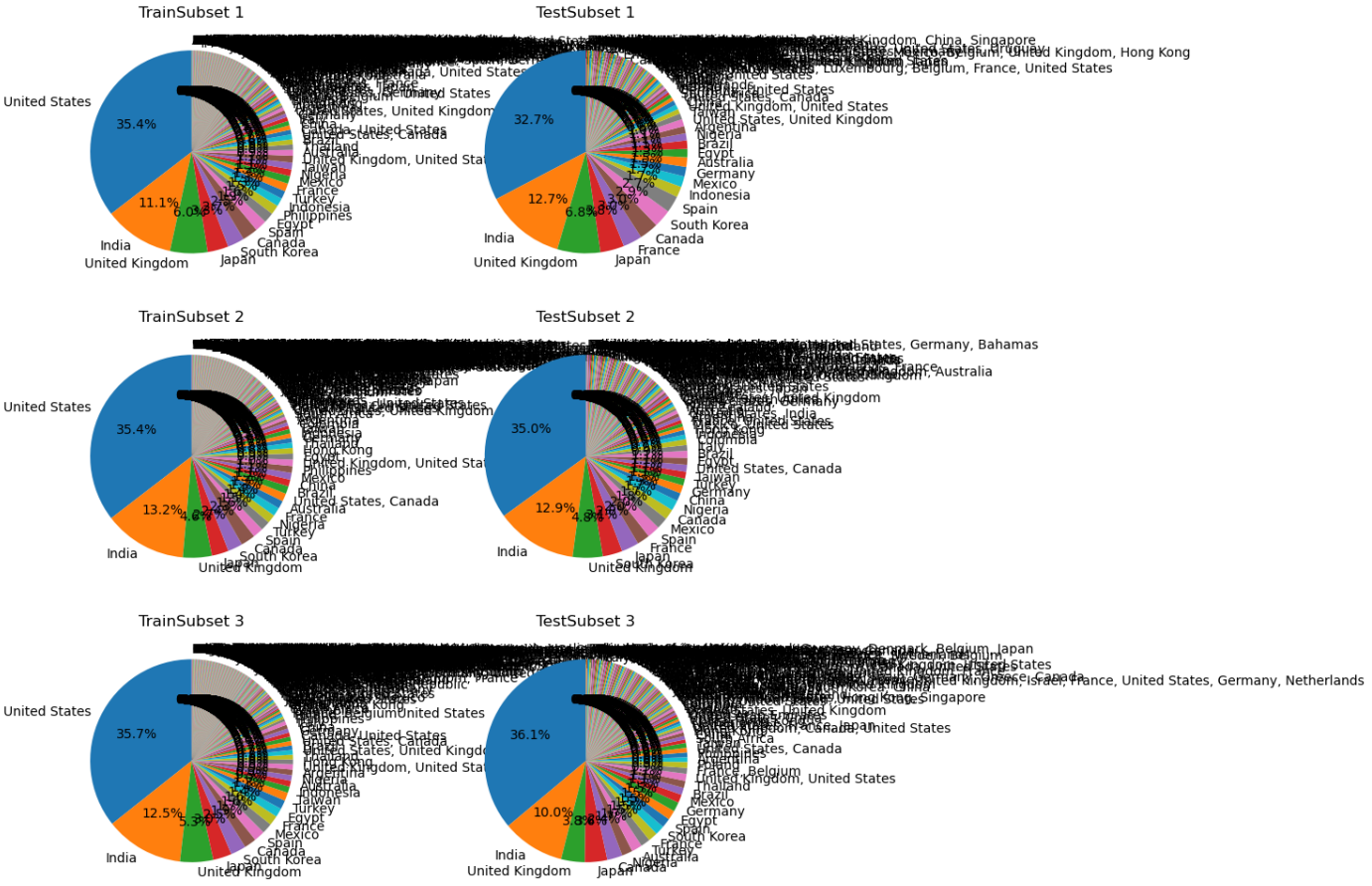
In [108…

```
# Dividere ciascun subset in training set e test set
train_subset1, test_subset1 = train_test_split(subset1, test_size=0.2,random_state=42)
train_subset2, test_subset2 = train_test_split(subset2, test_size=0.2,random_state=42)
train_subset3, test_subset3 = train_test_split(subset3, test_size=0.2,random_state=42)
# Creare il grafico con 6 torte
fig, axs = plt.subplots(3, 2, figsize=(10, 12))
# Funzione per disegnare una torta con etichette
def draw_pie(ax, data, title):
    ax.pie(data, labels=data.index, autopct='%1.1f%%', startangle=90)
    ax.set_title(title)
# Prima riga di torte (Subset 1)
draw_pie(axs[0, 0], train_subset1['country'].value_counts(normalize=True), 'TrainSubset

draw_pie(axs[0, 1], test_subset1['country'].value_counts(normalize=True), 'TestSubset 1'
# Seconda riga di torte (Subset 2)
draw_pie(axs[1, 0], train_subset2['country'].value_counts(normalize=True), 'TrainSubset
draw_pie(axs[1, 1], test_subset2['country'].value_counts(normalize=True), 'TestSubset 2'
# Terza riga di torte (Subset 3)
draw_pie(axs[2, 0], train_subset3['country'].value_counts(normalize=True), 'TrainSubset
draw_pie(axs[2, 1], test_subset3['country'].value_counts(normalize=True), 'TestSubset 3'
```

```
# Regolare lo spaziamento tra i subplots
plt.tight_layout()
# Mostrare il grafico
plt.show()
```

```
C:\Users\zetam\AppData\Local\Temp\ipykernel_4484\827625130.py:22: UserWarning: Tight lay
out not applied. tight_layout cannot make axes width small enough to accommodate all axe
s decorations
  plt.tight_layout()
```



2.9 Identificazione degli Outliers nell'Anno di Rilascio

In [121…

```
import pandas as pd
import matplotlib.pyplot as plt
# Lista con outliers da entrambi i lati
# Calcola la media e la deviazione standard
mean_value = df['release_year'].mean()
print('media anno:')
print(mean_value)
std_dev = df['release_year'].std()
print('deviazione standard:')
print(std_dev)
# Identifica gli outliers considerando ±3 * dev_std dalla media
outliers = df[(df['release_year'] > mean_value + 3 * std_dev) | (df['release_year'] < me
outliers
```

```
media anno:
2014.2047700170358
deviazione standard:
8.581060874548479
```

Out[121]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration
155	s156	Movie	Labyrinth	Jim Henson	David Bowie, Jennifer Connelly, Frank Oz, Kevi...	United Kingdom, United States	September 1, 2021	1986	PG	101 min
166	s167	Movie	Once Upon a Time in America	Sergio Leone	Robert De Niro, James Woods, Elizabeth McGover...	Italy, United States	September 1, 2021	1984	R	229 min
529	s530	Movie	Return of the Prodigal Son	Youssef Chahine	Majida El Roumi, Souheir El Morshidy, Shoukry ...	Egypt	July 6, 2021	1976	TV-MA	124 min
670	s671	Movie	Mobile Suit Gundam II: Soldiers of Sorrow	Yoshiyuki Tomino, Yoshikazu Yasuhiko	Toru Furuya, Shuichi Ikeda, Hirotaka Suzuoki, ...	NaN	June 19, 2021	1981	TV-14	133 min
1126	s1127	Movie	My Fair Lady	George Cukor	Audrey Hepburn, Rex Harrison, Stanley Holloway...	United States	April 1, 2021	1964	G	173 min
...	...	...	...	...	...	...	...	...	...	...
8569	s8570	Movie	The Young Vagabond	Sze Yu Lau	Chia-Hui Liu, Wong Yu, Jason Pai Piao, Lung We...	Hong Kong	August 16, 2018	1985	TV-14	85 min
8635	s8636	Movie	True Grit	Henry Hathaway	John Wayne, Glen Campbell, Kim Darby, Jeremy S...	United States	January 1, 2020	1969	G	128 min
8640	s8641	Movie	Tunisian Victory	Frank Capra, John Huston, Hugh Stewart, Roy Bo...	Burgess Meredith	United States, United Kingdom	March 31, 2017	1944	TV-14	76 min
8660	s8661	Movie	Undercover: How to Operate Behind Enemy Lines	John Ford	NaN	United States	March 31, 2017	1943	TV-PG	61 min
8739	s8740	Movie	Why We Fight: The Battle of Russia	Frank Capra, Anatole Litvak	NaN	United States	March 31, 2017	1943	TV-PG	82 min

74 rows × 12 columns

In [ ]: