Netflix_userbase_data_analysis

October 25, 2024

[2]:
```
#import the libraries

import pandas as pd
import sqlite3
import matplotlib.pyplot as plt
from sqlalchemy import create_engine
from streamlit import dataframe
import seaborn as sns
engine = create_engine('sqlite:///netflix_users.db',echo=False)
```

[3]:
```
df = pd.read_csv('netflix_userbase.csv')
df.head()
```

[3]:
```
   User ID Subscription Type  Monthly Revenue Join Date Last Payment Date  \
0        1             Basic               10  15-01-22          10-06-23
1        2           Premium               15  05-09-21          22-06-23
2        3          Standard               12  28-02-23          27-06-23
3        4          Standard               12  10-07-22          26-06-23
4        5             Basic               10  01-05-23          28-06-23

          Country  Age  Gender      Device Plan Duration
0   United States   28    Male  Smartphone       1 Month
1          Canada   35  Female      Tablet       1 Month
2  United Kingdom   42    Male    Smart TV       1 Month
3       Australia   51  Female      Laptop       1 Month
4         Germany   33    Male  Smartphone       1 Month
```

[4]:
```
df.isnull().sum()
df.columns = df.columns.str.replace(' ', '')
```

#Transfering the dataframe df into netflix_users.rb

[5]:
```
df.to_sql('netflix_userbase', con=engine, if_exists='replace')
```

[5]:
```
2500
```

[6]:
```
query_1 = "SELECT * FROM netflix_userbase"
```

```
pd.read_sql(query_1, con=engine)
```

[6]:
```
      index  UserID SubscriptionType  MonthlyRevenue  JoinDate  \
0         0       1            Basic              10  15-01-22
1         1       2          Premium              15  05-09-21
2         2       3         Standard              12  28-02-23
3         3       4         Standard              12  10-07-22
4         4       5            Basic              10  01-05-23
…     …     …              …             …       …
2495   2495    2496          Premium              14  25-07-22
2496   2496    2497            Basic              15  04-08-22
2497   2497    2498         Standard              12  09-08-22
2498   2498    2499         Standard              13  12-08-22
2499   2499    2500            Basic              15  13-08-22

     LastPaymentDate         Country  Age  Gender      Device PlanDuration
0           10-06-23   United States   28    Male  Smartphone      1 Month
1           22-06-23          Canada   35  Female      Tablet      1 Month
2           27-06-23  United Kingdom   42    Male    Smart TV      1 Month
3           26-06-23       Australia   51  Female      Laptop      1 Month
4           28-06-23         Germany   33    Male  Smartphone      1 Month
…               …         …  …    …       …          …
2495        12-07-23           Spain   28  Female    Smart TV      1 Month
2496        14-07-23           Spain   33  Female    Smart TV      1 Month
2497        15-07-23   United States   38    Male      Laptop      1 Month
2498        12-07-23          Canada   48  Female      Tablet      1 Month
2499        12-07-23   United States   35  Female    Smart TV      1 Month

[2500 rows x 11 columns]
```

0.1 What is the count of Users per Subscription Type?*

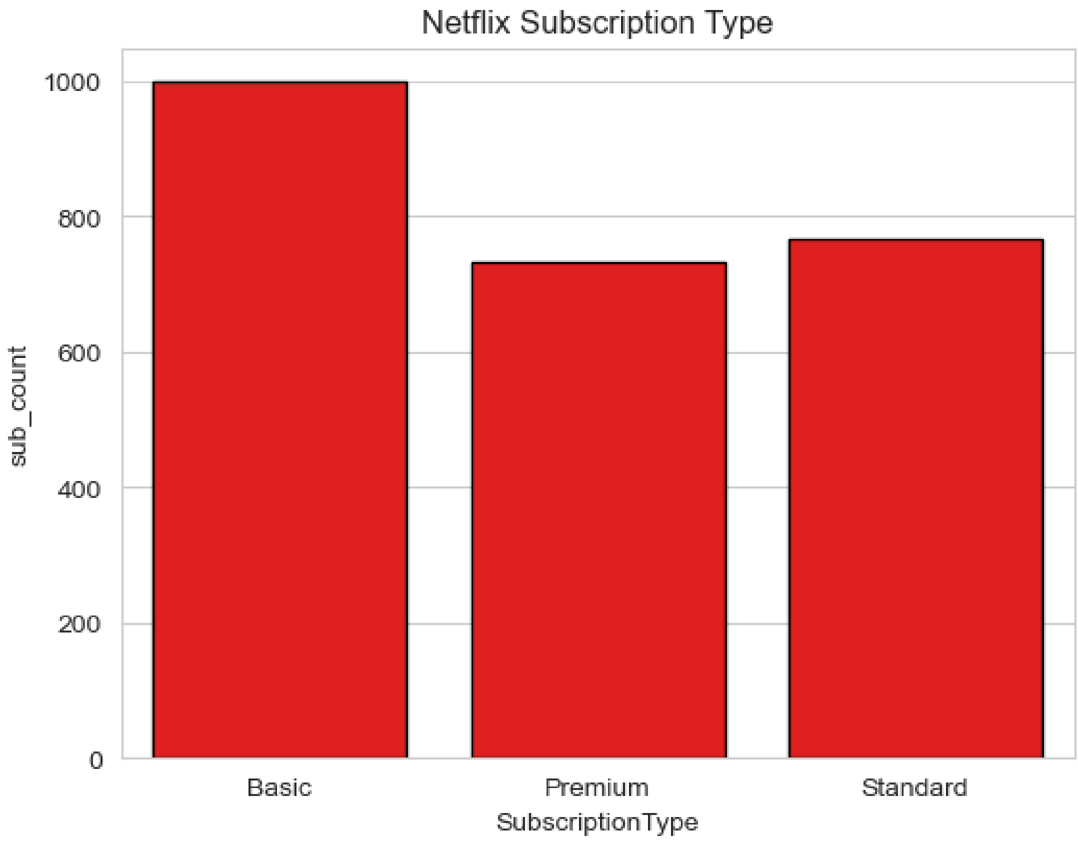
[7]:
```
query_2 = """ SELECT SubscriptionType,COUNT(*) AS sub_count FROM ␣
 ↪netflix_userbase
                GROUP BY SubscriptionType """
subscription_df = pd.read_sql(query_2, con=engine)
```

[8]:
```
subscription_df
```

[8]:
```
  SubscriptionType  sub_count
0            Basic        999
1          Premium        733
2         Standard        768
```

[9]:
```
sns.barplot(x='SubscriptionType', y='sub_count',␣
 ↪data=subscription_df,color='red',edgecolor='black')
plt.title("Netflix Subscription Type")
```

```
[9]: Text(0.5, 1.0, 'Netflix Subscription Type')
```



Netflix has more basic users than any other subscription type.

0.2 What is the count of users per device type?

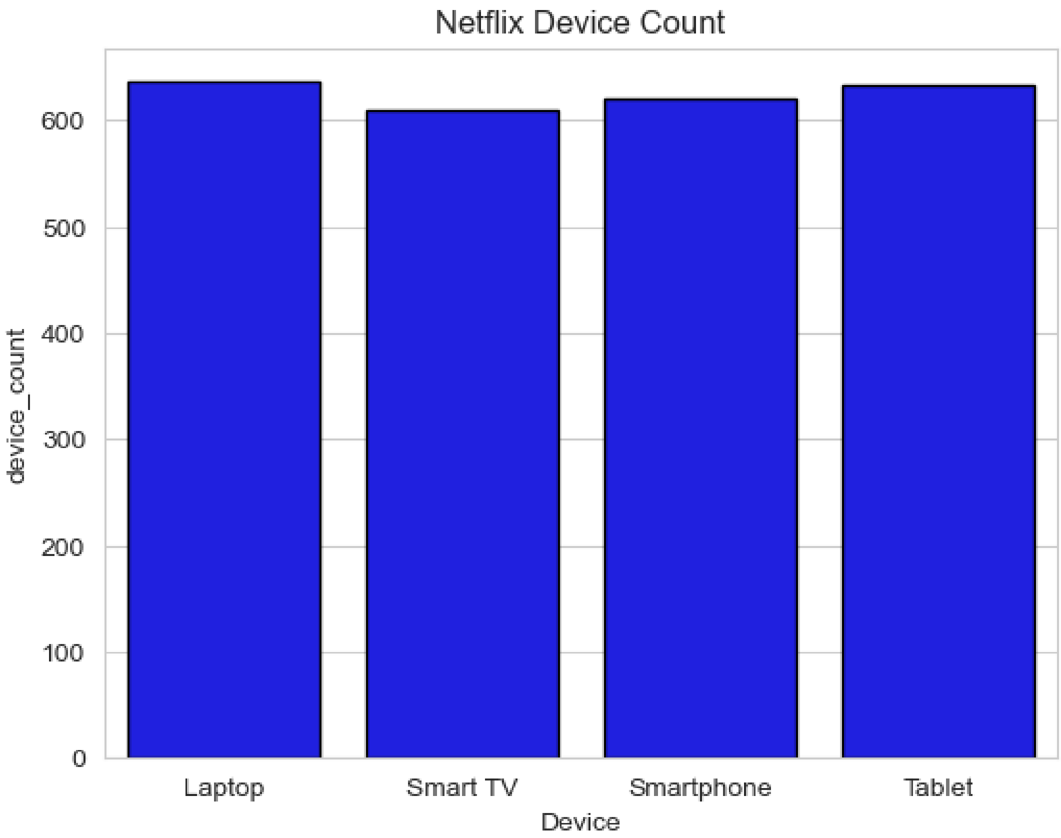
```
[10]: query_3 = """SELECT Device,COUNT(*) AS device_count FROM  netflix_userbase
                  GROUP BY Device """

      device_df = pd.read_sql(query_3, con=engine)
      device_df.head()
```

```
[10]:        Device  device_count
      0      Laptop           636
      1    Smart TV           610
      2  Smartphone           621
      3      Tablet           633
```

```
[11]: sns.barplot(x='Device', y='device_count',␣
       ↪data=device_df,color='blue',edgecolor='black')
```

```
plt.title("Netflix Device Count")
```

```
[11]: Text(0.5, 1.0, 'Netflix Device Count')
```



Netflix is watched on laptops more than any other medium.

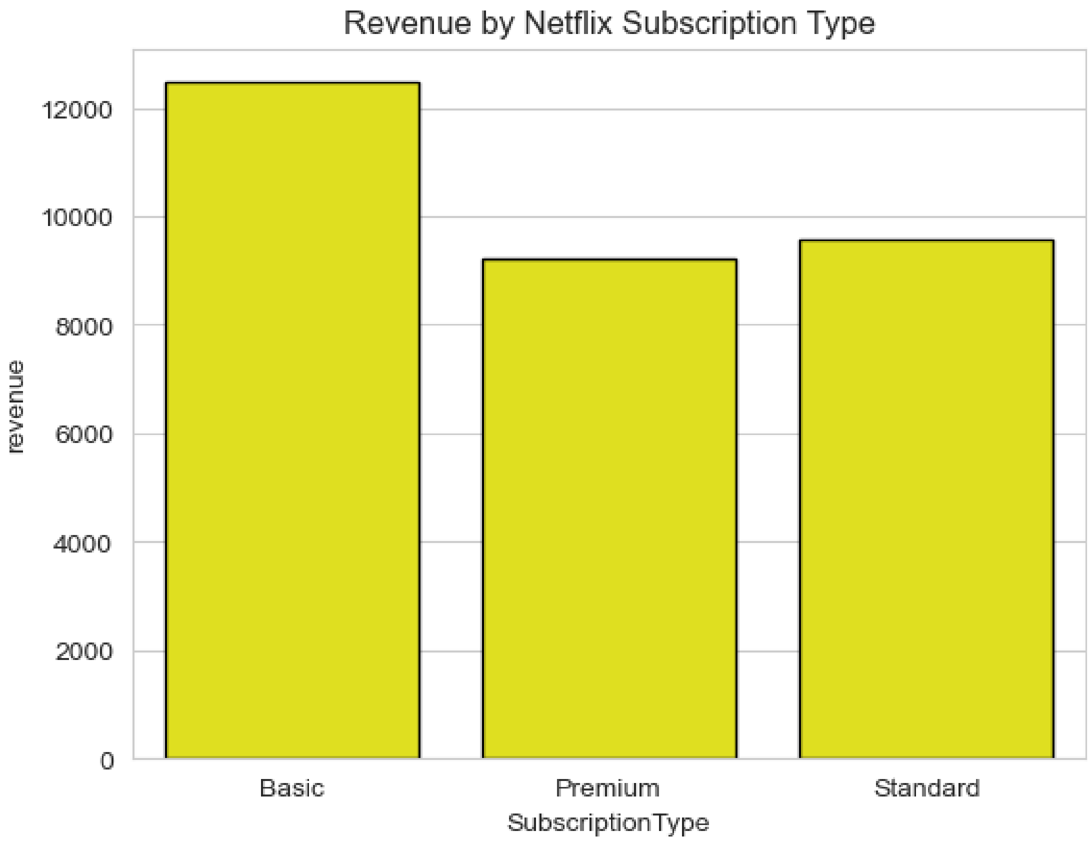
0.2.1 Which Subscription Type is the most profitable for Netflix.?

```
[12]: query_4 = """ SELECT SubscriptionType,SUM(MONTHLYREVENUE) AS revenue FROM␣
        ↪netflix_userbase
                      GROUP BY SubscriptionType """

      money_from_subs = pd.read_sql(query_4, con=engine)

      money_from_subs.head()
```

```
[12]:   SubscriptionType  revenue
      0            Basic    12469
      1          Premium     9229
      2         Standard     9573
```

```
[13]: sns.barplot(data=money_from_subs, x='SubscriptionType', y='revenue',␣
        ↪color='yellow',edgecolor='black')
      plt.title("Revenue by Netflix Subscription Type")
```

```
[13]: Text(0.5, 1.0, 'Revenue by Netflix Subscription Type')
```



Basic Plan pools in more revenue for Netflix.

0.3 Which gender uses Netflix the most?

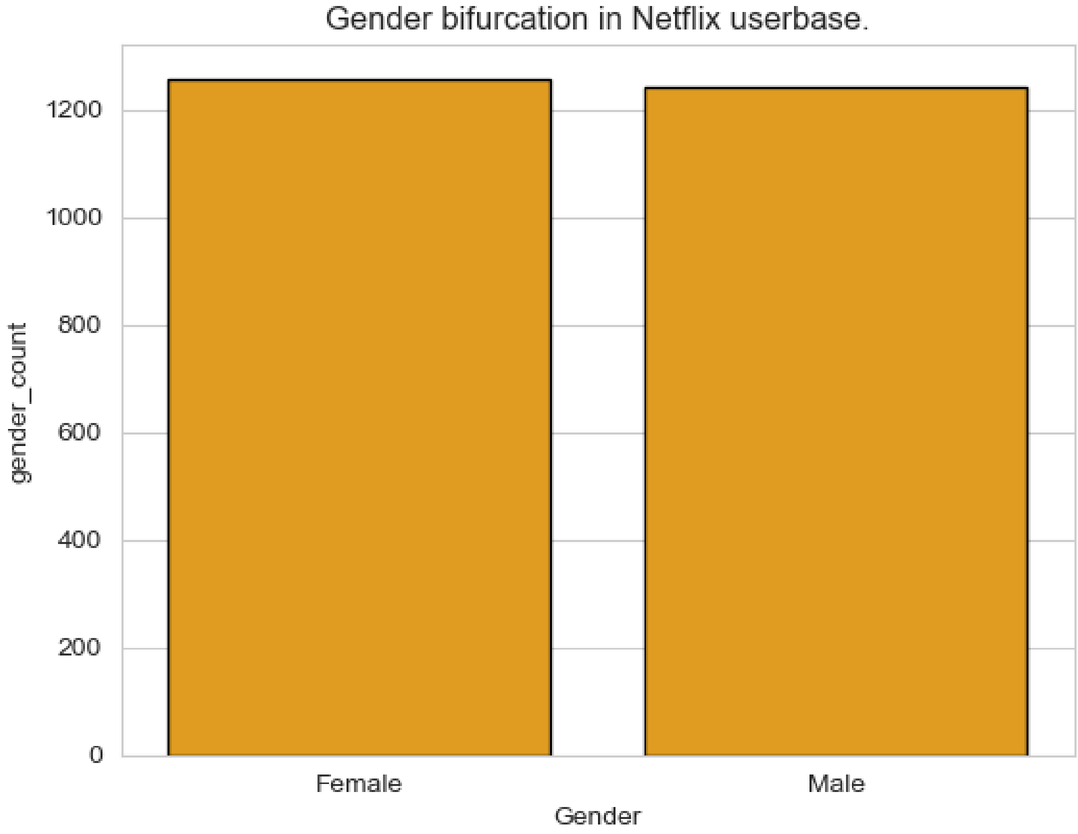
```
[14]: query_5 = """ SELECT Gender,COUNT(*) AS gender_count FROM  netflix_userbase
                    GROUP BY Gender """

      gender_df = pd.read_sql(query_5, con=engine)

      gender_df.head()
```

```
[14]:    Gender  gender_count
      0  Female          1257
      1    Male          1243
```

Netflix has more female subscribers than male subscribers.

```
[15]: sns.barplot(data=gender_df, x='Gender', y='gender_count',␣
       ↪color='orange',edgecolor='black')
      plt.title("Gender bifurcation in Netflix userbase.")
```

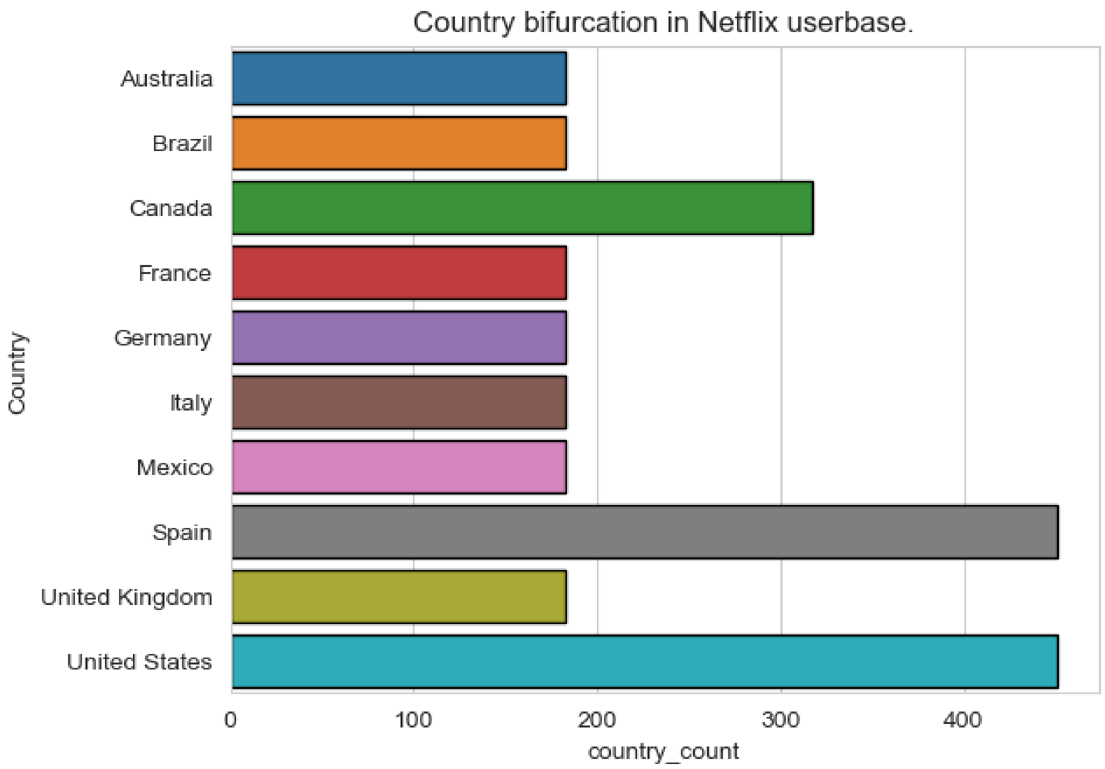
```
[15]: Text(0.5, 1.0, 'Gender bifurcation in Netflix userbase.')
```



0.4 Which Country has the most number of Netflix Users?

```
[16]: query_6 = """ SELECT Country,COUNT(*) as country_count FROM netflix_userbase
                    GROUP BY Country """
      country_df = pd.read_sql(query_6, con=engine)
      sns.barplot(country_df, y='Country', x='country_count',␣
       ↪hue='Country',edgecolor='black')
      plt.title("Country bifurcation in Netflix userbase.")
```

```
[16]: Text(0.5, 1.0, 'Country bifurcation in Netflix userbase.')
```



Spain has the most number of Netflix users according to the dataset

0.5 What is the average age of the user using Netflix?

[17]:
```
query_7 = "SELECT AVG(Age) FROM netflix_userbase"
pd.read_sql(query_7, con=engine)
```

[17]:
```
   AVG(Age)
0   38.7956
```

0.6 Which Country has the youngest user of Netflix ?

[25]:
```
query_8 = """ SELECT Country, MIN(Age) AS min_age FROM netflix_userbase
                GROUP BY Country ORDER BY min_age asc limit 1
"""

pd.read_sql(query_8, con=engine)
```

[25]:
```
         Country  min_age
0  United States       26
```

United States has the youngest user at 26 years old.

0.6.1 Which year did Netflix see a rise in joining of users?

```
[19]: query_8 = """ SELECT SUBSTRING(JoinDate,7,2) AS Year, COUNT(*) AS cnt FROM␣
       ↪netflix_userbase
                      GROUP BY Year """

      trend_of_signups = pd.read_sql(query_8, con=engine)
      trend_of_signups.head()
```

```
[19]:   Year   cnt
      0   21    14
      1   22  2448
      2   23    38
```

```
[20]: sns.lineplot(data=trend_ofsignups,x='Year',y='cnt')
      plt.title("Netflix signup trend from 2021-2023")
```

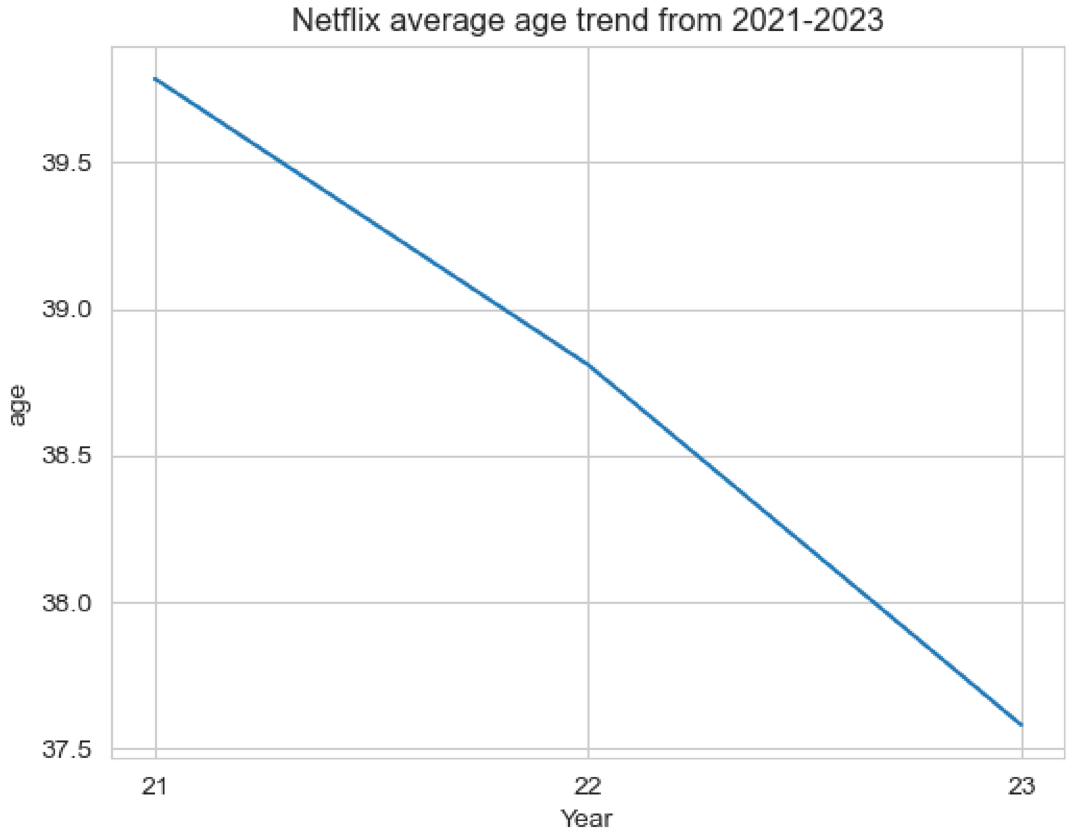
```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[20], line 1
----> 1 sns.lineplot(data=trend_ofsignups,x='Year',y='cnt')
      2 plt.title("Netflix signup trend from 2021-2023")

NameError: name 'trend_ofsignups' is not defined
```

The sign-ups for Netflix were at it's peak in 2022.However,there is a sharp decline post 2022.

0.6.2 Do we see any age-wise trends in singups?

```
[93]: query_9 = """ SELECT AVG(Age) as age,SUBSTRING(JoinDate,7,2) AS Year FROM␣
       ↪netflix_userbase
                      GROUP BY Year
       """

      trend_of_age = pd.read_sql(query_9, con=engine)

      sns.lineplot(data=trend_of_age,x='Year',y='age')
      plt.title("Netflix average age trend from 2021-2023")
```

```
[93]: Text(0.5, 1.0, 'Netflix average age trend from 2021-2023')
```



Interestingly,the average age of users of Netflix has dropped year on year post 2022.