Análise Exploratória - Netflix daily top 10

September 10, 2022

1 Análise Exploratória - Netflix daily top 10

Para esse case, utilizou-se do dataset disponível no Kaggle: https://www.kaggle.com/datasets/prasertk/netflix-daily-top-10-in-us

[2]:
```
import pandas as pd # biblioteca que fornece ferramentas para análise e␣
↪manipulação de dados
import datetime as dt # biblioteca que fornece as classes para manipulação de␣
↪datas e horas
```

[3]:
```
df = pd.read_csv('netflix daily top 10.csv')
```

Visualização da importação do dataframe que foi importado Visualizar as primeiras e últimas linhas do dataframe para ter noção de como esta a base que iremos trabalhar

[4]:
```
display(df)
```

```
           As of  Rank Year to Date Rank Last Week Rank  \
0     2020-04-01     1                 1              1
1     2020-04-01     2                 2              -
2     2020-04-01     3                 3              2
3     2020-04-01     4                 4              -
4     2020-04-01     5                 5              4
…          …   …               …             …
7095  2022-03-11     6                 5              1
7096  2022-03-11     7                 7              2
7097  2022-03-11     8                 8              -
7098  2022-03-11     9                 9              7
7099  2022-03-11    10                10              -

                            Title     Type Netflix Exclusive  \
0     Tiger King: Murder, Mayhem …  TV Show               Yes
1                           Ozark  TV Show               Yes
2                 All American  TV Show               NaN
3                 Blood Father    Movie               NaN
4                 The Platform    Movie               Yes
…                          …      …                 …
7095          Worst Roommate Ever  TV Show               Yes
7096            Vikings: Valhalla  TV Show               Yes
```

```
7097                        Shooter    Movie                NaN
7098                        Shrek 2    Movie                NaN
7099                          Shrek    Movie                NaN

     Netflix Release Date  Days In Top 10  Viewership Score
0            Mar 20, 2020               9                90
1            Jul 21, 2017               5                45
2            Mar 28, 2019               9                76
3            Mar 26, 2020               5                30
4            Mar 20, 2020               9                55
…                     …               …                 …
7095          Mar 1, 2022              10                81
7096         Feb 25, 2022              14               100
7097          Aug 1, 2014               3                 7
7098          Mar 1, 2022              10                33
7099          May 1, 2018               7                12

[7100 rows x 10 columns]
```

Descobrindo o período da análise (As of) Utilizando da biblioteca datetime

[5]:
```
data_inicio = pd.to_datetime(df['As of']).dt.date.min() # data mínima do␣
 ↪dataframe
print('Data de início:', data_inicio)
data_fim = pd.to_datetime(df['As of']).dt.date.max() # data máxima do dataframe
print('Data de fim:', data_fim)
```

```
Data de início: 2020-04-01
Data de fim: 2022-03-11
```

Verificar valores nulos e os tipos de dados

[6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7100 entries, 0 to 7099
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   As of                 7100 non-null   object
 1   Rank                  7100 non-null   int64
 2   Year to Date Rank     7100 non-null   object
 3   Last Week Rank        7100 non-null   object
 4   Title                 7100 non-null   object
 5   Type                  7100 non-null   object
 6   Netflix Exclusive     4599 non-null   object
 7   Netflix Release Date  7100 non-null   object
 8   Days In Top 10        7100 non-null   int64
 9   Viewership Score      7100 non-null   int64
dtypes: int64(3), object(7)
```

```
memory usage: 554.8+ KB
```

Verificar a existência de valores nulos

[7]:
```
df.isnull().sum()
```

[7]:
```
As of                      0
Rank                       0
Year to Date Rank          0
Last Week Rank             0
Title                      0
Type                       0
Netflix Exclusive       2501
Netflix Release Date       0
Days In Top 10             0
Viewership Score           0
dtype: int64
```

Entendendo melhor os valores nulos da coluna 'Netflix Exclusive'

[8]:
```
df['Netflix Exclusive'].value_counts()
```

[8]:
```
Yes    4599
Name: Netflix Exclusive, dtype: int64
```

Analisando as informações Estatísticas

[9]:
```
df.describe()
```

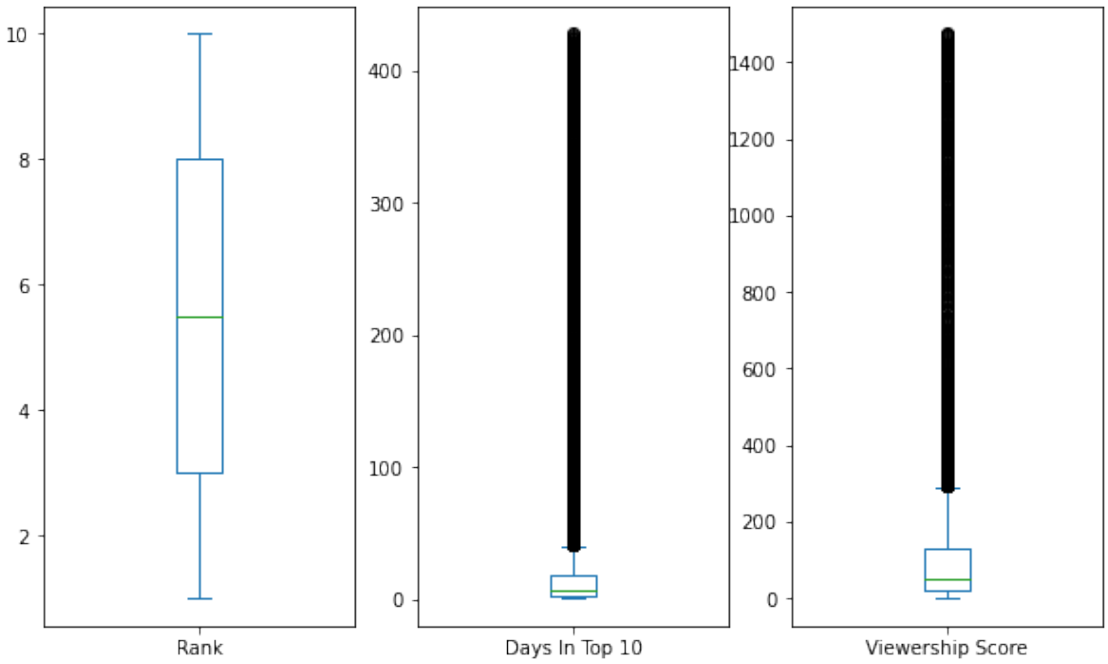
[9]:

	Rank	Days In Top 10	Viewership Score
count	7100.000000	7100.000000	7100.000000
mean	5.500000	24.123662	122.790141
std	2.872484	58.473789	213.861642
min	1.000000	1.000000	1.000000
25%	3.000000	3.000000	19.000000
50%	5.500000	7.000000	50.000000
75%	8.000000	18.000000	128.000000
max	10.000000	428.000000	1474.000000

Entendendo melhor as informações Estatísticas via boxplot

[10]:
```
df.plot(kind='box', figsize=(10,6), subplots=True)
```

[10]:
```
Rank                  AxesSubplot(0.125,0.125;0.227941x0.755)
Days In Top 10      AxesSubplot(0.398529,0.125;0.227941x0.755)
Viewership Score    AxesSubplot(0.672059,0.125;0.227941x0.755)
dtype: object
```



Observando os boxplots acima de Days In Top 10 vemos que existem muitos valores máximos acima de 100. Para isso vamos entender o que seriam esses outliers.

[12]:
```
df[df['Days In Top 10'] >= 100]
```

[12]:

	As of	Rank	Year to Date Rank	Last Week Rank	Title	Type	\
2886	2021-01-14	7	6	8	Cocomelon	TV Show	
2896	2021-01-15	7	7	10	Cocomelon	TV Show	
2909	2021-01-16	10	7	9	Cocomelon	TV Show	
2919	2021-01-17	10	10	9	Cocomelon	TV Show	
3019	2021-01-27	10	-	-	Cocomelon	TV Show	
…	…	…	…	…	…	…	
6674	2022-01-28	5	6	-	Cocomelon	TV Show	
6687	2022-01-29	8	5	8	Cocomelon	TV Show	
6718	2022-02-01	9	-	7	Cocomelon	TV Show	
6959	2022-02-25	10	-	-	Cocomelon	TV Show	
6998	2022-03-01	9	-	-	Cocomelon	TV Show	

	Netflix Exclusive	Netflix Release Date	Days In Top 10	Viewership Score
2886	NaN	Jun 1, 2020	100	287
2896	NaN	Jun 1, 2020	101	291
2909	NaN	Jun 1, 2020	102	292
2919	NaN	Jun 1, 2020	103	293
3019	NaN	Jun 1, 2020	104	294
…	…	…	…	…
6674	NaN	Jun 1, 2020	424	1466

```
6687              NaN          Jun 1, 2020              425              1469
6718              NaN          Jun 1, 2020              426              1471
6959              NaN          Jun 1, 2020              427              1472
6998              NaN          Jun 1, 2020              428              1474

[329 rows x 10 columns]
```

Entendendo os motivos do Title 'Cocomelon' e se realmente está certo

[13]:
```
df.Title.value_counts()
```

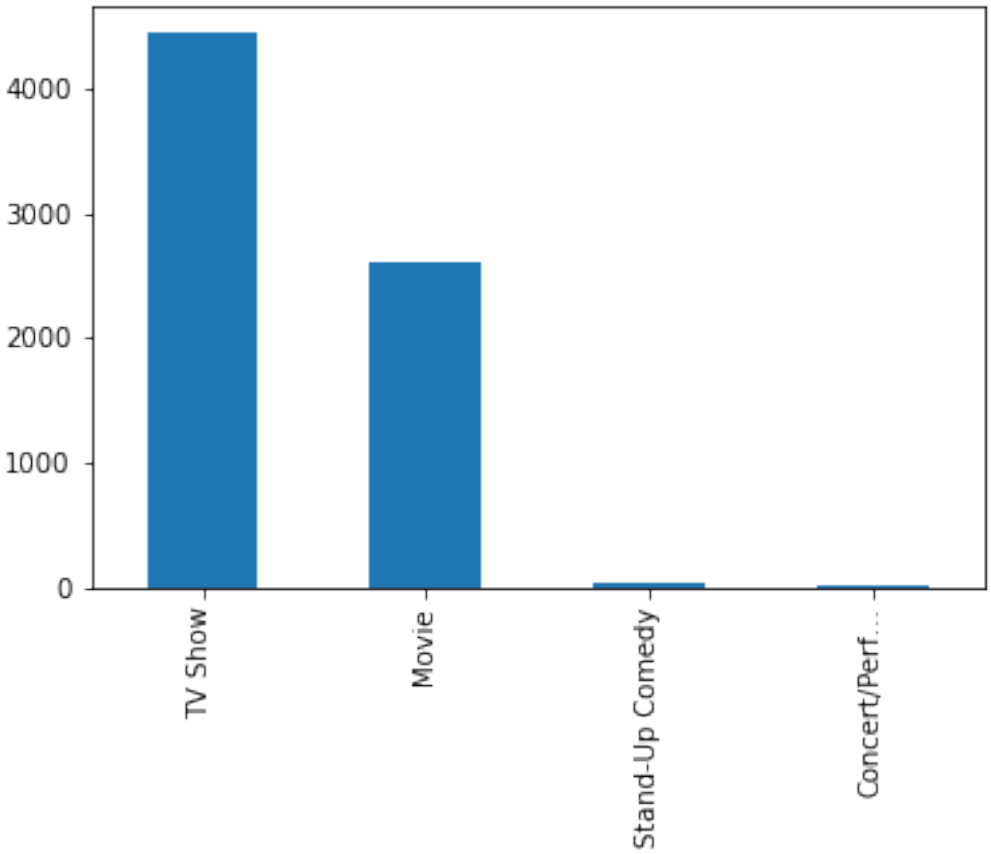
[13]:
```
Cocomelon                       428
Ozark                            85
Cobra Kai                        81
Manifest                         80
The Queenâ s Gambit              73
                                ...
The Office                        1
Animals on the Loose: A You…      1
Dark                              1
The Secret Life of Pets 2         1
Step Up Revolution                1
Name: Title, Length: 645, dtype: int64
```

Com isso, foi possível perceber que realmente o 'Cocomelon' ficou 428 dias no top 10 na Netflix. Uma coisa interessante a se pensar é que muitas das vezes a Análise Exploratória desmistifica o achismo. Ex: pensar que Cobra Kai é o mais visto, estando no top 10 por muito mais dias, sendo que no output acima não é.

Plotando gráficos de barras da coluna 'Type'

[16]:
```
df.Type.value_counts().plot(kind='bar')
```

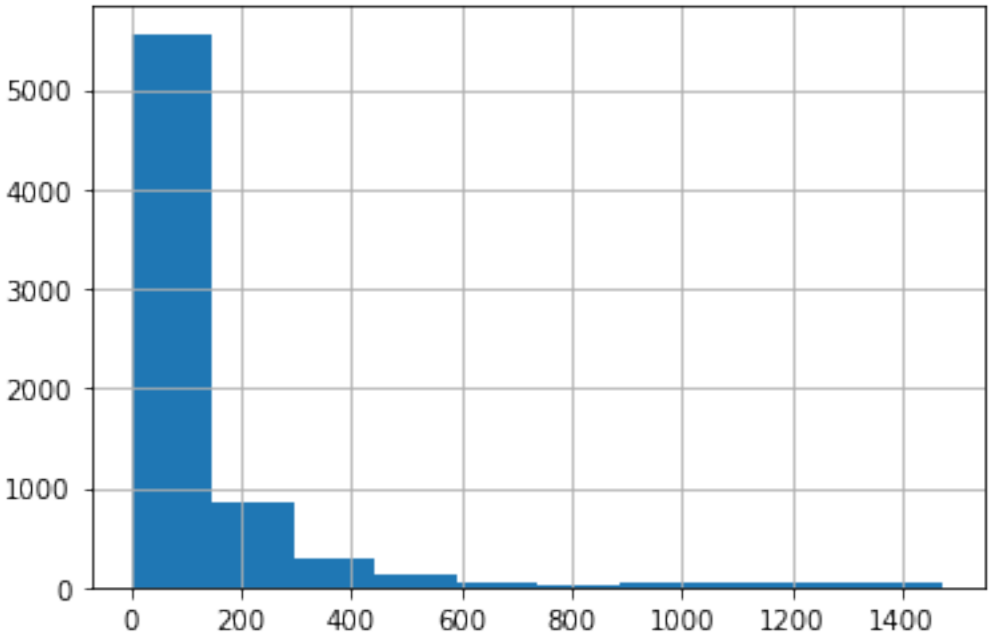
[16]:
```
<AxesSubplot:>
```



Entendendo a coluna 'Viewership Score'

[19]:
```
df['Viewership Score'].hist()
```

[19]:
```
<AxesSubplot:>
```



Mesmo assim não fica claro o que significa e o que se pode tirar da coluna 'Viewership Score'. Para isso o ideal é buscar outras informações sobre. No site que nos fornece os dados tem uma explicativa dessa coluna que nos diz o seguinte: 'The Viewership Score is a score assigned to each show based on its historical daily ranking, assigning 10 points for each no. 1 ranking, 9 points for each no. 2 ranking etc.' Ou seja, isso é uma pontuação recebida devido a posição que o filme/série ficou no ranking.

Verificando qual ganhou mais pontos

[23]:
```
df[df['Viewership Score']== df['Viewership Score'].max()]
```

[23]:

	As of	Rank	Year to Date Rank	Last Week Rank	Title	Type	Netflix Exclusive	Netflix Release Date	Days In Top 10	Viewership Score
6998	2022-03-01	9	-	-	Cocomelon	TV Show	NaN	Jun 1, 2020	428	1474

[ ]:
```

```