Analysis of Web Series Data

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
dataset = pd.read_excel(r"/content/2255872-anime_data.xlsx")
```

Overview of Data:

```
dataset.head(2)
```

	title	description	mediaType	eps	duration	ongoing	sznOfRelease	years_running	studio_primary	studios_colab	...	tag_Supernatural	tag_Magic	tag_Animal_Protagonists	tag_Ecchi	tag_M
0	Fullmetal Alchemist: Brotherhood	The foundation of alchemy is based on the law ...	TV	64	NaN	False	Spring	1	Bones	0	...	0	0	0	0	
1	your name.	Mitsuha and Taki are two total strangers livin...	Movie	1	107.0	False	is_missing	0	Others	0	...	1	0	0	0	

2 rows × 44 columns

Operations on the Dataset:

```
dataset.shape
```

```
(12101, 44)
```

```
dataset.eps.describe()
```

```
count    12101.000000
mean        13.393356
std         57.925097
min          1.000000
25%          1.000000
50%          2.000000
75%         12.000000
max       2527.000000
Name: eps, dtype: float64
```

```
dataset.columns
```

```
Index(['title', 'description', 'mediaType', 'eps', 'duration', 'ongoing',
       'sznOfRelease', 'years_running', 'studio_primary', 'studios_colab',
       'contentWarn', 'watched', 'watching', 'wantWatch', 'dropped', 'rating',
       'votes', 'tag_Based_on_a_Manga', 'tag_Comedy', 'tag_Action',
       'tag_Fantasy', 'tag_Sci_Fi', 'tag_Shounen', 'tag_Original_Work',
       'tag_Non_Human_Protagonists', 'tag_Drama', 'tag_Adventure',
       'tag_Family_Friendly', 'tag_Short_Episodes', 'tag_School_Life',
       'tag_Romance', 'tag_Shorts', 'tag_Slice_of_Life', 'tag_Seinen',
       'tag_Supernatural', 'tag_Magic', 'tag_Animal_Protagonists', 'tag_Ecchi',
       'tag_Mecha', 'tag_Based_on_a_Light_Novel', 'tag_CG_Animation',
       'tag_Superpowers', 'tag_Others', 'tag_missing'],
      dtype='object')
```

```
dataset.eps.describe()
```

```
count    12101.000000
mean        13.393356
std         57.925097
min          1.000000
25%          1.000000
50%          2.000000
75%         12.000000
max       2527.000000
Name: eps, dtype: float64
```

```
dataset[(dataset['eps']>24) & (dataset.duration.isna())].shape
```

```
(1493, 44)
```

```
dataset_excluding_out = dataset[dataset['eps']<50]
```

```
dataset_excluding_out['eps_brackets'] = pd.cut(dataset_excluding_out['eps'], bins = [1,10,20,30,40,50], labels = ['cat1','cat2','cat3','cat4','cat5'])
```

```
<ipython-input-13-82351de03e68>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dataset_excluding_out['eps_brackets'] = pd.cut(dataset_excluding_out['eps'], bins = [1,10,20,30,40,50], labels = ['cat1','cat2','cat3','cat4','cat5'])
```

```
dataset_excluding_out.shape
```

```
(11388, 45)
```

```
dataset_excluding_out.groupby(['eps_brackets']).duration.mean()
```

```
eps_brackets
cat1    13.556684
cat2     7.419295
cat3     7.184783
cat4     8.549020
```

```
cat5     8.823529
Name: duration, dtype: float64
```

```
dataset_excluding_out[dataset_excluding_out['eps_brackets']=='cat1'].shape
```

```
(1901, 45)
```

```
dataset[(dataset['eps']<24) & (~dataset.duration.isna())].describe()
```

	eps	duration	years_running	studios_colab	contentWarn	watched	watching	wantWatch	dropped	rating	...	tag_Supernatural	tag_Magic	tag_Animal_Protagoni
count	7098.000000	7098.000000	7098.000000	7098.000000	7098.000000	7098.000000	7098.000000	7098.000000	7098.000000	7098.000000	...	7098.000000	7098.000000	7098.000
mean	2.546210	25.080727	0.104959	0.034658	0.095661	1531.826289	42.525923	609.343054	22.749084	2.742419	...	0.057904	0.044520	0.066
std	3.611337	32.016127	0.556363	0.182924	0.294146	4699.844075	238.987630	1301.861782	72.586285	0.831137	...	0.233578	0.206261	0.248
min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.844000	...	0.000000	0.000000	0.000
25%	1.000000	4.000000	0.000000	0.000000	0.000000	41.000000	1.000000	28.000000	1.000000	2.072000	...	0.000000	0.000000	0.000
50%	1.000000	9.000000	0.000000	0.000000	0.000000	170.000000	5.000000	136.000000	5.000000	2.707000	...	0.000000	0.000000	0.000
75%	1.000000	30.000000	0.000000	0.000000	0.000000	914.000000	26.000000	622.000000	17.000000	3.387750	...	0.000000	0.000000	0.000
max	23.000000	163.000000	20.000000	1.000000	1.000000	115949.000000	15732.000000	21733.000000	2010.000000	4.663000	...	1.000000	1.000000	1.000

8 rows × 38 columns

```
dataset_excluding_out.groupby('mediaType').agg({'duration':'mean','mediaType' : 'count'})
```

	duration	mediaType
mediaType		
DVD Special	10.995798	802
Movie	57.869213	1928
Music Video	4.009412	1290
OVA	32.913809	1769
Other	7.219378	576
TV	7.130662	3308
TV Special	45.795181	504
Web	7.116523	1152
is_missing	17.555556	59

```
dataset.isna().sum()
```

```
title                            1
description                   4468
mediaType                        0
eps                              0
```

```
duration                      4636
ongoing                          0
sznOfRelease                     0
years_running                    0
studio_primary                   0
studios_colab                    0
contentWarn                      0
watched                          0
watching                         0
wantWatch                        0
dropped                          0
rating                           0
votes                            0
tag_Based_on_a_Manga             0
tag_Comedy                       0
tag_Action                       0
tag_Fantasy                      0
tag_Sci_Fi                       0
tag_Shounen                      0
tag_Original_Work                0
tag_Non_Human_Protagonists       0
tag_Drama                        0
tag_Adventure                    0
tag_Family_Friendly              0
tag_Short_Episodes               0
tag_School_Life                  0
tag_Romance                      0
tag_Shorts                       0
tag_Slice_of_Life                0
tag_Seinen                       0
tag_Supernatural                 0
tag_Magic                        0
tag_Animal_Protagonists          0
tag_Ecchi                        0
tag_Mecha                        0
tag_Based_on_a_Light_Novel       0
tag_CG_Animation                 0
tag_Superpowers                  0
tag_Others                       0
tag_missing                      0
dtype: int64
```

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12101 entries, 0 to 12100
Data columns (total 44 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   title                       12100 non-null  object 
 1   description                 7633 non-null   object 
 2   mediaType                   12101 non-null  object 
 3   eps                         12101 non-null  int64  
 4   duration                    7465 non-null   float64
 5   ongoing                     12101 non-null  bool   
 6   sznOfRelease                12101 non-null  object 
 7   years_running               12101 non-null  int64  
 8   studio_primary              12101 non-null  object 
 9   studios_colab               12101 non-null  int64  
 10  contentWarn                 12101 non-null  int64  
 11  watched                     12101 non-null  float64
 12  watching                   12101 non-null  int64
 13  wantWatch                  12101 non-null  int64
 14  dropped                    12101 non-null  int64
 15  rating                     12101 non-null  float64
 16  votes                      12101 non-null  int64
 17  tag_Based_on_a_Manga       12101 non-null  int64
 18  tag_Comedy                 12101 non-null  int64
 19  tag_Action                 12101 non-null  int64
 20  tag_Fantasy                12101 non-null  int64
 21  tag_Sci_Fi                 12101 non-null  int64
 22  tag_Shounen                12101 non-null  int64
 23  tag_Original_Work          12101 non-null  int64
 24  tag_Non_Human_Protagonists  12101 non-null  int64
 25  tag_Drama                  12101 non-null  int64
 26  tag_Adventure              12101 non-null  int64
 27  tag_Family_Friendly        12101 non-null  int64
 28  tag_Short_Episodes         12101 non-null  int64
 29  tag_School_Life            12101 non-null  int64
 30  tag_Romance                12101 non-null  int64
 31  tag_Shorts                 12101 non-null  int64
 32  tag_Slice_of_Life          12101 non-null  int64
 33  tag_Seinen                 12101 non-null  int64
 34  tag_Supernatural           12101 non-null  int64
 35  tag_Magic                  12101 non-null  int64
 36  tag_Animal_Protagonists    12101 non-null  int64
 37  tag_Ecchi                  12101 non-null  int64
 38  tag_Mecha                  12101 non-null  int64
 39  tag_Based_on_a_Light_Novel  12101 non-null  int64
 40  tag_CG_Animation           12101 non-null  int64
 41  tag_Superpowers            12101 non-null  int64
 42  tag_Others                 12101 non-null  int64
 43  tag_missing                12101 non-null  int64
dtypes: bool(1), float64(3), int64(35), object(5)
memory usage: 4.0+ MB
```

```
dataset.describe().T
```

	count	mean	std	min	25%	50%	75%	max
eps	12101.0	13.393356	57.925097	1.000	1.000	2.000	12.000	2527.000
duration	7465.0	24.230141	31.468171	1.000	4.000	8.000	30.000	163.000
years_running	12101.0	0.283200	1.152234	0.000	0.000	0.000	0.000	51.000
studios_colab	12101.0	0.051649	0.221326	0.000	0.000	0.000	0.000	1.000
contentWarn	12101.0	0.115362	0.319472	0.000	0.000	0.000	0.000	1.000
watched	12101.0	2862.605694	7724.347024	0.000	55.000	341.000	2026.000	161567.000
watching	12101.0	256.334435	1380.840902	0.000	2.000	14.000	100.000	74537.000
wantWatch	12101.0	1203.681431	2294.327380	0.000	49.000	296.000	1275.000	28541.000
dropped	12101.0	151.568383	493.931710	0.000	3.000	12.000	65.000	19481.000
rating	12101.0	2.949037	0.827385	0.844	2.304	2.965	3.616	4.702
votes	12101.0	2088.124700	5950.332228	10.000	34.000	219.000	1414.000	131067.000
tag_Based_on_a_Manga	12101.0	0.290802	0.454151	0.000	0.000	0.000	1.000	1.000
tag_Comedy	12101.0	0.272870	0.445453	0.000	0.000	0.000	1.000	1.000
tag_Action	12101.0	0.231221	0.421631	0.000	0.000	0.000	0.000	1.000
tag_Fantasy	12101.0	0.181555	0.385493	0.000	0.000	0.000	0.000	1.000
tag_Sci_Fi	12101.0	0.166267	0.372336	0.000	0.000	0.000	0.000	1.000
tag_Shounen	12101.0	0.144864	0.351978	0.000	0.000	0.000	0.000	1.000
tag_Original_Work	12101.0	0.135195	0.341946	0.000	0.000	0.000	0.000	1.000
tag_Non_Human_Protagonists	12101.0	0.112470	0.315957	0.000	0.000	0.000	0.000	1.000
tag_Drama	12101.0	0.106107	0.307987	0.000	0.000	0.000	0.000	1.000
tag_Adventure	12101.0	0.103793	0.305005	0.000	0.000	0.000	0.000	1.000
tag_Family_Friendly	12101.0	0.097017	0.295993	0.000	0.000	0.000	0.000	1.000
tag_Short_Episodes	12101.0	0.096934	0.295880	0.000	0.000	0.000	0.000	1.000
tag_School_Life	12101.0	0.092306	0.289470	0.000	0.000	0.000	0.000	1.000
tag_Romance	12101.0	0.092141	0.289237	0.000	0.000	0.000	0.000	1.000

```
dataset.drop(columns = ['title', 'description'], axis=1, inplace= True)
```

tag_Slice_of_Life	12101.0	0.080820	0.272569	0.000	0.000	0.000	0.000	1.000

```
dataset.head()
```

	mediaType	eps	duration	ongoing	sznOfRelease	years_running	studio_primary	studios_colab	contentWarn	watched	...	tag_Supernatural	tag_Magic	tag_Animal_Protagonists	tag_Ecchi	tag_Mec
0	TV	64	NaN	False	Spring	1	Bones	0	1	103707.0	...	0	0	0	0	
1	Movie	1	107.0	False	is_missing	0	Others	0	0	58831.0	...	1	0	0	0	

```
dataset.rating.describe()
```

```
count    12101.000000
mean         2.949037
std          0.827385
min          0.844000
25%          2.304000
50%          2.965000
75%          3.616000
max          4.702000
Name: rating, dtype: float64
```

```
dataset.dropna(inplace=True)
dataset.shape
```

```
(7465, 42)
```

```
dataset.columns
```

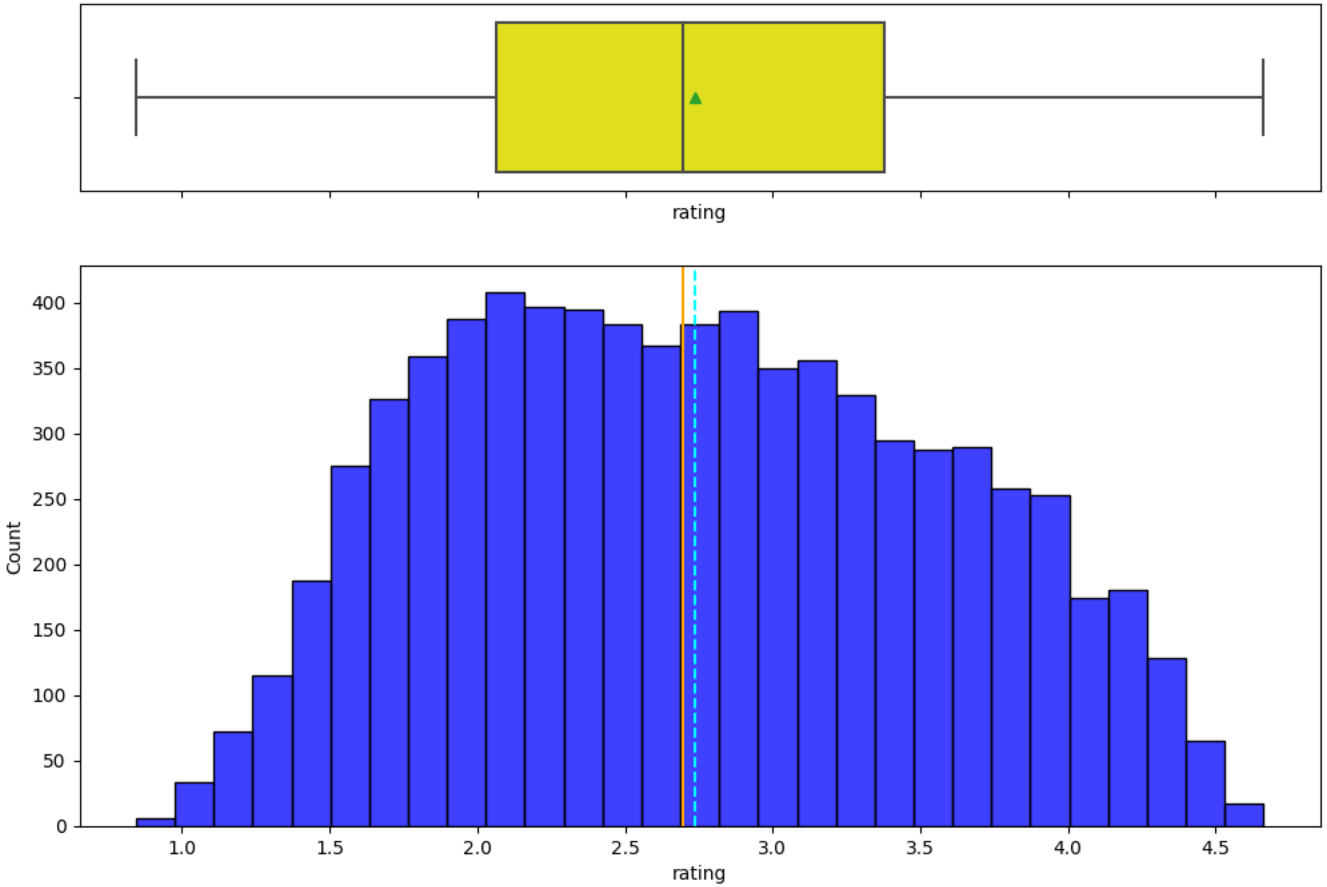
```
Index(['mediaType', 'eps', 'duration', 'ongoing', 'sznOfRelease',
       'years_running', 'studio_primary', 'studios_colab', 'contentWarn',
       'watched', 'watching', 'wantWatch', 'dropped', 'rating', 'votes',
       'tag_Based_on_a_Manga', 'tag_Comedy', 'tag_Action', 'tag_Fantasy',
       'tag_Sci_Fi', 'tag_Shounen', 'tag_Original_Work',
       'tag_Non_Human_Protagonists', 'tag_Drama', 'tag_Adventure',
       'tag_Family_Friendly', 'tag_Short_Episodes', 'tag_School_Life',
       'tag_Romance', 'tag_Shorts', 'tag_Slice_of_Life', 'tag_Seinen',
       'tag_Supernatural', 'tag_Magic', 'tag_Animal_Protagonists', 'tag_Ecchi',
       'tag_Mecha', 'tag_Based_on_a_Light_Novel', 'tag_CG_Animation',
       'tag_Superpowers', 'tag_Others', 'tag_missing'],
      dtype='object')
```

```
def continuos_univariate_analysis(data,feature,figsize=(12,8),kde=False,bins=None):
  f1, (ax_box,ax_hist)= plt.subplots(nrows=2,sharex=True, gridspec_kw={'height_ratios':(0.25,0.75)},figsize=figsize)
  sns.color_palette("viridis", as_cmap=True)
  sns.boxplot(data=data, x=feature , ax=ax_box, showmeans=True, color='yellow')
  sns.histplot(data=data, x=feature , ax=ax_hist, showmeans=True, color='crest', bins=bins, kde=kde) if bins else sns.histplot(
      data=data, x=feature , ax=ax_hist, kde=kde, color='blue')
  ax_hist.axvline(data[feature].mean(), color='cyan', linestyle='--')
  ax_hist.axvline(data[feature].median(), color='orange', linestyle="-")
```

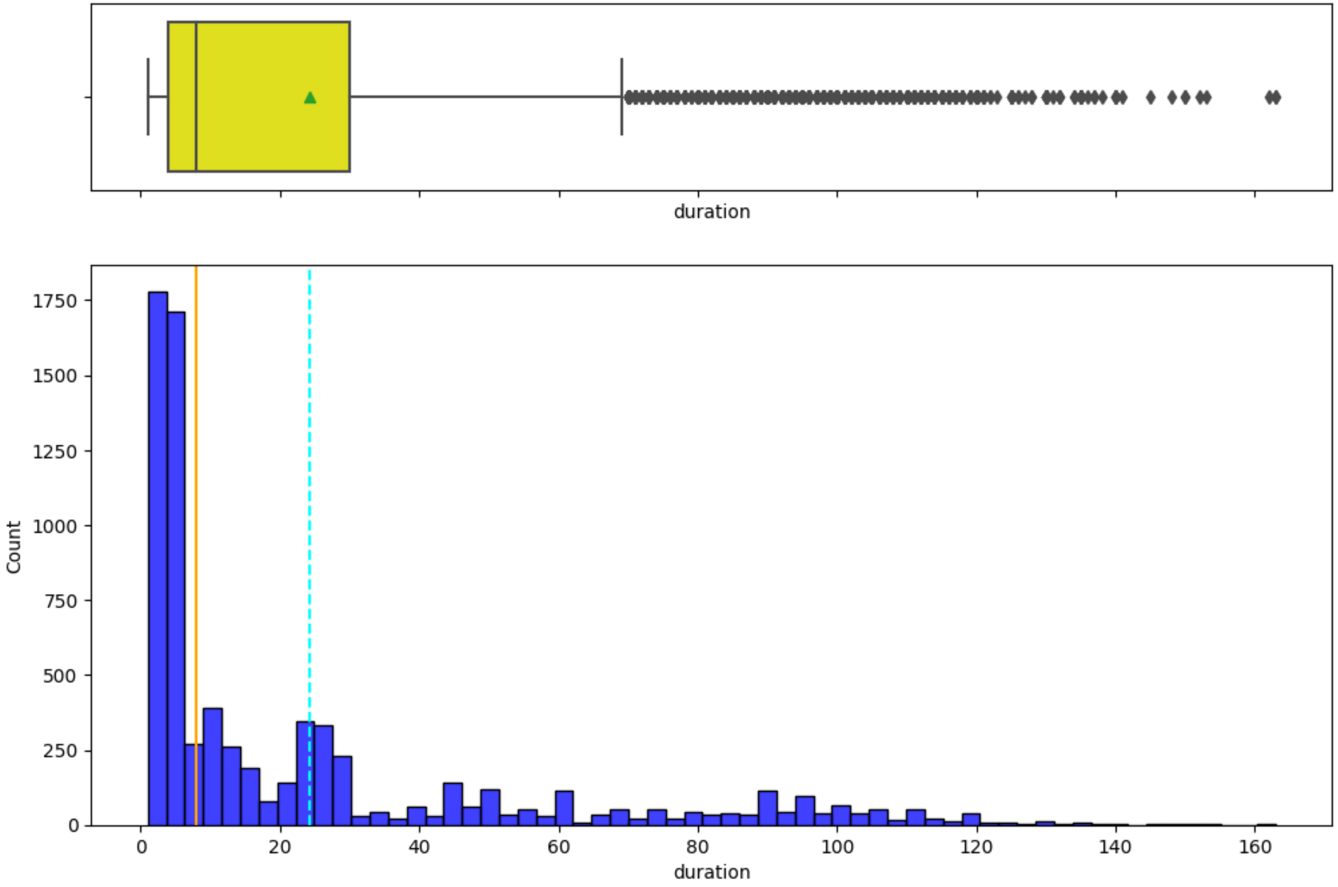
```
def discrete_univariate_analysis(data, features, perc=False, n=None):
    total = len(data[features])
    count  = data[features].nunique()
    if n is None:
        plt.figure(figsize=(count+1, 5))
    else:
        plt.figure(figsize = (n+1, 5))
    plt.xticks(rotation =90, fontsize=15)
    ax = sns.countplot(data = data, x = features, palette='flare', order = data[features].value_counts().index[:n].sort_values(ascending=False))
    for p in ax.patches:
        if perc == True:
            label = "{:.1f}%".format(100*p.get_height()/total)
        else:
            label = p.get_height()
        x = p.get_x() +p.get_width()/2
        y = p.get_height()
        ax.annotate(label, (x,y), ha ='center', va='center', size=12, xytext = (0,5), textcoords = "offset points")
    plt.show()
```

Box Plot and Histograph:

```
continuos_univariate_analysis(dataset,'rating')
```



```
continuos_univariate_analysis(dataset,'duration')
```



```
dataset[dataset['duration']>=80]['rating'].mean()
```

```
3.5694732254047326
```

```
dataset[dataset['duration']>=100]['rating'].mean()
```

```
3.729269121813031
```

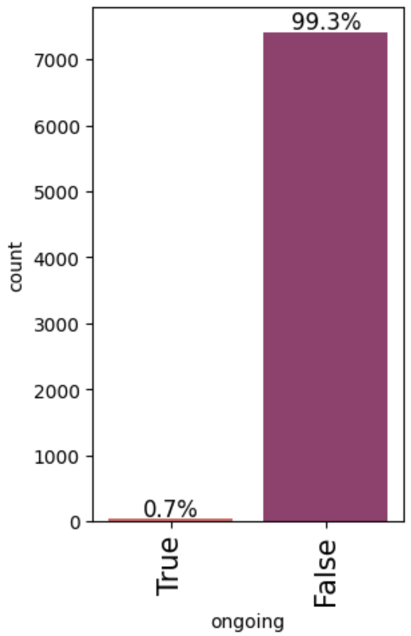
```
dataset[dataset['duration']>=110]['rating'].mean()
```

```
3.7585191256830606
```

```
dataset[(dataset['duration']>=5) & (dataset['duration']<=30)]['rating'].mean()
```

```
2.7890469755469756
```

```
discrete_univariate_analysis(dataset, "ongoing", perc= True)
```



```
dataset[dataset['ongoing']== True]['rating'].mean()
```

```
3.1624600000000003
```

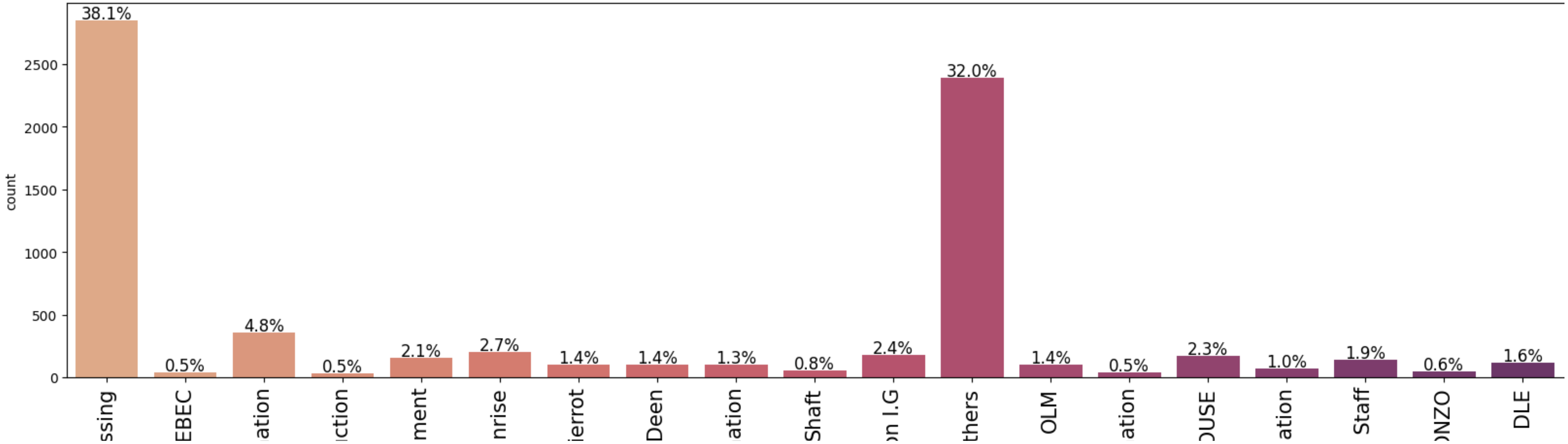
```
dataset[dataset['ongoing']== True]['duration'].mean()
```

```
8.94
```

```
discrete_univariate_analysis(dataset, "sznOfRelease", perc= True)
```

88.0%
6000
5000
4000
3000
count

```
discrete_univariate_analysis(dataset, "studio_primary", perc= True)
```

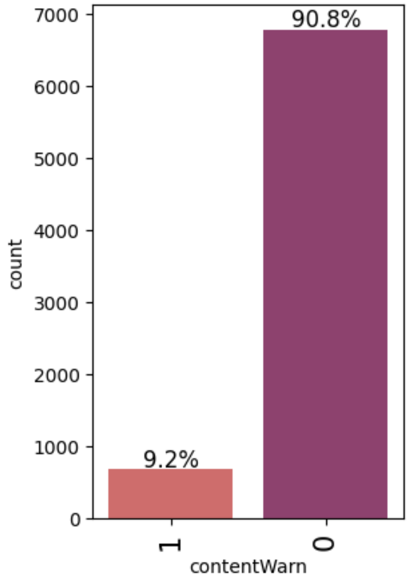


```
dataset[dataset['rating']>4]['studio_primary'].value_counts(normalize=True).mul(100).round(2)
```

```
Others                  38.25
Production I.G           8.42
is_missing               7.02
TMS Entertainment        5.96
MADHOUSE                 5.96
Sunrise                  4.91
Kyoto Animation          4.04
Studio Deen              3.68
Bones                    3.68
A-1 Pictures             3.68
Toei Animation           3.51
Shaft                    3.33
J.C. Staff               3.16
Studio Pierrot           2.46
XEBEC                    0.35
Tatsunoko Production     0.35
```

```
Nippon Animation         0.35
OLM                      0.35
Shin-Ei Animation        0.35
GONZO                    0.18
Name: studio_primary, dtype: float64
```

```
discrete_univariate_analysis(dataset, "contentWarn", perc= True)
```



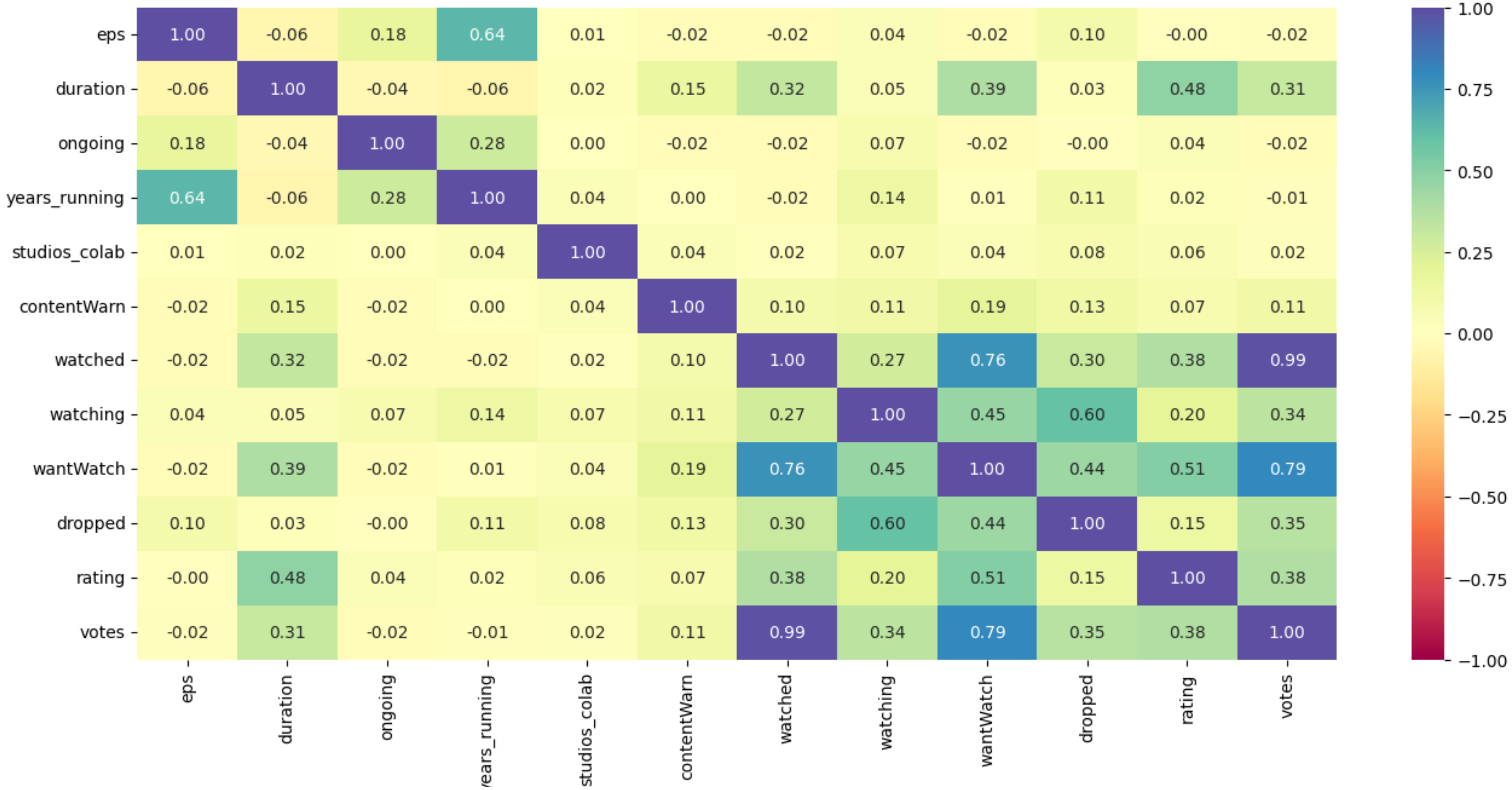
```
corr_cols = [item for item in dataset.columns if "tag" not in item]
```

```
corr_cols
```

```
['mediaType',
 'eps',
 'duration',
 'ongoing',
 'sznOfRelease',
 'years_running',
 'studio_primary',
 'studios_colab',
 'contentWarn',
 'watched',
 'watching',
 'wantWatch',
 'dropped',
 'rating',
 'votes']
```

```
plt.figure(figsize=(16,7))
sns.heatmap(dataset[corr_cols].corr(), annot=True, vmin=-1, vmax=1, fmt='.2f',cmap='Spectral')
plt.show()
```

```
<ipython-input-51-e207a43d782f>:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or spe
  sns.heatmap(dataset[corr_cols].corr(), annot=True, vmin=-1, vmax=1, fmt='.2f',cmap='Spectral')
```



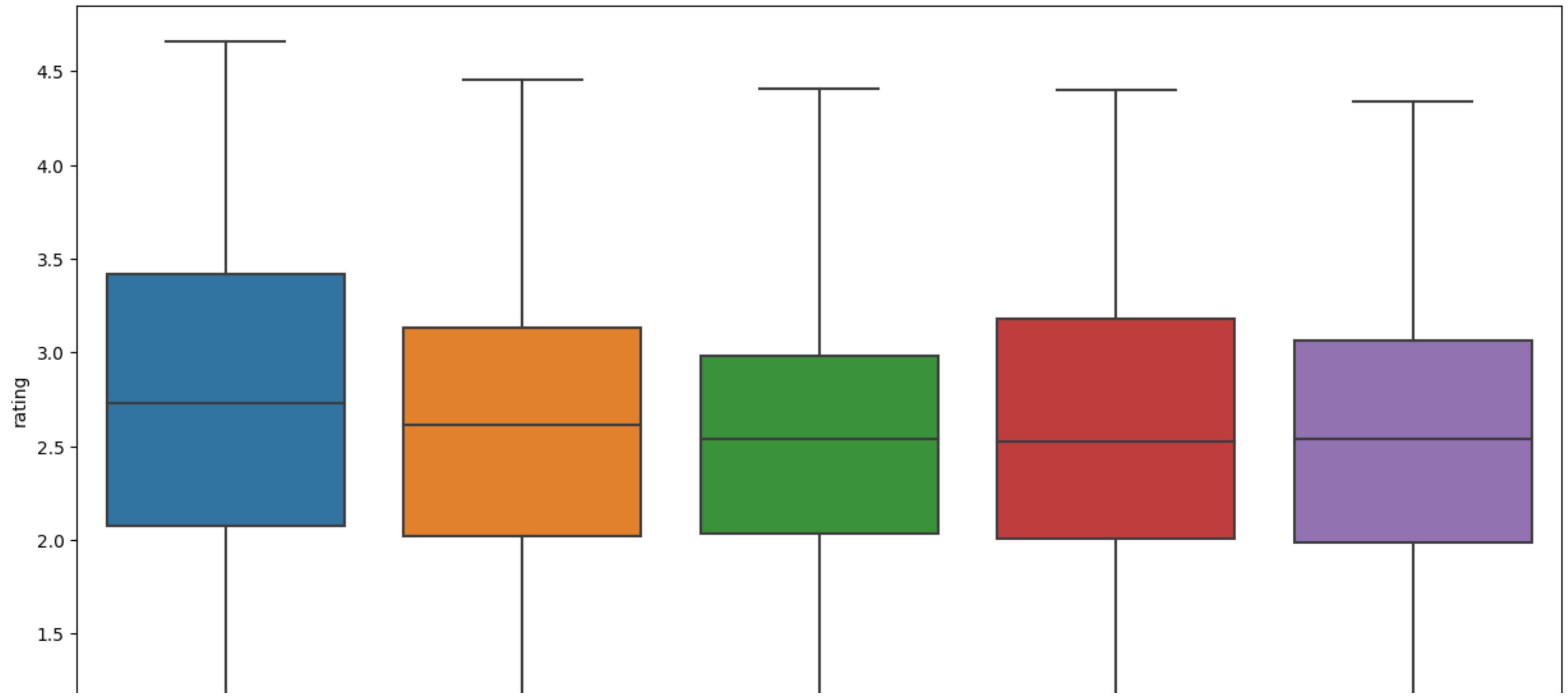
```
dataset.drop(columns=['eps','watched'], inplace= True)
```

```
dataset.shape
```

```
(7465, 40)
```

```
plt.figure(figsize=(15,8))
sns.boxplot(x='sznOfRelease', y='rating', data=dataset)
```

```
<Axes: xlabel='sznOfRelease', ylabel='rating'>
```



Model Building: Regression

```
x=dataset.drop(['rating'],axis=1)
y=dataset['rating']
```

```
x.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7465 entries, 1 to 12100
Data columns (total 39 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   mediaType                 7465 non-null   object
 1   duration                  7465 non-null   float64
 2   ongoing                   7465 non-null   bool
 3   sznOfRelease              7465 non-null   object
 4   years_running             7465 non-null   int64
 5   studio_primary            7465 non-null   object
 6   studios_colab             7465 non-null   int64
 7   contentWarn               7465 non-null   int64
 8   watching                  7465 non-null   int64
 9   wantWatch                 7465 non-null   int64
 10  dropped                   7465 non-null   int64
 11  votes                     7465 non-null   int64
 12  tag_Based_on_a_Manga        7465 non-null   int64
 13  tag_Comedy                  7465 non-null   int64
 14  tag_Action                  7465 non-null   int64
 15  tag_Fantasy                 7465 non-null   int64
 16  tag_Sci_Fi                  7465 non-null   int64
 17  tag_Shounen                 7465 non-null   int64
 18  tag_Original_Work           7465 non-null   int64
 19  tag_Non_Human_Protagonists  7465 non-null   int64
 20  tag_Drama                   7465 non-null   int64
 21  tag_Adventure               7465 non-null   int64
 22  tag_Family_Friendly         7465 non-null   int64
 23  tag_Short_Episodes          7465 non-null   int64
 24  tag_School_Life             7465 non-null   int64
 25  tag_Romance                 7465 non-null   int64
 26  tag_Shorts                  7465 non-null   int64
 27  tag_Slice_of_Life           7465 non-null   int64
 28  tag_Seinen                  7465 non-null   int64
 29  tag_Supernatural            7465 non-null   int64
 30  tag_Magic                   7465 non-null   int64
 31  tag_Animal_Protagonists     7465 non-null   int64
 32  tag_Ecchi                   7465 non-null   int64
 33  tag_Mecha                   7465 non-null   int64
 34  tag_Based_on_a_Light_Novel  7465 non-null   int64
 35  tag_CG_Animation            7465 non-null   int64
 36  tag_Superpowers             7465 non-null   int64
 37  tag_Others                  7465 non-null   int64
 38  tag_missing                 7465 non-null   int64
dtypes: bool(1), float64(1), int64(34), object(3)
memory usage: 2.2+ MB
```

```
x=pd.get_dummies(x, columns=x.select_dtypes(include=['object','category']).columns.tolist(),drop_first=True)
x.head()
```

	duration	ongoing	years_running	studios_colab	contentWarn	watching	wantWatch	dropped	votes	tag_Based_on_a_Manga	...	studio_primary_Shaft	studio_primary_Shin-Ei Animation	studio_primary_Studio Deen	s
1	107.0	False	0	0	0	1453	21733	124	43960	0	...	0	0	0	
2	130.0	False	0	0	1	946	17148	132	33752	1	...	0	0	0	
8	111.0	False	0	0	0	280	6624	150	6254	1	...	0	0	0	
27	125.0	False	0	0	0	589	12388	161	82752	0	...	0	0	0	
31	117.0	False	0	0	0	538	15651	130	26465	0	...	0	0	0	

5 rows × 69 columns

```
x.drop(columns='ongoing',inplace=True)
```

```
x.info()
```

```
 18  tag_Family_Friendly                  7465 non-null   int64
 19  tag_Short_Episodes                   7465 non-null   int64
 20  tag_School_Life                      7465 non-null   int64
 21  tag_Romance                          7465 non-null   int64
 22  tag_Shorts                           7465 non-null   int64
 23  tag_Slice_of_Life                    7465 non-null   int64
 24  tag_Seinen                           7465 non-null   int64
 25  tag_Supernatural                     7465 non-null   int64
 26  tag_Magic                            7465 non-null   int64
 27  tag_Animal_Protagonists              7465 non-null   int64
 28  tag_Ecchi                            7465 non-null   int64
 29  tag_Mecha                            7465 non-null   int64
 30  tag_Based_on_a_Light_Novel           7465 non-null   int64
 31  tag_CG_Animation                     7465 non-null   int64
 32  tag_Superpowers                      7465 non-null   int64
 33  tag_Others                           7465 non-null   int64
 34  tag_missing                          7465 non-null   int64
 35  mediaType_Movie                      7465 non-null   uint8
 36  mediaType_Music Video                7465 non-null   uint8
 37  mediaType_OVA                        7465 non-null   uint8
 38  mediaType_Other                      7465 non-null   uint8
 39  mediaType_TV                         7465 non-null   uint8
 40  mediaType_TV Special                 7465 non-null   uint8
 41  mediaType_Web                        7465 non-null   uint8
 42  mediaType_is_missing                 7465 non-null   uint8
 43  sznOfRelease_Spring                  7465 non-null   uint8
 44  sznOfRelease_Summer                  7465 non-null   uint8
 45  sznOfRelease_Winter                  7465 non-null   uint8
 46  sznOfRelease_is_missing              7465 non-null   uint8
 47  studio_primary_AIC                   7465 non-null   uint8
 48  studio_primary_Bones                 7465 non-null   uint8
 49  studio_primary_DLE                   7465 non-null   uint8
 50  studio_primary_GONZO                 7465 non-null   uint8
 51  studio_primary_J.C. Staff            7465 non-null   uint8
 52  studio_primary_Kyoto Animation       7465 non-null   uint8
 53  studio_primary_MADHOUSE              7465 non-null   uint8
 54  studio_primary_Nippon Animation      7465 non-null   uint8
 55  studio_primary_OLM                   7465 non-null   uint8
 56  studio_primary_Others                7465 non-null   uint8
 57  studio_primary_Production I.G        7465 non-null   uint8
 58  studio_primary_Shaft                 7465 non-null   uint8
 59  studio_primary_Shin-Ei Animation     7465 non-null   uint8
 60  studio_primary_Studio Deen           7465 non-null   uint8
 61  studio_primary_Studio Pierrot        7465 non-null   uint8
 62  studio_primary_Sunrise               7465 non-null   uint8
 63  studio_primary_TMS Entertainment     7465 non-null   uint8
 64  studio_primary_Tatsunoko Production  7465 non-null   uint8
 65  studio_primary_Toei Animation        7465 non-null   uint8
 66  studio_primary_XEBEC                 7465 non-null   uint8
 67  studio_primary_is_missing            7465 non-null   uint8
dtypes: float64(1), int64(34), uint8(33)
memory usage: 2.3 MB
```

```python
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error,r2_score, mean_absolute_error
```

```python
X_train, X_test,Y_train, Y_test= train_test_split(x,y, test_size=0.2, random_state=1)
```

```
print("Number of samples for train", X_train.shape[0])
print("Number of samples for test", X_test.shape[0])
```

```
    Number of samples for train 5972
    Number of samples for test 1493
```

```
lin_model= LinearRegression()
lin_model.fit(X_train, Y_train)
```

```
    ▾ LinearRegression
    LinearRegression()
```

```
def Model_performance(model,predictor,target):
  pred=model.predict(predictor)
  r2=r2_score(target,pred)
  rmse=np.sqrt(mean_squared_error(target,pred))

  results=pd.DataFrame({
      "RMSE":rmse,
      "R2 SCORE":r2
  }, index=[0])
  return results
```

```
print("Training Data Performance")
lin_model_train= Model_performance(lin_model, X_train, Y_train)
lin_model_train
```

```
    Training Data Performance
           RMSE  R2 SCORE
    0  0.580109  0.515527
```

```
print("Test Data Performance")
lin_model_test= Model_performance(lin_model, X_test, Y_test)
lin_model_test
```

```
    Test Data Performance
           RMSE  R2 SCORE
    0  0.566959  0.514441
```

```
x.columns
```

```
    Index(['duration', 'years_running', 'studios_colab', 'contentWarn', 'watching',
           'wantWatch', 'dropped', 'votes', 'tag_Based_on_a_Manga', 'tag_Comedy',
           'tag_Action', 'tag_Fantasy', 'tag_Sci_Fi', 'tag_Shounen',
           'tag_Original_Work', 'tag_Non_Human_Protagonists', 'tag_Drama',
           'tag_Adventure', 'tag_Family_Friendly', 'tag_Short_Episodes',
           'tag_School_Life', 'tag_Romance', 'tag_Shorts', 'tag_Slice_of_Life',
           'tag_Seinen', 'tag_Supernatural', 'tag_Magic',
           'tag_Animal_Protagonists', 'tag_Ecchi', 'tag_Mecha',
           'tag_Based_on_a_Light_Novel', 'tag_CG_Animation', 'tag_Superpowers',
           'tag_Others', 'tag_missing', 'mediaType_Movie', 'mediaType_Music Video',
           'mediaType_OVA', 'mediaType_Other', 'mediaType_TV',
           'mediaType_TV Special', 'mediaType_Web', 'mediaType_is_missing',
           'sznOfRelease_Spring', 'sznOfRelease_Summer', 'sznOfRelease_Winter',
           'sznOfRelease_is_missing', 'studio_primary_AIC', 'studio_primary_Bones',
           'studio_primary_DLE', 'studio_primary_GONZO',
           'studio_primary_J.C. Staff', 'studio_primary_Kyoto Animation',
           'studio_primary_MADHOUSE', 'studio_primary_Nippon Animation',
           'studio_primary_OLM', 'studio_primary_Others',
           'studio_primary_Production I.G', 'studio_primary_Shaft',
           'studio_primary_Shin-Ei Animation', 'studio_primary_Studio Deen',
           'studio_primary_Studio Pierrot', 'studio_primary_Sunrise',
           'studio_primary_TMS Entertainment',
           'studio_primary_Tatsunoko Production', 'studio_primary_Toei Animation',
           'studio_primary_XEBEC', 'studio_primary_is_missing'],
          dtype='object')
```

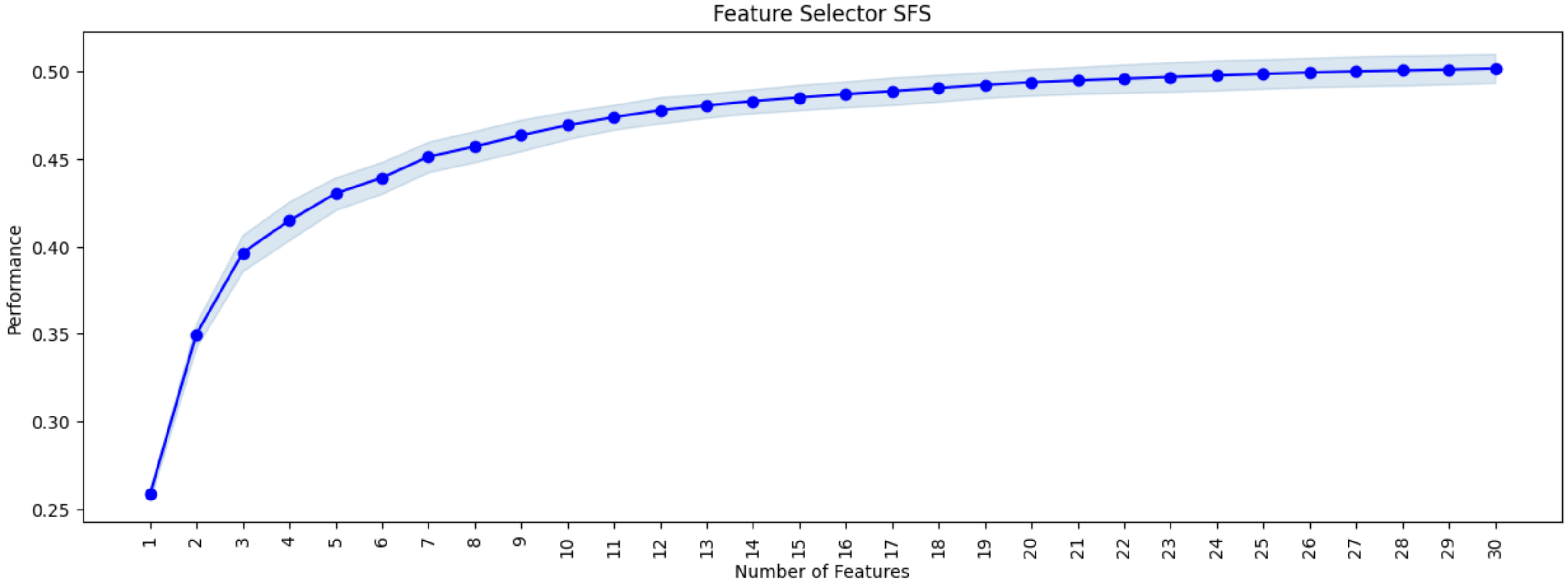
Feature Selection Technique

```
from mlxtend.feature_selection import SequentialFeatureSelector as SFS
reg= LinearRegression()
sfs = SFS(reg, k_features = X_train.shape[1],
          forward= True, floating=False, scoring='r2', n_jobs= -1,cv=5)
sfs=sfs.fit(X_train, Y_train)
```

```
from mlxtend.plotting import plot_sequential_feature_selection as plot_sfs
fig1= plot_sfs(sfs.get_metric_dict(),kind='std_err', figsize=(15,5))
plt.title("Feature Selector SFS")
plt.xticks(rotation=90)
plt.show()
```

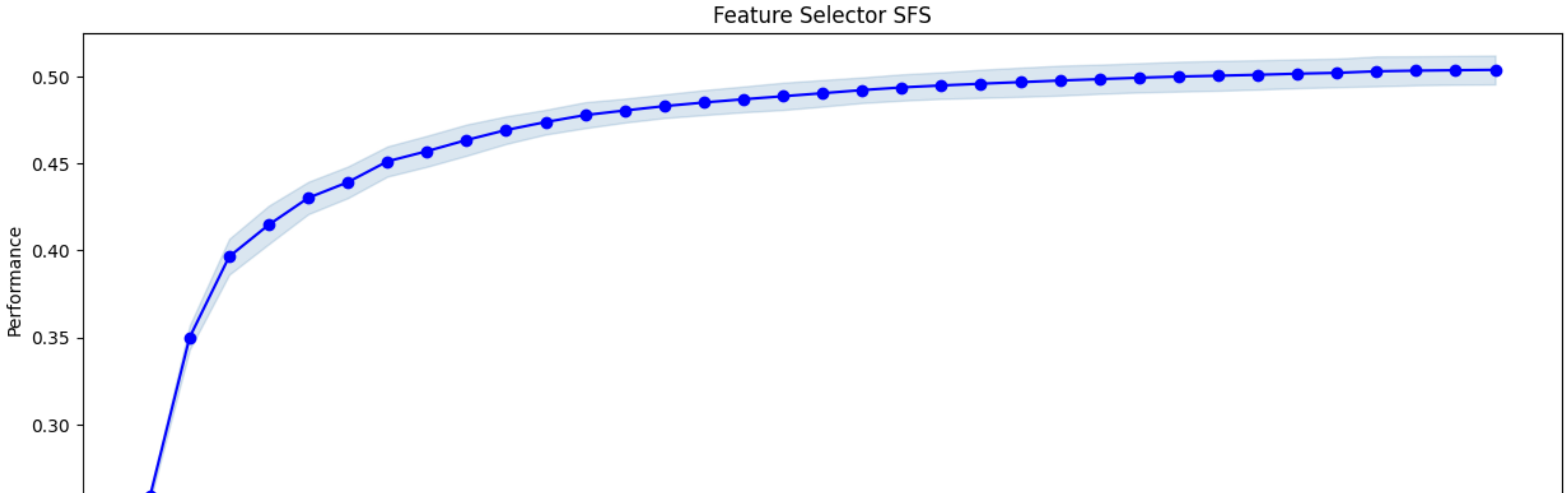
```
from mlxtend.feature_selection import SequentialFeatureSelector as SFS
reg= LinearRegression()
sfs = SFS(reg, k_features = 30,
          forward= True, floating=False, scoring='r2', n_jobs= -1,cv=5)
sfs=sfs.fit(X_train, Y_train)
```

```
from mlxtend.plotting import plot_sequential_feature_selection as plot_sfs
fig1= plot_sfs(sfs.get_metric_dict(),kind='std_err', figsize=(15,5))
plt.title("Feature Selector SFS")
plt.xticks(rotation=90)
plt.show()
```



```
from mlxtend.feature_selection import SequentialFeatureSelector as SFS
reg= LinearRegression()
sfs = SFS(reg, k_features = 35,
          forward= True, floating=False, scoring='r2', n_jobs= -1,cv=5)
sfs=sfs.fit(X_train, Y_train)
```

```
from mlxtend.plotting import plot_sequential_feature_selection as plot_sfs
fig1= plot_sfs(sfs.get_metric_dict(),kind='std_err', figsize=(15,5))
plt.title("Feature Selector SFS")
plt.xticks(rotation=90)
plt.show()
```



```
feature_index= list(sfs.k_feature_idx_)
print(feature_index)
```

```
[0, 1, 3, 5, 6, 7, 8, 10, 11, 12, 13, 16, 18, 19, 20, 22, 23, 24, 25, 30, 35, 37, 38, 39, 41, 48, 49, 56, 57, 60, 62, 63, 65, 66, 67]
```

```
X_train.columns[feature_index]
```

```
Index(['duration', 'years_running', 'contentWarn', 'wantWatch', 'dropped',
       'votes', 'tag_Based_on_a_Manga', 'tag_Action', 'tag_Fantasy',
       'tag_Sci_Fi', 'tag_Shounen', 'tag_Drama', 'tag_Family_Friendly',
       'tag_Short_Episodes', 'tag_School_Life', 'tag_Shorts',
       'tag_Slice_of_Life', 'tag_Seinen', 'tag_Supernatural',
       'tag_Based_on_a_Light_Novel', 'mediaType_Movie', 'mediaType_OVA',
       'mediaType_Other', 'mediaType_TV', 'mediaType_Web',
       'studio_primary_Bones', 'studio_primary_DLE', 'studio_primary_Others',
       'studio_primary_Production I.G', 'studio_primary_Studio Deen',
       'studio_primary_Sunrise', 'studio_primary_TMS Entertainment',
       'studio_primary_Toei Animation', 'studio_primary_XEBEC',
       'studio_primary_is_missing'],
      dtype='object')
```

```
X_train_final= X_train[X_train.columns[feature_index]]
X_test_final= X_test[X_test.columns[feature_index]]
```

```
lin_model_v2= LinearRegression()
lin_model_v2.fit(X_train_final, Y_train)
```

```
▾ LinearRegression
LinearRegression()
```

```
print("Training Data Performance")
lin_model_train= Model_performance(lin_model, X_train, Y_train)
```

```
lin_model_train
```

```
Training Data Performance
```

	RMSE	R2 SCORE
0	0.580109	0.515527

```
print("Training Data Performance")
lin_model_train= Model_performance(lin_model, X_test, Y_test)
lin_model_train
```

```
Training Data Performance
```

	RMSE	R2 SCORE
0	0.566959	0.514441

```
X_train.columns[feature_index]
```

```
Index(['duration', 'years_running', 'contentWarn', 'wantWatch', 'dropped',
       'votes', 'tag_Based_on_a_Manga', 'tag_Action', 'tag_Fantasy',
       'tag_Sci_Fi', 'tag_Shounen', 'tag_Drama', 'tag_Family_Friendly',
       'tag_Short_Episodes', 'tag_School_Life', 'tag_Shorts',
       'tag_Slice_of_Life', 'tag_Seinen', 'tag_Supernatural',
       'tag_Based_on_a_Light_Novel', 'mediaType_Movie', 'mediaType_OVA',
       'mediaType_Other', 'mediaType_TV', 'mediaType_Web',
       'studio_primary_Bones', 'studio_primary_DLE', 'studio_primary_Others',
       'studio_primary_Production I.G', 'studio_primary_Studio Deen',
       'studio_primary_Sunrise', 'studio_primary_TMS Entertainment',
       'studio_primary_Toei Animation', 'studio_primary_XEBEC',
       'studio_primary_is_missing'],
      dtype='object')
```

```
X_train.head(2)
```

	duration	years_running	studios_colab	contentWarn	watching	wantWatch	dropped	votes	tag_Based_on_a_Manga	tag_Comedy	...	studio_primary_Shaft	studio_primary_Shin-Ei Animation	studio_primary_Stu De
8843	1.0	0	0	0	0	17	0	14	0	0	...	0	0	
1599	24.0	0	0	0	24	1117	15	735	1	1	...	1	0	

2 rows × 68 columns

Amna Islam