```
!pip install category encoders
```

```
Collecting category
  Downloading category-0.1.0-py3-none-any.whl (17 kB)
Collecting encoders
  Downloading encoders-0.0.3-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (594 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 594.5/594.5 kB 5.3 MB/s eta 0:00:00
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from category) (1.23.5)
Collecting rust-category (from category)
  Downloading rust_category-0.2.0-cp310-cp310-manylinux_2_5_x86_64.manylinux1_x86_64.whl (975 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 975.7/975.7 kB 8.5 MB/s eta 0:00:00
Installing collected packages: rust-category, encoders, category
Successfully installed category-0.1.0 encoders-0.0.3 rust-category-0.2.0
```

```
import statsmodels.regression.linear_model as smf
```

```
import statsmodels.api as sm
```

```
import warnings
warnings.filterwarnings('ignore')
```

```
from sklearn.model_selection import cross_val_score, train_test_split
```

```
import seaborn as sns
```

```
import numpy as np
```

```
import pandas as pd
```

```
movies_df = pd.read_csv("/content/archive.zip", encoding = 'latin-1')
movies_df.head()
```

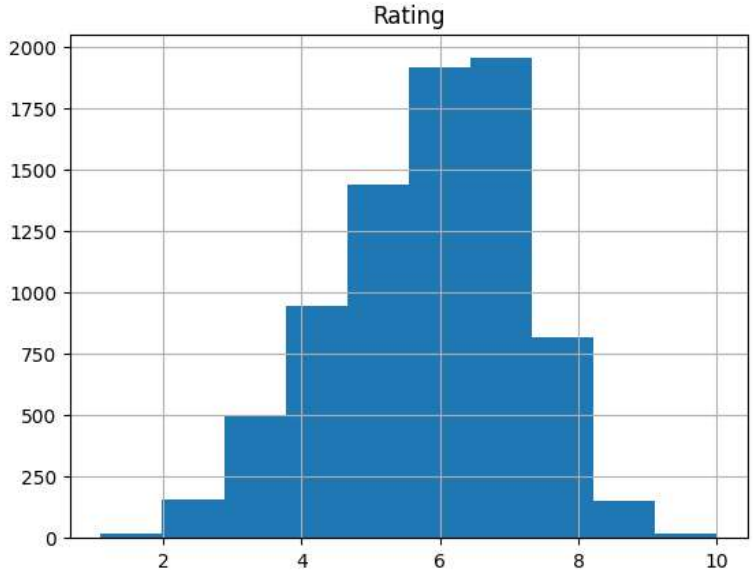
	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
0		NaN	NaN	Drama	NaN	NaN	J.S. Randhawa	Manmauji	Birbal	Rajendra Bhatia
1	#Gadhvi (He thought he was Gandhi)	(2019)	109 min	Drama	7.0	8	Gaurav Bakshi	Rasika Dugal	Vivek Ghamande	Arvind Jangid
2	#Homecoming	(2021)	90 min	Drama, Musical	NaN	NaN	Soumyajit Majumdar	Sayani Gupta	Plabita Borthakur	Roy Angana
3	#Yaaram	(2019)	110 min	Comedy, Romance	4.4	35	Ovais Khan	Prateik	Ishita Raj	Siddhant Kapoor
4	...And Once Again	(2010)	105 min	Drama	NaN	NaN	Amol Palekar	Rajat Kapoor	Rituparna Sengupta	Antara Mali

```
movies_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15509 entries, 0 to 15508
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Name      15509 non-null  object 
 1   Year      14981 non-null  object 
 2   Duration  7240 non-null   object 
 3   Genre     13632 non-null  object 
 4   Rating    7919 non-null   float64
 5   Votes     7920 non-null   object 
 6   Director  14984 non-null  object 
 7   Actor 1   13892 non-null  object 
 8   Actor 2   13125 non-null  object 
 9   Actor 3   12365 non-null  object 
dtypes: float64(1), object(9)
memory usage: 1.2+ MB
```

```
movies_df.hist()
```

```
array([[<Axes: title={'center': 'Rating'}>]], dtype=object)
```




```
rest_df = pd.read_csv('/content/archive.zip', encoding='latin1')
```

```
train_df, validation_df = train_test_split(rest_df, train_size = 0.75, random_state = 101)
train_df.head(2)
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
11868	Reshmi Sari	(1940)	NaN	NaN	NaN	NaN	G.P. Pawar	Vasant	NaN	NaN
10859	Pinjre Ke Panchhi	(1966)	NaN	Crime, Drama	5.8	13	Salil Choudhury	Balraj Sahni	Meena Kumari	Mehmood

```
validation_df.head(2)
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
14114	The Hundred Bucks	(2020)	102 min	Drama	6.8	10	Dushyant Pratap Singh	Kavita Tripathi	Rajesh Mishra	Dinesh Bawara
3386	Daman	(1951)	NaN	Drama,	NaN	NaN	Nanabhai Bhatt	Ajit	Nigar Sultana	Pran

```
test_df = movies_df
```

```
test_df[movies_df.duplicated(keep = False)]
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor 3
1249	Arab Ka Sona - Abu Kaalia	(1979)	NaN	Action	NaN	NaN	Master Bhagwan	Meena Rai	Dara Singh	NaN
1250	Arab Ka Sona - Abu Kaalia	(1979)	NaN	Action	NaN	NaN	Master Bhagwan	Meena Rai	Dara Singh	NaN
1768	Balidan	(1992)	NaN	Drama	NaN	NaN	NaN	NaN	NaN	NaN
1769	Balidan	(1992)	NaN	Drama	NaN	NaN	NaN	NaN	NaN	NaN
4722	First Time - Pehli Baar	(2009)	NaN	NaN	NaN	NaN	Raja Bundela	Zeenat Aman	Nitin Arora	Raj Babbar
4723	First Time - Pehli Baar	(2009)	NaN	NaN	NaN	NaN	Raja Bundela	Zeenat Aman	Nitin Arora	Raj Babbar
9712	Musafir	NaN	NaN	Thriller	NaN	NaN	Shiva Dagar	NaN	NaN	NaN
9713	Musafir	NaN	NaN	Thriller	NaN	NaN	Shiva Dagar	NaN	NaN	NaN
13068	Shivani	(2019)	NaN	Crime	NaN	NaN	Ugresh Prasad Ujala	Santosh	NaN	NaN
13069	Shivani	(2019)	NaN	Crime	NaN	NaN	Ugresh Prasad Ujala	Santosh	NaN	NaN
13307	Slumdog Karodpati	(2019)	118 min	Thriller	NaN	NaN	Rajesh Patole	Udhav Garje	Rahul Gavane	Govindrao
13308	Slumdog Karodpati	(2019)	118 min	Thriller	NaN	NaN	Rajesh Patole	Udhav Garje	Rahul Gavane	Govindrao

```
movies_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15509 entries, 0 to 15508
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Name      15509 non-null  object 
 1   Year      14981 non-null  object 
 2   Duration  7240 non-null   object 
 3   Genre     13632 non-null  object 
 4   Rating    7919 non-null   float64
 5   Votes     7920 non-null   object 
 6   Director  14984 non-null  object 
 7   Actor 1   13892 non-null  object 
 8   Actor 2   13125 non-null  object 
 9   Actor 3   12365 non-null  object 
dtypes: float64(1), object(9)
memory usage: 1.2+ MB
```

```
train_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 11631 entries, 11868 to 13151
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Name      11631 non-null  object 
 1   Year      11242 non-null  object 
 2   Duration  5490 non-null   object 
 3   Genre     10231 non-null  object 
```

```
 4   Rating    5969 non-null   float64
 5   Votes     5969 non-null   object
 6   Director  11240 non-null  object
 7   Actor 1   10390 non-null  object
 8   Actor 2   9829 non-null   object
 9   Actor 3   9281 non-null   object
dtypes: float64(1), object(9)
memory usage: 999.5+ KB
```

```
train_df.describe()
```

	Rating
count	5969.000000
mean	5.855336
std	1.377054
min	1.100000
25%	4.900000
50%	6.000000
75%	6.800000
max	10.000000

```
filled_df = pd.concat([train_df, validation_df], axis = 0)
filled_df.head(2)
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor
11868	Reshmi Sari	(1940)	NaN	NaN	NaN	NaN	G.P. Pawar	Vasant	NaN	Na

```
filled_df = filled_df.values
train_df = train_df.values
validation_df = validation_df.values
```

```
from sklearn.linear_model import ElasticNet, Lasso, LinearRegression, Ridge
from sklearn.metrics import r2_score, mean_squared_error
```

```
from sklearn.neighbors import KNeighborsRegressor
```

```
import lightgbm as lgb
import matplotlib.pyplot as plt
import os
import pandas as pd
import seaborn as sns
import statsmodels.api as sm
import statsmodels.regression.linear_model as smf
from sklearn.ensemble import GradientBoostingRegressor, RandomForestRegressor
from sklearn.linear_model import ElasticNet, Lasso, LinearRegression, Ridge
from sklearn.metrics import r2_score, mean_squared_error
from sklearn.neighbors import KNeighborsRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor
from statsmodels.formula.api import ols

sns.set()
%matplotlib inline
```

```
train_df, validation_df = train_test_split(rest_df, train_size = 0.75, random_state = 101)
train_df.head(2)
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Actor
11868	Reshmi Sari	(1940)	NaN	NaN	NaN	NaN	G.P. Pawar	Vasant	NaN	Na

```
test_df[movies_df.duplicated(keep = False)]
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-1-29635e2bd856> in <cell line: 1>()
----> 1 test_df[movies_df.duplicated(keep = False)]

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

SEARCH STACK OVERFLOW

```
movies_df.describe(include = 'all')
```

	Name	Year	Duration	Genre	Rating	Votes	Director	Actor 1	Actor 2	Act
count	15509	14981	7240	13632	7919.000000	7920	14984	13892	13125	123
unique	13838	102	182	485	NaN	2034	5938	4718	4891	48
top	Anjaam	(2019)	120 min	Drama	NaN	8	Jayant Desai	Ashok Kumar	Rekha	Pr
freq	7	410	240	2780	NaN	227	58	158	83	
mean	NaN	NaN	NaN	NaN	5.841621	NaN	NaN	NaN	NaN	N
std	NaN	NaN	NaN	NaN	1.381777	NaN	NaN	NaN	NaN	N
min	NaN	NaN	NaN	NaN	1.100000	NaN	NaN	NaN	NaN	N
25%	NaN	NaN	NaN	NaN	4.900000	NaN	NaN	NaN	NaN	N
50%	NaN	NaN	NaN	NaN	6.000000	NaN	NaN	NaN	NaN	N
75%	NaN	NaN	NaN	NaN	6.800000	NaN	NaN	NaN	NaN	N

```
train_df.shape
```

```
(11631, 10)
```

```
train_missing = list(train_df.isnull().sum())
val_missing = list(validation_df.isnull().sum())
test_misisng = list(test_df.isnull().sum())
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-2-92844f8bf91a> in <cell line: 1>()
----> 1 train_missing = list(train_df.isnull().sum())
      2 val_missing = list(validation_df.isnull().sum())
      3 test_misisng = list(test_df.isnull().sum())

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

SEARCH STACK OVERFLOW

```
train_missing_percent = list(train_df.isnull().sum() / len(train_df) * 100)
val_missing_percent = list(validation_df.isnull().sum() / len(validation_df) * 100)
test_misisng_percent = list(test_df.isnull().sum() / len(test_df) * 100)
```

```
sns.displot(movies_df.isnull())
plt.title("Missing values in dataset")
plt.show()
```


```
sns.heatmap(movies_df.isnull(), cmap = 'viridis')
plt.title("Missing values in dataset")
plt.show()
```