In [2]: import numpy as np import matplotlib.pyplot as plt import seaborn as sns df = pd.read\_csv(r'/home/nineleaps/Documents/IMDB-Movie-Data.csv') In [6]: df.head() **Unnamed:** Out[7]: Runtime Revenue Title Genre Description Director Actors Year Rating Votes Metascore (Minutes) (Millions) 12 A group of Chris Pratt, Vin Diesel, Guardians of Action, Adventure, Sci-Fi intergalactic criminals James Gunn 121 8.1 757074.0 333.13 NaN the Galaxy Bradley Cooper, Zoe S... are forced ... Following clues to the Noomi Rapace, Logan Prometheus Adventure, Mystery, Sci-Fi 2012 7.0 485820.0 1 2 origin of mankind, a Ridley Scott 124 126.46 65.0 NaN Marshall-Green, Michael Fa... te... Three girls are M. Night James McAvoy, Anya Taylor-2 3 Split Horror, Thriller kidnapped by a man 117 7.3 157606.0 138.12 62.0 NaN Shyamalan Joy, Haley Lu Richar... with a diag... In a city of humanoid Matthew Christophe animals, a hustling McConaughey,Reese 2016 270.32 NaN 3 Sing Animation, Comedy, Family 108 7.2 60545.0 59.0 Lourdelet Witherspoon, Seth Ma... thea... A secret government Will Smith, Jared Leto, Margot 5 Suicide Squad Action, Adventure, Fantasy agency recruits some David Ayer 6.2 393727.0 325.02 NaN Robbie, Viola D... of th... df.shape In [10]: (1000, 13)Out[10]: print('Number of rows:', df.shape[0]) print('Number of columns:', df.shape[1]) Number of rows: 1000 Number of columns: 13 In [13]: df.info() <class 'pandas.core.frame.DataFrame'> RangeIndex: 1000 entries, 0 to 999 Data columns (total 13 columns): Column Non-Null Count Dtype -----0 Rank 1000 non-null int64 1 Title 1000 non-null object 2 1000 non-null object Genre 3 Description 1000 non-null object 1000 non-null 4 object Director 1000 non-null 5 Actors object 6 1000 non-null object Year 7 Runtime (Minutes) 1000 non-null int64 8 1000 non-null float64 Rating 9 Votes 1000 non-null float64 Revenue (Millions) 873 non-null float64 10 936 non-null float64 Metascore float64 Unnamed: 12 2 non-null dtypes: float64(5), int64(2), object(6) memory usage: 101.7+ KB In [20]: print('Any missing values?', df.isnull().values.any()) Any missing values? True df.isnull().sum() Rank Title 0 0 Genre 0 Description 0 Director 0 Actors 0 Year Runtime (Minutes) 0 0 Rating 0 Votes Revenue (Millions) 127 Metascore 64 Unnamed: 12 998 dtype: int64 Missing values by Heatmap sns.heatmap(df.isnull()) <Axes: > Out[23]:

- 1.0

39

In [1]: **import** pandas **as** pd

78 117 156 195 - 0.8 234 273 312 351 390 0.6 429 468 507 546 585 0.4 624 663 702 741 780 0.2 819 858 897 936 Genre Actors Rating Rank Description Runtime (Minutes) Revenue (Millions) Unnamed: 12 Metascore Director Dropping a unnamed column which has maximum null values

Genre

## df1= df.drop(df.columns[12], axis=1)

Description

df1.head() In [28]:

Title

Out[28]: Rank

In [27]:

000[20].	ı	Rank	Title	Genre	Description	Director	Actors	Year	(Minutes)	Rating	Votes	(Millions)	Metascore
	0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S	2014	121	8.1	757074.0	333.13	76.0
	1	2	Prometheus	Adventure, Mystery, Sci-Fi	Following clues to the origin of mankind, a te	Ridley Scott	Noomi Rapace, Logan Marshall- Green, Michael Fa	2012	124	7.0	485820.0	126.46	65.0
	2	3	Split	Horror, Thriller	Three girls are kidnapped by a man with a diag	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar	2016	117	7.3	157606.0	138.12	62.0
	3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma	2016	108	7.2	60545.0	270.32	59.0
	4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D	2016	123	6.2	393727.0	325.02	40.0
In [31]:	per	centa	age_missing =	df1.isnull().sum()*	100/len(df1)								

Director

Revenue

Metascore

Runtime

Actors Year

Rating

Votes

0.0 0.0

0.0

0.0

Genre Description Director 0.0 Actors 0.0 Year Runtime (Minutes) 0.0 Rating 0.0 Votes 0.0 Revenue (Millions) 12.7 Metascore 6.4 dtype: float64 Dropping the rows of null values

percentage\_missing

Rank

Title

Out[31]:

In [33]:

## df1.dropna(axis=0, inplace= True)

percentage_mis	Dercentage_missing_new								
Rank	0.0								

percentage\_missing\_new = df1.isnull().sum()\*100/len(df1)

Out[36]: Title 0.0 Genre 0.0 Description

0.0 Director 0.0 Actors 0.0 0.0 Year Runtime (Minutes) 0.0 Rating 0.0 Votes 0.0 Revenue (Millions) 0.0 Metascore 0.0 dtype: float64 df1.shape (838, 12) Out[37]:

## Check for duplicate values dup\_value = df1.duplicated().any()

In [42]: print('Are there any dupliacted value in dataset?', dup\_value)

Are there any dupliacted value in dataset? False

In [ ]: # If there is any duplicated value then

# df1.drop\_duplicated()