Data Science 12th Project (Netflix Movies Dataset Cleaning)

Dr. Ahmed Elshaer

Mohamed Refaat - 211001955

Introduction

The dataset was acquired through web scraping of IMDB's top Netflix movies and shows page using Python's Beautiful Soup. This dataset serves as the foundation for developing a machine learning (ML) model aimed at predicting the genres of movies and shows.

I am keenly interested in working on this dataset owing to my avid interest in watching Netflix's movies and shows.

The dataset can be found on Kaggle at https://www.kaggle.com/datasets/bharatnatrayn/movies-dataset-for-feature-extracion-prediction?resource=download

The dataset comprises 9,999 records and encompasses 9 attributes.

The Attributes are:

- **MOVIES:** The movie's or the show's title.
- **YEAR:** The year(s) the movie or show has been running.
- **GENRE:** The movie's or the show's genre(s).
- **RATING:** The movie's or the show's IMDb fan rating out of 10.
- **ONE-LINE:** A short description of the movie or the show.
- **STARS:** A brief list of the director, main actors, and actresses.
- **VOTES:** The number of people who voted for the movie's or show's rating.
- **RunTime:** The duration of the movie in minutes or the show's average episode duration in minutes.
- **Gross:** The total worldwide earnings of the movie or the show in dollars.

Importing Necessary Libraries

In [ ]:

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

Loading the Dataset and Initial Exploration

In [ ]:

```
# Loading in the csv file as a pandas dataframe
Netflix_df = pd.read_csv("./Netflix Dataset.csv")
```

In [ ]:

```
# Printing the first 5 rows of the dataframe
Netflix_df.head()
```

Out[ ]:

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTin
0	Blood Red Sky	(2021)	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\n| \n Star...	21,062	121
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870	25
2	The Walking Dead	(2010–2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	885,805	44
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	414,849	23
4	Army of Thieves	(2021)	\nAction, Crime, Horror	NaN	\nA prequel, set before the events of Army of ...	\n Director:\nMatthias Schweighöfer\n| \n ...	NaN	Na

In [ ]:

```
# Describe your dataset using function describe() and explain the output in English
Netflix_df.describe()
```

Out[ ]:

	RATING	RunTime
count	8179.000000	7041.000000
mean	6.921176	68.688539
std	1.220232	47.258056
min	1.100000	1.000000
25%	6.200000	36.000000
50%	7.100000	60.000000
75%	7.800000	95.000000
max	9.900000	853.000000

The data description for the **Ratings** and **RunTime** columns is as follows:

- **RATING:**
    - Count: 8179 entries
    - Mean rating: 6.92 out of 10
    - Standard deviation: 1.22
    - Minimum rating: 1.1
    - 25th percentile rating: 6.2
    - Median rating (50th percentile): 7.1
    - 75th percentile rating: 7.8
    - Maximum rating: 9.9
- **RunTime:**
    - Count: 7041 entries
    - Mean duration: 68.69 minutes
    - Standard deviation: 47.26
    - Minimum duration: 1 minute
    - 25th percentile duration: 36 minutes
    - Median duration (50th percentile): 60 minutes
    - 75th percentile duration: 95 minutes
    - Maximum duration: 853 minutes

In [ ]:

```
# Check if there are any missing values in the dataset
Netflix_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9999 entries, 0 to 9998
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   MOVIES    9999 non-null   object 
 1   YEAR      9355 non-null   object 
 2   GENRE     9919 non-null   object 
 3   RATING    8179 non-null   float64
 4   ONE-LINE  9999 non-null   object 
 5   STARS     9999 non-null   object 
 6   VOTES     8179 non-null   object 
 7   RunTime   7041 non-null   float64
 8   Gross     460 non-null    object 
dtypes: float64(2), object(7)
memory usage: 703.2+ KB
```

In [ ]:

```
# Getting the original number of columns and rows
original_num_rows = Netflix_df.shape[0]
original_num_columns = Netflix_df.shape[1]
print("Number of records in the original dataset:", original_num_rows)
print("Number of attributes in the original dataset:", original_num_columns)
```

```
Number of records in the original dataset: 9999
Number of attributes in the original dataset: 9
```

Data Cleaning

In [ ]:

```
# Drop unnecessary columns in the data frame
Netflix_df[["Gross"]].isnull().sum()
```

Out[ ]:

```
Gross    9539
dtype: int64
```

In [ ]:

```
# Drop unnecessary columns in the data frame
Netflix_df.drop(
    ["Gross"], axis=1, inplace=True
)  # Deleting the "Gross" attribute since it doesn't contain enough data

Netflix_df.sample(5)
```

Out[ ]:

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	Ru
7420	Julie and the Phantoms	(2020–)	\nComedy, Drama, Family	9.1	\nJulie rehearses with Nick - while daydreamin...	\n Director:\nKabir Akhtar\n| \n Stars:\...	429	
7369	Feel Good	(2020–2021)	\nComedy, Drama, Romance	7.9	\nThe arrival of Noughties heart-throb Arnie R...	\n Director:\nAlly Pankiw\n| \n Stars:\n...	226	
7858	She-Ra and the Princesses of Power	(2018–2020)	\nAnimation, Action, Adventure	8.8	\nCatra orders Scorpia to find Entrapta's reco...	\n Director:\nChristina Manrique\n| \n S...	936	
1405	Mongol	(2007)	\nAction, Biography, Drama	7.2	\nThe story recounts the early life of Genghis...	\n Director:\nSergei Bodrov\n| \n Stars:...	45,408	
4154	Zinzana	(2015)	\nCrime, Drama, Thriller	7.0	\nTrapped in a prison cell ("Zinzana") in a re...	\n Director:\nMajid Al Ansari\n| \n Star...	1,428	

In [ ]:

```
# If a column name is not expressive, rename it
Netflix_df.rename(
    columns={
        "MOVIES": "Movies",
        "YEAR": "Year",
        "GENRE": "Genre",
        "RATING": "Fan Rating",
        "ONE-LINE": "Description",
        "STARS": "Main Cast",
        "VOTES": "Votes",
        "RunTime": "Run Time",
    },
    inplace=True,
)

Netflix_df.sample(5)
```

Out[ ]:

	Movies	Year	Genre	Fan Rating	Description	Main Cast	Votes	I T
1475	The Innocents	(2018)	\nDrama, Horror, Mystery	6.3	\nWhen teenagers Harry and June run away from ...	\n \n Stars:\nSorcha Groundsell,...	5,470	(
1966	The Great Hack	(2019)	\nDocumentary, Biography, History	7.1	\nThe Cambridge Analytica scandal is examined ...	\n Directors:\nKarim Amer, \nJehane Noujaim...	21,781	1
4567	Hidden Worlds	(2018)	\nDrama	7.2	\nWhile investigating an actress's supposed su...	\n \n Stars:\nNizar Nasser Seif,...	471	N
932	Kara Para Ask	(2014–2015)	\nAction, Crime, Drama	7.3	\nOmar is a police officer. After the death of...	\n \n Stars:\nEngin Akyürek, \nT...	5,300	9
7225	Family Business	(2019–)	\nComedy	7.8	\nAs the Hazans divvy up responsibilities for ...	\n Director:\nIgor Gotesman\n| \n Stars:...	71	N

In [ ]:

```python
# Find and remove duplicate values
Netflix_df.drop_duplicates(inplace=True)
removed_dup_num_rows = Netflix_df.shape[0]
print(original_num_rows - removed_dup_num_rows, "duplicate rows were dropped")
```

```
431 duplicate rows were dropped
```

In [ ]:

```python
# Handle missing data (Nan and None values)
Netflix_df.dropna(how="any", inplace=True)
```

In [ ]:

```python
# Display 5 random rows from the dataset
Netflix_df.sample(5)
```

Out[ ]:

	Movies	Year	Genre	Fan Rating	Description	Main Cast	Votes
4913	Revenge of the Electric Car	(2011)	\nDocumentary	7.2	\nDirector Chris Paine takes his film crew beh...	\n Director:\nChris Paine\n| \n Stars:\n...	1,808
6489	The Boondocks	(2005–2014)	\nAnimation, Action, Comedy	8.0	\nHuey and Riley team up to convince Granddad ...	\n Director:\nAnthony Bell\n| \n Stars:\...	381
1603	Pieles	(2017)	\nComedy, Drama, Fantasy	6.2	\nIn a strange world where people share numero...	\n Director:\nEduardo Casanova\n| \n Sta...	5,206
2106	Coisa Mais Linda	(2019–)	\nDrama, Romance	7.9	\nAfter the disappearance of her husband, Mari...	\n \n Stars:\nMaria Casadevall, ...	3,071
3250	Erotica 2022	(2020)	\nDrama, Mystery, Romance	3.4	\nFive stories about women's issues, loosely c...	\n Directors:\nKasia Adamik, \nOlga Chajdas...	223

In [ ]:

```
# Splitting the "Main Cast" column into "Director" and "Actors & Actresses" attribu
Netflix_df[["Director", "Actors & Actresses"]] = Netflix_df["Main Cast"].str.split(
    "Stars:", expand=True
)
```

In [ ]:

```
# Display 5 random rows from the dataset
Netflix_df.sample(5)
```

Out[ ]:

	Movies	Year	Genre	Fan Rating	Description	Main Cast	Votes	Ru Tim
232	Orphan Black	(2013–2017)	\nAction, Drama, Sci-Fi	8.3	\nA streetwise hustler is pulled into a compel...	\n \n Stars:\nTatiana Maslany, \...	104,501	44.
6960	Luna Nera	(2020–)	\nDrama, Fantasy, Thriller	4.7	\nWhile women's voices plague her, Ade struggl...	\n \n Stars:\nAntonia Fotaras, \...	128	40.
4835	Reversing Roe	(2018)	\nDocumentary	7.5	\nA deep historical look at one of the most co...	\n Directors:\nRicki Stern, \nAnne Sundberg...	1,160	99.
2764	City of Tiny Lights	(2016)	\nCrime, Drama, Mystery	5.5	\nIn the teeming, multicultural metropolis of ...	\n Director:\nPete Travis\n| \n Stars:\n...	2,279	110.
966	Cargo	(I) (2017)	\nDrama, Horror, Sci-Fi	6.3	\nAfter an epidemic spreads all over Australia...	\n Directors:\nBen Howling, \nYolanda Ramke...	39,943	105.

In [ ]:

```
# Dropping the "Main Cast" column since it won't be needed
Netflix_df.drop(["Main Cast"], axis=1, inplace=True)
```

In [ ]:

```
# Removing the "\n" and the "|" from the "Director", "Actors & Actresses", "Descrip
Netflix_df["Director"] = Netflix_df["Director"].str.replace("\n", "")
Netflix_df["Director"] = Netflix_df["Director"].str.replace("|", "")

Netflix_df["Actors & Actresses"] = Netflix_df["Actors & Actresses"].str.replace("\n
Netflix_df["Actors & Actresses"] = Netflix_df["Actors & Actresses"].str.replace("|"

Netflix_df["Description"] = Netflix_df["Description"].str.replace("\n", "")
Netflix_df["Description"] = Netflix_df["Description"].str.replace("|", "")

Netflix_df["Genre"] = Netflix_df["Genre"].str.replace("\n", "")
Netflix_df["Genre"] = Netflix_df["Genre"].str.replace("|", "")
```

```
Netflix_df.sample(5)
```

Out[ ]:

	Movies	Year	Genre	Fan Rating	Description	Votes	Run Time	Director
643	F Is for Family	(2015–)	Animation, Comedy, Drama	8.0	Follow the Murphy family back to the 1970s, wh...	31,263	30.0	
4649	Côpusu pâtî	(2015)	Horror	4.7	Facing goodbyes and graduation, Naomi Nakashim...	420	93.0	Director:Masafumi Yamada
4470	A Primeira Tentação de Cristo	(2019)	Comedy	4.6	Jesus, who's hitting the big 3-0, brings a sur...	3,267	46.0	Director:Rodrigo Van Der Put
6241	It's Not Over	(2014)	Documentary, Adventure, Drama	6.3	A fresh look on HIV/AIDS through the lens of M...	46	72.0	Director:Andrew Jenks
690	Le Guide de la famille parfaite	(2021)	Comedy, Drama	6.4	A couple in Québec deals with the pitfalls, pr...	611	102.0	Director:Ricardo Trogi

In [ ]:

```
# Reordering the columns
Netflix_df = Netflix_df[
    [
        "Movies",
        "Genre",
        "Description",
        "Run Time",
        "Fan Rating",
        "Votes",
        "Year",
        "Director",
        "Actors & Actresses",
    ]
]

Netflix_df.sample(5)
```

Out[ ]:

	Movies	Genre	Description	Run Time	Fan Rating	Votes	Year	Director
4228	Dolly Parton: A MusiCares Tribute	Documentary, Music	In a star-studded evening of music and memorie...	55.0	6.5	165	(2021 TV Movie)	Star:Brooke Weirick
6538	Avatar: The Last Airbender	Animation, Action, Adventure	Aang is tested as the Avatar when he must help...	25.0	7.0	3,270	(2005–2008)	Director:Giancarlo Volpe
3402	Fate/Extra Last Encore	Animation, Action, Fantasy	Waking up in a strange virtual world with no r...	24.0	5.4	664	(2018)	
2707	Private Life	Crime, Drama, Romance	In a world where data is no longer private, co...	70.0	6.6	498	(2020)	
3311	Dancing Queen	Music, Reality-TV	A docu-series that follows Justin Johnson, aka...	45.0	7.0	1,448	(2018)	

In [ ]:

```
# Sorting data with ascending based on rating
Netflix_df.sort_values(by="Fan Rating", axis=0, ascending=True)
```

Out[ ]:

	Movies	Genre	Description	Run Time	Fan Rating	Votes	Year	Dire
1166	Raketsonyeondan	Comedy, Drama, Sport	A city kid is brought to the countryside by hi...	80.0	1.1	25,629	(2021–)	
5365	Defcon 2012	Sci-Fi	On October 30, 2009 an independent filmmaker a...	92.0	1.8	377	(2010)	Directo Chris Ande
4332	Way of the Vampire	Action, Horror, Thriller	After defeating Dracula, Van Helsing is grante...	82.0	2.0	1,593	(2005)	Directors:S Nean Br Eduardo Du
5044	River of Darkness	Action, Horror, Thriller	When brutal murdering starts in a small rivers...	105.0	2.1	876	(2011)	Director:B Koe
3528	Sinister Squad	Action, Comedy, Fantasy	When a supernatural cult threatens Earth, Alic...	90.0	2.1	1,025	(2016)	Director:Jer M. In
...	...	...	...	...	...	...	...	
7159	Avatar: The Last Airbender	Animation, Action, Adventure	As the Fire Nation continues its assault on th...	24.0	9.7	2,770	(2005–2008)	Director:D F
8509	Avatar: The Last Airbender	Animation, Action, Adventure	Zuko battles his sister with Katara's help for...	92.0	9.8	5,283	(2005–2008)	Director:Joac Dos Sa
9892	Dexter	Crime, Drama, Mystery	Dexter and Arthur find themselves on a collisi...	51.0	9.8	11,638	(2006–2013)	Director:S

	Movies	Genre	Description	Run Time	Fan Rating	Votes	Year	Dire
8510	Avatar: The Last Airbender	Animation, Action, Adventure	Aang's moment of truth arrives. Can he defeat ...	92.0	9.9	8,813	(2005–2008)	Director:Joac Dos Sa
7640	BoJack Horseman	Animation, Comedy, Drama	BoJack reconnects with faces from his past.	26.0	9.9	12,369	(2014–2020)	Director:/ Wir

6779 rows × 9 columns

In [ ]:

```python
# Dropping rows with empty value of the "Director" attribute
Netflix_df = Netflix_df[Netflix_df["Director"].str.strip() != ""]
```

In [ ]:

```python
# Changing the "Year" column to the "Release Year" column and clean it
def extract_year(string):
    return string[1:5]


Netflix_df["Release Year"] = Netflix_df["Year"].apply(lambda x: extract_year(str(x)
```

In [ ]:

```python
# Dropping the "Year" column since it won't be needed
Netflix_df.drop(["Year"], axis=1, inplace=True)
```

In [ ]:

```python
# Display 5 random rows from the dataset
Netflix_df.sample(5)
```

Out[ ]:

	Movies	Genre	Description	Run Time	Fan Rating	Votes	Director	Actors & Actresses
2464	Away	Crime, Drama	A story set in the north English seaside town ...	105.0	6.8	2,079	Director:David Blair	Timothy Spall, Juno Temple, Hayley Squires, Ma...
9827	This Is a Robbery: The World's Biggest Art Heist	Documentary, Crime, History	An informant and a sting operation open up new...	50.0	6.8	137	Director:Colin Barnicle	David Nadolski, Anthony Romano, Carmello Merli...
2537	BAC Nord	Crime, Thriller	A police brigade working in the dangerous nort...	105.0	6.2	54	Director:Cédric Jimenez	Gilles Lellouche, François Civil, Karim Leklou...
327	Spirit Untamed	Animation, Adventure, Family	Lucky Prescott's life is changed forever when ...	88.0	5.4	1,741	Directors:Elaine Bogan, Ennio Torresan	Isabela Merced, Marsai Martin, Mckenna Grace, ...
1852	Naboer	Horror, Mystery, Thriller	John has just been left by his girlfriend Ingr...	75.0	6.5	9,475	Directors:Pål Sletaune, Tony Spataro	Kristoffer Joner, Cecilie A. Mosli, Julia Scha...

In [ ]:

```
# Changing the "Release Year" and Votes attribute from string to nunmeric value
Netflix_df["Release Year"] = pd.to_numeric(Netflix_df["Release Year"], errors="coer
Netflix_df["Votes"] = pd.to_numeric(Netflix_df["Votes"], errors="coerce")
```

In [ ]:

```
# Checking for null values and attributes data types after the data cleaning
Netflix_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 4856 entries, 0 to 9963
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Movies              4856 non-null   object 
 1   Genre               4856 non-null   object 
 2   Description         4856 non-null   object 
 3   Run Time            4856 non-null   float64
 4   Fan Rating          4856 non-null   float64
 5   Votes               2388 non-null   float64
 6   Director            4856 non-null   object 
 7   Actors & Actresses  4436 non-null   object 
 8   Release Year        4622 non-null   float64
dtypes: float64(4), object(5)
memory usage: 379.4+ KB
```

```
In [ ]: # Handle missing data (Nan and None values) after the data cleaning
        Netflix_df.dropna(how="any", inplace=True)
```

```
In [ ]: # Checking for null values and attributes data types after the data cleaning
        Netflix_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2034 entries, 157 to 9963
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Movies              2034 non-null   object 
 1   Genre               2034 non-null   object 
 2   Description         2034 non-null   object 
 3   Run Time            2034 non-null   float64
 4   Fan Rating          2034 non-null   float64
 5   Votes               2034 non-null   float64
 6   Director            2034 non-null   object 
 7   Actors & Actresses  2034 non-null   object 
 8   Release Year        2034 non-null   float64
dtypes: float64(4), object(5)
memory usage: 158.9+ KB
```

```
In [ ]: # Describe the dataset after the data cleaning
        Netflix_df.describe()
```

Out[ ]:

	Run Time	Fan Rating	Votes	Release Year
count	2034.000000	2034.00000	2034.000000	2034.000000
mean	60.605211	6.77999	348.632252	2015.805310
std	32.307244	1.29342	276.205341	8.457088
min	1.000000	1.80000	5.000000	1932.000000
25%	30.000000	6.02500	125.250000	2015.000000
50%	50.000000	7.00000	262.000000	2018.000000
75%	90.000000	7.70000	544.000000	2020.000000
max	252.000000	9.40000	999.000000	2021.000000

Run Time: The data includes 2,034 entries for the duration of movies or show episodes. The shortest runtime is 1 minute, the longest is 252 minutes, with an average runtime of approximately 60.61 minutes and a standard deviation of around 32.31 minutes.

Fan Rating: The dataset contains 2,034 ratings for movies or shows from fans. The ratings range from a minimum of 1.8 to a maximum of 9.4, with an average rating of about 6.78 and a standard deviation of roughly 1.29.

Votes: There are 2,034 recorded counts of votes for movies or shows, ranging from a minimum of 5 votes to a maximum of 999 votes. The average number of votes is approximately 348.63, with a standard deviation of around 276.21.

Release Year: The dataset spans release years from 1932 to 2021 for movies or shows. The earliest release is in 1932, the latest is in 2021, with the majority of entries falling between 2015 and 2020. The median release year is 2018.

In [ ]:

```
# Reordering the indices after the data cleaning
Netflix_df.index = np.arange(len(Netflix_df))
```

In [ ]:

```
# Displaying the first 20 records of the dataset after the data cleaning
Netflix_df.head(20)
```

Out[ ]:

	Movies	Genre	Description	Run Time	Fan Rating	Votes	Director	Acto Actr
0	Bartkowiak	Action, Crime, Sport	After his brother dies in a car crash, a disgr...	91.0	5.0	398.0	Director:Daniel Markowicz	Pawlo Dor Szy Bobr
1	Hostage House	Thriller	When an ambitious realtor and her daughter are...	85.0	3.7	315.0	Director:David Benullo	Jer Taylor Terra Sw
2	Cider no yô ni kotoba ga wakiagaru	Animation, Drama, Romance	A meeting and romance starts between two peopl...	87.0	7.0	941.0	Director:Kyohei Ishiguro	Some Ichi Sug Kimb
3	Deep	Drama, Mystery, Sci-Fi	Four insomniac med school students are lured i...	101.0	4.9	738.0	Directors:Sita Likitvanichkul, Jetarin Rat...	Par Rikulsur Lertsitt S
4	Fondeados	Comedy	Two directionless millennial bros get high and...	97.0	4.2	161.0	Director:Marcos Bucay	Esca Ri Pol N -
5	Father of the Bride Part 3 (ish)	Short, Family	Twenty-five years after the second movie's rel...	26.0	7.0	669.0	Director:Nancy Meyers	I Ke k C Kim Willi
6	A Perfect Fit	Comedy, Drama, Romance	Sparks fly when a fashion blogger in Bali meet...	112.0	5.2	294.0	Director:Hadrah Daeng Ratu	Nadya Refal Gio Abra
7	A Second Chance: Rivals!	Drama, Family, Sport	It's city girls vs country girls in the most c...	89.0	5.0	218.0	Director:Clay Glen	Emily M Stella S A Tuom

	Movies	Genre	Description	Run Time	Fan Rating	Votes	Director	Acto Actr
8	Le Guide de la famille parfaite	Comedy, Drama	A couple in Québec deals with the pitfalls, pr...	102.0	6.4	611.0	Director:Ricardo Trogi	F Az B Bell E Bie
9	Kidô senshi Gandamu: Senkô no Hasauei	Animation, Action, Drama	Hathaway Noa, still haunted by his past action...	95.0	6.8	734.0	Director:Shûkô Murase	Kenshô Reina Ju Suw Sô
10	La nuée	Drama, Fantasy, Horror	To save her farm from bankruptcy, a single mot...	101.0	6.5	338.0	Director:Just Philippot	Su Br S Kham Narbo
11	My Amanda	Drama, Romance	Two unusually close friends share every aspect...	89.0	5.7	214.0	Director:Alessandra de Rossi	Alessa de Piolo Pa
12	Gekijouban Bishoujo Senshi Sailor Moon Eternal	Animation, Action, Adventure	When a dark power enshrouds the Earth after a ...	160.0	7.1	932.0	Director:Chiaki Kon	Ko Mits Step Sheh Hig
13	Sangre en la boca	Drama, Sport	Ramón Alvia is a professional boxer who, altho...	97.0	5.6	647.0	Director:Hernán Belón	Leon Sbar E Don Érica
14	Lethal Seduction	Thriller	High-school senior Mark has never minded his o...	80.0	5.1	824.0	Director:Nancy Leopardi	Rur Am De Dina M
15	What Is Life Worth	Biography, Drama, History	An attorney in Washington D.C. battles against...	118.0	6.6	287.0	Director:Sara Colangelo	St Mi Keaton, Ryan, La

	Movies	Genre	Description	Run Time	Fan Rating	Votes	Director	Acto Actr
16	Jungle Beat: The Movie	Animation, Adventure, Comedy	A homesick alien crash-lands his spaceship nea...	88.0	5.5	760.0	Director:Brent Dawes	I Me I Ri Gavin
17	Sanglyusahoe	Drama	A deputy curator of a chaebol-funded art galle...	120.0	5.5	793.0	Director:Hyuk Byun	Park H Soo Ae Gyu Le
18	Arlo the Alligator Boy	Animation, Adventure, Comedy	A young humanoid alligator travels to the big ...	90.0	6.7	794.0	Director:Ryan Crego	Mich Woo Ryan C Dee Br
19	Dry Martina	Comedy, Drama	Martina was a famous singer in Argentina durin...	95.0	5.8	842.0	Director:Che Sandoval	Anto C Pa Cont Gerald

In [ ]:
```
# Showing the number of attributes and records before and after the data cleaning
clean_num_rows = Netflix_df.shape[0]
clean_num_columns = Netflix_df.shape[1]

print("The Uncleaned Dataset had",original_num_columns,"attributes, and after clean
print("Since the 'GROSS' attribute got dropped and the 'STARS' attribute got split
print()
print("The Uncleaned Dataset had",original_num_rows,"records, and after cleaning it
```

```
The Uncleaned Dataset had 9 attributes, and after cleaning it has 9 attributes
Since the 'GROSS' attribute got dropped and the 'STARS' attribute got split the numb
er didn't change

The Uncleaned Dataset had 9999 records, and after cleaning it has 2034 records
```

Pandas Aggregate Functions

In [ ]:
```
# Finding the total number of votes for the entire dataset
total_votes = int(Netflix_df["Votes"].sum())
print("The total number of votes for the entire dataset is", total_votes)
```

```
The total number of votes for the entire dataset is 709118
```

In [ ]:
```
# Finding the mean of the fan rating for the entire dataset
mean_fan_rating = Netflix_df["Fan Rating"].mean()
```

```python
mean_fan_rating = round(mean_fan_rating, 2) # Rounding the mean to 2 decimal places

print("The mean of the fan rating for the entire dataset is", mean_fan_rating)
```

```
The mean of the fan rating for the entire dataset is 6.78
```

In [ ]:

```python
# Finding the median of the release year for the entire dataset
median_release_year = int(Netflix_df["Release Year"].median())
print("The median of the release year for the entire dataset is", median_release_ye

# List 5 random movies from the median release year
print("5 random movies from the median release year:")
Netflix_df[Netflix_df["Release Year"] == median_release_year].sample(5)
```

```
The median of the release year for the entire dataset is 2018
5 random movies from the median release year:
```

Out[ ]:

	Movies	Genre	Description	Run Time	Fan Rating	Votes	Director	Actor Actres
309	Coborâm la prima	Drama	Several passengers remain locked in a subway s...	84.0	7.2	602.0	Director:Tedy Necula	Iu Bu Grumaze Ce Vict Co
415	Up North	Drama	Up North explores love, friendship, and family...	99.0	5.6	130.0	Director:Tope Oshin	Ada Re At Nafe B Mich
585	Dos Cataluñas	Documentary, History, News	A documentary that tackles the ideological con...	116.0	6.4	648.0	Directors:Álvaro Longoria, Gerardo Olivare...	Luis M Anson, Arrima Arta
1975	Paradise PD	Animation, Action, Comedy	Dusty gives Fitz a new identity: "Family Feud"...	23.0	6.7	149.0	Director:Lauren Andrews	Sa Cha Da Hern Tom Ke Kyle
1021	Good Girls	Comedy, Crime, Drama	Turner has Beth in checkmate; the women must c...	42.0	8.5	582.0	Director:Michael Weaver	Chris Hendri Retta, I Whitn Rer

In [ ]:

```python
# Finding the movie with the highest fan rating
highest_fan_rating = Netflix_df[Netflix_df["Fan Rating"] == Netflix_df["Fan Rating"
print("The movie with the highest fan rating is:")
highest_fan_rating.head(1)
```

```
The movie with the highest fan rating is:
```

Out[ ]:

	Movies	Genre	Description	Run Time	Fan Rating	Votes	Director	Actors & Actresses	Rele Y
1271	Julie and the Phantoms	Comedy, Drama, Family	Julie meets Luke's parents and shares a touchi...	24.0	9.4	489.0	Director:Kabir Akhtar	Madison Reyes, Charlie Gillespie, Owen Joyner,...	202

In [ ]:

```python
# Finding the movie with the lowest fan rating
lowest_fan_rating = Netflix_df[Netflix_df["Fan Rating"] == Netflix_df["Fan Rating"]
print("The movie with the lowest fan rating is:")
lowest_fan_rating
```

```
The movie with the lowest fan rating is:
```

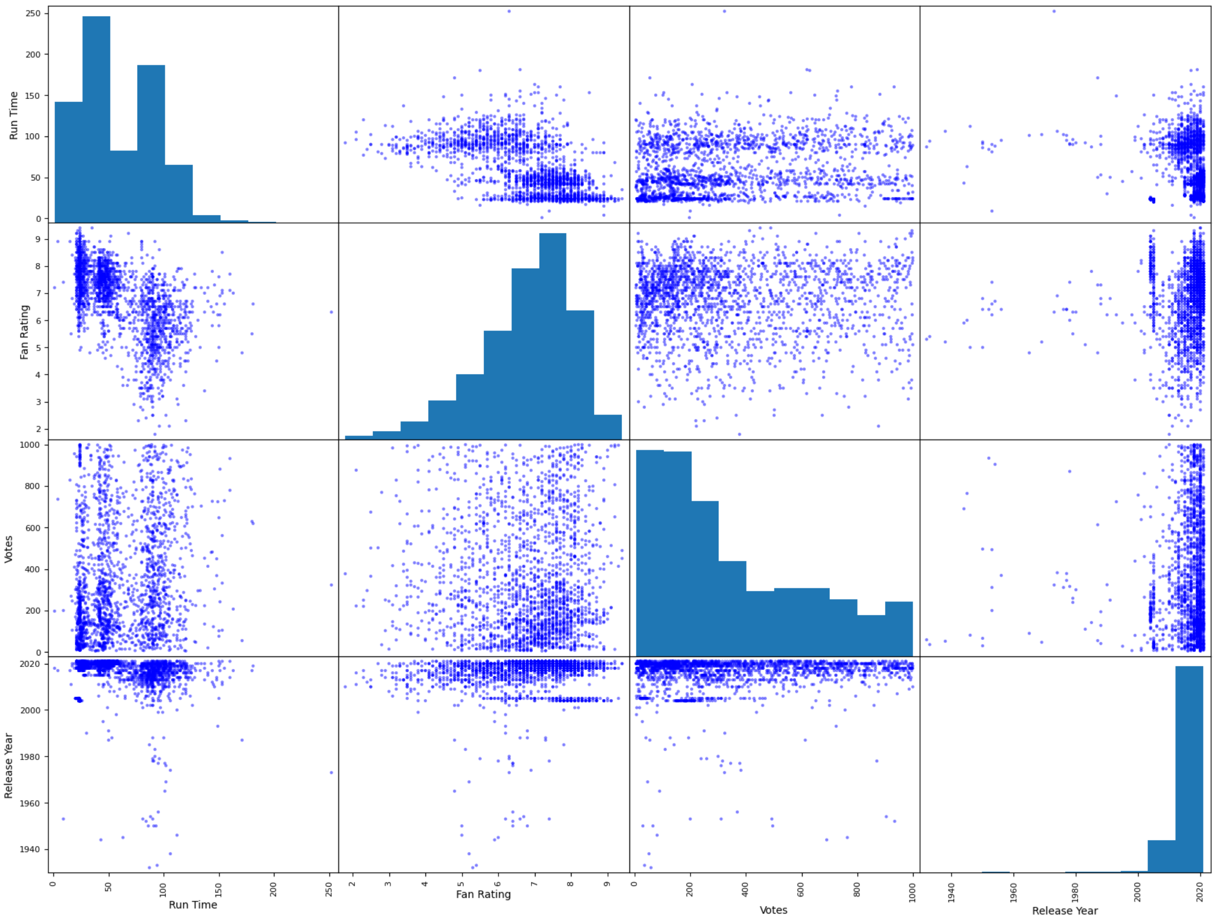
Out[ ]:

	Movies	Genre	Description	Run Time	Fan Rating	Votes	Director	Actors & Actresses	Release Year
609	Defcon 2012	Sci-Fi	On October 30, 2009 an independent filmmaker a...	92.0	1.8	377.0	Director:R. Christian Anderson	Shy Pilgreen, Dan Gruenberg, Brian Neil Hoff, ...	2010.0

Data Visualization

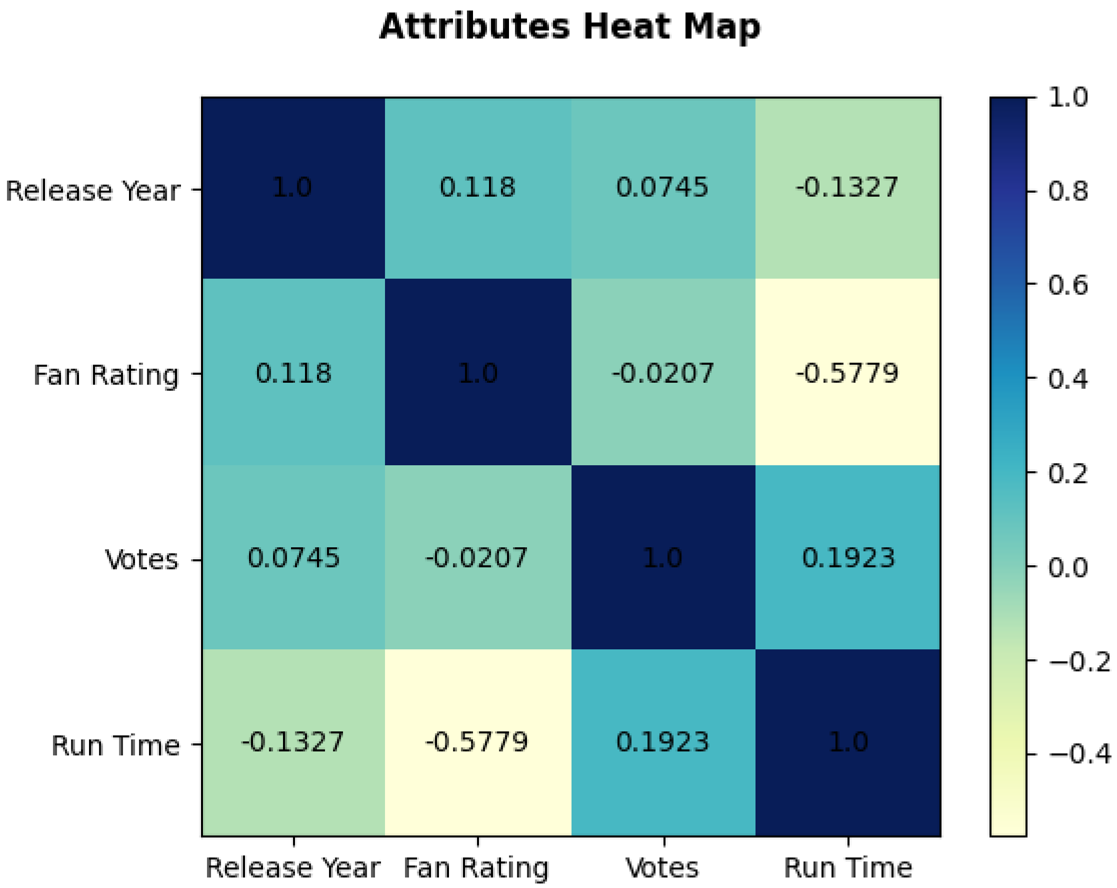
In [ ]:

```python
# Plotting the scatter matrix of the numerical values in the dataframe
pd.plotting.scatter_matrix(Netflix_df, figsize=(20, 15), color="blue")
plt.show()
```



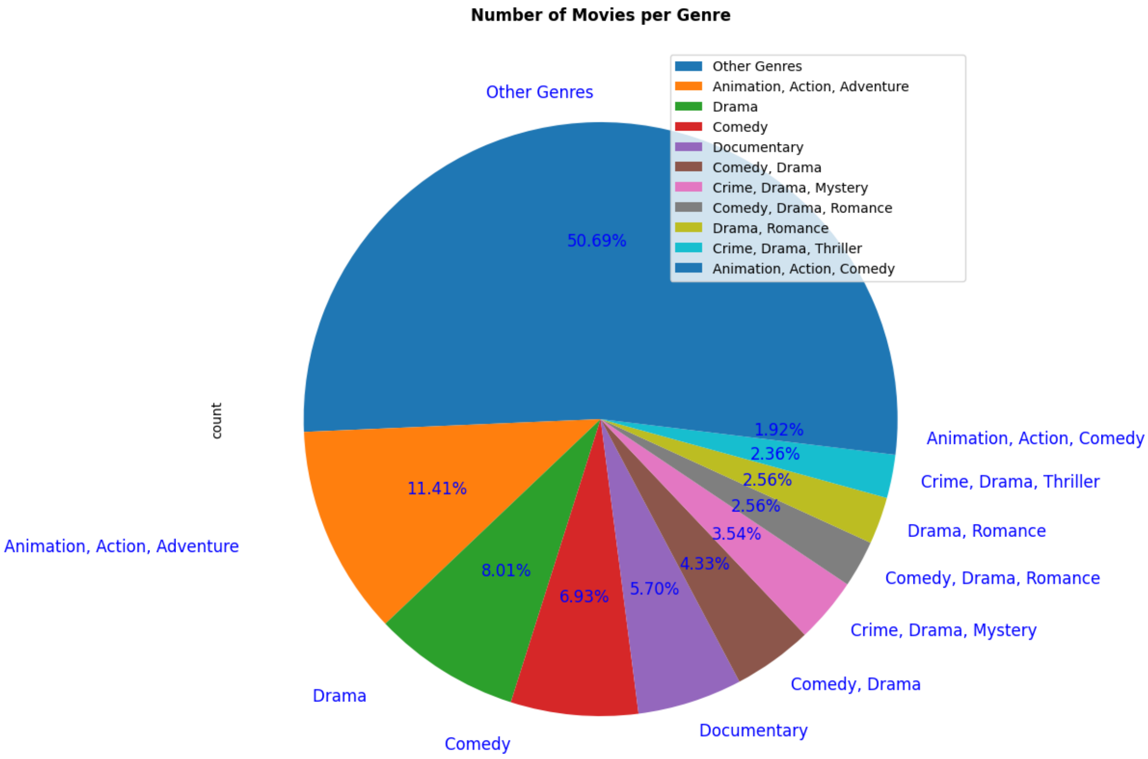
In [ ]:

```
# Obtaining the correlation matrix from the numericaL data in the dataframe
corr_mat = Netflix_df[["Release Year", "Fan Rating", "Votes", "Run Time"]].corr()

# Plotting the correlation matrix as a heat map
plt.imshow(corr_mat, cmap="YlGnBu", interpolation="nearest")

# Enabling the color bar
plt.colorbar()

# Disabling the grid lines
plt.grid(False)

# Annotating each square
for i in range(4):
    for j in range(4):
        plt.annotate(
            str(round(corr_mat.values[i][j], 4)),
            xy=(j, i),
            ha="center",
            va="center",
            color="black",
        )

# Setting the plot title
plt.title("\nAttributes Heat Map\n", weight="bold")

# Setting the tick frequency to fit the matrix
plt.xticks(range(len(corr_mat.columns)), corr_mat.columns)
plt.yticks(range(len(corr_mat.columns)), corr_mat.columns)

plt.show()
```



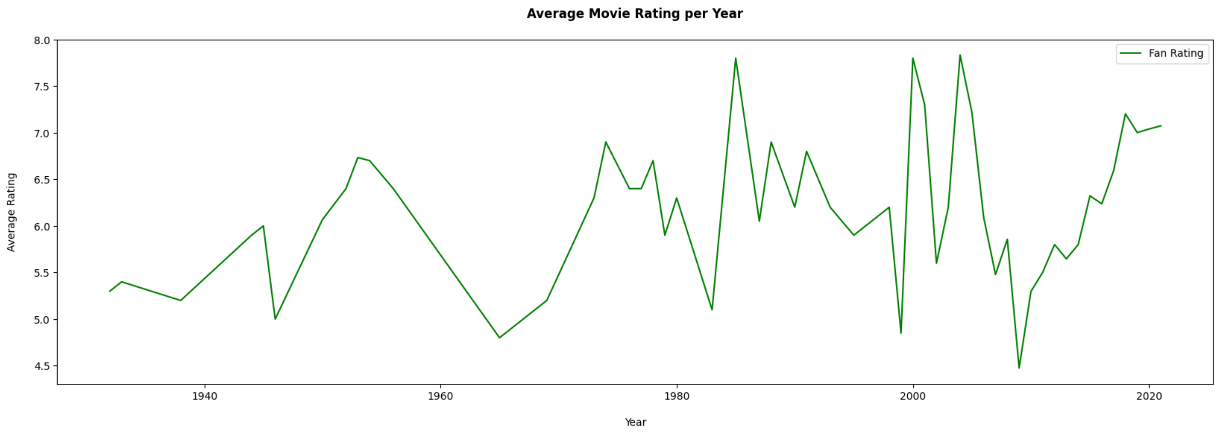
In [ ]:

```
# Plotting the histogram of the "Fan Rating" attribute

# Get the top 10 used genres
top_genres = Netflix_df["Genre"].value_counts().nlargest(10)

# Replace the other genres with "others"
genre_counts = (
    Netflix_df["Genre"].apply(lambda x: x if x in top_genres else "Other Genres").v
)

# Plotting the pie chart
genre_counts.plot.pie(
    figsize=(10, 10),
    autopct="%.2f%%",
    fontsize=12,
    textprops={"color": "blue"},  # !! Changed the color to black
)

# Setting the plot title
plt.title("\nNumber of Movies per Genre\n", weight="bold")

# Adding genre names to the legend
plt.legend(labels=genre_counts.index)
plt.show()
```



In [ ]:

```
# Plotting the histogram of the "Fan Rating" attribute

# Plot a time series of the rating of the movies over the years
Netflix_df.groupby("Release Year")["Fan Rating"].mean().plot(figsize=(20, 6), color

# Setting the plot title
plt.title("\nAverage Movie Rating per Year\n", weight="bold")

# Setting the x-axis label
plt.xlabel("\nYear")

# Setting the y-axis label
plt.ylabel("Average Rating\n")

plt.legend()
plt.show()
```


Average Movie Rating per Year
Fan Rating
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
Average Rating
1940
1960
1980
2000
2020
Year