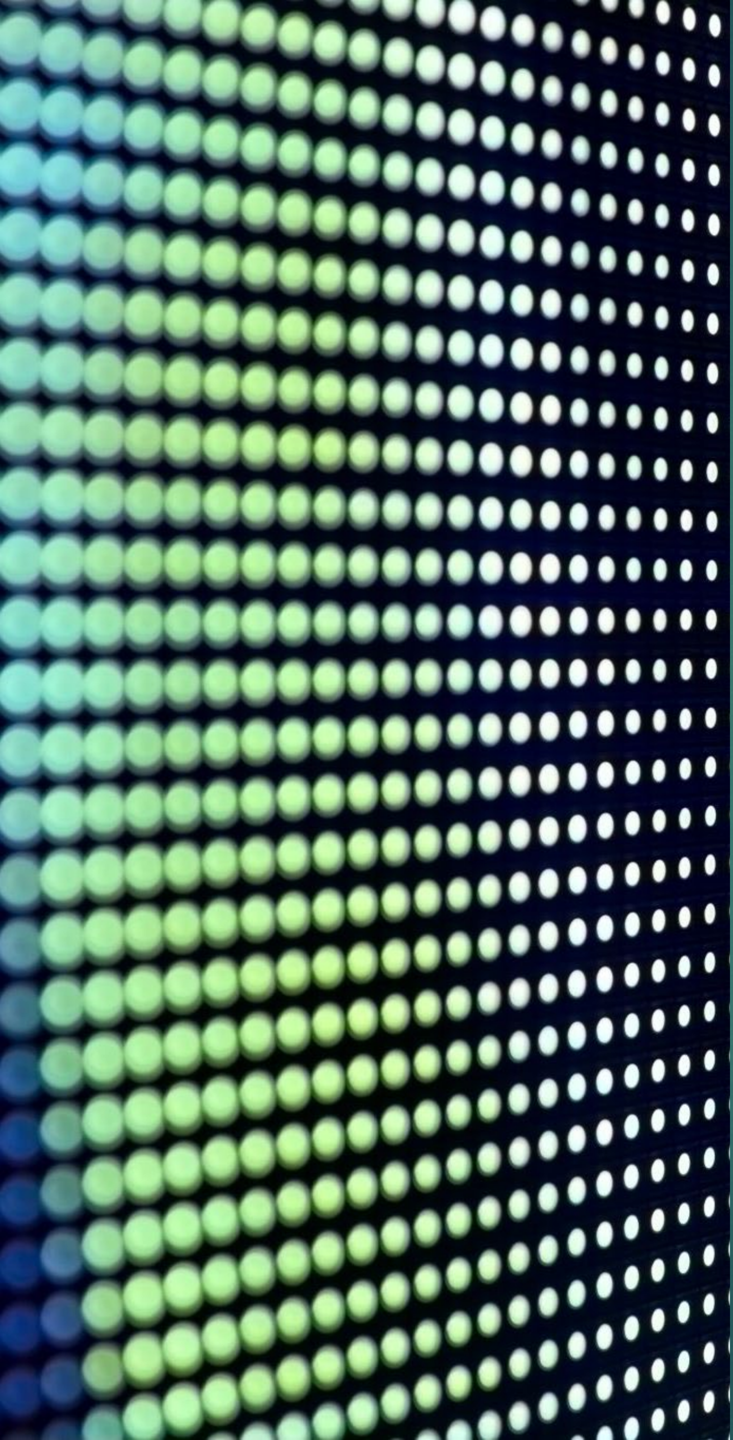


Data Cleaning Process for IMDB Movie Dataset

IMPROVING DATA QUALITY FOR BETTER ANALYSIS

CHIMEZIE NNABUIHE

Introduction

- This presentation covers the steps taken to clean the IMDB Movie Dataset
- (https://github.com/LearnDataSci/articles/blob/master/Python%20Pandas%20Tutorial%20A%20Complete%20Introduction%20for%20Beginners/IMDB-Movie-Data.csv)
- Data cleaning is a crucial step in data analysis to ensure accuracy and reliability.
  The key steps performed include handling missing values, data type conversion, removing duplicates, trimming whitespace, and splitting genres.

Initial Data Overview

- The IMDB Movie Dataset contains information about movies, including title, genre, description, director, actors, year, runtime, rating, votes, revenue, and Metascore.
- Initial inspection revealed missing values, inconsistent data types, and other issues that needed cleaning.

Step 1: Handling Missing Values

- Missing values were found in the 'Revenue (Millions)' and 'Metascore' columns.

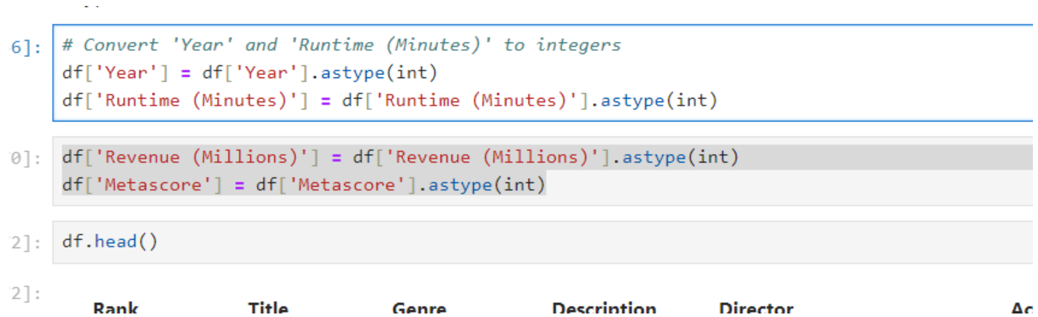
- To handle these, missing values were filled with the mean value of their respective columns, and converted to integers.

Code Snippet - Handling Missing Values

```
# Step 1: Handling Missing Values
revenue_mean = int(df['Revenue (Millions)'].mean())
metascore_mean = int(df['Metascore'].mean())

df['Revenue (Millions)'].fillna(revenue_mean, inplace=True)
df['Metascore'].fillna(metascore_mean, inplace=True)

df['Revenue (Millions)'] = df['Revenue (Millions)'].astype(int)
df['Metascore'] = df['Metascore'].astype(int)
```

Step 2: Data Type Conversion

- Certain columns had incorrect data types. Specifically, 'Year' and 'Runtime (Minutes)' were converted to integers, while 'Revenue (Millions)' and 'Metascore' were also converted to integers.



```
6]: # Convert 'Year' and 'Runtime (Minutes)' to integers
    df['Year'] = df['Year'].astype(int)
    df['Runtime (Minutes)'] = df['Runtime (Minutes)'].astype(int)
```

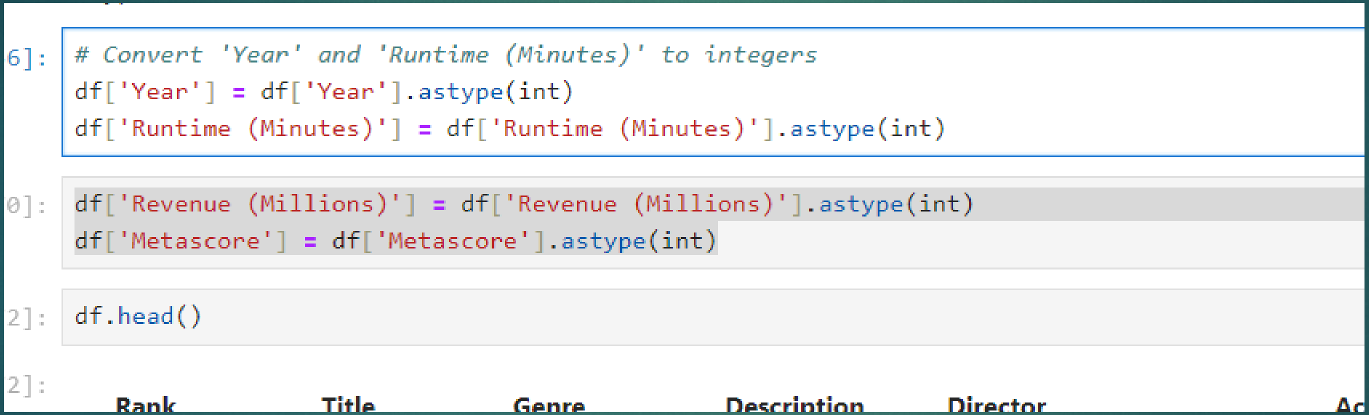
```
0]: df['Revenue (Millions)'] = df['Revenue (Millions)'].astype(int)
    df['Metascore'] = df['Metascore'].astype(int)
```

```
2]: df.head()
```

```
2]:
      Rank      Title      Genre      Description      Director      Ac
```

Code Snippet - Data Type Conversion

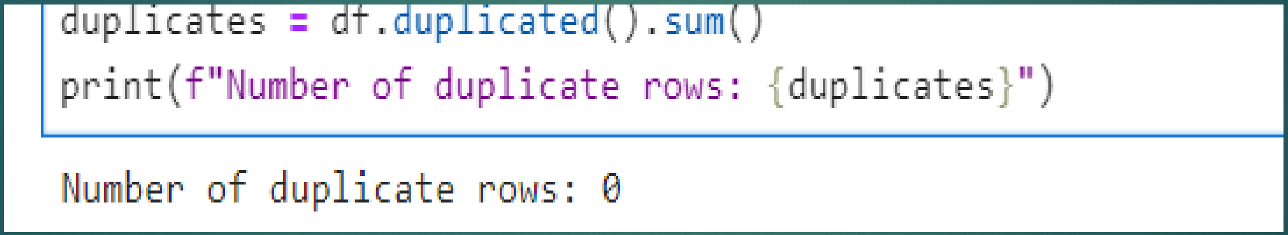
```
# Step 2: Data Type Conversion
df['Year'] = df['Year'].astype(int)
df['Runtime (Minutes)'] = df['Runtime (Minutes)'].astype(int)
```



```
6]: # Convert 'Year' and 'Runtime (Minutes)' to integers
    df['Year'] = df['Year'].astype(int)
    df['Runtime (Minutes)'] = df['Runtime (Minutes)'].astype(int)

0]: df['Revenue (Millions)'] = df['Revenue (Millions)'].astype(int)
    df['Metascore'] = df['Metascore'].astype(int)

2]: df.head()

2]:
        Rank        Title        Genre        Description        Director        Ac
```

Step 3: Removing Duplicates

- Duplicate rows can skew analysis results, so it's important to remove them. I checked for duplicates but none was found



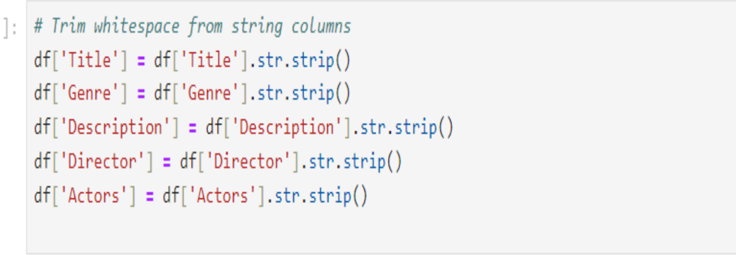
```
duplicates = df.duplicated().sum()
print(f"Number of duplicate rows: {duplicates}")
```

```
Number of duplicate rows: 0
```

Step 4: Trimming Whitespace

- Whitespace in string fields can cause issues during analysis. I trimmed leading and trailing whitespace from relevant columns.
- Code Snippet

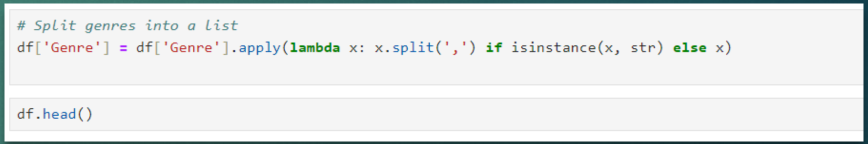
```
# Step 4: Trimming Whitespace
df['Title'] = df['Title'].str.strip()
df['Genre'] = df['Genre'].str.strip()
df['Description'] = df['Description'].str.strip()
df['Director'] = df['Director'].str.strip()
df['Actors'] = df['Actors'].str.strip()
```



```
]:  # Trim whitespace from string columns
    df['Title'] = df['Title'].str.strip()
    df['Genre'] = df['Genre'].str.strip()
    df['Description'] = df['Description'].str.strip()
    df['Director'] = df['Director'].str.strip()
    df['Actors'] = df['Actors'].str.strip()
```

Step 5: Splitting Genres

- The 'Genre' column contained multiple genres in a single string. We split this column into a list of genres for each movie.
- Splitting Genres df['Genre'] = df['Genre'].apply(lambda x: x.split(',') if isinstance(x, str) else x)



```
# Split genres into a list
df['Genre'] = df['Genre'].apply(lambda x: x.split(',') if isinstance(x, str) else x)
```

```
df.head()
```

Final Data Overview

- After cleaning, the dataset is now free of missing values, correct data types, no duplicates, trimmed whitespace, and split genres.
  The cleaned dataset is now ready for analysis.

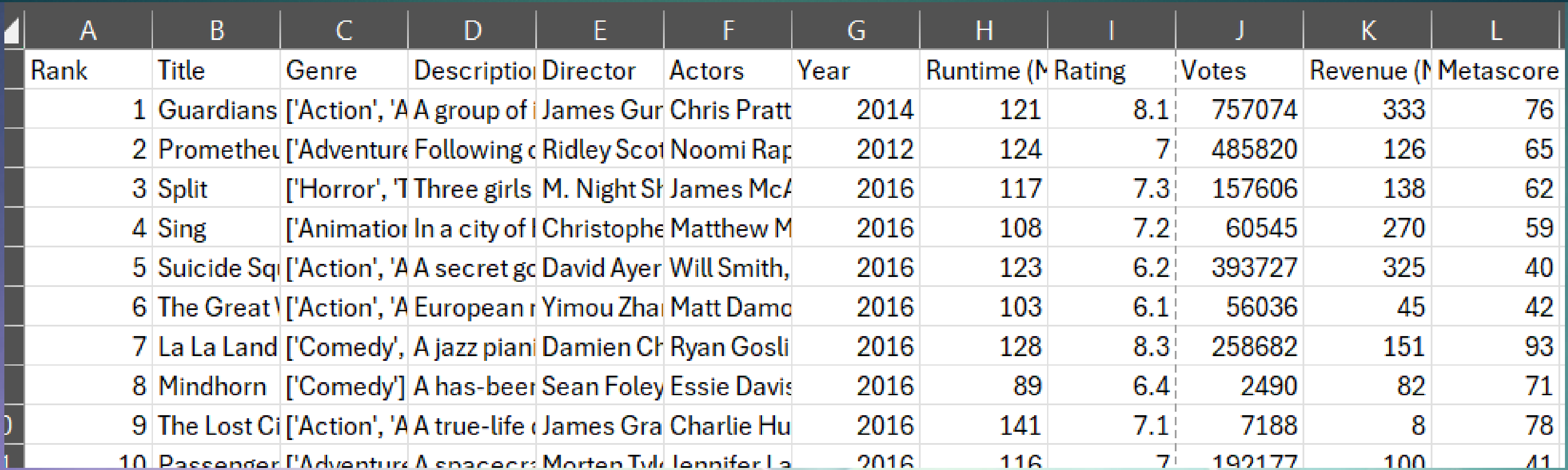
- Below is the Notebook used for the data cleaning.

Untitled3.ipynb

Conclusion

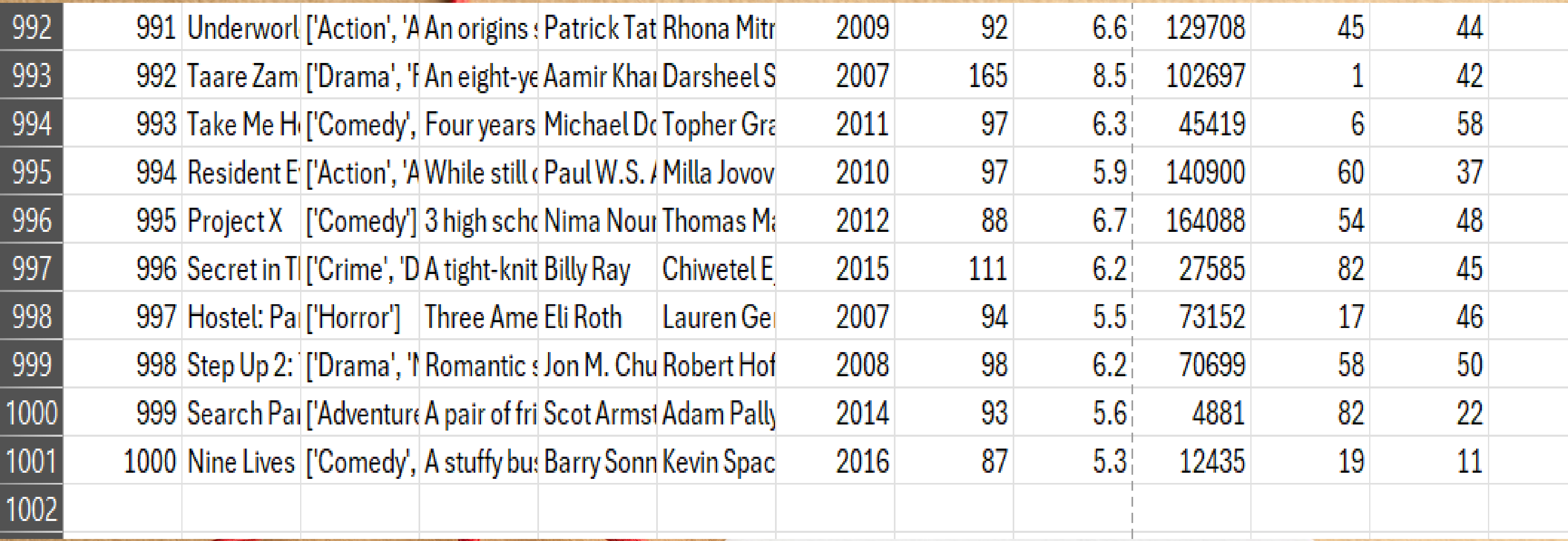
- The data cleaning process involved handling missing values, converting data types, removing duplicates, trimming whitespace, and splitting genres.
  These steps are crucial for ensuring data quality and reliability in any analysis.
  Thank you!
  Chimezie Nnabuihe

Cleaned IMDB Data Sample(Head(10))



A	B	C	D	E	F	G	H	I	J	K	L
Rank	Title	Genre	Descriptio	Director	Actors	Year	Runtime (M	Rating	Votes	Revenue (M	Metascore
1	Guardians	['Action', 'A	A group of i	James Gur	Chris Pratt	2014	121	8.1	757074	333	76
2	Prometheu	['Adventure	Following c	Ridley Scot	Noomi Rap	2012	124	7	485820	126	65
3	Split	['Horror', 'T	Three girls	M. Night Sh	James McA	2016	117	7.3	157606	138	62
4	Sing	['Animation	In a city of	Christophe	Matthew M	2016	108	7.2	60545	270	59
5	Suicide Sq	['Action', 'A	A secret go	David Ayer	Will Smith,	2016	123	6.2	393727	325	40
6	The Great \	['Action', 'A	European r	Yimou Zha	Matt Damo	2016	103	6.1	56036	45	42
7	La La Land	['Comedy',	A jazz piani	Damien Ch	Ryan Gosli	2016	128	8.3	258682	151	93
8	Mindhorn	['Comedy']	A has-beer	Sean Foley	Essie Davis	2016	89	6.4	2490	82	71
9	The Lost Ci	['Action', 'A	A true-life c	James Gra	Charlie Hu	2016	141	7.1	7188	8	78
10	Passenger	['Adventure	A spacecra	Morten Tyl	Jennifer La	2016	116	7	192177	100	41

Cleaned IMDB Data Sample(Tail(10))



992	991	Underworl	['Action', 'A	An origins s	Patrick Tat	Rhona Mitr	2009	92	6.6	129708	45	44
993	992	Taare Zam	['Drama', 'F	An eight-ye	Aamir Khai	Darsheel S	2007	165	8.5	102697	1	42
994	993	Take Me H	['Comedy',	Four years	Michael Do	Topher Gra	2011	97	6.3	45419	6	58
995	994	Resident E	['Action', 'A	While still c	Paul W.S. A	Milla Jovov	2010	97	5.9	140900	60	37
996	995	Project X	['Comedy']	3 high schc	Nima Nour	Thomas Ma	2012	88	6.7	164088	54	48
997	996	Secret in Tl	['Crime', 'D	A tight-knit	Billy Ray	Chiwetel E	2015	111	6.2	27585	82	45
998	997	Hostel: Pa	['Horror']	Three Ame	Eli Roth	Lauren Ge	2007	94	5.5	73152	17	46
999	998	Step Up 2:	['Drama', 'I	Romantic s	Jon M. Chu	Robert Hof	2008	98	6.2	70699	58	50
1000	999	Search Pai	['Adventure	A pair of fri	Scot Armst	Adam Pally	2014	93	5.6	4881	82	22
1001	1000	Nine Lives	['Comedy',	A stuffy bu	Barry Sonn	Kevin Spac	2016	87	5.3	12435	19	11
1002												