A

Course End Project Report on

Cinematic Insights

Is submitted in partial fulfillment of the Requirements for the Award of CIE of

DATA ANALYSIS AND VISUALIZATION-22ADE01

in

B.E, IV-SEM, INFORMATION TECHNOLOGY

Submitted by

G.Sai Varshini – 160122737006

T.Nikhitha – 160122737026

V.Akshitha - 160122737027

COURSE TAUGHT BY:

Dr Ramakrishna Kolikipogu Professor, Dept of IT.

DEPARTMENT OF INFORMATION TECHNOLOGY CHAITANYA

BHARATHI INSTITUTE OF TECHNOLOGY(A)

(Affiliated to OSmania University;Accredited by NBA,NAAC,ISO)

kokapet(V),GANDIPET(M),HYDERABAD-500075

Website:www.cbit.ac.in

2023-2024

CERTIFICATE

This is to certify that the course end project work entitled **”Your Project Title”** is submitted by **G .Sai Varshini (160122737006), T. Nikhitha (160122737026), and V.Akshitha (160122737027)** in partial fulfillment of the requirements for the award of CIE Marks of **DATA ANALYSIS AND VISUALIZATION (22ADE01)** of **B.E, IV-SEM, INFORMATION TECHNOLOGY** to CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY(A) affiliated to OSMANIA UNIVERSITY, Hyderabad is a record of bonafide work carried out by them under my supervision and guidance. The results embodied in this report have not been submitted to any other University or Institute for the award of any other Degree or Diploma.

Signature of Course Faculty
Dr Ramakrishna Kolikipogu
Professor of IT

Acknowledgement

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

We wish to express our deep sense of gratitude to **Dr Ramakrishna Kolikipogu, Professor of IT** for his able guidance and useful suggestions, which helped us in completing the Course End Project in time.

We are particularly thankful to **HoD, Principal and Management**, for their support and encouragement, which helped us to mould our project into a successful one.

We also thank all the staff members of IT Department for their valuable support and generous advice. Finally thanks to all our friends and family members for their continuous support and enthusiastic help.

G.Sai Varshini – 160122737006

T.Nikhitha – 160122737026

V.Akshitha - 160122737027

Abstract

"Cinematic Insights" presents a detailed exploration and visualization of the top 250 films according to IMDb. We have meticulously gathered data covering rankings, release years, genres, ratings, and box office earnings of these cinematic gems. Armed with this wealth of information, we're utilizing user-friendly tools like Pandas, NumPy, and Matplotlib to analyze and visualize the data in an engaging and insightful manner.

Our exploration takes us deep into the enchanting realm of cinema, where we uncover the secrets that make these movies truly exceptional. From unraveling the intricate relationships between ratings and box office success to discovering the audience's favorite genres, we're immersing ourselves in the magic of moviemaking.

Through vibrant charts, graphs, easy-to-understand analyses and compelling visualizations, this study investigates various aspects of the cinematic landscape. Whether you're a seasoned movie buff or simply curious about what makes a film great, join us as we embark on this thrilling adventure to unravel the mysteries behind the silver screen!

Keywords: Pandas, NumPy, Matplotlib, Visualization, Analyzation, Illustrated Charts, Realm of Cinema

Table of Contents

List of Figures

CHAPTER 1
Introduction

In the expansive universe of cinema, where storytelling transcends boundaries and captivates audiences, "Cinematic Insights" stands as a pioneering endeavor—a meticulous exploration and visualization of the IMDb Top 250 Movies. These films, revered for their cinematic prowess and cultural significance, represent the pinnacle of cinematic achievement, each holding a unique place in the hearts of viewers worldwide.

Driven by an insatiable curiosity and armed with a wealth of data covering rankings, release years, genres, ratings, and box office earnings, "Cinematic Insights" embarks on a journey to uncover the secrets behind these timeless cinematic treasures. Through the lens of data analysis and visualization, this study endeavors to shed light on the intricate dynamics that define the essence of these movies.

The foundation of "Cinematic Insights" rests upon a series of meticulously performed operations, each aimed at unraveling the mysteries of the silver screen:

1. Data Loading and Inspection:
- The journey begins with the loading and thorough inspection of the IMDb Top 250 Movies dataset, allowing for a comprehensive understanding of its structure and contents.

2. **Data Cleaning**:
- To ensure the integrity and reliability of the analysis, rigorous data cleaning operations are performed, including the identification and handling of missing values and duplicate entries.

3. **Data Exploration**:
- Delving into the depths of the dataset, "Cinematic Insights" conducts a detailed exploration, unveiling insights into the top-rated movies, their genres, and the temporal evolution of cinematic excellence.

4. **Data Visualization**:
- Armed with powerful visualization tools like Matplotlib and seaborn, "Cinematic Insights" brings the data to life, offering captivating visual representations of key trends and patterns within the dataset.

5. **Additional Operations**:
- Beyond exploration and visualization, "Cinematic Insights" conducts additional operations, including the categorization of movies based on their ratings and the identification of the most prevalent genres within the IMDb Top 250.

Through vibrant charts, compelling graphs, and insightful analyses, "Cinematic Insights" invites viewers to embark on a captivating journey into the heart of cinema. Whether you're a seasoned movie buff or an intrigued novice, this study promises to captivate your imagination and deepen your appreciation for the art of filmmaking. Join us as we unravel the mysteries behind the silver screen and celebrate the enduring magic of cinema through the lens of data analysis and visualization. Welcome to "Cinematic Insights," where every data point tells a story. and every visualization unveils a glimpse of cinematic brilliance.

CHAPTER 2
Methodology

In this section, we outline the methodologies employed in our research to achieve the objectives outlined in the preceding sections. Our approach encompasses a combination of quantitative and qualitative methods, each tailored to address specific aspects of the research inquiry. We adopt a multi-faceted methodology to comprehensively explore the research problem, gather relevant data, and derive meaningful insights. The methodologies utilized in this study include:

1. **Data collection and DATASET description**
2. **Data Analysis Methodology Overview**

2.1 Data collection and Dataset description

The dataset utilized in this study was sourced from the popular data science platform **Kaggle**, under the title "IMDB Top 250 Movies." Kaggle provides a diverse range of datasets for research and analysis purposes, and the "IMDB Top 250 Movies" dataset serves as a valuable resource for exploring and understanding trends in the film industry.

Dataset Description:

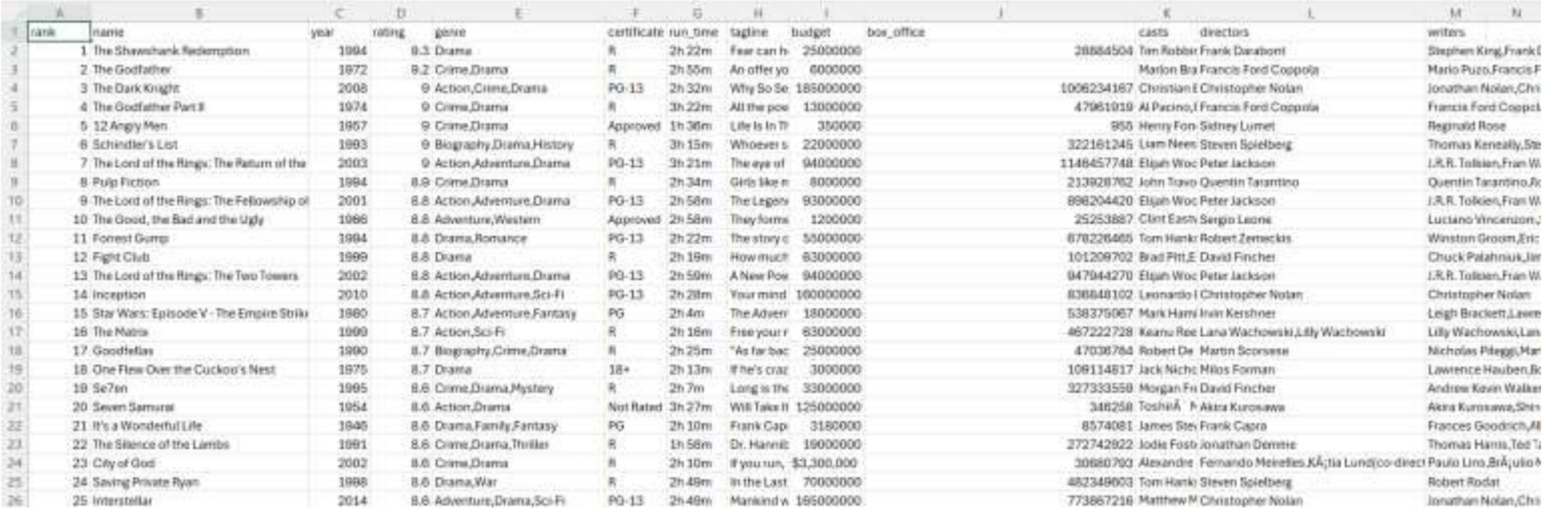
The dataset comprises various attributes related to the top-rated movies on IMDb, a renowned online database of film-related information.

Key attributes included in the dataset are as follows:

- **Rank**: The ranking of the movie within the IMDb Top 250 list.
- **Name**: The title of the movie.
- **Year**: The year of the movie's release.

- **Rating**: The IMDb rating of the movie.
- **Genre**: The genre(s) to which the movie belongs.
- **Certificate**: The age certification or content rating of the movie.
- **Runtime**: The duration of the movie in minutes.
- **Tagline**: A brief and catchy phrase associated with the movie.
- **Budget**: The estimated production budget of the movie.
- **Box Office**: The worldwide box office earnings of the movie.
- **Casts**: The main cast members of the movie.
- **Directors**: The director(s) of the movie.
- **Writers**: The writer(s) of the movie.

Figure 2.1



rank	name	year	rating	genre	certificate	run_time	tagline	budget	box_office	casts	directors	writers
1	The Shawshank Redemption	1994	9.3	Drama	R	2h 22m	Fear can h	25000000	28884504	Tim Robbir	Frank Darabont	Stephen King,Frank D
2	The Godfather	1972	9.2	Crime,Drama	R	2h 55m	An offer yo	6000000		Marlon Bra	Francis Ford Coppola	Mario Puzo,Francis F
3	The Dark Knight	2008	9	Action,Crime,Drama	PG-13	2h 32m	Why So Se	185000000	1006234167	Christian E	Christopher Nolan	Jonathan Nolan,Chri
4	The Godfather Part II	1974	9	Crime,Drama	R	3h 22m	All the pow	13000000	47961919	Al Pacino,I	Francis Ford Coppola	Francis Ford Coppol
5	12 Angry Men	1957	9	Crime,Drama	Approved	1h 36m	Life Is In Th	350000	955	Henry Fon	Sidney Lumet	Reginald Rose
6	Schindler's List	1993	9	Biography,Drama,History	R	3h 15m	Whoever s	22000000	322161245	Liam Neen	Steven Spielberg	Thomas Keneally,Ste
7	The Lord of the Rings: The Return of the	2003	9	Action,Adventure,Drama	PG-13	3h 21m	The eye of	94000000	1146457748	Elijah Woc	Peter Jackson	J.R.R. Tolkien,Fran W
8	Pulp Fiction	1994	8.9	Crime,Drama	R	2h 34m	Girls like m	8000000	213928762	John Travo	Quentin Tarantino	Quentin Tarantino,Ro
9	The Lord of the Rings: The Fellowship of	2001	8.8	Action,Adventure,Drama	PG-13	2h 58m	The Legen	93000000	898204420	Elijah Woc	Peter Jackson	J.R.R. Tolkien,Fran W
10	The Good, the Bad and the Ugly	1966	8.8	Adventure,Western	Approved	2h 58m	They forme	1200000	25253887	Clint Eastv	Sergio Leone	Luciano Vincenzoni,
11	Forrest Gump	1994	8.8	Drama,Romance	PG-13	2h 22m	The story c	55000000	678226465	Tom Hank:	Robert Zemeckis	Winston Groom,Eric
12	Fight Club	1999	8.8	Drama	R	2h 19m	How much	63000000	101209702	Brad Pitt,E	David Fincher	Chuck Palahniuk,Jim
13	The Lord of the Rings: The Two Towers	2002	8.8	Action,Adventure,Drama	PG-13	2h 59m	A New Pow	94000000	947944270	Elijah Woc	Peter Jackson	J.R.R. Tolkien,Fran W
14	Inception	2010	8.8	Action,Adventure,Sci-Fi	PG-13	2h 28m	Your mind	160000000	836848102	Leonardo I	Christopher Nolan	Christopher Nolan
15	Star Wars: Episode V - The Empire Strik	1980	8.7	Action,Adventure,Fantasy	PG	2h 4m	The Adven	18000000	538375067	Mark Hami	Irvin Kershner	Leigh Brackett,Lawre
16	The Matrix	1999	8.7	Action,Sci-Fi	R	2h 16m	Free your r	63000000	467222728	Keanu Ree	Lana Wachowski,Lilly Wachowski	Lilly Wachowski,Lan
17	Goodfellas	1990	8.7	Biography,Crime,Drama	R	2h 25m	"As far bac	25000000	47036784	Robert De	Martin Scorsese	Nicholas Pileggi,Mar
18	One Flew Over the Cuckoo's Nest	1975	8.7	Drama	18+	2h 13m	If he's craz	3000000	109114817	Jack Nichc	Milos Forman	Lawrence Hauben,Bo
19	Se7en	1995	8.6	Crime,Drama,Mystery	R	2h 7m	Long is the	33000000	327333559	Morgan Fr	David Fincher	Andrew Kevin Walker
20	Seven Samurai	1954	8.6	Action,Drama	Not Rated	3h 27m	Will Take It	125000000	346258	ToshiÅ M	Akira Kurosawa	Akira Kurosawa,Shin
21	It's a Wonderful Life	1946	8.6	Drama,Family,Fantasy	PG	2h 10m	Frank Cap	3180000	8574081	James Stev	Frank Capra	Frances Goodrich,Al
22	The Silence of the Lambs	1991	8.6	Crime,Drama,Thriller	R	1h 58m	Dr. Hannib	19000000	272742922	Jodie Fost	Jonathan Demme	Thomas Harris,Ted Ta
23	City of God	2002	8.6	Crime,Drama	R	2h 10m	If you run,	$3,300,000	30680793	Alexandre	Fernando Meirelles,KÃ¡tia Lund(co-direct	Paulo Lins,BrÃ¡ulio M
24	Saving Private Ryan	1998	8.6	Drama,War	R	2h 49m	In the Last	70000000	482349603	Tom Hank:	Steven Spielberg	Robert Rodat
25	Interstellar	2014	8.6	Adventure,Drama,Sci-Fi	PG-13	2h 49m	Mankind w	165000000	773867216	Matthew M	Christopher Nolan	Jonathan Nolan,Chri

Data collection from Kaggle ensures access to a curated and reliable dataset that aligns with the objectives of this study. The dataset's comprehensive nature, encompassing various aspects such as movie rankings, ratings, genres, and production details, facilitates a thorough analysis of the cinematic landscape represented within the IMDb Top 250 list.

By leveraging this dataset, the study aims to uncover insights into the factors contributing to the success and acclaim of top-rated movies, explore trends in cinematic preferences and genres, and delve into the dynamics of audience engagement and reception within the realm of cinema.

2.2 Data Analysis Methodology Overview

1. **Data Loading and Inspection**:

- Loaded the dataset from the CSV file using Pandas' `read_csv()` function.
- Inspected the first and last 10 rows of the dataset using the `head()` and `tail()` functions, respectively.
- Checked the shape of the dataset using the `shape` attribute to determine the number of rows and columns.
- Examined the information about the dataset, including data types and non-null values, using the `info()` function.

2. **Data Cleaning**:

- Checked for missing values in the dataset using the `isnull()` function to identify rows with null values.
- Dropped rows with missing data using the `dropna()` function to ensure data integrity.
- Checked for duplicate rows using the `duplicated()` function to identify and remove any duplicated entries.

3. **Data Exploration**:

- Calculated descriptive statistics of numerical variables in the dataset using the `describe()` function to obtain summary statistics such as mean, median, minimum, maximum, and quartiles.
- Extracted and analyzed the top 10 movies with the highest ratings using the `nlargest()` function to identify the highest-rated movies.
- Categorized movies based on their ratings into "Excellent," "Good," and "Average" categories using custom-defined criteria.
- Explored the number of movies in specific genres such as Action, Drama, Crime, Thriller, War, Romance, Adventure, Comedy, and Horror by filtering the dataset based on genre and counting the occurrences.

4. **Data Visualization**:

- Created histograms using Matplotlib or Seaborn to visualize the distributions of movie ratings and release years, providing insights into the spread and central tendencies of these variables.

- Plotted a bar chart to show the top 10 movie genres based on the number of movies, facilitating a visual comparison of genre popularity.
- Generated a scatter plot to explore the relationship between movie ratings and box office earnings, allowing for an analysis of potential correlations or trends between these variables.
- Created a pie chart to visualize the distribution of movie genres, providing a visual representation of genre composition within the dataset.
- Plotted a line graph to show the average movie rating over the years, highlighting trends or patterns in rating fluctuations over time.

5. **Additional Operations**:

- Calculated the mean, median, and standard deviation of movie ratings using descriptive statistics functions. Counted the number of movies in specific genres using both visualizations, such as bar charts, and print statements to provide numerical counts.

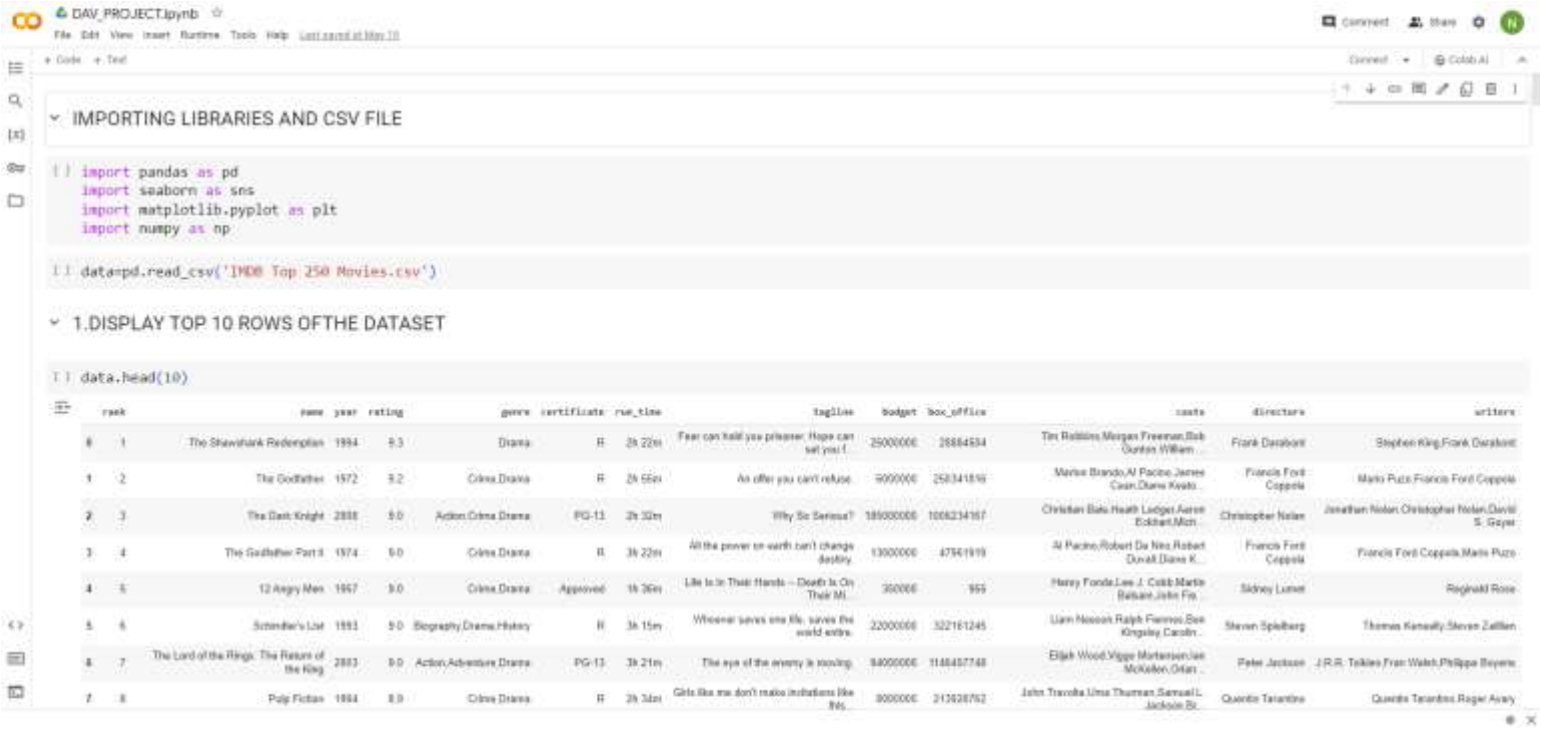
CHAPTER 3

System Architecture and Mathematical Analysis

3.1 Google Colab

Google Colaboratory, commonly known as Google Colab, is a free online cloud-based Jupyter notebook environment tailored for training machine learning and deep learning models. This article explores the functionalities, benefits, and features of Google Colab, elucidating its significance in the realm of data science and machine learning.

Figure 3.1: Google colab environment


DAV_PROJECT.ipynb
File Edit View Insert Runtime Tools Help Last saved at May 10
+ Code + Text
Comment
Share
Connect
Colab AI

IMPORTING LIBRARIES AND CSV FILE

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

```
data=pd.read_csv('IMDB Top 250 Movies.csv')
```

1.DISPLAY TOP 10 ROWS OFTHE DATASET

```
data.head(10)
```

	rank	name	year	rating	genre	certificate	run_time	tagline	budget	box_office	casts	directors	writers
0	1	The Shawshank Redemption	1994	9.3	Drama	R	2h 22m	Fear can hold you prisoner. Hope can set you f...	25000000	28884504	Tim Robbins,Morgan Freeman,Bob Gunton,William ...	Frank Darabont	Stephen King,Frank Darabont
1	2	The Godfather	1972	9.2	Crime,Drama	R	2h 55m	An offer you can't refuse.	6000000	250341816	Marlon Brando,Al Pacino,James Caan,Diane Keato...	Francis Ford Coppola	Mario Puzo,Francis Ford Coppola
2	3	The Dark Knight	2008	9.0	Action,Crime,Drama	PG-13	2h 32m	Why So Serious?	185000000	1006234167	Christian Bale,Heath Ledger,Aaron Eckhart,Mich...	Christopher Nolan	Jonathan Nolan,Christopher Nolan,David S. Goyer
3	4	The Godfather Part II	1974	9.0	Crime,Drama	R	3h 22m	All the power on earth can't change destiny.	13000000	47961919	Al Pacino,Robert De Niro,Robert Duvall,Diane K...	Francis Ford Coppola	Francis Ford Coppola,Mario Puzo
4	5	12 Angry Men	1957	9.0	Crime,Drama	Approved	1h 36m	Life Is In Their Hands -- Death Is On Their Mi...	350000	955	Henry Fonda,Lee J. Cobb,Martin Balsam,John Fie...	Sidney Lumet	Reginald Rose
5	6	Schindler's List	1993	9.0	Biography,Drama,History	R	3h 15m	Whoever saves one life, saves the world entire.	22000000	322161245	Liam Neeson,Ralph Fiennes,Ben Kingsley,Carolin...	Steven Spielberg	Thomas Keneally,Steven Zaillian
6	7	The Lord of the Rings: The Return of the King	2003	9.0	Action,Adventure,Drama	PG-13	3h 21m	The eye of the enemy is moving.	94000000	1146457748	Elijah Wood,Viggo Mortensen,Ian McKellen,Orlan...	Peter Jackson	J.R.R. Tolkien,Fran Walsh,Philippa Boyens
7	8	Pulp Fiction	1994	8.9	Crime,Drama	R	2h 34m	Girls like me don't make invitations like this...	8000000	213928762	John Travolta,Uma Thurman,Samuel L. Jackson,Br...	Quentin Tarantino	Quentin Tarantino,Roger Avary

3.1.1 What is Google Colab?

Google Colab offers a cloud-based environment accessible via any web browser, eliminating the need for local software installation. Users can leverage its computing resources, including CPUs, GPUs, and TPUs, facilitating efficient model training and execution.

3.2 Benefits of Google Colab

Accessibility: Users can access Google Colab from any location with internet connectivity, streamlining collaboration and workflow.

Power: The platform provides access to potent computing resources like GPUs and TPUs, enabling swift and effective model training.

Collaboration: Google Colab simplifies collaborative efforts by allowingreal-time editing and sharing of notebooks among team members.

Education: It serves as an invaluable educational tool for learning about machine learning and data science, offering a plethora of tutorials and resources.

3.2.1 Why Choose Google Colab?

Google Colab stands out as an ideal choice for students, data scientists, researchers, and enthusiasts due to its:

Ease of Use: With no setup requirements, users can swiftly start coding after creating an account.

Affordability: The platform is largely free to use, with paid plans available for more demanding tasks.

Flexibility: Users can seamlessly train models, process data, create visualizations, and collaborate with others, making it a versatile tool for various applications.

3.2.2 Notebook in Google Colab

In Google Colab, a notebook serves as a web-based environment for code creation and execution. Notebooks offer several advantages, including real-time code execution and visualization, support for markdown for documentation, and collaboration features, making them indispensable for data scientists and machine learning practitioners.

3.2.3 Google Colab Features

Google Colab boasts several features that enhance its usability and effectiveness:

Free Access to GPUs and TPUs: Users can leverage powerful computing resources without any additional cost.

Web-based Interface: The intuitive and user-friendly interface eliminates the need for local software installation.

Collaboration Tools: Multiple users can collaborate on the same notebook simultaneously, streamlining teamwork.

Markdown Support: Notebooks support markdown, enabling users to include formatted text, equations, and images alongside their code.

Pre-installed Libraries: Google Colab comes pre-installed with popular libraries and tools for machine learning and deep learning, such as TensorFlow and PyTorch, saving time on setup and configuration.

Google Colab emerges as a versatile and indispensable tool for machine learning and data science tasks, offering accessibility, power, and flexibility. Its user-friendly interface, collaborative features, and integration with powerful computing resources make it an invaluable asset for individuals and teams alike, driving innovation and progress in the field of machine learning and beyond.

3.3 Mathematical Analysis

A thorough analysis of the IMDb Top 250 Movies dataset was conducted, following a structured approach encompassing data loading, cleaning, exploration, visualization, and additional operations. We meticulously performed each step of the data analysis process to ensure a comprehensive understanding of the dataset and to extract valuable insights.

This involved loading the dataset from its source, meticulously cleaning it to remove any inconsistencies or missing values, exploring the data to uncover patterns and trends, visualizing key aspects to facilitate interpretation, and conducting additional operations to enhance the depth of our analysis. Through this systematic approach, we were able to gain meaningful insights into the IMDb Top 250 Movies dataset, empowering us to draw informed conclusions and make data-driven decisions

Data Loading and Inspection:

We used Pandas'read_csv()' function to load the dataset and inspected the first and last 10 rows using `head()` and `tail()` functions. The `shape` attribute was used to determine the dataset's dimensions, and the `info()` function provided insights into data types and non-null values.

Data Cleaning:

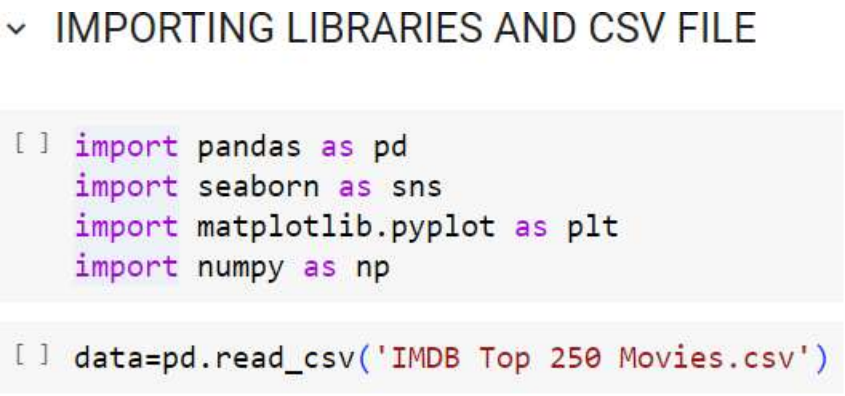
We identified missing values with the `isnull()` function and ensured data integrity by dropping rows with missing data using `dropna()`. Duplicate rows were identified and removed with the `duplicated()` function to eliminate redundancy.

Data Exploration:

Descriptive statistics were computed using the `describe()` function, revealing key metrics such as mean, median, and quartiles. We analyzed the top 10 movies with the highest ratings and categorized movies into "Excellent," "Good," and "Average" based on custom-defined criteria. Genre-specific exploration was conducted by filtering the dataset based on genre and counting occurrences.

Data Visualization:

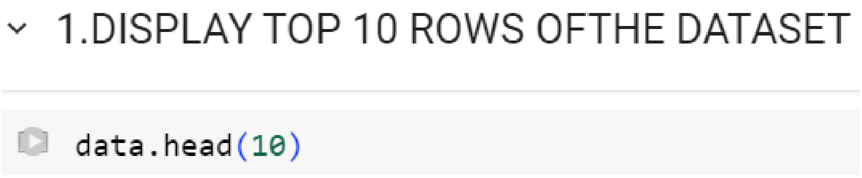
We created histograms to visualize distributions of movie ratings and release years. Bar charts displayed the top 10 movie genres based on movie counts, while scatter plots explored the relationship between ratings and box office earnings. Pie charts illustrated genre distribution, and line graphs depicted average movie ratings over time.



IMPORTING LIBRARIES AND CSV FILE

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

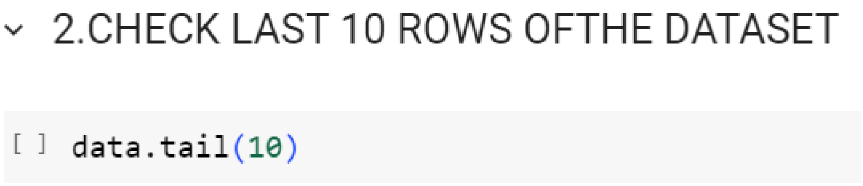
```
[ ] data=pd.read_csv('IMDB Top 250 Movies.csv')
```



1.DISPLAY TOP 10 ROWS OFTHE DATASET

```
data.head(10)
```

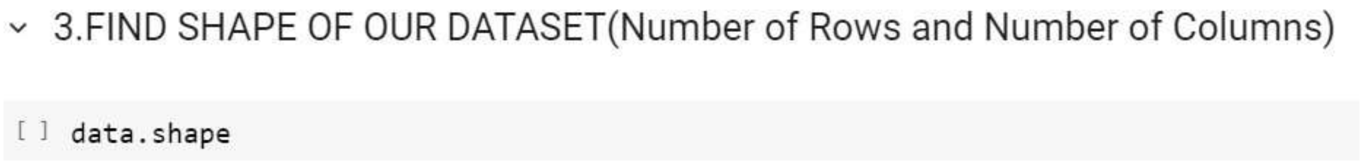
Figure 3.2



2.CHECK LAST 10 ROWS OFTHE DATASET

```
[ ] data.tail(10)
```

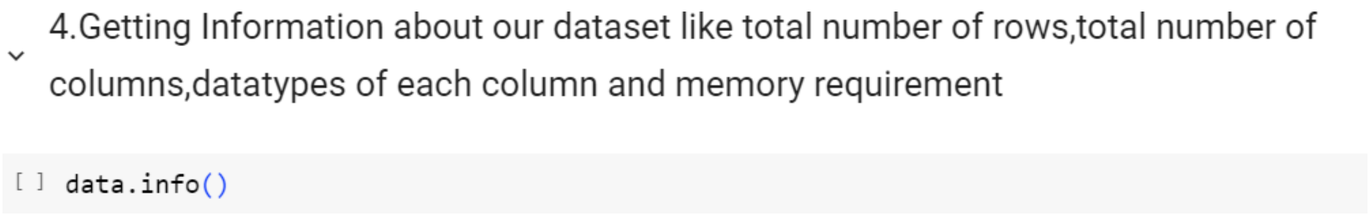
Figure 3.3



3.FIND SHAPE OF OUR DATASET(Number of Rows and Number of Columns)

```
[ ] data.shape
```

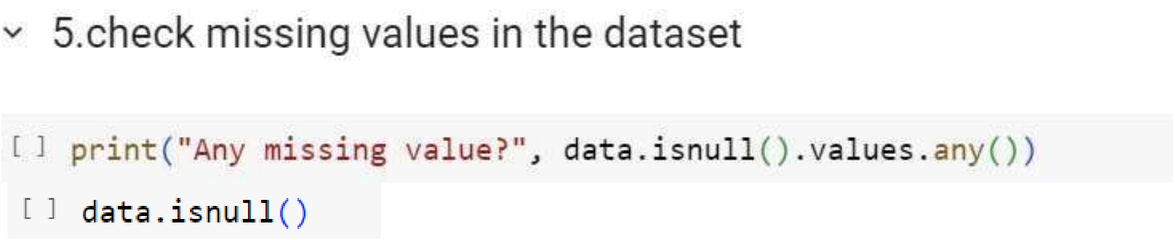
Figure 3.4



4.Getting Information about our dataset like total number of rows,total number of columns,datatypes of each column and memory requirement

```
[ ] data.info()
```

Figure 3.5

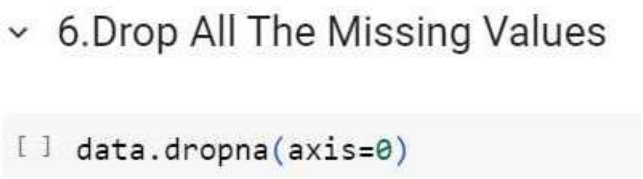


5.check missing values in the dataset

```
print("Any missing value?", data.isnull().values.any())
```

```
data.isnull()
```

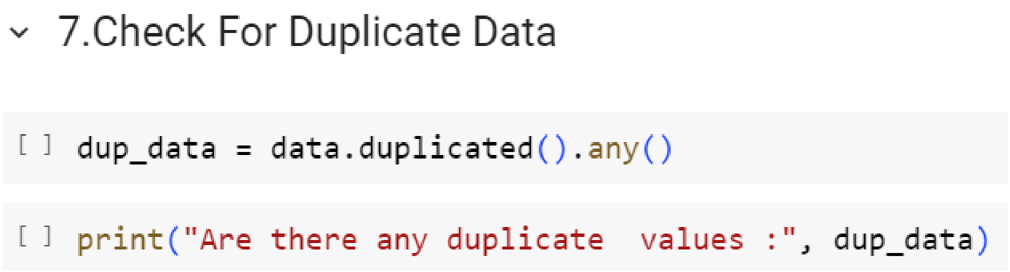
Figure 3.6



6.Drop All The Missing Values

```
data.dropna(axis=0)
```

Figure 3.7

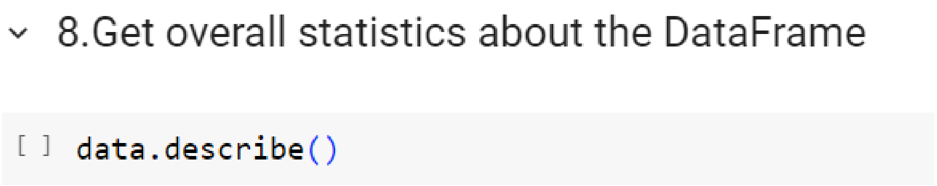


7.Check For Duplicate Data

```
dup_data = data.duplicated().any()
```

```
print("Are there any duplicate  values :", dup_data)
```

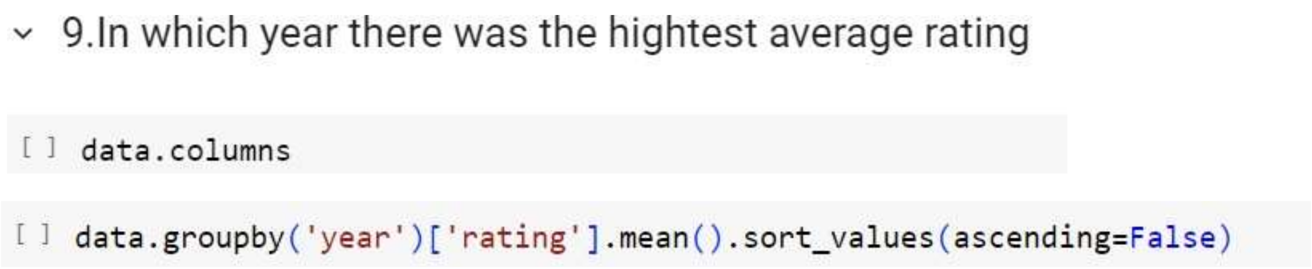
Figure 3.8



8.Get overall statistics about the DataFrame

```
data.describe()
```

Figure 3.9

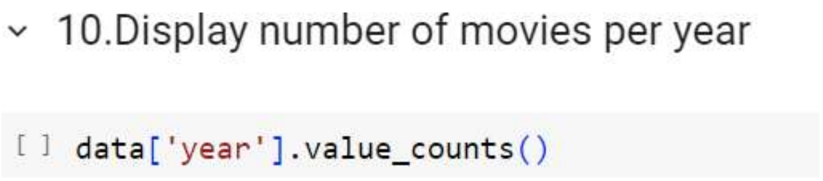


9.In which year there was the hightest average rating

```
data.columns
```

```
data.groupby('year')['rating'].mean().sort_values(ascending=False)
```

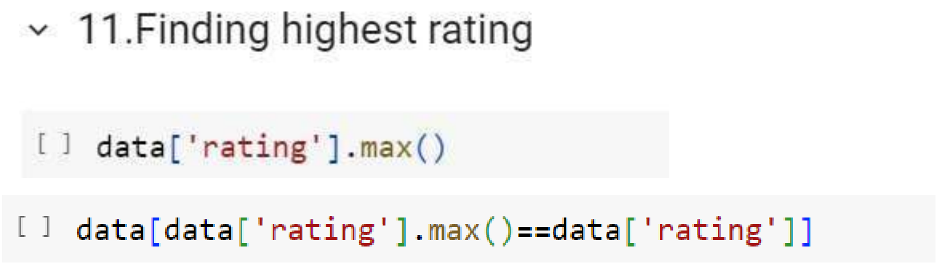
Figure 3.10



10.Display number of movies per year

```
data['year'].value_counts()
```

Figure 3.11

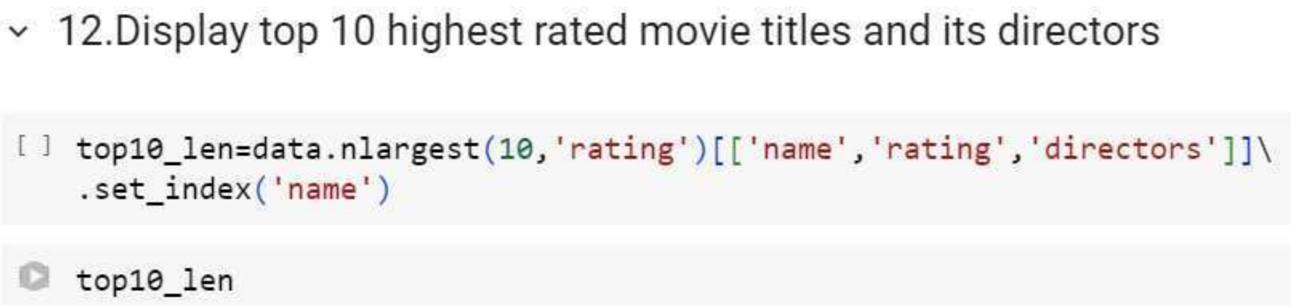


11.Finding highest rating

```
data['rating'].max()
```

```
data[data['rating'].max()==data['rating']]
```

Figure 3.12

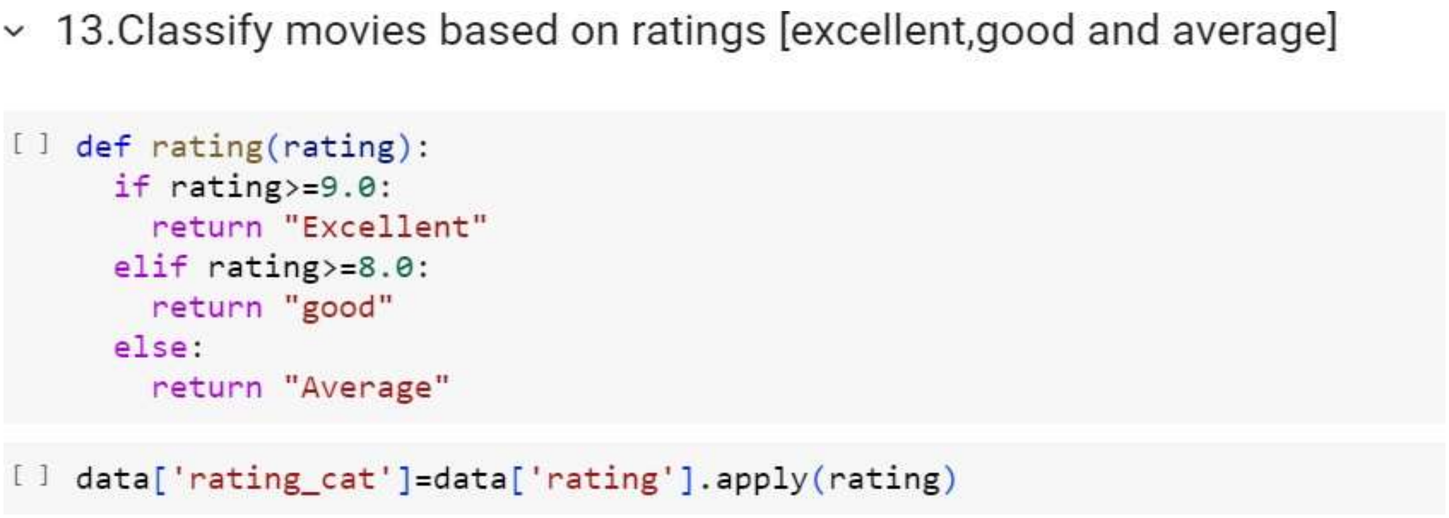


12.Display top 10 highest rated movie titles and its directors

```
top10_len=data.nlargest(10,'rating')[['name','rating','directors']]\
.set_index('name')
```

```
top10_len
```

Figure 3.13

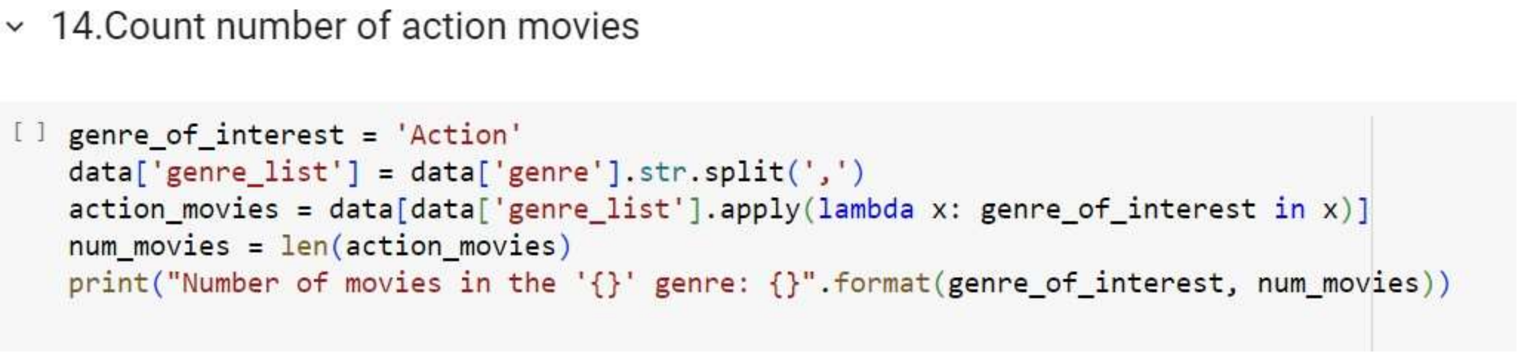


13.Classify movies based on ratings [excellent,good and average]

```
def rating(rating):
  if rating>=9.0:
    return "Excellent"
  elif rating>=8.0:
    return "good"
  else:
    return "Average"
```

```
data['rating_cat']=data['rating'].apply(rating)
```

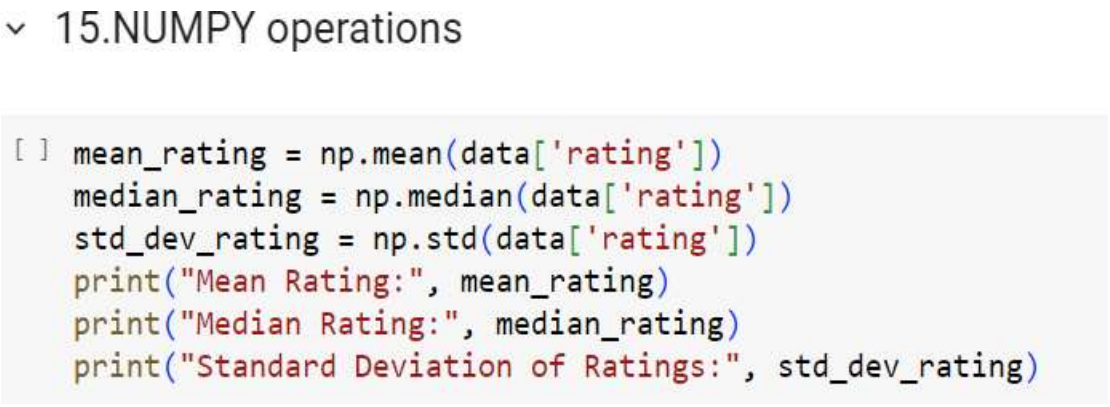
Figure 3.14



14.Count number of action movies

```
genre_of_interest = 'Action'
data['genre_list'] = data['genre'].str.split(',')
action_movies = data[data['genre_list'].apply(lambda x: genre_of_interest in x)]
num_movies = len(action_movies)
print("Number of movies in the '{}' genre: {}".format(genre_of_interest, num_movies))
```

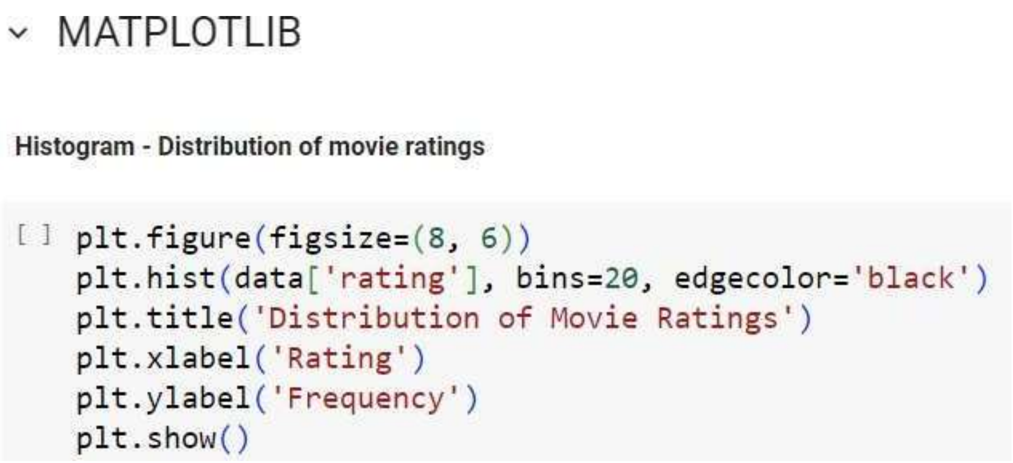
Figure 3.15



15.NUMPY operations

```
mean_rating = np.mean(data['rating'])
median_rating = np.median(data['rating'])
std_dev_rating = np.std(data['rating'])
print("Mean Rating:", mean_rating)
print("Median Rating:", median_rating)
print("Standard Deviation of Ratings:", std_dev_rating)
```

Figure 3.16

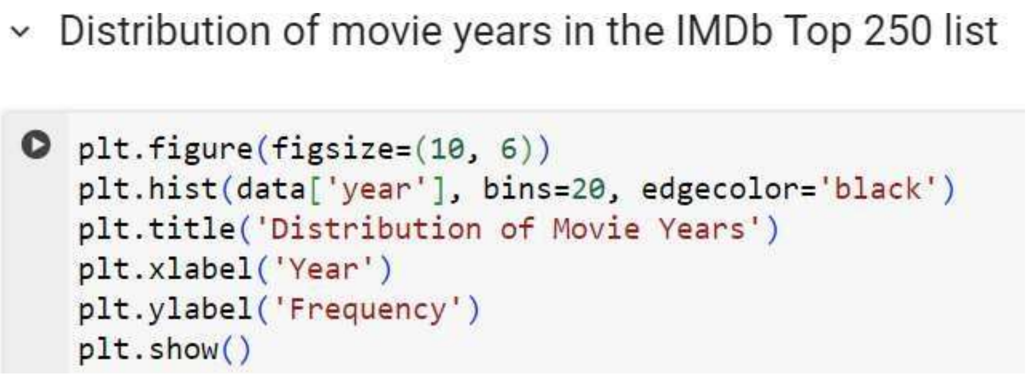


MATPLOTLIB

Histogram - Distribution of movie ratings

```
[ ] plt.figure(figsize=(8, 6))
    plt.hist(data['rating'], bins=20, edgecolor='black')
    plt.title('Distribution of Movie Ratings')
    plt.xlabel('Rating')
    plt.ylabel('Frequency')
    plt.show()
```

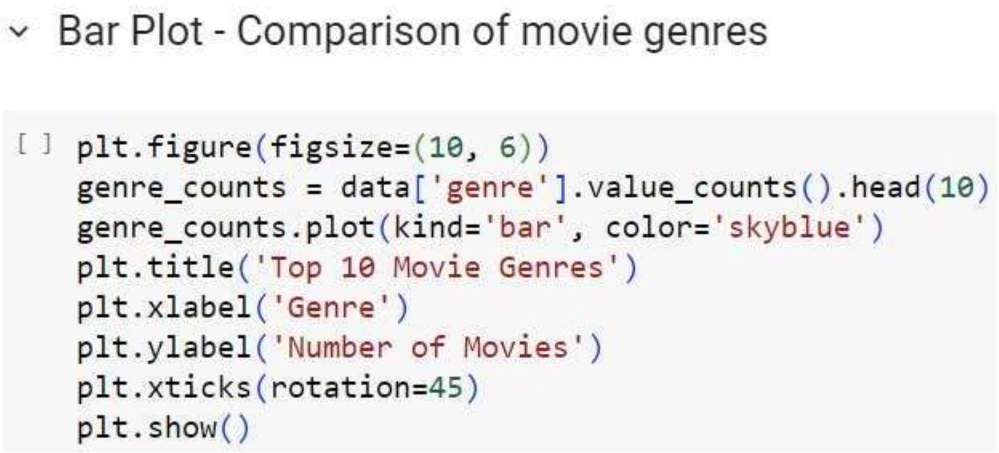
Figure 3.17



Distribution of movie years in the IMDb Top 250 list

```
plt.figure(figsize=(10, 6))
plt.hist(data['year'], bins=20, edgecolor='black')
plt.title('Distribution of Movie Years')
plt.xlabel('Year')
plt.ylabel('Frequency')
plt.show()
```

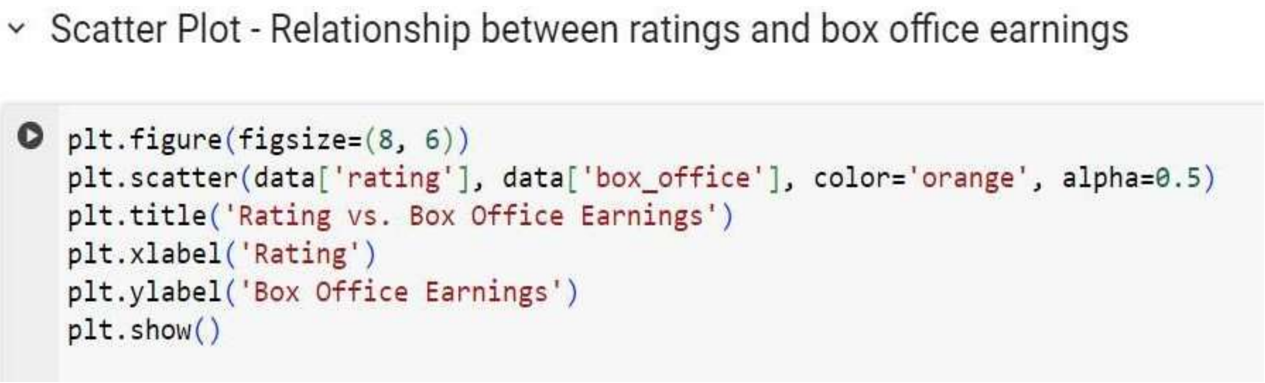
Figure 3.18



Bar Plot - Comparison of movie genres

```
[ ] plt.figure(figsize=(10, 6))
    genre_counts = data['genre'].value_counts().head(10)
    genre_counts.plot(kind='bar', color='skyblue')
    plt.title('Top 10 Movie Genres')
    plt.xlabel('Genre')
    plt.ylabel('Number of Movies')
    plt.xticks(rotation=45)
    plt.show()
```

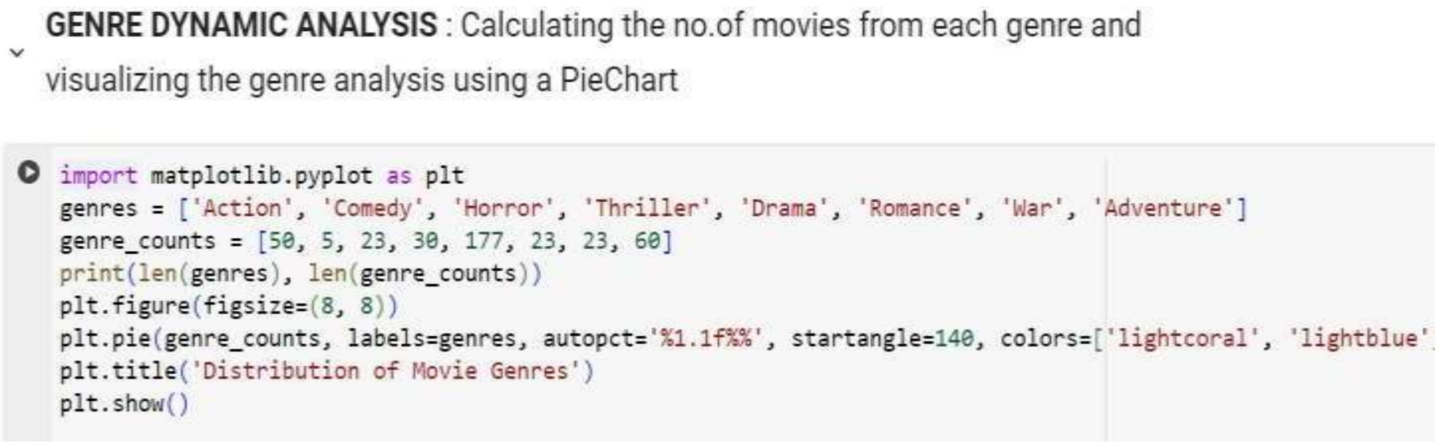
Figure 3.19



Scatter Plot - Relationship between ratings and box office earnings

```
plt.figure(figsize=(8, 6))
plt.scatter(data['rating'], data['box_office'], color='orange', alpha=0.5)
plt.title('Rating vs. Box Office Earnings')
plt.xlabel('Rating')
plt.ylabel('Box Office Earnings')
plt.show()
```

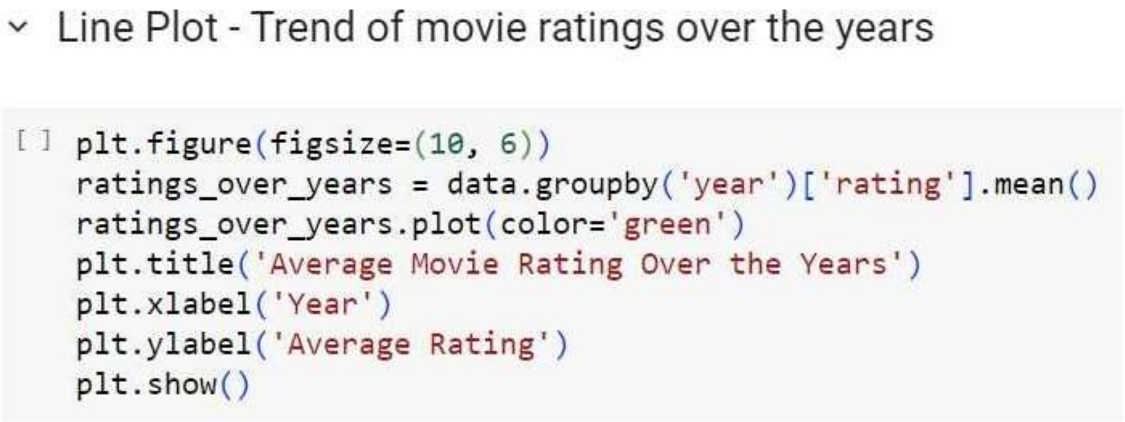
Figure 3.20



GENRE DYNAMIC ANALYSIS : Calculating the no.of movies from each genre and visualizing the genre analysis using a PieChart

```
import matplotlib.pyplot as plt
genres = ['Action', 'Comedy', 'Horror', 'Thriller', 'Drama', 'Romance', 'War', 'Adventure']
genre_counts = [50, 5, 23, 30, 177, 23, 23, 60]
print(len(genres), len(genre_counts))
plt.figure(figsize=(8, 8))
plt.pie(genre_counts, labels=genres, autopct='%1.1f%%', startangle=140, colors=['lightcoral', 'lightblue'
plt.title('Distribution of Movie Genres')
plt.show()
```

Figure 3.21



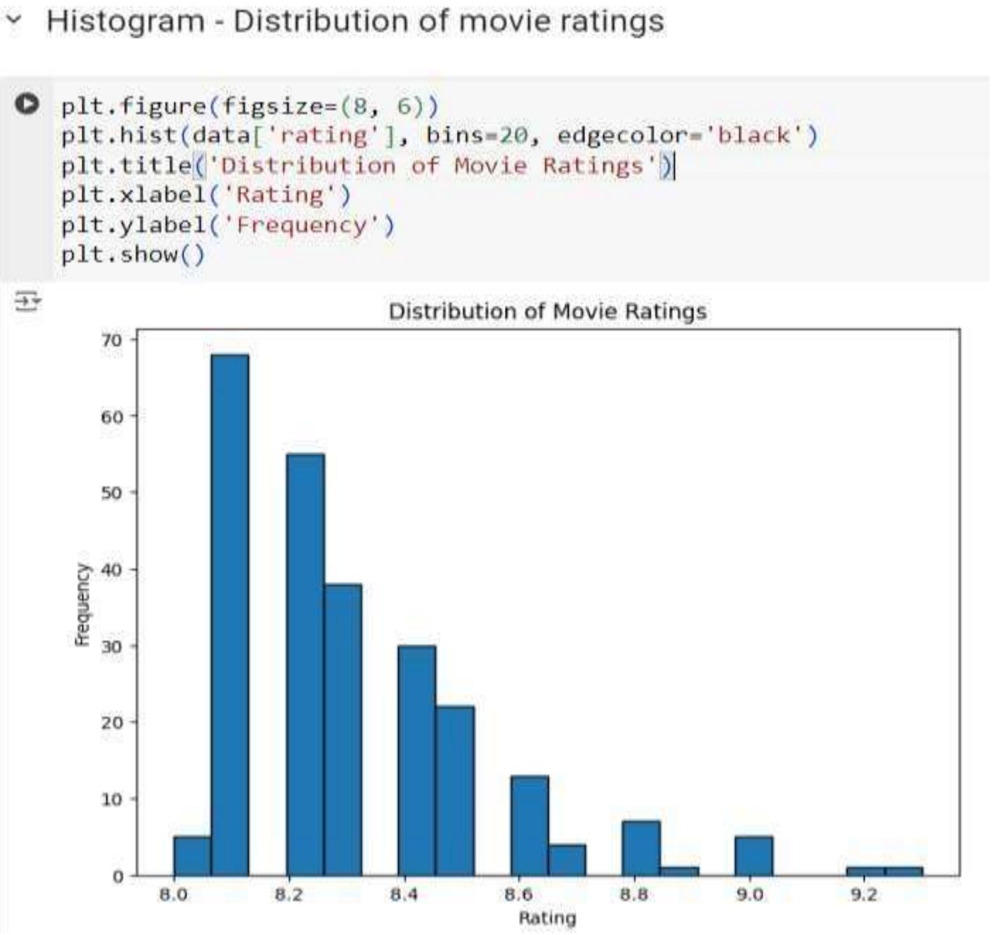
Line Plot - Trend of movie ratings over the years

```
plt.figure(figsize=(10, 6))
ratings_over_years = data.groupby('year')['rating'].mean()
ratings_over_years.plot(color='green')
plt.title('Average Movie Rating Over the Years')
plt.xlabel('Year')
plt.ylabel('Average Rating')
plt.show()
```

Figure 3.22

CHAPTER 4

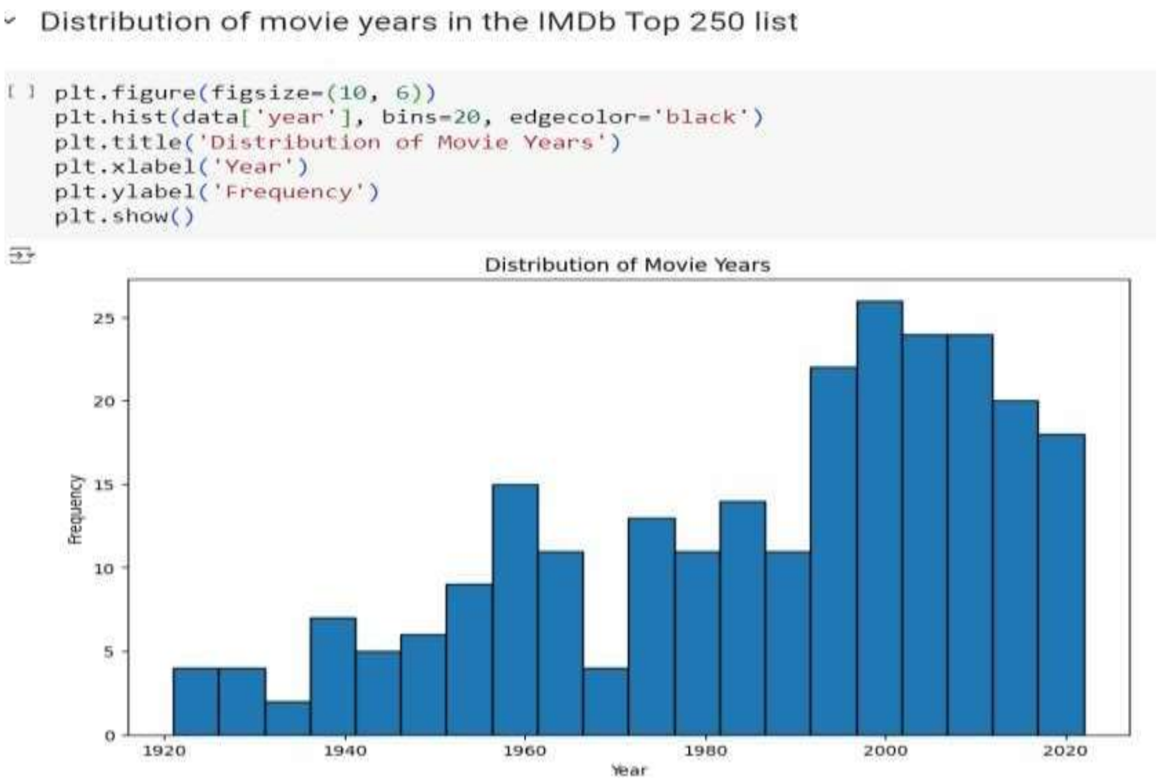
Result Analysis



Histogram - Distribution of movie ratings

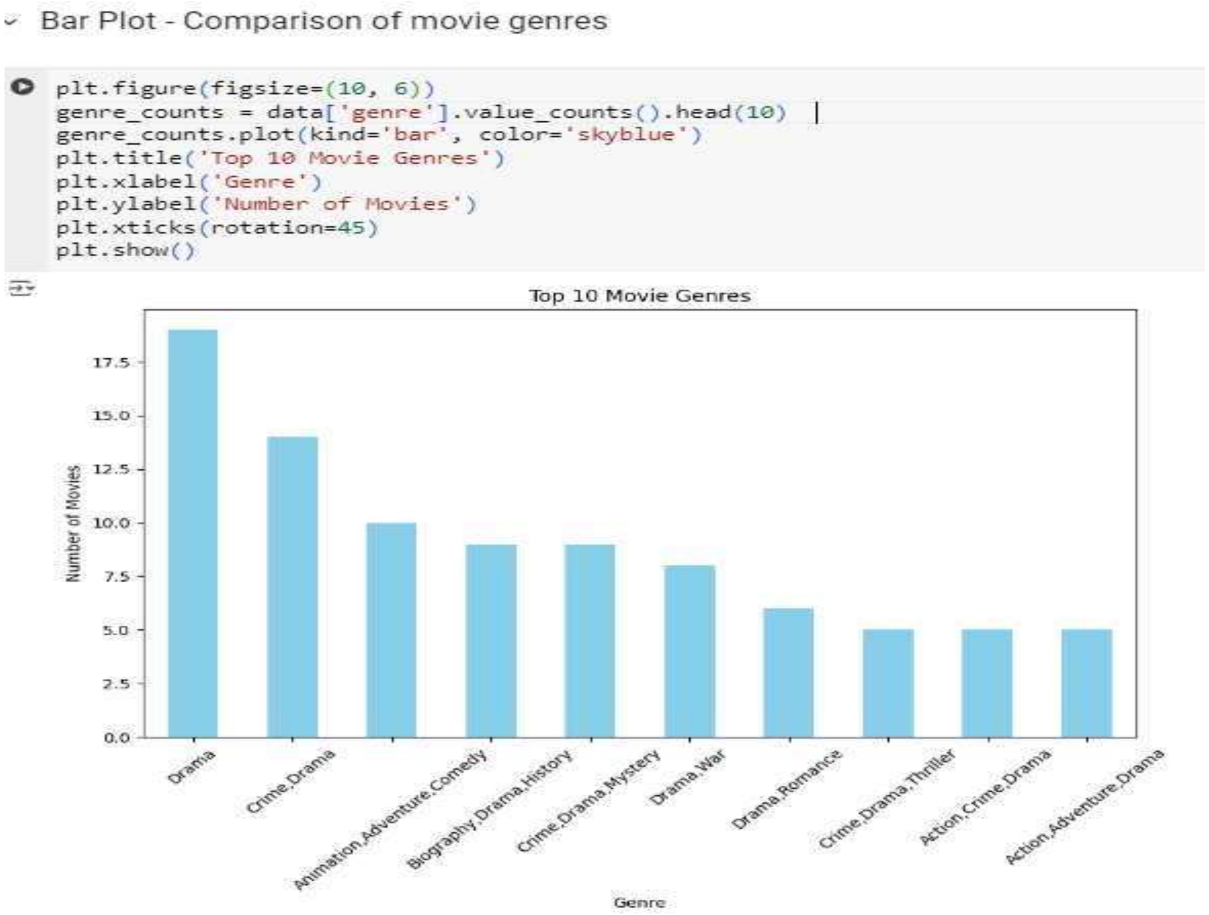
```
plt.figure(figsize=(8, 6))
plt.hist(data['rating'], bins=20, edgecolor='black')
plt.title('Distribution of Movie Ratings')
plt.xlabel('Rating')
plt.ylabel('Frequency')
plt.show()
```

Figure 4.1



Distribution of movie years in the IMDb Top 250 list

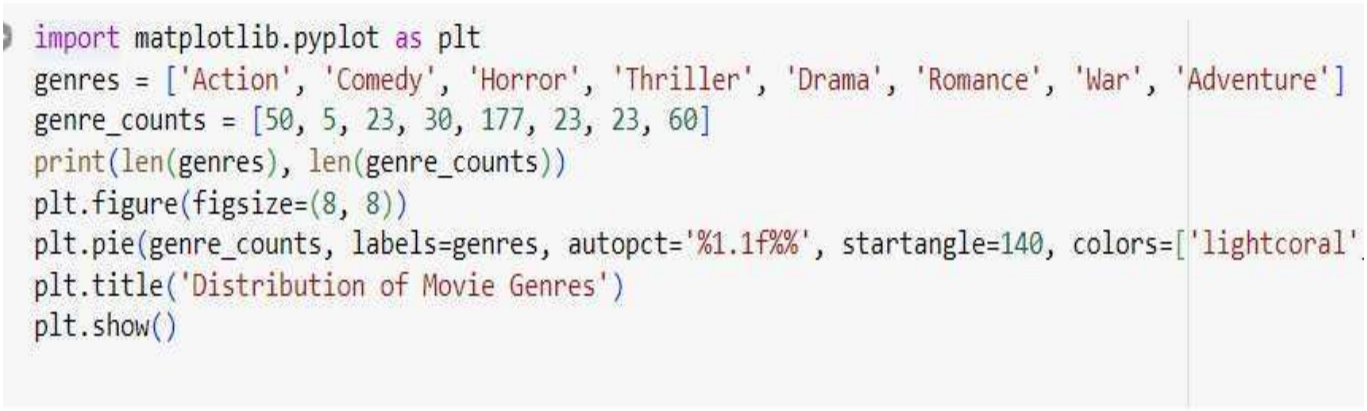
```
plt.figure(figsize=(10, 6))
plt.hist(data['year'], bins=20, edgecolor='black')
plt.title('Distribution of Movie Years')
plt.xlabel('Year')
plt.ylabel('Frequency')
plt.show()
```



Bar Plot - Comparison of movie genres

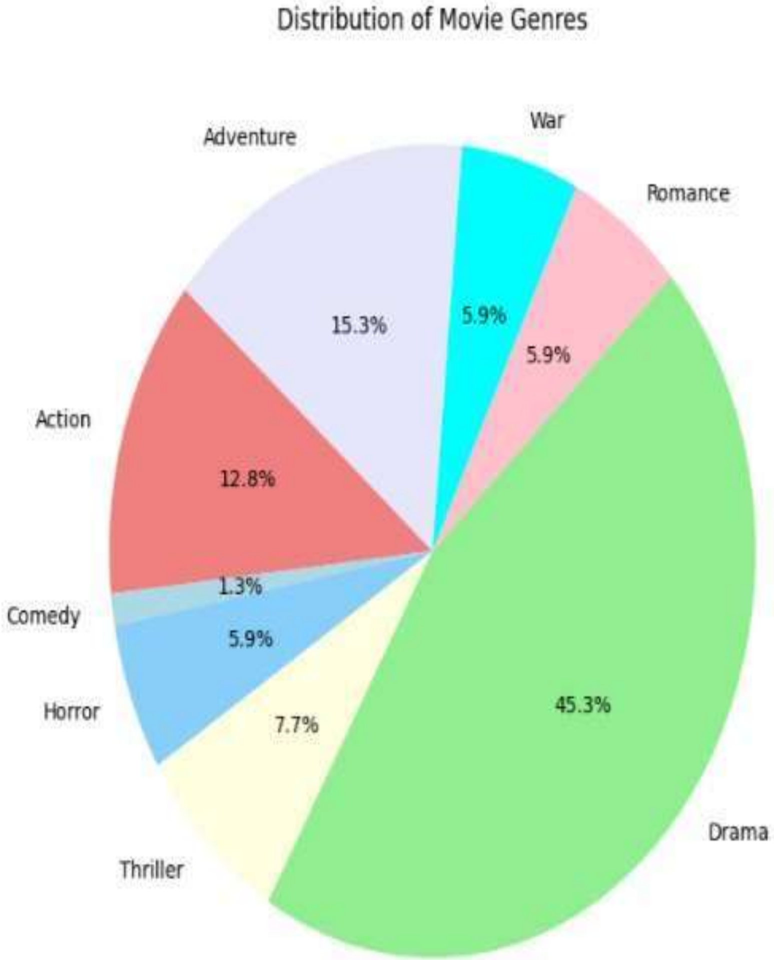
```
plt.figure(figsize=(10, 6))
genre_counts = data['genre'].value_counts().head(10)
genre_counts.plot(kind='bar', color='skyblue')
plt.title('Top 10 Movie Genres')
plt.xlabel('Genre')
plt.ylabel('Number of Movies')
plt.xticks(rotation=45)
plt.show()
```

GENRE DYNAMIC ANALYSIS : Calculating the no.of movies from each genre and visualizing the genre analysis using a PieChart

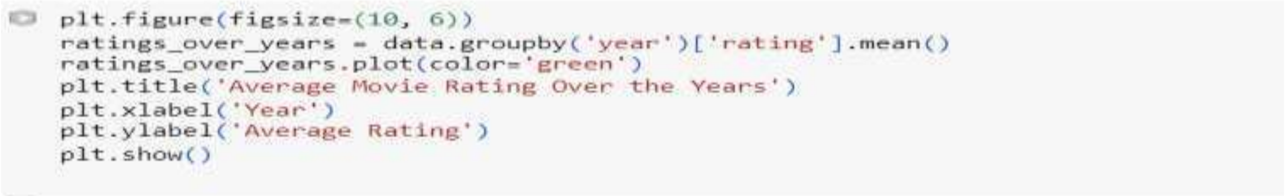


```
import matplotlib.pyplot as plt
genres = ['Action', 'Comedy', 'Horror', 'Thriller', 'Drama', 'Romance', 'War', 'Adventure']
genre_counts = [50, 5, 23, 30, 177, 23, 23, 60]
print(len(genres), len(genre_counts))
plt.figure(figsize=(8, 8))
plt.pie(genre_counts, labels=genres, autopct='%1.1f%%', startangle=140, colors=['lightcoral'
plt.title('Distribution of Movie Genres')
plt.show()
```

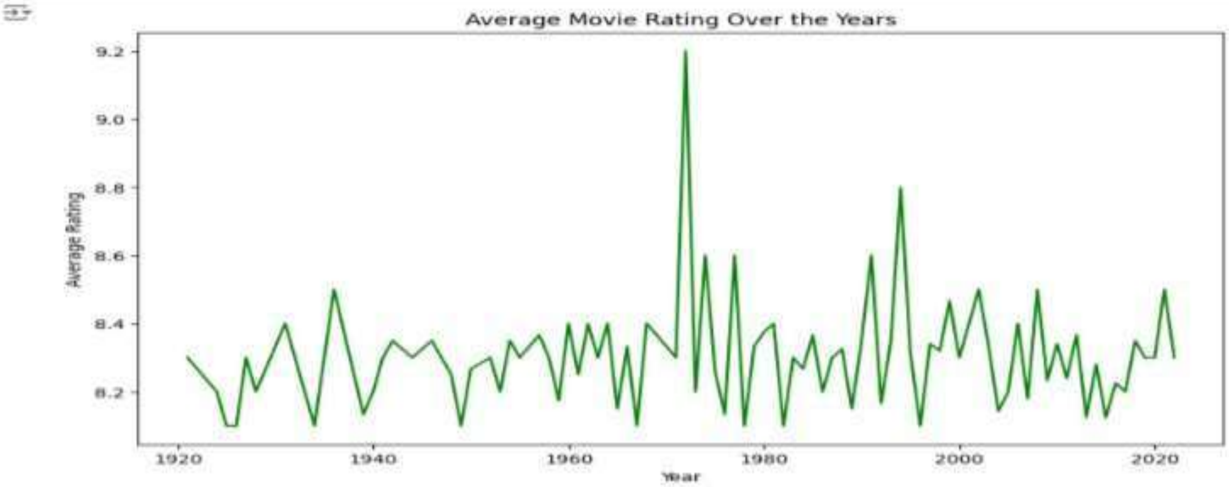
```
8 8
```



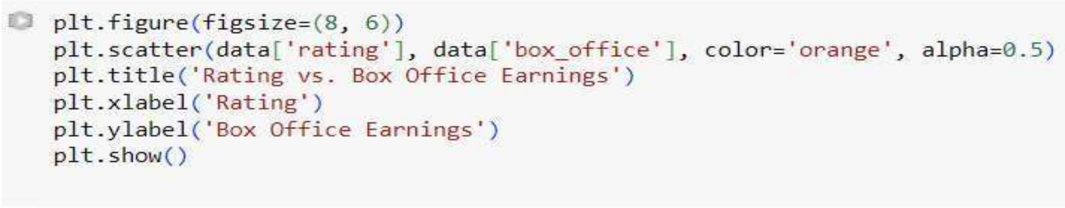
Line Plot - Trend of movie ratings over the years



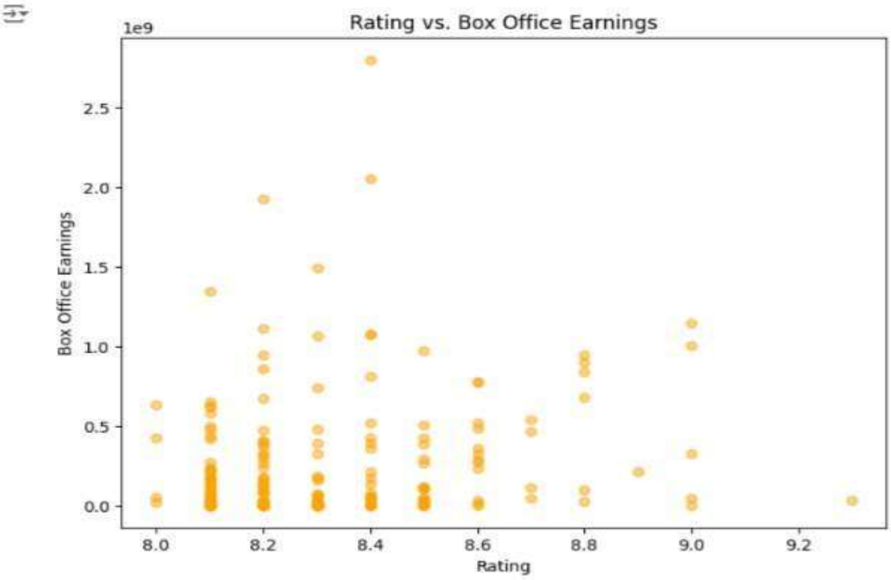
```
plt.figure(figsize=(10, 6))
ratings_over_years = data.groupby('year')['rating'].mean()
ratings_over_years.plot(color='green')
plt.title('Average Movie Rating Over the Years')
plt.xlabel('Year')
plt.ylabel('Average Rating')
plt.show()
```



Scatter Plot - Relationship between ratings and box office earnings



```
plt.figure(figsize=(8, 6))
plt.scatter(data['rating'], data['box_office'], color='orange', alpha=0.5)
plt.title('Rating vs. Box Office Earnings')
plt.xlabel('Rating')
plt.ylabel('Box Office Earnings')
plt.show()
```



CHAPTER 4
Conclusion

In summary, our comprehensive analysis of the IMDb Top 250 Movies dataset has provided valuable insights into the world of cinema. By meticulously following a structured approach encompassing data loading, cleaning, exploration, visualization, and additional operations, we have unearthed meaningful patterns and trends within the dataset. Through descriptive statistics, we gained a quantitative understanding of movie ratings and genres, while exploratory analyses revealed nuanced insights into audience preferences and cinematic trends. Our visualizations brought these findings to life, offering intuitive representations of data distributions and relationships.

Furthermore, our rigorous data cleaning processes ensured the reliability and integrity of our results, instilling confidence in our conclusions. Ultimately, this project highlights the importance of data-driven analysis in understanding complex phenomena such as movie ratings and genres, and underscores the value of systematic approaches in extracting actionable insights from large datasets.

References

[1] Matplotlib: A 2D Graphics Environment ,John D. Hunter

Computing in science & engineering (Print) 2007. 17993 Citations, 1 References.

Python Data Analytics: With Pandas, NumPy, and Matplotlib
Fabio Nelli 2023

[2] Research on Big Data Analysis Data Acquisition and Data Analysis
Hong Li
2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)

[3] Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory
Saeed Mian Qaisar

[4] Web-Based Visualization of Marine Environmental Data: Performance Analysis of a MatPlotLib Implementation.
Joseph A. James;Teng-Sheng Moh;Christopher A. Edwards