Foundation-Course-in-Data-Science

Board Infinity

A training report

Student Declaration

To whom so ever it may concern I, Venkata Siva Kalyan.Bavireddy**, hereby declare that the work done by me on**

“Foundation-Course-in DataScience”

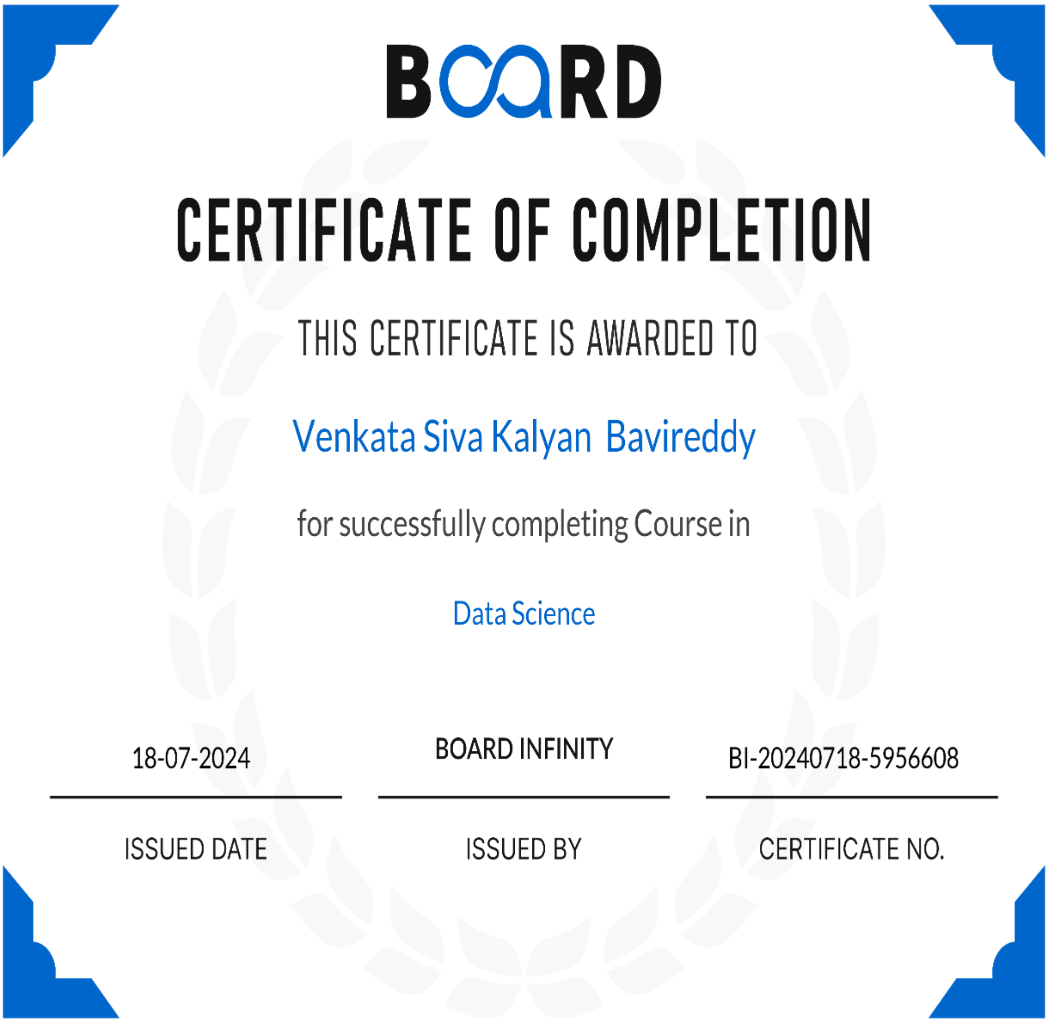
from **28/05/24** to **25/06/24** , is a record of original work for the partial fulfillment of the requirements for the award of the degree, **B. Tech Computer Science & Engineering.**

Venkata Siva Kalyan.Bavireddy

Signature of the student

S.NO	TOPIC
1	Certificate from organization
2	Module1: Advance Excel
3	Module2: Data Visuvalisation using tableau
4	Module3: Advance SQL
5	Module4: Programming for Data Science Python
6	Module5: Project
7	Conclusion

Certificate from organization.



BOARD

CERTIFICATE OF COMPLETION

THIS CERTIFICATE IS AWARDED TO

Venkata Siva Kalyan Bavireddy

for successfully completing Course in

Data Science

18-07-2024	BOARD INFINITY	BI-20240718-5956608
ISSUED DATE	ISSUED BY	CERTIFICATE NO.

Module 1:Advance Excel

The Rapid growth of Microsoft Excel 2010 marked a transformative phase for the software, particularly with the introduction of Power Pivot and Power Query.

Power Pivot enabled users to import, combine, and analyze large sets of data from multiple sources, using a powerful in-memory data model. This allowed for complex calculations and relationships, making it easier to perform data modeling directly within Excel.

Power Query enhanced data retrieval, offering robust tools for data transformation and cleaning before analysis. It simplified the process of importing data from various sources, such as databases, web services, and other files, ensuring users could work with clean, structured data.

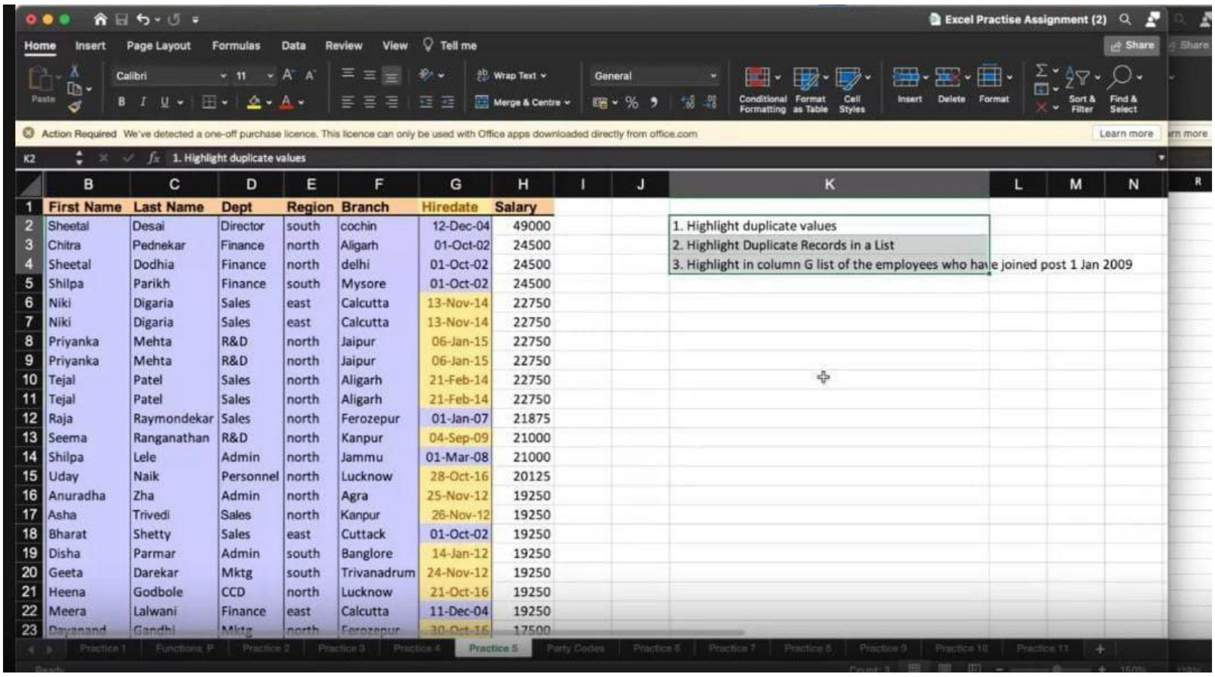
These features, integral to the Microsoft Power BI ecosystem, turned Excel into a more comprehensive business intelligence tool, making data analysis more accessible and efficient for users across various industries. Subsequent versions of Excel continued to build on these functionalities, further solidifying its role in advanced data analytics.

Advanced Excel skills are important in many industries because these industries often deal with a lot of data that needs to be analyzed and reported. Here's how different fields use these skills:

1. Financial Services: Banks and investment firms use Excel to create financial plans, analyze risks, and report on performance.
2. Management Consulting: Consulting firms analyze data and conduct market research using Excel to help businesses make better decisions.
3. Information Technology: IT companies use Excel to track projects, analyze data, and create complex spreadsheets for various tech tasks.
4. Healthcare: Hospitals and healthcare organizations use Excel to analyze patient information, forecast finances, and create operational reports.
5. Marketing and Advertising: Marketing companies track the performance of their campaigns and analyze data about customers using Excel.
6. Manufacturing and Logistics: Businesses in manufacturing and logistics use Excel for managing supply chains, tracking inventory, and planning production schedules.
7. Education and Research: Schools and research organizations rely on Excel for analyzing data, performing statistical calculations, and building models for studies.

In short, advanced Excel skills help professionals organize and make sense of data, which is crucial for making informed decisions and achieving business goals.

These are just a few examples, but in reality, advanced Excel skills are valued across a wide array of industries due to the software's widespread use for data management and analysis.



Excel Practise Assignment (2)
Home
Insert
Page Layout
Formulas
Data
Review
View
Tell me
Share
Action Required We've detected a one-off purchase licence. This licence can only be used with Office apps downloaded directly from office.com
Learn more
K2
Practice 5

1. Highlight duplicate values

	First Name	Last Name	Dept	Region	Branch	Hiredate	Salary
2	Sheetal	Desai	Director	south	cochin	12-Dec-04	49000
3	Chitra	Pednekar	Finance	north	Aligarh	01-Oct-02	24500
4	Sheetal	Dodhia	Finance	north	delhi	01-Oct-02	24500
5	Shilpa	Parikh	Finance	south	Mysore	01-Oct-02	24500
6	Niki	Digaria	Sales	east	Calcutta	13-Nov-14	22750
7	Niki	Digaria	Sales	east	Calcutta	13-Nov-14	22750
8	Priyanka	Mehta	R&D	north	Jaipur	06-Jan-15	22750
9	Priyanka	Mehta	R&D	north	Jaipur	06-Jan-15	22750
10	Tejal	Patel	Sales	north	Aligarh	21-Feb-14	22750
11	Tejal	Patel	Sales	north	Aligarh	21-Feb-14	22750
12	Raja	Raymondekar	Sales	north	Ferozepur	01-Jan-07	21875
13	Seema	Ranganathan	R&D	north	Kanpur	04-Sep-09	21000
14	Shilpa	Lele	Admin	north	Jammu	01-Mar-08	21000
15	Uday	Naik	Personnel	north	Lucknow	28-Oct-16	20125
16	Anuradha	Zha	Admin	north	Agra	25-Nov-12	19250
17	Asha	Trivedi	Sales	north	Kanpur	26-Nov-12	19250
18	Bharat	Shetty	Sales	east	Cuttack	01-Oct-02	19250
19	Disha	Parmar	Admin	south	Banglore	14-Jan-12	19250
20	Geeta	Darekar	Mktg	south	Trivanadrum	24-Nov-12	19250
21	Heena	Godbole	CCD	north	Lucknow	21-Oct-16	19250
22	Meera	Lalwani	Finance	east	Calcutta	11-Dec-04	19250
23	Dayanand	Gandhi	Mktg	north	Ferozepur	30-Oct-16	17500

1. Highlight duplicate values
2. Highlight Duplicate Records in a List
3. Highlight in column G list of the employees who have joined post 1 Jan 2009

Advanced Excel is essential for data analysis due to its powerful features that help users work with large amounts of data effectively. Here are the key roles it plays:

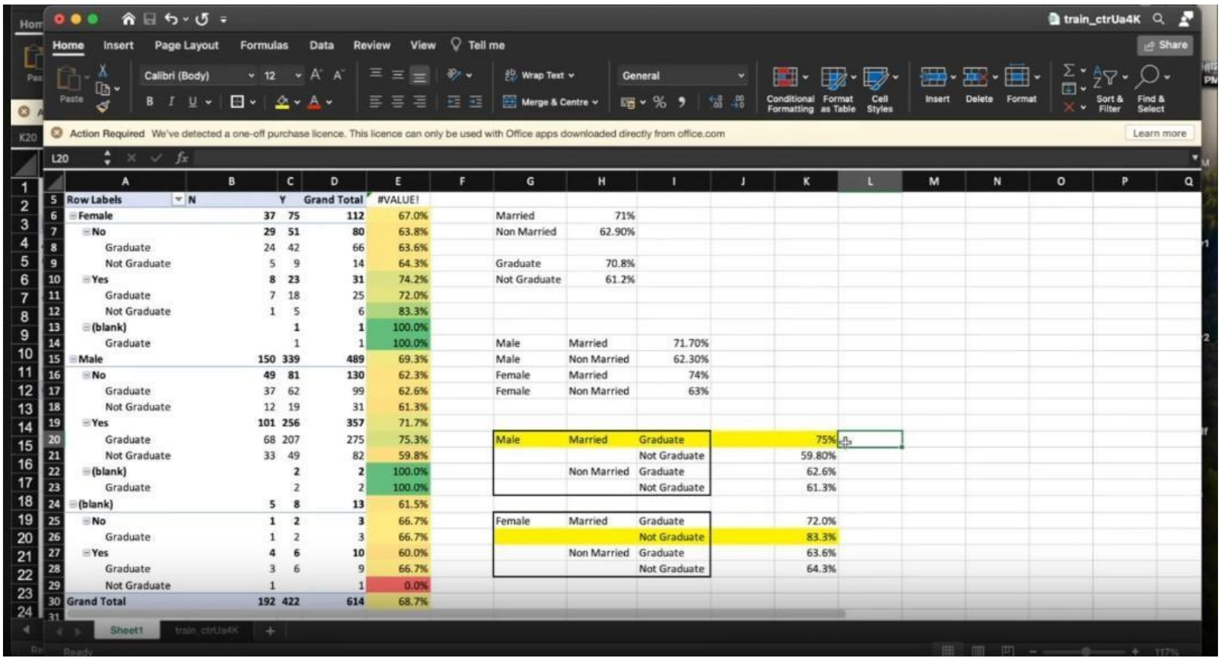
1. Data Cleaning and Transformation: Tools like Power Query and Power Pivot help users clean and structure raw data from different sources, making it ready for analysis.

2. Data Modeling and Calculation: Excel allows users to create complex data models and perform advanced calculations. Users can build custom formulas to analyze data and extract meaningful insights.

3. Visualization and Reporting: With advanced charting tools and features like PivotTables, Excel helps users visualize data trends and patterns. This enables the creation of detailed reports and dashboards, making data easier to understand and communicate.

4. Statistical Analysis: Excel supports various statistical operations, such as regression analysis, correlation, and t-tests, making it easier to analyze data statistically.

In summary, advanced Excel is a powerful platform for data analysis, offering a wide range of tools that help users explore, interpret, and present data effectively, which aids in informed decision-making.



Action Required We've detected a one-off purchase licence. This licence can only be used with Office apps downloaded directly from office.com

Row Labels	N	Y	Grand Total	#VALUE!
Female	37	75	112	67.0%
No	29	51	80	63.8%
Graduate	24	42	66	63.6%
Not Graduate	5	9	14	64.3%
Yes	8	23	31	74.2%
Graduate	7	18	25	72.0%
Not Graduate	1	5	6	83.3%
(blank)		1	1	100.0%
Graduate		1	1	100.0%
Male	150	339	489	69.3%
No	49	81	130	62.3%
Graduate	37	62	99	62.6%
Not Graduate	12	19	31	61.3%
Yes	101	256	357	71.7%
Graduate	68	207	275	75.3%
Not Graduate	33	49	82	59.8%
(blank)		2	2	100.0%
Graduate		2	2	100.0%
(blank)	5	8	13	61.5%
No	1	2	3	66.7%
Graduate	1	2	3	66.7%
Yes	4	6	10	60.0%
Graduate	3	6	9	66.7%
Not Graduate	1		1	0.0%
Grand Total	192	422	614	68.7%

Married	71%
Non Married	62.90%

Graduate	70.8%
Not Graduate	61.2%

Male	Married	71.70%
Male	Non Married	62.30%
Female	Married	74%
Female	Non Married	63%

Male	Married	Graduate	75%
		Not Graduate	59.80%
	Non Married	Graduate	62.6%
		Not Graduate	61.3%

Female	Married	Graduate	72.0%
		Not Graduate	83.3%
	Non Married	Graduate	63.6%
		Not Graduate	64.3%

Advantages of Advanced Excel:

1. Powerful Data Analysis: Advanced Excel has many tools to help users analyze large sets of data quickly and effectively, making it valuable for businesses.

2. Dynamic Reporting: Features like PivotTables and Power Query allow users to create interactive reports and visualizations that clearly communicate insights.

3. Automation and Efficiency: Users can automate repetitive tasks and complex calculations with tools like VBA (Visual Basic for Applications), saving time and boosting productivity.

4. Widely Used: Excel is commonly used across many industries, so knowing advanced Excel skills can open up job opportunities and help professionals in various roles.

Disadvantages of Advanced Excel:

1. Steep Learning Curve: It can take a lot of time and effort to learn how to use Excel's advanced features effectively.

2. Version Compatibility: Sometimes, files created with advanced features may not work properly on older versions of Excel, which can cause sharing issues.

3. Limited Scalability for Big Data: While great for medium-sized data, Excel can struggle with very large datasets, leading to slow performance or problems.

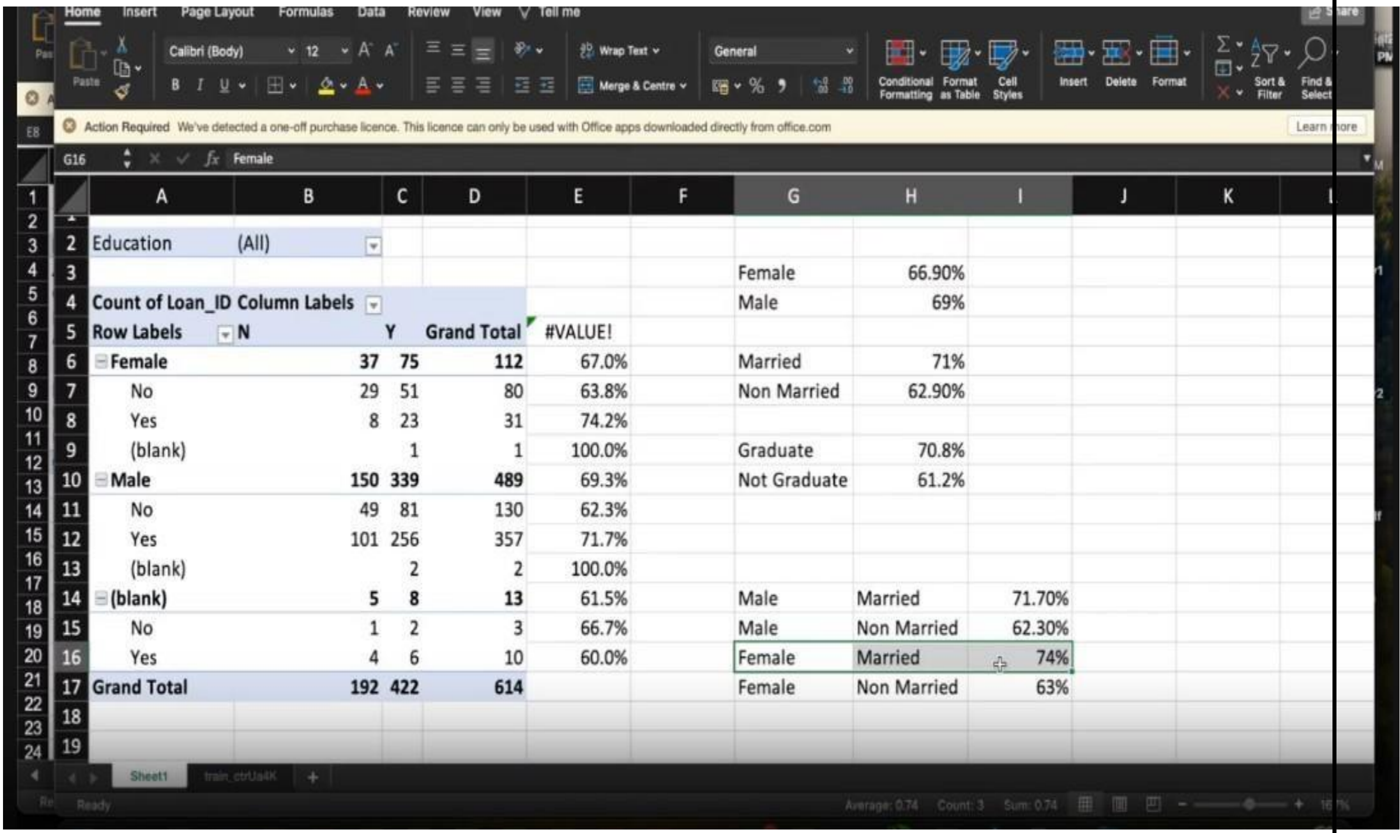
4. Overreliance on Manual Processes: If users depend too much on manual input, it can lead to mistakes, inconsistencies, and issues with data quality.

5. Cost of Advanced Features: Some advanced features may only be available in certain versions of Excel, which could mean extra costs for organizations.

Learning advanced Excel can be very beneficial for your career. Here’s how:

1. Better Data Analysis: You can analyze data in complex ways, like making predictions and modeling different scenarios. This helps you make smarter decisions.
2. Easier Reporting and Visualization: You can create interactive reports and charts that make data easy to understand and share, helping others see important information clearly.
3. Automating Tasks: You can automate repetitive tasks and create custom tools, which saves time and makes your work more efficient.
4. Career Growth: Many employers look for advanced Excel skills. Having these skills can give you an edge when applying for jobs and help you move up in your career.
5. Flexibility: Since Excel is widely used in many industries, your advanced skills can be applied to a variety of jobs and companies.

In short, learning advanced Excel helps you work with data better, boosts your job prospects, and allows for growth in data-focused careers.



Action Required We've detected a one-off purchase licence. This licence can only be used with Office apps downloaded directly from office.com

G16 *fx* Female

	A	B	C	D	E	F	G	H	I
2	Education	(All)							
3							Female	66.90%	
4	Count of Loan_ID	Column Labels					Male	69%	
5	Row Labels	N	Y	Grand Total	#VALUE!				
6	Female	37	75	112	67.0%		Married	71%	
7	No	29	51	80	63.8%		Non Married	62.90%	
8	Yes	8	23	31	74.2%				
9	(blank)		1	1	100.0%		Graduate	70.8%	
10	Male	150	339	489	69.3%		Not Graduate	61.2%	
11	No	49	81	130	62.3%				
12	Yes	101	256	357	71.7%				
13	(blank)		2	2	100.0%				
14	(blank)	5	8	13	61.5%		Male	Married	71.70%
15	No	1	2	3	66.7%		Male	Non Married	62.30%
16	Yes	4	6	10	60.0%		Female	Married	74%
17	Grand Total	192	422	614			Female	Non Married	63%

Module 2: Data Visualisation using tableau

Tableau Software was founded in 2003 by Chris Stolte, Pat Hanrahan, and Christian Chabot. Its goal is to help people see and understand their data better. The company introduced its first product, Tableau Desktop, in the same year, allowing users to connect to their data and visualize it easily.

Tableau is widely used for:

1. Data Visualization: Tableau is a powerful tool for creating interactive and shareable charts, graphs, and dashboards from your data. It has a user-friendly interface that makes it easy to work with different data sources.
2. Data Exploration and Analysis: Users can easily explore and analyze large amounts of data to find patterns, trends, and unusual points that can provide valuable insights.
3. Storytelling with Data: Tableau allows data scientists to tell a story with their data. They can build narratives around their findings, making it easier to communicate what the data means and what actions should be taken.

Many companies from different industries use Tableau for data visualization and analysis. Some famous companies that use Tableau include Verizon, Deloitte, Pfizer, and Netflix.

*Job Opportunities with Tableau:

There are several job roles that require Tableau skills, such as:

1. Tableau Developer: Creates and manages Tableau reports and dashboards.

2. Tableau Consultant: Provides expert advice on how to use Tableau effectively.

3. Business Intelligence Analyst: Analyzes data to support business decisions, often using Tableau tools.

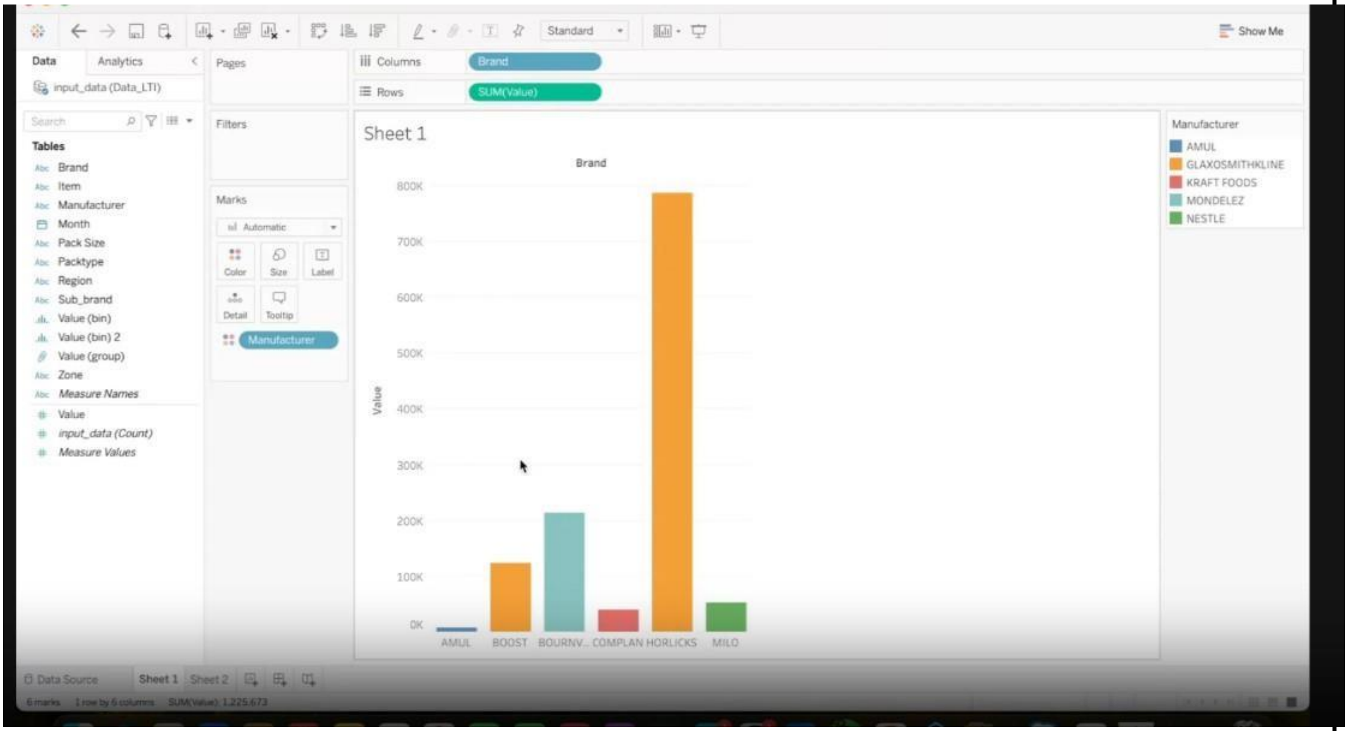
4. Data Visualization Specialist: Focuses on making data visually appealing and easy to understand.

5. Data Analyst: Examines data to find trends and insights, typically presenting results in Tableau.

*Popularity and Community.

Tableau is very well-known and considered one of the top tools for data visualization and analysis. Many data professionals choose Tableau because it is user-friendly, has powerful features, and allows for interactive visualizations.

The Tableau community is active and helpful, offering resources, online forums, and events for users to learn and connect. This helps both beginners and experienced users improve their Tableau skills.



Columns: Brand

Rows: SUM(Value)

Marks: Automatic

Manufacturer

6 marks 1 row by 6 columns SUM(Value): 1,225,673

Tableau has several important features that help you analyze and visualize data easily. Here are the main things you can do with Tableau:

Key Functions of Tableau

1. Connect to Data Sources: You can link Tableau to different data sources like Excel files, CSV files, databases (such as SQL Server and MySQL), and cloud services (like Google Analytics and Salesforce). This helps you bring in data from various places.

2. Data Preparation: After connecting to your data, Tableau offers tools to clean and organize it. You can change, filter, combine, and summarize data within Tableau without needing other programs.

3. Build Visualizations: One of Tableau's strengths is creating a variety of interactive and attractive visualizations. You can make different types of charts (like bar charts, line charts, scatter plots, and maps) as well as dashboards to show data clearly.

4. Analytics and Calculations: Tableau has built-in functions for calculations and statistics. You can create new formulas, summarize data, and add trend lines to help analyze it.

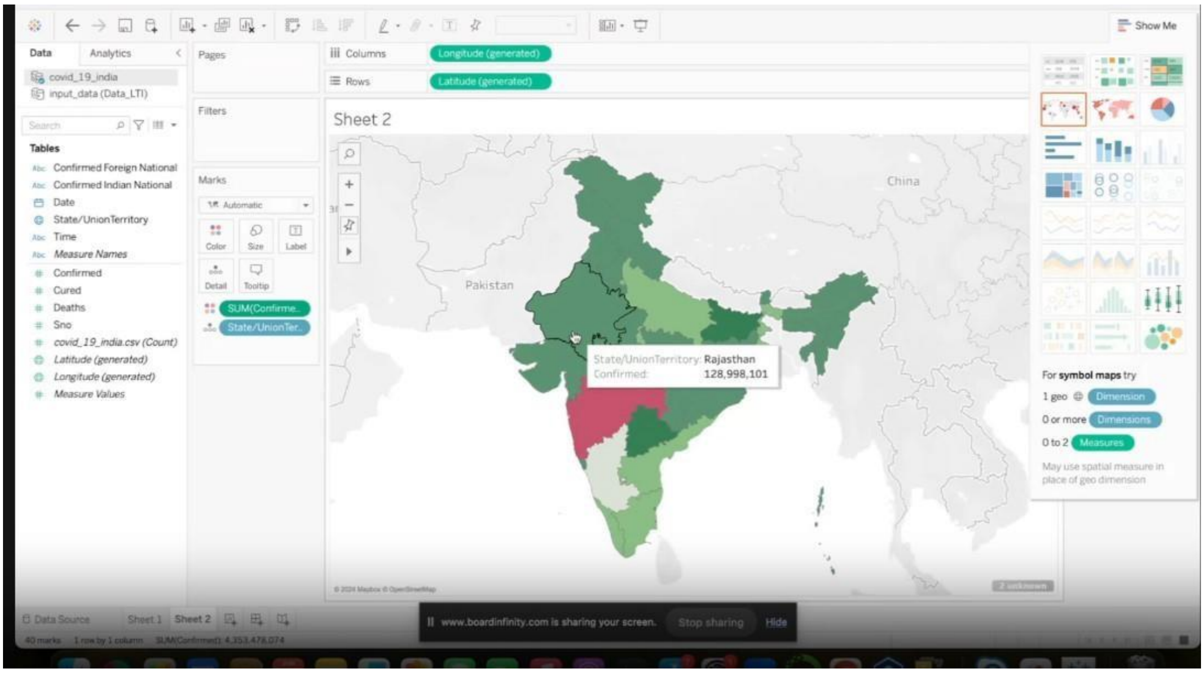
5. Dashboard Creation: You can put multiple visualizations together in interactive dashboards. These dashboards can have filters and buttons that let users explore the data themselves.

6. Sharing and Collaboration: Tableau makes it easy to share your work. You can publish your dashboards online so others can access them and also export visuals as images or PDFs to share easily.

7. Integration with Other Tools: Tableau can connect with other tools like R and Python, which adds more features for advanced analysis.

Summary

In short, Tableau provides powerful features that help people analyze and visualize data simply and effectively, allowing them to make better decisions based on that data.



Storytelling in Tableau is about using visualizations, dashboards, and notes to tell a clear and engaging story with data. It helps users share insights in a way that’s both structured and interesting. Here’s how you can achieve effective storytelling in Tableau:

Key Elements of Storytelling in Tableau

1. Flow and Structure: Start by identifying the main points of your data story. Arrange these points in a logical order, leading your audience from an introduction to the main insights and finally to actionable conclusions.

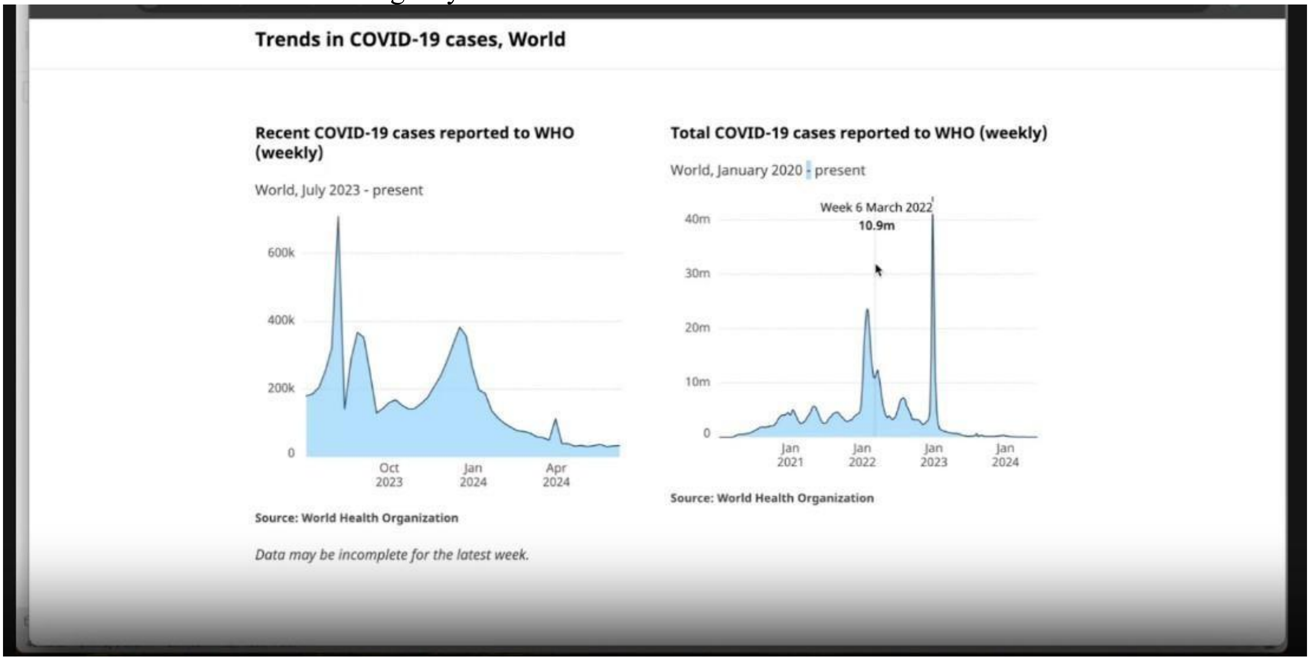
2. Storyboard Feature: Tableau has a storyboard feature that lets you create a series of related dashboards. You can organize these in a linear layout to present a cohesive narrative, helping guide the audience through the information.

3. Annotations and Commentary: You can add notes, callouts, and comments to your dashboards and visualizations. This might include text, images, or shapes that give extra context, highlight important findings, or explain the significance of the data.

4. Interactive Presentations: Tableau’s interactive features allow the audience to explore the data themselves during the presentation. This engagement helps them understand the story more deeply by interacting with the visualizations.

5. Incorporating Multimedia: You can include multimedia elements like images, videos, and web links in your Tableau presentations. This enriches the data story by adding context or background information that supports your points.

Summary
By using Tableau's features for visualization, interaction, and storytelling, you can create compelling data narratives. This approach helps inform and persuade your audience, empowering them to make better decisions based on the insights you share.



Trends in COVID-19 cases, World

Recent COVID-19 cases reported to WHO (weekly)

World, July 2023 - present

Source: World Health Organization

Data may be incomplete for the latest week.

Total COVID-19 cases reported to WHO (weekly)

World, January 2020 - present

Week 6 March 2022
10.9m

Source: World Health Organization

Key Points about Using Geography in Tableau.

1. **Show Spatial Relationships: By using maps, you can clearly illustrate how data varies by location. This helps highlight regional differences and trends.**

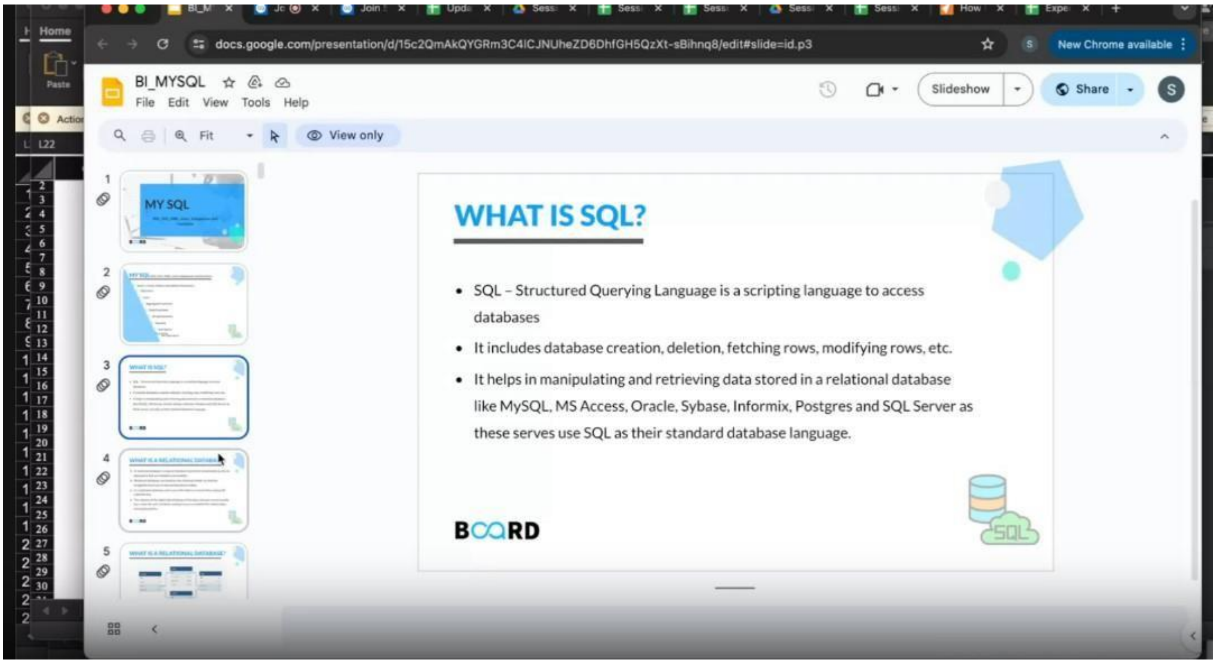
2. **Enhance Your Story: Tableau has strong mapping tools that let you create interactive maps. These maps can make your data story more engaging and help your audience understand the context better.**

3. **Visualize with Maps: You can integrate geographical data easily in Tableau. Creating maps allows you to focus on location-specific insights, making patterns more apparent.**

4. **Support Decision-Making: Understanding geographical context is crucial for strategic planning. By showing data on a map, you can provide valuable insights that guide important decisions.**

Module 3: **Advance SQL**

Certainly! Let's break down advanced SQL into more straightforward terms, focusing on its usage, future, job prospects, relevance in data science, the companies that use it, and its key functions.


docs.google.com/presentation/d/15c2QmAkQYGRm3C4ICJNUheZD6DhfGH5QzXt-sBihnq8/edit#slide=id.p3
New Chrome available
BI_MYSQL
File Edit View Tools Help
Slideshow
Share
Fit
View only
MY SQL

WHAT IS SQL?

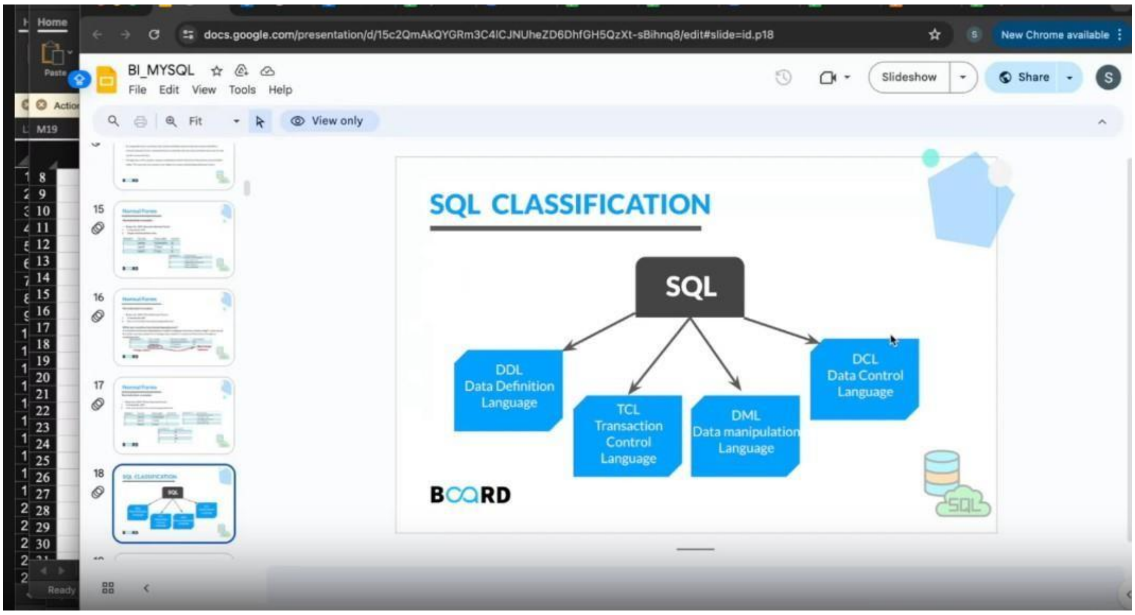
- SQL – Structured Querying Language is a scripting language to access databases
- It includes database creation, deletion, fetching rows, modifying rows, etc.
- It helps in manipulating and retrieving data stored in a relational database like MySQL, MS Access, Oracle, Sybase, Informix, Postgres and SQL Server as these serves use SQL as their standard database language.

BOARD

Usage of Advanced SQL:

Advanced SQL is all about doing more complex tasks with databases. Here are some key areas:

1. Complex Queries: You can mix multiple conditions and rules to get specific data.
2. Joins: Combining data from different tables to find related information.
3. Window Functions: Powerful tools that allow you to perform calculations over sets of rows, like calculating running totals or rankings without losing individual row context.
4. Common Table Expressions (CTEs): Temporary result sets that can be used within queries to simplify complex logic.
5. Stored Procedures and Triggers: Pre-written SQL commands that can automate tasks like updating data automatically when certain conditions are met.



SQL CLASSIFICATION

SQL

DDL Data Definition Language

TCL Transaction Control Language

DML Data manipulation Language

DCL Data Control Language

docs.google.com/presentation/d/15c2QmAkQYGRm3C4ICJNUheZD6DhfGH5QzXt-sBihnq8/edit#slide=id.p18
New Chrome available
BI_MYSQL
File Edit View Tools Help
Slideshow
Share
Fit
View only
BOARD
SQL

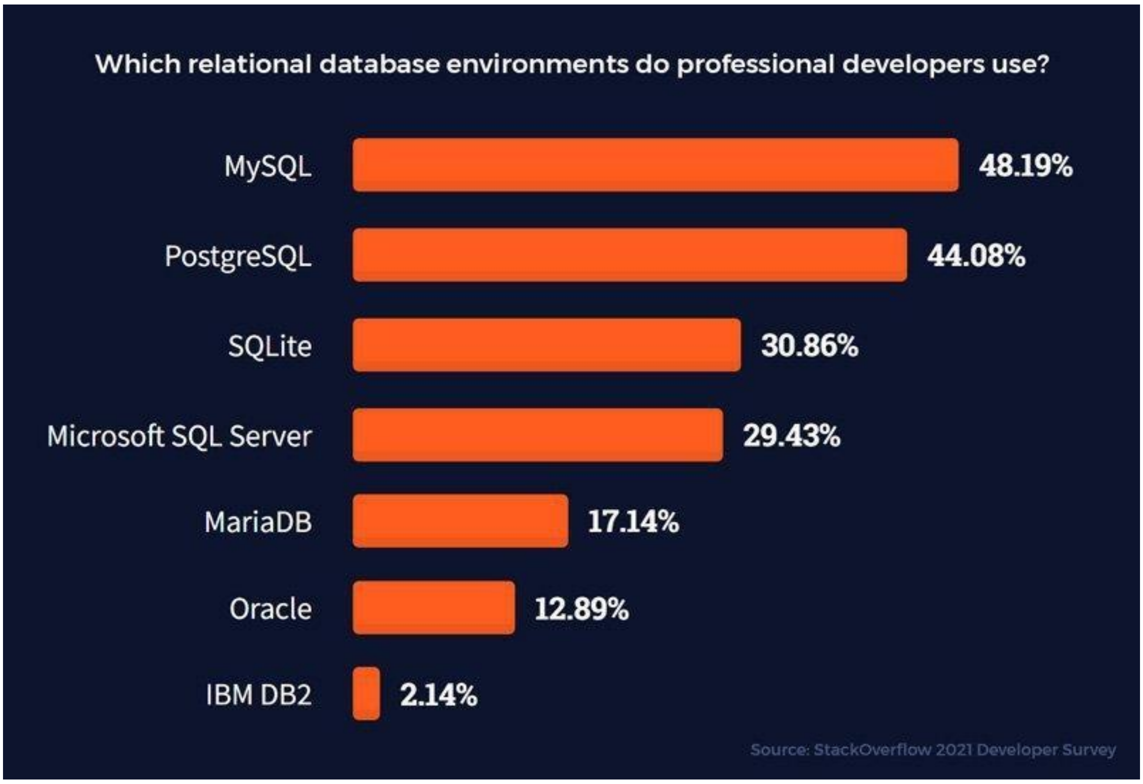
Future of Advanced SQL:

Advanced SQL is set to even more important. As data gets larger and more complex, companies will need advanced SQL to:

1. Work with big data and grow real-time analytics.

2. Integrate with machine learning tools.

3. Use in cloud systems, as more businesses move their data online.



Which relational database environments do professional developers use?

Source: StackOverflow 2021 Developer Survey

Current Job Opportunities:

Knowing advanced SQL opens the door to many career options, including:

1. SQL Developer: Writing and managing SQL code.
2. Database Administrator (DBA): Overseeing and maintaining databases.
3. Data Engineer: Building systems for data processing and analysis.
4. Data Scientist: Using data to create models and insights.

Role in Data Science:

In data science, advanced SQL helps with:

1. Data Transformation: Changing raw data into a format that's useful for analysis.
2. Data Extraction: Pulling data from databases for analysis or modeling.
3. Integration with Other Languages: Working alongside languages like Python or R for deeper analysis.

Companies Using SQL:

Many big companies rely on advanced SQL, such as:

1. Tech Giants: Google, Facebook, Amazon, and Netflix.
2. Financial Institutions: Banks and investment firms.
3. Healthcare Providers: Hospitals and clinics.

Functions and Properties:

Advanced SQL has many features that enhance its power:

1. Window Functions: For advanced calculations like rankings or moving averages.
2. Recursive Queries: Handling data that has a hierarchy, like employee management structures.
3. Performance Optimization Tools: Such as indexing, which helps data retrieval speed.

In summary, mastering advanced SQL equips you to handle complex data challenges effectively. It's a highly valued skill in the job market, especially within data-centric roles. As businesses harness more data, staying updated with advanced SQL techniques ensures that you are ready to meet modern data management demands.

Module 4: Programming for Data Science PYTHON

Definition of python:

Python is a popular programming language that is easy to read and write, making it great for many tasks. People use it for tasks like:

1. Building Websites: Creating web applications.
2. Automation: Automating repetitive tasks.
3. Scientific Work: Doing calculations and data analysis.
4. Data Science and AI: Analyzing data, making predictions, and training machines to learn from data.

Why Python is Important for Data Science

1. Easy to Learn: Python is simple for beginners to understand.
2. Versatile: You can use Python in different areas, like web development or data science.
3. Strong Community: There are many people using Python, so there are lots of resources and support available.
4. Great Libraries: Python has many libraries (pre-written code that you can use) that are perfect for data tasks, such as NumPy for calculations and pandas for handling data.

Job Opportunities with Python

Knowing Python opens up many job roles, including:

1. Data Scientist: Someone who analyzes data to find insights.
2. Machine Learning Engineer: Builds systems that learn from data.
3. Data Analyst: Looks at data and provides reports.
4. Python Developer: Writes software using Python.
5. AI Specialist: Works specifically on artificial intelligence projects.

Lots of big companies like Google, Facebook, Amazon, and Netflix look for people who know Python because they need it for their data projects.

What Python Can Do for Data Science

1. Cleaning: Python can help clean messy data by:

    1. Fixing missing Data values.
    2. Removing duplicate entries.
    3. Arranging data into a standard format.
2. Data Analysis: With libraries like pandas and NumPy, Python can:
    1. Analyze data and perform calculations.
    2. Create charts and graphs to visualize information.
3. Machine Learning: Python is great for building predictive models using libraries like scikit-learn and TensorFlow. It helps machines learn patterns from data.
4. Web Scraping: You can use Python to gather data from websites with tools like BeautifulSoup, allowing you to collect information for analysis.
5. Data Integration: Python can combine data from different sources and get it ready for analysis using ETL processes (Extract, Transform, Load).

Key Libraries in Python for Data Science

1. Pandas: For managing and analyzing data easily.
2. NumPy: For fast numerical calculations and handling arrays (like lists of numbers).
3. Matplotlib and Seaborn: For making visual representations like charts and graphs.
4. Scikit-learn: For various machine learning tasks, including classification and regression.
5. TensorFlow and Keras: For building deep learning models.
6. NLTK: For natural language processing (analyzing human language).

Conclusion

By learning Python, you gain a powerful tool for handling, analyzing, and drawing insights from data. It opens up many career paths and is essential in today's data-driven world.

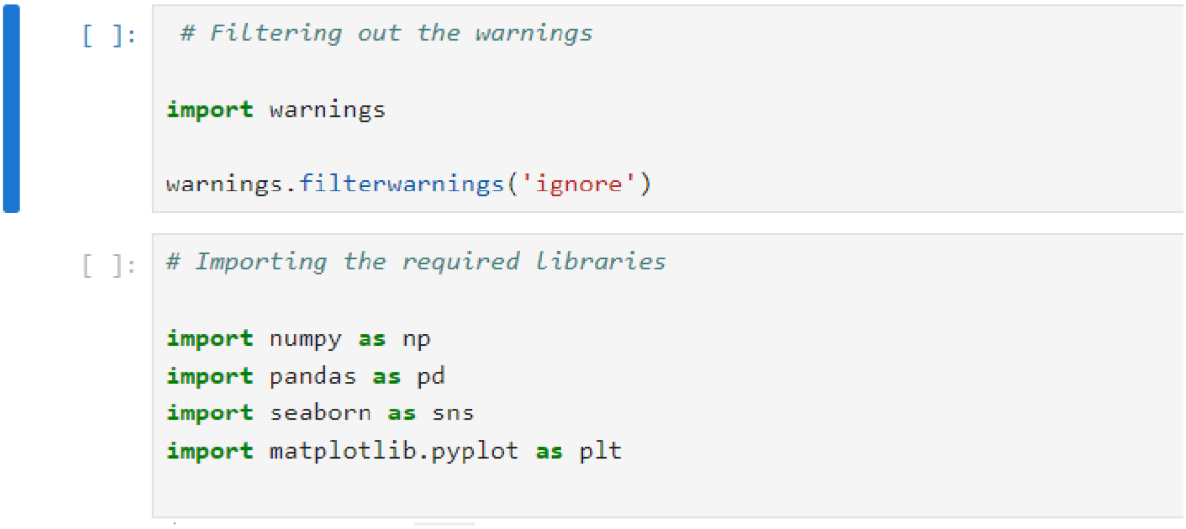
Module 5:Project

Assignment: Problem Statement:

You have the data for the 100 top-rated movies from the past decade along with various pieces of information about the movie, its actors, and the voters who have rated these movies online.

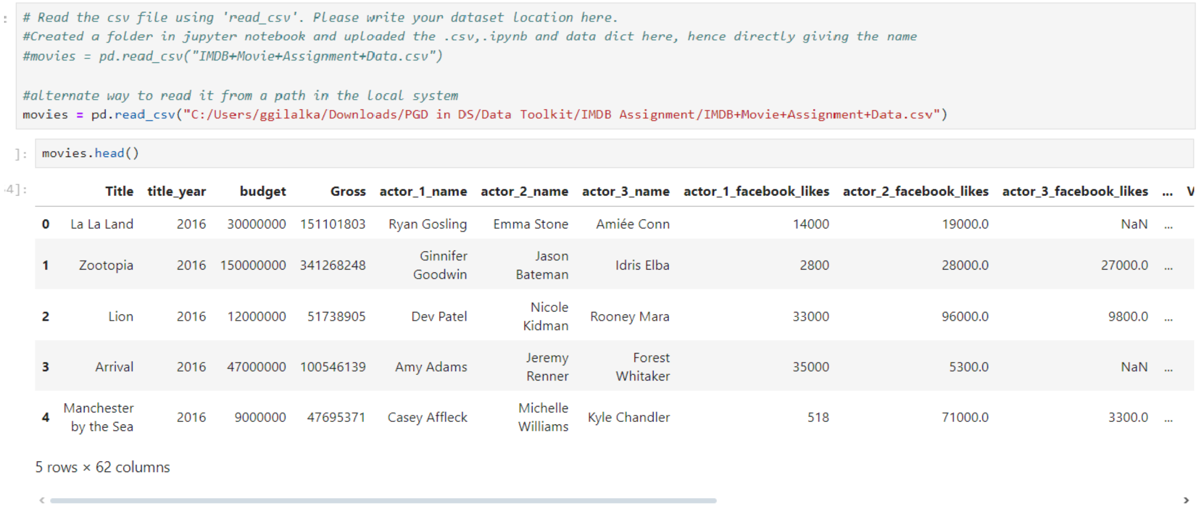
Criterion	Meets expectations	Does not meet expectations
Task 1 (~5%)	The commands are syntactically correct. The output of the code is correct in terms of the question and format. The data frame has been thoroughly inspected using the taught commands.	There are minor syntax errors in the code. The dataframe hasn't been thoroughly inspected before moving on to the next section.
Task 2 (~40%)	The commands are syntactically correct. The output of the code is correct in terms of the question and format. A new dataframe is created wherever it is asked to do so. In the case of dataframes, the results contain the same rows and columns as expected. Regarding plots, making appropriate charts with the mentioned libraries and getting the right trends. Writing clear and concise inferences for the charts wherever asked	There are minor syntax errors in the code. The functions/arguments used are only partially correct. After performing the operations for a subtask, the final result is not imported into the new said dataframe (if asked). In the case of the dataframes, the results either contain unnecessary rows/columns or miss the required ones. Using chart types which are not suitable for the required observations and wrong trends. Unclear and incorrect observations.
Task(~50%)	The commands are	There are minor syntax

	syntactically correct. The output of the code is correct in terms of the question and format. A new dataframe is created wherever it is asked to do so. In the case of dataframes, the results contain the same rows and columns as expected. Regarding plots, making appropriate charts with the mentioned libraries and getting the right trends. Writing clear and concise inferences for the charts wherever asked.	errors in the code. The functions/arguments used are only partially correct. After performing the operations for a subtask, the final result is not imported into the new said dataframe (if asked). In the case of the dataframes, the results either contain unnecessary rows/columns or miss the required ones. Using chart types which are not suitable for the required observations and wrong trends. Unclear and incorrect observations.
Adherence to coding guidelines (~5%)	The code is concise. Wherever appropriate, built-in functions are used instead of making the code longer (if-else statements, for loops,loc/iloc). If new variables are created, the names are descriptive and unambiguous. Following the variable/dataframe names mentioned in the question wherever it is provided. The code readability is good, with appropriate indentations. Charts are neatly formatted including proper chart sizes, annotations(if required) and labelling.	Long and complex code is used instead of shorter built-in functions wherever possible. The code readability is poor because of vaguely named variables or a lack of comments wherever necessary. Comments are not written, rendering the code difficult to understand. Unclear charts with no proper scales/legend.



```
[ ]:  # Filtering out the warnings

      import warnings

      warnings.filterwarnings('ignore')
```

```
[ ]:  # Importing the required libraries

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



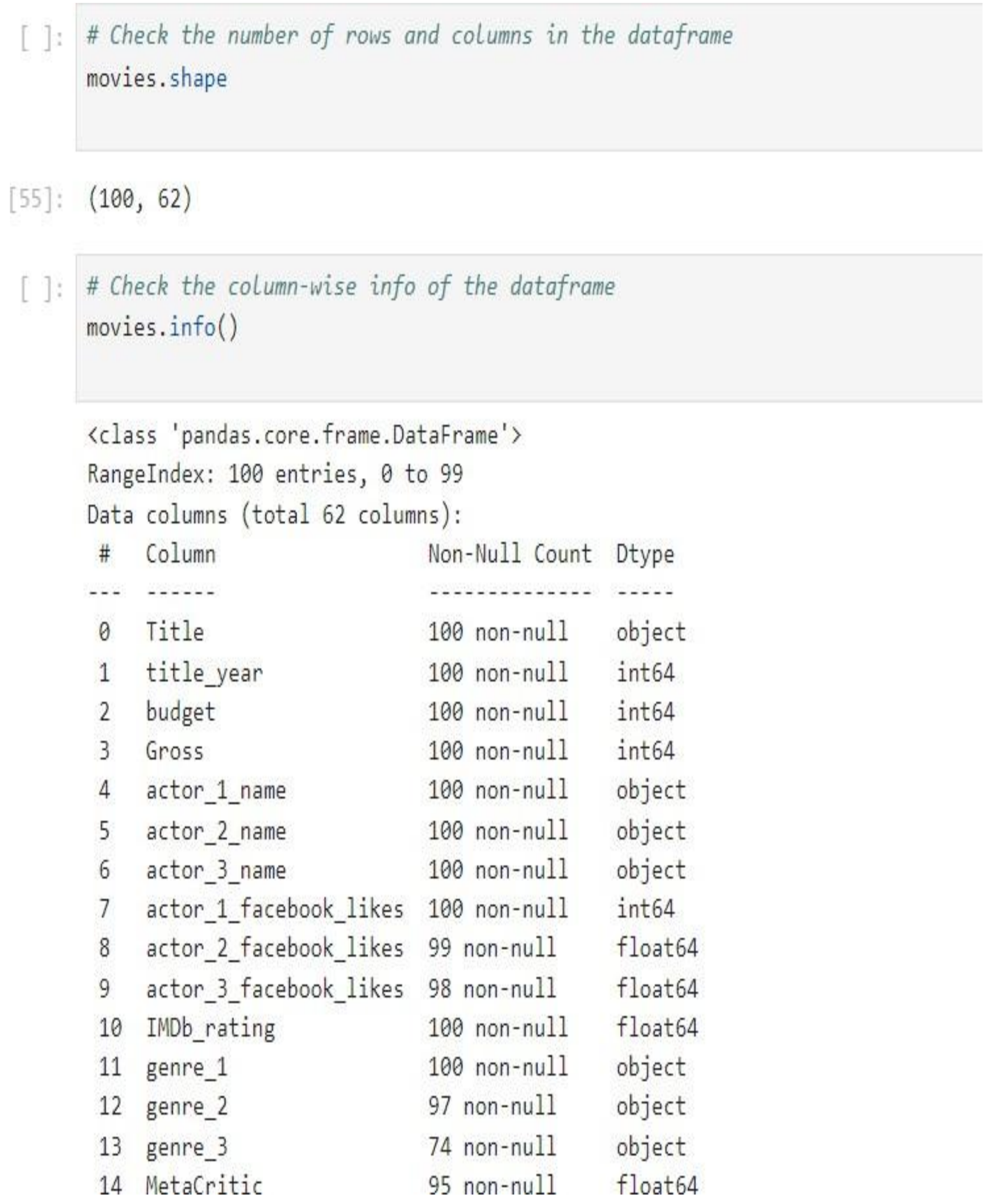
```
:  # Read the csv file using 'read_csv'. Please write your dataset location here.
   #Created a folder in jupyter notebook and uploaded the .csv,.ipynb and data dict here, hence directly giving the name
   #movies = pd.read_csv("IMDB+Movie+Assignment+Data.csv")

   #alternate way to read it from a path in the local system
   movies = pd.read_csv("C:/Users/ggilalka/Downloads/PGD in DS/Data Toolkit/IMDB Assignment/IMDB+Movie+Assignment+Data.csv")
```

```
]:  movies.head()
```

4]:

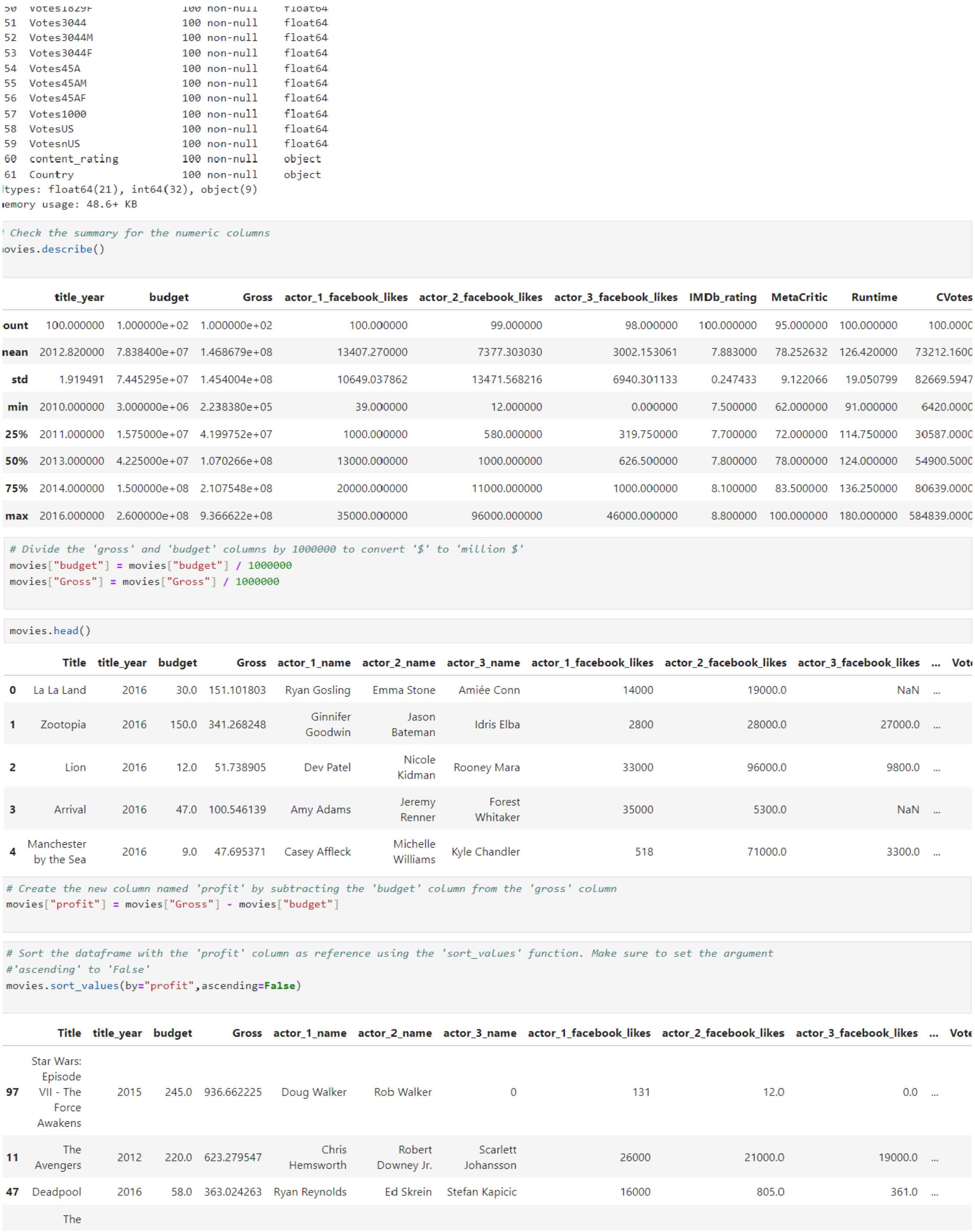
	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	V
0	La La Land	2016	30000000	151101803	Ryan Gosling	Emma Stone	Amiée Conn	14000	19000.0	NaN	...	
1	Zootopia	2016	150000000	341268248	Ginnifer Goodwin	Jason Bateman	Idris Elba	2800	28000.0	27000.0	...	
2	Lion	2016	12000000	51738905	Dev Patel	Nicole Kidman	Rooney Mara	33000	96000.0	9800.0	...	
3	Arrival	2016	47000000	100546139	Amy Adams	Jeremy Renner	Forest Whitaker	35000	5300.0	NaN	...	
4	Manchester by the Sea	2016	9000000	47695371	Casey Affleck	Michelle Williams	Kyle Chandler	518	71000.0	3300.0	...	

5 rows × 62 columns



```
[ ]: # Check the number of rows and columns in the dataframe
     movies.shape

[55]: (100, 62)

[ ]: # Check the column-wise info of the dataframe
     movies.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 100 entries, 0 to 99
     Data columns (total 62 columns):
      #   Column                  Non-Null Count  Dtype
     ---  ------                  --------------  -----
      0   Title                   100 non-null    object
      1   title_year              100 non-null    int64
      2   budget                  100 non-null    int64
      3   Gross                   100 non-null    int64
      4   actor_1_name            100 non-null    object
      5   actor_2_name            100 non-null    object
      6   actor_3_name            100 non-null    object
      7   actor_1_facebook_likes  100 non-null    int64
      8   actor_2_facebook_likes  99 non-null     float64
      9   actor_3_facebook_likes  98 non-null     float64
      10  IMDb_rating             100 non-null    float64
      11  genre_1                 100 non-null    object
      12  genre_2                 97 non-null     object
      13  genre_3                 74 non-null     object
      14  MetaCritic              95 non-null     float64
```

```
 15   Runtime                 100 non-null    int64
 16   CVotes10                100 non-null    int64
 17   CVotes09                100 non-null    int64
 18   CVotes08                100 non-null    int64
 19   CVotes07                100 non-null    int64
 20   CVotes06                100 non-null    int64
 21   CVotes05                100 non-null    int64
 22   CVotes04                100 non-null    int64
 23   CVotes03                100 non-null    int64
 24   CVotes02                100 non-null    int64
 25   CVotes01                100 non-null    int64
 26   CVotesMale              100 non-null    int64
 27   CVotesFemale            100 non-null    int64
 28   CVotesU18               100 non-null    int64
 29   CVotesU18M              100 non-null    int64
 30   CVotesU18F              100 non-null    int64
 31   CVotes1829              100 non-null    int64
 32   CVotes1829M             100 non-null    int64
 33   CVotes1829F             100 non-null    int64
 34   CVotes3044              100 non-null    int64
 35   CVotes3044M             100 non-null    int64
 36   CVotes3044F             100 non-null    int64
 37   CVotes45A               100 non-null    int64
 38   CVotes45AM              100 non-null    int64
 39   CVotes45AF              100 non-null    int64
 40   CVotes1000              100 non-null    int64
 41   CVotesUS                100 non-null    int64
 42   CVotesnUS               100 non-null    int64
 43   VotesM                  100 non-null    float64
 44   VotesF                  100 non-null    float64
 45   VotesU18                100 non-null    float64
 46   VotesU18M               100 non-null    float64
 47   VotesU18F               100 non-null    float64
 48   Votes1829               100 non-null    float64
 49   Votes1829M              100 non-null    float64
```



```
50  Votes1829F             100 non-null   float64
51  Votes3044              100 non-null   float64
52  Votes3044M             100 non-null   float64
53  Votes3044F             100 non-null   float64
54  Votes45A               100 non-null   float64
55  Votes45AM              100 non-null   float64
56  Votes45AF              100 non-null   float64
57  Votes1000              100 non-null   float64
58  VotesUS                100 non-null   float64
59  VotesnUS               100 non-null   float64
60  content_rating         100 non-null   object
61  Country                100 non-null   object
ltypes: float64(21), int64(32), object(9)
ıemory usage: 48.6+ KB
```

```
: Check the summary for the numeric columns
ıovies.describe()
```

	title_year	budget	Gross	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	IMDb_rating	MetaCritic	Runtime	CVotes
ount	100.000000	1.000000e+02	1.000000e+02	100.000000	99.000000	98.000000	100.000000	95.000000	100.000000	100.0000
nean	2012.820000	7.838400e+07	1.468679e+08	13407.270000	7377.303030	3002.153061	7.883000	78.252632	126.420000	73212.1600
std	1.919491	7.445295e+07	1.454004e+08	10649.037862	13471.568216	6940.301133	0.247433	9.122066	19.050799	82669.5947
min	2010.000000	3.000000e+06	2.238380e+05	39.000000	12.000000	0.000000	7.500000	62.000000	91.000000	6420.0000
25%	2011.000000	1.575000e+07	4.199752e+07	1000.000000	580.000000	319.750000	7.700000	72.000000	114.750000	30587.0000
50%	2013.000000	4.225000e+07	1.070266e+08	13000.000000	1000.000000	626.500000	7.800000	78.000000	124.000000	54900.5000
75%	2014.000000	1.500000e+08	2.107548e+08	20000.000000	11000.000000	1000.000000	8.100000	83.500000	136.250000	80639.0000
max	2016.000000	2.600000e+08	9.366622e+08	35000.000000	96000.000000	46000.000000	8.800000	100.000000	180.000000	584839.0000

```
# Divide the 'gross' and 'budget' columns by 1000000 to convert '$' to 'million $'
movies["budget"] = movies["budget"] / 1000000
movies["Gross"] = movies["Gross"] / 1000000
```

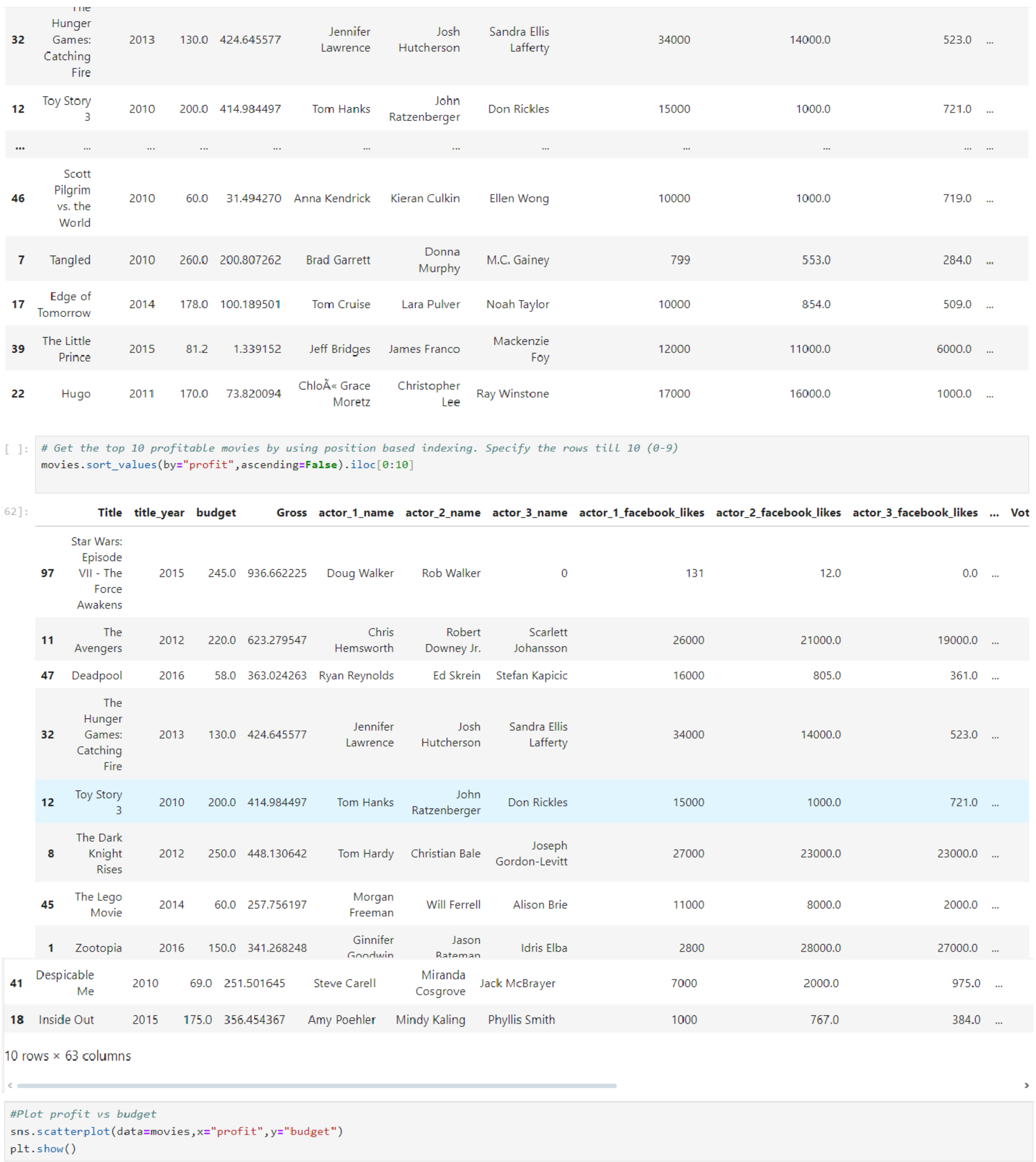
```
movies.head()
```

	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	Vot
0	La La Land	2016	30.0	151.101803	Ryan Gosling	Emma Stone	Amiée Conn	14000	19000.0	NaN	...	
1	Zootopia	2016	150.0	341.268248	Ginnifer Goodwin	Jason Bateman	Idris Elba	2800	28000.0	27000.0	...	
2	Lion	2016	12.0	51.738905	Dev Patel	Nicole Kidman	Rooney Mara	33000	96000.0	9800.0	...	
3	Arrival	2016	47.0	100.546139	Amy Adams	Jeremy Renner	Forest Whitaker	35000	5300.0	NaN	...	
4	Manchester by the Sea	2016	9.0	47.695371	Casey Affleck	Michelle Williams	Kyle Chandler	518	71000.0	3300.0	...	

```
# Create the new column named 'profit' by subtracting the 'budget' column from the 'gross' column
movies["profit"] = movies["Gross"] - movies["budget"]
```

```
# Sort the dataframe with the 'profit' column as reference using the 'sort_values' function. Make sure to set the argument
#'ascending' to 'False'
movies.sort_values(by="profit",ascending=False)
```

	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	Vote
97	Star Wars: Episode VII - The Force Awakens	2015	245.0	936.662225	Doug Walker	Rob Walker	0	131	12.0	0.0	...	
11	The Avengers	2012	220.0	623.279547	Chris Hemsworth	Robert Downey Jr.	Scarlett Johansson	26000	21000.0	19000.0	...	
47	Deadpool	2016	58.0	363.024263	Ryan Reynolds	Ed Skrein	Stefan Kapicic	16000	805.0	361.0	...	
	The											



32	The Hunger Games: Catching Fire	2013	130.0	424.645577	Jennifer Lawrence	Josh Hutcherson	Sandra Ellis Lafferty	34000	14000.0	523.0	...
12	Toy Story 3	2010	200.0	414.984497	Tom Hanks	John Ratzenberger	Don Rickles	15000	1000.0	721.0	...
...	...	...	...	...	...	...	...	...	...	...	...
46	Scott Pilgrim vs. the World	2010	60.0	31.494270	Anna Kendrick	Kieran Culkin	Ellen Wong	10000	1000.0	719.0	...
7	Tangled	2010	260.0	200.807262	Brad Garrett	Donna Murphy	M.C. Gainey	799	553.0	284.0	...
17	Edge of Tomorrow	2014	178.0	100.189501	Tom Cruise	Lara Pulver	Noah Taylor	10000	854.0	509.0	...
39	The Little Prince	2015	81.2	1.339152	Jeff Bridges	James Franco	Mackenzie Foy	12000	11000.0	6000.0	...
22	Hugo	2011	170.0	73.820094	ChloÃ« Grace Moretz	Christopher Lee	Ray Winstone	17000	16000.0	1000.0	...

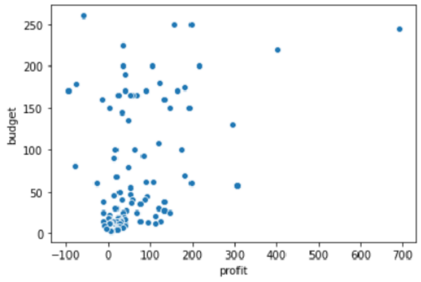
```
[ ]: # Get the top 10 profitable movies by using position based indexing. Specify the rows till 10 (0-9)
     movies.sort_values(by="profit",ascending=False).iloc[0:10]
```

62]:

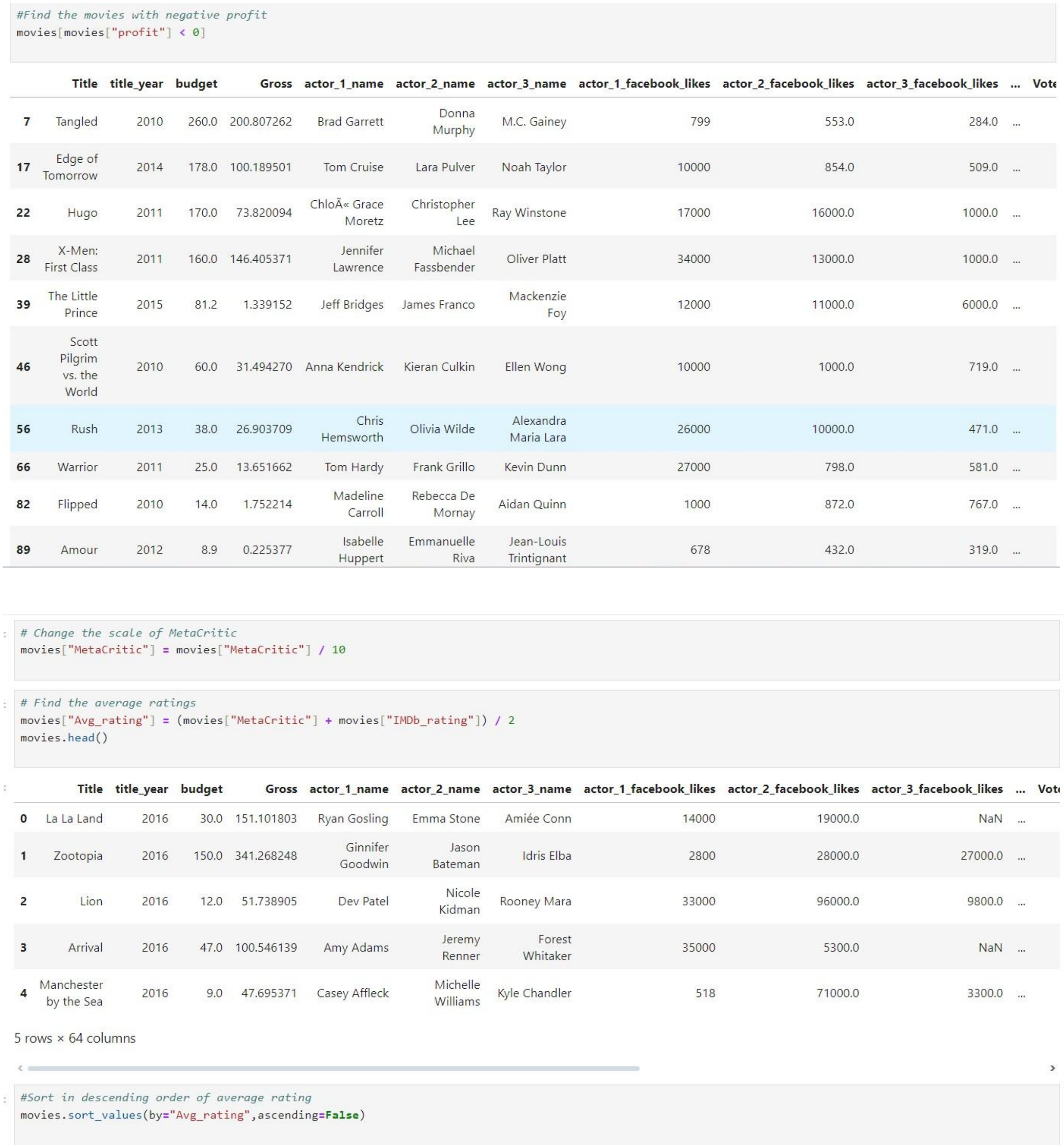
	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	Vot
97	Star Wars: Episode VII - The Force Awakens	2015	245.0	936.662225	Doug Walker	Rob Walker	0	131	12.0	0.0	...	
11	The Avengers	2012	220.0	623.279547	Chris Hemsworth	Robert Downey Jr.	Scarlett Johansson	26000	21000.0	19000.0	...	
47	Deadpool	2016	58.0	363.024263	Ryan Reynolds	Ed Skrein	Stefan Kapicic	16000	805.0	361.0	...	
32	The Hunger Games: Catching Fire	2013	130.0	424.645577	Jennifer Lawrence	Josh Hutcherson	Sandra Ellis Lafferty	34000	14000.0	523.0	...	
12	Toy Story 3	2010	200.0	414.984497	Tom Hanks	John Ratzenberger	Don Rickles	15000	1000.0	721.0	...	
8	The Dark Knight Rises	2012	250.0	448.130642	Tom Hardy	Christian Bale	Joseph Gordon-Levitt	27000	23000.0	23000.0	...	
45	The Lego Movie	2014	60.0	257.756197	Morgan Freeman	Will Ferrell	Alison Brie	11000	8000.0	2000.0	...	
1	Zootopia	2016	150.0	341.268248	Ginnifer Goodwin	Jason Bateman	Idris Elba	2800	28000.0	27000.0	...	
41	Despicable Me	2010	69.0	251.501645	Steve Carell	Miranda Cosgrove	Jack McBrayer	7000	2000.0	975.0	...	
18	Inside Out	2015	175.0	356.454367	Amy Poehler	Mindy Kaling	Phyllis Smith	1000	767.0	384.0	...	

10 rows × 63 columns

```
#Plot profit vs budget
sns.scatterplot(data=movies,x="profit",y="budget")
plt.show()
```



The dataset contains the 100 best performing movies from the year 2010 to 2016. However scatter plot tells a different story. You can notice that there are some movies with negative profit. Although good movies do incur losses, but there appear to be quite a few movie with losses. What can be the reason behind this? Lets have a closer look at this by finding the movies with negative profit.



```
#Find the movies with negative profit
movies[movies["profit"] < 0]
```

	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	Vote
7	Tangled	2010	260.0	200.807262	Brad Garrett	Donna Murphy	M.C. Gainey	799	553.0	284.0	...	
17	Edge of Tomorrow	2014	178.0	100.189501	Tom Cruise	Lara Pulver	Noah Taylor	10000	854.0	509.0	...	
22	Hugo	2011	170.0	73.820094	ChloÃ« Grace Moretz	Christopher Lee	Ray Winstone	17000	16000.0	1000.0	...	
28	X-Men: First Class	2011	160.0	146.405371	Jennifer Lawrence	Michael Fassbender	Oliver Platt	34000	13000.0	1000.0	...	
39	The Little Prince	2015	81.2	1.339152	Jeff Bridges	James Franco	Mackenzie Foy	12000	11000.0	6000.0	...	
46	Scott Pilgrim vs. the World	2010	60.0	31.494270	Anna Kendrick	Kieran Culkin	Ellen Wong	10000	1000.0	719.0	...	
56	Rush	2013	38.0	26.903709	Chris Hemsworth	Olivia Wilde	Alexandra Maria Lara	26000	10000.0	471.0	...	
66	Warrior	2011	25.0	13.651662	Tom Hardy	Frank Grillo	Kevin Dunn	27000	798.0	581.0	...	
82	Flipped	2010	14.0	1.752214	Madeline Carroll	Rebecca De Mornay	Aidan Quinn	1000	872.0	767.0	...	
89	Amour	2012	8.9	0.225377	Isabelle Huppert	Emmanuelle Riva	Jean-Louis Trintignant	678	432.0	319.0	...	

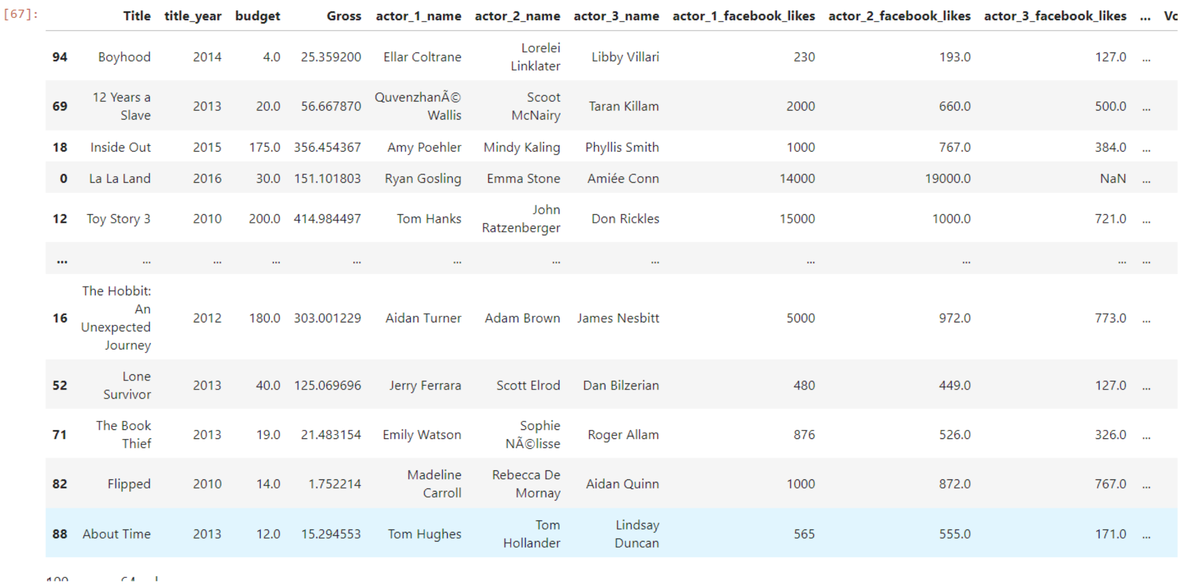
```
# Change the scale of MetaCritic
movies["MetaCritic"] = movies["MetaCritic"] / 10
```

```
# Find the average ratings
movies["Avg_rating"] = (movies["MetaCritic"] + movies["IMDb_rating"]) / 2
movies.head()
```

	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	Vot
0	La La Land	2016	30.0	151.101803	Ryan Gosling	Emma Stone	Amiée Conn	14000	19000.0	NaN	...	
1	Zootopia	2016	150.0	341.268248	Ginnifer Goodwin	Jason Bateman	Idris Elba	2800	28000.0	27000.0	...	
2	Lion	2016	12.0	51.738905	Dev Patel	Nicole Kidman	Rooney Mara	33000	96000.0	9800.0	...	
3	Arrival	2016	47.0	100.546139	Amy Adams	Jeremy Renner	Forest Whitaker	35000	5300.0	NaN	...	
4	Manchester by the Sea	2016	9.0	47.695371	Casey Affleck	Michelle Williams	Kyle Chandler	518	71000.0	3300.0	...	

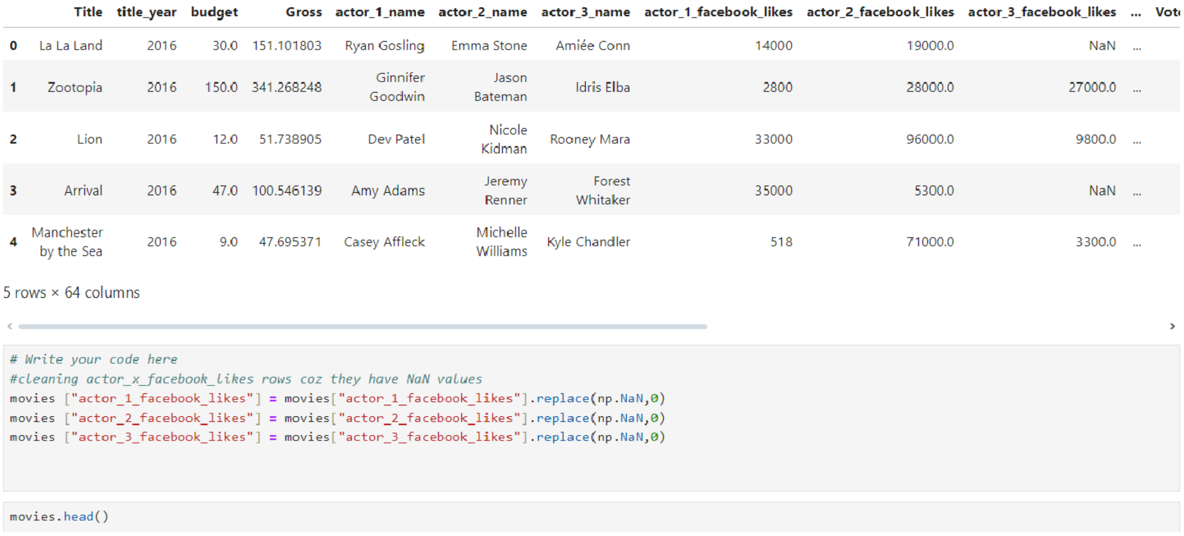
5 rows × 64 columns

```
#Sort in descending order of average rating
movies.sort_values(by="Avg_rating",ascending=False)
```



[67]:

	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	Vc
94	Boyhood	2014	4.0	25.359200	Ellar Coltrane	Lorelei Linklater	Libby Villari	230	193.0	127.0	...	
69	12 Years a Slave	2013	20.0	56.667870	QuvenzhanÃ© Wallis	Scoot McNairy	Taran Killam	2000	660.0	500.0	...	
18	Inside Out	2015	175.0	356.454367	Amy Poehler	Mindy Kaling	Phyllis Smith	1000	767.0	384.0	...	
0	La La Land	2016	30.0	151.101803	Ryan Gosling	Emma Stone	Amiée Conn	14000	19000.0	NaN	...	
12	Toy Story 3	2010	200.0	414.984497	Tom Hanks	John Ratzenberger	Don Rickles	15000	1000.0	721.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
16	The Hobbit: An Unexpected Journey	2012	180.0	303.001229	Aidan Turner	Adam Brown	James Nesbitt	5000	972.0	773.0	...	
52	Lone Survivor	2013	40.0	125.069696	Jerry Ferrara	Scott Elrod	Dan Bilzerian	480	449.0	127.0	...	
71	The Book Thief	2013	19.0	21.483154	Emily Watson	Sophie NÃ©lisse	Roger Allam	876	526.0	326.0	...	
82	Flipped	2010	14.0	1.752214	Madeline Carroll	Rebecca De Mornay	Aidan Quinn	1000	872.0	767.0	...	
88	About Time	2013	12.0	15.294553	Tom Hughes	Tom Hollander	Lindsay Duncan	565	555.0	171.0	...	

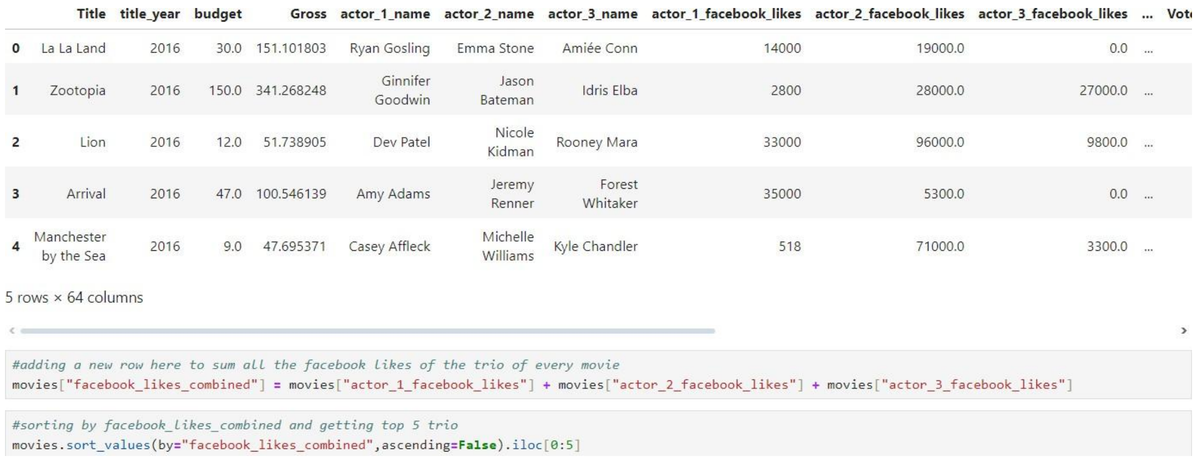


	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	Vote
0	La La Land	2016	30.0	151.101803	Ryan Gosling	Emma Stone	Amiée Conn	14000	19000.0	NaN	...	
1	Zootopia	2016	150.0	341.268248	Ginnifer Goodwin	Jason Bateman	Idris Elba	2800	28000.0	27000.0	...	
2	Lion	2016	12.0	51.738905	Dev Patel	Nicole Kidman	Rooney Mara	33000	96000.0	9800.0	...	
3	Arrival	2016	47.0	100.546139	Amy Adams	Jeremy Renner	Forest Whitaker	35000	5300.0	NaN	...	
4	Manchester by the Sea	2016	9.0	47.695371	Casey Affleck	Michelle Williams	Kyle Chandler	518	71000.0	3300.0	...	

5 rows × 64 columns

```
# Write your code here
#cleaning actor_x_facebook_likes rows coz they have NaN values
movies ["actor_1_facebook_likes"] = movies["actor_1_facebook_likes"].replace(np.NaN,0)
movies ["actor_2_facebook_likes"] = movies["actor_2_facebook_likes"].replace(np.NaN,0)
movies ["actor_3_facebook_likes"] = movies["actor_3_facebook_likes"].replace(np.NaN,0)
```

```
movies.head()
```



	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	Vote
0	La La Land	2016	30.0	151.101803	Ryan Gosling	Emma Stone	Amiée Conn	14000	19000.0	0.0	...	
1	Zootopia	2016	150.0	341.268248	Ginnifer Goodwin	Jason Bateman	Idris Elba	2800	28000.0	27000.0	...	
2	Lion	2016	12.0	51.738905	Dev Patel	Nicole Kidman	Rooney Mara	33000	96000.0	9800.0	...	
3	Arrival	2016	47.0	100.546139	Amy Adams	Jeremy Renner	Forest Whitaker	35000	5300.0	0.0	...	
4	Manchester by the Sea	2016	9.0	47.695371	Casey Affleck	Michelle Williams	Kyle Chandler	518	71000.0	3300.0	...	

5 rows × 64 columns

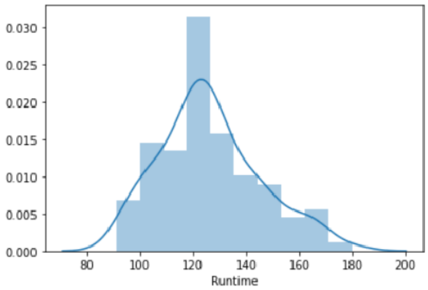
```
#adding a new row here to sum all the facebook likes of the trio of every movie
movies["facebook_likes_combined"] = movies["actor_1_facebook_likes"] + movies["actor_2_facebook_likes"] + movies["actor_3_facebook_likes"]
```

```
#sorting by facebook_likes_combined and getting top 5 trio
movies.sort_values(by="facebook_likes_combined",ascending=False).iloc[0:5]
```

	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	V
2	Lion	2016	12.0	51.738905	Dev Patel	Nicole Kidman	Rooney Mara	33000	96000.0	9800.0	...	
27	Inception	2010	160.0	292.568851	Leonardo DiCaprio	Tom Hardy	Joseph Gordon-Levitt	29000	27000.0	23000.0	...	
14	X-Men: Days of Future Past	2014	200.0	233.914986	Jennifer Lawrence	Peter Dinklage	Hugh Jackman	34000	22000.0	20000.0	...	
4	Manchester by the Sea	2016	9.0	47.695371	Casey Affleck	Michelle Williams	Kyle Chandler	518	71000.0	3300.0	...	
8	The Dark Knight Rises	2012	250.0	448.130642	Tom Hardy	Christian Bale	Joseph Gordon-Levitt	27000	23000.0	23000.0	...	

5 rows × 65 columns

```
# Runtime histogram/density plot
sns.distplot(movies["Runtime"])
plt.show()
```



`PopularR`.

```
# Write your code here
PopularR = movies[(movies["content_rating"] == "R") & (movies["CVotesU18"] > 0)].sort_values(by="CVotesU18",ascending=False).iloc[0:10]
```

```
# Create the dataframe df_by_genre
df_by_genre = movies.iloc[0:,11:14].join(movies.iloc[0:,16:60])
```

```
# Create a column cnt and initialize it to 1
df_by_genre["cnt"] = 1
```

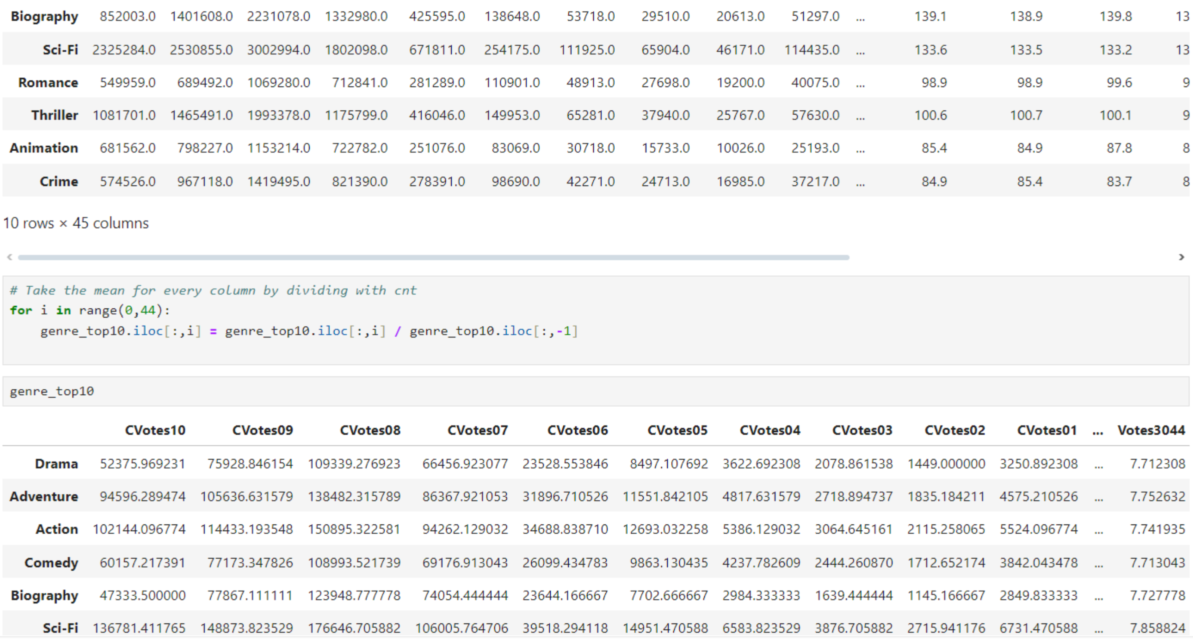
```
# Group the movies by individual genres
df_by_g1 = df_by_genre.groupby("genre_1").sum()
df_by_g2 = df_by_genre.groupby("genre_2").sum()
df_by_g3 = df_by_genre.groupby("genre_3").sum()
```

```
# Add the grouped data frames and store it in a new data frame
df_add = df_by_g1.add(df_by_g2,fill_value=0).add(df_by_g3,fill_value=0)
```

```
# Extract genres with atleast 10 occurences
genre_top10 = df_add[df_add["cnt"] > 10].sort_values(by="cnt",ascending=False)
```

```
genre_top10
```

	CVotes10	CVotes09	CVotes08	CVotes07	CVotes06	CVotes05	CVotes04	CVotes03	CVotes02	CVotes01	...	Votes3044	Votes3044M	Votes3044F	Votes4
Drama	3404438.0	4935375.0	7107053.0	4319700.0	1529356.0	552312.0	235475.0	135126.0	94185.0	211308.0	...	501.3	501.1	501.8	49
Adventure	3594659.0	4014192.0	5262328.0	3281981.0	1212075.0	438970.0	183070.0	103318.0	69737.0	173858.0	...	294.6	293.7	299.2	29
Action	3166467.0	3547429.0	4677755.0	2922126.0	1075354.0	393484.0	166970.0	95004.0	65573.0	171247.0	...	240.0	239.5	241.8	23
Comedy	1383616.0	1774987.0	2506851.0	1591069.0	600287.0	226852.0	97469.0	56218.0	39391.0	88367.0	...	177.4	177.4	178.3	17



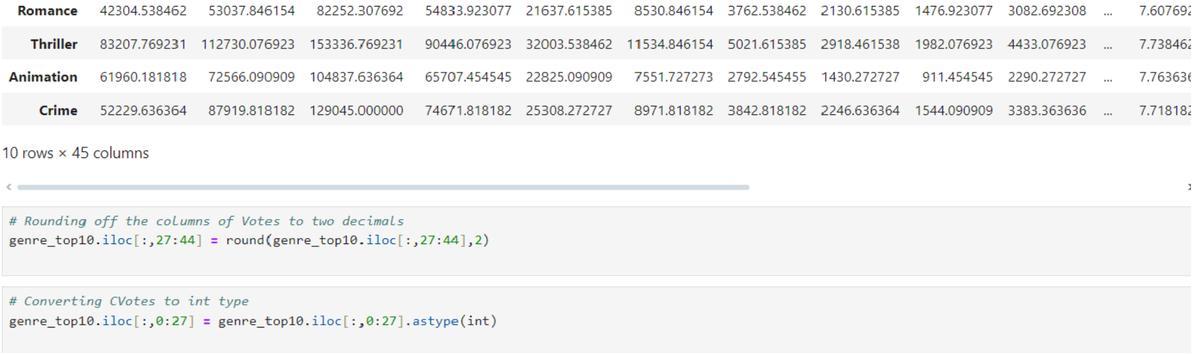
Biography	852003.0	1401608.0	2231078.0	1332980.0	425595.0	138648.0	53718.0	29510.0	20613.0	51297.0	...	139.1	138.9	139.8	13
Sci-Fi	2325284.0	2530855.0	3002994.0	1802098.0	671811.0	254175.0	111925.0	65904.0	46171.0	114435.0	...	133.6	133.5	133.2	13
Romance	549959.0	689492.0	1069280.0	712841.0	281289.0	110901.0	48913.0	27698.0	19200.0	40075.0	...	98.9	98.9	99.6	9
Thriller	1081701.0	1465491.0	1993378.0	1175799.0	416046.0	149953.0	65281.0	37940.0	25767.0	57630.0	...	100.6	100.7	100.1	9
Animation	681562.0	798227.0	1153214.0	722782.0	251076.0	83069.0	30718.0	15733.0	10026.0	25193.0	...	85.4	84.9	87.8	8
Crime	574526.0	967118.0	1419495.0	821390.0	278391.0	98690.0	42271.0	24713.0	16985.0	37217.0	...	84.9	85.4	83.7	8

10 rows × 45 columns

```
# Take the mean for every column by dividing with cnt
for i in range(0,44):
    genre_top10.iloc[:,i] = genre_top10.iloc[:,i] / genre_top10.iloc[:,-1]
```

```
genre_top10
```

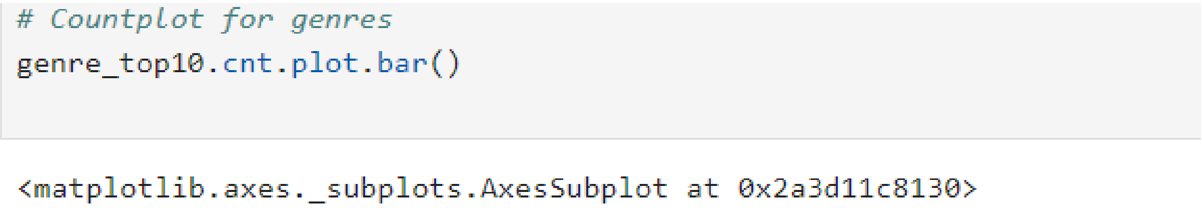
	CVotes10	CVotes09	CVotes08	CVotes07	CVotes06	CVotes05	CVotes04	CVotes03	CVotes02	CVotes01	...	Votes3044
Drama	52375.969231	75928.846154	109339.276923	66456.923077	23528.553846	8497.107692	3622.692308	2078.861538	1449.000000	3250.892308	...	7.712308
Adventure	94596.289474	105636.631579	138482.315789	86367.921053	31896.710526	11551.842105	4817.631579	2718.894737	1835.184211	4575.210526	...	7.752632
Action	102144.096774	114433.193548	150895.322581	94262.129032	34688.838710	12693.032258	5386.129032	3064.645161	2115.258065	5524.096774	...	7.741935
Comedy	60157.217391	77173.347826	108993.521739	69176.913043	26099.434783	9863.130435	4237.782609	2444.260870	1712.652174	3842.043478	...	7.713043
Biography	47333.500000	77867.111111	123948.777778	74054.444444	23644.166667	7702.666667	2984.333333	1639.444444	1145.166667	2849.833333	...	7.727778
Sci-Fi	136781.411765	148873.823529	176646.705882	106005.764706	39518.294118	14951.470588	6583.823529	3876.705882	2715.941176	6731.470588	...	7.858824
Romance	42304.538462	53037.846154	82252.307692	54833.923077	21637.615385	8530.846154	3762.538462	2130.615385	1476.923077	3082.692308	...	7.60769
Thriller	83207.769231	112730.076923	153336.769231	90446.076923	32003.538462	11534.846154	5021.615385	2918.461538	1982.076923	4433.076923	...	7.73846
Animation	61960.181818	72566.090909	104837.636364	65707.454545	22825.090909	7551.727273	2792.545455	1430.272727	911.454545	2290.272727	...	7.76363
Crime	52229.636364	87919.818182	129045.000000	74671.818182	25308.272727	8971.818182	3842.818182	2246.636364	1544.090909	3383.363636	...	7.71818



10 rows × 45 columns

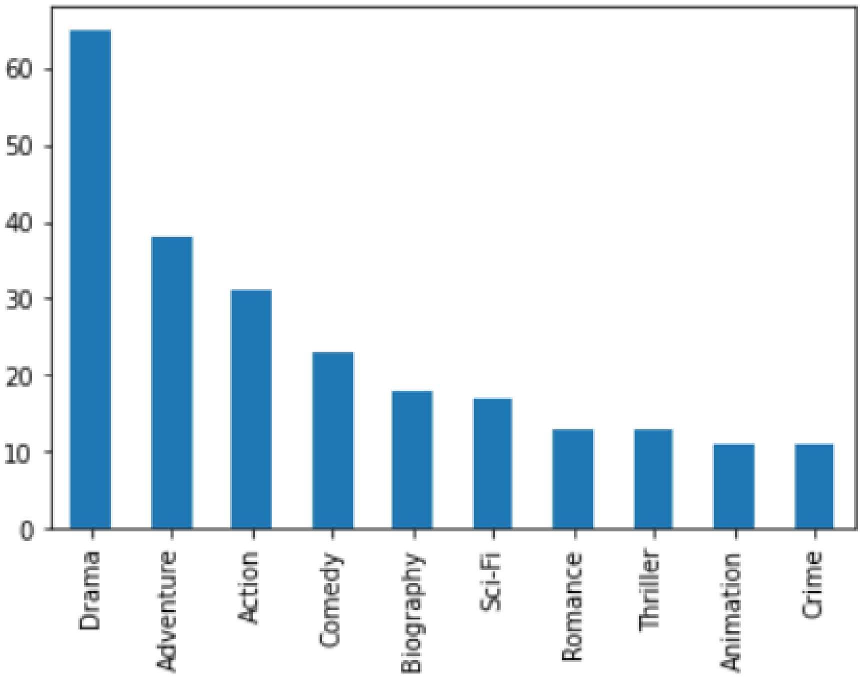
```
# Rounding off the columns of Votes to two decimals
genre_top10.iloc[:,27:44] = round(genre_top10.iloc[:,27:44],2)
```

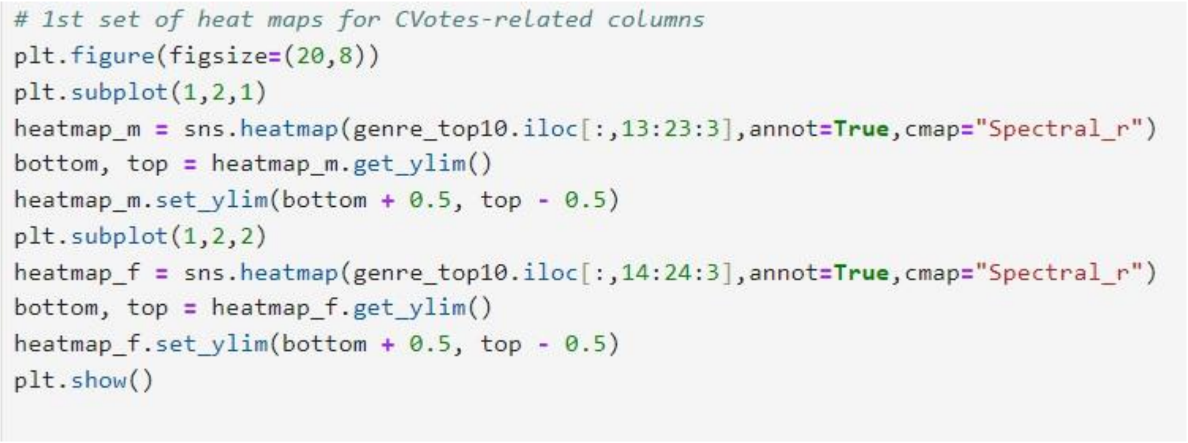
```
# Converting CVotes to int type
genre_top10.iloc[:,0:27] = genre_top10.iloc[:,0:27].astype(int)
```



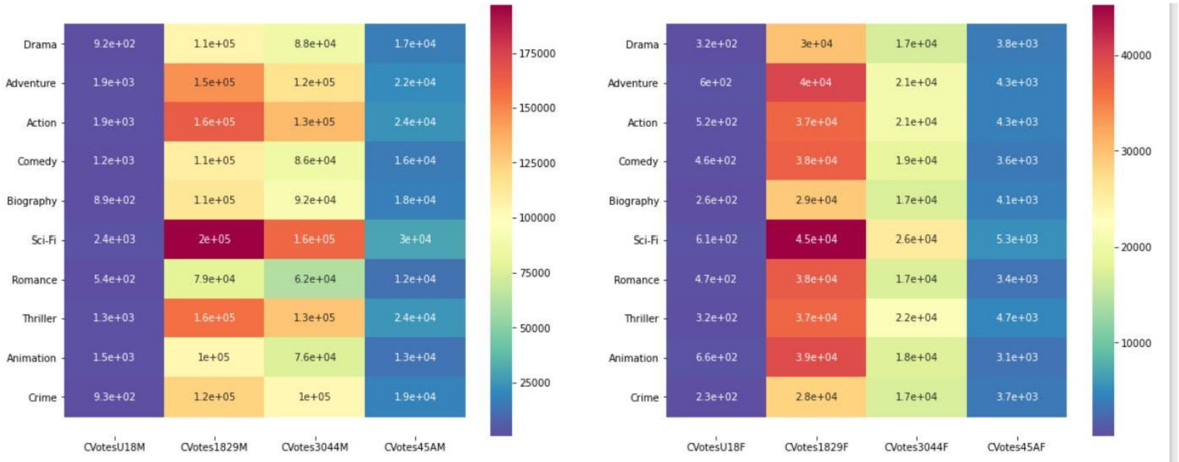
```
# Countplot for genres
genre_top10.cnt.plot.bar()
```

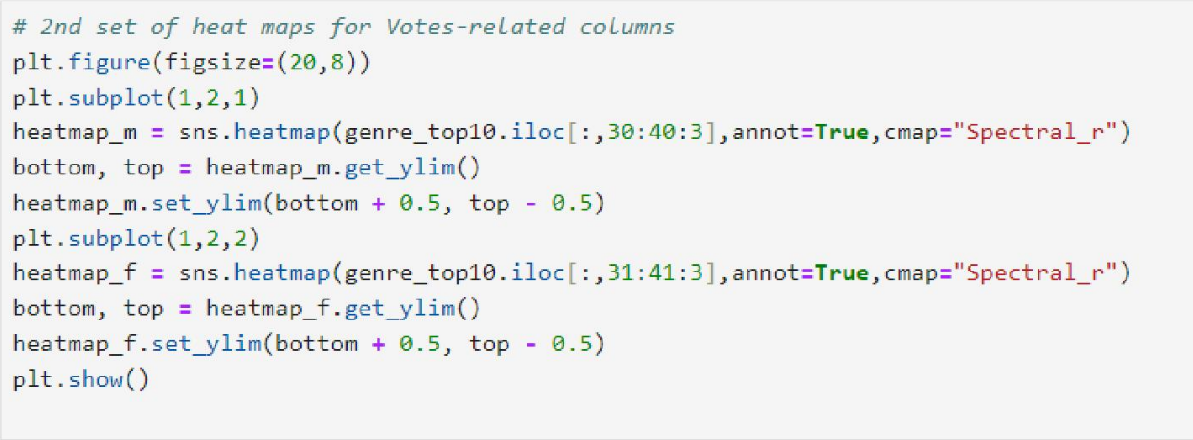
```
<matplotlib.axes._subplots.AxesSubplot at 0x2a3d11c8130>
```



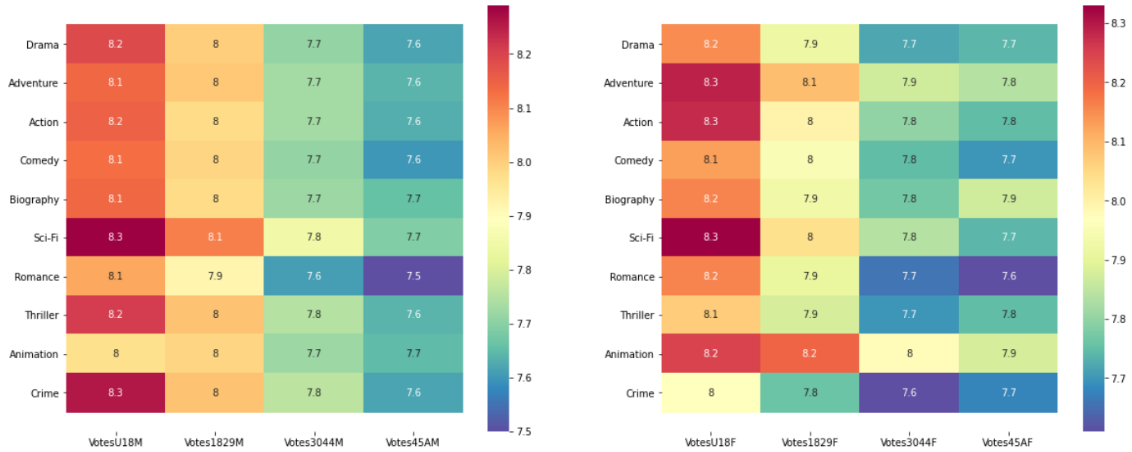


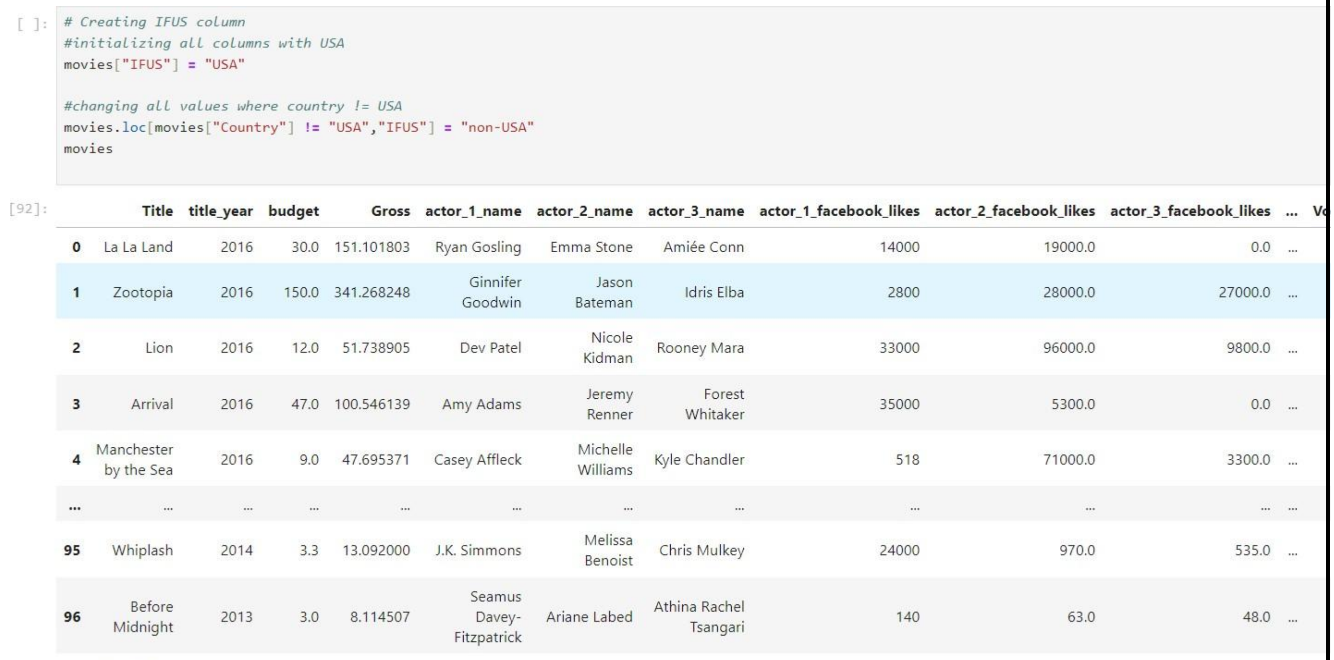
```
# 1st set of heat maps for CVotes-related columns
plt.figure(figsize=(20,8))
plt.subplot(1,2,1)
heatmap_m = sns.heatmap(genre_top10.iloc[:,13:23:3],annot=True,cmap="Spectral_r")
bottom, top = heatmap_m.get_ylim()
heatmap_m.set_ylim(bottom + 0.5, top - 0.5)
plt.subplot(1,2,2)
heatmap_f = sns.heatmap(genre_top10.iloc[:,14:24:3],annot=True,cmap="Spectral_r")
bottom, top = heatmap_f.get_ylim()
heatmap_f.set_ylim(bottom + 0.5, top - 0.5)
plt.show()
```





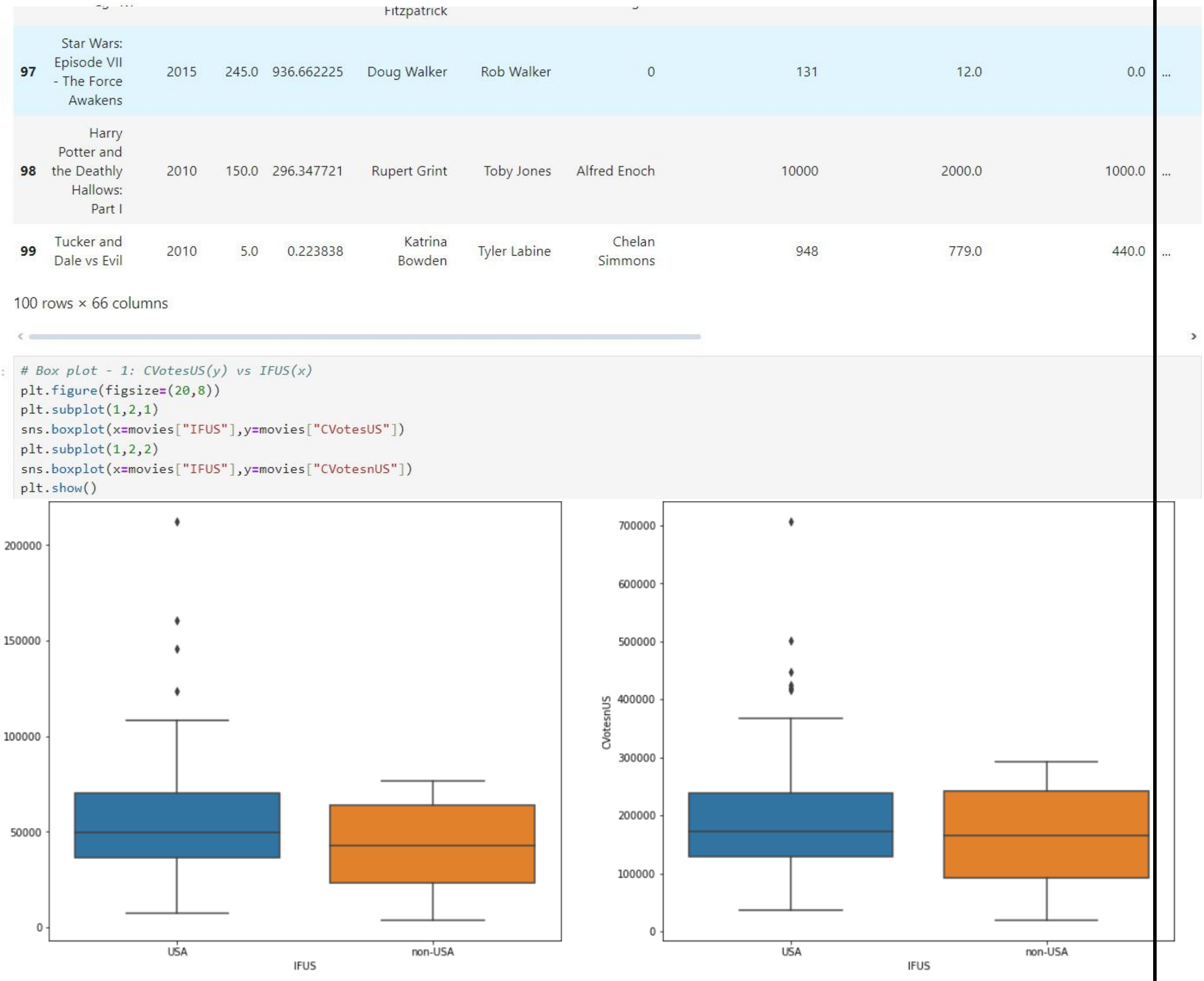
```
# 2nd set of heat maps for Votes-related columns
plt.figure(figsize=(20,8))
plt.subplot(1,2,1)
heatmap_m = sns.heatmap(genre_top10.iloc[:,30:40:3],annot=True,cmap="Spectral_r")
bottom, top = heatmap_m.get_ylim()
heatmap_m.set_ylim(bottom + 0.5, top - 0.5)
plt.subplot(1,2,2)
heatmap_f = sns.heatmap(genre_top10.iloc[:,31:41:3],annot=True,cmap="Spectral_r")
bottom, top = heatmap_f.get_ylim()
heatmap_f.set_ylim(bottom + 0.5, top - 0.5)
plt.show()
```





```
[ ]: # Creating IFUS column
     #initializing all columns with USA
     movies["IFUS"] = "USA"

     #changing all values where country != USA
     movies.loc[movies["Country"] != "USA","IFUS"] = "non-USA"
     movies
```

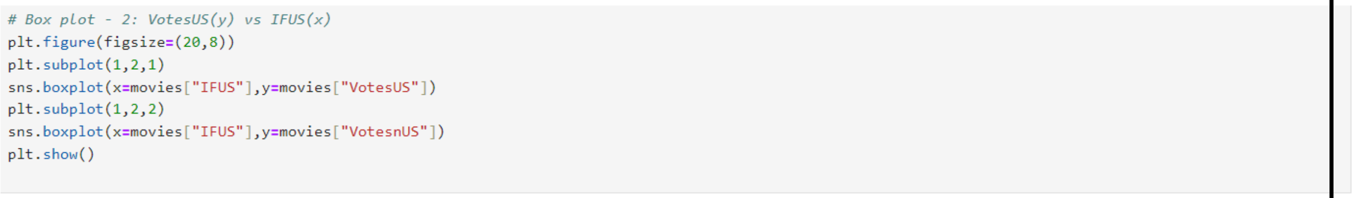
[92]:



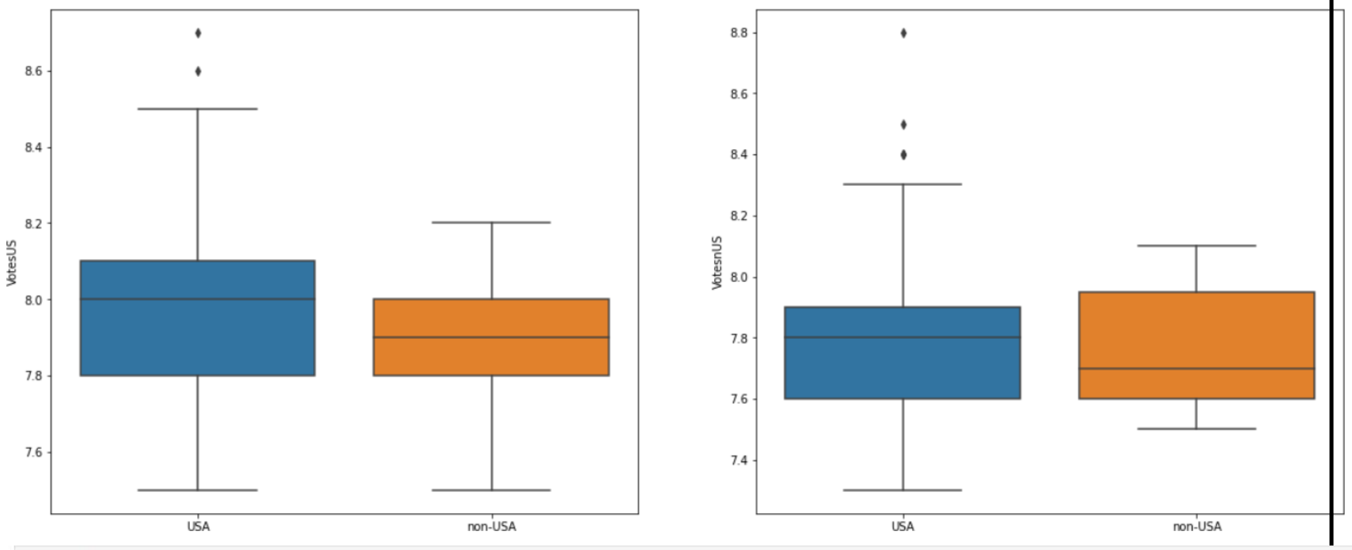
	Title	title_year	budget	Gross	actor_1_name	actor_2_name	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	actor_3_facebook_likes	...	Vo
0	La La Land	2016	30.0	151.101803	Ryan Gosling	Emma Stone	Amiée Conn	14000	19000.0	0.0	...	
1	Zootopia	2016	150.0	341.268248	Ginnifer Goodwin	Jason Bateman	Idris Elba	2800	28000.0	27000.0	...	
2	Lion	2016	12.0	51.738905	Dev Patel	Nicole Kidman	Rooney Mara	33000	96000.0	9800.0	...	
3	Arrival	2016	47.0	100.546139	Amy Adams	Jeremy Renner	Forest Whitaker	35000	5300.0	0.0	...	
4	Manchester by the Sea	2016	9.0	47.695371	Casey Affleck	Michelle Williams	Kyle Chandler	518	71000.0	3300.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
95	Whiplash	2014	3.3	13.092000	J.K. Simmons	Melissa Benoist	Chris Mulkey	24000	970.0	535.0	...	
96	Before Midnight	2013	3.0	8.114507	Seamus Davey-Fitzpatrick	Ariane Labed	Athina Rachel Tsangari	140	63.0	48.0	...	
97	Star Wars: Episode VII - The Force Awakens	2015	245.0	936.662225	Doug Walker	Rob Walker	0	131	12.0	0.0	...	
98	Harry Potter and the Deathly Hallows: Part I	2010	150.0	296.347721	Rupert Grint	Toby Jones	Alfred Enoch	10000	2000.0	1000.0	...	
99	Tucker and Dale vs Evil	2010	5.0	0.223838	Katrina Bowden	Tyler Labine	Chelan Simmons	948	779.0	440.0	...	

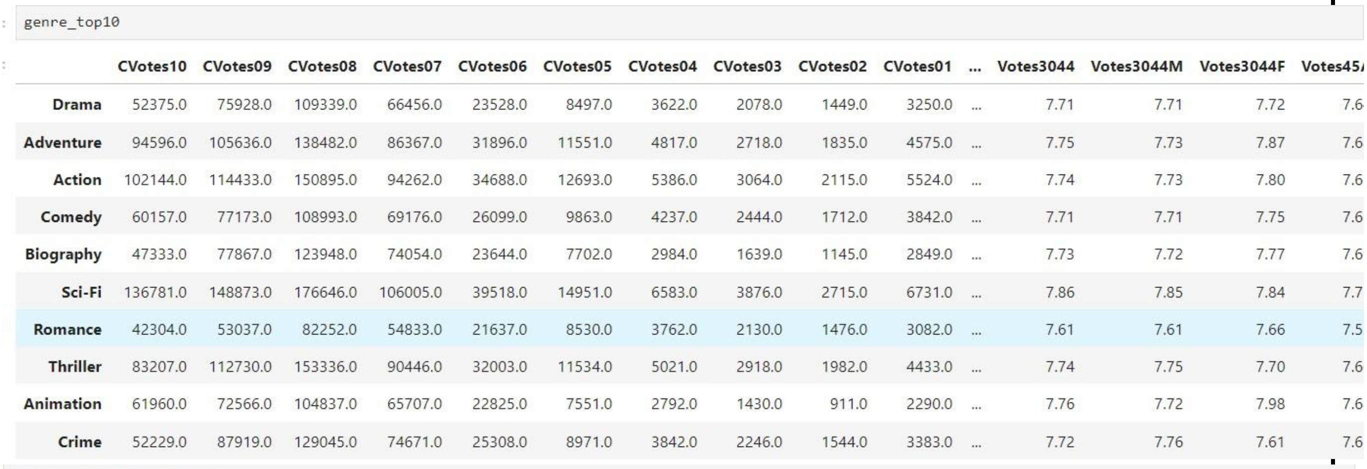
100 rows × 66 columns

```
# Box plot - 1: CVotesUS(y) vs IFUS(x)
plt.figure(figsize=(20,8))
plt.subplot(1,2,1)
sns.boxplot(x=movies["IFUS"],y=movies["CVotesUS"])
plt.subplot(1,2,2)
sns.boxplot(x=movies["IFUS"],y=movies["CVotesnUS"])
plt.show()
```



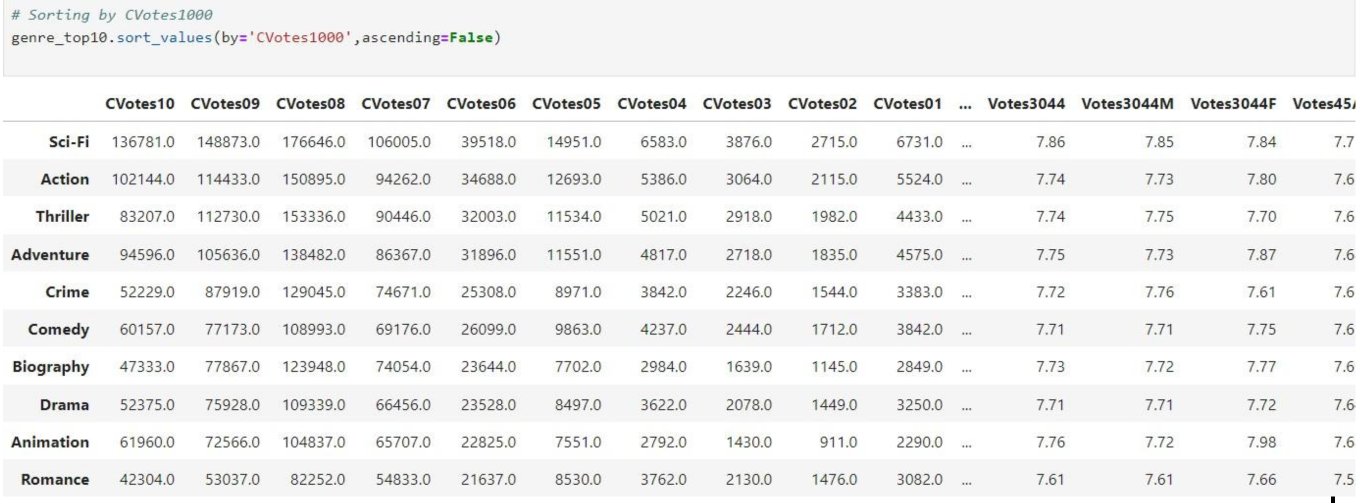
```
# Box plot - 2: VotesUS(y) vs IFUS(x)
plt.figure(figsize=(20,8))
plt.subplot(1,2,1)
sns.boxplot(x=movies["IFUS"],y=movies["VotesUS"])
plt.subplot(1,2,2)
sns.boxplot(x=movies["IFUS"],y=movies["VotesnUS"])
plt.show()
```





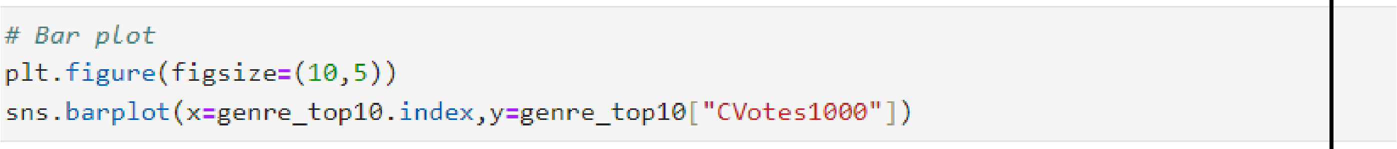
```
genre_top10
```

	CVotes10	CVotes09	CVotes08	CVotes07	CVotes06	CVotes05	CVotes04	CVotes03	CVotes02	CVotes01	...	Votes3044	Votes3044M	Votes3044F	Votes45A
Drama	52375.0	75928.0	109339.0	66456.0	23528.0	8497.0	3622.0	2078.0	1449.0	3250.0	...	7.71	7.71	7.72	7.6
Adventure	94596.0	105636.0	138482.0	86367.0	31896.0	11551.0	4817.0	2718.0	1835.0	4575.0	...	7.75	7.73	7.87	7.6
Action	102144.0	114433.0	150895.0	94262.0	34688.0	12693.0	5386.0	3064.0	2115.0	5524.0	...	7.74	7.73	7.80	7.6
Comedy	60157.0	77173.0	108993.0	69176.0	26099.0	9863.0	4237.0	2444.0	1712.0	3842.0	...	7.71	7.71	7.75	7.6
Biography	47333.0	77867.0	123948.0	74054.0	23644.0	7702.0	2984.0	1639.0	1145.0	2849.0	...	7.73	7.72	7.77	7.6
Sci-Fi	136781.0	148873.0	176646.0	106005.0	39518.0	14951.0	6583.0	3876.0	2715.0	6731.0	...	7.86	7.85	7.84	7.7
Romance	42304.0	53037.0	82252.0	54833.0	21637.0	8530.0	3762.0	2130.0	1476.0	3082.0	...	7.61	7.61	7.66	7.5
Thriller	83207.0	112730.0	153336.0	90446.0	32003.0	11534.0	5021.0	2918.0	1982.0	4433.0	...	7.74	7.75	7.70	7.6
Animation	61960.0	72566.0	104837.0	65707.0	22825.0	7551.0	2792.0	1430.0	911.0	2290.0	...	7.76	7.72	7.98	7.6
Crime	52229.0	87919.0	129045.0	74671.0	25308.0	8971.0	3842.0	2246.0	1544.0	3383.0	...	7.72	7.76	7.61	7.6



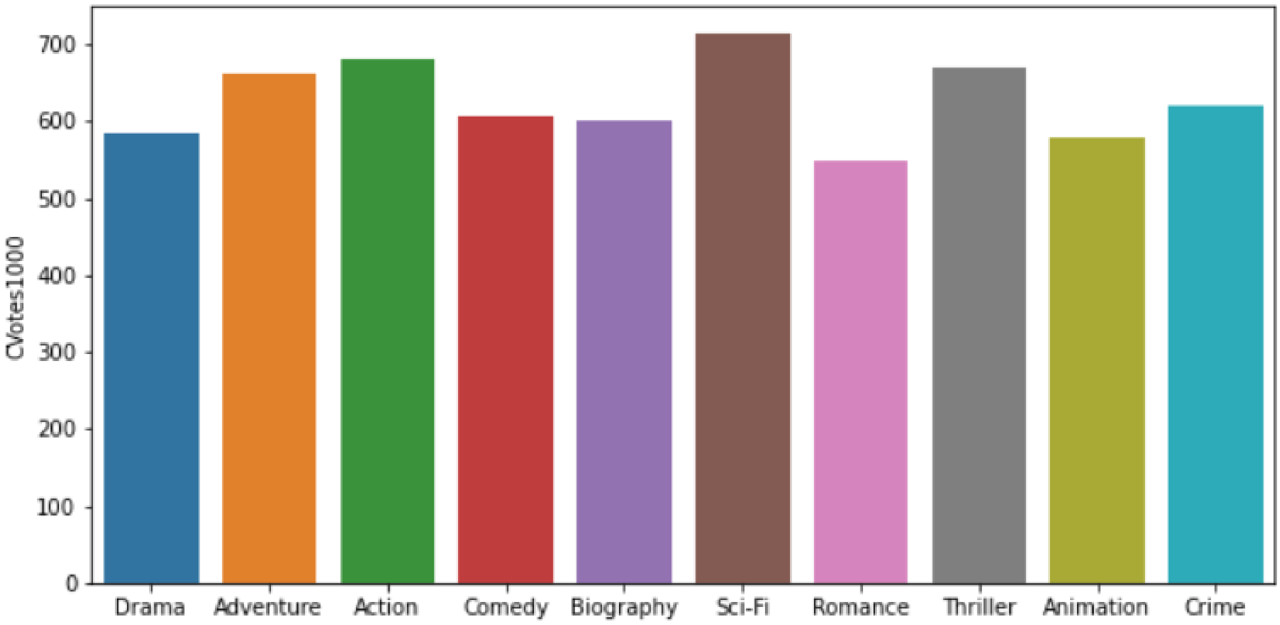
```
# Sorting by CVotes1000
genre_top10.sort_values(by='CVotes1000',ascending=False)
```

	CVotes10	CVotes09	CVotes08	CVotes07	CVotes06	CVotes05	CVotes04	CVotes03	CVotes02	CVotes01	...	Votes3044	Votes3044M	Votes3044F	Votes45A
Sci-Fi	136781.0	148873.0	176646.0	106005.0	39518.0	14951.0	6583.0	3876.0	2715.0	6731.0	...	7.86	7.85	7.84	7.7
Action	102144.0	114433.0	150895.0	94262.0	34688.0	12693.0	5386.0	3064.0	2115.0	5524.0	...	7.74	7.73	7.80	7.6
Thriller	83207.0	112730.0	153336.0	90446.0	32003.0	11534.0	5021.0	2918.0	1982.0	4433.0	...	7.74	7.75	7.70	7.6
Adventure	94596.0	105636.0	138482.0	86367.0	31896.0	11551.0	4817.0	2718.0	1835.0	4575.0	...	7.75	7.73	7.87	7.6
Crime	52229.0	87919.0	129045.0	74671.0	25308.0	8971.0	3842.0	2246.0	1544.0	3383.0	...	7.72	7.76	7.61	7.6
Comedy	60157.0	77173.0	108993.0	69176.0	26099.0	9863.0	4237.0	2444.0	1712.0	3842.0	...	7.71	7.71	7.75	7.6
Biography	47333.0	77867.0	123948.0	74054.0	23644.0	7702.0	2984.0	1639.0	1145.0	2849.0	...	7.73	7.72	7.77	7.6
Drama	52375.0	75928.0	109339.0	66456.0	23528.0	8497.0	3622.0	2078.0	1449.0	3250.0	...	7.71	7.71	7.72	7.6
Animation	61960.0	72566.0	104837.0	65707.0	22825.0	7551.0	2792.0	1430.0	911.0	2290.0	...	7.76	7.72	7.98	7.6
Romance	42304.0	53037.0	82252.0	54833.0	21637.0	8530.0	3762.0	2130.0	1476.0	3082.0	...	7.61	7.61	7.66	7.5



```
# Bar plot
plt.figure(figsize=(10,5))
sns.barplot(x=genre_top10.index,y=genre_top10["CVotes1000"])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2a3d3647ac0>
```



Information about the dataset:

Title: Title of movie Release

Date: The release date of the movie

Color/B&W: Movies Release Type

Genre: a style or category of art, music, or literature.

Language: Language in which movies was released

Country: Country where the movie was released

Rating: Rating for movie Lead

Actor: Lead actor in that movie Director

Name: Director name For that movie

Lead Actor FB Likes: Lead actors FB likes

Cast FB Likes: Cast actors FB likes

Director FB Likes: Director actors FB likes

Movie FB Likes: Movie actors FB likes

IMDb Score (1–10): IMDb actors score given

Total Reviews: Total reviews given to movie Duration (min) : Duration Movie in minute

Gross Revenue: Gross revenue, also known as gross income, is the sum of all money generated by a business, without taking into account any part of that total that has been or will be used for expenses

Budget: A budget is an estimation of revenue and expenses over a specified future period of time and is utilized by governments, businesses, and individuals. A budget is basically a financial plan for a defined period, normally a year that is known to greatly enhance the success of any financial undertaking.

Conclusion

In this project about analyzing IMDb movie data, we looked at movie ratings, cast details, and even fans' social media accounts. We also explored public reviews and created graphs to see which movie genres made the most money each year. From our findings, we discovered that sci-fi movies are the most popular among viewers, more so than romantic or mass appeal films. This analysis helped us understand what people like and how to make better business choices. We also calculated important statistics, like the average and range of IMDb scores for different genres. We used a scatter plot to see how movie budgets related to profits and identified the top ten highest-grossing movies, as well as those that lost money.

Overall, this analysis provided great insights into movie genres, ratings, and financial success, helping us understand what audiences prefer and current trends in the film industry.