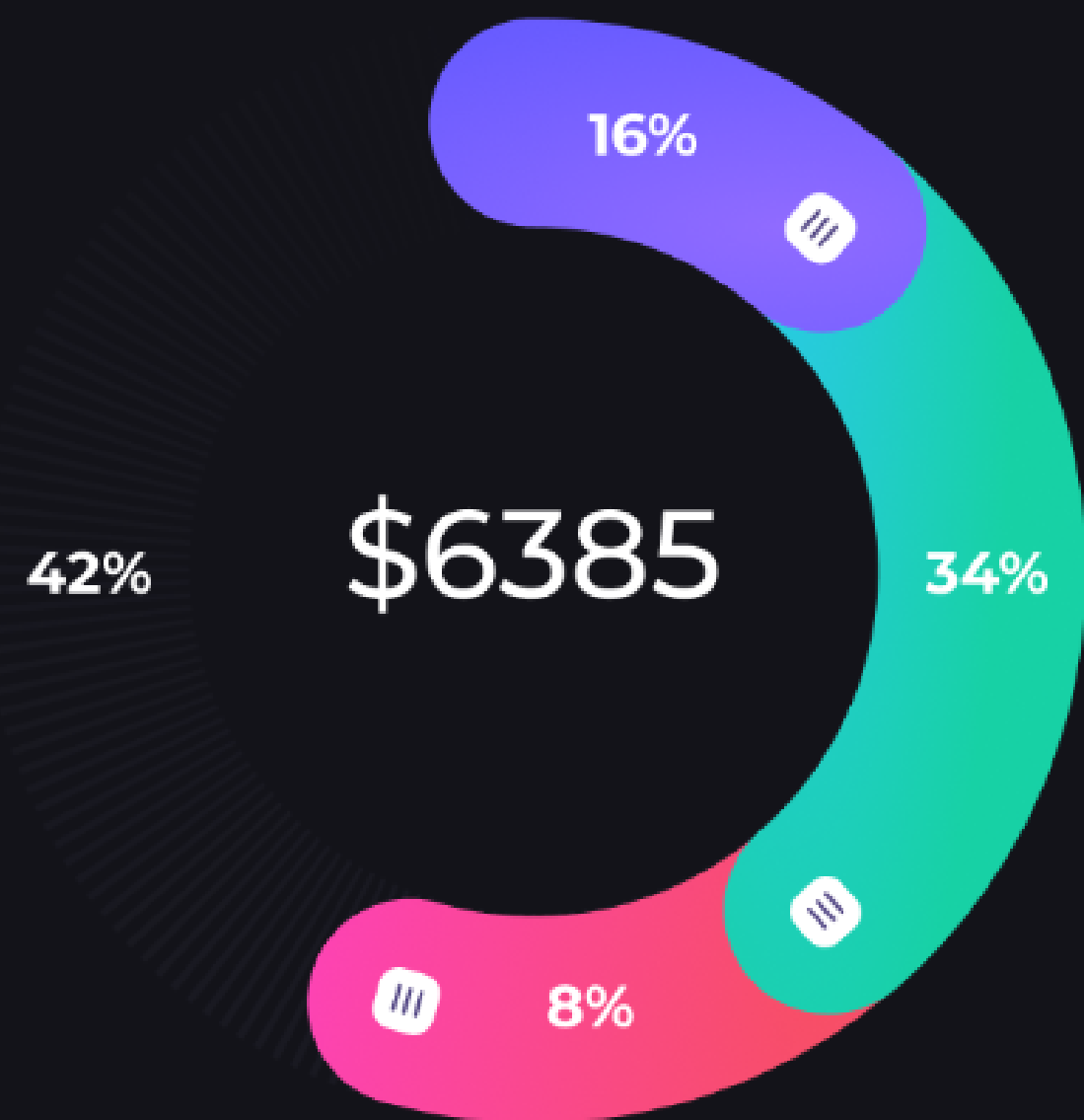
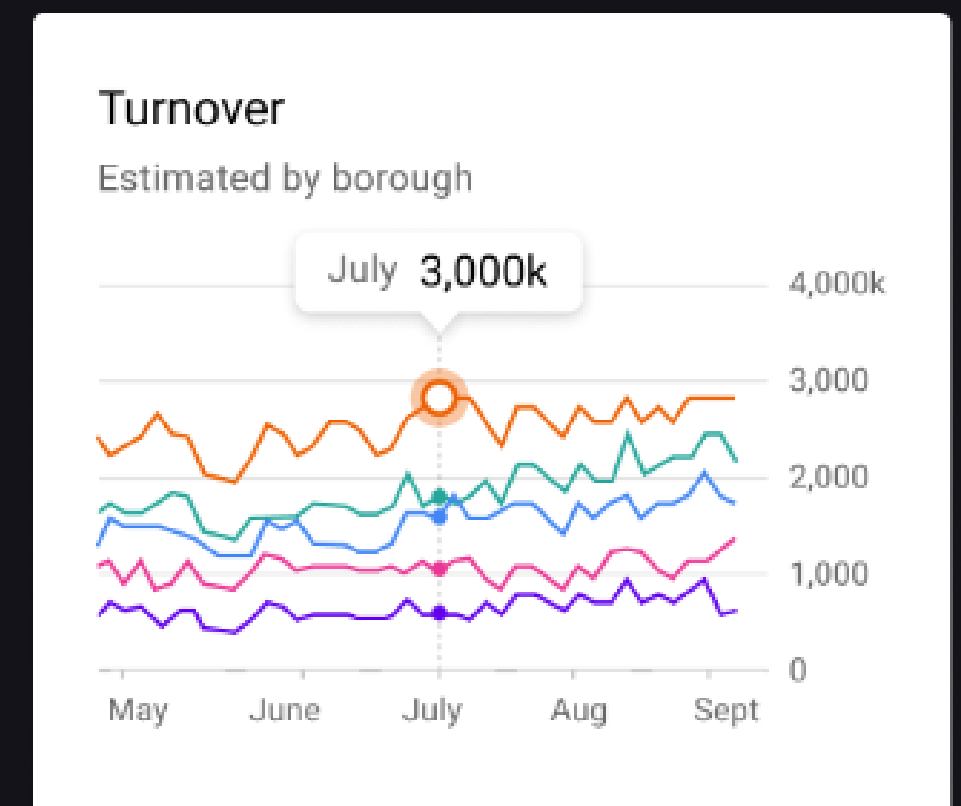
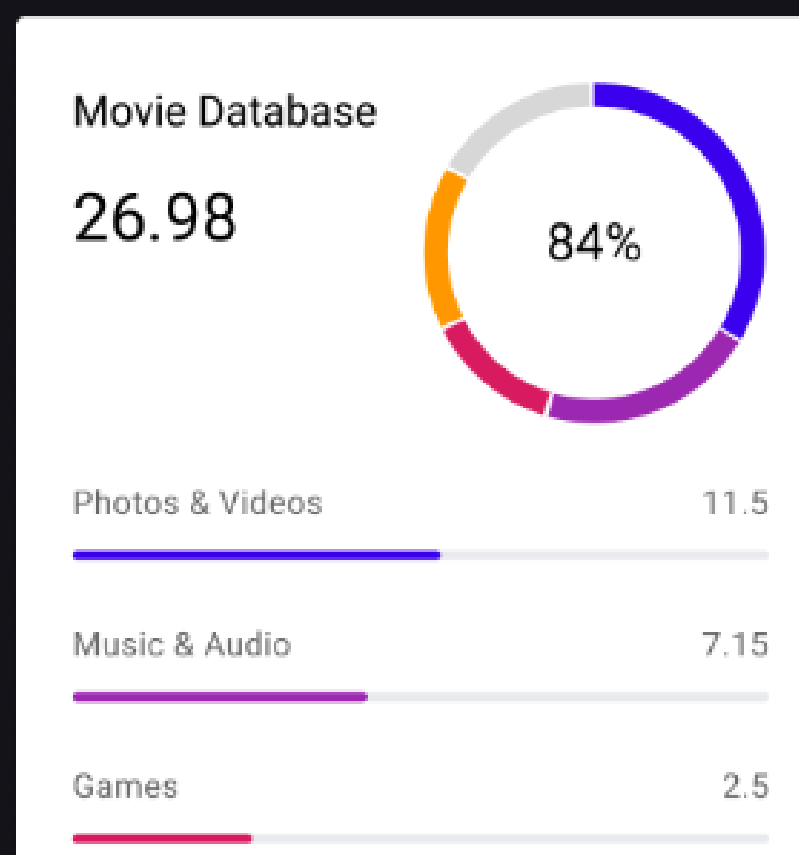


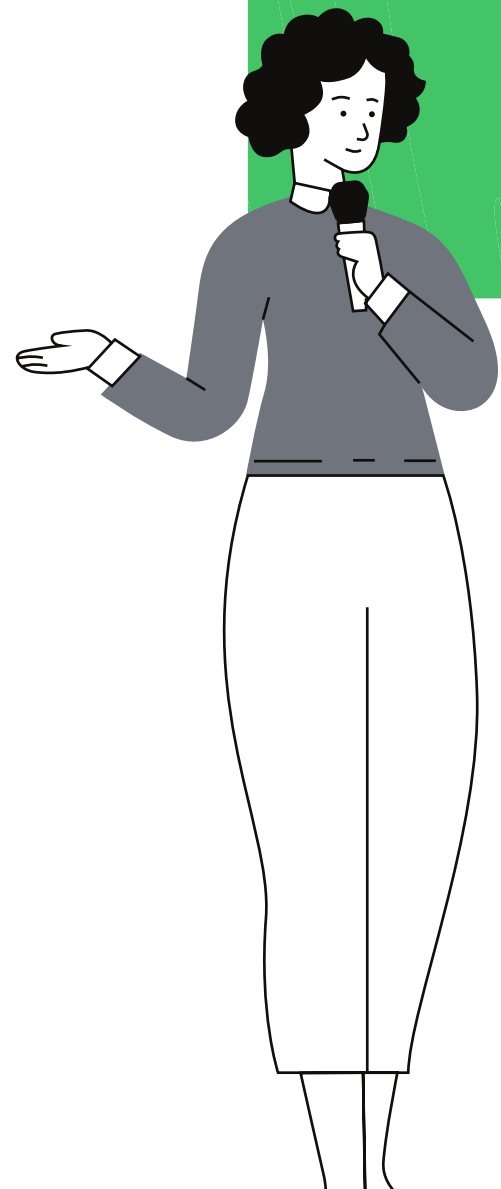
IMDB Movie Analysis

Project by Mayur Rajput



[Excel File Link](#)





Project's Agenda

1

Project description

2

Approach & Tech-Stack Used

3

Data Cleaning

4

Data Analysis

5

Interactive Dashboard

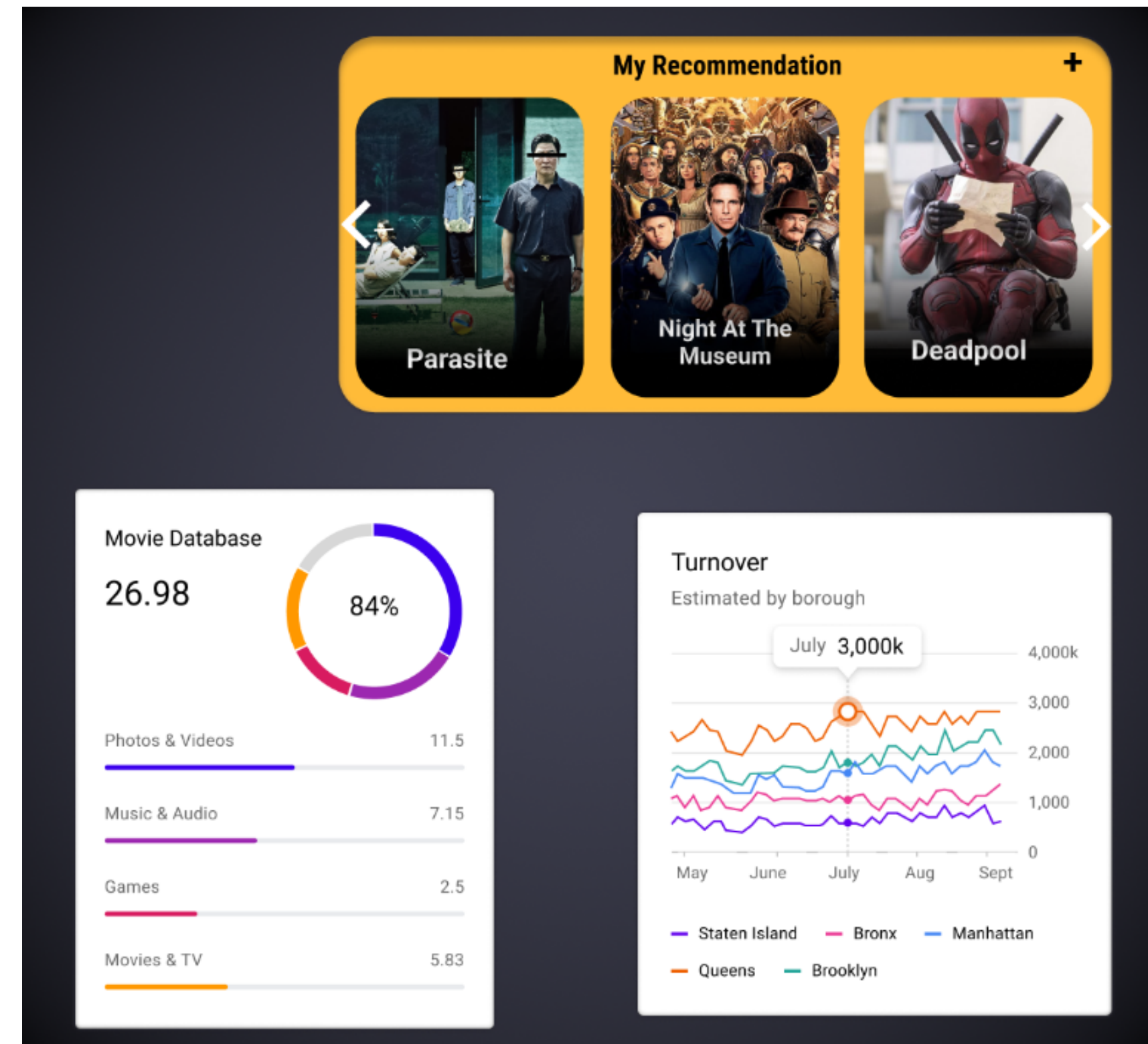
6

Insights & Results

[Excel File Link](#)

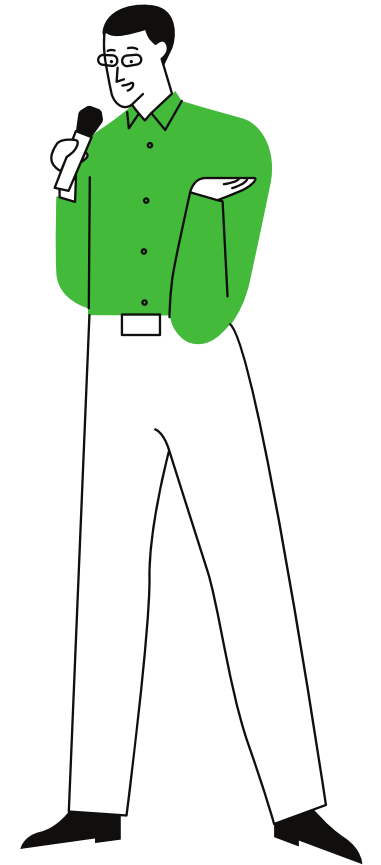
Project Description

The goal of this project is to use our knowledge of statistics and Excel to clean, analyze dataset and find out what factors makes a movie successful. Here, success can be defined by high IMDB ratings. These factors can be significantly used by movie producers, directors, and investors who want to understand what makes a movie successful and to make informed decisions in their future projects.



Approach

- **Data Cleaning:** This step involves cleaning the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering.



- **Data Analysis:** This step involves analyzing the data to find What factors influence the success of a movie on IMDB?

Tech Stack Used

Microsoft Excel

Microsoft Excel is used to handle handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering outliers, missing values and also used to perform analysis and answer certain questions that can help to understand what makes a movie successful and to make informed decisions in their future projects.

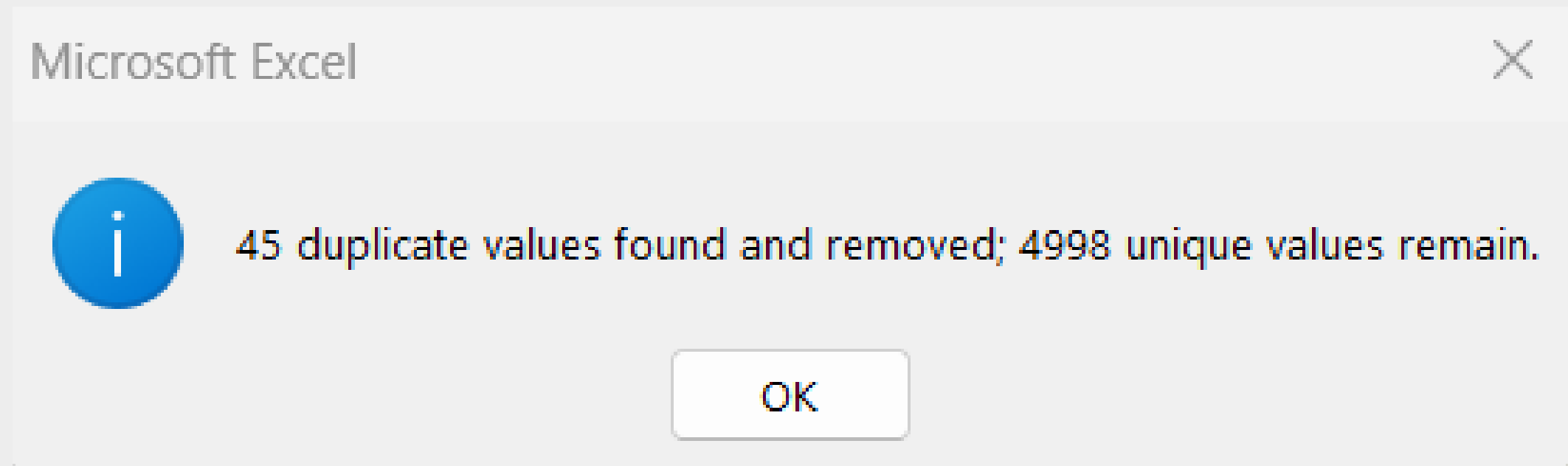
Canva

Canva is used to prepare this presentation

Data Cleaning

Removing duplicates from Dataset

- Total rows in Raw dataset provided – 5043
- Rows left after removing duplicates – 4998



Data Cleaning

Handling Missing data

a) Finding missing values in dataset:

Finding Missing Values in Dataset								Total Rows =		4998	Total Columns =		28									
19	103	49	15	103	23	13	7	874	0	7	0	0	0	23	13	152	0	20	12	5	301	487
0.4%	2.1%	1.0%	0.3%	2.1%	0.5%	0.3%	0.1%	17.5%	0.0%	0.1%	0.0%	0.0%	0.0%	0.5%	0.3%	3.0%	0.0%	0.4%	0.2%	0.1%	6.0%	9.7%
color	director	num_cr	duration	director	actor_3	actor_2	actor_1	gross	genres	actor_1	movie	num_vc	cast_to	actor_3	facenur	plot_ke	movie	num_us	language	country	content	budget
Color	James Cam	723	178	0	855	Joel David	1000	7.61E+08	Action Ad	CCH Pounc	Avatar	886204	4834	Wes Studi	0	avatar fut	http://ww	3054	English	USA	PG-13	2.37E+08
Color	Gore Verbi	302	169	563	1000	Orlando Bl	40000	3.09E+08	Action Ad	Johnny De	Pirates of	471220	48350	Jack Daver	0	goddess n	http://ww	1238	English	USA	PG-13	3E+08
Color	Sam Mend	602	148	0	161	Rory Kinne	11000	2E+08	Action Ad	Christoph	Spectre	275868	11700	Stephanie	1	bomb esp	http://ww	994	English	UK	PG-13	2.45E+08
Color	Christophe	813	164	22000	23000	Christian B	27000	4.48E+08	Action Thr	Tom Hardy	The Dark K	1144337	106759	Joseph Go	0	deception	http://ww	2701	English	USA	PG-13	2.5E+08
	Doug Walker			131		Rob Walke	131		Document	Doug Walk	Star Wars:	8	143		0		http://ww					
Color	Andrew Sta	462	132	475	530	Samantha	640	73058679	Action Ad	Daryl Saba	John Carte	212204	1873	Polly Walk	1	alien ame	http://ww	738	English	USA	PG-13	2.64E+08
Color	Sam Raimi	392	156	0	4000	James Fran	24000	3.37E+08	Action Ad	J.K. Simmc	Spider-Ma	383056	46055	Kirsten Du	0	sandman :	http://ww	1902	English	USA	PG-13	2.58E+08
Color	Nathan Gr	324	100	15	284	Donna Mu	799	2.01E+08	Adventure	Brad Garre	Tangled	294810	2036	M.C. Gaine	1	17th centu	http://ww	387	English	USA	PG	2.6E+08
Color	Joss Whed	635	141	0	19000	Robert Do	26000	4.59E+08	Action Ad	Chris Hem	Avengers:	462669	92000	Scarlett Jo	4	artificial in	http://ww	1117	English	USA	PG-13	2.5E+08
Color	David Yate	375	153	282	10000	Daniel Rad	25000	3.02E+08	Adventure	Alan Rickn	Harry Pott	321795	58753	Rupert Gri	3	blood boc	http://ww	973	English	UK	PG	2.5E+08
Color	Zack Snyder	673	183	0	2000	Lauren Col	15000	3.3E+08	Action Ad	Henry Cavi	Batman v	371639	24450	Alan D. Pul	0	based on c	http://ww	3018	English	USA	PG-13	2.5E+08
Color	Bryan Sing	434	169	0	903	Marlon Bra	18000	2E+08	Action Ad	Kevin Spac	Superman	240396	29991	Frank Lang	0	crystal epi	http://ww	2367	English	USA	PG-13	2.09E+08
Color	Marc Forst	403	106	395	393	Mathieu A	451	1.68E+08	Action Ad	Giancarlo	Quantum c	330784	2023	Rory Kinne	1	action her	http://ww	1243	English	UK	PG-13	2E+08
Color	Gore Verbi	313	151	563	1000	Orlando Bl	40000	4.23E+08	Action Ad	Johnny De	Pirates of	522040	48486	Jack Daver	2	box office	http://ww	1832	English	USA	PG-13	2.25E+08
Color	Gore Verbi	450	150	563	1000	Ruth Wils	40000	89289910	Action Ad	Johnny De	The Lone F	181792	45757	Tom Wilkin	1	horse out	http://ww	711	English	USA	PG-13	2.15E+08
Color	Zack Snyder	733	143	0	748	Christophe	15000	2.91E+08	Action Ad	Henry Cavi	Man of Ste	548573	20495	Harry Lenr	0	based on c	http://ww	2536	English	USA	PG-13	2.25E+08
Color	Andrew Ac	258	150	80	201	Pierfrance	22000	1.42E+08	Action Ad	Peter Dink	The Chron	149922	22697	Dami	4	brother br	http://ww	438	English	USA	PG	2.25E+08
Color	Joss Whed	703	173	0	19000	Robert Do	26000	6.23E+08	Action Ad	Chris Hem	The Aveng	995415	87697	Scarlett Jo	3	alien invas	http://ww	1722	English	USA	PG-13	2.2E+08
Color	Rob Marsh	448	136	252	1000	Sam Claflir	40000	2.41E+08	Action Ad	Johnny De	Pirates of	370704	54083	Stephen G	4	blackbear	http://ww	484	English	USA	PG-13	2.5E+08
Color	Barry Sonr	451	106	188	718	Michael St	10000	1.79E+08	Action Ad	Will Smith	Men in Bla	268154	12572	Nicole Sch	1	alien crim	http://ww	341	English	USA	PG-13	2.25E+08
Color	Peter Jack	422	164	0	773	Adam Brov	5000	2.55E+08	Adventure	Aidan Turn	The Hobbit	354228	9152	James Nes	0	army elf	http://ww	802	English	New Zeala	PG-13	2.5E+08
Color	Marc Web	599	153	464	963	Andrew Ga	15000	2.62E+08	Action Ad	Emma Sto	The Amazi	451803	28489	Chris Zylka	0	lizard out	http://ww	1225	English	USA	PG-13	2.3E+08

Data Cleaning

Handling Missing data

b) Performing CCA:

- Removing rows having missing values
- 1275 rows are removed from dataset
- 3723 Rows left in dataset

Finding Missing Values in Dataset								Total Rows =		3723		Total Columns =		28									
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	
color	director	num_cr	duratio	director	actor_3	actor_2	actor_1	gross	genres	actor_1	movie	num_vc	cast_to	actor_3	facenur	plot_ke	movie	num_us	language	country	content	budget	
Color	James Can	723	178	0	855	Joel David	1000	7.61E+08	Action Ad	CCH Pounc	Avatar	886204	4834	Wes Studi	0	avatar fut	http://ww	3054	English	USA	PG-13	2.37E+08	
Color	Gore Verbi	302	169	563	1000	Orlando Bl	40000	3.09E+08	Action Ad	Johnny De	Pirates of	471220	48350	Jack Daver	0	goddess n	http://ww	1238	English	USA	PG-13	3E+08	
Color	Sam Mend	602	148	0	161	Rory Kinne	11000	2E+08	Action Ad	Christoph	Spectre	275868	11700	Stephanie	1	bomb esp	http://ww	994	English	UK	PG-13	2.45E+08	
Color	Christophe	813	164	22000	23000	Christian B	27000	4.48E+08	Action Thi	Tom Hardy	The Dark K	1144337	106759	Joseph Go	0	deception	http://ww	2701	English	USA	PG-13	2.5E+08	
Color	Andrew St	462	132	475	530	Samantha	640	73058679	Action Ad	Daryl Saba	John Carte	212204	1873	Polly Walk	1	alien ame	http://ww	738	English	USA	PG-13	2.64E+08	
Color	Sam Raimi	392	156	0	4000	James Fran	24000	3.37E+08	Action Ad	J.K. Simmc	Spider-Ma	383056	46055	Kirsten Du	0	sandman :	http://ww	1902	English	USA	PG-13	2.58E+08	
Color	Nathan Gr	324	100	15	284	Donna Mu	799	2.01E+08	Adventure	Brad Garre	Tangled	294810	2036	M.C. Gaine	1	17th centu	http://ww	387	English	USA	PG	2.6E+08	
Color	Joss Whed	635	141	0	19000	Robert Do	26000	4.59E+08	Action Ad	Chris Hem	Avengers:	462669	92000	Scarlett Jo	4	artificial in	http://ww	1117	English	USA	PG-13	2.5E+08	
Color	David Yate	375	153	282	10000	Daniel Rad	25000	3.02E+08	Adventure	Alan Rickn	Harry Pott	321795	58753	Rupert Gri	3	blood boc	http://ww	973	English	UK	PG	2.5E+08	
Color	Zack Snyder	673	183	0	2000	Lauren Col	15000	3.3E+08	Action Ad	Henry Cavi	Batman v S	371639	24450	Alan D. Pul	0	based on c	http://ww	3018	English	USA	PG-13	2.5E+08	
Color	Bryan Sing	434	169	0	903	Marlon Bra	18000	2E+08	Action Ad	Kevin Spac	Superman	240396	29991	Frank Lang	0	crystal epi	http://ww	2367	English	USA	PG-13	2.09E+08	
Color	Marc Forst	403	106	395	393	Mathieu A	451	1.68E+08	Action Ad	Giancarlo	Quantum c	330784	2023	Rory Kinne	1	action her	http://ww	1243	English	UK	PG-13	2E+08	
Color	Gore Verbi	313	151	563	1000	Orlando Bl	40000	4.23E+08	Action Ad	Johnny De	Pirates of	522040	48486	Jack Daver	2	box office	http://ww	1832	English	USA	PG-13	2.25E+08	
Color	Gore Verbi	450	150	563	1000	Ruth Wils	40000	89289910	Action Ad	Johnny De	The Lone F	181792	45757	Tom Wilkin	1	horse out	http://ww	711	English	USA	PG-13	2.15E+08	
Color	Zack Snyder	733	143	0	748	Christophe	15000	2.91E+08	Action Ad	Henry Cavi	Man of Ste	548573	20495	Harry Lenr	0	based on c	http://ww	2536	English	USA	PG-13	2.25E+08	
Color	Andrew Ac	258	150	80	201	Pierfrance	22000	1.42E+08	Action Ad	Peter Dink	The Chron	149922	22697	Dami	4	brother br	http://ww	438	English	USA	PG	2.25E+08	
Color	Joss Whed	703	173	0	19000	Robert Do	26000	6.23E+08	Action Ad	Chris Hem	The Aveng	995415	87697	Scarlett Jo	3	alien invas	http://ww	1722	English	USA	PG-13	2.2E+08	
Color	Rob Marsh	448	136	252	1000	Sam Claflir	40000	2.41E+08	Action Ad	Johnny De	Pirates of	370704	54083	Stephen G	4	blackbearc	http://ww	484	English	USA	PG-13	2.5E+08	
Color	Barry Sonr	451	106	188	718	Michael St	10000	1.79E+08	Action Ad	Will Smith	Men in Bla	268154	12572	Nicole Sch	1	alien crim	http://ww	341	English	USA	PG-13	2.25E+08	
Color	Peter Jack	422	164	0	773	Adam Brov	5000	2.55E+08	Adventure	Aidan Turn	The Hobbit	354228	9152	James Nes	0	army elf l	http://ww	802	English	New Zeala	PG-13	2.5E+08	
Color	Marc Web	599	153	464	963	Andrew Ga	15000	2.62E+08	Action Ad	Emma Sto	The Amazi	451803	28489	Chris Zylka	0	lizard out	http://ww	1225	English	USA	PG-13	2.3E+08	
Color	Ridley Sco	343	156	0	738	William Hu	891	1.05E+08	Action Ad	Mark Addy	Robin Hoo	211765	3244	Scott Grim	0	1190s arc	http://ww	546	English	USA	PG-13	2E+08	

Data Analysis

a) Movie Genre Analysis:

Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, variance, standard deviation) of the IMDB scores.

Genre	Count of movies
Comedy	984
Action	951
Drama	659
Adventure	366
Crime	253
Biography	204
Horror	159
Animation	45
Fantasy	37
Documentary	26
Mystery	23
Sci-Fi	7
Family	3
Western	2
Musical	2
Romance	1
Thriller	1

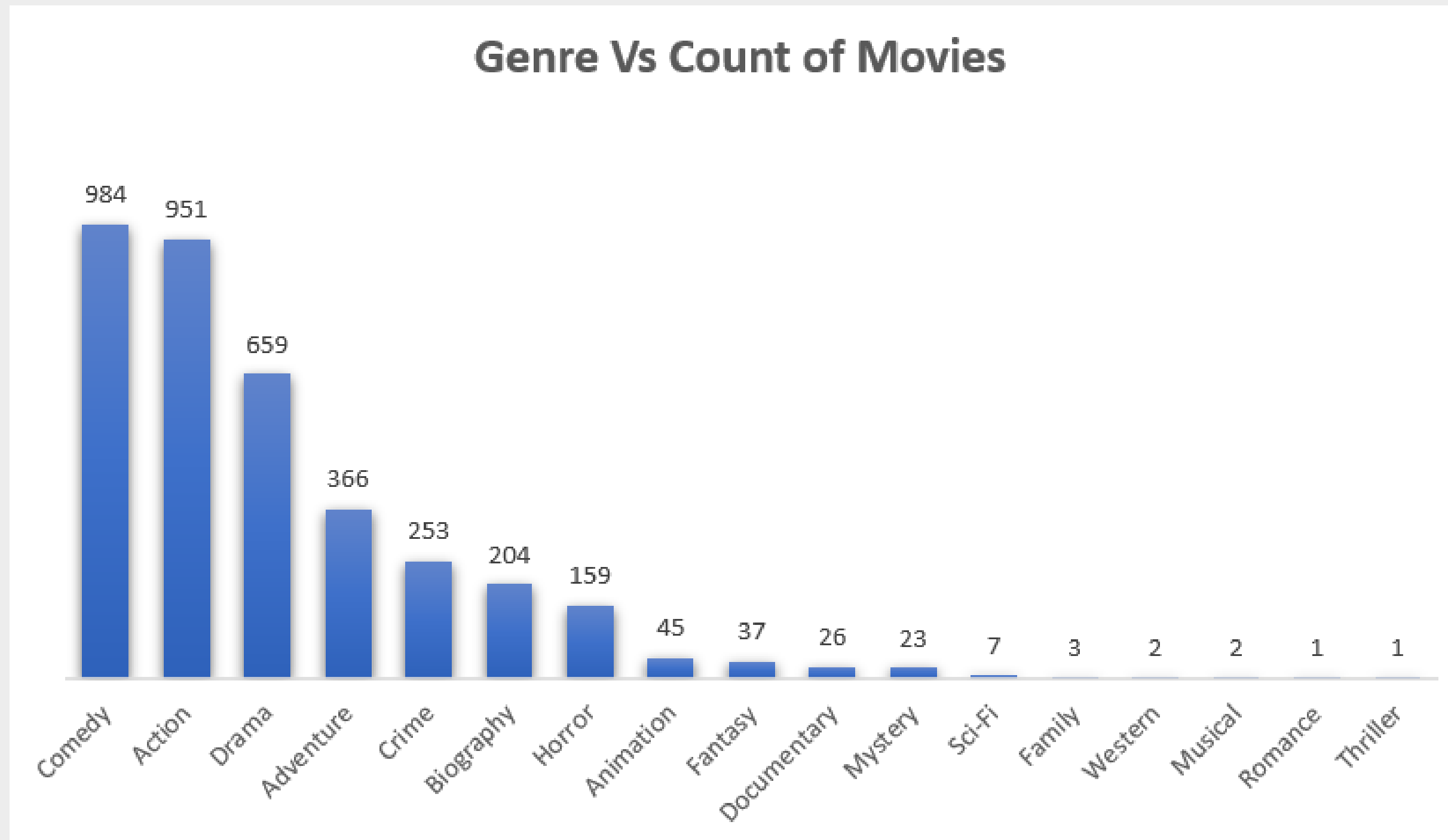
Descriptive Statistics based on IMDB Score for Genres							
Genres	Mean	Median	Mode	Std Deviat	Variance	Min	Max
Comedy	6.2	6.3	6.4	1.0	1.1	1.9	8.8
Action	6.3	6.3	6.1	1.0	1.1	2.1	9
Drama	6.8	6.9	6.7	0.9	0.8	2.1	8.8
Adventure	6.6	6.7	7.3	1.1	1.3	2.3	8.6
Crime	6.9	7	7.3	0.9	0.8	3.3	9.3
Biography	7.2	7.2	7	0.7	0.5	4.5	8.9
Horror	5.8	5.9	5.9	1.0	1.1	2.3	8.5
Animation	6.7	7	7.1	1.0	0.9	4.5	8
Fantasy	6.3	6.5	6.8	0.9	0.8	4.3	7.9
Documentary	6.8	7.45	7.5	1.7	2.8	1.6	8.5
Mystery	6.7	6.7	7.1	1.1	1.1	3.3	8.5
Sci-Fi	6.6	6.4	#N/A	1.0	1.1	5	8.2
Family	6.5	5.9	#N/A	1.0	1.0	5.7	7.9
Western	8.1	8.1	#N/A	0.8	0.6	7.3	8.9
Musical	6.8	6.75	#N/A	0.5	0.2	6.3	7.2
Romance	7.1	7.1	#N/A	0.0	0.0	7.1	7.1
Thriller	4.8	4.8	#N/A	0.0	0.0	4.8	4.8

- The most common genres of movies are Comedy, Action and Drama
- Movies in the Biography, Western and Romance genres tend to have higher average IMDb scores compared to other genres.

Data Analysis

a) Movie Genre Analysis:

Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.



Data Analysis

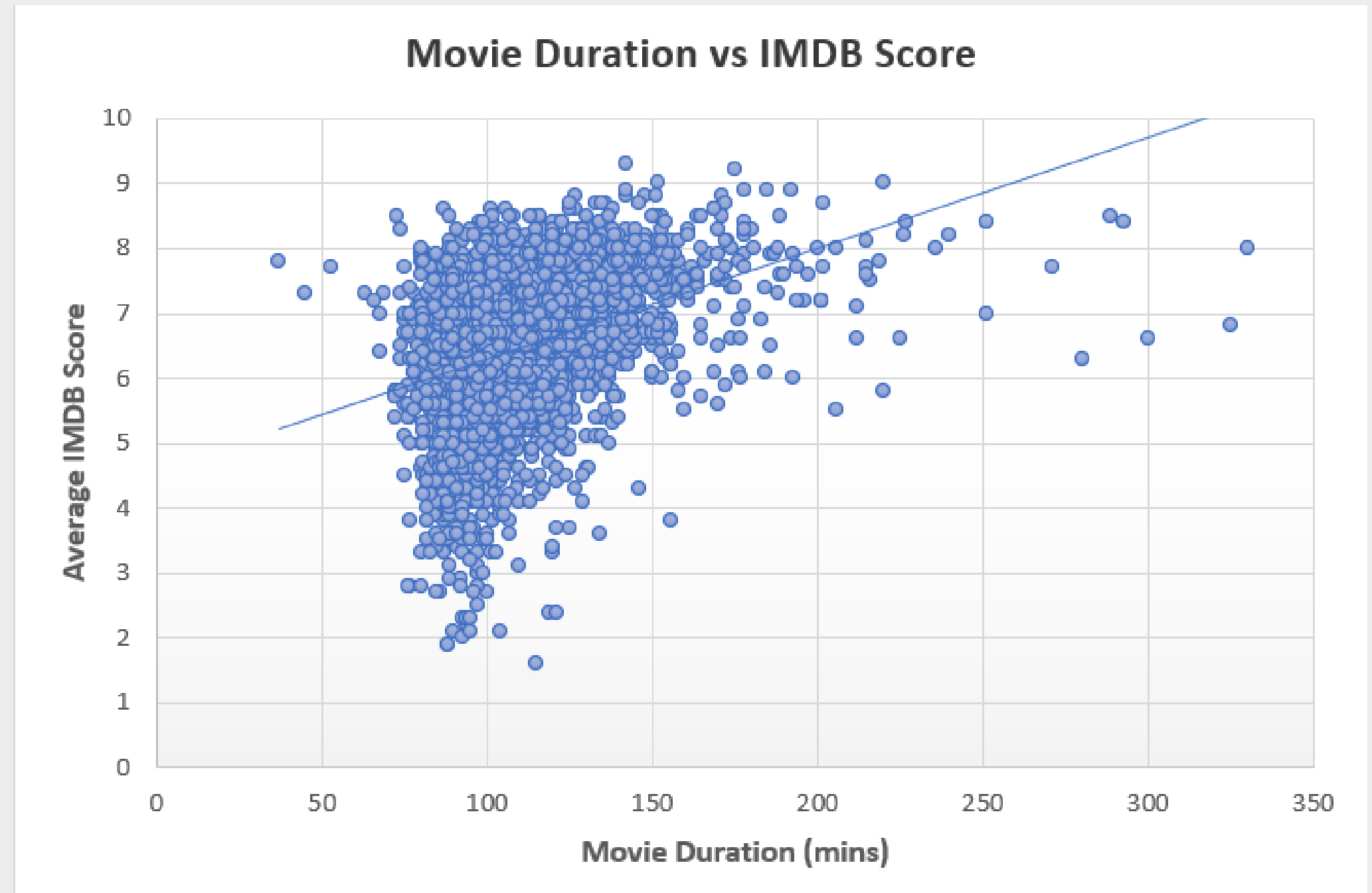
b) Movie Duration Analysis:

Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Duration (mins)	Count of movie_title
1-50	2
51-100	1388
101-150	2171
151-200	132
201-250	21
251-300	7
301-350	2

- Most common duration range for movies is 50– 150 mins.
- Movie duration doesn't strongly influence its IMDB score.

Average =	110.3
Median =	106.0
Std Deviation =	22.7



Data Analysis

c) Language Analysis:

Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Top 10 Most Common Languages	
Language	Count of movies
English	3566
French	34
Spanish	23
Mandarin	14
Japanese	10
German	10
Italian	7
Cantonese	7
Portuguese	5
Korean	5

- Most no of movies produced are in English language.
- English language has low average IMDB score as compared other languages like Japanese, German and it can be due quality of movie produced in these language is better than English language.


Descriptive Statistics of Language based on IMDB Score			
Language	Mean	Median	Std Deviation
English	6.4	6.5	1.0
French	7.4	7.3	0.5
Spanish	7.1	7.2	0.8
Mandarin	7.0	7.3	0.7
Japanese	7.7	8.0	0.9
German	7.8	7.8	0.7
Italian	7.2	7.0	1.1
Cantonese	7.3	7.3	0.3
Portuguese	7.8	8.0	0.9
Korean	7.7	7.7	0.5
Hindi	7.2	7.4	0.7
Norwegian	7.2	7.3	0.5
Persian	8.1	8.4	0.4
Danish	7.9	8.1	0.4
Thai	6.6	6.6	0.4
Dutch	7.6	7.8	0.3
Aboriginal	7.0	7.0	0.6
Dari	7.5	7.5	0.1
Indonesian	7.9	7.9	0.3
Kazakh	6.0	6.0	0.0
Filipino	6.7	6.7	0.0
Romanian	7.9	7.9	0.0
Bosnian	4.3	4.3	0.0
Vietnamese	7.4	7.4	0.0
Hebrew	8.0	8.0	0.0

Czech	7.4	7.4	0.0
Maya	7.8	7.8	0.0
Russian	6.5	6.5	0.0
Mongolian	7.3	7.3	0.0
Aramaic	7.1	7.1	0.0
None	8.5	8.5	0.0
Zulu	7.3	7.3	0.0
Arabic	7.2	7.2	0.0
Hungarian	7.1	7.1	0.0

Data Analysis

d) Director Analysis:

Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

Top 10 Director Based on Average IMDB Score			
Director name	 Count of movies	Average imdb_score	Overall Contribution (%)
Sergio Leone	3	8.4	0.17%
Christopher Nolan	8	8.4	0.45%
Pete Docter	3	8.2	0.17%
Hayao Miyazaki	4	8.2	0.22%
Quentin Tarantino	8	8.2	0.44%
Milos Forman	3	8.1	0.16%
David Lean	4	8.0	0.22%
Frank Darabont	4	8.0	0.21%
Denis Villeneuve	3	8.0	0.16%
James Cameron	7	7.9	0.37%

- The best directors are Sergio Leone & Christopher Nolan having average IMDB score of 8.4.
- To make movies successful directors can collaborate with these top 10 directors.

Data Analysis

e) Budget Analysis:

Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit.

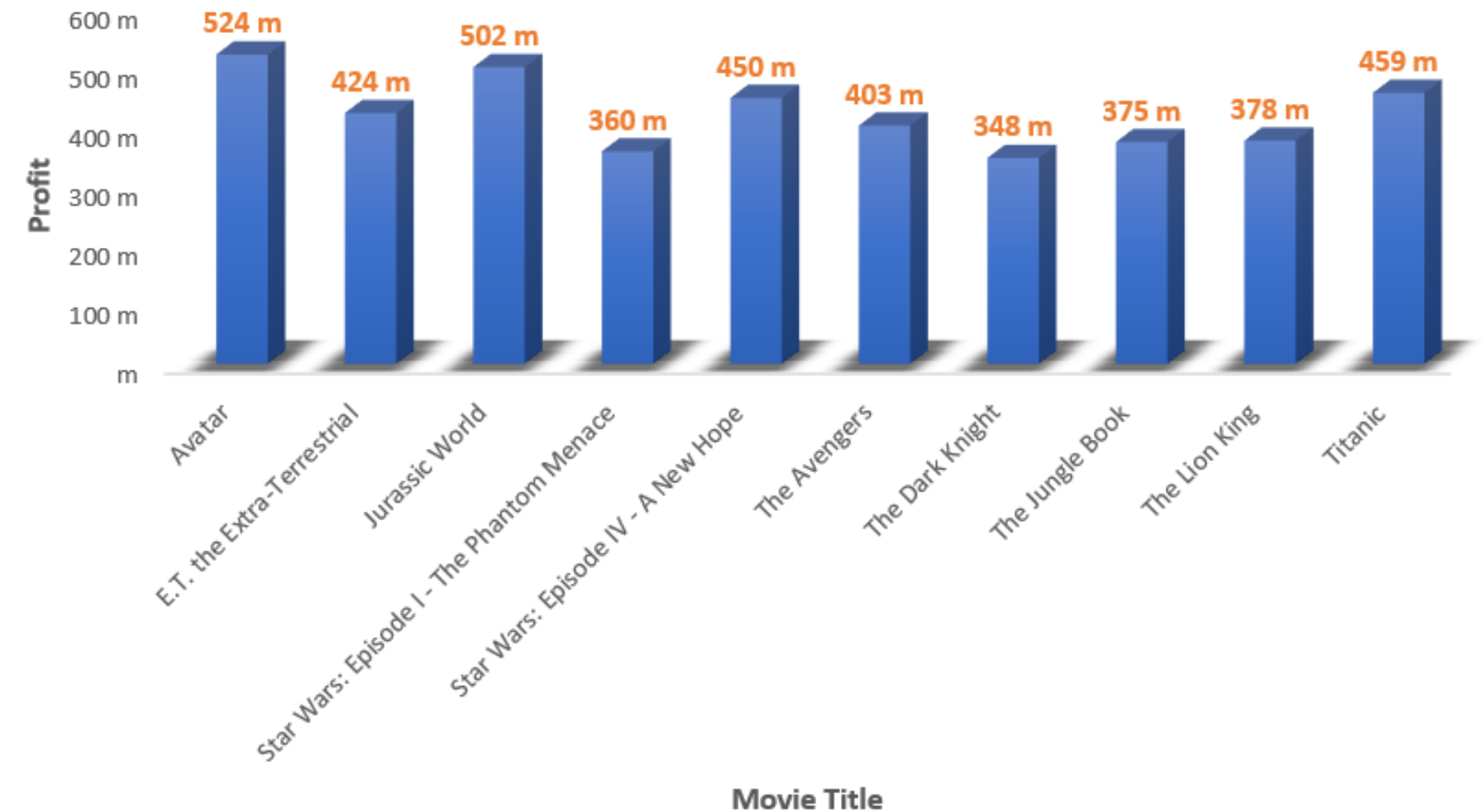
Top 10 Movies with Highest Profit

Movie Title	Profit
Avatar	523505847
E.T. the Extra-Terrestrial	424449459
Jurassic World	502177271
Star Wars: Episode I - The Phantom	359544677
Star Wars: Episode IV - A New Hope	449935665
The Avengers	403279547
The Dark Knight	348316061
The Jungle Book	375290282
The Lion King	377783777
Titanic	458672302

Correlation Coefficient

0.098318102

Top 10 Movies with Highest Profit



- Avatar is highest profitable movie having profit of 523 millions with budget of 237 millions.
- Correlation coefficient shows very weak correlation between budget and gross earning of movies. As long movie continues to provide best experience to viewers, gross earning of movies will increase.

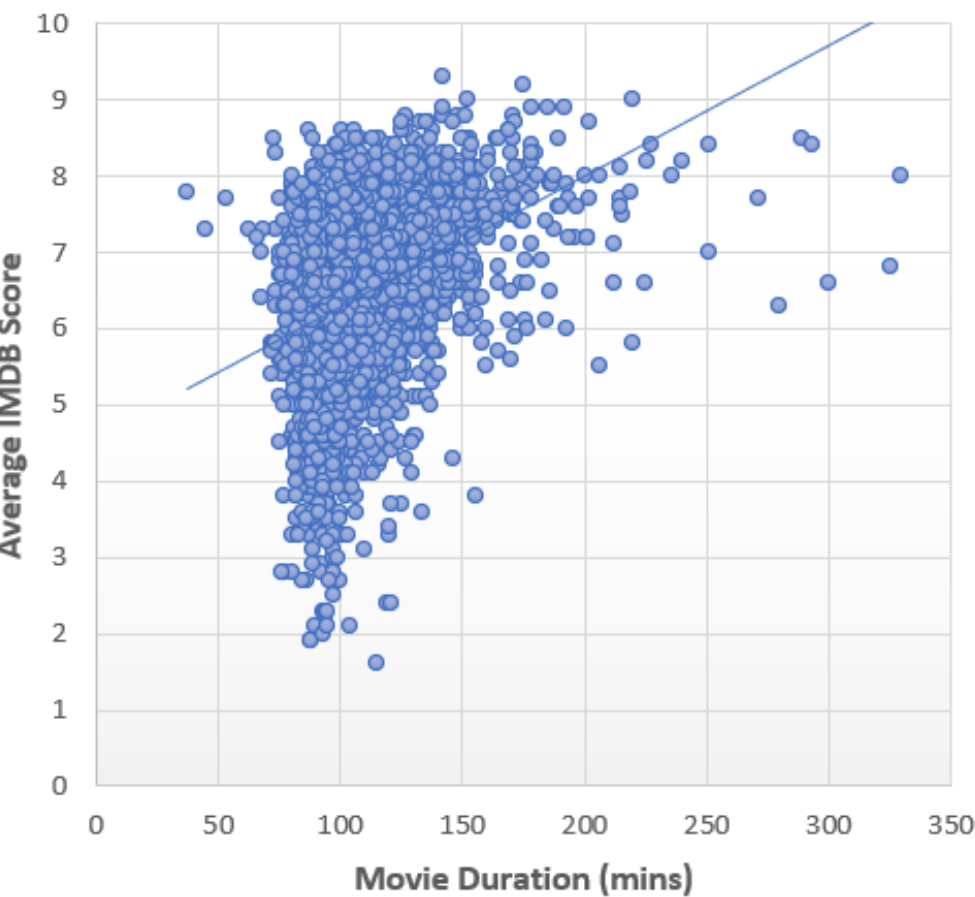
Interactive Dashboard:

IMDb IMDB Movie Analysis



Project By: Mayur Rajput

Movie Duration vs IMDB Score



Genre



Action

Adventure

Animation

Biography

Comedy

Crime

Title year



1927

1929

1933

1935

1936

1937

Country



Czech Republic

Denmark

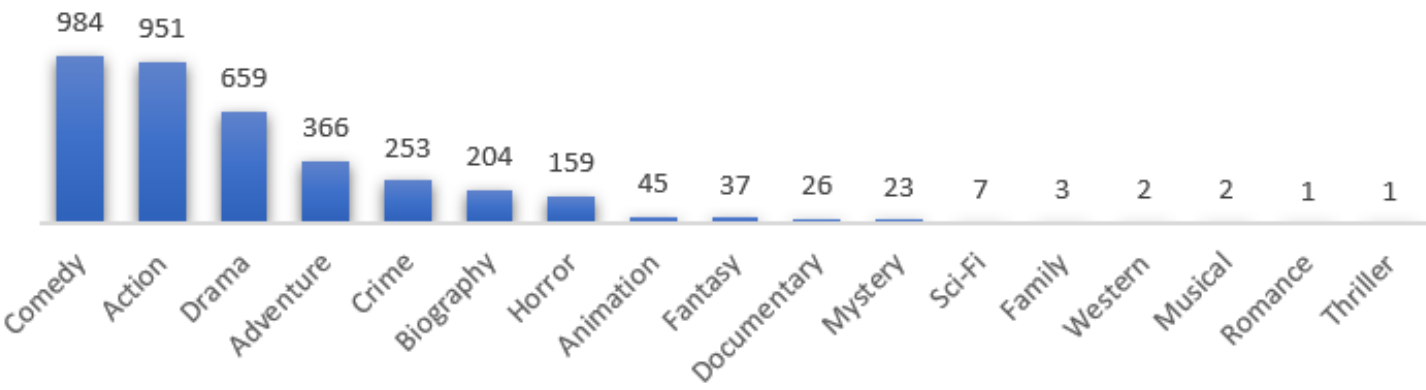
Finland

France

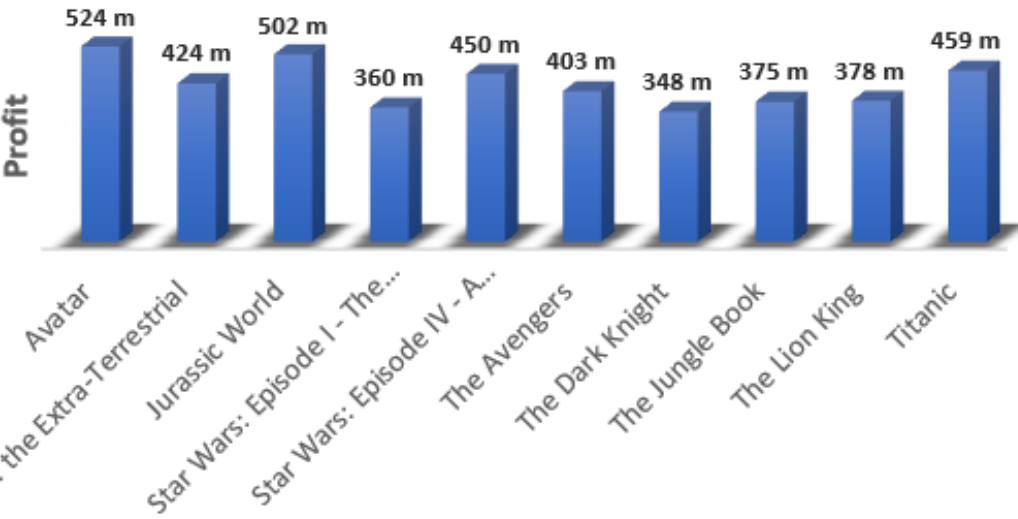
Georgia

Germany

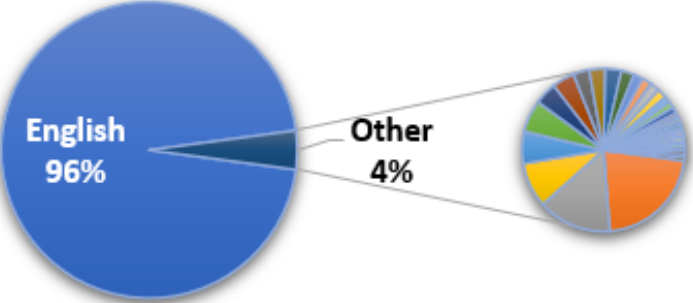
Genre Vs Count of Movies



Top 10 Movies with Highest Profit



Language Distribution



Insights :

1. The most common genres of movies are Comedy, Action and Drama. Movies in the Biography, Western and Romance genres tend to have higher average IMDb scores compared to other genres.
2. Most common duration range for movies is 50 – 150 mins.. Also Movie duration doesn't strongly influence its IMDB score.
3. Most no of movies produced are in English language. English language has low average IMDB score as compared other languages like Japanese, German and it can be due quality of movie produced in these language is better than English language.
4. The best directors are Sergio Leone & Christopher Nolan having average IMDB score of 8.4. To make movies successful directors can collaborate with these top 10 directors.
5. Avatar is highest profitable movie having profit of 523 millions with budget of 237 millions. Correlation coefficient shows very weak correlation between budget and gross earning of movies. As long movie continues to provide best experience to viewers, gross earning of movies will increase.

Results :

This project helped me to advance Excel skills and problem solving ability. Through this project I learned how to handle missing data and remove duplicates in data based on situation. Extensively worked on pivot table and charts which enabled me to give better representation of output in form of charts.