

Investigating IMBD Dataset

Overview

This paper is dedicated to find insights and interesting facts in the IMDB Data through simple Descriptive statistics and plotting visualizations

IMDB Dataset:

IMDB Data set is a csv file containing movies data from IMDB website from 1960 to 2015 with 21 columns and 10 thousand rows, it includes almost everything about the movie, from the movie overview to its most popular tagline, and it also contains the adjusted budget and revenue to account for the inflation.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  -
0   id                     10866 non-null  int64
1   imdb_id               10866 non-null  object
2   popularity             10866 non-null  float64
3   budget                10866 non-null  int64
4   revenue               10866 non-null  int64
5   original_title        10866 non-null  object
6   cast                  10790 non-null  object
7   homepage              2936 non-null  object
8   director              10822 non-null  object
9   tagline                8042 non-null  object
10  keywords              9373 non-null  object
11  overview              10862 non-null  object
12  runtime               10866 non-null  int64
13  genres                10843 non-null  object
14  production_companies  9836 non-null  object
15  release_date          10866 non-null  object
16  vote_count            10866 non-null  int64
17  vote_average          10866 non-null  float64
18  release_year          10866 non-null  int64
19  budget_adj            10866 non-null  float64
20  revenue_adj           10866 non-null  float64
dtypes: float64(4), int64(6), object(11)
memory usage: 1.7+ MB
```

```
0           The park is open.
1           What a Lovely Day.
2           One Choice Can Destroy You
3   Every generation has a story.
4           Vengeance Hits Home
Name: tagline, dtype: object
```

```
0   monster|dna|tyrannosaurus rex|velociraptor|island
1   future|chase|post-apocalyptic|dystopia|australia
2   based on novel|revolution|dystopia|sequel|dyst...
3   android|spaceship|jedi|space opera|3d
4   car race|speed|revenge|suspense|car
Name: keywords, dtype: object
```

Posed Questions:

- What is the correlation between genre , rating and popularity
- How the popularity ratings density curve changed over generation
- Assuming the movie popularity is related to the director popularity, what are the most popular directors in each generation and the movie they directed and what is the tagline of the movie
- Is there a correlation between popularity and rating
- Did movie revenue change overtime

Exploratory data analysis and visualization:

- Visualize summary statistics
- Relation between genre and popularity
- Relation between genre and rating
- Relation between genre and budget
- Relation between genre and revenue
- Most popular movies director names 70s
- Most popular movies director names 90s
- Most popular movies director names 2000s
- Plotting the previous three decades together
- Relation between popularity and rating considering genres
- Visualizing Revenue each year

Conclusion:

From the summary statistics it is clear that most of the movies are between 0.2 and 0.7 in popularity with outliers reaching up to 32 in popularity. Most of the globally recognized movies are above 2 in popularity but not necessarily high in rating and most all of them are outliers.

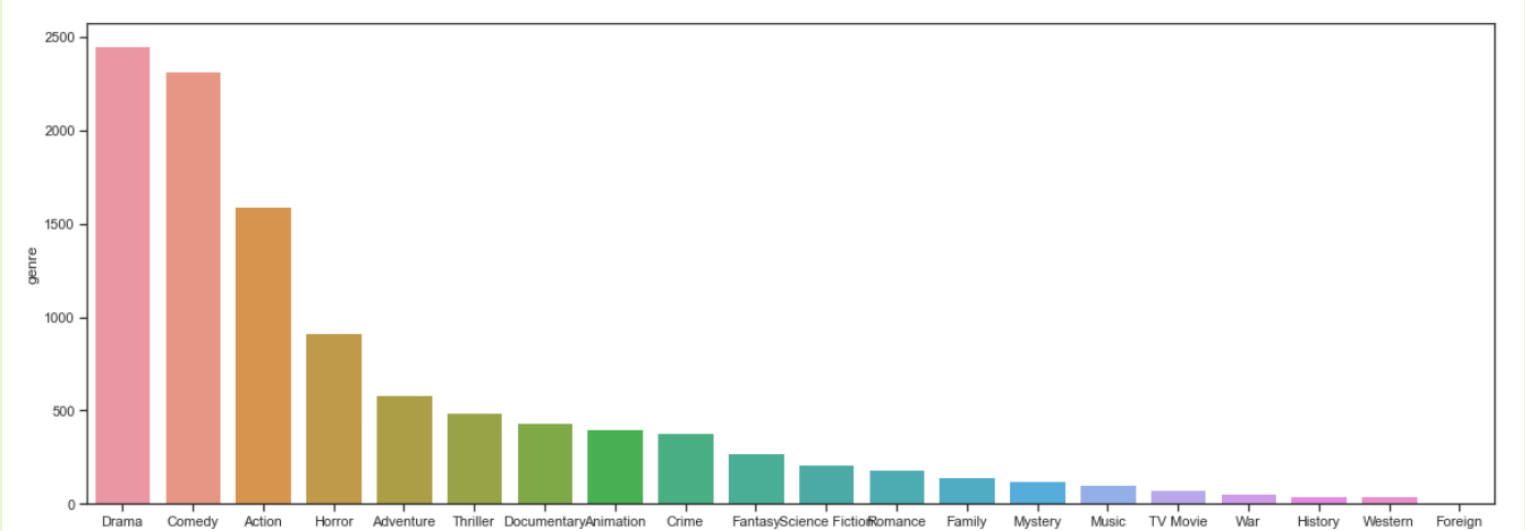
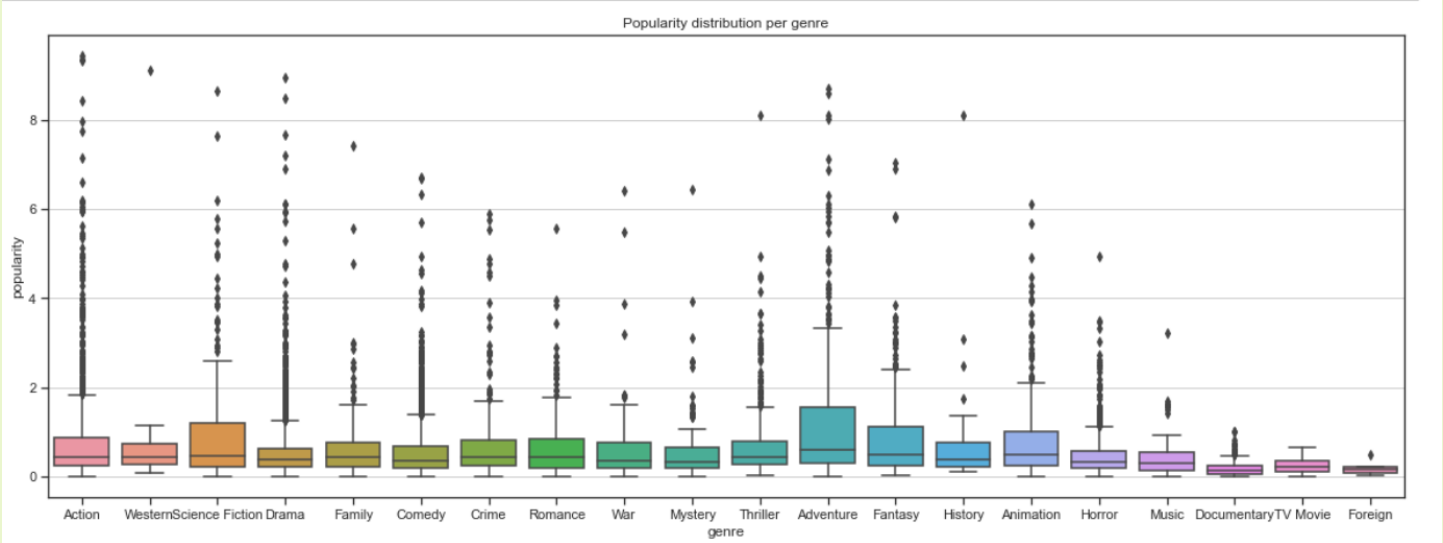
There are a few number movies before the 80s.

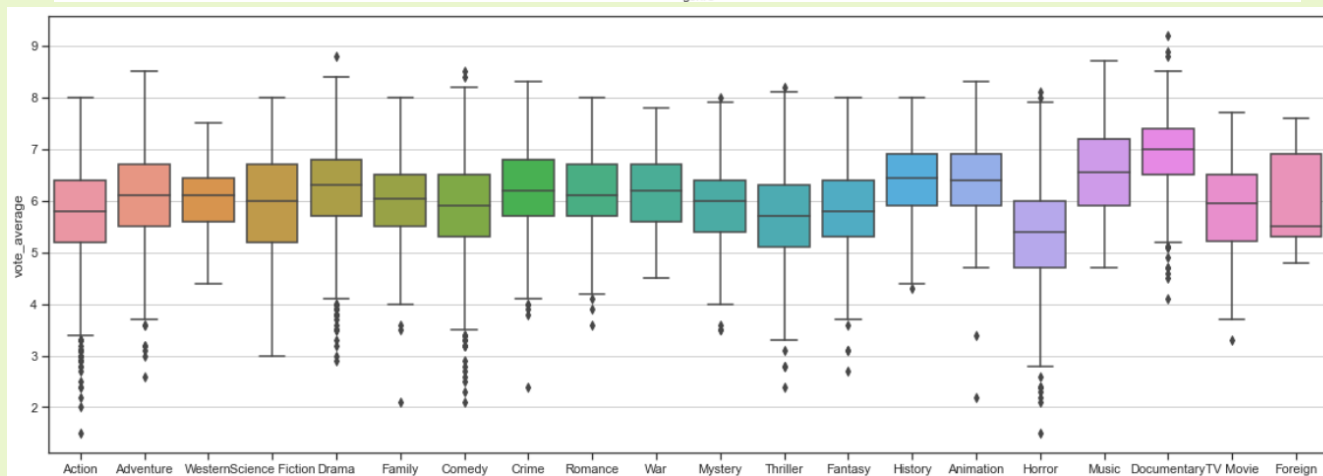
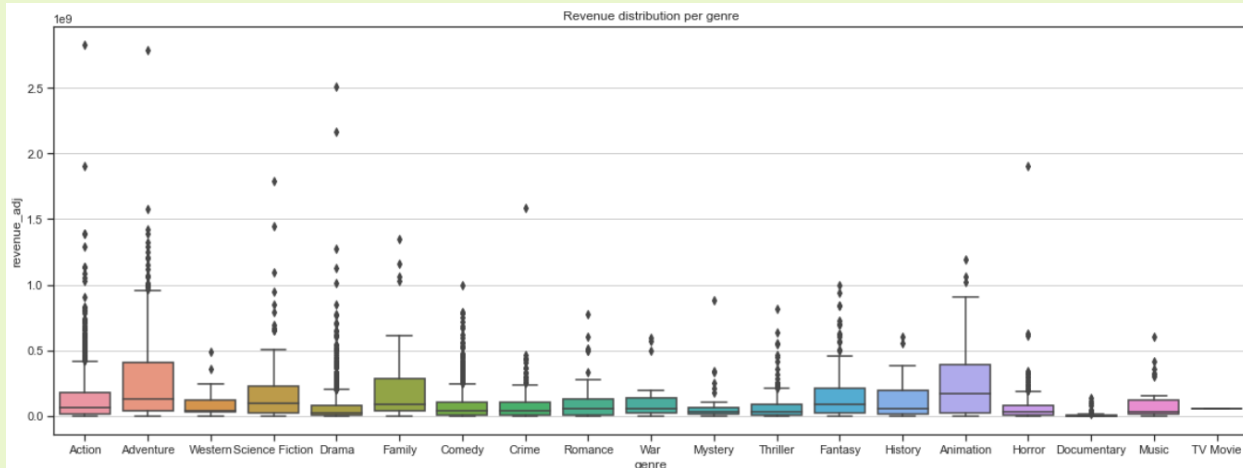
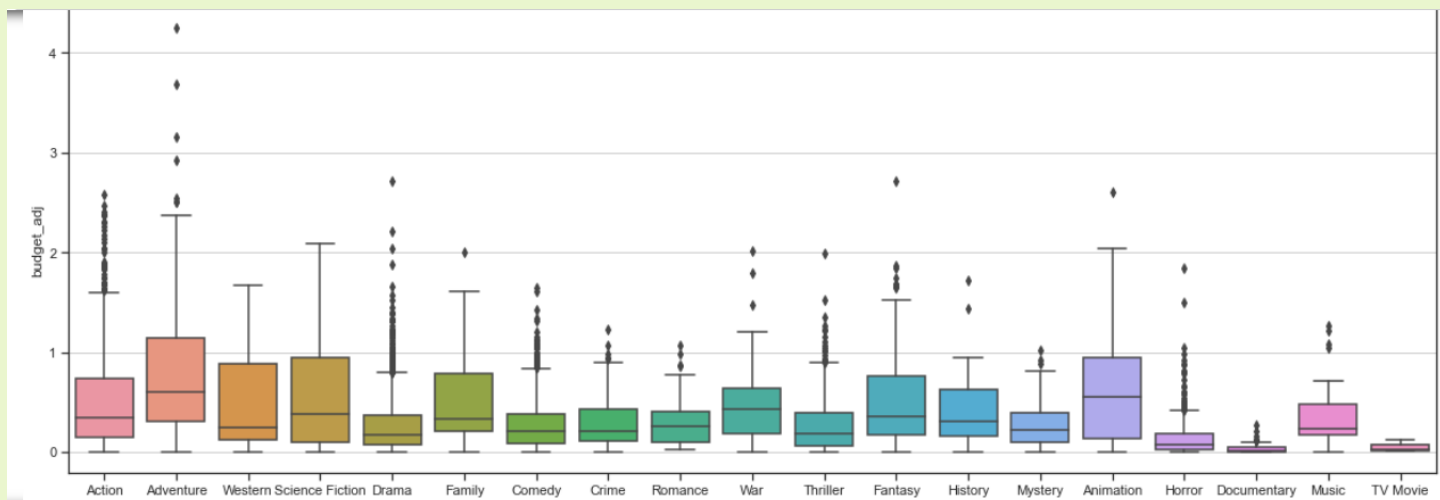
But most of the movies across time are under 2 in popularity. Top popular genres are Adventure, Science Fiction, Fantasy and Action. There are lots of unpopular action movies. First name appear genre is mostly drama, comedy and action respectively. Most of Action, comedy, Thriller, Fantasy and Horror movies are under 5.5 rating with horror movies on the top of this list.

Approximately 50% of the Drama genre is above 6.5 in Rating which is impressive regarding its frequent presence in the genres' first name, and also not having a very high budget, the Highest budget genre that stands out is the adventure genre but also high in revenue The revenue is not very reliable as there are a lot of missing data, but from the visualization it appears that the highest revenue was at the 70s however, if you took a look at the two plots in the 'Visualizing each year revenue' section you will observe a strange phenomenon, it appears in the line plot that the revenue increased as time goes on, which may be true, but looking at the boxplot, there are clearly a lot of outliers as the years approach 2015 and with very high revenue values which can be misleading.

Summary Statistics and Visualizations:

	popularity	runtime	vote_count	vote_average	release_year	budget_adj	revenue_adj
count	10866.000000	10866.000000	10866.000000	10866.000000	10866.000000	1.086600e+04	1.086600e+04
mean	0.646441	102.070863	217.389748	5.974922	2001.322658	1.755104e+07	5.136436e+07
std	1.000185	31.381405	575.619058	0.935142	12.812941	3.430616e+07	1.446325e+08
min	0.000065	0.000000	10.000000	1.500000	1960.000000	0.000000e+00	0.000000e+00
25%	0.207583	90.000000	17.000000	5.400000	1995.000000	0.000000e+00	0.000000e+00
50%	0.383856	99.000000	38.000000	6.000000	2006.000000	0.000000e+00	0.000000e+00
75%	0.713817	111.000000	145.750000	6.600000	2011.000000	2.085325e+07	3.369710e+07
max	32.985763	900.000000	9767.000000	9.200000	2015.000000	4.250000e+08	2.827124e+09





	director	original_title	vote_average	popularity	tagline
1386	James Cameron	Avatar	7.1	9.432768	Enter the World of Pandora.
1387	Quentin Tarantino	Inglourious Basterds	7.7	6.098027	Once upon a time in Nazi occupied France...
1388	Patrick Tatopoulos	Underworld: Rise of the Lycans	6.2	5.806897	Every war has a beginning.
1389	David Yates	Harry Potter and the Half-Blood Prince	7.3	5.076472	Dark Secrets Revealed
1390	Bob Peterson Pete Docter	Up	7.6	4.908902	Fly Up to Venezuela
...
7394	Sam Raimi	Spider-Man 3	5.8	2.520912	The battle within.
7395	Frank Darabont	The Mist	6.6	2.510571	Belief divides them, mystery surrounds them, b...
8661	Ridley Scott	Gladiator	7.7	4.271452	A Hero Will Rise.
8662	Christopher Nolan	Memento	7.9	3.095625	Some memories are best forgotten.
8663	Ron Howard	How the Grinch Stole Christmas	6.1	2.977558	He puts the mean in green.

	director	original_title	vote_average	popularity	tagline
1329	George Lucas	Star Wars	7.9	12.037933	A long time ago in a galaxy far, far away...
1330	Lewis Gilbert	The Spy Who Loved Me	6.2	2.379469	It's the BIGGEST. It's the BEST. It's BOND. An...
1331	John Lounsbery Wolfgang Reitherman Art Stevens	The Rescuers	6.6	1.719385	Two tiny agents vs. the world's wickedest woma...
7269	Francis Ford Coppola	The Godfather	8.3	5.738034	An offer you can't refuse.
7825	Ridley Scott	Alien	7.6	4.935897	In space no one can hear you scream.
7826	Francis Ford Coppola	Apocalypse Now	7.8	2.319520	This is the end...
7827	George Miller	Mad Max	6.5	2.012021	The Maximum Force Of The Future
7828	Robert Wise	Star Trek: The Motion Picture	6.0	1.823058	The human adventure is just beginning.
7829	Lewis Gilbert	Moonraker	5.7	1.691143	Outer space now belongs to 007.
9758	Francis Ford Coppola	The Godfather: Part II	8.1	3.264571	I don't feel I have to wipe everybody out, Tom...
9759	Roman Polanski	Chinatown	7.7	1.650848	NaN
9760	Guy Hamilton	The Man with the Golden Gun	6.1	1.619555	The man with the golden gun is ready to assass...
9805	MiloÅi Forman	One Flew Over the Cuckoo's Nest	8.0	3.258151	If he's crazy, what does that make you?
9806	Steven Spielberg	Jaws	7.3	2.563191	Don't go in the water.
9807	Terry Gilliam Terry Jones	Monty Python and the Holy Grail	7.6	2.234953	And now! At Last! Another film completely diff...
9923	Stanley Kubrick	A Clockwork Orange	7.7	3.072555	Being the adventures of a young man whose prin...
9924	Mel Stuart	Willy Wonka & the Chocolate Factory	7.2	2.430667	It's Scrumdiddlyumptious!

	director	original_title	vote_average	popularity	tagline
2409	David Fincher	Fight Club	8.1	8.947905	How much can you know about yourself if you've...
2410	Lilly Wachowski Lana Wachowski	The Matrix	7.8	7.753899	Welcome to the Real World.
2411	Sam Mendes	American Beauty	7.7	3.555720	Look closer.
2412	George Lucas	Star Wars: Episode I - The Phantom Menace	6.3	3.526029	Every generation has a legend. Every journey h...
2413	Stephen Sommers	The Mummy	6.3	2.943770	The legend you know. The adventure you have ye...
2414	Frank Darabont	The Green Mile	8.0	2.719461	Miracles do happen.
2415	M. Night Shyamalan	The Sixth Sense	7.5	2.578731	Not every gift is a blessing.
4177	Quentin Tarantino	Pulp Fiction	8.1	8.093754	Just because you are a character doesn't mean ...
4178	Frank Darabont	The Shawshank Redemption	8.4	7.192039	Fear can hold you prisoner. Hope can set you f...
4179	Robert Zemeckis	Forrest Gump	8.1	6.715966	The world will never be the same, once you've ...
4180	Roger Allers Rob Minkoff	The Lion King	7.7	4.782688	Life's greatest adventure is finding your plac...

