TMDB Movie Data Analysis

In this report, some findings related to the analysis of TMDB dataset are shared. Some questions were posed and their answers were found using EDA. This dataset contains the data of more than 10,000 movies from 1960 to 2015, such as revenue, budget, genre, cast, director and more.

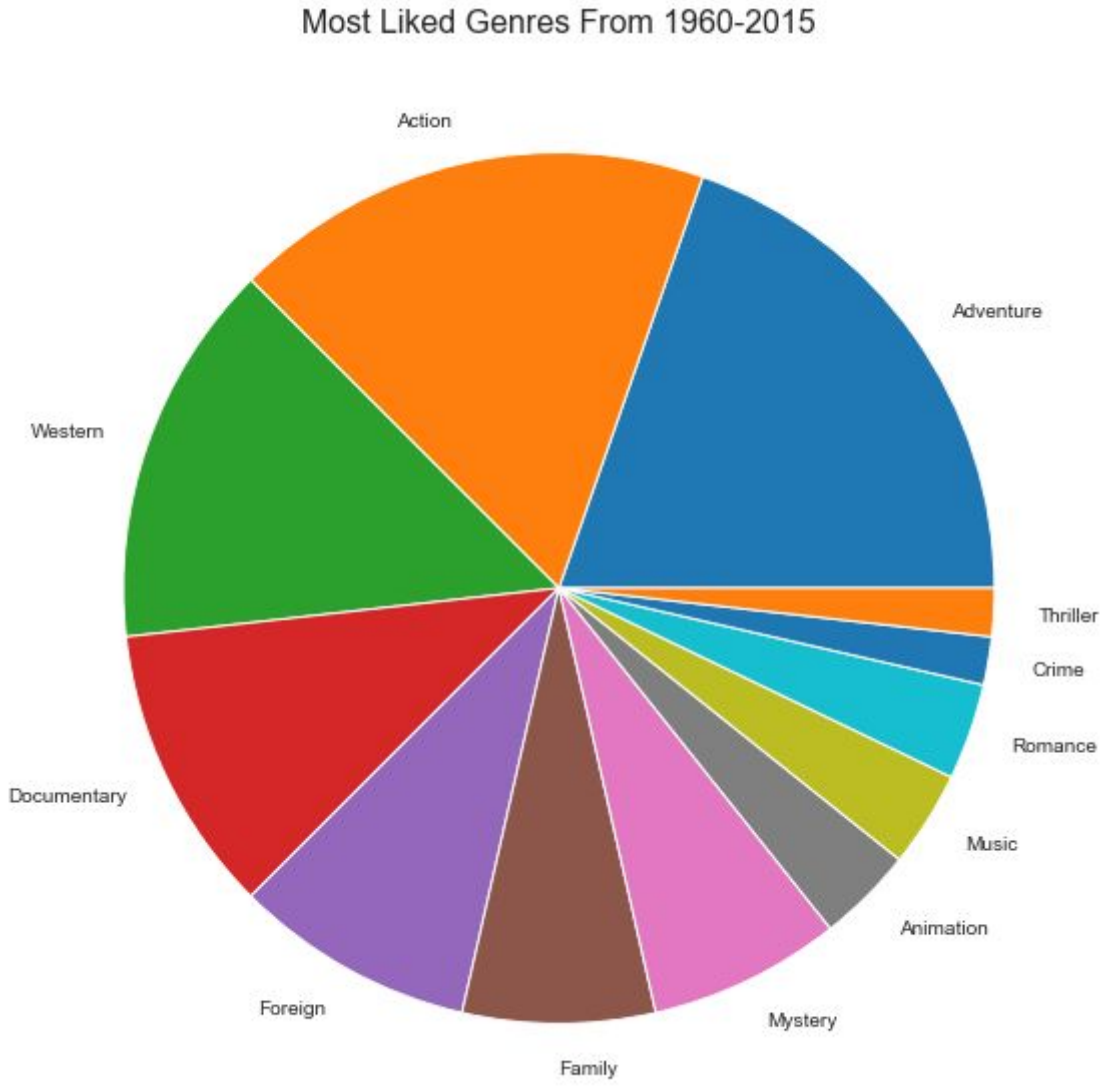
Question 1: Which genres are most popular from year to year?

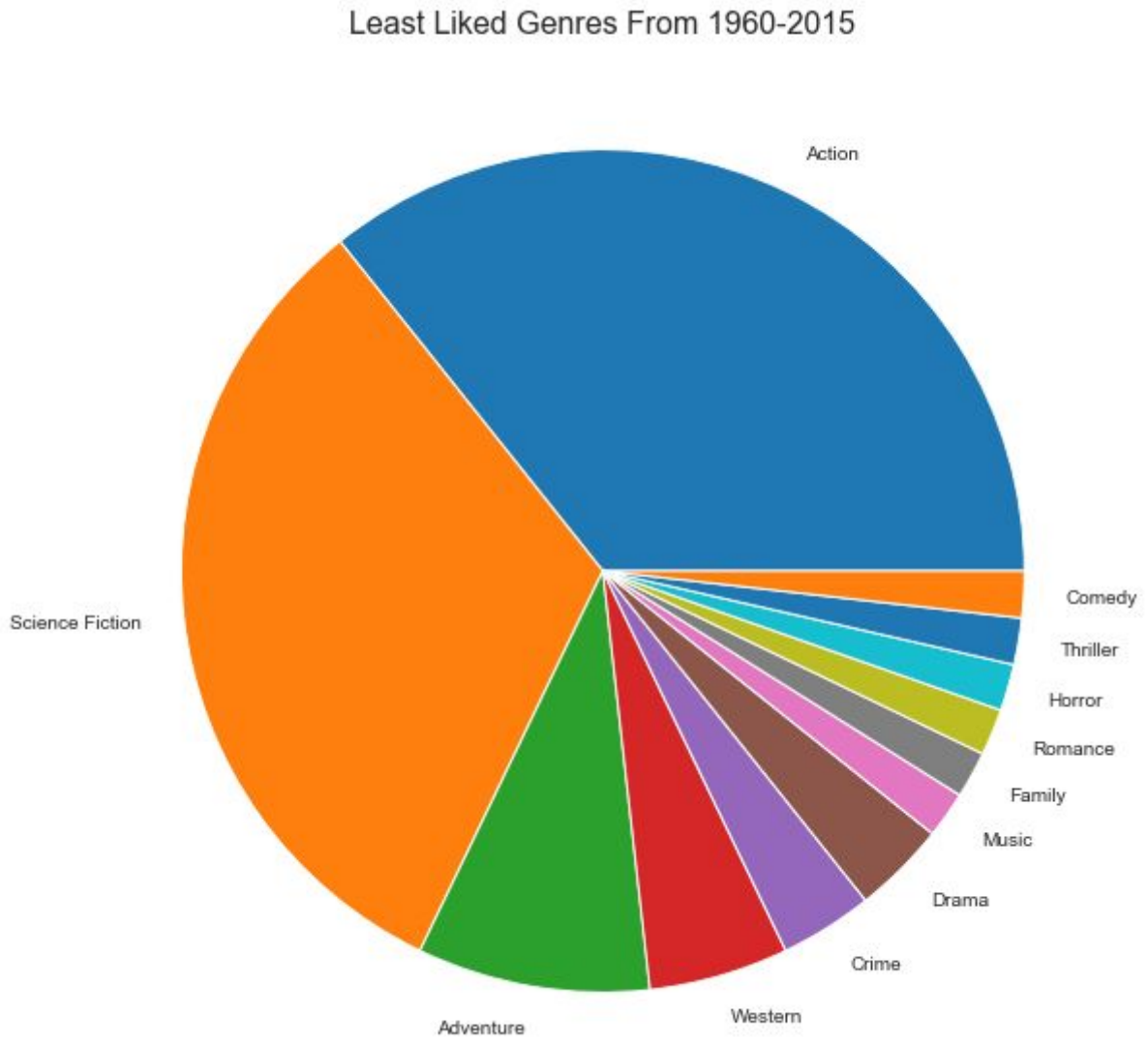
To find the most popular genre from year to year, the average ratings for each genre were calculated and the following table shows the most popular genre in each year.

Year	Best Genre	Worst Genre	Year	Best Genre	Worst Genre
1960	Adventure	Action	**1988**	Adventure	Science Fiction
1961	Mystery	Action	**1989**	Western	Thriller
1962	Western	Adventure	**1990**	Western	Action
1963	Adventure	Romance	**1991**	Foreign	Action
1964	Action	Science Fiction	**1992**	Western	Science Fiction
1965	Crime	Western	**1993**	Documentary	Science Fiction
1966	Animation	Science Fiction	**1994**	Documentary	Science Fiction
1967	Mystery	Science Fiction	**1995**	Mystery	Action
1968	Adventure	Action	**1996**	Foreign	Family
1969	Adventure	Action	**1997**	Thriller	Action
1970	Western	Crime	**1998**	Mystery	Science Fiction
1971	Action	Adventure	**1999**	Foreign	Science Fiction
1972	Action	Science Fiction	**2000**	Adventure	Science Fiction
1973	Romance	Adventure	**2001**	Documentary	Science Fiction
1974	Documentary	Action	**2002**	Western	Action
1975	Adventure	Action	**2003**	Western	Action
1976	Documentary	Action	**2004**	Foreign	Western
1977	Action	Music	**2005**	Foreign	Adventure
1978	Action	Science Fiction	**2006**	Action	Comedy
1979	Music	Science Fiction	**2007**	Action	Horror
1980	Action	Western	**2008**	Documentary	Action

1981	Music	Science Fiction	**2009**	Adventure	Crime
1982	Animation	Science Fiction	**2010**	Adventure	Action
1983	Family	Science Fiction	**2011**	Western	Action
1984	Romance	Action	**2012**	Adventure	Science Fiction
1985	Family	Adventure	**2013**	Adventure	Action
1986	Family	Action	**2014**	Action	Drama
1987	Action	Drama	**2015**	Family	Action

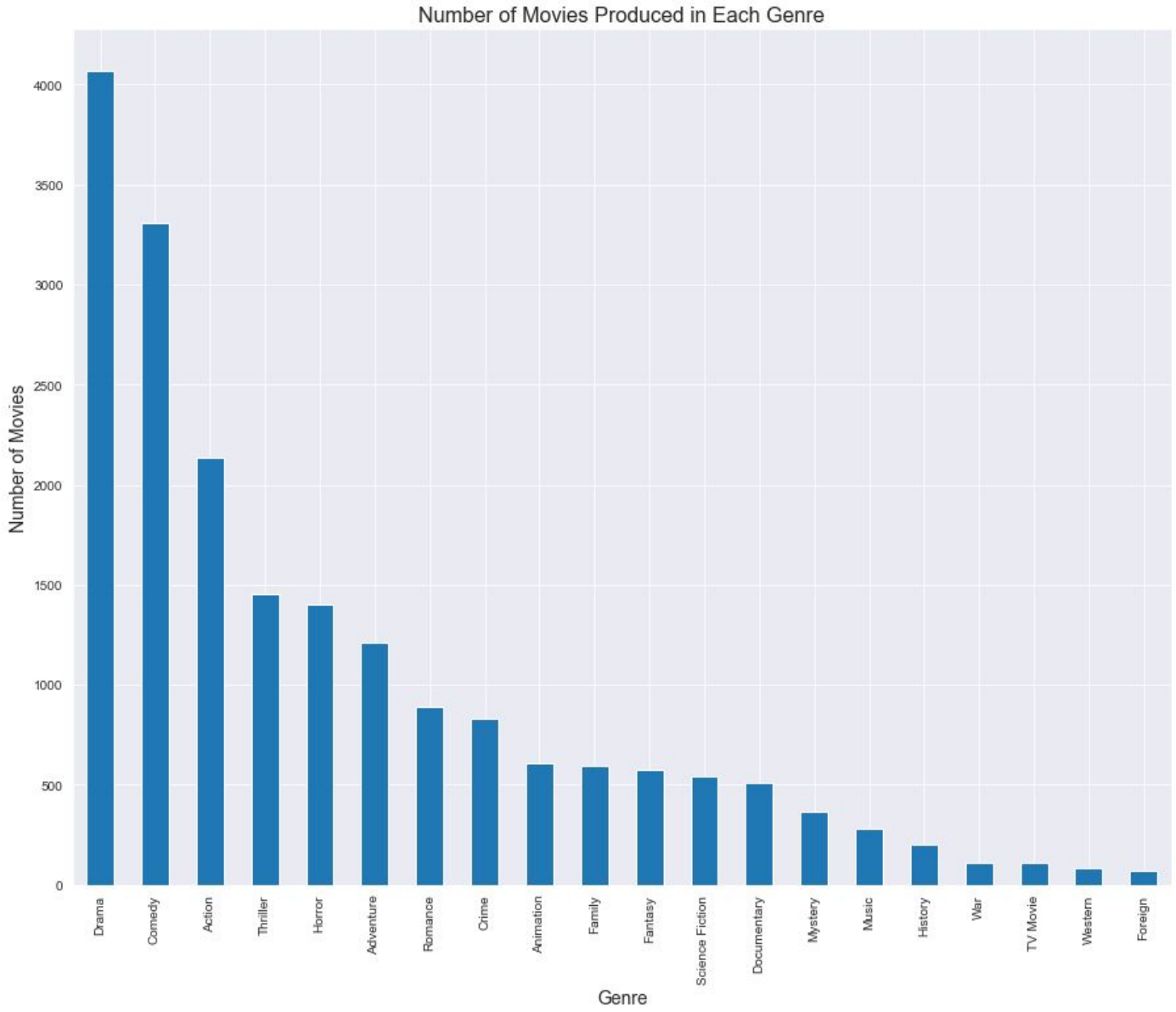
The following pie charts show which genres have dominated over the years.





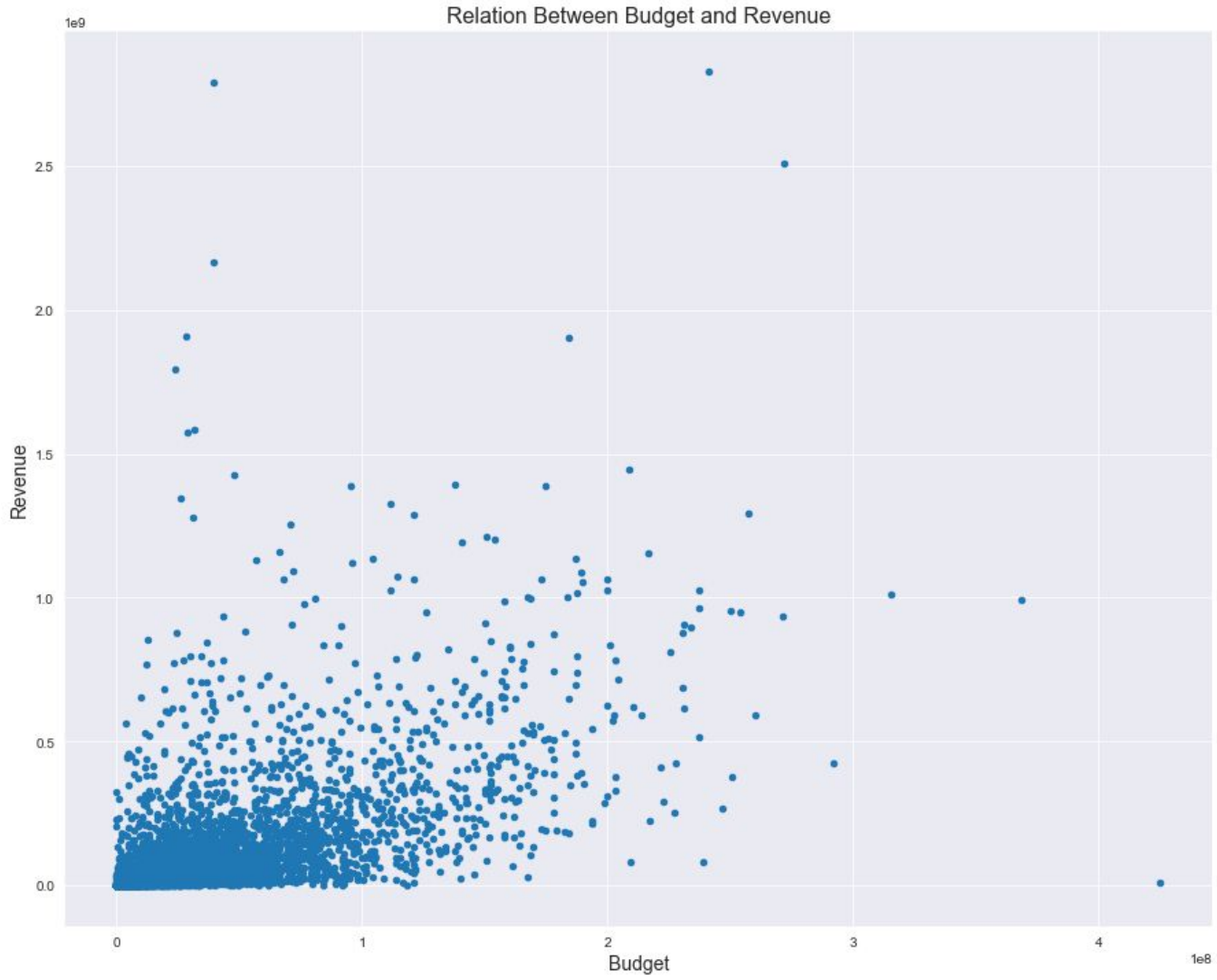
The above pie charts show that the ‘’Action’’ genre has been most liked and least liked, this makes sense because this particular genre is the third most produced genre from 1960 to 2015. Equal to ‘’Action’’ is ‘’Adventure’’ genre. One could argue that ‘’Adventure’’ movies are far more liked than ‘’Action’’ movies, because when looking at the number of movies produced in each genre, ‘’Adventure’’ is almost half the number of movies produced when compared to ‘’Action’’ movies.

A bar chart showing the number of movies produced in each genre in the next page.


Number of Movies Produced in Each Genre
Number of Movies
0
500
1000
1500
2000
2500
3000
3500
4000
Drama
Comedy
Action
Thriller
Horror
Adventure
Romance
Crime
Animation
Family
Fantasy
Science Fiction
Documentary
Mystery
Music
History
War
TV Movie
Western
Foreign
Genre

Question 2: What kinds of properties are associated with movies that have high revenues?

To find the properties associated with high revenue, mind immediately goes to how much money is spent on the production of the movie.
Here is a scatter plot between budget and revenue:



The correlation between those two variables equals to 0.57, so the amount of revenue a movie yields is somewhat related to the amount of money spent on the production of the movie.

Next step was to classify the movies into four categories:

- Low revenue
- Below average revenue
- Above average revenue
- High revenue

After classifying the movies, the average of each variable was calculated for each variable. This also shows which actor and director is associated with high revenue.

Revenue Bin	Average Runtime	Average Rate	Average Budget	FrequentActor	Frequent Director	Frequent Genre	Production Company
Low	103.22	5.96	16,000,828	Willem Dafoe	Richard Linklater	Drama	Warner Bros.
Below Average	106.40	6.07	29,470,726	Robert De Niro	Wes Craven	Drama	Universal Pictures
Above Average	109.60	6.19	44,490,498	Robert De Niro	Clint Eastwood	Drama	Warner Bros.
High	117.66	6.46	86,992,080	Tom Cruise	Steven Spielberg	Action	Warner Bros.

- The above table shows that “Tom Cruise” is the actor associated with high revenue, while “Steven Spielberg” is the director associated with high revenue.

- When it comes to the production companies, analysis shows that “Warner Bros” produces all kinds of movies, popular and unpopular. Also, “Action” movies yield the highest revenues.

- Runtime average for all movie revenue classifications is not that far from each other so it is not considered a factor.

- Low budget movies are not popular at all and it is shown that the average budget of a high revenue is almost five times more than that of a low revenue movie, and it is almost double the average budget of a above average revenue movie.

Data Wrangling Report

Data Quality Issues

- While exploring the data set it was noticed that some columns where not needed so using pandas drop method. The following columns were dropped:
    - Imdb_id
    - homepage
    - tagline
    - keywords
    - overview

- Release date column had a type string, so it was changed to datatime type using pandas to_datetime function.

- Checked for duplicate rows, and found one duplicate row so it was dropped using pandas drop_duplicates method.

- Replaced ‘0’ in [“budget”, “revenue”, “budget_adj”, “revenue_adj”, “runtime”] with NaN using pandas replace method.

Data Tidiness Issues

- Cast, production_companies and genres columns had multiple values in each column. These values where separated using the ‘|’ delimiter. The converted into lists which used to create a dataframe for each column and those dataframes were later concatenated to the main dataframe and columns with lists as data were dropped.