IMDB Data Analysis

By: Sakib Chughtai

Introduction - Background

- Global Box Office Valuation: $42.2 billion in 2019
- Top 3 continents by box office gross:
  - Asia Pacific - $17.8 Billion
  - US/Canada - $11.4 Billion
  - Europe, Middle East, North Africa - $10.3 Billion
- Main revenue streams
  - Theatrical Exhibition
  - Television broadcast rights
  - Home Video
- Theatrical Exhibition is the primary metric for assessing the success of a film, due to data availability.

Rank	Title	Gross ($)	Year
1	Avatar	2,923,847,066	2009
2	Avengers: Endgame	2,797,501,328	2019
3	Titanic	2,187,535,296	1997
4	Star Wars: The Force Awakens	2,068,223,624	2015
5	Avengers: Infinity War	2,048,359,754	2018
6	Spiderman: No Way Home	1,920,544,470	2021
7	Jurassic World	1,671,537,444	2015
8	The Lion King	1,656,943,394	2019
9	The Avengers	1,518,812,988	2012
10	Furious 7	1,516,045,911	2015

Business Understanding

- **Key Stakeholders:** Film and TV production companies looking to create and add offerings for customers.
- Other Stakeholders: Data Enthusiasts/Engineers looking to perform same analysis for other types of video entertainment such as Tv Shows.
- **Dataset** – excel file with 5000 IMDB movies with various features
- Next step: Understand and break down which features have an impact on the success metric (IMDB rating).

Analytics Approach

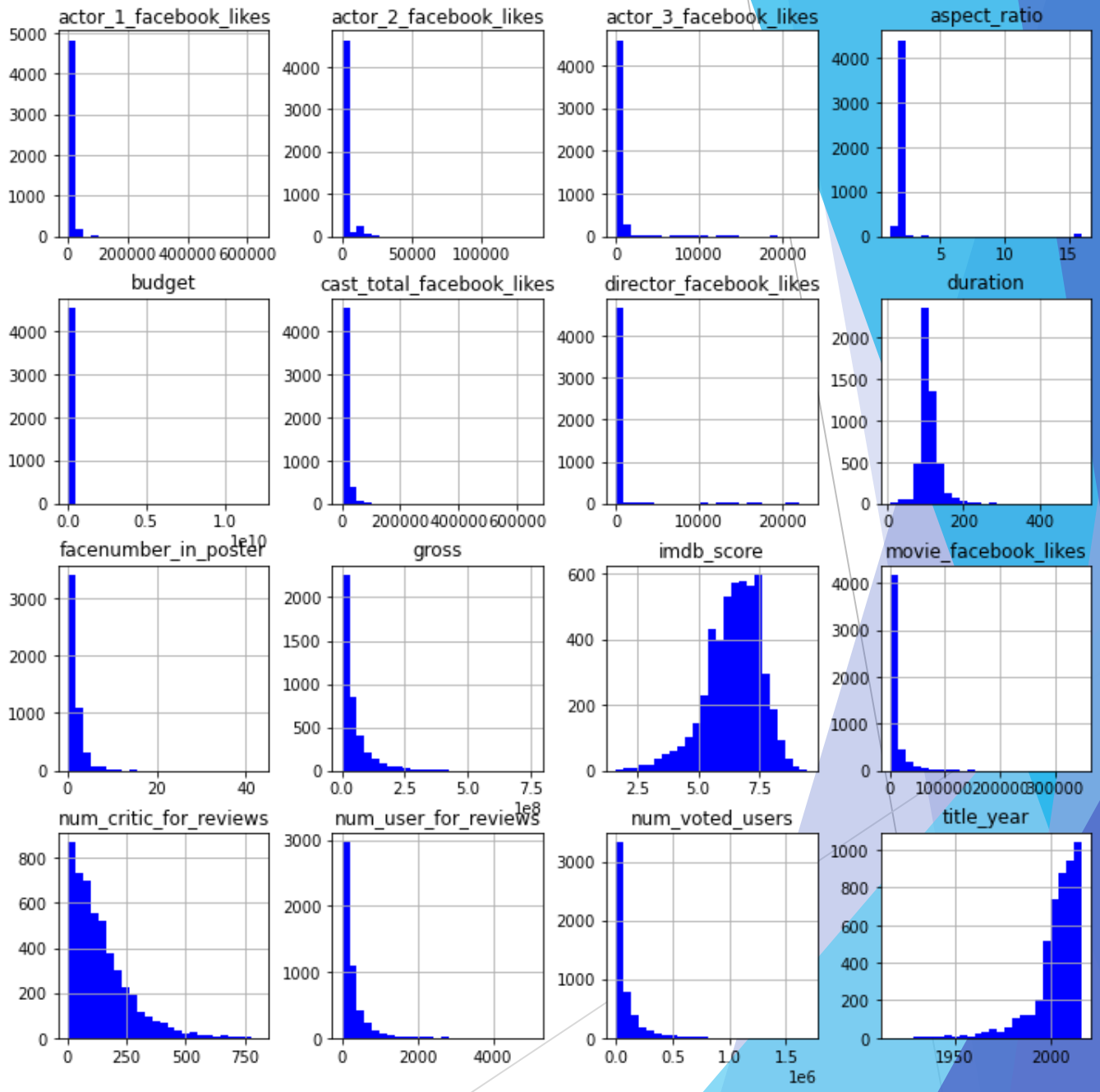
- Since we are predicting how successful a movie is by analysing historical data, we will use a **Descriptive Model**.
- Descriptive analytics provide a vision into the past and tells us what has happened and why.
- We will use Visualization tools to gain understanding of the dataset.
- **Linear Regression** – Using different features (X) to find the impact on our dependant variable Y (IMDB rating)
- **XGBoost** – High prediction performance for regression
- Regression Metrics: Mean Squared Error and Mean Absolute Error.

Tools – Python Libraries

- **Coding Environment:** Jupyter Notebook
- **Pandas, NumPy** - Data Cleaning and Manipulation
- **Matplotlib, Seaborn -** Data Visualization
- **Sci-kit Learn** – Data Modelling and Evaluation
- **XGBoost** – Data Modelling

Data Preparation

- Import necessary libraries
- Import excel dataset using pd.read_excel command
- Print dataframe shape to understand how data is structured
- Drop duplicates
- Remove redundant columns
- Use the split function to split up columns with multiple genres, plot keywords

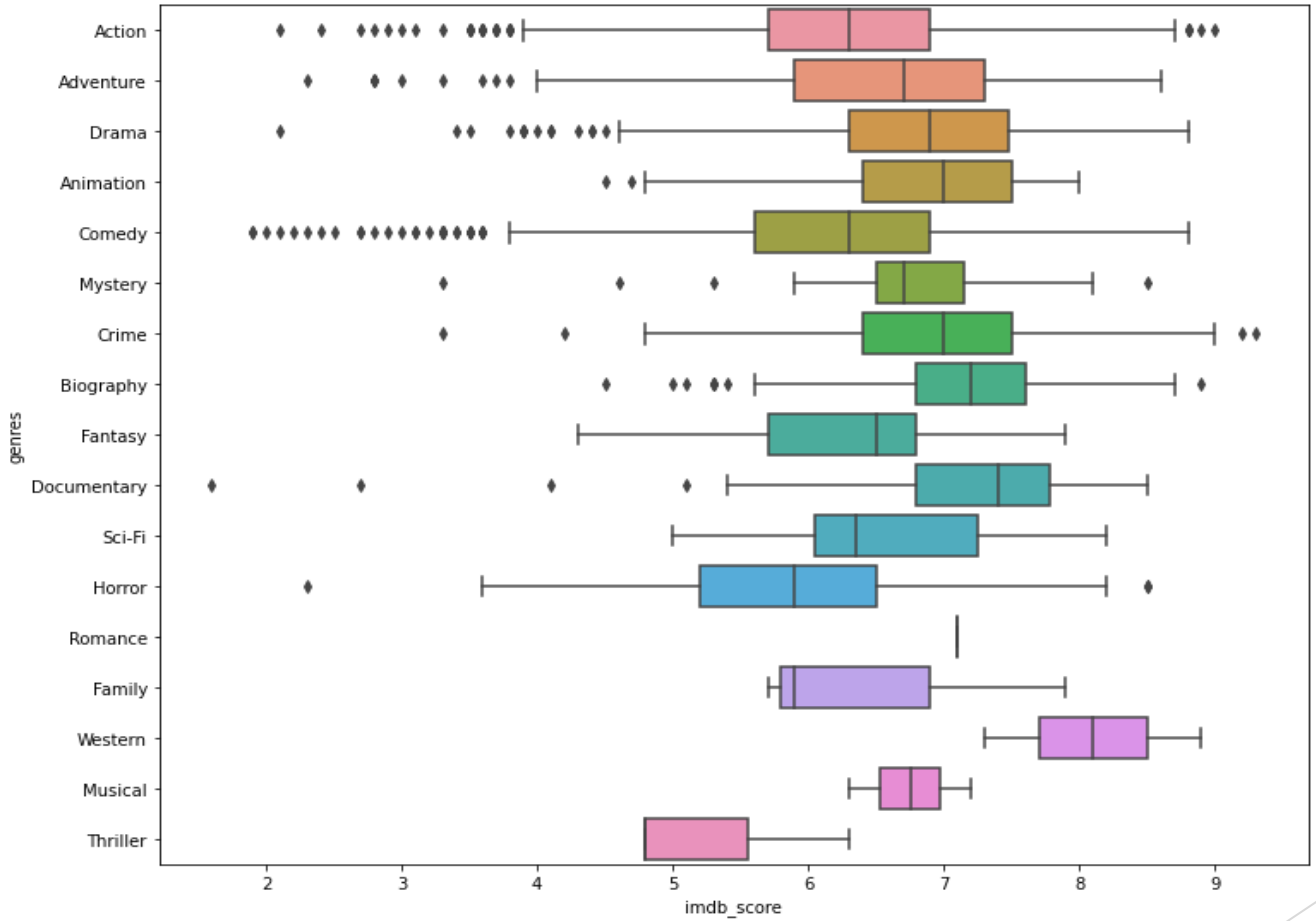


Data Preparation

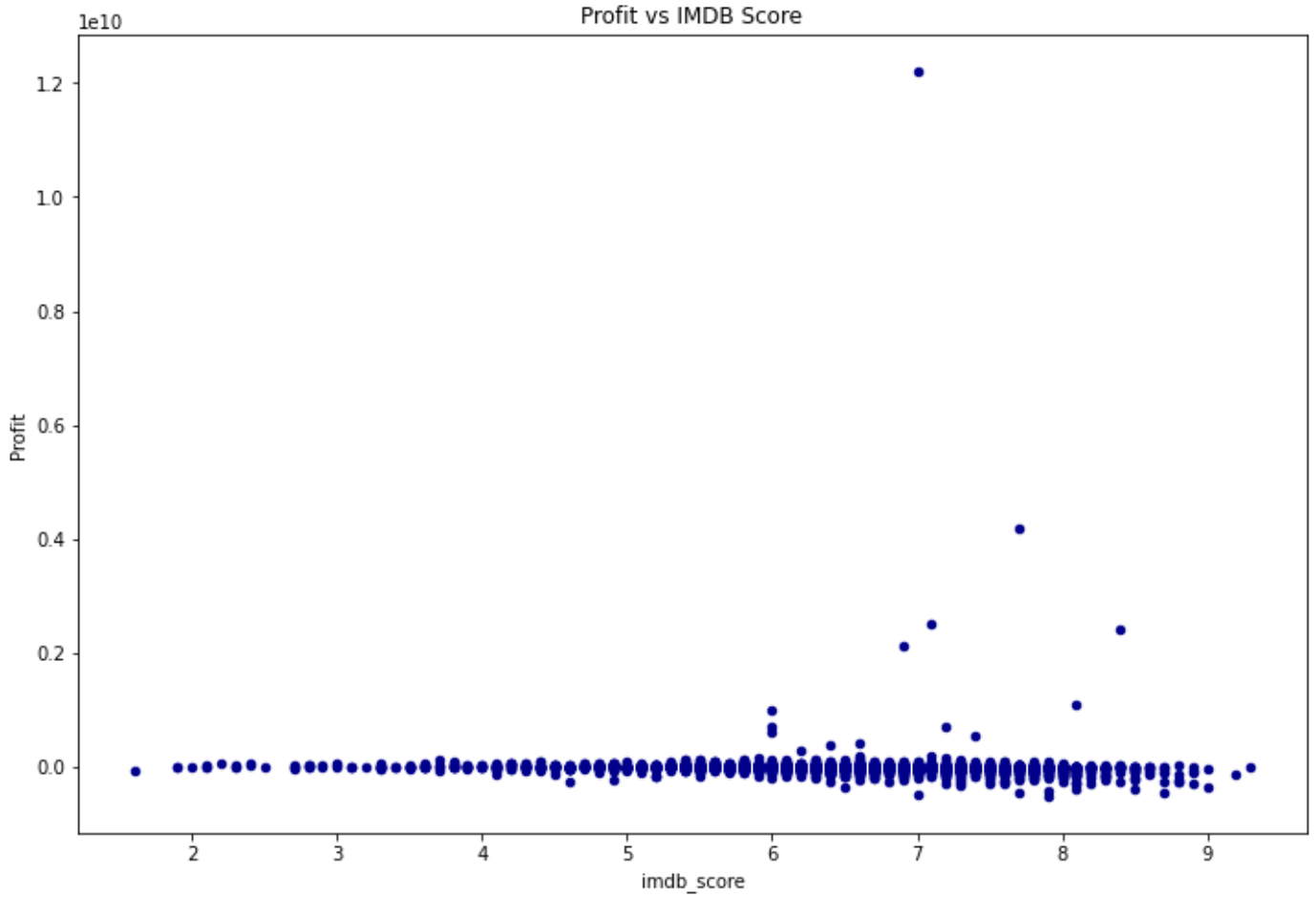
- Dealing with missing values
  - Print column names and number of Null values for each
  - Replace colour with most common colour (mode)
  - Remove null value for variables with names
  - Replace null value for numerical values with either Mean/Median
- Its important not to remove data when we can, as less data will affect the performance of our model.
- Data quality is essential.

Data Visualization

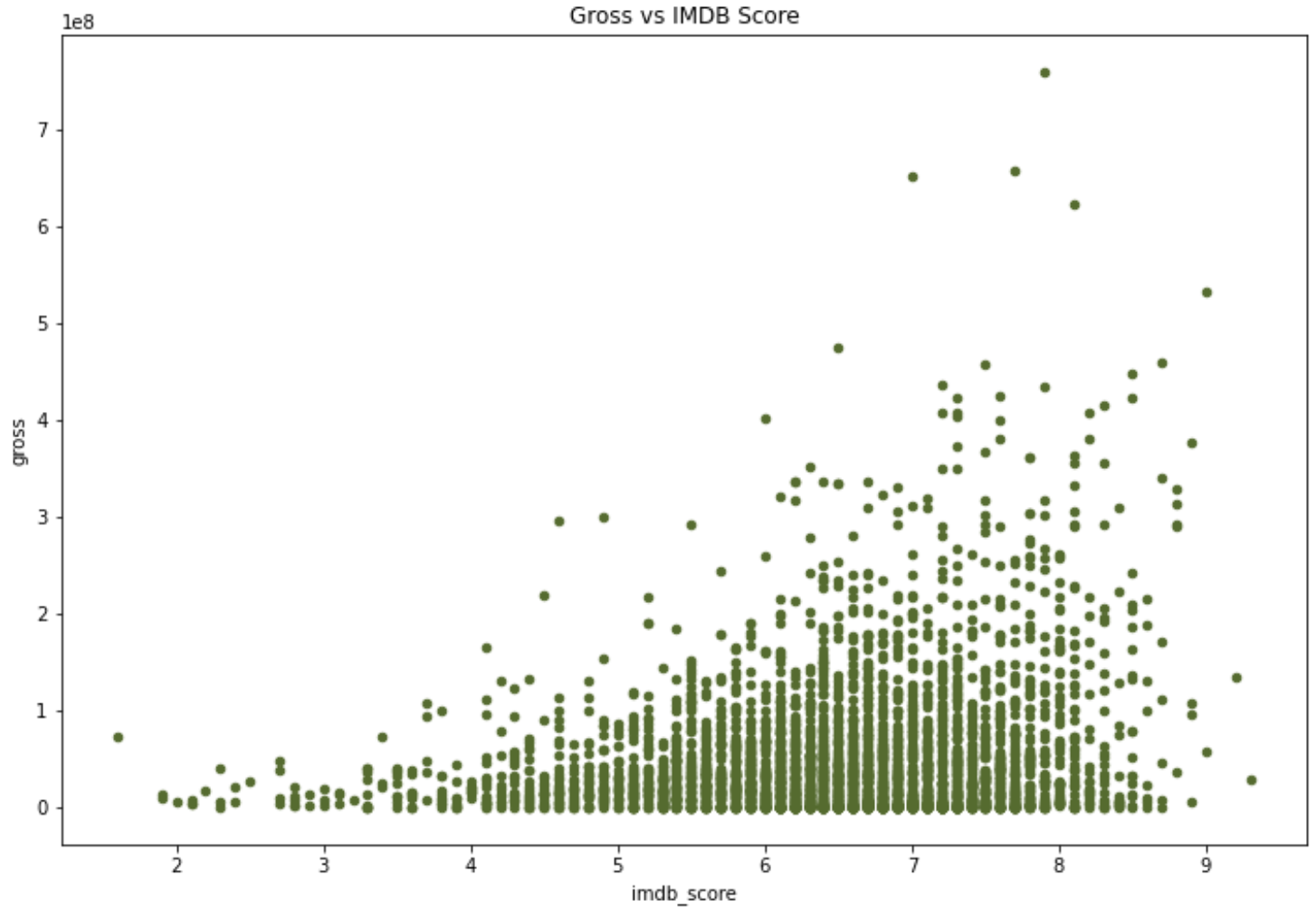
- Box Plot: Genres and their corresponding IMDB rating range



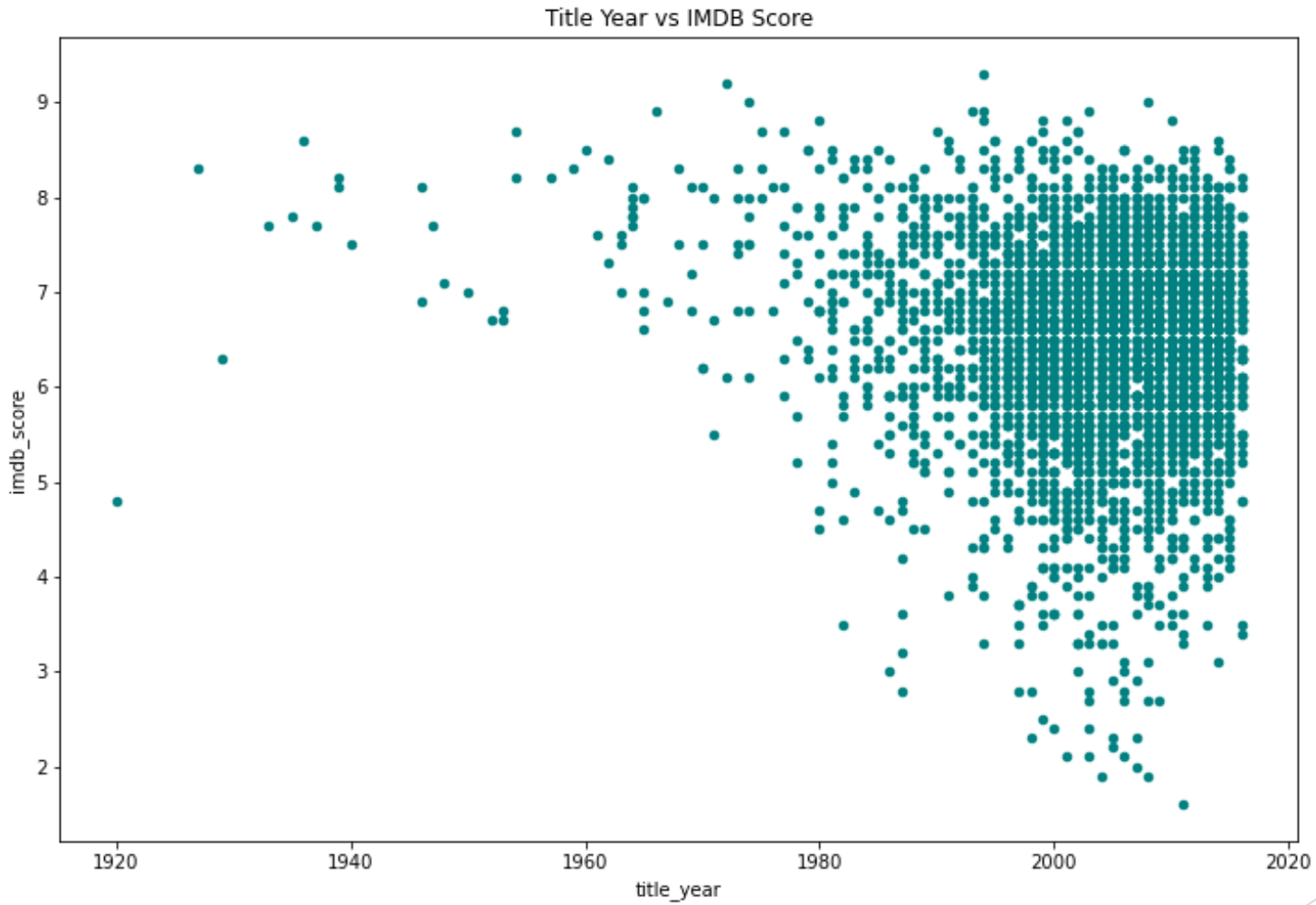
Profit vs IMDB Rating



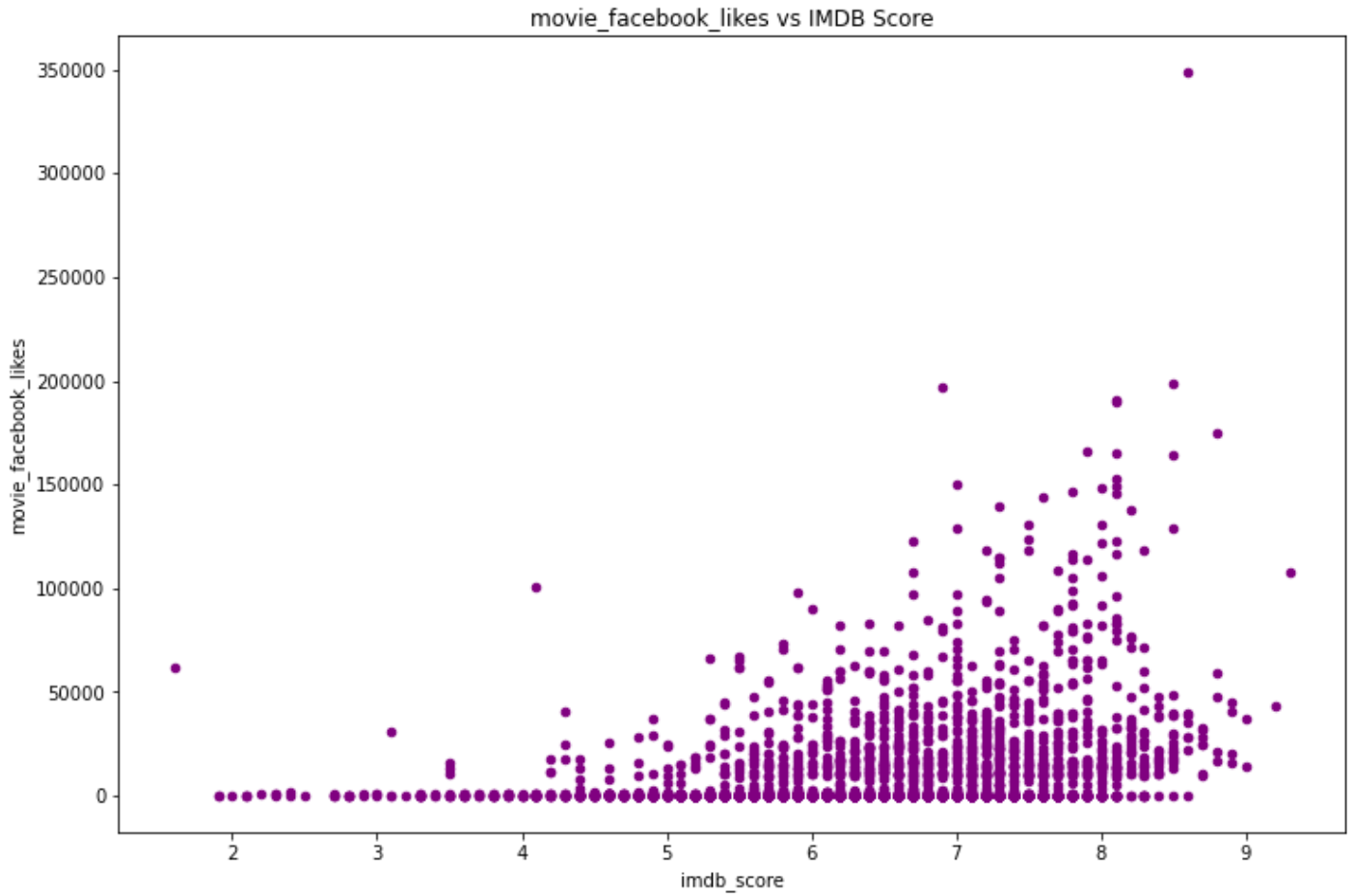
Gross vs IMDB Rating



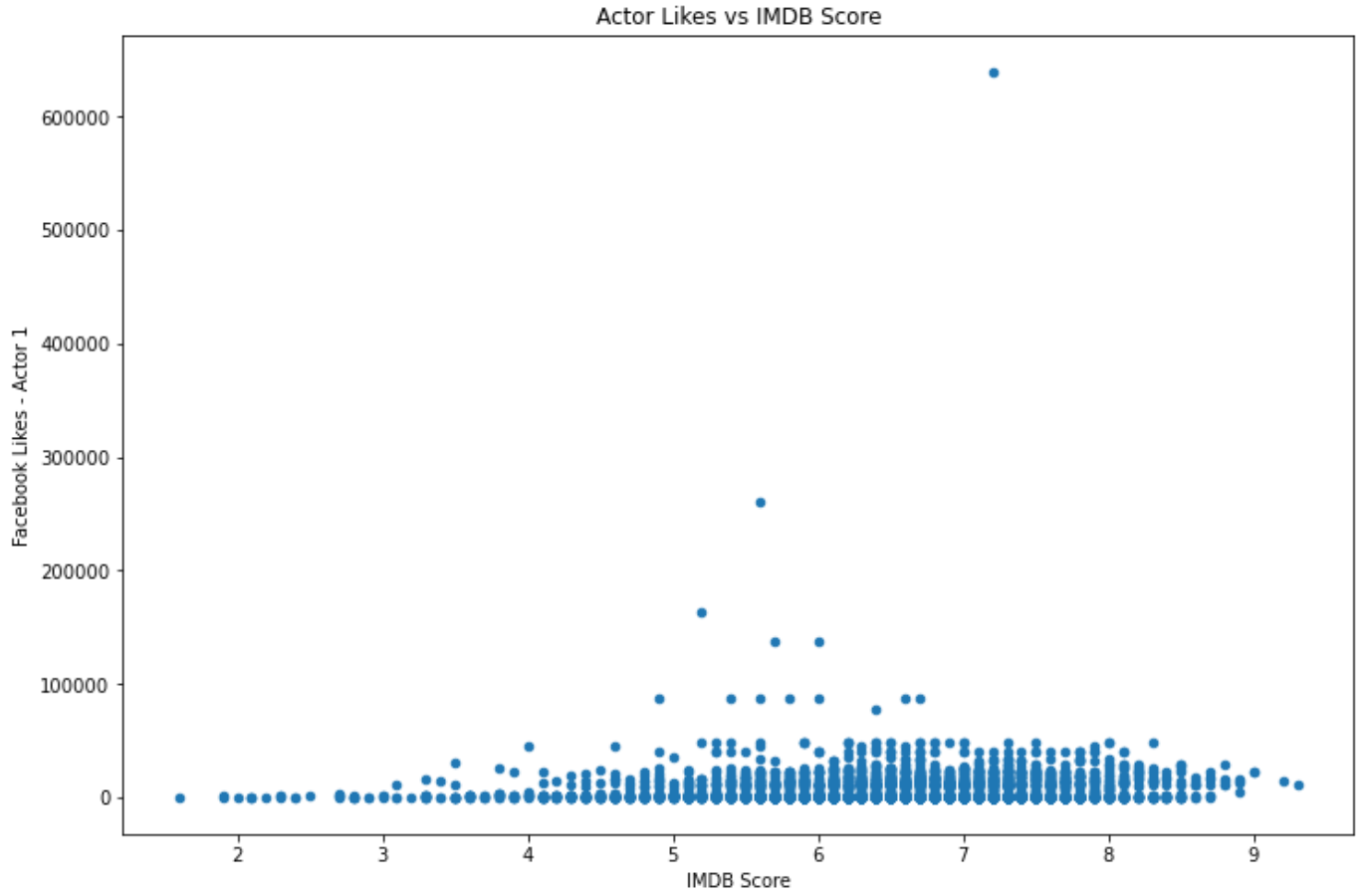
Title Year vs IMDB Rating



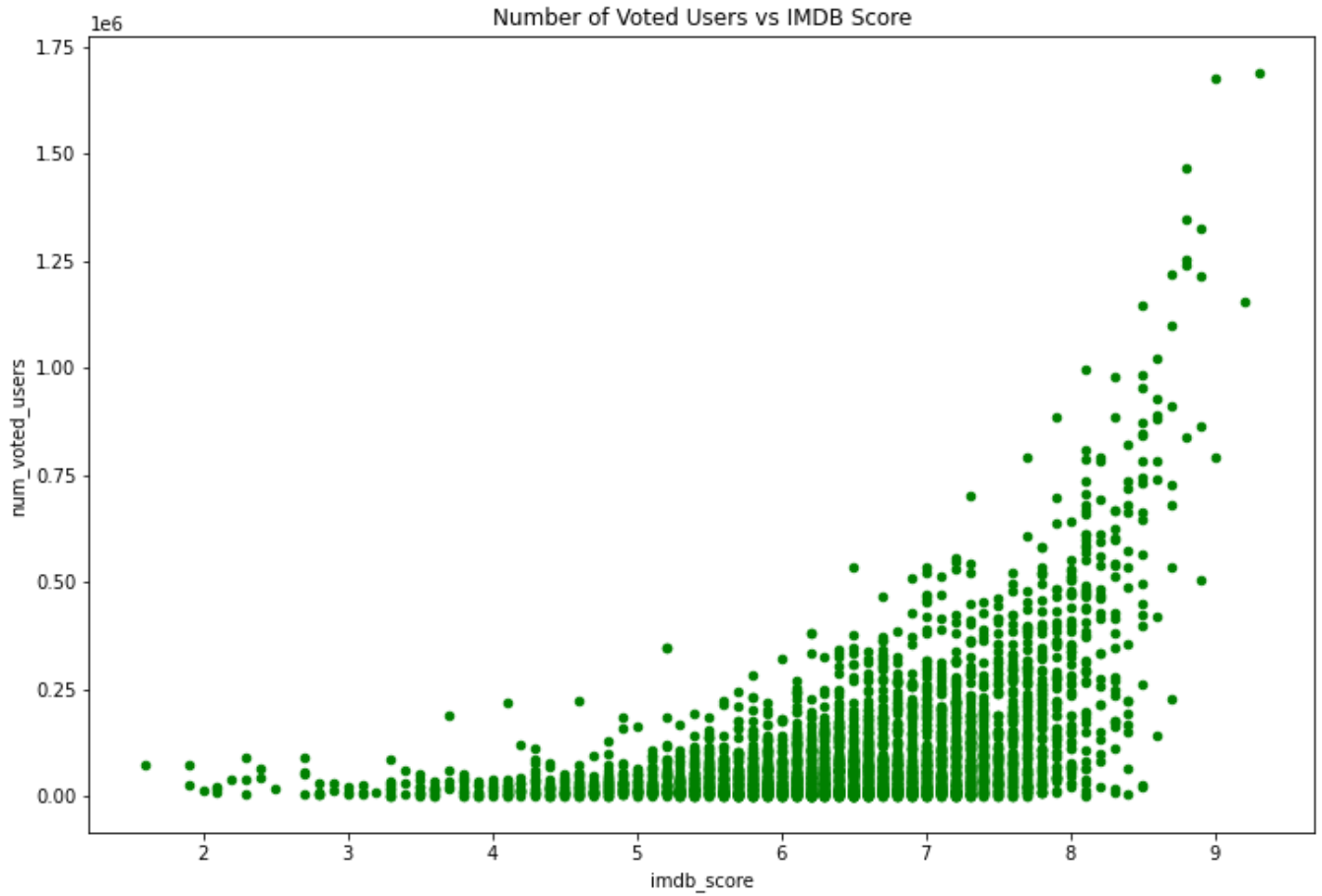
Movie Facebook Likes vs IMDB Rating



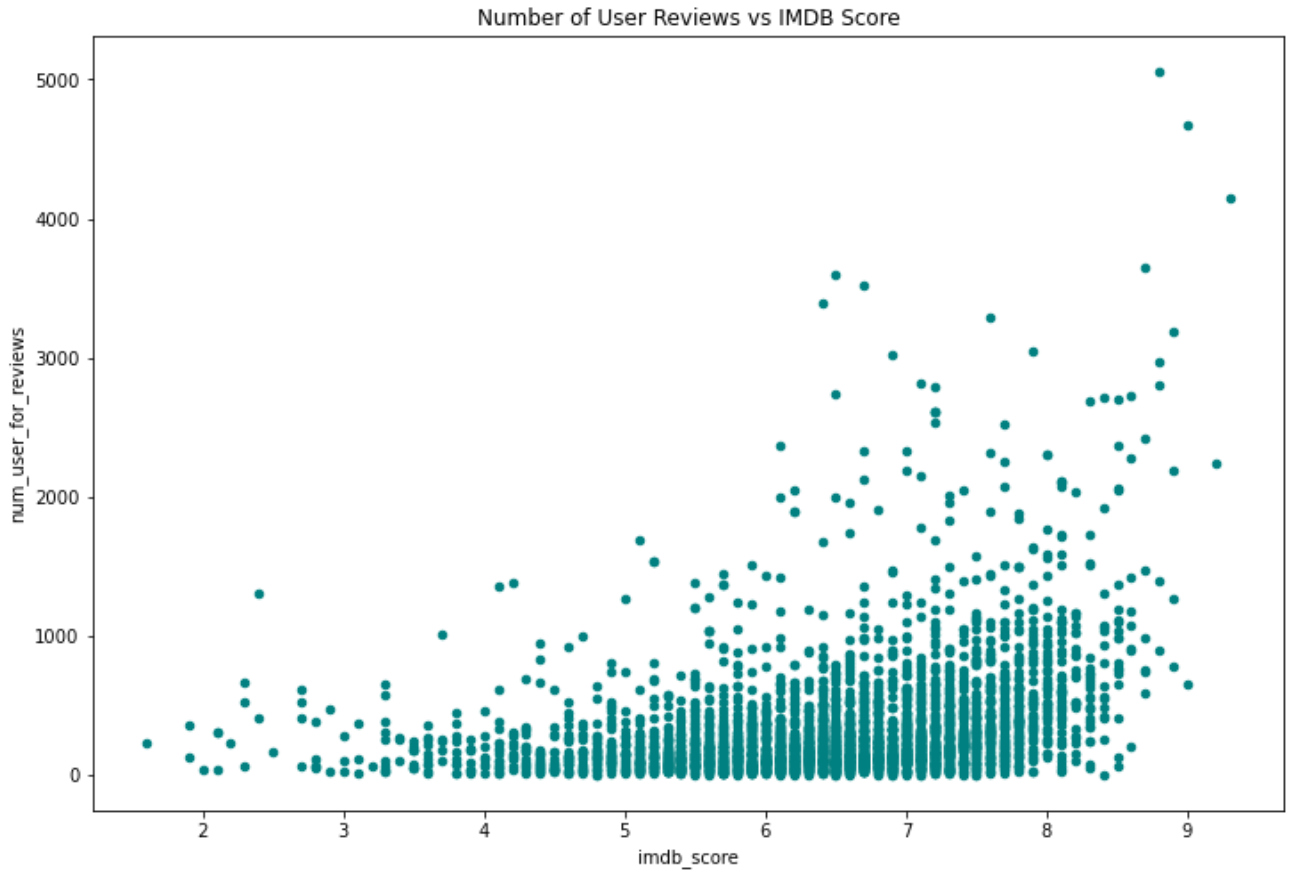
Actor 1 Likes vs IMDB Rating



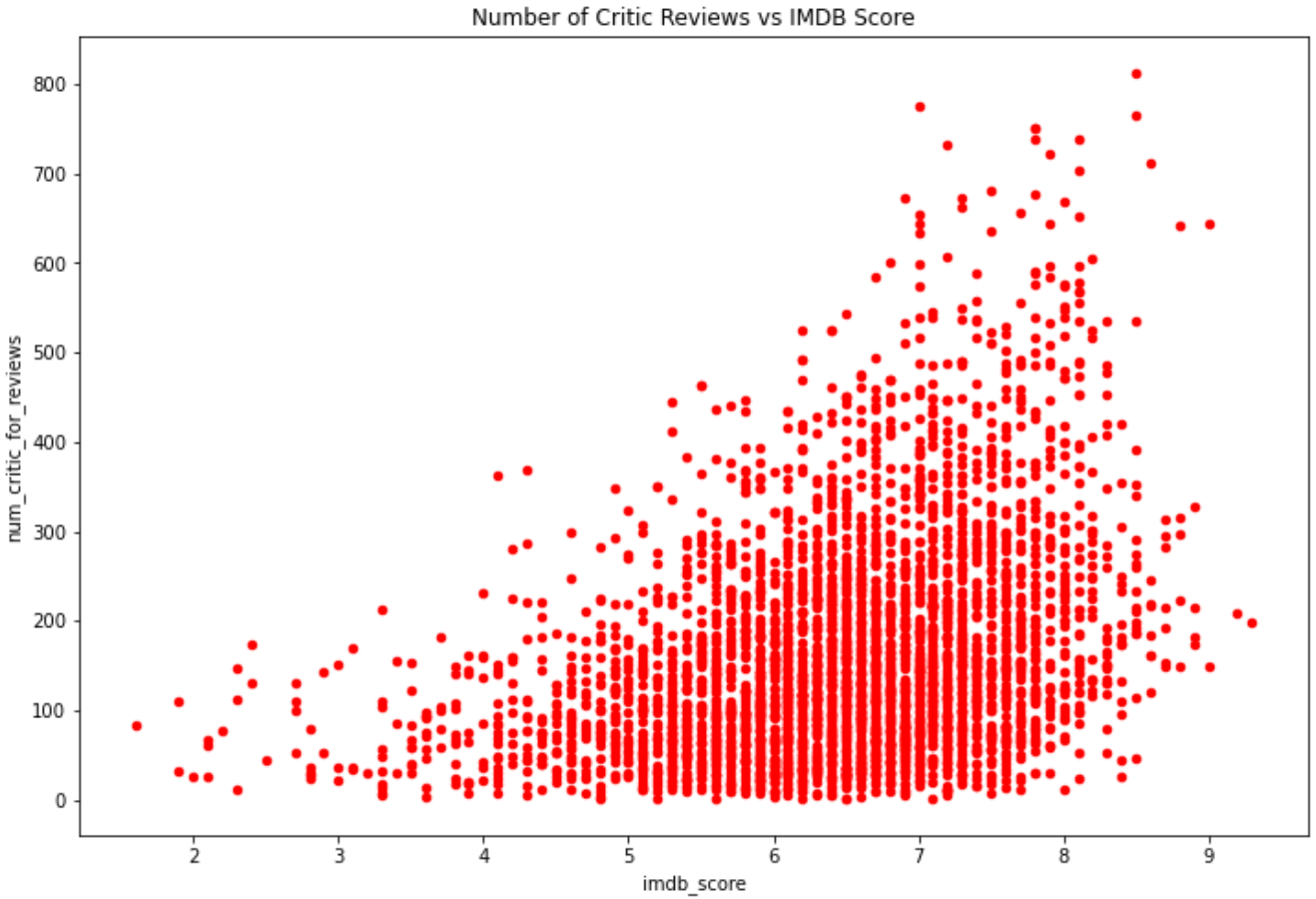
Number of Voted Users vs IMDB Rating



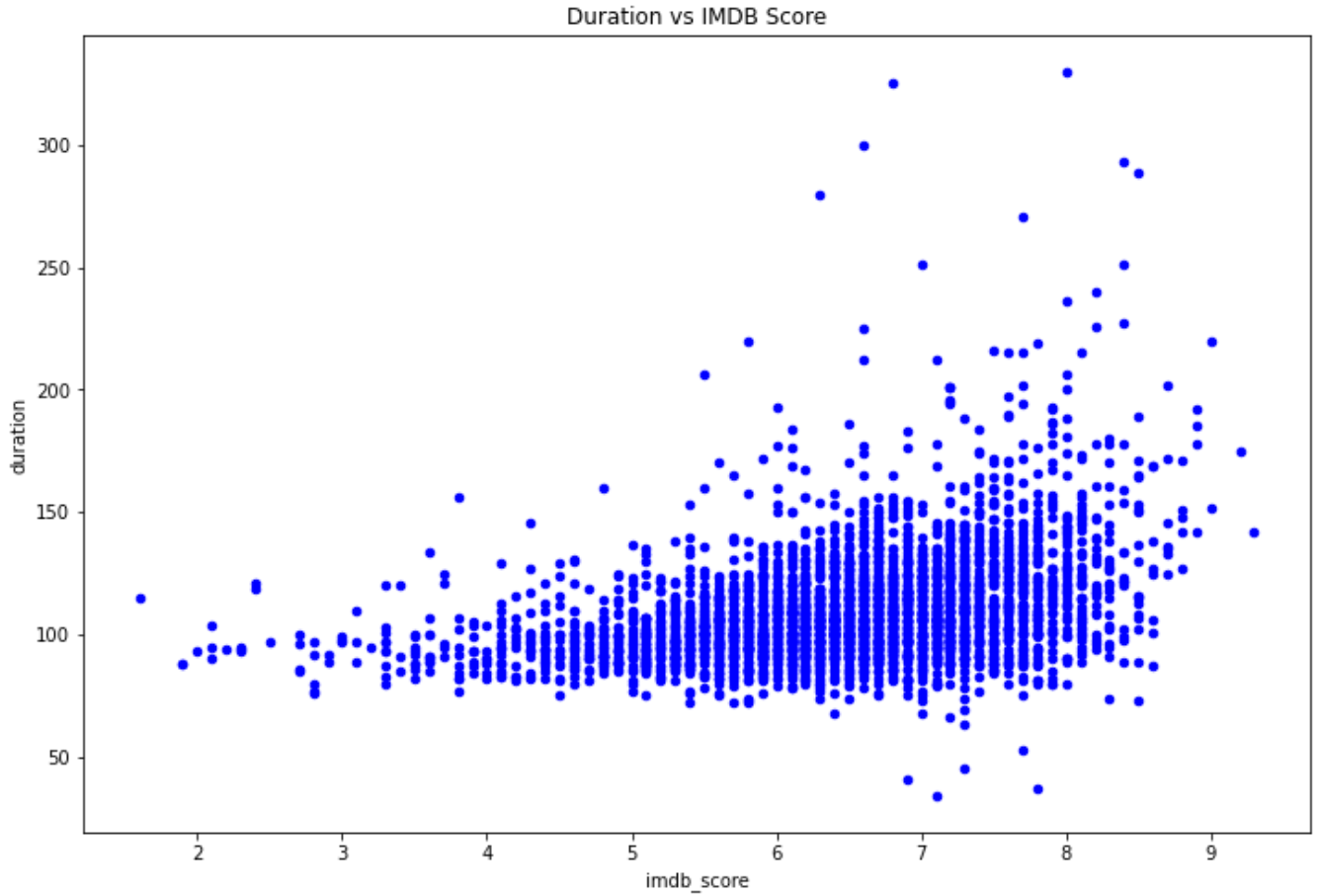
Number of User Reviews vs IMDB Rating



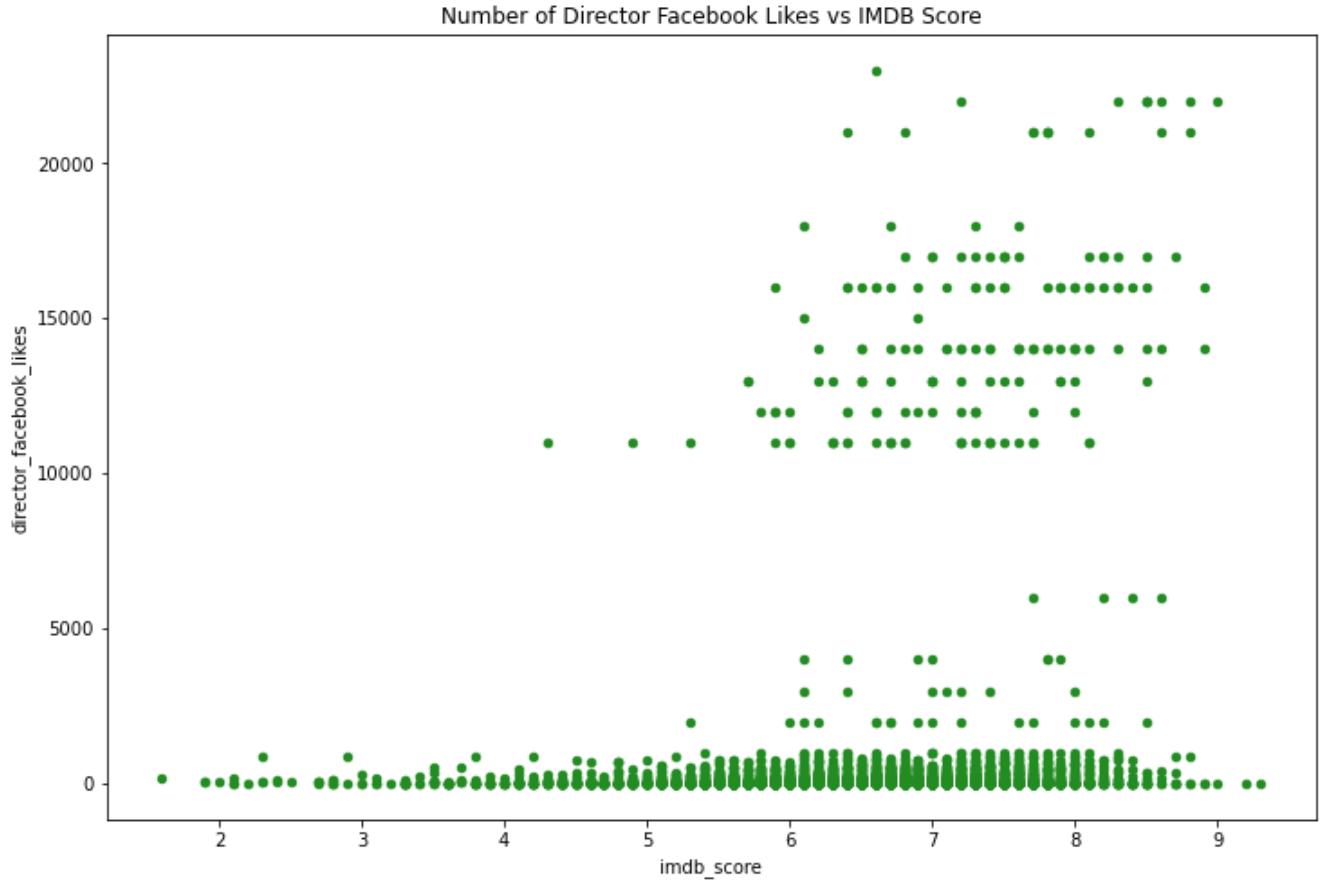
Number of Critic Reviews vs IMDB Rating



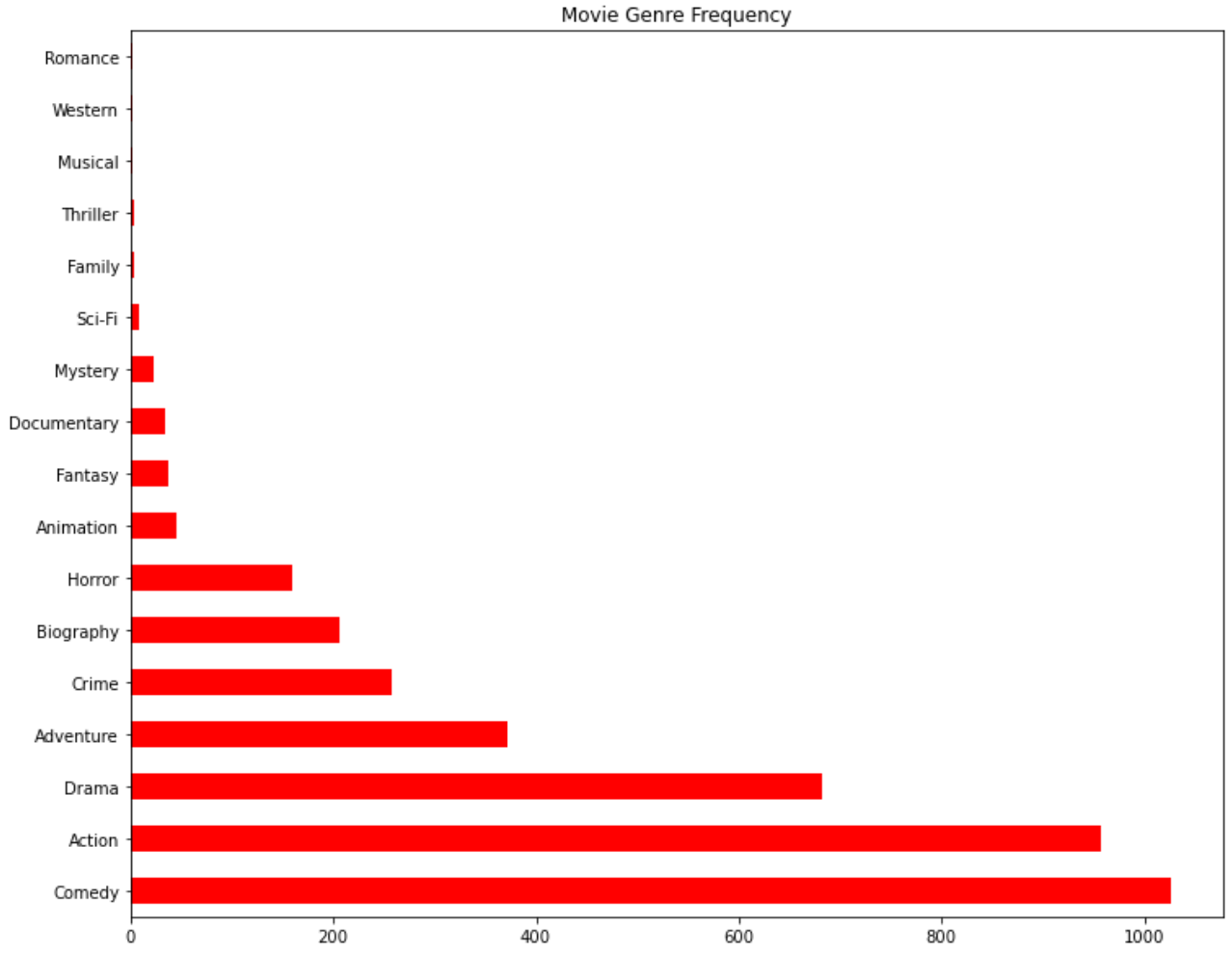
Duration vs IMDB Rating



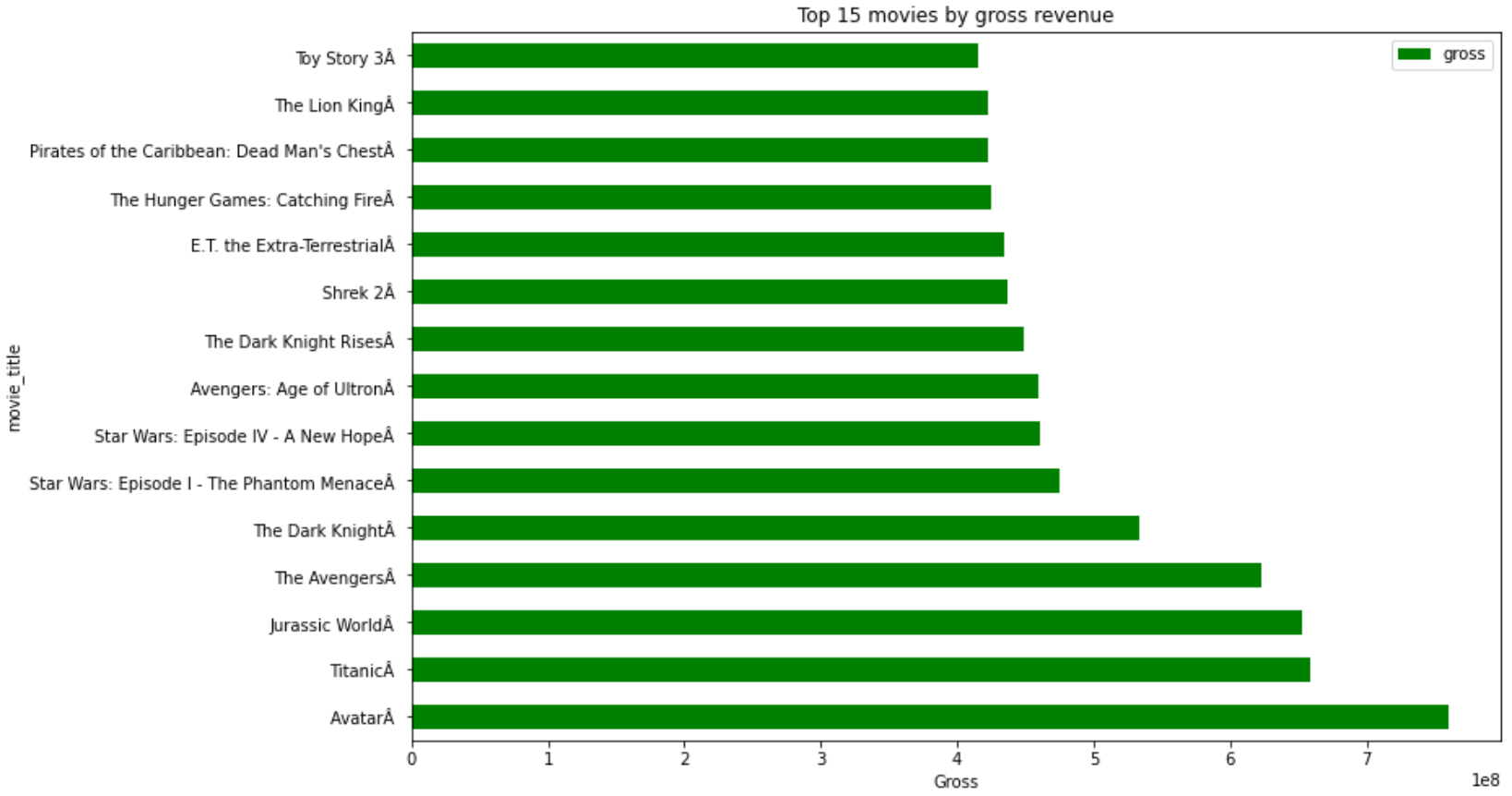
Director Likes vs IMDB Rating



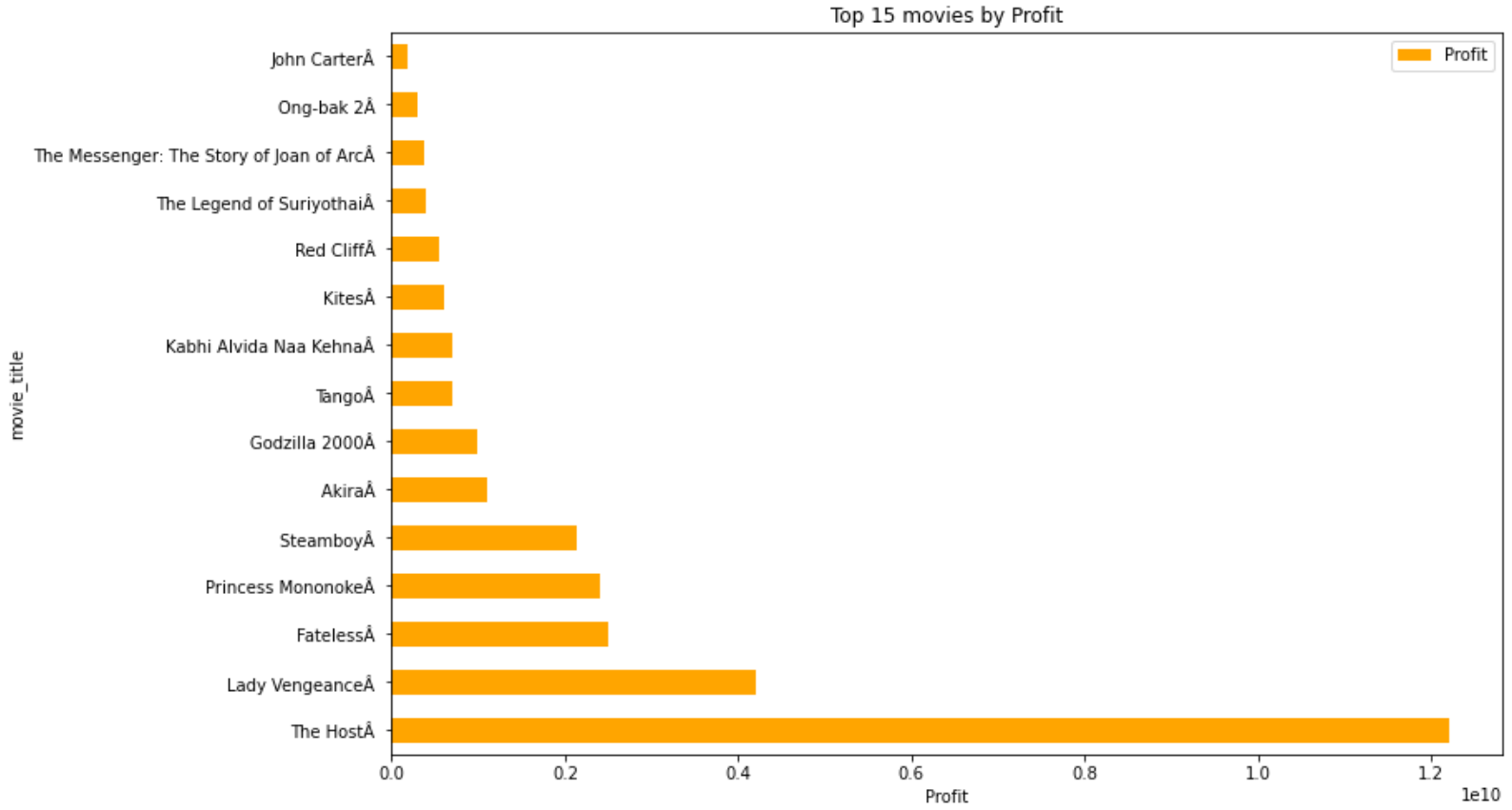
Frequency of Genres



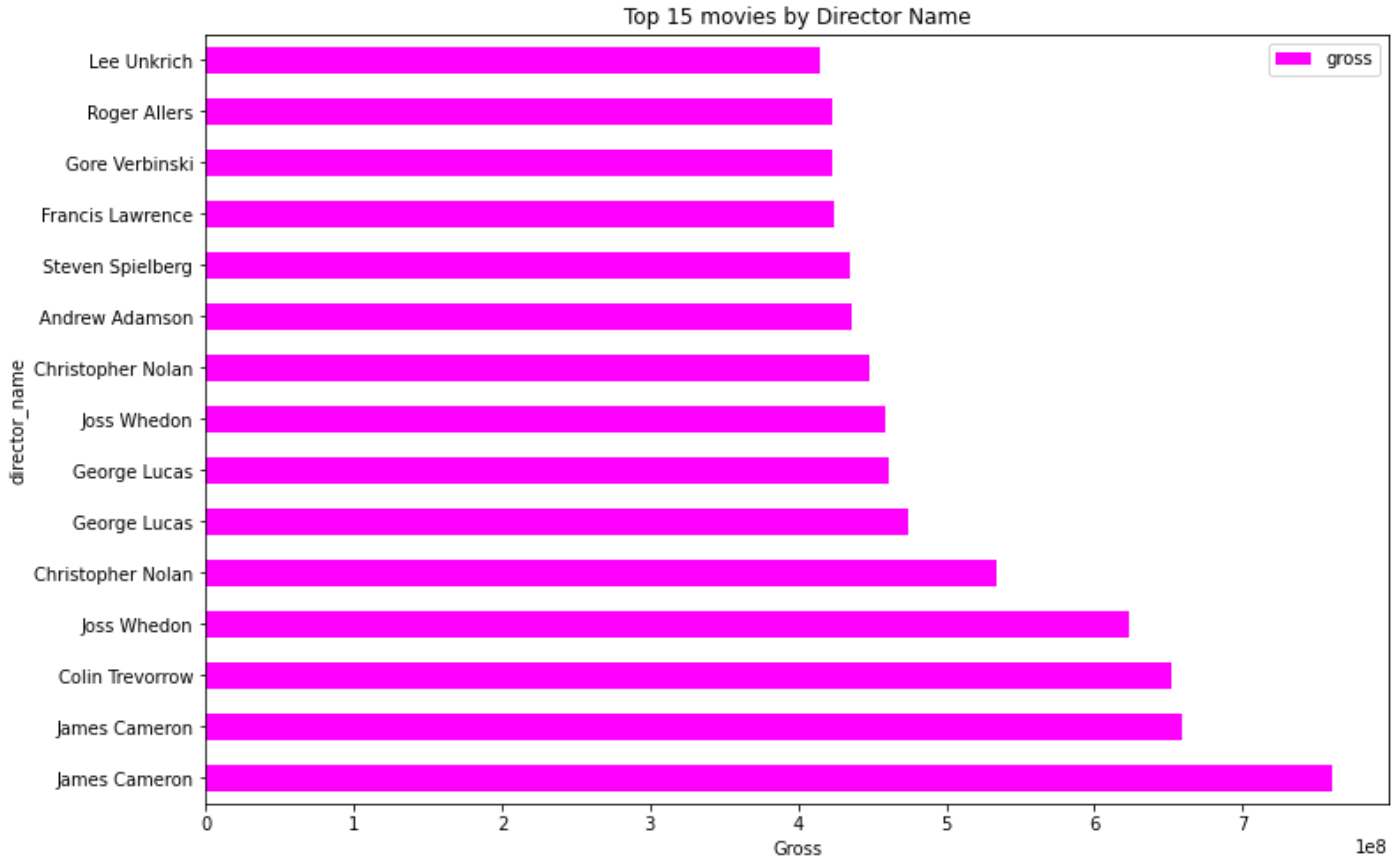
Top 15 movies by Gross



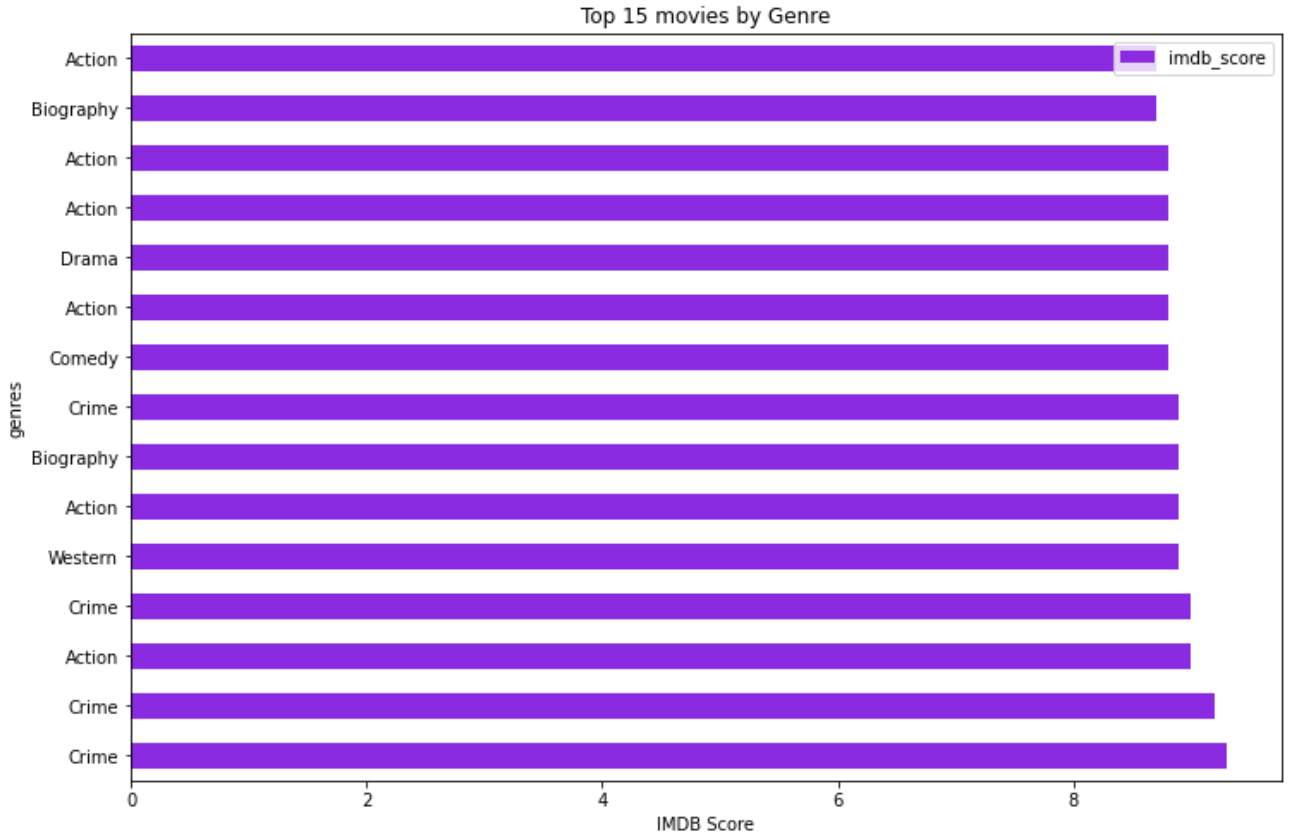
Top 15 movies by Profit



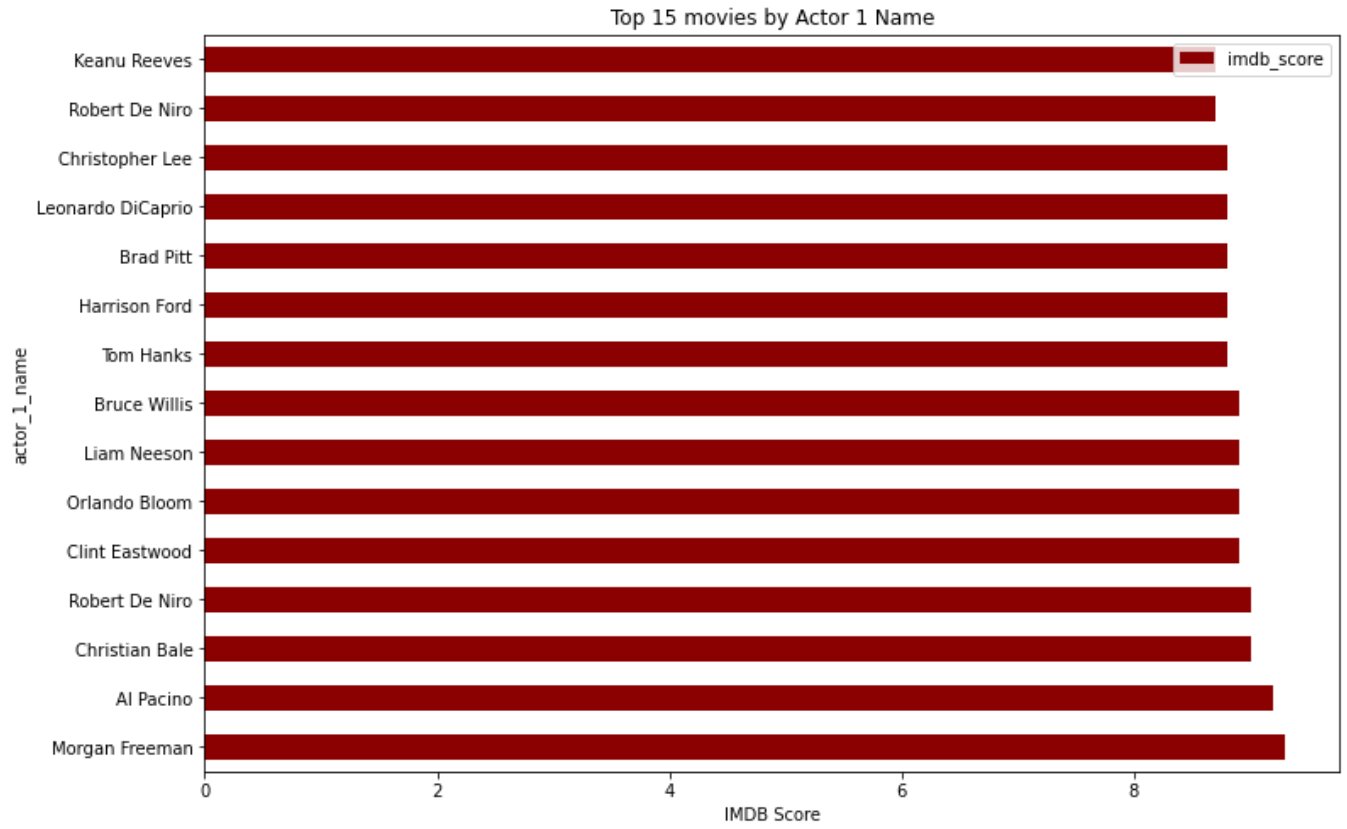
Top 15 Directors based on IMDB Rating



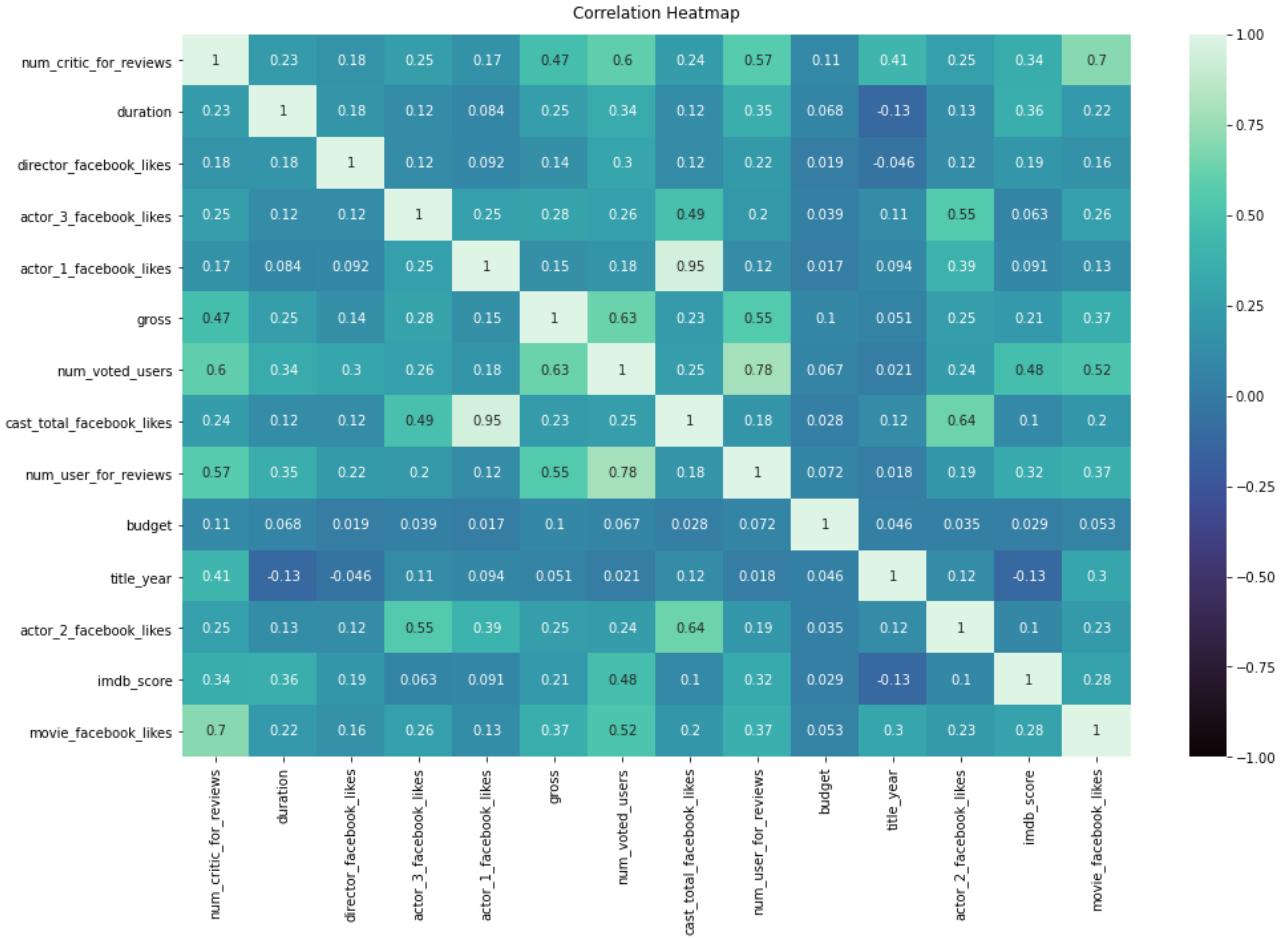
Top 15 Genres based on IMDB Rating



Top 15 Actors based on IMDB Rating



Final Data Preparation

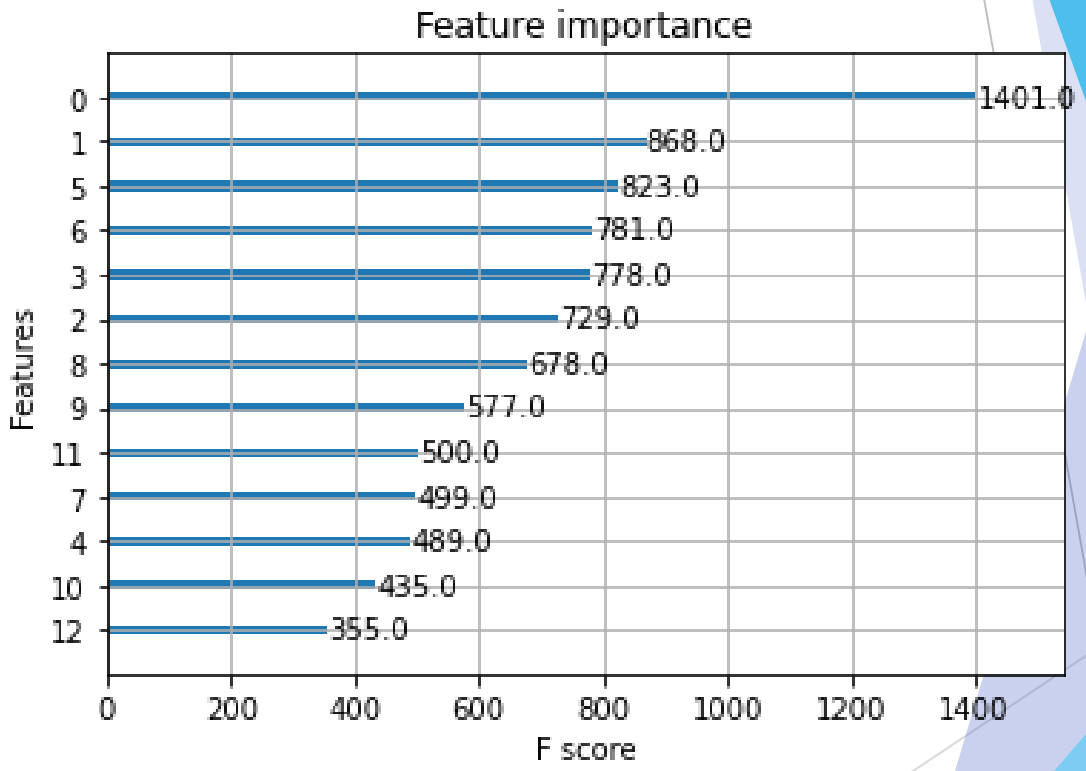


Modelling

- We must split data into training and test data so we can evaluate the performance of our algorithm on unseen data.
- We will use a 70/30 train test split.
- Perform feature scaling using MinMaxScaler to standardize the features, this improves Machine learning model performance.
- We do not use Standard Scaler because the data is not normally distributed
- **Linear Regression** Performance:
  - Mean Squared Error = 0.71, Mean Absolute Error = 0.65
- **XGBoost** Performance:
  - Mean Squared Error = 0.50, Mean Absolute Error = 0.51

Findings

- Most important features (in order):
  - Number of Critic Reviews
  - Duration
  - Gross
  - Number of Voted Users
  - Director Facebook Likes
  - Number of User for reviews
  - Budget
  - Actor Facebook Likes
  - Cast Facebook Likes
  - Title Year
  - Movie Facebook Likes



Conclusion

- **Goal:** Find the biggest factors that determine whether a movie is a success
- We defined objectives and broke the problem down to achieve the goal using the Data Science Methodology.
- We used Data Visualization techniques to gain insight into the correlations between different variables
- We also looked at the Top 15 instances for several features such as Genre, Gross, director Name
- Stakeholders now have more knowledge about which factors effect movies success and can take actionable steps to make sure they produce/show the right movies.

Improvements/Future Work

- Incorporate more modelling techniques such as Support Vector Machines, Random Forest
- Classification techniques – Logistic Regression
- Implement Neural Network to gain further hidden insights
  - Use techniques such as Optimization
  - Transfer Learning – using an already trained Neural Network on a new dataset
- More importantly:
  - Incorporate more datasets, different types of data
  - Incorporate big data tools/gather data from Data Lakes/Warehouses

References

- Notebook
  - https://github.com/sakibch/IMDB_movie_analysis/blob/main/movie-analysis.ipynb
- https://pandas.pydata.org/docs/
- https://matplotlib.org/stable/index.html
- https://scikit-learn.org/stable/user_guide.html
- https://seaborn.pydata.org/tutorial.html
- https://xgboost.readthedocs.io/en/stable/parameter.html

Thank You For Listening!