TASK

Exploratory Data Analysis on the Movies Data Set

Visit our website

Introduction

The movie data set can be used to study the evolution of movies over the past 100 years. We can gain insight into the most profitable and make predictions as to what could be the next big hit based on the Genre, production country, and the spoken language.

The dataset comprises statistical data collected about movies over the last 100 years from 1916 to 2017.
Key attributes include:
Budget, genres, id,popularity, production_countries, release_date, revenue, runtime, spoken_languages, title, vote_average, vote_count, release_year, profit

To conduct this analysis, we will utilize Python's data analysis and visualization libraries, including pandas, numpy, matplotlib, and seaborn. Techniques such as summary statistics, correlation analysis, and data visualization will be employed to uncover insights and highlight significant trends.

DATA CLEANING

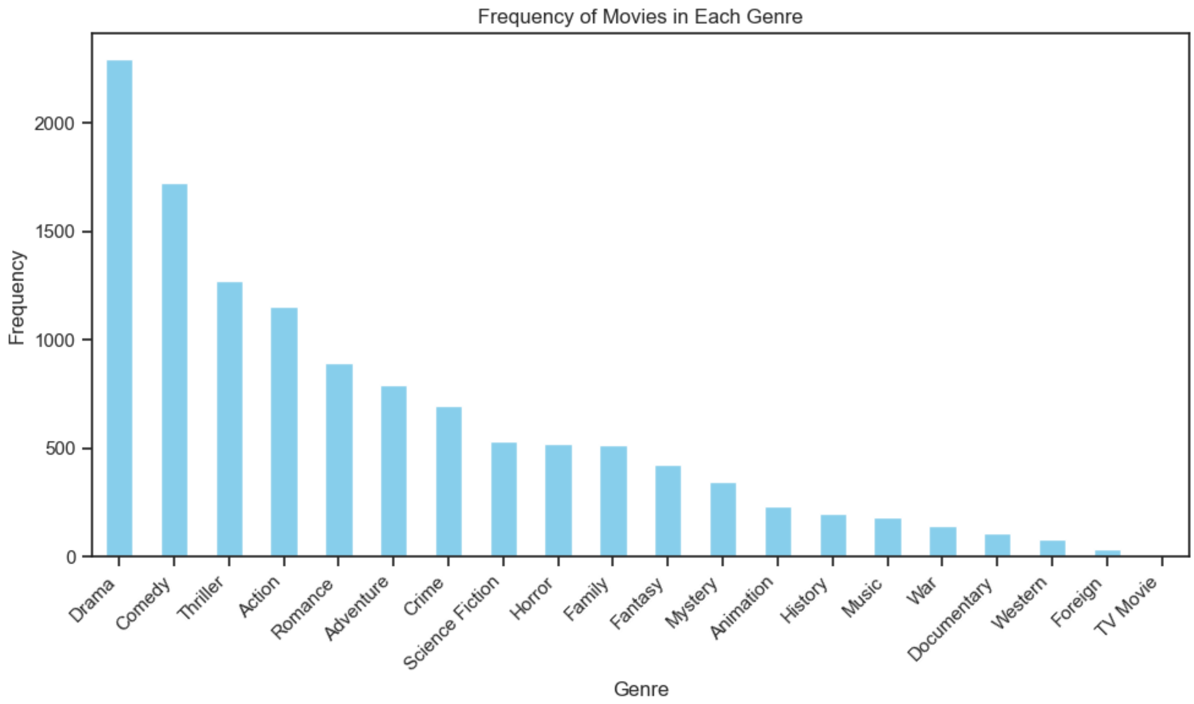
The following steps were taken to clean the data:

1. Redundant and unnecessary data was removed from the dataset. The columns removed were identified as not being relevant for the decision-making and evaluation of the data. (homepage, keywords, original_language, original_title, overview, production_companies, 'status', tagline)
2. The removal of duplicate values from the dataset.
3. The removal of rows of missing data.
4. The changing of columns datatype to make it easier to manipulate in the data exploration process. The release_date column has been changed to datetime data type and the budget and revenue columns have been changed to int64.
5. The genres, production_countries, and spoken_languages columns data was in JSON format. These were converted into a list of strings to make it easier to work with.

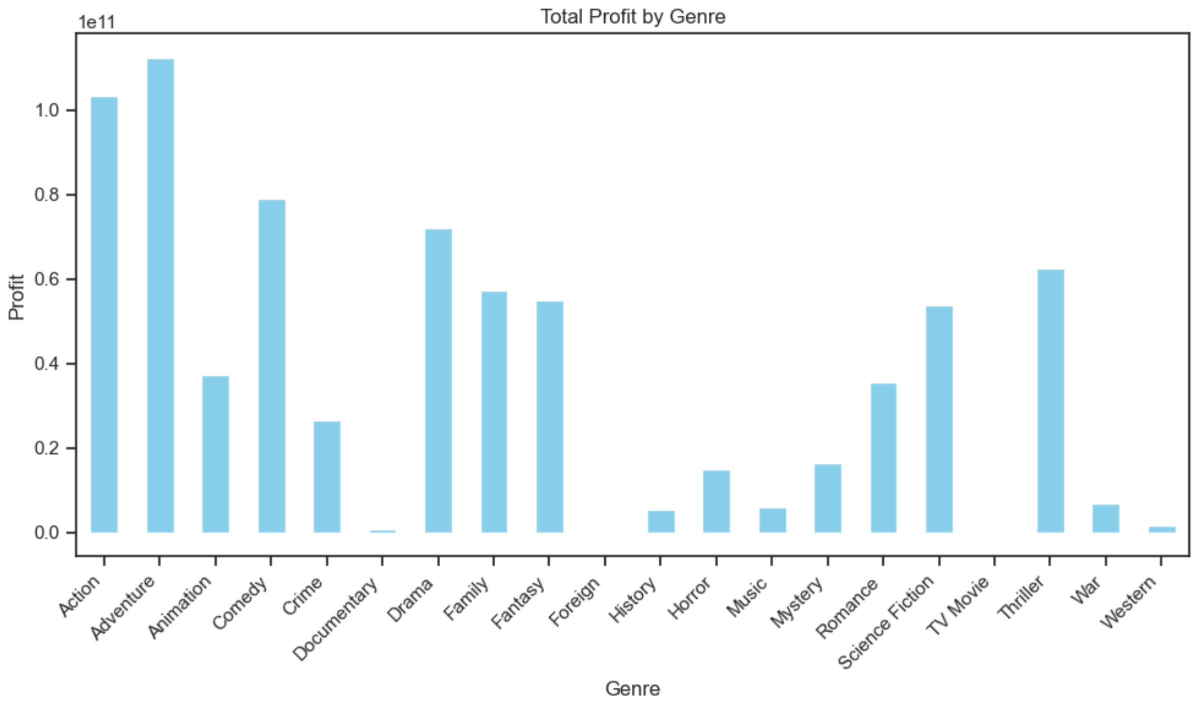
MISSING DATA

A check for missing data on the dataset was carried out. Upon inspection of the null check results the number of missing values was deemed small enough that the complete removal of the rows with the missing data would not affect the overall evaluation of the data.

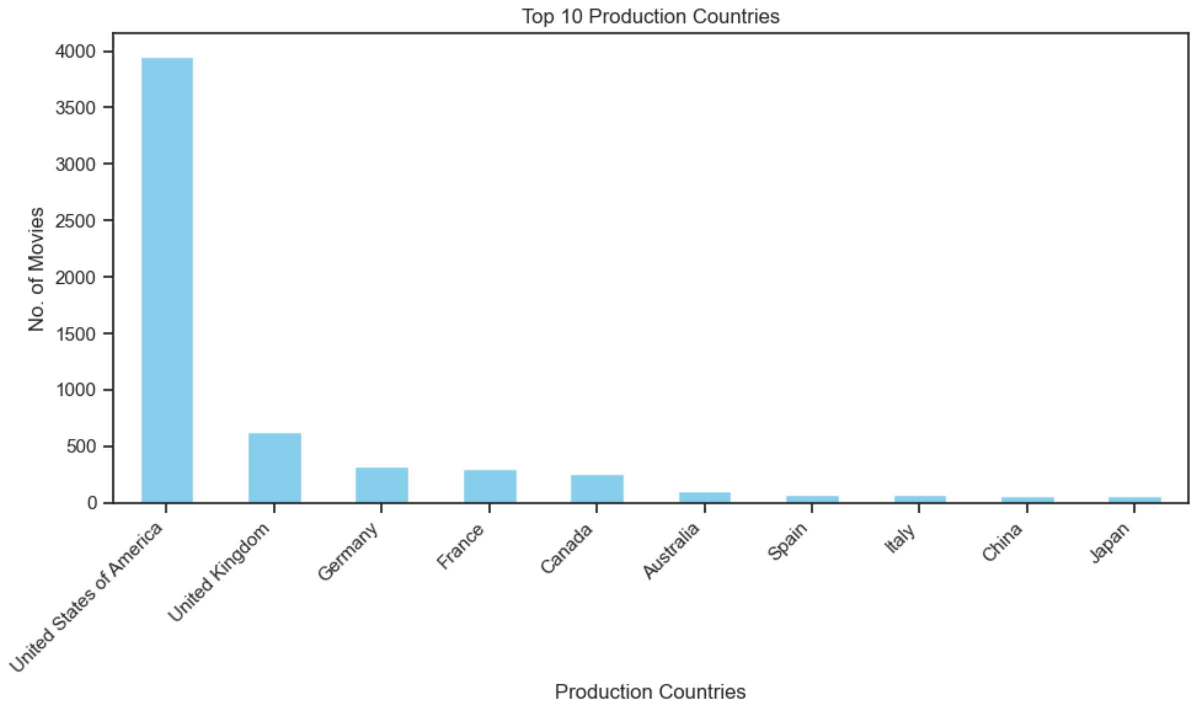
DATA STORIES AND VISUALISATIONS



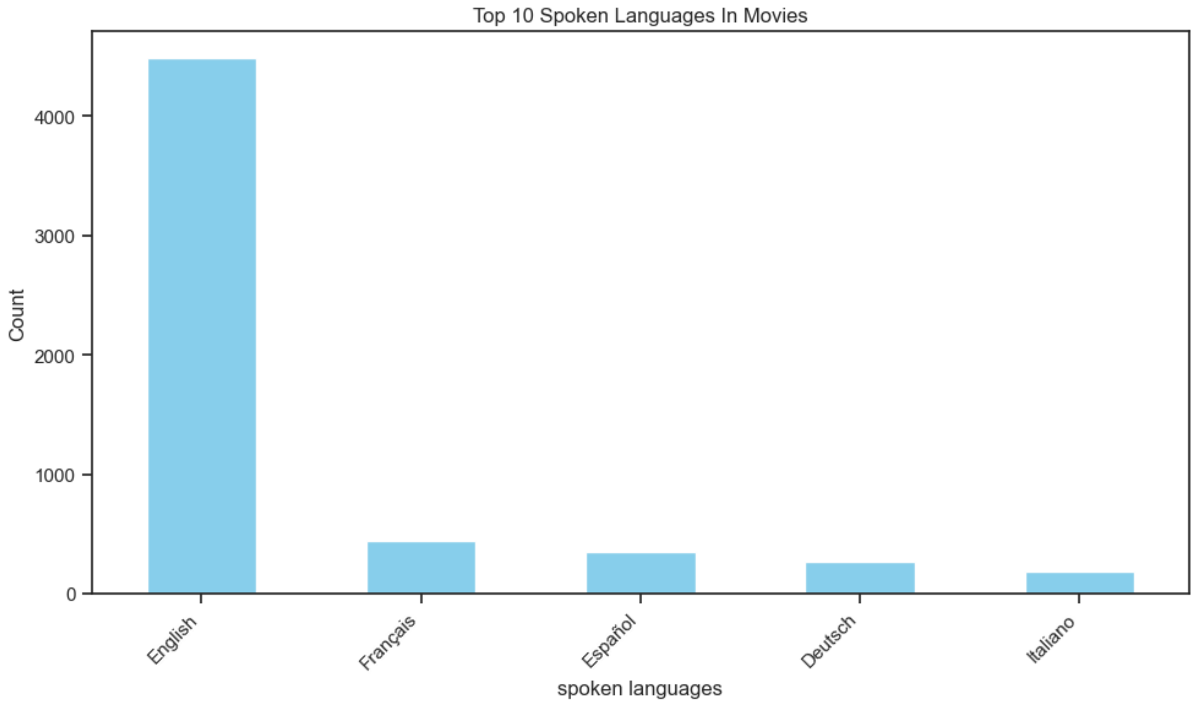
This graph displays the total number of movies made per genre for the past 100 years. Based on the graph you would assume that this would translate directly into the most profitable genres, but we will see from the next graph.



This graph shows us that the adventure genre followed by the action genre is the most profitable if we look at the ***frequency of movies in each genre*** we can see that while there might not be the most made movie genre when it applies to profitability it is about quality over quantity. Adventure movies such as the Star Wars series for example have a cult following and are deemed a timeless classic and can be enjoyed by every age group.



Now that we know what genre of movie is the most popular we can take a closer look into the most popular place to make movies. Based on the graph the United States of America far exceeds any other country for production. The assumption can be made from this is that they are the leaders in filmmaking and would have the largest base of expertise needed to make a great adventure movie. With that assumption, you could determine whether they would have all the necessary equipment (sound stages for example) to make said movie.



This graph shows us that English is the most popular language spoken in movies and inclusivity is important if we look purely from a profitability standpoint English far exceeds any other language to make a movie in.

In conclusion, based on the data presented these are the assumptions that can be made about making a profitable movie:

1. **The genre of the movie should be adventure.**
2. **The spoken language needs to be English**
3. **The production country should be the United States of America**

THIS REPORT WAS WRITTEN BY: SHEENA WEBER