TASK

Exploratory Data Analysis on the Movies Data Set

Visit our website

Introduction

This is a Data Analysis Capstone Project which will be dealing with analysing the Movies Data Set. In this data set are movies about 4800 from all genres including Animation, Action, Adventure and Science Fiction just to name a few. The data set has 20 columns with headings used to analyse each movie from, the budget, original language of the movie, its popularity, production companies, its title all the way to the vote count of how the people liked the movie.

In this document I will be sharing how I have analysed this data set using the knowledge I have gained in data cleaning, handling missing data and data visualisations.

DATA CLEANING

I started by dropping all the columns with redundant and unnecessary information, the kind of data that will not help in the analysis. I then followed with checking if there are any duplicate rows and dropped them if available.

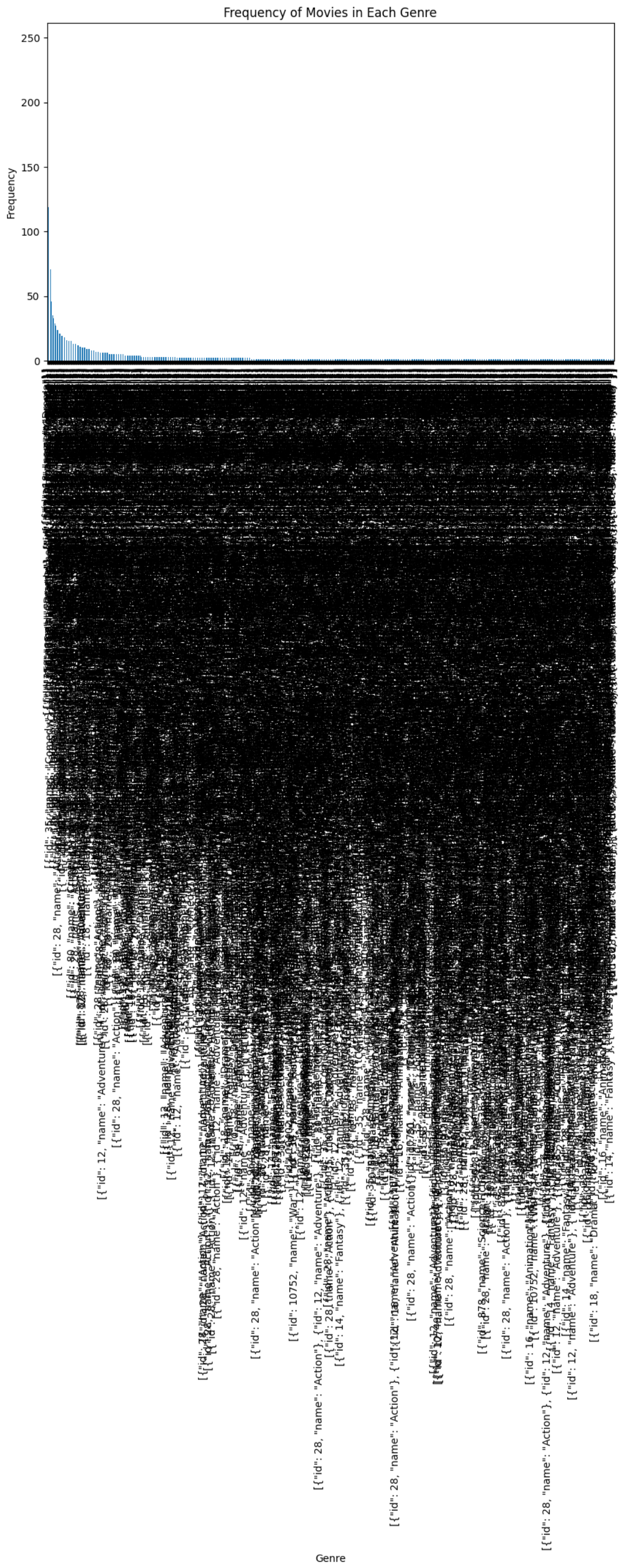
MISSING DATA

All the movies with zero budget were removed as that means that some values have not been recorded or there is some missing information. The rows with missing data were removed using ‘dropna’ function. I dropped these value so that I can get more accurate results.

DATA STORIES AND VISUALISATIONS

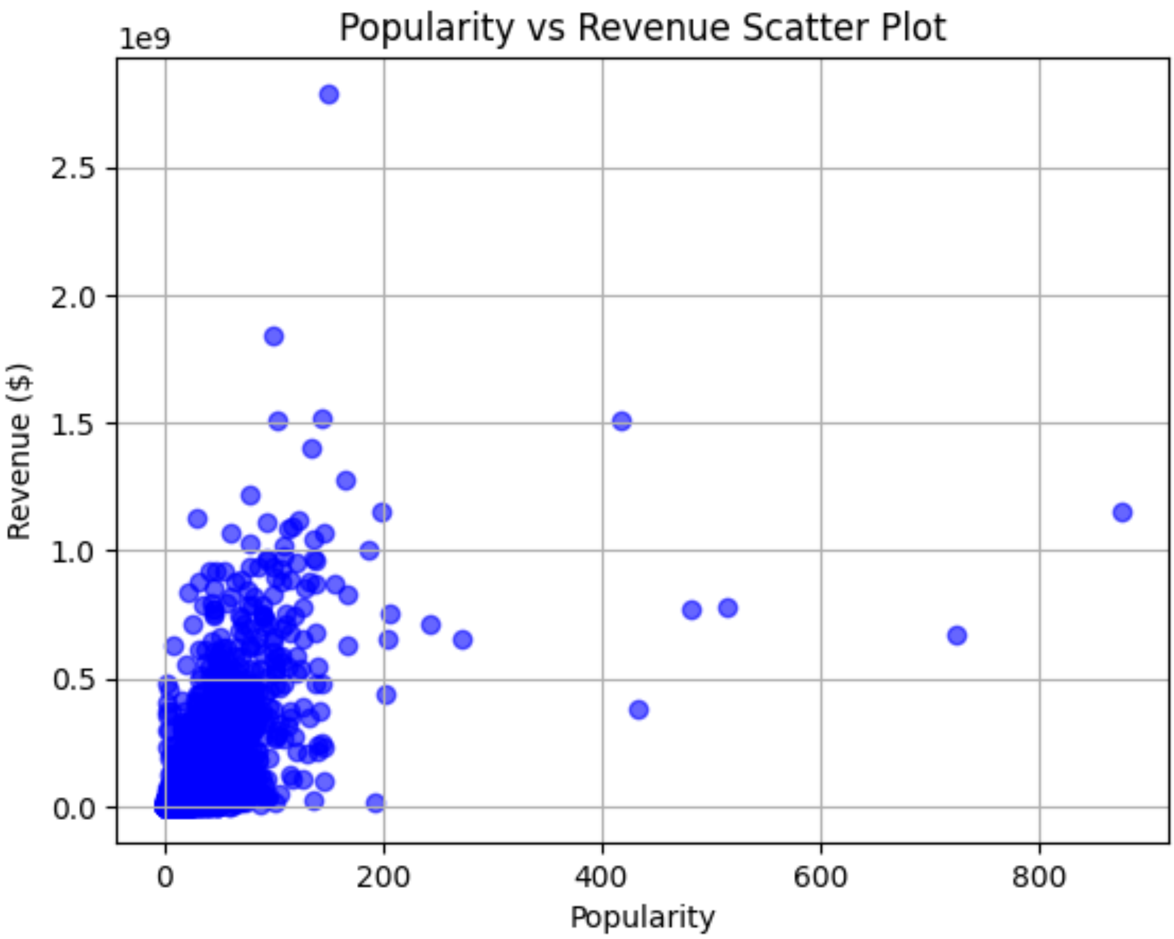
Frequency of movies in each genre

The bar plot explaining the frequency of movies in each genre The x-axis represents different genres. The y-axis represents the frequency of movies in each genre. Each bar's height indicates the number of movies in that genre. This bar plot helps identify the most and least common genres within the dataset.



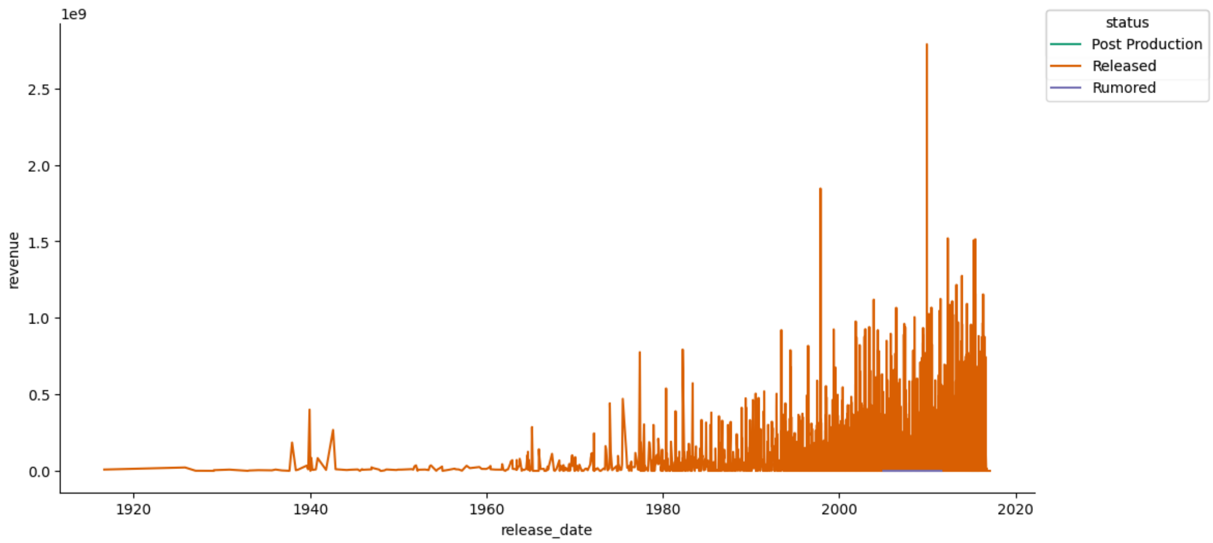
Popularity vs revenue

In the ‘popularity vs revenue’ scatter plot The x-axis represents the popularity of movies. The y-axis represents the revenue generated by movies. Each data point on the plot corresponds to a movie, showing its popularity and revenue relationship. Analysing this scatter plot can provide insights into whether highly popular movies tend to generate higher revenue, or if there are any outliers that defy this trend.



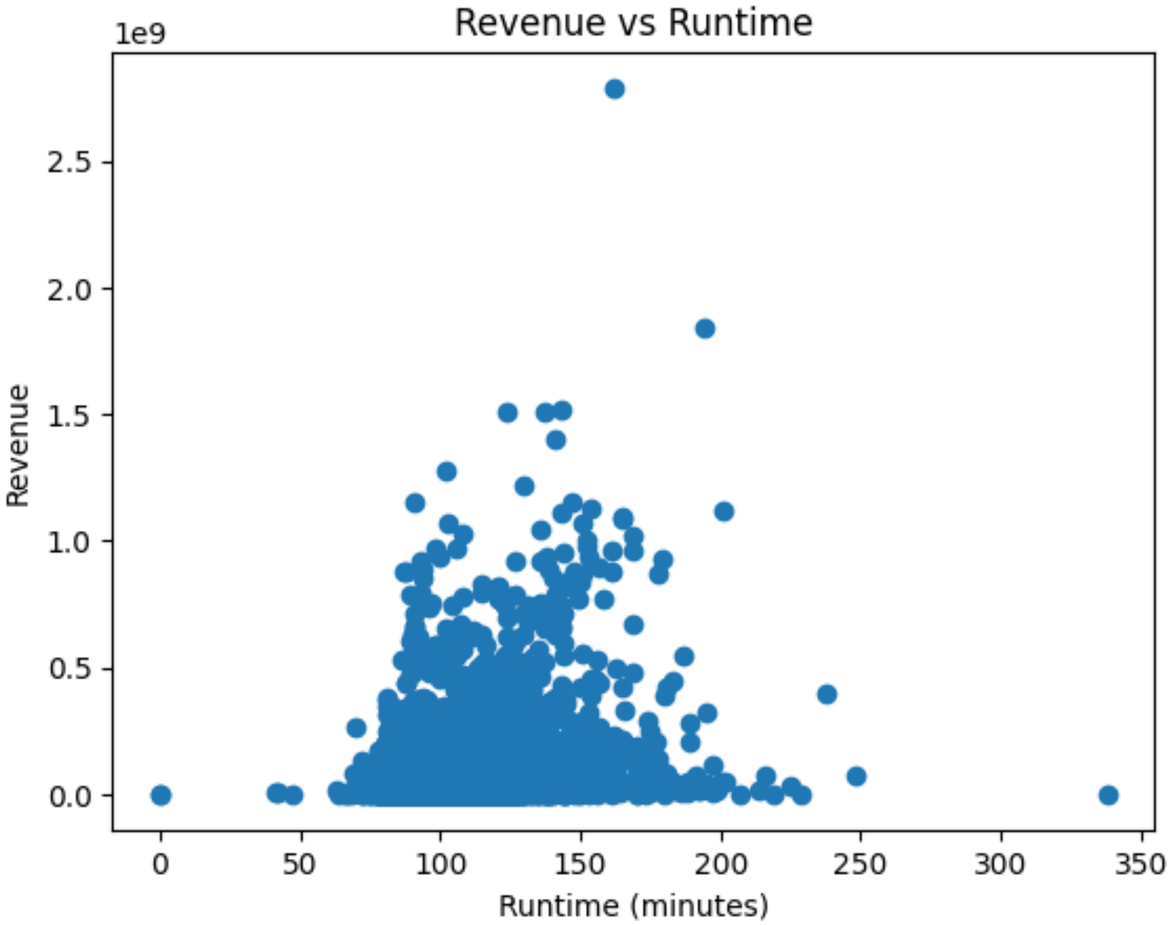
Release date vs revenue

In the plot below I am comparing the relationship between the date in which movies where released and its revenue. It seems as though movies did not make much revenue in the years before 1980, there after the revenue started hiking and from the 2000's onwards there was a great hike in revenue. There are many contributors to the great hike, such as accessibly of movies to people besides having to go to cinemas like in the olden days.



Revenue vs runtime

In the 'revenue vs runtime' scatter plot the positive correlation suggests that longer runtimes may be associated with higher revenues. By visualising revenue in relation to runtime, one can understand the potential impact of movie duration on revenue generation.



THIS REPORT WAS WRITTEN BY : NONOPHA GEGE