Wrangle & Analyze TMDB Movie Data

Wrangle Report

Introduction

The dataset that I will be wrangling is the TMDB Movie this is data contain information about 10,000 movies collected from the movie database.

Project Details

- Gathering data
- Assessing data
- Cleaning data

Gathering data

This data comes from one source TMDB with csv format contains about 10,000 movie with 21 columns contain movie information, this data download manually from Kaggle.

Assessing data

After gathering, the data is assessed for tidiness and quality as follows:

- A sample of data assessed is visually and summary of data types and non-null values is displayed, this allows to identify columns with incorrect data type and null values, Then IDs are checked for duplicates.
- There was two identifier I chose one of them.
- There was a tidiness problem with 4 columns that values are separated with | need to be in individual table.

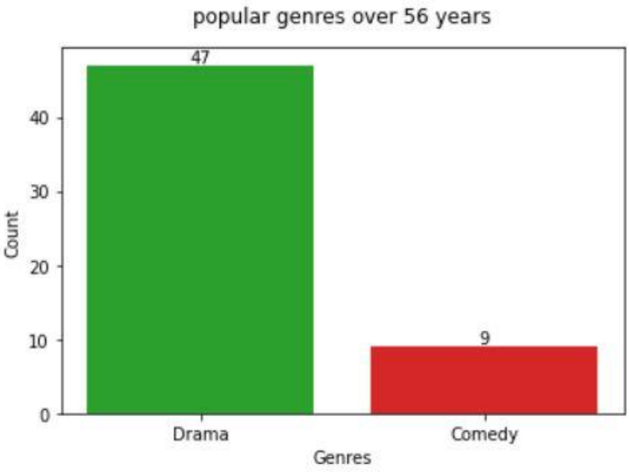
Cleaning data

- As first step, a copy of dataset is created for use through the cleaning.
- We have 1 duplicated id
- Convert the relation from one to one to one to many with some columns that was have a tidiness problem into new table.

Analyze & Insights Report

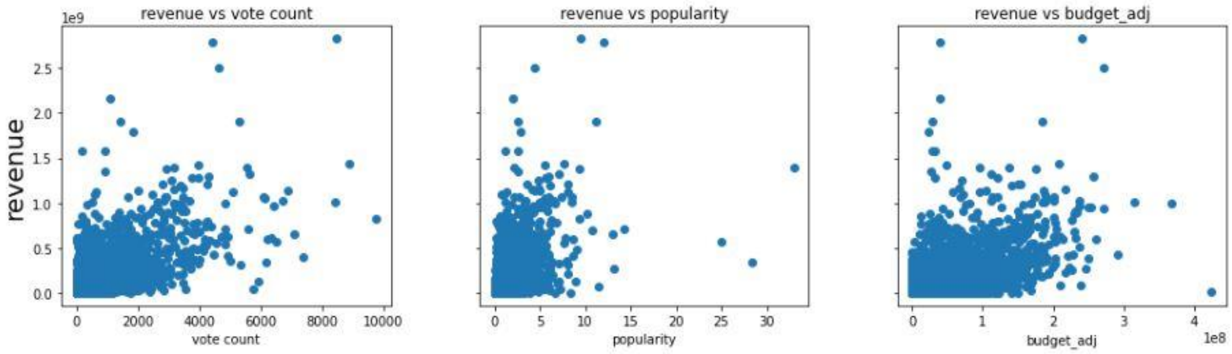
1. Popular genres over years?
2. The most correlation with revenue?
3. The most genre revenue & budget?
4. Top production revenue & budget?
5. Top production company that has product?
6. Revenue & budget & popularity over years?
7. Top 10 movies income?
8. Top 10 years revenue?
9. Top 10 movies losses (budget greater than revenue)?

1- Popular genres over years



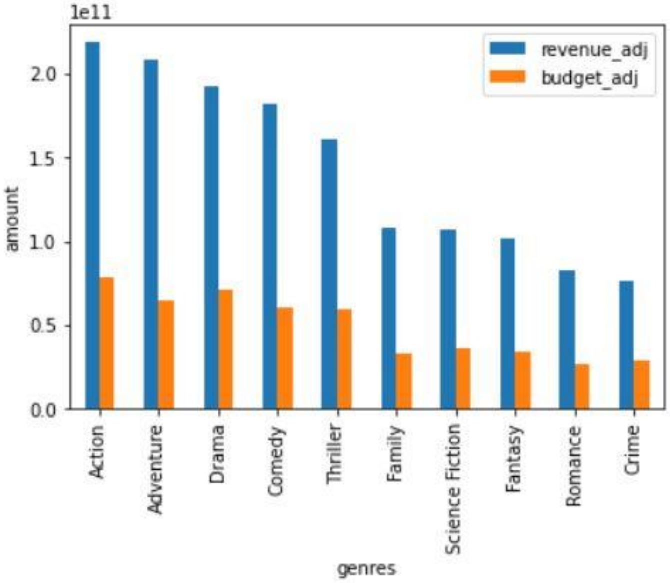
We analyze 56 years i found that the most popular genre over these years is **Drama by 47 years** and **Comedy by 9 years** these is the most popular genres from 1960 to 2015

2- The most correlation with revenue



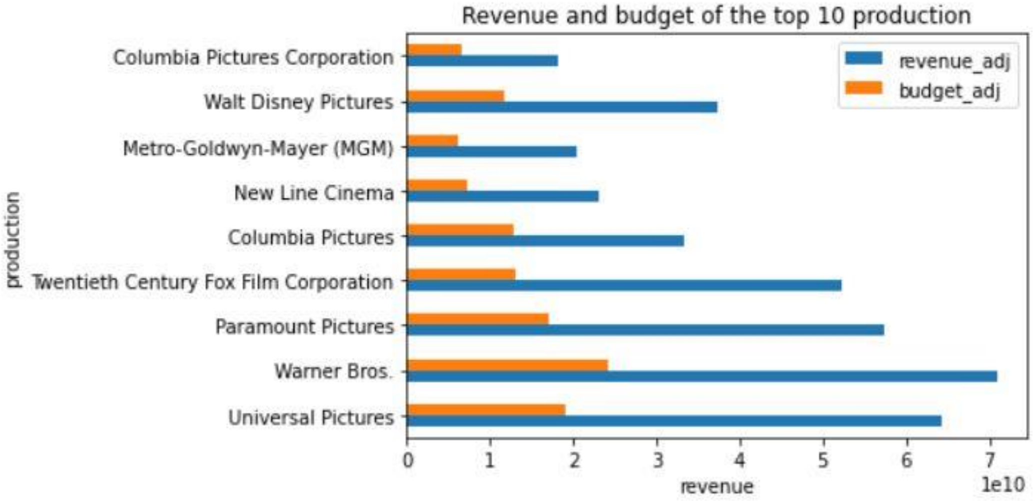
I found that the most correlation with revenue is **vote count** and **budget**

3- The most genre revenue & budget



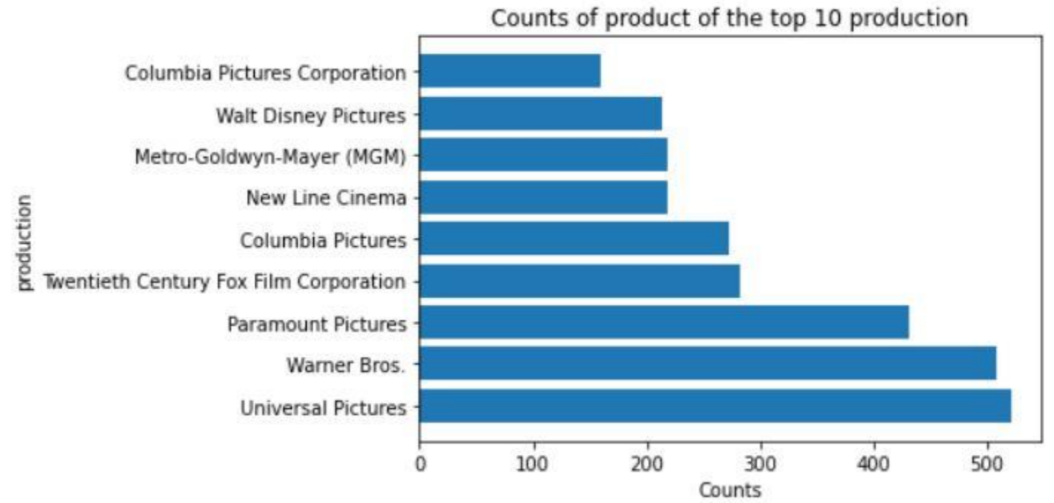
I found that the most genre revenue income and budget spend is **Action**

4- Top production revenue and budget



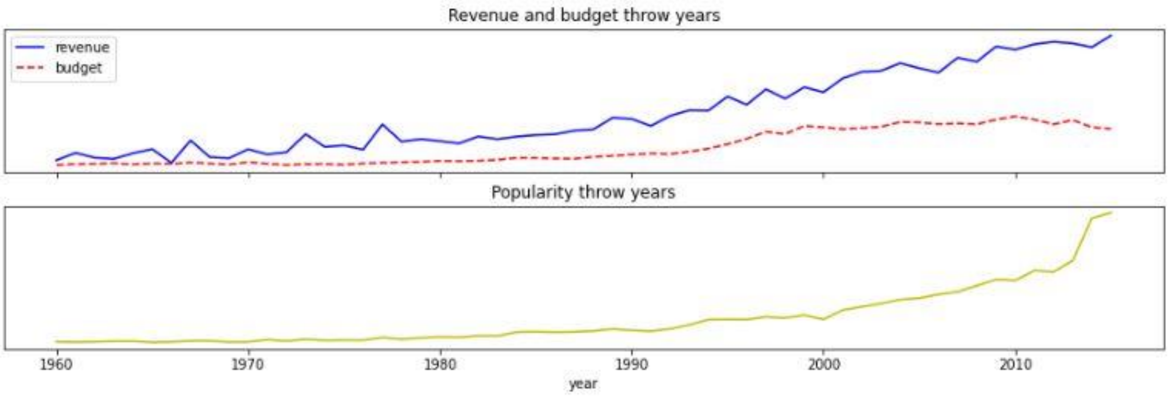
The top production of revenue and budget is **Warner Bros**

5- Top production company that has product



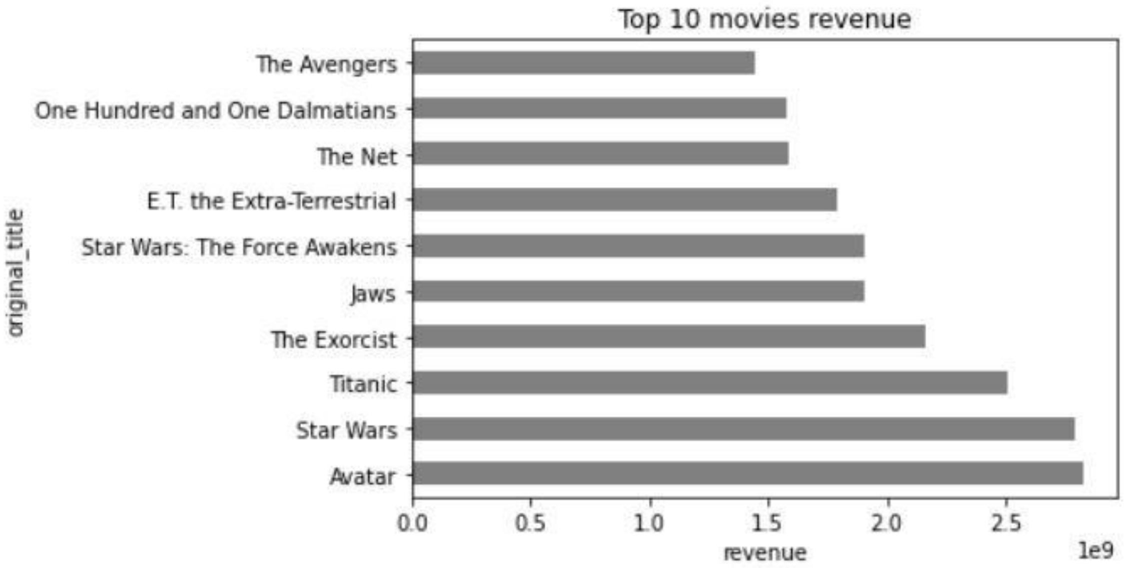
The top production of product is **Universal Picture**

6- Revenue & budget & popularity over years

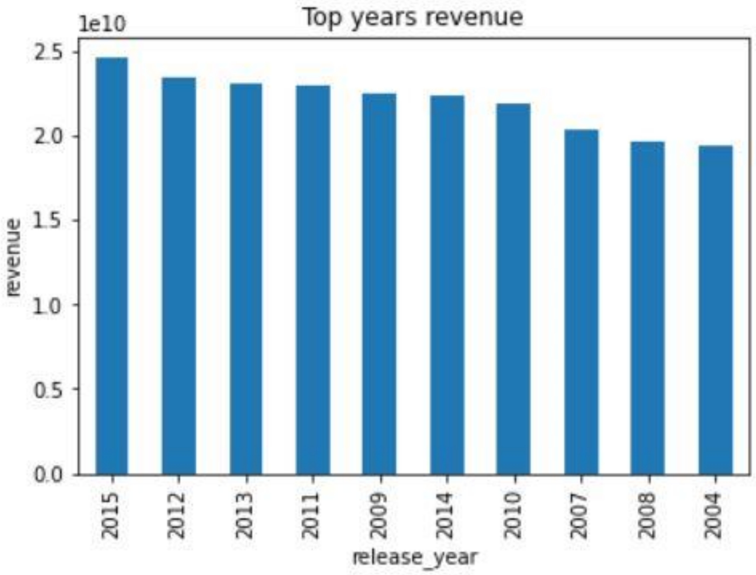


I found that revenue, budget and popularity go up over years

7- Top 10 movies income

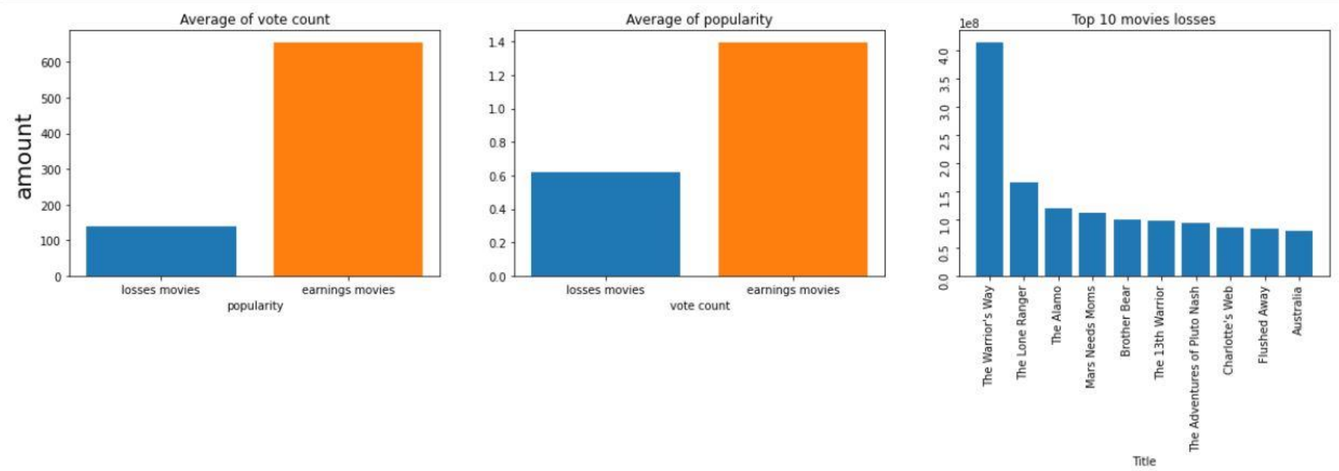


8- Top 10 years revenue



The top year got revenue was **2015**

9- Top 10 movies losses (budget greater than revenue)?



By calculating the average of **vote count** & **popularity** of two types of movie, movies that hasn't revenue and other have revenue, i found that the movies has losses got low popularity and low vote