This Analysis Done By

Zahra Abdelkareem Al-Hourani

Data Analyst Nanodegree Program

Udacity, One Million Arab Coder

June, 2021

Abstract

In this project, I focus on analyzing The Movie Database (TMDB), The Movie Database is a free and open-source database on movies and TV series, The difference with other databases is that TMDb is constantly updated by the community. The overall objective of this analysis is to provide in-depth insight into this dataset. This analysis allows us to draw a comprehensive picture in general about the given dataset, such as the genres that have the highest number of movies to understand the trend in the market. To this end, I use one type of analysis (descriptive analysis). In the descriptive analysis, I describe The TMDB dataset in general. The examined dataset is collected from the Kaggle website and it returns to the period from 1960 – 2015.

1 Introduction

1.1 Overview

In the movies datasbase (TMDB). The analysis of the dataset is very important to know the trends in the film industry. If the film has a low budget, its revenue tends to be low, So we need to know the genre with high revenue to invest in it to make more revenue. even if the genre has the highest number of movies does not mean it will bring high revenue.

1.2 Objective of Analysis

The purpose of our project was to gather and analyze detailed information on the TMDB dataset to provide insights and to answer some questions such as:

1. What kinds of properties are associated with movies that have high revenues?
2. Which genres are most popular from year to year?
3. How long is the average Runtime?
4. How does the Runtime change from year to year?
5. Who are the actors with the most movies?
6. Which Genres that has the most number of movies?
7. Which Production companies have the highest number of movies?
8. Which directors with the most movies?
9. How are the revenue level changes over years?
10. How are the Budget level changes over years?

2 Dataset

2.1Description

For this project, I used the TMDB dataset which is available publicly at the Kaggle data repository. This dataset contains real data about all Movies in the movies database from 1960 – 2015. The dataset contains one file "tmdb-movies.csv" which contains 21 columns and 10866 rows. (see Table 2.1.1) which contains various information about the dataset after I dropped some columns that I will not use in this study.

Table 2.1.1: The detailed attributes of "tmdb-movies.csv"

Attribute Name	Definition	Data Type (Numeric/Categora/Date)
id	Movie ID	Numeric / int
popularity	How much the popularity of specific film	Numeric / float
budget	How much does the film cost?	Numeric / int
revenue	How much revenue did we get from the film?	Numeric / int
Original_title	The title of the film	Categorical / nominal
cast	movie actors	Categorical / nominal
director	movie directors	Categorical / nominal
runtime	Film running time	Numeric / int
Genres	The type of the movie	Categorical / nominal
Production companies	Movie production company	Categorical / nominal
Vote_average	Average Rating of movie	Numeric / float
Vote_count	Count of movie raring	Numeric / int
release_year	Production year	Numeric / int

3 Methodology

The first next step after choosing the dataset is to make some preparation such as formatting and cleaning data. So I made some changes to the data as follows:

- Remove redundant data
- for zero values I chose to convert it to null then drop it
- Dropping NA, and Null values
- Split columns such as 'genres' , 'Production company ' , 'director' , 'cast' that contains ' | '

4 Results and Discussion

In this section, I will make some Descriptive analysis to explore our dataset.

Here first I will do a correlation map to know the relationship between the revenue and the other variables.

Let's start solving the questions I mentioned in the first section:

4.1 What kinds of properties are associated with movies that have high revenues?

To answer this question, I will make a correlation map so that I know the strength of the relationship of each variable with the revenue variable.

Figure 4.1.1: Correlation Map

Through this map, I noticed a very strong relationship between Revenue and Budget , Revenue and vote_count

and there is a strong relationship between Revenue and vote_average , Revenue and Runtime

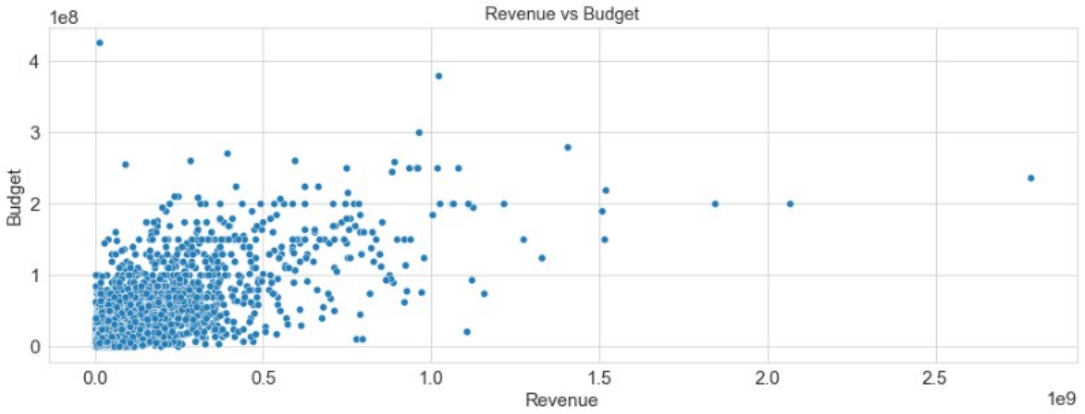

Figure 4.1.2: Revenue vs. Budget

Correlation between Revenue and Budget = 0.69 it is a very strong relationship which means the Movie with a high Budget receives high Revenue.

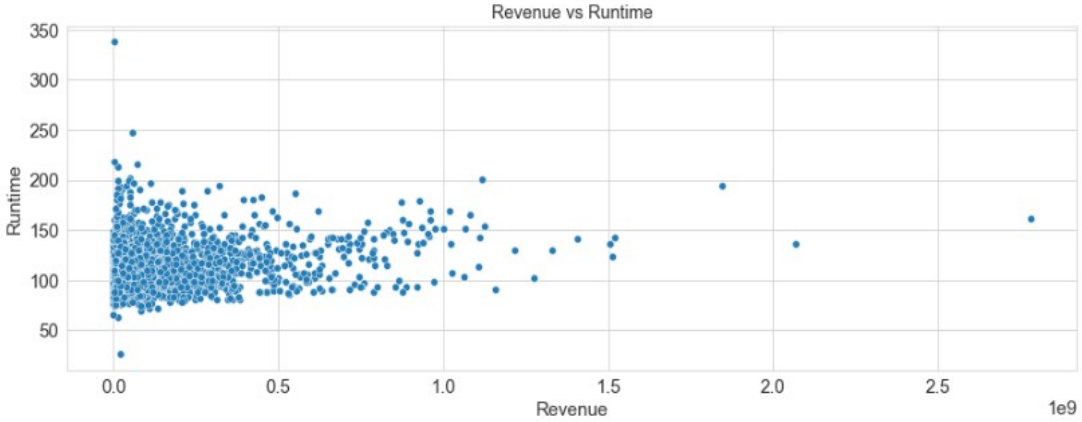

Figure 4.1.3: Revenue vs. Runtime

The correlation between Revenue and Runtime = 0.25 is a strong relationship but I can conclude from the scatter plot that the variables are not related to each other.

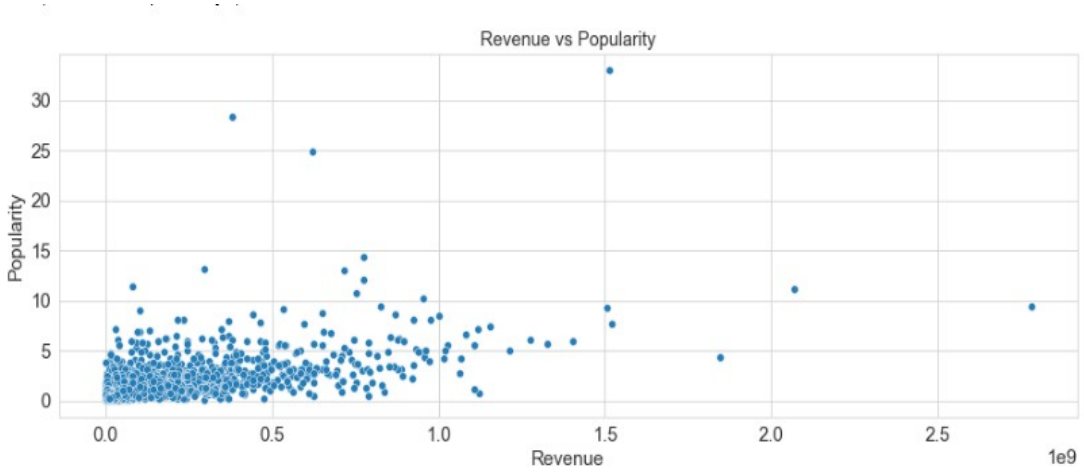

Figure 4.1.4: Revenue vs. Popularity

Correlation between Revenue and Popularity = 0.61 it is a very strong relationship which means the movie with high Popularity tends to bring high Revenue.

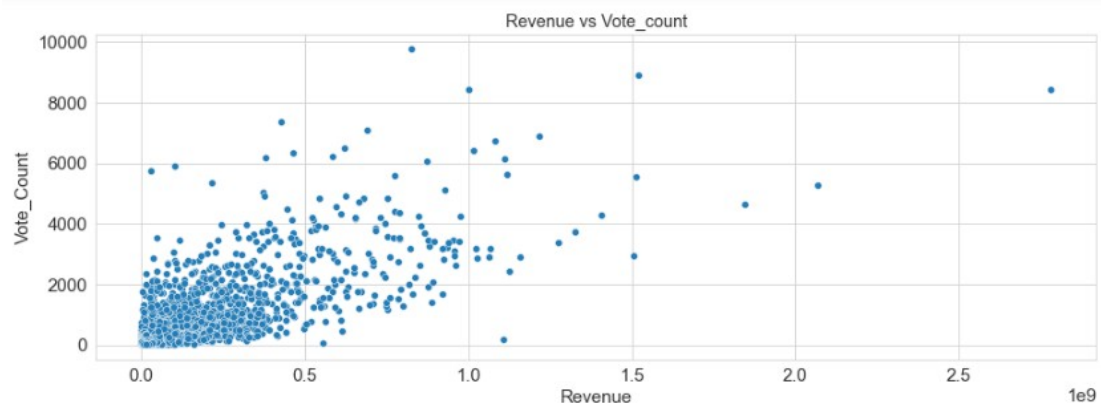

Figure 4.1.5: Revenue vs. Vote_Count

Correlation between Revenue and Popularity = 0.75 it is a very strong relationship which means the movie with a high vote count tend to bring high Revenue.

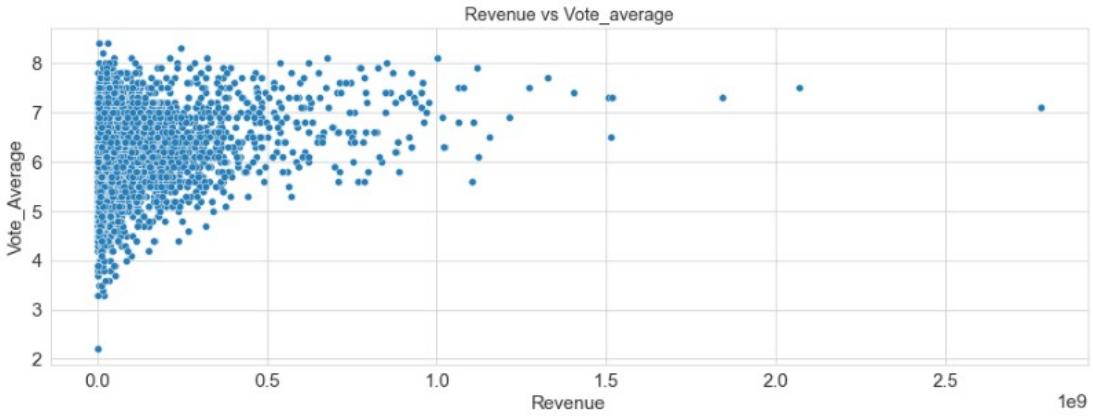

Figure 4.1.6: Revenue vs. Vote_Average

The correlation between Revenue and Runtime = 0.23 it is a strong relationship but I can conclude from the scatter plot that the variables are not related to each other.

4.2 Which genres are most popular from year to year?

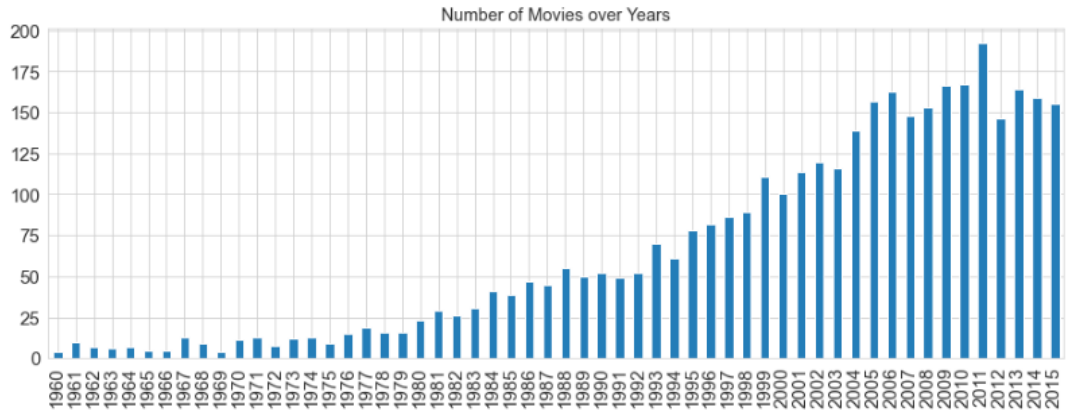

Figure 4.2.1: Number of Movies over Years

2011 was the year with most movies

4.3How long is the average Runtime?

```
#This code calculate the average of runtime
x = tmdb_dataset['runtime'].mean()
x
```

```
109.56132716888769
```

Figure 4.3.1: Average Runtime

From this simple code I can conclude that the Average Runtime = 109 minutes

4.4How does the Runtime change from year to year?

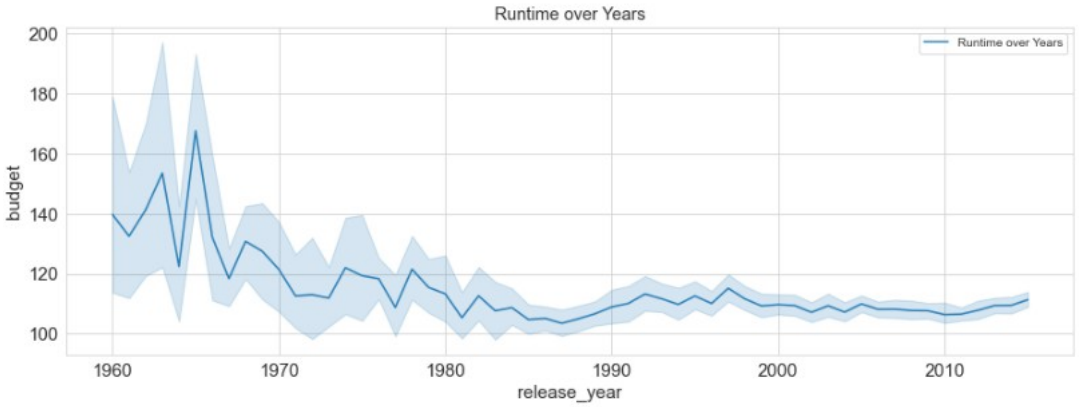

Figure 4.4.1: Runtime over Years

Referring to the Line Chart, I can conclude that the Runtime decreases over the years

4.5Which actor has the most number of movies?

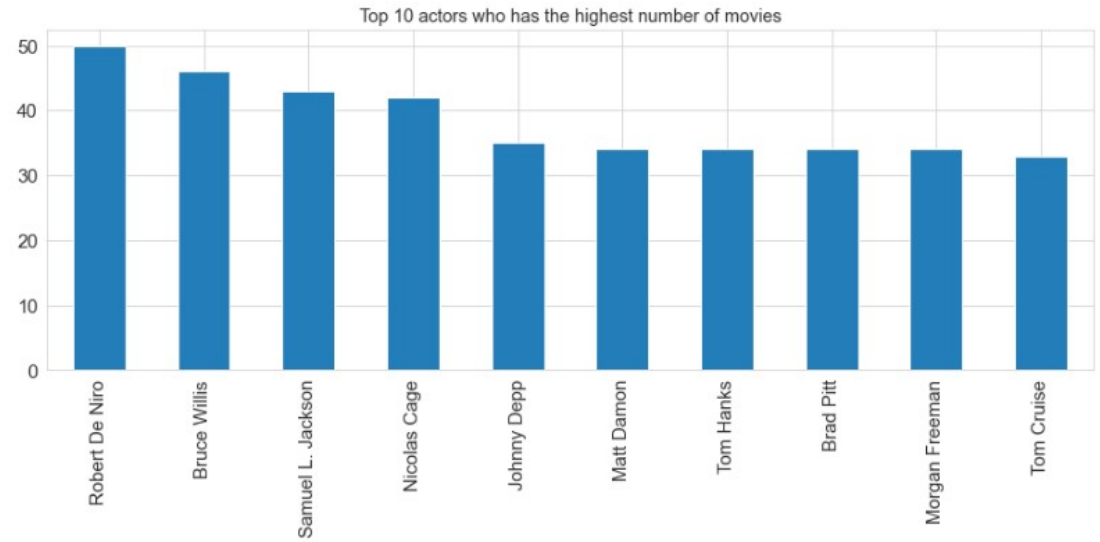

Figure 4.5.1: Actor with highest number of movies

In this Chart, I extracted the data of the top 10 Actors who have the highest number of movies and based on this data the one with most movies was ' Robert De Niro '

4.6Which Genres that has the most number of movies?

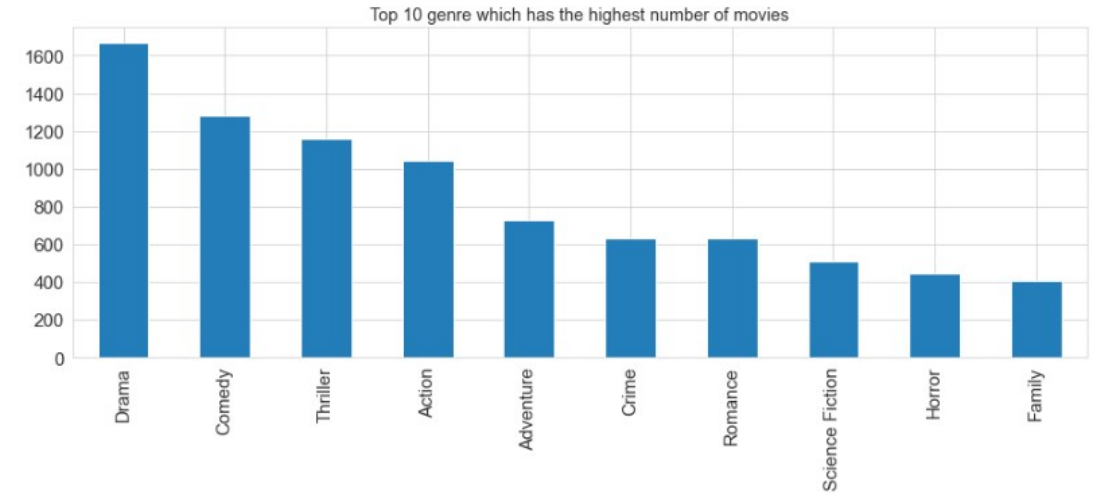

Figure 4.6.1: Genres with highest number of movies

In this Chart, I extracted the data of the top 10 Genres which have the highest number of movies Drama genres were the genre with the most number of movies

4.7 Which directors with the most movies?

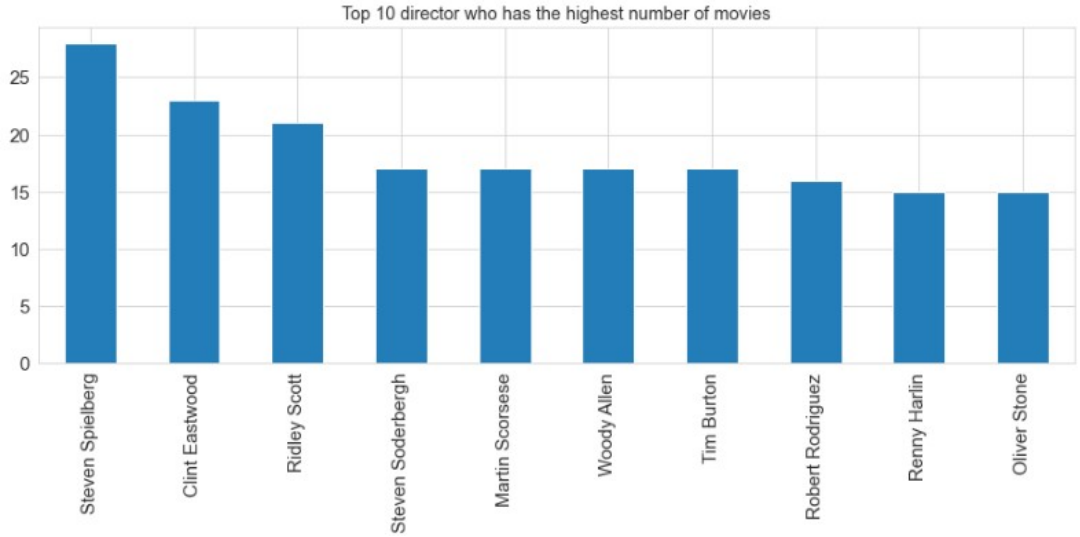

Figure 4.7.1: directors with highest number of movies

In this Chart, I extracted the data of the top 10 directors who have the highest number of movies and the director with most movies was ' Steven Spielberg

4.8Which Production companies have the highest number of movies?

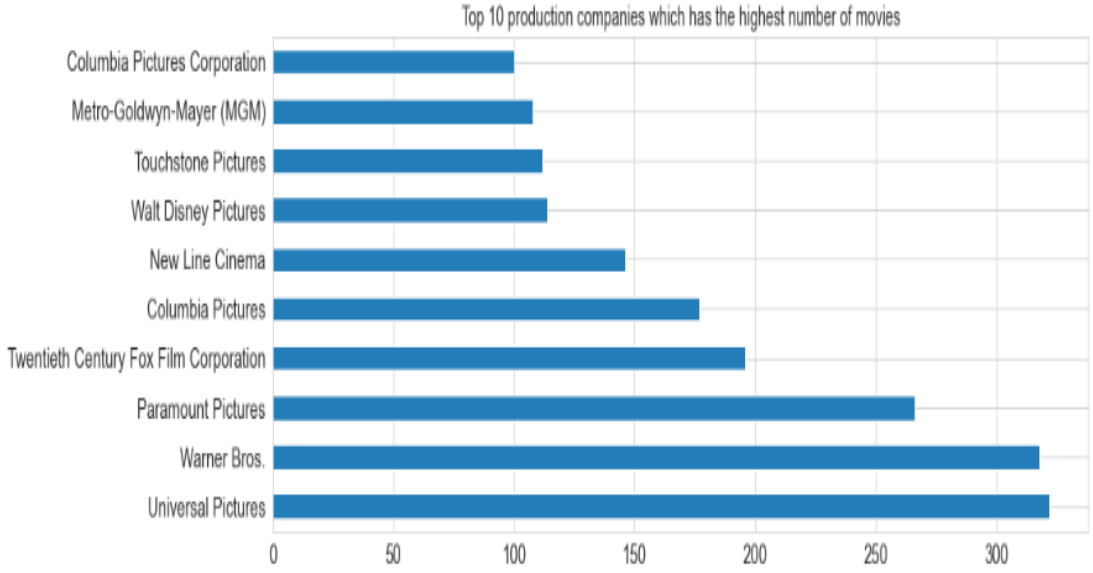

Figure 4.8.1: Production companies with highest number of movies

In this Chart, I extracted the data of the top 10 Production companies which have the highest number of movies and ' Columbia Pictures Corporation ' was the production company with the highest number of movies

4.9 How are the revenue level changes over years?

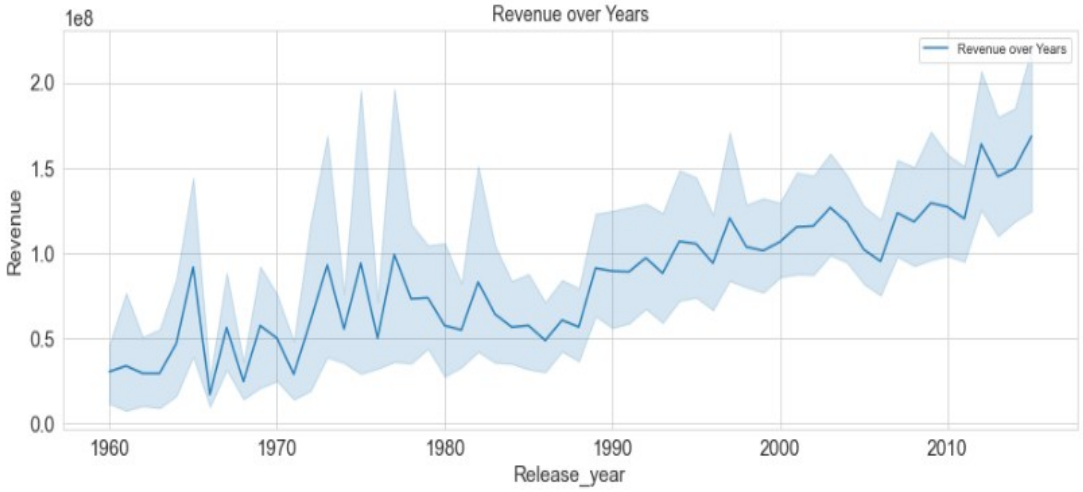

Figure 4.9.1: Changing in revenue over the years

Through the line chart, I conclude that the Revenue increases over the years

4.10 How are the Budget level changes over years?

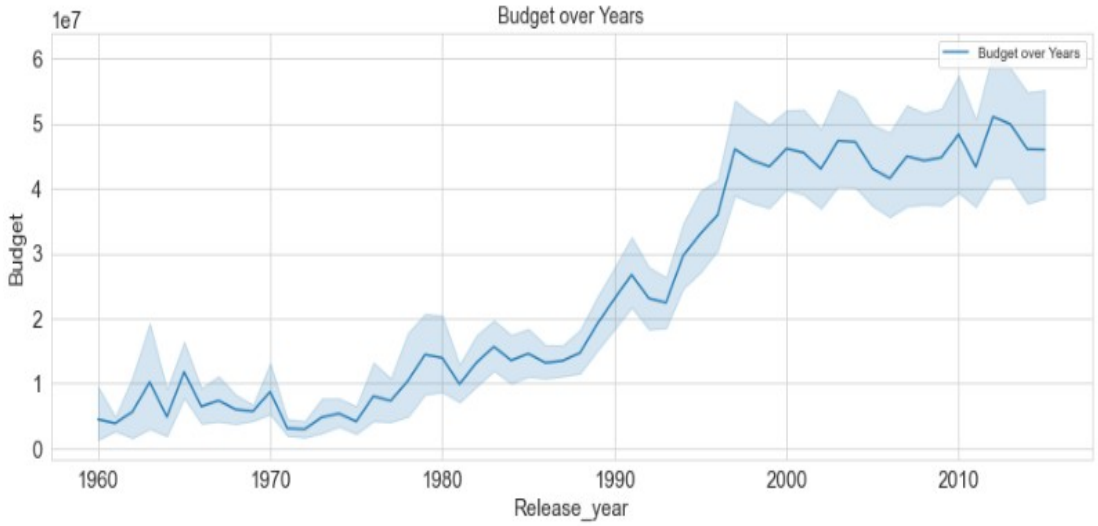

Figure 4.10.1: Changing in revenue over the years

Through the line chart, I conclude that the Budget increases over the years we conclude this earlier in the correlation map, there is a direct relationship between the revenue and the Budget.

5 conclusion

- There is a high correlation between Revenue and Budget = 0.69
- The popularity of the movie results in more revenue
- The high vote count results in more revenue
- The year with the highest number of movies was 2011
- The movies Runtime decreases over the years
- The revenue increases over the years so the film industry is profitable

6 Limitation

- The results of the analysis affected because of Missing values in the dataset
- there is a lot of zero values in the budget, revenue, and Runtime columns and thousands of rows was deleted, I dropped these columns so that affect the results
- Vote_average and popularity columns need to be explained in details like the criteria that they use to calculate them
- The currency must be specified in budget and revenue columns

7 References

1. Seaborn.pydata.org. 2021. *seaborn.heatmap — seaborn 0.11.1 documentation*. [online] Available at: <https://seaborn.pydata.org/generated/seaborn.heatmap.html> [Accessed 4 June 2021].
2. Seaborn.pydata.org. 2021. *seaborn.scatterplot — seaborn 0.11.1 documentation*. [online] Available at: <https://seaborn.pydata.org/generated/seaborn.scatterplot.html> [Accessed 4 June 2021].
3. Seaborn.pydata.org. 2021. *Controlling figure aesthetics — seaborn 0.11.1 documentation*. [online] Available at: <http://seaborn.pydata.org/tutorial/aesthetics.html> [Accessed 4 June 2021].
4. Matplotlib?, H., 2021. *How do I set the figure title and axes labels font size in Matplotlib?*. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/12444716/how-do-i-set-the-figure-title-and-axes-labels-font-size-in-matplotlib> [Accessed 4 June 2021].
5. Seaborn.pydata.org. 2021. *seaborn.lineplot — seaborn 0.11.1 documentation*. [online] Available at: <https://seaborn.pydata.org/generated/seaborn.lineplot.html> [Accessed 4 June 2021].
6. Matplotlib?, H., Seppänen, J. and M, G., 2021. *How do you change the size of figures drawn with Matplotlib?*. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/332289/how-do-you-change-the-size-of-figures-drawn-with-matplotlib> [Accessed 4 June 2021].
7. Pandas.pydata.org. 2021. *pandas.Series.str.cat — pandas 1.2.4 documentation*. [online] Available at: <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.str.cat.html> [Accessed 4 June 2021].
8. Python pandas: Why does df.iloc[:, :. and Kumar, M., 2021. *Python pandas: Why does df.iloc[:, :-1].values for my training data select till only the second last column?*. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/37512079/python-pandas-why-does-df-iloc-1-values-for-my-training-data-select-till> [Accessed 4 June 2021].
9. Pandas.pydata.org. 2021. *pandas.DataFrame.plot.bar — pandas 1.2.4 documentation*. [online] Available at: <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.plot.bar.html> [Accessed 4 June 2021].