TMDb movie data

Author: Aminat Owodunni

Overview

This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.

Aims and Objectives

This project aims to investigate this dataset and answer some specific questions.

Research Questions

1. What genres are the most frequent of all time?
2. Which genres/movies are the most popular of all time?
3. What genres/movies are most popular in the nineties?
4. What genres/movies are most popular in the millennium?
5. Which are the top-rated movies in the nineties?
6. Which are the top-rated movies in the millennium?
7. Which movie title had the highest budget?
8. Which movie title had the longest run time?
9. Which movie actors got the highest vote counts?
10. How is revenue trending over the period of time?
11. How runtime trends overtime?
12. Do top ratings movies always generate big revenue?
13. Do higher budget movies always generate big revenue?
14. What movies generated big revenues?
15. Can we provide a list of directors that generates big revenue?
16. Can we provide a list of production company that generates big revenue?

Comple code can be found on this GitHub repo

Data collection

The dataset was cleaned from original data on Kaggle.

Data importation and pre-processing

The data after being imported into the jupyter notebook was scrutinized thoroughly. The dataset was seen to contain both categorical and numerical variables. The raw dataset contains about 10865 rows and 21 columns.

Columns in the dataset

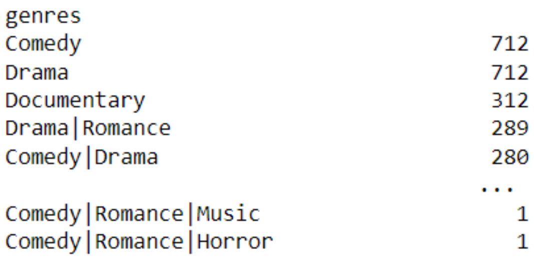
Id	- movie id
imdb_id	- movie online database id info.
popularity	- movie rate of demand
budget	- initial amount budgeted for movie
revenue	- initial revenue generated
original_title	- movie title

cast	- featured actors in movie
homepage	- movie homepage
director	- movie director
tagline	- slogan used to advertise movie
keywords	- keyword movie were found for
overview	- movie summary
runtime	- movie duration in minutes
genres	- movie category
production_companies	- company that produced the movie
release_date	- date movie is released to public
vote_count	- compilation of movie vote
vote_average	- movie average rating on a scale 0/10
release_year	- year movie is releases to public
budget_adj	- adjusted budget due to inflation
revenue_adj	- adjusted revenue due to inflation

Exploratory Data Analysis (EDA)

Exploratory data analysis is performed on the given dataset to gain more insights about the dataset in terms of its summary statistics and visualizations of various variables in the dataset.

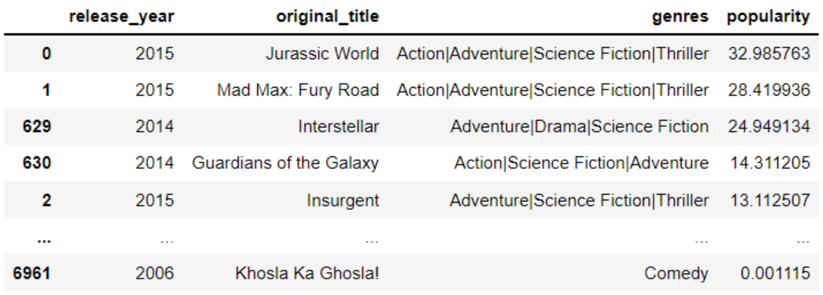
1. What genres are the most frequent of all time?



```
genres
Comedy                          712
Drama                           712
Documentary                     312
Drama|Romance                   289
Comedy|Drama                    280
                               ...
Comedy|Romance|Music              1
Comedy|Romance|Horror             1
```

The most frequent genres were **comedy** and **Drama** with a count of **712**.

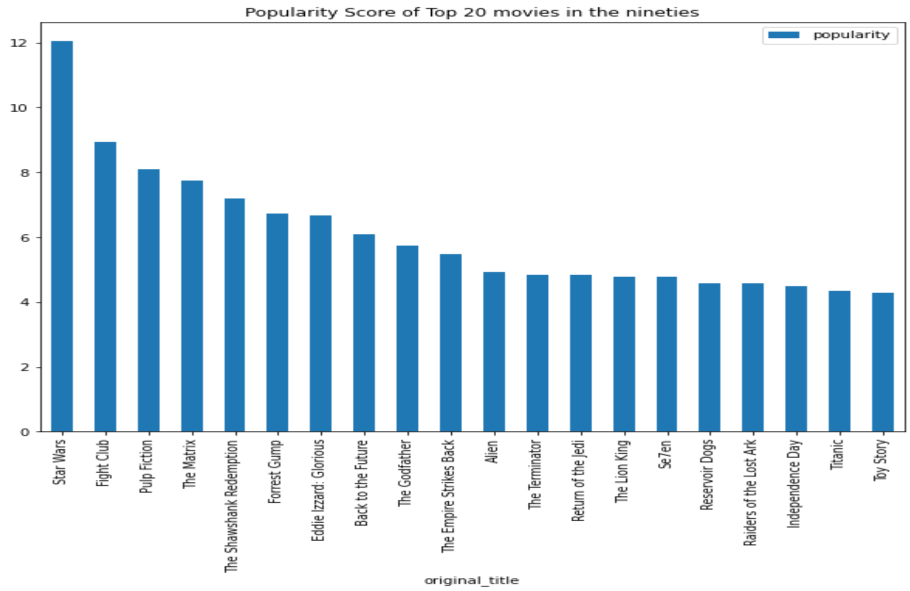
2. What genres/movies are most popular of all time?



	release_year	original_title	genres	popularity
0	2015	Jurassic World	Action|Adventure|Science Fiction|Thriller	32.985763
1	2015	Mad Max: Fury Road	Action|Adventure|Science Fiction|Thriller	28.419936
629	2014	Interstellar	Adventure|Drama|Science Fiction	24.949134
630	2014	Guardians of the Galaxy	Action|Science Fiction|Adventure	14.311205
2	2015	Insurgent	Adventure|Science Fiction|Thriller	13.112507
...	...	...	...	...
6961	2006	Khosla Ka Ghosla!	Comedy	0.001115

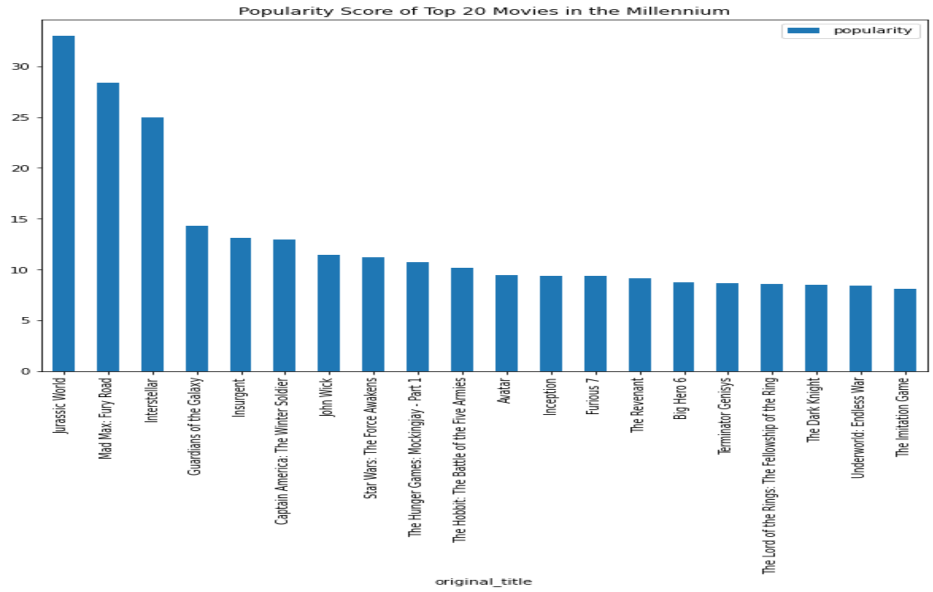
The popular genres all the time are **Action|Adventure|Science Fiction|Thriller** movies. **Jurassic world** was the most popular with a score of **32.985763**.

3. **What genres/movies that are most popular in the nineties**?



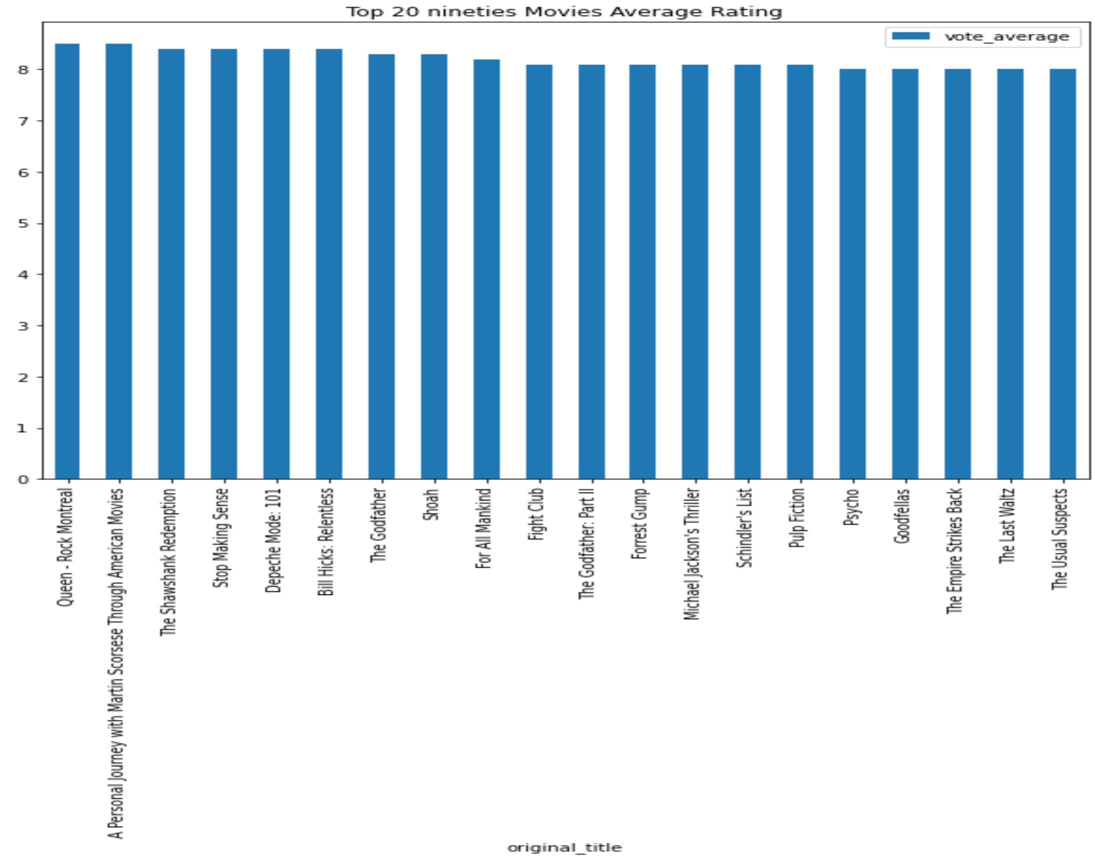
The most popular genres are in the nineties are; **Action|Adventure|Science Fiction movies. Star Wars**, a movie, released in **1977** was the most popular with a score of **12.037933**

4. **What genres/movies are most popular in the millennium?**



The most popular genres in the nineties are; **Action|Adventure|Science Fiction** movies. **Star Wars,** a movie, released in **1977** was the most popular with a score of **12.037933**

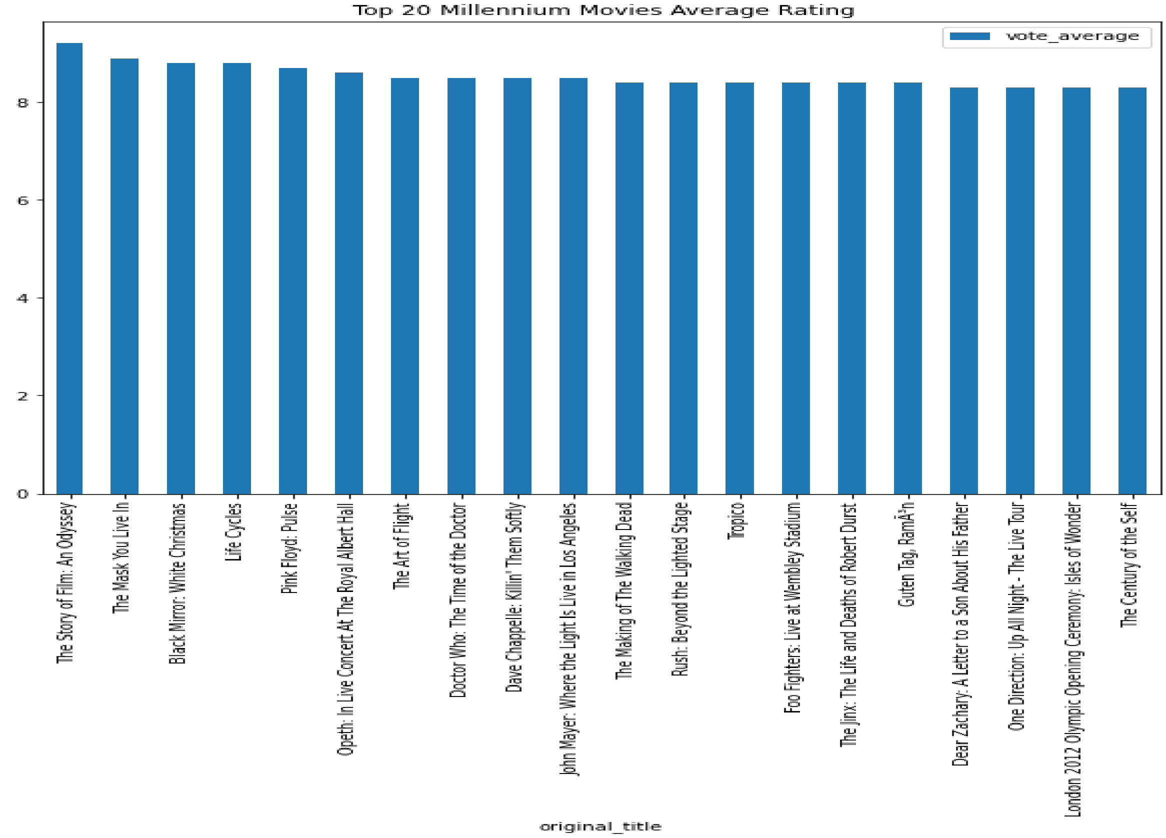
5. Which are the top-rated movies in the nineties?



Top rated genres in the nineties are movies of the category: **Documentary, Music, Drama, Crime**.

The two most rated movies are; **Queen _ Rock Montreal released in 1981** and **A Personal Journey with Martin Scorsese** released in **1995** both with average rating of **8.5.**

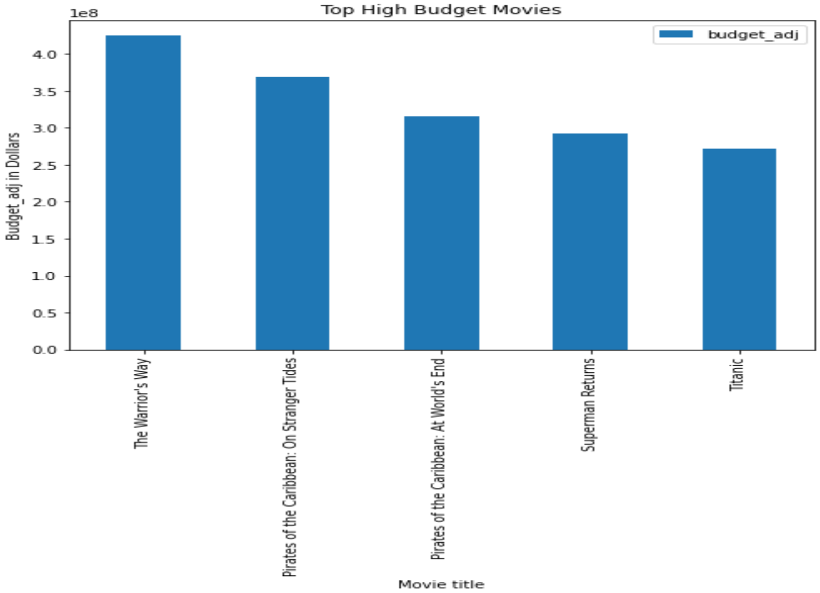
6. **Which are the top-rated movies in the millennnium?**



Top rated movies in the millennium are movies of category: **Drama, Horror, Mystery, Science Fiction, Thriller, Documentary**.

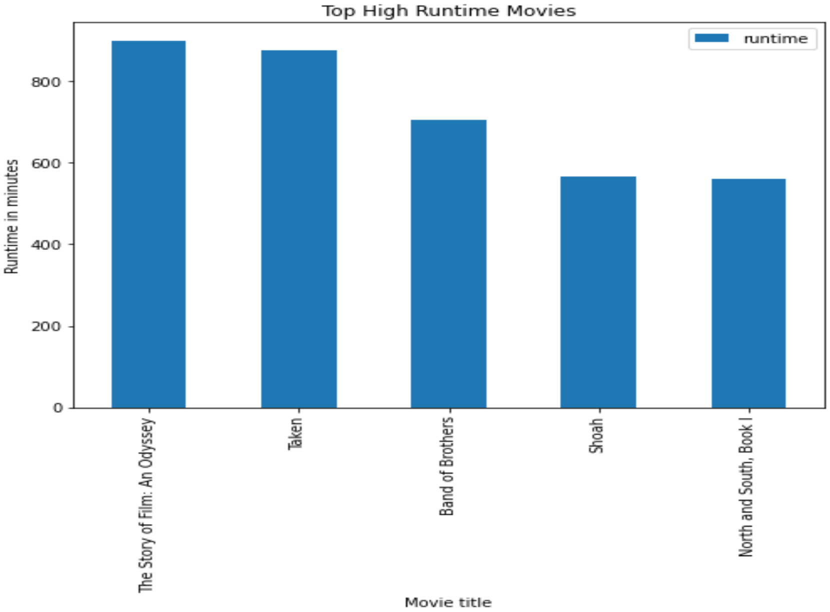
The most rated movie of the millennium is **The Story of Film: An Odyssey** released in **2011** with an average rating of **9.2**

7. **Which movie title had the highest budget?**



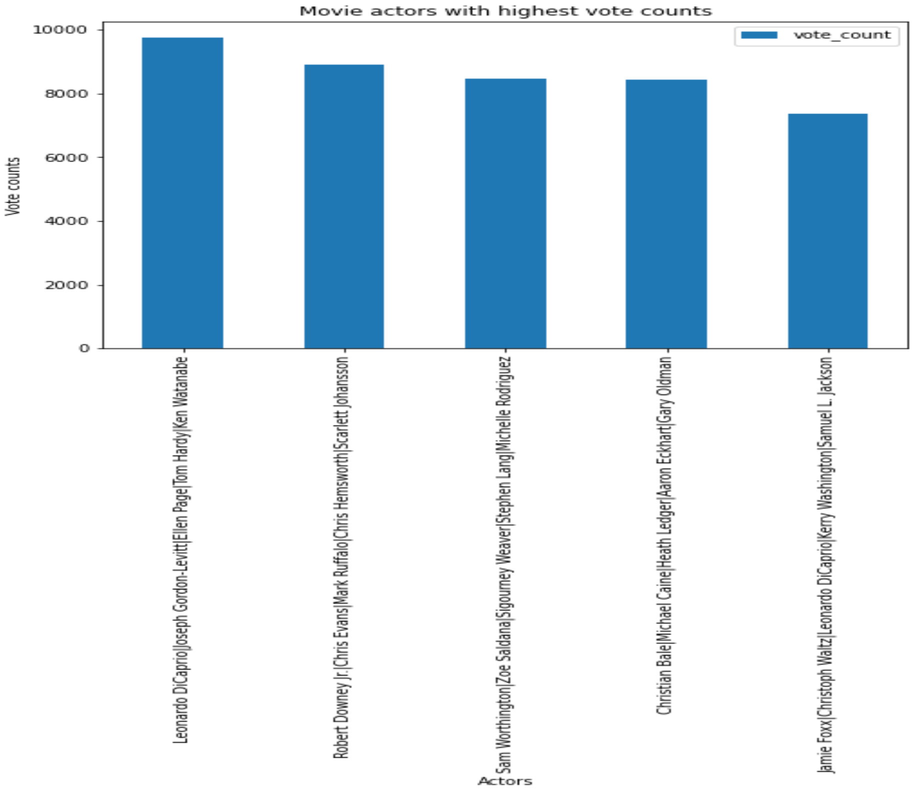
The Warrior's way was the movie with the highest budget

8. **Which movie title had the highest run time?**



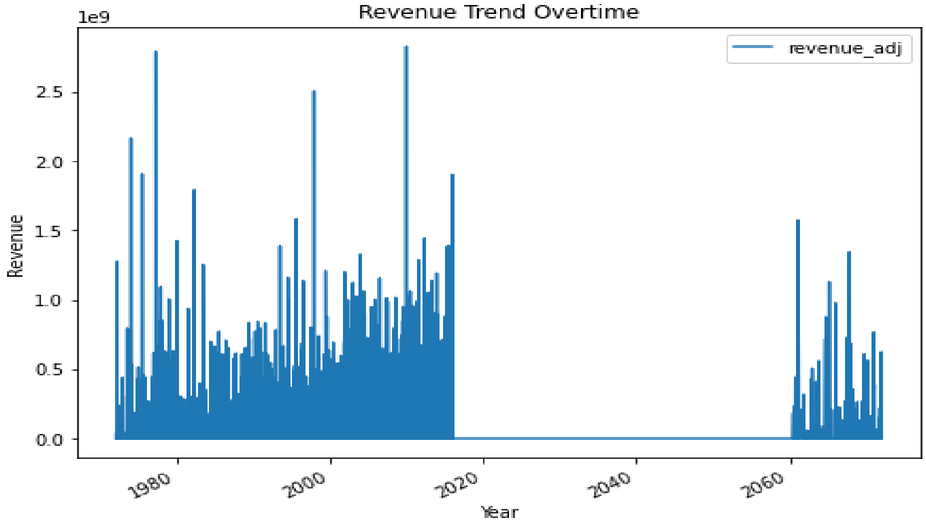
The Story of Film: An Odyssey released in **2011** had the highest runtime.

9. **Which movie actors got the highest vote counts?**



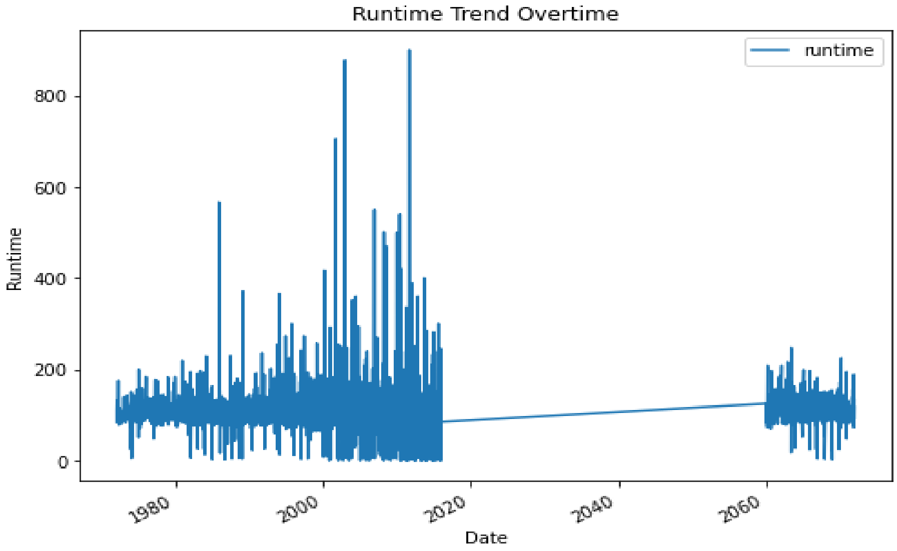
The movie actors with the highest vote counts are **Leonardo DiCaprio, Joseph Gordon-Levitt, Ellen Page, Tom Hardy, Ken Wantanabae**.

10. **How is revenue trending over the period of time?**



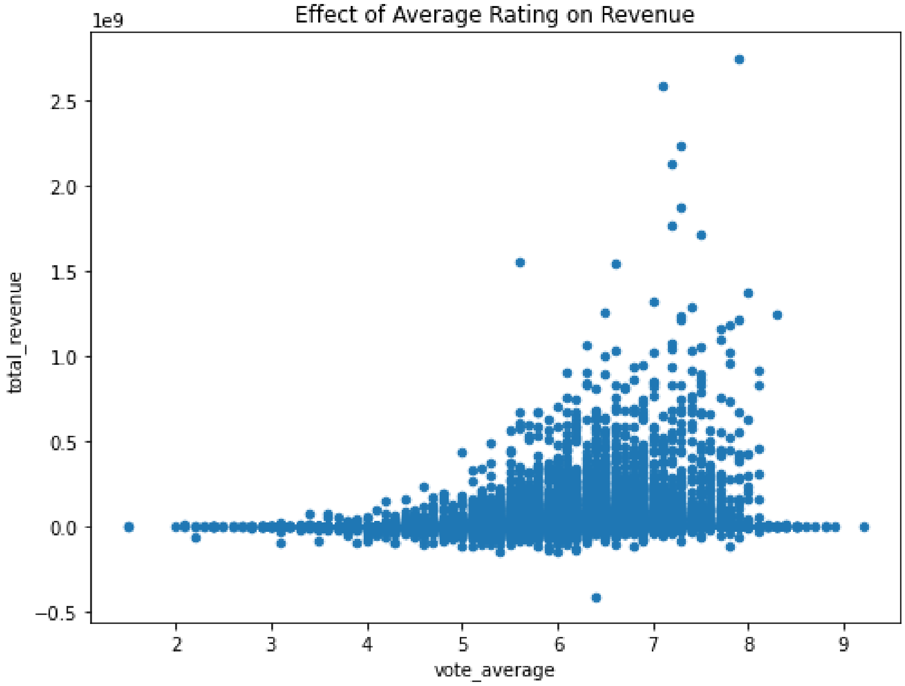
Revenue does not seem to follow a particular trend.

11. **How runtime trends overtime?**



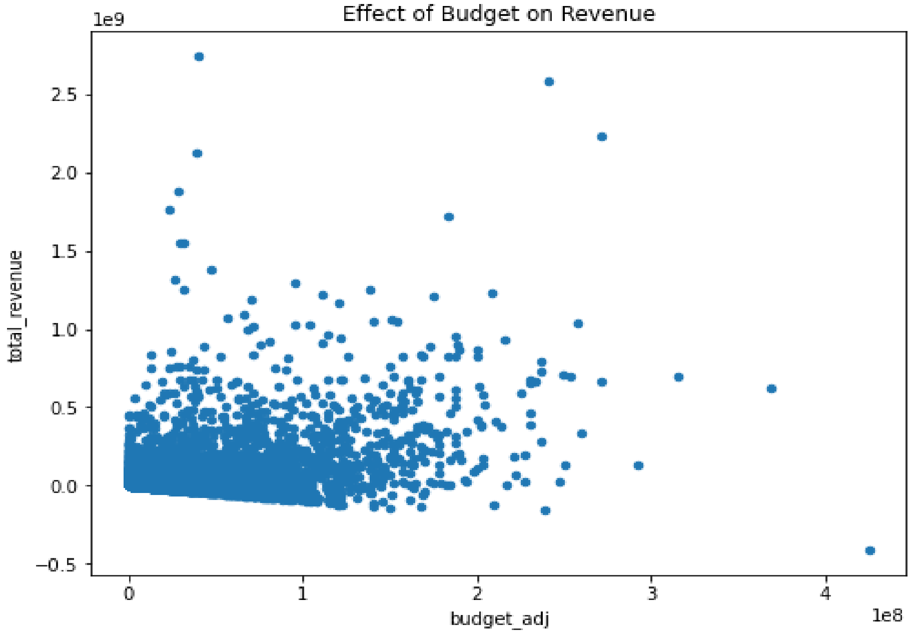
It can be seen that runtime increased over the years.

12. **Do top ratings movies always generate big revenue?**



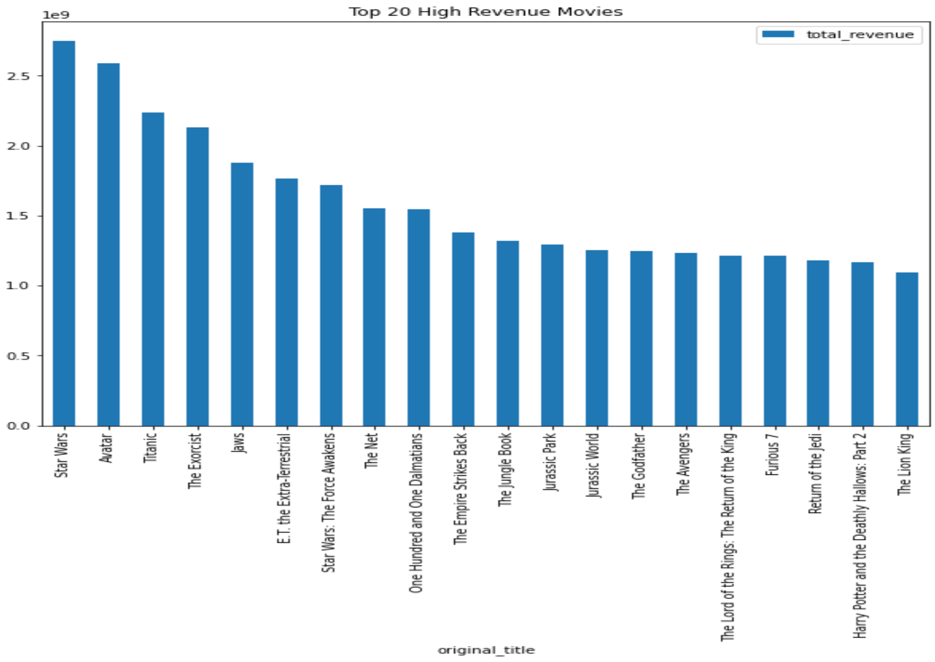
As seen from the plot, there is a relationship between movie rating and revenue. Top rated movies generate high revenue.

13. **Do higher budget movies always generate big revenue?**



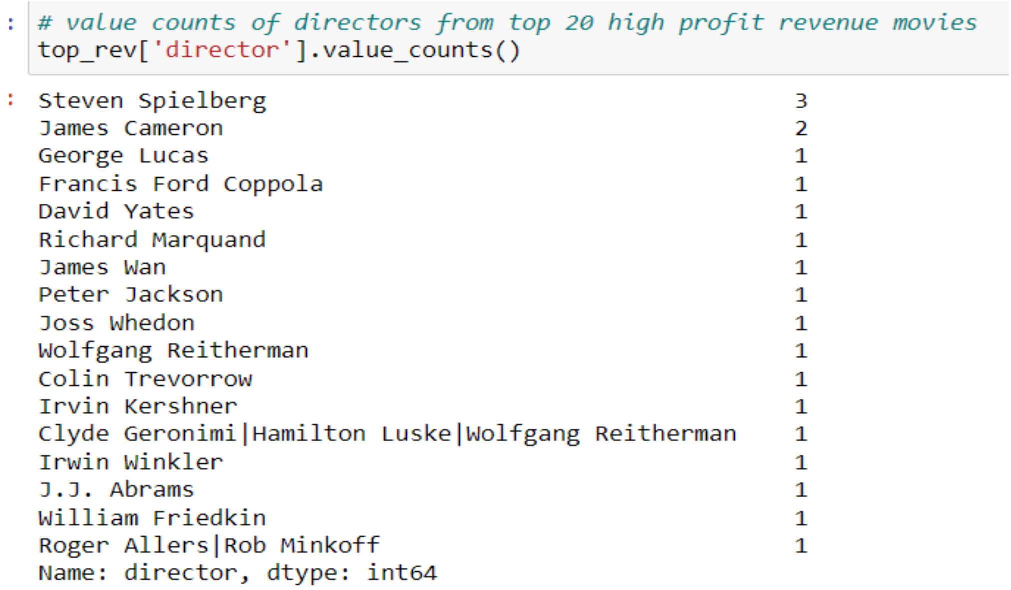
High budget movies generate high revenue.

14. What movies generated big revenues?



Star Wars was the movie that generated the biggest revenue.

15. Can we provide a list of directors that generates big revenue?

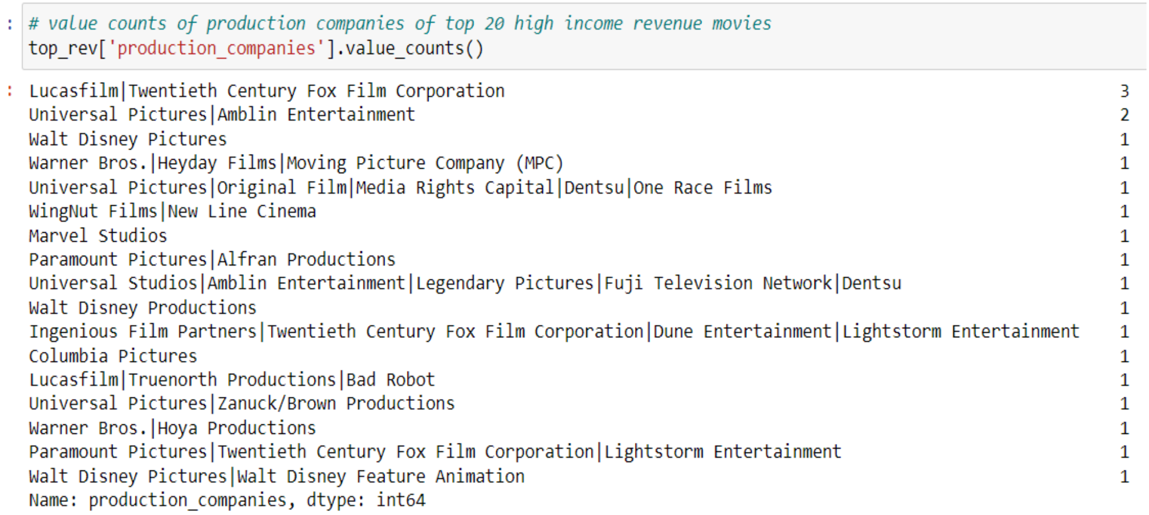


```
# value counts of directors from top 20 high profit revenue movies
top_rev['director'].value_counts()
```

```
Steven Spielberg                                      3
James Cameron                                         2
George Lucas                                          1
Francis Ford Coppola                                  1
David Yates                                           1
Richard Marquand                                      1
James Wan                                             1
Peter Jackson                                         1
Joss Whedon                                           1
Wolfgang Reitherman                                   1
Colin Trevorrow                                       1
Irvin Kershner                                        1
Clyde Geronimi|Hamilton Luske|Wolfgang Reitherman     1
Irwin Winkler                                         1
J.J. Abrams                                           1
William Friedkin                                      1
Roger Allers|Rob Minkoff                              1
Name: director, dtype: int64
```

Steven Spielberg and **James Cameron** are directors that generates big revenue

16. **Can we provide a list of production company that generates big revenue?**



```
# value counts of production companies of top 20 high income revenue movies
top_rev['production_companies'].value_counts()
```

```
Lucasfilm|Twentieth Century Fox Film Corporation                                                              3
Universal Pictures|Amblin Entertainment                                                                       2
Walt Disney Pictures                                                                                          1
Warner Bros.|Heyday Films|Moving Picture Company (MPC)                                                        1
Universal Pictures|Original Film|Media Rights Capital|Dentsu|One Race Films                                   1
WingNut Films|New Line Cinema                                                                                 1
Marvel Studios                                                                                                1
Paramount Pictures|Alfran Productions                                                                         1
Universal Studios|Amblin Entertainment|Legendary Pictures|Fuji Television Network|Dentsu                      1
Walt Disney Productions                                                                                       1
Ingenious Film Partners|Twentieth Century Fox Film Corporation|Dune Entertainment|Lightstorm Entertainment    1
Columbia Pictures                                                                                             1
Lucasfilm|Truenorth Productions|Bad Robot                                                                     1
Universal Pictures|Zanuck/Brown Productions                                                                   1
Warner Bros.|Hoya Productions                                                                                 1
Paramount Pictures|Twentieth Century Fox Film Corporation|Lightstorm Entertainment                            1
Walt Disney Pictures|Walt Disney Feature Animation                                                            1
Name: production_companies, dtype: int64
```

Most of the movies that generated high income were produced by **Lucasfilm, Twentieth Century Fox Film Corporation, Universal Pictures, Walt Disney, Paramount pictures, Warner Bros., Columbia pictures** among others.

Conclusion

The following were deduced from the dataset;

The most frequent genres were Comedy and Drama with a count of 712.
The popular genres all the time are Action|Adventure|Science Fiction|Thriller movies. Jurassic world was the most popular with a score of 32.985763.
The two most rated movies in the nineties were; Queen _Rock Montreal released in 1981 and A Personal Journey with Martin Scorsese released in 1995 both with average rating of 8.5.
The most rated movie of the millennium is The Story of Film: An Odyssey released in 2011 with an average rating of 9.2
The Warrior's way was the movie with the highest budget
The Story of Film: An Odyssey released in 2011 had the highest runtime.
The movie actors with the highest vote counts are Leonardo DiCaprio, Joseph Gordon-Levitt, Ellen Page, Tom Hardy, Ken Wantanabae.
Revenue does not seem to follow a particular trend.
It can be seen that runtime increased over the years.
There is a correlation between movie rating and revenue. Top rated movies generate high revenue.
High budget movies generate high revenue.
Star Wars was the movie that generated the biggest revenue.
Steven Spielberg and James Cameron are directors that generates big revenue.
Most of the movies that generated high income were produced by Lucasfilm, Twentieth Century Fox Film Corporation, Universal Pictures, Walt Disney, Paramount pictures, Warner Bros., Columbia pictures among others.

Limitation of the dataset: There are columns with more than 50% missing values e.g. homepage column. The release_date column has some erroneous values.