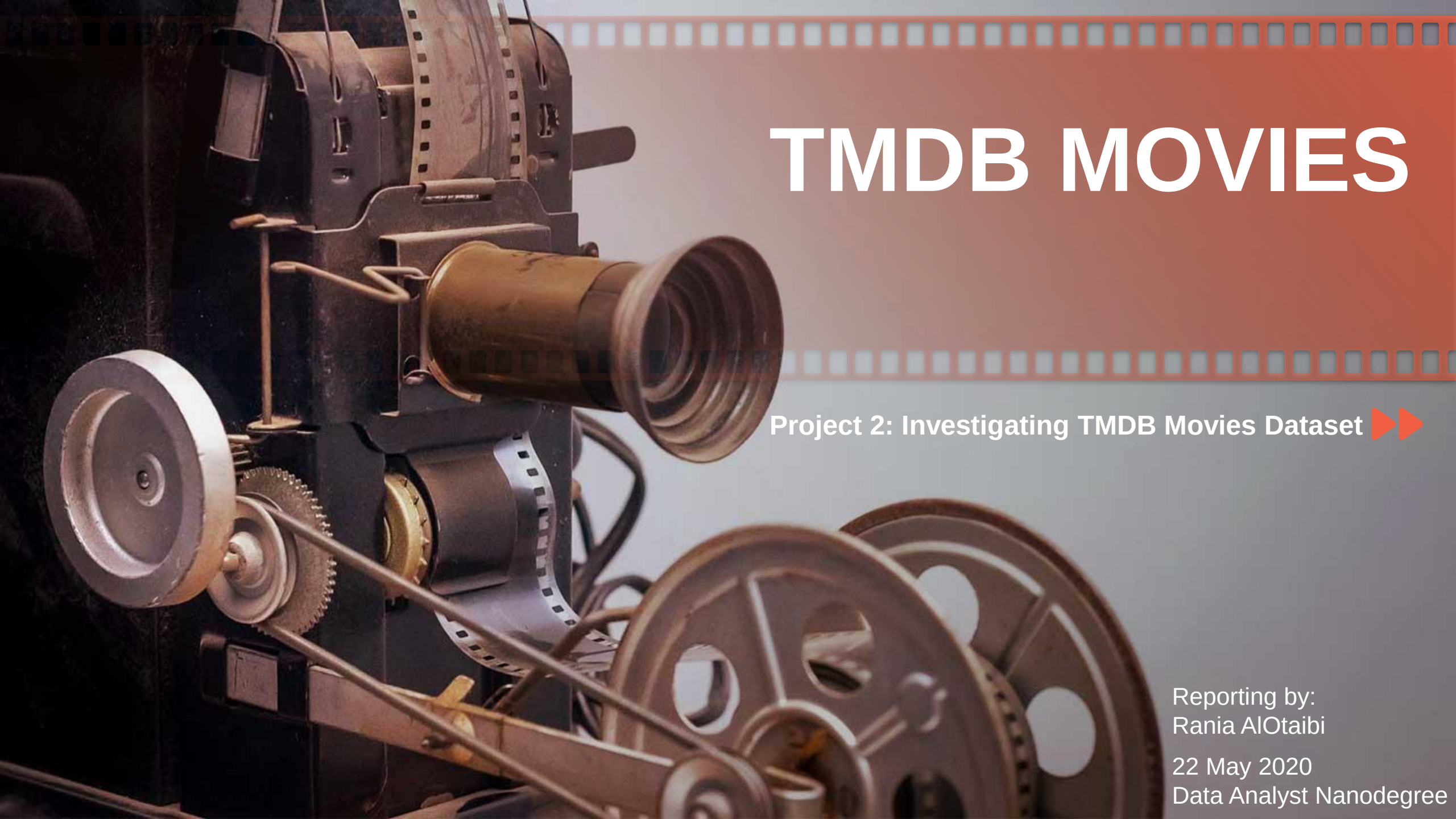
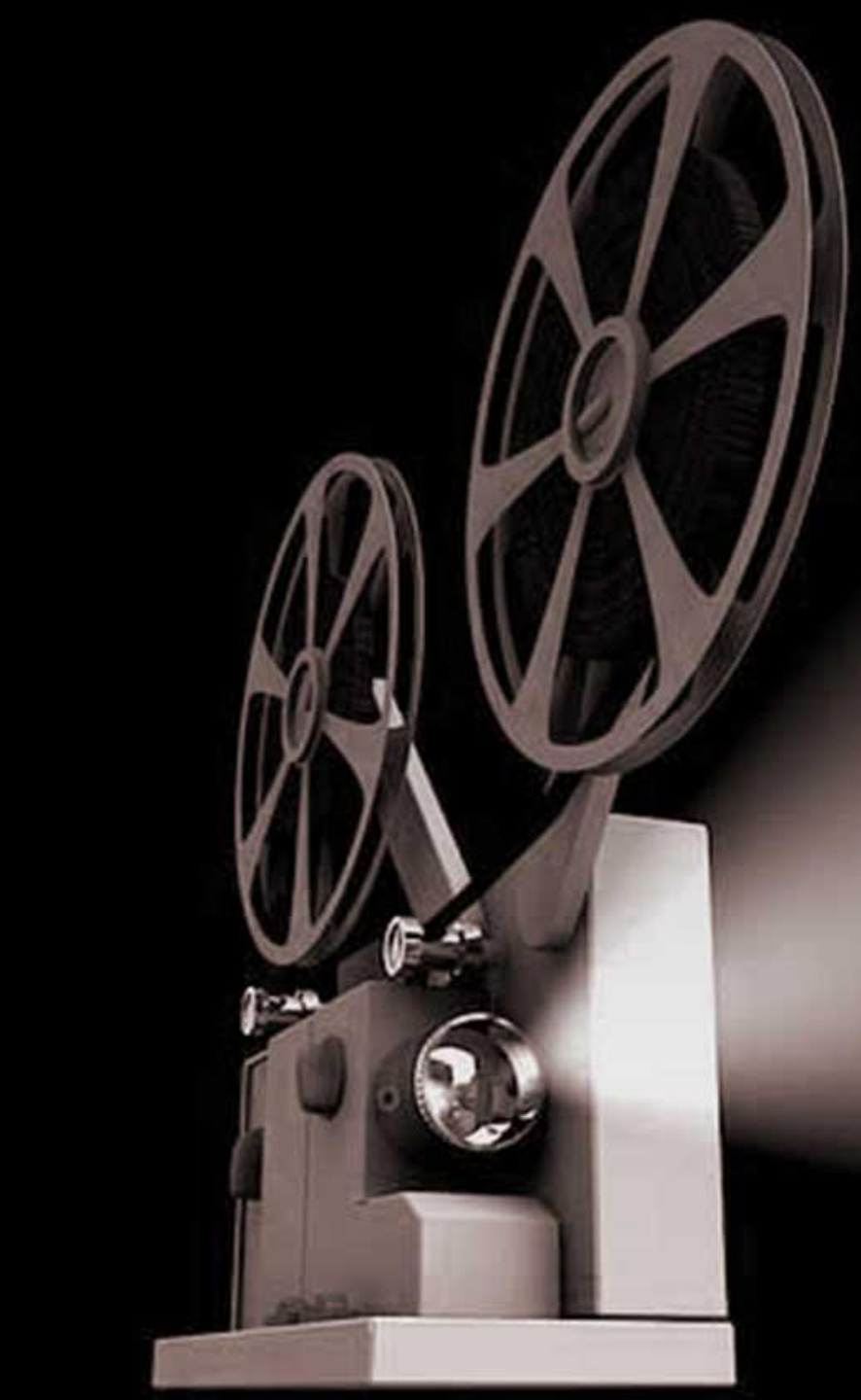


TMDB MOVIES

Project 2: Investigating TMDB Movies Dataset

Reporting by:
Rania AlOtaibi

22 May 2020
Data Analyst Nanodegree



Introduction:

In this project , I will analyze (TMDB) dataset and then communicate my findings about it. I will use the This TMDB Movies data set contains information about 10,000 movies collected from The Movie Database (TMDB),also there are various factors could be effect the quality of a films, as for instance the revenue , genre , budget etc..

Research Question To Analysis

Q1 What movies genres are most common in general?.

Q2 What are the highest 5 movies in profit ?

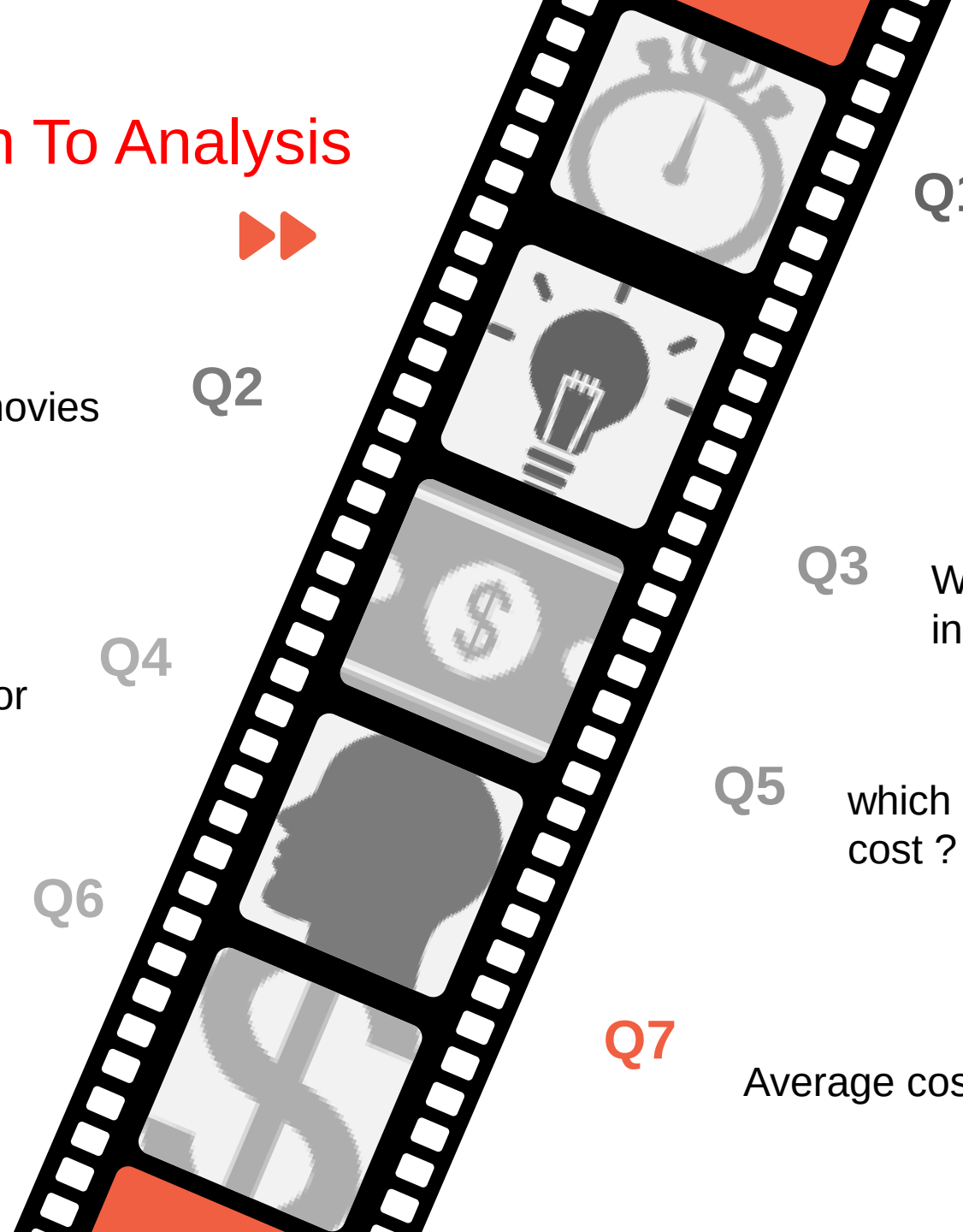
Q3 What are the highest 5 movies in expenses ?.

Q4 what are the top 5 director earned highest profit ?

Q5 which year that has the highest cost ?

Q6 Average Profit earned per movies ?

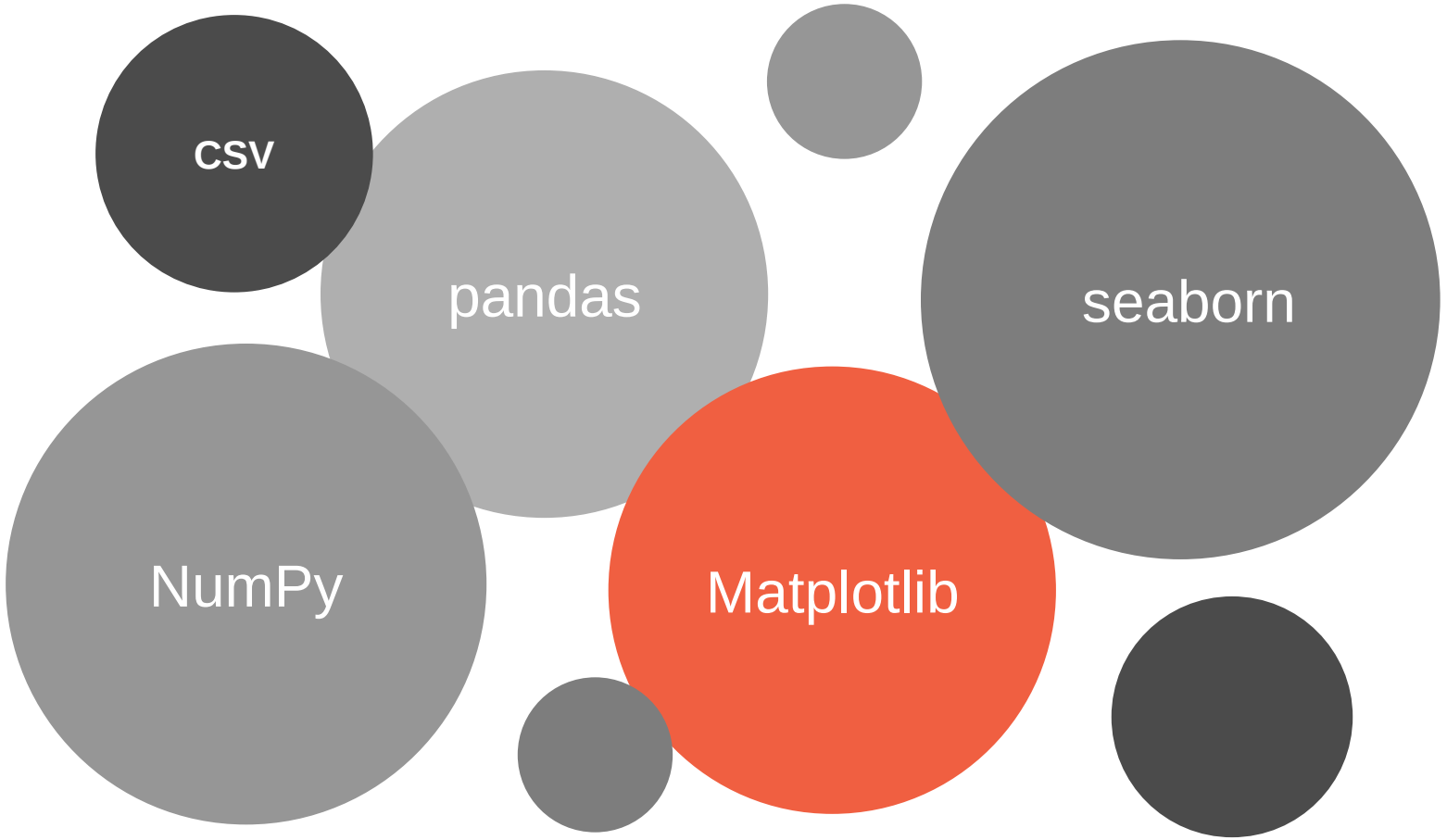
Q7 Average cost per movies ?.



Python libraries

Libraries

we'll use the Python libraries NumPy, pandas, seaborn and Matplotlib, which make writing data analysis code in Python a lot easier!



Data Wrangling:

The second step of the process is where we get the data included in the project resources, we will load in the data, check for cleanliness, and then trim and clean our dataset for analysis.

1- Read the data file tmdb-movies.csv.
2- Show all columns.
3- Show all datatypes.

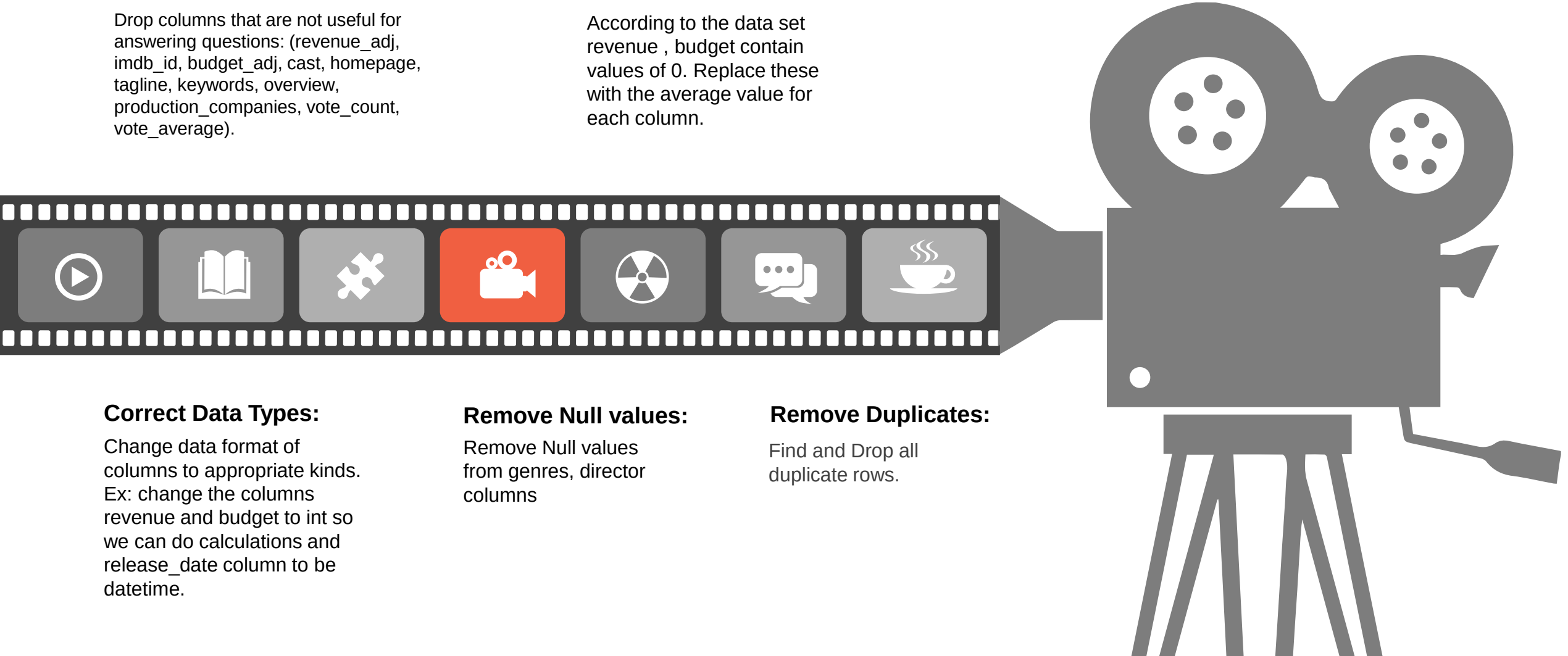
Data Cleaning

Remove Unnecessary Columns:

Drop columns that are not useful for answering questions: (revenue_adj, imdb_id, budget_adj, cast, homepage, tagline, keywords, overview, production_companies, vote_count, vote_average).

Dealing with 0 Values:

According to the data set revenue , budget contain values of 0. Replace these with the average value for each column.



Correct Data Types:

Change data format of columns to appropriate kinds. Ex: change the columns revenue and budget to int so we can do calculations and release_date column to be datetime.

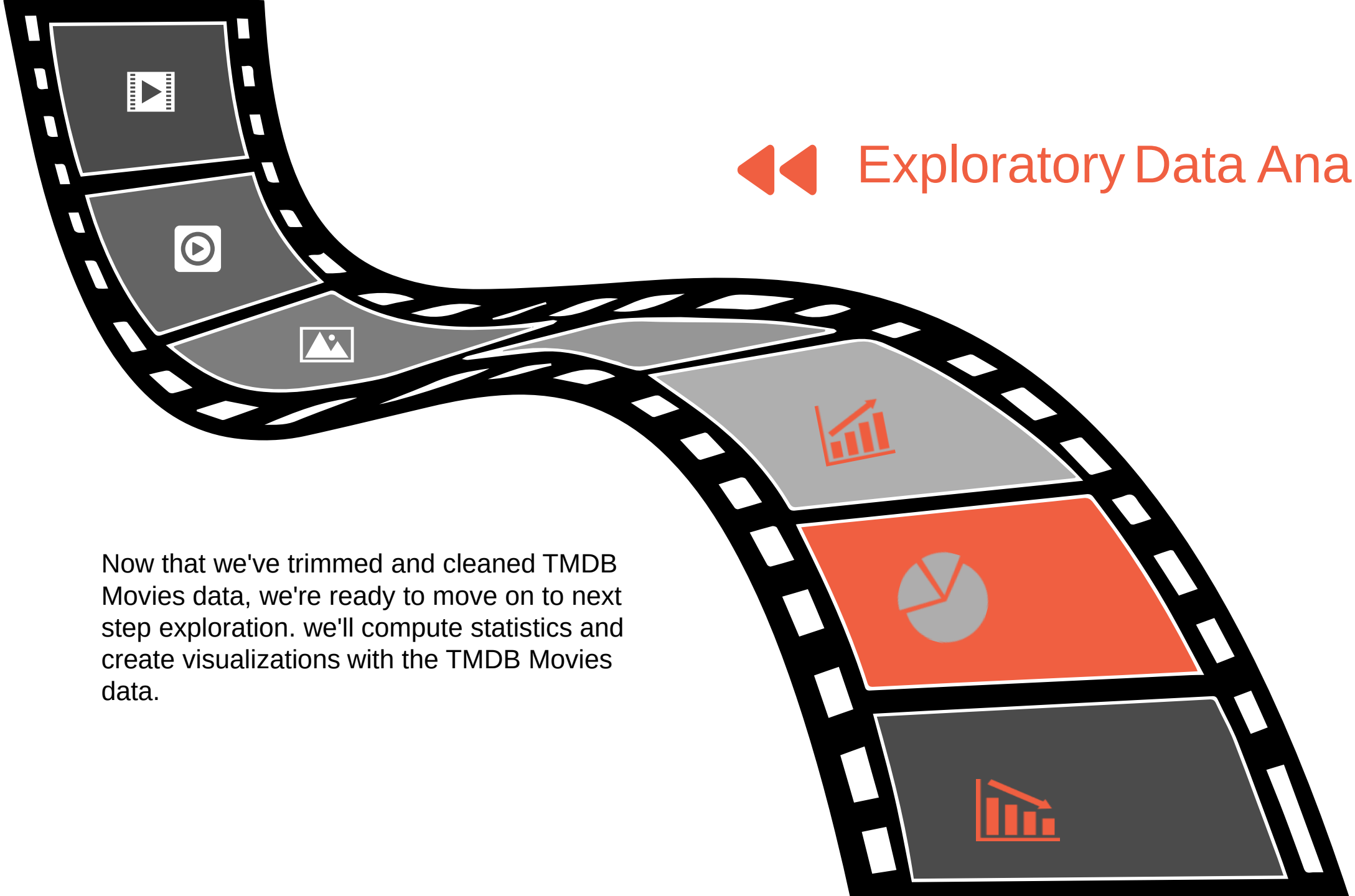
Remove Null values:

Remove Null values from genres, director columns

Remove Duplicates:

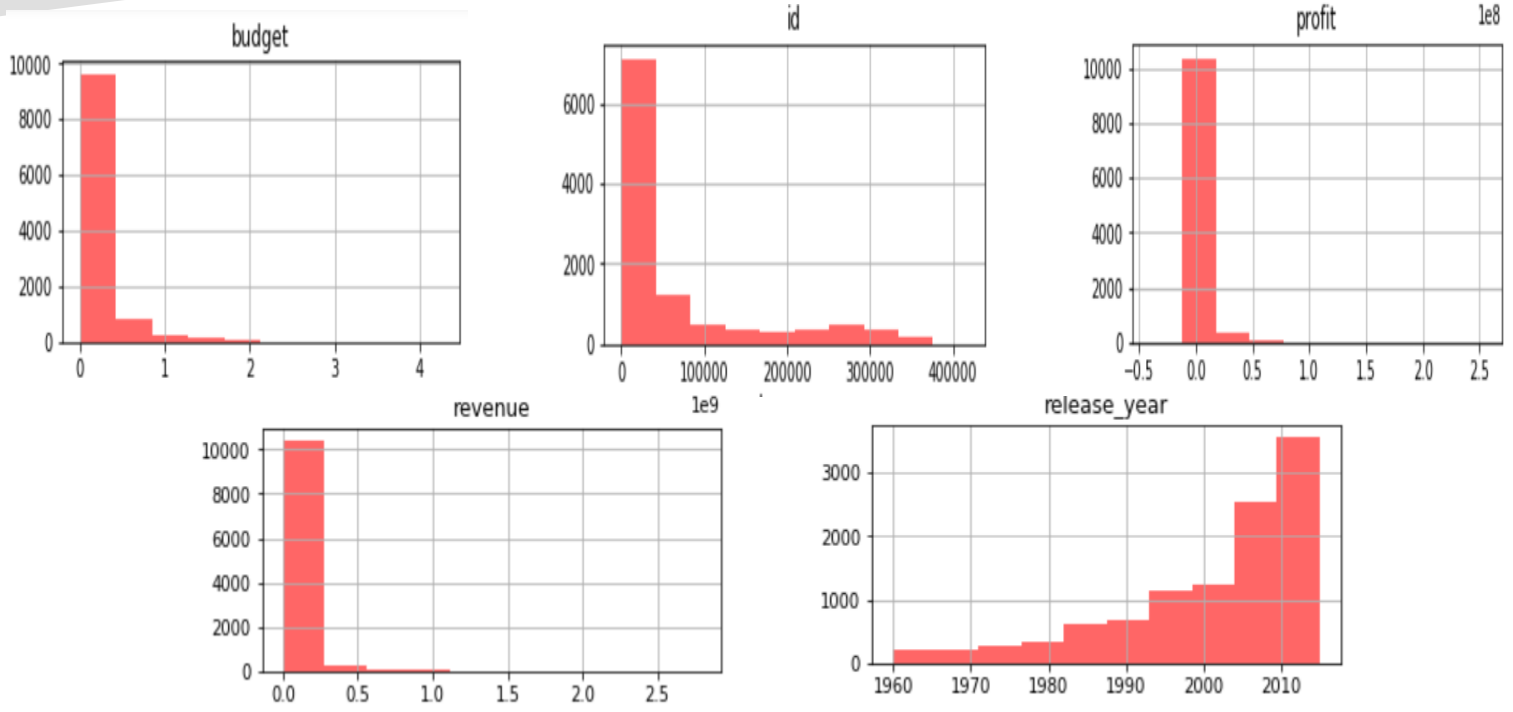
Find and Drop all duplicate rows.

Exploratory Data Analysis



Now that we've trimmed and cleaned TMDB Movies data, we're ready to move on to next step exploration. we'll compute statistics and create visualizations with the TMDB Movies data.

Explore Numerical Data With Histogram



Exploratory Data

What movies genres are most common in general?

Movies and Shows have more than one genre, so we need to split the genres column by the "|" character to the new column has one genre.

Cope Movies Data

create cope from original data set.

Split column

Deal with the column that have multiple values per cell, split column by the "|" character.

Single values

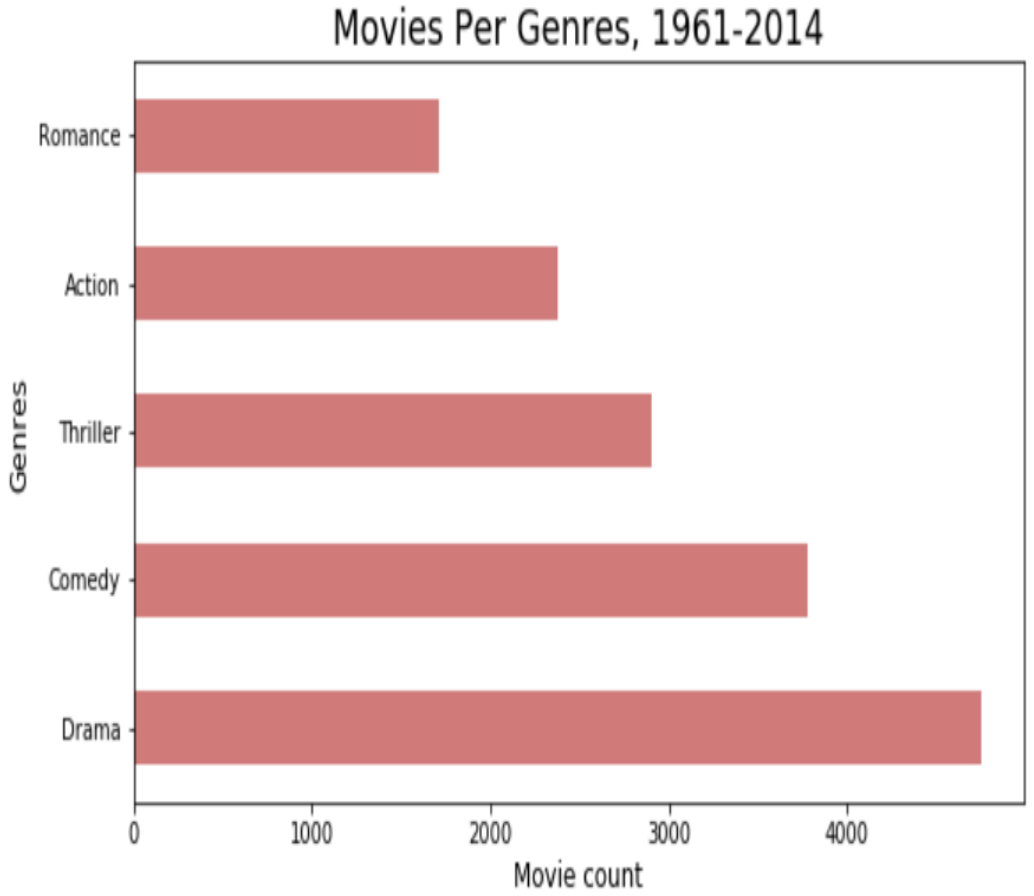
Check the column contains only single values.

Movies Per Genres

Used value_counts() for genres column and sort it by highest value to take top 5 genres.

Visualization

Use Bar chart to visualize top 5 genres.



This figure depict the number movies per genres from 1961 to 2014 and the Drama is most popular genre made 4754, followed by comedy and thriller made 3782 and 2904 respectively

Exploratory Data

What are the highest 5 movies in profit ?

We add a profit column: Profit = revenue (income) - budget (cost or expense)

Add new column

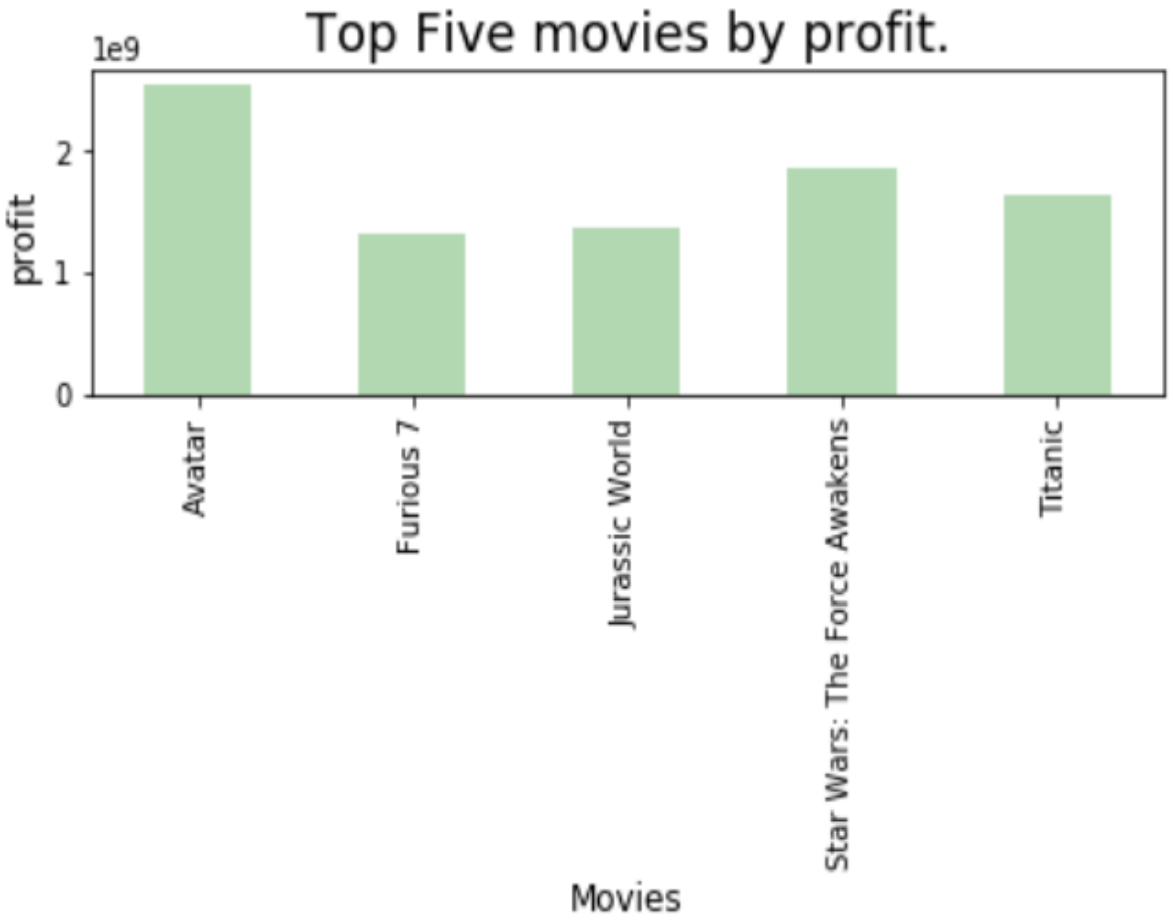
Profit = revenue (income) - budget (cost or expense)

Movies Per Profit

Used sum() for profit column and sort it by highest profit to take top 5 Movies.

Visualization

Use Bar chart to visualize top 5 Movies.



According to what the bar chart shows, top one movie that title Avatar had a largest profit reaching 2.5 billion.

Exploratory Data

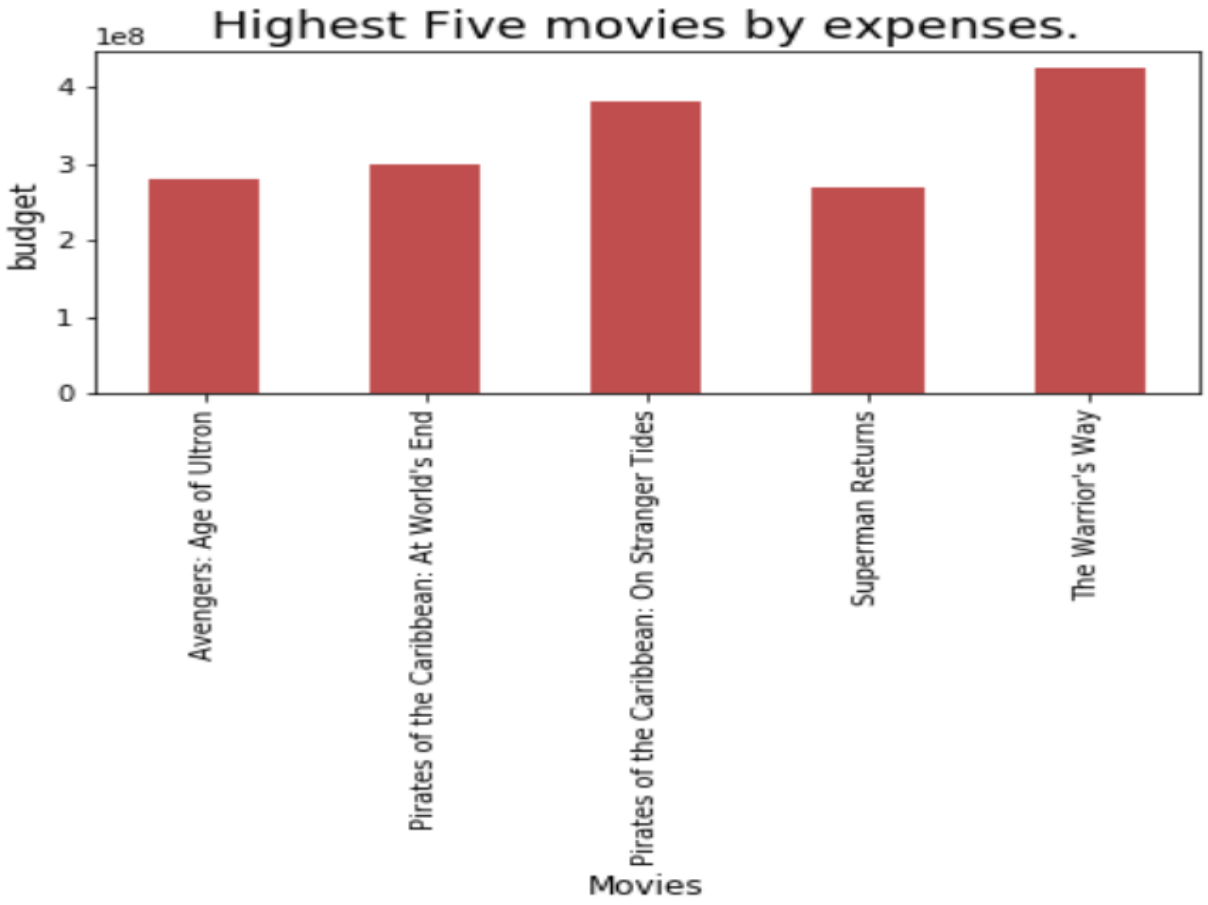
What are the highest 5 movies in expenses ?

Movies Per Expanses

Used sum() for budget column and sort it by highest expenses to take top 5 Movies.

Visualization

Use Bar chart to visualize top 5 Movies.



According to what the bar chart shows, top one movie that title 'The Warrior's Way' has a highest expenses reaching 425 million, while the 'On Stranger Tides' film was the second highest expense reaching 38 million

Exploratory Data

What are the top 5 director earned highest profit ?

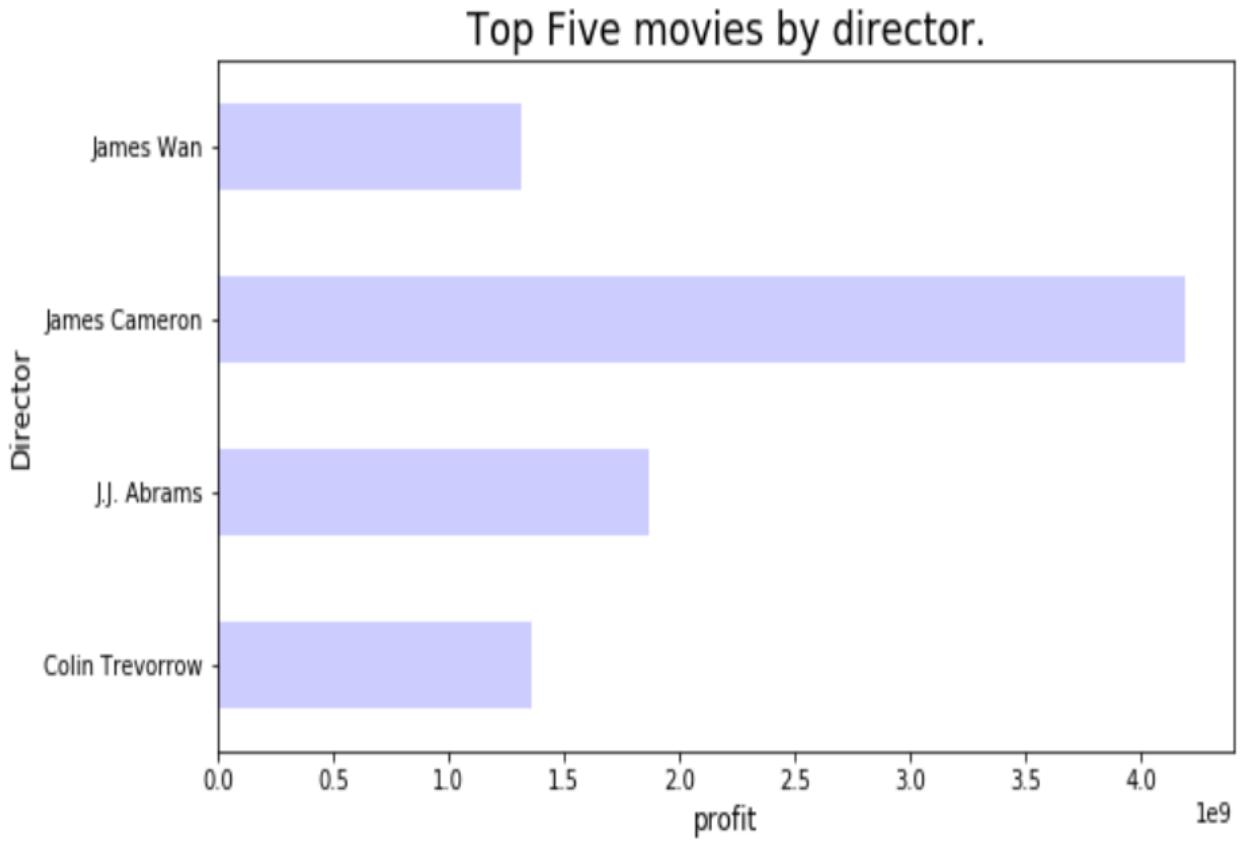
We add a profit column: Profit = revenue (income) - budget (cost or expense)

Profit Per Director

Used sum() for profit column and group by director and sort it by highest profit to take top 5 profit.

Visualization

Use Bar chart to visualize top 5 Profit.



According to what the bar chart shows, the first director 'James Cameron' has gain the significant profits.

Exploratory Data

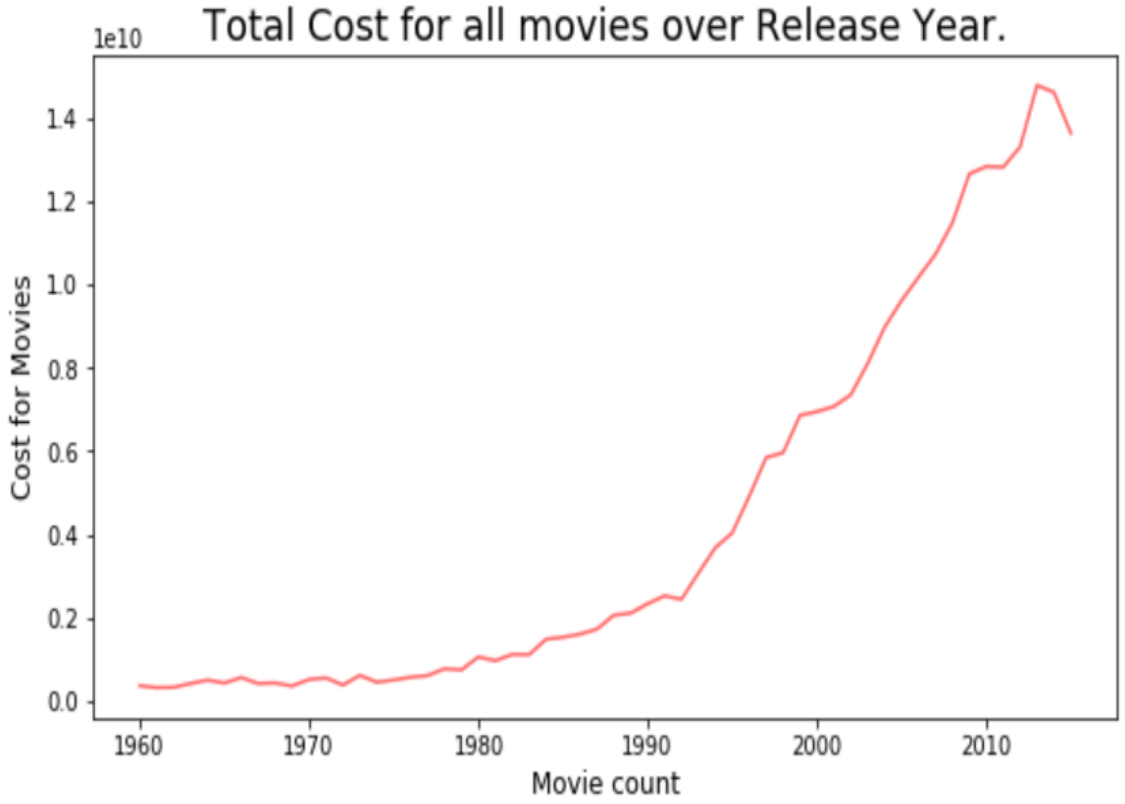
Which year that has the highest cost?

Cost Per Year

Used sum() for budget column and sort it by highest cost to take top years.

Visualization

Use Line chart to visualize cost per years.



It is clear from graph the ratio increased throughout the period, the cost of movies rocket significantly in 2010.

Exploratory Data

40M

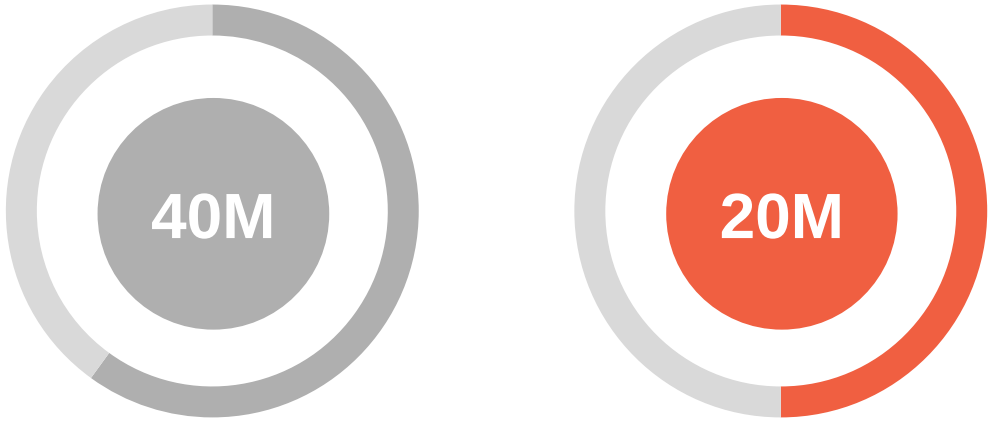
Average Profit Earned:

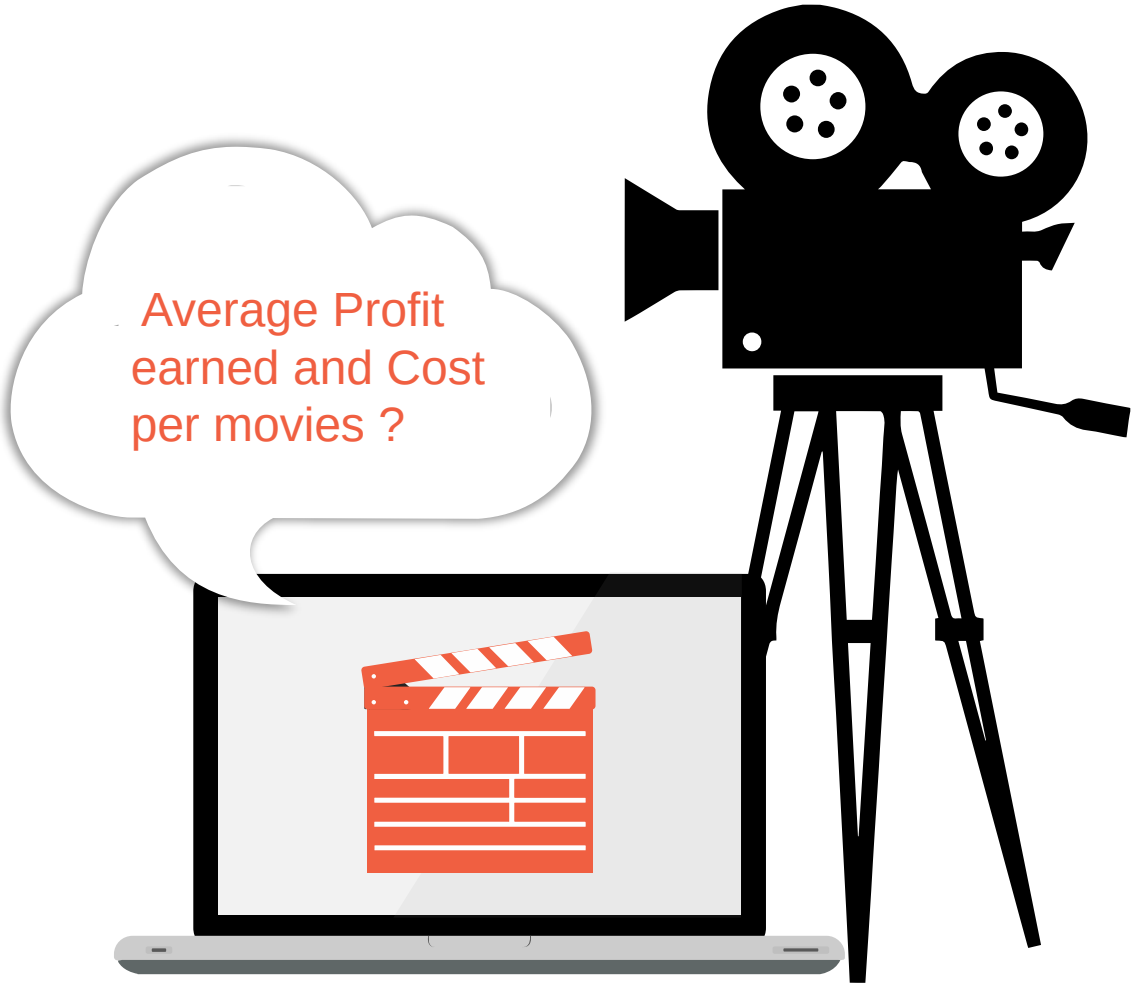
calculate the mean value for “profit" column by mean.()
So the movies have the Average profit around 40 million dollar.

20M

Average Cost:

calculate the mean value for “budget" column by mean.()
So the movies have the Average cost around 20 million dollar.





Conclusions & Limitations

First research question:

Shows the number of movies per genres from 1961 to 2014 and the Drama is most popular genre made 4754, followed by comedy and thriller made 3782 and 2904 respectively.

Second research question:

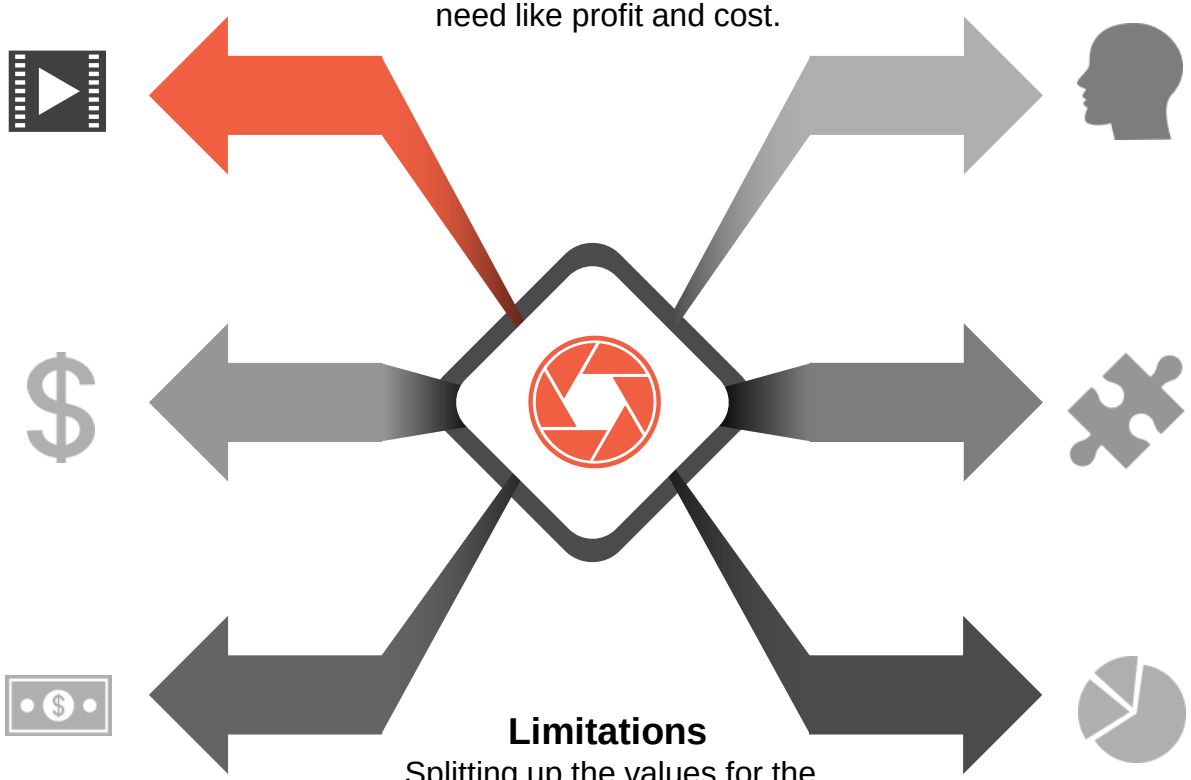
According to what the shows, top one movie that title Avatar had a largest profit reaching 2.5 billion.

Third research question:

According to what the shows, top one movie that title 'The Warrior's Way' has a highest expenses reaching 425 million, while the 'On Stranger Tides' film was the second highest expense reaching 38 million.

Conclusions

throughout this data analysis, I tray to show some business usual need like profit and cost.



Limitations

Splitting up the values for the genre columns really slowed down the processing so I Went back and created 2 copies of the data frame to perform those operations separately

Forth research question:

According to what the shows, the first director 'James Cameron' has gain the significant profits, James Cameron the director of Avatar movie as we see before.

Fifth research question:

It is clear from our analysis the ratio of cost increased throughout the period, the cost of movies rocket significantly in 2010.

Last 2 research questions:

The Average profit around 40 million dollar per movies, while average cost around 20 million dollar per movies.