Dataset

TMDb movie data

Problem statement

- What is the distribution of vote score average?
- What is the distribution of popularity?
- How many percent of movies have homepage?
- What is the ranking of revenue generating ability of actor/actress?
- Which genre is most popular over the years?

Investigation description

- Plot vote score average in histogram
- Plot popularity in histogram
- Count the movies with and without homepage
- Calculate the total revenue of movie that casted by the particular actor/actress
- Calculate the number of movies that contain the genre over years.

Data wrangling

- Remove duplicated row
- Change the column ‘release_date’ datatype to datetime
- Split the element of column ‘cast’, ‘production_companies’, ‘genres’, ‘director’ into list

Summary

- Distribution of vote average

  The vote score average is fall in a normal distribution with the mean of approximately 5.8

- Distribution of popularity

  The graph shows most of the popularity is fell in between 0 to 2.5

- Percentage of movies has home page

  The chart shows there are only 27% of movies have homepage

- Investigate revenue generating ability of actor/actress

  In summary, the most earning actor is Harrison Ford over the years which is more than 9 billions then followed by Tom Cruise, Tom Hanks, Emma Watson, Ian McKellen, Johnny Depp, Daniel Radcliffe, Rupert Grint, Robert Downey Jr., Ralph Fiennes.

- Investigate the genre porpularity over year

  Drame genre was the most popular genre which was mostly in the top rank over the year. However, Drama genre had the same popularity with comedy genre on year 1963 and 1966, and it was being passed by comedy genre on 1967, 1985, 1987, 1988, 1989, 1994, 2001 and 2003.

Limitation

The limitation of this project is there are some unrecorded data which would affect the result of the investigation. For example there are some of the records from the dataset do not have the cast. However this investigation could be an assumption with the current available data.