Project: Investigate a (TMDB 5000 Movie Dataset)

Introduction

this data set contains information about more than 10000 movies collected from the movie database , including a lot of information about these movies so we are going to discover the columns and values in this data set

we have 21 columns with the following titles:

['id', 'imdb_id', 'popularity', 'budget', 'revenue', 'original_title', 'cast', 'homepage', 'director', 'tagline', 'keywords', 'overview', 'runtime', 'genres', 'production_companies', 'release_date', 'vote_count', 'vote_average', 'release_year', 'budget_adj', 'revenue_adj']

Questions need to be answered:

After investigating the dataset we have multiple questions we will try to answer in our data analysis process :

1- what are the 10 longest movies in this dataset

2- what are the top 10 years in releasing movies

3- who are the top 10 directors by number of directed movies

Data Wrangling :

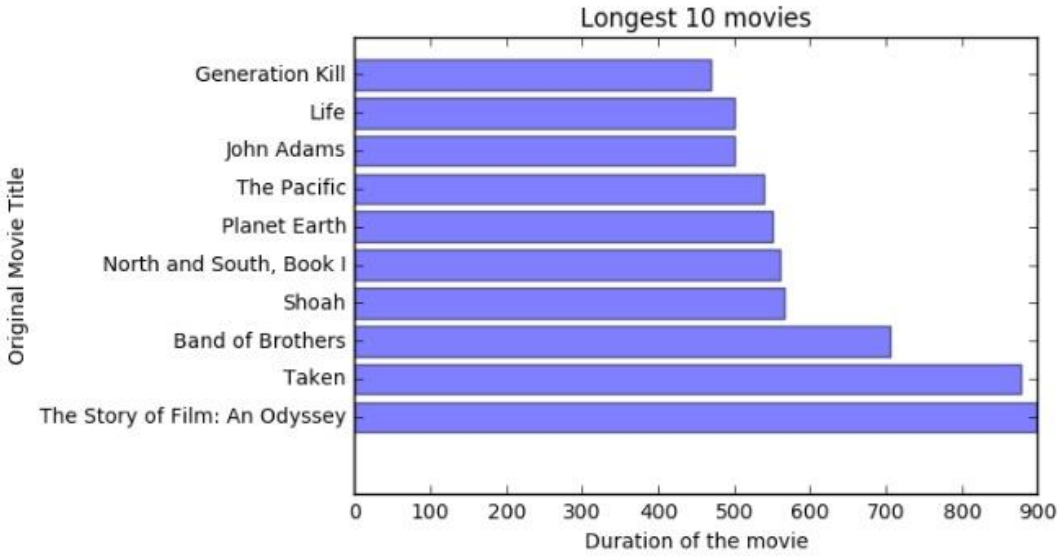
we can notice the year value in release_date column is shortened to be 2 numbers only for example : 11 instead of 2011 so we are going to change this value by removing the last 2 numbers in release_date column then we will add the correct year from release year column

Summary Statistics :

1- The 10 longest movies in this dataset are:

[The Story of Film: An Odyssey, Taken , Band of Brothers ,Shoah, (North and South, Book I) ,Planet Earth ,The Pacific ,John Adams , Life , Generation Kill]

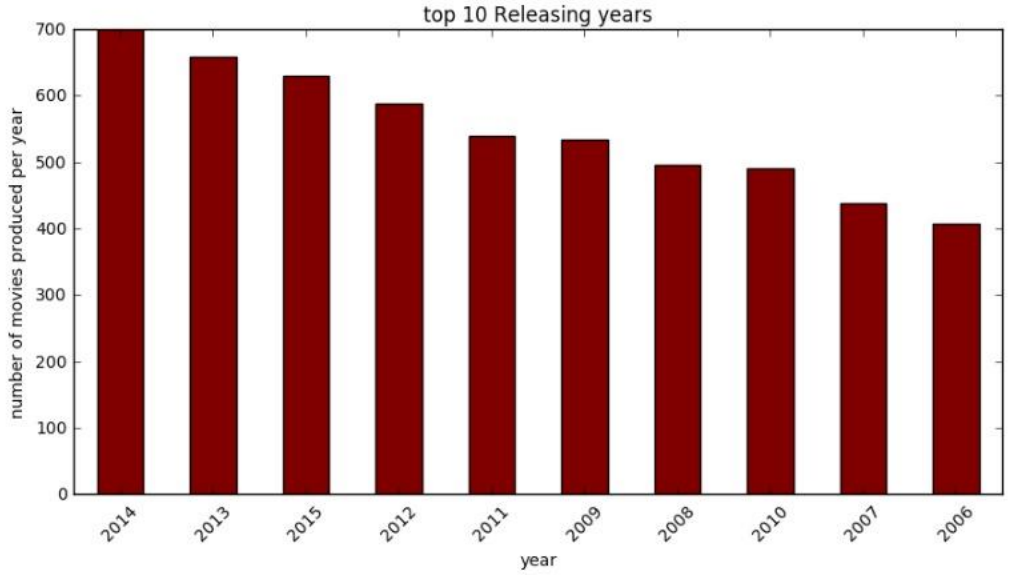
Bar Chart:



2- The top 10 years in releasing movies are:

`[2014,2013,2015,2012,2011,2009,2008,2010,2007,2006]`

Bar Chart:



3- the top 10 directors by number of directed movies:

[Woody Allen, Clint Eastwood, Martin Scorsese, Steven Spielberg, Ridley Scott, Steven Soderbergh, Ron Howard, Joel Schumacher, Brian de Palma, Barry Levinson]

Bar Chart: