Project 2
Investigating a dataset
TMBD movies

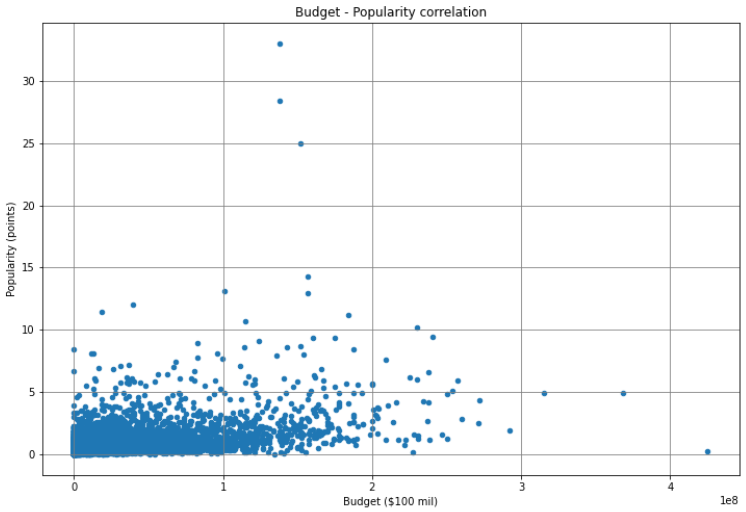
- TMBD movies dataset was selected for analysis. It was uploaded in a csv format to Jupyter Notebook on my local machine.
- Questions posed:

  1) Is budget associated with popularity? In other words, had movies with higher budget received a higher popularity score?

  2) Had the release year affected the revenue? Had older or newer movies generated higher revenues?

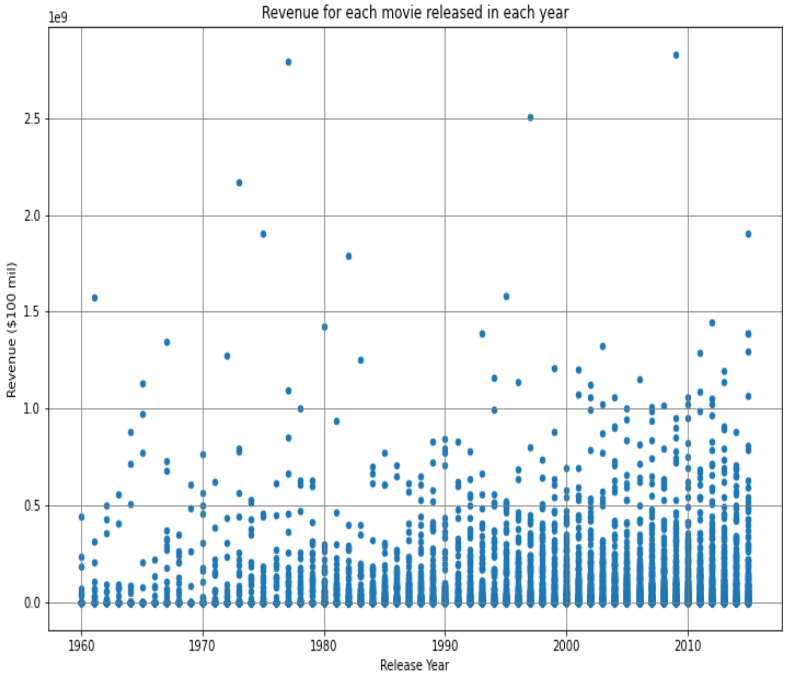
  3) Had runtime affected revenue? Had shorter or longer movies produced higher revenues?
- To answer the posed questions, I have plotted graphs (scatter and bar types). Scatter plots were selected since each dot represented a single row (movie), thus a better overview could be achieved. Bar chart, on the other hand, showed misleading conclusions since it mixed outliers with general trends. Only scatter plots were used to extract results. In some cases, outliers were disregarded to dig into general trends for the majority of the dataset. In the meantime, those disregarded outliers were also examined afterwards.

  On top of that, .info() and describe() functions were used to see overall statistics of the parameters.
- Before investigating the dataset, some data wrangling was performed. Numpy, pandas and matplotlib were imported. Then the data was read from csv into a dataframe. Then general information was checked with .info(), describe(), head() functions. Afterwards, some unnecessary columns were dropped, as well as 1 duplicate. After these manipulations, only one parameter still has some missing values. The rows with missing values were checked and dropped since no significant data was determined.
- - Q1 plot:



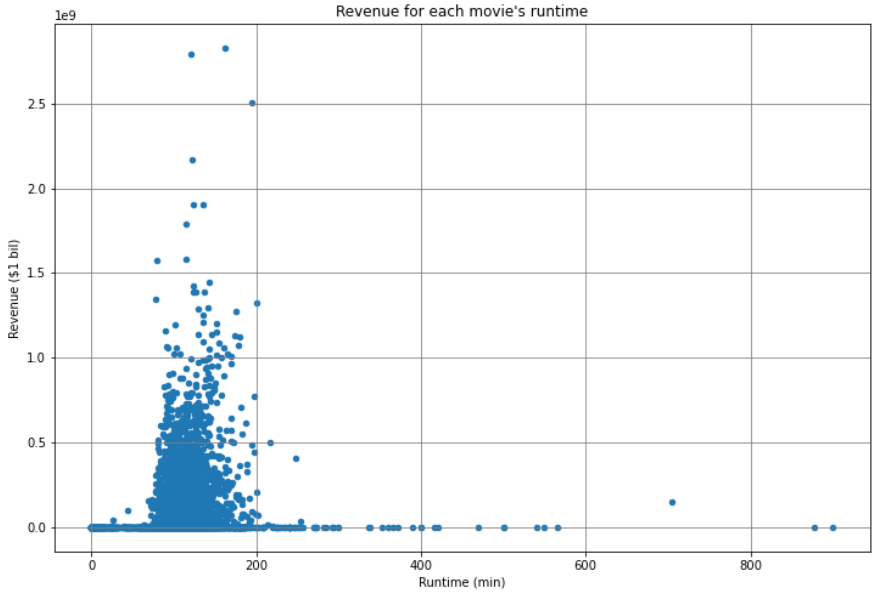
There is no direct correlation between a movie's budget and its success.

- Q2 plot:



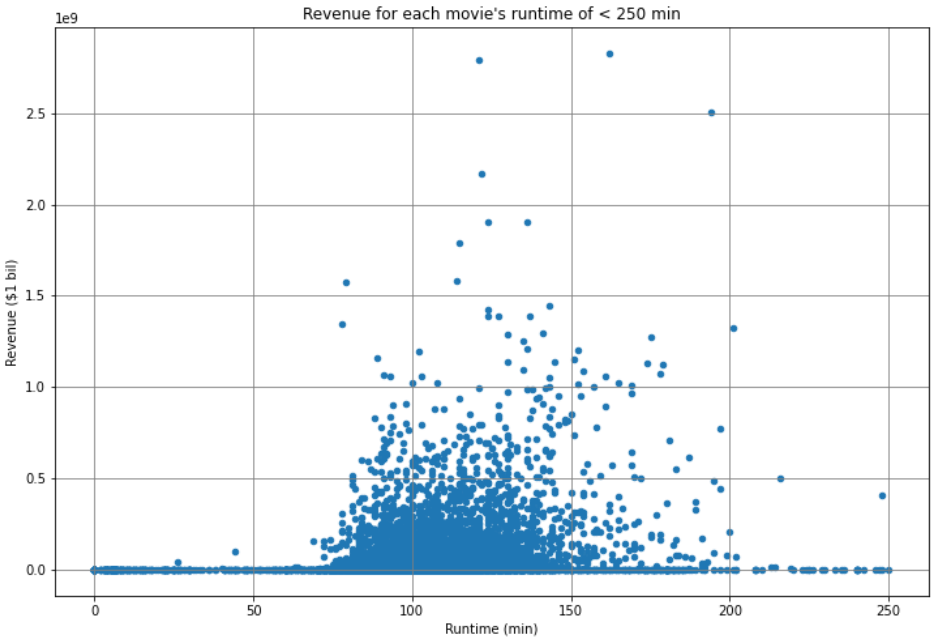
There is an overall positive correlation between a release year and movie's revenue with some outliers. Newer movies generated more revenue.

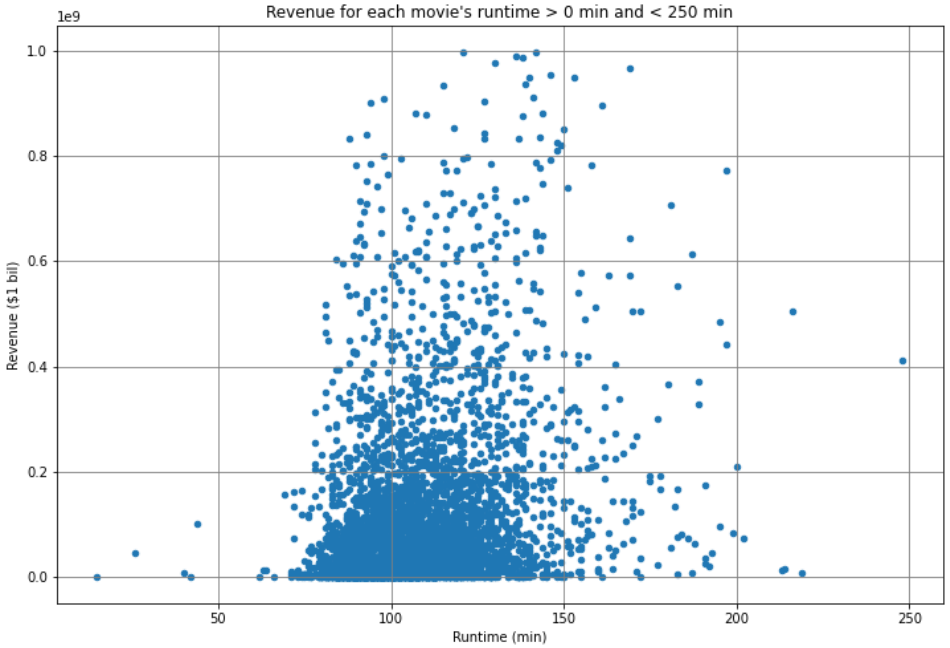
- Q3 plot:



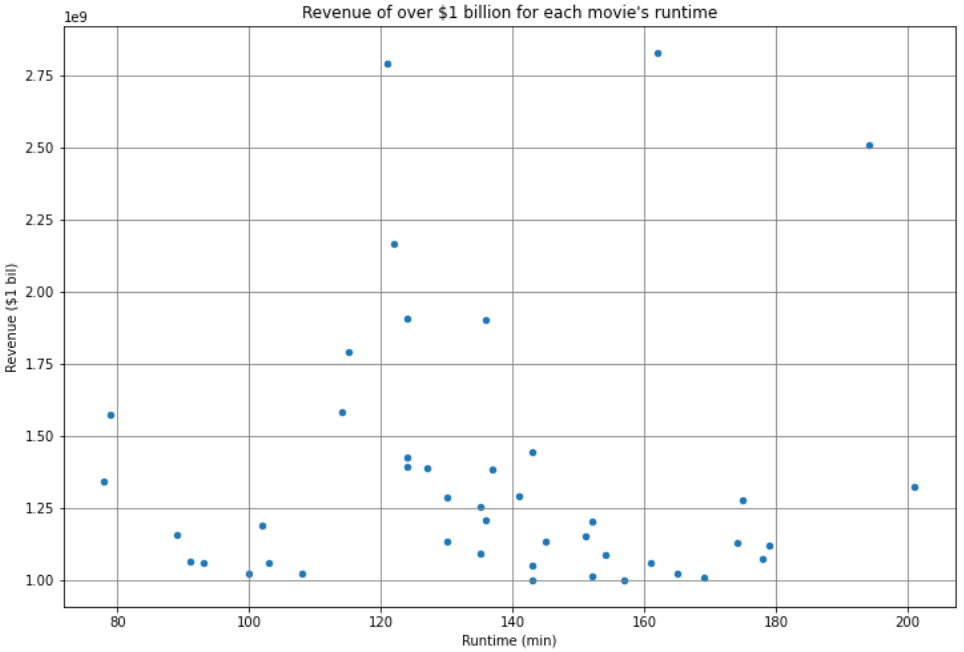
There is a general trend with normal distribution for movies with a non-zero revenue and revenue below 1 billion. Some outliers also exist.

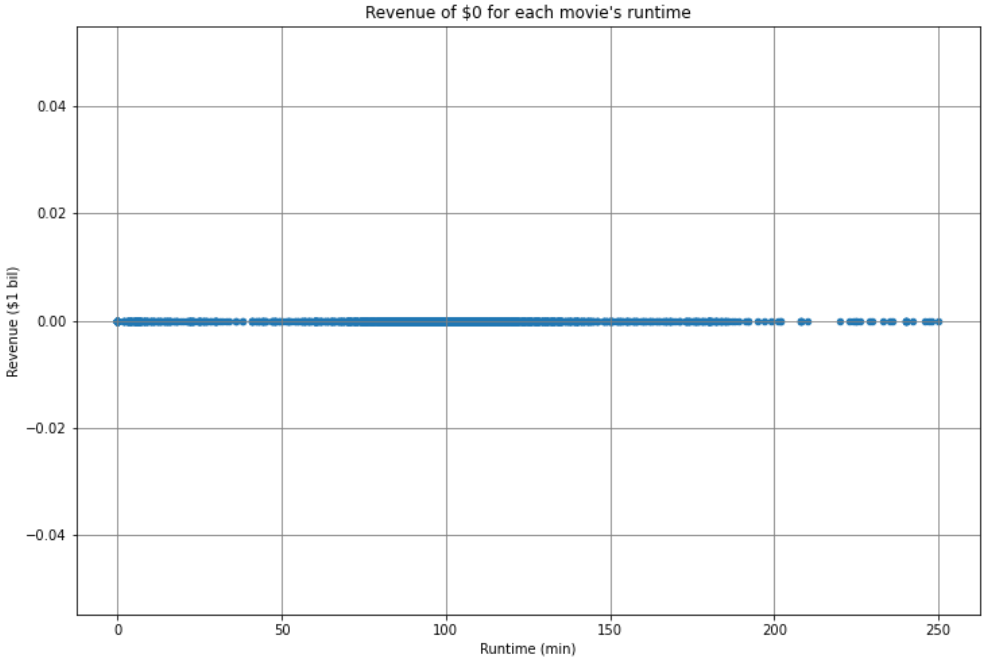
- Plots for general trends:





- Plots for outliers:





More detailed comments were left in the Jupyter Notebook project file.