Abduljalil Abiri – Udacity TMDB Movies dataset analysis

1) Dataset Analyzed: TMDB Movies Dataset

2) Questions posed
   a) The most profitable movie genre year by year
   b) Top 10 Most profitable movies and least profitable movies
   c) Properties associated with high revenue movies
   d) Top 20 Highest-grossing Actors on average

3) How each question was answered:
   a) The most profitable movie genre year by year:
      - Split the genre column and made each split genre a separate row
      - Grouped the dataset by year and found the highest genre by popularity for that year
      - Used a scatterplot to visualize the popularity of genres across several years (1-1960-2015)

   b) Top 10 Most profitable movies and least profitable movies
      - Created two datasets called dfTop10 and dfBottom10 respectively from the original movie dataset of just two columns (“original_title”,”net_profit”).
      - Sorted dfTop10 by net_profit in descending order and limited it to 10 rows. This gave the Top 10 most profitable movies
      - Sorted dfBottom10 by net_profit in ascending order and limited it to 10 rows. This gave the Top 10 least profitable movies

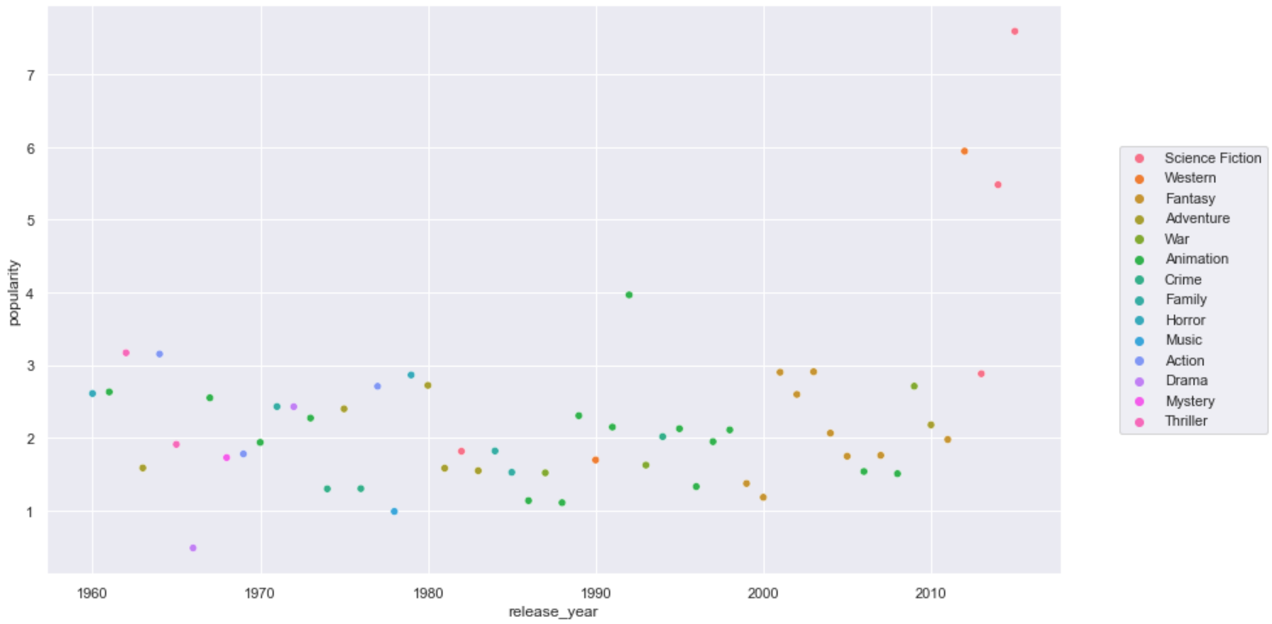
   c) Properties associated with high revenue movies
      - Calculated the mean revenue of the entire dataset. Then filtered the dataset to movies whose revenue that are higher than the revenue mean. Called this new dataset – dfHighProfit.
      - Found the columns of dfHighProfit which had the strongest correlation to revenue.
      - Made a scatterplot to visualize each columns’ correlation to revenue.

   d) Top 20 Highest-grossing Actors on average
      - Split the cast column and made each split cast a separate row. This is done so that each row has a single actor.
      - Grouped the dataset by “cast” column and calculated the average revenue generated per actor. Saved the result as a dataset called df1cast.
      - Grouped the dataset by “cast” column and calculated the number of movies per actor. Saved the result as a dataset called df1castsum.
      - Merged df1castsum into df1cast and limited the dataset to actors with at least 10 movies.
      - Sorted df1cast dataset by average revenue in descending order and limited the dataset to the first 20 rows.
      - Generated a horizontal bar plot to visualize the highest average revenue per movie for the actors in the top 20 rows.

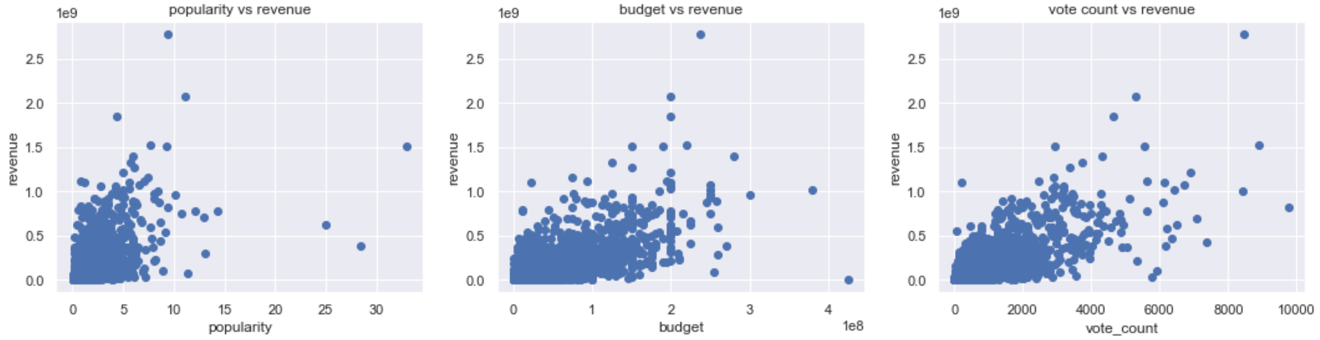
4) Data Wrangling Documentation
   a) Removed columns not needed in the dataset
   b) Removed all duplicate rows
   c) Removed rows whose revenue or budget were 0 or NaN.
   d) Changed the date column to datetime for easy time series analysis
   e) Created a new column “net_profit” that shows the profit made on each movie

5) Summary plots

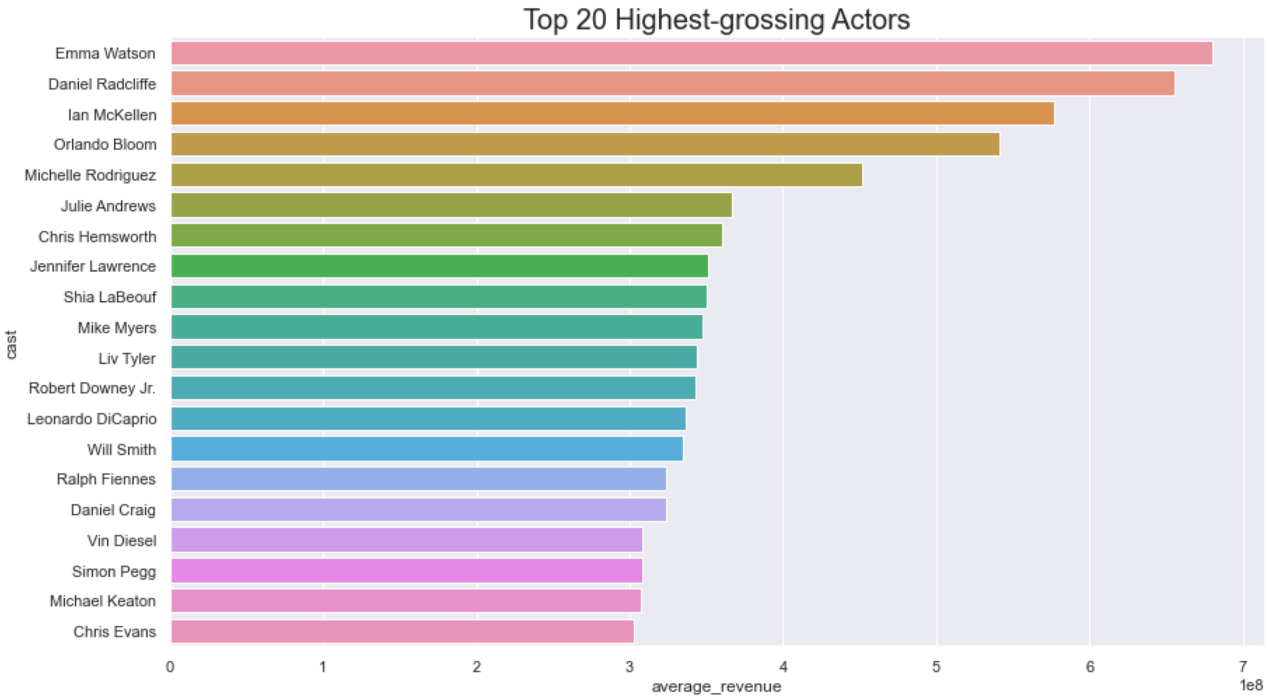
   a) Visualization of the most popular genre from 1960 to 2015



   b) Scatterplots of each chosen columns’ correlation to revenue.



c) A horizontal bar plot of the top 20 Highest-grossing Actors


Top 20 Highest-grossing Actors
cast
Emma Watson
Daniel Radcliffe
Ian McKellen
Orlando Bloom
Michelle Rodriguez
Julie Andrews
Chris Hemsworth
Jennifer Lawrence
Shia LaBeouf
Mike Myers
Liv Tyler
Robert Downey Jr.
Leonardo DiCaprio
Will Smith
Ralph Fiennes
Daniel Craig
Vin Diesel
Simon Pegg
Michael Keaton
Chris Evans
1e8
average_revenue