Data Appendix

This appendix contains all information about each data file and its variables. There were three data files used for the project including movie_reviews.csv, movie_data_initial.csv, and movie_data_final.csv.

Movie reviews data. This csv contains professional movie critic reviews for various movies between the 1940s-2020s. The dataset was found on Zenodo and downloaded to use the reviews for text analysis. There are over 17,000 reviews from over 2,500 movies in the dataset. Each row represents a review for a specific movie by an individual reviewer so the unit of observation is movie review.

- *Review_ID.* A unique identification number assigned to each review. This value is numeric.
- *Title.* The title of the movie that the review is referencing. This variable is categorical.
- *Year.* The year the review was published. The years range between 1913 and 2001. This variable is numeric.
- *Reviewer_Name.* The name of the reviewer who published each review. This variable is categorical.
- *Review_Text.* The full text and content of the review. This variable is categorical.
- *Rated.* The content rating of the movie being reviewed. This variable is categorical.
- *Year_API.* The year the movie was released. The years range between 1907 and 2023. This variable is numeric.
- *Genre.* The genres of the movie. Some movies contain multiple genres which are listed within this variable. This variable is categorical.
- *Directors.* The directors of the movie. If there are multiple directors, they are listed within this variable. This variable is categorical.
- *Writers*. The writers of the movie. If there are multiple writers, they are listed within this variable. This variable is categorical.
- *Actors.* The starring actors of the movie. This variable is categorical.
- *Plot.* The official plot of the movie. This variable is categorical.
- *First_Genre.* The first genre listed in the previous *Genre* variable. This variable is categorical.
- *First_Actor*. The first actor listed in the previous *Actors* variable. This variable is categorical.
- *First_Director*. The first director listed in the previous *Directors* variable. This variable is categorical.
- *First_Writer*. The first writer listed in the previous *Writers* variable. This variable is categorical.

- *First_Actor_Gender*. The gender of the actor listed in the previous *First_Actor* variable. This variable is categorical.
- *First_Director_Gender*. The gender of the director listed in the previous *First_Director* variable. This variable is categorical.
- *First_Writer_Gender*. The gender of the writer listed in the previous *First_Writer* variable. This variable is categorical.

Initial movie data. This dataset was created by merging the movie review data from Zenodo with web scraped data from Box Office Mojo. This was done by using SelectorGadget to scrape the data from Box Office Mojo and was read into R using the rvest package. After creating a dataframe with the variables *Title*, *Lifetime_Gross*, and *Release_Year* it was combined with the movie reviews data on the variable *Title* to create a comprehensive dataset. Each row represents a review for a specific movie by an individual reviewer so the unit of observation is movie review.

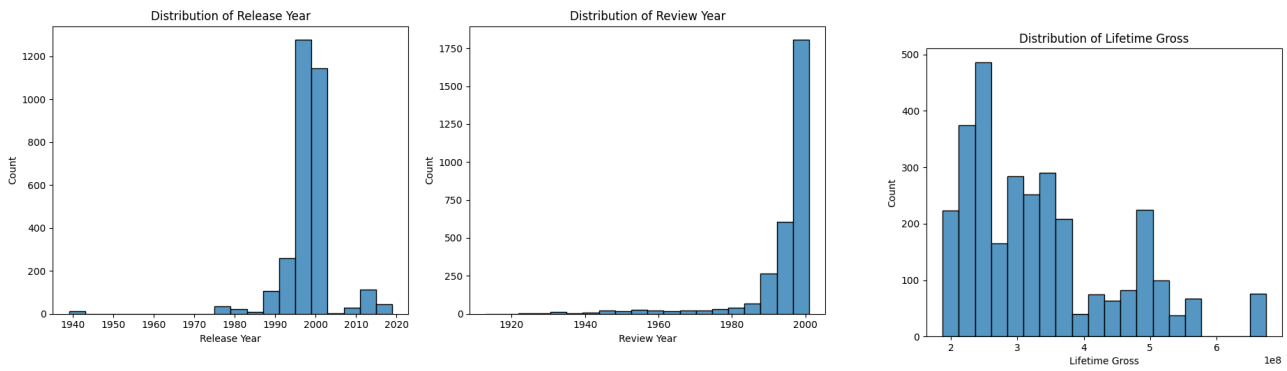
- *Review_ID.* A unique identification number assigned to each review. This value is numeric.
- *Title.* The title of the movie that the review is referencing. This variable is categorical.
- *Review_Year.* The year the review was published. The years range between 1913 and 2001. This variable is numeric.
- *Reviewer_Name.* The name of the reviewer who published each review. This variable is categorical.
- *Review_Text.* The full text and content of the review. This variable is categorical.
- *Rated.* The content rating of the movie being reviewed. This variable is categorical.
- *Year_API.* The year the movie was released. The years range between 1907 and 2023. This variable is numeric.
- *Genre.* The genres of the movie. Some movies contain multiple genres which are listed within this variable. This variable is categorical.
- *Directors.* The directors of the movie. If there are multiple directors, they are listed within this variable. This variable is categorical.
- *Writers*. The writers of the movie. If there are multiple writers, they are listed within this variable. This variable is categorical.
- *Actors.* The starring actors of the movie. This variable is categorical.
- *Plot.* The official plot of the movie. This variable is categorical.
- *First_Genre.* The first genre listed in the previous *Genre* variable. This variable is categorical.
- *First_Actor*. The first actor listed in the previous *Actors* variable. This variable is categorical.
- *First_Director*. The first director listed in the previous *Directors* variable. This variable is categorical.

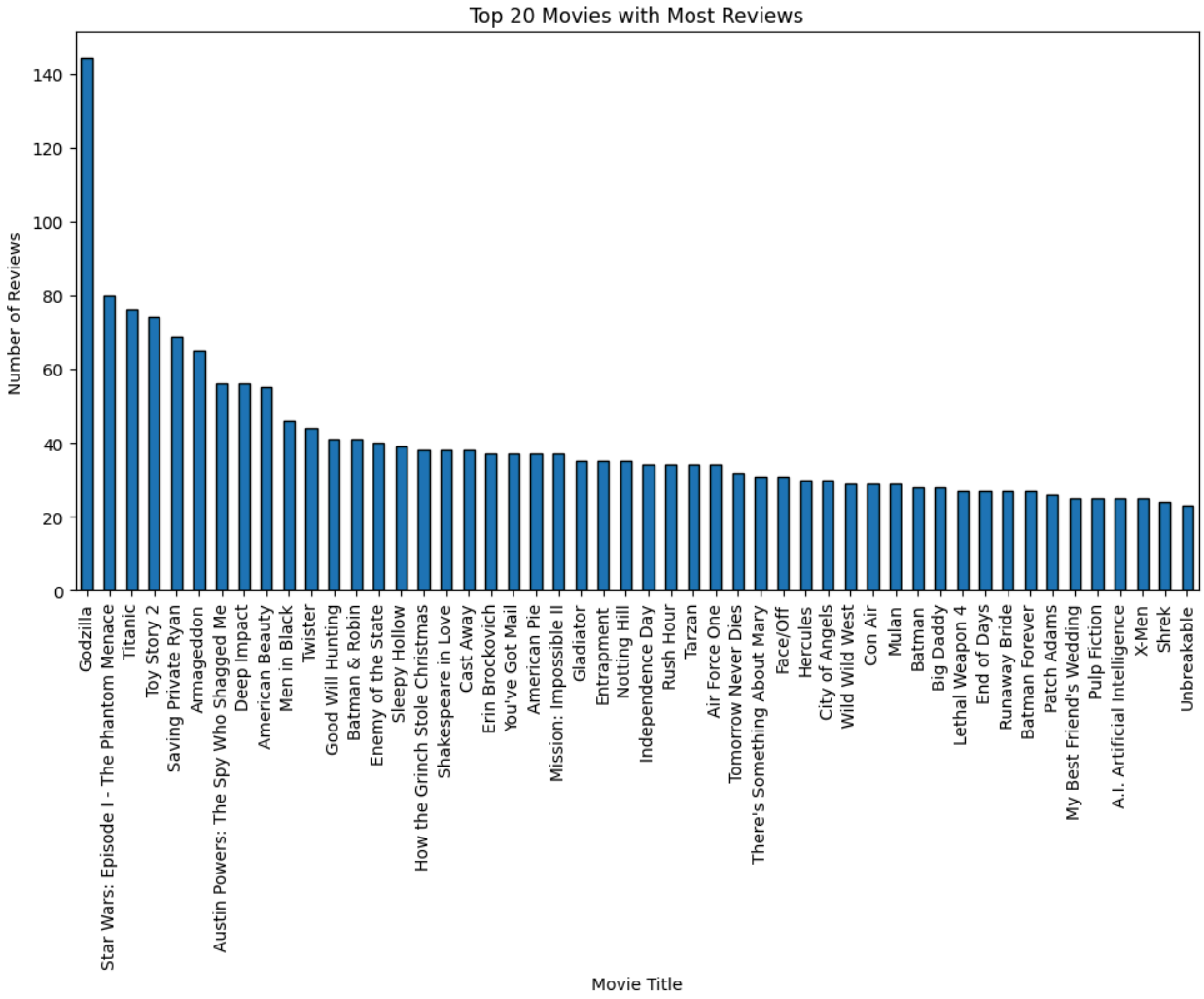
- *First_Writer*. The first writer listed in the previous *Writers* variable. This variable is categorical.
- *First_Actor_Gender*. The gender of the actor listed in the previous *First_Actor* variable. This variable is categorical.
- *First_Director_Gender*. The gender of the director listed in the previous *First_Director* variable. This variable is categorical.
- *First_Writer_Gender*. The gender of the writer listed in the previous *First_Writer* variable. This variable is categorical.
- *Lifetime_Gross.* The total revenue a movie has made in the box office since its release. This variable is categorical.
- *Release_Year.* The year the movie was released. This variable is categorical.

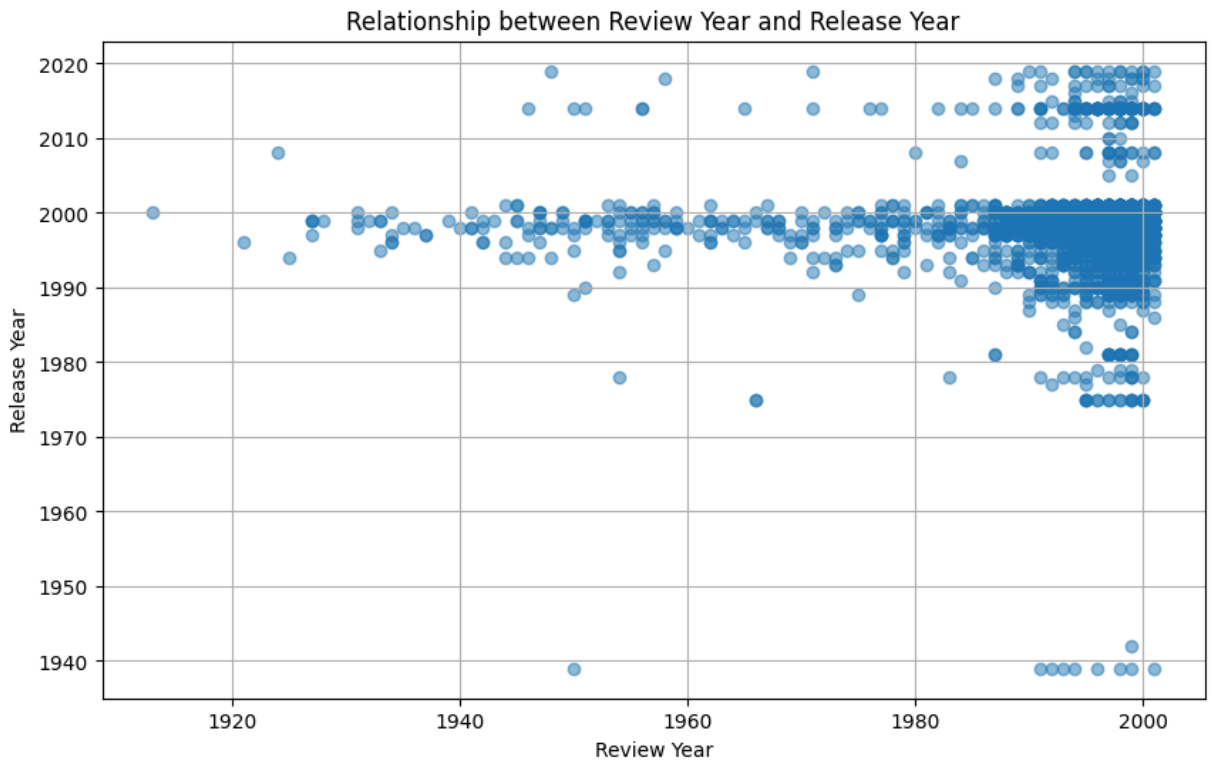
Final movie data. This dataset was created by cleaning the initial movie data. Most of the variables were removed, leaving only variables that were applicable to analysis. Some of the variables were also repeated between the datasets so duplicates were removed as well. The *Lifetime_Gross* and *Review_Year* variables were transformed into numeric values to make them more suitable for analysis. Each row represents a review for a specific movie by an individual reviewer so the unit of observation is movie review.

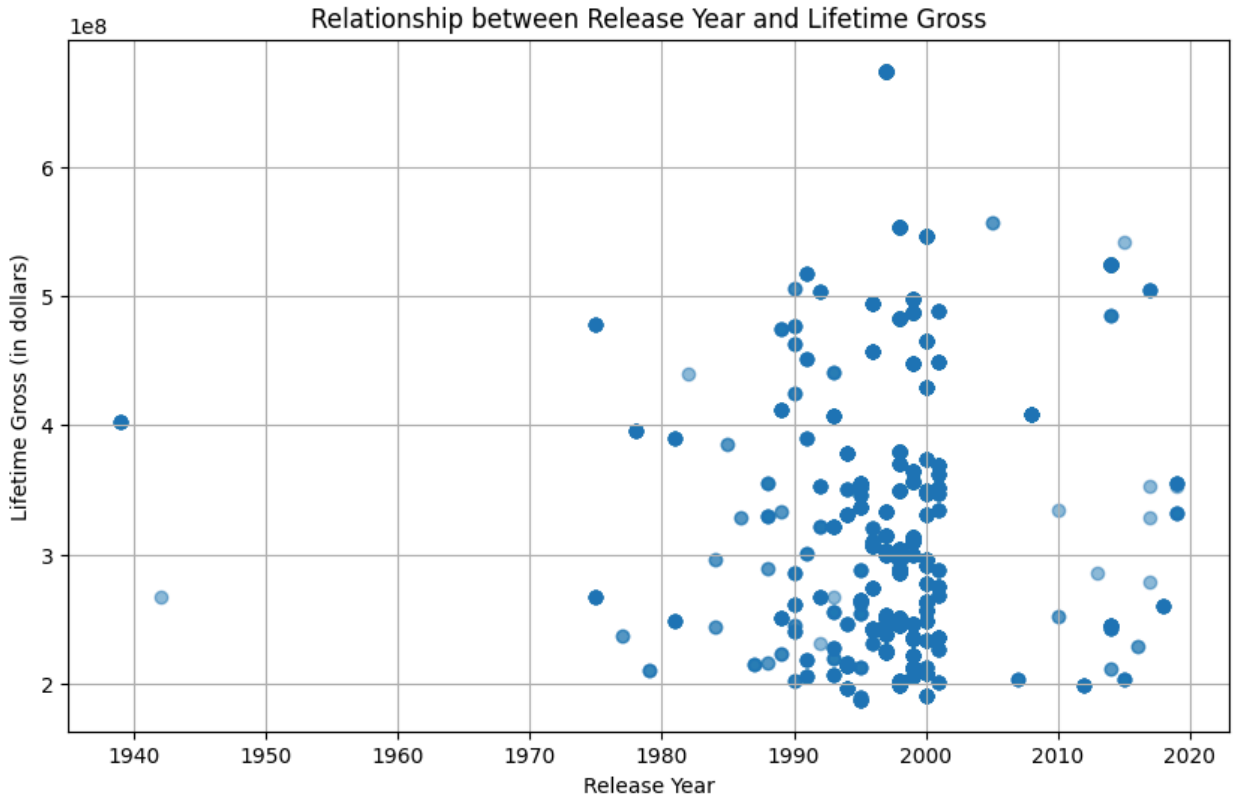
- *Title.* The title of the movie that the review is referencing. This variable is categorical.
- *Review_ID.* A unique identification number assigned to each review. This value is numeric.
- *Review_Text.* The full text and content of the review. This variable is categorical.
- *Lifetime_Gross.* The total revenue a movie has made in the box office since its release. The values range from 187,436,818 and 674,354,882. This variable is numeric.
- *Release_Year.* The year the movie was released. The years range from 1939 to 2019. This variable is numeric.

Movie data visualizations.




Top 20 Movies with Most Reviews
Number of Reviews
0
20
40
60
80
100
120
140
Godzilla
Star Wars: Episode I - The Phantom Menace
Titanic
Toy Story 2
Saving Private Ryan
Armageddon
Austin Powers: The Spy Who Shagged Me
Deep Impact
American Beauty
Men in Black
Twister
Good Will Hunting
Batman & Robin
Enemy of the State
Sleepy Hollow
How the Grinch Stole Christmas
Shakespeare in Love
Cast Away
Erin Brockovich
You've Got Mail
American Pie
Mission: Impossible II
Gladiator
Entrapment
Notting Hill
Independence Day
Rush Hour
Tarzan
Air Force One
Tomorrow Never Dies
There's Something About Mary
Face/Off
Hercules
City of Angels
Wild Wild West
Con Air
Mulan
Batman
Big Daddy
Lethal Weapon 4
End of Days
Runaway Bride
Batman Forever
Patch Adams
My Best Friend's Wedding
Pulp Fiction
A.I. Artificial Intelligence
X-Men
Shrek
Unbreakable
Movie Title


Relationship between Review Year and Release Year
Release Year
Review Year
2020
2010
2000
1990
1980
1970
1960
1950
1940
1920
1940
1960
1980
2000


Relationship between Release Year and Lifetime Gross
1e8
Lifetime Gross (in dollars)
Release Year
6
5
4
3
2
1940
1950
1960
1970
1980
1990
2000
2010
2020