Movie Rating Prediction EDA

Danielle Contreras, Aidan Feeley, Matt Poiesz

Research Question and Methods

Can we predict unreleased movies' IMDB ratings using statistical and machine learning methods?

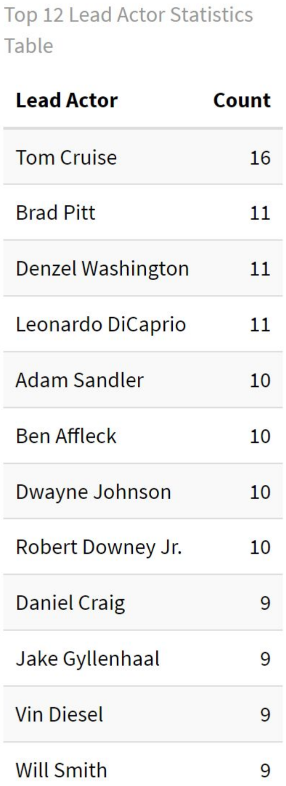
Methods:

- We will start out with **Rating** as the response. Using **Genre**, **Runtime**(minutes), **Budget**($1,000,000 >), **Director**, **Lead Actor**, **Votes**, **Rank**, and **Sequel/Remake**(Binary)) as the predictors.
- We will use **Multiple linear regression**, **KNN**, and **Random Forest** to predict ratings.

  * We will use some sort of variable selection and cross validation to evaluate the models.

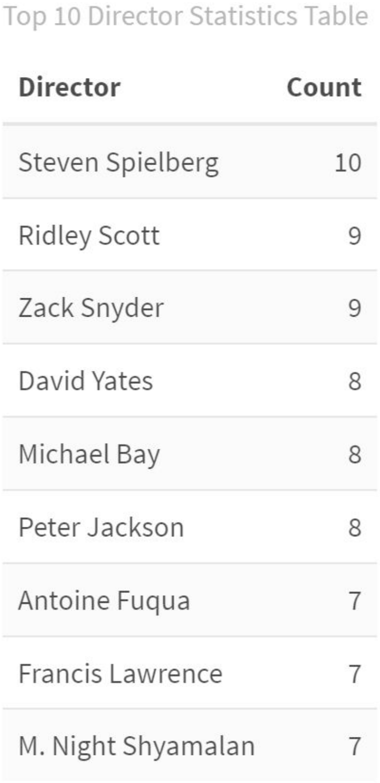
  * We will split our data 70/30 for training and test sets.

Summary Statistics (Tables)



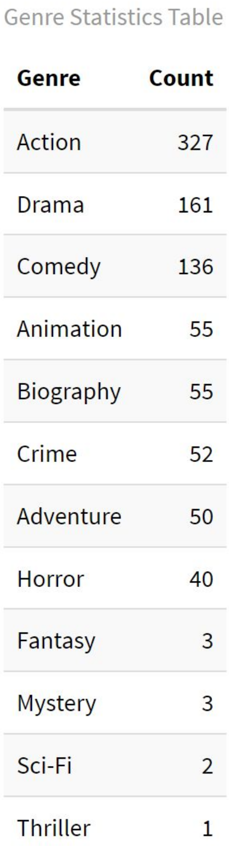
Top 12 Lead Actor Statistics Table

Lead Actor	Count
Tom Cruise	16
Brad Pitt	11
Denzel Washington	11
Leonardo DiCaprio	11
Adam Sandler	10
Ben Affleck	10
Dwayne Johnson	10
Robert Downey Jr.	10
Daniel Craig	9
Jake Gyllenhaal	9
Vin Diesel	9
Will Smith	9



Top 10 Director Statistics Table

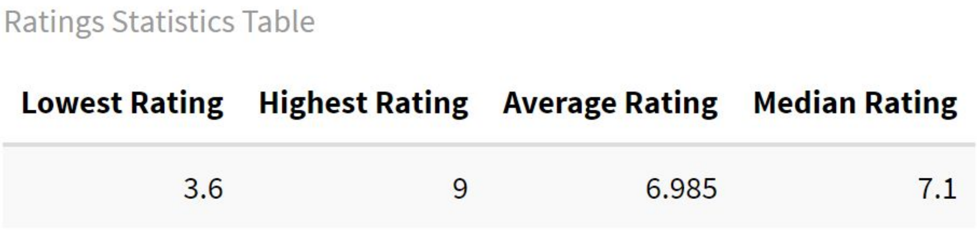
Director	Count
Steven Spielberg	10
Ridley Scott	9
Zack Snyder	9
David Yates	8
Michael Bay	8
Peter Jackson	8
Antoine Fuqua	7
Francis Lawrence	7
M. Night Shyamalan	7



Genre Statistics Table

Genre	Count
Action	327
Drama	161
Comedy	136
Animation	55
Biography	55
Crime	52
Adventure	50
Horror	40
Fantasy	3
Mystery	3
Sci-Fi	2
Thriller	1

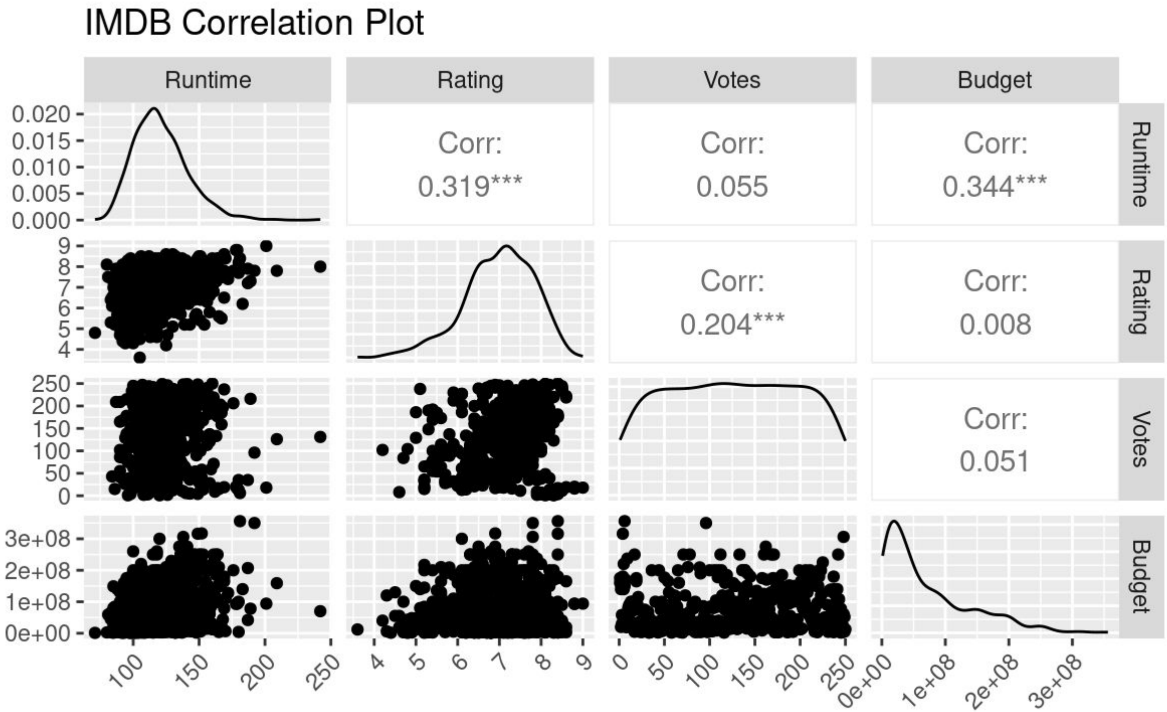
Summary Statistics



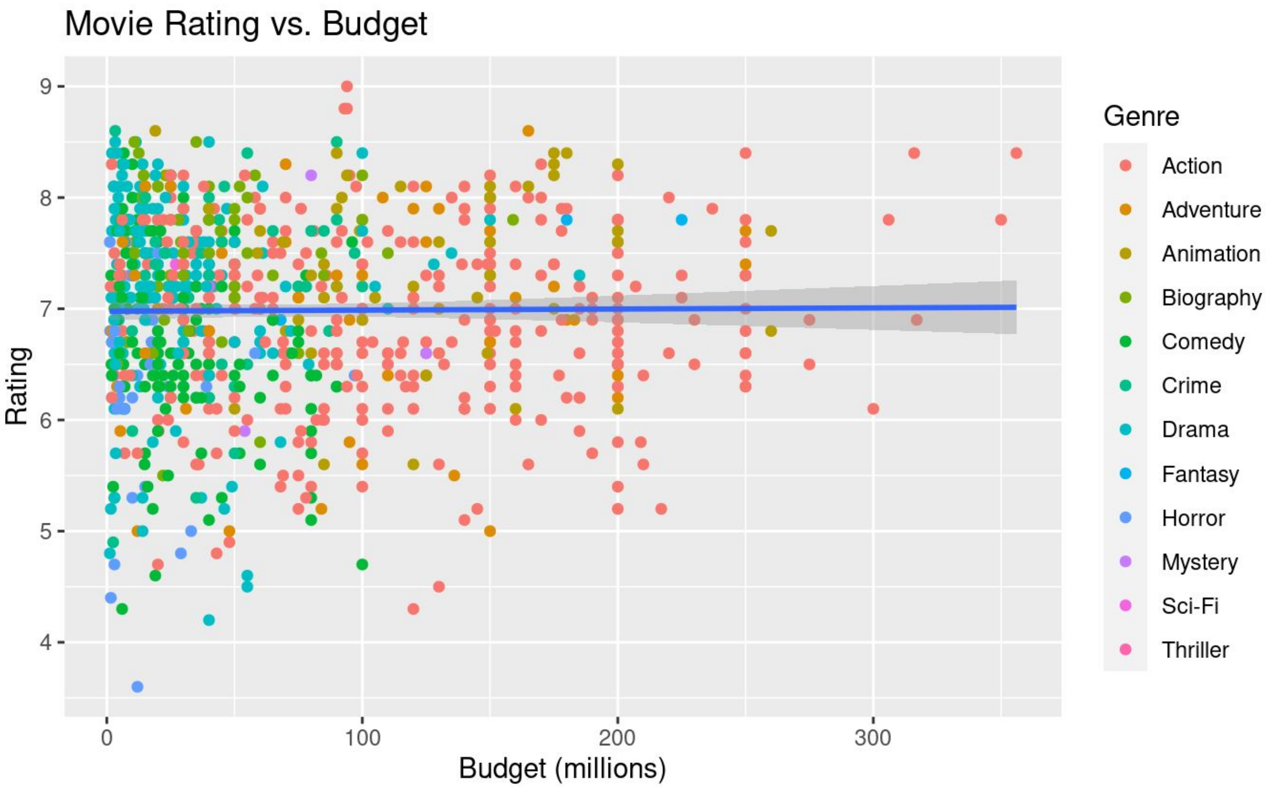
Ratings Statistics Table

Lowest Rating	Highest Rating	Average Rating	Median Rating
3.6	9	6.985	7.1

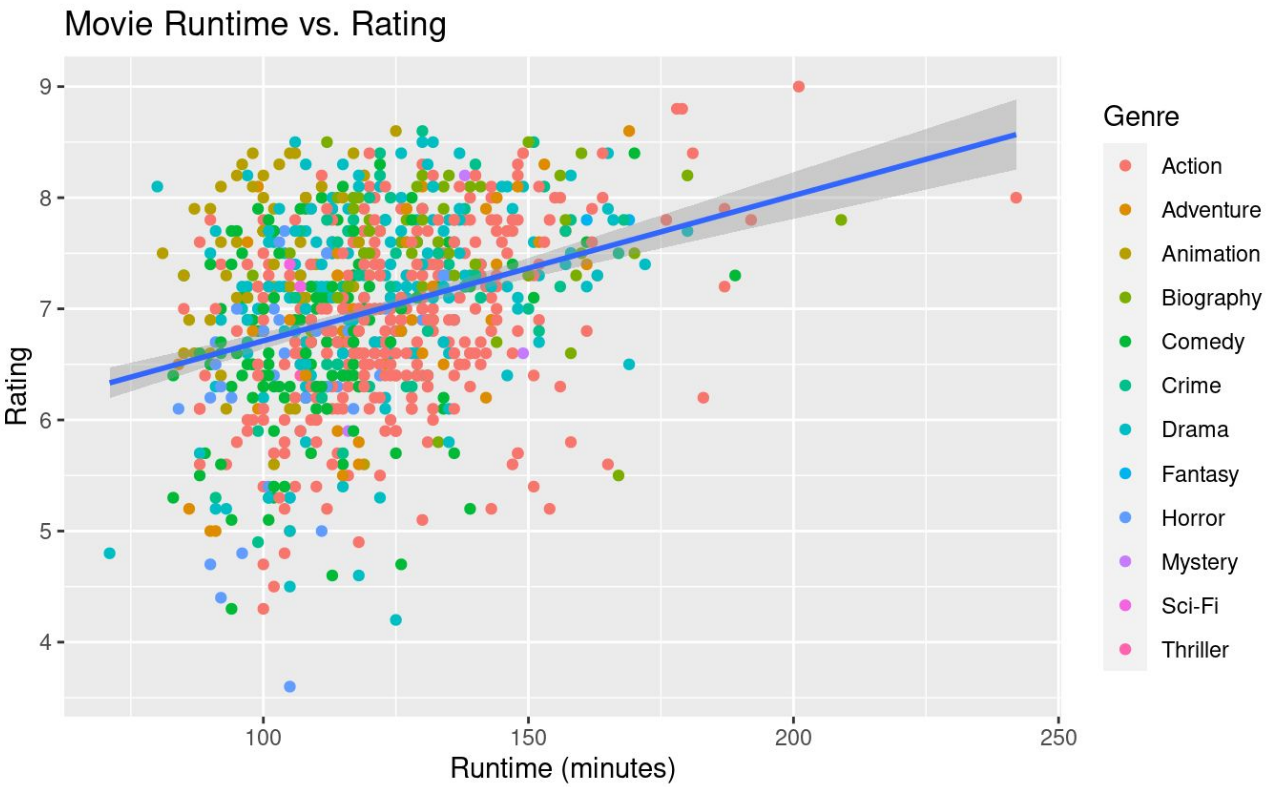
Correlation Plot



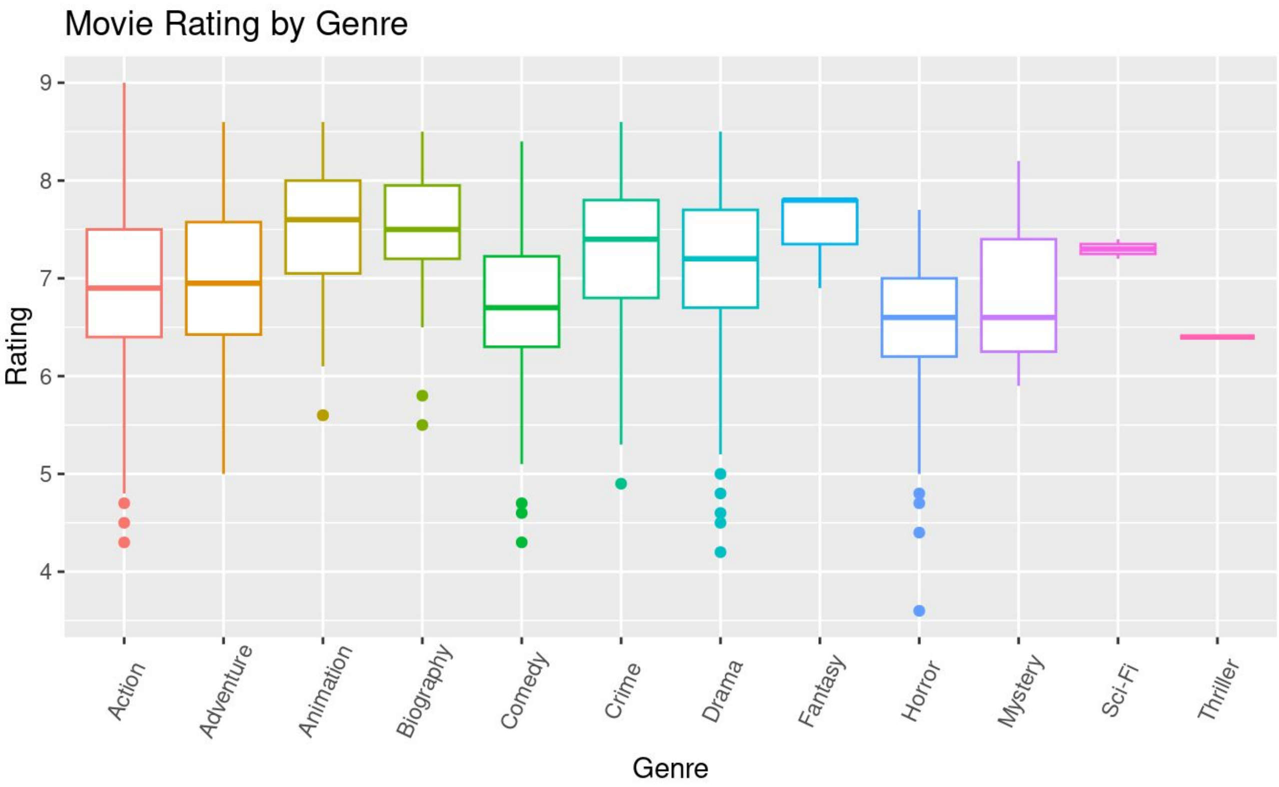
Rating vs. Budget



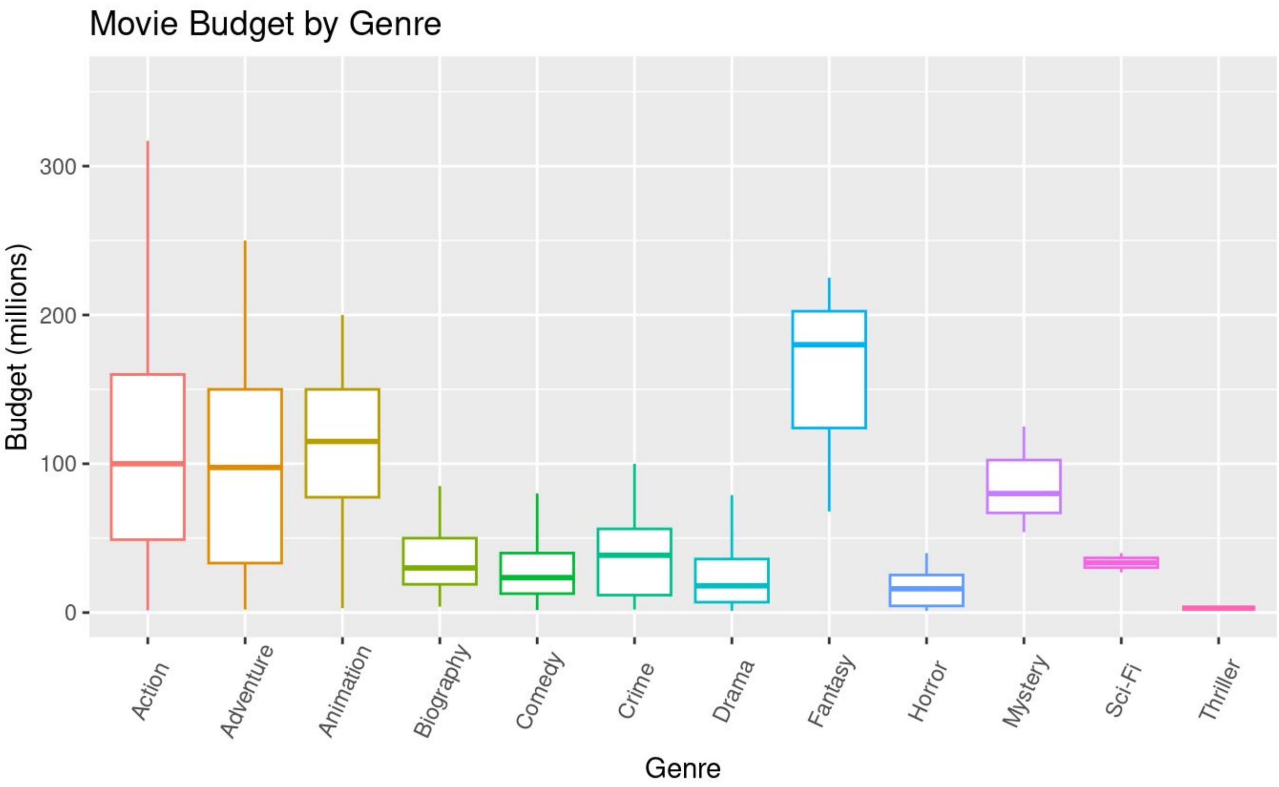
Rating vs. Runtime



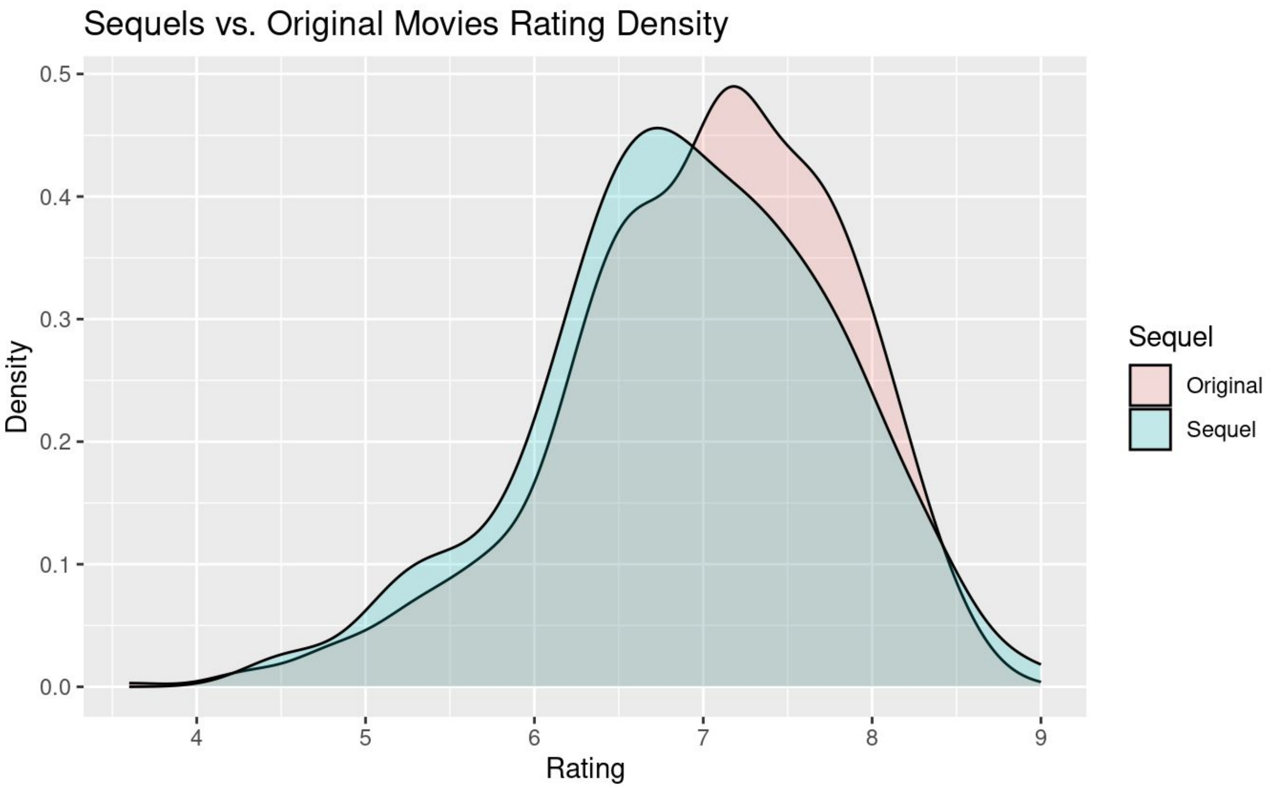
Rating by Genre



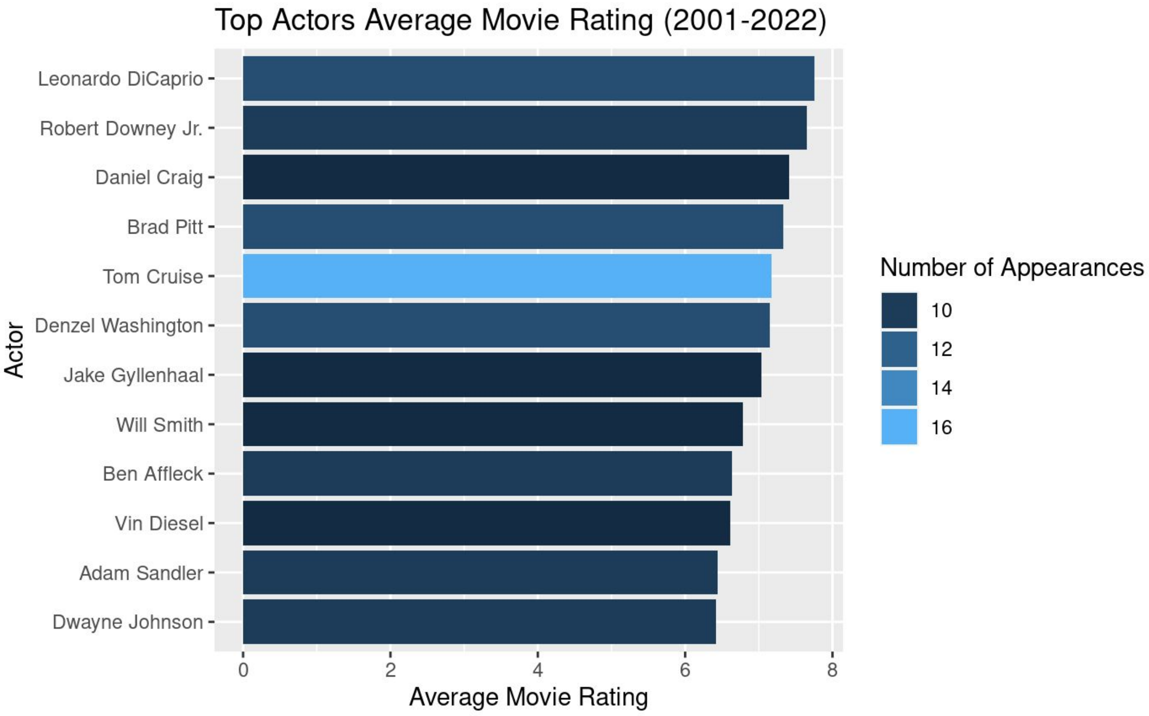
Budget by Genre



Density Plot



Average Movie Rating for Top Actors



Issues/Concerns

- Lots of actors and directors only have one data point
- Some genres only have one movie
- Possible to have a movie coming out that doesn’t include some of our categorical variables (Director, Lead actor/actress, genre)

Timeline

- **Now - April 1st** Explore our models(selecting variables and evaluating how well our models work)
- **April 21st** - Predict newly released movie ratings
  - We hope to use our model to predict movies coming out around April 21st
- **April 21st - May 1** write up final presentation and report