Machine Learning and Movies

By: Owen Doyle and Justin Self

Overview of Project

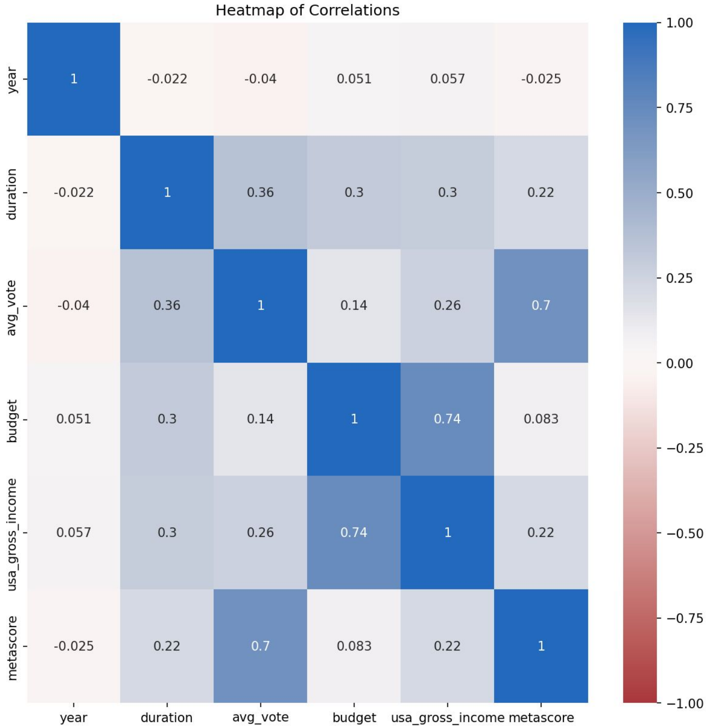
- Movies are, to most, a great form of entertainment
- We want to learn more about what makes a movie “enjoyable” (IMDb ratings)
- There are many applications of this research
  - Generating Revenue
  - Oscar Contenders
  - Making Enjoyable Movies in General

Data Collection/Cleansing

- Data set published on Kaggle by Stefano Leone
- Contained all movies with more than 100 votes on IMDb website
- Additional data cleansing was necessary (done in RStudio)
    - Made in USA
    - English
    - After 1990
    - Over 500 votes
    - Observations with null values removed
- New variable added for categorizing low, middle, and high ratings

Exploratory Data Analysis

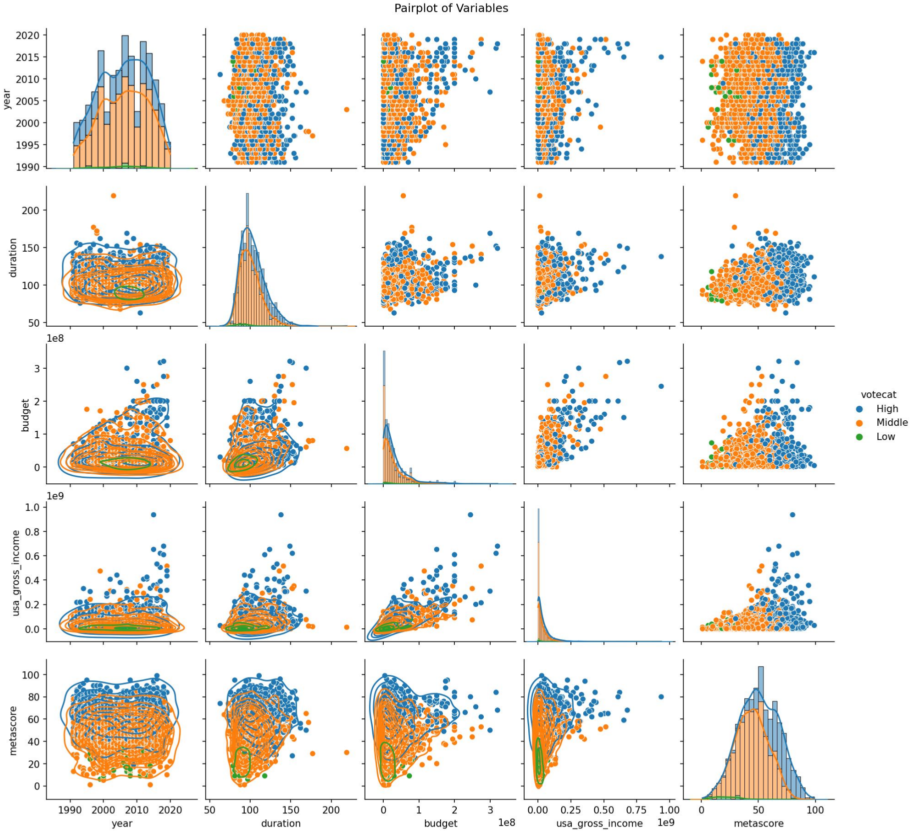
- Variables narrowed to some of the seemingly most important variables
- Duration, budget, USA gross income, and metascore had highest correlations with average vote
- Correlation between budget and average vote was surprisingly low



	year	duration	avg_vote	budget	usa_gross_income	metascore
year	1	-0.022	-0.04	0.051	0.057	-0.025
duration	-0.022	1	0.36	0.3	0.3	0.22
avg_vote	-0.04	0.36	1	0.14	0.26	0.7
budget	0.051	0.3	0.14	1	0.74	0.083
usa_gross_income	0.057	0.3	0.26	0.74	1	0.22
metascore	-0.025	0.22	0.7	0.083	0.22	1

Pairplot

- Same variables included in pairplot
- Metascore shows most differentiation in ratings
- USA gross income and duration also show some differentiation



Conclusions

- Metascore, duration, and USA gross income will be most important factors for our model
- Possible limitations of these variables
  - Metascore might be too similar to average rating
  - USA gross income cannot be determined until after the movie is released
- More variables related to actors might be helpful to making a more successful model
- Regression model will likely be most effective because we are trying to estimate relationship between independent variables and a dependent variable