Movie Gross Prediction- Part 2

Featuring the Same 3 People

What Have We Done

- Continuing to analyze the top 1000 rated movies
  - Remaining length is 750
- Data set is from IMDB and was found on Kaggle
- Inflation adjusted gross was made by multiplying by an inflation index centered around 1983
  - No change to a movie made in 1983, with older ones adjusted up and newer ones adjusted down

Cleaning the Data

- Removing “ min” from Runtime and converted to an integer
- Extra spaces
- Data points had to be removed
  - Those missing earnings or meta scores
  - Some columns that were found to be meaningless were removed
- Dummy Variables added to represent big name directors
  - Defined as 6+ movies on top 1000, but was later adjusted to 5

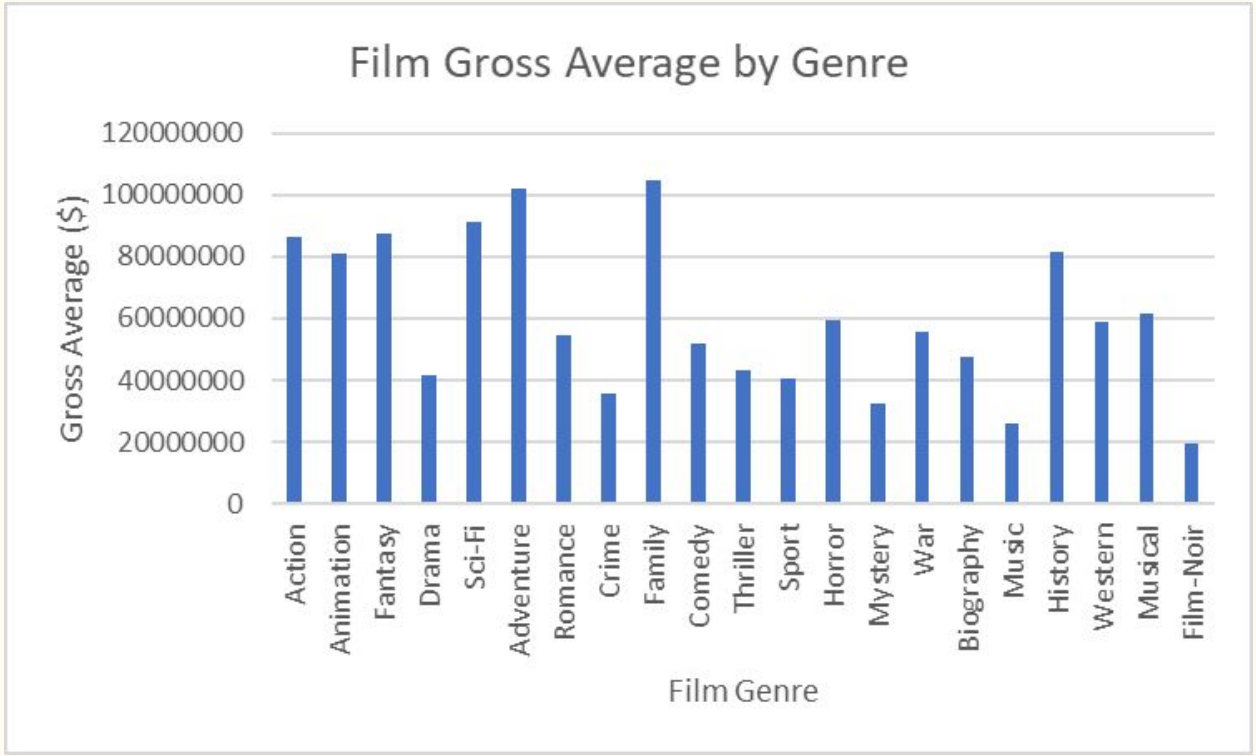
Data Analysis Correlation

- Averages and simple correlations were examined to determine what data points could be useful to us
  - Some surprising results

	Release Year vs Gross	Release Year vs Adjusted Gross	IMDB Rating vs Adjusted Gross	Number of Votes vs Adjusted Gross	Runtime vs Adjusted Gross	Runtime vs Adjusted Gross
Correlation Coefficient	0.236	-0.179	0.153	0.327	0.240	0.072
R^2 Value	0.055	0.031	0.023	0.106	0.057	0.0052

Data Analysis Averages

- Genre averages were examined with significant differences
  - Genre, Genre2, and Genre3 were taken into account
  - Ranging from Noir in last to Family in first
- Notable directors make a meaningful difference when the cutoff is at 5
  - Not if you put it at 6 due to James Cameron and Peter Jackson



Where Are We Going From Here

- Machine Learning with our variables
- Trim and adjust chosen variables as needed to get the best and simplest model
- Use plots to check whether fit is accurate or not