Movie Report UofT

By Brian Haley, Rustem Bogdanov, and Shaun MacLellan

Introductions:

Our goal with this project is to use multiple sources of information to identify unique insights with all movies that have been produced since 1901. We looked at the data gathered related to genre, director and year to see if it had any correlation with box office revenue and the budget of the film.

- IMDB
- Wikipedia

Methods: *Step by Step*

1. We imported pandas, MySQL, pymySQL
2. Imported CSV file from Wikipedia Movie Plots
3. Cleaned Wikipedia CSV file and dropped columns
4. Cleaned Wikipedia CSV file and renamed columns
5. Import CSV from IMDB data set. On movies
6. Cleaned IMDB CSV file and dropped columns
7. Cleaned IMDB CSV file and renamed columns
8. Merged IMDB and Wikipedia data sets on the join of IMDB index
9. Then imported into MySQL

Findings:

We were able to connect the plot and origin of the movie into a much larger dataset that is clean and able to be manipulated. .

Questions to Discuss:

1. What would be the strongest insights we could gather from the new data set?
2. Compared the budget of color to non-color movies?

3. How do the budgets of movies change with inflation taken into consideration?