




# FRESH OVER ROTTEN

THE VALUE OF A CRITIC'S PICK IN HOLLYWOOD

Sergio E. Betancourt  
STA2453 – Graduate Data Science Consultation  
November 27, 2018



# Motivation

- PwC<sup>1</sup> estimated global film revenue: USD ~\$38.3 billion (2016)
  - *Forecasted growth for the foreseeable future*
- In the US alone, estimated film revenue (2016): USD ~\$11.38 billion (30% of World's)
- Players: Production and distribution firms, cast and crew, governments and multinationals, etc.
- **Goal:** Understand the factors influencing film earnings, as well as film prestige in the shape of critics' reviews and reputable awards

<sup>1</sup> <https://www.statista.com>

# Direction

- What is the marginal financial contribution of a New York Times (NYT) critic's pick to a movie's revenue? How does box office performance differ across movie genres?
- Is there a relationship between critics' picks and earning a nomination to a prestigious award?
- What are the factors that contribute to earning a nomination and what is the role of review sentiment?

# Data

- Period: 2010 - 2017 (inclusive)
- Sources: New York Times' API<sup>2</sup> and IMBD's OMDB API<sup>3</sup>
- NYT information:
  - *Title*
  - *MPAA Rating*
  - *Critics' Pick (0 or 1)*
  - *Short Summary*
- OMDB information:
  - *Title*
  - *Box Office Earnings (taken as gross box office to present day)*
  - *Genre*
  - *Director*
  - *Actors*
  - *Awards*
  - *Various Critical Ratings*

<sup>2</sup>[https://developer.nytimes.com/movie\\_reviews\\_v2.json#/README](https://developer.nytimes.com/movie_reviews_v2.json#/README) ; <sup>3</sup><http://www.omdbapi.com>



# Questions vs Data

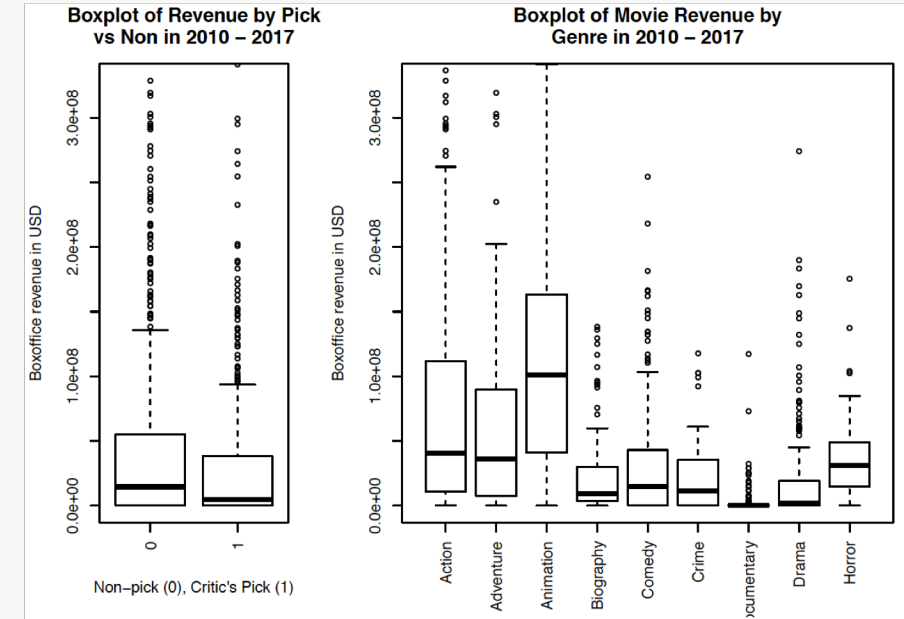
- Sourcing and choosing the data
  - *Legal vs illegal, useful vs not useful, granular vs general, etc.*
- Continuous process
  - *Adapting one for the sake of the other*
- Is what is at hand enough or are larger quantities needed? Redundancies?
- How to get more out of what is available?

# Extracting Meaning from Data

- Clean up and formatting
  - *Creating consistency, filtering, joining (good practices)*
- Missing Data
  - *Random vs non-random*
  - *Different title spellings in APIs*
- Feature Engineering
  - *Principal director and lead actor*
  - *Genre of each (M/F)*
  - *Nominated to at least one Oscar, Golden Globe, or BAFTA*
    - Captures wins, since winning implies nomination

# Exploration

- IDA
  - *Summary statistics*
  - *Tabulation*
  - *Rudimentary plots*
- Codification
- Sophistication versus ease of implementation
- Revisit research questions if needed
  - *Formalize plan of attack*



	Not Picked	Picked	Total	Proportion Picked (%)
Action	248	31	279	11
Adventure	47	26	73	36
Animation	51	25	76	33
Biography	47	34	81	42
Comedy	233	57	290	20
Crime	41	12	53	23
Documentary	106	65	171	38
Drama	180	97	277	35
Horror	31	11	42	26

# Analysis

- Marginal financial contribution to film revenue
  - *Log-Linear Regression*
  - *Review: diagnostics, transformations, coding, etc.*
- Variables contributing to probability of nomination
  - *Logistic Regression*
  - *Review: diagnostics, transformations, coding, etc.*
- Agreement of critic's picks and review language
  - *Sentiment Analysis*
  - *Binary Naïve Bayesian Classifier (New technique)*
  - *Integration of R and Python (reticulate)*
- NYT critics' picks and award nomination
  - *Categorical methods*

Specifics available in the full report:

[https://sergiosonline.github.io/files/Data\\_Science-Project\\_I\\_Report-20181113.pdf](https://sergiosonline.github.io/files/Data_Science-Project_I_Report-20181113.pdf)



# Results

- Marginal financial contribution to revenue (as multiplicative factors over overall mean)
  - *Genre*
    - Action (3.1), Documentary (.17), Drama (.4)
  - *Nomination to an award (7.15)*
  - *Rating*
    - Not Rated (.01), Restricted (.21)
- Genre contributor to the probability of earning a nomination
  - *Female lead actor (Almost doubles the odds of earning)*
- Agreement of critic's picks and review language
  - *Positive agreement, albeit less dramatic than expected*
- NYT critics' picks and award nomination
  - *Earning a pick yields almost 4 times the probability of earning a nomination*

Specifics available in the full report:

[https://sergiosonline.github.io/files/Data\\_Science-Project\\_I\\_Report-20181113.pdf](https://sergiosonline.github.io/files/Data_Science-Project_I_Report-20181113.pdf)

# Further Thoughts

- Goal: Straightforward and sound analysis
- Extensions to scope
  - *Longer time period*
  - *More variables from other sources, especially review content from other sites*
  - *Non-traditional media distributors and content*
- Sentiment analysis can be greatly expanded
  - *Complete review text directly from website (not legal...)*
  - *Non-naïve models in order*
- Many ways to extend research questions – When to stop?



Q+A

