PREDICTING MOVIE BOX OFFICE USING SOCIAL DATA

ANLY500 Group #6

Yuanjie Lei

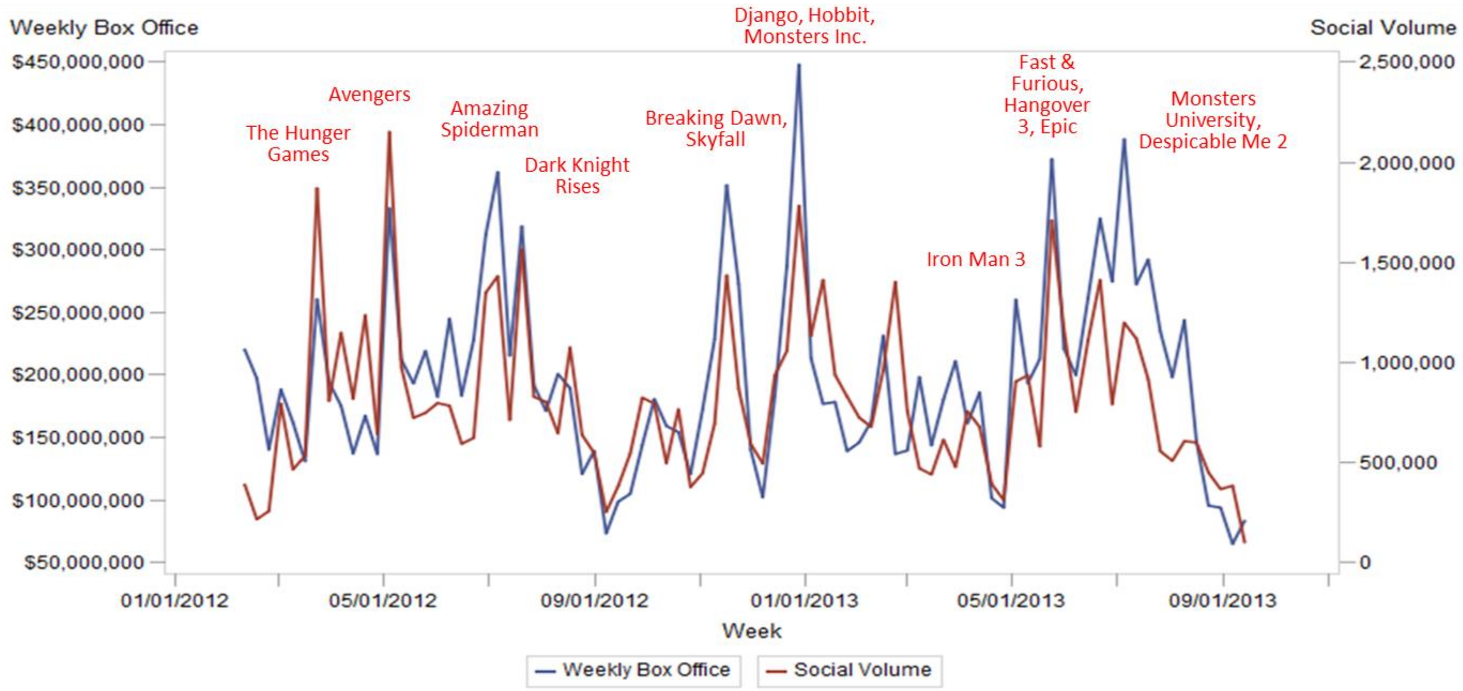
Fangya Tan

Shuqing Yang

Yanhong Ye

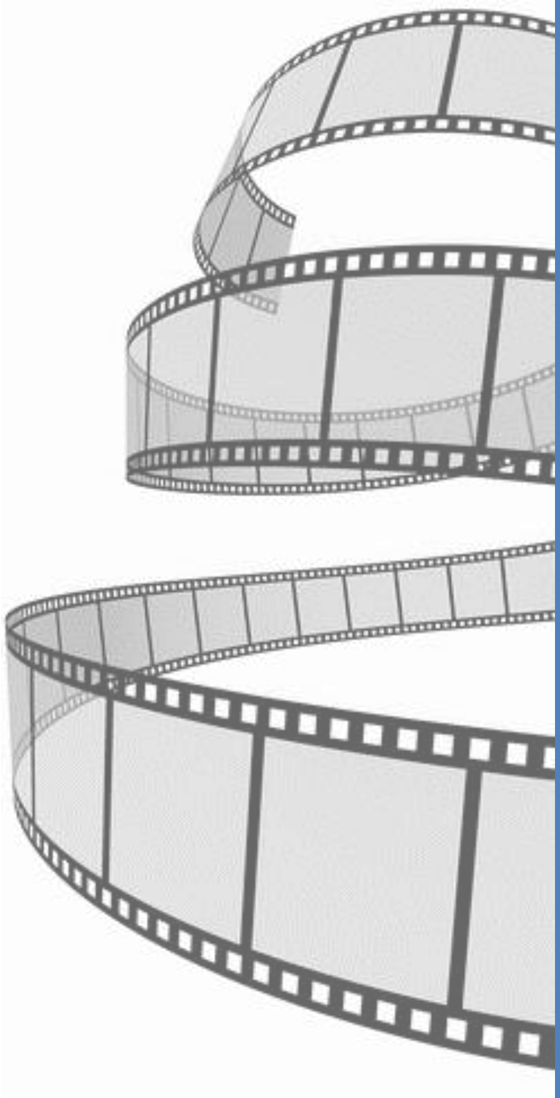
Introduction

- We found that the volume of social conversations of movies is highly correlated with box office.
  - E.g. The more people discussing a movie on Twitter, the higher the movie's box office.
- Our goal of this study is using social data to predict box office.



Research Question

- Study the correlation between Twitter data and/or other factors and movies' opening weekend box office in the US movie market.
- Can we use social network (Twitter) data to predict movies' opening weekend box office?



Data Sources

Movies Meta Data

- We collected movies meta data from some public sources online, key variables include movie, movie id, release date, genre, box office of the opening weekend, rating, theater counts. There are 417 movies in the dataset.

Box Office Mojo

IMDb

VARIETY Insight
A Service of Variety Media, LLC

Movies Volume Data

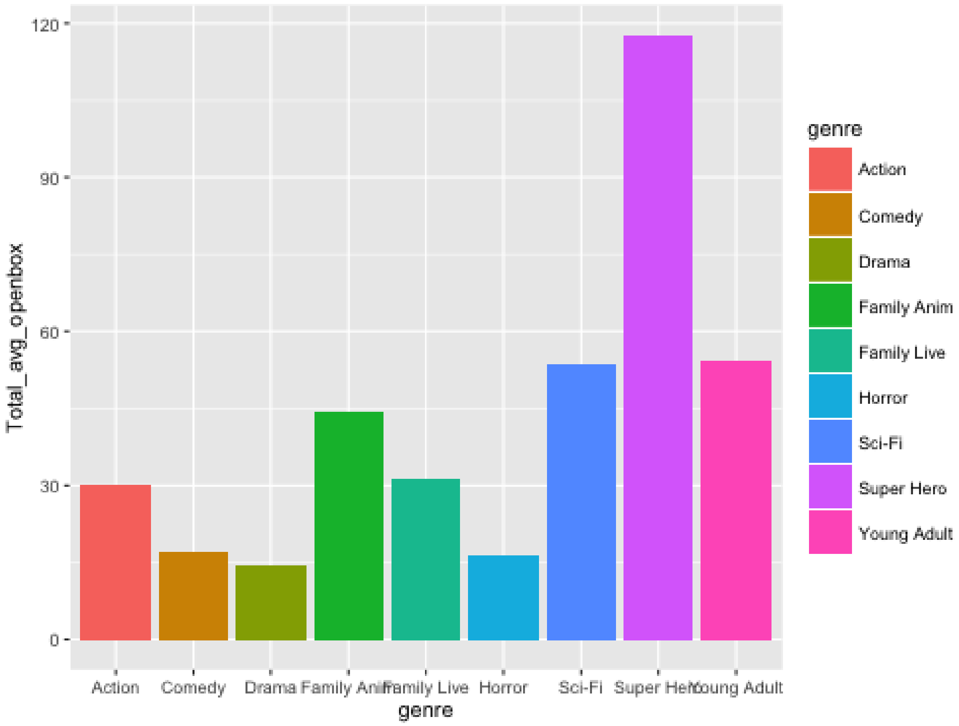
- We obtained the volume data from a social network analytics firm, 'Networked Insights'. The dataset has the daily conversation volume of movies on Twitter. Key variables include movie id, volume date, total volume, intent volume, and etc. The dataset has 84064 records.

NETWORKED INSIGHTS

Data Pre-processing

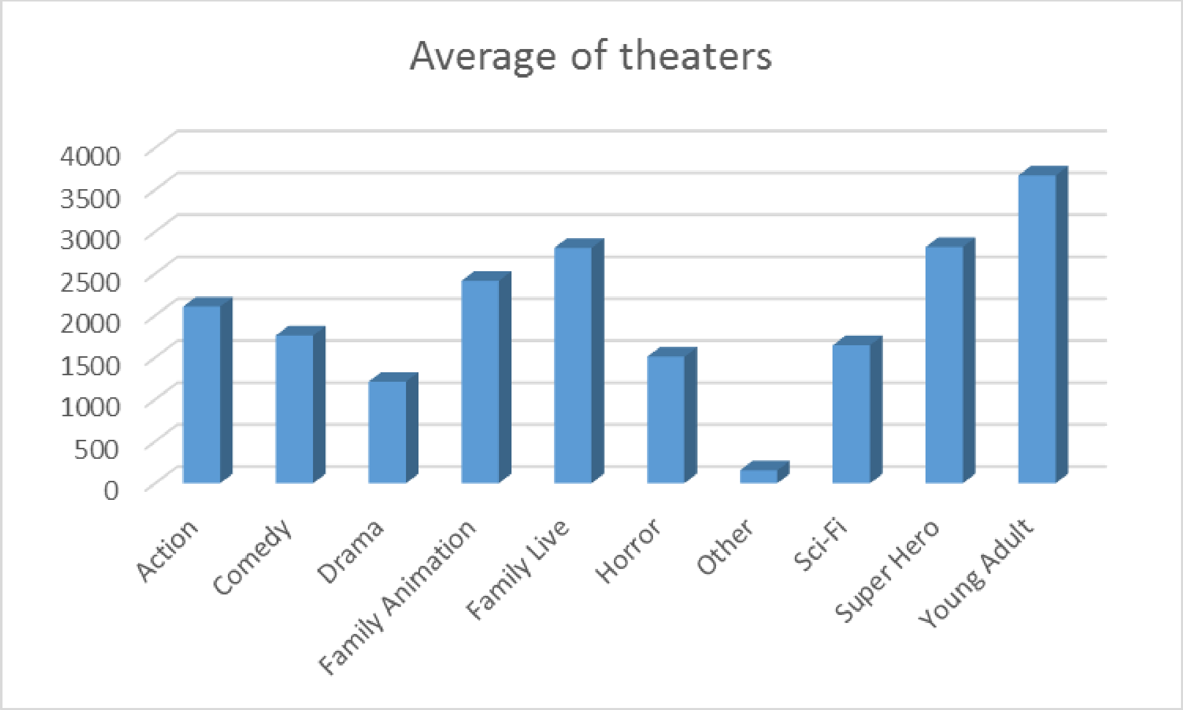
- Cleaned movie meta data, removed unreleased movies (box office is NA).
- Aggregated each daily volume variables to total volume. (post volume, intent volume, positive volume, negative volume, female volume, male volume, organic volume)
- Joined two datasets by movie id for EDA and model building.

EDA 1. Is there any differences in behavior between different genres?



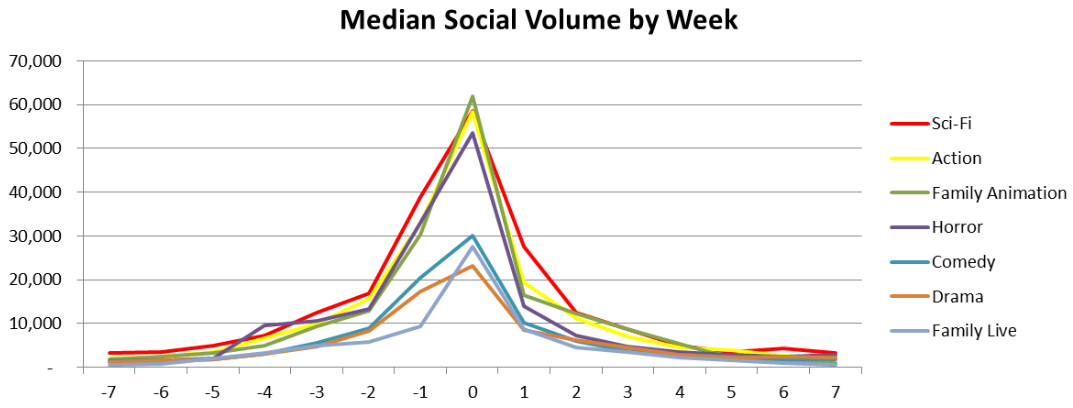
- Super Hero had the highest box office on opening weekend, following by Young Adult and Sci-Fi.
- Drama has the lowest box office on average.

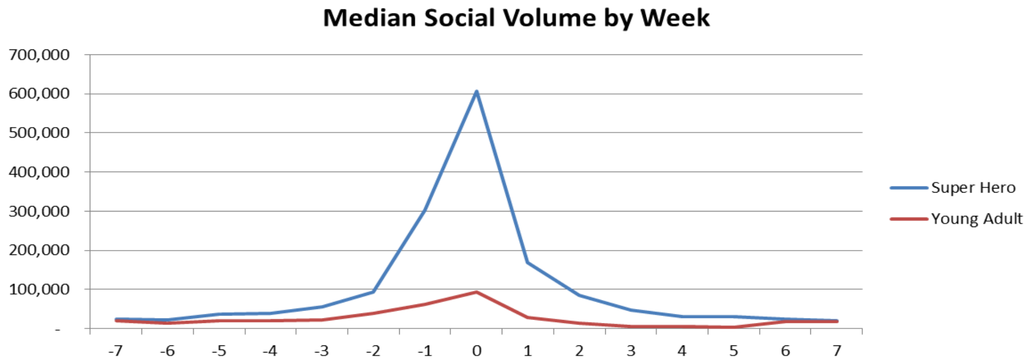
EDA 1. Is there any differences in behavior between different genres?



- Young Adult and Super Hero have the highest average theater counts.
- Drama has the lowest average theater counts.

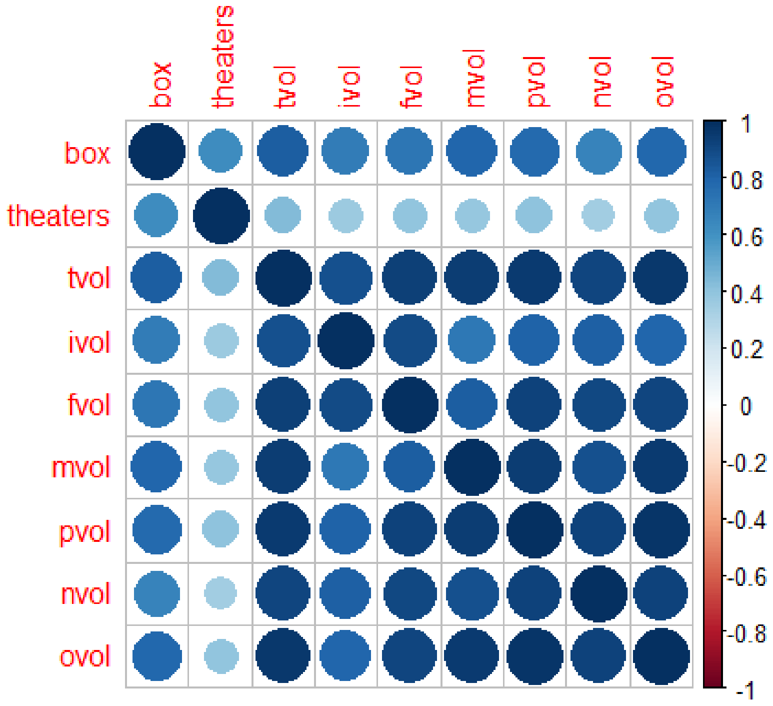
EDA 1. Is there any differences in behavior between different genres?


Median Social Volume by Week


Median Social Volume by Week

- We found significant growth of volume at week -4, likely due to TV & digital media campaigns.
- Most genres double their volume from week -2 to week -1.
- Super Hero and Young Adult movies see significantly higher volumes at all weeks followed by Sci-Fi, Action and Family Animation.

EDA 2. Which variables are correlated with box office?

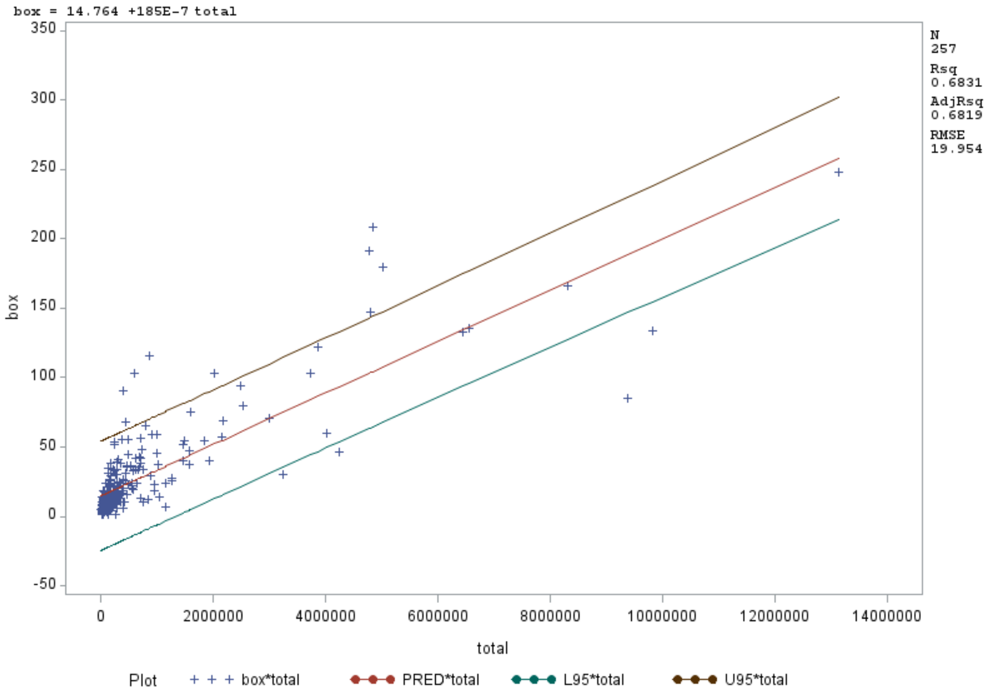


- Total volume, intent volume, and organic volume are highly correlated with box office. However, we should be aware of the multicollinearity issue since all 7 different types of volume are highly correlated with each other as well.
- Theater counts is also positively correlated to box office.

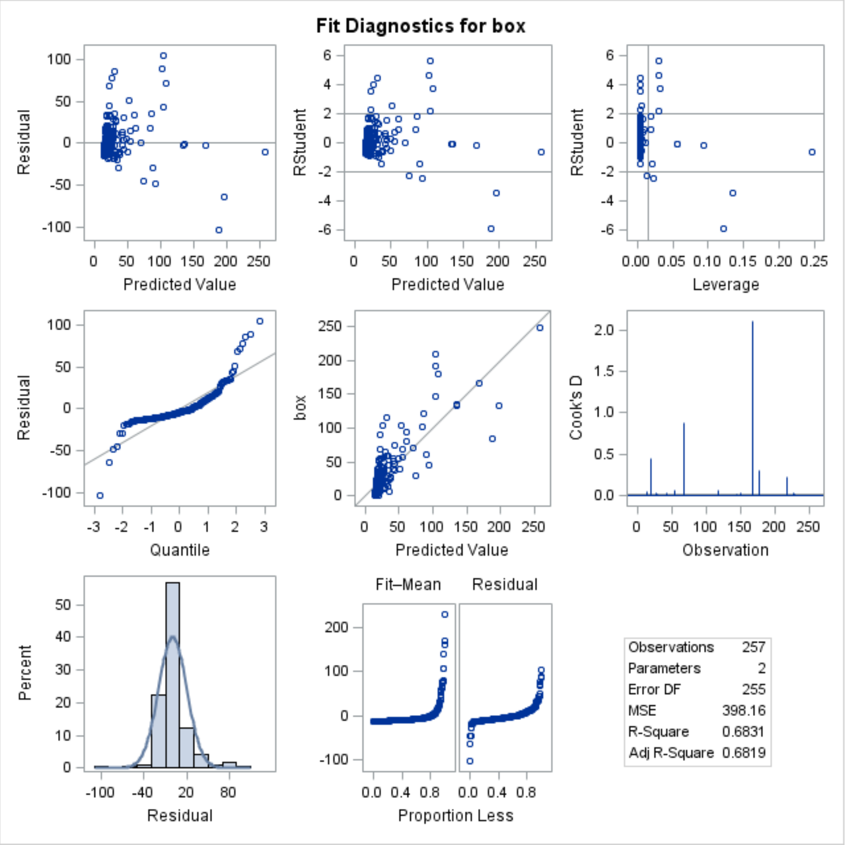
Model 1.Linear Regression Model

- Dependent Variable: Box Office
- Independent Variable: Sum of Post Volume
- A film's social volume alone is a strong indicator of opening weekend performance.
- This simple linear regression model using the sum of post volume as a predictor of opening weekend performance yields an R2 =68.31

Model 1.Linear Regression Model



box = 14.764 +185E-7 total



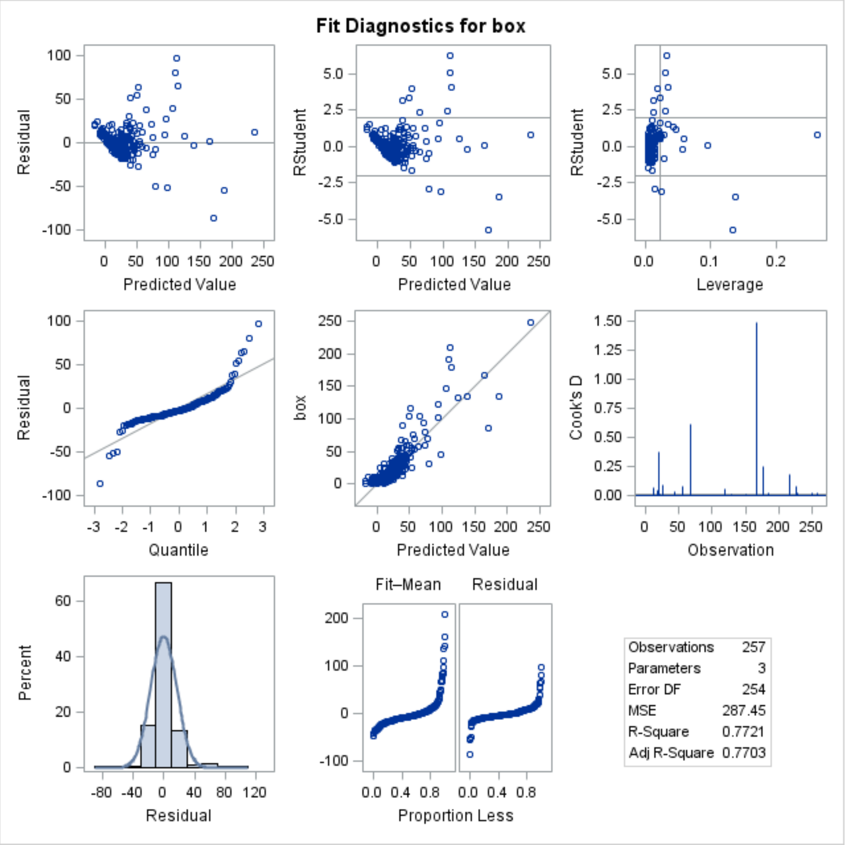
Fit Diagnostics for box

Observations	257
Parameters	2
Error DF	255
MSE	398.16
R-Square	0.6831
Adj R-Square	0.6819

Model 2.Multiple Linear Regression Model - Improving the Predictive Model

- Dependent Variable: Box Office
- Independent Variable: Sum of Post Volume, Theater Counts
- Predicted opening box revenue=-33.447+1.53e(-5)*total volume +0.0167*theater number
- Not all films are alike. Adding factor Theatre Counts greatly improves the accuracy of the model.
- R2 =

Model 2.Multiple Linear Regression Model - Improving the Predictive Model



Observations	257
Parameters	3
Error DF	254
MSE	287.45
R-Square	0.7721
Adj R-Square	0.7703

Example - Prediction of Tomorrowland

- Predicted opening box revenue=-33.447+1.53e(-5)*total volume +0.0167*theater number
- Total volume of Tomorrowland=250730, Theater number of Tomorrowland=3972
- Predicted opening box revenue=32.889 Million dollars
- Actual opening box revenue=33 Million dollars
- Standard error of opening box revenue=0.012

Limitation & Uncertainty

- Twitter data may not be a good representation of the population.
- Studio who invests more on promoting a movie may have higher volume than who invests less.
- We haven’t taken into account the sentiment of the post volume, a movie who has high negative post volume might result in lower box office.

Conclusion

- Social data is highly predictive of box office performance
- Social Volume *alone* explains 75% of the variance!
- Model prediction 10-weeks prior to release can predict opening weekend within 10 million 70% of the time.
- Studios can leverage these insights and make decisions based on social volume performance.