Tweevie

Ying J,
Seifert M.,
Zhang S.
Kwynn M .

Background

Movie box office is measured by the number of tickets sold, displaying the revenue raised by ticket sales daily. The projection and analysis of these earnings is very important for the creative and marketing industries, especially concerning the movie's following and its fans. Before the movie even premiers, twitter handles, hashtags, and the "drum roll" momentum has already commenced.

IAB (Interactive Advertising Bureau) reports that 42% of US smart-phone users under the age of 35 currently check their favorite social media apps before deciding on which movie to see.

Prior to "Black Panther" being released, it was ubiquitous on twitter and the social interactions foretold its success. Not only becoming the most tweeted-about movie of 2018, it became 9th in highest-grossing film in history.

Project Tweevie attempts to find the relationship between the Twitter-sphere to box office data.

Questions

- Which variables play an important role in predicting box office revenue?

- What are the relationships between box office and the various predictors?

- Which type of regression can best capture the relationship between box office and its predictors?

- How well do the regressions handle outliers?

Target Film Criteria

- Box office data is listed on Box Office Mojo
- Film is not a re-release or a limited release
- Plays in the US and is showing in more than 5 theaters
- Has an official hashtag that's reasonably exclusive to the film
- In theaters as of July 27th

Data Included

- 32 movies were included in the study
- Tweets mentioning the movies from July 18th to August 5th were retrieved using Tweepy
- Box office records from www.boxofficemojo.com were collected for the same time frame
    - Including both daily gross and number of theaters offering screening

Parameters

Independent Variables:

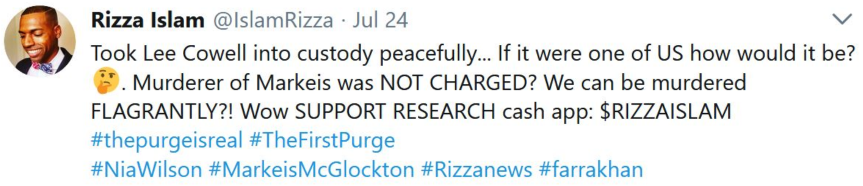
- Daily Twitter mentions
- IMDbPro STARMeter of top three billed actors from Wikipedia
- Days since release
- Weekday/Weekend
- Number of theater offering screening

Dependent Variable:

- Box office income per day

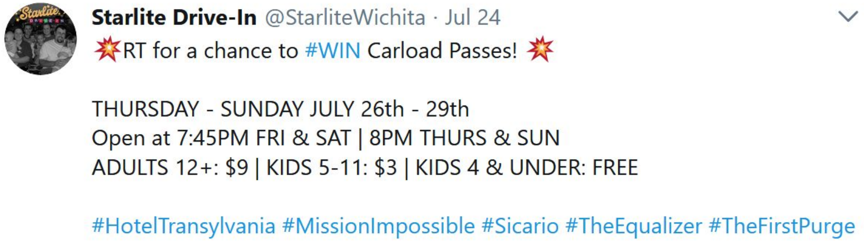
Cleaning Data

- For each movie, messages from the same twitter user is counted only once.
- Keywords/hashtags from unrelated tweets were added to a black list. Any tweet containing a blacklist word is excluded from the analysis.



Rizza Islam @IslamRizza · Jul 24

Took Lee Cowell into custody peacefully... If it were one of US how would it be? 🤔. Murderer of Markeis was NOT CHARGED? We can be murdered FLAGRANTLY?! Wow SUPPORT RESEARCH cash app: $RIZZAISLAM
#thepurgeisreal #TheFirstPurge
#NiaWilson #MarkeisMcGlockton #Rizzanews #farrakhan

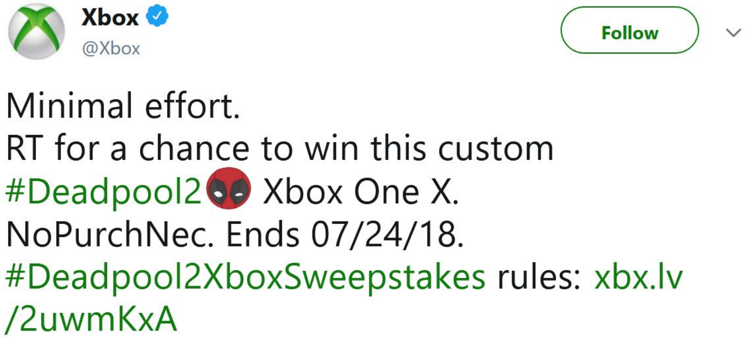


Starlite Drive-In @StarliteWichita · Jul 24

💥RT for a chance to #WIN Carload Passes! 💥

THURSDAY - SUNDAY JULY 26th - 29th
Open at 7:45PM FRI & SAT | 8PM THURS & SUN
ADULTS 12+: $9 | KIDS 5-11: $3 | KIDS 4 & UNDER: FREE

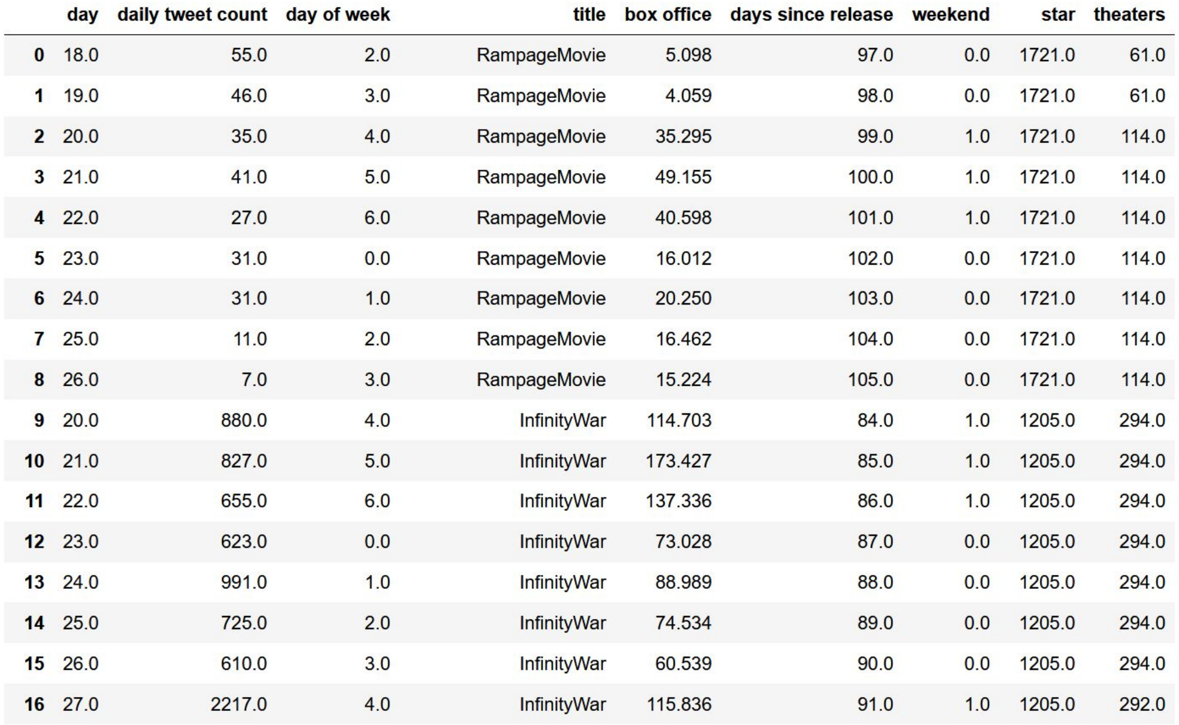
#HotelTransylvania #MissionImpossible #Sicario #TheEqualizer #TheFirstPurge



Xbox @Xbox

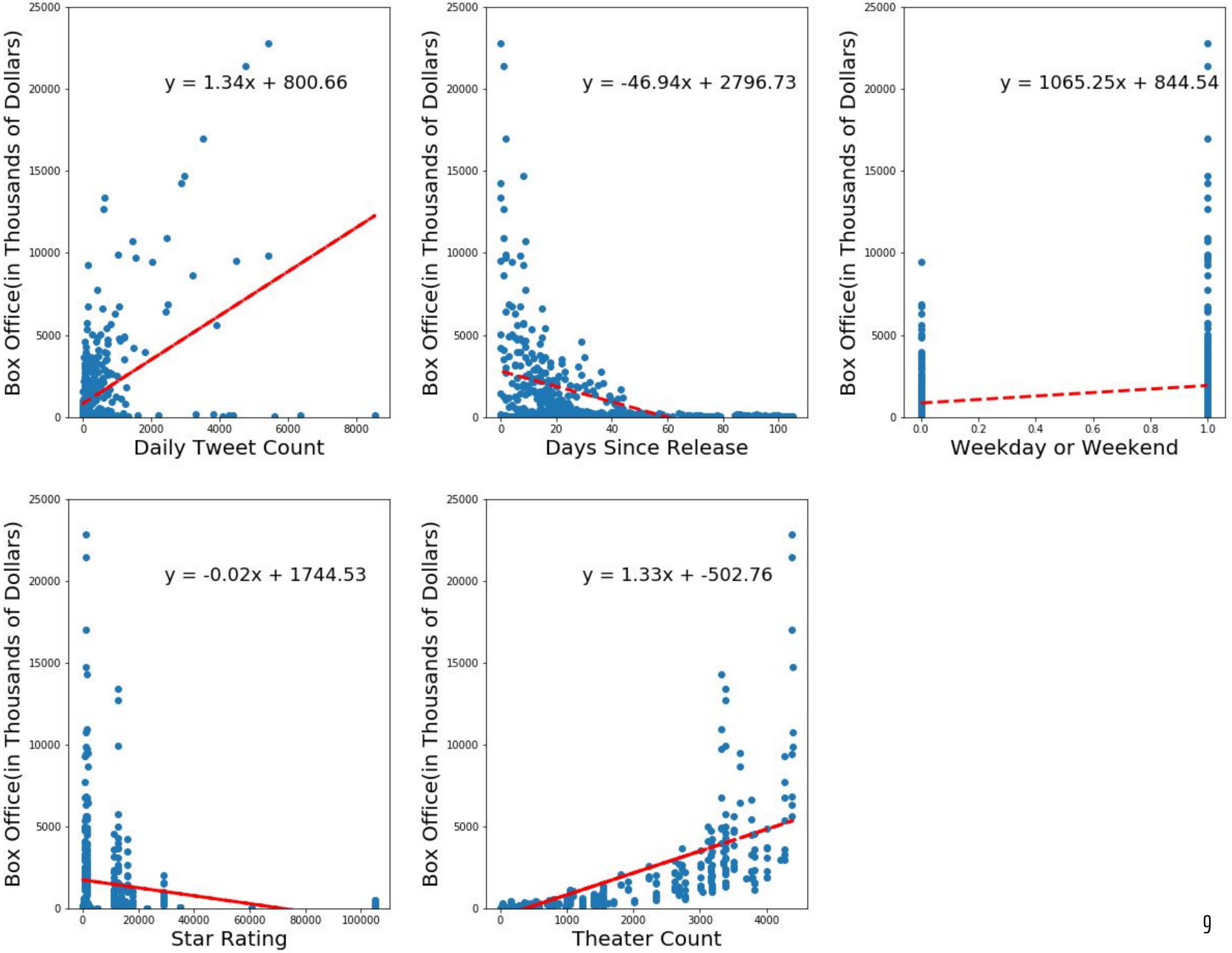
Minimal effort.
RT for a chance to win this custom #Deadpool2 Xbox One X.
NoPurchNec. Ends 07/24/18.
#Deadpool2XboxSweepstakes rules: xbx.lv/2uwmKxA

Example DataFrame



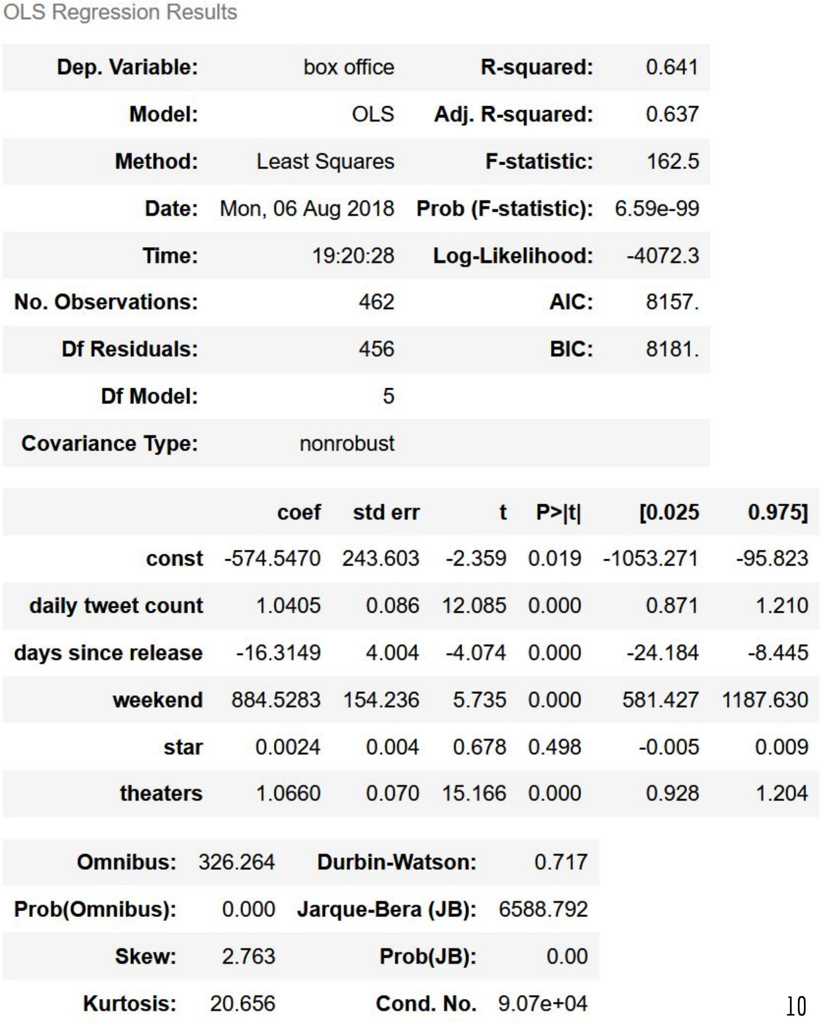
	day	daily tweet count	day of week	title	box office	days since release	weekend	star	theaters
0	18.0	55.0	2.0	RampageMovie	5.098	97.0	0.0	1721.0	61.0
1	19.0	46.0	3.0	RampageMovie	4.059	98.0	0.0	1721.0	61.0
2	20.0	35.0	4.0	RampageMovie	35.295	99.0	1.0	1721.0	114.0
3	21.0	41.0	5.0	RampageMovie	49.155	100.0	1.0	1721.0	114.0
4	22.0	27.0	6.0	RampageMovie	40.598	101.0	1.0	1721.0	114.0
5	23.0	31.0	0.0	RampageMovie	16.012	102.0	0.0	1721.0	114.0
6	24.0	31.0	1.0	RampageMovie	20.250	103.0	0.0	1721.0	114.0
7	25.0	11.0	2.0	RampageMovie	16.462	104.0	0.0	1721.0	114.0
8	26.0	7.0	3.0	RampageMovie	15.224	105.0	0.0	1721.0	114.0
9	20.0	880.0	4.0	InfinityWar	114.703	84.0	1.0	1205.0	294.0
10	21.0	827.0	5.0	InfinityWar	173.427	85.0	1.0	1205.0	294.0
11	22.0	655.0	6.0	InfinityWar	137.336	86.0	1.0	1205.0	294.0
12	23.0	623.0	0.0	InfinityWar	73.028	87.0	0.0	1205.0	294.0
13	24.0	991.0	1.0	InfinityWar	88.989	88.0	0.0	1205.0	294.0
14	25.0	725.0	2.0	InfinityWar	74.534	89.0	0.0	1205.0	294.0
15	26.0	610.0	3.0	InfinityWar	60.539	90.0	0.0	1205.0	294.0
16	27.0	2217.0	4.0	InfinityWar	115.836	91.0	1.0	1205.0	292.0

Exploratory Plots



Multiple Regression

A multiple linear regression was calculated to predict movie box offices based on daily twitter mentions, days past release, whether it’s a weekday or weekend, the combined star meter of lead actors, and number of theaters screening the film. A significant regression equation was found (F(5, 456) = 162.5, p = 0, with an R^2 of 0.64)



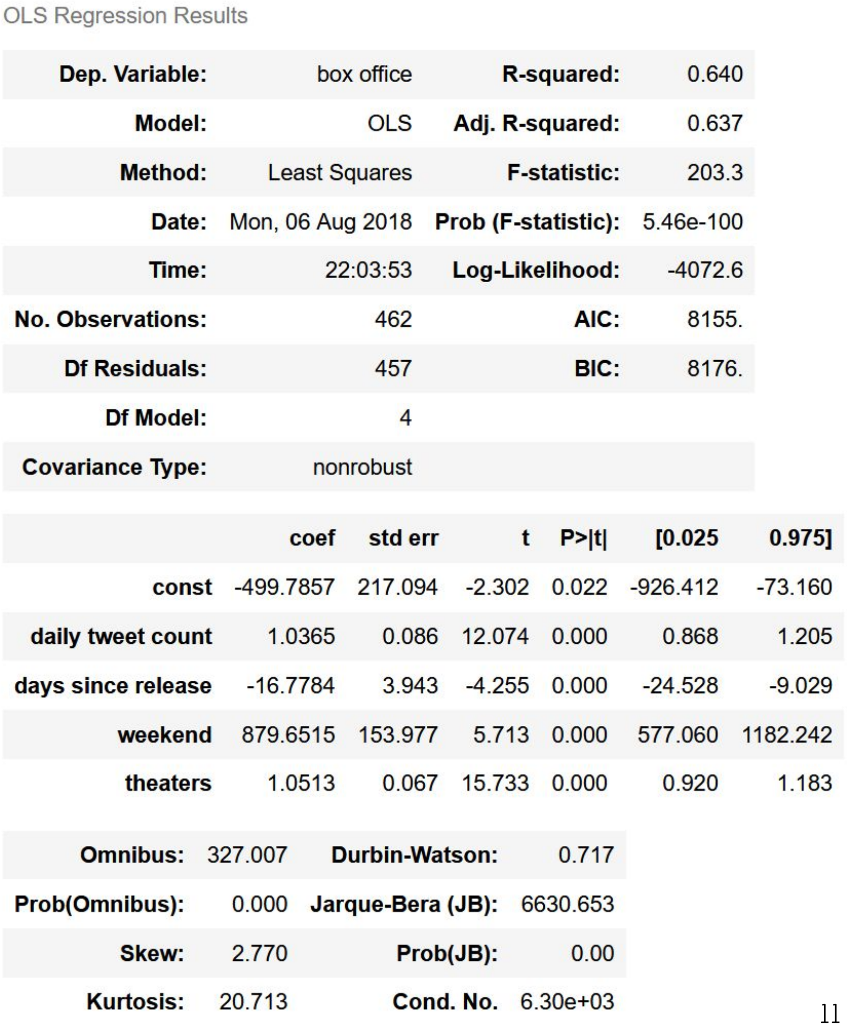
OLS Regression Results

Dep. Variable:	box office	**R-squared:**	0.641
Model:	OLS	**Adj. R-squared:**	0.637
Method:	Least Squares	**F-statistic:**	162.5
Date:	Mon, 06 Aug 2018	**Prob (F-statistic):**	6.59e-99
Time:	19:20:28	**Log-Likelihood:**	-4072.3
No. Observations:	462	**AIC:**	8157.
Df Residuals:	456	**BIC:**	8181.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-574.5470	243.603	-2.359	0.019	-1053.271	-95.823
daily tweet count	1.0405	0.086	12.085	0.000	0.871	1.210
days since release	-16.3149	4.004	-4.074	0.000	-24.184	-8.445
weekend	884.5283	154.236	5.735	0.000	581.427	1187.630
star	0.0024	0.004	0.678	0.498	-0.005	0.009
theaters	1.0660	0.070	15.166	0.000	0.928	1.204

Omnibus:	326.264	**Durbin-Watson:**	0.717
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6588.792
Skew:	2.763	**Prob(JB):**	0.00
Kurtosis:	20.656	**Cond. No.**	9.07e+04

Regression removing STARMeter as a predictor

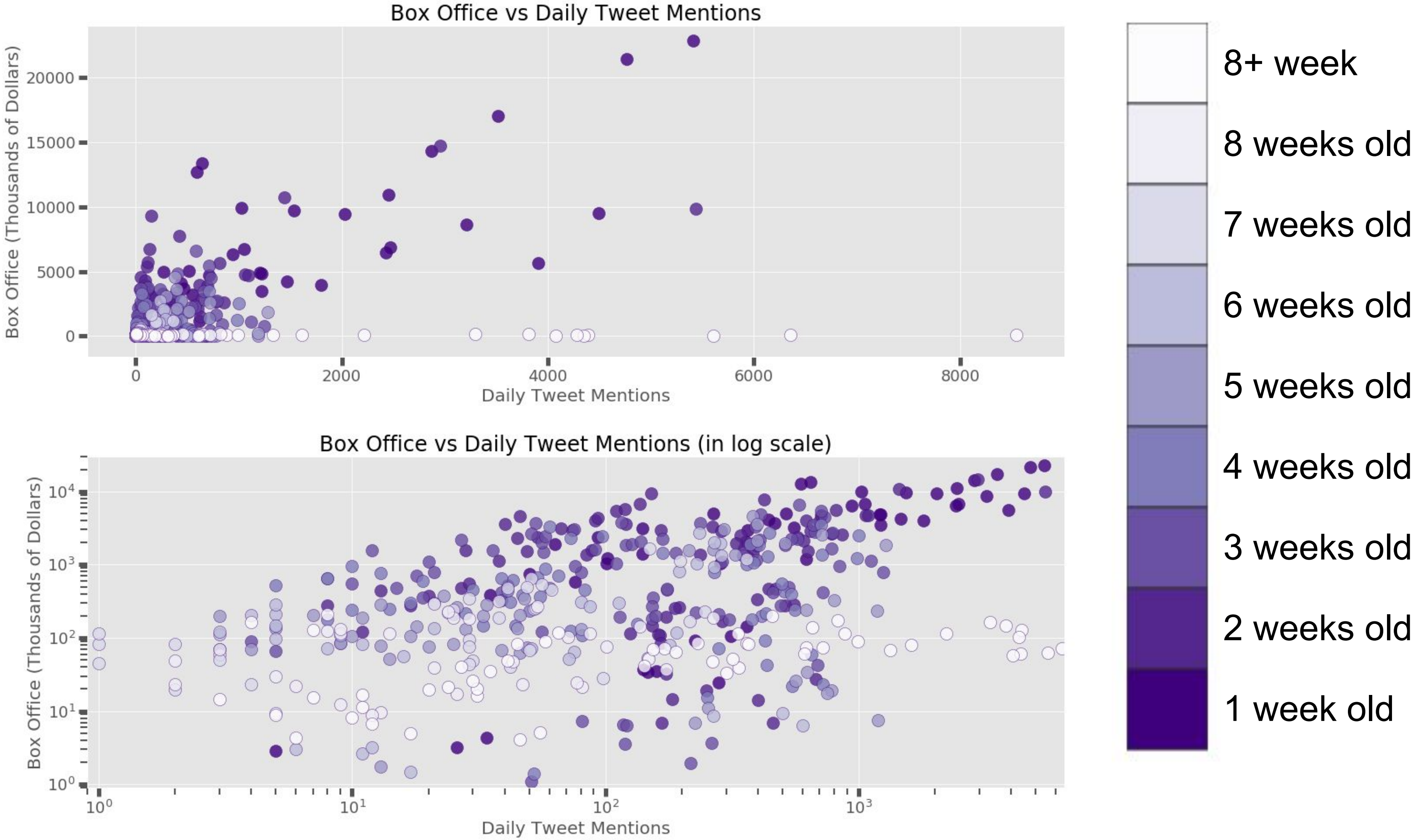


OLS Regression Results

Dep. Variable:	box office	**R-squared:**	0.640
Model:	OLS	**Adj. R-squared:**	0.637
Method:	Least Squares	**F-statistic:**	203.3
Date:	Mon, 06 Aug 2018	**Prob (F-statistic):**	5.46e-100
Time:	22:03:53	**Log-Likelihood:**	-4072.6
No. Observations:	462	**AIC:**	8155.
Df Residuals:	457	**BIC:**	8176.
Df Model:	4		
Covariance Type:	nonrobust		

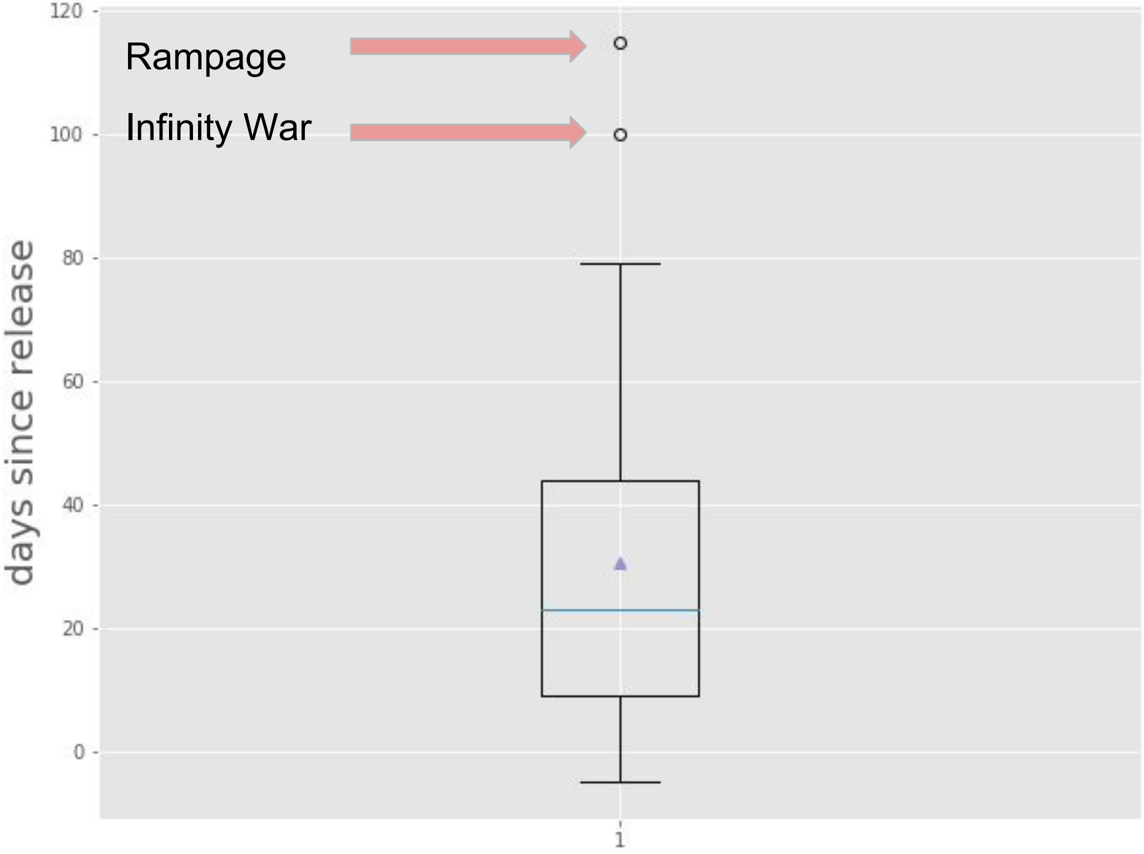
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-499.7857	217.094	-2.302	0.022	-926.412	-73.160
daily tweet count	1.0365	0.086	12.074	0.000	0.868	1.205
days since release	-16.7784	3.943	-4.255	0.000	-24.528	-9.029
weekend	879.6515	153.977	5.713	0.000	577.060	1182.242
theaters	1.0513	0.067	15.733	0.000	0.920	1.183

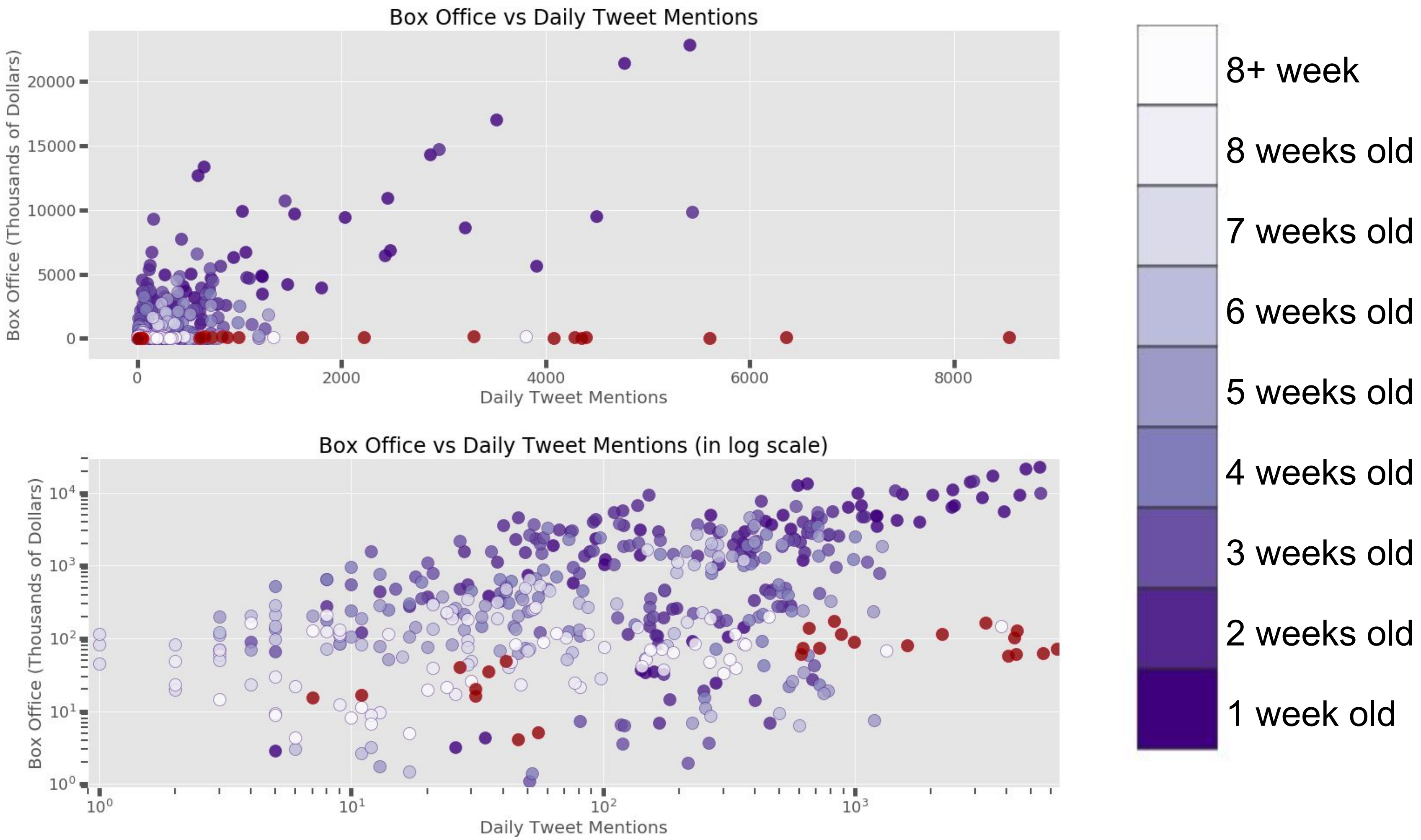
Omnibus:	327.007	**Durbin-Watson:**	0.717
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6630.653
Skew:	2.770	**Prob(JB):**	0.00
Kurtosis:	20.713	**Cond. No.**	6.30e+03


Box Office vs Daily Tweet Mentions

Box Office vs Daily Tweet Mentions (in log scale)

Days since release as of August 5th.





Box Office vs Daily Tweet Mentions

Box Office vs Daily Tweet Mentions (in log scale)

Optimizing Regression

These two outliers have been in theater for an unusually long amount of time because their box office overall performed exceptionally well and received ample amounts of attention.

Additionally, both films are based on already established franchises, so it's likely that the fanbase continue to mention the movie’s hashtag in tweets in discussions not directly related to seeing the movie itself. (e.g. Cosplay or Marvel giveaway.)

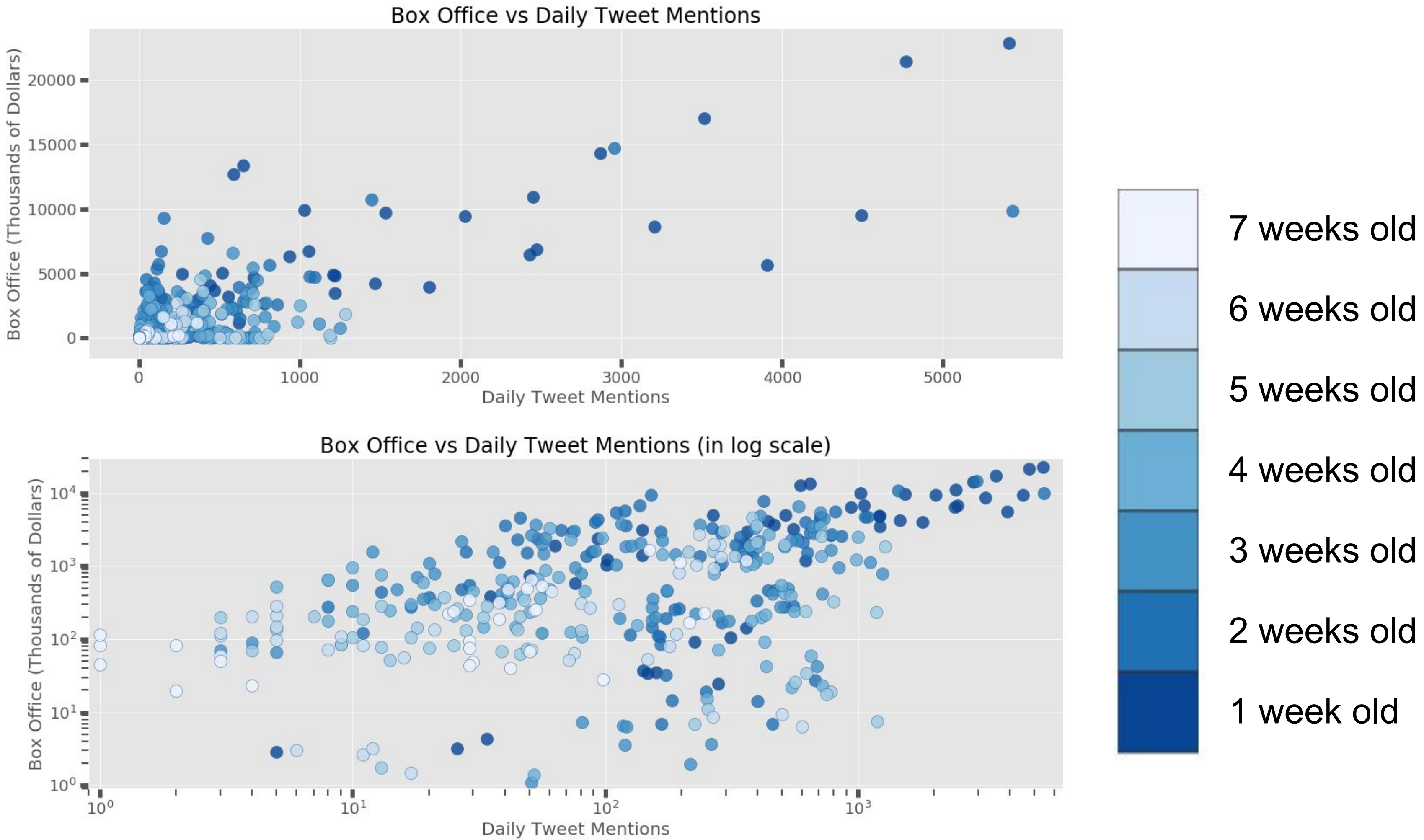
Thus the linear relationship between tweet mention and box office for these outliers are relatively weaker.

Optimizing Regression

R-squared for:

- All released films = 0.640
- all released films excluding Rampage & Infinity War = 0.743
- Released films less than 7 weeks old = 0.761
- Released films less than 6 weeks old = 0.763
- Released films less than 5 weeks old = 0.762
- Released films less than 4 weeks old = 0.769

No significant improvement in R-squared value beyond week 7. Subsequent linear regression analyses use data from film less than 7 weeks old



Box Office vs Daily Tweet Mentions

Box Office vs Daily Tweet Mentions (in log scale)

Assumptions of Multiple Regression

- Linear relationship between IV and DV
  - Residuals of the regression (errors between observed and predicted values) are normally distributed, i.e. independence of errors

- Homoscedasticity

- Normality

- No multicollinearity between IV

Multicollinearity

Tested using Correlation Matrix of IV. Determinant is closer to 1 if collinearity is low, closer to 0 if IV are collinear.

Determinant of correlation matrix including STARMeter = 0.425

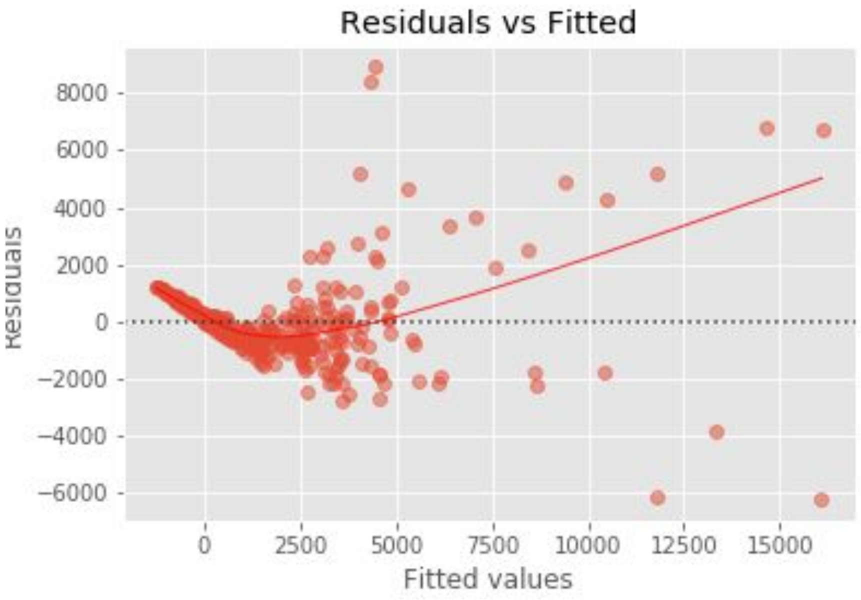
Determinant of correlation matrix excluding STARMeter = 0.658

Homoscedasticity & Normality of Residuals

Non-random distribution of residuals

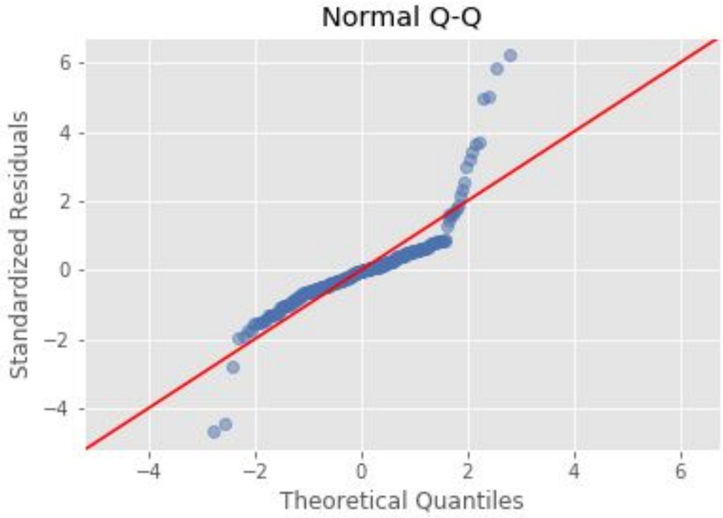
“Trumpet” shaped residual distribution implies heteroscedasticity

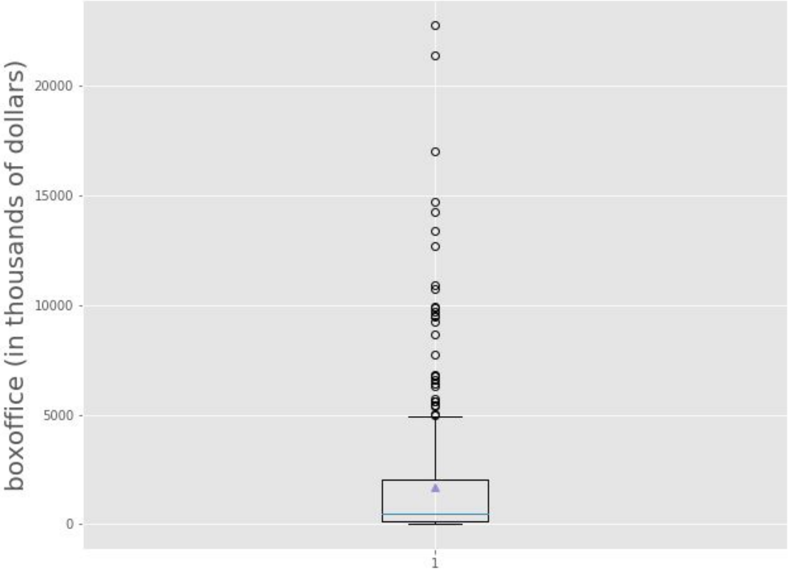
Curving trend suggests non-linear relationship between IV and DV



Normality

Upwards curving implies a long right tail skew in box office distribution





Multiple Regression Conclusion

- Multiple regression is a decent starting point in evaluating the relationship between various predictors and box office.
- IMDb's STARMeter is surprisingly irrelevant to box office. Indie films can do well too.
- Films that have been out for a while are harder to predict
- Our data set violated multiple assumptions of regression, the coefficients from the regression may not be able to predict box office with high accuracy.
- Overall, R-squared value of 0.76 is not bad for a simplistic model with coarse, publicly available data

Random Forest Regression

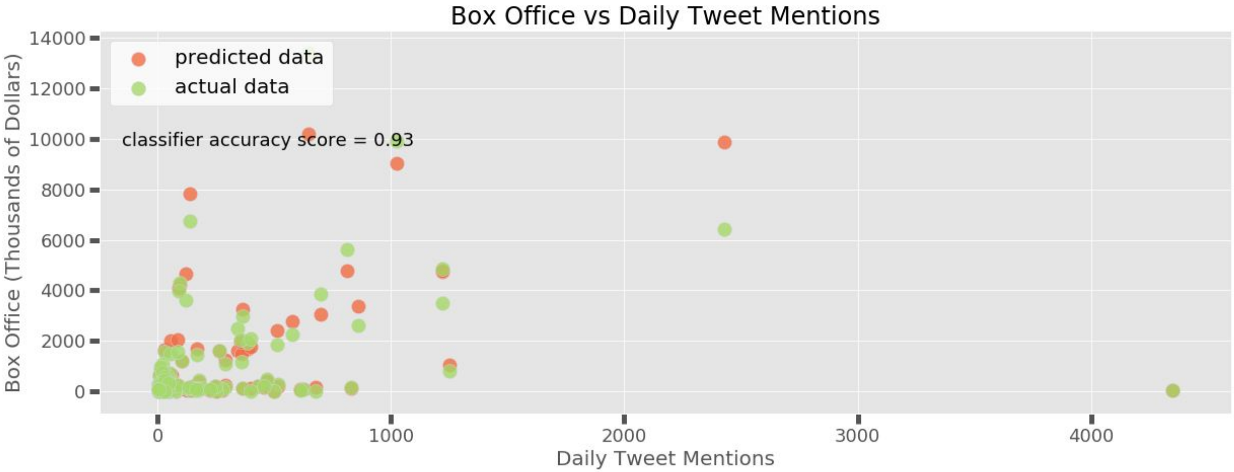
Advantages:

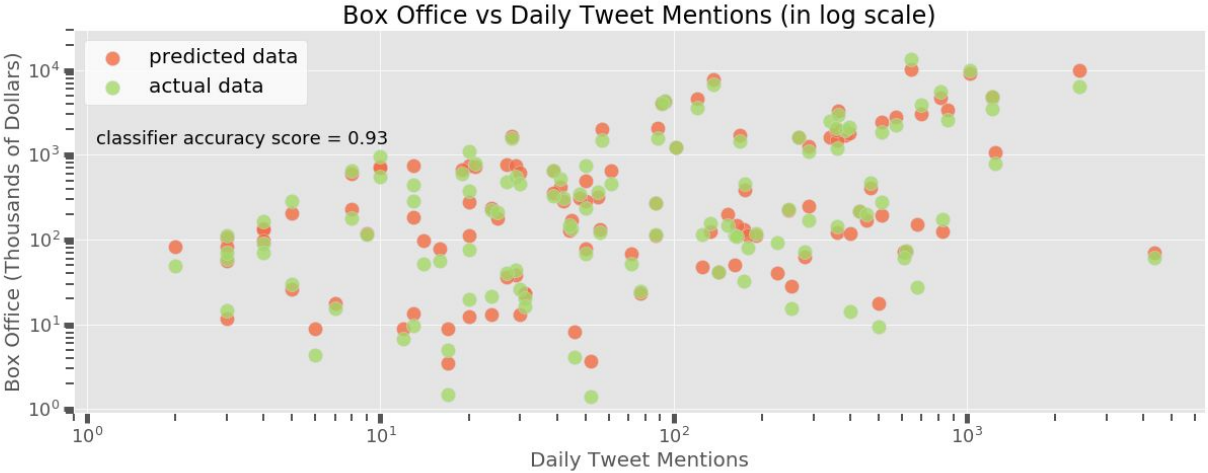
- Does not require linear relationship between IV and DV
- More relaxed with assumptions compared to multiple linear regression
- Outliers would not significantly skew model

Included all released films.

75% data were used to Train RF classifier.

The classifier made predictions about the remaining 25% data with an accuracy score of 0.93





K-Means Clustering

- Dividing the data set into “clusters” based on their multidimensional euclidean distance between each other.

- Data points in the same cluster are more similar to each other.

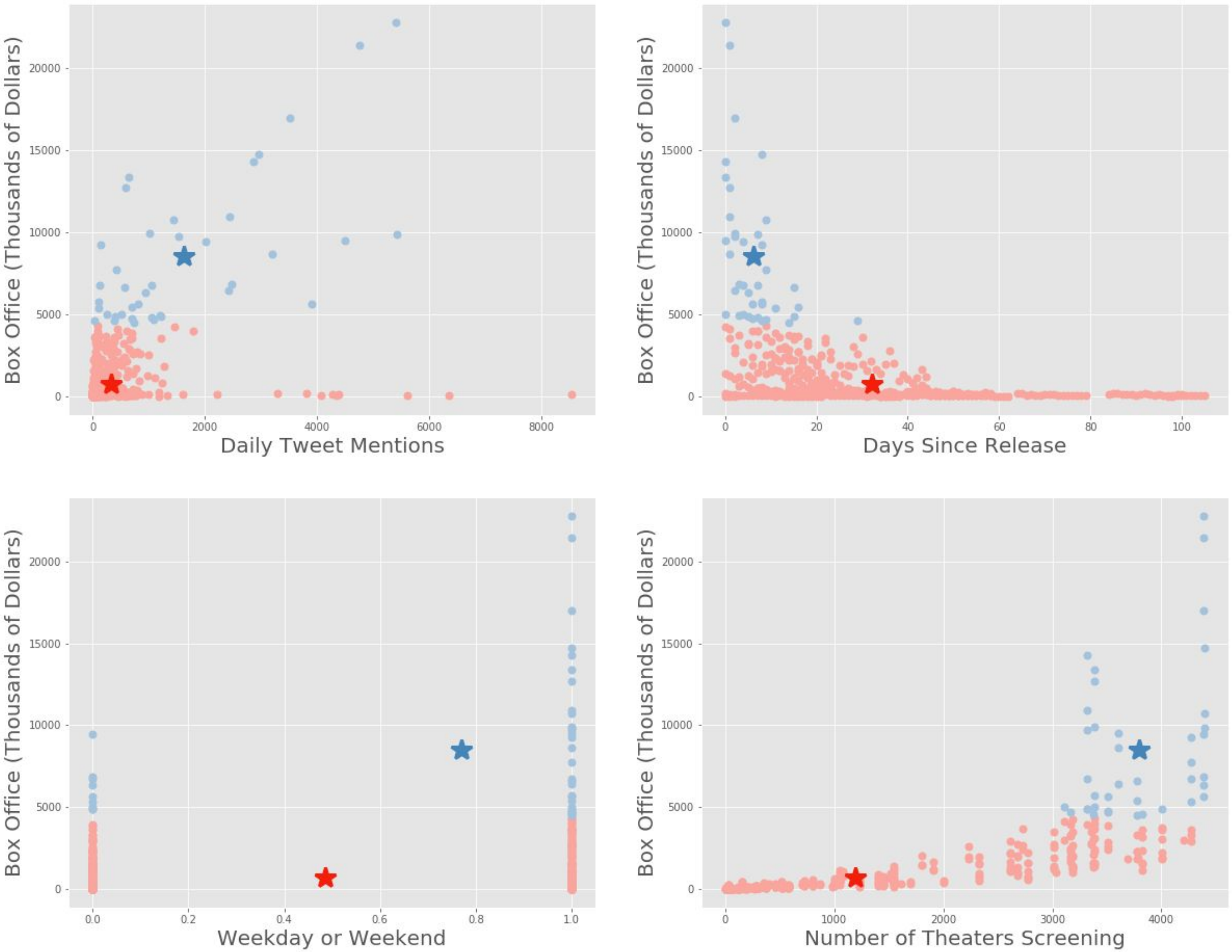
- RF might perform better for some of the clusters

2 clusters

Accuracy score:

0.934

0.733



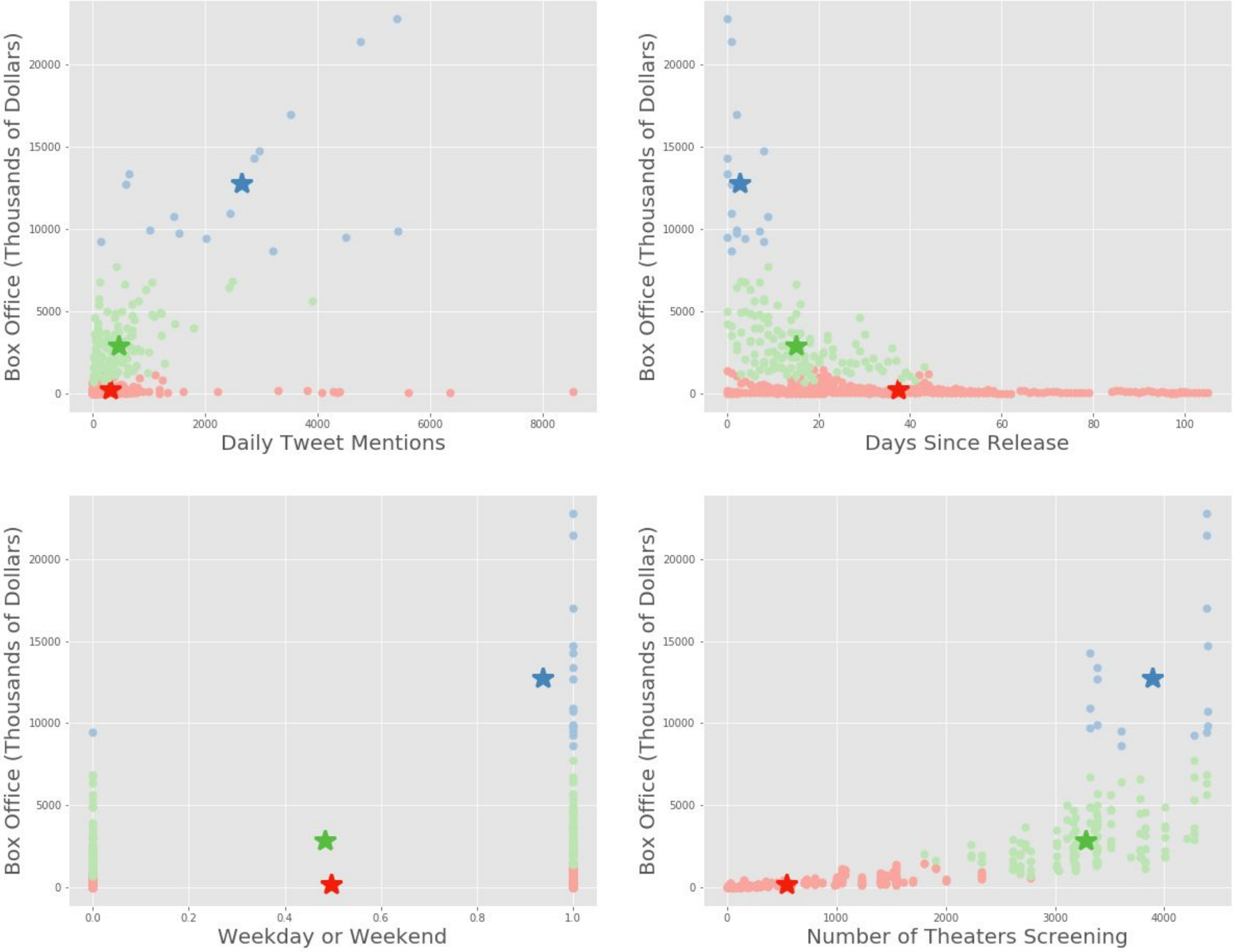
3 clusters

Accuracy score:

0.949

-1.165

0.789



4 clusters

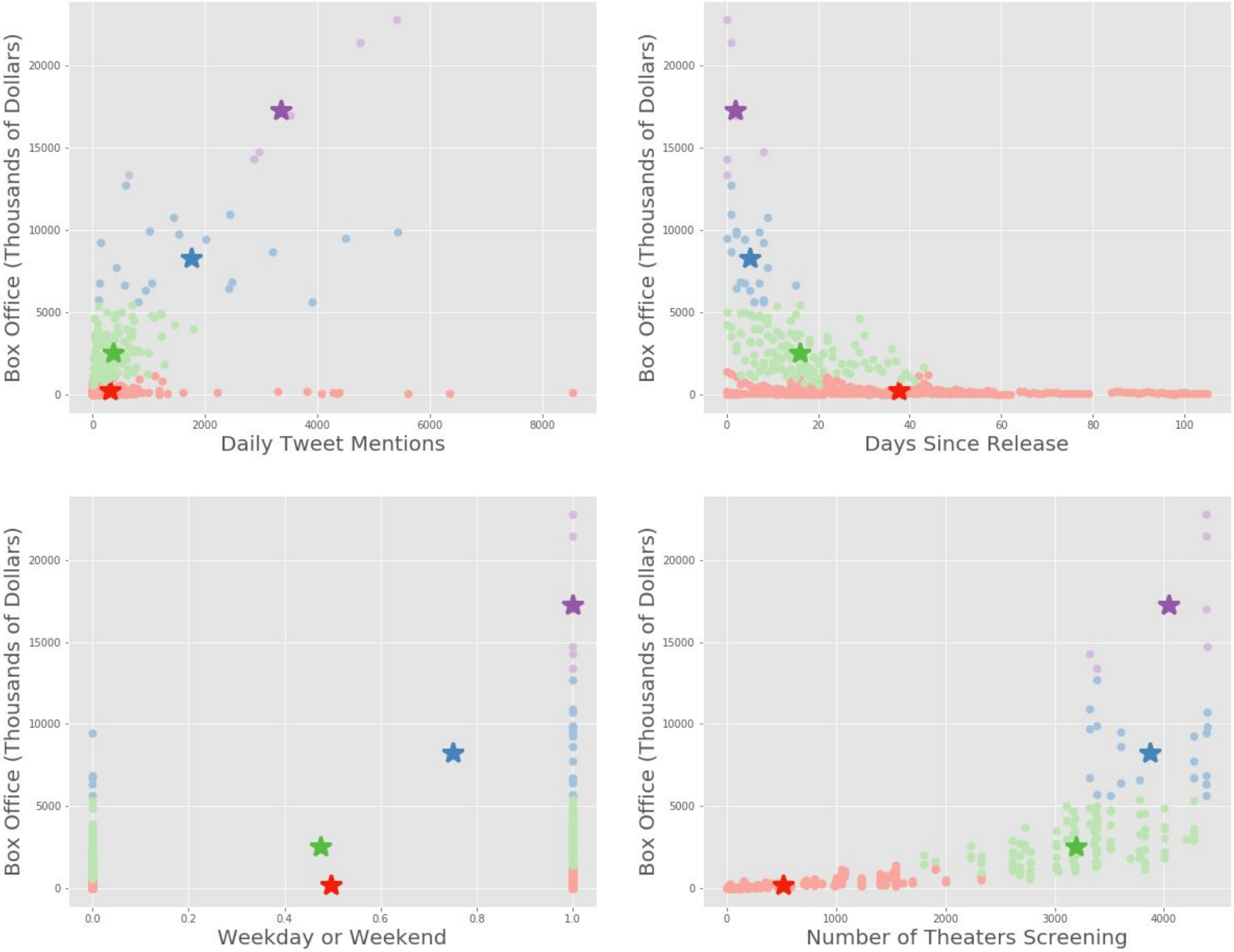
Accuracy score:

0.983

0.472

0.828

0.406



Conclusion

- Random Forest regression is highly accurate in predicting box office performances.

- Overall, K-means clustering did not boost the performance of the regression. One possible reason is that subdividing data sets reduces the amount of training the classifier receives. Other being there is no intrinsic cluster in our data.

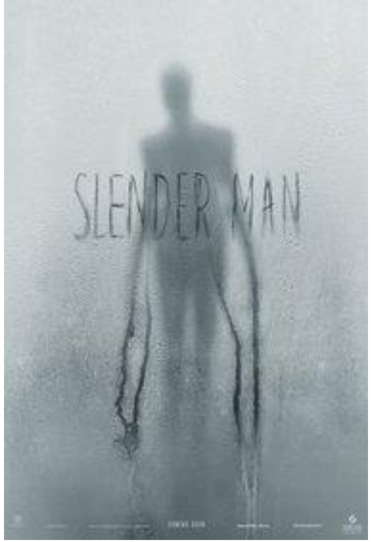
- Our simple, minimalistic model using publicly available data was able to predict box office with a fair degree of success

Our Crystal Ball for Friday, August 10th

Caveat: we don’t know how many theaters will be screening these films!

The Meg

$ 4000 per theater



Slender Man

$ 7500 per theater

BLACKKKLANSMAN
FROM PRODUCER JORDAN PEELE
BASED ON A CRAZY, OUTRAGEOUS, INCREDIBLE TRUE STORY!
AUGUST 10

BlacKkKlansman

$ 2800 per theater