

Movie Analysis

how could Microsoft enter the movie industry successfully?



key focus on financial variables



is there a variable that drives return on investment (ROI)?



which variable controls movie gross?

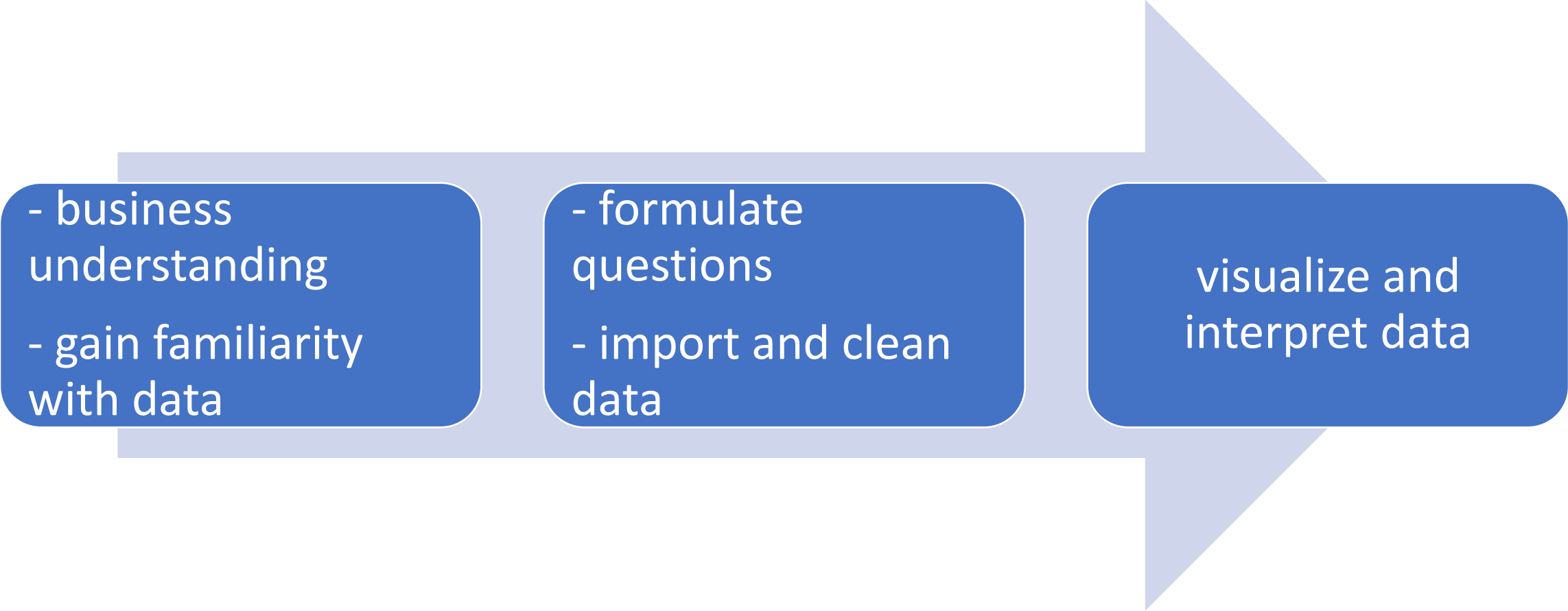


does a movie return more in domestic or foreign sales?



does a movies rating effect the rating assigned by critics?

methods



- business understanding
- gain familiarity with data

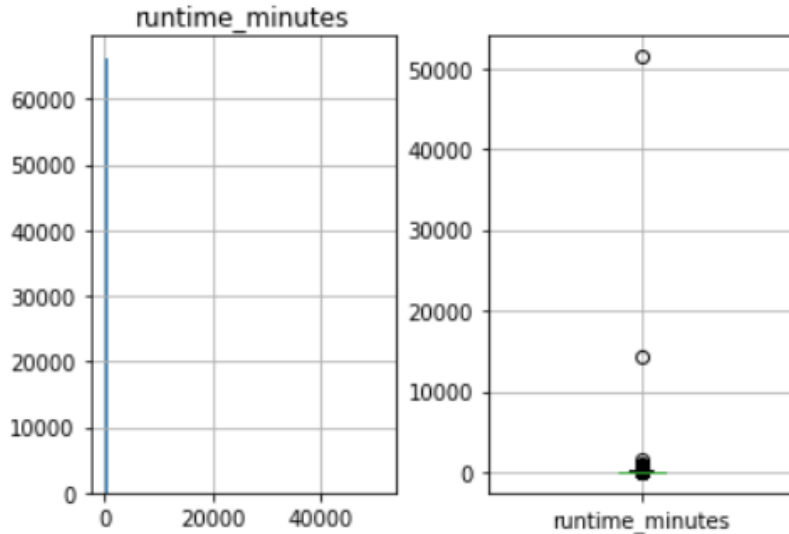
- formulate questions
- import and clean data

visualize and
interpret data

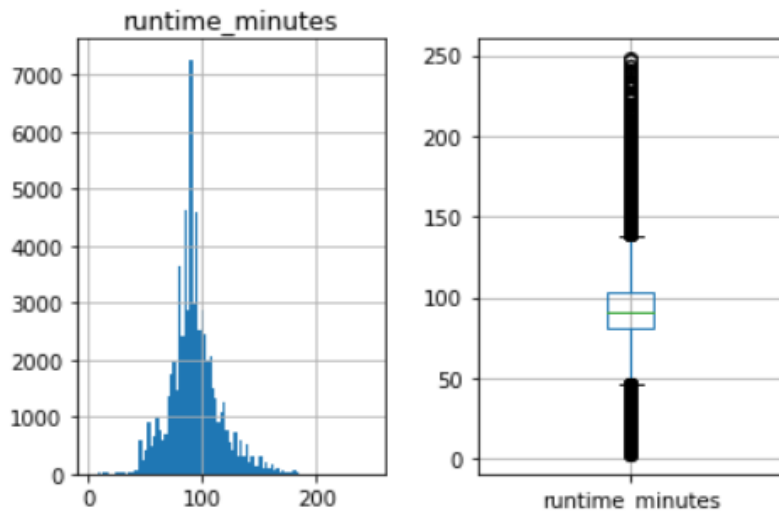
data cleaning

remove outliers, format data types, and remove missing values

before



after

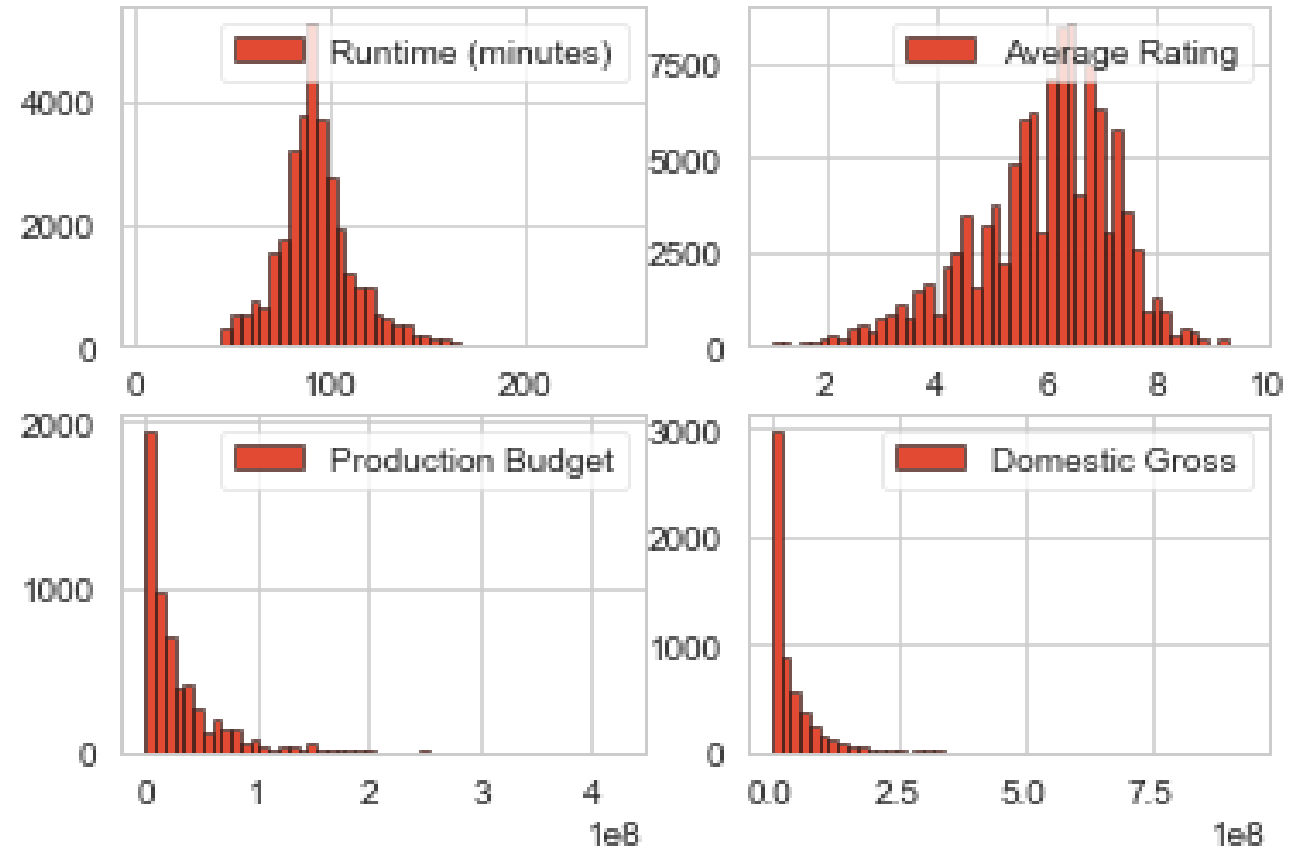
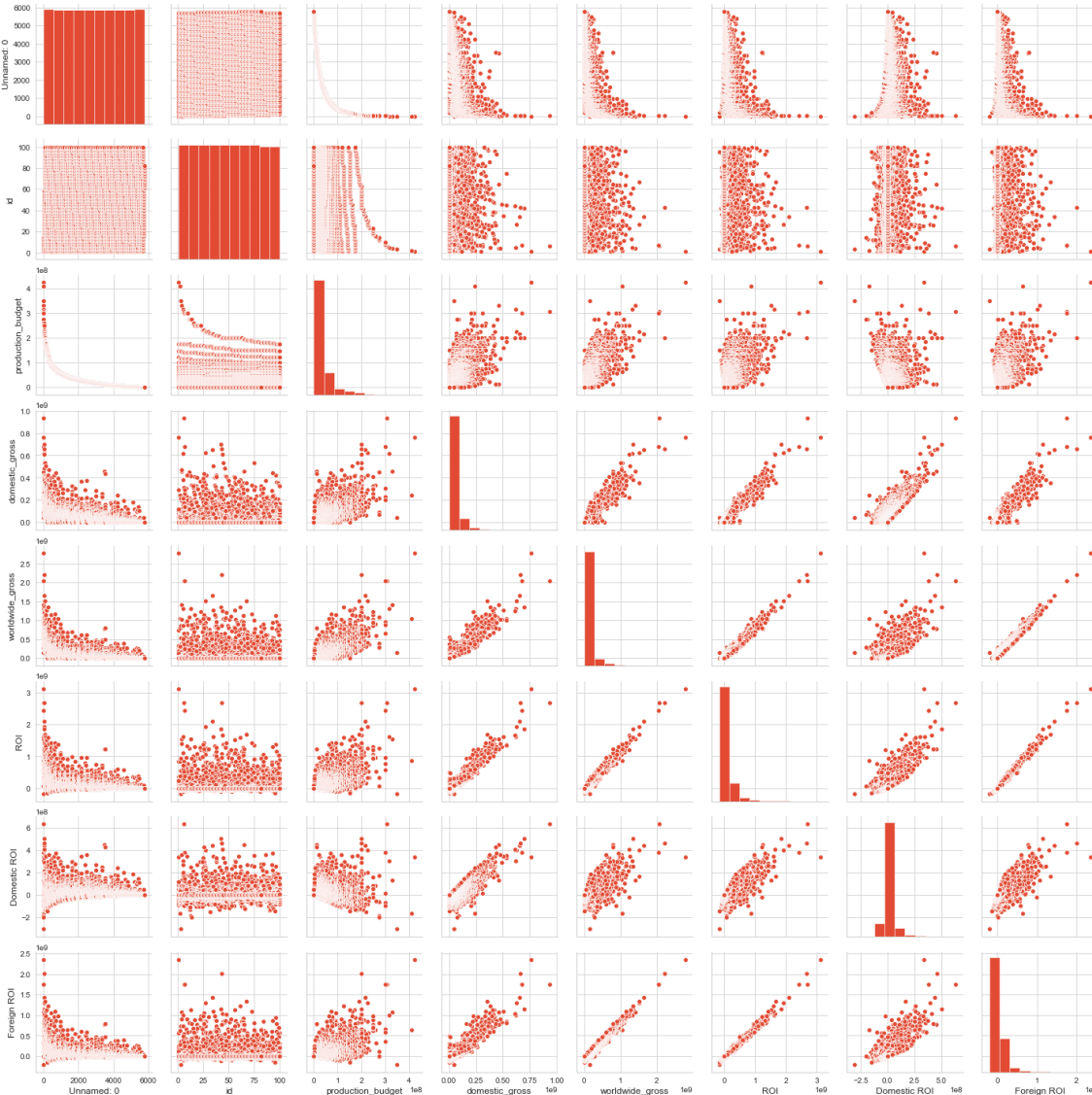


- removing outliers
 - Box office – filter using IQR
 - Runtime – using visualizations
- removing missing values
 - Drop null or NaN values
 - Drop columns with disproportionate amount of missing values
- formatting data types
 - Change data types to format such as int64 so analysis can be applied

data exploration

leveraging summary statistics to quickly find meaningful relationships

The Numbers Data Set



- python methods such as `.corr()` , `.describe()`, and visualizations like pairplots, histograms, and linear regression used to narrow in on relationships within the data

feature engineering

column addition, concatenation, and dummy variables

```
In [17]: Movie_Budgets['Domestic ROI'] = Movie_Budgets['domestic_gross'] - Movie_Budgets['production_budget']
```

```
In [18]: Movie_Budgets['Foreign ROI'] = Movie_Budgets['worldwide_gross'] - Movie_Budgets['production_budget']
```

```
In [19]: Movie_Budgets.head()
```

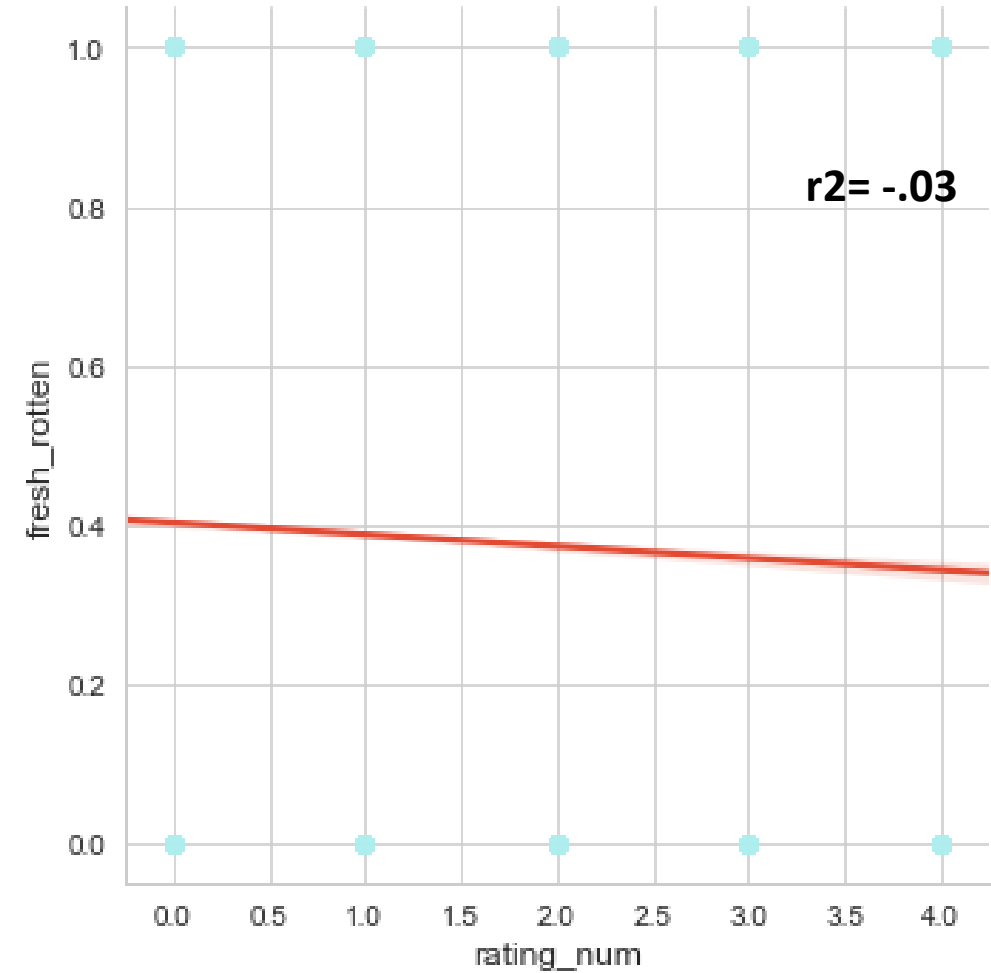
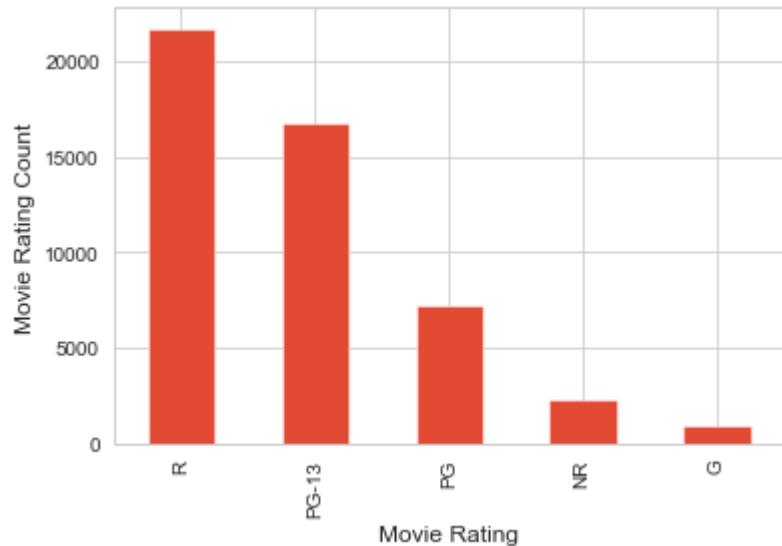
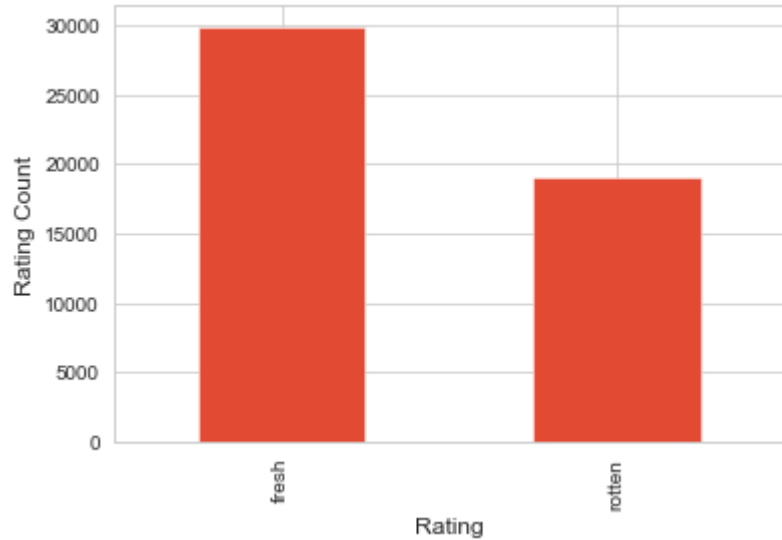
Out[19]:

	Unnamed: 0	id	release_date	movie	production_budget	domestic_gross	worldwide_gross	ROI	Domestic ROI	Foreign ROI
0	0	1	18-Dec-09	Avatar	425000000	760507625	2776345279	3111852904	335507625	2351345279
1	1	2	20-May-11	Pirates of the Caribbean: On Stranger Tides	410600000	241063875	1045663875	876127750	-169536125	635063875
2	2	3	7-Jun-19	Dark Phoenix	350000000	42762350	149762350	-157475300	-307237650	-200237650
3	3	4	1-May-15	Avengers: Age of Ultron	330600000	459005868	1403013963	1531419831	128405868	1072413963
4	4	5	15-Dec-17	Star Wars Ep. VIII: The Last Jedi	317000000	620181382	1316721747	1619903129	303181382	999721747

- column addition used to add and subtract variables
 - ROI
 - Domestic ROI
 - Foreign ROI
- dummy variables used to create numerical data from categorical data
 - rating 'fresh' or 'rotten'
 - movie rating
 - R, PG-13, PG, G, NR

should a movie target a specific rating?

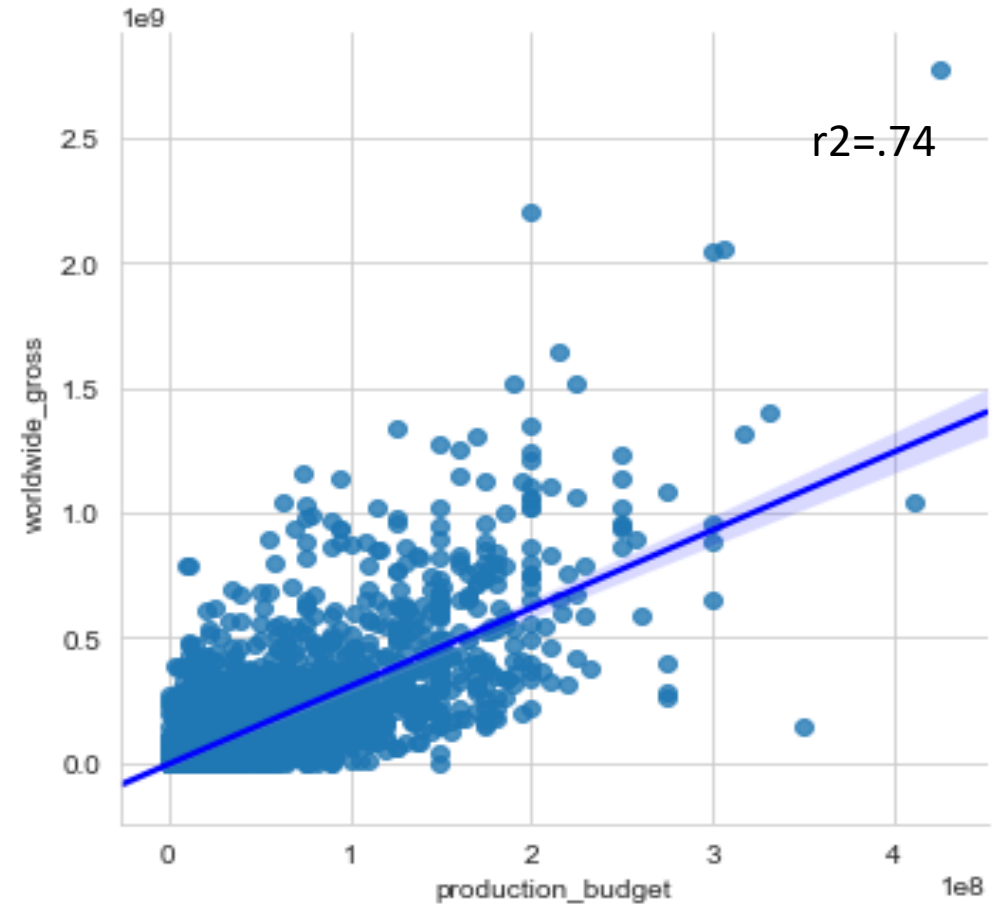
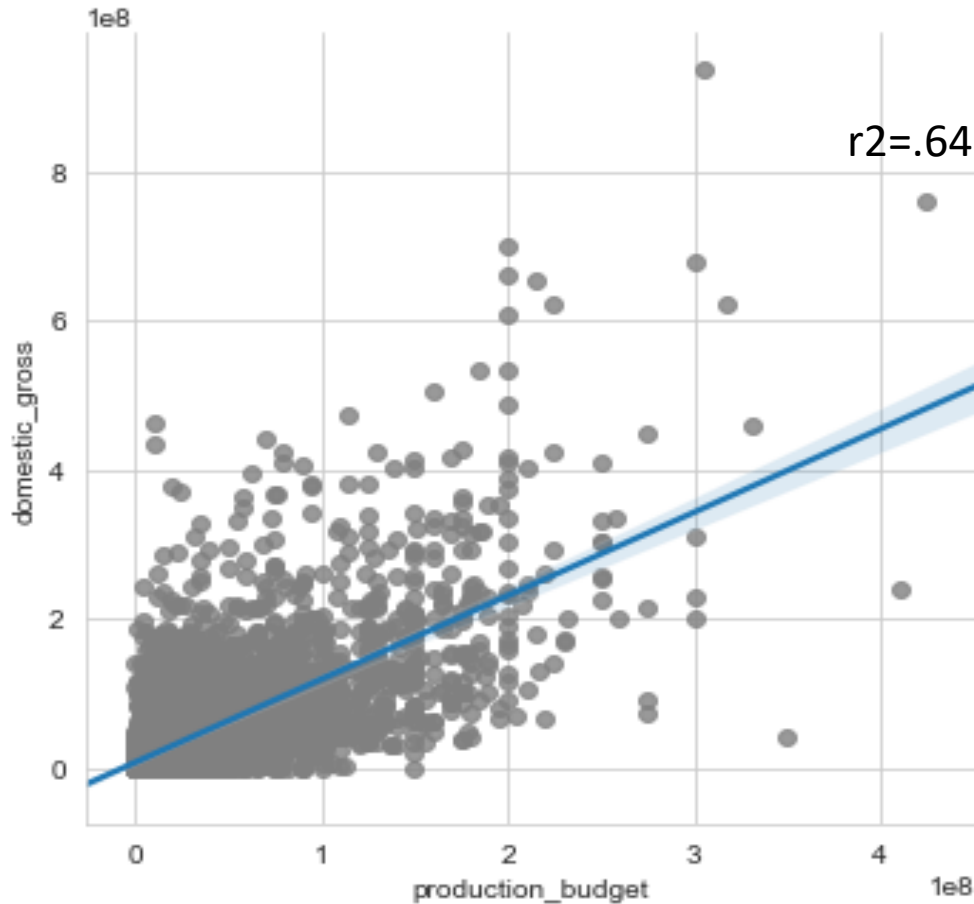
most movies are rated “R”, but rating does not matter



- there are a disproportionate amount of “R” rated movies, but movie rating has no correlation to how critics will review it.

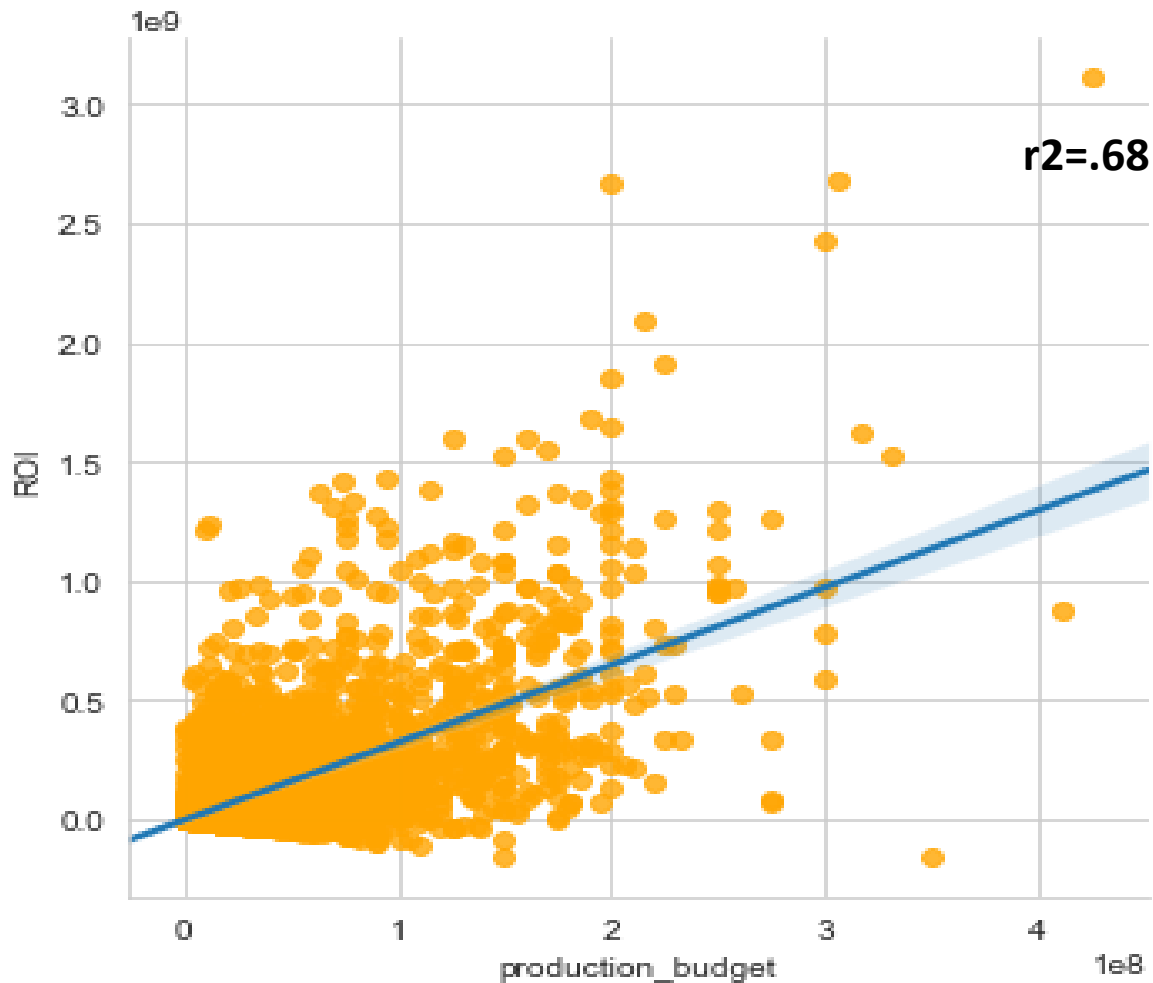
production budget and movie gross

production budget drives both foreign and domestic gross

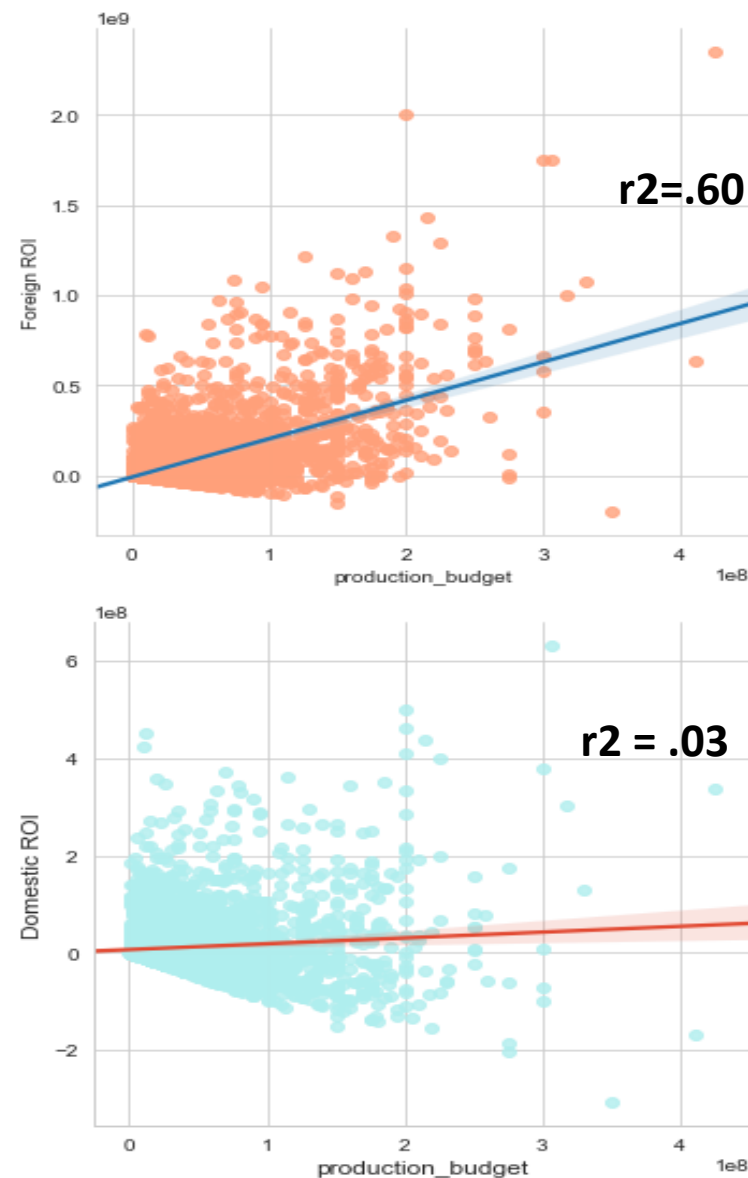


- production budget has a positive correlation with gross domestic and foreign

you have to spend money to make money
but where you release your movie matters



- positive correlation between production budget and ROI
but targeting large geographic area matters



recommendations and next steps

- key variable is production budget
 - higher ROI and movie gross
 - potentially correlates to higher end actors, film locations, and editing
- make sure the movie is released in foreign markets
 - Larger audience generates higher ROI than focusing on domestic release
- do not focus on a particular rating for the movie
 - No correlation between movie rating and how well critics rated the movie
- Next steps
 - Scrape more data
 - Analyze correlation between ROI and additional variables (genre, actor, director)