Movies

Team Members

Isaac Oshana, Kratika Sharma, Steven Wang

Dataset - Movies Metadata

This dataset contains a list of 45,000 + movies released on or before July 2017. The dataset has information regarding the movie's budget, revenue, rating, runtime, genre, producers and any other important information regarding the movie. There is also consumer information such as rating, popularity, estimated views and critic votes.

We will be looking at the financial performance of certain movies. To expand on that, we will see if there is any relationship between revenue and spending and their ratings. This will be used to evaluate how good the movies performed. Also, we will find out the correlation between variables, decision tree, choropleth graph of countries, clustering data and regression.

Data Cleaning

Original Data:

```
df.shape
```

```
(45466, 24)
```

	adult	belongs_to_collection	budget	genres	homepage	id	imdb_id	original_language	original_title	overview	...	release_date	revenue	runtime	spoken_languages	status	tagline	title	video	vote_average	vote_count
0	FALSE	{'id': 10194, 'name': 'Toy Story Collection', ...	30000000	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...	http://toystory.disney.com/toy-story	862	tt0114709	en	Toy Story	Led by Woody, Andy's toys live happily in his ...	...	10/30/1995	373554033.0	81.0	[{'iso_639_1': 'en', 'name': 'English'}]	Released	NaN	Toy Story	False	7.7	5415.0

- Special characters (!'@:,-#)
- NaN cells
- Multiple items for cell.
  - ex. Multiple production companies.
  - Took first item in the list.
- Repeat columns (ex. Original Language and Current Language)
- Irrelevant/Meaningless columns - Id, imbd_id, collection
- Non-Data related Columns - Homepage, Description, Overview

Cleaned File

Cleaned DF:

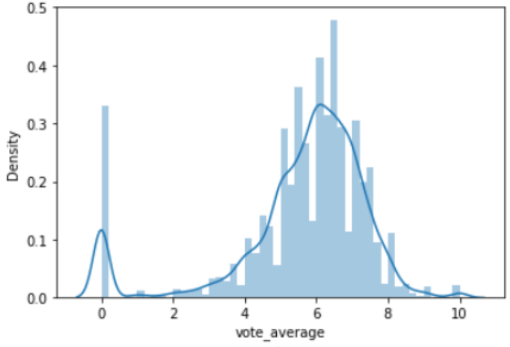
- Took first item from these Column:
  - First production company listed
  - First production country listed
  - First Language listed
- Removed irrelevant columns
- Special Characters removed
- Formatted floats, ints, NaN, etc. so could be used for data analysis

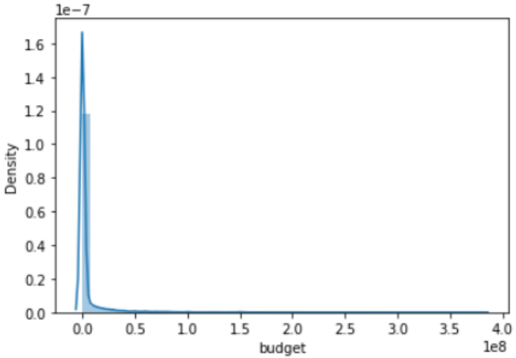
	adult	budget	original_language	original_title	popularity	release_date	revenue	runtime	status	title	vote_average	vote_count	production_country	production_company
0	False	30000000	en	Toy Story	21.946943	10/30/1995	373554033.0	81.0	Released	Toy Story	7.7	5415.0	'United States of America'	'Pixar Animation Studios'
1	False	65000000	en	Jumanji	17.015539	12/15/1995	262797249.0	104.0	Released	Jumanji	6.9	2413.0	'United States of America'	'TriStar Pictures'
2	False	0	en	Grumpier Old Men	11.712900	12/22/1995	0.0	101.0	Released	Grumpier Old Men	6.5	92.0	'United States of America'	'Warner Bros.'
3	False	16000000	en	Waiting to Exhale	3.859495	12/22/1995	81452156.0	127.0	Released	Waiting to Exhale	6.1	34.0	'United States of America'	'Twentieth Century Fox Film Corporation'
4	False	0	en	Father of the Bride Part II	8.387519	2/10/1995	76578911.0	106.0	Released	Father of the Bride Part II	5.7	173.0	'United States of America'	'Sandollar Productions'

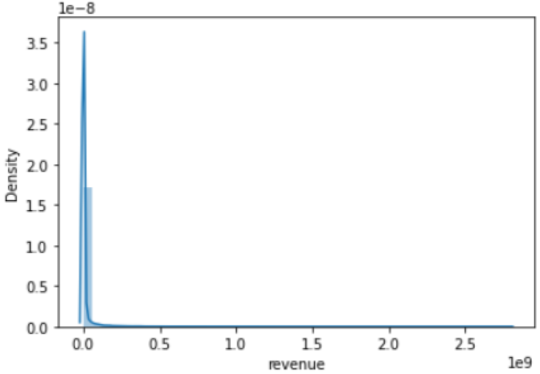
Main Points of Focus

- Financial Performance
    - Budgets, Revenue, Profits
    - Weighted Values
- Infrastructure
    - Production Companies
    - Production Countries
- Public Review
    - Viewerships
    - Reception

Graphical representation







Half the movies have a rating of less than or equal to 6

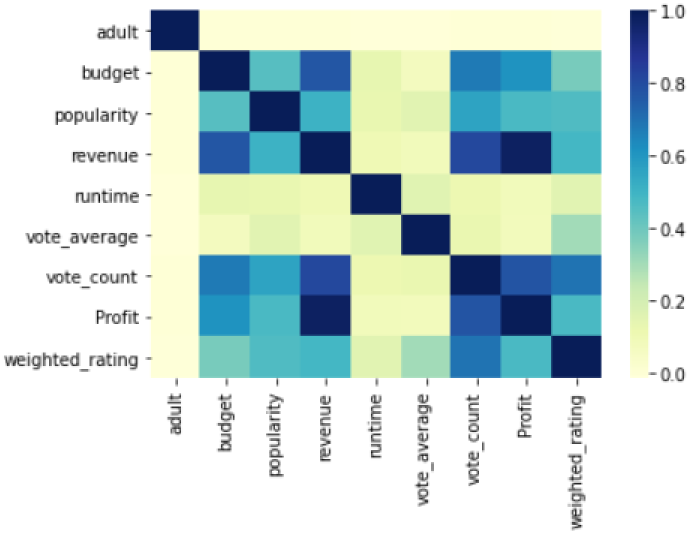
More than 75% of the movies have a budget smaller than 25 million dollar

Revenue is also decreasing as budget

Correlation

Below is the correlation heatmap:

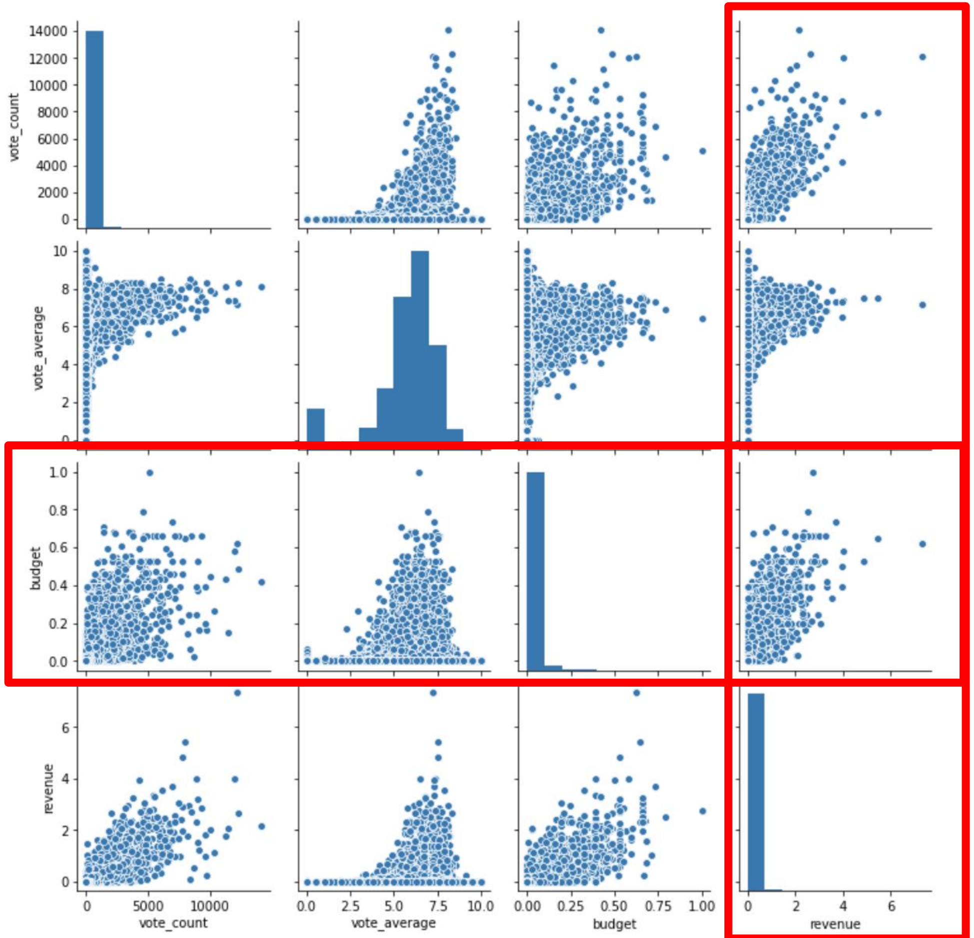
We can see that profit and revenue are highly correlated




```
cor[cor < 1].stack().nlargest(20)[::2]
```

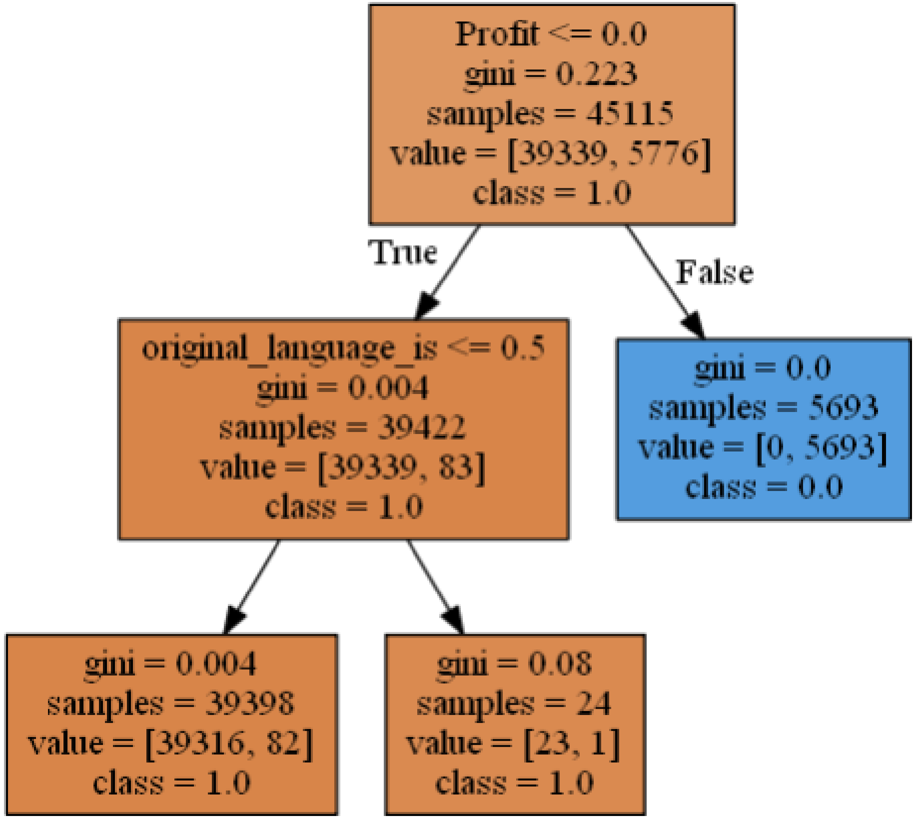
```
revenue      Profit             0.976896
             vote_count         0.812022
vote_count   Profit             0.775756
budget       revenue            0.768776
vote_count   weighted_rating    0.694526
budget       vote_count         0.676642
             Profit             0.614339
popularity   vote_count         0.559965
             revenue            0.506179
revenue      weighted_rating    0.489236
dtype: float64
```

Pairplot between vote count, vote average, budget and revenue



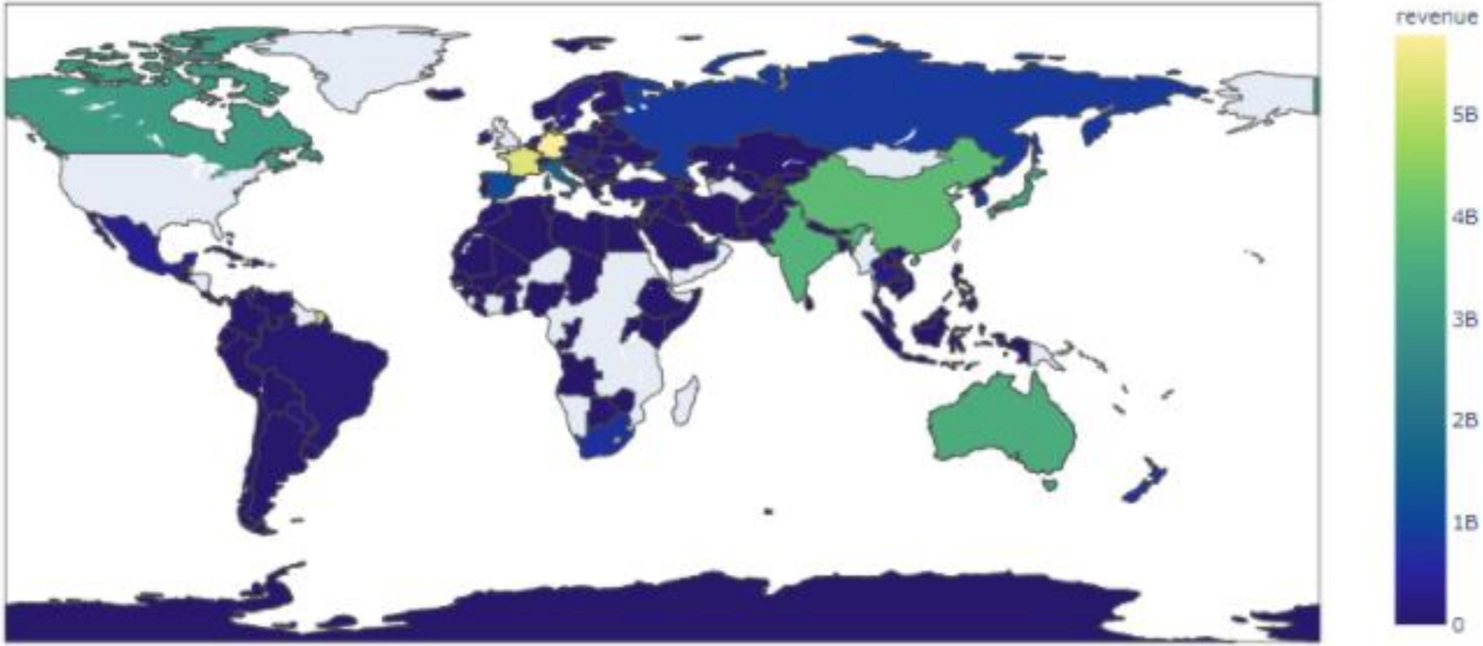
Machine Learning - Decision Tree

- We have made our profit column binary and defined dummy variable for all the categorical columns except profit.

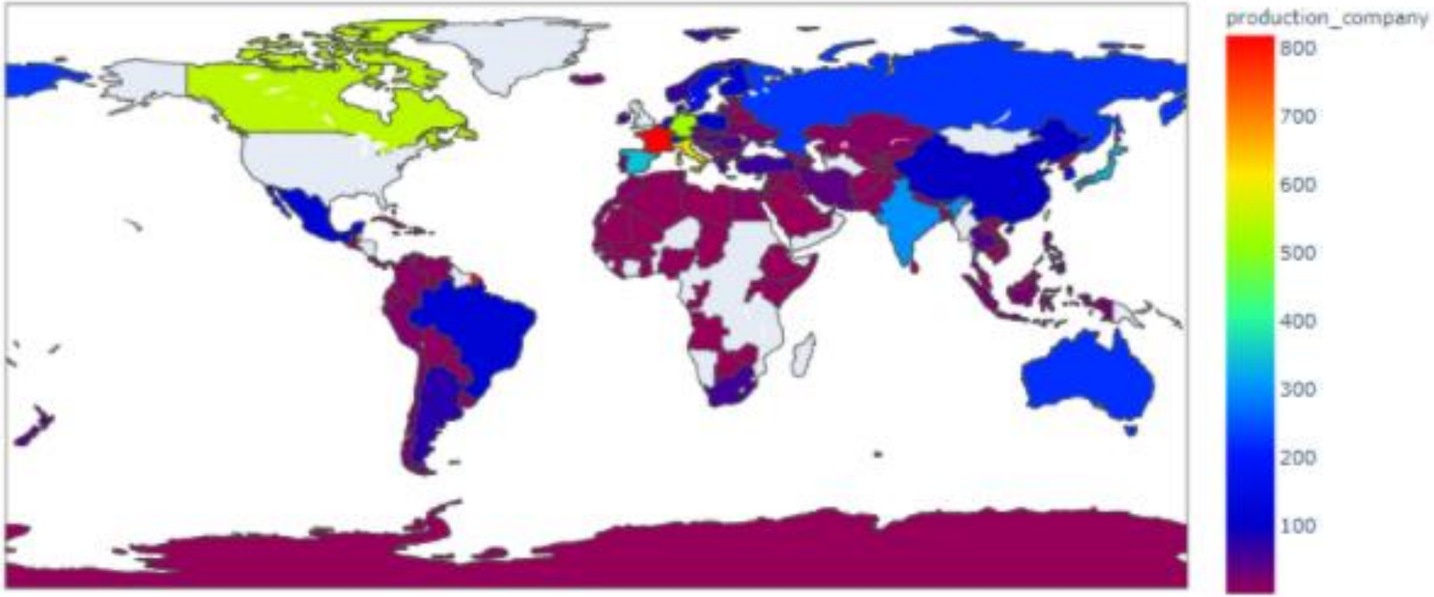


Visualization using Choropleth

Revenue by Country (Excludes USA & UK)



Number of Major Production Companies (Choropleth)



Popular Movies finding based on their rating (calculated IMDB rating style)

Formula Used:

Weighted Rating:

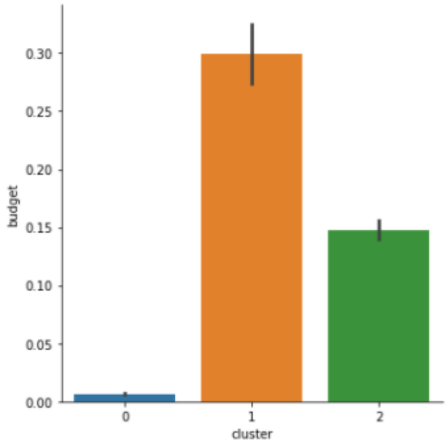
(vote_count / (vote_count + minvotes) * vote_average) + (minvotes / (minvotes + vote_count) * meanrating)

title	production_country	genres	Profit	popularity	weighted_rating
The Shawshank Redemption	USA	Drama	0.008793	51.645403	8.357762
The Godfather	USA	Drama	0.629122	41.109264	8.306364
The Dark Knight	USA	Drama	2.156733	123.167259	8.208383
Fight Club	USA	Drama	0.099615	63.869599	8.184911
Pulp Fiction	USA	Thriller	0.541918	140.950236	8.172169
Forrest Gump	USA	Comedy	1.639330	48.307194	8.069436
Schindler's List	USA	Drama	0.787804	41.725123	8.061056
Whiplash	USA	Drama	0.025768	64.299990	8.058076
Spirited Away	Japan	Fantasy	0.684013	41.048867	8.035658
The Empire Strikes Back	USA	Adventure	1.369474	19.470959	8.025820

title	production_country	genres	Profit	popularity	weighted_rating
Avatar	United Kingdom	Action	6.713066	185.070892	7.145295
Star Wars: The Force Awakens	USA	Action	4.797957	31.626013	7.403097
Titanic	USA	Drama	4.329037	26.889070	7.400463
Jurassic World	USA	Action	3.588234	32.790475	6.458748
Furious 7	USA	Action	3.463814	27.275687	7.144305
The Avengers	USA	Science	3.419889	89.887648	7.337813
Harry Potter and the Deathly Hallows: Part 2	USA	Family	3.202632	24.990737	7.749407
Avengers: Age of Ultron	USA	Action	2.961589	37.379420	7.200599
Frozen	USA	Animation	2.958471	24.248243	7.175762
Beauty and the Beast	USA	Family	2.902332	287.253654	6.714015

Clustering

We have divided it into 3 clusters. K means clustering for K=3, we can see that one cluster is the largest.



	budget	popularity	revenue	runtime	vote_average	vote_count	Profit	weighted_rating	Profit1	adult_False	...	original_language_zh	orig
cluster													
0	0.006451	2.510285	0.010748	93.557052	5.599872	45.558489	0.004297	5.661627	0.103880	0.999795	...	0.009311	
1	0.299835	36.515962	1.538087	127.492754	7.291787	5624.811594	1.238251	7.165625	1.000000	1.000000	...	0.000000	
2	0.147628	13.947426	0.506999	112.851103	6.637592	1685.854779	0.359371	6.413647	0.934743	1.000000	...	0.000919	

Regression - Measuring Predictive Performance

MAD Score	0.02195
MSE Score	0.0067
Prediction score	0.99
AUC Score	0.99

Conclusion

- Revenue and Budgets have a high positive correlation. Those two factors are some of the best determinants of profits.
- Vote counts and public response also determine how well movies perform financially.
- Countries with a strong movie infrastructure (Major production companies) tend to generate the most revenue.
    - There are outliers such as China, India and Australia that generate high revenue, but don’t have many production companies.
    - Some Countries that have many production companies don’t generate high revenue (South America, Northern Europe)
- Clustering showed that the movies with small runtime, revenue and budget performs less and generate less profit.
- Regression predictive analysis showed that our prediction score is 0.99 which is quite good for analysing the given data.

print(”Thank You”)