Investing in Film with Machine Learning

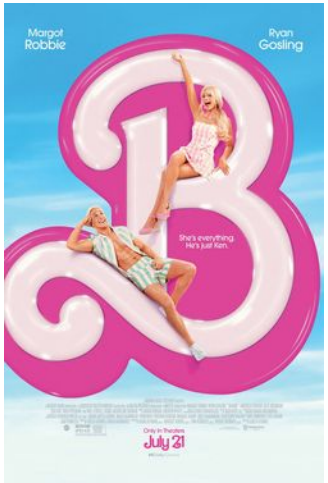
USC Marshall School of Business, Sep 2025

Investing in Film

Things we need to decide – “Valuation”

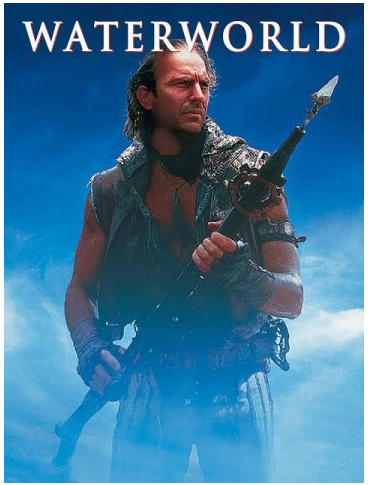
- How much to fund a movie
- How much to spend on marketing
- How to change our investment when we have a better understanding of its revenues

How much would you pay for these?



Barbie (2023)

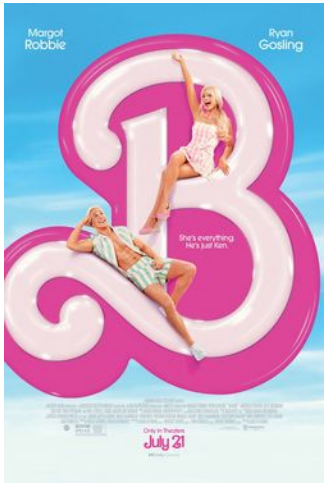
Box Office: $1.446B



Waterworld (1995)

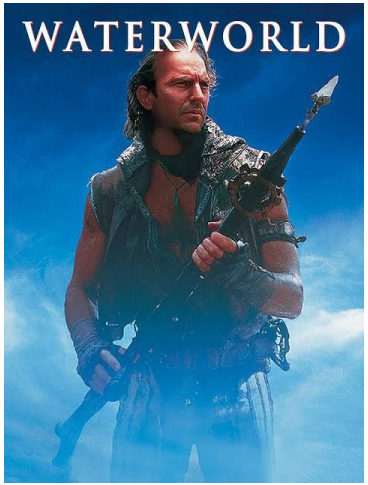
Box Office: $264M

How much would you pay for these?



Barbie (2023)

Box Office: $1.446B
Budget: $145M
Cost: $300M



Waterworld (1995)

Box Office: $264M
Budget: $175M
Cost: $235M

How does the film industry do this? - Comps

$?

Budget: $60M
Box Office: $469M

Toy-to-screen adaptation with heavy family appeal, successful cultural crossover.

Budget: $149M
Box Office: $822M

Female led franchise IP with broad appeal.

Budget: $160M
Box Office: $1.264B

Reimagining of a classic IP, heavily marketed, broad four-quadrant appeal.

How do we pick these “comps”?

Manually determined comps is a very subjective process that depends heavily on the individual’s knowledge and experience, mainly because it a domain that contains a wealth of available information

Production Elements

- Cast and Crew
- Genre
- Production budget
- Release timing

Market Factors

- Competition
- Platform strategy
- International appeal
- Merchandising potential

Start with averages

$852

$469M

$822M

$1.264B

ML can be used to

1. Help you pick a better set of comps that provide a more reliable box office prediction

2. More complex prediction than average

ML can be used to

1. **Help you pick a better set of comps that provide a more reliable box office prediction**

2. More complex prediction than average

Comps - K-Means Clustering

We have a lot of information about film (and TV shows) that you can download from multiple places

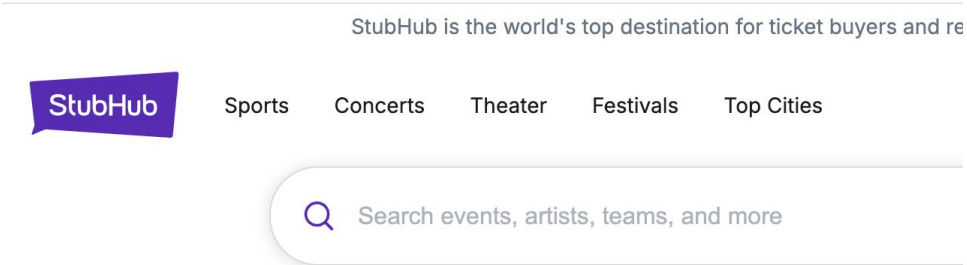
Comps - K-Means Clustering

If we make plots where we put different axes of these different pieces of metadata, you can see different titles will “cluster” together

Comps - K-Means Clustering

You can imagine there is a lot more information about movies that we can include that it will get cumbersome to try to plot them all together to find movies that are close to one another.

K-Means! (or any other similar clustering algorithm)

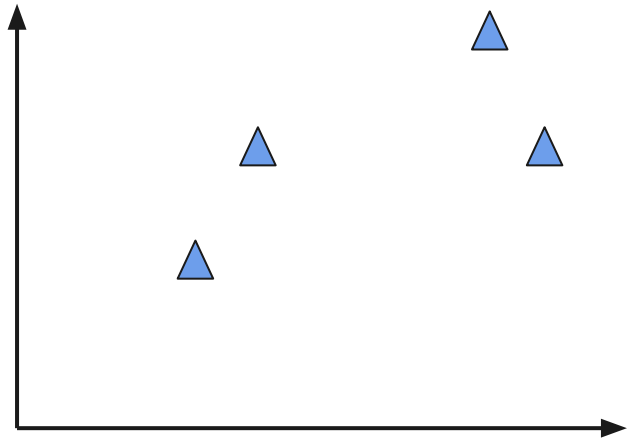


K-Means clustering

You can imagine there is a lot more information about movies that we can include that it will get cumbersome to try to plot them all together to find movies that are close to one another.

Step 1

- **Pick how many clusters you think there are**

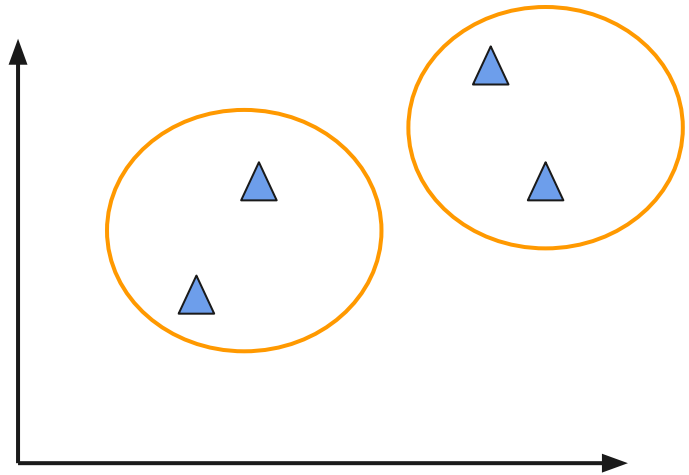


K-Means clustering

You can imagine there is a lot more information about movies that we can include that it will get cumbersome to try to plot them all together to find movies that are close to one another.

Step 1

- **Pick how many clusters you think there are**

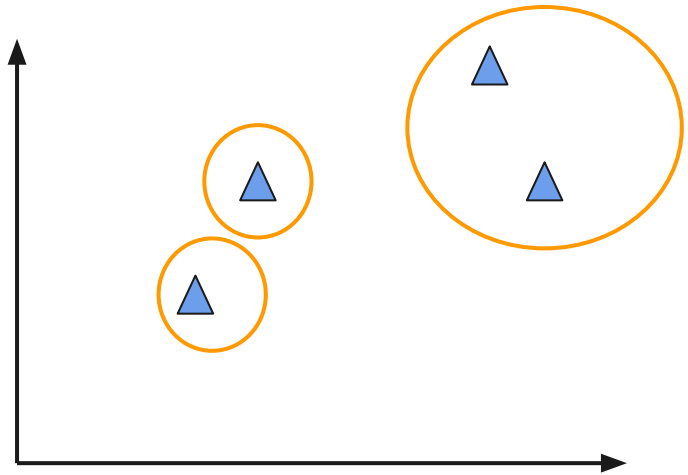


K-Means clustering

You can imagine there is a lot more information about movies that we can include that it will get cumbersome to try to plot them all together to find movies that are close to one another.

Step 1

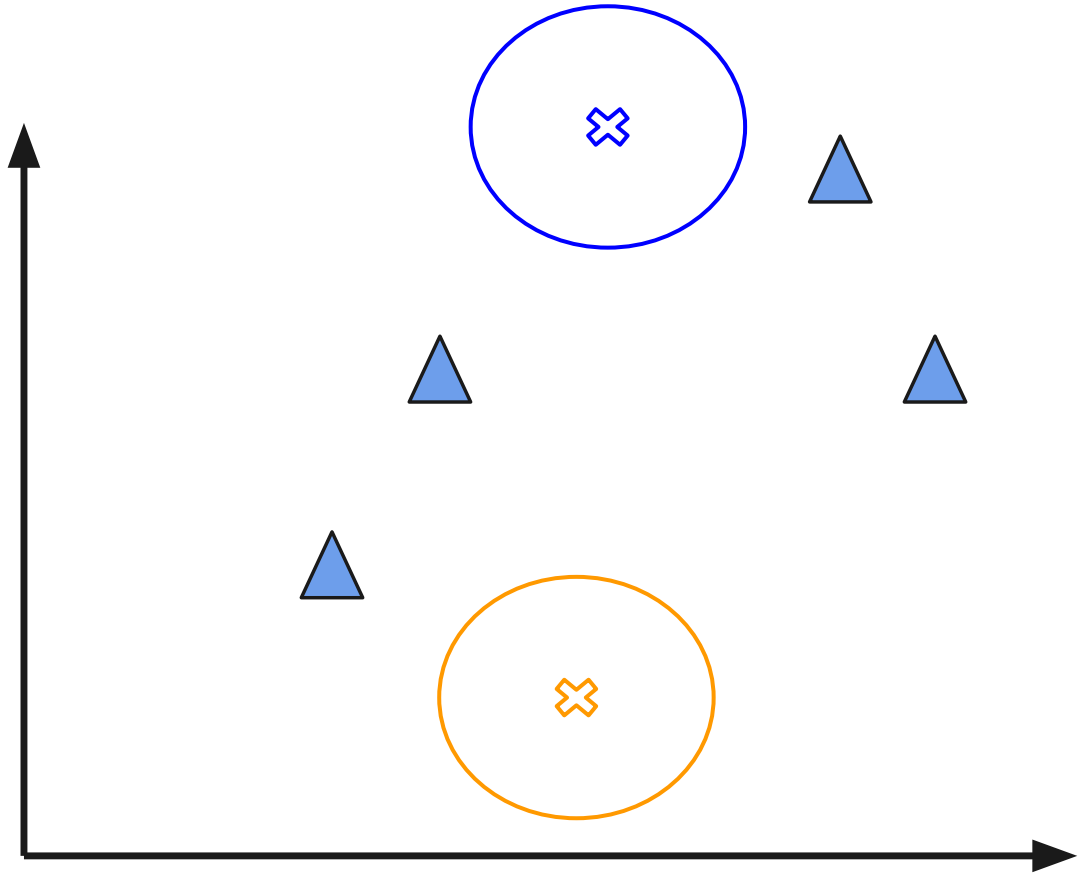
- **Pick how many clusters you think there are**



K-Means clustering

Step 2

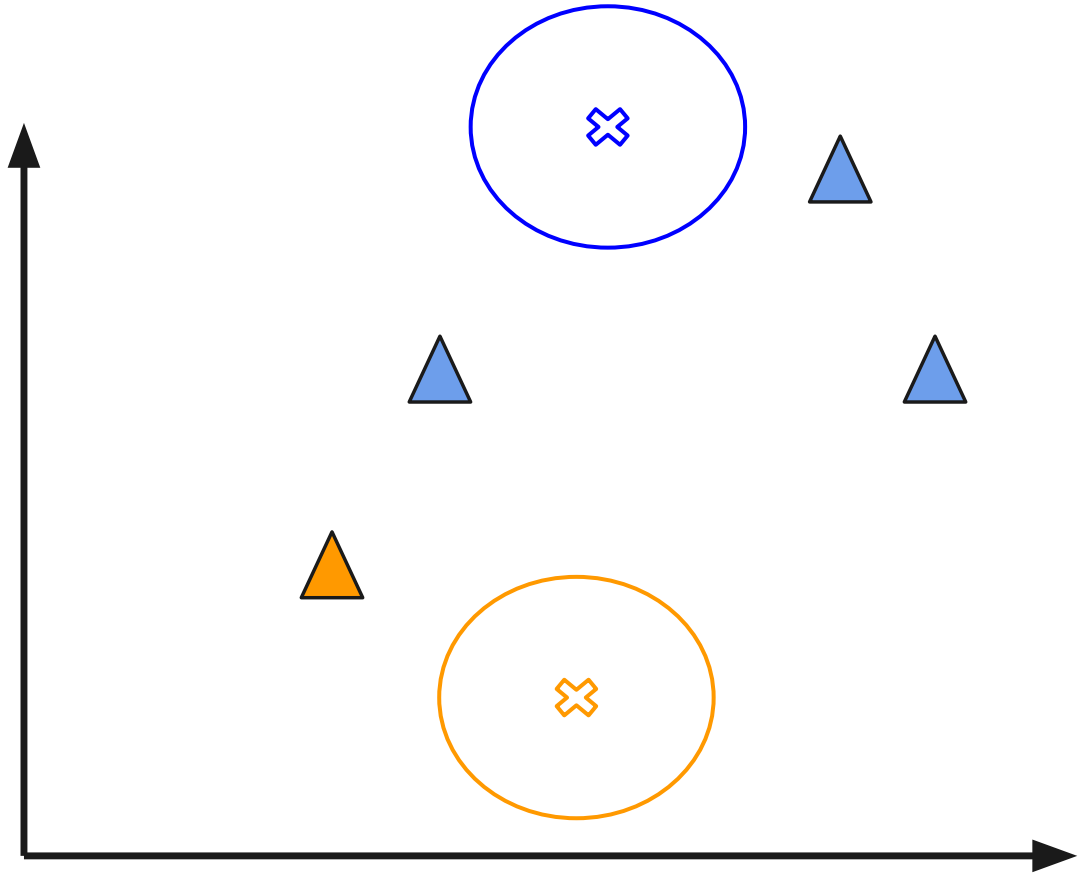
- Randomly set the central position of each cluster



K-Means clustering

Step 3

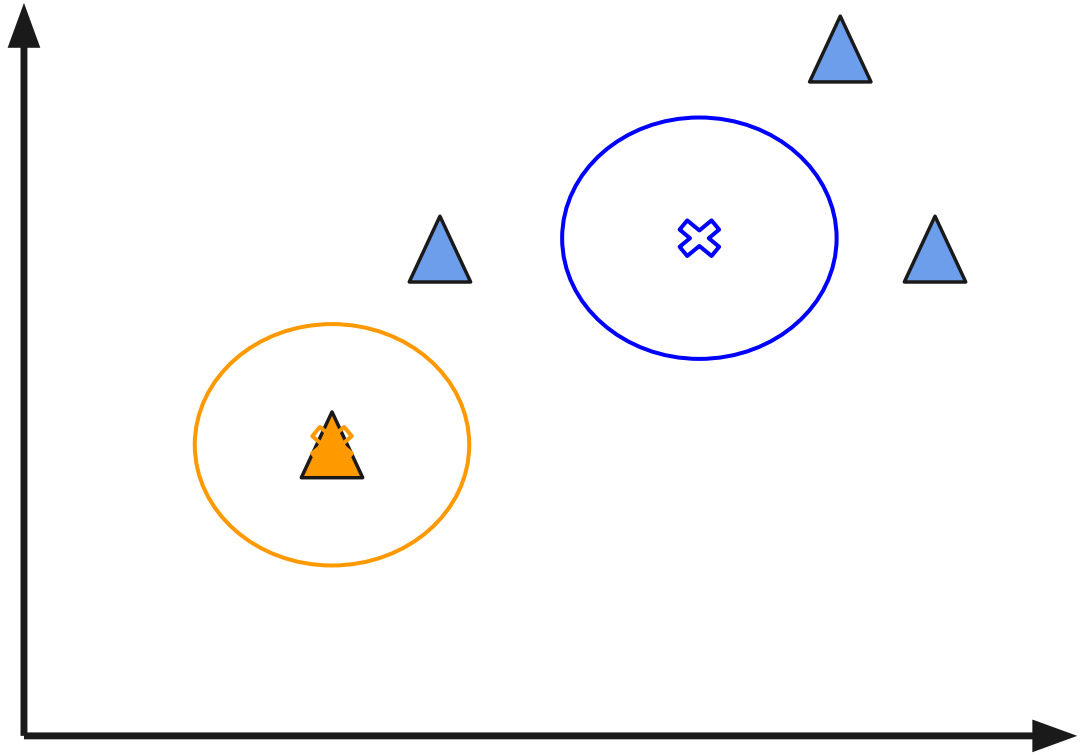
- Assign data points to the nearest cluster



K-Means clustering

Step 4

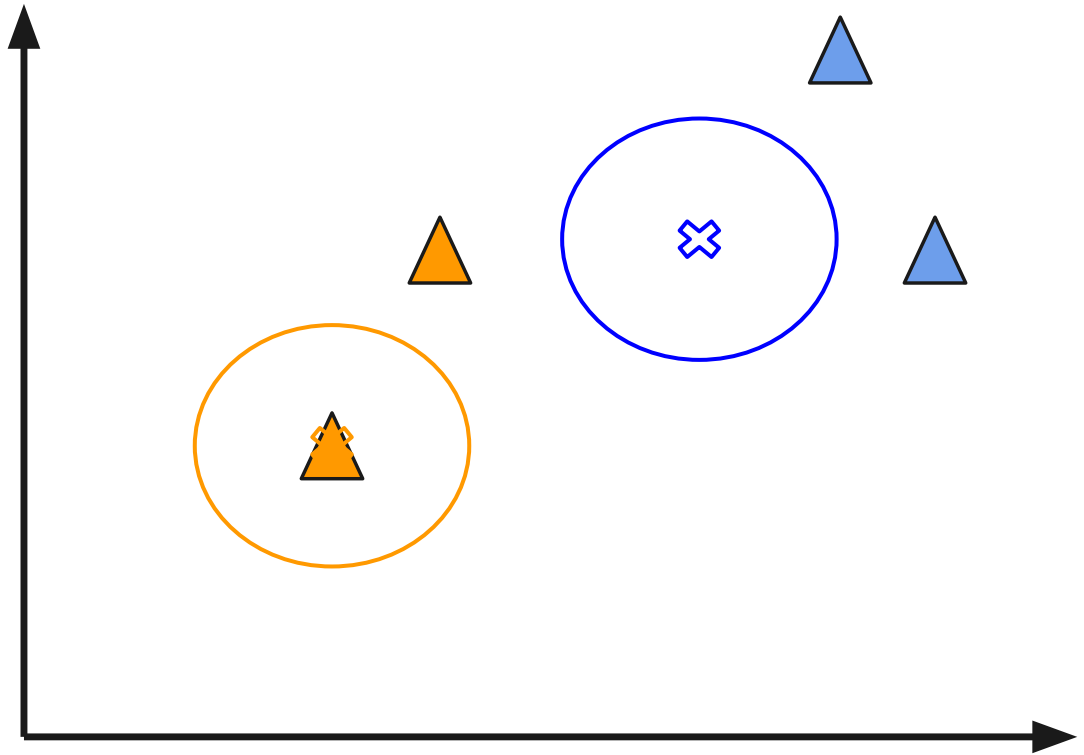
- We re-calculate the centers of the clusters
- New centroid is the average of all the data points



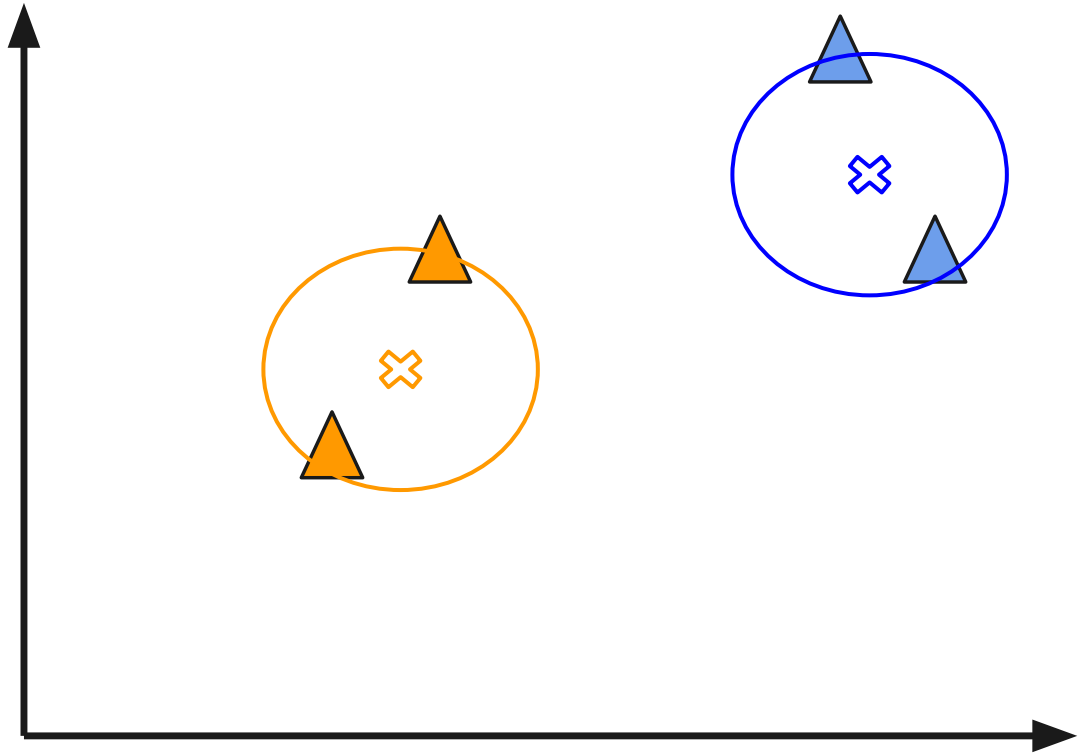
K-Means clustering

Step 5

- Repeat steps 3&4

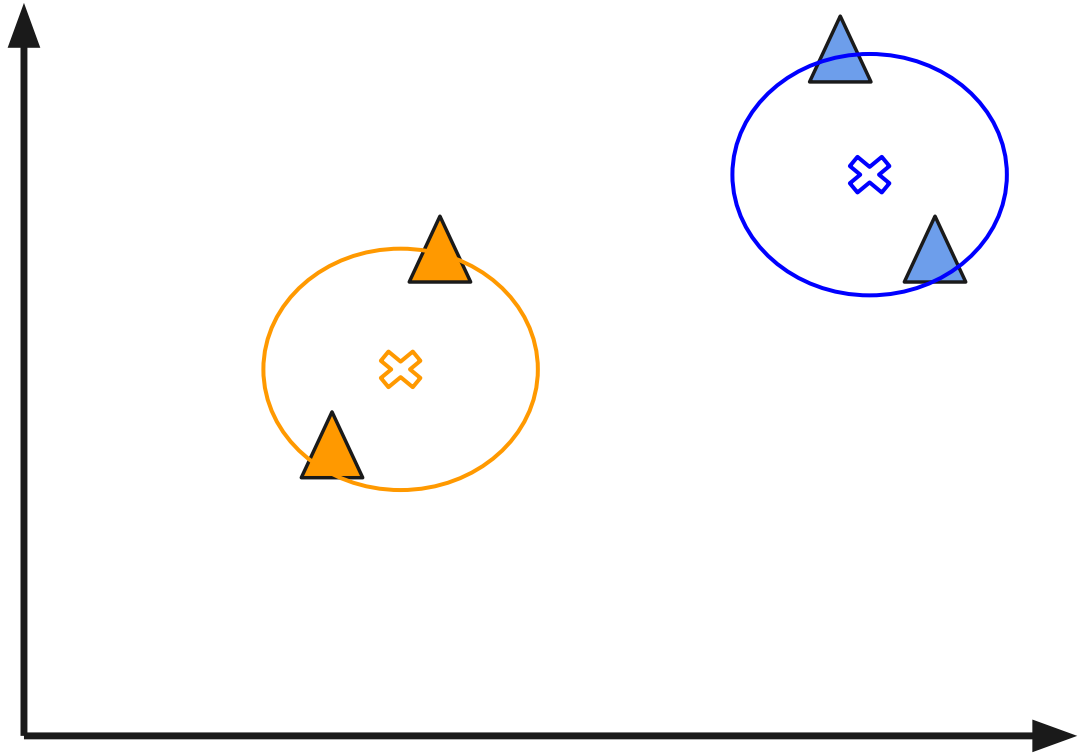


K-Means clustering



K-Means clustering

- We keep going until the centroid positions do not change substantially



Find the data you want to look at

MovieLens

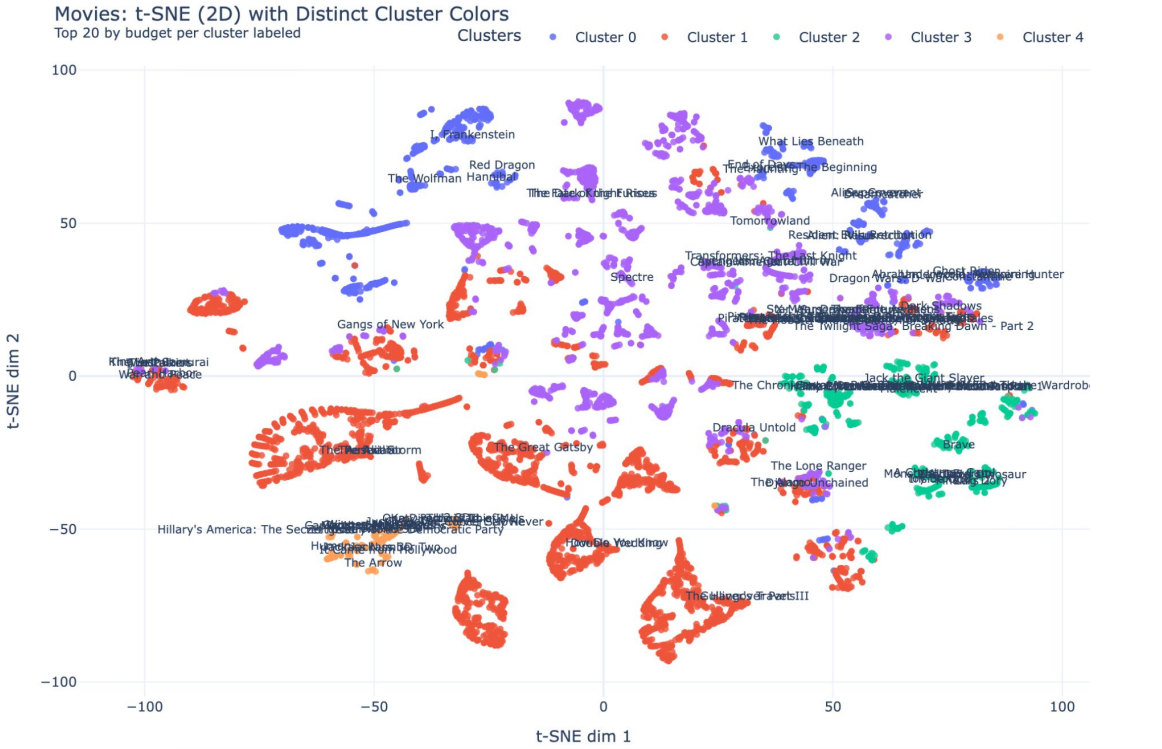
GroupLens Research has collected and made available rating data sets from the MovieLens web site (https://movielens.org). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

You can upload to ChatGPT to get it to tidy it up

Tidied Movies Metadata CSV

		title	release_date	budget	revenue	genres	runtime
1	0	Toy Story	1995-10-30	30000000.0	373554033.0	[{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}, {'id': 10751, 'name':	81.0
2	1	Jumanji	1995-12-15	65000000.0	262797249.0	[{'id': 12, 'name': 'Adventure'}, {'id': 14, 'name': 'Fantasy'}, {'id': 10751, 'name':	104.0
3	2	Grumpier Old Men	1995-12-22	0.0	0.0	[{'id': 10749, 'name': 'Romance'}, {'id': 35, 'name': 'Comedy'}]	101.0
4	3	Waiting to Exhale	1995-12-22	16000000.0	81452156.0	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'name': 'Drama'}, {'id': 10749, 'name':	127.0

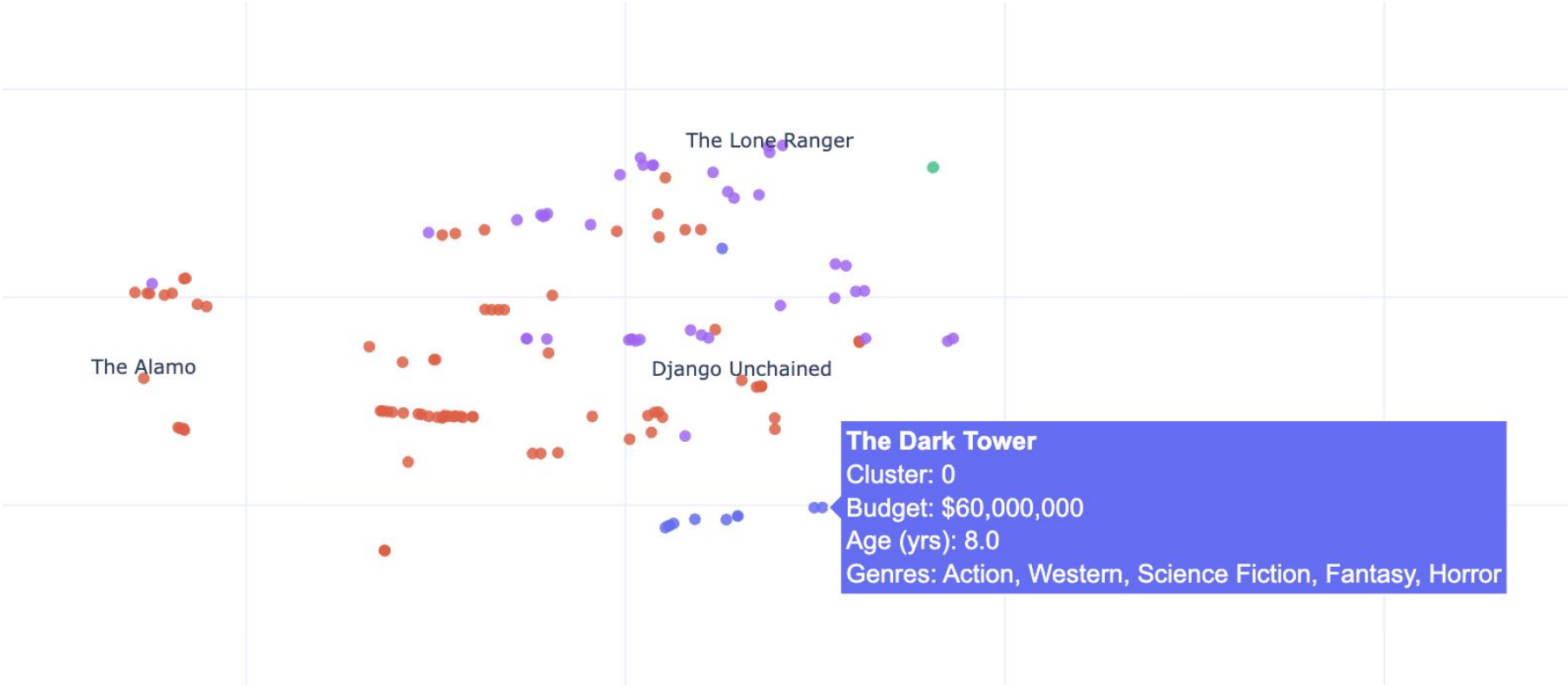
Ask it to run K-Means clustering for you and plot it



Quickly check it makes some sense…

If we have time

Quickly check it makes some sense...

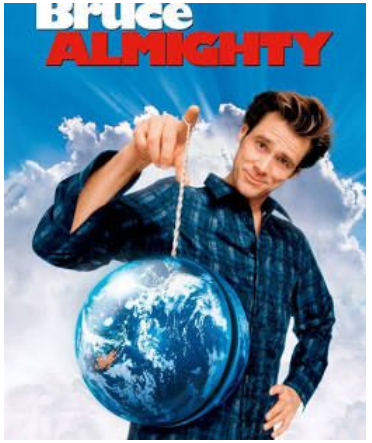


Then what about barbie?

$426M

Barbie (2023)

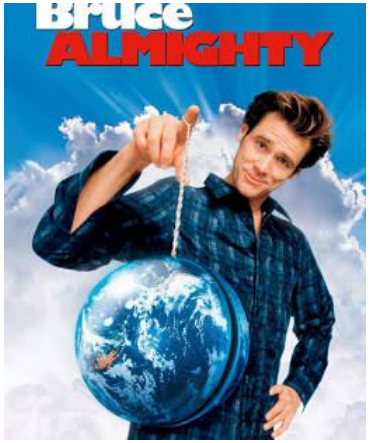
Budget: $145M
Genre: Comedy, Family
Age: 0



Could be better of course

$426M

- No franchise information
- No cast information
- No studio information



But I don’t make movies

Comps aren’t only used in the movie industry, there are many other places you’ll find it and can apply clustering techniques. A popular area to try it in is the stock market.

Calculating valuation of a company IPOing

TALA

Square

Revenue: $5.5B
Market Cap: $38B
Payment processing

Revenue: $31.0B
Market Cap: $290B
Digital payments

adyen

Revenue: $1.2B
Market Cap: $50B
Payment platform

Calculating house prices

Enter The Property Details Below To Calculate the ARV

18 Longlane Road, West Hartford, CT, USA

Baths	Beds	Sqft
2	3	1440

AVERAGE COMP PRICE PER SQ FOOT	SUBJECT PROPERTY ZESTIMATE	ESTIMATED AFTER REHAB VALUE
$161	$308,058	$232,046

Data provided by Zillow
Estimates for informational purposes only and may not be accurate

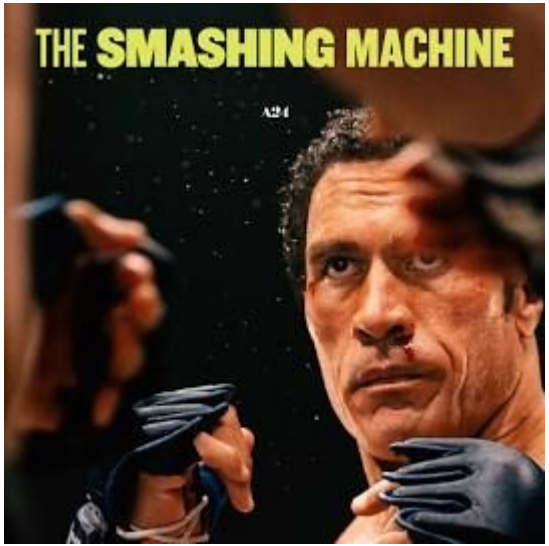
Please select your comps

	Address	Year Built	Beds	Baths	Sqft	Sold ($)	$/Sqft	Sale date
☑	85 Fairfield Rd, West Hartford CT	1957	3	3	2088	372500	178	03/27/2020
☑	14 Barksdale Rd, West Hartford CT	1957	3	2.5	1967	175000	89	02/21/2020

And many more

- Film
- Company valuations
- Real estate
- Compensation
- Art
- ….

Sounds pretty easy right?



October 3, 2025

https://tinyurl.com/usc-film