PROJECT PROPOSAL

PREDICTING MOVIE SUCCESS USING LOGISTIC REGRESSION

PRESENTED BY

PRASHANT KOIRALA
AASKA KOIRALA
AISHMITA YONZAN

PROBLEM STATEMENT

Can we predict whether a movie will be successful or not based on its metadata (e.g., budget, runtime, genre, and other pre-release attributes)?

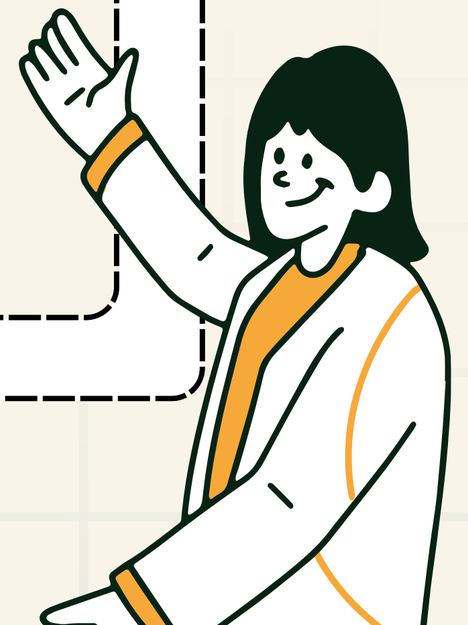
TYPE OF PROBLEM & STATISTICAL MODEL

Type of Problem:

- Binary Classification Problem
- Predictive Modeling Task

STATISTICAL TECHNIQUE CHOSEN

Logistic Regression



LOGISTIC REGRESSION

- A statistical model used for binary classification.
- Estimates the probability that a given input belongs to a particular category.
- Uses the sigmoid function to output a value between 0 and 1.
- Predicts outcomes like “Success (1)” or “Not Success (0)”.

KEY FEATURES

- It is ideal for binary outcomes (success vs. not success).
- It provides clear coefficients to understand how each feature (e.g., budget) impacts the probability of success.
- Suitable for a student project, as it's straightforward to implement and interpret.

LITERATURE REVIEW & DATASET SELECTION

DA, O., OM, S., AK, F., O, A., A, O., A, O., A, W., & M, Y. (2021). Movie success prediction using data mining. British Journal of Computer Networking and Information Technology, 4(2), 22–30. https://doi.org/10.52589/bjcnit-cqocirec

Zheng, Y. (2024). Predicting movie box office based on machine learning, deep learning, and statistical methods. Applied and Computational Engineering, 94(1), 20–32. https://doi.org/10.54254/2755-2721/94/2024melb0069

Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V., & Logesh, R. (2017). Predicting movie box office success using multiple regression and SVM. Journal of Science & Statistics, 182–186. https://doi.org/10.1109/iss1.2017.8389394

Velingkar, G., Varadarajan, R., Lanka, S., & M, A. K. (2022). Movie Box-Office success Prediction using Machine Learning. 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), 9, 1–6. https://doi.org/10.1109/icpc2t53885.2022.9776798

Mr, N. V., Mr, P. M., & Pb, S. B. (2014). Predicting movie success based on IMDB data. International Journal of Business Intelligents, 3(2). https://doi.org/10.20894/ijbi.105.003.002.004

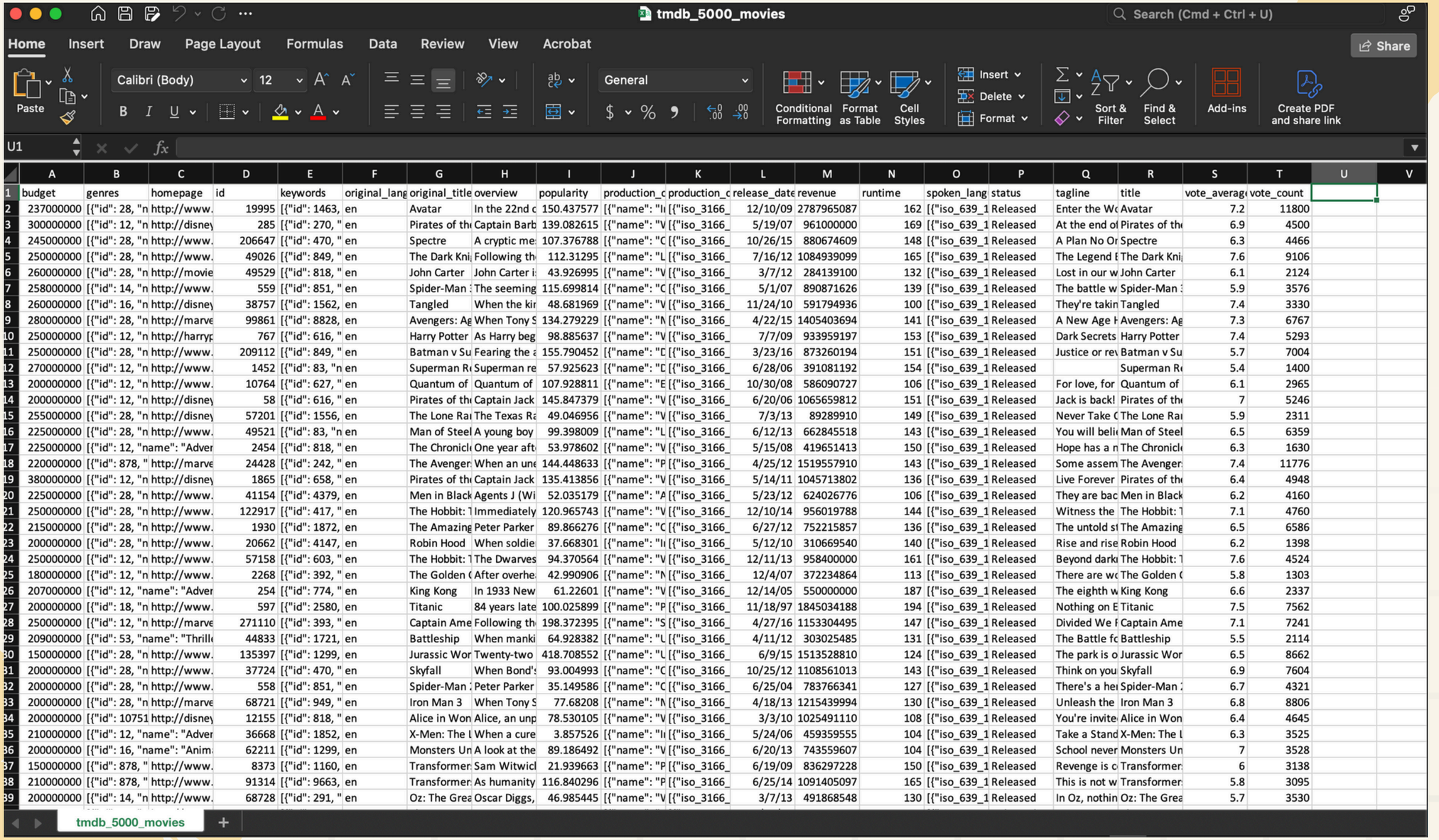
BRIEF OVERVIEW OF LITERATURE REVIEW

Proceedings Article • 10.1109/ISS1.2017.8389394 **3. Predicting movie box office success using multiple regression and SVM** V. Subramaniyaswamy, M. Viginesh Vaibhav, R. Vishnu Prasad +1 more 1 Dec 2017 66 26 Request PDF Podcast Chat 66	The paper you referenced is not the same as the one discussed here. This research focuses on predicting movie box office success using multiple regression and SVM, analyzing factors like trailer views, Wikipedia page views, and critic ratings.
Proceedings Article • 10.1109/ICPC2T53885.2022.9776798 **4. Movie Box-Office Success Prediction Using Machine Learning** Gaurang Velingkar, Rakshita Varadarajan, Sidharth Lanka +1 more 1 Mar 2022 66 4 Request PDF Podcast Chat 66	The paper focuses on predicting movie box office success using machine learning, specifically employing a Random Forest Regression model to analyze factors like genre, budget, and star power, aiding investors in making informed decisions before a movie's release.
• Proceedings Article • 10.1109/icpc2t53885.2022.9776798 **5. Movie Box-Office Success Prediction Using Machine Learning** 1 Mar 2022 66 8 PDF Podcast Chat	The paper focuses on predicting movie box office success using machine learning, specifically employing a Random Forest Regression model to analyze factors like genre, budget, and star power, aiding investors in making informed decisions before a movie's release.

DATASET SELECTION

- Source: TMDB / Kaggle (5,000+ movies)
- Features: Budget, Genre, Runtime, Country, Vote Count, Revenue

EXPLORATORY DATA ANALYSIS (EDA)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	budget	genres	homepage	id	keywords	original_lang	original_title	overview	popularity	production_c	production_c	release_date	revenue	runtime	spoken_lang	status	tagline	title	vote_averag	vote_count		
2	237000000	[{"id": 28, "n	http://www.	19995	[{"id": 1463,	en	Avatar	In the 22nd c	150.437577	[{"name": "I	[{"iso_3166_	12/10/09	2787965087	162	[{"iso_639_1	Released	Enter the W	Avatar	7.2	11800		
3	300000000	[{"id": 12, "n	http://disney	285	[{"id": 270, "	en	Pirates of the	Captain Barb	139.082615	[{"name": "V	[{"iso_3166_	5/19/07	961000000	169	[{"iso_639_1	Released	At the end of	Pirates of the	6.9	4500		
4	245000000	[{"id": 28, "n	http://www.	206647	[{"id": 470, "	en	Spectre	A cryptic me	107.376788	[{"name": "C	[{"iso_3166_	10/26/15	880674609	148	[{"iso_639_1	Released	A Plan No Or	Spectre	6.3	4466		
5	250000000	[{"id": 28, "n	http://www.	49026	[{"id": 849, "	en	The Dark Kni	Following th	112.31295	[{"name": "L	[{"iso_3166_	7/16/12	1084939099	165	[{"iso_639_1	Released	The Legend I	The Dark Kni	7.6	9106		
6	260000000	[{"id": 28, "n	http://movie	49529	[{"id": 818, "	en	John Carter	John Carter i	43.926995	[{"name": "V	[{"iso_3166_	3/7/12	284139100	132	[{"iso_639_1	Released	Lost in our w	John Carter	6.1	2124		
7	258000000	[{"id": 14, "n	http://www.	559	[{"id": 851, "	en	Spider-Man 3	The seeming	115.699814	[{"name": "C	[{"iso_3166_	5/1/07	890871626	139	[{"iso_639_1	Released	The battle w	Spider-Man 3	5.9	3576		
8	260000000	[{"id": 16, "n	http://disney	38757	[{"id": 1562,	en	Tangled	When the kir	48.681969	[{"name": "V	[{"iso_3166_	11/24/10	591794936	100	[{"iso_639_1	Released	They're takin	Tangled	7.4	3330		
9	280000000	[{"id": 28, "n	http://marve	99861	[{"id": 8828,	en	Avengers: Ag	When Tony S	134.279229	[{"name": "M	[{"iso_3166_	4/22/15	1405403694	141	[{"iso_639_1	Released	A New Age H	Avengers: Ag	7.3	6767		
10	250000000	[{"id": 12, "n	http://harryp	767	[{"id": 616, "	en	Harry Potter	As Harry beg	98.885637	[{"name": "V	[{"iso_3166_	7/7/09	933959197	153	[{"iso_639_1	Released	Dark Secrets	Harry Potter	7.4	5293		
11	250000000	[{"id": 28, "n	http://www.	209112	[{"id": 849, "	en	Batman v Su	Fearing the a	155.790452	[{"name": "D	[{"iso_3166_	3/23/16	873260194	151	[{"iso_639_1	Released	Justice or re	Batman v Su	5.7	7004		
12	270000000	[{"id": 12, "n	http://www.	1452	[{"id": 83, "n	en	Superman R	Superman re	57.925623	[{"name": "D	[{"iso_3166_	6/28/06	391081192	154	[{"iso_639_1	Released		Superman R	5.4	1400		
13	200000000	[{"id": 12, "n	http://www.	10764	[{"id": 627, "	en	Quantum of	Quantum of	107.928811	[{"name": "E	[{"iso_3166_	10/30/08	586090727	106	[{"iso_639_1	Released	For love, for	Quantum of	6.1	2965		
14	200000000	[{"id": 12, "n	http://disney	58	[{"id": 616, "	en	Pirates of the	Captain Jack	145.847379	[{"name": "V	[{"iso_3166_	6/20/06	1065659812	151	[{"iso_639_1	Released	Jack is back!	Pirates of the	7	5246		
15	255000000	[{"id": 28, "n	http://disney	57201	[{"id": 1556,	en	The Lone Rai	The Texas Ra	49.046956	[{"name": "V	[{"iso_3166_	7/3/13	89289910	149	[{"iso_639_1	Released	Never Take C	The Lone Rai	5.9	2311		
16	225000000	[{"id": 28, "n	http://www.	49521	[{"id": 83, "n	en	Man of Steel	A young boy	99.398009	[{"name": "L	[{"iso_3166_	6/12/13	662845518	143	[{"iso_639_1	Released	You will beli	Man of Steel	6.5	6359		
17	225000000	[{"id": 12, "name": "Adver		2454	[{"id": 818, "	en	The Chronicl	One year aft	53.978602	[{"name": "V	[{"iso_3166_	5/15/08	419651413	150	[{"iso_639_1	Released	Hope has a n	The Chronicl	6.3	1630		
18	220000000	[{"id": 878, "	http://marve	24428	[{"id": 242, "	en	The Avenger	When an une	144.448633	[{"name": "P	[{"iso_3166_	4/25/12	1519557910	143	[{"iso_639_1	Released	Some assem	The Avenger	7.4	11776		
19	380000000	[{"id": 12, "n	http://disney	1865	[{"id": 658, "	en	Pirates of the	Captain Jack	135.413856	[{"name": "V	[{"iso_3166_	5/14/11	1045713802	136	[{"iso_639_1	Released	Live Forever	Pirates of the	6.4	4948		
20	225000000	[{"id": 28, "n	http://www.	41154	[{"id": 4379,	en	Men in Black	Agents J (Wi	52.035179	[{"name": "A	[{"iso_3166_	5/23/12	624026776	106	[{"iso_639_1	Released	They are bac	Men in Black	6.2	4160		
21	250000000	[{"id": 28, "n	http://www.	122917	[{"id": 417, "	en	The Hobbit: T	Immediately	120.965743	[{"name": "V	[{"iso_3166_	12/10/14	956019788	144	[{"iso_639_1	Released	Witness the	The Hobbit: T	7.1	4760		
22	215000000	[{"id": 28, "n	http://www.	1930	[{"id": 1872,	en	The Amazing	Peter Parker	89.866276	[{"name": "C	[{"iso_3166_	6/27/12	752215857	136	[{"iso_639_1	Released	The untold st	The Amazing	6.5	6586		
23	200000000	[{"id": 28, "n	http://www.	20662	[{"id": 4147,	en	Robin Hood	When soldie	37.668301	[{"name": "I	[{"iso_3166_	5/12/10	310669540	140	[{"iso_639_1	Released	Rise and rise	Robin Hood	6.2	1398		
24	250000000	[{"id": 12, "n	http://www.	57158	[{"id": 603, "	en	The Hobbit: T	The Dwarves	94.370564	[{"name": "V	[{"iso_3166_	12/11/13	958400000	161	[{"iso_639_1	Released	Beyond dark	The Hobbit: T	7.6	4524		
25	180000000	[{"id": 12, "n	http://www.	2268	[{"id": 392, "	en	The Golden C	After overhe	42.990906	[{"name": "M	[{"iso_3166_	12/4/07	372234864	113	[{"iso_639_1	Released	There are wo	The Golden C	5.8	1303		
26	207000000	[{"id": 12, "name": "Adver		254	[{"id": 774, "	en	King Kong	In 1933 New	61.22601	[{"name": "V	[{"iso_3166_	12/14/05	550000000	187	[{"iso_639_1	Released	The eighth w	King Kong	6.6	2337		
27	200000000	[{"id": 18, "n	http://www.	597	[{"id": 2580,	en	Titanic	84 years late	100.025899	[{"name": "P	[{"iso_3166_	11/18/97	1845034188	194	[{"iso_639_1	Released	Nothing on E	Titanic	7.5	7562		
28	250000000	[{"id": 12, "n	http://marve	271110	[{"id": 393, "	en	Captain Ame	Following th	198.372395	[{"name": "S	[{"iso_3166_	4/27/16	1153304495	147	[{"iso_639_1	Released	Divided We F	Captain Ame	7.1	7241		
29	209000000	[{"id": 53, "name": "Thrill		44833	[{"id": 1721,	en	Battleship	When manki	64.928382	[{"name": "L	[{"iso_3166_	4/11/12	303025485	131	[{"iso_639_1	Released	The Battle fo	Battleship	5.5	2114		
30	150000000	[{"id": 28, "n	http://www.	135397	[{"id": 1299,	en	Jurassic Wor	Twenty-two	418.708552	[{"name": "L	[{"iso_3166_	6/9/15	1513528810	124	[{"iso_639_1	Released	The park is o	Jurassic Wor	6.5	8662		
31	200000000	[{"id": 28, "n	http://www.	37724	[{"id": 470, "	en	Skyfall	When Bond's	93.004993	[{"name": "C	[{"iso_3166_	10/25/12	1108561013	143	[{"iso_639_1	Released	Think on you	Skyfall	6.9	7604		
32	200000000	[{"id": 28, "n	http://www.	558	[{"id": 851, "	en	Spider-Man 3	Peter Parker	35.149586	[{"name": "C	[{"iso_3166_	6/25/04	783766341	127	[{"iso_639_1	Released	There's a hei	Spider-Man 3	6.7	4321		
33	200000000	[{"id": 28, "n	http://marve	68721	[{"id": 949, "	en	Iron Man 3	When Tony S	77.68208	[{"name": "M	[{"iso_3166_	4/18/13	1215439994	130	[{"iso_639_1	Released	Unleash the	Iron Man 3	6.8	8806		
34	200000000	[{"id": 10751	http://disney	12155	[{"id": 818, "	en	Alice in Won	Alice, an unp	78.530105	[{"name": "V	[{"iso_3166_	3/3/10	1025491110	108	[{"iso_639_1	Released	You're invite	Alice in Won	6.4	4645		
35	210000000	[{"id": 12, "name": "Adver		36668	[{"id": 1852,	en	X-Men: The L	When a cure	3.857526	[{"name": "I	[{"iso_3166_	5/24/06	459359555	104	[{"iso_639_1	Released	Take a Stand	X-Men: The L	6.3	3525		
36	200000000	[{"id": 16, "name": "Anim		62211	[{"id": 1299,	en	Monsters Un	A look at the	89.186492	[{"name": "V	[{"iso_3166_	6/20/13	743559607	104	[{"iso_639_1	Released	School never	Monsters Un	7	3528		
37	150000000	[{"id": 878, "	http://www.	8373	[{"id": 1160,	en	Transformer	Sam Witwicl	21.939663	[{"name": "P	[{"iso_3166_	6/19/09	836297228	150	[{"iso_639_1	Released	Revenge is c	Transformer	6	3138		
38	210000000	[{"id": 878, "	http://www.	91314	[{"id": 9663,	en	Transformer	As humanity	116.840296	[{"name": "P	[{"iso_3166_	6/25/14	1091405097	165	[{"iso_639_1	Released	This is not w	Transformer	5.8	3095		
39	200000000	[{"id": 14, "n	http://www.	68728	[{"id": 291, "	en	Oz: The Grea	Oscar Diggs,	46.985445	[{"name": "V	[{"iso_3166_	3/7/13	491868548	130	[{"iso_639_1	Released	In Oz, nothin	Oz: The Grea	5.7	3530		

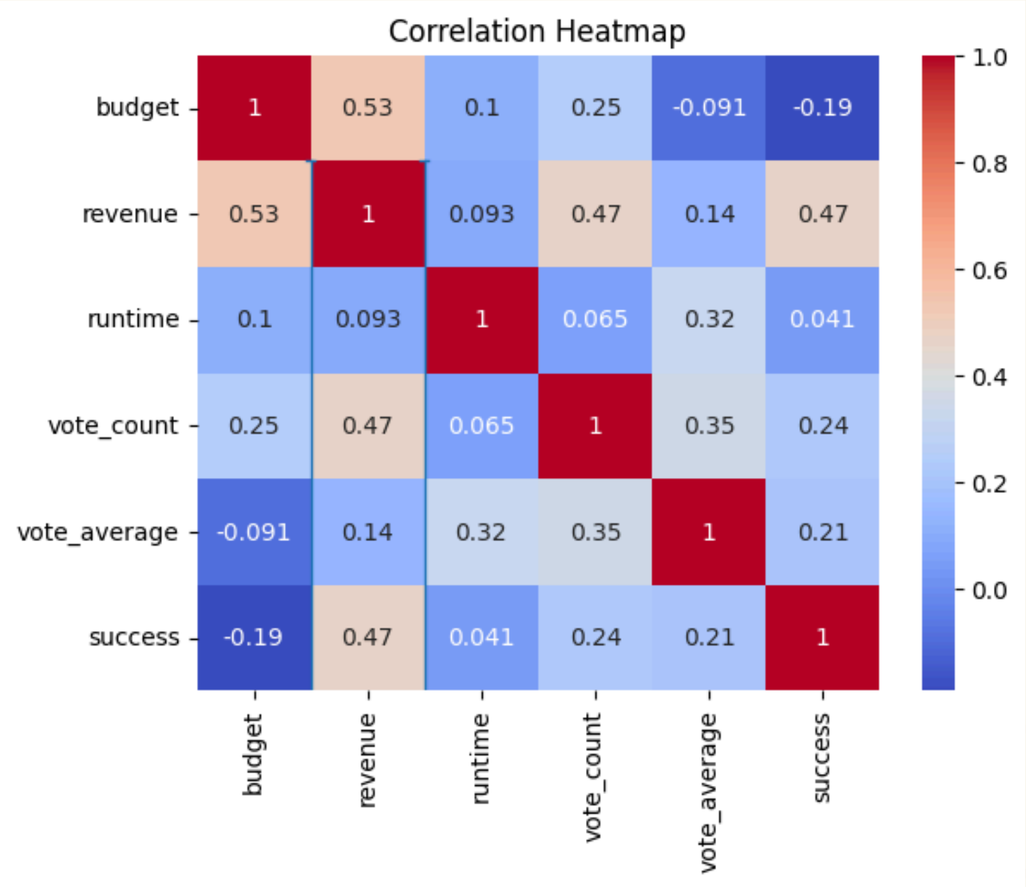
tmdb_5000_movies

EXPLORATORY DATA ANALYSIS (EDA)

Features Used for Analysis

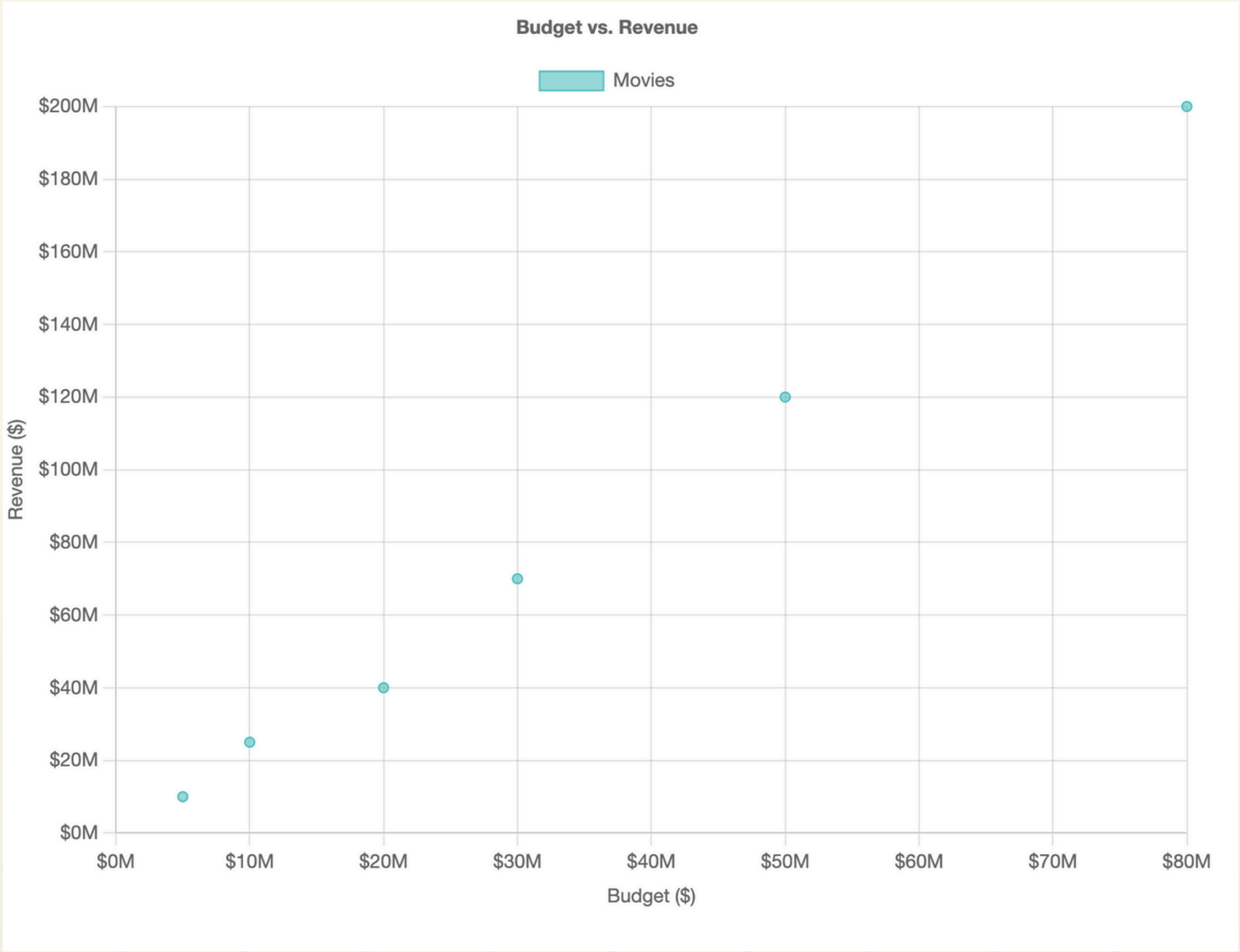
- **Numerical**: budget, revenue, runtime, year, vote_count (and possibly vote_average if not dropped).
- **Categorical**: certification_US, genre, country.
- **Target**: success (True/False, where True means revenue ≥ 2x budget).

WHY THOSE FEATURES?



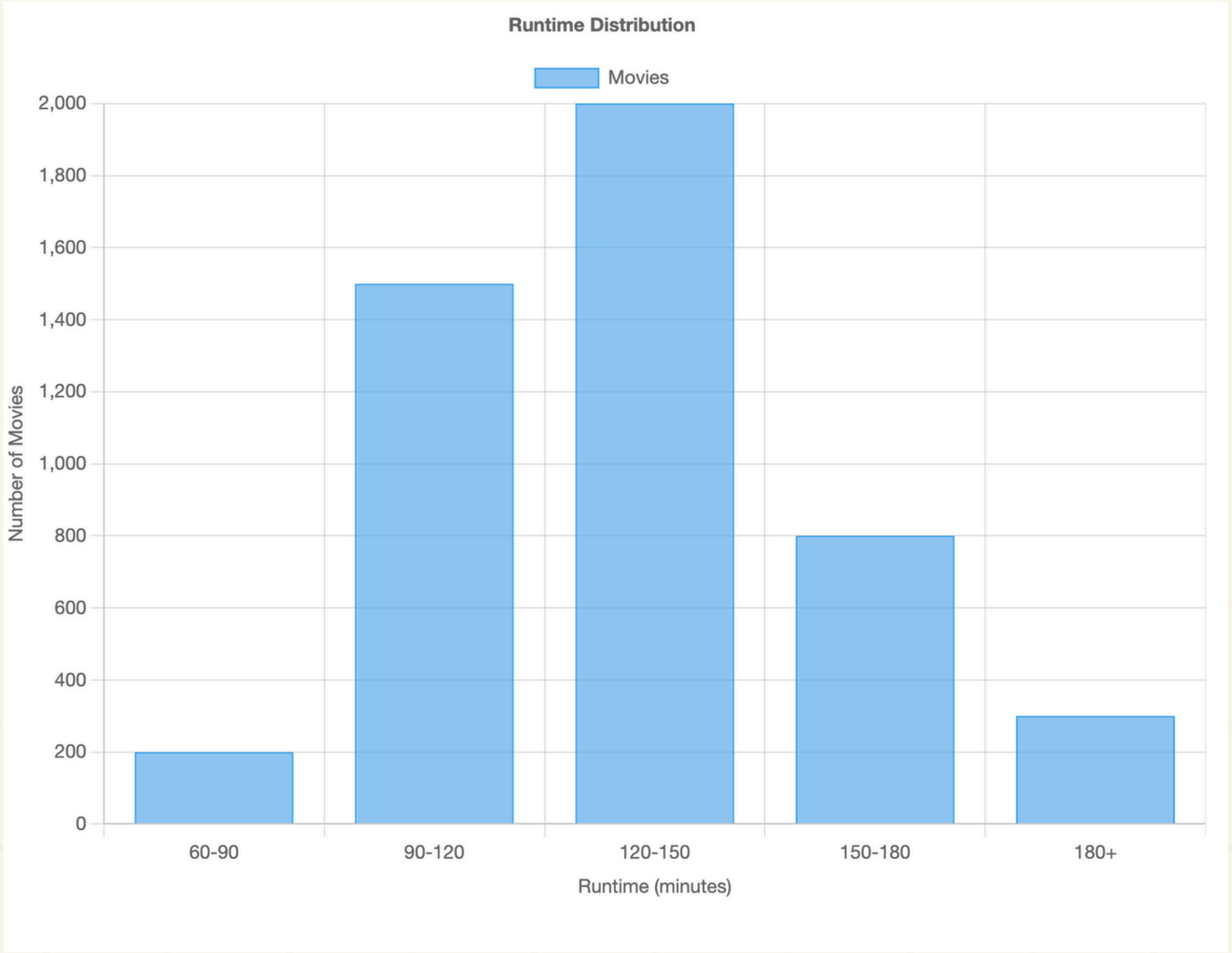
1. SUCCESS IS MOST TIED TO REVENUE (0.47) AND VOTE_COUNT (0.24), SO THESE FEATURES MIGHT PREDICT SUCCESS WELL.
2. BUDGET AND REVENUE ARE RELATED (0.53), MEANING BIGGER BUDGETS OFTEN LEAD TO HIGHER REVENUE.
3. RUNTIME HAS LITTLE CONNECTION TO OTHER FEATURES (ALL NEAR 0), SO IT MIGHT NOT BE A STRONG PREDICTOR OF SUCCESS.

WHY THOSE FEATURES?



PS: SHOWS IF HIGHER BUDGETS LEAD TO HIGHER REVENUE.

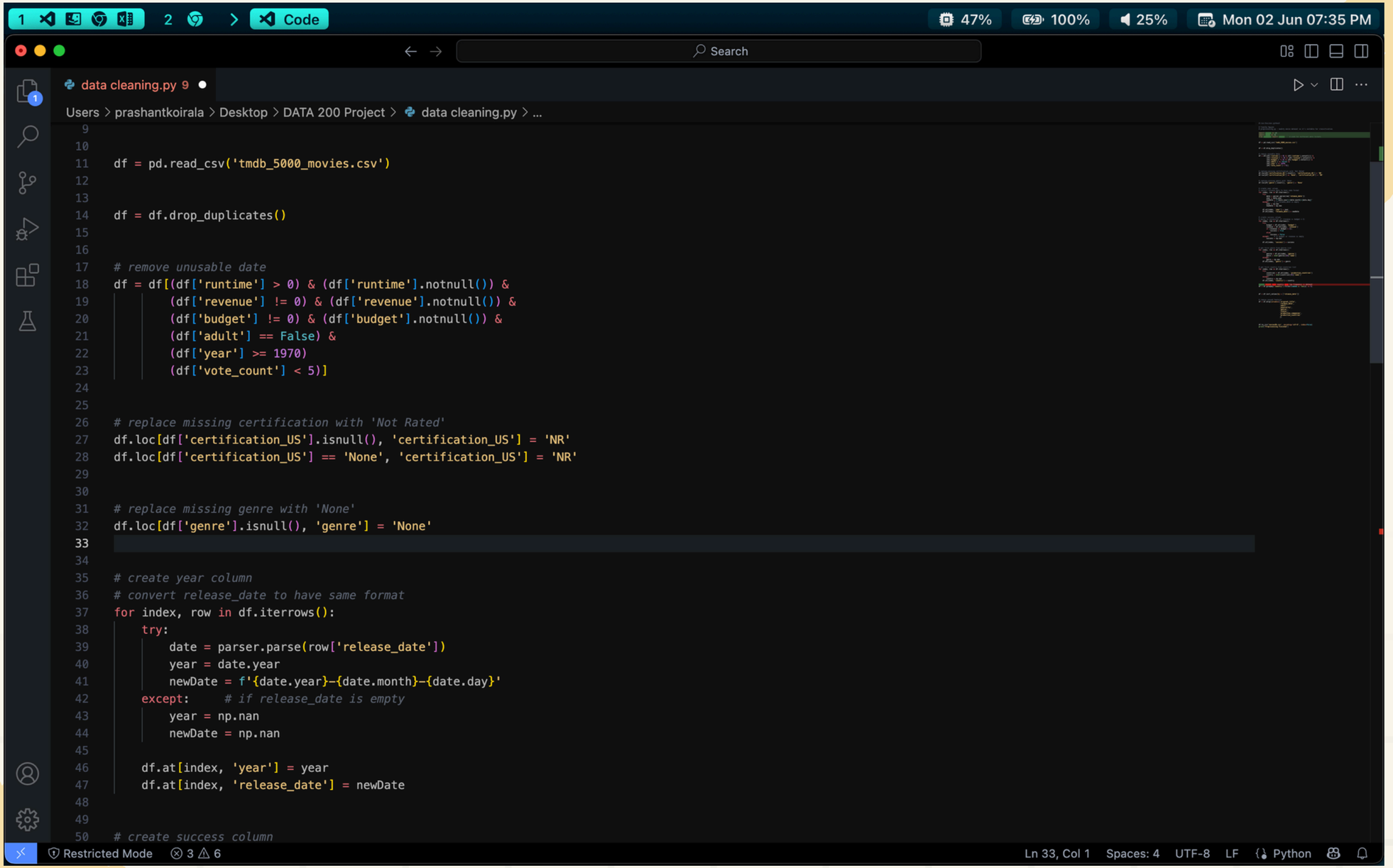
WHY THOSE FEATURES?



PS: SHOWS HOW RUNTIMES ARE DISTRIBUTED.

DATA CLEANING

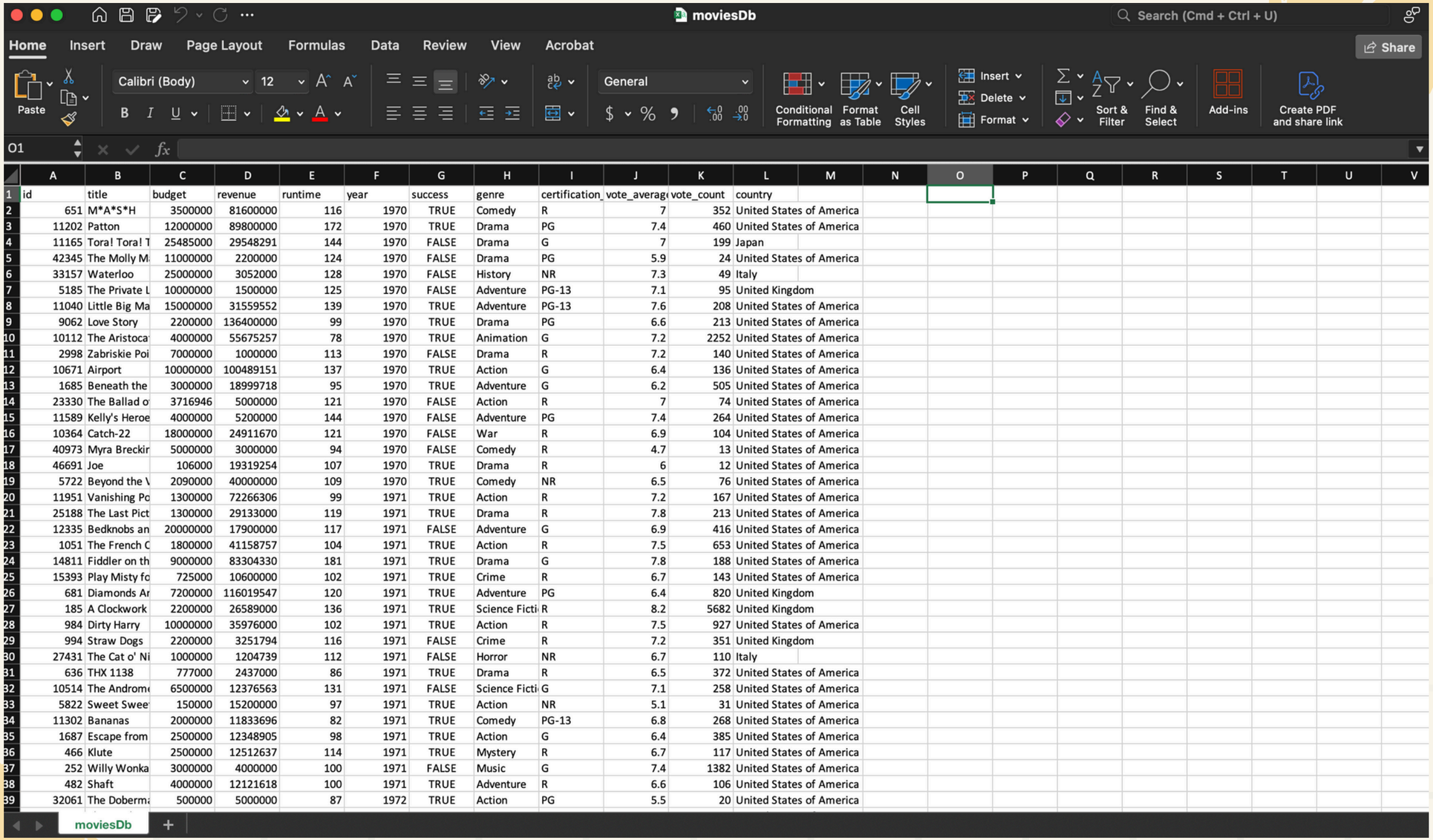
DATA CLEANING PROGRAM



DATA CLEANING

1. **Removed duplicates and movies with missing/zero budget, revenue, or runtime.**
2. **Replaced missing certification_US with “NR”; extracted first genre and country.**
3. **Filtered movies before 1970 and countries with < 5 occurrences.**
4. **Created success (True if revenue ≥ 2x budget) and year columns.**

AFTER DATA CLEANING



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	id	title	budget	revenue	runtime	year	success	genre	certification_	vote_averag	vote_count	country										
2	651	M*A*S*H	3500000	81600000	116	1970	TRUE	Comedy	R	7	352	United States of America										
3	11202	Patton	12000000	89800000	172	1970	TRUE	Drama	PG	7.4	460	United States of America										
4	11165	Tora! Tora! T	25485000	29548291	144	1970	FALSE	Drama	G	7	199	Japan										
5	42345	The Molly M	11000000	2200000	124	1970	FALSE	Drama	PG	5.9	24	United States of America										
6	33157	Waterloo	25000000	3052000	128	1970	FALSE	History	NR	7.3	49	Italy										
7	5185	The Private L	10000000	1500000	125	1970	FALSE	Adventure	PG-13	7.1	95	United Kingdom										
8	11040	Little Big Ma	15000000	31559552	139	1970	TRUE	Adventure	PG-13	7.6	208	United States of America										
9	9062	Love Story	2200000	136400000	99	1970	TRUE	Drama	PG	6.6	213	United States of America										
10	10112	The Aristoca	4000000	55675257	78	1970	TRUE	Animation	G	7.2	2252	United States of America										
11	2998	Zabriskie Poi	7000000	1000000	113	1970	FALSE	Drama	R	7.2	140	United States of America										
12	10671	Airport	10000000	100489151	137	1970	TRUE	Action	G	6.4	136	United States of America										
13	1685	Beneath the	3000000	18999718	95	1970	TRUE	Adventure	G	6.2	505	United States of America										
14	23330	The Ballad o	3716946	5000000	121	1970	FALSE	Action	R	7	74	United States of America										
15	11589	Kelly's Heroe	4000000	5200000	144	1970	FALSE	Adventure	PG	7.4	264	United States of America										
16	10364	Catch-22	18000000	24911670	121	1970	FALSE	War	R	6.9	104	United States of America										
17	40973	Myra Breckir	5000000	3000000	94	1970	FALSE	Comedy	R	4.7	13	United States of America										
18	46691	Joe	106000	19319254	107	1970	TRUE	Drama	R	6	12	United States of America										
19	5722	Beyond the V	2090000	40000000	109	1970	TRUE	Comedy	NR	6.5	76	United States of America										
20	11951	Vanishing Po	1300000	72266306	99	1971	TRUE	Action	R	7.2	167	United States of America										
21	25188	The Last Pict	1300000	29133000	119	1971	TRUE	Drama	R	7.8	213	United States of America										
22	12335	Bedknobs an	20000000	17900000	117	1971	FALSE	Adventure	G	6.9	416	United States of America										
23	1051	The French C	1800000	41158757	104	1971	TRUE	Action	R	7.5	653	United States of America										
24	14811	Fiddler on th	9000000	83304330	181	1971	TRUE	Drama	G	7.8	188	United States of America										
25	15393	Play Misty fo	725000	10600000	102	1971	TRUE	Crime	R	6.7	143	United States of America										
26	681	Diamonds Ar	7200000	116019547	120	1971	TRUE	Adventure	PG	6.4	820	United Kingdom										
27	185	A Clockwork	2200000	26589000	136	1971	TRUE	Science Ficti	R	8.2	5682	United Kingdom										
28	984	Dirty Harry	10000000	35976000	102	1971	TRUE	Action	R	7.5	927	United States of America										
29	994	Straw Dogs	2200000	3251794	116	1971	FALSE	Crime	R	7.2	351	United Kingdom										
30	27431	The Cat o' Ni	1000000	1204739	112	1971	FALSE	Horror	NR	6.7	110	Italy										
31	636	THX 1138	777000	2437000	86	1971	TRUE	Drama	R	6.5	372	United States of America										
32	10514	The Androm	6500000	12376563	131	1971	FALSE	Science Ficti	G	7.1	258	United States of America										
33	5822	Sweet Swee	150000	15200000	97	1971	TRUE	Action	NR	5.1	31	United States of America										
34	11302	Bananas	2000000	11833696	82	1971	TRUE	Comedy	PG-13	6.8	268	United States of America										
35	1687	Escape from	2500000	12348905	98	1971	TRUE	Action	G	6.4	385	United States of America										
36	466	Klute	2500000	12512637	114	1971	TRUE	Mystery	R	6.7	117	United States of America										
37	252	Willy Wonka	3000000	4000000	100	1971	FALSE	Music	G	7.4	1382	United States of America										
38	482	Shaft	4000000	12121618	100	1971	TRUE	Adventure	R	6.6	106	United States of America										
39	32061	The Doberma	500000	5000000	87	1972	TRUE	Action	PG	5.5	20	United States of America										

CONCLUSION

We successfully cleaned the TMDB dataset, creating a success variable (revenue ≥ 2x budget) and identifying key features like budget, revenue, and vote count. The correlation analysis showed revenue (0.47) and vote count (0.24) as strong predictors of success, with budget (0.19) having a weaker link. This supports our logistic regression approach, and further analysis will refine these findings.

REFERENCES

DA, O., OM, S., AK, F., O, A., A, O., A, O., A, W., & M, Y. (2021). Movie success prediction using data mining. British Journal of Computer Networking and Information Technology, 4(2), 22–30. https://doi.org/10.52589/bjcnit-cqocirec

Zheng, Y. (2024). Predicting movie box office based on machine learning, deep learning, and statistical methods. Applied and Computational Engineering, 94(1), 20–32. https://doi.org/10.54254/2755-2721/94/2024melb0069

Subramaniyaswamy, V., Vaibhav, M. V., Prasad, R. V., & Logesh, R. (2017). Predicting movie box office success using multiple regression and SVM. Journal of Science & Statistics, 182–186. https://doi.org/10.1109/iss1.2017.8389394

Velingkar, G., Varadarajan, R., Lanka, S., & M, A. K. (2022). Movie Box-Office success Prediction using Machine Learning. 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), 9, 1–6. https://doi.org/10.1109/icpc2t53885.2022.9776798

Mr, N. V., Mr, P. M., & Pb, S. B. (2014). Predicting movie success based on IMDB data. International Journal of Business Intelligents, 3(2). https://doi.org/10.20894/ijbi.105.003.002.004

THANK YOU