Predicting the success of movies

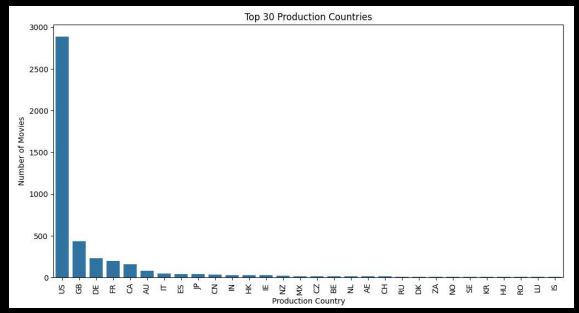
Made by:
Petra Nikolett T. Tóth

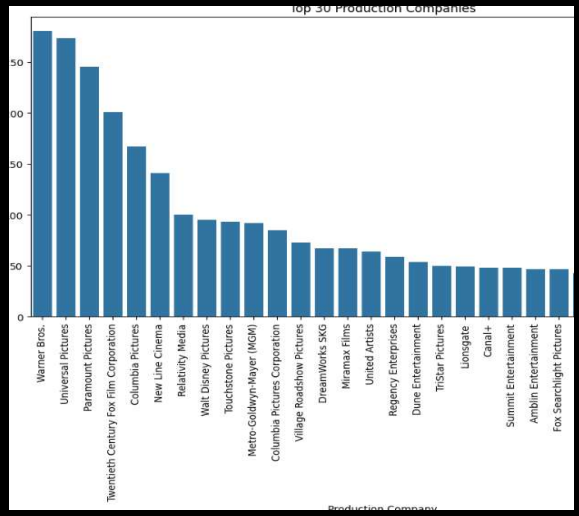
Chosen dataset: Movie ratings

Column1	Column2	Column6	Column9	Column10	Column11	Column12	Column13	Column14	Column15	Column19	Column20
budget	genres	original_language	popularity	production_companies	production_countries	release_date	revenue	runtime	spoken_languages	vote_average	vote_count
237000000	[{"id": 28, "nan	en	150.437577	[{"name": "Ingenious Filn	[{"iso_3166_1": "US", "n	2009-12-10	2787965087	162	[{"iso_639_1": "en", "r	7.2	11800
300000000	[{"id": 12, "nan	en	139.082615	[{"name": "Walt Disney P	[{"iso_3166_1": "US", "n	2007-05-19	961000000	169	[{"iso_639_1": "en", "r	6.9	4500
245000000	[{"id": 28, "nan	en	107.376788	[{"name": "Columbia Pic	[{"iso_3166_1": "GB", "n	2015-10-26	880674609	148	[{"iso_639_1": "fr", "n	6.3	4466
250000000	[{"id": 28, "nan	en	112.31295	[{"name": "Legendary Pic	[{"iso_3166_1": "US", "n	2012-07-16	1084939099	165	[{"iso_639_1": "en", "r	7.6	9106
260000000	[{"id": 28, "nan	en	43.926995	[{"name": "Walt Disney P	[{"iso_3166_1": "US", "n	2012-03-07	284139100	132	[{"iso_639_1": "en", "r	6.1	2124
258000000	[{"id": 14, "nan	en	115.699814	[{"name": "Columbia Pic	[{"iso_3166_1": "US", "n	2007-05-01	890871626	139	[{"iso_639_1": "en", "r	5.9	3576
260000000	[{"id": 16, "nan	en	48.681969	[{"name": "Walt Disney P	[{"iso_3166_1": "US", "n	2010-11-24	591794936	100	[{"iso_639_1": "en", "r	7.4	3330
280000000	[{"id": 28, "nan	en	134.279229	[{"name": "Marvel Studio	[{"iso_3166_1": "US", "n	2015-04-22	1405403694	141	[{"iso_639_1": "en", "r	7.3	6767
250000000	[{"id": 12, "nan	en	98.885637	[{"name": "Warner Bros."	[{"iso_3166_1": "GB", "n	2009-07-07	933959197	153	[{"iso_639_1": "en", "r	7.4	5293
250000000	[{"id": 28, "nan	en	155.790452	[{"name": "DC Comics", "	[{"iso_3166_1": "US", "n	2016-03-23	873260194	151	[{"iso_639_1": "en", "r	5.7	7004
270000000	[{"id": 12, "nan	en	57.925623	[{"name": "DC Comics", "	[{"iso_3166_1": "US", "n	2006-06-28	391081192	154	[{"iso_639_1": "en", "r	5.4	1400
200000000	[{"id": 12, "nan	en	107.928811	[{"name": "Eon Productic	[{"iso_3166_1": "GB", "n	2008-10-30	586090727	106	[{"iso_639_1": "en", "r	6.1	2965
200000000	[{"id": 12, "nan	en	145.847379	[{"name": "Walt Disney P	[{"iso_3166_1": "JM", "n	2006-06-20	1065659812	151	[{"iso_639_1": "en", "r	7.0	5246
255000000	[{"id": 28, "nan	en	49.046956	[{"name": "Walt Disney P	[{"iso_3166_1": "US", "n	2013-07-03	89289910	149	[{"iso_639_1": "en", "r	5.9	2311
225000000	[{"id": 28, "nan	en	99.398009	[{"name": "Legendary Pic	[{"iso_3166_1": "GB", "n	2013-06-12	662845518	143	[{"iso_639_1": "en", "r	6.5	6359
225000000	[{"id": 12, "nan	en	53.978602	[{"name": "Walt Disney",	[{"iso_3166_1": "CZ", "n	2008-05-15	419651413	150	[{"iso_639_1": "en", "r	6.3	1630
220000000	[{"id": 878, "na	en	144.448633	[{"name": "Paramount Pi	[{"iso_3166_1": "US", "n	2012-04-25	1519557910	143	[{"iso_639_1": "en", "r	7.4	11776
380000000	[{"id": 12, "nan	en	135.413856	[{"name": "Walt Disney P	[{"iso_3166_1": "US", "n	2011-05-14	1045713802	136	[{"iso_639_1": "en", "r	6.4	4948
225000000	[{"id": 28, "nan	en	52.035179	[{"name": "Amblin Entert	[{"iso_3166_1": "US", "n	2012-05-23	624026776	106	[{"iso_639_1": "en", "r	6.2	4160
250000000	[{"id": 28, "nan	en	120.965743	[{"name": "WingNut Film	[{"iso_3166_1": "NZ", "n	2014-12-10	956019788	144	[{"iso_639_1": "en", "r	7.1	4760
215000000	[{"id": 28, "nan	en	89.866276	[{"name": "Columbia Pic	[{"iso_3166_1": "US", "n	2012-06-27	752215857	136	[{"iso_639_1": "en", "r	6.5	6586
200000000	[{"id": 28, "nan	en	37.668301	[{"name": "Imagine Enter	[{"iso_3166_1": "GB", "n	2010-05-12	310669540	140	[{"iso_639_1": "en", "r	6.2	1398
250000000	[{"id": 12, "nan	en	94.370564	[{"name": "WingNut Film	[{"iso_3166_1": "NZ", "n	2013-12-11	958400000	161	[{"iso_639_1": "en", "r	7.6	4524
180000000	[{"id": 12, "nan	en	42.990906	[{"name": "New Line Cine	[{"iso_3166_1": "GB", "n	2007-12-04	372234864	113	[{"iso_639_1": "is", "n	5.8	1303

Cleaning the data

- excluding all unneccessary columns
- removing all rows with missing values (nothing, 0 or [])
- I checked if all columns were worth analysing, and removed the ones that seemed useless
- checked for weird values, removed the rows where I found any



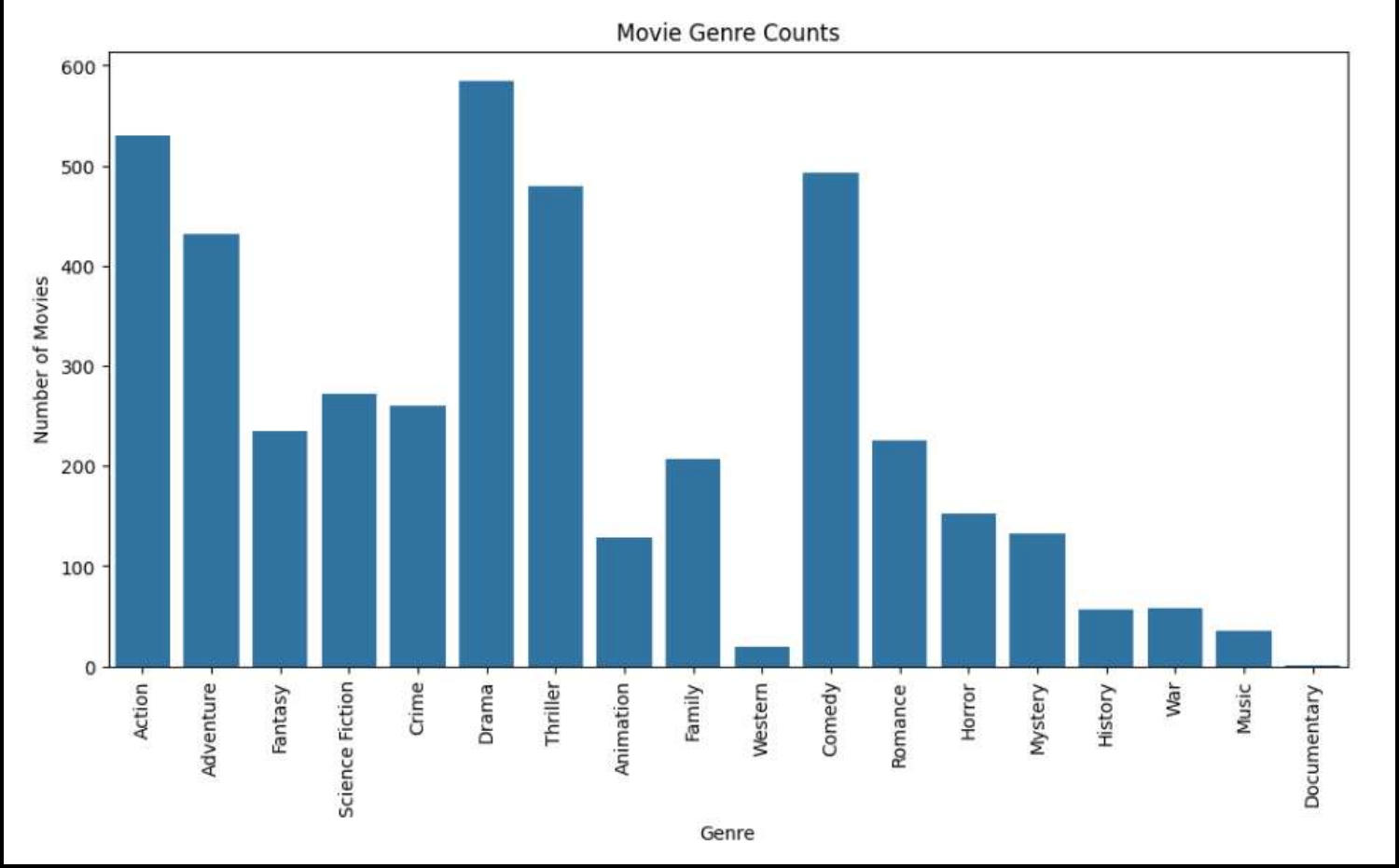


budget	1.000000
popularity	0.037073
revenue	5.000000
runtime	41.000000
vote_average	2.300000
vote_count	1.000000

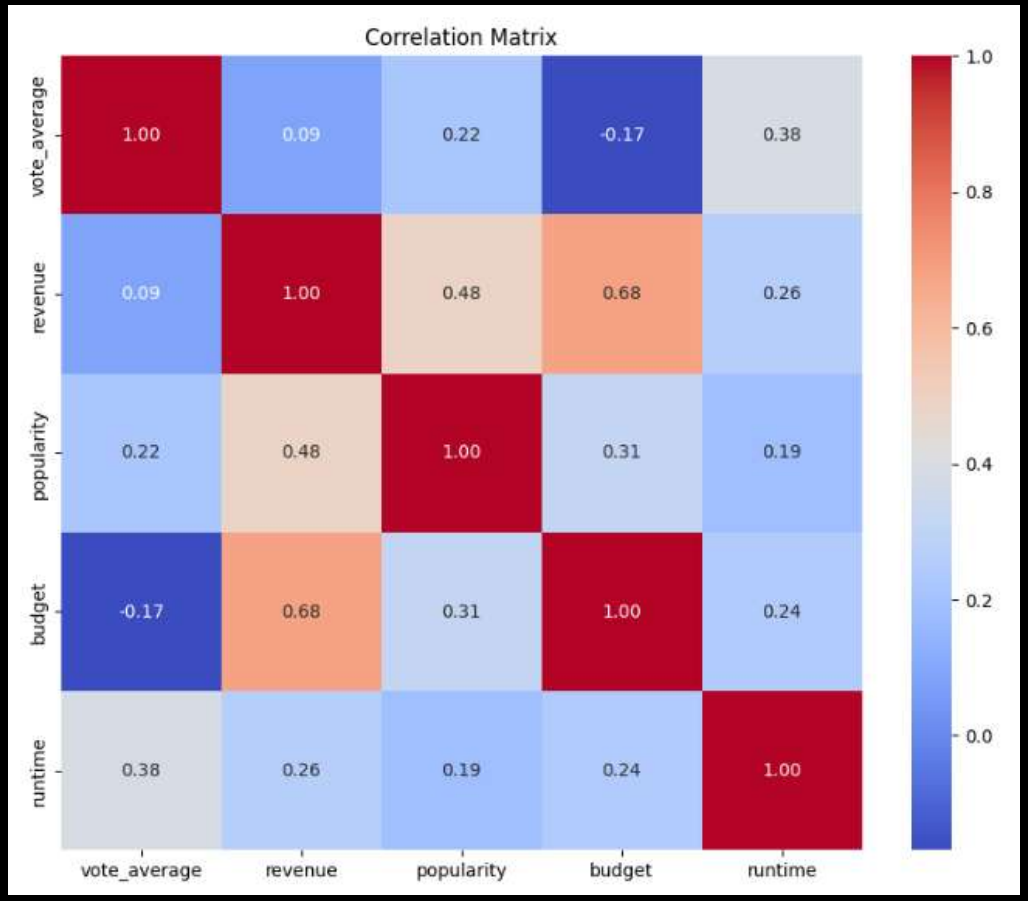
Transforming the genre column

ue	runtime	title	vote_average	vote_count	Action	...	Family	Western	Comedy	Romance	Horror	Mystery	History	War	Music	Documentary
87	162.0	Avatar	7.2	11800	1	...	0	0	0	0	0	0	0	0	0	0
00	169.0	Pirates of the Caribbean: At World's End	6.9	4500	1	...	0	0	0	0	0	0	0	0	0	0
09	148.0	Spectre	6.3	4466	1	...	0	0	0	0	0	0	0	0	0	0
99	165.0	The Dark Knight Rises	7.6	9106	1	...	0	0	0	0	0	0	0	0	0	0
00	132.0	John Carter	6.1	2124	1	...	0	0	0	0	0	0	0	0	0	0

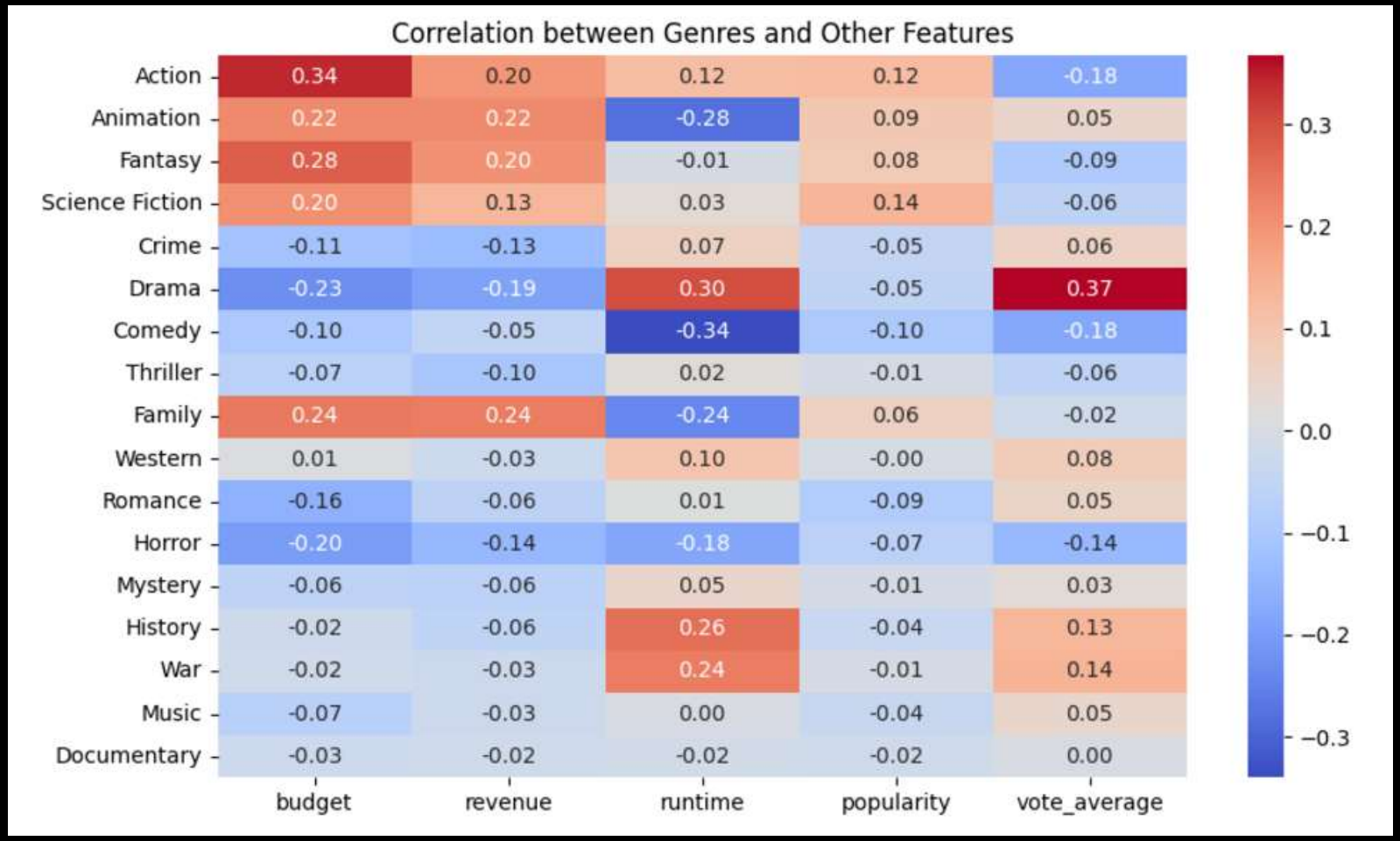
Genre counts



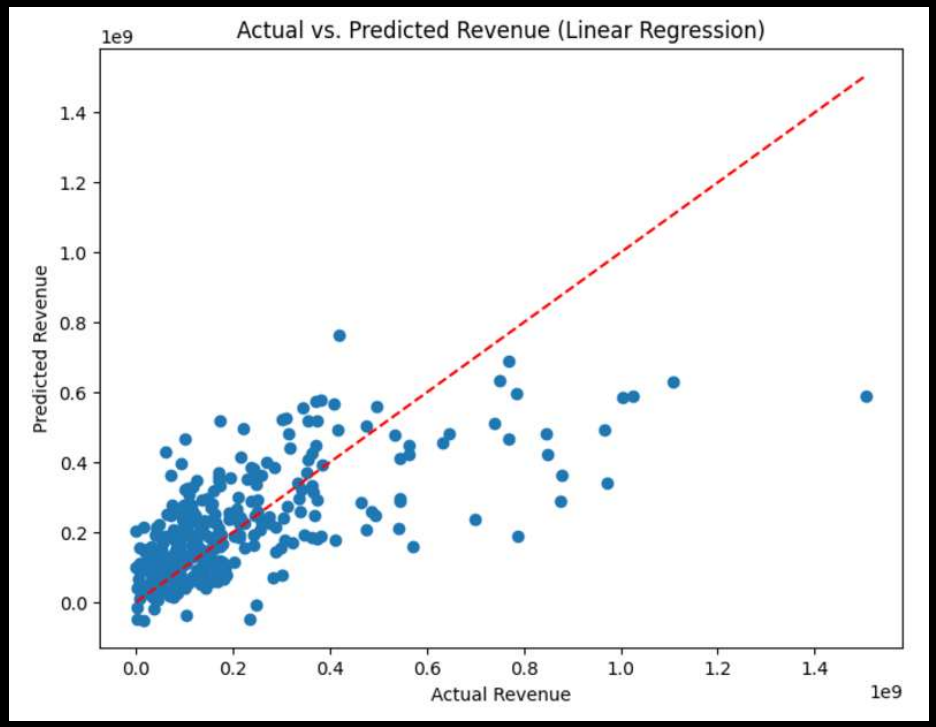
Checking correlations



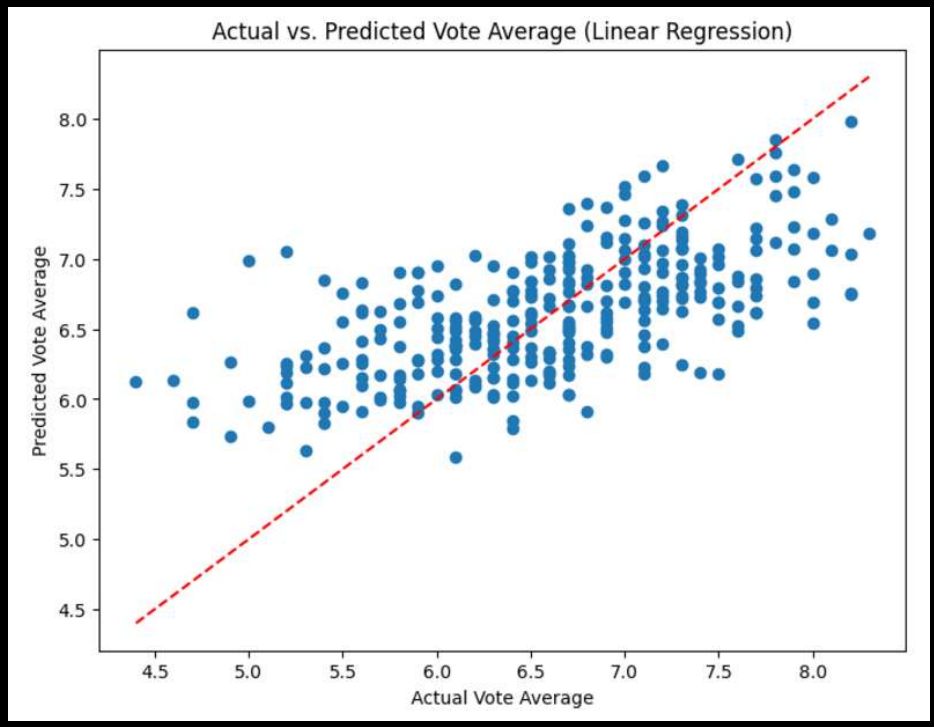
Checking correlations



Linear regression

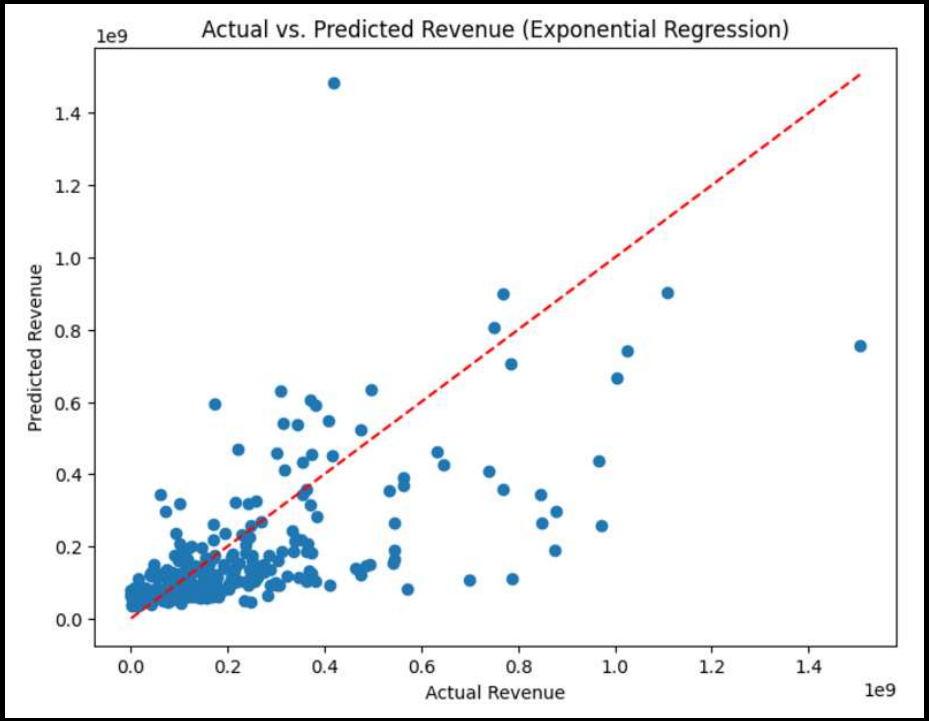


MSE: 2.5725146253851964e+16
R^2: 0.4523999014029635

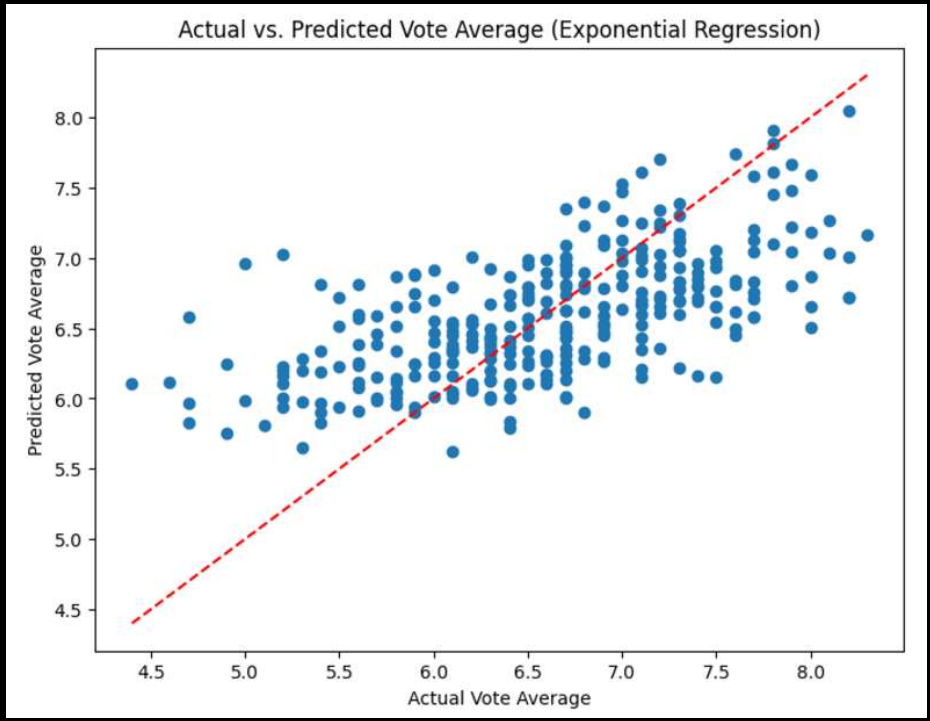


MSE: 0.37842152877483726
R^2: 0.3888472210721683

Exponential regression

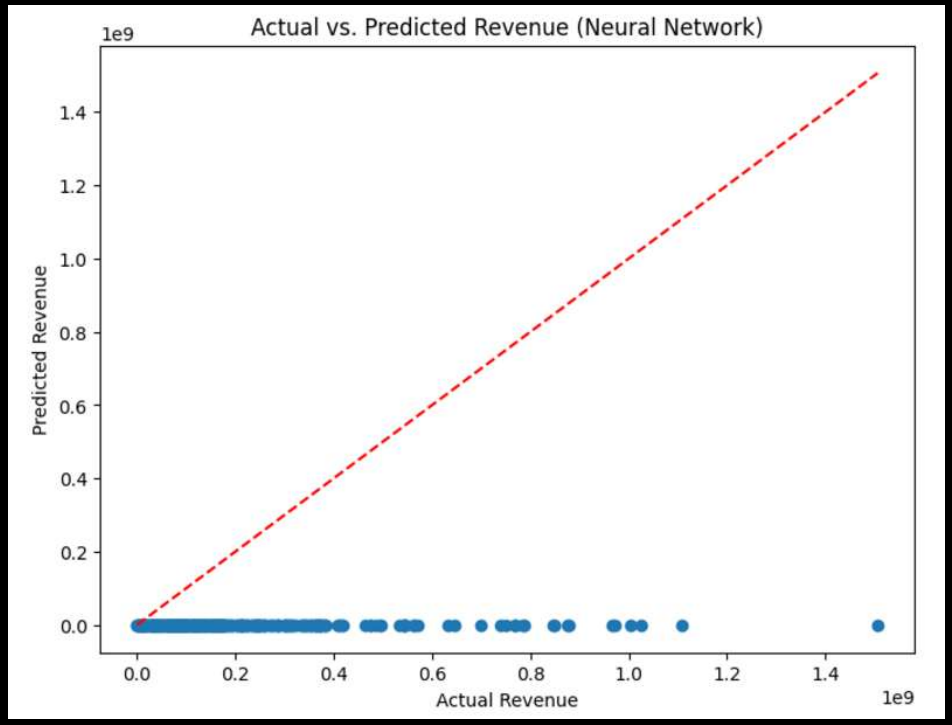


MSE: 2.945277666037281e+16
R^2: 0.37305143986259504

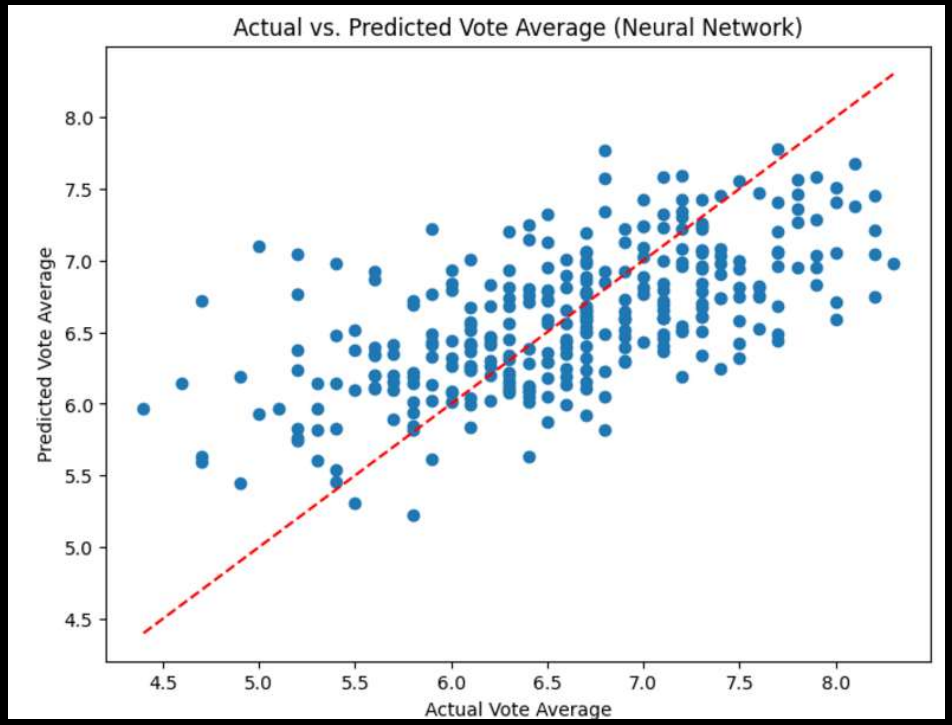


MSE: 0.3800860987194411
R^2: 0.3861589317703994

Neural network



MSE: 8.979925675450512e+16
R^2: -0.9115180969238281

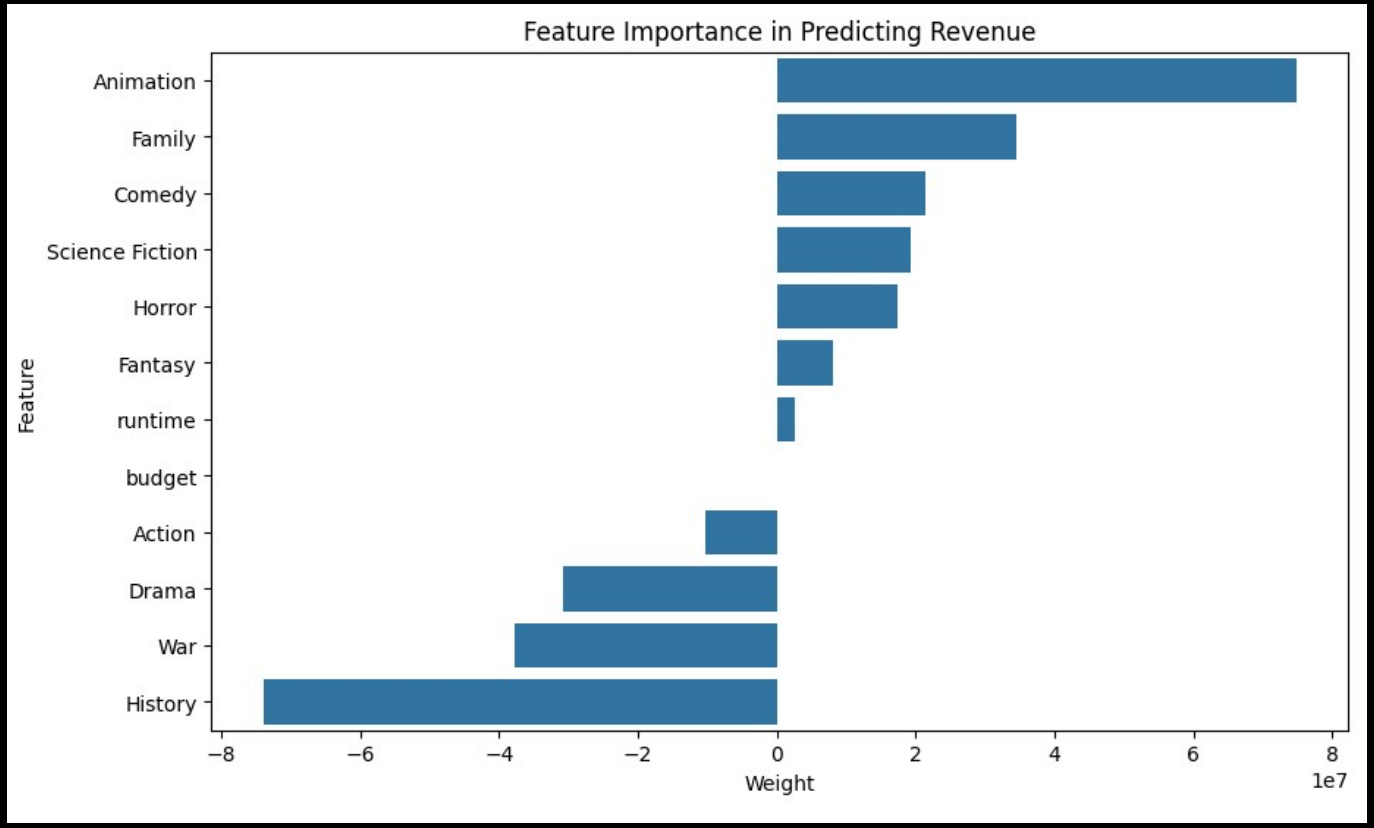


MSE: 0.3734845592289287 6
R^2: 0.39682045311113956

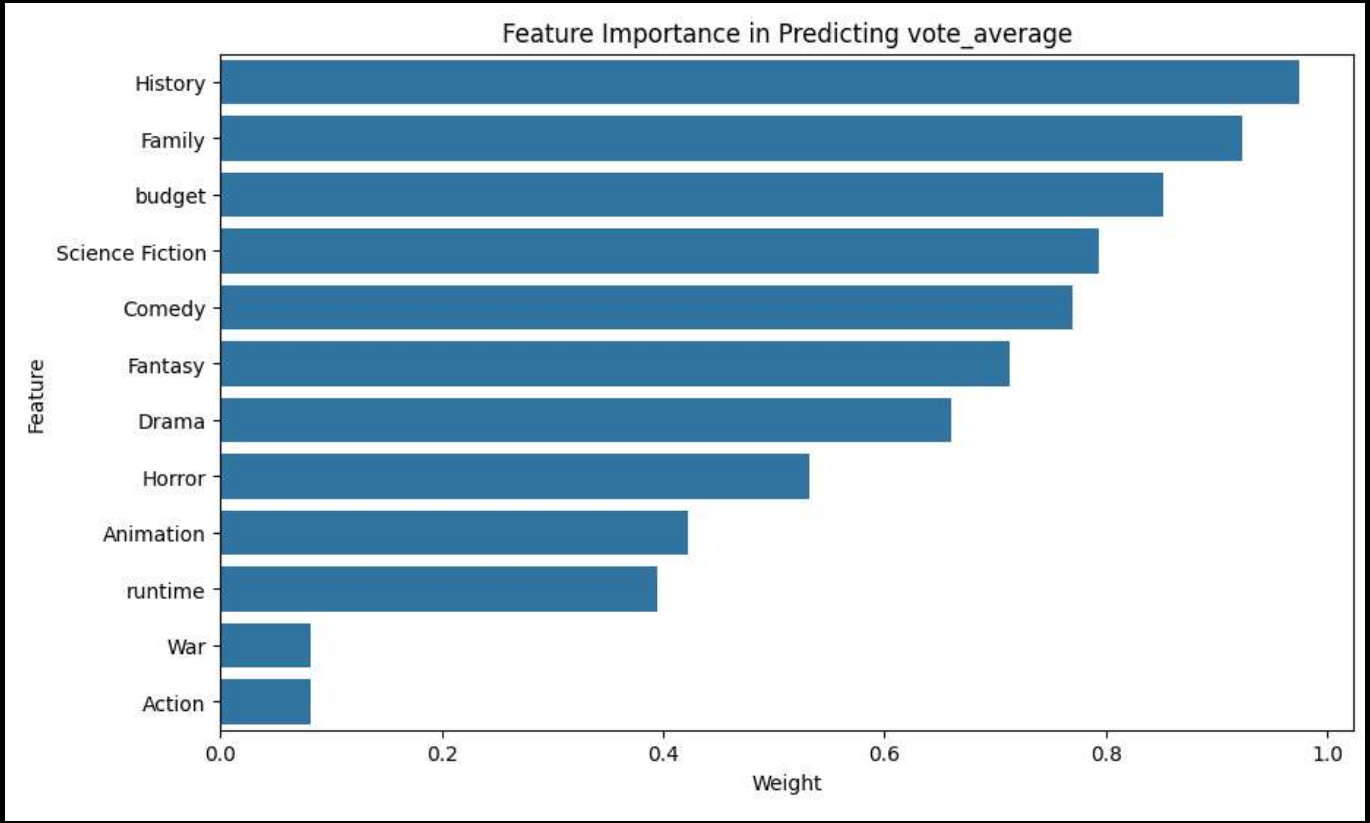
Evaluation

	Linear Regression	Exponential Regression	Neural Network
revenue MSE	***2.5725146253851964e+16***	2.9452776660372 81e+16	8.979925675450512e+16
revenue R^2	***0.4523999014029635***	0.37305143986259504	-0.911518096923 8281
vote_avarage MSE	0.37842152877483726	0.3800860987194411	***0.37348455922892876***
vote_avarage R^2	0.3888472210721683	0.3861589317703994	***0.39682045311113956***

Weights for Linear Regression predicting Revenue



Weights for Neural Network predicting Vote Avarage



Thanks for watching!