MOVIE DATABASE ANALYSIS

CENTRAL QUESTION

How do factors such as rating, domestic lifetime gross, foreign lifetime gross, and run-time affect a movie’s worldwide lifetime revenue (gross) and its ranking?

DATA ACQUISITION

How we scraped each dataset

BeautifulSoup4

- GET request, construct BS object
- Use find() and findall() method
- Construct dataframe from results

Discover API +Movie API

- Customized API keys by registration
- GET request, build dataframe from results
- Store results as JSON

Box Office Mojo

XPATH

- HTML Parser to get root
- Parse tree following path "*body/div//table*"
- Build LoL, then convert to dataframe

IMDB

	Title	Release Year	Rating
0	The Shawshank Redemption	1994	9.2
1	The Godfather	1972	9.2
2	The Dark Knight	2008	9.0
3	The Godfather Part II	1974	9.0
4	12 Angry Men	1957	9.0
5	Schindler's List	1993	8.9
6	The Lord of the Rings: The Return of the King	2003	8.9
7	Pulp Fiction	1994	8.8
8	The Lord of the Rings: The Fellowship of the Ring	2001	8.8
9	The Good, the Bad and the Ugly	1966	8.8
10	Forrest Gump	1994	8.8
11	Fight Club	1999	8.7
12	The Lord of the Rings: The Two Towers	2002	8.7
13	Inception	2010	8.7
14	Star Wars: Episode V - The Empire Strikes Back	1980	8.7
15	The Matrix	1999	8.7
16	Goodfellas	1990	8.7
17	One Flew Over the Cuckoo's Nest	1975	8.6
18	Se7en	1995	8.6
19	Seven Samurai	1954	8.6

TMDB

	page	results	total_pages	total_results
0	1	{'adult': False, 'backdrop_path': '/7ABsaBkO1j...	18889	377775
1	1	{'adult': False, 'backdrop_path': '/14QbnygCuT...	18889	377775
2	1	{'adult': False, 'backdrop_path': '/3rCuqnCQP7...	18889	377775
3	1	{'adult': False, 'backdrop_path': '/7ysBoVZhLP...	18889	377775
4	1	{'adult': False, 'backdrop_path': None, 'genre...	18889	377775
5	1	{'adult': False, 'backdrop_path': None, 'genre...	18889	377775
6	1	{'adult': False, 'backdrop_path': '/geYUecpFI2...	18889	377775
7	1	{'adult': False, 'backdrop_path': '/qjGrUmKW78...	18889	377775
8	1	{'adult': False, 'backdrop_path': '/3G1Q5xF40H...	18889	377775
9	1	{'adult': False, 'backdrop_path': '/jlGmlFOcfo...	18889	377775
10	1	{'adult': False, 'backdrop_path': '/uEwGFGtao9...	18889	377775
11	1	{'adult': False, 'backdrop_path': '/9e6wp707XM...	18889	377775
12	1	{'adult': False, 'backdrop_path': '/qkt1Qn9j6y...	18889	377775
13	1	{'adult': False, 'backdrop_path': '/mDfJG3LC3D...	18889	377775
14	1	{'adult': False, 'backdrop_path': '/b6ZJZHUdME...	18889	377775
15	1	{'adult': False, 'backdrop_path': '/1stUlsjawR...	18889	377775
16	1	{'adult': False, 'backdrop_path': '/urDWNffjwm...	18889	377775
17	1	{'adult': False, 'backdrop_path': '/cugmVwK0N4...	18889	377775
18	1	{'adult': False, 'backdrop_path': '/h6wflubjTv...	18889	377775
19	1	{'adult': False, 'backdrop_path': '/5rrGVmRUuC...	18889	377775

TMDB

	adult	backdrop_path	genre_ids	id	original_language	original_title	overview	popularity	poster_path	release_date	title	video	vote_average	vote_count
0	False	/7ABsaBkO1jA2psC8Hy4IDhkID4h.jpg	[28, 12, 14, 878]	19995	en	Avatar	In the 22nd century, a paraplegic Marine is di...	699.520	/jRXYjXNq0Cs2TcJjLkki24MLp7u.jpg	2009-12-15	Avatar	False	7.5	26715
1	False	/14QbnygCuTO0vl7CAFmPf1fgZfV.jpg	[28, 12, 878]	634649	en	Spider-Man: No Way Home	Peter Parker is unmasked and no longer able to...	416.358	/uJYYizSuA9Y3DCs0qS4qWvHfZg4.jpg	2021-12-15	Spider-Man: No Way Home	False	8.0	15998
2	False	/3rCuqnCQP7tZJ1rqqzSwxCzcW0w.jpg	[18, 10749]	247136	ja	M家の新妻　変態洗礼	Mikage will get married to Youiti next year, s...	359.864	/2oVfD5rUV2EElbQ11ds2Vf5nRaZ.jpg	2009-03-27	The Temptation of Kimono	False	5.4	7
3	False	/7ysBoVZhLPipvyZ8gyS9qvnPjUc.jpg	[18]	795514	en	The Fallout	In the wake of a school tragedy, Vada, Mia and...	331.519	/4ByHl9XRKR2iXbvF0ZilMRD1RcL.jpg	2021-03-17	The Fallout	False	7.5	425
4	False	None	[10749]	485470	ko	착한 형수2	If you give it once, a good brother-in-law who...	327.341	/3pEs4hmeHvTAsmx09whEaPDOQpq.jpg	2017-10-08	Nice Sister-In-Law 2	False	6.0	2
5	False	None	[27]	888838	en	The Long Dark Trail	After two impoverished teenage brothers manage...	325.087	/ebdDGnqQXDGfiggHSazaWCLF6Lf.jpg	2021-06-19	The Long Dark Trail	False	5.4	7
6	False	/geYUecpFl2AonDLhjyK9zoVFcMv.jpg	[16, 28, 14]	810693	ja	劇場版 呪術廻戦 0	Yuta Okkotsu is a nervous high school student ...	281.588	/3pTwMUEavTzVOh6yLN0aEwR7uSy.jpg	2021-12-24	Jujutsu Kaisen 0	False	8.3	686
7	False	/qjGrUmKW78MCFG8PTLDBp67S27p.jpg	[16, 28, 12, 14]	635302	ja	劇場版「鬼滅の刃」無限列車編	Tanjirō Kamado, joined with Inosuke Hashibira,...	247.391	/h8Rb9gBr48ODIwYUttZNYeMWeUU.jpg	2020-10-16	Demon Slayer -Kimetsu no Yaiba- The Movie: Mug...	False	8.3	2807
8	False	/3G1Q5xF40HkUBJXxt2DQgQzKTp5.jpg	[16, 35, 10751, 14]	568124	en	Encanto	The tale of an extraordinary family, the Madri...	234.609	/4j0PNHkMr5ax3IA8tjtxcmPU3QT.jpg	2021-10-13	Encanto	False	7.7	7634
9	False	/jlGmlFOcfo8n5tURmhC7YVd4Iyy.jpg	[28, 35, 12]	436969	en	The Suicide Squad	Supervillains Harley Quinn, Bloodsport, Peacem...	228.543	/kb4s0ML0iVZlG6wAKbbs9NAm6X.jpg	2021-07-28	The Suicide Squad	False	7.6	6783
10	False	/uEwGFGtao9YG2JolmdvtHLLVbA9.jpg	[99]	111332	en	Avatar: Creating the World of Pandora	The Making-of James Cameron's Avatar. It shows...	218.444	/sjf3xjuofCtDhZghJRzXlTiEjJe.jpg	2010-02-07	Avatar: Creating the World of Pandora	False	7.4	28
11	False	/9e6wp707XMouPG939o2fHunXXJR.jpg	[16, 12, 35, 10751, 14, 27]	639721	en	The Addams Family 2	The Addams get tangled up in more wacky advent...	215.800	/ld7YB9vBRp1GM1DT3KmFWSmtBPB.jpg	2021-10-01	The Addams Family 2	False	7.1	1004
12	False	/qkt1Qn9j6yw9rcJhvSu1p3wuiBm.jpg	[10751, 35, 14]	8871	en	How the Grinch Stole Christmas	Inside a snowflake exists the magical land of ...	211.736	/1WZbbPApEivA421gCOluuzMMKCk.jpg	2000-11-15	How the Grinch Stole Christmas	False	6.7	6201
13	False	/mDfJG3LC3Dqb67AZ52x3Z0jU0uB.jpg	[12, 28, 878]	299536	en	Avengers: Infinity War	As the Avengers and their allies have continue...	195.636	/7WsyChQLEftFiDOVTGkv3hFpyyt.jpg	2018-04-25	Avengers: Infinity War	False	8.3	25833
14	False	/b6ZJZHUdMEFECvGiDpJjlfUWela.jpg	[28, 12, 878]	284054	en	Black Panther	King T'Challa returns home to the reclusive, t...	195.461	/uxzzxijgPIY7slzFvMotPv8wjKA.jpg	2018-02-13	Black Panther	False	7.4	19932
15	False	/1stUIsjawROZxjiCMtqqXqgfZWC.jpg	[12, 14]	672	en	Harry Potter and the Chamber of Secrets	Cars fly, trees fight back, and a mysterious h...	194.601	/sdEOH0992YZ0QSxgXNIGLq1ToUi.jpg	2002-11-13	Harry Potter and the Chamber of Secrets	False	7.7	18992
16	False	/urDWNffjwmNi5lQaezw9GwqkUXa.jpg	[12, 14]	767	en	Harry Potter and the Half-Blood Prince	As Lord Voldemort tightens his grip on both th...	192.279	/z7uo9zmQdQwU5ZJHFpv2Upl30i1.jpg	2009-07-07	Harry Potter and the Half-Blood Prince	False	7.7	16840
17	False	/cugmVwK0N4aAcLibeIKN5jWDXSx.jpg	[16, 28, 12, 878]	768744	ja	僕のヒーローアカデミア THE MOVIE ワールド ヒーローズ ミッション	A mysterious group called Humarize strongly be...	182.478	/AsTlA7dj2ySGY1pzGSD0MoHFhEF.jpg	2021-08-06	My Hero Academia: World Heroes' Mission	False	7.6	319
18	False	/h6wflubjTvlYgFUfcRKmZOPBPzM.jpg	[878, 28, 12]	91314	en	Transformers: Age of Extinction	As humanity picks up the pieces, following the...	181.647	/7JV3srXxr4fOY58vvGeBRTSGOuF.jpg	2014-06-25	Transformers: Age of Extinction	False	5.9	6997
19	False	/5rrGVmRUuCKVbqUu41XlWTXJmNA.jpg	[12, 14, 10751]	674	en	Harry Potter and the Goblet of Fire	When Harry Potter's name emerges from the Gobl	177.553	/fECBtHlr0RB3foNHDiCBXeg9Bv9.jpg	2005-11-16	Harry Potter and the Goblet of Fire	False	7.8	17971

TMDB

	title	revenue	runtime
0	Titanic	2187463944	194
1	Star Wars: Episode I - The Phantom Menace	924317558	136
2	Jurassic Park	920100000	127
3	Independence Day	817400891	145
4	E.T. the Extra-Terrestrial	792965500	115
...	...	...	...
4489	Alvarez Kelly	0	106
4490	And the Ship Sails On	0	132
4491	Gulliver's Travels	0	76
4492	Small Wonders	0	77
4493	Small Faces	0	108

4494 rows × 3 columns

BOX OFFICE MO JO

Rank	Title	Worldwide Lifetime Gross	Domestic Lifetime Gross	Domestic	Foreign Lifetime Gross	Foreign	Year
1	Avatar	2922917914	785221649.0	26.9	2137696265	73.1	2009
2	Avengers: Endgame	2797501328	858373000.0	30.7	1939128328	69.3	2019
3	Titanic	2201647264	659363944.0	30.0	1542283320	70.0	1997
4	Star Wars: Episode VII - The Force Awakens	2069521700	936662225.0	45.3	1132859475	54.7	2015
5	Avengers: Infinity War	2048359754	678815482.0	33.1	1369544272	66.9	2018
6	Spider-Man: No Way Home	1916306995	814115070.0	42.5	1102191925	57.5	2021
7	Jurassic World	1671537444	653406625.0	39.1	1018130819	60.9	2015
8	The Lion King	1663250487	543638043.0	32.7	1119612444	67.3	2019
9	The Avengers	1518815515	623357910.0	41.0	895457605	59.0	2012
10	Furious 7	1515341399	353007020.0	23.3	1162334379	76.7	2015
11	Top Gun: Maverick	1487994195	717994195.0	48.2	770000000	51.8	2022
12	Frozen II	1450026933	477373578.0	32.9	972653355	67.1	2019
13	Avengers: Age of Ultron	1402809540	459005868.0	32.7	943803672	67.3	2015
14	Black Panther	1382248826	700426566.0	50.7	681822260	49.3	2018
15	Harry Potter and the Deathly Hallows: Part 2	1342359942	381447587.0	28.4	960912355	71.6	2011
16	Star Wars: Episode VIII - The Last Jedi	1332698830	620181382.0	46.5	712517448	53.5	2017
17	Jurassic World: Fallen Kingdom	1310466296	417719760.0	31.9	892746536	68.1	2018
18	Beauty and the Beast	1305611599	504481165.0	38.6	801130434	61.4	2017
19	Frozen	1304550716	400953009.0	30.7	903597707	69.3	2013
20	Incredibles 2	1243089244	608581744.0	49.0	634507500	51.0	2018

DATABASE DESIGN

Movie.db

IMDB

- Title
- Release Year
- Rating

TMDB

- Title
- Revenue
- Run-time

BoxOffice

- Rank
- Title
- World Lifetime Gross
- Domestic Lifetime Gross
- Domestic(%)
- Foreign Lifetime Gross
- Foreign(%)

CHALLENGES

in data acquisition and processing

Convert into list of tuples to be used in SQL

Integer types interpreted as BLOB (Binary Large Object) that cannot be read in SQLite

Long request time when scraping data

Requires permission when accessing TMDB

FINDINGS + KEY INSIGHTS

Fig1. Movie Ratings Distribution

Fig2. Runtime vs Revenue

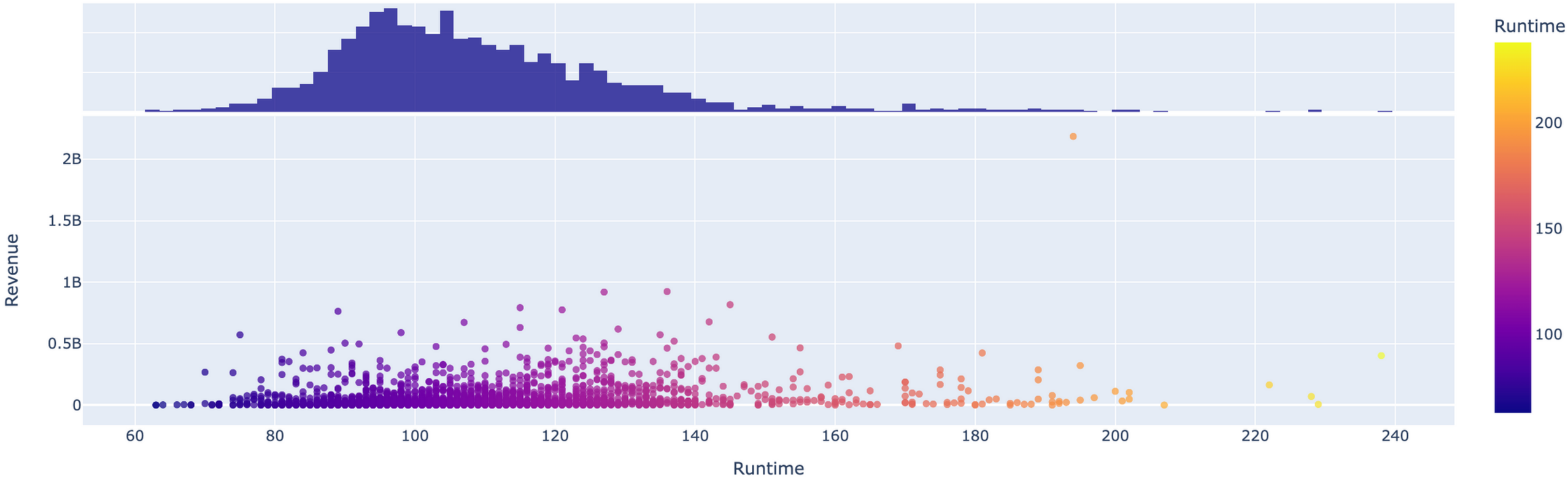
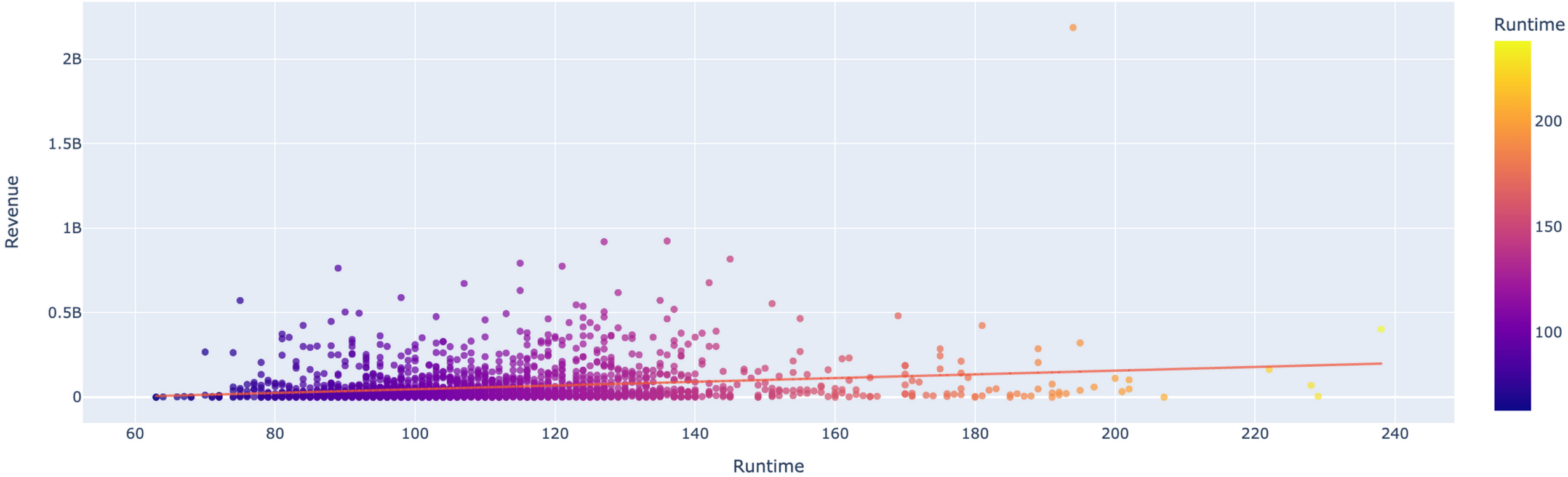


Fig3. Runtime vs Revenue



	px_fit_results
0	<statsmodels.regression.linear_model.Regressio...

OLS Regression Results

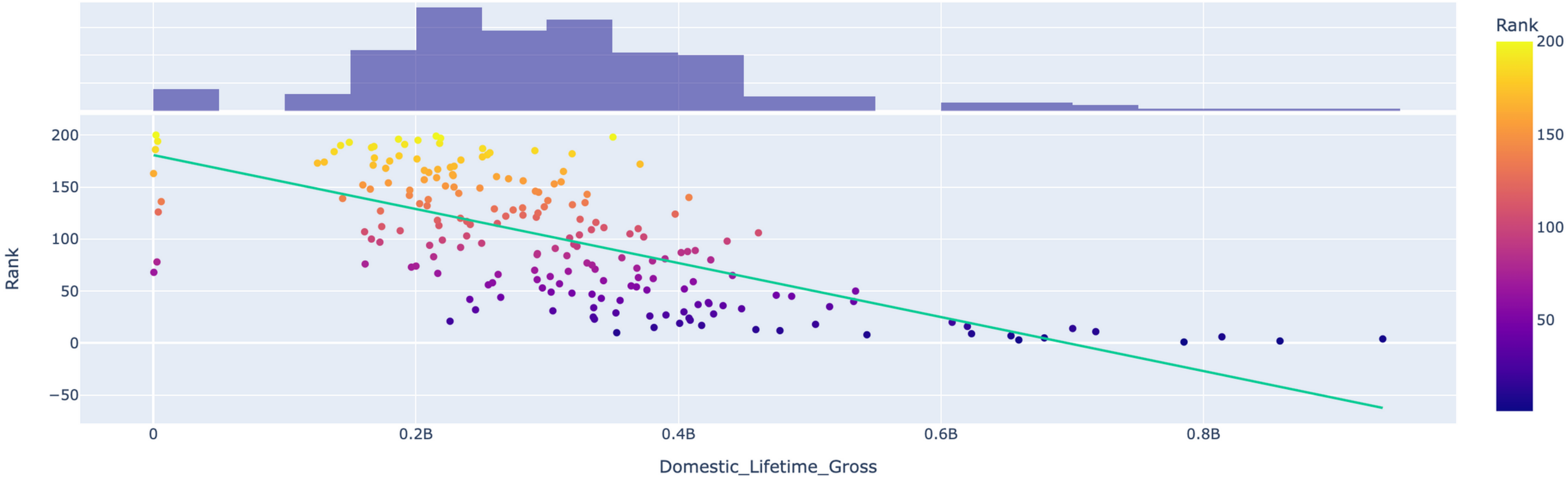
Dep. Variable:	y	**R-squared:**	0.046
Model:	OLS	**Adj. R-squared:**	0.045
Method:	Least Squares	**F-statistic:**	110.6
Date:	Sun, 11 Dec 2022	**Prob (F-statistic):**	2.69e-25
Time:	00:07:11	**Log-Likelihood:**	-45797.
No. Observations:	2302	**AIC:**	9.160e+04
Df Residuals:	2300	**BIC:**	9.161e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-6.247e+07	1.16e+07	-5.382	0.000	-8.52e+07	-3.97e+07
x1	1.098e+06	1.04e+05	10.517	0.000	8.93e+05	1.3e+06

Omnibus:	2666.817	**Durbin-Watson:**	1.188
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	491078.277
Skew:	5.708	**Prob(JB):**	0.00
Kurtosis:	73.637	**Cond. No.**	586.

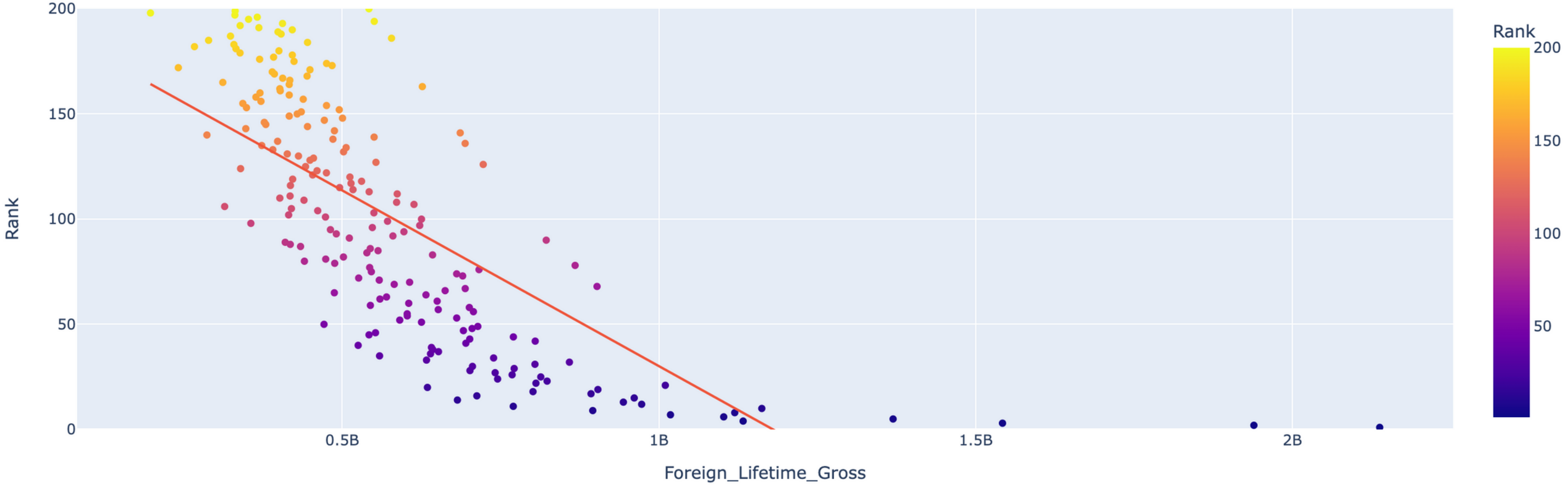
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fig4. Domestic Lifetime Gross vs Ranking



R squared = 0.46

Fig5. Foreign Lifetime Gross vs Ranking



R squared = 0.52