Final Report

SI 206: Introduction to Python

Kally Van, Angel Nguyen, Nina Wang

December 16, 2024

1. Goals for the Project

Initially, our project aimed to explore correlations between music popularity, streaming behavior, and regional user trends. However, difficulties in finding effective APIs and relevant datasets prompted us to change our objectives. We decided to analyze whether the success of the action movie genre—measured by ratings, box office revenue, and regional variations—is influenced by factors such as population density or budget.

Data we planned to gather:

- TMDB API (The Movie Database): Box office revenue, budget, and regional data
- OMDB API (The Open Movie Database): Rotten Tomatoes scores and box office revenue
- U.S. Census Bureau Data: Population data for U.S. regions (specifically looking at Midwest, South, Northeast, and West)

Using this new data, we plan to analyze specific trends in action movies, identify regional differences, and see potential correlations with demographic data.

Hypothesis: How does the action movie genre and box office revenue vary across different regions

2. Goals Achieved and Data Gathered

We successfully gathered and analyzed data from the following API sources and here is what they provided us:

- TMDB API: movie titles, ratings, revenue, budget, and regional data
- OMDB API: Rotten Tomatoes scores and box office revenue
- U.S. Census Bureau: population data for U.S. regions (Midwest, South, Northeast, and West) in 2021

Visualizations:

1. Financial Trends Chart (financial_trends.png):
    - Shows a decline in average movie revenue over time, even as total revenue increase
    - This decline suggests that there may have been an oversaturation of action movie production over time, which dilutes individual movie earnings

2. Ratings Bar Chart (ratings_bar_chart.png):
    - Displays average Rotten Tomatoes scores across regions, weighted by population percentages
    - The Northeast had the highest ratings, while the South had the lowest despite having the highest share (38.2%). Coastal regions like the Northeast and West have higher Rotten Tomatoes ratings (73.2% and 71.1%) despite fewer movies, while the South, with the largest population and most movies, has the lowest ratings (66.2%). This shows more movies don't always mean better ratings. The Midwest, with 17.3% of the population, has fewer movies, suggesting room for growth in that region.
3. Ratings Heatmap (ratings_heatmap.png):
    - Across all regions, there is a higher concentration that rated movies as "Good" (7.0–7.9/10), with the South showing the greatest variance. We also see that all regions don't have a strong preference of rating action movies "excellent" or "below average".
    - The South also had the largest proportion of "Good" ratings, highlighting regional rating differences but could be attributed to the South having a higher population density.
4. Revenue Pie Chart (revenue_pie_chart.png):
    - The pie chart reflects the estimated revenue each region would proportionally contribute to the total action movie revenue.
    - Across all regions, the South contributed the highest share of action movie revenue (38.2%). The Northeast contributed the least (17.3%).
    - We can attribute this to the proportion population each region has to the total population and can estimate that the distribution of each region would be roughly equal to proportion that region contributes to the total revenue of action movies.
5. Revenue Scatter Plot (revenue_scatter.png)
    - Demonstrates a positive correlation between TMDB ratings and revenue along with the movie budget.
    - The scatter plot demonstrates the relationship between a movie's TMDB rating and its revenue, with the color gradient of each point representing the movie's budget (yellower/greener for higher budgets, purple for lower budgets).
    - The visualization reveals that while higher-rated movies tend to earn more revenue, especially when combined with larger budgets, there's a variation in financial success across all rating levels, suggesting that a movie's commercial performance is influenced by other factors beyond just ratings and budget.

Our visualizations reveal that action movie success varies notably across U.S. regions, with audiences in the Northeast and West giving higher average ratings (7.32 and 7.11) compared to the South and Midwest

(6.62 and 6.83). While movie distribution naturally follows population size (South leading with 38.2% of movies), the financial success doesn't strictly follow this pattern, as the Northeast and West show stronger per-capita revenue despite smaller populations. The data suggests that both audience reception and box office performance of action movies are influenced by regional factors, though higher budgets and better ratings generally predict stronger financial outcomes across all regions.

3. Problems Faced

Throughout the project, we encountered several challenges:

1. Initial Data Challenges:
    - Our original goal of analyzing music popularity proved difficult due to limited meaningful connections between available datasets and APIs.
2. Scope Adjustments:
    - Narrowing our focus to action movies was necessary for a manageable analysis. Initially, we aimed to include multiple movie genres but decided that focusing on one genre would provide deeper insights.
3. API Key and Data Access Issues:
    - TheNumbers API: Access was unavailable, so we substituted overlapping data from TMDB.
    - Geonames: Web scraping challenges prompted us to switch to U.S. Census Bureau data for population metrics.
    - Trackt API: The client and secret keys provided were invalid, leading us to use OMDB for ratings and box office information.

Despite these obstacles, we successfully adapted our approach by using alternative data sources.

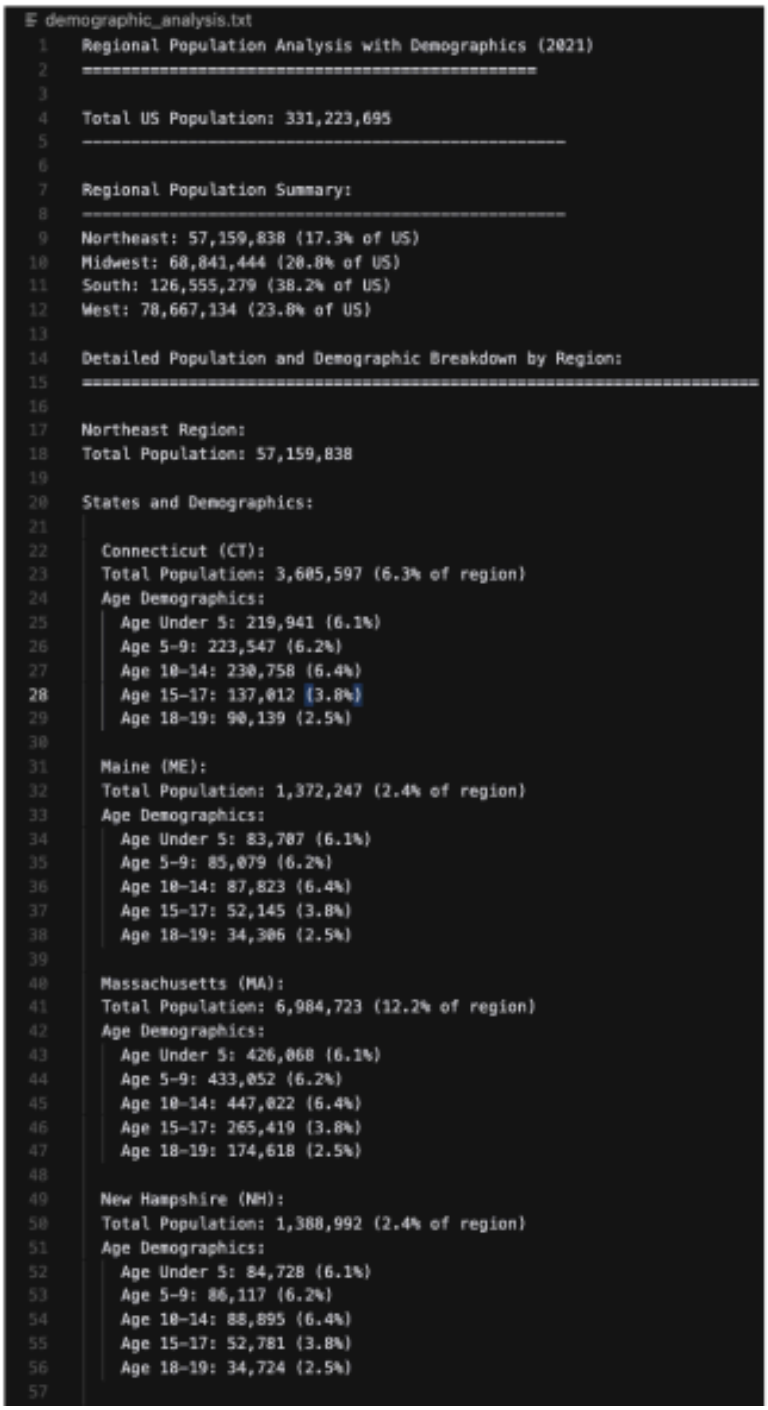
4. Calculations

Calculation 1: population_analysis.txt displays the total population split between 4 regions, each state in the regions, and the population in each state from Census Bureau API

population_analysis.txt

```
Regional Population Analysis (2021)
==================================

Population Statistics by Region:
--------------------------------------------------


Northeast Region:
Total Population: 57,159,838
States:
  Connecticut (CT): 3,605,597
  Maine (ME): 1,372,247
  Massachusetts (MA): 6,984,723
  New Hampshire (NH): 1,388,992
  New Jersey (NJ): 9,267,130
  New York (NY): 19,835,913
  Pennsylvania (PA): 12,964,056
  Rhode Island (RI): 1,095,610
  Vermont (VT): 645,570

Midwest Region:
Total Population: 68,841,444
States:
  Illinois (IL): 12,671,469
  Indiana (IN): 6,805,985
  Iowa (IA): 3,193,079
  Kansas (KS): 2,934,582
  Michigan (MI): 10,050,811
  Minnesota (MN): 5,707,390
  Missouri (MO): 6,168,187
  Nebraska (NE): 1,963,692
  North Dakota (ND): 774,948
  Ohio (OH): 11,780,017
  South Dakota (SD): 895,376
  Wisconsin (WI): 5,895,908

South Region:
Total Population: 126,555,279
States:
  Alabama (AL): 5,039,877
  Arkansas (AR): 3,025,891
  Delaware (DE): 1,003,384
  Florida (FL): 21,781,128
  Georgia (GA): 10,799,566
  Kentucky (KY): 4,509,394
  Louisiana (LA): 4,624,047
  Maryland (MD): 6,165,129
  Mississippi (MS): 2,949,965
  North Carolina (NC): 10,551,162
  Oklahoma (OK): 3,986,639
  South Carolina (SC): 5,190,705
  Tennessee (TN): 6,975,218
  Texas (TX): 29,527,941
  Virginia (VA): 8,642,274
  West Virginia (WV): 1,782,959

West Region:
Total Population: 78,667,134
States:
  Alaska (AK): 732,673
  Arizona (AZ): 7,276,316
  California (CA): 39,237,836
  Colorado (CO): 5,812,069
  Hawaii (HI): 1,441,553
  Idaho (ID): 1,900,923
  Montana (MT): 1,104,271
  Nevada (NV): 3,143,991
  New Mexico (NM): 2,115,877
  Oregon (OR): 4,246,155
  Utah (UT): 3,337,975
  Washington (WA): 7,738,692
  Wyoming (WY): 578,803

Total US Population: 331,223,695

Additional Statistics:
--------------------------------------------------
Northeast Population Percentage: 17.3%
Midwest Population Percentage: 20.8%
South Population Percentage: 38.2%
West Population Percentage: 23.8%
```

Calculation 2: demographic_analysis.txt displays the population and demographic breakdown between 4 regions and states (population summaries, age breakdowns between <5 - 19 years old)



demographic_analysis.txt

```
Regional Population Analysis with Demographics (2021)
==========================================

Total US Population: 331,223,695
------------------------------------------------

Regional Population Summary:
------------------------------------------------
Northeast: 57,159,838 (17.3% of US)
Midwest: 68,841,444 (20.8% of US)
South: 126,555,279 (38.2% of US)
West: 78,667,134 (23.8% of US)

Detailed Population and Demographic Breakdown by Region:
================================================================

Northeast Region:
Total Population: 57,159,838

States and Demographics:

  Connecticut (CT):
  Total Population: 3,605,597 (6.3% of region)
  Age Demographics:
    Age Under 5: 219,941 (6.1%)
    Age 5-9: 223,547 (6.2%)
    Age 10-14: 230,758 (6.4%)
    Age 15-17: 137,012 (3.8%)
    Age 18-19: 90,139 (2.5%)

  Maine (ME):
  Total Population: 1,372,247 (2.4% of region)
  Age Demographics:
    Age Under 5: 83,707 (6.1%)
    Age 5-9: 85,079 (6.2%)
    Age 10-14: 87,823 (6.4%)
    Age 15-17: 52,145 (3.8%)
    Age 18-19: 34,306 (2.5%)

  Massachusetts (MA):
  Total Population: 6,984,723 (12.2% of region)
  Age Demographics:
    Age Under 5: 426,068 (6.1%)
    Age 5-9: 433,052 (6.2%)
    Age 10-14: 447,022 (6.4%)
    Age 15-17: 265,419 (3.8%)
    Age 18-19: 174,618 (2.5%)

  New Hampshire (NH):
  Total Population: 1,388,992 (2.4% of region)
  Age Demographics:
    Age Under 5: 84,728 (6.1%)
    Age 5-9: 86,117 (6.2%)
    Age 10-14: 88,895 (6.4%)
    Age 15-17: 52,781 (3.8%)
    Age 18-19: 34,724 (2.5%)
```

Database Screenshots

Movies

Movie_ratings: TMDB & OMDB

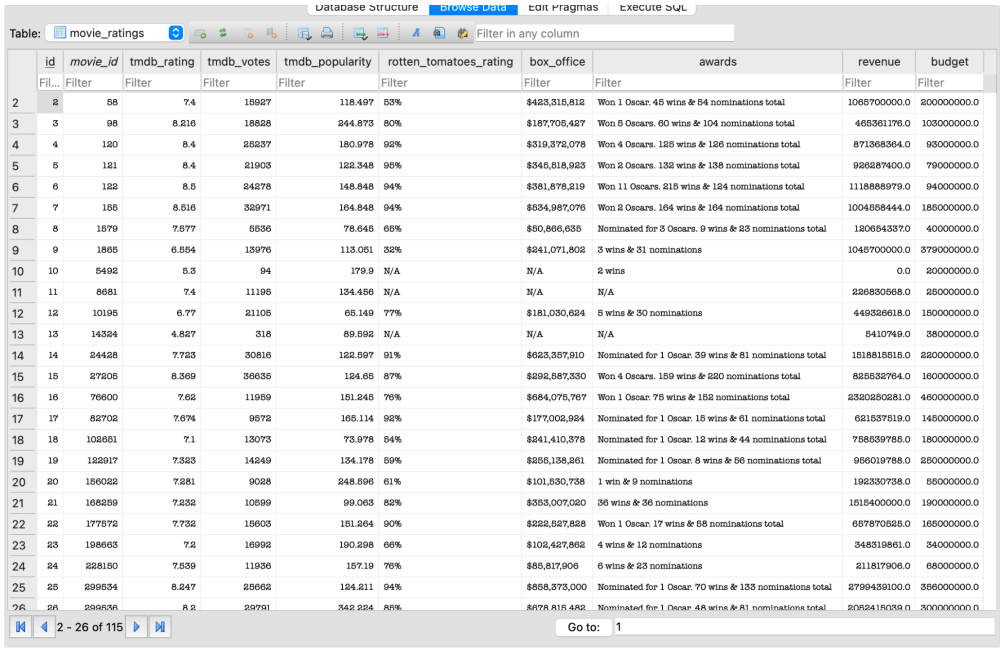


Table: movie_ratings

Filter in any column

	id	movie_id	tmdb_rating	tmdb_votes	tmdb_popularity	rotten_tomatoes_rating	box_office	awards	revenue	budget
	Fil...	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
2	2	58	7.4	15927	118.497	53%	$423,315,812	Won 1 Oscar. 45 wins & 54 nominations total	1068700000.0	200000000.0
3	3	98	8.216	18828	244.873	80%	$187,705,427	Won 5 Oscars. 60 wins & 104 nominations total	465361176.0	103000000.0
4	4	120	8.4	25237	180.978	92%	$319,372,078	Won 4 Oscars. 126 wins & 126 nominations total	871368364.0	93000000.0
5	5	121	8.4	21903	122.348	95%	$345,818,923	Won 2 Oscars. 132 wins & 138 nominations total	926287400.0	79000000.0
6	6	122	8.5	24278	148.848	94%	$381,878,219	Won 11 Oscars. 215 wins & 124 nominations total	1118888979.0	94000000.0
7	7	155	8.516	32971	164.848	94%	$534,987,076	Won 2 Oscars. 164 wins & 164 nominations total	1004558444.0	185000000.0
8	8	1579	7.577	5536	78.645	68%	$50,866,635	Nominated for 3 Oscars. 9 wins & 23 nominations total	120654337.0	40000000.0
9	9	1865	6.554	13976	113.051	32%	$241,071,802	3 wins & 31 nominations	1045700000.0	379000000.0
10	10	5492	5.3	94	179.9	N/A	N/A	2 wins	0.0	20000000.0
11	11	8681	7.4	11195	134.456	N/A	N/A	N/A	226830568.0	25000000.0
12	12	10195	6.77	21105	65.149	77%	$181,030,624	5 wins & 30 nominations	449326618.0	150000000.0
13	13	14324	4.827	318	89.592	N/A	N/A	N/A	5410749.0	38000000.0
14	14	24428	7.723	30816	122.597	91%	$623,357,910	Nominated for 1 Oscar. 39 wins & 81 nominations total	1518815515.0	220000000.0
15	15	27205	8.369	36635	124.65	87%	$292,587,330	Won 4 Oscars. 159 wins & 220 nominations total	825532764.0	160000000.0
16	16	76600	7.62	11959	151.245	76%	$684,075,767	Won 1 Oscar. 75 wins & 152 nominations total	2320250281.0	460000000.0
17	17	82702	7.674	9572	165.114	92%	$177,002,924	Nominated for 1 Oscar. 15 wins & 61 nominations total	621537519.0	145000000.0
18	18	102651	7.1	13073	73.978	54%	$241,410,378	Nominated for 1 Oscar. 12 wins & 44 nominations total	758539785.0	180000000.0
19	19	122917	7.323	14249	134.178	59%	$255,138,261	Nominated for 1 Oscar. 8 wins & 56 nominations total	956019788.0	250000000.0
20	20	156022	7.281	9028	248.596	61%	$101,630,738	1 win & 9 nominations	192330738.0	55000000.0
21	21	168259	7.232	10599	99.063	82%	$353,007,020	36 wins & 36 nominations	1515400000.0	190000000.0
22	22	177572	7.732	15603	151.264	90%	$222,527,828	Won 1 Oscar. 17 wins & 58 nominations total	657870525.0	165000000.0
23	23	198663	7.2	16992	190.298	66%	$102,427,862	4 wins & 12 nominations	348319861.0	34000000.0
24	24	228150	7.539	11936	157.19	76%	$85,817,906	6 wins & 23 nominations	211817906.0	68000000.0
25	25	299534	8.247	25662	124.211	94%	$858,373,000	Nominated for 1 Oscar. 70 wins & 133 nominations total	2799439100.0	356000000.0
26	26	299536	8.2	29791	342.224	85%	$678,815,482	Nominated for 1 Oscar. 48 wins & 81 nominations total	2052415039.0	300000000.0

2 - 26 of 115

Go to: 1

Movies :

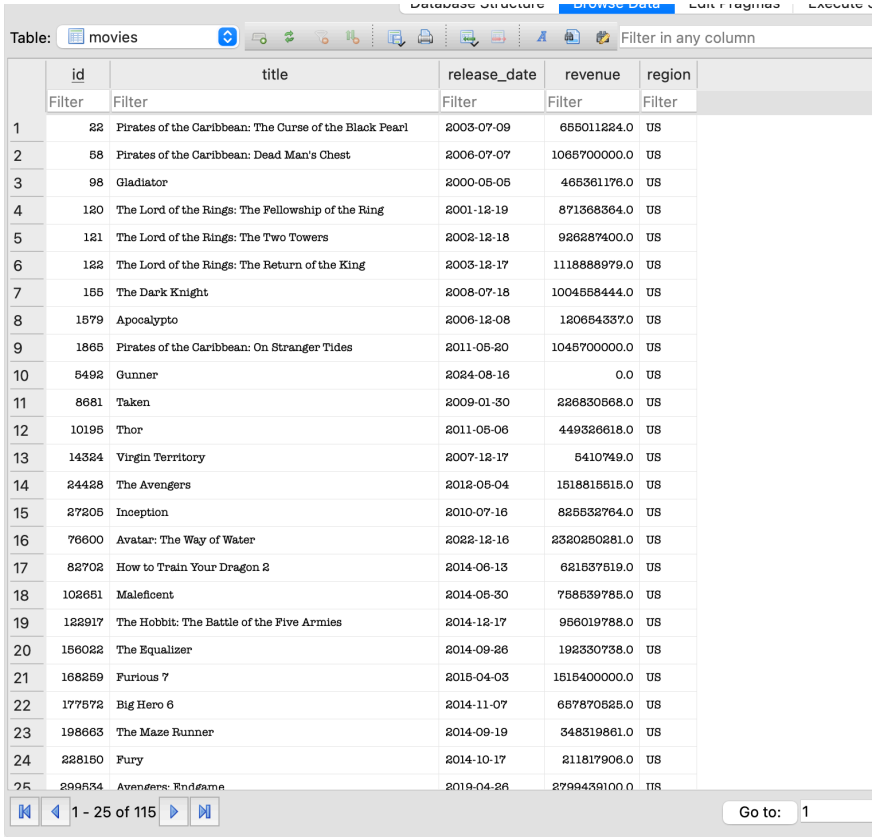


Table: movies

Filter in any column

	id	title	release_date	revenue	region
	Filter	Filter	Filter	Filter	Filter
1	22	Pirates of the Caribbean: The Curse of the Black Pearl	2003-07-09	655011224.0	US
2	58	Pirates of the Caribbean: Dead Man's Chest	2006-07-07	1065700000.0	US
3	98	Gladiator	2000-05-05	465361176.0	US
4	120	The Lord of the Rings: The Fellowship of the Ring	2001-12-19	871368364.0	US
5	121	The Lord of the Rings: The Two Towers	2002-12-18	926287400.0	US
6	122	The Lord of the Rings: The Return of the King	2003-12-17	1118888979.0	US
7	155	The Dark Knight	2008-07-18	1004558444.0	US
8	1579	Apocalypto	2006-12-08	120654337.0	US
9	1865	Pirates of the Caribbean: On Stranger Tides	2011-05-20	1045700000.0	US
10	5492	Gunner	2024-08-16	0.0	US
11	8681	Taken	2009-01-30	226830568.0	US
12	10195	Thor	2011-05-06	449326618.0	US
13	14324	Virgin Territory	2007-12-17	5410749.0	US
14	24428	The Avengers	2012-05-04	1518815515.0	US
15	27205	Inception	2010-07-16	825532764.0	US
16	76600	Avatar: The Way of Water	2022-12-16	2320250281.0	US
17	82702	How to Train Your Dragon 2	2014-06-13	621537519.0	US
18	102651	Maleficent	2014-05-30	758539785.0	US
19	122917	The Hobbit: The Battle of the Five Armies	2014-12-17	956019788.0	US
20	156022	The Equalizer	2014-09-26	192330738.0	US
21	168259	Furious 7	2015-04-03	1515400000.0	US
22	177572	Big Hero 6	2014-11-07	657870525.0	US
23	198663	The Maze Runner	2014-09-19	348319861.0	US
24	228150	Fury	2014-10-17	211817906.0	US
25	299534	Avengers: Endgame	2019-04-26	2799439100.0	US

1 - 25 of 115

Go to: 1

Omdb_movies: OMDB

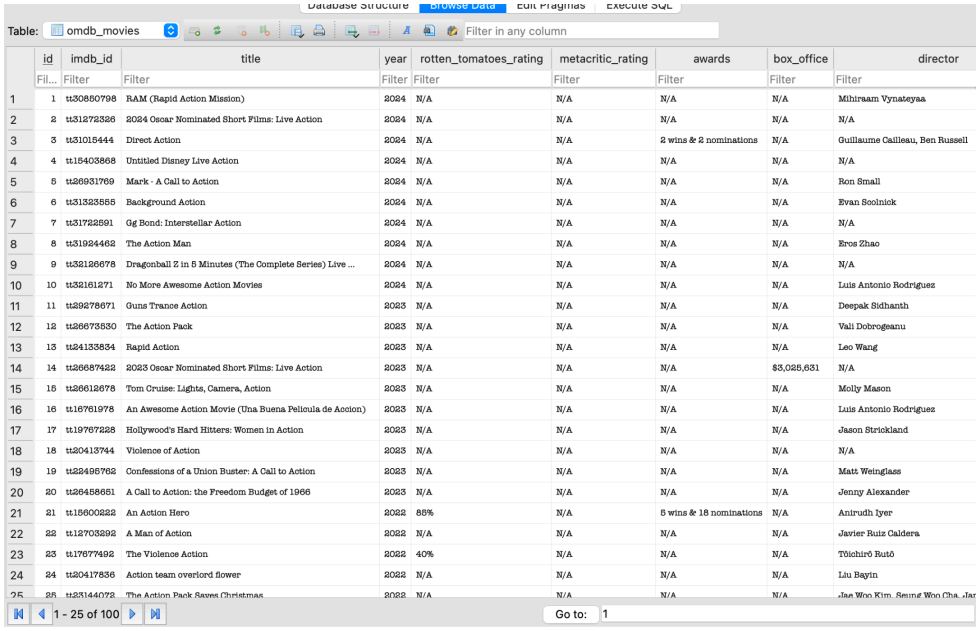


Table: omdb_movies — Filter in any column

	id	imdb_id	title	year	rotten_tomatoes_rating	metacritic_rating	awards	box_office	director
	Fil...	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	1	tt30850798	RAM (Rapid Action Mission)	2024	N/A	N/A	N/A	N/A	Mihiraam Vynateyaa
2	2	tt31272326	2024 Oscar Nominated Short Films: Live Action	2024	N/A	N/A	N/A	N/A	N/A
3	3	tt31015444	Direct Action	2024	N/A	N/A	2 wins & 2 nominations	N/A	Guillaume Cailleau, Ben Russell
4	4	tt15403868	Untitled Disney Live Action	2024	N/A	N/A	N/A	N/A	N/A
5	5	tt26931769	Mark - A Call to Action	2024	N/A	N/A	N/A	N/A	Ron Small
6	6	tt31323555	Background Action	2024	N/A	N/A	N/A	N/A	Evan Scolnick
7	7	tt31722591	Gg Bond: Interstellar Action	2024	N/A	N/A	N/A	N/A	N/A
8	8	tt31924462	The Action Man	2024	N/A	N/A	N/A	N/A	Eros Zhao
9	9	tt32126678	Dragonball Z in 5 Minutes (The Complete Series) Live ...	2024	N/A	N/A	N/A	N/A	N/A
10	10	tt32161271	No More Awesome Action Movies	2024	N/A	N/A	N/A	N/A	Luis Antonio Rodriguez
11	11	tt29278671	Guns Trance Action	2023	N/A	N/A	N/A	N/A	Deepak Sidhanth
12	12	tt28673530	The Action Pack	2023	N/A	N/A	N/A	N/A	Vali Dobrogeanu
13	13	tt24133834	Rapid Action	2023	N/A	N/A	N/A	N/A	Leo Wang
14	14	tt26687422	2023 Oscar Nominated Short Films: Live Action	2023	N/A	N/A	N/A	$3,025,631	N/A
15	15	tt26612678	Tom Cruise: Lights, Camera, Action	2023	N/A	N/A	N/A	N/A	Molly Mason
16	16	tt16761978	An Awesome Action Movie (Una Buena Película de Acción)	2023	N/A	N/A	N/A	N/A	Luis Antonio Rodriguez
17	17	tt19767228	Hollywood's Hard Hitters: Women in Action	2023	N/A	N/A	N/A	N/A	Jason Strickland
18	18	tt20413744	Violence of Action	2023	N/A	N/A	N/A	N/A	N/A
19	19	tt22496762	Confessions of a Union Buster: A Call to Action	2023	N/A	N/A	N/A	N/A	Matt Weinglass
20	20	tt26458651	A Call to Action: the Freedom Budget of 1966	2023	N/A	N/A	N/A	N/A	Jenny Alexander
21	21	tt15600222	An Action Hero	2022	88%	N/A	5 wins & 18 nominations	N/A	Anirudh Iyer
22	22	tt12703292	A Man of Action	2022	N/A	N/A	N/A	N/A	Javier Ruiz Caldera
23	23	tt17677492	The Violence Action	2022	40%	N/A	N/A	N/A	Tōichirō Rutō
24	24	tt20417836	Action team overlord flower	2022	N/A	N/A	N/A	N/A	Liu Bayin
25	25	tt23144072	The Action Pack Saves Christmas	2022	N/A	N/A	N/A	N/A	Jae Woo Kim, Seung Woo Cha, Jar

1 - 25 of 100 — Go to: 1

Tmdb_movies: TMDB

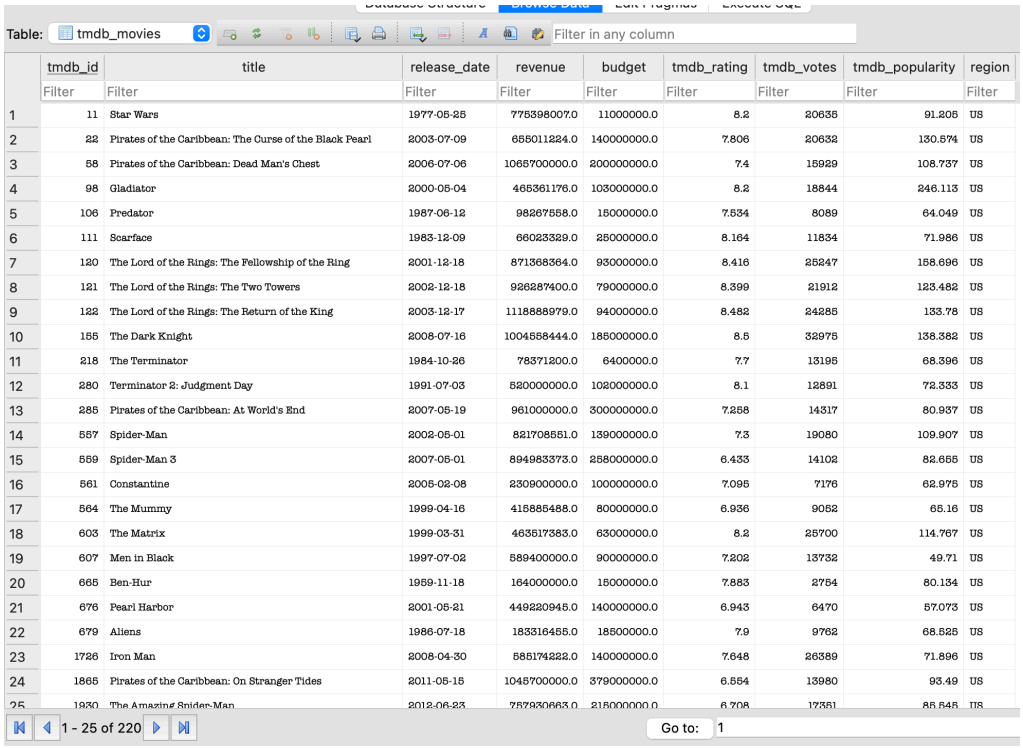


Table: tmdb_movies — Filter in any column

	tmdb_id	title	release_date	revenue	budget	tmdb_rating	tmdb_votes	tmdb_popularity	region
	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	11	Star Wars	1977-05-25	775398007.0	11000000.0	8.2	20635	91.205	US
2	22	Pirates of the Caribbean: The Curse of the Black Pearl	2003-07-09	655011224.0	140000000.0	7.806	20632	130.574	US
3	58	Pirates of the Caribbean: Dead Man's Chest	2006-07-06	1065700000.0	200000000.0	7.4	15929	108.737	US
4	98	Gladiator	2000-05-04	465361176.0	103000000.0	8.2	18844	246.113	US
5	106	Predator	1987-06-12	98267558.0	15000000.0	7.534	8089	64.049	US
6	111	Scarface	1983-12-09	66023329.0	25000000.0	8.164	11834	71.986	US
7	120	The Lord of the Rings: The Fellowship of the Ring	2001-12-18	871368364.0	93000000.0	8.416	25247	158.696	US
8	121	The Lord of the Rings: The Two Towers	2002-12-18	926287400.0	79000000.0	8.399	21912	123.482	US
9	122	The Lord of the Rings: The Return of the King	2003-12-17	1118888979.0	94000000.0	8.482	24285	133.78	US
10	155	The Dark Knight	2008-07-16	1004558444.0	185000000.0	8.5	32975	138.382	US
11	218	The Terminator	1984-10-26	78371200.0	6400000.0	7.7	13195	68.396	US
12	280	Terminator 2: Judgment Day	1991-07-03	520000000.0	102000000.0	8.1	12891	72.333	US
13	285	Pirates of the Caribbean: At World's End	2007-05-19	961000000.0	300000000.0	7.258	14317	80.937	US
14	557	Spider-Man	2002-05-01	821708551.0	139000000.0	7.3	19080	109.907	US
15	559	Spider-Man 3	2007-05-01	894983373.0	258000000.0	6.433	14102	82.655	US
16	561	Constantine	2005-02-08	230900000.0	100000000.0	7.095	7176	62.975	US
17	564	The Mummy	1999-04-16	415885488.0	80000000.0	6.936	9052	65.16	US
18	603	The Matrix	1999-03-31	463517383.0	63000000.0	8.2	25700	114.767	US
19	607	Men in Black	1997-07-02	589400000.0	90000000.0	7.202	13732	49.71	US
20	665	Ben-Hur	1959-11-18	164000000.0	15000000.0	7.883	2754	80.134	US
21	676	Pearl Harbor	2001-05-21	449220945.0	140000000.0	6.943	6470	57.073	US
22	679	Aliens	1986-07-18	183316455.0	18500000.0	7.9	9762	68.525	US
23	1726	Iron Man	2008-04-30	585174222.0	140000000.0	7.648	26389	71.896	US
24	1865	Pirates of the Caribbean: On Stranger Tides	2011-05-15	1045700000.0	379000000.0	6.554	13980	93.49	US
25	1930	The Amazing Spider-Man	2012-06-23	757930663.0	215000000.0	6.708	17351	86.546	US

1 - 25 of 220 — Go to: 1

U.S. Census:

Region_lookup

Table: region_lookup

	region_id	region_name
	Filter	Filter
1	1	Midwest
2	2	West
3	3	South
4	4	Northeast

Regions

Table: regions

Filter in any column

	id	state_id	region_id	population	age_group	age_population	percentage_of_state
	Fil...	Filter	Filter	Filter	Filter	Filter	Filter
1	1	36	3	3986639	Age Under 5	243184	6.1
2	2	36	3	3986639	Age 5-9	247171	6.2
3	3	36	3	3986639	Age 10-14	255144	6.4
4	4	36	3	3986639	Age 15-17	151492	3.8
5	5	36	3	3986639	Age 18-19	99665	2.5
6	6	27	1	1963692	Age Under 5	119785	6.1
7	7	27	1	1963692	Age 5-9	121748	6.2
8	8	27	1	1963692	Age 10-14	125676	6.4
9	9	27	1	1963692	Age 15-17	74620	3.8
10	10	27	1	1963692	Age 18-19	49092	2.5
11	11	11	2	1441553	Age Under 5	87934	6.1
12	12	11	2	1441553	Age 5-9	89376	6.2
13	13	11	2	1441553	Age 10-14	92259	6.4
14	14	11	2	1441553	Age 15-17	54779	3.8
15	15	11	2	1441553	Age 18-19	36038	2.5
16	16	41	1	895376	Age Under 5	54617	6.1
17	17	41	1	895376	Age 5-9	55513	6.2
18	18	41	1	895376	Age 10-14	57304	6.4
19	19	41	1	895376	Age 15-17	34024	3.8
20	20	41	1	895376	Age 18-19	22384	2.5
21	21	42	3	6975218	Age Under 5	425488	6.1
22	22	42	3	6975218	Age 5-9	432463	6.2
23	23	42	3	6975218	Age 10-14	446413	6.4
24	24	42	3	6975218	Age 15-17	265058	3.8
25	25	42	3	6975218	Age 18-19	174380	2.5

1 - 25 of 250

Go to: 1

state _lookup:

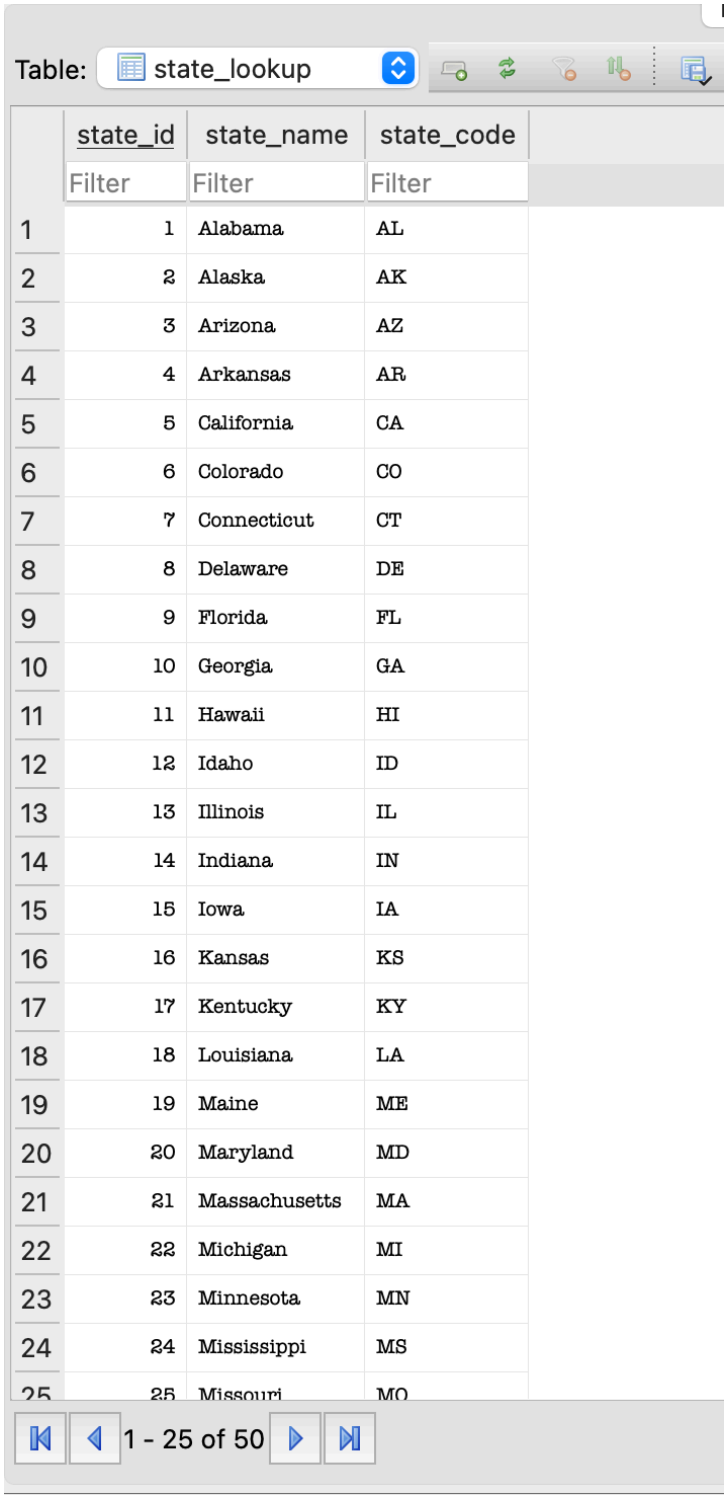


Table: state_lookup

	state_id	state_name	state_code
	Filter	Filter	Filter
1	1	Alabama	AL
2	2	Alaska	AK
3	3	Arizona	AZ
4	4	Arkansas	AR
5	5	California	CA
6	6	Colorado	CO
7	7	Connecticut	CT
8	8	Delaware	DE
9	9	Florida	FL
10	10	Georgia	GA
11	11	Hawaii	HI
12	12	Idaho	ID
13	13	Illinois	IL
14	14	Indiana	IN
15	15	Iowa	IA
16	16	Kansas	KS
17	17	Kentucky	KY
18	18	Louisiana	LA
19	19	Maine	ME
20	20	Maryland	MD
21	21	Massachusetts	MA
22	22	Michigan	MI
23	23	Minnesota	MN
24	24	Mississippi	MS
25	25	Missouri	MO

1 - 25 of 50

5. Visualization

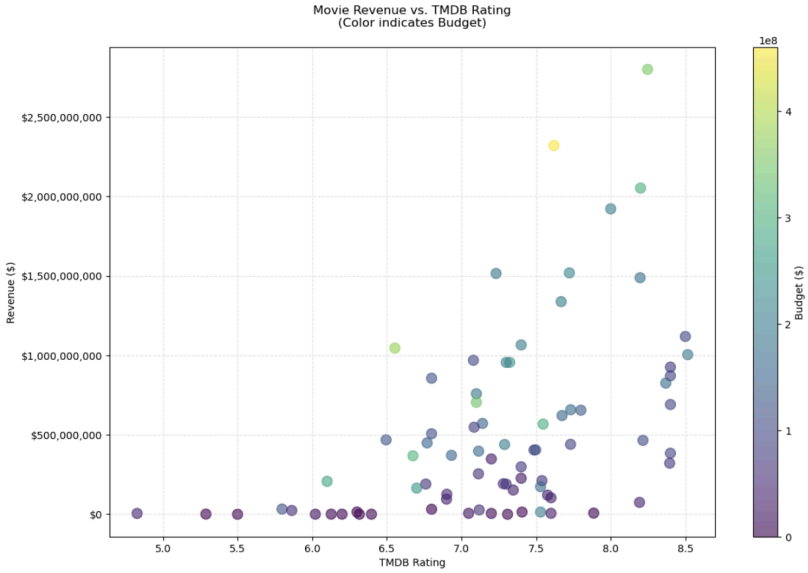


Figure 1. The scatter plot shows the relationship between a movie's TMDB rating and its revenue, with the color of each point indicating the movie's budget.

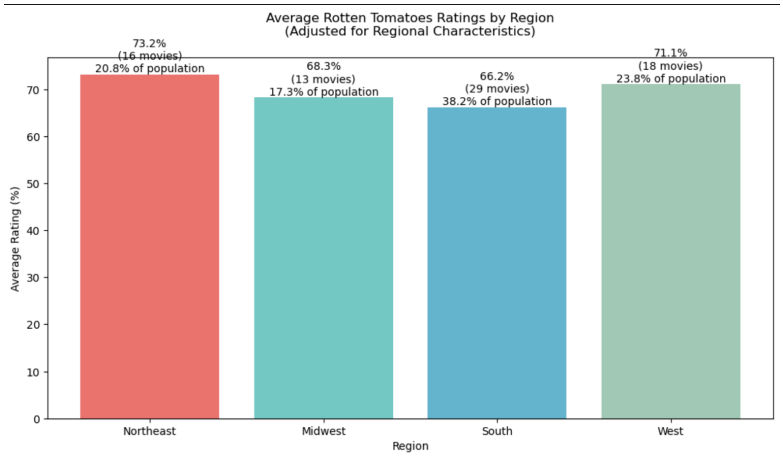


Figure 2. Bar chart of

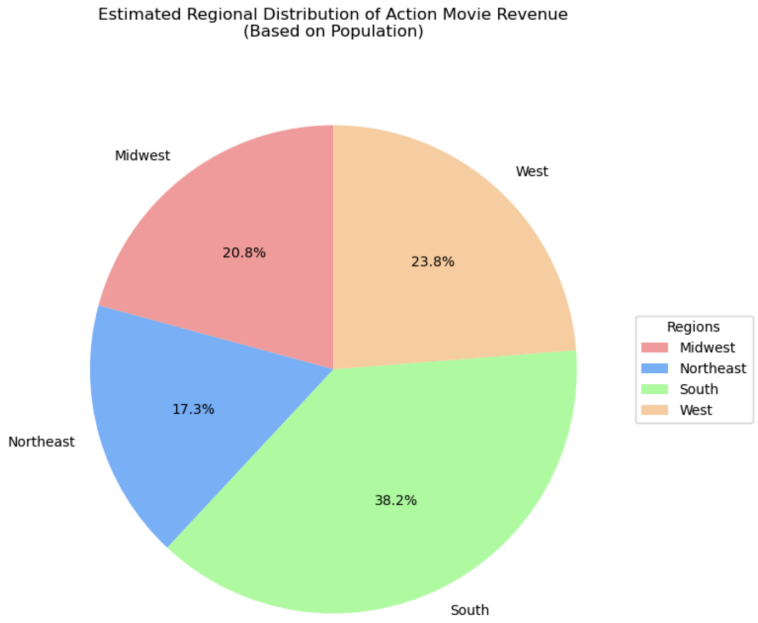


Figure 3. A pie chart showing the distribution of how much revenue each region contributes to the action movies presented in our dataset

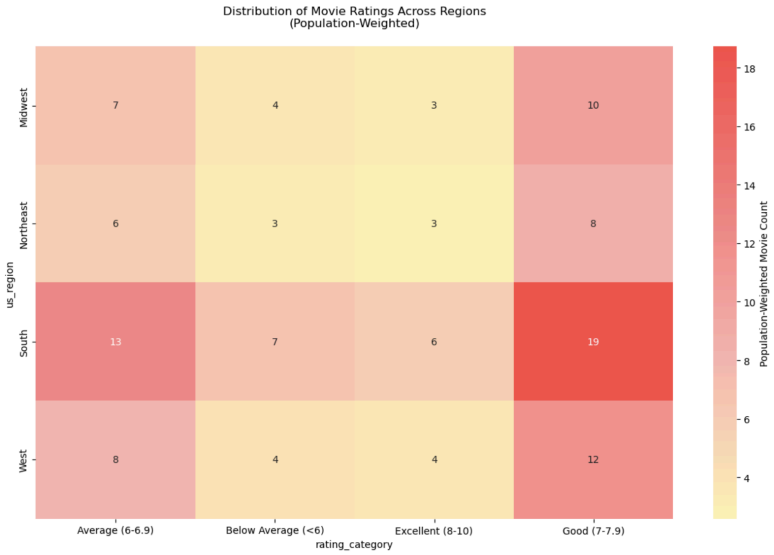


Figure 4. A heat map displaying the population density of each region rating action movies "Average," "Below Average," "Excellent," and "Good."

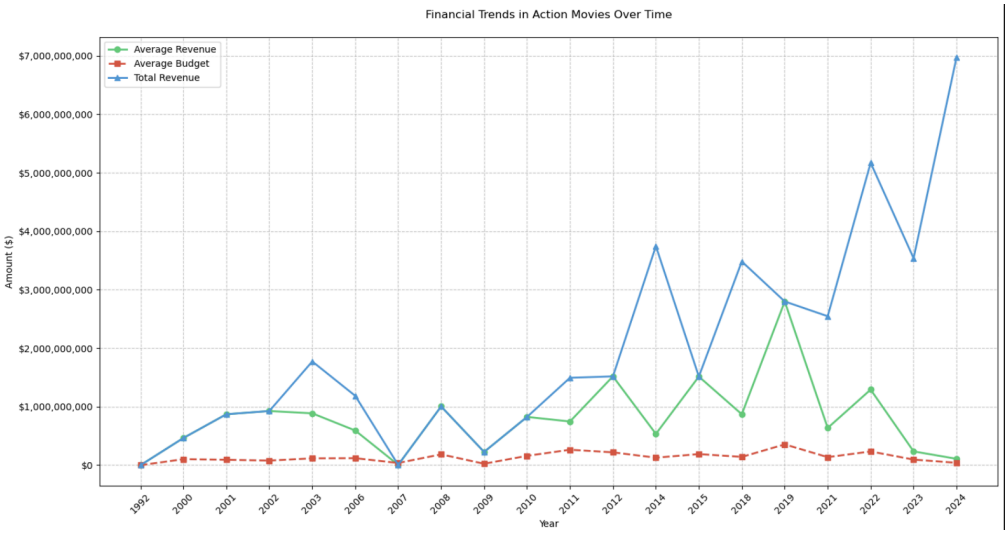


Figure 5. A line graph showing average budget, average revenue, and total revenue in billions over a period of time (1992-2024)

6. Instructions for running code:

Step 1: Make sure all required Python packages are installed

1. requests
2. pandas
3. tmdbv3api
4. python-dotenv
5. beautifulsoup4
6. matplotlib
7. seaborn
8. numpy

Step 2: Run movie data collection

movie_data_collector.py

- This creates the database

- Collects movie data from TMDB and OMDB APIs

- You'll see progress messages as it collects data

Step 3: Run census data collection

census_data.py

- This gets population data from Census API
- Shows population breakdowns by region
- You'll see each state being added with its population

Step 4: Run data analysis

process_movie_data.py

- Creates movie_analysis_results.txt
- Shows detailed statistics about movies
- You'll see a completion message when done

Step 5: Generate visualizations

visualizations.py

- Creates five visualization files:
- revenue_pie_chart.png
- ratings_bar_chart.png
- ratings_heatmap.png
- financial_trends.png
- revenue_scatter.png

Expected Output:

- A SQLite database file: movies.db
- An analysis text file: movie_analysis_results.txt & census_analysis.txt
- Five PNG image files with visualizations

Viewing Results:

- Open movie_analysis_results.txt in any text editor
- Open population_and_demographic_analysis.txt
- Open the PNG files to see the visualizations
- Each visualization shows different aspects of the movie data analysis
- The code should be run in exactly this order because each step depends on the data created in the previous steps.

7. Documentation:

File: census_data.py

Function: init_db()

Input: None

Output: Creates SQLite regions table with columns:

- id (PRIMARY KEY)
- country_code
- region_name
- us_region
- state_code
- population
- gdp_per_capita

Function: fetch_population_data()

Input:

- Census API endpoint data (api.census.gov/data/2021/pep/population)
- Population estimates by state
- US regions dictionary mapping states to regions

Output:

- Populated regions table in SQLite database
- Printed summary of population by region
- Regional population totals

Function: main()

Input: None

Output:

- Initialized database
- Fetched population data
- Generated visualizations

File: movie_data_collector.py

Function: init_db()

Input: None

Output: Creates two SQLite tables:

- movies table:
    - id, title, release_date, revenue, region
- movie_ratings table:
    - movie_id, tmdb_rating, tmdb_votes, tmdb_popularity
    - rotten_tomatoes_rating, box_office, awards
    - revenue, budget

Function: main()

Input:

- TMDB API data (movies, ratings, popularity)
- OMDB API data (Rotten Tomatoes ratings, box office)

Output:

- Populated movies and movie_ratings tables
- Printed summary of collected data
- Sample of movies with combined stats

File: process_movie_data.py

Function: calculate_movie_stats()

Input: Data from SQLite tables:

- movies: id, title, release_date, revenue, region
- movie_ratings: ratings, revenue, budget data
- regions: population, regional data

Output: movie_analysis_results.txt containing:

1. Yearly Statistics:
    - Movies per year
    - Average ratings
    - Financial metrics
2. Rotten Tomatoes Analysis:
    - Rating categories
    - Revenue analysis
3. TMDB Rating Performance:
    - Rating distribution
    - Financial metrics by rating
    - ROI calculations
4. US Market Analysis:
    - Population metrics
    - Movie distribution
    - Financial analysis
5. Regional Analysis:
    - Per-region statistics
    - Population-weighted metrics
    - Per capita calculations

File: visualizations.py

Function: create_revenue_pie_chart()

Input:

- Revenue data from movie_ratings table
- Regional data from regions table

Output: revenue_pie_chart.png showing regional revenue distribution

Function: create_rating_bar_chart()

Input:

- Rotten Tomatoes ratings from movie_ratings
- Regional population data from regions table

Output: ratings_bar_chart.png with regional rating averages

Function: create_ratings_heatmap()

Input:

- TMDB ratings from movie_ratings
- Regional data from regions table

Output: ratings_heatmap.png showing rating distribution

Function: create_financial_line_graph()

Input:

- Revenue and budget data from movie_ratings
- Release dates from movies table

Output: financial_trends.png showing trends over time

Function: create_revenue_scatter_plot()

Input:

- Revenue data from movie_ratings
- TMDB ratings from movie_ratings
- Regional data from regions table

Output: revenue_scatter.png showing rating-revenue relationship

Function: create_visualizations()

Input: None

Output:

- Calls all visualization functions
- Creates all five visualization files
- Prints progress messages

In conclusion, movie_data_collector.py sets up and populates movie data, census_data.py adds population data, process_movie_data.py generates the analysis and visualizations.py creates visual representations.

8. Concluding statement:

Based on our findings, our hypothesis that the action movie genre and box office revenue vary across different regions was partially supported. While regional differences in ratings and revenue contributions were evident, the data suggests that these variations are influenced by factors such as population density and other confounding factors of distribution. The South, with its largest population share, contributed the most to box office revenue but had the lowest average ratings, while the Northeast and West, despite fewer movies, showed higher average ratings. This indicates that while population size impacts revenue potential, it does not necessarily correlate with higher ratings or perceived movie quality. Moreover, the positive correlation between ratings and revenue highlights that higher-rated movies tend to perform better financially, regardless of region. These results display the varying aspects between demographic factors and the success of the action movie genre.

9. Documentation of resources:

Date	**Item Description**	**Location of Resource**	**Result (did it solve the issue?**
11/22/24	API access code	OMDB	Yes
11/22/24	API access code	TMDB	Yes
11/22/24	API access code	U.S. Census Bureau	Yes
11/25/24	Issue: Census API authentication was failing at times	ChatGPT	The API key was somehow recorded wrong, changed it to match our API key
11/30/24	Issue: Population data from Census API was returning	ChatGPT	Gave suggestions on how to fix by sorting state

	incorrect total US population		populations and implementing age group calculations correctly in census_data.py
12/1/24	Issue: Movie titles with special characters were causing encoding errors	ChatGPT	Added UTF-8 encoding handling for all text processing operations
12/3/24	Issue: We didn't know what packages to use for the visuals and what visuals were already shown during lecture	SI 206 Lecture 23 Plotly slides	Slides showed that we can use Plotly to create different types of graphs and it gave us some ideas of what we wanted to implement
12/4/24	Issue: ModuleNotFoundError when running the visualization script	ChatGPT	Created a requirements.txt file and installed all necessary packages using conda/pip.
12/4/24	Issue: TMDB API rate limiting causing incomplete data collection	ChatGPT	Added time delays and error handling in the script.
12/5/24	Issue: Rotten Tomatoes ratings stored with '%' symbol caused calculation errors	ChatGPT	Cleaned data by removing '%' and converting to float.
12/6/24	Issue: The population analysis file wasn't updating with new data	ChatGPT	The file handling process was adjusted to ensure the file opens correctly for writing, saves the new data, and then closes properly after the update.
12/6/24	Issue: Movie revenue calculations showed incorrect regional distributions	ChatGPT	Adjusted the formula to account for population differences between regions.
12/7/24	Issue: Database tables had duplicate entries for movies	ChatGPT	Added UNIQUE constraints and JOIN conditions in SQL queries to prevent duplicates
12/12/24	Issue: U.S. Census database didn't have 100 entries, and we wanted suggestions on how to increase it to at least 100	ChatGPT	Gave a suggestion to collect data points that accounted for age demographics

12/13/24	Issue: After the grading session, we were advised to create separate tables so that there are no duplicate strings in the data tables (State & region names). We wanted suggestions on how we can do this.	ChatGPT	Gave us suggestions on what tables to create such as a table just for the states and assigning the state different id and the same with regions.