Does higher IMDB rating lead to higher revenue?

Team: Cason Chen, Rajvir Clair, Jimmy He, Yi-Hua Huang, Jingya Hu, Islam Sultan, Eva Wen

Agenda

1. Problem Definition
2. Ideal Experiment
3. Data Cleaning
4. Descriptive Statistics
5. Regression Model and Results
6. Key Takeaways & Conclusion

Problem Definition

Does higher IMDb rating relate to higher revenue?

H_0: There is no association.

Ha: Higher IMDb ratings are associated with higher revenue.

- **IMDb Rating ★ ★ ★**
- **A platform where users score movies from 1 to 10**

IMDb

- **Box Office Revenue 💰**
- **Inflation Adjusted**

Understanding this relationship matters for anyone **making creative**, **production**, or **investment** decisions in the film industry

Ideal Experiment

Title	Year	**Rating**	Metascore	Genre	Vote	Director	Runtime	Description	cpi_factor
Movie A	2014	**9**	75	Action	100000	Mike	150	-	1.060705
Movie B	2014	**6**	75	Action	100000	Mike	150	-	1.060705

Two Movies Identical in Every Way

- **Same director**
- **Same genre**
- **Same runtime**
- **Same budget & marketing**
- **Same release year**
- **Same distribution strategy**

... BUT IMDb Rating

Why This Perfect Experiment Is Impossible

- **Ratings depend on:**
  - **Early audience composition**
  - **Marketing exposure**
  - **Genre appeal**
  - **Critical reviews (Metascore)**
  - **Overall popularity (Votes)**
- **Revenue can also influence ratings (reverse causality)**

Data Cleaning & Descriptive Statistics

Using Python to cleaning data and visualization

Data Cleaning

Step 1

- Treat missing value as "NA" and drop all NA rows

Step 2

- Keep movies from 2014–2018 and CPI-adjust to Real Revenue

Step 3

- Create dummy variable for top 4 popular genres: Drama / Comedy / Action / Adventure

Step 4

- Keep movies with at least one of these four genres and return cleaning dataset

Descriptive Statistics

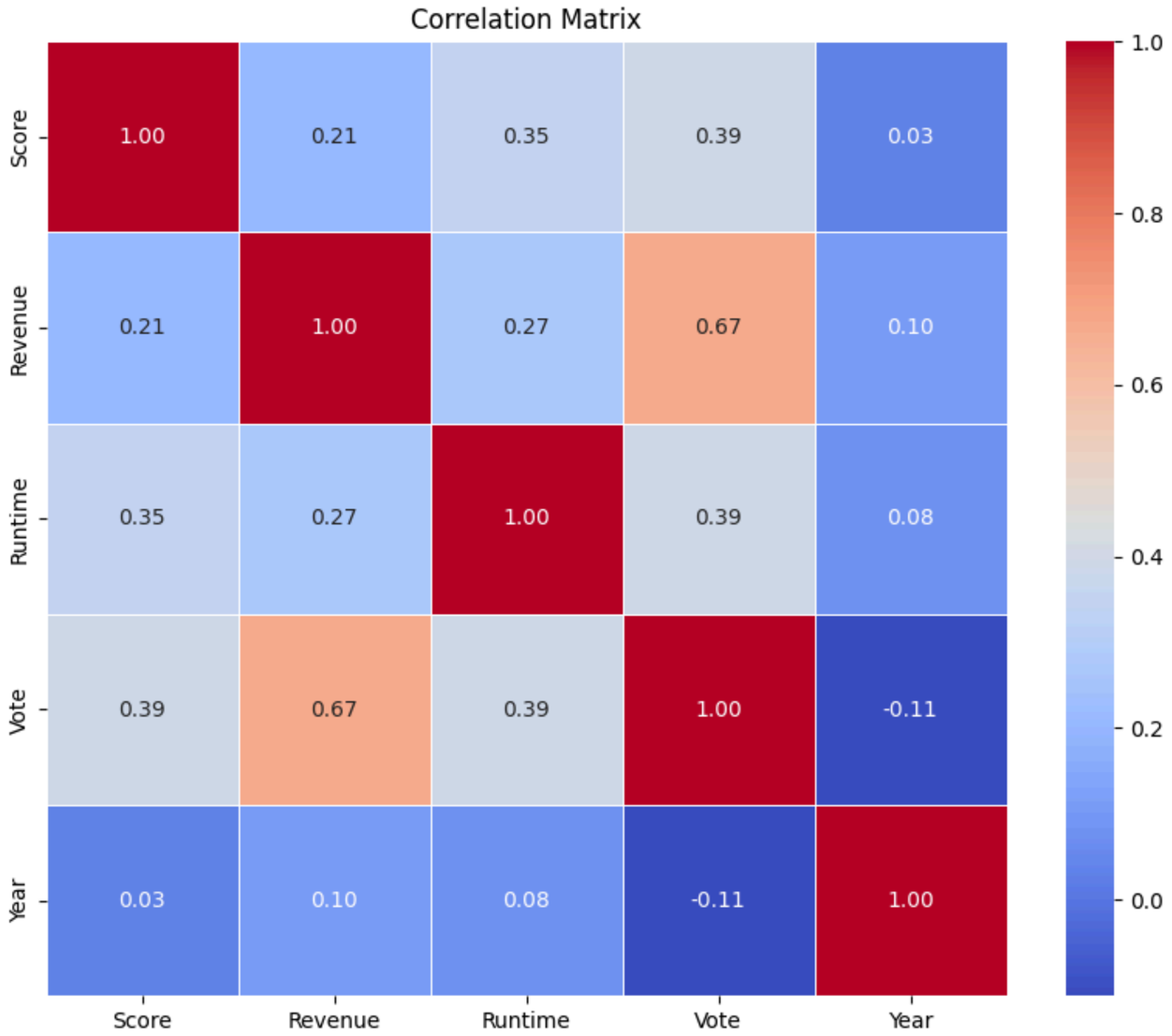
Callouts

- N = 929 films
- Avg Score: 6.6
- Avg Runtime: 110 min
- Avg Revenue: $51M

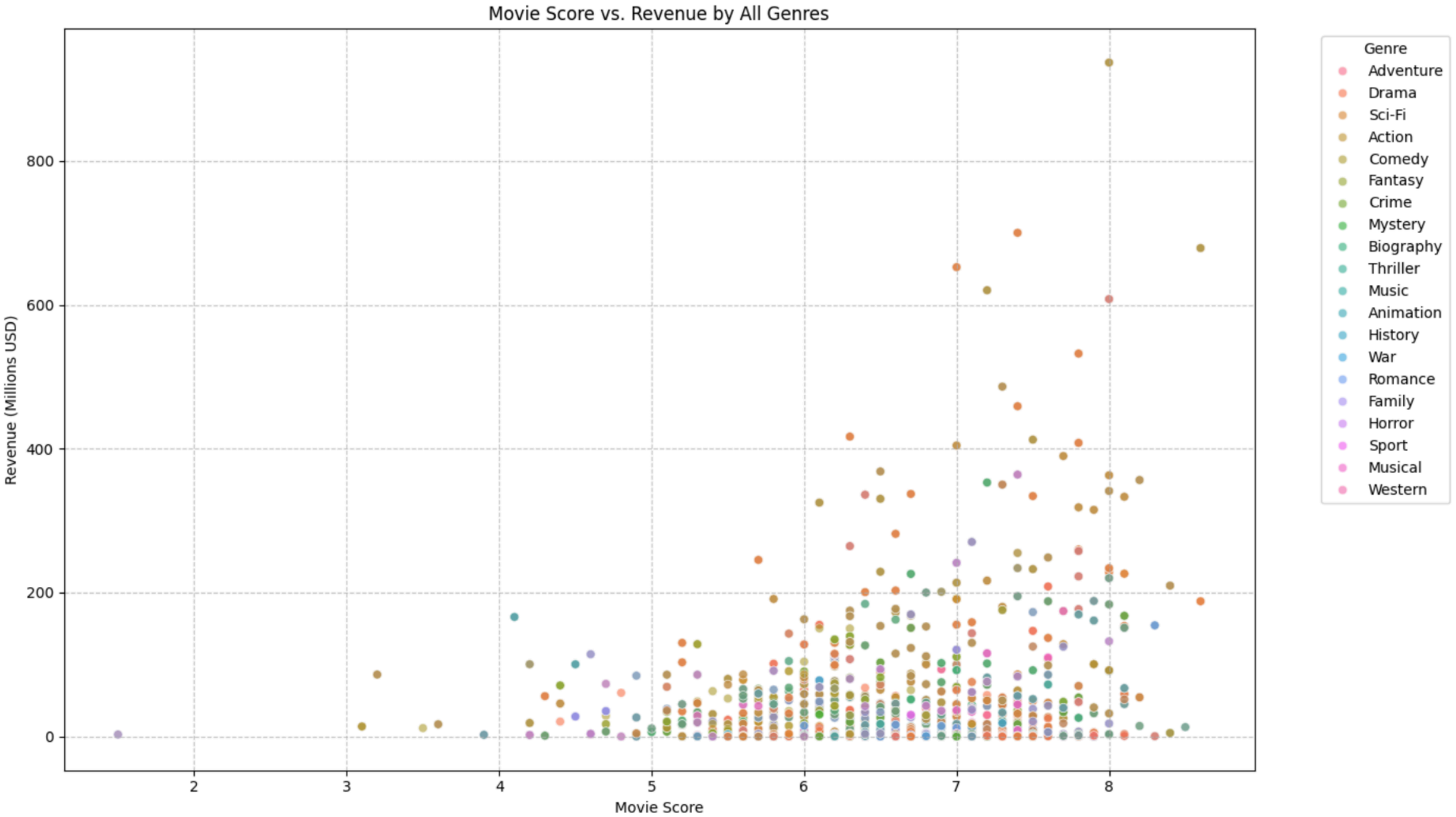
Mini Insights

- Revenue ↔ Votes: 0.67
- Score ↔ Revenue: 0.21–0.39
- Runtime has weak impact

Descriptive Statistics



Descriptive Statistics



Descriptive Statistics

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.20787483
R Square	0.04321194
Adjusted R Sc	0.04217981
Standard Erro	93.737921
Observations	929

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	367873.439	367873.439	41.8666101	1.5805E-10
Residual	927	8145361.6	8786.79784		
Total	928	8513235.04			

	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-100.92315	24.0641162	-4.1939273	3.0047E-05	-148.14962	-53.696692	-148.14962	-53.696692
Rating	23.4387969	3.6224418	6.47044126	1.5805E-10	16.3296594	30.5479344	16.3296594	30.5479344

Final Regression Model Preview

- **Revenue:** movie's ticket sales
- **Rating:** IMDB rating, purely voted by everyone
- **Four Popular Genres**: Action, Adventure, Comedy, Drama
- **Genres' Interaction Terms:** show how different genres affect the main relationship
- **Metascore:** critic's score of the movie
- **Votes**: number of votes for each IMDB rating

Regression Statistics					
Multiple R	0.76528695				
R Square	0.58566411				
Adjusted R Sc	0.58069389				
Standard Erro	62.0210223				
Observations	929				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	4985896.23	453263.294	117.834567	6.059E-167
Residual	917	3527338.8	3846.6072		
Total	928	8513235.04			
	Coefficients	*Standard Erro*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-100.21471	53.9975793	-1.8559112	0.06378687	-206.18789
Rating	20.2587854	8.53314157	2.37412977	0.01779541	3.51203146
Drama	184.063244	48.7940526	3.77224753	0.00017216	88.3022649
Comedy	78.6510285	41.8030298	1.88146718	0.06022497	-3.3896889
Action	-104.35273	45.416181	-2.2976994	0.02180328	-193.48446
Adventure	-22.806178	42.794071	-0.5329285	0.59421222	-106.79187
Drama*Ratir	-34.655934	7.50570319	-4.6172801	4.4425E-06	-49.386285
Comedy*Rat	-12.436989	6.25491598	-1.9883543	0.04706943	-24.712602
Action*Ratin	16.1515728	6.98032116	2.31387245	0.02089501	2.45231325
Adventure*R	9.28941673	6.53849999	1.42072597	0.15573637	-3.5427448
Metascore	0.16439314	0.16282814	1.00961133	0.31294786	-0.1551659
Vote in millio	380.624676	20.5120849	18.5561184	1.651E-65	340.368595

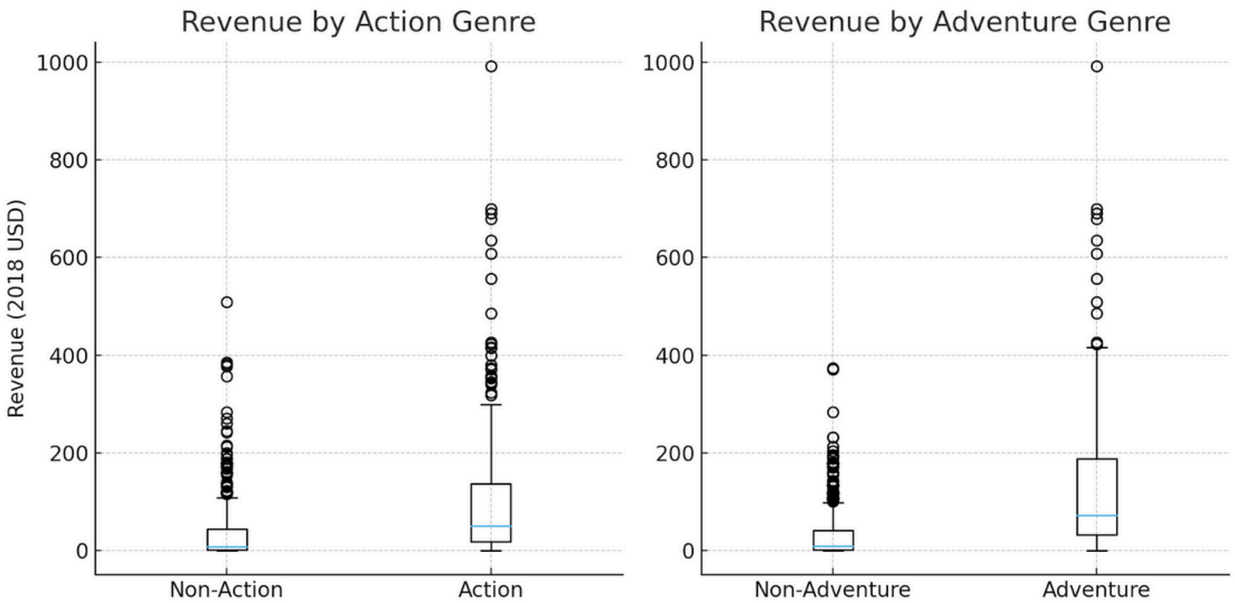
Action & Adventure

Descriptive Highlights (n = 929)

Genre	Avg Revenue	Avg Votes	Avg Rating
Action	**$105M**	**0.153M**	6.46
Non-Action	$33M	0.062M	6.64
Adventure	**$133M**	**0.167M**	6.6
Non-Adventure	$29M	0.065M	6.59

Action & Adventure earn 3–4× higher revenue despite similar or lower ratings.

Revenue Distribution by Genre



Blockbusters cluster in Action & Adventure; non-genre films remain low.

Regression Insights

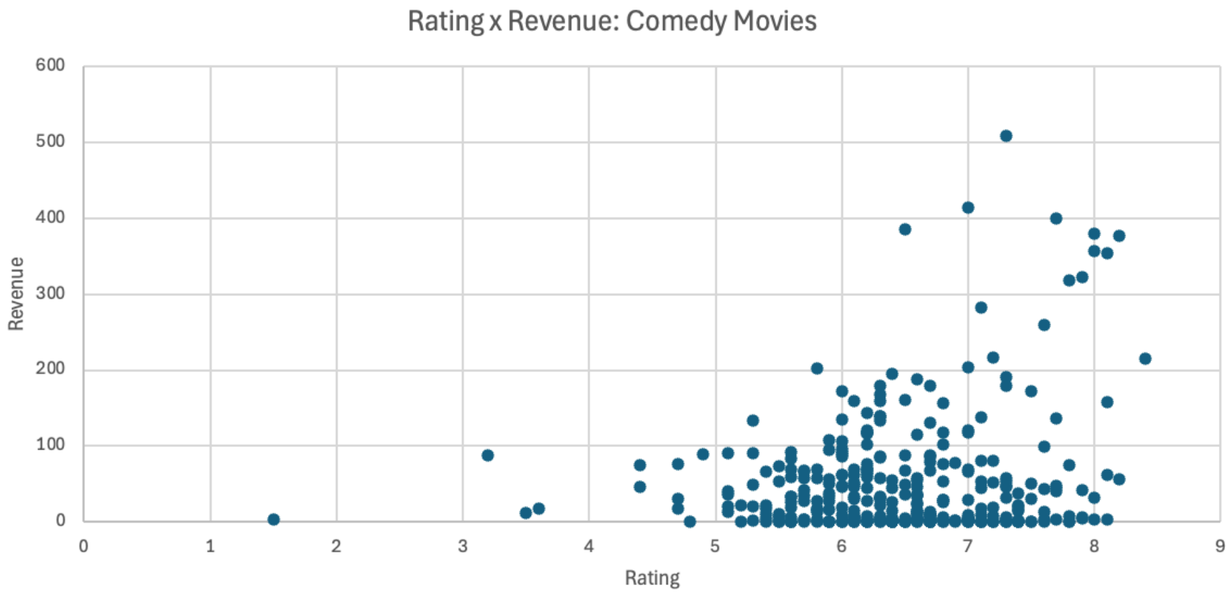
Action × Rating is significant (p ≈ 0.02)
Well-rated Action films earn ~$36M per rating point vs $20M for others.
→ Action amplifies revenue.

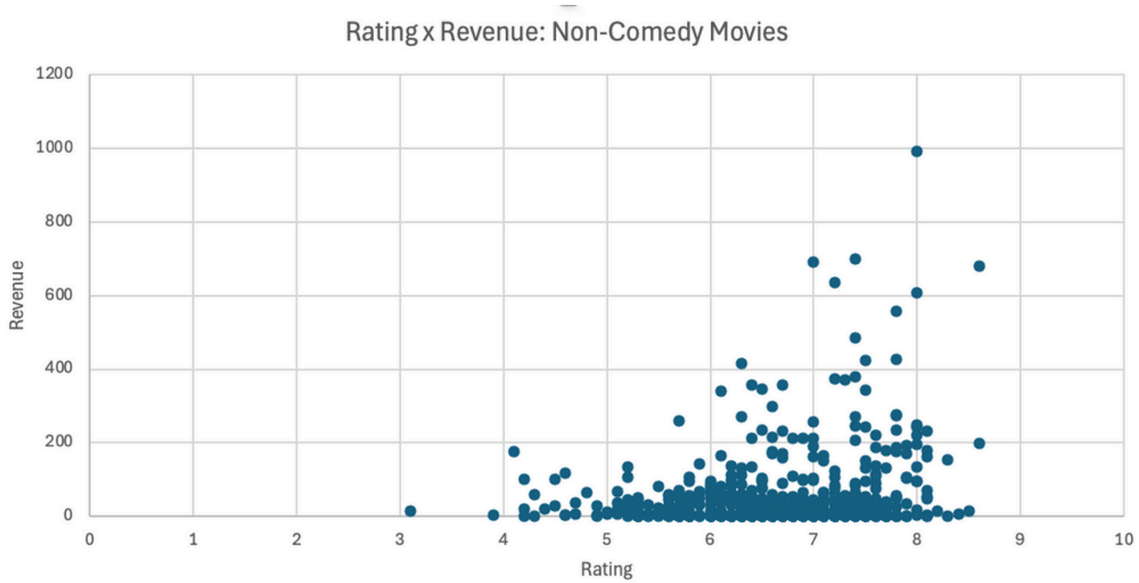
Adventure is not significant
No meaningful effect after controlling for rating & votes.
→ Adventure's uplift disappears statistically.

Comedy & Drama

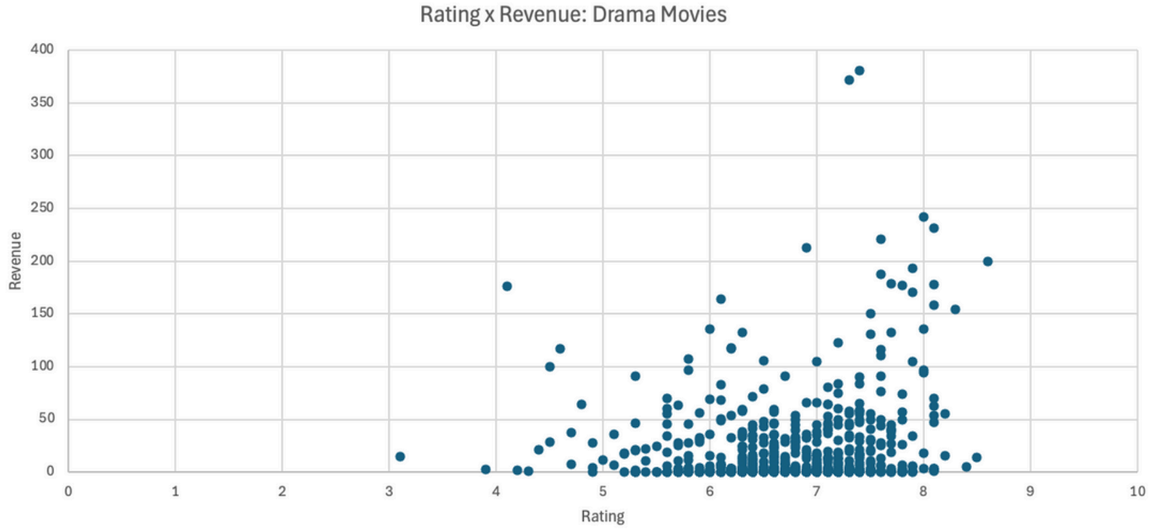
Genre	Avg Revenue	Avg Votes	Avg Rating
Comedy	**$51M**	**0.07M**	6.5
Non-Comedy	$55M	0.1M	6.7
Drama	**$24M**	**0.07M**	6.7
Non-Drama	$105M	0.13M	6.3

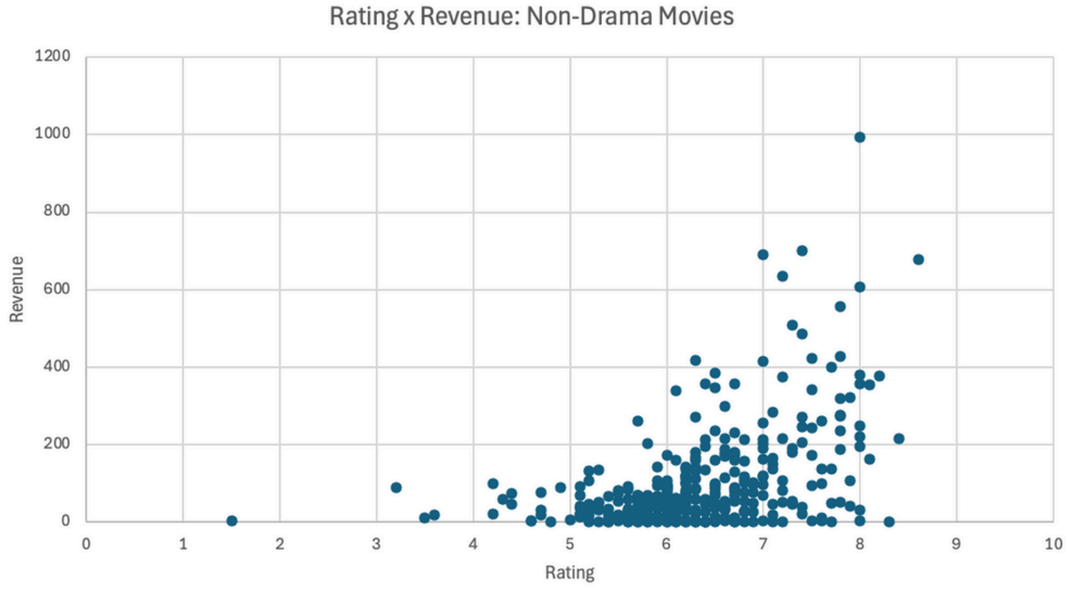
Comedy



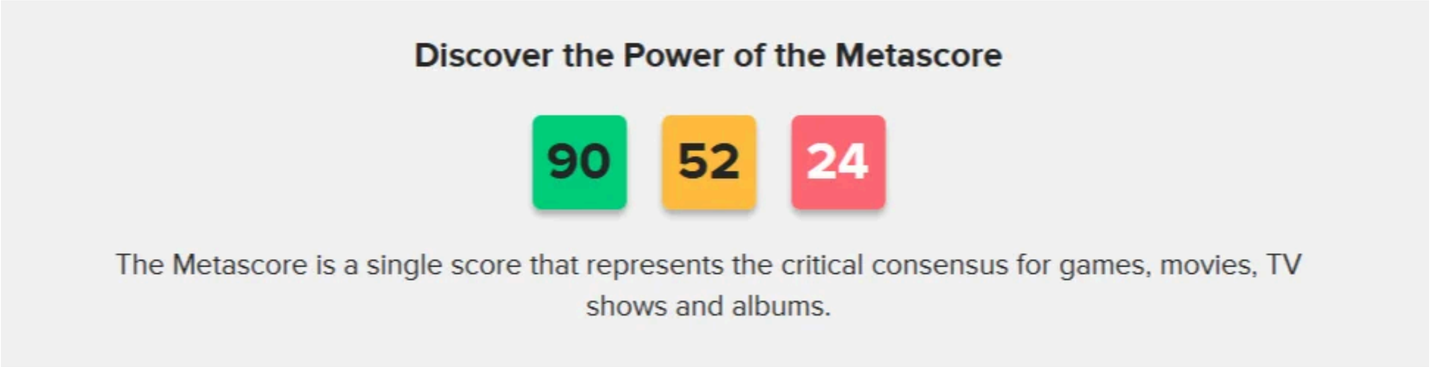


Drama





Metascore Analysis



Is it really rating affecting revenue? Or it's quality, represented by metascore, making a difference ?

Adding Metascore to Main Regression

Adjusted R^2 0.0422 to 0.0487

Rating P 1.58E-10 to 1.89E-10

Regression Statistics	
Multiple R	0.20787483
R Square	0.04321194
Adjusted R Sc	0.04217981
Standard Erro	93.737921
Observations	929

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	367873.439	367873.439	41.8666101
Residual	927	8145361.6	8786.79784	
Total	928	8513235.04		

	Coefficients	*Standard Erro*	*t Stat*	*P-value*
Intercept	-100.92315	24.0641162	-4.1939273	3.0047E-05
Rating	23.4387969	3.6224418	6.47044126	1.5805E-10

Regression Statistics	
Multiple R	0.22319333
R Square	0.04981526
Adjusted R Sc	0.04776303
Standard Erro	93.4643186
Observations	929

ANOVA

	df	*SS*	*MS*	*F*
Regression	2	424089.029	212044.514	24.273665
Residual	926	8089146.01	8735.57884	
Total	928	8513235.04		

	Coefficients	*Standard Erro*	*t Stat*	*P-value*
Intercept	-123.47281	25.5875379	-4.8255056	1.6326E-06
Metascore	-0.6050852	0.23852505	-2.5367785	0.01135096
Rating	32.2112824	5.00041918	6.44171642	1.896E-10

Adding Metascore to Main Regression

	Coefficients	*Standard Erro*	*t Stat*	*P-value*
Intercept	-100.21471	53.9975793	-1.8559112	0.06378687
Rating	20.2587854	8.53314157	2.37412977	0.01779541
Drama	184.063244	48.7940526	3.77224753	0.00017216
Comedy	78.6510285	41.8030298	1.88146718	0.06022497
Action	-104.35273	45.416181	-2.2976994	0.02180328
Adventure	-22.806178	42.794071	-0.5329285	0.59421222
Drama*Ratir	-34.655934	7.50570319	-4.6172801	4.4425E-06
Comedy*Rat	-12.436989	6.25491598	-1.9883543	0.04706943
Action*Ratin	16.1515728	6.98032116	2.31387245	0.02089501
Adventure*R	9.28941673	6.53849999	1.42072597	0.15573637
Metascore	0.16439314	0.16282814	1.00961133	**0.31294786**
Vote in millio	380.624676	20.5120849	18.5561184	1.651E-65

Interpretation: Metascore helps explains movie's revenue, but it's not directly related to revenue. The relationship between rating and revenue becomes more significant with metascore separated

Votes Analysis

Adding Votes increases model explanatory power 10× and reverses the rating effect.

Adding Votes to Main Regression

Adjusted R^2	0.0422 to 0.462
Rating P	1.58E-10 to 0.01015

Regression Statistics	
Multiple R	0.20787483
R Square	0.04321194
Adjusted R Sc	0.04217981
Standard Erro	93.737921
Observations	929

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	367873.439	367873.439	41.8666101
Residual	927	8145361.6	8786.79784	
Total	928	8513235.04		

	Coefficients	*Standard Erro*	*t Stat*	*P-value*
Intercept	-100.92315	24.0641162	-4.1939273	3.0047E-05
Rating	23.4387969	3.6224418	6.47044126	1.5805E-10

Regression Statistics	
Multiple R	0.679942
R Square	0.462322
Adjusted R	0.46116
Standard E	70.30774
Observatic	929

ANOVA

	df	*SS*	*MS*	*F*
Regressior	2	3935852	1967926	398.1095
Residual	926	4577383	4943.178	
Total	928	8513235		

	Coefficients	*tandard Erro*	*t Stat*	*P-value*
Intercept	57.02995	18.98259	3.00433	0.002733
Vote in mil	526.624	19.60165	26.86631	5.1E-118
Rating	-7.6052	2.9525	-2.57585	0.010153

Adding Votes to Main Regression

	Coefficients	*Standard Erro*	*t Stat*	*P-value*
Intercept	-100.21471	53.9975793	-1.8559112	0.06378687
Rating	20.2587854	8.53314157	2.37412977	0.01779541
Drama	184.063244	48.7940526	3.77224753	0.00017216
Comedy	78.6510285	41.8030298	1.88146718	0.06022497
Action	-104.35273	45.416181	-2.2976994	0.02180328
Adventure	-22.806178	42.794071	-0.5329285	0.59421222
Drama*Ratir	-34.655934	7.50570319	-4.6172801	4.4425E-06
Comedy*Rat	-12.436989	6.25491598	-1.9883543	0.04706943
Action*Ratin	16.1515728	6.98032116	2.31387245	0.02089501
Adventure*R	9.28941673	6.53849999	1.42072597	0.15573637
Metascore	0.16439314	0.16282814	1.00961133	**0.31294786**
Vote in millio	380.624676	20.5120849	18.5561184	1.651E-65

Interpretation:

- Votes – movie's audience size and overall popularity – the strongest predictor

Rating and Votes are correlated

- Votes are included, most of the popularity-based variation is absorbed by Votes
- Rating's coefficient shrinks, becomes **slightly negative, and its p-value increases.**

Interpret and Predict with model

Removing Insignificant Variables?
Adjusted R^2 Dropped

Regression Statistics	
Multiple R	0.76528695
R Square	0.58566411
Adjusted R S	**0.58069389**
Standard Err	62.0210223
Observations	929

Remove Metascore

Regression Statistics	
Multiple R	0.76498598
R Square	0.58520355
Adjusted R S	**0.58068507**
Standard Err	62.0216747
Observations	929

Indirectly explain the data

Revenue=
−100.21+20.25(Rating)+184.06(Drama)+78.65(Comedy)−104.35(Action)−22.80(Adventure)
−34.66(Drama×Rating)
−12.44(Comedy×Rating)+16.15(Action×Rating)+9.29(Adventure×Rating)+0.16(Metascore)+380(Votes)

Movie	Predicted Revenue	Real Revenue	Error
Inside Out	302.46	377.65	75.19
Arrival	164.04	105.2	-58.84
Dunkirk	170.28	192.97	22.69

- **Genre Sensitivity:** Movies like **Inside Out**, more accurately described as animation+family movie, is not accurately explained by formula
- **Special Labels**: **Arrival** is a Sci-fi, non-maintream movie, but the model treats it as drama movie with strong engagement
- **Typical Blockbuster:** Movies with clear genre and public labels, like Dunkirk (Award-winning Action Movie), fits the model's structure