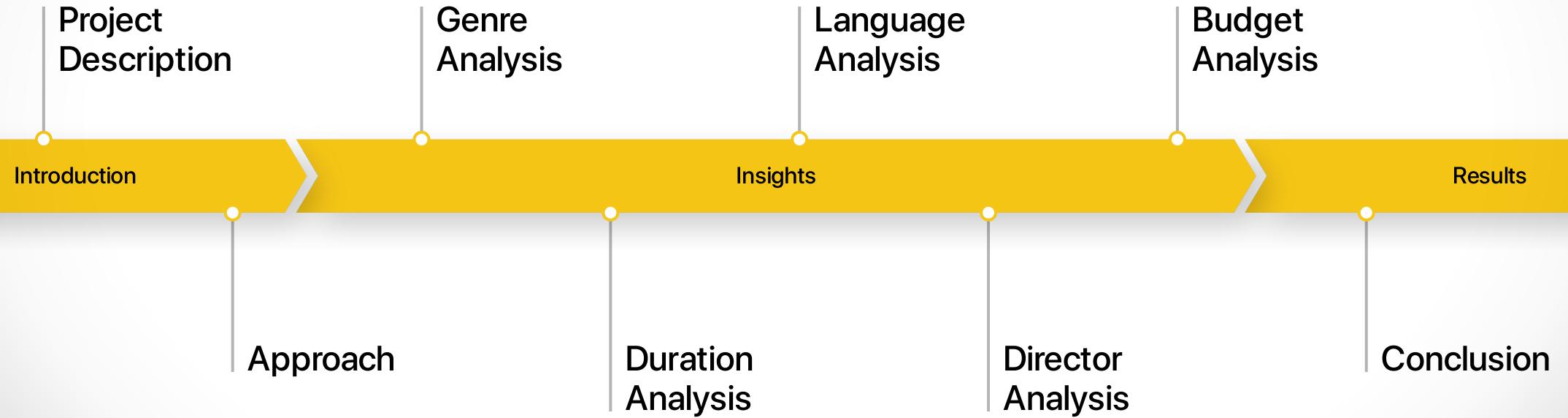


IMDb Movie Analysis

Siddhant Tripathi



agenda



project description

1

Analyzing historical movie
data

3

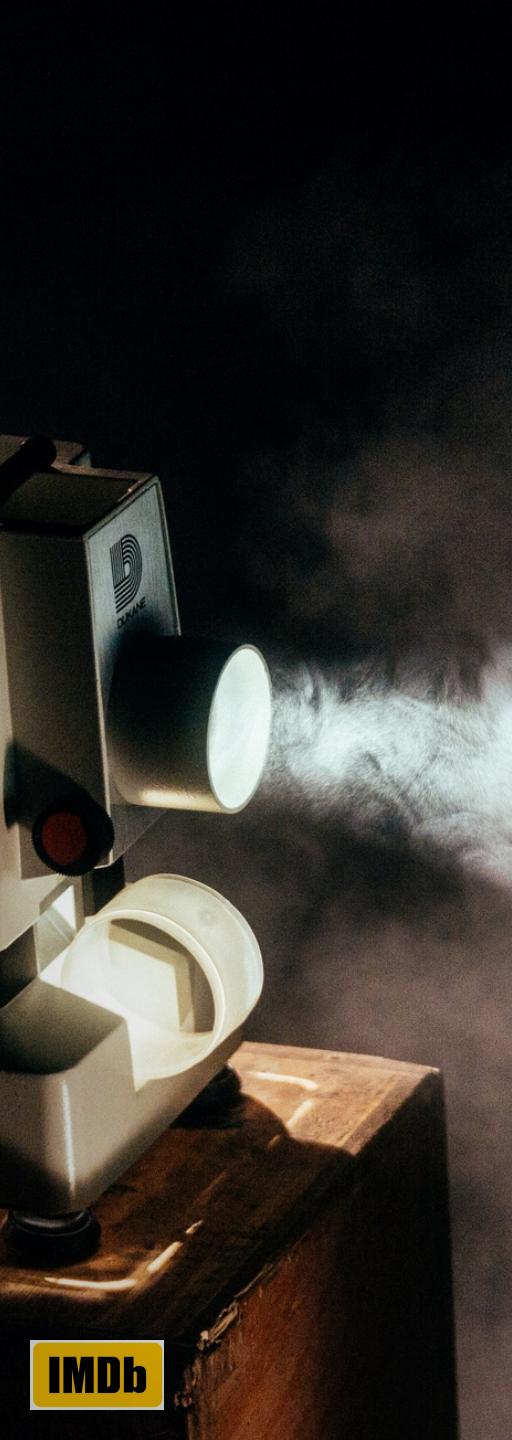
Analyzing factors impacting
movie ratings

2

Handling outliers and
duplicates

4

Summarizing findings to
actionable insights



project approach

- **Tech-Stack Used**
MS Excel for connecting to CSV data using PowerQuery and creating PivotTable and charts
- **Data preprocessing**
Removing unnecessary columns and duplicates to improve data quality and performance
- **Feature extraction**
Split multi-value genre field into separate columns, counted movies in each genre (with replacement) to enable genre-specific analysis



project insights



Genre Analysis

Top genres by movie count



Duration Analysis

Score-duration descriptive stats



Language Analysis

Language distribution



Director Analysis

Top directors by movies & score



Budget Analysis

Budget-gross correlation

genre analysis

Top 10 genres by number of movies

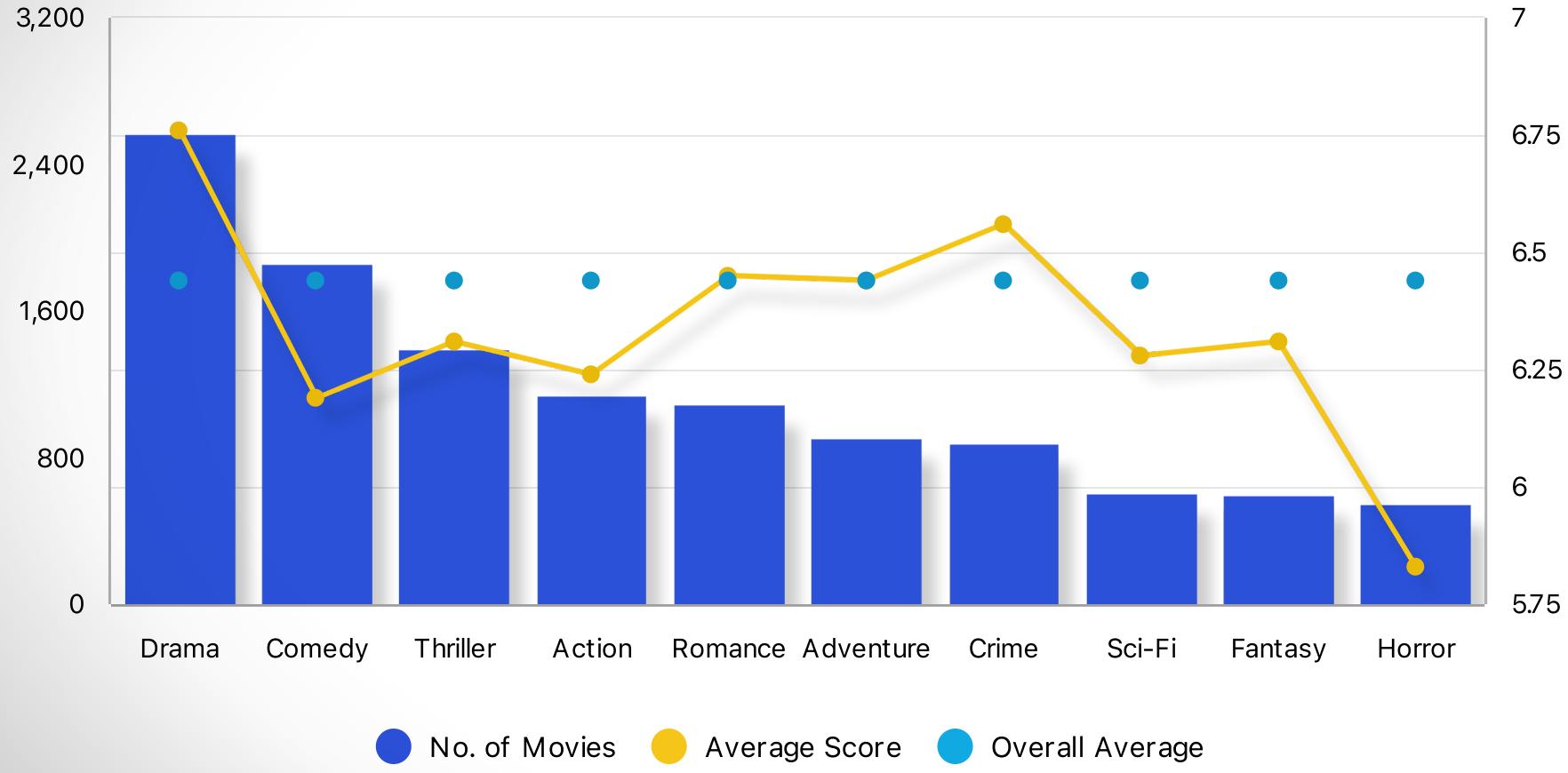
TOP
10

genres	movies	average score	median score	mode score	range	variance	std dev
Drama	2,571	6.76	6.9	7.2	7.3	0.91	0.95
Comedy	1,862	6.19	6.3	6.7	7.8	1.19	1.09
Thriller	1,393	6.31	6.4	6.1	6.8	1.11	1.05
Action	1,143	6.24	6.3	6.1	7.4	1.24	1.11
Romance	1,097	6.45	6.5	6.5	6.5	1.00	1.00
Adventure	914	6.44	6.6	6.7	7	1.27	1.13
Crime	883	6.56	6.6	6.6	6.9	1.05	1.02
Sci-Fi	611	6.28	6.4	6.7	6.9	1.46	1.21
Fantasy	604	6.31	6.4	6.7	7.2	1.34	1.16
Horror	556	5.83	5.9	6.2	6.5	1.26	1.12



Note: Movies with multiple genres were counted under each listed genre

genre analysis



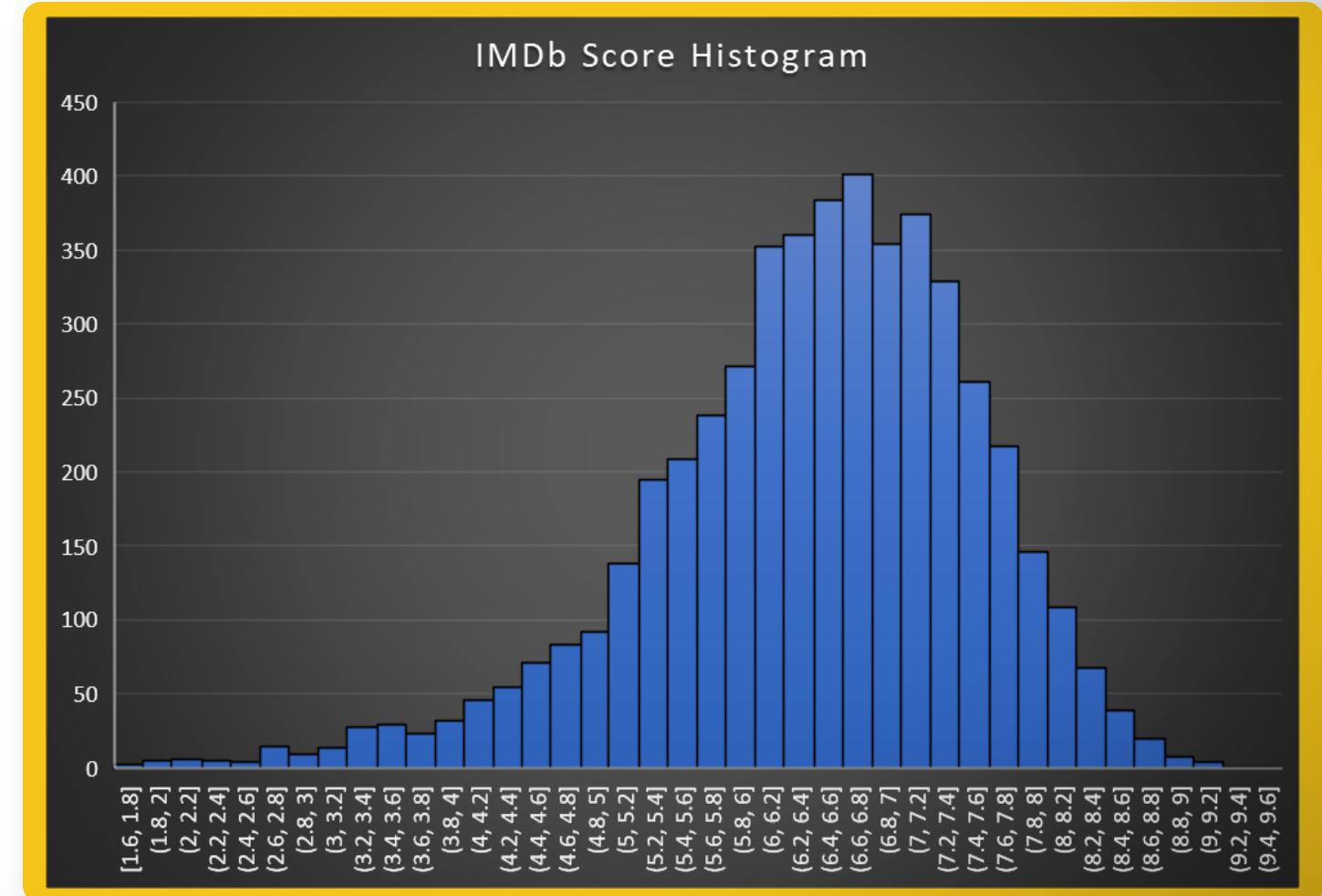
- Identified top 10 genres by movie count; Drama had most movies, Horror had least
- Drama and Crime outperformed overall avg IMDb score; Romance/Adventure at par with avg; Horror performed worst
- Sci-Fi had highest std dev of top 10, indicating most polarizing genre
- All top 10 genres had wide score ranges (≥ 6.5), suggesting presence of outliers

duration analysis

Score distribution

Mean score	6.44
Median score	6.6
Mode score	6.7
Min score	1.6
Max score	9.5
Std Dev	1.12
Pearson's second coeff	-0.42
Kurtosis	0.94
Excess Kurtosis	-2.058

- Scores normally distributed
- Moderate negative skew
- Platykurtic (short-tailed) - fewer outliers

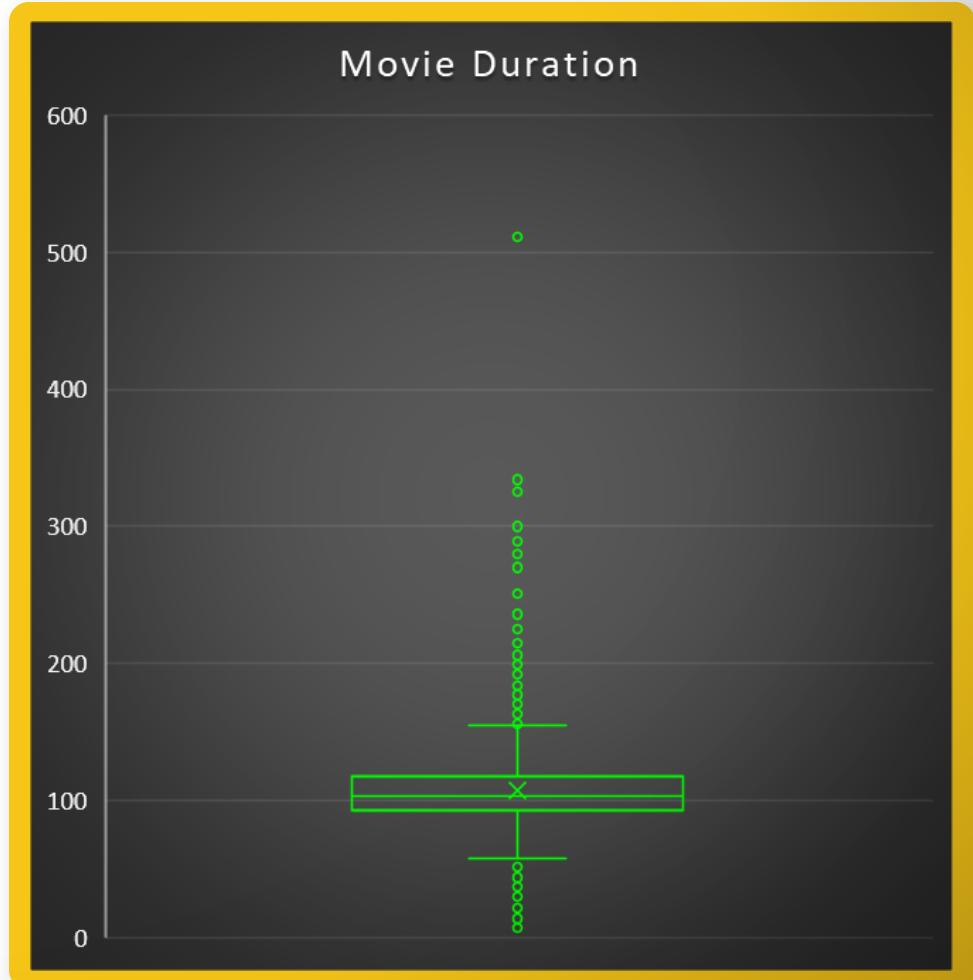


duration analysis

Duration distribution

Mean duration	107.21
Median duration	103
Mode duration	90
Min duration	7
Max duration	511
Std Dev	25.25
Pearson's second coeff	0.50
Kurtosis	22.57
Excess Kurtosis	19.568

- Duration ranges from 7 (short film) to 511 (series)
- Moderate positive skew
- Leptokurtic (heavy-tailed) - large number of outliers as seen in box plot

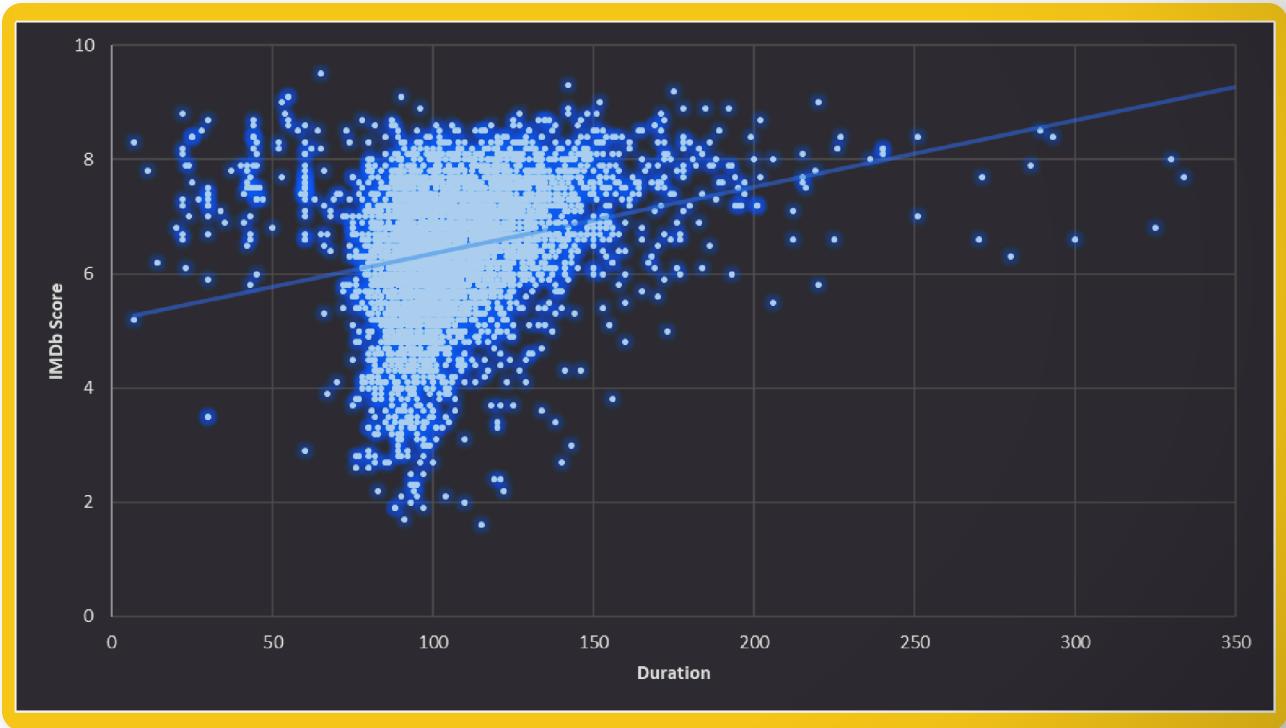


duration analysis

Score-duration correlation

Correl	0.2626183
Correl (excluding duration outliers)	0.3181839

- Most of the movies have a duration of 50-150
- Weak positive correlation between duration and score
- No significant change in correlation after excluding duration outliers



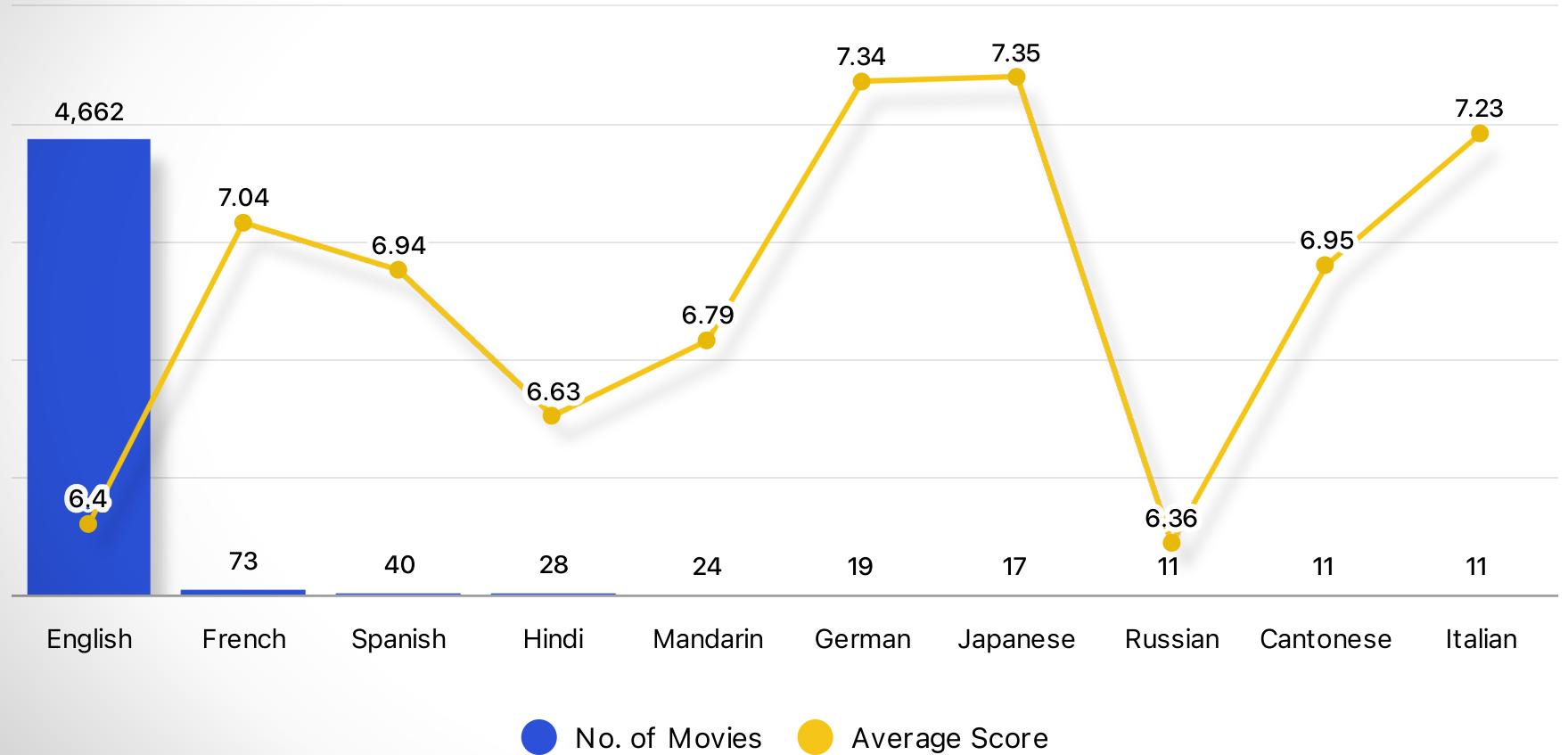
language analysis

Top 10 languages by number of movies



language	movies	average score	median score	std dev
English	4,662	6.40	6.5	1.12
French	73	7.04	7.2	0.72
Spanish	40	6.94	7.15	0.84
Hindi	28	6.63	6.95	1.37
Mandarin	24	6.79	7.05	1.02
German	19	7.34	7.6	0.93
Japanese	17	7.35	7.5	0.97
Russian	11	6.36	6.5	1.32
Cantonese	11	6.95	7.2	0.67
Italian	11	7.23	7.3	1.19

language analysis



- Identified top 10 languages by movie count; English dominated with more than 93% of the database
- German and Japanese had highest avg IMDb scores; English and Russian had lowest
- Hindi and Russian had highest std dev of scores among top languages

director analysis

Considering directors with at least 5 movies

Top 5 directors by score

director	movies	average score	percentile rank
Christopher Nolan	8	8.43	0.995
Quentin Tarantino	8	8.20	0.987
Frank Capra	5	8.06	0.977
Stanley Kubrick	7	8.00	0.971
James Cameron	7	7.91	0.969

Among directors with at least 5 movies, Christopher Nolan ranks the highest followed by Tarantino

Top 5 directors by movie count

director	movies	average score	percentile rank
Steven Spielberg	26	7.48	0.889
Woody Allen	22	7.01	0.755
Martin Scorsese	20	7.66	0.932
Clint Eastwood	20	7.23	0.835
Ridley Scott	17	7.07	0.762

- Spielberg ranks in top 12% and Scorsese in top 7% by avg IMDb score despite high volume of 26 and 20 films resp.
- Their acclaimed filmographies demonstrate ability to produce highly-rated films consistently over prolific careers.

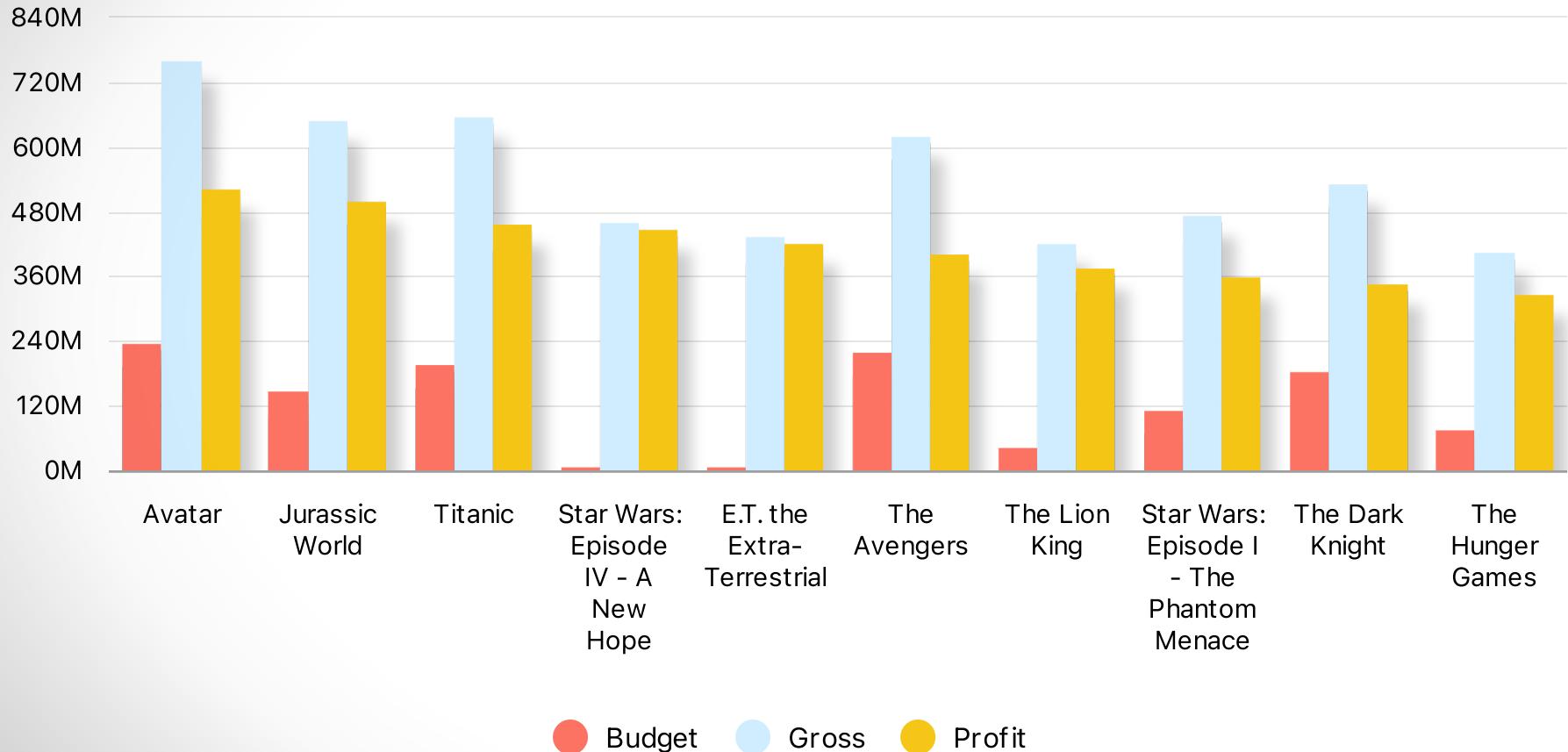
budget analysis

Top 10 movies by profit

TOP
10

	movie	budget	gross	profit
	Avatar	237M	761M	524M
	Jurassic World	150M	652M	502M
	Titanic	200M	659M	459M
	Star Wars: Episode IV - A New Hope	11M	461M	450M
	E.T. the Extra-Terrestrial	11M	435M	424M
	The Avengers	220M	623M	403M
	The Lion King	45M	423M	378M
	Star Wars: Episode I - The Phantom Menace	115M	475M	360M
	The Dark Knight	185M	533M	348M
	The Hunger Games	78M	408M	330M

budget analysis



Budget-Profit correlation

Overall = 0.101

Budget \leq 300M = 0.648

- Stronger correlation between budget and earnings for lower budget films (\leq 300M)
- Weaker, nonlinear correlation for mega-budget blockbusters ($>$ 300M)
- Budget less predictive of earnings past 300M threshold
- Diminishing returns of higher budgets on earnings

Conclusion



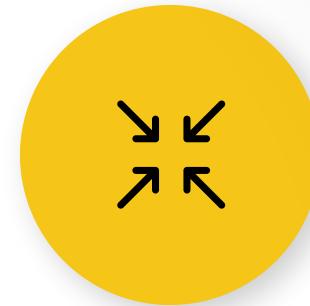
Key Insights

- Genre performance vs average
- Budget-profit correlation threshold
- Top directors' quality & volume



Skills Demonstrated

- Feature extraction for analysis
- Correlation and distribution skewness and kurtosis
- Formulas, pivot tables and visualizations



Impact

Enhanced data analytics skills for manipulation, visualization, statistical analysis, quantification of relationships

Project Workbook available [Here](#)