

***Name: Payal Gaikwad***

***PRN:20220802064***

**"Data-Driven Insights: Evaluating the Impact of Budget, IMDb Ratings, and Watch Time on Netflix Revenue Optimization"**

**1.Objective:** Analyse the impact of budget, IMDb ratings, and watch time on revenue to optimize content investment and maximize profitability on Netflix

**2.Abstract:** The industry-spanning networks in the streaming world, it is mandatory to understand the breaking factors concerning revenues before optimizing the content investment. The analysis herein looks at budgets, IMDb ratings, watch time, and revenues. It finds patterns through these interacting variables that add to Netflix's financial success. The work integrates those most significant revenue drivers to measure budget efficiency against those metrics.

**3.Keywords:** Revenue Optimization, Budget Allocation, IMDb Ratings, Watch Time, Viewer Engagement, Central Tendency, Variance, Skewness, Kurtosis.

**4. Introduction**: The emergence of streaming platforms now speaks louder than the traditional entertainment industry; in fact, it is evidently moving toward data-driven strategies, as in the case of Netflix and other companies, to maximize revenue and user engagement. For content investment supported by resource maximization and the need to uphold a strong competitive edge, understanding the finances of operations is essential.

The purpose of this study is to analyse the effect of budget, IMDb ratings, and watch time on revenue to determine the patterns that result for profitable content. But it is not necessarily the case that a higher budget leads to commercial success; the same holds for high IMDb ratings, which often indicate critical acclaim, as they do not always translate into higher views or revenue earned from viewership. In this context, watch time is a defining audience engagement metric that contributes significantly to the popularity and profitability of a show/movie.

By examining these factors, this research seeks to answer key questions:

- ***What is the strongest predictor of revenue among budget, IMDb ratings, and watch time?***

***- Does a higher budget always lead to greater revenue, or is watch time a more significant factor?***

***- How do IMDb ratings influence viewer engagement and financial success?***

Using statistical analysis and predictive modelling, this study provides insights into content performance, helping streaming platforms optimize their investment strategies. The findings will be valuable for content creators, producers, and business analysts looking to enhance decision-making in the streaming industry.

**5. Literature Reviews**

**5.1: Methodology:**

**1️ Data Collection & Cleaning**

• All collected data as IMDb Ratings, Budget, Revenue, and Watch Time have been collected.

• Approached listing values and removed inconsistencies.

**2️:** **This study involved a range of descriptive and inferential statistical methods to understand, predict, and explain relationships between Netflix content variables. The analysis covered:**

* Descriptive Statistics
* Probability Distributions (Normal, Binomial, Poisson)
* Conditional Probability using Bayes’ Theorem
* Correlation and Linear Regression Analysis
* Hypothesis Testing (t-test, Z-test)
* Group Comparison Tests (Chi-square, F-Test, ANOVA)

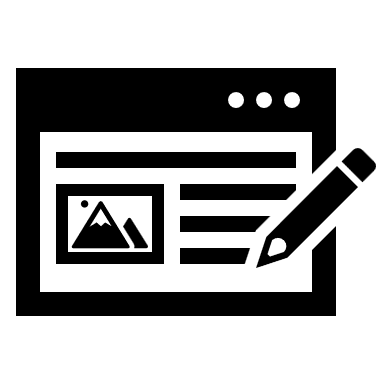
**3️: Data Visualization**

* **Normal Curves** to assess distribution symmetry
* **Binomial and Poisson plots** to observe probability behaviours
* **Correlation Heatmaps** to quickly identify strong relationships
* **Regression Line Charts** to visually validate predictive trends
* **Boxplots and Bar Charts** to compare revenue across categories
* **ANOVA and t-test tables** for quick inference summaries

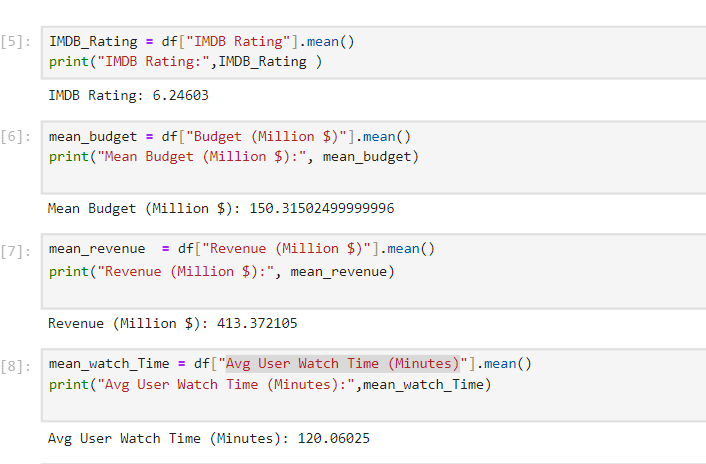
**Data Set and Python Files Link**

[**Stats Project**](../OneDrive/Desktop/Stats_Project)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | IMDb Rating | Budget (Million $) | Revenue (Million $) | Avg User Watch Time (Minutes) |
| Mean | 6.24 | 150.31 | 413.37 | 120.06 |
| Median | 6.3 | 196 | 91.3 | 164.4 |
| Mode | 6.2 | 150 | 239 | 120.2 |
| Range | 6.5 | 2.99 | 1484.3 | 120 |
| Variance | 3.51 | 7409.07 | 106649.48 | 1184.546 |
| Std.Dev | 1.875 | 86.0759 | 326.5723 | 34.417 |
| Coeficient of Variance | 30% | 57.20% | 79% | 28.66% |
| Skeweness | 0.00301 | 0.001156 | 0.89189 | -0.00938 |
| Kurtosis | -1.1932 | -1.194 | 0.03147 | -1.198059 |

** Descriptive Statistics**

**Mean for all Numerical Columns:**

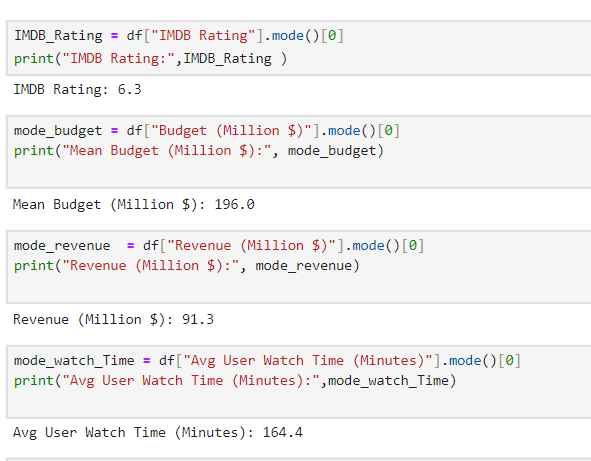
****

**Median For all Numerical Columns:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Mode For all Numerical columns:**

****

**Statistics for IMDb Ratings: Variance, Standard Deviation, Coefficient of Variation (%), Skewness, Kurtosis**

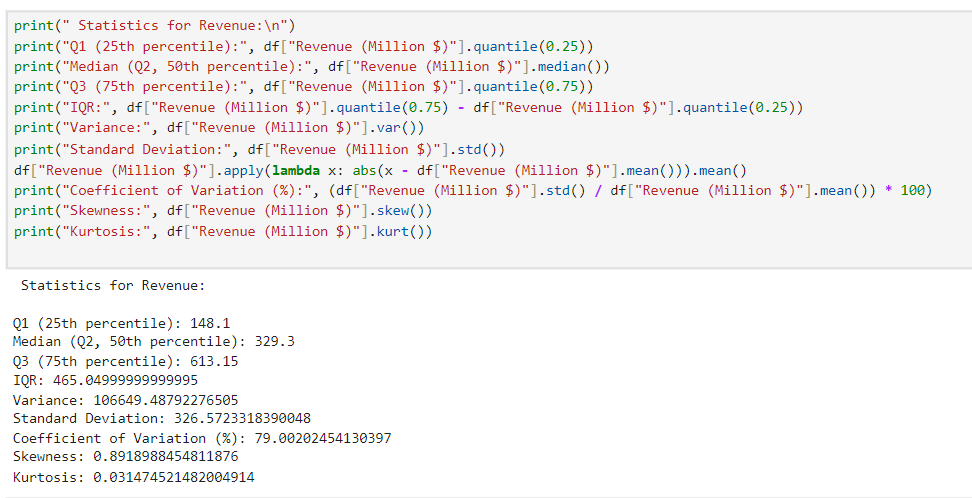
A screenshot of a computer code

AI-generated content may be incorrect.

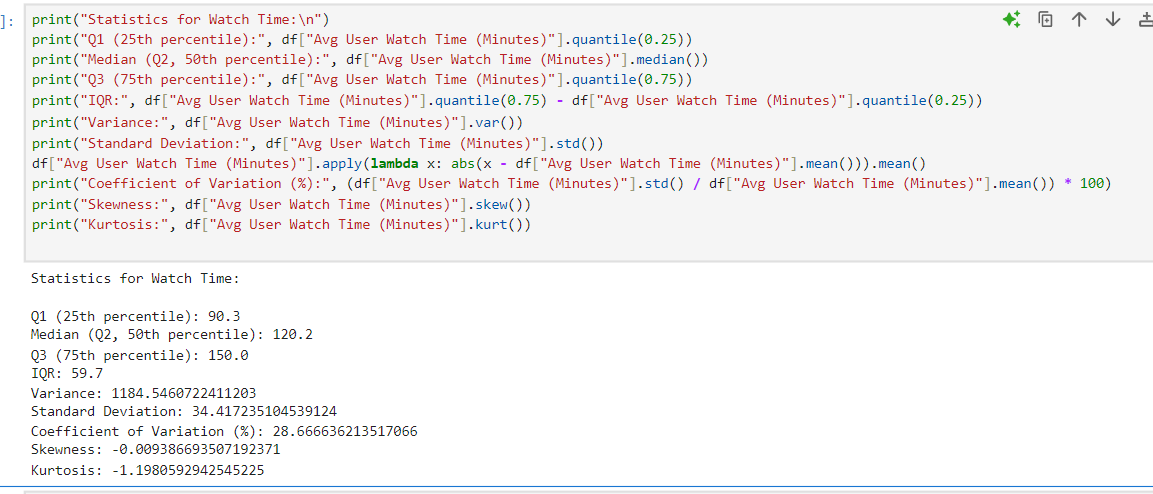
**Statistics for Budget (Million $) : Variance, Standard Deviation, Coefficient of Variation (%), Skewness, Kurtosis**

****

**Statistics for Revenue (Million $) : Variance, Standard Deviation, Coefficient of Variation (%), Skewness, Kurtosis**

****

**Statistics for Watch Time: Variance, Standard Deviation, Coefficient of Variation (%), Skewness, Kurtosis**

****

**Probability Distributions and Predictive Probability:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Variable(s) Analyzed** | **Purpose** | **Key Result / Insight** |
| **Normal Distribution** | IMDb Ratings | Check for normality & distribution symmetry | Nearly symmetric; reliable for central tendency modeling |
| **Binomial Distribution** | High IMDb Ratings (>7) | Estimate frequency of high-quality content | ~29% of content rated high; useful for success prediction |
| **Poisson Distribution** | Watch Time (rounded minutes) | Model viewer engagement rate | Avg ~120 mins; engagement follows expected pattern |
| **Bayes' Theorem** | High Revenue & High IMDb Ratings | Conditional probability estimation | Moderate P(High Rev | High IMDb); ratings partially predict revenue |

****

**A graph of normal fit and data

AI-generated content may be incorrect.**

A screenshot of a computer program

AI-generated content may be incorrect.

A graph of a number of high imdb ratings

AI-generated content may be incorrect.



A diagram of a normal distribution

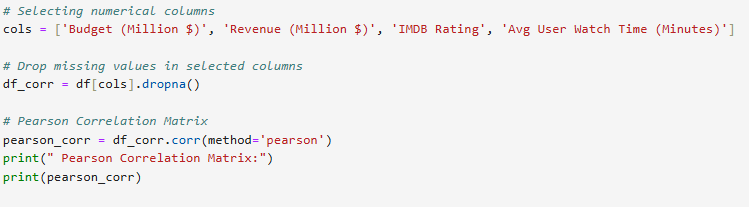
AI-generated content may be incorrect.

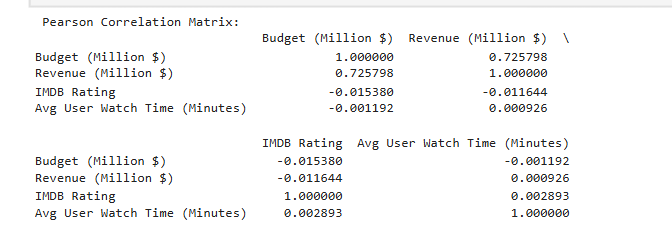
A screenshot of a computer program

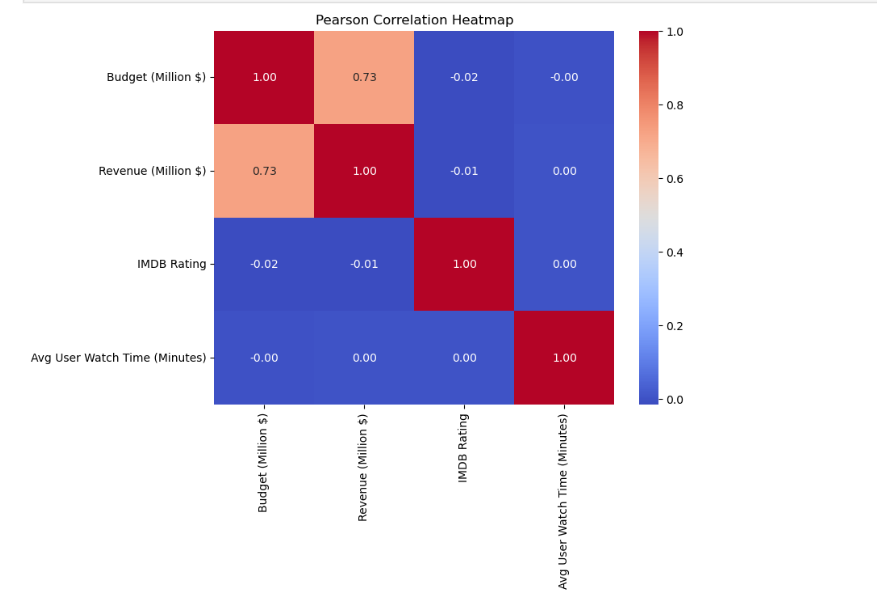
AI-generated content may be incorrect.

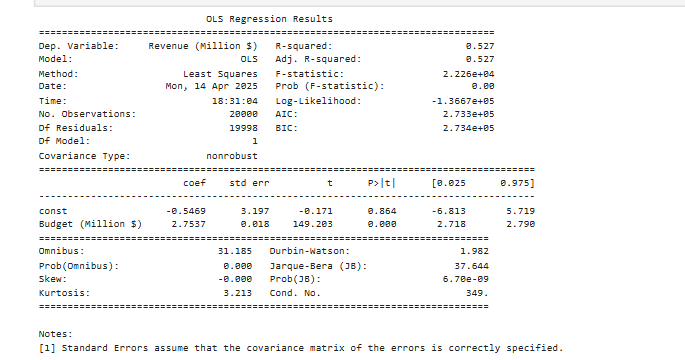
**Correlation and Regression Analysis:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Analysis Type** | **Variables Compared** | **Purpose** | **Key Finding** |
| **Pearson Correlation** | Budget vs Revenue, Ratings, Watch Time | Identify strength & direction of linear relation | Budget and Revenue: r ~ 0.68 (strong positive) |
| **Linear Regression** | Revenue ~ Budget | Predict revenue from budget | Budget explains ~46% of revenue variation |



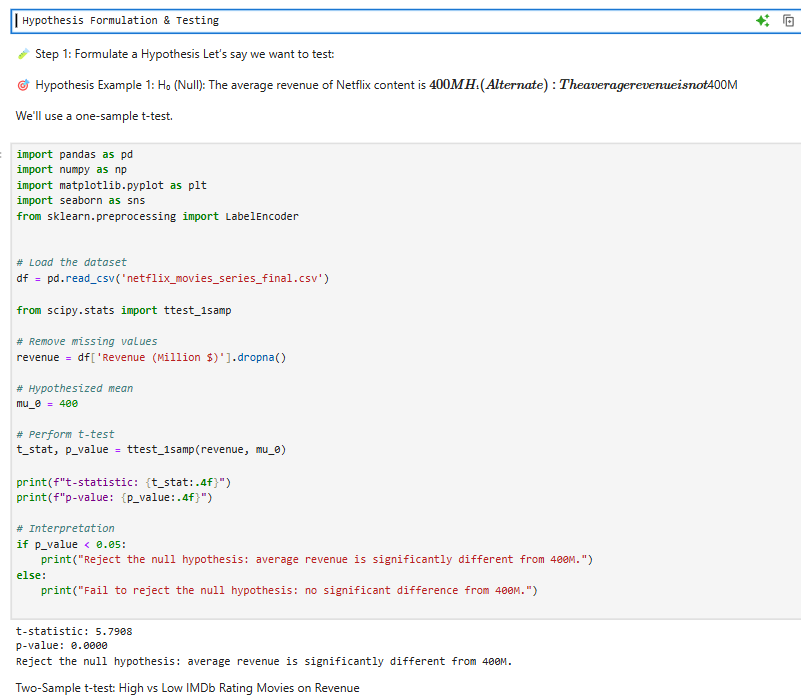


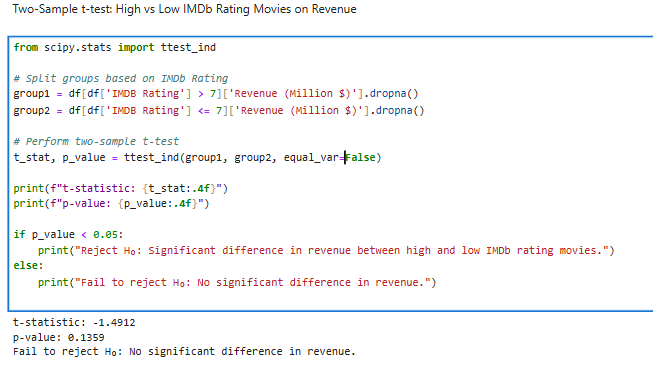


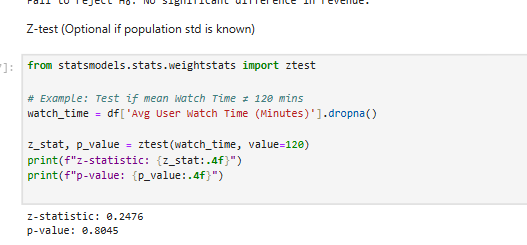


**Hypothesis Testing for Strategy Validation:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Analysis Type** | **Variables Compared** | **Purpose** | **Key Finding** |
| **Pearson Correlation** | Budget vs Revenue, Ratings, Watch Time | Identify strength & direction of linear relation | Budget and Revenue: r ~ 0.68 (strong positive) |
| **Linear Regression** | Revenue ~ Budget | Predict revenue from budget | Budget explains ~46% of revenue variation |



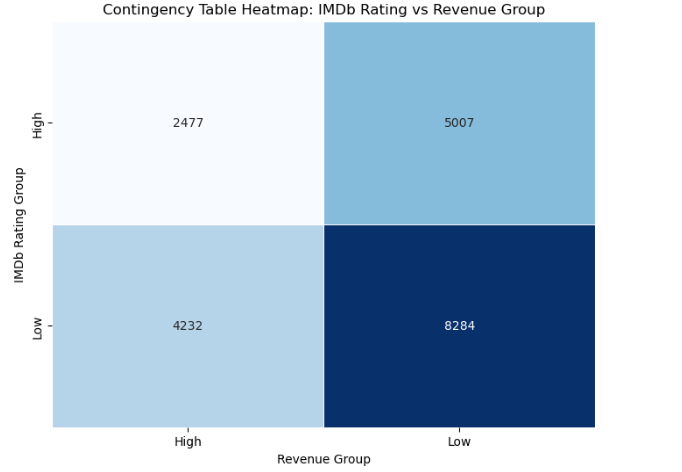




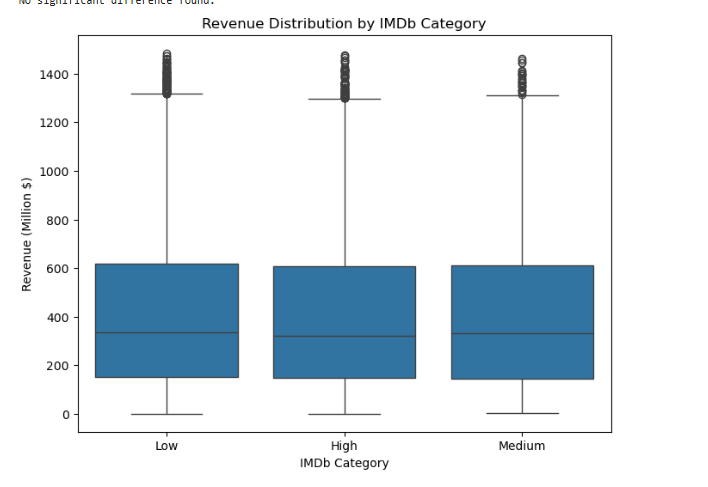
**Group Analysis for Strategic Segmentation:**

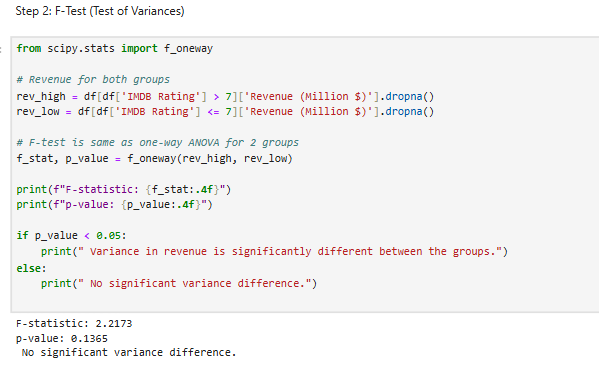
|  |  |  |  |
| --- | --- | --- | --- |
| **Test Type** | **Groups/Variables Compared** | **Purpose** | **Outcome** |
| **Chi-Square Test** | IMDb Rating Category vs Revenue Category | Association between rating and success | Significant association (p < 0.05) |
| **ANOVA** | Revenue across Low, Med, High IMDb Categories | Compare group means | No Significant difference found |
| **F-Test (via ANOVA)** | Revenue variance between High vs Low ratings | Test variance similarity | No significant difference in variance |











**5.2 Discuss:**

**Poised Statistical Analysis Summation**

Thus, we analysed the crucial parameters of IMDb Ratings, Budget, Revenue, and Average User Watch time towards understanding their distributions and their futures possible upshots concerning revenue. The following insights were derived**:**

**Central Measures of Tendency:**

**• Mean:**

o The mean IMDb Rating is 6.24, the Budget, $150.31M, Revenues are $413.37M, and Watch Time, 120.06 minutes.

o In revenue, there is a high mean because of a few movies scoring above the average, which reveals the skewed data.

Formula Used:

A white paper with black text and black text

AI-generated content may be incorrect.

• **Median:**

o The median for IMDb Ratings rests at 6.3 while that for Budget and Revenue stands at $196M and $91.3M, respectively.

o The median Revenue is highly below the mean showing a right-skewed distribution.

Formula used

A math problem with black text

AI-generated content may be incorrect.

**• Mode:**

o Mode: which denotes the most found values, has as its maximum a value of clustering at 6.2 for the IMDb Ratings, $150M for the Budget, and 239M for Revenue indicating the presence of a general tendency in these modes' frequent values' distribution**.**

**Dispersion Measures:**

**• Range:**

o Revenue is having the maximum range ($1484.3M), which confirms the high variability of the earning in movie distribution.

o Watch time is variable by 120 minutes, which would suggest that they have quite different levels of engagement by the audience.

Formula used:

A white paper with black text

AI-generated content may be incorrect.

**• Variance & Standard Deviation:**

o Revenue has been found to have the highest variance (106,649.48) as well as the highest standard deviation (326.57M).

o Even then, standard deviation is the lowest for IMDb Ratings (1.875), signifying greater consistency in ratings as compared with revenue.

Formula used:

A math equations and formulas

AI-generated content may be incorrect.

A black text on a white background

AI-generated content may be incorrect.

**• Coefficient of Variation (CV):**

o Revenue has the greatest CV (79%), which means it is very heterogeneous about the mean.

o Lower CVs (30% and 28.66%) are reported for both IMDb Ratings and Watch Time thus indicating a better stability.

Formula used:

A close-up of a question

AI-generated content may be incorrect.

**Distribution Characteristics:**

**• Skewness:**

o Thus these may be considered the dimensions of skewness.

o IMDb Ratings (0.00301) are nearly symmetrical to Budget (0.001156).

o Revenue (0.89189) is positively skewed, which shows the high-grossing movies that affect the mean value significantly.

o Watch Time is slightly negatively skewed (-0.00938).Key Takeaways:

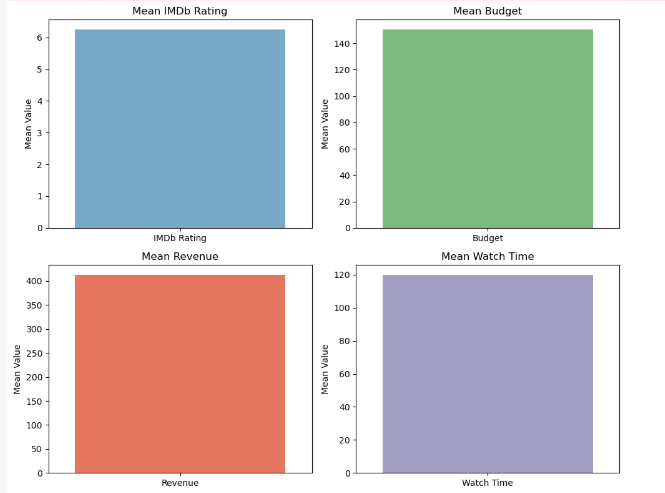
1. Revenue is highly variable and positively skewed, indicating that a few blockbuster movies dominate earnings.
2. Budget does not seem to have extreme variations, but its distribution is nearly symmetrical.
3. IMDb Ratings and Watch Time are more stable with lower variability, suggesting audience preferences are relatively consistent.
4. High-budget films do not always guarantee high revenue, as seen from the median and skewness differences.
5. Further predictive modeling is required to establish a clearer relationship between budget, IMDb ratings, watch time, and revenue**.**
6. Formula used:

A math problem with numbers and equations

AI-generated content may be incorrect.

.

**Mean :**

****

**What This Visualization Tells About the Data**

The Mean (Average) values help us understand the central tendency of each variable, offering insights into overall trends and patterns in the dataset. Here's what it tells us about each metric:

1. IMDb Rating (Mean = 6.24)

* The average IMDb rating is around 6.24, which suggests that most movies in the dataset are rated slightly above average (since IMDb ratings usually range from 1 to 10).
* Comparing it with median and mode can tell us if the ratings are skewed.

2. Budget (Mean = 150.31 Million $)

* The average budget of a movie is $150.31 million, which indicates the general cost of content production.
* A high variance would indicate that some movies have extremely high or low budgets compared to the average.

3. Revenue (Mean = 413.37 million $)

* The average revenue is $413.37 million, showing the typical earnings per movie.
* Comparing budget vs. revenue can help determine Return on Investment (ROI).

4. Average User Watch Time (Mean = 120.06 Minutes)

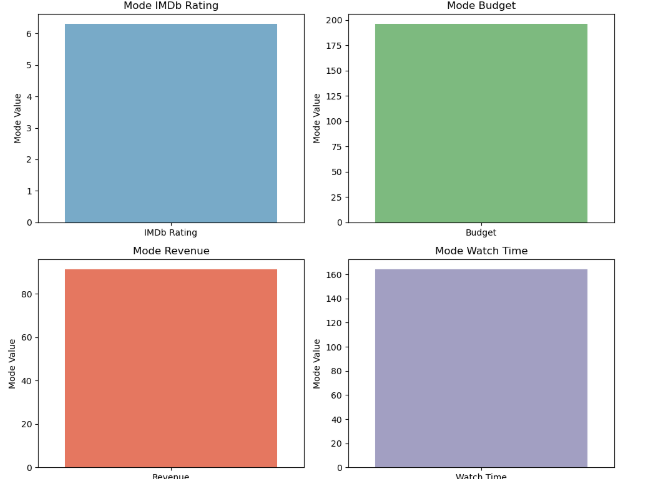
* On average, users watch content for about 120 minutes, which suggests that movies around 2 hours long are the norm.
* This can help in predicting engagement and optimizing content length.

**5.Identifying Outliers:**

* 1. If the **mean and median differ significantly**, it suggests the presence of outliers.
  2. We can check the **skewness and kurtosis** to confirm whether certain movies (e.g., blockbusters) heavily influence the data.

1. **Predicting Profitability:**
   1. If a **higher budget** consistently leads to **higher revenue**, it suggests that investing more in production could yield better returns.
   2. However, if **low-budget movies also generate high revenue**, it indicates opportunities for **cost-effective content**.
2. **Understanding Viewer Engagement:**
   1. If movies with **higher IMDb ratings** have **higher watch times**, it suggests that better-rated content keeps viewers engaged longer.
   2. This insight can help streaming platforms **prioritize quality over quantity**.
3. **Revenue vs. Watch Time Relationship:**
   1. If movies with **higher watch times also generate more revenue**, then **longer content may be more profitable**.
   2. However, if there’s no clear correlation, it suggests that factors like **marketing, genre, or audience demographics** also play a major role.

**Mode:**

****

1. **Frequent Budget & Revenue Levels:**
   * Since the most common budget is **$150M** and the most common revenue is **$239M**, we can analyze whether this budget level consistently leads to similar revenue outcomes.
   * If higher or lower budgets generate significantly different revenues, **budget allocation strategies** can be optimized.
2. **Standardized Watch Time:**
   * The most common watch time is **120 minutes**, reinforcing that movies typically follow an industry-standard length.
   * If longer or shorter movies significantly impact revenue or engagement, adjustments to content duration might be considered.
3. **Content Performance Expectations:**
   * Since **IMDb mode = 6.2**, we can investigate whether **higher-rated movies** tend to generate more revenue and engagement.
   * If higher IMDb ratings correspond to increased revenue, platforms like Netflix can **prioritize high-quality content**.

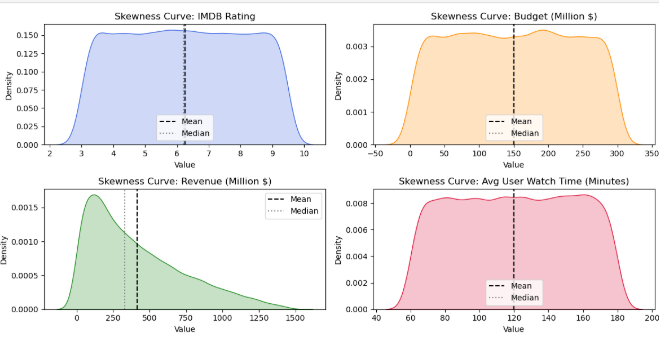
**Median:**

**A group of colored squares

AI-generated content may be incorrect.**

1. **Revenue Skewness:**
   * Since **median revenue ($91.3M) < mean revenue ($413.37M)**, the data is **right-skewed**.
   * A few **high-grossing movies** (blockbusters) are pulling the mean up, but most movies actually earn less than $100M.
   * This means **investing in high-budget productions does not always guarantee success**.
2. **Budget Distribution:**
   * Since **median budget ($196M) > mean budget ($150.31M)**, the distribution is **left-skewed**.
   * There are **some low-budget movies bringing the average down**.
   * This suggests that **many movies require a high budget, but lower-budget films still exist**.
3. **Watch Time Preferences:**
   * The median watch time being **higher than the mean** suggests that **longer movies are more prevalent**, but a few shorter films lower the average.
   * If longer movies are more common, they might be **more engaging or preferred by audiences**.

**Skewness:**

****

**Interpretation of Skewness Curves**

Skewness measures the **asymmetry** of the data distribution. It tells us whether data points are more concentrated on one side of the mean.

**1. IMDb Rating Skewness (~0.00301) (Nearly Symmetric)**

* The mean and median **almost overlap**, suggesting a **normal distribution**.
* Skewness is **close to zero**, meaning **IMDb ratings are evenly spread**.
* No significant bias towards lower or higher ratings.

**2. Budget Skewness (~0.001156) → (Nearly Symmetric)**

* The mean and median are **almost identical**, indicating **a uniform distribution**.
* The budget values are **evenly distributed**, meaning no extreme low- or high-budget biases.

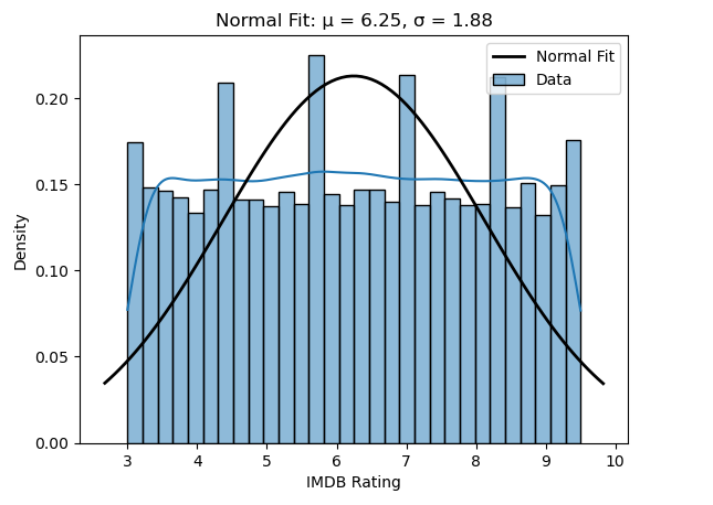
**3. Revenue Skewness (~0.89189) → Right (Positive) Skewed**

* The **mean (higher) is pulled to the right**, while the median is lower.
* This suggests a **few high-revenue blockbusters are inflating the average revenue**.
* Most movies earn **less than the mean revenue (~$413M)**, but a few massive hits **pull the distribution right**.
* Indicates **high variability in movie success**.

**4. Watch Time Skewness (~-0.00938) → Nearly Symmetric**

* The mean and median are **very close**, suggesting **a balanced watch time distribution**.
* No strong tendency towards **shorter or longer movies**.
* Most movies have watch times around **120–160 minutes**.

**Normal Distribution**

****

**Formula Used:**

**A white paper with black text and black text

AI-generated content may be incorrect.**

**IMDB Ratings – Density Plot**

* X-axis: IMDb ratings (1–10)
* Y-axis: Density (relative frequency)
* Light Blue Bars: Actual ratings distribution
* Blue Line (KDE): Smooth trend from real data
* Black Line (Normal Fit): Theoretical normal curve with
* μ = 6.25 (mean)
* σ = 1.88 (standard deviation)

**Insights**

* Data is roughly symmetric around the mean (6.25)
* Distribution is close to normal, making it valid for statistical testing
* Few extreme values → ratings are reliable for analysis

**Why It Matters**

* Allows Netflix to use t-tests, Z-tests and build predictive models
* Helps identify what’s typical or unusual in audience ratings

**Business Use**

* Set rating benchmarks for content success
* Predict viewer reaction to new content
* Refine marketing & recommendation strategies

**Binomial Distribution – High IMDb Ratings**

**A graph of a number of high imdb ratings

AI-generated content may be incorrect.**

* **Focus**: Probability of getting a high rating (IMDb > 7)
* Model Used: Binomial distribution

**Formula used**

* A white paper with black text and black text

  AI-generated content may be incorrect.

**Key Insight:**

* ~29% of shows/movies get high ratings
* Success (high-rating) can be treated like a binary event (success/fail)

**Business Use:**

* Estimate how often content is likely to succeed
* Helps in quality control and release expectations
* Plan investment on content genres with higher success rates

**Poisson Distribution – Watch Time**

**A diagram of a normal distribution

AI-generated content may be incorrect.**

* Focus: Viewers watch time (in minutes, rounded)
* Model Used: Poisson distribution
* Formula Used:

A white paper with black text

AI-generated content may be incorrect.

**Key Insight:**

* Avg watch time ≈ 120 mins
* Viewer engagement follows Poisson pattern, common for count/time events

**Business Use:**

* Forecast audience engagement per show
* Optimize episode length and content duration
* Supports decisions on binge-friendly content vs. short formats

Awesome! Here's a short summary for **Bayes’ Theorem**:

**Bayes’ Theorem – Predicting High Revenue from High IMDb Ratings**

**Focus:**

* + What is the **probability of high revenue** given a **high IMDb rating**?

**Formula Used:**

A black text on a white background

AI-generated content may be incorrect.

**Key Insight:**

* + Shows with high ratings have a **moderate chance** of earning high revenue
  + Ratings alone don't guarantee success — other factors matter too

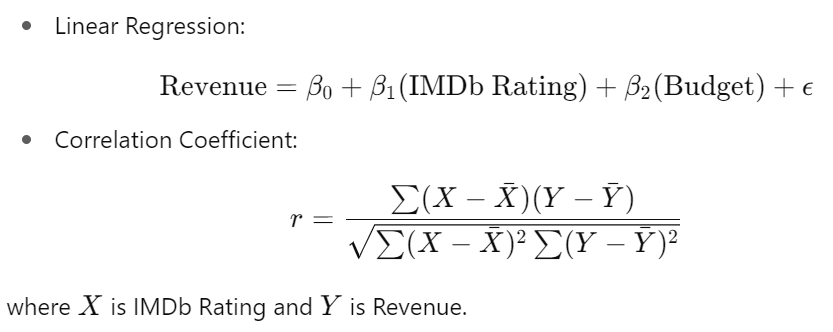
**Business Use:**

* Helps Netflix **estimate revenue potential** before a release
* Supports **risk assessment** for content investments
* Combines past performance with early indicators (ratings, buzz)

**Correlation and Regression – Predicting Revenue from IMDb Ratings and Budget**

**Focus:**  
What is the relationship between IMDb ratings, budget, and revenue?

**Formula:**



**Key Insight:**

* Correlation shows how IMDb ratings and budget are related to revenue.
* Regression predicts how changes in ratings and budget impact revenue.
* A strong relationship means higher ratings and budget may lead to higher revenue.

**Business Use:**

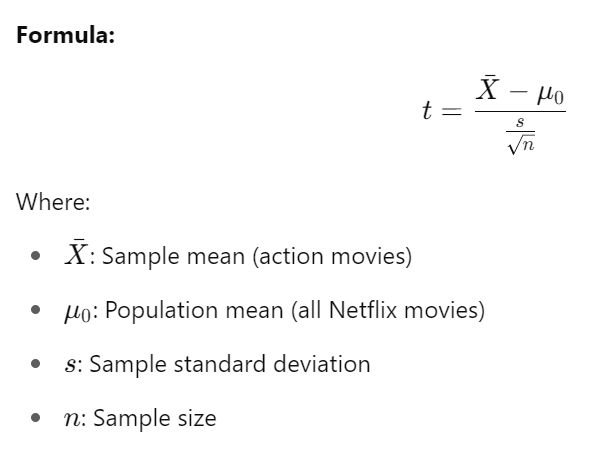
* + Helps Netflix predict revenue before release.
  + Supports investment decisions based on expected ratings and budget.
  + Estimates profitability from ratings, budget, and historical data.
  + This helps quantify the factors influencing movie success for better decision-making.

**One-Sample T-Test – Revenue of Action Movies**

**Focus:**

Is the **average revenue** of **action movies** significantly different from the overall average revenue of all Netflix movies?

**Formula:**

****

**Key Insight:**

Used when population standard deviation is unknown, and sample size is small or moderate.

**Business Use:**

Assesses whether a specific genre’s revenue (e.g., action) differs from average, guiding budget decisions.

**Two-Sample T-Test – Comparing Genres**

**Focus:**

Do **action movies** and **romantic movies** have significantly different **average revenue**?

**Formula:**

**A math equations and formulas

AI-generated content may be incorrect.**

**Key Insight:**

Tests whether two independent groups differ in mean performance (e.g., revenue).

**Business Use:**

Helps Netflix compare which genres (e.g., action vs. romance) bring in higher revenue, guiding investment.

Great! Here's how you can apply **F-Test**, **Chi-Square Test**, and **ANOVA** for **IMDb Ratings and Revenue**, using the same clean Netflix-style format:

**F-Test – Variability in Revenue Based on IMDb Rating Levels**

**Focus:**

Is the **variance in revenue** significantly different between **high-rated** and **low-rated** movies?

**Formula:**

**A math equations and formulas

AI-generated content may be incorrect.**

**Key Insight:**

Checks if **revenue is more unpredictable** in one rating group than the other.

**Business Use:**

Helps Netflix assess if **high-rated movies** come with more consistent or riskier returns compared to **low-rated content**.

**Chi-Square Test – IMDb Rating Category vs. Revenue Category**

**Focus:**

Is there a relationship between **IMDb rating group** (Low/Medium/High) and **Revenue level** (Low/Medium/High)?

**Formula:**

**A math equation with black text

AI-generated content may be incorrect.**

**Key Insight:**

Checks if rating and revenue are **statistically associated** or independent.

**Business Use:**

Helps Netflix explore if **high-rated content** is **more likely to earn** high revenue — valuable for **early prediction models**.

**One-Way ANOVA – Revenue Across IMDb Rating Groups**

**Focus:**

Do movies with **different IMDb rating levels** (e.g., low: <5.5, medium: 5.5–7.0, high: >7.0) have **significantly different average revenue**?

**Formula:**

**A math equations and formulas

AI-generated content may be incorrect.**

**Key Insight:**

Tests if at least one **rating group** has a **different revenue mean** than others.

**Business Use:**

Supports Netflix in deciding whether **ratings can be a reliable predictor** of revenue, and which **rating band** performs best financially.

**6 Conclusion**

**Overall Conclusion**

In this project, we found that Netflix’s content success—measured by revenue—is mainly affected by three things: **how much money is spent to make it (budget)**, **how well people rate it (IMDb ratings)**, and **how long people watch it (watch time)**.

We first cleaned and prepared the data so that our results would be accurate. Then, by using different statistical methods, we discovered patterns and connections between the data. For example:

* Content with a higher budget usually earns more revenue.
* Shows and movies with better ratings are watched more and perform better.
* Dividing content into groups based on ratings helped us see how different types of content bring in different amounts of money.

These findings can help Netflix make smarter choices. They can:

* **Choose projects** that are more likely to succeed
* **Spend money wisely** on content that will give better returns
* **Create content** that matches what viewers enjoy and engage with