University of Washington - Tacoma

IMDB MOVIE DATA ANALYSIS PROJECT

By: Quinn Duong

Email: qduong@uw.edu

Instructor: Dan Heinig, MBA

Class: TBANLT 485

ABSTRACT

This project aims to predict movie revenue based on independent attributes from the IMDB Movies dataset. The data will be carefully prepared, including handling missing values and ensuring data consistency, to ensure accurate analysis. The project involves data preprocessing, exploratory data analysis, feature engineering, and machine learning techniques. The project will focus on understanding which independent attributes have the most significant impact on revenue prediction. Various machine learning algorithms, such as regression models and ensemble methods, will be trained and evaluated using appropriate evaluation metrics to determine the most effective model for revenue prediction. Through the predictive model, insights can be gained into the factors that contribute to a movie's success in terms of revenue generation. This information can be valuable for industry professionals, production companies, and investors, providing guidance for decision-making and resource allocation.

DATA DESCRIPTION REPORT

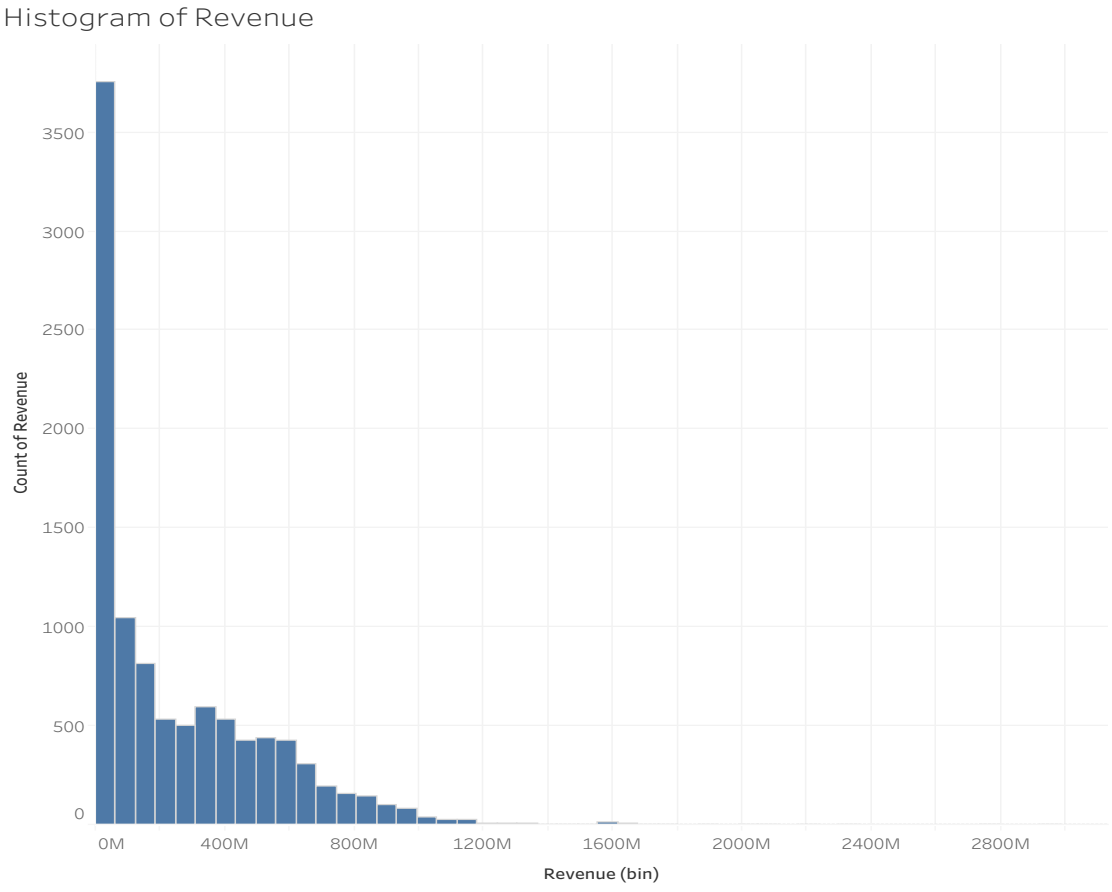
Dictionary

- Revenue: The amount of money generated by the movie at the box office or through other means of distribution.
- Names: The names of the movies.
- Date_x: The date associated with the movie, which could be the release date or another relevant date.
- Genre: The category or genre of the movie, such as action, comedy, drama, etc.
- Overview: A brief summary or description of the movie.
- Crew: The people involved in the production of the movie, including the director, writers, producers, and other crew members.
- Orig_title: The original title of the movie, which might be different from the translated or localized title.
- Status: The current status of the movie, which could indicate if it's released, in production, or some other stage.
- Orig_lang: The original language in which the movie was produced.
- Country: The country or countries associated with the movie's production.
- Score: The rating or score assigned to the movie by IMDB or other rating systems.
- Budget_x: The budget allocated for the production of the movie.

Univariate Properties

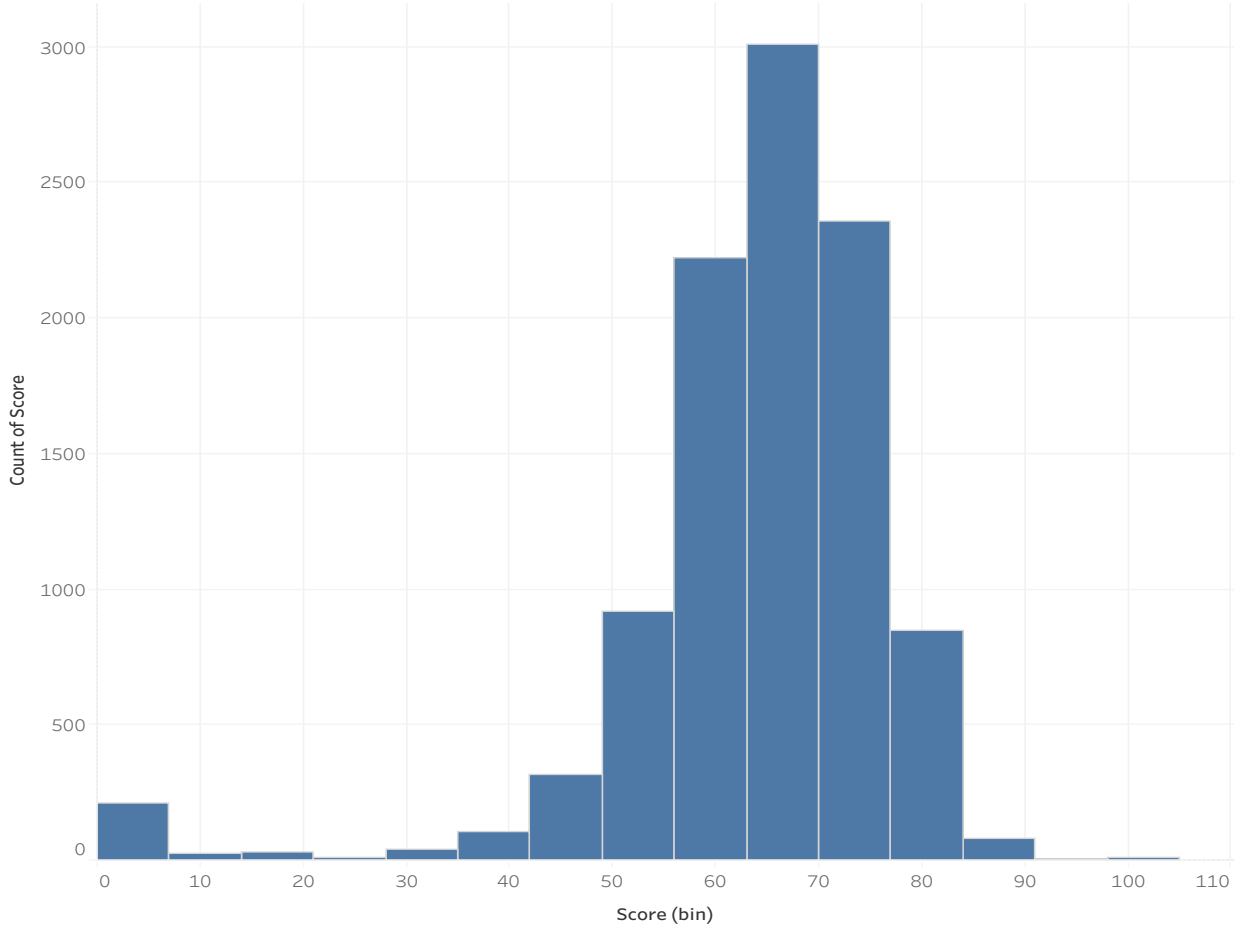
Feature	Var Type	Data Type	Count	Missing	%	Unique	Min	Q1	Med	Q3	Max	Mean	SD	Skew	Kurt
Revenue	Label	Numeric	10178	0	0	8227	0	28,588,985	152,934,877	417,802,077	2,923,706,026	253,140,093	277,788,049	1.54	4.08
Names	Feature	Categorical	10178	0	0	9656									
Date_x	Feature	Categorical	10178	0	0	5688									
Genre	Feature	Categorical	10178	85	0.0084	2303									
Overview	Feature	Categorical	10178	0	0	9905									
Crew	Feature	Categorical	10178	56	0.0055	9927									
Orig_title	Feature	Categorical	10178	0	0	9732									
Status	Feature	Categorical	10178	0	0	3									
Orig_lang	Feature	Categorical	10178	0	0	54									
Country	Feature	Categorical	10178	0	0	60									
Score	Feature	Numeric	10178	0	0	79	0	59	65	71	100	63.5	13.5	-2.39	8.79
Budget_x	Feature	Numeric	10178	0	0	2316	1	15,000,000	50,000,000	105,000,000	460,000,000	64,882,379	57,075,645	0.88	0.45

Histograms



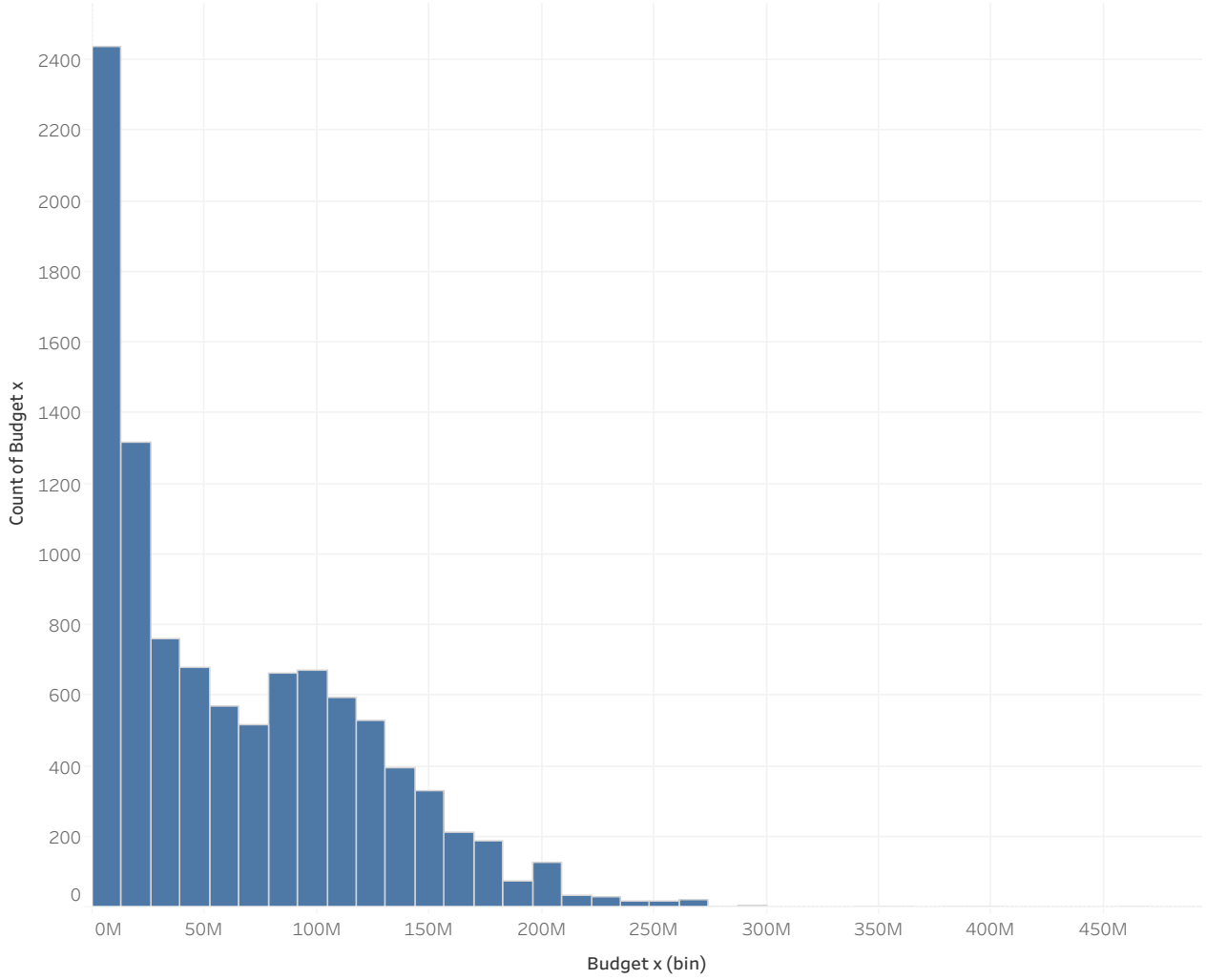
The trend of count of Revenue for Revenue (bin).

Histogram of Score



The trend of count of Score for Score (bin).

Histogram of Budget_x



The trend of count of Budget x for Budget x (bin).

DATA EXPLORATION REPORT

This report details the relationship between each potential feature and the label “Revenue”

SUMMARY TABLE

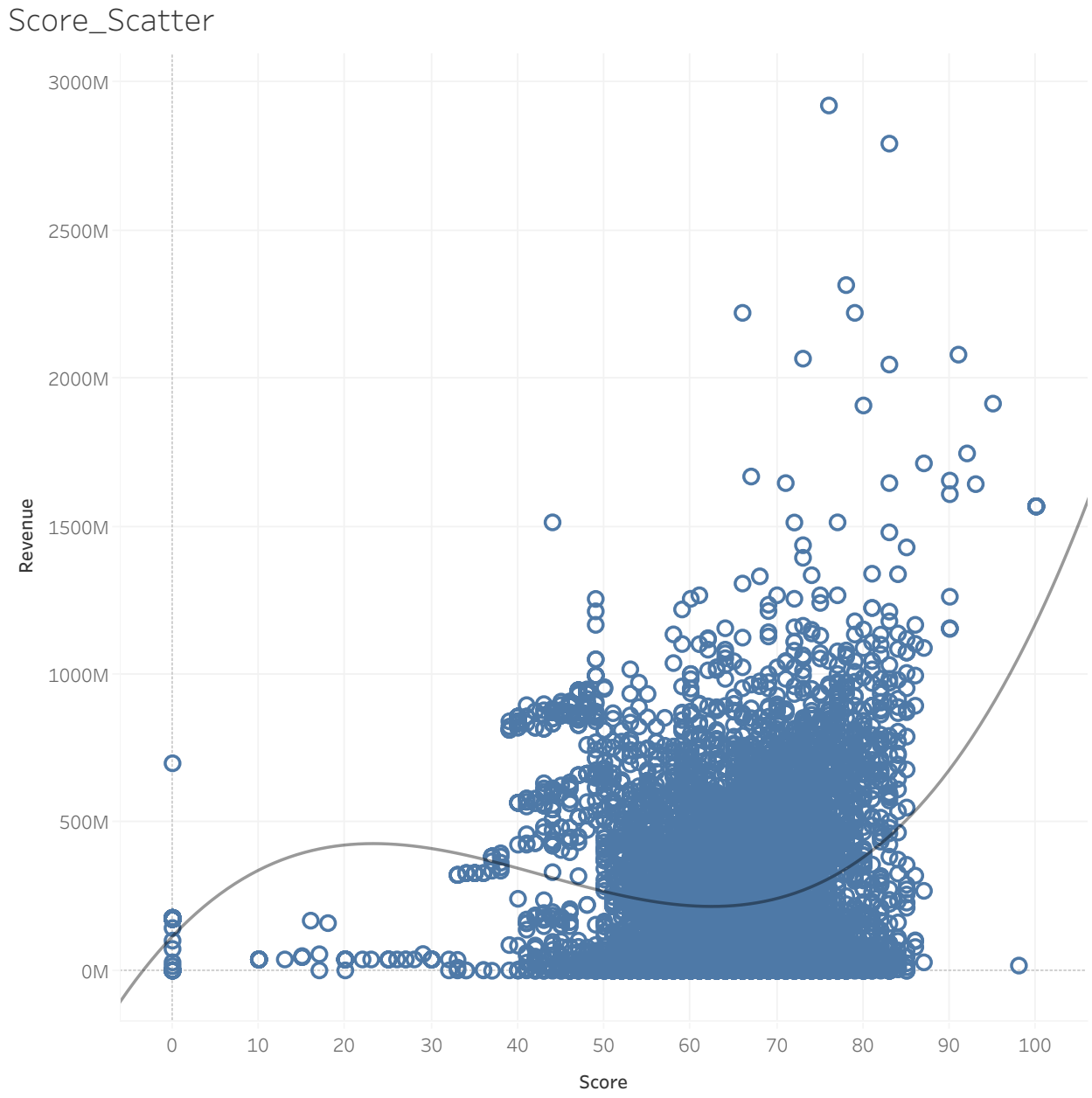
Feature	Analysis	Effect size	P-value
Genre	F-stat	56.2825183	0.09760088
Status	F-stat	4.60360861	0.0100365
Orig_lang	F-stat	1.036249	0.52531402
Country	F-stat	0.193352	0.893399
Score	R squared	0.0540439	<0.0001
Budget_x	R squared	0.48948	< 0.0001

CORRELATION MATRIX:

	Score	*Budget_x*	*Revenue*
Score	1		
Budget_x	-0.2354982	1	
Revenue	0.09649035	0.6738236	1

Score

H#: We expect that Score will have a positive effect on Revenue because movies with higher scores or ratings tend to attract more viewers or generate more interest, which can result in increased revenue.



Score vs. Revenue.

Regression Equation: Revenue = 7062.92*Score^3 + -909440*Score^2 + 3.09036e+07*Score + 1.11419e+08

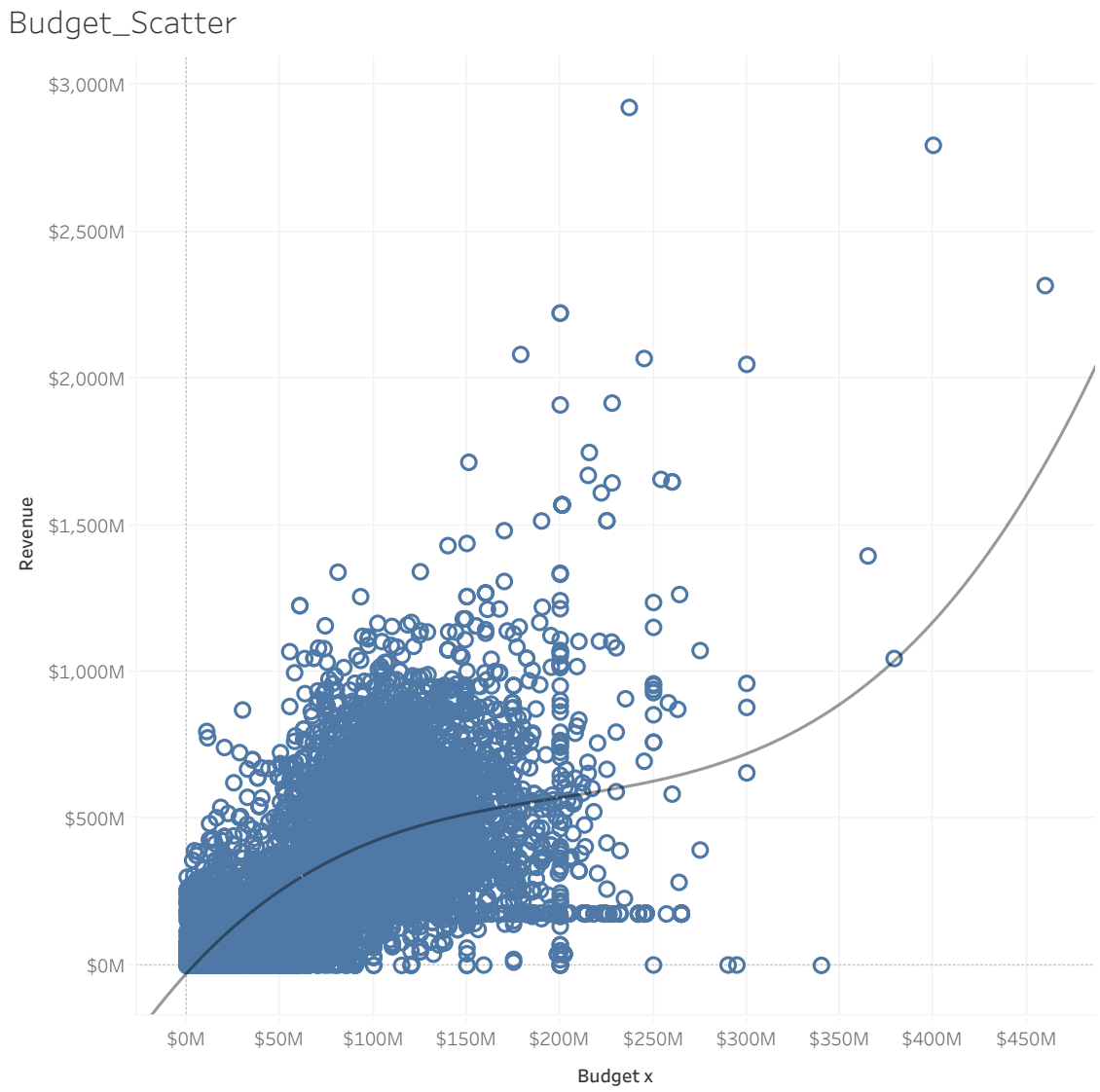
R^2: 0.0540439

p-value: <0.0001

Summary: There is a strong effect of Score on Revenue that is very reliable. We tested all non-linear transformations and found that Polynomial produced the best R^2 value of 0.0540439 and a p value of < 0.0001

Budget

H#: We expect that Budget will have a positive effect on Revenue because a higher budget allows for better production values including advanced special effects, high-quality cinematography, and immersive sound design, marketing efforts, and overall quality, which can attract more viewers and lead to increased revenue.



Budget x vs. Revenue.

Regression Equation: Revenue = 4.99415e-17*Budget x^3 + -2.99863e-08*Budget x^2 + 6.99551*Budget x + -3.0297e+07

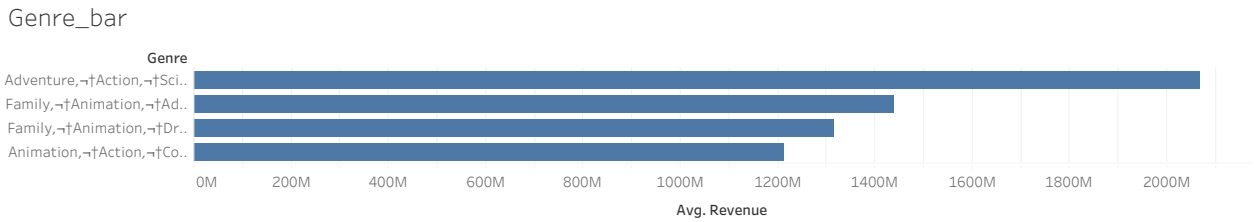
R2: 0.48948

p-value: <0.0001

Summary: There is a strong effect of Budget on Revenue that is very reliable. We tested all non-linear transformations and found that Polynomial produced the best R^2 value of 0.48948 and a p value of < 0.0001

Genre

H#: We expect that there will be an effect by Genre on Revenue because different genres have varying degrees of popularity and audience appeal, which can influence the revenue generated by a movie. For example, certain genres, such as action, adventure, comedy, and superhero films, have historically attracted large audiences and generated significant revenue.



Average of Revenue for each Genre. The view is filtered on Genre, which keeps Adventure,¬†Action,¬†Science Fiction,¬†Fantasy, Animation,¬†Action,¬†Comedy,¬†Mystery,¬†Crime,¬†Fantasy, Family,¬†Animation,¬†Adventure,¬†Comedy,¬†Fantasy and Family,¬†Animation,¬†Drama.

***Adventure, Action, Science Fiction, Fantasy* mean:** 2068223624

***Family, Animation, Adventure, Comedy, Fantasy* mean:** 1437862795

***Family, Animation, Drama* mean:** 169650000

***Animation, Action, Comedy, Mystery, Crime, Fantasy* mean:** 1213425727

One-way ANOVA *F*: 56.2825183

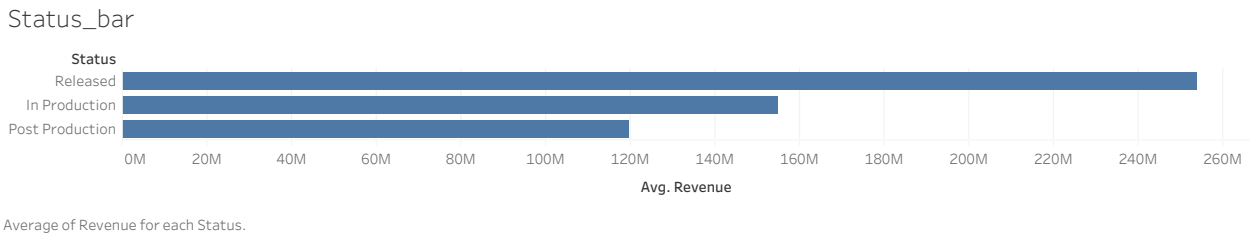
p-value: 0.09760088

	DF	**SS**
Between	**(k-1) = 3**	2.7566E+18
Within	**(n-k) = 1**	1.6326E+16
Total	**(dfb + dfw) = 4**	2.773E+18

Summary: There *is* an effect by *Genre* on *Revenue* with a f-stat value of 56.2825183 and a p value of 0.09760088 (<0.5)

Status

H#: We expect that there will be an effect by Status on Revenue because the status refers to the current stage of the movie, which could indicate if it's released, in production, or at some other stage of development, impacting its revenue potential. For example, Movies that are released during peak moviegoing seasons or strategically timed to coincide with holidays or other events tend to have better revenue prospects.



Average of Revenue for each Status.

***Released* mean:** 253703704

***In Production* mean:** 154868097

***Post Production* mean:** 119669447

One-way ANOVA *F*: 4.60360861

p-value: 0.0100365

	DF	**SS**
Between	**(k-1) = 2**	7.0998E+17
Within	**(n-k) =** 10175	7.8461E+20
Total	**(dfb + dfw)** = 10177	7.8532E+20

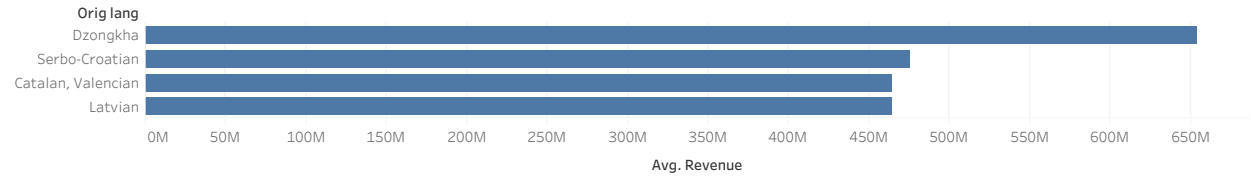
Summary: There *is* an effect by *Status* on *Revenue* with a f-stat value of 4.60360861 and a p value of 0.0100365 (<0.5)

Original language

H#: We expect that there will be an effect by Language on Revenue because movies produced in widely spoken languages, such as English, Mandarin, Hindi, or Spanish, have a larger potential audience and can attract more viewers, leading to higher revenue; or movies produced in a specific language may carry cultural significance or reflect the cultural context of a particular

region, attracting audiences from that region who are more likely to connect with the movie, increasing their interest and potential ticket sales.

Original Language



Average of Revenue for each Orig lang. The view is filtered on Orig lang, which keeps Catalan, Valencian, Dzongkha, Latvian and Serbo-Croatian.

***Dzongkha* mean:** 653988016.4

***Serbo-Croatian* mean:** 475701740.6

***Catalan, Valencian* mean:** 464231672.3

***Latvian* mean:** 463972282.2

One-way ANOVA *F*: 1.0362487

p-value: 0.52531402

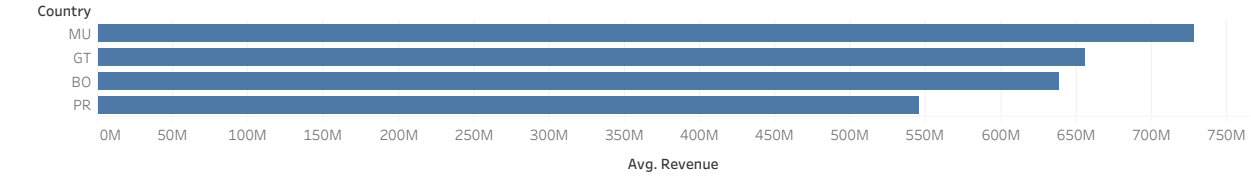
	DF	**SS**
Between	**(k-1) = 3**	2.9425E+16
Within	**(n-k) = 2**	1.8931E+16
Total	**(dfb + dfw) = 5**	4.8356E+16

Summary: There *is* no effect by *Language* on Revenue with a f-stat value of 1.0362487 and a p value of 0.52531402 since we reject the null when the p-value is higher than 0.5

Country

H#: We expect that there will be an effect by Country on Revenue because

Country_bar



Average of Revenue for each Country. The view is filtered on Country, which keeps BO, GT, MU and PR.

***MU* mean: 728608266**

***GT* mean: 655664751.8**

***BO* mean: 638332462.6**

***PR* mean: 545316307.9**

One-way ANOVA *F*: 0.193352043

p-value: 0.893398995

	DF	SS
Between	**(k-1) = 3**	2.54325E+16
Within	**(n-k) = 2**	8.769E+16
Total	**(dfb + dfw) = 5**	1.1312E+17

Summary: There *is no* effect by *Country* on Revenue with a f-stat value of 0.193352043 and a p value of 0.893398995 since we reject the null when the p-value is higher than 0.5.

MODEL TESTING

Regression Models

Algorithm	Feature Scoring Method	List of actual n features selected	R squared for the model	RMSE for the model
Linear Regression	Permutation Feature Importance	Budget_x, Score, Genre, Orig_lang, Country, Status	0.462166	199741517.7532
Linear Regression	Permutation Feature Importance	Budget_x, Score, Genre, Orig_lang, Country	0.454759	201112209.2440
Linear Regression	Permutation Feature Importance	Budget_x, Score, Genre, Orig_lang	0.457208	200659999.9526
Bayesian Linear Regression	Permutation Feature Importance	Budget_x, Score, Country, Genre, Orig_lang, Status	0.501381	192321776.9978
Bayesian Linear Regression	Permutation Feature Importance	Budget_x, Score, Country,Gerne, Orig_lang	0.49426	193690269.1353
Bayesian Linear Regression	Permutation Feature Importance	Budget_x, Score, Country, Gerne	0.493663	193804535.2
Neural Network Regression	Permutation Feature Importance	Budget_x, Score, Country, Genre, Status, Orig_lang	-0.314941	312318286.9
Neural Network Regression	Permutation Feature Importance	Budget_x, Score, Country, Status Orig_lang	-0.156993	299408729.8
Neural Network Regression	Permutation Feature Importance	Budget_x, Score, Status, Orig_lang	-0.141126	297618042.2
Decision Forest Regression	Permutation Feature Importance	Budget_x, Country, Score, Genre, Orig_lang, Status	0.511011	190455715.6
Decision Forest Regression	Permutation Feature Importance	Budget_x, Country, Score, Genre	0.527407	187235394.4
Decision Forest Regression	Permutation Feature Importance	Budget_x, Country, Score	0.665959	161024723.4
Decision Forest Regression with Tuner	Permutation Feature Importance	Budget_x, Country, Score	0.659922	162473094.4

Text Analytics Models

Algorithm	List of actual n features selected	N-gram size	Weighting function	Feature scoring method	Number of desired features	R2	RMSE
Linear Regression	Budget_x, Score, Genre, Orig_lang, Country, Status, Crew, Orig_title, Overview, Names, Date_x, Crew	1	Binary Weight	Fisher Score	20	0.525545	192246515.09040
Linear Regression	Budget_x, Score, Genre, Orig_lang, Country, Status, Crew, Orig_title, Overview, Names, Date_x, Crew	2	Binary Weight	Fisher Score	20	0.525502	192255296.41317
Linear Regression	Budget_x, Score, Genre, Orig_lang, Country, Status, Crew, Orig_title, Overview, Names, Date_x, Crew	4	Binary Weight	Fisher Score	20	0.52559	192237312.82006
Linear Regression	Budget_x, Score, Genre, Orig_lang, Country, Status, Crew, Orig_title, Overview, Names, Date_x, Crew	4	TF-IDF Weight	Fisher Score	20	0.52559	192237312.81991
Linear Regression	Budget_x, Score, Genre, Orig_lang, Country, Status, Crew, Orig_title, Overview, Names, Date_x, Crew	4	Graph Weight	Fisher Score	20	0.525203	192315779.89714
Linear Regression	Budget_x, Score, Genre, Orig_lang, Country, Status, Crew, Orig_title, Overview, Names, Date_x, Crew	4	Binary Weight	Fisher Score	40	0.525659	192223479.446459
Linear Regression	Budget_x, Score, Genre, Orig_lang, Country, Status, Crew, Orig_title, Overview, Names, Date_x, Crew	4	TF-IDF Weight	Fisher Score	40	0.525659	192223479.446306

FINDINGS

Feature Selection Method

PFI												
Budget_x	1.05536											
Score	0.108317											
Country	0.020903											
Genre	0.010982											
Status	0.007065											
Orig_lang	0.00029											
Date_x	-0.000195											
Names	-0.000695											
Orig_title	-0.002953											
Crew	-0.004224											
Overview	-0.004714											
Variable											R squared	RMSE
Budget_x	Score	Country	Genre	Status	Orig_lang	Date_x	Names	Orig_title	Crew	Overview	0.530254	191290159
Budget_x	Score	Country	Genre	Status	Orig_lang						0.462166	199741518
FBFS												
Feature Scoring Method	**Number of Desired Features**	**List of actual n features selected**	**R squared for the model**	**RMSE for the model**								
Pearson Correlation	4	Budget_x, Date_x, Score	0.511673	190326666								
Pearson Correlation	2	Budget_x, Date_x	0.455971	200888548.8								
Mutual Information	10	Budget_x, Country, Genre, Date_x, Score, Orig_lang, Names, Orig_title, Overview, Status	0.511673	190326665.8								
Mutual Information	6	Budget_x, Country, Genre, Date_x, Score, Orig_lang	0.454947	201077648.1								
Chi Squared	10	Overview, Orig_title, Names, Genre, Budget_x, Country, Orig_lang, Date_x, Score, Status	0.511673	190326665.8								
Chi Squared	6	Overview, Orig_title, Names, Genre, Budget_x, Country	0.445983	202724225.9								

For regression models, to identify which variables has the most impact on dependent variable (revenue), I conducted model testing comparing two feature selection methods: FBFS and PFI.

Overall, the results indicate that the Filter Based Feature Selection (FBFS) method achieved an R-squared value of 0.511673 and an RMSE of 190326665.8, while Permutation Feature Importance (PFI) method resulted in an R-squared value of 0.530254 and an RMSE of 191290158.9. In this scenario, it can be argued that PFI is better for feature selection since its ability to select features that result in smaller prediction errors suggests that it captured important and relevant information for predicting revenue.

With the PFI method, I filtered out independent variables with negative scores (Date_x, Names, Orig_title, Crew, Overview) after running the first experiment including all the variables except for Crew/Cash variable (I will analyze it to predict Revenue in the text analytic model), since by filtering out the independent variables with negative scores, we can reduce the complexity of the model and focusing on the features that have a more meaningful impact on predicting revenue. Also, running the regression models including all features is time consuming, which is not efficient for regression model testing. Although, the r-squared is lower than the first model's, the second model has a much lower overfitting.

For text analytics models, while analyze Crew to predict Revenue, I also include all the features to avoid bias since I don't want to exclude any attributes that may have an important impact on the dependent variable. Moreover, it would be easier and less time consuming to process this big amount of data.

Regression models

- Among the models evaluated, the Decision Forest Regression model produced the best results for the regression analysis. It achieved the highest R-squared values in all trials among those models, indicating that it explained a substantial amount of the variance in the dependent variable (revenue). Additionally, it had the lowest RMSE, suggesting that it made more accurate predictions with smaller errors compared to other models. Followed by, the Bayesian Linear Regression model performed slightly better than the Linear Regression model, achieving a higher R-squared value of 0.501381 and a slightly lower RMSE of 192321776.9978. However, it did not outperform the Decision Forest Regression model.
- Based on the results of the Decision Forest Regression model, we can conclude that Budget, Country and Score has the largest impact on Revenue with a highest R-squared of 0.665959 and a lowest RMSE of 161024723.4. Besides these three attributes, Genre and Original language also contribute to predicting Revenue.
- Using Tuner did not improve the model's performance.

Text Analytics – Extract N-gram size

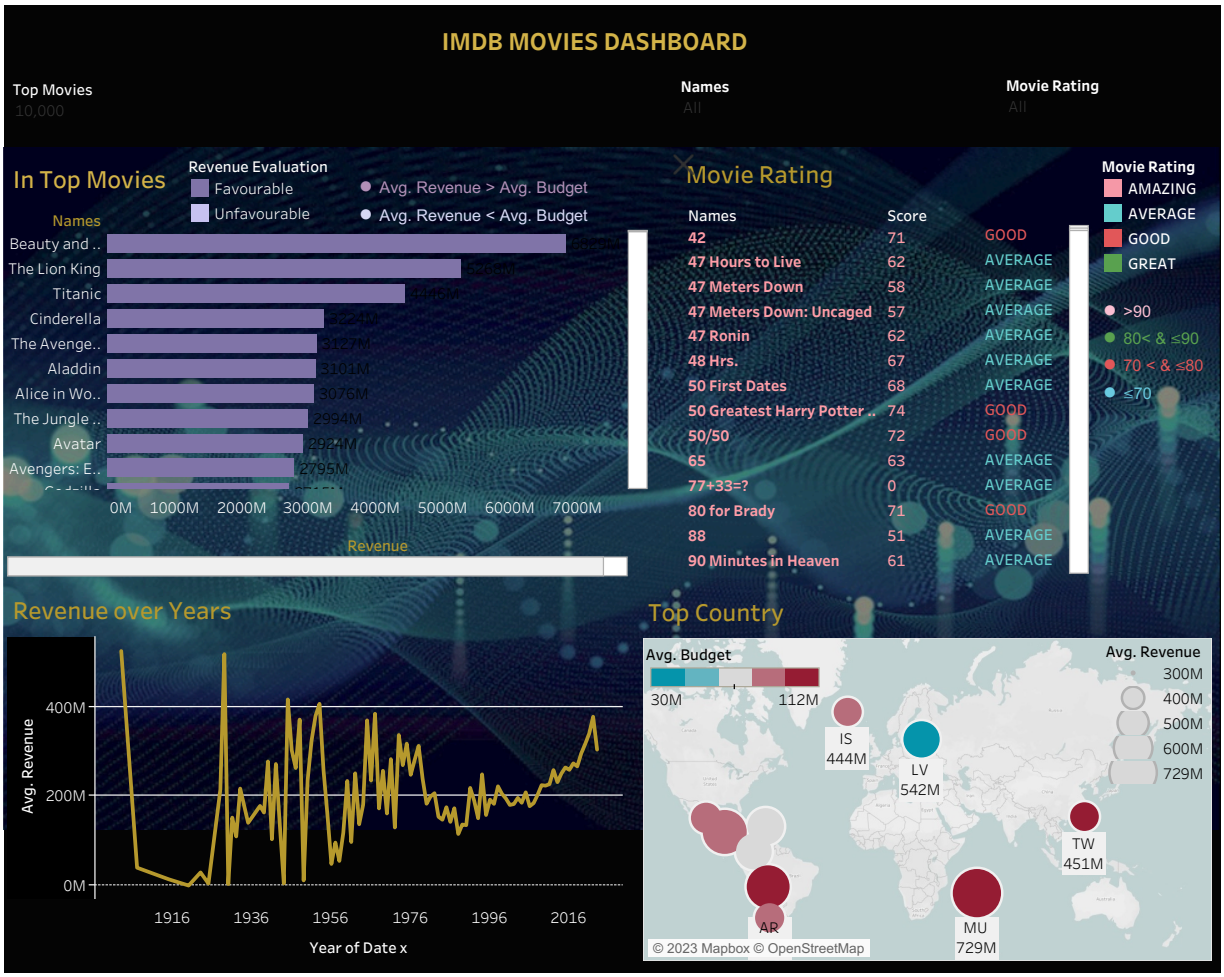
- During the process of analyzing text (Crew/Cash) to predict Revenue, I attempted to test the text analytics model including all features with 3 different n-gram sizes (1,2, 4), 3

different weighting functions (Binary Weight, TF-IDF Weight, Graph Weight), and 2 different numbers of desired features (20, 40) so that I could determine which combination of parameters gives the best outcome.

- Based on the results, Combination 6 and 7 give similar results (N-gram size = 4, Weighting function = TF-IDF Weight or Binary Weight, Number of desired features = 40) produced the best text analysis results. They both achieved the highest R-squared value of 0.525659, indicating that they explained a slightly higher amount of variance in the dependent variable compared to the other combinations. They also had the slightly lowest RMSE of 192223479.446306, suggesting that it made slightly more accurate predictions.
- Some additional insights from the testing conducted include:
  - The choice of N-gram size did not have a significant impact on the model's performance. Combinations with N-gram sizes of 1, 2, and 4 produced similar R-squared values and RMSE.
  - Weighting functions such as Binary Weight and TF-IDF Weight resulted in similar performance, with slight variations in R-squared and RMSE values.
  - Increasing the number of desired features from 20 to 40 did not lead to substantial improvements in the model's performance. The R-squared values slightly decreased, while the RMSE values increased.

Key Insights

Tableau Dashboard:



- Based on the dashboard, here are some key insights that I achieved:

- Avatar has the highest average revenue and probably gained a significant profit since its revenue is much higher than its initial budget. Followed by, Avengers: Endgame, Avatar: The Way of Water, and Titanic also have high revenues.
- I tried to evaluate each movie based off its score and categorized into groups of amazing (>90), great (80<&<=90), good (70< &<=80), average (>70), and then used filter "wildcard match" that can help stakeholders to search for a specific movie's rating. In this section, it also shows revenue of each movie so that viewers can identify the relationship between score and revenue.
- Next, I used a line graph to demonstrates the growing trend of revenue over 100 years. It shows that in 1903, we have the highest average revenue of $525M, which is very an interesting fact since technology was very poor during that time.
- In this dashboard, we can also identify top 10 countries that have the highest average revenues, which are Iceland($444M), Latvia($542M), Taiwan($451M), Mauritius ($729M), …

Data Limitation & Additional data: There are missing values (crew and genre attribute), outliers, or inconsistent data entries, impacting the validity of the results. Using regression models has limited my feature selection, which might cause bias and unfair conclusion. The dataset includes a limited number of variables, which may not capture all the relevant factors influencing revenue. Additional variables such as marketing budget, production cost, audience demographics, or competition could provide valuable insights into revenue prediction.

Additional questions that could be addressed with this data:

1. How does the budget of a movie influence its revenue? Is there a correlation between higher budgets and higher revenues?
2. What is the impact of the original language and country of origin on revenue? Are certain languages or countries associated with higher revenue?
3. How does the score or rating of a movie affect its revenue? Is there a relationship between critical acclaim and financial success?

RECOMMENDATIONS

- Increase the budget: The analysis indicated that budget has a significant impact on revenue. Allocating a higher budget for movie production can potentially lead to increased production quality, better marketing campaigns, and wider distribution, all of which can contribute to higher revenue.
- Focus on countries with high revenue potential: The analysis also identified that the country of origin has an impact on revenue. By identifying countries where the movies tend to perform well financially, production companies can prioritize targeting those markets with tailored marketing strategies and distribution efforts.
- Improve movie scores: The analysis revealed a positive relationship between movie scores and revenue. Investing in high-quality scripts, talented directors, and skilled production crews can help improve the overall quality and reception of the movies, leading to better scores and increased revenue.
- Collaborate with successful crew members and talents: Based on the analysis, crew members play a significant role in revenue generation. Building relationships and collaborations with experienced and successful crew members, directors, writers, and actors can enhance the overall quality and marketability of the movies, leading to increased revenue.

Appendices:

Feature Selection:

Feature Selection Method – PFI:
https://gallery.cortanaintelligence.com/Experiment/Feature-Selection-Method-PFI

Feature Selection Method – FBFS:

https://gallery.cortanaintelligence.com/Experiment/Feature-Selection-Method-FBFS

Model Testing:

Regression Models:

https://gallery.cortanaintelligence.com/Experiment/Model-Testing-PFI

Text Analytics Models:

https://gallery.cortanaintelligence.com/Experiment/Text-Analytics-67