Data Benders

2023

IMDb Movie Rating Prediction

Team

Halil Kolatan

Muhammed Maral

Table of contents

Introduction	3
Data Collection	4
Exploratory Data Analysis	5
Data Cleaning	6
Feature Engineering	7
Train-Validation-Test Split	9
Model Development	10
Model Evaluation	11
Conclusion	12
Future Work	13

Introduction

Objectives:

1. Unleashing the Power of Data
2. Understanding the importance of movie Ratings
3. Motivation towards predicting IMDb movie Ratings

Unleashing the **Power** of Data

$42.5B

Global Revenue

200

Update Per min

7.5M

Titles in Database

Importance of **Movie Ratings**

10,000

Movies Analyzed

%95

Decision Making

%20 ↑

Box-office Revenue

Motivation towards **Rating Prediction**

Audience Preferences

Accurate Planning

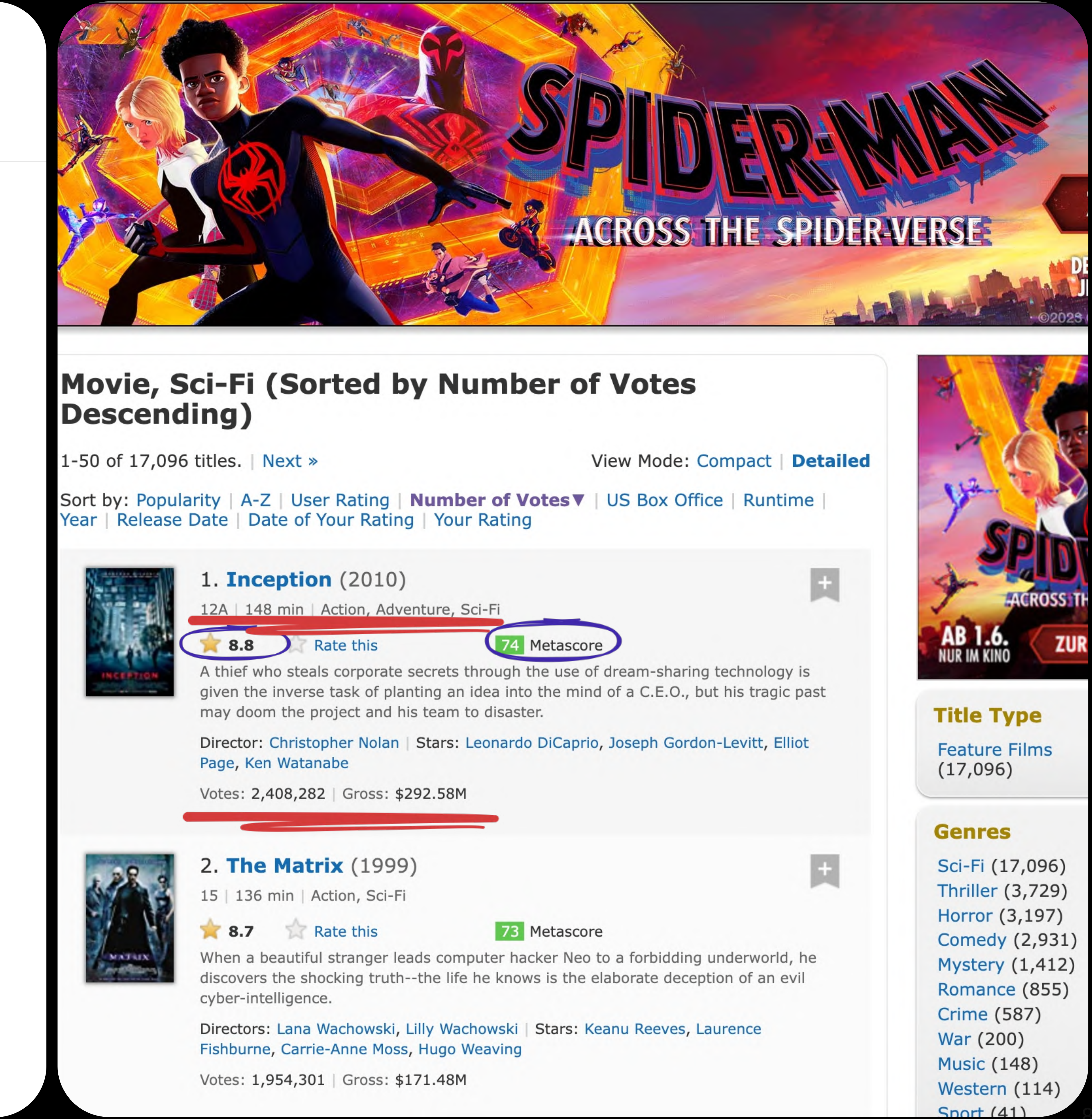
Improving the UX

Data Collection

Tools Used:

Scraped Data:

- Title
- Publish Year
- Runtime
- Genre
- IMDb Rating
- Metascore
- Movie Votes
- Budget
- Gross Revenue



Movie, Sci-Fi (Sorted by Number of Votes Descending)

1-50 of 17,096 titles. | Next »

View Mode: Compact | **Detailed**

Sort by: Popularity | A-Z | User Rating | **Number of Votes▼** | US Box Office | Runtime | Year | Release Date | Date of Your Rating | Your Rating

1. **Inception** (2010)

12A | 148 min | Action, Adventure, Sci-Fi

★ 8.8 ☆ Rate this 74 Metascore

A thief who steals corporate secrets through the use of dream-sharing technology is given the inverse task of planting an idea into the mind of a C.E.O., but his tragic past may doom the project and his team to disaster.

Director: Christopher Nolan | Stars: Leonardo DiCaprio, Joseph Gordon-Levitt, Elliot Page, Ken Watanabe

Votes: 2,408,282 | Gross: $292.58M

2. **The Matrix** (1999)

15 | 136 min | Action, Sci-Fi

★ 8.7 ☆ Rate this 73 Metascore

When a beautiful stranger leads computer hacker Neo to a forbidding underworld, he discovers the shocking truth--the life he knows is the elaborate deception of an evil cyber-intelligence.

Directors: Lana Wachowski, Lilly Wachowski | Stars: Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving

Votes: 1,954,301 | Gross: $171.48M

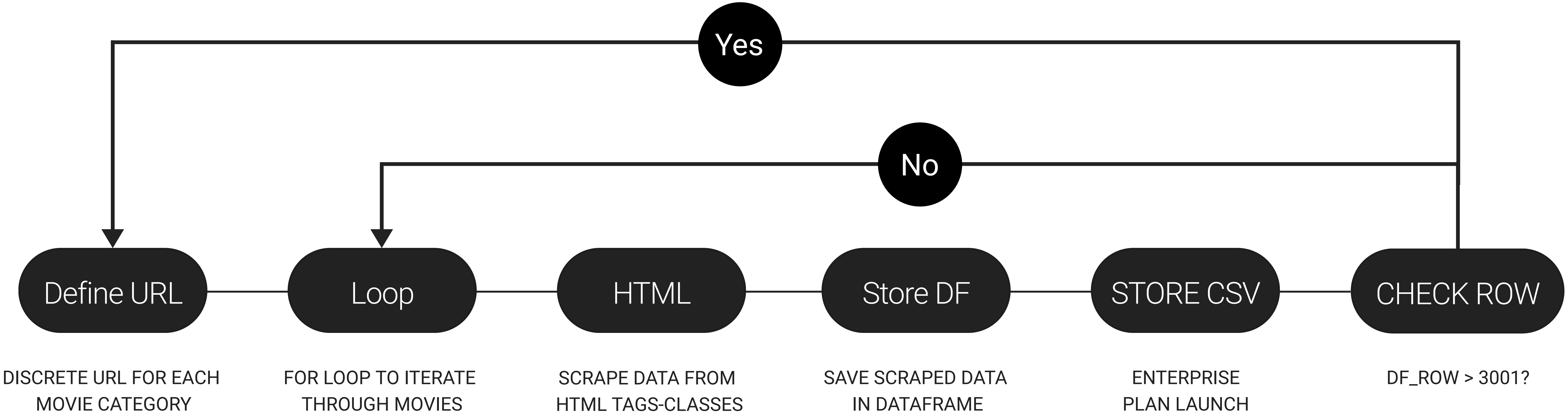
Title Type

Feature Films (17,096)

Genres

Sci-Fi	(17,096)
Thriller	(3,729)
Horror	(3,197)
Comedy	(2,931)
Mystery	(1,412)
Romance	(855)
Crime	(587)
War	(200)
Music	(148)
Western	(114)
Sport	(41)

Web Scraping Flowchart



Yes

No

Define URL — DISCRETE URL FOR EACH MOVIE CATEGORY

Loop — FOR LOOP TO ITERATE THROUGH MOVIES

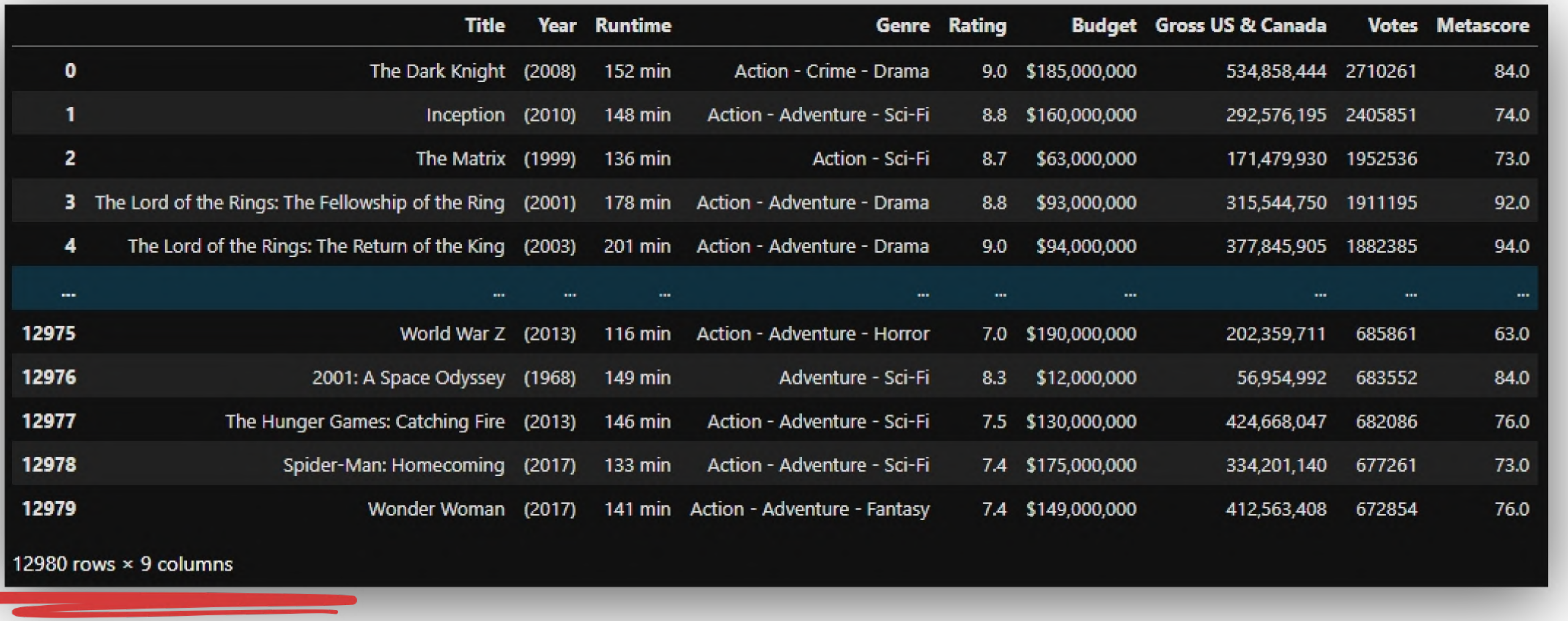
HTML — SCRAPE DATA FROM HTML TAGS-CLASSES

Store DF — SAVE SCRAPED DATA IN DATAFRAME

STORE CSV — ENTERPRISE PLAN LAUNCH

CHECK ROW — DF_ROW > 3001?

Web Scraping DataFrame



	Title	Year	Runtime	Genre	Rating	Budget	Gross US & Canada	Votes	Metascore
0	The Dark Knight	(2008)	152 min	Action - Crime - Drama	9.0	$185,000,000	534,858,444	2710261	84.0
1	Inception	(2010)	148 min	Action - Adventure - Sci-Fi	8.8	$160,000,000	292,576,195	2405851	74.0
2	The Matrix	(1999)	136 min	Action - Sci-Fi	8.7	$63,000,000	171,479,930	1952536	73.0
3	The Lord of the Rings: The Fellowship of the Ring	(2001)	178 min	Action - Adventure - Drama	8.8	$93,000,000	315,544,750	1911195	92.0
4	The Lord of the Rings: The Return of the King	(2003)	201 min	Action - Adventure - Drama	9.0	$94,000,000	377,845,905	1882385	94.0
...	...	...	...	...	...	...	...	...	...
12975	World War Z	(2013)	116 min	Action - Adventure - Horror	7.0	$190,000,000	202,359,711	685861	63.0
12976	2001: A Space Odyssey	(1968)	149 min	Adventure - Sci-Fi	8.3	$12,000,000	56,954,992	683552	84.0
12977	The Hunger Games: Catching Fire	(2013)	146 min	Action - Adventure - Sci-Fi	7.5	$130,000,000	424,668,047	682086	76.0
12978	Spider-Man: Homecoming	(2017)	133 min	Action - Adventure - Sci-Fi	7.4	$175,000,000	334,201,140	677261	73.0
12979	Wonder Woman	(2017)	141 min	Action - Adventure - Fantasy	7.4	$149,000,000	412,563,408	672854	76.0

12980 rows × 9 columns

Data Cleaning

Removed Rows

Rows with no US Revenue

Rows with no Budget

Duplicated Movies

Rows with no Metascore

New Movies

New Dataframe Rows

12980 → 4852 rows

Edited Columns

Year

(2018) object → 2018 int64

Runtime

158 min object → 158 int64

Budget

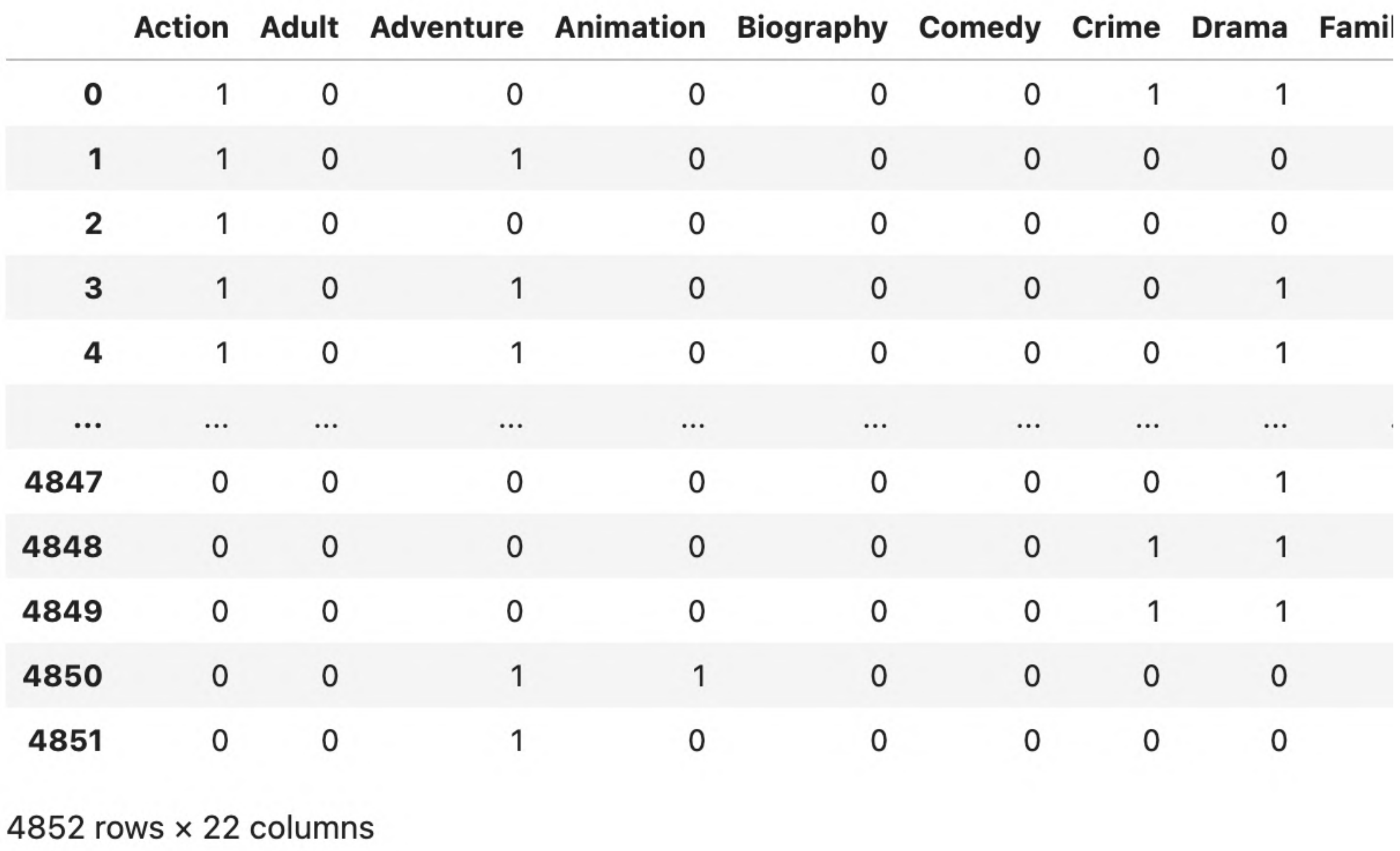
$185,000,000 → 185000000 fl64

Gross US Canada

534,858,444 → 534858444 fl64

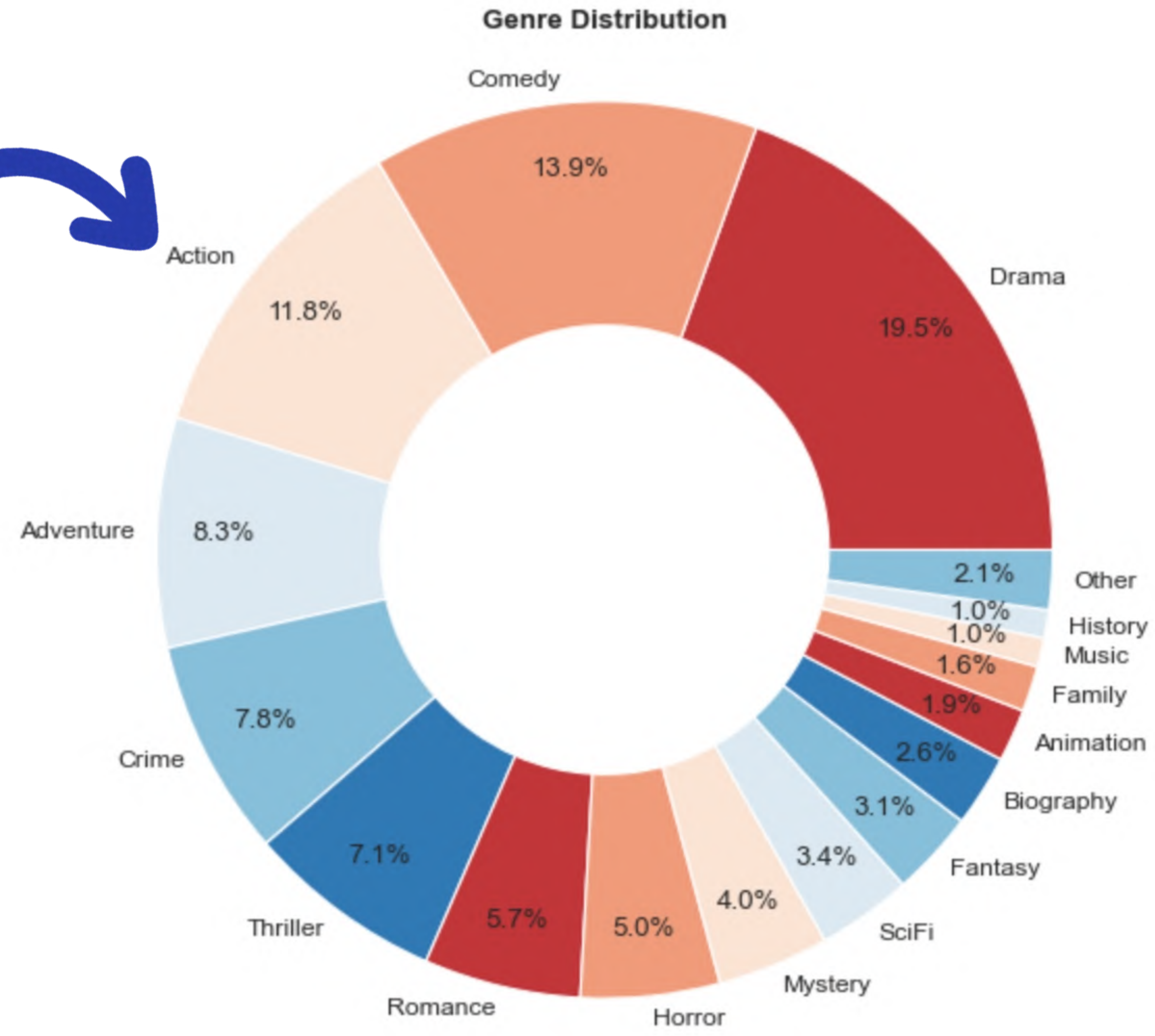
Exploratory Data Analysis

Applying Dummy Variable to the Movie Genre's



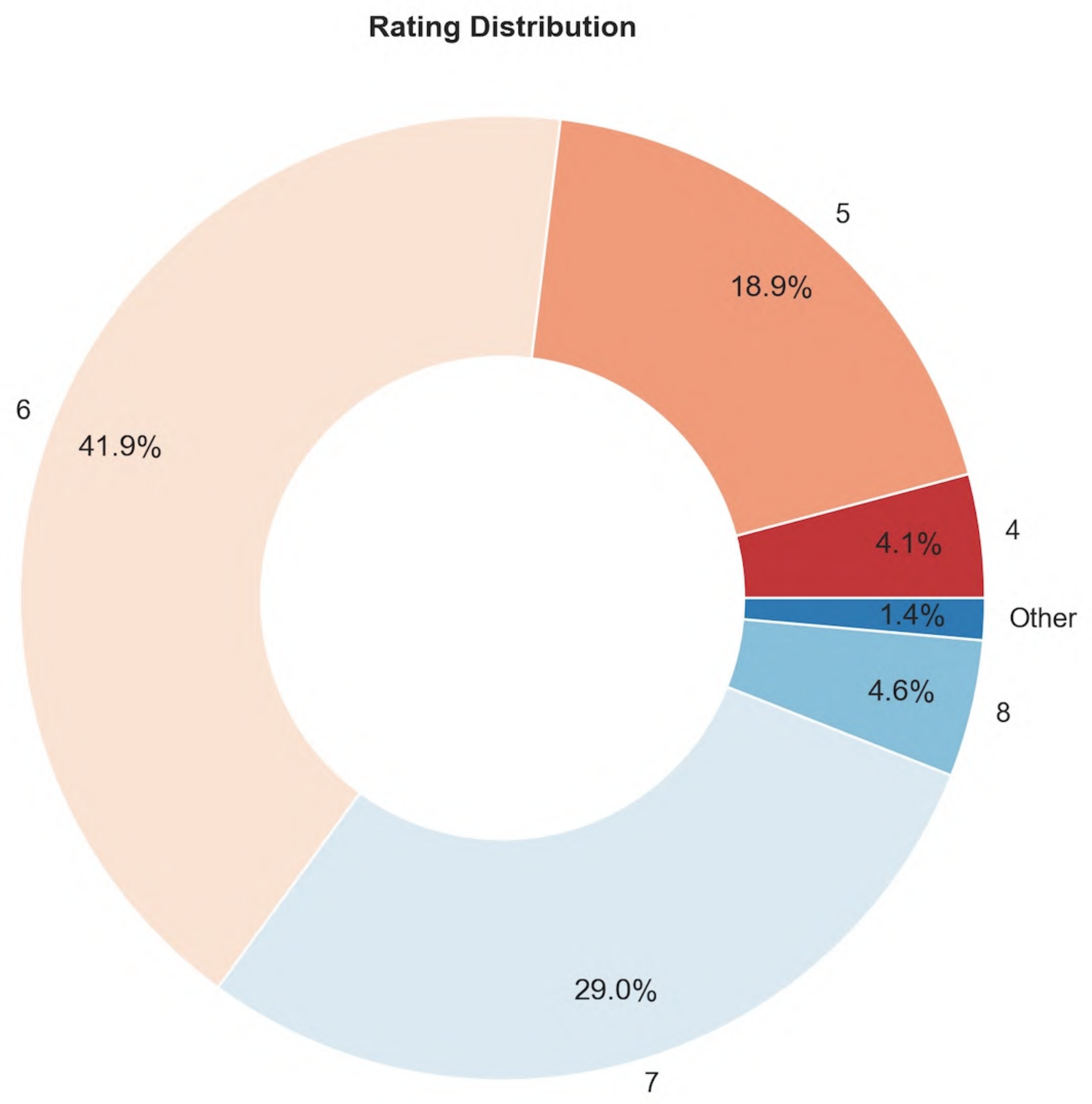
	Action	Adult	Adventure	Animation	Biography	Comedy	Crime	Drama	Famil
0	1	0	0	0	0	0	1	1	
1	1	0	1	0	0	0	0	0	
2	1	0	0	0	0	0	0	0	
3	1	0	1	0	0	0	0	1	
4	1	0	1	0	0	0	0	1	
...	...	...	...	...	...	...	...	...	
4847	0	0	0	0	0	0	0	1	
4848	0	0	0	0	0	0	1	1	
4849	0	0	0	0	0	0	1	1	
4850	0	0	1	1	0	0	0	0	
4851	0	0	1	0	0	0	0	0	

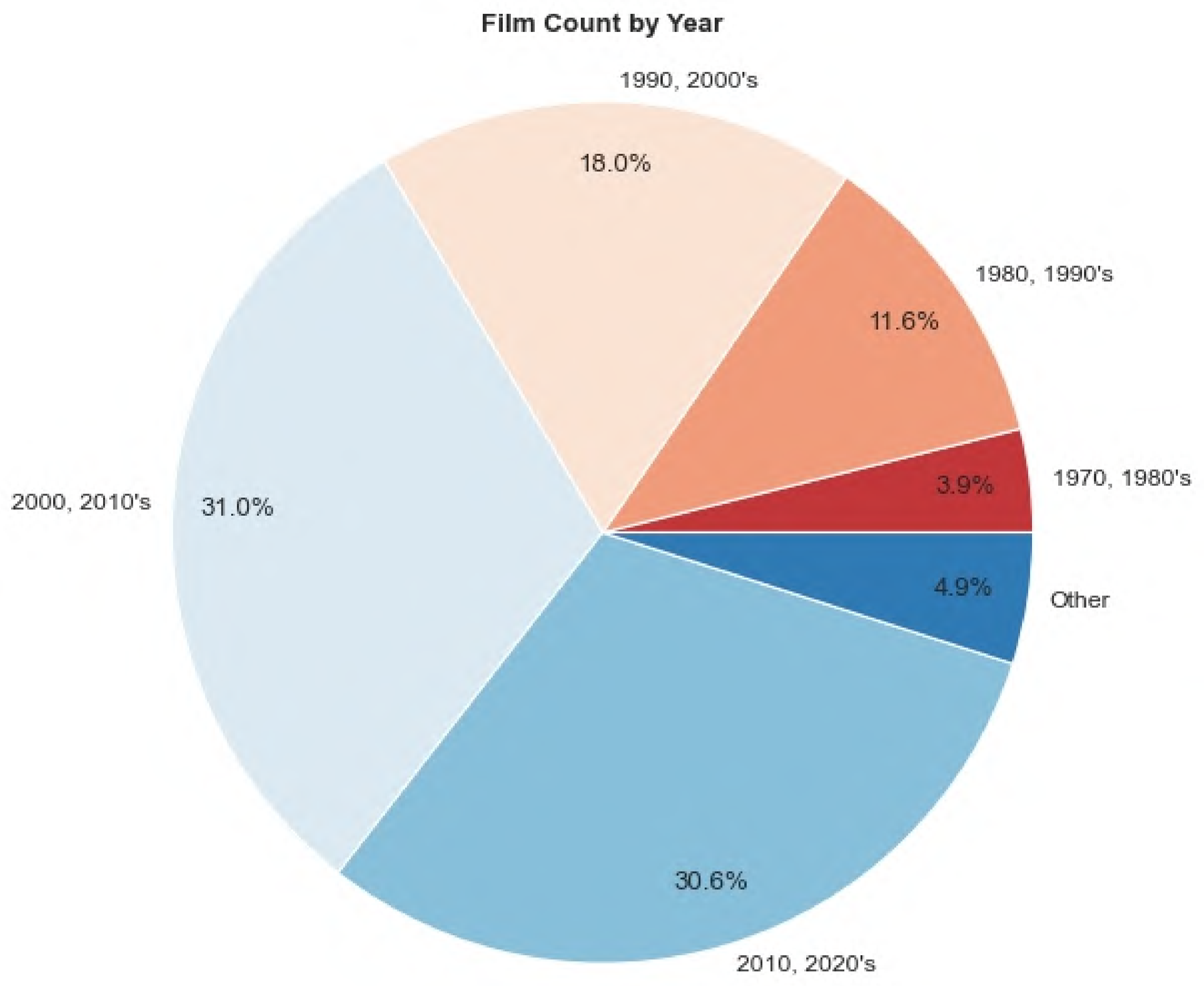
4852 rows × 22 columns



Genre Distribution

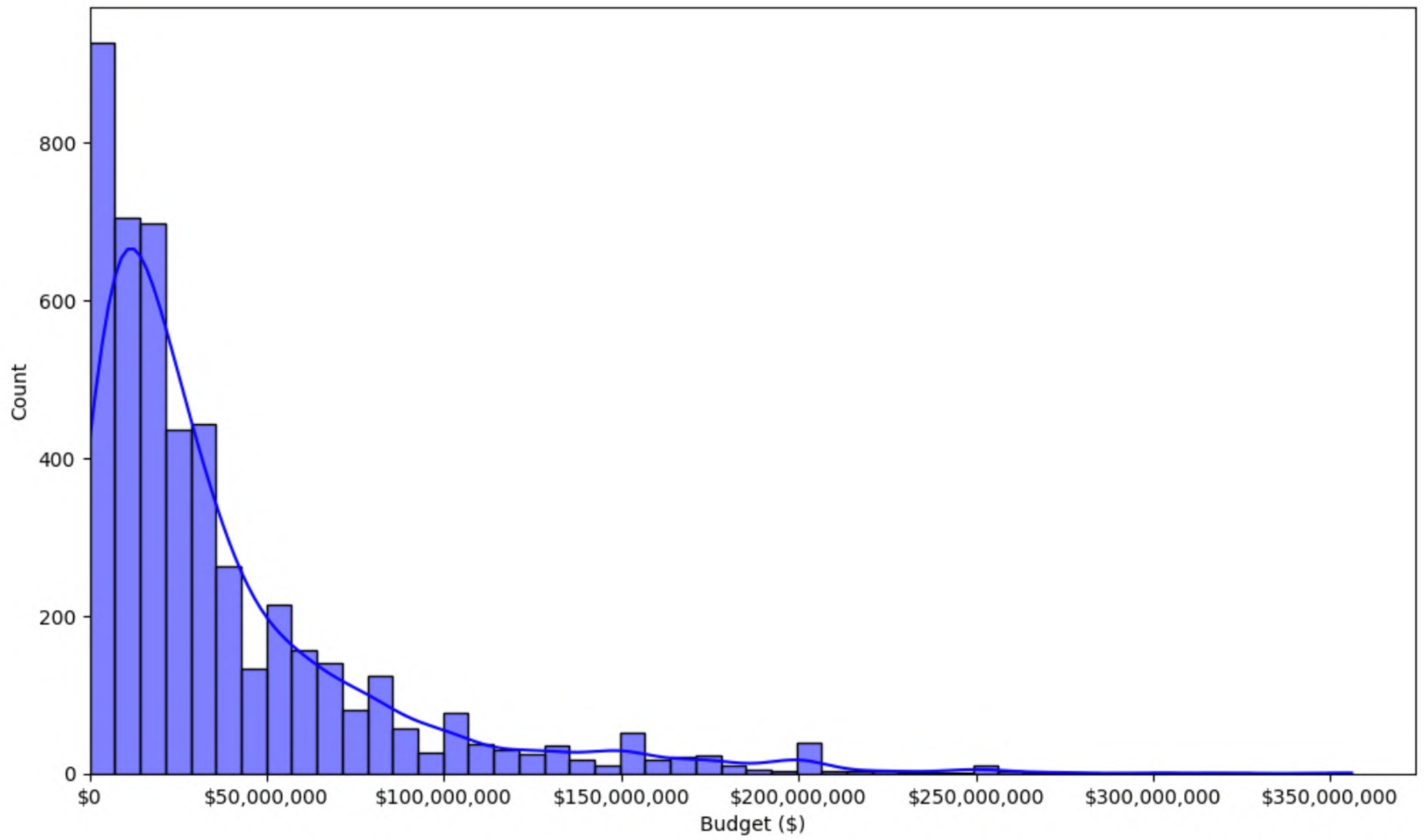
Visualizations

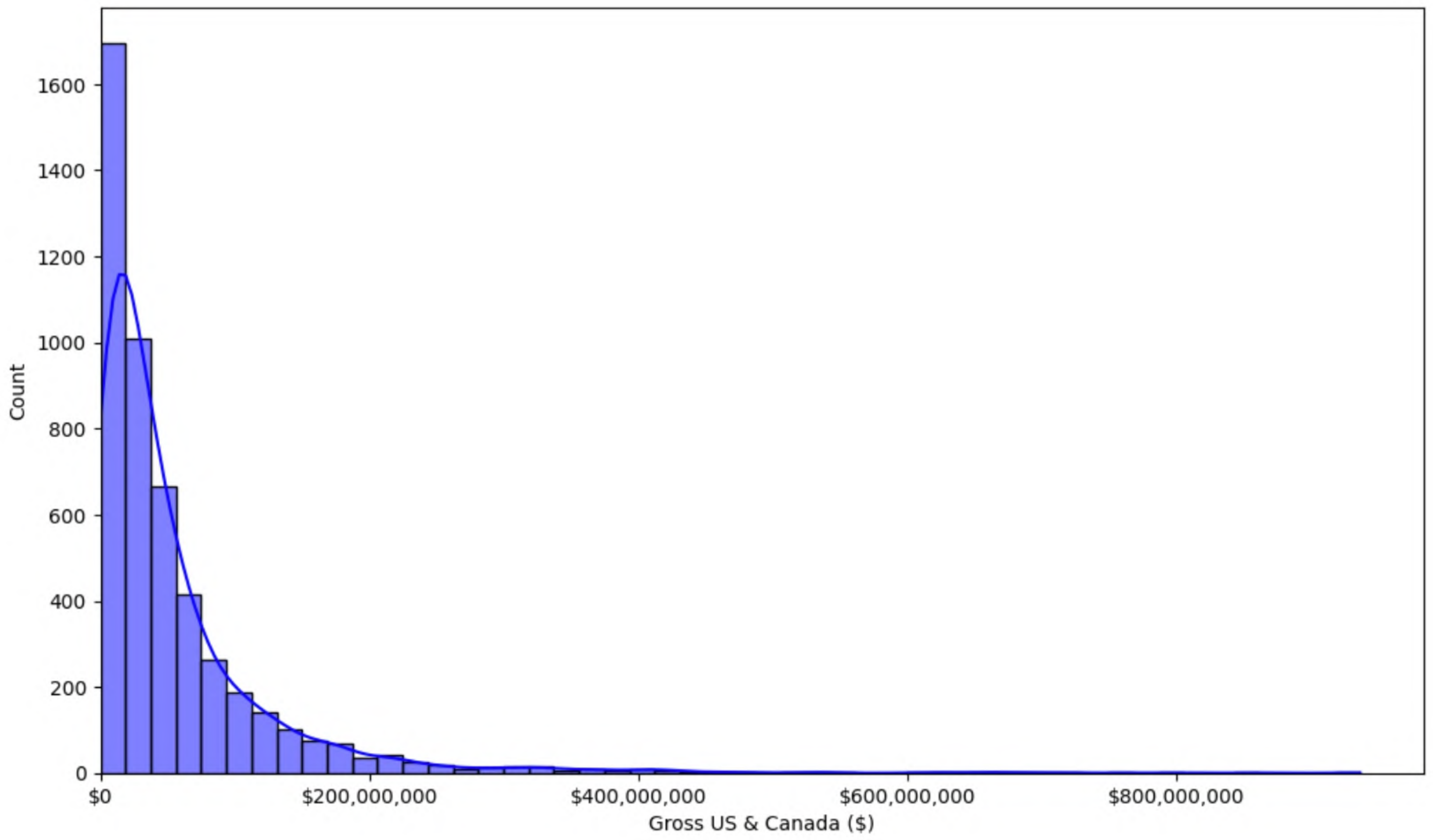




Try Pitch

Visualizations





Feature Engineering

Gross US&Canada %30 - %40

Gross Worldwide %60 - %70

New Column:

Estimated WorldWide Gross

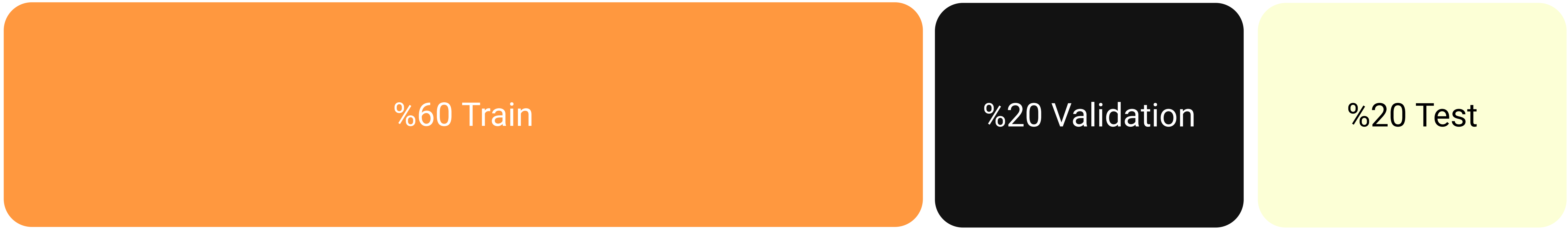
New Column:

Score = (Rating + (Metascore/10)) / 2

Estimated WorldWide Gross - Budget

Estimated Revenue

Train-Validation-Test Split



%60 Train

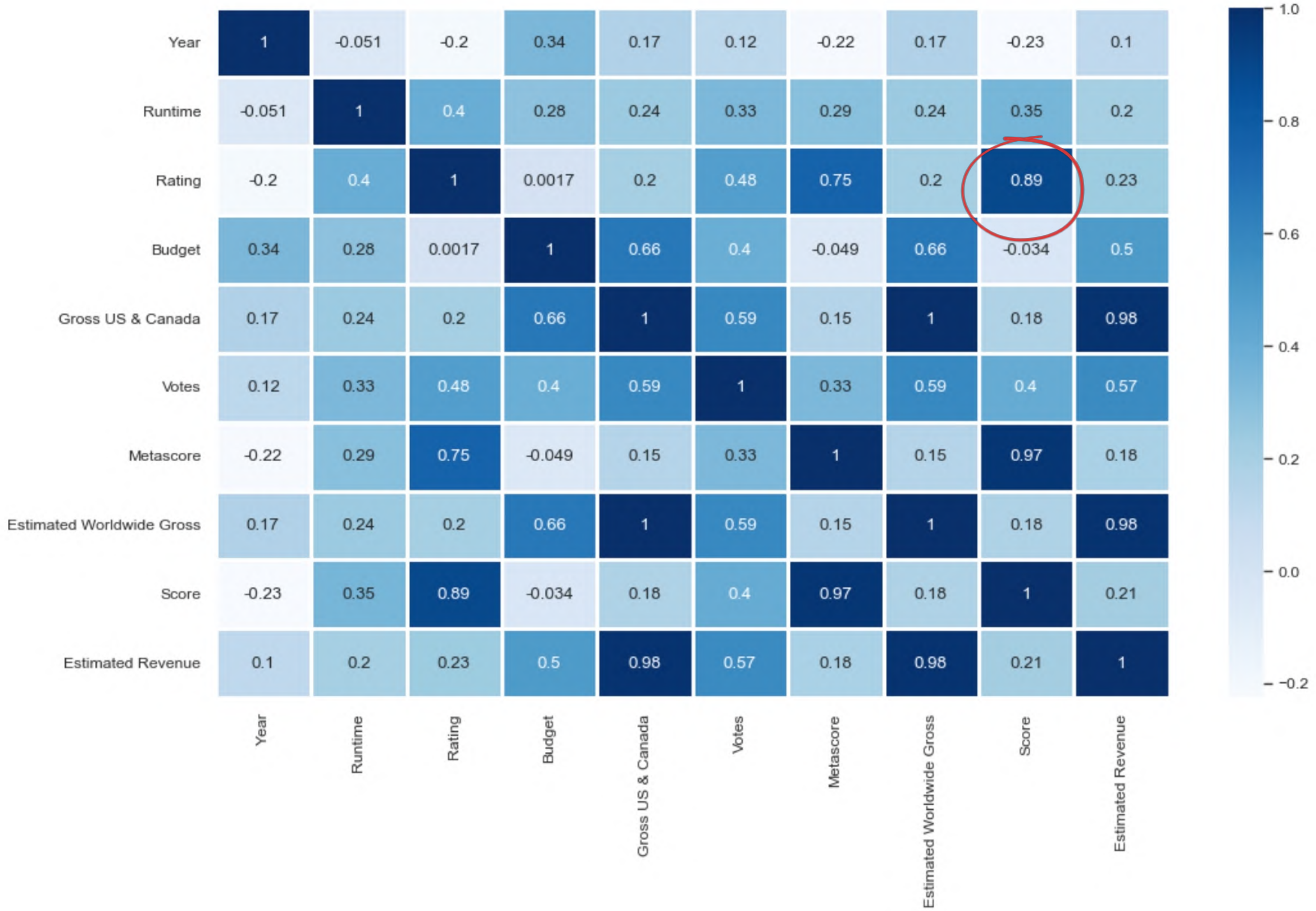
%20 Validation

%20 Test

Feature Engineering

New Column:

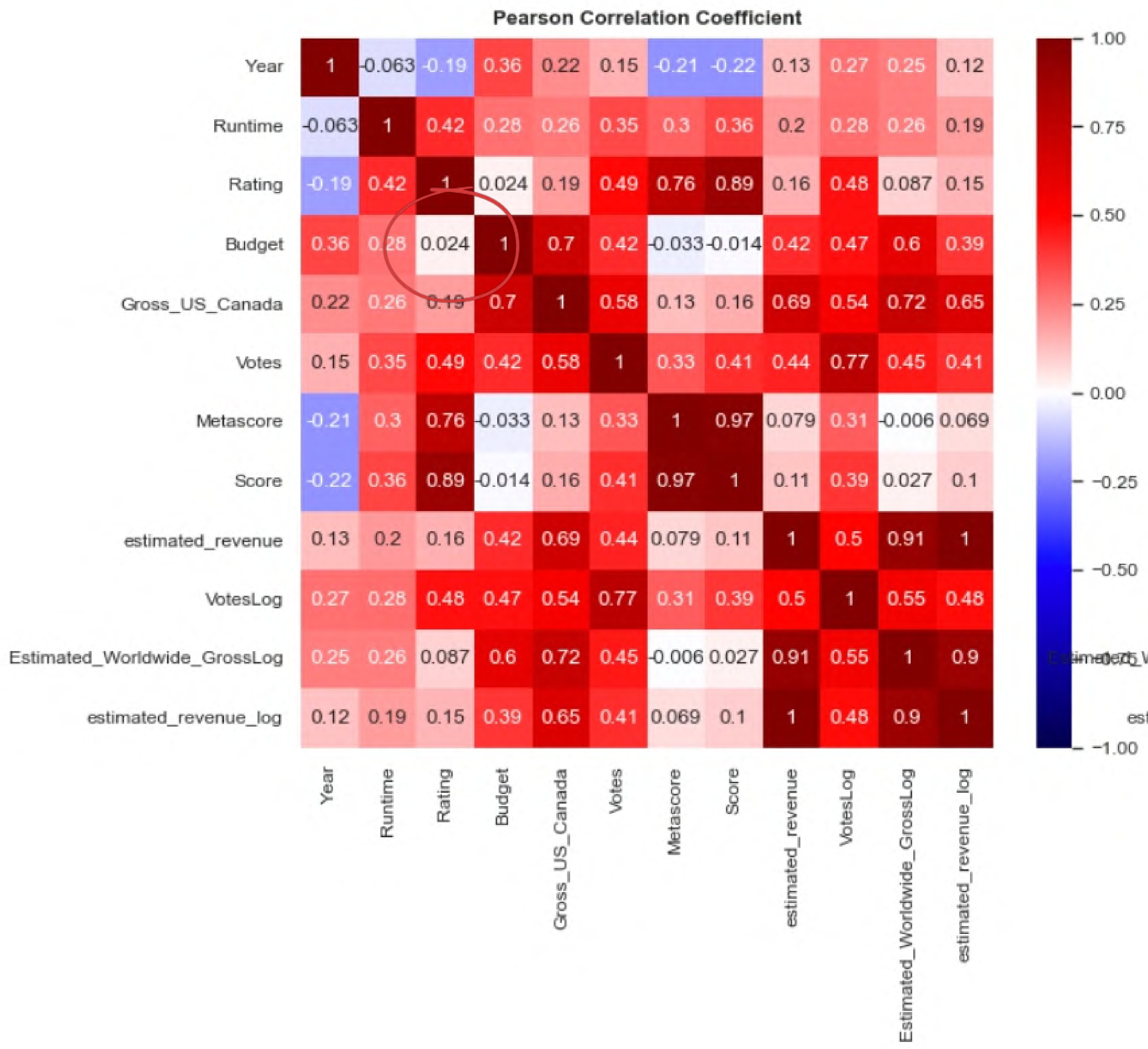
~~Score = (Rating + [illegible]tascore/10)) / 2~~



	Year	Runtime	Rating	Budget	Gross US & Canada	Votes	Metascore	Estimated Worldwide Gross	Score	Estimated Revenue
Year	1	-0.051	-0.2	0.34	0.17	0.12	-0.22	0.17	-0.23	0.1
Runtime	-0.051	1	0.4	0.28	0.24	0.33	0.29	0.24	0.35	0.2
Rating	-0.2	0.4	1	0.0017	0.2	0.48	0.75	0.2	0.89	0.23
Budget	0.34	0.28	0.0017	1	0.66	0.4	-0.049	0.66	-0.034	0.5
Gross US & Canada	0.17	0.24	0.2	0.66	1	0.59	0.15	1	0.18	0.98
Votes	0.12	0.33	0.48	0.4	0.59	1	0.33	0.59	0.4	0.57
Metascore	-0.22	0.29	0.75	-0.049	0.15	0.33	1	0.15	0.97	0.18
Estimated Worldwide Gross	0.17	0.24	0.2	0.66	1	0.59	0.15	1	0.18	0.98
Score	-0.23	0.35	0.89	-0.034	0.18	0.4	0.97	0.18	1	0.21
Estimated Revenue	0.1	0.2	0.23	0.5	0.98	0.57	0.18	0.98	0.21	1

Result = Multicollinearity

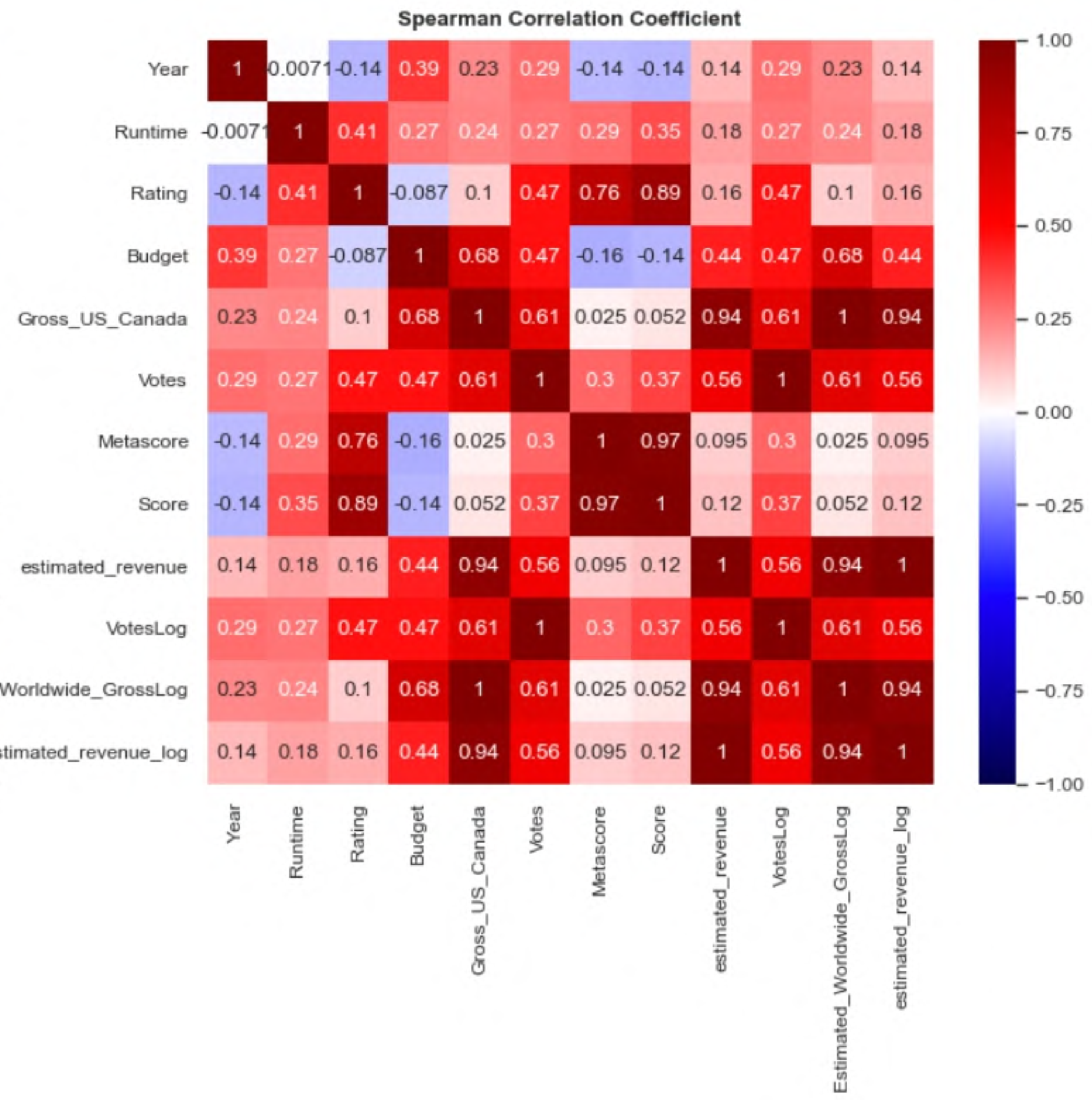
Comparison of Pearson Spearman



Pearson Correlation Coefficient

	Year	Runtime	Rating	Budget	Gross_US_Canada	Votes	Metascore	Score	estimated_revenue	VotesLog	Estimated_Worldwide_GrossLog	estimated_revenue_log
Year	1	-0.063	-0.19	0.36	0.22	0.15	-0.21	-0.22	0.13	0.27	0.25	0.12
Runtime	-0.063	1	0.42	0.28	0.26	0.35	0.3	0.36	0.2	0.28	0.26	0.19
Rating	-0.19	0.42	1	0.024	0.19	0.49	0.76	0.89	0.16	0.48	0.087	0.15
Budget	0.36	0.28	0.024	1	0.7	0.42	-0.033	-0.014	0.42	0.47	0.6	0.39
Gross_US_Canada	0.22	0.26	0.19	0.7	1	0.58	0.13	0.16	0.69	0.54	0.72	0.65
Votes	0.15	0.35	0.49	0.42	0.58	1	0.33	0.41	0.44	0.77	0.45	0.41
Metascore	-0.21	0.3	0.76	-0.033	0.13	0.33	1	0.97	0.079	0.31	-0.006	0.069
Score	-0.22	0.36	0.89	-0.014	0.16	0.41	0.97	1	0.11	0.39	0.027	0.1
estimated_revenue	0.13	0.2	0.16	0.42	0.69	0.44	0.079	0.11	1	0.5	0.91	1
VotesLog	0.27	0.28	0.48	0.47	0.54	0.77	0.31	0.39	0.5	1	0.55	0.48
Estimated_Worldwide_GrossLog	0.25	0.26	0.087	0.6	0.72	0.45	-0.006	0.027	0.91	0.55	1	0.9
estimated_revenue_log	0.12	0.19	0.15	0.39	0.65	0.41	0.069	0.1	1	0.48	0.9	1

Linear Relationship



Spearman Correlation Coefficient

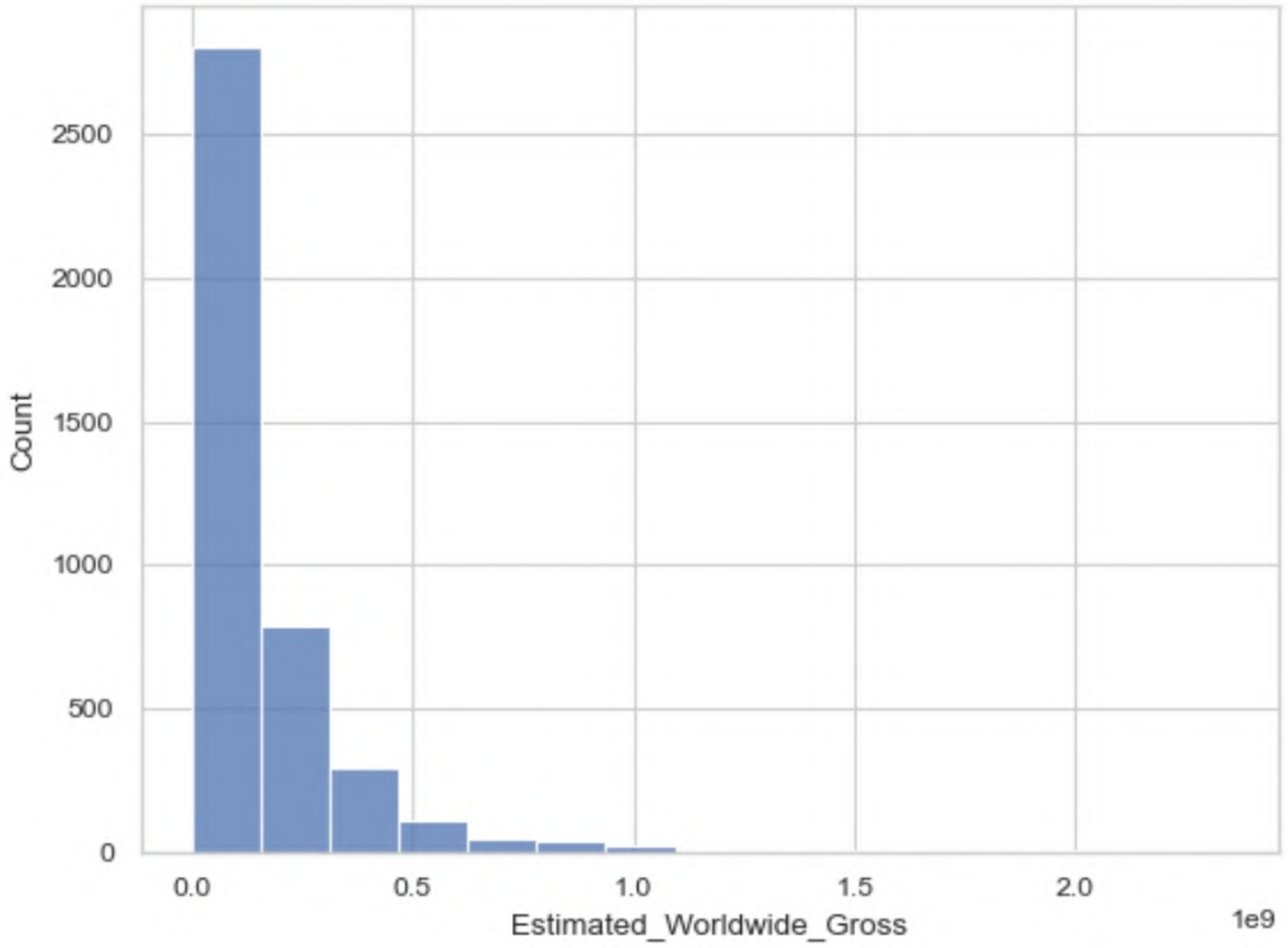
	Year	Runtime	Rating	Budget	Gross_US_Canada	Votes	Metascore	Score	estimated_revenue	VotesLog	Estimated_Worldwide_GrossLog	estimated_revenue_log
Year	1	-0.0071	-0.14	0.39	0.23	0.29	-0.14	-0.14	0.14	0.29	0.23	0.14
Runtime	-0.0071	1	0.41	0.27	0.24	0.27	0.29	0.35	0.18	0.27	0.24	0.18
Rating	-0.14	0.41	1	-0.087	0.1	0.47	0.76	0.89	0.16	0.47	0.1	0.16
Budget	0.39	0.27	-0.087	1	0.68	0.47	-0.16	-0.14	0.44	0.47	0.68	0.44
Gross_US_Canada	0.23	0.24	0.1	0.68	1	0.61	0.025	0.052	0.94	0.61	1	0.94
Votes	0.29	0.27	0.47	0.47	0.61	1	0.3	0.37	0.56	1	0.61	0.56
Metascore	-0.14	0.29	0.76	-0.16	0.025	0.3	1	0.97	0.095	0.3	0.025	0.095
Score	-0.14	0.35	0.89	-0.14	0.052	0.37	0.97	1	0.12	0.37	0.052	0.12
estimated_revenue	0.14	0.18	0.16	0.44	0.94	0.56	0.095	0.12	1	0.56	0.94	1
VotesLog	0.29	0.27	0.47	0.47	0.61	1	0.3	0.37	0.56	1	0.61	0.56
Estimated_Worldwide_GrossLog	0.23	0.24	0.1	0.68	1	0.61	0.025	0.052	0.94	0.61	1	0.94
estimated_revenue_log	0.14	0.18	0.16	0.44	0.94	0.56	0.095	0.12	1	0.56	0.94	1

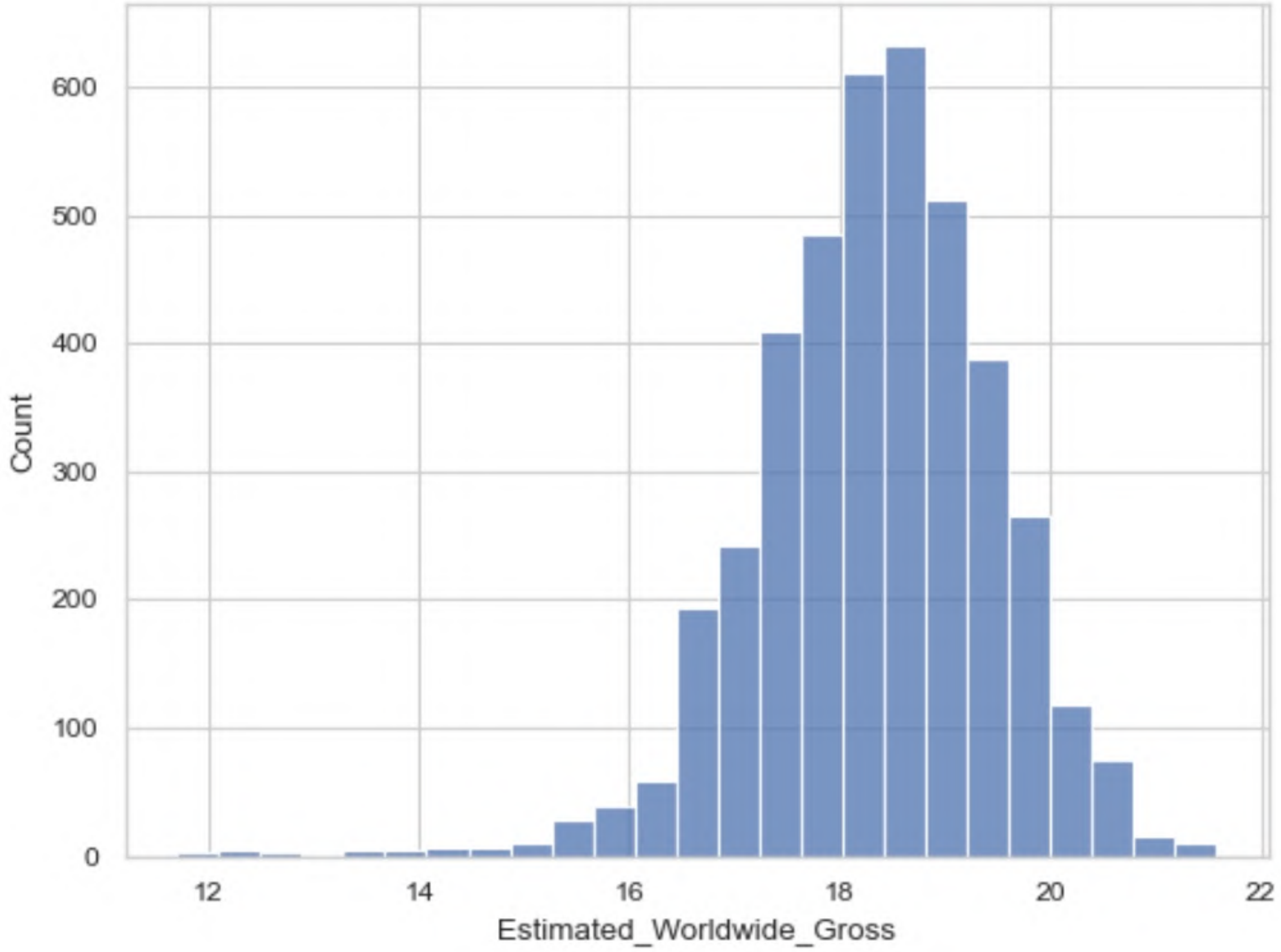
Rank Correlation

Model Development

- Analyze the behavior of the dataset.

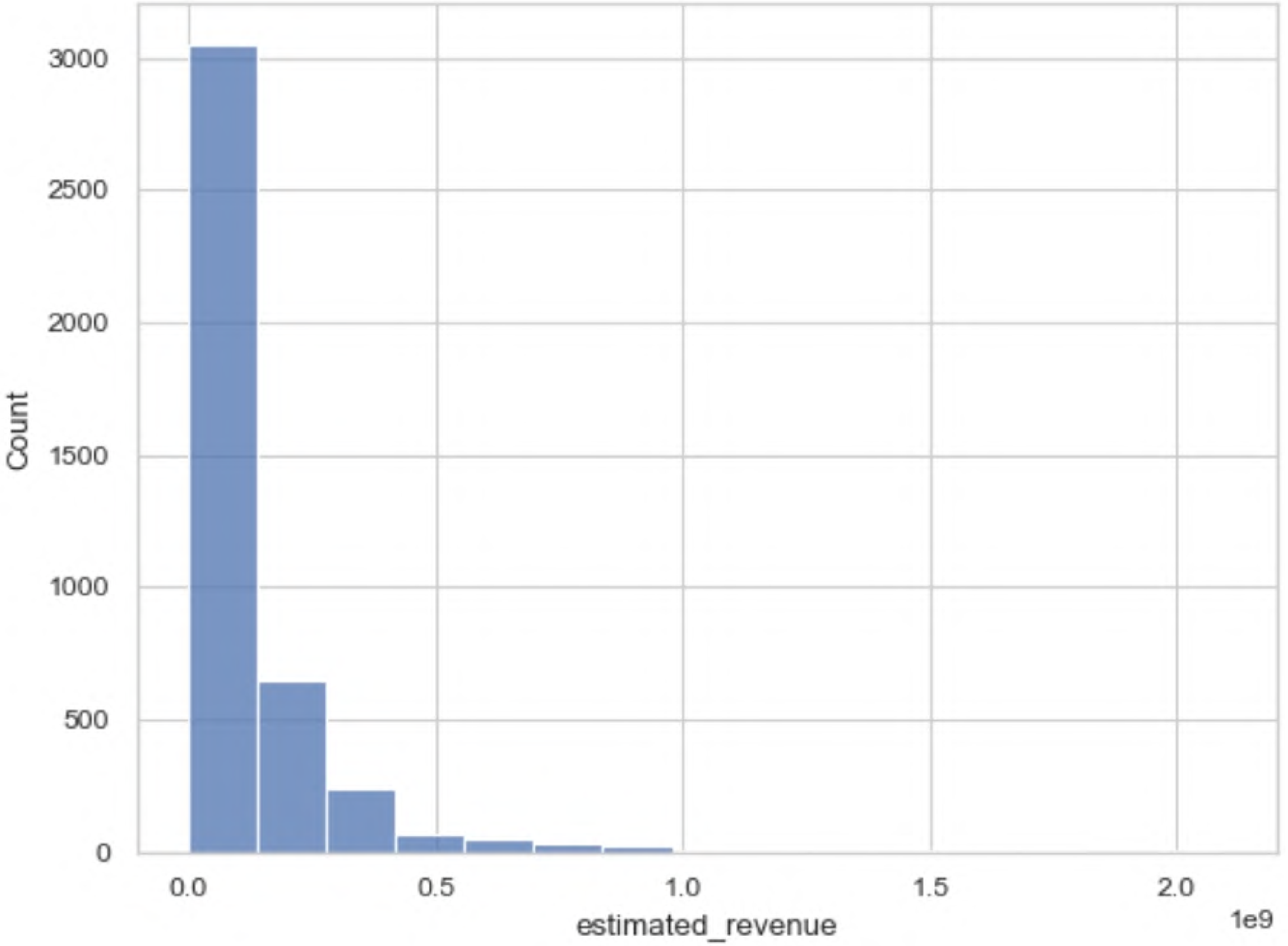
- Filter&Manipulate The behavior

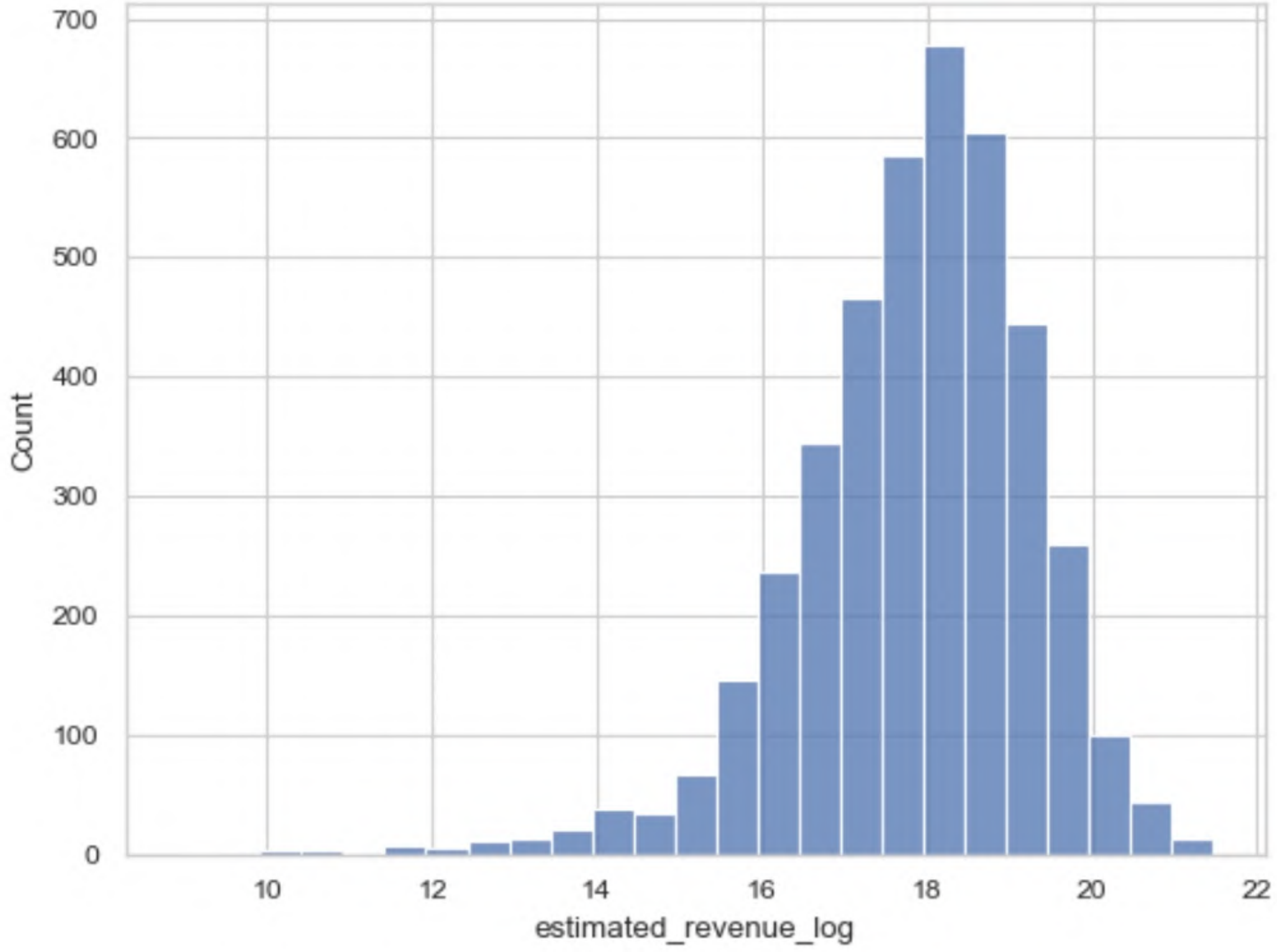




Model Development

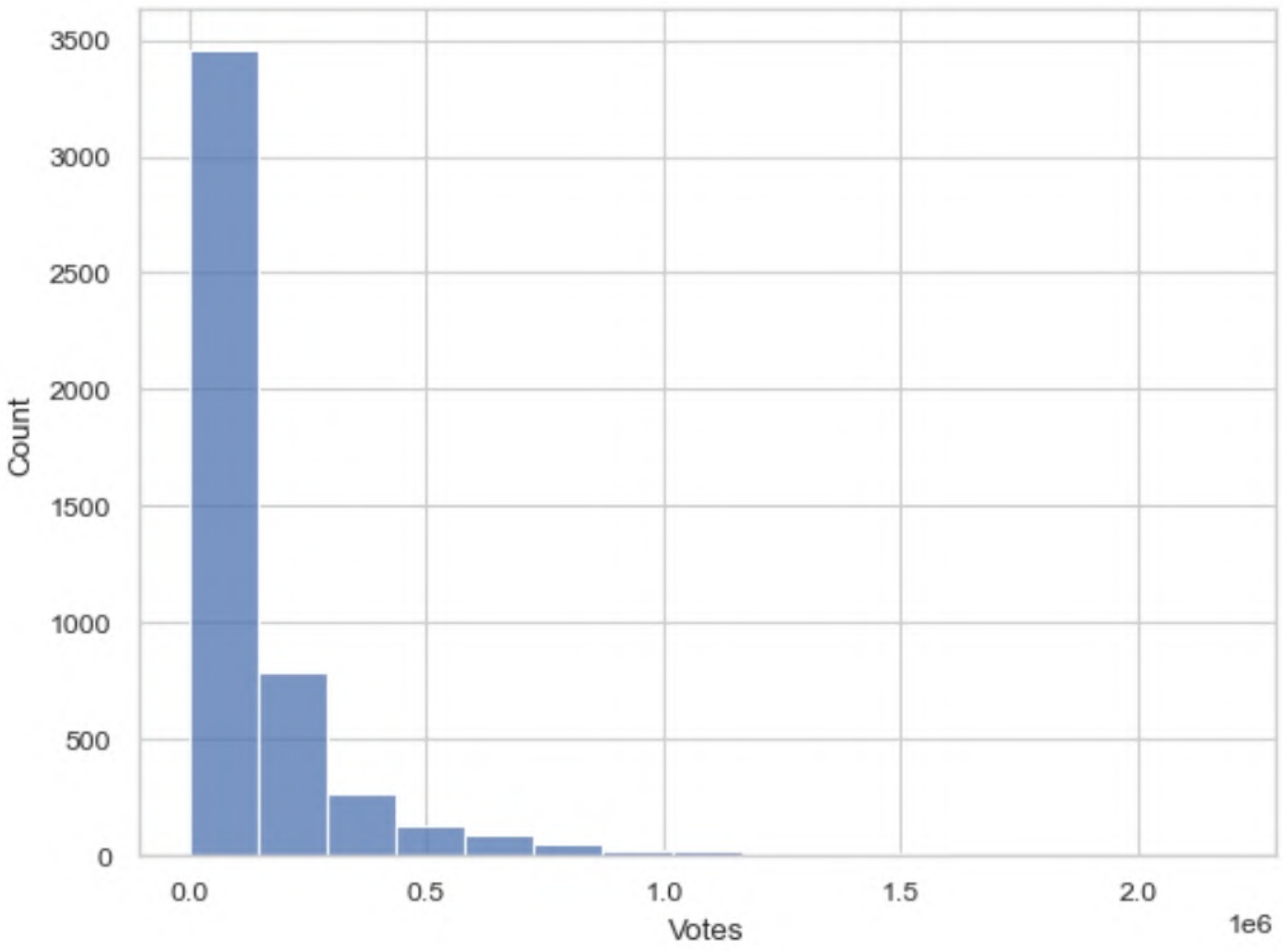
- Analyze the behavior of the dataset. ⟶
- Filter&Manipulate The behavior

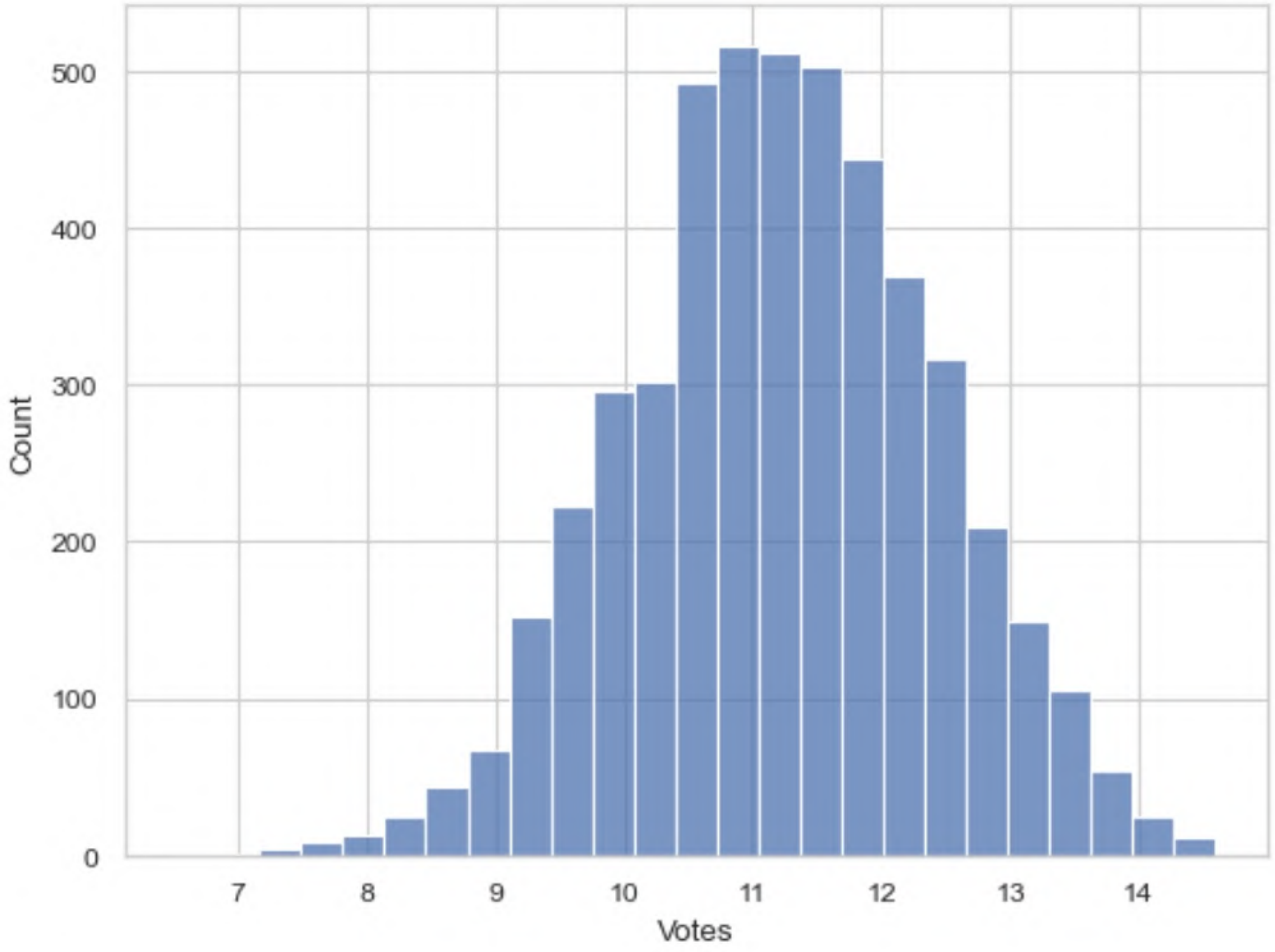




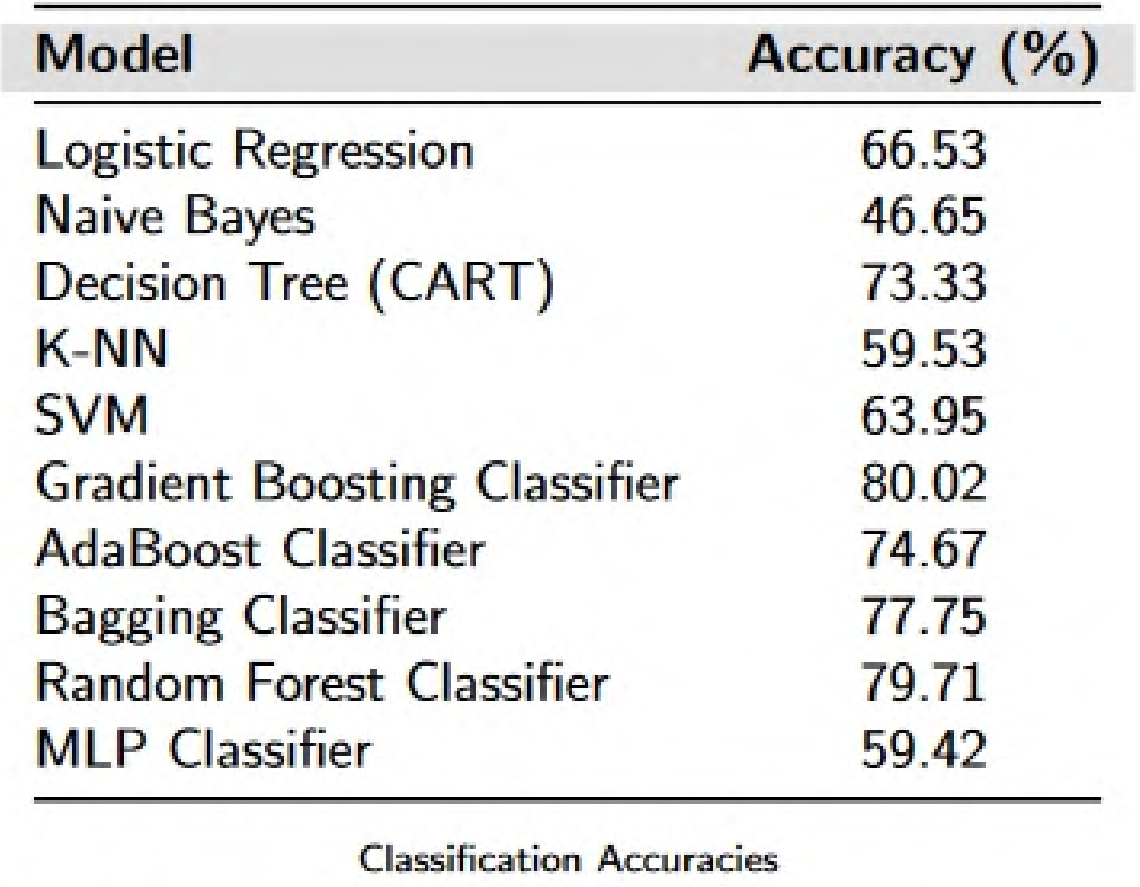
Model Development

- Analyze the behavior of the dataset. →
- Filter&Manipulate The behavior



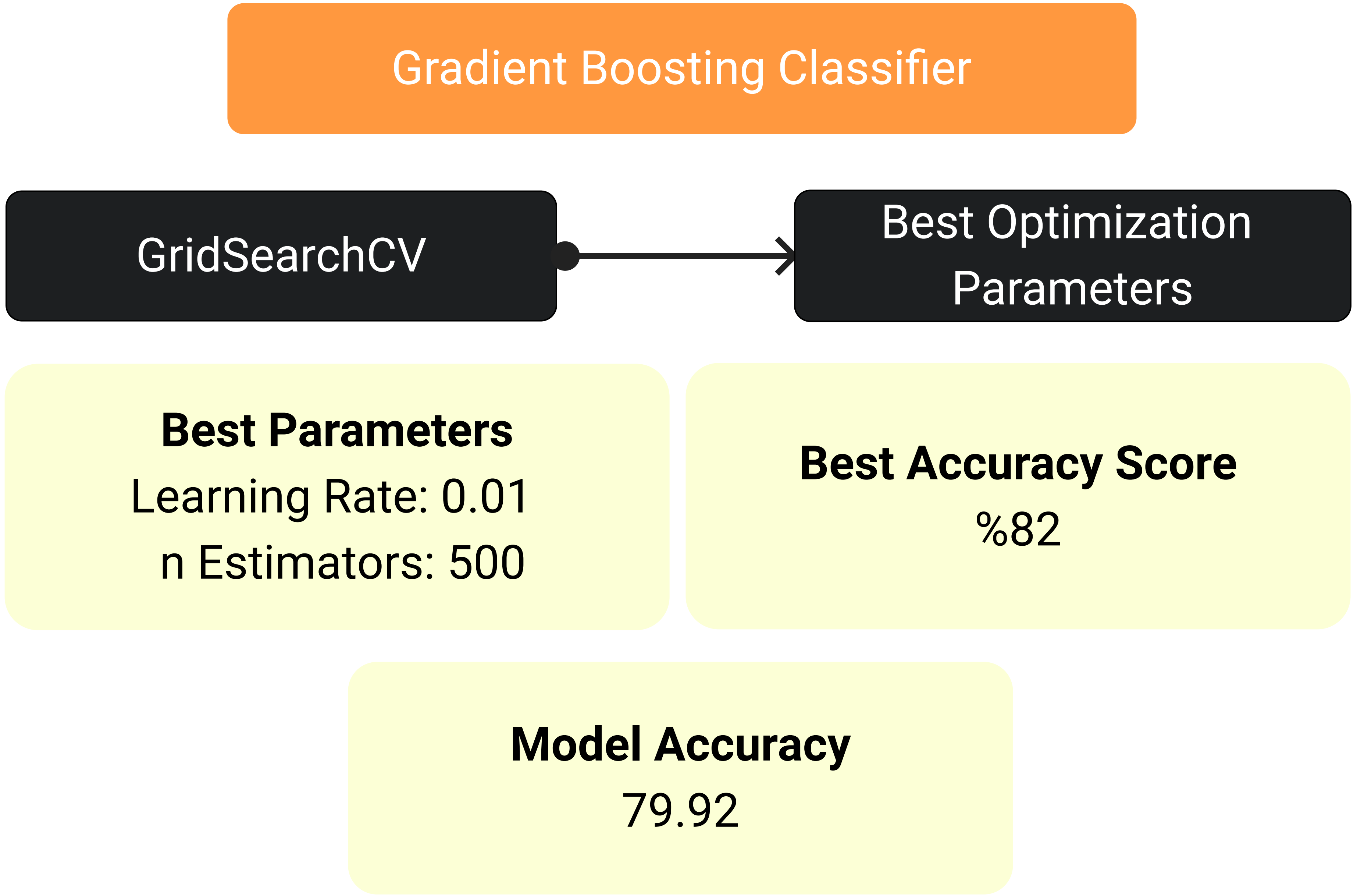


Multiple-Model Techniques



Model	Accuracy (%)
Logistic Regression	66.53
Naive Bayes	46.65
Decision Tree (CART)	73.33
K-NN	59.53
SVM	63.95
Gradient Boosting Classifier	80.02
AdaBoost Classifier	74.67
Bagging Classifier	77.75
Random Forest Classifier	79.71
MLP Classifier	59.42

Classification Accuracies



Gradient Boosting Classifier

GridSearchCV → Best Optimization Parameters

Best Parameters
Learning Rate: 0.01
n Estimators: 500

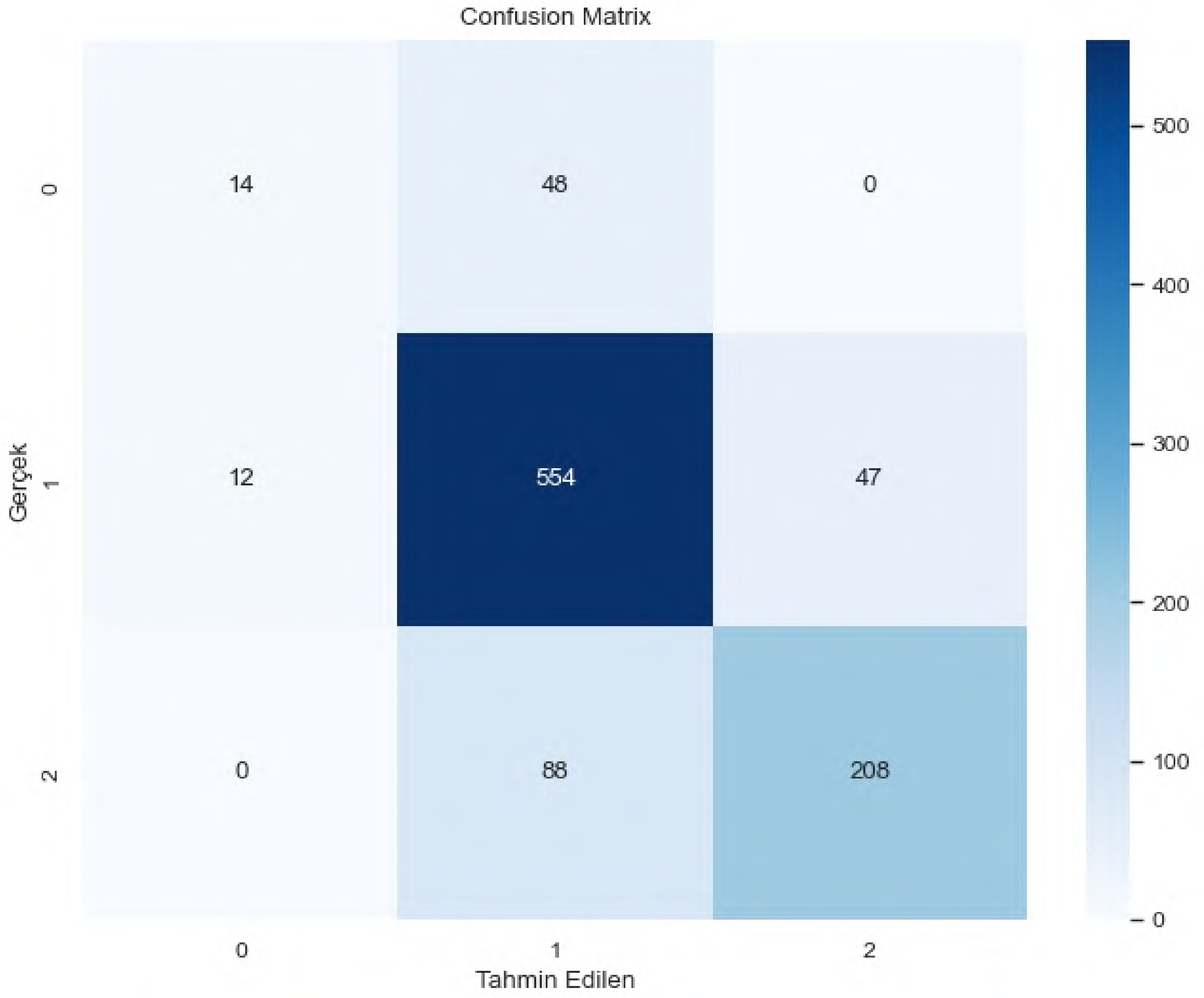
Best Accuracy Score
%82

Model Accuracy
79.92

Gradient Boosting Classifier

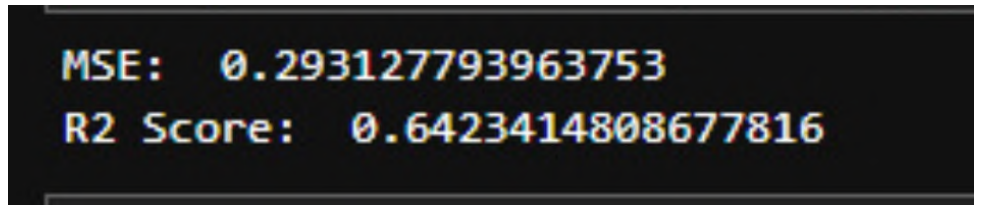
Model Control

Accuracy	%79.92
Precision	%78.99
Recall	%79.92
F1-Score	%78.73



Model Evaluation - Train

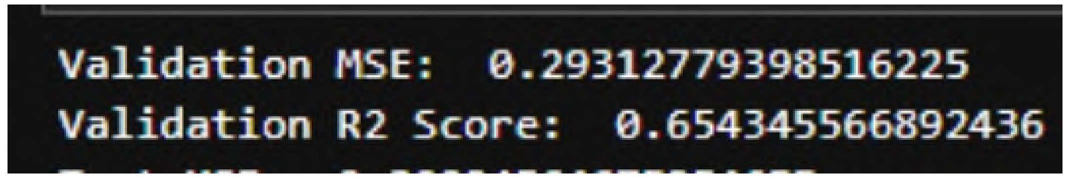
%60 Train



```
MSE:  0.293127793963753
R2 Score:  0.6423414808677816
```

Model Evaluation - Validation

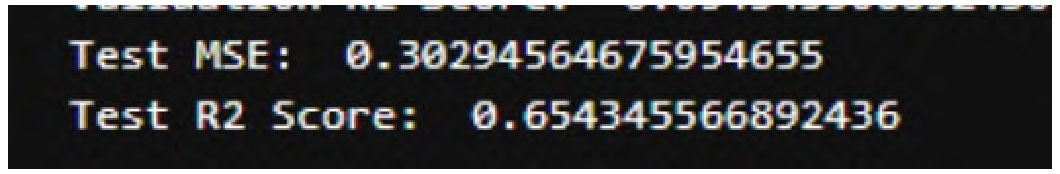
%20 Validation



```
Validation MSE:  0.29312779398516225
Validation R2 Score:  0.654345566892436
```

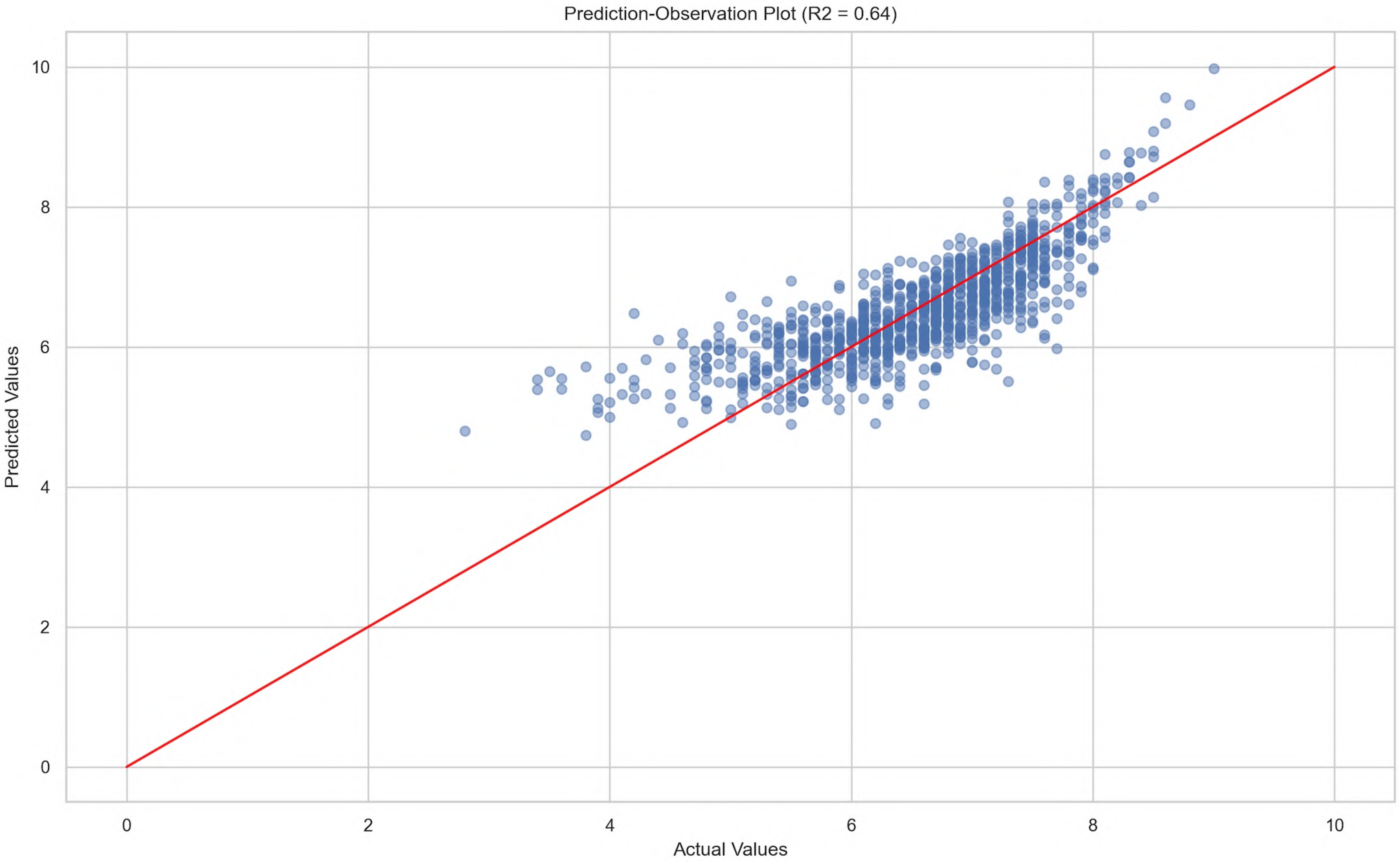
Model Evaluation - Test

%20 Test



```
Test MSE:  0.30294564675954655
Test R2 Score:  0.654345566892436
```

Model Development



Conclusion & Future Work

Features: "Year", "Runtime", "Gross US & Canada", "Votes", "Metascore", "Estimated Revenue", "Budget"

Target: Rating

Success Rate: %65.5

Future Work

Director Data

Actor Data