01

REGRESSION ANALYSIS

"MOVIE INDUSTRY"

TEAM 5 - BAN 1

VINCENT | TOSHIAKI | ARJUN

ABOUT THIS ANALYSIS

- Movie industry
- Over 7000 records
- Information on:
  - Votes
  - Runtime
  - Gross
  - Budget

- Univariate and Bivariate Analysis.
- Outliers, missing values.
- Shape before: (7668, 6)
- Shape after: (4502, 52)
- Regression with scaler

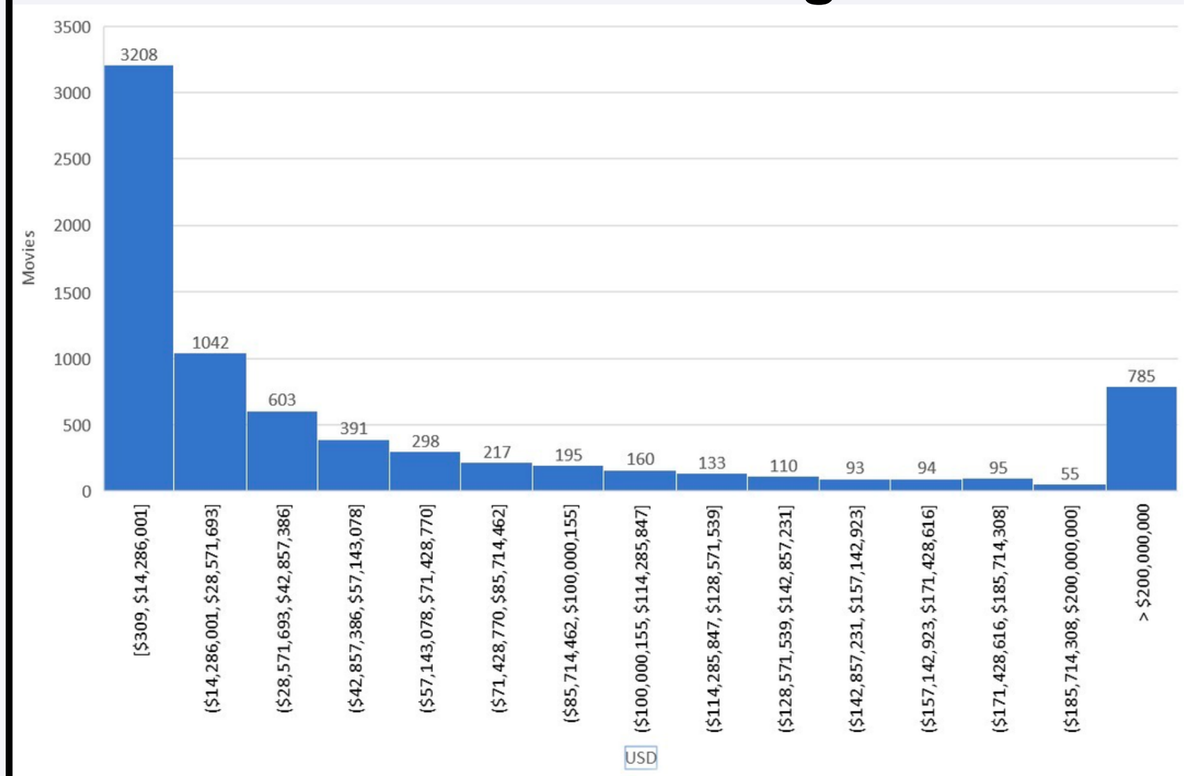
Source: www.kaggle.com/datasets/danielgrijalvas/movies/code

BUSINESS CASE

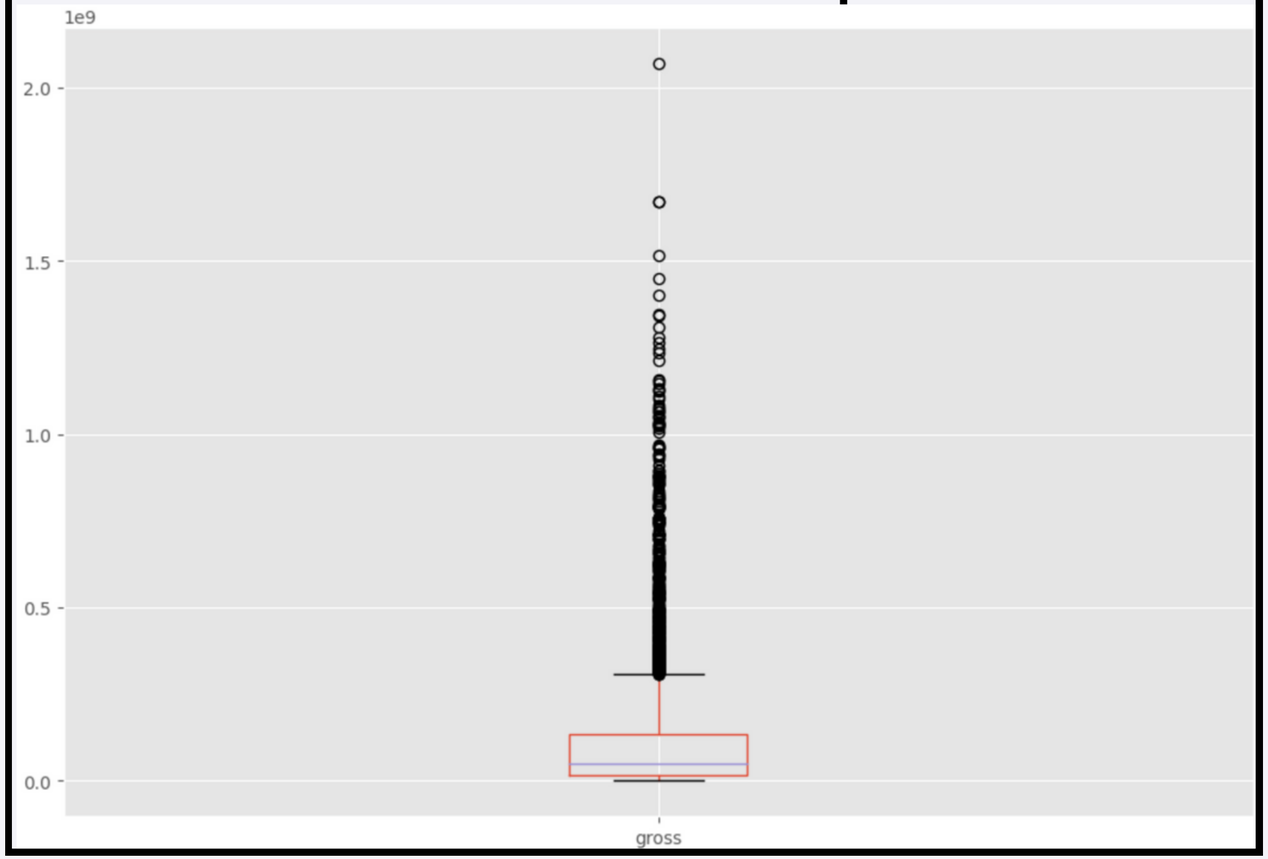
What are the driving factors of a good-scoring movie?

UNIVARIATE ANALYSIS

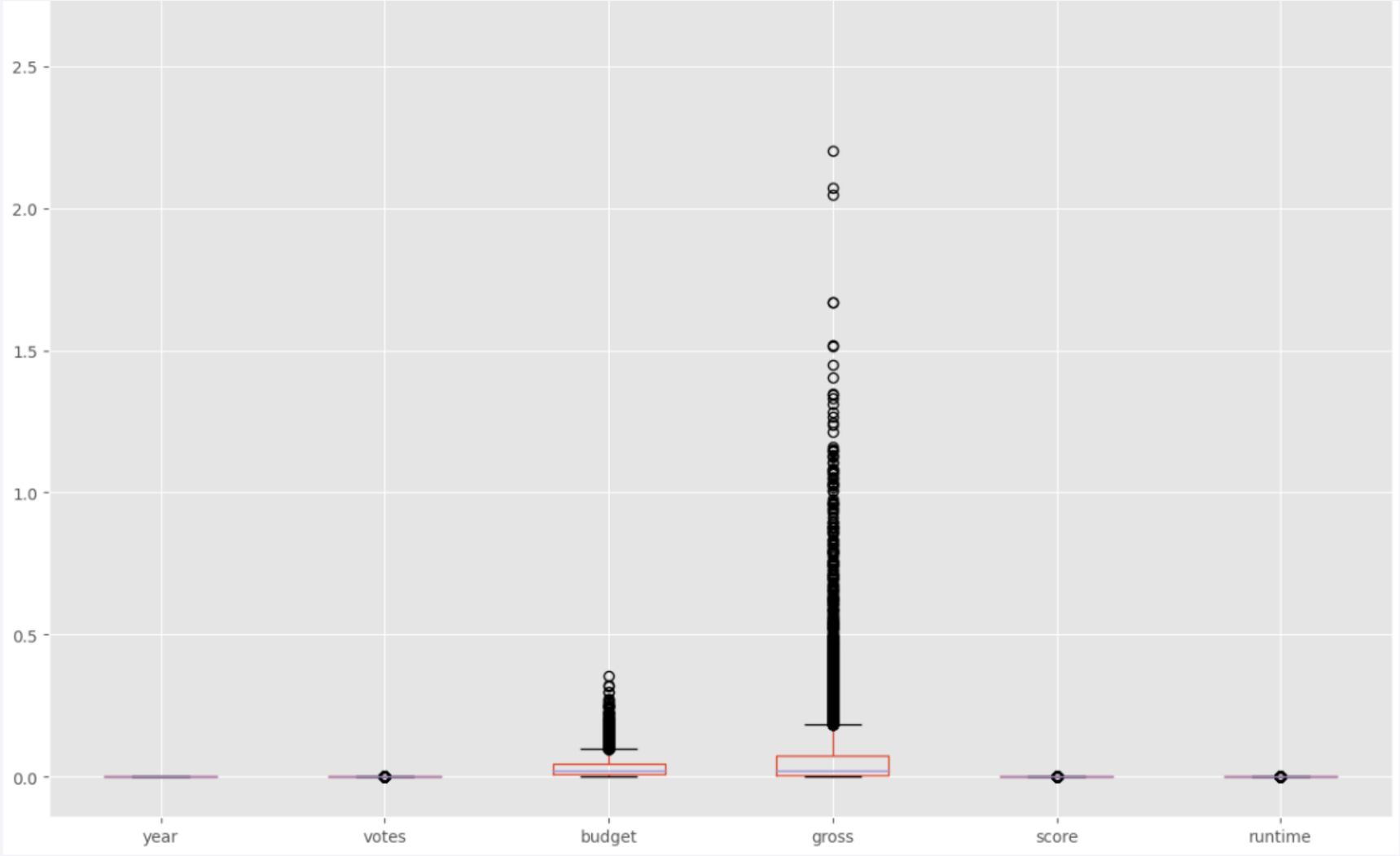
Gross Revenue Histogram



Gross Revenue Boxplot



UNIVARIATE ANALYSIS



UNIVARIATE ANALYSIS GROSS

MIN	MAX
$2,554	$2,069,521,700

MEAN	MEDIAN	MODE
$115,114,071.48	$49,843,011	$140,000,000

UNIVARIATE ANALYSIS

VOTES

MEAN	115,066
MEDIAN	52,000
MODE	NaN
MIN	195
MAX	2,400,000

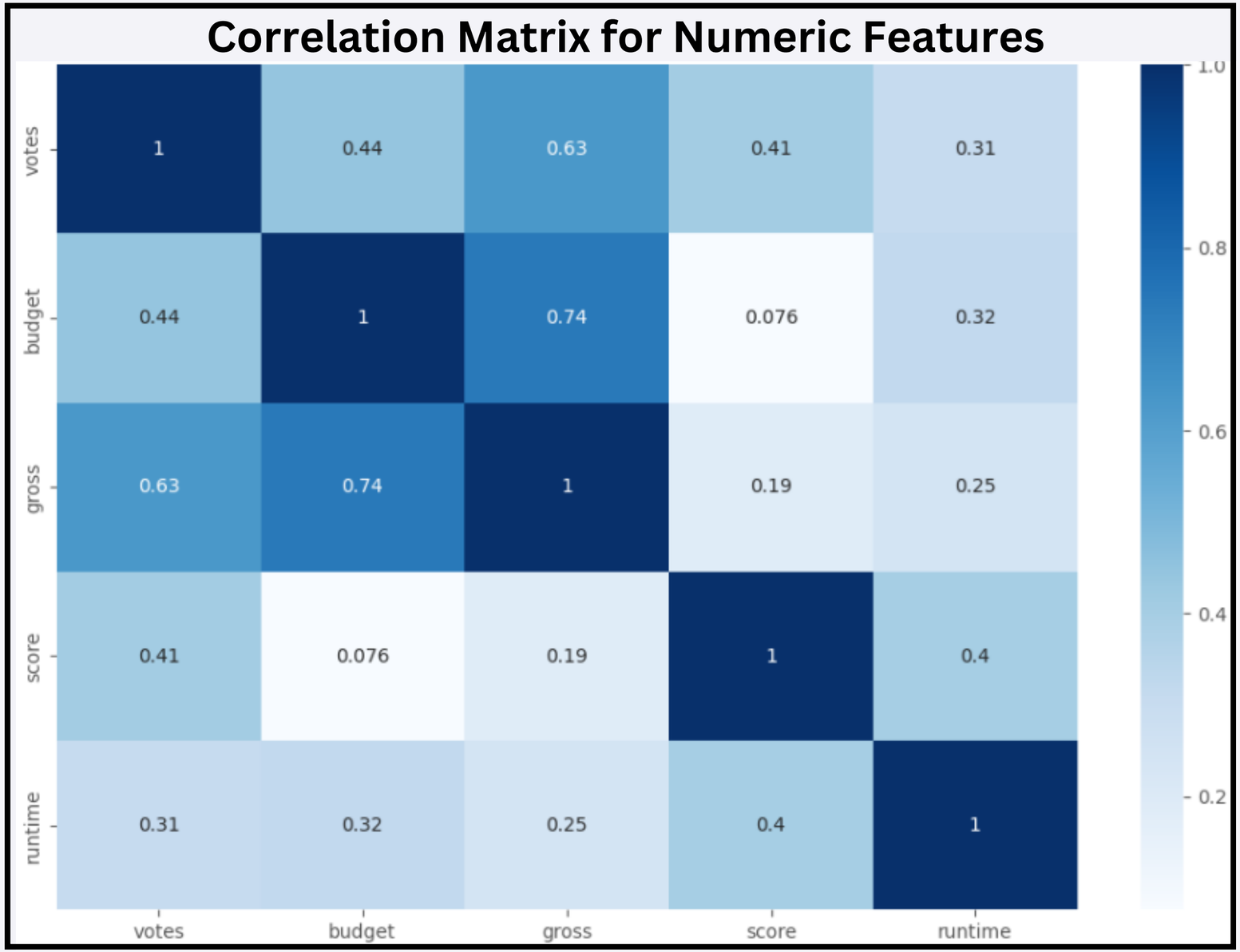
BUDGET

MEAN	39,945,720
MEDIAN	250,000,000
MODE	30,000,000
MIN	6,000
MAX	356,000,000

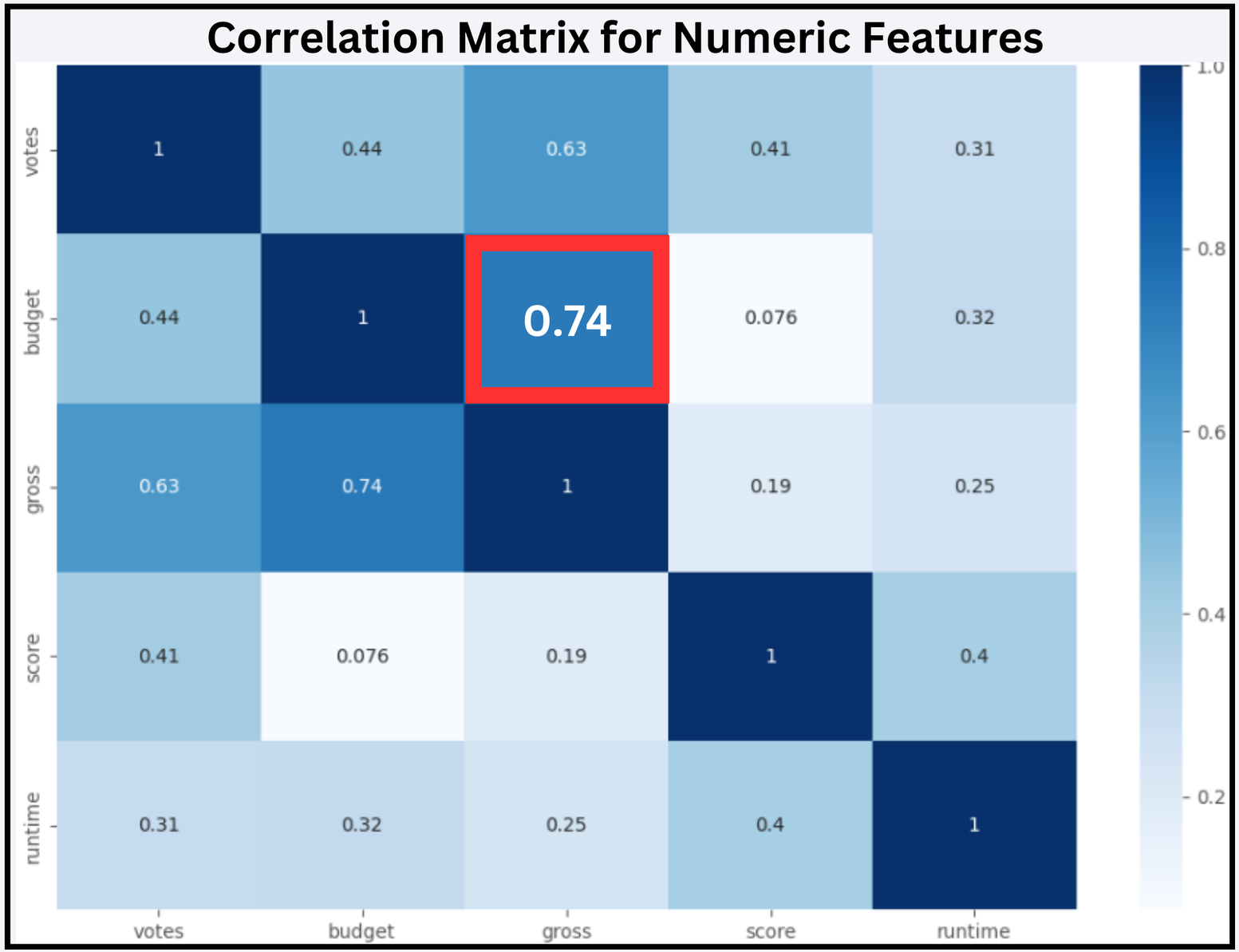
RUNTIME

MEAN	109
MEDIAN	106
MODE	101
MIN	63
MAX	271

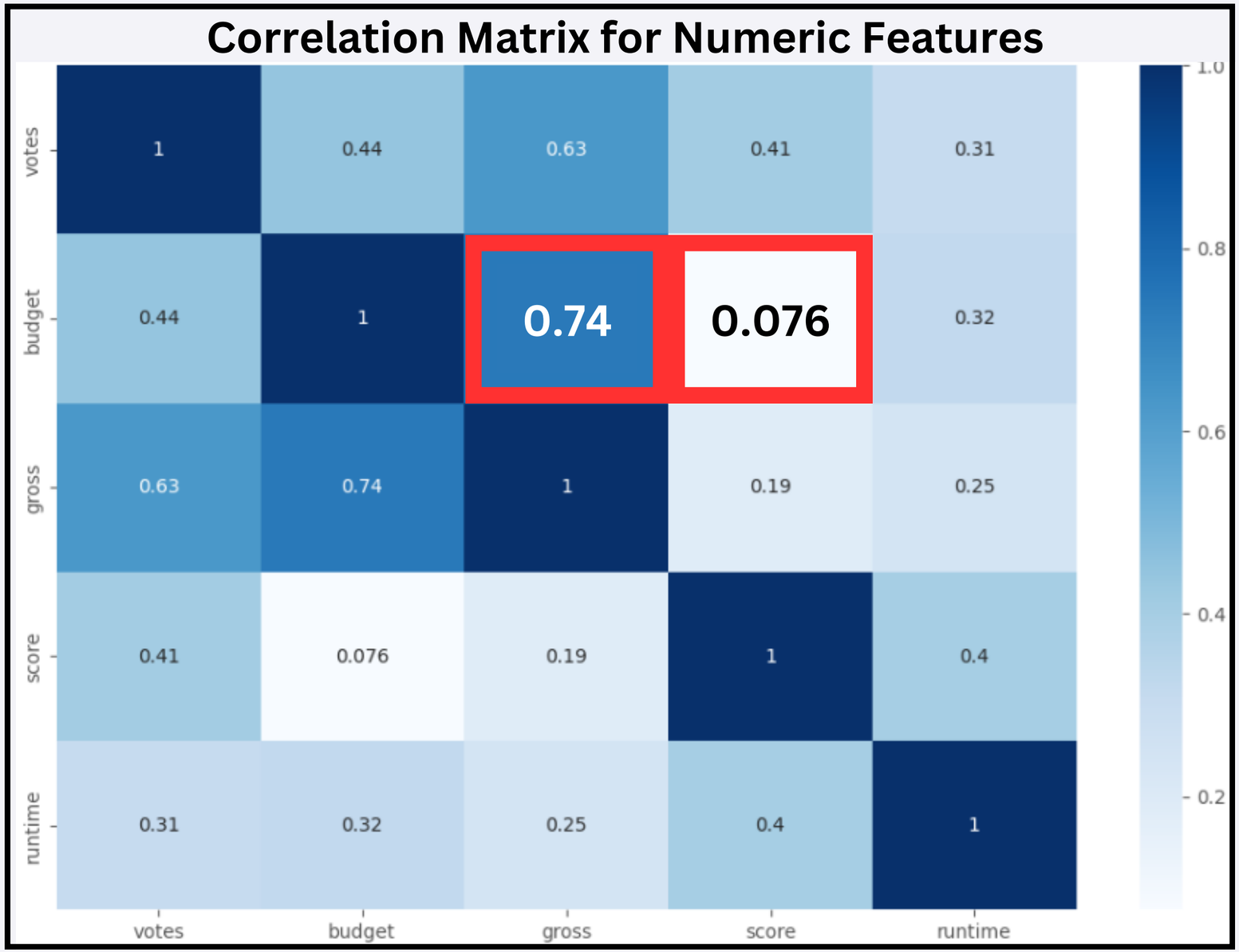
BIVARIATE ANALYSIS



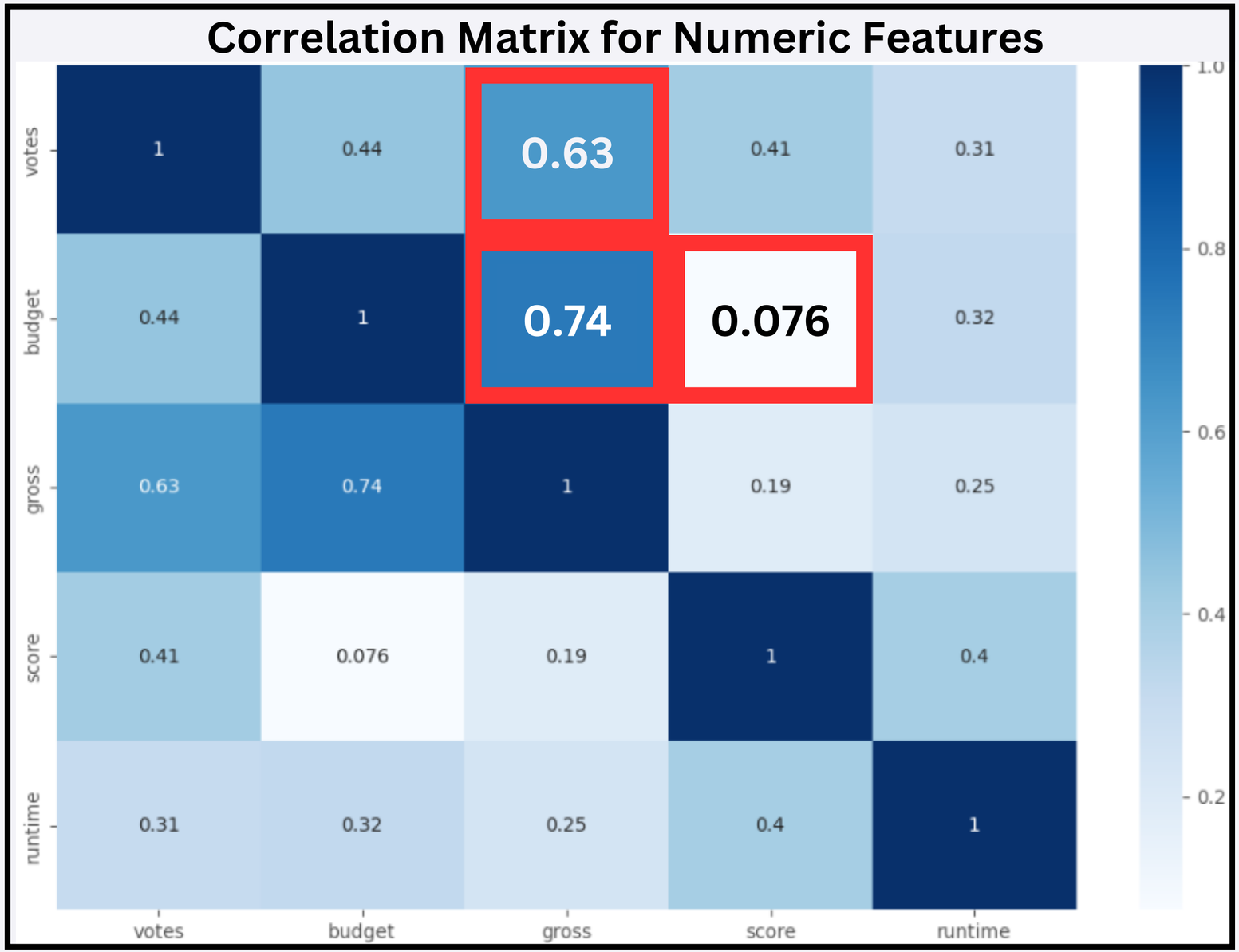
BIVARIATE ANALYSIS



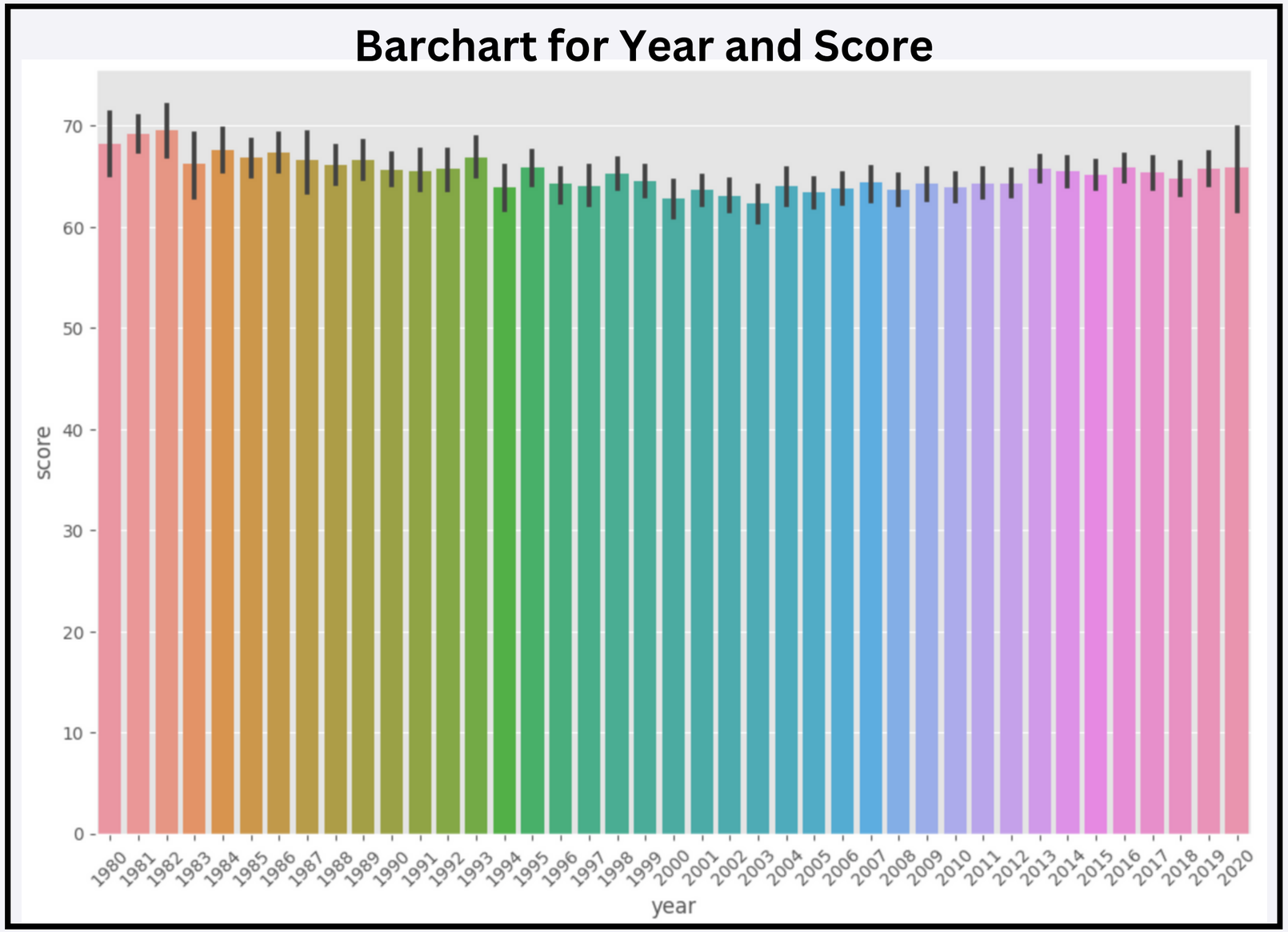
BIVARIATE ANALYSIS



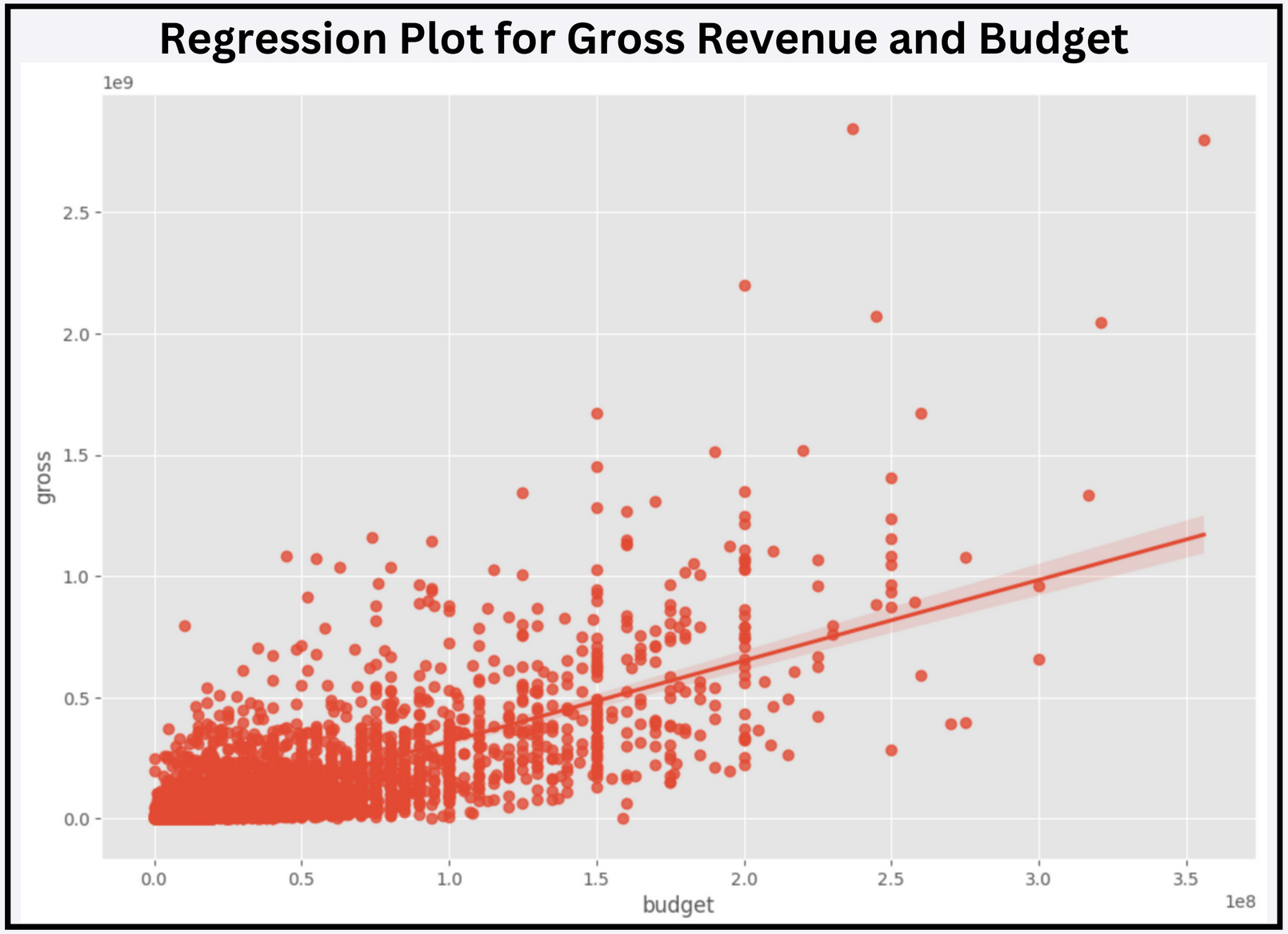
BIVARIATE ANALYSIS



YEAR V/S SCORE



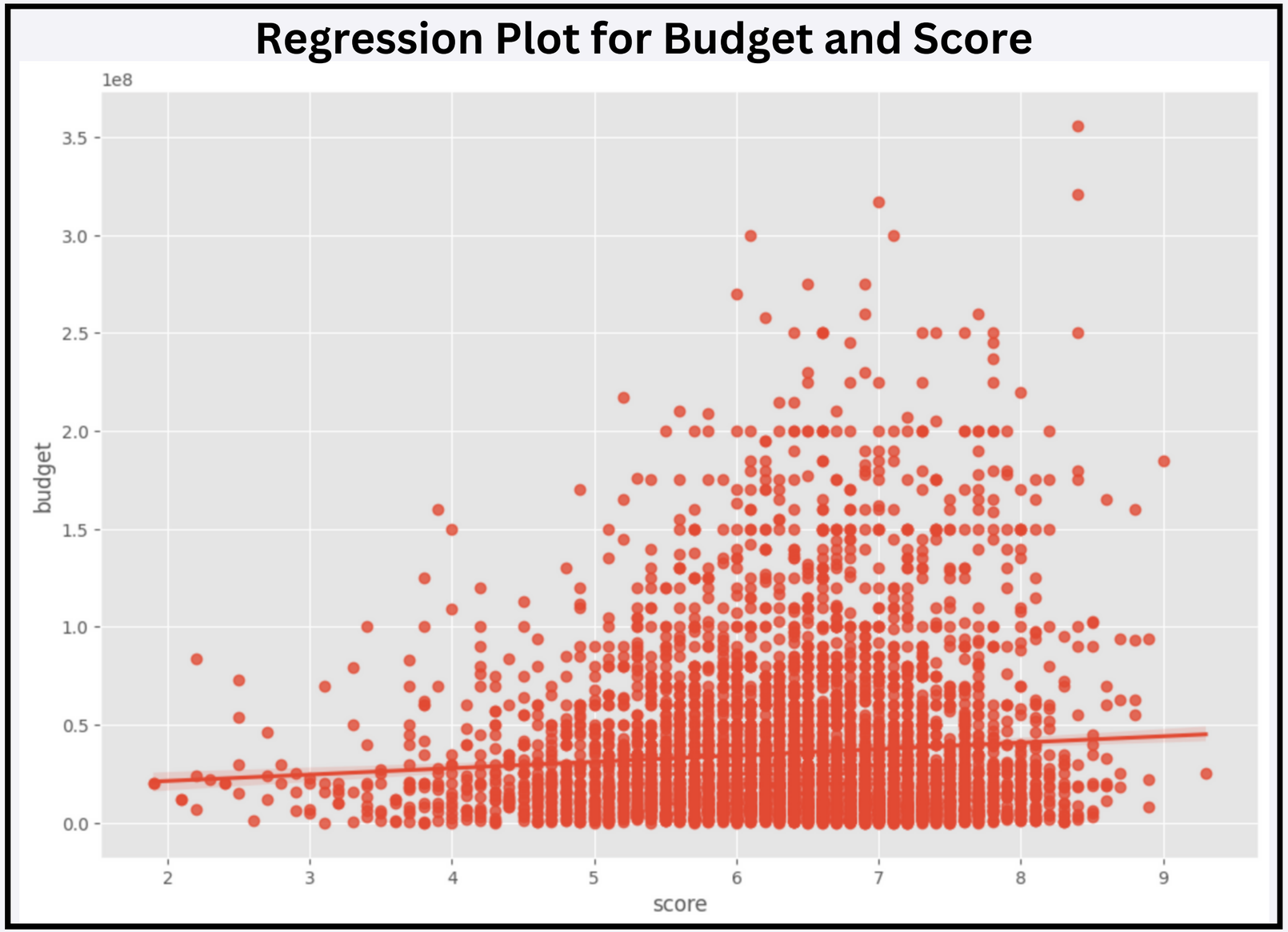
GROSS V/S BUDGET



CORRELATION

0.74

BUDGET V/S SCORE



CORRELATION

0.076

REGRESSION ANALYSIS

	Model	Score
2	Linear Regression 3	57.68
1	Linear Regression 2	42.91
0	Linear Regression 1	38.59

- Country Dummies
- Scaled Data
- Holidays
- Famous Writer, Director, Star

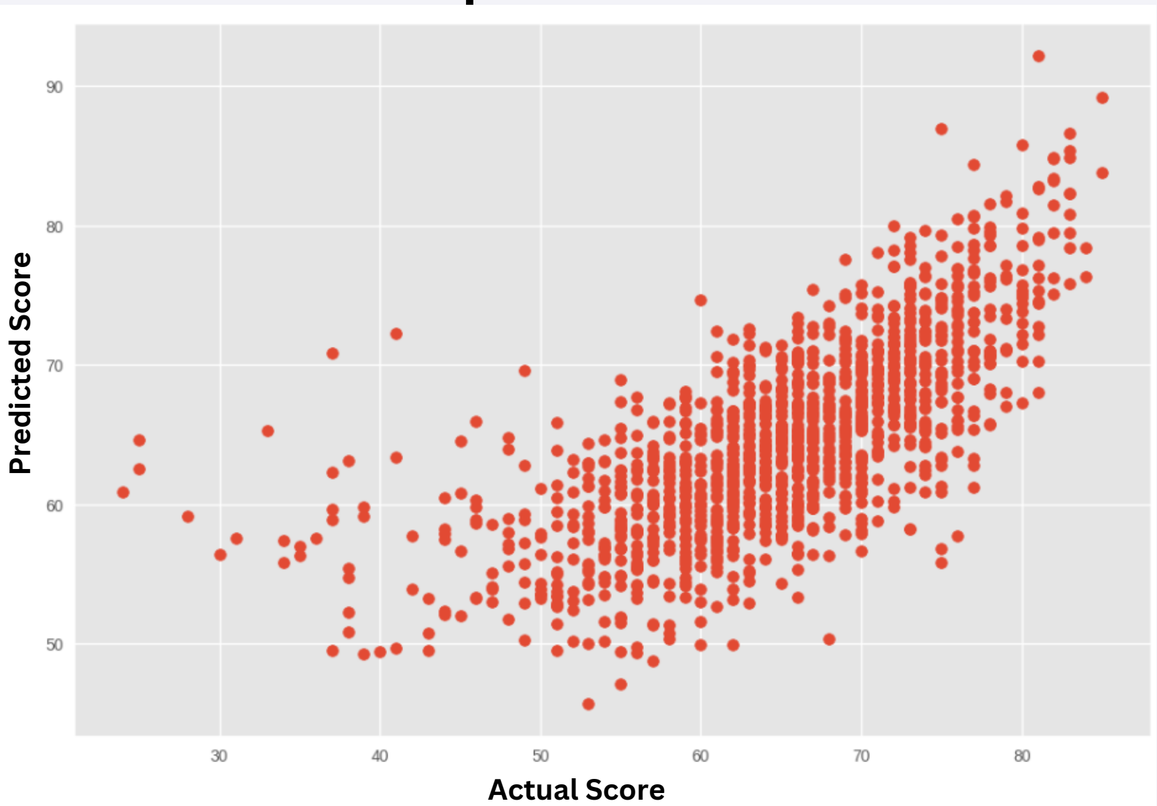
- Genre Dummies

- Numeric Variables

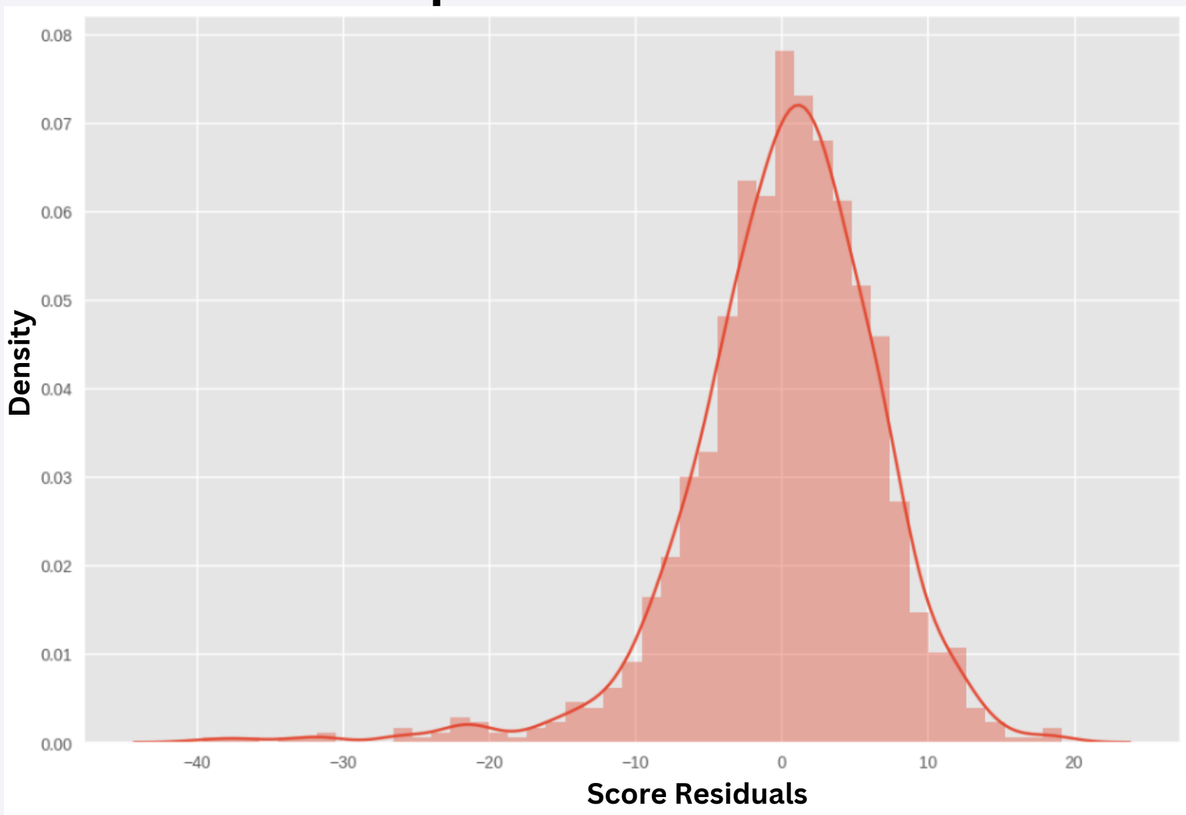
p-values were all under 0.005

RESIDUAL PLOTS

Scatterplot of Residuals



Distplot of Residuals



FINDINGS & RECOMMENDATIONS

FINDINGS

1. Budget ≠ Score
2. Marvel Studios

RECOMMENDATIONS

1. USA or Not ?
2. Feedback is Key
3. Best Genres

LEARNINGS

1. **Bias** is closer to a project than we realize.
   As for us with the **year** example.

2. **Collaboration** is a must know skill.

3. **Data can be normalized** which makes for **better regression analysis.**