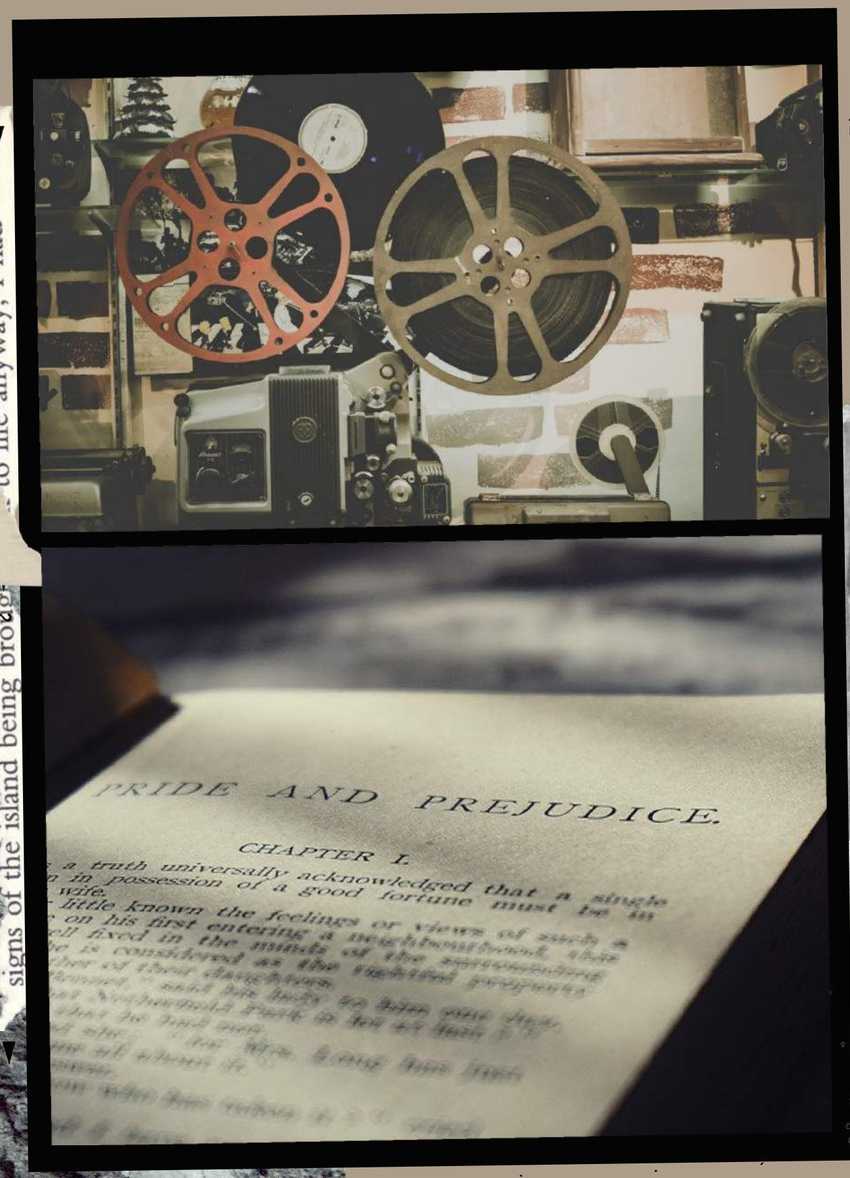
Linear Regression With Movies Based On Novels.

Outline:

- Problem Statement and Motivation.
- Data Collection.
- Data Cleaning.
- Methodology.
- Result.
- Future Steps.



Motivation

Create a Linear regression model that can predict World Wild Gross of movies based on novels by determining the features most influential to their success.

Data Collection:

We scraping three web site to collect our data by merge the flowing :

1- IMDB website:

Https://www.imdb.com/search/keyword/?keywords=based-on_novel

Title, Rating, Runtime, IMDB scores, Genre.

Total of 3450 dataset.

2-BoxOfficeMojo website.

https://www.boxofficemojo.com

The theaters, Total gross.

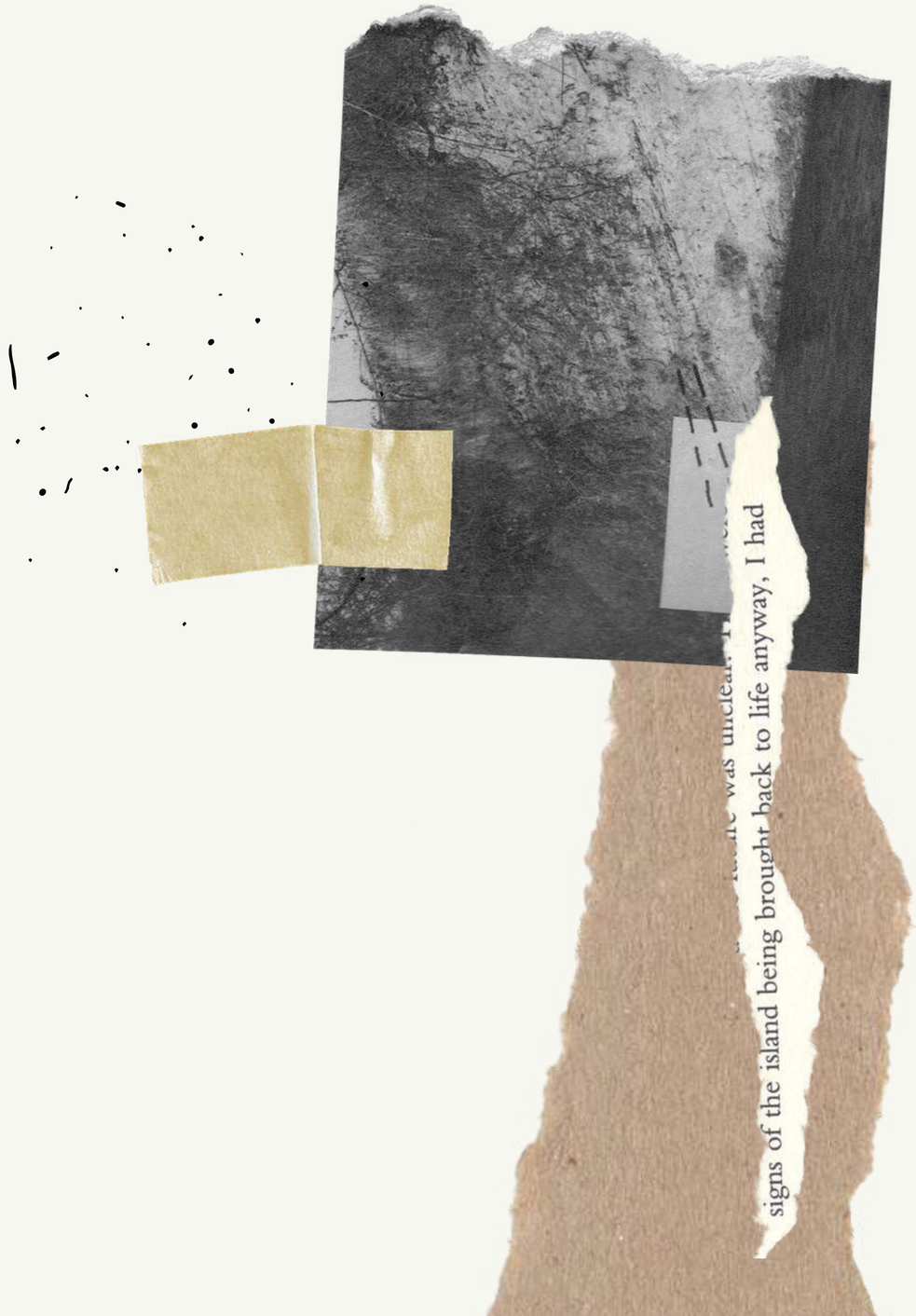
Total of 999 dataset.

3- The Numbers website.

https://www.the-numbers.com

Title, Budget, Worldwide gross.

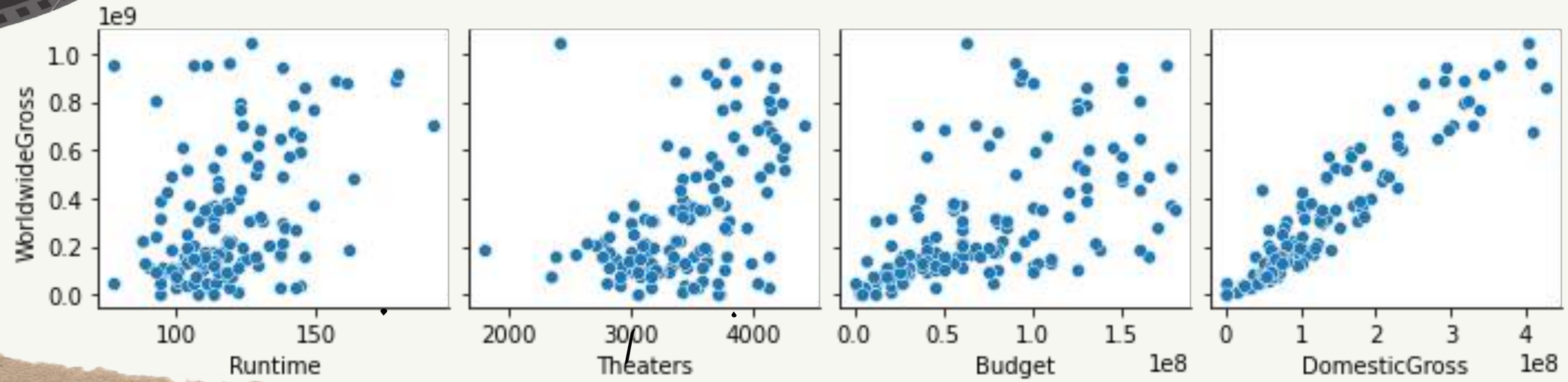
Total of 6100 dataset.

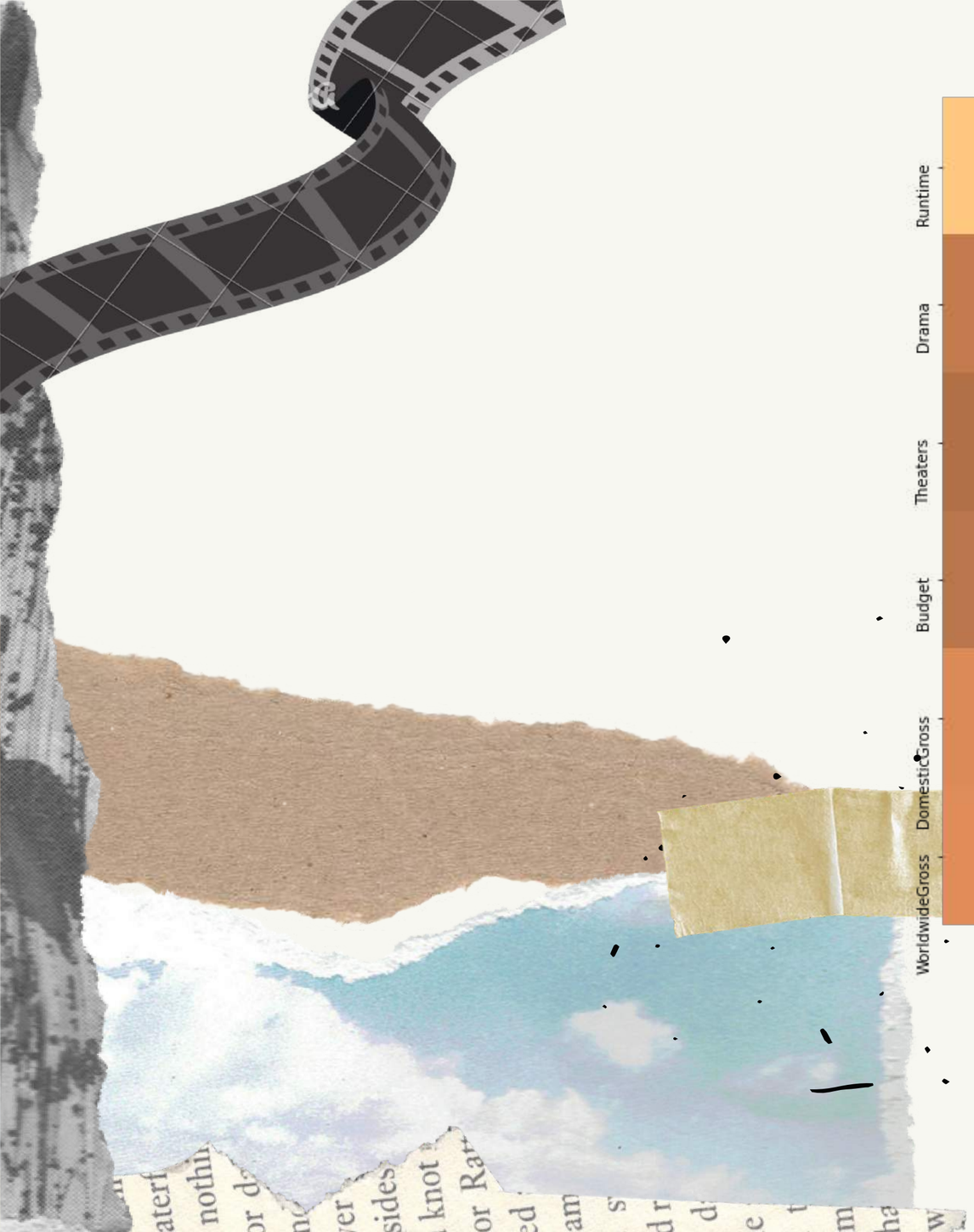


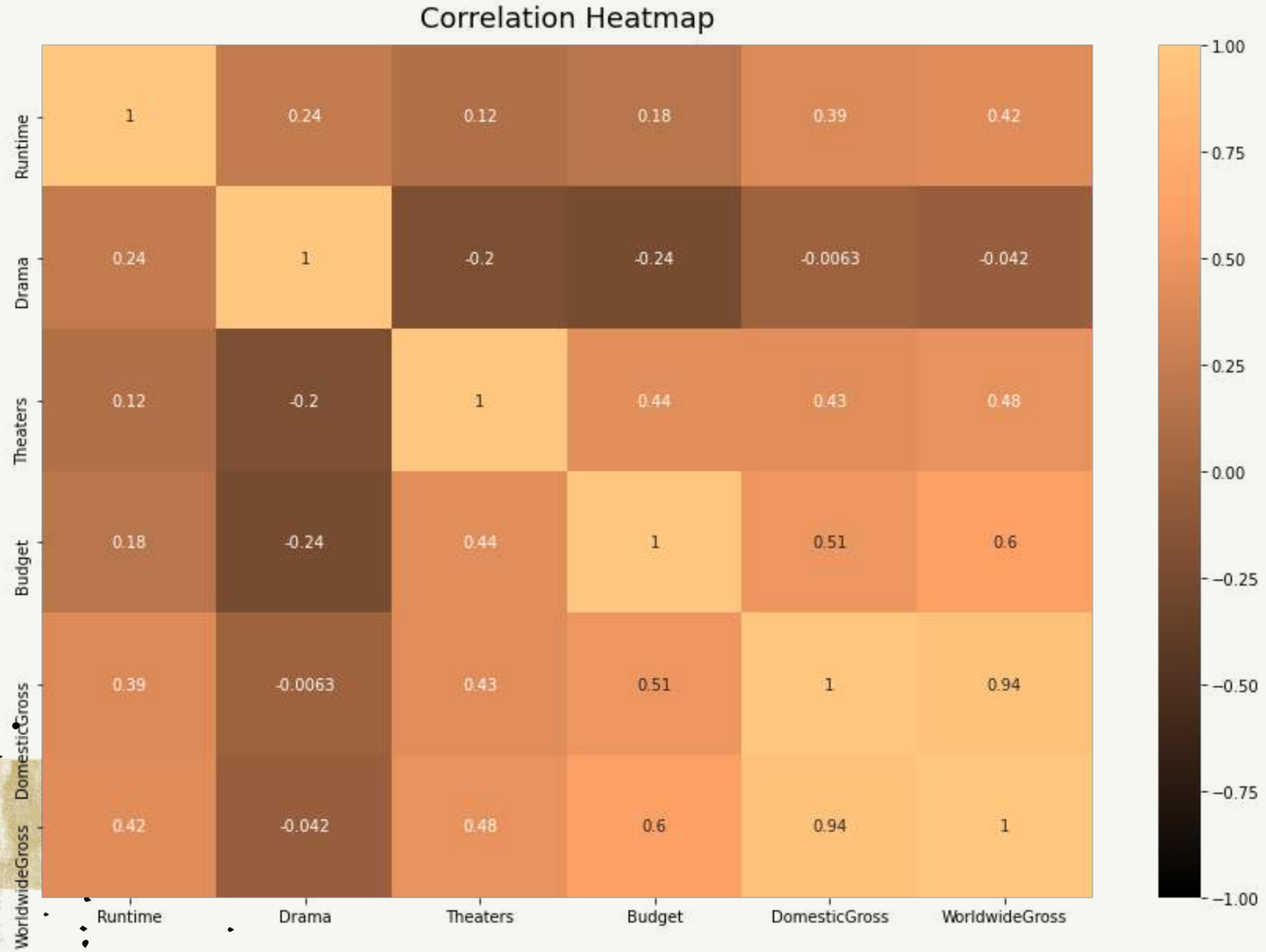
Data clean

1. Handling miss value
2. Remove duplicated value and nan value
3. Convert formatting for runtime ,date resident features

Features pair plot






Correlation Heatmap
Runtime
Drama
Theaters
Budget
DomesticGross
WorldwideGross
1
0.24
0.12
0.18
0.39
0.42
0.24
1
-0.2
-0.24
-0.0063
-0.042
0.12
-0.2
1
0.44
0.43
0.48
0.18
-0.24
0.44
1
0.51
0.6
0.39
-0.0063
0.43
0.51
1
0.94
0.42
-0.042
0.48
0.6
0.94
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Methodology

- split data train ,test, validation
- Liner regression
- LASSO Regularization
- Feature engineering

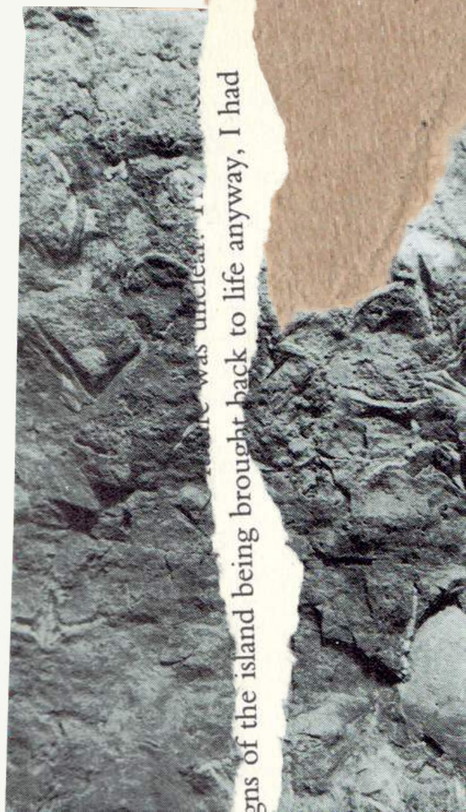
Create new variable "Primetime", indicating if a movie was released during a typically successful time of year (summer or holiday season).

Result

X= Rating, Runtime, IMDB_score, Budget, Theaters, Primetime.

Y= Worldwide Gross.

the metrics	Score	R^2
Linear Regression	0.54	0.31
Lasso	0.54	0.31

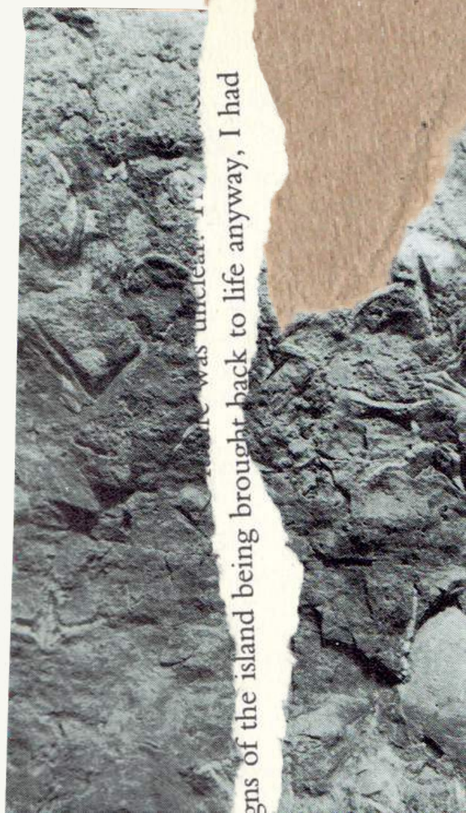


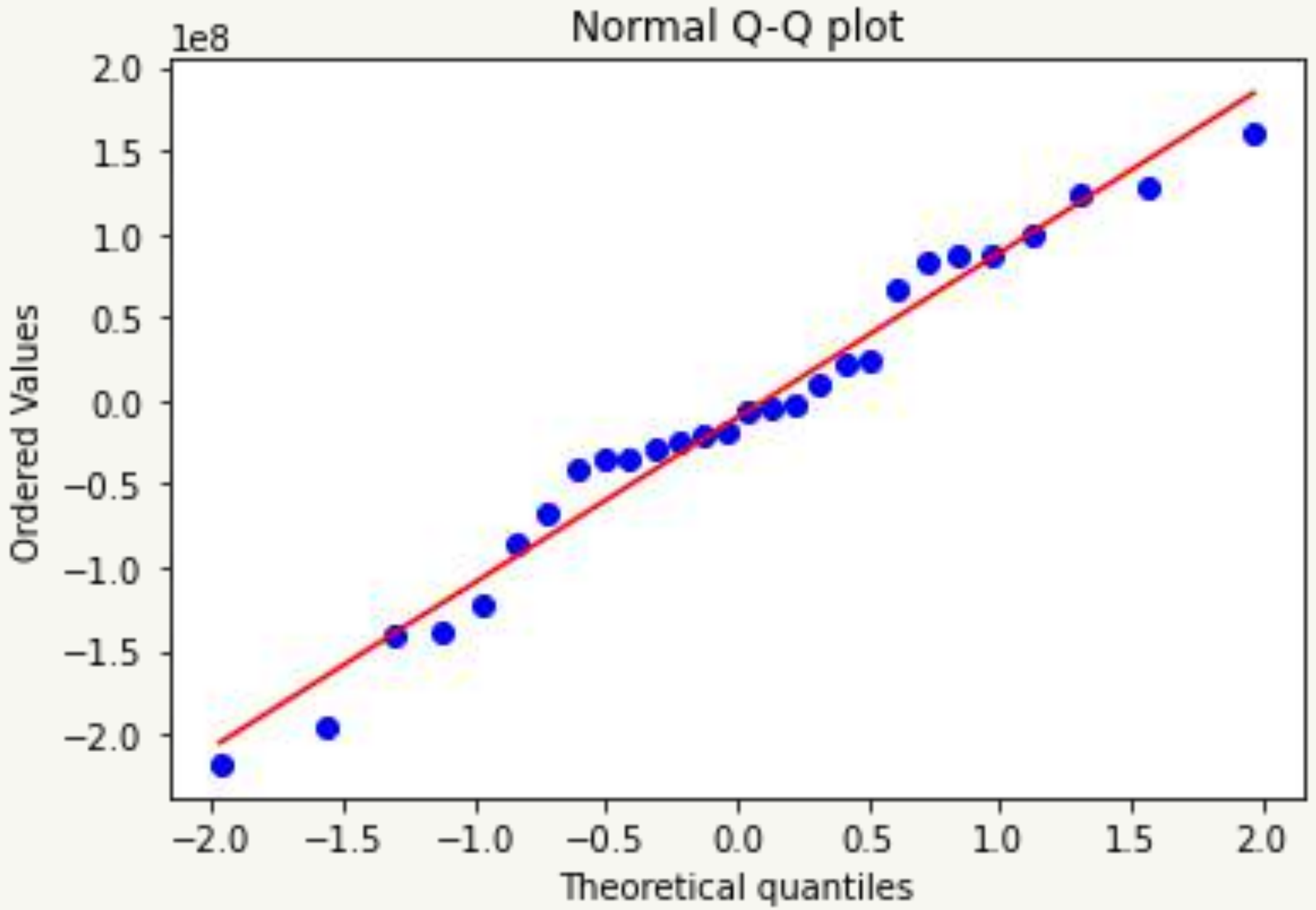
Result

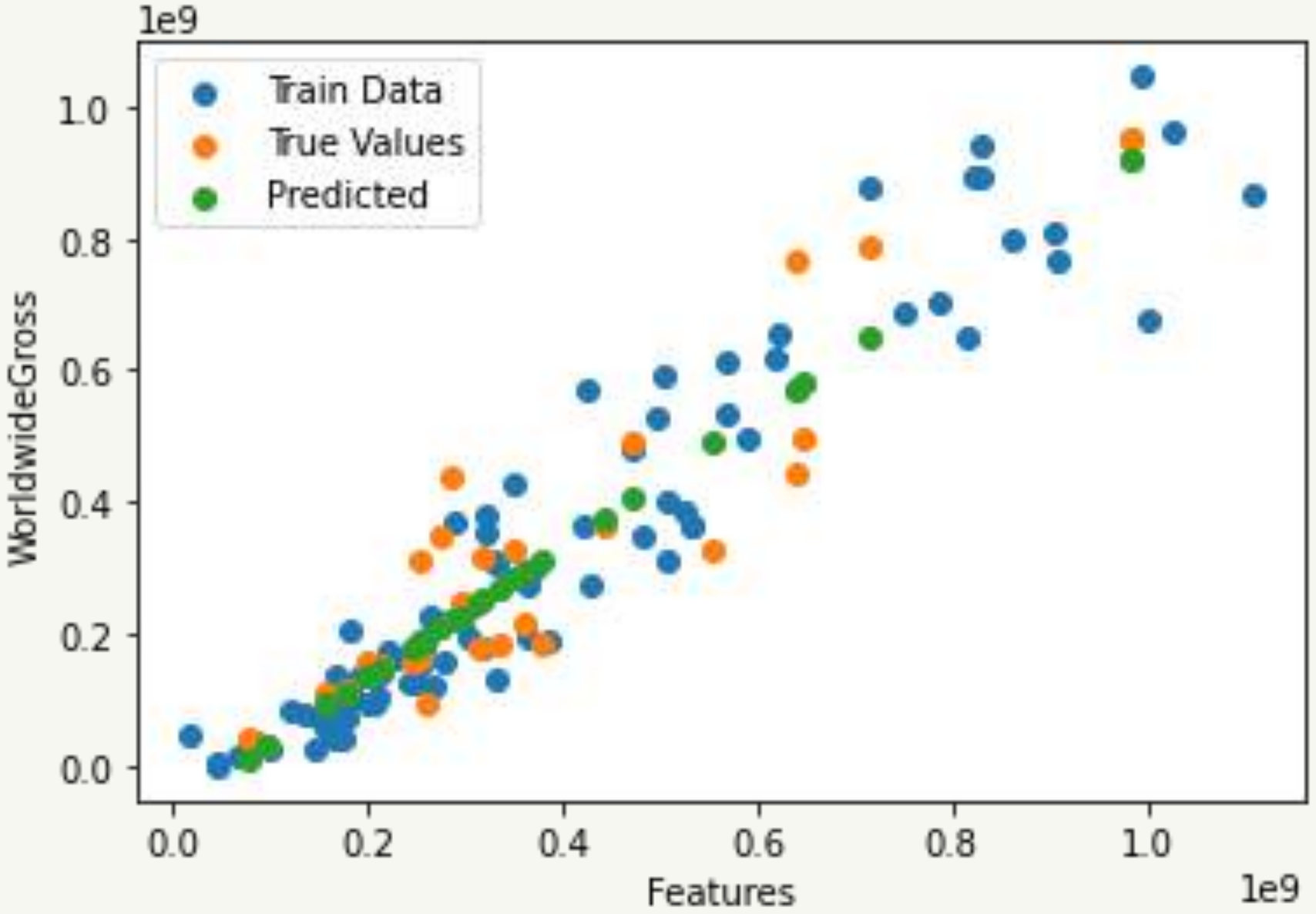
X= Budget, Primetime, Theaters, Domestic Gross.

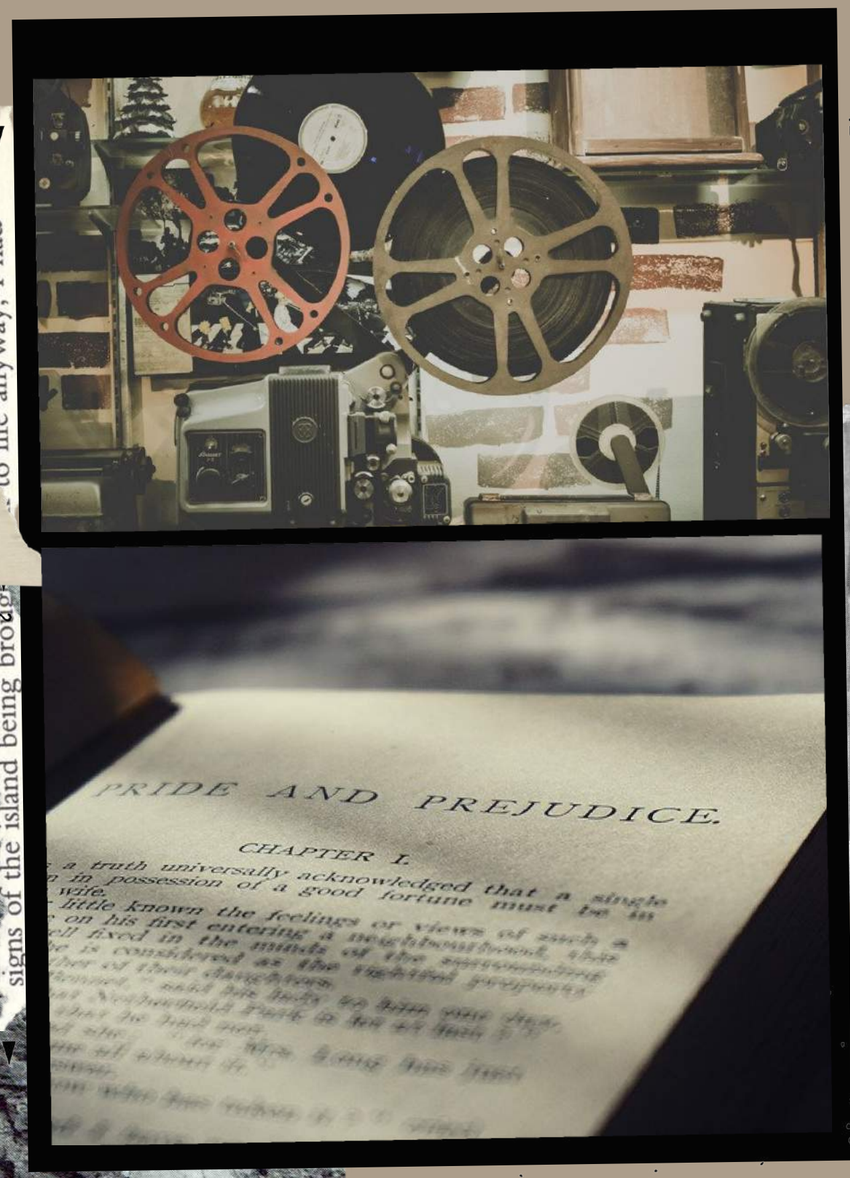
Y= Worldwide Gross.

the metrics	Score	R^2
Linear Regression	0.92	0.85
Lasso	0.92	0.85




Normal Q-Q plot
1e8
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Ordered Values
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Theoretical quantiles


1e9
Train Data
True Values
Predicted
1.0
0.8
0.6
0.4
0.2
0.0
WorldwideGross
0.0
0.2
0.4
0.6
0.8
1.0
Features
1e9



Future steps

We will improve our prediction model by add more new features

Define different movie’s category

THANK YOU

by:
Rana Alqahtani
Ohoud Albabtain
Khulud Amer