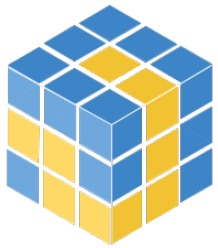
Revenue Predict Of movie

By Afnan Alsirhani and Mada abudash.

Objectives:

- Web **scraping** provides an effective way to extract large amounts of data quickly for further analysis.
- The goal of this project is to **takes data** from **IMDb**, **The Number** websites and uses **regression models** to predict **Worldwide Gross** of movies Based on features of an movies on these websites.

Tools:



NumPy

statsmodel

Pandas

BeautifulSoup

Methodology

Data retrieval

Data preparation

Feature selection

Model fitting and selection

Data retrieval

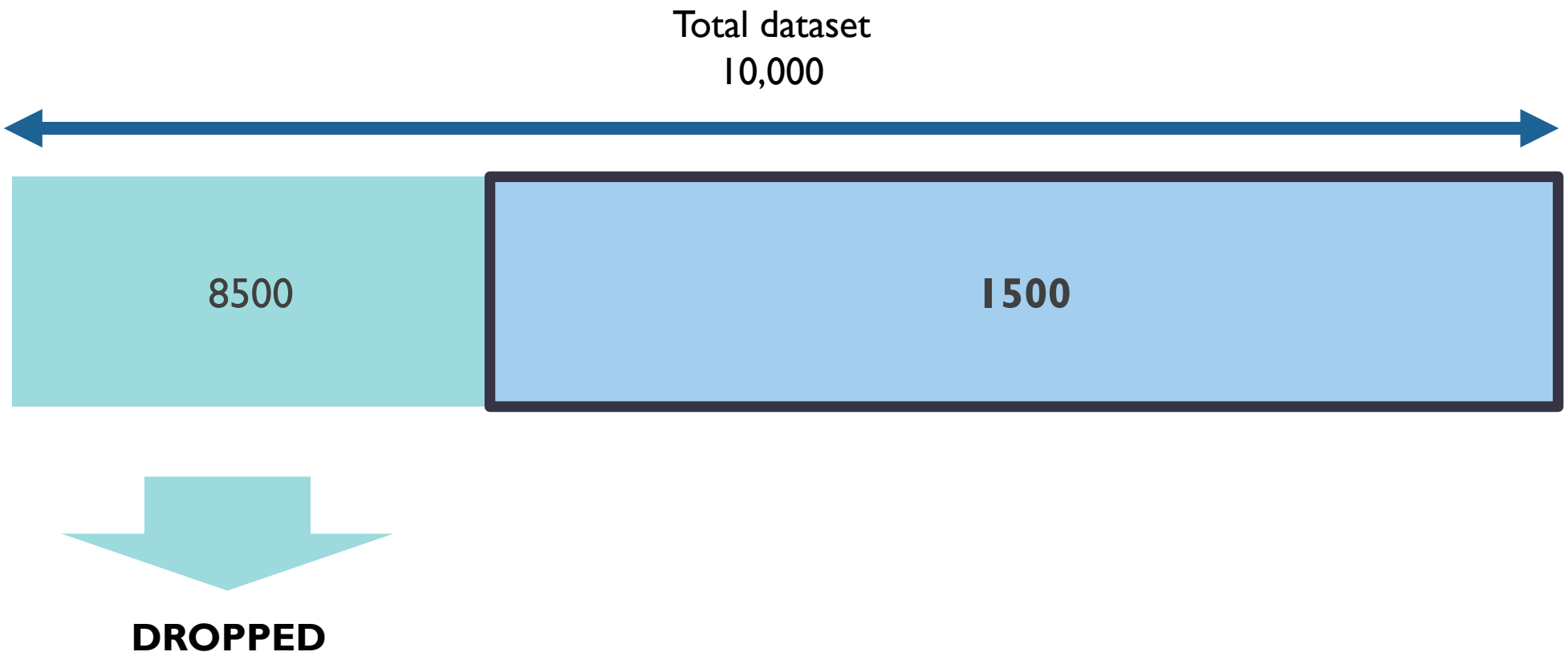
Obtained via 

web scraping

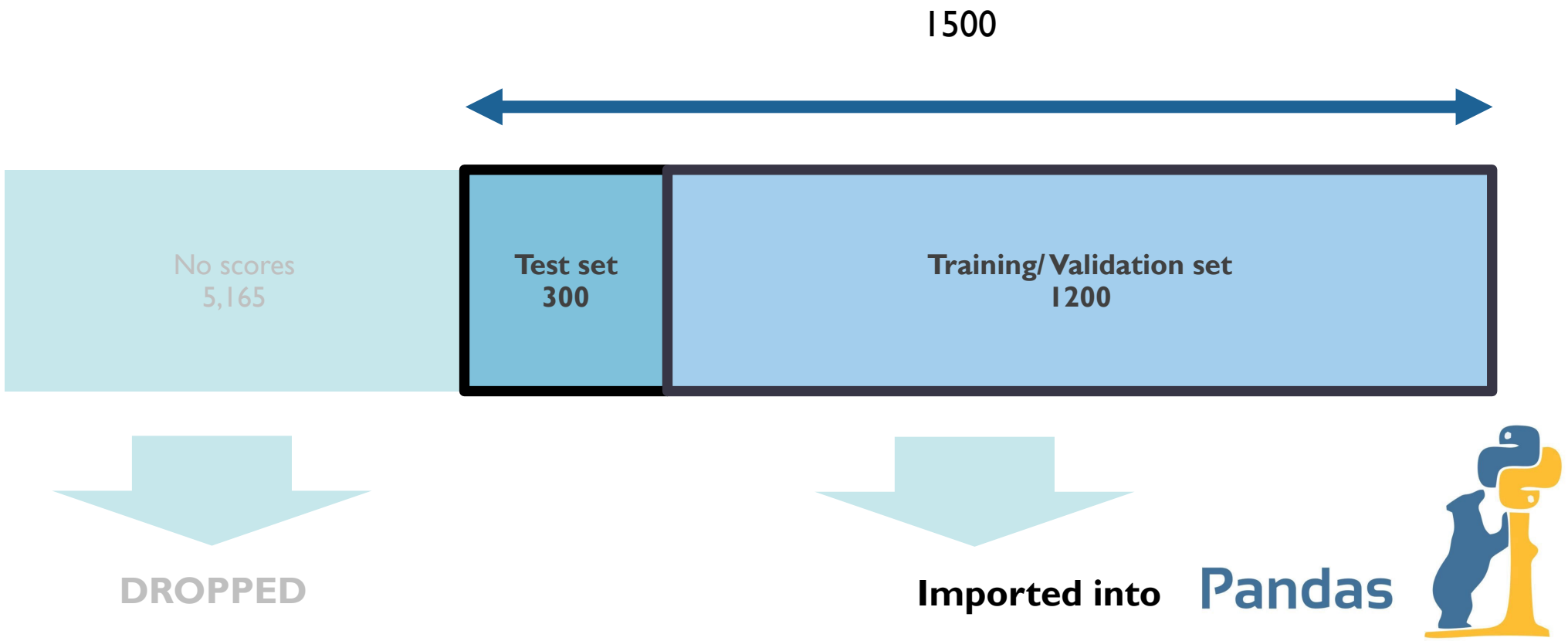
Total dataset
10,000

Retrieved in 335 batches of 50 between 23 and 24 April 2020

Data retrieval



Data retrieval



List of initial features

No	Feature	Description
1	Title	Name of Movie
2	Production Budget	is a financial plan that lists the number of units to be manufactured during a period.
3	Domestic Revenue	"Domestic" refers to gross box-office revenue from North America (U.S., Canada, and Puerto Rico), unless otherwise noted. "International" covers the rest of the world.
4	worldwide Revenue	The global Revenue
5	Genres	Type of movie
6	Runtime	Duration per minutes
7	MPAA Rating	Another type of movie
8	IMDb Rating	Score of the movie giving by IMDb
9	Actor Score	Number of stars participate in movie.
10	Release Data	is a fixed **date** on which a product is due to become available for the public to see or buy

Data preparation

Removed & handled nulls

Dropped all duplicated.

Dropped unneeded columns

Treated categorical data

Deal with Timeseries

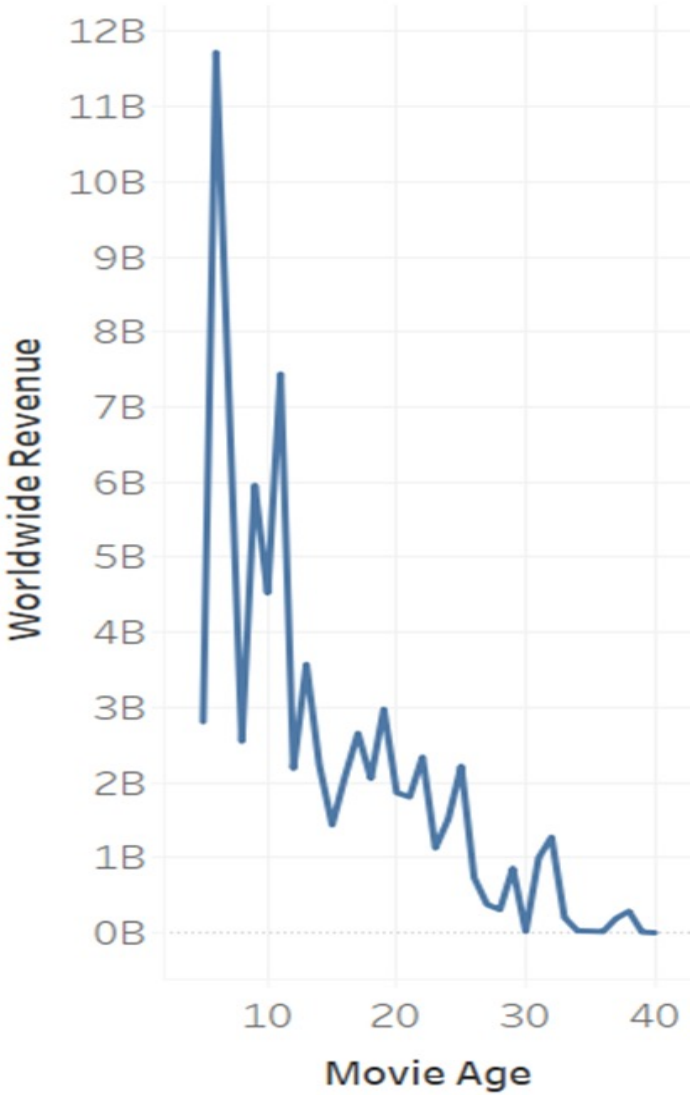
Visualised with Seaborn

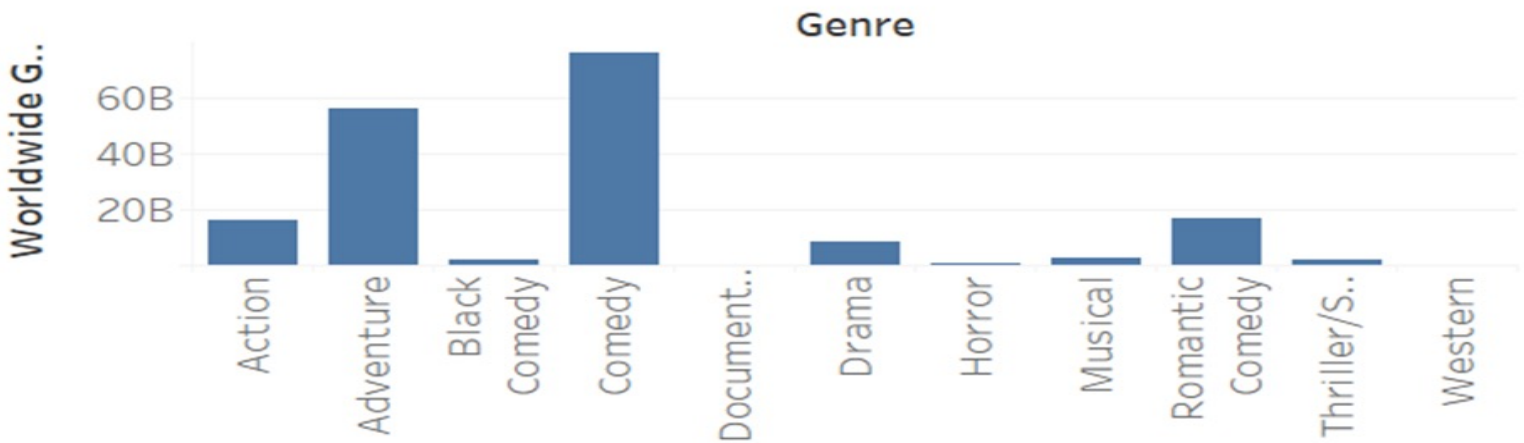
Worldwide Gross

1,577,585 48,045,556,983

distrubution of movies country and revenue

age of moveis by revenue





Features Engineering :

Genres → dummy variables	Consolidated into 18 total
Actor Score → values based on number of actors	From 0 to 4

Feature selection :

Runtime, Budget ,Domestic Revenue ,IMDB Rating, Actor Score , Action , Family, Adventure , Animation ,Biography, Comedy ,Crime, Drama ,Fantasy ,History, Horror ,Music, Mystery, Romance ,Sport ,Thriller, War ,Western ,movie age)

Model selection

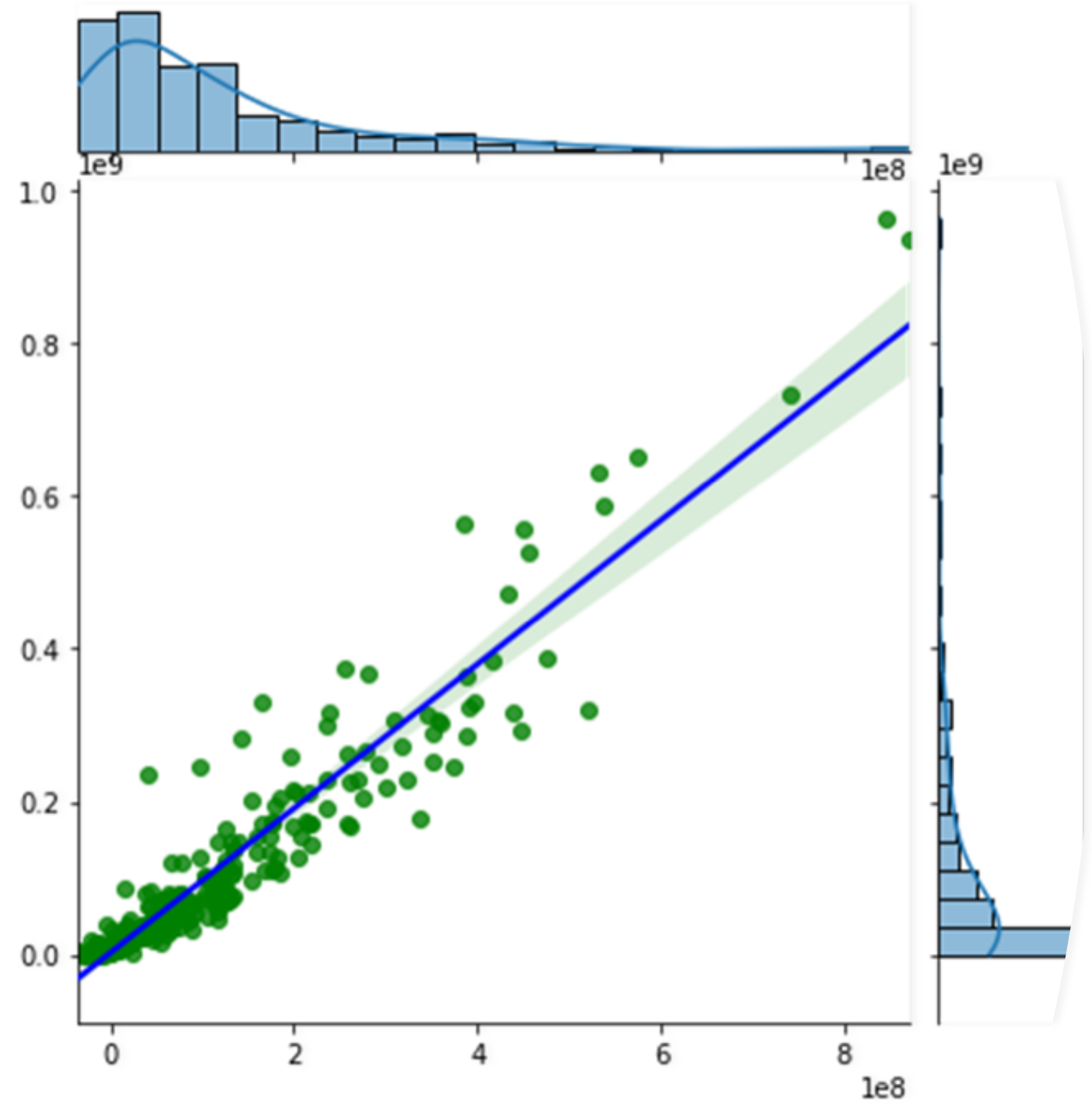
Models	Metrices
• Linear Regression • Random Forest Regressor Regression • K Neighbors Regression • Decision Tree • Polynomial Regression • fold cross validation performed on Linear Regression	• R^2 • **RMSE** • **SS_Residual** • **SS_Total**


Compare Regression Between Models
Degree of Metrics
80M
60M
40M
20M
0
RMSE
r_squared
Linear Regression with validation data
Linear Regression with test data
Random Forest Regression for val
Random Forest Regression for test data
KNN all columns for val
KNN all columns for test
Decision Tree Regressor all columns

Model selection

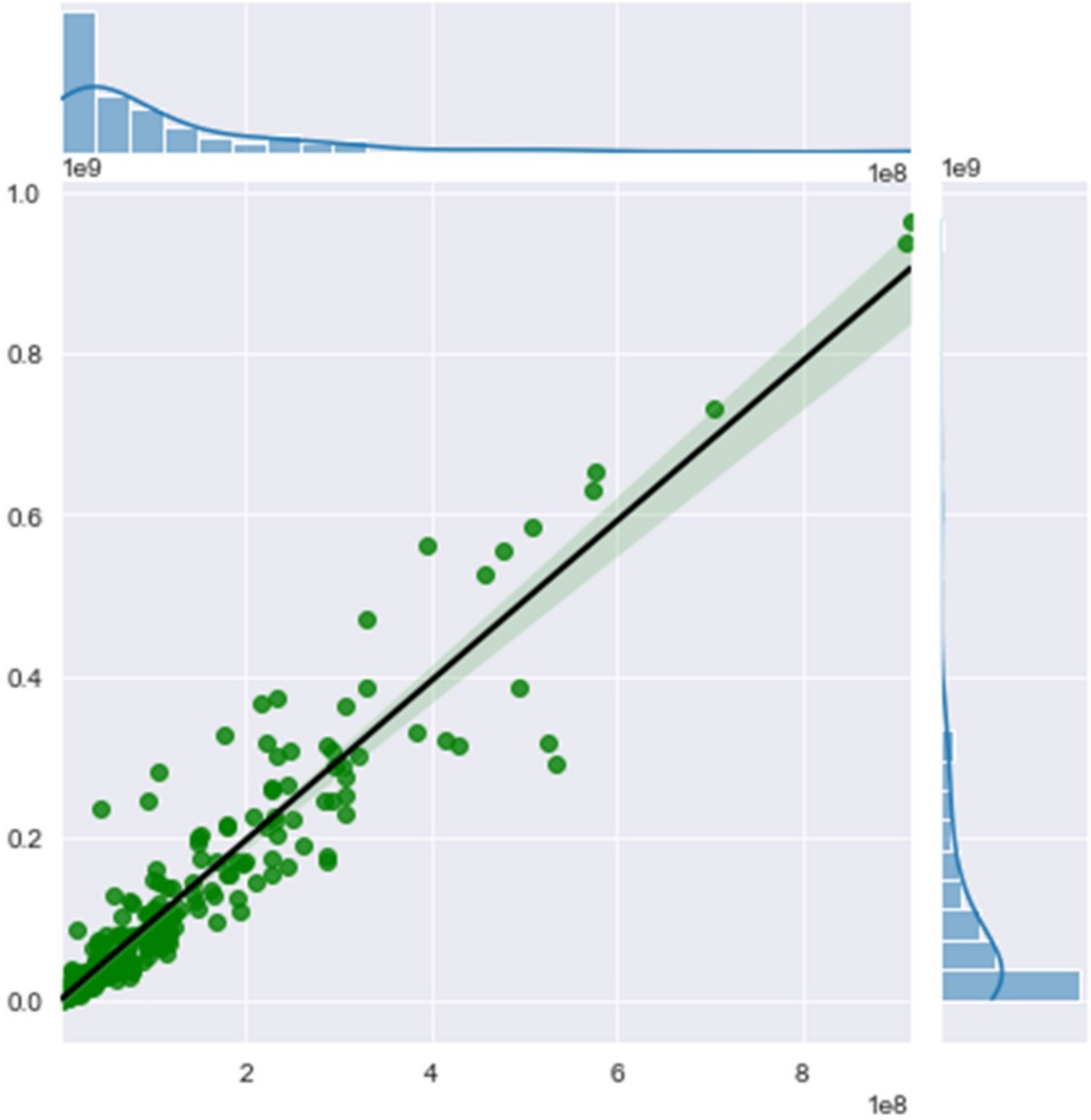
	Linear Regression	K Neighbors Regression
R2 for train data	0.8750633	0.89928
R^2 For validation	0.877486	0.875927
R^2 For Test	0.888242	0.891178
R2 for validation with polynomial	0.877486	0.870994
R2 for test with polynomial	0.888242	0.899107
RMSE For validation	5.324018e+07	5.357784e+07
RMSE For Test	5.901892e+07	5.823860e+07

K Neighbors is our better model



Linear Regression

- the predict in y and in the x axis is the test data , data using linear Regression.
- The line represent how the model fit in our data , as you can see the some of data blots are scattered and away from the line.



KNN Regression

- In the y axis is the predict and the x axis is the test data using KNN Regression.
- The line represent how the model fit in our data.
- KNN proved to be a better model for our data set as you can clearly see from the data blot which follows the line perfectly.

Future Work:

- More data for IMBD
- Group Categorical Features like Genre
- More sophisticated models need to be tested

Thank you for your attention

- Any questions?