

## Data Collection and Preprocessing Phase

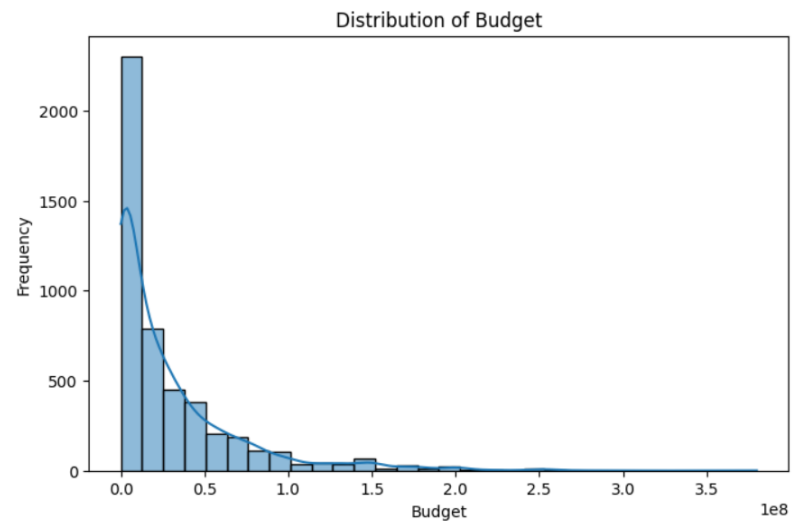
Date	23 September 2024
Team ID	LTVIP2024TMID24986
Project Title	Movie Box Office Gross Prediction using Machine Learning
Maximum Marks	6 Marks

### Data Exploration and Preprocessing Report

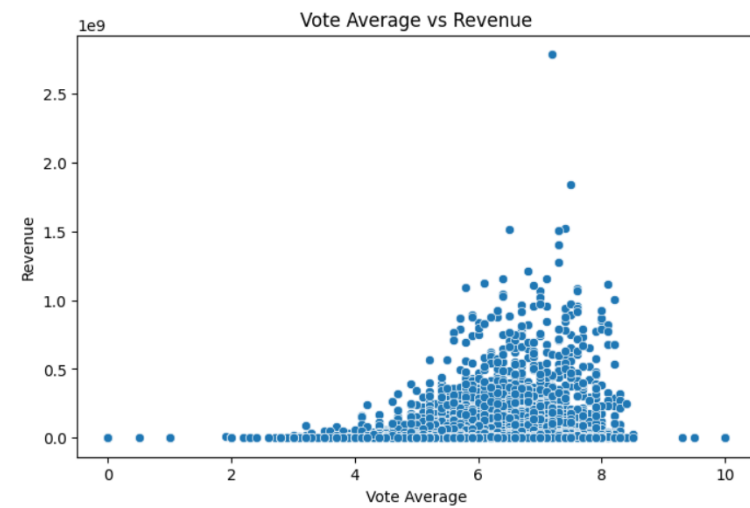
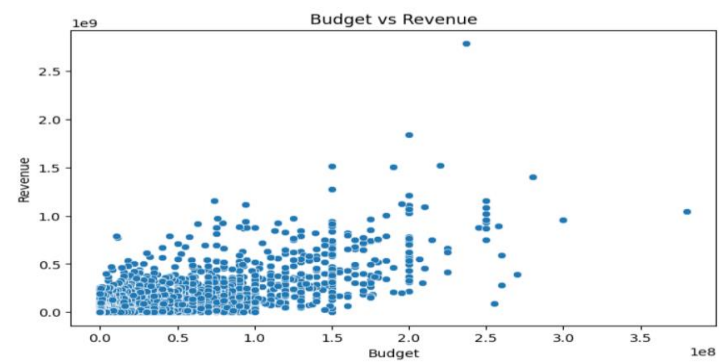
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	<u>Dimension:</u> 4083rows × 23columns
	<u>Descriptive statistics:</u>

## Univariate Analysis



## Bivariate Analysis



## Multivariate Analysis



## Outliers and Anomalies

-

## Data Preprocessing Code Screenshots

### Loading Data

```
credits=pd.read_csv("/content/tmdb_5000_credits.csv")
movies_df=pd.read_csv("/content/tmdb_5000_movies.csv")

credits.head()
```

	movie_id	title	cast	crew
0	19995	Avatar	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...
2	206647	Spectre	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "54805967c3a36829b5002c41", "de...
3	49026	The Dark Knight Rises	[{"cast_id": 2, "character": "Bruce Wayne / Ba...	[{"credit_id": "52fe4781c3a36847f81398c3", "de...
4	49529	John Carter	[{"cast_id": 5, "character": "John Carter", "c...	[{"credit_id": "52fe478ac3a36847f813eaa3", "de...

### Handling Missing Data

```
from sklearn.preprocessing import LabelEncoder
from collections import Counter as c
cat=['director','genres']
for i in movies_box[cat]:
    print("LABEL ENCODING OF:",i)
    LE = LabelEncoder()
    print(c(movies_box[i]))
    movies_box[i] = LE.fit_transform(movies_box[i])
    print(c(movies_box[i]))
```

Data Transformation	<pre> movies['log_revenue'] = np.log1p(movies['revenue']) movies['log_budget'] = np.log1p(movies['budget'])  movies_box = movies.drop(['homepage', 'id', 'keywords', 'original_language', 'original_title',                           'overview', 'production_countries', 'release_date', 'spoken_languages',                           'status', 'tagline', 'title_x', 'title_y', 'cast',                           'log_revenue', 'log_budget'], axis = 1)  movies_box=movies_box.drop(['production_companies'],axis=1)  movies_box.isnull().sum() </pre> <div> <a href="#">+ Code</a> <a href="#">+ Text</a> </div>
Feature Engineering	Attached the codes in final submission.
Save Processed Data	-