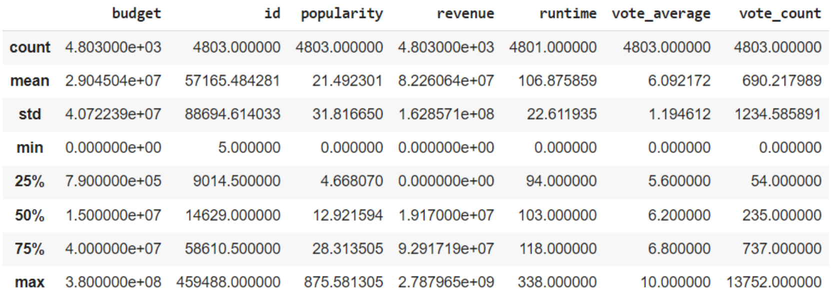
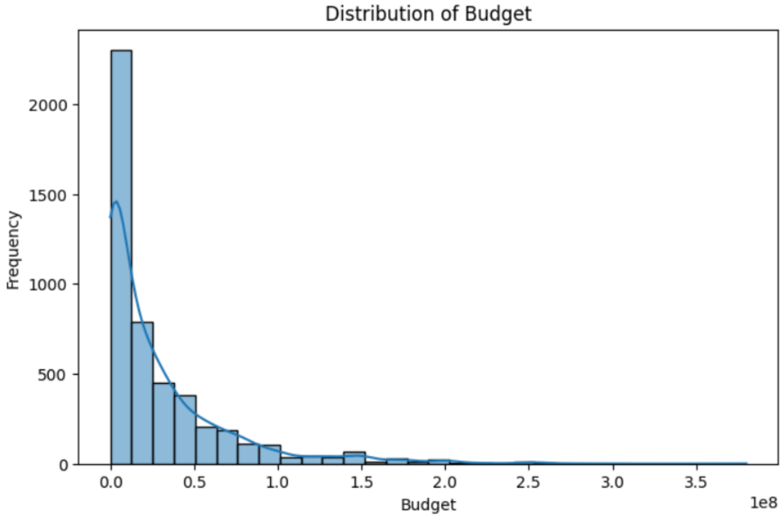
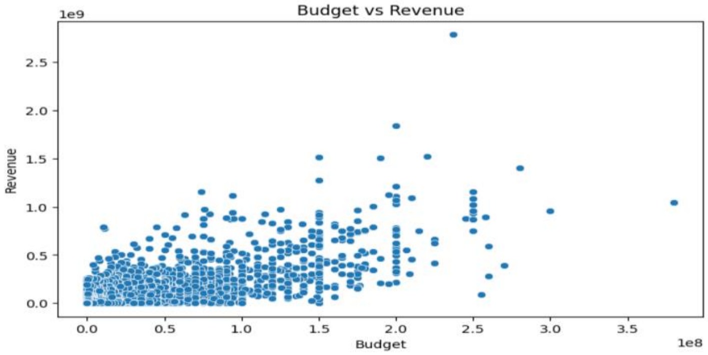
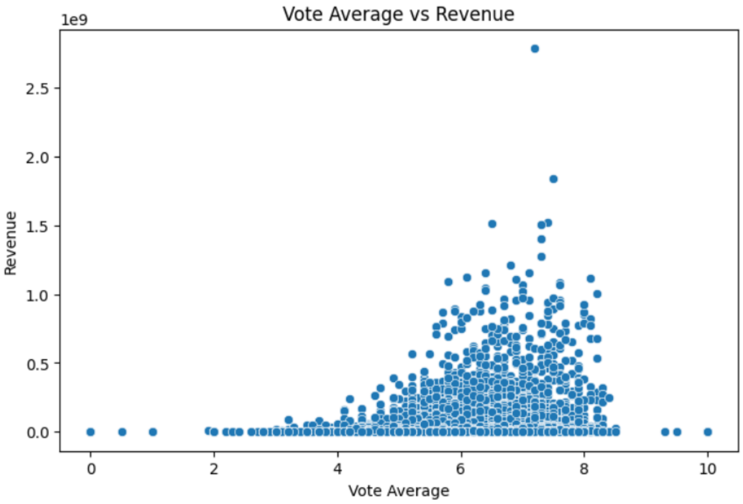
# Data Collection and Preprocessing Phase

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

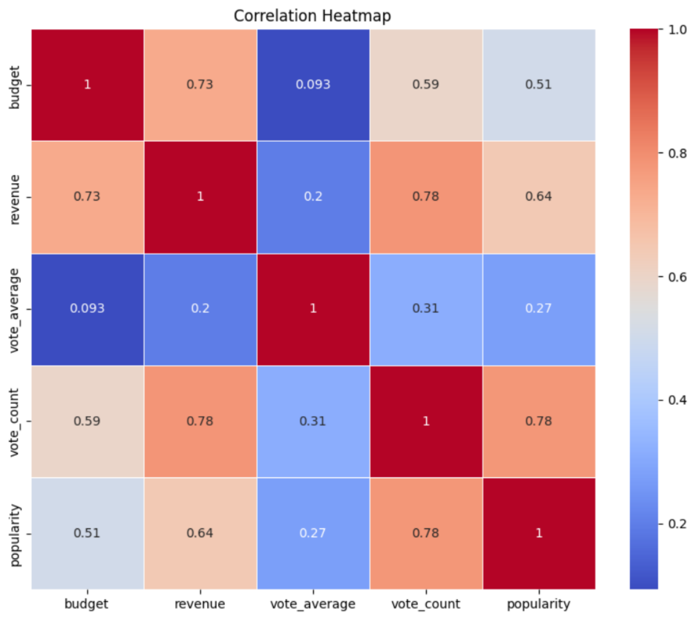
## Data Exploration and Preprocessing Report

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.

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>Dimension:<br>4083rows × 23columns<br>Descriptive statistics:<br>
<table>
<tr><th></th><th>budget</th><th>id</th><th>popularity</th><th>revenue</th><th>runtime</th><th>vote_average</th><th>vote_count</th></tr>
<tr><td>count</td><td>4.803000e+03</td><td>4803.000000</td><td>4803.000000</td><td>4.803000e+03</td><td>4801.000000</td><td>4803.000000</td><td>4803.000000</td></tr>
<tr><td>mean</td><td>2.904504e+07</td><td>57165.484281</td><td>21.492301</td><td>8.226064e+07</td><td>106.875859</td><td>6.092172</td><td>690.217989</td></tr>
<tr><td>std</td><td>4.072239e+07</td><td>88694.614033</td><td>31.816650</td><td>1.628571e+08</td><td>22.611935</td><td>1.194612</td><td>1234.585891</td></tr>
<tr><td>min</td><td>0.000000e+00</td><td>5.000000</td><td>0.000000</td><td>0.000000e+00</td><td>0.000000</td><td>0.000000</td><td>0.000000</td></tr>
<tr><td>25%</td><td>7.900000e+05</td><td>9014.500000</td><td>4.668070</td><td>0.000000e+00</td><td>94.000000</td><td>5.600000</td><td>54.000000</td></tr>
<tr><td>50%</td><td>1.500000e+07</td><td>14629.000000</td><td>12.921594</td><td>1.917000e+07</td><td>103.000000</td><td>6.200000</td><td>235.000000</td></tr>
<tr><td>75%</td><td>4.000000e+07</td><td>58610.500000</td><td>28.313505</td><td>9.291719e+07</td><td>118.000000</td><td>6.800000</td><td>737.000000</td></tr>
<tr><td>max</td><td>3.800000e+08</td><td>459488.000000</td><td>875.581305</td><td>2.787965e+09</td><td>338.000000</td><td>10.000000</td><td>13752.000000</td></tr>
</table>
</td></tr>
</table>

| | |
|---|---|
| Univariate Analysis |  |
| Bivariate Analysis |  <br> |

Multivariate Analysis

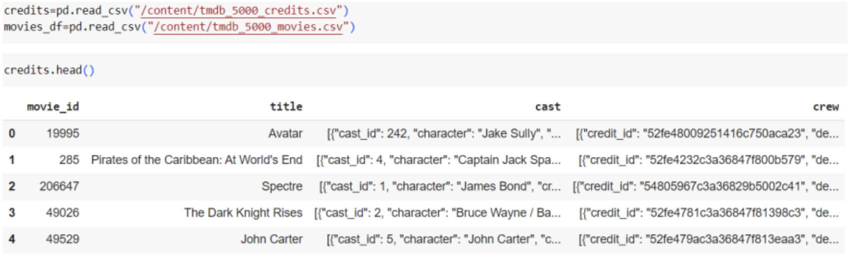


Outliers and Anomalies

-

**Data Preprocessing Code Screenshots**

Loading Data

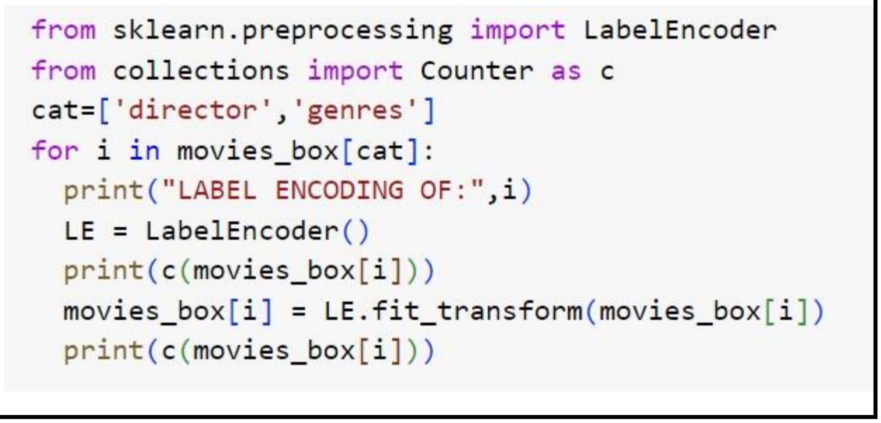


```
credits=pd.read_csv("/content/tmdb_5000_credits.csv")
movies_df=pd.read_csv("/content/tmdb_5000_movies.csv")
```

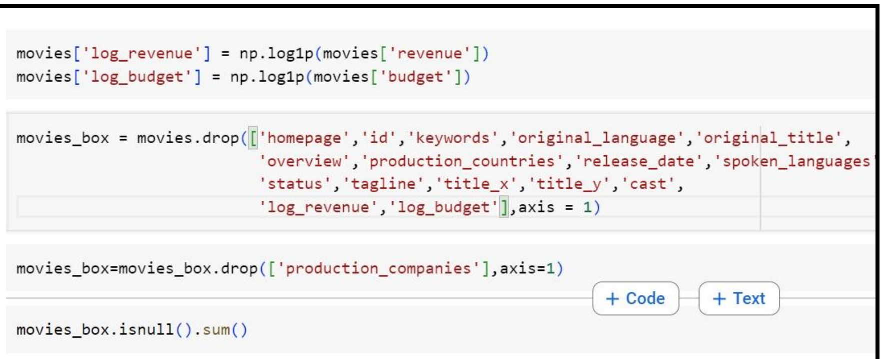
```
credits.head()
```

```
   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": "52fe479ac3a36847f813eaa3", "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]))
```

<table>
<tr><td>Data Transformation</td><td>


<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>

</td></tr>
<tr><td>Feature Engineering</td><td>Attached the codes in final submission.</td></tr>
<tr><td>Save Processed Data</td><td>-</td></tr>
</table>