TMDB Movie Data Analysis

Introduction

This data set contains information about 10,866 movies collected from The Movie Database (TMDB) between 1960 to 2015.

Steps 1:
Questions that I could direct from the given datasedt.

1. What genres are more popular overall?
2. What genres are more popular throughout the decade?
3. What are the properties associated with higher revenue?
4. Which actors have starred in most movies?
5. Who has directed the most movies?
6. What are the most popular movies?

Step 2:
Data Wangling, the loading of the dataset, modifying the data, data cleaning, Removing outliers, removing duplicates data.

Step 3:
Data Exploration; finding patterns and creating better features for exploration.

Importation of libraries to be used

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import cm
import seaborn as sns
```

Data Wrangling

Loading the dataset

```
tmdb = pd.read_csv('tmdb dataset.csv')
print(f"Number of Observations in tmdb dataset: {tmdb.shape}")
print(tmdb)
```

The code above loads the data into a dataframe, to assess the quality of the data. From the dataframe, the properties of the data and descriptive statistics was generated.

Data Cleaning

Modification of the data set, removal of extraneous data and duplicates, and adding new information to the data.

I dropped extraneous columns and duplicates that aren't relevant with the analysis as shown below;

```
tmdb.drop(['imdb_id', 'homepage', 'tagline', 'overview', 'runtime', 'budget_adj', 'revenue_adj'], axis=1,
inplace=True)
print(tmdb.head())
```

```
tmdb.drop_duplicates(inplace=True)
print(sum(tmdb.duplicated()))
```

I modified released_date and more information was derived from it giving us the date, month and year. Thereafter released_date was dropped from the dataframe.

Addition of new columns was performed, as listed below:

Date

Month

Year

Profit

Profitable ratio

Revenue rating

Decade

Also in the data, cast, genres and Director are separated by '|' character. Split function was used to separate each values.
The code as shown below:

```
# create separate dataframes for each: genres, cast, and director.
tmdb['genres'].str.contains('|')
tmdb['genres'].nunique()

split_genre = tmdb_split_genre['genres'].str.split('|').apply(pd.Series,
1).stack().reset_index(level=1, drop=True)
split_genre.name = 'genre_split'
tmdb_split_genre = tmdb_split_genre.drop(['genres'], axis=1).join(split_genre)
print(tmdb_split_genre)
```

Repeat prodecure for cast and director. Then check out for dupilicate using the code below:

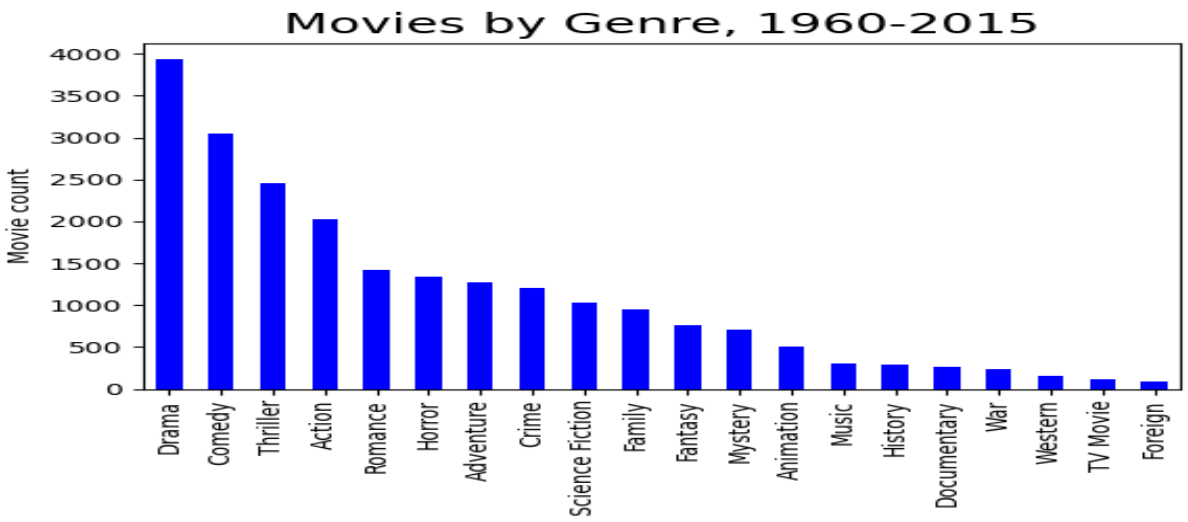
```
# check for duplicates and view the info for the new dataset.
print(tmdb_split_genre.info())
print(tmdb_split_genre.shape)
print(sum(tmdb_split_genre.duplicated()))
```

Exploratory Analysis

1 . What genres are most popular overall?

```
tmdb_split_genre['genre_split'].value_counts().plot(kind='bar', color='blue')
plt.title('Movies by Genre, 1960-2015', size=18)
plt.xlabel('Genre', size=12)
plt.ylabel('Movie count', size=12)
plt.show()
```

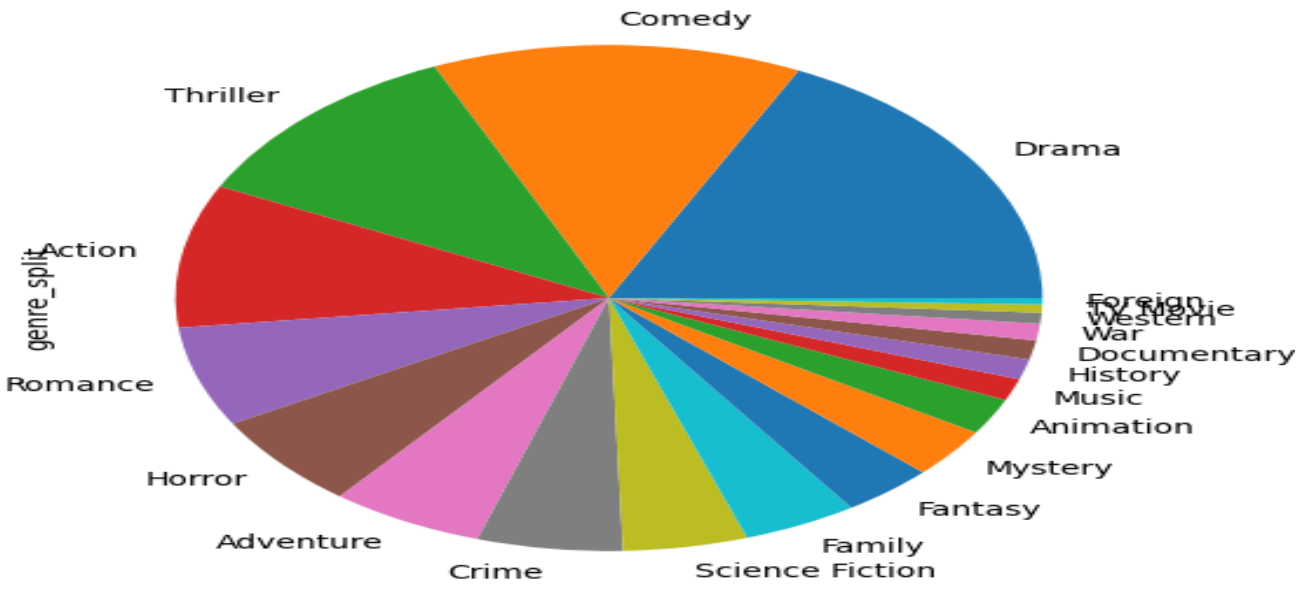
Output:



Piechart

```
tmdb_split_genre['genre_split'].value_counts().plot(kind='pie', figsize=(6, 6))
plt.show()
```

Output:



The Charts above showed Drama, Comedy, Thriller, and Action are the most popular genres in general. The pie chart is a better visual since we can assess that these top 4 genres make up about 50% of all movies. Foreign, TV Movies and Westerns are the least popular genres.

2. What genres are most popular throughout the decades?

```
genres_decades = tmdb_split_genre.groupby(['decades'])['genre_split'].value_counts()
genres_decades_largest = genres_decades.groupby(level=0).nlargest(3).reset_index(level=0, drop=True)
print(genres_decades_largest)
```

Output:

```
decades              genre_split
sixties              Drama          178
                     Comedy         114
                     Action          84
seventies            Drama          245
                     Thriller       166
                     Action         128
eighties             Comedy         405
                     Drama          402
                     Action         261
nineties             Drama          811
                     Comedy         693
                     Thriller       472
two_thousands        Drama         1381
                     Comedy        1111
                     Thriller       869
two_thousand_tens    Drama          917
                     Thriller       629
                     Comedy         610
```

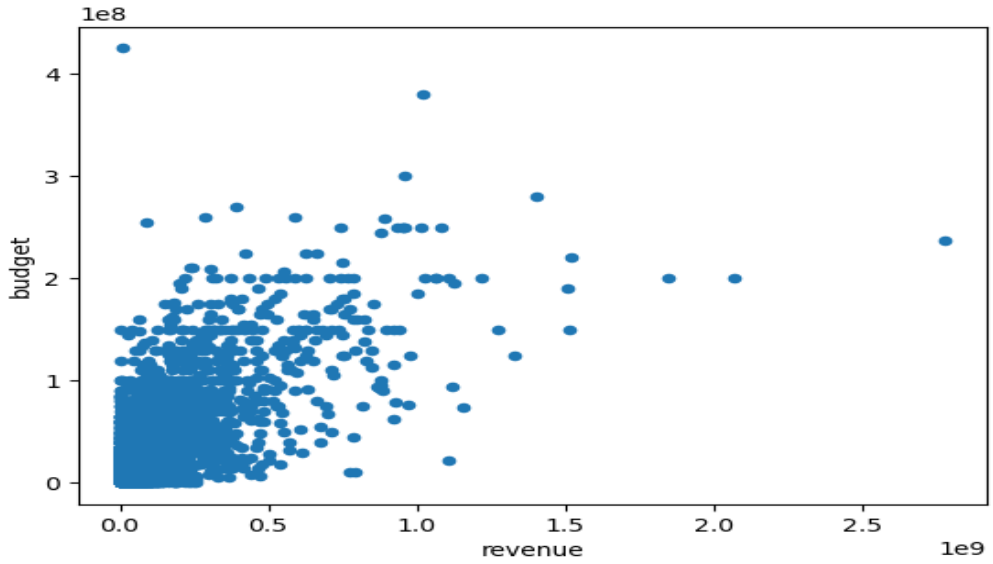
The figure above showed the top 3 most popular genre throughout the decade, Drama is the most popular genre throughout in each decades except in the eighties where comedy was the most popular.

3. What properties are associated with higher revenues?

General scatter plots of revenue vs budget, profit, profitability_ratio and popularity.

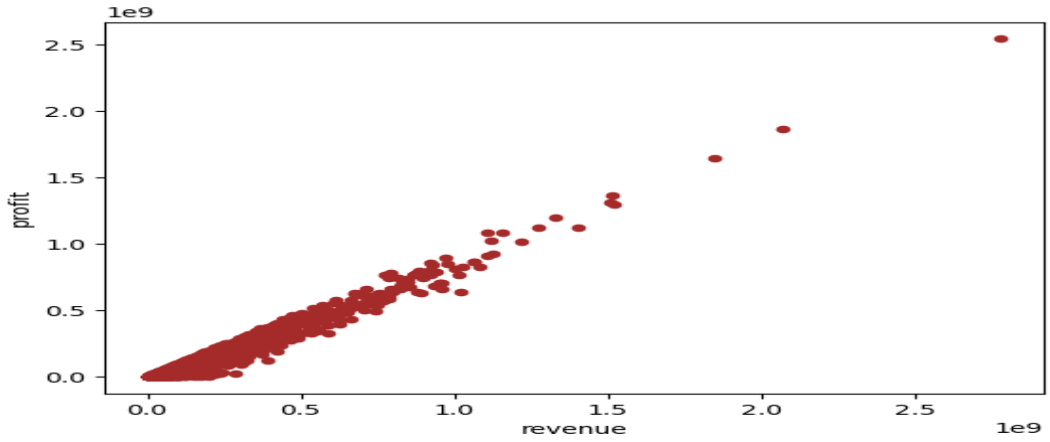
```
# Revenue vs Budget
tmdb.plot(x='revenue', y='budget', kind='scatter')
plt.show()
```

Output



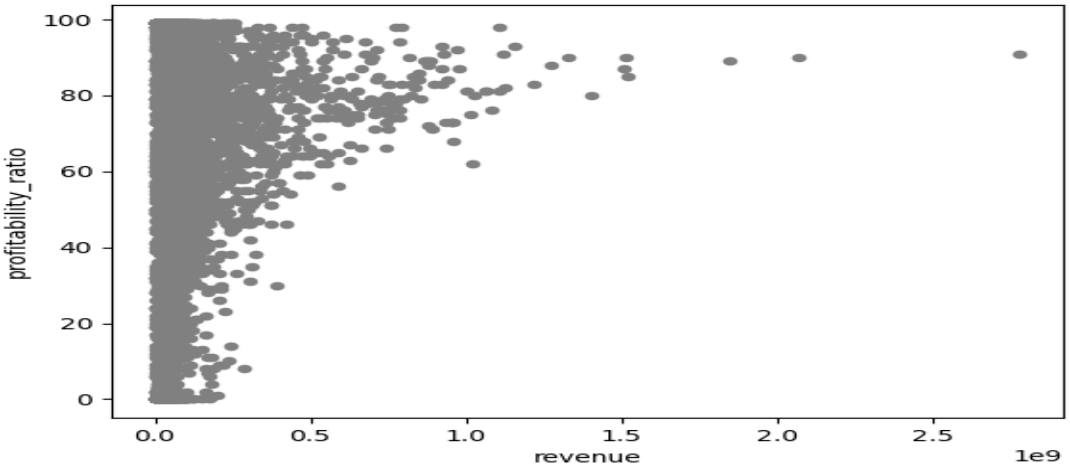

```
tmdb.plot(x='revenue', y='profit', kind='scatter', color='brown')
plt.show()
```

Output



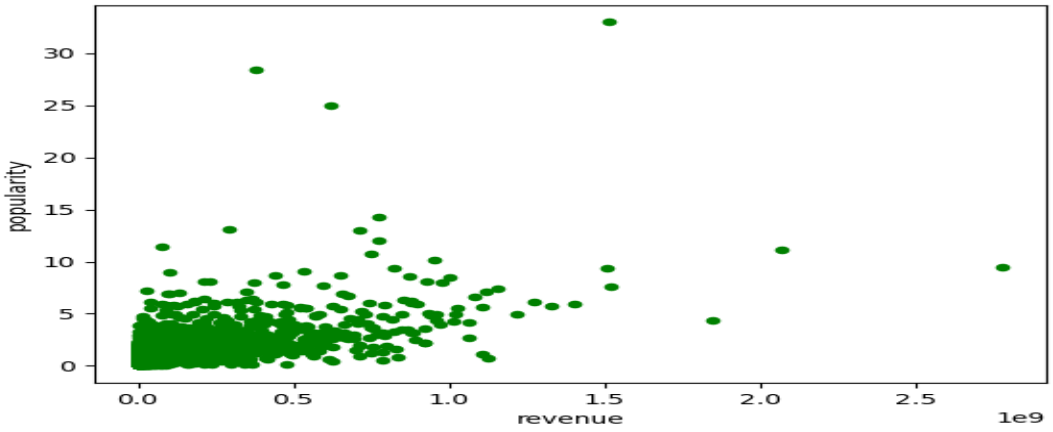
```
# Revenue vs profitability ratio
tmdb.plot(x='revenue', y='profitability_ratio', kind='scatter', color='grey')
plt.show()
```

Output:




```
# Revenue vs popularity
tmdb.plot(x='revenue', y='popularity', kind='scatter', color='green')
plt.show()
```

Output



The figures above shows:

Revenue and budget have a weak positive correlation.

Revenue and profit have a strong positive correlation.

Revenue and profitability ratio have a weak positive correlation.

Revenue and popularity have positive correlation, movies with higher revenues tend to be more popular.

4. Which actors have starred in the most movies?

```
cast = tmdb_split_cast['cast_split'].value_counts().head(20)
print(cast)
```

Output:

```
Name: genre_split, dtype: int64
Robert De Niro          68
Samuel L. Jackson       65
Bruce Willis            61
Nicolas Cage            59
Michael Caine           50
Robin Williams          48
Morgan Freeman          47
John Goodman            47
Tom Hanks               46
John Cusack             46
Alec Baldwin            45
Julianne Moore          44
Liam Neeson             44
Susan Sarandon          43
Johnny Depp             43
Dennis Quaid            43
Gene Hackman            43
Clint Eastwood          43
Nicole Kidman           42
Willem Dafoe            42
```

From the data set, the figure above shows the first 20 actors that have starred in most movies, Robert De Niro is the most starred actor with 68 movies.

5. Who has directed the most movies?

```
director = tmdb_split_director['director_split'].value_counts().head(20)
print(director)
```

output

```
Woody Allen             41
Clint Eastwood          33
Steven Spielberg        30
Martin Scorsese         27
Ridley Scott            23
Steven Soderbergh       22
Ron Howard              22
Joel Schumacher         20
John Carpenter          19
Brian De Palma          19
David Cronenberg        19
Tim Burton              19
Barry Levinson          19
Robert Rodriguez        18
Wes Craven              18
Mike Nichols            18
Roman Polanski          17
Francis Ford Coppola    17
Renny Harlin            17
Oliver Stone            17
```

Woody Allen directed 41 movies, the most from the dataset.

6. What are the most popular movies?

```
popular_movies = tmdb[['popularity', 'original_title']].sort_values(by='popularity', ascending=False).head(10)
print(popular_movies)
```

Output:

```
      popularity                          original_title
0     32.99          Jurassic World
1     28.42          Mad Max: Fury Road
629   24.95          Interstellar
630   14.31          Guardians of the Galaxy
2     13.11          Insurgent
631   12.97          Captain America: The Winter Soldier
1329  12.04          Star Wars
632   11.42          John Wick
3     11.17          Star Wars: The Force Awakens
633   10.74          The Hunger Games: Mockingjay - Part 1
```

The data above shows the the top 10 popular movies, Jurrassic World with 33.99% is the most popular movie.