Project - 2

**Uncovering surprising Facts from World Wide Movie Database using Data Cleaning & Data Visualization**

**OVERVIEW**

A project to overlook at the movie’s database and interpret various finding using Data cleaning, Data wrangling and Data Visualization

**Software Requirements**

1. Programming Language: Python

2. Environment: Jupyter Notebooks / Google Collab

3. Database: CSV (export type)

4. Operation System: Windows XP or above

5. Libraries Used: Pandas, Folium, Seaborn, Scikit, SKLEARN, Wordcount

6.Datasets used: TMDB Dataset

1. **Open a New Notebook and import the required libraires and read the csv file**

|  |  |
| --- | --- |
|  | import numpy as np  import pandas as pd  pd.set\_option('max\_columns', None)  import matplotlib.pyplot as plt  import seaborn as sns  plt.style.use('ggplot')  import datetime  from wordcloud import WordCloud |

**Description:**

First, we need to import the libraries that are useful. Matplotlib is a plotting library in python language and its numerical mathematics extension is NumPy. It helps in plotting points. We use seaborn for representing graphs.

1. **Loading the training & testing Dataset**

data = pd.read\_csv('/content/data.csv')

**Description:**

Here we load the data using pandas pd.read\_csv function. It extracts the data from csv files.

1. **Visualizing the Distribution of Revenue with & without Log**

|  |  |
| --- | --- |
| fig, axis = plt.subplots(figsize=(16,8))  plt.subplot(1, 2, 1)  sns.distplot(data['revenue'], kde=False);  plt.title('Distrinution of revenue');  plt.subplot(1, 2, 2)  sns.distplot(np.log1p(data['revenue']), kde=False);  plt.title('Distribution of log revenue')  Output:  **Description:**  The film industry uses **box office revenue**, or, the amount of money generated through movie ticket sales, to measure the commercial success of the films it produces. Ticket sales also account for a sizable portion of the film industry's total revenues. In the above piece of code, we find the distribution of revenue. We use distributed plot to represent the data. Seaborn provides a high-level interface for drawing attractive and informative statistical graphics. In the first distribution we find some error in representing the data, so we use logarithmic scale using NumPy. It resolves an issue with visualizations that skew towards large values in a dataset. Hence, we get some understanding from second plot i.e., more the movies more revenue. This is called Data Wrangling.   1. **Finding the Relationship between Movie Revenue & Budget**   data['log\_revenue'] = np.log1p(data['revenue'])  data['log\_budget'] = np.log1p(data['budget'])  plt.figure(figsize=(16, 8))  plt.subplot(1, 2, 1)  sns.scatterplot(data['budget'], data['revenue'])  plt.title('Revenue vs budget');  plt.subplot(1, 2, 2)  sns.scatterplot(data['log\_budget'], data['log\_revenue'])  plt.title('log transfromation of revenue vs budget');  **Output:**    **Description:**  To find the relation between the movie budget and revenue we use scatterplot. It observes and visually display the relationship between variables. In the above plot we can visualize one thing that revenue is dependent of movie budget. Movie with highest budget gets high revenue. |  |

1. **Impact of Film’s Revenue with or without Homepage**

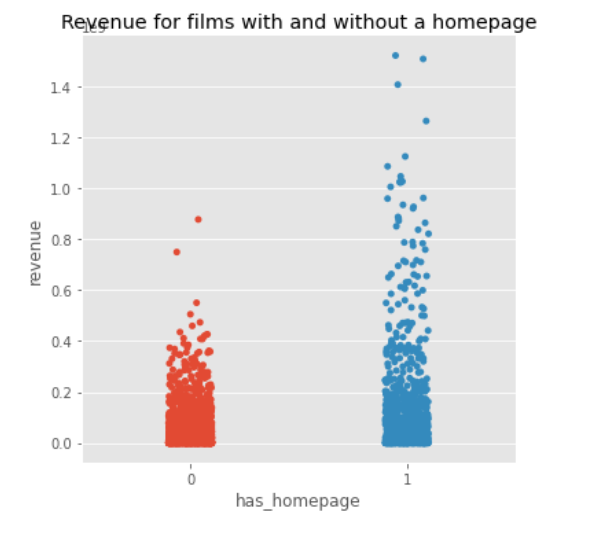
data['has\_homepage'] = 0

data.loc[data['homepage'].isnull() == False, 'has\_homepage'] = 1

sns.catplot(x='has\_homepage', y='revenue', data=data);

plt.title('Revenue for films with and without a homepage');

**Output:**

****

**Description:**

Here we collect the data of movies which has homepage and assigned it as 1, those movies which doesn’t have homepage assigned it as 0. Then to represent the data with homepage and without homepage we used categorical plot. From the result we get to know movie having homepages has more revenue.

1. **Films Revenue in various Languages**

language\_data = data.loc[data['original\_language'].isin(data['original\_language'].value\_counts().head(10).index)]

plt.figure(figsize=(16,8))

plt.subplot(1, 2, 1)

sns.boxplot(x='original\_language', y = 'revenue', data=language\_data )

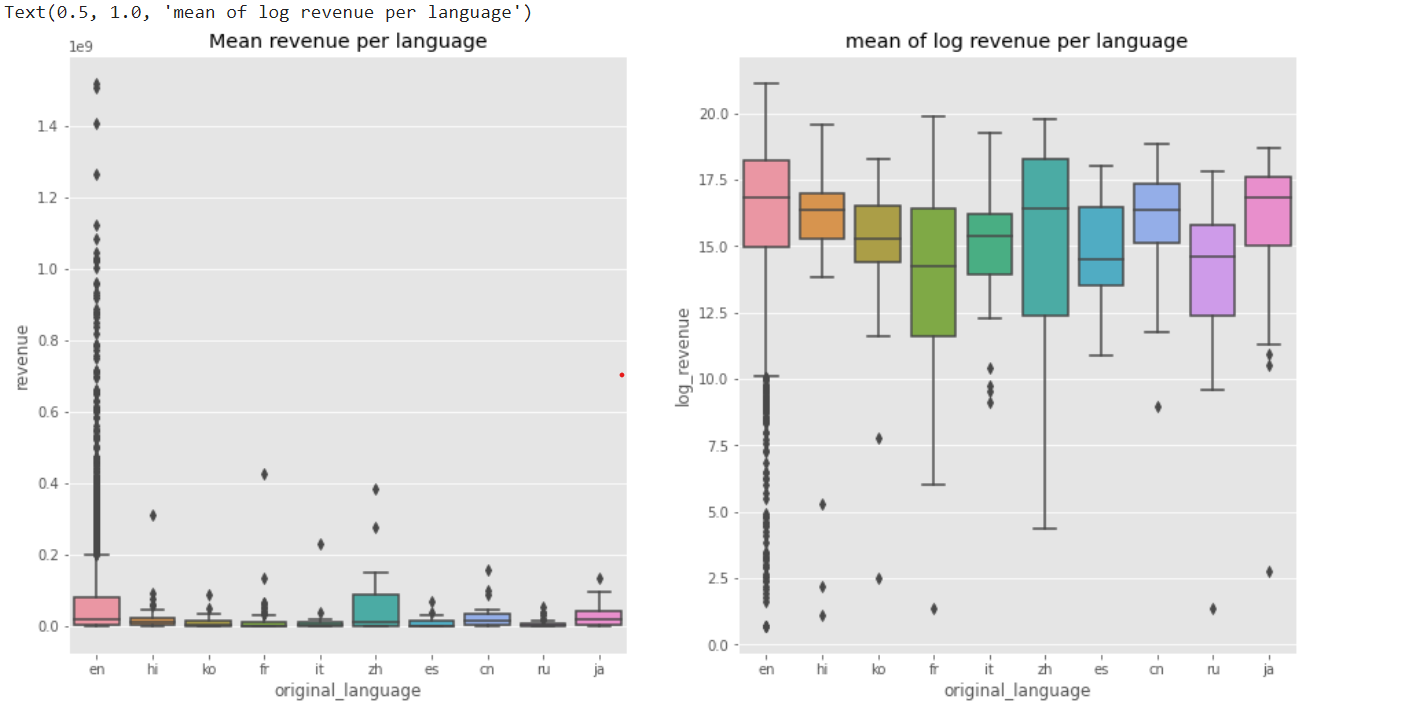
plt.title('Mean revenue per language')

plt.subplot(1, 2, 2)

sns.boxplot(x='original\_language', y = 'log\_revenue', data=language\_data)

plt.title('mean of log revenue per language')

Output:



Description:

Here we store original languages in language\_data, then we use box plot to represent the data taking x-axis as languages and y-axis as revenue. Box plot is used to summarize a set of data measured on an interval scale. It shows the shape of distribution, its central value, mean and its variability. From above plot we can conclude the most of the movies are made zh (language code).

1. **Frequent Words in Movie Titles**

plt.figure(figsize=(12, 12))

text =  ' '.join(data['original\_title'].values)

wordcloud = WordCloud(max\_font\_size=None,

                     background\_color ='white',

                     width =1200, height =1000).generate(text)

plt.imshow(wordcloud)

plt.title('Top word across movie titles')

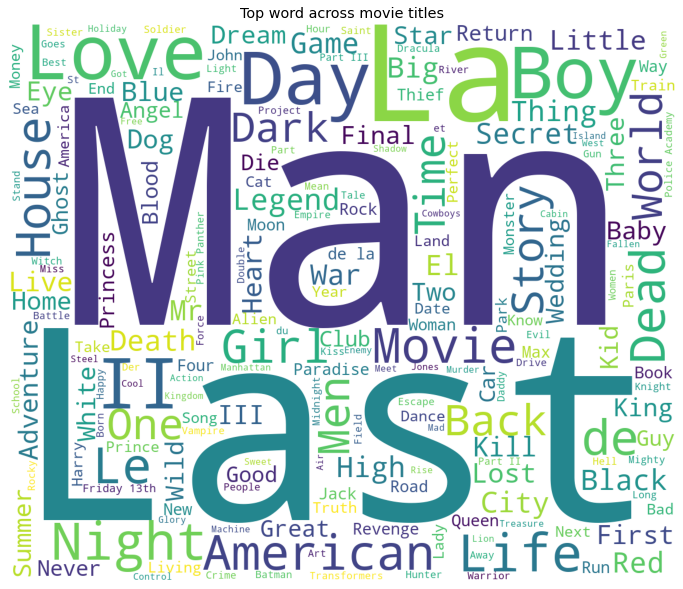
plt.axis('off')

plt.show()

**Description:**

To find most frequent words used in movie title we use word cloud. It highlights the popular word or most frequently used words. Word cloud can only access text but not float values or integers.

**Output:**

****

1. **Frequent Words in Movie Overviews**

plt.figure(figsize=(12, 12))

text = ‘ '.join(data['overview'].fillna('').values)

wordcloud = WordCloud(max\_font\_size=None,

                     background\_color ='white',

                     width =1200, height =1000).generate(text)

plt.imshow(wordcloud)

plt.title('Top word across movie overviews')

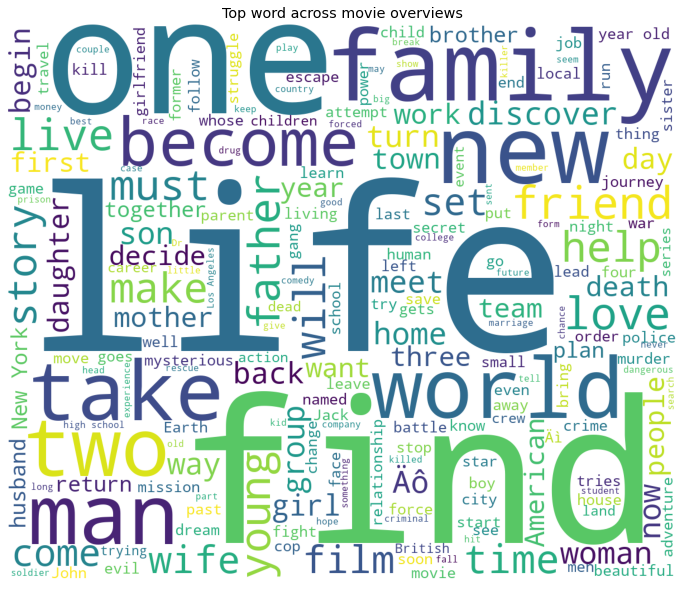
plt.axis('off')

plt.show()

**Description:**

We use. fillna() inbuilt function in pandas because if csv file contains some null values it is represented as Nan/Na in data frames, which leads to error. To get rid of that we use .fillna() it just removes the null values.

**Output:**

****

**Conclusion:**

We can get answers to all those questions that comes to our mind about popularity, revenue, budget, etc…

Using Data Visualization and Data Wrangling we get many more facts about movies by representing it in its best plot.

|  |
| --- |
|  |

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |