PHASE-1 PROJECT

PRESENTATION

BUSINESS CASE

Microsoft firm wants to venture into Movies industry by creating new movie studio They need: concrete information before venturing into the investment in order to make a viable decision on

- Which type of films are doing well in box_office.
- Films that are currently doing well in box_office.

IMPORTING NECESSARY LIBRARIES

The following libraries are usefull in data Anlysis

- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- import seaborn as sns

LOADING DATA INTO PANDAS

Loading 'tmdb.movies.csv' Data

Converting the CSV file format into pandas represents data in tabular format which is more easier to handle and manipulate huge data and assigning it to the object name data_ movies.

# CONVERTING AND ASSIGNING DATA INTO PANDAS.
data_movies=pd.read_csv('tmdb.movies.csv')
as shown below

	Genre_Ids	Id	Original_Language	Popularity	Release_Date	Title	Vote_Average	Vote_Count	Year	month	
0	[12, 14, 10751]	12444	En	33.533	2010-11-19	Harry Potter and the Deathly Hallows: Part 1	7.7	10788	2010	11	
1	[14, 12, 16, 10751]	10191	En	28.734	2010-03-26	How to Train Your Dragon	7.7	7610	2010	3	
2	[12, 28,	10138	En	28.515	2010-	Iron	6.8	12368	2010	5	

CROSSCHECKING BASIC STRUCTURE OF OUR DATASET USING THE INBUILT METHODS BEFORE CLEANING.¶

- ✓ data_movies.shape >>shows the structure of dataset
- ✓ data_movies.columns >>previewing column names incase of any changes
  data_movies.head() >>showing the firts 5 dataset
- ✓ data_movies.tail() showing the last 5 dataset
- ✓ data_movies.infor >>shows basic information on dataset
  data_movies.isnull()>>missing values of the dataset
- ✓ data_movies. isna().sum >>shows total sum of missing values of the dataset
  data_movies. describe >>shows the statistical summary of the dataset
  data_movies.dtypes >>shows type of data in our dataset

(26517, 10)

OBSERVATION

code for checking the dataset shape(structure)

data_movies.shape
(26517,10)

- Above results shows our dataset contains (26517 rows and 10 columns)

#code syntax below display first 5 information of the dataset.

- data_movies.head()

#code syntax below display first 5 information of the dataset. Data_movies.head()

Unnamed: 0	genre_ids	id	original_language	original_title	popularity	release_date	title	vote_average	vote_count	
0	0	[12, 14, 10751]	12444	en	Harry Potter and the Deathly Hallows: Part 1	33.533	2010-11-19	Harry Potter and the Deathly Hallows: Part 1	7.7	10788
1	1	[14, 12, 16, 10751]	10191	en	How to Train Your Dragon	28.734	2010-03-26	How to Train Your Dragon	7.7	7610
2	2	[12, 28, 878]	10138	en	Iron Man 2	28.515	2010-05-07	Iron Man 2	6.8	12368
3	3	[16, 35, 10751]	862	en	Toy Story	28.005	1995-11-22	Toy Story	7.9	10174
4	4	[28, 878, 12]	27205	en	Inception	27.920	2010-07-16	Inception	8.3	22186

Code syntax below display last 5 information of the dataset

data_movies.tail()

Unnamed: 0	genre_ids	id	original_language	original_title	popularity	release_date	title	vote_average	vote_count	
26512	26512	[27, 18]	488143	en	Laboratory Conditions	0.6	2018-10-13	Laboratory Conditions	0.0	1
26513	26513	[18, 53]	485975	en	_EXHIBIT_84xxx_	0.6	2018-05-01	_EXHIBIT_84xxx_	0.0	1
26514	26514	[14, 28, 12]	381231	en	The Last One	0.6	2018-10-01	The Last One	0.0	1
26515	26515	[10751, 12, 28]	366854	en	Trailer Made	0.6	2018-06-22	Trailer Made	0.0	1
26516	26516	[53, 27]	309885	en	The Church	0.6	2018-10-05	The Church	0.0	1

code syntax below display basic information of the dataset data_movies.info

It gives the summary information of the dataset

data_movies.isnull().sum()

- Crosschecking sum of missing values in our data

Crosschecking missing values data_movies.isna()

- Code above crosschecks sum of missing values in dataset

Displaying the statiscal summary of the dataset data_movies.describe()

Unnamed: 0	id	popularity	vote_average	vote_count	
count	26517.00000	26517.000000	26517.000000	26517.000000	26517.000000
mean	13258.00000	295050.153260	3.130912	5.991281	194.224837
std	7654.94288	153661.615648	4.355229	1.852946	960.961095
min	0.00000	27.000000	0.600000	0.000000	1.000000
25%	6629.00000	157851.000000	0.600000	5.000000	2.000000
50%	13258.00000	309581.000000	1.374000	6.000000	5.000000
75%	19887.00000	419542.000000	3.694000	7.000000	28.000000
max	26516.00000	608444.000000	80.773000	10.000000	22186.000000

Displaying type of data types

- data_movies.dtypes #dataset contains integer, object and float data types.
- The method above is used for displaying datatypes.

DATA CLEANING

Data cleaning involves cleaning data by removing in balanced data, inconsistencies data, misspelled text, duplicates in order to have a cleaning data

The method below are essential in data cleaning

Factors to crosscheck when performing data cleaning

1.Loading data to crosscheck unnecessary factors which makes our dataset uncleaned

2.Removing whitespaces>>(data_movies.columns.str.strip)

3.data_movies.duplicates() >>checking any duplicates

4.checking mispelled text

5.data_movies.astype() checking wrong data types assigned incase

6.data_movies >>checking outdated data

7.data_movies >> checking imbalanced data

8.checking outliers

9.data_movies.dtypes >>checking valid type of data in our dataset

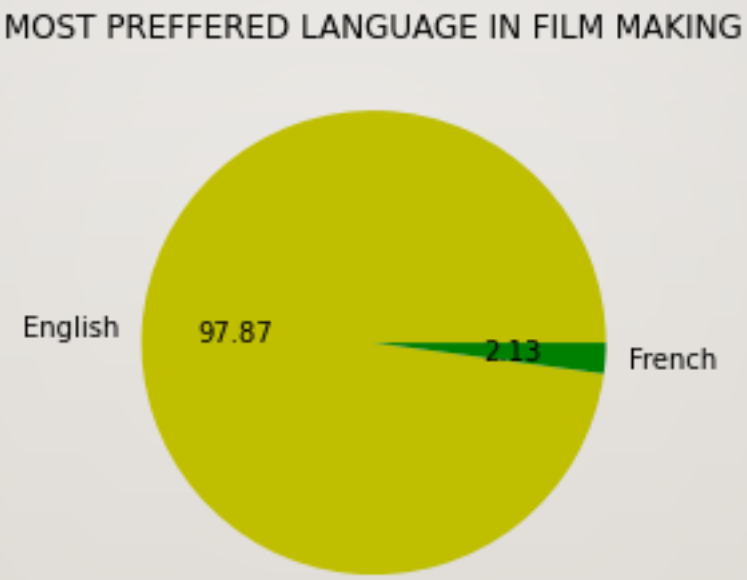
10.checking missing values.

EXPLORATORY DATA ANALYSIS

After crosschecking and cleaning data now we can question data the right questions relating to our business problems by using statistical graphics and other visualization methods since data presented in a pictorial format is much easier to interpret and make decisions, basing our factors on the following aspects in order to come up with a concrete solution which is viable, this factors are:

- *Language.
- *Release month or date.
- *Genre.
- *Popularity
- *Vote count.

PIE CHART


MOST PREFFERED LANGUAGE IN FILM MAKING
English
97.87
2.13
French

OBSERVATIONS

- From the pie chart above we can observe that English and French languages are used in making films.

RECOMMENDATIONS

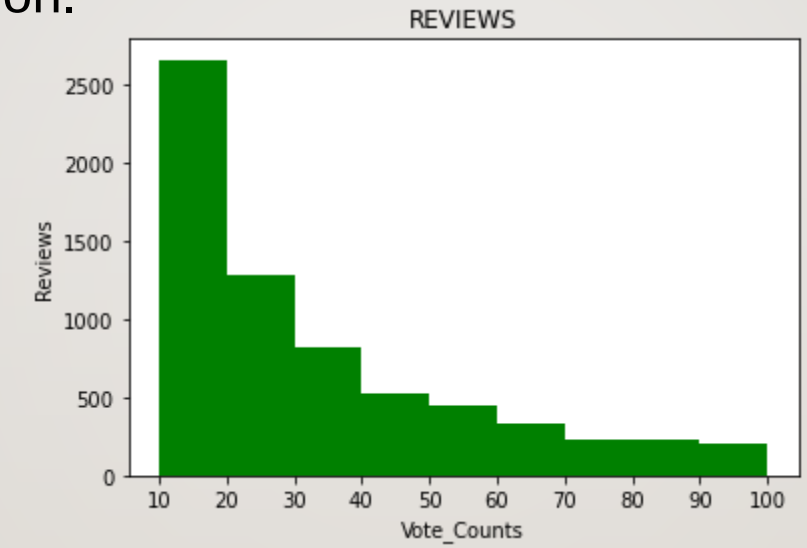
After a clear observation, I prefer both languages to be used in film making but English to take a bigger portion basing on the pie chart above where English takes 97.87% which is widely spoken in most countries and French 2.13% which is spoken in few countries.

CONCLUSIONS

- Basing on the business problem I conclude by clarifying both Languages to be used to capture all viewers but English to dominate the bigger part than French in order to hit the Box office market and perform well.

HISTOGRAM

- Histogram is pictorial representation of data, organized according to user specified range for easy interpretation.



OBSERVATIONS

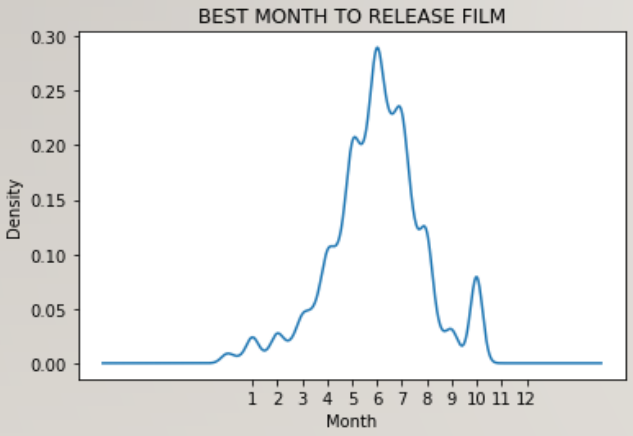
- From the Histogram shown above it indicates movie vote_count.

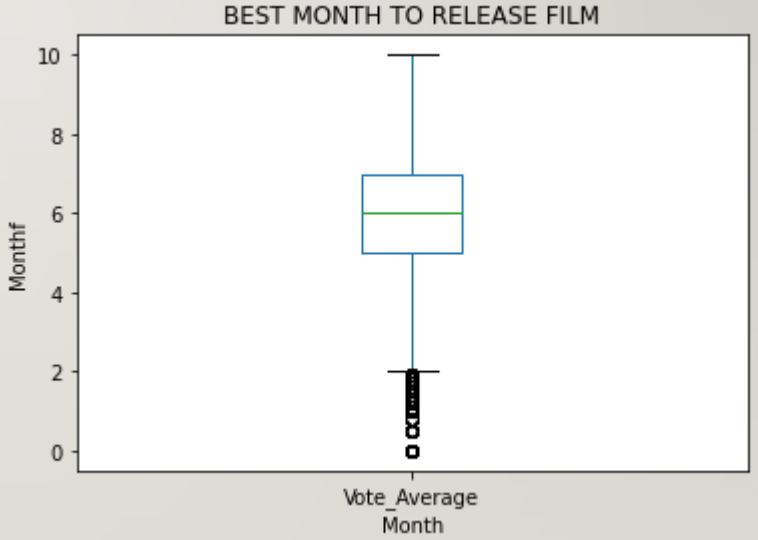
CONCLUSIONS

- Basing on our business understanding viewers votes or reviews is very important factor and must be consider for a movie to perform well and make profits, from the Histogram above the higher the reviews the increase chances of success.

BEST MONTH TO RELEASE FILM

- As shown below the best months to release films averagely falls on the middle months across all the months.





END OF PRESENTATION

- Hope the presentation will be viable to the company in decision making in order to venture into film industry.

- Presented by George Odhiambo Oduor