WILLY ANGOLE

PHASE 1 PROJECT WRITE UP

# 1.BUSINESS UNDERSTANDING

Link to jupyter notebook

Link to github repository

**UNDERSTANDING OF THE PROBLEM**

The problem faced by Microsoft in the creation of a new movie studio is the type of films that are currently doing the best at the box office. This role has been delegated to me to explore the given data in order to ascertain the best selling movies. The solution for this problem will be achieved by finding out the best selling movies, the time that this movies sell best and finally the potential writers and directors to perform the task.

**2.DATA UNDERSTANDING**

This is done by assessing the data. I assessed the six data sets provided by the mocrosoft in order to provide a solution for studio set. This was possible by unzipping the data set provided and actually loading the data in the jupyter notebook, opening the files and trying to see which data sets are related there. I got three files that could help me to answer questions that could enable me to solve the problem. The files that aided in project success werr:

1.Movie budget ,,,this file contained expenditure on the movies and returns hence being of significance in calculating profit. The very same file contained date for the project, this helped so much in identifying the best months of the year that profits were high.

2.im\_db ,,,,this contained movie genres with high ratings hence being recomendable for finding the best movie genre.

3.Movie\_info contained names of the writers who have managed to write many movies and this was essential because it help in getting potential employees for the company.

The very same table contained directors who have directed many movies , this is essentia in selecting potential directors for the company to hire.

**3.DATA ANALYSIS**

**EXPLORATORY DATA ANALYSIS**

**3.1 DATA CLEANING**

This is a process of ensuring that the data you are working on lacks missing values as well as duplicates.

This was possible due to availability of inbuilt python modules which are imported to enable to handle the task. This includes:

\*importing matplotlib for plotting

\*import numpy for numerical operation

\*importing searbon for advanced and cool plotting

Using of statistical and data manipulation techniques was applied, this helps in having a look at a bigger picture of the data. This displayed relationship between variables and it helped in coming up with the desired recommendations for the project.

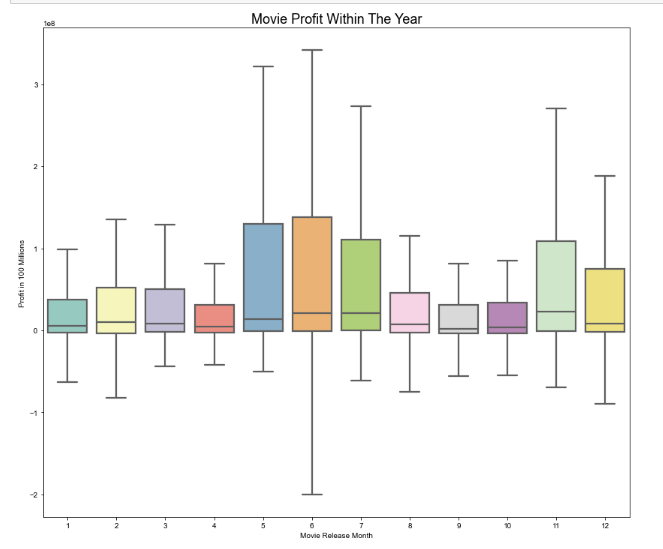
**3.2 HANDLING OF MISSING VALUES**

The first file for movie budget information contained no missing value, the second file which was im\_db contained missing values. This was handled by dropping the rows with missing values , this was best technique to handle it.

The third file which was movie\_info had missing values but this was handled by filling the missing data with most appearing characters in each column. Removing could have not been the best option because the missing values were small in number.

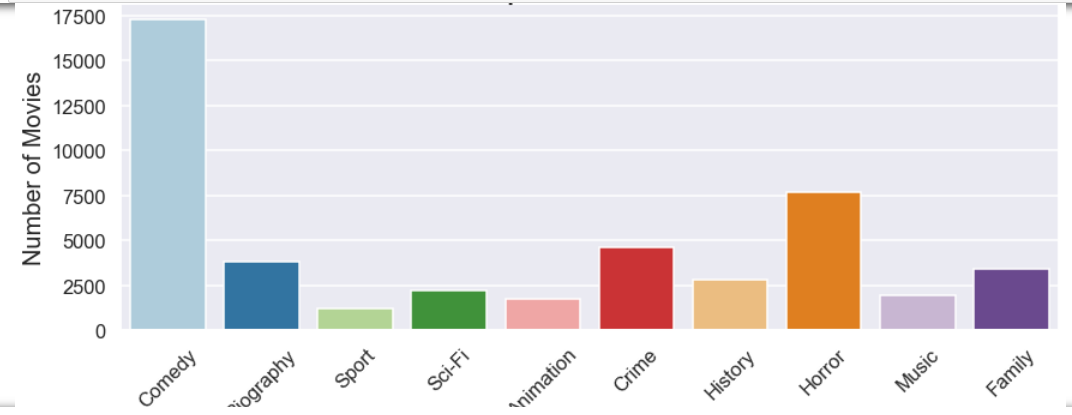
**3.3 DUPLICATE VALUES**

There were no duplicates values in the data were removedwhile cleaning data to avoid issue many outliers which could have hindered the decision making while recommending the results.

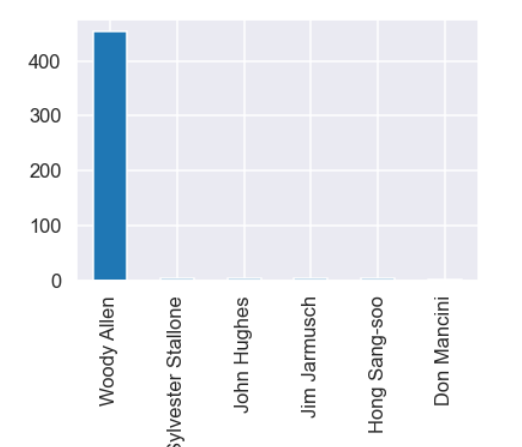
****

**The above graphs shows relationship between months of the year and profit yielded.**

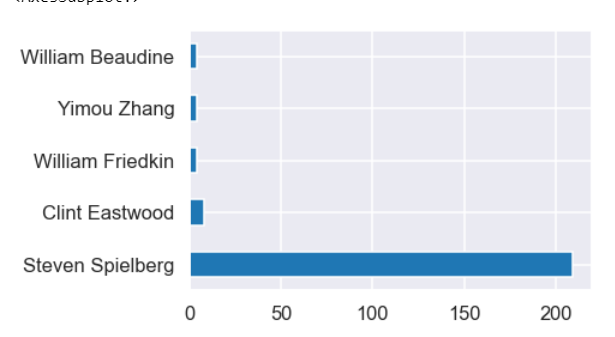
**The profit are high during April, May, Jine and December, this are the best times to release movies.**

****

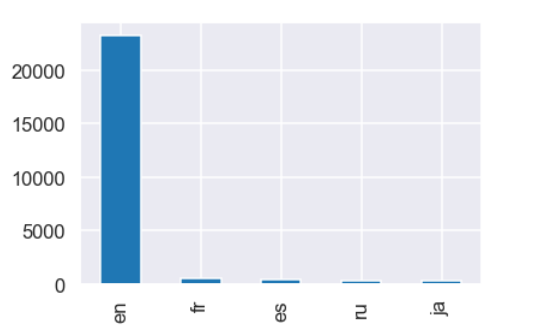
The bar graph above shows relationship between number of movies and genre of movies ,,the most liked genre is comedy then horror movies.



In the above bar graph shows movies writers with their value counts,,,,Woody Allen happens to be the written the highest number of movies.



The above horizontal bar graph shows the director who has directed many movies and that is Steven Spielberg.



The above graph shows best movie languages liked by most viewers, This shows that most people like to watch movies English movies.

**4 RECOMMENDATIONS**

#### **Based on the above data Analysis I recoomend microsoft to:**

### 1.Release comedy movies, horror movies, crime movies and family movies in order for them to gain high profit

### 2.Release movies in the month of April, may, june and December because to in order to yield high profit, most people are likely to be on holidays watching and enjoying the stated movies genres

### 3.Hire writers who have written alot of movies because they have got experience.

### 4.Hire directors who have directed many movies since they have got experience

#### **5.The most liked movie language is English**