MICROSOFT'S NEW STUDIO PROJECT

BUSINESS UNDERSTANDING

PROBLEM STATEMENT

The problem statement is to determine the genre of films are currently doing best in the market. To know which films are leading I will use the exploratory data analysis on several datasets to generate insights for Microsoft new movie studio and know what type of films that are leading.

UNDERSTANDING THE PROBLEM

Microsoft wants to start a new movie studio but they don't know a lot about movies. They want to know what types of films are currently doing the best at the box office. This will help the head of Microsoft's new movie studio to decide what type of film to create.

DATA UNDERSTANDING

DATA COLLECTION

The data was collected from https://www.imdb.com/ and https://www.boxofficemojo.com/ . The two datasets will help us get more details about movies and see which films are currently doing the best at the box office.

DATA DESCRIPTION

Data description for imdb

The data below was collected from two tables, that is movie basics and movie rating collected from imdb database. The two tables were joined since they have great benefit on our analysis.

columns	Description
movie_id	The identity of the movie
average_rating	The average rating a movie is given
num_votes	The number of votes a movie is given
primary_title	The title that is displayed on the imdb database
original_title	The title that the movie was originally given
start_year	The time a movie was published
runtime_minutes	The time that a movie take to complete
genres	The type of movie

Data description for bom.movie_gross.csv.gz

title	The title of the movie
studio	The studio that released the movie

domestic_gross	The total amount of money the movie is generating in its country
foreign_gross	The total amount of money the movie is generating in its other countries
year	The year the movie was released

DEFINING THE METRIC FOR SUCCESS

EXPERIMENTAL DESIGN

1. Loading Datasets and Preparing the Data.
2. Data Cleaning to deal with null values and Outliers.
3. Exploratory Data Analysis (Univariate and Bivariate Analysis)
4. Conclusions and Recommendation

DATA PREPARATION

SELECTING DATA

We will use the relevant columns to know the currently leading films.

DATA CLEANING

Data cleaning is done to make sure there is accuracy, Completeness, Consistency and Uniformity of the Data.

The first thing done is renaming the columns in a way that they can be read. The next thing is to know the data type of each column so that we can be able to do the analysis.

The missing values were checked and were found. In the bom.movie_gross.csv.gz I replaced the missing values for foreign gross column and domestic gross with the median of the values that were present. I then dropped the rows where the studio column had null values.

In the imdb dataset where I joined two columns, I replaced the null values in the genres column with 'Missing'. I then dropped the remaining columns with null values.

The two datasets were found to have no duplicates.

DATA ANALYSIS

UNIVARIATE DATA ANALYSIS FOR bom.movie_gross.csv.gz

Numerical Data

There were many outliers in the dataset, domestic_gross (408) and foreign_gross (619). They were too many to remove since they would affect the accuracy of the analysis. The many outliers showed that the data was not normally distributed.

Categorical Data

The category that I focused on is the studio. IFC studio seems to be leading in film production, Microsoft should watch out on it.

Summary Statistics

	Domestic gross	Foreign gross	title	year	studio
Mean	28561064.15730337	52623864.15819042	_	_	_
Median	1400000.0	18900000.0	_	_	_

Mode	1400000.0	18900000.0	Bluebeard	2015	IFC
Range	936699900.0	960499400.0	_	_	_
Variance	4461119781012780.0	1.211637017121155e+16	_	_	_
Standard deviation	66791614.601031914	110074384.71875076	_	_	_

Univariate Analysis Recommendation

The data is heavily skewed to the left. There are several outliers on the right side this is why initially decided to keep the outliers. I have decided to use domestic_gross , studio and foreign_gross

BIVARIATE DATA ANALYSIS FOR bom.movie_gross.csv.gz

Numeric

Strong positive correlation was found among domestic_gross and foreign_gross. They were linear correlations with Pearson's coefficients greater than 0.78.

UNIVARIATE DATA ANALYSIS FOR imdb dataset

Numerical Data

There were many outliers in the dataset average_rating (1327), num_votes (10472) and runtime_minutes (3588). They were too many to remove since they would affect the accuracy of the analysis.

Categorical Data

The category that I focused on is the genres. Drama seems to be leading genre in the industry currently.

Summary Statistics

Statistics	runtime_minutes	genres
Mean	94.6540400990398	_
Mode	90.0	Drama

BIVARIATE DATA ANALYSIS FOR imdb dataset

There was weak positive correlation between num votes and runtime minutes. There was no correlation between average_rating and runtime minutes.

CONCLUSION

In conclusion there was no relationship between the rating and runtime minutes for the movie. IFC is the top leading studio generating the highest income. Domestic gross and foreign gross are highly correlated meaning an increase in domestic gross leads to an increase foreign gross.

RECOMMENDATION

The head of new Microsoft studio should watch IFC and get the technique they use to edit their movie and also the genres they mostly

produce since it generates the highest gross income both locally and internationally.

The head of new Microsoft studio should focus on the following genres: Drama, Documentary, Comedy, Comedy Drama and Horror. This are the top five leading genres currently. The should produce one genre at a time and see the ratings since its just the beginning.

I would recommend the head of new Microsoft studio that the maximum amount of time for a movie should be 95 minutes.