MOVIES ANALYSIS PROJECT

1. BUSINESS UNDERSTANDING

Microsoft company wants to get into the movie business and in order to do this, they need to know somethings before they get started. The aim of this project is to analyze data from various movie studios such as IMDB so as to know the factors to consider to enhance success of the movie business.

This analysis is going to answer the following questions among other questions that come up during analysis;

1. Which are the most produced genres in the movie industry?
2. What is the budget allocated for each genre?
3. What is the relationship between the budget allocated and the gross profit of a genre?

2. DATA UNDERSTANDING

2.1 DATA USED

In this analysis, I'll use three datasets to come up with recommendations for Microsoft. These datasets are from Box Office and IMDB. The datasets are;

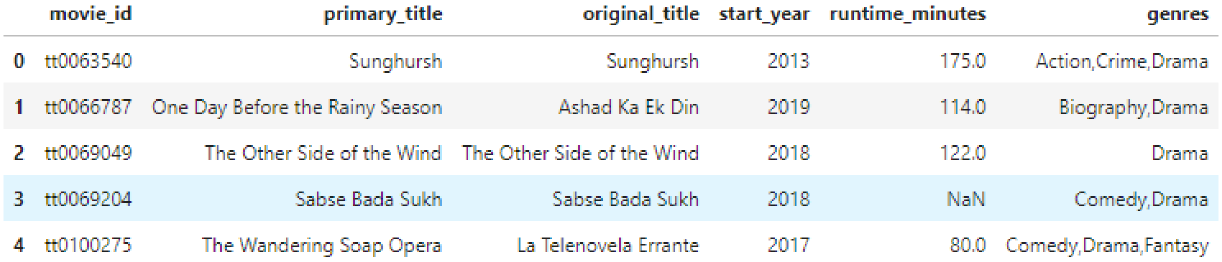
1. Im.db – This is a sqlite3 database. From this database I'll use only to tables which are movie_basics and movie_ratings.
2. Bom.movie_gross.csv – This is a csv file that contains movies and the domestic gross incomes and foreign gross incomes. It also contains the studios where the movies are produced.
3. tn.movie_budgets.csv – This is a csv file that shows the budgets for previously produced movies.

The above datasets were loaded into the pandas data frame so as to make the analysis of the data easier.

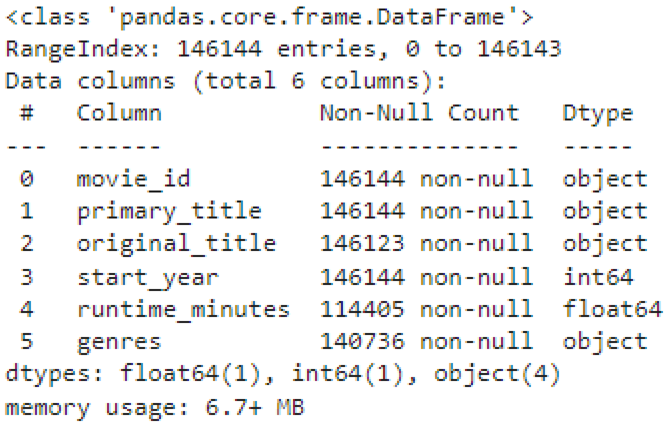
2.2 DATA DESCRIPTION.

The following is how the data looks like;

1. Movie_basics ;

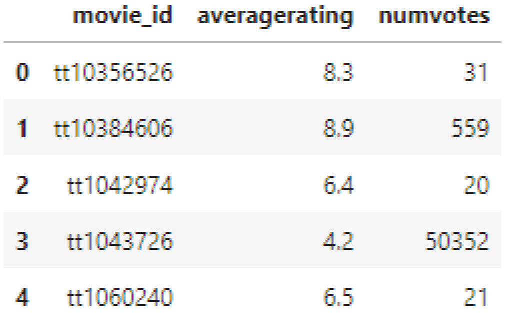


	movie_id	primary_title	original_title	start_year	runtime_minutes	genres
0	tt0063540	Sunghursh	Sunghursh	2013	175.0	Action,Crime,Drama
1	tt0066787	One Day Before the Rainy Season	Ashad Ka Ek Din	2019	114.0	Biography,Drama
2	tt0069049	The Other Side of the Wind	The Other Side of the Wind	2018	122.0	Drama
3	tt0069204	Sabse Bada Sukh	Sabse Bada Sukh	2018	NaN	Comedy,Drama
4	tt0100275	The Wandering Soap Opera	La Telenovela Errante	2017	80.0	Comedy,Drama,Fantasy

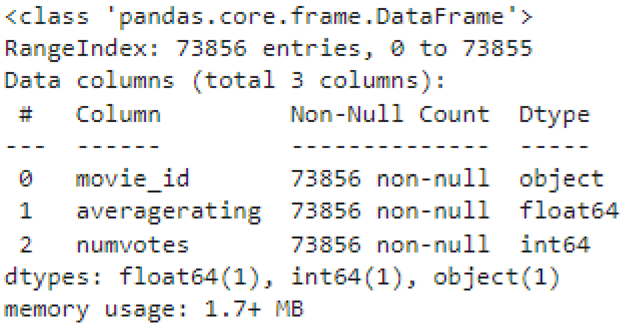


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 146144 entries, 0 to 146143
Data columns (total 6 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   movie_id         146144 non-null  object 
 1   primary_title    146144 non-null  object 
 2   original_title   146123 non-null  object 
 3   start_year       146144 non-null  int64  
 4   runtime_minutes  114405 non-null  float64
 5   genres           140736 non-null  object 
dtypes: float64(1), int64(1), object(4)
memory usage: 6.7+ MB
```

2. Movie_ratings;

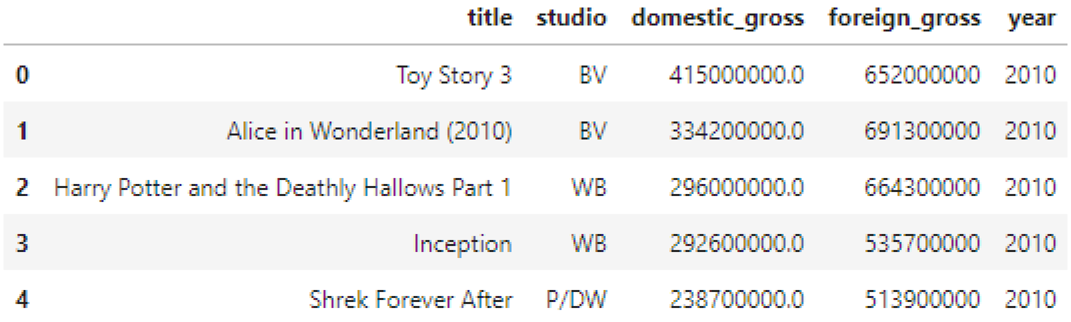


	movie_id	averagerating	numvotes
0	tt10356526	8.3	31
1	tt10384606	8.9	559
2	tt1042974	6.4	20
3	tt1043726	4.2	50352
4	tt1060240	6.5	21

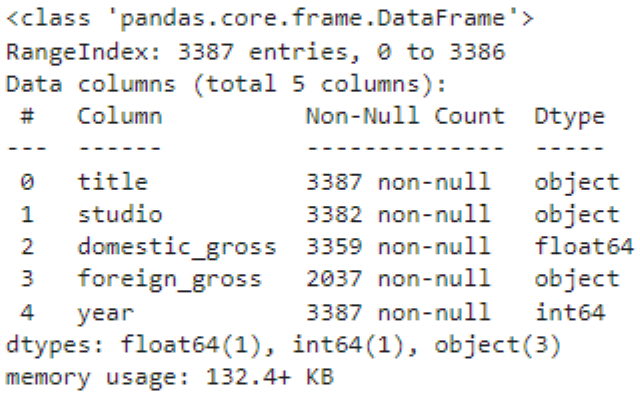


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 73856 entries, 0 to 73855
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   movie_id       73856 non-null  object 
 1   averagerating  73856 non-null  float64
 2   numvotes       73856 non-null  int64  
dtypes: float64(1), int64(1), object(1)
memory usage: 1.7+ MB
```

3. Bom.movie_gross.csv

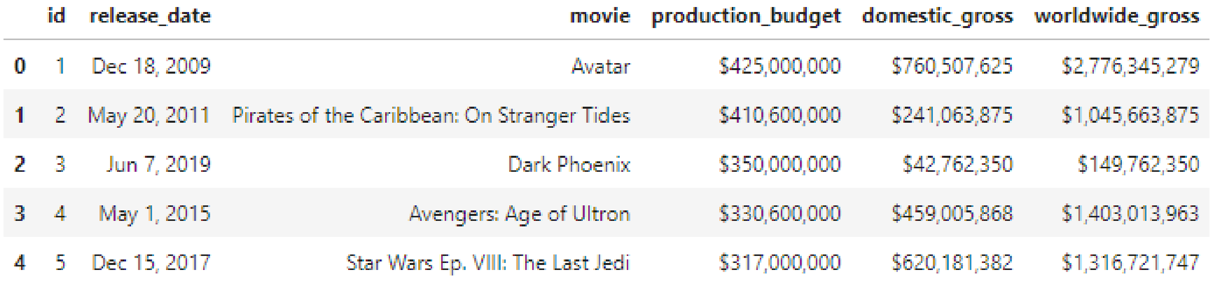


	title	studio	domestic_gross	foreign_gross	year
0	Toy Story 3	BV	415000000.0	652000000	2010
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010
3	Inception	WB	292600000.0	535700000	2010
4	Shrek Forever After	P/DW	238700000.0	513900000	2010

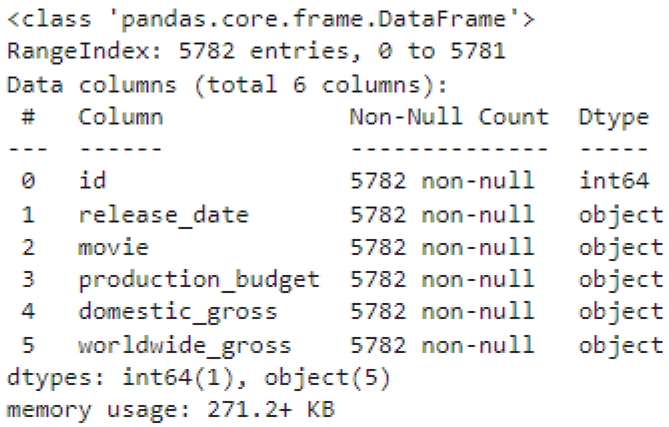


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3387 entries, 0 to 3386
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   title           3387 non-null   object 
 1   studio          3382 non-null   object 
 2   domestic_gross  3359 non-null   float64
 3   foreign_gross   2037 non-null   object 
 4   year            3387 non-null   int64  
dtypes: float64(1), int64(1), object(3)
memory usage: 132.4+ KB
```

4. tn.movie_budgets.csv;



	id	release_date	movie	production_budget	domestic_gross	worldwide_gross
0	1	Dec 18, 2009	Avatar	$425,000,000	$760,507,625	$2,776,345,279
1	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$410,600,000	$241,063,875	$1,045,663,875
2	3	Jun 7, 2019	Dark Phoenix	$350,000,000	$42,762,350	$149,762,350
3	4	May 1, 2015	Avengers: Age of Ultron	$330,600,000	$459,005,868	$1,403,013,963
4	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi	$317,000,000	$620,181,382	$1,316,721,747



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5782 entries, 0 to 5781
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   id                 5782 non-null   int64 
 1   release_date       5782 non-null   object
 2   movie              5782 non-null   object
 3   production_budget  5782 non-null   object
 4   domestic_gross     5782 non-null   object
 5   worldwide_gross    5782 non-null   object
dtypes: int64(1), object(5)
memory usage: 271.2+ KB
```

3. DATA PREPARATION

This section involves following the following steps;

1. Check for missing values.
2. If there are missing values deal with them accordingly either by dropping rows with missing values or replacing the missing values with the mean or median.
3. Check for duplicates.
4. If there are any duplicates, remove them.
5. Check to see if there are outliers.
6. If there are outliers deal with them accordingly.
7. Obtain the columns that are needed for the analysis.
8. Discard columns that you don’t need for analysis.

3.1 DATA CLEANING

This was done to ensure that we have a dataset that is clean, accurate, consistent and uniform. I first checked if there were any missing values in the dataset. The dataset had a lot of missing values. I replaced the missing values with the mean for the numerical columns. For the categorical columns, I dropped the rows with null values.

In this step, I also converted some columns to a numerical datatype so as to be able to plot graphs using the datasets. For example, in the budgets data, I had to convert the columns production budget, domestic gross and worldwide gross form an object datatype to a numerical datatype.

All the datasets had no duplicates.

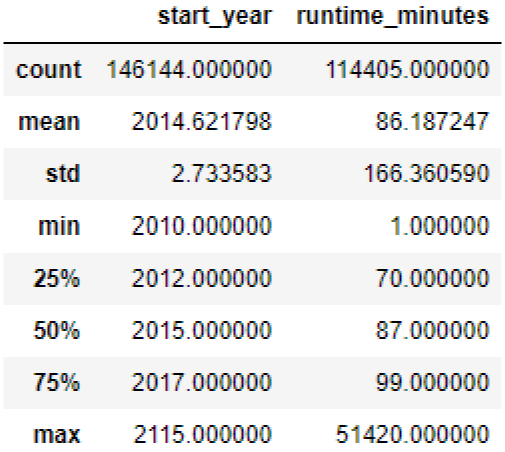
4. DATA ANALYSIS

In this step, I did exploratory data analysis (EDA) where I explored all the datasets above so as to see the relationship between the datasets.

I also joined some data frames so as to perform a deeper analysis on the data. After further analysis of the data, I plotted some visualizations to help in coming up with the recommendations for Microsoft.

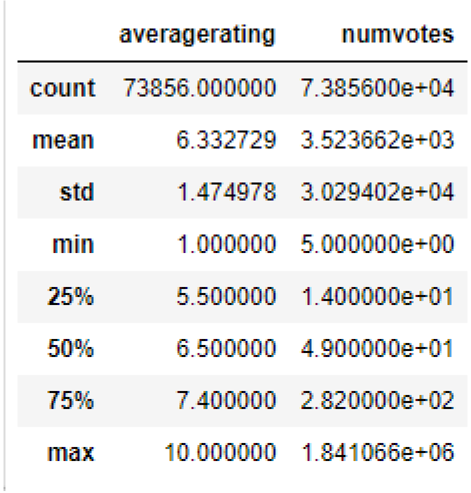
Summary statistics for the data;

a) Movie_basics



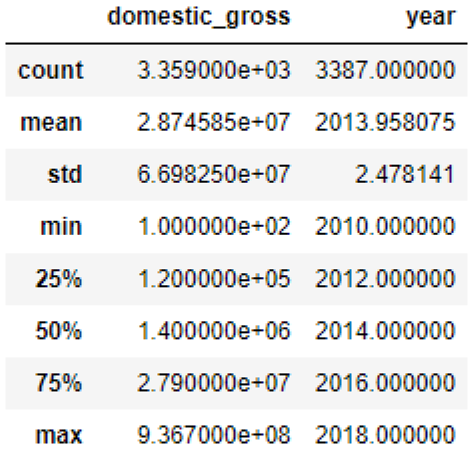
	start_year	runtime_minutes
count	146144.000000	114405.000000
mean	2014.621798	86.187247
std	2.733583	166.360590
min	2010.000000	1.000000
25%	2012.000000	70.000000
50%	2015.000000	87.000000
75%	2017.000000	99.000000
max	2115.000000	51420.000000

b) Movie ratings



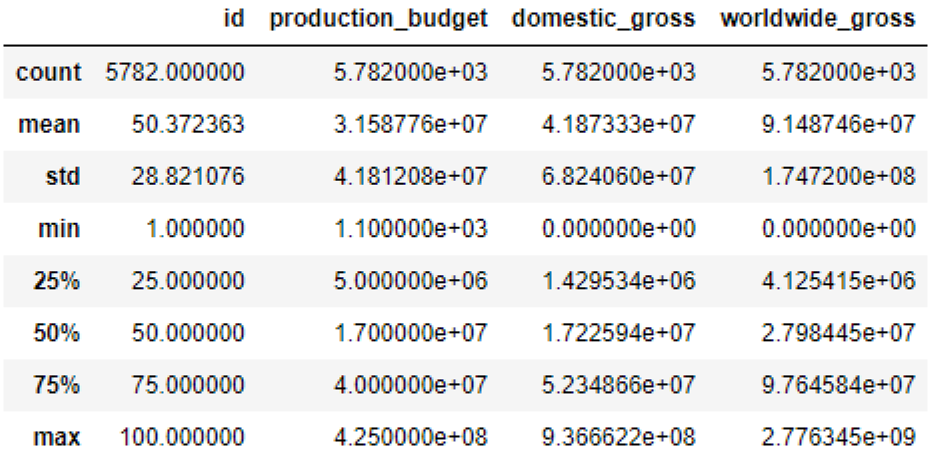
	averagerating	numvotes
count	73856.000000	7.385600e+04
mean	6.332729	3.523662e+03
std	1.474978	3.029402e+04
min	1.000000	5.000000e+00
25%	5.500000	1.400000e+01
50%	6.500000	4.900000e+01
75%	7.400000	2.820000e+02
max	10.000000	1.841066e+06

c) Bom.movie_gross.csv



	domestic_gross	year
count	3.359000e+03	3387.000000
mean	2.874585e+07	2013.958075
std	6.698250e+07	2.478141
min	1.000000e+02	2010.000000
25%	1.200000e+05	2012.000000
50%	1.400000e+06	2014.000000
75%	2.790000e+07	2016.000000
max	9.367000e+08	2018.000000

d) tn.movie_budgets.csv



	id	production_budget	domestic_gross	worldwide_gross
count	5782.000000	5.782000e+03	5.782000e+03	5.782000e+03
mean	50.372363	3.158776e+07	4.187333e+07	9.148746e+07
std	28.821076	4.181208e+07	6.824060e+07	1.747200e+08
min	1.000000	1.100000e+03	0.000000e+00	0.000000e+00
25%	25.000000	5.000000e+06	1.429534e+06	4.125415e+06
50%	50.000000	1.700000e+07	1.722594e+07	2.798445e+07
75%	75.000000	4.000000e+07	5.234866e+07	9.764584e+07
max	100.000000	4.250000e+08	9.366622e+08	2.776345e+09

CONCLUSIONS

After a lot of analysis of the data and visualization of the data, it is noted that some combination of genres in a movie have a high rating as well as a high profit after the movie has been released.

It is also noted that allocating the budget for movie also affects the profit you earn from the movies. If too little is budgeted, then the profit will also be little. The more resources you put into producing a movie, the higher the profit you earn from the movie.

RECOMMENDATIONS

- Microsoft should consider producing movies of genres 'Adventure, Drama', 'Comedy, Drama' or 'Adventure, Comedy'. A combination of these genres seem to have a very high rating and also have a high profit.
- Consider using a budget of over 10 million dollars. The more resources you allocate to a movie, the higher the profit.
- Microsoft should not concentrate on just producing a lot of movies because the number of movies produced by a studio does not contribute to an increase in the income of the studio. They should aim at producing the genres with a high rating as well as a high profit.