1.1 PROJECT OVERVIEW

1.2. Business understanding overview

As Microsoft you want to start a movie studio and you also want to know the kind of films to create. I personally think looking at datasets given; Box Office Mojo, IMDB, Rotten Tomatoes, The Movie DB and The Numbers that have the current trend of the movie world would be a great deal. After getting ideal datasets I will then use Exploratory data analysis to look into them and find out everything happening now in the movie world.

1.3 Business Understanding(Objectives)

This is a really new thing Microsoft is doing and the main thing data analysis is here to do is:

- bring out the types of films that are currently performing in box-office.
- Translate the findings from the performing films into actionable insights that the head of Microsoft's new movie studio can use to help decide what type of films to create.
- Bring out any new findings that would boost the content creation as a start.

Tentatively this will be judged a success if:

- By bringing out the types of films currently performing in box office makes Microsoft have a snippet of what they will start with in content creation.

Business Understanding (Assessing the situation)

This is Microsoft’s first attempt to start venturing into the movie world and therefore I will assume the task of a data analyst and use data from Box Office Mojo, IMDB and The Numbers which Microsoft has given me to try and bring out the current types of films being watched and also translate them into actionable insights that will then help the head of movies to know what type of films to create.

Risks, aside from the monetary outlays for the consultants and the time spent by on the study, there is not a great deal of immediate risk in this venture.

2. DATA UNDERSTANDING

2.1. Data Understanding of the Box Office Mojo dataset

This data was collected from the bom.movies_gross dataset. The program Box Office Mojo by imdbpro contains the movies that we are to use in this data.

2.Data Description

Column	Description
title	These are the titles of the movies
studio	These are the various studios that produce these movies

Domestic_gross	This is the standard measure of the value added created through the production of movies in a country.
Foreign_gross	This is the value added through the production of movies overseas.
year	This is the specific year of release of each movies

Missing Values in the bom.movies_gross Dataset

Column	No. of missing values
Title	0
Studio	5
Domestic_gross	28
Foreign_gross	1350
year	0

Most of the columns in this dataset will be features from which correlation of the dataset is found,this will help in giving the head of movies actionable insights on what Microsoft really needs to do in the movie world.

DATA UNDERSTANDING OF THR NUMBERS DATASET

I collected this data from the numbers dataset. This dataset shows us the top performing movies in terms of the gross that is produced and the production budget.

Data description

Column Name	**Description**
id	This is the relevant identifier for the movies
Release_date	This is the specific date a movie was released or will be released
movie	The name of a movie
Production_budget	The total money used for production of a movie
Domestic_gross	The total money produced locally by a movie from its country of origin
Worldwide_gross	The total money produced globally by a movie

The production budget column is very relevant for this analysis as after we have known the various types of films to produce we will want to approximate a budget or production. The domestic gross column and the worldwide gross column will come in handy to explain to us what to expect after making a certain investment on a movie because by Microsoft venturing to the movie word they are there to make money definitely and we need the numbers to give us some sense of direction.

DATA UNDERSTANDING OF THE IMDB DATASET

This data was collected from the IMDB dataset. The program IMDB contains the movies that we are to use in analysis. The movies have been placed in structured tables in relation to each other.

This dataset has 8 tables:

- movie_basics
- directors
- known_for
- movie_akas
- movie_ratings
- persons
- principals
- writers

The table movie_basics has a column “genre” that we are to use to tell Microsoft the current types of films that are more popular. The table movie_ratings also contains a column called “averagerating” that will tell us the most popular movies as per now. Combining these two datasets will help me give the Head of movies in Microsoft actionable insights that Microsoft can use instantly to help them join the movie world.

I joined the movie_basics and movie ratings tables in order to have a good overview of the average rating and the genre.

2.3 DESCRIBING THE QUESTION

2.3.1. SPECIFYING THE QUESTION

Determining the Genre of the movie that is bringing a lot of profit and specifying how ratings of the genres affect how a movie makes its profits. To investigate this, our hypothesis will be:

1. The Null hypothesis will be that the popularity of a genre also means that the movie has a high average rating and is creating a good domestic gross and foreign gross.
2. The Alternate hypothesis will be that when a genre is not popular its average rating, domestic gross and foreign gross are low.

2.3.2 DEFINING THE METRIC FOR SUCCESS

It will be deemed successful if the Null hypothesis fails to be rejected, i.e. if it is true.

2.3.3. EXPERIMENTAL DESIGN

1. . Loading Datasets and Preparing the Data.
2. Data Cleaning to deal with Anomalies and Outliers.
3. Exploratory Data Analysis (Univariate and Bivariate Analysis).
4. Hypothesis Testing to Implement the Solution.
5. Conclusions and Recommendations

HYPOTHESIS TESTING PROCEDURE

The procedure to be followed will be:

- Specifying the null and alternate hypothesis

For IMDB dataset for example most of the genres that have a rating of above 5 have an element of drama in them, meaning I have to be keen with the ratings and what it has to bring about. Also to notice is that there are some movies with a high rating and have a low number of votes. This means some individuals only remember to press the ratings button and fail to vote it, or else it could mean that some genres are just overrated

DATA PREPARATION

SELECTING THE DATA FROM THE BOX OFFICE MOJO DATASET.

I will use the studio, domestic gross and foreign gross columns as they are relevant to the study. The other columns will be used to look out for correlation.

SELECTING DATA FROM THE NUMBERS DATASET

I will use the movie, production budget, domestic gross and worldwide gross as they are really needed to have a good analysis. The other columns will be used to look out for correlation.

SELECTING THE DATA FROM THE IMDB DATASET.

I Decided to merge the movie basics and movie ratings table, so that I may know the genres and their specific ratings. This is much critical to the study. The other tables will be used for correlation.

DATA CLEANING

This was done to ensure the Validity, Accuracy, Completeness, Consistency and Uniformity of the Data.

1.DATA CLEANING OF THE BOX OFFICE MOJO DATASET

The first step was to check for the data types of each column. This is to make sure the relevant columns have the appropriate data types for analysis. Then to check for the missing values to avoid later problems. I found out that the studio, domestic gross and foreign gross columns have missing values.
The foreign gross column was in string format yet it represents currency so I had to coerce it to numeric before beginning to remove the missing values. I then decided to change the missing values into a percentage so that I can easily find out their impact on the data. The foreign gross column had a higher percentage, followed by the domestic gross then the studio column.
The domestic gross column had few null values dropping the null values would not be appropriate, but filling the null values with the median would make more sense.
The studio column had also few null values and because it is a string it was suitable to fill the missing values with “missing” would be appropriate so that we just know that a certain row had no studio. We have almost half of the dataset with missing values and they are from the foreign gross column, it will be wise if we drop the rows with the missing values from our dataset.
The last step was to confirm that there are no null values after the data cleaning procedure.
Having done the missing values successfully, the second step was to check for the outliers as they may really affect my analysis. The domestic gross column had 180 outliers while the foreign gross column had 260 outliers. The total outliers in the whole dataset were 2032,

removing the outliers would really affect the dataset as most of the rows will be removed, the proper decision here was to retain the outliers.

2.DATA CLEANING OF THE NUMBERS DATASET

The columns were then checked to see if they were of the appropriate types / dtypes. The production budget, domestic gross and foreign gross columns were in string format yet we need them in numeric form to enable numeric calculations. The three columns were then converted to numeric form to enable calculations. After this, missing values in the datasets were checked for and were found to be none. The data was also found to be consistent there being no duplicated data although the dataset had 5782 outliers. Removing these outliers would really affect the outliers and hence deciding to retain the outliers was the most suitable decision.

3.DATA CLEANING OF THE IMDB DATASET

I merged two tables movie ratings and movie basics.
The first thing done was to rename the columns in the merged table (merge imdb) to make them uniform and readable. The columns were then checked to see if they were of the appropriate types / dtypes. After this, missing values in the datasets were checked for and two columns runtime minutes and genres had missing values. I then found the shape of the merge imdb table. I replaced the values in the genres column with “missing” because it is a categorical data. I dropped the missing values in the runtime minutes’ column because it is not part of what I will use for the analysis. The data was also found to be consistent there being no duplicated data. One of the worrying things was the number of outliers, the whole dataset had a total of 66236 outliers. This meant I could not remove them as they would have a huge impact on the analysis.

DATA ANALYSIS

EXPLORATORY DATA ANALYSIS

EDA (BOX OFFICE MOJO DATASET)

(a). Numerical

The missing values in the studio column were replaced with “missing” to simply show that a certain row had no studio. The domestic gross column missing values were replaced with the median while the foreign gross column rows that had missing values were dropped.
There were a lot of outliers, domestic_gross(180) and foreign_gross(26) columns. These are too many to remove as this will affect the accuracy of the data analysis, and the result could be inconclusive and/or incorrect. The outliers suggest that the data could possibly be data that does not have a normal distribution.

(b). Categorical

The studio column is the category of interest. By finding the count we get to know the most popular studios. Universal studios, Fox studios and WB studios are the top three studios.

(c). Summary statistics

	Domestic_gross	Foreign_gross	Year
Count	2.032000e+03	2.032000e+03	2032.000000
Mean	4.505974e+07	7.505704e+07	2013.486713
std	7.602265e+07	1.375294e+08	2.591852
Min	4.000000e+02	6.000000e+02	2010.000000
25%	6.912500e+05	3.775000e+06	2011.000000

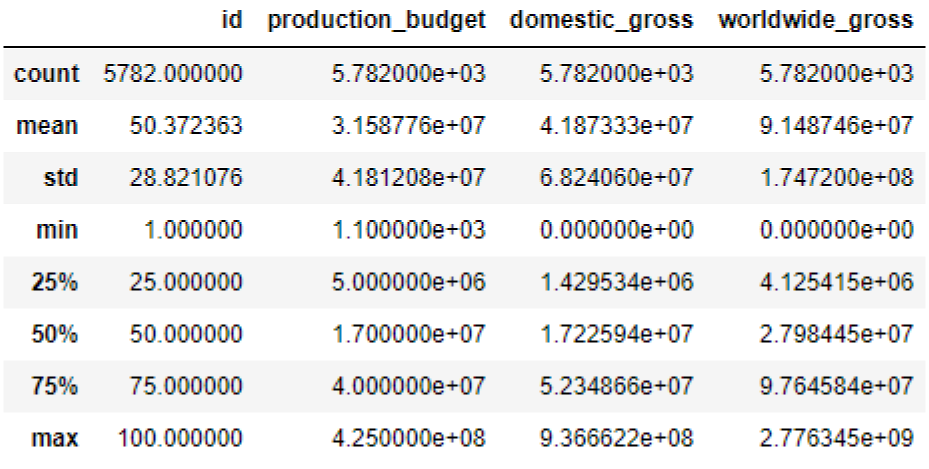
50%	1.535000e+07	1.890000e+07	2013.000000
75%	5.482500e+07	7.505000e+07	2016.000000
max	7.001000e+08	9.605000e+08	2018.000000

(d). Box Office Mojo Analysis Recommendation
Using the studio column, Universal studios is the top, followed by Fox studios and then WB studio. This means that for Microsoft to succeed they need to take note of the top performers in the industry. This will help Microsoft to know where to edge them out and make themselves more relevant than the other studios. In future analysis I would recommend to use the studio and its Gross income. We were limited here because of the many outliers we have found.

EDA THE NUMBERS DATASET
(a)Numerical
Actually the essence of this dataset is the gross income and the production budget. The production budget column had 431 outliers, the domestic gross column had 463 outliers and the worldwide gross had 604 outliers. Removing these outliers would really impact the analysis, so the option was to change the numeric columns in string to integer to enable calculations.
I then used the movie and gross column to know how the much money the movies produce both locally and worldwide. In the domestic gross, we had Star Wars, avatar and black panther giving the highest domestic gross while in the worldwide gross we had Avatar, Titanic and Star Wars leading. I added a new column gross income to give us the gross from domestic gross and foreign gross. Comparing the movie and its production budget was also a relevant idea just to help us take various precautions.

(b)Categorical
This dataset was mainly focused on numeric data rather than categorical data. The only categorical data here was the movie column which was directly related to the numerical columns.

(c). Summary statistics



	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

(d). The Numbers Dataset Analysis

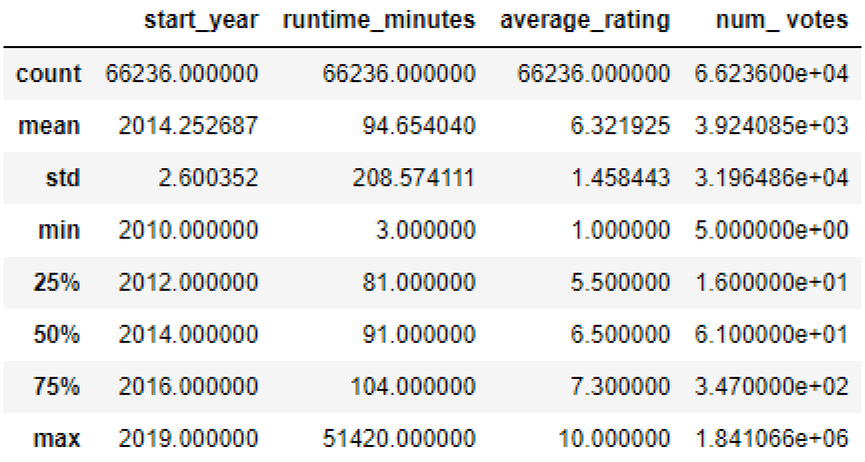
Basing on the columns I analyzed during this analysis, the production budget for a movie might lead to the gross income being high. Avatar's production budget is very high so is its worldwide gross and domestic gross. Apparently, some movies which also have a high production budget do not have a high gross. This would be a very good thing that Microsoft should watch out for as having a high production budget does not guarantee the success of a movie in terms of the gross that comes out of it.The analysis also showed that when a movie is performing well locally chances of it failing worldwide is very low.Micrososft should really focus also on this as also they will be expecting returns on the money they've put in.

EDA THE IMDB DATASET

The most important data that was being analyzed from here was categorical data, the focus was on the movie basics and movie rating tables which I merged to get one table merge imdb that could give an overview of a genre and its average rating. Having had a lot of outliers in some columns like number of votes it would be really difficult to use them in the analysis as this would also largely impact the question of study. The genre that was being produced the most was Drama, Documentary and comedy. Also worth noting was that the top 10 genres being produced had an element of drama in them example; comedy-drama, drama-romance.

I also tried to sort the average rating in order to find the top genres being watched and it was worth noting that the top genre was documentary, comedy, drama, adventure and crime also had some elements although Documentary is the genre with a high rating. Most of the genres also had an average runtime minutes of 90.0 minutes about 1 hr. 30 minutes which is of much importance to Microsoft. Most genres were also averaging a rating of 6 meaning Microsoft should take note of this.

Summary Statistics



	start_year	runtime_minutes	average_rating	num_ votes
count	66236.000000	66236.000000	66236.000000	6.623600e+04
mean	2014.252687	94.654040	6.321925	3.924085e+03
std	2.600352	208.574111	1.458443	3.196486e+04
min	2010.000000	3.000000	1.000000	5.000000e+00
25%	2012.000000	81.000000	5.500000	1.600000e+01
50%	2014.000000	91.000000	6.500000	6.100000e+01
75%	2016.000000	104.000000	7.300000	3.470000e+02
max	2019.000000	51420.000000	10.000000	1.841066e+06

THE IMDB DATASET ANALYSIS

Microsoft should focus so much on the genre they produce, for instance the most produced genre in the analysis was Drama, Documentary and Comedy. But then the genre that had highest ratings was Documentary with comedy and drama also having some elements. To Microsoft this means that producing a genre that is not related to documentary will lead to it not being rated highly. This shows that when someone enters like the imdb website and start looking for a good genre to watch they will be convinced by the ratings on Documentaries much more and they will definitely have interest in it. Worthwhile finding out is the influence the rating has on the gross that comes out of the movie as Drama is the most produced genre.

HYPOTHESIS TESTING

To investigate this, our hypothesis will be:

1. The Null hypothesis will be that the popularity of a genre also means that the movie has a high average rating and is creating a good domestic gross and foreign gross.
2. The Alternate hypothesis will be that when a genre is not popular its average rating, domestic gross and foreign gross are low.

HYPOTHESIS TESTING RESULTS

In the analysis especially in the imdb dataset it was crystal clear that the most produced genre was drama, documentary and comedy. While finding the genre that had a high average rating documentary was the top with instances of drama and comedy. This cements the null hypothesis that by a genre being popular it attributes to its average rating.

CONCLUSION

When a genre has a high average rating the probability of it being watched more is high. The genres that are being produced the most are drama, documentary and comedy. This means most of the studios have focused on this and it is what is performing in the movie world highly.

RECOMMENDATIONS

Microsoft should take note of the top performing studios; Uni, Fox and WB. These are top studios from the gross income they generate meaning they also put a lot of work to this and before you enter any space you must first know how the others have been doing it and if we could maybe get the genre produced by each studio it would be better so that as Microsoft we enter the space focusing on certain genres.

The production budget for a movie might lead to the gross income being high. Avatar’s production budget is very high so is its worldwide gross and domestic gross. Apparently, some movies which also have a high production budget do not have a high gross. This would be a very good thing that Microsoft should watch out for as having a high production budget does not guarantee the success of a movie in terms of the gross that comes out of it. The analysis also showed that when a movie is performing well locally chances of it failing worldwide is very low. Microsoft should really focus also on this as also they will be expecting returns on the money they’ve put in.

As Microsoft enters the movie world they should focus on top genres first; Drama, documentary and Comedy. Also take a program at a time to produce as it is just a start. Microsoft should pay a lot of attention also to the ratings they get as this maybe sending a signal to just shift to another. In future analysis it would be better if we rate a genre and the gross income. It would also be better to get the directors as this is a key factor in the movie world.