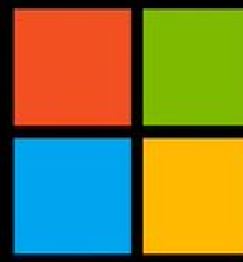
DATA SCIENCE

PHASE 1 PROJECT



Microsoft

Moringa School

James Irungu Ndiritu

TABLE OF CONTENTS

Problem Statement

Microsoft sees all the big companies creating original video content and they want to get in on the fun. They have decided to create a new movie studio, but they don’t know anything about creating movies. You are charged with exploring what types of films are currently doing the best at the box office. You must then translate those findings into actionable insights that the head of Microsoft's new movie studio can use to help decide what type of films to create.

Data Sources

1. bom.movie_gross.csv
2. title.basics.csv
3. title.basics.csv

Understanding Data

```
# Check that there are the correct number of rows
movie_gross.shape[0]
```

```
3387
```

```
# Check that there are the correct number of rows
title_basics.shape[0]
```

```
146144
```

```
# Check that there are the correct number of rows
title_rating.shape[0]
```

```
73856
```

```
#summary statistics of title_basics
title_basics.describe()
```

Understanding Data

```
#summary statistics of title_basics
movie_gross.describe()
```

	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

```
#summary statistics of title_basics
title_rating.describe()
```

Understanding Data

```
movie_gross['foreign_gross'].replace(',','',inplace=True,regex=True)
```

```
movie_gross.dtypes
```

```
title              object
studio             object
domestic_gross    float64
foreign_gross      object
year                int64
dtype: object
```

```
#replace the missing values of foreign_gross and domestic_gross with 0.
movie_gross['domestic_gross'] = movie_gross.domestic_gross.fillna(0.0)
movie_gross['foreign_gross']  = movie_gross.foreign_gross.fillna(0.0)

#replace missing values of studio with 'Unknown'
movie_gross['studio'] = movie_gross.studio.fillna('Unknown')
```

Understanding Data

```
# Function CorrectDataType to correct wrong data types
#Changing foreign_gross : object to float
def CorrectDataType(data, cols, dtype):
    for col in cols:
        data[col] = data[col].astype(dtype)
        return data.head()

CorrectDataType(movie_gross, [ 'foreign_gross'], 'float')
movie_gross.dtypes
```

```
title              object
studio             object
domestic_gross    float64
foreign_gross     float64
year                int64
dtype: object
```

Check and Removal of Duplicated Values

```
#Duplicated values
#title_basics[title_basics.duplicated(keep=False, subset=['primary_title','start_year'])].sort_values(by='start_year')
title_basics[title_basics.duplicated(keep=False, subset=['primary_title','genres','start_year'])].sort_values(by=['primary_title'
```

Understanding Data

```
AllDatasetsWithSpliGenres['genres'] = AllDatasetsWithSpliGenres['genres'].str.split(',')
#transform each element in the genres list to a row
AllDatasetsWithSpliGenres = AllDatasetsWithSpliGenres.explode('genres')
```

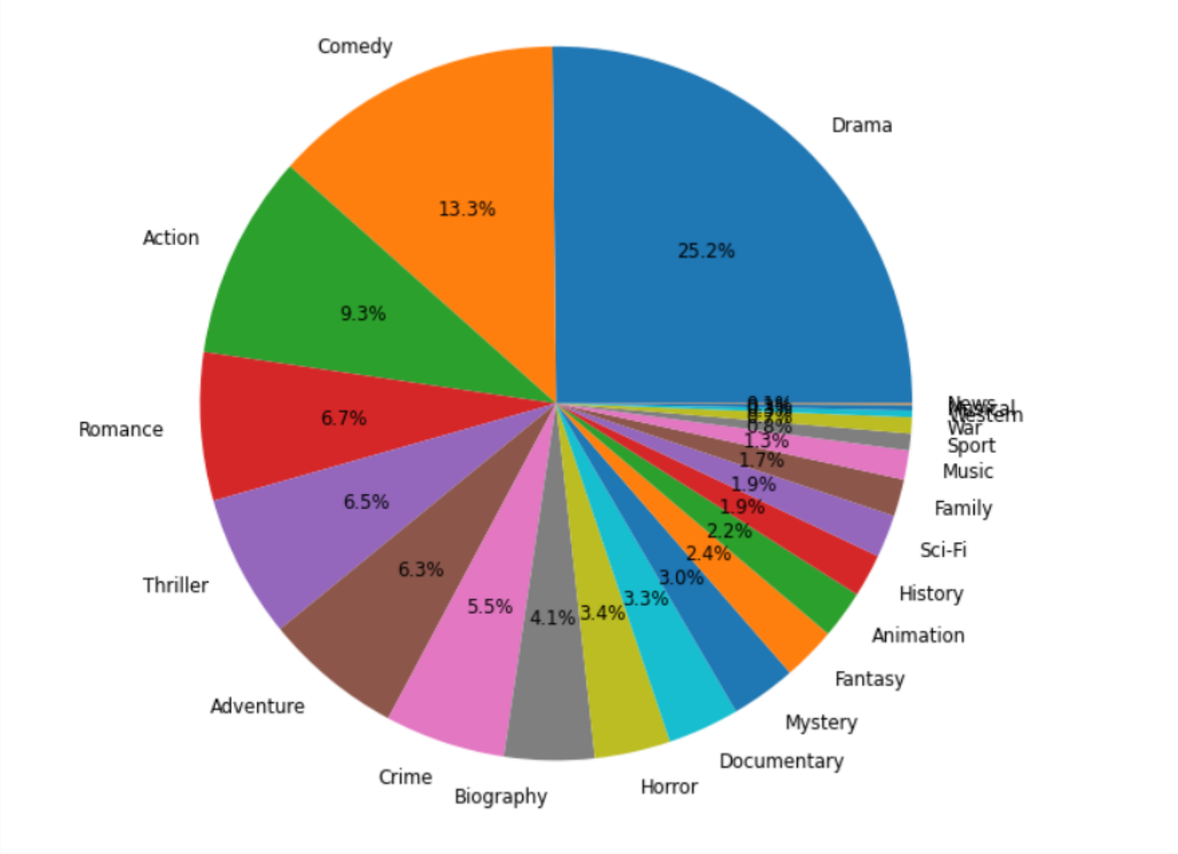
```
display(AllDatasetsWithSpliGenres)
```

```
#Reset Index
AllDatasetsWithSpliGenres=AllDatasetsWithSpliGenres.reset_index(drop=True)
```

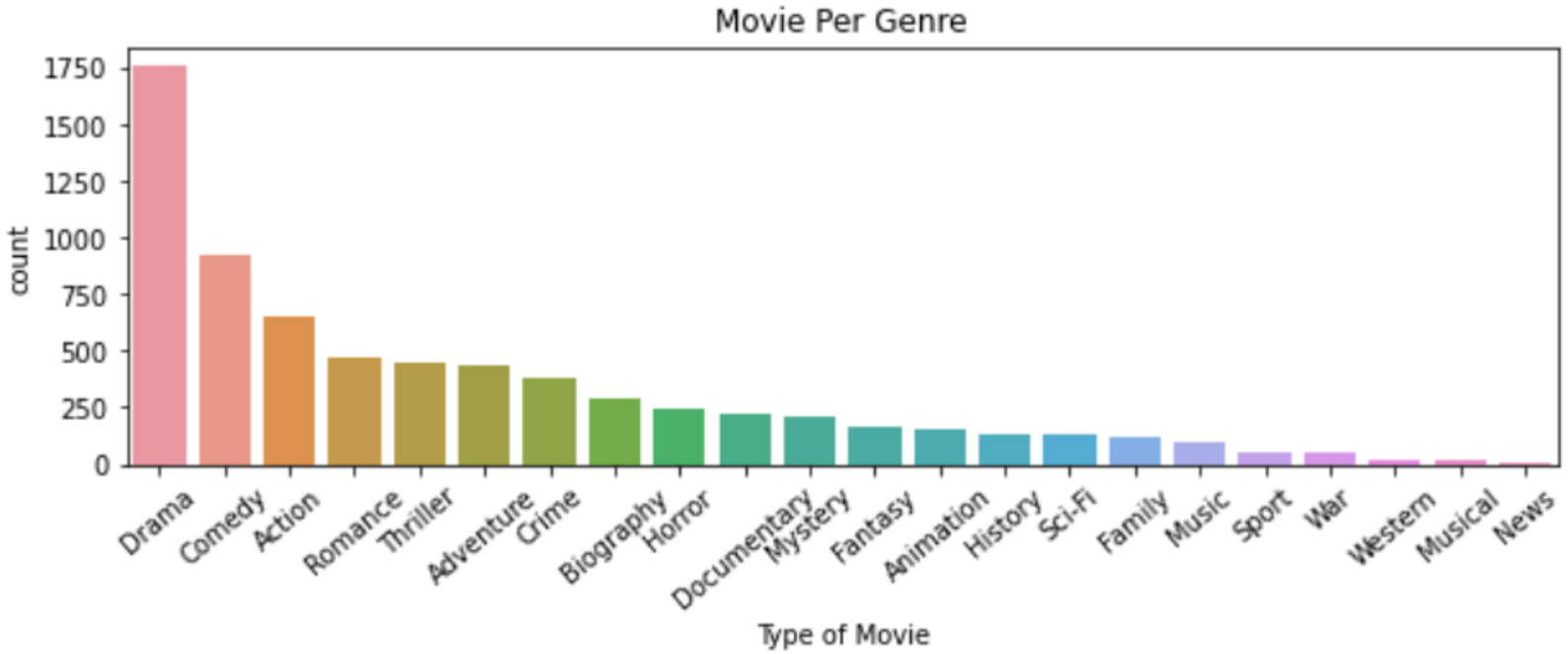
```
#value count of column genre
AllDatasetsWithSpliGenres['genres'].value_counts()
```

```
Drama          1756
Comedy          926
Action          646
Romance         468
Thriller        453
Adventure       439
Crime           382
Biography       285
Horror          240
Documentary     227
Mystery         207
Fantasy         170
Animation       152
History         136
Sci-Fi          135
Family          117
```

Data Analysis And Findings

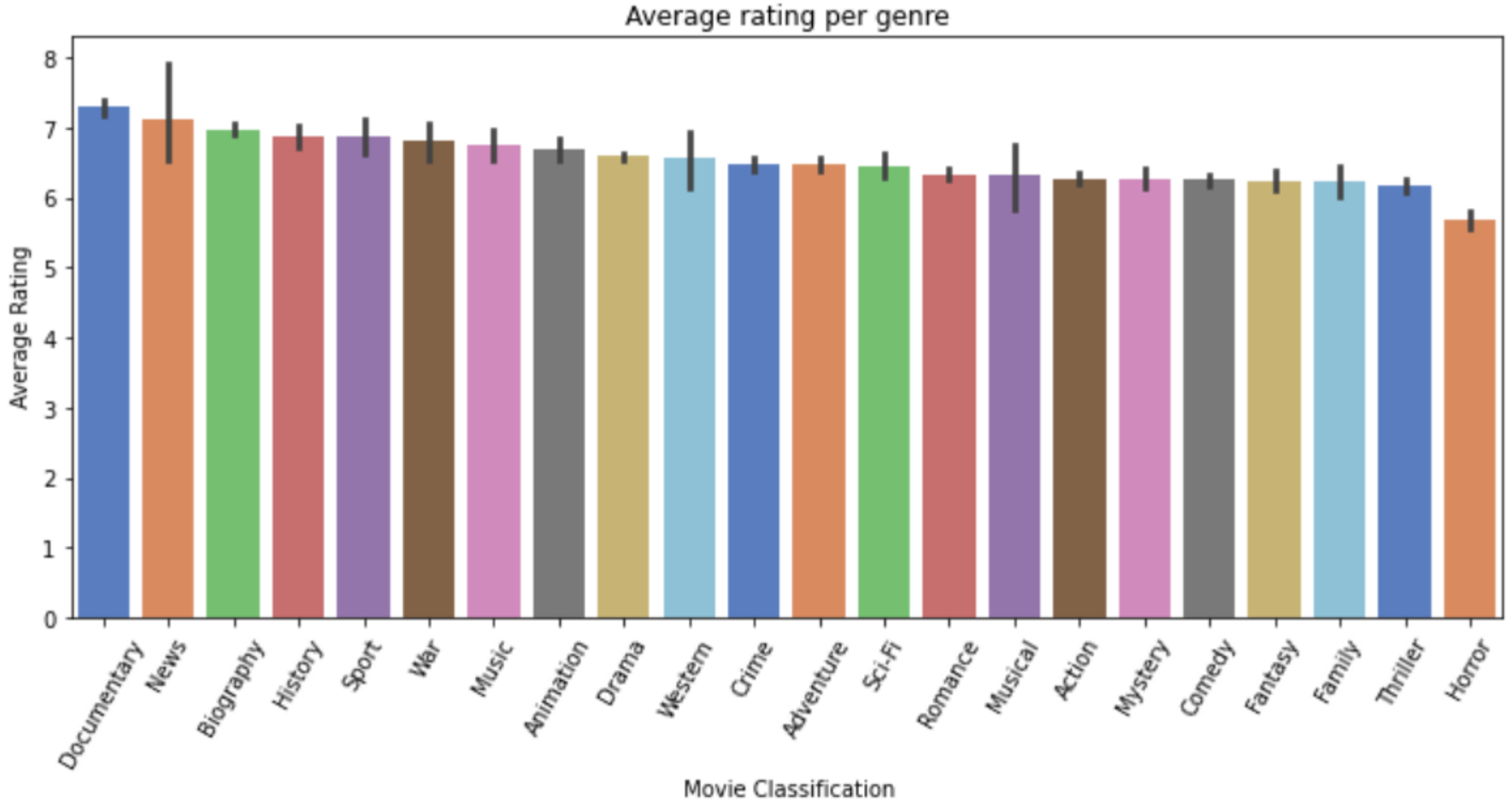


Data Analysis And Findings



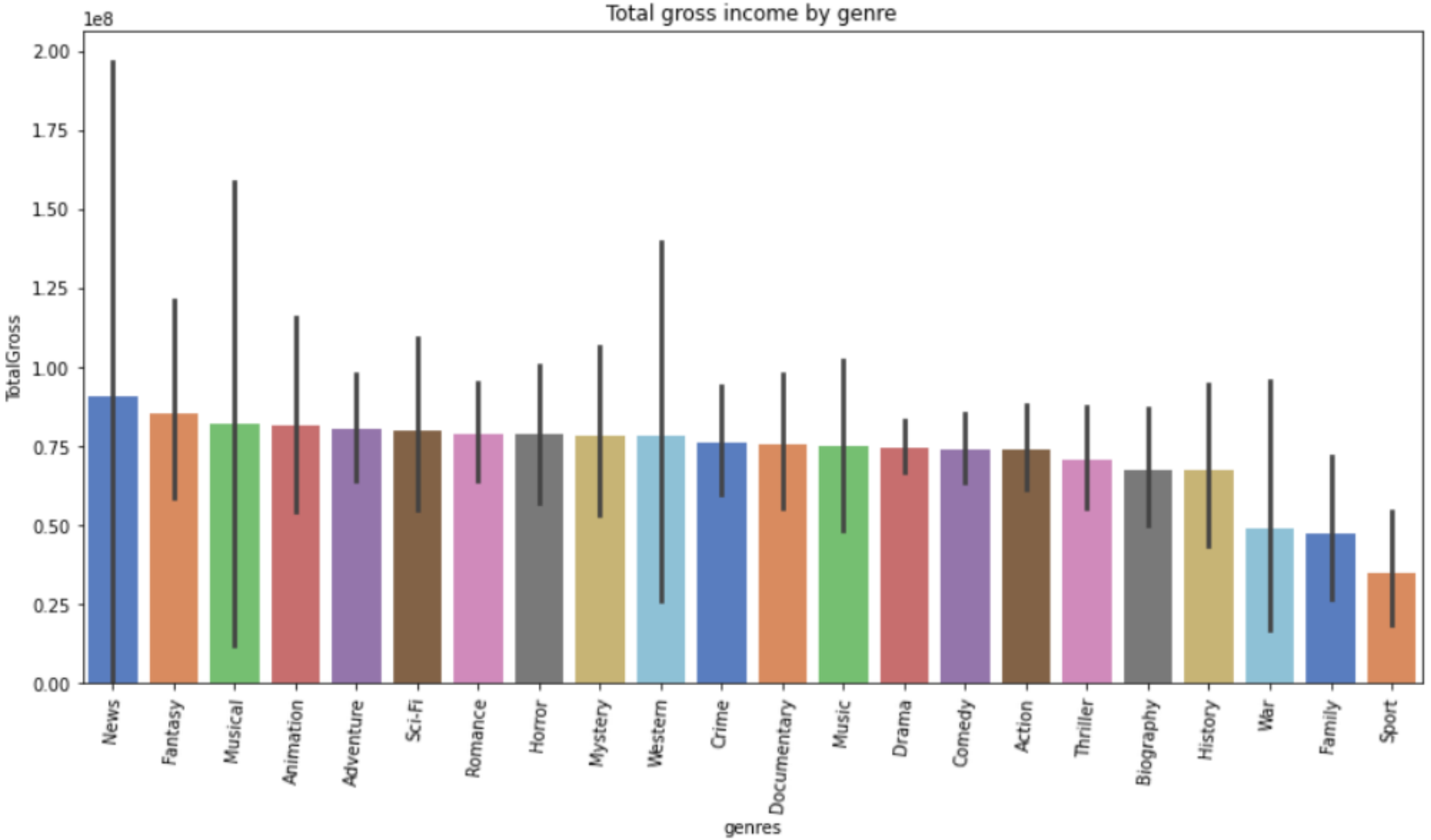
It's clear that drama and comedy genres have the highest number of movies released during this period.News,musical and western genres had the least release.

Data Analysis And Findings



There is no big difference between the highest rated genre and the least. Documentary & News genre have the highest rating.On the other side horror and Thriller are rated least.

Data Analysis And Findings



Conclusions And Recommendations

- ***Conclusion 1***

Documentary & News genres have the highest rating.On the other hand horror and Thriller are rated least

- ***Conclusion 2***

News,fantasy,animations seems to earn higher income compared to others.War,family and sport have the least gross income.

- ***Conclusion 3***

Income from domestic sales is less compared to foreign sales.

- ***Conclusion 4***

High rating movies earns more than less rated movies

Recomendations

1.Microsoft should consider News,fantasy and animation movies as the highest earners.

2.Microsoft should focus more on foreign market inorder to maximis