DETERMINING THE MOST APPRROPRIATE FILMS FOR A MOVIE STUDIO

Business Understanding

The problem question involves the use of variables to determine which types of films are appropriate for developing a movie studio. To investigate on the best films for a production company we aim towards working with various variables which include; types of films, consumer preferences and income generated. We aim towards understanding how these variables contribute to success in the production sector.

Data Understanding

```
#import pandas for data cleaning and manipulation
import pandas as pd
#import numpy for numerical operations
import numpy as np
# import seaborn and matplotlib for data visualization
import matplotlib.pyplot as plt
import seaborn as sns

#connecting to the SQL database using sqlite3
import sqlite3
conn = sqlite3.connect ("im.db")
cur = conn.cursor

# extracting the dataset to represent genres from the SQL data base
films = pd.read_sql ("SELECT* from movie_basics;", conn)
films.head()
```

```
    movie_id                    primary_title
original_title  \
0  tt0063540                        Sunghursh
Sunghursh
1  tt0066787  One Day Before the Rainy Season            Ashad Ka Ek
Din
2  tt0069049       The Other Side of the Wind  The Other Side of the
Wind
3  tt0069204                  Sabse Bada Sukh            Sabse Bada
Sukh
4  tt0100275         The Wandering Soap Opera       La Telenovela
Errante

   start_year  runtime_minutes                genres
0        2013            175.0    Action,Crime,Drama
1        2019            114.0       Biography,Drama
2        2018            122.0                 Drama
3        2018              NaN          Comedy,Drama
4        2017             80.0  Comedy,Drama,Fantasy
```

```
#SQL database to represent consumer preferences
preferences = pd.read_sql ("SELECT* from movie_ratings;", conn)
preferences
```

```
        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
...           ...            ...       ...
73851   tt9805820            8.1        25
73852   tt9844256            7.5        24
73853   tt9851050            4.7        14
73854   tt9886934            7.0         5
73855   tt9894098            6.3       128

[73856 rows x 3 columns]
```

```
income = pd.read_csv('zippedData/bom.movie_gross.csv.gz')
income.head()
```

```
                                         title studio  domestic_gross
\
0                                  Toy Story 3     BV     415000000.0

1                   Alice in Wonderland (2010)     BV     334200000.0

2  Harry Potter and the Deathly Hallows Part 1     WB     296000000.0

3                                    Inception     WB     292600000.0

4                          Shrek Forever After   P/DW     238700000.0


  foreign_gross  year
0     652000000  2010
1     691300000  2010
2     664300000  2010
3     535700000  2010
4     513900000  2010
```

Data preparation

Joining tables to obtain the required columns

```
#obtaining the required columns from the SQL database
combined = pd.read_sql("""
SELECT movie_basics.start_year, movie_basics.genres,
movie_ratings.numvotes
FROM movie_basics
```

```
INNER JOIN movie_ratings
ON movie_basics.movie_id = movie_ratings.movie_id
ORDER BY numvotes DESC
;
""",conn)
combined
```

```
       start_year                    genres  numvotes
0            2010  Action,Adventure,Sci-Fi   1841066
1            2012           Action,Thriller   1387769
2            2014   Adventure,Drama,Sci-Fi   1299334
3            2012             Drama,Western   1211405
4            2012  Action,Adventure,Sci-Fi   1183655
...           ...                       ...       ...
73851        2018                    Comedy         5
73852        2018             Comedy,Horror         5
73853        2019                   Romance         5
73854        2019               Documentary         5
73855        2019                      None         5

[73856 rows x 3 columns]
```

Data Description

The variables of interest to our study were selected to determine how they contribute to the success of a production sector. The types of films available was measured using genres, customer preferences was measured using number of votes and income generated by existing movie studios were measured using domestic gross.

Data Cleaning

```
#checking for missing values
income.isnull().sum()
```

```
title                0
studio               5
domestic_gross      28
foreign_gross     1350
year                 0
dtype: int64
```

Dealing with missing data

```
income =
income['domestic_gross'].fillna(income['domestic_gross'].median())
income
```

```
0       415000000.0
1       334200000.0
2       296000000.0
3       292600000.0
4       238700000.0
```

```
          ...
3382        6200.0
3383        4800.0
3384        2500.0
3385        2400.0
3386        1700.0
Name: domestic_gross, Length: 3387, dtype: float64
```

```
income = income['foreign_gross'].fillna(000000)
income
```

```
0       652000000
1       691300000
2       664300000
3       535700000
4       513900000
          ...
3382            0
3383            0
3384            0
3385            0
3386            0
Name: foreign_gross, Length: 3387, dtype: object
```

```
income = income['studio'].fillna('unknown')
income
```

```
0               BV
1               BV
2               WB
3               WB
4             P/DW
           ...
3382         Magn.
3383            FM
3384          Sony
3385    Synergetic
3386         Grav.
Name: studio, Length: 3387, dtype: object
```

```
#ensuring the missing values in income have been eliminated
income.isna()
```

```
0       False
1       False
2       False
3       False
4       False
        ...
3382    False
3383    False
```

```
3384    False
3385    False
3386    False
Name: studio, Length: 3387, dtype: bool
```

```
#checking for duplicates
income.duplicated().isna()
```

```
0       False
1       False
2       False
3       False
4       False
        ...
3382    False
3383    False
3384    False
3385    False
3386    False
Name: studio, Length: 3387, dtype: bool
```

Checking for missing values in our joined tables

```
#checking for missing values
combined.isnull().sum()
```

```
start_year      0
genres        804
numvotes        0
dtype: int64
```

```
#filling the missing values with 'unknown' string
combined = combined['genres'].fillna('unknown')
combined
```

```
0        Action,Adventure,Sci-Fi
1                Action,Thriller
2         Adventure,Drama,Sci-Fi
3                  Drama,Western
4        Action,Adventure,Sci-Fi
                  ...
73851                     Comedy
73852              Comedy,Horror
73853                    Romance
73854                Documentary
73855                    unknown
Name: genres, Length: 73856, dtype: object
```

```
#confirming that the missing values have been eliminated
combined.isna().sum()
```

```
0
```

```
#checking for duplicates
combined.duplicated().isna()
```

```
0        False
1        False
2        False
3        False
4        False
         ...
73851    False
73852    False
73853    False
73854    False
73855    False
Name: genres, Length: 73856, dtype: bool
```

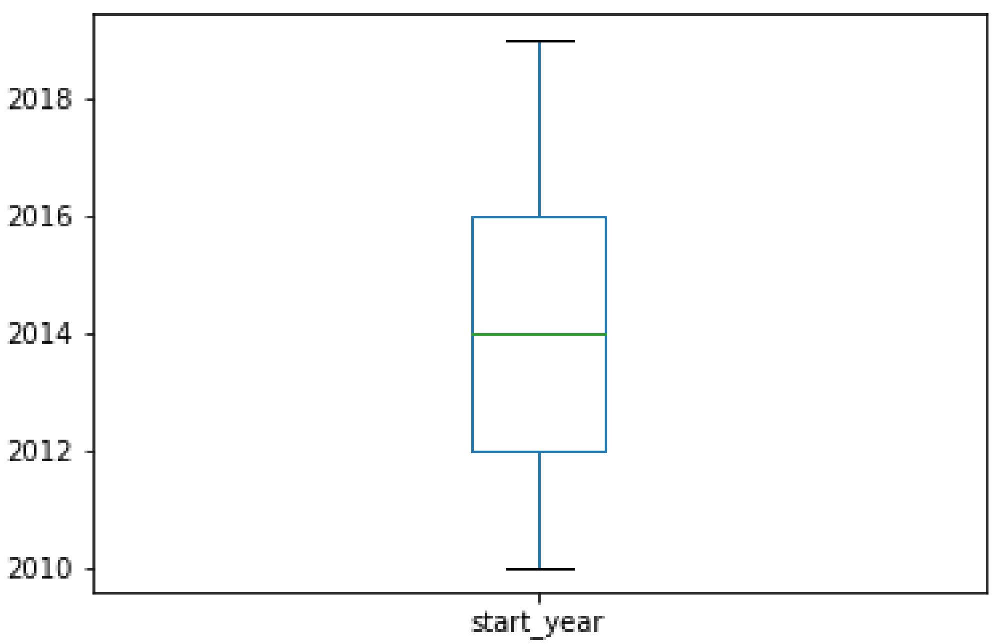
Data Analysis

Exploratory Data Analysis

Univariate Data Analysis

```
#plotting a boxplot on consumer preferences
combined.plot('numvotes', kind="box")
```

```
<AxesSubplot:>
```



The boxplot shows that there are different views between consumers on the types of films they prefer. This evidently proves that the different opinions between consumers is worth to be considered in our analysis. The boxplot also displays a symmetric hence normal distribution since the median lies in the middle and the whiskers are almost the same on both sides of the boxplot.

```
#obtaining basic summary statistics on consumer preferences
combined['numvotes'].describe()
```

```
count    7.385600e+04
mean     3.523662e+03
std      3.029402e+04
min      5.000000e+00
25%      1.400000e+01
50%      4.900000e+01
75%      2.820000e+02
max      1.841066e+06
Name: numvotes, dtype: float64
```

Consumer prefrences have a mean of 3.52 and a standard deviation of 3.03. This indiates that the data is reliable for our analysis since it is not highly spread out in reference to our standard deviation value.

```
income.info()
```

```
<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
```

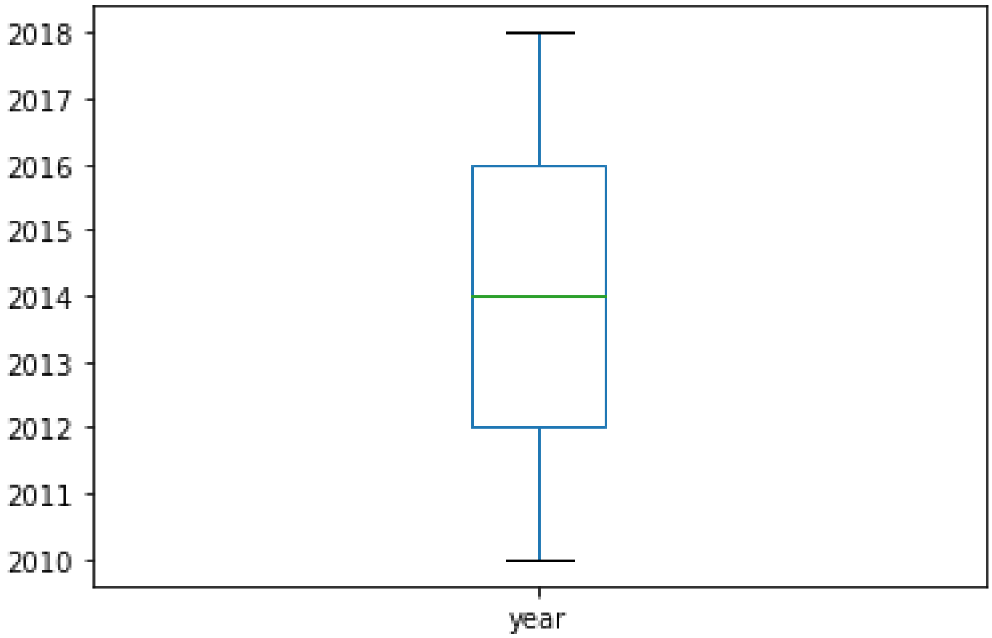
```
income['domestic_gross'].describe()
```

```
count    3.359000e+03
mean     2.874585e+07
std      6.698250e+07
min      1.000000e+02
25%      1.200000e+05
50%      1.400000e+06
75%      2.790000e+07
max      9.367000e+08
Name: domestic_gross, dtype: float64
```

Income generated has a mean of 2.87 and a standard deviation of 6.69. This indicates that the data is highly spread out in reference to our standard deviation of 6.69.

```
#plotting a boxplot on incomes generated by existing movie studios
income.plot('domestic_gross', kind = 'box')
```

```
<AxesSubplot:>
```



The box plot shows that there are different amounts earned by movie studios which mainly depends on the films produced. The whiskers are almost the same size on both sides of the box hence the distribution is normal. This is evidently supported by the median which lies in the middle of the boxplot.

Bivariate Data Analysis

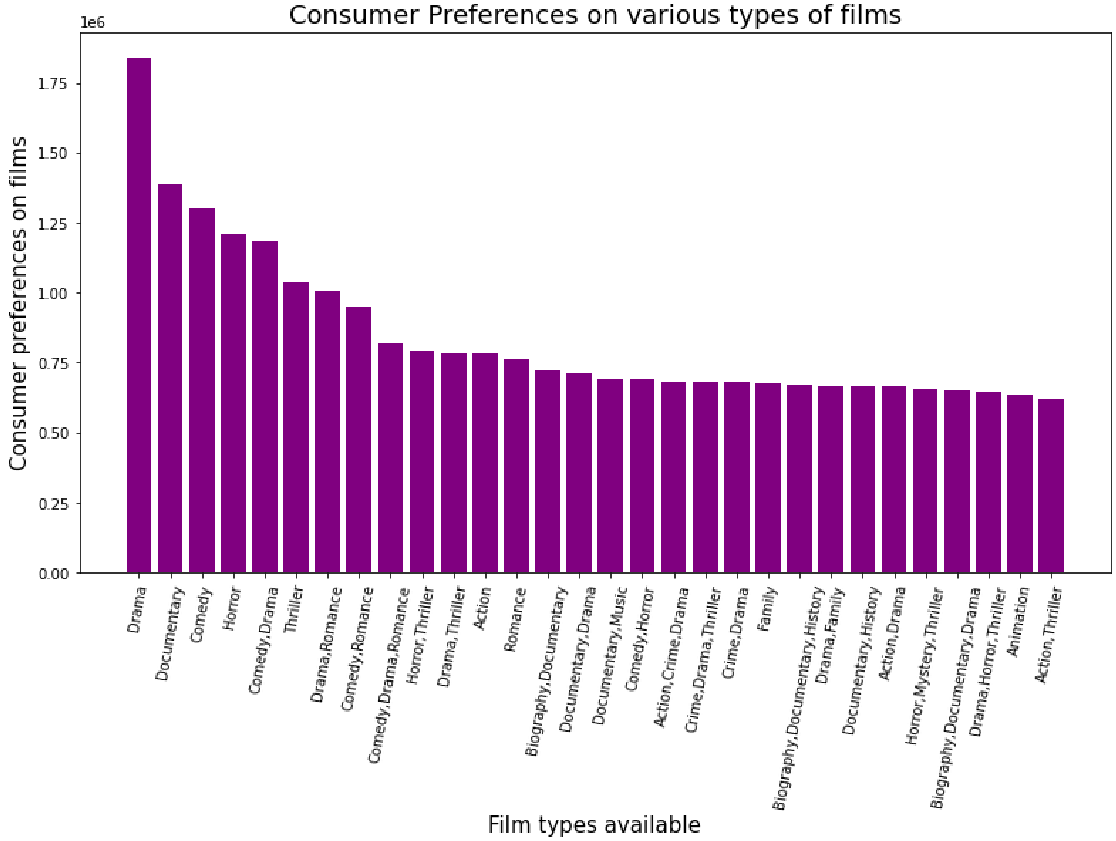
Plotting a graph to diplay the relationship between genres and number of votes

```
#setting the number of genres to 30
film_types = combined['genres'].value_counts().head(30)

#setting the number of votes to 30
film_preferences = combined['numvotes'].head(30)

#plotting the relationship between genres and number of votes
fig, ax = plt.subplots(figsize=(13, 7))
x= film_types.index
y= film_preferences.values
ax.bar(x, y , color= 'purple')
#labelling the axis
plt.xlabel('Film types available', fontsize=15)
plt.ylabel('Consumer preferences on films', fontsize=15)
#rotating the x axis
plt.xticks(rotation = '80')
#giving the graph a title
plt.title('Consumer Preferences on various types of films',
fontsize=18)
```

```
Text(0.5, 1.0, 'Consumer Preferences on various types of films')
```



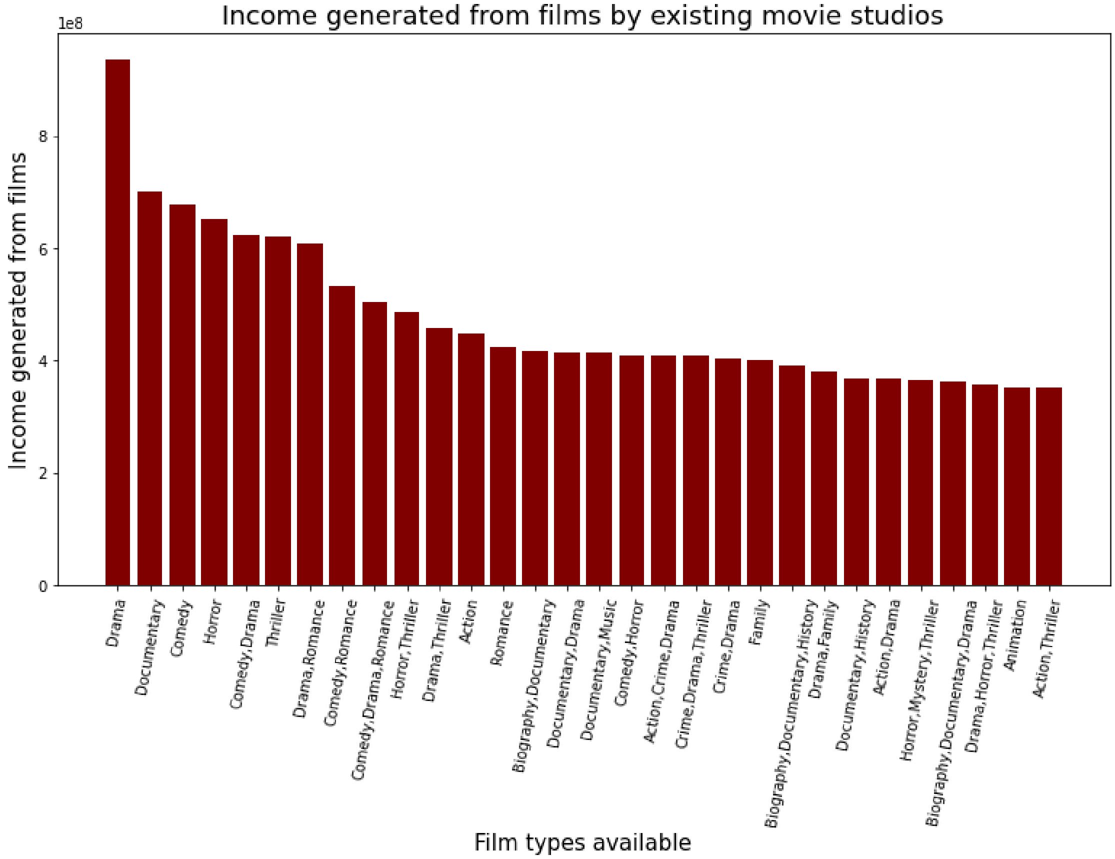
Films based on drama, documetary and comedy are the most preferred by consumers. A movie studio that majors in producing these films is therefore likely to thrive in the production sector. Films based on Action, thriller and animation are less likely preferred by consumers hence producing these films is likely to lead to losses.

Plotting a graph to display the realtionship between genres and income generated

```
generated =
income['domestic_gross'].sort_values(ascending=False).head(30)

fig, ax = plt.subplots(figsize=(13, 7))
x= film_types.index
y= generated.values
ax.bar(x, y , color= 'maroon')
#labelling the axis
plt.xlabel('Film types available', fontsize=15)
plt.ylabel('Income generated from films', fontsize=15)
#rotating the x axis
plt.xticks(rotation = '80')
#giving the graph a title
plt.title('Income generated from films by existing movie studios',
fontsize=18)
```

```
Text(0.5, 1.0, 'Income generated from films by existing movie studios')
```



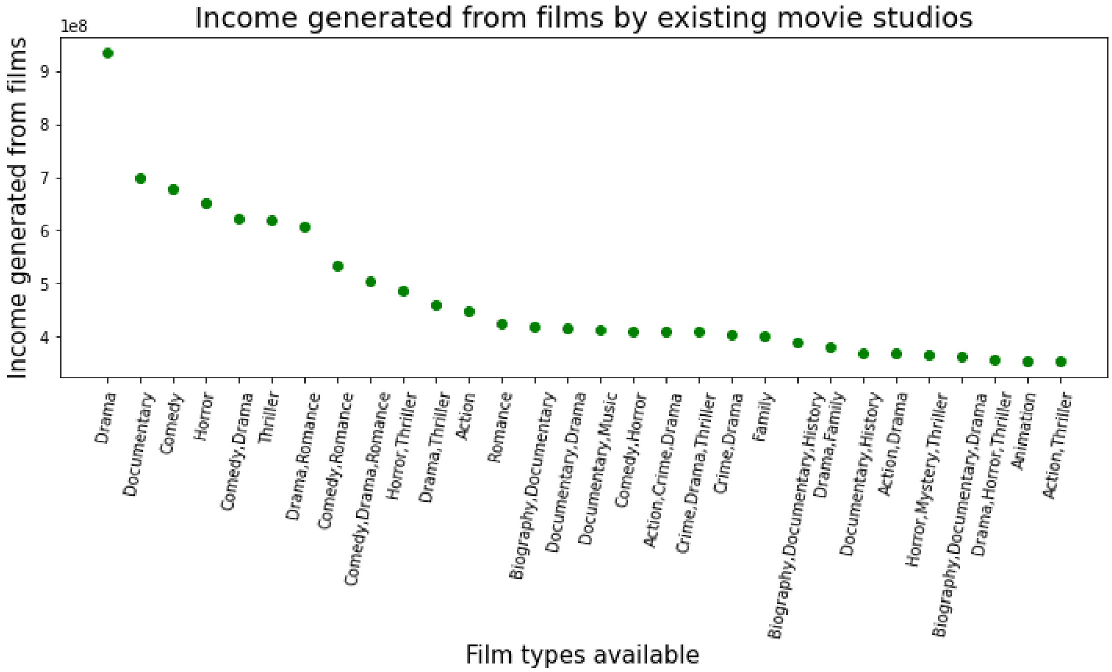
Existing movie studios that base their films on drama, documentary and comedy are generating huge incomes. This explains the association that consumers have in the production sector. The more a film is viewed, the more income for the production sector while the less a film is viewed, the lower the income generated.

Determining Correlation

Testing for correlation between types of films and income generated using scatter plots

```
fig, ax = plt.subplots(figsize=(12, 4))
x= film_types.index
y= generated.values
ax.scatter(x, y , color= 'green')
plt.xlabel('Film types available', fontsize=15)
plt.ylabel('Income generated from films', fontsize=15)
#rotating the x axis
plt.xticks(rotation = '80')
#giving the graph a title
plt.title('Income generated from films by existing movie studios',
fontsize=18)
```

```
Text(0.5, 1.0, 'Income generated from films by existing movie studios')
```

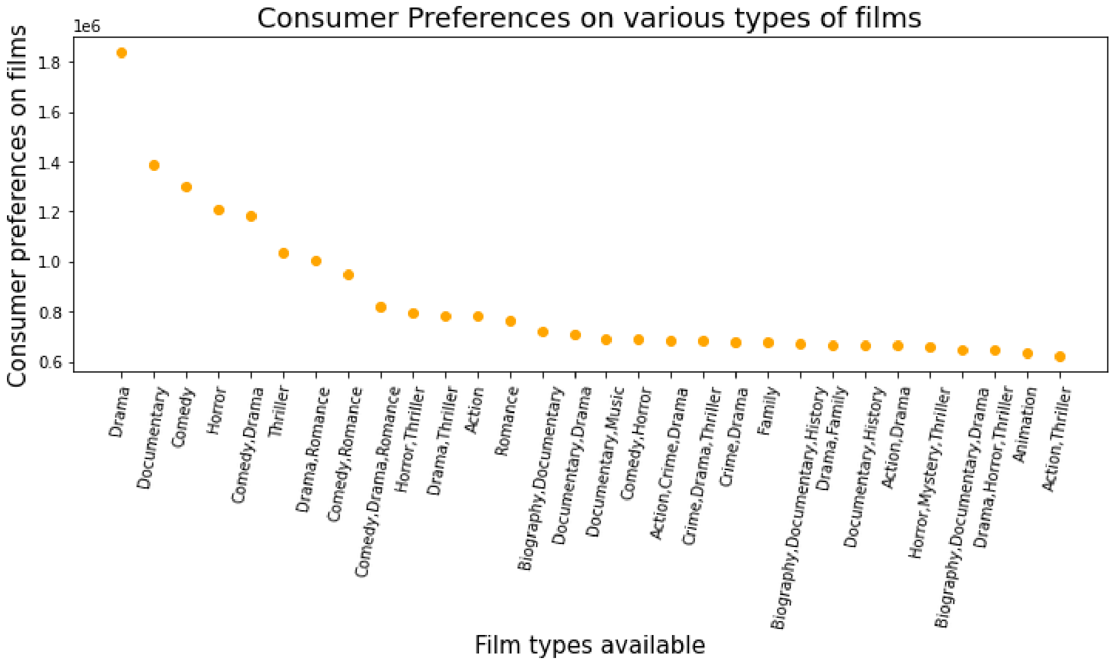


The scatter plot displays a perfect negative correlation between the types of films available and the income generated by movie studios.

Testing for correlation between types of films and number of votes using scatter plots

```
#plotting the relationship between genres and number of votes
fig, ax = plt.subplots(figsize=(12, 4))
x= film_types.index
y= film_preferences.values
ax.scatter(x, y , color= 'orange')
#labelling the axis
plt.xlabel('Film types available', fontsize=15)
plt.ylabel('Consumer preferences on films', fontsize=15)
#rotating the x axis
plt.xticks(rotation = '80')
#giving the graph a title
plt.title('Consumer Preferences on various types of films',
fontsize=18)
```

```
Text(0.5, 1.0, 'Consumer Preferences on various types of films')
```



The scatter plot displays a positive negative correlation between the types of films available and consumer preferences. This evidently shows that the higher a film is preferred, the higher the views gained while the less a film is preferred, the less the views.

Conclusion

This research shows that there exists a relationship between film types, consumer preferences and income generated by movie studios. Films based on drama, documentary and comedy are the best option for film production since existing movie studios are generating huge incomes through accumulated views by consumers. The higher the content from these films is viewed by consumers, the higher the income earned by production companies.

Recommendations

A movie studio should implement film production based on drama, documentary and comedy in order to generate high revenues since they are likely to gain views on their production. Consumers play a huge role in ensuring the success of movie studios hence regular surveys should be conducted to investigate on their interests in film content. The use of technology advanced devices when conducting film production to ensure the content displayed is of high quality.