Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Contents

INTRODUCTION 3

DATA PREPARATION 3

DATA CLEANING 3

MISSING DATA 3

DATA TYPE AND FORMAT 4

DATA STORIES AND VISUALISATIONS 4

INTRODUCTION

The dataset named ‘movies.csv’ includes a total of 20 columns, which are 'budget', 'genres', 'homepage', 'id', 'keywords', 'original_language', 'original_title', 'overview', 'popularity', 'production_companies', 'production_countries', 'release_date', 'revenue', 'runtime', 'spoken_languages', 'status', 'tagline', 'title', 'vote_average', and 'vote_count'. There are 4803 entries with a mixed datatype including int64, object, and float64.

Activities carried out in the exploratory project includes:

- Load the dataframe in
- Clean the data
- Remove duplicate rows
- Discard entries with a zero movie budget
- Manipulate certain columns to the correct data type
- Answer the questions about the data

DATA PREPARATION

DATA CLEANING

The following steps were carried out for ‘data cleaning’:

1. Identify columns that are redundant or unnecessary
   Based on the data set, it seems the following columns are not needed for the analysis and are removed from the data set: ['keywords', 'homepage', 'status', 'tagline', 'original_language', 'overview', 'production_companies', 'original_title'].
2. Remove duplicate rows
   Duplicate rows were removed using .drop_duplicates(), if any.

MISSING DATA

Some movies in the database have zero budget or zero revenue which implies that their values have not been recorded or some information is missing. Such entries from the dataframe were discarded, including:

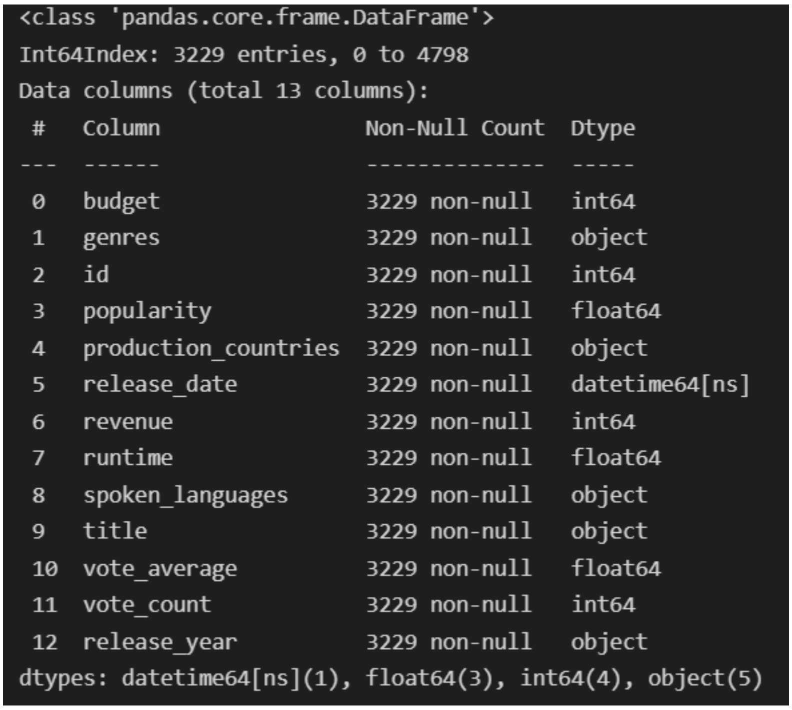
For column [‘budget’, ‘revenue’], relevant lines/records with a null value were dropped.

DATA TYPE AND FORMAT

In order to manipulate columns easily, it is important to make use of the python objects and make sure they are in a usable format. The following steps were carried out:

1. Through data set overview, it was noted that the ‘release data’ column is not in Date format. pd.to_datetime() method was applied to convert the column into correct date format. A new column named ‘release_year’ was also being extracted from ‘release_date’ and being added in the data set.
2. In addition to this, columns [‘budget’, ‘revenue’] were also converted into integer ‘int64’ for easier visualisation in the next steps.
3. Columns ‘[‘genres’, ‘spoken_languages’, ‘production_countries’] were in the JSON format. These have been flattened for easier interpretation in next steps for data analysis.

After the data preparation procedures, the final prepared data set include a total of 13 columns, detailed information is presented below:



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3229 entries, 0 to 4798
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   budget                3229 non-null   int64
 1   genres                3229 non-null   object
 2   id                    3229 non-null   int64
 3   popularity            3229 non-null   float64
 4   production_countries  3229 non-null   object
 5   release_date          3229 non-null   datetime64[ns]
 6   revenue               3229 non-null   int64
 7   runtime               3229 non-null   float64
 8   spoken_languages      3229 non-null   object
 9   title                 3229 non-null   object
 10  vote_average          3229 non-null   float64
 11  vote_count            3229 non-null   int64
 12  release_year          3229 non-null   object
dtypes: datetime64[ns](1), float64(3), int64(4), object(5)
```

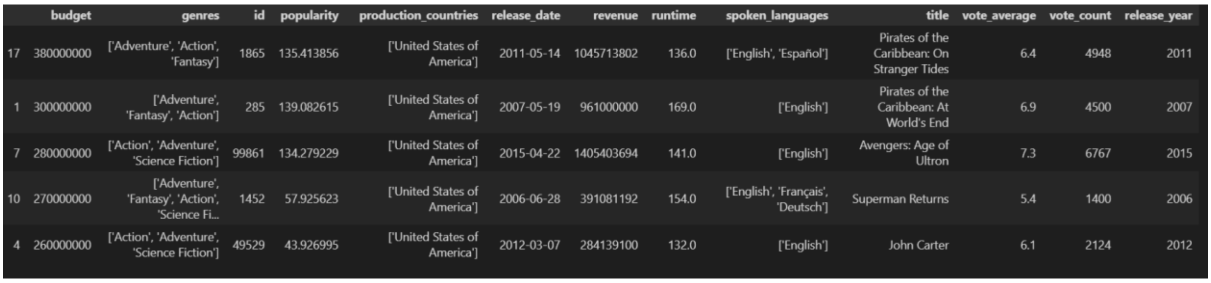
DATA STORIES AND VISUALISATIONS

Data visualisation and story-telling were carried out by identify of relationships between variables/features.

1. Which are the 5 most expensive movies? How do the most expensive and cheapest movies compare? Exploring the most expensive movies help you

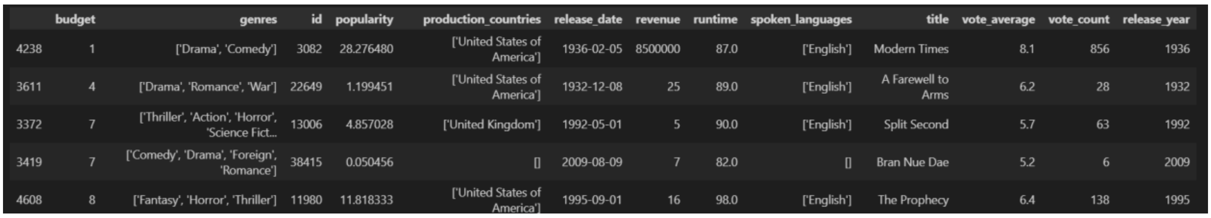
explore if some movies are worth the money spent on them based on their performance and revenue generated.

1) Information of the 5 most expensive movies:



	budget	genres	id	popularity	production_countries	release_date	revenue	runtime	spoken_languages	title	vote_average	vote_count	release_year
17	380000000	['Adventure', 'Action', 'Fantasy']	1865	135.413856	['United States of America']	2011-05-14	1045713802	136.0	['English', 'Español']	Pirates of the Caribbean: On Stranger Tides	6.4	4948	2011
1	300000000	['Adventure', 'Fantasy', 'Action']	285	139.082615	['United States of America']	2007-05-19	961000000	169.0	['English']	Pirates of the Caribbean: At World's End	6.9	4500	2007
7	280000000	['Action', 'Adventure', 'Science Fiction']	99861	134.279229	['United States of America']	2015-04-22	1405403694	141.0	['English']	Avengers: Age of Ultron	7.3	6767	2015
10	270000000	['Adventure', 'Fantasy', 'Action', 'Science Fi...	1452	57.925623	['United States of America']	2006-06-28	391081192	154.0	['English', 'Français', 'Deutsch']	Superman Returns	5.4	1400	2006
4	260000000	['Action', 'Adventure', 'Science Fiction']	49529	43.926995	['United States of America']	2012-03-07	284139100	132.0	['English']	John Carter	6.1	2124	2012

2) Information of the 5 cheapest movies:



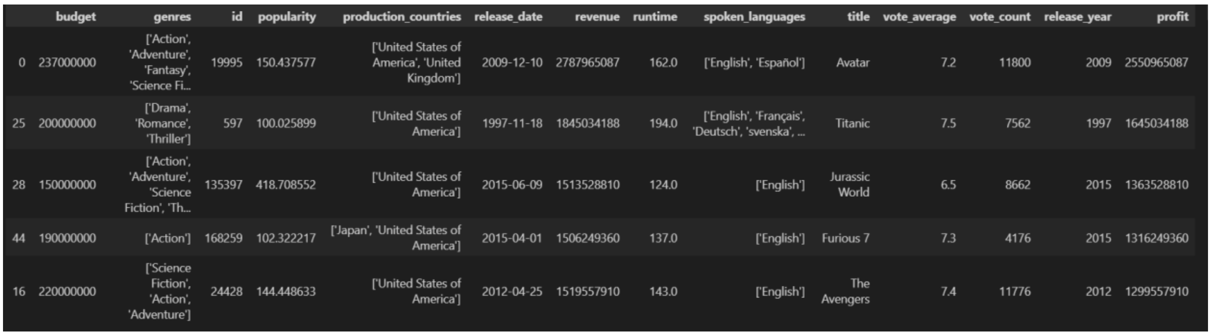
	budget	genres	id	popularity	production_countries	release_date	revenue	runtime	spoken_languages	title	vote_average	vote_count	release_year
4238	1	['Drama', 'Comedy']	3082	28.276480	['United States of America']	1936-02-05	8500000	87.0	['English']	Modern Times	8.1	856	1936
3611	4	['Drama', 'Romance', 'War']	22649	1.199451	['United States of America']	1932-12-08	25	89.0	['English']	A Farewell to Arms	6.2	28	1932
3372	7	['Thriller', 'Action', 'Horror', 'Science Fict...	13006	4.857028	['United Kingdom']	1992-05-01	5	90.0	['English']	Split Second	5.7	63	1992
3419	7	['Comedy', 'Drama', 'Foreign', 'Romance']	38415	0.050456	[]	2009-08-09	7	82.0	[]	Bran Nue Dae	5.2	6	2009
4608	8	['Fantasy', 'Horror', 'Thriller']	11980	11.818333	['United States of America']	1995-09-01	16	98.0	['English']	The Prophecy	6.4	138	1995

3) Comparison

   From above information, it seems that the most expensive movies do not always mean that they will have a vote_average and guaranteed good return. For the top 5 most expensive movies, though they all have some profit, the overall profitability((revenue-budget)/budget) is not as good as some of the cheaper movies. Movies with lower investment could also result in a good return in profit.

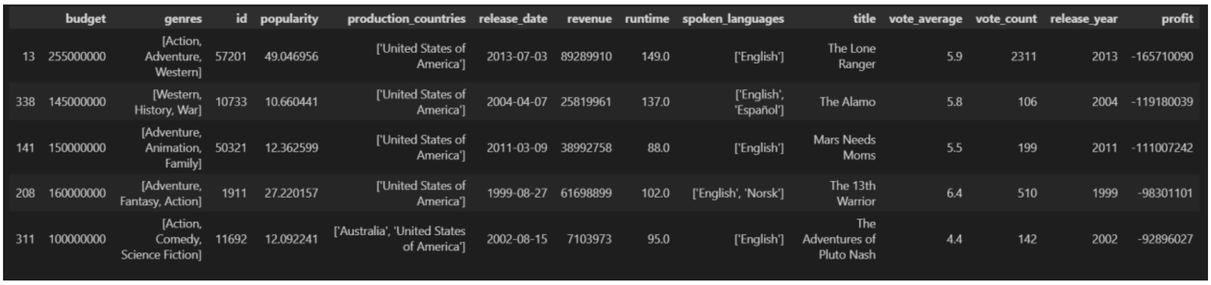
2. What are the top 5 most profitable movies? Compare the min and max profits. The comparison helps us identify the different approaches which failed and succeeded. Subtracting the budget from the revenue generated, will return the profit earned.

   1) In order to study how profitable a movie is, a new column named [‘profit’] was generated and added into the data set.
   2) The top 5 most profitable movies:



	budget	genres	id	popularity	production_countries	release_date	revenue	runtime	spoken_languages	title	vote_average	vote_count	release_year	profit
0	237000000	['Action', 'Adventure', 'Fantasy', 'Science Fi...	19995	150.437577	['United States of America', 'United Kingdom']	2009-12-10	2787965087	162.0	['English', 'Español']	Avatar	7.2	11800	2009	2550965087
25	200000000	['Drama', 'Romance', 'Thriller']	597	100.025899	['United States of America']	1997-11-18	1845034188	194.0	['English', 'Français', 'Deutsch', 'svenska', ...	Titanic	7.5	7562	1997	1645034188
28	150000000	['Action', 'Adventure', 'Science Fiction', 'Th...	135397	418.708552	['United States of America']	2015-06-09	1513528810	124.0	['English']	Jurassic World	6.5	8662	2015	1363528810
44	190000000	['Action']	168259	102.322217	['Japan', 'United States of America']	2015-04-01	1506249360	137.0	['English']	Furious 7	7.3	4176	2015	1316249360
16	220000000	['Science Fiction', 'Action', 'Adventure']	24428	144.448633	['United States of America']	2012-04-25	1519557910	143.0	['English']	The Avengers	7.4	11776	2012	1299557910

   3) The bottom least profitable movies:

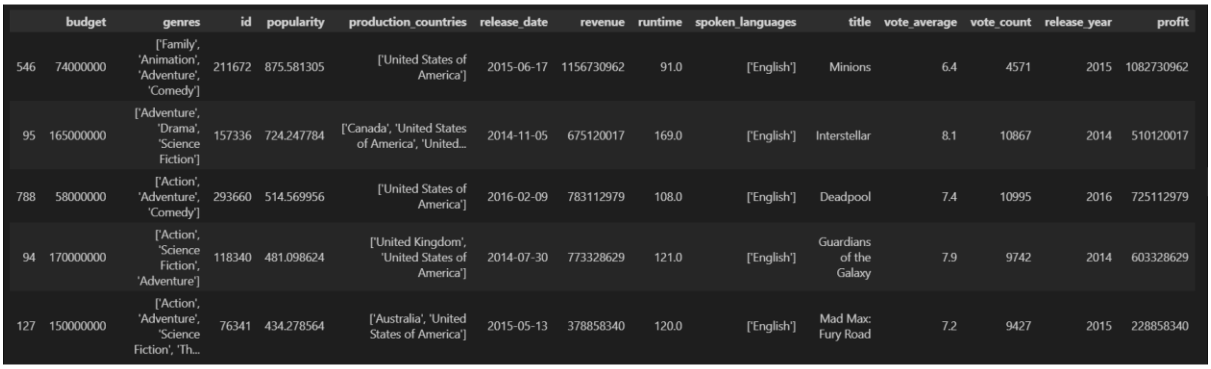


	budget	genres	id	popularity	production_countries	release_date	revenue	runtime	spoken_languages	title	vote_average	vote_count	release_year	profit
13	255000000	[Action, Adventure, Western]	57201	49.046956	['United States of America']	2013-07-03	89289910	149.0	['English']	The Lone Ranger	5.9	2311	2013	-165710090
338	145000000	[Western, History, War]	10733	10.660441	['United States of America']	2004-04-07	25819961	137.0	['English', 'Español']	The Alamo	5.8	106	2004	-119180039
141	150000000	[Adventure, Animation, Family]	50321	12.362599	['United States of America']	2011-03-09	38992758	88.0	['English']	Mars Needs Moms	5.5	199	2011	-111007242
208	160000000	[Adventure, Fantasy, Action]	1911	27.220157	['United States of America']	1999-08-27	61698899	102.0	['English', 'Norsk']	The 13th Warrior	6.4	510	1999	-98301101
311	100000000	[Action, Comedy, Science Fiction]	11692	12.092241	['Australia', 'United States of America']	2002-08-15	7103973	95.0	['English']	The Adventures of Pluto Nash	4.4	142	2002	-92896027

4) From above data, it seems that the most profitable movie(max) have very similar investment with the least profitable movie(min), which are 237,000,000 and 255,000,000. However, one has a profit of 2,550,965,087 and the other is -165,710,090. Both movies are under similar ‘genres’ type and very close runtime and are both available in English. It could also be told from the tables that their main differences are with columns ‘popularity’, ‘vote_average’, and ‘vote_count’. The most profitable movie(max) has much higher values of these features compared with the least profitable movie(min).

3. Find the most talked about movies.

1) This is defined based on the value of ‘popularity’ column, using .sort_values() method.
2) The most talked about movies:

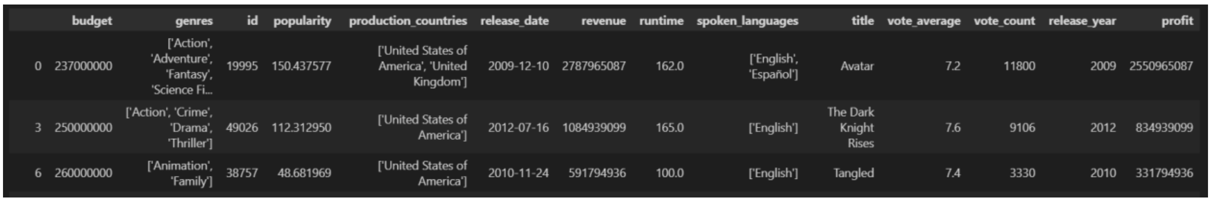


	budget	genres	id	popularity	production_countries	release_date	revenue	runtime	spoken_languages	title	vote_average	vote_count	release_year	profit
546	74000000	['Family', 'Animation', 'Adventure', 'Comedy']	211672	875.581305	['United States of America']	2015-06-17	1156730962	91.0	['English']	Minions	6.4	4571	2015	1082730962
95	165000000	['Adventure', 'Drama', 'Science Fiction']	157336	724.247784	['Canada', 'United States of America', 'United...	2014-11-05	675120017	169.0	['English']	Interstellar	8.1	10867	2014	510120017
788	58000000	['Action', 'Adventure', 'Comedy']	293660	514.569956	['United States of America']	2016-02-09	783112979	108.0	['English']	Deadpool	7.4	10995	2016	725112979
94	170000000	['Action', 'Science Fiction', 'Adventure']	118340	481.098624	['United Kingdom', 'United States of America']	2014-07-30	773328629	121.0	['English']	Guardians of the Galaxy	7.9	9742	2014	603328629
127	150000000	['Action', 'Adventure', 'Science Fiction', 'Th...	76341	434.278564	['Australia', 'United States of America']	2015-05-13	378858340	120.0	['English']	Mad Max: Fury Road	7.2	9427	2015	228858340

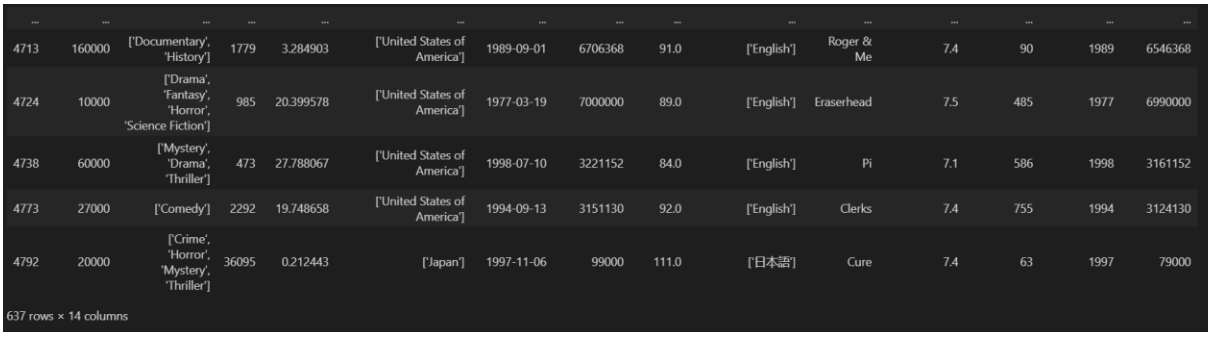
3) It can be seen from above table that the most talked about movie does not necessarily have the highest ‘vote_average’ and ‘vote_count’. But their ‘vote_average’ value is generally high and have positive profit.

4. Find movies which are rated above 7.

1) Use conditional method to select movies with rating higher than 7
2) Details shown below:



	budget	genres	id	popularity	production_countries	release_date	revenue	runtime	spoken_languages	title	vote_average	vote_count	release_year	profit
0	237000000	['Action', 'Adventure', 'Fantasy', 'Science Fi...	19995	150.437577	['United States of America', 'United Kingdom']	2009-12-10	2787965087	162.0	['English', 'Español']	Avatar	7.2	11800	2009	2550965087
3	250000000	['Action', 'Crime', 'Drama', 'Thriller']	49026	112.312950	['United States of America']	2012-07-16	1084939099	165.0	['English']	The Dark Knight Rises	7.6	9106	2012	834939099
6	260000000	['Animation', 'Family']	38757	48.681969	['United States of America']	2010-11-24	591794936	100.0	['English']	Tangled	7.4	3330	2010	331794936



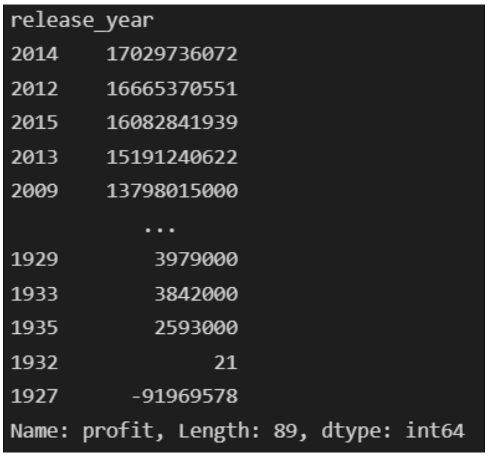
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4713	160000	['Documentary', 'History']	1779	3.284903	['United States of America']	1989-09-01	6706368	91.0	['English']	Roger & Me	7.4	90	1989	6546368
4724	10000	['Drama', 'Fantasy', 'Horror', 'Science Fiction']	985	20.399578	['United States of America']	1977-03-19	7000000	89.0	['English']	Eraserhead	7.5	485	1977	6990000
4738	60000	['Mystery', 'Drama', 'Thriller']	473	27.788067	['United States of America']	1998-07-10	3221152	84.0	['English']	Pi	7.1	586	1998	3161152
4773	27000	['Comedy']	2292	19.748658	['United States of America']	1994-09-13	3151130	92.0	['English']	Clerks	7.4	755	1994	3124130
4792	20000	['Crime', 'Horror', 'Mystery', 'Thriller']	36095	0.212443	['Japan']	1997-11-06	99000	111.0	['日本語']	Cure	7.4	63	1997	79000

637 rows × 14 columns

3) From results above, there are a total number of 637 movies with rating higher than 7 in the dataset.

5. Which year did we have the most profitable movies?

1) Use .groupby([‘release_year’]) method to sort [‘profile’] columns in decreasing order.
2) Results obtained is illustrated below:

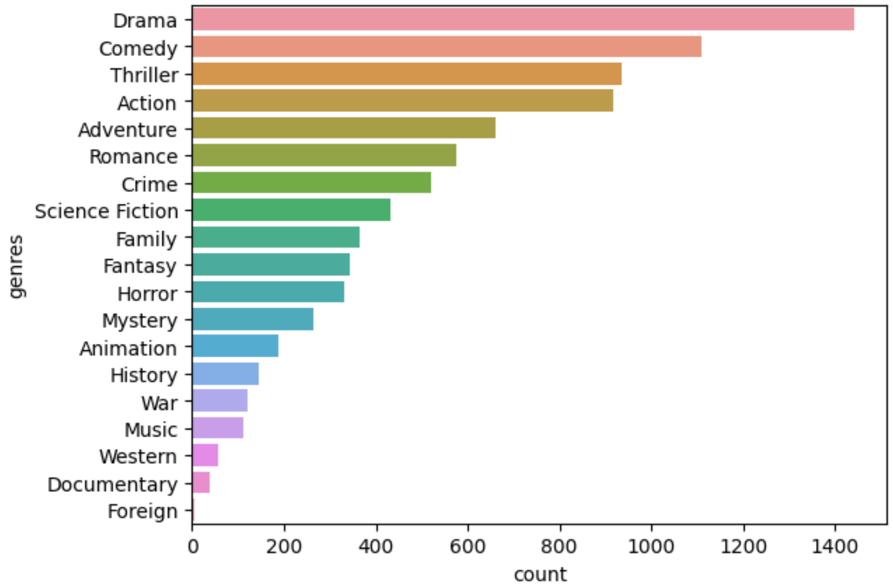


```
release_year
2014    17029736072
2012    16665370551
2015    16082841939
2013    15191240622
2009    13798015000
           ...
1929        3979000
1933        3842000
1935        2593000
1932             21
1927      -91969578
Name: profit, Length: 89, dtype: int64
```

3) From results above, it can be told that the most successful release year is in 2014, and the year 1927 was the least successful. There are probably more insights to draw from these data if knowing what has happened in the world and what movies are released in these years and try to correlate them. However, this will not be covered in this exploratory study.

6. What are the most successful genres?

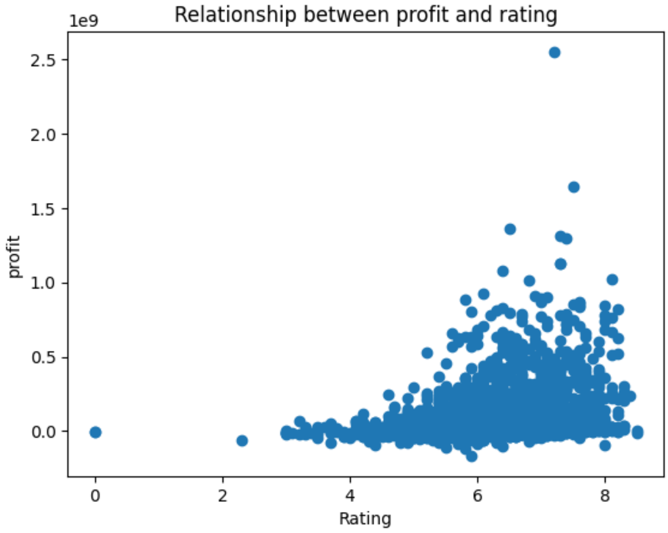
1) There could be multi-ways to interpret this. In this exploratory, this was explored by counting how many movies were released in total per genre type.
2) A bar plot was then created:



3) From above data, it can be found that the top 3 most successful genres are ‘Drama’, ‘Comedy’, ‘Thriller’ respectively.

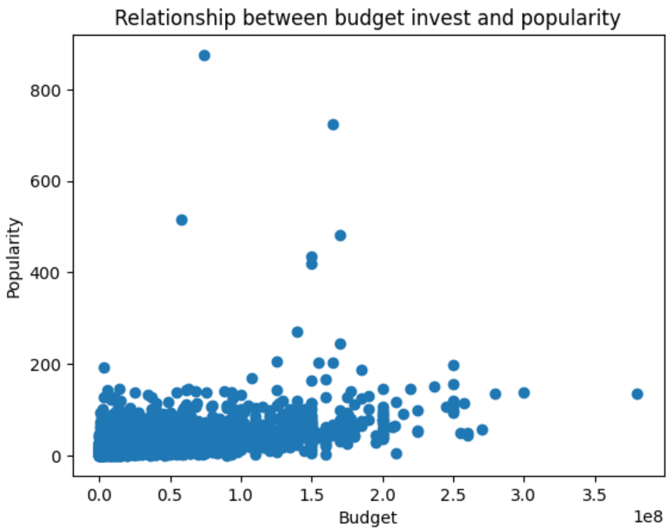
7. Three more interesting visualisations.

1) If the movie has high rating, is it always profitable?



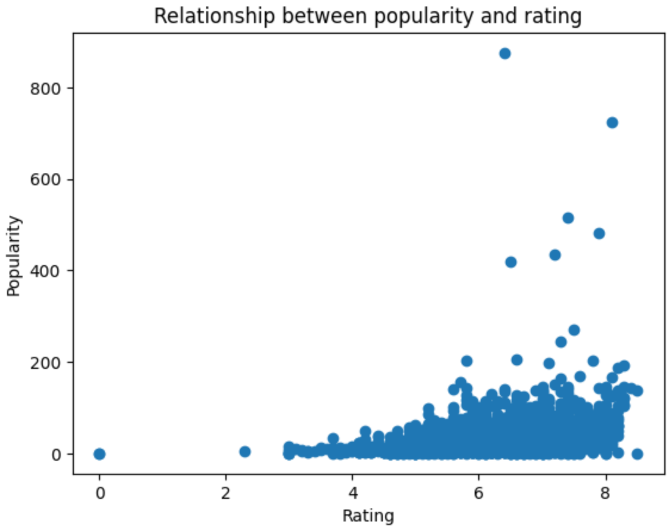
From the result, high rating may not always make the movie profitable. However, there is a trend that the higher rating the movie has, the more likely that it will be profitable. It can also be noticed from the graph that when rating is less than 4, the movie is generally not very much profitable.

2) If invest more money in a movie, will this make it to be more popular?



From the result, movie budget does not seem to have any significant influence on its popularity. Therefore, investing more money in a movie, does not mean that it will be successful.

3) If a movie has high rating, does this mean that it is very popular?



From the result, it seems that in general when a movie has higher rating, it is more likely to be popular. This might not be indefinite for high rating movies, but it can be noticed from the graph that movies with rating lower than 4 have very low popularity.

THIS REPORT WAS WRITTEN BY : Feifei Zhang