finalproject_faseela

December 20, 2022

1 Disney dataset exploratory analysis to find top genre and directors in terms of revenue

Author: Faseela Kuttikkattu Karim

1.1 Foreword

This notebook will be showing some exploratory data analysis for the Disney's Character Success dataset located here. Here I am analyzing the Disney's revenue dataset to find top directors and genres in terms of revenue they have generated.

2 Introduction

2.1 Question(s) of interests

In this analysis, I will be investigating a question associated with the collection `Disney's Character Success` datasets. **Question:** I am interested in finding out which genres and directors created maximum revenue for Disney based on the available information in the given datasets. I would expect the **Comedy** movies generate maximum revenue. The reason because I am exploring this particular question is because I am curious about knowing what kind of movies people are interested in and generating revenue.

2.2 Dataset description

- **disney_movies_total_gross.csv**
  - This file contains information about Disney's gross revenue for different movies and genre of each movies.
- **disney-director.csv**
  - This file includes information about movies and directors.

3 Methods and Results

Since I am only interested in revenue generated by different genre movies and directors I need to use tables that contain information on disney movies total revenue and disney directors. This implies that I need to use the **disney_movies_total_gross** and the **disney-director** tables. First I will pull the data and convert them as dataframes. Then I will wrangle and clean the dataset. Importantly I need to convert total_gross column from `object` type to `integer` inorder to find the sum. Then sort the table and plot the graph. To find the top 10 directors I need to merge

disney-director and disney_movies_total_gross datasets. For that I need to create an external function and save it in another `.py` file, call it when required. Then I need to plot a bar graph showing top 10 directors and revenue generated by them. I will make a function for testing my external function by creating another `.py` file.

However, before moving further, let us import the tables and do some basic visualizations.

[1]:
```
# Lets import all the required libraries needed for this analysis
import altair as alt
import pandas as pd

# import all the required files
movie_revenue = pd.read_csv("data/disney_movies_total_gross.csv")
directors = pd.read_csv("data/disney-director.csv")
```

Lets see what the tables look like.

[2]:
```
movie_revenue.head()
```

[2]:

	movie_title	release_date	genre	MPAA_rating	total_gross	inflation_adjusted_gross
0	Snow White and the Seven Dwarfs	Dec 21, 1937	Musical	G	$184,925,485	$5,228,953,251
1	Pinocchio	Feb 9, 1940	Adventure	G	$84,300,000	$2,188,229,052
2	Fantasia	Nov 13, 1940	Musical	G	$83,320,000	$2,187,090,808
3	Song of the South	Nov 12, 1946	Adventure	G	$65,000,000	$1,078,510,579
4	Cinderella	Feb 15, 1950	Drama	G	$85,000,000	$920,608,730

[3]:
```
directors.head()
```

[3]:

	name	director
0	Snow White and the Seven Dwarfs	David Hand
1	Pinocchio	Ben Sharpsteen
2	Fantasia	full credits
3	Dumbo	Ben Sharpsteen
4	Bambi	David Hand

Lets get some information about the tables.

[4]:
```
movie_revenue.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 579 entries, 0 to 578
Data columns (total 6 columns):
```

```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   movie_title              579 non-null    object
 1   release_date             579 non-null    object
 2   genre                    562 non-null    object
 3   MPAA_rating              523 non-null    object
 4   total_gross              579 non-null    object
 5   inflation_adjusted_gross  579 non-null    object
dtypes: object(6)
memory usage: 27.3+ KB
```

The movie_revenue table has 579 rows and 6 columns. Every movie_title has a release date,total gross. But genre is available only for 562 movies.

[5]:
```
directors.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56 entries, 0 to 55
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   name      56 non-null     object
 1   director  56 non-null     object
dtypes: object(2)
memory usage: 1.0+ KB
```

The directors table has 56 rows and 2 columns. Every movie name provided with director name.

[6]:
```
directors_df=pd.DataFrame(directors)
```

Data cleaning step to Remove the movies without genre

[7]:
```
movies_with_genre=movie_revenue.dropna(subset=["genre"])
movies_with_genre
```

[7]:
```
                         movie_title  release_date      genre MPAA_rating  \
0    Snow White and the Seven Dwarfs  Dec 21, 1937    Musical           G
1                          Pinocchio   Feb 9, 1940  Adventure           G
2                           Fantasia  Nov 13, 1940    Musical           G
3                   Song of the South  Nov 12, 1946  Adventure           G
4                          Cinderella  Feb 15, 1950      Drama           G
..                                ...           ...       ...         ...
574         The Light Between Oceans   Sep 2, 2016      Drama       PG-13
575                   Queen of Katwe  Sep 23, 2016      Drama          PG
576                   Doctor Strange   Nov 4, 2016  Adventure       PG-13
577                            Moana  Nov 23, 2016  Adventure          PG
578     Rogue One: A Star Wars Story  Dec 16, 2016  Adventure       PG-13

      total_gross inflation_adjusted_gross
```

```
0    $184,925,485           $5,228,953,251
1     $84,300,000           $2,188,229,052
2     $83,320,000           $2,187,090,808
3     $65,000,000           $1,078,510,579
4     $85,000,000             $920,608,730
..            …                      …
574   $12,545,979            $12,545,979
575    $8,874,389             $8,874,389
576  $232,532,923           $232,532,923
577  $246,082,029           $246,082,029
578  $529,483,936           $529,483,936

[562 rows x 6 columns]
```

Data wrangling step to convert total_gross column into integer type and grouped the movies based on their genre and found out total revenue of each genre.

[8]:
```
moviesgenre_data = pd.DataFrame(movies_with_genre)
moviesgenre_data=moviesgenre_data.astype({'total_gross': 'string'})
total=moviesgenre_data['total_gross'].str.split('$',n=1,expand=True)
moviesgenre_data['total_gross']=total[1]
moviesgenre_data
moviesgenre_data1=moviesgenre_data
moviesgenre_data2=moviesgenre_data
moviesgenre_data1['total_gross']=moviesgenre_data['total_gross'].str.
 ↪replace(',','')
moviesgenre_data2['total_gross']=moviesgenre_data1['total_gross'].astype(int)
moviesgenre_totalrevenue=moviesgenre_data2.groupby('genre')['total_gross'].
 ↪sum().reset_index().sort_values('total_gross',ascending=False)
moviesgenre_totalrevenue
```

[8]:
```
                  genre  total_gross
1             Adventure  16389069453
3                Comedy   8119619678
0                Action   4184563282
6                 Drama   4106972970
10    Thriller/Suspense   1406806519
8               Musical   1157284155
9       Romantic Comedy   1152206855
11              Western    359011459
5           Documentary    180685619
4   Concert/Performance    103456466
2          Black Comedy     97543212
7                Horror     87068872
```

Now that we have it in the proper format, we can generate a bar plot to visualize it.

[9]:
```
# Use altair to generate a bar plot
genre_revenue_plot = (
    alt.Chart(moviesgenre_totalrevenue, width=500, height=300)
    .mark_bar()
    .encode(
        x=alt.X("genre", title="Genre",sort='-y'),
        y=alt.Y("total_gross", title="total_gross"),
    )
    .properties(title="Revenue generated by movies of different genre")
)
genre_revenue_plot
```

[9]:
```
alt.Chart(…)
```

The graph shows the revenue from highest to lowest. Maximum revenue generated by adventure movies and least revenue was from horror movies when we consider the data from 1937 till 2016.

Now lets try to find out top 10 directors based on the revenue generated by their movies,to do this, I will import and use the script I created with a custom function that takes in two dataframes directors_df and moviesgenre_data2 which contains information about revenue and directors respectively, and merge these two by a 'movie_title' column.

[10]:
```
import table_merge as ps
directors_df=directors_df.assign(movie_title=directors_df['name'])
merged_data=ps.merge_table(directors_df,moviesgenre_data2,"movie_title")
merged_data=merged_data.drop(columns=['name'])
merged_data=merged_data.dropna(subset=["total_gross"])
merged_director_revenue=merged_data.groupby('director')['total_gross'].sum().
 ↪reset_index().sort_values('total_gross',ascending=False)
merged_director_revenue.head(10)
```

[10]:

	director	total_gross
25	Wolfgang Reitherman	966009582.0
20	Ron Clements	840214815.0
4	Chris Buck	571829828.0
19	Roger Allers	422780140.0
10	Gary Trousdale	403143238.0
7	Clyde Geronimi	343655718.0
3	Byron Howard	341268248.0
23	Wilfred Jackson	286151353.0
13	Mark Dindal	224683238.0
9	Don Hall	222527828.0

Now that we have it in the proper format, we can generate a bar plot to visualize it.

[11]:
```
# Use altair to generate a bar plot
director_revenue_plot = (
    alt.Chart(merged_director_revenue.head(10), width=500, height=300)
```

```
    .mark_bar()
    .encode(
        x=alt.X("director", title="Director",sort='-y'),
        y=alt.Y("total_gross", title="total_gross"),
    )
    .properties(title="Top 10 Directors in terms of revenue generated by their␣
 ↪movies")
)
director_revenue_plot
```

[11]:
```
alt.Chart(…)
```

The graph shows the revenue from highest to lowest. Maximum revenue generated by movies directed by Mr.Wolfgang Reitherman and Mr.Don Hall is in the 10th position.

Sample data creation and testing is done on the script as shown below

[12]:
```
import test_table_merge as pt
pt.test_merge_table()
```

4 Discussions

In this work, I analyzed Disney's Character Success dataset(https://data.world/kgarrett/disney-character-success-00-16) and tried to compute which genere of movies generated the most revenue and which directors were successful in generating the most revenue. `disney_movies_total_gross.csv`, `disney-director.csv` are the two tables I used out of five tables provided in the Disney's dataset. Before answering the main question, I did some exploratory data analysis to see key information about the data and the datatype of different colums. I found that the gross revenue for each movie is given and there was some 'NaN' values, so cleaned the table to remove NaN value. But the important point was all the numbers in `string` datattype and it would be difficult to use it as it is. After the initial analysis I came to know that data cleaning is required and some wrangling steps are necessary to further process the data. For the `disney-director` table as well there were some movies which was not a part of the `disney_movies_total_gross.csv` table. So table merging and further cleaning was required to be able to do my second part of question. Once I made the properly wrangled and cleaned data I went ahead with analysing the genre of movies which generated maximum revenue. It was quite surpricing to know that my initial assumptions about `comedy` genre was wrong and I gained much more insights after this exploratory analysis. I found that `adventure` movies generated most revenue (**16.3 Billion**) and `comedy` comes after that(**8.1 Billion**).`Action` and `drama` comes after that (**8.1 Billion**), and both genres generates almost equal revenue for Disney. Horror movies only have a (**87 million**) revenue wchich comes last.

It is quite interesting to find that **Wolfgang Reitherman** is the director with the most gross earning movie in the given list. The earning from his direction is almost a billion(**.966 Billion**). However, **Ron Clements** the second director in the list made about (**.84 Billion**), which is also a great earning and which is expected considering he is the director of popular movies like The `Little Mermaid (1989)`, `Aladdin (1992)` and `Hercules (1997)`. The last one in the top 10 list is **Don Hall** with (**.22 Billion**).

Another question that would be interesting for the given dataset is the change in the gross revenues of movies over the years. This is interesting because Disney has so many competitors over the years and new movie platforms.

5 References

The data is taken is from https://data.world/kgarrett/disney-character-success-00-16, also checked about directors and movies in wikipedia. Used pandas documentation for reference https://pandas.pydata.org/docs/reference/

5.1 Resources used

- [Data Source] https://data.world/kgarrett/disney-character-success-00-16