final_project_submission_suvinmajithia

December 12, 2023

0.0.1 Disney Movies: A Data-Driven Exploration of Movie Revenue and Genre Dynamics

Foreword

The objective of this notebook is conducting some data analysis for the Disney dataset located here. Here I am analyzing the Disney dataset to scrutinize the correlation between movie revenues and their genres. Through application of Python scripts, unit tests, and the principles of reproducibility, this report offers an interesting exploration of my findings.

0.0.2 Introduction

0.0.3 Question(s) of interests

In this analysis, I will be solving a question about which movie genre has generated more gross revenue for Disney. I am also interested in finding out which genre has the most impact associated with it and which genre has more number of movies produced by Disney. This is interesting because the Disney movies are based on various themes. I would expect the 'Comedy' genre to have the most impact overall.

0.0.4 Dataset description

The below datasets were taken directly from this website . The Walt Disney Company, commonly known as Disney, is an American multinational mass media and entertainment conglomerate that is headquartered at the Walt Disney Studios complex in Burbank, California. The Disney dataset is composed of 5 tables, disney-characters.csv, disney-director.csv, disney-voice-actors.csv, disney_revenue_1991-2016.csv and disney_movies_total_gross.csv that contains information about different Disney characters, Disney movies directors, Disney movie characters voice artists, annual gross revenue of the Disney company and the total gross and inflation adjusted gross revenue generated by different Disney movies. I will be using the disney_movies_total_gross tables as formally described below: disney_movies_total_gross.csv This file contains information on the movie title, release date, MPAA rating, genre, total gross revenue and inflation adjusted gross revenue of the Disney movies.

0.0.5 Methods and Results

Since I am only interested in computing the genre and its impact based on revenue and other factors, I will need to use the table that contains information on genre and inflation adjusted gross revenue. This implies that I will need to use the disney_movies_total_gross table.

However, firstly, let us import the tables and do some basic visualizations.

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

# Import all the required 5 disney tables/files
movie_total_data = pd.read_csv("data/disney_movies_total_gross.csv")
revenue_data = pd.read_csv("data/disney_revenue_1991-2016.csv")
characters_data = pd.read_csv("data/disney-characters.csv")
director_data = pd.read_csv("data/disney-director.csv")
voice_actors_data = pd.read_csv("data/disney-voice-actors.csv")
```

Lets see what all the tables look like.

[2]:
```python
# Checking the first few rows of all the tables
movie_total_data.head()
```

[2]:
```
                        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

    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
```

[3]:
```python
revenue_data.head()
```

[3]:
```
   Year  Studio Entertainment[NI 1]  Disney Consumer Products[NI 2]  \
0  1991                      2593.0                           724.0
1  1992                      3115.0                          1081.0
2  1993                      3673.4                          1415.1
3  1994                      4793.0                          1798.2
4  1995                      6001.5                          2150.0

   Disney Interactive[NI 3][Rev 1]  Walt Disney Parks and Resorts  \
0                              NaN                         2794.0
1                              NaN                         3306.0
2                              NaN                         3440.7
3                              NaN                         3463.6
4                              NaN                         3959.8
```

```
  Disney Media Networks  Total
0                   NaN   6111
1                   NaN   7502
2                   NaN   8529
3                   359  10414
4                   414  12525
```

[4]:
```
characters_data.head()
```

[4]:
```
                      movie_title       release_date        hero  \
0  \nSnow White and the Seven Dwarfs  December 21, 1937  Snow White
1                      \nPinocchio   February 7, 1940   Pinocchio
2                       \nFantasia  November 13, 1940         NaN
3                           Dumbo   October 23, 1941       Dumbo
4                         \nBambi    August 13, 1942       Bambi

     villian                        song
0  Evil Queen  Some Day My Prince Will Come
1   Stromboli      When You Wish upon a Star
2  Chernabog                          NaN
3  Ringmaster                    Baby Mine
4      Hunter               Love Is a Song
```

[5]:
```
director_data.head()
```

[5]:
```
                              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
```

[6]:
```
voice_actors_data.head()
```

[6]:
```
         character      voice-actor                         movie
0     Abby Mallard      Joan Cusack                Chicken Little
1  Abigail Gabble     Monica Evans                The Aristocats
2         Abis Mal  Jason Alexander          The Return of Jafar
3              Abu     Frank Welker                       Aladdin
4         Achilles             None  The Hunchback of Notre Dame
```

Lets get some other information about the disney_movies_total_gross.csv table.

[7]:
```
movie_total_data.info()
movie_total_data['inflation_adjusted_gross'] =␣
↪movie_total_data['inflation_adjusted_gross'].str.
↪replace(r'\D','',regex=True).astype(float)
```

```
<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
```

Our disney_movies_total_gross has some null values in the genre column so let's explore them in detail.

[8]:
```
# Some of the genre data has NA values...we need to deep dive in it
movie_total_data[movie_total_data[['genre']].isna().any(axis=1)]
```

[8]:
```
                                 movie_title  release_date genre MPAA_rating  \
20   The Many Adventures of Winnie the Pooh  Mar 11, 1977   NaN         NaN   
22                 Herbie Goes to Monte Carlo  Jun 24, 1977   NaN         NaN   
23                             The Black Hole  Dec 21, 1979   NaN         NaN   
24                           Midnight Madness   Feb 8, 1980   NaN         NaN   
25               The Last Flight of Noah’s Ark  Jun 25, 1980   NaN         NaN   
26                   The Devil and Max Devlin   Jan 1, 1981   NaN         NaN   
121                                   Newsies   Apr 8, 1992   NaN          PG   
122                               Passed Away  Apr 24, 1992   NaN       PG-13   
128               A Gun in Betty Lou's Handbag  Aug 21, 1992   NaN       PG-13   
146                            Bound by Honor  Apr 16, 1993   NaN           R   
155                       My Boyfriend's Back   Aug 6, 1993   NaN       PG-13   
156                               Father Hood  Aug 27, 1993   NaN       PG-13   
168                            Red Rock West  Jan 28, 1994   NaN           R   
251                           The War at Home  Nov 20, 1996   NaN           R   
304                                 Endurance  May 14, 1999   NaN          PG   
350                   High Heels and Low Lifes  Oct 26, 2001   NaN           R   
355                       Frank McKlusky C.I.   Jan 1, 2002   NaN         NaN   

     total_gross  inflation_adjusted_gross  
20            $0                       0.0  
22   $28,000,000               105847527.0  
23   $35,841,901               120377374.0  
24    $2,900,000                 9088096.0  
25   $11,000,000                34472116.0  
26   $16,000,000                48517980.0  
121   $2,706,352                 5497481.0  
122   $4,030,793                 8187848.0  
```

```
128   $3,591,460                  7295423.0
146   $4,496,583                  9156084.0
155   $3,218,882                  6554384.0
156   $3,268,203                  6654819.0
168   $2,502,551                  5170709.0
251      $34,368                    65543.0
304     $229,128                   380218.0
350     $226,792                   337782.0
355           $0                        0.0
```

Now replacing the null values with the actual 'genre' of the movies, we can ignore the movies which have less than 10 million+ in revenue for our analysis, we will replace the null values with specified genre.

[9]:
```
# Herbie Goes to Monte Carlo  - Action genre
movie_total_data.loc[movie_total_data['movie_title']=='Herbie Goes to Monte␣
↪Carlo','genre'] = 'Action'
# The Black Hole - Action genre
movie_total_data.loc[movie_total_data['movie_title']=='The Black Hole','genre']␣
↪= 'Action'
# The Last Flight of Noah's Ark -  Adventure genre
movie_total_data.loc[movie_total_data['movie_title']=='The Last Flight of␣
↪Noah's Ark','genre'] = 'Adventure'
# The Devil and Max Devlin - Comedy genre
movie_total_data.loc[movie_total_data['movie_title']=='The Devil and Max␣
↪Devlin','genre'] = 'Comedy'


# dropping NaN genre values from the dataframe
movie_total_data= movie_total_data.dropna(subset=['genre'])
movie_total_data = movie_total_data.reset_index()
```

Checking for more information on the disney_movies_total_gross table after replacing null values with the specified values.

[10]:
```
movie_total_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 566 entries, 0 to 565
Data columns (total 7 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   index                     566 non-null     int64
 1   movie_title               566 non-null     object
 2   release_date              566 non-null     object
 3   genre                     566 non-null     object
 4   MPAA_rating               513 non-null     object
 5   total_gross               566 non-null     object
```

```
 6   inflation_adjusted_gross  566 non-null    float64
dtypes: float64(1), int64(1), object(5)
memory usage: 31.1+ KB
```

We will now check which movie genre performed well for Disney movies based on the table used disney_movies_total_gross.

[11]:
```
# Checking out which movie genre performed well for Disney
movie_total_data['inflation_adjusted_gross'] =␣
 ↪movie_total_data['inflation_adjusted_gross'].astype(float)
movie_total_data['genre'] = movie_total_data['genre'].astype(str)

movie_genre_group = pd.DataFrame(movie_total_data.
 ↪groupby('genre')['inflation_adjusted_gross'].sum().
 ↪sort_values(ascending=False))


# Reset the index so we can plot using altair
movie_genre_group = movie_genre_group.reset_index()
movie_genre_group
```

[11]:

	genre	inflation_adjusted_gross
0	Adventure	2.459574e+10
1	Comedy	1.545804e+10
2	Musical	9.657566e+09
3	Drama	8.195804e+09
4	Action	5.725162e+09
5	Thriller/Suspense	2.151691e+09
6	Romantic Comedy	1.788873e+09
7	Western	5.167099e+08
8	Documentary	2.034884e+08
9	Black Comedy	1.567305e+08
10	Horror	1.404831e+08
11	Concert/Performance	1.148217e+08

Plotting the bar graph using Altair to check which genre has generated most inflation adjusted revenue.

[12]:
```
# Use altair to generate a bar plot
num_parts_plot = (
    alt.Chart(movie_genre_group, width=500, height=500)
    .mark_bar()
    .encode(
        x=alt.X("genre", title="Genre"),
        y=alt.Y("inflation_adjusted_gross", title="Gross revenue in $"),
    )
    .properties(title="Genre and their revenue")
)
```

```
num_parts_plot
```

[12]: alt.Chart(…)

From the above visualization, it is shown that the 'Adventure' genre has generated the most revenue. But the picture is not over yet, let's explore further…

[13]:
```
movie_genre_group.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12 entries, 0 to 11
Data columns (total 2 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   genre                     12 non-null     object 
 1   inflation_adjusted_gross  12 non-null     float64
dtypes: float64(1), object(1)
memory usage: 320.0+ bytes
```

Now we will plot another graph to visualize which genre has the most number of movies produced by Disney. Therefore, let us count the occurrences of different genre and creating another bar graph to represent that.

[14]:
```
# Checking the count of genres in the dataset
movie_genre_count = movie_total_data.groupby('genre').count()
movie_genre_count


# Reset the index so we can plot using altair
movie_genre_count = movie_genre_count.reset_index()
movie_genre_count
```

[14]:
```
                  genre  index  movie_title  release_date  MPAA_rating  \
0                Action     42           42            42           36   
1             Adventure    130          130           130          119   
2          Black Comedy      3            3             3            3   
3                Comedy    183          183           183          162   
4   Concert/Performance      2            2             2            2   
5           Documentary     16           16            16           16   
6                 Drama    114          114           114          103   
7                Horror      6            6             6            5   
8               Musical     16           16            16           15   
9        Romantic Comedy     23           23            23           22   
10    Thriller/Suspense     24           24            24           23   
11              Western      7            7             7            7   

    total_gross  inflation_adjusted_gross  
0            42                        42  
```

1	130	130
2	3	3
3	183	183
4	2	2
5	16	16
6	114	114
7	6	6
8	16	16
9	23	23
10	24	24
11	7	7

Plotting subsequent bar graph based on the above analysis

[15]:
```python
# Use Altair to generate a bar plot
genre_parts_plot = (
    alt.Chart(movie_genre_count, width=500, height=500)
    .mark_bar()
    .encode(
        x=alt.X("genre", title="Genre"),
        y=alt.Y("index", title="Count of genre"),
    )
    .properties(title="Genre and their count")
)
genre_parts_plot
```

[15]: alt.Chart(…)

The graph here shows that the 'Comedy' genre has the most movies made by Disney, with the second being 'Adventure'.

[16]:
```python
#Importing the custom function
import project_function as pf

final_data = pf.avg_frame(movie_genre_group,movie_genre_count)

# resetting the index
final_data = final_data.reset_index()
final_data
```

[16]:

	genre	inflation_adjusted_gross	index	avg_count
0	Adventure	2.459574e+10	130	1.891980e+08
1	Comedy	1.545804e+10	183	8.447019e+07
2	Musical	9.657566e+09	16	6.035979e+08
3	Drama	8.195804e+09	114	7.189302e+07
4	Action	5.725162e+09	42	1.363134e+08
5	Thriller/Suspense	2.151691e+09	24	8.965379e+07
6	Romantic Comedy	1.788873e+09	23	7.777708e+07

```
7                 Western              5.167099e+08       7  7.381571e+07
8             Documentary              2.034884e+08      16  1.271803e+07
9            Black Comedy              1.567305e+08       3  5.224349e+07
10                 Horror              1.404831e+08       6  2.341385e+07
11   Concert/Performance               1.148217e+08       2  5.741084e+07
```

The revenue numbers and the count calculated in the previous line of code have very differentiating values, therefore, getting the average of gross_total vs count and plotting another graph to gain better insights.

[17]:
```python
# Use altair to generate a bar plot
avg_parts_plot = (
    alt.Chart(final_data, width=500, height=500)
    .mark_bar()
    .encode(
        x=alt.X("genre", title="Genre"),
        y=alt.Y("avg_count", title="Avg $ amount of genre"),
    )
    .properties(title="Genre and their Average $ gross revenue")
)
avg_parts_plot
```

[17]: alt.Chart(…)

Based on the graph, I encountered a rather astonishing result, that shows that the 'Musical' genre has produced the most revenue effect.

[18]:
```python
# Checking out the test cases

import test_function as ttf
ttf.test_custom_agg()
```

0.0.6 Discussions

In this work, I analyzed the Disney dataset and tried to compute which genre has the most impact in terms of revenue. I did some exploratory data analysis to find that the genre of the Disney movies that is most produced is 'Comedy', most popular amongst fans and impactful is 'Musical', and the one that has brought in the most revenue for Disney movies is 'Adventure'.

It is quite unexpected to find that the 'Musical' genre is the most popular amongst fans and has generated the highest revenue effect, as discussed earlier I had expected 'Comedy' to be the most popular genre.

Impact of such findings would be recommending Disney to make more 'Musical' genre based movies. I would like to have the data to see the original budget of the movie, to get more better insights and findings.

0.0.7 References

0.0.8 Resources used

Data Source

1. This Disney database used in this work was borrowed from the following website: https://data.world/kgarrett/disney-character-success-00-16/workspace/data-dictionary
2. The dataset description part involves introduction of Disney movie production borrowed from Wikipedia.