Final Project

INFO 2648

Jesse
Hemingway

DATA

- Box Office performance for top 200 movies from 2010-2023
- https://www.kaggle.com/datasets/parthdande/movies-box-office-collection-data-2000-2024
- Data is relatable
- Movies were a key part of my childhood

This data was pulled from Kaggle and shows the box office performance for the top 200 movies world-wide from the years 2010-2023.

I chose this data because it's regarding a relatable entertainment medium that we all have enjoyed or have key memories of in our lives.

Weekly family movie nights were a key part of my childhood growing up and our vast movie collection and meticulous organization and indexing of said collection was a pride and joy of all the members of my family.

DATA - VALUES

- Rank
- Release Group
- Worldwide
- Domestic
- Domestic_percent
- Foreign
- Foreign_percent
- Year

The data includes:
- Rank: movie rating, starting at zero where zero is #1
- Release Group: movie title
- release year
- Worldwide: box office performance worldwide
- Domestic: domestic performance
- Domestic_percent: domestic performance percentage of worldwide performance
- Foreign: foreign/international performance
- Foreign_percent: foreign/international performance percentage
- Year: release year

A good film is when the price of dinner, the theatre admission, and the babysitter were worth it.

Alfred Hitchcock

LET'S ANSWER

As we review the data, we will be looking to answer three questions:

Did movies perform better domestically or internationally from 2010-2023?

Which year grossed the most total worldwide?

What were the top 10 movies of 2010-2023?

Did movies perform better domestically or internationally from 2010-2023?

As we review the data, we will be looking to answer three questions:

Did movies perform better domestically or internationally from 2010-2023?

Which year grossed the most total worldwide?

What were the top 10 movies of 2010-2023?

Which year grossed the most total worldwide?

As we review the data, we will be looking to answer three questions:

Did movies perform better domestically or internationally from 2010-2023?

Which year grossed the most total worldwide?

What were the top 10 movies of 2010-2023?

What were the top 10 movies of 2010-2023?

As we review the data, we will be looking to answer three questions:

Did movies perform better domestically or internationally from 2010-2023?

Which year grossed the most total worldwide?

What were the top 10 movies of 2010-2023?

EXPLORATORY DATA ANALYSIS

Before we can answer these questions, I had to start with exploratory data analysis to better understand what the data included and how to best manipulate and represent it.

INFO AND DATA TYPES

#	Column	Non-Null Count	Dtype
0	Rank	2800	int64
1	Release Group	2800	object
2	Worldwide	2800	object
3	Domestic	2800	object
4	Domestic_percent	2800	object
5	Foreign	2800	object
6	Foreign_percent	2800	object
7	year	2800	int64

RangeIndex: 2800, 0 to 2799

8 Data Columns

Dtypes: int64 (2), object (6)

Memory Usage: 175.1+ KB

I imported my data set as Movies.

Using the info function, I was able to identify the 8 columns and data types for Movies.

2 columns had an int64 datatype, while 6 columns had an object data type.

There are no cells with a null or missing value.

INFO AND DATA TYPES

#	Column	Non-Null Count	Dtype
0	Rank	2800	int64
1	Release Group	2800	object
2	Worldwide	2800	object
3	Domestic	2800	object
4	Domestic_percent	2800	object
5	Foreign	2800	object
6	Foreign_percent	2800	object
7	year	2800	int64

RangeIndex:
2800, 0 to 2799

8 Data Columns

Dtypes: int64 (2), object (6)

Memory Usage:
175.1+ KB

Using the info function, I was able to identify the 8 columns and data types.

2 columns had an int64 datatype, while 6 columns had an object data type.

There are no cells with a null or missing value.

SUBSETS

0	2010
1	2010
2	2010
3	2010
4	2010
...	...
2795	2023
2796	2023
2797	2023
2798	2023
2799	2023

```
year_sub = movies['year']
```

Understanding the questions I wanted to answer, I knew creating several subsets would help to better visualize the groups I would need to create later.

The first subset was years - dividing the data by the years the films were released, giving me a total of 14 sub data groups for the years.

SUBSETS

	Rank	Release Group	Domestic	Domestic_percent	year
0	0	Toy Story 3	41,50,04,880	38.90%	2010
1	1	Alice in Wonderland	33,41,91,110	32.60%	2010
2	2	Harry Potter and the Deathly Hollows: Part 1	29,59,83,305	30.80%	2010
3	3	Inception	29,25,76,195	35.30%	2010
4	4	Shrek Forever After	23,87,36,787	31.70%	2010
...	...	...	...	...	...
2795	195	On the Wandering Paths	0	0	2023
2796	196	Weekend Rebels	0	0	2023
2797	197	Ransomed	1,42,101	1.80%	2023
2798	198	Checker Tobi und die Reise zu den fliegenden Fl√°ssen	0	0	2023
2799	199	The Silent Service	0	0	2023

```
domestic_sub = movies[['Rank','Release Group','Domestic','Domestic_percent','year']]
```

The next subset was domestic - showing only the data for the domestic values.

SUBSETS

	Rank	Release Group	Foreign	Foreign_percent	year
0	0	Toy Story 3	65,19,64,823	61.10%	2010
1	1	Alice in Wonderland	69,12,76,000	67.40%	2010
2	2	Harry Potter and the Deathly Hollows: Part 1	66,43,00,000	69.20%	2010
3	3	Inception	53,56,82,500	64.70%	2010
4	4	Shrek Forever After	51,38,64,080	68.30%	2010
...	...	...	...	...	...
2795	195	On the Wandering Paths	81,87,125	100%	2023
2796	196	Weekend Rebels	81,84,539	100%	2023
2797	197	Ransomed	79,59,533	98.20%	2023
2798	198	Checker Tobi und die Reise zu den fliegenden Fl√°ssen	79,53,344	100%	2023
2799	199	The Silent Service	78,36,539	100%	2023

```
foreign_sub = movies[['Rank','Release Group','Foreign','Foreign_percent','year']]
```

The final subset was foreign - showing only the data for international values.

The domestic and foreign subsets revealed that not all movies were released globally. While this information was interesting to note and could be helpful in the analyzing of the data, this was not used for this particular purpose, as it did not lend to answering any of the questions.

CHANGING DATA TYPES

Before we can answer these questions, I had to start with exploratory data analysis to better understand what the data included and how to best manipulate and represent it.

DATA TYPES

#	Column	Non-Null Count	Dtype
0	Rank	2800	int64
1	Release Group	2800	object
2	Worldwide	2800	int64
3	Domestic	2800	int64
4	Domestic_percent	2800	object
5	Foreign	2800	int64
6	Foreign_percent	2800	object
7	year	2800	int64

```
movies['Worldwide'] = movies['Worldwide'].str.replace(',', '').astype(int)
movies['Domestic'] = movies['Domestic'].str.replace(',', '').astype(int)
movies['Foreign'] = movies['Foreign'].str.replace(',', '').astype(int)
```

To be able to utilize our numerical data and answer our questions, the data for Worldwide, Domestic, and Foreign earnings had to be changed to int64.

The data provided in these columns included commas, so the code had to include a replacement function, in addition to the astype function.

The resulting data info is shown here.

GROUPING AND AGGREGATION

Grouping and aggregation was the final step needed to sort the data in a way that would allow me to answer my questions.

GROUP BY YEAR

	Rank	Release Group	Worldwide	Domestic	Domestic_percent	Foreign	Foreign_percent	year
0	0	Toy Story 3	1066969703	415004880	38.90%	651964823	61.10%	2010
1	1	Alice in Wonderland	1025467110	334191110	32.60%	691276000	67.40%	2010
2	2	Harry Potter and the Deathly Hallows: Part 1	960283305	295983305	30.80%	664300000	69.20%	2010
3	3	Inception	828258695	292576195	35.30%	535682500	64.70%	2010
4	4	Shrek Forever After	752600867	238736787	31.70%	513864080	68.30%	2010
...	...	...	...	...	...	...	...	...
2605	5	Spider0Man: Across the Spider0Verse	690615475	381311319	55.20%	309304156	44.80%	2023
2606	6	Wonka	632302312	218402312	34.50%	413900000	65.50%	2023
2607	7	The Little Mermaid	569626289	298172056	52.30%	271454233	47.70%	2023
2608	8	Mission: Impossible 0 Dead Reckoning Part One	567535383	172135383	30.30%	395400000	69.70%	2023
2609	9	Elemental	496444308	154426697	31.10%	342017611	68.90%	2023

```
year_grp = movies.groupby(['year'])
```

Grouping by year allowed me to create a separate data group, which could more easily be manipulated to calculate certain totals and apply needed functions.

So now that we've created a group, let's get around to answering our questions.

Did movies perform better domestically or internationally from 2010-2023?

DOMESTIC VS. FOREIGN

1	33,41,91,110
2	29,59,83,305
3	29,25,76,195
4	23,87,36,787
...	...
2795	0
2796	0
2797	1,42,101
2798	0
2799	0

domestic_sub['Domestic'].sum

Series.sum of 0 41,50,04,880

1	69,12,76,000
2	66,43,00,000
3	53,56,82,500
4	51,38,64,080
...	...
2795	81,87,125
2796	81,84,539
2797	79,59,533
2798	79,53,344
2799	78,36,539

foreign_sub['Foreign'].sum

Series.sum of 0 65,19,64,823

By using the Domestic and Foreign subsets we created at the start and finding the sum of each, we are able to determine that movies performed better internationally.

Which year grossed the most total worldwide?

HIGHEST GROSSING YEAR

year	median	mean	max	sum
2019	63349461.0	1.771393e+08	2799439100	35427854208
2017	60999583.0	1.710457e+08	1332539889	34209146582
2018	73714586.5	1.706797e+08	2048359754	34135939193
2015	72145657.0	1.637093e+08	2068223624	32741853312
2016	75353022.5	1.618727e+08	1153296293	32374542002
2014	72084134.0	1.500526e+08	1104054072	30010529897
2012	61663992.0	1.475370e+08	1518812988	29507398427
2013	68552764.5	1.453878e+08	1280802282	29077554568
2011	62458932.5	1.346257e+08	1341511219	26925147932
2010	60706384.5	1.278503e+08	1066969703	25570061807
2023	32431771.5	1.078035e+08	1445638421	21560692459
2022	21745864.0	1.043728e+08	2320250281	20874564107
2021	26511473.0	9.124814e+07	1912233593	18249627473
2020	12023268.0	4.020858e+07	461421559	8041716369

```
year_grp['Worldwide'].agg(['median', 'mean', 'max', 'sum']).sort_values(by=['sum'], ascending=[False])
```

By using aggregation and sorting on my year group, we are able to identify that 2019 was the highest grossing year.

What were the top 10 movies of 2010-2023?

TOP 10 MOVIES OF 2010-2023

	Rank	Release Group	Worldwide	Domestic	Domestic_percent	Foreign	Foreign_percent	year
1800	0	Avengers: Endgame	2799439100	858373000	30.70%	1941066100	69.30%	2019
2400	0	Avatar: The Way of Water	2320250281	684075767	29.50%	1636174514	70.50%	2022
1000	0	Star Wars: Episode VII 0 The Force Awakens	2068223624	936662225	45.30%	1131561399	54.70%	2015
1600	0	Avengers: Infinity War	2048359754	678815482	33.10%	1369544272	66.90%	2018
2200	0	Spider0Man: No Way Home	1912233593	804793477	42.10%	1107440116	57.90%	2021
1001	1	Jurassic World	1670400637	652270625	39%	1018130012	61%	2015
1801	1	The Lion King	1656943394	543638043	32.80%	1113305351	67.20%	2019
400	0	The Avengers	1518812988	623357910	41%	895455078	59%	2012
1002	2	Furious 7	1515047671	353007020	23.30%	1162040651	76.70%	2015
2401	1	Top Gun: Maverick	1495696292	718732821	48.10%	776963471	51.90%	2022

```
top10 = movies.sort_values(by=['Worldwide'], ascending=[False])
top10.head(10)movies.head(10)
```

We go back to our original Movies dataset and sort by the Worldwide values to find our top movies for 2010-2023.

Did any surprise you? Any you thought should have been in the top 10?

I found it interesting that the top 10 movies over the 14 years, were not all number 1's from their release year.

Thank you!

Jesse
Hemingway