

MOVIE STUDIO SUGGESTIONS

Table of Contents

| | |
|--|---|
| CHAPTER 1 BUSINESS UNDERSTANDING | 2 |
| 1.1. OVERVIEW | 2 |
| 1.2. PROBLEM STATEMENT | 2 |
| CHAPTER 2 DATA UNDERSTANDING | 2 |
| 2.1. DATA UNDERSTANDING | 2 |
| A. BOX OFFICE MOJO | 2 |
| 2.2. DATA PREPARATION | 5 |
| CHAPTER 3 DATA ANALYSIS..... | 5 |
| CHAPTER 4 CONCLUSION AND RECOMMENDATION..... | 6 |

CHAPTER 1 BUSINESS UNDERSTANDING

1.1. OVERVIEW

Microsoft intend to enter the movie studio scene and they have undertaken to research the industry to ensure that the content they create is marketable. They have charged the Data Scientist to give them insights on the market trend so that they can make a decision on what types of films to create. This will be achieved by analyzing what types of films are doing best at the box office and translating the findings into actionable insights that can then be rendered to the head of Microsoft's new movie studio.

1.2. PROBLEM STATEMENT

The main goal of the project is to find the types of films doing best at the box office. This shall be achieved by analyzing the following:

- i) What genres brought in the most income?
- ii) What were the budgets and profits of the movies?
- iii) What directors brought in the most income and best ratings?
- iv) Was there a relationship between runtime and ratings?
- v) What are the top studios?

CHAPTER 2 DATA UNDERSTANDING

2.1. DATA UNDERSTANDING

A. BOX OFFICE MOJO

The Box Office Mojo data had the following properties:

- i) The shape was (3387, 5)
- ii) The column information was:

```
bom_movie_data.info()
✓ 0.1s

<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
```

- iii) The data description was:

```
bom_movie_data.describe()
```

✓ 0.2s

| | domestic_gross | year |
|-------|----------------|-------------|
| count | 3.359000e+03 | 3387.000000 |
| mean | 2.874585e+07 | 2013.958075 |
| std | 6.698250e+07 | 2.478141 |
| min | 1.000000e+02 | 2010.000000 |
| 25% | 1.200000e+05 | 2012.000000 |
| 50% | 1.400000e+06 | 2014.000000 |
| 75% | 2.790000e+07 | 2016.000000 |
| max | 9.367000e+08 | 2018.000000 |

iv) The first five values for this data were:

```
bom_movie_data.head()
```

✓ 0.1s

| | title | studio | domestic_gross | foreign_gross | year |
|---|---|--------|----------------|---------------|------|
| 0 | Toy Story 3 | BV | 415000000.0 | 652000000 | 2010 |
| 1 | Alice in Wonderland (2010) | BV | 334200000.0 | 691300000 | 2010 |
| 2 | Harry Potter and the Deathly Hallows Part 1 | WB | 296000000.0 | 664300000 | 2010 |
| 3 | Inception | WB | 292600000.0 | 535700000 | 2010 |
| 4 | Shrek Forever After | P/DW | 238700000.0 | 513900000 | 2010 |

v) The last five values for this data were:

```
bom_movie_data.tail()
```

✓ 0.7s

| | title | studio | domestic_gross | foreign_gross | year |
|------|-----------------------------|------------|----------------|---------------|------|
| 3382 | The Quake | Magn. | 6200.0 | NaN | 2018 |
| 3383 | Edward II (2018 re-release) | FM | 4800.0 | NaN | 2018 |
| 3384 | El Pacto | Sony | 2500.0 | NaN | 2018 |
| 3385 | The Swan | Synergetic | 2400.0 | NaN | 2018 |
| 3386 | An Actor Prepares | Grav. | 1700.0 | NaN | 2018 |

B. THE NUMBERS

The Numbers data had the following properties:

- i) The shape was (5782,6)
- ii) The column information was:

```
tn_data.info()
```

✓ 0.7s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5782 entries, 0 to 5781
Data columns (total 6 columns):
#   Column              Non-Null Count  Dtype
---  -
0   id                   5782 non-null   int64
1   release_date         5782 non-null   object
2   movie                 5782 non-null   object
3   production_budget    5782 non-null   object
4   domestic_gross       5782 non-null   object
5   worldwide_gross      5782 non-null   object
dtypes: int64(1), object(5)
memory usage: 271.2+ KB
```

iii) The data description was:

```
tn_data.describe()
```

✓ 0.1s

| | id |
|-------|-------------|
| count | 5782.000000 |
| mean | 50.372363 |
| std | 28.821076 |
| min | 1.000000 |
| 25% | 25.000000 |
| 50% | 50.000000 |
| 75% | 75.000000 |
| max | 100.000000 |

iv) The first five values of this data were:

```
tn_data.head()
```

✓ 0.1s

| | id | release_date | movie | production_budget | domestic_gross | worldwide_gross |
|---|----|--------------|---|-------------------|----------------|-----------------|
| 0 | 1 | Dec 18, 2009 | Avatar | \$425,000,000 | \$760,507,625 | \$2,776,345,279 |
| 1 | 2 | May 20, 2011 | Pirates of the Caribbean: On Stranger Tides | \$410,600,000 | \$241,063,875 | \$1,045,663,875 |
| 2 | 3 | Jun 7, 2019 | Dark Phoenix | \$350,000,000 | \$42,762,350 | \$149,762,350 |
| 3 | 4 | May 1, 2015 | Avengers: Age of Ultron | \$330,600,000 | \$459,005,868 | \$1,403,013,963 |
| 4 | 5 | Dec 15, 2017 | Star Wars Ep. VIII: The Last Jedi | \$317,000,000 | \$620,181,382 | \$1,316,721,747 |

v) The last five values of this data were:

```
tn_data.tail()
```

✓ 0.1s

| | id | release_date | movie | production_budget | domestic_gross | worldwide_gross |
|------|----|--------------|-------------------------------|-------------------|----------------|-----------------|
| 5777 | 78 | Dec 31, 2018 | Red 11 | \$7,000 | \$0 | \$0 |
| 5778 | 79 | Apr 2, 1999 | Following | \$6,000 | \$48,482 | \$240,495 |
| 5779 | 80 | Jul 13, 2005 | Return to the Land of Wonders | \$5,000 | \$1,338 | \$1,338 |
| 5780 | 81 | Sep 29, 2015 | A Plague So Pleasant | \$1,400 | \$0 | \$0 |
| 5781 | 82 | Aug 5, 2005 | My Date With Drew | \$1,100 | \$181,041 | \$181,041 |

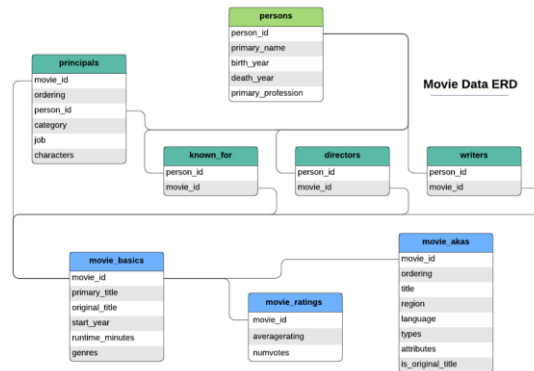
C. IMDB

The IMDB data was in SQL format and the tables associated with it were:

- i) principals
- ii) persons
- iii) known_for
- iv) directors
- v) writers
- vi) movie_basics
- vii) movie_ratings

viii) movie_akas

The entity relationship diagram (ERD) was as follows:



2.2. DATA PREPARATION

Data cleaning on The Numbers dataset was done. The dollar sign and comma sign were removed and their types changed to float so that calculations can be done on the production_budget, domestic_gross, worldwide_gross columns. There were no missing values found in this dataset so no action against them was taken and the analysis was done on this set. However, there were some duplicates and they were dropped.

```

tn_data.info()
✓ 0.9s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5782 entries, 0 to 5781
Data columns (total 6 columns):
#   Column                Non-Null Count  Dtype
---  -
0   id                    5782 non-null   int64
1   release_date          5782 non-null   object
2   movie                 5782 non-null   object
3   production_budget     5782 non-null   float64
4   domestic_gross        5782 non-null   float64
5   worldwide_gross       5782 non-null   float64
dtypes: float64(3), int64(1), object(2)
    
```

The individual tables on the IMDB dataset were checked for missing values and duplicates and they were handled differently. For columns with a lot of missing values, the columns were dropped. If the missing values were not too many, the missing records were replaced with the mean of the column for integers and floats or the mode for string data.

Bom movies' missing data were filled with the mean of the respective column. However, some missing records were dropped from the dataset.

CHAPTER 3 DATA ANALYSIS

After analysis, it was found that:

- i) The genre with the highest rating was Family and Adventure

- ii) The movies released for sale to the worldwide audience did better than those released domestically
- iii) The top 10 directors were found to be:
 - a. Reinhard Kungel
 - b. Erik Matti
 - c. Mark Adams
 - d. Ana Reiper
 - e. Aydin Bulut
 - f. Valeria Testagrossa
 - g. Fancesco Longo
 - h. Davide Pesca
 - i. Ignas Joynas
- iv) The movie length should range from 40 to 120 minutes.
- v) The top studios that they would be in competition with are:
 - a. Studio BV
 - b. Fox
 - c. Warner Bros (WB)
 - d. Universal Pictures (Uni)
 - e. Sony

CHAPTER 4 CONCLUSION AND RECOMMENDATION

The results above informed the recommendations to Microsoft for as they open the movie studio. These would ensure they produce films that will be consumed by many and have directors that would ensure they have high ratings. Also they have a budget that they can use while making the films so that they can measure their expected profits depending on the genre they pick.