|  |  |
| --- | --- |
| IMDB  Data Profiling | Abstract  To analyze and gain insights from the IMDB data. Utilizing advanced data profiling through Alteryx.  Monalika Pradhan, Akhilesh Dongre, Jatin Madan, Shivani Shekhawat  Group 4 |

**Introduction**

This report comprehensively analyzes Food Facility Inspections in California, utilizing advanced data profiling techniques in Alteryx. This analysis aimed to gain deep insights into the inspection data, enabling informed decision-making and ensuring food safety standards are met across various facilities in California.

The dataset is taken from :

<https://data.sonomacounty.ca.gov/Health/Food-Facility-Inspections/8r44-w5qd>

**Data Profiling using Alteryx**

Performed data profiling using Alteryx, used tools like AutoField, Field Summary, Basic data profile, unique tools to identify appropriate data types, min-max values, and data cleansing opportunities.

A screenshot of a computer

Description automatically generated

1. Avatar Dataset

A screenshot of a computer

Description automatically generated

The dataset provided contains information about the domestic box office performance of the movie "Avatar". Here's a brief overview of the data, which is similar to the other 8 movie datasets:

* **tconst**: An identifier for the movie.
* **title**: The movie title.
* **Date**: The date of the box office record.
* **Rank**: The rank of the movie in terms of box office revenue for that day.
* **Gross**: The gross revenue for that day.
* **%YD** (% Year-to-Date): The percentage change in revenue compared to the same day in the previous year.
* **%LW** (% Last Week): The percentage change in revenue compared to the same day in the previous week.
* **Theaters**: The number of theaters showing the movie.
* **Per Theater:** The revenue per theater.
* **Total Gross**: The cumulative total gross revenue up to that date.
* **Days**: The number of days since the movie's release.

Alteryx Tools:

* Auto Field

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Size** |
| tconst | String | 9 |
| title | String | 6 |
| Date | String | 12 |
| Rank | Int64 | 8 |
| Gross | String | 11 |
| %YD | String | 4 |
| %LW | String | 4 |
| Theaters | String | 5 |
| Per Theater | String | 6 |
| Total Gross | String | 12 |
| Days | Int16 | 2 |

* Field Summary

| Record | Report |
| --- | --- |
| 1 | **String/Character Fields**   | **Name** | **% Missing** | **Unique Values** | **Shortest Value** | **Longest Value** | **Min Value Count** | **Max Value Count** | **Remarks** | | --- | --- | --- | --- | --- | --- | --- | --- | | Total Gross | 0.0% | 318 | $26,752,099 | $109,497,762 | 1 | 1 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | Rank | 0.0% | 17 | 1 | 10 | 1 | 176 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | %YD | 0.6% | 155 | 2 | -34% | 1 | 7 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | Date | 0.0% | 318 | Jan 1, 2010 | Dec 18, 2009 | 1 | 1 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | Theaters | 0.0% | 44 | 9 | 3,452 | 2 | 28 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | Per Theater | 0.0% | 293 | $60 | $7,750 | 1 | 3 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | Days | 0.0% | 318 | 1 | 100 | 1 | 1 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | Gross | 0.0% | 317 | $571 | $26,752,099 | 1 | 2 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | %LW | 4.4% | 116 | -2% | -14% | 1 | 14 | Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together. | | title | 0.0% | 1 | Avatar | Avatar | 318 | 318 |  | | tconst | 0.0% | 1 | tt0499549 | tt0499549 | 318 | 318 |  | |

* Basic Data Profile

We have added the basic data profile report separately, where we can get insights about the nulls, non-null, min, max, data type, shortest and longest values, etc. Below is the snipped of the report:

A screenshot of a computer

Description automatically generated

**Basic Findings**

**Rank**: The movie's rank varied from 1 to 55 during its box office run, and it contains “-” values as well

**Gross**: Daily gross earnings ranged from a minimum of $357 to a maximum of about $28.27 million.

**Per Theater**: Earnings per theater varied between $60 and $8,181.

**Total Gross**: The cumulative gross earnings increased over time, reaching approximately $760.51 million.

**Days**: The dataset covers 336 days from the movie's release.

**%LW and %YD** contains null values

**Insights and Observations**

Strong Initial Performance: "Avatar" had a very strong opening, indicated by the high initial daily and per-theater gross.

Longevity in Theaters: The movie remained in theaters for a significant duration (336 days), highlighting its lasting appeal.

Consistent Top Rankings: Despite fluctuations, the movie consistently ranked well during its theatrical run.

Revenue Stability: After the initial spike, the total gross showed stability, indicating a steady influx of viewers over an extended period.

1. The Numbers - Domestic Box Office - Avengers\_ Age of Ultron (2015)

A screenshot of a computer

Description automatically generated

Inconsistencies and Useful Information:

The numerical values in columns like Gross, Per Theater, and Total Gross are stored as strings, possibly with currency symbols, which might require conversion to numerical types for analysis.

The Date column is also in object format, suggesting the need for conversion to a datetime format for time series analysis.

The dataset provides detailed information about the daily box office performance of "Avengers: Age of Ultron", which can be useful for analyzing trends over time, such as how the movie's revenue changed from day to day, weekly trends, and its overall financial success.

Data Description:

The description provides a count of values, unique values, top (most frequent) values, and frequency for categorical data.

Numerical summary (mean, standard deviation, min, max, quartiles) is not applicable for most columns due to their object data type, except for Days.

1. The Numbers - Domestic Box Office - Avengers\_ Endgame (2019)

Data Types: Except for 'Days', which is an integer, all other columns are of the object (string) type.

Missing Values:

'%YD': 2 missing values.

'%LW': 8 missing values.

'Per Theater': 1 missing value.

Unique Values:

The 'tconst' and 'title' columns have only one unique value each, indicating the dataset is specific to "Avengers: Endgame".

The 'Date' column has 141 unique values, suggesting data for each day of the movie's run.

'Gross', 'Total Gross', 'Per Theater', '%YD', and '%LW' show a variety of unique entries, reflecting the daily variations in box office performance.

Range of Data:

The 'Days' column ranges from 0 to 140, indicating the dataset covers a 141-day period of the movie's box office run.

Duplicates: There are no duplicate rows in the dataset.

Descriptive Statistics: Since most columns are of string type, traditional descriptive statistics (like mean, standard deviation) are not applicable. However, the frequent occurrence of specific values in columns like 'Rank' and 'Theaters' can be of interest.

1. The Numbers - Domestic Box Office - Avengers\_ Infinity War (2018)

Data Inconsistencies:

Data types in the datasets are consistent.

Date formats are consistent with each other.

1. The Numbers - Domestic Box Office - Black Panther (2018)

Missing Values:

%YD (Year-to-Date) column has 2 missing values.

%LW (Last Week) column has 8 missing values.

Per Theater column has 1 missing value.

Days column has 1 missing value.

Data Types:

Most columns are of object type (probably containing strings), except for Days, which is a float.

Columns like Gross, Total Gross, and Per Theater that appear to represent financial figures are currently in string format due to the presence of symbols like $. Similarly, columns like Theaters might contain numeric values but are in string format possibly due to the use of commas.

Basic Statistics:

The dataset is quite uniform in terms of tconst and title, indicating it is specific to the movie "Black Panther".

There are 176 unique dates, suggesting daily records.

The Rank column has 16 unique values with '-' being the most frequent.

Financial columns (Gross, Per Theater, Total Gross) show a wide range of values, but exact figures are not discernible from the summary statistics due to their string format.

The Days column ranges from 1 to 175, indicating the data spans a period of 175 days.

1. The Numbers - Domestic Box Office - Spider-Man No Way Home

Data Overview:

The dataset contains 128 entries with 11 columns.

* Missing Values:

'%YD' (Year-over-Year percentage change) has 2 missing values.

'%LW' (Last Week percentage change) has 8 missing values.

'Days' (number of days since release) has 1 missing value.

* Data Types and Consistency:

Many columns which appear to contain numerical data ('Gross', 'Theaters', 'Per Theater', 'Total Gross') are actually stored as objects (likely due to the presence of characters like '$' and ',').

'Date' is also stored as an object, which suggests it might not be in a proper date format.

'Rank' appears as an object which might include non-numeric values.

* Unique Values in Categorical Columns:

'tcont' has only one unique value ('tt10872600'), indicating it's the same for the entire dataset.

'title' has only one unique value ('Spider-Man: No Way Home'), suggesting the dataset is specific to this title.

'Date' has multiple unique values in a range of formats (e.g., 'Dec 16, 2021', 'Jan 1, 2022'), indicating the data spans different dates.

Other columns like 'Rank', 'Gross', '%YD', '%LW', 'Theaters', 'Per Theater', and 'Total Gross' have multiple unique entries.

1. The Numbers - Domestic Box Office - Star Wars\_ Episode VII - The Force Awakens

Statistical Summary (for the Days column):

Count: 120

Mean: 59.5 days

Standard Deviation: 34.78 days

Min: 0 days

25th Percentile: 29.75 days

Median: 59.5 days

75th Percentile: 89.25 days

Max: 119 days

Inconsistencies and Unique Patterns:

Rank: Contains a variety of values including 'P', numerical ranks, and '-'. The '-' might indicate missing or non-applicable ranks.

Theaters: Varies from 0 to 4,134. The '0' value could represent pre-release or special cases.

Data Format Issues:

Many columns that should be numeric are in string format (e.g., Gross, Theaters, Per Theater, Total Gross). This could be due to the presence of symbols like $ and , in these fields.

1. The Numbers - Domestic Box Office - The Avengers

Date Range: The dataset covers a period of about four and a half months.

Gross Earnings:

The average daily gross earning is approximately $6.57 million.

The maximum daily gross earning recorded is over $80 million.

Theaters:

On average, "The Avengers" was shown in about 2,515 theaters.

The maximum number of theaters showing the movie at any given time was 4,349.

Per Theater Earnings:

The average gross earning per theater is around $1,687.

The highest gross earning per theater reached $18,582.

Total Gross:

The average cumulative gross over the recorded period is around $531.99 million.

The maximum cumulative gross recorded is over $621 million.

Days Since Release:

The dataset includes data for up to 133 days since the movie's release.

The average number of days since release for the data points is approximately 52 days.

1. The Numbers - Domestic Box Office – Titanic

Date Range:

The dataset covers nearly 20 years, from December 19, 1997, to December 14, 2017.

Gross Earnings:

The average daily gross earning is approximately $2.43 million.

The maximum daily gross earning recorded is over $13 million.

Theaters:

On average, "Titanic" was shown in about 2,123 theaters.

The maximum number of theaters showing the movie at any given time was 3,265.

Per Theater Earnings:

The average gross earning per theater is around $929.

The highest gross earning per theater reached $4,667.

Total Gross:

The average cumulative gross over the recorded period is around $493.52 million.

The maximum cumulative gross recorded is over $659 million.

Days Since Release:

The dataset includes data for up to 7,301 days (20 years) since the movie's release.

The average number of days since release for the data points is approximately 1,740 days.

For the Days column, we have “,” in the values which need to be handled further.

**JSON Files:**

**Name\_basics**

Missing Values: The columns DI\_CreateDatetime, DI\_JobId, and DI\_JobName have missing values (6 instances each). All other columns are complete.

Data Types: The birthYear column is of type integer, while deathYear is of type object. This inconsistency might be due to the presence of non-numeric values (e.g., '\N') in the deathYear column.

Birth and Death Years: The birthYear values seem consistent, but deathYear contains only the placeholder value '\N', which suggests either missing information or that the subjects are all living.

Profession and Titles: The primaryProfession and knownForTitles columns have structured lists (e.g., comma-separated values), which might need further parsing depending on the use case.

**Title\_basics**

No Missing Values: There are no missing values in this dataset.

Data Types: Most columns are of type object, including startYear, endYear, and runtimeMinutes, which might be more useful as numeric types. The isAdult column is an integer, which is appropriate.

Year and Runtime Values: The startYear column contains years in a consistent format. However, the endYear column has placeholder values ('\N'), suggesting these titles do not have an end year (likely because they are movies). The runtimeMinutes column has numeric values in string format.

Genres: The genres column contains comma-separated values, similar to the knownForTitles column in the name basics dataset.