**DATA PROFILING DOCUMENTATION**

Let’s understand the data that is provided. Our overall data comprises of 9 TSV and sql dump having 6 tables.

Firstly, let’s understand about the data in 9 TSVs.

**1.TSV**

The following are the 9 TSVs provided along with the number of records and fields that each TSV contains.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **TSV Name** | **Number of Fields** | **Number of Records** |
| 1 | The Numbers - Domestic Box Office - Avatar | 11 | 318 |
| 2 | The Numbers - Domestic Box Office – Avengers\_Age of Ultron (2015) | 11 | 79 |
| 3 | The Numbers - Domestic Box Office – Avengers\_Endgame (2019) | 11 | 141 |
| 4 | The Numbers - Domestic Box Office – Avengers\_Infinity War (2018) | 11 | 141 |
| 5 | The Numbers - Domestic Box Office - Black Panther (2018) | 11 | 176 |
| 6 | The Numbers - Domestic Box Office - Spider-Man No Way Home | 11 | 128 |
| 7 | The Numbers - Domestic Box Office - Star Wars\_Episode VII - The Force Awakens | 11 | 120 |
| 8 | The Numbers - Domestic Box Office - The Avengers | 11 | 97 |
| 9 | The Numbers - Domestic Box Office - Titanic | 11 | 265 |

All the TSVs comprises of data related to individual movies comprising of the following 11 fields along with their description

|  |  |
| --- | --- |
| **FIELD** | **DESCRIPTION** |
| tconst | A unique identifier used by databases for movie or TV show titles. |
| Title | The name of the movie or TV show. |
| Date | The release date or the specific date to which the data pertains. |
| Rank | The ranking of the movie based on box office sales or viewer ratings for a given period. (ranging from 1-10) |
| Gross | Box office revenue of the movie for a specific period. |
| %YD | Percentage change in gross revenue compared to the same date in the previous year. |
| %LW | Percentage change in gross revenue compared to the previous week. |
| Theaters | The number of theaters showing the movie. |
| Per Theater | The average gross revenue per theater. |
| Total Gross | The total box office revenue earned by the movie to date. |
| Days | The number of days the movie has been in release in theaters. |

A screenshot of a computer

Description automatically generated

When analyzing the 9 TSV’s dynamically since all the TSVs had the same structure, we got the following results :

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Field** | **Datatype** | **OK %** | **NOT OK %** | **Unique %** | **Null %** | **Empty %** | **Min Length** | **Max Length** | **Remarks** |
| tconst | String | 100% | 0.00% | 0.61% | 0.00% | 0.00% | 9 | 10 | No inconsistencies recorded. But for one TSV i.e., Spiderman no way home , had its “tconst” column named as “tcont” but alteryx took it under tconst either way. |
| Title | String | 100% | 0.00% | 0.61% | 0.00% | 0.00% | 6 | 42 | No inconsistencies observed |
| Summary | String | 100% | 0.00% | 90.31% | 0.00% | 0.00% | 11 | 12 | No inconsistencies observed |
| Rank | String | 67.24% | 32.76% | 1.37% | 0.00% | 0.00% | 1 | 6 | Contained inconsistenjcies such as “-“ and “P” in the data. |
| Gross | String | 100% | 0.00% | 99.8% | 0.00% | 0.00% | 4 | 12 | No inconsistencies observed |
| %YD | String | 98.43% | 0.00% | 17.34% | 1.57% | 0.07% | 1 | 4 | There were 22 null values and 1 empty row |
| %LW | String | 92.63% | 0.00% | 12.15% | 6.96% | 0.41% | 3 | 4 | There were 102 null values and 6 were empty. |
| Theatres | String | 100% | 0.00% | 13.04% | 0.00% | 0.00% | 1 | 5 | No inconsistencies observed |
| Per Theatre | String | 99.66% | 0.00% | 66.01% | 0.34% | 0.00% | 3 | 7 | There were 5 null values observed |
| Total Gross | String | 100% | 0.00% | 100% | 0.00% | 0.00% | 10 | 12 | No inconsistencies observed |
| Days | String | 99.80% | 0.00% | 27.17% | 0.20% | 0.00% | 1 | 5 | There were 3 null values observed. |