IMDb Data Profiling Report

MINAL NARANJE – 002087516 | **ADITI DESHMUKH –** 002055568 | **SAMEER MALVE -** 002316478

As part of our data processing and transformation efforts, our team conducted a thorough data profiling analysis on the seven IMDb datasets. This process involved evaluating the structure, quality, consistency, and integrity of the data before implementing any transformations in **Alteryx and YData Profiling**. Below, we present our findings for each dataset individually, highlighting potential issues and outlining the steps we plan to take to ensure data quality.

---

1. name.basics.tsv (Personnel Data)

1.1 Data Overview

The name.basics.tsv file provides details about various individuals involved in the film and entertainment industry, such as actors, directors, and writers. It contains a total of **14,195,120 rows**, with each row representing a unique person.

The key columns and their respective data types are:

- **nconst (string):** A unique identifier for each individual.
- **primaryName (string):** The full name of the person.
- **birthYear (string):** The year the individual was born.
- **deathYear (string):** The year of death (if applicable).
- **primaryProfession (string):** A comma-separated list of professions associated with the individual.
- **knownForTitles (string):** A list of tconst identifiers indicating the movies or TV shows the person is known for.

1.2 Data Completeness & Missing Values

- The **birthYear column has missing values for approximately 13.6 million records**, meaning that birth years are unavailable for 95.8% of individuals.
- Similarly, **deathYear has missing values for 14 million records (98.5%)**, likely because many individuals are still alive.
- The **primaryProfession column has around 2.7 million missing values (19.6%)**, indicating that certain individuals do not have a listed profession.
- The **knownForTitles field is missing for approximately 1.6 million records (11.4%)**, which suggests that some people do not have notable work associated with them.

1.3 Data Consistency & Accuracy

- The **nconst column is unique**, and we did not find any duplicate identifiers, which is a good indication of data integrity.
- Another key issue is that **knownForTitles references some tconst values that do not exist in title.basics.tsv**, which means there may be foreign key integrity issues.

1.4 Data Distribution & Summary

- The most commonly occurring **professions** in the dataset include **actors, directors, and writers**, which aligns with expectations for IMDb data.

- We noticed **outliers in the birthYear column**, with some values dating back to the **1800s**, which is likely a data entry error.

1.5 Key Relationships Across Datasets

- The knownForTitles column in this dataset **links to title.basics.tsv via the tconst field**, but we identified several mismatches.
- This dataset also connects with title.principals.tsv and title.crew.tsv to map personnel to specific movies and TV shows.

1.6 Data Issues & Remediation Plan

- **Fixing birthYear and deathYear inconsistencies** by ensuring that deathYear is always greater than or equal to birthYear.
- **Validating knownForTitles** against title.basics.tsv to remove or correct unmatched records.
- **Handling missing values** by filling them with default values or applying data imputation techniques.

---

2. title.basics.tsv (Title Details)

2.1 Data Overview

The title.basics.tsv file contains **11,464,895 records** and serves as the main catalog of movies, TV series, and other productions. It provides metadata about each title, such as its type, release year, and genre.

Key columns:

- **tconst (string):** A unique identifier for each title.
- **titleType (string):** Specifies whether the title is a **movie, short, series, etc.**.
- **primaryTitle (string):** The main title used for display.
- **originalTitle (string):** The title in its original language.
- **isAdult (integer):** A binary flag indicating whether the content is intended for adults.
- **startYear (integer):** The release year of the title.
- **endYear (integer):** Applicable for TV series, indicating when it ended.
- **runtimeMinutes (integer):** The duration of the title in minutes.
- **genres (string):** A comma-separated list of genres.

2.2 Data Completeness & Missing Values

- **endYear is missing for 99.2% of records**, which suggests that most titles are either movies or ongoing TV shows.
- **runtimeMinutes has missing values for 68.5% of the records**, which affects the ability to analyze movie durations.
- **genres is missing for 4.4% of records**, meaning some titles have no genre classification.

2.3 Data Consistency & Accuracy

- We found **3,500 duplicate movie titles**, which might cause issues in reporting.
- The **tconst field is unique and well-maintained**.

- Some records have **incorrect startYear values**, including placeholder values such as \\N.

2.4 Data Distribution & Summary

- The dataset is **dominated by movies (65%)**, followed by TV series (20%), shorts (10%), and other content (5%).
- The most common **genres** are **Drama, Comedy, and Documentary**.
- **Runtime anomalies** were detected, with some movies exceeding **500 minutes**, which is unusually long.

2.5 Key Relationships Across Datasets

- **Joins with title.ratings.tsv** using tconst to match movies with ratings.
- **Links with title.akas.tsv** to provide alternate titles.
- **References title.crew.tsv** to connect movies with their directors and writers.

2.6 Data Issues & Remediation Plan

- **Remove duplicate movie titles** to maintain data integrity.
- **Fix runtimeMinutes inconsistencies** by filling missing values with reasonable estimates based on titleType.
- **Standardize genres and correct missing values**.

---

3. title.crew.tsv (Directors & Writers)

3.1 Data Overview

This dataset contains **11,464,885 records** and provides information on the **directors and writers** associated with each title.

Key columns:

- **tconst (string):** The unique title identifier.
- **directors (string):** A comma-separated list of director nconst values.
- **writers (string):** A comma-separated list of writer nconst values.

3.2 Data Completeness & Missing Values

- **33% of the records are missing directors**.
- **58% of the records are missing writers**, meaning many productions do not have associated writers in the dataset.

3.3 Data Consistency & Accuracy

- Some **director and writer nconst values do not exist in name.basics.tsv**, affecting referential integrity.

3.4 Data Distribution & Summary

- The **most frequently credited directors** include **Steven Spielberg, Alfred Hitchcock, and Martin Scorsese**.
- The **most frequently credited writers** include **Quentin Tarantino and Christopher Nolan**.

3.5 Key Relationships Across Datasets

- This dataset connects to **name.basics.tsv** via nconst.

3.6 Data Issues & Remediation Plan

- Identify missing **nconst references** and correct or remove invalid values.
- **Handle movies with missing directors and writers** appropriately.

---

4. title.episode.tsv (Episode Details)

4.1 Data Overview

The title.episode.tsv dataset contains information about TV series episodes, linking them to their parent series. It consists of **8,815,771 rows**, with each row representing an individual episode.

Key columns:

- **tconst (string):** The unique identifier for each episode.
- **parentTconst (string):** The tconst of the parent TV show.
- **seasonNumber (integer):** The season number in which the episode appears.
- **episodeNumber (integer):** The episode number within the season.

4.2 Data Completeness & Missing Values

- **13% of records have missing season numbers**, making it difficult to determine the episode's placement within a series.
- **17% of records have missing episode numbers**, indicating incomplete sequencing.
- **parentTconst is missing for 0.4% of records**, which means some episodes are not linked to a parent series.

4.3 Data Consistency & Accuracy

- **Duplicate records found** where multiple episodes share the same tconst.
- **parentTconst values do not always exist in title.basics.tsv**, causing referential integrity issues.
- Some **season and episode numbers contain non-numeric characters**, likely due to data input errors.

4.4 Data Distribution & Summary

- The **average number of episodes per series is around 20**, but some shows have **over 1,000 episodes**, which may require verification.
- **Season 1 has the highest number of episodes**, likely because many shows don't continue beyond their first season.

4.5 Key Relationships Across Datasets

- The parentTconst field should map to tconst in **title.basics.tsv**.
- This dataset is crucial for **analyzing episode count trends and TV series longevity**.

4.6 Data Issues & Remediation Plan

- **Fix incorrect season and episode numbering** to maintain proper sequencing.

- **Remove duplicate records** to avoid redundancy.
- **Ensure all parentTconst values exist in title.basics.tsv** before integration.

---

5. title.ratings.tsv (Movie Ratings)

5.1 Data Overview

The title.ratings.tsv dataset contains IMDb ratings for movies and TV shows. It consists of **1,536,010 rows**, with each row representing a title with a rating.

Key columns:

- **tconst (string):** The unique identifier for the title.
- **averageRating (float):** The IMDb rating on a scale of 1-10.
- **numVotes (integer):** The number of votes received for the title.

5.2 Data Completeness & Missing Values

- **No missing values** were found in this dataset.
- However, **some titles in title.basics.tsv are missing from title.ratings.tsv**, meaning they do not have IMDb ratings.

5.3 Data Consistency & Accuracy

- **No duplicate tconst values**, ensuring data uniqueness.
- **Outliers detected** where some titles have an **unrealistically high number of votes (millions),** likely indicating biased voting patterns.

5.4 Data Distribution & Summary

- The **average rating across all titles is around 6.5**, with **most movies rated between 5 and 8**.
- **Highly rated titles have lower vote counts**, while **popular movies tend to have ratings around 7.0**.

5.5 Key Relationships Across Datasets

- The tconst field should map to title.basics.tsv.
- This dataset is essential for **trend analysis and ranking of movies and TV shows**.

5.6 Data Issues & Remediation Plan

- **Investigate extreme values in vote counts** to detect manipulation.
- **Ensure all movies and shows in title.basics.tsv exist in title.ratings.tsv**, if possible.

---

6. title.akas.tsv (Alternate Titles)

6.1 Data Overview

The title.akas.tsv dataset contains alternative titles of movies and TV shows in different regions and languages. It includes **51,409,880 rows**.

Key columns:

- **titleId (string):** The unique identifier for the title.

- **ordering (integer):** The display order of the alternate title.
- **title (string):** The alternate title.
- **region (string):** The country/region where this title is used.
- **language (string):** The language of the alternate title.
- **isOriginalTitle (integer):** Flag indicating if this is the original title.

6.2 Data Completeness & Missing Values

- **22% of records have missing region codes**, affecting country-specific title analysis.
- **35% of records have missing language codes**, making language-based filtering difficult.

6.3 Data Consistency & Accuracy

- **Some region codes are not valid country codes**, requiring standardization.
- **Duplicate alternate titles exist**, where the same title is repeated for multiple regions.

6.4 Data Distribution & Summary

- **English titles dominate** the dataset, followed by Spanish, French, and German titles.
- **The same movie often has multiple alternate titles**, especially in non-English-speaking regions.

6.5 Key Relationships Across Datasets

- titleId should map to tconst in title.basics.tsv.
- **This dataset helps in region-specific and multilingual analysis of movie popularity.**

6.6 Data Issues & Remediation Plan

- **Standardize region and language codes** based on ISO country codes.
- **Remove duplicate entries** for better clarity.

---

7. title.principals.tsv (Cast & Crew Details)

7.1 Data Overview

The title.principals.tsv dataset contains information about **cast and crew members** involved in movies and TV shows. It consists of **90,984,102 rows**.

Key columns:

- **tconst (string):** The unique identifier for the title.
- **ordering (integer):** The ranking order of the person’s credit in the title.
- **nconst (string):** The unique identifier for the person.
- **category (string):** The person's role (e.g., actor, director, writer).
- **job (string):** The specific job title (e.g., cinematographer, composer).
- **characters (string):** The character name played (for actors).

7.2 Data Completeness & Missing Values

- **10% of records have missing job values**, mostly for actors.

- **15% of records have missing characters values**, meaning some actors are listed without character names.

7.3 Data Consistency & Accuracy

- **Duplicate records detected** where the same person is credited multiple times for the same title.
- **Some nconst values do not exist in name.basics.tsv**, affecting referential integrity.

7.4 Data Distribution & Summary

- **Most records belong to actors**, followed by directors and producers.
- **Some movies have over 100 credited individuals**, which might require filtering for primary cast members.

7.5 Key Relationships Across Datasets

- tconst maps to title.basics.tsv, linking movies to cast/crew members.
- nconst maps to name.basics.tsv, linking people to their roles.

7.6 Data Issues & Remediation Plan

- **Ensure all nconst values exist in name.basics.tsv** before transformation.
- **Remove redundant credits** for clarity.

Final Summary

Our team has identified **various data quality issues** across these datasets, including missing values, duplicate records, and referential integrity concerns. To address these:

- We are **cleaning, standardizing, and validating data using Alteryx and YData Profiling**.
- We are **enforcing primary and foreign key constraints** in **Snowflake** to maintain data integrity.
- We are **documenting all transformations** to ensure consistency in reporting.

---

8. language_codes_profile

8.1 Data Overview

The **language_codes_clean.csv** dataset contains ISO language codes along with their corresponding language names. It consists of **183 rows**, with each row representing a unique language and its associated ISO codes.

Key columns:

- **ISO Language Names (string):** The official name of the language.
- **Set 1 (string):** Primary ISO 639-1/2/3 codes for the language.
- **Set 2T (string):** ISO 639-2/T codes (terminological).
- **Set 2B (string):** ISO 639-2/B codes (bibliographic).
- **Set 3 (string):** ISO 639-3 codes.
- **Scope (categorical):** Specifies if the language is an **Individual language or a Macrolanguage**.
- **Type (categorical):** Specifies if the language is **Living, Ancient, Constructed, or Historical**.
- **Endonym (string):** The native name of the language.

- **Other Names (string):** Alternate or commonly used names.
- **Notes (string):** Additional information about the language.

8.2 Data Completeness & Missing Values

- **288 missing values (15.7%)** were found across multiple fields.
- **No missing values** in the **ISO Language Names** or **Set 1, Set 2T, Set 2B, and Set 3 columns**.
- **Significant missing values in the "Other Names" (67.8%) and "Notes" (89.6%) columns**, which may impact analysis.

8.3 Data Consistency & Accuracy

- **No duplicate ISO Language Names**, ensuring data uniqueness.
- **Distinct values for all sets of ISO codes**, maintaining referential integrity.
- **Standardized format for ISO codes** to prevent inconsistencies.

8.4 Data Distribution & Summary

- **Most languages belong to the "Living" category (173 out of 183).**
- **4 languages are classified as "Ancient", 1 as "Constructed", and 1 as "Historical".**
- **149 languages are classified as "Individual" and 34 as "Macrolanguages".**
- **Word clouds indicate high frequency of commonly spoken languages such as "Norwegian", "Gaelic", and "Haitian".**

8.5 Key Relationships Across Datasets

- The **ISO language codes** may be used for **language-based analysis and classification** in other datasets.
- **Mapping to other datasets can help in multilingual text processing, translation services, and NLP applications.**

8.6 Data Issues & Remediation Plan

- **Address missing values** in "Other Names" and "Notes" by referencing external sources.
- **Ensure consistency in endonyms** by applying language-specific standardization rules.
- **Validate ISO codes** against the official **ISO 639 registry** for accuracy.