Understanding of the Data

TITLE.BASICS.TSV:

This report provides a formal assessment of the title.basics.tsv dataset structure and identifies necessary preparatory steps for robust Exploratory Data Analysis (EDA). The analysis is crucial for ensuring data integrity and optimizing performance before statistical and temporal investigations.

Statistic	Value	Implication for EDA
Total Records	**12,097,750**	Due to the high row count, initial processing (such as grouping and aggregation) should prioritize efficient data types to prevent memory overflow.
Memory Footprint	**≈ 4.07 GB**	Suggests that optimizing object (string) columns and avoiding memory-intensive operations (like large cross-joins) without prior filtering is necessary.

Columns Data And its Understanding:

Column Name	Current Type	Data Quality Issue	Required Transformation
runtimeMinutes	object	**Critical Non-Numeric Data.** The column contains non-convertible characters (typically \N), preventing direct numerical analysis.	**Coercion:** Must be converted to a numeric type (float64 or Int64) with errors coerced to **NaN**.
startYear	float64	Presence of NaN values	**Casting:** Convert to Int64 (nullable

Column Name	Current Type	Data Quality Issue	Required Transformation
		forces a non-integer type.	integer) after ensuring all non-NaN values are whole numbers.
genres	object	**Multi-Label String Format.** Values are delimited by commas (e.g., "Action,Drama"), rendering direct frequency counting impossible.	**Parsing:** Must be split and "exploded" to create a standard, single-label row format for accurate genre distribution analysis.
titleType	category	Highly optimized categorical data.	**None.** Ideal for immediate segmentation (e.g., comparing runtime of 'movie' vs. 'short').
isAdult	int8	Binary flag (0 or 1).	**Casting:** Convert to a boolean type (bool) for cleaner filtering and logical operations.

Missing Data:

Handling missing values is paramount, as different columns exhibit different—and expected—levels of data incompleteness.

Column	Approximate Null Count	Missingness Percentage	Interpretation and Strategy
runtimeMinutes	**≈ 7.8 Million**	**$\approx 64\%$**	**High Information Loss.** Analysis involving duration must be conducted on a substantial **subset** of

Column	Approximate Null Count	Missingness Percentage	Interpretation and Strategy
			the data. Imputation (filling with mean/median) is generally inadvisable due to the vast difference between runtimes of "shorts" and "features."
endYear	**≈ 11.9 Million**	**$\approx 98\%$**	**Expected Incompleteness.** This primarily applies to ongoing TV series and movies. **Strategy:** This column should only be utilized when calculating the *lifespan* of completed TV shows; it should otherwise be ignored for general analysis.
genres	**$\approx 530,000$**	**$\approx 4.4\%$**	**Low Incompleteness.** Missing values here can be safely handled by either dropping the records or assigning them a categorical placeholder (e.g., "Unspecified") without compromising overall genre statistics.

Title.Rating.TSV:

Statistic	Value	Implication for EDA
Total Records	**$1,605,930$**	Only $\approx 13\%$ of the 12 million titles in the BASICS file have a rating. This confirms that the majority of titles (shorts, episodes, etc.) are unrated.

Statistic	Value	Implication for EDA
Memory Footprint	**≈ 120 MB**	Excellent memory efficiency, making this table ideal for direct loading and joining operations.
Missing Data	**Zero**	The dataset is exceptionally clean, eliminating the need for missing value imputation before analysis.

Column Analysis:

Column Name	Data Type	Analytical Role & Integration Strategy
tconst	category	**Primary Join Key.** This column acts as the foreign key required to link rating metrics back to descriptive features (e.g., genre, runtime) in the TITLE.BASICS.TSV file.
averageRating	float32	**Quality Metric.** This is the simple arithmetic mean of all votes. **EDA Tip:** This metric is easily skewed by low vote counts and must be weighted (see Section 3).
numVotes	int32	**Reliability Metric.** Indicates the statistical reliability of the averageRating.

Title.Crew.Tsv:

The CREW file maintains the same dimensionality as the BASICS file, indicating that every title record has a corresponding crew record, even if the crew fields themselves are null.

Statistic	Value	Implication for EDA
Total Records	**$12,097,750$**	This file must be joined with TITLE.BASICS.TSV using the **tconst** key. Since it has the same number of rows, it suggests a 1:1 relationship with the basics file.

Statistic	Value	Implication for EDA
Memory Footprint	**≈ 2.03 GB**	Moderate memory usage for its size, but efficiency is hampered by all three columns being loaded as `object` (string) types.

Column Analysis:

Column Name	Current Type	Data Quality Issue	Required Transformation
tconst	object	**Suboptimal Type.** The unique identifier (join key) is loaded as a generic string.	**Casting:** Convert to `category` for memory efficiency and optimal join performance with other IMDB tables.
directors	object	**Multi-Value Field.** Contains comma-separated IMDB Person IDs (e.g., `nm0005690,nm0721526`) which are critical for network analysis.	**Parsing & Exploding:** Must be split into lists and then "exploded" to create one row per title-director relationship, enabling director-centric analysis (e.g., director influence).
writers	object	**Multi-Value Field.** Contains comma-separated IMDB Person IDs.	**Parsing & Exploding:** Similar to directors, this field must be processed for writer-centric network analysis.

Missing Data:

Column	Approximate Null Count	Missingness Percentage	Interpretation and Strategy
directors	**≈ 5.3 Million**	**$\approx 44\%$**	**Significant Missingness.** A large portion of titles lack director information. **Strategy:** Filter for non-null values before running director network analysis or

Column	Approximate Null Count	Missingness Percentage	Interpretation and Strategy
			calculating director-based statistics.
writers	**≈ 6.1 Million**	**$\approx 50\%$**	**Significant Missingness.** Half the dataset lacks writer data. This is expected, as many shorts or non-narrative titles may not credit a writer.

NAME.BASICS.TSV:

This is the largest personnel-related file, containing nearly 15 million records, representing the unique individuals tracked in the IMDB database.

Statistic	Value	Implication for EDA
Total Records	**$14,905,821$**	This table acts as the master lookup dictionary for all individuals. It must be joined with the exploded TITLE.CREW.TSV file to transform nconst IDs into readable names.
Memory Footprint	**≈ 3.11 GB**	Moderate memory usage for its size. The primary goal is to convert the object type nconst to category for improved join performance.

Column Analysis:

Column Name	Current Type	Data Quality Issue	Required Transformation
nconst	object	**Key Column Suboptimal Type.** The unique personnel ID is loaded as a generic string.	**Casting:** Convert to category for memory efficiency and optimal join performance.

Column Name	Current Type	Data Quality Issue	Required Transformation
primaryName	object	**Identification Field.** Contains the individual's full name.	**None.** Ready for use in visualizations, tables, and top-list rankings.
birthYear	float64	**Temporal Data.** Loaded as float due to NaN values (e.g., birth year unknown, or living).	**Casting:** Convert to Int64 (nullable integer). Used to calculate age at death or career span.
deathYear	float64	**Temporal Data.** Primarily NaN for living individuals.	**Casting:** Convert to Int64 (nullable integer). Used to calculate career longevity for deceased individuals.
primaryProfession	object	**Multi-Value Field.** Contains comma-separated professions (e.g., "actor,director,producer").	**Parsing & Exploding:** Must be split and "exploded" to analyze the distribution of individuals across different professions (e.g., how many actors are also directors).

Missing Data:

Column	Approximate Null Count	Missingness Percentage	Interpretation and Strategy
birthYear	**$\approx 660,000$**	**$\approx 4.4\%$**	**Low Missingness.** Missing records should be excluded from age-based cohort analysis but can be tolerated for name-based lookups.
deathYear	**≈ 14.65 Million**	**$\approx 98\%$**	**Expected Incompleteness.** The vast majority of individuals in the dataset are assumed to be living or their death date is unrecorded.
primaryProfession	**≈ 2.97 Million**	**$\approx 20\%$**	**Significant Missingness.** One-fifth of the records lack a profession tag. **Strategy:** Exclude these records from professional distribution analyses, or categorize them as 'Unspecified.'

TITLE.PRINCIPALS.TSV:

Statistic	Value	Implication for EDA
Total Records	**$\approx 500,000$**	This is a significantly smaller, curated list compared to the ≈ 12 million records in TITLE.BASICS.TSV. It likely only lists the top 10 or so principals per title.

Statistic	Value	Implication for EDA
Columns	**6**	It is a highly detailed linking table, containing two foreign keys (tconst and nconst) and four descriptive fields for the relationship.

Column Analysis:

Column Name	Current Type	Data Quality Issue	Required Transformation
tconst	object	**Title Join Key.** The unique title identifier is loaded as a generic string.	**Casting:** Convert to category for efficient joining with TITLE.BASICS.TSV.
nconst	object	**Personnel Join Key.** The unique person identifier is loaded as a generic string.	**Casting:** Convert to category for efficient joining with NAME.BASICS.TSV.
ordering	int64	**Rank/Importance.** Indicates the order in which the principal is credited.	**Casting:** Convert to a smaller integer type (e.g., int8) if the maximum ordering value is small, to save memory.
characters	object	**String Array.** Contains character names, often encapsulated in an array-like string (e.g., ["Self"]).	**Parsing & Cleaning:** Must clean/parse the array brackets ([] and quotes) and split the string to extract readable character names for analysis.
job	object	**Specific Role.** Provides further detail beyond category (e.g., if category is 'producer', job might be 'executive producer').	**Cleaning & Casting:** Handle NaN values and convert to category to analyze job distribution.
category	object	**Broad Role.** Specifies the broad type of	**Casting:** Convert to category for efficient

Column Name	Current Type	Data Quality Issue	Required Transformation
		principal (e.g., 'actor', 'director', 'writer').	analysis of role distribution.

Relationship Analysis:

Relationship	Join Type	Records Involved	Records Matched	Interpretation
Basics $\leftrightarrow$ Ratings	Primary Key Join	12.1M $\leftrightarrow 1.6$M	1.6 Million	Confirms that only **13.2%** of all titles have associated rating data.
Basics $\leftrightarrow$ Crew	Left Join Potential	12.1M $\leftrightarrow 12.1$M	12.1 Million	Confirms a **$1:1$** relationship. Every title in the basics file has a corresponding entry in the crew file (even if director/writer fields are null).
People ID (nconst)	Lookup Key	14.9 Million Total	N/A	This key will be used to enrich the final merged dataset by joining with name_basics to retrieve primary name and profession.

The three-step merge strategy prioritizes quality and measurable performance over simple coverage of all titles.

Step	Operation	Key	Output Dataset	Analytical Justification
1	**title_basics INNER JOIN ratings**	tconst	≈ 1.6 Million Rows	**Filters for Quality.** The INNER JOIN discards 87% of the data, retaining only titles for which quality and popularity metrics (averageRating, numVotes) are available.
2	**Result LEFT JOIN crew**	tconst	≈ 1.6 Million Rows	**Adds Personnel.** Ensures all 1.6 million rated titles retain their records and gains director/writer IDs. Missing director/writer IDs will appear as NaN in the crew columns.
3	**Result JOIN name_basics**	nconst	Varies (Post-Explosion)	**Enriches Data.** This final step is essential for converting numerical IDs into readable names, making the crew analysis usable (e.g., director nm0005690 becomes 'William K.L. Dickson').