Dataset Description for ExoHabitAI

1. Overview

This dataset comes from the **NASA Exoplanet Archive**. It includes detailed information on confirmed exoplanets, focusing on those discovered by the **TESS** (Transiting Exoplanet Survey Satellite) mission and marked as primary records.

- Total records: 39,212 exoplanets
- Total attributes: 289 columns
- Granularity: One row corresponds to one confirmed exoplanet
- Temporal coverage: Discoveries from 1992 to 2026
- Discovery constraint: disc_facility contains “TESS” and default_flag = 1

The dataset combines planetary, orbital, stellar, observational, and administrative parameters, making it one of the most thorough structured resources for exoplanet studies.

2. Attribute Groups and Their Roles

2.1 Identification and Cross-Referencing Attributes

Key columns:

- pl_name (official exoplanet designation),
- hostname (name of the host star system),
- pl_letter (alphabetical order of planet discovery in the system),
- hd_name, hip_name (Henry Draper and Hipparcos catalog identifiers),
- tic_id (TESS Input Catalog identifier),
- gaia_dr2_id, gaia_dr3_id (Gaia mission astrometric identifiers)

These attributes **uniquely identify each planet** and allow for cross-matching with external astronomical catalogs. While they are important for traceability and verification, they do not add analytical value in predictive or statistical modeling tasks.

2.2 Discovery and Detection Metadata

Key columns:

- discoverymethod (primary technique used to detect the planet),
- disc_year (year the planet was confirmed),
- disc_facility (observatory or mission responsible for discovery),
- disc_telescope (telescope used for detection),
- disc_instrument (specific instrument used),

- Detection flags such as tran_flag, rv_flag, micro_flag (binary indicators of detection techniques applied)

These attributes describe how a planet was detected and are **useful for understanding survey bias** but are **not intrinsic planetary properties**.

2.3 Planetary Physical Characteristics

Key columns:

- Size: pl_rade (planetary radius measured in Earth radii)
- Mass: pl_bmasse, pl_masse, pl_msinie (all in Earth masses)
- Density: pl_dens (bulk density of the planet)
- Thermal/Energy: pl_eqt (estimated equilibrium temperature), pl_insol (stellar energy received relative to Earth)

Radius data is more complete than mass data due to TESS's transit method, making density values sparse and mass measurements redundant. These attributes are key to planetary classification (rocky, gaseous, mini-Neptune) and are **most relevant for studies focused on habitability.**

2.4 Orbital and Transit Parameters

Key columns:

pl_orbper, pl_orbsmax, pl_orbeccen, pl_orbincl, pl_trandep, pl_trandur, pl_tranmid, pl_imppar

These parameters describe orbital geometry and transit behavior, which directly affect stellar energy received**, climate stability**, and observational accuracy. However, many associated uncertainty and limit-flag columns have a **high number of missing values.**

2.5 Host Star Properties

Key columns:

st_teff, st_rad, st_mass, st_lum, st_met, st_age, st_logg

These attributes describe the **physical and evolutionary properties** of the host star, which determine the **radiation environment** and **long-term conditions** experienced by its orbiting planets.

3. Attributes That Can Be Safely Ignored

To reduce dataset complexity without losing essential information, the following attributes can be excluded:

3.1 Redundant Unit Conversions

Jupiter-unit duplicates (e.g., pl_radj, pl_massj, pl_bmassj) that repeat Earth-unit values.

3.2 Administrative and Reference Fields

Citation and update metadata (disc_refname, pl_refname, st_refname, sy_refname, rowupdate, releasedate) that track provenance only.

3.3 Error Margins and Limit Flags (Optional)

Columns ending in err1, err2, and lim, which represent uncertainties and are unnecessary for baseline analysis.

3.4 Catalog and Positional Identifiers

Sky coordinates and catalog IDs (ra, dec, glon, gaia_dr2_id, etc.) not required for planetary characterization.

4. Conclusion

- The dataset has many dimensions and is sparse, with many attributes having significant **missing values**.
- **Radius-based parameters are more complete** than mass-based measurements because of the Transit detection bias.
- There is a strong observational bias since most planets were discovered using the Transit method.
- Only a **specific subset of planetary and stellar attributes** is needed for meaningful analysis.
- Removing redundant units, administrative metadata, and sparsely populated uncertainty columns makes it easier to use without impacting scientific validity.