Milestone 1: Data Collection, Description, and Preprocessing Strategy

Dataset Description

Data Source
The primary dataset is sourced from the **NASA Exoplanet Archive (Planetary Systems Table)**. This is a public astronomical archive containing data on all confirmed exoplanets.

- **Total Raw Observations:** ~39,000 rows
- **Unique Confirmed Planets:** ~6,100 planets
- **Format:** CSV (Comma Separated Values)

The "Duplicate" Nature of Astronomical Data

The raw dataset contains approximately 39,000 entries for only 6,100 planets. This is because the archive records **historical scientific publications**. If a specific planet (e.g., *Kepler-186 f*) has been observed by five different research teams over a decade, it appears five times in the dataset with slightly different measurement values.

For Machine Learning purposes, treating these rows as independent data points would introduce **Data Leakage** and **Bias**. Therefore, a critical part of our data description involves defining the "Best Representative Row" for each planet.

Feature (Selected Attributes)

From the 280+ columns available in the raw data, we have selected the following features based on their relevance to planetary habitability physics.

Planetary Parameters

Feature Name (Raw)	Standardized Name	Unit	Description
pl_name	**planet_name**	String	Unique identifier for the exoplanet.
pl_rade	**radius_earth**	Earth Radii (R_E)	The radius of the planet compared to Earth. Crucial for determining if a planet is rocky or gaseous.

Feature Name (Raw)	Standardized Name	Unit	Description
pl_masse	**mass_earth**	Earth Masses (M_E)	The mass of the planet. Determines gravity and atmosphere retention.
pl_orbper	**orbital_period**	Days	Time taken to complete one orbit around the host star.
pl_orbsmax	**semimajor_axis**	AU	Average distance from the star. Determines the thermal environment.
pl_eqt	**eq_temp_k**	Kelvin (K)	Theoretical equilibrium temperature of the planetary surface.
pl_dens	**density**	g/cm^3	Bulk density, used to infer composition (Iron, Rock, Water, Gas).

Stellar Parameters

Feature Name (Raw)	Standardized Name	Unit	Description
hostname	**host_star_name**	String	Name of the host star.
st_teff	**star_temp_k**	Kelvin (K)	Effective surface temperature of the star.
st_lum	**star_luminosity**	log(Solar)	Total energy output of the star relative to the Sun.
st_spectype	**star_spectype**	String	Spectral classification (e.g., G2V, M3). Indicates star age, size, and radiation stability.

Data Cleaning & Deduplication

Allowing multiple rows for the same planet violates this assumption and causes the model to memorize specific planets rather than learning generalizable physics.

Implementation:

1. **Completeness Sorting:** We calculate the number of **NaN** (missing) values for every row.
2. **Selection:** For every unique planet name, we retain only the row with the **least missing data**.
3. **Result:** The dataset is reduced from ~39,000 observations to ~6,100 unique, high-quality planetary profiles.

Missing Data Handling (Imputation Strategy)

Astronomical data often suffers from missing values due to observational limitations (e.g., Mass is harder to measure than Radius).

Physics-Based Recovery (Unit Conversion)
Before statistical imputation, we recover real data stored in alternative units. The NASA archive often stores data in "Jupiter Units" when "Earth Units" are missing.

- **Formula:** $Radius_{Earth} \approx 11.209 \times Radius_{Jupiter}$
- **Formula:** $Mass_{Earth} \approx 317.8 \times Mass_{Jupiter}$
- This will give us more ground real values of the dataset

Statistical Imputation
For remaining gaps, we use **Median Imputation**.

- **Why:** Exoplanet data is heavily right-skewed (power-law distribution). The Mean is sensitive to outliers (massive gas giants), whereas the Median provides a robust central tendency for typical planets.

Outlier Detection & Handling

Outliers are often errors to be removed. In Exoplanetary Science, outliers are often **real, massive objects** (Hot Jupiters, Brown Dwarfs).

- **Lower Bound (Physics Floor):** We enforce strict physical limits. Values ≤ 0 for Mass, Radius, or Temperature are removed as they represent measurement errors.
- **Upper Bound (No Capping):** We explicitly **do not** remove or cap massive planets.
    - A planet with $100 M_E$ is physically valid. Removing it would bias the model against gas giants.
    - So, We use robust scaling methods (see Section 3.5) to handle these large values without deleting them.

Feature Engineering

We derive new "synthetic" features that combine multiple physical parameters to give the ML model stronger signals regarding habitability.

1. **Habitability Score (ESI Proxy):**
    - A calculated index based on the geometric mean of a planet's similarity to Earth in terms of Radius and Temperature.
    - *Formula:* $Score = \sqrt{(1 - | \frac{R - R_\oplus}{R + R_\oplus} |)^{0.57} \times (1 - | \frac{T - 288}{T + 288} |)^{1.07}}$
2. **Stellar Compatibility:**

    - Maps Spectral Types (O, B, A, F, G, K, M) to a numerical score.
    - G and K stars (Sun-like) receive high scores (1.0 - 0.9). M stars (volatile red dwarfs) receive medium scores. O/B stars (short-lived) receive low scores.

3. **Orbital Stability:**
    - A logarithmic interaction feature between Orbital Period and Semi-Major Axis to represent the orbital dynamics of the system.

Categorical Encoding & Feature Scaling

Encoding:

- **Feature: star_class** (derived from **star_spectype**).
- **One-Hot Encoding**.
- **Justification:** Spectral types are nominal categories without a strict linear ordinal relationship suitable for regression.

Scaling:

- **RobustScaler**.
- Standard scalers (Z-score) use Mean and Variance. Because we retained massive "Monster" planets (outliers), the Mean is distorted. **RobustScaler** uses the Median and Interquartile Range (IQR), ensuring that Earth-like planets are scaled appropriately even in the presence of massive Gas Giants.

Target Variable Creation

Since "Habitability" is not a direct column in the raw data, we generate a ground-truth label for supervised learning.

- **Label: habitable_binary** (0 or 1).
- **Criteria (Conservative):**
    1. **Radius:** $0.5 R_E$ to $1.6 R_E$ (Likely Rocky Surface).
    2. **Temperature:** $200 K$ to $330 K$ (Potential for Liquid Water).

Conclusion

The preprocessing pipeline outlined above transforms raw, noisy, and duplicated astronomical data into a clean, physics-compliant dataset.