🌌 ExoHabitAI: Dataset Documentation & Analysis Report

Project: ExoHabitAI – Predictive Modeling for Exoplanet Habitability

Author: Samridhi Gupta

Role: Machine Learning Intern, Infosys Springboard

Date: February 15, 2026

---

1. Executive Summary

This document provides a comprehensive technical analysis of the dataset used for the **ExoHabitAI** project. The primary objective is to develop a Machine Learning model capable of predicting the habitability potential of exoplanets based on their physical and stellar characteristics.

The analysis transforms raw astronomical data into a refined, "ML-ready" format, identifying key candidates for life by engineering new metrics such as the **Earth Similarity Index (ESI)**.

---

2. Dataset Overview 🌍

The foundation of this project is the **NASA Exoplanet Archive**, specifically the Planetary Systems Composite Data table. This archive is the global standard for confirmed exoplanet discoveries.

Dataset Profile	Details
Source Name	NASA Exoplanet Archive (Planetary Systems Composite)
Original Dimensions	**39,386** Observations × **289** Features
Refined Analysis Scope	**6,107** Unique Exoplanets
Primary Discovery Methods	Transit, Radial Velocity

Key Insight: To ensure data integrity, we filtered for the default_flag = 1 parameter. This reduces redundancy by selecting only the single most accurate set of parameters for each confirmed planet, removing duplicate or outdated observations.

---

3. The Data Dictionary (Key Features) 📖

While the raw dataset contains 289 columns, the **ExoHabitAI** model focuses on **8 critical features** that directly influence a planet's potential to host life.

Planetary Parameters

- **pl_name (Planet Name):** Unique identifier for the exoplanet.
- **pl_rade (Planet Radius):** Measured in **Earth Radii ($R_{\oplus}$)**. This is the primary indicator of composition (e.g., Rocky vs. Gas Giant).
- **pl_bmasse (Planet Mass):** Measured in **Earth Mass ($M_{\oplus}$)**. Essential for calculating density and gravity.
- **pl_orbper (Orbital Period):** The time (in days) it takes the planet to complete one orbit around its star.
- **pl_orbsmax (Semi-Major Axis):** The average distance from the host star in **Astronomical Units (AU)**.

Stellar & Environmental Parameters

- **pl_eqt (Equilibrium Temperature):** The theoretical surface temperature of the planet in **Kelvin (K)**.
- **st_teff (Stellar Effective Temperature):** The surface temperature of the host star.
- **st_rad (Stellar Radius):** The size of the host star in **Solar Radii ($R_{\odot}$)**.

---

4. Data Quality & Missing Values Analysis ⚠️

Real-world astronomical data is rarely perfect. A thorough gap analysis was conducted to identify limitations and define imputation strategies.

Missing Data Profile

- **High Severity:**
  - **Insolation Flux (pl_insol):** Missing in **85%** of observations.
  - **Planetary Density (pl_dens):** Missing in **81%** of observations.
- **Moderate Severity:**
  - **Planet Mass (pl_bmasse):** Missing in **~50%** of observations (mostly for Transit discoveries where only radius is known).

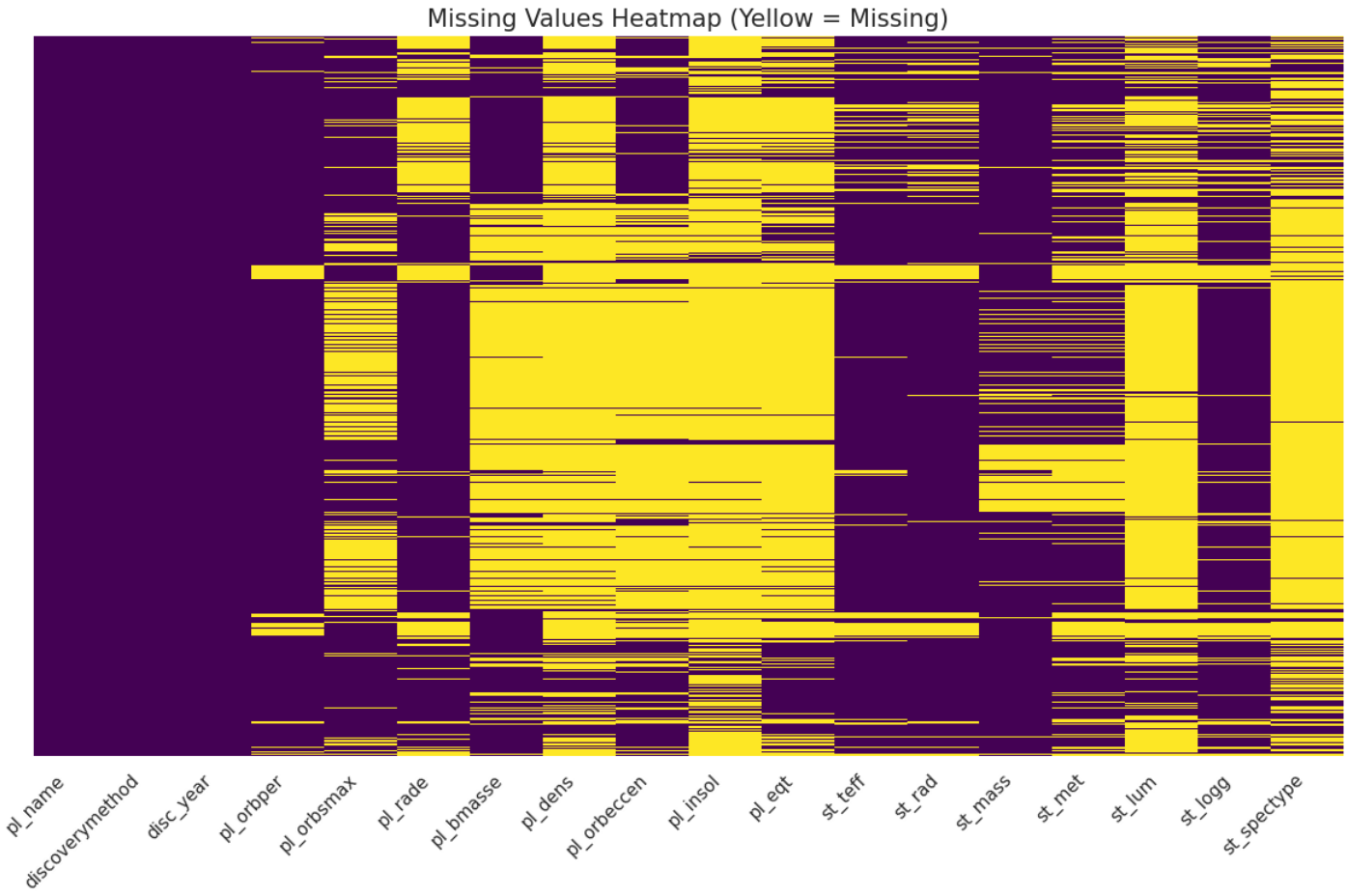

Figure 1: Heatmap visualizing data gaps. Yellow bands indicate missing values, highlighting the need for advanced imputation strategies rather than simple deletion.

The Fix: Physics-Based Imputation

Instead of dropping these valuable rows or filling them with zeros (which would imply a "dead" star), we utilized **Astrophysical Laws** for imputation:

1. **Stefan-Boltzmann Law:** Used to derive **Luminosity** and **Insolation Flux** where Stellar Radius and Temperature were available.
2. **Mass-Radius Relations:** Used to estimate missing mass for rocky planets based on probabilistic density models.

---

5. Statistical Summary: The "Range of Life" 📊

The dataset represents a vast diversity of worlds, from scorching "Hot Jupiters" to frozen "Super-Earths."

Parameter	Minimum	Maximum	Observation
Planet Radius	0.3 $R_{\oplus}$	87.2 $R_{\oplus}$	Ranges from Mars-sized objects to brown dwarfs.

Parameter	Minimum	Maximum	Observation
Orbital Period	~2 hours	>1000 years	Includes "Ultra-Short Period" planets and distant giants.
Equilibrium Temp	50 K	4,050 K	Spans from frozen nitrogen oceans to vaporized rock.

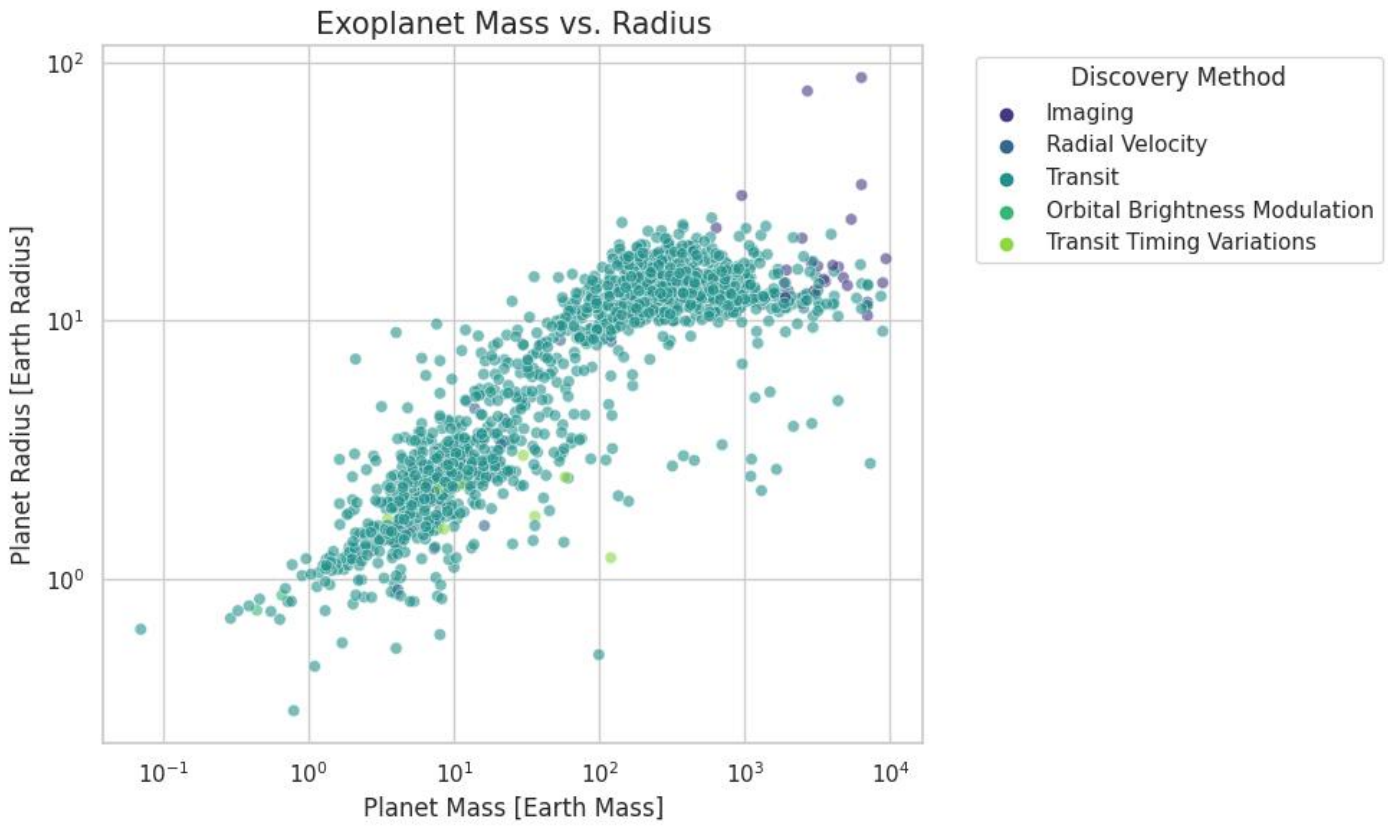

Figure 2: The "Exoplanet Population Diagram." The logarithmic scale clearly separates the dense cluster of Rocky Worlds (bottom left) from the massive Gas Giants (top right).

---

6. Target Variable Definition (The "Habitability" Logic) 🎯

Since the raw NASA data does not contain a "Habitable" label, a custom target variable was engineered for Supervised Learning.

Criteria for Is_Habitable Class:

A planet is classified as a potential candidate (**1**) if it meets **ALL** of the following:

1. **Composition:** Must be Rocky (Radius < 1.6 $R_{\oplus}$ **OR** Mass < 10 $M_{\oplus}$).
2. **Temperature:** Must fall within the liquid water range (180 K < T_{eq} < 320 K).
3. **Insolation:** Must receive stable stellar energy (0.25 < Flux < 2.2 $S_{\oplus}$).

The Earth Similarity Index (ESI)

To provide a granular ranking, we implemented the ESI formula:

$$ESI = \prod_{i=1}^{n} \left(1 - \left| \frac{x_i - x_{i0}}{x_i + x_{i0}} \right| \right)^{\frac{w_i}{n}}$$

Where x represents planetary properties (Radius, Density, Temperature) compared to Earth's reference values.

---

7. Discovery Methods & Data Sources

Understanding *how* data is collected helps explain biases (e.g., Transit method favors large planets close to stars).

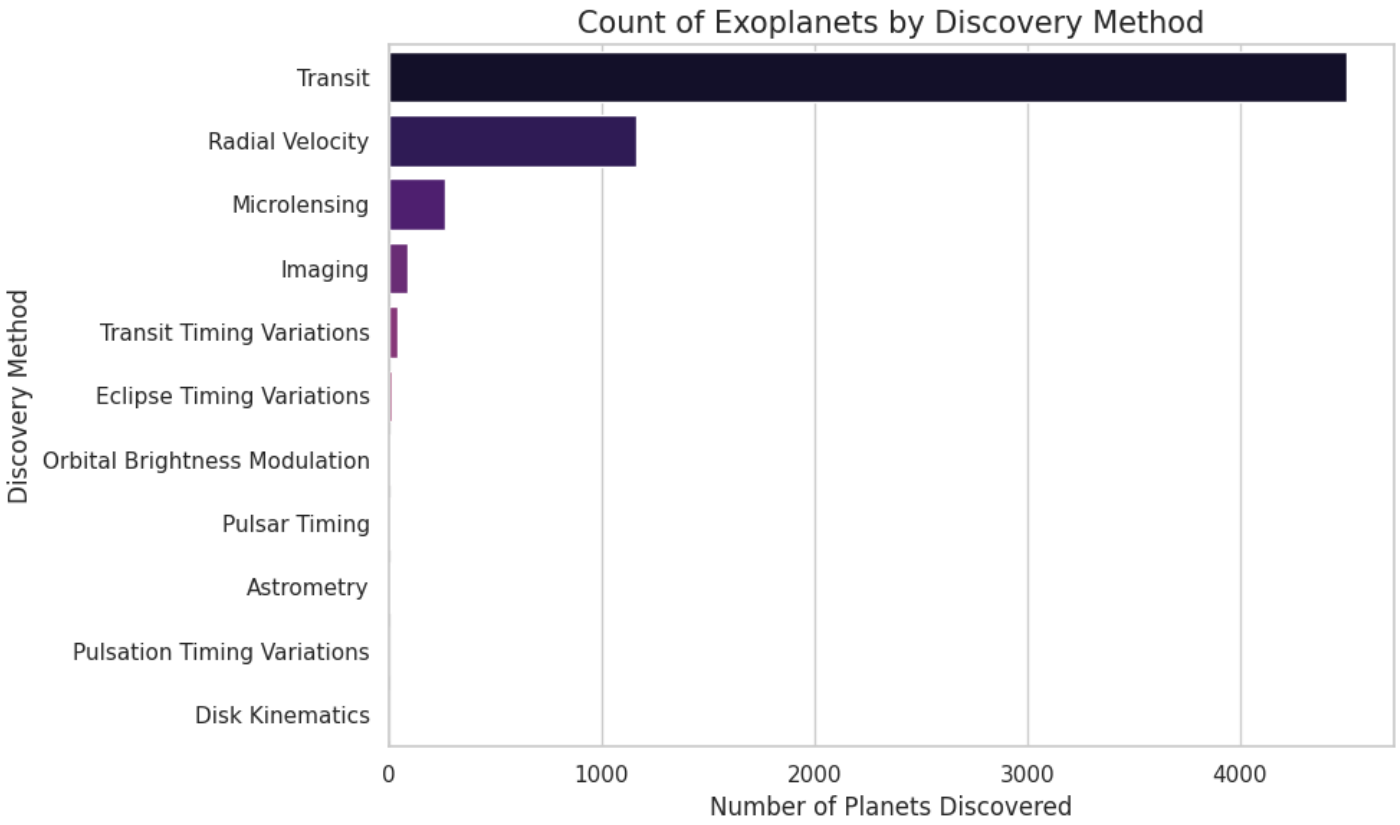

Figure 3: Distribution of Exoplanet Discoveries. The "Transit" method accounts for the vast majority of detections, followed by "Radial Velocity."

---

8. Conclusion & Next Steps

The dataset has been successfully cleaned, engineered, and validated. With **6,107** unique observations and a clearly defined target variable, the project is ready to proceed to **Milestone 2: Machine Learning Model Development**.

The next phase will involve training **Random Forest** and **XGBoost** classifiers to predict the habitability of identifying new candidates from partial data.
