Dataset Description and Exploratory Data Analysis

1.Dataset Overview

The dataset used in this project is sourced from the NASA Exoplanet Archive, a reliable public repository containing confirmed exoplanet data collected from multiple astronomical missions.

- Source: NASA Exoplanet Archive
- Total Records: 720 exoplanets
- Total Features: 92 columns
- Format: Structured tabular data (CSV)

2. Objective of Using the Dataset

The objective of using this dataset is to explore planetary and stellar features that influence exoplanet habitability.

This analysis forms the foundation for building a machine learning model to predict potentially habitable exoplanets.

3. Tools and Technologies Used

Dataset exploration was performed using:

- Programming Language: Python
- Environment: VS Code
- Interface: Jupyter Notebook
- Libraries:
    - pandas – data loading and analysis
    - matplotlib – data visualization

4. Dataset Structure and Inspection

Initial exploration was conducted using:

- df.head() – to view sample records

- df.shape() – to understand dataset dimensions
- df.info() – to inspect column names, data types, and non-null values

Observations:

- The dataset contains 720 rows and 92 columns
- Data includes both:
    - Numerical features (radius, mass, temperature, orbital parameters)
    - Categorical features (planet name, discovery method, stellar type)

6. Missing Value Analysis

Missing values were analyzed using:

- df.isnull().sum()

Key Findings:

- Missing values exist in important parameters such as Planetary mass, Stellar insolation, Equilibrium temperature, Stellar spectral type

8. Visual Data Exploration

A histogram of planetary radius was plotted.

Observations:

- Highest concentration of planets lies in the **0–5 Earth radii range**
- Earth-sized planets are fewer compared to larger planets
- Larger planets dominate due to observational detection bias

9. Identification of Habitability-Related Features

Based on exploration, key habitability features include:

Planetary Features:

- Planet radius
- Planet mass
- Orbital distance

- Equilibrium temperature
- Stellar insolation

Host Star Features:

- Stellar temperature
- Stellar mass
- Stellar radius

10. Key Inferences from the Dataset

- The dataset is feature-rich but contains missing values, requiring preprocessing.
- Most detected exoplanets are larger than Earth, reflecting detection bias.
- Earth-like planets are relatively rare in the observed data.
- Both planetary and stellar properties are essential for habitability analysis.
- The dataset is well-suited for applying machine learning techniques to predict habitability.