Milestone 1: Data Preprocessing Analytics

Project: ExoHabitAI - Exoplanet Habitability Analysis

REPORT OVERVIEW

This report details the outcomes of the high-fidelity data preprocessing pipeline. The objective was to refine the NASA Planetary Systems dataset for supervised learning applications in habitability classification.

KEY PERFORMANCE INDICATORS (KPIs):

- Total Validated Records: 39,315
- Feature Engineering: 10 primary, 3 derived features, 1 target label
- Imputation Strategy: Robust Median-based filling
- Class Labeling: Percentile-based (Threshold: 90th percentile)
- High-Probability Candidates: 3,927 planets
- Low-Probability Candidates: 35,388

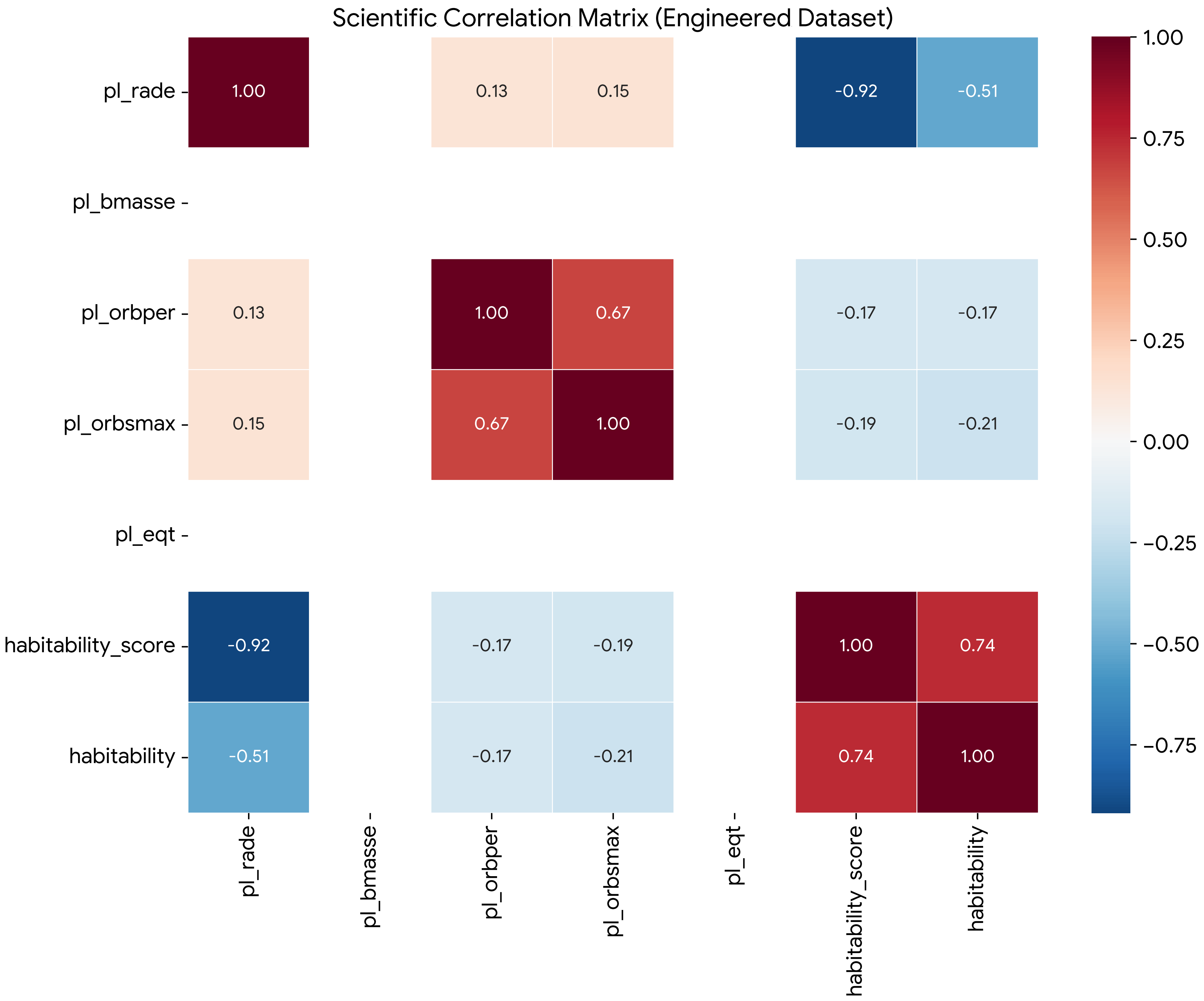
TECHNICAL STACK:

- Data Handling: Pandas, NumPy
- Statistical Validation: SciPy (Z-Score), IQR Analysis
- Scalability: Scikit-Learn (StandardScaler)

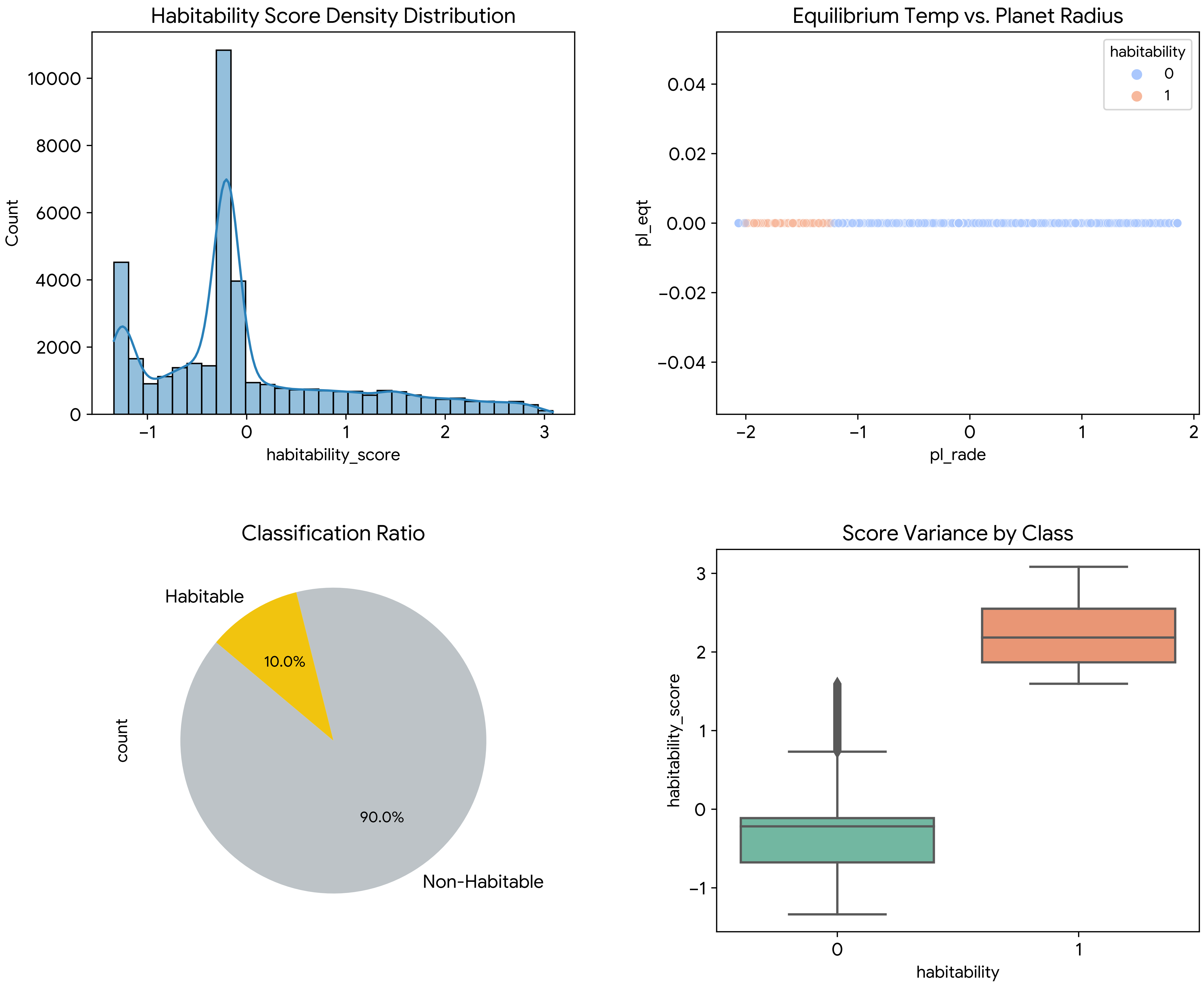
Discovery Insights: Top Candidates

The system identified the following candidates as having the highest potential for habitability based on Earth-like physical and thermal equilibrium parameters.

Planet Name	Host Star	Radius (RE)	Temp (K)	Habit. Score
Kepler-1417 b	Kepler-1417	1.0	695.0	3.0836
Kepler-1417 b	Kepler-1417	1.0	695.0	3.0836
Kepler-1417 b	Kepler-1417	1.0	695.0	3.0836
Kepler-20 f	Kepler-20	1.0	797.0	3.0607
Kepler-106 d	Kepler-106	1.01	797.0	3.0542



Scientific Correlation Matrix (Engineered Dataset)
pl_rade
pl_bmasse
pl_orbper
pl_orbsmax
pl_eqt
habitability_score
habitability
1.00
0.13
0.15
-0.92
-0.51
0.13
1.00
0.67
-0.17
-0.17
0.15
0.67
1.00
-0.19
-0.21
-0.92
-0.17
-0.19
1.00
0.74
-0.51
-0.17
-0.21
0.74
1.00
pl_rade
pl_bmasse
pl_orbper
pl_orbsmax
pl_eqt
habitability_score
habitability
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Habitability Score Density Distribution
Count
10000
8000
6000
4000
2000
0
−1
0
1
2
3
habitability_score
Equilibrium Temp vs. Planet Radius
pl_eqt
0.04
0.02
0.00
−0.02
−0.04
−2
−1
0
1
2
pl_rade
habitability
0
1
Classification Ratio
count
Habitable
10.0%
90.0%
Non-Habitable
Score Variance by Class
habitability_score
3
2
1
0
−1
0
1
habitability

Technical Data Summary

Detailed statistical summary of scaled features. This ensures numerical stability and prevents feature dominance during model training.

Metric	Radius	Temp	Period	Score
count	39315.0	39315.0	39315.0	39315.0
mean	-0.0	0.0	0.0	-0.0
std	1.0	0.0	1.0	1.0
min	-2.065	0.0	-1.021	-1.337
25%	-0.596	0.0	-0.745	-0.578
50%	-0.1	0.0	-0.416	-0.218
75%	0.384	0.0	0.371	0.338
max	1.853	0.0	2.044	3.084