**NASA Exoplanet Archive Data Analysis Report**

**Student Name:** Thi Minh Khue Bui   
**Student ID:** 47509694

**Executive Summary**

This report presents a comprehensive analysis of the NASA Exoplanet Archive dataset containing 9,903 confirmed exoplanet records. The analysis demonstrates proficiency across multiple weeks of course material (Weeks 7-11) including Python Standard Library, NumPy, Matplotlib, and Pandas. Key findings reveal evolving detection capabilities, distinct exoplanet populations, and strong correlations between stellar and planetary properties.

## Data Cleaning

**Data Quality Assessment**

The original dataset required extensive cleaning to ensure reliable analysis. Multiple data quality issues were systematically identified and addressed:

**Primary Issues Identified:**

1. **Missing Values:** Significant gaps in physical parameters, with completeness ranging from 16% (Stellar spectral class) to 100% (Basic identifiers)
2. **Duplicate Records:** Multiple measurement entries for the same exoplanet from different studies and time periods
3. **Invalid Physical Values:** Negative values for quantities that must be positive (masses, radii, temperatures, distances)
4. **Extreme Outliers:** Values likely representing measurement errors or data entry mistakes
5. **Invalid Discovery Years:** Entries with unrealistic discovery dates outside the plausible range (1990-2025)

**Systematic Cleaning Process:**

Using a combination of **Week 10-11 Pandas** operations and **Week 8 NumPy** statistical methods:

* **Duplicate Removal:** Applied *df.drop\_duplicates()* to eliminate exact duplicate entries while preserving legitimate multiple measurements
* **Physical Validation:** Removed rows with negative values for physical quantities using boolean indexing
* **Outlier Detection:** Implemented 3×IQR method using NumPy percentile calculations to identify and remove extreme outliers while preserving legitimate extreme values
* **Temporal Validation:** Filtered discovery years to reasonable bounds (1990-2025) based on exoplanet discovery history

**Strategic Decision: Retention of High-Missing-Value Columns**

A critical decision was made to **retain columns with high missing value percentages** (up to 84% missing for Stellar spectral class) rather than dropping them entirely. This approach was chosen for several important reasons:

1. **Astronomical Data Reality:** Missing values in exoplanet research are often due to measurement difficulty rather than data quality issues. For example, planet temperatures are extremely challenging to measure and require specific observational conditions.
2. **Preserving Rare Information:** Columns like stellar spectral class, despite 84% missing values, provide crucial astrophysical context for the 16% of cases where data exists. Removing these columns would eliminate valuable scientific information.
3. **Method-Specific Limitations:** Different discovery methods naturally produce different types of missing data. Transit photometry easily measures planet radius but struggles with mass, while radial velocity does the opposite. Retaining all columns preserves the full analytical potential.
4. **Statistical Power Consideration:** With 8,669 final records, even columns with 60-80% missing values still provide 1,700-3,500 data points - sufficient for meaningful statistical analysis.

**Alternative Approach Rejected:** Complete case analysis (retaining only rows with all variables complete) would have reduced the dataset to fewer than 500 records, severely limiting statistical power and introducing severe selection bias toward only the most thoroughly studied objects.

**Results:** Retained 85.2% of original data (8,669 of 9,903 records), ensuring high data quality while maintaining statistical power and preserving the full scope of available astronomical information.

**Additional Cleaning Opportunities:**

* Cross-validation with external astronomical databases for measurement verification
* Uncertainty analysis to flag measurements with unusually high error margins
* Consistency validation between related physical parameters using astrophysical relationships
* Standardization of facility names and stellar classification systems

**Column Retention Strategy:** Rather than dropping columns with high missing value percentages, a strategic decision was made to retain all original variables. This preserves the maximum analytical potential while acknowledging that different research questions may require different subsets of the available data. Analysis techniques were adapted to handle missing data appropriately rather than discarding potentially valuable information.

## Numerical Analysis

**Statistical Summary**

The numerical analysis focused on seven key physical parameters using NumPy statistical functions:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Count | Mean | Median | Std Dev | Min | Max |
| Orbital period (days) | 6156 | 16.6 | 8.36 | 22.7 | 0.177 | 156 |
| Planet radius (R\_E) | 4172 | 2.2 | 2.06 | 1.02 | 0.42 | 6.59 |
| Planet mass (M\_E) | 1148 | 47.8 | 10.7 | 84.9 | 0.193 | 423 |
| Planet temperature (K) | 2362 | 854 | 786 | 376 | 104 | 2330 |
| Stellar temperature (K) | 5719 | 5340 | 5530 | 808 | 2890 | 8720 |
| Stellar mass (M\_sol) | 5200 | 0.901 | 0.925 | 0.238 | 0.117 | 1.78 |
| Stellar distance (pc) | 6429 | 459 | 351 | 408 | 1.83 | 2010 |

**Statistical Insights**

**Distribution Characteristics:** Most planetary parameters exhibit right-skewed distributions (mean > median), indicating the presence of a few extremely large values alongside many smaller ones. This is particularly pronounced for planet mass (coefficient of variation = 1.77) and orbital period (CV = 1.37), suggesting these quantities span several orders of magnitude. Stellar parameters show more symmetric distributions, with stellar temperature being slightly left-skewed, indicating our sample is biased toward stars slightly cooler than average.

**Physical Implications:** The wide range in planetary masses (2,190× span from minimum to maximum) reflects the diversity of exoplanet types, from small rocky worlds to massive gas giants. The typical exoplanet in our sample orbits closer to its star than Earth does to the Sun (median orbital period = 8.4 days vs. Earth's 365 days), highlighting the observational bias toward close-in planets that are easier to detect. The stellar sample represents primarily main-sequence stars like our Sun, with masses clustered around 0.9 solar masses.

## Simple Plot Analysis

A graph of a graph showing the evolution of an experiment

AI-generated content may be incorrect.

**Plot Description**

The scatter plot displays the logarithm of planet mass (in Earth masses) versus discovery year for 1,148 exoplanets with complete mass and discovery year data. A clear downward trend line (slope = -0.052 log(M\_E)/year) demonstrates the evolution of detection capabilities over three decades.

**Analysis**

This plot reveals dramatic improvements in exoplanet detection sensitivity over time. The strong negative correlation (r = -0.588) indicates that astronomers can now detect progressively smaller planets. In the early years (1995-2005), only massive planets exceeding 100 Earth masses were detectable, primarily gas giants discovered through radial velocity techniques. Modern surveys routinely find sub-Earth mass planets, representing a ~1000-fold improvement in mass sensitivity. This trend suggests detection capabilities improve by approximately one order of magnitude every 20 years, driven by advances in photometric precision, instrumentation, and analysis techniques.

The script would show that **we're getting better at finding smaller planets over time**. Early discoveries were mostly giant planets (easier to detect), but newer techniques let us find Earth-sized worlds. This creates a downward trend in the typical planet mass vs discovery year.

## Multi-variable Plot part b

**Investigation Rationale**

The multi-variable analysis explores the complex relationships between stellar properties, planetary characteristics, and discovery methods by examining four variables simultaneously: stellar temperature (x-axis), planet temperature (y-axis), planet mass (bubble size), and discovery method (color coding). This investigation tests whether different discovery techniques are sensitive to different types of planetary systems and whether planetary temperatures correlate with stellar properties in ways that might reveal atmospheric or orbital characteristics.

The choice of these variables addresses fundamental questions in exoplanet science: Do planets around hotter stars have higher temperatures? Are certain discovery methods biased toward specific stellar or planetary types? How does the planet mass influence the relationship between stellar and planetary temperatures? Understanding these relationships is crucial for correcting observational biases and interpreting the true distribution of exoplanet properties.

## Multi-variable Plot part c

**A graph showing a graph of a number of red dots

AI-generated content may be incorrect.**

**Analysis and Conclusions**

The multi-variable plot reveals several important patterns in exoplanet characteristics and discovery biases. Most strikingly, planetary temperatures are consistently lower than stellar temperatures (most points fall below the diagonal equality line), which is physically expected since planets receive only a fraction of their star's energy output and lose heat through radiation to space.

The moderate positive correlation between stellar and planetary temperatures (r = 0.472) indicates that hotter stars do tend to host warmer planets, but the relationship is not deterministic, suggesting that orbital distance and atmospheric properties play crucial roles. Different discovery methods show distinct selection biases: Transit photometry efficiently detects planets across all stellar temperatures and finds predominantly small to medium-mass worlds, while Radial Velocity techniques favor planets around cooler stars and tend to detect more massive planets due to the stronger gravitational signals they produce.

## Extension Task

**Extension Task a: Advanced Machine Learning Implementation**

**Methodology and Technical Implementation**

The extension analysis employs K-means clustering, an unsupervised machine learning technique not covered in standard coursework, to identify natural groupings within the exoplanet population. This approach goes beyond descriptive statistics by using algorithmic pattern recognition to discover hidden structures in multidimensional data.

**Technical Implementation:**

* **Feature Selection:** Five key parameters (orbital period, planet radius, planet mass, stellar temperature, stellar mass)
* **Data Preprocessing:** Log transformation of skewed variables followed by standardization (mean=0, std=1)
* **Optimal Cluster Detection:** Elbow method analysis testing k=2 to k=7 clusters
* **Statistical Validation:** ANOVA testing to confirm significant differences between clusters

**Advanced Features Demonstrated**

This analysis incorporates several sophisticated techniques beyond basic coursework:

1. **Scikit-learn Integration:** Utilizes StandardScaler and KMeans from sklearn
2. **Optimal Parameter Selection:** Implements elbow method for data-driven cluster number selection
3. **Statistical Validation:** Applies scipy.stats ANOVA testing for cluster significance
4. **Multidimensional Scaling:** Handles 5-dimensional feature space with appropriate normalization

**Extension Task b: Scientific Value and Insights**

A graph of a graph and a graph of a graph

AI-generated content may be incorrect.

**Scientific Value**

The clustering analysis adds substantial value by revealing that exoplanets naturally segregate into three distinct populations, each with characteristic properties that likely reflect different formation mechanisms or evolutionary pathways. This unsupervised approach objectively identifies patterns that might be missed in traditional parameter-by-parameter analysis.

**Cluster Characteristics:**

* **Cluster 1 (185 planets):** Small, close-in planets around solar-type stars
* **Cluster 2 (131 planets):** Intermediate-mass planets with moderate orbital periods
* **Cluster 3 (64 planets):** Large planets in wider orbits around more massive stars

**Novel Insights**

The machine learning approach reveals that exoplanet diversity is not continuous data but shows clear population structures. All five parameters show statistically significant differences between clusters (p < 0.001), indicating these groupings represent genuine physical distinctions rather than random noise. This population structure suggests that planet formation operates through discrete mechanisms or occurs in distinct stellar environments, providing targets for future theoretical modeling and observational follow-up studies.

**Implications for Exoplanet Science:** These results support theories of multiple planet formation pathways and suggest that observational surveys should be designed to sample all three populations adequately. The distinct clustering also implies that statistical studies of exoplanet occurrence rates should account for population structure rather than treating all planets as a homogeneous sample.