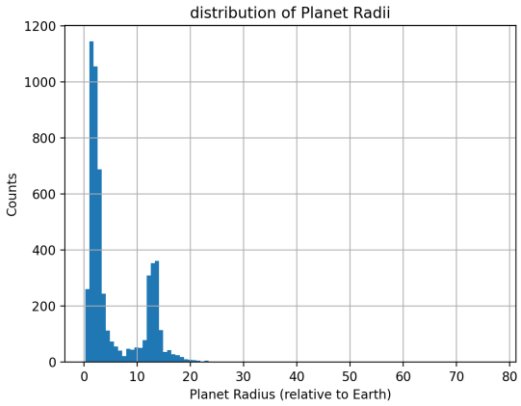
DATASCI200 Section 6 Project 2 – Exploratory Data Analysis

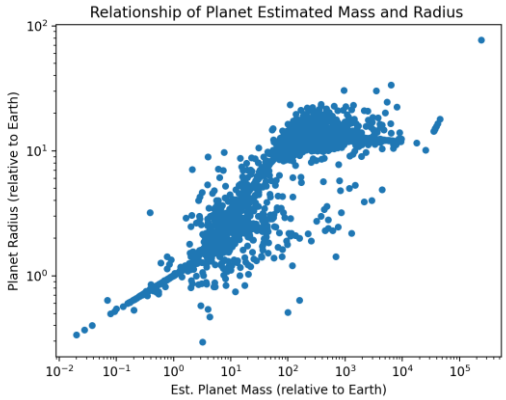
Team Members: Cedar Frost, Sammy Lee, Spencer Morris

GitHub repo: https://github.com/UC-Berkeley-I-School/Project2_Frost_Lee_Morris

Primary Dataset: The dataset of interest is called the Planetary Systems Composite Parameters Planet Data, hosted by the NASA Exoplanet Archive. Each row corresponds to a single confirmed exoplanet (i.e. one which has been verified using a physically robust measurement technique such as radial velocity) , and the full dataset contains 84 columns, each of which corresponds to a physical property. The organization has combined reported properties of a single exoplanet from various studies into a single curated dataset: where discrepancies of a measured property exist, the organization has an established process to choose a most appropriate measurement. The dataset table was downloaded from here on 4 April 2023.

Preliminary Figures:





Key Variables: The full dataset contains 84 variables, but many of these are related to nuanced astronomical data, measurement error, or metadata. We restrict our analysis to the key astrophysical properties related to the exoplanet and its host star, as well as some critical discovery metadata. The following fields were defined as of primary interest:

- Discovery Data
  - **discoverymethod** ← Observational technique by which planet was discovered
  - **disc_year** ← Year the planet was discovered
  - **pl_controv_flag** ← Flag indicated if the planet discovery is controversial
- Planet Data
  - **pl_name** ← Unique identifier for each planet
  - **pl_orbper** ← Orbital period [unit: Earth days]
  - **pl_rade** ← Radius [unit: relative to Earth's radius]
  - **pl_bmasse** ← Mass (minimum; best estimate) [unit: relative to Earth's mass]
  - **pl_dens** ← Density [unit: g/cm^3]
- Star Data
  - **hostname** ← Unique identifier for each star
  - **sy_dist** ← Distance [unit: parsecs]
  - **st_mass** ← Mass [unit: relative to Sun's mass]
  - **st_spectype** ← Spectral Type [unitless; Morgan-Keenann classification]

- **st_lum** ← Luminosity [unit: relative to Sun's luminosity log10(Solar)]
- **st_teff** ← Temperature [unit: Kelvin]
- **st_met** ← Metallicity (proxy for system composition and age) [unit: dex]
- **st_age** ← Age [unit: Gya]

Plan: Our final report will include a brief overview of the nature and transformations of the dataset as well as key observational principles. We have identified a number of preliminary questions of interest, but will limit the scope of the final report based on the most interesting findings. The following questions and sub-questions have been identified for investigation during the exploratory data analysis:

- Dataset Transformation
  - Where are the null values in the dataset?
  - What additional filtering should be imposed on the dataset during analysis?
    - Should only planets with a full set of certain fields be included?
    - Should planets flagged as controversial be excluded?
  - Are there any artifacts or outliers that stand out?
    - What is the nature of the multiple linear regions evident in the planet radius/mass plot?
- Discovery data
  - How have planet discoveries changed over time?
    - Has one observing method become dominant and if so why might that be?
  - Do the different discovery techniques produce similar distributions or have unique biases for…
    - controversial flag?
    - planet mass, radius, period?
    - stellar distance, mass, luminosity, spectral type, metallicity?
- Stellar data
  - What is the relationship between a star's metallicity and its age?
  - Is there a relationship between a star's age and its number of planets?
- Planet data
  - How does the sample of detected planets compare to the densities of Earth and Jupiter?
  - What are the closest Earth analogues that have been detected?
  - Are there any planets within the habitable zone around sun-like stars?
    - Note: this would require a thorough analysis with clearly stated assumptions, and may be beyond our scope