Project Proposal

Tags	
Assign	
Due	@April 7, 2023
Status	In Progress

▼ Introduction

▼ Definition

An exoplanet is any planet beyond our solar system. Most orbit other stars, but free-floating exoplanets, called rogue planets, orbit the galactic center and are untethered to any star.

Most of the exoplanets discovered so far are in a relatively small region of our galaxy, the Milky Way. We know from NASA’s Kepler Space Telescope that there are more planets than stars in the galaxy.

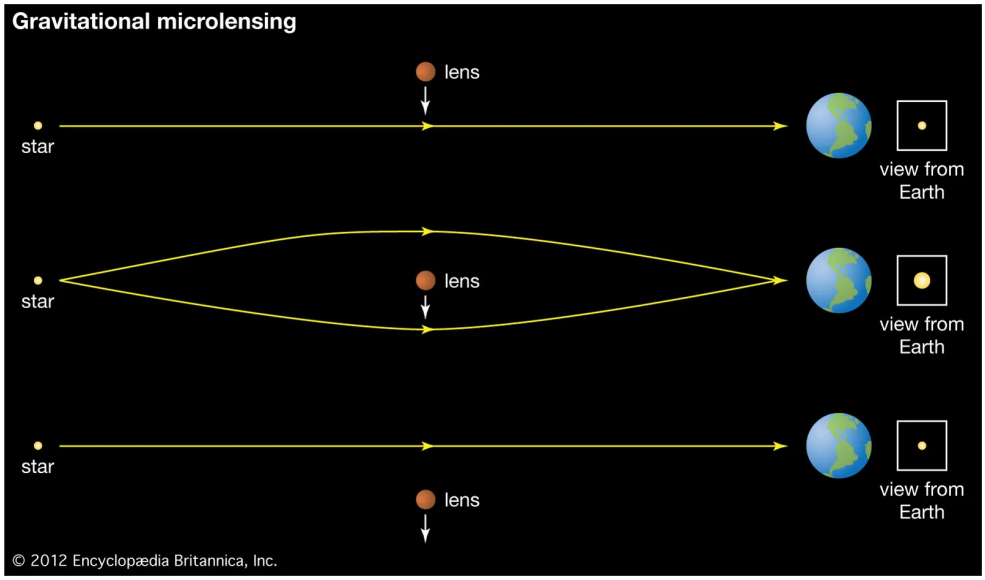
▼ Composition and types

By measuring exoplanets’ sizes (diameters) and masses (weights), we can see compositions ranging from very rocky (like Earth and Venus) to very gas-rich (like Jupiter and Saturn). Exoplanets are made up of elements similar to those of the planets in our solar system, but their mixes of those elements may differ. Some planets may be dominated by water or ice, while others are dominated by iron or carbon.

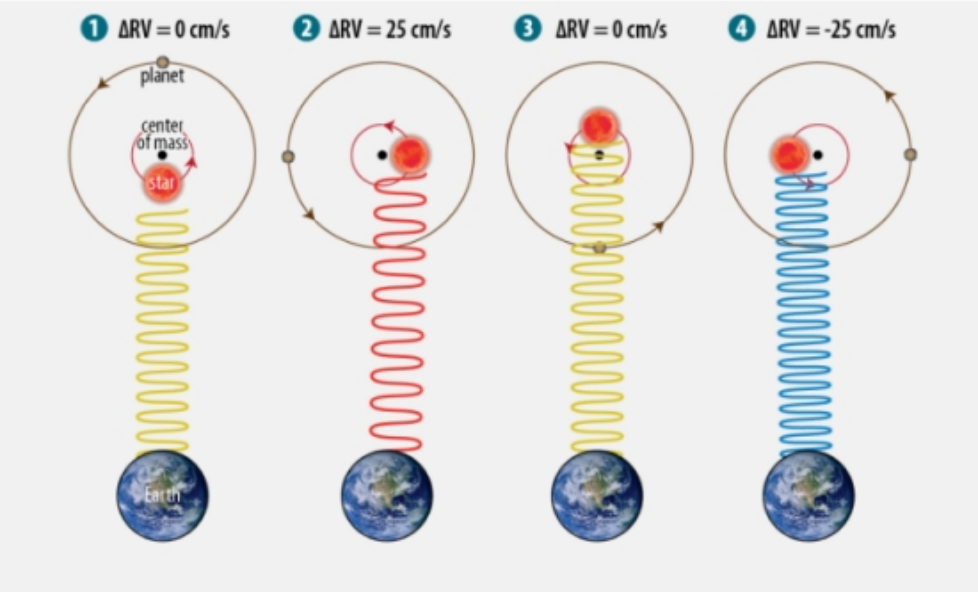
The classes of exoplanets include terrestrial planets made mostly of rock, gas giants with a gas envelope and a small solid core, super Earths with a thick atmosphere and a rocky surface, mini-Neptunes with an icy surface, rocky core, and a thick atmosphere, and hot Jupiters which are gas giants but are very close to host star.

▼ Discovery methods

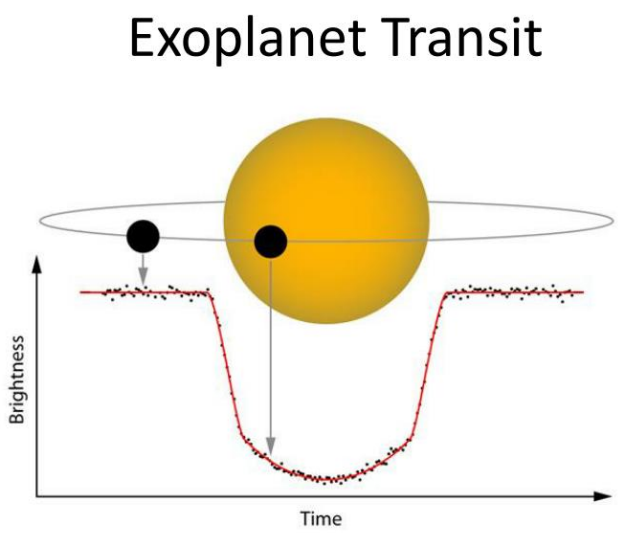
Most exoplanets are found through indirect methods: measuring the dimming of a star that happens to have a planet pass in front of it, called the transit method, or monitoring the spectrum of a star for the tell-tale signs of a planet pulling on its star and causing its light to subtly Doppler shift. Space telescopes have found thousands of planets by observing “transits,” the slight dimming of light from a star when its tiny planet passes between it and our telescopes. Other detection methods include gravitational lensing, the so-called “wobble method.”



Microlensing approach



Radial velocity approach



http://www.astro.caltech.edu/people/faculty/wasp10_transit_600.jpg

Transit approach

*Taken from https://exoplanets.nasa.gov/what-is-an-exoplanet/overview/

▼ Datasets

Exoplanet dataset is retrieved from http://exoplanetarchive.ipac.caltech.edu on Sun Feb 26 04:31:47 2023 West Standard Time:

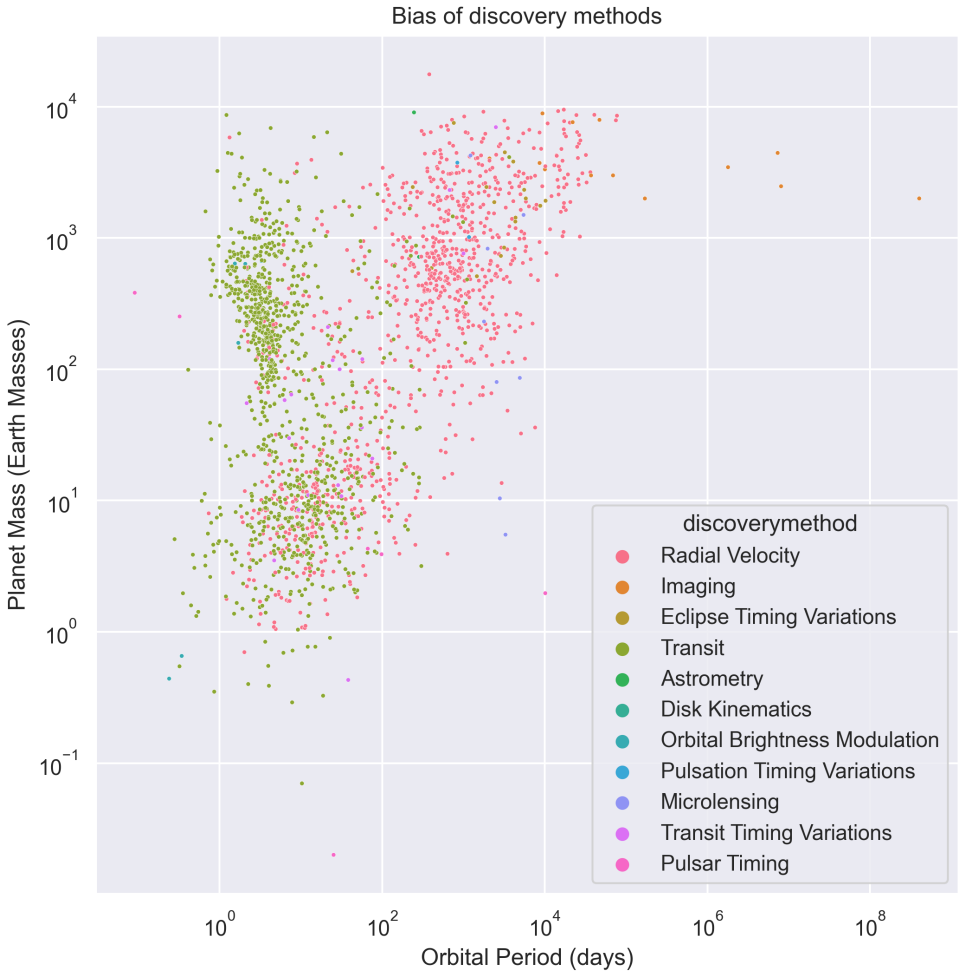
- Confirmed exoplanets: 5272 rows, 92 columns
- Planet emission spectrum: 574 rows, 14 columns
- Planet transit spectrum: 5745 rows, 22 columns

Depending on models used in later analysis, some columns are subject to be dropped or cleaned. But we keep all columns for now in case for new models. The following analysis is based on the confirmed exoplanets dataset.

▼ Recaps from EDA

From our exploratory data analysis. We developed the following insights about the data.

▼ Bias of discovery method



Dominant discovery methods are transit, radial velocity and microlensing. But those methods are not equally biased towards different types of planets. From the scatterplot, we can see transit has bias

- On planets with short periods and small semi-major axes because they transit their host star more frequently.
- On larger planets are more likely to be detected since they block more light.
- On planets whose orbital planes pass earth. Otherwise transit would not be observed.

Radial velocity has bias

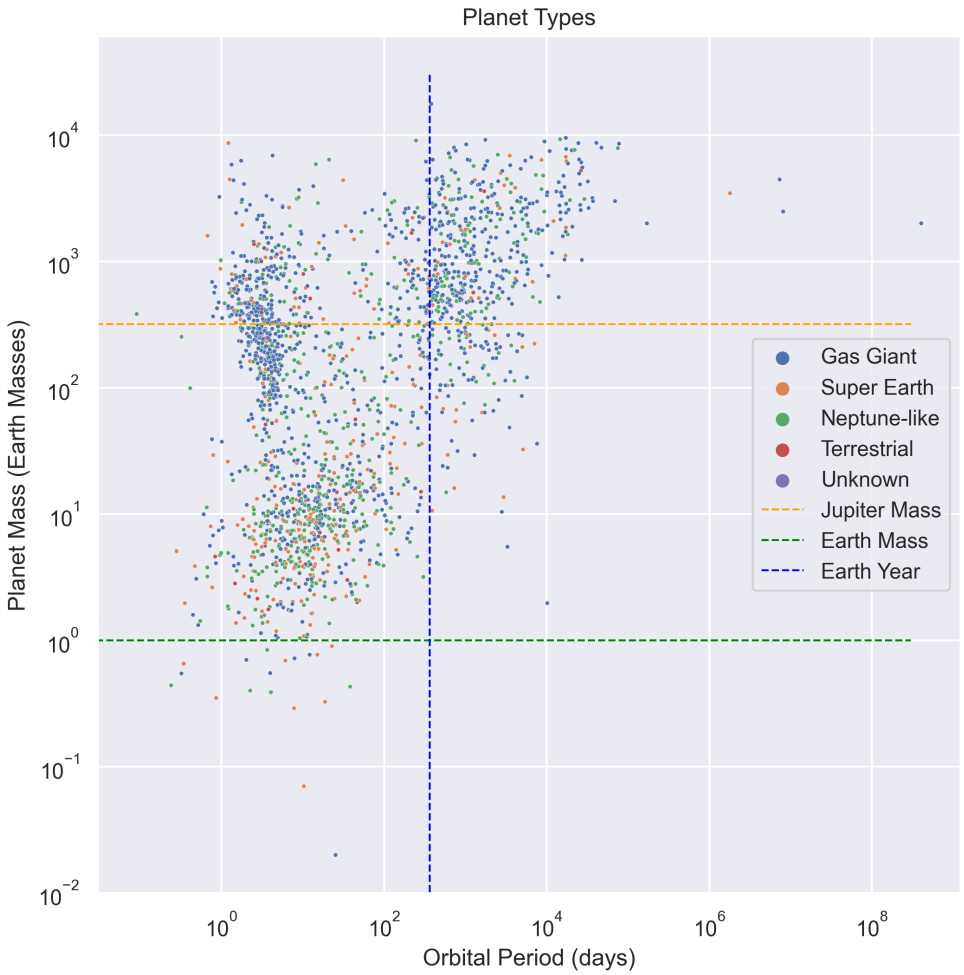
- On massive planets that are close to host stars.
- On planets whose orbital plane close to the earth. The constraint on orbital plane is less strict than transit method.

Micro-lensing has bias

- On massive planets bc massive planets have stronger lensing effect

- On planets distant from host stars (Complementary to transit and radial velocity methods)
- On planets with specific orbital period, orientation and eccentricity because lensing effect is strongest when the planet resides in the middle between background starlight and the earth

▼ Planet types



Unsurprisingly, gas giants tend to be most massive. Super-earth and Neptune-like planets tend to have intermediate mass.

But there are several counter-intuitive findings:

- Hot Jupiters are common.
- Earth-like planets are much rarer than super-earths.
- Most planets have less period than earth-year.

The above findings are contradictory to our solar system. This could be result of the bias of discovery methods or the fact that our solar system is actually a special case.

▼ Problem Statement

Based on the result of EDA. There are a few problems that we think are worth further studying. For each problem we plan to use various models, tune hyper parameters and select the best performing model.

▼ Habitability

Habitability is a major question in exoplanet research. What makes a planet habitable depends on several factors.

First, we need to estimate planets within the "Goldilocks Zone". This zone refers to the range of stellar radiation received from the host star that must be within the correct range to support life. To calculate the stellar flux, we use the star's luminosity and the distance between the star and the exoplanet. The range of stellar fluxes that should result in a "Goldilocks" label is typically taken as the range which allows liquid water.

Second, we consider planet size and mass. We need to determine the surface gravity and the ability to hold onto an atmosphere. Generally, planets with a size similar to Earth (0.8 to 1.5 Earth radii) are considered potentially habitable. Perhaps a potentially habitable planet should have a mass between 0.5 and 5 Earth masses.

Third, we need to estimate atmosphere composition quality using existing features. The composition of a planet's atmosphere is crucial in determining habitability. A habitable planet should have an atmosphere rich in water vapor, nitrogen, and trace amounts of other gases like carbon dioxide and methane. Atmospheric composition information are deducible from planet transit and emission spectrum.

▼ Planet types

Over 5000 Exoplanets have been discovered till now based on various methods such as radial velocity, microlensing, imaging etc., Moreover, the discovered planets have certain characteristics that make them an ‘exoplanet’ such as their mass, radius, orbital period among others. However, not all these planets can be deemed to be the same type. While some are super-earths, others are Gas Giants, Neptune-like and Terrestrial in nature. The goal of this problem is to categorize the exoplanets into different groups using clustering methods and to compare and evaluate these groups against the known planet types.

▼ Likely discovery methods

Discovery methods have strong bias. It has been shown in EDA . The following table summarizes bias of each discovery method on each predictor

	Mass	Radius	Orbit radius	Orbit orientation
Transit	Large	Large	Very short	Strictly coplanar

	Mass	Radius	Orbit radius	Orbit orientation
Radial velocity	Large	Large	Short	Coplanar
Micro-lensing	Large	Large	Long	No bias
Direct imaging	Large	Large	Short	No bias

Build a bayesian model that takes planet characteristics as input and predict the most likely discovery method.

Predictors include

- Planet mass
- Orbit radius
- Stellar mass

▼ Planet and Host star relation

Based on the result of EDA . We plan to study the relation between host star and planets, build a regression models that predict the distribution of planetary characteristics based on stellar characteristics.

Planetary characteristics:

- Orbit radius
- Mass
- Radius (a reflection of density)
- Type

Useful stellar characteristics

- Stellar mass
- Spectral type (a reflection of stellar surface temperature)
- Metallicity (a reflection of stellar age and generation)

Stellar characteristics are not available in the original dataset. We need to either download the information from another dataset or web-scrape the data and make a joint table with planetary data.

▼ Methods

Candidate models

- k-NN

- Logistic regression
- LDA, QDA
- Decision tree random forest
- Neural network

For the questions in Habitability , we will be using regression methods to estimate quantities such as planet mass, atmosphere composition and stellar flux. Moreover, we will also be using classification techniques to determine whether a given exoplanet is in the habitable zone or not. Feature selection will be performed through feature importance metrics from tree-based models and/or stepwise logistic regression will be used to determine the useful features in classifying ‘habitability’.

For the questions in Planet types , different clustering methods such as centroid-based (K-means), Hierarchical, density-based (DBSCAN) will be evaluated using the silhouette coefficient. Also, the results will be compared against the known planet types (Gas-Giant etc.). There are approximately 50 variables in the dataset where not all are useful. Moreover, clustering algorithms may suffer from the curse of dimensionality if all these variables are chosen. To circumvent this problem, we would be exploring dimensionality reduction techniques such as PCA to reduce the number of dimensions in the dataset.

For questions in Planet and Host star relation , for each type of planet, we plan to further divide them into subsets based on unsupervised clustering method. For example, Gas giants can be divided into Jupiter-like and Hot Jupiter. In each subset, we study the the mean of stellar properties such as stellar mass, age, and number of host stars. Then develop a Bayesian model to predict planet types and characteristics based on stellar attributes.