Seth Beckett & Carson Stoker

CS 6830

Final Project Proposal

Description

We decided to look at something that may not be immediately important to humanity, but could eventually determine whether or not we as a species face extinction one day. Our project will include analyzing a large repository of data that NASA has on exoplanets to determine which are potentially habitable, determine which are closest to us, and then build a model that can classify future exoplanets as habitable or uninhabitable based on relevant data.

Motivation

Climate change is everywhere. It seems that every year the weather is getting more and more extreme, and soon we may hit a point where climate change damage becomes irreversible. It's not completely far-fetched to say that at some point within the next 100 years, many parts of the Earth that are currently habitable will become uninhabitable. What's the solution? To ruin a different planet of course!

While ruining a different planet was mostly a joke, we are unlikely to survive long-term as a species if we can't adapt and mitigate climate change issues. Our project is based on the assumption that we will be able to handle those issues, or have the technology before those issues destroy us to become an interplanetary species. Finding other habitable planets (or making other planets habitable) is vital for humans to be able to continue to evolve as a species and continue pushing the boundaries of what is possible. Another reason for identifying habitable planets is to search for the existence of alien life.

Schedule

Our schedule isn't set in stone, but we tentatively hope to split our remaining time into four main sections: finishing data exploration and descriptive statistics, creating a classification model for predicting habitability, making our final presentation, and writing our final reports. Since our in-class presentation could be as soon as April 19th, we

hope to finish up our EDA and stats section by the end of Friday. By the end of Monday we will need to have our classification portion done, leaving Tuesday to create and finalize our presentation for Wednesday the 19th. We each will then need to finish our individual reports by the 29th.

Division of Responsibilities

We will split all the responsibilities equally.

Potential Problems & Alternative Approaches

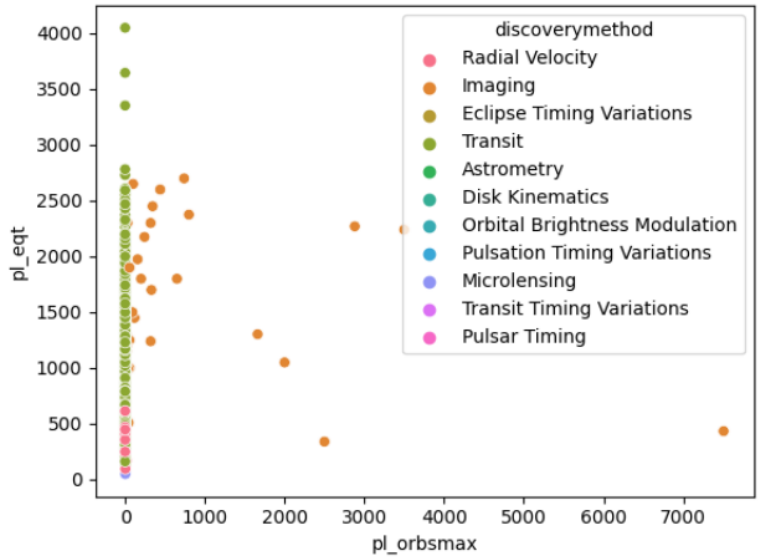
Many of the measurements are given with uncertainty intervals, and so we might not be accurately classifying a planet as habitable or uninhabitable if only using the center value. It also doesn't clarify the confidence level for those intervals, which makes it more difficult to interpret.

Another potential problem is that there are multiple different catalogs for exoplanets, with different amounts of "confirmed" exoplanets. Depending on the database and parameters used for determining an exoplanet, one could get varying results for prediction. Additionally, there are many different techniques that are used to discover new exoplanets in distance star systems. These techniques differ in what data they can produce and also contain biases towards discovery of planets with certain properties. Simply sifting through all of this data will take a good deal of time and hopefully provide valuable insights. One other challenge could be the imbalances in the data, with many more inhabitable planets than potentially habitable ones.

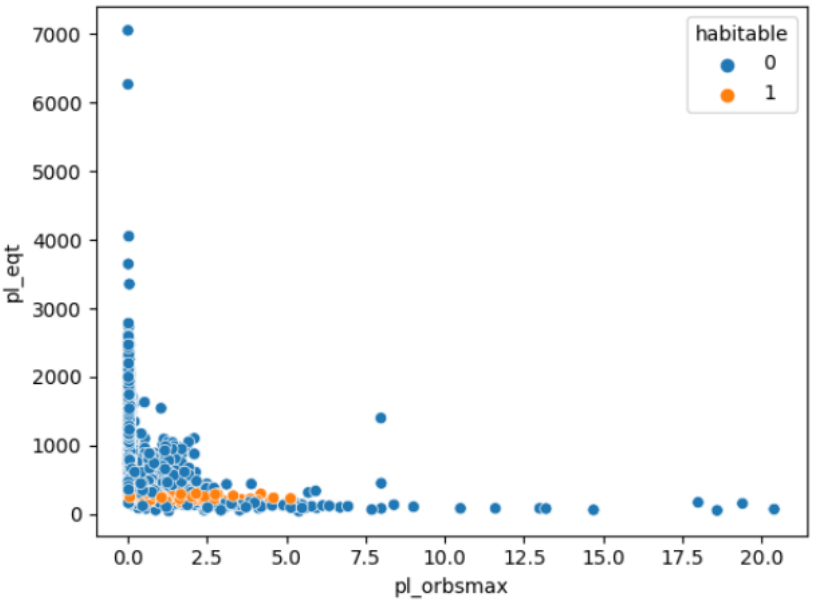
Preliminary Results

Thus far, we have done some preliminary exploration of the data. Some of the most relevant attributes of the data we want to focus on include mass, temperature, and radius of the planets and stars in a system, orbital parameters such as period and max orbital distance, and gravity. Below is an image of 5000+ exoplanets with the equilibrium temperature of the planet on the y axis and the semi major axis (max distance) of the orbit on the x axis. The data points are colored by the technique used in their discovery. As we can see in the graph and the two tables below. Some of the properties vary

greatly by discovery method. This is due to the bias mentioned early. Understanding the techniques and why they produce these biases will be important and provide an opportunity to look more into the data. Additionally, we can see the same graph, but colored by habitability below that.



discoverymethod	pl_orbsmax
Astrometry	0.499825
Disk Kinematics	130.000000
Eclipse Timing Variations	4.271083
Imaging	498.703870
Microlensing	2.484399
Orbital Brightness Modulation	0.013667
Pulsar Timing	4.897800
Pulsation Timing Variations	1.700000
Radial Velocity	2.118700
Transit	0.118125
Transit Timing Variations	0.820778



discoverymethod	pl_eqt
Disk Kinematics	57.948803
Eclipse Timing Variations	434.794838
Imaging	718.730648
Orbital Brightness Modulation	5152.640777
Pulsation Timing Variations	474.713915
Radial Velocity	481.665549
Transit	928.180437
Transit Timing Variations	643.197994