# Final Project Report: Determining the Habitability of Exoplanets

## 1. Introduction:

For my final project, I attempted to determine the the habitability of exoplanets from a known dataset from the NASA Exoplanet Archive (The dataset can be accessed at: https://exoplanetarchive.ipac.caltech.edu/). This dataset contained information about the properties of thousands of exoplanets, including their orbital period, orbital radius, surface temperature, etc.

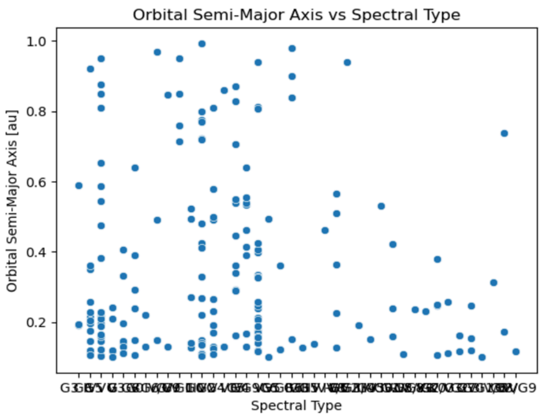
## 2. Method:

### (i) Filtering:

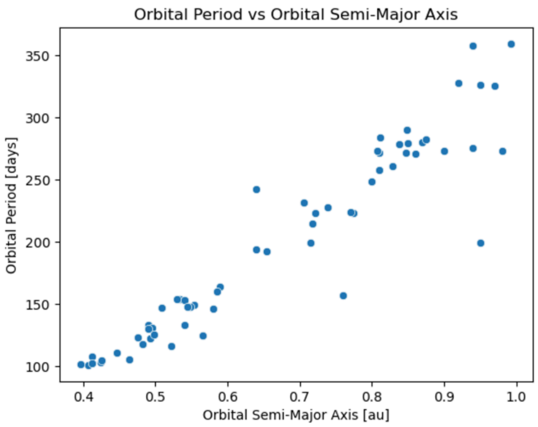
I decided to filter the dataset by comparing the properties of the exoplanets and their stars with that of the Earth and Sun. Initially, I was going to set 6 or 7 filters, but I soon found out that the dataset was rapidly thinning out after about 2 to 3 filters, so I had to make do with them. There was also the problem of the NaN values in some of the columns I was concerned with, so I decided to use the dropna() function so that I would not encounter any problems later on, but I realized that it drastically reduced the amount of data I could work with. So instead, I calculated the average of the numerical columns and used the fillna() function to replace the NaN values with the averages.

### (ii) Plotting:

After applying each filter, I plotted a graph to see how many data points I had left to work with and also to understand what was happening to the dataset after each filter. One of the graphs is shown below:

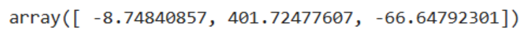


# 3. Results:



After plotting the orbital period vs orbital semi-major axis graph, it seemed like the equation for the line of best fit would be that of a linear curve, but after multiple attempts to find the curve, I realized that the best fit line was in fact a quadratic function. The comparatively small-scaled x axis and the large-scaled y axis made it seem this way.

I generated random points using the mean and the standard deviation and created a curve-fitting function that accepted parameters for a quadratic function. Below was the list of parameters that were returned after entering in some initial guesses.

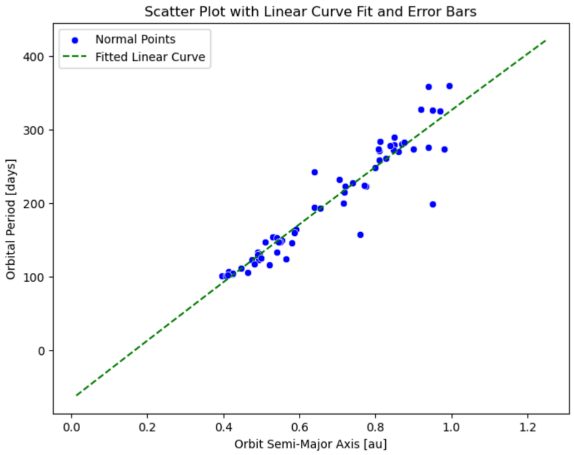


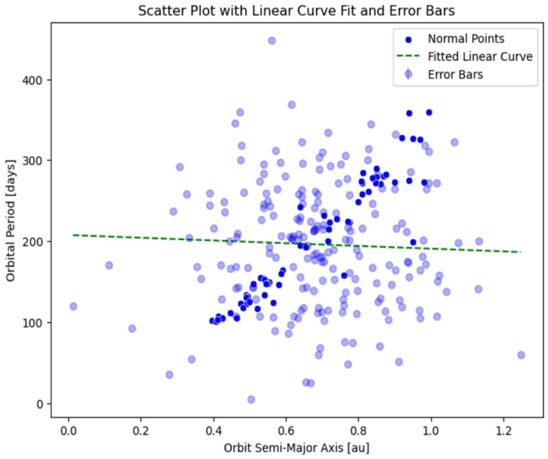
```
array([ -8.74840857, 401.72477607, -66.64792301])
```

The equation of fit was determined to be:

$$491.55334327748x^2 - 264.64505276399x + 137.60685480922$$

And the quadratic curve and the error bars are shown in the graphs below:





# 4. Discussion:

Unfortunately, the error bars for the random data were not plotting correctly and I was unable to plot the outliers because I ran into a lot of issues. I used the method of Interquartile Range to detect outliers, but for some reason, whenever I tried plotting the outliers and residuals, the program gave an error or showed nothing at all.

# 5. Conclusion:

Overall, I am satisfied with the best fit line and the equation of fit that I managed to obtain, but in the future, I might use a better and more reliable method to detect outliers. In hindsight, I should have thought of different ways to filter that dataset in the first place, since I ran out of data points after just 2 to 3 filters, but I should keep in mind that most datasets are quite small to begin with, so determining the habitability of exoplanets with an insufficient amount of data is quite difficult.