Mass-Radius Relationship of Rocky Exoplanets

Introduction

For this project I have modeled the relationship between mass and radius of rocky exoplanets. In this report, I will discuss the process I took to find a model that shows the correlation well.

Libraries and Data

To find a model of the relationship, I first needed to import a few Python libraries. I have imported pandas, matplotlib.pyplot, numpy, and scipy.optimize in order to work with a dataframe and curve-fitting.

For the data, I specifically need the mass and radius of exoplanets, so I have retrieved data from [NASA Exoplanet Archive].

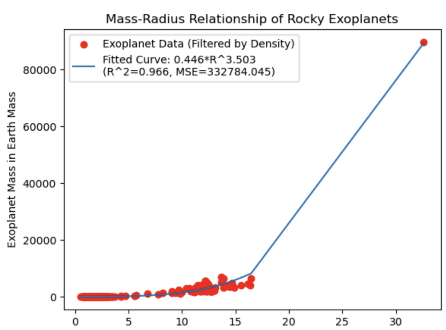
Filtering Data and Curve-Fitting

Step 1: (Filtering by Density)

In order to determine which planets are rocky exoplanets, I have calculated the density using the formula below. Then I have only kept the data of exoplanets with density greater than 5 since rock exoplanets are denser than other types of exoplanets such as icy and gas.

Density= $\frac{M}{(4/3)\pi R^2}$, $M = Mass\ and\ R = Radius$

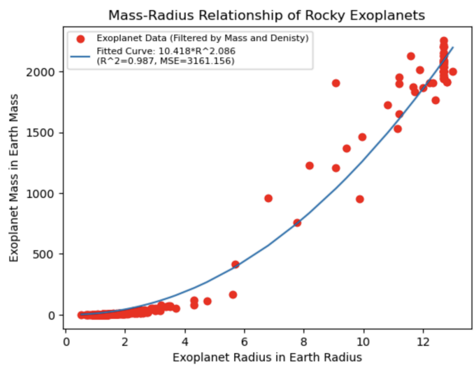
After filtering, I have plotted the data and tried to fit it to a power model in which radius is the x-value and mass is the y-value. However, creating a power model fit, it was apparent that the R^2 value was low. Thus, I have calculated the Interquartile Range (IQR) in order to find the outlier and filter them out. The result of the curve-fit is the graph below (with the R^2 value and MSE value to consider error).



The fitted curve has a r^2 value near 1, which indicates that the model is a good fit. However, the MSE value was extremely high, which can make sense due to the high value. In addition, there is one data point that has an extremely high mass that skews off from the concentrated data on the bottom left of the graph. Thus, I thought zooming on the concentrated data may help improve the model.

Step 2: (Filtering by Density and Mass)

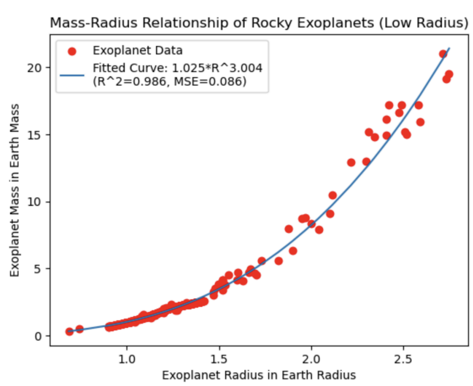
In order to zoom in on the concentrated data, I filter the data even more by removing the outliers in terms of mass as well; I have used IQR to remove outliers. The result is the graph below.

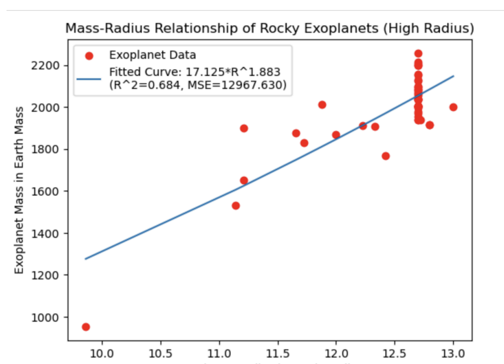


It is apparent that the r^2 value improved slightly since it is closer to 1. In addition, the MSE value has decreased significantly, which implies that this model is significantly better. However, I noticed that the data was concentrated in mainly two separate areas. From the graph, it is apparent that the data was concentrated when the radius was low and high; radius less than 3 (initially thought it was 4 but after zooming in it was 3) and radius greater than 10. Thus, I have zoomed in on the two concentrated areas and then removed the outliers by density again. I have removed outliers again for the two areas since when I first removed outliers, the data consisted of both low and high radius, which outliers specifically within low or high radius would not be removed; in short, outliers are relative.

Step 3: Divide by Low and High Radius

The graphs below are the fitted curve for low and high radius.





It is apparent that the fitted curve for the low radius is a good model since the r^2 value is close and the MSE value is near 0. However, the model for the high radius is obviously terrible due to the r^2 and MSE value. This difference between low and high radius could be due to the lack of data for high radius rocky exoplanets compared to low radius; this lack of data for high radius would explain the MSE value in step 2. Thus, an accurate model of the mass and radius relationship of rocky exoplanets would be the low radius model.

Result

Through all of this, an accurate model of the relationship between mass and radius can be generated for only rocky exoplanets that have a relatively low radius (less than 3 times earth's radius). Thus, the model is M= $1.025R^{3.004}$ in which M is for mass and R for radius; all in terms of earth radius. Furthermore, I have generated a random list of radius to plot the model.