Project: Exoplanets dataset analysis

Ugo LABBE

December 3, 2023

1 Introduction

This project on exoplanets and their different characteristics will include a short description of the dataset, including simple statistics as well as graphics to help visualize the data.

I will then include different analysis that I found to be useful on the dataset, to understand better how certain things work.

2 Dataset description and visualization

2.1 Descriptive statistics

The Exoplanets dataset is made of 1013 different planets, with 7 variables: The planet variables are ratios over earth (for example, the variable planet mass is the planet mass over the earth mass) The orbital variables are the relation between the planet and their star And the star variables are ratios over the sun (for example, the variable star mass is the star mass over the sun mass). Here is a descriptive table with some statistics:

	planet_mass	planet_radius	orbital_period	orbital_radius	star_mass	star_radius	star_luminosity
count	1013.000000	1013.000000	1013.000000	1013.000000	1013.000000	1013.000000	1013.000000
mean	343.563980	9.033790	18.851294	0.101647	0.992595	1.125280	3.206746
std	811.766416	5.612988	65.343955	0.162509	0.265571	0.497601	1.113675
min	0.889840	0.884800	0.719573	0.015260	0.330000	0.328000	1.020237
25%	10.900540	2.817920	3.118601	0.041010	0.830000	0.794000	2.434663
50%	133.476000	10.673600	4.542169	0.054900	1.000000	1.038000	3.145093
75%	349.580000	13.596800	11.024540	0.096000	1.180000	1.390000	3.902029
max	11440.800000	20.888000	1071.232280	1.890000	1.720000	4.230000	6.365151

Figure 1: Descriptive statitics on the exoplanets dataset.

2.2 Data visualization

I decided to plot the distribution of both the mass of the planets and the mass of the stars.

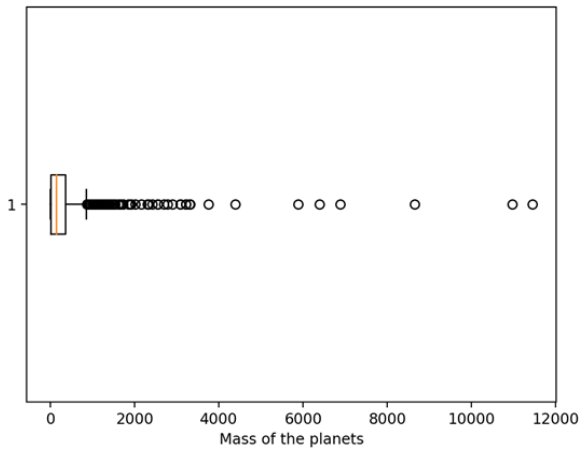


Figure 2: Boxplot of the distribution of planets mass.

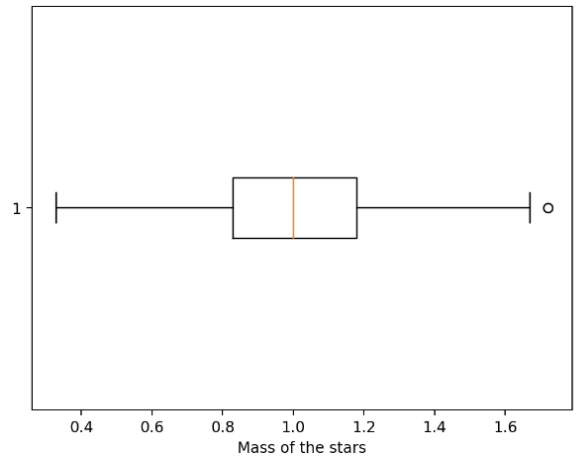
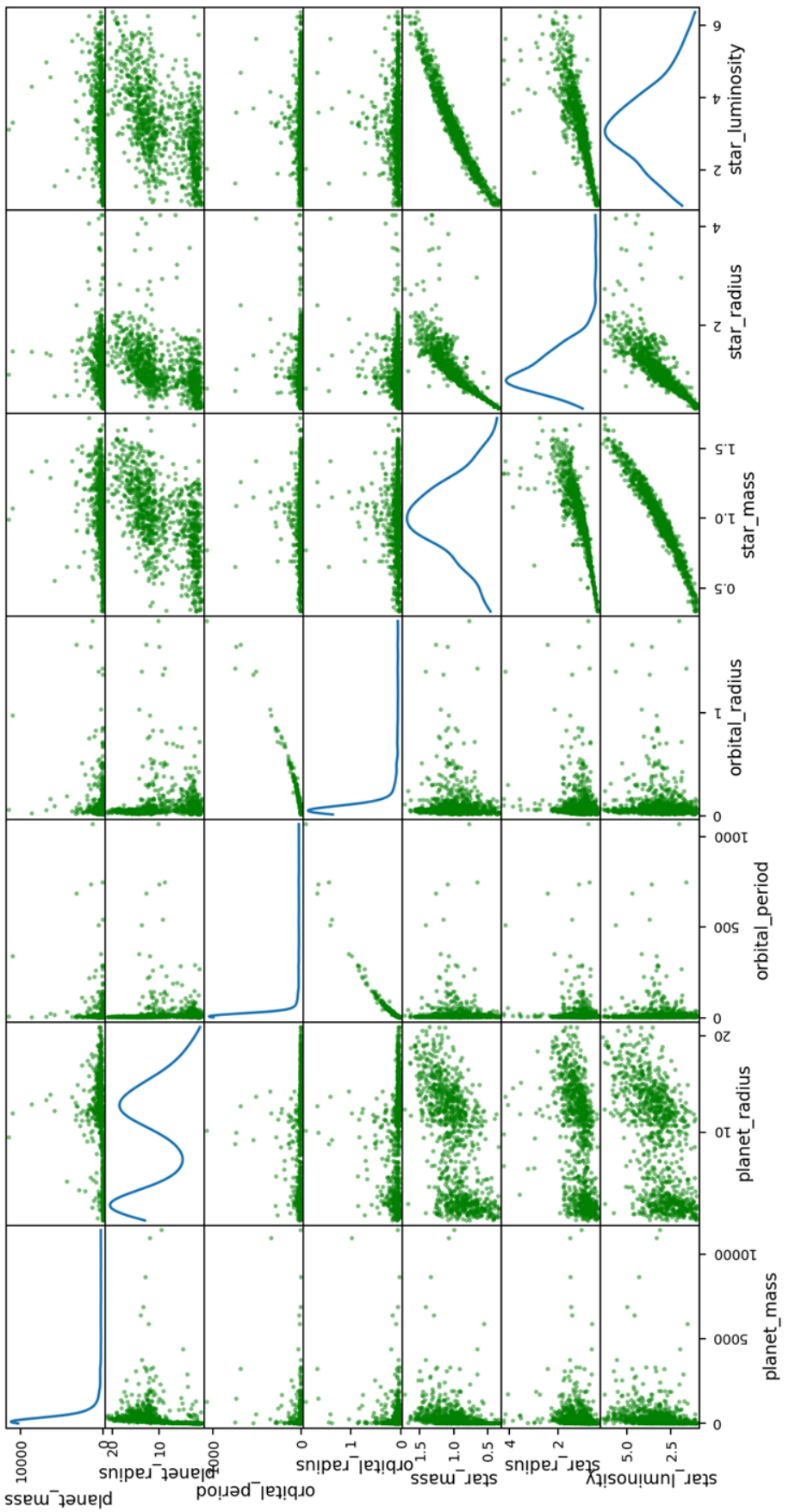


Figure 3: Boxplot of the distribution of stars mass.

If we want to visualize more data we can use a pair plot, which will highlight the different correlations between each variables, as well as their distribution:


planet_mass
planet_radius
orbital_period
orbital_radius
star_mass
star_radius
star_luminosity

We can see that some variables are highly correlated, it is something useful that we will use for later experiments such as linear regression. We can also observe that the star mass and the star luminosity seems to follow quite well a normal distribution.

3 Linear Regression

In a first time, we will try to predict the variable Star Luminosity with the variable Star mass. We the predicted value, we can compare the actual data to our linear regression. We can also plot a 95% confidence interval to the graph.

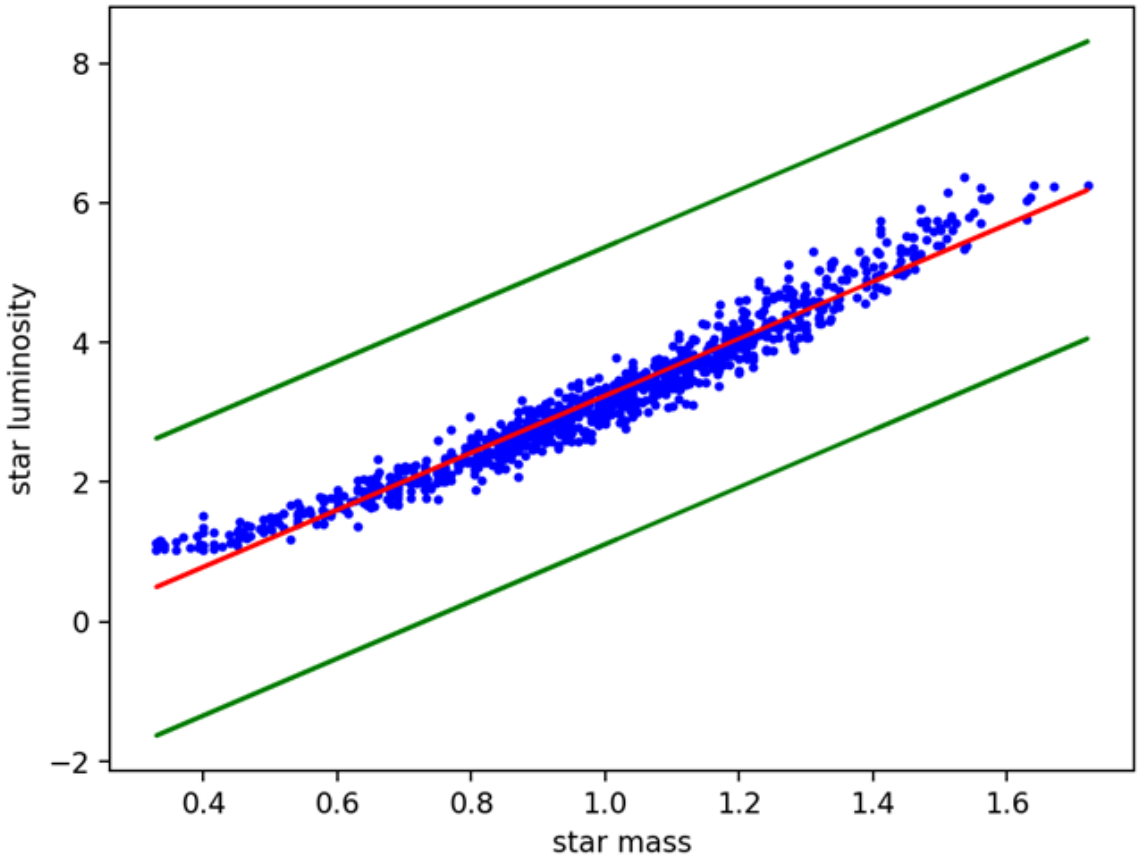


Figure 4: Linear model with predicted variable Star Luminosity with the variable Star mass.

We see here that the model predicts really well a line where most of the data point falls into. Now if we use the star radius instead of the star we obtain a less convincing result and see that there quite a lot of planets falling outside the 95% confidence interval.

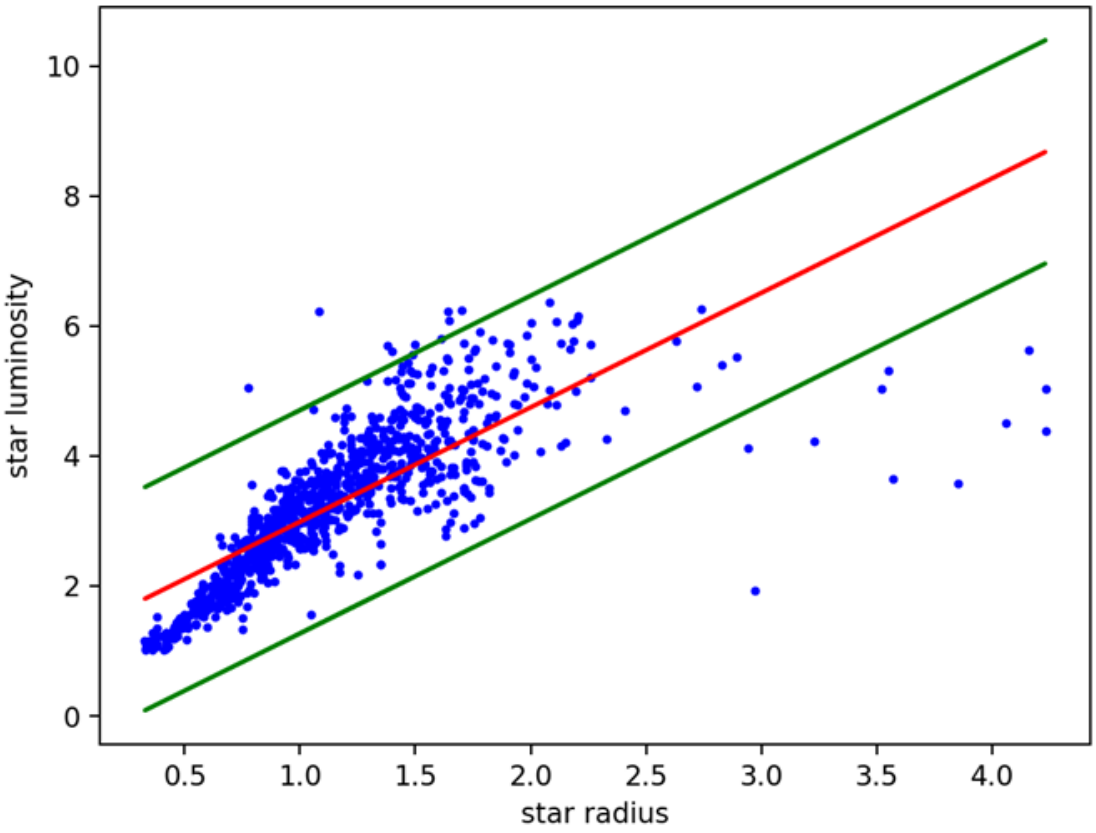


Figure 5: Linear model with predicted variable Star Luminosity with the variable Star radius.

It would then be interesting to create a multiple linear regression model with these two variables, as well as using all variables and compare a metric, such as the R^2.

- First model (with mass of the star) : 0.9513127335789652
- Second model (with radius of the star) : 0.6196728849965294
- Third model (with both variables) : 0.9518959702765545
- Fourth model (with all variables) : 0.9519796068415349

The closer to 1 the R^2 value is, the better the model is to predict accurately the star luminosity (in our example). We see that the second model is by far the worst model while the first, third and last model are doing a good job (really close to 1). The last model as the highest R^2 so it's better to use this one to predict the star luminosity.

4 Habitable zone distance

We know that for a planet to be considered as habitable, its orbital radius needs to fall in between this range:

$$Orbital_radius \in [0.8 * \sqrt{starluminosity}; 2 * \sqrt{starluminosity}]$$

We checked with our dataset and actually found 3 planets that satisfy it. Planet of index 60, 427 and 431 (the red dots on the following graph):

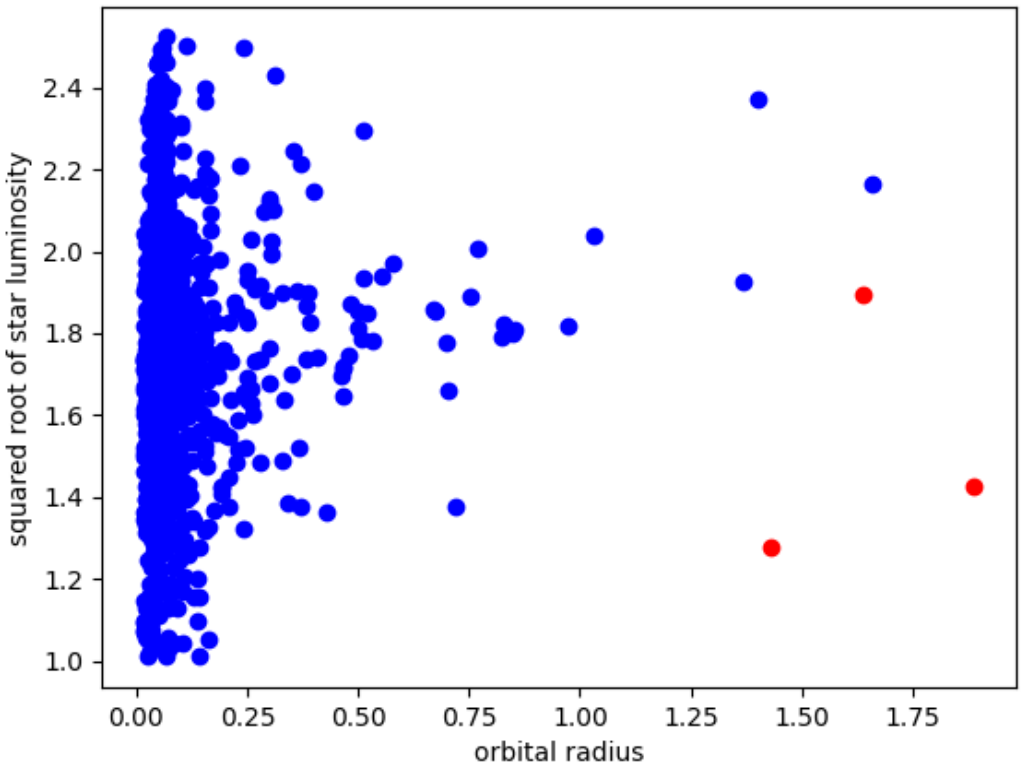


Figure 6: Scatter plot of the orbital radius with the star luminosity.

It would be then interesting to look into them as they seem quite far for any other points. We might find some other characteristics. For example, their placement in the scatter plot with the mass of the planet and the planet radius:

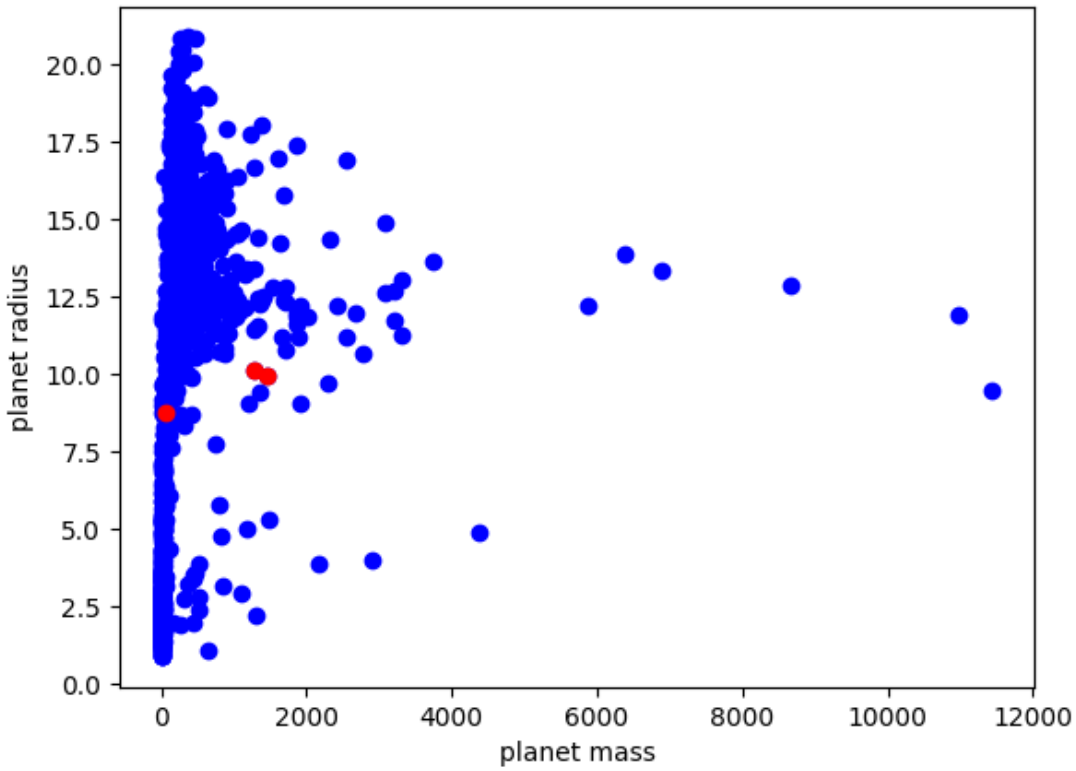


Figure 7: Scatter plot of the planet mass with the planet radius.

We see that they don't really stand out, however they are fairly close in term of planet radius (between 8 and around 10) In a second time, we could observe their position on the graph of orbital period with orbital radius.

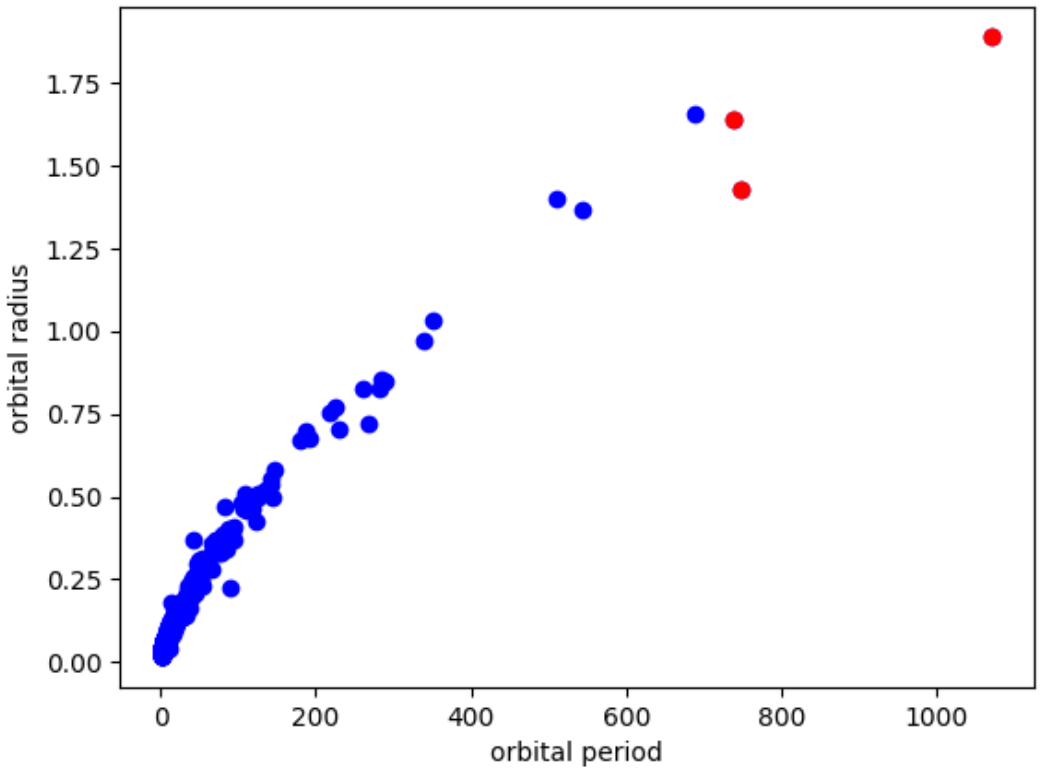


Figure 8: Scatter plot of the orbital period with the orbital radius.

Here we see that they really stand out further from any other planets. We could understand that the orbital radius, the orbital period and the planet radius play a certain role in the habitability of a certain planet, as we observe closeness between the 3 planets discovered.