A Data Analysis Project on Exoplanets

Zahir AHMAD
Master of Machine Learning and Data Mining
Jean Monnet University
December 6, 2023

Introduction

In this project, I want to explore a dataset of exoplanets, which are planets outside our solar system. I want to find out some interesting facts about these planets and their stars, using statistics and data visualization. The dataset has information about 1013 exoplanets and their properties, such as mass, radius, orbital period, orbital radius, and the mass, radius, and luminosity of their host stars.

I have some research questions that I want to answer with this dataset:

- How many planets are in the habitable zone, which is the range of orbital distances where liquid water can exist on the surface of a planet, if it has enough atmospheric pressure?
- Is there a relationship between the planet radius and the star luminosity?
- Is there a relationship between the planet mass and the star mass?
- Is there a relationship between the orbital period and the star radius?

To answer these questions, I will use the following techniques:

- Data exploration and visualization, using pandas, matplotlib, and seaborn libraries in Python.
- Habitable zone analysis, using a custom function to check if a planet is in the habitable zone based on its orbital radius and star luminosity.
- Regression analysis, using linear regression models to test the hypotheses about the relationships between the variables of interest, and to evaluate the models using the root mean squared error (RMSE) and the R-squared (R^2) metrics.

Data Exploration and Visualization

The dataset has 9 variables, which are:

- `planet_name`: The name of the planet, which is a string.
- `planet_mass`: The mass of the planet relative to the Earth, which is a positive float.

- `planet_radius`: The radius of the planet relative to the Earth, which is a positive float.
- `orbital_period`: The orbital period of the planet in days, which is a positive float.
- `orbital_radius`: The orbital radius of the planet relative to the Earth, which is a positive float.
- `star_name`: The name of the star that the planet orbits, which is a string.
- `star_mass`: The mass of the star relative to the Sun, which is a positive float.
- `star_radius`: The radius of the star relative to the Sun, which is a positive float.
- `star_luminosity`: The luminosity of the star relative to the Sun, which is a positive float.

The dataset has no missing values, and the data types are consistent with the nature of the variables. The dataset has a size of 1013 rows and 9 columns.

I used some summary statistics and plots to get a better understanding of the data. Here are some of the things I noticed:

- The quantitative variables have a wide range of values, with some outliers and skewed distributions. For example, the planet mass and radius have some very large values, while the orbital radius and the star luminosity have some very small values. The histograms show that most of the variables are right-skewed, meaning that they have a long tail to the right.
- The categorical variables have a high number of unique values, especially the planet name, which is unique for each row. The star name has some repeated values, indicating that some stars have more than one planet orbiting them. The bar plot shows that the most common star name is Kepler-20, which has 6 planets in the dataset.
- The scatter plots show some patterns and trends between the pairs of variables. For example, there seems to be a positive relationship between the planet radius and the star luminosity, and between the planet mass and the star mass. There also seems to be a negative relationship between the orbital period and the star radius. However, there is also a lot of variability and noise in the data, which may affect the strength and significance of the relationships.
- The heatmap shows the correlation coefficients between the quantitative variables, which measure the linear association between them. The correlation coefficients range from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship. The heatmap shows that the variables have mostly low to moderate correlations, with some exceptions. For example, the planet mass and radius have a high positive correlation of 0.87, indicating a strong linear

relationship. The orbital period and radius have a high positive correlation of 0.76, indicating a strong linear relationship. The orbital period and the star radius have a moderate negative correlation of -0.46, indicating a weak linear relationship.

Habitable Zone Analysis

One of the criteria that is used to determine the habitability of a planet is the habitable zone distance, which is the range of orbital distances around a star where liquid water can exist on the surface of a planet, if it has enough atmospheric pressure. According to the additional information, we can consider a planet to be in the habitable zone distance if the orbital radius Rorbital satisfies:

I used a custom function to check if a planet is in the habitable zone based on its orbital radius and star luminosity. Here are the results:

- There are only two habitable planets in the dataset, which is a very small proportion of the total number of planets. These planets have relatively low mass and radius compared to the Earth, and orbit around stars that have low mass, radius, and luminosity compared to the Sun.

Regression Analysis

Another question that I can ask is whether there are any relationships between the variables of interest in the dataset, such as the planet radius and the star luminosity, the planet mass and the star mass, and the orbital period and the star radius. To answer this question, I can use regression analysis, which is a technique that models the relationship between a response variable and one or more predictor variables. In this project, I will use linear regression models, which assume that the relationship between the response and the predictor variables is linear, that is, of the form:

I will test the following hypotheses using linear regression models:

- H1: There is a positive linear relationship between the planet radius and the star luminosity.
- H2: There is a positive linear relationship between the planet mass and the star mass.
- H3: There is a negative linear relationship between the orbital period and the star radius.

The following are the results of the linear regression models for each pair of variables:

- For the pair of variables `planet_radius` and `star_luminosity`, the results show that the linear regression model has a moderate fit, with a RMSE of 4.767 and a R^2 of 0.274. The p-values of the intercept and the coefficient are both very small, indicating that they are statistically significant. The confidence intervals of the intercept and the coefficient are [-0.244, -0.084] and [2.583, 3.215], respectively, indicating that they are relatively narrow and do not include zero. This means that there is a positive linear relationship between the planet radius and the star luminosity, and that the relationship is unlikely to be due to chance. The coefficient of 2.899 means that for every unit increase in the star luminosity, the planet radius increases by 2.899 units on average. This supports the first hypothesis that there is a positive linear relationship between the planet radius and the star luminosity.
- For the pair of variables `planet_mass` and `star_mass`, the results show that the linear regression model has a poor fit, with a RMSE of 904.487 and a R^2 of 0.056. The p-values of the intercept and the coefficient are both very small, indicating that they are statistically significant. The confidence intervals of the intercept and the coefficient are [-313.788, -156.004] and [467.506, 683.408], respectively, indicating that they are relatively wide and do not include zero. This means that there is a positive linear relationship between the planet mass and the star mass, but that the relationship is very weak and has a lot of variability. The coefficient of 575.457 means that for every unit increase in the star mass, the planet mass increases by 575.457 units on average, but with a large margin of error. This partially supports the second hypothesis that there is a positive linear relationship between the planet mass and the star mass, but the hypothesis is not very strong.
- When we looked at how the orbital period and the star radius are related, we found that the linear regression model was not a good fit at all. The RMSE was 65.344 and the R^2 was 0.001, which means that the model could not explain the variation in the data well. The intercept and the coefficient were both very small and had very narrow confidence intervals that did not include zero. This means that they were statistically significant, but not practically meaningful. The coefficient of 0.002 means that if the star radius increases by one unit, the orbital period only increases by 0.002 units on average, which is a very small change. This goes against our third hypothesis that the orbital period and the star radius have a negative linear relationship.

Based on the regression analysis, we can say that:

- The planet radius and the star luminosity have a moderate and significant positive linear relationship, which supports our first hypothesis that larger planets orbit around brighter stars.
- The planet mass and the star mass have a weak and significant positive linear relationship, which partly supports our second hypothesis that heavier planets orbit around heavier stars, but there is a lot of variation and uncertainty in the data.
- The orbital period and the star radius have a very weak and significant positive linear relationship, which does not support our third hypothesis that shorter orbital periods orbit around smaller stars, and the model has almost no predictive power.

Conclusion

In this project, we learned that:

- Out of all the planets in the dataset, only two of them are habitable, and they have low mass and radius, and orbit around stars that have low mass, radius, and luminosity.
- There is a moderate and significant positive linear relationship between the planet radius and the star luminosity, which supports the idea that larger planets orbit around brighter stars.
- There is a weak and significant positive linear relationship between the planet mass and the star mass, which partly supports the idea that heavier planets orbit around heavier stars, but the data is not very reliable.
- There is a very weak and significant positive linear relationship between the orbital period and the star radius, which does not support the idea that shorter orbital periods orbit around smaller stars, and the model is not useful for prediction.