Isa Maguire

Data Science Proj 2 Report

Abstract

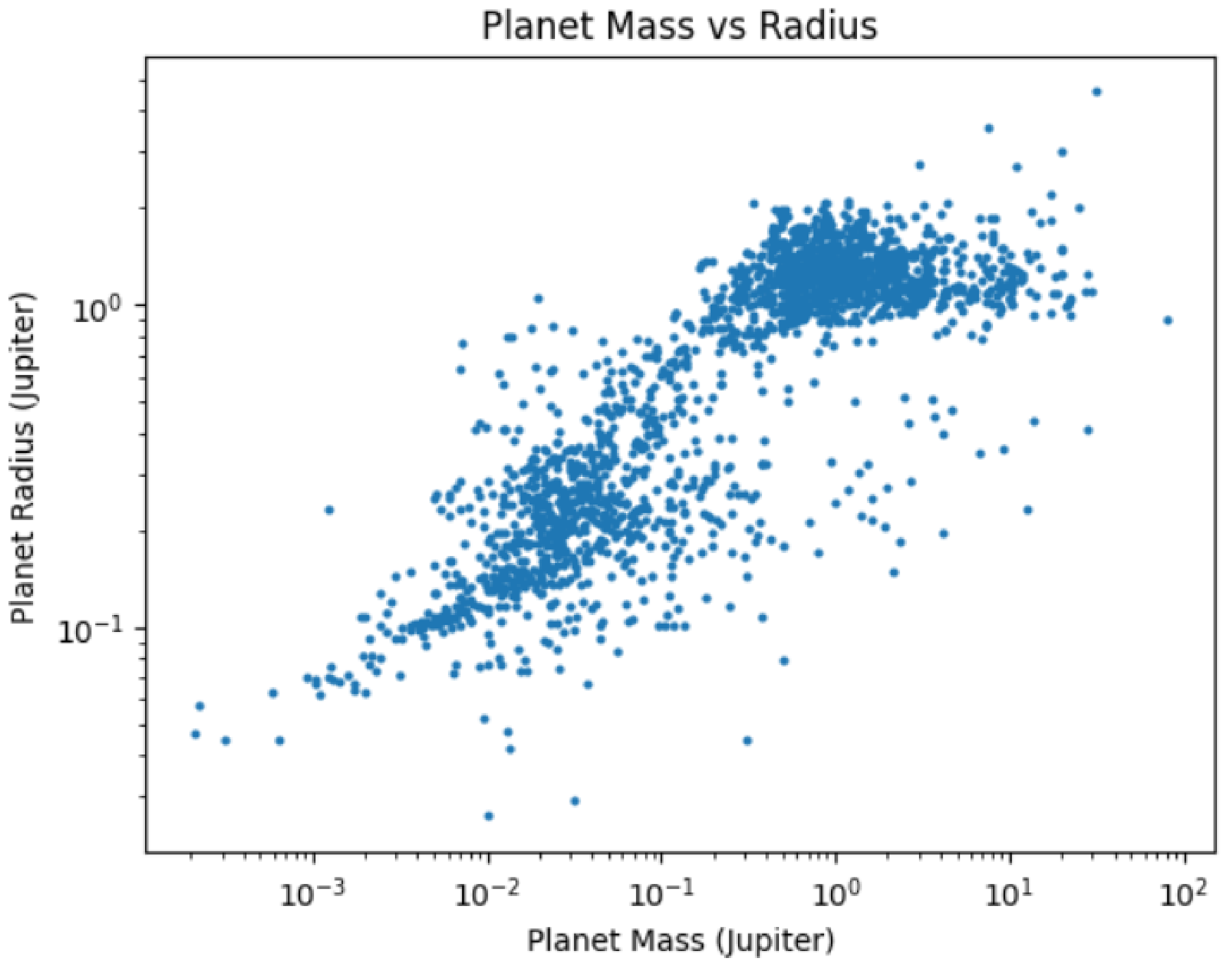
Does the amount of energy that reaches a planet from its star have any effect on its density? If we could guess what kinds of planets would form based on the energy they receive from their star, we could then predict what a solar system would possibly look like. To investigate this relationship, I plotted the two against each other and measured the Kendall Rank Correlation Coefficient and P value to test my hypothesis: Like seen in our own solar system, stars usually form dense rocky planets closer in and lighter gas planets further out. My evaluation revealed an extremely low p-value below an alpha of 0.05 with a slight negative correlation, the opposite of what I hypothesized. I concluded that this could be due to there not being enough density measurements for colder planets.

Background

A solar system is first born when a disc of dust spins around the newly formed star. As the disc cools and the dust clumps together, planets are created. In our own solar system, we see a pattern of the more dense rocky planets being closer to the star and the less dense gas planets being further away. Proving consistency in how solar systems are formed could help us understand places where we aren't able to measure planets yet. If there is a proven relationship and we see a star with a gas planet that is getting x amount of energy from its star, we could predict that there may be smaller planets with closer orbits. To measure energy that reaches the planet, I used planet equilibrium temperature. This measurement assumes that each planet is a blackbody which means it absorbs all the energy that hits it, and calculates what the temperature of that planet would be based on that assumption. That means every planet has the same criteria for evaluation, and things like the atmosphere of the planet won't be taken into account for the temperature. All planets will be consistently measured by the amount of energy that reaches it from the star. I placed this on my x axis and the density of the planet on the y axis. Another important variable in this project was the spectral type of the star. Our sun is a G2-V star: G the type of star which in this case is yellow dwarf, 2 is a classification for temperature within that type, and V means it is a main sequence star which is burning hydrogen.

Data

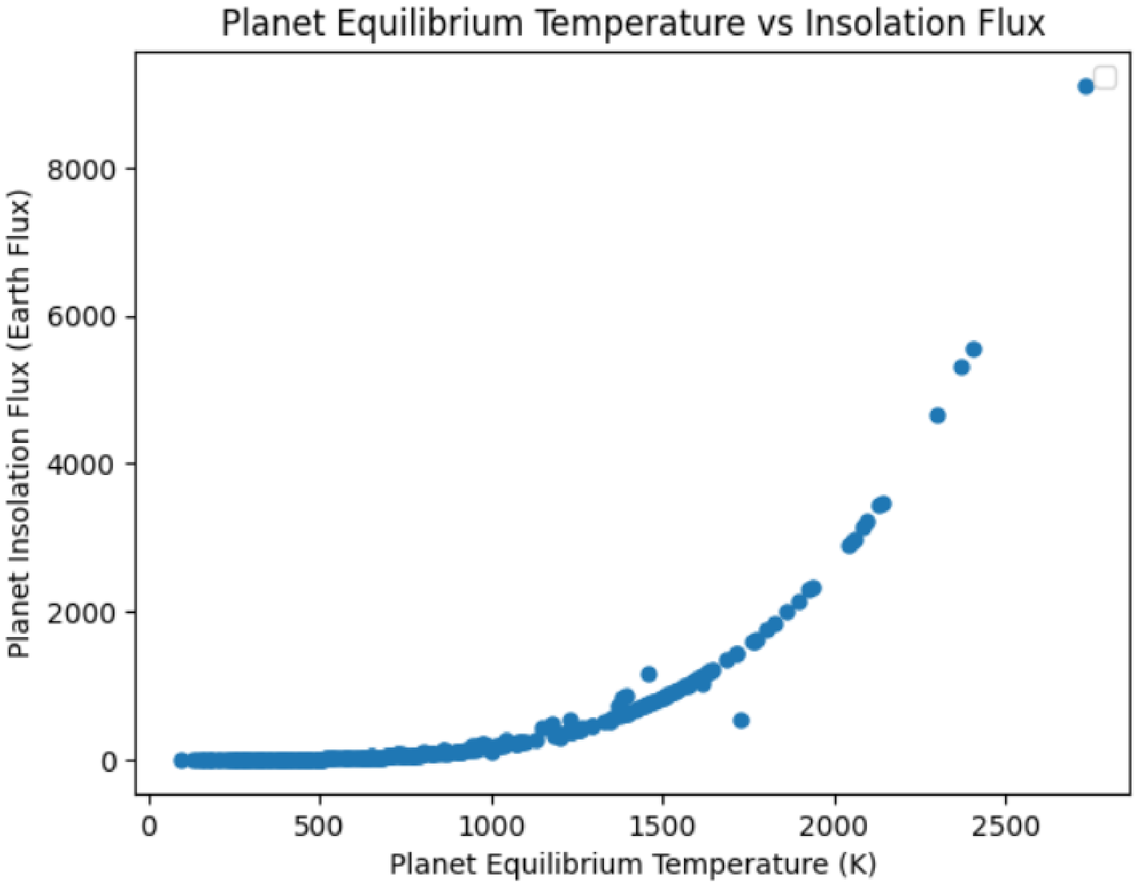
The exoplanet database is NASA's constantly updated resource where we can find every known exoplanet (rows) and some of its properties as well as information about the stars they orbit (columns). In one of our previous encounters with this dataset, we plotted planet mass and radius against each other to measure three correlation values between the two. Here is an example of one of the plots we made with the exoplanet database.



The data I worked with in this investigation was only the subset that contained measurements for both equilibrium temperature and density. Out of 5312 planets this was only 389. Originally I planned to use Insolation Flux as a measure of energy that reaches the planet, but that subset only contained 197 planets. Equilibrium temperature is directly proportional to the Insolation Flux so

I decided that substituting the two wouldn't hurt the outcome. Here is a plot of the two against each other.

-



It is possible that with such a small number of points in these subsets, we won't get a sample that represents the entire database or planets in the universe. We may need more data and improved detection methods to make a grand assumption like the one in my hypothesis, but the database currently contains all the known data on exoplanets that NASA has and some of them don't have measurements for equilibrium temperature, density, or both.

Procedure

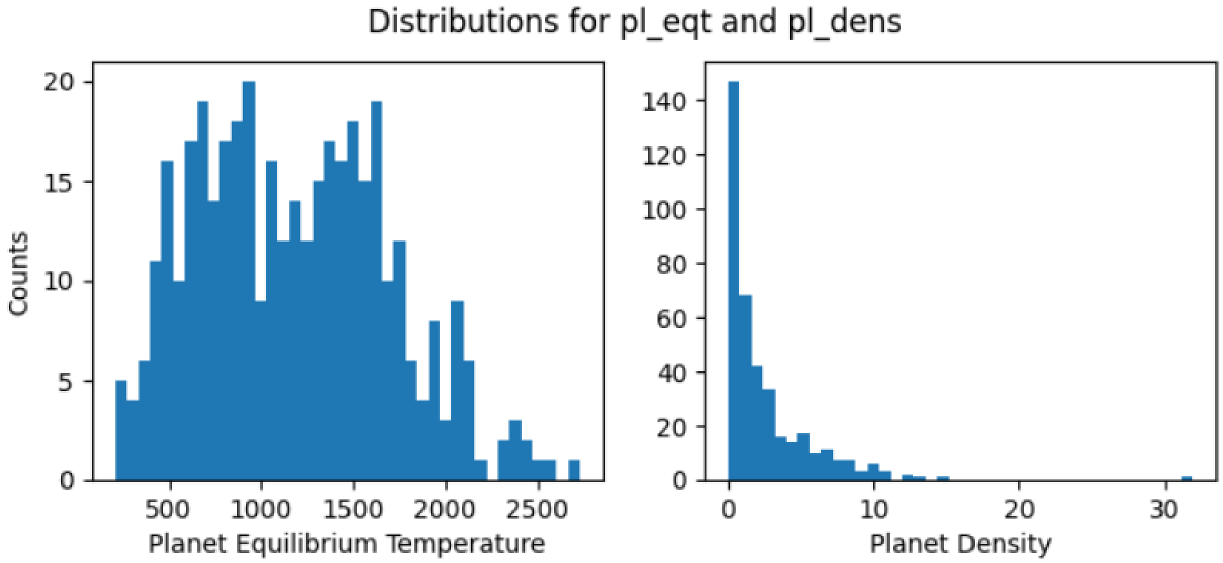
To capture the subset of data I wanted, I used one line of code and the dataframe.dropna() method to only use the planets that had measurements for both planet equilibrium temperature and density. I then created arrays out of each column and plotted their distributions in a histogram, using the techniques we learned in lab 8 (tuples of axes, subplots) to put them next to each other in the same figure. Since neither one of these has a normal distribution, and also because planet density has significant outliers, we are automatically unable to use Pearson's r to measure correlation. This means we need to use Kendall’s coefficient, which is not parametric and measures the level of concordance or discordance between the points. I then created the scatter plot and included the measurements for our solar system to see if they followed the same trend as the exoplanets.

Once this was done I created another dataframe with the same columns plus the stellar spectral type using the same dropna() method. I decided that I might be able to get better data if I took the spectral type into account because the number and type of planets in a system can change when the star becomes larger at the end of its life. For this reason I only looked at stars with spectral class V which is the same as our sun. The sun’s full spectral type is G2-V, where G is the star type (yellow dwarf), the number represents temperature, and the V means that the star is a main sequence star. This means that the star is burning hydrogen, and also that it hasn’t expanded to the point of altering its planets. This filtering resulted in only 152 points. I used the same plotting methods to show the distributions, which ended up being similar to the ones before. For the same reasons I chose to use Kendall’s coefficient to measure the correlation.

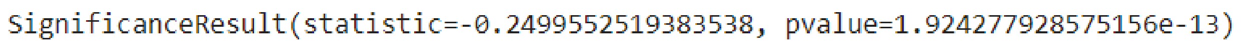
The last thing I did was get the distributions for each column of the exoplanet archive separately. Earlier when I filtered by rows that had both measurements I only got 389, but there were 820 measurements for equilibrium temperature and 522 for density. I used the same methods as above to plot the two next to each other.

Discussion and Analysis

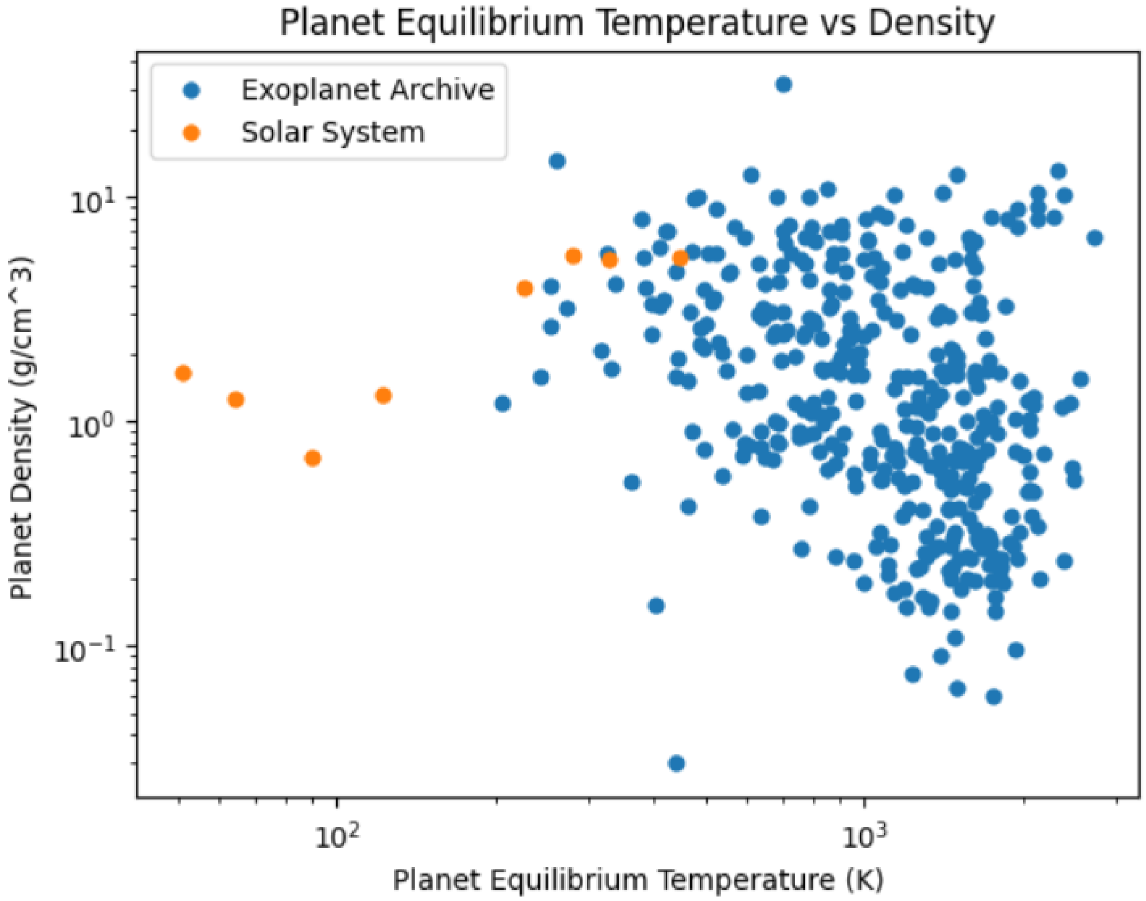
I hypothesized that planets with higher equilibrium temperatures would also have higher densities, which means my null hypothesis is that there is no relationship between the two. My first step was to determine how to measure correlation. I wanted to use Pearson's coefficient but first I had to see if my data passed the necessary requirements. I plotted the distributions of planet equilibrium temperature and density to check.



Planet equilibrium temperature has a bimodal distribution and planet density was highly skewed to the right. This automatically disqualified me from using Pearson, so I had to use Kendall Rank to measure correlation. Below is the scatter plot with our solar system's planets included, as well as the Kendall Coefficient and p value.

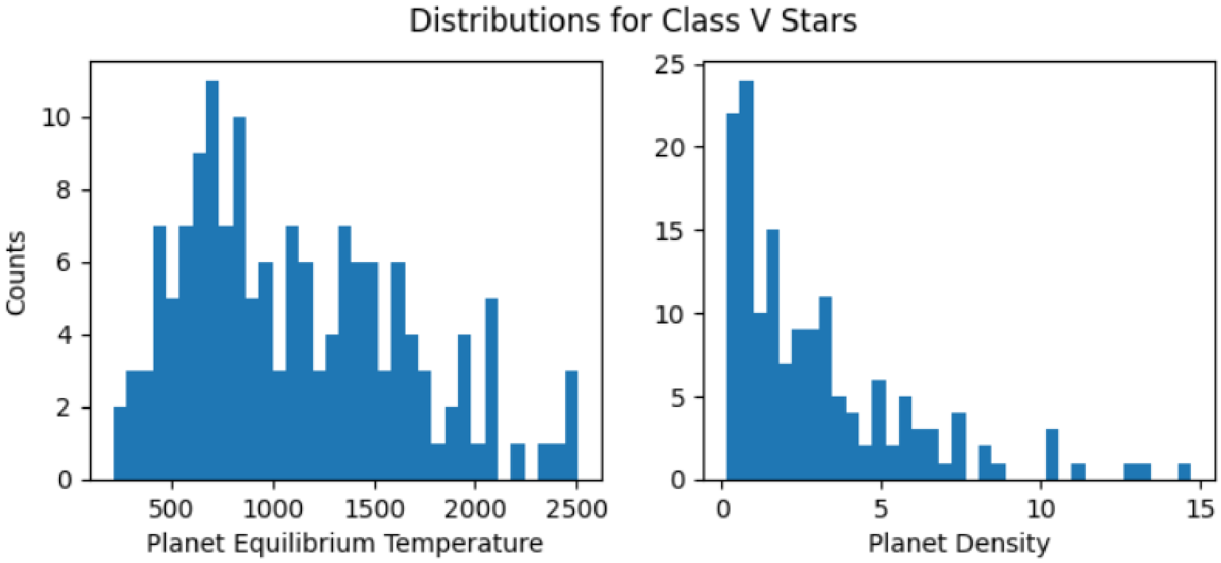


```
SignificanceResult(statistic=-0.2499552519383538, pvalue=1.924277928575156e-13)
```

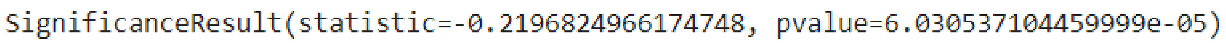


With a P-value way below our alpha of 0.05, I had to reject the null hypothesis and conclude that there was in fact a relationship between equilibrium temperature and planet density. The issue was that my correlation coefficient was -0.24, which meant that there was actually the opposite relationship that I originally thought even though the strength of that relationship was weak. I thought that maybe I needed to be more critical with the data that I included, so I decided to only use planets orbiting stars that were in the same life stage as the sun. As a star ages it can expand which will envelop the inner planets and possibly strip the gas off the gas planets. This would mean that we didn’t get to see that system as it formed in its early life. To account for this I filtered the existing data to only include the stars like the sun in their main sequence with class V.

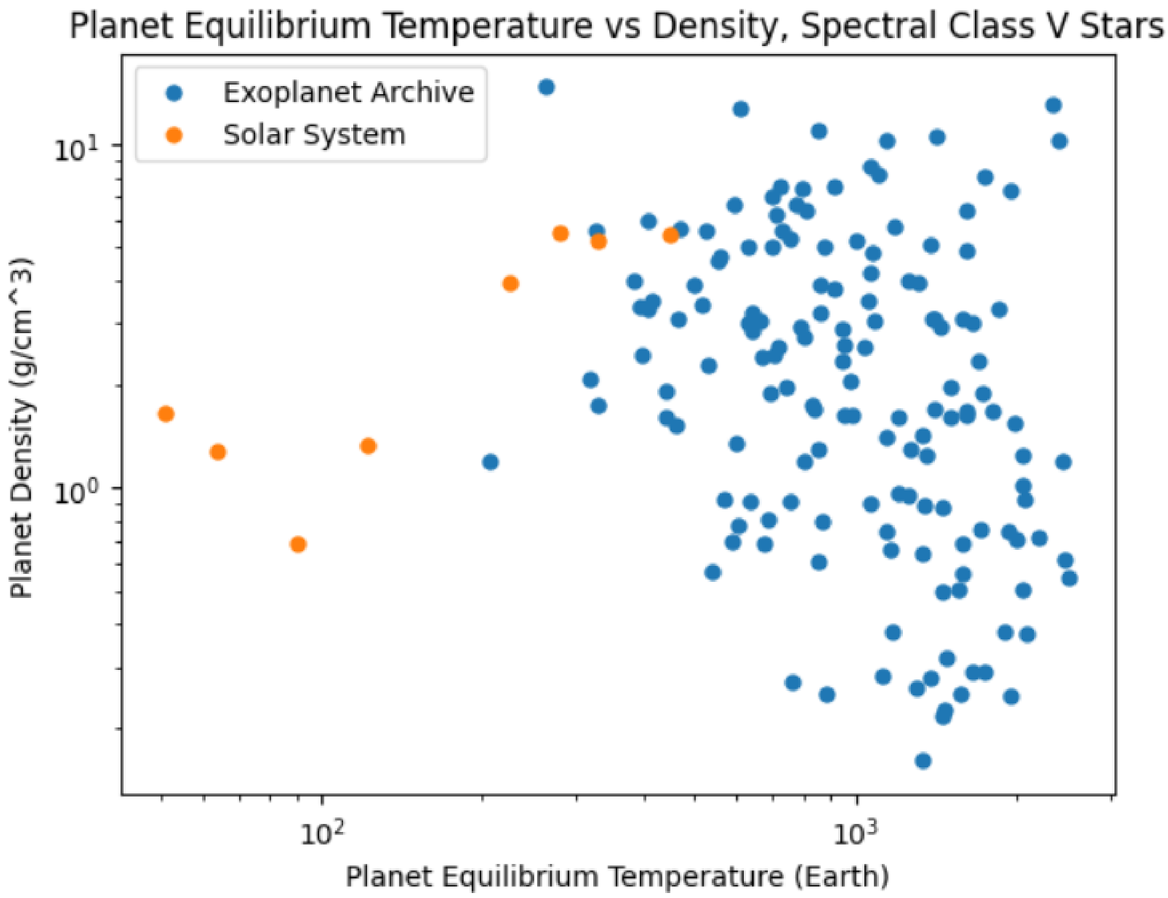
I wanted to see if the distributions for the two original variables changed so I created another side-by side plot.



Again, we cannot use Pearson to measure correlation because these two distributions are not normal. I plotted the two against each other again and used the Kendall coefficient to measure correlation.

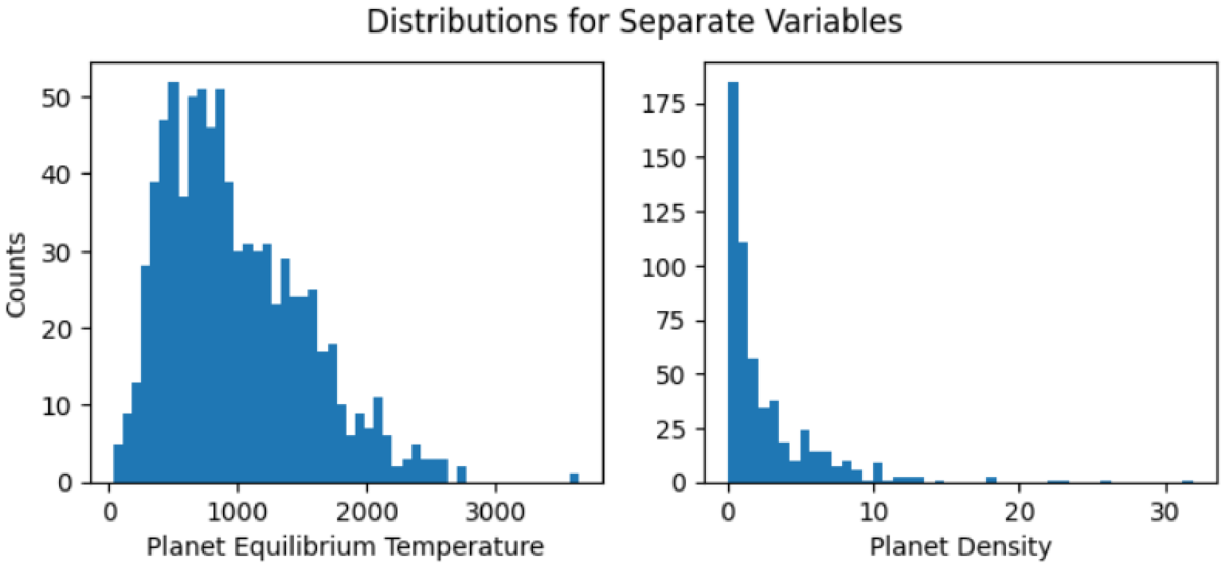


```
SignificanceResult(statistic=-0.2196824966174748, pvalue=6.030537104459999e-05)
```



The correlation and P-value for class V stars was nearly identical to that of all stars, which indicated no change based on star type. Again I had to refute the null hypothesis to determine that there exists a weak relationship between the two, it’s just the opposite of what I thought.

One issue that glares when I see these scatterplots is the lack of points near the gas planets in our own solar system. We don’t have any planets in the graph with equilibrium temperatures as low as theirs were. I wanted to see if this was true for all data, so I graphed the distributions of each investigated column of the exoplanet archive.



Here we can see that there are in fact measurements for colder planets in the range of our gas giants, but they didn't make the cut when we took points that only had measurements for both. This leads me to think that we aren't able to measure density for planets that receive a certain amount of energy, or are a certain distance away from the star. With the measurements we have so far it can be concluded that there is a weak negative relationship between planet equilibrium temperature and density, but this could change as we are able to gain density measurements for colder planets.