For my final project, I decided to visualize some of the data from the Caltech exoplanet repository online. I initially wanted to create a network from the data and link planets together based on shared characteristics or habitability. Network analysis is very useful for pulling information from interconnected systems, such as finding vulnerabilities in power line networks or social networks. This doesn't really apply here as there are no actual connections between these planets apart from the fact that they may share some characteristics or exist within the same solar system. Instead, I decided to just use certain visual aspects to represent some of these characteristics making it easier to understand some of this data.

First, I downloaded the confirmed exoplanets database from the Caltech website into a comma separated value(csv) file. I began sifting through the data looking at what I actually had at my disposal. I quickly realized that there was less data for each planet than I had assumed. I was hoping there would be information about atmospheres and albedo. It makes sense that this data was absent as these things are very hard to detect. However, I did have some very useful information about these planets that I could use to draw some of my own conclusions from. This information includes planet radius, semimajor axis, planet mass, planet density, stellar mass, stellar radius, and more. There is all sorts of useful information that can be calculated from this such as stellar luminosity, the habitable zone of the system, and an estimate of the composition of each planet.

At first, I was going to parse the data into my python program directly from the excel file. This would have been a lot of work though, so I was happy to find an API on the website that gave me a direct line into the database from my code. From there I sorted out what data I wanted from each planet and what was more or less useless to what I wanted to do. The most useful data was the density of the planet. Unfortunately, most of the planets didn't have this data. About 400 out of the 4000 total planets had density values available which narrowed the dataset down nicely into something a little quicker to process.

I wrote three functions, one to determine the stellar luminosity of each host star, one to take that luminosity and determine the habitable zone with an arbitrary albedo of 0.3 (again, albedo was not provided but would have been very useful information). This was the most far-fetched calculation made, as there were assumptions made about the albedo and there are other factors that did not go into account such as the atmosphere of the planet, oxidation of surface minerals, weathering, condensation of water, and more.

The last function I wrote was to estimate the composition of each planet based on the given density. This function compared their densities against the densities of common materials such as iron, lead, and gold, as well as against the densities of planets here in our solar system like Earth, Mars, Neptune, and Jupiter. This function determined that many of the high mass planets are hot jupiters and the low mass planets are mostly terrestrial with similar compositions to Earth and Mars. I then wrote a function that took all of this data and put it into a csv file so it could be read by my network visualization software, Cytoscape.

In Cytoscape I played around with different styles, trying to find something that brought out the most information. I finally decided to sort the planets based on mass, give them an individual size based on radius, and give them a color that corresponded to their composition. This really made it clear that the larger planets are mainly Jupiter and Neptune sized planets. I also created another diagram with the same layout, but the color was based on the method that was used to find them. Most were found with the transit method and only a few (of the limited dataset) were found with other methods such as orbital brightness modulation, pulsar timing, and radial velocity.

The most interesting planet I found in this dataset was the planet orbiting the only star found with pulsar timing, PSR J1749-1438 b. The reason this planet initially piqued my interest was because I was curious to see what the density of it was given its proximity to the pulsar. It turned out to be one of the densest planets in the dataset with a density of 23 g/cm^3. After some research I discovered that this planet is likely the remnant of another star that avoided total destruction during the supernova that lead to the eventual formation of the pulsar. Additionally, its composition is likely crystalline carbon with a density greater than that of diamonds. That's right - it's a diamond planet. It's too bad it's only 0.004 AU away from a pulsar.

Overall this project was very useful for exploring the ways exoplanets are discovered and how certain aspects are determined such as habitability and composition. Visualizing the data made it much easier to get a grasp of some of the information that was a calculated or taken from the Caltech database.