**Any surprises from your domain from these data?**

As I spend more time exploring the \*\*NASA Exoplanet Archive” [1] that roughly 10-12 new exoplanets are discovered *every week.* I find that to be astounding; it is amazing how far we can see into the universe now.

**The dataset is what you thought it was?**

The Kepler Objects of Interest (KOI) dataset is downloadable as a csv. True confession: I typically look to Microsoft Excel to be my first exploratory tool. I know it well, and with the exception of very large datasets, it does contain lots of exploratory and analytic functionality.

Upon opening the csv the first thing I noticed was that the first 144 rows were dedicated to variable definitions. I deleted them from the csv file and saved them to a separate file.

The Kepler Name column had missing values, but only for the False Positives; which makes sense as they are not exoplanets.

**Have you had to adjust your approach or research questions?**

I believe the complexity of the dataset is going to pose a challenge. It contains over 140 variables, which, I knew at the outset will likely require investigating feature reduction. However, the variables belong to a number of categories:

* Transit Properties
* Threshold-Crossing Event (TCE) Information
* Stellar Parameters
* KIC Parameters
* Pixel-Based KOI Vetting Statistics

I will need to research how I might factor in the category/variables relationship.

Another option is to fit the model to subsets of the original dataset, where each subset is specific to a category.

**Is your method working?**

**What challenges are you having?**

**References**

[1] https://exoplanetarchive.ipac.caltech.edu/index.html