David Kinney DSC 680

Spring 2021 Professor Catherine Williams

**Assignment 2.2**

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

As I spent more time exploring the \*\*NASA Exoplanet Archive” [1] I was a bit astounded that roughly 10-12 new exoplanets are discovered *every week.* 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 dataset 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. However, there were about twenty completely empty columns that I dropped.

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

I believe the complexity of the dataset might pose a challenge. It contains over 140 variables (although as I previously mentioned, twenty of those were empty), which I knew at the outset will likely require applying feature reduction.

The variables belong to a number of categories:

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

In light of this, I’ve subsetted the original dataset into separate datasets for each category. I am going to take two passes at Dimensionality Reduction by applying PCA to each subset. If that proves fruitful, I will then merge the subsets into another dataset, and fit PCA against that as well.

**Is your method working?**

As the saying goes, “so far, so good”, but it’s been my experience that you can only do so much wrangling, exploration, feature reduction and extraction, and still not really know if a method is working until you compile that first model and pull the trigger. I’m still at the stage of trying to get the final dataset in the best shape possible. I ended the week by attempting to apply Seaborn’s “pairplot” to each of the subsets, but it kept locking up on me. I’d really like take a look at those prior to running PCA, but in the interest of time, I may put that aside for now and on Monday start on the Dimensionality Reduction phase.

**What challenges are you having?**

My favorite movie du jour is “Interstellar”. I must’ve watched it ten times by now, and I still sob at the end. I’ve read the book and the screenplay and I listen to the soundtrack a couple times week. So, please indulge me if I cadge a quote from Professor Brand: ““I’m not afraid of death. I’m an old physicist. I’m afraid of time.” OK, maybe a little melodramatic for the question at hand, but when you reach my age, time really does become something you think about constantly.

Which is all just to say, the only challenge I have right now is time. Or, to personalize that a bit more, time *management* which—like data visualization—is not something at which I excel. Old coders tend to get easily distracted. All that being said, I’m aware now, after nine classes, that it typically takes me the first couple weeks to get into the groove. So, I have high hopes. 😊

**References**

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