# Project 4: Peer Review Response

Kurt Hamblin

**Review Of:** Yoni Brande

May 6

## Review of Document

In terms of clustering algorithm nuances, there's a ton and you could write a whole thesis on all the different clustering algorithms and their specific advantages/disadvantages.The main downside to k-means clustering is that it requires you to decide the number of clusters from the start, whereas more advanced clustering algorithms decide on this during the fit. It also struggles particularly with outliers and with clusters of varying size/density, and doesn't perform well with extremely high-dimensional data because the euclidean distance metric converges to essentially a constant number with high dimensions. Looking at the Mass period plots in your document, I would personally suspect poor performance from k-means due to the data overlap and varying density/shape, and would personally probably use some form of density-based clustering. For the sake of the project though, I think k-means is perfectly fine, and a good exercise because you can explore the performance of the algorithm. I don't personally think GMM is worth the complexity here, and I would suspect that density-based clustering would perform better due to the differences in cluster shapes in the plot- I would really just stick with k-means for the sake of your sanity for a short project like this. Also, if you're trying to find subsets of populations, algorithms that don't require you to specify the number of clusters are arguably essential since it allows you to truly explore the data rather than constrain it, but if you know off the bat that there are X number of subset populations, something like k-means that forces you to specify the number of clusters would be good to identify all members of those X clusters.

In order to differentiate between "real" clusters and those originating from obersvational effects, I think you would need to train a classifiction network of some kind on datasets of known exoplanet subgroups. You could then classify exoplanets, and the remainder would be either new subsets of exoplanets or those originating from observational effects. Depending on how severe the observational effects are, some form of outlier detection could be extremely effective (where you train on normal "real" exoplanets and pick out the strange ones). In fact, if you are unsure of the existence of outliers, it would likely be beneficial to perform outlier detection/removal first, so that you can then cluster and eventually classify exoplanet subsets. It's not clear to me whether your plan is simply unsupervised data exploration or actual source classification, so I've talked about them both (I'm just not familiar with the data so that's solely due to my lack of knowledge, sorry if my responses are confusing).