Critique for Areeba

Cady Stringer and David Aarhus

What does the plan do well?

- She considered computational time and chose to use LOOCV, which will avoid bias and is perfect because she has a small dataset. She justified it well.
- Her use of cross validation will create more accurate models.
- She Z scored all continuous variables which makes sense because you'd want everything to be on the same scale, especially with clustering models.
- She has a strong justification for using GM models, especially the soft assignment aspect and how that will help her understand her clusters better.
- For linear regression, using a scatterplot is a great data visualization because it can show you if your variables have a linear relationship, and if they do then a linear regression is the perfect model to use.

What could be improved and why?

- For one model she justifies K fold CV, and for one she justifies LOOCV. This is confusing because if she's working with the same dataset for each model, then she should probably use the same cross validation for each model unless there's a big reason not to.
- It is simple to use linear regression like she said, but if the data doesn't have a linear relationship then this kind of model won't be able to capture the true relationship between variables. We think a linear model is a great starting point, but she may need to modify her approach and choose a different model to use instead depending on how well her model performs.

- We think she should further elaborate on what she's looking for on her visualizations. For example, in her scatterplots and barplots, what variables will she be comparing? Also what variables will she use for a scatterplot and a barplot? Will they be the same variables? Different variables?
- She should specify if she means that her dataset is small in the number of rows or in the number of columns. If she isn't working with a lot of attributes, dimensionality reduction may not be the best option. She could use Ridge or Lasso to see which variables are more important, but if she already doesn't have many predictors reducing the number of columns will take away some of her prediction power in her models.

What are some limitations of the data/analysis plan?

- A small dataset limits the impact of her results. We aren't sure exactly how small the data set is, but if it's too small then her findings may not have any real-world implications because they wouldn't be statistically significant.

Overall, we like your idea Areeba and think this will create a really cool project! Good luck!