*ST599 In-person Updates:*

**Thursday 4/9**

Ethan Sun and Jeff met and talked about roles and possible directions.  (1 hour)

We discussed dividing the training data into two parts, so as to have some data left for verification.  We can use the remaining data to run the following analyses:

* PCA on categories:   What is the variability like within the types of stars assigned by hand?  Ethan believes this will be easy to implement.
* K means clustering:  How can the test set be grouped?

Our goal would be to create clusters within the training data then use it to investigate create clusters within the larger dataset.

* Hierarchical clustering:  Are there similarities or relationships among the 25 predictors?

Roles were assigned (randomly):

Sun is the Coordinator and Checker,  Ethan is the recorder, Jeff is the monitor.

Future tasks:

* ~~We still need to submit our roles to Sharmodeep, along with our Team Expectations.~~
* ~~Get everyone up and running on Git.~~
* Try out some of the above analysis.
* Understand what the variables and labels of stars mean (should be available in the dataset documentation).

**Monday 4/20**

6:00pm: everyone arrived at Jeff’s office.

6:05pm: reviewed Jeff’s work  
 -k-means clustering  
 -used dplyr to summarize 85-dimensional means of each of the 25 star classes  
 -first algorithm Hartigan-Wong: sensitive to empty clusters, threw an error!  
 -second algorithm Lloyd: not sensitive, still finds empty clusters, but gives results

6:25pm: reviewed Ethan’s work  
 -locating NAs, finding a rpoper method to address them  
 -pca analysis by star class

6:50pm: reviewed Sun’s work  
 -NA imputation  
 -hierarchical clustering ideas

Additional discussion (online) may be found at

<https://docs.google.com/document/d/1UVtS5TiKHJ54Uf--8a2xtbH5eWwD0fYvLLrws5mnP6o/edit>