Project Proposal: Data Dragons

Team members: Katherine Min, Heleyna Tucker, Josh Garman

Questions/Datasets to be used:

1. **How well can we predict caffeine consumption (in mg) will be above the daily recommended limit (400mg)?**

- (Use Total Nutrient Intakes, First Day (P_DR1TOT) and Demographic data for 2017-March 2020) Possibly use more datasets.

2. **How well can we predict sleep hours, do they get the recommended amount (7 > hours) a night?**

- Use Total Nutrient Intakes, First Day (P_DR1TOT), Sleep Disorders (P_SLQ) from 2017-March 2020, possibly use more datasets as well.

Response variables: Caffeine (mg) and sleep hours

Methods of Investigation:

Both of our questions are binary categorical predictions. Our initial likely candidates for modeling approaches are KNN, LDA/QDA, logistic regression, and perhaps some form of tree methodology (e.g. random forest). These approaches make the most sense given that our questions at hand are categorical in nature. Obviously we would likely use some form of cross-validation for tuning hyperparameters in the instance of trees, KNN, and potentially regularization in logistic regression. We may also consider SVM, however it is similar to LDA/QDA and may be somewhat superfluous if done in tandem.