**SEIS 763-01 Spring 2019**

**Project Proposal**

**Team: James Dedon, Eric Helander, Dan Ward, Lee Williams**

**What data set is being used?**

We propose to use the Eating & Health (EH) Module of the American Time Use Survey (ATUS) from 2014. The dataset can be found at: <https://www.kaggle.com/bls/eating-health-module-dataset#ehresp_2014.csv>, with the data dictionary stored here: <https://www.bls.gov/tus/ehmintcodebk1416.pdf>. In particular, we fill focus on the EH Respondent file (ehresp\_2014), which includes a wide range of information collected from respondents regarding their health, diet, activities, and economic situation.

This data set includes 11,212 observations and 37 variables. Of the 37 variables, approximately 25 are likely to be useful either as predictors or target values. The majority (n=27) of the variables are categorical and have been encoded to numeric values (e.g., 1: Yes, 2: No). Of these, the greatest number of possible valid values is 5. Six of the variables are continuous. All of the missing values have already been encoded to one of three possible values: `-1` for blank, `-2` for a value unknown to the respondent, and `-3` for a respondent's refusal to answer.

We initially considered the [Food Environment Atlas](https://catalog.data.gov/dataset/food-environment-atlas-f4a22) data set, which contains county-level data related to food environment, community characteristics, and health. While this dataset also captured some key features, it only contained around 3500 observations, which proved too small for the project. We are going to entertain the idea of using this smaller dataset as a supplemental data source to provide additional insight into the health of the American population as a result of diet.

**Is there a specific reason you picked this data set?**

We are interested in analyzing what factors are related to overall health. This dataset includes many key features that we anticipate will be relevant factors in overall health, such as amount of soda consumed, times exercising each week, number of times eating fast food, as well as several key economic features of the respondents.

**What is the question of interest?**

Broadly speaking, we are interested in factors related to positive and negative health outcomes. More specifically, we plan to evaluate the following:

1. Which behavioral (such as shopping patterns), diet (such as soda consumption and whether an individual eats meat), and activity (such as frequency of exercise) factors are most salient in predicting an individual has a healthy body mass index (BMI)?
2. What is the role of socioeconomic factors (such as income level and participation in WIC or SNAP programs) in relation to individuals having a healthy BMI?
3. What clusters arise that group behavioral, diet, activity, and health data?

**What methods do you plan to use?**

We plan to use feature selection techniques to determine which variables have the greatest predictive power. Since many of our predictor variables are categorical, we will end up making frequent use of one hot encoding. We will consider various regression techniques to predict an individual's BMI (a continuous value). We will also likely try breaking BMI into underweight, healthy, overweight, and obese categories, for which we would try a variety of classification algorithms. Classification, likewise, would be useful for predicting an individual's self-assigned health status (Excellent, Very Good, Good, Fair, or Poor). For any given question, we will try a variety of algorithms and try to eliminate the features that have negligible relevance and find the optimum hyperparameters.