# 2.2 Course Project: Milestone 2--Data Selection and Project Proposal

## Introduction

**Background**

**The American Time Use Survey (ATUS) Eating and Health (EH) Module was fielded from 2006 and 2008, and again in 2014 and 2016. The EH Module data files contain information related to eating, meal preparation and health in the United States. [1] This overview is provided by the United States Department of Agriculture (USDA) Economic Research Service website:**

**Individual decisions about how to use the 24 hours in a day have short- and long-term implications for income and earnings, health, and other aspects of well-being. Understanding time use patterns can provide insight into economic behaviors associated with eating patterns as well as the diet and health status of individuals. Knowing more about eating patterns, grocery shopping, and meal preparation, as well as understanding whether participants in food and nutrition assistance programs face different time constraints than nonparticipants can inform the design of food assistance and nutrition policies and programs. [2]**

**Problem Statement**

It’s more or less generally accepted that “eating badly” results in poor health. However, over the course of my lifetime, what constitutes eating healthy has changed, sometimes drastically (such as consuming fat, dairy, etc.). I will analyze the effects of meal preparation, fresh vs. fast food and snacking patterns on health metrics, such as weight and BMI.

**Scope**

The scope of the project will revolve mainly around the 2014-2016 Eating and Health Module data; however, this may expand if other interesting datasets are discovered during the project’s lifecycle, as mentioned below in the **Data Sources** section.

## Preliminary Requirements

**Technical Approach**

1. Data Gathering – this step will follow the retrieval of the datasets mentioned below.
2. Exploratory Data Analysis – initial descriptive statistics and visualization to get a feel for the datasets.
3. Data Cleaning – based on the results of EDA, clean and transform the data as deemed necessary.
4. Modeling – examine different models for best fit, applying best in class validation processes.
5. Predictive Analysis

**Data sources or plan for data** [1]

There are 3 datasets from 2014:

1. The **EH Respondent** file, which contains information about EH respondents, including general health and body mass index. There are 37 variables.
2. The **EH Activity** file, which contains information such as the activity number, whether secondary eating occurred during the activity, and the duration of secondary eating. There are 5 variables.
3. The **EH Replicate weights** file, which contains miscellaneous EH weights. There are 161 variables.

In addition, the USDA houses a number of other datasets that might bear reviewing; they can be found here: <https://www.ers.usda.gov/data-products/eating-and-health-module-atus/>

**Analysis**

Using various descriptive statistics and visualization techniques, I will analyze the effects of eating habits on respondents’ health factors.

**Requirement Development**

I plan to rely almost exclusively on **Jupyter Lab**, building a notebook leveraging the Python language. This will allow me to output presentation-quality documentation at the conclusion of the project via Markdown and the **nbconvert** Jupyter extension. In addition, if any **R** coding is required, I can add that into the same notebook via available Python libraries.

**Model Deployment**

For this project deployment will be a static approach, as the data used is based on historical survey data; there is no frequently published data. Therefore, deployment will be informal, and there will be no need for ongoing monitoring, as this is a batch-centric model, and not online.

**Testing and Evaluation**

At the lowest level, testing will be performed by breaking the data up into training and testing datasets. This will allow me to evaluate whether or not my model(s) generalize well. Delving a little deeper, depending on the model I end up with it will be possible to perform evaluation by utilizing *cross-validation.* More specifically, *k-fold* cross validation takes the concept of training/testing datasets one step further, by splitting the data up into *k* “folds”. The model is then trained on each fold, and the performance metrics are averaged.

## Expected Results

My expectation is that we will see a direct correlation between eating habits/food sourcing and overall health characteristics (another example of garbage in/garbage out)?

## Execution and Management of Project

**Project Plan**

At this point in time, my plan is simply to follow the Milestone requirements and timeline.

* Week 1: Milestone 1 Due (Team Information/Communication Plan)
* Week 2: Milestone 2 Due (Data Selection and Project Proposal) & Peer Review
* Week 4: Peer Review
* Week 5: Milestone 3 Due (Preliminary Analysis)
* Week 6: Peer Review
* Week 9: Milestone 4 Due (Project Presentation & Status) & Peer Review
* Week 10: Peer Review
* Week 12: Milestone 5 Due (Final project paper and presentation) & Peer Review (Due Saturday!)

**Project Risk**

A preliminary viewing of the data reveals rather robust and voluminous data, which is a good thing. However, the risk in having a large number of features, spread across multiple datasets, is not being able to “see the forest for the trees”. A challenge is going to be to intelligently evaluate the features across all three datasets and retain just those features that bring value to the project.

**References**

[1] <https://www.kaggle.com/bls/eating-health-module-dataset>

[2] <https://www.ers.usda.gov/data-products/eating-and-health-module-atus/>