**Predicting Obesity**

Not a Piece of Cake

|  |  |  |  |
| --- | --- | --- | --- |
| James Dedon  GPS  University of St. Thomas  St. Paul MN USA  [dedo3671@stthomas.edu](mailto:dedo3671@stthomas.edu) | Eric Helander  GPS  University of St. Thomas  St. Paul MN USA  [hela6565@stthomas.edu](mailto:hela6565@stthomas.edu) | Dan Ward  GPS  University of St. Thomas  St. Paul MN USA  [ward1516@stthomas.edu](mailto:ward1516@stthomas.edu) | Lee Williams  GPS  University of St. Thomas  St. Paul MN USA  [lee.williams@stthomas.edu](mailto:lee.williams@stthomas.edu) |

# ABSTRACT

In a bold attempt to tackle unanswered or frequently challenged questions in America’s battle with obesity, the authors attempt to create a novel approach to processing data from the American Time Use Survey (ATUS) Eating & Health (EH) module and train models to shed light upon what may be perhaps unknown factors in determining a person’s body mass index classification.

Using techniques different than those covered over the course of the semester, an attempt was made to produce a robust tool for pre-processing numerical and categorical data and apply feature selection techniques and test the results on a variety of classifiers models.

The outputs of the pipeline models seemed weaker than what the authors had come to expect from the examples and datasets covered over the course of the semester, but this was to be expected and the authors attempt to explain the causes.

A decision tree classifier model was also fitted and visualized to address the initial problem and motivation for the analysis of the EH module. This produces a solution to the problem where we sought to identify the most significant factors in the survey that may have an effect on BMI.

# CCS CONCEPTS

• Computing methodologies~Feature selection, Computing methodologies~Supervised learning by regression, Computing methodologies~Supervised learning by classification, Computing methodologies~Classification and regression trees, Computing methodologies~Support vector machines, Computing methodologies~Gaussian processes [14]

# KEYWORDS

BMI, General Health, Kaggle, ATUS, Decision Tree, scikit-learn, OneHotEncoder, NumPy, data preparation, categorical variables, numerical variables, regression, SVM, Classification

# PROBLEM

Our initial problem of interest was to determine which behavioral, lifestyle, and socio-economic factors could be used to predict good or poor health, as measured by either BMI (‘ERBMI’) or a subjective self-rating (‘EUGENHTH’). Attributes of interest included dietary (e.g., amount of soda consumed and frequency of fast food consumption), behavioral (e.g., times exercising per week), and a variety of economic data points (e.g., income in relation to the poverty threshold).

# MOTIVATION

Being able to identify health situations that are predictive of negative health outcomes could be instrumental in helping connect individuals to interventions before irreversible health conditions have fully developed. This has implications for not only individuals but for society more broadly. In 2017, healthcare spending in the U.S. averaged $10,749 per person, or 17.9 percent of Gross Domestic Product.[1]

Poor health results in financial consequences at virtually all levels of the economy. Individuals and households pay directly for treatment and indirectly through rising healthcare costs; employers face higher costs in providing healthcare for their workers; and the government can end up footing the bill to provide healthcare to its citizens. Insurance companies face larger and more frequent claims and must refactor their models and increase premiums for those that are causing greater risks to their profitability. Furthermore, as households spend more on healthcare, fewer resources are available for other goods and services or for saving and investing[2].

Body mass index (BMI) is a simple measure of obesity. BMI is calculated based on an individual’s height and weight. Specifically, or . BMI is used to estimate an individual’s body fat and risk for a variety of diseases. While BMI is not a perfect measure, as it does not factor in the composition of a person’s weight, meaning lean mass vs body fat, in general, it is a good predictor. A higher BMI is associated with heart disease, type 2 diabetes, breathing issues, and some types of cancer.[3]

# DATASET

## Background

We selected the respondent file (‘ehresp\_2014’) from the Eating & Health (EH) Module of the 2014 American Time Use Survey (ATUS)

The data for ATUS comes from nearly 200,000 interviews and is the foremost federal survey that focuses on the entire spectrum of non-market activities. ATUS is used extensively for analysis and research related to economics, health and safety, work and leisure, and conducting comparisons across countries. Refer to the later section, Analysis of Other Research Using This Dataset, for specific examples[4].

ATUS is conducted by the U.S. Census Bureau and is sponsored by the Bureau of Labor Statistics. The EH Module, in particular, is sponsored by the U.S. Department of Agriculture’s Economic Research Service[5]. The questionnaire used for the EH Module can be accessed at the footnoted link[6].

According to the 2014-16 Eating & Health Module User's Guide (2016 Edition)[7], the ATUS EH Module “collects additional data to analyze relationships among time use patterns and eating patterns, nutrition, and obesity; food and nutrition assistance programs; and grocery shopping and meal preparation."

We retrieved the dataset from Kaggle[8] and the data dictionary from the Bureau of Labor Statistics[5]. As is typical of Kaggle, finding and accessing the dataset was without difficulty. The data dictionary was instrumental in helping us understand the variables included in the dataset and for making informed decisions regarding invalid values.

Since we will be referring to specific variables throughout this paper, a brief overview of their naming conventions is worthwhile. The first character of a variable’s name indicates for which module the data point was originally collected. All variables in our dataset were collected either for the EH Module (with an initial character of ‘E’ for ‘eating’) or the ATUS interview itself (with an initial character of ‘T’ for ‘time’). The second and (sometimes) third characters indicate the type of variable. The options for the second and third characters are as follows:

* ‘U’: unedited variable
* ‘E’: edited variable
* ‘R’: recode
* ‘RT’: summary variable
* ‘X:’ allocation flag
* ‘T’: topcode flag

The remaining characters comprise a short descriptive name for the variable. Our primary target variable, ‘ERBMI’ will serve as an example.

* The first character, ‘E’, indicates this value was collected as part of the EH Module.
* The second character, ‘R’, indicates this variable is a recode: It was calculated based on a combination of other values (in this case, the respondent’s height and weight).
* The remaining characters, ‘BMI’, indicate this variable refers to the respondent’s body mass index.

## Structure

The ATUS EH Respondent File contains one record per respondent[5] and includes 11,212 observations and 37 variables. The dataset includes 575 instances where ‘ERBMI’ contains a negative value. We excluded these records for much of our analysis, resulting in 10,637 observations that were used for training and testing. The majority of the variables (n = 27) are categorical and have been encoded to numeric values (e.g., ‘1’: ‘Yes’, ‘2’: ‘No’). Of the 6 continuous variables, only 1 (‘ERBMI’) contains decimal values; the other 5 contain integers only. Some of these continuous variables have been bottomcoded and topcoded. ‘EUWGT’ (a respondent’s weight in pounds) for instance, is topcoded to ‘340’ pounds and bottomcoded to ‘98’ pounds. If a respondent provided a value outside this range, such as ‘97’ pounds, it would be coded as ‘98’. For these situations, a separate variable (‘ETWGT’, in the case of ‘EUWGT’) indicates whether bottomcoding, topcoding, or neither occurred.

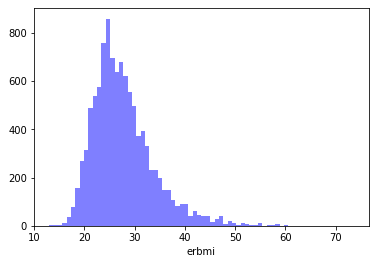
The dataset contained no null or blank values. Any missing values had been pre-coded to ‘-1’ for ‘Blank’, ‘-2’ for ‘Don’t know’, and ‘-3’ for ‘Refused’. We considered excluding these invalid values or imputing some other value. Upon further consideration, however, we determined that there may be predictive value in differentiating between whether a respondent refused to answer a question versus the value being blank. A value may be blank for any number of reasons; a respondent’s refusal to answer a question may indicate that the value, if coded, would be more likely to be extreme, or it may indicate a character trait that is associated with positive or negative health outcomes. Future analysis could focus on the predictive power of leaving these values as coded or opting for an imputed value.

Some continuous numeric columns do contain a significant number of ‘0’ values. We decided to leave these values unchanged.Future analysis could seek to determine whether ‘0’ is a valid value for these instances, or whether a different value (e.g., the mean derived from the non-zero values in that column) would be more appropriate.

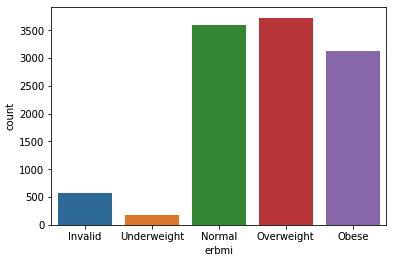
A further opportunity for future analysis could be to determine a threshold of maximum allowed invalid (e.g., ‘0’ or negative) values for either an instance or a column in the training data. If the instance or column exceeds that threshold, exclude it entirely.

## Initial Review of the Data

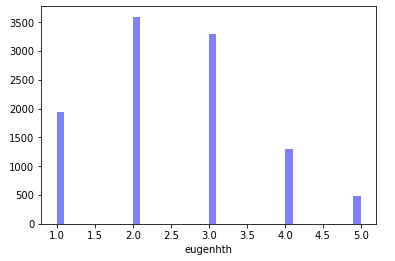
When doing an initial review of our data we started with a focus on our two target variables, BMI (ERBMI) and General Health (EUGENHTH). Looking at the cleansed attribute BMI the distribution is relatively normal with a slight right skew. There is a clear peak around the mid 20’s, with a much longer tail in the upper BMI’s representing those individuals with BMI’s in the Obese range.



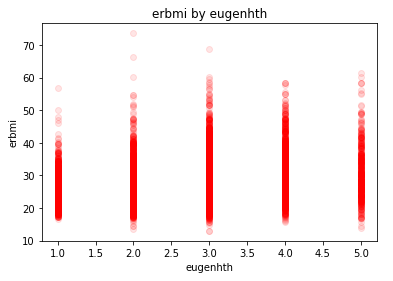
When analyzing the raw BMI scores we categorized them into the buckets of Invalid (a response of -1 to -3), Underweight (BMI <18.5), Normal Weight (18.5<BMI<24.9), Overweight (24.9<BMI<29.9), Obese (30<BMI). For our analysis we removed the Invalid responses, and as shown in the graph, this was a small percentage of the overall responses. Looking further at this distribution in the following graph, shows that the Underweight category has significantly fewer respondents than all of the other categories, with Normal, Overweight, and Obese having a relatively similar number of respondents. Other than the under-represented underweight category, this was largely the expected population distribution.



Moving on to our next target variable of General Health (EUGENHTH) we found that responses came in a relatively normal distribution. General Health was recorded as an integer with the following values. Excellent Health (1), Very Good Health (2), Good Health (3), Fair Health (4), and Poor Health (5). With roughly half of the respondents rating themselves as either in Excellent Health or in Very Good Health. This was a minor surprise as only roughly 35% of the respondents were in the normal BMI range.



The next analysis we did was to plot General Health against BMI to see if there was a correlation between a person's BMI and how they rated themselves in overall health. When reviewing this information it is clear that there is a connection between lower BMI and a person reporting themself as having excellent health or even very good health. With that said there are some rather significant outliers, most notably in those that rated themselves as in very good health (a rating of 2.) As the scale moved from Good Health (3) to Poor Health (5), there is significantly less variation in the minimum and maximum BMI between these general health categories. Our assumption is that between these categories the subjective nature of the general health category is causing the lack of variation of overall BMI range between these general health categories, or that factors like age and genetics play a role in the impact on one's health at the higher BMI scores.



## Preparation

Originally, we began loading the dataset in the manner demonstrated in class, loading it into a dataframe followed by isolating the target variable and passing the data into numpy arrays, but this quickly highlighted many difficulties in managing a larger number of variables, mostly categorical.

Therefore, instead of using the scikit-learn ‘train\_test\_split()’ function, we opted for creating a separate ‘split\_data()’ function to split the data using a numpy ‘permutation’. This enabled us to more easily, at a later time, select a specific subset (e.g., the ‘train’ or ‘test’ subset) from the original dataset.

We used the scikit-learn StandardScaler for the continuous numeric attributes and one hot encoded the categorical attributes.

The scikit-learn OneHotEncoder did not lend itself toward one hot encoding this many variables. Instead, we used the pandas ‘get\_dummies()’ function, with ‘drop\_first’ set to ‘True’. This resulted in 92 attributes. Working with this large number of variables and with NumPy arrays - which do not retain pandas DataFrame column headings - proved challenging for interpretability.

Our initial goal was to be able to explain behavioral, lifestyle, and socio-economic factors are related to good or poor health. As the data became divorced from its original state, we shifted our attention to developing the best model (measured by accuracy), recognizing that our model would effectively be a black box.

After completing the initial data preparation (ingestion, scaling, and loading), we attempted to apply a few different models to predict either BMI or BMI class (i.e., ‘Underweight’, ‘Normal’, ‘Overweight’, ‘Obese’). Simple models, such as logistic regression, ran for a reasonable duration. More complex models, such as a support vector machine, were quite slow. Working with this large number of attributes not only contributed to long-running code, but also led to poor accuracy.

We attempted to move the notebooks to Google Cloud Engine, but scaling the machines there with more cores did not have the desired effect of speeding up the process, but instead reduced total CPU load as more cores were added, so we stuck with processing the notebooks on local machines.

For a start, we implemented a ‘select\_k\_best\_categorical\_features()’ function as a wrapper for SelectKBest from ‘sklearn.feature\_selection’ to select the specified (‘k’) number of best features, as measured by a passed-in score function. We went with the chi-square score function, as that seemed the best of those available to select features that are relevant in classification models. A consequence of using chi-square, however, is that this only worked for selecting the best categorical features, and not continuous numerical features.

Of the classifier models that were fitted using our filtered features, the highest performing was the Logistic Regression classifier. It reached a 46% accuracy using 91 features.

Of the SVM kernels, the rbf, sigmoid, and poly kernels failed to be compatible with the data being produced by the pipeline. Those kernels failed with a runtime error stating that the classifier does not produce ‘\_coef’ or ‘feature\_importances\_’. We unfortunately learned this too late to correct before submission.

## Recursive Feature Elimination with Cross Validation

We implemented a ‘get\_optimum\_features()’ function as a wrapper for RFECV (Recursive Feature Elimination with Cross Validation) from ‘sklearn.feature\_selection’. RFECV takes parameters for the estimator object, a cross-validation strategy or splitter, and a scorer. The function then returns an object with attributes for the ranking of features and the score that each feature received. We started out by supplying it with a logistic regression estimator and a support vector machine with a linear kernel, both using a StratifiedKFold cross validation technique with 2 splits. After running the logistic regression estimator through the ‘get\_optimum\_features()’ function, we ended up with an accuracy in the mid-90%s and fewer than 10 features. The function took approximately a minute to run. We then tried the support vector machine. It took many minutes to run and produced an accuracy near - though not exceeding - the accuracy from the linear regression estimator.

Similar to the challenge encountered with pipelines, we soon discovered that many models do not support feature selection via RFECV because, as the error states, “The classifier does not expose ‘coef\_’ or ‘feature\_importances\_ attributes.” As a result, we only ended up using RFECV for logistic regression and a support vector machine with a linear kernel.

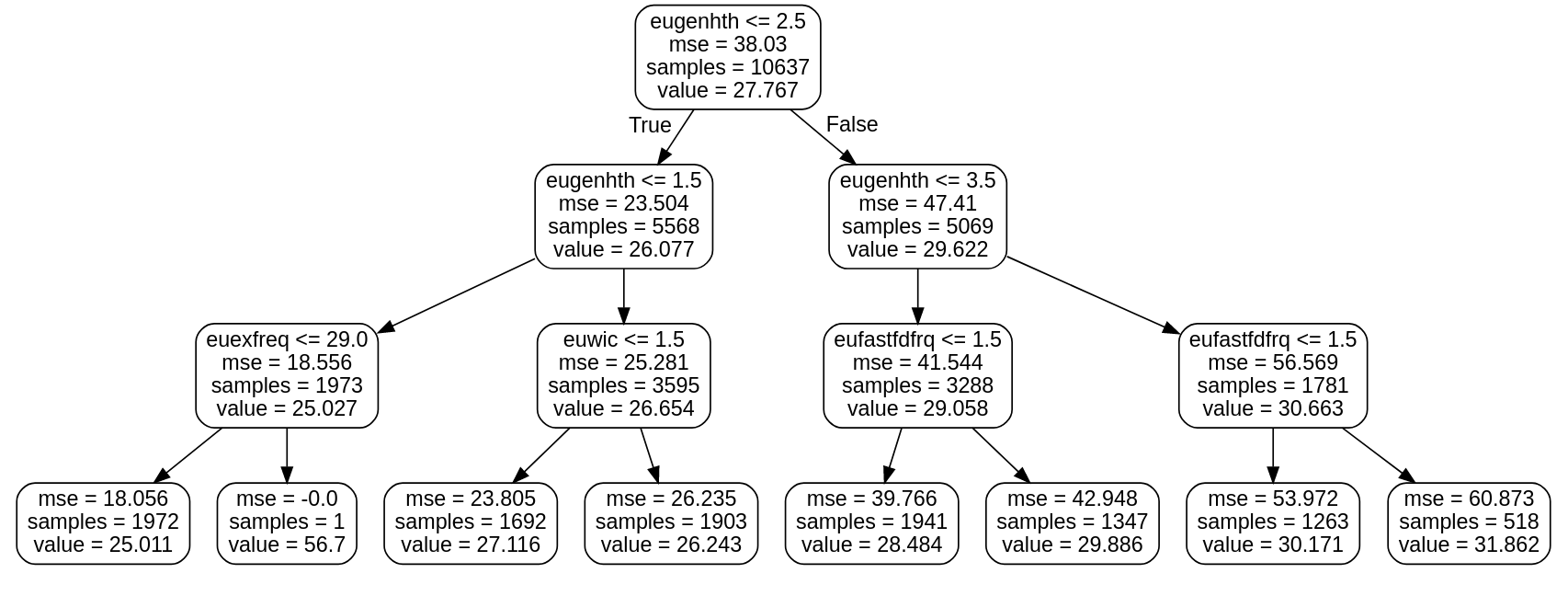
After seeing accuracies in the upper 90% range, we eventually realized that our models were incorporating the height and weight features which should give nearly a perfect score for a regression model. Removing those two features, our model accuracies dropped to the mid 40% range. At first this was somewhat disappointing, but after further consideration, this was somewhat to be expected, as we are dealing with bucketed categories that fall along a continuous scale. Edge cases in these buckets would account for those misclassifications. Linear regression models on the entire dataset produced a mean absolute error of 2.43 in BMI while factoring height and weight, which would account for these edge cases, as each bucket has roughly a 7 point range.

We rudimentarily verified results by manually selecting features from the original dataset based on their SelectKBest fit scores and ran all of the classifier models against them. This was done in the classifierVerification notebook. Using this approach, the best we could achieve by manually selecting features was in the upper 30% range, which falls below the scores achieved by the automated notebook.

## Decision Trees

As we gained familiarity with the data we found that some of the data was hierarchical: certain questions were asked as ‘follow-ups’ to others. For example, one variable encodes a response to the question ‘If you drank any soft drinks in the last week, were any of them ‘diet’ soft drinks?’ This variable is connected with the variable encoding the consumption of soft drinks, and suggested a tree structure in their analysis.

Also, the goal of our project was to discern from the study data which lifestyle and socioeconomic factors are most responsible for health outcomes. Decision tree algorithms choose the most effective variable for the next node in the tree, giving us the analysis we want.



One of the biggest benefits of a decision tree classification is the possibility of visualizing the outputs to present a hierarchical display of features with the highest relevance in deciding the classification. Using the graphviz package, we can create pleasant visualizations as seen in the DT notebook.

## Data Preparation for Tree Classification

The first data preparation step was to replace all the negative data values, as those indicated various types of missing data, with zeros as a form of dimensionality reduction. Looking at the reasons for a non-response may have value in a future, in-depth study of a particular factor, but at this stage we are looking at the larger picture.

We then removed all rows for which there was missing data for body mass index or for general health, as those were the two target variables we were looking for. This did reduce our data points by about 6%, but we wanted to focus on predicting outcomes, not predicting non-responses.

We also removed variables that represented metadata about how data was edited by the data collectors, or which year was used for calculating poverty thresholds. We also removed height and weight, as they are the inputs to the formula for Body Mass Index, and would only add redundancy to our factor analysis.

The Body Mass Index, once calculated, is usually interpreted by subdividing it’s value into ranges interpreted as ‘Underweight,’ ‘Healthy’, ‘Overweight’ and ‘Obese,’ we we replaced the given numeric BMI data with its respective BMI category .

A further edit of the data was to simplify the variables representing respondents’ income. Two variables compared a respondent’s income to 130% of the poverty threshold, and another to 180% of the poverty threshold. Presumably this was to reduce potential non-responses to a question about a sensitive matter.

A third variable was set by the survey authors as a combination of the first two. This variable had five response values. The two encoded income less than 130% of poverty threshold and greater than 180% of income. The remaining three encoded various combinations of greater than 130% and less than 180%, but did not stratify the income within that range, so these three responses were combined into one response.

Most of the remaining variables were categorical, so those we expanded into one hot vectors. The few remaining numeric variables did not require further processing, as Decision Trees are not hindered by non-normalized data.

## Tree Classification to Identify predictors of General Health

The first tree we constructed was one that would look at which variables best explain the respondents’ report of their general health class. We tried different tree depths, and found that after a height of five there was an insignificant gain in information, so we limited the tree to a depth of five.

The two factors that most explained data variance were, respectively: Healthy weight and low income. Persons who were not overweight and whose income was not less than 130% of the poverty threshold tended to have reported themselves to be in Excellent or Very Good health.

For persons who were overweight and did have low income, the next most explanatory variable was if they exercised more than three times a week. Those who did not exercise more than three times a week reported Good health if they often (more than 4.5 hours per week) engage in secondary eating; that is eating while performing another primary activity. An example of this would be eating chips while watching a television program.

However, as discussed below, exercise was the major predictor of Body Mass Index classes, so future studies may wish to remove exercise frequency as a predictor of general health as it is already present in the BMI predictor.

But, of those sampled in the overweight BMI range with a significant exercise frequency, those who did eat while performing other activities more than 4.5 hours per week tended to report their health as Fair or Poor.

Another prediction chain leading to Poor health was being Overweight, not in low income, but sometimes not getting enough to eat, and spending less than 3 hours per week sitting down to eat.

Overall, respondents tended to report themselves in Very Good or Excellent health. Less than 25% rated their health as Fair or Poor. This may reflect that the value was self-reported, that respondents did not have a clear metric, were over optimistic about their health, or were unaware of health issues. Or it may be that, overall, persons in the United States are in very good health.

## Tree Classification to Identify predictors of Body Mass Index

We also used a tree classifier to find the most significant predictors of Body Mass Index classes. We did again experiment with different tree heights, and settled on a height of four. After that there was little information gain.

The results were stark. The top predictor of a Healthy BMI was if the respondent exercised three or less times per week or not. Of the 7121 samples where the respondent did exercise three times or less, only four samples were expected to be Healthy. Their additional predictors were that those four were not recipients of the SNAP program and they spent more than 207 minutes per week in eating as the primary activity.

Of the 3486 samples who responded that they exercised four or more times per week, 3100 were expected to be classed as having a Healthy BMI. 35 samples were still expected to be Obese. Those respondents reported they sometimes did not have enough food to eat.

Interestingly, of those who exercised frequently and reported drinking soft drinks in the last week, those who didn’t drink diet soft drinks were more likely to have a Healthy BMI.

# ANALYSIS OF OTHER RESEARCH USING THIS DATASET

As we mentioned earlier in the introduction, the American Time Use Survey (ATUS) is conducted by the U.S. Bureau of Labor statistics to collect data on Americans’ use of their time and participation in different activities. The data set consists of single-year data files from 2003 to 2017, as well as multi-year data files. In addition to the comprehensive annual sets, there are also separate module files for well-being, leave, and eating & health. The significance of these datasets is that they contains a nationally representative samples, which may prove to be elusive in surveys of smaller scale, conducted by organizations other than top-level federal agencies with fewer resources.

The ATUS eating & health modules are particularly relevant for a wide variety of research because of their inclusion of numerical body mass index (BMI) for each participant, as well as self reported data of respondents’ health in a 5 category feature ranging from poor to excellent and a wide variety of other features that expose behavioral patterns, economic status, and lifestyle choices that may or may not be linked to positive or negative health outcomes.

The explosion of obesity and decrease in overall health of wide swaths of the American population over the past few decades has created significant demand for and funding of research to identify the causes of said changes, as well as solutions and other insights that may be useful to formulating meaningful public policies at all levels of government.

We identified four such projects and will in this section describe the motivations and techniques used to reach their findings.

## Paper #1

The first one we came across was “The Role of Time Use Behaviors in the Risk of Obesity among Low-Income Mothers”[9], which sought to identify the relationships between obesity and motherhood/income in a sample consisting of women ages 18-55. Like our analysis, the researchers primarily focused on the BMI feature as their primary dependent variable and categorized it into the respective weight categories assigned by the U.S. Centers for Disease Control and National Institute for Health.

Differing from our analysis, however, the authors excluded male respondents from their analysis and split the weight category into two groups, normal and overweight/obese. They excluded altogether the respondents that had BMIs that fell into the underweight category. Their decision to proceed as such was guided by a need to increase statistical power of their findings and rationalized by recognizing the similarities in health risks in both the overweight and obese population segments.

Also differing from our models, the authors relied on only household income and motherhood status of respondents. However, they used age, marital status, education, urban residence, region, race, immigration and employment status to adjust their models.

Like our findings, the researchers found income to have a significant inverse association with a BMI and weight categorization.

## Paper #2

A second paper we found, “Occupational Characteristics and the Obesity Wage Penalty”, by Jennifer Bennett Shinall of Vanderbilt University[10], used the dataset to demonstrate the wage penalties that obese women pay in the U.S. workplace.

The author matched data from the EHM to the Occupational Information Network (O\*Net) of the U.S. Department of Labor to calculate the penalty in wages that obese women pay for their weight in comparison to non-obese women and men, both obese and non obese, in the workforce.

Interestingly, the author analyzes how different groups of people sort themselves into different professions and occupational characteristics of jobs with different requirements and levels of exposure.

## Paper #3

The third analysis that we came across was “Americans’ Eating Patterns and Time Spent on Food: The 2014 Eating & Health Module Data”[11] by Karen S. Hamrick and Ket McClelland. In their research, they analyzed the dataset to identify patterns in behavior in Americans relating to the acquisition, preparation, and consumption of different foods and beverages, as well as other behavioral patterns, such as watching television and exercise. There was not much analysis beyond summarizing the 2014 results in a way that the presentation of their results seemed to be focused on presenting the EHM data set as a viable data source for future research and the publication of their findings by the EHM creators further supports this assumption. It is notable that a large amount of analysis focuses on the BMI data as a dependent variable in various charts throughout the paper. Nonetheless, this paper serves as an excellent high level analysis of the 2014 EHM data set.

## Paper #4

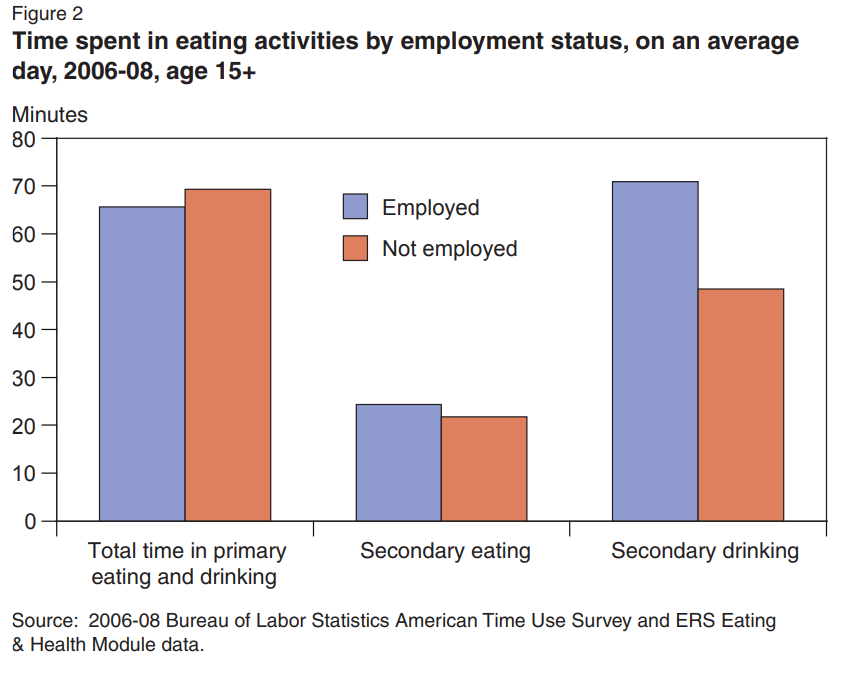
Fourth, another paper by Karen Hamrick, along with Margaret Andrews of the Economic Research Service, U.S. Department of Agriculture titled “SNAP Participants’ Eating Patterns over the Benefit Month: A Time Use Perspective.”[12] This research focused on the benefits cycle of the Supplemental Nutrition Assistance Program (SNAP), widely known as food-stamps, and its impact on SNAP recipients dietary habits ad different times of the month. The authors developed a logistic regression model that predicted the likelihood of a person going without food and extrapolated those estimates to conclude that there is a direct correlation between the distribution cycle of SNAP benefits and the probability of a person going without food on a given day of the calendar month.

## Paper #5

Finally, another Hamrick, et al. paper titled “How Much Time Do Americans Spend on Food?”[13] analyzed earlier (2006-08) EHM data to draw a picture of Americans’ habits with eating and drinking, as well as relationships of those habits with their health status and BMI. This paper did some analysis on different activities and their relationships to BMI and reported health.

The results however, are nothing surprising. For example people who eat as a secondary activity while driving, cleaning, working, or grooming had a lower average BMI than those who ate while watching TV.

The authors of paper 5 do an excellent job of presenting results that add additional understanding to the data in question, however. In the following figure they present the differences in eating habits of employed and unemployed persons.



This analysis is relevant to the motivations behind the our undertaking of this project, but like most of the other papers reviewed, the author presented a broad picture of the overall results and did not discuss any statistical or ML methodologies used to get their results.

## Summary of other research

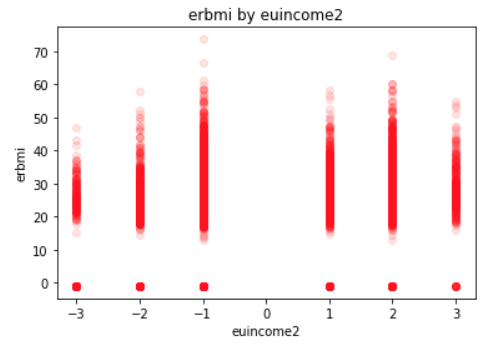
Only one of the five papers reviewed utilized the dataset to conduct any regression or categorization of the data as we have done so in our own investigations. Most of the reviewed research simply performed high level analysis of the data to draw vague pictures of the sampled populations with watered down policy proposals.

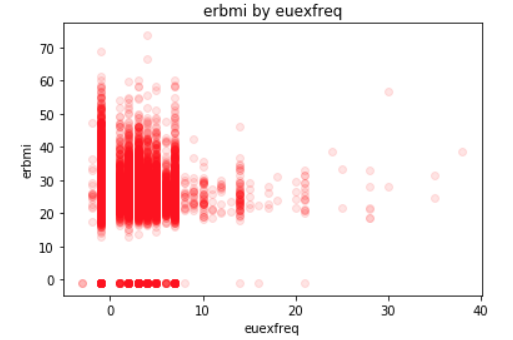
As expected, further investigation is needed in all cases.

# LESSONS LEARNED

Starting out this project it seemed as though gaining expected outcomes would be a simple matter of plugging in the data and the models would be able to predict BMI with ease. Drinking soda or snacking a lot is surely going to lead to higher BMI, right?

We found that’s not quite always the case. Individual analysis of the vast majority of the categorical features shows that while BMI ranges may vary from category to category, the means remain relatively similar. Of the three valid responses for euincome2 (1,2,3), which measures the respondents’ income in relation to 130% of the poverty threshold, variation is negligible.





Slight trends can be seen in the numerical features, but those were expected and didn’t lead to any new insights. For example, strong correlations were visible between height and weight and BMI (above), but as you now know, those are directly related. Trends in exercise frequency, on the other hand show dense clustering in the zero to eight range of exercise frequency with a slight downward trend up until around ten, but an upward trend in the outliers to the right. This could be explained by those with higher BMI feeling compelled to increase their exercise regimens.

Our main takeaway from this project is that categorical features are a finicky beast and not always so easy to handle in large numbers. The processes to manage these features, correctly processing them and maintaining functional code is perhaps the most significant achievement of this project.

By taking many of the key concepts learned throughout the semester, we sought out new approaches to handle the tasks at hand, such as utilizing the pandas get\_dummies() function to handle larger numbers of categorical features and integrating StratifiedKFold() into the estimator loop that generates our feature selection / hyperparameter exploration outputs.

While we would have liked to have created a bulletproof pipeline for the data, we in the end found that these types of things can get very complicated very quickly. Nonetheless, we are happy with the results that our notebooks produced.

While we don’t expect this work to lead to any substantial grants for future research or major scientific awards, the experience of creating this work along with the struggles and realizations truly served as a learning experience that really puts into perspective the complexities of machine learning in the real world.

# REFERENCES

[1] CMS.gov, (12/11/2018). National Health Expenditure Data - Historical. Retrieved from <https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nationalhealthaccountshistorical.html>

[2] World Health Organization (2009), WHO Guide to Identifying the Economic Consequences of Disease and Injury. Retrieved from <https://www.who.int/choice/publications/d_economic_impact_guide.pdf?ua=1>.

[3] National Heart, Lung, and Blood Institute. Assessing Your Weight and Health Risk. Retrieved from <https://www.nhlbi.nih.gov/health/educational/lose_wt/risk.htm>

[4] United States Department of Labor, Bureau of Labor Statistics(June 28, 2018). American Time Use Survey. Retrieved from <https://www.bls.gov/tus/overview.htm#1>.

[5] United States Department of Labor, Bureau of Labor Statistics(July, 2017). American Time Use Survey (ATUS) Data Dictionary: 2014-16 Eating & Health Module Data Variables collected in ATUS Eating and Health Module July 2017.. Retrieved from <https://www.bls.gov/tus/ehmintcodebk1416.pdf>.

[6] United States Department of Labor, Bureau of Labor Statistics(December, 2017). American Time Use Survey Eating & Health Module 2014-16 Questionnaire December 2017. Retrieved from <https://www.bls.gov/tus/ehmquestionnaire1416.pdf> .

[7] United States Department of Agriculture, Economic Research Service (May 2016). 2014-16 Eating & Health Module User’s Guide (2016 Edition). Retrieved from <https://www.ers.usda.gov/publications/pub-details/?pubid=42817> .

[8] United States Bureau of Labor Statistics (2014). Eating & Health Module Dataset, American Time Use Survey (ATUS) Eating & Health Module Files from 2014. Retrieved from <https://www.kaggle.com/bls/eating-health-module-dataset#ehresp_2014.csv> .

[9] Margaret Gough, Adam M. Lippert and Molly A. Martin. 2019. The Role of Time Use Behaviors in the Risk of Obesity among Low-Income Mothers. In Women's Health Issues 29-1 (2019) 23-30. <https://www.whijournal.com/article/S1049-3867(17)30727-2/pdf>.

[10] Jennifer Bennett Shinall. 2016. Occupational Characteristics and the Obesity Wage Penalty. In Social Science Research Network. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2379575>.

[11] Karen S. Hamrick and Ket McClelland. 2016. Americans’ Eating Patterns and Time Spent on Food: The 2014 Eating & Health Module Data. In Economic Information Bulletin Number 158. <https://ageconsearch.umn.edu/record/262141/files/60064_eib158.pdf>.

[12] Karen S. Hamrick and Margaret Andrews. 2016. SNAP Participants’ Eating Patterns over the Benefit Month: A Time Use Perspective. In PLoS ONE 11(7). <https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0158422&type=printable>.

[13] Karen S. Hamrick, Margaret Andrews, Joanne Guthrie, David Hopkins and Ket McClelland. 2011. How Much Time Do Americans Spend on Food? Economic Information Bulletin Number 86. <https://foodtechconnect.com/wp-content/uploads/2011/11/WhereEat.pdf>.

[14] ACM Digital Library . <https://dl.acm.org/ccs/ccs.cfm?id=0&lid=0>.