Kari Palmier
CSC 478 –Final Project

National Health and Nutrition Examination Survey Data Analysis Project

Introduction

- I was interested in understanding the different factors that contribute to a person’s health
  - Demographic, medical conditions, food purchasing and eating habits
- The dataset I chose was the Kaggle version of the National Health and Nutrition Examination Survey dataset from California between 2013 and 2014.
  - Kaggle had multiple csv files for demographic, questionnaire, medication, and examination data
  - Kaggle provided links to codebooks with variable meanings
- I decided to investigate what contributes to a person’s weight, specifically being obese through predictive models
- I also performed clustering to see what trends are present among the variables chosen

Data Preparation

Data Preparation

- Merged together variables from the Kaggle demographics and questionnaire csv files
  - Used the common SEQN in both files to perform merge
- Several variables contained Refused and Don’t Know survey answers
  - Removed all rows containing these values
- Two variables about the number of certain meals people ate contained one value that represented over 21
  - Removed rows with these entries because there is no way to know the actual values
- Several categorical variables contained NaN values
  - Removed all rows with categorical NaN (cannot replace these)
- Number of fast food meals had NaN values for 22% of its entries
  - Removed these rows because mean may not correctly represent data

Data Preparation

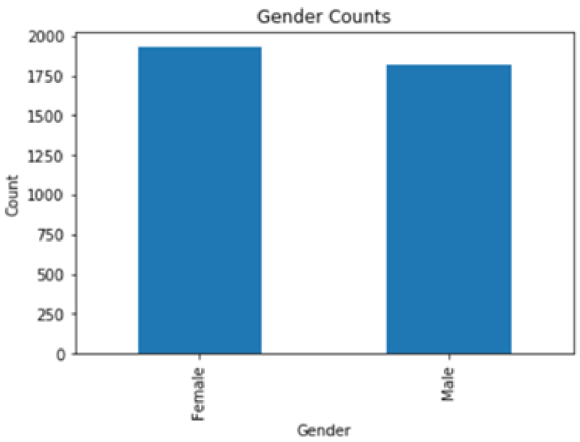
- Replaced all remaining NaN values in continuous values with mean of the variable
- Removed two income range categories that overlapped the other brackets
  - Were two ranges for less than $20,000 and one for over $20,000
  - Examples of other ranges were $15,000 to $20,000 and $35, 000 to $45,000
- Created a BMI continuous variable from existing weight and height
  - Converted weight to kg and height to m
  - BMI = weight (kg) / (height (m))^2
  - Was used as the target variable for linear regression
- Used BMI to create an obesity indicator variable
  - Values were 1 for obese entries (BMI >= 30)
  - Values were 0 for not obese entries (BMI < 30)
  - Was used as class target variable in classification

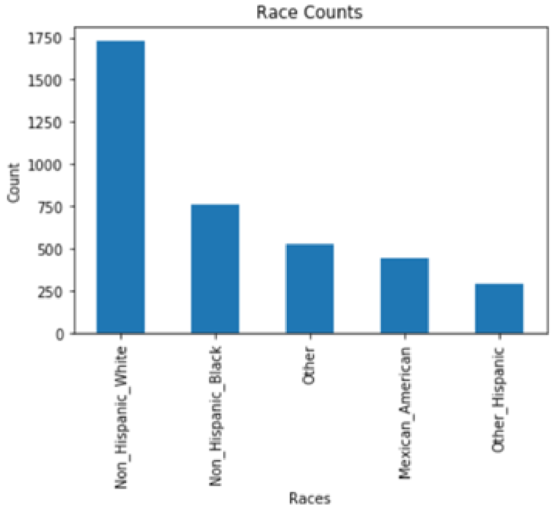
Exploratory Analysis

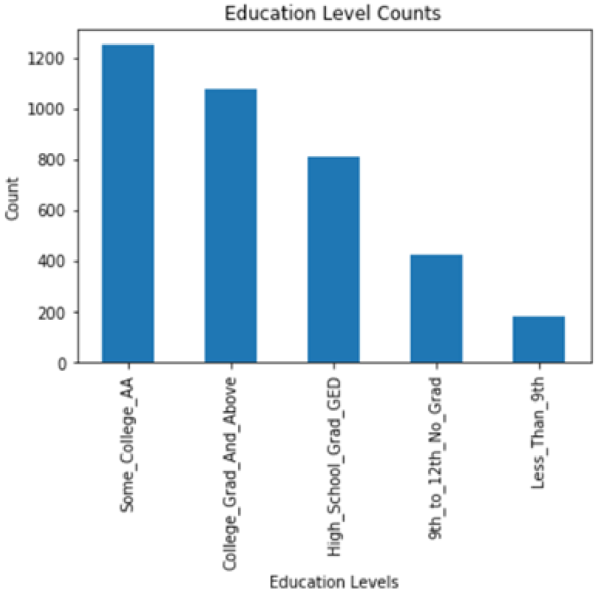
Exploratory Analysis

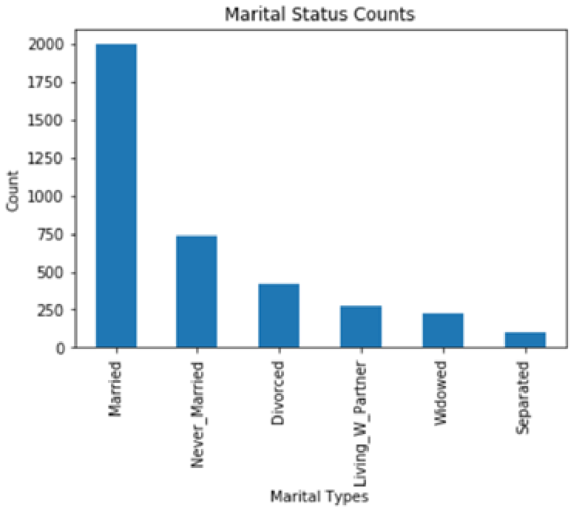
- Income bracket of $100,000 and above had a much higher number of respondents than the other income brackets
- Significantly more respondents were non-Hispanic white than other races
- A much larger number of respondents were married than the other marital statuses
- More respondents had some college education or an associate degree
  - Second highest education level was college graduate and above
- Distribution of men and women was almost equal (slightly more women)
- Age is evenly distributed until 65, then dips down until 80
  - Spikes at 80 because anyone over 80 was counted as 80
- BMI calculated was skewed toward lower BMI values
- The obesity indicator was significantly imbalanced with many more not obese entries (96.7% was not obese)

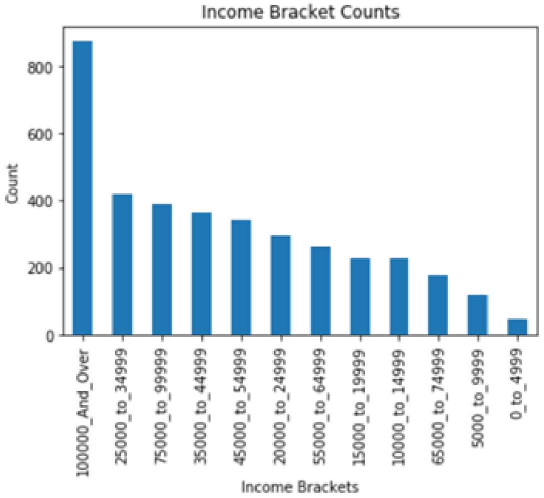
Exploratory Analysis (cont.)

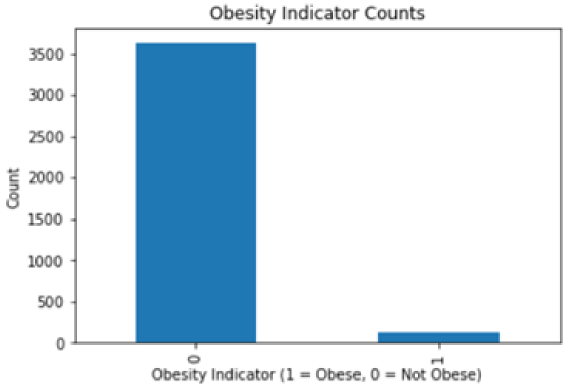




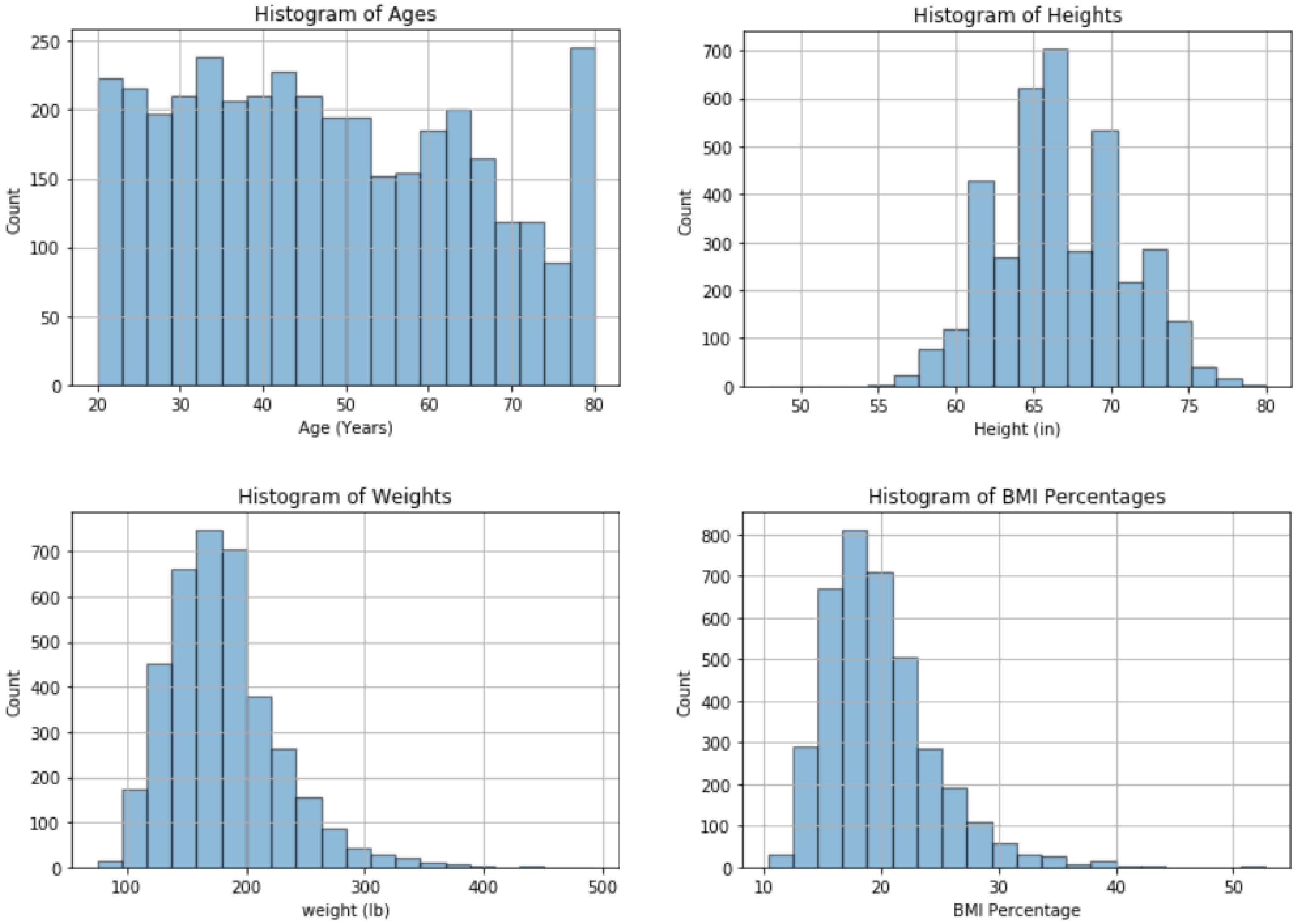




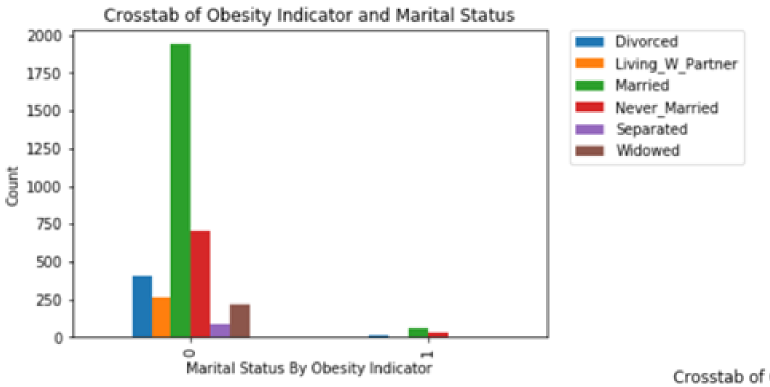


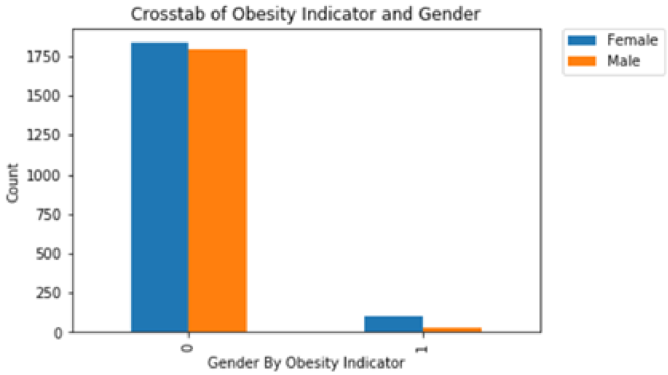


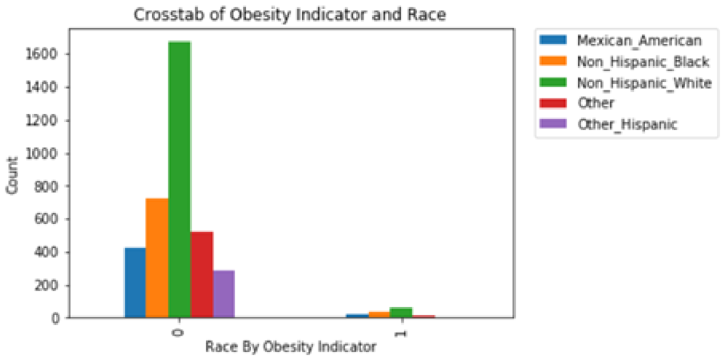
Exploratory Analysis (cont.)

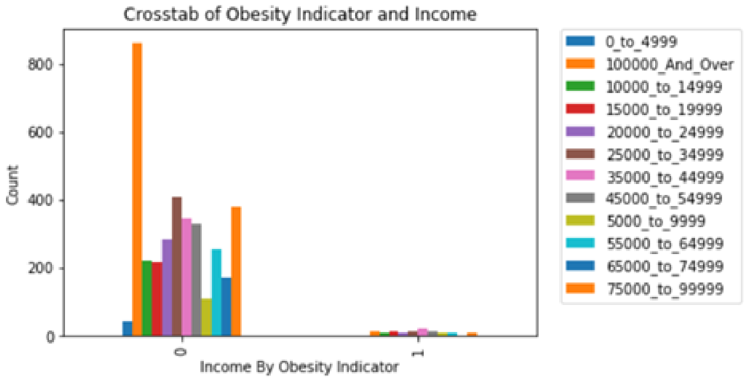


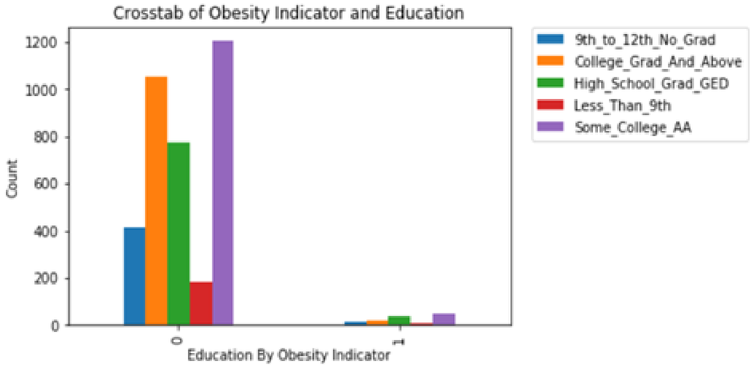
Exploratory Analysis (cont.)











Classification

Classification Procedure

- Initial models created
  - Decision tree with equal class weights, k nearest neighbor, naïve bayes Gaussian, naïve bayes multinomial, and linear discriminant analysis
  - Used sklearn functions for all modelling
- Class variable used was the obesity indicator (1 for BMI >= 30, 0 for BMI < 30)
- Models created to attempt to handle class imbalance
  - Decision tree with balanced class weights, random forest ensemble, and adaboost ensemble
- Created training and target datasets
  - Training had target removed, as well as weight, height, and BMI
- Split training and target datasets into 80% training, 20% testing randomly
- Created normalized split training and testing datasets for KNN
- Feature selection was performed in two-stages for all initial models (except LDA) and the decision tree with balanced class weights
  - First stage is to find the optimal percent of features based on max accuracy
  - Second stage is to eliminate any of the selected features with p values over 0.05

Classification Procedure

- Performed model selection for decision tree, k nearest neighbor, and ensemble models to find optimum model parameters
  - Used sklearn grid search algorithm for model selection
- Performed cross validation on training data of model with reduced features and model parameters
  - Used cross validation training and testing accuracies to alter model parameters
- Created full final training model and used to predict testing data
  - Compared the final training and testing accuracy to determine the final model performance (overfit, underfit, etc.)
- Created all the initial models except LDA with and without feature selection
  - LDA was only created with all features
- Created all initial models with the original dataset (no height and weight)
- Created all initial models with height and weight included in dataset
  - Investigate effect of height and weight on performance of classification

Classification Results

- All classification models using dataset without height and weight did not predict the test cases properly
  - Initial models all did not properly predict obese cases due to class imbalance during cross validation, final full train modelling, and test set prediction
  - Decision tree with balanced class weights models classified obese cases well, but misclassified 17% of not obese cases as obese
  - Ensemble random forest and adaboost models properly classified cross validation and full training data, but misclassified all obese cases in testing data
    - The ensemble methods overfit to the training data
- Adding height and weight into the dataset improved the initial model decision tree performance
  - No significant improvement in other initial models
- Decision tree models with height and weight in the dataset had uninteresting results
  - Decision tree model with height and weight using feature selection only used weight and gender for all nodes (most were weight)
  - Decision tree model with height and weight using all features only used weight for all nodes

Clustering

Clustering Process

- Used the k-means and cluster performance functions from sklearn
- Clustering was performed with the entire training dataset (before it was split into testing and training)
- Performed clustering with unnormalized and normalized training data
- Calculated the optimal k value by finding the sum of squared errors for each cluster model for different k values, then finding the knee of the sum of squared error versus k plot
  - unnormalized data k = 8, normalized data k = 12
- Performed clustering with k = 2, then used the obesity indicator class variable to measure the completeness and homogeneity
- Performed clustering of 80% training data, then used cluster assignments to predict test case classes
- Performed PCA on unsplit normalized training
  - Selected 38 components that captured 90% of total variance
- Found optimal k of PCA components to be k = 11
- Performed k = 2 clustering of PCA to find completeness and homogeneity

Clustering Results

- The sum of squared error was significantly less using normalized data versus unnormalized data for all clustering cases
- The sum of squared error was not significantly different between the normalized data and PCA component clustering
- The homogeneity and completeness scores of all clustering were very low
  - Unnormalized scores were worse (0.032% completeness, 0.19% homogeneity)
  - Normalized and PCA scores were similar
    - Normalized completeness = 3.02%, homogeneity = 13.2%
    - PCA completeness = 1.18%, homogeneity = 18.6%
  - Poor scoring is due to only using k = 2 (all optimal k values were 8 and above)
- The prediction accuracy of the 20% testing data was 41.15% for unnormalized and 31.56% for normalized
- Performed clustering with a dataset containing weight and height
  - Optimal k, completeness, and homogeneity scores did not change
  - Prediction accuracy of 20% testing data improved to 60.72% for unnormalized and 68.7% for normalized

Clustering Results

- One cluster was healthy women (no weight issues, no diseases, no high blood pressure, no diabetes, etc.) who had smoked over 100 cigarettes in their life
- Two clusters (one men and one women) contained all healthy variables (no weight issues, no diseases, etc.) with non-Hispanic white race, high income to poverty ratio, college graduate or above, married, and income of $100,000 or above
  - Implies these other variables may be related to healthy people of both genders
- One cluster of women with high age, high income to poverty ratio, race of non-Hispanic white, and married that had weight issues (doctor had told them they were overweight, needed to lose weight, and needed exercise), high blood pressure, and thyroid issues
  - Weight issues could be related to the high age, blood pressure and/or thyroid issues since the rest of the variables were similar to healthy women.
- One cluster of men with high age, high income to poverty ratio, race of non-Hispanic white, and married that had weight issues (doctor had told them they were overweight, needed to lose weight, and needed exercise), and high blood pressure
  - May mean weight issues in men were related to age and/or high blood pressure

Linear Regression

Regression Procedure

- Initial models created
  - Standard linear regression, ridge regression, and lasso regression
  - Used sklearn functions for all modelling
- Target used was the continuous BMI variable
- Created training and target datasets
  - Training had target removed, as well as weight, height, and obesity indicator
- Split training and target datasets into 80% training, 20% testing randomly
- Created normalized split training and testing datasets
- Feature selection was performed in two-stages for all models
  - First stage is to find the optimal percent of features based on min RMSE
  - Second stage is to eliminate any of the selected features with p values over 0.05
- Performed model selection for ridge and lasso models to find optimum model parameters
  - Used sklearn grid search algorithm for model selection

Regression Procedure

- Performed cross validation on training data of model with reduced features and model parameters
  - Used cross validation training and testing RMSE to alter model parameters
- Created full final training model and used to predict testing data
  - Compared the final training and testing RMSE to determine the final model performance (overfit, underfit, etc.)
- Created all the models with and without normalized training data
- Created all models with the original dataset (no height and weight)
- Created all models with height and weight included in dataset
  - Investigate effect of height and weight on performance of regression

Regression Results

- All regression models using dataset without height and weight had most RMSE values were between 12% and 13.5% (full training was 3.5%)
  - Same for cross validation and test data prediction (full training was 3.5%)
  - Same for normalized and unnormalized data sets
- All regression models using dataset without height and weight had low R squared values (0.41)
  - Approx. 60% of variance of BMI was not captured by model variables
- Lasso regression did not follow an acceptable pattern for RMSE as the percentage of features increased
  - RMSE should decrease as percentage of features increases
  - RMSE had a non-decreasing step pattern for unnormalized data and was constant for normalized
- Ridge alpha found through model selection was 4.996
- All regression models using dataset with height and weight had RMSE values were between 1.5% and 3%
- All regression models using dataset with height and weight had high R squared values (0.87)

Regression Results

- For coefficient analysis, looked at coefficient > 0.3 and < -0.3
- Positive predictive coefficients for regression with dataset without height and weight:
  - non-Hispanic black, Mexican-American, no doctor visits in the past year, 16 or more doctor visits in the past year
- Negative predictive coefficients for regression with dataset without height and weight:
  - other race, doctor not they're overweight, doctor not saying they needs to lose weight, income of $100,000 and above, no asthma, no high blood pressure
- Positive predictive coefficients for regression with dataset with height and weight:
  - female, Mexican-American, other race, person been told they are overweight by a doctor, no doctor visits (normalized ridge also had very high weight coefficient)
- Negative predictive coefficients for regression with dataset with height and weight:
  - male, some college/associate, college graduate and above, and doctor not saying they're overweight

Conclusions

- All classification models had issues due to the high class imbalance of the obesity indicator class variable (96.7% of cases were not obese)
  - Trying class weighting and ensemble methods did not solve the issue
  - Class imbalance is most likely too extreme for oversampling the obese or under sampling the not obese
  - Could try to create a new, more balanced class of people overweight and above (BMI > 25) and possibly even another for underweight (BMI < 18.5) instead
- Decision tree classification improved with adding height and weight
  - Results were not interesting since main variable used was weight
- Clustering performed best on normalized data
  - PCA did not improve clustering
  - Trying to score clustering did not work because K = 2 was too low
- Regression resulted in low R squared without weight and height
- Adding weight and height increased R squared
  - Target BMI was calculated from weight, so results may be uninteresting