Capstone Project 2 - Milestone Report
Predicting Health using the National Health and Nutrition Survey

Problem Statement

There are many factors that impact the health and wellness of human society. It should be possible to reduce the occurrence of particular health concerns if those factors with high impact can be identified. The goal of this project is to use statistical inference and machine learning to explore if predictions in certain illnesses can be made related to habits, nutrition, BMI, and bloodwork.

Proposed Solution

Perform exploratory analysis and predictive modeling from the National Health and Nutrition Survey (NHANES) dataset that assesses the health and nutritional status of people in the United States.

- Accurate predictive models can help people in general to benefit if looking for ways to better their life.
- It can extend to professionals in the medical, nutritional, and fitness arenas to discuss the potential effects of life choices relating to health conditions.
- This could lead to people wanting to research particular findings in more depth in order to better their lives. Potentially adding longevity and making a positive difference for our population at large.

Datasets

A collection of datasets from the National Health and Nutrition Survey (NHANES) was obtained from Kaggle. https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey.
This information was used in the project to study variables that could potentially help predict health conditions and improve human lives.
More information on the NHANES survey can be found on the Center for Disease Control and Prevention's website https://www.cdc.gov/Nchs/Nhanes/about_nhanes.htm.

CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

Wrangling Steps Performed:

Data, Demographics, and Labs

1. Three csv files were downloaded and read into a normalized pandas dataframe. They were merged together to create one dataframe.

2. Columns were explored to determine which would be useful on this project. Those identified for use were then renamed from their initial codes to identifiable word strings.
3. A feature generation for BMI at age 25 was created using the height and weight found in the data.
4. The data was checked and scanned for any null information present. This was later used when looking at categories for those having asthma and those having arthritis. Null values had their entire rows removed when those columns were used for explorations.

Dealing with Outliers