Final Project Checkpoint 1

1. GitHub Repository

   https://github.com/heavens-potato/320-final-project

2. Dataset

   We are using the "Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System" dataset from the Centers for Disease Control and Prevention (CDC). The dataset consists of surveys of participants on several factors, such as their fruit/vegetable consumption, frequency of intensive physical activity, overweight classification, etc. It also contains other categorical data like stratification categories, as well as numerical data like sample size and confidence levels. The dataset was published by the CDC on July 21, 2016, and has been most recently updated on September 12, 2025, as of September 25, 2025.

   APA Citation:

   Centers for Disease Control and Prevention. (2016). *Nutrition, physical activity, and obesity - behavioral risk factor surveillance system*. Centers for Disease Control and Prevention. Updated September 25, 2025. https://data.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7/about_data

3. Why this dataset

   We chose this data set because we want to study the risk factors of obesity with a large-scale, reliable dataset. We want to see if an ML model will be able to show correlations between certain factors and obesity. This issue pertains to us because of the negative health effects obesity has on one's life, and we would like more insights on what habits contribute to healthier lives. This dataset also provides a lot of information with 106,000 rows and 33 columns of data, which gives us a lot to work with to analyze and make important conclusions about the topic.