Research Proposal

Janet Zhuang
Ange Olson
Nick Benevento

1. **Research Topic:** We have chosen to study indicators and prevalence of heart disease in the United States. Lifestyle factors (exercise, mental health, drinking/smoking habits), socioeconomic factors, and other existing health conditions all may be either a) causally linked or b) correlated with heart disease in an individual. The ability to predict whether or not an individual is at risk of heart disease can enable physicians and individuals to take early preventative action to reduce deaths as a result of heart attacks, stroke, and related diseases.
2. **SMART Question(s):** The CDC notes that three key risk factors for heart disease are smoking, high blood pressure, and high cholesterol–however, roughly half of Americans fall into at least one of these categories. Given this, are these three risk factors good predictors for heart disease, or are there factors that may be less prevalent and give more of an indication of risk of disease?
3. **Dataset:**
   - The link to our dataset is here
   - Originally, the data came from the CDC Behavioral Risk Factor Surveillance System (BRFSS). The data was cleaned by the author on Kaggle Kamil Pytlak.
   - The dataset consists of 18 columns and over 300,000 observations.
4. **GitHub:** https://github.com/NickBenevento/DATS_6103_Project
5. **Proposed Modeling Methods:** We plan to use a logistic regression model to predict the probability that an individual will be either at risk of heart disease (1) or not at risk (0). As an additional method, we will use a support vector machine (SVM) with a linear kernel in order to compare and contrast our results.