Business Project by Alexandria Zeng

Abstract: The client for my project is the CDC. The goal of this project was to collect information about risk factors for diabetes to support a new public health campaign. Diabetes is a top public health issue: it is the 7th leading cause of death in the US. I worked with the diabetes health indicators dataset from Kaggle. I found that being overweight is a significant contributor to diabetes, cholesterol, and heart disease. This information will be used to create a classification model to target areas in the US with the greatest prevalence of diabetes.

Design: Based on the CDC's website, over 37 million US adults have diabetes (11.3% of the population). Diabetes is the 7th leading cause of death in the US and in the last 20 years the number of adults diagnosed with diabetes has doubled. The goal is to collect data about risk factors of diabetes and use it to create a classification model. For the business side, recommendations will be provided to plan public health funding around these high risk factors.

Data: I used the diabetes health indicators dataset that was cleaned by a user on Kaggle. On Kaggle there are three versions of the data. I used the balanced version that had a 50/50 split between individuals with and without diabetes. There were ~70,000 rows and 21 variables. Survey responses and the features fall roughly into three categories: demographics, lifestyle, and other health issues (heart disease, cholesterol, etc.)

Algorithms: Most of the variables were binary so I created calculated fields for each of those variables.

Tools:

- Excel for data cleaning and Tableau for visualizations

Communication:

- I plan to clean up my code and upload it to my Github.