The first step for analysis would be to build out an analytical base table, ABT, of the variables included in the dataset. The ABTs can be split into continuous features vs. categorical features.  ABTs include mean, median, mode, and percentage of missing values. The ABTs assist in developing a plan on how to handle data quality issues.

For the NHANEs dataset, it was apparent that some features had a large number of missing values. Features that had over 30% of values missing were removed from the dataset. Once the features with a high % of missing values were removed, 38 features remained. On the remaining features, imputation was used for the missing values. Categorial variables used mode imputation, whereas continuous features used mean imputation.

Matplotlib was used for data exploration. Histograms were used to visualize the distribution of continuous features. Box plots were used to see if there were relationships between certain continuous and categorical features.

Finally, a decision tree classification method was used for making predications if a person would have diabetes. Along with predictions, the decision tree method was used to develop a features importance table. Lastly, the predictions were visualized using a confusion matrix.