DIABETES EDA

3/2/23

SUMMARY

The dataset that was used for exploratory data analysis (EDA) was procured from Kaggle utilizing data collected by the National Institute of Diabetes and Digestive and Kidney Diseases. The question being explored was to determine what health related measurements may predict an outcome of a diabetes diagnosis.

Analysis was conducted with the following variables: age, glucose, insulin, body mass index (BMI), number of pregnancies, diabetes pedigree function, and the diabetes outcome. Potential outliers were found with the patient records according to the recorded BMI and number of pregnancies. These would need to be removed for further analysis. When determining potential predictors of the diabetes outcome, it was found that glucose, diabetes function pedigree, BMI, and age may be statistically significant.

IMPROVEMENTS

The dataset may not include enough statistically significant variables to accurately predict a diabetes outcome. Additional variables should be collected and analyzed to answer the question stated. Conducting additional tests outside of the requirements would have also assisted in understanding the relationships between the variables.

Some of the variables that might have an impact on a diabetes diagnosis that were not evaluated are lifestyle factors (exercise, stress, etc.), hormonal diseases, pancreas health, and medications taken. The dataset was also limited to female patients at least 21 years old of Pima Indian heritage. There should be a more varied patient demographic for further analysis.

ASSUMPTIONS

At the beginning of the project, one of the assumptions that I made was that genetic risk would be one of the most significant factors in a diabetes outcome; however, after conducting testing it showed some correlation but other factors such as glucose and BMI had a larger impact.

CHALLENGES

One of the biggest challenges in the project I had was determining the correlation between the variables. I tried several different combinations and chose to document the ones that had the highest correlation, and I still found a negligible correlation between the variables. When conducting the regression analysis, it did impact the performance on what provided the best outcome. This may have been solved with additional variables for consideration.