The impact of diabetes on overall health

By: Jedediah Brough

As I began the EDA of the dataset, the first thing that I noticed was that there the dataset needed to be cleaned. There were no missing values or partial data, but it was very obvious that there were a lot of outliers that needed to be handled. There were several observations with values that didn’t make any sense. Some observations had 0 blood pressure, others had a BMI of 0, and there were other extreme values. I did some research into these and found that a blood pressure of 0 would mean that your heart had stopped. A healthy range for BMI is between 18 and 25, and the lowest I could find was around 12. Most of the data was normally distributed with one mode or exponentially distributed.

As I worked through the analysis, I felt that I wasn’t able to look much into the outcome variable that indicated diabetes in individuals. I was able to use the outcome variable as a sort of filter but I wasn’t able to Outcome being a discrete variable, I wasn’t able to create a histogram or any other graph to view that data.

I would have enjoyed if the Outcome variable distinguished type 1 vs type 2 diabetes. Type 2 diabetes is much more manageable through lifestyle adjustments like change in diet and an increase in activity level. Type 1 is more difficult to manage and is strictly treated through insulin injections. Variables like BMI and blood pressure can be more indicative of lifestyle and are more likely to be related to type 2 diabetes. Differentiating between type 1 and type 2 diabetes would have made the analysis on BMI and blood pressure more meaningful.

As stated, the assumption was made that BMI and blood pressure are strongly connected to the diabetes outcome. While that is not incorrect, the assumption that they relate equally to both types of diabetes is incorrect.

I did not fully understand the diabetes pedigree function. I understood that it related to a family history of diabetes but I do not know how it was calculated or how indicative it was. Because of this, I really did not use the pedigree function much during the overall analysis.