Exploratory Data Analysis of Diabetes and Pregnancy

Samantha Traversi
MDS 546

Hypothesis

There is a correlation between the number of times a woman has been pregnant and the severity of the health factors linked to diabetes. The measured health factors include the level of plasma glucose concentration two hours after an oral glucose test, insulin concentration two hours after a serum insulin test, diastolic blood pressure, tricep skinfold thickness, and body mass index.

Data Sources

The information in the dataset used in this analysis was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases in a study of the Pima Indian community. The purpose of the dataset was to take diagnostic measurements and make diagnostic predictions on whether or not a patient has diabetes. Since the original publication, this data has become public domain and for this analysis it was sourced from Kaggle.

Diabetes Definitions and Background Information

Diabetes - The data in this set does not distinguish between Type 1 Diabetes (an autoimmune disease that prevents the body from making its own insulin), Type 2 Diabetes (a condition where the body is unable to use insulin properly, typically caused by poor diet and lifestyle choices), and Gestational Diabetes (high blood sugar levels during pregnancy). Instead of making distinctions between the three types of diabetes, this data exploration aims to determine whether or not the number of pregnancies a woman has had has an effect on the severity of her diabetes. Before delving into the data, it is important to define a few terms and highlight how these conditions relate to a woman's diabetic condition.

Plasma Glucose concentration at two hours in an oral glucose tolerance test - Glucose a sugar found in the blood that serves as the body's main source of energy. An oral glucose test is performed when the patient drinks a glucose-rich beverage after eight to twelve hours of fasting, and their blood is tested two hours after consuming the glucose drink to determine whether the body is capable of metabolising the sugar. Once the glucose drink has been consumed, the plasma glucose levels should rise and then insulin should begin moving the glucose into the cells of the body. The longer it takes for the plasma glucose levels to return to normal, the stronger the indicator that the body is not properly utilizing insulin. Test results below 140 milligrams per deciliter typically indicate normal blood sugar levels, 140 to 199 milligrams per

deciliter typically indicate impaired glucose tolerance, and 200 milligrams per deciliter or higher typically indicate diabetes.

2-Hour serum insulin (mu U/ml) - A two hour serum insulin test is used to measure the level of insulin in the blood two hours after administering glucose. A normal range of serum insulin is between 16 and 166 mu U/ml. Abnormally low insulin levels typically indicate Type 1 Diabetes, while abnormally high insulin levels typically indicate Type 2 Diabetes.

Diabetes Pedigree Function - this measurement evaluates the genetic relationships to the subject and provides a "provides a synthesis of the diabetes mellitus history in relatives and the genetic relationship of those relatives to the subject." Since my hypothesis focused on pregnancy and the correlation between the number of pregnancies and the other health metrics, I disregarded this column.

Diastolic Blood Pressure (mm Hg) - Diastolic blood pressure measures the force of the blood pushing against the artery walls when the heart is at rest and filling with blood. Individuals with higher plasma glucose levels tend to have higher blood pressure as the glucose affects the mobility of the blood within the arteries. It is recommended to keep diastolic blood pressure below 80 millimeters of mercury (mm Hg)

Triceps Skin Fold Thickness (mm) - Using calipers, the skin on the back of the left arm is pinched and lifted, then the width of the skin fold is measured in millimeters. The larger the value of skin fold thickness, the higher the levels of fat distribution in the limbs. The standard normal value for a woman of about 30% body fat is 18 mm.

Body Mass Index (weight in kg/(height in m)^2) - This measurement evaluates a person's weight in relation to their height, and is used to determine whether a person is underweight, of a healthy weight, overweight, or obese. According to the Center for Disease Control, a healthy BMI is between 18.5 and 24.9.

Age (years) - This is the age in years of the individual. In this particular data set, all subjects are between 21 and 81 years old. Since my focus was on pregnancy and health factors related to diabetes, not age, I disregarded this column in the analysis; however, the range of this column should be considered for an overall understanding of the subjects.

Class variable - This column is a binary measurement. Individuals without diabetes are measured as 0, and those with diabetes are measured as 1.

Initial Data Evaluation

This particular dataset, derived from the original, larger dataset, was originally a .csv file with nine columns and 768 rows (one subject per row). Each of the columns was dedicated to values medically linked to diabetes.

There are a few constraints to this data. First of all, the subjects are entirely of Pima Indian heritage, so future tests need to be conducted to determine whether this analysis applies to the global population, or if there are genetic, cultural, or environmental factors that affect this particular community.

Second, this dataset was selected from a much larger database. In this analysis, we are assuming that the original data was sampled randomly from the Pima Indian community. Equally importantly, we are also assuming that the subjects in this dataset have been sampled randomly from the original dataset.

Upon initial review of the data, I decided to explore the potential correlation of the number of pregnancies on severity of health factors that have been medically linked to diabetes. The suggestion we're looking for is whether the number of pregnancies a woman has had exacerbates the gravity of diabetes or heightens the risk of conditions related to diabetes. We are unable to prove causation with this correlation, we are just trying to determine whether more studies should be conducted.

Since the focus was the number of pregnancies leading to increased levels of medical measurements, I decided to eliminate two columns from this dataset: Diabetes Pedigree Function and Age (years). The Diabetes Pedigree Function is a number created to measure genetic prevalence of diabetes around the subject, which can't be affected by the number of pregnancies. Age is also not affected by the number of pregnancies. Removing these two columns allowed me to have a sharper focus on the correlation between the number of pregnancies and the other health metrics.

There were also quite a few zero-values in the columns. For the Class Variable column, the zeros serve as the boolean value of "False" for the condition of diabetes in the subject, so those values needed to remain. For the Number of Pregnancies column, the zeros serve to show the subjects which had never been pregnant before, so those values needed to remain as well. For all other columns, the zero values appeared to indicate missing data. Instead of removing all subjects with null values, which would have removed a significant percentage of the subjects in this dataset, I chose to replace the null values with the mean of each row.

Finally, before analyzing the data within this set, I reviewed the information of diabetic and non-diabetic subjects together, then also grouped into the diabetic subset and the non-diabetic subset to make comparisons between the groups of data.

Statistical Methods

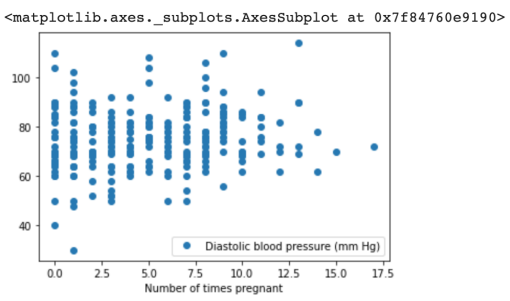
Table 1 (below) shows that though it appears that there may be a small difference between the number of times a woman has been pregnant and diabetes, the p value is actually significantly less than 0.05, so we can conclude that there is some relationship between diabetes and the number of times a woman has been pregnant, specifically if the number is above 4 pregnancies.

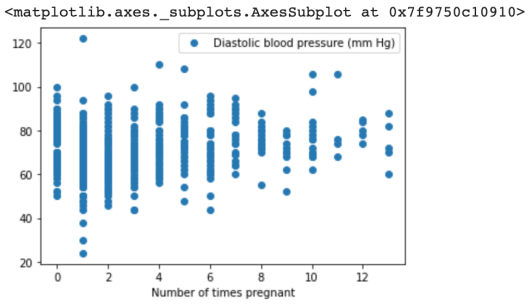
Table 1: T-test for the Differences Between Means of Number of Times Pregnant Among Diabetic and Non-Diabetic Subjects

	Diabetic Subjects	Non-Diabetic Subjects
Mean	4.866	3.298
Sample Standard Deviation	3.741	3.017
Number of Observations	268	500
Difference Between the Means	1.568	
df	766	
Standard Error	0.249	
t-Statistic	6.300	
p value	0.0001	

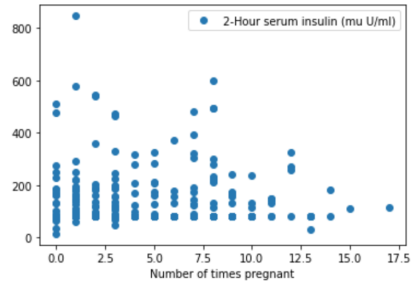
We also wanted to explore whether the number of pregnancies a woman has had is correlated with the risk level of each health factor linked to diabetes. Below are a series of plots with the diabetic subjects on the left and the non-diabetic subjects on the right.

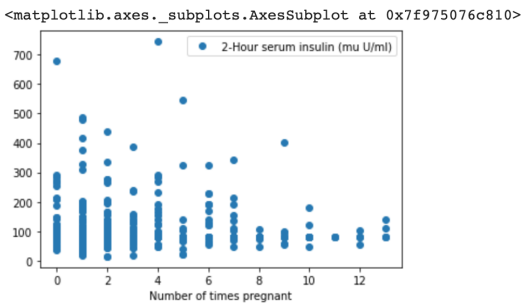
Number of Pregnancies and Diastolic Blood Pressure





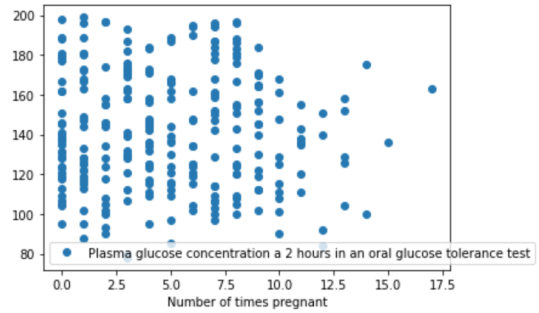
Number of Pregnancies and 2-Hour Serum Insulin Test



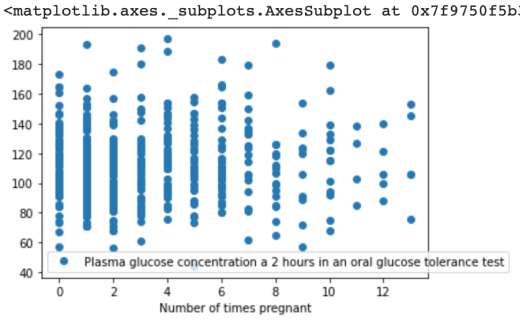


Number of Pregnancies and 2-Hour Plasma Glucose Oral Test

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f84760e8590>
```

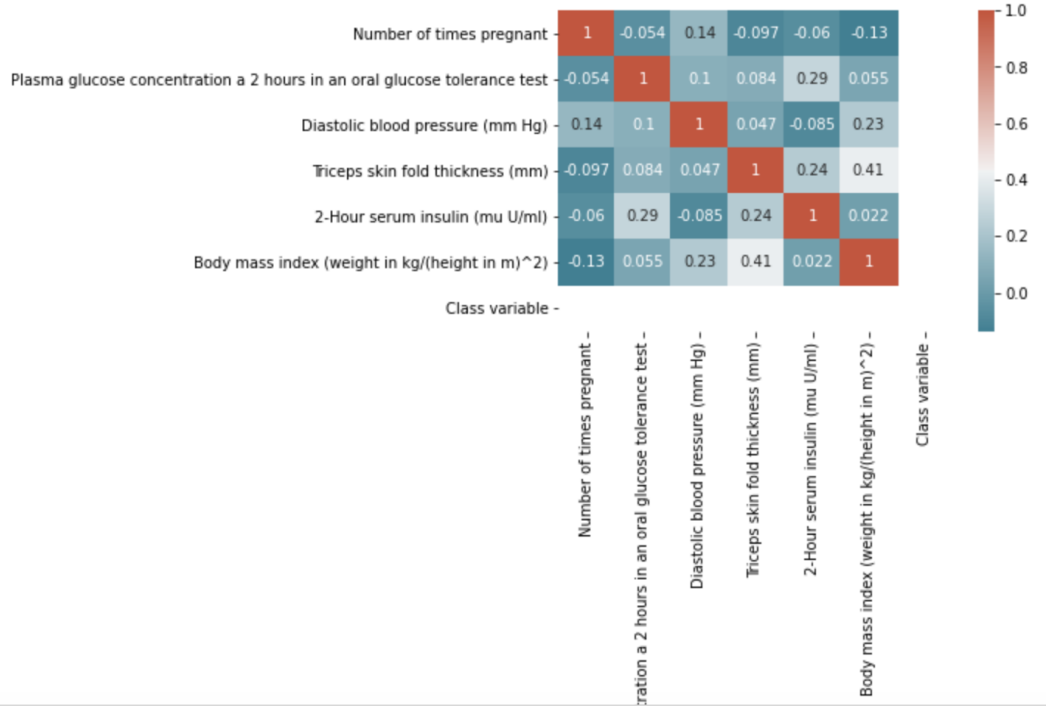



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9750f5b390>
```

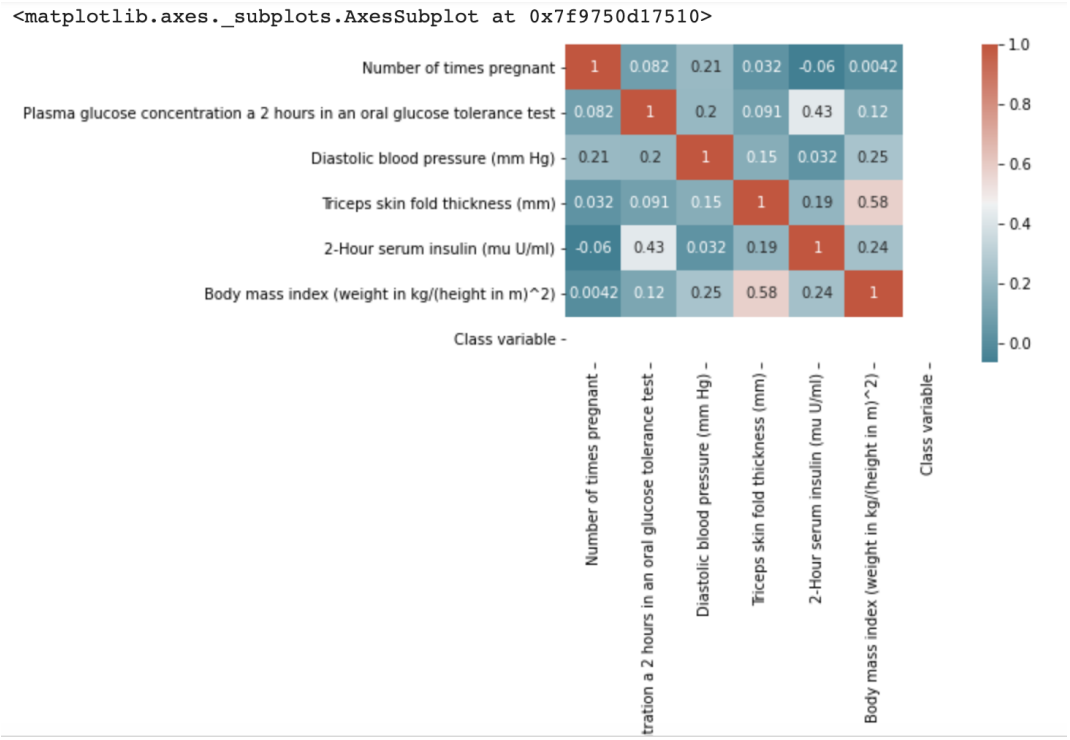


Heatmap: Number of Pregnancies and Health Measurements Linked to Diabetes (Diabetic)

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f84760e8d50>
```



Heatmap: Number of Pregnancies and Health Measurements Linked to Diabetes (Non-Diabetic)



Conclusion

In conclusion, the occurrence of diabetes and the number of pregnancies a woman has had appear to be correlated; however, more studies need to be performed in order to determine causality. There is not a strong enough relationship between the number of pregnancies and health factors related to diabetes in order to determine that with an increased number of pregnancies, health risks linked to diabetes also increase. The heatmaps do show correlations between the Plasma Glucose test and the Serum Insulin test; however both tests use a combination of fasting and glucose to measure quantity of glucose in the plasma and quantity of insulin in the blood, which correlate medically (insulin and how the body uses it has a direct influence on the level of glucose in the bloodstream).

Possibilities for Further Study

Further statistical analysis should be done with this dataset to include the Diabetes Pedigree Function and Age in Years. If blocking out the Diabetes Pedigree Function has an effect on the correlation between the number of pregnancies a woman has had and the health data linked to diabetes, it would be worth continued study to determine how genetic and familial connections to diabetes affects the other health factors of a subject, and how that increases over number of pregnancies.

Also, since Type 2 Diabetes develops over years of diet, exercise, and other health choices that a subject makes, the age should also be taken into further consideration in future studies. It is likely that there will be a difference in measured health conditions within different age ranges of subjects.

Furthermore, this dataset was originally developed to determine how health metrics can be used to diagnostically predict diabetes within subjects. The data in this set can be further analyzed to make health predictions and recommendations that could improve the lives of the individuals in the Pima Indian community.

Works Cited

National Institute of Diabetes and Digestive and Kidney Diseases. (2021). *Diabetes.csv*. Retrieved from https://www.kaggle.com/uciml/pima-indians-diabetes-database

Center for Disease Control. (2021). *Diabetes.* Retrieved from https://www.cdc.gov/diabetes/index.html

American Diabetes Association. (2021). *Diabetes Overview.* Retrieved from https://www.diabetes.org/diabetes

Editor. (2019, January 5). *Blood Sugar Level Ranges.* Retrieved from https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html

The British Diabetic Association. (2021). *Diabetes and Blood Pressure.* Retrieved from https://www.diabetes.org.uk/guide-to-diabetes/managing-your-diabetes/blood-pressure

U.S. National Library of Medicine. (2021). *Insulin in Blood.* Retrieved from https://medlineplus.gov/lab-tests/insulin-in-blood/#:~:text=This%20test%20measures%20the%20amount,body's%20main%20source%20of%20energy.

Shanker, M., Hu, M., Hung, M. (1999, November 13). *Estimating Probabilities of Diabetes Mellitus Using Neural Networks.* Retrieved from http://www.personal.kent.edu/~mshanker/personal/Zip_files/sar_2000.pdf