**INF2178 Mid-term Project**

Tong Wu

1. **Abstract**

Type 2 diabetes is an impairment of the way the body regulates and uses sugar (glucose) for fuel. This long-term (chronic) condition causes too much sugar to circulate in the blood.[[1]](#footnote-0) Ultimately, high blood sugar levels can lead to disturbances in the circulatory, nervous, and immune systems.[[2]](#footnote-1) Type 2 diabetes used to be called adult-onset diabetes, which data shows a strong age-related relationship - type 2 diabetes is more common in older adults. This project examines the impact of blood sugar and age on type 2 diabetes in a person's diagnosis of type 2 diabetes.

Type 2 diabetes, on the other hand, cannot be cured, but strategies like losing weight, eating healthy, and exercising can help you manage the disease.[[3]](#footnote-2) In this study, we used the BMI parameter to analyze whether a healthier lifestyle could reduce the chance of developing type 2 diabetes. The dataset "diabetes.csv" contains the outcome of type 2 diabetes as a binary variable to demonstrate whether each patient was diagnosed with diabete. We use a regression model and N way ANOVA to analyze the relationship between the independent variable and the dependent variable. In other words, in this study, we explored the relationship between "glucose" as a dependent variable and independent variables: "BMI" (lifestyle-related diabetes risk) and "Age" (age-related diabetes risk) ).

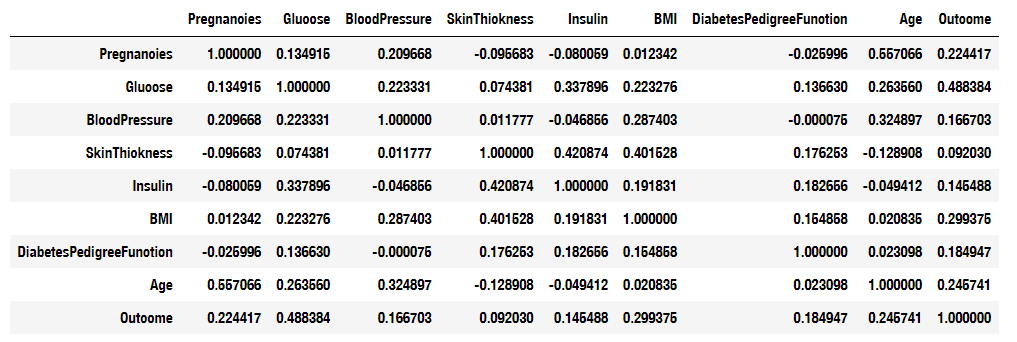
Correlation Analysis, ANOVA, and OLS regression analysis were used in this paper to examine the number of pregnancies, blood glucose (blood sugar level), blood pressure, skin thickness, insulin, BMI (body mass index), diabetes spectrum function (genetic/family history risk), and The influence of factors such as age on diabetes.

The original "diabetes.csv" dataset does not serve the above research question. Therefore, some transformation and extraction operations were taken to prepare a clean dataset suitable for exploring the interaction between the dependent variable "Glucose" and the set of predictors ("BMI", "Age").

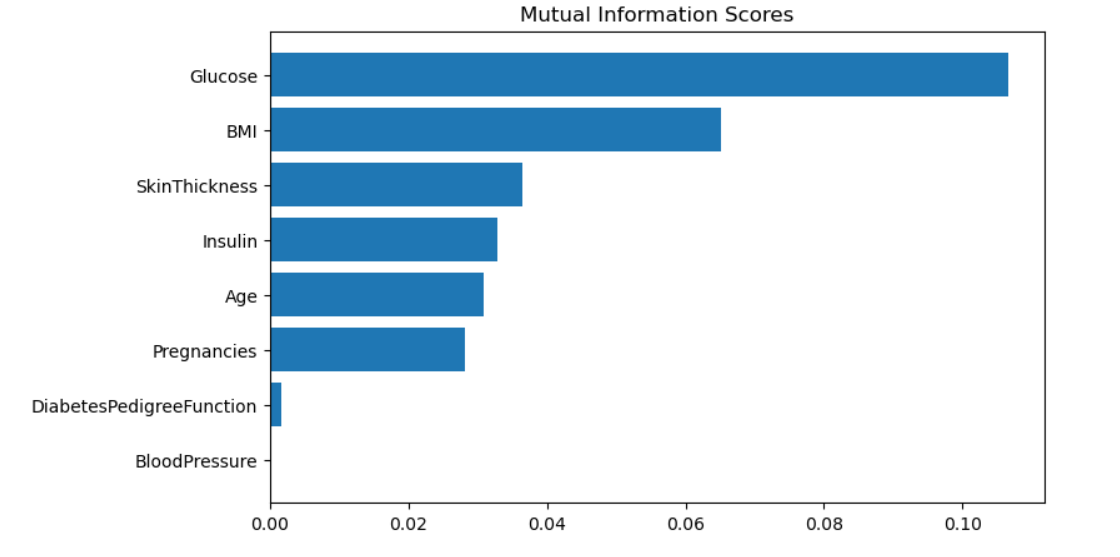
1. **Data Cleaning**

First, we cleaned the data. There are many 0 values in this dataset to represent invalid data. The minimum value of glucose, blood pressure, body mass index in the dataset is zero, which is impossible. So we count how many zeros exist, which means how many abnormal values in the dataset in each column.

For the Glucose, BloodPressure, BMI can remove rows with abnormal values, while insulin and SkinThickness are unusable. Since too many rows with zeros(abnormal values), imputation might not help, another idea is to create an indicator. Total count of records with abnormal values is 44. The original dataset has 768 rows, after dropping rows with zero in Glucose, BloodPressure, BMI, remaining 724 rows.

**Figure 1**: Pearson Correlation Analysis with the Raw Dataset.

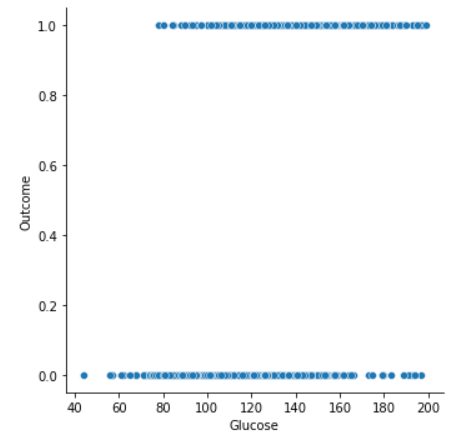
Then we will examine the correlation matrix. From the chart above, we could get the result that the Glucose has the highest correlation level with the outcome. The BMI and Age follows right after the Glucose. This high correlation is consistent with our hypothesis and lays the foundation for our further analysis. According to correlation matrix, select Glucose, BMI, Age as input variables, these variables are not quite correlated.



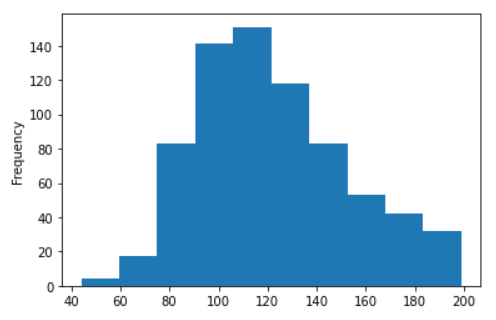
**Figure 2**: Mutual Information Score Chart with the Raw Dataset

Then we will cross examine functions to calculate mutual information score to measure the strength of association between words x and y. We can acknowledge from the chart that, Glucose, Age BMI are highly related to the outcome.

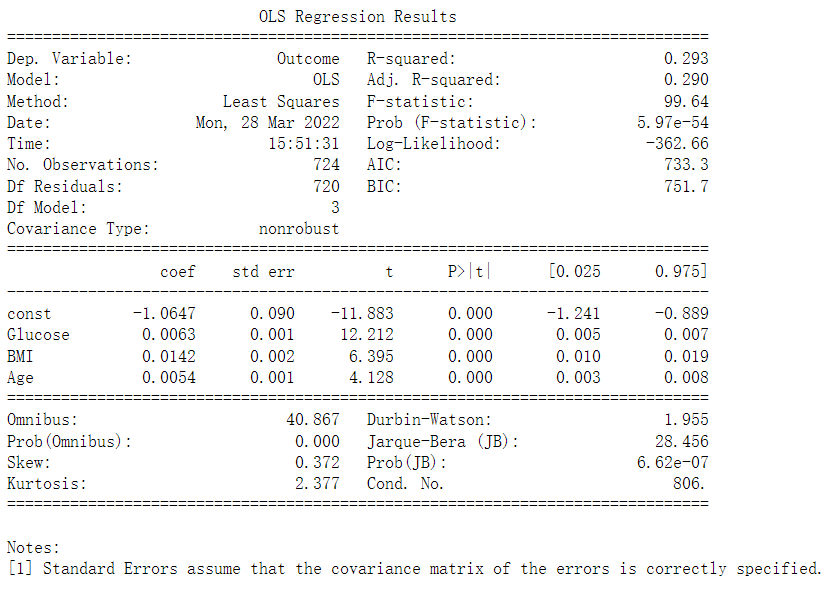
Then we print the scattered plot and histogram to visualize the data. There is no obvious linear pattern in the scattered plot, it doesn’t show a cause-and-effect relationship exists.



**Figure 3**: Scattered Plot



**Figure 4**: Histogram



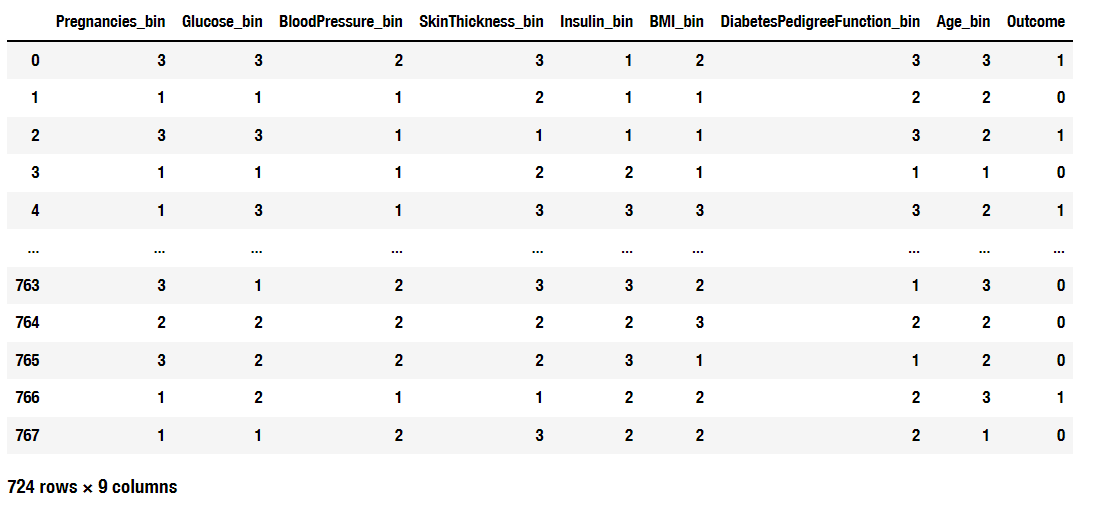
**Figure 5**: Pearson Correlation Analysis for All Factors Present with the Raw Dataset.

By conducting the OLS regression, we could get the conclusion that the model is valid with high F score, and F-statistic is much less than 5%. It’s enough to reject the null hypothesis:, which means it's a significant model. However, the adjusted R square is quite low, only 0.29. It means that only 29% of variance is explained by this model. As the raw data is continuous variables, it’s unable to conduct anova test, have to bin the input variables

1. **Categorical variables**

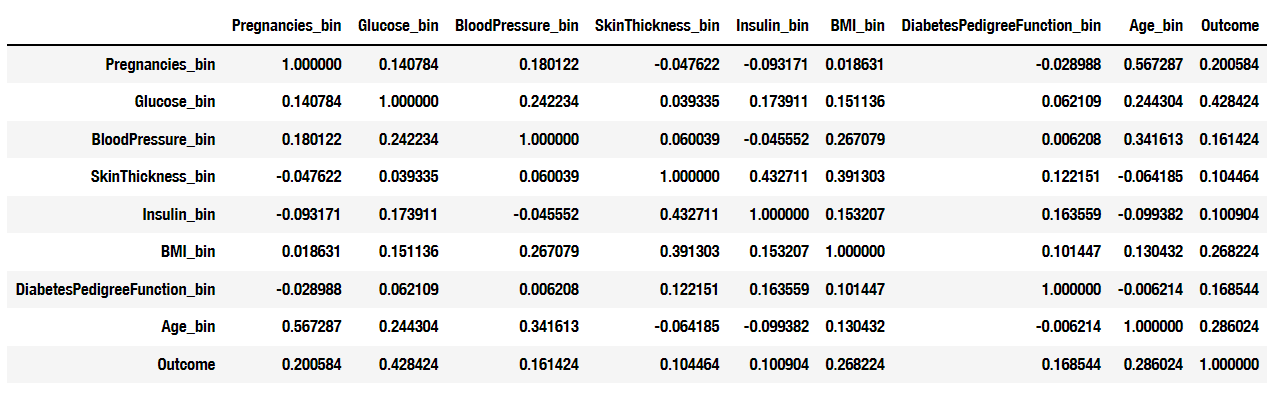
In this experiment, we chose to use categorical variables to observe the relationship between each variable and the outcome.

I defined a function to binning continuous variables into categorical variables with 3 levels (low, medium, high). The dataset will be tested with N way ANOVA, and we need to ensure the binning is balanced ahead of time. It means that under each category, the counts are the same and the data transformation is completed. From Fig 6, we can have an overview of the binned data. Then we will analyze the categorical variables by the correlation matrix again for the colineary.



**Figure 6**: Information score with transformed independent variables.

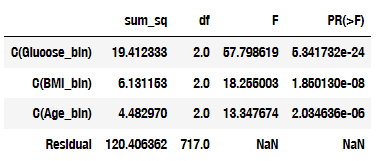
According to the correlation matrix, select Glucose, Age, BMI as input variables, these variables are not quite correlated from the chart.



**Figure 7**: Pearson Correlation Analysis with transformed independent variables.

We conduct the OLS Regression model again to ensure the model’s significance. The conclusion of the OLS Regression chart shows that the model is valid with high F score, and F-statistic is much less than 5%. It’s enough to reject the null hypothesis: the model is not significant, which means it's a significant model. However, the adj R square is quite low, only 0.251, meaning only 25.1% of variance is explained by this model.

1. **ANOVA Test**

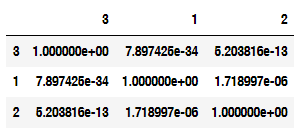


**Figure 8**: ANOVA Test

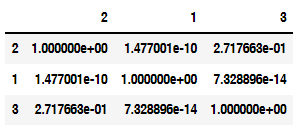
From the ANOVA Test, we can see that Glucose has the highest impact on the incidence of diabetes, as high as 19.41. It is 3 times the influence of BMI and 4 times that of Age. From this, we can conclude that in terms of preventing diabetes, the first thing we need is to detect and control blood sugar levels to avoid sudden diabetes caused by eating high-glycemic foods. Second, the higher the body fat percentage, the greater the possibility of developing diabetes. And with age, the likelihood of developing diabetes also increases. The correlation between the two variables is not high, which means that older age is not more likely to be obese. Therefore, the elderly need to pay more attention to healthy diet and living habits, and maintaining a good BMI can help the elderly better control the onset of diabetes.

1. **Post Hoc Analysis**

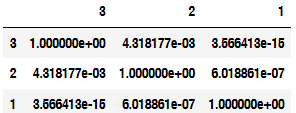
Finally, we used post hoc analysis to test for differences in means to ensure a complete ANOVA analysis process. The post hoc analysis of the three variables shows that the conclusion of the ANOVA test is still valid.



**Figure 9**: Post hoc analysis for Glucose\_bin



**Figure 10**: Post hoc analysis for BMI\_bin



**Figure 11**: Post hoc analysis for Age\_bin

1. **Summary**

In this study, we first extracted the available rows after cleaning the data, and then performed Pearson Correlation Analysis. After OLS Regression Analysis, we binned continuous variables into categorical variables with 3 levels (low, medium, high) and repeated the above process. Finally, we conducted ANOVA Test and got the correlation conclusion. Finally, Post Hoc Analysis is used to ensure the accuracy of the conclusions.

Major risk factors for diabetes include age, obesity, family history, pregnancy, and more. Experiments based on this data analysis did support an increase in the risk of developing diabetes with increasing age and BMI. While blood sugar level is one of the important indicators to measure type 2 diabetes, based on correlation analysis and established knowledge, we can say that by adopting a healthier lifestyle, a person's risk of developing diabetes will be reduced. As people age, people with higher BMIs, known as obese people, are at higher risk and need extra attention to regular checkups and sugar intake. For the elderly, we recommend more balanced nutritional intake and proper exercise to prevent diabetes.

1. “Type 2 Diabetes.” Mayo Clinic. Mayo Foundation for Medical Education and Research, January 20, 2021. https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193. [↑](#footnote-ref-0)
2. Ibid. [↑](#footnote-ref-1)
3. Ibid. [↑](#footnote-ref-2)