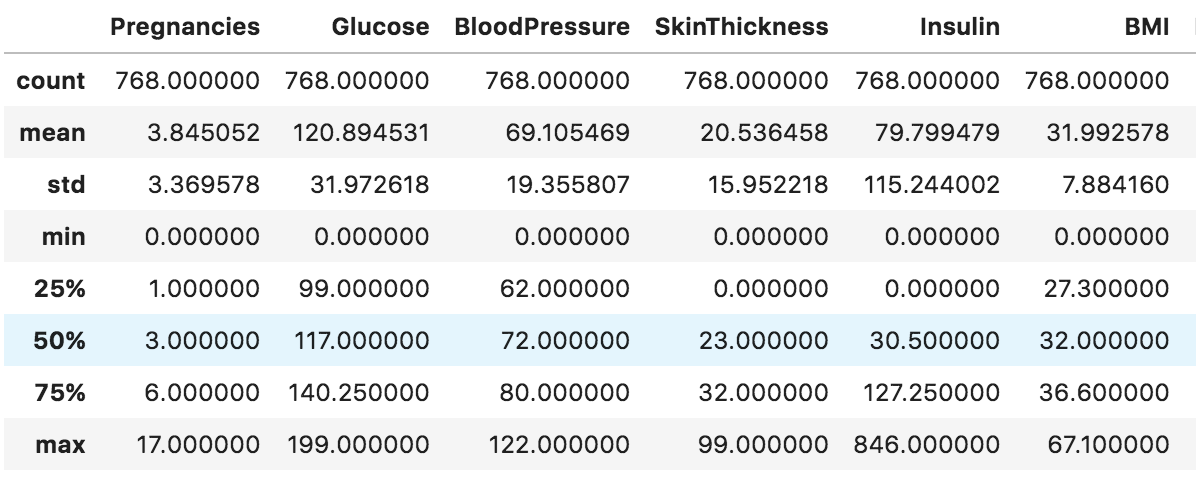
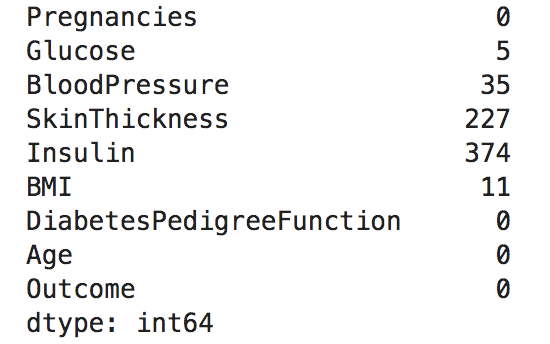
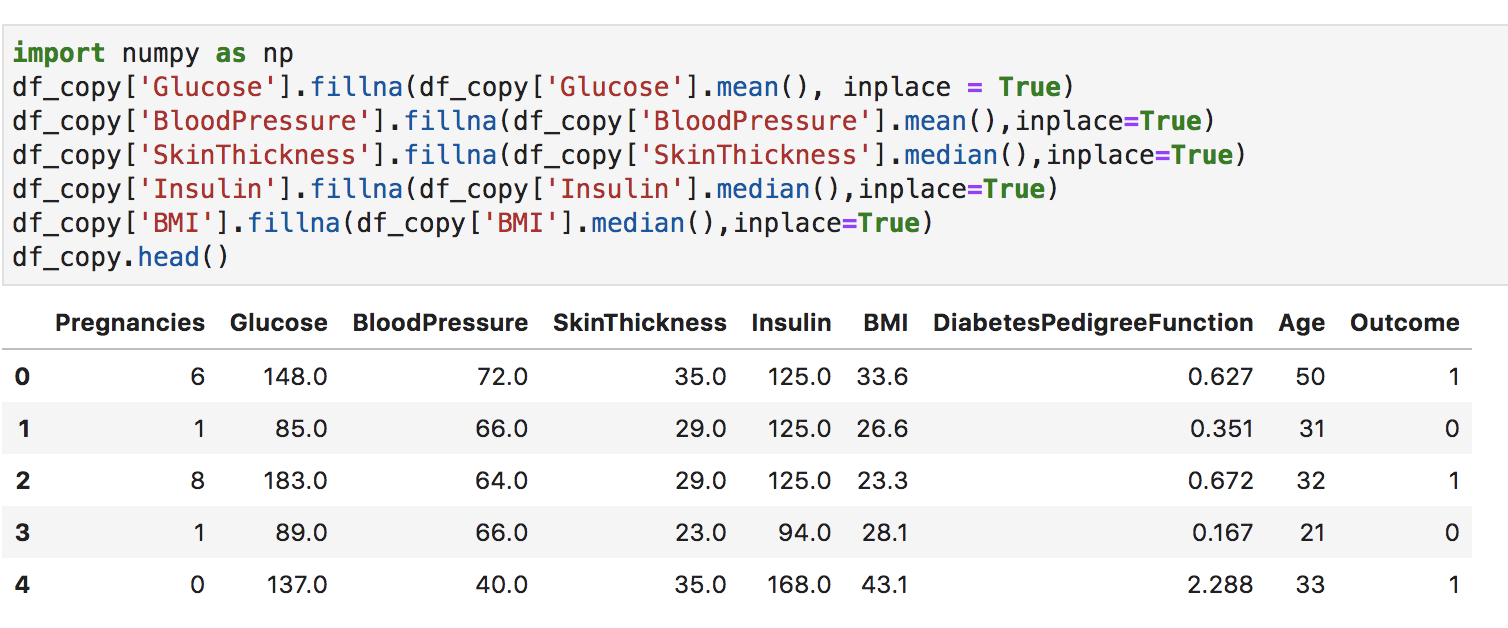
* **Clean Diabetes dataset**

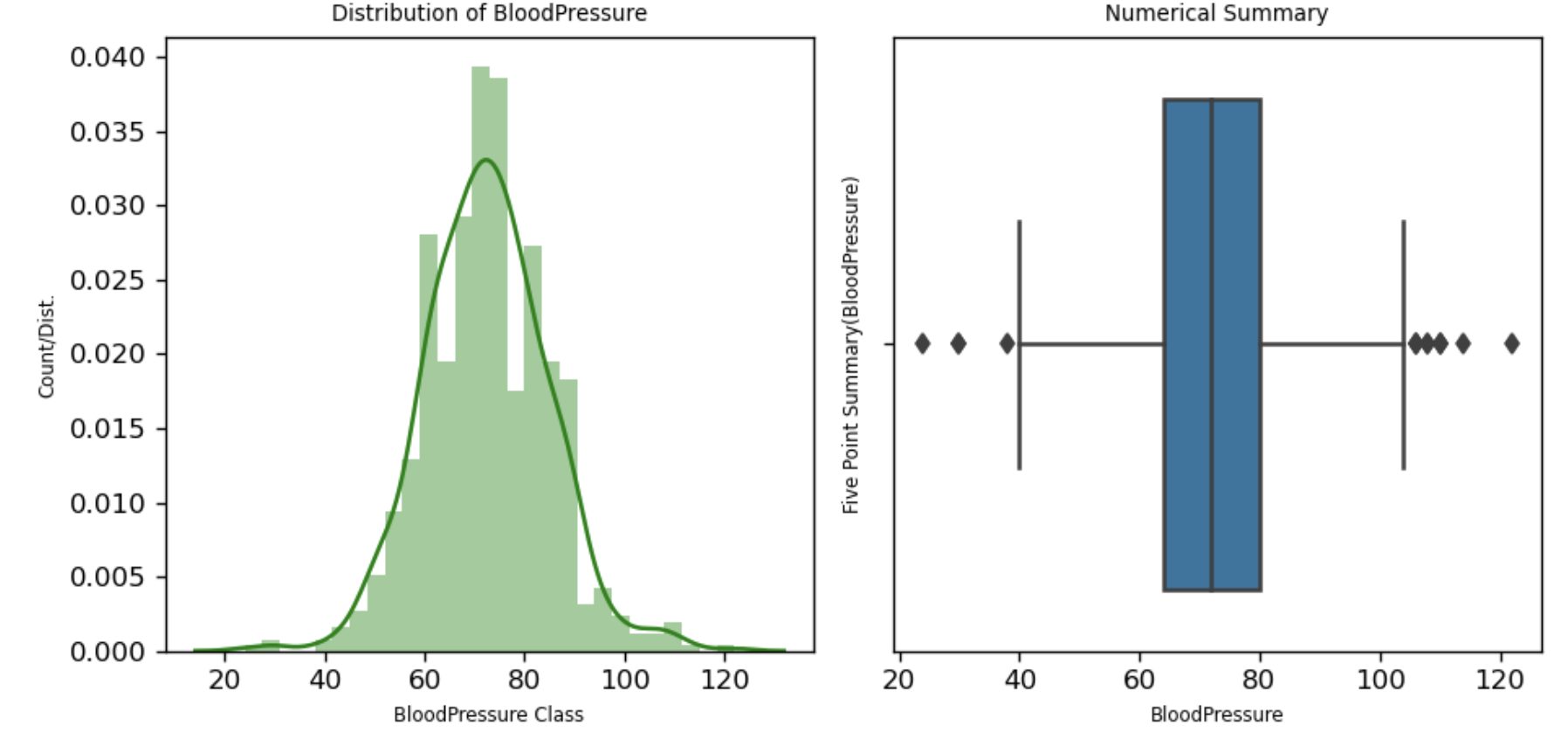
The diabetes dataset has 9 columns, 8 independent variables, and 1 dependent variable. There are 768 entries with the integer and float datatypes. When using the describe function to check the dataset information, several columns have a minimum of 0, which does not make sense. They are Glucose, BloodPressure, SkinThickness, Insulin, and BMI. Pregnancies here do make sense, which could indicate no pregnancy. 

Summing up the null values from these columns, I noticed that the null value accounts for a considerable portion in the column for BloodPressure, SkinThickness, and Insulin, 5%, 30%,49%, respectively.

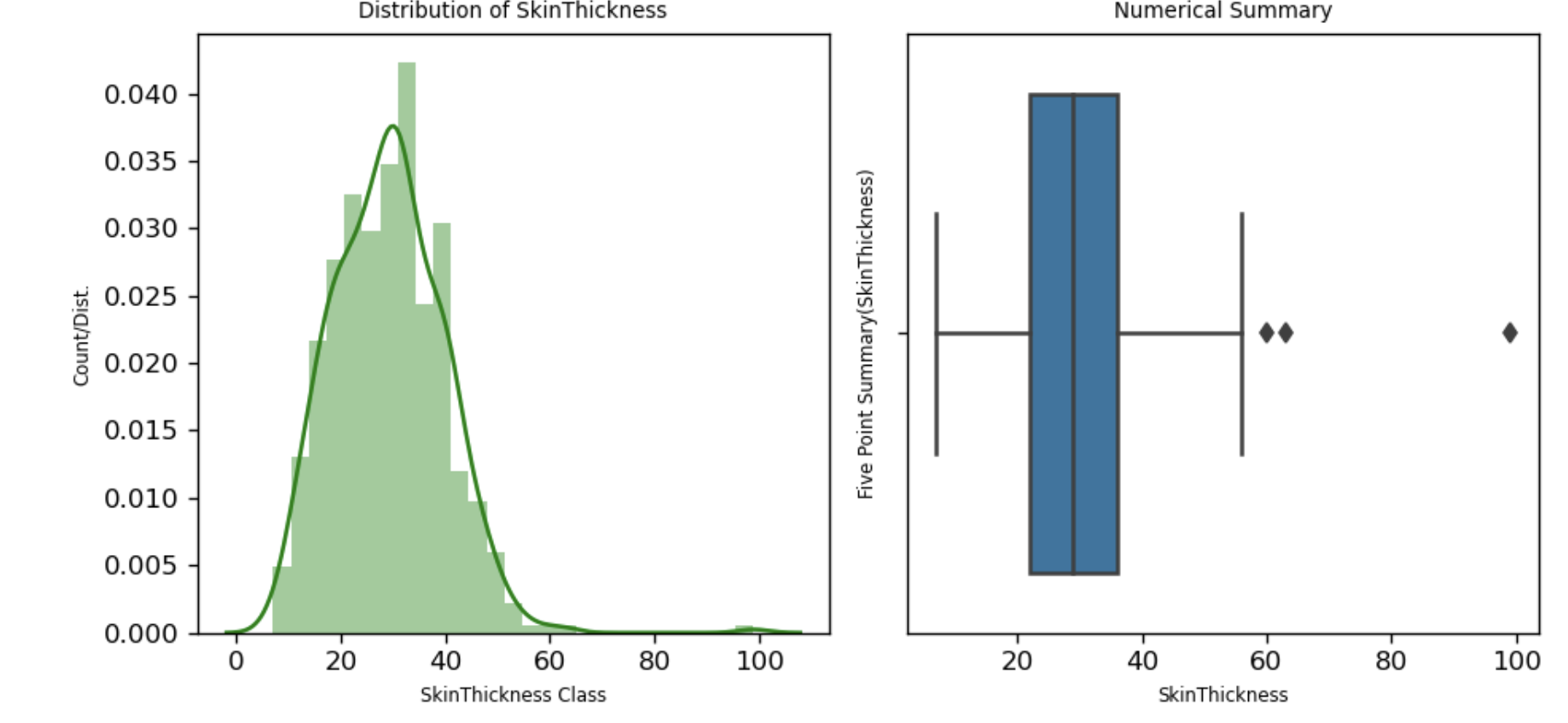


With this high percentage and keeping the total entry amount, I suggest filling in the values with either mean or median. I then checked these 5 columns’ distributions and box plots, I realized to fill in the mean to the missing values for the right-skewed distribution and fill in the rest with the median to keep them bell shapes so the modified data can be applied to the prediction models. 

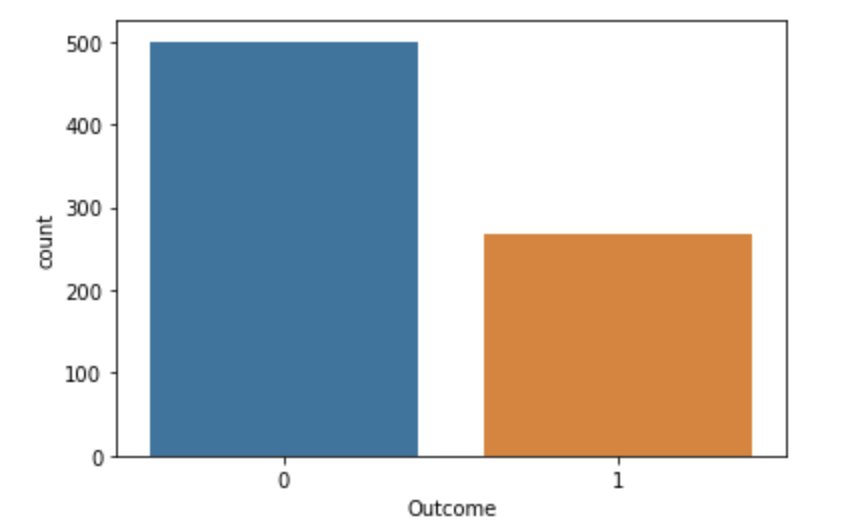
For example, from the blood pressure distribution and box plot diagram, we can tell the distribution displays a tall bell shape. There should be outliers on both ends, which are also displayed on the box plot. The mean should replace the missing values, making the diagram normalized. This method also applies to glucose levels.

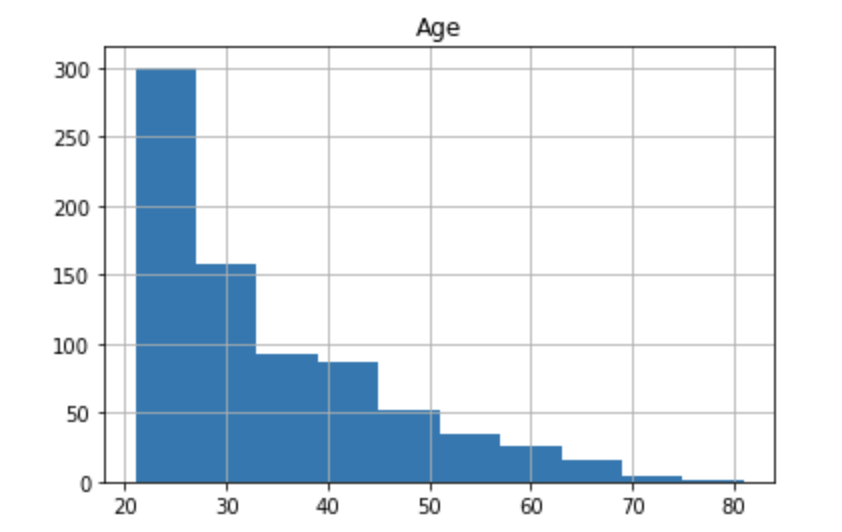


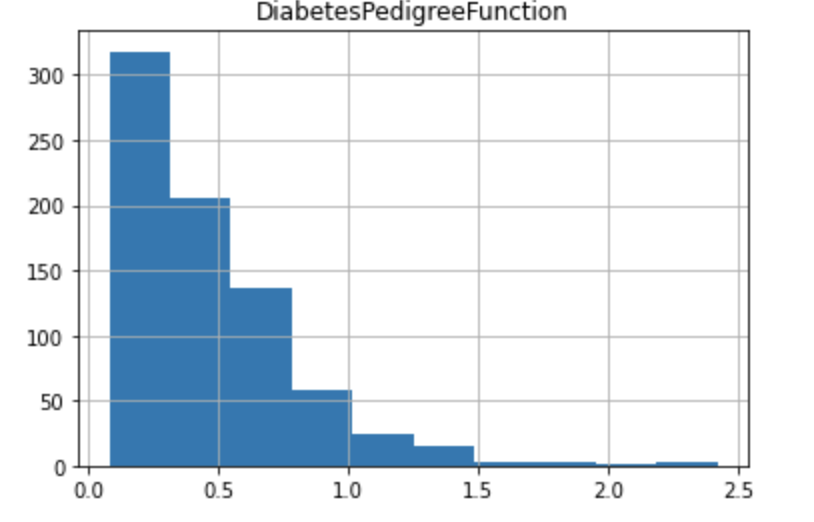
The other example is skin thickness level. We can tell the distribution displays are skewed to the right. There should be outliers on the right side, which is also displayed on the box plot. The median should replace the missing values. Because the median in a right-skewed diagram is higher than the mean, it can help the diagram normalized. Insulin and BMI level also follow this filling in missing value pattern.

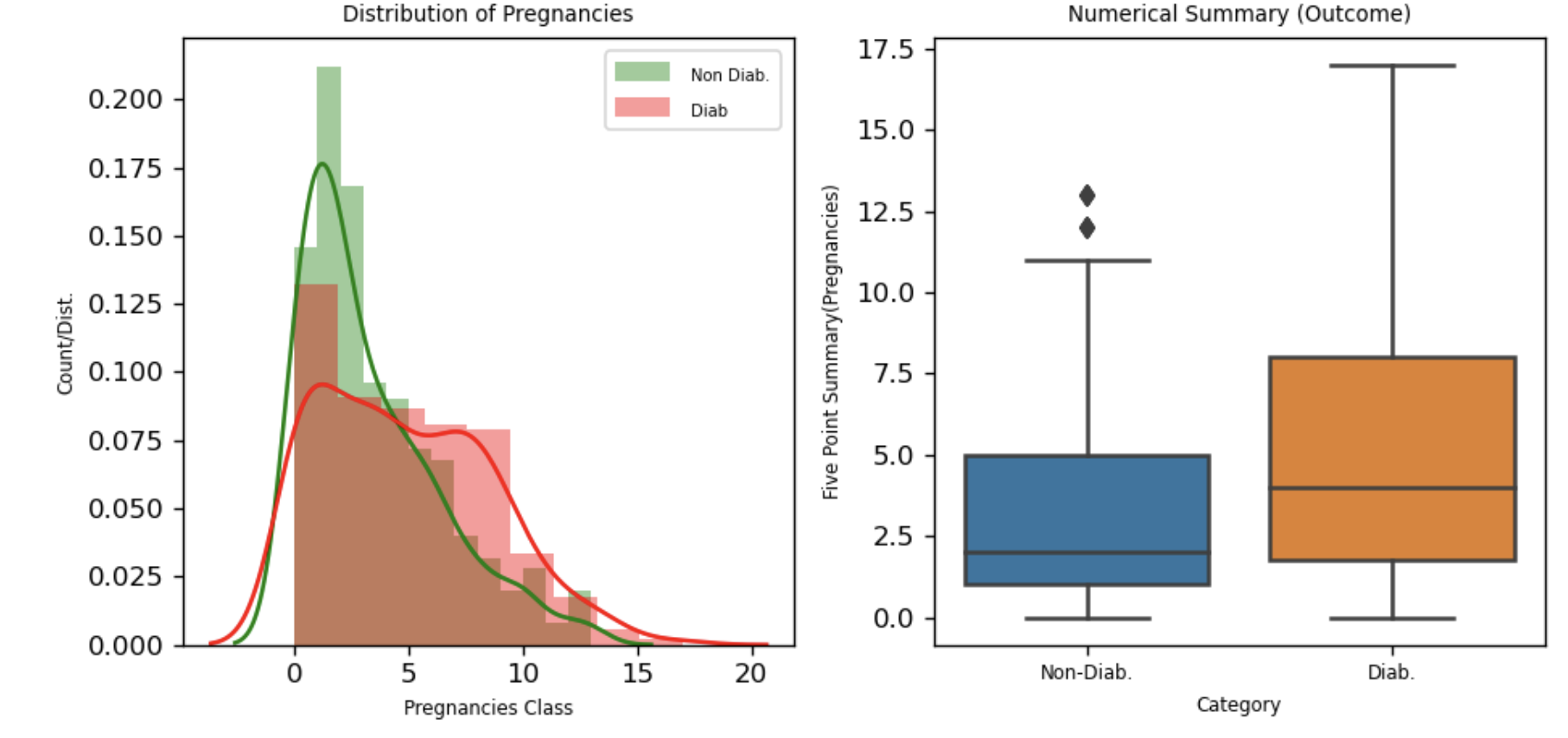


* **Conduct Exploratory Data Analysis using Visualizations and t-tests to test your initial hypotheses/research questions. Your RQs can and will morph and change appropriately as you go through the process of analysis.**

The outcome in the graph has two values, 0 and 1. ‘0’ here means no diabetes, and ‘1’ here means having diabetes. The graph shows that no diabetes has a higher number than with diabetes. 

The age and Diabetes Pedigree Function are both right-skewed in the graph. From the information description, they look normal.



I then discuss how variables respond to the outcomes from the comparison distribution diagrams and box plots. From pregnancies distribution, the distribution diagram shows the distribution skewed to the right with the outliers are present on the higher end. Both box plots show that people with longer pregnancy time have a higher risk of diabetes. 

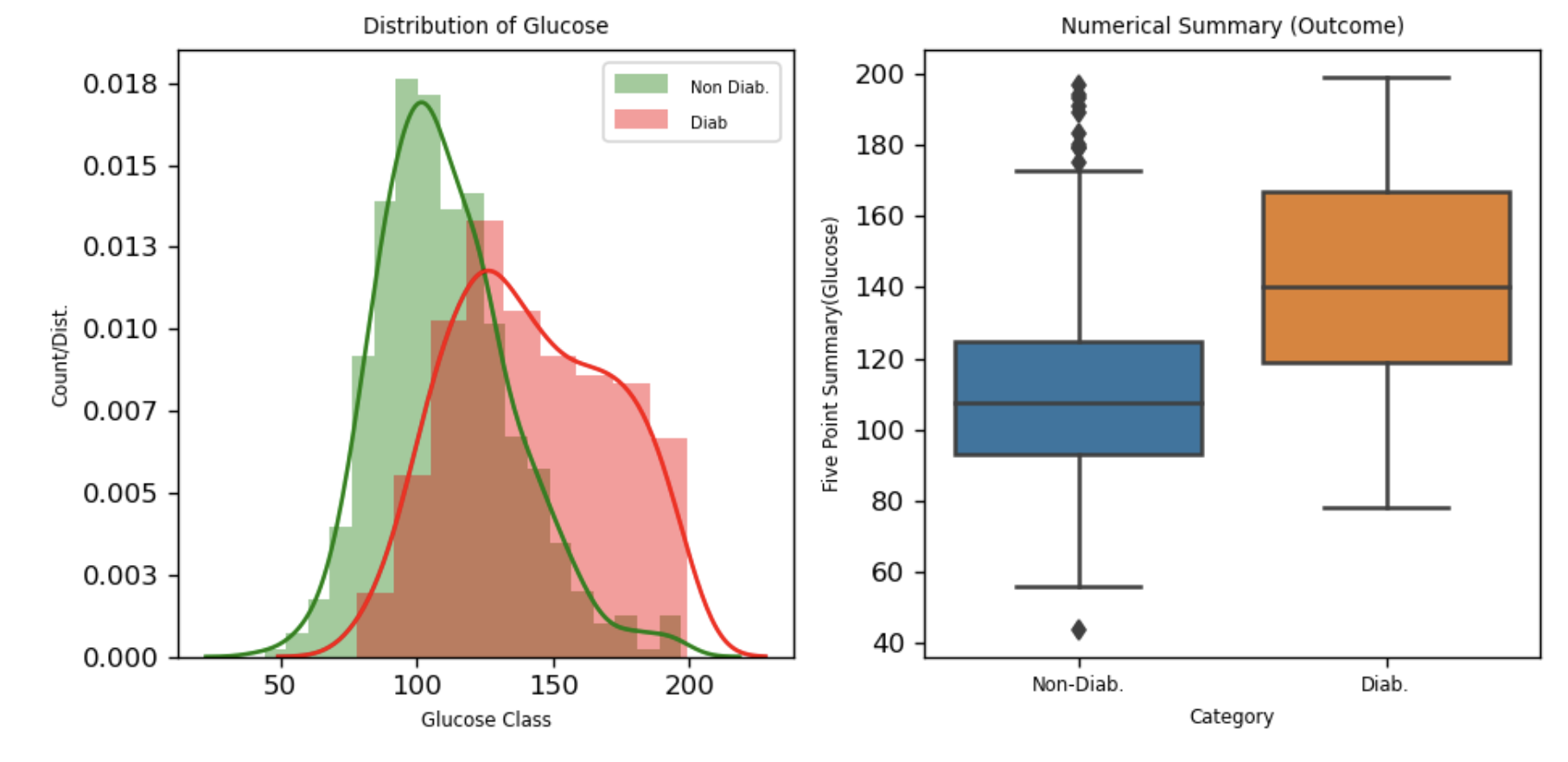
**Research question**: Considering with 95% confidence, do diabetic patients have longer mean pregnancies time than non-diabetic patients?

H0: μ1 − μ2 is 0

Ha: μ1 − μ2 is not 0

μ1 here means mean pregnancies time of Diabetic Patients, while μ2 means mean pregnancies time of Non-Diabetic Patients. The significance level is 5%.

I randomly choose 30 samples from non-diabetes and diabetes pregnancy data. The p-value is 5.065e-10, which is smaller than the 0.05 significant level. Thus, we reject the null. The conclusion for this research question is that we support a significant difference between the mean pregnancy time for diabetic and non-diabetic patients.

The non-diabetes with glucose distribution diagram shows the distribution skewed to the right, with the outliers present on the higher end. Both box plots show that people with higher glucose levels have a higher risk of diabetes. 

**Research question**: Considering with 95%, do diabetic patients have higher glucose level than non-diabetic patients?

H0: μ1 − μ2 is 0

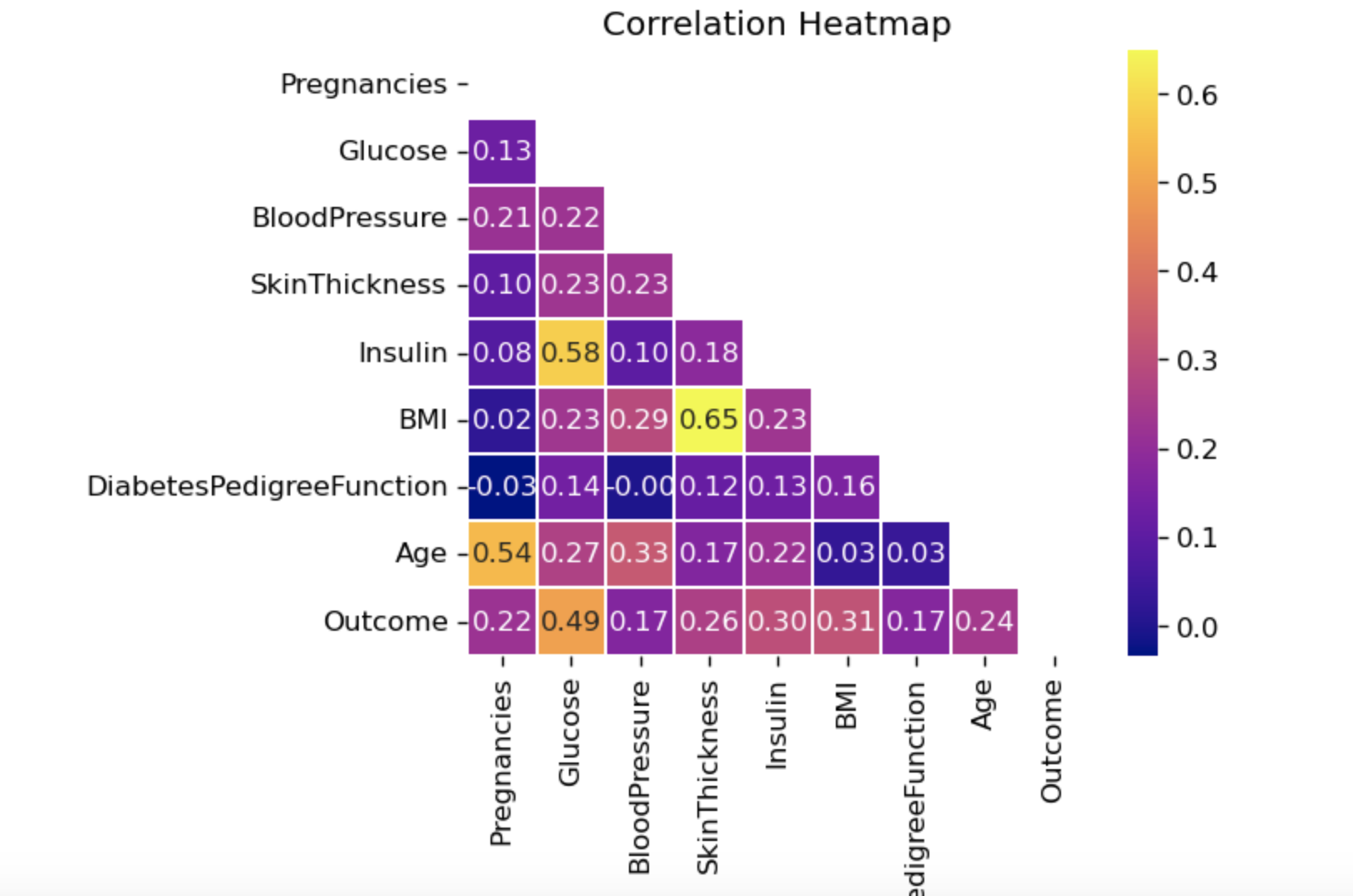
Ha: μ1 − μ2 is not 0

μ1 here means mean glucose level of Diabetic Patients, while μ2 means mean glucose level of Non-Diabetic Patients. The significance level is 5%.

I randomly choose 30 samples from non-diabetes and diabetes glucose data. The p-value is 5.62e-05, which is smaller than the 0.05 significant level. Thus, we reject the null. The conclusion for this research question is that we support a significant difference between the mean glucose level for diabetic and non-diabetic patients.

In addition to the Pregnancy Time and glucose, I also conduct the independent t-test on BloodPressure, SkinThickness, Insulin, BMI, and Age variables to see if the higher these variables are, the higher risk patients will have diabetes. The p-value of all above variables are lower than 5%, which are 3.709e-06, 1.793e-09, 1.21e-08, 8.337e-19, 2.21e-11, respectively. We reject the null hypothesis and support a significant difference between the BloodPressure level, SkinThickness Insulin, BMI, and Age mean for diabetic and non-diabetic patients.

I also conduct the multivariable analysis. I do not think there are strong linear relationships between any two variables from the heat map diagram below.

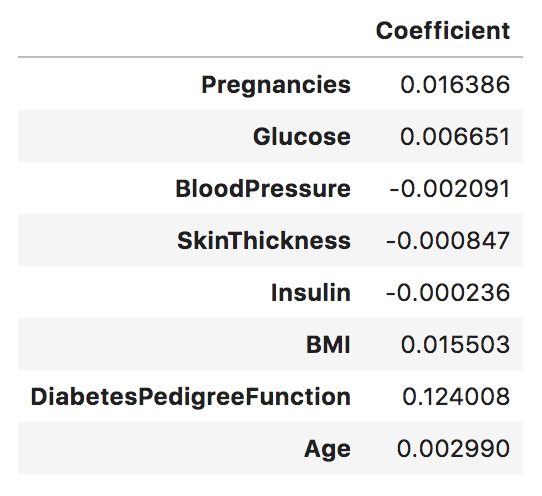


* **Conduct Multiple Linear Regression.**

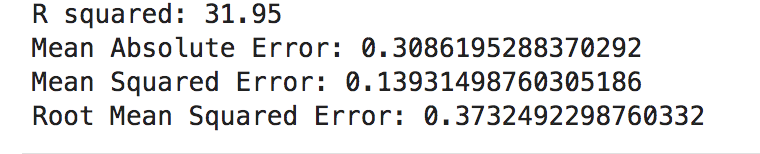
Before conducting the multiple linear regression, I divided the dataset into two parts, the training dataset, and the testing dataset. The training dataset is 80% of the original dataset, and the testing dataset is 20% of the original dataset. The hypothesis of the multiple linear regression model is there will be no significant prediction of outcomes by pregnancies, glucose, blood pressure, skin thickness, Insulin, BMI, DiabetesPedigreeFunction, and Age. After applying the multiple linear models, the coefficient result or R^2 is returned, which presents how strong relationship between the independent variables and the dependent variable.

In the coefficient table, for every one percent increase in pregnancies time, there is an associated 1.6 percent increase in getting diabetes; for every one percent increase in glucose level, there is an associated .67 percent increase in getting diabetes; for every one percent increase in blood pressure level there is an associated .2 percent decrease in getting diabetes; for every one percent increase in skin thickness level there is an associated .08 percent decrease in getting diabetes; for every one percent increase in insulin level there is an associated .02 percent decrease in getting diabetes; for every one percent increase in BMI level there is an associated 1.5 percent increase in getting diabetes; for every one percent increase in Diabetes Pedigree Function

level, there is an associated 12.4 percent increase in getting diabetes; for every one percent increase in Age, there is an associated 0.3 percent increase in getting diabetes.



The value of R square of the multiple linear regression is 31.95, which means only 31.95% of data fit the regression model.



Because the independent variable has small values, the mean absolute error, mean squared error, and root mean squared error shows small values. Typically, the lower these values, the better the model performs. However, RMSE is 0.37, greater than 10% of the outcome’s mean. This also says that the model is not accurate enough. One reason could be not enough data. The second reason could be that the assumption is not good enough. Also, it could be because I did not choose the good features.

* **You will need to wrangle continuous columns into categorical ones that make sense. Make sure you cite why you did what you did.**

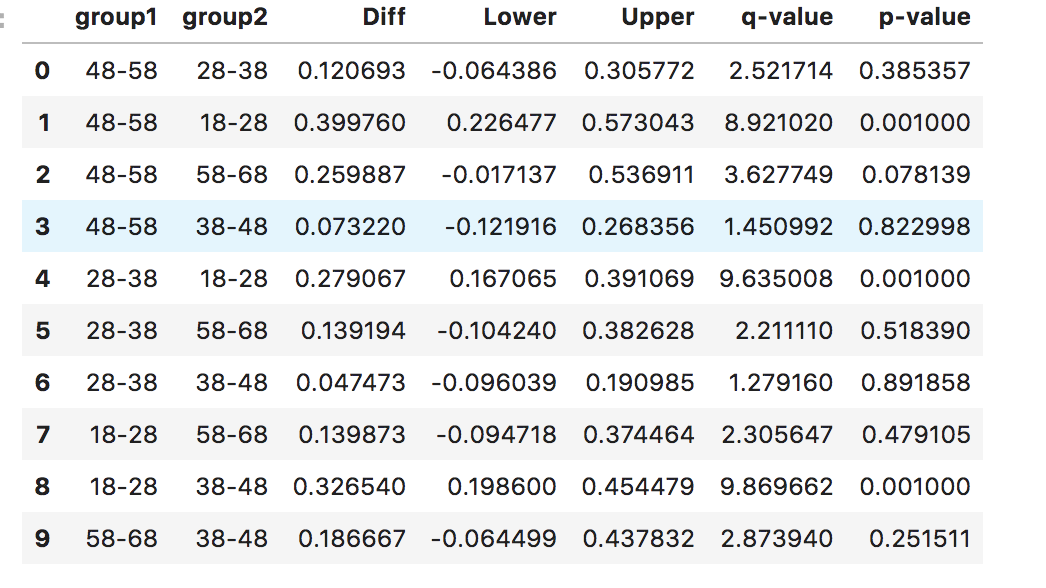
I also conduct an ANOVA test on this dataset. Before conducting the Anova testing, I check the unique value in the dataset. Outcome and pregnancy are both category variables, which are not part of my research questions. I used a t-test to see why the age mean has a significant difference between diabetes and non-diabetes. I split the age into 5 parts, 18-28,28-38,38-48,48-58,58-68. Each range represents a generation of people who could have diabetes. Then, I can examine if each age range has any relationship with the diabetes outcomes.

* **ANOVA modeling**

I did one-way ANOVA first to see if age group means are equal, given the 95% confidence interval.

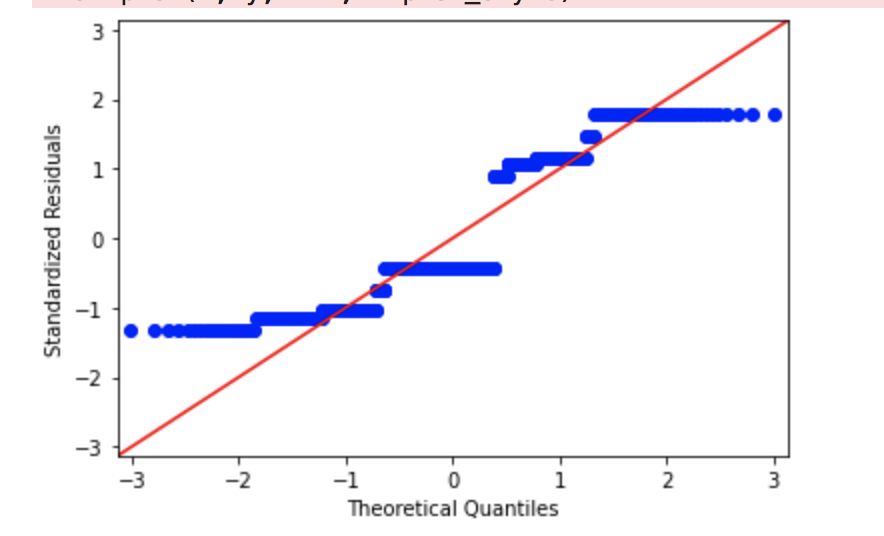
* Null hypothesis: Age groups means are equal (no variation in means of groups)  
  H0: μ18-28=μ28-38= μ38-48= μ48-58=μ58-68
* Alternative hypothesis: At least, one age group mean is different from other groups  
  H1: All μ are not equal

With one-way ANOVA, the p-value is 0.003, less than 0.05. I conclude that there is significant differences among age groups.

Because there is no way to know which age group is significantly different from others. Thus, I did the post hoc comparison analysis. 

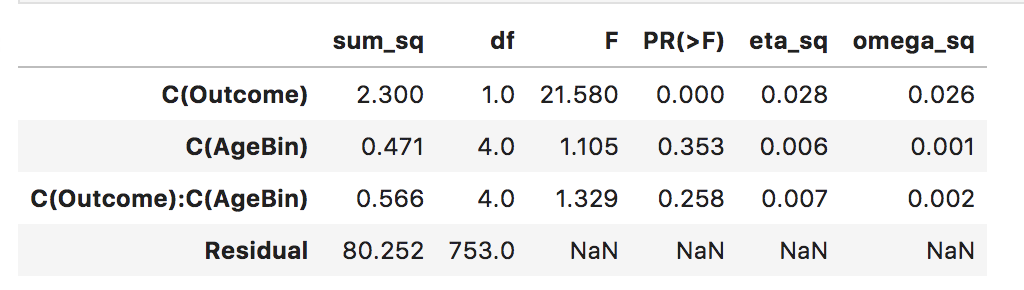
Above results from Tukey’s HSD suggests that 1,4,8 pairwise comparisons for age groups rejects null hypothesis (p < 0.05) and indicates statistical significant differences.

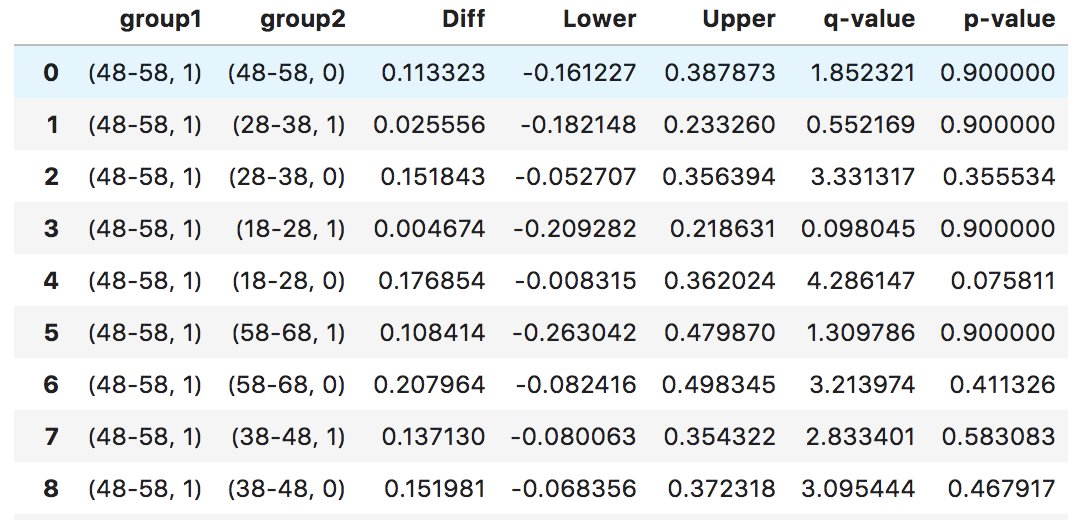
In addition to the Tukey’s HSD, I also generate the QQ plot to from standardized residuals, where outliers can be seen in the graph.

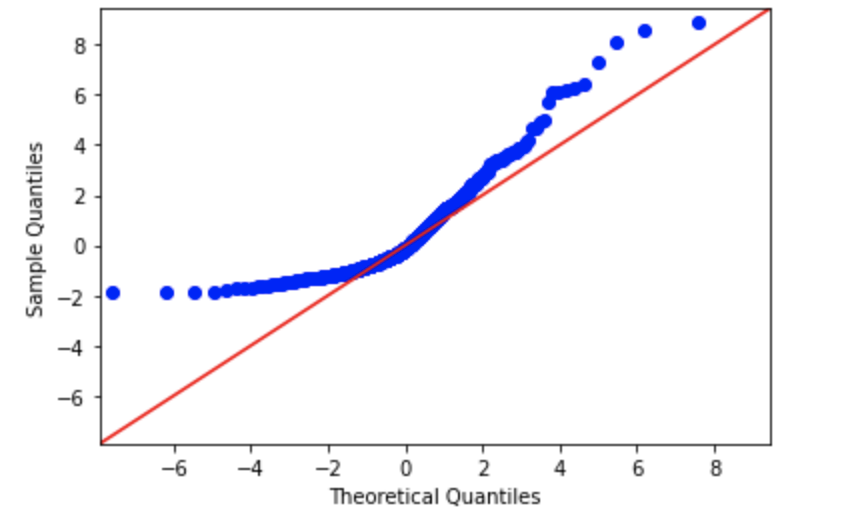


As the standardized residuals do not lie around the 45-degree line, it suggests that the residuals are approximately not all normally distributed. For the next step, the Shapiro-Wilk test can be used to check the normal distribution of residuals. But, this part will not include the Shapiro-Wilk test.

I also conduct two a two-way ANOVA test. Using two-way ANOVA, i can simultaneously evaluate how age group and diabetes outcome affect the Diabetes Pedigree Function. Here I need to test three hypotheses 1) effect of age group on Diabetes Pedigree Function 2) effect of diabetes outcome on Diabetes Pedigree Function, and 3) effect of both variable's interactions on Diabetes Pedigree Function.



From the above diagram, the p-value obtained from ANOVA analysis for diabetes outcome is statistically significant (p<0.05). Age and interaction are more significant values than 0.05. We conclude that diabetes outcome significantly affects the Diabetes Pedigree Function. Age group and interaction do not significantly affect the Diabetes Pedigree Function.

I only take a screenshot of part comparison results. From the diagram above, we can see 4 has a lower value than 0.05, where we can reject the null. Patients within 48 to 58 age range with diabetes and within 18 to 28 age group without diabetessignificantly affect the Diabetes Pedigree function. 

In the residual plot, standardized residuals lie around the 45-degree line, it suggests that the residuals are not approximately normally distributed.

Compared to multiple linear regression, the ANOVA model can better explain the relationship of variables. I gradually learned when to use an independent t-test, paired t-test, one-way and two-way ANOVA from the statistical analysis. I think these methods are essential to help select the features, so the model can perform better for predictions with convenient features. In addition to two-way ANOVA, I think N-way ANOVA also should be conducted. However, this paper does not include this. Hopefully, I can conduct the 3 or 4-factor ANOVAs in this experiment.