Project Summary

A study is being done to determine predictive factors contributing to the presence of diabetes in subjects. Data was gathered on 8 different factors that could possibly affect whether subjects had diabetes or not. The organizers of the study would like the base model to consist of only the age factor. My job is to conduct an analysis on what factor, besides age, best predicts diabetes presence.[1] The dataset can be found here: https://www.kaggle.com/datasets/aemyjutt/diabetesdataanslysis. There are 768 data points in total.

Key Questions

The main question I will be trying to solve is which factor in addition to age creates the best prediction model for the presence of diabetes. Then I will analyze whether this additional factor model is the best model or if the organizers should stick with the original base model. I also hope to paint a picture of the data and possible patterns overall.

Data Preparation

This dataset was already very workable and required no changes according to our needs. I will be analyzing all factors where data was collected, so no columns need to be removed. There were also no duplicates in the dataset.

Approach

To perform my analysis, I first started with calculating some summary statistics of each factor to gain more insight into the data. A summary of this data can be found in Table 1. Then, I generated several linear models starting with the base model and then a model consisting of age and the factor I am analyzing. I then calculated the sum of squared errors (SSE) for each model to use in calculating the coefficient of partial determination (R^2). I will be using the coefficient to determine the best additional factor. Once I have determined the best additional factor, I will compare that model to the base model using hypothesis testing and ANOVA. Lastly, I generated informative graphs to visualize our results.

[1] This situation is purely hypothetical

Tables & Plots

Table 1: Summary Statistics

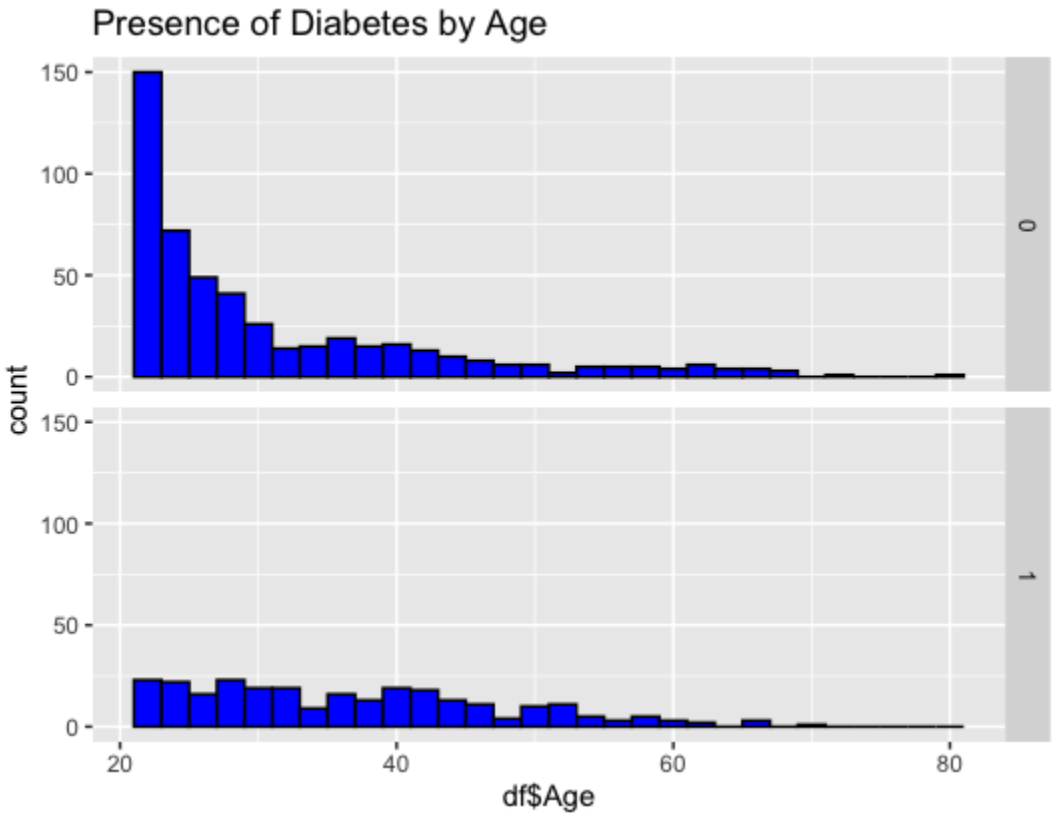
	Mean	Standard Deviation	Correlation w/ Outcomes Column	Range
Age	33.2409	11.7602	0.2384	60
Pregnancies	3.8451	3.3696	0.2219	17
Glucose	120.8945	31.9726	0.4666	199
Blood Pressure	69.1055	19.3558	0.0651	122
Skin Thickness	20.5365	15.9522	0.0748	99
Insulin	79.7995	115.2440	0.1305	846
BMI	31.9926	7.8842	0.2927	67.1
Diabetes Pedigree Function	0.4719	0.3313	0.1738	2.342

Table 2: Linear Models

	B_0 (Intercept)	B_1 (Age)	B_2 (Additional Factor)	R^2
Age (Base model)	0.0276	0.0097	N/A	N/A
Pregnancies Model	0.0525	0.0068	0.0185	0.0128
Glucose Model	-0.6007	0.0050	0.0065	0.1857
Blood Pressure Model	0.0159	0.0096	0.0002	0.00007
Skin Thickness Model	-0.0516	0.0101	0.0031	0.0112
Insulin Model	-0.0269	0.0099	0.0006	0.0210

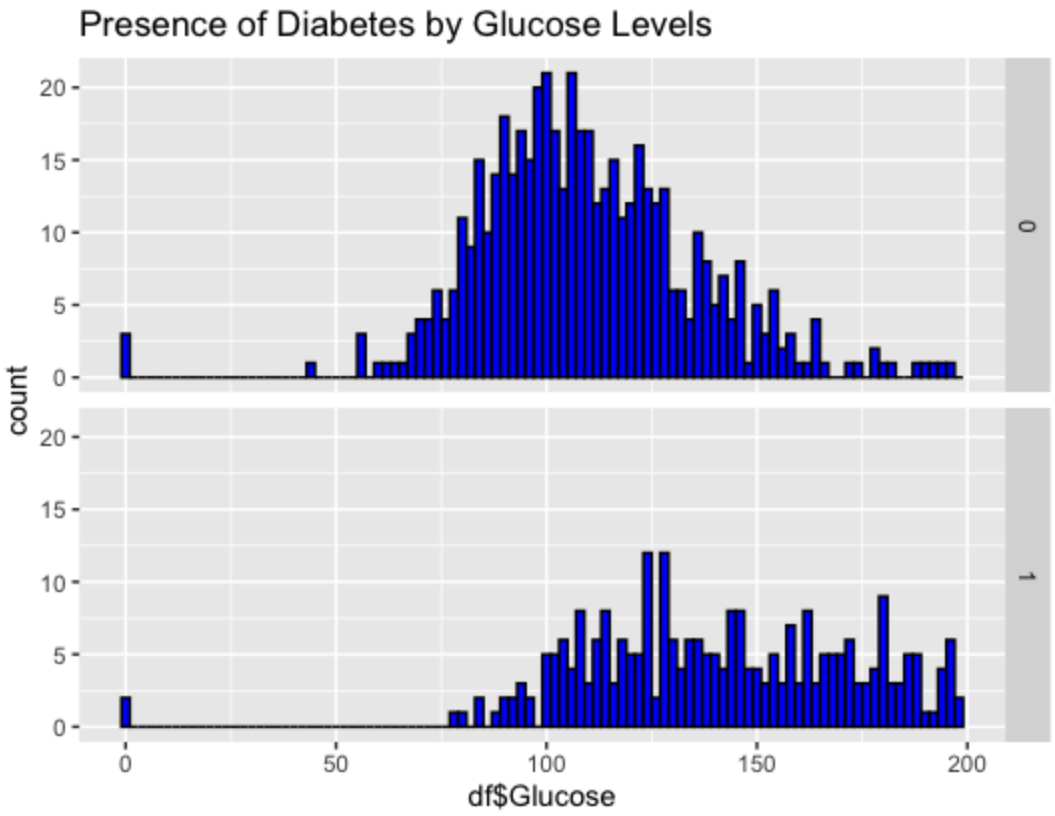
BMI Model	-0.5090	0.0092	0.0172	0.0857
Diabetes Pedigree Model	-0.0776	0.0094	0.2390	0.0292

Figure 1:



* 1 represents presence of diabetes in subject, 0 represents no presence

Figure 2:



Results & Takeaways

From Table 1, we can see some interesting results. Most importantly, we can see the differences in correlation with the "Outcomes" column that represents the presence of diabetes. Glucose has the highest correlation, indicating that it will possibly be the best additional factor. In Table 2, we can see the coefficients of our models and the corresponding R^2. Based on the values of R^2 I found, glucose would be the best factor alongside age in predicting the probability of getting diabetes as it has the highest R^2 with a value of 0.1857. I then conducted a hypothesis test where I compared the full (glucose) and reduced (base) models by generating the ANOVA table. From the ANOVA table, we get a F-statistic of 174.51 and a p-value < 0.0001, so we can conclude the glucose model is better. Our glucose coefficient tells us that for every 1 unit increase in glucose level, the chance of having diabetes increases by 0.65%. Out of the several plots I generated, I included two of them in this report. Figure 1 is a histogram of presence of diabetes by age. From the plot, we can see that age does not seem to have a large impact on the chance of getting diabetes. Figure 2 is a histogram of presence of diabetes by glucose levels.

From where the data is centered, we can see that as glucose levels increase, the chance of diabetes rises.

The goal of this report was to conduct a statistical analysis of factors contributing to the presence of diabetes in subjects. Through many steps, we were able to conclude that glucose is the best predictor alongside age in predicting whether a subject would have diabetes or not. We also calculated summary statistics and generated plots that helped give insights of the data itself.