Written Report

Team 4: Angela Chen, Chenye Zhu, Amy Liu, Justin Pan

2024-11-02

Introduction and EDA

Introduction

Diabetes is a chronic disease characterized by persistently high blood glucose levels due to insufficient insulin production (Type 1, ~ 5% of cases) or ineffective insulin use (Type 2, 90-95% of cases)(Bullard et al. 2018). The International Diabetes Federation estimates 10.5% of adults aged 20-79 lives with diabetes, with nearly half unaware of their condition. Common diagnostic tests include plasma glucose, which measures glucose after fasting, and the A1C test, which measures average glucose over 2-3 months. High-risk groups for Type 2 diabetes include individuals aged 35 or older, women who have experienced gestational diabetes, those with a family history, people who are overweight or obese, and certain racial and ethnic groups (NIDDK 2023).

Recent studies have advanced our understanding of the risk factors and dynamics of diabetes. For example, research conducted by Shuguang Hospital in China found significant interactions between age and other risk factors, influencing diabetes risk among middle-aged and elderly populations in Shanghai (Yan et al. 2023). Another study by Hubert Kolb and Stephan Martin focused on how lifestyle and environmental factors can increase body mass index (BMI) and lead to the loss of beta-cell function, a direct precursor to diabetes. (Kolb and Martin 2017)

This study investigates the relationship between glucose levels, various risk factors, and the likelihood of developing Type 2 diabetes in adult women. This project will enable us to understand how demographic and environmental factors contribute to diabetes prevalence among adult women, enhancing diabetes prevention and management.

Data

Our dataset includes 2,768 observations from the National Institute of Diabetes and Digestive and Kidney Diseases and Frankfurt Hospital, including anthropometric measurements (glucose level, BMI, blood pressure, etc.) alongside a binary diabetes outcome (1 - diabetic, 0 - not diabetic).

In our analysis, we focus on several key variables due to their established correlation with diabetes risk. Elevated fasting plasma glucose levels are a primary indicator of diabetes risk, with studies such as Zhao et al. (2019) confirming that higher levels significantly increase the likelihood of developing type 2 diabetes. Blood pressure is another critical factor; hypertension is not only commonly associated with diabetes but also increases the risk of cardiovascular complications,

making its management vital (Boer et al. 2017). Body Mass Index (BMI) is strongly linked to an increased risk of diabetes, as obesity contributes to insulin resistance, underscoring the importance of BMI in diabetes risk assessments (Klein et al., 2022).

Additionally, the Diabetes Pedigree Function highlights a genetic predisposition through a family history of diabetes, which is often included in risk prediction models. This function scores the probability of diabetes based on family history, with a range from 0.08-2.42 in our data. Age is another significant factor, with the prevalence of type 2 diabetes increasing notably after age 45, necessitating regular screenings for older adults. The history of gestational diabetes in women suggests a higher risk of developing type 2 diabetes later, marking it as a critical factor in risk assessments. Lastly, abnormal insulin levels can indicate beta-cell dysfunction and insulin resistance, both precursors to diabetes, which is why measuring insulin is crucial in understanding an individual's metabolic state. (Joshi and Dhakal 2021) As indicated in Collier's research, an increase in skin thickness is significantly related to the duration of diabetes, therefore, it will be worthy to explore this variable. (Collier et al. 1989)

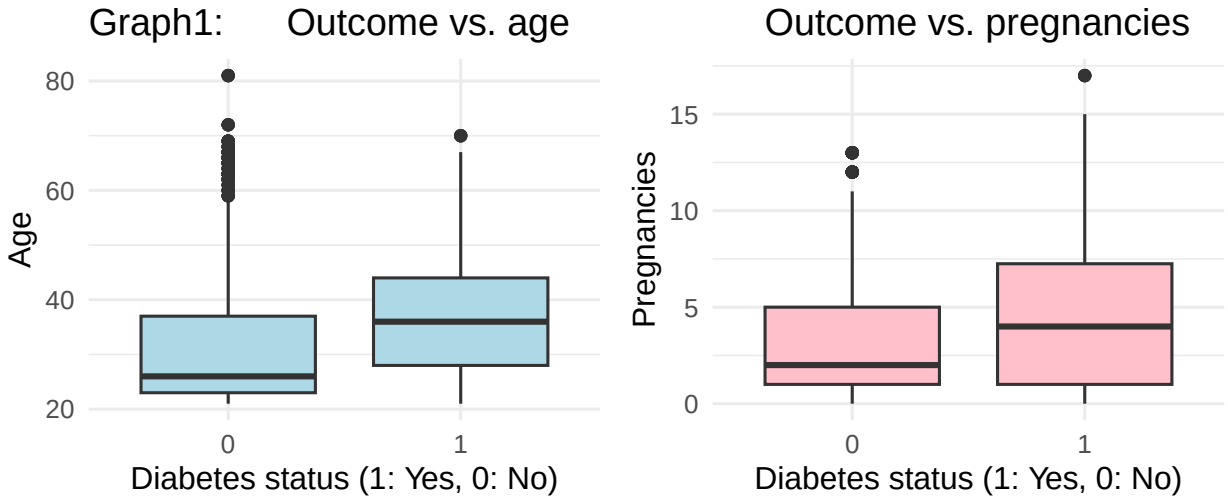
Exploratory Data Analysis

Table 1: Distribution of Diabetes Outcomes

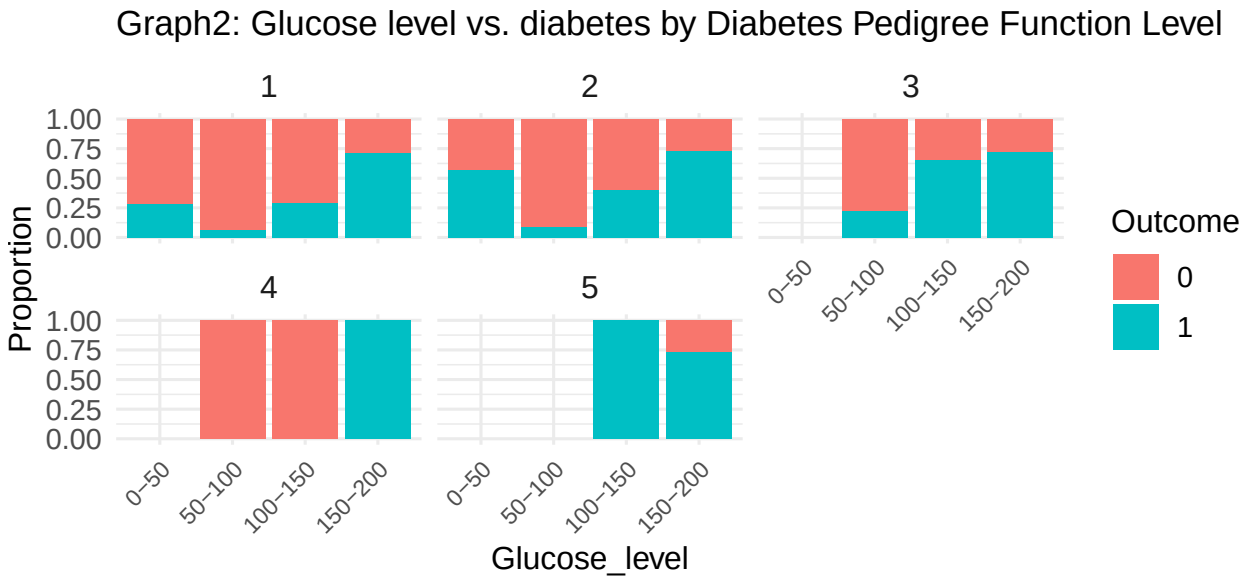
Non-diabetic: 0	Diabetic: 1
1816	952

From our explanatory data analysis, we found that around 34% of our sample are patients diagnosed with diabetes. Our outcome variable is diabetes status, with `1` indicating yes and `0` indicating no.

We explored the relationship between explanatory and our diabetes outcome variables. We included the effect of age and pregnancy status here as an example and found the median age and number of pregnancies is higher for those with diabetes compared to those without. The variability for age is similar across both statuses, but there is a significant number of outliers for those without diabetes. The variability for pregnancies is higher for those with diabetes.



We also conducted EDA to explore interesting interaction effects such as glucose levels and the diabetes pedigree function, as previous literature has emphasized including family history in diabetes risk assessment, where glucose levels vary based on genetic background (Elbein, Maxwell, and Schumacher 1991).



We divided the diabetes pedigree functions (0.078-0.242) into five levels and glucose into ranges (0-50, 50-100, 100-150, 150-200) to analyze their interaction. At pedigree level 1, there seems to be a moderate correlation between glucose levels and diabetes risk. Level 2 shows a stronger relationship, with higher diabetes prevalence even at lower glucose ranges. Level 3 displays a steep increase in diabetes risk as glucose levels rise. Level 5 consistently shows high diabetes risk across all glucose ranges, with near-complete presence at lower glucose ranges. These patterns demonstrate that te impact of glucose levels on diabetes risk is not uniform but rather changes based on diabetes pedigree function level. See more interaction effect exploration in the Appendix (Interaction Explorations EDA).

Additionally, we found that those with diabetes had much higher median levels of glucose. We also found that those with diabetes had slightly higher median blood pressure, median BMI, median age, and median skin thickness. We found that those with diabetes had lower median insulin, however. There was similar variability between diabetes/non-diabetes groups for blood pressure and BMI, and higher variability for diabetes group for the insulin and skin thickness. Please see appendix for EDA visualizations with all other variables (Response vs. explanatory EDA).

Methodology

Train-test data split

People with diabetes are 1/3 of the whole data set, thus we decide to do 80-20 of the train-test split (a common heuristic in machine learning). Since we have sufficient samples of each, this

is a relatively balanced choice that allows us to achieve a balance between training and testing accuracy.

Since our outcome is binary (1 & 0), we need to predict whether our selected predictor variables could predict whether the patient has diabetes or not. Thus, we apply logistic regression.

Assumptions for Logistic Regression

Empirical logit analysis indicated that most variables met the linearity assumption with our response variable, except for `Insulin` , which was excluded due to insufficent data caused by a significant number of zero values (Testing linearity assumption in Appendix). The dataset, comprising randomly selected Pima Indian women and those in Frankfurt hospital aged 21 and older, supports assumptions for randomness and independence, as individuals were sampled without evident bias (age, blood pressure, etc.) and each observation represents unique measurements. These validations justify proceeding with logistic regression analysis.

Initial model

Initially, we included all variables in our model (besides `Insulin`) to see which ones were/were not significant. We found that `SkinThickness` had a p-value much larger than 0.05, so we decided to exclude it from the model. As seen in our model output, every predictor variable is statistically significant. We also checked the VIF between all variables, and found that all were below 10, indicating low multicollinearity. (Full Model)

Table 2: Initial Model Exploration

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-7.843	0.399	-19.667	0.000	-8.640	-7.076	NA
Glucose	0.032	0.002	16.230	0.000	0.028	0.036	1.046
BloodPressure	-0.009	0.003	-3.206	0.001	-0.015	-0.004	1.124
BMI	0.076	0.008	9.591	0.000	0.060	0.091	1.091
DiabetesPedigreeFunction	0.837	0.171	4.885	0.000	0.502	1.174	1.005
Age	0.015	0.005	2.798	0.005	0.004	0.025	1.445
Pregnancies	0.124	0.019	6.585	0.000	0.087	0.161	1.397

We also computed the AIC and BIC for our model, which is 2173.037 and 2212.955. We will use these values to compare with other models in the following section.

Table 3: Model Summary Statistics

Null Deviance	df Null	Log Likelihood	AIC	BIC	Deviance	df Residual	Observations
2856.989	2213	-1079.518	2173.037	2212.955	2159.037	2207	2214

Exploring Interaction Effects

From our EDA, we saw prominent differences between each glucose level's proportion of diabetes by diabetes pedigree function levels, indicating a potential interaction effect. Therefore, we decided to fit a model including this effect. This model initially indicated VIF values of 17.07 and 19.29 for the diabetes pedigree function and its interaction term with age, which comes unsurprising given the known multicollinearity issues that arise when including interaction terms. (See Original Interaction Term Output).

For this reason, we decided to mean-center the data to address the multicollinearity (high VIF) introduced by adding the interaction term between Glucose and the Pedigree function. Mean-centering adjusts the predictors to have zero mean, which helps to minimize the correlation between them and their interaction terms. This adjustment is important for interpreting the interaction effects of DiabetesPedigreeFunction and Glucose without the confounding influence of high multicollinearity. Check Mathematical Explanation of Mean Center Reducing the Multi-Collinearity

Table 4: Model Summary for interaction Glucose_c x DiabetesPedigreeFunction_c

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-3.666	0.343	-10.700	0.000	-4.347	-3.003	NA
Glucose_c	0.033	0.002	16.526	0.000	0.029	0.037	1.114
BloodPressure	-0.010	0.003	-3.235	0.001	-0.015	-0.004	1.123
BMI	0.079	0.008	9.864	0.000	0.064	0.095	1.105
DiabetesPedigreeFunction_c	1.067	0.167	6.376	0.000	0.739	1.396	1.159
Age	0.013	0.005	2.485	0.013	0.003	0.024	1.444
Pregnancies	0.129	0.019	6.810	0.000	0.092	0.167	1.406
Glucose_c:DiabetesPedigreeFunction_c	-0.024	0.005	-5.171	0.000	-0.033	-0.015	1.242

Without multicollinearity, we then checked the AIC/BIC for this transformed model:

Table 5: Model Summary Statistics

Null Deviance	df Null	Log Likelihood	AIC	BIC	Deviance	df Residual	Observations
2856.989	2213	-1067.806	2151.611	2197.232	2135.611	2206	2214

Our AIC here is 2151.611 and our BIC is 2197.232. Our AIC here is lower than that of the model without the interaction term (2173.037), indicating that our new model with the interaction term fits the data better relative to its complexity. Our BIC here is also lower (compared to the previous model's 2212.955, showing that the interaction term likely adds value to the model without overfitting.

We then conducted a drop-in deviance test to see if this interaction term is necessary. From the test, we can see that the p value is 0, meaning we have evidence to suggest our mean-centered `Glucose_c*DiabetesPedigreeFunction_c` is statistically significant in predicting diabetes outcome.

Table 6: Drop-in-deviance-test

term	df.res	dev.res	df	deviance	p.value
Outcome ~ Glucose_c + BloodPressure + BMI + DiabetesPedigreeFunction_c + Age + Pregnancies	2207	2159.037	NA	NA	NA
Outcome ~ Glucose_c + BloodPressure + BMI + DiabetesPedigreeFunction_c + Age + Pregnancies + Glucose_c * DiabetesPedigreeFunction_c	2206	2135.611	1	23.425	0

We also evaluated other interaction effects including `Glucose_c*BMI_c`, `Glucose_c*BloodPressure_c`, `Glucose_c*Age_c`,`Pregnancies_c*Age_c`, `Pregnancies*DiabetesPedigreeFunction.` (See Interaction Term Model Testing). All of the interaction terms are significant, so we compared AIC and BIC values for models including each interaction effect.

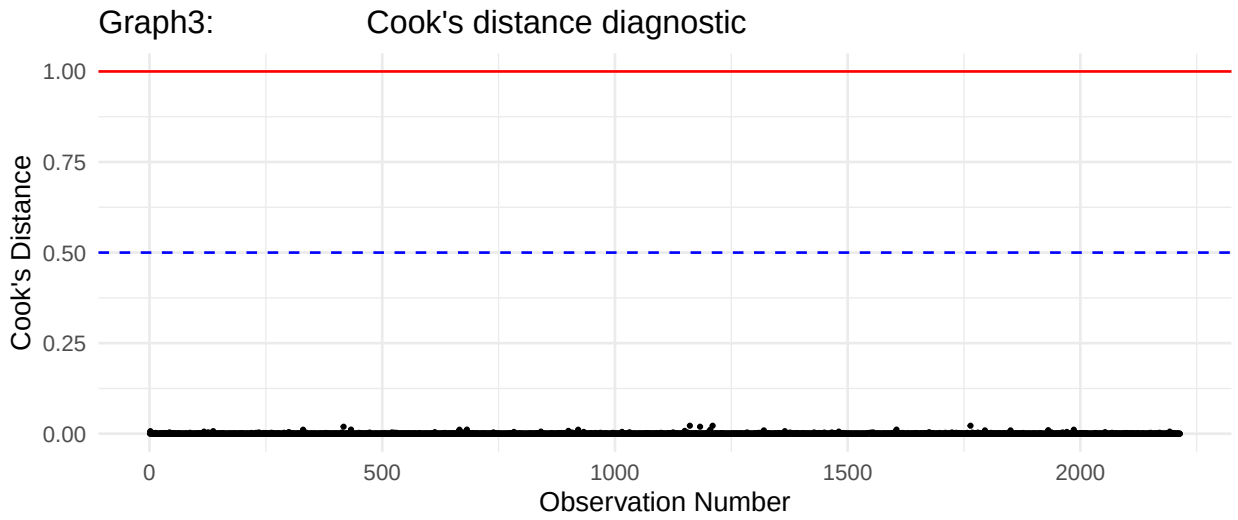
Table 7: Model Comparison Statistics

Model	AIC	BIC
Glucose_c*DiabetesPedigreeFunction_c	2151.611	2197.232
Glucose_c*Age_c	2154.896	2200.517
Glucose_c*BMI_c	2158.004	2203.624
Glucose_c*BloodPressure_c	2170.291	2215.291
Pregnancies_c*Age_c	2151.774	2197.395
Pregnancies*DiabetesPedigreeFunction	2173.835	2219.455

Our AIC/BIC table of all interaction models indicate that the model including an interaction between Pedigree Function and Glucose provides us the smallest AIC/BIC values. The inclusion of this variable is reasonable due to the well-documented relationship between a patient's genetic predisposition and glucose metabolism to diabetes risk. Thus, we believe that this model is ideal for testing in our project.

Cook's Distance

We calculated Cook's distance for each observation in our final model configuration to ensure the strength and reliability of our model.



As illustrated, none of the data points exceed the commonly used threshold of 0.5, indicating there are no influential points in this model. This suggests that there is no single data point that may skew the overall model, allowing us to conclude that our model is a good fit.

Results

Interpretation

Table 8: Model Interpretation

term	estimate	std.error	statistic	p.value
(Intercept)	-3.666	0.343	-10.700	0.000
Glucose_c	0.033	0.002	16.526	0.000
BloodPressure	-0.010	0.003	-3.235	0.001
BMI	0.079	0.008	9.864	0.000
DiabetesPedigreeFunction_c	1.067	0.167	6.376	0.000
Age	0.013	0.005	2.485	0.013
Pregnancies	0.129	0.019	6.810	0.000
Glucose_c:DiabetesPedigreeFunction_c	-0.024	0.005	-5.171	0.000

Our analysis confirms that all predictors except age in our model are statistically significant ($p < 0.05$), showing a mix of positive and negative effects on diabetes risk. As the number of pregnancies increases by 1 pregnancy, the odds of having diabetes increases by 13% ($e^{0.129}$) holding all else constant. Similarly, when the mean-centered diabetes pedigree function increases by 1 unit, the odds of having diabetes is multiplied by a factor of 2.90 ($e^{1.067}$) holding all else constant. BMI also plays a role in increasing the risk of diabetes, with BMI increasing the odds by about 8.2% ($e^{0.079}$) . Thus, our model indicates that women with more pregnancies, a stronger family predisposition, and higher BMI may be more at-risk of having diabetes.

On the other hand, Blood Pressure has a slight negative association with diabetes risk. This finding is particularly intriguing as it suggests that higher blood pressure may slightly reduce the risk of diabetes, counterintuitive to what might be expected. The magnitude of this effect is quite small (each unit increase in blood pressure is associated with only about a 1% decrease in diabetes odds), and examining our EDA plots reveals considerable variability in the blood pressure-diabetes relationship. This might suggest that this finding may reflect limitations in our dataset or unmeasured confounding factors. For instance, patients with known high blood pressure may be under closer medical supervision and lifestyle management, potentially affecting their diabetes risk, thus making our model having a negative coefficient. Additionally, our merged dataset from two different sources (Pima Indian and Frankfurt hospital populations) may introduce complexities in blood pressure measurement and recording that affect this relationship.

There is also a significantly negative interaction between Glucose_c and DiabetesPedigreeFunction_c (Glucose_c:DiabetesPedigreeFunction_c), suggesting that the effect of glucose on diabetes risk differs by family diabetes history; more specifically, as diabetes pedigree function goes up, the impact of glucose on diabetes risk slightly decreases, holding all else constant. This demonstrates how strong genetic predisposition to diabetes can lead to disease development even without severely elevated glucose levels.

Model Testing

We applied our model on the testing set and fit an ROC curve with threshold of 0.5 and calculated the area under curve. Our ROC curve's AUC value of 0.831 suggests that our model with mean-centered variables and interaction terms has good predictive ability in predicting diabetes status, correctly ranking a randomly chosen positive case higher than a randomly chosen negative case 83.1% of the time. The curve's shape also confirms a strong predictive performance.

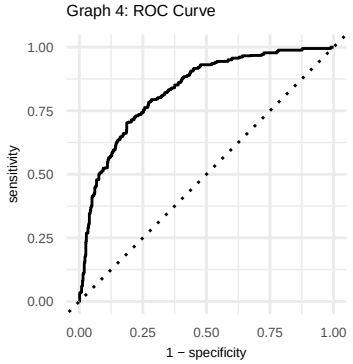


Table 9: Model ROC-AUC Score

Metric	Estimator	Estimate
roc_auc	binary	0.831

We then used a confusion matrix to summarize our model's predictability. We initially chose a threshold of 0.5, and arrived at an accuracy rating of 79.6%, sensitivity rate of 62.7%, and specificity

rate of 88.1%. The lower sensitivity rate indicates that only 62.7% of people with diabetes were correctly classified by our model as having diabetes.

Table 10: Confusion Matrix (Threshold: 0.5)

	Actual 0	Actual 1
0	325	69
1	44	116

Despite the high accuracy (proportion of individuals diabetes status correctly classified) and specificity rate (proportion of people without diabetes correctly classified), we decided to lower our threshold and chose 0.25. We believe that a higher sensitivity rate is more important than specificity because we prioritize correctly classifying patients with diabetes as having diabetes. Adopting this approach minimizes harm to potential patients, as misdiagnosing a patient without diabetes as diabetic, while not ideal, typically involves reversible risks once the error is identified. In contrast, incorrectly diagnosing a patient who actually has diabetes as healthy can delay necessary treatment, leading to severe complications and increased mortality risk.

Table 11: Confusion Matrix (Threshold: 0.25)

	Actual 0	Actual 1
0	235	24
1	134	161

By using a threshold of 0.25, our sensitivity rate increases to 87%. Our accuracy rate is now 71.5%, and our specificity rate is 63.69%. With this threshold, 87% of the people with diabetes are correctly classified as having diabetes.

Discussion

Summary

In our study, we examined how various factors affect the risk of diabetes by using a logistic regression model. Our findings highlight the significant role of genetic predisposition, as measured by the Diabetes Pedigree Function, and lifestyle factors like high BMI and number of pregnancies in increasing the likelihood of diabetes. Interestingly, our model revealed a negative association between blood pressure and diabetes risk, suggesting that higher blood pressure may slightly reduce the odds of diabetes. We also concluded that the negative interaction effect between glucose levels and genetic predisposition is significant, demonstrating the complexity of risk factors and points to the need for further research to fully understand the underlying mechanisms. Our approach prioritizes sensitivity in diagnosis and recognizes the severe implications of misdiagnosing diabetes, even if that signifies a slight decline in the accuracy of our model.

Limitations

Our study uses a dataset that merges information from the Pima population and the Frankfurt hospital, which introduces challenges related to the generalizability and consistency in measurements. While merging these data sources does increase the diversity and size of our sample, it increases the difficulty in comparing these measurements. For instance, differences in equipment or how tests like blood pressure and glucose levels are conducted might influence our results. These potential inconsistencies need to be considered, as they could limit how much we can generalize our findings to other populations or contexts.

In addition, even though the merged dataset is frequently cited by researchers, there is limited information on how the data from the Pima Indians and the Frankfurt hospital were combined. Without detailed documentation on the merging process, it's unclear whether the datasets were integrated using a perfectly consistent criteria. The lack of information also extends to how data normalization was conducted, especially for a binary outcome like diabetes, which is often diagnosed based on a range of measurements.

Finally, we excluded insulin level from our model because of the lack of data. As indicated by Joshi, insulin levels does play a role in the outcome of diabetes. Therefore, to address this variable, we may need to explore alternative modeling strategies or transformations of the insulin variable.

Future Research

In our model, we excluded Skin Thickness due to its high p value, as indicated in our initial model. This is the only variable that was removed from the dataset and is not a significant predictor of the likelihood of developing Type 2 diabetes in women. Although previous studies indicating that skin thickness was related to duration of diabetes, it would be worth it to further explore this variable's relationship with predicting an individual's diabetes outcome.

In addition, to address the negative association between blood pressure and diabetes outcomes identified in our study, future research involving a larger, more diverse dataset to validate our findings and take account into potential confounders may be essential. Assembling data that includes additional lifestyle factors, such as diet, physical activity, smoking habits, and alcoholic consumption can also enhance our understanding their impact on metabolic health and blood pressure, and thus, be able to reveal more complex relationships between predictors and diabetes outcome.

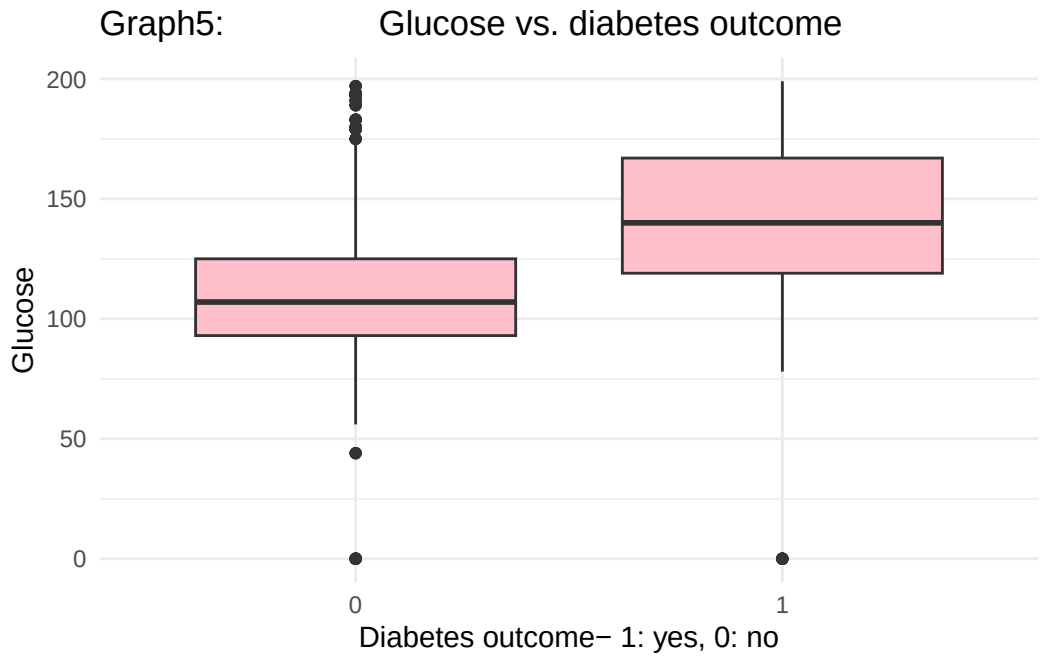
Because of our merged data set, by analyzing the data from the Pima Indians and the Frankfurt hospital independently, we can uncover more risk factors and health outcomes specific to each group. This approach can reveal how environmental, lifestyle, and genetic factors uniquely contribute to the diabetes risk, as these populations are very different, with the Pima peoples having one of the highest diabetes rate in the world while Germany being around the same as the global average.

Finally, analyzing longitudinal studies (using data collected for A1C tests, which measure average glucose over 2-3 months) would allow us to track the progression of diabetes over time, revealing how the disease develops and responds to different interventions. This would improve our understanding of both blood pressure-diabetes dynamics and long-term risk factor impacts, ultimately advancing personalized diabetes prevention and treatment strategies.

Appendix

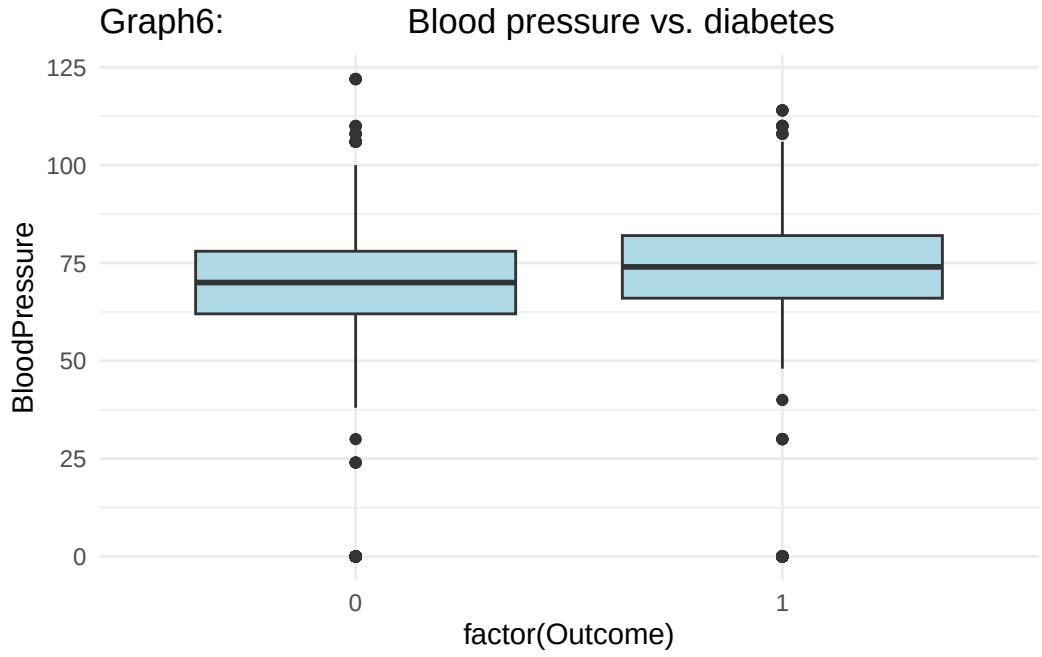
Response vs. explanatory EDA

Relationship between Outcome and Glucose:

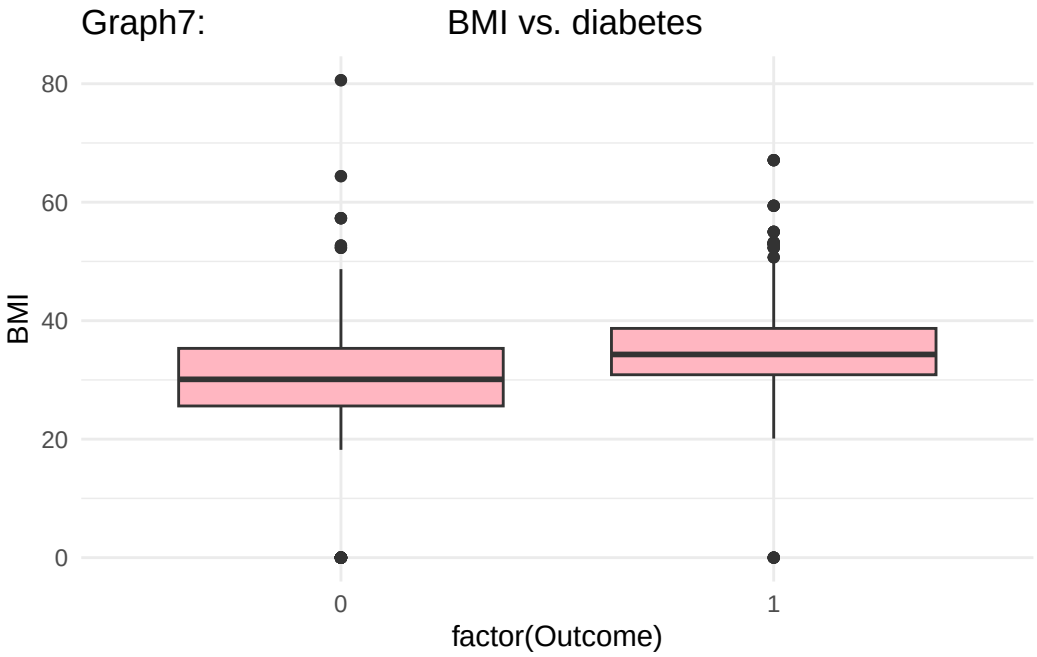


This could tell us that Glucose has a strong effect on having diabetes.

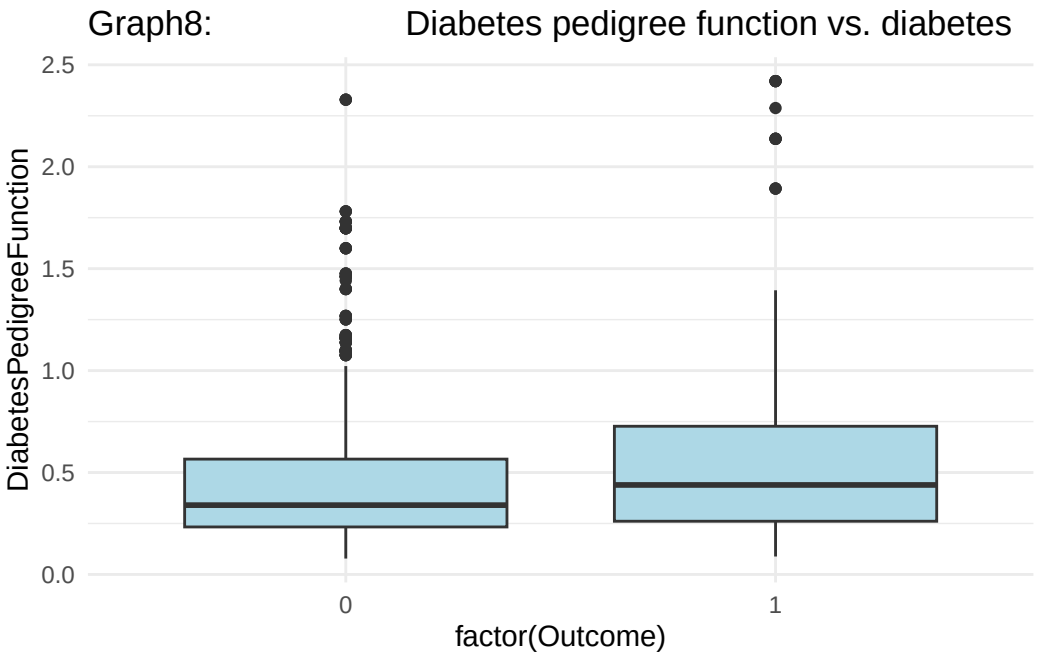
Relationship between Outcome and Blood Pressure:



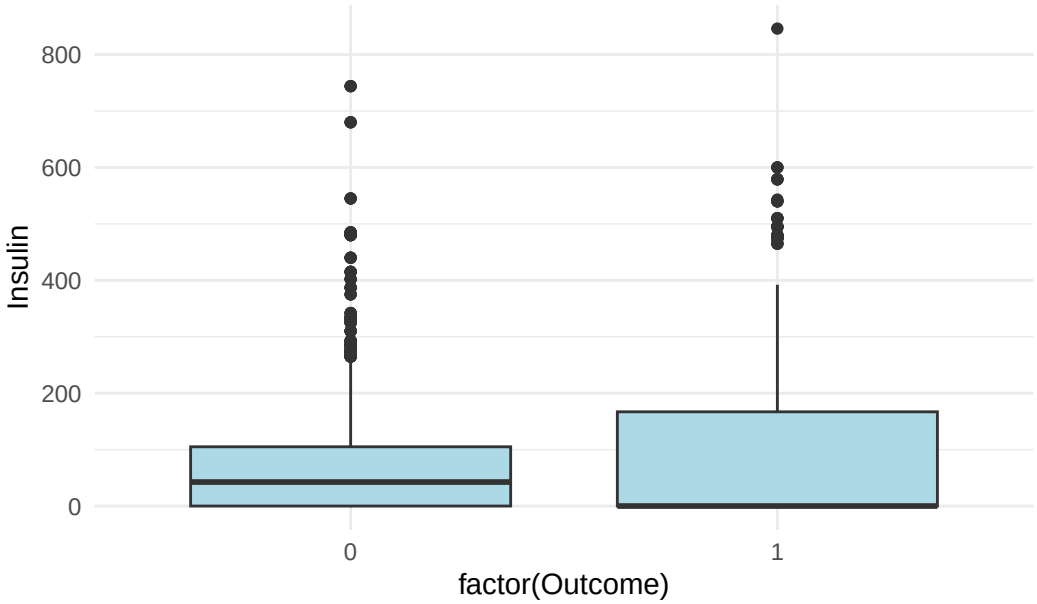
Relationship between Outcome and BMI:



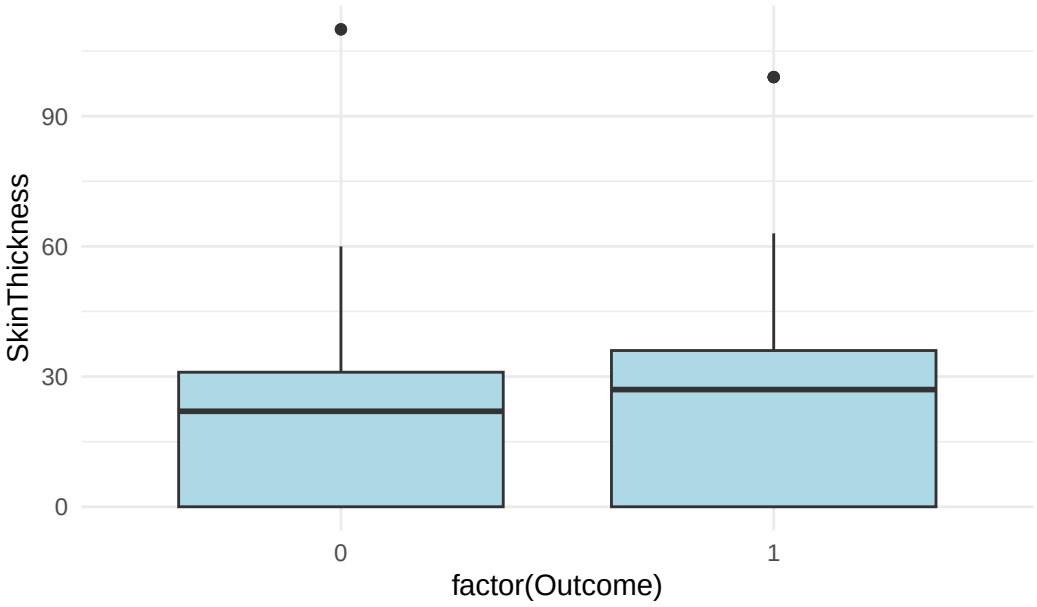
Relationship between Outcome and DiabetesPedigreeFunction:



Graph9: Insulin vs. diabetes

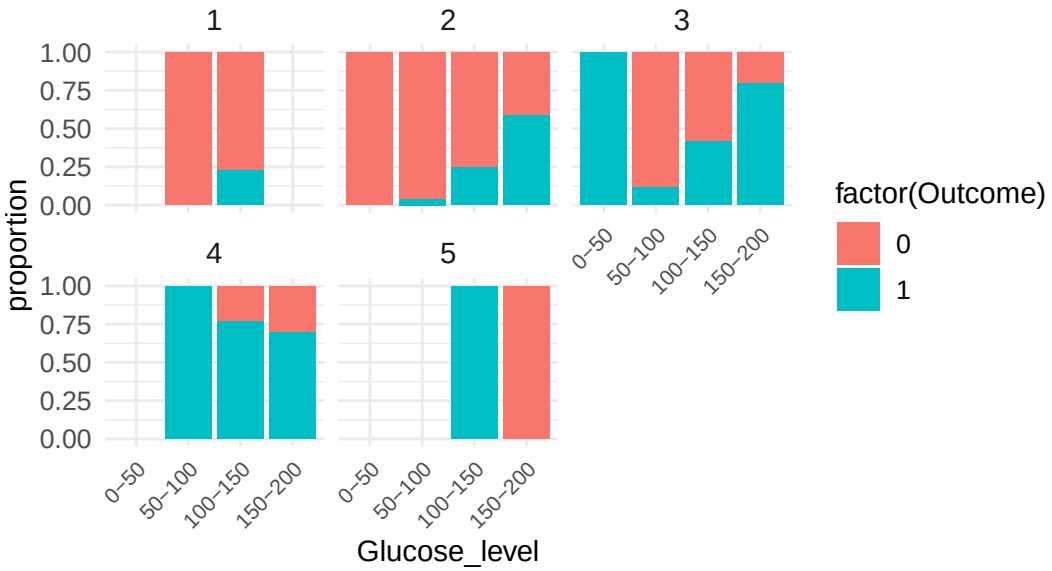


Graph10: Skin thickness vs. diabetes



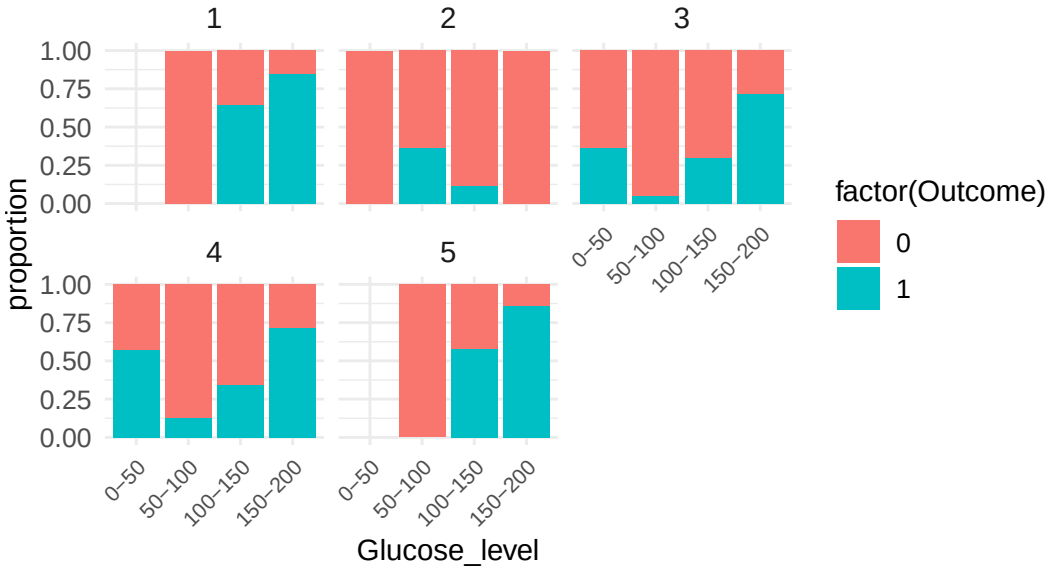
Interaction Explorations EDA

Graph11: Glucose level vs. diabetes by BMI level



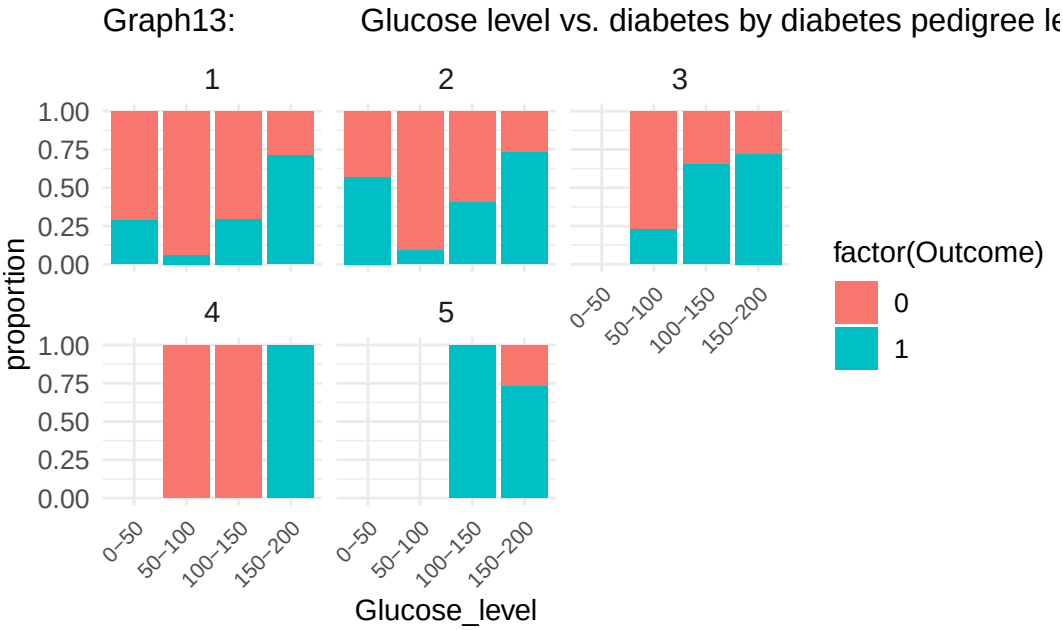
The proportion of individuals with diabetes varies noticeably across the BMI categories. For higher BMI levels (e.g., groups 4 and 5), the likelihood of diabetes increases more dramatically with higher glucose levels than it does in lower BMI levels (e.g., group 1).

Graph12: Glucose level vs. diabetes by blood pressure level



The proportions of outcomes differ depending on blood pressure levels. For certain blood pressure levels, the likelihood of diabetes (proportion of Outcome = 1) increases with higher glucose levels.

However, the pattern is not consistent across all blood pressure groups. This suggests that blood pressure levels may play a role in moderating the relationship between glucose levels and diabetes.



In groups 4 and 5 (higher diabetes pedigree levels), individuals with higher glucose levels (121–200) are overwhelmingly associated with diabetes (Outcome = 1). These groups show a striking shift in the proportion toward Outcome = 1 at higher glucose levels, as compared to more gradual increases in groups 1, 2, and 3, indicating a potential interaction effect.

Full Model

Table 12: Model Exploration

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-7.870	0.401	-19.632	0.000	-8.671	-7.099
Glucose	0.032	0.002	16.240	0.000	0.028	0.036
BloodPressure	-0.009	0.003	-3.014	0.003	-0.015	-0.003
BMI	0.078	0.008	9.333	0.000	0.062	0.094
DiabetesPedigreeFunction	0.854	0.173	4.940	0.000	0.517	1.195
Age	0.014	0.005	2.658	0.008	0.004	0.025
Pregnancies	0.124	0.019	6.573	0.000	0.087	0.161
SkinThickness	-0.003	0.004	-0.778	0.436	-0.010	0.004

Table 13: VIF for initial model

names	x
Glucose	1.046

names	x
BloodPressure	1.158
BMI	1.210
DiabetesPedigreeFunction	1.024
Age	1.474
Pregnancies	1.396
SkinThickness	1.244

Adding Interaction Terms

For each interaction term, we check the interaction term through adding it to the model, check the multicollinearity, mean-center the interaction variables, and then conduct a drop-in deviance test.

Original Interaction Term Output

Table 14: Model Summary for interaction Glucose x DiabetesPedigreeFunction

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-9.584	0.549	-17.452	0.000	-10.678	-8.524	NA
Glucose	0.045	0.003	13.405	0.000	0.038	0.051	3.044
BloodPressure	-0.010	0.003	-3.235	0.001	-0.015	-0.004	1.123
BMI	0.079	0.008	9.864	0.000	0.064	0.095	1.105
DiabetesPedigreeFunction	3.977	0.643	6.189	0.000	2.702	5.229	17.076
Age	0.013	0.005	2.485	0.013	0.003	0.024	1.444
Pregnancies	0.129	0.019	6.810	0.000	0.092	0.167	1.406
Glucose:DiabetesPedigreeFunction	-0.024	0.005	-5.171	0.000	-0.033	-0.015	19.279

Mathematical Explanation of Mean Center Reducing the Multi-Collinearity

Consider a model with variables X_1, X_2, and their interaction X_1X_2. The model can be written as:

$y \sim X_1 + X_2 + X_1X_2$

After mean-centering, each variable becomes its deviation from the mean:

$y \sim (X_1 - \bar{X_1}) + (X_2 - \bar{X_2}) + (X_3 - \bar{X_3})$

where $X_3 = X_1X_2$

The correlation between mean-centered variables can be expressed as:

$\text{Corr}(X_1, X_2) = \text{Cor}(X_1 - \bar{X_1}, X_2 - \bar{X_2})$

For the interaction term:

$$\text{Corr}(X_1, X_1X_2) = \text{Cor}(X_1 - \bar{X_1}, X_1X_2 - \overline{X_1X_2})$$

$$\neq \text{Cor}(X_1 - \bar{X_1}, (X_1 - \bar{X_1})(X_2 - \bar{X_2}))$$

This transformation reduces multicollinearity because the centered interaction term $(X_1X_2 - \overline{X_1X_2})$ is less correlated with the individual centered variables $(X_1 - \bar{X_1})$ and $(X_2 - \bar{X_2})$ than in the uncentered case.

Interaction Term Model Testing

Glucose X BMI

Checking the interaction effect between Glucose and BMI.

Table 15: Model Summary for interaction Glucose x BMI

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-12.898	1.306	-9.876	0.000	-15.473	-10.354	NA
Glucose	0.073	0.010	7.244	0.000	0.053	0.092	27.236
BloodPressure	-0.010	0.003	-3.356	0.001	-0.016	-0.004	1.118
BMI	0.225	0.037	6.090	0.000	0.153	0.298	24.229
DiabetesPedigreeFunction	0.908	0.172	5.281	0.000	0.573	1.247	1.019
Age	0.012	0.005	2.260	0.024	0.002	0.023	1.455
Pregnancies	0.128	0.019	6.724	0.000	0.091	0.165	1.405
Glucose:BMI	-0.001	0.000	-4.212	0.000	-0.002	-0.001	53.422

Transforming the variable through mean-centering.

Table 16: Model Summary for interaction Gluocose_c x BMI_c

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-1.456	0.252	-5.774	0.000	-1.952	-0.963	NA
Glucose_c	0.034	0.002	16.351	0.000	0.030	0.039	1.205
BloodPressure	-0.010	0.003	-3.356	0.001	-0.016	-0.004	1.118
BMI_c	0.082	0.008	10.104	0.000	0.066	0.098	1.171
DiabetesPedigreeFunction	0.908	0.172	5.281	0.000	0.573	1.247	1.019
Age	0.012	0.005	2.260	0.024	0.002	0.023	1.455
Pregnancies	0.128	0.019	6.724	0.000	0.091	0.165	1.405
Glucose_c:BMI_c	-0.001	0.000	-4.212	0.000	-0.002	-0.001	1.272

Table 17: Model Summary Statistics

Null Deviance	df Null	Log Likelihood	AIC	BIC	Deviance	df Residual	Observations
2856.989	2213	-1071.002	2158.004	2203.624	2142.004	2206	2214

Table 18: Drop-in-deviance-test

term	df.res	dev.res	df	deviance	p.value
Outcome ~ Glucose_c + BloodPressure + BMI_c + DiabetesPedigreeFunction + Age + Pregnancies	2207	2159.037	NA	NA	NA
Outcome ~ Glucose_c + BloodPressure + BMI_c + DiabetesPedigreeFunction + Age + Pregnancies + Glucose_c * BMI_c	2206	2142.004	1	17.033	0

Glucose X Blood Pressure

Checking the interaction effect between Glucose and BloodPressure:

BloodPressure is no longer statistically significant here, probably due to the high multi-collinearity.

Table 19: Model Summary for interaction Glucose x BloodPressure

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-10.085	1.152	-8.758	0.000	-12.424	-7.909	NA
Glucose	0.050	0.009	5.624	0.000	0.033	0.068	21.340
BloodPressure	0.023	0.015	1.461	0.144	-0.007	0.054	28.712
BMI	0.075	0.008	9.380	0.000	0.059	0.090	1.094
DiabetesPedigreeFunction	0.840	0.171	4.910	0.000	0.506	1.177	1.005
Age	0.014	0.005	2.748	0.006	0.004	0.025	1.445
Pregnancies	0.125	0.019	6.649	0.000	0.088	0.162	1.394
Glucose:BloodPressure	0.000	0.000	-2.111	0.035	0.000	0.000	54.072

Table 20: Model Summary for interaction Glucose_c x BloodPressure_c

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-4.580	0.346	-13.245	0.000	-5.267	-3.911	NA
Glucose_c	0.033	0.002	16.199	0.000	0.029	0.037	1.093
BloodPressure_c	-0.008	0.003	-2.492	0.013	-0.014	-0.002	1.188
BMI	0.075	0.008	9.380	0.000	0.059	0.090	1.094
DiabetesPedigreeFunction	0.840	0.171	4.910	0.000	0.506	1.177	1.005
Age	0.014	0.005	2.748	0.006	0.004	0.025	1.445
Pregnancies	0.125	0.019	6.649	0.000	0.088	0.162	1.394
Glucose_c:BloodPressure_c	0.000	0.000	-2.111	0.035	0.000	0.000	1.119

After the transformation, BloodPressure_c is significant again.

Table 21: Model Summary Statistics

Null Deviance	df Null	Log Likelihood	AIC	BIC	Deviance	df Residual	Observations
2856.989	2213	-1077.145	2170.291	2215.911	2154.291	2206	2214

Table 22: Drop-in-deviance-test

term	df.res	dev.res	df	deviance	p.value
Outcome ~ Glucose_c + BloodPressure_c + BMI + DiabetesPedigreeFunction + Age + Pregnancies	2207	2159.037	NA	NA	NA
Outcome ~ Glucose_c + BloodPressure_c + BMI + DiabetesPedigreeFunction + Age + Pregnancies + Glucose_c * BloodPressure_c	2206	2154.291	1	4.746	0.029

Glucose X Age

Checking the interaction between the Glucose and Age

Table 23: Model Summary for Glucose x Age

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-10.968	0.816	-13.440	0.000	-12.586	-9.385	NA
Glucose	0.057	0.006	9.511	0.000	0.045	0.069	9.749
BloodPressure	-0.009	0.003	-3.183	0.001	-0.015	-0.004	1.125
BMI	0.072	0.008	9.087	0.000	0.057	0.088	1.098
DiabetesPedigreeFunction	0.801	0.173	4.635	0.000	0.464	1.142	1.007
Age	0.108	0.021	5.149	0.000	0.066	0.148	23.732
Pregnancies	0.116	0.019	6.243	0.000	0.080	0.152	1.372
Glucose:Age	-0.001	0.000	-4.586	0.000	-0.001	0.000	35.554

Table 24: Model Summary for Glucose_c x Age_c

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-3.314	0.321	-10.326	0.000	-3.951	-2.693	NA
Glucose_c	0.034	0.002	16.467	0.000	0.030	0.038	1.148
BloodPressure	-0.009	0.003	-3.183	0.001	-0.015	-0.004	1.125
BMI	0.072	0.008	9.087	0.000	0.057	0.088	1.098
DiabetesPedigreeFunction	0.801	0.173	4.635	0.000	0.464	1.142	1.007
Age_c	0.023	0.005	4.193	0.000	0.012	0.033	1.585
Pregnancies	0.116	0.019	6.243	0.000	0.080	0.152	1.372
Glucose_c:Age_c	-0.001	0.000	-4.586	0.000	-0.001	0.000	1.292

Table 25: Model Summary Statistics

Null Deviance	df Null	Log Likelihood	AIC	BIC	Deviance	df Residual	Observations
2856.989	2213	-1069.448	2154.896	2200.517	2138.896	2206	2214

Table 26: Drop-in-deviance-test

term	df.res	dev.res	df	deviance	p.value
Outcome ~ Glucose_c + BloodPressure + BMI + DiabetesPedigreeFunction + Age_c + Pregnancies	2207	2159.037	NA	NA	NA
Outcome ~ Glucose_c + BloodPressure + BMI + DiabetesPedigreeFunction + Age_c + Pregnancies + Glucose_c * Age_c	2206	2138.896	1	20.141	0

Pregnancies X Age

check the interaction between pregnancy and age:

Table 27: Model Summary for interaction Age x Pregnancies

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-8.989	0.478	-18.796	0.000	-9.945	-8.069	NA
Glucose	0.032	0.002	16.294	0.000	0.028	0.036	1.051
BloodPressure	-0.009	0.003	-2.978	0.003	-0.014	-0.003	1.128
BMI	0.078	0.008	9.717	0.000	0.062	0.093	1.106
DiabetesPedigreeFunction	0.835	0.172	4.849	0.000	0.499	1.175	1.006
Age	0.044	0.008	5.481	0.000	0.029	0.060	3.340
Pregnancies	0.418	0.065	6.442	0.000	0.292	0.547	16.305
Age:Pregnancies	-0.008	0.002	-4.758	0.000	-0.011	-0.004	23.102

Table 28: Model Summary for interaction Age_c x Pregnancies_c

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-6.900	0.380	-18.168	0.000	-7.659	-6.169	NA
Glucose	0.032	0.002	16.294	0.000	0.028	0.036	1.051
BloodPressure	-0.009	0.003	-2.978	0.003	-0.014	-0.003	1.128
BMI	0.078	0.008	9.717	0.000	0.062	0.093	1.106
DiabetesPedigreeFunction	0.835	0.172	4.849	0.000	0.499	1.175	1.006
Age_c	0.017	0.005	3.149	0.002	0.006	0.027	1.409
Pregnancies_c	0.168	0.021	7.994	0.000	0.127	0.210	1.716
Age_c:Pregnancies_c	-0.008	0.002	-4.758	0.000	-0.011	-0.004	1.415

Table 29: Model Summary Statistics

Null Deviance	df Null	Log Likelihood	AIC	BIC	Deviance	df Residual	Observations
2856.989	2213	-1067.887	2151.774	2197.395	2135.774	2206	2214

Table 30: Drop-in-deviance-test

term	df.res	dev.res	df	deviance	p.value
Outcome ~ Glucose + BloodPressure + BMI + DiabetesPedigreeFunction + Age_c + Pregnancies_c	2207	2159.037	NA	NA	NA
Outcome ~ Glucose + BloodPressure + BMI + DiabetesPedigreeFunction + Age_c + Pregnancies_c + Pregnancies_c * Age_c	2206	2135.774	1	23.263	0

Pregnancies X DiabetesPedigreeFunction

check the interaction effect between Pregnancies and DiabetesPedigreeFunction

Table 31: Model Summary for interaction DiabetesPedigreeFunction x Pregnancies

term	estimate	std.Error	statistic	p.value	conf.low	conf.high	VIF
(Intercept)	-7.747	0.407	-19.023	0.000	-8.560	-6.964	NA
Glucose	0.032	0.002	16.227	0.000	0.028	0.036	1.046
BloodPressure	-0.010	0.003	-3.256	0.001	-0.015	-0.004	1.127
BMI	0.076	0.008	9.602	0.000	0.061	0.092	1.094
DiabetesPedigreeFunction	0.623	0.259	2.408	0.016	0.119	1.133	2.287
Age	0.015	0.005	2.862	0.004	0.005	0.025	1.450
Pregnancies	0.096	0.032	3.009	0.003	0.034	0.159	3.973
DiabetesPedigreeFunction:Pregnancies	0.057	0.052	1.092	0.275	-0.045	0.161	4.567

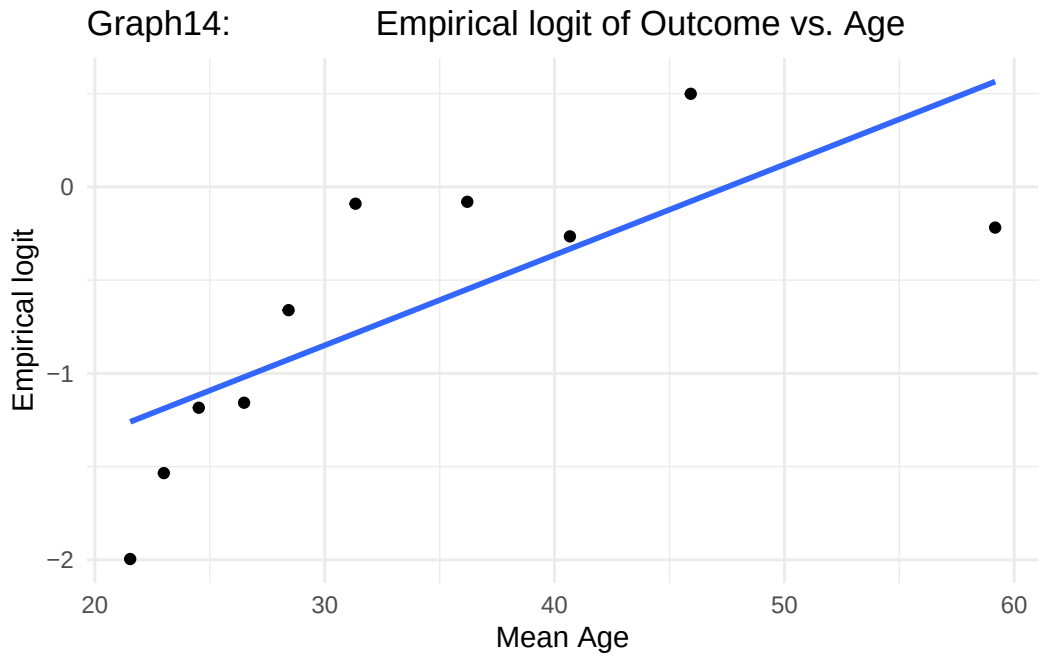
Table 32: Model Summary Statistics

Null Deviance	df Null	Log Likelihood	AIC	BIC	Deviance	df Residual	Observations
2856.989	2213	-1078.917	2173.835	2219.455	2157.835	2206	2214

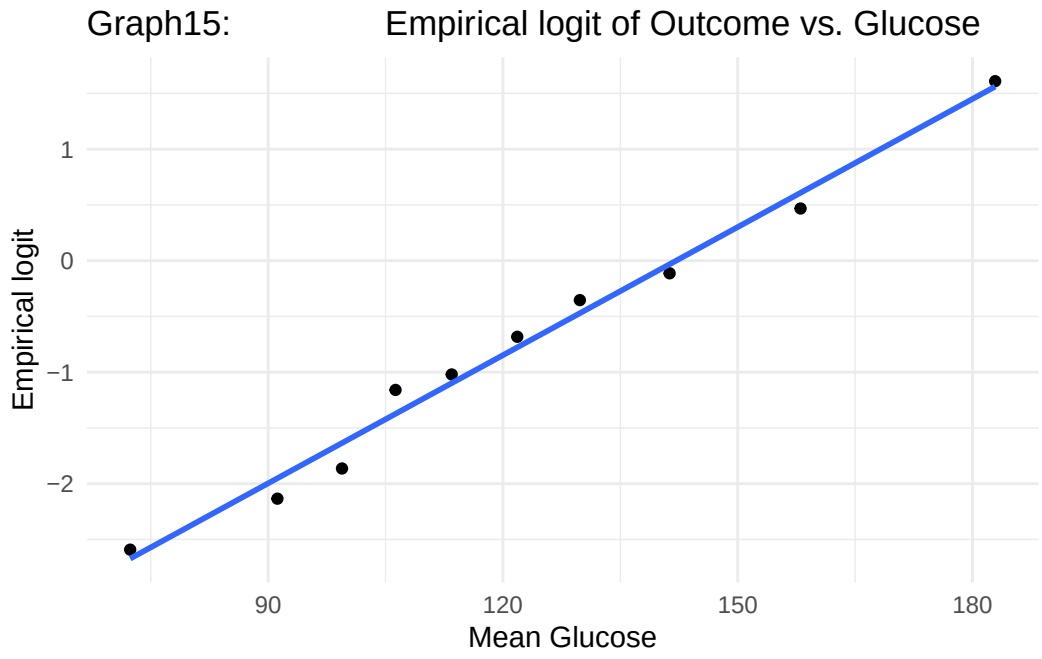
Testing linearity assumption

Use empirical logit graph between outcome and age:

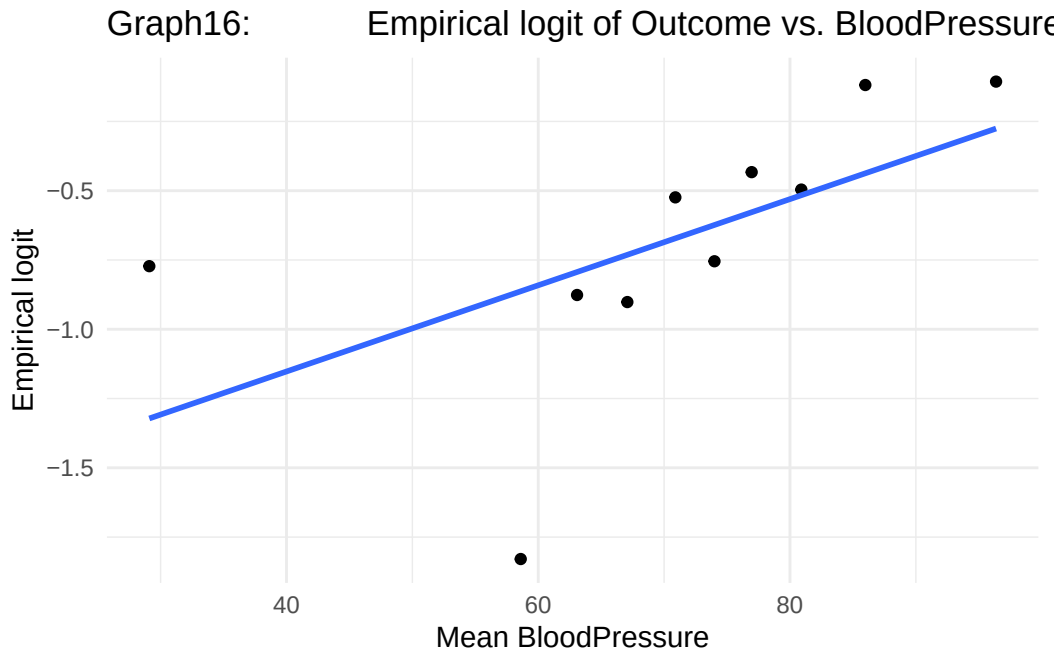
Age vs. Outcome



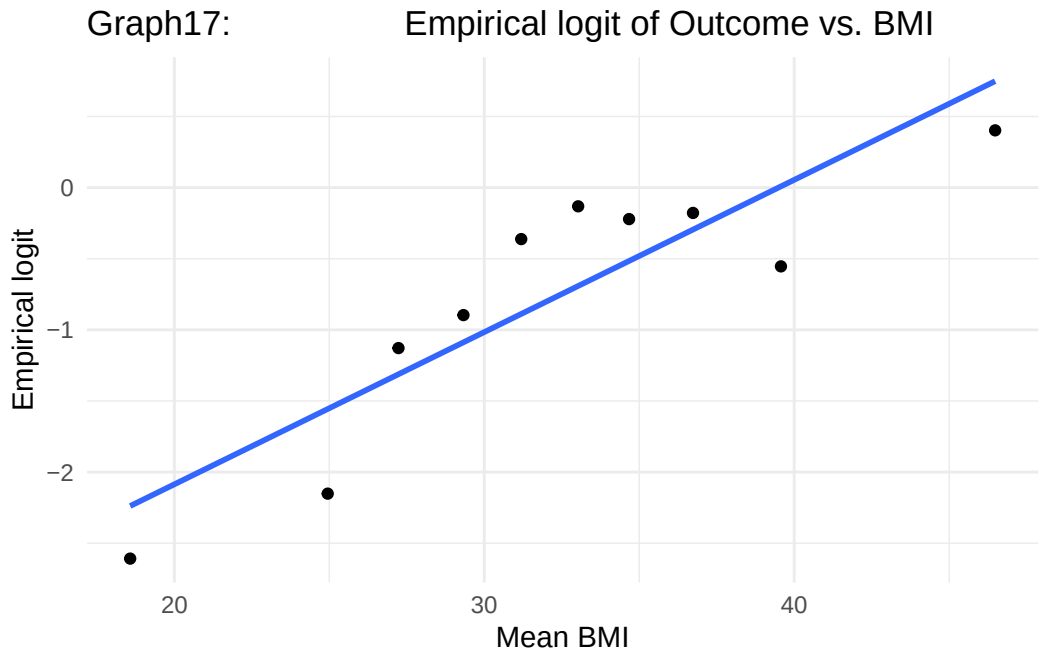
Glucose vs. Outcome



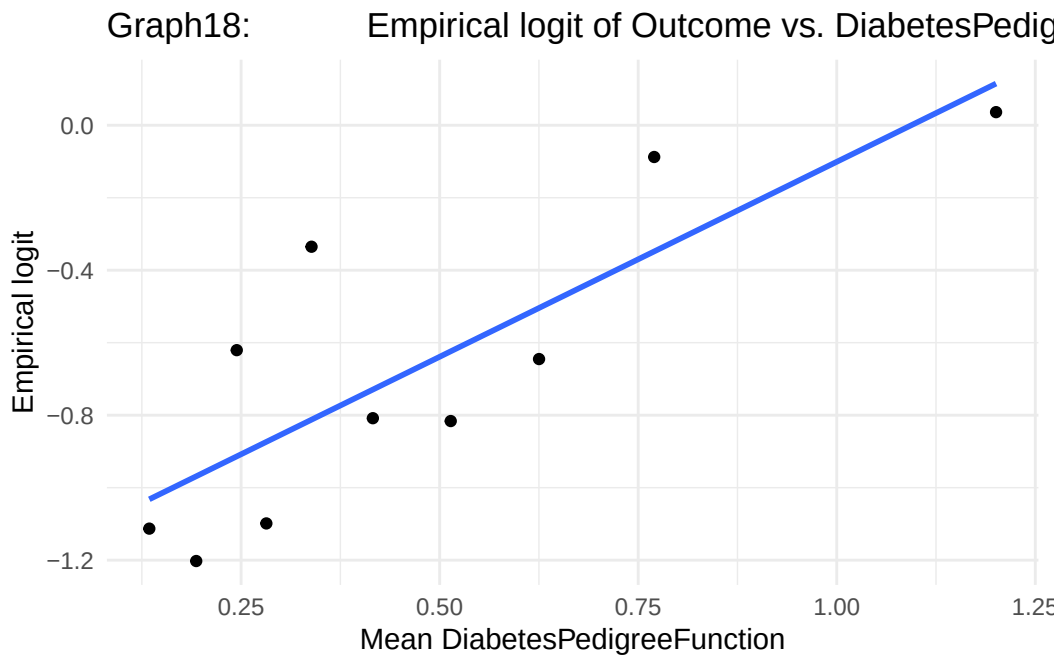
BloodPressure vs. Outcome:



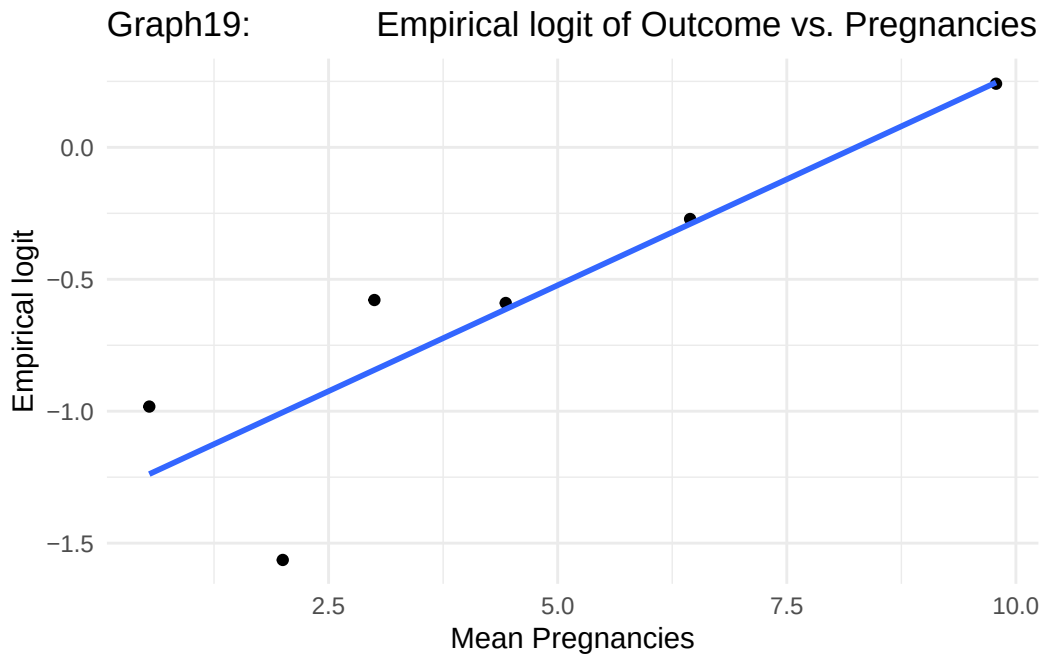
BMI vs. Outcome



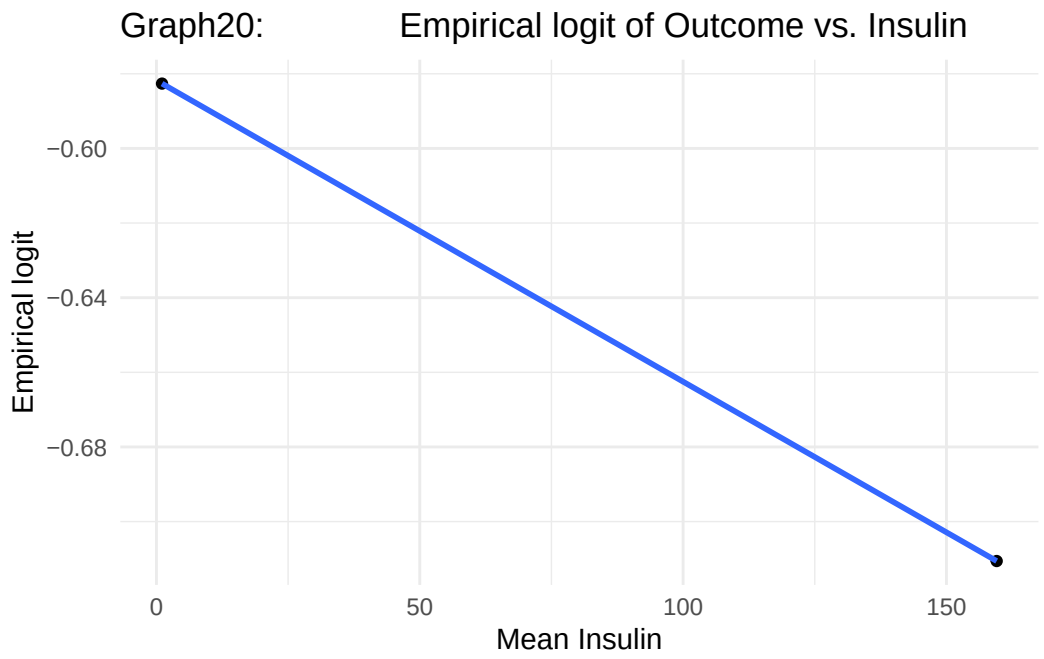
DiabetesPedigreeFunction vs. Outcome:



Pregnancies vs. Outcome:



Insulin vs. Outcome:



There is not enough data for insulin in the data set to assume a linear relationship.

```
[1] 1330
```

1,330 of the `Insulin` data points are 0, making it heavily skewed.

Citations

Boer, Ian H. de, Sripal Bangalore, Athanase Benetos, Andrew M. Davis, Erin D. Michos, Paul Muntner, Peter Rossing, Sophia Zoungas, and George Bakris. 2017. "Diabetes and Hypertension: A Position Statement by the American Diabetes Association." *Diabetes Care* 40 (9): 1273–84. https://doi.org/10.2337/dci17-0026.

Bullard, Kai McKeever, Catherine C. Cowie, Sarah E. Lessem, Sharon H. Saydah, Andy Menke, Linda S. Geiss, Trevor J. Orchard, Deborah B. Rolka, and Giuseppina Imperatore. 2018. "Prevalence of Diagnosed Diabetes in Adults by Diabetes Type — United States, 2016." *MMWR. Morbidity and Mortality Weekly Report* 67 (12): 359–61. https://doi.org/10.15585/mmwr.mm6712a2.

Collier, Andrew, Alan W Patrick, Derek Bell, David M Matthews, Cecilia C A Maclntyre, David J Ewing, and Basil F Clarke. 1989. "Relationship of Skin Thickness to Duration of Diabetes, Glycemic Control, and Diabetic Complications in Male IDDM Patients." *Diabetes Care* 12 (5): 309–12. https://doi.org/10.2337/diacare.12.5.309.

Elbein, Steven C, Teresa M Maxwell, and Mary Catherine Schumacher. 1991. "Insulin and Glucose Levels and Prevalence of Glucose Intolerance in Pedigrees With Multiple Diabetic Siblings." *Diabetes* 40 (8): 1024–32. https://doi.org/10.2337/diab.40.8.1024.

Joshi, Ram D., and Chandra K. Dhakal. 2021. "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches." *International Journal of Environmental Research and Public Health* 18 (14): 7346. https://doi.org/10.3390/ijerph18147346.

Kolb, Hubert, and Stephan Martin. 2017. "Environmental/Lifestyle Factors in the Pathogenesis and Prevention of Type 2 Diabetes." *BMC Medicine* 15 (1). https://doi.org/10.1186/s12916-017-0901-x.

Yan, Zihui, Mengjie Cai, Xu Han, Qingguang Chen, and Hao Lu. 2023. "The Interaction Between Age and Risk Factors for Diabetes and Prediabetes: A Community-Based Cross-Sectional Study." *Diabetes, Metabolic Syndrome and Obesity* Volume 16 (January): 85–93. https://doi.org/10.2147/dmso.s390857.