Stat 171 Final Project Report

Eve Fraczkiewicz, Sai Shrinidi Krishnan, Seona Magdum, William Pham, Tianle Qi

March 20, 2025

University of California Riverside

Introduction:

In 2021, 38.4 million Americans, or 11.6% of the population, were diagnosed with diabetes. It was also listed as the eighth leading cause of death in the United States (Statistics About Diabetes | ADA). It is clear we are facing a health crisis, so it is important to identify factors that lead to diabetes and accurately predict a person's diagnosis.

Our goal for this project is to investigate factors that may contribute to a diabetes diagnosis by fitting a generalized linear model with various lifestyle and health variables and interpreting these predictor coefficients. We aim to determine the least and most significant factors for predicting a diabetes diagnosis to further our understanding of how to detect and mitigate this disease.

Data Description:

The dataset, derived from the National Health and Nutrition Examination Survey (NHANES), contains 300 observations with nine variables. The response variable is a binary variable depicting diabetes status (1 for diagnosed and 0 for not). The predictor variables are four numeric variables, age, BMI, systolic blood pressure, and cholesterol, and four categorical variables, gender (male or female), smoker status (1 for smokes, 0 for doesn't smoke), family history of diabetes (1 for family history 0 for no family history), and physical activity status (1 for is physically active 0 for is not).

Data Cleaning:

The gender column contained "Male" and "Female" string values, so we replaced them with 0 and 1, respectively, for use in our model. There were no N/A or missing values, so we used all 300 observations.

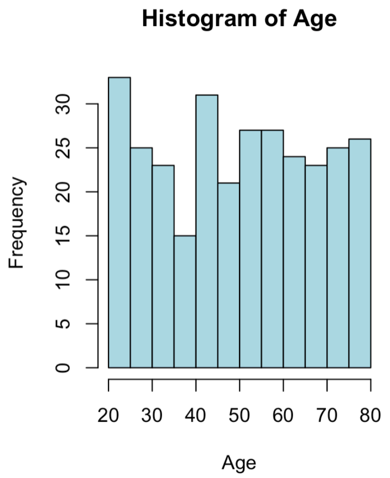
Summary Statistics:

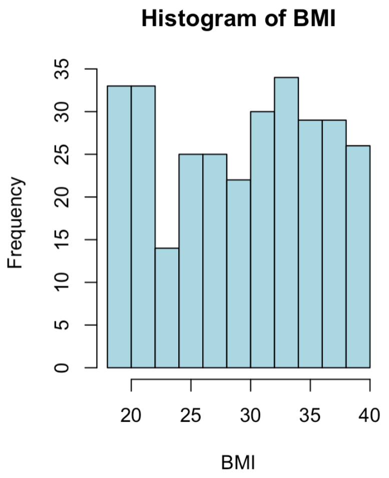
Statistic	Age	BMI	Smoker	Physical Activity	Diabetes Status	Cholesterol	Systolic BP	Family History
Minimum	20	18.11	0	0	0	151.1	90.29	0
1st Qua.	34	23.43	0	0	0	184.2	113.23	0
Median	51	29.79	0	1	1	223.1	137.42	0
Mean	49.93	29.1	0.26	0.6233	0.5333	222.8	136	0.4067
3rd Qua.	64	34.58	1	1	1	259.7	158.6	1
Maximum	79	39.79	1	1	1	299.5	179.85	1
Correlation*	0.474	0.153	0.113	0.024	0.009	-0.018	-0.041	-0.092

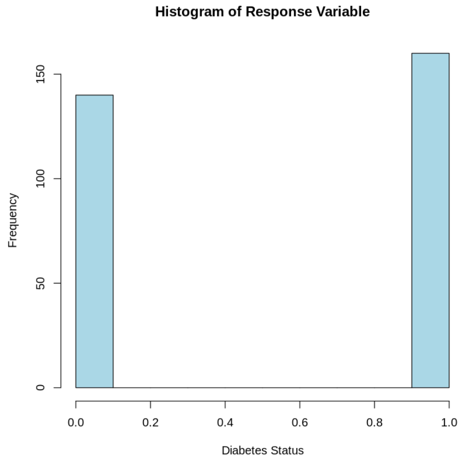
* Correlation of each variable with Diabetes Status. Highlighted correlation values indicate significant correlations.

The table above shows that the age and BMI ranges are pretty broad, indicating that the sample collected is diverse. The means show that most subjects did not smoke and were physically active. A little under half of the subjects had a family history of diabetes, and about half of the subjects were diagnosed with diabetes.

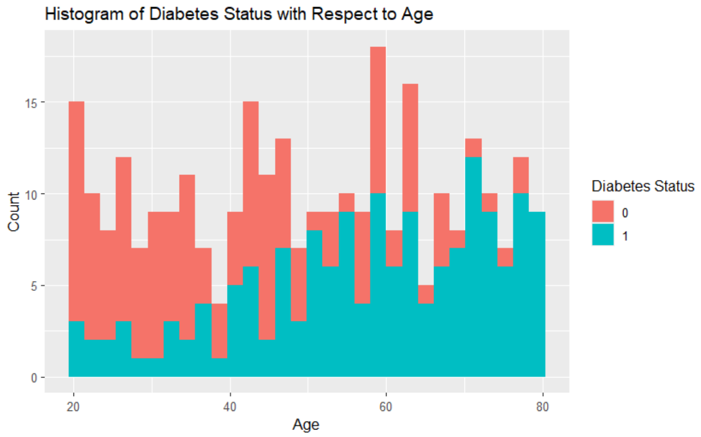
Exploratory Data Analysis:

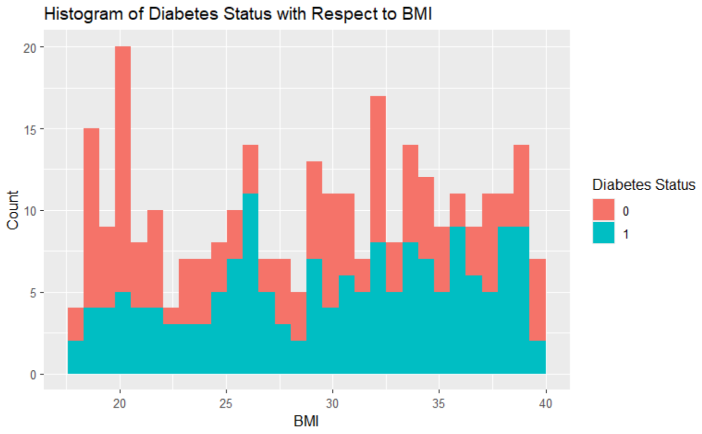




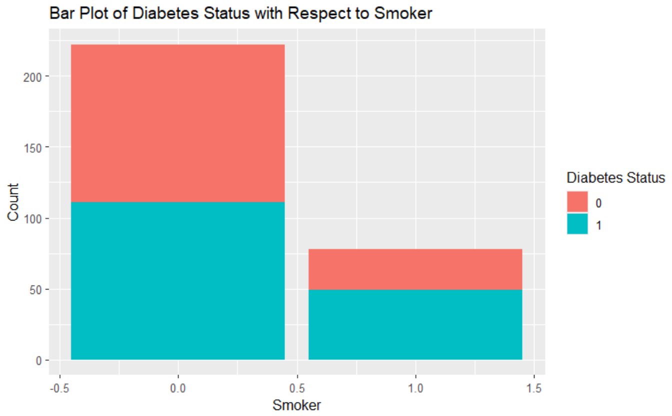


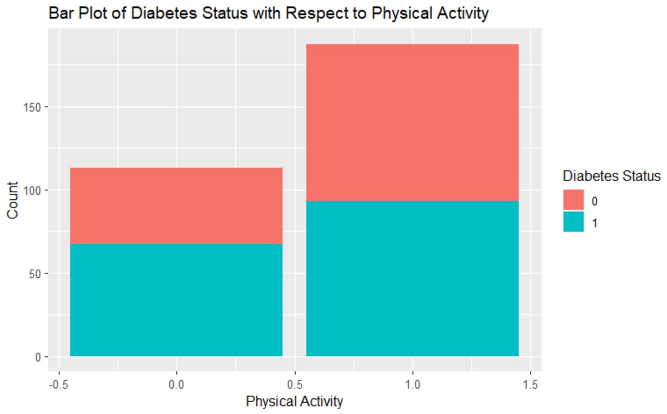
The histogram for age reveals that almost every age range was represented equally within the data as most of the ranges had around 20 to 25 observations in each, except the 35 to 40 age range with around 15 observations, the 20 to 25 age range with around 35 observations, and the 40 to 45 age range with around 30 observations. The BMI histogram seems to have a weak bimodal shape, revealing more observations at the lower and higher ends of BMI, with fewer observations being seen from 17 to 30 (especially at 17). The histogram of our response variable shows that a little over half of the subjects were diagnosed with diabetes, but the split between them is still relatively even.





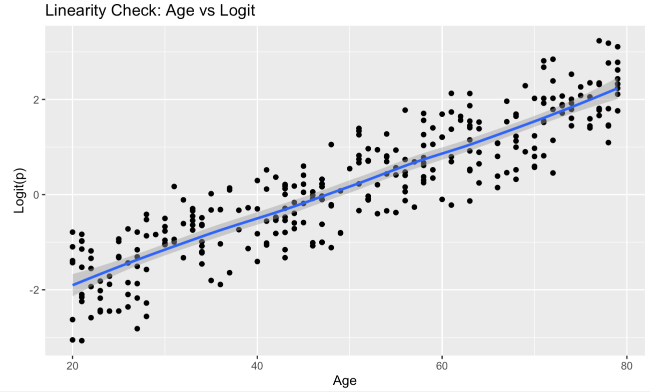
Taking a look at Diabetes Status compared to Age and BMI, we see that older individuals and those with higher BMI tend to have diabetes compared to younger people and those with a lower BMI.



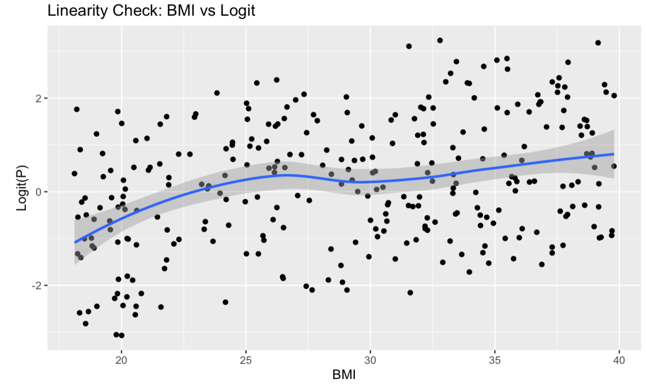


The split between those diagnosed with diabetes and those who have not is somewhat unbalanced for Smokers and Physical Activity. It's not severe enough to warrant under/oversampling, but we should consider it when creating and interpreting our model.

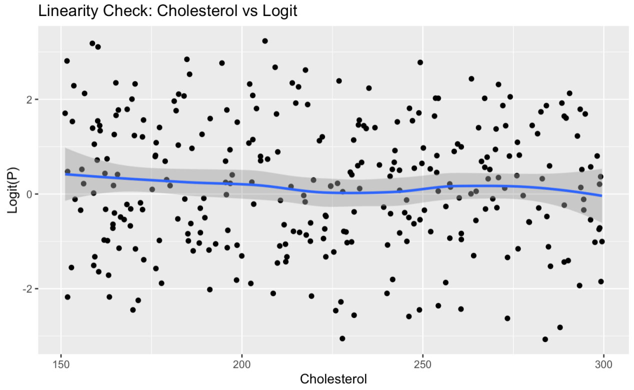
Linearity:



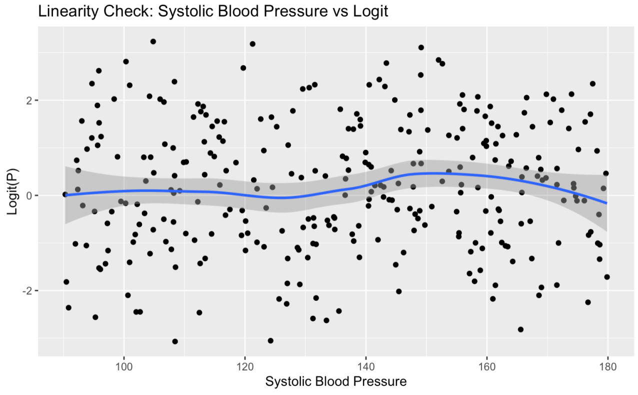
Based on the logit versus age plot, the linearity assumption for age seems to be satisfied, as the loess curve looks straight with no curvature. This suggests that no higher-order terms for age will be needed in the final model.



The logit versus BMI plot suggests that the linearity assumption for BMI is violated, as the Loess curve shows some curvature. The way the Loess curve looks suggests that a cubed term of BMI could potentially be included in the final model.

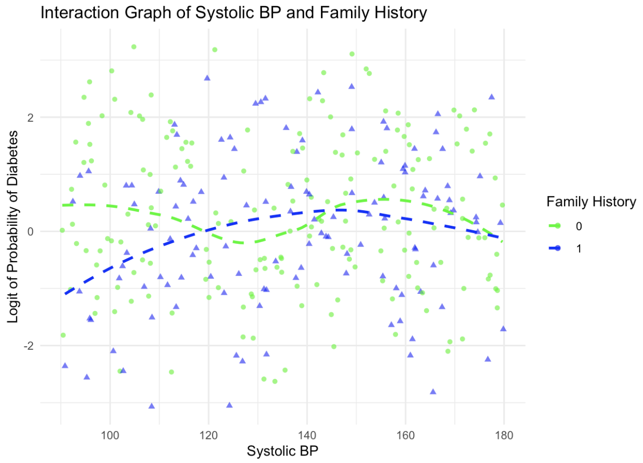


The logit versus cholesterol plot seems to satisfy the linearity assumption, as the Loess curve looks mostly straight. This suggests that no higher-order terms for cholesterol will be needed in the final model.

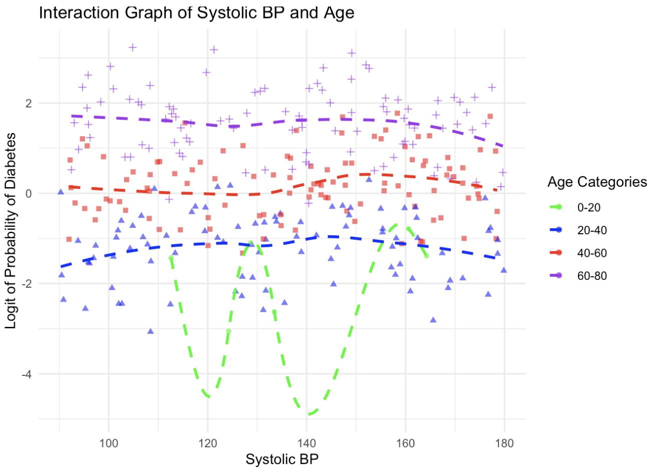


The logit versus systolic blood pressure plot suggests that the linearity assumption has been violated, as the Loess curve shows some curvature. This suggests that a higher-order term for systolic blood pressure will be needed in the final model.

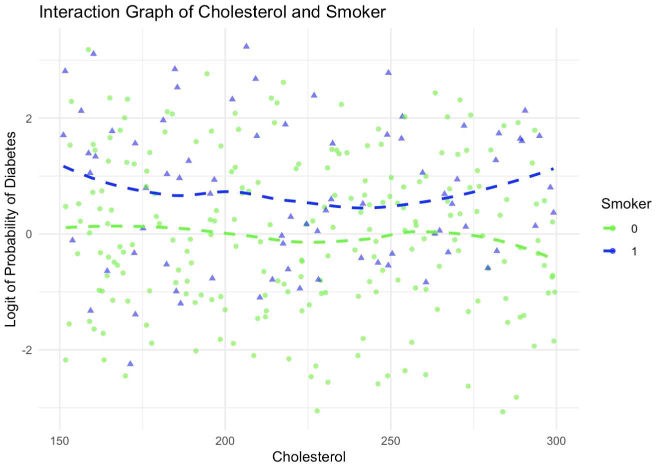
Model Structure:



The plot shows no interaction between family history and systolic blood pressure. The Loess curves are in similar areas and follow similar trends, meaning there is no difference in systolic blood pressure based on family history. This suggests that the final model should not include an interaction term between family history and systolic blood pressure.



Based on the plot, we can see an interaction between age and systolic blood pressure as the loess curves for the different age ranges are in other areas and follow different trends. This suggests that the final model could include an interaction term between age and systolic blood pressure.



The plot shows an interaction between smoker status and cholesterol levels. The cholesterol curves differ between smokers and non-smokers, as the curves are in different areas and follow different trends. This suggests that an interaction term between smoker status and cholesterol levels could potentially be included in the final model.

What we observed in plot one was also in other interactions, such as the one between gender and systolic blood pressure and the one between BMI and family history. This suggests that these models should also not be included in the final model. All these interaction terms were left out of the model-building process. While we only showed the plot for two interactions, several more interactions, all included in the model-building stage, were not shown.

Methodology:

Since our response variable is binary, we will use a logistic regression generalized linear model. Our model equation is:

$$\log(\frac{p_i}{1-p_i}) = \mathbf{X}_i^T\beta, \text{where } \mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T, \beta = (\beta_1, \ldots, \beta_p)^T, \text{and } i = 1, \ldots, n$$

Where log(pi / 1 - pi) is the link function for our model, X_i represents our predictor variables, and β represents the coefficients of X_i.

Before we begin model selection, we need to determine if the assumptions for our base model are satisfied. Since we are doing a logistic regression model, we will check for binary response, dispersion, linearity, independence, and multicollinearity.

Model Assumptions For The Base Model:

Binary Response - Our response variable is Diabetes Status, binary, where 1 indicates a diabetes diagnosis and 0 indicates no diagnosis. Thus, the binary response assumption is satisfied.

Independence -

Durbin-Watson Test			
Lag	**Autocorrelation**	**D-W Statistic**	**p-value**
1	0.007771225	1.980057	0.876

We perform the Durbin-Watson test on a basic main effects model to check for independence between residuals. The test reveals that the residuals are uncorrelated or independent as the p-value is greater than an α = 0.05, leading us to fail to reject the null hypothesis of the residuals being uncorrelated and concluding that they are independent.

Dispersion - Since the independence assumption was passed based on the Durbin-Watson test above, we do not need to check the dispersion assumption as it is linked to independence.

Linearity - The linearity check we did above in the EDA section revealed some violations in the linearity assumptions. We combatted this by adding higher-order terms to our first model that reflected the trends seen on the plots.

Multicollinearity -

Multicollinearity VIF Values							
Age	**BMI**	**Gender**	**Smoker**	**Cholesterol**	**Systolic BP**	**Physical Activity**	**Family History**
1.108152	1.072568	1.018301	1.022086	1.022328	1.007883	1.005718	1.011940

To test for Multicollinearity between our predictor variables, we checked the VIF values. All values are less than 5, indicating no significant multicollinearity issues for these predictors, and we can be sure that they are independent of each other. We do not need to worry about variables that are highly correlated with each other and run into inaccurate outcomes while determining the result. Since no multicollinearity exists, we do not need to centralize or use ridge regression.

Roadmap:

To determine the most effective model for predicting diabetes status, we tested multiple logistic regression models. Multiple iterations, quadratic terms, and variable selection techniques were all applied during this process. We examined the p-values, residual deviance, and AIC values to determine the most optimal model. For each model we also checked the Pearson residuals for each numeric predictor to ensure that the linearity assumption wasn't violated and performed a Wald test to check if there was at least one useful predictor within the model.

Starting with the primary effect model, we had an output showing That Age, BMI, and Smoking were significant predictors, along with an AIC of 341.38.

In model 1, we added quadratic terms (BMI^3 & Systolic BP^3), and the purpose was to see any impact in predicting diabetes status. With an increase in AIC of 356.8 and higher p-values for BMI^3 and Systolic BP^3), quadratic terms were not improving the overall prediction of the model with overfitting concerns of being too complex. For the linearity test, the loess curves for each residual plot were relatively straight, so the model passes the linearity assumption. After performing a Wald test, the p-value we get is 1.9e-6. Since this is lesser than α = 0.05, we reject the null hypothesis and conclude that there is at least one significant predictor within the model.

In model 2, we removed quadratic terms while keeping the interaction terms to see their importance. AIC did offer a better fit of 353.5, and residual deviance remained similar from 310.86 to 311.30. This model was a better fit but not significantly better than the last. For the linearity test, the loess curves for each residual plot were relatively straight, so the model passes the linearity assumption. After performing a Wald test, the p-value we get is 6.3e-7. Since this is lesser than α = 0.05, we reject the null hypothesis and conclude that there is at least one significant predictor within the model.

In model 3, we removed all the interaction terms with p-values greater than 0.7 from the last model, keeping only interaction terms with a higher chance of being statistically significant. AIC improved to 347.54, and the Wald test showed a chi-squared value of 66.6 with 17 degrees of freedom and a p-value of 8.3e-08, indicating smoker and gender remain non-significant and systolic blood pressure and physical activity remain non-significant after removing interaction terms. For the linearity test, the loess curves for each residual plot were relatively straight, so the model passes the linearity assumption. After performing a Wald test, the p-value we get is 8.3e-8. Since this is lesser than α = 0.05, we reject the null hypothesis and conclude that there is at least one significant predictor within the model.

In model 4, we removed all the interaction terms with p-values greater than 0.5 from the last model, with the AIC decreasing to 342.41, meaning the model improved compared to model 3. This indicates that Smoker:BMI is not a critical interaction term in our model. The Wald test showed a chi-squared value of 66 with 14 degrees of freedom and a p-value of 1e-08, which indicates that the rest of the predictors also contribute to the significance of our model. For the linearity test, the loess curves for each residual plot were relatively straight, so the model passes the linearity assumption. After performing a Wald test, the p-value we get is 1.0e-8. Since this is lesser than α = 0.05, we reject the null hypothesis and conclude that there is at least one significant predictor within the model.

In model 5, we removed all the interaction terms with p-values greater than 0.3 from the last model, keeping only those with a higher chance of being statistically significant. AIC improved to 338.28, and the Wald test showed a chi-squared value of 65.1 with 10 degrees of freedom and a p-value of 3.8e-10, indicating that the terms removed were insignificant. For the linearity test, the loess curves for each residual plot were relatively straight, so the model passes the linearity assumption. After performing a Wald test, the p-value we get is 3.8e-10. Since this is lesser than α = 0.05, we reject the null hypothesis and conclude that there is at least one significant predictor within the model.

In model 6, we removed all terms with a p-value greater than 0.15: cholesterol, physical activity, family history, and their interaction terms. Although the model's AIC decreased from 338.28 to 333.83, we did achieve a much simpler and cleaner model by removing the terms that were not contributing to the overall fit. For the linearity test, the loess curves for each residual plot were relatively straight, so the model passes the linearity assumption. After performing a Wald test, the p-value we get is 4.2e-12. Since this is lesser than α = 0.05, we reject the null hypothesis and conclude that there is at least one significant predictor within the model.

We also decided to look at a reduced model 6 by removing smoker and BMI:smoker as the p-value for smoker was 0.01929 in model 6, which was above our 0.15 threshold. AIC increased from 333.83 to 337.12, indicating our model has slightly worsened. Using the likelihood ratio test between the full and reduced model 6, we can see a p-value of 0.0069, which is less than 0.05, indicating that the full model 6 is a better fit than the reduced model 6. Smoking is statistically significant to the overall model. For the linearity test, the loess curves for each residual plot were relatively straight, so the model passes the linearity assumption. After performing a Wald test, the p-value we get is 1.7e-12. Since this is lesser than α = 0.05, we reject the null hypothesis and conclude that there is at least one significant predictor within the model.

Results:

Throughout the model selection process, we started with a full model consisting of quadratic and interaction terms, then removed insignificant ones. We ended up with model 6 for our final model because it provided the best fit and had the lowest AIC score of 333.28.

Final Model:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_2 x_4 + \beta_6 x_2 x_3$$

$$x_1 = \text{Age} \quad x_3 = \text{Smoker}$$

$$x_2 = \text{BMI} \quad x_4 = \text{Systolic Blood Pressure}$$

Predictor	Estimate	Std. Error	z-value	p-value
(Intercept)	-11.41	3.664	-3.115	0.00184
Age	0.06978	0.009070	7.694	1.43e-14
BMI	0.2863	0.1227	2.333	0.01964
Smoker	-2.042	1.589	-1.285	0.19878
Systolic_BP	0.04684	0.02509	1.866	0.06199
BMI:Systolic_BP	-0.001704	0.0008653	-1.969	0.04896
BMI:Smoker	0.09127	0.05315	1.717	0.08595

Interpretations:

The estimated coefficient for Age is about 0.06978, which means that as age has a one-unit increase, the odds of having diabetes increase by about 7.2%. The estimated coefficient for BMI is about 0.2863, which means that as BMI increases by one unit, the odds of diabetes increase by about 33%. The estimated coefficient for Smokers is about -2.042, which means that smokers are 87% less likely to be diagnosed with diabetes. The estimated coefficient for Systolic BP is about 0.04684, which means that as an individual's systolic BP increases by one unit, the odds of having diabetes increase by about 4.8%. The estimated coefficient for the BMI and Systolic BP interaction term is about -0.001704, which means that for every one-unit increase in Systolic BP, the odds of a BMI's effect on diabetes decrease by about 0.17%. The estimated coefficient for the BMI and Smoker interaction term is 0.09127, which means that for people who smoke, the odds of their BMI also impacting their risk of developing diabetes increases by about 9.56%.

The p-values for Age, BMI, and the interaction term between BMI and Systolic BP are all less than α = 0.05, meaning they are significant to the model. The p-values for Smoker, Systolic BP, and the interaction term between BMI and Smoker are all slightly above α = 0.05, meaning they are insignificant to the model. However, we included them as the individual predictors in another significant interaction term, and the LRT revealed that we should not remove the interaction term between BMI and Smoker.

Assumptions Check:

Binary Response - Our response variable is Diabetes Status, binary, where 1 indicates a diabetes diagnosis and 0 indicates no diagnosis. Thus, the binary response assumption is satisfied.

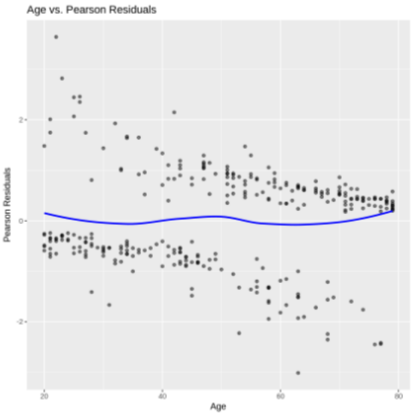
Independence -

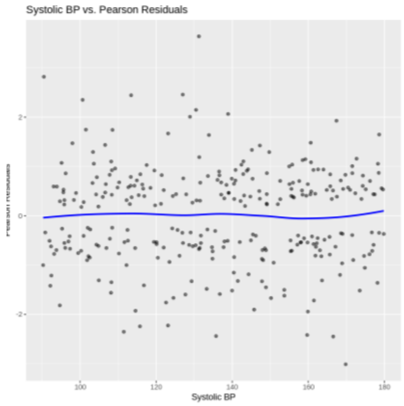
Durbin-Watson Test			
Lag	**Autocorrelation**	**D-W Statistic**	**p-value**
1	0.02691249	1.942591	0.602

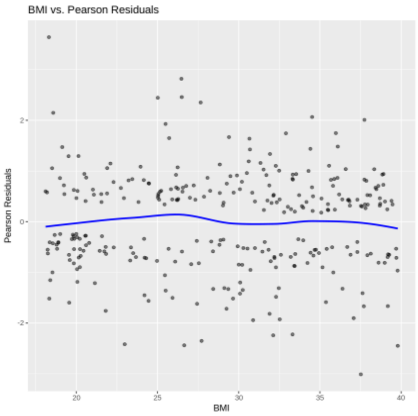
We perform the Durbin-Watson test on our final model to check for independence between residuals. The test reveals that the residuals are uncorrelated or independent, as the p-value is greater than an $\alpha = 0.05$. Thus, we fail to reject the null hypothesis of the uncorrelated residuals and conclude that they are independent.

Dispersion - Since the independence assumption was passed based on the Durbin-Watson test above, we do not need to check the dispersion assumption as it is linked to independence.

Linearity - The linearity assumption is satisfied for all the predictor variables in our final model, as the loess curves are primarily straight.







Wald Test -

Wald Test		
Chi-squared	**Degrees of Freedom**	**p-value**
65.0	6	4.2e-12

We performed the Wald test to check for any useful predictors within our model. Since the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that there is at least one significant predictor.

Discussion and Conclusion:

Based on the final model, we can conclude that Age and BMI are the two stronger predictors of Diabetes. As people age, the risk of developing diabetes increases. According to WebMD, the older we are as humans, the more susceptible we are to getting diabetes. The same goes with BMI; the more a person's BMI increases, the more they are putting themselves at risk for getting diagnosed with diabetes. Age and BMI are strong predictors in our model, and they tie into studies showing that age and BMI are significant factors for diabetes. Systolic BP alone indicates that it has a weak effect on Diabetes. Systolic BP tends to be high when a person has diabetes but may not be a proper predictor on its own to indicate diabetes. However, due to its interaction with BMI being significant, it suggests that having a high BMI and a high Systolic BP can together increase the risk of diabetes. Smoking alone, as well as its interaction with BMI, shows that it does not correlate at all with Diabetes. However, studies have shown that smoking alone can increase a person's risk for diabetes, which then suggests that the data collected may not be an accurate representation of smoking and diabetes.

We encountered a few limitations: our models became more complex, with higher AIC values and too many non-significant terms, leading to overfitting. Another concern to address is possible biases in variables, such as smoking or physical activity, as they could be self-reported and could lead to misrepresentation of the final model. Variables such as blood pressure or cholesterol levels can also fluctuate based on external temporary factors such as stress or diet. Possible improvements for the future could be to move forward with the undersampling of smokers and physical activity levels and see how it would impact the final model's performance. When looking at coefficients, we noticed that smokers are 87% less likely to be diagnosed with diabetes, contradicting research showing smoking increases the chance of diabetes. Based on this statement, more analysis needs to be done, and a possible improvement could be to adjust variables like diet to isolate the smoking variable. We can also reassess all estimated coefficients for interaction terms and see if they indicate a meaningful relationship between the variables or if they result from imbalances in the dataset.

Author Contributions:

Eve Fraczkiewicz: Introduction, Data Cleaning, EDA, Model Building
Sai Shrinidi Krishnan: Introduction, Summary Statistics, Model Structure, Discussion
Seona Magdum: Introduction, Discussion
William Pham: EDA, Model Assumptions, Results
Tianle Qi: EDA, Model Structure Roadmap, Results

References:

CDC: Smoking and Diabetes. (n.d.). Retrieved March 19, 2025, from https://www.cdc.gov/tobacco/campaign/tips/diseases/diabetes.html

Diabetes in America: Prevalence, statistics, and economic impact. (n.d.). Retrieved March 19, 2025, from https://diabetes.org/about-diabetes/statistics/about-diabetes

European Society of Cardiology: Body Mass Index is a More Powerful Risk Factor for Diabetes Than Genetics. (August 31, 2020). Retrieved March 19, 2025, from https://www.escardio.org/The-ESC/Press-Office/Press-releases/Body-mass-index-is-a-more-powerful-risk-factor-for-diabetes-than-genetics

WebMD: How Age Relates to Type 2 Diabetes. (June 15, 2024). Retrieved March 19, 2025, from https://www.webmd.com/diabetes/diabetes-link-age