Food Insecurity and Sociodemographic Factors on Diabetes

Group 43: Pey-Tzer Chio, Christina Gregis, Ju Ho Lee, Naomi Shields

Summary

Research Question: To what extent do food insecurity and socioeconomic determinants influence diabetes prevalence in the United States?

Problem Description:

There are limited longitudinal studies discussing the relationship between factors including food insecurity (FI) and type 2 diabetes (T2DM) (Gu et al., 2025). Some studies have demonstrated worse glycemic control in those with FI; however, there is limited evidence that this disparity widens over time (Gu et al., 2025). With limited studies, it creates a problem that impacts the overall health of the American adult population. We will further explore the relationships between FI and sociodemographic factors on diabetes prevalence in the United States.

Goals & Major Findings:

Our project applied 4 main approaches to analyze Nutritional Health and Nutrition Examination Survey (NHANES) data: 2 logistic regression models, a decision tree model, and a Poisson regression model. We found that increasing food insecurity significantly raises diabetes risk, with low food security associated with a 44.5% higher odds of diabetes diagnosis. SNAP participation was linked to the 44.3% increased risk of diabetes. There were clear demographic patterns that showed Mexican Americans have the highest diabetes prevalence, with women having a 29.4% lower diabetes risk than men, and risk increased substantially with age, especially over 40. Our decision tree revealed that recent blood tests, age, prediabetes history, and education level were the strongest predictors of diabetes. Our Poisson regression model examined respondents across poverty categories, highlighting how food assistance program participation relates to diabetes across poverty levels. Our conclusions on food insecurity's relationship to diabetes remained consistent across all our models. These findings demonstrate how food access and socioeconomic factors explain diabetes prevalence among Americans.

Data Sources:

We analyzed 5 primary data sources, collected from the Centers for Disease Control and Prevention (CDC)'s National Center for Data Statistics:

1) *P_DEMO*: Demographics: This contains demographic variables, including age, gender, race, and ethnicity.
2) *P_DIQ*: Diabetes Questionnaire: This provides self-reported diabetes information, including diabetes diagnosis, prediabetes history, and blood testing results.
3) *P_FSQ*: Food Security Questionnaire: This captures household food household food security, using the United States Department of Agriculture (USDA)'s criteria
4) *P_DBQ*: Diet Behavior and Nutrition: This records eating patterns, including non-home meals, fast food consumption, and frozen foods meal frequency.
5) P_INQ: Income Questionnaire: This contains household income and participation in food assistance programs, like Supplemental Nutrition Assistance Program (SNAP) and emergency food banks.

All data sources are separate files available from 2017-2020 NHANES database (Centers for Disease Control and Prevention, n.d.-a). We selected relevant variables, based on topics we explored in our literature survey. In the original data files, the variables had unclear names, such as “DID010”. The data variables could only be identified in a data lookup table. When we merged our four data sets, we renamed the variables to be clear, concise, intuitive names, like “diabetes_diag” for diabetes diagnosis. Our final data file includes the updated variable names to reflect each variable’s descriptions available on the NHANES variable lookup table (Centers for Disease Control and Prevention, n.d.-b).

Introduction

Diabetes remains one of the most pressing public health challenges in the United States, with its burden disproportionately affecting low-income and food-insecure populations. The intersection of food insecurity (FI) and diabetes is complex, shaped by systemic, environmental, and socioeconomic factors. While previous studies have examined the relationship between FI and health outcomes, much of the research has been limited to analyses of single indicators, failing to capture the complex interactions of determinants of diabetes. Our analysis builds upon previous research by employing multiple statistical methods—including logistic regression, decision trees, and Poisson regression—to examine diabetes prevalence. We expect to identify the key determinants that significantly impact diabetes risk and develop predictive and explanatory models for determinants of diabetes and inform improvements in healthcare policies.

The following literature review summarizes findings from 20 academic sources on the links between food insecurity and diabetes. These studies span multiple themes: the nutritional and behavioral impacts of food insecurity, socioeconomic and demographic determinants, the effectiveness of food assistance programs, and the structural barriers to healthy food access. They provide essential context for our analysis of the variables influencing diabetes risk.

Literature Review:

Food Insecurity – Community Challenges and Nutritional Intervention (Christina)

Researchers examined food security challenges across six rural communities in six different states: Arkansas, Montana, North Carolina, Oregon, Texas, and West Virginia). The study identified how rural communities have higher poverty rates and less access to healthy food. Rural participants discussed managing food budgets, rationing their daily food intake, and prioritizing their children's needs above their own, while in economic distress. Rural residents were more likely to find coping mechanisms within their own households, before seeking community assistance. This behavior is partly rooted in the social stigma that Americans have about receiving aid from food assistance programs, as also highlighted in a similar study conducted in Pittsburg (Byker Shanks et al., 2022).

Research confirms these nutritional challenges between poverty and health. A team of researchers analyzed the Third National Health and Nutrition Survey (NHANES III) from 1988-1994 via a logistic regression model and compared the nutritional health of Americans from food insecure and food secure backgrounds. The study argued that dietary supplements and fortified foods alone do not compensate for inadequate nutrition. Younger, food-insufficient Americans

were 1.39 times more likely to have calcium intake below 50 percent of the recommended levels. Meanwhile, older, food-insufficient Americans are 3.66 times likely more likely to have insufficient iron levels. These findings highlight that limited food access compromises an individual's immune system and increases the risk of chronic disease. The researchers emphasized the importance of targeted interventions, like food stamps and nutritional education, for improving diets among food-insecure communities (Dixon et al., 2001).

Similarly, another study used logistic regression to analyze the 1999-2002 National Health and Nutrition Examination Survey (NHANES), revealing the paradoxical relationship between food insecurity and diabetes. Individuals experiencing severe food insecurity had 2.1 times higher odds of diabetes compared to food-secure individuals, even after adjusting for physical activity and sociodemographic factors. This can be explained by individuals with food insecurity having higher consumption rates of inexpensive food alternatives that offered little nutritional value. This relationship remained significant after adjusting for Body Mass Index (BMI), highlighting how food insecurity's association with diabetes exists even if obesity is not considered. The study proposed an evolutionary mechanism, where chronic food scarcity triggers insulin resistance – originally an adaptive measure to conserve muscle protein – that may ultimately lead to diabetes (Seligman et al., 2007).

To examine this relationship more closely, a study of diabetic patients in the San Francisco and Chicago safety net clinics created multiple linear regression models to evaluate the relationship between food insecurity and glycemic control via Hemoglobin A1c (HbA1c). HbA1c is a common metric for measuring long-term glucose levels in diabetic patients. 46% of diabetes patients were food-insecure, with a higher prevalence among younger patients. Food-insecure patients were more likely to have poor glycemic control. Diabetic patients' difficulty following a diabetic diet and emotional distress from diabetes explained over half the relationship between food insecurity and glycemic control. This study reveals how critical patients' food access is for managing diabetes (Seligman et al., 2012).

To address these challenges, we can analyze the USDA's Supplemental Nutrition Assistance Program (SNAP). SNAP is the nation's largest nutrition program, committed to alleviating food scarcity and feeding families in low-income households. Researchers in a Pittsburg Hill study used linear regression to examine SNAP participants in Pittsburg neighborhoods with and without a new supermarket. The goal was to explain how food access affected SNAP participants' health. Increasing food access, via the new supermarket, significantly improved SNAP participants' food security and diet. These also participants had reduced sugar consumption and a decreased BMI, compared to SNAP participants without the new supermarket access. The study showed that despite the stigma associated with enrolling in food assistance programs, food access significantly improved participants' health outcomes (Cantor et al., 2020).

Correlation between Food Security and the Prevalence of Diabetes (Pey)

FI is the socioeconomic condition of limited access to adequate food and can pose a barrier to maintaining a healthy diet, which can impact the management of various chronic conditions such as diabetes. A cross-sectional analysis was conducted among 2075 adults aged 20 years and older with diabetes and participated in the 2018-2018 NHANES survey. This

analysis studied the association between household FI/diet quality and several biomarkers including A1c, blood pressure, and cholesterol (ABC). The findings showed that among the participants, 17.6% had FI/low diet quality, 14.2% had FI/high diet quality, while 33.1% had food security/low diet quality, and 35.2% had food security/high diet quality. A1c levels below 5.7% are considered normal, 5.7%-6.4% indicate prediabetes, and 6.5% or greater indicate diabetes. Those with FI/low diet quality were at a higher risk of an A1c ≥7.0% (adjusted odds ratio=1.85, 95% confidence interval 1.23 to 2.80) and A1c ≥8.0% (adjusted odds ratio=1.79, 95% confidence interval 1.04 to 3.08)). Similarly, those with FI/high diet quality and food security/low diet quality were also at risk of an elevated A1c. Overall, regardless of diet quality, the study showed a significant association between FI, elevated A1c, and other metabolic outcomes (Casagrande et al., 2022).

On the other hand, researchers of a 2-year longitudinal study examined the association of three food security categories (persistently secure, intermittently secure, and persistently insecure) as well as changes in diet quality, weight, and glycemia in adults with prediabetes and T2DM. The participants of this study were recruited between December 2019 to December 2020 and consisted of individuals aged 21-62 and enrolled in the Massachusetts Flexible Services (Flex), a Medicaid-funded program, from five community health centers. The results of the study showed that among the 188 participants, about a third of them fell into each category with the highest group in the persistently insecure category. More specifically, 32.4% were persistently food secure, with 33% intermittently secure, and 34.5% persistently insecure. The study found that there were no differences in change in diet quality, BMI, or hemoglobin A1c in relation to FI or evidence suggesting that FI was associated with higher A1c; however, more longitudinal studies are needed (Gu et al., 2025).

Another analysis was conducted using data gathered from the NHANES survey between 2005 and 2014 consisting of 27,218 adults aged 20 years and older from the United States showed that about 12% of adults have diabetes with a quarter undiagnosed. Among this, those with prediabetes were 39% more likely to be FI and those with diabetes were 58% more likely to be FI. As individuals with lower socioeconomic status have limited access to the healthcare system, an increase in diabetes screening for FI populations is necessary. These findings are critical for primary care and strategies, such as community screening events and tailored medical interventions, can improve health outcomes (Walker et al., 2018).

Likewise, an analysis from NHANES (2011-2016) was conducted with 1682 adults 20 years of age and older with T2DM assessing food security using multinomial regression models. The results indicated that those with poor diet quality and FI had higher odds to have an elevated A1c ([AOR = 6.12] for HbA1c = 8-<9% and [AOR = 6.7] for HbA1c =>9%). Additionally, minority status was associated with an A1c ≥9% (Black [AOR = 1.71] and Hispanic [AOR = 2.95]). Overall, poor diet quality was found to be associated with poor glycemic control among adults with T2DM in the US with FI increasing the odds of having poor glycemic control while those with a poor diet but had food security did not (Shaheen et al., 2021).

Despite limited longitudinal studies suggesting a link between food insecurity and elevated A1c, there is good evidence that suggest that food programs, such as federal nutrition assistance programs, can impact intermediate outcomes like dietary intake and food security. FI and diabetes are linked through multiple pathways with research connecting lower socioeconomic status and increased diabetes diagnosis and complications. For example, low-income households encounter barriers such as high food prices and limited access to healthy food. Due to this, FI individuals may consume fewer nutrient dense foods, such as fruits and vegetables, leading to poorer overall diet quality. Over long periods of time, this dietary pattern can increase insulin resistance and blood glucose levels. In addition, individuals experiencing FI may also have to make difficult decisions between food and other necessities, such as medication, which can make diabetes management more challenging (Levi et al., 2023).

Sociodemographic and Health Determinants of Diabetes (Ju Ho)

Food insecurity affected 11.8% of American households (15 million people) in 2017. The relationship between food insecurity and obesity has been studied for decades, with the first published paper in 1995 by WH Dietz. Since then, the topic has grown in importance and the NIH has invested in a spectrum of obesity research over the past 15 years to understand the complicated factors that contribute to obesity, food insecurity, and nutrition. The NIH continues to prioritize food insecurity research to improve health, quality of life, lifespan, reduce illnesses and health disparities in America (Brown et al., 2019).

Food security is traditionally defined by four pillars: availability, access, utilization, and stability. Clapp, Moseley, Burlingame, and Termine (2021) propose two additional pillars—agency and sustainability. Agency refers to individuals' ability to make decisions about their food (what people eat and how they access it), while sustainability emphasizes ensuring food systems remain viable long-term without depleting natural resources. This expanded framework provides a more comprehensive view of food insecurity, considering not only access to food but also the ability to make healthy food choices and the long-term sustainability of food systems, both of which influence the prevalence of obesity (Clapp et al., 2021).

Lee, Zhao, Reesor-Oyer, Cepni, and Hernandez (2020) explore the bidirectional relationship between food insecurity and housing instability, showing that food-insecure families are 62% more likely to experience housing instability, and families with housing instability are 40% more likely to face future food insecurity. This cycle of food insecurity and housing instability underscores how interconnected social determinants of health are. The study calls for integrated programs that address both issues simultaneously, particularly for vulnerable populations, such as those with children or individuals experiencing mental health challenges like anxiety and depression (Lee et al., 2020).

The long-term consequences of food insecurity on health are also well-documented. Leung and Wolfson (2021) found that food insecurity is associated with poorer diet quality in older adults, which can contribute to the development of chronic diseases. Similarly, Banerjee, Radak, and Dunn (2020) demonstrate that food insecurity is linked to significantly higher mortality rates, particularly from cardiovascular diseases. These studies highlight the profound impact food insecurity can have on health over time, underscoring the need for early intervention to reduce its long-term health effects (Leung & Wolfson, 2021).

The relationships between food insecurity, sociodemographic factors, diabetes, and long-term health outcomes are complex and multifaceted. Previous research has typically concentrated on examining a single variable in isolation with respect to food insecurity, often overlooking the broader array of interconnected factors. This narrow focus has limited the understanding of the many variables that can contribute to food insecurity and its resulting health outcomes. A more comprehensive approach that considers multiple social determinants of health and health factors, is essential for advancing the understanding of food insecurity's impact on diabetes and developing more holistic interventions to improve public health (Banerjee et al., 2020).

The Structural Impact of Food Insecurity on Diabetes Management (Naomi)

Food insecurity is not just an individual challenge—it is a systemic issue deeply rooted in historical policies and socioeconomic disparities. Redlining, restrictive zoning laws, and

economic disinvestment have led to the formation of food deserts, where access to affordable, nutritious food is severely limited. These structural barriers have contributed to higher diabetes prevalence, increased complications, and elevated mortality rates, particularly among low-income and minority populations (Hill-Briggs et al., 2020).

Beyond food availability, the financial burden of diabetes management is exacerbated by food insecurity. Individuals with unstable access to food often struggle to afford balanced meals, leading to poor glycemic control. Research highlights two key dimensions of food scarcity: food insecurity, which stems from financial constraints, and physical food access, which depends on the geographic availability of nutritious options. Both factors are strongly linked to higher HbA1c levels, making diabetes more difficult to manage (Berkowitz et al., 2018).

Neighborhood characteristics also play a critical role in shaping diabetes risk, particularly for youth. While type 2 diabetes (T2D) was historically considered an adult disease, its prevalence among younger populations—especially in low-income and minority communities—has been rising. Socioeconomic status, rurality, and limited food access create an environment that fosters chronic disease. While much research has focused on individual-level risk factors such as obesity and maternal health, recent studies emphasize the need to address broader social and environmental influences (Liese et al., 2018).

Dietary patterns remain one of the most significant factors in diabetes prevention and management, but food insecurity limits the ability to maintain a healthy diet. Consuming whole grains, fruits, vegetables, legumes, and nuts while reducing refined grains, processed meats, and sugary beverages is associated with better glycemic control. However, economic disparities often dictate dietary choices, necessitating policy interventions such as agricultural subsidies for healthy foods and taxation on unhealthy products to improve accessibility (Ley et al., 2014).

Additionally, food-insecure individuals often face difficult financial trade-offs, such as choosing between purchasing food, medication, or other essential expenses. This economic strain forces reliance on cheaper, calorie-dense foods that contribute to poor diabetes management. Addressing these challenges requires systemic interventions, including healthcare screenings for food insecurity and policies that improve access to culturally appropriate, nutritious foods. From a predictive standpoint, food insecurity is a critical variable in identifying individuals at higher risk of requiring medical care, emphasizing the broader economic and structural dimensions of diabetes outcomes (Gucciardi et al., 2014).

Analysis

Method #1 (Pey-Tzer) — *Food Insecurity, Demographics, and Diabetes Prevalence Logistic Regression:*

The logistic regression model will explore how certain explanatory variables (predictors), such as food insecurity, impact the prevalence of diabetes diagnosis (response). Since food insecurity is related to many other factors, variables to also consider include race, dietary quality, and socioeconomic status. Assessing these variables can help us better interpret the correlation between FI and diabetes. Logistic regression models are useful for predicting binary outcomes using one or more continuous or categorical predictors, which would be useful in this case as there are several categorical predictors, and we are trying to predict the probability of

diabetes diagnosis based on those predictors. The data source includes a dataset from the NHANES survey (Centers for Disease Control and Prevention, 2025a).

Results #1 (Pey-Tzer) — *Logistic Regression:*

We chose to use a logistic regression model to understand the association between the prevalence of diabetes and food insecurity. Since the goal is to identify individuals at the highest risk for diabetes based on food security and economic factors, it was important that we delve into various factors related to food security and socioeconomic status.

Explanatory variables:

1. "household_fs_cat" (*categorical*): household food security category for the last 12 months (**1 - full food security, 2- marginal food security, 3- low food security, 4- very low food security**)
2. "SNAP_current" (*categorical*): whether any member of the household receives SNAP benefits (**1 - yes, 2 - no, 7- refused, 9 - don't know**)
3. "Poverty_index" (*numerical continuous*): family monthly poverty level index, a ratio of monthly family income to the HHS poverty guidelines specific to family size
4. "Age" (*numerical*): age in years at screening
5. "hispanic_origin" (*categorical*): reported race and Hispanic origin information (**1- Mexican American, 2- Other Hispanic, 3- Non-Hispanic White, 4- Non-Hispanic Black, 5- Other Race**)
6. "non_hispanic_origin" (*categorical*): reported race and Hispanic origin information, with Non-Hispanic Asian category (**1- Mexican American, 2- Other Hispanic, 3- Non-Hispanic White, 4- Non-Hispanic Black, 6- Non-Hispanic Asian, 7- Other Race**)

Response variable: "diabetes_diag" (*categorical*): whether the participant has been told by a healthcare professional if they have diabetes (**1- yes, 2- no, 3- borderline, 7- refused, 9- don't know**)

Exploratory Data Analysis:

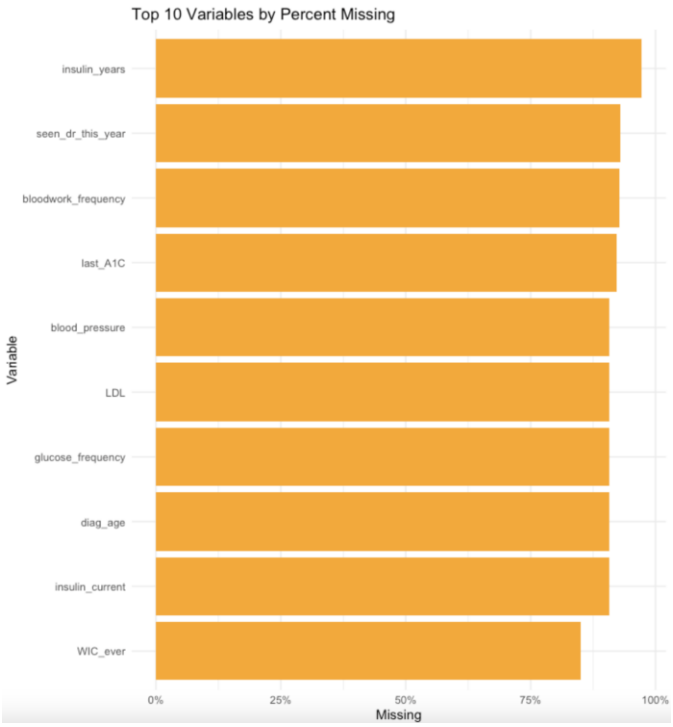


Figure 1. Top 10 variables with missing data

Upon selecting the predictors and response, we created a table to identify the top variables in the whole dataset that had missing values to ensure that none of the ones we selected had a significant number of NA values. As seen in the bar chart, none of the of the selected predictors are listed in the *Top 10 Variables by Percent Missing*.

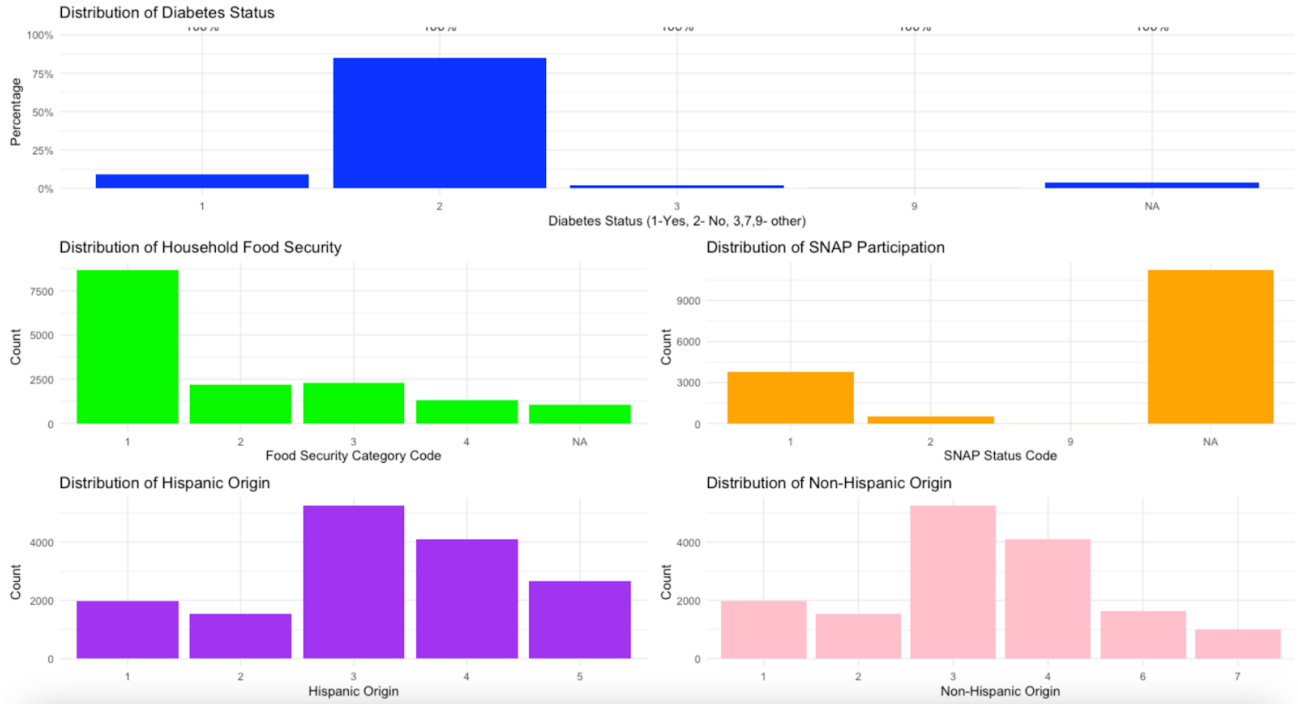


Figure 2. Distribution of selected variables

We then visualized the univariate distribution of the selected variables. As seen in the charts above, most of the data points fall within category 2 (no diabetes) for the *diabetes_diag* response variable. For the categorical predictors, most of the points fall within category 1 (full food security) for household food security, N/A for SNAP participation, 3 for Hispanic origin and Non-Hispanic Origin (non-Hispanic White).

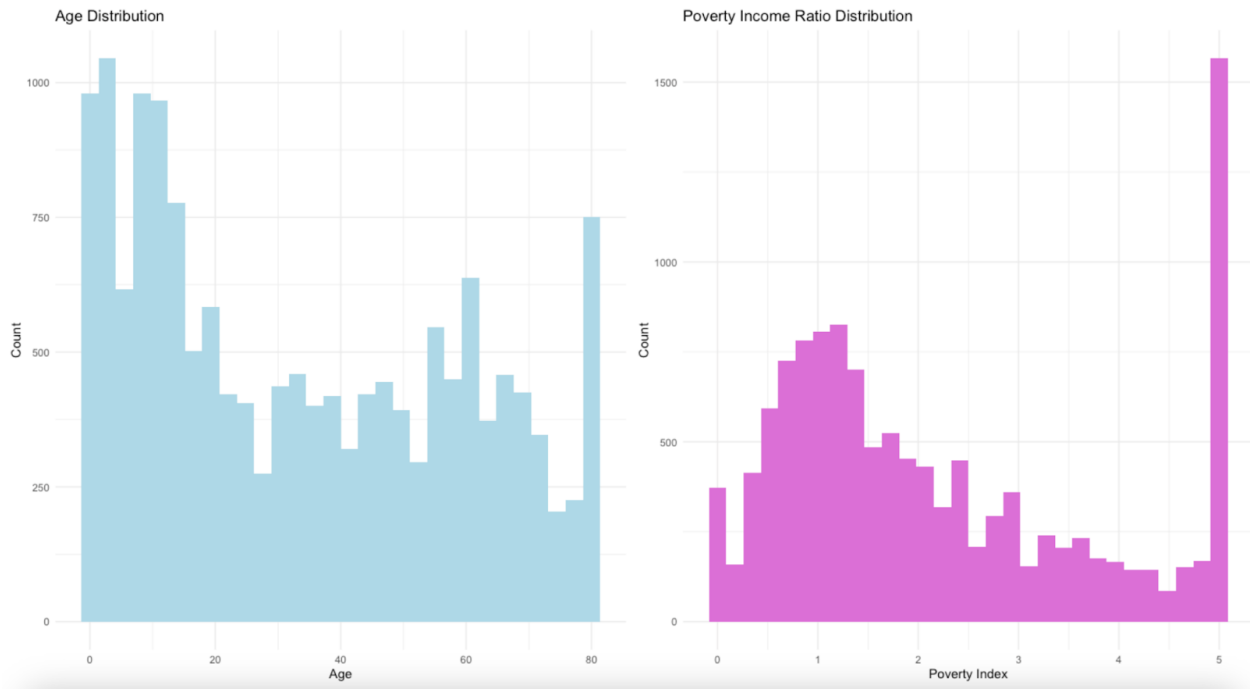


Figure. 3 Distribution of numerical predictors

For the numerical predictors, we can see a relatively spread of age range, with more individuals in the younger adult category, and a relatively right-skewed chart for the poverty index level with more individuals around index level 2 and spike at 5. Overall, this tells us that we need to transform the response to a binary variable and impute the other predictors with NA values.

The response variable selected was *diabetes_diag* as we are trying to see the relationship between food insecurity and the prevalence of diabetes in the US. The response variable was transformed into a binary variable (yes- 1 (originally 1), no- 0 (originally 2, 3, 7, 9)), and made into a new column named *diabetes_binary*. Then, we completed a bivariate analysis between

each selected predictors and response.

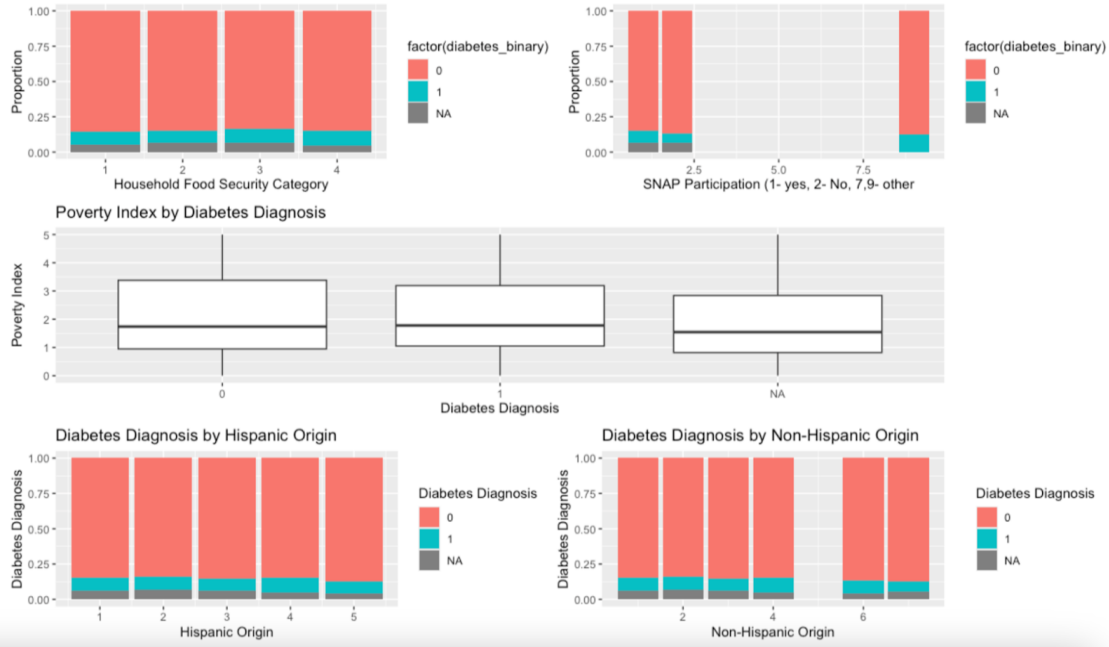


Figure 4. Correlation between response (*diabetes_diag*) and categorical predictors

We can see that the majority of the data points for each category in *household_fs_cat, SNAP_current, hispsanic_origin,* and *non_hispanic_origin* do not have diabetes (0). For the *poverty_index* box-and-whiskers plot, we can see that the median and overall range of the index level is roughly the same for all categories. With the median line closer to the bottom of each box, it suggests that this distribution is most likely right-skewed, but does not show a strong difference between individuals with and without diabetes and poverty level.

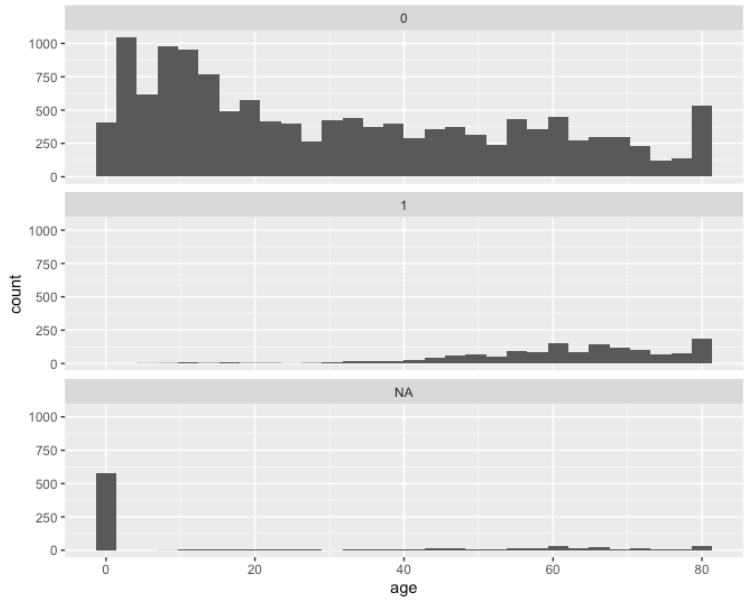


Figure 5. Correlation between response (*diabetes_diag*) and predictor *age*

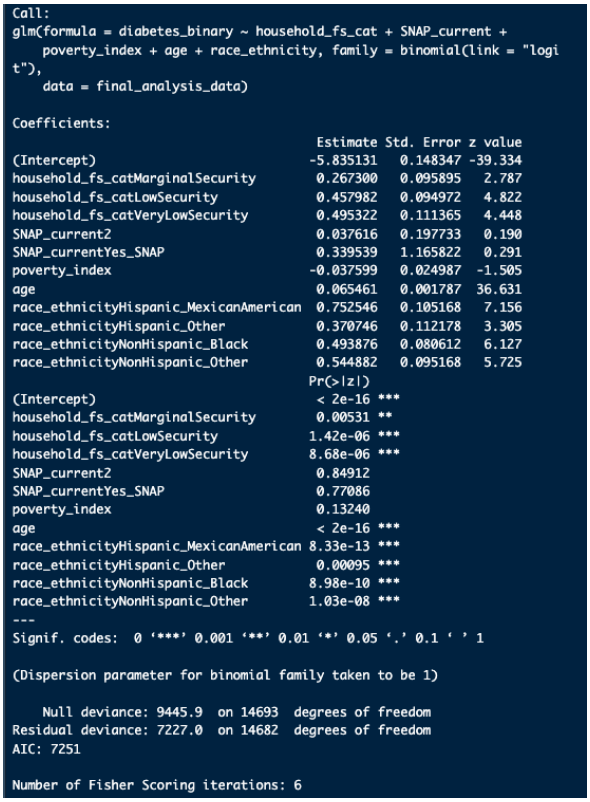
Lastly, in the *age* predictor, we can see that most of the data points fall within the no diabetes category with a relatively spread

Data Cleaning: In total, there were 16,122 NA values between each of the selected seven variables. The categorical variables (*household_fs_cat, SNAP_current, hispanic_origin, and non_hispanic_origin*) were changed to factors so that it could be categorized based on the levels indicated on the survey (ie. *household_fs_cat*: 1 = full food security). Since there were several NA values, we imputed the mode for the categorical data and imputed the median for the numerical variables (*poverty_index*, and *age*). After imputing values for the predictors, there were 1440 remaining NA values and finally a total of 0 NA values after removing missing values from the response variable.

For a clearer interpretation, we relabeled the categorical values with its respective category names (ie. 1 -> full food security). Additionally, we combined the race columns (hispanic_origina and non_hispanic_origin) into one (race_ethnicity) and relabeled them with its category name for easier interpretation.

The final columns selected include diabetes_binary, household_fs_cat, SNAP_current, poverty_index, age, race_ethnicity. After analyzing the distribution of the response variable in the final cleaned dataset, there were a total of 13249 (90.2%) without diabetes, and 1445 (9.8%) with diabetes.

Modeling: We then built the logistic regression model based on the transformed response variable and final predictors as indicated above.



```
Call:
glm(formula = diabetes_binary ~ household_fs_cat + SNAP_current +
    poverty_index + age + race_ethnicity, family = binomial(link = "logi
t"),
    data = final_analysis_data)

Coefficients:
                                  Estimate Std. Error z value
(Intercept)                     -5.835131   0.148347 -39.334
household_fs_catMarginalSecurity  0.267300   0.095895   2.787
household_fs_catLowSecurity       0.457982   0.094972   4.822
household_fs_catVeryLowSecurity   0.495322   0.111365   4.448
SNAP_current2                     0.037616   0.197733   0.190
SNAP_currentYes_SNAP              0.339539   1.165822   0.291
poverty_index                    -0.037599   0.024987  -1.505
age                               0.065461   0.001787  36.631
race_ethnicityHispanic_MexicanAmerican  0.752546   0.105168   7.156
race_ethnicityHispanic_Other      0.370746   0.112178   3.305
race_ethnicityNonHispanic_Black   0.493876   0.080612   6.127
race_ethnicityNonHispanic_Other   0.544882   0.095168   5.725
                                 Pr(>|z|)
(Intercept)                       < 2e-16 ***
household_fs_catMarginalSecurity  0.00531 **
household_fs_catLowSecurity      1.42e-06 ***
household_fs_catVeryLowSecurity  8.68e-06 ***
SNAP_current2                     0.84912
SNAP_currentYes_SNAP              0.77086
poverty_index                     0.13240
age                               < 2e-16 ***
race_ethnicityHispanic_MexicanAmerican 8.33e-13 ***
race_ethnicityHispanic_Other      0.00095 ***
race_ethnicityNonHispanic_Black  8.98e-10 ***
race_ethnicityNonHispanic_Other  1.03e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9445.9  on 14693  degrees of freedom
Residual deviance: 7227.0  on 14682  degrees of freedom
AIC: 7251

Number of Fisher Scoring iterations: 6
```

Figure 6. Logistic Regression Summary

The reference values for the categorical variables are as follows, *household_fs_cat*- Full security, *SNAP_current*- no SNAP, and *race_ethnicity*- non-Hispanic White. Using the logistic summary, we determined which predictors were significant as evidenced by a very low p-values. The confidence interval (CI) also helped to determine which predictors had strong significance against the response. Then, using this information, we created line plots to show the estimated probability of the predictors with significant correlation. The logistic summary indicates that compared to reference value, as food security worsens the likelihood (log-odds) of having diabetes increases significantly as evidenced by gradually decreasing p-values for marginal security, low security, and very low security, respectively. The summary shows that age has a higher log-odds of diabetes as evidenced by a very low p-value. Also, the *race_ethnicity* indicates that all other race/ethnicities have significantly higher likelihood of developing diabetes as compared to its reference value(non-Hispanic White). Non significant predictors include *SNAP_current* (p = 0.798) and poverty_index (p = 0.132). In addition to the model summary, the odds ratio and confidence intervals were calculated. The confidence interval (CI) *household_fs_cat*, *age*, and *race_ethnicity* are above 1, which indicate strong significance between the response and these predictors.

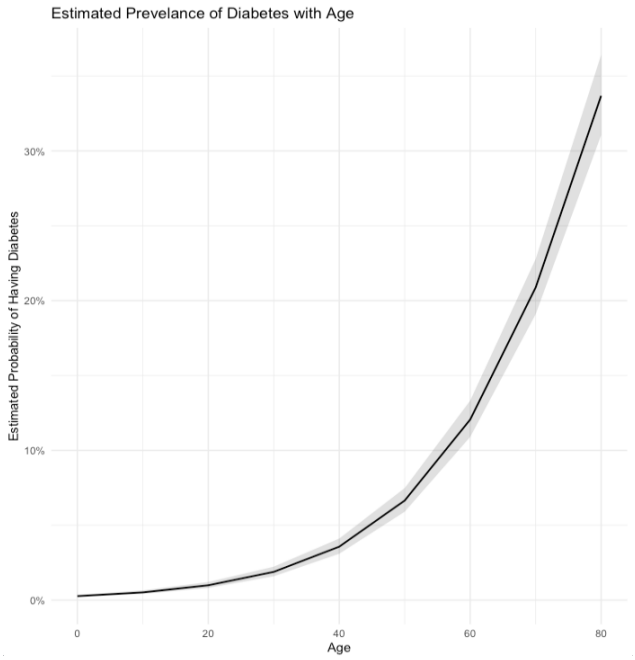


Figure 7. Estimated Prevalence of Diabetes and Age

The plot for *age* and *diabetes_binary* shows that there is a strong positive relationship between age and probability of developing diabetes with a gradual increase that accelerates around age 40 (note there is a shaded-grey area around the solid line representing the CI level).

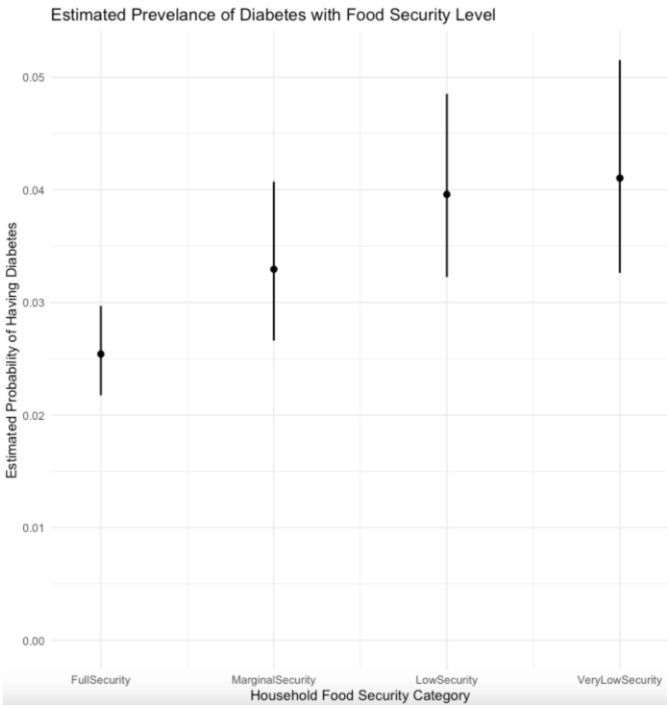


Figure 8. Estimated Prevalence of Diabetes with Food Security Level

The plot for *household_fs_cat* and *diabetes_binary* shows the estimated probability (black dot) and CI level (solid line) that increases as food insecurity worsens.

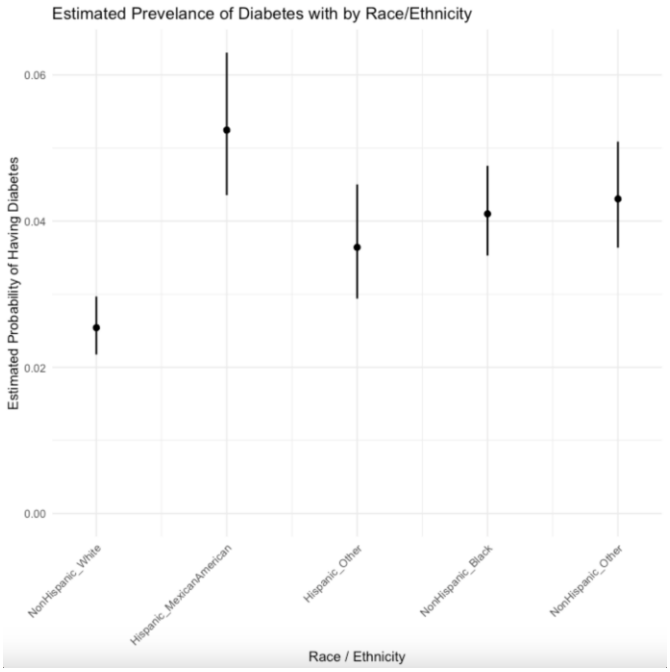


Figure 9. Estimated Prevalence of Diabetes by Race/Ethnicity

The plot for *race_ethnicity* and *diabetes_binary* shows that all other groups have a higher estimated probability of diabetes as compared to the reference group with Hispanic Mexican-

American having the highest prevalence of diabetes diagnosis. Based on this plot, we can see a clear difference in diabetes prevalence between different races in the US.

Overall, there seems to be a statistically significant association between diabetes diagnosis (*diabetes_binary*) and food security (*household_fs_cat*). There also seemed to be a clear positive association between diabetes diagnosis and age. Additionally, based on the model and plot, there seems to be a distinct disparity between diabetes diagnosis and races. Lastly, the bivariate association between poverty index and diabetes diagnosis was unclear.

Method #2 (Christina) — Food Security, Meal Patterns & Diabetes Logistic Regression:

The logistic regression model will explore how one's economic background, such as Supplemental Nutrional Assistance Program (SNAP) participation and poverty levels (predictors), relate to diabetes prevalence (response). This binary classification approach will find a person's probability of having diabetes. By analyzing the relationship between food accessibility and diabetes via odds ratios, we can determine how SNAP utilization and economic hardship explain diabetes likelihood. This approach offers advantages over previous studies by capturing a person's economic situation and diabetes risk, while accounting for factors influencing nutritional assistance participation.

We have a large data set: the full data set is 15, 560 rows and 39 columns. Each row represents a NHANES surveyed participant – a U.S. citizen, while each column represents a person's response about their health and background. We want to select variables to consider for our initial model. This is to narrow down our choice to variables we are interested in for the initial Logistic Regression model. We will later perform feature selection to retain relevant features.

In the literature survey, we explored how nutrition and local communities' roles in aiding lower income individuals contributed to food security and health outcomes. With the logistic regression model, we explored how food security and the community's impact further in explaining a person's diabetes diagnosis. We created an initial data summary and explored the number of missing values in each feature. Of the 39 possible variables in our merged data set, we had selected 10 variables that are relevant to our logistic regression model analysis. These variables have minimal missing values, compared to other NHANES survey variables:

Explanatory:

- **SNAP_ever**: Whether the respondent ever received SNAP benefits (Yes = 1, No = 2)
- **emergency_food**: Respondent received emergency food from food banks, food pantries, or soup kitchens (Yes = 1, No = 2)
- **household_fs_cat**: Food security category (1 = Fully Secure, 2 = Marginal Food Security, 3= Low Food Security, 4 = Very Low Food Security)
- **non_home_meals**: Number of weekly meals not consumed at home
- **fast_food_meals**: Number of weekly fast food meals consumed
- **frozen_meals**: Number of frozen meals consumed in the past 30 days
- **poverty_cat**: Poverty Category. This is based on the ratio of the household's income to the poverty threshold. (1 = Below Poverty Threshold, 2 = At or Up to 130% Above Poverty Threshold, 3 = Above 130% Poverty Threshold)
- **gender**: Gender of the respondent (1 = Male, 2 = Female)
- **age**: Age of respondent in years

Response: **diabetes_diag**: Diabetes diagnosis by a healthcare professional (Yes/No)

Results #2 (Christina) — Food Security, Meal Patterns & Diabetes Logistic Regression:

Data Preparation: The logistic regression model focused on 10 main features in the full model, including NHANES survey respondents' nutritional assistance status, economic background, meals consumed, and their demographics. To begin our analysis, we investigated our data's missing values.

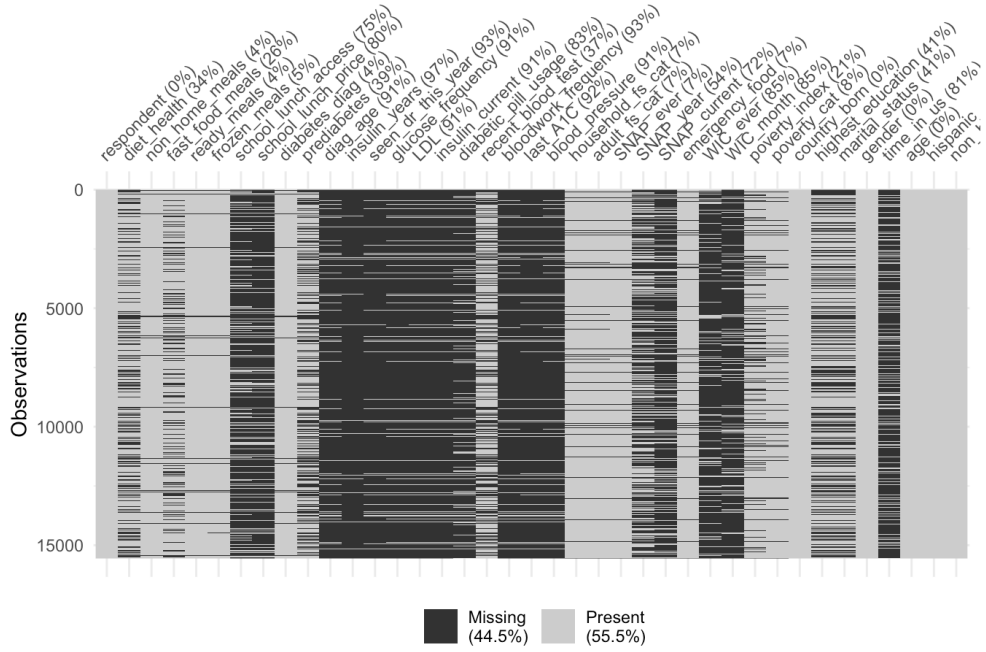


Figure 1: Initial Cleaned Data Set

There was initially a large amount of missing data. Only 55.5% of the full data set was originally present. 44.5% of the full data set is missing. This may be because respondents did not choose to respond to a question that was not applicable to them.

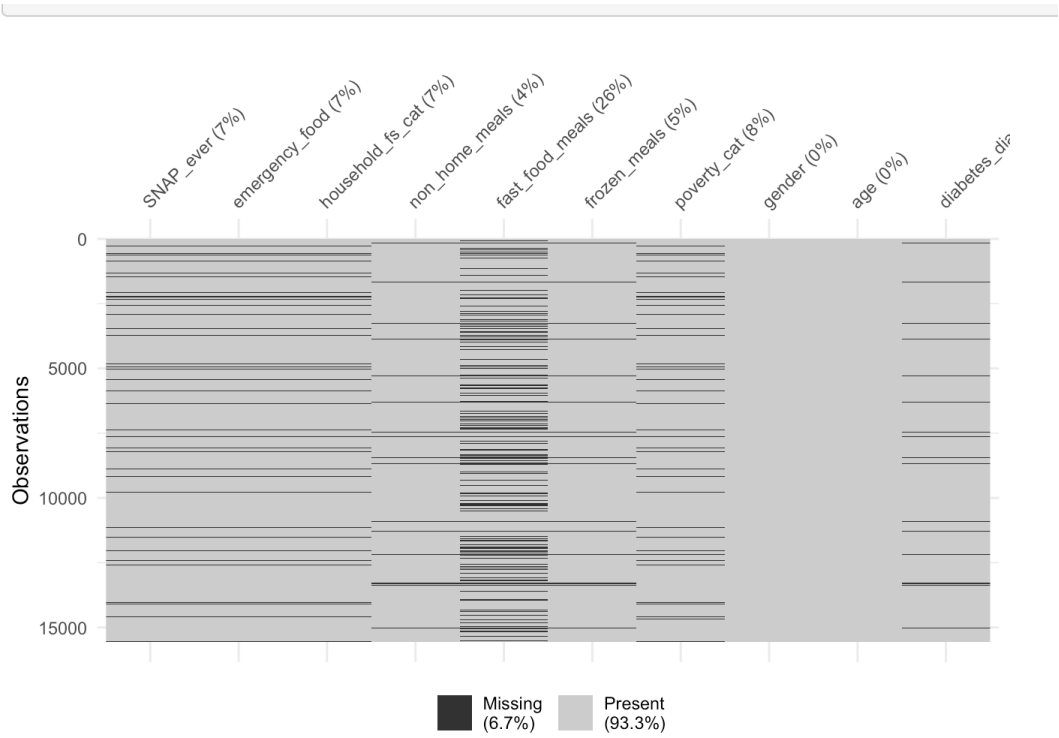


Figure 2: Cleaned Data Set, before special value encoding adjustment

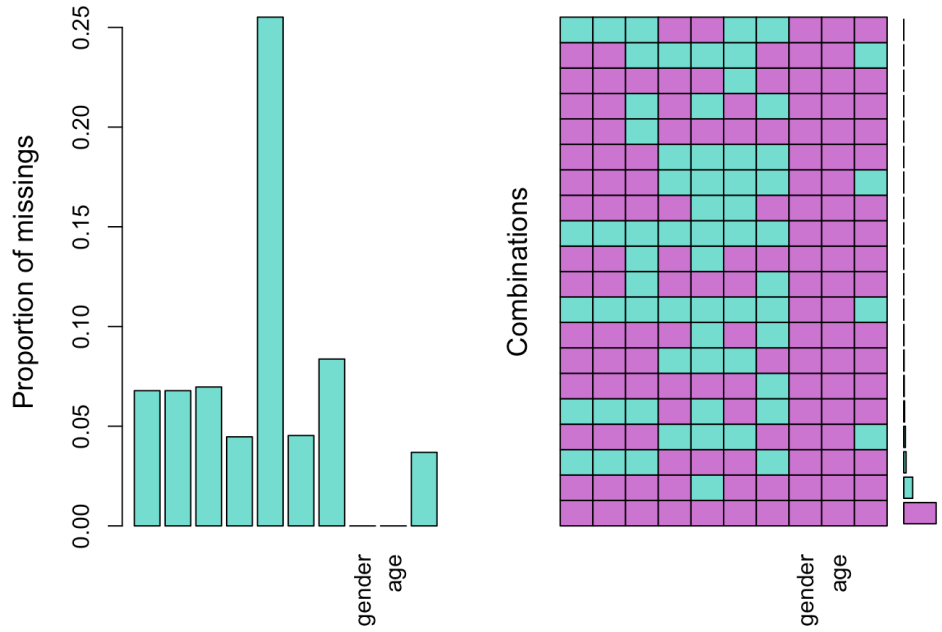


Figure 3: Cleaned Data Set Proportion of Missing Values, before special value encoding adjustment

For our initially selected variables, only 6.7% of our data is missing. Based on the graphs above, gender and age are the most complete variables, while fast food meals are 26% incomplete.

Special-Encoded Values: NHANES has encoded special values for when respondents did not answer a question that was used to generate a column. This is resulting in extreme values, like “9999” and “9”, that we can see in each variable’s summary, despite their variables’ data descriptions (Ley et al., 2014).

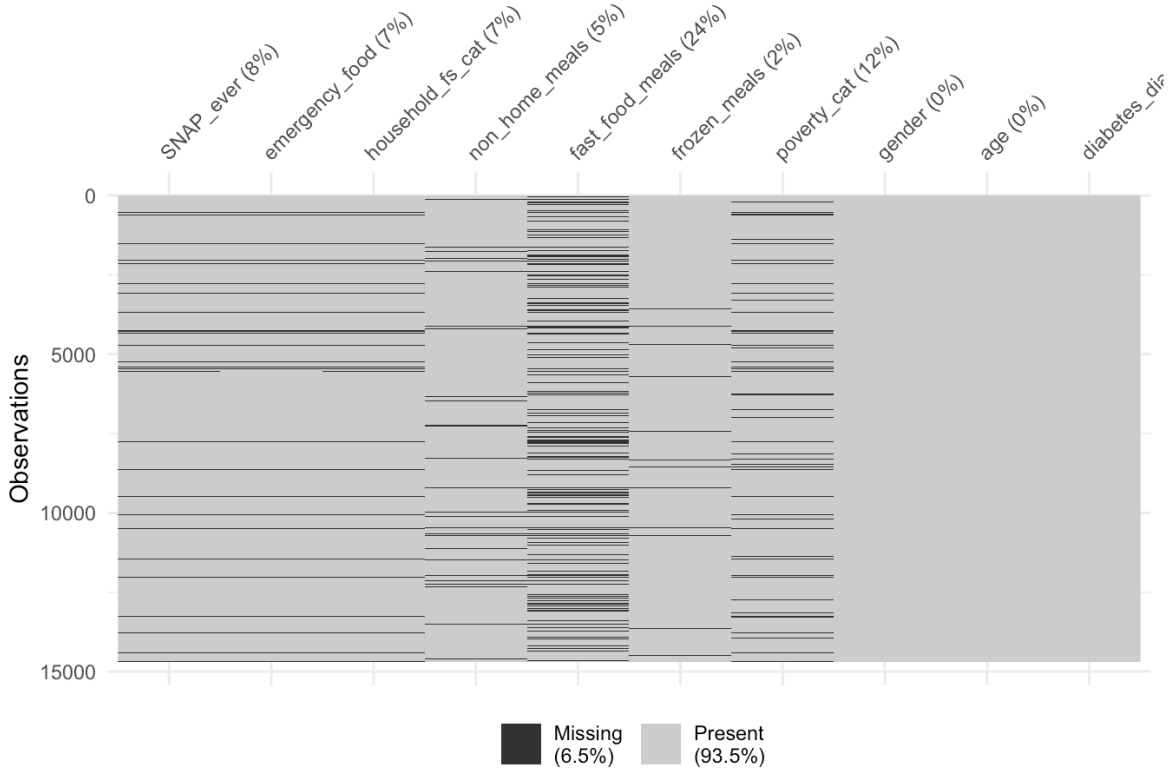


Figure 4: Cleaned Data Set, after special value encoding adjustment

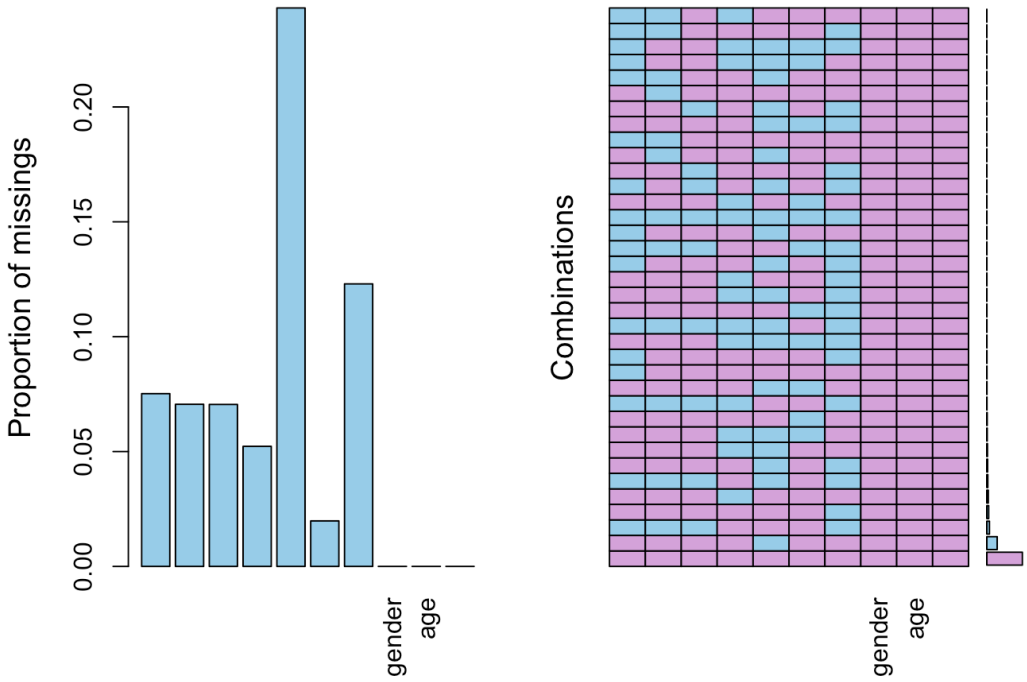


Figure 5: Cleaned Data Set Proportion of Missing Values, after special value encoding adjustment

To address the NHANES special encoded missing values, we can replace each special variable with a missing value (NA). This resulted in increasing our complete data by 0.2% to 93.5%, compared to our cleaned data set with relevant features selected. Overall, we increased our complete data percentage by 38%, compared to the initial data set.

SNAP_ever	emergency_food	household_fs_cat	non_home_meals	fast_food_meals
No	No	Full	1	0
Yes	No	Full	1	0
No	No	Full	3	2
No	No	Full	4	0
No	No	Full	3	0
No	No	Marginal	3	3

frozen_meals	poverty_cat	gender	age	diabetes_diag
5	Above	Male	2	No
1	Below	Female	13	No
2	Above	Male	2	No
0	Above	Female	21	No
0	Below	Female	18	No
0	Below	Male	2	No

Table: Cleaned Data Frame, with renamed factors

We also filled in each factor level, based on NHANES descriptions. For instance, Diabetes Diagnosis was changed to “Yes” and “No”, instead of "1” or “2”, to improve interpretability. Once this data cleaning was complete, the remaining NAs were removed. This approach ensures our predictions reflect actual reported behavior, rather than relying on imputed values that could add artificial patterns into our analysis. The final data set consisted of 9,322 rows and 10 columns.

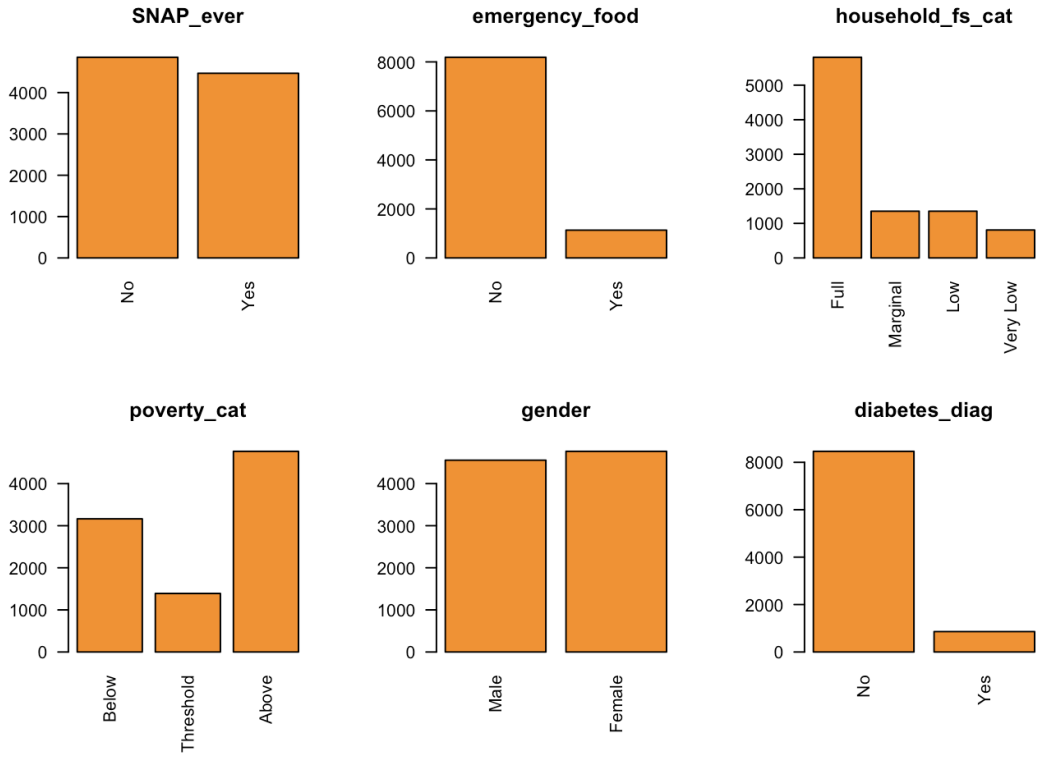


Figure 6: Categorical Features

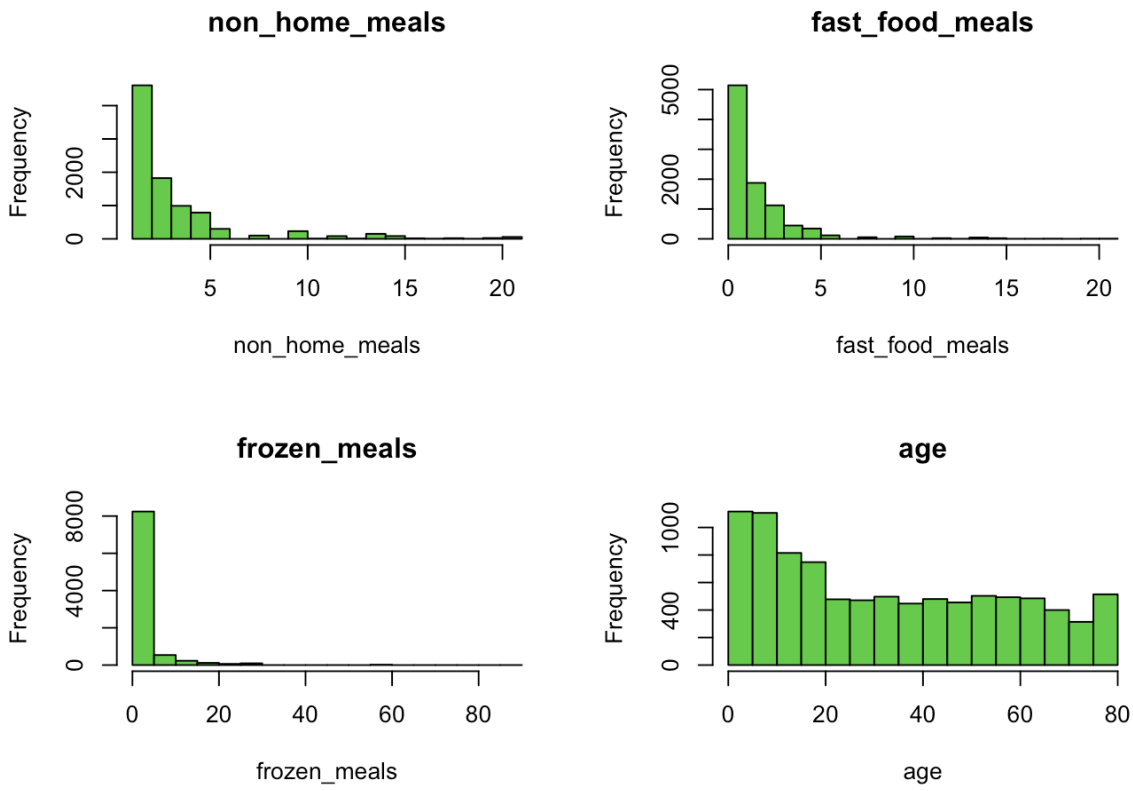


Figure 7: Continuous Features

Exploratory Data Analysis: Next, we explored our features. For diabetic diagnosis, the bar plot shows class imbalance between diabetics and non-diabetics; about 10% of respondents are diabetic. The visualizations reveal that most Americans are food secure, although approximately half of the respondents at poverty or below level. While about half the respondents received SNAP assistance, most respondents did not utilize food bank aid. We also evaluated meal patterns. Most non-home, fast food, and frozen meals are not consumed frequently, as shown by the right-skewed distributions. The graphs show that demographics are fairly balanced.

	GVIF	Df	GVIF^(1/(2*Df))
SNAP_ever	1.366784	1	1.169095
emergency_food	1.266498	1	1.125388
household_fs_cat	1.555760	3	1.076442
non_home_meals	1.935039	1	1.391057
fast_food_meals	1.985057	1	1.408920
frozen_meals	1.018075	1	1.008997
poverty_cat	1.426317	2	1.092833
gender	1.014247	1	1.007098
age	1.136317	1	1.065982

Table: VIF Scores

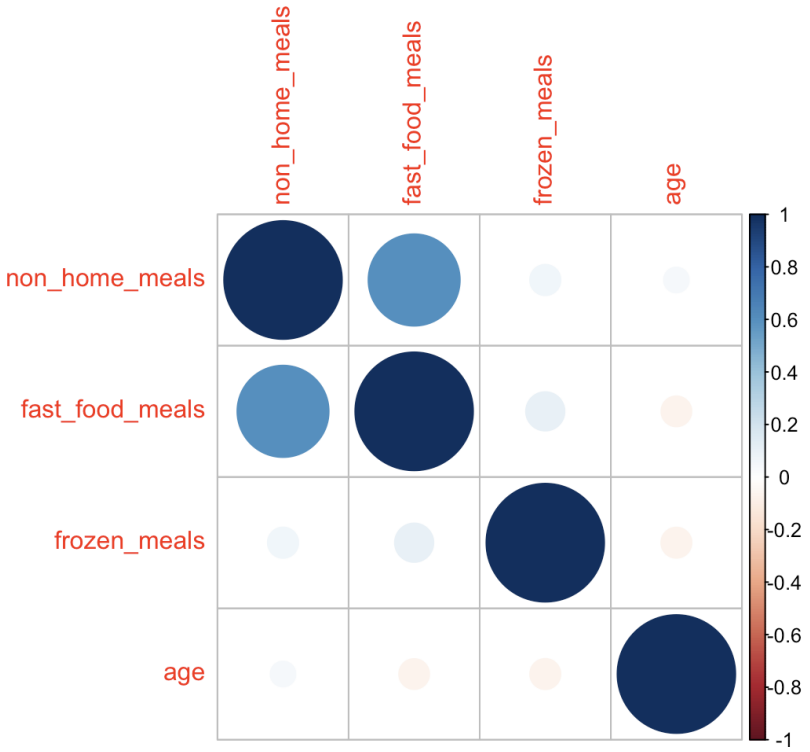


Figure 8: Correlation Plot

In general, there is no correlation among most continuous predictors. However, there is a positive, moderate correlation (0.6014576) between non-home meals and fast food meals. Since the VIFs are relatively low (non-home meals = 1.935039 and fast food meals = 1.985057), we can conclude that this is not a significant reason to remove both predictors from the model. We decided to keep these two features because they represent distinct eating patterns – meals outside of home can capture longer meals at restaurants or friend's home, while fast food meals capture quick meals. There is no problematic correlation among the features. Therefore, regularization is not needed.

Model Fitting: We created various models, beginning with a full model with all predictors included. Stepwise Selection was implemented to reduce the number of features in our full model. When comparing models, the Stepwise Regression model was favored, as the initial stepwise model had a lower AIC (AIC: 4340.3) than the full model (AIC: 4347.1).

For the Stepwise model, all variables are significant (alpha = 0.05): Intercept, Age, Gender (Female), SNAP Ever Recipient (Yes), Household Food Security Category (Marginal, Low, & Very Low), Fast-food meals, and Non-home meals.

Our Stepwise Model has the same significant features as the full model. We have 7 significant features, including the intercept. Compared to the full model, the following coefficients were not included: Emergency Meals (Yes), Frozen Meals and Poverty Category (Threshold & Above)

Compared to the full model, the stepwise model included all features except: Emergency Meals (Yes), Frozen Meals and Poverty Category (Threshold & Above).

Cross Validation: 10-fold Cross Validation was performed on both the full and stepwise regression models. This allowed us to evaluate the performance of the Stepwise regression model compared to the full model. Cross validation was initially performed for both models using a classification threshold of 0.5. To increase our model performance, cross validation also allowed us to find the optimal thresholds for classifying diabetes patients. The optimal cross validation threshold was determined to be 0.65, for both the full cross validation and Stepwise cross validation models. For the Stepwise cross validation model, the overall significance test (with p-value = 0) shows that the Stepwise Cross Validation model has significant explanatory power. Both the Deviance (0.4626126) and Pearsons (0.7464227) overdispersion parameters are less than the overdispersion threshold of 2. Thus, the Stepwise Cross Validation Model is not overdispersed.

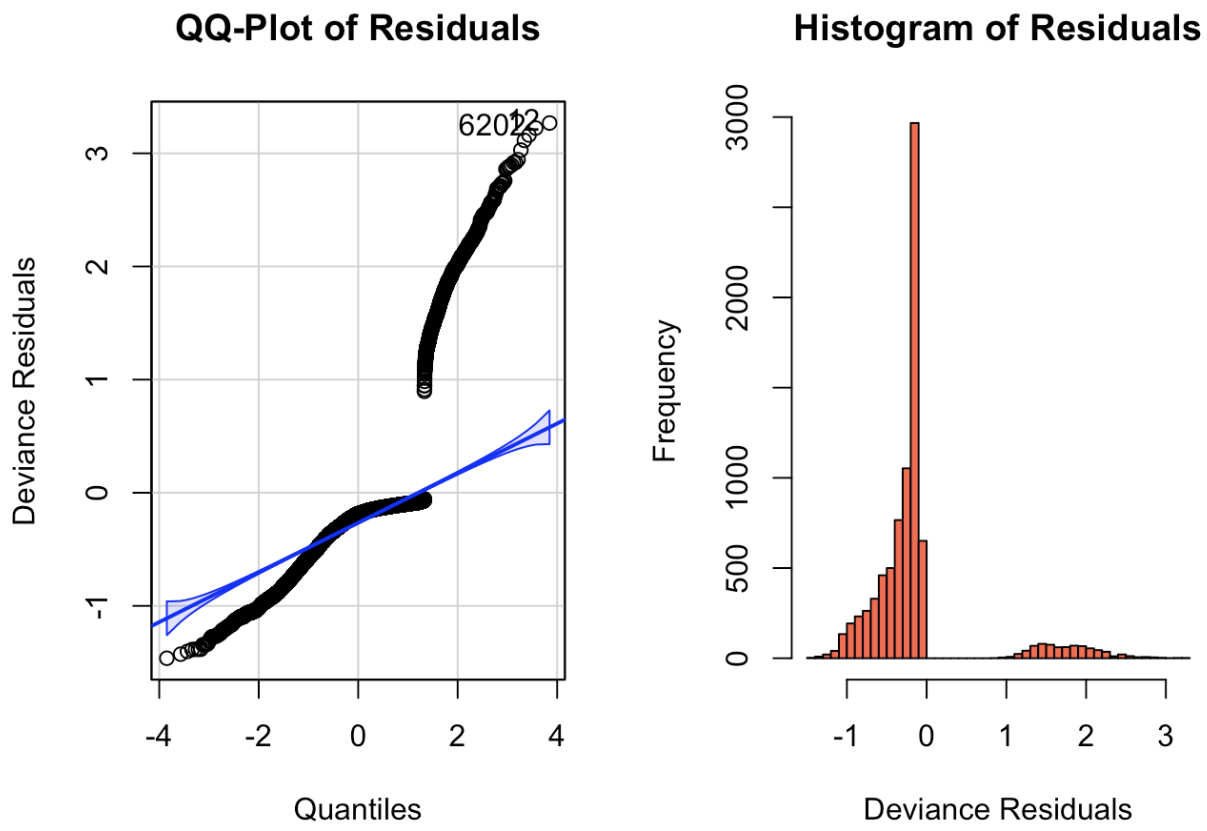


Figure 9: Model Assessment

Model Assessment: The QQ-Plot demonstrates that most residuals do not follow the reference line and follow a bimodial distribution. Similarly, the residual plot does not follow a normal distribution and is possibly bimodial. The Normality Plot and Deviance Residuals Plots both indicate that the residuals are not normally distributed. However, the Deviance and Pearsons Goodness-of-Fit tests Since the p-values of the Goodness-of-Fit tests are 1; we accept the null hypothesis and conclude the model is a good fit. Considering the overall regression and Goodness-of-Fit Deviance and Pearsons tests indicate the model is a good fit, the violations to Goodness-of-Fit graph assumptions may be because diabetes diagnosis are imbalanced.

	Log_Odds	Odds_Ratio	Percent_Change
(Intercept)	-5.64804346	0.003524406	-99.6%
age	0.06762136	1.069960103	7%
SNAP_everYes	0.36671835	1.442991443	44.3%
genderFemale	-0.34823153	0.705935415	-29.4%
household_fs_catMarginal	0.29965514	1.349393371	34.9%
household_fs_catLow	0.35419009	1.425026049	42.5%
household_fs_catVery Low	0.25352092	1.288554334	28.9%
fast_food_meals	0.09096440	1.095230011	9.5%
non_home_meals	-0.04653943	0.954526921	-4.5%

Table 1: Odds Ratio

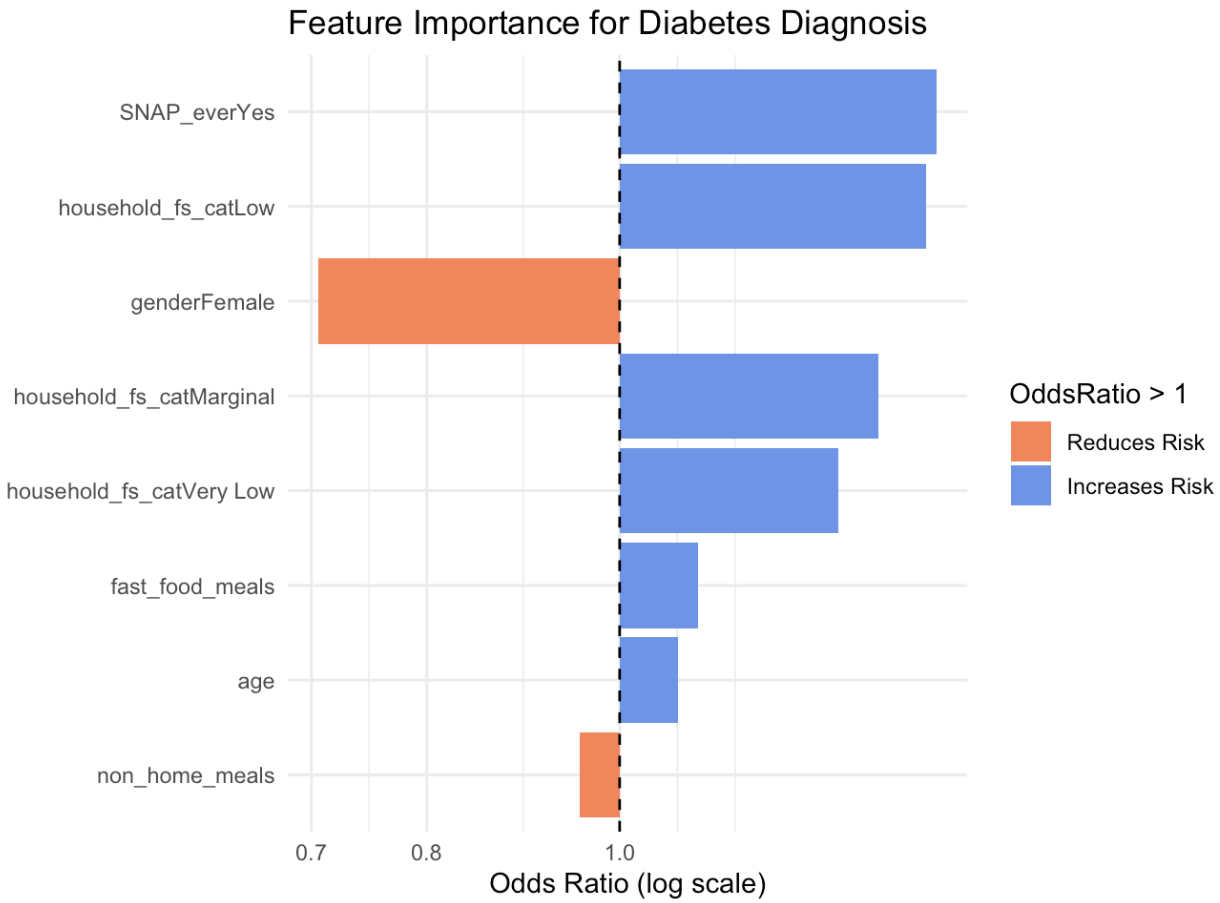


Figure 10: Feature Importance

We can evaluate the odds ratio and feature importance for our Stepwise Cross Validation model. Based on the odds ratio feature importance table and graph above, the low and very low household food security categories have the strongest positive impact on diabetes diagnosis. SNAP benefit recipients and marginal household food security category also have a strong risk, while fast food and age minimally contribute to an increase diabetes risk. SNAP participation is associated with a 44.3% increase in odds of having diabetes. Meanwhile, being female is associate with a 29.4% decrease in odds of having diabetes. On the other hand, being female has a strong reduce risk of diabetes diagnosis. Non-home meals have a slight negative impact in reducing diabetes risk.

Logistic Regression Model (Threshold)	**Sensitivity**	**Specificity**	**Accuracy**
Full CV (Default = 0.50)	0.03604651	0.9940912	0.9057069
Full CV (Optimal = 0.65)	0.001162791	0.9998818	0.9077451

Stepwise CV (Default = 0.50)	0.03604651	0.9943276	0.9059215
Stepwise CV (Optimal = 0.65)	**0.001162791**	**1**	**0.9078524**
Stepwise CV (Threshold F1 Score = 0.16)	**0.6674419**	**0.8317183**	**0.816563**

Table 2: Model Comparison

Model Comparison: When comparing models, we must consider that our data is heavily imbalanced for our response variable, diabetes diagnosis. Approximately 10% of respondents are diabetic. We can evaluate whether lowering the threshold can better classify diabetes patients. This would be better if we were aiming to capture as many diabetic patients as possible. However, if we are looking at the survey results holistically, our previous threshold would remain effective. The table's results above are based on each model's confusion matrix. The best performing model (in blue above), measured by accuracy, is the Stepwise Cross Validation Model, with an accuracy of ~90.79%. All models (excluding the last one) have a maximum ~0.2% difference in accuracy.

We can broaden our approach to aid medical professions better understand diabetic patients. Diabetic patients are the minority class. Compared to the Full Cross Validation Model, the Stepwise Cross Validation Model has decreased sensitivity and increased specificity. In a healthcare setting, prioritizing sensitivity is more critical than maximizing accuracy. Misclassifying a diabetic patient as non-diabetic (false-negative) may delay essential treatment, leading to increased health risks. Thus, a model with higher sensitivity, with a high accuracy, provides more clinical value than maximizing a model's accuracy. This method balances precision and recall and helps us find the optimal value that will ensure we better capture the diabetic patients.

Lowering our classification threshold to 0.16, as found by F1 score, dramatically improved the sensitivity and specificity of our Stepwise Cross Validation Model. This Stepwise Cross Validation model has only a 10% lower accuracy than the other high performing models, yet sensitivity improved by ~67%. This threshold change helps with our imbalanced data. While the lowered threshold does not lead to the best accuracy, as the threshold of 0.65 allowed, this model is better at identifying diabetes respondents.

Method #3 (Ju Ho) – Decision Trees

A decision tree model can be used to classify whether an individual will have diabetes by evaluating health indicators and demographic variables. The model can identify the most important variables, and we can evaluate model performance using a confusion matrix. This approach expands the scope by considering additional factors that can have a more significant impact on health outcomes. Decision trees offer greater transparency and are more naturally equipped to manage categorical variables.

Table 1: Descriptions of Predictor Variables

Variable Name	Description	Type
age	Age in years	Numerical
gender	Gender	Categorical (binary)
marital_status	Marital status (married, divorced, never married)	Categorical
highest_education	Highest completed education	Categorical
poverty_index	Measure of income	Numerical (continuous)
diet_health	Respondent's opinion of their diet health	Categorical
non_home_meals	Number of meals not home prepared in a week	Numerical
fast_food_meals	Number of fast food meals in a week	Numerical
ready_meals	Number of ready to eat grocery meals outside the home in a month	Numerical
frozen_meals	Number of frozen meals in a month	Numerical
prediabetes	Whether or not respondent has had prediabetes	Categorical (binary)
recent_blood_test	Whether or not respondent has had a blood test in the past 3 years	Categorical (binary)
diabetes_diag (response variable)	Whether or not respondent has diabetes	Categorical (binary)

Analysis #3 (Ju Ho) – Decision Trees

Diabetes can be influenced by a wide range of determinants across demographics, income, clinical indicators, and dietary patterns. We used a decision tree model with 12 variables (references in Table 1) from the combined NHANES dataset to analyze the relationships between diabetes and these predictor variables.

Data Cleaning: Many of these variables had categories for a small number of missing responses and values greater than the common range (for example, number of meals > 21 in a week). These were replaced with NA and excluded from the dataset. Other missing values were imputed using the mode of each variable. The categorical variables were factorized and the categorical variable values which were indicated using numbers were renamed for better interpretability. The final dataset had 11756 observations and 13 columns (12 predictors, 1 response).

Data Exploration: Exploratory analysis of the numerical variables shows that the proportion of people with diabetes increases as age increases. The proportion of people with diabetes looks consistent across poverty index levels and the four meal type variables.

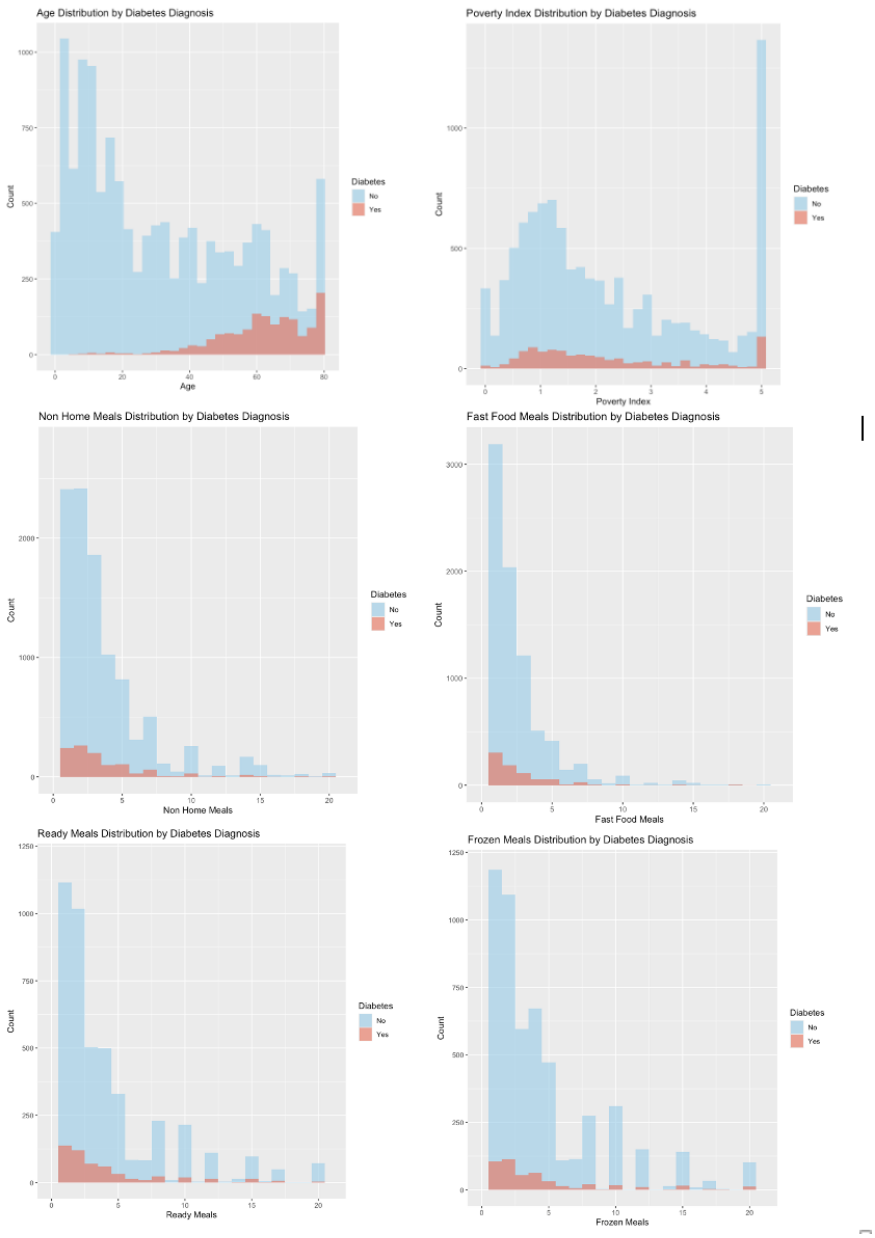
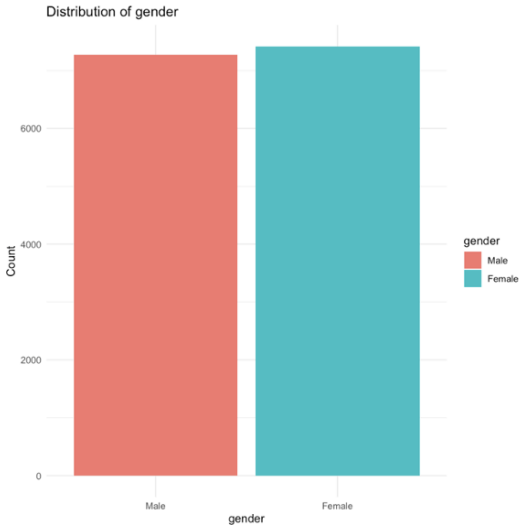
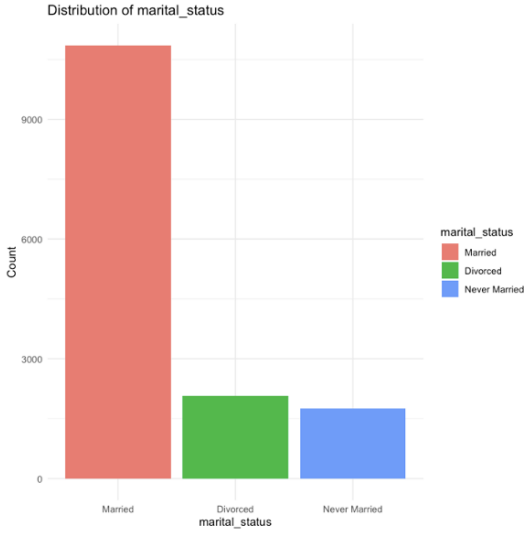
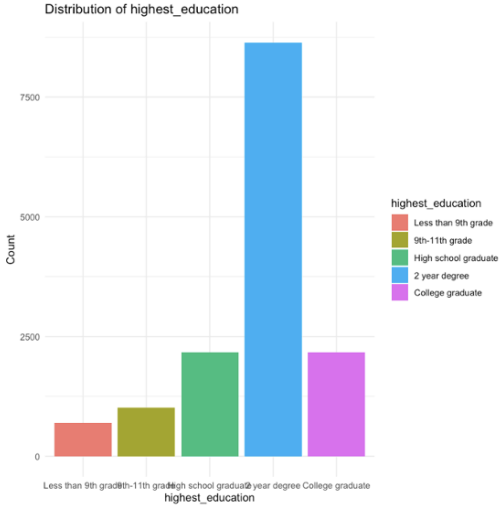


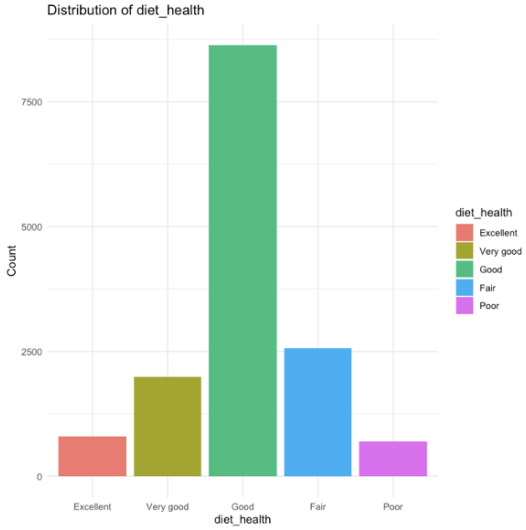
Figure 1: Distributions of Numerical Variables

The response variable is imbalanced, as we have a small proportion of people with diabetes compared to the size of the dataset. The other demographics variables are also imbalanced, but variables such as highest_education and diet_health are likely to be a representation of the population, as most people have post-secondary education, and many people would self-identify as having average diet health.


Distribution of gender
Count
6000
4000
2000
0
Male
Female
gender
gender
Male
Female


Distribution of marital_status
Count
9000
6000
3000
0
Married
Divorced
Never Married
marital_status
marital_status
Married
Divorced
Never Married


Distribution of highest_education
Count
7500
5000
2500
0
Less than 9th grade
9th-11th grade
High school graduate
2 year degree
College graduate
highest_education
highest_education
Less than 9th grade
9th-11th grade
High school graduate
2 year degree
College graduate


Distribution of diet_health
Count
7500
5000
2500
0
Excellent
Very good
Good
Fair
Poor
diet_health
diet_health
Excellent
Very good
Good
Fair
Poor

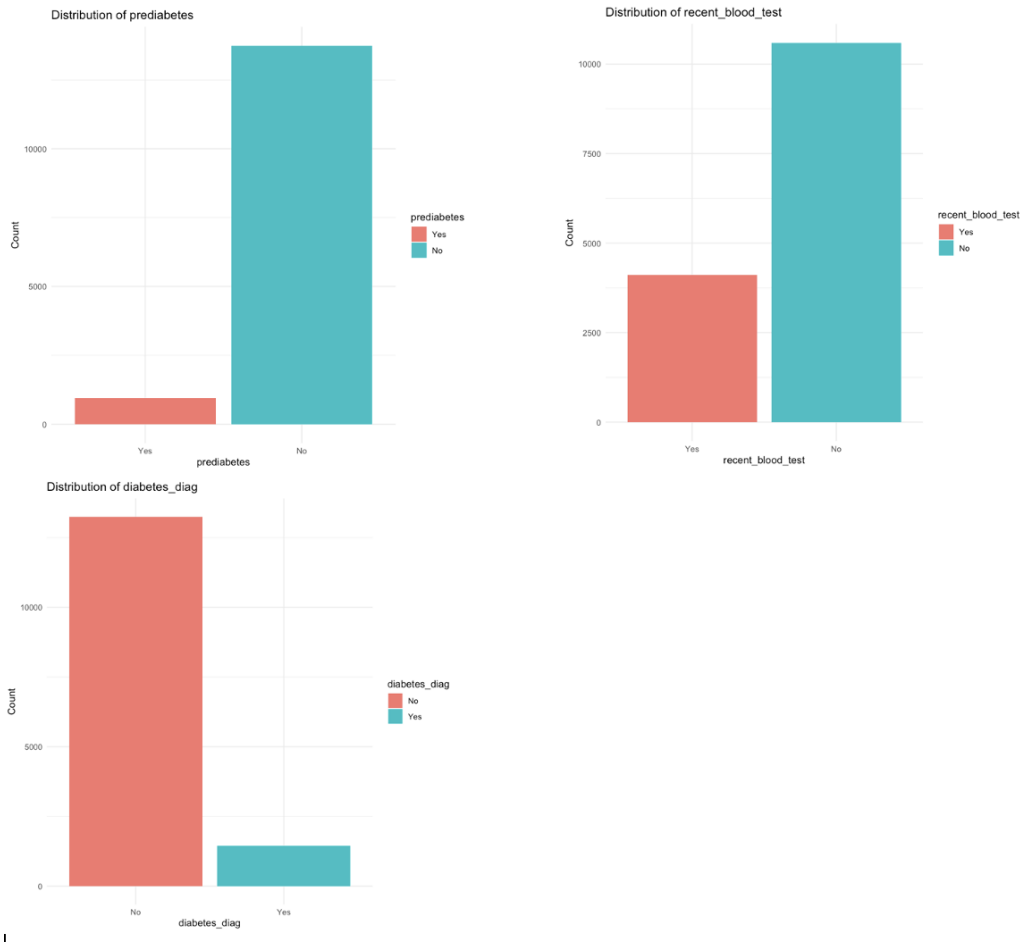


Figure 2: Distributions of Categorical Variables

Modeling: The cleaned data was split into a train-test split of 80-20. A decision tree model was trained using all variables with 5-fold cross-validation. The mean accuracy is 91%, but the No Information Rate (NIR) is 0.9057, which is the accuracy that would be achieved by always predicting the majority class. This highlights the problem of response variable imbalance in the training dataset, as only 10% of the response variable is the minority class (has diabetes) and 90% is the majority class (does not have diabetes). The low sensitivity indicates that the model correctly identifies 32.85% of the people who have diabetes and is missing around 67% of true positives. The specificity is high at 96.47% so the model has a low false positive rate and has a high rate of correctly identifying people who do not have diabetes. Overall, the initial model is heavily biased toward the majority class and accuracy is not a good measure for model performance.

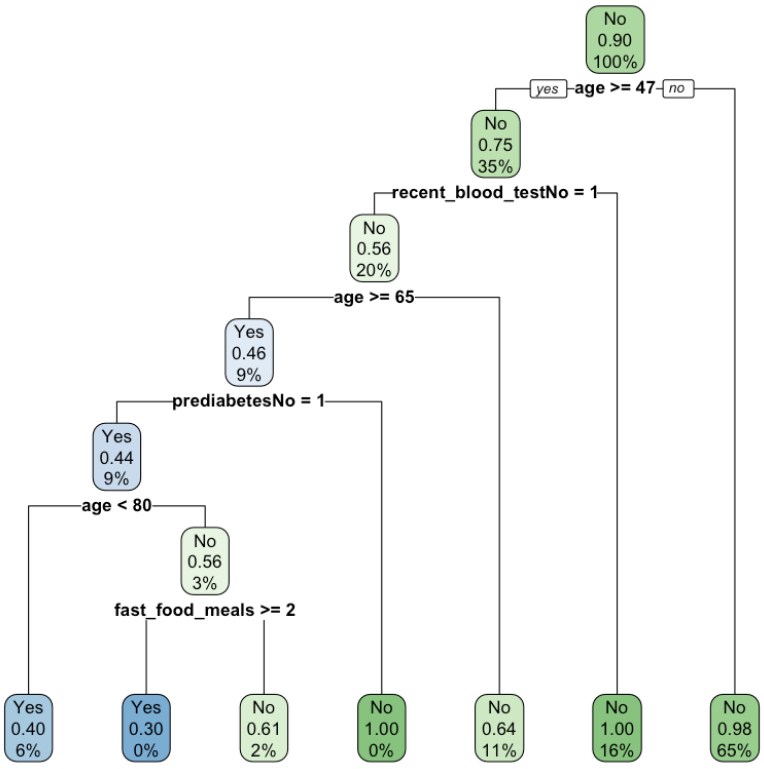


Figure 3: Decision Tree Model with All Predictor Variables

The most important variables are whether a person has had a recent blood test, age, prediabetes, and education. Marital status and diet health are moderately important, and the poverty index and diet choices variables are least important. The least important predictors can be removed to improve efficiency and reduce model overfitting.

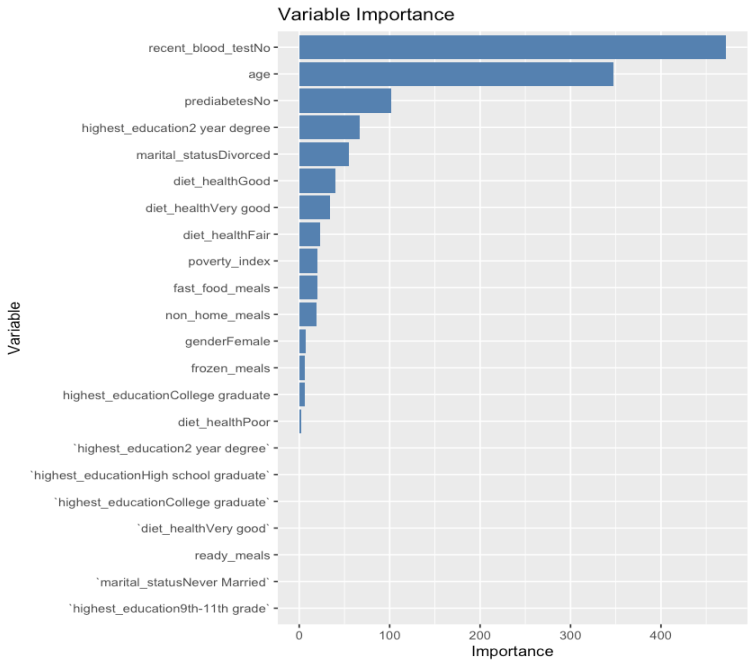


Figure 4: Variable Importance in Decision Tree

A second model was fit without the following variables: poverty_index, non_home_meals, fast_food_meals, frozen_meals, ready_meals. To account for the weight imbalance in the response variable, the data points with diabetes was given double the weight of the data points without diabetes.

Metric	Model 1	Model 2	Better
Accuracy	0.9047	0.885	Model 1
Sensitivity/Recall	0.32852	0.85120	Model 2
Specificity	0.96467	0.88012	Model 1
Balanced Accuracy	0.64660	0.86982	Model 2
F1 Score	0.394	0.575	Model 2

Table 2: Confusion Matrix Results of Model 1 and Model 2

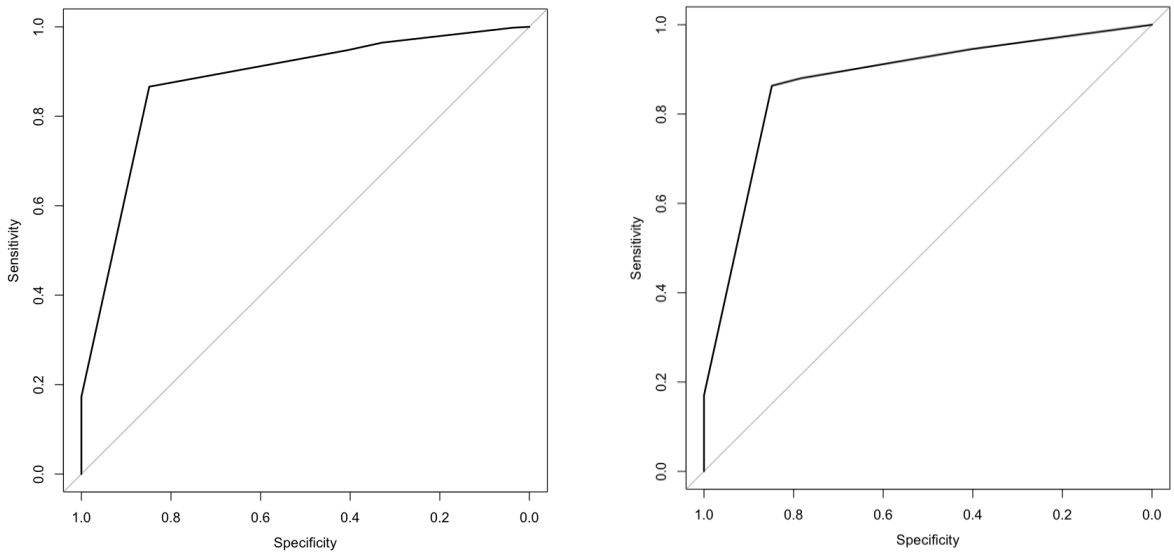


Figure 5: ROC Curves for Models 1 and 2

Model 1 is better at identifying the majority class but worse at predicting individuals who have diabetes. Despite the higher accuracy, the class imbalance makes this value misleading, and the balanced accuracy is lower than model 2. Model 2 has higher balanced accuracy and sensitivity, it is better at identifying true positives, which can predict real cases of diabetes. There is not much difference between the ROC curves, which means that model2 still struggles to handle the response variable imbalance. In this healthcare context, it is more important to accurately predict true positives than true negatives, as it is more important to identify when someone has diabetes.

Method #4 (Naomi) — Poisson Regression:

To examine the relationship between food insecurity and diabetes prevalence, a Poisson regression model was used to analyze aggregated data from the CDC, focusing on poverty index categories. The outcome variable represents the count of diabetes cases within each poverty group, with total population size as an offset to account for varying exposure across groups. The model includes predictors such as food insecurity indicators, dietary patterns, and demographic characteristics, all of which may influence diabetes risk.

The analysis considers both economic and physical barriers to food access, capturing how these factors contribute to health disparities in diabetes prevalence. Variables like SNAP participation, emergency food reliance, and dietary habits such as the number of fast food and frozen meals consumed, reflect broader issues of food insecurity that are shaped by socioeconomic conditions and access to nutritious food. These elements are particularly relevant in understanding how economic strain and food availability impact health outcomes in underserved communities.

This methodology highlights the importance of structural factors in shaping diabetes risk, where food insecurity, socioeconomic status, and neighborhood characteristics interact to influence health. By modeling diabetes cases across different poverty levels, the analysis identifies vulnerable groups and emphasizes the need for systemic interventions that address food

access, economic disparities, and social determinants of health. The following variables were created and used in the model using data from the NHANES survey:

Explanatory Variables (Aggregated at the Poverty Index Category Level):

- **avg_non_home_meals:** The average number of meals per week that were not consumed at home among respondents in each poverty category group.
- **avg_fast_food_meals**: The average number of fast food meals consumed per week among respondents in each group.
- **avg_ready_meals**: The average number of ready-to-eat meals (e.g., prepared supermarket meals) consumed per week in each group.
- **avg_frozen_meals**: The average number of frozen meals consumed in the past 30 days per respondent in each group.
- **pct_prediabetes**: The percentage of respondents in each group who reported a prediabetes diagnosis.
- **SNAP_ever_count:** The number of individuals in each group who reported ever receiving SNAP (Supplemental Nutrition Assistance Program) benefits.
- **pct_emergency_food:** The percentage of respondents in each group who reported receiving emergency food from food banks, pantries, or soup kitchens.
- **pct_female:** The percentage of female respondents in each poverty group (based on gender = 2).
- **avg_age:** The average age of respondents in each group.
- **poverty_category:** A categorical variable based on income-to-poverty threshold ratios, grouped into bins such as “1.0–1.1”, “1.1–1.2”, etc.

Offset (Exposure):

- **total_people:** The total number of survey respondents in each poverty category group. This is used as an offset in the Poisson model to account for differences in group size.

Response Variable:

- **diabetes_cases:** The number of individuals in each poverty category group who reported having been diagnosed with diabetes by a healthcare professional.

Analysis #4 (Naomi) — Poisson Regression:

The Poisson regression model focused on 10 main variables in the full model, including aggregates of NHANES survey respondents’ nutritional assistance status, economic background, meals consumed, and their demographics. To begin the analysis, the data had to be cleaned and manipulated so that it would be suited for Poisson regression.*Data Preparation:* The data cleaning and manipulation process involved several steps to ensure the dataset was properly prepared for analysis. First, rows with missing or invalid values for critical variables like diabetes_diag and poverty_category were removed to maintain data integrity. The poverty_category variable was then created by categorizing the continuous poverty_index variable into more granular groups based on income relative to the poverty threshold. These categories range from "<1.0" to "4.9-5.0" to reflect different levels of poverty in the dataset. To further refine the data, placeholder values such as 9999 and 7777, which indicated missing or uncertain responses, were replaced with NA. Additionally, summary statistics for dietary habits (e.g., avg_non_home_meals, avg_fast_food_meals), age, and gender were aggregated within

each poverty category. This aggregation ensured consistency and reduced variability across the dataset. The cleaned and manipulated dataset was then ready for exploratory data analysis.

Exploratory Data Analysis: For the exploratory data analysis (EDA), I began by examining the relationship between various variables and the prevalence of diabetes to uncover trends and patterns that might inform the model development process. First, I created a scatter plot to explore the association between the number of people receiving SNAP sistance (SNAP_ever_count) and the count of diabetes cases. By including a linear regression line with geom_smooth(method = "lm"), I was able to visually assess that there was a linear relationship between SNAP participation and diabetes cases. This plot helped me identify correlations, suggesting that higher SNAP participation might have an impact on diabetes prevalence rates compared to those with lower participation. This could indicate that food insecurity, linked to reliance on SNAP, might be a contributing factor to diabetes cases, warranting further investigation.

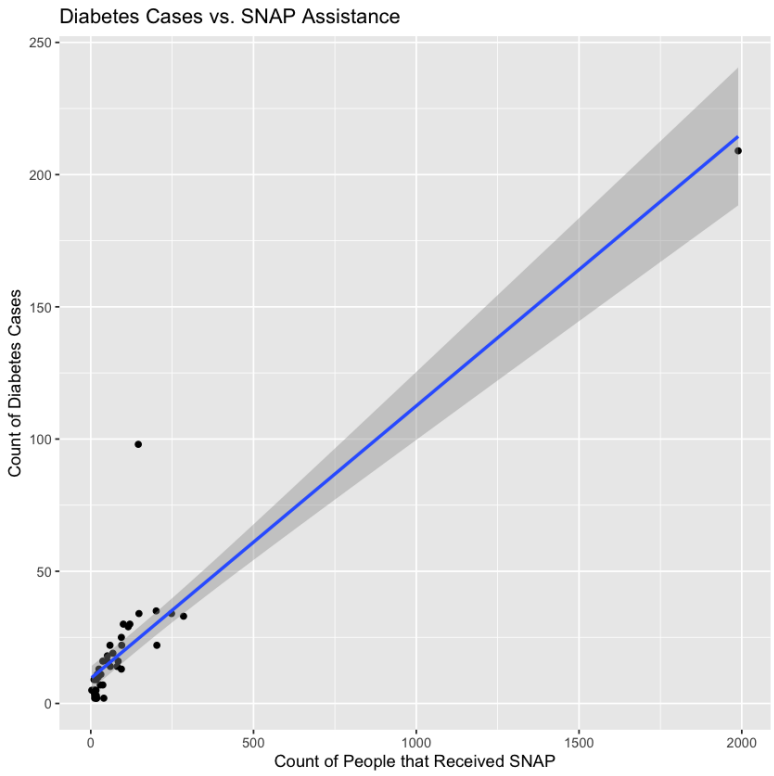


Figure 1: Scatter plot of Diabetes Cases vs. SNAP Assistance

Next, I introduced a new variable, diabetes_rate, which was calculated by dividing the number of diabetes cases by the total population. This new metric allowed for a more standardized measure of diabetes prevalence across different populations, enabling easier comparisons. To further analyze diabetes prevalence in relation to socioeconomic factors, I created a bar chart to visualize how diabetes rates varied across different poverty categories (poverty_category). This

bar chart provided insights into whether poverty, as indicated by the poverty category, had a clear association with higher or lower diabetes rates. The visualization showed whether individuals in lower poverty categories experienced a higher prevalence of diabetes, supporting the hypothesis that financial and food insecurity might be key determinants of diabetes outcomes.

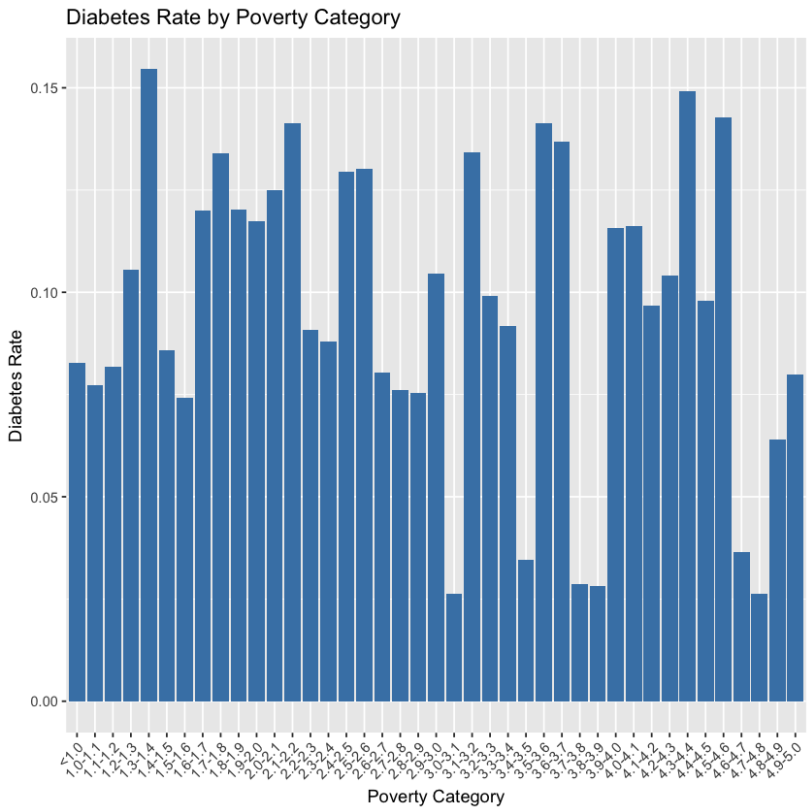


Figure 2: Histogram of Diabetes Case Count per Poverty Index Category

The results from these visualizations and the initial exploration of relationships were important for understanding the broader context of food insecurity and diabetes prevalence before moving on to more formal statistical modeling. By visualizing these relationships, I was able to identify trends and guide the interpretation of the relationships between food insecurity, poverty, and diabetes outcomes.

Modeling: In the modeling phase, a Poisson regression model was employed to predict diabetes cases based on various independent variables. I used log(total_people) as an offset in the model to account for the different population sizes across areas, ensuring that the predicted diabetes cases were not influenced by the overall population size. The independent variables included measures related to food consumption, food insecurity, and demographic factors such as age and gender. The model was fit using the poisson family with a log link function, which is appropriate for count data like the number of diabetes cases per poverty index category.

The coefficients of this initial model, model1, provided important insights into the relationships between the predictors and the outcome. For instance, avg_non_home_meals had a statistically significant negative relationship with diabetes cases (p = 0.040), suggesting that as the average number of non-home meals increases, diabetes cases tend to decrease. This could be interpreted as non-home meals potentially being linked to healthier eating behaviors, or the fact that individuals who eat away from home might have access to better food options. On the other hand, variables such as avg_fast_food_meals, avg_ready_meals, and SNAP_ever_count did not show significant associations with diabetes cases, suggesting that these factors may not have a direct impact in this model or that their effect is weak relative to the other predictors. The avg_age variable, however, exhibited a strong positive relationship with diabetes cases (p < 0.001), indicating that older populations are at a higher risk for diabetes, which aligns with established health trends.

```
Call:
glm(formula = diabetes_cases ~ avg_non_home_meals + avg_fast_food_meals +
    avg_ready_meals + avg_frozen_meals + pct_prediabetes + SNAP_ever_count +
    pct_emergency_food + pct_female + avg_age, family = poisson(link = "log"),
    data = train_summary, offset = log(total_people))

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.630e+00  1.102e+00  -3.293  0.00099 ***
avg_non_home_meals  -2.280e-01  1.112e-01  -2.050  0.04032 *
avg_fast_food_meals  1.235e-01  1.823e-01   0.678  0.49808
avg_ready_meals     -2.545e-02  9.516e-02  -0.267  0.78913
avg_frozen_meals    -1.093e-02  1.042e-01  -0.105  0.91643
pct_prediabetes     -7.880e-03  1.498e-02  -0.526  0.59901
SNAP_ever_count     -4.787e-06  8.057e-05  -0.059  0.95263
pct_emergency_food   8.644e-03  7.566e-03   1.143  0.25323
pct_female          -4.316e-03  1.248e-02  -0.346  0.72940
avg_age              5.512e-02  1.301e-02   4.237 2.26e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 70.396  on 40  degrees of freedom
Residual deviance: 37.032  on 31  degrees of freedom
AIC: 235.05

Number of Fisher Scoring iterations: 4
```

Figure 3: Summary of Model1

I assessed the overall significance of model1, finding that the likelihood ratio test indicated a highly significant improvement in model fit compared to the null model (p-value ≈ 0.0001), suggesting that the predictors included in the model meaningfully contributed to explaining the variance in diabetes cases. I also checked for overdispersion, with a value of 1.19 for model1, which based on a threshold of 2 indicates that overdispersion is not present.

After fitting the initial model, I performed stepwise regression to simplify model1 and improve interpretability. The stepwise model, model2, included only avg_non_home_meals and avg_age, both of which were statistically significant predictors. The AIC of model2 (224.14) was slightly lower than that of model1 (235.05), suggesting that the simpler model might provide a more accurate representation of the data.

```
Call:
glm(formula = diabetes_cases ~ avg_non_home_meals + avg_age,
    family = poisson(link = "log"), data = train_summary, offset = log(total_people))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.962951   0.240451 -12.322  < 2e-16 ***
avg_non_home_meals -0.328052   0.076564  -4.285 1.83e-05 ***
avg_age            0.044118   0.008027   5.496 3.88e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 70.396  on 40  degrees of freedom
Residual deviance: 40.117  on 38  degrees of freedom
AIC: 224.14

Number of Fisher Scoring iterations: 4
```

Figure 4: Summary of Model2

The likelihood ratio test between the two models indicated that the additional variables in model1 did not provide a significant improvement ($p \approx 0.88$), further supporting the sufficiency of the simpler model for predicting diabetes cases.

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	38	40.11657	NA	NA	NA
2	31	37.03247	7	3.084101	0.8771147

Table 1: Likelihood Ratio Test Output

In terms of model performance, the RMSE for model1 was 2.32, and for model2, it was slightly lower at 2.25, indicating that both models provided relatively similar predictions. The model deviance values (37.03 for model1 and 40.12 for model2) were also comparable, further confirming the adequacy of the simpler model.

Overall, the analysis suggests that avg_non_home_meals and avg_age are the most important predictors for counts of diabetes cases, while other factors like food insecurity and SNAP assistance may not play a significant role in this dataset. These findings will be helpful in guiding future health interventions and resource allocation strategies.

Explanation of Changes

Research Question: Our original research question was "Can we identify individuals at the highest risk for diabetes based on food security and economic indicators?" As our analysis expanded, we revised this question to shift the focus from identifying individuals to understanding the broader relationship between food insecurity, socioeconomic factors, and diabetes. The revised research question, *"To what extent do food insecurity and socioeconomic determinants influence diabetes prevalence in the United States?*" better aligns with our logistic regression, decision tree, and Poisson regression models. This revision captures our aim to explore both individual features, along with food access and economic status patterns, to explain diabetes diagnosis.

Data: In our original Analysis Plan, we planned to use 3 data files, derived from the United States Department of Agriculture (USDA) and Centers for Disease Control (CDC)'s National Health and Nutrition Examination Survey (NHANES). The sources were:

1) USDA's Food Access Atlas (Economic Research Service, n.d.).

2) CDC's Chronic Illness Indicators (U.S. Chronic Disease Indicators, 2024).

3) CDC's Behavioral Risk & Diabetes Prevalence (Centers for Disease Control and Prevention, n.d.-c).

We intended to merge these files on a shared U.S. states column. As our project progressed, we decided to focus exclusively on the NHANES data and selected 5 relevant NHANES data files that incorporated diabetes diagnosis, food security, and socioeconomic factors. This revised approach made merging easier, expanded our data size to 15,560 rows rather than 1 row per state. The revised approach also prevented data loss, as we merged on a shared respondent ID column.

Poisson Regression Plan (Method #4): The original plan outlined in the Analysis Plan for the Poisson regression model was to examine the relationship between food insecurity and diabetes prevalence at the state level. Since we changed the data sources, the data no longer had state level information available. The approach was then adapted appropriately so that a Poisson regression model could be created using the NHANES data from the CDC.

Logistic Regression Plan (Method #1): The original plan outlined in the Analysis Plan was to use Linear Regression modeling to explore changes in biometrics, such as glycemic levels or HbA1c (response) and changes in FI (predictor). Since food insecurity is related to many other factors, variables to also consider include demographics, dietary quality, and socioeconomic status. We then chose to use a logistic regression instead of a linear regression model as initially indicated in the analysis plan because it is useful for predicting binary outcomes using multiple predictors.

Conclusions

The results from our four models (2 logistic regression, 1 decision tree, 1 Poisson regression) all concluded a similar result, which was that factors related to food insecurity and demographics were highly correlated to diabetes prevalence.

Successes:

The explanatory models conducted were the logistic regression model (method #1) and Poisson regression model, while the predictive models included logistic regression model (method #2) and decision tree. Logistic regression model (Method #2) attained ~90.79% accuracy through Stepwise 10-fold Cross Validation. The decision tree showed that Model 2 had a higher balanced accuracy and sensitivity, which is better at identifying true positives for predicting real cases of diabetes. Methods 2 and 3 yielded the best results of approximately 91% accuracy, but due to the highly imbalanced response variable, it may not be the best measure of performance as it gives a misleading impression of model performance. Sensitivity is a better measure for detecting true positive cases of diabetes, which is the primary interest in our healthcare context. Method 2's stepwise CV model had the highest sensitivity at 67% and Method 3's decision tree model with class weightings had the highest sensitivity at 85%. Through careful metric selection and model tuning, we were successful in developing models to predict diabetes using food security, demographics, and health indicator variables.

- The *Food Insecurity, Demographics, and Diabetes Prevalence Logistic Regression Model (Method #1)* showed statistical significance between the response variable *(diabetes_binary*) and predictors *household_fs_cat, age,* and *race_ethnicity (*p-value <0.05*)*. There was a positive association between diabetes and age and distinct disparity between each race group. The analysis also showed a non-significant relationship between the response and predictors *poverty_index* and *SNAP_current.* This information helps to identify targeted interventions, such as pre-diabetes screening, to reduce the prevalence of diabetes in a population.
- The *Food Security, Meal Patterns, & Diabetes Logistic Regression Model (Method #2)* reveals significant food insecurity impacts on diabetes. Through data cleaning and deciphering special encoded NHANES records, this method attained ~90.79% accuracy through Stepwise 10-fold Cross Validation. VIF assessment confirmed that there is minimal multicollinearity among predictors. To better aid healthcare diagnosis for diabetes, f1-score optimization achieved a better cross validation threshold and dramatically improved sensitivity from ~3.60% to ~66.74%. The feature importance plot visually demonstrated that SNAP participation and low food security categories most strongly influence diabetes risk, with the odds ratio table showing marginal, low, and very low food insecurity increase diabetes by ~34.9%, ~42.5%, and ~28.9%, respectively, while SNAP participation is associated with a ~44.3% higher risk of diabetes. The feature importance also demonstrated that being female strongly decreases diabetes risk, with the odds ratio table showing a ~29.4% decreased risk for diabetes.
- *Sociodemographic and Health Determinants of Diabetes (Method #3):* The second model was able to make significant improvement in predicting true instances of diabetes, resulting in sensitivity of 85%. In our healthcare context, it is more important to correctly predict diabetes than it is to correctly predict that an individual does not have diabetes. The models also identified that age and whether or not a person has had recent health

checkups (blood tests, testing for prediabetes), are the most important determinants of diabetes.

- *The Structural Impact of Food Insecurity on Diabetes Management (Method #4)*: This analysis successfully utilized a Poisson regression model with an offset to account for varying population sizes when predicting diabetes cases across poverty index categories. The initial model revealed a statistically significant overall fit ($p < 0.001$) and provided insight into how factors like non-home meal frequency and average age relate to diabetes case counts. After stepwise model selection, the final model retained only these two predictors, both of which were statistically significant and interpretable: more frequent non-home meal consumption was associated with fewer diabetes cases, while higher average age was associated with more cases. The model demonstrated reasonable predictive performance on the test data, with a low RMSE of approximately 2.25. Furthermore, the dispersion statistic was close to 1 (1.06), indicating that the Poisson assumption of no overdispersion was met.

Limitations/Further Investigation:

We integrated multiple datasets from the NHANES survey to build upon previous research to address any gaps in prior studies. Although these surveys allow us to integrate several socioeconomic and demographic covariates to provide a more comprehensive overview on the relationship between food insecurity and diabetes prevalence, self-reported data can introduce selection bias. Additionally, multicollinearity among some of these covariates may affect overall model performance. To address these limitations in the future, we could apply sampling weights to adjust for the probability of bias.

- The *Food Insecurity, Demographics, and Diabetes Prevalence Logistic Regression Model (Method #1)* encountered several challenges including significant amounts of missing data that required imputing. Additionally, the data that we gathered from the NHANES survey were self-reported can introduce selection bias. Additionally, multicollinearity among some of these covariates may affect overall model performance. To address these limitations in the future, we could apply sampling weights to adjust for the probability of bias.
- The *Food Security, Meal Patterns, & Diabetes Logistic Regression Model (Method #2)* faced substantial challenges. The data set had significant class imbalance, with ~10% of the respondents were diabetic. Data cleaning required substantial reduction from 15,560 to 9,322 respondents. While the model had a good fit via goodness of fit tests, the QQ-Plot and Histogram of Deviance Residuals suggests there is a binomial distribution, or significant class imbalance that the model is not fully capturing for diabetic classifications. The model faced a significant trade-off with performance accuracy and diabetes detection, as the best model had a ~90.79% accuracy and ~0.12% sensitivity, while optimizing f1-score to change the classification threshold created a model with ~81.66% accuracy and ~66.74% sensitivity
- *Sociodemographic and Health Determinants of Diabetes (Method #3):* The decision tree models revealed the challenges of the imbalanced response variable, as the models were not good at predicting the minority class of people who had diabetes. To account for this imbalance, a second model was trained, with double the weight for the minority class, which performed much better in predicting the cases of individuals with diabetes, but did not show significant improvement in the ROC curves. Due to the imbalanced classes, evaluating models with sensitivity/recall is more interpretable for this problem.

Further investigation can be done to better adapt to the data imbalance and explore other ways of imputing missing values for categorical variables.

- *The Structural Impact of Food Insecurity on Diabetes Management (Method #4)*: Despite the model's strengths, there are a few notable limitations. Many variables in the full model were not statistically significant, and the stepwise-selected model explained a relatively small proportion of variance in diabetes case counts, suggesting that other unobserved or omitted variables may play a more substantial role.
    - The data was also not ideal for Poisson regression since it was not count based and it had to be manipulated to be suitable.

References

Banerjee, S., Radak, T., & Dunn, P. (2020). Food insecurity and mortality in American adults: Results from the NHANES-linked mortality study. Journal of Nutrition Education and Behavior, 22(2), 174-181. https://doi.org/10.1177/1524839920945927

Berkowitz, S. A., Andrew J. Karter, Giselle Corbie-Smith, Hilary K. Seligman, Sarah A. Ackroyd, Lily S. Barnard, Steven J. Atlas, Deborah J. Wexler; Food Insecurity, Food "Deserts," and Glycemic Control in Patients With Diabetes: A Longitudinal Analysis. *Diabetes Care* 1 June 2018; 41 (6): 1188–1195. https://doi.org/10.2337/dc17-1981

Brown, A. G. M., Esposito, L. E., Fisher, R. A., Nicastro, H. L., Tabor, D. C., & Walker, J. R. (2019). Food insecurity and obesity: Research gaps, opportunities, and challenges. *Translational Behavioral Medicine, 9*(5), 980–987. https://doi.org/10.1093/tbm/ibz117

Byker Shanks, C., Andress, L., Hardison-Moody, A., Jilcott Pitts, S., Patton-Lopez, M., Prewitt, T. E., Dupuis, V., Wong, K., Kirk-Epstein, M., Engelhard, E., Hake, M., Osborne, I., Hoff, C., & Haynes-Maslow, L. (2022). Food Insecurity in the Rural United States: An Examination of Struggles and Coping Mechanisms to Feed a Family among Households with a Low-Income. *Nutrients*, *14*(24), 5250. https://doi.org/10.3390/nu14245250

Cantor, J., Beckman, R., Collins, R. L., Dastidar, M. G., Richardson, A. S., & Dubowitz, T. (2020). SNAP Participants Improved Food Security And Diet After A Full-Service Supermarket Opened In An Urban Food Desert. *Health affairs (Project Hope)*, *39*(8), 1386–1394. https://doi.org/10.1377/hlthaff.2019.01309

Casagrande, S. S., Bullard, K. M., Siegel, K. R., & Lawrence, J. M. (2022). Food insecurity, diet quality, and suboptimal diabetes management among US adults with diabetes. BMJ Open Diabetes Research & Care, 10(5), e003033. https://doi.org/10.1136/bmjdrc-2022-003033

Centers for Disease Control and Prevention. (n.d.). 2017-March 2020 pre-pandemic questionnaire data - continuous NHANES. Retrieved March 28, 2025, from https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&Cycle=2017-2020

Centers for Disease Control and Prevention. (n.d.). NHANES 2017-March 2020 pre-pandemic questionnaire variable list. Retrieved March 28, 2025, from https://wwwn.cdc.gov/nchs/nhanes/search/variablelist.aspx?Component=Questionnaire&Cycle=2017-2020

Centers for Disease Control and Prevention. (n.d.). BRFSS: Graph of current prevalence of diabetes. Retrieved March 10, 2025, from https://data.cdc.gov/Behavioral-Risk-Factors/BRFSS-Graph-of-Current-Prevalence-of-Diabetes

Centers for Disease Control and Prevention. (2025). *Behavioral Risk Factor Surveillance System (BRFSS) public use data*. U.S. Department of Health & Human Services.

https://data.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factor-Surveillance-System-BRFSS-P/dttw-5yxu

Centers for Disease Control and Prevention. (2025). NHANES NNYFS: Dietary supplement use 24-hour - Individual dietary supplements data documentation, codebook, and frequencies. www.cdc.gov/Nchs/Nnyfs/Y_DS1IDS.htm

Clapp, J., Moseley, W. G., Burlingame, B., & Termine, P. (2021). The case for a six-dimensional food security framework. Food Policy, 102, 102164.

Dixon, L. B., Winkleby, M. A., & Radimer, K. L. (2001). Dietary intakes and serum nutrients differ between adults from food-insufficient and food-sufficient families: Third National Health and Nutrition Examination Survey, 1988–1994. The Journal of Nutrition, 131(4), 1232-1246. https://doi.org/10.1093/jn/131.4.1232

Economic Research Service. (n.d.). Food access research atlas - Download the data. U.S. Department of Agriculture. Retrieved March 10, 2025, from https://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data

Gucciardi, E., Vahabi, M., Norris, N. *et al.* The Intersection between Food Insecurity and Diabetes: A Review. *Curr Nutr Rep* 3, 324–332 (2014). https://doi.org/10.1007/s13668-014-0104-4

Kristine D. Gu et al., Nutrition & Diabetes. (2025). *Association of food insecurity with changes in diet quality, weight, and glycemia over two years in adults with prediabetes and type 2 diabetes on Medicaid. https://www.nature.com/articles/s41387-024-00273-7*

Hill-Briggs, F., Adler, N. E., Berkowitz, S. A., Chin, M. H., Gary-Webb, T. L., Navas-Acien, A., Thornton, P. L., & Haire-Joshu, D. (2020). Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes care*, *44*(1), 258–279. Advance online publication. https://doi.org/10.2337/dci20-0053

Lee, C. Y., Zhao, X., Reesor-Oyer, L., Cepni, A. B., & Hernandez, D. C. (2020). Bidirectional relationship between food insecurity and housing instability. Journal of the Academy of Nutrition and Dietetics, 120(11), 1837–1845. https://doi.org/10.1016/j.jand.2020.08.081

Leung, C. W., & Wolfson, J. A. (2021). Food insecurity among older adults: 10-year national trends and associations with diet quality. *Journal of the American Geriatrics Society, 69*(4), 964–971. https://doi.org/10.1111/jgs.16971

Levi, R., Bleich, S. N., & Seligman, H. K. (2023). Food Insecurity and Diabetes: Overview of intersections and potential dual solutions. Diabetes Care, 46(9), 1599–1608. https://doi.org/10.2337/dci23-0002

Ley, S. H., Hamdy, O., Mohan, V., & Hu, F. B. (2014). Prevention and management of type 2 diabetes: Dietary components and nutritional strategies. The Lancet, 383(9933), 1999–2007. https://doi.org/10.1016/S0140-6736(14)60613-9

Liese, A. D., Lamichhane, A. P., Garzia, S. C. A., Puett, R. C., Porter, D. E., Dabelea, D., D'Agostino, R. B., Standiford, D., & Liu, L. (2018). Neighborhood characteristics, food deserts, rurality, and type 2 diabetes in youth: Findings from a case-control study. *Health & Place, 50*, 81–88. https://doi.org/10.1016/j.healthplace.2018.01.004

Seligman, H. K., Bindman, A. B., Vittinghoff, E., Kanaya, A. M., & Kushel, M. B. (2007). Food insecurity is associated with diabetes mellitus: results from the National Health Examination and Nutrition Examination Survey (NHANES) 1999-2002. *Journal of general internal medicine*, *22*(7), 1018–1023.

Seligman, H. K., Elizabeth A. Jacobs, Andrea López, Jeanne Tschann, Alicia Fernandez; Food Insecurity and Glycemic Control Among Low-Income Patients With Type 2 Diabetes. *Diabetes Care* 1 February 2012; 35 (2): 233–238. https://doi.org/10.2337/dc11-1627

Shaheen, M., W. Kibe, L., & M. Schrode, K. (2021). Dietary quality, food security and glycemic control among adults with diabetes. Clinical Nutrition ESPEN, 46, P336-342. https://www.clinicalnutritionespen.com/article/S2405-4577(21)01063-9/fulltext

U.S. Chronic Disease Indicators. (2024, February 13). Data.gov; Centers for Disease Control and Prevention. https://catalog.data.gov/dataset/u-s-chronic-disease-indicators
Walker, R. J., Grusnick, J., Garacci, E., Mendez, C., & Egede, L. E. (2018). Trends in Food Insecurity in the USA for Individuals with Prediabetes, Undiagnosed Diabetes, and Diagnosed Diabetes. Journal of General Internal Medicine, 34(1), 33–35. https://doi.org/10.1007/s11606-018-4651-z

Appendix

GitHub: https://github.gatech.edu/ISyE6414-SP2025/runic_ogres/tree/main

All members contributed equally to all aspects of the project. Many of our individual portions have our names labelled as subsections throughout our analysis. Our code is collected on GitHub, in the repository above. On GitHub, the *data* folder contains all the data files and code used to create our final, merged data set: *final_data.csv.* We each used the final_data.csv file for our individual analysis.

Our individual project contributions are further elaborated on below:

Christina: I created the *Food Security, Meal Patterns & Diabetes Logistic Regression Model* (Methods & Results #2). I also sent out meeting invites and provided meeting summaries for important group meetings. My individual code is uploaded to my folder on GitHub. My code contains detailed comments, visuals, and interpretations on my individual analysis. These findings are also found throughout the report. I contributed to the Summary Section, including the revised research question, data section, and goals & major findings subsections. In the literature section, I wrote an analysis under the subsection *Food Insecurity – Community Challenges and Nutritional Intervention.* In the Conclusion section, I explained my model's successes and limitations. I maintained the references section. I also contributed to the Explanation of Changes, under the research question and data subsections, and the Appendix preface.

Ju Ho: My analysis used decision tree models to analyze demographics, dietary behaviour, and health indicators in relation to diabetes. Twelve predictor variables with less than 50% missing values were selected to be included in the initial model. Some of the chosen variables had missing responses and out of bounds responses such as 9999 when a respondent declined a response. These issues were resolved with mode imputation for categorical variables and out of bounds responses were omitted from the final dataset. An initial model was trained with all variables included. A second model was trained to attempt to correct the imbalanced data, which performed better to predict diabetes cases. As a group member, I identified data sets for use in our analysis, contributed equally to our literature review, individual analysis, introduction and successes and limitations of the project.

Naomi: I created the *Structural Impact of Food Insecurity on Diabetes Management Poisson Regression Model* (Methods & Results #4). I also combined the P_DBQ.xpt, P_DEMO.xpt, P_DIQ.xpt, P_FSQ.xpt, and P_INQ.xpt files from the CDC NHANES Survey website into the final_dataset.csv file that we all used in our analysis. This code can be found in the data folder on our GitHub in the cdc_data_merging.ipynb file. My individual code is uploaded to my folder on GitHub, and this contains detailed comments and interpretations on my individual analysis. These findings are also found throughout the report. I also contributed to the literature review section and analysis section under the subsection *Structural Impact of Food Insecurity on*

Diabetes Management. I added notes on my model's successes and limitations in the Conclusion section. I added to the Explanation of Changes to explain my slight change in methodology from what I originally proposed in the Initial Analysis Plan.

Pey-Tzer: I created the Logistic Model (Method & Results #1) to analyze the correlation between food insecurity and diabetes prevalence using demographic, dietary behaviors, and socioeconomic factors. I created the initial research question and contributed to the Problem Description subsection in the Summary. My individual code is stored in my folder on GitHub, and this contains visualizations of my findings. I originally intended to create a multiple linear regression model and later chose to create a logistic regression model. I explained these changes under the Explanation of Changes section, under Logistic Regression Plan (Method #1). I contributed to the Literature Section, under the subsection *Correlation between Food Security and the Prevalence of Diabetes*. I wrote the Conclusion section, including the Successes, Limitations/Further Research, to summarize our project's findings.