Analysing Factors Contributing to Type II Diabetes Risk Status

Sean Mulherin and Ramiro Lobo

STATS 411 Multivariate Statistical Analysis
Dr. Maria Cha
Winter 2025

UCLA College | Physical Sciences Statistics & Data Science

Introduction

Background

- Diabetes is a chronic disease that describes the body's inability to properly regulate insulin levels
- Our bodies use insulin to break down glucose obtained from the foods/drinks we consume
- There are two main types
  - Type I: not preventable - typically inherited genetically or triggered via viral infection
  - Type II: preventable - typically a result of lifestyle habits

Introduction

Motivation

- Abnormal glucose levels can lead to many health issues, including:
  - Heart Disease
  - Kidney Disease
  - Cataracts
  - Nerve Damage
- As of 2021, 38 million people have diabetes - 11.6% of population (American Diabetes Association)
- Diabetes is the 8th leading cause of death in the U.S.
- Total costs span $412.9 billion (Emily et al., 2022)

Introduction

Objective

- The goal is to help prevent the onset of type II diabetes.

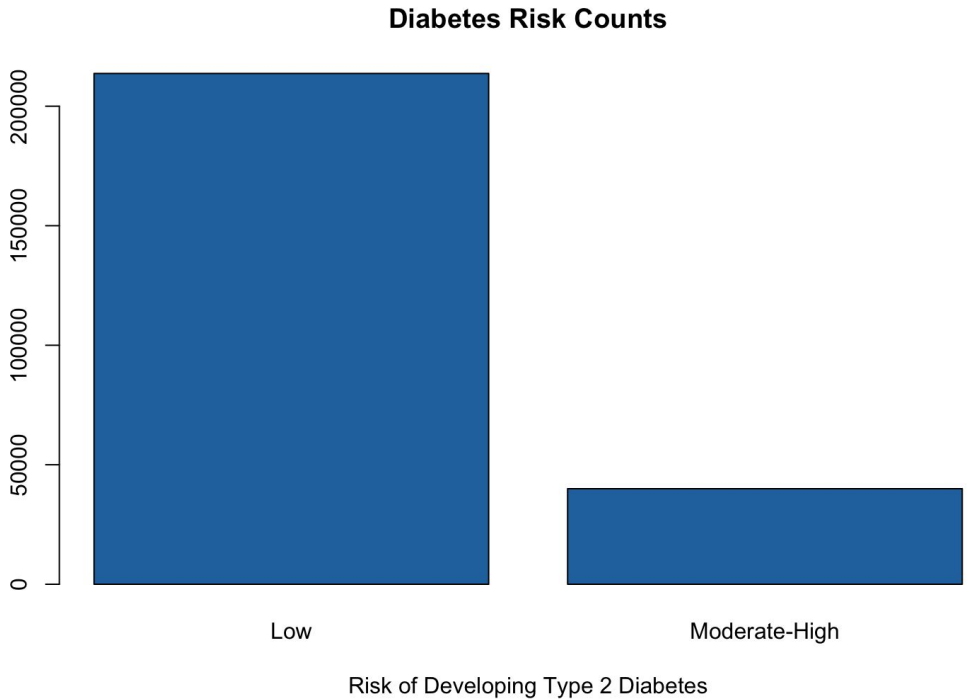
- To accomplish this, we leverage rigorous statistical analysis to identify factors that affect the development of type II diabetes, including:
  - Logistic Regression Analysis
  - Principal Component Analysis
  - Factor Analysis

Data

- Every year, the Center for Disease Control (CDC) conducts a behavioral survey of Americans
  - Behavioral Risk Factor Support Survey (BRFSS) collects health-related information via telephone surveys of over 400,000 Americans from each state
- Predictors
  - 21 variables of class numeric, ordinal, and categorical
  - **Biological factors**: blood pressure, cholesterol, BMI
  - **Lifestyle factors**: age, sex, smoker, diet, exercise, alcohol consumption
  - **Social factors**: income, education, mental health
- Response
  - Binary: low or medium-high risk of developing type II diabetes
- 253,680 total observations

Data

EDA


Diabetes Risk Counts
0
50000
100000
150000
200000
Low
Moderate-High
Risk of Developing Type 2 Diabetes

Data

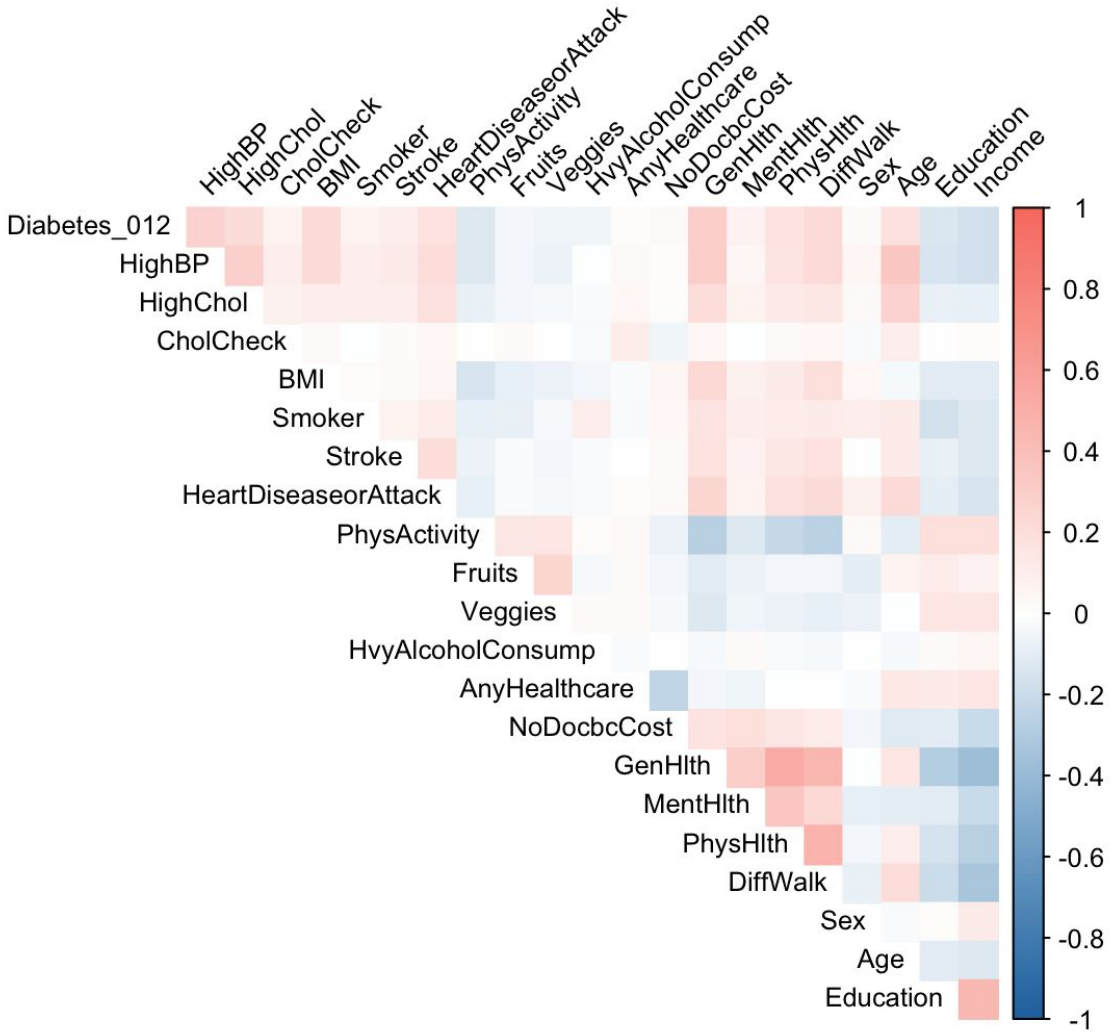
EDA

Highest Positive Correlations with Diabetes Risk Status (> 0.2)

- GenHlth
- HighBP
- BMI
- DiffWalk
- HighChol

Highest Negative Correlations with Diabetes Risk Status (< -0.12)

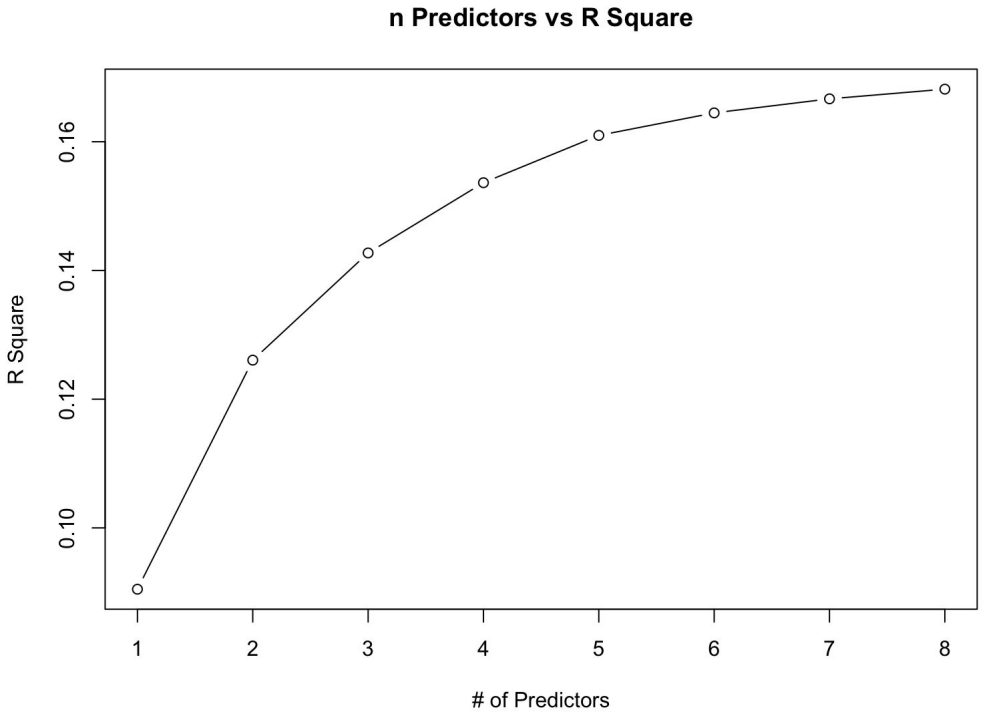
- Income
- Education
- PhysActivity



Methodology - Regression

Notable Factor Coefficients

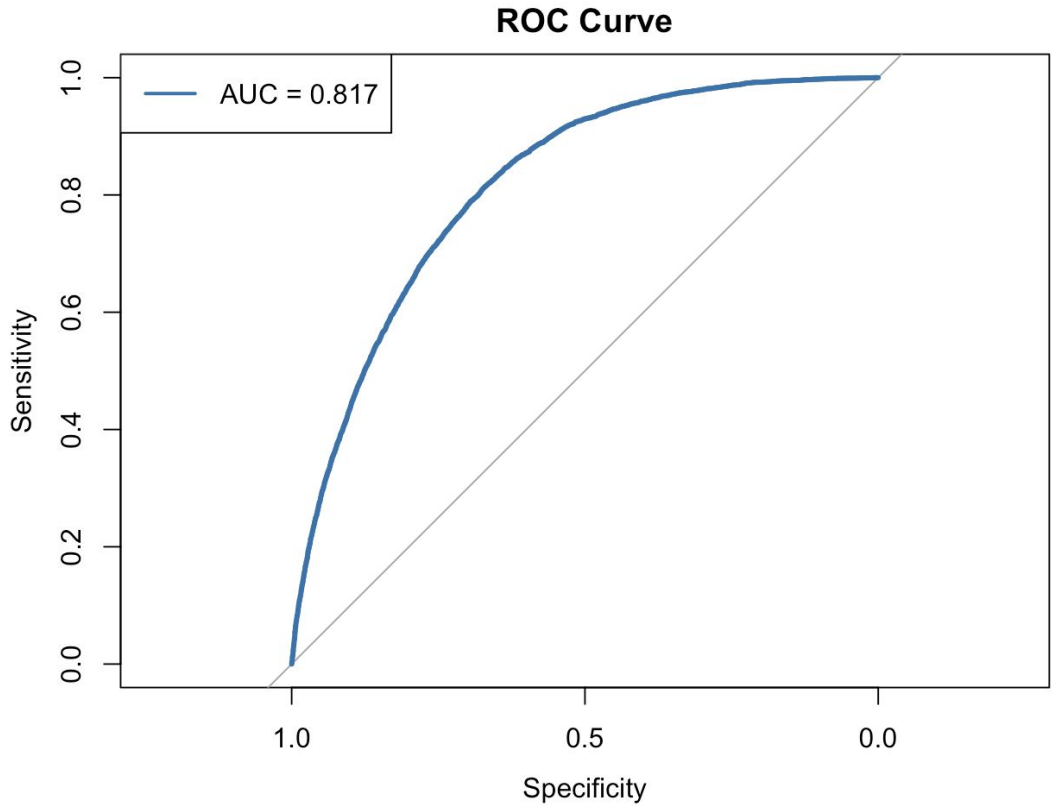
- CholCheck = 1.21
- HighBP = 0.71
- HvyAlcoholConsump = 0.66
- HighChol = 0.60
- GenHlth = 0.51
- Sex = 0.24



Methodology - Regression Cont.

Utilized the Youden Index to find the optimal cut-off threshold. (Ruopp et. al, 2008)

Optimal Threshold = 0.133



Methodology - Regression Cont.

We want to confidently capture medium-high risk patients, so we prioritize Recall

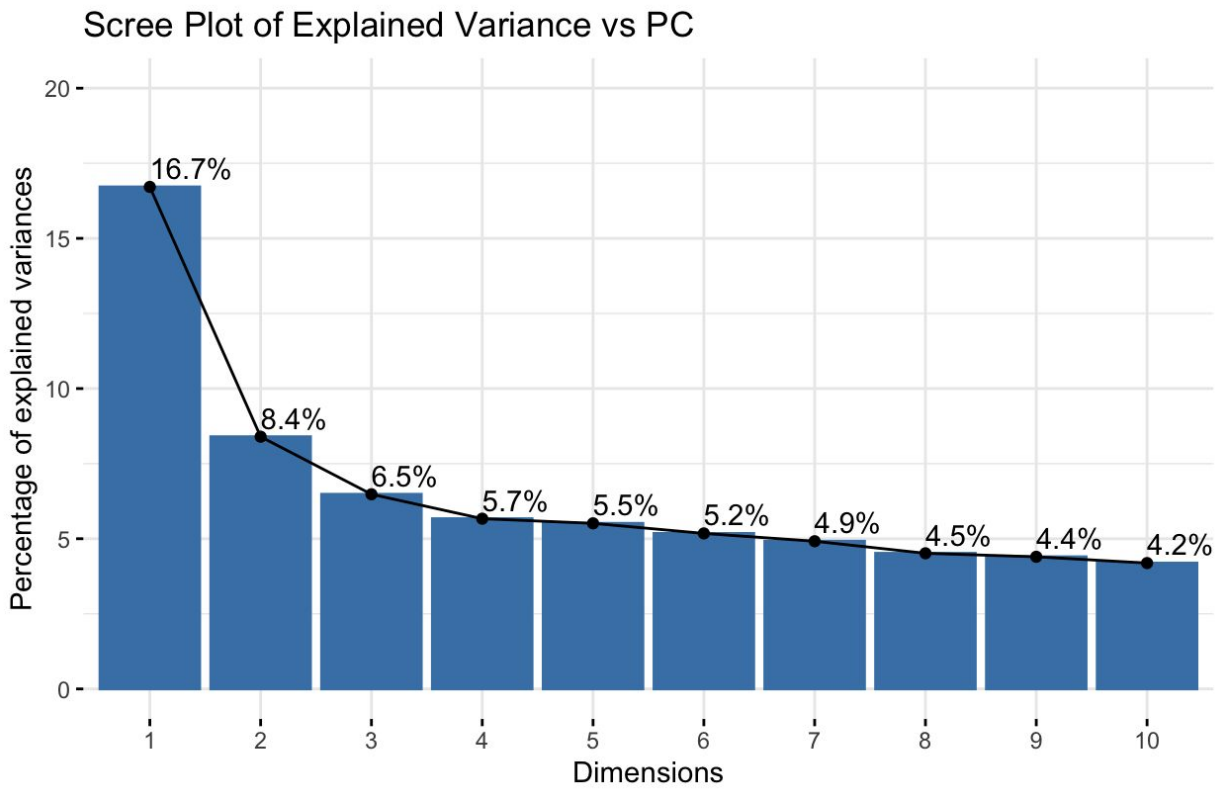
- Accuracy = 69.5%
- Precision = 31.8%
- Recall = 81.2%
- F1 Score = 45.7%

		Observed	
		Low Risk	**Medium-High Risk**
Predicted	**Low Risk**	28745	1508
	Medium-High Risk	13969	6514

Methodology - PCA

Top 10 PC's explain 66% of variance

Top 14 PC's explain 80% of variance

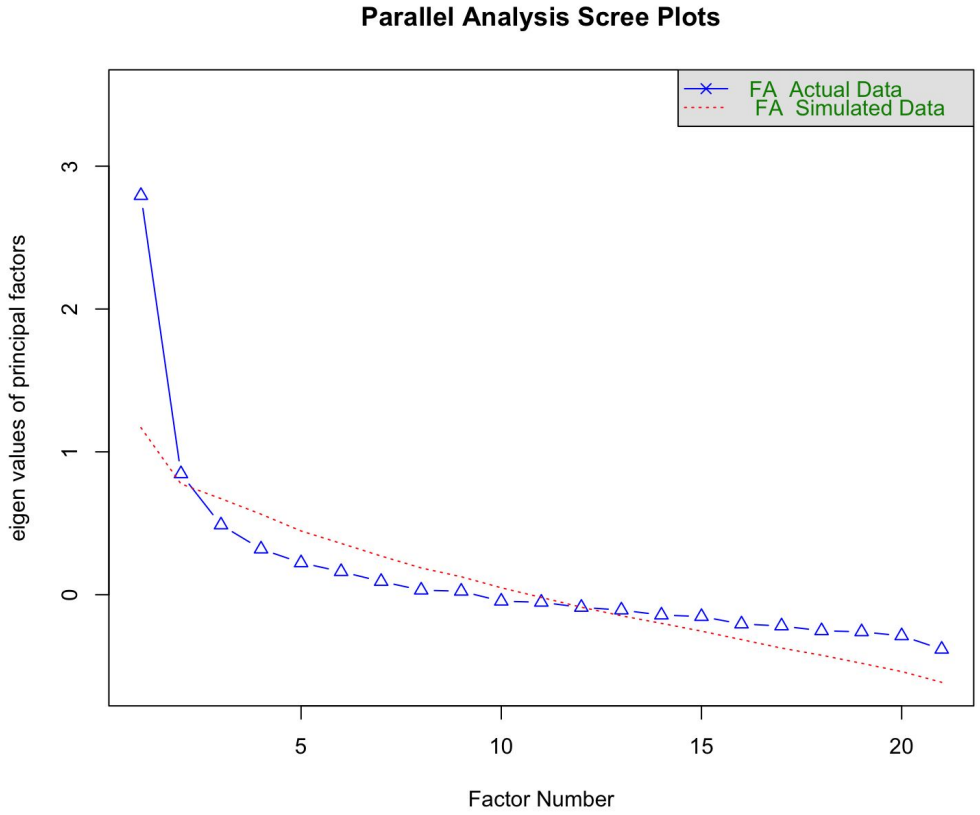


Methodology - Factor Analysis

Two factors are identified as the ideal amount to explain the variance.

Factor 1 is likely to represent the general health of a person.

Factor 2 is likely to represent biological factors, predominantly cholesterol health.



Conclusion

- Biological Factors:
  - High Cholesterol has the strongest adverse effect
  - High Blood Pressure has the second strongest adverse effect
  - Males are more likely to develop type II diabetes than females
  - As Age increases, the risk increases

- Lifestyle Factors:
  - Heavy Alcohol Consumption has the strongest adverse effect
    - Males > 14 drinks/week
    - Females > 7 drinks/week
  - Marginal negative effects:
    - Income, Education, Physical Activity, Diet, and Smoking

Conclusion

American Diabetes Association. (Nov, 2023). *Diabetes facts and statistics*. Diabetes.org. Retrieved March 1, 2025, from https://diabetes.org/about-diabetes/statistics/about-diabetes

Emily D. Parker, Janice Lin, Troy Mahoney, Nwanneamaka Ume, Grace Yang, Robert A. Gabbay, Nuha A. ElSayed, Raveendhara R. Bannuru; Economic Costs of Diabetes in the U.S. in 2022. Diabetes Care 2 January 2024; 47 (1): 26–43. https://doi.org/10.2337/dci23-0085

Centers for Disease Control and Prevention. (2025, February). 2023 BRFSS survey data and documentation. https://www.cdc.gov/brfss/annual_data/annual_2023.html

Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. Biom J. 2008 Jun;50(3):419-30. doi: 10.1002/bimj.200710415. PMID: 18435502; PMCID: PMC2515362.

Analyzing Factors Contributing to Type II Diabetes Risk Status

Project Report by
Sean Mulherin, Ramiro Lobo

March 14, 2025

Stats 411 Final Project

University of California Los Angeles
Department of Statistics & Data Science

1 Introduction

Diabetes is a chronic disease that describes the body's inability to properly regulate insulin levels. Insulin is used by the human body to break down blood glucose, or blood sugar, obtained from dietary consumption. There are two main types of diabetes. Type I diabetes occurs when a person's pancreas does not produce enough insulin. It usually appears in childhood and is not preventable. Type II diabetes is a metabolic disorder in which the body becomes resistant to or does not produce enough insulin to break down blood sugar. This type of disease is largely preventable and often is related to lifestyle factors such as diet and activity levels. Abnormal glucose levels produced by both types can lead to various negative health outcomes, including heart disease, kidney disease, cataracts, and nerve damage. In 2021, 38 million people in the United States have been diagnosed with diabetes, approximately 11.6% of the total population [1]. It represents the eighth leading cause of death in the United States, costing Americans approximately $412 billion in 2022 [2].

In this report, we employ rigorous statistical analysis with the objective of identifying factors that impact the development of Type II diabetes and strengthening our understanding of its pathology. In particular, we employ the following methods: Logistic Regression Analysis, Principal Component Analysis, and Factor Analysis. These multivariate statistical methods allow us to account for a large number of variables in our analysis of potential indicators for the development of Type II diabetes.

2 Data

The data comes from the United States Center for Disease Control's (CDC) Behavioral Risk Factor Support Survey (BRFSS). The BRFSS is an ongoing system of telephone surveys that collect data on health-related behaviors, health conditions, and health care access among adults from all 50 states and additional territories that choose to participate. This report analyzes the most recent BRFSS data from 2023 which contains 433,323 records in total.

Our analysis focuses on analyzing the factors that contribute to a person being at moderate to high risk of developing Type II diabetes. In our analysis, we utilized 21 factors as predictors, grouped into the following categories and subcategories:

- **Biological:** presence of high blood pressure, presence of high cholesterol, BMI, cholesterol check, stroke, heart disease/attack, general health, physical health
- **Lifestyle**: age, sex, smoking, diet, exercise, alcohol consumption, fruit consumption, vegetable consumption, healthcare, difficulty walking,
- **Social**: income, education, mental health

These factors were used to model a binary response variable indicating whether a respondent self-identified as being at low or moderate-high risk of developing type II diabetes. From the initial sample of over 400,000, there were 253,680 observations containing data for the desired factors and response variable.

2.1 Exploratory Data Analysis

From the 253,680 individuals in the data, 39,977 (15.7%) were labeled as being at moderate-high risk of contracting type II diabetes, compared to 213,703 (84.2%) individuals who were labeled as low-risk (see Figure 1).

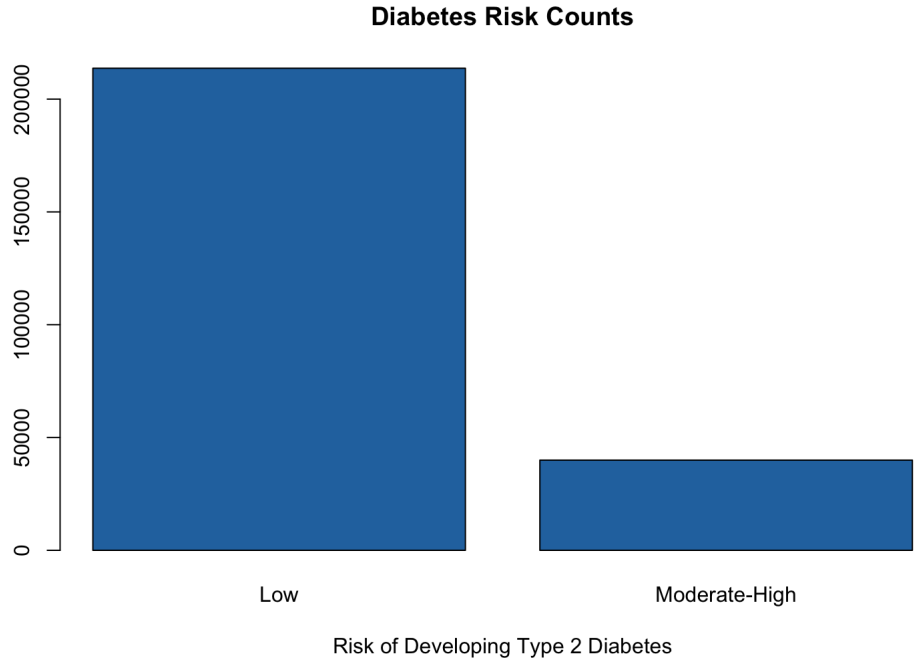


Figure 1: Imbalance in Type II diabetes case counts, with 15.7% of observations falling in the positive, moderate-high risk category.

In Figure 4, we see that there are five variables that have a positive correlation with diabetes risk greater than 0.2: General Health (where a higher value indicates a lower reported health score), diagnosis of high blood pressure, BMI, difficulty walking, and diagnosis of high cholesterol. In contrast, there were three factors with a negative correlation less than -0.12: income, education, and physical activity (see Figure 4 in the Appendix).

3 Analysis & Results

In our study, three statistical methods were used to model the factors associated with being at risk of contracting Type II diabetes: Logistic Regression, Principal Component Analysis, and Factor Analysis. Below we represent the results from each method on our dataset.

3.1 Logistic Regression

The full results of the logistic regression model can be seen in Table 2 in the appendix section. Of the 21 predictors used, only three were not significant at the 1% level, whether the respondent identified as a smoker, vegetable consumption, and mental health. The remaining 18 factors were statistically significant in our regression model.

The factor with the highest coefficient was an indicator of when the survey respondents last checked their cholesterol (namely, *CholCheck*), where a higher value indicates that more time has passed since their last cholesterol examination. Each additional year since their last cholesterol examination was associated with approximately a threefold increase in the odds of being classified as moderate-high risk of developing Type

II diabetes. Additionally, the presence of high blood pressure nearly doubles the risk while high cholesterol increases the odds by a factor of 1.8. Many lifestyle factors, such as fruit/vegetable consumption and physical activity, were associated with a slight decrease in the odds of being at moderate-high risk. Moreover, heavy alcohol consumption is shown to have a strong detrimental effect on risk factor, almost doubling the risk of developing Type II diabetes.

To accurately identify patients who are at moderate-high risk, we prioritized the recall of our model. Using the Youden Index, we found the optimal cut-off threshold to be 0.133 [3]. This threshold resulted in an overall accuracy of 69.5% and a recall of 81.2%, ensuring that our model accurately captures at-risk individuals. This also resulted in a lower precision of 31.8% and an overall F1 score of 45.7%.

3.2 Principal Component Analysis

In addition to using logistic regression, we also conducted Principal Component Analysis (PCA) with a varimax rotation for improved interpretability. Since the predictor variables were measured on different scales, we conducted our analysis using the correlation matrix. After conducting our analysis, we found that the first 10 principal components explained about 66% of the variance in our data and the first 14 principal components explained about 80% of the overall variance (see Figure 2). Considering the PC loadings, as shown in Table 1, the first principal component weighs the following features most heavily: GenHlth, MentHlth, PhysHlth, DiffWalk. Hence, the first principal component seems to explain the overall health of an individual. The second principal component weighs the following features the most heavily: HighBP, HighChol, Age. It seems this component is predominantly tracking an individual's biological health. In addition, the third main component heavily weighs the consumption of fruits and vegetables, tracking the diet of an individual. Lastly, the fourth principal component places the highest weight on Education and Income essentially taking into account the socioeconomic status of an individual.

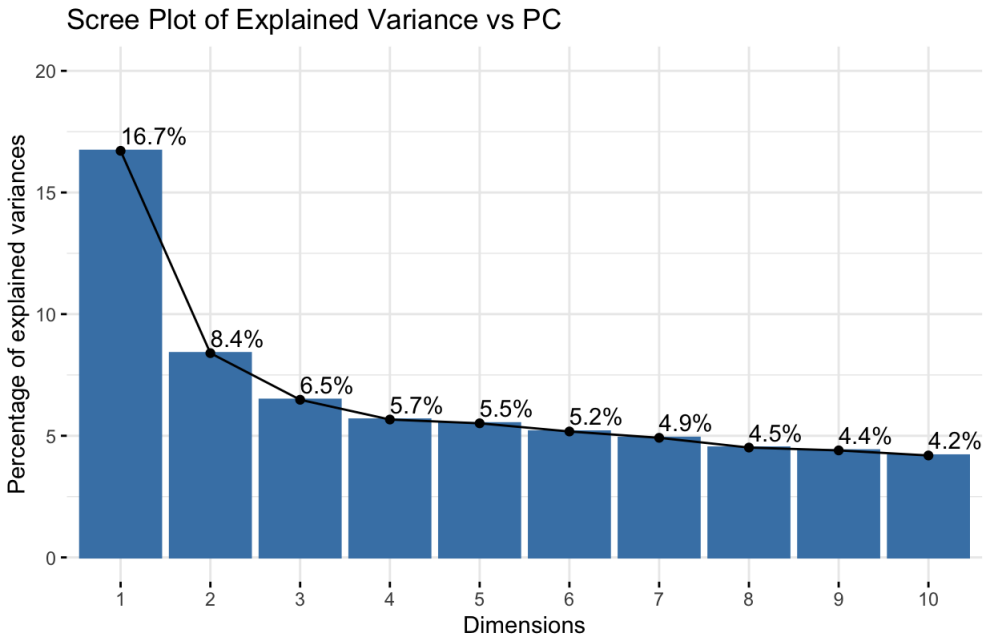


Figure 2: Proportion of variance explained by each principal component

Feature	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
HighBP	-0.09	-0.67	-0.01	0.12	-0.04	-0.31	0.12	0.06	0.00	0.04
HighChol	-0.12	-0.78	-0.11	-0.10	-0.02	0.00	-0.04	0.06	-0.10	-0.00
CholCheck	-0.01	-0.07	0.01	0.04	0.03	-0.04	0.03	0.95	0.12	-0.01
BMI	-0.14	-0.08	-0.02	0.04	-0.06	-0.87	-0.04	0.05	-0.06	-0.01
Smoker	-0.21	-0.10	0.05	0.35	-0.57	0.21	-0.10	-0.00	0.06	0.33
Stroke	-0.10	-0.01	-0.04	0.04	0.07	-0.01	0.88	0.03	-0.02	0.07
HeartDiseaseorAttack	-0.20	-0.30	0.02	0.05	-0.27	0.06	0.52	-0.00	-0.01	-0.10
PhysActivity	0.28	0.03	0.27	-0.32	-0.11	0.30	-0.02	0.21	-0.19	-0.07
Fruits	0.05	0.00	0.76	-0.01	0.12	0.07	0.03	0.01	0.03	-0.11
Veggies	0.02	0.04	0.77	-0.15	-0.02	-0.03	-0.06	-0.00	-0.01	0.10
HvyAlcoholConsump	-0.04	-0.00	0.02	0.06	-0.02	0.01	-0.03	0.01	0.03	-0.93
AnyHealthcare	-0.12	-0.07	-0.02	-0.23	0.05	0.07	-0.03	0.16	0.70	-0.01
NoDocbcCost	-0.24	-0.02	-0.03	0.04	0.06	0.03	0.00	0.02	-0.73	0.02
GenHlth	-0.65	-0.22	-0.07	0.27	-0.08	-0.21	0.15	0.02	-0.06	-0.04
MentHlth	-0.69	0.06	-0.10	-0.07	0.05	0.10	-0.09	0.07	-0.24	0.08
PhysHlth	-0.80	-0.05	0.01	0.09	-0.02	-0.04	0.11	-0.01	0.01	-0.04
DiffWalk	-0.63	-0.15	0.03	0.23	0.05	-0.18	0.18	-0.05	0.10	-0.04
Sex	0.12	0.01	-0.13	-0.16	-0.81	-0.15	0.09	-0.02	-0.02	-0.11
Age	-0.00	-0.66	0.11	0.24	0.03	0.16	0.16	-0.05	0.33	-0.04
Education	0.07	0.03	0.11	-0.80	0.06	0.07	-0.01	-0.05	0.09	0.02
Income	0.27	0.10	0.07	-0.69	-0.13	-0.01	-0.09	-0.01	0.22	0.07

Table 1: Loadings of Principal Components

3.3 Factor Analysis

Finally, we conducted a factor analysis on our dataset. Similarly to our principal component analysis, we use the correlation matrix due to the different measurement scales of our variables. In order to uncover any latent factors contributing to an increased risk of disease contraction we use maximum likelihood estimation (MLE) to estimate the factor loadings rather than the principal components. Using the `fa.parallel()` function in `R` we find that a model with two factors explains the most variance in our analysis (see Figure 3).

As a result, our final factor analysis model uses two factors with a varimax rotation. The full factor loadings can be seen in Table 3 in the Appendix. The first factor can be interpreted as an individual's overall reported health and socioeconomic status factor. In the first factor, the variables with loads greater than 0.4 are General Health, Physical Health, Difficulty Walking, and Mental Health. This means that higher values, or worse self-reported health, are correlated with having a moderate to high risk of Type II diabetes. The variables with loadings less than -0.3 are Income, Education, and Physical Activity. These loadings indicate that being at a higher socioeconomic level, indicated by higher income and education, as well as higher levels of physical activity are negatively correlated with being at moderate-high risk of Type II disease.

The second factor can be interpreted as a biological measurement factor or cardiovascular health factor. The highest positive loadings on this factor, those greater than 0.4, include: age, high blood pressure, and high cholesterol. This suggests that an individual's age, high blood pressure, and high cholesterol are correlated with a greater risk of developing Type II diabetes. The variables with negative loadings are similar to those of the first factor, but are not as significant as those of the first factor. Individuals with lower income and education levels tend to report worse health outcomes, which aligns with known socioeconomic disparities in diabetes prevalence.

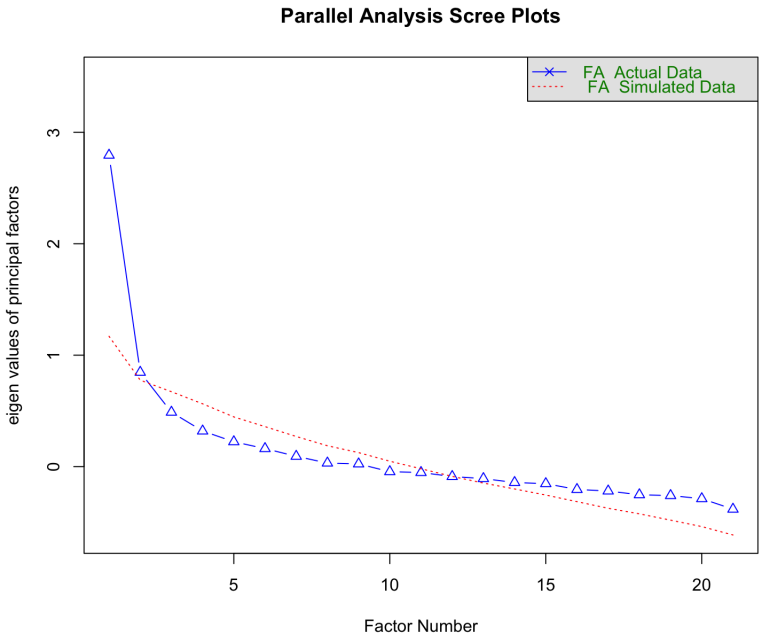


Figure 3: Scree Plot of Factors

4 Conclusion

This study highlights key factors contributing to Type II diabetes risk, emphasizing the role of cardiovascular health indicators such as high blood pressure and high cholesterol. The logistic regression results underscore the importance of frequent cholesterol checks, suggesting that proactive screening can help identify at-risk individuals early. Factor analysis further supports the interplay between socioeconomic status and health outcomes, reinforcing the notion that lower income and education levels correlate with a higher likelihood of developing Type II diabetes.

While these findings offer valuable insights, several limitations must be acknowledged. The reliance on self-reported data introduces potential biases, as individuals may underreport unhealthy behaviors. Future research could explore longitudinal studies to examine how lifestyle interventions impact diabetes risk over time.

From a policy perspective, these findings advocate for increased accessibility to preventive healthcare services, particularly among lower-income populations. Public health initiatives should focus on educating individuals about the importance of regular cholesterol screening and promoting lifestyle modifications such as increased physical activity and improved diet to improve overall cardiovascular health. Addressing these risk factors through targeted interventions can significantly reduce the incidence of Type II diabetes and improve overall public health outcomes.

References

[1] American Diabetes Association. Diabetes facts and statistics. `https://diabetes.org/about-diabetes/statistics/about-diabetes`, nov 2023. Retrieved March 1, 2025.

[2] Emily D. Parker, Janice Lin, Troy Mahoney, Nwanneamaka Ume, Grace Yang, Robert A. Gabbay, Nuha A. ElSayed, and Raveendhara R. Bannuru. Economic costs of diabetes in the u.s. in 2022. *Diabetes Care*, 47(1):26–43, 11 2023.

[3] Whitcomb BW Schisterman EF Ruopp MD, Perkins NJ. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom J*, 2008.

Appendix

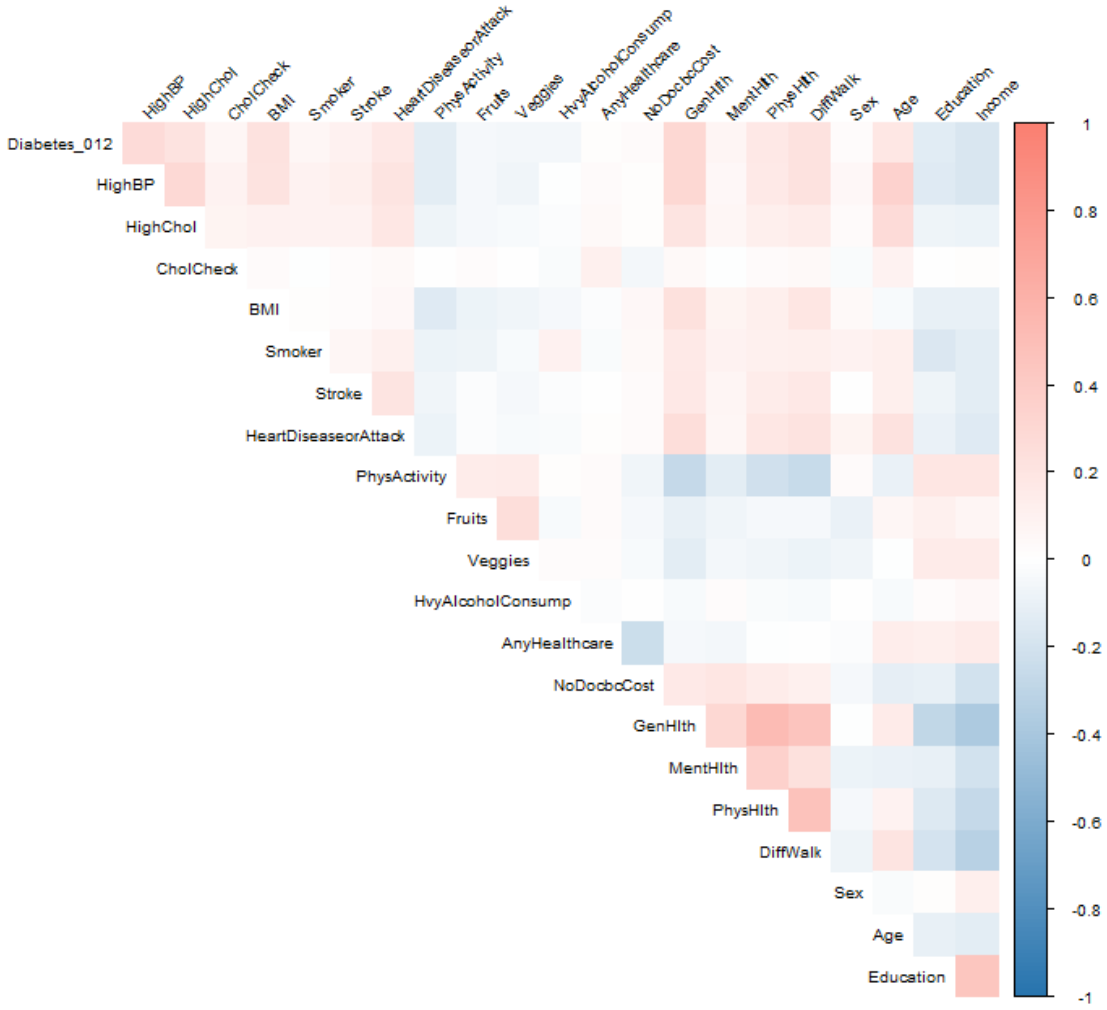


Figure 4: This figure contains a correlation plot between diabetes risk and all related predictors.

Table 2: Logistic Regression Coefficients for Type II Diabetes Risk Model

Predictor	**Estimate**	**Std. Error**	**z value**	**p-value**	**Sig.**
(Intercept)	-5.69009	0.08700	-65.407	$< 2e-16$	***
HighBP	0.77568	0.01534	50.578	$< 2e-16$	***
HighChol	0.60731	0.01433	42.390	$< 2e-16$	***
CholCheck	1.22626	0.06888	17.803	$< 2e-16$	***
BMI	0.80103	0.01449	55.263	$< 2e-16$	***
Smoker	-0.01517	0.01399	-1.084	0.278394	
Stroke	0.16715	0.02703	6.184	$< 6.25e-10$	***
HeartDiseaseorAttack	0.29060	0.01911	15.208	$< 2e-16$	***
PhysActivity	-0.10689	0.01519	-7.037	$< 1.96e-12$	***
Fruits	-0.03497	0.01446	-2.419	0.015566	*
Veggies	-0.04060	0.01684	-2.411	0.015918	*
HvyAlcoholConsump	0.70614	0.03892	18.144	$< 2e-16$	***
AnyHealthcare	0.08510	0.03448	2.468	0.013580	*
NoDocbcCost	0.09029	0.02384	3.788	0.000152	***
GenHlth	0.91850	0.01642	55.930	$< 2e-16$	***
MentHlth	-0.03640	0.01561	-2.332	0.019689	*
PhysHlth	0.11891	0.01532	7.764	$< 8.21e-15$	***
DiffWalk	0.32610	0.01700	19.180	$< 2e-16$	***
Sex	0.18486	0.01422	12.999	$< 2e-16$	***
Age	0.51072	0.01540	33.158	$< 2e-16$	***
Education	-0.14573	0.01569	-9.288	$< 2e-16$	***
Income	-0.27590	0.01763	-15.649	$< 2e-16$	***

Table 3: Factor Loadings from Factor Analysis

Variable	**Factor 1**	**Factor 2**
HighBP	0.48	0.29
HighChol	0.35	0.24
CholCheck	0.08	0.16
BMI	0.27	-0.04
Smoker	0.24	0.01
Stroke	0.26	0.07
HeartDiseaseorAttack	0.37	0.17
PhysActivity	-0.36	0.11
Fruits	-0.15	0.13
Veggies	-0.19	0.09
HvyAlcoholConsump	0.04	0.03
AnyHealthcare	-0.04	0.27
NoDocbcCost	0.18	-0.34
GenHlth	0.60	-0.13
MentHlth	0.20	-0.35
PhysHlth	0.42	-0.22
DiffWalk	0.57	-0.08
Sex	-0.02	0.10
Age	0.32	0.51
Education	-0.35	0.13
Income	-0.42	0.09