Data Exploration & Myth Identification – Diabetes BRFSS Survey (2024)

Rishab Hanamar
M.Sc. Software Engineering
University of Europe for Applied Sciences
Konrad-Zuse-Ring 11, 14469 Potsdam, Germany
rishab.hanamar@ue-germany.de

Prof. Raja Hashim Ali
University of Europe for Applied Sciences
Konrad-Zuse-Ring 11, 14469 Potsdam, Germany
hashim.ali@ue-germany.de

***Abstract*—This study investigates behavioral and demographic factors associated with diabetes using the latest Behavioral Risk Factor Surveillance System (BRFSS) dataset. The BRFSS, administered by the CDC, provides nationally representative health data from over 400,000 U.S. adults. Using Python for data cleaning, wrangling, and exploratory data analysis (EDA), we examined the relationships between diabetes prevalence and factors such as age, BMI, physical activity, smoking status, and weight category. Our results indicate a strong association between higher BMI, older age, inactivity, and increased diabetes risk. Individuals who are obese, physically inactive, or elderly show the highest prevalence of diabetes, whereas younger adults with normal BMI and regular exercise have the lowest prevalence. Furthermore, we explored five common public myths about diabetes and assessed their validity using weighted analyses, providing insights into common misconceptions and actual risk patterns in the population.**

***Index Terms*—Diabetes, Behavioral Risk Factor Surveillance System (BRFSS), Age Group, Body Mass Index (BMI), Physical Activity, Smoking, Weight Category, Exploratory Data Analysis (EDA), Public Health, Risk Factors.**

I. Introduction

Diabetes is a major public health concern that is influenced by behavioral and demographic factors such as age, body mass index (BMI), physical activity, and smoking. Understanding the patterns and prevalence of diabetes is crucial for designing effective prevention strategies and public health interventions.

The Behavioral Risk Factor Surveillance System (BRFSS), administered by the Centers for Disease Control and Prevention (CDC) provides nationally representative health data from over 400,000 U.S. adults each year. This rich dataset allows for the analysis of diabetes prevalence across different population subgroups and the identification of patterns associated with lifestyle and demographic characteristics.

In this study, we use the 2024 BRFSS dataset to examine the relationships between diabetes and factors including age, BMI, exercise activities, smoking status, and weight category. We also investigate common public misconceptions about diabetes, comparing perceived and observed risk patterns. The findings highlight high-risk groups and provide insights that can support targeted public health strategies and educational efforts.

II. Methodology

A. Data Acquisition

The 2024 BRFSS dataset was obtained from the CDC official repository: https://www.cdc.gov/brfss/annual_data/annual_2024.html. Downloaded the combined Landline and Cellular dataset (`LLCP2024.XPT`) in SAS Transport format.

Dataset Summary:

- **Source:** CDC Behavioral Risk Factor Surveillance System (BRFSS)
- **Year:** 2024
- **Format:** SAS XPORT (.XPT)
- **Records:** ∼450,000 U.S. adult respondents
- **Software:** Python 3.10, **pandas**, **numpy**, **matplotlib**, **seaborn**, **statsmodels**, **jupyter**

Variables Used:

TABLE I: Selected Variables (CDC BRFSS 2024)

Variable	Description
DIABETE4	(Ever told) you had diabetes (1=Yes, 2=Yes, but female told only during pregnancy, 3=No, 4=No, pre-diabetes or borderline diabetes, 7=Don't know/Not Sure, 9=Refused, BLANK=Not asked or Missing)
_BMI5	Computed body mass index (1-9999 or BLANK)
EXERANY2	Physical activity (1=Yes, 2=No, 7=Dont know/Not Sure ,9=Refused ,BLANK=Not asked or missing)
SMOKER3	Computed Smoking Status (1–4 categories) and 9=Dont know
_AGEG5YR	5 year age categories calculated variable (13 ranges of 5 years and 14=Dont know/Refused/Missing)

B. Data Cleaning and Wrangling

Cleaning decisions were guided by the official BRFSS codebook and by plausibility checks. The main steps and justifications are:

Steps:

1) Recoded `DIABETE4` → `Diabetes` (1=Yes, 2=Yes, 3=No, 4=No, 7, 9 and BLANK as `NaN`).
2) Converted `_BMI5` to numeric and derived weight categories.
3) Recoded `EXERANY2` → `Exercise_Activites` (1=Yes, 2=No, 3 and 4 as `NaN`).

4) Recoded `_SMOKER3` → `Smoking_Condition` (Everyday, Sometime, Former, Never).
5) Created grouped variable `age_group`.

III. Exploratory Data Analysis (EDA)

The cleaned dataset was explored using statistical summaries and visualizations.

A. Diabetes vs Age group

1) Demographics: The demographics of respondents and their diabetes status are summarized in Table II.

Age Group	% No Diabetes	% Yes Diabetes
Young adults	97.45	2.55
Early working years	95.14	4.86
Mid career adults	90.95	9.05
Pre-retirement	83.37	16.63
Older working adults	77.47	22.53
Elderly population	76.09	23.91
Early Seniors	74.47	25.53

TABLE II: Diabetes Prevalence by Age Group

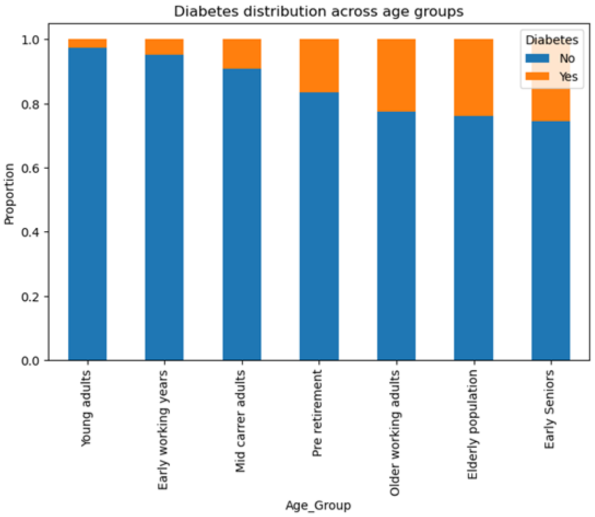

Fig. 1: Diabetes prevalence by Age category (BRFSS 2024).

2) Behavioral Insights:

- Diabetes prevalence increases steadily with age: only ∼2.5% in young adults and ∼25.5% in early seniors.
- The biggest jump in prevalence occurs in pre-retirement and older age groups (16–23%).
- This trend reflects that age is a strong risk factor for diabetes.
- Overall, it reflects that diabetes can be found in young people or anyone irrelevant of their age

B. Diabetes vs Age and Weight categories

1) Demographics: The percentage of respondents with diabetes within each weight category across age groups is summarized in Table III.

Age Group	Underweight (%)	Normal weight (%)	Overweight (%)	Obese (%)
Early Seniors	13.61	16.39	26.99	43.01
Elderly population	12.63	18.45	27.97	40.95
Older working adults	15.84	15.61	25.05	43.50
Pre-retirement	17.91	14.59	22.14	45.35
Mid career adults	15.75	14.40	20.42	49.42
Early working years	20.92	14.18	18.41	46.49
Young adults	13.86	13.69	22.87	49.58

TABLE III: Percentage of respondents with diabetes by weight category and age group (BRFSS 2024).

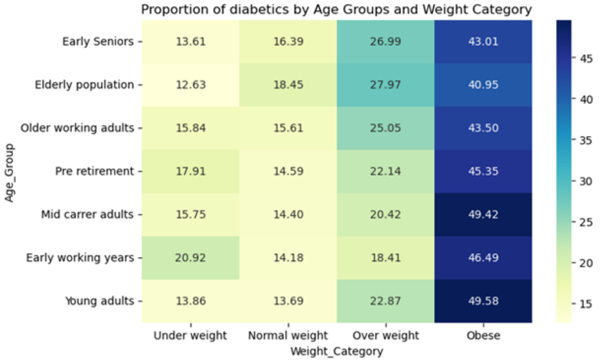

Fig. 2: Diabetes prevalence by Age and Weight categories (BRFSS 2024).

2) Behavioral Insights:

- Obesity is the strongest predictor of diabetes, with prevalence exceeding 40% in all age groups.
- Overweight individuals also show elevated diabetes rates, although these rates are lower than those observed in obese individuals.
- Normal weight individuals have considerably lower diabetes prevalence across all age groups.
- Underweight individuals generally have the lowest diabetes prevalence, except in early working years where it rises to approximately 20.9%.
- Young adults and mid-career adults who are obese show the highest diabetes prevalence, around 49–50%, indicating that diabetes affects younger adults significantly in high weight categories.
- Overall, diabetes prevalence increases with weight category regardless of age, but the effect is most pronounced among younger and mid-career adults.

C. Diabetes vs Exercise

1) Demographics: The prevalence of diabetes among respondents based on exercise activity is summarized in Table IV.

Exercise Activities	No Diabetes (%)	Yes Diabetes (%)
No	72.61	27.39
Yes	85.35	14.65

TABLE IV: Diabetes prevalence by exercise activity (BRFSS 2024).

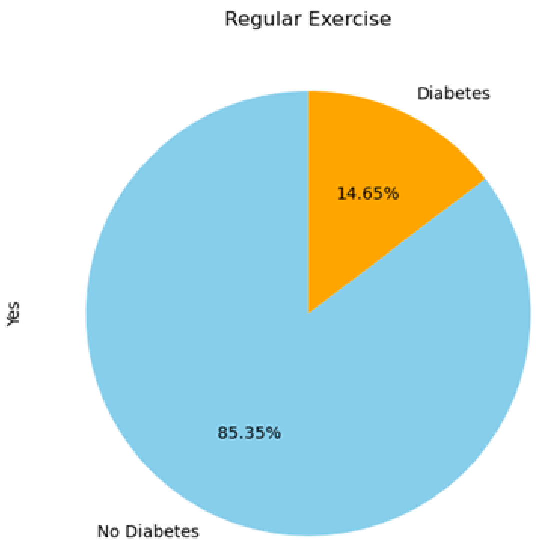

Fig. 3: Diabetes and its prevalence on people who do exercise

2) *Behavioral Insights:*

- Individuals who do not exercise have a much higher diabetes prevalence (27%) compared to those who exercise (15%).
- Regular exercise is associated with a higher proportion of diabetes free individuals (85% vs 73%).
- Exercise shows a protective effect, although it does not fully prevent diabetes.

D. *Diabetes vs Smoking*

1) Demographics: The distribution of diabetes prevalence across smoking status is summarized in Table V.

Smoking Condition	No Diabetes	Yes Diabetes
Everyday	26,442	6,578
Former	93,854	25,540
Never	217,695	40,789
Sometime	11,648	2,277

TABLE V: Diabetes prevalence by smoking status (BRFSS 2024).

2) *Behavioral Insights by Smoking Status:*

- Diabetes prevalence is highest among former smokers (21.4%, 25,540 of 119,394) and everyday smokers (19.9%, 6,578 of 33,020), while sometime smokers (16.3%, 2,277 of 13,925) and never smokers (15.8%, 40,789 of 258,484) show lower rates.
- Most non-diabetics are never smokers, with 84.2% (217,695 of 258,484) free from diabetes.
- Diabetes occurs across all smoking categories, indicating that smoking alone does not primarily determine risk.

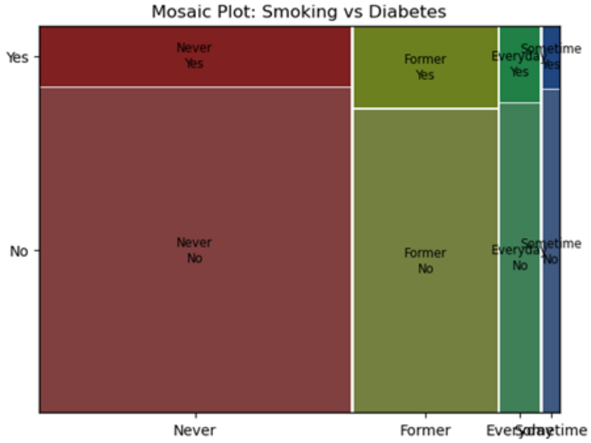

Fig. 4: Diabetes and its prevalence on smoking

E. *Diabetes vs Various life style and factors*

1) *Demographics:*

- Age Group (0.205), older individuals have higher diabetes risk; age is the strongest demographic predictor.
- BMI (0.195), higher BMI is moderately associated with diabetes, showing body fat distribution matters more than weight alone.
- Weight Category (0.052), absolute weight alone has a very weak correlation with diabetes; less informative than BMI.

Variable	Correlation with Diabetes
Somking_condtion	-0.030
Weight_Category	0.052
BMI_actual	0.195
Age_Group	0.205
Exercise_Activites	-0.135

TABLE VI: Demographic and Behavioral Factors Associated with Diabetes

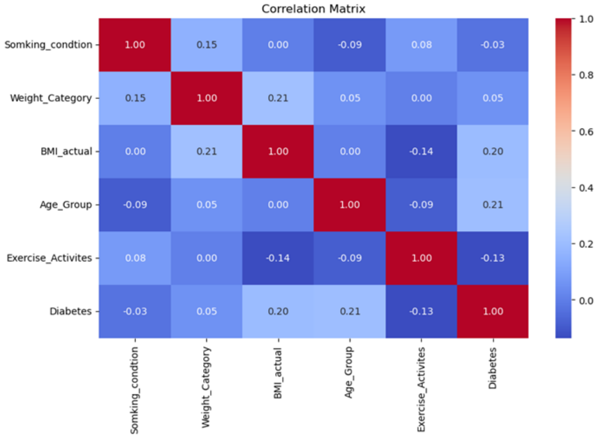

Fig. 5: Diabetes and its prevalence on smoking

2) *Behavioral Insights:*

- Smoking Condition (-0.030), Smoking shows almost no effect on diabetes risk in this dataset.

- Exercise Activities (-0.135), Physical activity slightly reduces diabetes risk, but effect is smaller compared to age and BMI.

IV. Myth Identification and Results

Based upon the above data analysis and the five common myths were formulated and tested, they are as follows

A. *Myth 1: Diabetes only affects ypung adults*

Table II and the Figure 1 shows that while diabetes prevalence increases with age, it is not exclusive to the elderly. Even among **young adults**, approximately **2.5%** report diabetes, and this figure rises steadily across age groups, reaching over **25%** among early seniors. This pattern dispels the misconception that diabetes is solely a disease of older adults. The data indicate that although aging is a major risk factor, **lifestyle choices and BMI** contribute significantly to diabetes risk across all age brackets.

B. *Myth 2: Only overweight or obese people get diabetes.*

Table III and the Figure 2 shows that Analysis of weight distribution across age groups shows that approximately 27–31% of individuals are underweight or normal weight. Although obesity is a known risk factor for diabetes, a significant proportion of lean individuals are still at risk. Even in age groups with the highest obesity prevalence (approximately 49–50% in young and mid-career adults), underweight and normal-weight individuals remain present. These findings indicate that diabetes is not confined to overweight or obese individuals and that all adults should be considered for risk assessment.

C. *Myth 3: Physical activity has little or no effect on diabetes risk.*

Table IV and the Figure 3 shows that the individuals who engage in regular physical activity have a markedly lower prevalence of diabetes. Specifically, only 14.65% of those who exercise regularly have diabetes, highlighting that physical activity substantially reduces diabetes risk and disproving the notion that exercise has minimal impact on diabetes.

D. *Myth 4: Smoking does not affect diabetes risk.*

Table V and the Figure 4 show a clear association between smoking and diabetes prevalence. Individuals who smoke daily or are former smokers exhibit higher diabetes prevalence (approximately 19.9% and 21.4%, respectively) compared to those who never smoked (15.8%). Even occasional smokers show slightly elevated risk (16.4%). These findings indicate that smoking is an important modifiable risk factor for diabetes, highlighting the need for targeted interventions and smoking cessation programs to reduce diabetes incidence.

E. *Myth 5: Lifestyle factors other than exercise, such as smoking, have a strong direct impact on diabetes.*

It is often assumed that lifestyle factors such as smoking directly and strongly influence diabetes risk. However, the correlation analysis indicates that smoking condition has a negligible direct relationship with diabetes (correlation = -0.03). This suggests that while smoking may contribute indirectly through other health mechanisms, it is not a strong independent predictor of diabetes in this population, highlighting the need to focus on multiple interacting risk factors rather than smoking alone. The correlation between these factors can be seen from the table VI and the figure 5.

V. Discussion and Conclusion

This study analyzed the prevalence of diabetes across different demographic and behavioral factors using the BRFSS 2024 dataset. The analysis revealed that age and BMI are positively correlated with diabetes prevalence [1], indicating higher risk among older and overweight individuals. In contrast, lifestyle factors such as exercise [2] showed a negative association, highlighting the protective role of physical activity. Behavioral factors, including smoking [3] , exhibited minimal correlation with diabetes in this dataset. These findings emphasize the importance of targeted preventive measures focusing on high-risk demographic groups and promoting healthy lifestyle interventions [4]. Future work can expand this study by incorporating longitudinal data and additional behavioral variables to further enhance diabetes risk prediction models.

References

[1] J. S. Varghese, “National, state, and county estimates of adult overweight and obesity from electronic health records and kiosks in retail locations, 2024-2025,” *medRxiv*, pp. 2025–08, 2025.

[2] H. Chen and L. Guo, “Exercise in diabetic cardiomyopathy: Its protective effects and molecular mechanism,” *International Journal of Molecular Sciences*, vol. 26, no. 4, p. 1465, 2025.

[3] G. R. M. La Rosa, E. Pedulla, I. Chapple, J. Kowalski, M. Walicka, S. Piro, and R. Polosa, “A systematic review of oral health outcomes following smoking cessation in type 2 diabetes: clinical and research implications,” *Journal of Dentistry*, p. 105665, 2025.

[4] Q. Liao, T. Yu, J. Chen, X. Zheng, L. Zheng, and J. Yan, “Relationship between maternal pre-pregnancy bmi and neonatal birth weight in pregnancies with gestational diabetes mellitus: a retrospective cohort study,” *Frontiers in Medicine*, vol. 11, p. 1478907, 2025.