Combined Impact of Physical Activity and Diet on Reducing Diabetes Risk

data-to-paper

October 13, 2024

Abstract

Diabetes is a pressing global health concern, necessitating effective prevention strategies. Although physical activity and diet are recognized as key modifiers of diabetes risk, their synergistic effects require further examination. This study investigates these interactions using data from the CDC's 2015 Behavioral Risk Factor Surveillance System, involving over 250,000 individuals. Through logistic regression analysis, our research identifies a significant reduction in diabetes risk associated with the combined effect of regular physical activity and fruit consumption. In contrast, the addition of vegetable intake did not notably enhance this protective association. Furthermore, age and BMI emerged as prominent risk factors, whereas socioeconomic factors like income and education played a protective role. These findings advocate for lifestyle interventions that integrate physical activity with dietary improvements, particularly fruit consumption, as a strategic approach to diabetes prevention. Nonetheless, the moderate explanatory power of our model suggests the necessity for a broader investigation into additional contributing factors and underscores the importance of a holistic approach to disease prevention.

Introduction

Diabetes mellitus, notably type 2 diabetes, remains a significant public health challenge worldwide due to its increasing prevalence and associated complications, including cardiovascular diseases and diminished quality of life [1, 2]. Lifestyle modifications, particularly physical activity and diet, have been underscored in various studies as effective strategies for mitigating the risk of developing type 2 diabetes [3, 4]. Recent guidelines consistently highlight the importance of integrating these lifestyle interventions into diabetes prevention programs [5]. Despite these recommendations, there remains considerable need for further elucidation of interactive effects between physical activity and specific dietary components, such as fruit and vegetable consumption, in diabetic risk reduction [2, 6].

Extensive research has explored the relationships between obesity, diet, and physical activity with diabetes, yet there is still ambiguity regarding the combined effects of these variables [7, 8]. Prior studies have established the independent benefits of exercise and healthy eating habits, but it is less clear how these factors might work together to further inform preventive strategies [9, 10]. Furthermore, while socioeconomic factors have been recognized for their potential protective roles, the interplay with lifestyle modifications needs clarification [11, 12]. Clarifying these aspects can improve our understanding of diabetes etiology and inform comprehensive public health interventions aimed at mitigating diabetes risk.

In this context, the current study aims to dissect the synergistic effects of physical activity and fruit and vegetable intake on the risk of diabetes using data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) [13]. This dataset offers a vast sample size and comprehensive variables, making it an excellent resource for examining these intricate relationships [14, 15]. Unlike previous studies, which frequently focused on isolated lifestyle factors, this research emphasizes their combined impact and incorporates socioeconomic variables into the analysis [16, 17]. In doing so, our study fills a pivotal gap by providing insights into how integrated lifestyle changes can be effectively promoted as preventive measures against diabetes.

Methodologically, we utilized logistic regression models to examine the influence of physical activity and dietary habits on diabetes risk, while also considering demographic and socioeconomic factors [18]. By including interaction terms for physical activity and fruit and vegetable consumption, this study comprehensively evaluates their potential combined effects [19]. The analysis revealed significant associations between physical activity coupled with fruit consumption and reduced diabetes risk, a finding that supports integrating fruit intake into physical activity recommendations [20, 21]. Ultimately, this study underscores the complexities of diabetes prevention, demonstrating how targeted lifestyle modifications can serve as strategic public health interventions.

Results

To begin exploring the relationship between lifestyle factors and diabetes risk, we conducted a descriptive analysis of key variables. The descriptive statistics presented in Table 1 provide an overview of participants' engagement in physical activity, fruit, and vegetable consumption, along with diabetes diagnosis and demographic factors such as age, BMI, income, and education level. On average, 75.65% of participants reported engaging in physical activity in the past 30 days, while 63.43% and 81.14% consumed one or more fruits and vegetables daily, respectively. The mean BMI of participants was 28.38, indicating a trend toward overweight in the population studied. Notably, 13.93% of participants reported having been diagnosed with diabetes.

Table 1: Descriptive Statistics of Key Variables Related to Diabetes Risk

	mean	std
Physical Activity	0.7565	0.4292
Fruit Consumption	0.6343	0.4816
Vegetable Consumption	0.8114	0.3912
Diabetes Diagnosis	0.1393	0.3463
Age Group	8.032	3.054
Body Mass Index	28.38	6.609
Income Level	6.054	2.071
Education Level	5.05	0.9858

Physical Activity: Engaged in physical activity in the past 30 days, 1: Yes, 0: No
Fruit Consumption: Consumes one or more fruits daily, 1: Yes, 0: No
Vegetable Consumption: Consumes one or more vegetables daily, 1: Yes, 0: No
Diabetes Diagnosis: Diagnosis of diabetes, 1: Yes, 0: No
Age Group: Age category (1=18-24, ..., 13=80 or older)
Body Mass Index: Body Mass Index, kg/m^{2}
Income Level: Income scale (1: <=10k, ..., 8: >75k)
Education Level: Education scale (1=Never attended school, ..., 6=College)

Subsequently, we applied logistic regression analysis to evaluate the association between physical activity, dietary habits, and diabetes risk. The analysis, depicted in Figure 1, assesses the interaction effects of physical activity and consumption of fruits and vegetables, while controlling for confounding variables, including standardized age, BMI, income, and education. The results demonstrate that physical activity has a protective effect, with the interaction term between physical activity and fruit consumption

(PA $*$ Fruits) showing a significant odds ratio less than one (-0.1718, $P < 10^{-6}$). However, the interaction between physical activity and vegetable consumption (PA $*$ Veggies) did not reach conventional significance levels (-0.05864, $P = 0.055$). Regression analysis further identified standardized age and BMI as strong predictors of diabetes risk, with age (0.6304, $P < 10^{-6}$) and BMI (0.5443, $P < 10^{-6}$) being positively associated with diabetes prevalence.

Finally, in summary, our analysis of data on 253680 individuals, of whom 35346 reported having diabetes, highlights the complexity of diabetes risk and the role of modifiable lifestyle factors. The logistic regression model achieved a pseudo R-squared value of 0.1226, suggesting that additional variables may influence diabetes risk beyond those captured in our model. Overall, the results underscore the significance of physical activity, particularly when coupled with fruit consumption, as a preventive strategy against diabetes.

Discussion

This study aimed to elucidate the combined impact of physical activity and dietary habits, particularly fruit and vegetable intake, on diabetes risk using data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset. Recent literature has highlighted lifestyle modifications, including physical activity and dietary improvements, as effective preventive measures against type 2 diabetes [1, 2]. Prior studies have noted the independent benefits of such interventions but have called for a deeper understanding of their interactive effects on diabetes risk [3, 6]. Our analysis sought to bridge this gap by employing logistic regression models to assess the synergistic effects of these lifestyle factors, while incorporating socioeconomic variables.

Methodologically, we utilized logistic regression to explore the effects of physical activity and fruit and vegetable consumption on diabetes risk, with interaction terms included to capture potential synergies. Our comprehensive analysis of 253,680 respondents from the BRFSS dataset indicated that physical activity and fruit consumption jointly reduced diabetes risk, as evidenced by a significant interaction effect (PA $*$ Fruits: odds ratio less than one, $P < 10^{-6}$). Conversely, the interaction between physical activity and vegetable consumption did not exhibit significant protective association ($P = 0.055$). These findings align partially with existing literature which recognizes the individual benefits of physical activity and healthy diets [22, 23], yet offer novel insights into their combined effect, a subject less

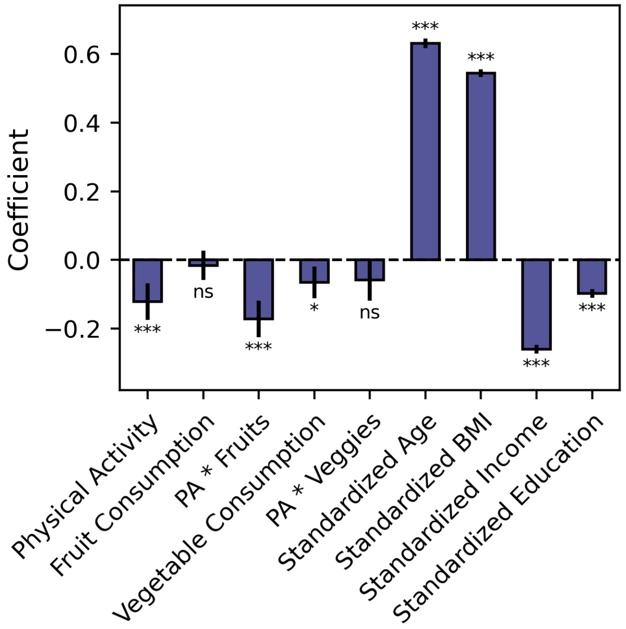

Figure 1: Logistic Regression Analysis for Interaction Between Physical Activity and Dietary Habits on Diabetes Risk The figure illustrates the interaction effects between physical activity and dietary habits on diabetes risk, omitting the intercept term. Physical Activity: Engaged in physical activity in the past 30 days, 1: Yes, 0: No. Fruit Consumption: Consumes one or more fruits daily, 1: Yes, 0: No. Vegetable Consumption: Consumes one or more vegetables daily, 1: Yes, 0: No. Std. Error: Standard error of the coefficient. CI Lower: Lower bound of 95% confidence interval. CI Upper: Upper bound of 95% confidence interval. PA * Fruits: Interaction term between Physical Activity and Fruit Consumption. PA * Veggies: Interaction term between Physical Activity and Vegetable Consumption. Standardized Age: Age standardized for logistic regression. Standardized BMI: BMI standardized for logistic regression. Standardized Income: Income standardized for logistic regression. Standardized Education: Education level standardized for logistic regression. Significance: ns p >= 0.01, * p < 0.01, ** p < 0.001, *** p < 0.0001.

frequently detailed in current research [7, 8].

Despite the robust dataset and rigorous analytical approach, this study is not without limitations. First, the cross-sectional nature of the BRFSS data precludes the establishment of causal relationships. Longitudinal studies are necessary to confirm the temporal direction of the observed associations. Second, self-reported data on lifestyle and health behaviors may be subject to recall bias, potentially affecting the reliability of the findings. Furthermore, although the inclusion of socioeconomic factors provides a broader context for diabetes risk, other unmeasured confounding variables, such as genetic predispositions, could moderate our results. Finally, while the pseudo R-squared value provides some indication of model fit, it underscores the multifactorial nature of diabetes, suggesting additional parameters may need exploration.

In conclusion, this study underscores the significance of integrating physical activity with dietary modifications, particularly fruit consumption, in diabetes prevention efforts. Our findings advocate for public health strategies that align with previous recommendations yet further emphasize the value of combined lifestyle interventions [11, 12]. Moving forward, future research should aim to establish causative links through prospective cohort studies and consider a more holistic approach that includes genetic, environmental, and psychosocial factors. Such comprehensive assessments will be crucial for the formulation of effective, targeted interventions to mitigate the burgeoning burden of diabetes globally.

Methods

Data Source

The present study utilizes the dataset derived from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), managed by the Centers for Disease Control and Prevention (CDC). This annual telephone survey gathers comprehensive health-related data from more than 400,000 individuals across the United States, focusing on health-related risk behaviors, chronic health conditions, and the utilization of preventive services. The dataset employed consists of 253,680 responses, encompassing 22 distinct variables related to diabetes risk factors, such as high blood pressure, cholesterol levels, body mass index, smoking habits, dietary patterns, physical activity, and demographic characteristics. The dataset is notably free from missing values, providing a robust basis for examining the interaction effects of various lifestyle factors on diabetes risk.

Data Preprocessing

In preparation for analysis, the dataset underwent a standardization process to facilitate the comparative evaluation of key continuous variables, specifically body mass index (BMI), age, income, and education. These variables were standardized to enable direct interpretability of their coefficients in the logistic regression model. Additionally, interaction terms were created to explore potential synergistic effects of physical activity with fruit and vegetable consumption on diabetes risk. The interaction terms were defined as the product of binary indicators for physical activity and the consumption of fruits and vegetables.

Data Analysis

The analysis commenced with the computation of summary statistics, capturing the mean and standard deviation for a subset of variables pertinent to diabetes risk. The primary analytical approach involved logistic regression modeling to ascertain the influence of physical activity and dietary habits on the likelihood of diabetes diagnosis. The model integrated interaction terms to assess whether the combination of physical activity with fruit or vegetable intake provided additive benefits in diabetes risk mitigation. Regression outcomes were evaluated by estimating coefficients for the predictor variables, with additional calculations for their respective standard errors, confidence intervals, and significance levels. Furthermore, model fit was assessed using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and pseudo R-squared, providing insights into the explanatory power achieved through the inclusion of both lifestyle and demographic variables.

Code Availability

Custom code used to perform the data preprocessing and analysis, as well as the raw code outputs, are provided in Supplementary Methods.

References

[1] S. Colberg, R. Sigal, J. Yardley, M. Riddell, D. Dunstan, P. Dempsey, E. Horton, Kristin Castorino, and D. Tate. Physical activity/exercise and diabetes: A position statement of the american diabetes association. *Diabetes Care*, 39:2065 – 2079, 2016.

[2] S. Bird and J. Hawley. Update on the effects of physical activity on insulin sensitivity in humans. *BMJ Open Sport Exercise Medicine*, 2, 2017.

[3] R. Lehmann, V. Kaplan, R. Bingisser, K. Bloch, and G. Spinas. Impact of physical activity on cardiovascular risk factors in iddm. *Diabetes Care*, 20:1603 – 1611, 1997.

[4] Pmella Goveia, Wilson Can-Montaez, D. Santos, G. Lopes, R. Ma, B. Duncan, P. Ziegelman, and M. Schmidt. Lifestyle intervention for the prevention of diabetes in women with previous gestational diabetes mellitus: A systematic review and meta-analysis. *Frontiers in Endocrinology*, 9, 2018.

[5] D. Prcoma, G. Oliveira, A. F. Simo, . Dutra, O. Coelho, M. C. Izar, R. Pvoa, I. Giuliano, A. C. A. Alencar Filho, C. Machado, C. Scherr, F. Fonseca, R. S. Santos Filho, T. Carvalho, . Avezum, R. Esporcatte, B. Nascimento, D. Brasil, G. P. Soares, P. B. Villela, R. M. Ferreira, W. Martins, A. Sposito, B. Halpern, J. Saraiva, L. Carvalho, M. Tambascia, O. Coelho-Filho, A. Bertolami, H. Corra Filho, H. Xavier, J. Faria-Neto, M. Bertolami, V. Giraldez, A. Brando, Audes D M Feitosa, C. Amodeo, D. S. M. Souza, E. Barbosa, M. V. Malachias, W. Souza, F. Costa, I. Rivera, L. Pellanda, M. A. F. D. Silva, A. Achutti, A. R. Langowiski, C. Lantieri, J. Scholz, S. Ismael, J. C. Ayoub, Luiz Csar Nazrio Scala, M. Neves, P. Jardim, S. Fuchs, T. Jardim, E. Moriguchi, J. C. Schneider, M. Assad, S. Kaiser, A. M. Lottenberg, C. Magnoni, M. Miname, Roberta Soares Lara, A. Herdy, CLAUDIO GIL ARAUJO, M. Milani, Miguel Morita Fernandes da Silva, R. Stein, F. Lucchese, F. Nobre, Hermilo Borba Griz, L. Magalhes, Mario Henrique Elesbo de Borba, M. Pontes, and R. Mourilhe-Rocha. Updated cardiovascular prevention guideline of the brazilian society of cardiology - 2019. *Arquivos Brasileiros de Cardiologia*, 113:787 – 891, 2019.

[6] L. Bazzano, Tricia Y. Li, Kamudi J. Joshipura, and F. Hu. Intake of fruit, vegetables, and fruit juices and risk of diabetes in women. *Diabetes Care*, 31:1311 – 1317, 2008.

[7] K. V. Hjerkind, J. Stenehjem, and T. Nilsen. Adiposity, physical activity and risk of diabetes mellitus: prospective data from the population-based hunt study, norway. *BMJ Open*, 7, 2017.

[8] I. Vuori. Health benefits of physical activity with special reference to interaction with diet. *Public Health Nutrition*, 4:517 – 528, 2001.

[9] R. Bailey, J. Singleton, and J. Majersik. Association of obesity and diabetes with physical activity and fruit and vegetable consumption in stroke survivors. *Family practice*, 2020.

[10] Zixin Zeng, Yuqian Bian, Y. Cui, Donghui Yang, Yafeng Wang, and Chuanhua Yu. Physical activity dimensions and its association with risk of diabetes in middle and older aged chinese people. *International Journal of Environmental Research and Public Health*, 17, 2020.

[11] S. Stringhini, Adam G. Tabk, T. Akbaraly, S. Sabia, M. Shipley, M. Marmot, Eric J. Brunner, David Batty, P. Bovet, M. Kivimki, Open Access, Adam G. Tabk, T. Akbaraly, M. Shipley, M. Marmot, Eric J. Brunner, and G. Batty. Contribution of modifiable risk factors to social inequalities in type 2 diabetes: prospective whitehall ii cohort study. *The BMJ*, 345, 2012.

[12] A. Krist, K. Davidson, C. Mangione, M. Barry, M. Cabana, A. Caughey, Katrina E. Donahue, Chyke A Doubeni, J. Epling, Martha Y. Kubik, S. Landefeld, G. Ogedegbe, L. Pbert, Michael Silverstein, M. Simon, Chien-Wen Tseng, and J. Wong. Behavioral counseling interventions to promote a healthy diet and physical activity for cardiovascular disease prevention in adults with cardiovascular risk factors: Us preventive services task force recommendation statement. *JAMA*, 324 20:2069–2075, 2020.

[13] L. Moore, K. Dodd, F. Thompson, K. Grimm, Sonia A. Kim, and K. Scanlon. Using behavioral risk factor surveillance system data to estimate the percentage of the population meeting us department of agriculture food patterns fruit and vegetable intake recommendations. *American journal of epidemiology*, 181 12:979–88, 2015.

[14] Xingyou Zhang, J. Holt, Shumei Yun, Hua Lu, K. Greenlund, and J. Croft. Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American journal of epidemiology*, 182 2:127–37, 2015.

[15] Elie S. Al Kazzi, B. Lau, Tianjing Li, E. Schneider, M. Makary, and S. Hutfless. Differences in the prevalence of obesity, smoking and alcohol

in the united states nationwide inpatient sample and the behavioral risk factor surveillance system. *PLoS ONE*, 10, 2015.

[16] Nandita Bhan, I. Kawachi, M. Glymour, and S. Subramanian. Time trends in racial and ethnic disparities in asthma prevalence in the united states from the behavioral risk factor surveillance system (brfss) study (1999-2011). *American journal of public health*, 105 6:1269–75, 2015.

[17] Elizabeth L. Tung, Arshiya A. Baig, E. Huang, N. Laiteerapong, and Kao-Ping Chua. Racial and ethnic disparities in diabetes screening between asian americans and other adults: Brfss 20122014. *Journal of General Internal Medicine*, 32:423–429, 2017.

[18] Ram Joshi and Chandra Dhakal. Predicting type 2 diabetes using logistic regression and machine learning approaches. *International Journal of Environmental Research and Public Health*, 18, 2021.

[19] N. Senaviratna and T. M. J. A. Cooray. Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics*, 2019.

[20] T. Gress, F. Nieto, E. Shahar, M. Wofford, and F. Brancati. Hypertension and antihypertensive therapy as risk factors for type 2 diabetes mellitus. *The New England Journal of Medicine*, 342:905–912, 2000.

[21] K. Meyer, L. Kushi, D. Jacobs, J. Slavin, T. Sellers, and A. Folsom. Carbohydrates, dietary fiber, and incident type 2 diabetes in older women. *The American journal of clinical nutrition*, 71 4:921–30, 2000.

[22] S. R. Mortensen, P. Kristensen, A. Grntved, M. Ried-Larsen, C. Lau, and S. Skou. Determinants of physical activity among 6856 individuals with diabetes: a nationwide cross-sectional study. *BMJ Open Diabetes Research & Care*, 10, 2022.

[23] Yasmin Aridi, Jacqueline L. Walker, E. Roura, and O. Wright. Adherence to the mediterranean diet and chronic disease in australia: National nutrition and physical activity survey analysis. *Nutrients*, 12, 2020.

A Data Description

Here is the data description, as provided by the user:

```
## General Description
The dataset includes diabetes related factors extracted from
    the CDC’s Behavioral Risk Factor Surveillance System (BRFSS
    ), year 2015.
The original BRFSS, from which this dataset is derived, is a
    health-related telephone survey that is collected annually
    by the CDC.
Each year, the survey collects responses from over 400,000
    Americans on health-related risk behaviors, chronic health
    conditions, and the use of preventative services. These
    features are either questions directly asked of
    participants, or calculated variables based on individual
    participant responses.

## Data Files
The dataset consists of 1 data file:

### "diabetes_binary_health_indicators_BRFSS2015.csv"
The csv file is a clean dataset of 253,680 responses (rows) and
     22 features (columns).
All rows with missing values were removed from the original
    dataset; the current file contains no missing values.

The columns in the dataset are:

#1 ‘Diabetes_binary‘: (int, bool) Diabetes (0=no, 1=yes)
#2 ‘HighBP‘: (int, bool) High Blood Pressure (0=no, 1=yes)
#3 ‘HighChol‘: (int, bool) High Cholesterol (0=no, 1=yes)
#4 ‘CholCheck‘: (int, bool) Cholesterol check in 5 years (0=no,
     1=yes)
#5 ‘BMI‘: (int, numerical) Body Mass Index
#6 ‘Smoker‘: (int, bool) (0=no, 1=yes)
#7 ‘Stroke‘: (int, bool) Stroke (0=no, 1=yes)
#8 ‘HeartDiseaseorAttack‘: (int, bool) coronary heart disease (
    CHD) or myocardial infarction (MI), (0=no, 1=yes)
#9 ‘PhysActivity‘: (int, bool) Physical Activity in past 30
    days (0=no, 1=yes)
#10 ‘Fruits‘: (int, bool) Consume one fruit or more each day (
    0=no, 1=yes)
#11 ‘Veggies‘: (int, bool) Consume one Vegetable or more each
    day (0=no, 1=yes)
#12 ‘HvyAlcoholConsump‘ (int, bool) Heavy drinkers (0=no, 1=yes
    )
```

```
#13 ‘AnyHealthcare‘ (int, bool) Have any kind of health care
    coverage (0=no, 1=yes)
#14 ‘NoDocbcCost‘ (int, bool) Was there a time in the past 12
    months when you needed to see a doctor but could not
    because of cost? (0=no, 1=yes)
#15 ‘GenHlth‘ (int, ordinal) self-reported health (1=excellent,
     2=very good, 3=good, 4=fair, 5=poor)
#16 ‘MentHlth‘ (int, ordinal) How many days during the past 30
    days was your mental health not good? (1 - 30 days)
#17 ‘PhysHlth‘ (int, ordinal) Hor how many days during the past
     30 days was your physical health not good? (1 - 30 days)
#18 ‘DiffWalk‘ (int, bool) Do you have serious difficulty
    walking or climbing stairs? (0=no, 1=yes)
#19 ‘Sex‘ (int, categorical) Sex (0=female, 1=male)
#20 ‘Age‘ (int, ordinal) Age, 13-level age category in
    intervals of 5 years (1= 18 - 24, 2= 25 - 29, ..., 12= 75 -
     79, 13 = 80 or older)
#21 ‘Education‘ (int, ordinal) Education level on a scale of 1
    - 6 (1=Never attended school, 2=Elementary, 3=Some high
    school, 4=High school, 5=Some college, 6=College)
#22 ‘Income‘ (int, ordinal) Income scale on a scale of 1 to 8 (
    1= <=10K, 2= <=15K, 3= <=20K, 4= <=25K, 5= <=35K, 6= <=50K,
     7= <=75K, 8= >75K)
```

B Data Exploration

B.1 Code

The Data Exploration was carried out using the following custom code:

```
import pandas as pd

# Load the dataset
data = pd.read_csv("diabetes_binary_health_indicators_BRFSS2015
    ↪ .csv")

# Open the output file
with open("data_exploration.txt", "w") as file:

    # Data Size
    file.write("# Data Size\n")
    data_shape = data.shape
    file.write(f"Number of rows: {data_shape[0]}\n")
    file.write(f"Number of columns: {data_shape[1]}\n\n")
```

```
# Summary Statistics
file.write("# Summary Statistics\n")
summary_stats = data.describe().transpose()
file.write(f"{summary_stats}\n\n")

# Categorical Variables
file.write("# Categorical Variables\n")
categorical_columns = ["Diabetes_binary", "HighBP", "
    ↪ HighChol", "CholCheck", "Smoker", "Stroke",
                       "HeartDiseaseorAttack", "
                          ↪ PhysActivity", "Fruits", "
                          ↪ Veggies", "HvyAlcoholConsump",
                       "AnyHealthcare", "NoDocbcCost", "
                          ↪ DiffWalk", "Sex", "Age", "
                          ↪ Education", "Income"]

for column in categorical_columns:
    counts = data[column].value_counts()
    file.write(f"Column: {column}\n")
    file.write(f"Most common value: {counts.idxmax()} with
        ↪ {counts.max()} occurrences\n\n")

# Missing Values
file.write("# Missing Values\n")
missing_values_count = data.isnull().sum().sum()
if missing_values_count > 0:
    file.write(f"There are {missing_values_count} missing
        ↪ values in the dataset.\n")
else:
    file.write("# Not Applicable\n")

# Unique categorical values check
file.write("# Unique Values in Categorical Variables\n")
for column in categorical_columns:
    unique_values = data[column].nunique()
    file.write(f"Column: {column}, Unique Values: {
        ↪ unique_values}\n")

# Any other summary deemed relevant
file.write("# Correlation with 'Diabetes_binary'\n")
correlation = data.corr()['Diabetes_binary']
file.write(f"{correlation}\n\n")
```

B.2 Code Description

In this data exploration process, the provided code performs a series of analytical steps on the diabetes-related dataset to understand its structure, characteristics, and relationships between variables. The exploratory data analysis (EDA) can be summarized as follows:

B.3 Data Loading and Initialization

The dataset is imported as a pandas DataFrame for ease of manipulation. The initial step ensures that all necessary data points are readily available for evaluation and that the DataFrame's structure aligns with the expectations of 253,680 rows and 22 features.

B.4 Data Size

The code calculates the dimensions of the dataset, which includes the number of rows and columns. This step is crucial to confirm the expected size of the dataset, allowing researchers to ensure that the dataset is complete and that no truncation errors occurred during data importation.

B.5 Summary Statistics

Summary statistics for each feature in the dataset are calculated and exported. This includes measures such as mean, standard deviation, min, and max. These statistics provide a quick insight into the distribution and range of values for numerical features, which is useful for identifying potential outliers and understanding data central tendency.

B.6 Analysis of Categorical Variables

For features that are categorized as categorical variables, such as binary indicators or ordinal scales, the code analyzes and stores the frequency of each unique value. Additionally, the most common value in each categorical variable is identified, which helps in understanding the dominant demographic or characteristic within the dataset.

B.7 Checking for Missing Values

A check for missing values is performed to confirm the integrity of the dataset. Since missing values can significantly impact the results of any analysis, ensuring their non-existence (as the dataset purports no missing values from its pre-processing description) confirms the reliability of subsequent analyses.

B.8 Unique Values in Categorical Variables

The number of unique values in each categorical column is assessed. This step provides insight into the variability and potential heterogeneity within each categorical feature, which aids in future modeling decisions, such as the choice of encoding techniques for machine learning algorithms.

B.9 Correlation Analysis

Finally, the code computes the Pearson correlation coefficient of each feature with respect to the target variable, 'Diabetes_binary'. This correlation analysis assists in identifying which features have a linear relationship with the prevalence of diabetes, guiding feature selection and hypothesis generation for further studies.

Overall, this methodical exploration provides a foundational understanding of the dataset, setting the stage for more detailed analysis and model-building efforts.

B.10 Code Output

data_exploration.txt

```
# Data Size
Number of rows: 253680
Number of columns: 22

# Summary Statistics
                      count    mean    std min 25% 50% 75% max
Diabetes_binary      253680  0.1393 0.3463   0   0   0   0   1
HighBP               253680   0.429 0.4949   0   0   0   1   1
HighChol             253680  0.4241 0.4942   0   0   0   1   1
CholCheck            253680  0.9627 0.1896   0   1   1   1   1
BMI                  253680   28.38  6.609  12  24  27  31  98
Smoker               253680  0.4432 0.4968   0   0   0   1   1
Stroke               253680 0.04057 0.1973   0   0   0   0   1
HeartDiseaseorAttack 253680 0.09419 0.2921   0   0   0   0   1
PhysActivity         253680  0.7565 0.4292   0   1   1   1   1
Fruits               253680  0.6343 0.4816   0   0   1   1   1
Veggies              253680  0.8114 0.3912   0   1   1   1   1
HvyAlcoholConsump    253680  0.0562 0.2303   0   0   0   0   1
AnyHealthcare        253680  0.9511 0.2158   0   1   1   1   1
NoDocbcCost          253680 0.08418 0.2777   0   0   0   0   1
GenHlth              253680   2.511  1.068   1   2   2   3   5
MentHlth             253680   3.185  7.413   0   0   0   2  30
PhysHlth             253680   4.242  8.718   0   0   0   3  30
```

```
DiffWalk             253680  0.1682 0.3741   0   0   0   0   1
Sex                  253680  0.4403 0.4964   0   0   0   1   1
Age                  253680   8.032  3.054   1   6   8  10  13
Education            253680    5.05 0.9858   1   4   5   6   6
Income               253680   6.054  2.071   1   5   7   8   8

# Categorical Variables
Column: Diabetes_binary
Most common value: 0 with 218334 occurrences

Column: HighBP
Most common value: 0 with 144851 occurrences

Column: HighChol
Most common value: 0 with 146089 occurrences

Column: CholCheck
Most common value: 1 with 244210 occurrences

Column: Smoker
Most common value: 0 with 141257 occurrences

Column: Stroke
Most common value: 0 with 243388 occurrences

Column: HeartDiseaseorAttack
Most common value: 0 with 229787 occurrences

Column: PhysActivity
Most common value: 1 with 191920 occurrences

Column: Fruits
Most common value: 1 with 160898 occurrences

Column: Veggies
Most common value: 1 with 205841 occurrences

Column: HvyAlcoholConsump
Most common value: 0 with 239424 occurrences

Column: AnyHealthcare
Most common value: 1 with 241263 occurrences

Column: NoDocbcCost
Most common value: 0 with 232326 occurrences

Column: DiffWalk
Most common value: 0 with 211005 occurrences
```

```
Column: Sex
Most common value: 0 with 141974 occurrences

Column: Age
Most common value: 9 with 33244 occurrences

Column: Education
Most common value: 6 with 107325 occurrences

Column: Income
Most common value: 8 with 90385 occurrences

# Missing Values
# Not Applicable
# Unique Values in Categorical Variables
Column: Diabetes_binary, Unique Values: 2
Column: HighBP, Unique Values: 2
Column: HighChol, Unique Values: 2
Column: CholCheck, Unique Values: 2
Column: Smoker, Unique Values: 2
Column: Stroke, Unique Values: 2
Column: HeartDiseaseorAttack, Unique Values: 2
Column: PhysActivity, Unique Values: 2
Column: Fruits, Unique Values: 2
Column: Veggies, Unique Values: 2
Column: HvyAlcoholConsump, Unique Values: 2
Column: AnyHealthcare, Unique Values: 2
Column: NoDocbcCost, Unique Values: 2
Column: DiffWalk, Unique Values: 2
Column: Sex, Unique Values: 2
Column: Age, Unique Values: 13
Column: Education, Unique Values: 6
Column: Income, Unique Values: 8
# Correlation with 'Diabetes_binary'
Diabetes_binary               1
HighBP                   0.2631
HighChol                 0.2003
CholCheck               0.06476
BMI                      0.2168
Smoker                  0.06079
Stroke                   0.1058
HeartDiseaseorAttack     0.1773
PhysActivity            -0.1181
Fruits                 -0.04078
Veggies                -0.05658
HvyAlcoholConsump      -0.05706
AnyHealthcare           0.01626
NoDocbcCost             0.03143
GenHlth                  0.2936
```

```
MentHlth                 0.06932
PhysHlth                  0.1713
DiffWalk                  0.2183
Sex                      0.03143
Age                       0.1774
Education                -0.1245
Income                   -0.1639
Name: Diabetes_binary, dtype: float64
```

C Data Analysis

C.1 Code

The Data Analysis was carried out using the following custom code:

```
# IMPORT
from my_utils import df_to_latex, df_to_figure
import pandas as pd
import numpy as np
import statsmodels.formula.api as smf
import pickle
from sklearn.preprocessing import StandardScaler

# LOAD DATA
df = pd.read_csv("diabetes_binary_health_indicators_BRFSS2015.
   ↪ csv")

# DATASET PREPARATIONS
# No dataset preparations are needed.

# DESCRIPTIVE STATISTICS
## Table df_desc_stat:
caption = "Descriptive statistics of key variables related to
   ↪ diabetes risk"
# Calculate mean and standard deviation
desc_stats = df[['PhysActivity', 'Fruits', 'Veggies', '
   ↪ Diabetes_binary', 'Age', 'BMI', 'Income', 'Education']].
   ↪ agg(['mean', 'std']).transpose()
df_to_latex(desc_stats, 'df_desc_stat', caption=caption)

# PREPROCESSING
# Standardize BMI, Age, Income, and Education for the logistic
   ↪ regression analysis
scaler = StandardScaler()
df[['Standardized_BMI', 'Standardized_Age', '
   ↪ Standardized_Income', 'Standardized_Education']] = scaler
   ↪ .fit_transform(
    df[['BMI', 'Age', 'Income', 'Education']]
)

# Creating interaction terms
df[’PhysActivity_Fruits’] = df[’PhysActivity’] * df[’Fruits’]
df[’PhysActivity_Veggies’] = df[’PhysActivity’] * df[’Veggies’]

# ANALYSIS
## Table df_logistic_reg:
caption = "Logistic regression analysis for the interaction
    ↪ between physical activity, fruit and vegetable
    ↪ consumption on diabetes risk"
# Perform logistic regression
model = smf.logit(formula=’Diabetes_binary ~ PhysActivity *
    ↪ Fruits + PhysActivity * Veggies + Standardized_Age +
    ↪ Standardized_BMI + Standardized_Income +
    ↪ Standardized_Education’, data=df).fit()
summary_frame = model.summary2().tables[1]  # Extract relevant
    ↪ regression summary table
# Format for LaTeX output
summary_frame = summary_frame[["Coef.", "Std.Err.", "P>|z|", "
    ↪ [0.025", "0.975]"]]
summary_frame.columns = ["Coefficient", "Standard Error", "P-
    ↪ value", "CI Lower", "CI Upper"]
df_to_latex(summary_frame, ’df_logistic_reg’, caption=caption)

# SAVE ADDITIONAL RESULTS
additional_results = {
    ’Total number of observations’: len(df),
    ’Number of diabetes cases’: df[’Diabetes_binary’].sum(),
    ’Logistic model AIC’: model.aic,
    ’Logistic model BIC’: model.bic,
    ’Model pseudo R-squared’: model.prsquared,
}
with open(’additional_results.pkl’, ’wb’) as f:
    pickle.dump(additional_results, f)
```

C.2 Provided Code

The code above is using the following provided functions:

```
def df_to_latex(df,
        filename: str, caption: str,
    ):
    """
    Saves a DataFrame ‘df‘ and creates a LaTeX table.
    ‘filename‘, ‘caption‘: as in ‘df.to_latex‘.
    """
```

```
def df_to_figure(
        df, filename: str, caption: str,
        x: Optional[str] = None, y: List[str] = None,
        kind: str = ’bar’,
        logx: bool = False, logy: bool = False,
        y_ci: Optional[List[str]] = None,
        y_p_value: Optional[List[str]] = None,
    ):
    """
    Save a DataFrame ‘df‘ and create a LaTeX figure.
    Parameters, for LaTex embedding of the figure:
    ‘filename‘: Filename for the figure.
    ‘caption‘: Caption for the figure.

    Parameters for df.plot():
    ‘x‘: Column name for x-axis (index by default).
    ‘y‘: List of m column names for y-axis (m=1 for single plot
       ↪ , m>1 for multiple plots).
    ‘kind‘: only bar is allowed.
    ‘logx‘ / ‘logy‘ (bool): log scale for x/y axis.

    ‘y_ci‘: Confidence intervals for errorbars.
        List of m column names indicating confidence intervals
           ↪ for each y column.
        Each element in these columns must be a Tuple[float,
           ↪ float], describing the lower and upper bounds of
           ↪ the CI.

     ‘y_p_value‘: List of m column names (List[str]) containing
        ↪  numeric p-values of the corresponding y columns.
        ↪ These numeric values will be automatically converted
        ↪  by df_to_figure to stars (’***’, ’**’, ’*’, ’ns’)
        ↪ and plotted above the error bars.

    If provided, the length of ‘y_ci‘, and ‘y_p_value‘ should
       ↪ be the same as of ‘y‘.

    Example:
    Suppose, we have:

    df_lin_reg_longevity = pd.DataFrame({
        ’adjusted_coef’: [0.4, ...], ’adjusted_coef_ci’:
           ↪ [(0.35, 0.47), ...], ’adjusted_coef_pval’:
           ↪ [0.012, ...],
        ’unadjusted_coef’: [0.2, ...], ’unadjusted_coef_ci’:
           ↪ [(0.16, 0.23), ...], ’unadjusted_coef_pval’:
           ↪ [0.0001, ...],
    }, index=[’var1’, ...])
```

```
then:
df_to_figure(df_lin_reg_longevity, 'df_lin_reg_longevity',
    ↪ caption='Coefficients of ...', kind='bar',
    y=['adjusted_coef', 'unadjusted_coef'],
    y_ci=['adjusted_coef_ci', 'unadjusted_coef_ci'],
    y_p_value=['adjusted_coef_pval', 'unadjusted_coef_pval
        ↪ '])
"""
```

C.3 Code Description

C.4 Data Import and Loading

The dataset of interest is imported using the pandas library from a CSV file containing diabetes-related health indicators gathered from the BRFSS 2015 survey. This dataset includes a large sample of observations with multiple features pertinent to diabetes risk factors.

C.5 Descriptive Statistics

Descriptive statistics, specifically the mean and standard deviation, were computed for key continuous variables including physical activity, fruit and vegetable consumption, diabetes status, age, BMI, income, and education. This statistical summary provides initial insights into the data's central tendency and dispersion. The summary statistics are formatted into a LaTeX table for inclusion in reports or publications.

C.6 Data Preprocessing

For the purpose of subsequent logistic regression analysis, several continuous variables (BMI, Age, Income, and Education) are standardized. Standardization is performed using 'StandardScaler' from the sklearn library, which centers each variable by subtracting the mean and scales it by the standard deviation. This transformation facilitates the interpretation of the regression coefficients and ensures comparability across variables with different units and scales.

Furthermore, interaction terms between physical activity and dietary components (fruits and vegetables) are created. These terms allow the analysis to assess whether the combined effect of physical activity and diet is synergistic or antagonistic in relation to diabetes risk.

C.7 Logistic Regression Analysis

A logistic regression model is constructed using the statsmodels library to analyze the relationship between diabetes status and the main predictors, including the interaction terms (physical activity with fruits, and physical activity with veggies), and adjusted for age, BMI, income, and education (all standardized). The model estimation yields coefficients which indicate the strength and direction of association, with accompanying standard errors and confidence intervals for the coefficients. The regression results are reformatted into a LaTeX table for formal presentation.

C.8 Results Storage

Additional results, including model diagnostics and overall fit statistics such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the pseudo R-squared value, are computed and stored. The total number of observations and the number of diabetes cases are also captured. These summarized results are saved using the ‘pickle‘ module for potential future reference and reporting.

The overall methodology applies statistical techniques to illuminate the interplay between physical activity, diet, and diabetes risk, ensuring findings are robustly captured and effectively communicated for further scrutiny and usage in research.

C.9 Code Output

df_desc_stat.pkl

```
                  mean    std
PhysActivity    0.7565 0.4292
Fruits          0.6343 0.4816
Veggies         0.8114 0.3912
Diabetes_binary 0.1393 0.3463
Age              8.032  3.054
BMI              28.38  6.609
Income           6.054  2.071
Education         5.05 0.9858
```

df_logistic_reg.pkl

```
                       Coefficient Standard Error   P-value CI
                           Lower CI Upper
Intercept                   -1.846        0.02121         0
    -1.888   -1.805
```

```
PhysActivity                -0.1213        0.02734   9.13e-06
    -0.1749  -0.06772
Fruits                     -0.01567        0.02177      0.472
    -0.05833     0.027
PhysActivity:Fruits         -0.1718        0.02695   1.86e-10
    -0.2246   -0.1189
Veggies                    -0.06512         0.0235    0.00558
    -0.1112  -0.01907
PhysActivity:Veggies       -0.05864        0.03056      0.055
    -0.1185  0.001259
Standardized_Age             0.6304       0.007265          0
    0.6161    0.6446
Standardized_BMI             0.5443       0.005749          0
    0.533    0.5555
Standardized_Income         -0.2601       0.006498          0
    -0.2729   -0.2474
Standardized_Education     -0.09721       0.006513   2.24e-50
    -0.11  -0.08445
```

additional_results.pkl

```
{
    'Total number of observations': 253680,
    'Number of diabetes cases': 35346,
    'Logistic model AIC': 1.797e+05,
    'Logistic model BIC': 1.799e+05,
    'Model pseudo R-squared': 0.1226,
}
```

D LaTeX Table Design

D.1 Code

The LaTeX Table Design was carried out using the following custom code:

```
# IMPORT
import pandas as pd
from my_utils import df_to_latex, df_to_figure, is_str_in_df,
   ↪ split_mapping, AbbrToNameDef

# PREPARATION FOR ALL TABLES AND FIGURES
all_mapping: AbbrToNameDef = {
    # Key metrics and their explanations
    'PhysActivity': ('Physical Activity', 'Engaged in physical
       ↪ activity in the past 30 days, 1: Yes, 0: No'),
    'Fruits': ('Fruit Consumption', 'Consumes one or more
       ↪ fruits daily, 1: Yes, 0: No'),
    'Veggies': ('Vegetable Consumption', 'Consumes one or more
       ↪ vegetables daily, 1: Yes, 0: No'),
    'Diabetes_binary': ('Diabetes Diagnosis', 'Diagnosis of
       ↪ diabetes, 1: Yes, 0: No'),
    'Age': ('Age Group', 'Age category (1=18-24, ..., 13=80 or
       ↪ older)'),
    'BMI': ('Body Mass Index', 'Body Mass Index, kg/m$^2$'),
    'Income': ('Income Level', 'Income scale (1: <=10k, ..., 8:
       ↪  >75k)'),
    'Education': ('Education Level', 'Education scale (1=Never
       ↪ attended school, ..., 6=College)'),
    'Coefficient': ('Coefficient', None),
    'Standard Error': ('Std. Error', 'Standard error of the
       ↪ coefficient'),
    'P-value': ('P-value', None),
    'CI Lower': ('CI Lower', 'Lower bound of 95% confidence
       ↪ interval'),
    'CI Upper': ('CI Upper', 'Upper bound of 95% confidence
       ↪ interval'),
    'Intercept': ('Intercept', 'The intercept in the logistic
       ↪ regression model'),
    'PhysActivity:Fruits': ('PA * Fruits', 'Interaction term
       ↪ between Physical Activity and Fruit Consumption'),
    'PhysActivity:Veggies': ('PA * Veggies', 'Interaction term
       ↪ between Physical Activity and Vegetable Consumption')
       ↪ ,
    'Standardized_Age': ('Standardized Age', 'Age standardized
       ↪ for logistic regression'),
    'Standardized_BMI': ('Standardized BMI', 'BMI standardized
       ↪ for logistic regression'),
    'Standardized_Income': ('Standardized Income', 'Income
       ↪ standardized for logistic regression'),
    'Standardized_Education': ('Standardized Education', '
       ↪ Education level standardized for logistic regression'
       ↪ ),
}

# Process df_desc_stat
df_desc_stat = pd.read_pickle('df_desc_stat.pkl')
# Format values: Not Applicable
# Rename rows and columns
mapping = dict((k, v) for k, v in all_mapping.items() if
   ↪ is_str_in_df(df_desc_stat, k))
abbrs_to_names, glossary = split_mapping(mapping)
df_desc_stat.rename(columns=abbrs_to_names, index=
   ↪ abbrs_to_names, inplace=True)
df_to_latex(
    df_desc_stat, 'df_desc_stat_formatted',
    caption="Descriptive Statistics of Key Variables Related to
        ↪  Diabetes Risk",
    glossary=glossary
)

# Process df_logistic_reg
df_logistic_reg = pd.read_pickle(’df_logistic_reg.pkl’)
# Format values: Not Applicable
# Remove Intercept
df_logistic_reg.drop(index=’Intercept’, inplace=True)
# Prepare Confidence Intervals as tuples
df_logistic_reg[’CI’] = list(zip(df_logistic_reg[’CI Lower’],
    ↪ df_logistic_reg[’CI Upper’]))
# Rename rows and columns
mapping = dict((k, v) for k, v in all_mapping.items() if
    ↪ is_str_in_df(df_logistic_reg, k))
abbrs_to_names, glossary = split_mapping(mapping)
df_logistic_reg.rename(columns=abbrs_to_names, index=
    ↪ abbrs_to_names, inplace=True)
df_to_figure(
    df_logistic_reg, ’df_logistic_reg_formatted’,
    caption="Logistic Regression Analysis for Interaction
        ↪ Between Physical Activity and Dietary Habits on
        ↪ Diabetes Risk",
    note="The figure illustrates the interaction effects
        ↪ between physical activity and dietary habits on
        ↪ diabetes risk, omitting the intercept term.",
    glossary=glossary,
    kind=’bar’,
    y=[’Coefficient’],
    y_ci=[’CI’],  # a single column with (lower, upper) tuple
        ↪ values
    y_p_value=[’P-value’]
)
```

D.2 Provided Code

The code above is using the following provided functions:

```
def df_to_latex(df,
        filename: str, caption: str,
        note: str = None,
        glossary: Dict[Any, str] = None,
    ):
    """
    Saves a DataFrame ‘df‘ and creates a LaTeX table.
    ‘filename‘, ‘caption‘: as in ‘df.to_latex‘.
    ‘note‘: Note to be added below the table caption.
    `glossary`: Glossary for the table.
    """


def df_to_figure(
        df, filename: str, caption: str,
        note: str = None, glossary: Dict[Any, str] = None,
        x: Optional[str] = None, y: List[str] = None,
        kind: str = 'bar',
        logx: bool = False, logy: bool = False,
        y_ci: Optional[List[str]] = None,
        y_p_value: Optional[List[str]] = None,
        xlabel: str = None, ylabel: str = None,
    ):
    """
    Save a DataFrame `df` and create a LaTeX figure.
    Parameters, for LaTex embedding of the figure:
    `filename`: Filename for the figure.
    `caption`: Caption for the figure.
    `note`: Note to be added below the figure caption.
    `glossary`: Glossary for the figure.

    Parameters for df.plot():
    `x`: Column name for x-axis (index by default).
    `y`: List of m column names for y-axis (m=1 for single plot
       ↪ , m>1 for multiple plots).
    `kind`: only bar is allowed.
    `logx` / `logy` (bool): log scale for x/y axis.
    `xlabel`: Label for the x-axis.
    `ylabel`: Label for the y-axis.

    `y_ci`: Confidence intervals for errorbars.
        List of m column names indicating confidence intervals
           ↪ for each y column.
        Each element in these columns must be a Tuple[float,
           ↪ float], describing the lower and upper bounds of
           ↪ the CI.

     `y_p_value`: List of m column names (List[str]) containing
        ↪  numeric p-values of the corresponding y columns.
        ↪ These numeric values will be automatically converted
        ↪  by df_to_figure to stars ('***', '**', '*', 'ns')
        ↪ and plotted above the error bars.

    If provided, the length of `y_ci`, and `y_p_value` should
       ↪ be the same as of `y`.

    Example:
    Suppose, we have:
```

```
    df_lin_reg_longevity = pd.DataFrame({
        'adjusted_coef': [0.4, ...], 'adjusted_coef_ci':
            ↪ [(0.35, 0.47), ...], 'adjusted_coef_pval':
            ↪ [0.012, ...],
        'unadjusted_coef': [0.2, ...], 'unadjusted_coef_ci':
            ↪ [(0.16, 0.23), ...], 'unadjusted_coef_pval':
            ↪ [0.0001, ...],
    }, index=['var1', ...])

    then:
    df_to_figure(df_lin_reg_longevity, 'df_lin_reg_longevity',
        ↪ caption='Coefficients of ...', kind='bar',
        y=['adjusted_coef', 'unadjusted_coef'],
        y_ci=['adjusted_coef_ci', 'unadjusted_coef_ci'],
        y_p_value=['adjusted_coef_pval', 'unadjusted_coef_pval
            ↪ '])
    """


def is_str_in_df(df: pd.DataFrame, s: str):
    return any(s in level for level in getattr(df.index, '
        ↪ levels', [df.index]) + getattr(df.columns, 'levels',
        ↪ [df.columns]))

AbbrToNameDef = Dict[Any, Tuple[Optional[str], Optional[str]]]

def split_mapping(abbrs_to_names_and_definitions: AbbrToNameDef
    ↪ ):
    abbrs_to_names = {abbr: name for abbr, (name, definition)
        ↪ in abbrs_to_names_and_definitions.items() if name is
        ↪ not None}
    names_to_definitions = {name or abbr: definition for abbr,
        ↪ (name, definition) in abbrs_to_names_and_definitions.
        ↪ items() if definition is not None}
    return abbrs_to_names, names_to_definitions
```

D.3 Code Output

df_desc_stat_formatted.pkl

```
\begin{table}[h]
\caption{Descriptive Statistics of Key Variables Related to
   Diabetes Risk}
\label{table:df-desc-stat-formatted}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrr}
\toprule
 & mean & std \\
\midrule
\textbf{Physical Activity} & 0.7565 & 0.4292 \\
\textbf{Fruit Consumption} & 0.6343 & 0.4816 \\
\textbf{Vegetable Consumption} & 0.8114 & 0.3912 \\
\textbf{Diabetes Diagnosis} & 0.1393 & 0.3463 \\
\textbf{Age Group} & 8.032 & 3.054 \\
\textbf{Body Mass Index} & 28.38 & 6.609 \\
\textbf{Income Level} & 6.054 & 2.071 \\
\textbf{Education Level} & 5.05 & 0.9858 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Physical Activity}: Engaged in physical activity
   in the past 30 days, 1: Yes, 0: No
\item \textbf{Fruit Consumption}: Consumes one or more fruits
   daily, 1: Yes, 0: No
\item \textbf{Vegetable Consumption}: Consumes one or more
   vegetables daily, 1: Yes, 0: No
\item \textbf{Diabetes Diagnosis}: Diagnosis of diabetes, 1:
   Yes, 0: No
\item \textbf{Age Group}: Age category (1=18-24, ..., 13=80 or
   older)
\item \textbf{Body Mass Index}: Body Mass Index, kg/m$^2$
\item \textbf{Income Level}: Income scale (1: $<$=10k, ..., 8:
   $>$75k)
\item \textbf{Education Level}: Education scale (1=Never
   attended school, ..., 6=College)
\end{tablenotes}
\end{threeparttable}
\end{table}
```

df_logistic_reg_formatted.pkl

```
\begin{figure}[htbp]
\centering
\includegraphics{df_logistic_reg_formatted.png}
\caption{Logistic Regression Analysis for Interaction Between
   Physical Activity and Dietary Habits on Diabetes Risk
The figure illustrates the interaction effects between physical
     activity and dietary habits on diabetes risk, omitting the
     intercept term.
Physical Activity: Engaged in physical activity in the past 30
   days, 1: Yes, 0: No.
Fruit Consumption: Consumes one or more fruits daily, 1: Yes,
   0: No.
Vegetable Consumption: Consumes one or more vegetables daily,
    1: Yes, 0: No.
Std. Error: Standard error of the coefficient.
CI Lower: Lower bound of 95\% confidence interval.
CI Upper: Upper bound of 95\% confidence interval.
PA * Fruits: Interaction term between Physical Activity and
    Fruit Consumption.
PA * Veggies: Interaction term between Physical Activity and
    Vegetable Consumption.
Standardized Age: Age standardized for logistic regression.
Standardized BMI: BMI standardized for logistic regression.
Standardized Income: Income standardized for logistic
    regression.
Standardized Education: Education level standardized for
    logistic regression.
Significance: ns p $>$= 0.01, * p $<$ 0.01, ** p $<$ 0.001, ***
     p $<$ 0.0001.}
\label{figure:df-logistic-reg-formatted}
\end{figure}
% This latex figure presents "df_logistic_reg_formatted.png",
% which was created from the df:
%
% index,"Coefficient","Std. Error","P-value","CI Lower","CI
    Upper","CI"
% "Physical Activity",-0.1213,0.02734,9.13e-06,-0.1749,
    -0.06772,(-0.1749, -0.06772)
% "Fruit Consumption",-0.01567,0.02177,0.472,-0.05833,0.027,(
    -0.05833, 0.027)
% "PA * Fruits",-0.1718,0.02695,<1e-06,-0.2246,-0.1189,(
    -0.2246, -0.1189)
% "Vegetable Consumption",-0.06512,0.0235,0.00558,-0.1112,
    -0.01907,(-0.1112, -0.01907)
% "PA * Veggies",-0.05864,0.03056,0.055,-0.1185,0.001259,(
    -0.1185, 0.001259)
% "Standardized Age",0.6304,0.007265,<1e-06,0.6161,0.6446,(
    0.6161, 0.6446)
% "Standardized BMI",0.5443,0.005749,<1e-06,0.533,0.5555,(
    0.533, 0.5555)
% "Standardized Income",-0.2601,0.006498,<1e-06,-0.2729,
    -0.2474,(-0.2729, -0.2474)
% "Standardized Education",-0.09721,0.006513,<1e-06,-0.11,
    -0.08445,(-0.11, -0.08445)
%
% To create the figure, this df was plotted with the command:
%
% df.plot(kind='bar', y=['Coefficient'])
%
```

```
% Confidence intervals for y-values were then plotted based on
   column: [’CI’].
%
% P-values for y-values were taken from column: [’P-value’].
%
% These p-values were presented above the data points as stars
   (with significance threshold values indicated in the figure
    caption).
```

E Calculation Notes

- 0.7565 * 100 = 75.65

  percentage of participants who engaged in physical activity

- 0.6343 * 100 = 63.43

  percentage of participants who consume one or more fruits daily

- 0.8114 * 100 = 81.14

  percentage of participants who consume one or more vegetables daily

- 0.1393 * 100 = 13.93

  percentage of participants diagnosed with diabetes