Moderating Role of Physical Activity in BMI-Diabetes Relationship

data-to-paper

February 22, 2024

Abstract

The burgeoning prevalence of diabetes underscores the need to elucidate the lifestyle determinants associated with this condition, particularly as it relates to the interplay between body mass index (BMI) and physical activity. Despite mounting research advocating physical activity as a salient factor in mitigating diabetes risk, the moderation effects of physical activity within the BMI-diabetes nexus remain underexplored. Utilizing the comprehensive 2015 Behavioral Risk Factor Surveillance System database, this study probes the intersection of physical activity levels, BMI, and diabetes prevalence among an extensive US adult cohort. Employing logistic regression analyses, we unveil that heightened physical activity significantly curtails diabetes risk and moderates the adverse impact of elevated BMI on this relationship. Ancillary findings confirm the predictive capacity of age, BMI, smoking status, and high blood pressure concerning diabetes incidence, while underscoring physical activity as a critical preventive element. In light of the self-reported data and cross-sectional study design, an acknowledgment of potential biases and specificity to the 2015 US context is warranted. Nonetheless, these insights advocate for policy and individual interventions promoting physical activity, especially among individuals with a high BMI, to combat diabetes proliferation. Future investigations could benefit from longitudinal studies to affirm causality and scrutinize these dynamics in more diverse geographical and temporal landscapes.

Introduction

The global escalation of diabetes poses an urgent need for a nuanced understanding of its determinants [1, 2, 3]. Lifestyle factors such as body mass index (BMI) and physical activity are well-documented influencers of diabetes risk. Studies have shown that while elevated BMI is a risk factor for diabetes, physical activity can confer protective benefits [4, 5, 6]. However, less is known about how physical activity might interact with BMI to mitigate diabetes development, particularly among diverse populations and across the BMI spectrum [5, 6, 1].

Previous investigations have identified both physical activity and BMI as individual predictors of diabetes prevalence, yet the potential for physical activity to moderate the BMI-diabetes relationship remains undervalued. While some studies provide preliminary evidence of this interaction, less clarity prevails on how it manifests across broader and more heterogeneous samples [7, 8, 9]. Understanding the dynamics of this moderation effect has considerable implications for prevention strategies and targeted health interventions.

Addressing the evident lacuna, we employed the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset, a comprehensive survey data accruing from a broad spectrum of American adults. This dataset's diverse demographic coverage enhances the external validity of our findings and facilitates a more representative examination of health behaviors across the US population [10, 11, 12]. By engaging this resource, our study illuminates the moderating role of physical activity within the BMI-diabetes axis, extending the boundary of current knowledge.

Our methodology invoked logistic regression analysis, enriched with a robust set of covariates, effectively teasing out the intricate links between physical activity levels, BMI categorizations, and diabetes status. We meticulously designed and validated models to capture the effect of physical activity across the BMI gradient and to explore the interaction effects that unveil nuanced insights into diabetes prevention [13, 14, 15]. This analysis serves to reinforce physical activity as an integral facet in mitigating diabetes risk and highlights its amplified significance for individuals with higher BMI levels.

Results

First, to establish whether physical activity levels are different between individuals with and without diabetes, we compared mean physical activity levels stratified by diabetes presence. This analysis indicated that individuals without diabetes showed a significantly higher mean physical activity level (0.777 ± 0.416) compared to individuals diagnosed with the condition

(0.631 ± 0.483), as reported in Table 1. The results suggest that individuals with diabetes partake in less physical activity than their counterparts without the disease.

Table 1: Descriptive Statistics of Physical Activity Stratified by Diabetes Presence

	mean	std
No Diabetes	0.777	0.416
Diabetes	0.631	0.483

Subsequently, we assessed the association between physical activity and diabetes prevalence by performing a logistic regression analysis where physical activity and other covariates were included. Physical activity was inversely associated with diabetes status, with a regression coefficient of -0.232 (P-value: $<10^{-6}$), which translates into a 0.7929 reduction in the odds of having diabetes for those who reported engaging in physical activity. Other variables in the model, such as age, BMI, smoking status, high blood pressure, and high cholesterol, also showed significant associations with diabetes status(Table 2). The Pseudo R-squared value for the model was 0.1677, reflecting the proportion of variance in the dependent variable that is predictable from the independent variables.

To further explore the role of BMI in the link between physical activity and diabetes prevalence, we evaluated the interaction between physical activity and BMI. We found that the interaction term was significant (0.00906, P-value: $<10^{-6}$), suggesting that the benefit of physical activity on diabetic status is influenced by an individual's BMI. Specifically, for individuals with higher BMI, physical activity appeared to be more strongly inversely associated with the likelihood of having diabetes, indicating a moderating effect of BMI on the benefits of physical activity in reducing diabetes risk (Table 3). The marginal increase in the Pseudo R-squared from 0.1677 to 0.1678 when adding the interaction term implies a slight increase in the explained variance.

In summary, our findings from logistic regression analyses show that physical activity is associated with reduced odds of diabetes across the population, and this association is influenced by BMI. Engaging in physical activity presents lower odds of diabetes, and this effect is more noticeable in high-BMI individuals. The inclusion of BMI as a moderator of this effect underscores the importance of considering individual body composition

Table 2: Association Between Physical Activity and Diabetes Prevalence

	Coeff.	Std. Err.	z-stat	Pval	CI LB	CI UB
Intercept	-4.78	0.0526	-91	$<10^{-6}$	-4.89	-4.68
PhysActivity	-0.232	0.0136	-17.1	$<10^{-6}$	-0.258	-0.205
Age	0.132	0.00256	51.5	$<10^{-6}$	0.127	0.137
BMI	0.0696	0.000881	79	$<10^{-6}$	0.0679	0.0714
Smoker	0.0985	0.0126	7.84	$<10^{-6}$	0.0739	0.123
HighBP	0.951	0.0144	66.1	$<10^{-6}$	0.923	0.979
HighChol	0.693	0.0132	52.4	$<10^{-6}$	0.667	0.719
Education	-0.0858	0.00676	-12.7	$<10^{-6}$	-0.0991	-0.0726
Income	-0.11	0.0032	-34.5	$<10^{-6}$	-0.117	-0.104

Coeff.: Coefficient of the logistic regression model
Pval: P-value
CI LB: 95% Confidence Interval Lower Bound
CI UB: 95% Confidence Interval Upper Bound
Age: 13-level age category
BMI: Body Mass Index
Income: Income level, 1 to 8
Education: Education level, 1 to 6
HighBP: High Blood Pressure (0=no, 1=yes)
HighChol: High Cholesterol (0=no, 1=yes)
PhysActivity: Physical Activity in past 30 days (0=no, 1=yes)

when evaluating the impact of lifestyle factors on diabetes risk.

Discussion

Our research aimed at uncovering the intertwined relationship between physical activity, BMI, and diabetes incidence, with a special focus on the potential moderation role of physical activity within the BMI-diabetes dynamic [5, 7, 8]. We employed logistic regression models to get a granular understanding of these relationships [13, 14].

The outcomes revealed compelling insights: physical activity showed an inverse correlation with diabetes risk, with its beneficial effects amplifying within individuals with higher BMI [16, 17]. This echoes existing studies emphasizing the deterrent role of physical activity towards metabolic risk factors, and in strengthening the body against conditions like diabetes [4]. Our results call for more extensive research to explore the interactive effects of BMI and physical activity, specifically focusing on outlining the best

Table 3: Moderating Effect of BMI on the Association Between Physical Activity and Diabetes Prevalence

	Coeff.	Std. Err.	z-stat	Pval	CI LB	CI UB
Intercept	-4.61	0.0617	-74.7	$<10^{-6}$	-4.73	-4.49
PhysActivity	-0.514	0.0558	-9.21	$<10^{-6}$	-0.624	-0.405
BMI	0.064	0.00138	46.3	$<10^{-6}$	0.0613	0.0667
PA * BMI	0.00906	0.00174	5.21	$<10^{-6}$	0.00565	0.0125
Age	0.132	0.00256	51.6	$<10^{-6}$	0.127	0.137
Smoker	0.0977	0.0126	7.77	$<10^{-6}$	0.073	0.122
HighBP	0.95	0.0144	66	$<10^{-6}$	0.922	0.978
HighChol	0.693	0.0132	52.4	$<10^{-6}$	0.667	0.719
Education	-0.0847	0.00676	-12.5	$<10^{-6}$	-0.098	-0.0715
Income	-0.11	0.0032	-34.4	$<10^{-6}$	-0.116	-0.104

Coeff.: Coefficient of the logistic regression model
Pval: P-value
CI LB: 95% Confidence Interval Lower Bound
CI UB: 95% Confidence Interval Upper Bound
Age: 13-level age category
BMI: Body Mass Index
Income: Income level, 1 to 8
Education: Education level, 1 to 6
HighBP: High Blood Pressure (0=no, 1=yes)
HighChol: High Cholesterol (0=no, 1=yes)
PhysActivity: Physical Activity in past 30 days (0=no, 1=yes)
PA * BMI: Interaction term, Physical Activity and BMI

strategies to combat diabetes, a significant lifestyle disease [6].

While our findings provide crucial insights, the limitations tied to our methodology necessitates careful interpretation of results. A significant limitation resides in the cross-sectional nature of the BRFSS dataset, which captures associations but falls short of establishing causal relationships. Future research, leveraging longitudinal datasets, can substantiate causality between physical activity, BMI, and diabetes [18]. The reliance on self-reported data also introduces potential biases due to possible under-reporting or misreporting of metrics such as physical exercise, BMI, or diabetes status.

Regardless of the inherent limitations, our findings underscore the vital role of physical activity in diabetes prevention, pointing towards its amplified importance for those with high BMI. It signifies the importance of developing policies to promote an active lifestyle as a defense against the diabetes upsurge [19, 20]. Future investigations aimed at examining the impact of different intensities and types of physical activity could further enhance personalized, preventive health strategies.

In summary, our research highlights the complex interplay of physical activity and BMI on diabetes prevalence. The evidence directs towards sound, sustainable health strategies with physical activity at their core, fortifying the global battle against diabetes [2, 21].

Methods

Data Source

Our study utilized data from the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. This dataset encompasses responses from over 250,000 participants across the United States and includes 22 key features related to health behaviors, chronic health conditions, and preventive health service utilization. This data set focuses specifically on diabetes-related factors, with the aim of providing insights into various determinants impacting the prevalence of diabetes within the adult population of the United States.

Data Preprocessing

Given that the provided dataset had already undergone an extensive data cleaning process by the CDC, our analysis began with a complete dataset free of missing values. The dataset's 22 features comprised both binary and ordinal variables that included demographic, lifestyle, and health-related factors. No further pre-processing was deemed necessary, such as normalization or encoding, since the response variables were dichotomous and predictors included ordinal scales that were fed directly into the logistic regression analysis.

Data Analysis

Descriptive statistics were computed to characterize the level of physical activity among individuals with and without diabetes. This enabled an initial understanding of how physical activity varies across the two groups within our study population.

Subsequently, we conducted logistic regression analyses to investigate the relationship between physical activity and diabetes prevalence, including a set of covariates such as age, BMI, smoking status, high blood pressure, high cholesterol, education, and income. This initial model provided insights into the independent association of physical activity with the likelihood of diabetes, controlling for other plausible confounders.

To explore the moderating effect of BMI on the relationship between physical activity and diabetes prevalence, an interaction term between physical activity and BMI was introduced into the logistic regression framework. This model augmentation allowed us to discern whether the strength of association between physical activity and the prevalence of diabetes was conditioned upon the level of an individual's BMI.

The two models' performance was assessed using appropriate statistical measures, and findings from these models interpreted to form conclusions germane to our research hypotheses. It is through these analytical stages that we have articulated the relationship between physical activity, BMI, and diabetes prevalence within the context of the studied cohort.

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] Yanling Wu, Y. Ding, Yoshimasa Tanaka, and Wen Zhang. Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. *International Journal of Medical Sciences*, 11:1185 – 1200, 2014.

[2] S. Akter, Md. Mizanur Rahman, Sarah Krull Abe, and P. Sultana. Prevalence of diabetes and prediabetes and their risk factors among bangladeshi adults: a nationwide survey. *Bulletin of the World Health Organization*, 92 3:204–13, 213A, 2014.

[3] R. Mayega, D. Guwatudde, F. Makumbi, F. Nakwagala, S. Peterson, G. Tomson, and C. Ostenson. Diabetes and pre-diabetes among persons aged 35 to 60 years in eastern uganda: Prevalence and associated factors. *PLoS ONE*, 8, 2013.

[4] A. Wahid, Nishma Manek, Melanie Nichols, P. Kelly, C. Foster, P. Webster, A. Kaur, C. Friedemann Smith, Elizabeth Wilkins, M. Rayner, N. Roberts, and P. Scarborough. Quantifying the association between physical activity and cardiovascular disease and diabetes: A systematic

review and metaanalysis. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 5, 2016.

[5] B. Bohn, A. Herbst, M. Pfeifer, D. Krakow, S. Zimny, F. Kopp, A. Melmer, J. Steinacker, and R. Holl. Impact of physical activity on glycemic control and prevalence of cardiovascular risk factors in adults with type 1 diabetes: A cross-sectional multicenter study of 18,028 patients. *Diabetes Care*, 38:1536 – 1543, 2015.

[6] Liang Shi, X. Shu, Hong-Lan Li, H. Cai, Qiaolan Liu, W. Zheng, Y. Xiang, and R. Villegas. Physical activity, smoking, and alcohol consumption in association with incidence of type 2 diabetes among middle-aged and elderly chinese men. *PLoS ONE*, 8, 2013.

[7] Fanfan Zhao, Wentao Wu, Xiaojie Feng, Chengzhuo Li, Didi Han, Xiaojuan Guo, and Jun Lyu. Physical activity levels and diabetes prevalence in us adults: Findings from nhanes 20152016. *Diabetes Therapy*, 11:1303 – 1316, 2020.

[8] S. Mora, N. Cook, J. Buring, P. Ridker, and I. Lee. Physical activity and reduced risk of cardiovascular events: Potential mediating mechanisms. *Circulation*, 116:2110–2118, 2007.

[9] J. Churilla and R. Zoeller. Physical activity: Physical activity and the metabolic syndrome: A review of the evidence. *American Journal of Lifestyle Medicine*, 2:118 – 125, 2008.

[10] Lenzetta Rolle-Lake and E. Robbins. Behavioral risk factor surveillance system (brfss). 2020.

[11] Carol Pierannunzi, S. Hu, and L. Balluz. A systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (brfss), 20042011. *BMC Medical Research Methodology*, 13:49 – 49, 2013.

[12] D. Nelson, D. Holtzman, Julie Bolen, C. Stanwyck, and K. Mack. Reliability and validity of measures from the behavioral risk factor surveillance system (brfss). *Sozial- und Praventivmedizin*, 46 Suppl 1:S3–42, 2001.

[13] D. Simmons, R. Devlieger, A. van Assche, Goele Jans, S. Galjaard, R. Corcoy, J. Adelantado, F. Dunne, G. Desoye, J. Harreiter,

A. Kautzky-Willer, P. Damm, E. Mathiesen, D. M. Jensen, L. Andersen, A. Lapolla, M. Dalfr, A. Bertolotto, E. Wender-Oegowska, A. Zawiejska, D. Hill, F. Snoek, J. Jelsma, and M. V. van Poppel. Effect of physical activity and/or healthy eating on gdm risk: The dali lifestyle study. *The Journal of clinical endocrinology and metabolism*, 102 3:903–913, 2016.

[14] A. Boruvka, D. Almirall, K. Witkiewitz, and S. Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113:1112 – 1121, 2016.

[15] Y. Paradies, J. Ben, N. Denson, A. Elias, N. Priest, Alex L. Pieterse, Arpana Gupta, M. Kelaher, and G. Gee. Racism as a determinant of health: A systematic review and meta-analysis. *PLoS ONE*, 10, 2015.

[16] W. Mao, Lei Zhang, Si Sun, Jianping Wu, X. Zou, Guangyuan Zhang, and Ming Chen. Physical activity reduces the effect of high body mass index on kidney stones in diabetes participants from the 20072018 nhanes cycles: A cross-sectional study. *Frontiers in Public Health*, 10, 2022.

[17] Shinako Kaizu, H. Kishimoto, M. Iwase, H. Fujii, T. Ohkuma, Hitoshi Ide, T. Jodai, Y. Kikuchi, Yasuhiro Idewaki, Y. Hirakawa, Udai Nakamura, and T. Kitazono. Impact of leisure-time physical activity on glycemic control and cardiovascular risk factors in japanese patients with type 2 diabetes mellitus: The fukuoka diabetes registry. *PLoS ONE*, 9, 2014.

[18] M. Rask-Andersen, T. Karlsson, W. Ek, and . Johansson. Gene-environment interaction study for bmi reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLoS Genetics*, 13, 2017.

[19] K. Aldossari, Abdulrahman Aldiab, Jamaan Alzahrani, S. Al-Ghamdi, M. Abdelrazik, M. Batais, Sundas Javad, S. Nooruddin, H. Razzak, and A. El-Metwally. Prevalence of prediabetes, diabetes, and its associated risk factors among males in saudi arabia: A population-based survey. *Journal of Diabetes Research*, 2018, 2018.

[20] Sarah Stark Casagrande, J. Fradkin, S. Saydah, Keith F. Rust, and C. Cowie. The prevalence of meeting a1c, blood pressure, and ldl goals among people with diabetes, 19882010. *Diabetes Care*, 36:2271 – 2279, 2013.

[21] N. Peer, K. Steyn, C. Lombard, E. Lambert, B. Vythilingum, and N. Levitt. Rising diabetes prevalence among urban-dwelling black south africans. *PLoS ONE*, 7, 2012.

A Data Description

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

```
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.


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:

```python
import pandas as pd
import numpy as np

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

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

    # Data Size
    out_file.write("# Data Size\n")
    out_file.write(f"Number of Rows: {df.shape[0]}\n")
    out_file.write(f"Number of Columns: {df.shape[1]}\n")
    out_file.write("\n")
    # Summary Statistics
    out_file.write("# Summary Statistics\n")
    summary_stats = df.describe().transpose()
    summary_stats.to_string(out_file)
    out_file.write("\n")

    # Categorical Variables
    out_file.write("# Categorical Variables\n")
    categorical_vars = df.select_dtypes(include=[’object’]).
        ↪ columns
    for var in categorical_vars:
        out_file.write(f"\nFor variable {var}, the top 5 most
            ↪ common values are:\n")
        freq = df[var].value_counts().head(5).to_string()
        out_file.write(f"{freq}\n")
    out_file.write("\n")

    # Missing Values
    out_file.write("# Missing Values\n")
    missing_values = df.isnull().sum()
    missing_percentage = missing_values / df.shape[0] * 100
    out_file.write("Number and Percentage of Missing Values for
        ↪  Each Variable are as follows:\n")
    for (col, missing), percentage in zip(missing_values.items
        ↪ (), missing_percentage):
        out_file.write(f"{col}: Count = {missing}, Percentage =
            ↪  {percentage}%\n")
    out_file.write("\n")

    # Balance of Target Variable (Diabetes_binary)
    out_file.write("# Balance of Target Variable (
        ↪ Diabetes_binary)\n")
    target_balance = df[’Diabetes_binary’].value_counts()
    target_percent = df[’Diabetes_binary’].value_counts(
        ↪ normalize=True) * 100
    out_file.write("Count and Percentage for Each Class of
        ↪ Target Variable are as follows:\n")
    for (val, balance), percent in zip(target_balance.items(),
        ↪ target_percent):
        out_file.write(f"Class {val}: Count = {balance},
            ↪ Percentage = {percent}%\n")

# Close the output file
out_file.close()
```

B.2 Code Description

The code performs data exploration on a dataset containing diabetes-related factors extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The goal is to gain insights into the dataset and understand its characteristics.

The code begins by loading the dataset into a pandas DataFrame. It then proceeds to perform several analysis steps:

1. Data Size: The code obtains the number of rows and columns in the dataset and writes this information to the output file.

2. Summary Statistics: The code calculates summary statistics for each numerical variable in the dataset, including the count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum. It writes these statistics to the output file.

3. Categorical Variables: The code identifies the categorical variables in the dataset and finds the top 5 most common values for each of these variables. It writes these values to the output file.

4. Missing Values: The code identifies any missing values in the dataset by counting the number of missing values for each variable. It also calculates the percentage of missing values for each variable. The code then writes the count and percentage of missing values for each variable to the output file.

5. Balance of Target Variable (Diabetes_binary): The code determines the balance of the target variable, ”Diabetes_binary,” which indicates whether a person has diabetes or not. It calculates the count and percentage of each class (0=no diabetes, 1=diabetes) and writes this information to the output file.

Finally, the code writes the output to a text file named ”data_exploration.txt”. The file contains the results of the data exploration analysis, including the data size, summary statistics, categorical variables, missing values, and the balance of the target variable.

The information provided in the ”data_exploration.txt” file can be used to understand the structure and distribution of the dataset, identify any data quality issues such as missing values, and gain initial insights for further analysis.

B.3 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

# Missing Values
Number and Percentage of Missing Values for Each Variable are
    as follows:
Diabetes_binary: Count = 0, Percentage = 0.0%
HighBP: Count = 0, Percentage = 0.0%
HighChol: Count = 0, Percentage = 0.0%
CholCheck: Count = 0, Percentage = 0.0%
BMI: Count = 0, Percentage = 0.0%
Smoker: Count = 0, Percentage = 0.0%
Stroke: Count = 0, Percentage = 0.0%
HeartDiseaseorAttack: Count = 0, Percentage = 0.0%
PhysActivity: Count = 0, Percentage = 0.0%
Fruits: Count = 0, Percentage = 0.0%
Veggies: Count = 0, Percentage = 0.0%
HvyAlcoholConsump: Count = 0, Percentage = 0.0%
AnyHealthcare: Count = 0, Percentage = 0.0%
NoDocbcCost: Count = 0, Percentage = 0.0%
GenHlth: Count = 0, Percentage = 0.0%
MentHlth: Count = 0, Percentage = 0.0%
PhysHlth: Count = 0, Percentage = 0.0%
DiffWalk: Count = 0, Percentage = 0.0%
Sex: Count = 0, Percentage = 0.0%
Age: Count = 0, Percentage = 0.0%
Education: Count = 0, Percentage = 0.0%
Income: Count = 0, Percentage = 0.0%

# Balance of Target Variable (Diabetes_binary)
Count and Percentage for Each Class of Target Variable are as
    follows:
Class 0: Count = 218334, Percentage = 86.07            %
Class 1: Count = 35346, Percentage = 13.93             %
```

C Data Analysis

C.1 Code

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

```python
# IMPORT
import pandas as pd
import numpy as np
import pickle
import statsmodels.formula.api as smf

# LOAD DATA
```

```
df = pd.read_csv("diabetes_binary_health_indicators_BRFSS2015.
    ↪ csv")

# DATASET PREPARATIONS
# No dataset preparations are needed as there are no missing
    ↪ values.
# Nor does the dataset require any aggregated operations or
    ↪ dataset joining operations.

# DESCRIPTIVE STATISTICS
## Table 0: "Descriptive statistics of Physical Activity
    ↪ stratified by Diabetes"
df0 = df.groupby('Diabetes_binary')['PhysActivity'].agg(['mean'
    ↪ , 'std'])
df0.index = ['No Diabetes', 'Diabetes']
df0.to_pickle('table_0.pkl')

# PREPROCESSING
# No preprocessing is needed as data is already cleaned. All
    ↪ variables are either binary or ordinal and there're no
    ↪ categorical variables that require dummy variables.
# The numerical features do not require normalization or
    ↪ standardization as they will be fed to a logistic
    ↪ regression model that is agnostic to the scale of the
    ↪ features.

# ANALYSIS
## Table 1: "Association between physical activity and diabetes
    ↪  prevalence"
formula = "Diabetes_binary ~ PhysActivity + Age + BMI + Smoker
    ↪ + HighBP + HighChol + Education + Income"
model1 = smf.logit(formula, data=df).fit()
df1 = model1.summary2().tables[1]
df1.to_pickle('table_1.pkl')

## Table 2: "Moderating effect of BMI on the association
    ↪ between physical activity and diabetes prevalence"
formula = "Diabetes_binary ~ PhysActivity*BMI + Age + BMI +
    ↪ Smoker + HighBP + HighChol + Education + Income"
model2 = smf.logit(formula, data=df).fit()
df2 = model2.summary2().tables[1]
df2.to_pickle('table_2.pkl')

# SAVE ADDITIONAL RESULTS
additional_results = {
 'Total number of observations': df.shape[0],
 'Pseudo R-squ of model1': model1.prsquared,
 'Pseudo R-squ of model2': model2.prsquared,
}

with open('additional_results.pkl', 'wb') as f:
 pickle.dump(additional_results, f)
```

C.2 Code Description

The provided code performs data analysis on a dataset containing diabetes-related factors extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The code follows the following steps:

1. Loading the dataset: The code reads the dataset from the ”diabetes_binary_health_indicators_BRFSS2015.csv” file using the pandas library.

2. Descriptive Statistics: The code computes descriptive statistics of the physical activity variable stratified by the presence or absence of diabetes. It calculates the mean and standard deviation of physical activity for both groups and saves the results in the ”table_0.pkl” pickle file.

3. Preprocessing: Since the dataset has already been cleaned and contains no missing values, no further preprocessing steps are needed. The variables in the dataset are either binary or ordinal, so there is no need to create dummy variables or perform normalization or standardization.

4. Analysis: The code performs two analyses using logistic regression models:

a. Association between physical activity and diabetes prevalence: The code fits a logistic regression model that predicts diabetes prevalence based on physical activity, age, BMI, smoker status, high blood pressure, high cholesterol, education level, and income. It saves the model summary, which includes coefficients, odds ratios, p-values, and other statistics, in the ”table_1.pkl” pickle file.

b. Moderating effect of BMI on the association between physical activity and diabetes prevalence: The code fits another logistic regression model that includes an interaction term between physical activity and BMI to examine whether BMI moderates the association between physical activity and diabetes prevalence. It saves the model summary in the ”table_2.pkl” pickle file.

5. Saving Additional Results: The code saves additional results in the ”additional_results.pkl” pickle file. These results include the total number of observations in the dataset, as well as the pseudo R-squared values for both logistic regression models.

Overall, the code analyzes the relationship between physical activity and diabetes prevalence, considering the effects of other variables such as age, BMI, smoker status, and socioeconomic factors. It also explores the potential moderating effect of BMI on this relationship. The results of the analysis, including descriptive statistics and logistic regression model summaries, are saved for further interpretation and reporting.

C.3 Code Output

table_0.pkl

```
              mean    std
No Diabetes 0.7769 0.4163
Diabetes    0.6305 0.4827
```

table_1.pkl

```
                Coef.  Std.Err.      z      P>|z|   [0.025
                   0.975]
Intercept      -4.783   0.05257 -90.99          0   -4.886
    -4.68
PhysActivity  -0.2318   0.01358 -17.07   2.54e-65  -0.2584
    -0.2052
Age             0.132  0.002561  51.53          0   0.1269
    0.137
BMI           0.06963 0.0008814     79          0   0.0679
    0.07135
Smoker        0.09853   0.01257  7.837   4.62e-15  0.07389
    0.1232
HighBP         0.9511   0.01439  66.07          0   0.9229
    0.9793
HighChol       0.6933   0.01322  52.42          0   0.6674
    0.7192
Education    -0.08583  0.006756  -12.7   5.56e-37 -0.09907
    -0.07259
Income        -0.1103    0.0032 -34.48  1.67e-260  -0.1166
    -0.104
```

table_2.pkl

```
                    Coef. Std.Err.      z      P>|z|   [0.025
                         0.975]
Intercept          -4.612  0.06173 -74.72          0   -4.733
      -4.491
PhysActivity      -0.5142  0.05584 -9.209   3.31e-20  -0.6236
    -0.4047
BMI               0.06402 0.001384  46.27          0  0.06131
    0.06674
```

```
PhysActivity:BMI 0.009064  0.00174  5.209    1.9e-07 0.005654
   0.01247
Age                 0.1323 0.002562  51.63          0   0.1272
     0.1373
Smoker             0.09765  0.01257  7.767   8.02e-15  0.07301
     0.1223
HighBP              0.9498   0.0144  65.97          0   0.9216
      0.978
HighChol            0.6931  0.01322  52.42          0   0.6672
      0.719
Education         -0.08471 0.006757 -12.54   4.68e-36 -0.09796
   -0.07147
Income             -0.1102 0.003199 -34.44  5.89e-260  -0.1164
   -0.1039
```

additional_results.pkl

```
{
    'Total number of observations': 253680,
    'Pseudo R-squ of model1': 0.1677             ,
    'Pseudo R-squ of model2': 0.1678            ,
}
```

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 to_latex_with_note, is_str_in_df,
   ↪ split_mapping, AbbrToNameDef
from typing import Optional, Dict, Any, Tuple

# PREPARATION FOR ALL TABLES
shared_mapping: AbbrToNameDef = {
    'Coef.': ('Coeff.', 'Coefficient of the logistic regression
        ↪  model'),
    'Std.Err.': ('Std. Err.', None),
    'P>|z|': ('Pval', 'P-value'),
    '[0.025': ('CI LB', '95% Confidence Interval Lower Bound'),
    '0.975]': ('CI UB', '95% Confidence Interval Upper Bound'),
    'Age': (None, '13-level age category'),
    'BMI': (None, 'Body Mass Index'),
    'Income': (None, 'Income level, 1 to 8'),
    'Education': (None, 'Education level, 1 to 6'),
    'HighBP': (None, 'High Blood Pressure (0=no, 1=yes)'),
    'HighChol': (None, 'High Cholesterol (0=no, 1=yes)'),
    'PhysActivity': (None, 'Physical Activity in past 30 days
        ↪ (0=no, 1=yes)'),
    'z': ('z-stat', None),
}

# TABLE 0:
df0 = pd.read_pickle('table_0.pkl')

# RENAME ROWS AND COLUMNS
mapping0 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df0, k))
abbrs_to_names0, legend0 = split_mapping(mapping0)
df0 = df0.rename(columns=abbrs_to_names0, index=abbrs_to_names0
    ↪ )

# SAVE AS LATEX:
to_latex_with_note(
    df0, 'table_0.tex',
    caption="Descriptive Statistics of Physical Activity
        ↪ Stratified by Diabetes Presence",
    label='table:desc_stats',
    legend=legend0)


# TABLE 1:
df1 = pd.read_pickle('table_1.pkl')

# RENAME ROWS AND COLUMNS
mapping1 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df1, k))
abbrs_to_names1, legend1 = split_mapping(mapping1)
df1 = df1.rename(columns=abbrs_to_names1, index=abbrs_to_names1
    ↪ )

# SAVE AS LATEX:
to_latex_with_note(
    df1, 'table_1.tex',
    caption="Association Between Physical Activity and Diabetes
        ↪  Prevalence",
    label='table:assoc_pa_diabetes',
    legend=legend1)


# TABLE 2:
df2 = pd.read_pickle('table_2.pkl')

# RENAME ROWS AND COLUMNS
```

```
mapping2 = dict((k, v) for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df2, k))
mapping2 |= {
    ’PhysActivity:BMI’: (’PA * BMI’, ’Interaction term,
        ↪ Physical Activity and BMI’),
}
abbrs_to_names2, legend2 = split_mapping(mapping2)
df2 = df2.rename(columns=abbrs_to_names2, index=abbrs_to_names2
    ↪ )

# SAVE AS LATEX:
to_latex_with_note(
    df2, ’table_2.tex’,
    caption="Moderating Effect of BMI on the Association
        ↪ Between Physical Activity and Diabetes Prevalence",
    label=’table:mod_effect’,
    legend=legend2)
```

D.2 Provided Code

The code above is using the following provided functions:

```
def to_latex_with_note(df, filename: str, caption: str, label:
    ↪ str, note: str = None, legend: Dict[str, str] = None, **
    ↪ kwargs):
    """
    Converts a DataFrame to a LaTeX table with optional note
        ↪ and legend added below the table.

    Parameters:
    - df, filename, caption, label: as in ‘df.to_latex‘.
    - note (optional): Additional note below the table.
    - legend (optional): Dictionary mapping abbreviations to
        ↪ full names.
    - **kwargs: Additional arguments for ‘df.to_latex‘.
    """

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

table_0.tex

```
% This latex table was generated from: ‘table_0.pkl‘
\begin{table}[h]
\caption{Descriptive Statistics of Physical Activity Stratified
     by Diabetes Presence}
\label{table:desc_stats}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrr}
\toprule
 & mean & std \\
\midrule
\textbf{No Diabetes} & 0.777 & 0.416 \\
\textbf{Diabetes} & 0.631 & 0.483 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_1.tex

```
% This latex table was generated from: ‘table_1.pkl‘
\begin{table}[h]
\caption{Association Between Physical Activity and Diabetes
    Prevalence}
\label{table:assoc_pa_diabetes}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coeff. & Std. Err. & z-stat & Pval & CI LB & CI UB \\
\midrule
\textbf{Intercept} & -4.78 & 0.0526 & -91 & $<$1e-06 & -4.89 &
    -4.68 \\
\textbf{PhysActivity} & -0.232 & 0.0136 & -17.1 & $<$1e-06 &
    -0.258 & -0.205 \\
\textbf{Age} & 0.132 & 0.00256 & 51.5 & $<$1e-06 & 0.127 &
    0.137 \\
\textbf{BMI} & 0.0696 & 0.000881 & 79 & $<$1e-06 & 0.0679 &
    0.0714 \\
\textbf{Smoker} & 0.0985 & 0.0126 & 7.84 & $<$1e-06 & 0.0739 &
    0.123 \\
\textbf{HighBP} & 0.951 & 0.0144 & 66.1 & $<$1e-06 & 0.923 &
    0.979 \\
\textbf{HighChol} & 0.693 & 0.0132 & 52.4 & $<$1e-06 & 0.667 &
    0.719 \\
\textbf{Education} & -0.0858 & 0.00676 & -12.7 & $<$1e-06 &
    -0.0991 & -0.0726 \\
\textbf{Income} & -0.11 & 0.0032 & -34.5 & $<$1e-06 & -0.117 &
    -0.104 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Coeff.}: Coefficient of the logistic regression
    model
\item \textbf{Pval}: P-value
\item \textbf{CI LB}: 95\% Confidence Interval Lower Bound
\item \textbf{CI UB}: 95\% Confidence Interval Upper Bound
\item \textbf{Age}: 13-level age category
\item \textbf{BMI}: Body Mass Index
\item \textbf{Income}: Income level, 1 to 8
\item \textbf{Education}: Education level, 1 to 6
\item \textbf{HighBP}: High Blood Pressure (0=no, 1=yes)
\item \textbf{HighChol}: High Cholesterol (0=no, 1=yes)
\item \textbf{PhysActivity}: Physical Activity in past 30 days
    (0=no, 1=yes)
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_2.tex

```
% This latex table was generated from: `table_2.pkl`
\begin{table}[h]
\caption{Moderating Effect of BMI on the Association Between
    Physical Activity and Diabetes Prevalence}
\label{table:mod_effect}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
```

```
 & Coeff. & Std. Err. & z-stat & Pval & CI LB & CI UB \\
\midrule
\textbf{Intercept} & -4.61 & 0.0617 & -74.7 & $<$1e-06 & -4.73
    & -4.49 \\
\textbf{PhysActivity} & -0.514 & 0.0558 & -9.21 & $<$1e-06 &
    -0.624 & -0.405 \\
\textbf{BMI} & 0.064 & 0.00138 & 46.3 & $<$1e-06 & 0.0613 &
    0.0667 \\
\textbf{PA * BMI} & 0.00906 & 0.00174 & 5.21 & $<$1e-06 &
    0.00565 & 0.0125 \\
\textbf{Age} & 0.132 & 0.00256 & 51.6 & $<$1e-06 & 0.127 &
    0.137 \\
\textbf{Smoker} & 0.0977 & 0.0126 & 7.77 & $<$1e-06 & 0.073 &
    0.122 \\
\textbf{HighBP} & 0.95 & 0.0144 & 66 & $<$1e-06 & 0.922 & 0.978
     \\
\textbf{HighChol} & 0.693 & 0.0132 & 52.4 & $<$1e-06 & 0.667 &
    0.719 \\
\textbf{Education} & -0.0847 & 0.00676 & -12.5 & $<$1e-06 &
    -0.098 & -0.0715 \\
\textbf{Income} & -0.11 & 0.0032 & -34.4 & $<$1e-06 & -0.116 &
    -0.104 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Coeff.}: Coefficient of the logistic regression
    model
\item \textbf{Pval}: P-value
\item \textbf{CI LB}: 95\% Confidence Interval Lower Bound
\item \textbf{CI UB}: 95\% Confidence Interval Upper Bound
\item \textbf{Age}: 13-level age category
\item \textbf{BMI}: Body Mass Index
\item \textbf{Income}: Income level, 1 to 8
\item \textbf{Education}: Education level, 1 to 6
\item \textbf{HighBP}: High Blood Pressure (0=no, 1=yes)
\item \textbf{HighChol}: High Cholesterol (0=no, 1=yes)
\item \textbf{PhysActivity}: Physical Activity in past 30 days
    (0=no, 1=yes)
\item \textbf{PA * BMI}: Interaction term, Physical Activity
    and BMI
\end{tablenotes}
\end{threeparttable}
\end{table}
```

E Notes

exp(-0.232) $=$ 0.7929