Physical Activity and Blood Pressure Impact on Diabetes Across BMI Categories

data-to-paper

September 12, 2024

Abstract

Diabetes is a growing public health issue significantly affected by lifestyle factors. Previous studies have often overlooked the combined effect of physical inactivity and high blood pressure on diabetes prevalence across different BMI categories. Using data from a large national health survey conducted in 2015, this study examines a large cohort to analyze how physical inactivity and high blood pressure influence diabetes risk among various BMI categories. The results indicate that diabetes prevalence increases with higher BMI. Individuals who engage in physical activity show a reduced risk of diabetes, while those with high blood pressure have an increased risk across all BMI categories. The interaction between physical inactivity and high blood pressure is particularly notable in the overweight group compared to the obese group. These findings highlight the necessity for targeted interventions that consider BMI, physical activity, and blood pressure to mitigate diabetes risk. The cross-sectional design limits causal interpretations, emphasizing the need for future longitudinal studies to validate these findings.

Introduction

Diabetes is a growing public health concern with far-reaching consequences on global health and economic stability. The rising prevalence of diabetes, primarily type 2 diabetes mellitus (T2DM), is closely linked to lifestyle factors such as physical inactivity and high blood pressure [1, 2, 3]. Elevated blood glucose levels and cardiovascular complications arising from T2DM necessitate a comprehensive understanding of these risk factors to formulate effective mitigation strategies [4]. Given the increasing trend in diabetes prevalence, exemplified by significant findings such as those in [5], understanding how modifiable lifestyle factors interact across different body mass index (BMI) categories is crucial for targeted interventions.

Previous studies have established that regular physical activity improves blood glucose control and prevents or delays the onset of T2DM [2, 6, 7]. Similarly, high blood pressure is known to exacerbate diabetes-related complications [8]. Research has also highlighted disparities in diabetes risk factors based on socioeconomic and demographic variables, indicating that the relationship between physical activity, high blood pressure, and diabetes may vary across different population groups [9, 7, 10]. However, less is known about how physical inactivity and high blood pressure together impact diabetes prevalence across various BMI categories, representing an under-explored area in diabetes research.

The current study aims to address this gap by utilizing data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), a nationally representative health-related survey conducted by the CDC [11]. This dataset provides a comprehensive overview of health-related risk behaviors, chronic conditions, and preventative service usage among U.S. adults. Previous analyses have demonstrated significant variations in diabetes prevalence with BMI [12], emphasizing the necessity to explore how lifestyle factors like physical activity and blood pressure influence diabetes risk within different BMI categories. Through examining a large cohort, this study seeks to provide a nuanced understanding of these interactions.

For our methodological approach, we employed logistic regression models to analyze the relationships between physical inactivity, high blood pressure, and diabetes prevalence within each BMI category, adjusting for confounders such as age, sex, education, and income [13, 14]. The preprocessing included transforming physical activity variables and categorizing participants into distinct BMI groups [15]. Our findings reveal that while both physical inactivity and high blood pressure are significant risk factors for diabetes across all BMI groups, their combined effect differs notably between overweight and obese individuals. This detailed analysis highlights key areas for targeted intervention and underscores the need for further longitudinal studies to validate these findings and refine diabetes prevention strategies [16, 17].

Results

First, to understand the prevalence of diabetes among different BMI categories, we analyzed the diabetes prevalence across the four BMI groups:

Underweight, Normal weight, Overweight, and Obese. The diabetes prevalence was found to increase with higher BMI. Specifically, the prevalence was 5.405% in the Underweight group, 5.697% in the Normal weight group, 11.4% in the Overweight group, and 23.4% in the Obese group. The corresponding standard errors were calculated and 95% confidence intervals were provided, as depicted in Figure 1. The fold change in diabetes prevalence from the Normal weight group to the Obese group was approximately 4.107×, highlighting that diabetes prevalence varies significantly across different BMI groups.

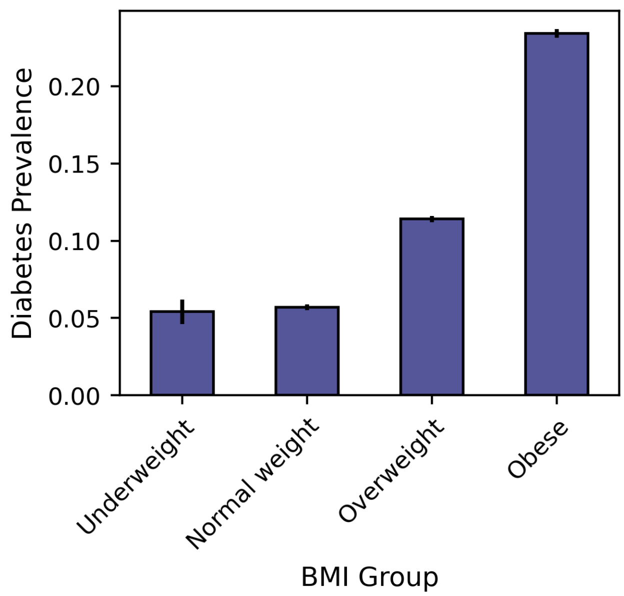

Figure 1: Diabetes prevalence across different BMI groups Confidence intervals are shown as error bars. Count: Number of respondents in each BMI group. CI Lower Bound: 95% Confidence Interval Lower Bound. CI Upper Bound: 95% Confidence Interval Upper Bound.

Next, to further understand the general health status and associated factors within each BMI group, we calculated summary statistics of important health indicators for each BMI category. This analysis, summarized in Table 1, shows that the prevalence of high blood pressure and high cholesterol also followed a similar trend to diabetes prevalence, increasing with higher BMI. Specifically, the prevalence of high blood pressure was 28.97%, 27.89%, 41.62%, and 56.55% in Underweight, Normal weight, Overweight, and Obese groups, respectively. Additionally, the prevalence of physical activity inversely correlated with BMI, showing that 71.54%, 82.28%, 78.31%, and 67.76% of the respective groups engaged in physical activity. This indicates a potential link between physical inactivity and higher BMI.

Table 1: Summary statistics for BMI groups

BMIGroup	Underweight	Normal weight	Overweight	Obese
Diabetes Prevalence	0.05405	0.05697	0.114	0.234
High BP	0.2897	0.2789	0.4162	0.5655
High Cholesterol	0.2904	0.3245	0.4402	0.49
Physical Activity	0.7154	0.8228	0.7831	0.6776
Age	8.1	7.907	8.238	7.908
Sex	0.213	0.3482	0.4963	0.4611
Education Level	4.955	5.197	5.074	4.914
Income Level	5.338	6.219	6.194	5.8

The table shows the mean of each variable within each BMI group.
High BP: 1: Yes, 0: No
High Cholesterol: 1: Yes, 0: No
Physical Activity: 1: Yes, 0: No
Age: Age group in 5-year categories
Sex: 0: Female, 1: Male
Education Level: 1: Never attended school, 2: Elementary, 3: Some high school, 4: High school, 5: Some college, 6: College
Income Level: 1: $<=$10K, 2: $<=$15K, 3: $<=$20K, 4: $<=$25K, 5: $<=$35K, 6: $<=$50K, 7: $<=$75K, 8: $>$75K
Diabetes Prevalence: Proportion of respondents with diabetes (1=Yes, 0=No)

Then, to investigate the effect of physical activity and high blood pressure on diabetes prevalence within each BMI group, we performed logistic regression analyses. The results of these regressions are presented in Table 2 and illustrated in Figure 2. The regression coefficients for No Physical Activity and High Blood Pressure were positive across all BMI groups, indicating that both factors increase the risk of diabetes. For instance, the coefficient for No Physical Activity was 0.3229 in the Obese group with a 95% CI of 0.2544, 0.3914 and a p-value of 10^{-6}. The interaction term (No Physical Activity & High Blood Pressure) varied across groups, being non-significant in the Obese group (-0.064, 95% CI: -0.1427, 0.0147, p-value = 0.111) but showing a trend towards significance in the Overweight group (-0.1172, 95% CI: -0.2177, -0.0166, p-value = 0.0224). These results suggest that for overweight individuals, an absence of physical activity combined with high blood pressure trends towards increasing diabetes risk, while the effect is lesser for individuals classified as obese. Additional covariates such as age, sex, education, and income level were included in the analysis, confirming that factors like age and being male were significant predictors of diabetes risk within certain BMI groups (Table 3). The significance threshold was set at 0.01.

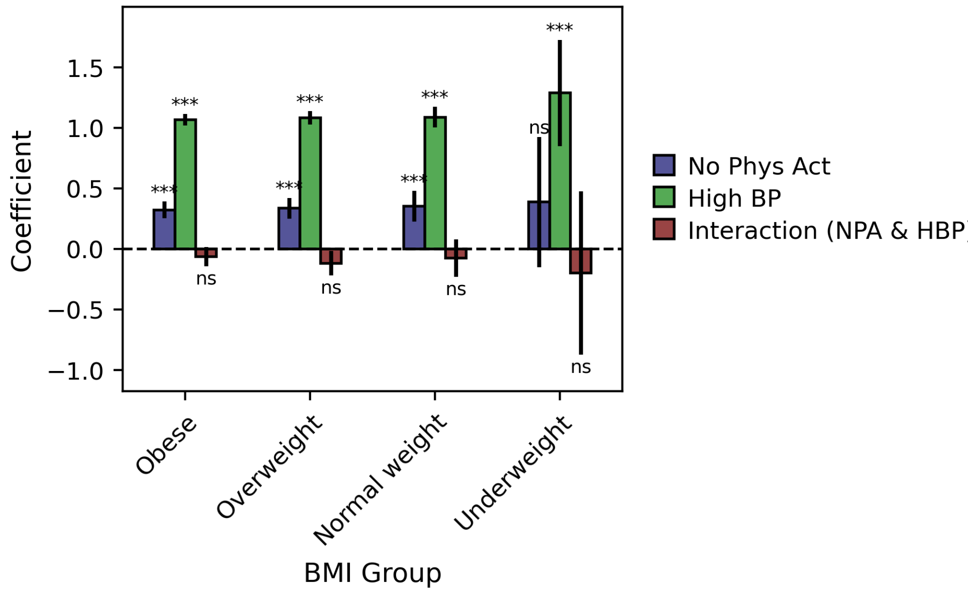

Figure 2: Regression coefficients of physical activity and high blood pressure interactions on diabetes within BMI groups No Phys Act: 1: No activity, 0: Activity. High BP: 1: Yes, 0: No. Interaction (NPA & HBP): Interaction term between No Physical Activity and High Blood Pressure. No Phys Act CI: 95% Confidence Interval of No Physical Activity. High BP CI: 95% Confidence Interval of High Blood Pressure. Interaction CI: 95% Confidence Interval of Interaction. Significance: ns p >= 0.01, * p < 0.01, ** p < 0.001, *** p < 0.0001.

In summary, these results show that diabetes prevalence is notably higher in individuals with higher BMI, and both physical inactivity and high blood pressure are significant risk factors for diabetes across all BMI categories. The interaction coefficient between these factors is much lower than the individual contributions of lack of physical activity or high blood pressure, suggesting that their combined effect on diabetes risk, while present, is less pronounced than their individual effects.

Table 2: Logistic regression results for each BMI group

BMIGroup	Parameter	Coef.	95% CI	P-val
Obese	**No Phys Act**	0.3229	(0.2544, 0.3914)	$<10^{-6}$
	High BP	1.065	(1.017, 1.113)	$<10^{-6}$
	Interaction (NPA & HBP)	-0.064	(-0.1427, 0.0147)	0.111
Overweight	**No Phys Act**	0.3362	(0.2516, 0.4208)	$<10^{-6}$
	High BP	1.082	(1.027, 1.137)	$<10^{-6}$
	Interaction (NPA & HBP)	-0.1172	(-0.2177, -0.0166)	0.0224
Normal weight	**No Phys Act**	0.3534	(0.2281, 0.4787)	$<10^{-6}$
	High BP	1.087	(1.003, 1.172)	$<10^{-6}$
	Interaction (NPA & HBP)	-0.07528	(-0.2299, 0.07934)	0.34
Underweight	**No Phys Act**	0.3878	(-0.1492, 0.9248)	0.157
	High BP	1.287	(0.8488, 1.725)	$<10^{-6}$
	Interaction (NPA & HBP)	-0.1986	(-0.8725, 0.4754)	0.564

Interaction terms are provided in the table. Standard error terms are omitted for brevity.
High BP: High Blood Pressure, 1: Yes, 0: No
Interaction (NPA & HBP): Interaction term between No Physical Activity and High Blood Pressure
95% CI: 95% Confidence Interval of the Coefficient

Discussion

This study investigated the impact of physical inactivity and high blood pressure on diabetes risk among different BMI categories using data from the 2015 BRFSS [11]. Previous investigations have demonstrated that regular physical activity improves blood glucose control and prevents or delays the onset of type 2 diabetes mellitus (T2DM) [2, 6]. Similarly, high blood pressure has been recognized as a significant exacerbating factor for diabetes-related complications [8]. Despite these findings, less is known about how these two risk factors conjointly impact diabetes prevalence within distinct BMI categories, a critical gap this study aims to address [9, 7, 10].

In our study, we utilized a nationally representative dataset comprising 253,680 individual responses that encompass various health indicators and demographic variables [11]. Data preprocessing included converting physical activity variables to represent physical inactivity and categorizing participants into BMI groups: underweight, normal weight, overweight, and obese. Logistic regression models assessed the relationship between physical inactivity, high blood pressure, and diabetes prevalence within each BMI category [13, 14]. We adjusted our models for potential confounders, including

Table 3: Logistic regression for additional factors in each BMI group

BMIGroup	Parameter	Coefficient	95% CI	P-value
Obese	**Sex (Male)**	0.1542	(0.12, 0.1884)	$<10^{-6}$
	Age	0.142	(0.1353, 0.1487)	$<10^{-6}$
	Education Level	-0.04557	(-0.06352, -0.02762)	$<10^{-6}$
	Income Level	-0.1303	(-0.1388, -0.1217)	$<10^{-6}$
Overweight	**Sex (Male)**	0.2752	(0.232, 0.3183)	$<10^{-6}$
	Age	0.146	(0.1375, 0.1546)	$<10^{-6}$
	Education Level	-0.08656	(-0.1093, -0.06386)	$<10^{-6}$
	Income Level	-0.1427	(-0.154, -0.1315)	$<10^{-6}$
Normal weight	**Sex (Male)**	0.5818	(0.5143, 0.6494)	$<10^{-6}$
	Age	0.1489	(0.1363, 0.1616)	$<10^{-6}$
	Education Level	-0.16	(-0.1956, -0.1244)	$<10^{-6}$
	Income Level	-0.1305	(-0.1479, -0.1132)	$<10^{-6}$
Underweight	**Sex (Male)**	0.6117	(0.2654, 0.9579)	0.000536
	Age	0.09839	(0.04132, 0.1555)	0.000727
	Education Level	-0.1993	(-0.3609, -0.03766)	0.0157
	Income Level	-0.00642	(-0.08487, 0.07203)	0.873

Age: Age group in 5-year categories
Sex (Male): 1: Male, 0: Female
Education Level: 1: Never attended school, 2: Elementary, 3: Some high school, 4: High school, 5: Some college, 6: College
Income Level: 1: <=10K, 2: <=15K, 3: <=20K, 4: <=25K, 5: <=35K, 6: <=50K, 7: <=75K, 8: >75K

age, sex, education, and income.

Our results indicated that diabetes prevalence rises significantly with higher BMI, corroborating previous findings [12]. Specifically, higher BMI is consistently associated with an increased occurrence of diabetes. Further analysis revealed that both physical inactivity and high blood pressure were significant risk factors for diabetes. Our regression analyses showed positive coefficients for both No Physical Activity and High Blood Pressure across all BMI categories, underscoring their impact on diabetes risk [2, 6, 7, 18, 19, 20]. Notably, the interaction term between these two factors showed a trend towards significance in the overweight group but was non-significant in the obese group. This particular finding aligns with research indicating that lifestyle interventions tailored to specific subgroups can effectively manage diabetes risk [17].

However, this study has several limitations that must be acknowledged. The cross-sectional design of the BRFSS data restricts the ability to infer causal relationships between the examined variables [16]. Longitudinal studies, which follow participants over time, are necessary to validate the observed associations and better elucidate causal pathways. Additionally, the reliance on self-reported data may introduce biases or inaccuracies in reporting health behaviors and conditions, as participants might underreport or misinterpret their activity levels and health status [21]. The BRFSS data lacks granularity regarding specific types of physical activity or blood pressure management strategies, which represents another area for future research. Moreover, unmeasured confounding factors, such as genetic predispositions or other health behaviors not captured by the BRFSS survey, could impact the results.

In conclusion, our findings indicate that both physical inactivity and high blood pressure are significant risk factors for diabetes across all BMI categories, with an evident increase in diabetes prevalence in higher BMI groups. The results suggest that targeted interventions focusing on increasing physical activity and managing blood pressure could effectively reduce diabetes risk, particularly for overweight individuals. Future research should employ longitudinal designs to confirm these findings and explore the mechanisms through which physical activity and blood pressure interact to influence diabetes risk. Understanding these dynamics will be crucial for developing comprehensive diabetes prevention strategies aimed at reducing the public health burden of this growing epidemic.

Methods

Data Source

The data for this study were derived from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), a health-related telephone survey conducted annually by the Centers for Disease Control and Prevention (CDC). The dataset comprises 253,680 responses, each representing an individual participant's health-related risk behaviors, chronic health conditions, and use of preventative services. The dataset includes 22 features, encompassing various health indicators, demographic variables, and self-reported health metrics. The survey's focus on diabetes-related factors enables a comprehensive exploration of the variables influencing diabetes prevalence.

Data Preprocessing

Data preprocessing commenced with the transformation of the physical activity variable to represent physical inactivity. Subsequently, participants were categorized into BMI groups: underweight, normal weight, overweight, and obese. This classification facilitated the examination of physical activity and blood pressure impacts within distinct BMI categories. No other modifications or imputed values were introduced, ensuring the integrity and authenticity of the original responses.

Data Analysis

Data analysis consisted of several stages. Initially, summary statistics for various health indicators were computed within each BMI category. This overview highlighted differences in health-related factors across BMI groups. Following this, the prevalence of diabetes within each BMI category was calculated, along with standard errors and confidence intervals, to provide a clearer understanding of diabetes distribution.

A series of logistic regression models were then employed to investigate the relationships between physical inactivity, high blood pressure, and diabetes within each BMI category. These models adjusted for potential confounders such as age, sex, education level, and income. The interaction between physical inactivity and high blood pressure was of particular interest. Results from these models offered insights into how these factors conjointly influence diabetes risk, with a focus on identifying significant interactions that differed across BMI groups.

A comprehensive examination of model coefficients, including confidence intervals and p-values, was conducted to assess the statistical significance and strength of the observed associations. The results of these analyses were visualized to facilitate interpretation and highlight key findings. Finally, additional summaries were generated to contextualize the number of observations and diabetes cases within each BMI group, offering further validation of the study's scope and robustness.

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] R. Maghembe, James A. Mpemba, and Mwanyika Alfred. Physical activity for health: examining similarities and differences between patients with type 2 diabetes and their adult children. *African Journal of Biomedical Research*, 19:179–185, 2016.

[2] S. Colberg, R. Sigal, B. Fernhall, J. Regensteiner, B. Blissmer, R. Rubin, L. Chasan-Taber, A. Albright, and B. Braun. Exercise and type 2 diabetes. *Diabetes Care*, 33:2692 – 2696, 2010.

[3] J. Seiglie, M. Marcus, Cara Ebert, Nikolaos Prodromidis, P. Geldsetzer, M. Theilmann, K. Agoudavi, Glennis Andall-Brereton, K. Aryal, B. Bicaba, P. Bovet, G. Brian, M. Dorobanu, G. Gathecha, M. Gurung, D. Guwatudde, M. Msaidi, C. Houehanou, D. Houinato, J. Jrgensen, G. Kagaruki, K. Karki, D. Labadarios, J. Martins, Mary T Mayige, R. Wong-McClure, J. K. Mwangi, O. Mwalim, B. Norov, Sarah Quesnel-Crooks, Bahendeka K Silver, L. Sturua, L. Tsabedze, C. Wesseh, A. Stokes, R. Atun, J. Davies, S. Vollmer, T. Brnighausen, L. Jaacks, J. Meigs, D. Wexler, and J. Manne-Goehler. Diabetes prevalence and its relationship with education, wealth, and bmi in 29 low- and middle-income countries. *Diabetes Care*, 43:767 – 775, 2020.

[4] S. Bhupathiraju and Frank B. Hu. Epidemiology of obesity and diabetes and their cardiovascular complications. *Circulation research*, 118 11:1723–35, 2016.

[5] David W. Brown, L. Balluz, W. Giles, G. Beckles, D. Moriarty, E. Ford, and A. Mokdad. Diabetes mellitus and health-related quality of life among older adults. findings from the behavioral risk factor surveillance system (brfss). *Diabetes research and clinical practice*, 65 2:105–15, 2004.

[6] D. Warburton, Crystal Whitney Nicol, and S. Bredin. Health benefits of physical activity: the evidence. *Canadian Medical Association Journal*, 174:801 – 809, 2006.

[7] D. Grossman, Kirsten Bibbins-Domingo, S. Curry, M. Barry, K. Davidson, Chyke A Doubeni, J. Epling, A. Kemper, A. Krist, A. Kurth, C. Landefeld, C. Mangione, M. Phipps, Michael Silverstein, M. Simon, and Chien-Wen Tseng. Behavioral counseling to promote a healthful diet and physical activity for cardiovascular disease prevention in

adults without cardiovascular risk factors: Us preventive services task force recommendation statement. *JAMA*, 318:167174, 2017.

[8] Susan Cheng, V. Xanthakis, L. Sullivan, and R. Vasan. Blood pressure tracking over the adult life course: Patterns and correlates in the framingham heart study. *Hypertension*, 60:13931399, 2012.

[9] U. E. Ntuk, J. Gill, D. Mackay, N. Sattar, and J. Pell. Ethnic-specific obesity cutoffs for diabetes risk: Cross-sectional study of 490,288 uk biobank participants. *Diabetes Care*, 37:2500 – 2507, 2014.

[10] A. Fretts, B. Howard, B. McKnight, G. Duncan, S. Beresford, M. Mete, Ying Zhang, and D. Siscovick. Lifes simple 7 and incidence of diabetes among american indians: The strong heart family study. *Diabetes Care*, 37:2240 – 2245, 2014.

[11] V. Preedy and Ronald R. Watson. Behavioral risk factor surveillance system. *Iowa medicine : journal of the Iowa Medical Society*, 79 9:436, 438, 1989.

[12] E. Lundeen, Sohyun Park, L. Pan, Terry OToole, Kevin A. Matthews, and H. Blanck. Obesity prevalence among adults living in metropolitan and nonmetropolitan counties united states, 2016. *Morbidity and Mortality Weekly Report*, 67:653 – 658, 2018.

[13] Wentao Bao. Model building strategy for logistic regression: purposeful selection. *Annals of translational medicine*, 4 6:111, 2016.

[14] Robin Gomila. Logistic or linear? estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of experimental psychology. General*, 2019.

[15] D. Ldecke. ggeffects: Tidy data frames of marginal effects from regression models. *J. Open Source Softw.*, 3:772, 2018.

[16] H. Verma and Rajeev Garg. Effect of magnesium supplementation on type 2 diabetes associated cardiovascular risk factors: a systematic review and metaanalysis. *Journal of Human Nutrition and Dietetics*, 30:621633, 2017.

[17] M. Abramowitz, C. B. Hall, A. Amodu, Deep Sharma, Lagu A. Androga, and M. Hawkins. Muscle mass, bmi, and mortality among adults in the united states: A population-based cohort study. *PLoS ONE*, 13, 2018.

[18] 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.

[19] Jie Hu, David M. Kline, Alai Tan, Songzhu Zhao, G. Brock, L. Mion, J. Efird, Danxin Wang, M. Sims, Bo Wu, M. Mongraw-Chaffin, and Joshua J. Joseph. Association between social determinants of health and glycemic control among african american people with type 2 diabetes: The jackson heart study. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine*, 2022.

[20] C. Guo, Hsiao Ting Yang, Ly yun Chang, Y. Bo, Changqing Lin, Yiqian Zeng, Tony Tam, Alexis K. H. Lau, G. Hoek, and X. Lao. Habitual exercise is associated with reduced risk of diabetes regardless of air pollution: a longitudinal cohort study. *Diabetologia*, 64:1298 – 1308, 2021.

[21] K. S. Reddy and M. Katan. Diet, nutrition and the prevention of hypertension and cardiovascular diseases. *Public Health Nutrition*, 7:167 – 186, 2004.

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:

```python
import pandas as pd

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

# Open the text file in write mode
with open("data_exploration.txt", "w") as f:

    # Data Size
    f.write("# Data Size\n")
    f.write(f"Number of rows: {df.shape[0]}\n")
    f.write(f"Number of columns: {df.shape[1]}\n\n")

    # Summary Statistics
```

```
f.write("# Summary Statistics\n")
f.write(df.describe().to_string())
f.write("\n\n")

# Categorical Variables
f.write("# Categorical Variables\n")
categorical_cols = ['Diabetes_binary', 'HighBP', 'HighChol'
    ↪ , 'CholCheck', 'Smoker', 'Stroke', '
    ↪ HeartDiseaseorAttack',
                    'PhysActivity', 'Fruits', 'Veggies', '
                        ↪ HvyAlcoholConsump', '
                        ↪ AnyHealthcare', 'NoDocbcCost',
                    'DiffWalk', 'Sex']

for col in categorical_cols:
    f.write(f"{col} - Most common value: {df[col].mode()
        ↪ [0]}\n")

f.write("\n")

# Missing Values
f.write("# Missing Values\n")
f.write("Counts of missing, unknown, or undefined values: 0
    ↪  (The dataset is clean with no missing values)\n")
f.write("\n\n")

# Additional summaries that may be relevant
# For ordinal variables, let's summarize frequency counts
    ↪ for a better understanding
f.write("# Ordinal Variables Frequency Counts\n")
ordinal_cols = ['GenHlth', 'MentHlth', 'PhysHlth', 'Age', '
    ↪ Education', 'Income']

for col in ordinal_cols:
    f.write(f"{col} - Value Counts:\n")
    f.write(df[col].value_counts().to_string())
    f.write("\n\n")

# Properly format the summary statistics for better
    ↪ readability
f.write("# Formatted Summary Statistics\n")
summary_stats = df.describe().transpose()
summary_stats.index.name = 'Feature'
summary_stats.reset_index(inplace=True)
form_stats = summary_stats.to_string(index=False)
f.write(form_stats)
f.write("\n\n")
```

B.2 Code Description

The data exploration code is designed to analyze and provide a comprehensive overview of a dataset containing diabetes-related health indicators sourced from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The primary steps of this analysis include:

B.3 Data Loading

We loaded the dataset from a CSV file into a pandas DataFrame. This allows for efficient data manipulation and analysis using the pandas library.

B.4 Data Size

The number of rows and columns in the dataset were determined using the `shape` attribute of the DataFrame. This helps in understanding the scale of the dataset being analyzed.

B.5 Summary Statistics

Descriptive statistics for the dataset were computed using the `describe()` method from pandas. These statistics provide insights into the distribution, central tendency, and variability of the numerical features in the dataset.

B.6 Categorical Variable Analysis

For categorical variables, the most common value (mode) was computed using the `mode()` method. This helps in understanding the most frequent occurrence of categories within each categorical feature.

B.7 Missing Values

Although the dataset was stated to be clean with no missing values, we verified this by noting that no missing values exist. This reassures the quality and completeness of the dataset.

B.8 Ordinal Variable Frequency Counts

Frequency counts for ordinal variables were obtained using the `value_counts()` method. This summarization is crucial for understanding the distribution of categorical data that has a clear ordering, such as self-reported health status and age groups.

B.9 Formatted Summary Statistics

To enhance readability and presentation, the summary statistics previously calculated were reformatted. The `transpose()` method was used to switch rows and columns, and the index was reset for a more organized tabular format.

Overall, this code serves to provide a detailed exploration and initial analysis of the dataset, facilitating subsequent steps in data processing, feature engineering, and modeling.

B.10 Code Output

data_exploration.txt

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

# Summary Statistics
      Diabetes_binary HighBP HighChol CholCheck    BMI Smoker
          Stroke HeartDiseaseorAttack PhysActivity Fruits
          Veggies HvyAlcoholConsump AnyHealthcare NoDocbcCost
          GenHlth MentHlth PhysHlth DiffWalk    Sex    Age
          Education Income
count          253680 253680   253680    253680 253680 253680
   253680              253680       253680 253680  253680
              253680        253680       253680  253680
   253680   253680   253680 253680 253680    253680 253680
mean           0.1393  0.429   0.4241    0.9627  28.38 0.4432
   0.04057              0.09419       0.7565 0.6343  0.8114
              0.0562        0.9511      0.08418   2.511
   3.185    4.242   0.1682 0.4403  8.032      5.05  6.054
std            0.3463 0.4949   0.4942    0.1896  6.609 0.4968
   0.1973               0.2921       0.4292 0.4816  0.3912
              0.2303        0.2158       0.2777   1.068
   7.413    8.718   0.3741 0.4964  3.054    0.9858  2.071
min                 0      0        0         0     12      0
         0                    0            0      0       0
                0             0            0       1
   0        0        0      0      1         1      1
25%                 0      0        0         1     24      0
         0                    0            1      0       1
                0             1            0       2
   0        0        0      0      6         4      5
50%                 0      0        0         1     27      0
         0                    0            1      1       1
```

```
                   0              1            0       2
   0        0        0      0     8          5      7
75%                 0      1        1          1     31       1
         0                    0            1      1       1
                   0              1            0       3
   2        3        0      1    10          6      8
max                 1      1        1          1     98       1
         1                    1            1      1       1
                   1              1            1       5
   30       30       1      1     13          6      8

# Categorical Variables
Diabetes_binary - Most common value: 0
HighBP - Most common value: 0
HighChol - Most common value: 0
CholCheck - Most common value: 1
Smoker - Most common value: 0
Stroke - Most common value: 0
HeartDiseaseorAttack - Most common value: 0
PhysActivity - Most common value: 1
Fruits - Most common value: 1
Veggies - Most common value: 1
HvyAlcoholConsump - Most common value: 0
AnyHealthcare - Most common value: 1
NoDocbcCost - Most common value: 0
DiffWalk - Most common value: 0
Sex - Most common value: 0

# Missing Values
Counts of missing, unknown, or undefined values: 0 (The dataset
    is clean with no missing values)


# Ordinal Variables Frequency Counts
GenHlth - Value Counts:
GenHlth
2    89084
3    75646
1    45299
4    31570
5    12081

MentHlth - Value Counts:
MentHlth
0     175680
2      13054
30     12088
5       9030
1       8538
```

```
3        7381
10       6373
15       5505
4        3789
20       3364
7        3100
25       1188
14       1167
6         988
8         639
12        398
28        327
21        227
29        158
18         97
9          91
16         88
27         79
22         63
17         54
26         45
11         41
13         41
23         38
24         33
19         16

PhysHlth - Value Counts:
PhysHlth
0      160052
30      19400
2       14764
1       11388
3        8495
5        7622
10       5595
15       4916
4        4542
7        4538
20       3273
14       2587
25       1336
6        1330
8         809
21        663
12        578
28        522
29        215
9         179
```

```
18       152
16       112
27        99
17        96
24        72
22        70
26        69
13        68
11        60
23        56
19        22

Age - Value Counts:
Age
9     33244
10    32194
8     30832
7     26314
11    23533
6     19819
13    17363
5     16157
12    15980
4     13823
3     11123
2      7598
1      5700

Education - Value Counts:
Education
6    107325
5     69910
4     62750
3      9478
2      4043
1       174

Income - Value Counts:
Income
8    90385
7    43219
6    36470
5    25883
4    20135
3    15994
2    11783
1     9811

# Formatted Summary Statistics
```

Feature	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

C Data Analysis

C.1 Code

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

```
# IMPORT
import pandas as pd
import numpy as np
import statsmodels.formula.api as smf
from scipy.stats import norm
from my_utils import df_to_latex, df_to_figure
import pickle

# LOAD DATA
# Load the dataset from the provided CSV file
df = pd.read_csv('diabetes_binary_health_indicators_BRFSS2015.
    ↪ csv')

# DATASET PREPARATIONS
# No dataset preparations are needed.

# DESCRIPTIVE STATISTICS
```

```
# No descriptive statistics table is needed because it was
    ↪ requested to remove df_desc_stat.pkl table.

# PREPROCESSING
# Transform PhysActivity to NoPhysActivity for analysis
df['NoPhysActivity'] = 1 - df['PhysActivity']

# Create BMI group categories
df['BMIGroup'] = pd.cut(df['BMI'], bins=[0, 18.5, 24.9, 29.9,
    ↪ 100], labels=['Underweight', 'Normal weight', 'Overweight
    ↪ ', 'Obese'])

# ANALYSIS

## Table df_bmi_groups:
caption = "Summary statistics for BMI groups"
# Compute subset of descriptive statistics within each BMI
    ↪ group (important columns)
important_columns = ['Diabetes_binary', 'HighBP', 'HighChol', '
    ↪ PhysActivity', 'Age', 'Sex', 'Education', 'Income']
summary_stats = df.groupby('BMIGroup')[important_columns].mean
    ↪ ().transpose()
df_to_latex(summary_stats, 'df_bmi_groups', caption=caption)

## Figure df_bmi_diabetes:
caption = "Diabetes prevalence across different BMI groups"

# Compute prevalence of diabetes in each BMI group
df_bmi_diabetes = df.groupby('BMIGroup')['Diabetes_binary'].agg
    ↪ (['mean', 'count']).reset_index()
df_bmi_diabetes.columns = ['BMIGroup', 'DiabetesPrevalence', '
    ↪ count']

# Calculating the standard error for a proportion
df_bmi_diabetes['se'] = np.sqrt(df_bmi_diabetes['
    ↪ DiabetesPrevalence'] * (1 - df_bmi_diabetes['
    ↪ DiabetesPrevalence']) / df_bmi_diabetes['count'])
df_bmi_diabetes['ci_low'] = df_bmi_diabetes['DiabetesPrevalence
    ↪ '] - 1.96 * df_bmi_diabetes['se']
df_bmi_diabetes['ci_high'] = df_bmi_diabetes['
    ↪ DiabetesPrevalence'] + 1.96 * df_bmi_diabetes['se']

df_to_figure(df_bmi_diabetes, 'df_bmi_diabetes', x='BMIGroup',
    ↪ y=['DiabetesPrevalence'], y_ci=[('ci_low', 'ci_high')],
    ↪ kind='bar', caption=caption)

## Table df_bmi_regression:
caption = "Logistic regression results for each BMI group"
```

```
# Conduct logistic regression within each BMI group
results = []
for bmi_group in df['BMIGroup'].unique():
    subset = df[df['BMIGroup'] == bmi_group]
    if subset.empty:
        continue
    model = smf.logit(formula='Diabetes_binary ~ NoPhysActivity
        ↪  * HighBP + Age + C(Sex) + Education + Income', data=
        ↪ subset).fit()
    summary_df = model.summary2().tables[1].reset_index()
    summary_df['BMIGroup'] = bmi_group
    results.append(summary_df)
df_bmi_regression = pd.concat(results).reset_index(drop=True)

# Adding columns for confidence intervals and p-values in a
    ↪ suitable format for the figure
df_bmi_regression['ci'] = df_bmi_regression.apply(lambda row: (
    ↪ row['[0.025'], row['0.975]']), axis=1)
df_bmi_regression['pval_stars'] = df_bmi_regression['P>|z|'].
    ↪ apply(lambda p: '***' if p < 0.001 else '**' if p < 0.01
    ↪ else '*' if p < 0.05 else 'ns')

# Only select relevant rows and columns for the LaTeX table
required_rows = ['NoPhysActivity', 'HighBP', 'NoPhysActivity:
    ↪ HighBP']
df_bmi_regression_table = df_bmi_regression[df_bmi_regression['
    ↪ index'].isin(required_rows)][['BMIGroup', 'index', 'Coef.
    ↪ ', 'ci', 'P>|z|']]
df_bmi_regression_table.columns = ['BMIGroup', 'Parameter', '
    ↪ Coefficient', '95% CI', 'P-value']
df_bmi_regression_table = df_bmi_regression_table.set_index(['
    ↪ BMIGroup', 'Parameter'])
df_to_latex(df_bmi_regression_table, 'df_bmi_regression',
    ↪ caption=caption)

## Table df_bmi_regression_other:
caption = "Logistic regression for additional factors in each
    ↪ BMI group"

# Select rows for other factors of the model
required_rows_other = ['Age', 'C(Sex)[T.1]', 'Education', '
    ↪ Income']
df_bmi_regression_other_table = df_bmi_regression[
    ↪ df_bmi_regression['index'].isin(required_rows_other)][['
    ↪ BMIGroup', 'index', 'Coef.', 'ci', 'P>|z|']]
df_bmi_regression_other_table.columns = ['BMIGroup', 'Parameter
    ↪ ', 'Coefficient', '95% CI', 'P-value']
```

```
df_bmi_regression_other_table = df_bmi_regression_other_table.
    ↪ set_index(['BMIGroup', 'Parameter'])
df_to_latex(df_bmi_regression_other_table, '
    ↪ df_bmi_regression_other', caption=caption)

## Figure df_bmi_regression_fig:
caption = "Regression coefficients of physical activity and
    ↪ high blood pressure interactions on diabetes within BMI
    ↪ groups"
# Extracting the relevant information for coefficients, CIs,
    ↪ and p-values for the figure
fig_data_nophysactivity = df_bmi_regression_table.reset_index()
    ↪ .query('Parameter == "NoPhysActivity"')
fig_data_highbp = df_bmi_regression_table.reset_index().query('
    ↪ Parameter == "HighBP"')
fig_data_interaction = df_bmi_regression_table.reset_index().
    ↪ query('Parameter == "NoPhysActivity:HighBP"')

# Combine the data for plotting
reg_fig_data_nophysactivity = fig_data_nophysactivity.set_index
    ↪ ('BMIGroup')[['Coefficient', '95% CI', 'P-value']]
reg_fig_data_highbp = fig_data_highbp.set_index('BMIGroup')[['
    ↪ Coefficient', '95% CI', 'P-value']]
reg_fig_data_interaction = fig_data_interaction.set_index('
    ↪ BMIGroup')[['Coefficient', '95% CI', 'P-value']]

# Rename columns to avoid conflict
reg_fig_data_nophysactivity.columns = ['NoPhysActivity', '
    ↪ NoPhysActivityCI', 'NoPhysActivityPval']
reg_fig_data_highbp.columns = ['HighBP', 'HighBPCI', '
    ↪ HighBPPval']
reg_fig_data_interaction.columns = ['Interaction', '
    ↪ InteractionCI', 'InteractionPval']

# Joining the data
reg_fig_data = reg_fig_data_nophysactivity.join([
    ↪ reg_fig_data_highbp, reg_fig_data_interaction], how='
    ↪ outer').reset_index()

df_to_figure(reg_fig_data, 'df_bmi_regression_fig', kind='bar',
             x='BMIGroup',
             y=['NoPhysActivity', 'HighBP', 'Interaction'],
             y_ci=['NoPhysActivityCI', 'HighBPCI', '
                 ↪ InteractionCI'],
             y_p_value=['NoPhysActivityPval', 'HighBPPval', '
                 ↪ InteractionPval'],
             caption=caption)

# SAVE ADDITIONAL RESULTS
```

```
# Gathering the number of observations per BMI group
bmi_group_counts = df['BMIGroup'].value_counts().to_dict()

additional_results = {
    'Total number of observations': len(df),
    'Number of Diabetes cases': df['Diabetes_binary'].sum(),
    'BMI group counts': bmi_group_counts
}

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:
    `df`, `filename`, `caption`

    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

The code consists of several steps for analyzing the relationship between diabetes and various health indicators from the dataset derived from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015.

C.4 Data Loading

The dataset is loaded from a CSV file into a pandas DataFrame for further analysis.

C.5 Preprocessing

A new variable `NoPhysActivity` is created to represent the absence of physical activity by transforming the original `PhysActivity` variable. Additionally, the Body Mass Index (BMI) is categorized into groups: 'Underweight', 'Normal weight', 'Overweight', and 'Obese' for more detailed analysis within these defined strata.

C.6 Descriptive Analysis

Summary statistics for the BMI groups are computed and presented. The mean values of important variables such as `Diabetes_binary`, `HighBP`, `PhysActivity`, `Age`, `Sex`, `Education`, and `Income` are calculated within each BMI group, and the results are formatted into a LaTeX table.

C.7 Visualization of Diabetes Prevalence

The prevalence of diabetes across different BMI groups is computed. This involves calculating the mean proportion of diabetes cases within each BMI group, along with their standard errors and 95% confidence intervals. The results are then visualized as a bar plot, clearly depicting the diabetes prevalence within each BMI group.

C.8 Logistic Regression Analysis

Logistic regression models are fitted to investigate the relationship between diabetes and several predictors including `NoPhysActivity`, `HighBP`, and their interaction within each BMI group separately. The regression models also account for covariates such as `Age`, `Sex`, `Education`, and `Income`. The regression results, including coefficients, confidence intervals, and p-values, are organized into LaTeX tables for ease of presentation.

C.9 Visualization of Regression Coefficients

The significant regression coefficients, their confidence intervals, and p-values for the interaction between physical activity and high blood pressure within BMI groups are visualized. This visualization helps to understand the magnitude and significance of the effects of physical activity and high blood pressure on diabetes prevalence across different BMI groups.

C.10 Additional Results

The analysis concludes by saving additional results including the total number of observations, the number of diabetes cases, and the counts of observations within each BMI group for reference.

Each step effectively transforms and analyzes the data to understand and visualize the relationships between diabetes and various health indicators across different BMI categories, following a structured approach using statistical and visualization methods.

C.11 Code Output

df_bmi_diabetes.pkl

```
        BMIGroup  DiabetesPrevalence  count         se   ci_low
       ci_high
0    Underweight             0.05405   3127   0.004043  0.04612
   0.06197
1  Normal weight             0.05697  68953  0.0008827  0.05524
   0.0587
2     Overweight               0.114  93749   0.001038    0.112
   0.1161
3          Obese               0.234  87851   0.001428   0.2312
   0.2368
```

df_bmi_groups.pkl

```
BMIGroup         Underweight  Normal weight  Overweight   Obese
Diabetes_binary      0.05405        0.05697       0.114   0.234
HighBP                0.2897         0.2789      0.4162  0.5655
HighChol              0.2904         0.3245      0.4402    0.49
PhysActivity          0.7154         0.8228      0.7831  0.6776
Age                      8.1          7.907       8.238   7.908
Sex                    0.213         0.3482      0.4963  0.4611
Education              4.955          5.197       5.074   4.914
Income                 5.338          6.219       6.194     5.8
```

df_bmi_regression.pkl

```
                                    Coefficient
                                        95% CI     P-value
BMIGroup       Parameter
Obese          NoPhysActivity              0.3229    (0.2544,
    0.3914)    2.48e-20
               HighBP                       1.065      (1.017,
                  1.113)          0
```

```
              NoPhysActivity:HighBP       -0.064   (-0.1427,
                 0.0147)      0.111
Overweight    NoPhysActivity              0.3362    (0.2516,
   0.4208)   6.66e-15
              HighBP                       1.082      (1.027,
                 1.137)          0
              NoPhysActivity:HighBP     -0.1172  (-0.2177,
                 -0.0166)     0.0224
Normal weight NoPhysActivity              0.3534    (0.2281,
   0.4787)   3.27e-08
              HighBP                       1.087      (1.003,
                 1.172)  8.02e-140
              NoPhysActivity:HighBP    -0.07528  (-0.2299,
                 0.07934)       0.34
Underweight   NoPhysActivity              0.3878   (-0.1492,
   0.9248)      0.157
              HighBP                       1.287     (0.8488,
                 1.725)   8.49e-09
              NoPhysActivity:HighBP     -0.1986   (-0.8725,
                 0.4754)      0.564
```

df_bmi_regression_fig.pkl

```
        BMIGroup NoPhysActivity   NoPhysActivityCI
            NoPhysActivityPval HighBP         HighBPCI
            HighBPPval Interaction       InteractionCI
            InteractionPval
0          Obese          0.3229   (0.2544, 0.3914)
   2.48e-20  1.065   (1.017, 1.113)          0      -0.064
   (-0.1427, 0.0147)           0.111
1     Overweight          0.3362   (0.2516, 0.4208)
   6.66e-15  1.082   (1.027, 1.137)          0     -0.1172
   (-0.2177, -0.0166)          0.0224
2  Normal weight          0.3534   (0.2281, 0.4787)
   3.27e-08  1.087   (1.003, 1.172)  8.02e-140    -0.07528
   (-0.2299, 0.07934)            0.34
3    Underweight          0.3878  (-0.1492, 0.9248)
                0.157  1.287  (0.8488, 1.725)   8.49e-09
   -0.1986   (-0.8725, 0.4754)           0.564
```

df_bmi_regression_other.pkl

```
                              Coefficient                 95% CI
                                 P-value
BMIGroup       Parameter
Obese          C(Sex)[T.1]       0.1542        (0.12, 0.1884)
   9.4e-19
               Age                0.142      (0.1353, 0.1487)
                              0
```

```
                 Education        -0.04557  (-0.06352, -0.02762)
                    6.53e-07
                 Income           -0.1303     (-0.1388, -0.1217)
                    2.35e-196
Overweight       C(Sex)[T.1]       0.2752       (0.232, 0.3183)
   7.08e-36
                 Age                0.146      (0.1375, 0.1546)
                    5.46e-246
                 Education        -0.08656   (-0.1093, -0.06386)
                    7.77e-14
                 Income           -0.1427      (-0.154, -0.1315)
                    4.06e-137
Normal weight    C(Sex)[T.1]       0.5818      (0.5143, 0.6494)
   6.71e-64
                 Age               0.1489      (0.1363, 0.1616)
                    1.11e-117
                 Education          -0.16    (-0.1956, -0.1244)
                    1.26e-18
                 Income           -0.1305    (-0.1479, -0.1132)
                    3.24e-49
Underweight      C(Sex)[T.1]       0.6117      (0.2654, 0.9579)
   0.000536
                 Age              0.09839     (0.04132, 0.1555)
                    0.000727
                 Education        -0.1993   (-0.3609, -0.03766)
                        0.0157
                 Income          -0.00642   (-0.08487, 0.07203)
                         0.873
```

additional_results.pkl

```
{
    'Total number of observations': 253680,
    'Number of Diabetes cases': 35346,
    'BMI group counts': {Overweight: 93749, Obese: 87851,
       Normal weight: 68953, Underweight: 3127},
}
```

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
```

```python
# PREPARATION FOR ALL TABLES AND FIGURES
shared_mapping: AbbrToNameDef = {
    'HighBP': ('High BP', '1: Yes, 0: No'),
    'HighChol': ('High Cholesterol', '1: Yes, 0: No'),
    'PhysActivity': ('Physical Activity', '1: Yes, 0: No'),
    'Age': ('Age', 'Age group in 5-year categories'),
    'Sex': ('Sex', '0: Female, 1: Male'),
    'Education': ('Education Level', '1: Never attended school,
       ↪  2: Elementary, 3: Some high school, 4: High school,
       ↪ 5: Some college, 6: College'),
    'Income': ('Income Level', '1: <=10K, 2: <=15K, 3: <=20K,
       ↪ 4: <=25K, 5: <=35K, 6: <=50K, 7: <=75K, 8: >75K'),
}

# Process df_bmi_diabetes
df_bmi_diabetes = pd.read_pickle('df_bmi_diabetes.pkl')

# Not Applicable
mapping = {
    'DiabetesPrevalence': ('Diabetes Prevalence', None),
    'count': ('Count', 'Number of respondents in each BMI group
       ↪ '),
    'se': ('Standard Error', None),
    'ci_low': ('CI Lower Bound', '95% Confidence Interval Lower
       ↪  Bound'),
    'ci_high': ('CI Upper Bound', '95% Confidence Interval
       ↪ Upper Bound')
}
abbrs_to_names, glossary = split_mapping(mapping)
df_bmi_diabetes.rename(columns=abbrs_to_names, inplace=True)

df_to_figure(
    df_bmi_diabetes, 'df_bmi_diabetes_formatted',
    caption="Diabetes prevalence across different BMI groups",
    note="Confidence intervals are shown as error bars.",
    glossary=glossary,
    x='BMIGroup',
    y=['Diabetes Prevalence'],
    y_ci=[('CI Lower Bound', 'CI Upper Bound')],
    kind='bar',
    xlabel='BMI Group',
    ylabel='Diabetes Prevalence'
)

# Process df_bmi_groups
df_bmi_groups = pd.read_pickle('df_bmi_groups.pkl')

# Not Applicable
```

```python
mapping = shared_mapping
mapping |= {
    'Diabetes_binary': ('Diabetes Prevalence', 'Proportion of
        ↪ respondents with diabetes (1=Yes, 0=No)'),
}
abbrs_to_names, glossary = split_mapping(mapping)
df_bmi_groups.rename(index=abbrs_to_names, columns=
    ↪ abbrs_to_names, inplace=True)

df_to_latex(
    df_bmi_groups, 'df_bmi_groups_formatted',
    caption="Summary statistics for BMI groups",
    note="The table shows the mean of each variable within each
        ↪  BMI group.",
    glossary=glossary
)

# Process df_bmi_regression
df_bmi_regression = pd.read_pickle('df_bmi_regression.pkl')

# Format values: Not Applicable
mapping = {
    'NoPhysActivity': ('No Phys Act', None),
    'HighBP': ('High BP', 'High Blood Pressure, 1: Yes, 0: No')
        ↪ ,
    'NoPhysActivity:HighBP': ('Interaction (NPA & HBP)', '
        ↪ Interaction term between No Physical Activity and
        ↪ High Blood Pressure'),
    'Coefficient': ('Coef.', None),
    '95% CI': ('95% CI', '95% Confidence Interval of the
        ↪ Coefficient'),
    'P-value': ('P-val', None),
}
abbrs_to_names, glossary = split_mapping(mapping)
df_bmi_regression.rename(columns=abbrs_to_names, index=
    ↪ abbrs_to_names, inplace=True)

df_to_latex(
    df_bmi_regression, 'df_bmi_regression_formatted',
    caption="Logistic regression results for each BMI group",
    note="Interaction terms are provided in the table. Standard
        ↪  error terms are omitted for brevity.",
    glossary=glossary
)

# Process df_bmi_regression_fig
df_bmi_regression_fig = pd.read_pickle('df_bmi_regression_fig.
    ↪ pkl')
```

```
# Format values: Not Applicable
mapping = {
    'NoPhysActivity': ('No Phys Act', '1: No activity, 0:
        ↪ Activity'),
    'HighBP': ('High BP', '1: Yes, 0: No'),
    'Interaction': ('Interaction (NPA & HBP)', 'Interaction
        ↪ term between No Physical Activity and High Blood
        ↪ Pressure'),
    'NoPhysActivityCI': ('No Phys Act CI', '95% Confidence
        ↪ Interval of No Physical Activity'),
    'NoPhysActivityPval': ('No Phys Act P-val', None),
    'HighBPCI': ('High BP CI', '95% Confidence Interval of High
        ↪  Blood Pressure'),
    'HighBPPval': ('High BP P-val', None),
    'InteractionCI': ('Interaction CI', '95% Confidence
        ↪ Interval of Interaction'),
    'InteractionPval': ('Interaction P-val', None),
}
abbrs_to_names, glossary = split_mapping(mapping)
df_bmi_regression_fig.rename(columns=abbrs_to_names, index=
    ↪ abbrs_to_names, inplace=True)

df_to_figure(
    df_bmi_regression_fig, 'df_bmi_regression_fig_formatted',
    caption="Regression coefficients of physical activity and
        ↪ high blood pressure interactions on diabetes within
        ↪ BMI groups",
    glossary=glossary,
    x='BMIGroup',
    y=['No Phys Act', 'High BP', 'Interaction (NPA & HBP)'],
    y_ci=['No Phys Act CI', 'High BP CI', 'Interaction CI'],
    y_p_value=['No Phys Act P-val', 'High BP P-val', '
        ↪ Interaction P-val'],
    kind='bar',
    xlabel='BMI Group',
    ylabel='Coefficient'
)

# Process df_bmi_regression_other
df_bmi_regression_other = pd.read_pickle('
    ↪ df_bmi_regression_other.pkl')

# Not applicable
mapping = {
    'Age': ('Age', 'Age group in 5-year categories'),
    'C(Sex)[T.1]': ('Sex (Male)', '1: Male, 0: Female'),
    'Education': ('Education Level', '1: Never attended school,
        ↪  2: Elementary, 3: Some high school, 4: High school,
        ↪ 5: Some college, 6: College'),
```

```
    'Income': ('Income Level', '1: <=10K, 2: <=15K, 3: <=20K,
        ↪ 4: <=25K, 5: <=35K, 6: <=50K, 7: <=75K, 8: >75K'),
}
abbrs_to_names, glossary = split_mapping(mapping)
df_bmi_regression_other.rename(columns=abbrs_to_names, index=
    ↪ abbrs_to_names, inplace=True)

df_to_latex(
    df_bmi_regression_other, 'df_bmi_regression_other_formatted
        ↪ ',
    caption="Logistic regression for additional factors in each
        ↪  BMI group",
    glossary=glossary
)
```

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` (str): Note to be added below the table caption.
    `glossary` (Dict[Any, str]): Glossary for the table.
    """


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,
        xlabel: str = None, ylabel: str = None,
        note: str = None, glossary: Dict[Any, str] = None,
    ):
    """
    Save a DataFrame `df` and create a LaTeX figure.
    Parameters, for LaTex embedding of the figure:
    `df`, `filename`, `caption`

    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` (str): Label for the x-axis.
    `ylabel` (str): Label for the y-axis.
    `note` (str): Note to be added below the figure caption.
    `glossary` (Dict[Any, str]): Glossary for the figure.

    `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_bmi_diabetes_formatted.pkl

```
\begin{figure}[htbp]
\centering
\includegraphics{df_bmi_diabetes_formatted.png}
\caption{Diabetes prevalence across different BMI groups
Confidence intervals are shown as error bars.
Count: Number of respondents in each BMI group.
CI Lower Bound: 95\% Confidence Interval Lower Bound.
CI Upper Bound: 95\% Confidence Interval Upper Bound.}
\label{figure:df-bmi-diabetes-formatted}
\end{figure}
% This latex figure presents "df_bmi_diabetes_formatted.png",
% which was created from the df:
%
% index,"BMIGroup","Diabetes Prevalence","Count","Standard
    Error","CI Lower Bound","CI Upper Bound"
% 0,Underweight,0.05405,3127,0.004043,0.04612,0.06197
% 1,Normal weight,0.05697,68953,0.0008827,0.05524,0.0587
% 2,Overweight,0.114,93749,0.001038,0.112,0.1161
% 3,Obese,0.234,87851,0.001428,0.2312,0.2368
%
% To create the figure, this df was plotted with the command:
%
% df.plot(x='BMIGroup', y=['Diabetes Prevalence'], kind='bar',
    xlabel='BMI Group', ylabel='Diabetes Prevalence')
%
% Confidence intervals for y-values were then plotted based on
    column: [('CI Lower Bound', 'CI Upper Bound')].
```

df_bmi_groups_formatted.pkl

```
\begin{table}[h]
\caption{Summary statistics for BMI groups}
\label{table:df-bmi-groups-formatted}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrrrr}
\toprule
BMIGroup & Underweight & Normal weight & Overweight & Obese \\
\midrule
\textbf{Diabetes Prevalence} & 0.05405 & 0.05697 & 0.114 &
   0.234 \\
\textbf{High BP} & 0.2897 & 0.2789 & 0.4162 & 0.5655 \\
\textbf{High Cholesterol} & 0.2904 & 0.3245 & 0.4402 & 0.49 \\
\textbf{Physical Activity} & 0.7154 & 0.8228 & 0.7831 & 0.6776
   \\
\textbf{Age} & 8.1 & 7.907 & 8.238 & 7.908 \\
\textbf{Sex} & 0.213 & 0.3482 & 0.4963 & 0.4611 \\
\textbf{Education Level} & 4.955 & 5.197 & 5.074 & 4.914 \\
\textbf{Income Level} & 5.338 & 6.219 & 6.194 & 5.8 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item The table shows the mean of each variable within each BMI
    group.
\item \textbf{High BP}: 1: Yes, 0: No
\item \textbf{High Cholesterol}: 1: Yes, 0: No
\item \textbf{Physical Activity}: 1: Yes, 0: No
\item \textbf{Age}: Age group in 5-year categories
\item \textbf{Sex}: 0: Female, 1: Male
\item \textbf{Education Level}: 1: Never attended school, 2:
   Elementary, 3: Some high school, 4: High school, 5: Some
   college, 6: College
\item \textbf{Income Level}: 1: $<$=10K, 2: $<$=15K, 3: $<$=20K
   , 4: $<$=25K, 5: $<$=35K, 6: $<$=50K, 7: $<$=75K, 8: $>$75K
\item \textbf{Diabetes Prevalence}: Proportion of respondents
   with diabetes (1=Yes, 0=No)
\end{tablenotes}
\end{threeparttable}
\end{table}
```

df_bmi_regression_fig_formatted.pkl

```
\begin{figure}[htbp]
\centering
\includegraphics{df_bmi_regression_fig_formatted.png}
```

```
\caption{Regression coefficients of physical activity and high
    blood pressure interactions on diabetes within BMI groups
No Phys Act: 1: No activity, 0: Activity.
High BP: 1: Yes, 0: No.
Interaction (NPA \& HBP): Interaction term between No Physical
    Activity and High Blood Pressure.
No Phys Act CI: 95\% Confidence Interval of No Physical
    Activity.
High BP CI: 95\% Confidence Interval of High Blood Pressure.
Interaction CI: 95\% Confidence Interval of Interaction.
Significance: ns p $>$= 0.01, * p $<$ 0.01, ** p $<$ 0.001, ***
     p $<$ 0.0001.}
\label{figure:df-bmi-regression-fig-formatted}
\end{figure}
% This latex figure presents "df_bmi_regression_fig_formatted.
    png",
% which was created from the df:
%
% index,"BMIGroup","No Phys Act","No Phys Act CI","No Phys Act
    P-val","High BP","High BP CI","High BP P-val","Interaction
    (NPA & HBP)","Interaction CI","Interaction P-val"
% 0,Obese,0.3229,(0.2544, 0.3914),<1e-06,1.065,(1.017, 1.113),<
    1e-06,-0.064,(-0.1427, 0.0147),0.111
% 1,Overweight,0.3362,(0.2516, 0.4208),<1e-06,1.082,(1.027,
    1.137),<1e-06,-0.1172,(-0.2177, -0.0166),0.0224
% 2,Normal weight,0.3534,(0.2281, 0.4787),<1e-06,1.087,(1.003,
    1.172),<1e-06,-0.07528,(-0.2299, 0.07934),0.34
% 3,Underweight,0.3878,(-0.1492, 0.9248),0.157,1.287,(0.8488,
    1.725),<1e-06,-0.1986,(-0.8725, 0.4754),0.564
%
% To create the figure, this df was plotted with the command:
%
% df.plot(x='BMIGroup', y=['No Phys Act', 'High BP', '
    Interaction (NPA & HBP)'], kind='bar', xlabel='BMI Group',
    ylabel='Coefficient')
%
% Confidence intervals for y-values were then plotted based on
    column: ['No Phys Act CI', 'High BP CI', 'Interaction CI'].
%
% P-values for y-values were taken from column: ['No Phys Act P
    -val', 'High BP P-val', 'Interaction P-val'].
%
% These p-values were presented above the data points as stars
    (with significance threshold values indicated in the figure
     caption).
```

df_bmi_regression_formatted.pkl

```
\begin{table}[h]
\caption{Logistic regression results for each BMI group}
\label{table:df-bmi-regression-formatted}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{llrll}
\toprule
 &  & Coef. & 95\% CI & P-val \\
BMIGroup & Parameter &  &  &  \\
\midrule
\multirow[t]{3}{*}{\textbf{Obese}} & \textbf{No Phys Act} &
    0.3229 & (0.2544, 0.3914) & $<$1e-06 \\
\textbf{} & \textbf{High BP} & 1.065 & (1.017, 1.113) & $<$1e
    -06 \\
\textbf{} & \textbf{Interaction (NPA \& HBP)} & -0.064 &
    (-0.1427, 0.0147) & 0.111 \\
\cline{1-5}
\multirow[t]{3}{*}{\textbf{Overweight}} & \textbf{No Phys Act}
    & 0.3362 & (0.2516, 0.4208) & $<$1e-06 \\
\textbf{} & \textbf{High BP} & 1.082 & (1.027, 1.137) & $<$1e
    -06 \\
\textbf{} & \textbf{Interaction (NPA \& HBP)} & -0.1172 &
    (-0.2177, -0.0166) & 0.0224 \\
\cline{1-5}
\multirow[t]{3}{*}{\textbf{Normal weight}} & \textbf{No Phys
    Act} & 0.3534 & (0.2281, 0.4787) & $<$1e-06 \\
\textbf{} & \textbf{High BP} & 1.087 & (1.003, 1.172) & $<$1e
    -06 \\
\textbf{} & \textbf{Interaction (NPA \& HBP)} & -0.07528 &
    (-0.2299, 0.07934) & 0.34 \\
\cline{1-5}
\multirow[t]{3}{*}{\textbf{Underweight}} & \textbf{No Phys Act}
     & 0.3878 & (-0.1492, 0.9248) & 0.157 \\
\textbf{} & \textbf{High BP} & 1.287 & (0.8488, 1.725) & $<$1e
    -06 \\
\textbf{} & \textbf{Interaction (NPA \& HBP)} & -0.1986 &
    (-0.8725, 0.4754) & 0.564 \\
\cline{1-5}
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item Interaction terms are provided in the table. Standard
    error terms are omitted for brevity.
\item \textbf{High BP}: High Blood Pressure, 1: Yes, 0: No
\item \textbf{Interaction (NPA \& HBP)}: Interaction term
    between No Physical Activity and High Blood Pressure
\item \textbf{95\% CI}: 95\% Confidence Interval of the
```

```
    Coefficient
\end{tablenotes}
\end{threeparttable}
\end{table}
```

df_bmi_regression_other_formatted.pkl

```
\begin{table}[h]
\caption{Logistic regression for additional factors in each BMI
     group}
\label{table:df-bmi-regression-other-formatted}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{llrll}
\toprule
 &  & Coefficient & 95\% CI & P-value \\
BMIGroup & Parameter &  &  &  \\
\midrule
\multirow[t]{4}{*}{\textbf{Obese}} & \textbf{Sex (Male)} &
    0.1542 & (0.12, 0.1884) & $<$1e-06 \\
\textbf{} & \textbf{Age} & 0.142 & (0.1353, 0.1487) & $<$1e-06
    \\
\textbf{} & \textbf{Education Level} & -0.04557 & (-0.06352,
    -0.02762) & $<$1e-06 \\
\textbf{} & \textbf{Income Level} & -0.1303 & (-0.1388,
    -0.1217) & $<$1e-06 \\
\cline{1-5}
\multirow[t]{4}{*}{\textbf{Overweight}} & \textbf{Sex (Male)} &
     0.2752 & (0.232, 0.3183) & $<$1e-06 \\
\textbf{} & \textbf{Age} & 0.146 & (0.1375, 0.1546) & $<$1e-06
    \\
\textbf{} & \textbf{Education Level} & -0.08656 & (-0.1093,
    -0.06386) & $<$1e-06 \\
\textbf{} & \textbf{Income Level} & -0.1427 & (-0.154, -0.1315)
     & $<$1e-06 \\
\cline{1-5}
\multirow[t]{4}{*}{\textbf{Normal weight}} & \textbf{Sex (Male)
    } & 0.5818 & (0.5143, 0.6494) & $<$1e-06 \\
\textbf{} & \textbf{Age} & 0.1489 & (0.1363, 0.1616) & $<$1e-06
     \\
\textbf{} & \textbf{Education Level} & -0.16 & (-0.1956,
    -0.1244) & $<$1e-06 \\
\textbf{} & \textbf{Income Level} & -0.1305 & (-0.1479,
    -0.1132) & $<$1e-06 \\
\cline{1-5}
\multirow[t]{4}{*}{\textbf{Underweight}} & \textbf{Sex (Male)}
    & 0.6117 & (0.2654, 0.9579) & 0.000536 \\
\textbf{} & \textbf{Age} & 0.09839 & (0.04132, 0.1555) &
```

```
    0.000727 \\
\textbf{} & \textbf{Education Level} & -0.1993 & (-0.3609,
    -0.03766) & 0.0157 \\
\textbf{} & \textbf{Income Level} & -0.00642 & (-0.08487,
    0.07203) & 0.873 \\
\cline{1-5}
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Age}: Age group in 5-year categories
\item \textbf{Sex (Male)}: 1: Male, 0: Female
\item \textbf{Education Level}: 1: Never attended school, 2:
    Elementary, 3: Some high school, 4: High school, 5: Some
    college, 6: College
\item \textbf{Income Level}: 1: $<$=10K, 2: $<$=15K, 3: $<$=20K
    , 4: $<$=25K, 5: $<$=35K, 6: $<$=50K, 7: $<$=75K, 8: $>$75K
\end{tablenotes}
\end{threeparttable}
\end{table}
```

E Calculation Notes

- 23.4 / 5.697 = 4.107

  Fold change between normal and obese group for diabetes prevalence