Analysis of
Diabetes Health Indicators Dataset

Name: Harpreet Kaur
College : Southern Alberta Institute of Technology
Date: Feb 15, 2025

TABLE OF CONTENT

Analysis of

Diabetes Health Indicators Dataset

ARTICLE

Harpreet Kaur

Abstract: This study analyzes diabetes health indicators using a dataset from the UCI Machine Learning Repository to understand trends, risk factors, and disparities across different demographics in the U.S. Diabetes is a chronic condition leading to serious health complications such as heart disease, vision loss, and kidney failure. The research aims to identify key risk factors, promote early diagnosis, and support data-driven decision-making for public health policies. A Power BI dashboard was developed to visualize significant risk factors, including BMI, Blood Pressure, Cholesterol, Age, and Walking Difficulty, revealing higher diabetes rates among older individuals with high BMI and mobility issues. The findings provide actionable insights for targeted health interventions and policy improvements, contributing to better management and prevention of diabetes.

Introduction

Diabetes is a chronic disease affecting millions in the U.S., leading to serious health complications like heart disease, vision loss, and kidney failure. Early diagnosis and effective management are crucial for reducing its impact. This study analyzes diabetes health indicators using a dataset from the UCI Machine Learning Repository to identify key risk factors, including BMI, Blood Pressure, Cholesterol, Age, and Walking Difficulty. Using Power BI for data visualization, the research highlights trends and disparities across different demographics, providing insights to support public health decision-making and targeted interventions for diabetes prevention and management.

Data set Overview

The Diabetes Health Indicators Dataset is sourced from the Behavioral Risk Factor Surveillance System (BRFSS) 2015, a health-related telephone survey conducted annually by the Centers for Disease Control and Prevention (CDC) in the United States. The dataset contains 253,680 cleaned survey responses, providing insights into diabetes prevalence, risk factors, and health disparities across different demographics. It focuses on three classes of the target variable **Diabetes_012**:

- **0**: No diabetes or diabetes only during pregnancy
- **1**: Prediabetes
- **2**: Diabetes

This dataset is imbalanced, with the majority of respondents falling into the no-diabetes category. It includes 21 feature variables, such as **HighBP (High Blood Pressure), HighChol (High Cholesterol), CholCheck (Cholesterol Check), BMI (Body Mass Index), Smoker, Stroke, HeartDiseaseorAttack, PhysActivity (Physical Activity), and Fruits (Daily Fruit Consumption)**. These features are crucial for predicting diabetes risk using data-driven models.

Dataset link : https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

Features of Dataset:

1) Diabetes_012 : 0 = no diabetes 1 = prediabetes 2 = diabetes
2) HighBP : 0 = no high BP 1 = high BP
3) High Chol : 0 = no high cholesterol 1 = high cholesterol
4) CholCheck : 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
5) BMI : Body mass Index
6) Smoker : Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
7) Stroke : (Ever told) you had a stroke. 0 = no 1 = yes
8) HeartDiseaseorAttack : coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
9) PhysActivity : physical activity in past 30 days - not including job 0 = no 1 = yes
10) Fruits : Consume Fruit 1 or more times per day 0 = no 1 = yes
11) Veggies : Consume Vegetables 1 or more times per day 0 = no 1 = yes
12) HvyAlcoholConsump : Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per
13) AnyHealthcare : Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
14) NoDocbcCost : 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 : Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
16) MentHlth : Now thinking about your mental health, which includes stress, depression, and problems with emotions
17) PhysHlth : Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30
18) DiffWalk : Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
19) Sex : 0 = female 1 = male
20) Age : 13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
21) Education : Education level (EDUCA see codebook) scale 1-6 1 = Never attended school or only kindergarten 2 = Grades
22) Income : Income scale (INCOME2 see codebook) scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

Purpose and Importance

The purpose of this article is to analyze diabetes health indicators to identify key risk factors such as BMI, Blood Pressure, Cholesterol, Age, and Walking Difficulty using a dataset from the UCI Machine Learning Repository. By leveraging Power BI for data visualization, the study aims to provide insights into trends and disparities across different demographics, supporting early detection and risk assessment. This is crucial as diabetes leads to severe health complications, and effective management relies on timely diagnosis. The findings contribute to data-driven public health policies and targeted interventions, ultimately helping to reduce the disease burden.

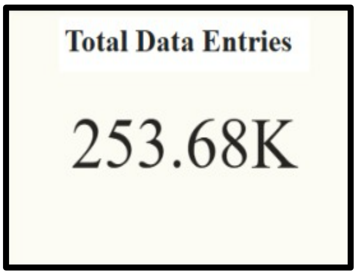
Purpose of Analysis

The purpose of this analysis is to identify and understand the key risk factors associated with diabetes using the provided dataset. By examining various health indicators such as BMI, blood pressure, cholesterol levels, physical activity, and other lifestyle factors, the goal is to identify trends and correlations that can contribute to early diagnosis and effective management of diabetes. This analysis aims to gain insights into the factors that influence diabetes prevalence across different demographics and provide actionable recommendations to support healthcare policies, interventions, and public health initiatives aimed at reducing the burden of diabetes and improving overall health outcomes.

Key Questions :

1) IDENTIFYING THE MOST IMPORTANT FEATURES RELATED TO DIABETES?
2) HOW DO THE KEY HEALTH INDICATORS (e.g. BMI, BP) IMPACT DIABETIC RISK?
3) WHICH AGE GROUP IS MOST AFFECTED BY DIABETES?

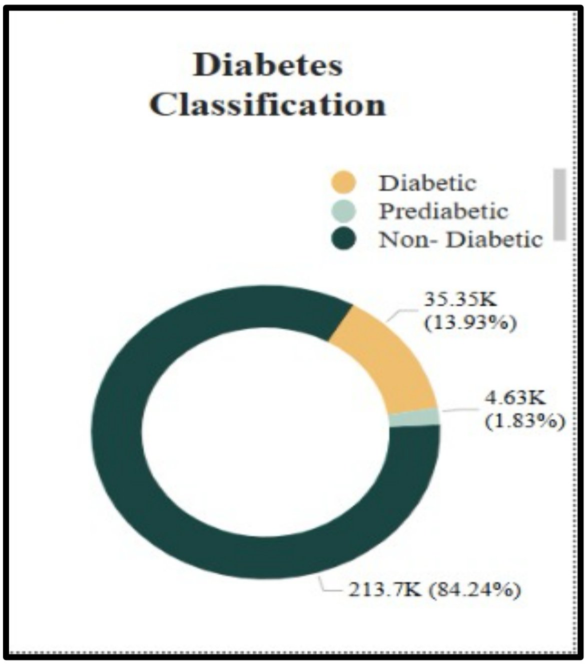
Understanding the data :



Total Data Entries

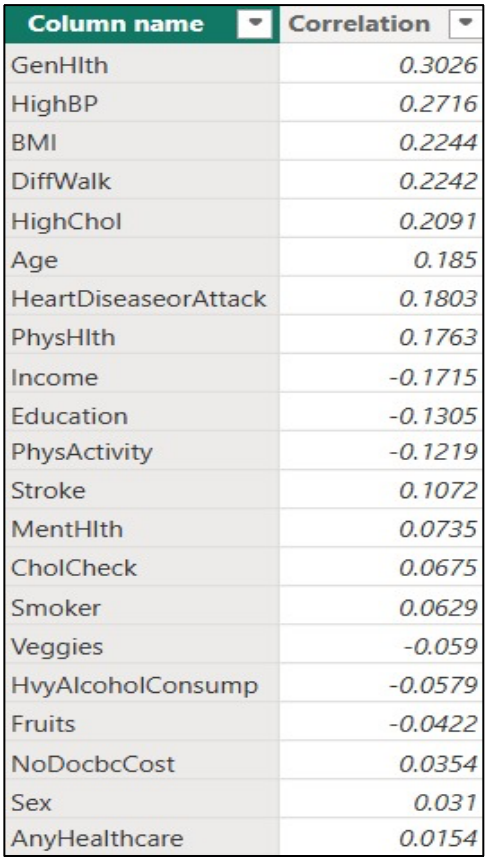
253.68K

To gain a better understanding of the dataset, I first analyzed the total number of data entries and the classification of diabetes status among them. Using Power BI, I presented the total number of entries as a **card**, which revealed that the dataset consists of **253.68k survey responses**. This helped establish the scope and size of the data available for analysis.



To further explore the distribution of diabetes status, I used a **doughnut chart** to visualize the classification of these entries into three categories: diabetic, prediabetic, and non-diabetic. The chart showed that **35.35k (13.93%)** of the entries were classified as diabetic, **4.63k (1.83%)** as prediabetic, and the majority, **213.7k (84.24%)**, as non-diabetic. This distribution highlights the class imbalance present in the dataset, with non-diabetic cases significantly outnumbering diabetic and prediabetic cases. Understanding this imbalance is crucial for accurate predictive modeling and ensures that the analysis accounts for potential biases in the data.

IDENTIFYING THE MOST IMPORTANT FEATURES RELATED TO DIABETES?



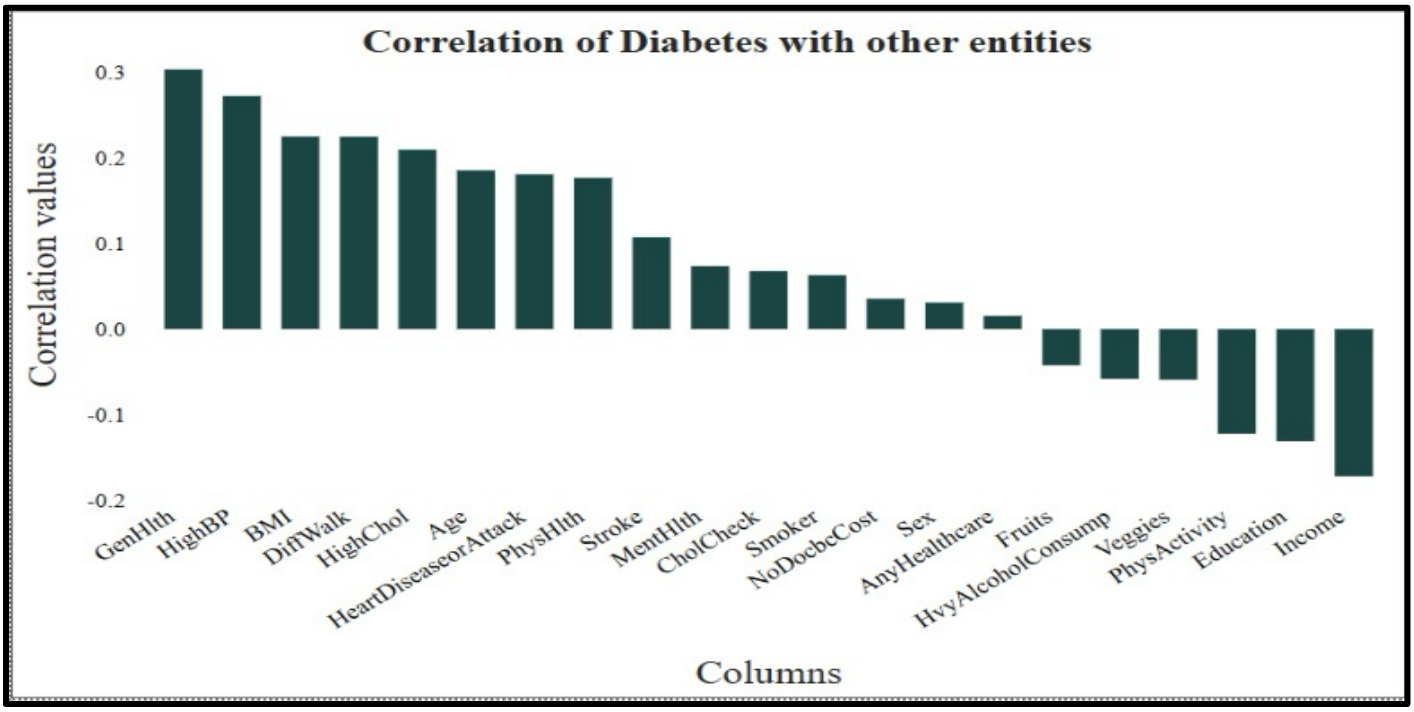
Column name	Correlation
GenHlth	0.3026
HighBP	0.2716
BMI	0.2244
DiffWalk	0.2242
HighChol	0.2091
Age	0.185
HeartDiseaseorAttack	0.1803
PhysHlth	0.1763
Income	-0.1715
Education	-0.1305
PhysActivity	-0.1219
Stroke	0.1072
MentHlth	0.0735
CholCheck	0.0675
Smoker	0.0629
Veggies	-0.059
HvyAlcoholConsump	-0.0579
Fruits	-0.0422
NoDocbcCost	0.0354
Sex	0.031
AnyHealthcare	0.0154

To identify the most important features related to diabetes, I performed a correlation analysis on the dataset. Correlation measures the strength and direction of a linear relationship between two variables, helping to determine how strongly one variable influences another. A correlation coefficient ranges from -1 to 1, where:

- **1** indicates a perfect positive correlation (both variables move in the same direction),
- **-1** indicates a perfect negative correlation (the variables move in opposite directions),
- **0** indicates no linear relationship between the variables.

I calculated the correlation between all pairs of variables in the dataset, creating a correlation matrix. This matrix highlighted which features had strong relationships with the target variable (diabetes status). For a clearer understanding, I created a table to display these correlation values, which allowed me to identify the most influential features in predicting diabetes risk.

Additionally, I visualized these correlations using a bar graph, which helped highlight the variables with the highest positive or negative correlation to diabetes. This visualization made it easier to pinpoint the most important health indicators, such as BMI, blood pressure, and cholesterol levels, that have a significant impact on the likelihood of developing diabetes. These insights can guide further analysis, predictive modeling, and the development of targeted health interventions.

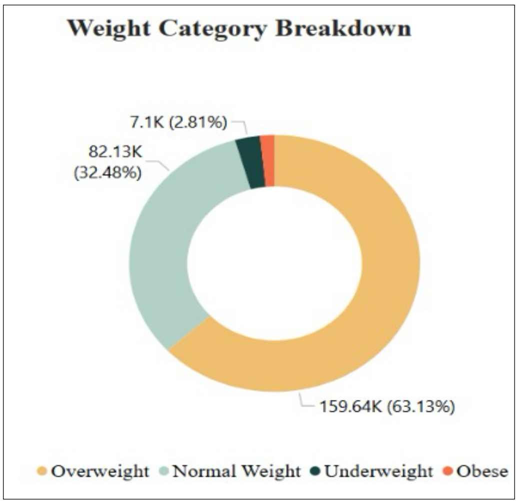


HOW DO THE KEY HEALTH INDICATORS (e.g. BMI, BP) IMPACT DIABETIC RISK?

Diabetes Risk Across Weight Distribution

To understand the relationship between weight distribution and diabetes risk, I analyzed the data by categorizing it into four weight groups: **Underweight, Normal Weight, Overweight, and Obese**.
In this analysis, the dataset included a **BMI (Body Mass Index)** column, which I utilized to categorize weight into four distinct groups. Specifically, individuals with a **BMI ≤ 19** were classified as **Underweight**, those with a **BMI between 20 and 25** were categorized as **Normal Weight**, a **BMI between 26 and 50** was considered **Overweight**, and the remaining entries were labeled as **Obese**. This categorization helped in examining the relationship between weight distribution and diabetes risk more effectively.

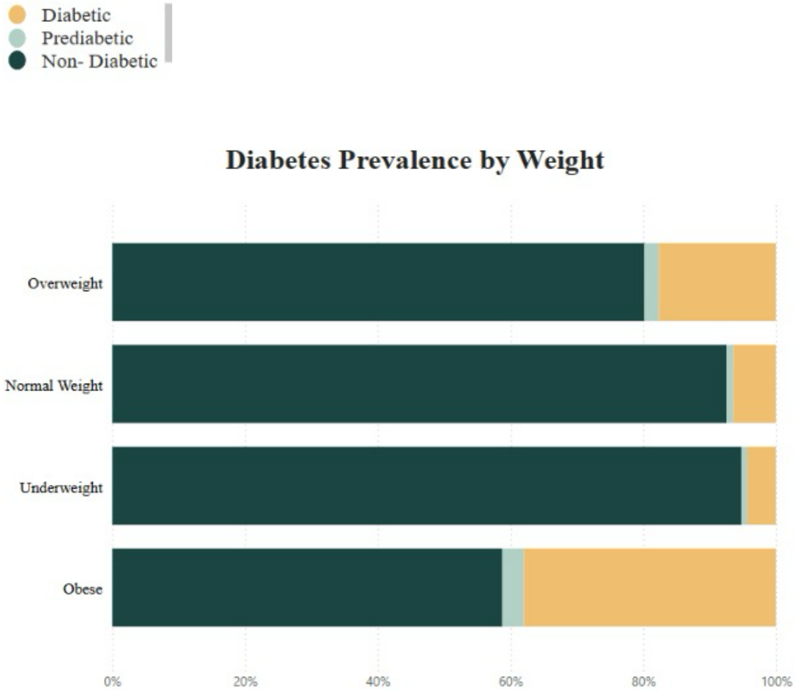
The dashboard presents this analysis using two visualizations:



1. **Weight Category Breakdown**
   The doughnut chart provides a comprehensive view of the weight distribution across the dataset. It shows that:

- **63.13%** (159.64k) of the individuals are **Overweight**.
- **32.48%** (82.13k) fall under the **Normal Weight** category.
- Only **2.81%** (7.1k) are **Underweight**, while the remaining portion is classified as **Obese**.

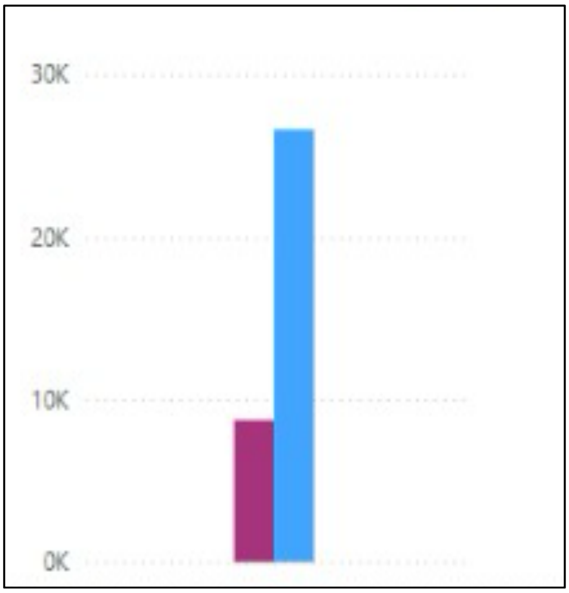
2. **Diabetes Prevalence by Weight**:



The bar chart shows the proportion of diabetic and non-diabetic cases within each weight category. It reveals that diabetes prevalence is notably higher in the **Obese** category compared to the others. Conversely, the **Underweight** group has the lowest prevalence of diabetes. This suggests a strong correlation between obesity and diabetes risk.

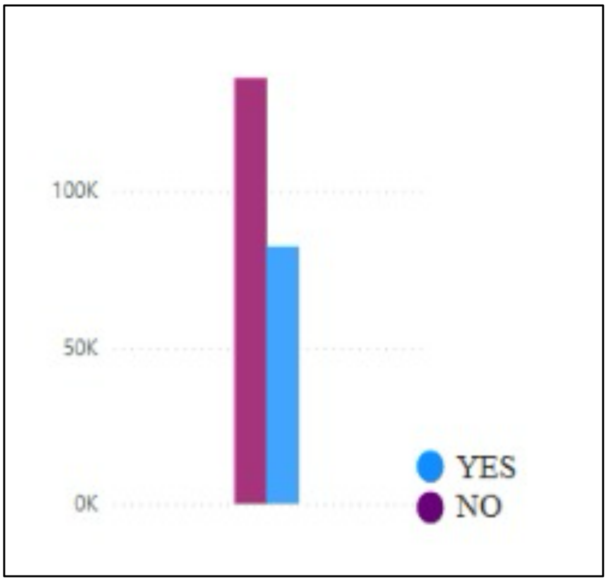
Influence of High Blood Pressure

Diabetic Data :



The bar chart illustrates the distribution of **diabetic individuals** based on their **blood pressure levels**. It compares two groups: those with **High Blood Pressure** (represented by the sky blue bar) and those with **Normal Blood Pressure** (represented by the maroon bar). The chart clearly shows that the **number of diabetic individuals with High Blood Pressure** is significantly higher than those with Normal Blood Pressure. The sky blue bar is notably taller, indicating a larger affected population.

Non - Diabetic Data:

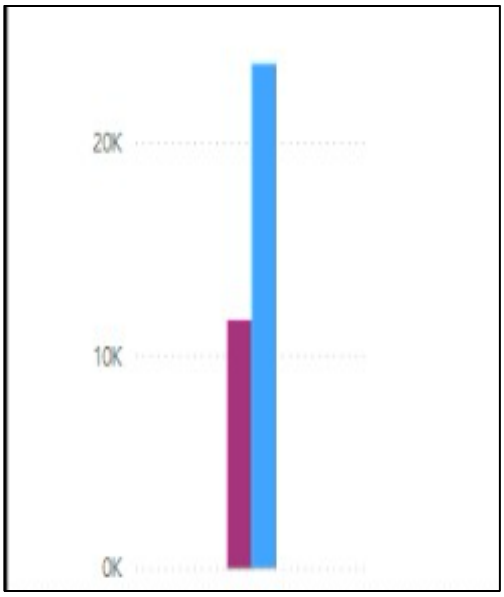


The bar chart illustrates the distribution of **non-diabetic individuals** based on their **blood pressure levels**. It compares two groups: those with **High Blood Pressure** (indicated by the blue bar labeled as "YES") and those with **Normal Blood Pressure** (indicated by the maroon bar labeled as "NO"). The maroon bar is significantly taller, indicating that the majority of non-diabetic individuals have **Normal Blood Pressure**. In contrast, the blue bar is shorter, showing a comparatively smaller group with High Blood Pressure.

The comparison of blood pressure trends between diabetic and non-diabetic individuals reveals a significant association between High Blood Pressure and diabetes. Diabetic individuals are more likely to have High Blood Pressure, underscoring the importance of monitoring and managing blood pressure to reduce diabetes-related complications. In contrast, Normal Blood Pressure is more prevalent among non-diabetic individuals, suggesting a protective factor against diabetes. This contrast highlights the crucial role of blood pressure management in diabetes prevention and emphasizes the need for lifestyle interventions to maintain healthy blood pressure levels.

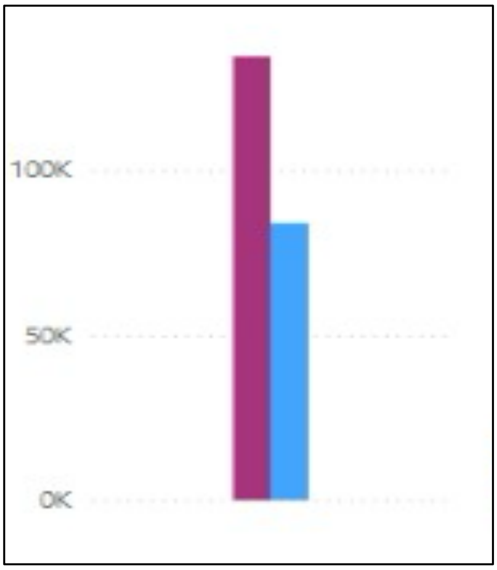
Influence of High Blood Pressure

Diabetic Data :



For diabetic individuals, the graph shows that **high cholesterol** levels are more prevalent, as indicated by the taller blue bar. In contrast, **normal cholesterol** levels are less common, represented by the shorter purple bar. This pattern suggests a strong association between diabetes and elevated cholesterol levels, emphasizing the importance of effective cholesterol management for **diabetic patients**. It highlights the need for regular monitoring and lifestyle changes to reduce the risk of cardiovascular complications associated with high cholesterol in diabetes.

Non - Diabetic Data:

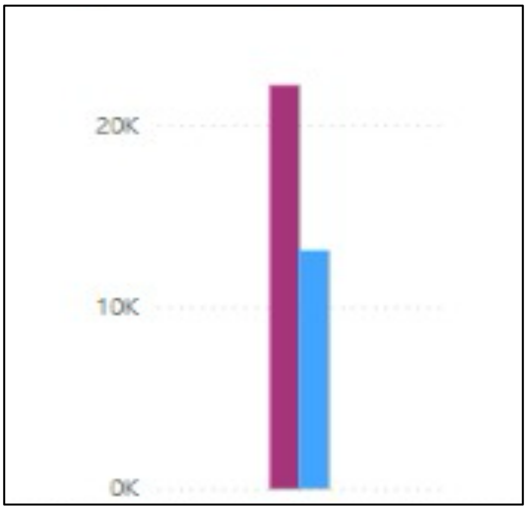


For non-diabetic individuals, the graph reveals that **normal cholesterol** levels are more common, as shown by the taller purple bar. However, a significant portion of non-diabetics also has **high cholesterol**, indicated by the shorter blue bar. This suggests that while **normal cholesterol** is predominant, **high cholesterol** is still present among **non-diabetic populations**. It underscores the importance of maintaining healthy cholesterol levels as a preventive measure, even for those without diabetes, to lower the risk of cardiovascular diseases.

Comparing both graphs, it is evident that **high cholesterol** is more common among **diabetic individuals** than **non-diabetics**. This reinforces the link between diabetes and elevated cholesterol levels, underlining the importance of regular cholesterol screening and lifestyle modifications for diabetic patients. The contrast also suggests that maintaining normal cholesterol levels may serve as a protective factor against diabetes-related complications.

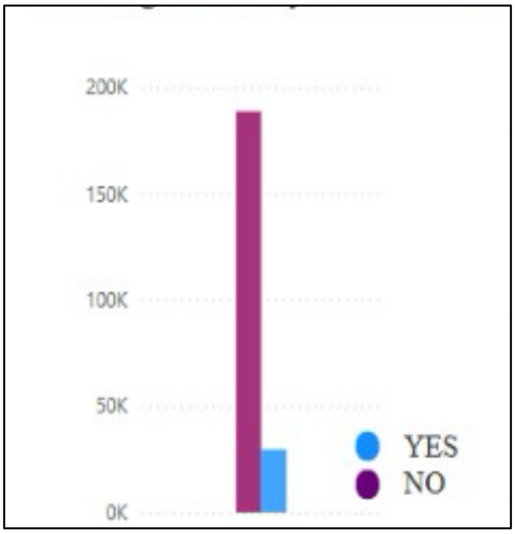
Walking Difficult Distribution among Diabetic and Non-Diabetic individuals

Diabetic Data :



- Approximately 22,000 individuals without walking difficulties (shown in purple)
- About 13,000 individuals reporting walking difficulties (shown in blue)
- This indicates that roughly 37% of diabetic individuals experience walking difficulties

Non - Diabetic Data:



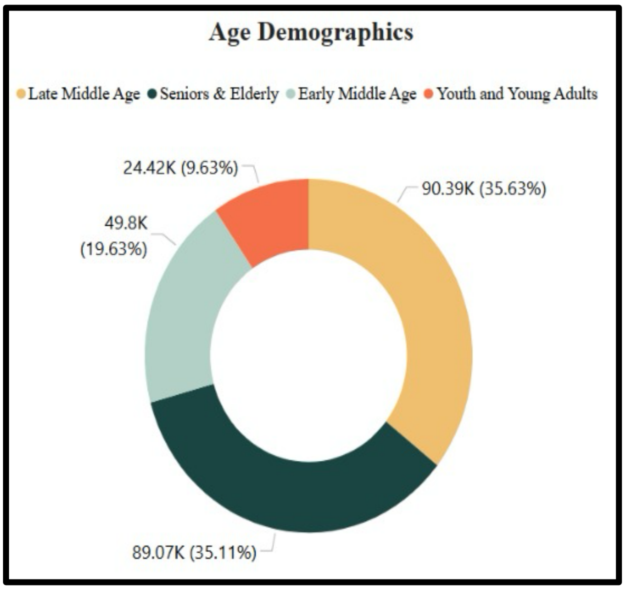
- Approximately 180,000 individuals without walking difficulties (shown in purple)
- About 35,000 individuals reporting walking difficulties (shown in blue)
- This translates to roughly 16% of non-diabetic individuals experiencing walking difficulties

Key Insights:

1. The proportion of individuals experiencing walking difficulties is significantly higher in the diabetic population (37%) compared to the non-diabetic population (16%)
2. This suggests a potential correlation between diabetes and mobility challenges
3. The stark difference in percentages (more than double) indicates that diabetes might be a contributing factor to walking difficulties, possibly due to:
   - Diabetes-related complications like neuropathy
   - Circulation problems in lower extremities
   - Overall impact on physical mobility and strength

This analysis reveals that individuals with diabetes are more likely to experience walking difficulties compared to their non-diabetic counterparts, highlighting the importance of mobility management and physical activity programs in diabetes care.

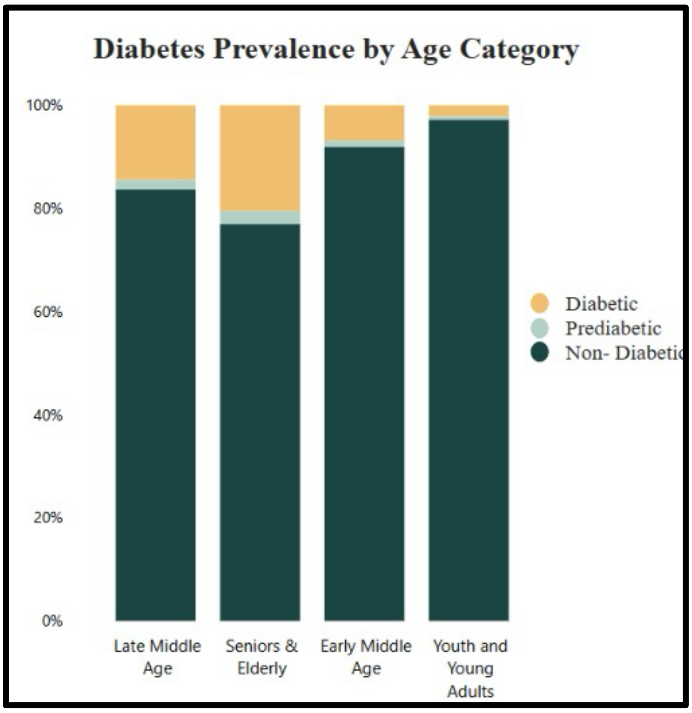
WHICH AGE GROUP IS MOST AFFECTED BY DIABETES?



This donut chart illustrates the overall distribution of age categories in the dataset, showing:

- Late Middle Age represents 35.63% (90.39K individuals)
- Seniors & Elderly make up 35.11% (89.07K individuals)
- Early Middle Age comprises 19.63% (49.8K individuals)
- Youth and Young Adults account for 9.63% (24.42K individuals)

The chart reveals that Late Middle Age and Seniors & Elderly together constitute over 70% of the dataset, indicating a significant representation of older age groups in the study.



This stacked bar chart shows the proportion of diabetic, prediabetic, and non-diabetic individuals across different age groups:

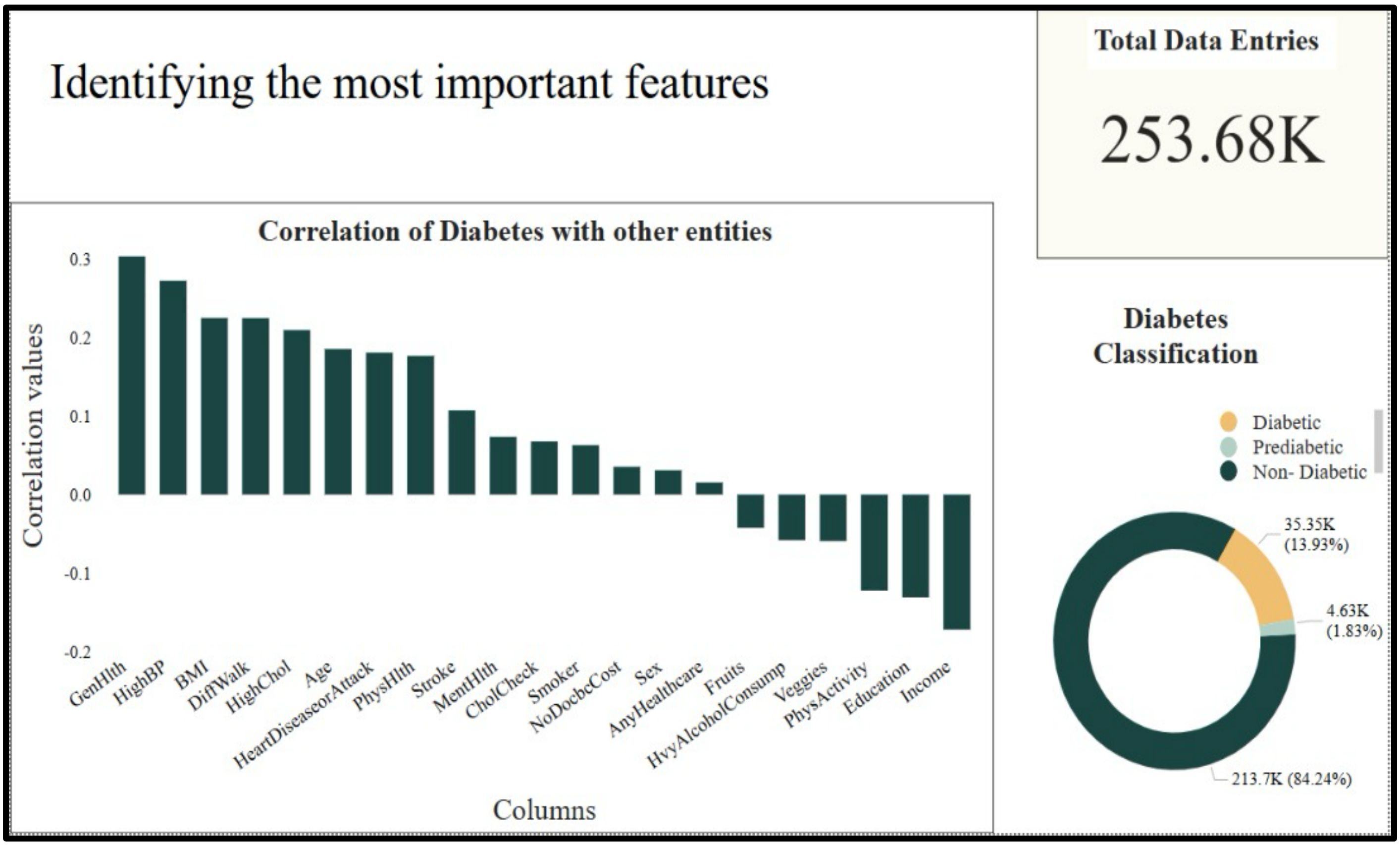
- Seniors & Elderly show the highest proportion of diabetes, with approximately 40% of individuals being diabetic
- Late Middle Age follows with roughly 15% diabetic cases
- Early Middle Age and Youth/Young Adults show progressively lower rates of diabetes
- The proportion of non-diabetics (shown in dark green) increases as age decreases, with Youth and Young Adults having the highest percentage of non-diabetics

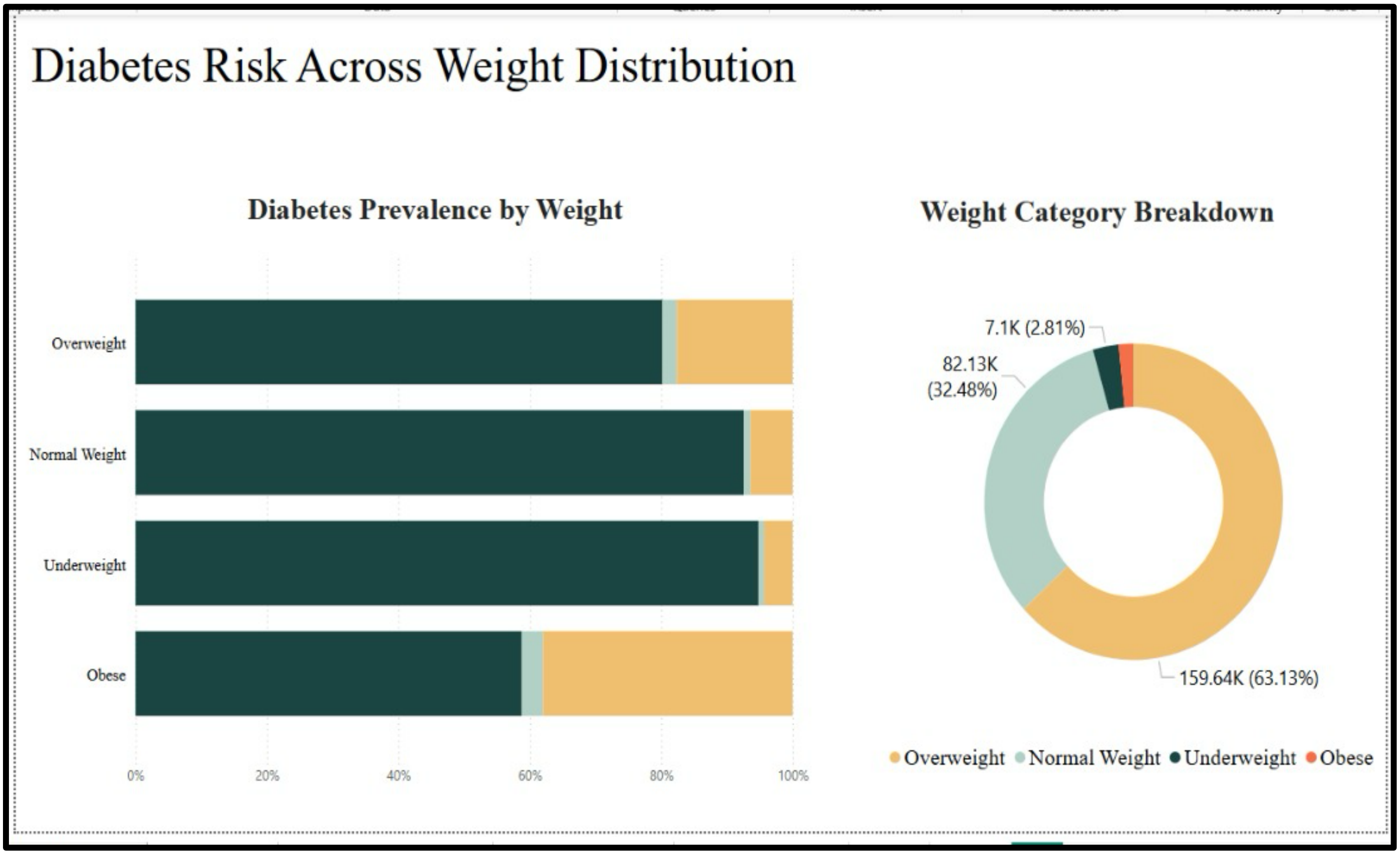
The Seniors & Elderly age group is most affected by diabetes, showing both:

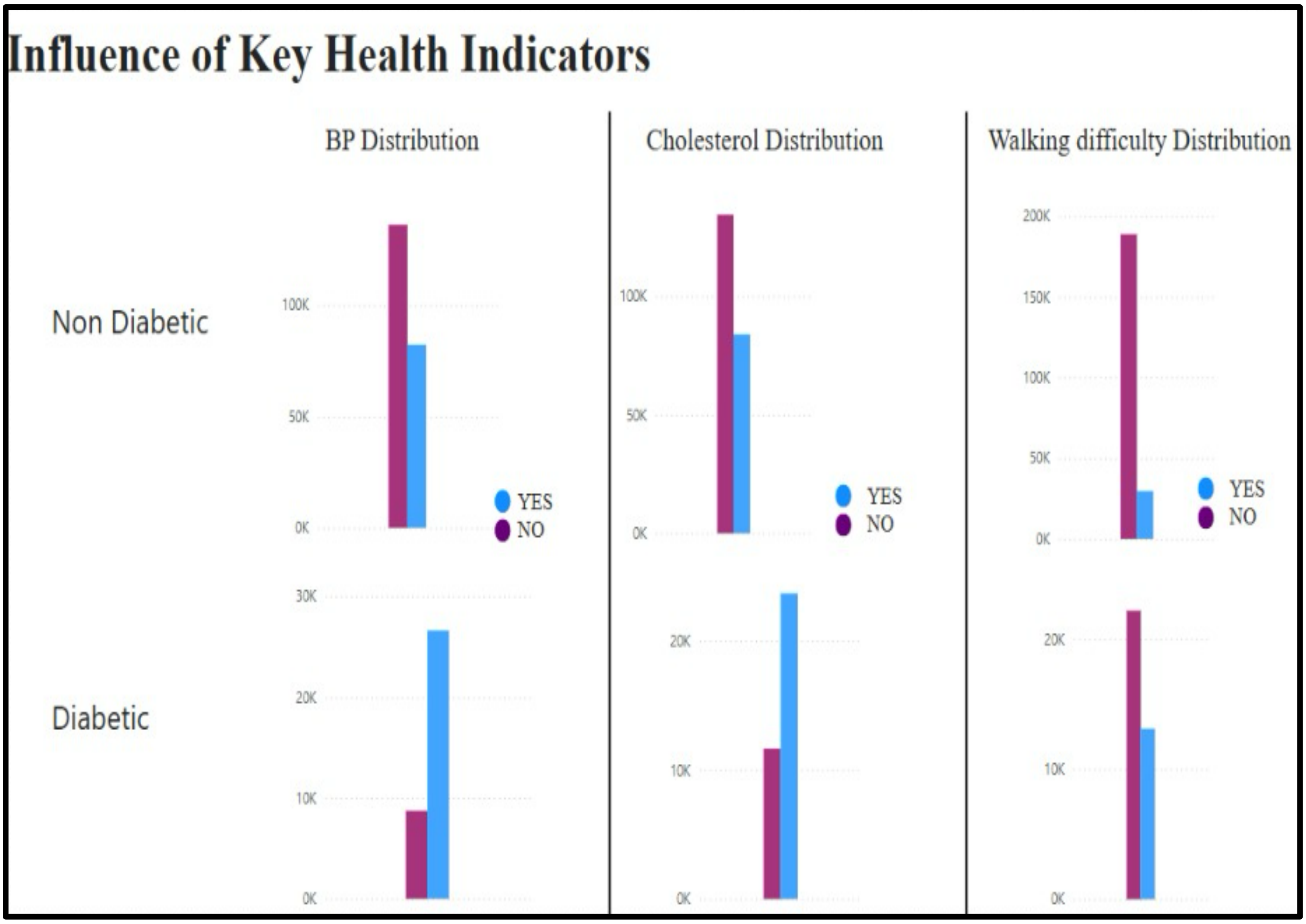
1. High representation in the population (35.11% of total subjects)
2. Highest proportion of diabetic cases (approximately 40% of their age group)

This indicates a clear correlation between advancing age and diabetes prevalence, with the Seniors & Elderly population being particularly vulnerable to the condition. The combination of these visualizations effectively demonstrates that while Late Middle Age and Seniors & Elderly groups have similar population sizes, the Seniors & Elderly group has a notably higher diabetes prevalence rate, making them the most affected age group in the study.

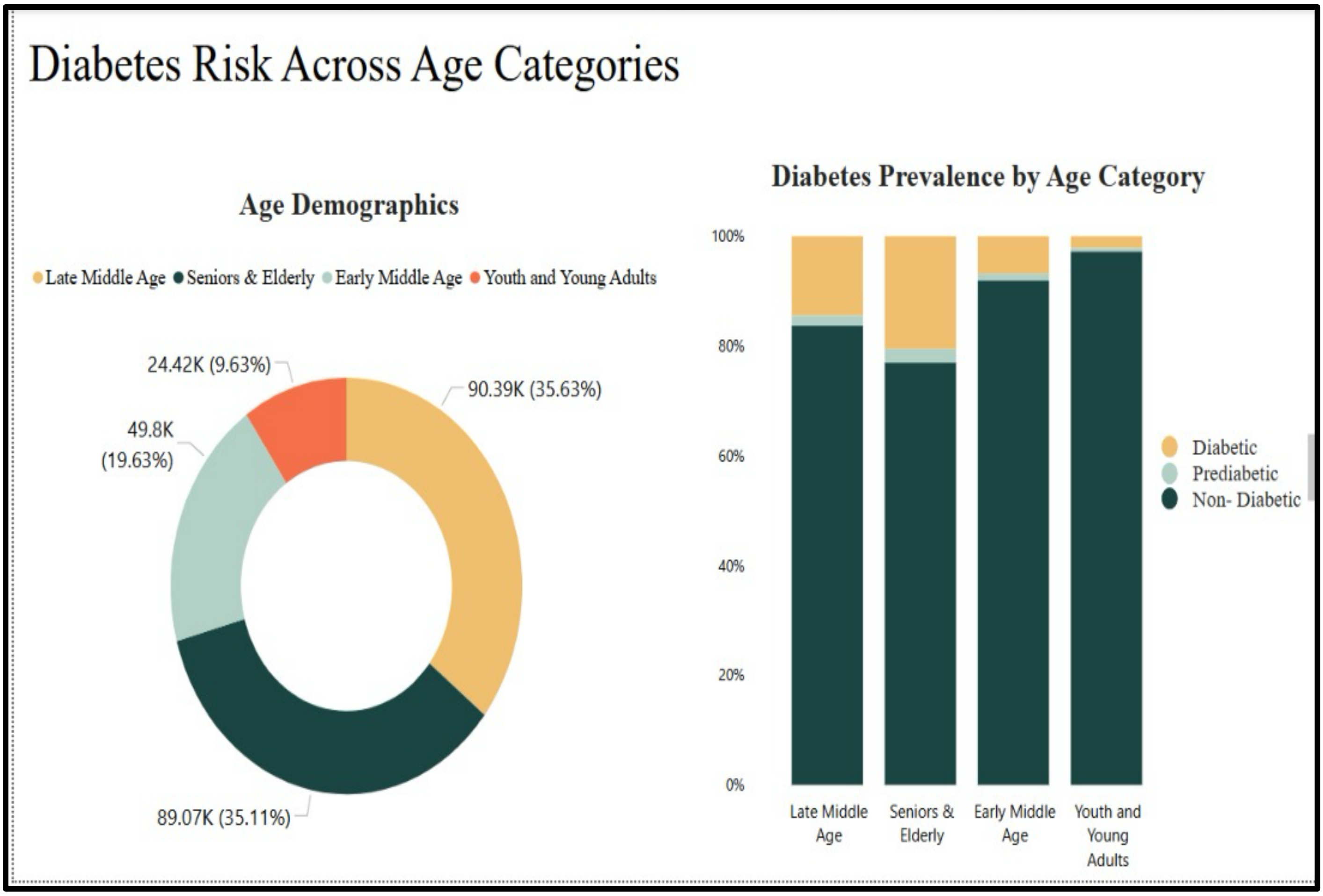
Dashboards:


Identifying the most important features
Total Data Entries
253.68K
Correlation of Diabetes with other entities
Correlation values
0.3
0.2
0.1
0.0
-0.1
-0.2
GenHlth
HighBP
BMI
DiffWalk
HighChol
Age
HeartDiseaseorAttack
PhysHlth
Stroke
MentHlth
CholCheck
Smoker
NoDocbcCost
Sex
AnyHealthcare
Fruits
HvyAlcoholConsump
Veggies
PhysActivity
Education
Income
Columns
Diabetes Classification
Diabetic
Prediabetic
Non- Diabetic
35.35K (13.93%)
4.63K (1.83%)
213.7K (84.24%)


Diabetes Risk Across Weight Distribution
Diabetes Prevalence by Weight
Overweight
Normal Weight
Underweight
Obese
0%
20%
40%
60%
80%
100%
Weight Category Breakdown
7.1K (2.81%)
82.13K (32.48%)
159.64K (63.13%)
Overweight
Normal Weight
Underweight
Obese



Influence of Key Health Indicators



Diabetes Risk Across Age Categories

Conclusion

This diabetes dashboard provides valuable insights into the key risk factors associated with diabetes, highlighting the significant influence of BMI, Blood Pressure, Cholesterol, Age, and Walking Difficulty. The analysis reveals that older age groups, individuals with high BMI, and those experiencing mobility issues show notably higher rates of diabetes. These findings emphasize the importance of monitoring these health metrics to identify at-risk populations and implement targeted preventive measures.

Additionally, the strong correlation between high blood pressure and cholesterol levels with diabetes underscores the need for effective lifestyle interventions and medical management to reduce diabetes-related complications. While Income and Education were found to be negatively correlated with diabetes, they were excluded from the analysis to maintain a focused examination of more direct health-related metrics. By prioritizing the most impactful risk factors, this dashboard offers practical insights for healthcare professionals and policymakers to develop tailored strategies for diabetes prevention and management.