SHARNDEEP KAUR
MENTOR: JUNAID QAZI

AN ARTICLE

ON

“ANALYSIS OF DIABETES HEALTH INDICATORS (DATASET)”

SUBMITTED BY:

SHARNDEEP KAUR

INDEX:

SHARNDEEP KAUR
MENTOR: JUNAID QAZI

ABSTRACT:

This study examines diabetes health indicators using a dataset from the UCI Machine Learning Repository to identify trends, risk factors, and disparities among different demographics in the U.S. Diabetes is a chronic disease that can lead to severe complications, including heart disease, vision impairment, and kidney failure. The research focuses on recognizing key risk factors, promoting early diagnosis, and supporting data-driven public health decisions. A Power BI dashboard was created to visualize critical factors such as BMI, blood pressure, cholesterol, age, and walking difficulty. The analysis highlights higher diabetes rates among older individuals with elevated BMI and mobility challenges. These findings offer valuable insights for targeted diabetes interventions and policy enhancements, improving prevention and management strategies.

INTRODUCTION:

Diabetes is a chronic condition that affects millions in the U.S., often resulting in severe complications such as heart disease, vision impairment, and kidney failure. Early detection and effective management are essential in mitigating its impact. This study examines 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. Leveraging Power BI for data visualization, the research uncovers trends and disparities across various demographics, offering insights to inform public health decisions and guide targeted interventions for diabetes prevention and management.

DATASET OVERVIEW:

The Diabetes Health Indicators Dataset is derived from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), an annual health-related telephone survey conducted by the Centers for Disease Control and Prevention (CDC) in the United States. It comprises 253,680 cleaned survey responses, offering valuable insights into diabetes prevalence, risk factors, and health disparities across various demographics. The dataset classifies individuals into three categories based on the target variable **Diabetes_012**:

- **0**: No diabetes or gestational diabetes (during pregnancy)
- **1**: Prediabetes
- **2**: Diabetes

Notably, the dataset is imbalanced, with most respondents classified as non-diabetic. 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 play a vital role in 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:

This article examines 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. Leveraging Power BI for data visualization, the study uncovers trends and disparities across different demographics, aiding in early detection and risk assessment. Since diabetes can lead to serious health complications, timely diagnosis is crucial for effective management. The findings provide valuable insights to inform data-driven public health policies and targeted interventions, ultimately helping to reduce the overall disease burden.

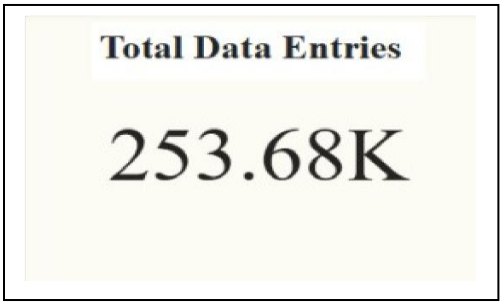
PURPOSE OF ANALYSIS:

This analysis seeks to identify and understand the key risk factors associated with diabetes using the provided dataset. By evaluating health indicators such as BMI, blood pressure, cholesterol levels, physical activity, and other lifestyle factors, the study aims to uncover trends and correlations that support early diagnosis and effective disease management. Additionally, it explores how diabetes prevalence varies across different demographics, providing actionable insights to inform healthcare policies, interventions, and public health initiatives. Ultimately, the findings contribute to reducing the burden of diabetes and enhancing overall health outcomes.

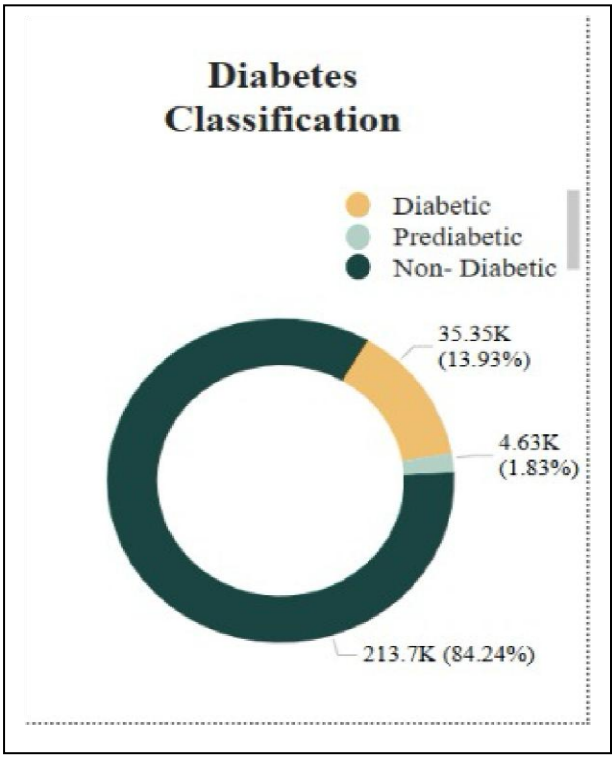
Key Questions:

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:



To gain a clearer understanding of the dataset, I first examined the total number of data entries and the distribution of diabetes classifications. Using Power BI, I presented the total number of entries as a card, showing that the dataset includes 253.68k survey responses. This helped establish the dataset's scale and provided a foundation for further analysis



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

.

IDENTIFYING THE MOST IMPORTANT FEATURES RELATED TO DIABETES?

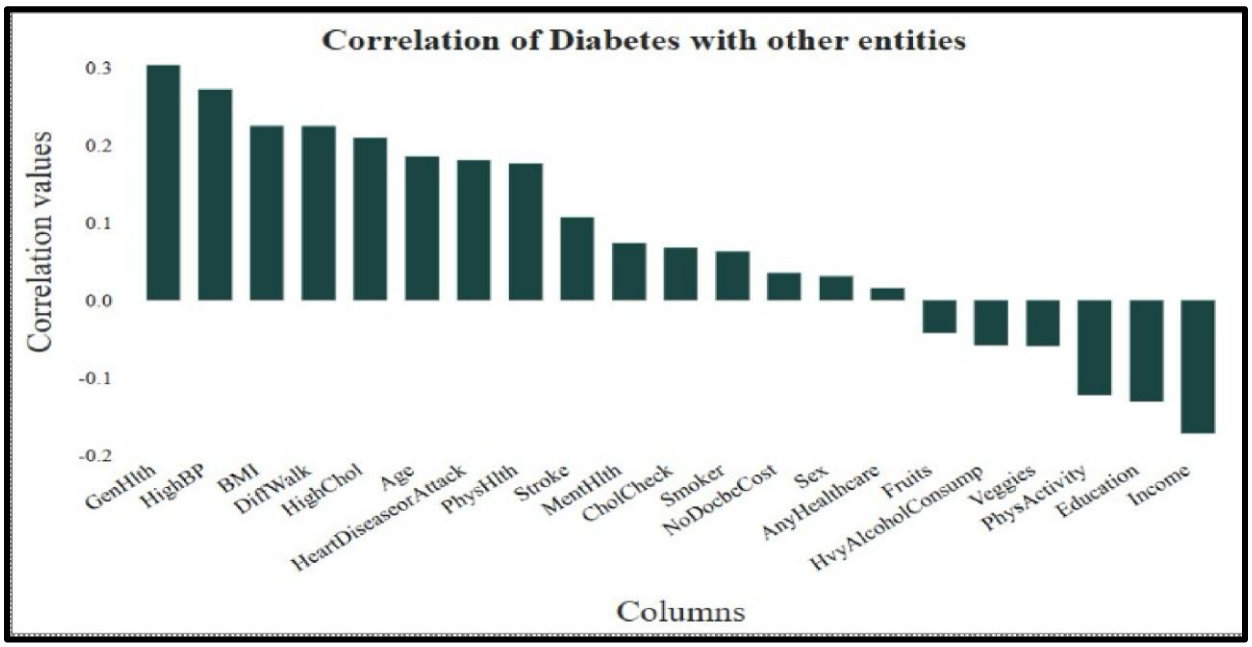
Column name	Correlation
GenHlth	0.3026
HighBP	0.2716
BMI	0.2244
DiffWalk	0.2242
HighChol	0.2091
Age	0.185
HeartDiseaseorAttack	0.1803
PhysHlth	0.1763
Income	-0.1715
Education	-0.1305
PhysActivity	-0.1219
Stroke	0.1072
MentHlth	0.0735
CholCheck	0.0675
Smoker	0.0629
Veggies	-0.059
HvyAlcoholConsump	-0.0579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

To identify the key features associated with diabetes, I conducted a correlation analysis on the dataset. Correlation measures the strength and direction of a linear relationship between two variables, helping to assess how one variable influences another. The correlation coefficient ranges from -1 to 1, where:

- 1 represents a perfect positive correlation (both variables change in the same direction),
- -1 represents 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 and created a correlation matrix. This matrix highlighted the features most strongly related to the target variable (diabetes status). To make the results more accessible, I created a table displaying these correlation values, which helped identify the most significant features for predicting diabetes risk.

I also visualized these correlations using a bar graph, which emphasized the variables with the strongest positive or negative correlations to diabetes. This visualization made it easier to identify key health indicators, such as BMI, blood pressure, and cholesterol levels, that significantly influence the likelihood of developing diabetes. These insights can inform further analysis, predictive modeling, and the creation of targeted health interventions.



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

Diabetes Risk Across Weight Distribution

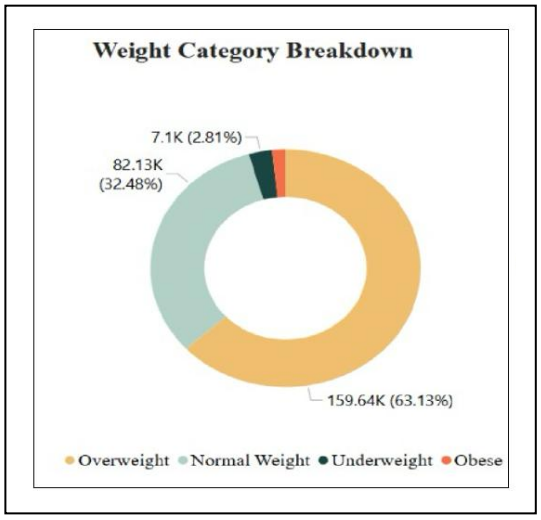
To explore the relationship between weight distribution and diabetes risk, I categorized the data into four weight groups: Underweight, Normal Weight, Overweight, and Obese.

The dataset included a BMI (Body Mass Index) column, which I used to classify individuals into distinct weight categories. Specifically:

- BMI ≤ 19 was classified as **Underweight**,
- BMI between 20 and 25 as **Normal Weight**,
- BMI between 26 and 50 as **Overweight**,
- Remaining entries were labeled as **Obese**.

This classification allowed for a more effective analysis of the relationship between weight distribution and diabetes risk.

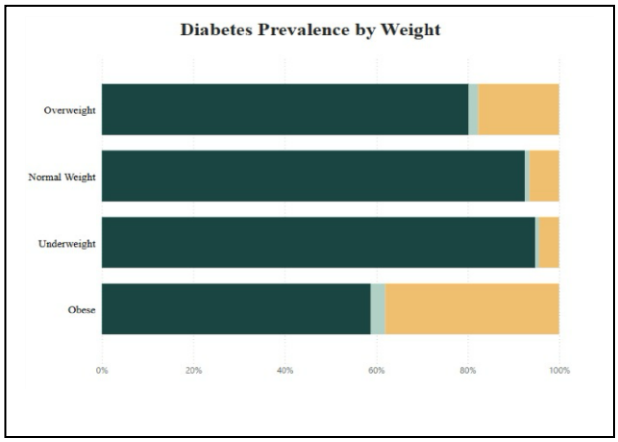
The dashboard presents this analysis using two visualizations



Weight Category Breakdown

The doughnut chart provides a comprehensive view of the weight distribution across the dataset. It shows that:

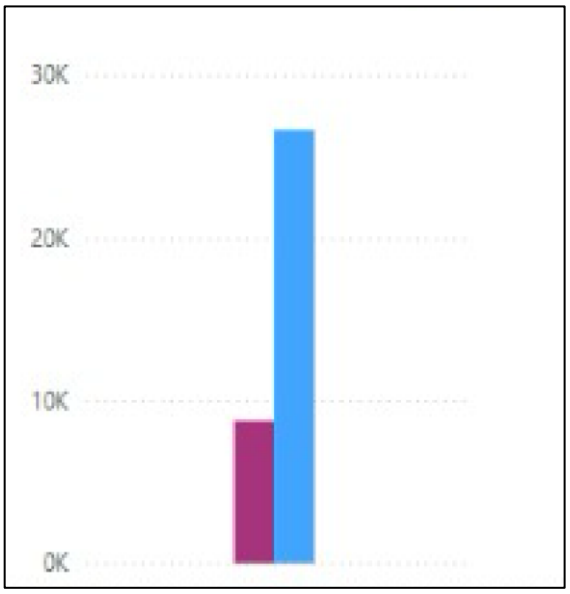
- **63.13%** (159.64k) of the individuals are **Overweight**.
- **32.48%** (82.13k) fall under the **Normal Weight** category.
- Only **2.81%** (7.1k) are **Underweight**, while the remaining portion is classified as **Obese**.



The bar chart illustrates the proportion of diabetic and non-diabetic cases across different weight categories. It shows that diabetes prevalence is significantly higher among individuals in the **Obese** category compared to other groups. In contrast, the **Underweight** group has the lowest diabetes prevalence. This indicates a strong correlation between obesity and an increased risk of diabetes.

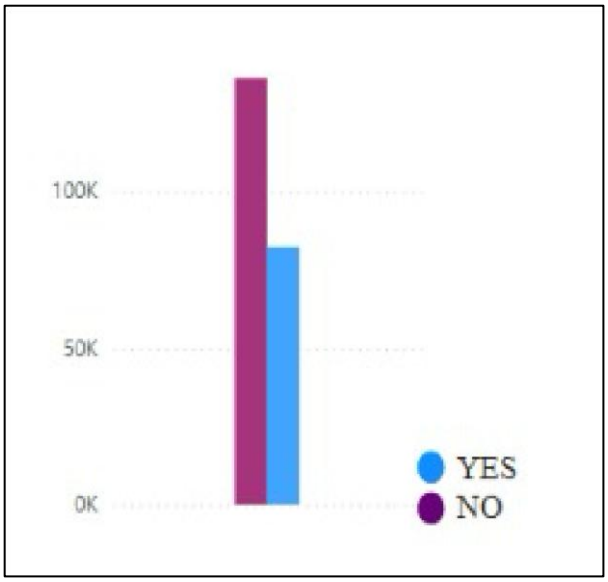
INFLUENCE ON BLOOD PRESSURE:

DiabetesData :



The bar chart depicts the distribution of diabetic individuals based on their blood pressure levels, comparing 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 highlights that diabetes is significantly more prevalent among individuals with High Blood Pressure. The sky blue bar is noticeably taller, indicating a larger affected population in this group.

Non - Diabetes Data:



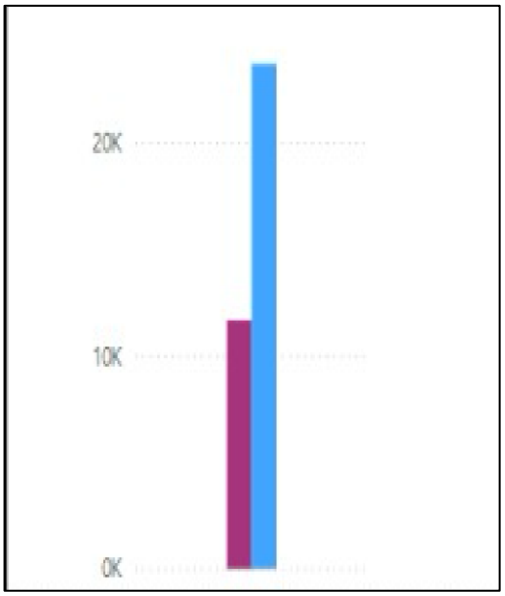
The bar chart represents the distribution of non-diabetic individuals based on their blood pressure levels, comparing two groups: those with **High Blood Pressure** (represented by the blue bar labeled "YES") and those with **Normal Blood Pressure** (represented by the maroon bar labeled "NO"). The maroon bar is noticeably taller, indicating that most non-diabetic individuals have Normal Blood Pressure. In contrast, the shorter blue bar reflects a smaller proportion of non-diabetic individuals with High Blood Pressure.

The comparison of blood pressure trends between diabetic and non-diabetic individuals reveals a strong association between **High Blood Pressure** and diabetes. Diabetic individuals are more likely to have High Blood Pressure, emphasizing the importance of monitoring and managing blood pressure to reduce diabetes-related complications. Conversely, **Normal Blood Pressure** is more common among non-diabetic individuals, suggesting it may serve as a protective factor against diabetes. This contrast underscores the critical role of blood pressure management in diabetes prevention and highlights the need for lifestyle interventions to maintain healthy blood pressure levels.

SHARNDEEP KAUR
MENTOR: JUNAID QAZI

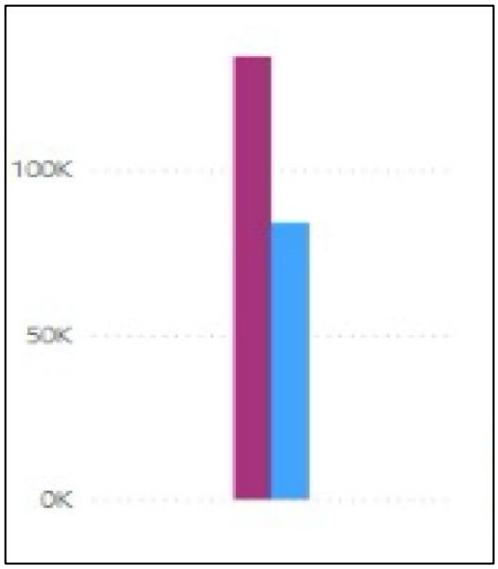
INFLUENCE OF CHOLESTROL:

Diabetes Data:



For diabetic individuals, the graph indicates that **high cholesterol levels** are more prevalent, as shown by the taller blue bar, while **normal cholesterol levels** are less common, represented by the shorter purple bar. This pattern suggests a strong link 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 modifications to reduce the risk of cardiovascular complications associated with high cholesterol in diabetes.

Non - Diabetes Data:



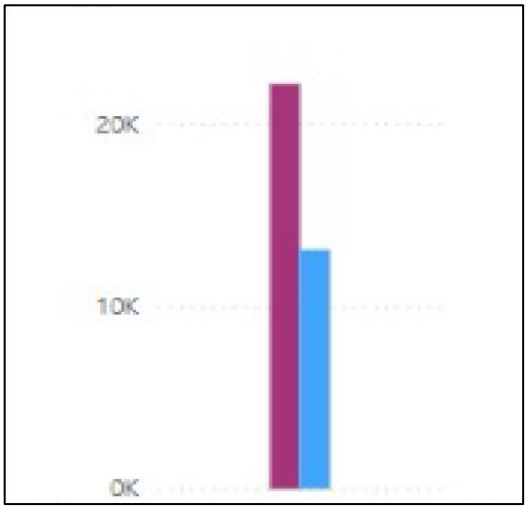
For non-diabetic individuals, the graph shows that **normal cholesterol levels** are more prevalent, as indicated by the taller purple bar. However, a notable portion of non-diabetics also has **high cholesterol**, represented by the shorter blue bar. This suggests that while normal cholesterol is more common, high cholesterol remains present among non-diabetic populations. It highlights the importance of maintaining healthy cholesterol levels as a preventive measure, even for those without diabetes, to reduce the risk of cardiovascular diseases.

A comparison of both graphs clearly shows that **high cholesterol** is more prevalent among diabetic individuals than non-diabetics. This reinforces the connection between diabetes and elevated cholesterol levels, emphasizing the need for regular cholesterol screening and lifestyle modifications for diabetic patients. Additionally, the contrast suggests that maintaining **normal cholesterol levels** may act as a protective factor against diabetes-related complications.

SHARNDEEP KAUR
MENTOR: JUNAID QAZI

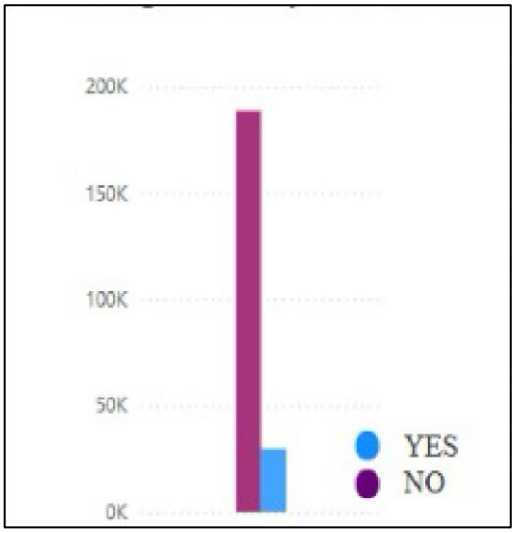
WALKING DIFFICULTY IN DIABETIC AND NON- DIABETIC INDIVIDUALS:

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:



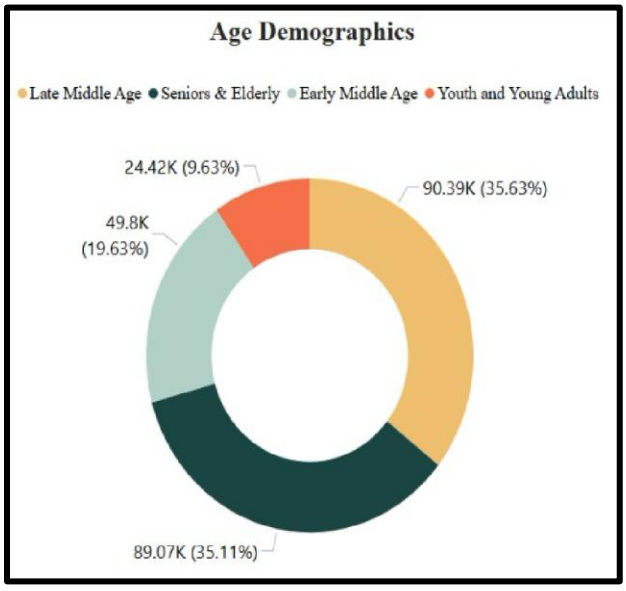
- Approximately 180,000 individuals without walking difficulties (shown in purple)
- About 35,000 individuals reporting walking difficulties (shown in blue)
- This translates to roughly 16% of non-diabetic individuals experiencing walking difficulties

Key Insights:

1. **Walking difficulties** are significantly more common among diabetic individuals (37%) compared to non-diabetics (16%).
2. This suggests a possible correlation between **diabetes and mobility challenges**.
3. The substantial difference (more than double) indicates that diabetes may contribute to walking difficulties, potentially due to:
    - **Diabetes-related complications** such as neuropathy
    - **Circulation issues** affecting the lower extremities
    - **Reduced physical mobility and strength** associated with the condition

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.

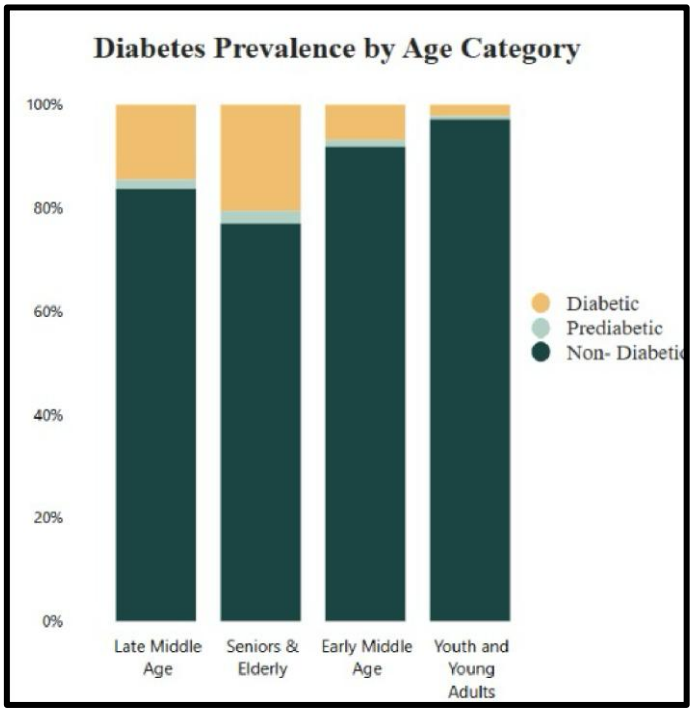
WHICH AGE GROUP IS MOST AFFECTED BY DIABETES?



This donut chart illustrates the distribution of age categories within the dataset, highlighting:

- **Late Middle Age**: 35.63% (90.39K individuals)
- **Seniors & Elderly**: 35.11% (89.07K individuals)
- **Early Middle Age**: 19.63% (49.8K individuals)
- **Youth & Young Adults**: 9.63% (24.42K individuals)

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



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

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

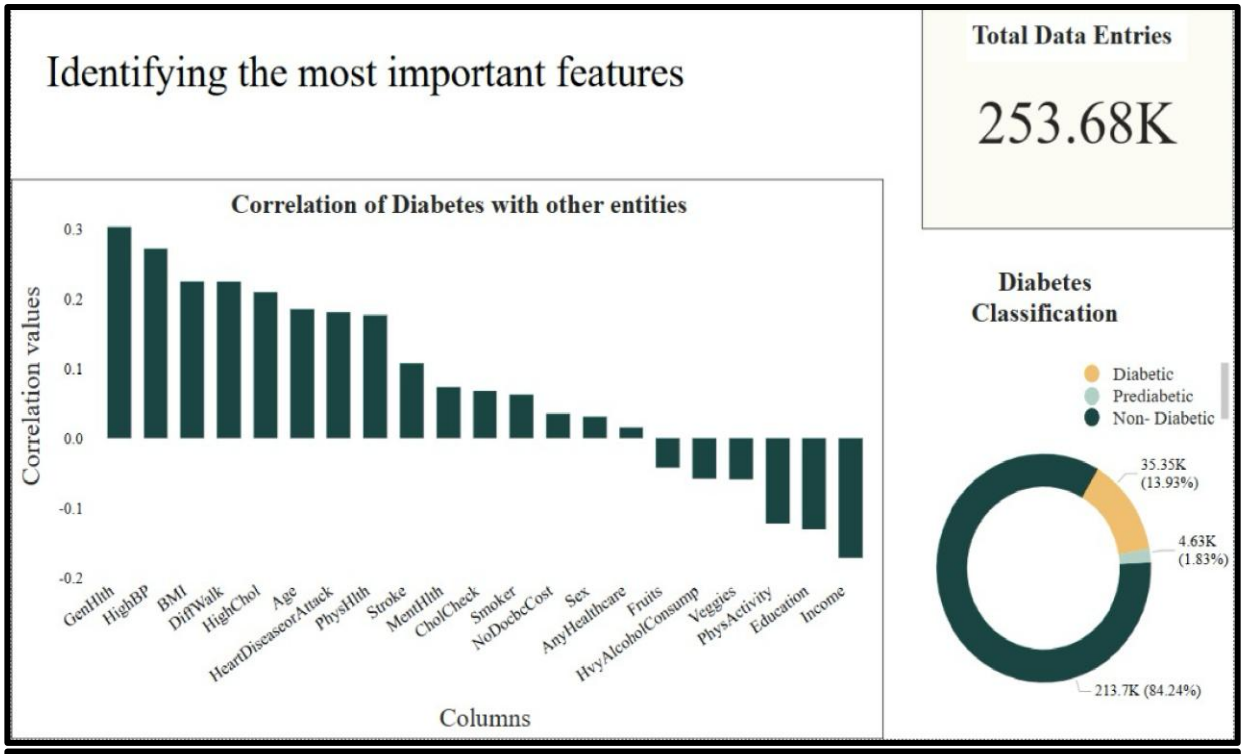
The **Seniors & Elderly** group is the most affected by diabetes, as they:

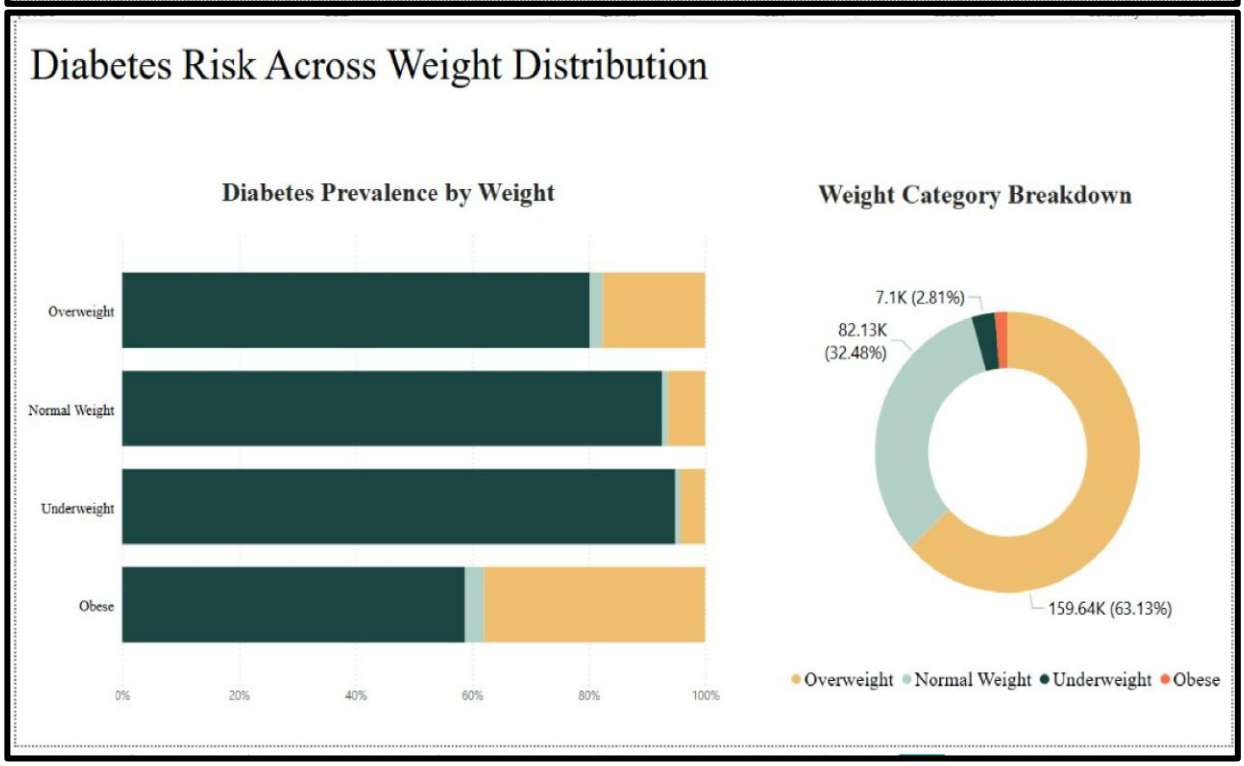
1. Represent a significant portion of the dataset (**35.11% of total subjects**).
2. Have the highest proportion of diabetic individuals (**around 40% within their age group**).

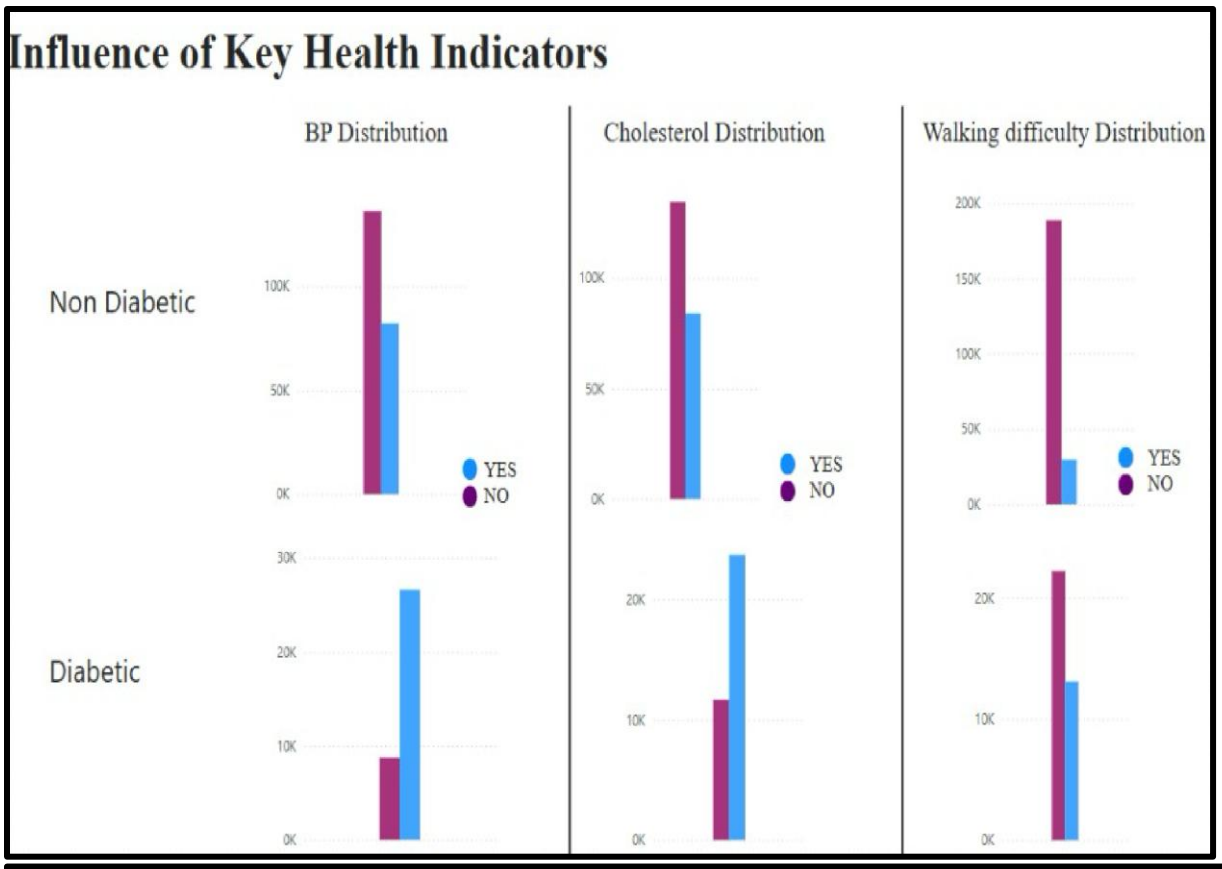
This highlights a clear correlation between **age and diabetes prevalence**, with **older individuals being more vulnerable** to the condition. The visualizations also emphasize that while **Late Middle Age and Seniors & Elderly** groups are similarly represented in the dataset, the **Seniors & Elderly group has a significantly higher diabetes prevalence**, making them the most impacted age group in the study.

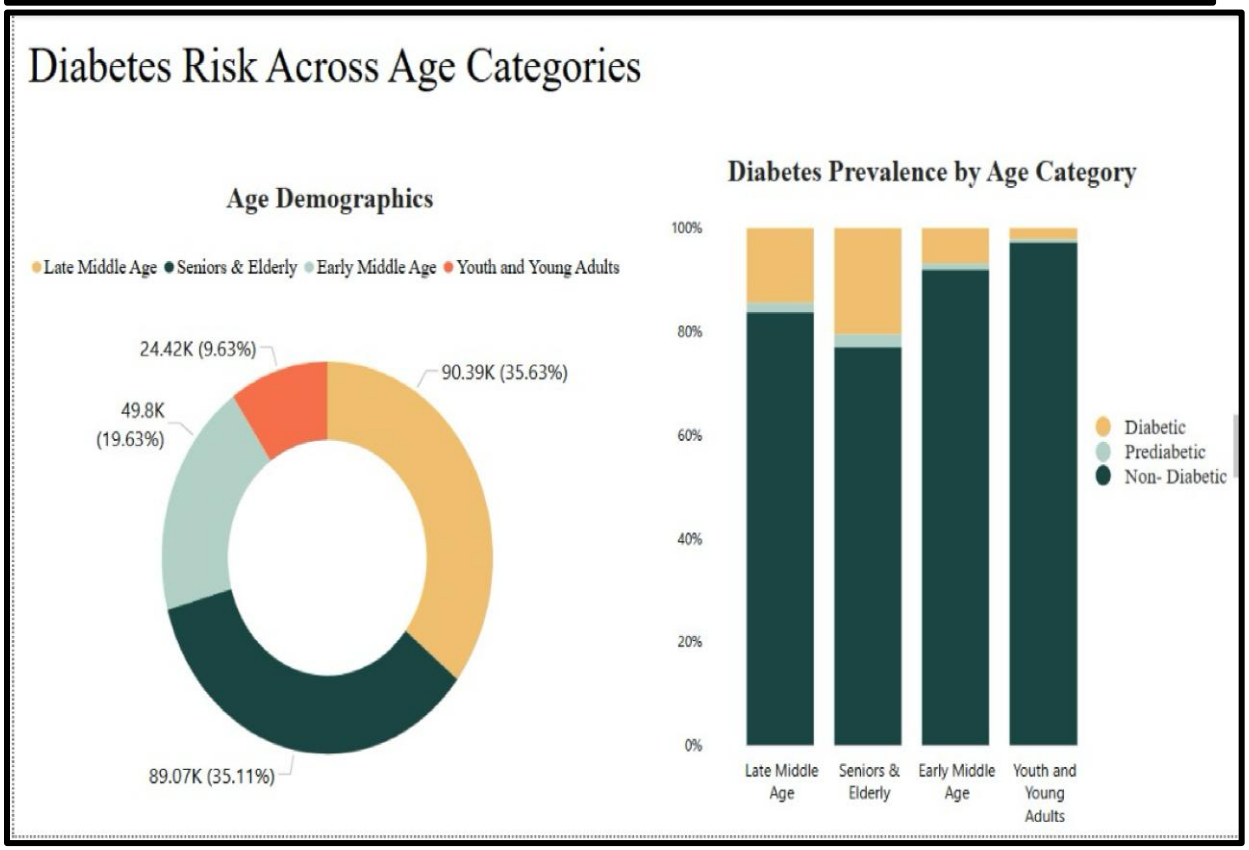
SHARNDEEP KAUR
MENTOR: JUNAID QAZI

Dashboards:






Influence of Key Health Indicators
BP Distribution
Cholesterol Distribution
Walking difficulty Distribution
Non Diabetic
Diabetic
YES
NO
100K
50K
0K
200K
150K
30K
20K
10K


Diabetes Risk Across Age Categories
Age Demographics
Late Middle Age
Seniors & Elderly
Early Middle Age
Youth and Young Adults
24.42K (9.63%)
90.39K (35.63%)
49.8K
(19.63%)
89.07K (35.11%)
Diabetes Prevalence by Age Category
100%
80%
60%
40%
20%
0%
Late Middle Age
Seniors & Elderly
Early Middle Age
Youth and Young Adults
Diabetic
Prediabetic
Non- Diabetic

Conclusion

This **diabetes dashboard** provides valuable insights into the key risk factors contributing to diabetes, emphasizing the significant impact of **BMI, Blood Pressure, Cholesterol, Age, and Walking Difficulty**. The analysis highlights that **older adults, individuals with high BMI, and those with mobility challenges** exhibit higher diabetes rates, reinforcing the need for proactive health monitoring to identify at-risk populations and implement targeted preventive strategies.

Furthermore, the strong correlation between **high blood pressure and elevated cholesterol levels** with diabetes underscores the importance of **lifestyle modifications and medical interventions** to minimize diabetes-related complications. While **Income and Education** were found to have a negative correlation with diabetes, they were excluded from this analysis to maintain a **focused evaluation of direct health-related factors**.

By prioritizing the most influential risk factors, this dashboard offers **practical insights** for healthcare professionals and policymakers, aiding in the development of **effective diabetes prevention and management strategies**.