University at Buffalo

Diabetes Prediction Analysis

Group Name: Viz wizards

Group Number: 1

Professor Keaton

University at Buffalo

March 3rd, 2025

Group Project

By:

Sadia Alam–50398627

Nimisha Nooti- 50602756

Saumya Kumari- 50603672

Vaishnavi Varanasi – 50590668

Project Report

Team Introduction:

Our group's name is Viz wizards and we are a 4 member group. All of our members are Business Analytics students.

- **Sadia** is interested in data analytics and is also interested in making dashboards for a career. She wants to become a Clinical Business Analyst, or a Healthcare Data Analyst, or even a Marketing Analyst.
- **Saumya** is interested in Data Analytics and aims to apply analytics to healthcare using data to improve decisions and processes.
- **Vaishnavi** is a Statistics enthusiast who finds joy in applying statistical concepts to real-world scenarios making data-driven insights both practical and exciting.
- **Nimisha** is interested in Data Engineering and wants to specialize in designing scalable data pipelines, data process optimization, and managing cloud technologies.

Dataset Description:

Patients' demographic and medical information, as well as whether they have diabetes or not, are included in the diabetes_prediction_dataset.csv file. The dataset is called the diabetes prediction dataset. It includes a number of characteristics, including age, gender, blood glucose level, HbA1c level, smoking history, heart disease, hypertension, and body mass index (BMI). The dataset was found in kaggle and has a size of 9 columns with 100,001 rows. This dataset has a lot of valuable and insightful data. Using the dataset, machine learning models can be built to predict a patient's risk of developing diabetes based on their demographic information and medical history.
Here is a link to the dataset:
https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Visualization objectives:

From the data we hope to understand what factors could have an effect on diabetes. Our target audience for this dataset are healthcare providers and insurance

companies. The healthcare professionals can look at the data and use it to assess which patients would be more at risk and more prone to diabetes. They could also use the data to provide preventative and individualized treatment programs. Insurance companies could use this data to determine how pricing should be set for higher risk individuals. They could offer people who are more prone to diabetes some additional checkups.

The questions we hope to answer are:

- What factors may make individuals prone to diabetes?
- Does age and gender have any effect on diabetes? If so, how?
- Can the BMI of an individual affect whether or not they are at risk of diabetes? If so, how?
- What is the effect of smoking on diabetes?
- Does having a history of other conditions such as hypertension and heart disease make someone prone to diabetes?
- What factors do not have any effect on diabetes?
- Does the risk of diabetes increase with age? Are there specific age groups that are at higher risk?
- At what age does the risk of diabetes start increasing significantly?
- Is there a particular age group where diabetes is common?
- What proportion of people with heart disease also have diabetes?
- Are people with hypertension more likely to develop diabetes?
- How often do people with normal BMI still develop diabetes?
- Is there a "danger zone" for blood glucose levels where diabetes risk sharply increases?

There is a potential to uncover different patterns by visualising this data. Some of them are:

1. **Identify high risk groups:** Understand which demographic(like Age & Gender) and lifestyle(like smoking) factors contribute most to diabetes risk.

2. We can examine how conditions like **hypertension** and **heart disease** link to diabetes.
3. We can determine thresholds in **BMI, HbA1c** and **blood glucose** levels that are strongly associated with diabetes.
4. We can detect a few **unexpected cases** as well(anomaly/ outliers detection).

Planned Visualization - Outcomes:

1. To identify potential outliers, skewness, or normality in the data, We can use **box plots or histograms** to represent the distribution of age, BMI, HbA1c, and blood glucose levels.
2.To compare Prevalence of Diabetes Among Different Groups - We can compare the incidence of diabetes by sex, smoking status, and history of hypertension or heart disease with **bar charts or grouped box plots.**
3. In order to identify Correlations Between Features - We can use **heat maps or scatter plots** to explore correlations between HbA1c, blood glucose, and BMI levels. We can also identify if certain features (e.g., hypertension, heart disease) are strongly correlated with diabetes.
4. To describe Trends Across Age Groups we can use **line plots or stacked bars** to illustrate the rise in risk for diabetes and the role played by hypertension and BMI.
5. In order to compare Blood Glucose and HbA1c Levels we can plot **scatter plots or density plots** to see the correlation between HbA1c levels and blood glucose levels.
6. We can recognize High-Risk Patient Groups by using **heat maps or clustered bar charts** to display high-risk groups by combining a number of factors (e.g., high BMI and advanced age, smoking status, and hypertension).

Plan: Meeting Platform- Google Meet

Week & Dates	Key Tasks	Meeting Schedule

Week 1: Feb 17 – Feb 23	**Data Preprocessing:** Clean the dataset, handle missing values, and check distributions.	Friday- Feb 21st
	Create Initial Visualizations: Develop first drafts of key visualizations based on dataset attributes.	
Week 2: Feb 24 – Mar 1	**Improve & Expand Visualizations:** Enhance clarity, refine aesthetics and optimize insights.	Wednesday- Feb 26th
	Write Narrative Explanations: Describe how each visualization serves the data story.	
Week 3: Mar 2 – Mar 9	**Finalize Visualizations & Insights:** Polish visualizations for presentation quality.	Friday- March 7th
	Complete Final Presentation Deck: Integrate visuals, insights and key takeaways.	

Dataset- Diabetes dataset

Exploratory Data Analysis (EDA):

1. **Initial exploration:**
    - Dataset contains 100,000 rows and 9 columns.
    - There are no missing values in in the dataset.

    The different Column Types are:

    - Categorical Variables: gender, smoking history
    - Numerical Variables: age, BMI, HbA1c_level, blood_glucose_level
    - Binary Variables (0/1): hypertension, heart_disease, diabetes
    - Target Variable: diabetes

2. **Further Analysis:**

    Duplicate Rows: 3,854 duplicates found.

    Numerical Data Insights:

    - **Age:** Ranges from 0.08 to 80 years.

- **BMI:** Goes up to **95.69**, which is unusually high.
- **Blood Glucose Levels:** Maximum is **300**, possibly indicating some outliers.
- **HbA1c Level:** Ranges from **3.5 to 9.0**.

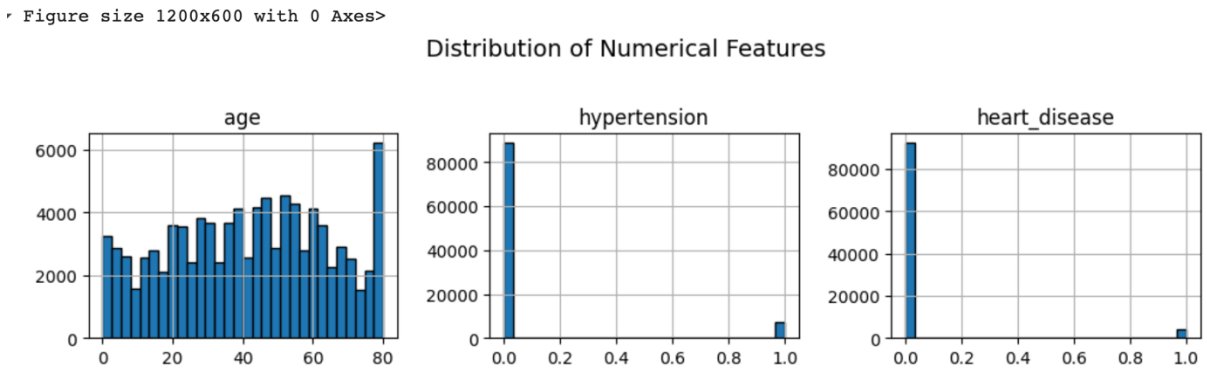
Categorical Data Insights:

- **Gender:** Includes Female, Male, and Other.
- **Smoking History:** Contains never, No Info, current, former, ever, not current.
- **Diabetes Cases:** 8.5% of the dataset has diabetes (1), while 91.5% does not (0).

3. **<u>Univariate Analysis Findings:</u>**

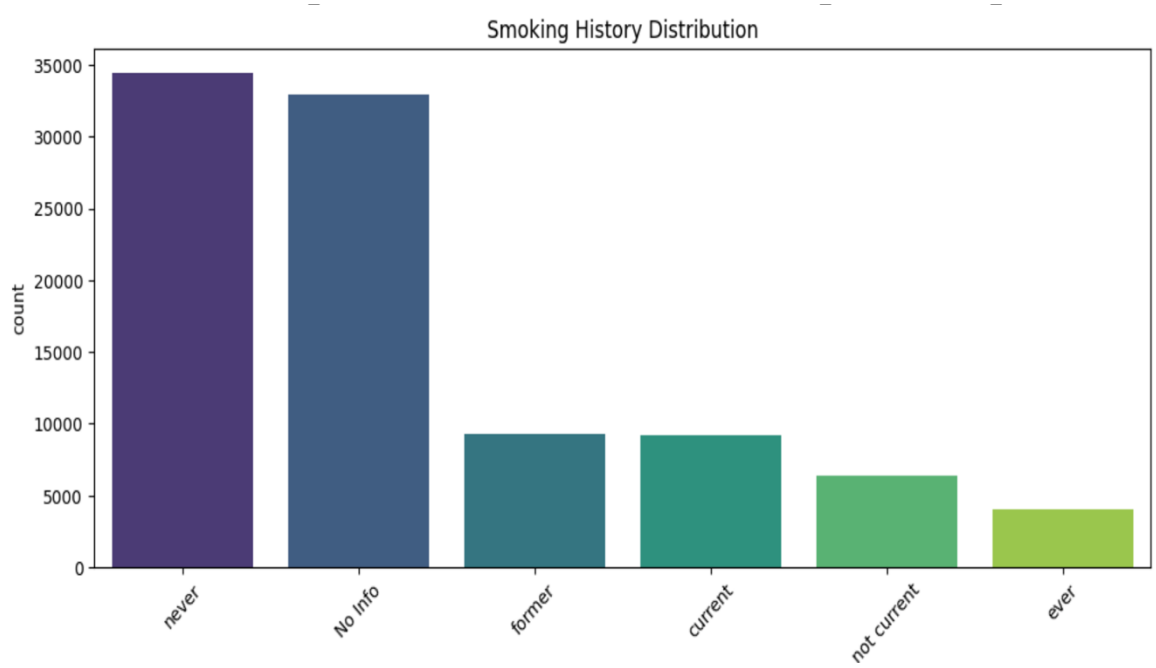
Numerical Distributions:

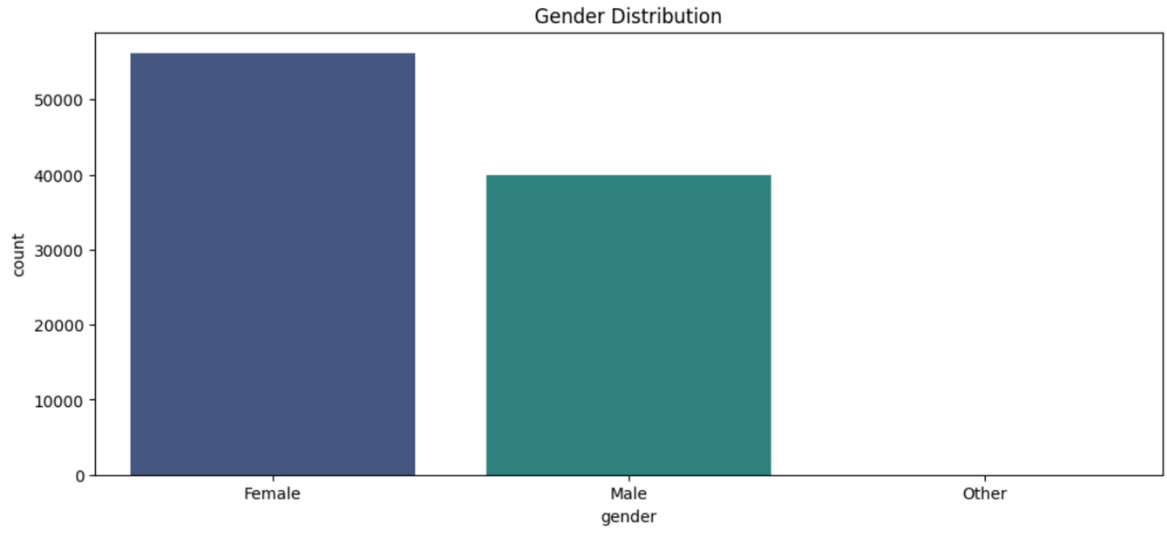
- **Age:** Bimodal distribution with peaks around young and middle-aged groups.
- **BMI:** Most values are between 20-30, but some extreme outliers above 90.
- **Blood Glucose & HbA1c Levels:** Right-skewed, indicating some high values.



Categorical Distributions:

- **Gender:** Majority are Male and Female, with a small fraction labelled Other.
- **Smoking History:** No Info and never are common, followed by former and current.





4. **<u>Outlier and Bivariate Analysis Findings:</u>**

Outliers:

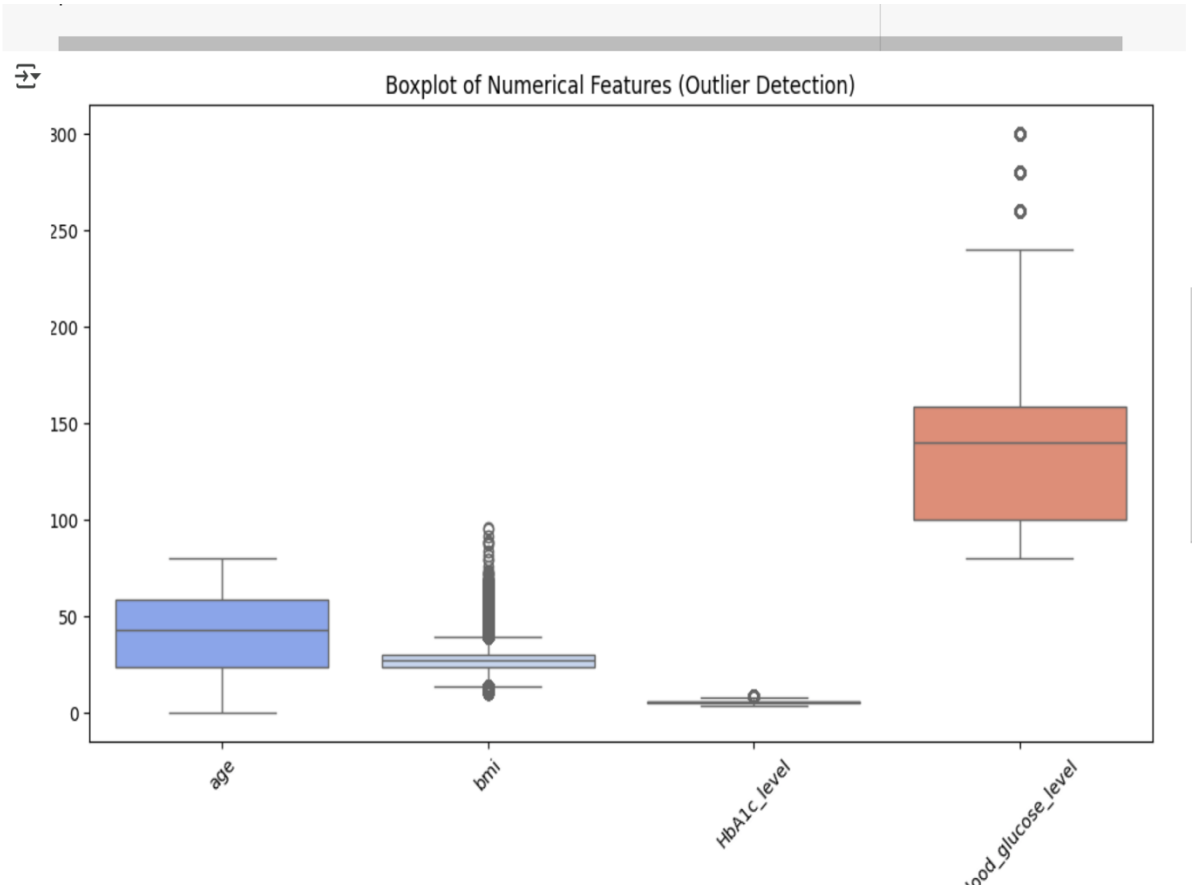
- BMI has extreme outliers (values above 50).
- Blood Glucose Levels have some very high values (above 250).
- Age has very low values (below 1 year), which might be data entry errors.

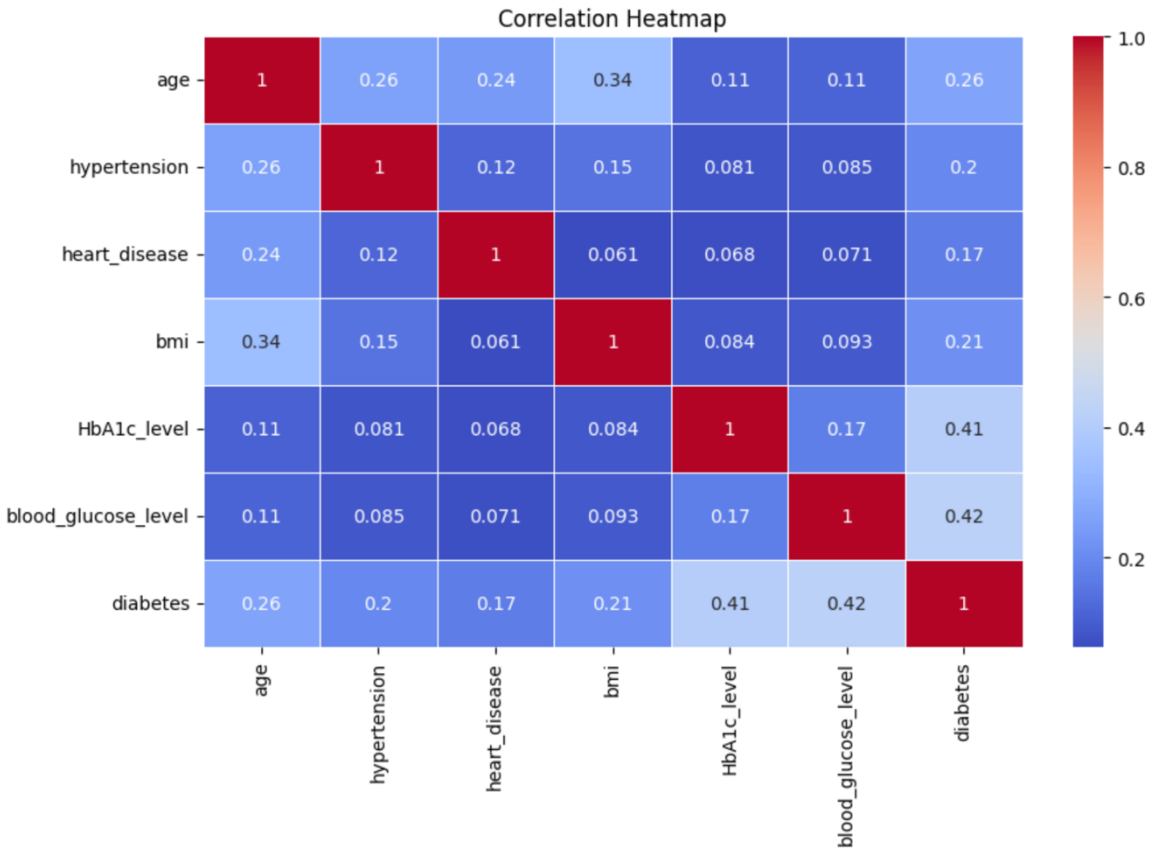
Correlation Insights:

- Blood Glucose Level and Diabetes have a strong positive correlation.
- HbA1c Level also shows a moderate correlation with Diabetes.
- BMI has a weaker relationship with Diabetes.

Diabetes Distribution:

- More males than females have diabetes.
- Higher BMI and Blood Glucose Levels are linked to diabetes.





Summary of the dataset:

This dataset contains **96,146 records** and **9 columns**, with no missing values. It primarily focuses on **diabetes prediction**, using both numerical and categorical features.

1. **Duplicate Records:** Initially, the dataset had 100,000 rows, but 3,854 duplicates were removed.
2. **Categorical Features:** The dataset includes **gender** and **smoking history** as categorical variables.
3. **Numerical Features:** Key numerical attributes are **age, BMI, blood glucose level, HbA1c level, hypertension, and heart disease**.
4. **Target Variable:** The dataset aims to predict **diabetes**, which is represented as a binary variable (0 for non-diabetic, 1 for diabetic).
5. **Diabetes Distribution:**
    - **91.18%** of individuals do not have diabetes.
    - **8.82%** of individuals are diabetic.

Visualisations Plan & Insights from the Visuals:

Story we wish to tell: **Understanding Diabetes Risk Factors and Trends**

We want to explore **what factors contribute to diabetes, how they interact, and what patterns emerge in the dataset**.

Chart1: Age and Diabetes Occurrence (Bar Chart):

1. **Purpose:** Show how **the prevalence of diabetes changes with age groups**.
2. **Chart Type: Bar Chart** (Best for comparison across groups)
3. **Design Considerations:**
    a. Labelling age groups clearly (e.g., **0-10, 10-19, 20-29 +**) and labelling **percent of people with diabetes**.
    b. Highlighting the age group **most affected by diabetes** using an emphasis on color.
4. **Expected Insight:** Does diabetes risk **increase with age**? If yes, at what stage does it spike?

Insights from the graph:

Age and Diabetes Prevalence – Bar Chart

Diabetes Prevalence is Highest in the 40-70 Age Group

- The dark colors highlight that **these age group have the largest number of diabetic individuals**.

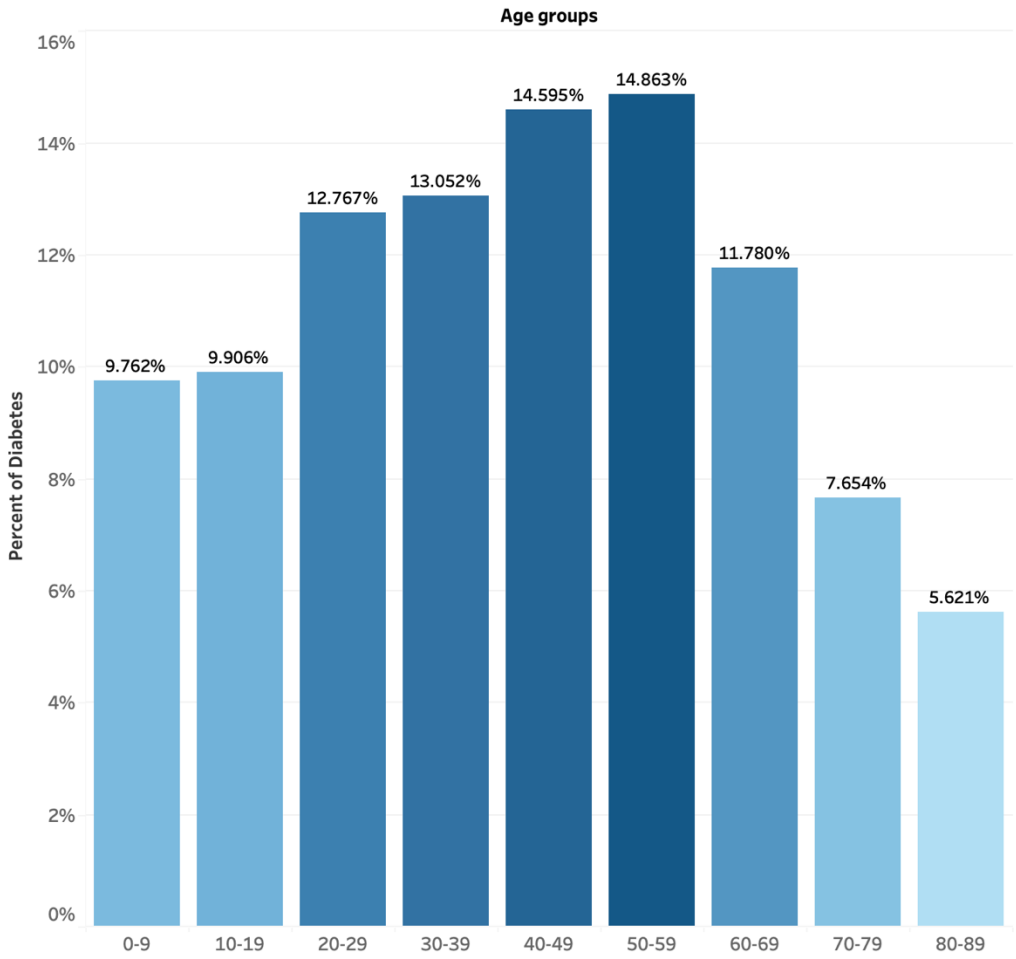
Younger Age Groups (20-40) Also Show a High Number of Diabetic Cases

- In younger age groups, people in their **20s and 30's** also have diabetes.

Lowest Diabetes occurrence in the 0-19 and 70-89 Groups

- People who are **younger like children and teens** are less diabetic.
- People who are **older in their 70s and 80s** also have a low count of diabetes.

The Prevalence of diabetes amongst different age groups



Why I chose this visual?

- **Bar charts** are a great way to show how data is distributed through a certain factor. In this case that would be how diabetes is shown through different age groups. It also aligns with What I want to emphasize? It tells a clear story i.e., my data answers a question.
- **The main insight is evident:** The prevalence of diabetes increases with age. However, surprisingly elderly individuals, people past their 70's, tend to have less occurrence of diabetes.
- **Annotation is used effectively** to highlight the exact percentage of people who have diabetes in each age group. Each age group is displayed in **bins,** this makes it so that division between different age groups is easy to understand
- **The title and axis labels are clear**, making interpretation easy.

- **Color is used strategically**, darker colors indicate a higher prevalence of age groups with diabetes, while lighter colors indicate lower prevalence of diabetes.

How would this chart be used:

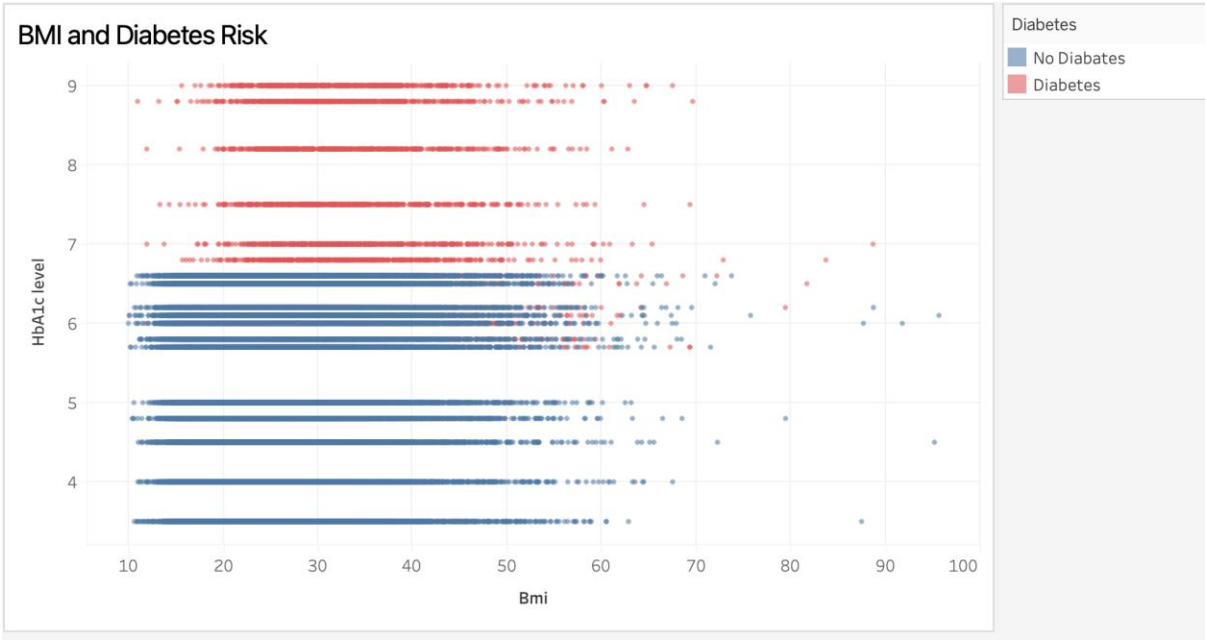
- **Healthcare workers,** could use this data to prevent the occurrence of diabetes in certain age group. Preventative care could be utilized to keep people from getting diabetes in the first place. This chart could prevent risk.

Chart2: BMI and Diabetes Risk (Scatter Plot)

1. **Purpose:** Show the relationship between BMI and diabetes
2. **Chart Type:** Scatter Plot (Best for correlation analysis).
3. **Design Considerations**:
    a. Plotting BMI on the x-axis and HbA1c levels on the y-axis.
    b. Using color to separate diabetic vs. non-diabetic individuals.
4. **Expected Insight**: Does a **higher BMI increase diabetes risk**? Is there a **BMI threshold where diabetes becomes significantly more common**?

Insights from the graph:

BMI and Diabetes Risk (Scatter Plot)



HbA1c Levels are Higher in Diabetic Individuals

- Diabetic individuals (red dots) mostly have HbA1c levels above 6.5, confirming that higher HbA1c levels strongly correlate with diabetes.
- Non-diabetics (blue dots) cluster around 4-6 HbA1c levels, which aligns with normal blood sugar ranges.

Higher BMI is Linked to Increased Diabetes Risk

- The majority of diabetic individuals (red) have BMI values above 25, reinforcing the well-known link between obesity and diabetes risk.
- While not all individuals with high BMI have diabetes, the proportion of diabetics increases as BMI rises.

Distinct Clustering and Outliers

- Most non-diabetic individuals have BMI between 18-30 and HbA1c below 6.5.
- Diabetics show a more spread-out distribution, particularly for higher BMI values (above 30).
- Some outliers exist at extremely high BMI values (above 70-90), which could be due to measurement errors or rare cases.

Why I Chose This Visual?

- This is a **relationship visualization**, making **scatter plots the best choice**. It effectively **shows trends between BMI and diabetes prevalence**.
- **Position**: The scatter plot accurately places individuals along **BMI (X-axis) and HbA1c levels (Y-axis)**.
- **Color Differentiation: Red** for diabetics, ensuring immediate visibility of affected individuals. **Gray** for non-diabetics, providing contrast without overwhelming the graph. **Opacity Reduces Overlapping**: Helps prevent clusters from looking too dense.
- **Follows Readability and Clarity Principles**
    - **No clutter**, as only relevant data points are displayed.
    - **Gridlines are minimal**, keeping the focus on data points.
    - **Direct labelling with axis titles and a clear heading** ensures **ease of interpretation**.
- **Provides a clear answer to the question:** Does a higher BMI increase diabetes risk? Is there a BMI threshold where diabetes becomes significantly more common?- Higher BMI is linked to increased diabetes risk, with diabetes cases rising significantly beyond BMI 25. The strongest increase in diabetes prevalence occurs at BMI 30 and above (Obesity range). While BMI alone does not cause diabetes, it is a major contributing factor in metabolic health. BMI threshold where diabetes prevalence increases significantly appears to be around 25-30.

Chart 3: Diabetes Prevalence Among Individuals with Heart Disease & Hypertension (Bubble Chart)

1. **Purpose:** Show the relationship between heart disease, hypertension, and diabetes prevalence.
2. **Chart Type:** Bubble (Best for compositional analysis).
3. **Design Considerations**:

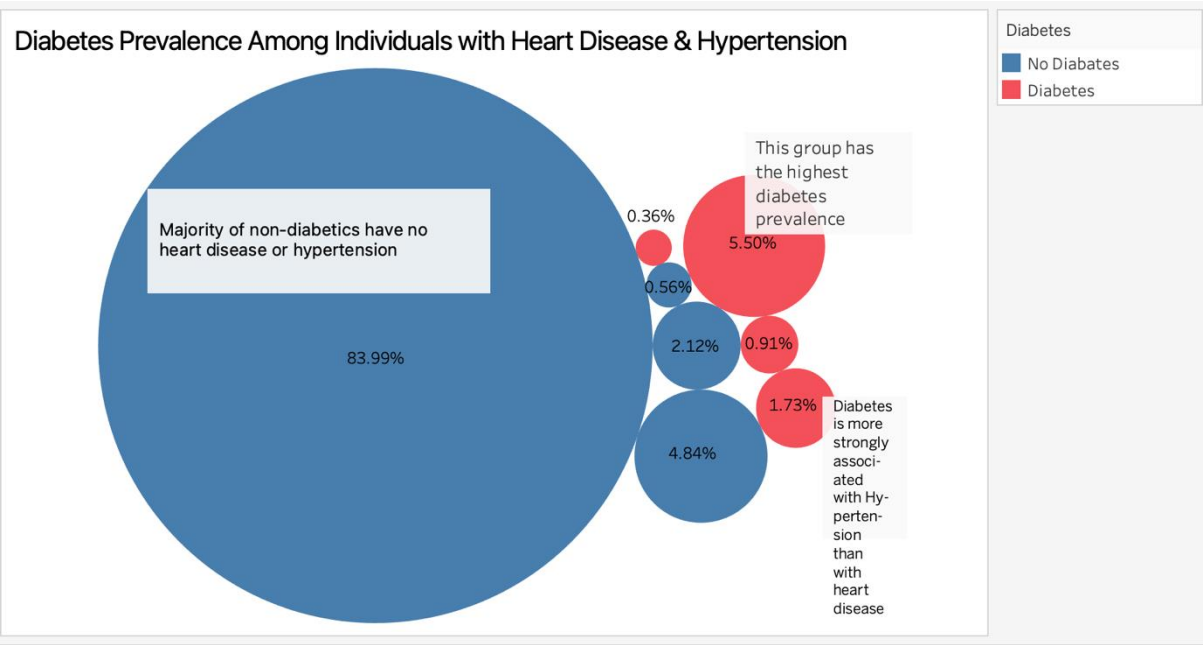
a. Positioning the largest bubble (83.99%) centrally ensures that it is the first thing the viewer notices.
b. Smaller bubbles are placed around it, allowing for an easy comparison of diabetes prevalence among different risk groups
c. Diabetics → Red (Higher Risk), Non-Diabetics → Blue (Lower Risk), This contrast immediately draws attention to the high-risk diabetic groups.
d. Following the **best practice of direct labelling** instead of using a legend, making the chart easier to interpret.
e. The **Z-pattern layout** guides the eye from the **largest bubble → high-risk groups → annotations.**

4. **Expected Insight**: How do hypertension and heart disease impact diabetes prevalence?

Insights from the graph:

Diabetes Prevalence Among Individuals with Heart Disease & Hypertension (Bubble Chart)

- **Most non-diabetics (83.99%) have neither hypertension nor heart disease.** This confirms that **the majority of the population does not have these conditions or diabetes**.



- **Diabetes is still prevalent even in individuals with no hypertension or heart disease (5.50%).** This highlights that **diabetes can develop independently of these conditions**, but the risk is lower.
- **Hypertension is more strongly linked to diabetes than heart disease alone.** The red bubble(1.73%) clarifies that **diabetes cases are more concentrated in individuals with hypertension** rather than heart disease.

Why I Chose This Visual?

Bubble Chart was chosen because:

- It is best for Composition Analysis
- Bubble charts are ideal for showing proportions and making comparisons between categories easy.
- The bubble sizes accurately reflect the distribution of diabetics and non-diabetics across different health conditions.
- It shows clear Visual Hierarchy
- Largest bubble = Most non-diabetics, instantly emphasizing the baseline population.

- Smaller bubbles highlight how diabetes prevalence changes with additional conditions.

o Effective Use of Color & Labels

- Red for diabetics, blue for non-diabetics, ensuring instant differentiation.
- Direct percentage labelling improved readability without requiring extra legends.

o Annotations Enhance Storytelling

- Key insights are highlighted in text boxes, reducing the need for additional interpretation.
- The viewer immediately understands where diabetes prevalence is highest.

Chart4: Diabetes Prevalence Across Smoking History Categories (Square Tree Map):

1. **Purpose:** This visualization highlights the relationship between **smoking history** and **diabetes prevalence** by comparing individuals across different smoking categories.

2. **Chart Type:** Square Tree map (Best for Proportional Comparison). Square tree map charts are effective for comparing proportions within categories. The size of each square represents the number of individuals within that smoking category. The color contrast between diabetics (red) and non-diabetics (blue) makes it easy to distinguish between the two groups.

3. **Design Considerations:**
    a. **Positioning for Focus**
        i. Larger squares (e.g., **Never Smoked & No Info groups**) are placed **centrally** to ensure they **catch the viewer's attention first**.
        ii. Smaller squares are arranged around them to provide context.
    b. **Effective Use of Color**
        i. **Diabetics → Red (Higher Risk)**
        ii. **Non-Diabetics → Blue (Lower Risk)**

iii. This contrast helps viewers quickly understand how smoking history correlates with diabetes prevalence.

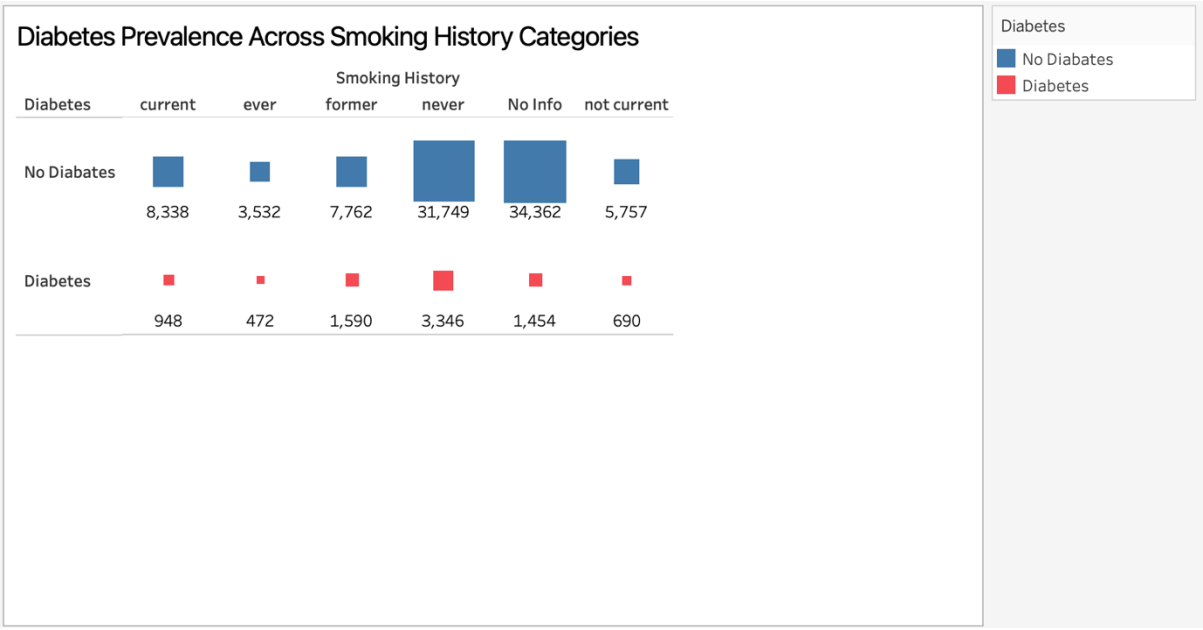
c. **Labelling for Clarity**

i. **Direct labels** are placed inside each square rather than relying on a legend, improving readability.

ii. Each label shows the **exact number of individuals** in each category.

d. **Guided Flow with Z-Pattern Layout**

e. The eye naturally follows a **Z-pattern**, starting from the largest blue square (No Info category) and moving through the **Never Smoked** category to other smoking history groups.



Insights from graph:

- **Never Smoked Group is Surprisingly Large.** A **large portion of individuals who never smoked are non-diabetic** (31,749), indicating that smoking is not necessarily a major predictor of diabetes.
- However, **diabetes is still present (3,346 cases)** in this group, showing that **other factors contribute to diabetes risk.**
- **No Info Category is Significant**

- Many individuals have **no smoking history recorded**, and their diabetes rate is **relatively low** (1,454 cases).
- This suggests that missing data or other lifestyle factors might play a role in the diabetes risk assessment.
- **Former Smokers Have a Considerable Diabetes Presence**
- While **7,762 former smokers** are non-diabetic, **1,590 are diabetic**.
- This suggests that **smoking history may contribute to long-term diabetes risk, even after quitting.**
- **Current Smokers Have a Lower Reported Diabetes Rate**
- Only **948 current smokers** have diabetes compared to **8,338 non-diabetics**, suggesting either:
    - A **selection bias (healthy smoker effect)** where individuals who smoke but develop diabetes may quit.
    - Other lifestyle or genetic factors may be more relevant to diabetes risk.

Why I Chose This Visual?

1. **Best for Compositional Analysis**
    - A tree map is ideal for understanding the **proportions of diabetics and non-diabetics** in each smoking category.
    - The square sizes accurately reflect the **distribution** of individuals within each group.

2. **Clear Visual Hierarchy**
    - The **largest category (Never Smoked, No Info)** is **centrally placed**, ensuring that the most important data is seen first.
    - Smaller groups are arranged around it to maintain a **logical structure**.3

3. **Effective Use of Color & Labels**
    - **Red (Diabetics) vs. Blue (Non-Diabetics)** ensures instant recognition.

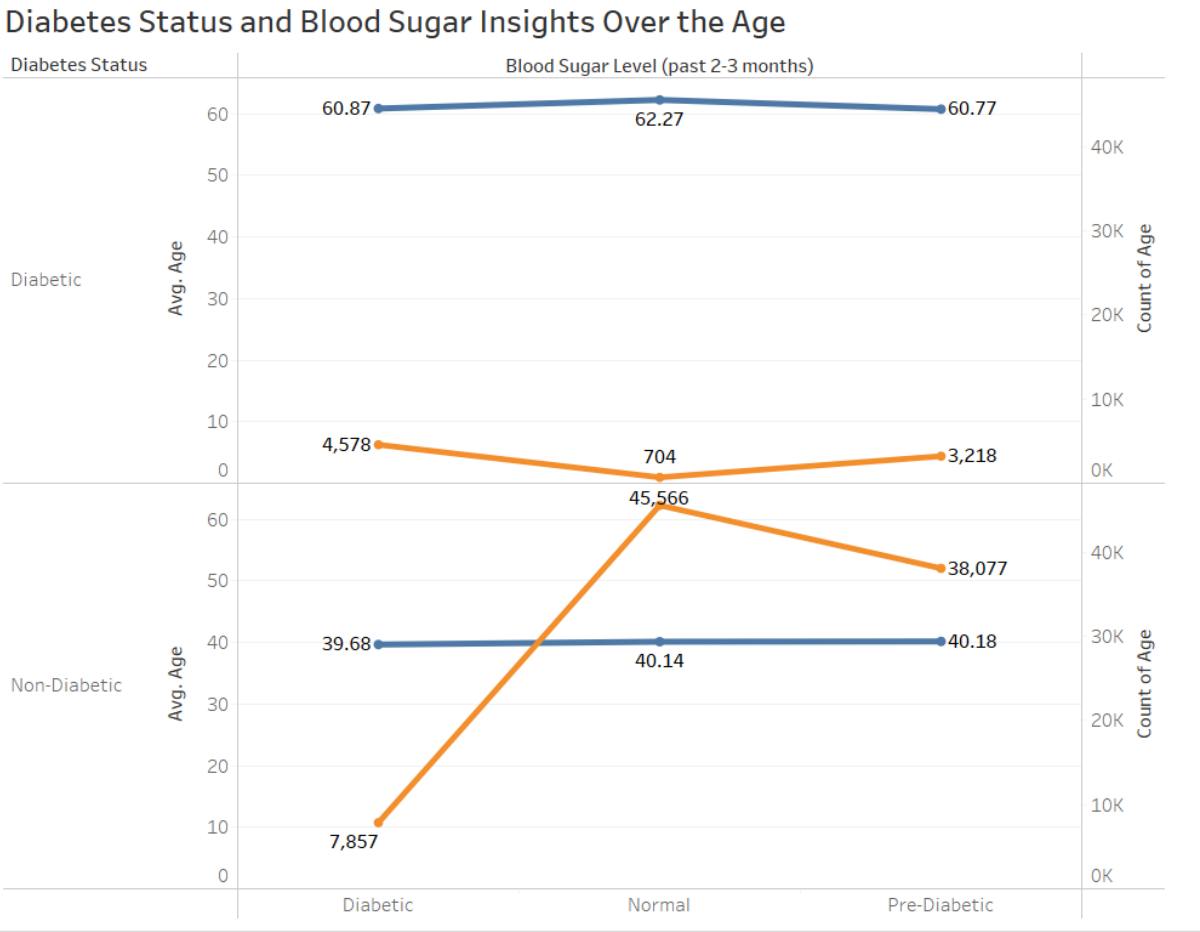
- **Direct number labels** make interpretation quick and easy.

4. **Storytelling Through Annotations**
   - Highlights **unexpected insights**, like the fact that **diabetes exists even in non-smokers**.
   - Guides viewers to **question assumptions** about smoking and diabetes risk.

Chart- 5: Diabetes Status and Blood Sugar Insights Over the Age –Measuring blood sugar level using HbA1c level.

1. **Purpose**: The purpose of the visualization is to contrast diabetic and non-diabetic patients' blood glucose levels and avg age in the past 2-3 months. It also wishes to divide individuals into normal, pre-diabetic, and diabetic categories based on their blood glucose levels.

2. **Chart Type**: The chart type is likely a 'Line Chart', which would easily allow comparison between blood sugar and average age in diabetic and non-diabetic groups.



3. **Design Considerations**: The use of varying colours to make it more discernible between the diabetic and the non-diabetic groups. Careful labelling of the X-axis (age, blood sugar values) and the Y-axis (frequency or numbers). Having a legend to separate diabetic, pre-diabetic, and normal categories. Ensuring the data range (age and blood sugar levels between 0-70) is properly labelled.

4. **Expected Insight**: This visualization shows data pertaining to diabetes status, with a focus on blood sugar levels for the last 2-3 months and mean age. Some of the following insights can be inferred:
    - **Blood Sugar Levels:** The blood sugar levels are measured using HbA1c level (<5.6: normal, 5.7-6.4: pre diabetic and >6.4-diabetic) for diabetic and non-diabetic patients ranging from 0 to 70 age. This indicates a comparison of blood sugar control in these two groups.

- **Average Age:** The average age of the individuals in the dataset ranges between 0 to 70 years. This explains the False Positive and False Negative of the HbA1c test by comparing it to the actual diabetes status.
- **Comparative Analysis:** This viz compares blood sugar and age in diabetic versus non-diabetic individuals, showing trends or differences between these metrics.
- **Potential Trends:** There are trends showing how blood sugar changes with age or how diabetes prevalence changes by age groups. At the age of 39-41 patients are non-diabetic but still there HbA1c levels are varying. We can notice a greater number of patients with normal HbA1c level are stated as non- diabetic. At the age of 60- 63 patients are mostly diabetic and there HbA1c levels are

5. **Why I Chose This Visual?**

   This visual (Line Chart) was chosen because it effectively compares key metrics (blood sugar levels and age) between diabetic and non-diabetic individuals. It also provides a clear classification of diabetes status - false positive and False negative values in HbA1c test level, which is crucial for understanding the data distribution and identifying potential health trends. The comparative nature of the chart makes it easy to spot differences and patterns at a glance.

Targeted Audiences:

- Physicians, nurses, and endocrinologists who should have information about blood sugar level and diabetes status at different age groups to provide quality patient care.
- Public health planners and policymakers who require information on diabetes and pre-diabetes incidence to plan and implement efficient health interventions and programs.
- Clinicians who research diabetes trends, risk factors, and the effect of age on diabetes prevalence.

- People who have diabetes or pre-diabetes and their care providers who could benefit from being aware of how blood sugar changes with age and diabetes status.
- Experts who educate the public about diabetes prevention and control, highlighting through statistics the importance of blood glucose testing and identifying risk factors.
- Insurers who need data on diabetes prevalence and risk factors to develop health insurance products and ascertain risk.
- Companies that produce drugs and medications for diabetes, which need data on the population stricken to guide research and marketing.

Reflection:

From this assignment, we learned that certain factors like age, smoking status, blood sugar levels, heart disease, and hypertension may or may not influence diabetes. Initially, we assumed that all these factors would make diabetes inevitable. However, after analysing the data and creating visualizations, we discovered some unexpected insights.

For instance, we were confident that age would directly correlate with a higher risk of diabetes. Surprisingly, elderly individuals (70+) had lower diabetes rates than younger individuals. We also assumed that smoking would have a direct impact on diabetes, but the data revealed that many non-smokers had a history of smoking.

Some of our initial hypotheses were confirmed. High BMI and elevated **HbA1c** levels were strongly linked to increased diabetes risk, which aligned with our expectations. Additionally, we found that diabetic patients generally had higher blood sugar levels, further supporting our assumptions.

Creating the visualizations required multiple iterations and revisions. Choosing the right charts and graphs was challenging and involved a lot of trial and error. Each team member ensured that we used diverse visualization types and explored different factors

to avoid redundancy. We also carefully selected a **theme and color scheme** that represented health and wellness.

Initially, our charts were not as effective as we had hoped. However, through peer feedback and collaboration, we refined our visualizations to better convey our findings. It was important for us to tell a compelling story with our data, ensuring that our visualizations were **informative and meaningful** for the audience.

This assignment provided **hands-on experience** in creating a visualization dashboard. It also gave us **real-world exposure** to team collaboration and designing effective visualizations. Through this process, we gained valuable insights into how data can be transformed into a **compelling and impactful narrative**.