Group 1

Ayush Patel – 202318036
Mitul Dudhat - 202318024
Siddharth Kadam – 202318015
Sreejesh S. Nair – 202318001

Diabetes Risk Analysis Dashboard Report

I. **Project Goal**:

The goal of this project is to analyse health and demographic data to identify key risk factors associated with diabetes and visualize their impact across different populations. By examining attributes such as age, BMI, gender, smoking history, hypertension, and heart disease, we aim to uncover patterns and correlations that highlight high-risk groups, ultimately supporting targeted diabetes prevention and intervention strategies.

II. Objectives and Narrative

- **Objective**:

1. Identify and visualize critical factors that contribute to diabetes risk using statistical data and visualization techniques.
2. Highlight the influence of individual factors (e.g., age, BMI, gender, chronic health conditions) on diabetes prevalence.
3. Examine combined effects of multiple risk factors to understand their cumulative impact on diabetes likelihood.
4. Convey a narrative that emphasizes significant predictors of diabetes across demographic groups.
5. Illustrate how different risk factors, individually or collectively, affect diabetes risk for targeted prevention insights.

Story:

The report's narrative centres on understanding diabetes risk through a layered approach. Starting from isolated risk factors (like smoking, hypertension, heart disease) to exploring cumulative risks (chronic risk scores), age-related patterns, BMI distributions, and gender-specific variations, this report provides a comprehensive picture of diabetes predictors. Visualization methods are used to create a coherent story showing how diabetes risk factors escalate and interact across different population subsets.

III. Dataset

Brief Description:

The dataset is structured in a tabular format for diabetes prediction and health analysis. It includes both categorical and numerical attributes, with a total of 100000 records. Each record represents an individual's health metrics, capturing various attributes that influence diabetes risk.

Variables

- **Age**: Represents the age of each individual in the dataset. There are 102 unique age values, ranging from 0 to 80 years.
- **Gender**: A categorical attribute, gender classifies each individual as either "Male" or "Female," with two possible levels.
- **BMI (Body Mass Index)**: BMI is a quantitative attribute showing the body mass index of each individual, with values ranging from 10 to 96. It has a cardinality of 4247.
- **Smoking History**: This categorical attribute indicates whether an individual is a smoker, with six levels: "Current", "Ever", "Never", "Former", "Not Current" and "No Info".
- **Hypertension**: A categorical attribute that denotes whether the individual has high blood pressure. It has two levels: "1" for presence and "0" for absence of hypertension.
- **Heart Disease**: This categorical attribute indicates the presence of heart disease, with two levels: "1" for presence and "0" for absence of heart disease.
- **Blood Glucose Level**: A quantitative attribute representing the fasting blood glucose level for each individual, with values ranging from 80 to 300.
- **HbA1c Level**: This quantitative attribute measures the average blood glucose level over the past three months (measured by HbA1c). The range for HbA1c values is from 3.0 to 10.0.

Derived Variables

- **Chronic Risk Score**: This derived quantitative variable combines heart disease, hypertension, and smoking history into a cumulative risk score, indicating overall diabetes risk. The score is calculated as follows:

- 0.5×Heart Disease+0.3×Hypertension+0.2×Smoking History Score0.5×Heart Disease +0.3×Hypertension+0.2×Smoking History Score.
- The **Smoking History Score** is determined based on smoking history:
  - *Current smoker*: 1
  - *Ever smoked*: 0.75
  - *Former smoker*: 0.5
  - *Not current smoker*: 0.25
  - *Never smoked*: 0

- This formula weights heart disease as the most significant contributor, followed by hypertension and smoking history. The Chronic Risk Score ranges from 0 to 1, depending on the presence and severity of each factor.

- **Smoking Category**: This derived variable categorizes smoking history into two groups based on smoking data availability:

  - *Yes*: If smoking history is anything other than "No Info" or "Never"
  - *No*: If smoking history is "No Info" or "Never" This categorization helps simplify analysis by indicating whether an individual has a known history of smoking.

- **Age Group**: This derived categorical variable segments age into distinct groups:
  - *Infant*: Age < 1
  - *Children*: Age 1–12
  - *Adolescents*: Age 13–17
  - *Adult*: Age 18–65
  - *Old*: Age > 65 This grouping allows for a clearer understanding of diabetes risk across different life stages.
- **BMI Categories**: This derived categorical variable categorizes individuals based on BMI ranges:
  - *Under 25 BMI*: BMI ≤ 25
  - *25-30*: BMI between 25 and 30
  - *Over 30 BMI*: BMI > 30 This categorization helps visualize the risk associated with different BMI levels.
- **Weight Category**: Derived using BMI values, this variable categorizes individuals into different weight classifications:
  - *Underweight*: BMI < 18.5
  - *Normal weight*: 18.5 ≤ BMI < 24.9
  - *Overweight*: 25 ≤ BMI < 29.9
  - *Obese*: 30 ≤ BMI < 34.9
  - *Severely Obese*: BMI ≥ 35 This classification offers insights into how weight categories relate to diabetes risk.

IV. Context and Problems

Context:

Diabetes is a chronic, potentially life-threatening condition that has reached epidemic proportions worldwide, with a range of contributing health conditions and lifestyle choices. Understanding the factors that elevate diabetes risk can significantly enhance preventive care, allowing healthcare providers to identify high-risk individuals and implement tailored interventions. This project aims to leverage data visualization to uncover and validate relationships between several health conditions and diabetes risk factors, with a focus on identifying actionable insights that can help in risk assessment and potentially inform healthcare interventions.

Background:

This analysis utilizes a dataset comprising health records for 100,000 individuals, each containing data on demographics and health-related metrics such as gender, age, hypertension, heart disease, smoking history, BMI, HbA1c levels, and blood glucose levels, along with a diabetes indicator. By examining these variables through data visualization techniques, we aim to reveal insights into the relationships between various health conditions and diabetes, identifying patterns that could support healthcare professionals in making informed decisions. Specifically, we'll investigate how factors like hypertension, heart disease, smoking habits, obesity, age, and gender influence diabetes risk.

Specific Problem:

The major goal of this project is to analyse how individual and combined risk factors, such as hypertension, heart disease, smoking behaviours, BMI, and demographic characteristics like age and gender, influence diabetes risk. Furthermore, the study intends to analyse the relative significance of HbA1c levels vs blood glucose as indicators of diabetes risk.

The analysis will focus on visual comparisons and insights regarding:

Type 1:

Key Questions to Explore:

1. **How do cardiovascular health and lifestyle factors affect diabetes risk?**
   - *Hypothesis (H1):* Individuals with a higher "Chronic Risk Score"—a combination of hypertension, heart disease, and smoking—are more likely to develop diabetes, with heart disease having the strongest individual impact.
2. **Does BMI significantly impact diabetes risk across different demographics?**

- *Hypothesis (H2):* Individuals with a BMI over 30 are at a higher risk of diabetes compared to those with a BMI under 25, regardless of age and gender.

3. **Does gender play a role in influencing diabetes-related biomarkers?**
   - *Hypothesis (H3):* Gender and obesity levels significantly affect blood glucose and HbA1c levels, with males and those classified as severely obese showing higher levels. Furthermore, HbA1c may serve as a more reliable predictor of diabetes risk compared to blood glucose alone.

Type 2:

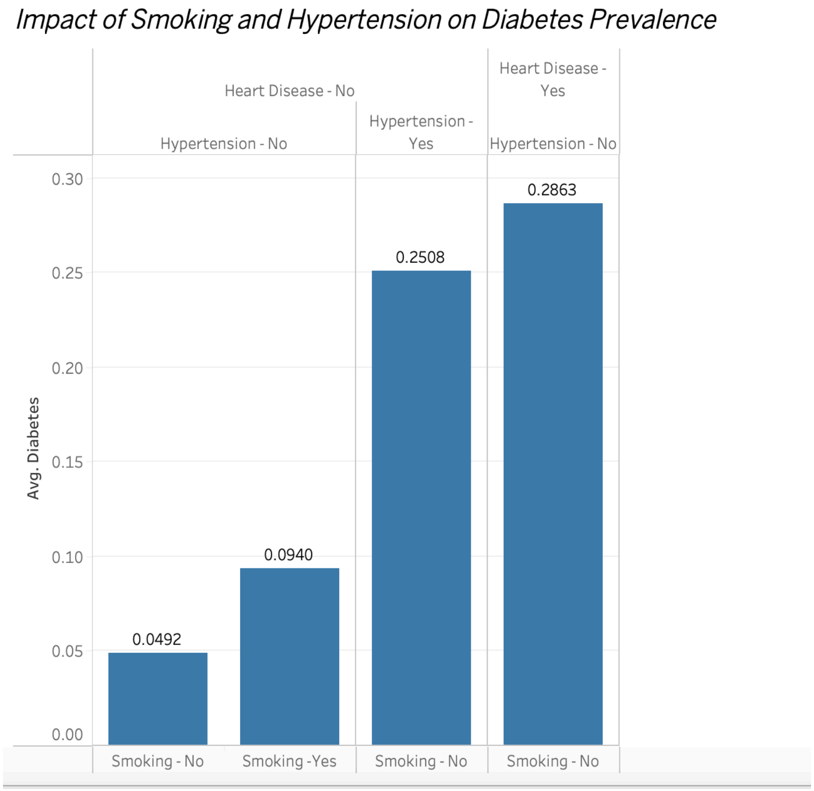
Key Questions for Exploration

1. **Risk Factors and Diabetes Correlation**
   - How do specific health factors—heart disease, hypertension, and smoking history—influence the likelihood of diabetes?
   - Is there a combined effect of these factors, or does one factor, like heart disease, play a more critical role?
2. **BMI as a Predictor of Diabetes**
   - Are individuals with high BMI more likely to develop diabetes than those with lower BMI, regardless of their age or gender?
   - How does obesity in males and females influence blood glucose and HbA1c levels?
3. **Demographic Influences on Blood Glucose and HbA1c Levels**
   - Does gender impact blood glucose and HbA1c levels, and is HbA1c a more reliable indicator of diabetes risk than blood glucose alone?
   - Do males and severely obese individuals show higher levels of blood glucose and HbA1c, as hypothesized?

V. Techniques Used and Tasks Performed for Each Dashboard

DASHBOARD 1: Relationship Between Risk Factors and Diabetes Incidence

1. Individual impact of Smoking, Hypertension and Heart Disease on Diabetes Prevalence



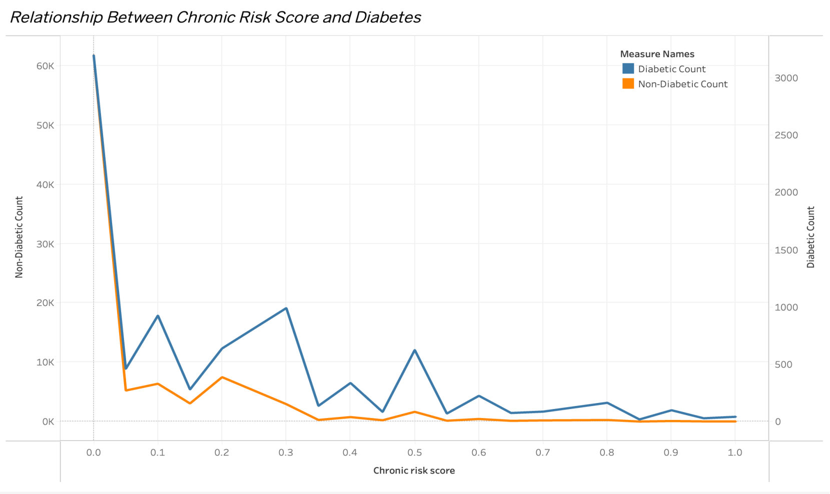
◊ **Idiom**: **Bar chart**

- **Three key, One Value**
    - **Categorical attribute**: **Smoking status combined with Heart Disease and Hypertension** (e.g., "Smoking - No" with "Hypertension - No" and "Heart Disease - Yes").
    - **One quantitative attribute**: **Average Diabetes Prevalence** (Avg. Diabetes) shown along the y-axis.

- **Mark(1D)**:
    - **Lines** representing each category combination's average diabetes prevalence.

- **Channels**:
  - **Vertical position(Length)**: Used to express the quantitative value of Avg. Diabetes for each category combination.
  - **Spatial regions**: One per mark, where each bar is horizontally separated and vertically aligned, with clear distinction between category groups.
    - **Ordering by Avg. Diabetes**: Order each bar in ascending order based on Average Diabetes Percentage.
- **Task**:
  - **Comparison and value lookup** to understand the impact of smoking, heart disease, and hypertension on diabetes prevalence across different subgroups.
- **Scalability**:
  - Suitable for **dozens to hundreds of levels** for the categorical attribute (key) and **hundreds for quantitative values** (bars). This chart currently has a small number of categories, making it easy to interpret and compare values.

2. Relationship Between Chronic Risk Score and Diabetes



◊ **Idiom**: Dual-axis line chart

- **Two key, One value**
  - **Categorical attribute**:
    - **Diabetes Status** (Line Categories): Two categories represented by the lines, "Diabetic Count" and "Non-Diabetic Count".
    - **Count** (Y-axis for both measures): The primary measure displayed on both Y-axes is the count of Diabetic and Non-Diabetic instances

- **Quantitative attribute**:
    - **Chronic Risk Score** (X-axis): This is a continuous variable showing risk scores ranging from 0 to 1.

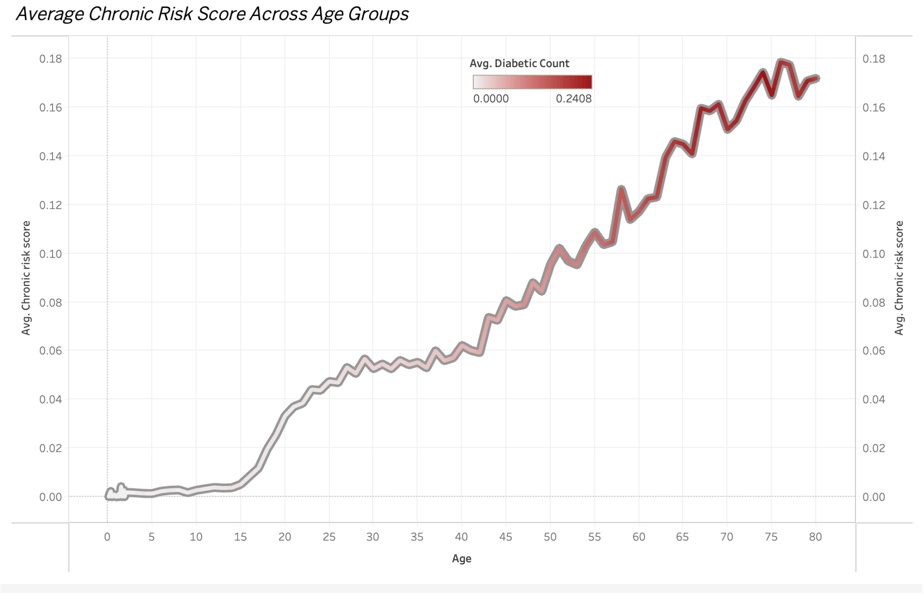
- **Marks (1D)**:
    - Lines for each category, representing the trend of counts over Chronic Risk Score.

- **Channels**:
    - **Color**: Differentiates between "Diabetic Count" (blue) and "Non-Diabetic Count" (orange).
    - **Position**: The height of each line represents the count for each Chronic Risk Score category.

- **Task**
    - The primary task here is **comparison** of Diabetic and Non-Diabetic Counts across different Chronic Risk Scores, allowing for insight into how risk levels relate to diabetes status.

- **Scalability**
    - This chart design scales well for datasets with a moderate range of Chronic Risk Scores and Diabetes statuses. However, for large more nuanced categories or risk intervals.

3. Average Chronic Risk Score Across Age Groups



Idiom: Single-axis line chart

- Zero key, Three Value

- **categorical attribute**: None – The chart does not use categorical attributes or grouped data.
- **quantitative attribute**:
  - **Age (X-axis):** This is a continuous variable showing age groups, ranging from 0 to around 80 years.
  - **Average Chronic Risk Score (Y-axis):** A continuous measure representing the average chronic risk score for each age group,
  - **Average Diabetic Count (Color Gradient):** Represented by the color intensity along the line, where a darker color indicates a higher average diabetic count for each age group.

- **Mark(1D)**:
  - **Lines:** A single line connects data points across the age range, representing the trend in average chronic risk score as age increases.
- **Channels**:

  Position:
  - **X-axis** – Represents **Age**, showing the trend of chronic risk across different age groups.
  - **Y-axis** – Represents **Average Chronic Risk Score**, with higher values indicating a higher risk.

  Color (Intensity/Gradient):
  - Represents **Average Diabetic Count**, with color intensity increasing alongside the diabetic count for each age group.

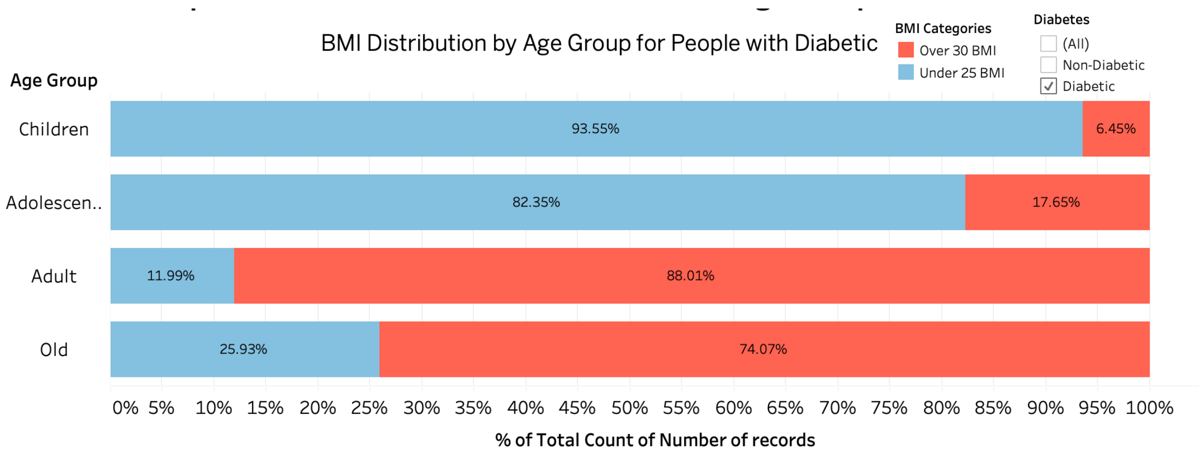
- **Task**:
  - Identify the trend of average chronic risk score across different age groups.
  - **Observe diabetic prevalence** alongside chronic risk, using color intensity to see where diabetic counts are higher.

- **Scalability**:
  - **Data Range:** This chart is suitable for datasets with continuous age and risk values across a reasonable range (e.g., age 0–80, chronic risk score up to 0.2).

Additional Attributes: If necessary, additional health indicators or risk factors could be added as different line markers or color gradients

DASHBOARD 2: Impact of BMI on Diabetes Across Age Groups and Gender

1. BMI Distribution by Age Group for People with Diabetes



Idiom: **Stacked Bar Chart**

- Two key, One Value

- **Categorical attributes**:
  - Age Group (Children, Adolescent, Adult, Old)
  - BMI Categories (Over 30 BMI, Under 25 BMI)
- **Quantitative attribute**: Percentage of total count of records for each BMI category within the age groups.

Marks(1D)

- **Vertical stack of line marks**: Each horizontal bar represents an age group, with the stacked segments representing different BMI categories.

Channels

- **Horizontal position (Length):** Represents the percentage of records for each BMI category.
- **Color (hue)**: Distinguishes the BMI categories (blue for Under 25 BMI and red for Over 30 BMI).
- **Spatial regions**:
  - **Aligned region**: The entire bar is aligned to the percentage of the total count, showing the full scale of BMI distribution.
  - **Unaligned region**: The inner segments (BMI categories) within each bar are unaligned but form part of the whole bar.

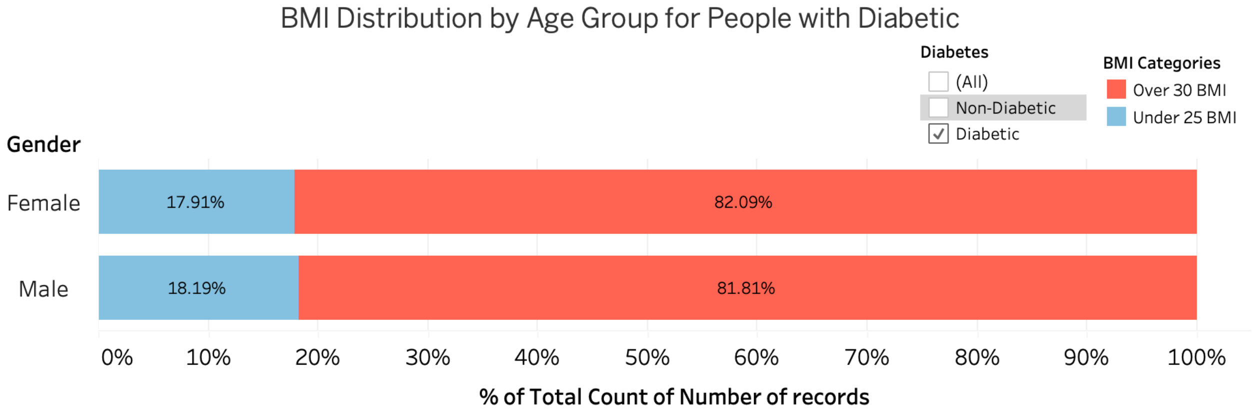
Tasks:

- **Part-to-whole relationship**: The chart shows how much of the total diabetic population in each age group falls into the two BMI categories.

Scalability:

- **For the stacked key attribute (BMI categories)**: The chart can handle 10-12 levels or segments before becoming overcrowded.
- **For the main key attribute (Age group)**: The chart is scalable to dozens or even hundreds of age groups (bars).

2. BMI Distribution by Age Group for People with Diabetes



Idiom: **Stacked bar chart**

- Two key, One Value

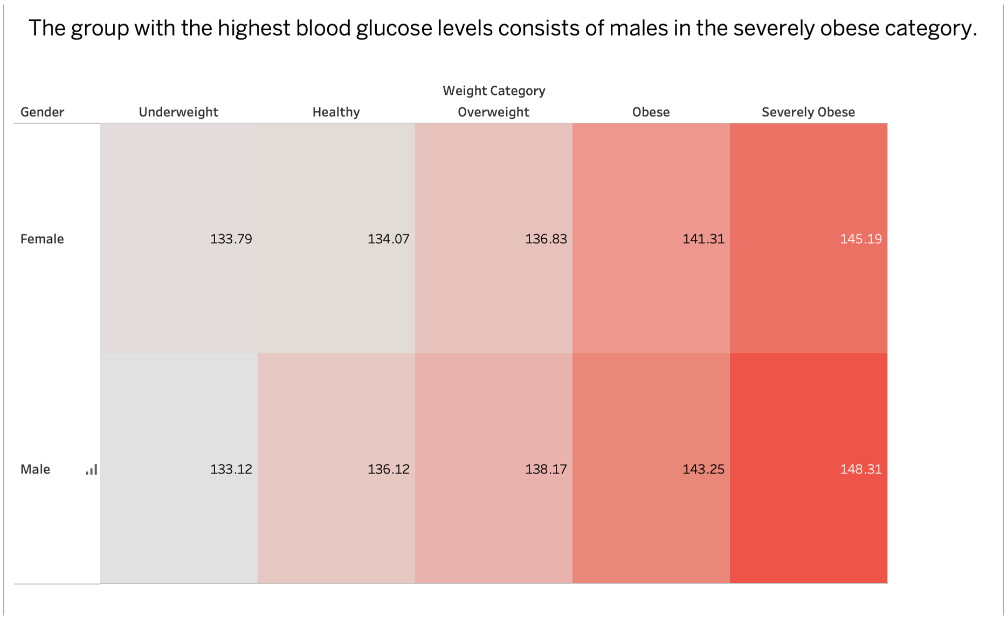
- **Categorical attributes**:
  - **Gender** (Male, Female)
  - **BMI Categories** (Under 25 BMI, over 30 BMI).
- **Quantitative attribute**: Percentage of total count of records for each BMI category within the age groups

- **Mark(1D)**:
  - Vertical stack of **line marks** representing each **BMI category** within **Gender** groups.

- **Channels**:
    - H**orizontal position (Length)**: Encodes the percentage each BMI category contributes to the total for each gender.
    - **Color (hue)**: Differentiates the BMI categories (blue for Under 25 BMI, red for Over 30 BMI).
    - **Spatial regions**:
        - **Aligned region**: The start of each bar is aligned at 0% for the lowest bar component (Under 25 BMI).
        - **Unaligned region**: Subsequent segments in the stack (Over 30 BMI) are unaligned, showing the part-to-whole relationship.
- **Task**:
    - **Part-to-whole relationship**: The chart shows how much of the total diabetic population by gender falls into the two BMI categories.
- **Scalability**:
    - **For the stacked key attribute (BMI categories)**: The chart can handle 10-12 levels (segments) before becoming overcrowded.
    - **For the main key attribute (Gender)**: it can accommodate dozens to hundreds of levels (bars).

DASHBOARD 3: Gender & weight impacts on Blood Glucose & HbA1c

1. Blood Glucose by Gender & Weight



The group with the highest blood glucose levels consists of males in the severely obese category.

	Weight Category				
Gender	Underweight	Healthy	Overweight	Obese	Severely Obese
Female	133.79	134.07	136.83	141.31	145.19
Male	133.12	136.12	138.17	143.25	148.31

Idiom: **Heatmap**

- Two Key, One Value

- Two categorical attributes: **Gender** (Male, Female) and **Weight Category** (Underweight, Healthy, Overweight, Obese, Severely Obese).
- One quantitative attribute: **Blood Glucose Level**.

- Marks(0D)

- **Point Marks**: Each cell is represented by a colored square, where the **color intensity** indicates the blood glucose level.
- **Rectilinear Layout**: A 2D matrix layout with a grid-like structure enables quick scanning across both categorical attributes.

- Channels

- **Color (Quantitative Encoding)**: Use a single-hue, diverging color map,
  - Lower saturation (lighter shades) represents lower blood glucose levels.
  - Higher saturation (darker shades) highlights higher blood glucose levels.
- This approach leverages a consistent hue to avoid distractions, allowing saturation levels to emphasize quantitative differences in blood glucose.

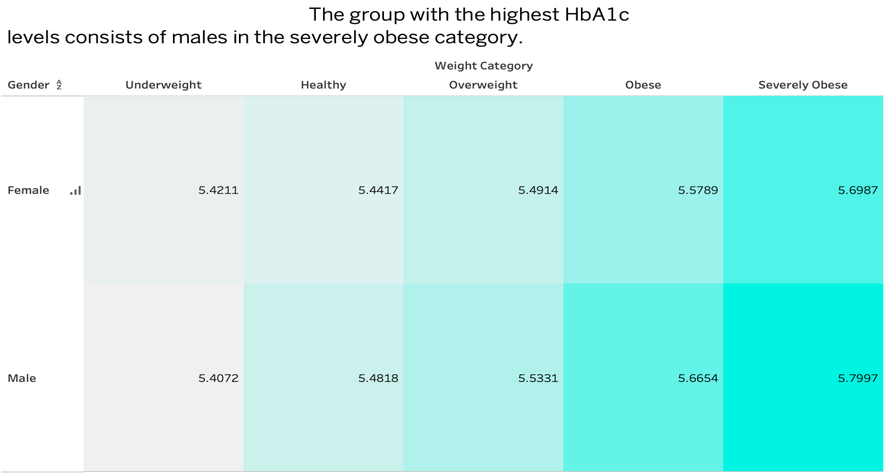
- Task

- **Pattern Identification**: Easily identify which combinations of gender and weight category have higher or lower blood glucose levels.
- **Outlier Detection**: a notably darker cell in the **"Male" and "Severely Obese"** cell indicates this group's higher blood glucose level.

- Scalability

- Designed to handle **up to 1 million data points** by summarizing them into categorical groups.
- Capable of displaying hundreds of categorical levels on both axes and 10+ quantitative levels for blood glucose.

2. HbA1c by Gender and Weight



The group with the highest HbA1c levels consists of males in the severely obese category.

	Weight Category				
Gender	Underweight	Healthy	Overweight	Obese	Severely Obese
Female	5.4211	5.4417	5.4914	5.5789	5.6987
Male	5.4072	5.4818	5.5331	5.6654	5.7997

Idiom: Heatmap

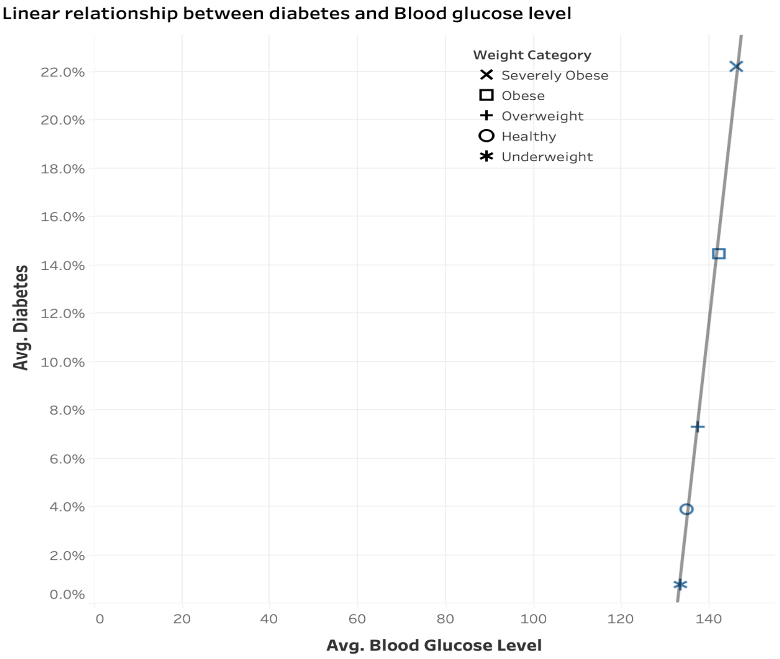
- **Two Key, One Value**:
  - The heatmap represents **Gender** (Male, Female) and **Weight Category** (Underweight, Healthy, Overweight, Obese, Severely Obese) as categorical axes, with **HbA1c levels** as the quantitative value.
- **Marks (0D)**
  - **Point Marks**: Each cell is a colored square indicating the HbA1c level for the corresponding gender and weight category.

- **Rectilinear Layout**: The heatmap has a 2D matrix layout that provides a structured overview, enabling easy comparison between genders across different weight categories.

- **Channels**
  - **Color (Quantitative Encoding)**:
    - The heatmap uses a **single-hue, diverging color map** where:
      - **Lower HbA1c levels** are represented by **lighter shades**.
      - **Higher HbA1c levels** are represented by **darker shades**.
    - This color intensity helps in distinguishing the HbA1c levels across groups without introducing color distractions.

- **Task**
  - **Pattern Identification**:
    - The heatmap makes it straightforward to identify patterns based on gender and weight category. both males and females show an increase in HbA1c levels as weight category moves from Underweight to Severely Obese.
  - **Outlier Detection**:
    - The darkest cell, corresponding to **Male in the Severely Obese category**, highlights an outlier with the highest HbA1c level (5.7997). This suggests that males in this weight category have the highest HbA1c levels in the dataset.
- **Scalability**
    - Designed to handle **up to 1 million data points** by summarizing them into categorical groups.
    - Capable of displaying hundreds of categorical levels on both axes and 10+ quantitative levels for HbA1c.

3. Regression of Diabetes & Blood glucose



Idiom: **Scatterplot**

One Key, Two Values

- **Quantitative attributes**: Average Blood Glucose Level (X-axis) and Average Diabetes Percentage (Y-axis)
- **categorical attributes**: Each point on the graph represents a pair of values (glucose level and diabetes percentage) with different weight category.

Marks(0D):

- **Points with different shapes:** The points on the scatterplot represent different weight categories, distinguished by unique shapes (e.g., cross, square, plus, circle, asterisk).
- Each data point represents the relationship between blood glucose level and diabetes

Channels:

- **Horizontal position** : Represents the average blood glucose level.
- **Vertical position**: Represents the average diabetes percentage.
- **Shape**: Use distinct shapes to distinguish different **Weight Categories** (e.g., Underweight, Healthy, Overweight, Obese, Severely Obese).

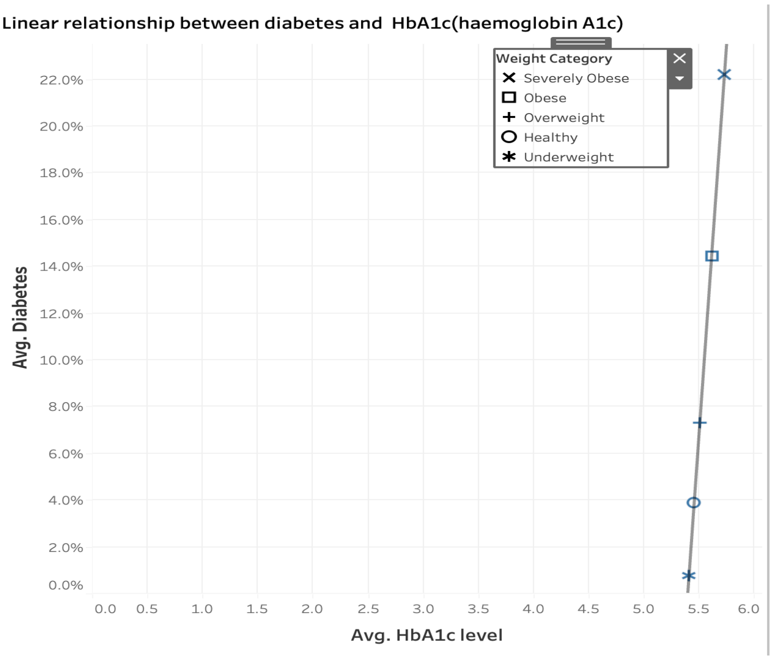
Tasks:

- **Find trends**: The chart shows a linear relationship, where higher glucose levels are associated with a higher percentage of diabetes with weight category.
- **Detect outliers**: The scatterplot may reveal any points that deviate from the linear trend.
- **Correlation**: The positive slope indicates a positive correlation between glucose levels and diabetes percentage.

Scalability:

- This scatterplot design supports **hundreds of data points** while distinguishing points by weight categories.
- Effective for both high- and low-density datasets, provided shape encoding are used to maintain readability.

4. Regression of Diabetes &HbA1c



Idiom: **Scatterplot**

-- One Key, Two Values

- **Quantitative attributes**: Average HbA1c Level (X-axis) and Average Diabetes Percentage (Y-axis)

- **Categorical attribute**: Each point on the graph represents a pair of values (glucose level and diabetes percentage) with different weight category

Marks(0D):

- **Points with different shapes**: The points on the scatterplot represent different weight categories, distinguished by unique shapes (e.g., cross, square, plus, circle, asterisk).
- Each data point represents the relationship between HbA1c Level and diabetes percentage.

Channels:

- **Horizontal position** : Represents the average HbA1c level.
- **Vertical position**: Represents the average diabetes percentage.
- **Shape**: Distinguishes between weight categories (e.g., severely obese, obese, overweight, healthy, underweight).

Tasks:

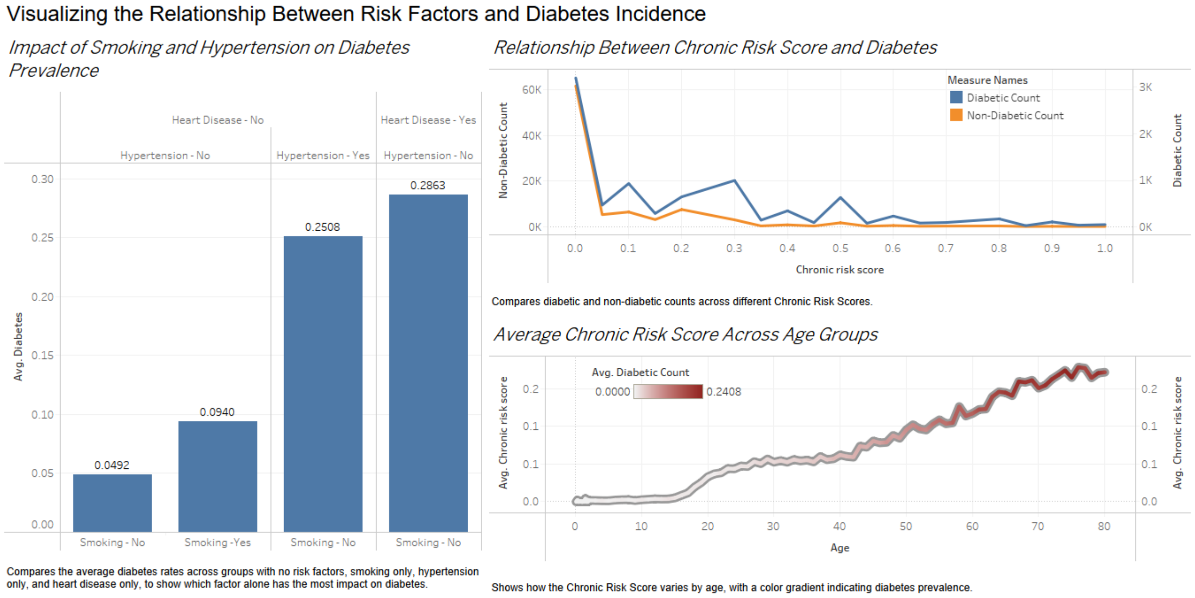
- **Find trends**: The chart shows a linear relationship, where higher HbA1c levels are associated with a higher percentage of diabetes.
- **Detect outliers**: Points that deviate from the trend would indicate outliers.
- **Correlation**: The positive slope indicates a strong correlation between HbA1c levels and diabetes percentage, regardless of weight category.

Scalability:

- This scatterplot design supports **hundreds of data points** while distinguishing points by weight categories.
- Effective for both high- and low-density datasets, provided shape encoding are used to maintain readability.

VI. Visualization Approach and Rationale

H1 : Risk Factors and Diabetes Incidence Dashboard



Visualizing the Relationship Between Risk Factors and Diabetes Incidence

Impact of Smoking and Hypertension on Diabetes Prevalence

Compares the average diabetes rates across groups with no risk factors, smoking only, hypertension only, and heart disease only, to show which factor alone has the most impact on diabetes.

Relationship Between Chronic Risk Score and Diabetes

Compares diabetic and non-diabetic counts across different Chronic Risk Scores.

Average Chronic Risk Score Across Age Groups

Shows how the Chronic Risk Score varies by age, with a color gradient indicating diabetes prevalence.

Plot 1. Task: Analyse Individual Risk Factors on Diabetes Prevalence

Approach & Rationale: To examine how each risk factor independently affects diabetes prevalence, We segmented the data into groups based on smoking, hypertension, and heart disease status. This approach enables us to observe the unique impact of each factor without overlapping influences.

Visualization Choice: Bar Chart

- **Reasoning:** The bar chart is ideal for comparing categorical data, allowing us to directly observe the relative impact of each risk factor on diabetes prevalence. Smoking, hypertension, and heart disease groups are easily distinguishable, making it clear which factor or combination of factors is most strongly associated with diabetes.
- **Impact Analysis:**
  - **Smoking Alone:** The diabetes prevalence among non-smokers without other risk factors is around 4.92%, while for smokers without other risk factors, it doubles to approximately 9.4%, indicating smoking's moderate impact on diabetes risk.
  - **Hypertension Alone:** Individuals with hypertension but without other risk factors have a much higher diabetes prevalence at 25.08%, highlighting hypertension as a significant predictor of diabetes.

    - **Heart Disease Alone:** Diabetes prevalence is highest (28.63%) among those with heart disease alone, supporting the hypothesis that heart disease has the strongest impact on diabetes risk.

Key Takeaway: While each factor contributes to diabetes risk, heart disease and hypertension emerge as the strongest individual predictors.

Plot 2. Task: Explore the Combined Impact of Risk Factors Using Chronic Risk Score

Approach & Rationale: To understand the cumulative effect of multiple risk factors (smoking, hypertension, heart disease),We created a "Chronic Risk Score" that combines these elements. This metric allows us to observe how overall risk correlates with diabetes incidence.

Visualization Choice: Line Chart with Dual Axes

- **Reasoning:** A line chart with dual y-axes effectively shows the contrasting trends of diabetic and non-diabetic counts across the continuous variable of chronic risk score. This approach visually separates the diabetic and non-diabetic populations, allowing for easy comparison across varying risk levels.
- **Impact Analysis:**
    - **Low Chronic Risk Scores:** At lower chronic risk scores, the non-diabetic count is high, suggesting minimal diabetes risk.
    - **Higher Chronic Risk Scores:** As the chronic risk score rises, non-diabetic counts gradually decline, while diabetic counts increase slightly, supporting the hypothesis that a higher cumulative risk score is associated with a greater likelihood of diabetes.

Key Takeaway: The chronic risk score effectively gauges diabetes risk, with higher scores indicating elevated likelihood of diabetes, consistent with the hypothesis.

Plot 3. Task: Visualize Age-Related Patterns in Chronic Risk Scores and Diabetes Prevalence

Approach & Rationale: Since age can influence the likelihood of developing risk factors, We examined how chronic risk scores vary by age to reveal age-related patterns in diabetes prevalence. This approach provides insights into the cumulative effect of age and associated risk factors.

Visualization Choice: Line Chart with Color Gradient

- **Reasoning:** The line chart with a color gradient is well-suited to represent changes across the continuous variable of age while simultaneously displaying diabetes prevalence through color intensity. This design helps highlight age groups with heightened diabetes risk.
- **Impact Analysis:**

- **Younger Age Groups (0-30):** Chronic risk scores are low, and diabetes prevalence is minimal, indicating lower risk factor accumulation in this age range.
- **Middle-Aged and Older Groups (30-80):** Chronic risk scores increase with age, as does diabetes prevalence, particularly after age 60, suggesting that accumulated health risks contribute to diabetes susceptibility in older age groups.

Key Takeaway: Chronic risk scores rise with age, and this increase is associated with higher diabetes prevalence, illustrating the cumulative effect of age-related risk factors on diabetes incidence.

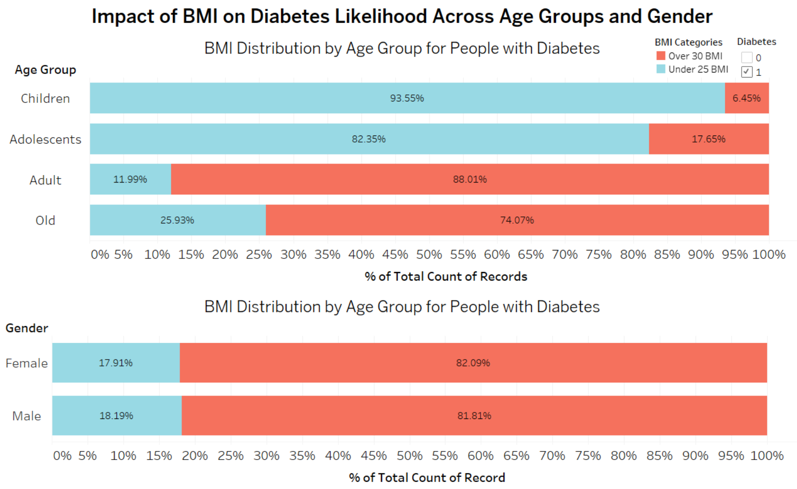
Hypothesis Testing

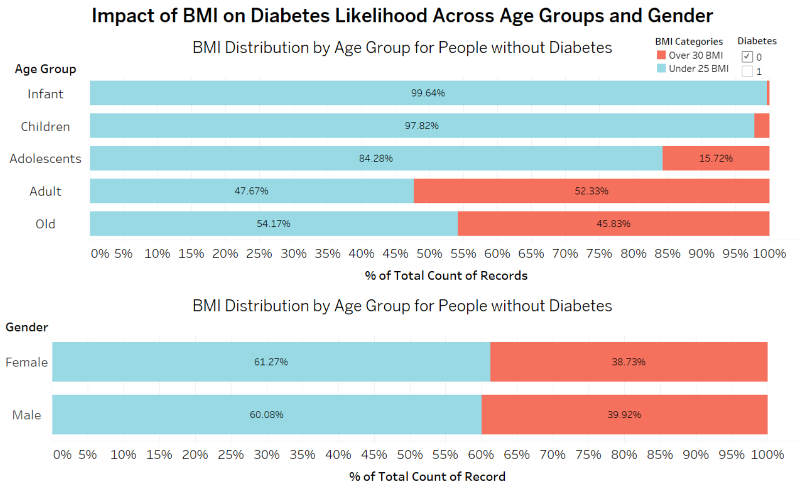
Hypothesis: Patients with higher *Chronic Risk Scores*—a combined measure of hypertension, heart disease, and smoking habits—are more likely to have diabetes. Each health condition contributes differently to the overall risk, with heart disease having the strongest impact.

Conclusion on Hypothesis:

- **Accepted or Rejected:** The hypothesis is **accepted**. The visualizations show a positive correlation between higher chronic risk scores and diabetes incidence, confirming that patients with elevated risk factors are more likely to develop diabetes. Additionally, the bar chart confirms heart disease's strong impact on diabetes risk, supporting the hypothesis that heart disease has the most substantial influence among the individual risk factors.

H2 : Impact of BMI on Diabetes Prevalence Across Age Groups and Gender





Plot 1. Impact of BMI on Diabetes Likelihood Across Age Groups and Gender (With Diabetes)

- **Approach & Rationale**: To examine the distribution of BMI among people with diabetes, we segmented the data by age groups and gender. This segmentation allows us to see how BMI correlates with diabetes prevalence across different demographics.
- **Visualization Choice**: Stacked Bar Chart
- **Reasoning**: The stacked bar chart provides a clear comparison of BMI categories (over 30 vs. under 25) within each age group and gender, allowing us to observe trends in diabetes prevalence based on BMI.
- **Impact Analysis**:
    - **Age Group**:
        - **Children and Adolescents**: The majority of children (93.55%) and adolescents (82.35%) with diabetes have a BMI under 25, indicating that BMI may have a smaller impact on diabetes in these age groups.
        - **Adults and Older Individuals**: Adults (88.01%) and older individuals (74.07%) with diabetes predominantly have a BMI over 30, suggesting a stronger association between high BMI and diabetes in these age groups.
    - **Gender**:
        - **Females**: Among females with diabetes, 82.09% have a BMI over 30, supporting the idea that higher BMI contributes to diabetes risk.
        - **Males**: Similarly, 81.81% of males with diabetes fall into the BMI over 30 category, showing a consistent pattern across genders.
- **Key Takeaway**: Adults and older individuals with diabetes are more likely to have a high BMI (over 30), highlighting the significant role of BMI in diabetes prevalence for these age groups, regardless of gender.

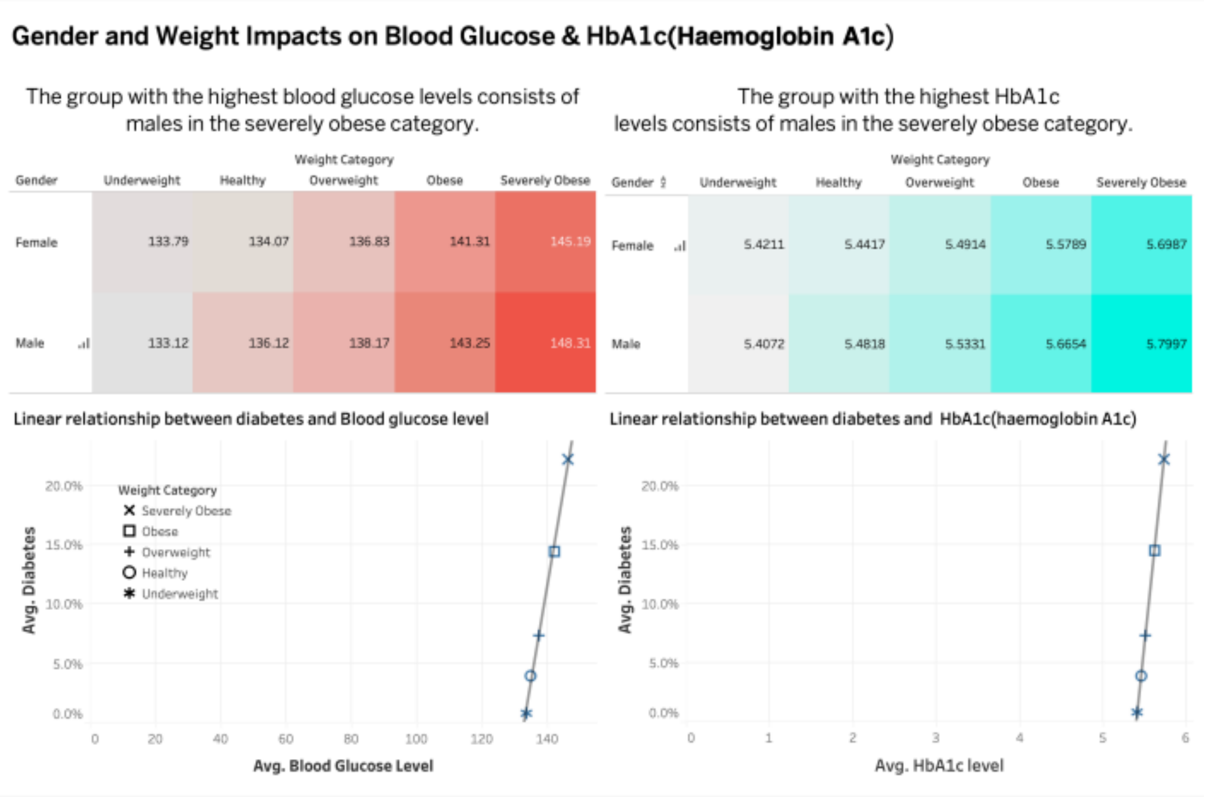
Plot 2. Impact of BMI on Diabetes Likelihood Across Age Groups and Gender (Without Diabetes)

- **Approach & Rationale**: We analysed BMI distribution among people without diabetes, segmented by age groups and gender, to observe if the patterns in BMI vary for non-diabetic individuals compared to those with diabetes.
- **Visualization Choice**: Stacked Bar Chart
- **Reasoning**: Using the same chart type as for the diabetic group enables a direct visual comparison of BMI trends across age groups and genders for those without diabetes.
- **Impact Analysis**:
    - **Age Group**:
        - **Infants and Children**: Nearly all infants (99.64%) and children (97.82%) without diabetes have a BMI under 25, indicating low BMI in younger age groups without diabetes.
        - **Adolescents**: 84.28% of non-diabetic adolescents have a BMI under 25, slightly lower than children but still indicating a trend of lower BMI in younger non-diabetic populations.
        - **Adults and Older Individuals**: Adults (52.33%) and older individuals (45.83%) without diabetes have a nearly even distribution between BMI over 30 and under 25, suggesting that high BMI is less strongly associated with diabetes absence in these groups.
    - **Gender**:
        - **Females**: Among females without diabetes, 61.27% have a BMI under 25, contrasting with females with diabetes who predominantly have a BMI over 30.
        - **Males**: Similarly, 60.08% of non-diabetic males have a BMI under 25, showing a consistent lower-BMI trend among non-diabetic individuals of both genders.
- **Key Takeaway**: Non-diabetic individuals across all age groups are more likely to have a lower BMI (under 25), especially in younger age groups. This contrasts with the higher BMI trend seen in diabetic adults and older individuals.

Hypothesis Testing

- **Hypothesis**: People with a BMI over 30 are more likely to have diabetes compared to those with a BMI under 25, regardless of their age or gender.
- **Conclusion on Hypothesis**: **Accepted**. The analysis supports the hypothesis, showing that individuals with diabetes are predominantly in the BMI over 30 category, especially in adults and older individuals, and this trend is consistent across genders.

H3 : Gender & Weight Impacts on Diabetes Risk



Gender and Weight Impacts on Blood Glucose & HbA1c(Haemoglobin A1c)

The group with the highest blood glucose levels consists of males in the severely obese category.

Gender	Weight Category				
	Underweight	Healthy	Overweight	Obese	Severely Obese
Female	133.79	134.07	136.83	141.31	145.19
Male	133.12	136.12	138.17	143.25	148.31

The group with the highest HbA1c levels consists of males in the severely obese category.

Gender	Weight Category				
	Underweight	Healthy	Overweight	Obese	Severely Obese
Female	5.4211	5.4417	5.4914	5.5789	5.6987
Male	5.4072	5.4818	5.5331	5.6654	5.7997

Linear relationship between diabetes and Blood glucose level

Linear relationship between diabetes and HbA1c(haemoglobin A1c)

Plot 1. Gender and Weight Impacts on Blood Glucose Levels

- **Approach & Rationale**: The data was segmented by gender and weight categories, defined using BMI values to create groups (underweight, normal weight, overweight, obese, and severely obese). This approach provides clarity on how gender and weight influence blood glucose levels independently.
- **Visualization Choice**: Heatmap Table
- **Reasoning**: The heatmap table allows for a color-coded comparison across weight categories and gender, making it easy to see which groups have higher blood glucose levels. This format highlights that males and severely obese individuals have particularly elevated levels.
- **Impact Analysis**:
  - **Highest Blood Glucose Levels**: Severely obese males have the highest average blood glucose level (148.31), followed by severely obese females (145.19). This finding supports the hypothesis that both gender and severe obesity are associated with higher blood glucose levels.
  - **Gender Differences**: Across all weight categories, males generally show higher average blood glucose levels than females.
- **Key Takeaway**: Gender and weight category contribute to variations in blood glucose levels, with males and severely obese individuals exhibiting the highest levels. This supports the hypothesis that these groups have an elevated diabetes risk based on blood glucose levels.

Plot 2. Gender and Weight Impacts on HbA1c Levels

- **Approach & Rationale**: HbA1c levels were examined by gender and weight categories to determine patterns. The weight categories were based on BMI ranges, providing consistency across analyses.
- **Visualization Choice**: Heatmap Table
- **Reasoning**: The heatmap's color gradients provide a clear visual comparison of HbA1c levels across gender and weight groups, making it easy to identify that males and severely obese individuals have the highest HbA1c levels.
- **Impact Analysis**:
  - **Highest HbA1c Levels**: Severely obese males have the highest HbA1c level at 5.7997, followed by severely obese females at 5.6987. This reinforces the hypothesis that males and severely obese individuals tend to have higher HbA1c levels, further indicating a higher diabetes risk.
  - **Gender and Weight Influence**: Males consistently show higher HbA1c levels than females within each weight category.
- **Key Takeaway**: Gender and weight significantly influence HbA1c levels, with the highest levels observed in severely obese males. This aligns with the hypothesis, suggesting HbA1c as a robust diabetes risk indicator in high-risk groups.

Plot 3. Linear Relationship Between Diabetes and Blood Glucose Level

- **Approach & Rationale**: A scatter plot with a trend line was used to analyze the relationship between blood glucose levels and diabetes prevalence across weight categories, helping to evaluate blood glucose as a predictor of diabetes risk.
- **Visualization Choice**: Scatter Plot with Trend Line
- **Reasoning**: The scatter plot visually shows the relationship between blood glucose levels and diabetes prevalence, with a trend line to represent the linear correlation. The weight categories are marked with distinct symbols for easier differentiation.
- **Impact Analysis**:
  - **Trend Observation**: Diabetes prevalence increases with higher blood glucose levels, especially in severely obese individuals, who show a diabetes prevalence approaching 20%.
  - **Statistical Significance**: The p-value for blood glucose as a predictor of diabetes is 0.0001355, which is statistically significant (below the 0.05 threshold), confirming the reliability of the observed trend.
- **Key Takeaway**: Higher blood glucose levels are strongly associated with diabetes prevalence, especially in severely obese individuals. The low p-value supports blood glucose as a statistically significant predictor of diabetes risk.

Plot 4. Linear Relationship Between Diabetes and HbA1c Level

- **Approach & Rationale**: The scatter plot examines the correlation between HbA1c levels and diabetes prevalence to evaluate HbA1c's strength as a diabetes predictor, especially in comparison to blood glucose levels.
- **Visualization Choice**: Scatter Plot with Trend Line
- **Reasoning**: This scatter plot with a trend line visually demonstrates the stronger linear relationship between HbA1c levels and diabetes prevalence. Each weight category is distinctly marked, making it easy to observe differences.
- **Impact Analysis**:
  - **Stronger Trend**: The slope of the line is steeper for HbA1c than for blood glucose, indicating a stronger correlation between HbA1c levels and diabetes prevalence, particularly in severely obese individuals (where diabetes prevalence exceeds 20%).
  - **Statistical Significance**: The p-value for HbA1c as a predictor is less than 0.0001, indicating an even stronger statistical significance compared to blood glucose.
- **Key Takeaway**: HbA1c levels have a stronger correlation with diabetes prevalence than blood glucose levels, as shown by both the steeper trend line and lower p-value. This aligns with the hypothesis that HbA1c is a more reliable indicator of diabetes risk.

Hypothesis Testing

- **Hypothesis**: Gender and weight category influence blood glucose and HbA1c levels, with males and those in the severely obese category showing higher levels. HbA1c may also be a stronger indicator of diabetes risk than blood glucose levels alone.
- **Conclusion on Hypothesis**:
  - **Accepted**: The data strongly supports the hypothesis. Both the heatmaps and scatter plots reveal that males and severely obese individuals have higher blood glucose and HbA1c levels. The low p-values confirm that both blood glucose and HbA1c are statistically significant predictors of diabetes risk, with HbA1c showing a slightly stronger relationship due to its lower p-value.
  - **Reason**: The heatmaps confirm that severely obese males have the highest levels of blood glucose and HbA1c, consistent with the hypothesis. The steeper linear relationship and lower p-value for HbA1c further support the conclusion that HbA1c is a stronger indicator of diabetes risk.

VII. Conclusion:

The Diabetes Risk Analysis Dashboard project provides critical insights into diabetes risk factors, supporting healthcare professionals in risk assessment. This report outlines the approach taken, highlights findings from the data visualizations, and offers actionable recommendations based on the analysis. By examining the influence of individual and combined health conditions—such as heart disease, hypertension, smoking, and BMI—the dashboard underscores significant predictors of diabetes, with HbA1c emerging as a particularly strong indicator.