Introduction to the Dataset

The Diabetes Dataset is a widely used resource in data science, particularly for tasks involving machine learning and health informatics. It's designed to facilitate the development and testing of models aimed at predicting outcomes related to diabetes. This dataset typically includes several features or variables that are considered relevant for studying diabetes, such as glucose concentration, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, age, and the outcome (whether the individual has diabetes or not).

Data Source

The specific dataset you mentioned is hosted on Kaggle, a platform that serves as a hub for data scientists and machine learning practitioners to find and publish datasets, explore models, and engage with a wide community on data science projects. Kaggle datasets are often used for competitions, educational purposes, and research projects. The Diabetes Dataset on Kaggle, provided by the user "mathchi," is accessible for download and use under the terms specified on the platform, usually for educational and research purposes.

Dataset Column description

Pregnancies: The number of times the patient has been pregnant. This feature is relevant because gestational diabetes is a common condition, and the number of pregnancies can influence the risk of developing type 2 diabetes later in life.

Glucose: The plasma glucose concentration a 2 hours in an oral glucose tolerance test. High blood glucose levels may indicate an impaired ability to process sugar, which is a hallmark of diabetes.

BloodPressure: Diastolic blood pressure (mm Hg). While not directly related to diabetes, high blood pressure is commonly found in people with diabetes and can complicate the condition.

SkinThickness: Triceps skin fold thickness (mm). This measure can be an indicator of body fat or insulin resistance. In some datasets, this variable might have many missing values or zeros, indicating that it was not measured.

Insulin: 2-Hour serum insulin (mu U/ml). Insulin levels are directly related to diabetes, as insulin is the hormone responsible for regulating blood glucose levels.

BMI: Body mass index (weight in kg/(height in m)^2). BMI is a key factor in diabetes risk, as obesity is a major risk factor for developing type 2 diabetes.

DiabetesPedigreeFunction: A function that scores the likelihood of diabetes based on family history. It's a measure of the genetic influence on the patient's likelihood of developing the disease.

Age: Age in years. Age is another risk factor, as the risk of developing type 2 diabetes increases with age.

Outcome: The class variable (0 or 1) indicating whether the patient has diabetes (1) or not (0). This is the target variable for predictive modeling.

Research Questions

When working with the Diabetes Dataset, several research questions can be formulated to explore the data and potentially uncover insights related to diabetes. Here are a few examples:

Risk Factor Analysis: What are the most significant risk factors associated with diabetes, based on the features available in the dataset?

Feature Relationships: How do the various features (like glucose levels, BMI, age, etc.) relate to each other, and do these relationships differ between individuals with and without diabetes?

Demographic Analysis: Is there a demographic pattern (such as age) that is more prevalent in the diabetic population within this dataset?

Data Cleaning Documentation

This documentation details the data cleaning process undertaken on the Diabetes Dataset. The focus was on addressing placeholder zeros in several key columns: BMI, BloodPressure, Glucose, SkinThickness, and Insulin. These zeros are not physiologically plausible and were treated as missing values.

Initial Checks

Initial validation confirmed there were no NA, NULL, or empty string values across the dataset. However, a targeted check revealed **placeholder zeros** in five critical columns, necessitating a tailored approach for each.

Zeros were found in the following columns:

BMI (11 zeros)

BloodPressure (35 zeros)

Glucose (5 zeros)

SkinThickness (227 zeros)

Insulin (374 zeros)

1. Handling Zeros in BMI, blood pressure, and Glucose

For columns with relatively fewer zeros (BMI, blood pressure, Glucose), median imputation was deemed appropriate due to its simplicity and the ability to preserve the central tendency without being affected by outliers.

2. Handling Zeros in SkinThickness and Insulin

Given the higher number of zeros in SkinThickness and Insulin, a more sophisticated method was required to impute these missing values without introducing bias. Predictive mean matching (PMM) via the mice package was selected for its robustness and ability to handle multivariate data.

Post-Cleaning Validation

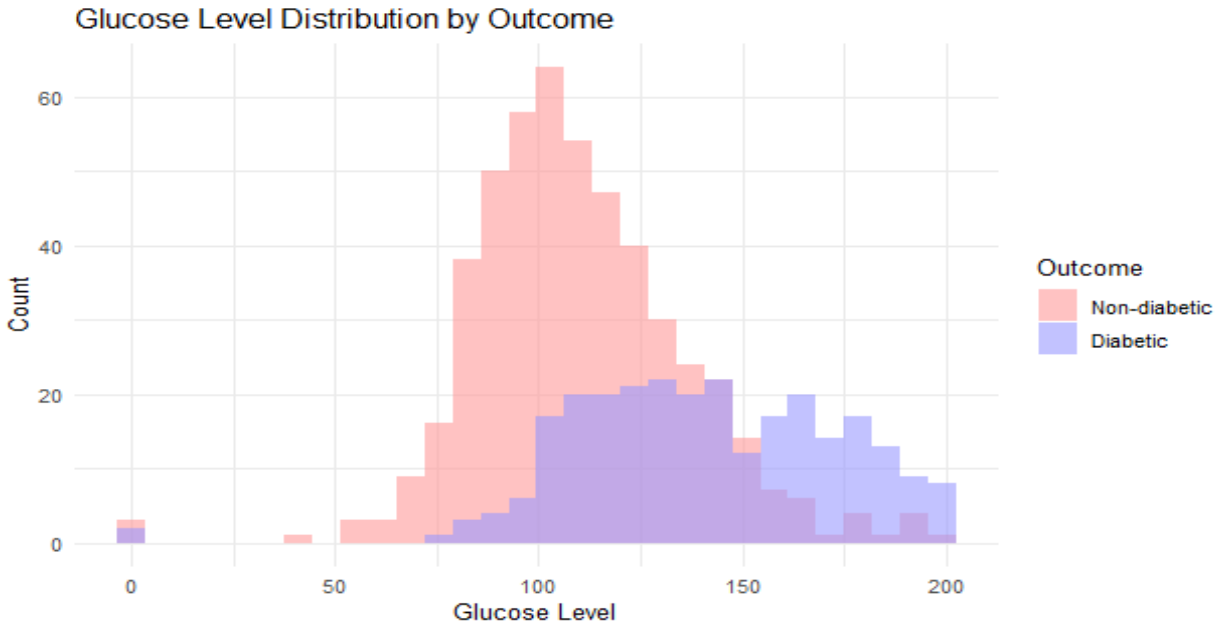
After imputation, the dataset was re-evaluated to ensure no zeros remained in the specified columns, validating the effectiveness of the cleaning process. The integrity and distribution of the data were also assessed to ensure the imputation did not introduce any unintended biases.

Conclusion

The data cleaning process for the Diabetes Dataset involved careful consideration of each variable's physiological plausibility and the proportion of placeholder zeros. Median imputation was applied for columns with fewer zeros to maintain simplicity, while predictive mean matching was used for columns with more extensive missing data, offering a balance between sophistication and practicality. This meticulous approach ensures the dataset is primed for accurate and reliable analysis.

Data Visualization section

Research Question Number 1 Risk Factor Analysis



Distribution Shapes: The distribution for non-diabetics (presumably in purple) peaks at lower glucose levels and appears normally distributed, centering around the normal glucose range. In contrast, the distribution for diabetics (presumably in pink) is right-skewed, indicating higher glucose levels overall.

Glucose Level Peaks: The non-diabetic peak is around the normal fasting glucose level range, possibly between 70 and 100 mg/dL. The diabetic peak is shifted towards the right, likely in the range indicative of diabetes (possibly around 120 to 140 mg/dL), suggesting that individuals with diabetes have higher glucose levels on average.

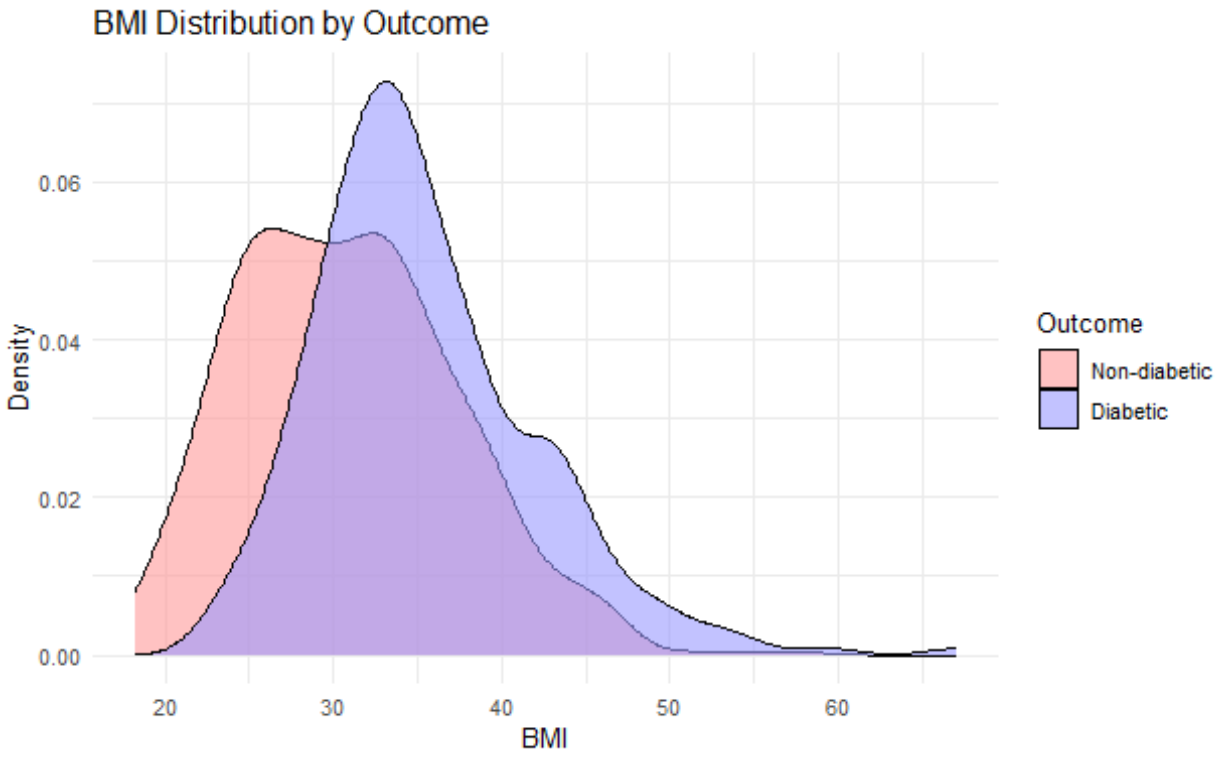
Overlap of Distributions: There is an overlap between the two distributions, showing that there is a range of glucose levels where both non-diabetic and diabetic individuals are present. This might reflect the complexity of diabetes diagnosis, where glucose levels alone don't always perfectly differentiate between outcomes.

Width of the Distributions: The non-diabetic distribution is narrower, which suggests less variability in glucose levels among non-diabetic individuals. The diabetic distribution is wider, showing more variability, which could be due to a variety of factors influencing glucose levels in diabetics.

Counts: The highest count of individuals for both categories appears to be in the normal to slightly elevated glucose level range, indicating that most people in the dataset, whether diabetic or not, have glucose levels that don't reach the extremes.

From these observations, it can be concluded that higher glucose levels are associated with diabetes, corroborating the medical understanding that elevated glucose levels are a hallmark of diabetes. However, due to the overlap in glucose levels between the two groups, additional factors likely play a role in the diagnosis of diabetes.

BMI Distribution



Peak Distribution: The peak of the BMI for non-diabetic individuals (in purple) is in the lower range than for diabetic individuals (in pink). This suggests that on average, non-diabetic individuals in this dataset have a lower BMI.

Spread and Shape: The distribution for non-diabetics is somewhat more pointed and has a narrower spread, indicating less variability in BMI among non-diabetics. On the other hand, the diabetic group shows a broader distribution, suggesting a wider range of BMI values among diabetic individuals.

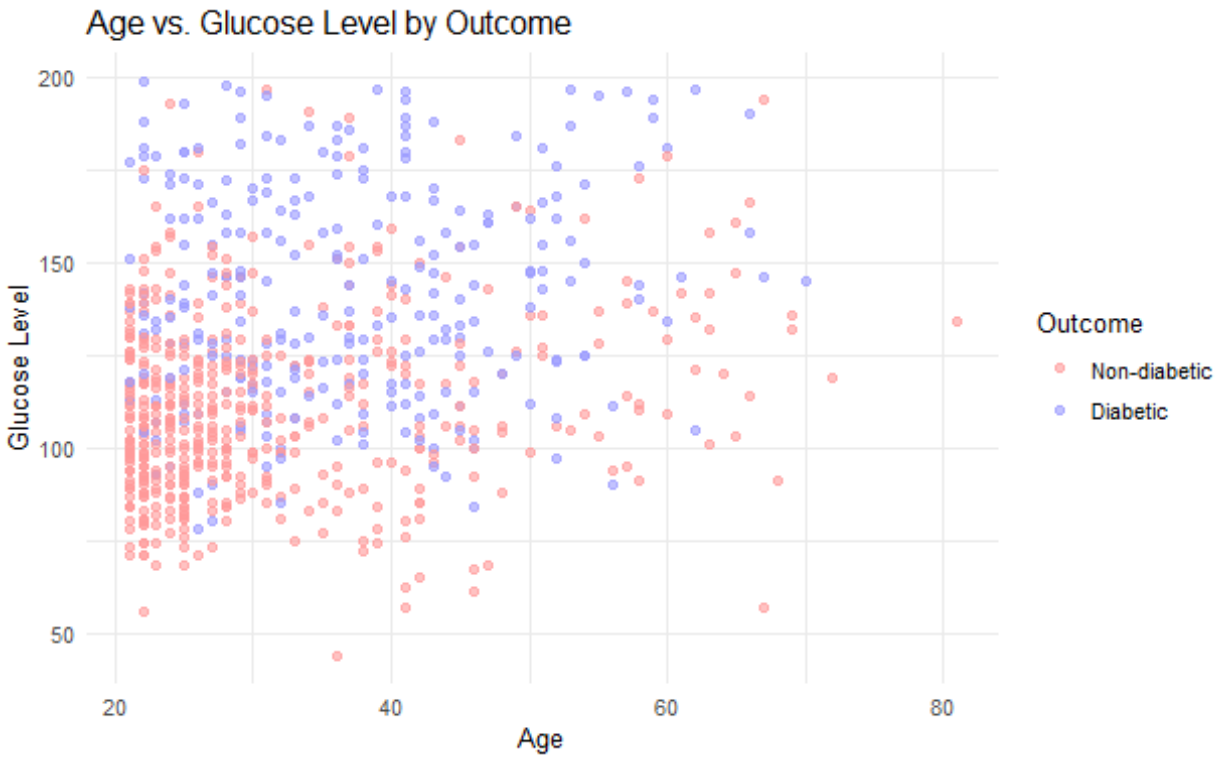
Overlap of Distributions: There is considerable overlap in BMI values between diabetic and non-diabetic individuals. This implies that while there is a general trend for higher BMI among diabetics, there is not a clear-cut BMI value that separates the two groups.

Density Values: The density plot shows that a substantial proportion of both non-diabetic and diabetic individuals have BMI values within what is considered a normal to overweight range, but there is a noticeable shift towards higher BMI values in the diabetic group.

BMI as a Risk Factor: Higher BMI values are more commonly associated with the diabetic group, aligning with the understanding that a higher BMI can be a risk factor for diabetes.

From this visualization, it's apparent that BMI is distributed differently between non-diabetic and diabetic individuals, with a trend towards higher BMI values in the diabetic group. However, due to the significant overlap, BMI alone is not a definitive predictor of diabetes, and it should be considered in conjunction with other factors.

Age vs glucose levels



The plot shows two distinct groups: non-diabetic individuals represented by one color (presumably blue) and diabetic individuals represented by another color (presumably red).

Glucose Levels Across Ages: For both diabetics and non-diabetics, glucose levels are spread across all ages. This indicates that while age is a variable, it's not the sole determinant of glucose levels.

Concentration of Diabetic Individuals: There's a noticeable concentration of diabetic individuals with higher glucose levels across the age spectrum, not confined to any specific age group.

Non-Diabetic Glucose Levels: Non-diabetic individuals mostly cluster at lower glucose levels, though there is some overlap with the diabetic individuals' glucose levels, especially in the middle range.

Age Distribution: The ages of the individuals in the dataset are distributed broadly from young adults to the elderly. There's no clear indication that glucose levels increase with age in a significant manner for non-diabetics.

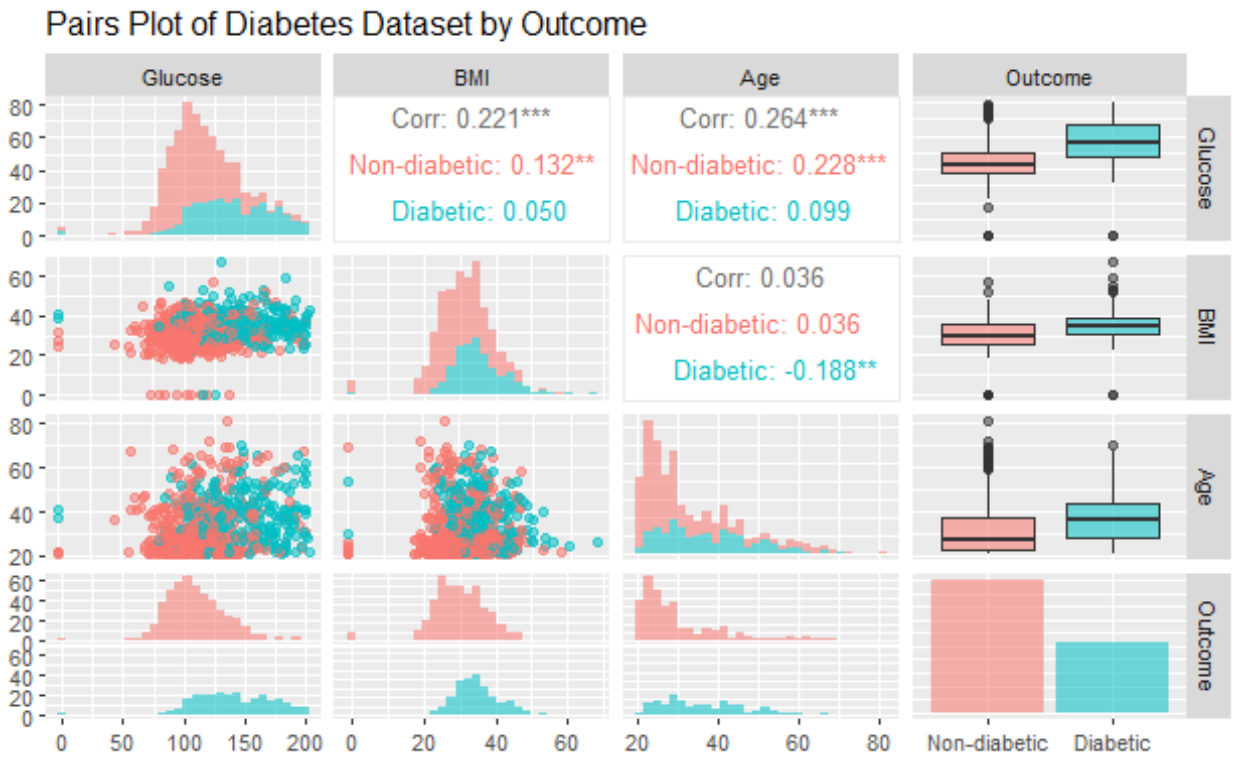
Glucose Level Variability: Diabetic individuals show more variability in glucose levels, with many data points above the standard clinical threshold for diabetes diagnosis (usually around 126 mg/dL fasting).

Overlap and Age-Related Trends: Some younger individuals have high glucose levels typically associated with diabetes, and some older individuals have glucose levels in the non-diabetic range, showing that the relationship between age and glucose level is not straightforward.

From this scatter plot, it's evident that higher glucose levels are associated with diabetes, but there's not a strict age dependency. Instead, diabetes affects a broad range of ages, with individuals showing elevated glucose levels not confined to older age groups. The visualization reinforces that age alone is not a reliable predictor of diabetes, and there must be other factors at play that influence glucose levels in individuals.

Research Question Number 2 Feature Relationships

Relationship between glucose, age, and BMI



The pairs plot presents a matrix of scatter plots, histograms, and correlation coefficients for three variables (Glucose, BMI, and Age) by diabetes status (non-diabetic or diabetic):

Glucose and BMI:

There is a positive correlation between Glucose and BMI for the entire dataset, stronger in non-diabetics than in diabetics.

Scatter plots show a denser clustering for non-diabetics at lower values of both BMI and Glucose, while diabetic individuals are more spread out across higher values of both variables.

Glucose and Age:

The correlation between Glucose and Age is relatively weak. It's more pronounced in non-diabetic individuals compared to diabetic individuals.

The scatter plot does not display a clear trend between age and glucose levels for either group, suggesting other factors may contribute more significantly to glucose levels than age alone.

BMI and Age:

The correlation between BMI and Age is very weak across both non-diabetics and diabetics, with a very slight negative correlation in diabetics.

This implies that there is no strong relationship between age and BMI within this dataset.

Histograms (Diagonal):

Histograms show the distribution of each variable separated by outcome. The distributions of Glucose and BMI are visibly different between non-diabetics and diabetics, with diabetics tending to have higher values.

Age distribution is similar between both outcomes, suggesting age by itself is not distinctly different between non-diabetic and diabetic individuals in this dataset.

Correlation Coefficients (Upper Diagonal):

The numeric correlation coefficients are displayed above the diagonal, showing the degree of linear relationship between the variable pairs.

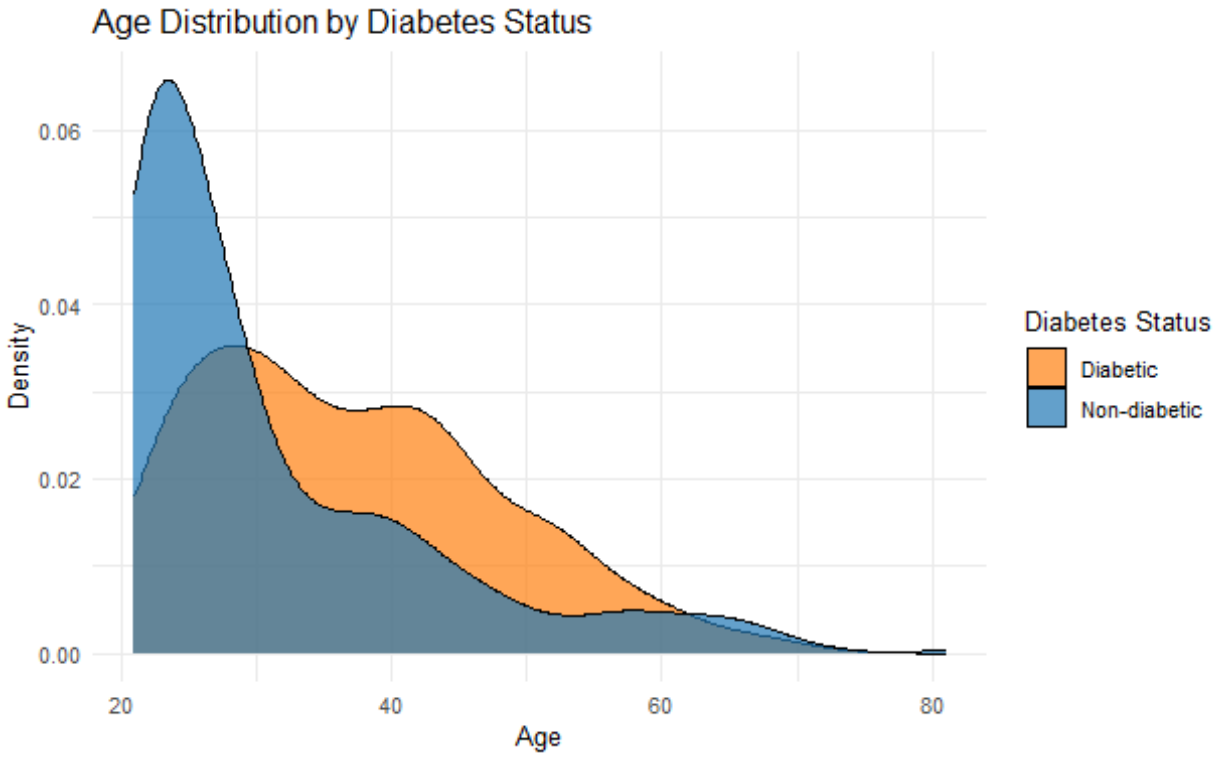
The significant correlations are marked with asterisks, with more asterisks denoting higher levels of statistical significance.

For non-diabetic individuals, there are significant positive correlations between BMI & Glucose and Age & Glucose. In contrast, diabetic individuals show a very weak positive correlation for BMI & Glucose and a weak negative correlation for BMI & Age.

The visualization suggests that while there are relationships between these variables, they are not uniform across outcomes. Higher glucose levels and BMI are more prevalent in diabetic individuals, while age does not seem to play a significant role in distinguishing between the outcomes. The correlations and distributions indicate that diabetes status is associated with both BMI and glucose, but the effect of age is less clear. It also highlights the complexity of the relationships among these variables in the context of diabetes.

Research Question Number 3

Age Distribution by diabetes status



Age Distribution for Non-diabetics: The blue area, presumably representing non-diabetic individuals, has a peak in the younger age range and a steep decline as age increases. This suggests that within this dataset, a larger proportion of younger individuals are classified as non-diabetic.

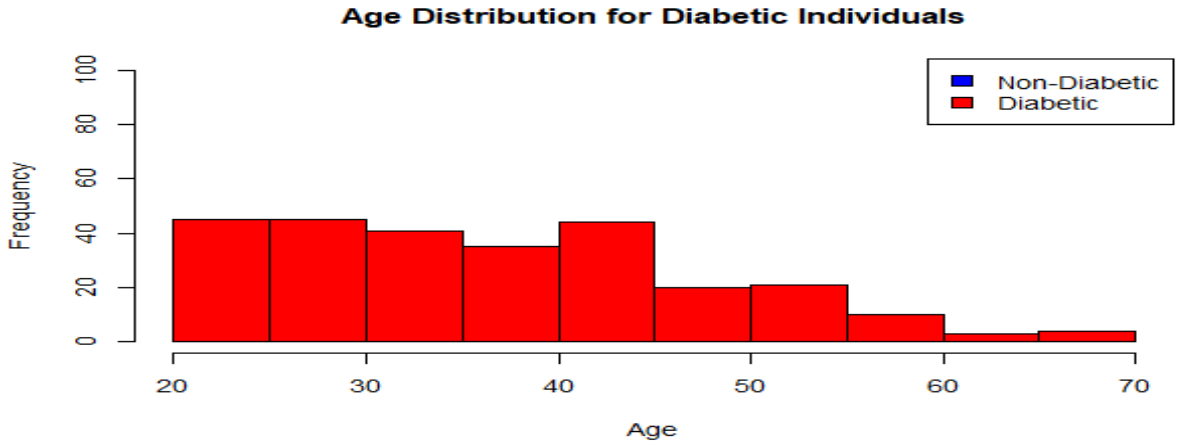
Age Distribution for Diabetics: The orange area, presumably representing diabetic individuals, shows a flatter distribution, with a broader spread across age ranges, peaking at a later age than the non-diabetics. This indicates that diabetes is more evenly distributed across different ages, but with a tendency towards middle-aged and older adults.

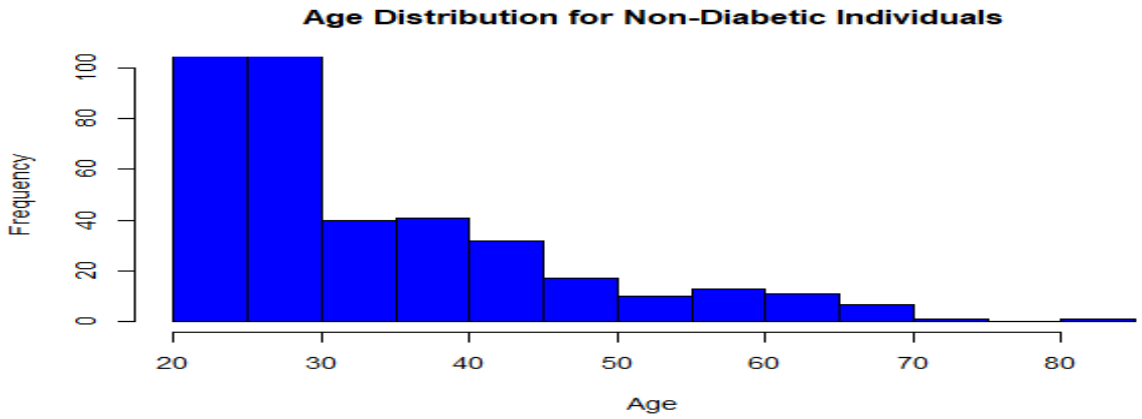
Overlap Between Groups: There is a notable overlap in the middle age range, indicating that middle-aged individuals in the dataset are distributed across both diabetic and non-diabetic categories.

Higher Ages: The diabetic group shows a longer tail into the higher ages compared to the non-diabetic group, which drops off more sharply. This could suggest that the proportion of diabetic individuals does not decrease as sharply with age as it does for non-diabetics.

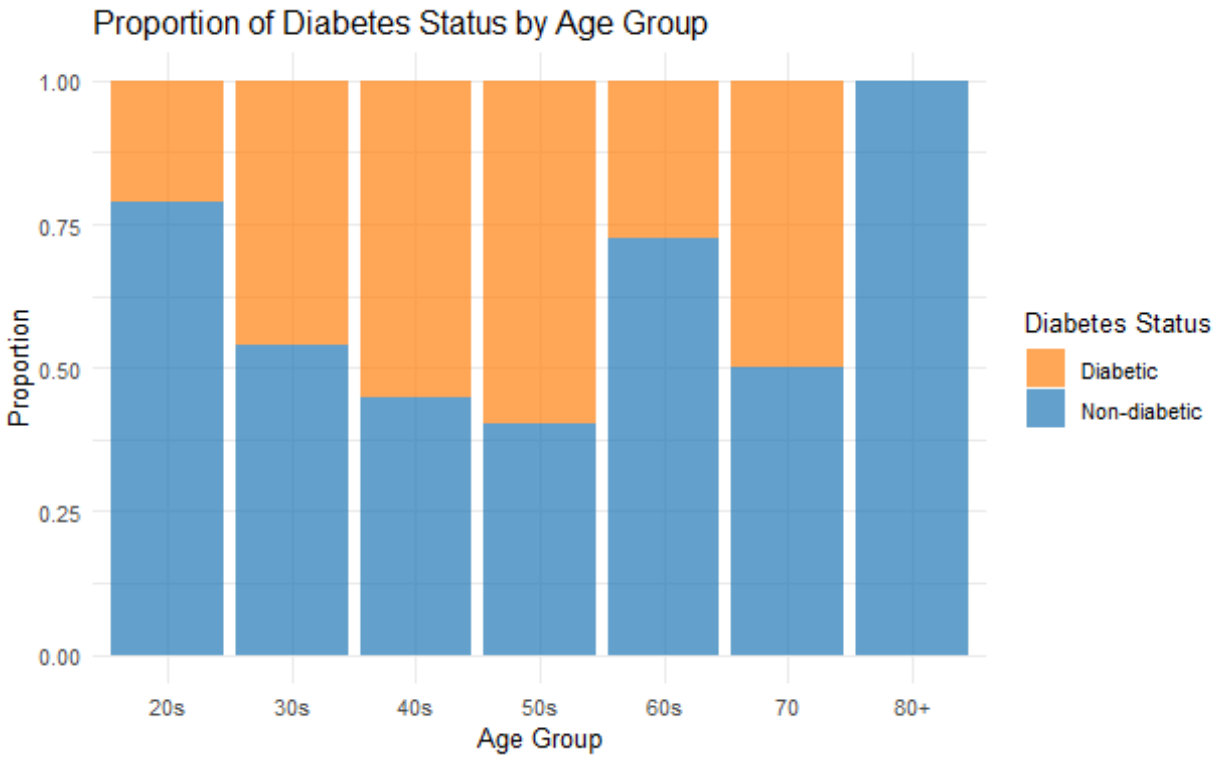
Diabetes Prevalence: While this plot alone doesn't show prevalence rates, the shift towards older ages in the diabetic group aligns with the general understanding that the prevalence of diabetes increases with age.

From this visualization, one might infer that age is associated with diabetes status, with the likelihood of being diabetic increasing with age. However, since there is significant overlap in the middle age range, age alone is not a distinct separator between diabetics and non-diabetics. Other factors in conjunction with age likely contribute to the risk of diabetes.





proportion of diabetic to non-diabetic individuals across different age groups



Proportion Stability: Across the age groups from the 20s to 60s, the proportion of individuals with diabetes (shown in orange) to those without (shown in blue) remains relatively consistent.

Increase in Proportion with Age: There is a noticeable increase in the proportion of individuals with diabetes in the 70s and even more so in the 80+ age group.

Age as a Risk Factor: The chart suggests that while age is a risk factor for diabetes, its impact becomes more pronounced in the older age categories. Younger age groups show a significant number of individuals with diabetes, but the proportion grows in the older populations.

Younger Age Groups: The younger age groups (20s to 60s) show a substantial presence of diabetes, indicating that the condition affects a wide range of ages.

Oldest Age Group: The 80+ age group has the highest proportion of diabetic to non-diabetic individuals compared to the other age groups, aligning with the understanding that the risk of diabetes increases with age.

Conclusion of Diabetes Dataset Analysis

The analysis of the Diabetes Dataset revealed several important insights into the relationship between various health metrics and the presence of diabetes. The data cleaning process was a critical step that involved identifying and rectifying issues with implausible zero values in key variables, ensuring the quality and integrity of the data for accurate analysis.

Key Findings from Data Visualization:

Glucose Levels: As expected, glucose levels were significantly higher in individuals diagnosed with diabetes compared to non-diabetic individuals. The distribution was right-skewed for diabetics, with a wider range of higher glucose values.

BMI Trends: The density plots highlighted a higher BMI range among diabetic individuals, suggesting an association between increased body mass index and diabetes.

Age Dynamics: Although age is a known risk factor for diabetes, the scatter plots and density distributions did not show a strong direct relationship between age and glucose levels within the diabetic and non-diabetic groups. Instead, age distribution was relatively even across the dataset, with a slight increase in diabetes prevalence in the older population (70+ years).

Correlation Between Metrics: The pairs plot illustrated that while there were positive correlations between BMI and glucose levels, they were not uniform across diabetic and non-diabetic individuals. Age showed a very weak correlation with BMI and glucose levels, indicating that the interplay between these factors and diabetes is complex.

Proportion of Diabetic Status Across Ages: Diabetes was present across all age groups, with the proportion of diabetic individuals increasing slightly in the older age groups.

The analysis suggests that diabetes is influenced by a combination of factors, with glucose levels and BMI being the most indicative of diabetic status within this dataset. Age alone did not emerge as a decisive factor; however, there is an indication that the risk of diabetes increases with age, especially after 70 years.