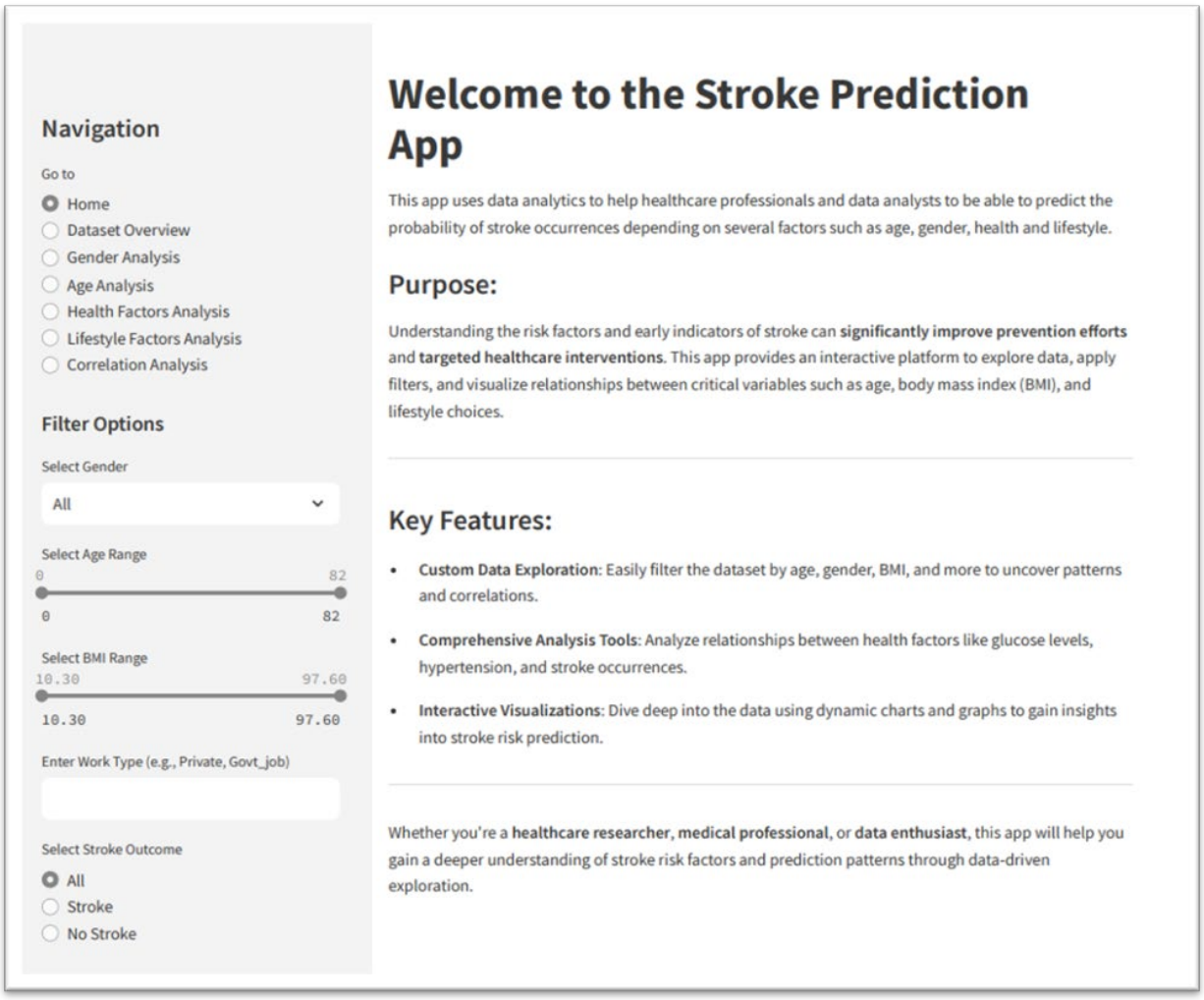
4. Description of the Streamlit application



The Stroke Prediction App is designed for healthcare professionals, researchers, and data analysts. The main objective of the application is to help predict stroke occurrences based on several factors such as age, health status and lifestyle choices.

The **Navigation Panel** located on the left side includes options like:

- Home: This is the introduction to the application.
- Dataset Overview: Overview of the data used in the analysis
- Gender Analysis: Allows analysis based on gender differences.
- Age Analysis: Explores age-related factors in stroke prediction.

- Health Factors Analysis: Looks into health variables like blood pressure or BMI.
- Lifestyle Factors Analysis: Examines lifestyle choices (e.g., smoking, physical activity) and their impact.
- Correlation Analysis: Investigates the relationships between various factors

The **Filter Options** panel located on the left side as well provides several filtering options, such as:

- Select Gender: Drop-down menu to filter data by gender.
- Select Age Range: Slider to set a specific age range.
- Select BMI Range: Slider to define the BMi range for data filtering.
- Enter Work Type: Text input for specifying work types (e.g., private, government jobs).
- Select Stroke Outcome: Option to filter the dataset by stroke outcomes (All Stroke, No Stroke).

The **Main Panel** is the central part of the application which displays information about its purpose and key features:

Purpose: To use data analytics for better understanding and predicting stroke risk factors, thereby supporting targeted healthcare interventions.

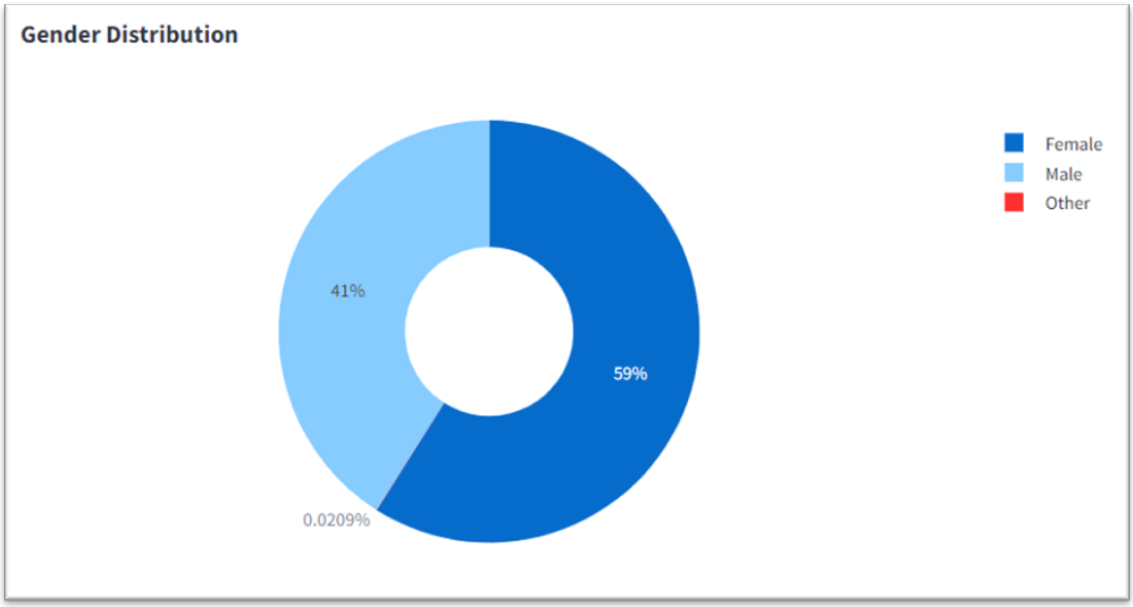
Key Features:

- **Custom Data Exploration:** Users can filter the dataset by various parameters to explore patterns and correlations.
- **Comprehensive Analysis Tools:** The application helps analyse relationships between health factors (e.g., glucose levels, hypertension) and stroke occurrences.
- **Interactive Visualisations:** The application includes dynamic charts and graphs for in-depth risk analysis.

The application therefore aims to support professionals in exploring stroke data comprehensively, offering insights into prevention and targeted intervention strategies.

5. Data Analysis

a. Gender Analysis



The graph that is being displayed above is a doughnut chart that displays the gender distribution for the dataset used in the Stroke Prediction Application. The data analysis is as follows:

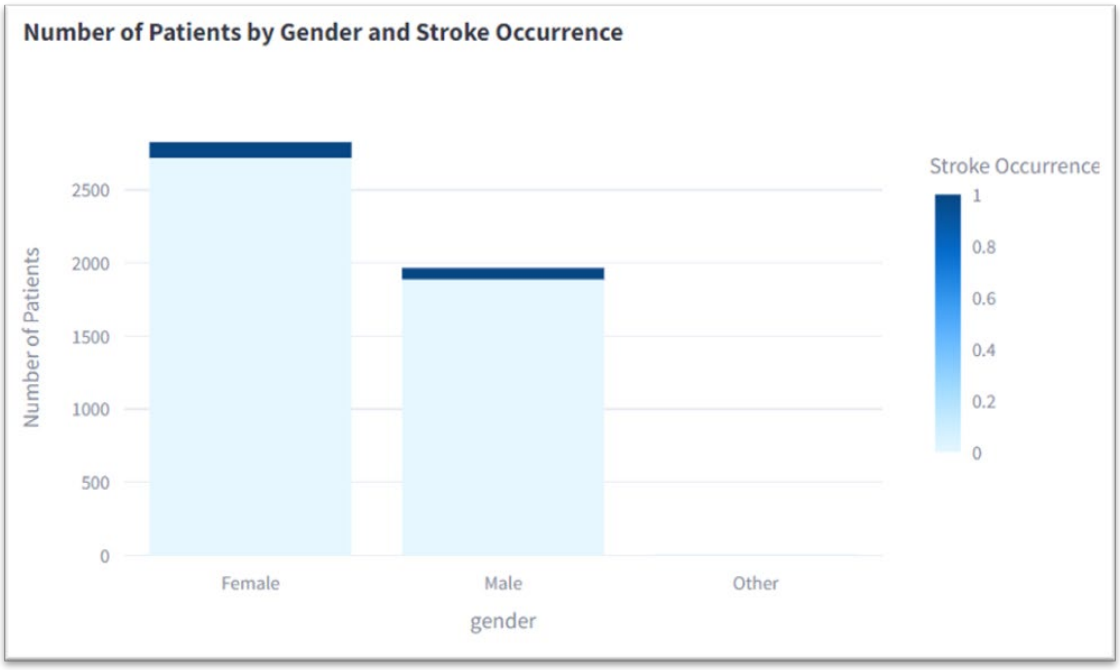
Breakdown by Gender:
Female: 59% of the total population analysed is represented by the greatest area of the chart, which is depicted in blue.

Male: Making up 41% of the total, the second-largest component is depicted in a darker shade of blue.

Other: A very small percentage of the population (0.0209%), shown by a tiny red portion, identifies as non-binary or as a gender other than male and female.

The important takeaways from the gender analysis is that the majority of the population is more than half, at 59% are females. Though still less than the female population, men also make up a sizable part of the population at 41%. There is a small representation of gender diversity, but it is

almost negligible at 0.0209%. The analysis highlights a gender imbalance, with more females in the group studied and very little representation of non-binary or other gender identities.



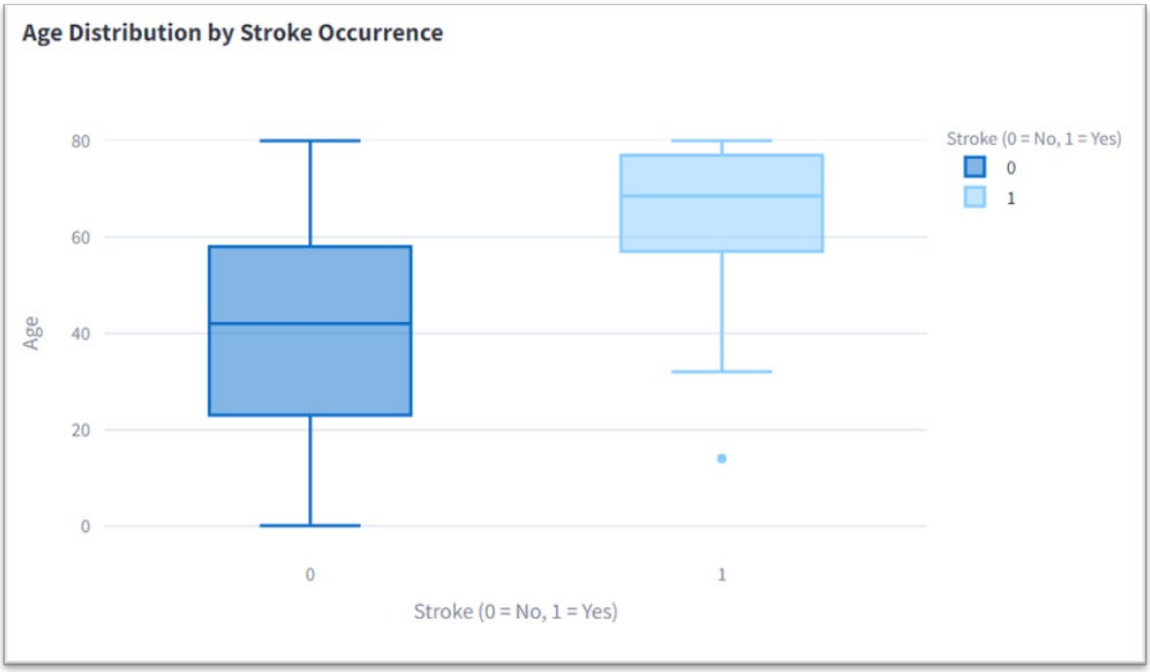
This graph uses a bar chart to show the number of patients by gender and stroke occurrence.

Important Findings in the data analysis for Female patients show that in total, there are more than 2500 female patients. This group's stroke occurrence score is somewhat below 1, indicating that female patients may have a comparatively high stroke incidence. While male patients are less than 2000 in total, which suggests that there are fewer male patients than female patients. The incidence of stroke in the male group is marginally lower than in the female's group, but it is still rather close to 1. Indicating a considerable incidence of stroke among male patients as well. In the other gender, the number of patients is minuscule or non-existent. The very small sample size is probably the reason why the stroke occurrence in the group is not well represented or noticeable in the chart.

In accordance with the gender distribution displayed in the preceding figure, there are more female patients than male patients in the data. Patients of both sexes experience a high incidence of stroke, with a modest increase in female patients. The stroke occurrence is shown on the graph by a darker

hue that is close to 1, which indicates that almost all of the group's members are suffering from strokes.

b. Age Analysis



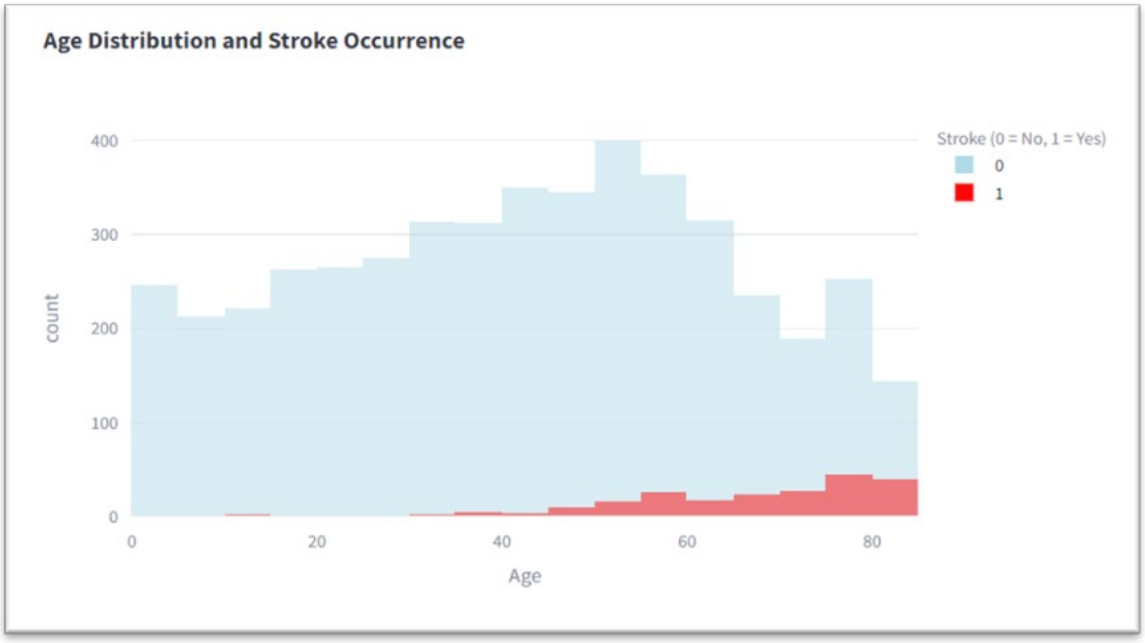
The plot displays two box plots next to each other. Patients without a stroke are represented by the left plot (Stroke = 0), and patients with a stroke are represented by the right plot (Stroke = 1).

Absent Stroke (0):
It shows that the median age of stroke-free individuals is between 40 and 45 years old. Within the interquartile range (IQR), which roughly corresponds to the ages of 25 to 60 years, the centre 50% of patients fall. The age spread is shown by the whiskers, which reach down to around 20 years and up to about 80 years.

Regarding Stroke (1):
Patients who experience a stroke typically have a much higher median age of approximately 70 years. Stroke patients have a narrower IQR, roughly falling between 60 and 80 years old, indicating

that older people account for the majority of stroke cases. While the upper whisker extends into the 80s, the bottom whisker drops to around 40 years old, indicating that younger stroke patients do exist, albeit they are less common.

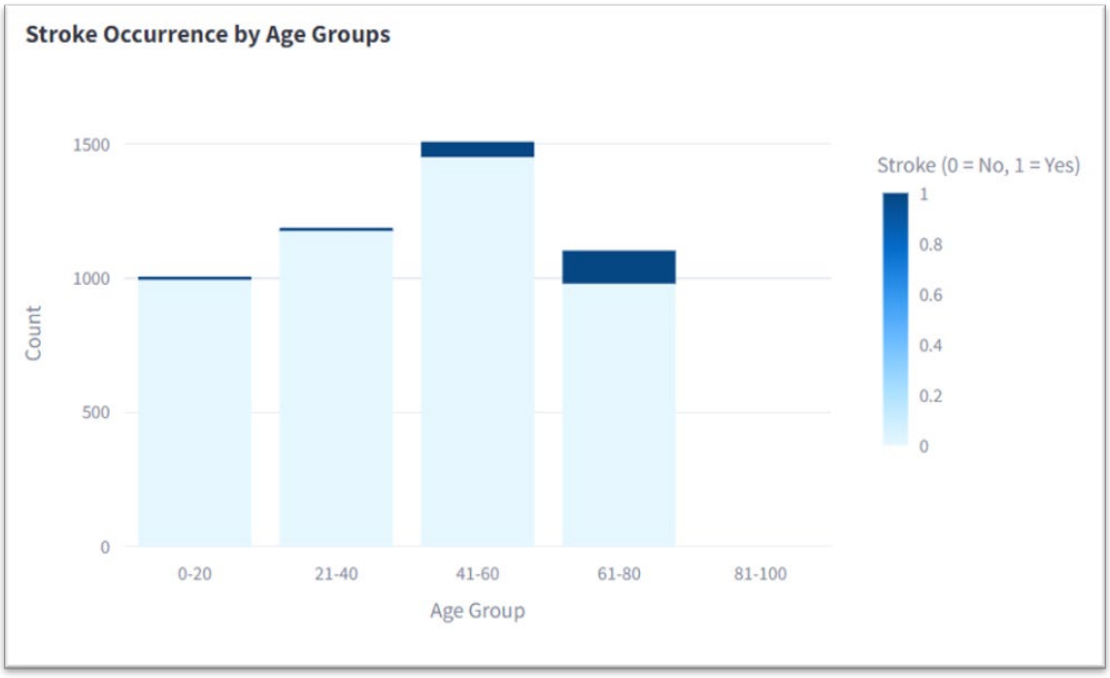
From the above analysis, it can be concluded that stroke patients are typically substantially older, with a median age of about 70. Patients without a stroke are often younger, in the range of 40-45. This suggests that the risk of stroke is strongly correlated with advancing age.



The complete population's age distribution is displayed on the graph, with the red areas denoting stroke instances (Stroke = 1) and the blue bars reflecting no stroke (Stroke = 0). With peaks between the ages of 40-60, the bulk of the population is concentrated between 20 and 70. Stroke incidence (red) begins to show up in tiny numbers at 40 years of age and rises in frequency with age, peaking in people 60 years of age and older.

In summary, stroke incidence is uncommon in younger people and begins to climb around age 40, peaking at age 60. The results from the box plot are supported by the histogram, which shows that strokes grow more common in elderly populations. The two graphs show that the likelihood of having a stroke increases with age. The majority of stroke victims are over 60, with younger people

less likely to be affected. Those who have had a stroke typically have a much greater median age than people who have not.



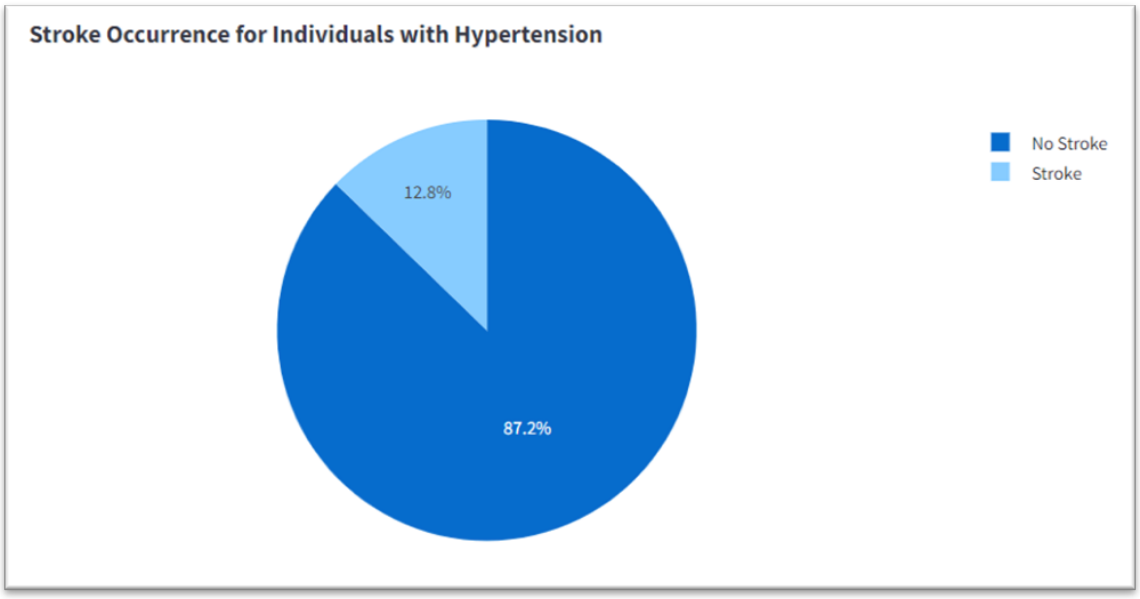
The graph shows the distribution of stroke occurrences across different age groups.

Count of Individuals: Each bar's height represents the number of individuals in that specific age group. For example, the 41-60 age group has the highest count, indicating a larger population in this category.

Stroke Occurrence (Color Gradient): The colour gradient indicates the proportion of stroke occurrences within each age group. Darker shades of blue correspond to a higher occurrence of strokes. In the 41-60 age group, there is a notable dark blue portion at the top, indicating a higher number of stroke cases. The age groups 61-80 and 21-40 also show some presence of stroke occurrences, as indicated by the dark blue sections at the top of their bars. The age group 0-20 and 81-100 have a minimal presence of dark blue, suggesting fewer stroke occurrences.

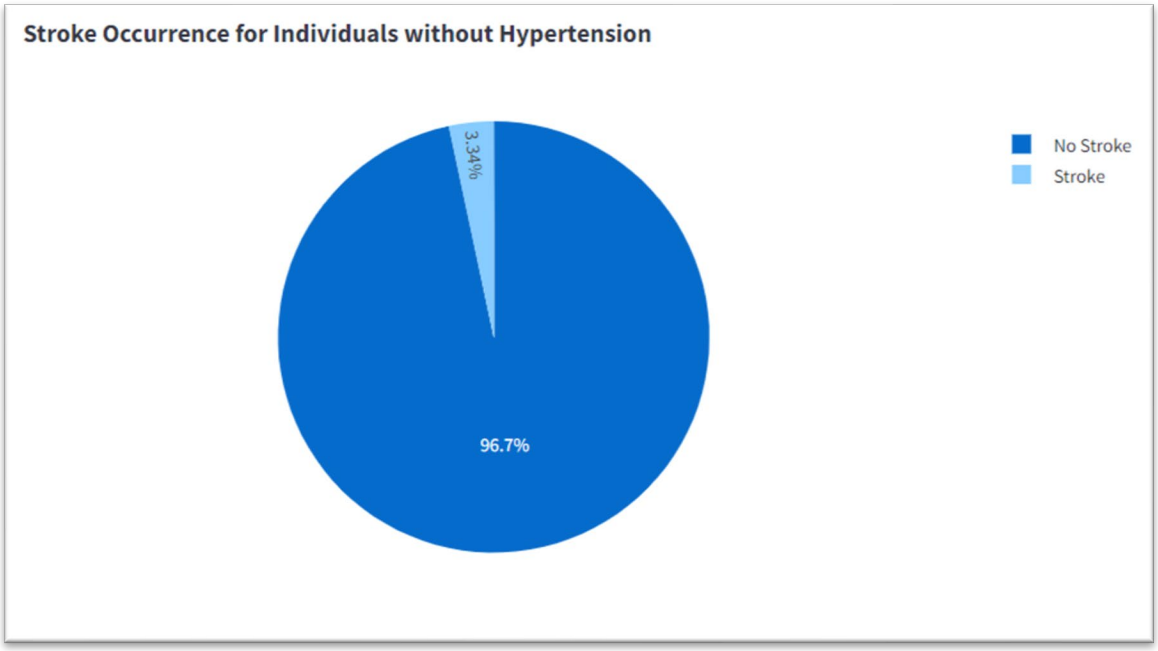
The graph therefore reveals that stroke occurrences are most prevalent in the 41-60 age group, as indicated by the darker blue shade at the top of the bar. This information can be valuable in targeting age groups for stroke prevention and healthcare interventions.

c. Health Factors Analysis

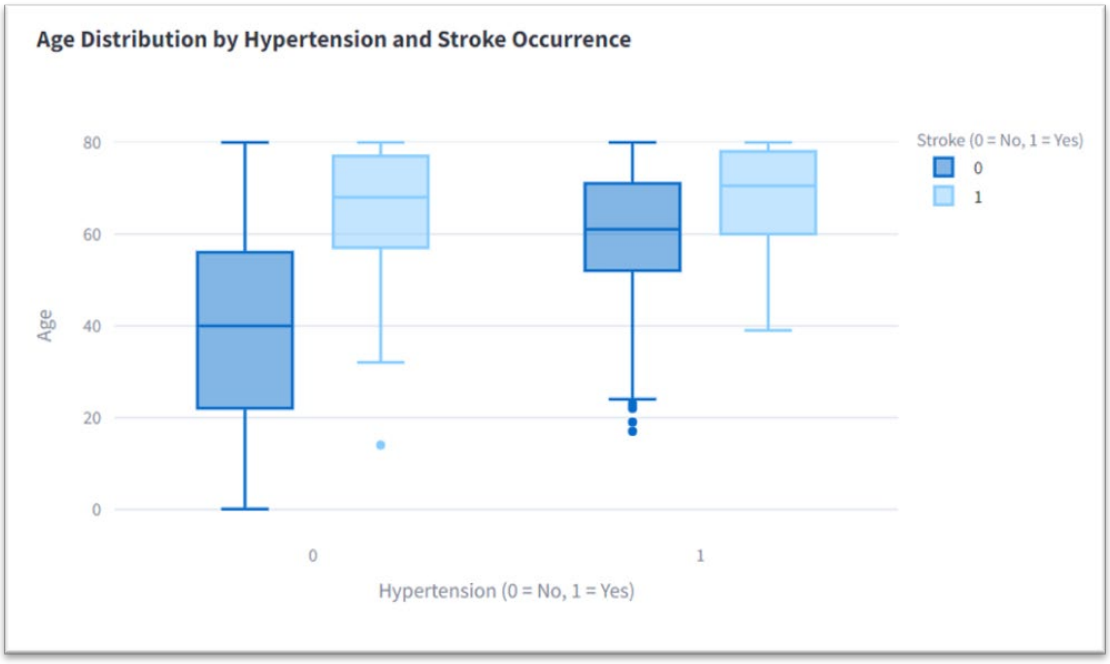


The chart shows that 87.2% of individuals with hypertension do not experience a stroke, while 12.8% of those with hypertension have suffered a stroke.

This suggests that having hypertension increases the risk of stroke. While the majority of individuals with hypertension do not experience a stroke, the 12.8% who do represent a significant proportion compared to individuals without hypertension



The chart shows that 97% of individuals without hypertension do not suffer from strokes, while only 3% have had a stroke. This indicates that individuals without hypertension are far less likely to experience a stroke compared to those with hypertension.



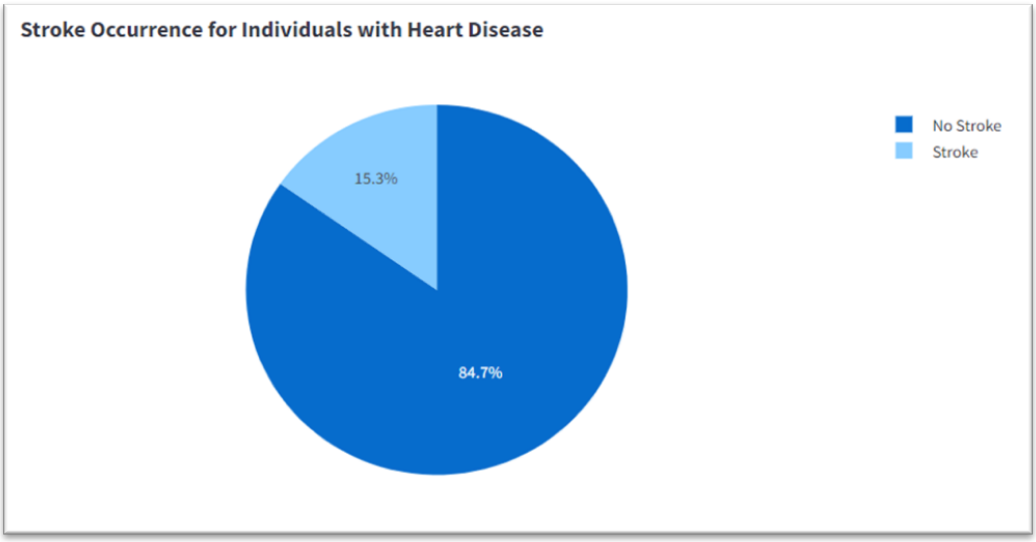
Individuals without Hypertension (0):

- The age distribution for individuals who have not had a stroke skews slightly younger, with the middle 50% of ages ranging between approximately 30 and 60 years.
- Those who have experienced a stroke in this group tend to be older, with a median age around 70 years, suggesting that stroke occurrences in this group are more common in older individuals.

Individuals with Hypertension (1):

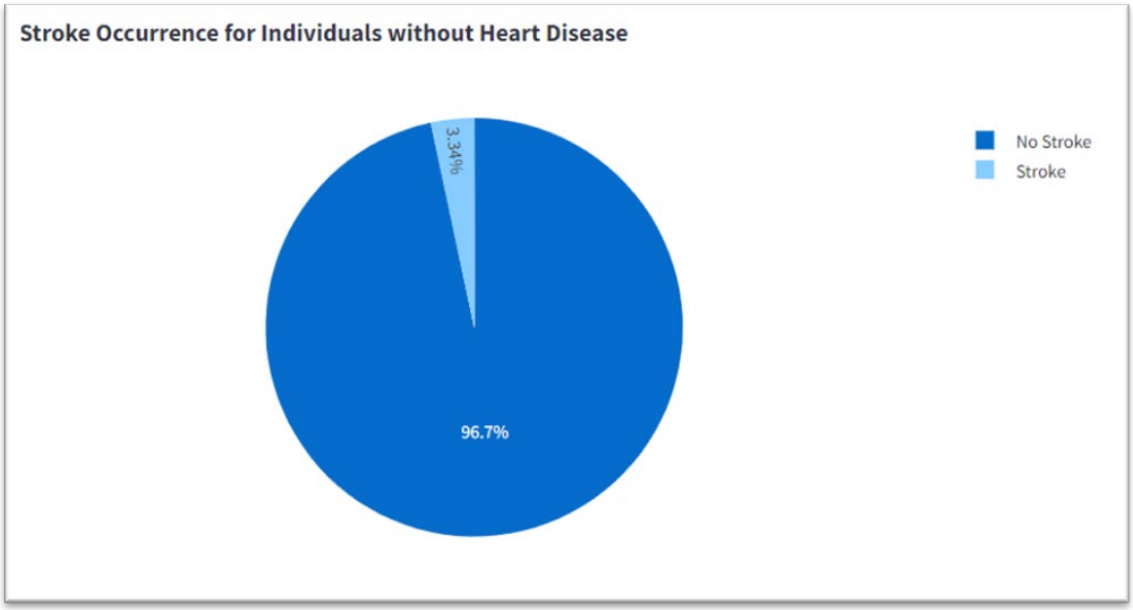
- The age distribution for individuals with hypertension who have not experienced a stroke is generally older, with a median age around 55 years. This implies that hypertension tends to affect individuals at later stages in life.
- Those with hypertension who have experienced a stroke tend to be older as well, with a median age close to 70 years, but their age distribution is tighter, meaning stroke occurrence is concentrated among older individuals with hypertension.

We can conclude that individuals with hypertension, regardless of stroke history, tend to be older, reinforcing the idea that age plays a significant role in developing hypertension. The combination of older age and hypertension increases the likelihood of suffering a stroke, as seen in the narrower and higher distribution of ages among individuals who have had strokes.



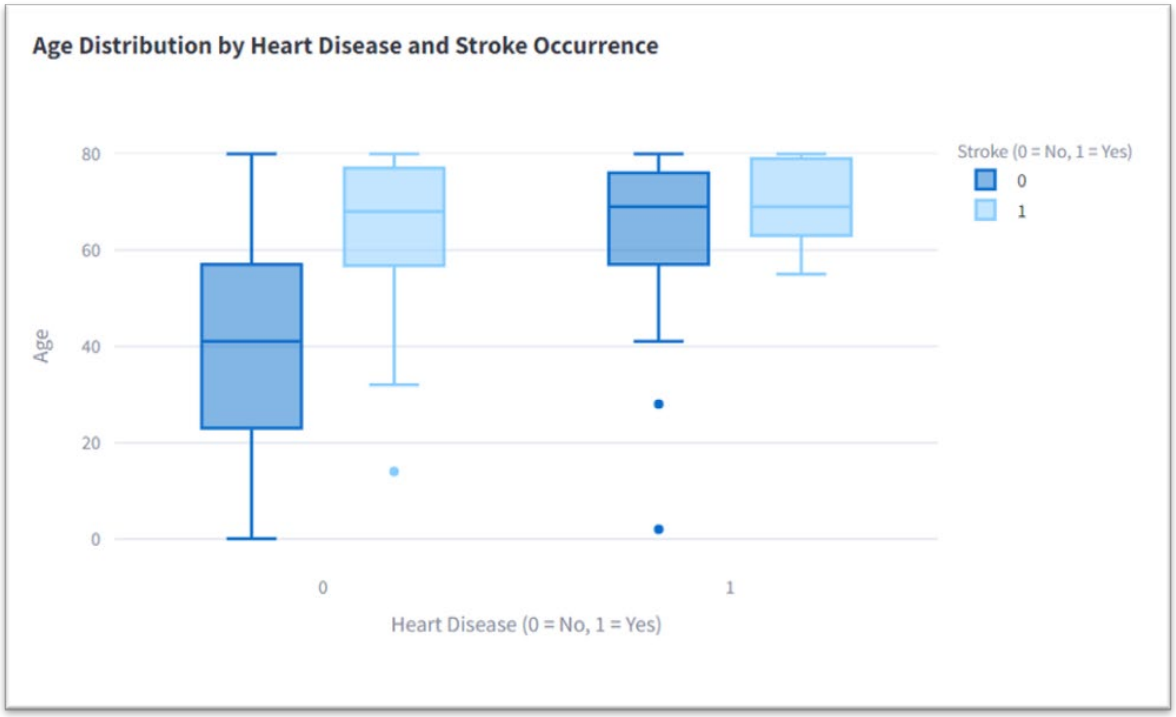
The chart above shows that 84.7% of individuals with heart disease have not experienced a stroke, while 15.3% have suffered a stroke.

This indicates that people with heart disease have a higher risk of stroke compared to those without heart disease. Although the majority of individuals with heart disease do not experience a stroke, the 15.3% who do represent a significant portion of the population. This suggests that heart disease increases vulnerability to stroke.



The chart shows that 96.7% of individuals without heart disease have not suffered a stroke, and only 3.3% have experienced a stroke.

This suggests that individuals without heart disease are significantly less likely to experience a stroke. The low percentage of stroke occurrences in this group shows that heart disease is a critical factor influencing stroke risk



Individuals without Heart Disease (0):

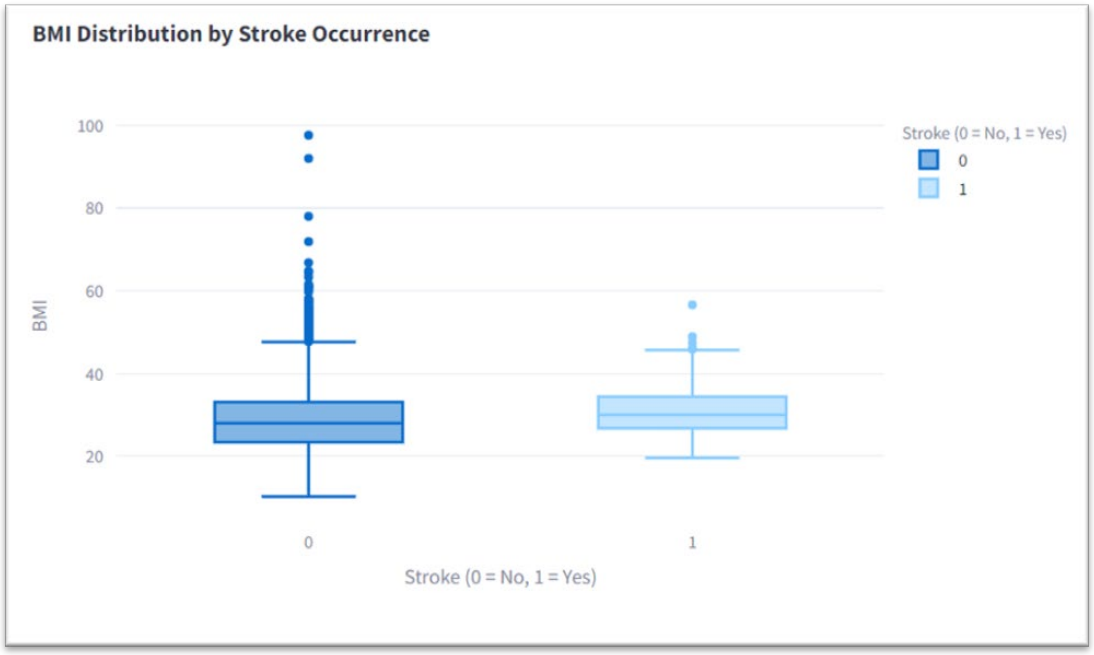
- The age distribution of individuals without heart disease who have not had a stroke is relatively wide, ranging from 20 to 80 years. The median age is around 45 years.
- For those without heart disease who have experienced a stroke, the age range is older, with the middle 50% falling between 55 and 75 years, and a median age around 70 years. This suggests that strokes occur more often in older individuals without heart disease.

Individuals with Heart Disease (1):

- In individuals with heart disease who have not had a stroke, the age distribution is slightly older, with a median age of 60 years, and a range of 40 to 80 years. This implies that heart disease is more prevalent in older individuals.
- Those with heart disease who have had a stroke tend to be older as well, with a tighter range between 60 and 75 years. The median age for this group is also around 65 to 70 years.

We can conclude that people without heart disease generally experience strokes at older ages, with a noticeable jump in risk for individuals over 55. For those with heart disease, the risk of stroke increases significantly as they age, especially after the age of 60.

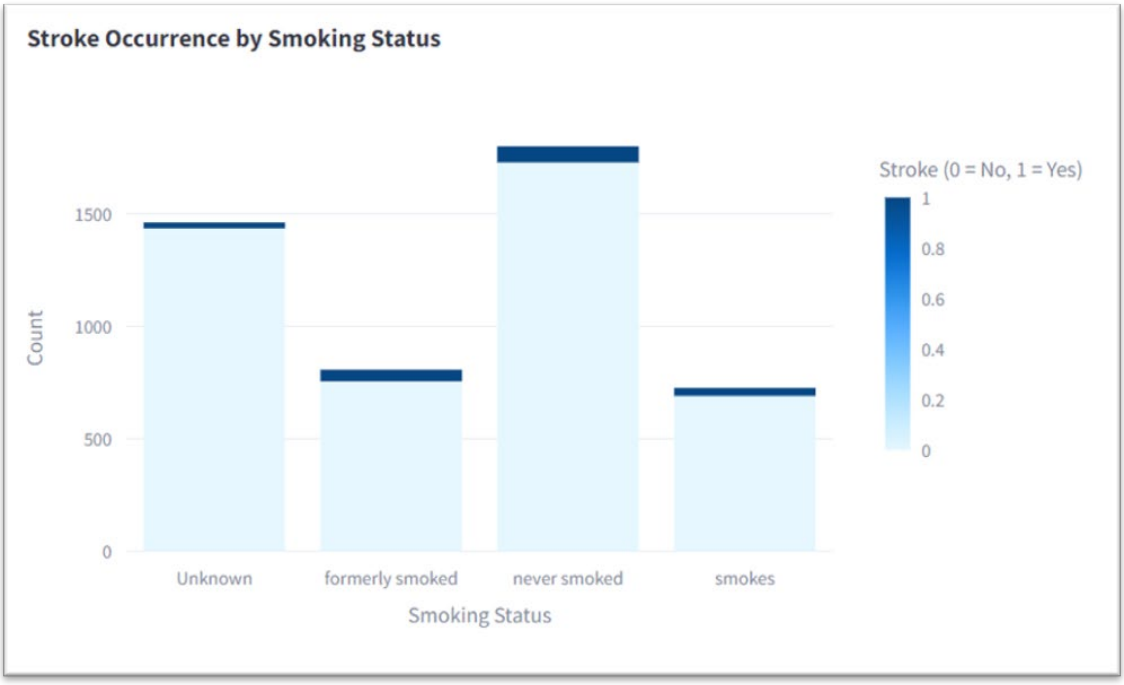
d. Lifestyle Factors Analysis



Non-Stroke Group (0): The BMI is distributed widely, with a few outliers above 50 BMI, suggesting a higher range of body weight variance among those who did not experience a stroke.

Stroke Group 1: The BMI has a smaller interquartile range, fewer outliers, and a less distributed distribution. The majority of the BMI readings in this cluster fall between 20 and 40.

Interpretation: Based on the boxplot, it appears that there is no significant direct correlation between BMI and the incidence of stroke. It's crucial to keep in mind, though, that BMI may indirectly raise the risk of stroke by affecting other conditions like diabetes or hypertension. There is no discernible change in BMI values between the stroke and non-stroke groups, despite the fact that the BMI distribution is marginally closer among stroke patients.

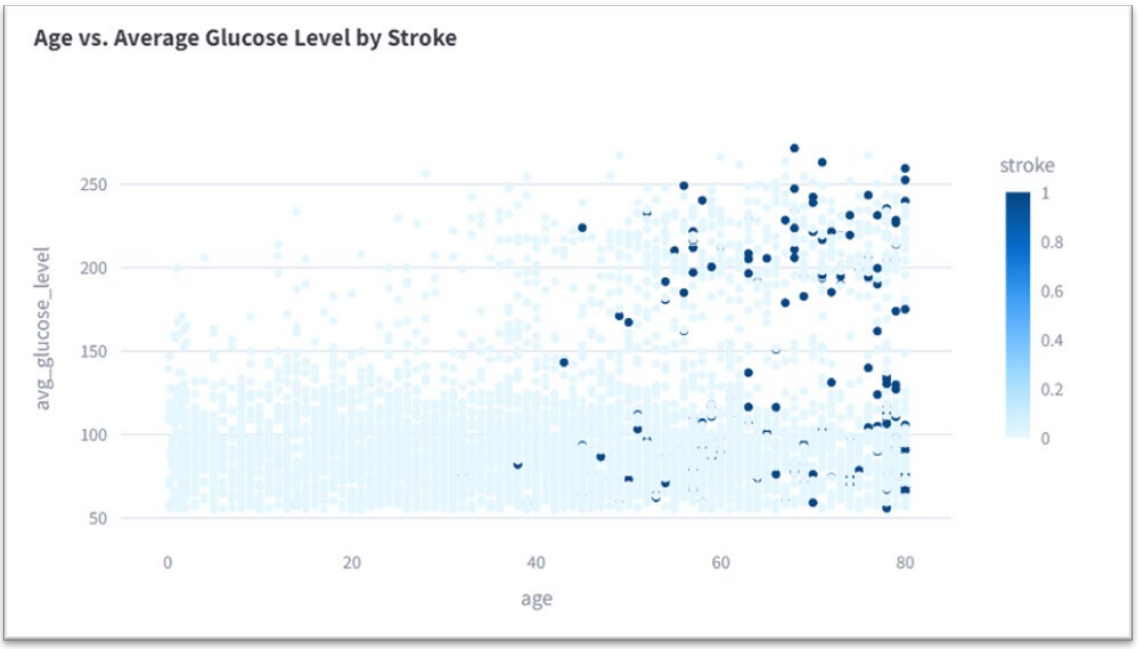


Unknown: Most of the population has an unknown smoking status, and just a small percentage of them have had strokes. Those who had previously smoked made up a smaller group than those who had "never smoked," but their percentage of stroke victims was significantly greater. The largest group, those who have never smoked, had a lower percentage of strokes than those who have smoked in the past.

Smokers: Compared to the "never smoked" group, the number of smokers is lower, but the risk of stroke is slightly greater.

Interpretation: According to the visualization, those who "smoke" now or "formerly smoked" are somewhat more likely to have a stroke than people who "never smoked." The difference isn't as obvious as one might predict, though. Although there is some uncertainty in the interpretation due to the size of the "unknown" category, smoking seems to be a risk factor for stroke overall, particularly for people who have a history of smoking.

e. Correlation Analysis

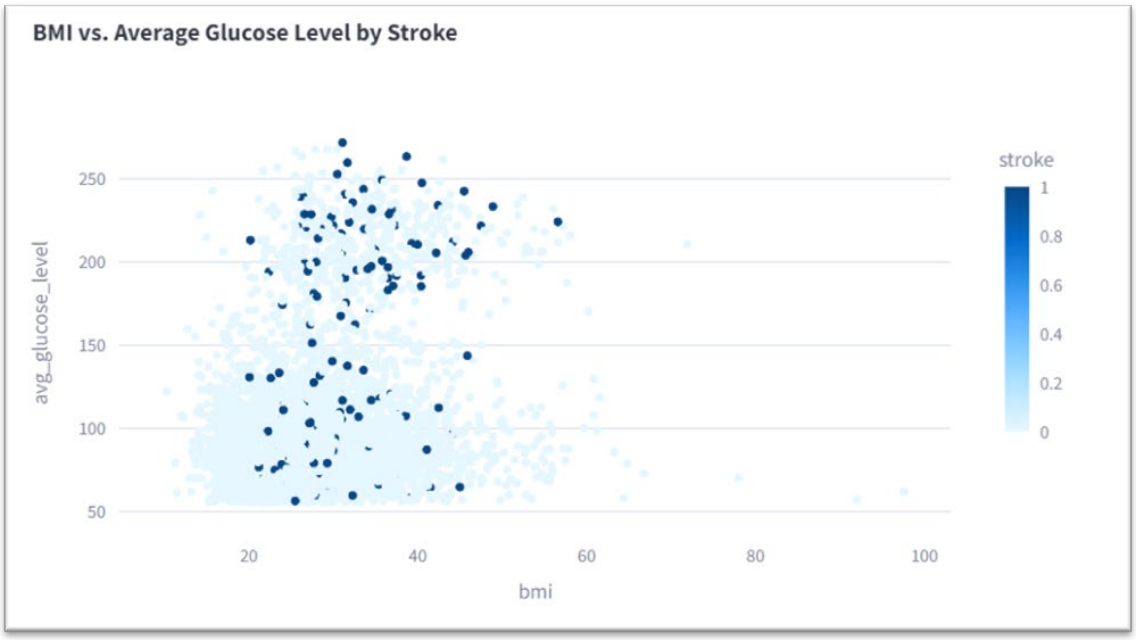


There is a moderate positive correlation between age and glucose level (0.2333), indicating that as age increases, the average glucose level tends to increase. This is expected because with aging, individuals may develop conditions like diabetes, which affects glucose regulation.

The correlation between age and stroke (0.2199). Age emerges as the most significant factor linked to stroke risk. The positive correlation suggests that as individuals grow older, their likelihood of experiencing a stroke increase. This aligns with established medical research, which consistently shows that stroke is more prevalent among older adults. Age-related physiological changes, such as arterial stiffening, reduced blood flow, and increased prevalence of chronic conditions (e.g., hypertension, heart disease), all contribute to this heightened risk.

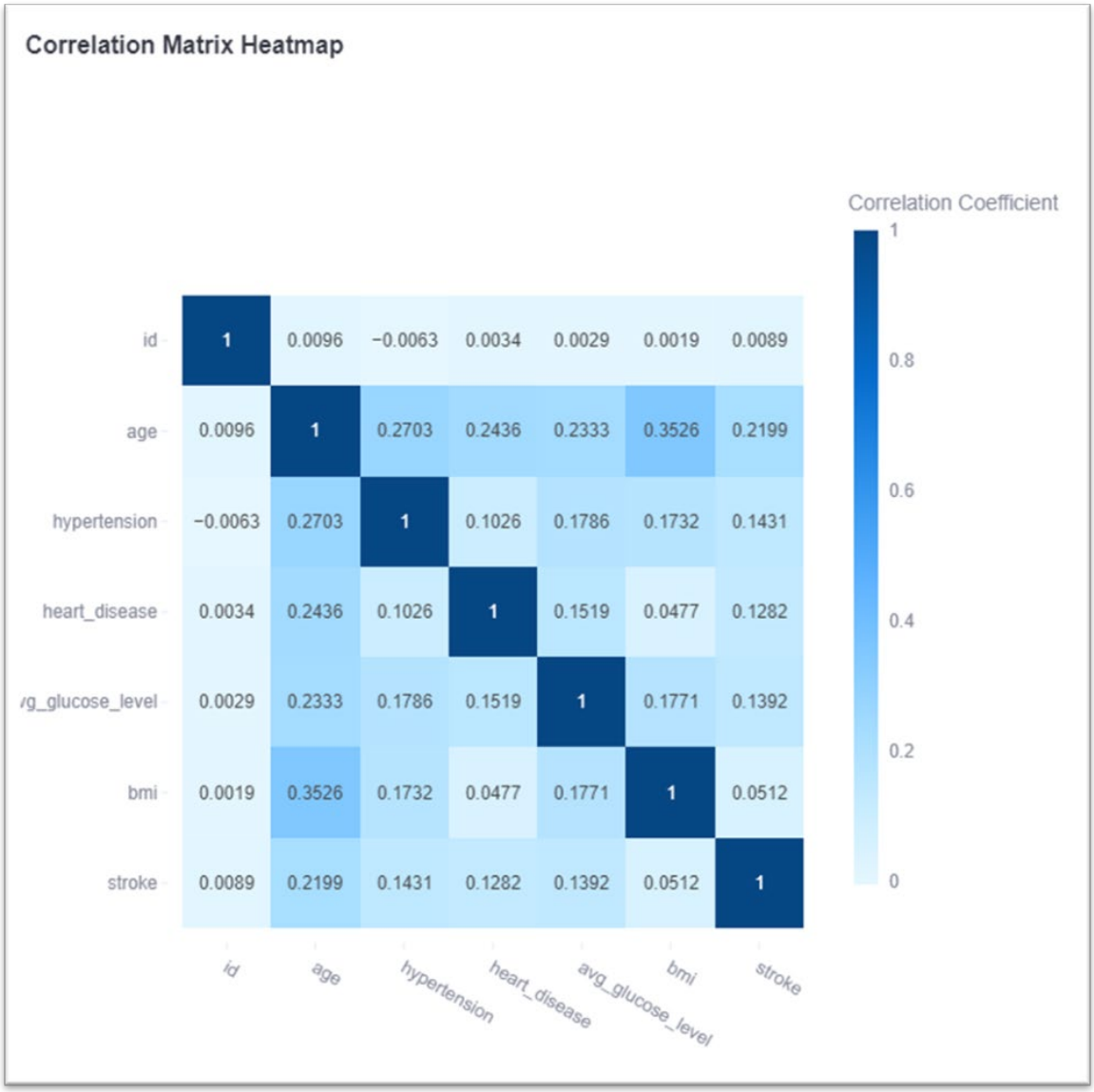
The correlation between average glucose level and stroke (0.1392), shows a weak but statistically significant correlation with stroke occurrences. Elevated glucose levels, which are commonly associated with conditions like diabetes, seem to marginally increase the risk of stroke. While the correlation is not as strong as that of age, it highlights a secondary risk factor. High glucose levels

could contribute to arterial damage over time, increasing the likelihood of stroke, especially in those with long-term or poorly managed diabetes.



The correlation between BMI and average glucose level is weak (0.1171), suggesting a minor positive relationship between higher BMI and glucose levels. Higher BMI is often associated with conditions like obesity and metabolic syndrome, which can affect glucose regulation.

The correlation between BMI and stroke (0.0512) is very weak, indicating that BMI alone may not be a strong predictor of stroke in this dataset. Other factors like age or lifestyle could play a more significant role.



This matrix suggests that age is the strongest predictor of stroke risk, followed by hypertension, based on the dataset. Average glucose level also has a weak positive correlation with stroke, but its impact is less pronounced compared to hypertension. BMI and other health factors such as heart disease show much weaker correlations with stroke. However, the combined effect of these variables should still be considered in a multivariate model for more accurate predictions, as interactions between them may influence overall stroke risk.

For individuals using the app to assess their stroke risk:

- Age is the dominant risk factor, so older users should take immediate action to manage stroke risk factors.
- Managing blood sugar levels is crucial, particularly for people with diabetes or prediabetes.
- While BMI does not strongly influence stroke risk in isolation, it may still contribute to overall health complications that can raise stroke risk.
- Users with hypertension or heart disease should not disregard these conditions, even though their direct correlation to stroke is weak in this dataset.

For healthcare providers or analysts using the app:

- Age should be given high importance when developing predictive models for stroke risk.
- A multifactorial approach is needed to assess stroke risk comprehensively, considering how different factors like age, glucose levels, and cardiovascular health interact with one another.

In summary, the correlation analysis equips app users with a clear understanding of which health factors to focus on for stroke prevention. Age, glucose levels, and managing hypertension and heart disease are key takeaways, guiding both individuals and healthcare providers toward more informed decision-making about stroke risk management.