Univariate Analysis

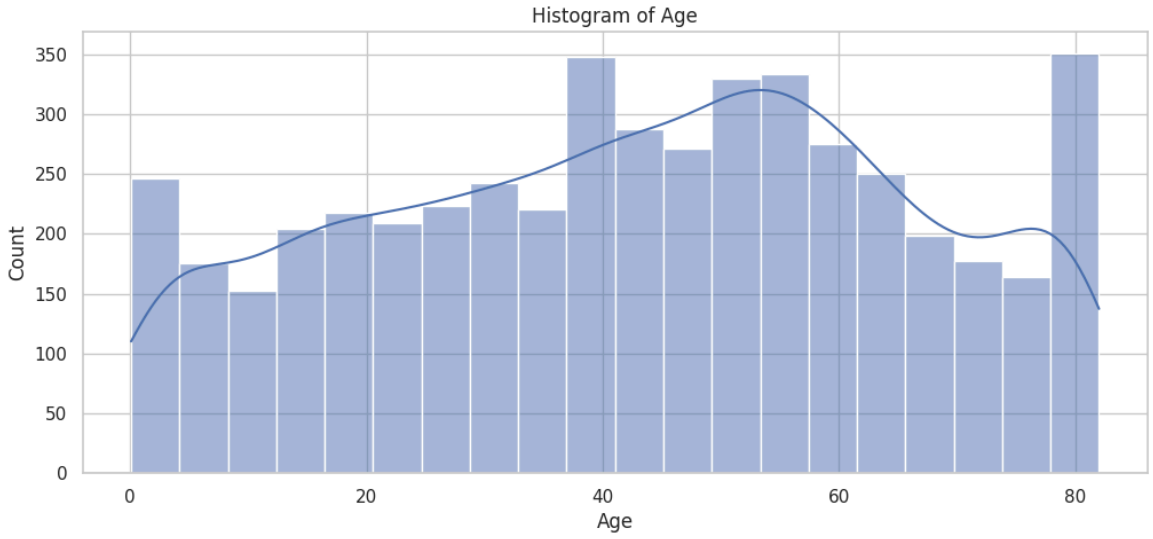
Numeric variables:

1) age:

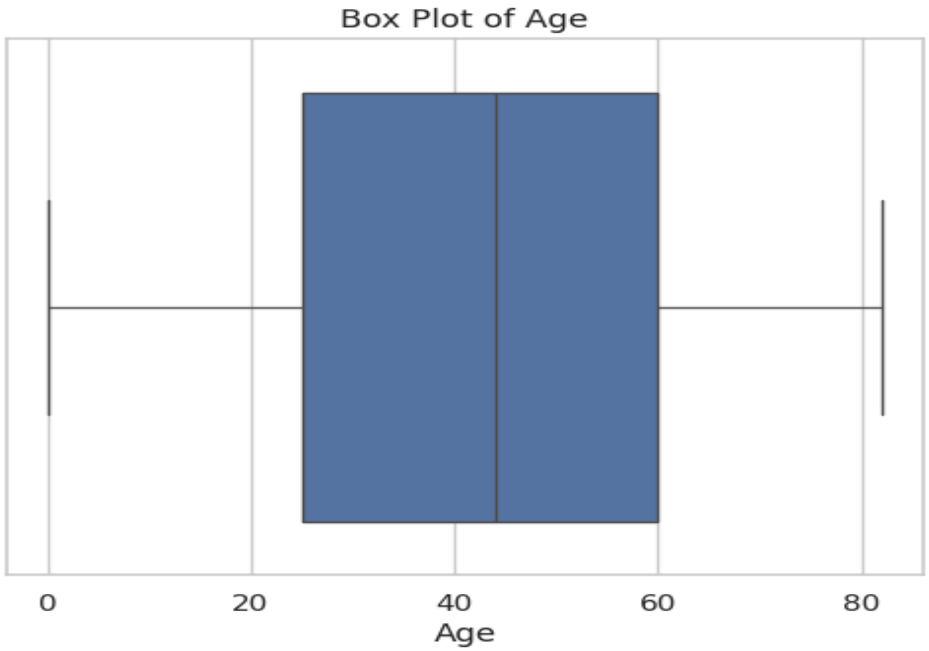
Summary statistics:

Summary statistics for age:
count 4873.000000
mean 42.868467
std 22.587424
min 0.080000
25% 25.000000
50% 44.000000
75% 60.000000
max 82.000000
Name: age, dtype: float64

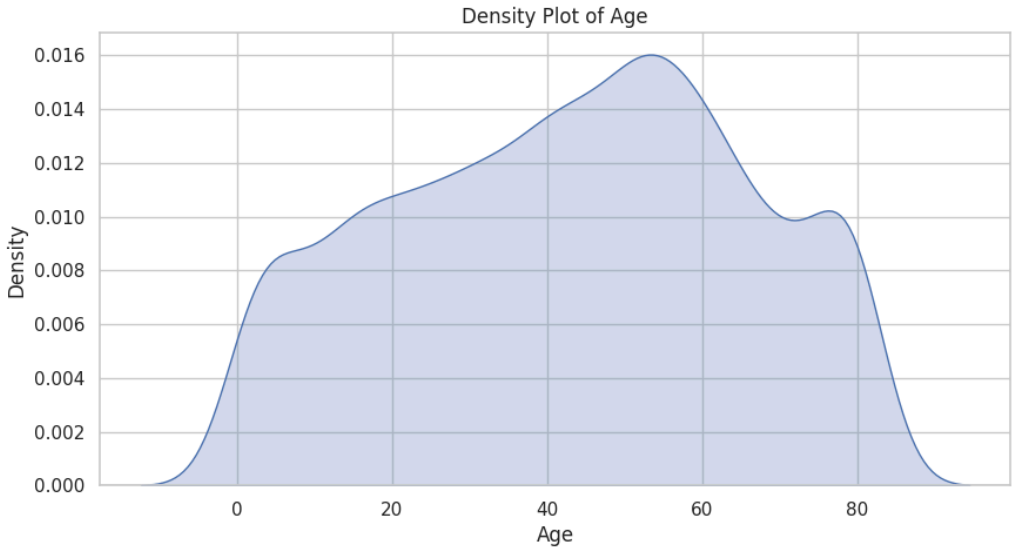
Histogram:



Boxplot:



Density plot:



Distribution: The age variable spans from 0.08 to 82 years, indicating a wide age range. The distribution skews very slightly to the left, as the mean (42.87) and median (44) are very close, showing a balanced spread with a slight leaning towards older ages. This can be seen in all three plots

Central Tendency: The mean age is 42.87, and the median is close at 44, highlighting a central concentration around middle age in this dataset.

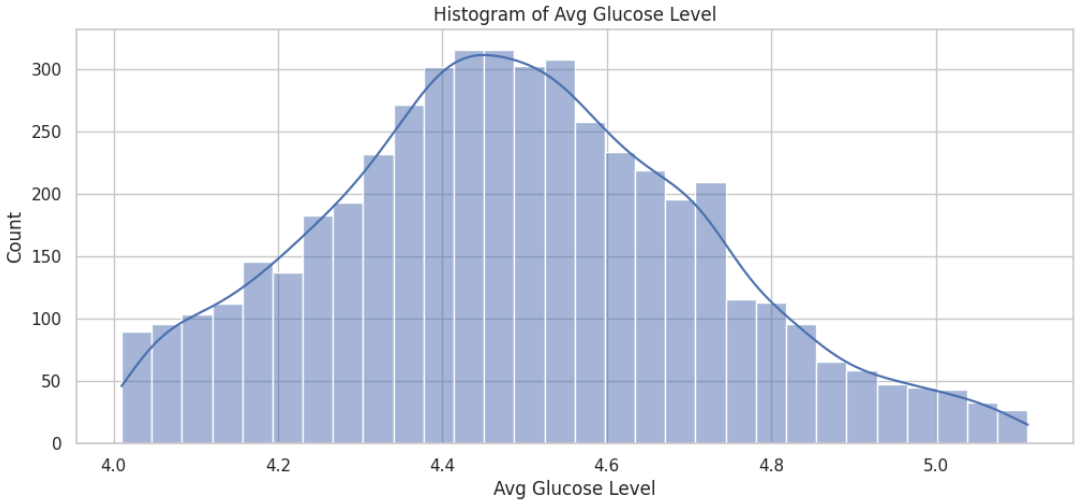
Variability: The standard deviation of 22.59 suggests high variability in the age group. With the 25th percentile at 25 years and the 75th at 60, most participants are between young adulthood and senior years. This variability can reveal patterns across age groups in relation to other health indicators.

2) avg_glucose_level:

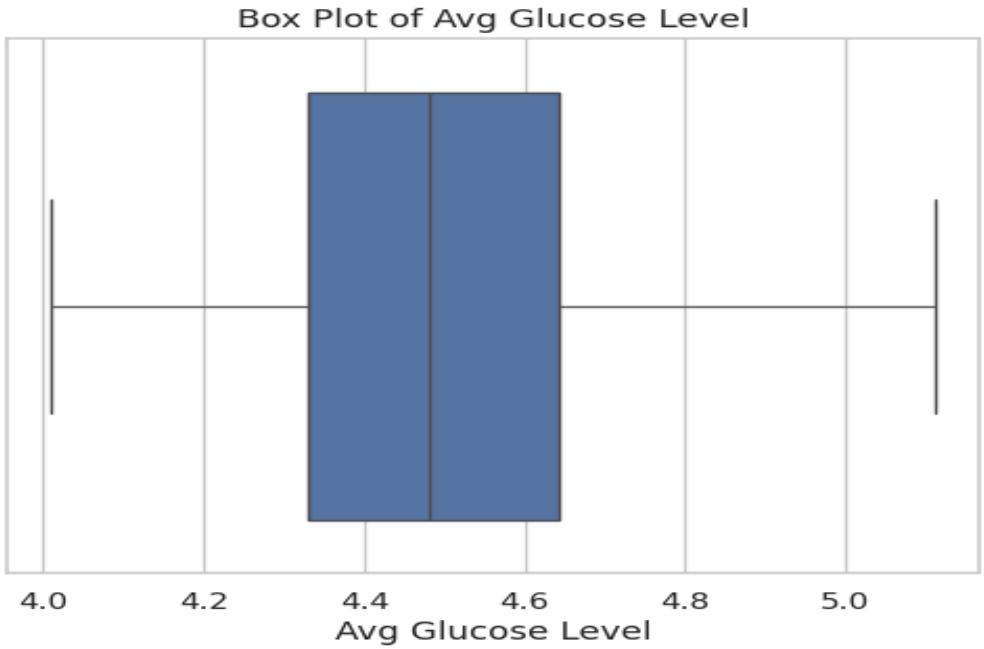
Summary statistics:

Summary statistics for avg_glucose_level:
count 4873.000000
mean 4.488129
std 0.231261
min 4.009513
25% 4.328758
50% 4.479947
75% 4.642273
max 5.111928
Name: avg_glucose_level, dtype: float64

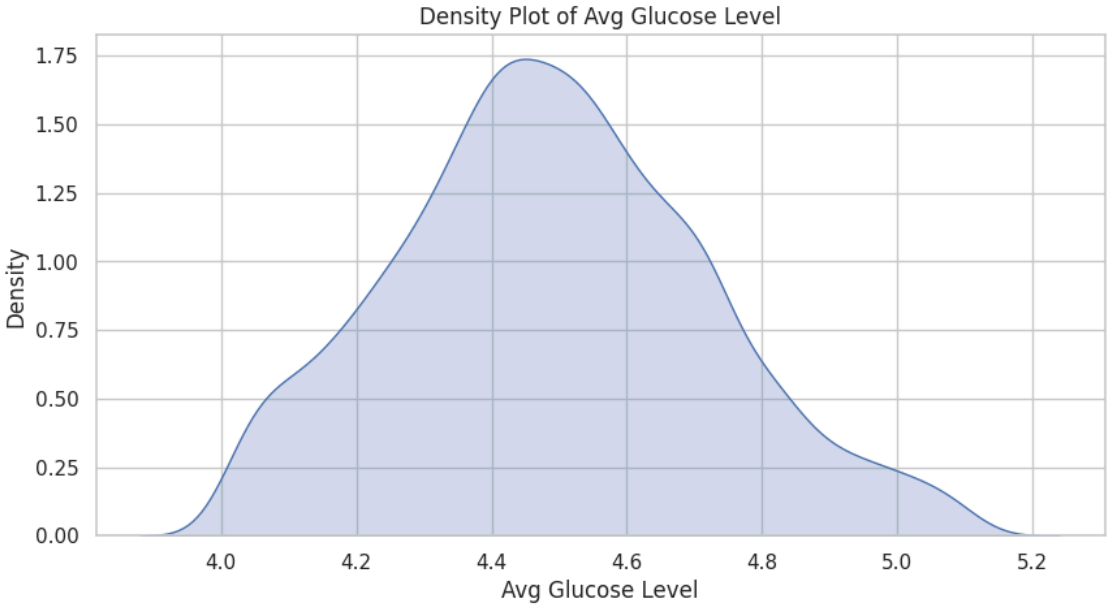
Histogram:



Boxplot:



Density plot:



Distribution: The average glucose level has a minimum of 4.01 and a maximum of 5.11, showing a narrow, condensed range. Given the mean (4.49) and median (4.48) are nearly identical, the distribution appears symmetrical.

Central Tendency: The mean of 4.49 aligns closely with the median of 4.48, suggesting the values are clustered closely around this central point, in a bell-shaped curve. It can be seen from the density plot and histogram.

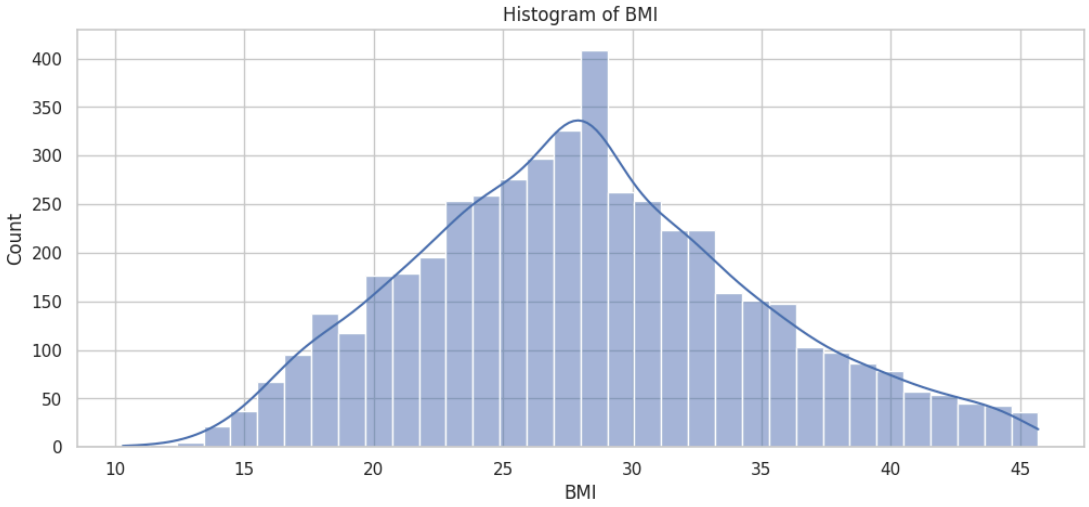
Variability: A low standard deviation of 0.23 indicates minimal spread among values, suggesting that glucose levels are quite uniform across the sample. This low variability helps identify any small deviations that may still be impactful when compared to health outcomes.

3) bmi:

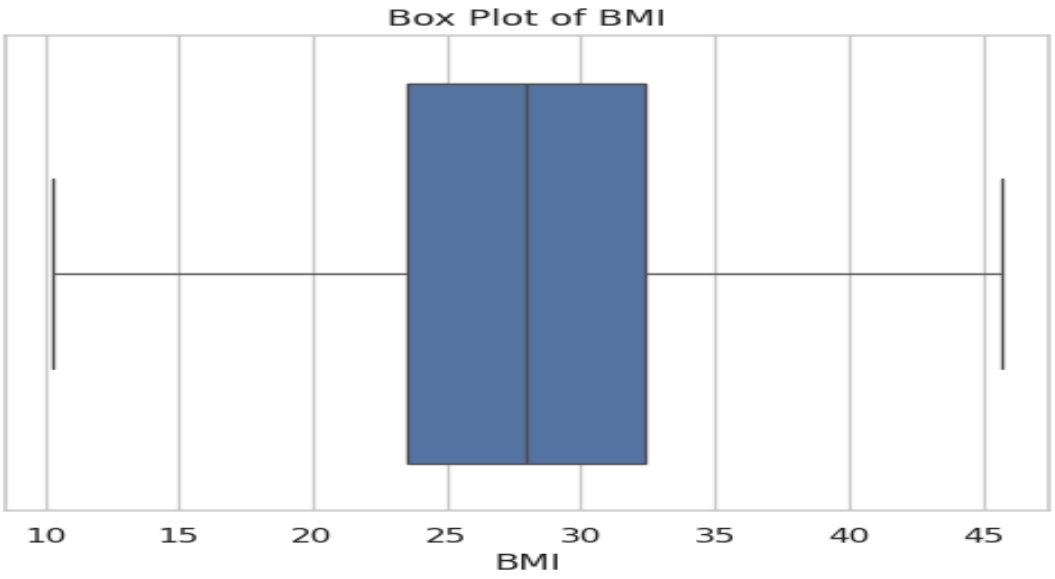
Summary statistics:

Summary statistics for bmi:
count 4873.000000
mean 28.217669
std 6.651141
min 10.300000
25% 23.500000
50% 28.000000
75% 32.400000
max 45.700000
Name: bmi, dtype: float64

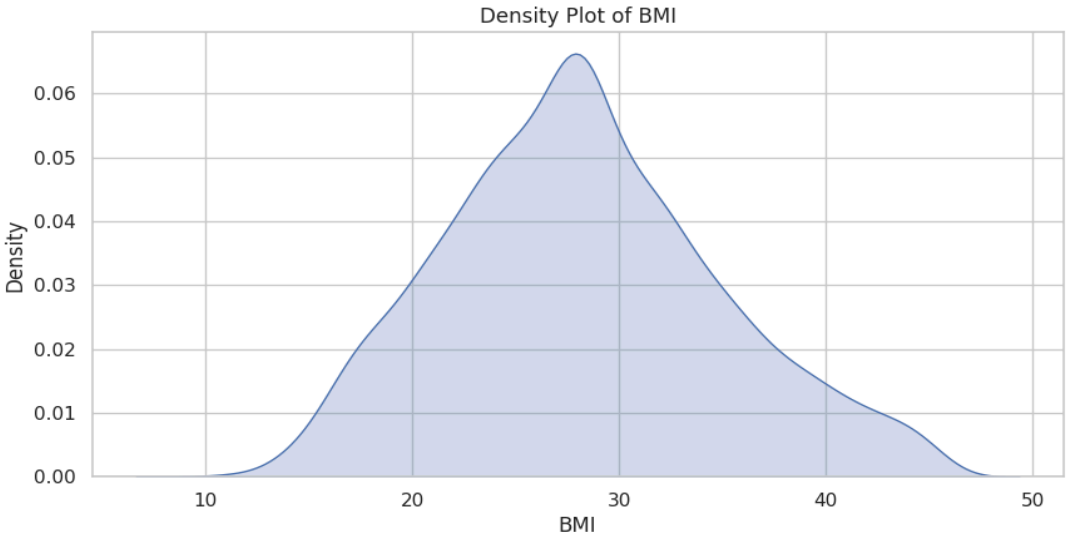
Histogram:



Boxplot:



Density plot:



Distribution: BMI values range widely from 10.3 to 45.7. The mean is 28.22, and the median is 28, indicating a roughly symmetric distribution.

Central Tendency: The average BMI of 28.22 places the central tendency in the overweight category, with a tight alignment between the mean and median.

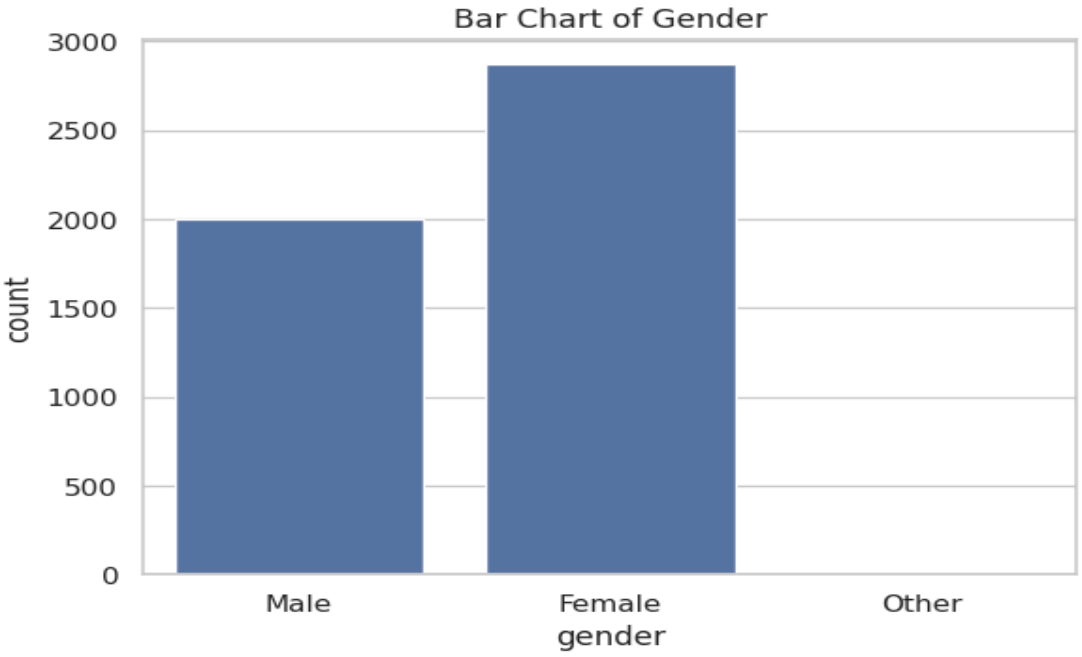
Variability: The standard deviation of 6.65 shows moderate variability in BMI. The values in the 25th percentile (23.5) to the 75th percentile (32.4) span normal to obese categories. This variable provides insight into the overall body composition trends and potential health risks within the group.

Categorical variables :

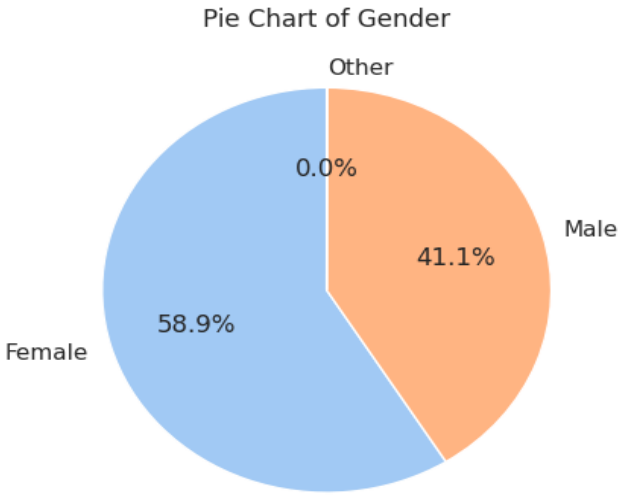
1) gender:

Word cloud:

Bar chart:



Pie chart:

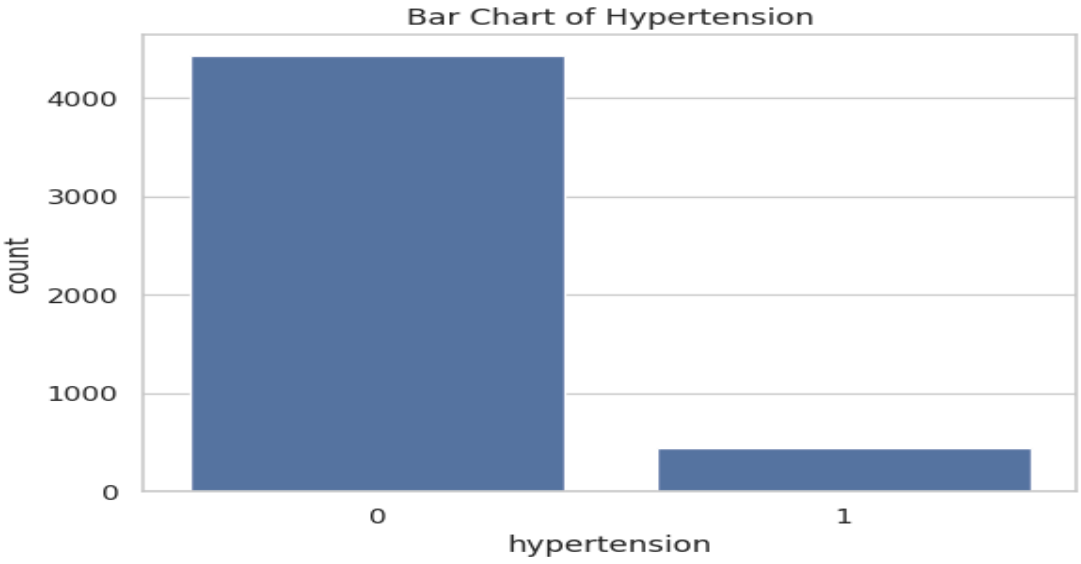


Distribution: In the **gender** variable, females make up 58.9% of the dataset, and males represent 41.1%. This imbalance shows a higher representation of females, which may impact analysis involving gender comparisons as it skews toward female health data.

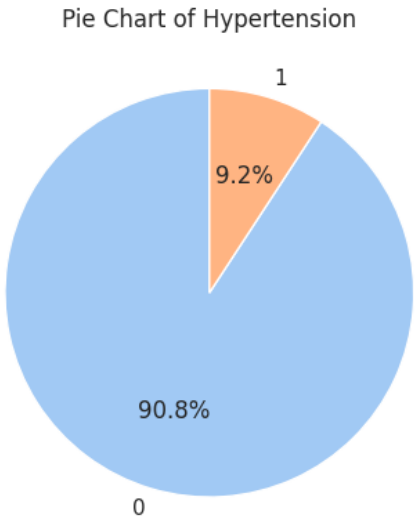
Central Tendency and Variability: Here, the mode is "female." The variability is low, but the 17.8% difference indicates a meaningful skew toward female representation. This distribution could influence findings on gender-specific health outcomes in the analysis.

2) hypertension:

Bar chart:



Pie chart:

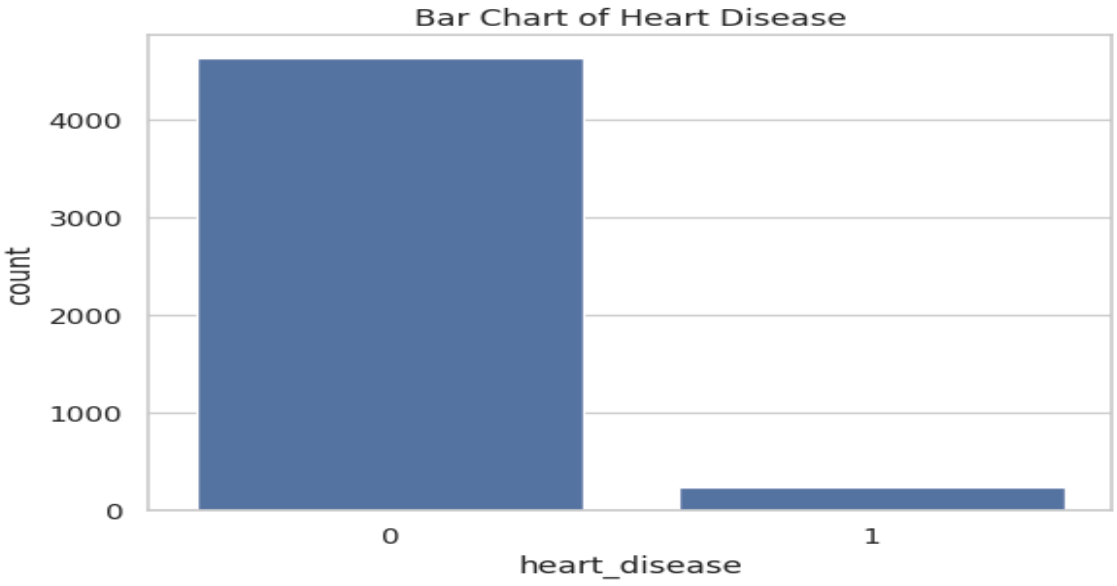


Distribution: For **hypertension**, 90.8% of individuals do not have hypertension, while 9.2% do. This distribution suggests that hypertension is relatively rare in this sample, which may limit the statistical power of hypertension-related insights.

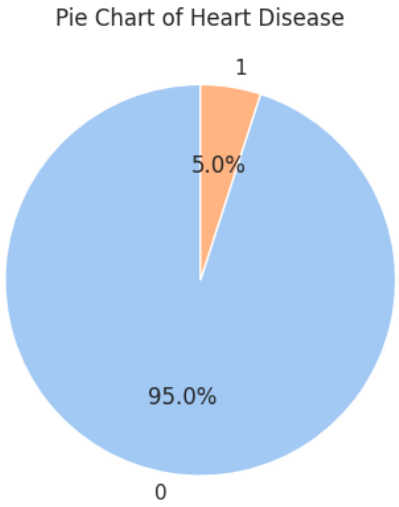
Central Tendency and Variability: The mode is "No" for hypertension, indicating that most participants do not have the condition. Variability is low, as the dataset is heavily skewed toward individuals without hypertension, which restricts detailed comparison between those with and without the condition in relation to stroke risk.

3) heart_disease:

Bar chart:



Pie chart:

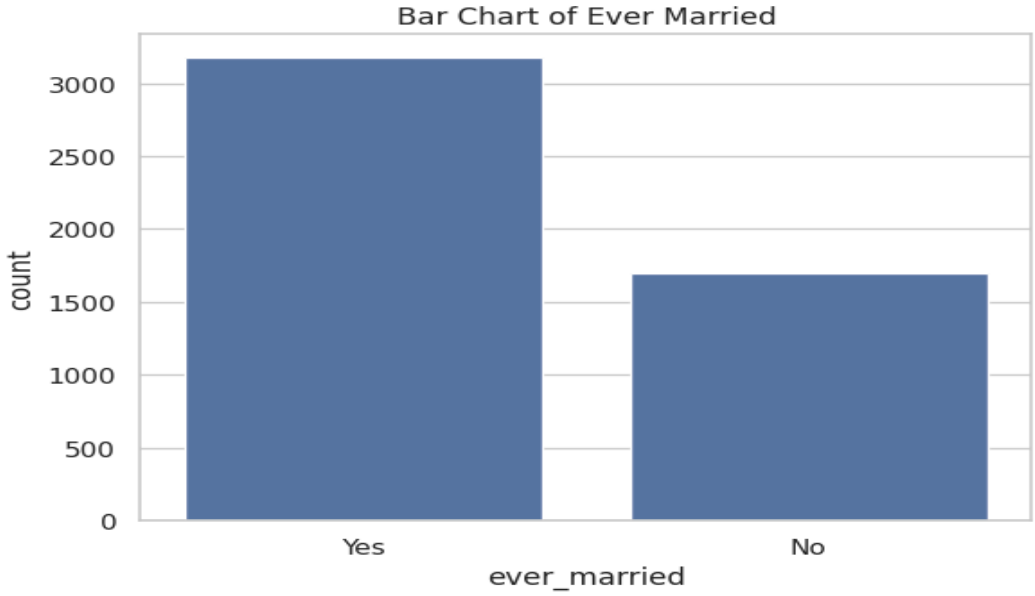


Distribution: In the **heart_disease** variable, 95% of participants do not have heart disease, while only 5% do. This low prevalence of heart disease suggests that it is an uncommon condition within the dataset.

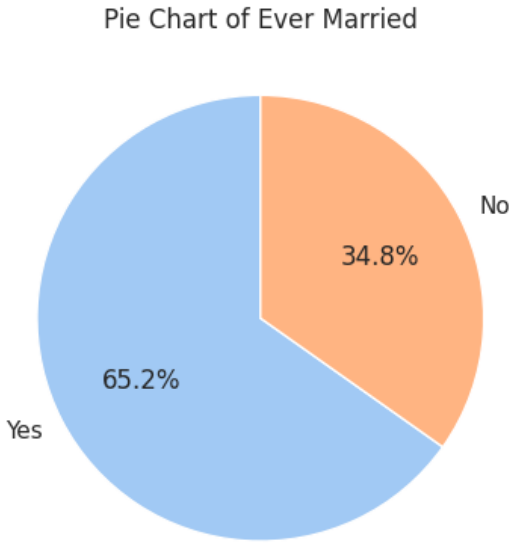
Central Tendency and Variability: The mode is "No" for heart disease, showing that the condition is uncommon in this population. The variability is very low, with a strong skew toward individuals without heart disease, which limits the dataset's ability to explore heart disease as a potential risk factor for stroke comprehensively.

4) ever_married:

Bar chart:



Pie chart:

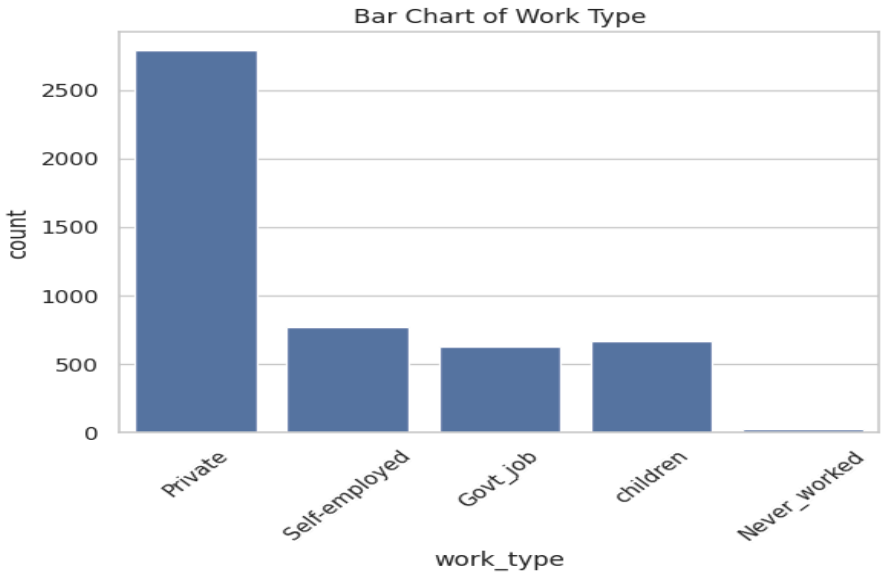


Distribution: For the **ever_married** variable**,** 65.2% of participants have been married, while 34.8% have not. This distribution indicates that a majority of participants have been married, which might correlate with other demographic variables such as age.

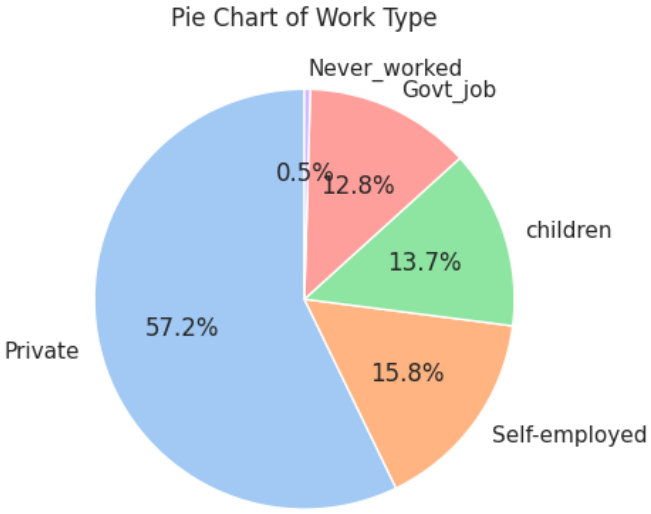
Central Tendency and Variability: The mode is "Yes" for ever married, as it is the most common category. There is moderate variability, with a significant proportion of individuals in each category, allowing for comparisons between marital status and health outcomes, such as any potential impact on stroke risk.

5) work_type:

Bar chart:



Pie chart:

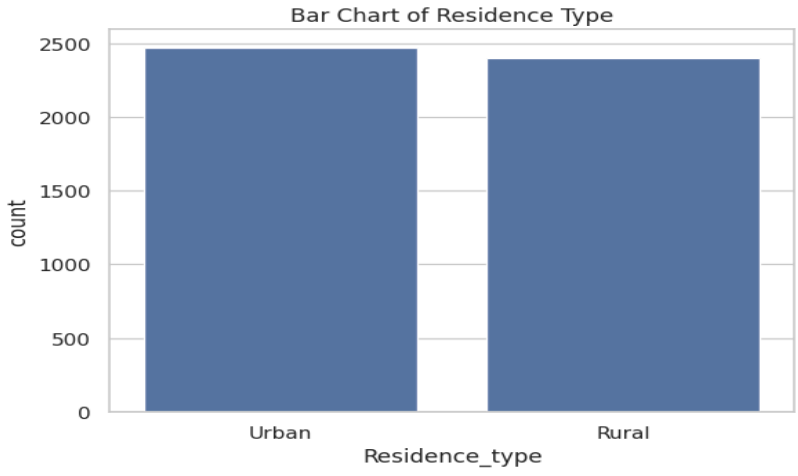


Distribution: The **work_type** variable is divided into multiple categories: 57.2% work in private employment, 15.8% are self-employed, 13.7% are children (thus not employed), 12.8% work in government jobs, and 0.5% are in never worked. This indicates that private employment is the most common work type, with notable diversity across categories.

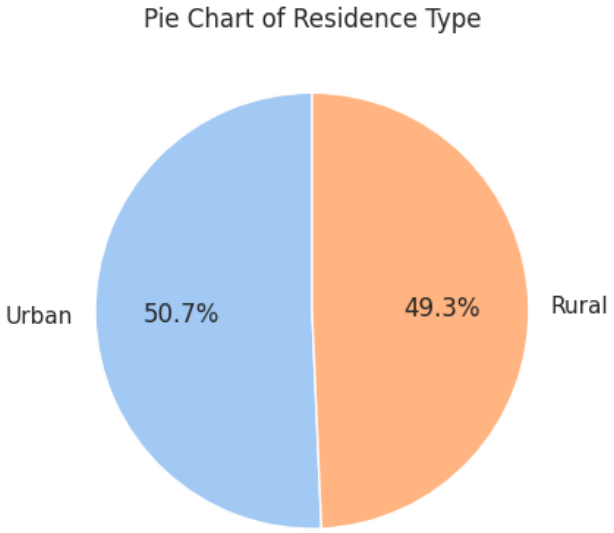
Central Tendency and Variability: The mode is "Private" work, as over half of the participants are employed in this sector. There is high variability, with substantial representation across different work types, enabling in-depth analysis of how various employment types might correlate with health and lifestyle factors.

6) Residence_type:

Bar chart:



Pie chart:

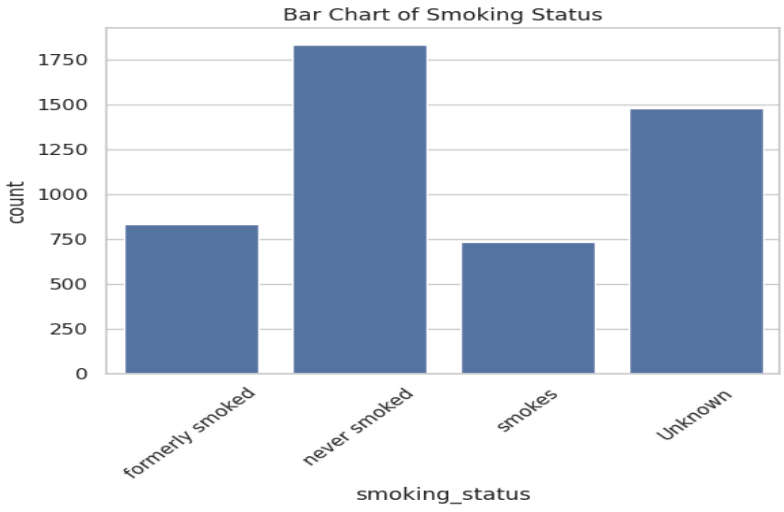


Distribution: In **Residence_type**, 50.7% of participants are urban residents, and 49.3% are rural residents. This nearly equal distribution offers a balanced view of urban and rural settings, which can yield meaningful comparisons on how living environment impacts health.

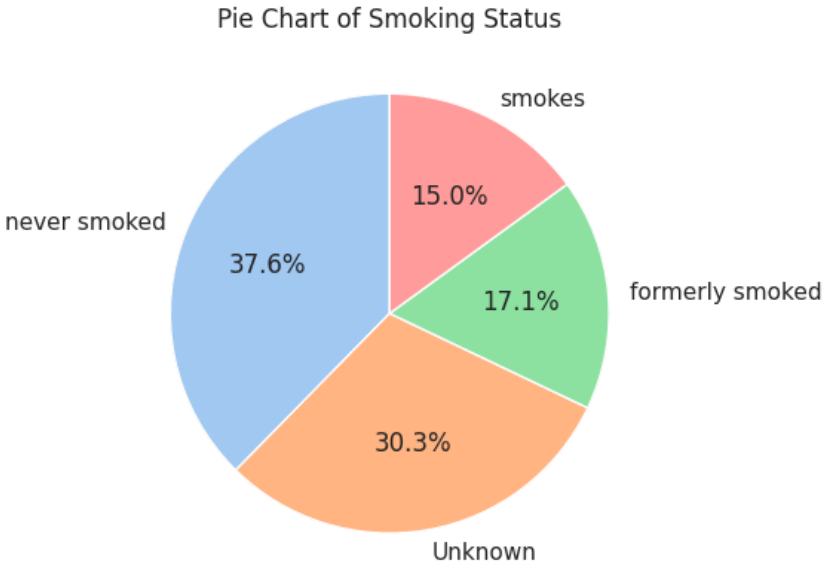
Central Tendency and Variability: With a nearly equal split, there is no dominant mode, and the variability is high. This balance allows for comprehensive urban-rural analyses, offering valuable insights into whether stroke risk or health factors differ significantly between residence types.

7) smoking_status:

Bar chart:



Pie chart:



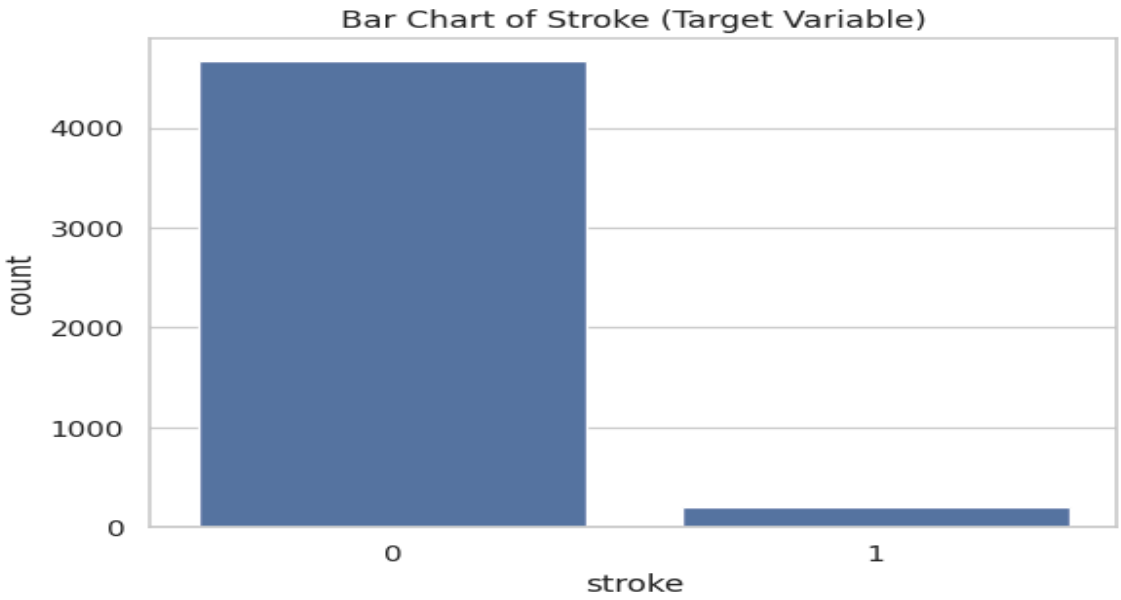
Distribution: **Smoking_status** is split into 37.6% who have never smoked, 30.3% unknown, 15% who currently smoke, and 17.1 formerly smoked. This breakdown shows a significant portion of participants with some smoking history (either former or current).

Central Tendency and Variability: The mode is "Never smoked," but the distribution is spread across categories, indicating moderate variability. The representation of former and current smokers allows for detailed comparisons, potentially revealing the impact of smoking on stroke risk and health outcomes. The large proportion of unknown limits the accuracy and completeness of any analyses on smoking's impact, potentially skewing results and weakening conclusions about smoking-related risk factors for stroke.

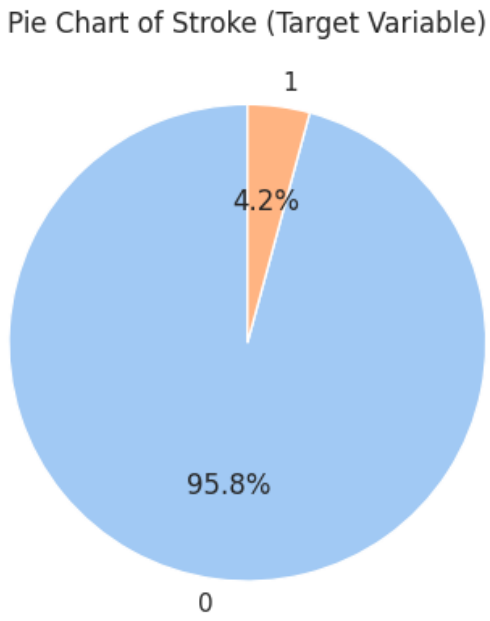
Target variable:

stroke:

Bar chart:



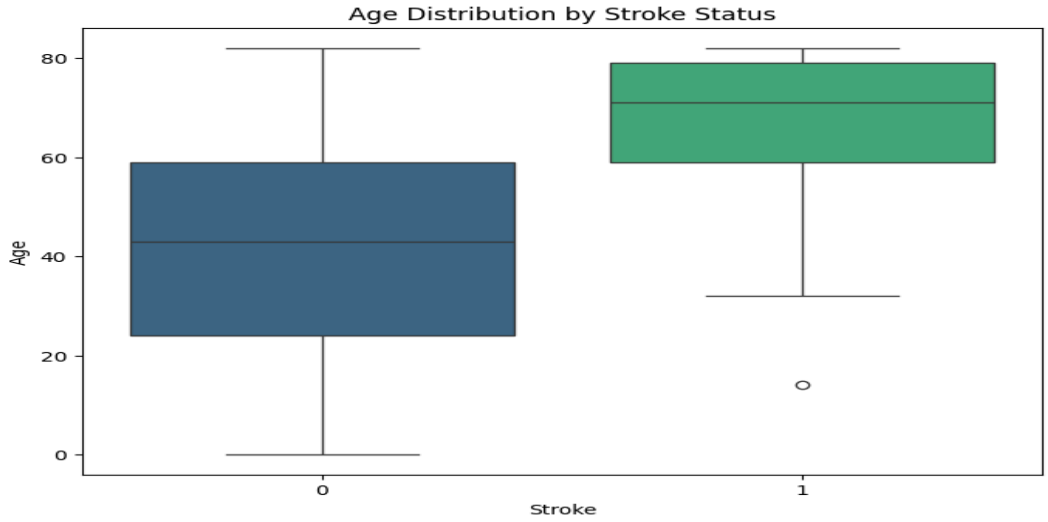
Pie chart:



The plots for the target variable (stroke) provide a clear view of its distribution within the dataset, indicating the proportion of individuals who have experienced a stroke versus those who haven't. This univariate insight is essential for understanding the prevalence of stroke in the sample, which helps contextualize analyses of potential risk factors across other variables.

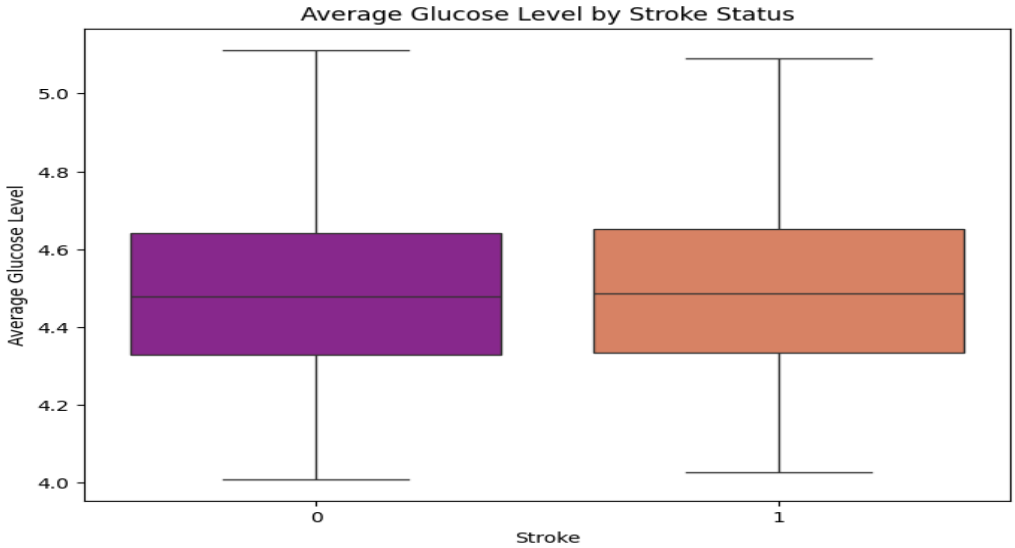
Bivariate Analysis

1. Age and Stroke



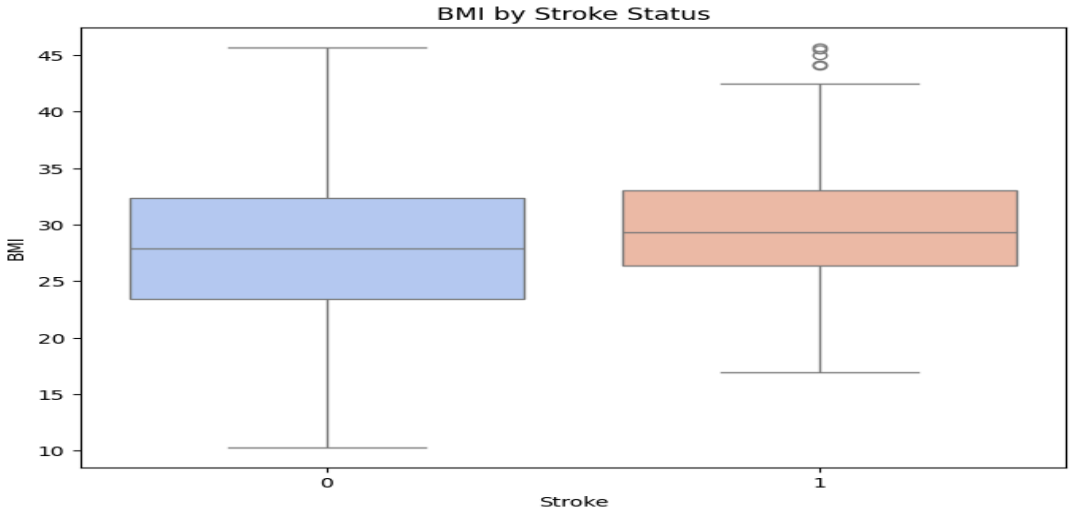
The box plot shows that individuals with strokes (1) are generally older compared to those without strokes (0). This indicates a strong dependency between age and the likelihood of having a stroke, suggesting that age is a significant risk factor for strokes.

2. Avg Glucose Level and Stroke



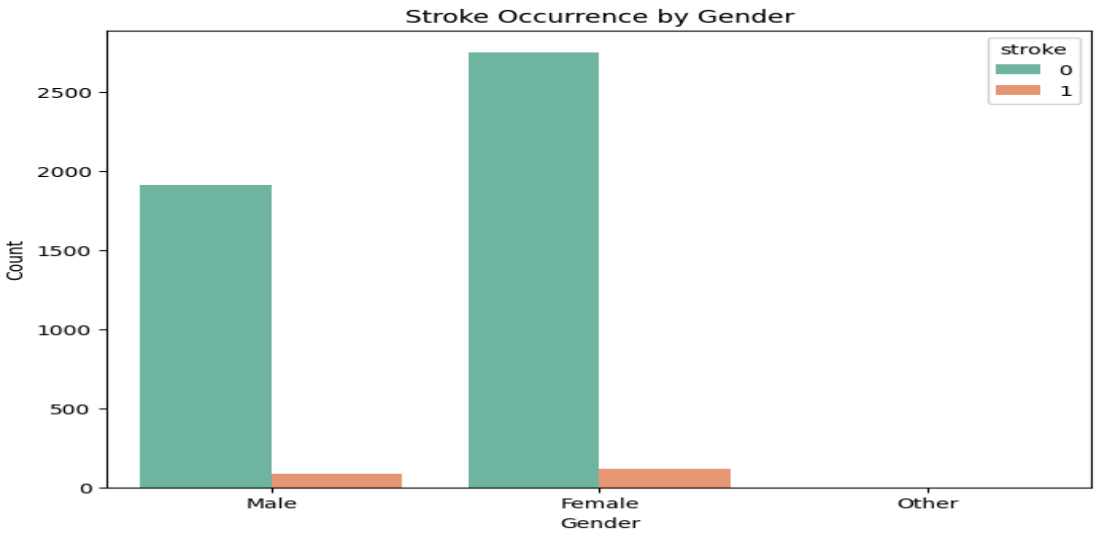
The box plot indicates no significant difference between the glucose levels of individuals with strokes (1) and those without strokes (0). This suggests that average glucose level may not have a strong correlation with stroke occurrence in this dataset.

3. BMI and Stroke

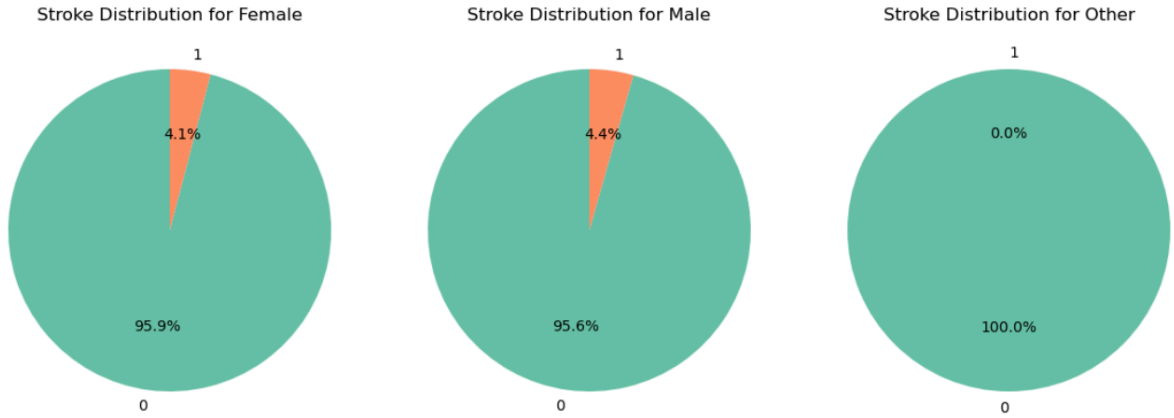


The box plot shows that individuals with strokes (1) tend to have slightly higher BMI values than those without strokes (0). Outliers at the higher end for stroke cases suggest that extremely high BMI may be associated with an increased risk of stroke, though the effect is not pronounced. The presence of outliers indicates some individuals with very high BMI are more likely to have strokes.

4. Gender and Stroke

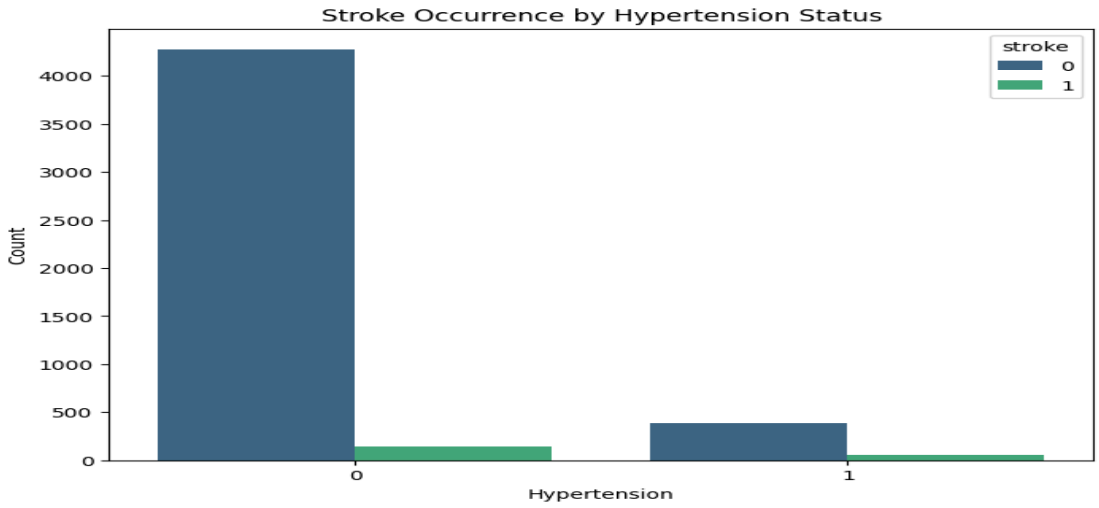


The bar graph shows that while the majority of cases in both genders are "no stroke," females have higher counts than males for both stroke and no-stroke categories. This suggests that gender distribution in the dataset is skewed toward females, but there's no clear indication of a strong gender-based correlation with stroke occurrence.

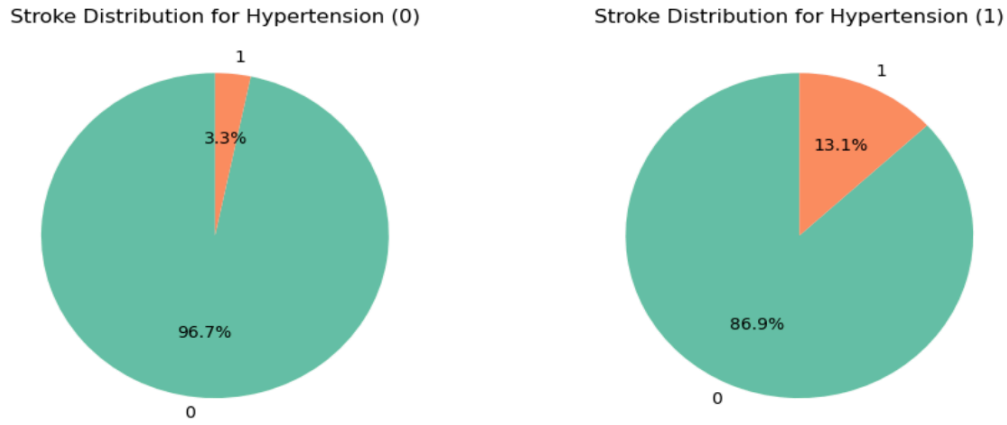


These pie charts do not show strong relation between stroke occurrence and gender.

5. Hypertension and Stroke

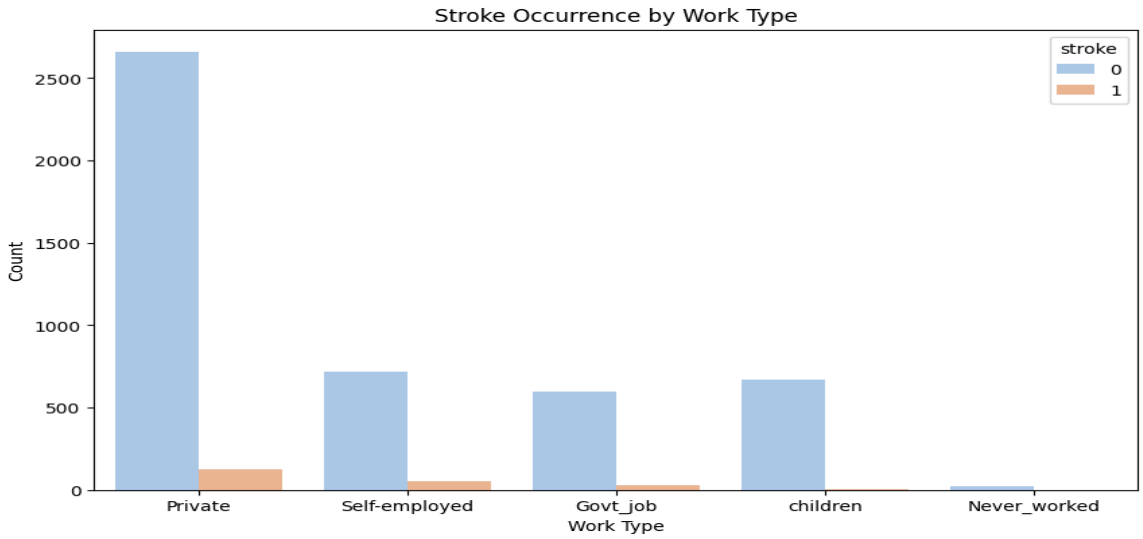


The bar graph indicates that more stroke cases occur in individuals without hypertension. However, the proportion of no strokes among individuals with no hypertension is also much higher compared to those with hypertension. Therefore it is difficult to determine if there is a definitive correlation between the two since data is skewed towards people with no hypertension.

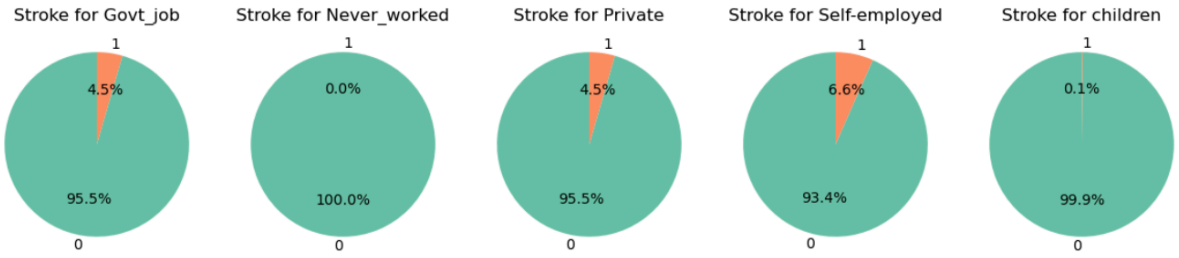


These pie charts show that people with hypertension are more likely to have stroke as compared to people without hypertension.

6. Work Type and Stroke

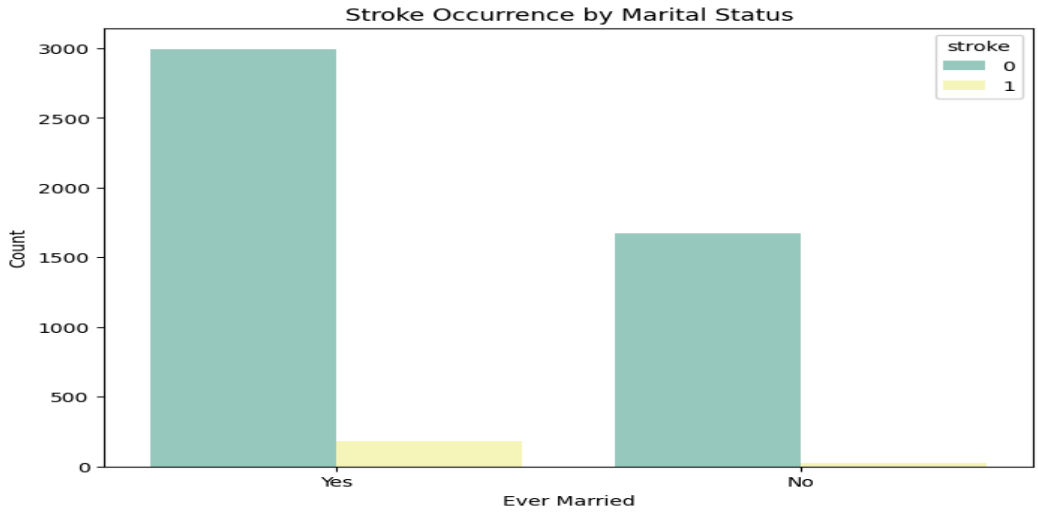


The uneven distribution of categories in this chart makes it difficult to draw meaningful conclusions about the correlation between this factor and stroke occurrence. Further analysis, such as calculating stroke rates or adjusting for category size, would be needed to assess the true relationship.

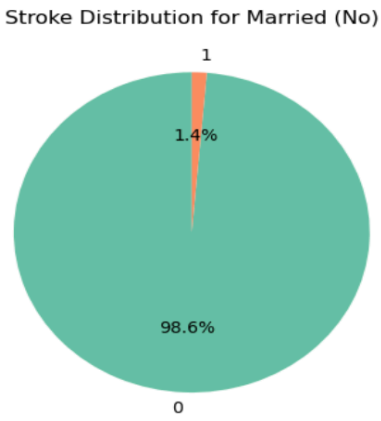


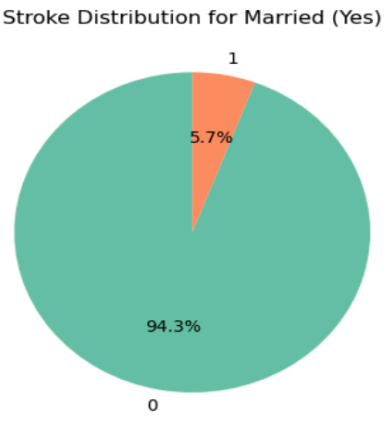
These pie charts do not show strong relation between stroke occurrence and work type. However, they show that children and people who never worked have very weak to no relation with stroke occurrence.

7. Ever Married and Stroke



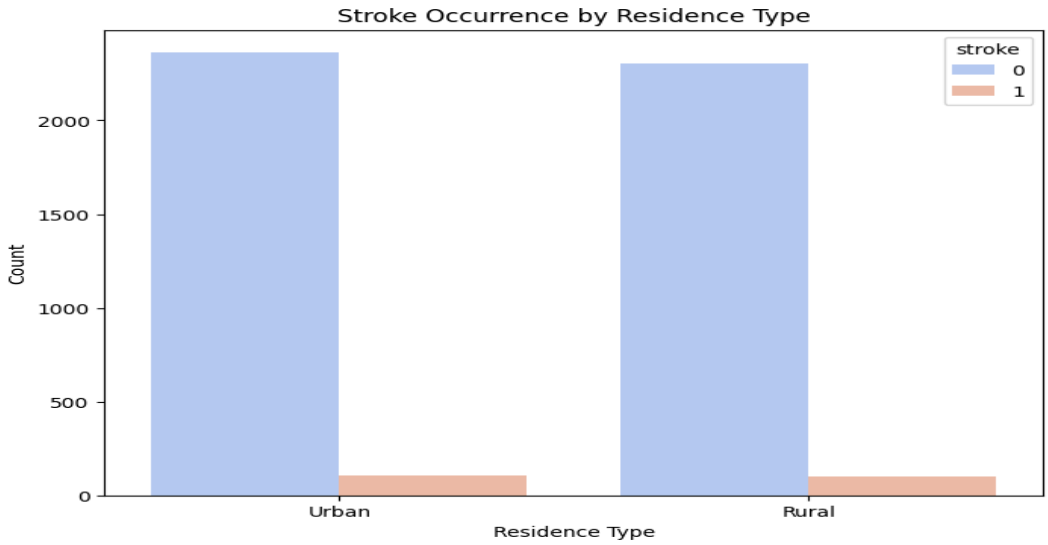
The bar chart shows higher stroke and non-stroke occurrences in the "married" category compared to "non-married." This likely reflects the larger number of married individuals in the dataset, rather than a direct link between marriage and stroke. However could still point to slightly more chances of stroke in the married category.



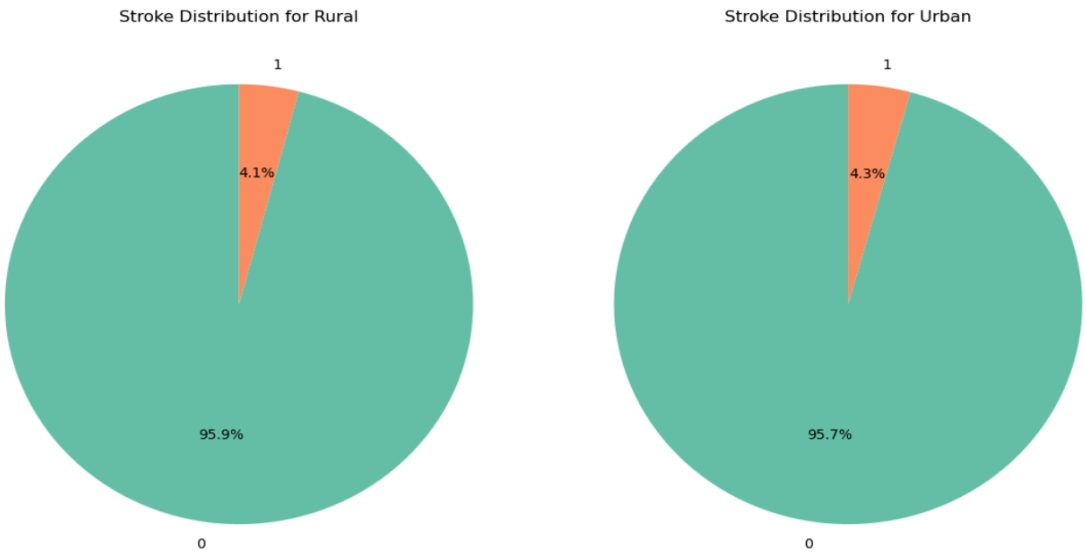


These pie charts show that married people are more likely to have stroke as compared to unmarried people.

8. Residence Type and Stroke

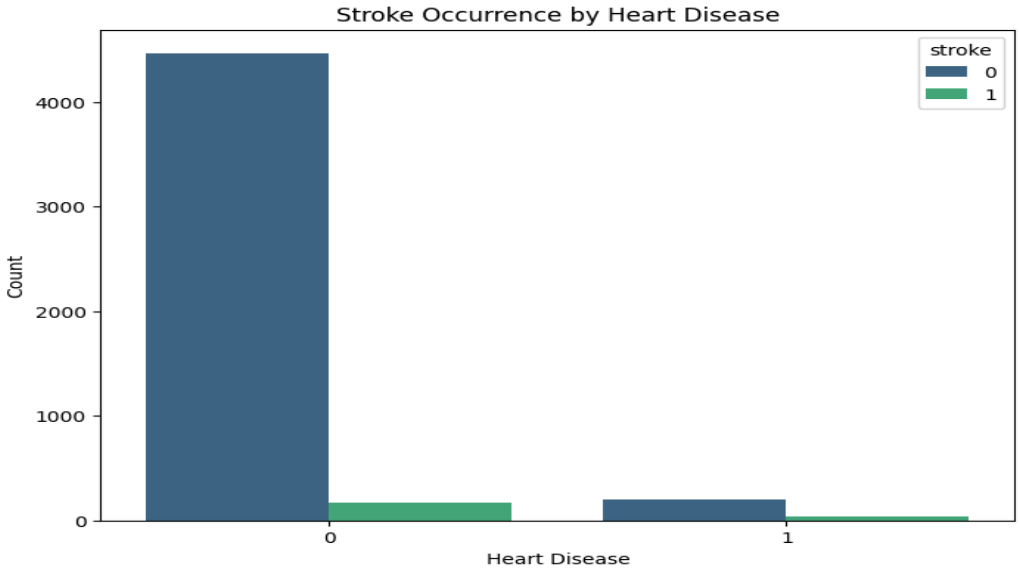


There does not appear to be any correlation between stroke occurrence and residence type from this graph.

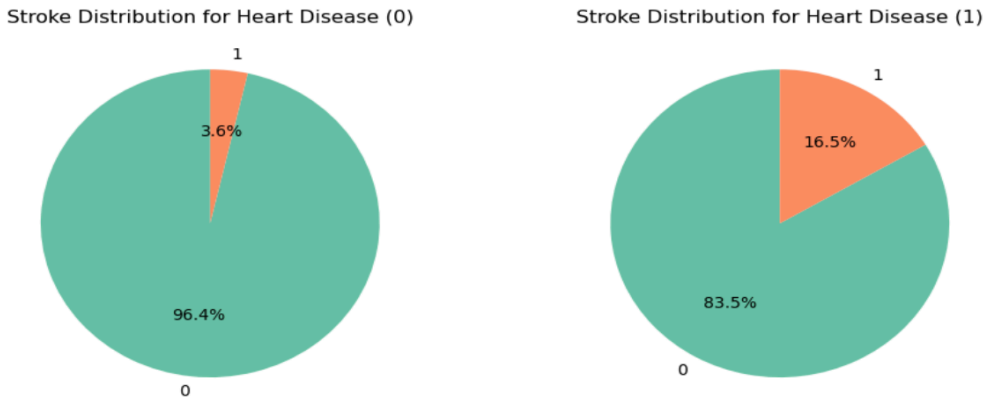


These pie charts do not show strong relation between stroke occurrence and residence type.

9. Heart Disease and Stroke

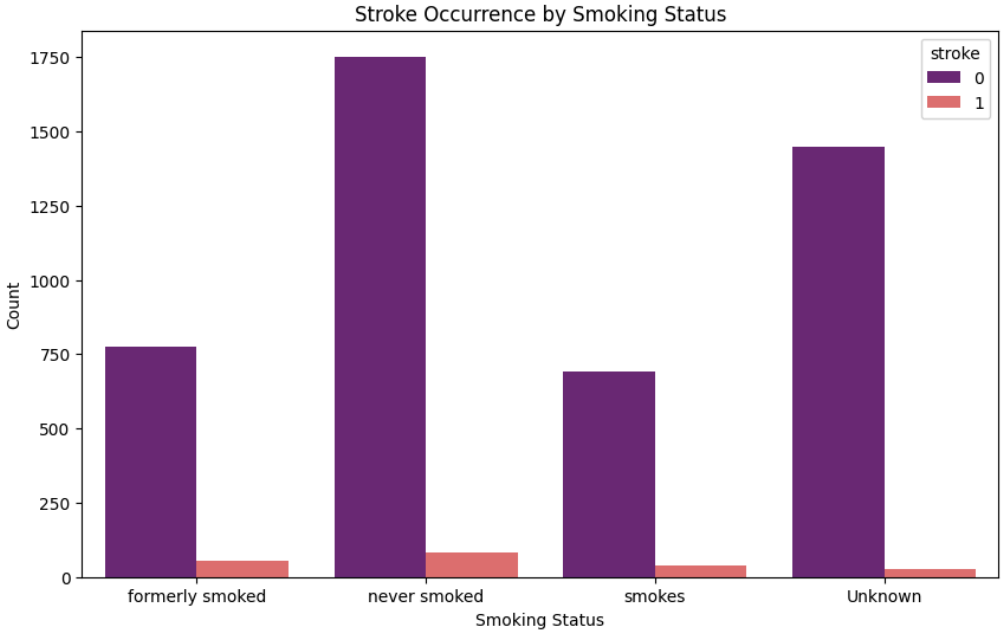


This bar graph suggests that data is skewed towards people without heart disease so no meaningful insights can be derived from this.

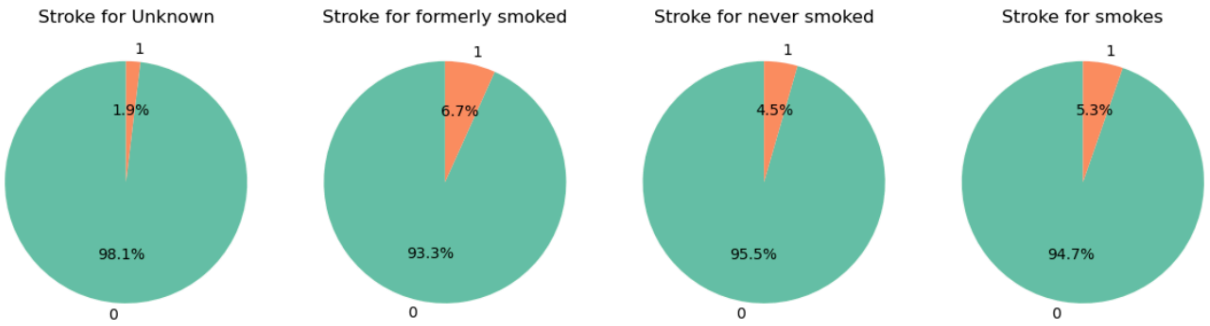


These pie charts show that people with heart disease are more likely to have stroke as compared to people without heart disease.

10. Smoking Status and Stroke

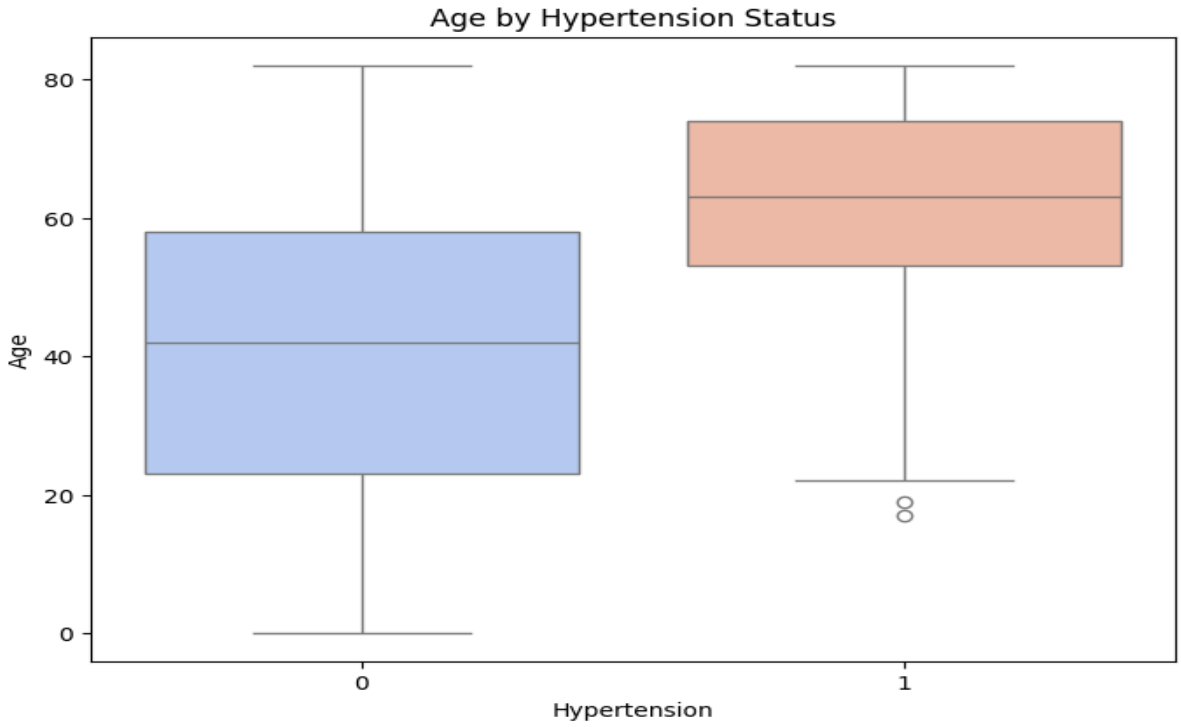


The uneven distribution of categories in this chart makes it difficult to draw meaningful conclusions about the correlation between this factor and stroke occurrence. Further analysis, such as calculating stroke rates or adjusting for category size, would be needed to assess the true relationship.



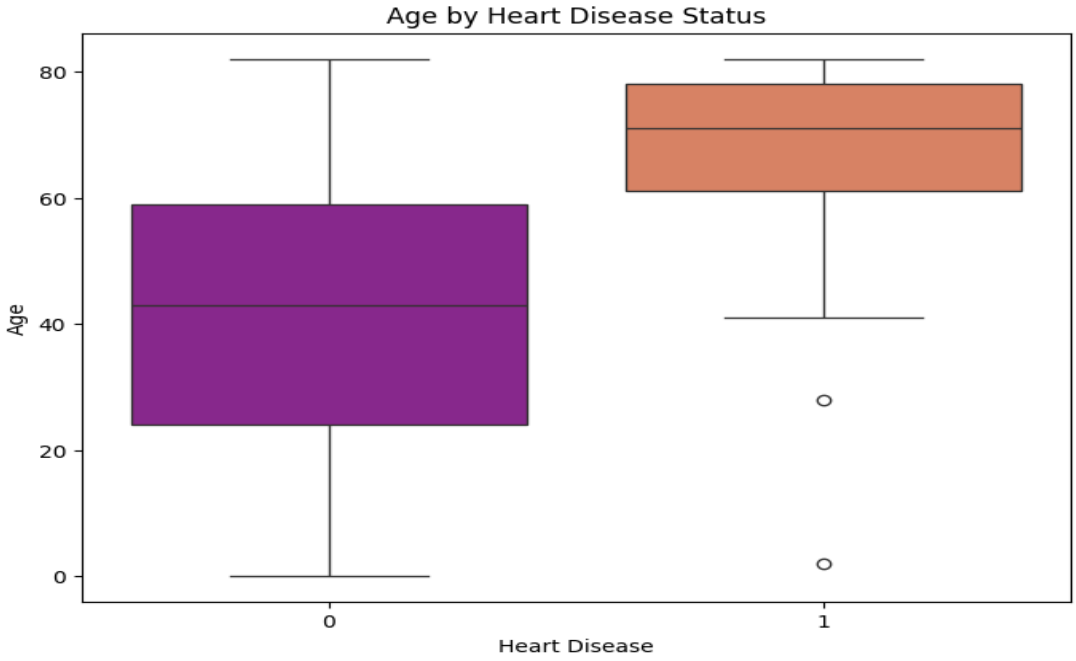
These pie charts do not show strong relation between stroke occurrence and smoking status. However, ‘formerly smoked’ and ‘smokes’ have slightly more chance than ‘never smoked’.

11. Age and Hypertension



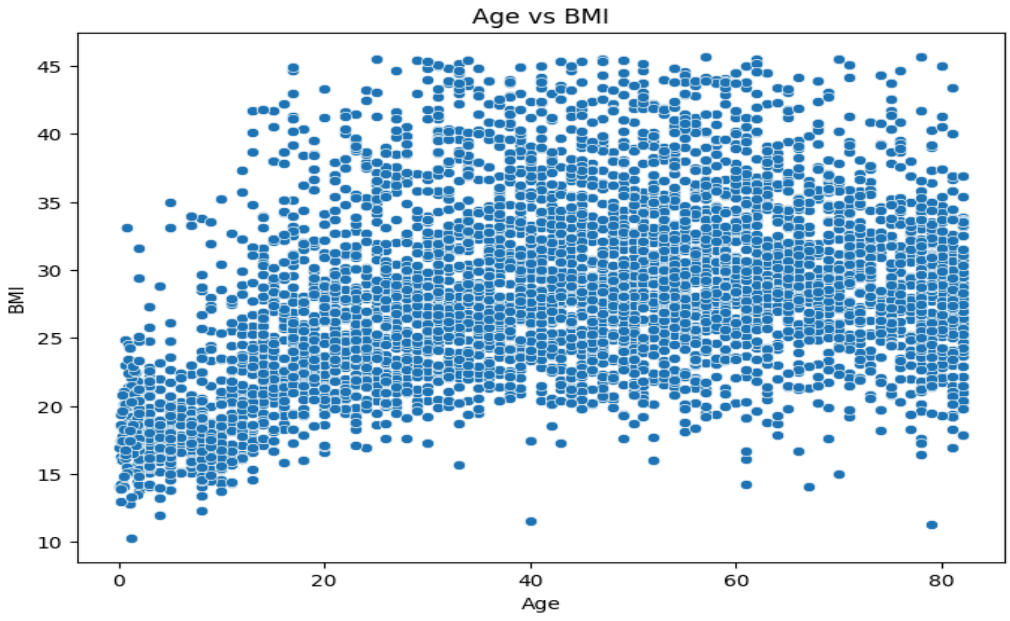
The box plot shows that individuals with hypertension (1) are generally older compared to those without hypertension (0). This indicates a strong dependency between age and the likelihood of having hypertension, suggesting that age is a significant risk factor for hypertension.(with the exception of some outliers)

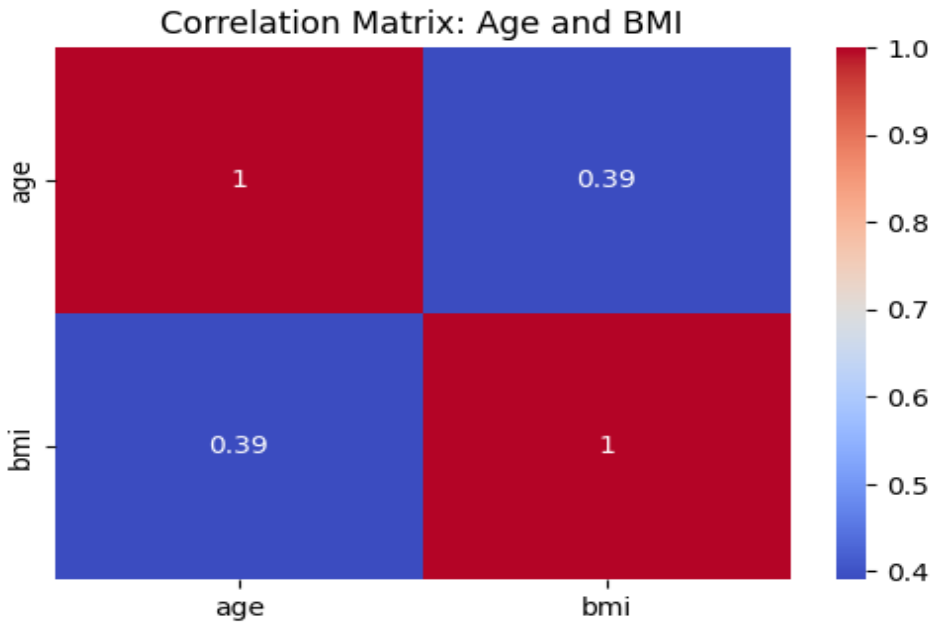
12. Age and Heart Disease



The box plot shows that individuals with heart disease (1) are generally older compared to those without heart disease (0). This indicates a strong dependency between age and the likelihood of having heart disease, suggesting that age is a significant risk factor for heart disease (with the exception of some outliers).

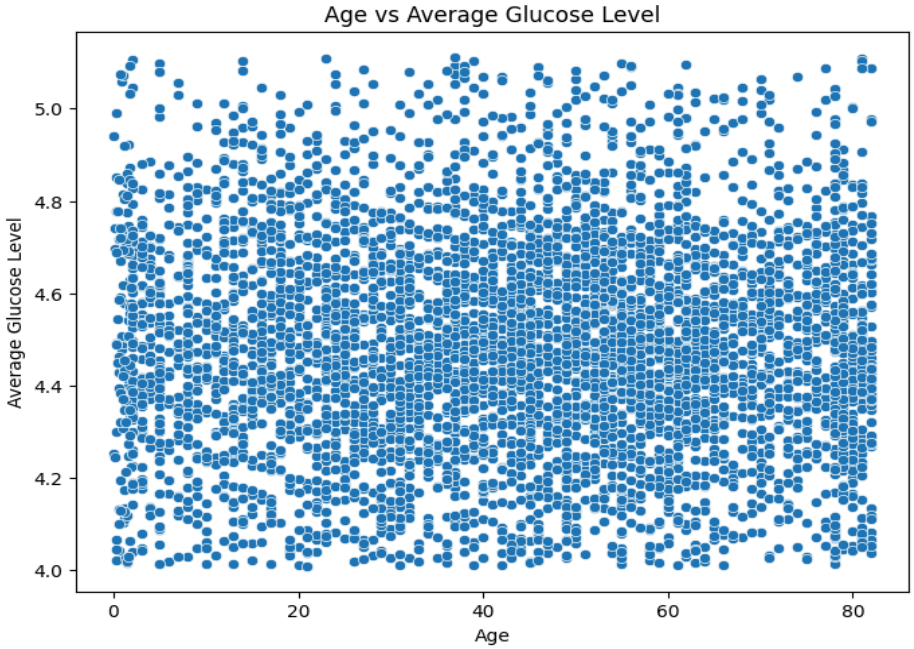
13. Age and BMI

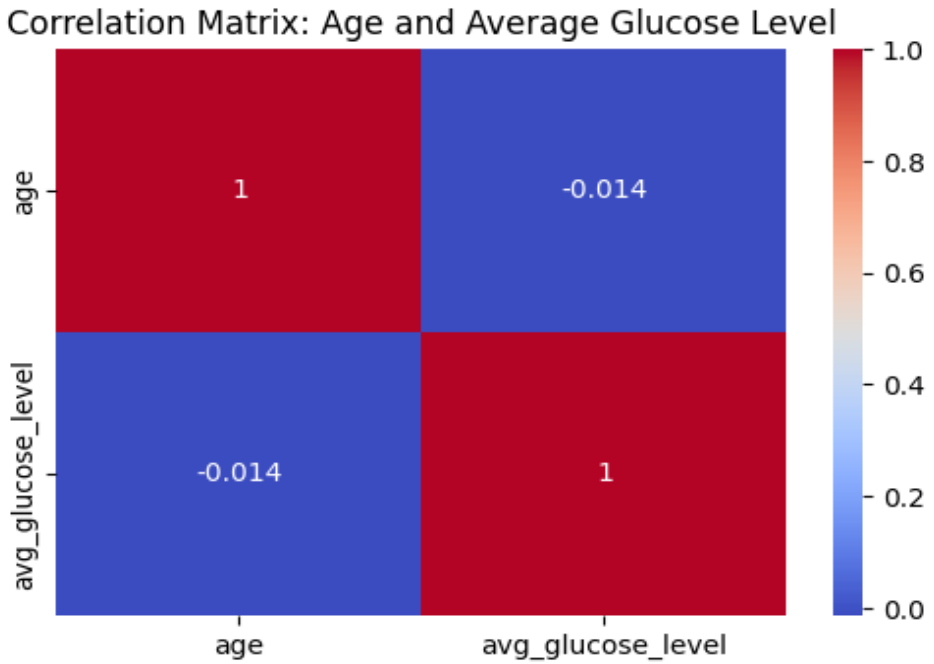




The scatter plot shows no clear pattern, indicating a weak relationship between the two variables. The correlation matrix, with a value of 0.39, suggests a weak positive correlation between age and BMI. However, this correlation is not strong enough to suggest a meaningful connection.

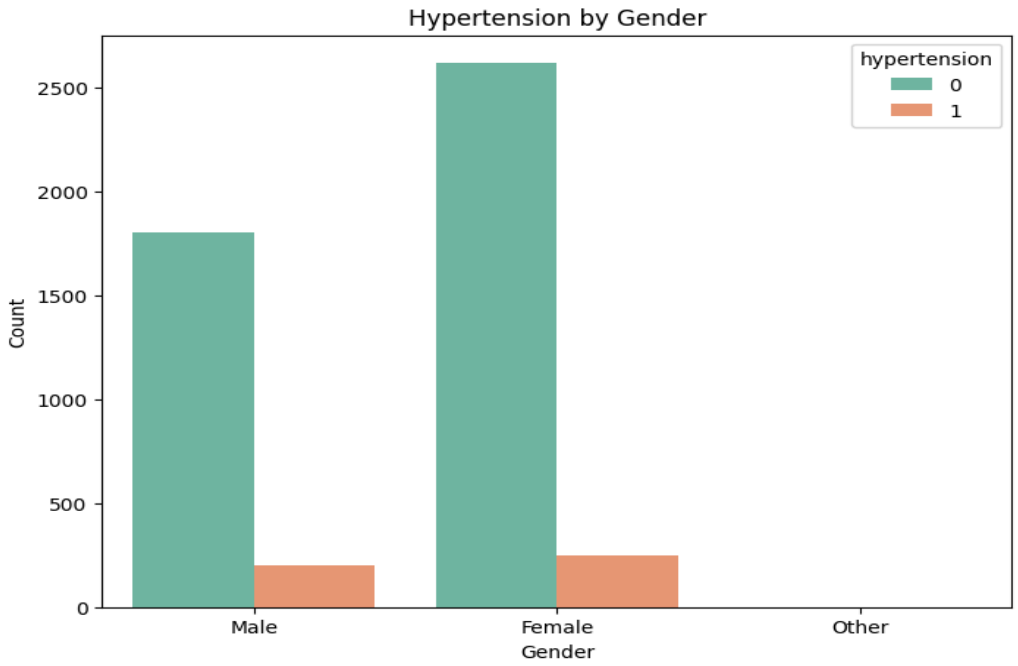
14. Age and Glucose Level

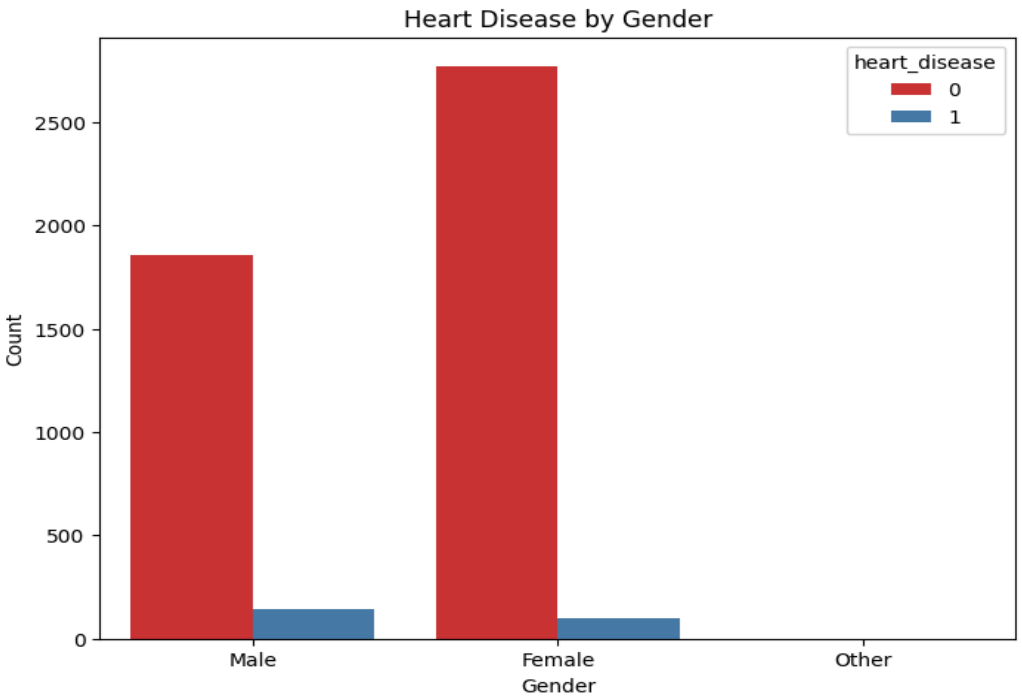




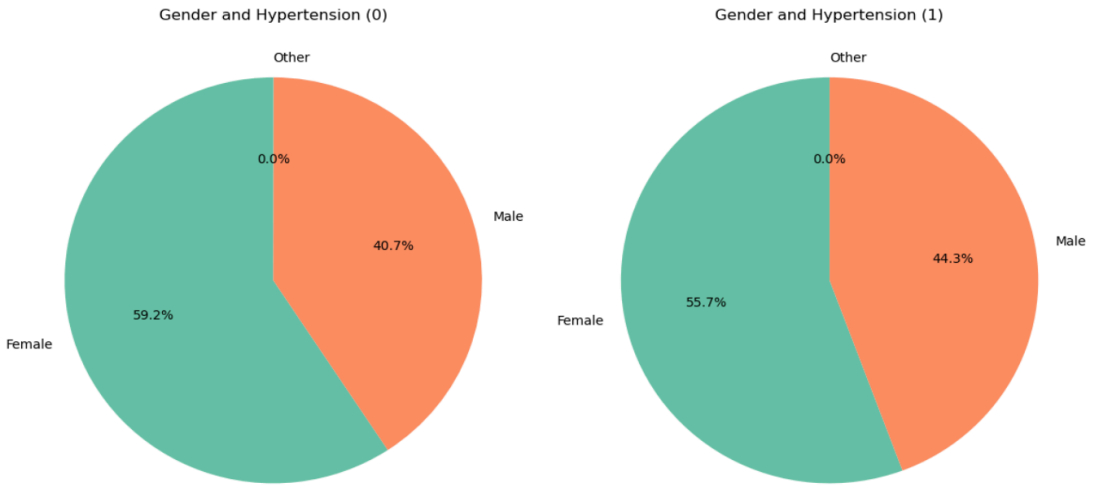
The scatter plot shows no clear trend, indicating a very weak or no relationship between the two variables. The correlation matrix, with a value of -0.14, confirms a very weak negative correlation, suggesting that age has little to no impact on glucose levels.

15. Gender and Hypertension/Heart Disease

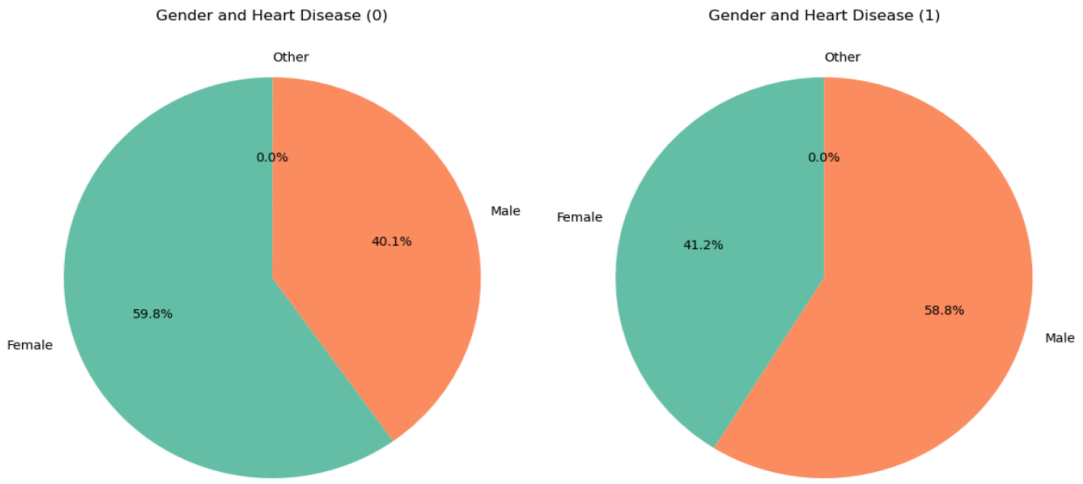




The uneven distribution of categories in these charts makes it difficult to draw meaningful conclusions about the correlation between these factors.

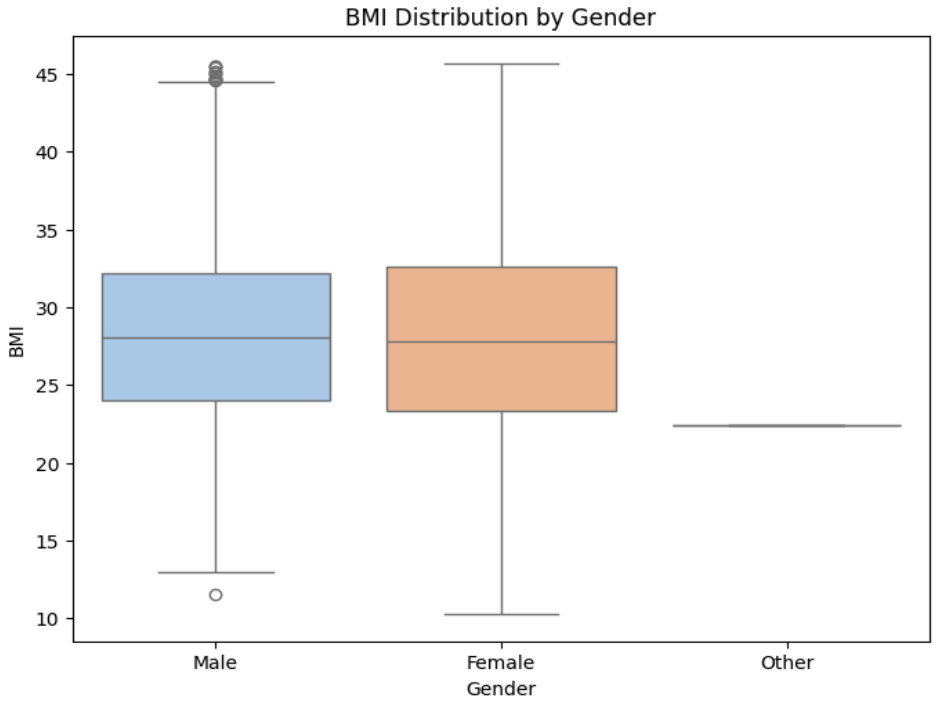


These pie charts do not show any important relation between gender and hypertension.

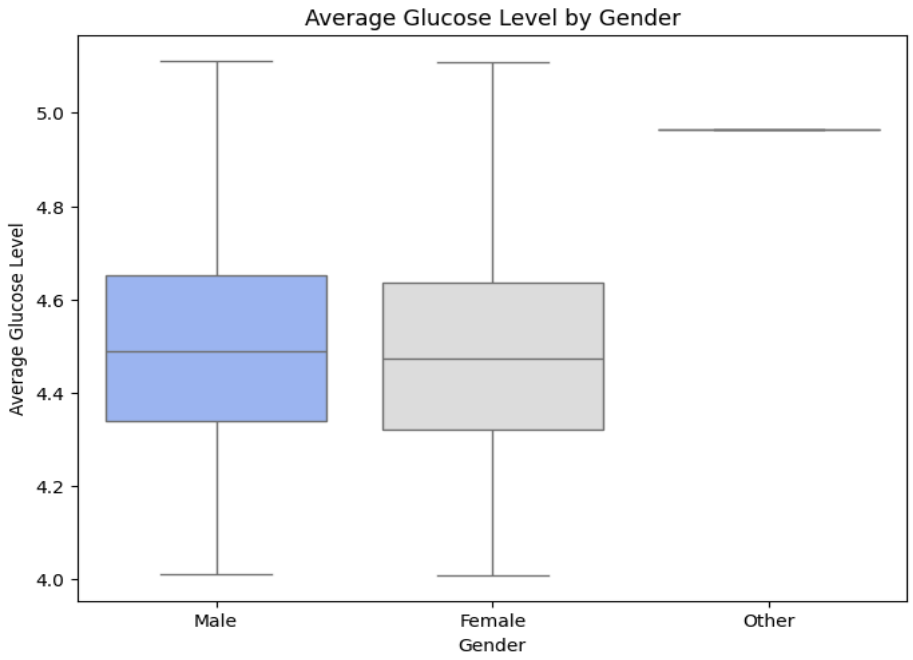


These pie charts show relation between gender and heart disease where heart disease is more common in males.

16. Gender and BMI/Glucose Level

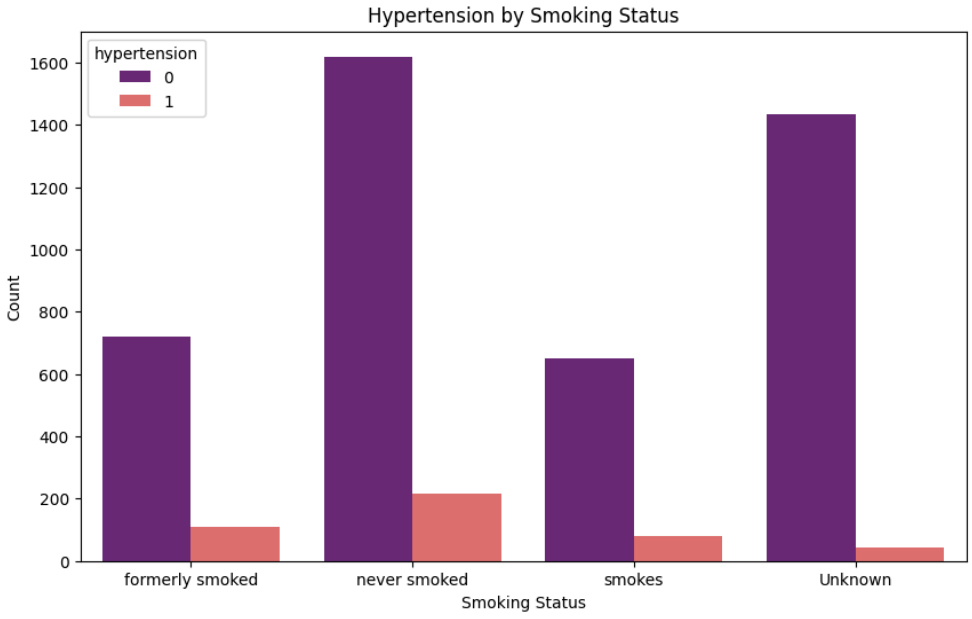


There doesn't seem to be any correlation between gender and bmi. With the exception of some outliers the overall distribution of is same across both genders

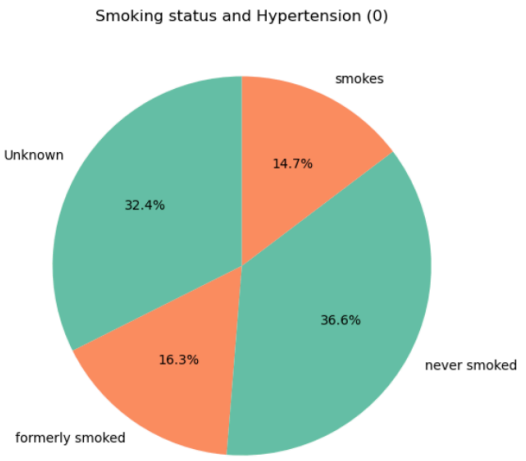


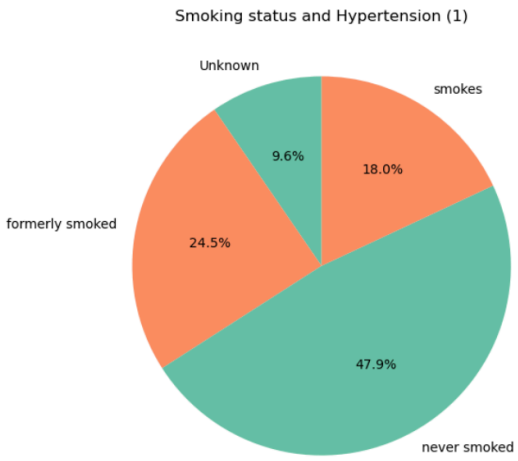
Females have slightly lower average glucose levels but not significant enough to determine a correlation between gender and glucose levels

17. Smoking Status and Hypertension



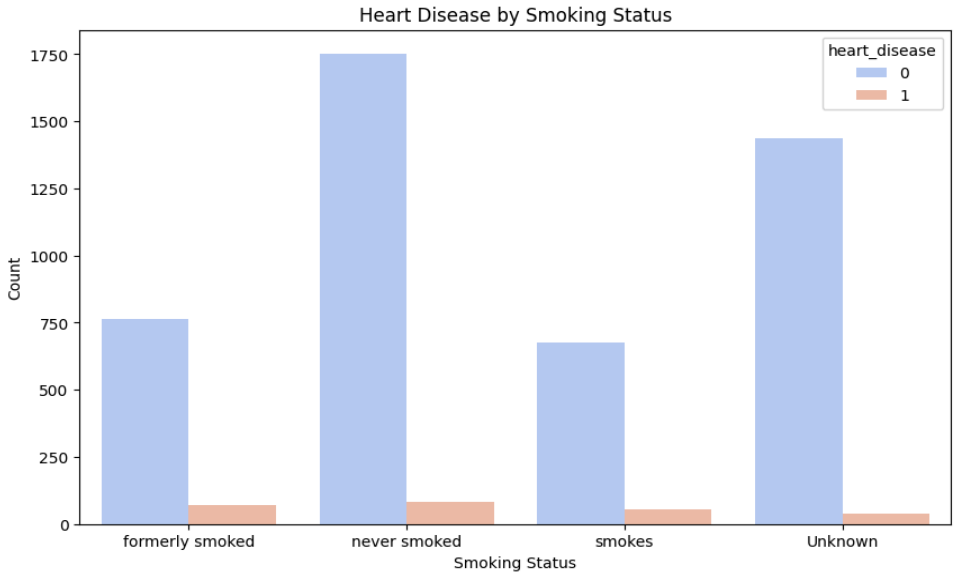
The uneven distribution of categories in this chart makes it difficult to draw meaningful conclusions about the correlation between these factors.



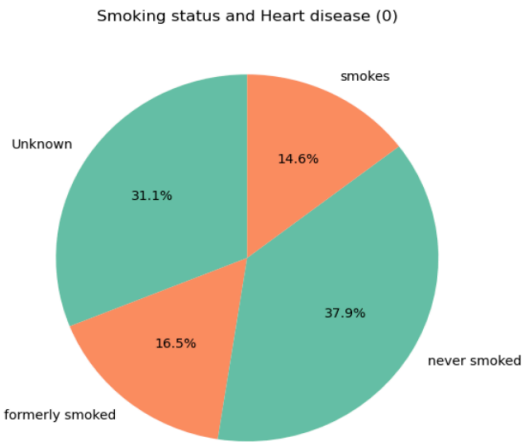


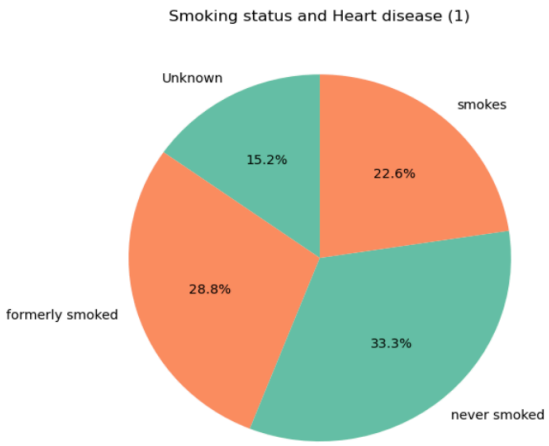
These pie charts do not provide any important information.

18. Smoking Status and Heart Disease



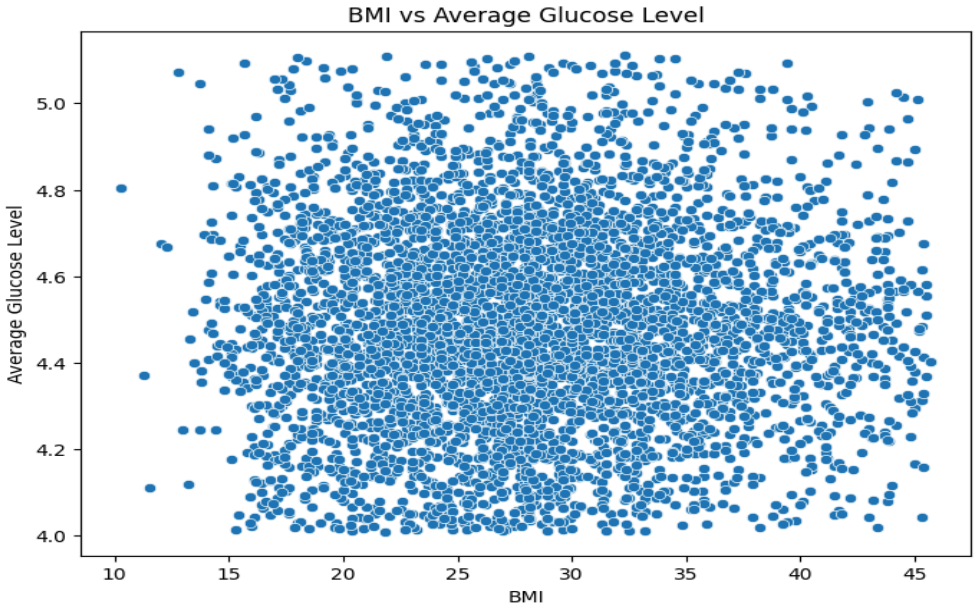
The uneven distribution of categories in this chart makes it difficult to draw meaningful conclusions about the correlation between these factors.

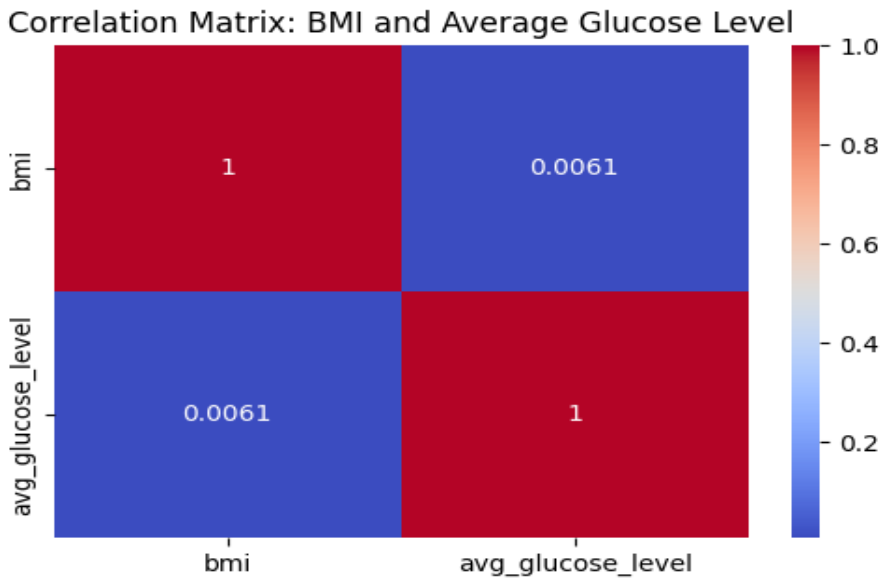




These pie charts do not provide any important information.

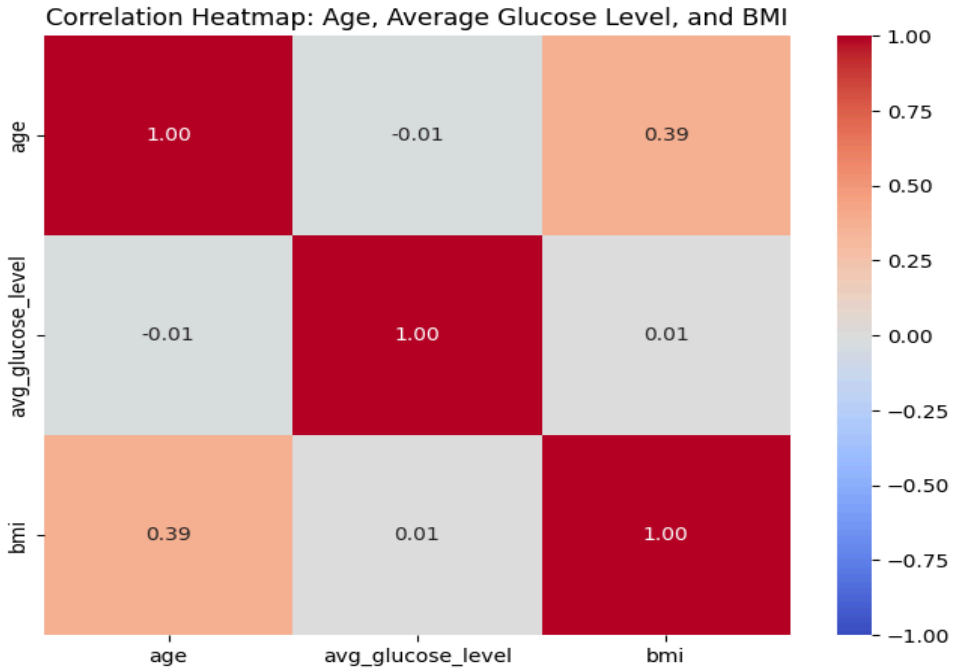
19. BMI and Glucose Level



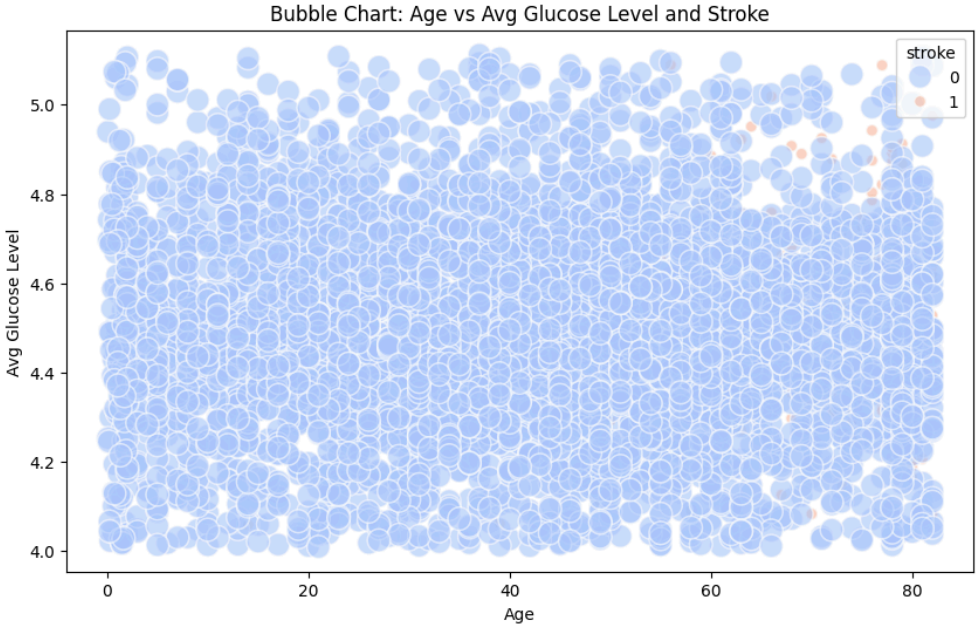


The scatter plot shows no clear pattern, indicating a very weak or no apparent relationship between the two variables. The correlation matrix, with a value of 0.0061, suggests a very very weak positive correlation between avg glucose level and BMI. It is not strong enough to suggest a meaningful connection.

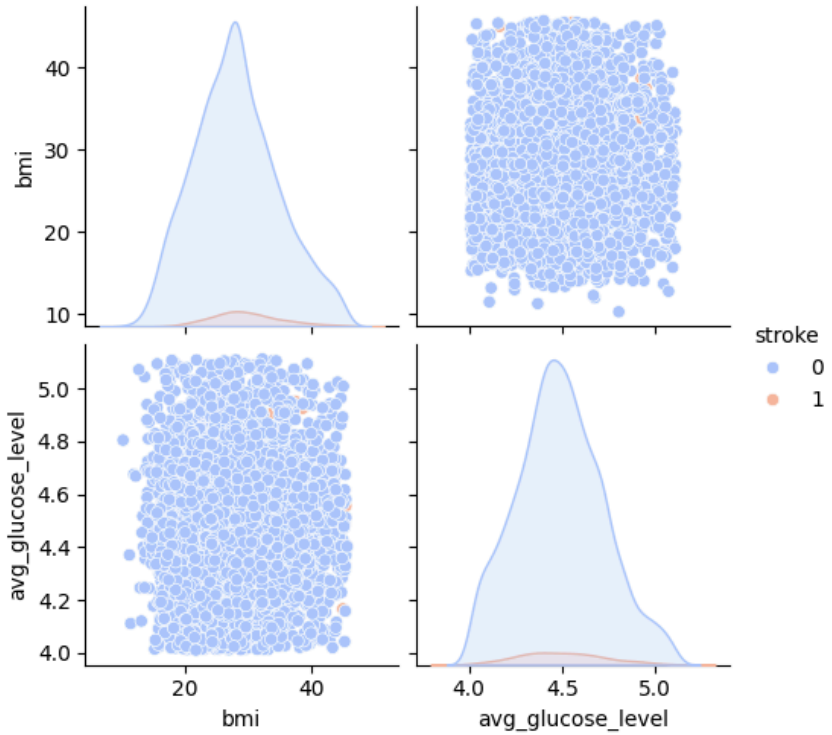
Multivariate Analysis



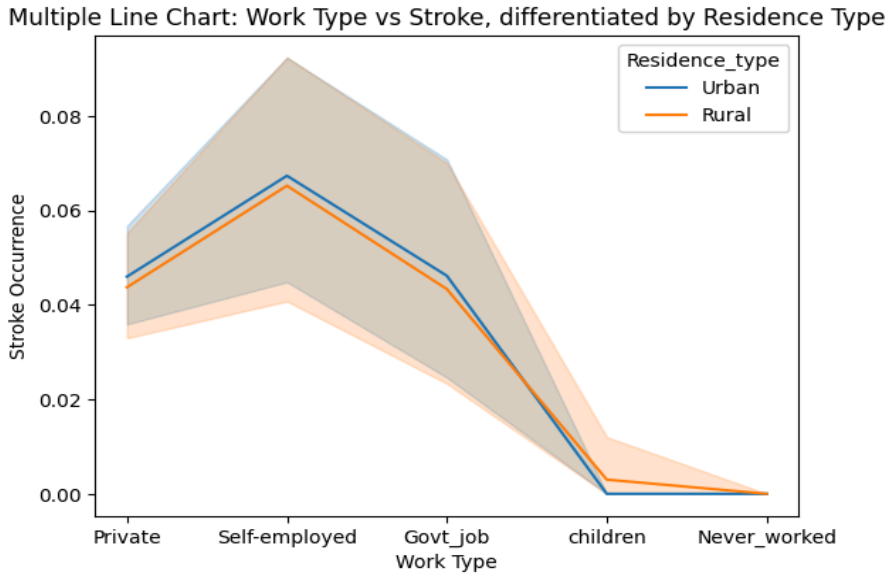
Age and BMI have a weak positive correlation (0.39), while age and average glucose level (-0.01), as well as average glucose level and BMI (0.01), show very weak correlations.



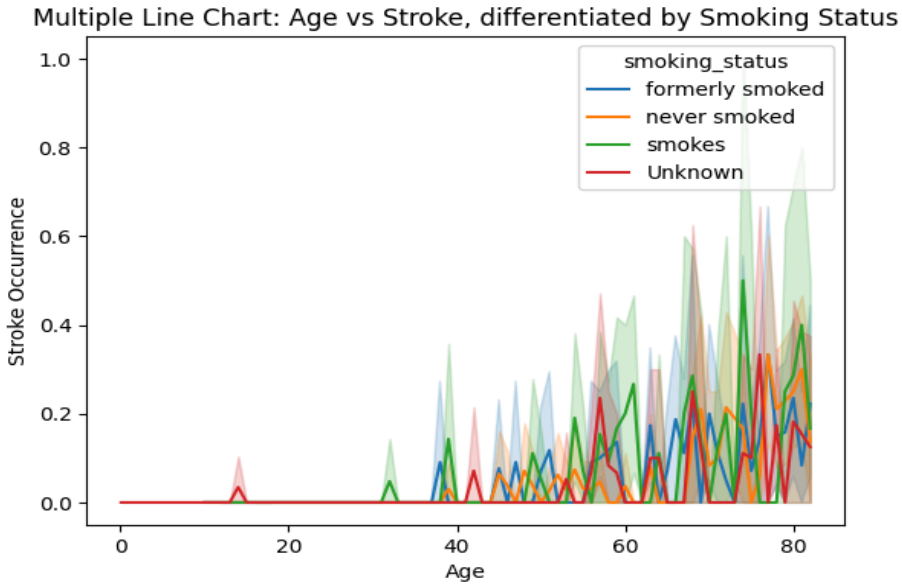
Stroke cases (in orange) seem more prevalent in the older age range, while average glucose levels vary widely across ages with no distinct clustering pattern.



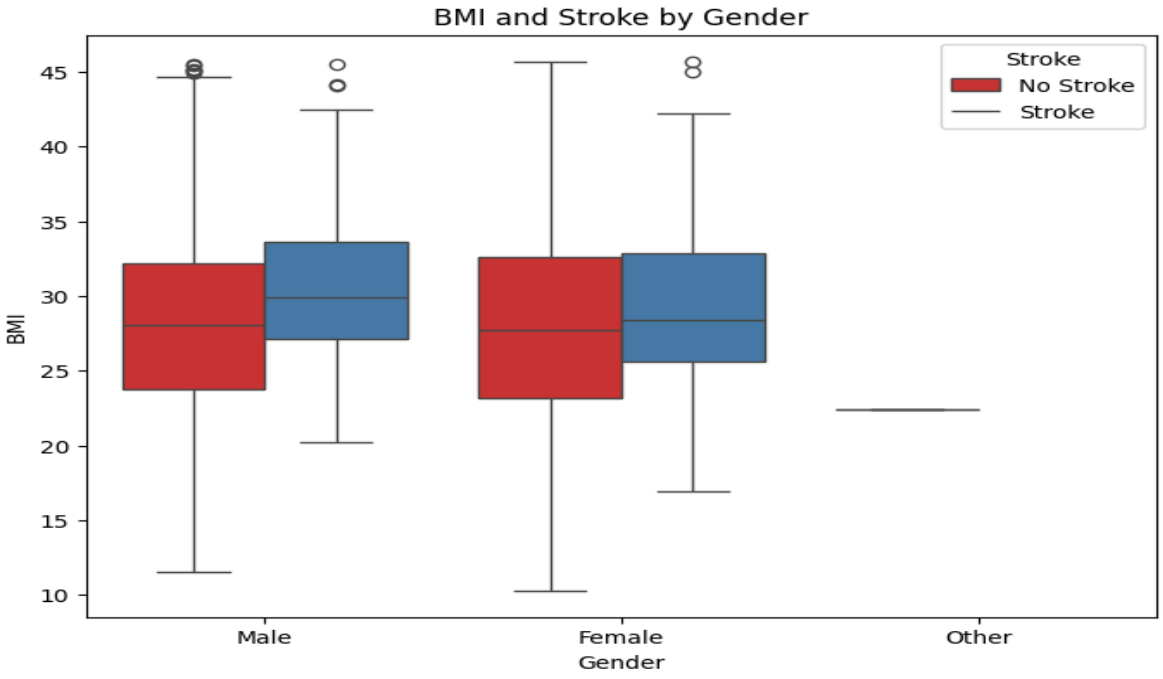
Stroke cases are relatively rare and scattered across all BMI and average glucose levels, with no clear pattern. Most data points cluster around BMI values of 20-40 and average glucose levels of 4.0-5.0. There is no strong relationship between BMI and average glucose level, and stroke cases (in orange) do not concentrate in any specific range of these variables.



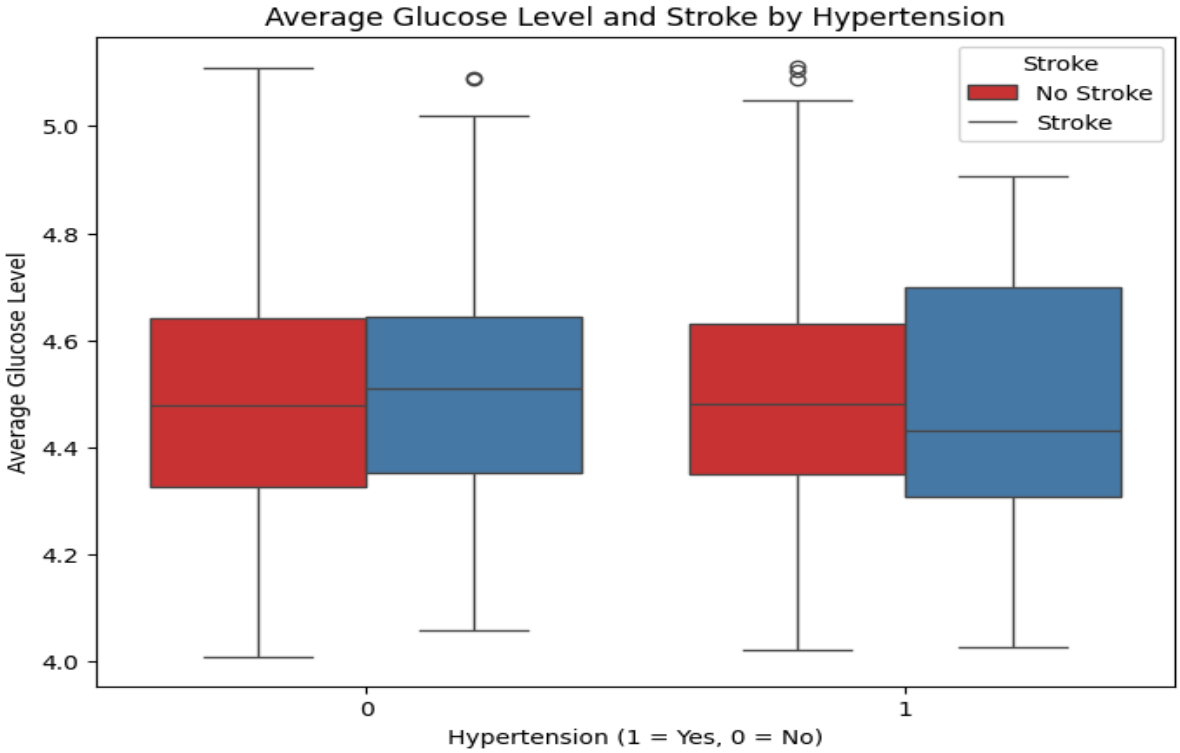
This shows stroke occurrences by work type, separated by urban and rural residence. Self-employed individuals have the highest stroke rates, while children and those who never worked have the lowest.



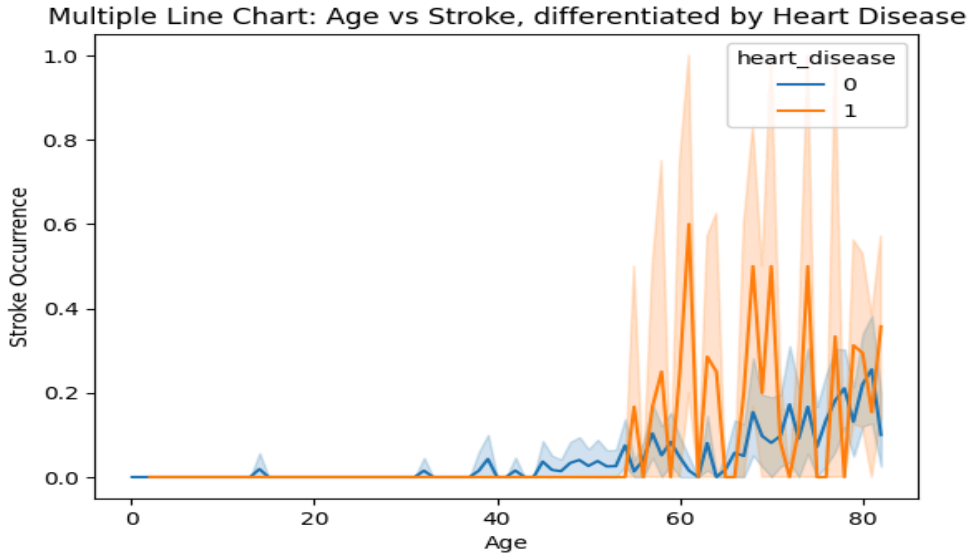
Illustrates stroke rates increasing with age, differentiated by smoking status. Stroke occurrences rise notably after age 40, with smokers showing higher rates at older ages.



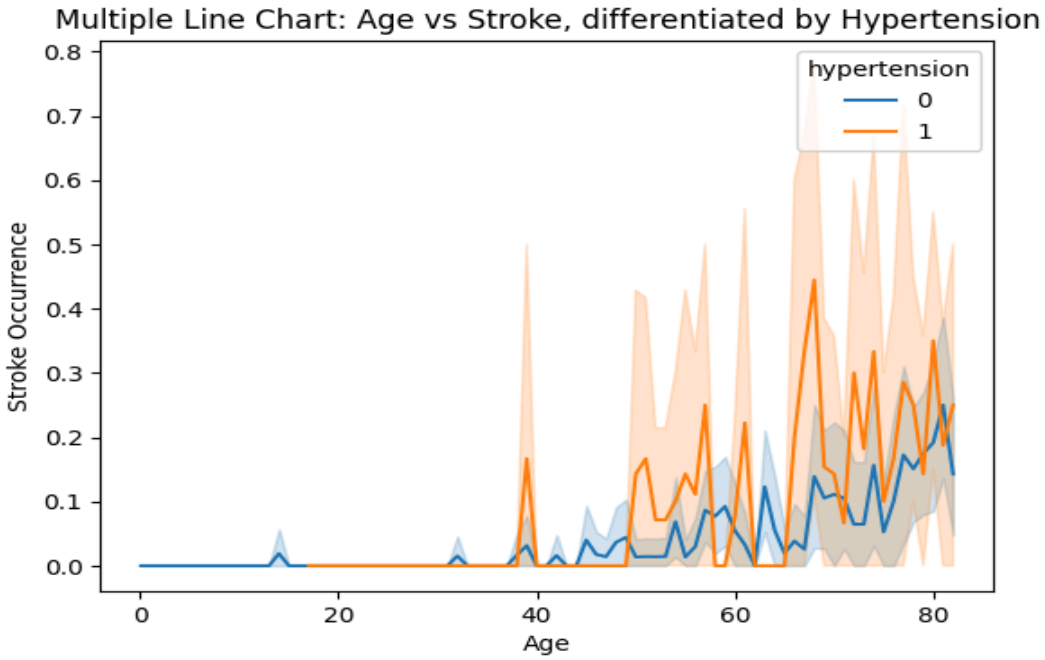
Shows the distribution of BMI for individuals with and without strokes, separated by gender. It shows that males and females with strokes tend to have higher BMIs compared to those without strokes. The BMI range is fairly consistent across genders, but outliers are more evident in the male group.



This shows the distribution of average glucose levels for individuals with and without strokes, separated by hypertension status. It indicates that those with and without hypertension have similar glucose level distributions, but stroke occurrences are slightly more spread out at higher levels in both groups.



Stroke occurrence rises with age and is higher among those with heart disease (orange line), especially noticeable after age 40. This trend is similar to that in hypertensive individuals, with a peak around age 60-80, indicating a strong link between heart disease and stroke occurrence in older adults.



Stroke occurrence increases with age and is more frequent in individuals with hypertension (orange line) than those without (blue line), particularly noticeable after age

40. Hypertensive individuals show a higher variability and a rising stroke risk with age, peaking around age 60-80.

Descriptive Analysis

1. What is the distribution of residence types (Urban vs. Rural)?

```
residence_distribution = df['Residence_type'].value_counts()
print("Distribution of residence types (Urban vs. Rural):\n", residence_distribution)
```

```
Distribution of residence types (Urban vs. Rural):
 Residence_type
Urban    2470
Rural    2403
Name: count, dtype: int64
```

The distribution of residence types in the dataset is almost evenly split between Urban and Rural areas. This balanced distribution suggests that the dataset represents both residence types fairly equally, minimizing bias in analyses or predictions related to residence type.

2. What is the average BMI of patients who experienced a stroke compared to those who didn't?

```
average_bmi_by_stroke = df.groupby('stroke')['bmi'].mean()
print("Average BMI of patients who experienced a stroke compared to those who didn't:\n", average_bmi_by_stroke)
```

```
Average BMI of patients who experienced a stroke compared to those who didn't:
 stroke
0    28.139632
1    29.994634
Name: bmi, dtype: float64
```

The average BMI (Body Mass Index) of patients who experienced a stroke is slightly higher than those who did not. This indicates that BMI may be a contributing factor or associated risk for stroke.

3. How do hypertension prevalence differ among genders?

```
hypertension_by_gender = pd.crosstab(df['gender'], df['hypertension'])
print("Hypertension prevalence by gender (%):\n", hypertension_by_gender)
```

```
Hypertension prevalence by gender (%):
 hypertension     0    1
gender
Female        2621  250
Male          1802  199
Other            1    0
```

Out of 2871 females, about 8.7% have hypertension while out of 2001 males, about 10% have hypertension. This analysis suggests that gender might play a role in hypertension prevalence, with males showing a higher rate.

4. Which work type has the highest occurrence of stroke?

```
stroke_by_work_type = pd.crosstab(df['work_type'], df['stroke'])

work_type_with_highest_stroke = stroke_by_work_type[1].idxmax()
print("Work type with the highest occurrence of stroke:", work_type_with_highest_stroke)
```

```
Work type with the highest occurrence of stroke: Private
```

People working in the Private sector are more likely to experience strokes compared to other work types. This could be attributed to factors such as stress, work environment, or lifestyle associated with private sector jobs.

5. For patients with heart disease, what is the average age and BMI?

```
heart_disease_avg = df[df['heart_disease'] == 1][['age', 'bmi']].mean()
print("Average age and BMI for patients with heart disease:\n", heart_disease_avg)
```

```
Average age and BMI for patients with heart disease:
 age    68.275720
bmi    30.206996
dtype: float64
```

Patients with heart disease have an average age of approximately 68.28 years. This indicates that heart disease is more prevalent in older individuals, aligning with the general understanding that the risk of heart-related conditions increases with age. Moreover, the average BMI of patients with heart disease is about 30.21. A BMI of 30 or above is classified as obese, suggesting a strong correlation between obesity and heart disease.

6. How does marital status correlate with stroke occurrence across different work types?

```
marital_status_work_stroke = pd.crosstab(
    [df['ever_married'], df['work_type']],
    df['stroke'])

print("Correlation of marital status with stroke occurrence across different work types:\n", marital_status_work_stroke)
```

```
Correlation of marital status with stroke occurrence across different work types:
 stroke                       0    1
ever_married work_type
No           Govt_job       105    5
             Never_worked    22    0
             Private        777   12
             Self-employed  100    5
             children       669    1
Yes          Govt_job       492   23
             Private       1885  113
             Self-employed  618   46
```

For unmarried individuals, stroke prevalence is low for individuals who never worked while those employed in the private sector have slightly higher prevalence of stroke. For married individuals, stroke prevalence is lowest in the government sector and highest in the private sector. Overall, stroke prevalence is low for unmarried individuals as compared to married.

7. What is the distribution of stroke occurrences across genders?

```
stroke_distribution_by_gender = pd.crosstab(df['gender'], df['stroke'])
print("Distribution of stroke occurrences across genders:\n", stroke_distribution_by_gender)
```

```
Distribution of stroke occurrences across genders:
 stroke     0    1
gender
Female  2754  117
Male    1913   88
Other      1    0
```

Out of 2871 females, about 4% experienced a stroke while out of 2001 males, about 4.3% had a stroke. This analysis suggests that gender might play a role in stroke prevalence, with males showing a higher rate.

8. What is the correlation between average glucose level, BMI, age, and stroke occurrence?

```
correlation_data = df[['avg_glucose_level', 'bmi', 'age', 'stroke']]

correlation_matrix = correlation_data.corr()
print("Correlation matrix between avg_glucose_level, BMI, age, and stroke occurrence:\n", correlation_matrix)
```

```
Correlation matrix between avg_glucose_level, BMI, age, and stroke occurrence:
                    avg_glucose_level       bmi       age    stroke
avg_glucose_level           1.000000  0.006053 -0.013543  0.009842
bmi                         0.006053  1.000000  0.390506  0.055994
age                        -0.013543  0.390506  1.000000  0.232465
stroke                      0.009842  0.055994  0.232465  1.000000
```

The correlation coefficient between average glucose level and stroke occurrence is 0.0098, indicating a very weak or negligible relationship. This suggests that glucose level alone may not strongly predict stroke. Age and bmi have a moderate positive correlation of 0.39 aligning with the notion that bmi increases with age. Also, the correlation coefficient between age and stroke is 0.232, which is a moderate positive correlation. This suggests that older individuals are more likely to experience a stroke.