STROKE DATASET

ANALYSIS REPORT

Duong Lam Tuan Anh

This analysis was made with the use of Python.
The relevant codes can be found here.

OUTLINE

1. OBJECTIVE OF THE ANALYSIS

Determine the risk factors of stroke.

By examining the relationship between age, hypertension, heart disease, average glucose level, BMI, smoking status, and stroke, this analysis aim to determine the factors that are associated with a higher risk of stroke.

2. DATASET DESCRIPTION

Stroke Dataset

- ***gender***: categorical variable indicating the gender of the individual (male or female).

- ***age***: continuous variable indicating the age of the individual.

- ***hypertension***: binary variable indicating whether the individual has hypertension or not (1 for yes, 0 for no).

- ***heart_disease***: binary variable indicating whether the individual has heart disease or not (1 for yes, 0 for no).

- ***ever_married***: categorical variable indicating whether the individual is married or not (yes or no).

- ***work_type***: categorical variable indicating the type of work the individual does (private, self-employed, government job, children, or never worked).

- ***residence_type***: categorical variable indicating whether the individual lives in an urban or rural area.

- ***avg_glucose_level***: continuous variable indicating the average glucose level in the individual's blood.

- ***bmi***: continuous variable indicating the body mass index of the individual.

- ***smoking_status***: categorical variable indicating the smoking status of the individual (formerly smoked, never smoked, smokes).

- ***stroke***: binary variable indicating whether the individual had a stroke or not (1 for yes, 0 for no).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29065 entries, 0 to 29064
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   gender             29062 non-null  object 
 1   age                29031 non-null  float64
 2   hypertension       29063 non-null  float64
 3   heart_disease      29057 non-null  float64
 4   ever_married       29053 non-null  object 
 5   work_type          29051 non-null  object 
 6   residence_type     29053 non-null  object 
 7   avg_glucose_level  29021 non-null  float64
 8   bmi                29040 non-null  float64
 9   smoking_status     29048 non-null  object 
 10  stroke             29063 non-null  float64
dtypes: float64(6), object(5)
memory usage: 2.4+ MB
```

2. DATASET DESCRIPTION

Have a quick look at the dataset!

	gender	age	hypertension	heart_disease	ever_married	work_type	residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	58.0	1.0	0.0	Yes	Private	Urban	87.96	39.2	never smoked	0.0
1	Female	70.0	0.0	0.0	Yes	Private	Rural	69.04	35.9	formerly smoked	0.0
2	Female	52.0	0.0	0.0	Yes	Private	Urban	77.59	17.7	formerly smoked	0.0
3	Female	75.0	0.0	1.0	Yes	Self-employed	Rural	243.53	27.0	never smoked	0.0
4	Female	32.0	0.0	0.0	Yes	Private	Rural	77.67	32.3	smokes	0.0
...	...	...	...	...	...	...	...	...	...	...	...
29060	Female	10.0	0.0	0.0	No	children	Urban	58.64	20.4	never smoked	0.0
29061	Female	56.0	0.0	0.0	Yes	Govt_job	Urban	213.61	55.4	formerly smoked	0.0
29062	Female	82.0	1.0	0.0	Yes	Private	Urban	91.94	28.9	formerly smoked	0.0
29063	Male	40.0	0.0	0.0	Yes	Private	Urban	99.16	33.2	never smoked	0.0
29064	Female	82.0	0.0	0.0	Yes	Private	Urban	79.48	20.6	never smoked	0.0

3. DATA CLEANING

- **No irrelevant data** was detected.
- **No duplicate entries** were identified.
- The missing values don't show any clear patterns, which means they are randomly distributed **(MCAR)**. This suggests that the missing values can be ignored or treated as insignificant.

  → *Missing values will be deleted.*

	Missing Values	Percentage
gender	3	0.010322
age	34	0.116979
hypertension	2	0.006881
heart_disease	8	0.027525
ever_married	12	0.041287
work_type	14	0.048168
residence_type	12	0.041287
avg_glucose_level	44	0.151385
bmi	25	0.086014
smoking_status	17	0.058490
stroke	2	0.006881

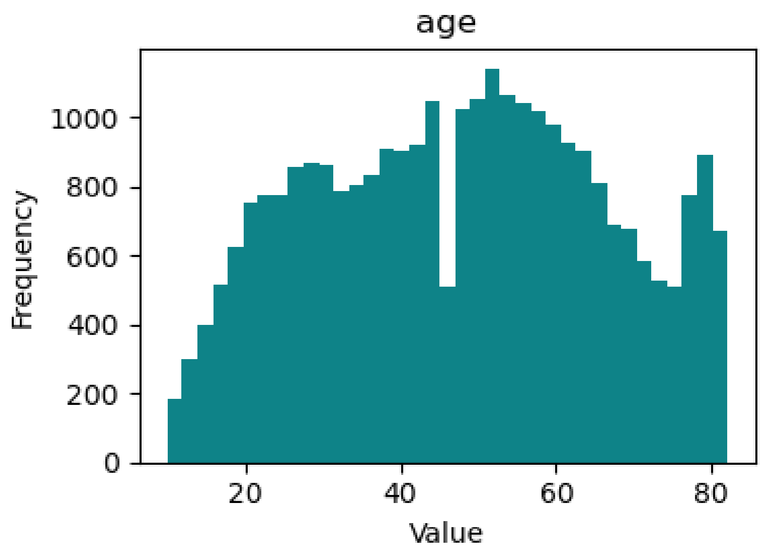
4.
CONTINUOUS VARIABLES

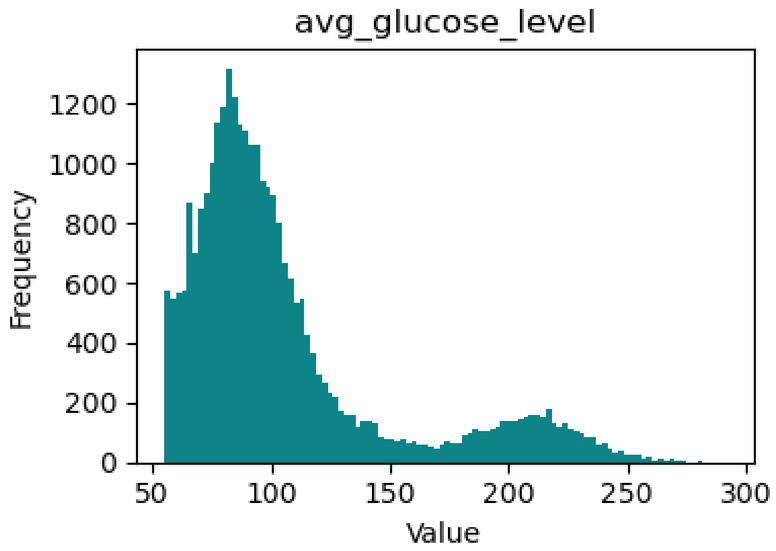
age, avg_glucose_level, bmi

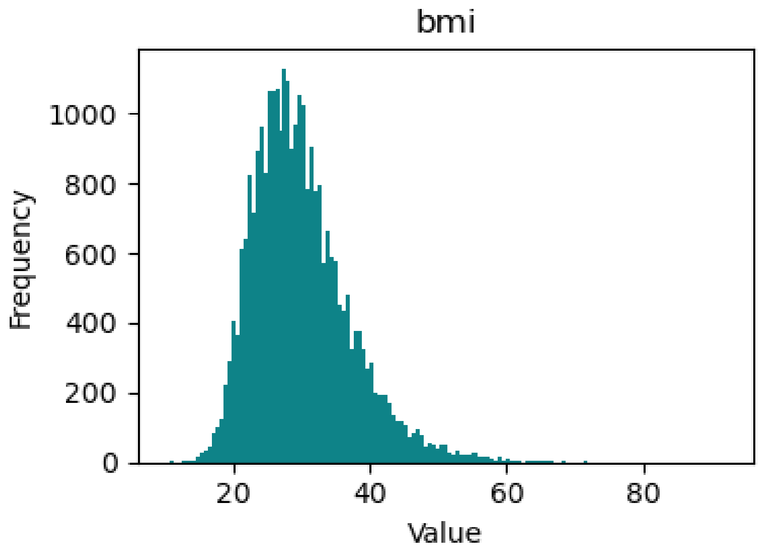
DESCRIPTIVE ANALYSIS

THE DIFFERENCES IN MEAN, MODE, AND MEDIAN SUGGEST NON-NORMAL DISTRIBUTIONS.

	count	mean	median	mode	range	min	max	std	var
age	28916	47.668281	48.0	51.0	72.0	10.0	82.0	18.732721	350.914824
avg_glucose_level	28916	106.388265	92.12	87.15	236.04	55.01	291.05	45.266083	2049.018249
bmi	28916	30.049658	28.9	26.8, 27.6	81.9	10.1	92.0	7.190541	51.703887





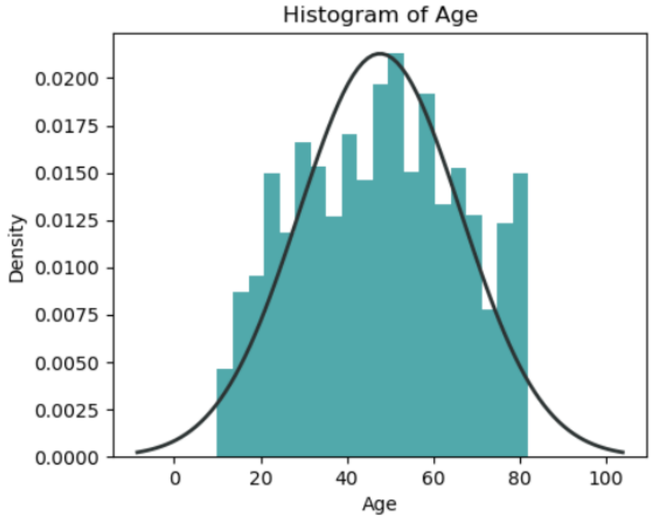


NORMALITY TESTS

THE DIFFERENCES IN MEAN, MODE, AND MEDIAN SUGGEST NON-NORMAL DISTRIBUTIONS.

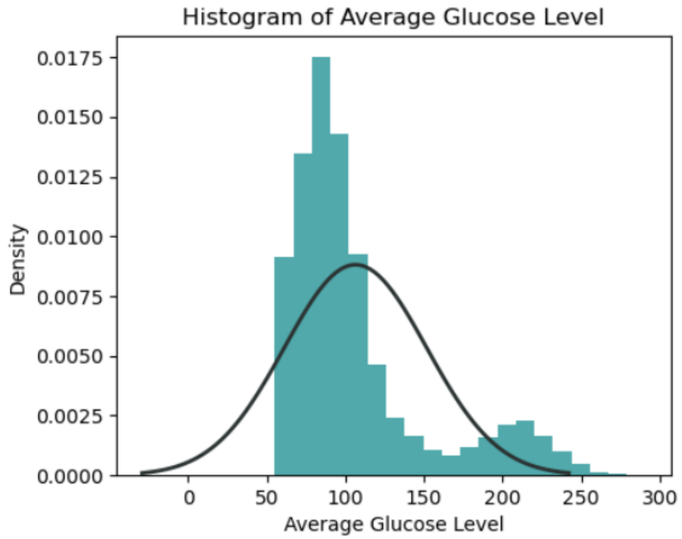
age

No. of values	28916
Mean	47.66828
Mode	51
Median	48
Skewness	-0.00276
Std. Error of Skewness	-1.623791
Kurtosis	-0.96793
Std. Error of Kurtosis	-0.023334



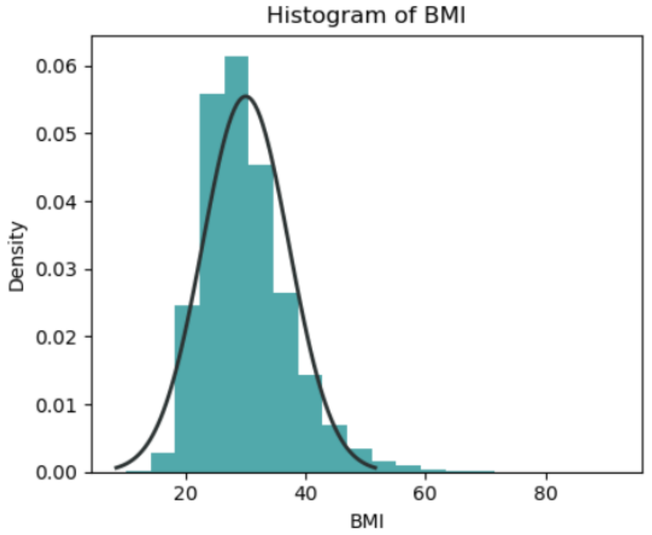
avg_glucose_level

No. of values	28916
Mean	106.388264
Mode	87.15
Median	92.12
Skewness	1.570659
Std. Error of Skewness	0.009236
Kurtosis	1.653999
Std. Error of Kurtosis	-0.007915



bmi

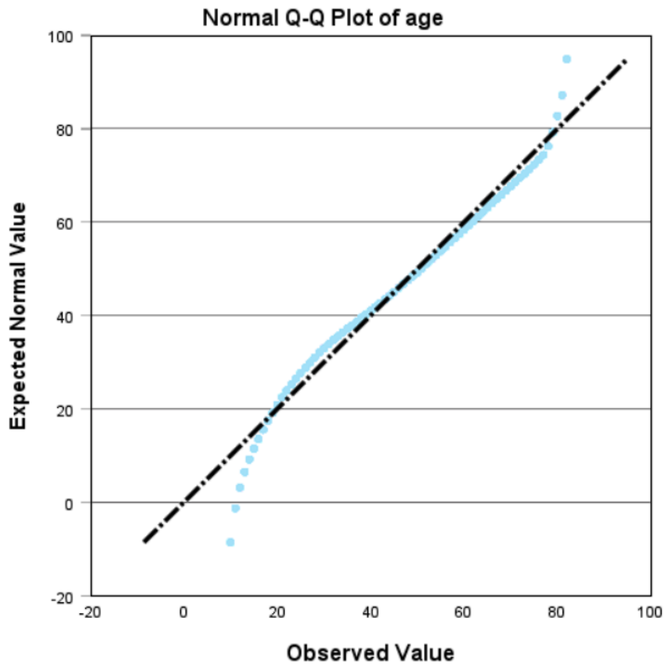
No. of values	28916
Mean	30.049657
Mode	51
Median	26.8, 27.6
Skewness	1.072204
Std. Error of Skewness	0.006305
Kurtosis	2.228752
Std. Error of Kurtosis	-0.004535



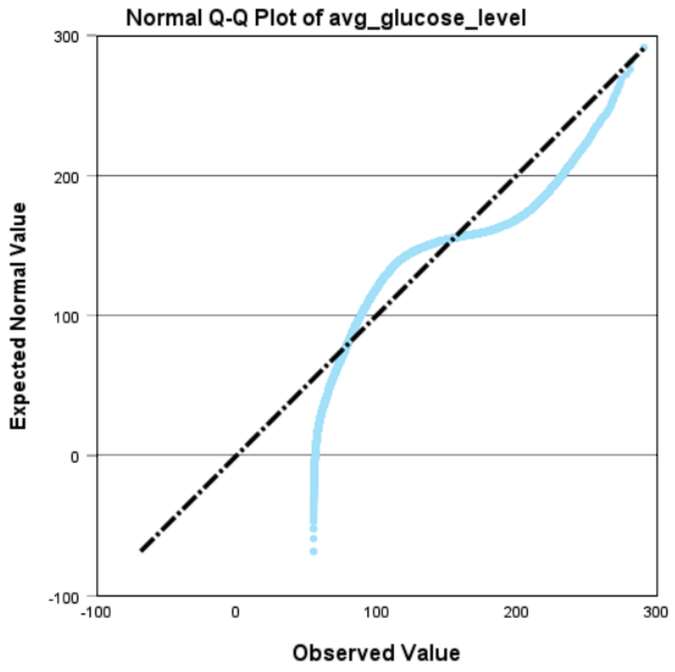
NORMALITY TESTS

Q-Q PLOTS SUGGEST NON-NORMAL DISTRIBUTIONS.

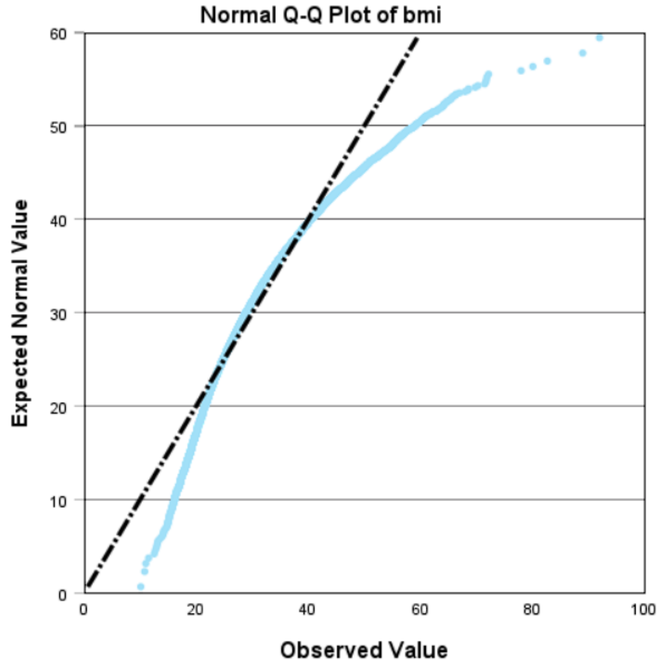
age



avg_glucose_level



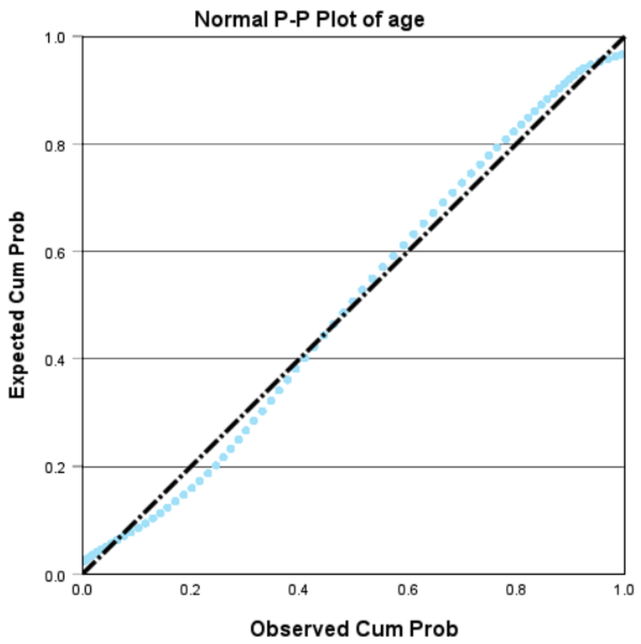
bmi



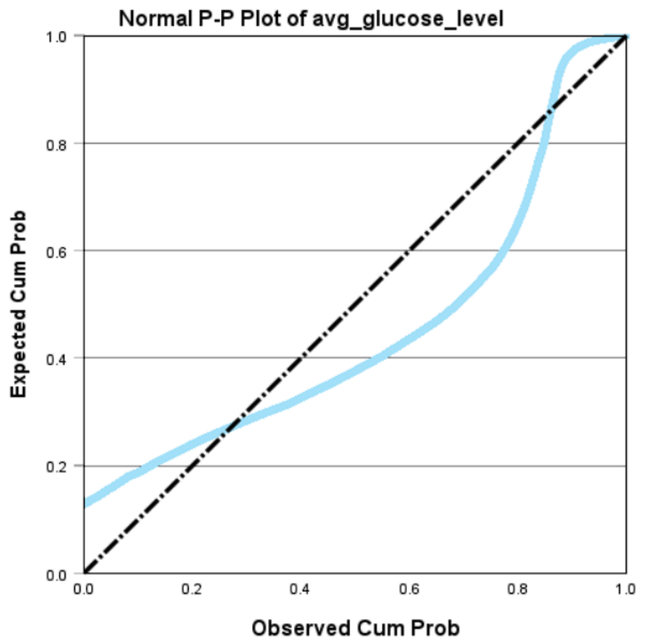
NORMALITY TESTS

P-P PLOTS SUGGEST NON-NORMAL DISTRIBUTIONS.

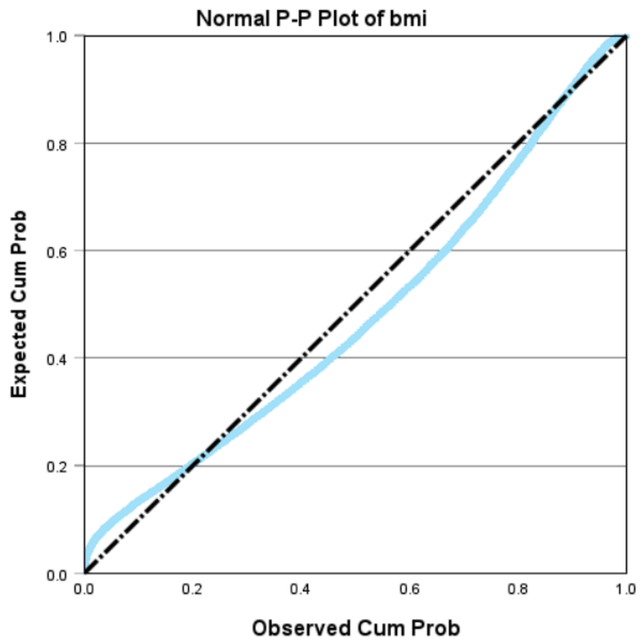
age



avg_glucose_level



bmi



These P-P plots are drawn on SPSS.

NORMALITY TESTS

KOLMOGOROV-SMIRNOV TESTS SUGGEST NON-NORMAL DISTRIBUTIONS.

Given the dataset's large size of 28,916 rows, the **Kolmogorov-Smirnov test** is preferable over the Shapiro-Wilk test for checking normality.

Hypothesis:

H0: The sample comes from a population with a normal distribution

H1: The sample does not come from a population with a normal distribution

Variable	age	avg_glucose_level	bmi
Statistics	1.0	1.0	1.0
p-value	0.0	0.0	0.0
Statistic Location	10.0	55.01	10.1
Statistic Sign	-1	-1	-1

↓

age:

p_value < 0.05

→ Reject H0, Accept H1

→ *'age'* is not normally distributed

↓

avg_glucose_level:

p_value < 0.05

→ Reject H0, Accept H1

→ *'avg_glucose_level'* is not normally distributed

↓

bmi:

p_value < 0.05

→ Reject H0, Accept H1

→ *'bmi'* is not normally distributed

POWER & EFFECT SIZE *(MANN-WHITNEY U TEST)*

AGE AND AVERAGE GLUCOSE LEVELS DIFFERENCES ARE LIKELY DETECTABLE, WHILE BMI DIFFERENCES ARE UNLIKELY.

Moving forward, variables were divided into ***stroke (1)*** and ***non-stroke (0)*** groups to investigate differences using non-parametric tests. Due to non-normal distribution of continuous variables and independent groups, the **Mann-Whitney U test** was the suitable choice.

‘age’ *Effect size:* 0.17841242497426993 *Power:* 1.0	→ **A substantial age difference** between stroke and non-stroke individuals is **highly probable.**
‘avg_glucose_level’ *Effect size:* 0.39959474393445016 *Power:* 1.0	→ **A noticeable difference in average glucose levels** between stroke and non-stroke individuals is **highly probable.**
‘bmi’ *Effect size:* 0.4986453336330313 *Power:* 1.0	→ **A negligible bmi difference** between stroke and non-stroke individuals is **highly unlikely.**

HYPOTHESIS TESTING *(MANN-WHITNEY U TEST)*

AGE IS A SIGNIFICANT PREDICTOR OF STROKE RISK.

Hypothesis:

H0: The age distributions of individuals who have had a stroke and those who have not are not significantly different.

H1: The age distributions of individuals who have had a stroke and those who have not are significantly different.

'age' by 'stroke' U value: 12749267 p_value: 0	***p_value < 0.05*** → Reject H0, Accept H1 → The age distribution disparity between stroke and non-stroke groups is **statistically significant.**

→ Age is a significant predictor of stroke risk.

HYPOTHESIS TESTING *(MANN-WHITNEY U TEST)*

AVERAGE GLUCOSE LEVEL IS A SIGNIFICANT PREDICTOR OF STROKE RISK.

Hypothesis:

H0: *The average glucose level distributions of individuals who have had a stroke and those who have not are not significantly different.*

H1: *The average glucose level distributions of individuals who have had a stroke and those who have not are significantly different.*

'avg_glucose_level' by 'stroke' U value: 9316994.5 p_value: 0	**p_value < 0.05** → Reject H0, Accept H1 → The average glucose level distribution disparity between stroke and non-stroke groups is **statistically significant.**

→ Average glucose level is a significant predictor of stroke risk.

HYPOTHESIS TESTING *(MANN-WHITNEY U TEST)*

BMI IS NOT A SIGNIFICANT PREDICTOR OF STROKE RISK.

Hypothesis:

H0: The BMI distributions of individuals who have had a stroke and those who have not are not significantly different.

H1: The BMI distributions of individuals who have had a stroke and those who have not are significantly different.

'bmi' by 'stroke' U value: 7779943 p_value: 0.91	**p_value = 0.91 > 0.05** → Reject H1, Accept H0 → The BMI distribution disparity between stroke and non-stroke groups is **not statistically significant.**

→ BMI is not a significant predictor of stroke risk.

CORRELATION ANALYSIS - ‘AGE’ & ‘STROKE’

‘AGE’ AND ‘STROKE’ ARE SIGNIFICANTLY CORRELATED.

For calculating the correlation between the continuous variables (‘age’, ‘avg_glucose_level’ & ‘bmi’) and the categorical variable ‘stroke’, the ***point-biserial correlation coefficient*** is used.

Hypothesis:

H0: There is no significant correlation between *‘age’* and *‘stroke’*

H1: There is a significant correlation between the *‘age’* and *‘stroke’*

Correlation coefficient: 0.15426411087630276

p_value: 1.826429907943924e-153

p_value < 0.05

→ Reject H0 → Accept H1

→ There is a significant correlation between the ‘age’ and ‘stroke’.

CORRELATION ANALYSIS - ‘AGE’ & ‘STROKE’

‘AVG_GLUCOSE_LEVEL’ AND ‘STROKE’ ARE SIGNIFICANTLY CORRELATED.

Hypothesis:

H0: There is no significant correlation between ‘*avg_glucose_level*’ and ‘*stroke*’

H1: There is a significant correlation between ‘*avg_glucose_level*’ and ‘*stroke*’

Correlation coefficient: 0.07525378841310744

p_value: 1.3635622454206563e-37

p_value < 0.05

→ Reject H0 → Accept H1

→ There is a significant correlation between the ‘avg_glucose_level’ and ‘stroke’.

CORRELATION ANALYSIS - ‘AGE’ & ‘STROKE’

‘BMI’ AND ‘STROKE’ ARE NOT SIGNIFICANTLY CORRELATED.

Hypothesis:

H0: There is not significant correlation between *‘bmi’* and *‘stroke’*

H1: There is a significant correlation between *‘bmi’* and *‘stroke’*

Correlation coefficient: -0.004175207957119117

p_value: 0.47773253332879284

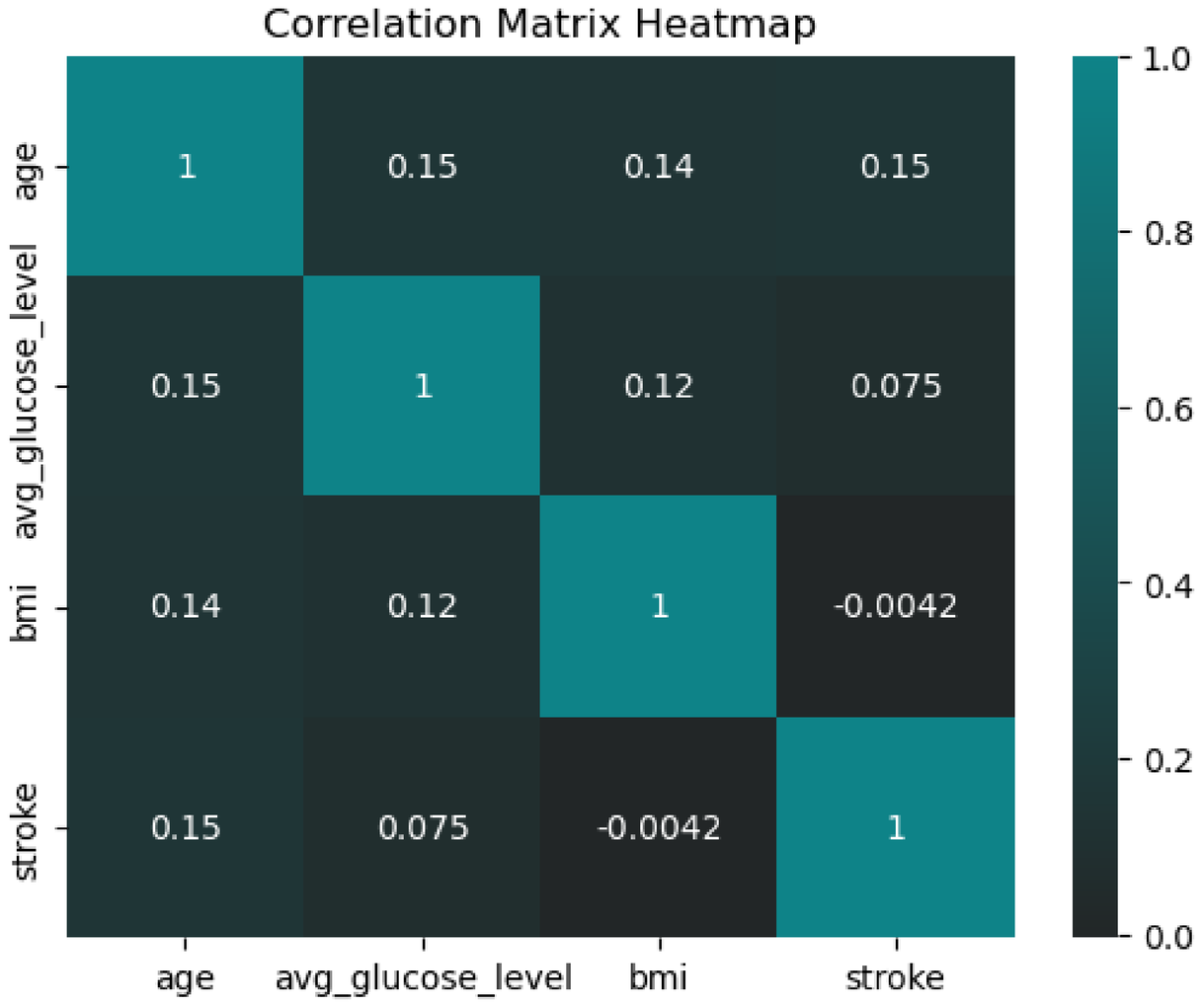
p_value = 0.48 > 0.05

→ Reject H1 → Accept H0

→ There is no significant correlation between the ‘bmi’ and ‘stroke’.

CORRELATION HEATMAP

AGE HAS THE MOST SIGNIFICANT CORRELATION WITH STROKE, WHILE BMI DOES NOT.

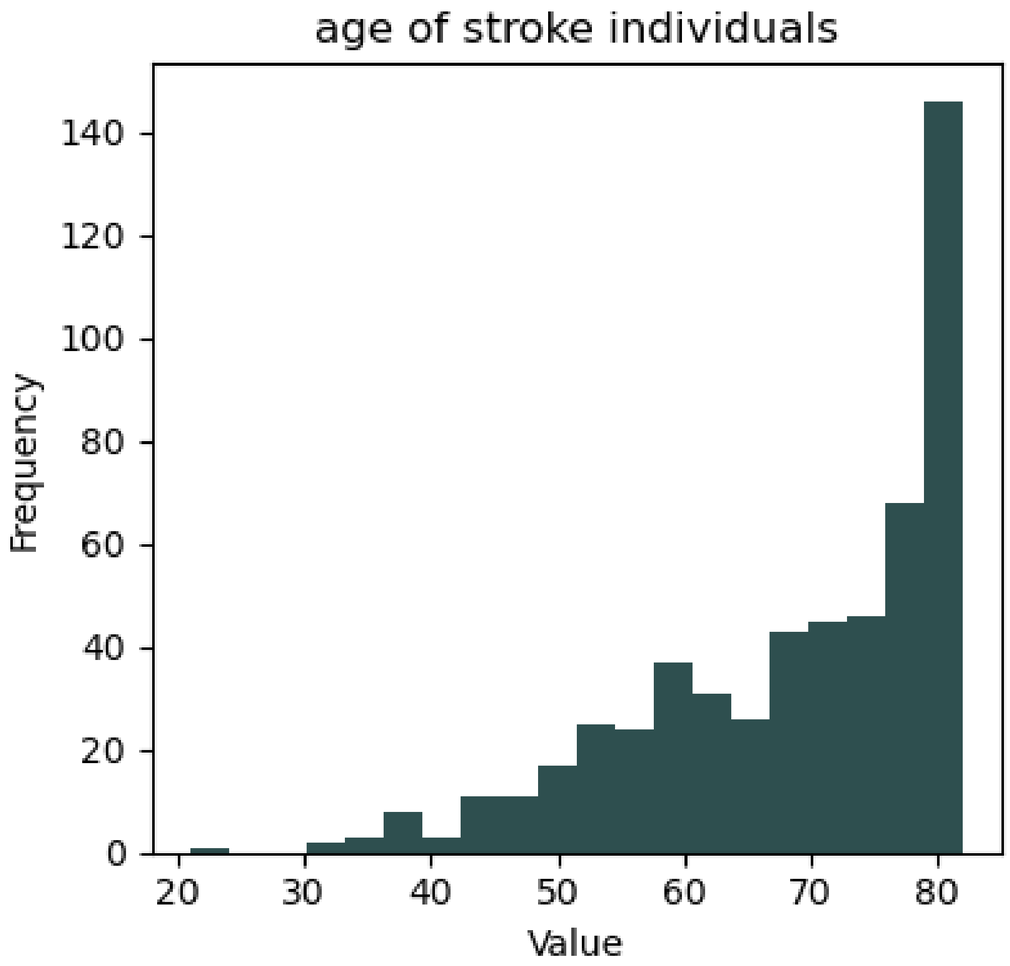


Notes:

- **Point-biserial correlation** measures the correlation between 'stroke' (binary variable) and 'age', 'avg_glucose_level', and 'bmi' (continuous variables).
- **Spearman correlation** assesses the correlation between 'age', 'avg_glucose_level', and 'bmi' (continuous variables).
- 'bmi' and 'stroke' show no statistically significant correlation.

AGE * STROKE - IMPLICATIONS

THE LIKELIHOOD OF STROKE RISES STARTING AT THE AGE OF 50 AND INTENSIFIES REMARKABLY AFTER REACHING 80 YEARS OLD.



The graph presented focuses on the **age distribution of individuals who have previously had a stroke** *(stroke = 1)*, building upon the established disparities in age distribution between stroke and non-stroke individuals, and the observed correlation between age and the occurrence of strokes.

Implications:

- Stroke risk tends to **increase with age.**

After the age of

50+

stroke risk increases

notably

After the age of

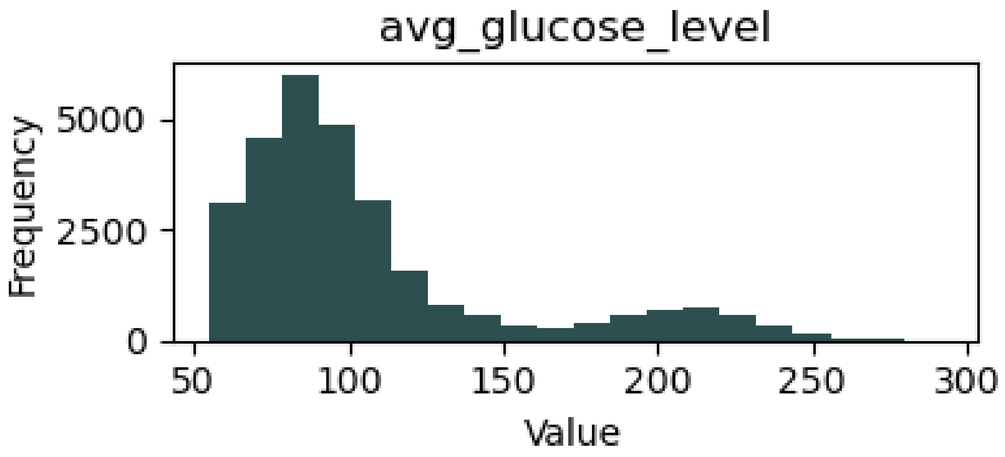
80+

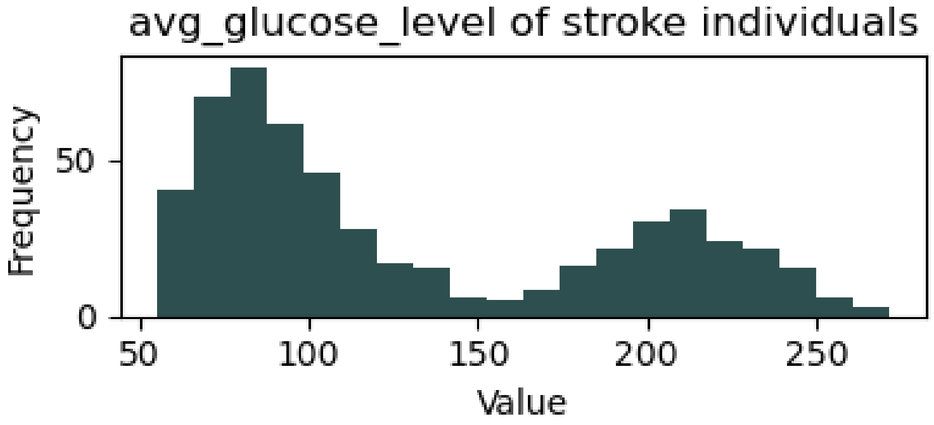
stroke risk escalates

exceptionally

AVG_GLUCOSE_LEVEL * STROKE - IMPLICATIONS

INDIVIDUALS WITH AVERAGE GLUCOSE LEVEL EXCEEDING 175MG/DL FACE A HIGHER STROKE RISK.





The graph above depicts the **average glucose levels distribution for the entire dataset**, while the graph below focuses on **average glucose levels distribution among individuals who have had a stroke.** Both graphs exhibit a *similar pattern within the 70-100 mg/dL range*, indicating a **possible dataset bias** towards that range in terms of the number of values.

Implications:

- The stroke risk tends to increase when the average glucose level is **outside the range of 110 mg/dL to 175 mg/dL.**

- Given the biasedness, it becomes evident that individuals with an **average glucose level exceeding 175 mg/dL face an elevated stroke risk.**

Z-SCORE ANALYSIS

PATIENTS >45 YEARS OLD AND/OR WITH AVERAGE GLUCOSE LEVEL <80.43 MG/DL SHOULD BE MORE CONCERNED ABOUT STROKE.

The z-score analysis is performed on the dataframe df1, which is a subset obtained from the original dataframe by querying for rows where stroke equals 1.

age_z_score: 94.88% of data points have an age greater than 45 (z-score = -1.9815831121459675).

→ Precautions should be taken for patients aged 45 years or older.

glucose_z_score: 35.65% of data points have an average glucose level smaller than 88.29 (z-score = -0.6981781291938679).

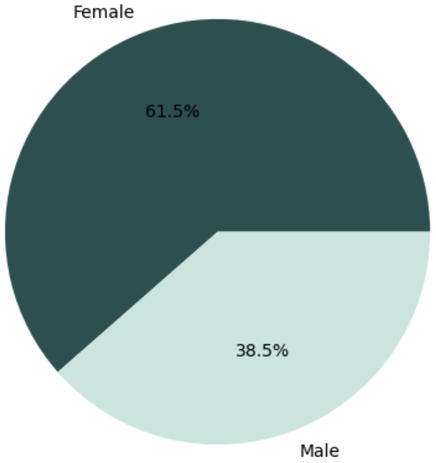
→ Precautions should be taken for patients whose average glucose level smaller than 88.29 mg/dL.

5.
CATEGORICAL VARIABLES

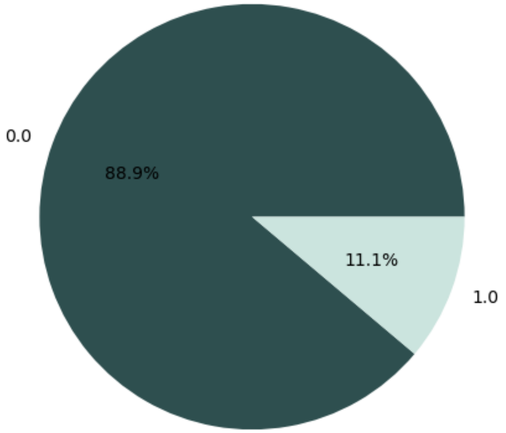
gender, hypertension, heart_disease, ever_married, work_type, residence_type, smoking_status

DESCCRIPTIVE ANALYSIS

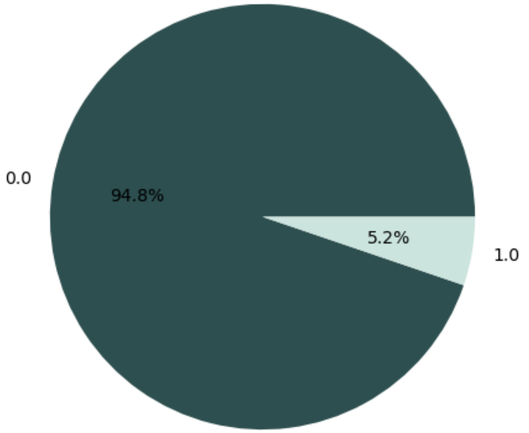
Gender



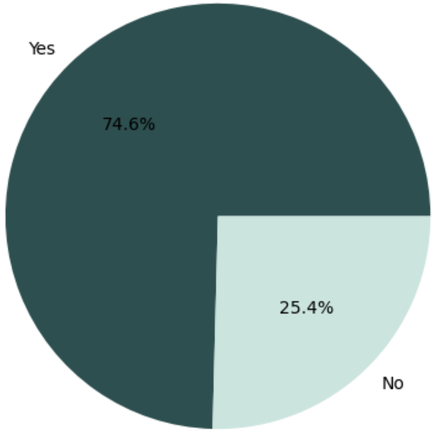
Hypertension



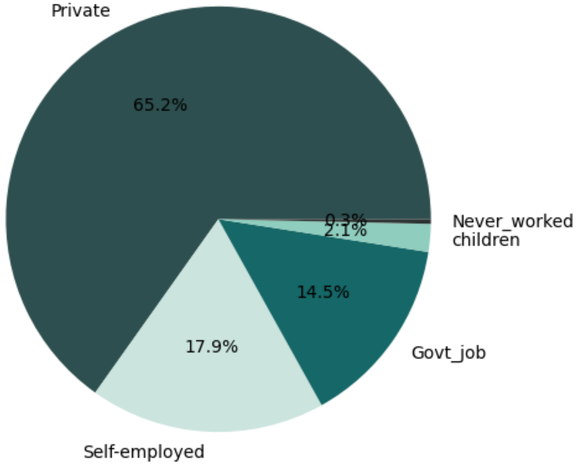
Heart Disease



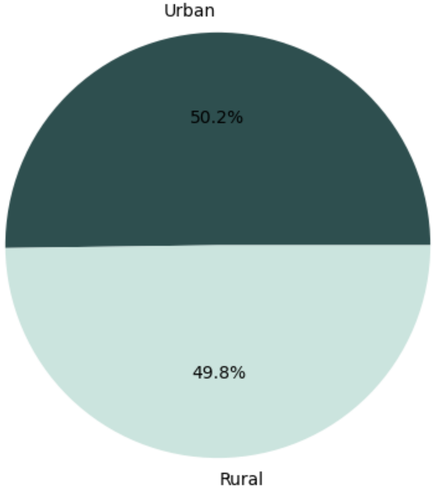
Ever Married



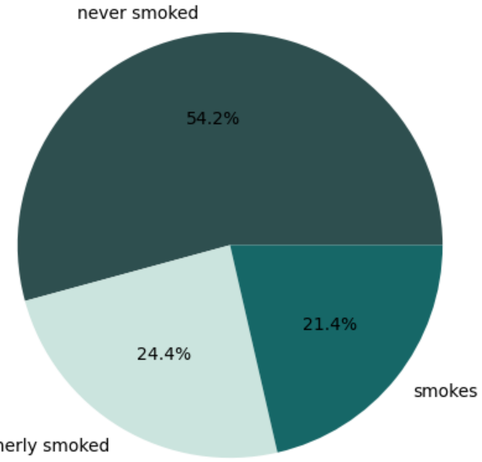
Work Type



Residence Type



Smoking Status



CHI-SQUARE HYPOTHESIS TESTING

THE VARIABLES GENDER, HYPERTENSION, HEART DISEASE, AND MARITAL STATUS ARE ALL FOUND TO HAVE A SIGNIFICANT CORRELATION WITH STROKE.

Hypothesis:

H0: There is no statistically significant correlation between the observed variable and stroke.

H1: There is a statistically significant correlation between the observed variable and stroke.

Variable	Chi-square	df	p-value	Conclusion
gender	4.4080	1	0.0358	p_value = 0.0358 < 0.05 → Reject H0 → Accept H1 → There is a **significant correlation** between 'gender' and 'stroke'.
hypertension	179.0933	1	0.0000	p_value < 0.05 → Reject H0 → Accept H1 → There is a **significant correlation** between 'hypertension' and 'stroke'.
heart_disease	319.2124	1	0.0000	p_value < 0.05 → Reject H0 → Accept H1 → There is a **significant correlation** between 'heart disease' and 'stroke'.
ever_married	65.2294	1	0.0000	p_value < 0.05 → Reject H0 → Accept H1 → There is a **significant correlation** between 'ever-married' and 'stroke'.

CHI-SQUARE HYPOTHESIS TESTING

THE VARIABLES WORK TYPE AND SMOKING STATUS EXHIBIT A SIGNIFICANT CORRELATION WITH STROKE, WHEREAS RESIDENCE TYPE DOES NOT SHOW A SIGNIFICANT CORRELATION.

Hypothesis:

H0: There is no statistically significant correlation between the observed variable and stroke.

H1: There is a statistically significant correlation between the observed variable and stroke.

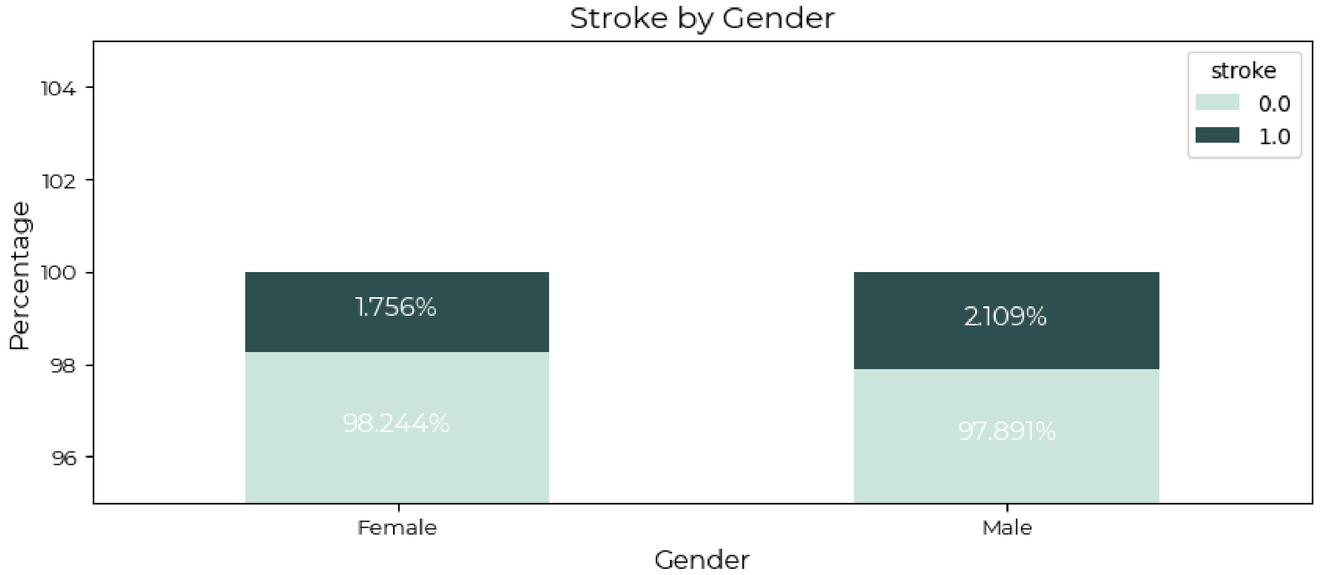
Variable	Chi-square	df	p-value	Conclusion
work_type	75.8109	4	0.0000	p_value < 0.05 → Reject H0 → Accept H1 → There is a **significant correlation** between 'work-type' and 'stroke'.
residence_type	0.0596	1	0.8072	p_value = 0.8072 > 0.05 → Accept H0 → There is **no significant correlation** between residence_type and stroke.
smoking_status	21.9462	2	0.0000	p_value < 0.05 → Reject H0 → Accept H1 → There is a **significant correlation** between 'smoking_status' and 'stroke'.

→ gender, hyper_tension, heart_disease, ever_married, work_type and smoking_status all have significant correlation with 'stroke', while only residence_type exhibits no such correlation.

CROSS-TABULATION: GENDER * STROKE

MEN APPEAR TO HAVE A HIGHER SUSCEPTIBILITY TO STROKE COMPARED TO WOMEN.

gender \ stroke	0	1
female	98.244%	1.756%
male	97.891%	2.109%
Total	98.108%	1.892%



1.8% of ***female*** had a stroke.

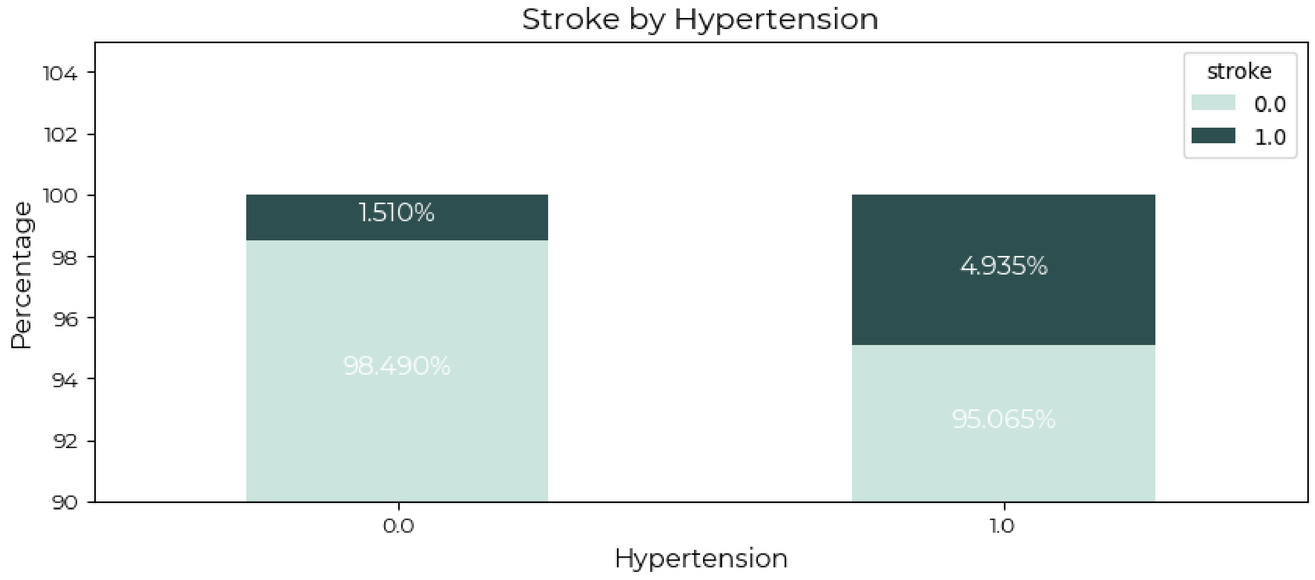
2.1% of ***male*** had a stroke.

→ **Men** appear to have a **higher susceptibility** to stroke compared to women.

CROSS-TABULATION: HYPERTENSION * STROKE

HAVING HYPERTENSION IS ASSOCIATED WITH AN INCREASED RISK OF STROKE COMPARED TO INDIVIDUALS WITHOUT HYPERTENSION.

hypertension \ stroke	0	1
0	98.490%	1.510%
1	95.065%	4.935%
Total	98.108%	1.892%



1.5% of people ***without hypertension*** had a stroke.

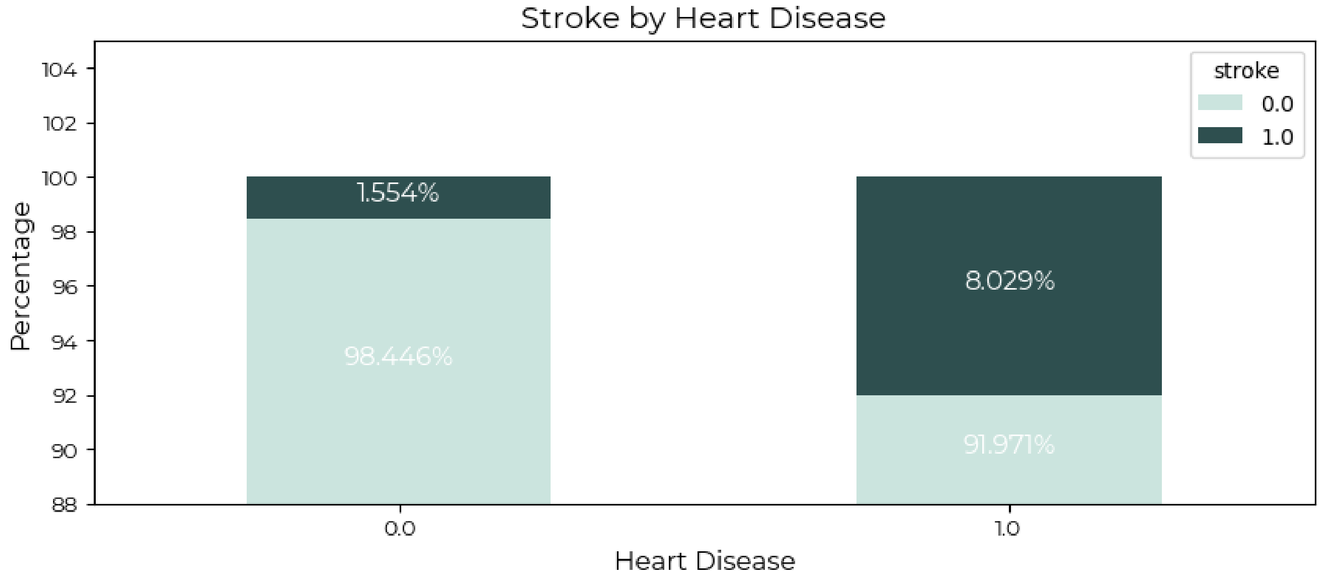
4.9% of people ***with hypertension*** had a stroke.

→ **Individuals with hypertension** have a **higher stroke risk** compared to those without hypertension.

CROSS-TABULATION: HEART_DISEASE * STROKE

HEART DISEASE IS STRONGLY ASSOCIATED WITH A SIGNIFICANTLY HIGHER RISK OF STROKE.

heart_disease \ stroke	0	1
0	98.446%	1.554%
1	91.971%	8.029%
Total	98.108%	1.892%



1.6% of people ***without heart diseases*** had a stroke.

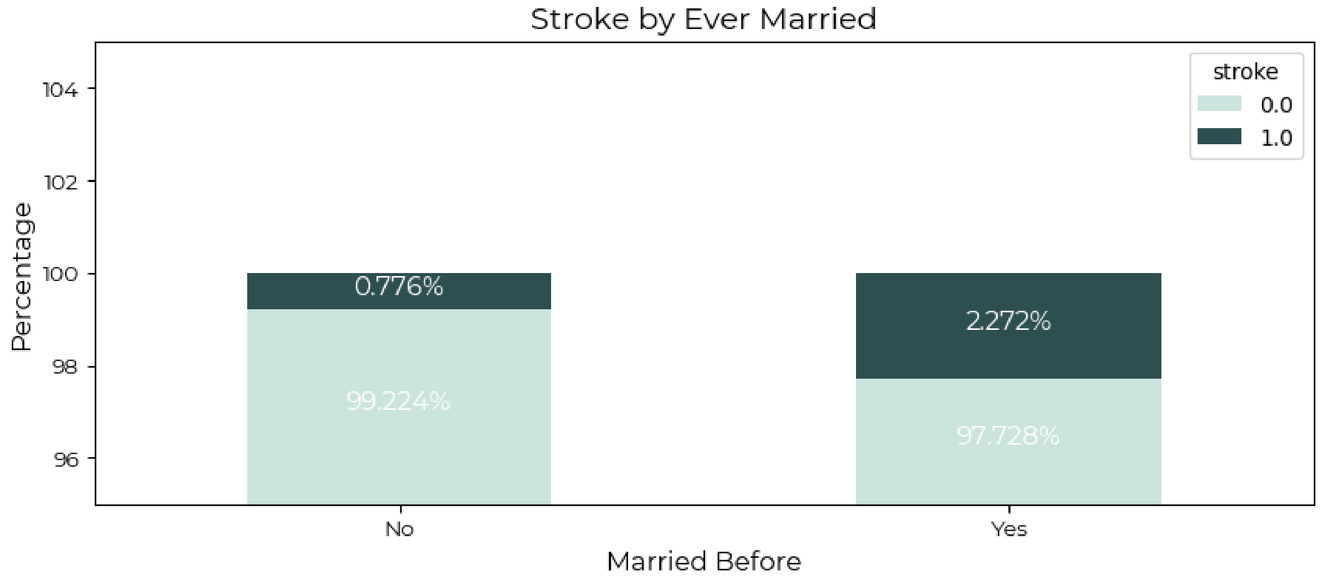
8% of people ***with a heart disease*** had a stroke.

→ **Individuals with a heart disease** have a **significantly higher** risk of stroke compared to those without any heart diseases.

CROSS-TABULATION: EVER_MARRIED * STROKE

MARRIED INDIVIDUALS ARE AT A GREATER RISK OF STROKE COMPARED TO THOSE WHO ARE SINGLE.

stroke / ever_married	0	1
No	99.224%	0.776%
Yes	97.728%	2.272%
Total	98.108%	1.892%



0.8% of single individuals had a stroke.

2.3% of married individuals had a stroke.

→ **Married individuals** are at a **greater risk of stroke** compared to those who are single.

CROSS-TABULATION: WORK_TYPE * STROKE

INDIVIDUALS WHO HAVE NEVER WORKED BEFORE DO NOT HAVE A RISK OF STROKE, WHEREAS SELF-EMPLOYMENT CARRIES THE HIGHEST RISK OF STROKE.

stroke / work_type	0	1
children	100%	0%
Never_worked	100%	0%
Govt_job	98.421%	1.579%
Private	98.350%	1.650%
Self-employed	96.718%	3.282%
Total	98.108%	1.892%

No children had had a stroke.

No individuals who had never worked had a stroke.

1.6% of government workers had a stroke.

1.7% of private sector workers had a stroke.

3.3% of self-employed individuals had a stroke.

→ Individuals who have **never worked** before **do not have a risk** of stroke, whereas **self-employment carries the highest risk of stroke.**

CROSS-TABULATION: SMOKING_STATUS * STROKE

SMOKING HAS A MINIMAL IMPACT ON THE RISK OF STROKE.

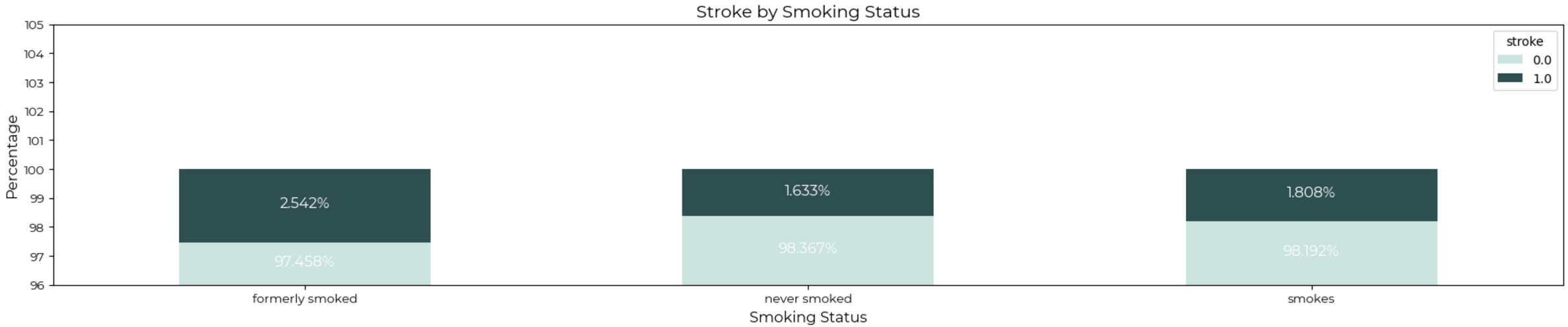
smoking_status \ stroke	0	1
formerly smoked	97.458%	2.542%
never smoked	98.367%	1.633%
smokes	98.192%	1.808%
Total	98.108%	1.892%

1.6% of individuals who never smoked had a stroke.

1.8% of frequent smokers had a stroke

2.5% of former smokers had a stroke.

→ The risk of stroke is **slightly elevated** in **smokers and former smokers** compared to non-smokers, and this association is **further influenced by another variable.**



(EVER_MARRIED * WORK_TYPE) * STROKE

NO SIGNIFICANT ASSOCIATION IS OBSERVED BETWEEN MARITAL STATUS, WORK TYPE, AND STROKE RISK.

stroke / ever_married * work_type	0	1
(No, Govt_job)	99.224%	0.776%
(No, Never_worked)	100%	0%
(No, Private)	99.402%	0.598%
(No, Self-employed)	97.059%	2.941%
(No, children)	100%	0%
(Yes, Govt_job)	98.238%	1.762%
(Yes, Never_worked)	100%	0%
(Yes, Private)	97.950%	2.050%
(Yes, Self-employed)	96.667%	3.333%
Total	98.108%	1.892%

p-value = 1

→ Implication:

- **Working individuals who are married** have a slightly higher risk of stroke.

- However, it is important to note that this observation is **not statistically significant.**

HEALTH_ISSUE_COUNT * STROKE

THE GREATER NUMBER OF HEALTH ISSUES A PERSON EXPERIENCES, THE HIGHER THE LIKELIHOOD OF HAVING A STROKE.

stroke / health_issue_count	0	1
0	98.871%	1.129%
1	98.637%	1.363%
2	97.271%	2.729%
3	95.261%	4.739%
4	89.796%	10.204%
Total	95.967%	4.033%

A new column named **'health_issue_count'** was created to determine the cumulative number of health issues experienced by each individual.

These health issues encompass **hypertension**, **heart diseases**, **low glucose levels** *(below 70)*, **high glucose levels** *(above 100)*, **very high glucose levels** *(above 150)*, **obesity** *(BMI greater than 30)*, and **super obesity** *(BMI greater than 35)*.

→ Implication:

More health issues are associated with a higher likelihood of stroke, and as the number of health issues increases, the gap in stroke risk widens.

HEALTH_ISSUE_COUNT * STROKE

THIS RELATIONSHIP IS STATISTICALLY SIGNIFICANT.

Hypothesis:

H0: There is not significant correlation between health_issue_count and stroke.

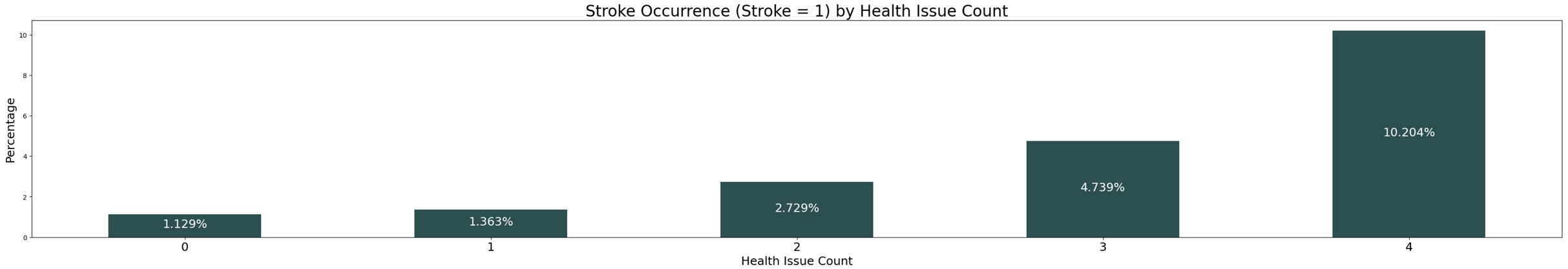
H1: There is a significant correlation between health_issue_count and stroke.

Chi-square statistic	14.428780990059204
df	5
p-value	0.013103342274608626

p_value < 0.05

→ Reject H0 → Accept H1

→ There is a **significant correlation** between ‘health_issue_count’ and ‘stroke’.



(SMOKING_STATUS * HEALTH_ISSUE_COUNT) * STROKE

AS HEALTH ISSUES INCREASE, PRIOR AND FREQUENT SMOKERS HAVE A SIGNIFICANTLY HIGHER STROKE RISK THAN NON-SMOKERS.

	stroke	0	1
smoking_status	**health_issue_count**	**0**	**1**
formerly smoked	0	98.32	1.68
	1	98.06	1.94
	2	96.93	3.07
	3	94.67	5.33
	4	89.66	10.34
never smoked	0	99.09	0.91
	1	98.78	1.22
	2	97.42	2.58
	3	95.89	4.11
	4	90.91	9.09
smokes	0	98.81	1.19
	1	98.87	1.13
	2	97.36	2.64
	3	94.78	5.22
	4	88.24	11.76

Chi-square statistic	44.53253222858696
df	14
p-value	4.8566060379867334e-05

→ Implication:

- Having **two or more health issues increases the likelihood of stroke** compared to individuals who have never smoked, particularly among those who had a history of smoking or were frequent smokers.

- The **magnitude** of this difference becomes **more pronounced** with an **increasing number of health issues.**

- This finding is **statistically significant.**

(WORK_TYPE * HEALTH_ISSUE_COUNT) * STROKE

AS HEALTH ISSUES INCREASE, STROKE RISK IS HIGHER FOR GOVERNMENT AND PRIVATE SECTOR WORKERS COMPARED TO THE SELF-EMPLOYED.

	stroke	0	1
work_type	health_issue_count	0	1
Govt_job	0	99.37	0.63
	1	98.90	1.10
	2	97.59	2.41
	3	95.87	4.13
	4	88.24	11.76
Private	0	99.03	0.97
	1	98.90	1.10
	2	97.47	2.53
	3	95.59	4.41
	4	86.81	13.19
Self-employed	0	97.42	2.58
	1	97.16	2.84
	2	96.30	3.70
	3	94.22	5.78
	4	97.44	2.56

Children and never-worked individuals are excluded from display as they had no strokes in the dataset.

Chi-square statistic	102.73258742189476
df	21
p-value	9.453664471442912e-13

→ Implication:

- Individuals with **more health issues** working in **government** or **private sectors** have a **higher risk of stroke.**

-This finding is **statistically significant.**

6.
LOGISTIC REGRESSION

6. LOGISTIC REGRESSION

INDIVIDUALLY, AGE, AVERAGE GLUCOSE LEVEL, HYPERTENSION, HEART DISEASES, AND FREQUENT SMOKING CAN INDEPENDENTLY INFLUENCE THE RISK OF STROKE.

Logistic Regression Result

	coef	std err	z	P>|z|	[0.025	0.975]
const	-8.5709	0.396	-21.657	0.000	-9.347	-7.795
age	0.0725	0.004	18.971	0.000	0.065	0.080
hypertension	0.4445	0.101	4.417	0.000	0.247	0.642
heart_disease	0.6150	0.114	5.417	0.000	0.393	0.838
avg_glucose_level	0.0038	0.001	4.909	0.000	0.002	0.005
bmi	-0.0084	0.007	-1.180	0.238	-0.022	0.006
gender_Male	0.0248	0.092	0.271	0.787	-0.155	0.205
ever_married_Yes	-0.1578	0.146	-1.079	0.280	-0.444	0.129
work_type_Never_worked	-5.3873	38.432	-0.140	0.889	-80.712	69.938
work_type_Private	0.0574	0.139	0.412	0.680	-0.216	0.330
work_type_Self-employed	-0.0360	0.152	-0.237	0.813	-0.334	0.262
work_type_children	-19.8231	3.38e+04	-0.001	1.000	-6.62e+04	6.61e+04
residence_type_Urban	0.0191	0.088	0.216	0.829	-0.154	0.192
smoking_status_never smoked	0.0254	0.103	0.248	0.804	-0.176	0.227
smoking_status_smokes	0.2680	0.127	2.110	0.035	0.019	0.517

Variables that exhibit a p-value less than 0.05 indicate a significant influence on the dependent variable 'stroke'.

Implication:

Individually, factors such as *age*, *average glucose level*, *hypertension*, *heart diseases*, and *frequent smoking* can each independently influence the risk of stroke.

6. LOGISTIC REGRESSION

THERE ARE SUPERIOR PREDICTORS OF STROKE RISK OUTSIDE OF THE DATASET.

Logistic Regression Results			
Dep. Variable:	stroke	No. Observations:	28916
Model:	Logit	Df Residuals:	28901
Method:	MLE	Df Model:	14
converged:	False	Pseudo R-squ.:	0.1607
Coveriance Type:	nonrobust	Log-Likelihood:	-2276.2
		LL-Null:	-2712.1
		LLR p-value:	4.539e-177
		Cox & Snell R-squ.:	0.0297024

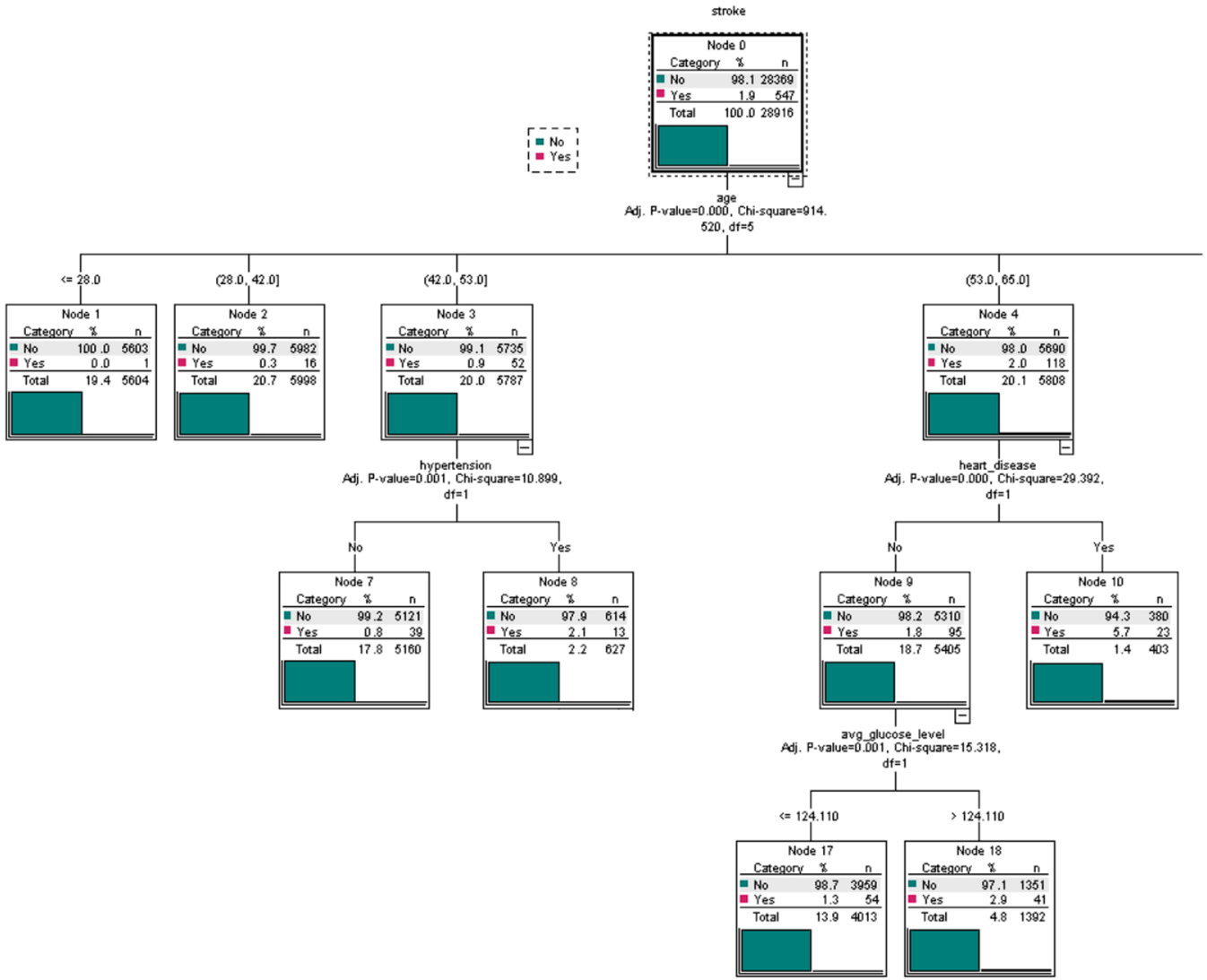
The Cox & Snell R Square of 0.0297 suggests the model only minimally predicts stroke risk, indicating **superior predictors outside of the dataset.**

Alternatively, **health issues combined** may have a greater impact on stroke risk than individually.

7.
DECISION TREE ANALYSIS

7. DECISION TREE ANALYSIS

THIS ANALYSIS IS RUN ON SPSS



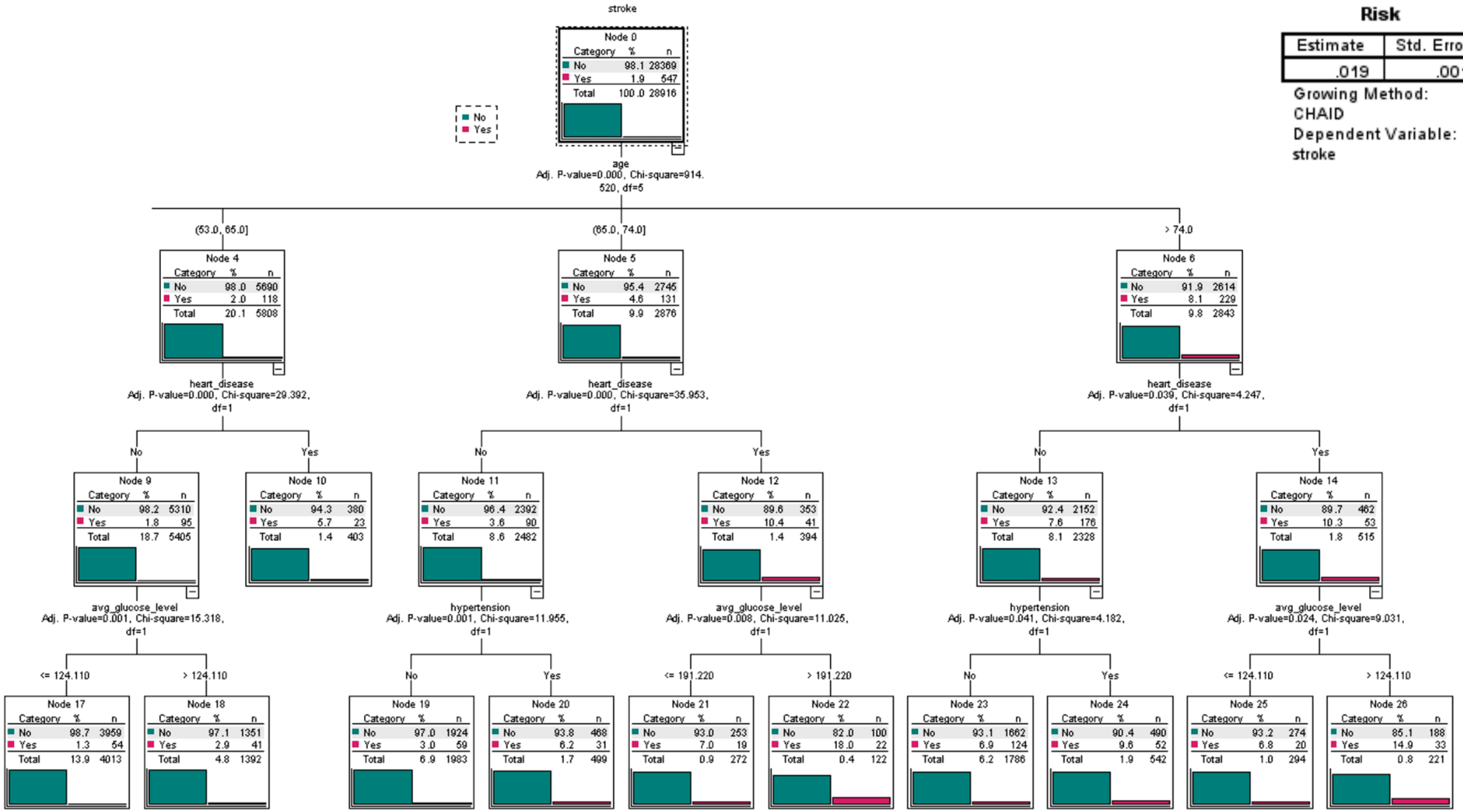
Risk

Estimate	Std. Error
.019	.001

Growing Method: CHAID
Dependent Variable: stroke

7. DECISION TREE ANALYSIS

THIS ANALYSIS IS RUN ON SPSS



Risk

Estimate	Std. Error
.019	.001

Growing Method: CHAID
Dependent Variable: stroke

7. DECISION TREE ANALYSIS

IMPLICATIONS

Age-specific stroke risk factors:

- Between *42-53 years old*, ***hypertension*** has the most significant impact on stroke risk.

- Between *53-65 years old*, ***heart diseases*** have the most significant impact on stroke risk. However, for those without heart diseases, an ***average glucose level greater than 124 mg/dL*** poses the most risk to stroke.

- Between *66-74 years old*, **heart diseases** have the most significant impact on stroke risk, and an ***average glucose level greater than 191 mg/dL*** increases that risk. However, for those without heart diseases, ***hypertension*** seems to be a significant predictor of stroke.

- *74 years old and above*, ***heart diseases*** have the most significant impact on stroke risk, and an ***average glucose level greater than 124 mg/dL*** increases that risk. However, for those without heart diseases, ***hypertension*** seems to be a significant predictor of stroke.

8. RELIABILITY TEST

IMPLICATIONS

Cronbach's alpha: 0.9877888225914057

The internal consistency reliability of the variables in the dataframe is not satisfactory.

However,

in the context of the stroke prediction dataset, **reliability and validity are not directly applicable concepts**, as they typically pertain to psychological or survey research with multiple items measuring a single construct.

9. IMPLICATIONS

- Stroke risk is highly correlated with age, exhibiting a significant **increase in probability after the age of 50 and a notable escalation at 80 years old.**
- Elevated glucose levels, especially **above the range of 110 mg/dL to 175 mg/dL**, can increase the risk of stroke.
- **Heart diseases and hypertension** have a profound impact on stroke risk.
- **Self-employed workers** are at a higher risk of stroke, whereas individuals with **multiple health issues** face higher risk in **government or private sector jobs.**
- Smoking becomes increasingly influential on stroke risk as **the number of health issues increases.**
- **More health issues are associated with a higher likelihood of stroke**, and as the number of health issues increases, the gap in stroke risk widens.
- Each factor individually, including **age, average glucose level, hypertension, heart diseases, and smoking**, can impact stroke risk. However, when combined, multiple health issues may have a greater impact on stroke risk compared to individual factors.
- **BMI and residence type do not significantly contribute** to stroke risk.

9. IMPLICATIONS VS. SECONDARY FINDINGS

- Health conditions that increase the risk of stroke: **High blood pressure, heart disease, diabetes**, etc...
- Age: “The older you are, the more likely you are to have a stroke. **The chance of having a stroke about doubles every 10 years after age 55**”.
- “In a meta-analysis of 32 studies, **cigarette smoking was found to be a significant independent contributor to stroke** with about a 50% increase in risk compared to non-smokers, and smoking contributed to 12% to 14% of all stroke deaths”.
- “Longitudinal studies have demonstrated that there is a **strong consistent association between blood pressure and stroke** in both men and women, for fatal and non fatal stroke and at all ages”.
- “Evidence derived from longitudinal studies has demonstrated that **marital strain, job stress** and depression are associated with coronary heart disease, mortality and stroke”.
- “**Coronary artery disease increases your risk for stroke**, because plaque builds up in the arteries and blocks the flow of oxygen-rich blood to the brain”.

10. REFERENCES

- Centers for Disease Control and Prevention (CDC). (2023, February 16). Know Your Risk for Stroke. Retrieved June 23, 2023, from https://www.cdc.gov/stroke/risk_factors.htm
- Johns Hopkins Medicine (2023, March 9). Stroke. https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke
- NCBI (2007, May 1). Risk Factors for Stroke. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3006180/
- AHA (2007, May 1). The Impact of Multiple Health Conditions on Stroke Risk. https://www.ahajournals.org/doi/full/10.1161/01.STR.27.5.819
- The Journal of Clinical Endocrinology & Metabolism (2020, August 1). Association of Elevated Blood Glucose with Stroke Risk: A Systematic Review and Meta-Analysis. https://academic.oup.com/endo/article/159/8/3120/5051605
- AHA (1998, March 1). Elevated Blood Glucose and Stroke Risk. https://www.ahajournals.org/doi/10.1161/01.str.0000115297.92132.84

THANK YOU FOR YOUR TIME!

Duong Lam Tuan Anh