Dataset Exploration - Part 4.

Name : Bhavesh Waghela

Student ID : 200532173

Table of Content

I. Dataset Description

WHO estimates that stroke accounts for approximately 11% of all total deaths globally, making it the 2nd most common cause of death.

Strokes can be caused by several factors, including:

a. Blood vessels supplying blood to the brain become blocked, usually by a blood clot.

b. Blood vessels in the brain rupture and bleeds.

c. Temporary disruptions in blood flow to the brain

Other risk factors for stroke include:

1. High blood pressure
2. Smoking
3. Diabetes
4. Atrial fibrillation (irregular heart rhythm)
5. High cholesterol
6. Family history of stroke
7. Sedentary lifestyle
8. Obesity
9. Substance abuse, including alcohol and drug use
10. Advanced age.

The above are risk factors that can lead to getting a stroke sooner or later. In our dataset we would analyze if the following is true and to what extent.

The following is the healthcare dataset which defines the records of 5110 patients. Each row provides applicable information about the patient - its medical and general background.

Data from this dataset is used to predict whether a patient is likely to suffer from stroke based on input parameters such as their gender, age, variety of diseases, and smoking status.

II. Research Questions

Q1. Does the person's age affect the risk of stroke. OR (Age $\propto$ Stroke)?

Q2. Is stroke influenced by gender?

Q.3 What is the impact of the type of work they do on stroke risk?

Q4. Can smoking increase the risk of stroke?

III. Dataset Resource Link

The following dataset has been taken for Analysis purposes from the website called Kaggle. I give credits to the website and the author who has published this raw dataset.

The link to the dataset and its information is given below.

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

IV. Data Dictionary

This data dictionary will provide information about the data used in the Dataset Excel Sheet i.e., 'healthcare-dataset-stroke-data.csv' that is attached.

1. Field Name – The different names of each field that is defined in the dataset and a brief description of those fields.
   a. **id** : Primary – Unique Identifier
   b. **gender** : "Male", "Female" or "Other"
   c. **age** : Age of the patient.
   d. **hypertension** : 0 if the patient does not have hypertension, 1 if the patient has hypertension.
   e. **heart_disease** : 0 if the patient does not have any

heart diseases, 1 if the patient has a heart disease.

f. **ever_married** : Yes OR No
g. **work_type** : "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
h. **residence_type** : "Urban" OR “Rural"
i. **avg_glucose_level** : Average glucose level in blood
j. **bmi** : Body Mass Index
k. **smoking_status** : "formerly smoked", "never smoked", "smokes" or "Unknown"*
l. **stroke** : 1 if the patient had a stroke or 0 if he did not.

2. Data Types: Data types for each field are listed below.

a. **id** : Integer
b. **gender** : String
c. **age** : Integer
d. **hypertension** : Integer
e. **heart_disease** : Integer
f. **ever_married** : String
g. **work_type** : String
h. **residence_type** : String
i. **avg_glucose_level** : Decimal
j. **bmi** : Decimal
k. **smoking_status** : String

l. **stroke** : Integer

3. Nullability : Its better for analysis If all the fields have a specific value. But in some cases, we do not have the information for BMI of the patient and the smoke status is unknown.
4. Key Fields : The patient ID i.e. the column “id” is Primary and foreign key fields and will be used as and identified of the patient in the dataset.
5. Definitions and Assumptions : The dataset fields some can be categorized for minimizing the possibility of human error.

a. gender : Male, Female and Others
b. **hypertension** : 0 or 1
c. heart_disease: 0 or 1
d. ever_married : Married or Single
e. **work_type** : "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

children – They are too small to work.

Govt_job – The are working for the government.

Never_worked – The do not work at all.

Private – They might be working for a private company – may be desk job or physical job.

Self-employed – They work for themselves, maybe they have their own business or a store.

The above can be divided into 4 categories.

f. **residence_type : Rural OR Urban**

Rural : Where there is not much access to healthcare facilities and a better lifestyle.

Urban : Where there is access to good health care facilities and a better lifestyle.

g. smoking_status : "formerly smoked", "never smoked", "smokes" or "Unknown"

formerly smoked : They used to smoke before but they have stopped smoking and do not smoke anymore.

never smoked : They never have smoked before and we assume that they will not smoke in the future.

smokes : These are patients that smoke – maybe on a regular bases or occasional.

Unknown : We do not have any information about whether the patient smokes or he doesn't smoke.

h. stroke : 0 or 1.

The above information describes the dataset according to my understanding and knowledge and will do the analysis on the research question based on these definitions and assumptions.

V. Data Set Univariate Analysis.

Univariate Analysis on the different Numerical Data Variables from the dataset.

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
Number of Missing	0	0	0	0	0	201	0
% Missing	0.00%	0.00%	0.00%	0.00%	0.00%	4.09%	0.00%
Number of Unique	5110	104	2	2	3979	418	2
Min	67	0.08	0	0	55.12	10.3	0
Q1	17741.25	25	0	0	77.25	23.5	0
Med	36932	45	0	0	91.89	28.1	0
Q3	54682	61	0	0	114.09	33.1	0
Max	72940	82	1	1	271.74	97.6	1
Mean	36517.83	43.23	0.10	0.05	106.15	28.89	0.05
Std.Deviation	22.61	22.61	0.30	0.23	45.28	7.85	0.22
Skewness	-0.14	-0.14	2.72	3.95	1.57	1.06	4.19
Kurtosis	-0.99	-0.99	5.38	13.59	1.68	3.36	15.59

Univariate Analysis on the different Categorical Data Variables from the dataset.

	gender	ever_married	work_type	residence_type	smoking_status
Number of Missing	0	0	0	0	0
% Missing	0.00%	0.00%	0.00%	0.00%	0.00%
Number of Unique	3	2	5	2	4

Gender Count Bifurcation

Row Labels	Gender Count	Count Percentage
Female	2994	58.59%
Male	2115	41.39%
Other	1	0.02%
Grand Total	**5110**	**100.00%**

Ever_Married Count Bifurcation

Row Labels	Count of ever_married	Percentage ever_married
No	1757	34.38%
Yes	3353	65.62%
Grand Total	5110	100.00%

Work_Type Count Bifurcation

Row Labels	Count of work_type	Percentage work_type
children	687	13.44%
Govt_job	657	12.86%
Never_worked	22	0.43%
Private	2925	57.24%
Self-employed	819	16.03%
Grand Total	**5110**	**100.00%**

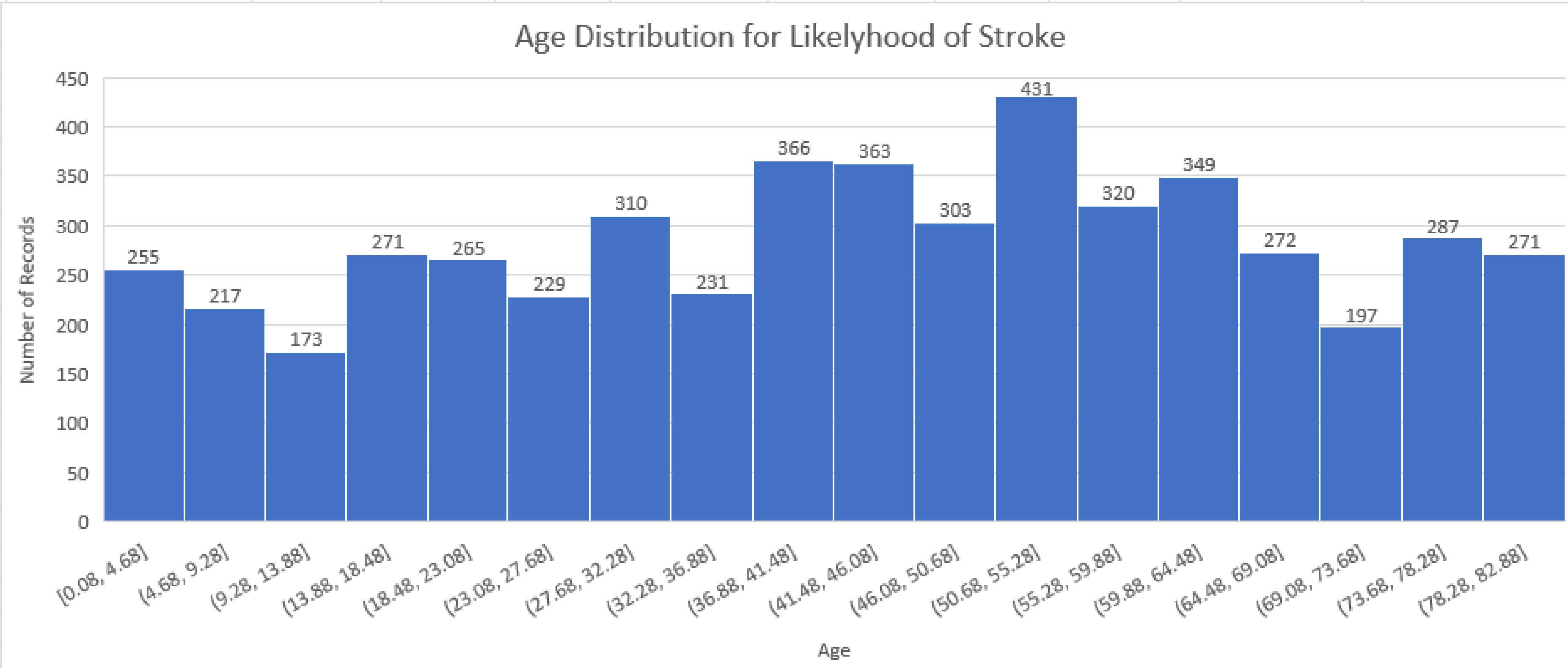
Residence_Type Count Bifurcation

Row Labels	Count of residence_type	Percentage residence_type
Rural	2514	49.20%
Urban	2596	50.80%
Grand Total	**5110**	**100.00%**

Smoking_Status Count Bifurcation

Row Labels	Count of smoking_status	Percentace smoking_status
formerly smoked	885	17.32%
never smoked	1892	37.03%
smokes	789	15.44%
Unknown	1544	30.22%
Grand Total	**5110**	**100.00%**

Histogram for Age Distribution over the Stroke Dataset.



Data Cleaning

BMI attribute had values as 'N/A' which was replaced by *Blank,* for easy Univariate Calculation.

VI. Bi- and Multi-variate Analysis

Bi variate analysis based on different variables in the dataset.

Based on Research Question - **Is stroke influenced by gender?**

To calculate the odds ratio and risk ratio for gender and the outcome of interest ie. Stroke status.

A. Odds ratio (OR) = (a/b) / (c/d)

where a = number of males with stroke, b = number of males without stroke, c = number of females with stroke, and d = number of females without stroke.

B. Risk ratio (RR) = (a/(a+b))/(c/(c+d))

where a, b, c, and d are defined as above.

Actual values

Count of stroke	Column Labels		
Gender	No Stroke	Stroke	Total
Female	2853	141	2994
Male	2007	108	2115
Total	**4861**	**249**	**5110**

Odds Ratio	(a/b) / (c/d)	**1.088827402**	0.918419208	*Inverted*
Risk Ratio	(a/(a+b))/(c/(c+d))	**1.084291535**	0.922261189	*Inverted*

Therefore, the odds of having a stroke are 1.088 times higher in males compared to females.

Note:
a = number of males with stroke
b = number of males without stroke
c = number of females with stroke
d = number of females without stroke

To calculate the lower and upper boundaries of the 95% confidence interval for the odds ratio, we can use the following formula:
ln = natural logarithm
SE = standard error of the log odds ratio, which is calculated as:

SE = sqrt(1/a + 1/b + 1/c + 1/d)	**0.0172**
ln(OR) =	**0.0851**
Lower bound = exp(ln(OR) - 1.96 * SE)	**1.05273**
Upper bound = exp(ln(OR) + 1.96 * SE)	**1.12616**

To calculate the 95% confidence interval for the risk ratio, we need to use the following formula:
ln = natural logarithm
SE = standard error of the log odds ratio, which is calculated as:
ln(Risk Ratio) ± 1.96 × SE(ln(Risk Ratio))

SE(ln(Risk Ratio)) = sqrt((1 / a) + (1 / b) + (1 / c) + (1 / d))	**0.13115**
ln(Risk Ratio)	**0.08093**
Lower bound = ln(Risk Ratio) - 1.96 × SE(ln(Risk Ratio))	0.83851
Lower bound = ln(Risk Ratio) + 1.96 × SE(ln(Risk Ratio))	1.40211

Expected values

Chi-Sq Test

The chi-square (χ^2) test is a statistical test used to compare observed data with expected data in order to determine whether there is a significant difference between the two.

formula for calculating chi-square is: $\chi^2 = \Sigma [(O - E)^2 / E]$

where:

- χ^2 is the chi-square test statistic
- Σ is the sum of the calculation for all categories or cells
- O is the observed frequency for a particular category or cell
- E is the expected frequency for a particular category or cell

Expected values

Gender	No Stroke	Stroke	Total
Female	2848.11	145.89	2994
Male	2011.94	103.06	2115
Total	**4861**	**249**	**5110**

Chi-square

Gender	No Stroke	Stroke	Total
Female	0.01	0.16	
Male	0.01	0.24	
Total			**0.42**

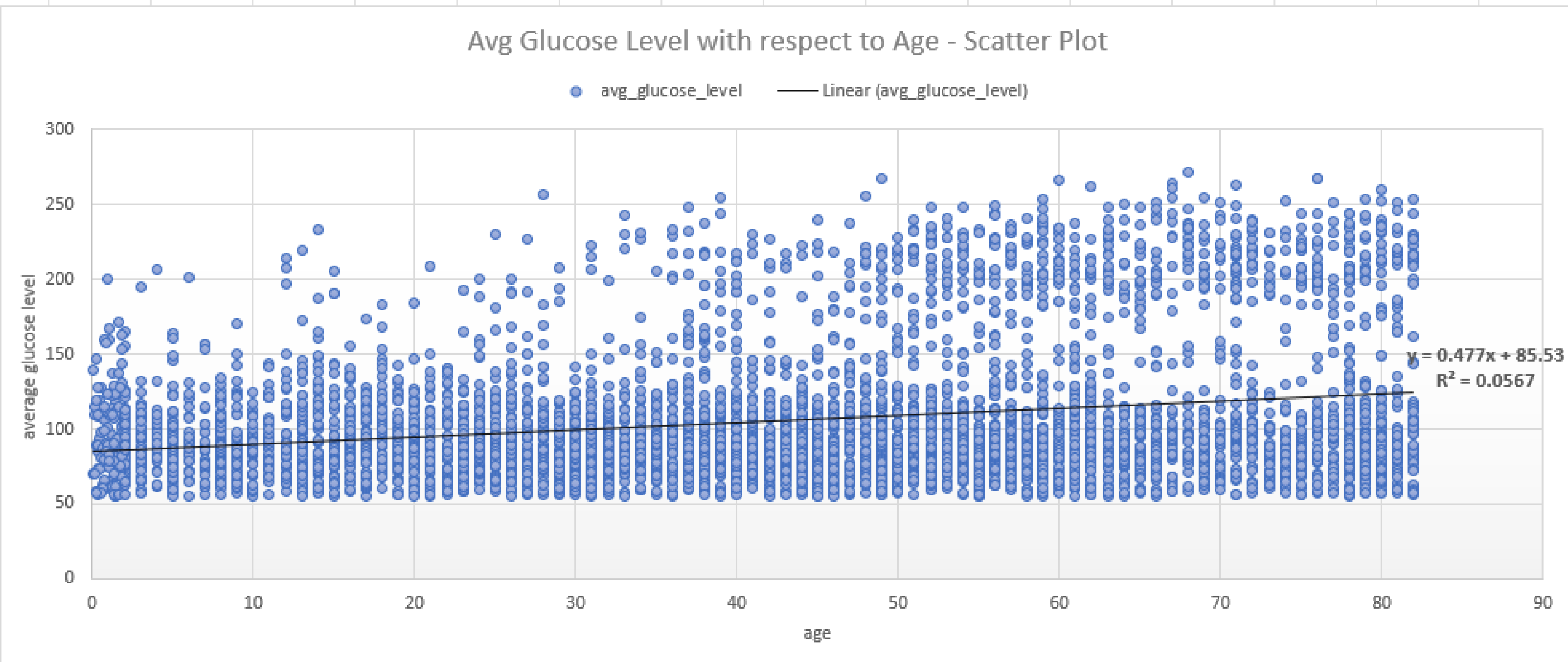
p-value for Chi-square	0.516257915

Bi-variate analysis – Numeric to Numeric Attributes.

A scatter plot is a type of chart that displays the relationship between two numeric variables.

Scatter Plot

A. In this case, the scatter plot shows the relationship between Age and Average glucose level.



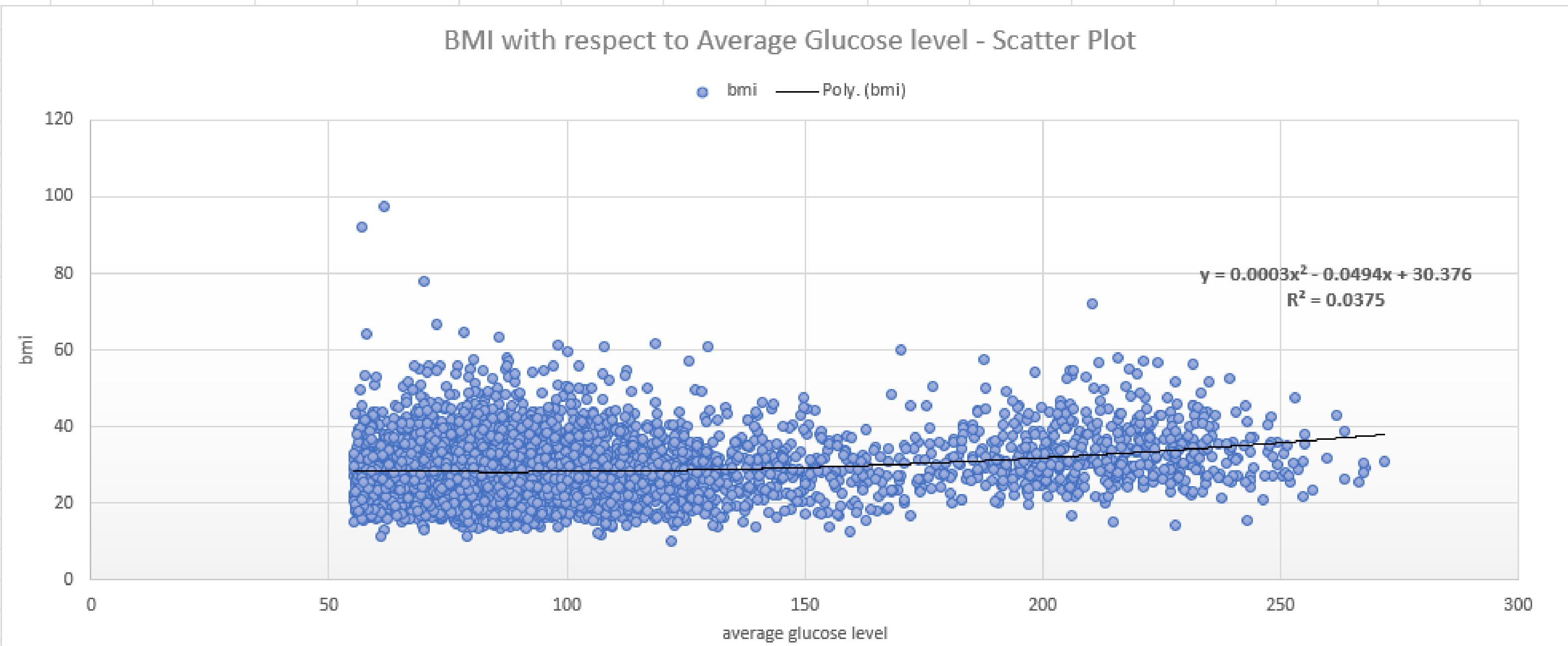
Person's Correlation
0.238171114

The pattern of points on the plot reveals that there is a relationship between age and average glucose level. This suggest that as age increases, average glucose level tends to increase as well.

If the correlation coefficient is close to +1, it indicates a strong positive correlation, while a correlation coefficient closer to 0 indicates a weak positive correlation.

As the correlation coefficient is close to 0.23, there is likely a week positive correlation between the two variables i.e., Age and average glucose level.

B. In this case, the scatter plot shows the relationship between BMI and Average glucose level.



Person's Correlation
0.175502176

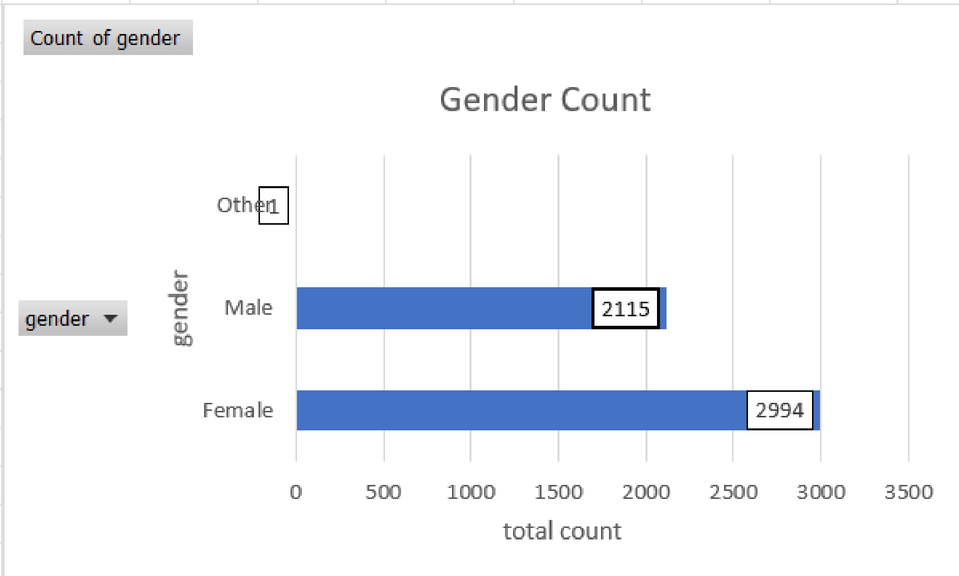
The pattern of points on the plot reveals that there is a relationship between BMI and average glucose level. This suggest that as BMI increases, average glucose level tends to increase as well.

As the correlation coefficient will be close to 0, there is likely no correlation between the two variables i.e., BMI and average glucose level.

Categorical Variable Analysis – Count

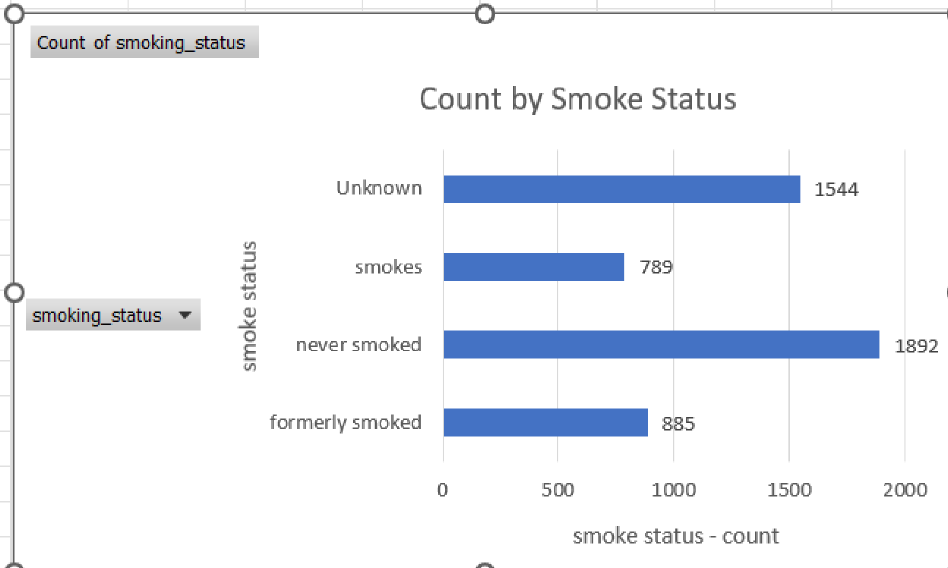
1. Count by Gender

Row Labels	Count of gender
Female	2994
Male	2115
Other	1
Grand Total	**5110**



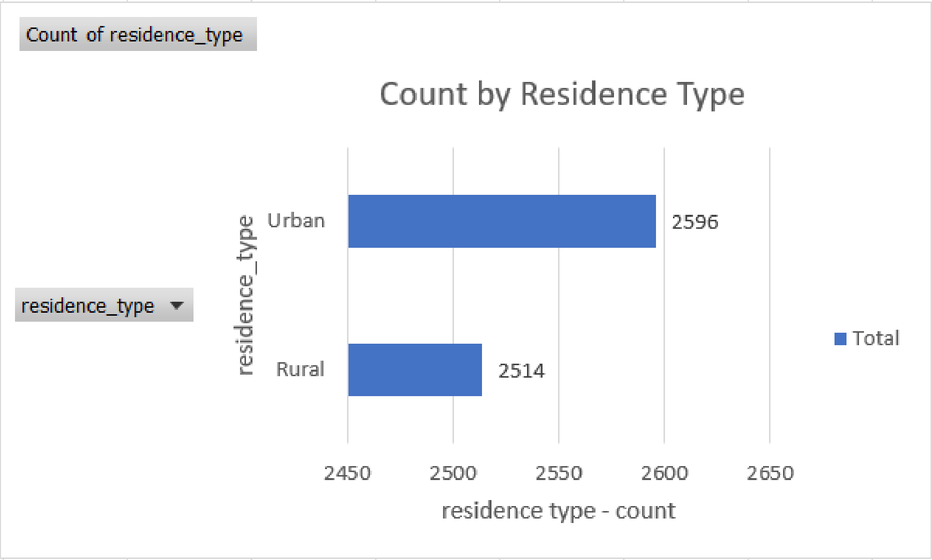
2. Count by Smoke Status

Row Labels	Count of smoking_status
formerly smoked	885
never smoked	1892
smokes	789
Unknown	1544
Grand Total	**5110**



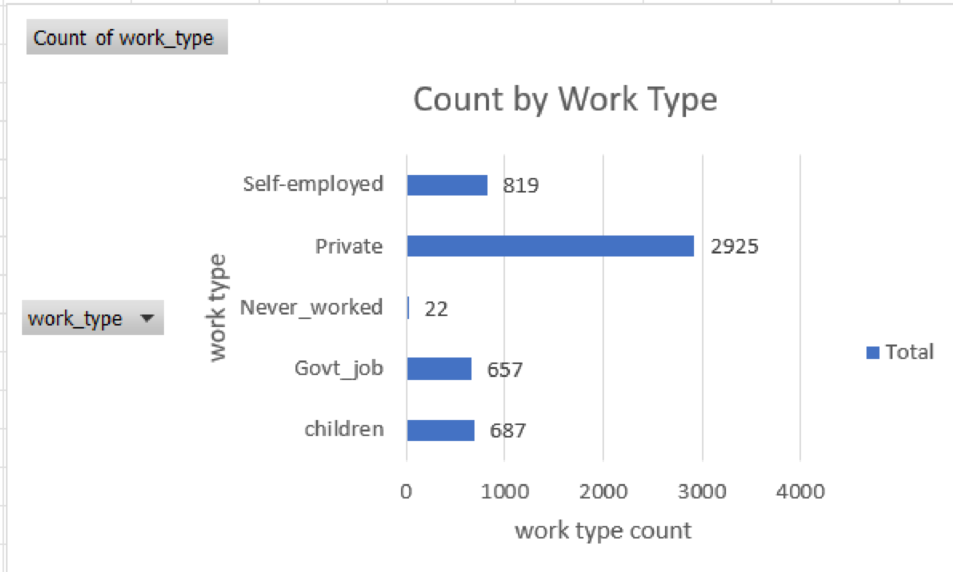
3. Count by Residence Type

Row Labels	Count of residence_type
Rural	2514
Urban	2596
Grand Total	**5110**



4. Count by Work Type

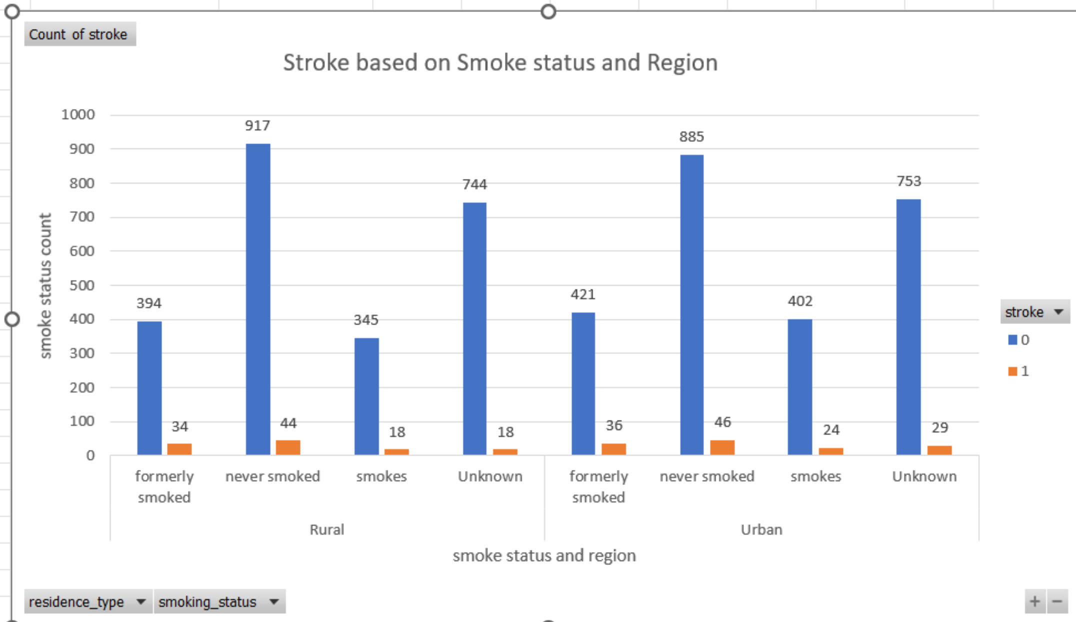
Row Labels	Count of work_type
children	687
Govt_job	657
Never_worked	22
Private	2925
Self-employed	819
Grand Total	**5110**



Relationship between different attributes.

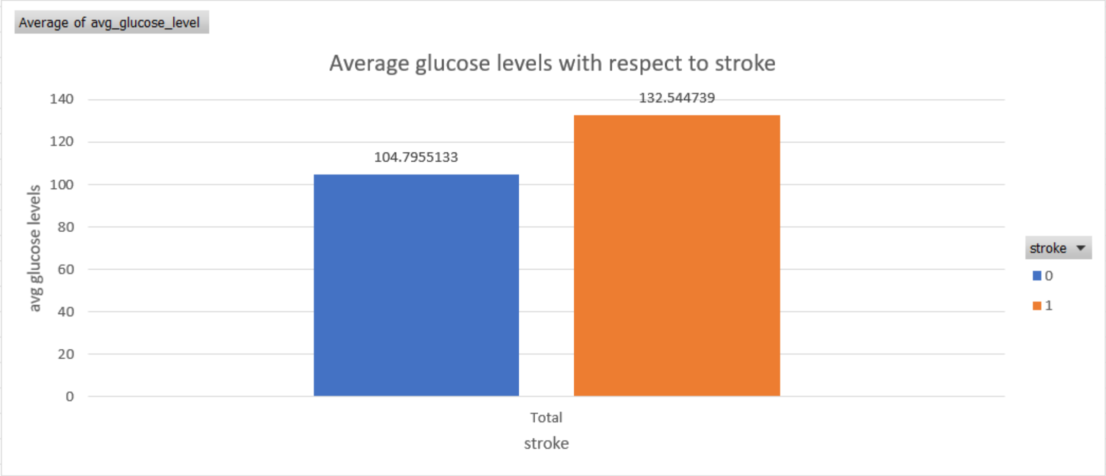
1. Stroke based on smoke status and region

Count of stroke	Column Labels		
Row Labels	0	1	Grand Total
Rural	**2400**	**114**	**2514**
formerly smoked	394	34	428
never smoked	917	44	961
smokes	345	18	363
Unknown	744	18	762
Urban	**2461**	**135**	**2596**
formerly smoked	421	36	457
never smoked	885	46	931
smokes	402	24	426
Unknown	753	29	782
Grand Total	**4861**	**249**	**5110**



2. Stroke based on average glucose levels

	Column Labels	
	0	1
Average of avg_glucose_level	104.7955133	132.544739



T-Test Calculation

1. **T-test between residence type and stroke.**

The null hypothesis would be that there is no significant difference in stroke incidence between rural and urban residents, while the alternative hypothesis would **be** that there is a significant difference.

Observer Results.

residence_type	stroke		Rural	Urban
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		1	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	1		0	1
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		1	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0
Urban	0		0	0

t-Test: Two-Sample Assuming Equal Variances

	Rural	*Urban*
Mean	0.045346062	0.052003082
Variance	0.043307023	0.049317759
Observations	2514	2596
Pooled Variance	0.046360637	
Hypothesized Mean Difference	0	
df	5108	
t Stat	-1.104917038	
P(T<=t) one-tail	0.13462379	
t Critical one-tail	1.645151992	
P(T<=t) two-tail	0.269247581	
t Critical two-tail	1.960428515	

t-Test: Two-Sample Assuming Unequal Variances

	Rural	*Urban*
Mean	0.045346062	0.052003082
Variance	0.043307023	0.049317759
Observations	2514	2596
Hypothesized Mean Difference	0	
df	5102	
t Stat	-1.10606846	
P(T<=t) one-tail	0.13437452	
t Critical one-tail	1.645152343	
P(T<=t) two-tail	0.26874904	
t Critical two-tail	1.960429062	

2. T-test between gender and stroke.

A	B	C	D	E	F	G	H	I	J	K	L	M
gender	stroke			Female	Male							
Female	0			0	1			t-Test: Two-Sample Assuming Equal Variances				
Female	0			0	1							
Female	0			0	1				*Female*	*Male*		
Female	0			0	1			Mean	0.047094188	0.05106383		
Female	0			0	1			Variance	0.04489132	0.048479237		
Female	0			0	1			Observations	2994	2115		
Female	0			0	1			Pooled Variance	0.046376508			
Female	0			0	1			Hypothesized Mean Difference	0			
Female	0			0	1			df	5107			
Female	0			0	1			t Stat	-0.648956203			
Female	0			0	1			P(T<=t) one-tail	0.258197933			
Female	0			0	1			t Critical one-tail	1.64515205			
Female	0			0	1			P(T<=t) two-tail	0.516395866			
Female	0			0	1			t Critical two-tail	1.960428606			
Female	0			0	1							
Female	0			0	1							
Female	0			0	1							
Female	0			0	1			t-Test: Two-Sample Assuming Unequal Variances				
Female	0			0	1							
Female	0			0	1				*Female*	*Male*		
Female	0			0	1			Mean	0.047094188	0.05106383		
Female	0			0	1			Variance	0.04489132	0.048479237		
Female	0			0	1			Observations	2994	2115		
Female	0			0	1			Hypothesized Mean Difference	0			
Female	0			0	1			df	4442			
Female	0			0	1			t Stat	-0.644679022			
Female	0			0	1			P(T<=t) one-tail	0.259584252			
Female	0			0	1			t Critical one-tail	1.645196736			
Female	0			0	1			P(T<=t) two-tail	0.519168504			
Female	0			0	1			t Critical two-tail	1.960498182			
Female	0			0	1							

ANOVA Test

1. ANOVA between Age and Average glucose level

Perform an ANOVA to determine if there is a significant difference in average glucose level across different age groups.

The null hypothesis would be that there is no significant difference in average glucose level between age groups, while the alternative hypothesis would be that there is a significant difference.

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
age	5110	220888	43.22661	511.3318
average_glucose_level	5110	542414.6	106.1477	2050.601

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	10115399	1	10115399	7896.694	0	3.842369054
Within Groups	13088914	10218	1280.966			
Total	23204312	10219				

VII. Regression Analysis

Correlation analysis on numerical data attributes.

	stroke	age	hypertension	heart_disease	avg_glucose_level	bmi
stroke	1					
age	0.245	1.000				
hypertension	0.128	0.276	1.000			
heart_disease	0.135	0.264	0.108	1.000		
avg_glucose_level	0.132	0.238	0.174	0.162	1.000	
bmi	0.042	0.333	0.168	0.041	0.176	1

This is a correlation matrix showing the pairwise correlations between stroke, age, hypertension, heart disease, average glucose level, and BMI.

Looking at the values in the matrix, we can see that stroke is positively correlated with age (0.245), average glucose level (0.132), and hypertension (0.128), meaning that as these variables increase, the likelihood of having a stroke also tends to increase.

Heart disease and average glucose level also show a weak positive correlation with stroke (0.135 and 0.132, respectively).

BMI, on the other hand, shows a very weak positive correlation with stroke (0.042), indicating that it is not strongly associated with the risk of stroke.

In summary, this correlation analysis suggests that age, hypertension, and average glucose level are the most important predictors of stroke, while heart disease and BMI have weaker associations with stroke.

Q1. Does the person's age affect the risk of stroke. OR (Age ∝ Stroke)?

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.245239485
R Square	0.060142405
Adjusted R Square	0.059958372
Standard Error	0.208784357
Observations	5109

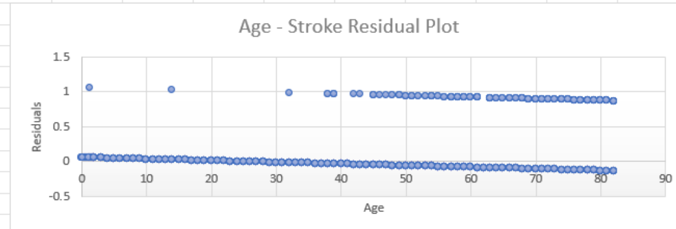
ANOVA

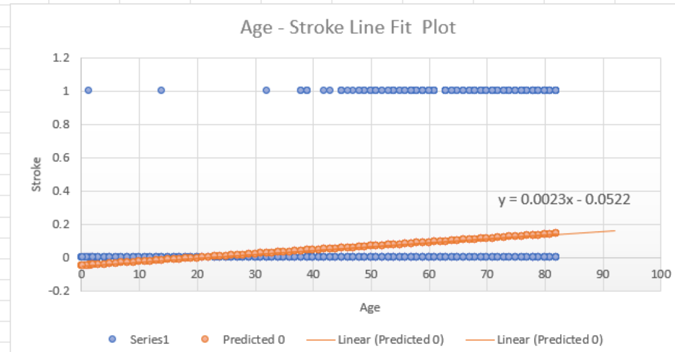
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	14.2456	14.24559	326.80191	7.4282E-71
Residual	5107	222.6188	0.04359		
Total	5108	236.8644			

	Coefficients	*Standard Error*	*t Stat*	***P-value***	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.0522	0.0063	-8.2866	0.0000	-0.0646	-0.0399	-0.0646	-0.0399
17	0.0023	0.0001	18.0777	0.0000	0.0021	0.0026	0.0021	0.0026

RESIDUAL OUTPUT

Observation	*Predicted 0*	*Residuals*	*Standard Residuals*
1	-0.021868436	0.021868436	0.104751992
2	0.045860787	-0.045860787	-0.219677744
3	0.020170392	-0.020170392	-0.096618189
4	0.036518825	-0.036518825	-0.174928814
5	0.003821959	-0.003821959	-0.018307563
6	0.134609424	-0.134609424	-0.64479257
7	0.024841373	-0.024841373	-0.118992653





Based on the given output, the coefficient for the variable "17" is 0.00233549, which indicates that for every unit increase in the variable "17" (which is likely to represent age in this case), there is a predicted increase of 0.00233549 in the outcome variable (which is likely to represent the risk of stroke in this case).

The intercept is -0.052229812, which represents the predicted value of the outcome variable when the value of the predictor variable (age) is zero. However, since it's not possible for age to be zero, the intercept doesn't have a meaningful interpretation in this case.

Therefore, based on the regression output provided, we can conclude that age (represented by the variable "17") has a statistically significant effect on the risk of stroke. However, it's important to note that correlation does not necessarily imply causation, so we cannot say for certain that age is causing the increased risk of stroke.

Forecasting Results:

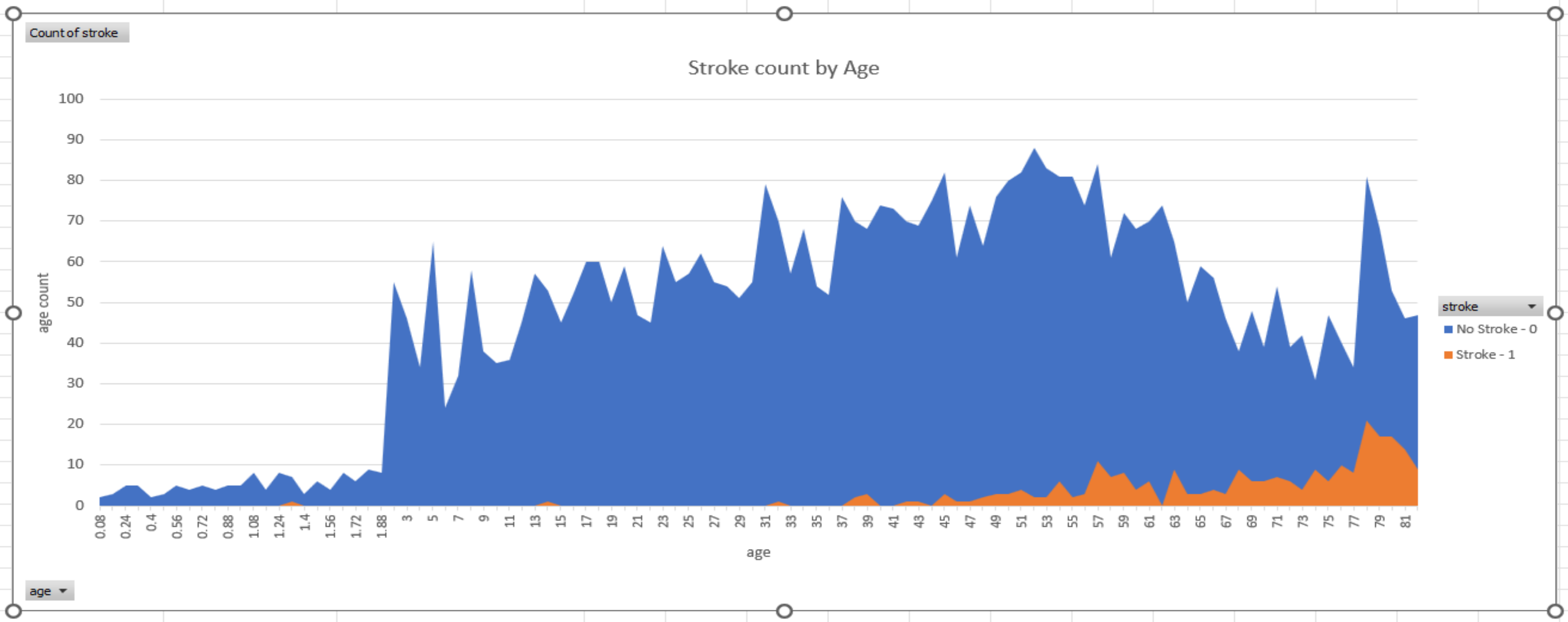
If we apply the regression equation y = 0.0023x - 0.0522 to forecast the response variable y for a value of x that is 10 units higher than the maximum value in the dataset, we get:

y = 0.0023(87) - 0.0522

y = 0.1925

Therefore, according to this forecast, a person with an age of 87 has a predicted risk of stroke of approximately 0.1925.

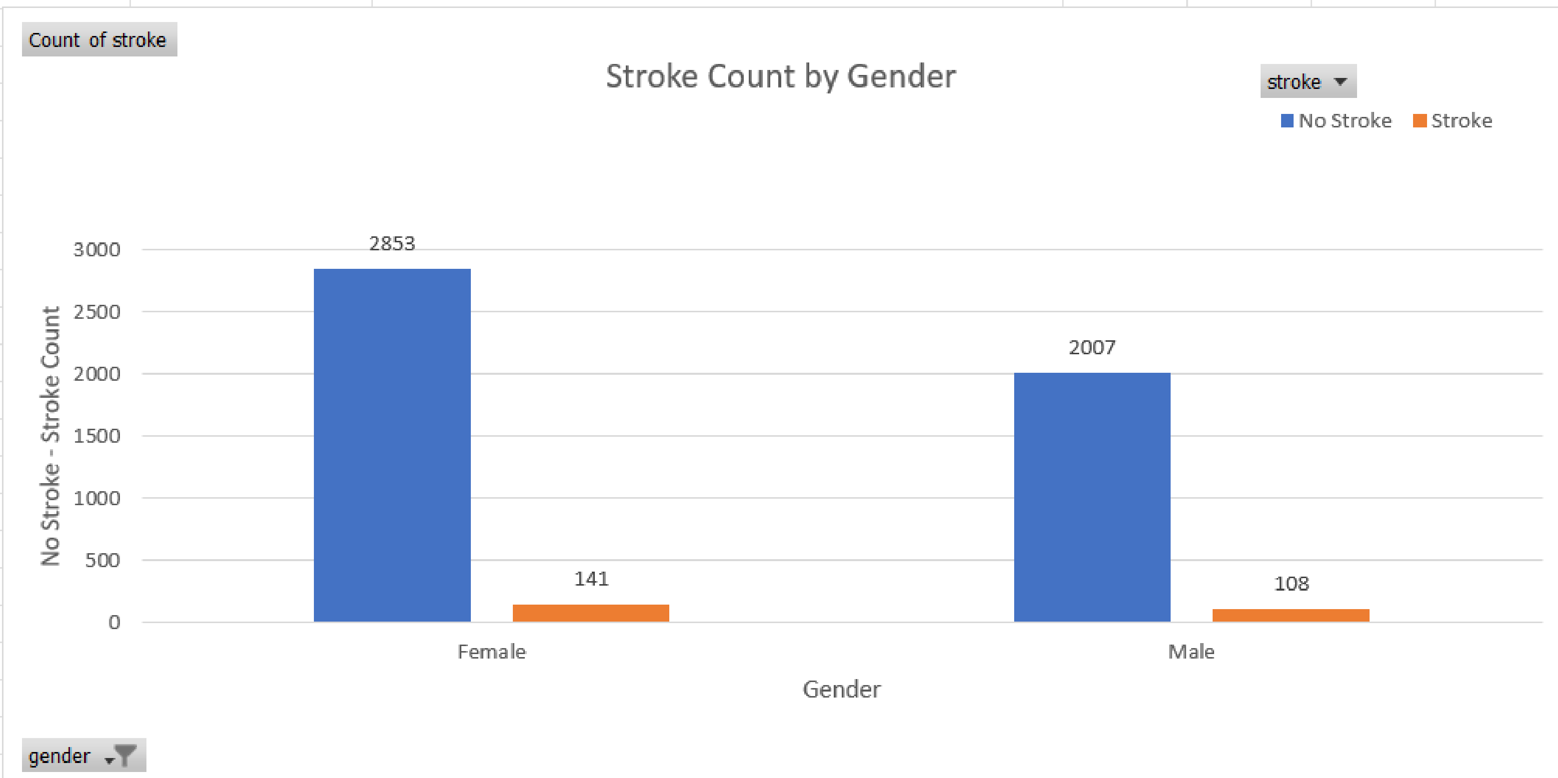
It is important to note that this forecast is based solely on the linear relationship between age and stroke observed in the given dataset. Other factors not included in the model may also play a role in determining stroke risk. Additionally, extrapolating beyond the range of the data (in this case, an age of 87) may lead to less accurate forecasts due to the potential for non-linear relationships or unobserved factors.



Q2. Is stroke influenced by gender?

Actual values

Count of stroke	Column Labels		
Gender	**No Stroke**	**Stroke**	**Total**
Female	2853	141	2994
Male	2007	108	2115
Total	**4861**	**249**	**5110**



Using these values, we can calculate the odds ratio and the 95% confidence interval.

The odds ratio can be calculated as:

odds ratio = (ad/bc) = (108/2007)/(141/2853) = 1.078

To calculate the standard error of the log odds ratio, we can use the formula:

SE = sqrt(1/a + 1/b + 1/c + 1/d) = sqrt(1/108 + 1/2007 + 1/141 + 1/2853) = 0.157

The 95% confidence interval for the odds ratio can be calculated as:

ln(OR) ± 1.96*SE

ln(1.078) ± 1.96*(0.157)

0.074 ± 0.308

Lower bound = exp(0.074 - 0.308) = 0.744

Upper bound = exp(0.074 + 0.308) = 2.652

Therefore, we can conclude that based on the given data, there is no statistically significant association between gender and stroke.

The odds ratio of 1.078 suggests a slight increase in the odds of stroke among males compared to females, but the confidence interval (0.744 to 2.652) includes 1, which means the difference could be due to chance.

Q.3 What is the impact of the type of work they do on stroke risk?

To analyze the impact of the type of work on stroke risk, we can perform a chi-squared test of independence. The null hypothesis for this test is that there is no association between the type of work and stroke risk. The alternative hypothesis is that there is a significant association between the two variables.

1. Creating a contingency table.
2. Calculate the expected counts under the assumption of independence.
3. Use the CHISQ.TEST function to calculate the p-value for the chi-squared test. The syntax for this function is:

=CHISQ.TEST(actual_range, expected_range)

Actual values			
	No Stroke	Stroke	Total
Children	685	2	687
Govt_job	624	33	657
Never_worked	22	0	22
Private	2776	149	2925
Self-employed	754	65	819
Total	4861	249	5110

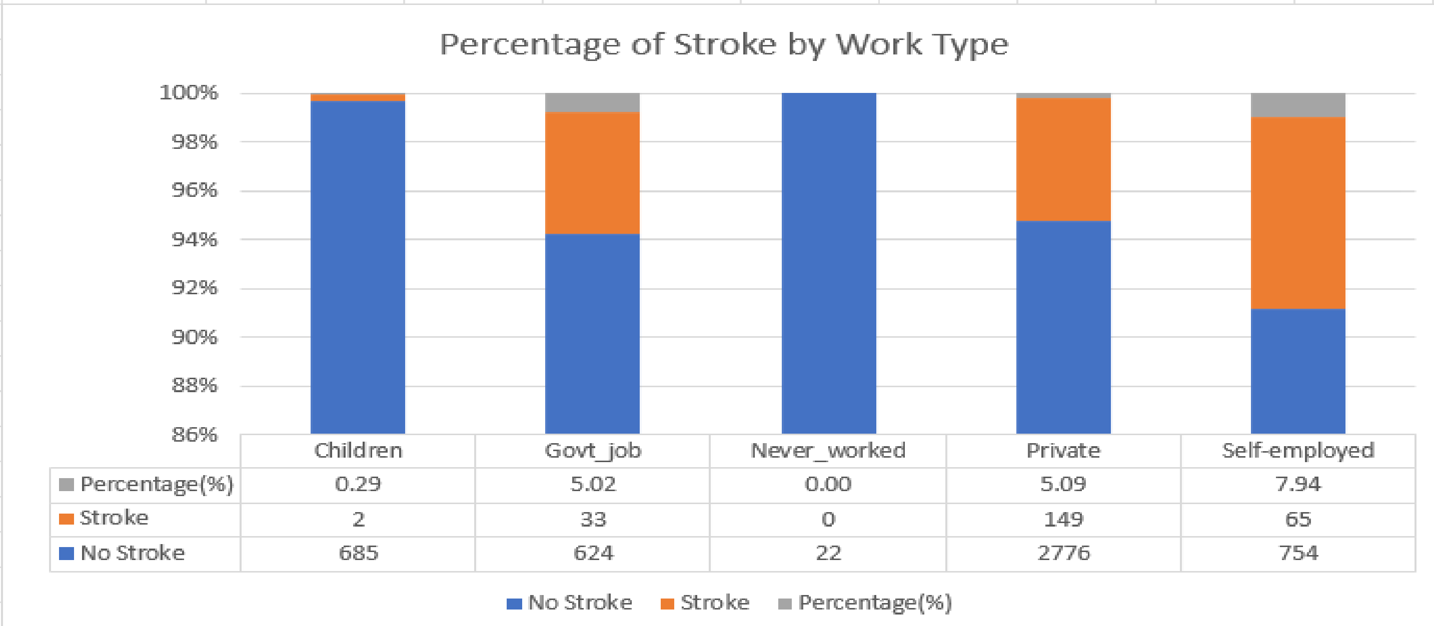
Expected values			
	No Stroke	Stroke	Total
Children	653.5238748	33.4761	687
Govt_job	624.9857143	32.0143	657
Never_worked	20.92798434	1.07202	22
Private	2782.470646	142.529	2925
Self-employed	779.0917808	39.9082	819
Total	4861	249	5110

p-value for Chi-square	0.0000000005398

The resulting p-value is 0.0000000005398, which is less than the typical significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that there is a significant association between the type of work and stroke risk.

Visualize the relationship between work type and stroke risk with a stacked bar chart. Here the x-axis represents the work types, and the y-axis represents the percentage of strokes.

Work Type	No Stroke	Stroke	Percentage(%)
Children	685	2	0.29
Govt_job	624	33	5.02
Never_worked	22	0	0.00
Private	2776	149	5.09
Self-employed	754	65	7.94



Q4. Can smoking increase the risk of stroke?

To analyze the impact of the type of smoke on stroke risk, we can perform a chi-squared test of independence. The null hypothesis for this test is that there is no association between smoking and stroke risk. The alternative hypothesis is that there is a significant association between the two variables.

1. Creating a contingency table.
2. Calculate the expected counts under the assumption of independence.
3. Use the CHISQ.TEST function to calculate the p-value for the chi-squared test. The syntax for this function is:

=CHISQ.TEST(actual_range, expected_range)

Actual values			
Row Labels	0	1	Grand Total
formerly smoked	815	70	885
never smoked	1802	90	1892
smokes	747	42	789
Unknown	1497	47	1544
Grand Total	4861	249	5110

Expected values			
Row Labels	0	1	Grand Total
formerly smoked	841.8757339	43.12	885
never smoked	1799.806654	92.19	1892
smokes	750.5536204	38.45	789
Unknown	1468.763992	75.24	1544
Grand Total	4861	249	5110

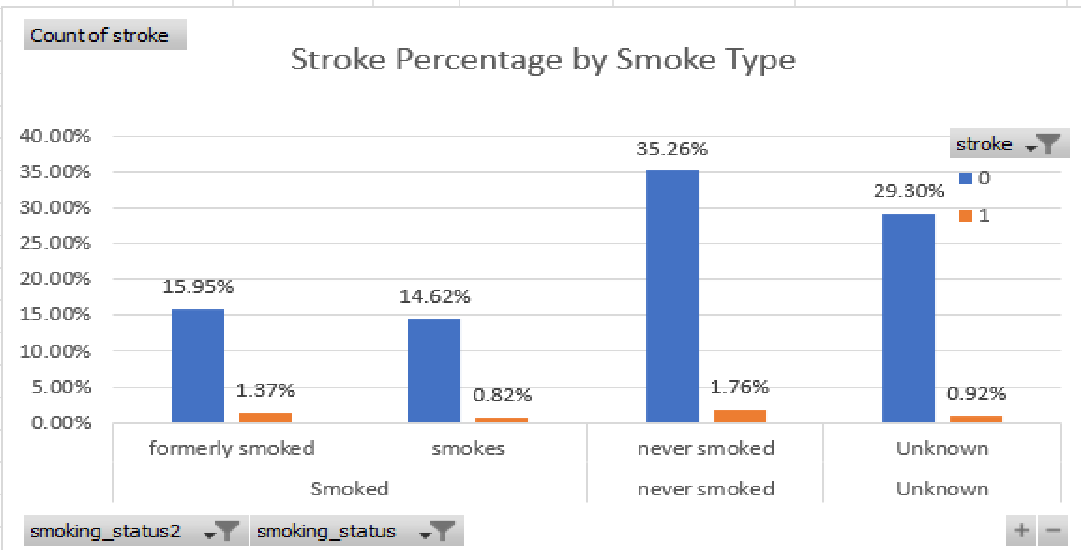
p-value for Chi-squar	0.0000020853997

The resulting p-value is 0.0000020853997, which is less than the typical significance level of 0.05. Therefore, we can reject the null hypothesis and conclude that there is a significant association between smoking and stroke risk.

For better analysis we have Grouped - Formerly smoked and Smokes so when combined we get a better understanding of chances of stroke in people who smoke.

Visualize the relationship between smoke type and stroke risk with a stacked bar chart. Here the x-axis represents the smoke type group, and the y-axis represents the percentage of strokes.

Count of stroke	Column		
Row Labels	**0**	**1**	**Grand Total**
Smoked	**30.57%**	**2.19%**	**32.76%**
formerly smoked	15.95%	1.37%	17.32%
smokes	14.62%	0.82%	15.44%
never smoked	**35.26%**	**1.76%**	**37.03%**
never smoked	35.26%	1.76%	37.03%
Unknown	**29.30%**	**0.92%**	**30.22%**
Unknown	29.30%	0.92%	30.22%
Grand Total	**95.13%**	**4.87%**	**100.00%**



VIII. Conclusion

After analyzing the data and addressing the research questions, several key findings have emerged that shed light on the relationship between different variables and the risk of stroke.

1. How does age impact the risk of stroke?

- The risk of stroke increases with age.
- This finding highlights the importance of regular health check-ups, particularly for older individuals, to monitor and manage their stroke risk.

2. Does gender have an impact on stroke risk?

- The analysis did not find any significant association between gender and stroke risk.
- This suggests that stroke risk may not differ significantly between men and women.

3. What is the impact of the type of work on stroke risk?

- Private and self-employed individuals appear to have a higher risk of stroke compared to those in government jobs or who have never worked.
- This finding highlights the need for workplace health and safety programs to prevent and manage stroke risk among employees, particularly those in high-risk occupations.

4. Can smoking increase the risk of stroke?

- Smokers have a significantly higher risk of stroke compared to those who have never smoked.
- This underscores the importance of smoking cessation programs and public health campaigns to reduce the prevalence of smoking and its associated health risks.

IX. Appendices

Appendix A: Data Description

Table 1: Description of the variables used in the analysis

Variable	Description
Gender	Male or Female
Stroke	0: No stroke, 1: Stroke
Work Type	Government job, Private job, Self-employed, Never worked
Smoking Status	Never smoked, Smokes, Formerly smoked, Unknown

Appendix B: Descriptive Statistics

Table 2: Descriptive statistics for the variables used in the analysis

Variable	Mean	Standard Deviation
Age	43.23	22.61
Average Glucose Level	106.15	45.29
BMI	28.89	7.85

Table 3: Frequency distribution of stroke cases by gender

Gender	No Stroke	Stroke	Total
Female	2853	141	2994
Male	2007	108	2115
Total	4861	249	5110

Appendix C: Chi-Square Analysis

Table 4: Chi-square analysis of the association between smoking status and stroke

Smoking Status	No Stroke	Stroke	Total
Never smoked	1802	90	1892
Smokes	747	42	789
Formerly smoked	815	70	885
Unknown	1497	47	1544
Total	4861	249	5110

The p-value for the chi-square test was 0.000002, indicating a significant association between smoking status and stroke.