

DATA ANALYSIS TO IDENTIFY RISK FACTORS FOR STROKE

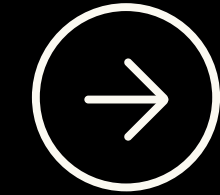


Prepared by:

1. Mostafa Fathy Mahmoud Abdel Rahim
2. Abanoub Fahem Fawzy Fahem
3. Michael Mamdouh Sedrak
4. Mohamed Ayman Sobhi

Supervisor: Abdullah Kamal

AGENDA



1 INTRODUCTION

2 RESEARCH PROBLEM

3 Global Stroke Rate

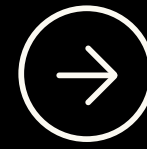
4 DATA USED

5 METHODOLOGY

6 Dashboard

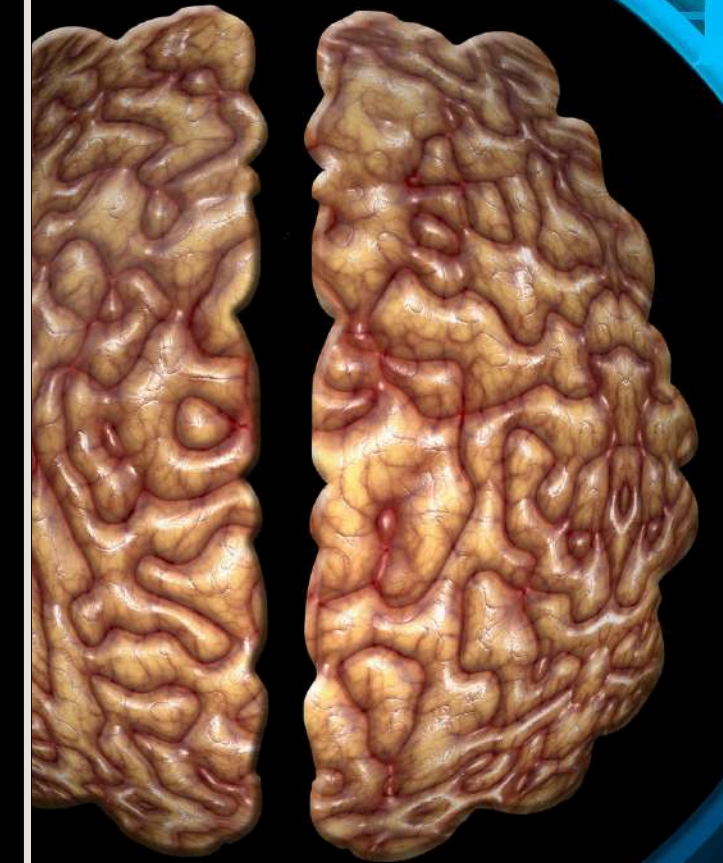
7 RESULTS

8 RECOMMENDATIONS



INTRODUCTION

- Stroke Is A Major Health Concern globally.
- The aim of this project is to analyze data to identify the main risk factors that contribute to Stroke .

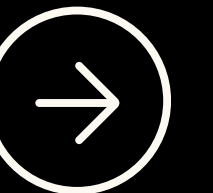




RESEARCH PROBLEM

Why is this research important?

- Stroke rates are increasing, making it crucial to understand its risk factors.
- Key research questions:
 - What are the most significant risk factors for stroke?
 - How can we use data to predict and mitigate these risks?





Global Stroke Rate

Data indicate an increasing trend in stroke cases globally between 2018 and 2021, with the number of cases increasing from 11.2 million in 2018 to 12.3 million in 2021. In 2022, there was a slight decrease to 11.9 million, but it increased again in 2023 and 2024 to reach 12.4 million cases.



DATA USED



- Data source: Patients' data from open datasets.
- Characteristics of the data:
 - Number of samples: 4981
 - Key variables: Age, Gender, Hypertension, Heart disease, Married or single , Work type, Residence type , Avg glucose level, Bmi , Smoking status, hospital visits

METHODOLOGY

DATA COLLECTION



the essential first step in research, providing accurate and relevant information for analysis.

DATA CLEANING



Removing Missing Values And Correcting

EXPLORATORY DATA ANALYSIS (EDA)



Examining variables, visualizations, and relationships.

MODELING



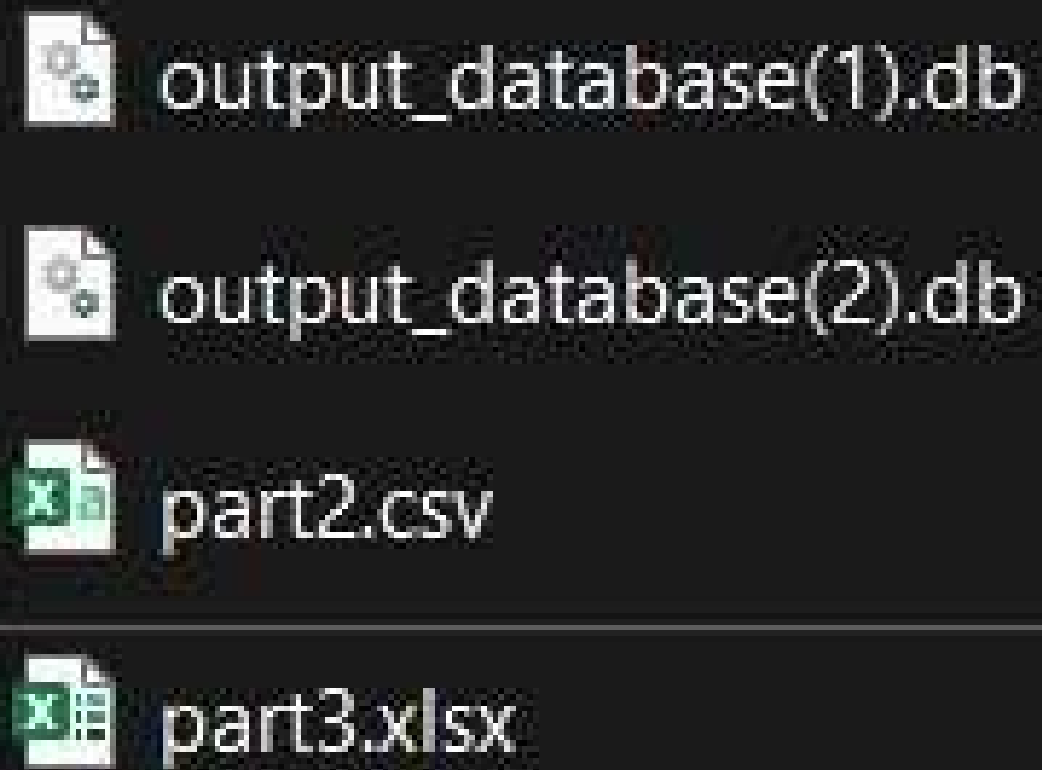
Using correlation or regression analysis to identify significant relationships.

- Split stroke data into three parts and save each part in a different format: CSV, Excel, and SQL.

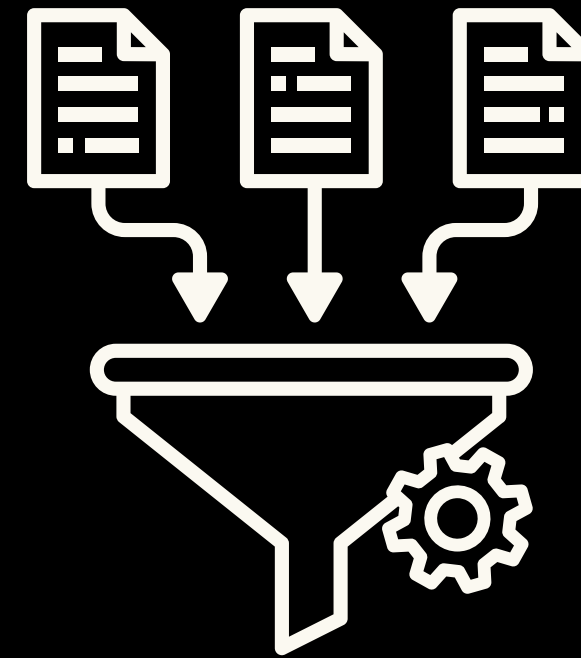
```
1 file_path = '/content/brain_stroke.csv'
2 df = pd.read_csv(file_path)
3
4
5 chunk_size = math.ceil(len(df) / 3)
6
7
8 df1 = df.iloc[:chunk_size]
9 df2 = df.iloc[chunk_size:2*chunk_size]
10 df3 = df.iloc[2*chunk_size:]
11
12 df2.to_csv('part2.csv', index=False)
13 df3.to_excel('part3.xlsx', index=False)
```

```
1 import sqlite3
2
3 conn = sqlite3.connect('output_database.db')
4
5 df1.to_sql('brain_stroke', conn, if_exists='replace', index=False)
6 conn.close()
```


- Each file stores a portion of the original data in a different format to facilitate analysis or integration with other systems.



output_database(1).db
output_database(2).db
part2.csv
part3.xlsx



- The purpose of this line of code is to create a new DataFrame (df) that combines all the stroke data pieces into one continuous dataset, by arranging them vertically. This is useful for performing analyses or operations on the entire dataset at once after splitting it into parts.

```
[ ] 1 df=pd.concat([df1,df2,df3],axis=0)
```

- We used Power Query.

Sheet1 - Power Query Editor

File Home Transform Add Column View

Close & Load Refresh Preview Properties Advanced Editor Manage Query

Choose Columns Remove Columns Keep Rows Remove Rows Sort Split Column Group By Data Type: Text Use First Row as Headers Replace Values Merge Queries Append Queries Combine Files Manage Parameters Data source settings New Source Recent Sources Enter Data New Query

Queries [1] Sheet1

Column1

1 MALE;54.0;0.0;0.0;Yess;govt_job;Rural;87.85;31.1;SMOKES;1.0

2 FEMALE;78.0;0.0;0.0;Yess;PRIVATE;Urban;106.74;33.0;FORMERLY SM...

3 NAN;81.0;nan;nan;Yess;NAN;Rural;80.13;nan;nver smoked;1.0

4 NAN;78.0;0.0;0.0;NO;self-employed;Urban;90.19;26.9;nver smoked;1.0

5 mle;78.0;nan;nan;Yess;nan;Rural;134.8;33.6;UNKNOWN;1.0

6 FEMALE;63.0;0.0;0.0;Yess;govt_job;Rural;106.58;23.9;UNKNOWN;1.0

7 FEMALE;unknown;nan;0.0;Yess;private;nan;high;91.45363605560401;...

8 femael;-5;nan;nan;Yess;nan;Rural;high;91.45363605560401;formarly s...

9 MALE;63.0;0.0;0.0;Yess;private;Urban;nan;30.7;NEVER SMOKED;1.0

10 nan;48.0;0.0;0.0;Yess;PRIVATE;Urban;74.11;error;NEVER SMOKED;1.0

11 femael;81.0;0.0;0.0;Yess;NAN;Urban;95.84;21.5;nver smoked;1.0

12 FEMALE;79.0;nan;1.0;NO;PRIVATE;nan;205.33;error;smoeks;nan

13 mle;149;0.0;0.0;Yess;self-employed;Urban;nan;27.1;nver smoked;1.0

14 FEMALE;55.0;1.0;1.0;Yess;PRIVATE;Urban;210.4;40.0;smoeks;1.0

15 NAN;73.0;0.0;0.0;Yess;govt_job;Rural;219.73;28.6;nan;nan

16 MALE;unknown;1.0;1.0;Yess;nan;Urban;250.89;91.45363605560401;s...

17 NAN;79.0;0.0;0.0;Yess;PRIVATE;Rural;72.73;error;NAN;1.0

18 FEMALE;63.0;0.0;0.0;Yess;nan;nan;high;91.45363605560401;formarly ...

19 NAN;81.0;0.0;0.0;NO;govt_job;Urban;nan;91.45363605560401;SMOKE...

20 mle;-5;nan;0.0;NO;private;nan;0.7713802524021462;31.9;SMOKES;1.0

21 femael;unknown;0.0;1.0;Yess;private;Urban;nan;31.0;unknown;1.0

22 femael;81.0;0.0;0.0;Yess;PRIVATE;Rural;nan;27.5;NEVER SMOKED;1.0

23 FEMALE;39.0;0.0;0.0;Yess;SELF-EMPLOYED;Urban;97.76;29.6;smoeks;...

24 femael;80.0;nan;nan;NO;private;nan;66.03;35.4;NEVER SMOKED;1.0

25 femael;nan;nan;0.0;Yess;self-employed;Rural;81.95;16.9;NEVER SMOK...

26 NAN;79.0;0.0;0.0;Yess;private;Urban;0.7713802524021462;21.5;NAN;...

27 NAN;73.0;0.0;0.0;Yess;SELF-EMPLOYED;Urban;70.94;nan;nan;1.0

28 FEMALE;77.0;1.0;0.0;Yess;SELF-EMPLOYED;Urban;199.84;28.0;formarl...

Query Settings

PROPERTIES

Name Sheet1

All Properties

APPLIED STEPS

Source

Navigation

Changed Type

- In Excel, we used the column splitting method to organize the data.

The screenshot displays the Microsoft Power Query Editor window. The ribbon at the top includes tabs for File, Home, Transform, Add Column, and View. The Transform tab is active, showing options like Split Column, Group By, and Replace Values. A data table with a single column named 'Column1' is visible, containing various text entries. A dialog box titled 'Split Column by Delimiter' is open in the center, prompting the user to specify a delimiter. The 'Semicolon' is selected in the 'Select or enter delimiter' dropdown. Under 'Split at', the option 'Each occurrence of the delimiter' is selected. The 'Quote Character' is set to double quotes. The 'Advanced options' section includes a checkbox for 'Split using special characters' and an 'Insert special character' button. The 'Query Settings' pane on the right shows the query name 'Sheet1' and a list of applied steps: 'Source', 'Navigation', and 'Changed Type'. The status bar at the bottom indicates '1 COLUMN, 999+ ROWS' and 'Column profiling based on top 1000 rows'. A timestamp 'PREVIEW DOWNLOADED AT 8:01 PM' is visible in the bottom right corner.

Sheet1 - Power Query Editor

File Home Transform Add Column View

Close & Load Refresh Preview Manage Query

Choose Columns Remove Columns Keep Rows Remove Rows Sort Split Column Group By Data Type: Text Use First Row as Headers Replace Values Merge Queries Append Queries Combine Files Manage Parameters Data source settings New Source Recent Sources Enter Data

Queries [1] Sheet1

Column1

MALE;54.0;0.0;0.0;Yess;govt_job;Rural;8
FEMALE;78.0;0.0;0.0;Yess;PRIVATE;Urban
NAN;81.0;nan;nan;Yess;NAN;Rural;80.13
NAN;78.0;0.0;0.0;NO;self-employed;Urban
mle;78.0;nan;nan;Yess;nan;Rural;134.8;3
FEMALE;63.0;0.0;0.0;Yess;govt_job;Rural
FEMALE;unknown;nan;0.0;Yess;private;nan
femael;-5;nan;nan;Yess;nan;Rural;high;9
MALE;63.0;0.0;nan;Yess;private;Urban;nan
nan;48.0;0.0;0.0;Yess;PRIVATE;Urban;74
femael;81.0;0.0;0.0;Yess;NAN;Urban;95.0
FEMALE;79.0;nan;1.0;NO;PRIVATE;nan;20
mle;149;0.0;0.0;Yess;self-employed;Urban
FEMALE;55.0;1.0;1.0;Yess;PRIVATE;Urban
NAN;73.0;0.0;nan;Yess;govt_job;Rural;21
MALE;unknown;1.0;1.0;Yess;nan;Urban;nan
NAN;79.0;0.0;0.0;Yess;PRIVATE;Rural;72
FEMALE;63.0;0.0;0.0;Yess;nan;nan;high;5
NAN;81.0;0.0;0.0;NO;govt_job;Urban;nan
mle;-5;nan;0.0;NO;private;nan;0.771380
femael;unknown;0.0;1.0;Yess;private;Urban;nan;31.0;unknown;1.0
femael;81.0;0.0;0.0;Yess;PRIVATE;Rural;nan;27.5;NEVER SMOKED;1.0
FEMALE;39.0;0.0;0.0;Yess;SELF-EMPLOYED;Urban;97.76;29.6;smoeks;...
femael;80.0;nan;nan;NO;private;nan;66.03;35.4;NEVER SMOKED;1.0
femael;nan;nan;0.0;Yess;self-employed;Rural;81.95;16.9;NEVER SMOK...
NAN;79.0;0.0;0.0;Yess;private;Urban;0.7713802524021462;21.5;NAN;...
NAN;73.0;0.0;0.0;Yess;SELF-EMPLOYED;Urban;70.94;nan;nan;1.0
FEMALE;77.0;1.0;0.0;Yess;SELF-EMPLOYED;Urban;199.84;28.0;formarl...

Table.TransformColumnTypes(Sheet1_Sheet,{{"Column1", type text}})

Split Column by Delimiter

Specify the delimiter used to split the text column.

Select or enter delimiter

Semicolon

Split at

☐ Left-most delimiter

☐ Right-most delimiter

☒ Each occurrence of the delimiter

Advanced options

Quote Character

"

☐ Split using special characters

Insert special character

OK Cancel

Query Settings

PROPERTIES

Name

Sheet1

All Properties

APPLIED STEPS

Source

Navigation

Changed Type

1 COLUMN, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 8:01 PM

Sheet1 - Power Query Editor

File

Home

Transform

Add Column

View

Close & Load

Refresh Preview

Advanced Editor

Manage

Choose Columns

Remove Columns

Keep Rows

Remove Rows

Sort

Split Column

Group By

12 Replace Values

Data Type: Text

Use First Row as Headers

Merge Queries

Append Queries

Combine Files

Manage Parameters

Data source settings

New Source

Recent Sources

Enter Data

Close

Query

Manage Columns

Reduce Rows

Transform

Combine

Parameters

Data Sources

New Query

Queries [1]
Sheet1

✕

✓

fx

= Table.TransformColumnTypes(#"Split Column by Delimiter",{{"Column1.1", type text}, {"Column1.2", type text}, {"Column1.3", type text},

	Column1.1	Column1.2	Column1.3	Column1.4	Column1.5	Column1.6	Column1.7
1	MALE	54.0	0.0	0.0	Yess	govt_job	Rural
2	FEMALE	78.0	0.0	0.0	Yess	PRIVATE	Urban
3	NAN	81.0	nan	nan	Yess	NAN	Rural
4	NAN	78.0	0.0	0.0	NO	self-employed	Urban
5	mle	78.0	nan	nan	Yess	nan	Rural
6	FEMALE	63.0	0.0	0.0	Yess	govt_job	Rural
7	FEMALE	unknown	nan	0.0	Yess	private	nan
8	femael	-5	nan	nan	Yess	nan	Rural
9	MALE	63.0	0.0	nan	Yess	private	Urban
10	nan	48.0	0.0	0.0	Yess	PRIVATE	Urban
11	femael	81.0	0.0	0.0	Yess	NAN	Urban
12	FEMALE	79.0	nan	1.0	NO	PRIVATE	nan
13	mle	149	0.0	0.0	Yess	self-employed	Urban
14	FEMALE	55.0	1.0	1.0	Yess	PRIVATE	Urban
15	NAN	73.0	0.0	nan	Yess	govt_job	Rural
16	MALE	unknown	1.0	1.0	Yess	nan	Urban
17	NAN	79.0	0.0	0.0	Yess	PRIVATE	Rural
18	FEMALE	63.0	0.0	0.0	Yess	nan	nan
19	NAN	81.0	0.0	0.0	NO	govt_job	Urban
20	mle	-5	nan	0.0	NO	private	nan
21	femael	unknown	0.0	1.0	Yess	private	Urban
22	femael	81.0	0.0	0.0	Yess	PRIVATE	Rural
23	FEMALE	39.0	0.0	0.0	Yess	SELF-EMPLOYED	Urban
24	femael	80.0	nan	nan	NO	private	nan
25	femael	nan	nan	0.0	Yess	self-employed	Rural
26	NAN	79.0	0.0	0.0	Yess	private	Urban
27	NAN	73.0	0.0	0.0	Yess	SELF-EMPLOYED	Urban
28							

Query Settings

PROPERTIES

Name

Sheet1

All Properties

APPLIED STEPS

Source

Navigation

Changed Type

Split Column by Delimiter

Changed Type1



- Change data type

Sheet1 - Power Query Editor

FileHomeTransformAdd ColumnView

Close & Load

Refresh Preview

Close

Properties

Advanced Editor

Manage

Choose Columns

Remove Columns

Manage Columns

Keep Rows

Remove Rows

Reduce Rows

A-Z

Z-A

Sort

Split Column

Group By

Transform

Data Type: Text

Use First Row as Headers

Replace Values

Merge Queries

Append Queries

Combine Files

Combine

Manage Parameters

Parameters

Data source settings

Data Sources

New Source

Recent Sources

Enter Data

New Query

Queries [1]

Sheet1

= Table.TransformColumnTypes("#Split Column by Delimiter",{{"Column1.1", type text}, {"Column1.2", type text}, {"Column1.3", type text},

	Column1.1	Column1.2	Column1.3	Column1.4	Column1.5	Column1.6	Column1.7
1	MALE	54.0	0.0	0.0	Yess	govt_job	Rural
2	FEMALE	78.0	0.0	0.0	Yess	PRIVATE	Urban
3	NAN	81.0	nan	nan	Yess	NAN	Rural
4	NAN	78.0	0.0	0.0	NO	self-employed	Urban
5	mle	78.0	nan	nan	Yess	nan	Rural
6	FEMALE	63.0	0.0	0.0	Yess	govt_job	Rural
7	FEMALE	unknown	nan	0.0	Yess	private	nan
8	femael	-5	nan	nan	Yess	nan	Rural
9	MALE	63.0	0.0	nan	Yess	private	Urban
10	nan	48.0	0.0	0.0	Yess	PRIVATE	Urban
11	femael	81.0	0.0	0.0	Yess	NAN	Urban
12	FEMALE	79.0	nan	1.0	NO	PRIVATE	nan
13	mle	149	0.0	0.0	Yess	self-employed	Urban
14	FEMALE	55.0	1.0	1.0	Yess	PRIVATE	Urban
15	NAN	73.0	0.0	nan	Yess	govt_job	Rural
16	MALE	unknown	1.0	1.0	Yess	nan	Urban
17	NAN	79.0	0.0	0.0	Yess	PRIVATE	Rural
18	FEMALE	63.0	0.0	0.0	Yess	nan	nan
19	NAN	81.0	0.0	0.0	NO	govt_job	Urban
20	mle	-5	nan	0.0	NO	private	nan
21	femael	unknown	0.0	1.0	Yess	private	Urban
22	femael	81.0	0.0	0.0	Yess	PRIVATE	Rural
23	FEMALE	39.0	0.0	0.0	Yess	SELF-EMPLOYED	Urban
24	femael	80.0	nan	nan	NO	private	nan
25	femael	nan	nan	0.0	Yess	self-employed	Rural
26	NAN	79.0	0.0	0.0	Yess	private	Urban
27	NAN	73.0	0.0	0.0	Yess	SELF-EMPLOYED	Urban
28							

Query Settings

PROPERTIES

Name

Sheet1

All Properties

APPLIED STEPS

Source

Navigation

Changed Type

Split Column by Delimiter

Changed Type1



AutoSave

Off

output.xlsx • Saved to this PC

Search

M

FileHomeInsertPage LayoutFormulasDataReviewViewAutomateHelp

Get

From Text/CSV

From Web

From Table/Range

From Picture

Recent Sources

Existing Connections

Get & Transform Data

Refresh

All

Queries & Connections

Properties

Workbook Links

Queries & Connections

Stocks

Currencies

Data Types

Sort

Filter

Sort & Filter

Clear

Reapply

Advanced

Data Tools

What-If Analysis

Forecast Sheet

Forecast

Group

Ungroup

Subtotal

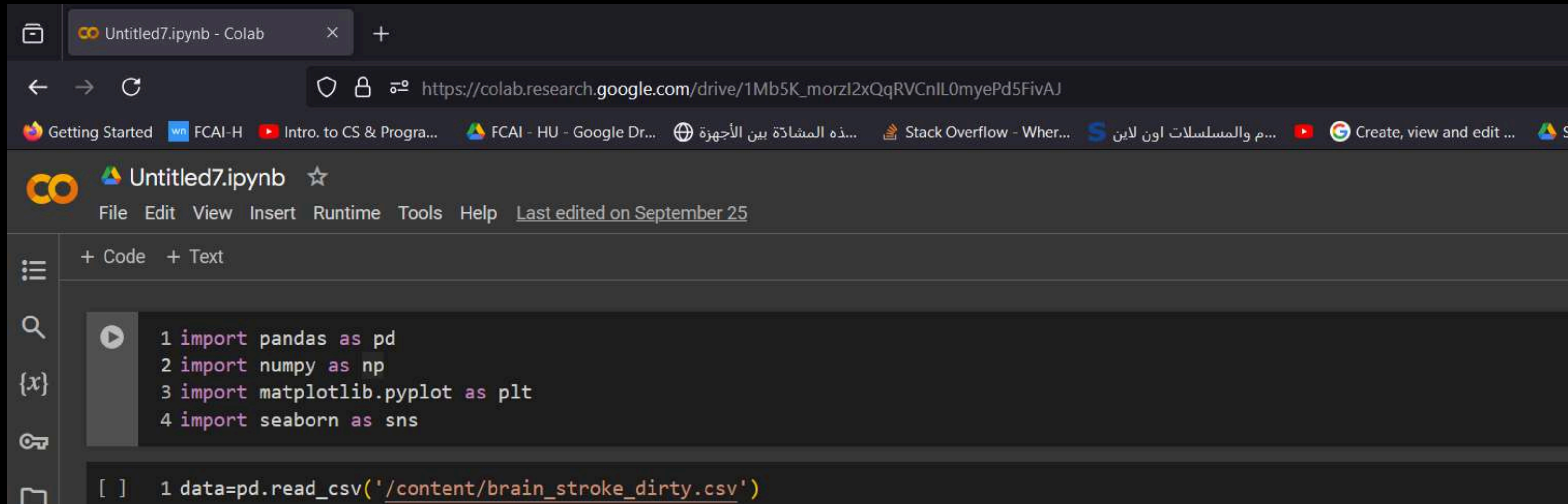
Outline

CommentsShare

L5

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	gender	age	hypertension	heart_disease	ever_married	work_type	residence_type	avg_glucose_level	bmi	smoking_status	stroke						
2	MALE	54.0	0.0	0.0	Yess	govt_job	Rural	87.85	31.1	SMOKES	1.0						
3	FEMALE	78.0	0.0	0.0	Yess	PRIVATE	Urban	106.74	33.0	FORMERLY SMOKED	nan						
4	NAN	81.0	nan	nan	Yess	NAN	Rural	80.13	nan	nver smoked	1.0						
5	NAN	78.0	0.0	0.0	NO	self-employed	Urban	90.19	26.9	nver smoked	1.0						
6	mle	78.0	nan	nan	Yess	nan	Rural	134.8	33.6	UNKNOWN	1.0						
7	FEMALE	63.0	0.0	0.0	Yess	govt_job	Rural	106.58	23.9	UNKNOWN	1.0						
8	FEMALE	unknown	nan	0.0	Yess	private	nan	high	91.45363605560401	NAN	1.0						
9	femael	-5	nan	nan	Yess	nan	Rural	high	91.45363605560401	formarly smoked	1.0						
10	MALE	63.0	0.0	nan	Yess	private	Urban	nan	30.7	NEVER SMOKED	1.0						
11	nan	48.0	0.0	0.0	Yess	PRIVATE	Urban	74.11	error	NEVER SMOKED	1.0						
12	femael	81.0	0.0	0.0	Yess	NAN	Urban	95.84	21.5	nver smoked	1.0						
13	FEMALE	79.0	nan	1.0	NO	PRIVATE	nan	205.33	error	smoeks	nan						
14	mle	149	0.0	0.0	Yess	self-employed	Urban	nan	27.1	nver smoked	1.0						
15	FEMALE	55.0	1.0	1.0	Yess	PRIVATE	Urban	210.4	40.0	smoeks	1.0						
16	NAN	73.0	0.0	nan	Yess	govt_job	Rural	219.73	28.6	nan	nan						
17	MALE	unknown	1.0	1.0	Yess	nan	Urban	250.89	91.45363605560401	smoeks	1.0						
18	NAN	79.0	0.0	0.0	Yess	PRIVATE	Rural	72.73	error	NAN	1.0						
19	FEMALE	63.0	0.0	0.0	Yess	nan	nan	high	91.45363605560401	formarly smoked	1.0						
20	NAN	81.0	0.0	0.0	NO	govt_job	Urban	nan	91.45363605560401	SMOKES	1.0						
21	mle	-5	nan	0.0	NO	private	nan	0.7713802524021462	31.9	SMOKES	1.0						
22	femael	unknown	0.0	1.0	Yess	private	Urban	nan	31.0	unknown	1.0						
23	femael	81.0	0.0	0.0	Yess	PRIVATE	Rural	nan	27.5	NEVER SMOKED	1.0						
24	FEMALE	39.0	0.0	0.0	Yess	SELF-EMPLOYED	Urban	97.76	29.6	smoeks	1.0						
25	femael	80.0	nan	nan	NO	private	nan	66.03	35.4	NEVER SMOKED	1.0						
26	femael	nan	nan	0.0	Yess	self-employed	Rural	81.95	16.9	NEVER SMOKED	1.0						
27	NAN	79.0	0.0	0.0	Yess	private	Urban	0.7713802524021462	21.5	NAN	1.0						

- Pandas and NumPy libraries were imported for data analysis and mathematical operations, along with Matplotlib and Seaborn for creating visualizations and statistical plots



The screenshot displays a Google Colab notebook titled "Untitled7.ipynb". The browser address bar shows the URL https://colab.research.google.com/drive/1Mb5K_morz12xQqRVCnIL0myePd5FivAJ. The notebook interface includes a menu bar with options: File, Edit, View, Insert, Runtime, Tools, and Help. Below the menu, there are tabs for "+ Code" and "+ Text". The code editor shows the following Python code:

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
```

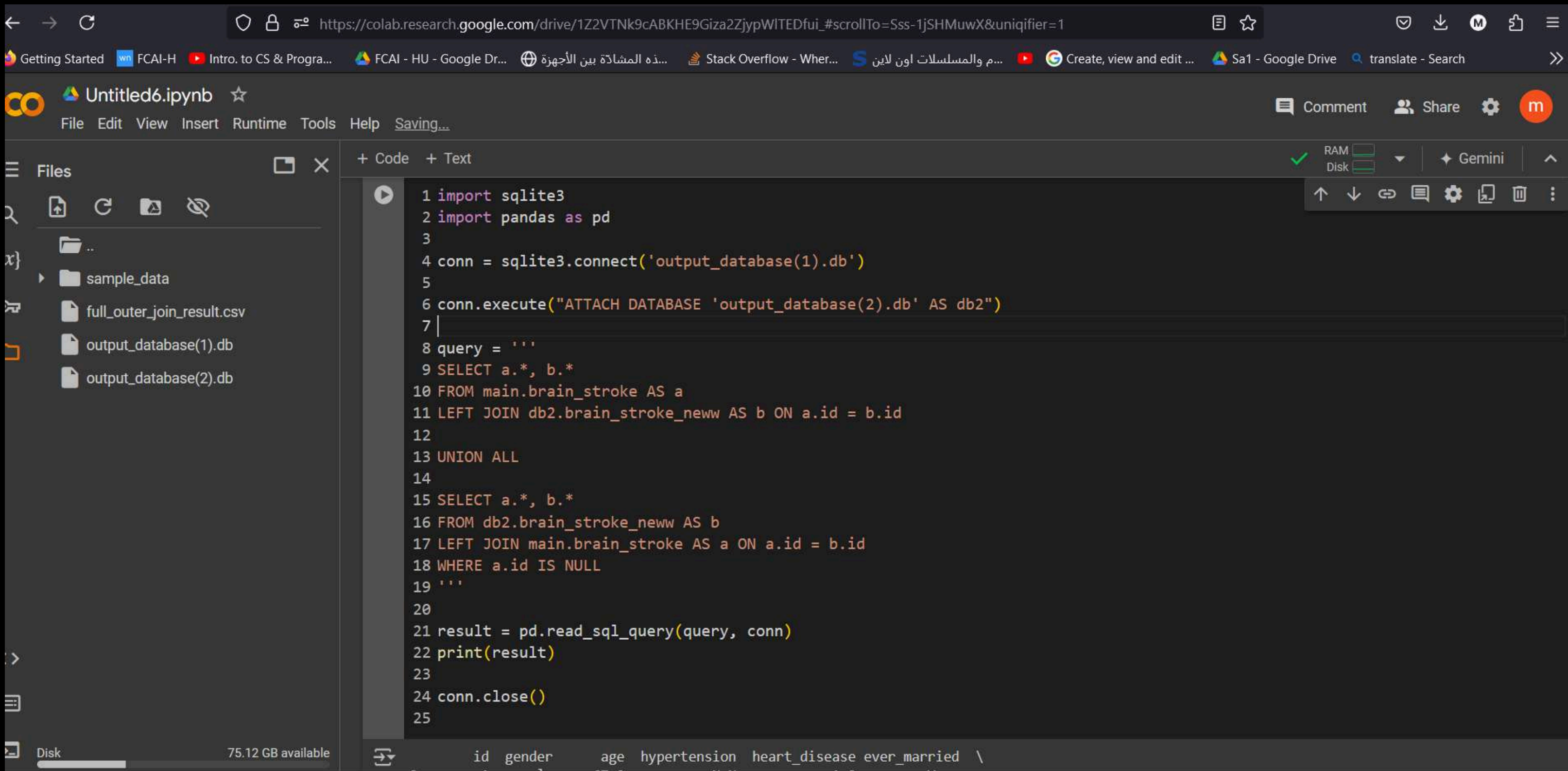
Below the code editor, there is a cell containing the command to read a CSV file:

```
[ ] 1 data=pd.read_csv('/content/brain_stroke_dirty.csv')
```

The right sidebar of the notebook is partially visible, showing icons for file management, search, and other tools.



- In short, the code merges data from two different tables in two different databases based on the id column, and displays all the data while keeping the unmatched rows from the second table.



The screenshot shows a Google Colab notebook titled "Untitled6.ipynb". The left sidebar displays a file explorer with a folder named "sample_data" containing files: "full_outer_join_result.csv", "output_database(1).db", and "output_database(2).db". The main area contains a Python script that connects to two SQLite databases, performs a left join of "main.brain_stroke" and "db2.brain_stroke_neww" on the "id" column, and uses pandas to read and print the result. The script is as follows:

```
1 import sqlite3
2 import pandas as pd
3
4 conn = sqlite3.connect('output_database(1).db')
5
6 conn.execute("ATTACH DATABASE 'output_database(2).db' AS db2")
7
8 query = '''
9 SELECT a.*, b.*
10 FROM main.brain_stroke AS a
11 LEFT JOIN db2.brain_stroke_neww AS b ON a.id = b.id
12
13 UNION ALL
14
15 SELECT a.*, b.*
16 FROM db2.brain_stroke_neww AS b
17 LEFT JOIN main.brain_stroke AS a ON a.id = b.id
18 WHERE a.id IS NULL
19 '''
20
21 result = pd.read_sql_query(query, conn)
22 print(result)
23
24 conn.close()
25
```

At the bottom of the notebook, a preview of the resulting DataFrame is shown with columns: id, gender, age, hypertension, heart_disease, ever_married, and \.

- We processed the gender column by replacing inconsistent values (e.g., 'MALE' with 'Male'), analyzed the distribution of gender categories, and determined the most frequent gender value (mode). Additionally, we verified the data types of the dataset columns

```
1 data.gender.value_counts()
```

gender	count
femael	1284
FEMALE	1271
mle	918
MALE	916
NAN	429

dtype: int64

```
1 data['gender']=data['gender'].replace('MALE','Male')
```

```
1 data.dtypes
```

gender	object

```
[18] 1 data.gender.value_counts()
```

gender	count
Female	2555
Male	1834
NAN	429

dtype: int64

```
1 data.gender.mode()[0]
```

'Female'

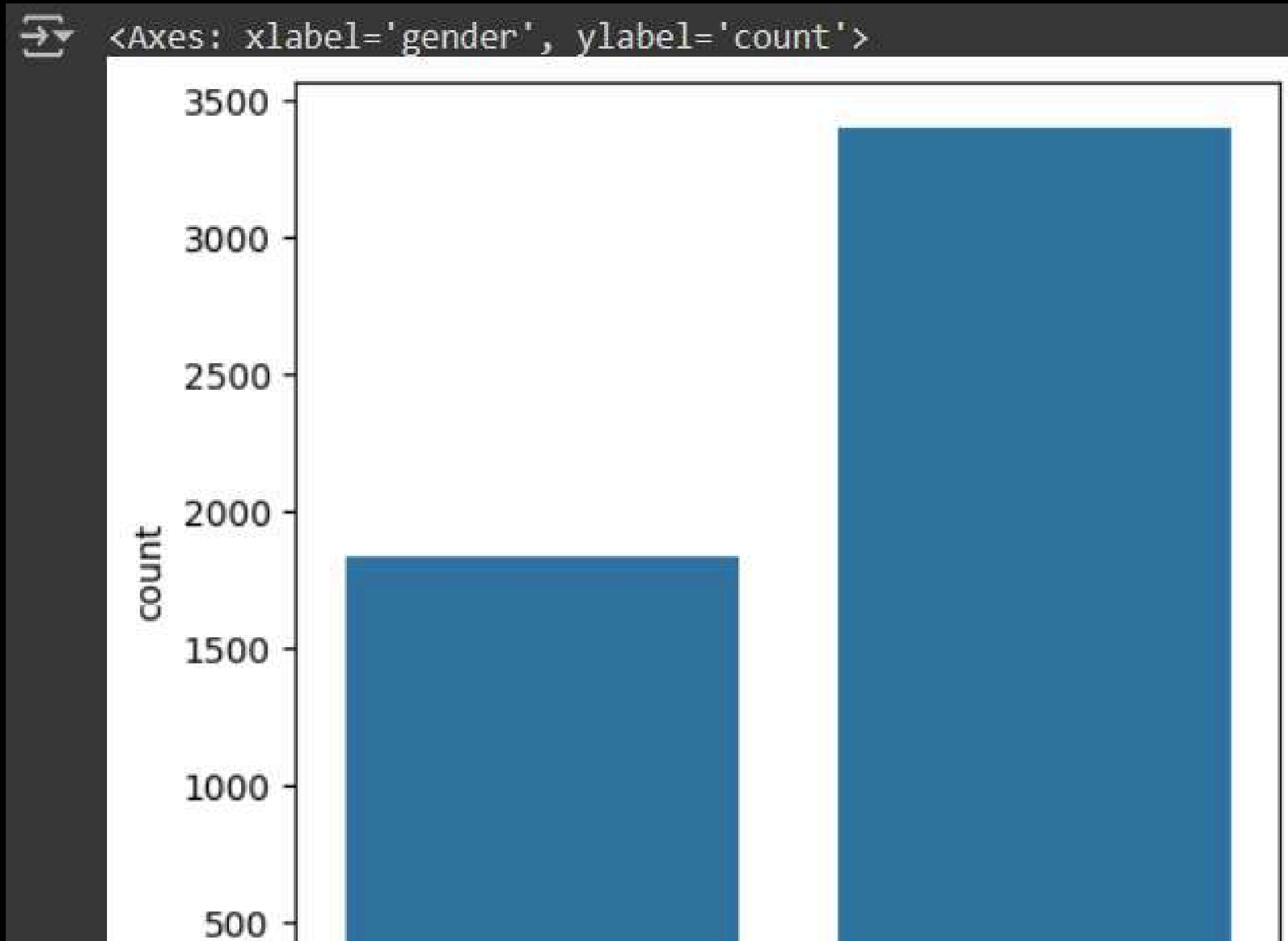
```
[25] 1 data['gender']=data['gender'].replace('NAN',data.gender.mode()[0])
```

```
1 data.gender.value_counts()
```

gender	count
Female	3398
Male	1834

dtype: int64

- We used Seaborn to visualize the distribution of gender categories by creating a count plot. The figure size was set to 5x5 for a clear representation.



- We converted the 'age' column to numeric values, handling any errors by coercing invalid entries into missing values. Missing ages were filled with the column's mean. We calculated the total number of missing values, as well as the maximum, minimum, mean, and median ages to gain a better understanding of the dataset

```
[50] 1 data['age'] = pd.to_numeric(data['age'], errors='coerce')
```

```
[51] 1 data["age"] = data["age"].astype(float)
```

```
1 data.dtypes
```

	0
gender	object
age	float64

```
data['age'].fillna(data['age'].mean(), inplace=True)
```

```
data.age.isna().sum()
```

```
[57] 1 data['age'].max()
```

```
149.0
```

```
1 data['age'].min()
```

```
-5.0
```

```
[53] 1 data.age.mean()
```

```
49.4743810194973
```

```
1 data.age.median()
```

```
47.0
```

- We calculated the Interquartile Range (IQR) to detect outliers in the 'age' column. Then, we set lower and upper bounds to identify extreme values. Outliers below and above these bounds were replaced with the mean age. Finally, we recalculated the maximum and minimum ages to ensure the data was cleaned of outliers.

```
1 Q1 = data['age'].quantile(0.25)
2 Q3 = data['age'].quantile(0.75)
3 IQR = Q3 - Q1
4 lower_bound = Q1 - 1.5 * IQR
5 upper_bound = Q3 + 1.5 * IQR
6 print(lower_bound, upper_bound)
7
```

```
6.5 76.5
```

```
1 data['age'] = data['age'].apply(lambda x: data['age'].mean() if x < lower_bound else x)
2 data['age'] = data['age'].apply(lambda x: data['age'].mean() if x > upper_bound else x)
```

```
1 data['age'].max()
```

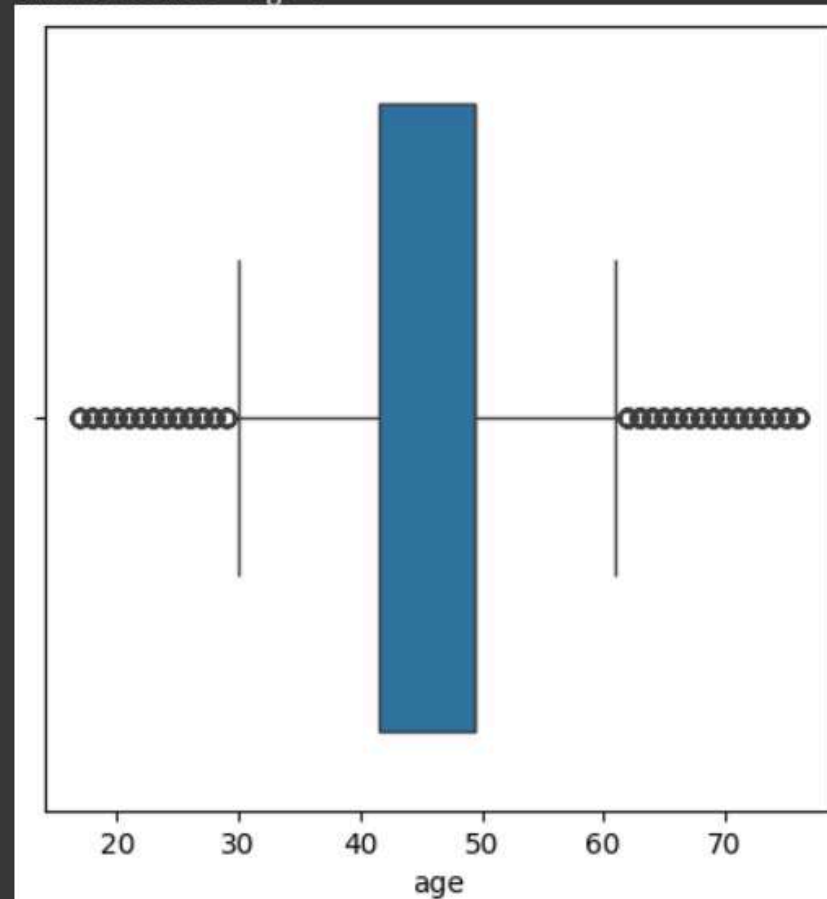
```
6.0
```

```
1 data['age'].min()
```

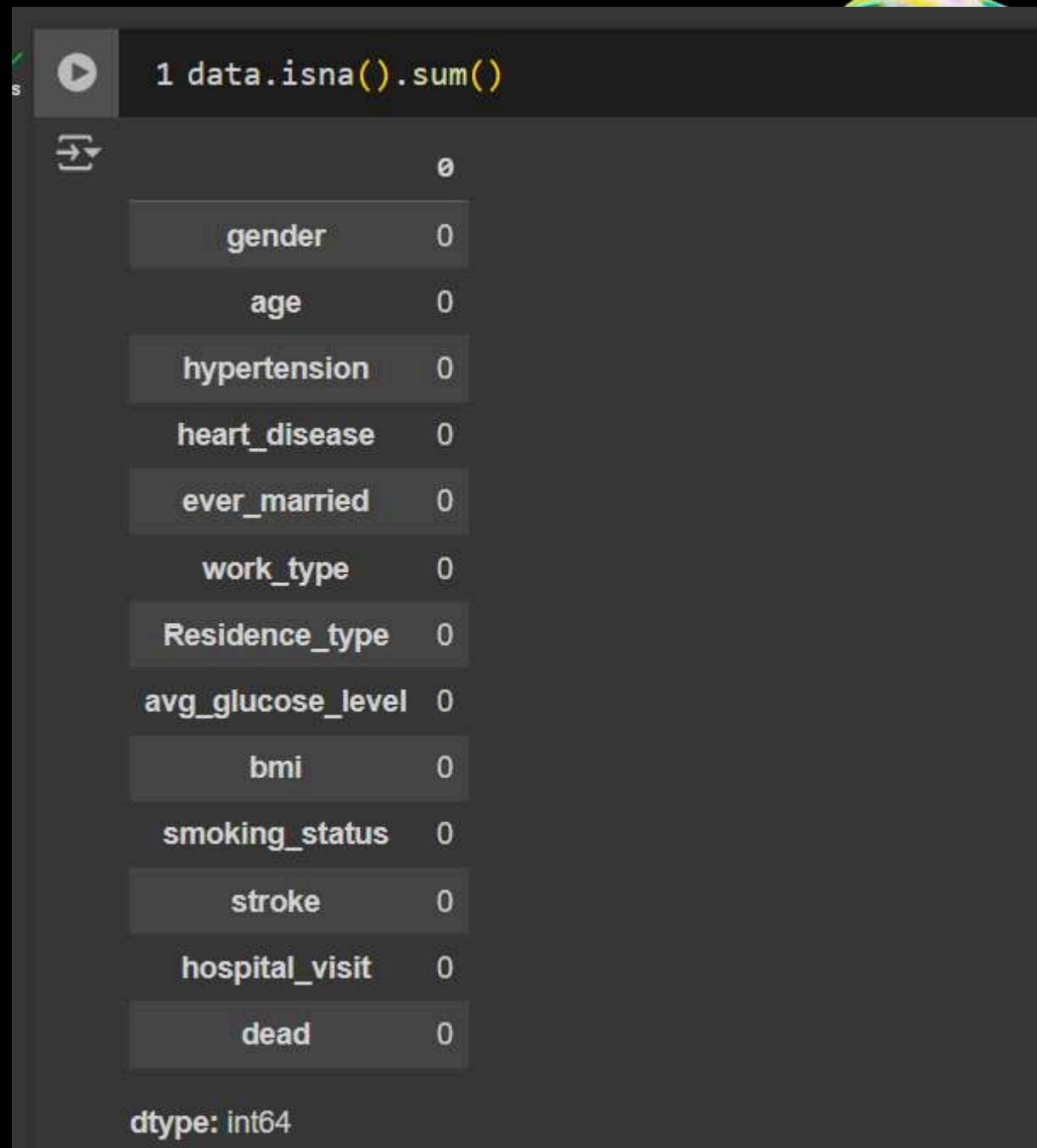
```
6.0
```

```
1 plt.figure(figsize=(5,5))
2 sns.boxplot(data=data, x='age')
```

<Axes: xlabel='age'>



- We identified the total number of missing values in the dataset and saved the results of a full outer join operation to a CSV file for further analysis



The image shows a Jupyter Notebook interface. At the top, there is a code cell with the text `1 data.isna().sum()`. Below the code cell, there is an output area displaying a table of missing values for various features. The table has two columns: the feature name and the count of missing values. All counts are 0. At the bottom of the output area, it says `dtype: int64`.

	0
gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
bmi	0
smoking_status	0
stroke	0
hospital_visit	0
dead	0

dtype: int64

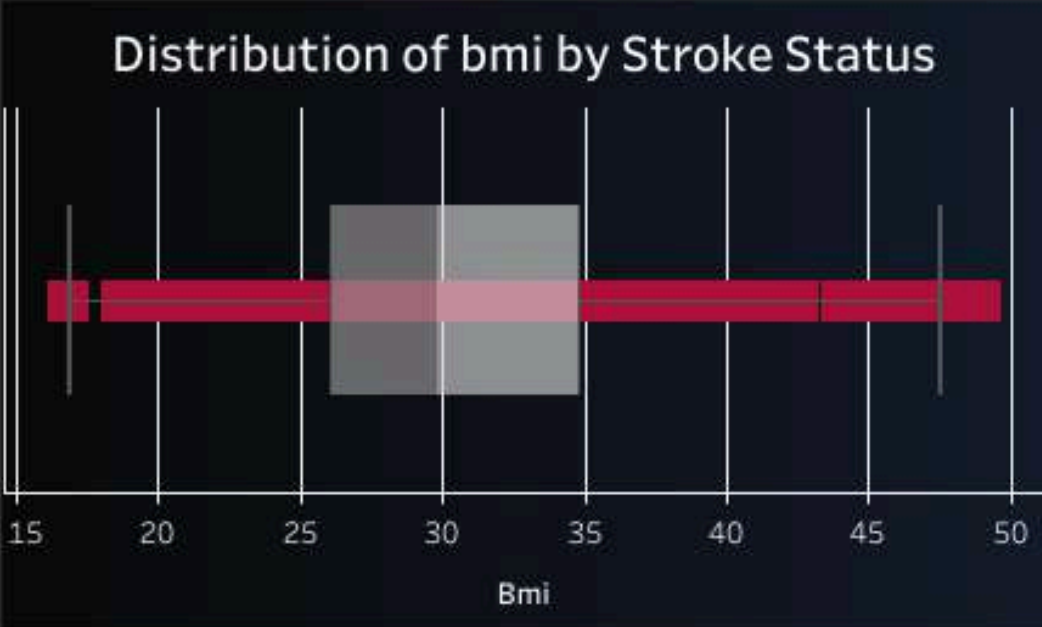
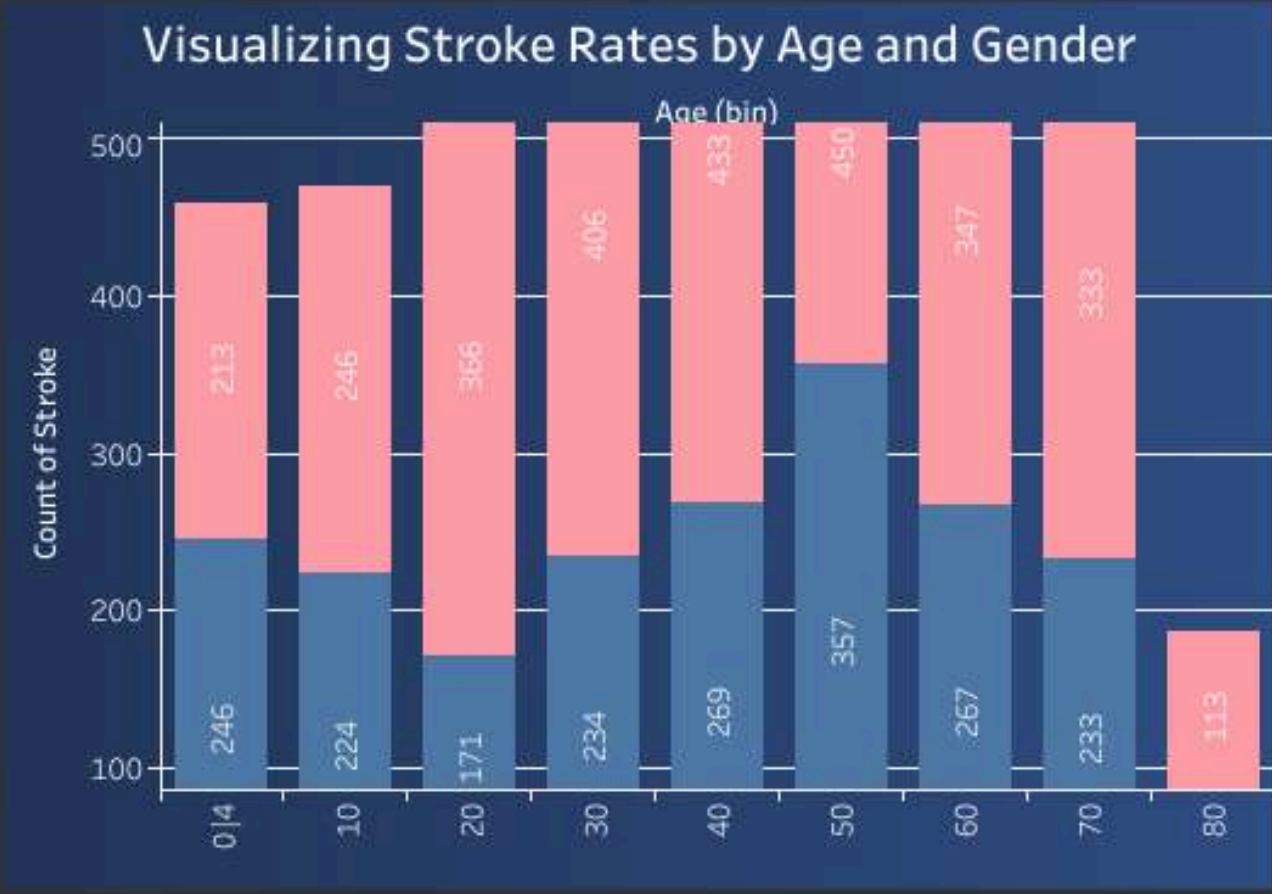
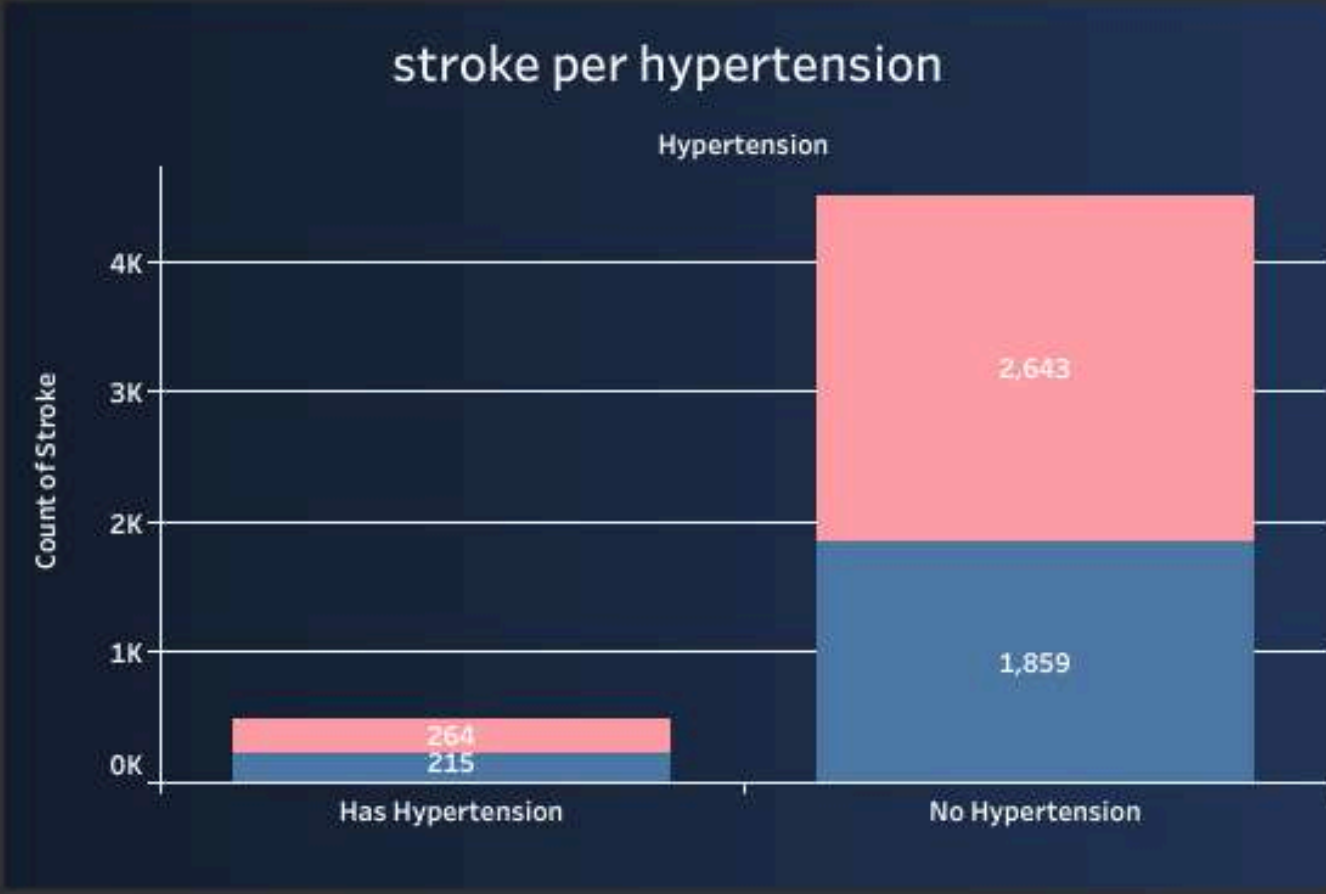
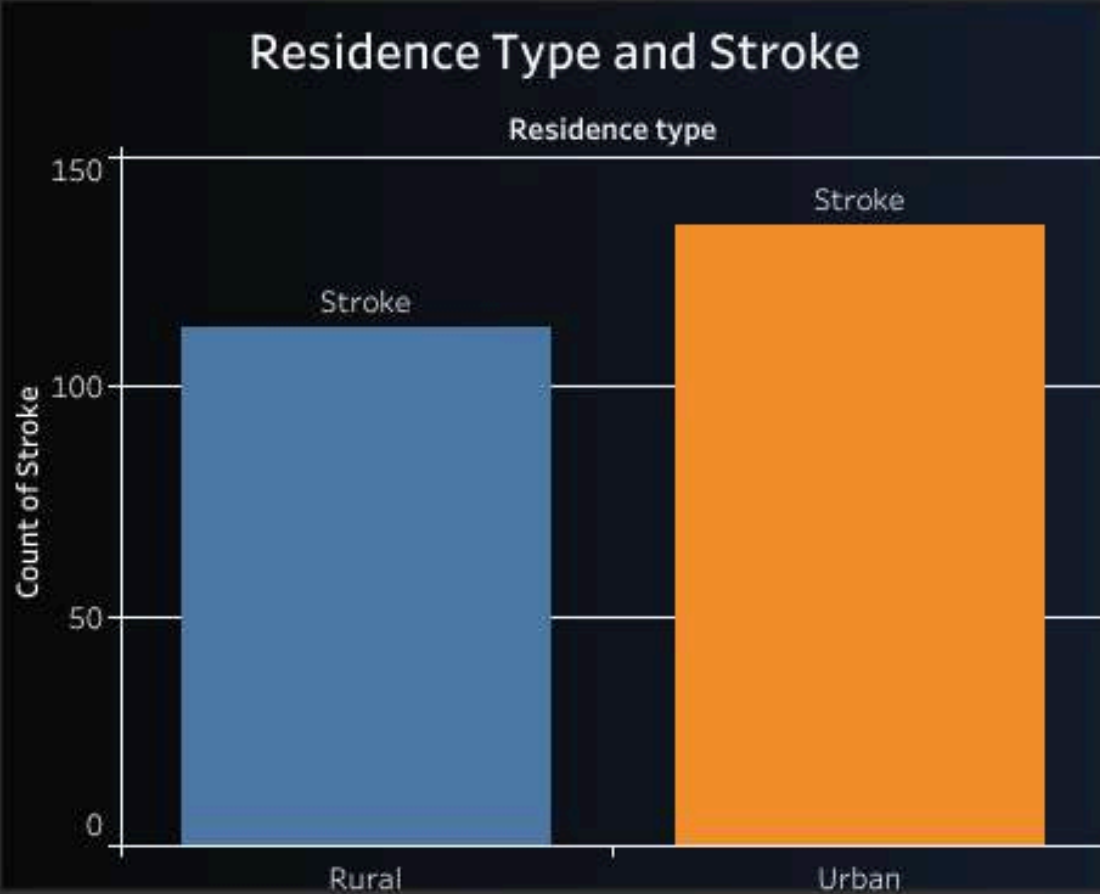
```
2 result.to_csv('full_outer_join_result.csv', index=False)  
3
```



WE USED TABLEAU

Brain Stroke Data Analysis

Avg. Age	Avg. Avg Glucose Level	Avg. Bmi	Avg. Hospital Visits	Number Of Records
43	106	28	6	4,981



Brain Stroke Data Analysis

Avg. Age

43

Avg. Avg Glucose Level

106

Avg. Bmi

28

Avg. Hospital Visits

6

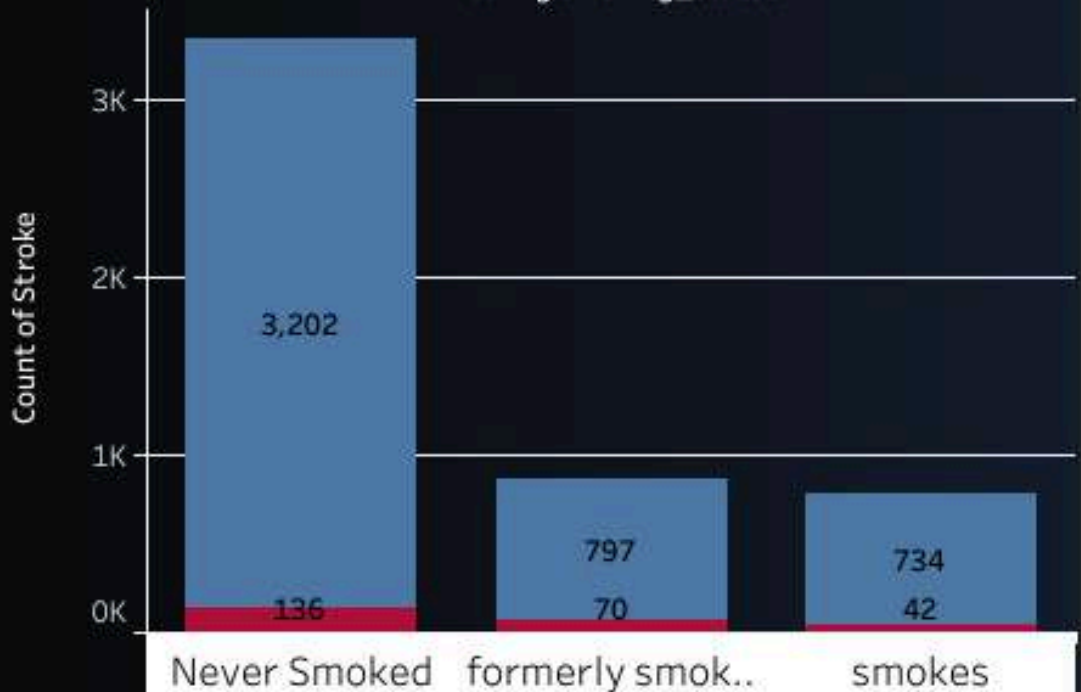
Number Of Records

4,981



Stroke Status by Smoking Status

filtering smoking_status

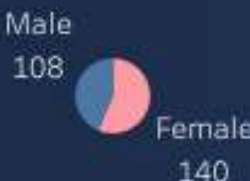
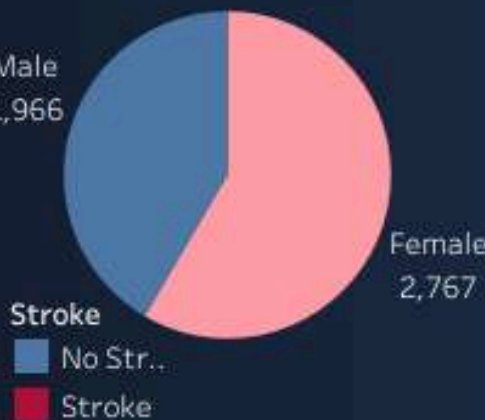


Proportion of Stroke Cases by Gender: Male vs Female

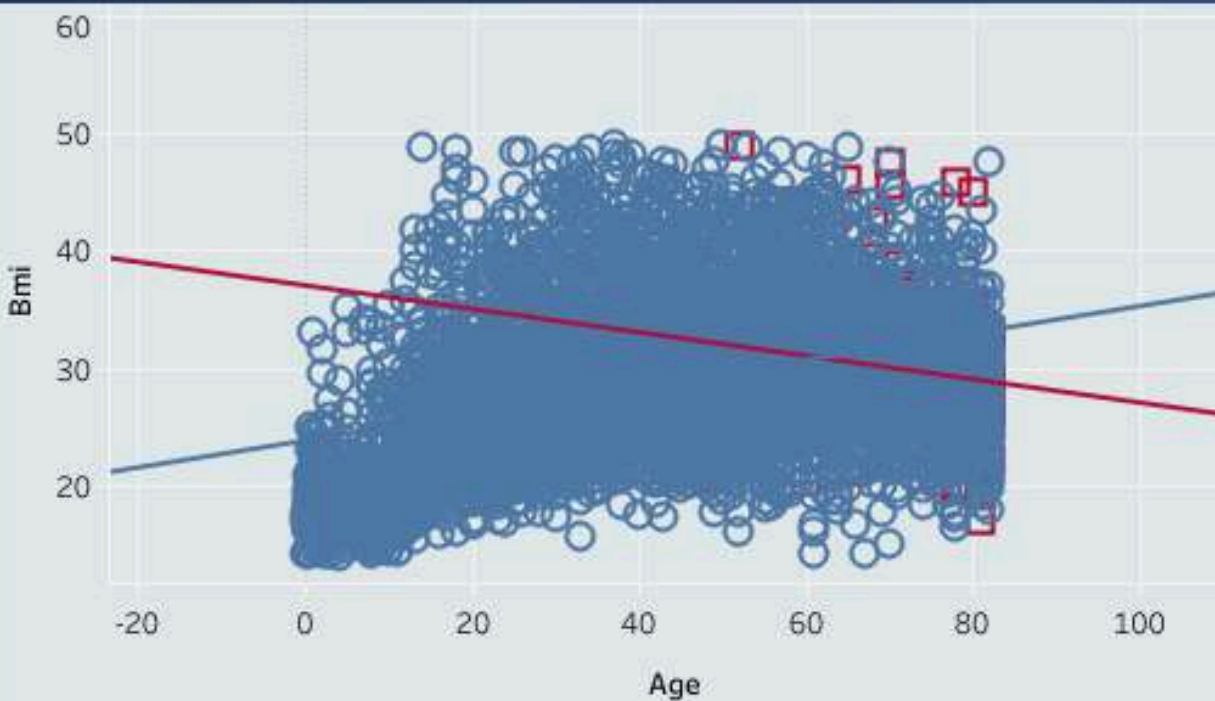
No Stroke

Stroke

Stroke

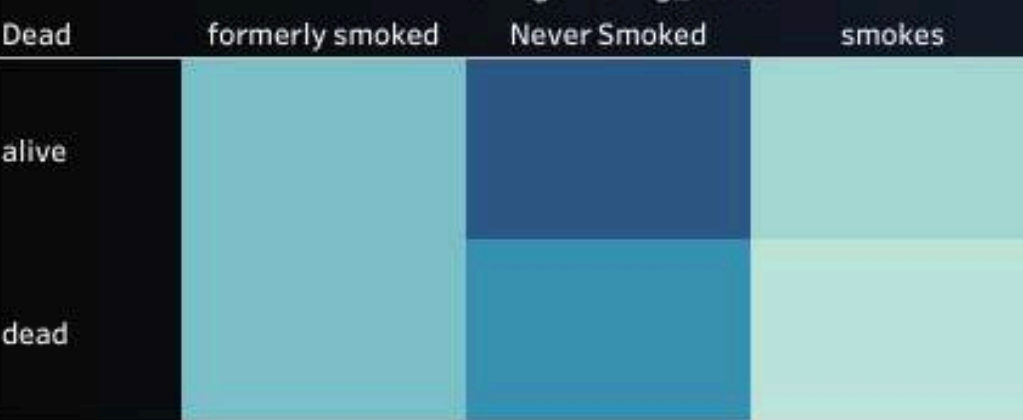


Exploring the Correlation of BMI and Age with Stroke Risk

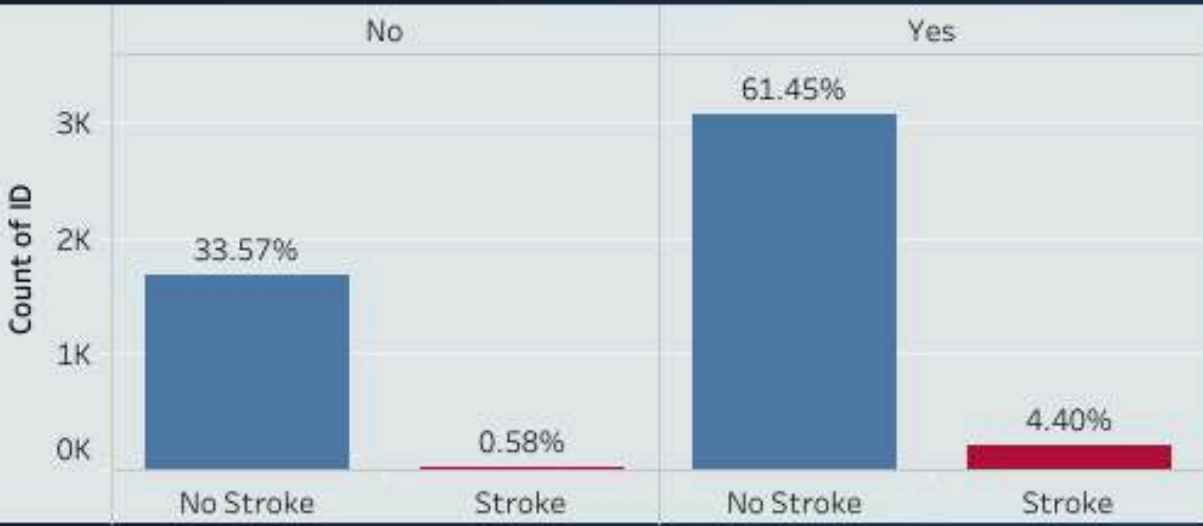


Exploring the Impact of Age and Avg Glucose Level on Stroke Risk

filtering smoking_status



Social status on stroke



Heart disease and stroke risk

	Stroke	
	No Stroke	Stroke
Heart Disease		
Has Heart Disease	82.91%	17.09%
No Heart Disease	95.73%	4.27%

RESULTS

- Key findings:

- Factor 1: Gender: Women are more likely to suffer from strokes compared to men, representing the majority of stroke cases.
- Factor 2: Body Mass Index (BMI): Obesity or being overweight (BMI between 25-35) is a significant risk factor that increases the likelihood of stroke.
- Factor 3: Age: Stroke rates clearly increase with age, with a peak in the 40 - 50 age group.
- Factor 4: Hospital Visits: There is a noticeable relationship between an increased number of hospital visits and higher stroke rates, suggesting that individuals with chronic health conditions may be at higher risk.
- Factor 5: Smoking: Current smokers and those who have previously quit are at a higher risk of stroke compared to non-smokers, making smoking cessation essential for reducing this risk.
- Factor 6: Residence: Urban residents exhibit a higher risk of stroke compared to rural residents, highlighting the need for health awareness initiatives in cities and improved healthcare access in rural areas.
- Factor 7: Individuals with hypertension are at a higher risk of stroke compared to those without it.
- Factor 8: Married individuals had a higher stroke incidence (4.40%) than unmarried individuals (0.58%).

RECOMMENDATIONS

• Based on the analysis, focus on preventing these key risk factors:

1. **Focus on Weight Management:** Encourage individuals, especially those with a BMI between 25-35 (overweight/obese), to adopt healthier lifestyles that include regular exercise and balanced diets to reduce the risk of stroke.
2. **Women's Health Initiatives:** Given that women are at a higher risk of stroke, especially those with higher BMI, targeted health programs should be developed to raise awareness and provide preventive measures for women.
3. **Regular Health Monitoring:** Encourage individuals with chronic health conditions or those who frequently visit hospitals to undergo regular stroke risk assessments, as frequent hospital visits may indicate underlying health risks.
4. **Promote Preventive Healthcare:** Educating people on managing hypertension, glucose levels, and maintaining an overall healthy lifestyle could significantly reduce stroke risk.
5. **Customized Care for Older Adults:** While age does not directly increase stroke risk, regular health check-ups and tailored preventive strategies for the elderly should be maintained, especially to monitor BMI and other risk factors like blood pressure.
6. **Implement comprehensive smoking cessation programs** that provide support and resources to help individuals quit smoking, along with public awareness campaigns about the dangers of smoking and its direct link to stroke risk.
7. **Enhance effective access to healthcare services** in urban areas through integrated health centers, and implement awareness programs in rural communities to address stroke risk factors and promote healthy lifestyles.
8. **Stroke occurrence is higher among married individuals (4.40%).** Investigating factors like stress or lifestyle in married life could help in stroke prevention

By focusing on these areas, the risk of stroke can be significantly mitigated, improving overall health outcomes.

• Apply these insights in healthcare policy to reduce stroke risks.



CONCLUSION

- Summary of the results and their practical applications.
How this analysis can help in improving healthcare and stroke prevention.



THANK YOU

for your time and attention

