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

INTRODUCTION

- Stroke Is A Major Health Concernglobally.
- 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

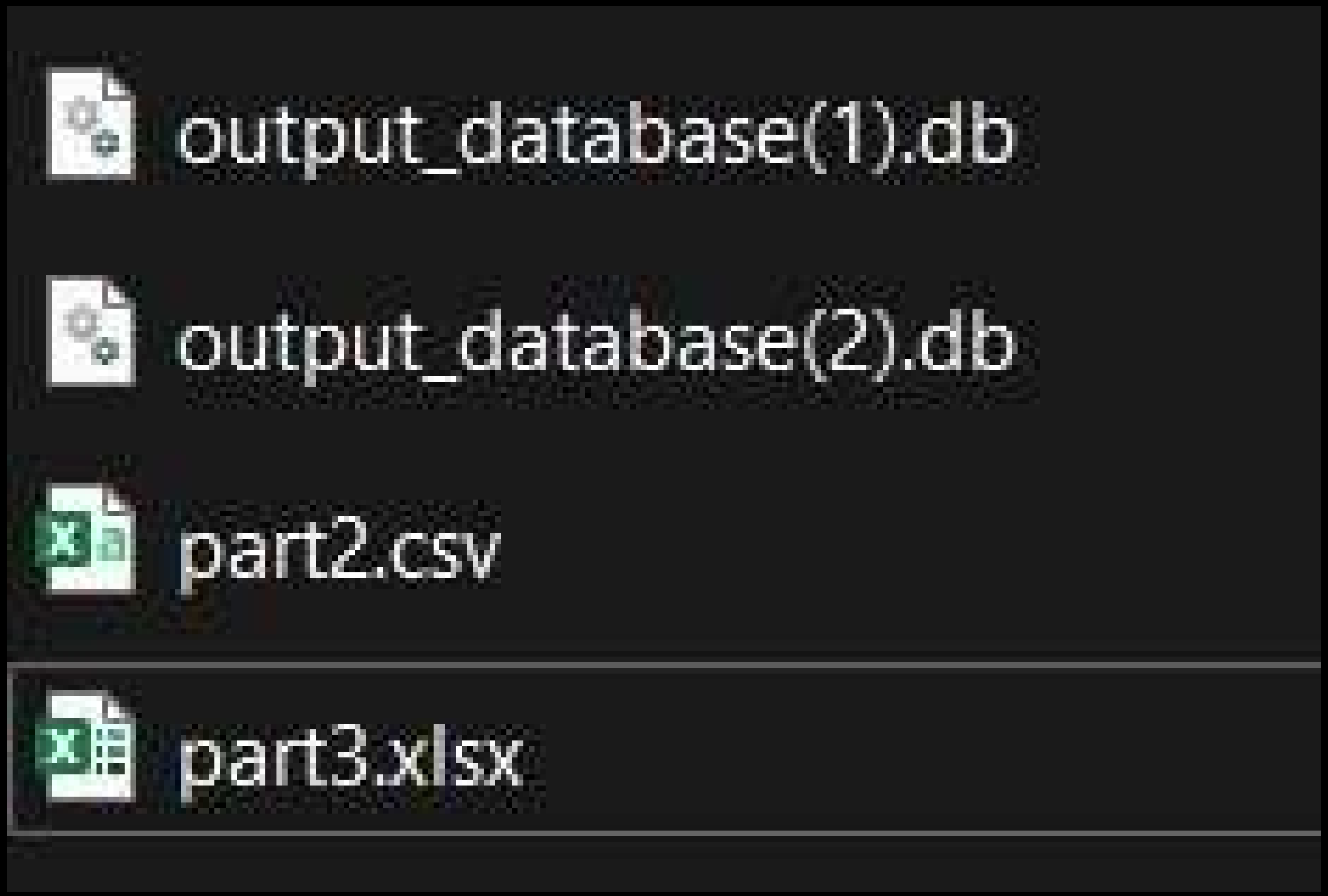
Using correlation or regression analysis to identify significant relationships.

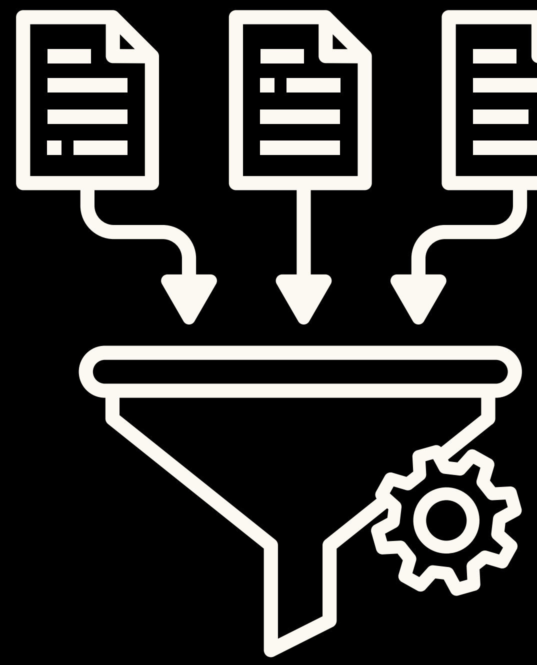
- Split stroke data into three parts and save each part in a different format: CSV, Excel, and SQL.

```python
file_path = '/content/brain_stroke.csv'
df = pd.read_csv(file_path)


chunk_size = math.ceil(len(df) / 3)


df1 = df.iloc[:chunk_size]
df2 = df.iloc[chunk_size:2*chunk_size]
df3 = df.iloc[2*chunk_size:]

df2.to_csv('part2.csv', index=False)
df3.to_excel('part3.xlsx', index=False)
```

```python
import sqlite3

conn = sqlite3.connect('output_database.db')

df1.to_sql('brain_stroke', conn, if_exists='replace', index=False)
conn.close()
```

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

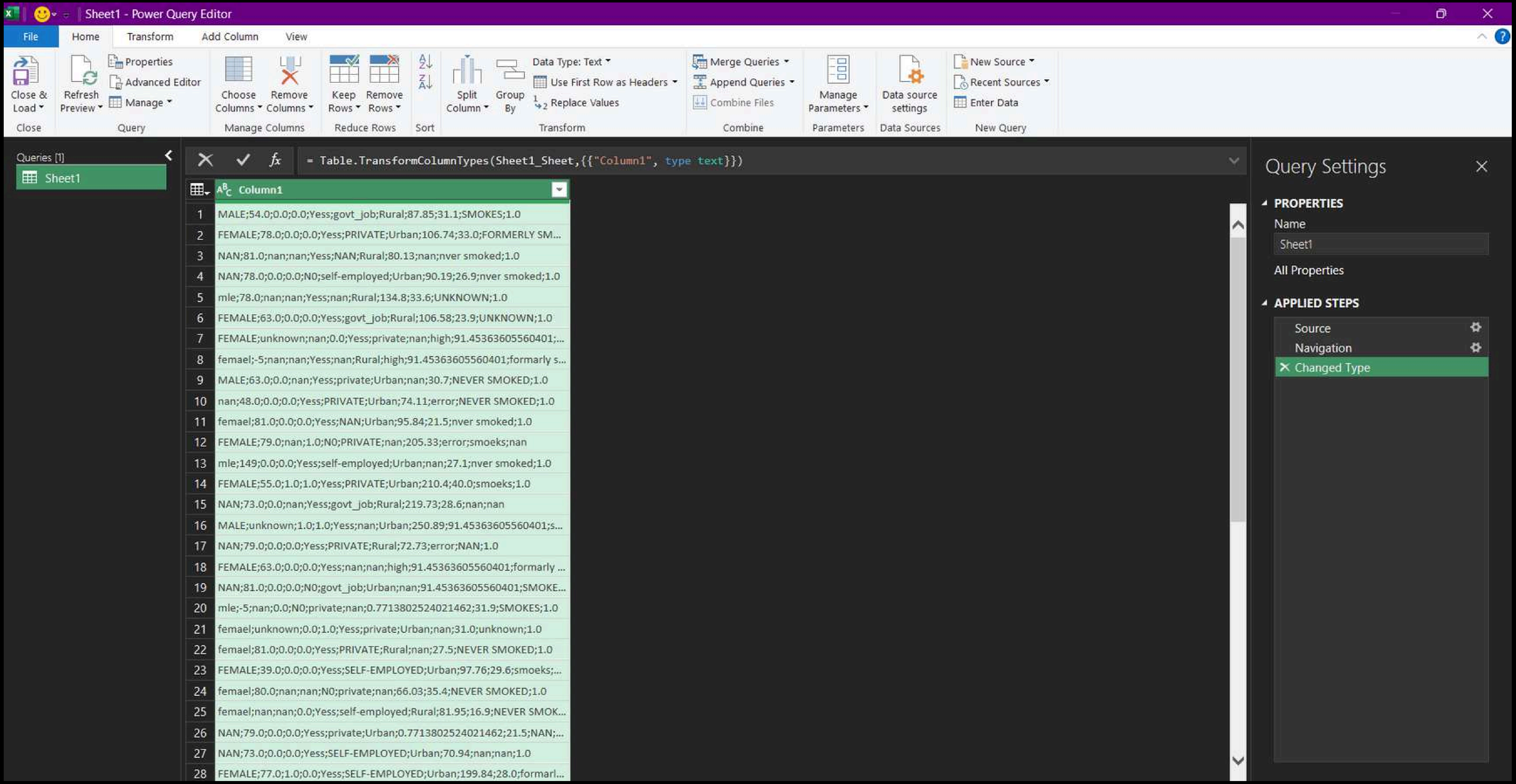




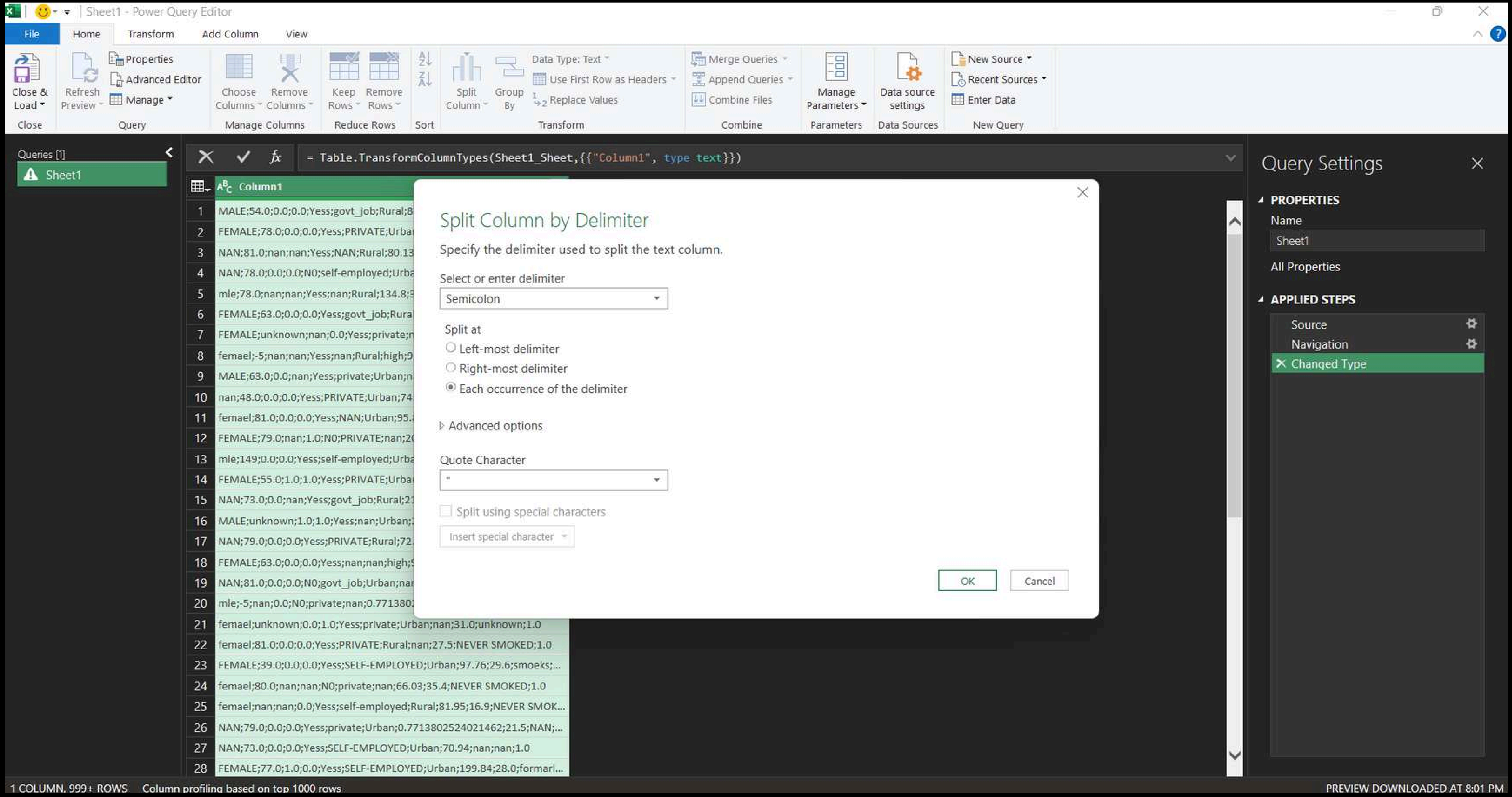
- The purpose of this line of code is to create a new DataFrame (df) that combines all the stroke data pieces into one continuous dataset, by arranging them vertically. This is useful for performing analyses or operations on the entire dataset at once after splitting it into parts.

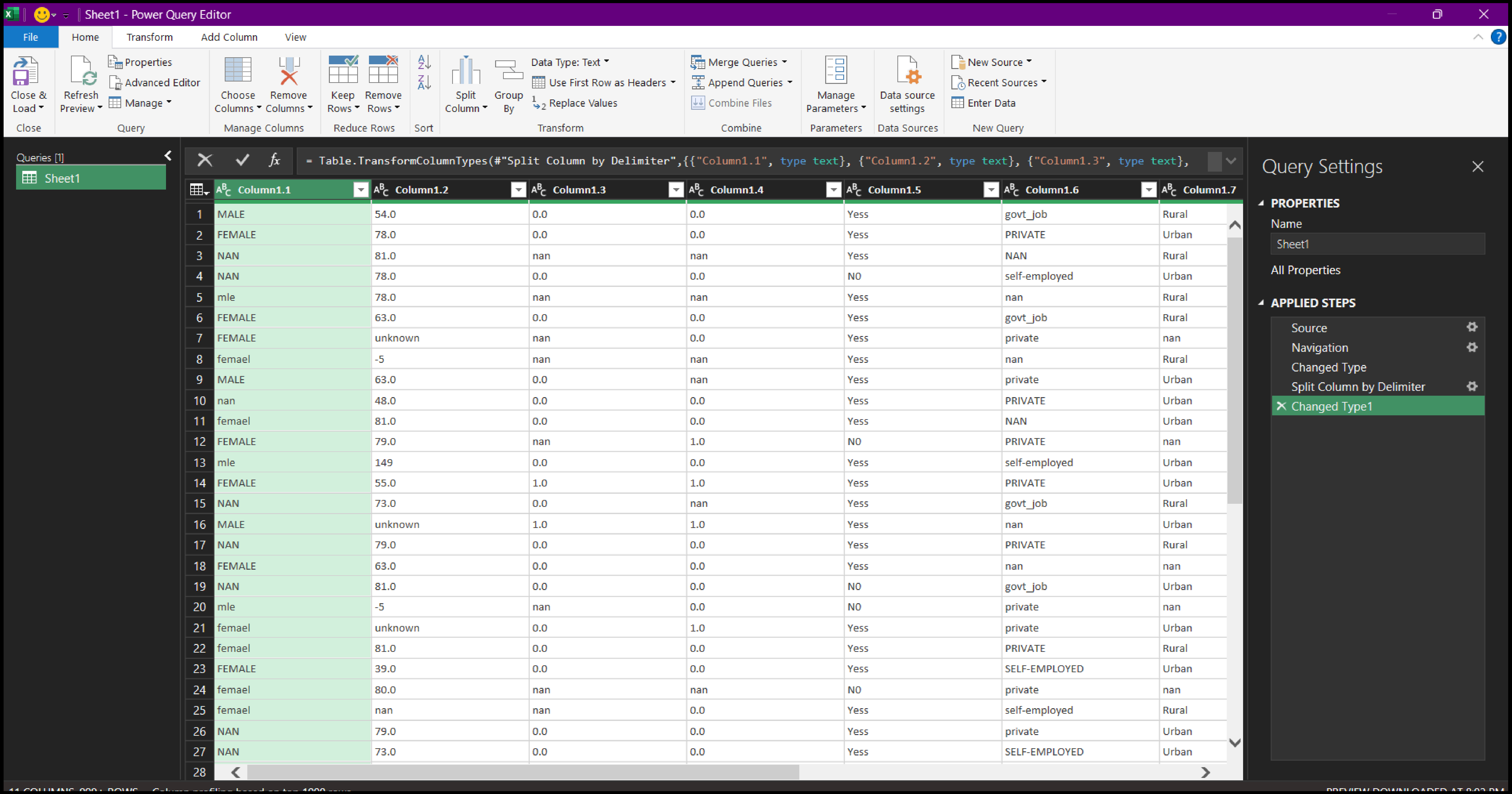
```
[ ]   1 df=pd.concat([df1,df2,df3],axis=0)
```

- We used Power Query.



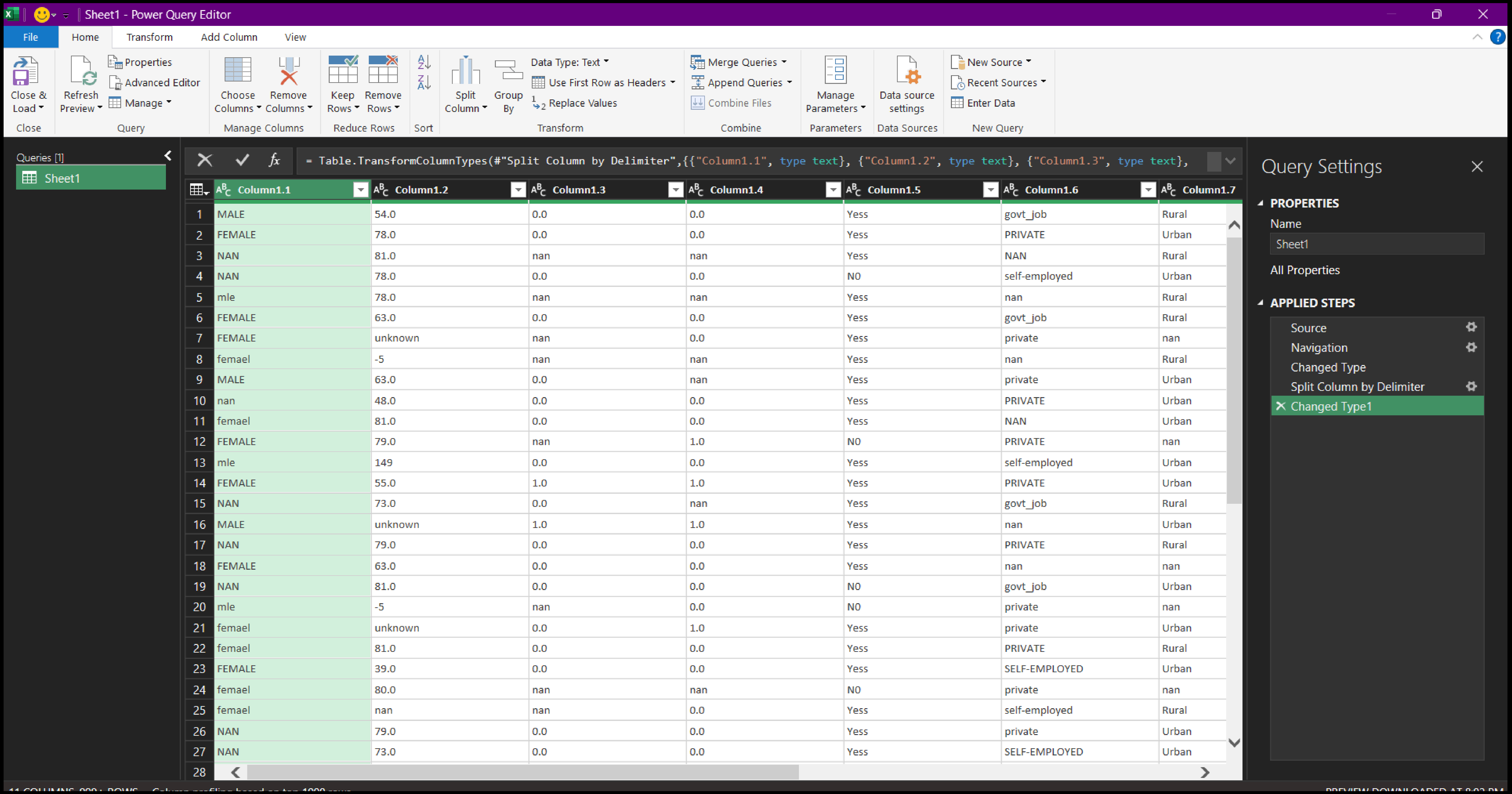
- In Excel, we used the column splitting method to organize the data.





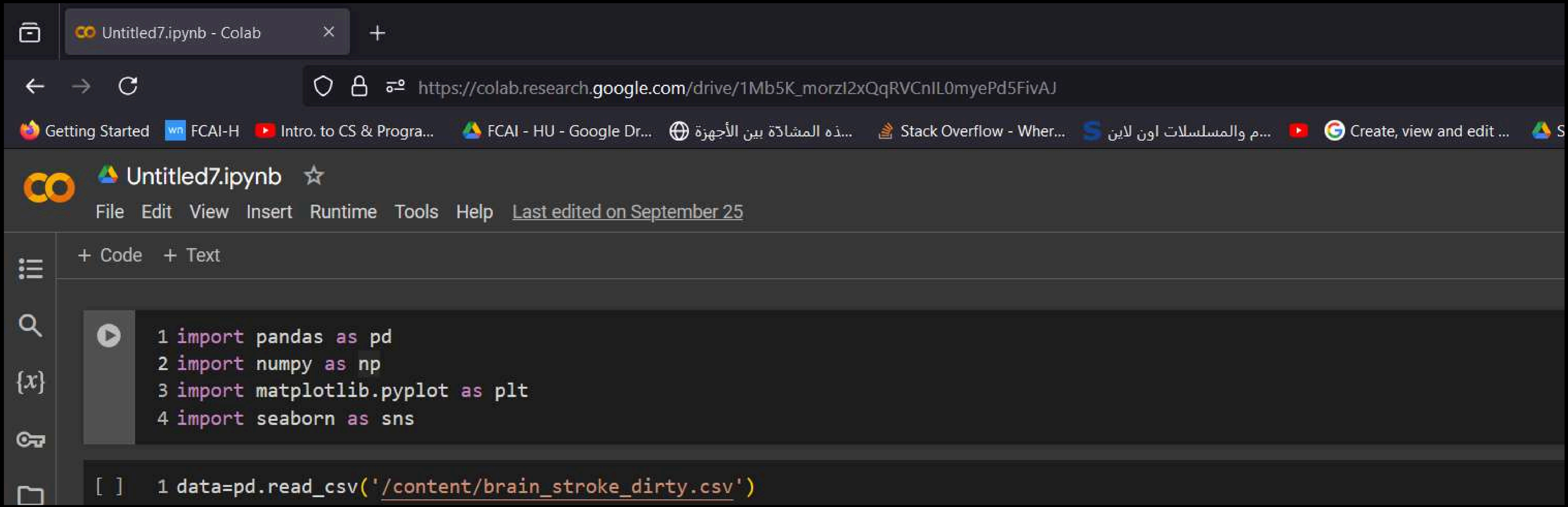
	Column1.1	Column1.2	Column1.3	Column1.4	Column1.5	Column1.6	Column1.7
1	MALE	54.0	0.0	0.0	Yess	govt_job	Rural
2	FEMALE	78.0	0.0	0.0	Yess	PRIVATE	Urban
3	NAN	81.0	nan	nan	Yess	NAN	Rural
4	NAN	78.0	0.0	0.0	NO	self-employed	Urban
5	mle	78.0	nan	nan	Yess	nan	Rural
6	FEMALE	63.0	0.0	0.0	Yess	govt_job	Rural
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							

- Change data type

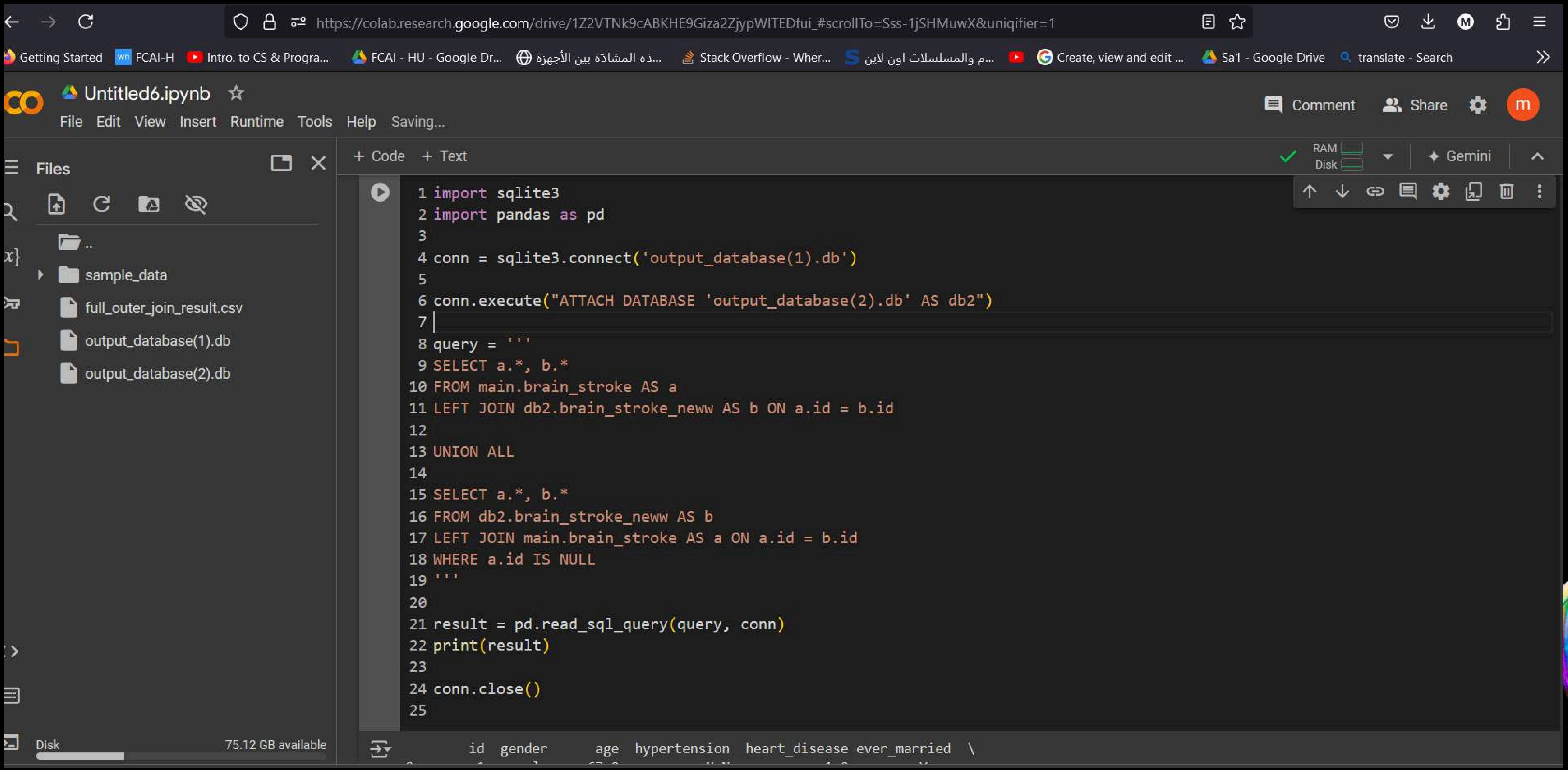


gender	age	hypertension	heart_disease	ever_married	work_type	residence_type	avg_glucose_level	bmi	smoking_status	stroke
MALE	54.0	0.0	0.0	Yess	govt_job	Rural	87.85	31.1	SMOKES	1.0
FEMALE	78.0	0.0	0.0	Yess	PRIVATE	Urban	106.74	33.0	FORMERLY SMOKED	nan
NAN	81.0	nan	nan	Yess	NAN	Rural	80.13	nan	nver smoked	1.0
NAN	78.0	0.0	0.0	N0	self-employed	Urban	90.19	26.9	nver smoked	1.0
mle	78.0	nan	nan	Yess	nan	Rural	134.8	33.6	UNKNOWN	1.0
FEMALE	63.0	0.0	0.0	Yess	govt_job	Rural	106.58	23.9	UNKNOWN	1.0
FEMALE	unknown	nan	0.0	Yess	private	nan	high	91.45363605560401	NAN	1.0
femael	-5	nan	nan	Yess	nan	Rural	high	91.45363605560401	formarly smoked	1.0
MALE	63.0	0.0	nan	Yess	private	Urban	nan	30.7	NEVER SMOKED	1.0
nan	48.0	0.0	0.0	Yess	PRIVATE	Urban	74.11	error	NEVER SMOKED	1.0
femael	81.0	0.0	0.0	Yess	NAN	Urban	95.84	21.5	nver smoked	1.0
FEMALE	79.0	nan	1.0	N0	PRIVATE	nan	205.33	error	smoeks	nan
mle	149	0.0	0.0	Yess	self-employed	Urban	nan	27.1	nver smoked	1.0
FEMALE	55.0	1.0	1.0	Yess	PRIVATE	Urban	210.4	40.0	smoeks	1.0
NAN	73.0	0.0	nan	Yess	govt_job	Rural	219.73	28.6	nan	nan
MALE	unknown	1.0	1.0	Yess	nan	Urban	250.89	91.45363605560401	smoeks	1.0
NAN	79.0	0.0	0.0	Yess	PRIVATE	Rural	72.73	error	NAN	1.0
FEMALE	63.0	0.0	0.0	Yess	nan	nan	high	91.45363605560401	formarly smoked	1.0
NAN	81.0	0.0	0.0	N0	govt_job	Urban	nan	91.45363605560401	SMOKES	1.0
mle	-5	nan	0.0	N0	private	nan	0.7713802524021462	31.9	SMOKES	1.0
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	1.0
femael	80.0	nan	nan	N0	private	nan	66.03	35.4	NEVER SMOKED	1.0
femael	nan	nan	0.0	Yess	self-employed	Rural	81.95	16.9	NEVER SMOKED	1.0
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



- 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.



- 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')
```

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

gender	count
Female	2555
Male	1834
NAN	429

dtype: int64

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

'Female'

```
1 data.dtypes
```

	0
gender	object

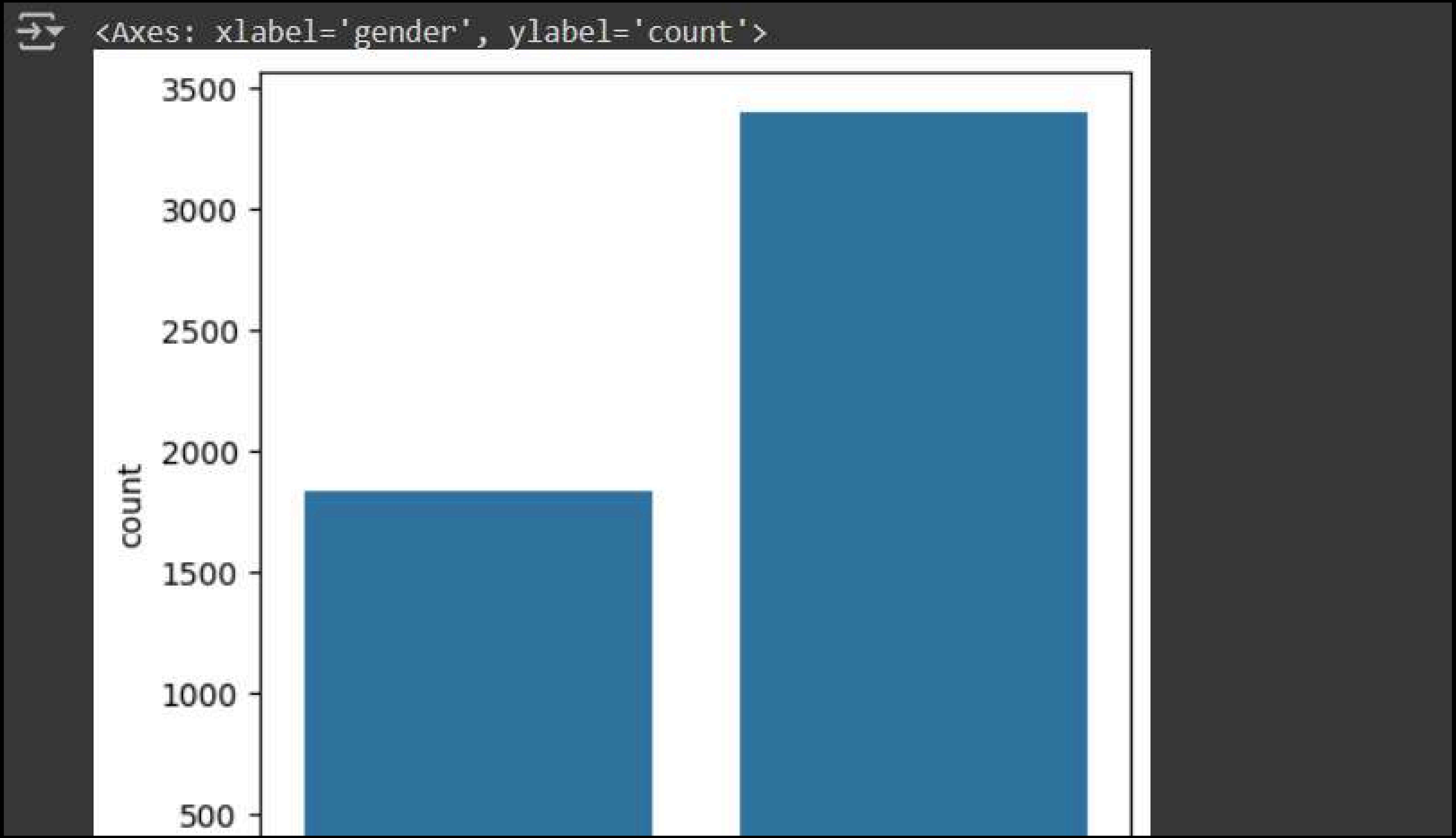
```
[25] 1 data['gender']=data['gender'].replace('NAN',data.gender.mode()[0])
```

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

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

```
data.dtypes
```

	0
gender	object
age	float64

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

```
a.age.isna().sum()
```

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

149.0

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

-5.0

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

49.4743810194973

```
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.

```
Q1 = data['age'].quantile(0.25)
Q3 = data['age'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
print(lower_bound,upper_bound)
```

```
.5 76.5
```

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

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

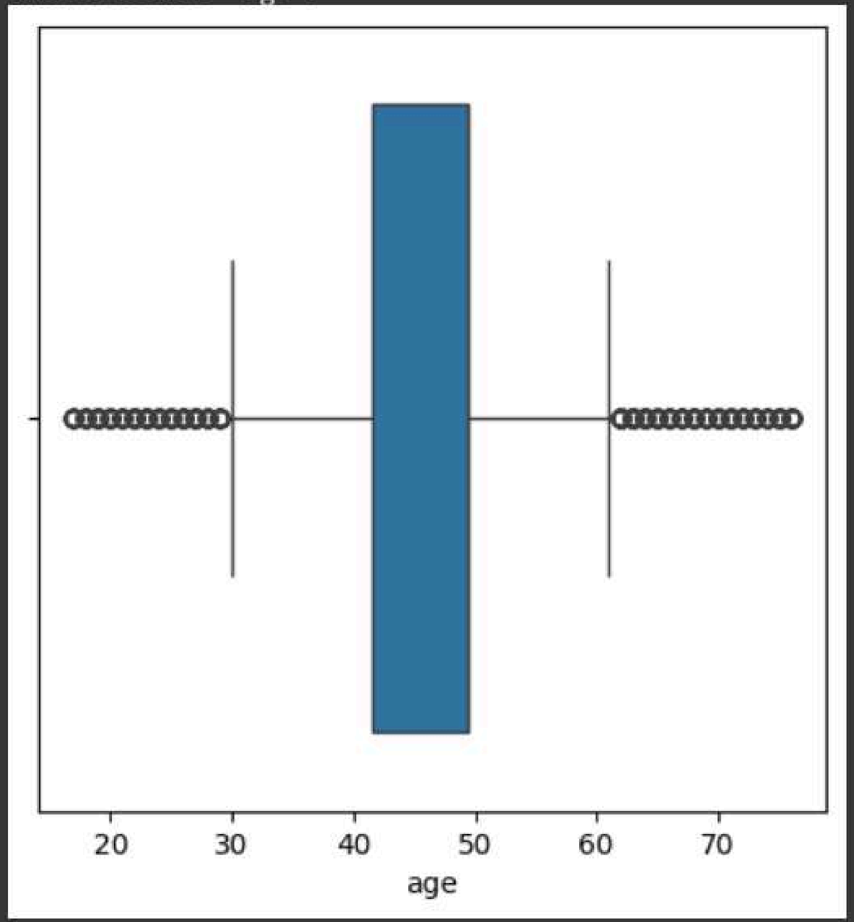
```
.0
```

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

```
.0
```

```
plt.figure(figsize=(5,5))
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

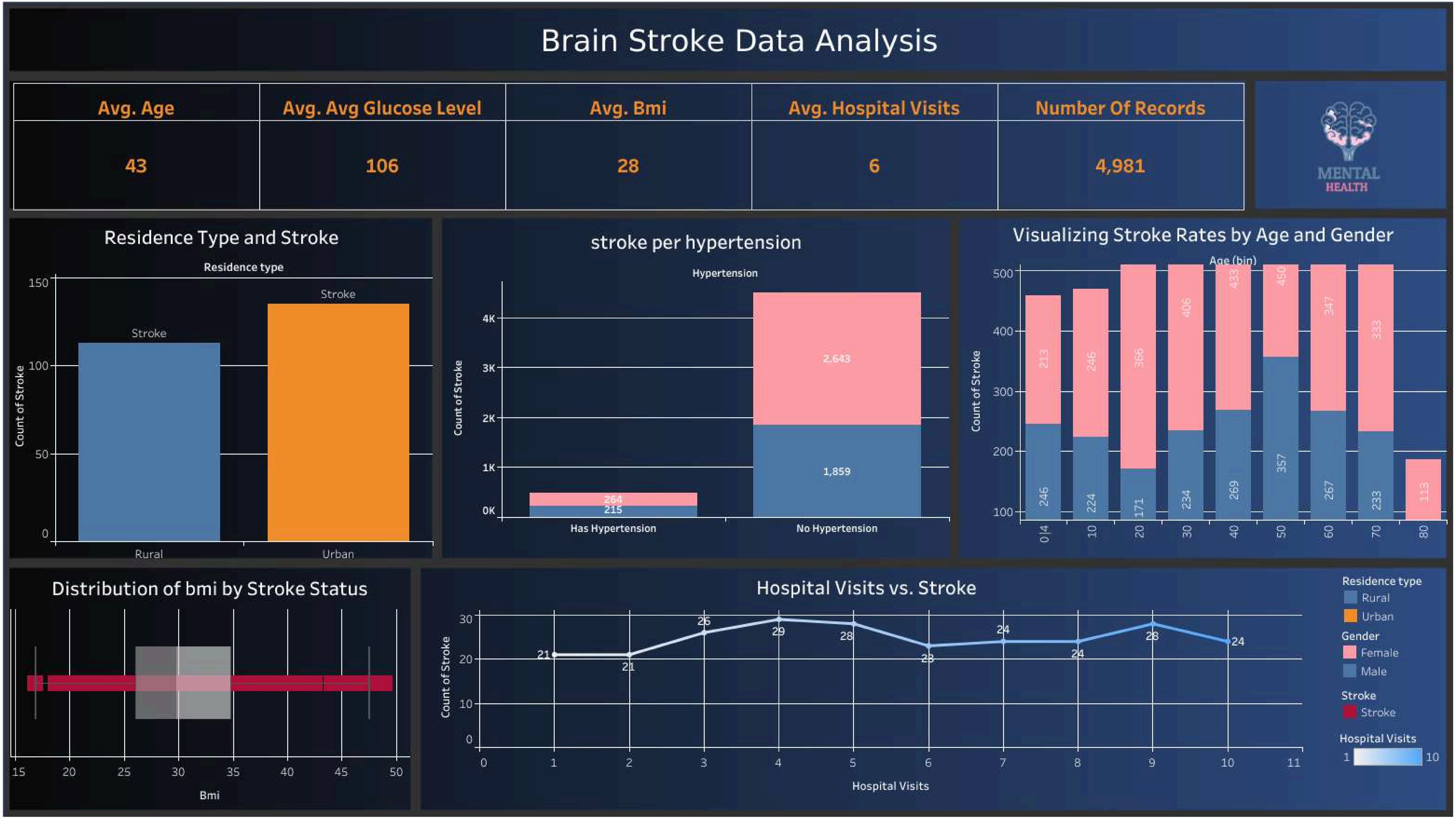
```
data.isna().sum()
```

	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

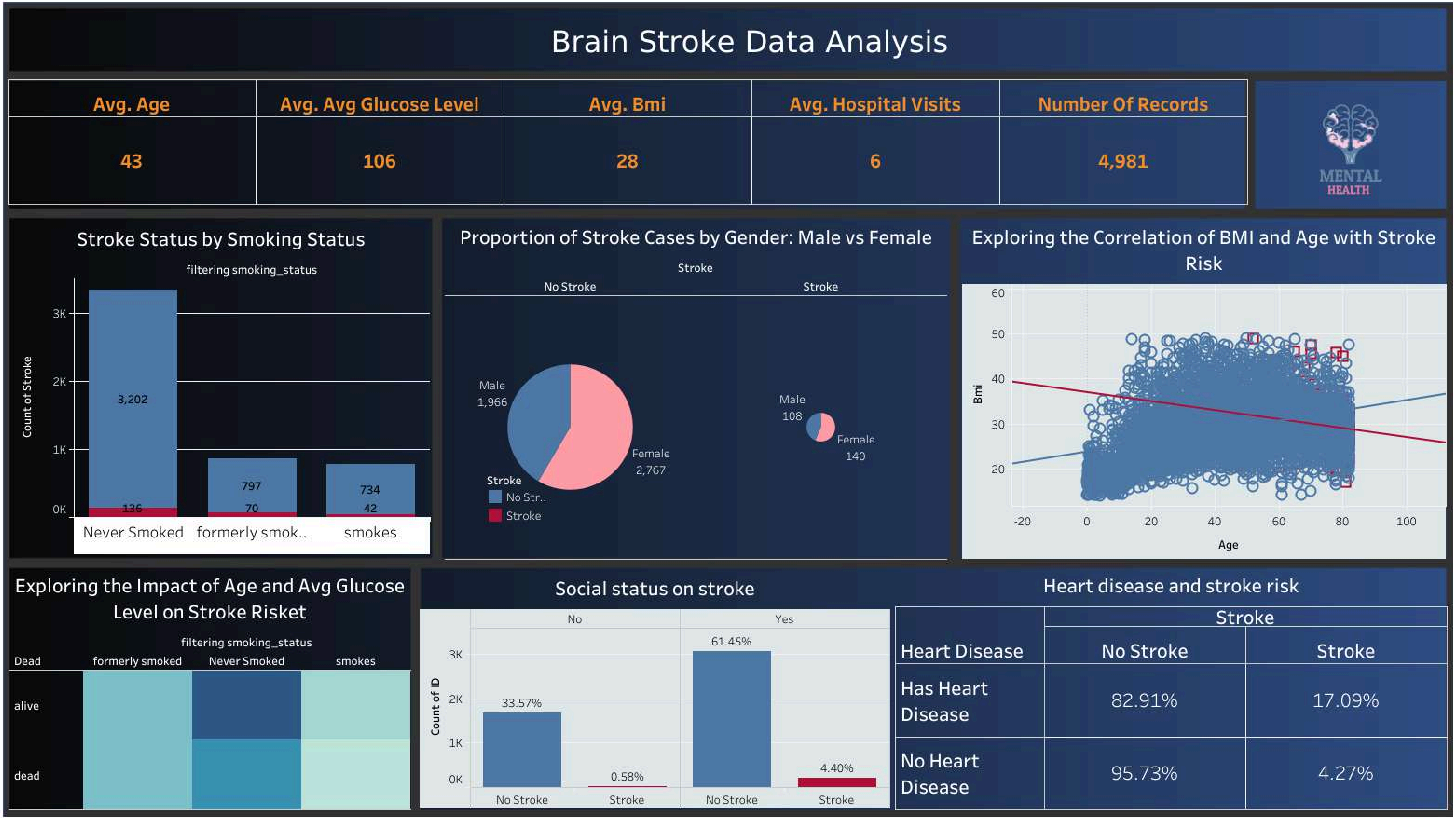
dtype: int64

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

WE USED TABLEAU



Brain Stroke Data Analysis
Avg. Age
43
Avg. Avg Glucose Level
106
Avg. Bmi
28
Avg. Hospital Visits
6
Number Of Records
4,981
MENTAL
HEALTH
Residence Type and Stroke
Residence type
Stroke
Stroke
Count of Stroke
Rural
Urban
stroke per hypertension
Hypertension
2,643
1,859
264
215
Count of Stroke
Has Hypertension
No Hypertension
Visualizing Stroke Rates by Age and Gender
Age (bin)
Count of Stroke
Distribution of bmi by Stroke Status
Bmi
Hospital Visits vs. Stroke
Count of Stroke
Hospital Visits
Residence type
Rural
Urban
Gender
Female
Male
Stroke
Stroke
Hospital Visits


Brain Stroke Data Analysis
Avg. Age
43
Avg. Avg Glucose Level
106
Avg. Bmi
28
Avg. Hospital Visits
6
Number Of Records
4,981
MENTAL
HEALTH
Stroke Status by Smoking Status
filtering smoking_status
Count of Stroke
3K
2K
1K
0K
3,202
136
797
70
734
42
Never Smoked
formerly smok..
smokes
Proportion of Stroke Cases by Gender: Male vs Female
Stroke
No Stroke
Stroke
Male
1,966
Female
2,767
Male
108
Female
140
Stroke
No Str..
Stroke
Exploring the Correlation of BMI and Age with Stroke Risk
Bmi
60
50
40
30
20
-20
0
20
40
60
80
100
Age
Exploring the Impact of Age and Avg Glucose Level on Stroke Risket
filtering smoking_status
Dead
formerly smoked
Never Smoked
smokes
alive
dead
Social status on stroke
No
Yes
Count of ID
3K
2K
1K
0K
33.57%
0.58%
61.45%
4.40%
No Stroke
Stroke
No Stroke
Stroke
Heart disease and stroke risk
Stroke
Heart Disease
No Stroke
Stroke
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

<u>• Based on the analysis, focus on preventing these key risk factors:</u>

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