AMRITA VISHWA VIDYAPEETHAM

Kochi Campus

ASSIGNMENT

Subject: DATA MINING & APPLICATIONS

Title of Assignment: LAB SHEET

Submitted by: SREELAKSHMI V

Roll No: KH.EN.P2MCA25158

Course: MCA - AI & DS / 2025

Semester: SEM 1

Submitted to: Dr. SANYEETHA I

Date of Submission: 4-11-25

Signature of Faculty

LAB ASSIGNMENT

1. Download any data set in CSV format from Kaggle.com and customize the data set with your own requirement
2. Write a summary of the dataset.
3. Do the following in the dataset

a) Load the dataset into a data frame and print the entire dataset
b) Display the size of the dataset
c) Print the first 10 rows of the dataset
d) Print the last 10 rows of the dataset
e) Display the information about data in the dataset
f) Display the statistical view of the dataset
g) Display the datatype of data in the dataset
h) Print the null values of all columns in the dataset
i) Display the no. of unique values in the dataset
j) Print the % of null values in each columns.
k) Fill the missing values in the dataset with mean of the columns.
l) Sort the values in the dataset based on the third columns
m) Rename the second column name with a new name.
n) Print the categorical and numerical columns in the dataset.
o) Drop the outliers in the dataset if any and show the dataset

ANSWERS

2. Summary of the Dataset

The data set contains 5,110 patient records. In this there are 12 columns, including demographics, clinical factors and life style information. These are used to predict stroke.

Source : Stroke Prediction Dataset
|| clinical feature for predicting stroke events

Features :

Numerical : age, avg_glucose_level, bmi
Categorical : gender, hypertension, heart_disease, ever_married
work_type, Residence_type, smoking_status, stroke

3 Codes

```
import pandas as pd
import numpy as np
a) df = pd.read_csv("health care-dataset- stroke-data.csv")
   print(df)
b) print(df.shape)
c) print(df.head(10))
d) print(df.tail(10))
e) print(df.info())
f) print(df.describe(include="all"))
g) print(df.dtypes)
h) print(df.isnull().sum())
i) print(df.nunique())
j) print((df.isnull().sum/len(df))*100)
k) for col in df.columns:
     if df[col].dtypes in ['float64','int64']:
        df.fillna(df[col].mean(), inplace=True)
   print(df.isnull().sum())
```

l)
```
third_col = df.columns[2]
df_sorted = df.sort_values(by=third_col)
print(df_sorted)
```

m)
```
second_col = df.columns[1]
new_col_name = second_col + "_new"
df.rename(columns = {second_col: new_col_name}, inplace=True)
print(df.head())
```

n)
```
cat_cols = df.select_dtypes(include = ['Object']).columns
print(cat_cols)
num_cols = df.select_dtypes(include = ['int64', 'float64'])
                .columns.
print(num_cols)
```

o)
```
Q1 = df[num_cols].quantile(0.25)
Q3 = df[num_cols].quantile(0.75)
IQR = Q3 - Q1
df_no_outliers = df[~((df[num_cols] < (Q1 - 1.5 * IQR)) |
        (df[num_cols] > (Q3 + 1.5 * IQR))).any(axis=1)]
print(df_no_outliers.head())
print(df_no_outliers.shape())
```

DATASET ANALYSIS REPORT

Figure 1: Entire Dataset

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	nan	never smoked	1
31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59.0	0	0	Yes	Private	Rural	76.15	nan	Unknown	1
60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81.0	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61.0	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54.0	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78.0	0	1	Yes	Private	Urban	219.84	nan	Unknown	1
5317	Female	79.0	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1

Figure 2: Size of the Dataset

Dataset Size: (5110, 12)

Figure 3: First 10 Rows of Dataset

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	nan	never smoked	1
31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59.0	0	0	Yes	Private	Rural	76.15	nan	Unknown	1
60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1

Figure 4: Last 10 Rows of Dataset

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
68398	Male	82.0	1	0	Yes	Self-employed	Rural	71.97	28.3	never smoked	0
36901	Female	45.0	0	0	Yes	Private	Urban	97.95	24.5	Unknown	0
45010	Female	57.0	0	0	Yes	Private	Rural	77.93	21.7	never smoked	0
22127	Female	18.0	0	0	No	Private	Urban	82.85	46.9	Unknown	0
14180	Female	13.0	0	0	No	children	Rural	103.08	18.6	Unknown	0
18234	Female	80.0	1	0	Yes	Private	Urban	83.75	nan	never smoked	0
44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.2	40.0	never smoked	0
19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

Figure 5: Dataset Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 5110 non-null   int64  
 1   gender             5110 non-null   object 
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64  
 4   heart_disease      5110 non-null   int64  
 5   ever_married       5110 non-null   object 
 6   work_type          5110 non-null   object 
 7   Residence_type     5110 non-null   object 
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object 
 11  stroke             5110 non-null   int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

Figure 6: Statistical View of Dataset

Statistic	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.0	5110.0	5110.0	5110.0	5110.0	4909.0	5110.0
mean	36517.8294	43.2266	0.0975	0.054	106.1477	28.8932	0.0487
std	21161.7216	22.6126	0.2966	0.2261	45.2836	7.8541	0.2153
min	67.0	0.08	0.0	0.0	55.12	10.3	0.0
25%	17741.25	25.0	0.0	0.0	77.245	23.5	0.0
50%	36932.0	45.0	0.0	0.0	91.885	28.1	0.0
75%	54682.0	61.0	0.0	0.0	114.09	33.1	0.0
max	72940.0	82.0	1.0	1.0	271.74	97.6	1.0

Figure 7: Data Types of Dataset Columns

Column	Data Type
id	int64
gender	object
age	float64
hypertension	int64
heart_disease	int64
ever_married	object
work_type	object
Residence_type	object
avg_glucose_level	float64
bmi	float64
smoking_status	object
stroke	int64

Figure 8: Null Values per Column

Column	Null Count
id	0
gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
bmi	201
smoking_status	0
stroke	0

Figure 9: Number of Unique Values per Column

Column	Unique Count
id	5110
gender	3
age	104
hypertension	2
heart_disease	2
ever_married	2
work_type	5
Residence_type	2
avg_glucose_level	3979
bmi	418
smoking_status	4
stroke	2

Figure 10: Percentage of Null Values per Column

Column	% Nulls
id	0.0
gender	0.0
age	0.0
hypertension	0.0
heart_disease	0.0
ever_married	0.0
work_type	0.0
Residence_type	0.0
avg_glucose_level	0.0
bmi	3.9334637964774952
smoking_status	0.0
stroke	0.0

Figure 11: Dataset After Filling Missing Values with Mean

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	28.893236911794666	never smoked	1
31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59.0	0	0	Yes	Private	Rural	76.15	28.893236911794666	Unknown	1
60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81.0	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61.0	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54.0	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78.0	0	1	Yes	Private	Urban	219.84	28.893236911794666	Unknown	1
5317	Female	79.0	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1

Figure 12: Dataset Sorted by Third Column (age)

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
29955	Male	0.08	0	0	No	children	Rural	70.33	16.9	Unknown	0
47350	Female	0.08	0	0	No	children	Urban	139.67	14.1	Unknown	0
22877	Male	0.16	0	0	No	children	Urban	114.71	17.4	Unknown	0
8247	Male	0.16	0	0	No	children	Urban	109.52	13.9	Unknown	0
41500	Male	0.16	0	0	No	children	Rural	69.79	13.0	Unknown	0
11371	Male	0.24	0	0	No	children	Urban	89.28	14.2	Unknown	0
53279	Male	0.24	0	0	No	children	Rural	118.87	16.3	Unknown	0
64974	Male	0.24	0	0	No	children	Urban	58.35	18.6	Unknown	0
69222	Male	0.24	0	0	No	children	Urban	57.09	19.4	Unknown	0
42500	Male	0.24	0	0	No	children	Rural	146.97	18.5	Unknown	0
68382	Male	0.32	0	0	No	children	Urban	127.78	20.8	Unknown	0
61511	Female	0.32	0	0	No	children	Rural	73.71	16.2	Unknown	0
13857	Male	0.32	0	0	No	children	Urban	89.04	17.8	Unknown	0
37622	Female	0.32	0	0	No	children	Urban	108.63	19.6	Unknown	0
66772	Female	0.32	0	0	No	children	Rural	55.86	16.0	Unknown	0

Figure 13: Renamed Column 'gender' → 'gender_new’

id	gender_new	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	nan	never smoked	1
31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59.0	0	0	Yes	Private	Rural	76.15	nan	Unknown	1
60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81.0	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61.0	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54.0	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78.0	0	1	Yes	Private	Urban	219.84	nan	Unknown	1
5317	Female	79.0	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1

Figure 14: Categorical and Numerical Columns

```
Categorical Columns:
['gender', 'ever_married', 'work_type', 'Residence_type', 'smoking_status']

Numerical Columns:
['id', 'age', 'hypertension', 'heart_disease', 'avg_glucose_level', 'bmi', 'stroke']
```

Figure 15: Dataset After Dropping Outliers (New Shape: 3824 rows × 12 columns

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
30669	Male	3.0	0	0	No	children	Rural	95.12	18.0	Unknown	0
16523	Female	8.0	0	0	No	Private	Urban	110.89	17.6	Unknown	0
56543	Female	70.0	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0
46136	Male	14.0	0	0	No	Never_worked	Rural	161.28	19.1	Unknown	0
52800	Female	52.0	0	0	Yes	Private	Urban	77.59	17.7	formerly smoked	0
15266	Female	32.0	0	0	Yes	Private	Rural	77.67	32.3	smokes	0
10460	Female	79.0	0	0	Yes	Govt_job	Urban	77.08	35.0	Unknown	0
63884	Female	37.0	0	0	Yes	Private	Rural	162.96	39.4	never smoked	0
37893	Female	37.0	0	0	Yes	Private	Rural	73.5	26.1	formerly smoked	0
67855	Female	40.0	0	0	Yes	Private	Rural	95.04	42.4	never smoked	0
25774	Male	35.0	0	0	No	Private	Rural	85.37	33.0	never smoked	0
19584	Female	20.0	0	0	No	Private	Urban	84.62	19.7	smokes	0
24447	Female	42.0	0	0	Yes	Private	Rural	82.67	22.5	never smoked	0
49589	Female	44.0	0	0	Yes	Govt_job	Urban	57.33	24.6	smokes	0
47175	Female	49.0	0	0	Yes	Private	Rural	60.22	31.5	smokes	0