Assignment 5 – Data Wrangler

Table of Contents

I. Data Description

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

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.

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

Assignment 5 – Data Wrangler

Data Attributed 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"*

Assignment 5 – Data Wrangler

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

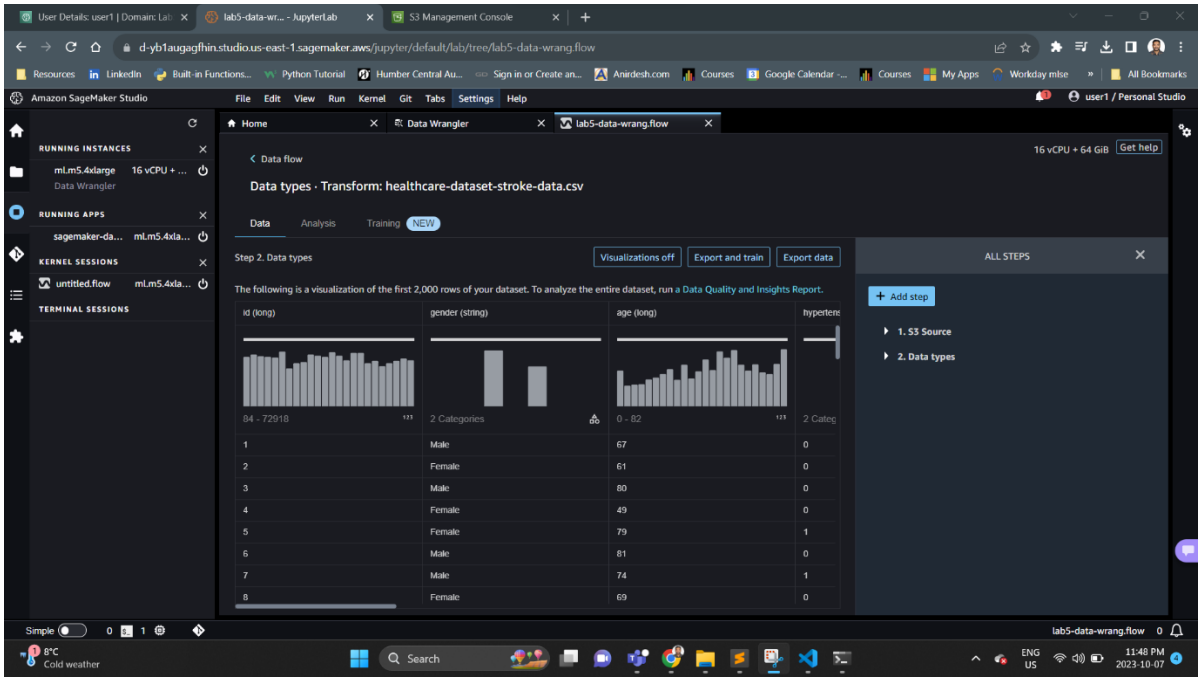
   l. stroke : Integer

The above information describes the dataset according to my understanding and knowledge and will do the analysis on it with Data Wrangler in AWS based on these data definitions.

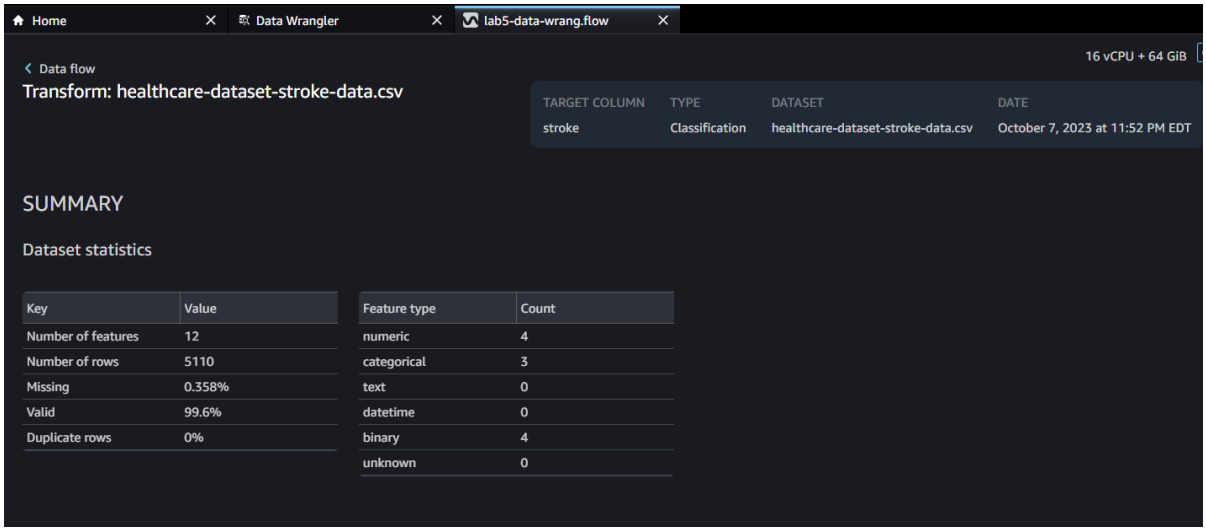
Assignment 5 – Data Wrangler

II. Analysis and Visualization in Data Wrangler

1. Load the dataset to S3 bucket, import the dataset to Data Wrangler from S3 bucket and create a flow diagram.



2. Generate a quick report for getting the summary of the dataset using inbuild feature of Data Wrangler which gives information on missing values and number or outliers.



Assignment 5 – Data Wrangler

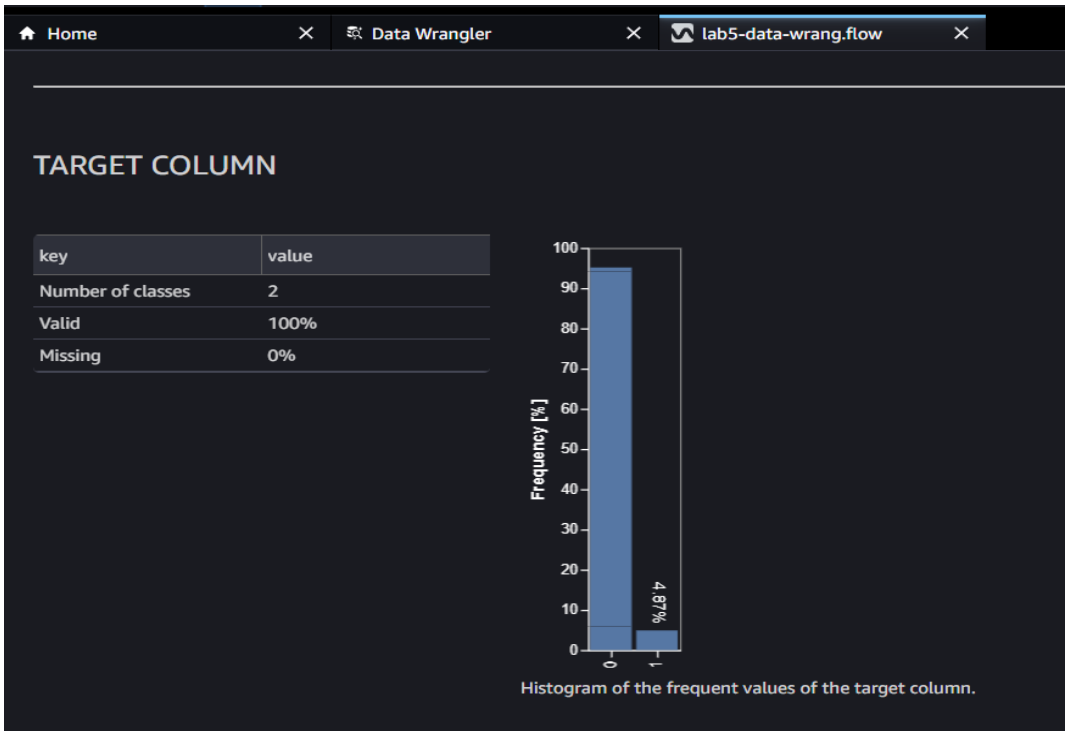
Table Summary: TableSummary

Last generated: Sunday, October 8, 2023 at 12:10 AM EDT

Refresh

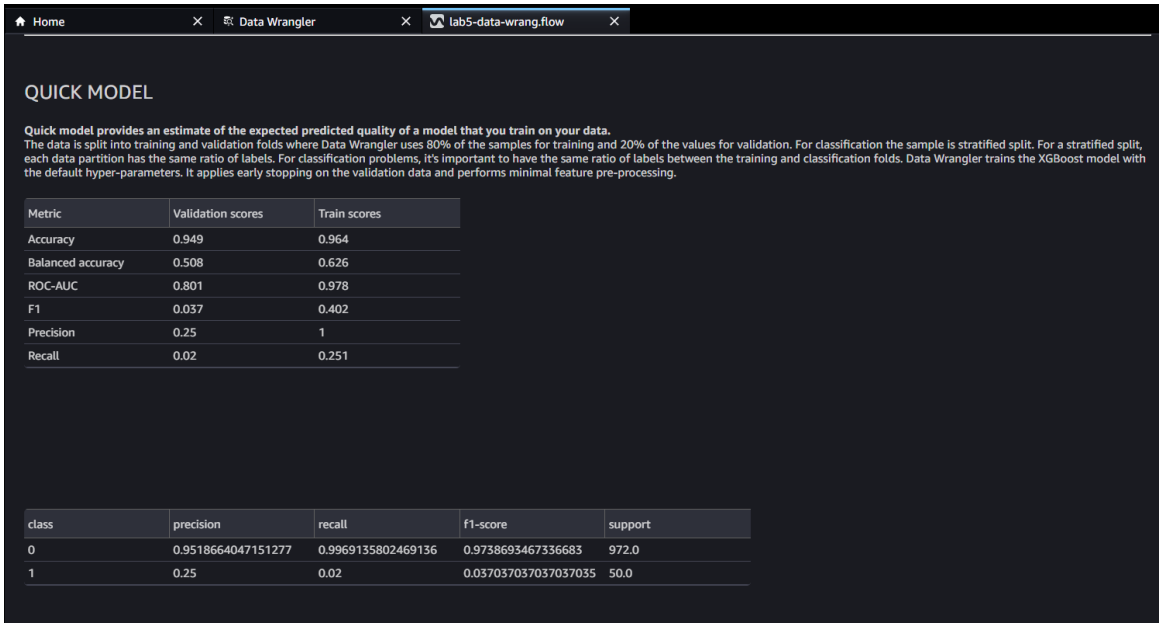
summary	id	gender	age	hypertension	heart_diseas
count	5110	5110	5110	5110	5110
mean	36517.82935420744	None	43.21526418786693	0.0974559686888454	0.05401174
stddev	21161.72162482715	None	22.633865752854746	0.296606674233791	0.22606298
min	67	Female	0	0	0
max	72940	Other	82	1	1

3. Target Column selection – ‘stroke’ analysis



4. Data Wrangler can run a quick model on the dataset and give insights into the stroke prediction and also creates visualization.

Assignment 5 – Data Wrangler

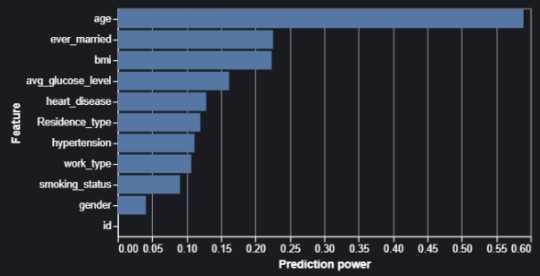


Metric	Validation scores	Train scores
Accuracy	0.949	0.964
Balanced accuracy	0.508	0.626
ROC-AUC	0.801	0.978
F1	0.037	0.402
Precision	0.25	1
Recall	0.02	0.251

class	precision	recall	f1-score	support
0	0.9518664047151277	0.9969135802469136	0.9738693467336683	972.0
1	0.25	0.02	0.037037037037037035	50.0

5. Feature Summary Report

Feature	Prediction power	Type	Valid	Missing	High severity warnings	Medium severity warni…
age	0.59	numeric	100%	0%	0	0
ever_married	0.226	binary	100%	0%	0	0
bmi	0.224	numeric	96.1%	3.93%	0	0
avg_glucose_level	0.162	numeric	100%	0%	0	0
heart_disease	0.129	binary	100%	0%	0	0
Residence_type	0.12	binary	100%	0%	0	0
hypertension	0.112	binary	100%	0%	0	0
work_type	0.107	categorical	100%	0%	0	0
smoking_status	0.0904	categorical	100%	0%	0	0
gender	0.0408	categorical	100%	0%	0	0
id	0	numeric	100%	0%	0	0

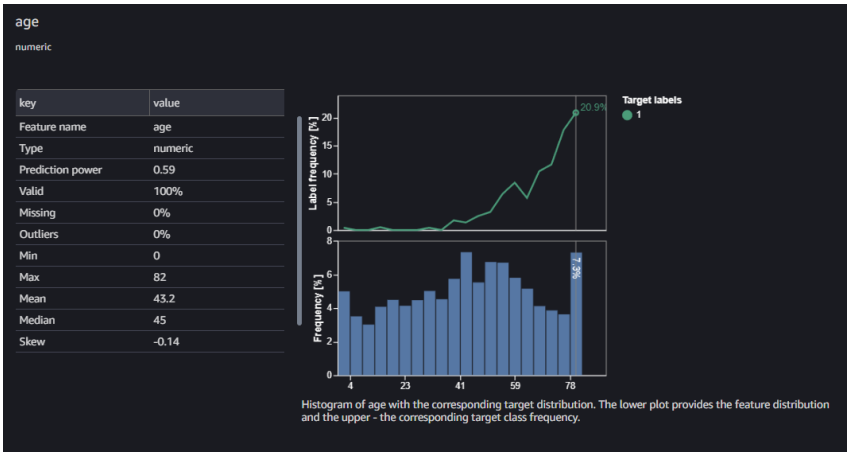


Prediction power of the features. Higher prediction power scores, toward 1, indicate columns that are more useful for predicting the target on their own. Lower scores, toward 0 point to columns that contain little useful information for predicting the target on their own.

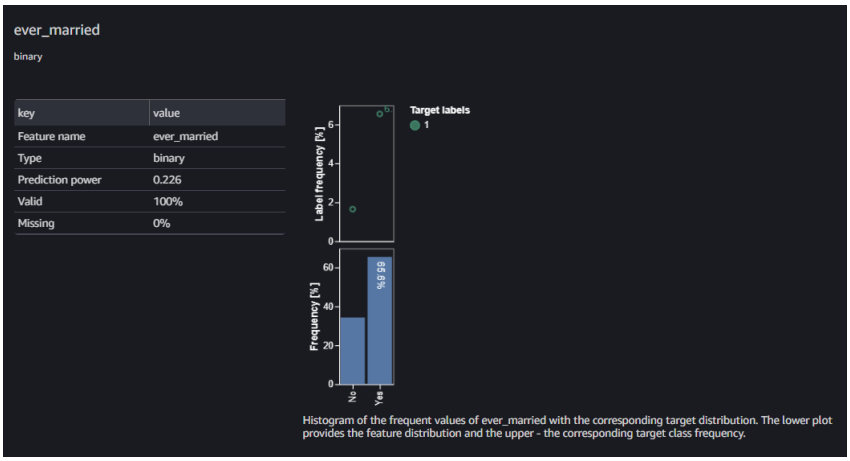
Assignment 5 – Data Wrangler

6. Feature that are most important for consideration into predicting the output variable stroke.

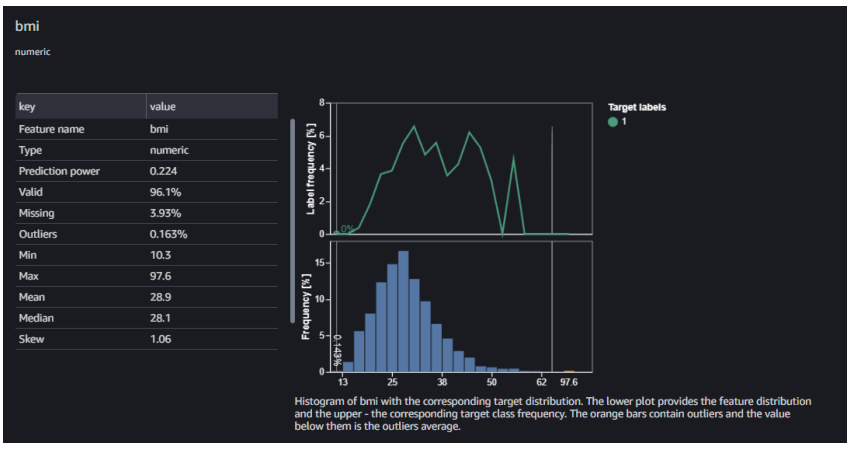
   a. Age Distribution



   b. Ever Married



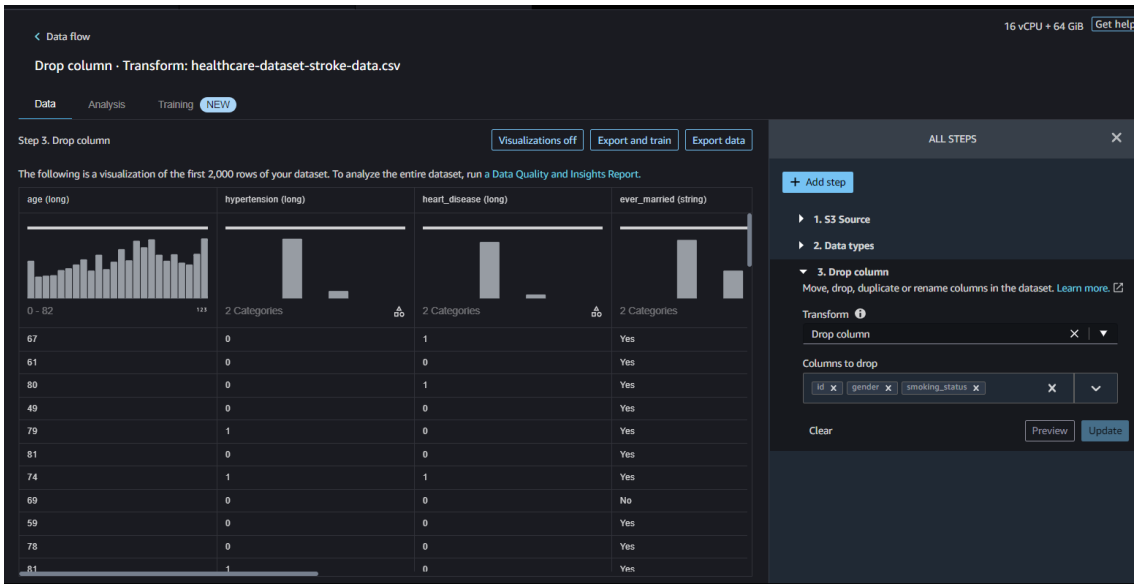
   c. BMI Levels



Assignment 5 – Data Wrangler

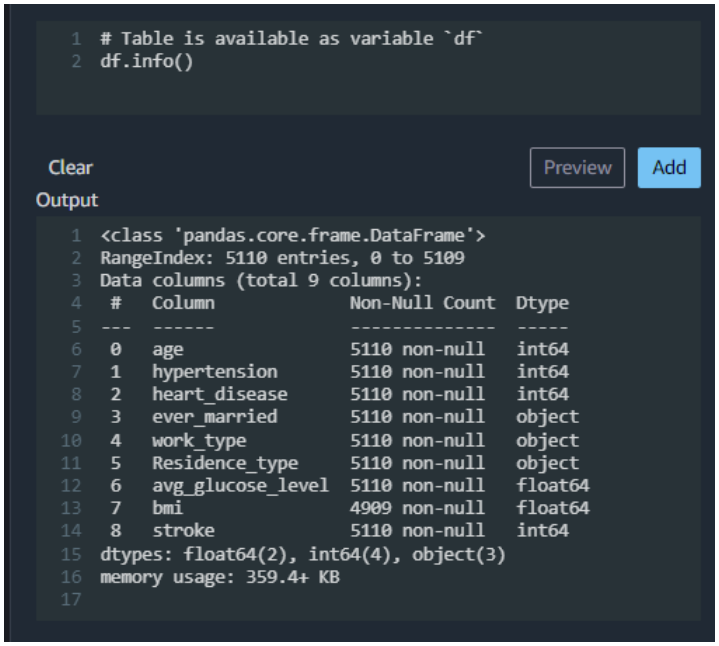
III. Transformations using Data Wrangler.

1. Drop Columns Transformations.



Columns – ‘id’, ‘gender’ and ‘smoke_status’ were dropped as they did not have any significant impact on the model taking consideration of Feature summary report.

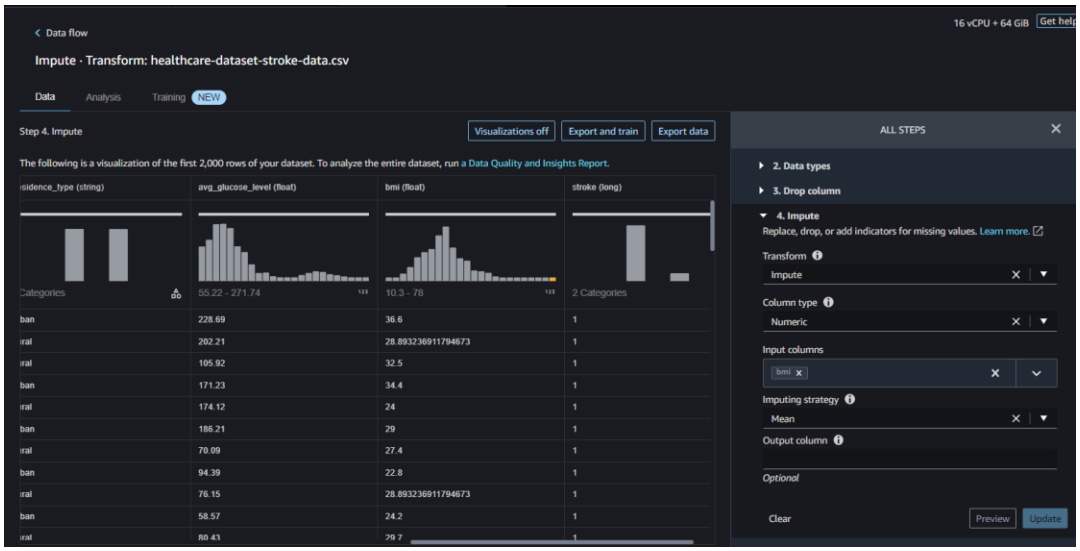
2. Identify Missing Values using custom code.

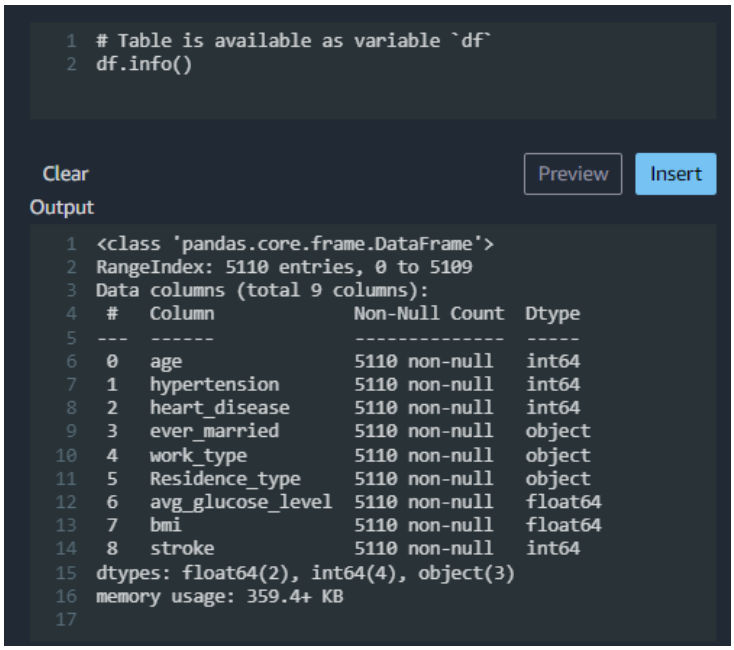


Running a python code df.info() we found the information that the ‘bmi’ column has some missing values.

Assignment 5 – Data Wrangler

3. Impute Missing Values with Mean of the column BMI.

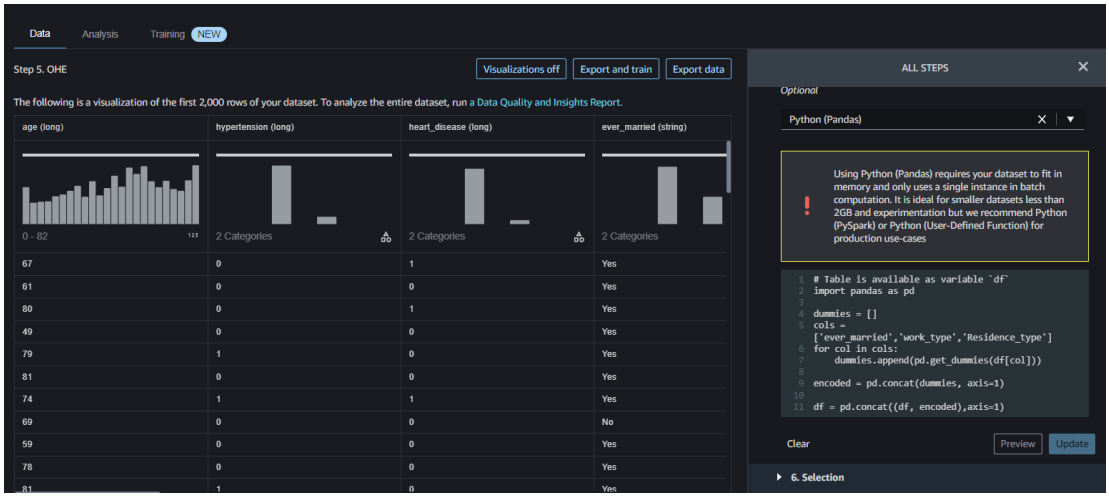




As the ‘bmi’ column is important for the prediction of the target column ‘stroke’ we are not removing those columns instead we would impute the missing values of bmi with the mean of the column.

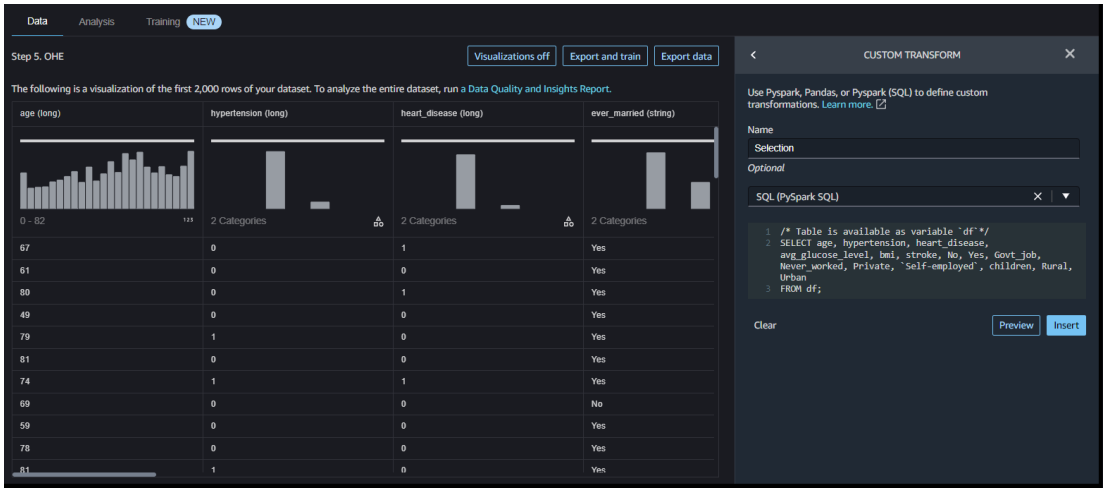
Assignment 5 – Data Wrangler

4. Perform One Hot Encoding on categorial variables.



As there are multiple categorical variables in the dataset and I wish to use those variables as features in a machine learning model I have applied One-hot-encoding on few categories such as ‘ever_married’, ‘work_type’ and ‘Residence_type’.

5. Select Required columns for further analysis.



After performing One-hot-encoding the columns that I though that are required for the input to the machine learning model are been retained and the remaining column have been dropped. I have used PySpark SQL for querying the data frame and select what is required.