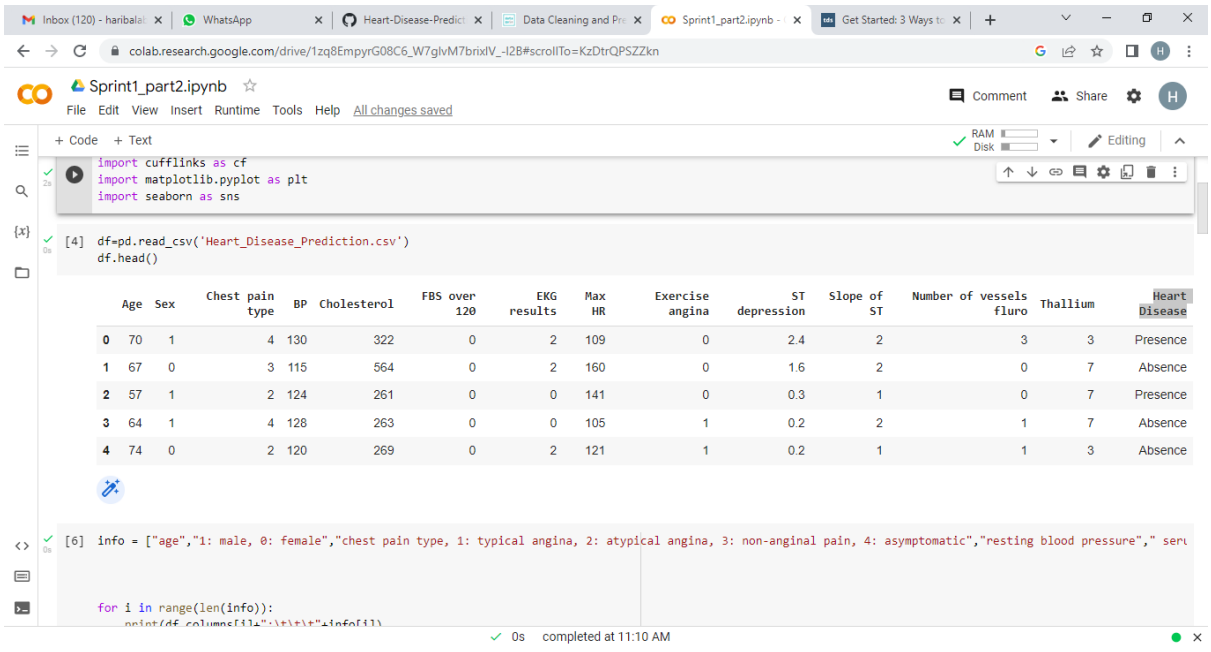
SPRINT 1-PART 2

Date	27 October 2022
Team ID	PNT2022TMID53213
Project Name	Visualizing and Predicting Heart Diseases with an Interactive Dash Board

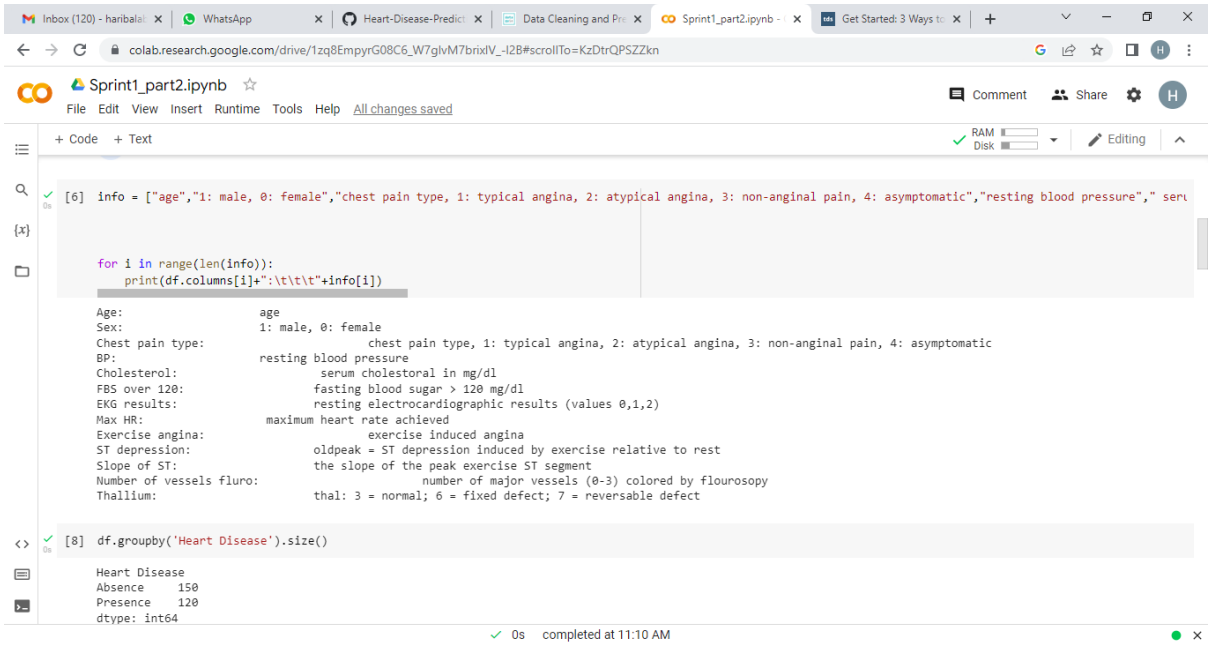
1)Description:

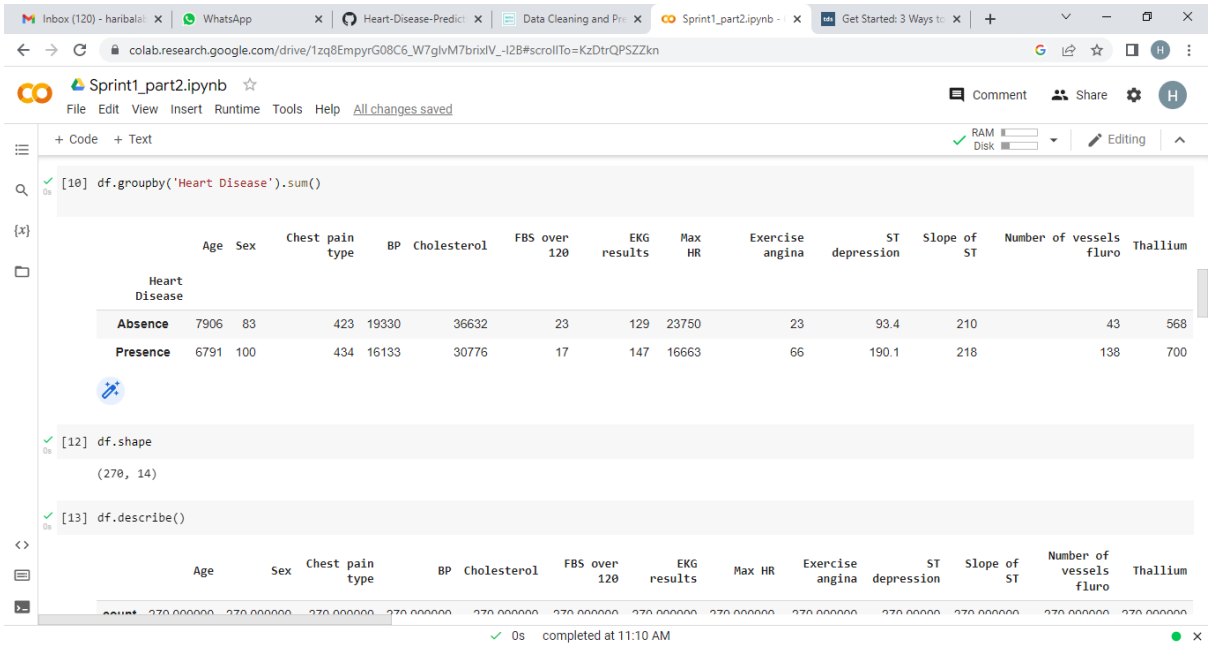
->It tells about the dataset and their columns and values, and their variation in it.

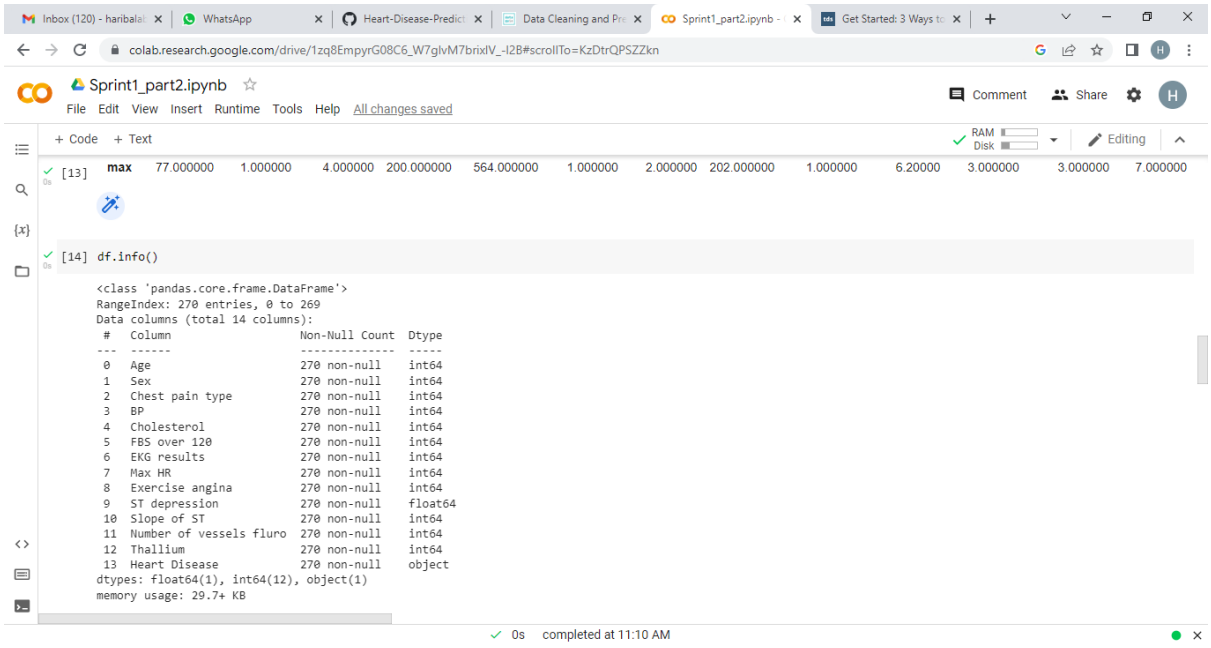
->To help members of your organization quickly identify datasets that might be useful for them, provide a concise, informative description of your dataset in the dataset's settings. Users will see this description in the info tooltip next to the dataset's name in the datasets hub, as well as on the dataset's details page. Providing a meaningful description helps foster dataset reuse. For instance, based on a dataset's description, users may decide to explore reports that are based on the dataset, or to create their own reports based on the dataset.

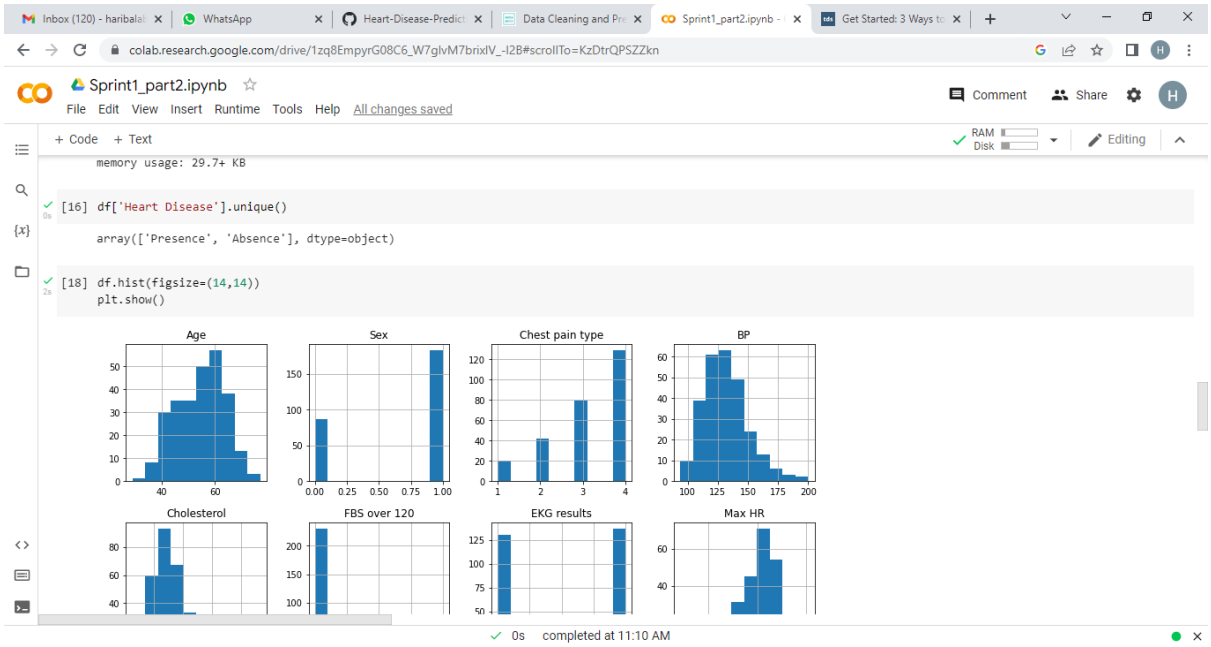


	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence


Sprint1_part2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
[6] info = ["age","1: male, 0: female","chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic","resting blood pressure"," seru
for i in range(len(info)):
print(df.columns[i]+":\t\t\t"+info[i])
Age: age
Sex: 1: male, 0: female
Chest pain type: chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
BP: resting blood pressure
Cholesterol: serum cholestoral in mg/dl
FBS over 120: fasting blood sugar > 120 mg/dl
EKG results: resting electrocardiographic results (values 0,1,2)
Max HR: maximum heart rate achieved
Exercise angina: exercise induced angina
ST depression: oldpeak = ST depression induced by exercise relative to rest
Slope of ST: the slope of the peak exercise ST segment
Number of vessels fluro: number of major vessels (0-3) colored by flourosopy
Thallium: thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
[8] df.groupby('Heart Disease').size()
Heart Disease
Absence 150
Presence 120
dtype: int64
0s completed at 11:10 AM


Sprint1_part2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
[10] df.groupby('Heart Disease').sum()
Age Sex Chest pain type BP Cholesterol FBS over 120 EKG results Max HR Exercise angina ST depression Slope of ST Number of vessels fluro Thallium
Heart Disease
Absence 7906 83 423 19330 36632 23 129 23750 23 93.4 210 43 568
Presence 6791 100 434 16133 30776 17 147 16663 66 190.1 218 138 700
[12] df.shape
(270, 14)
[13] df.describe()
Age Sex Chest pain type BP Cholesterol FBS over 120 EKG results Max HR Exercise angina ST depression Slope of ST Number of vessels fluro Thallium
0s completed at 11:10 AM





2)Preprocessing:

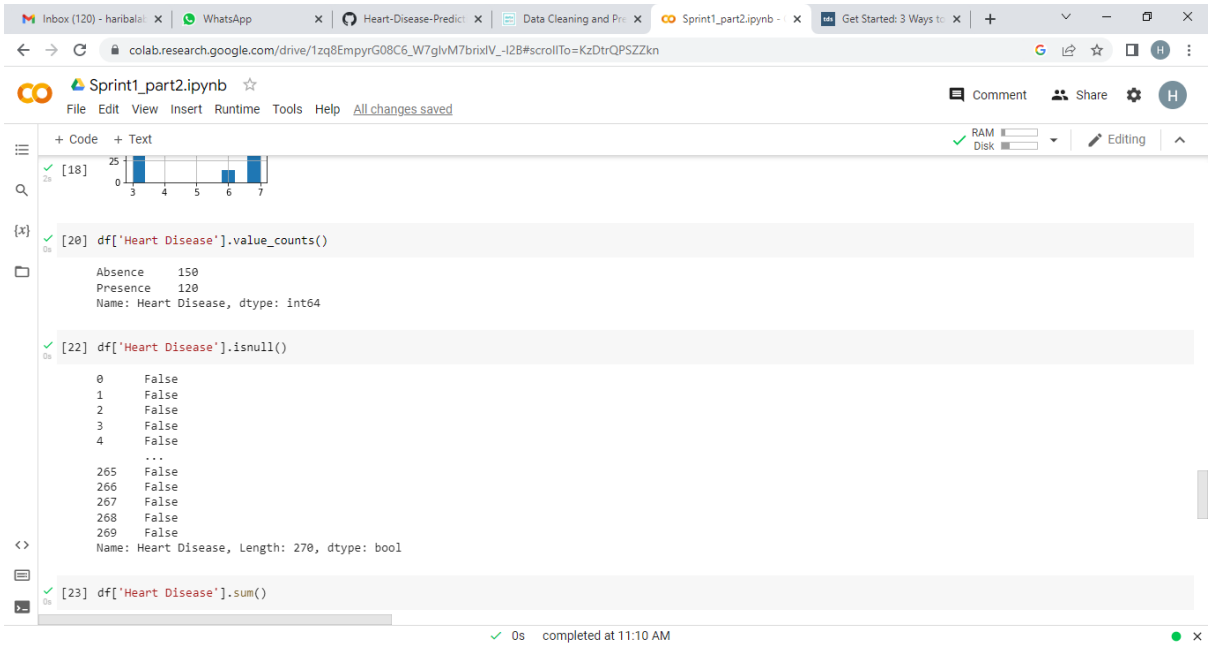
->Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

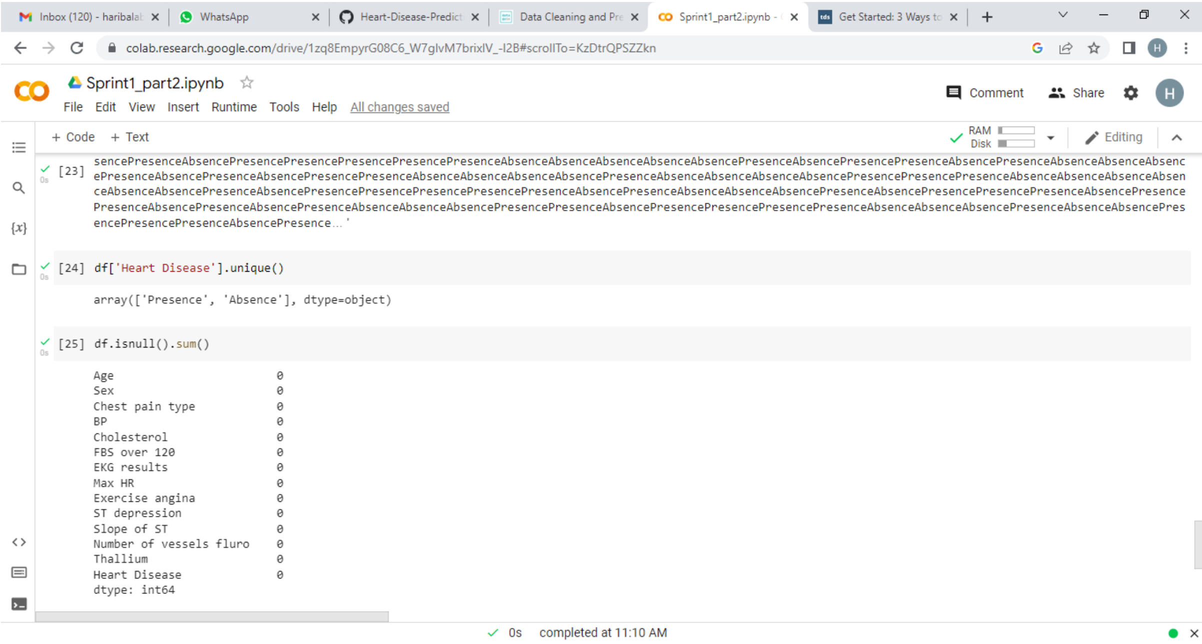
When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

->A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

Steps:

- ->**Getting the dataset**
- **Importing libraries**
- **Importing datasets**
- **Finding Missing Data**
- **Encoding Categorical Data**
- **Splitting dataset into training and test set**
- **Feature scaling**





3)Cleaning:

->Data scientists spend a large amount of their time cleaning datasets and getting them down to a form with which they can work. In fact, a lot of data scientists argue that the initial steps of obtaining and cleaning data constitute 80% of the job.

Therefore, if you are just stepping into this field or planning to step into this field, it is important to be able to deal with messy data, whether that means missing values, inconsistent formatting, malformed records, or nonsensical outliers.

->Cleaning is not needed as this dataset already ready to solve the problem.