

# DATA PREPROCESSING

# Dataset

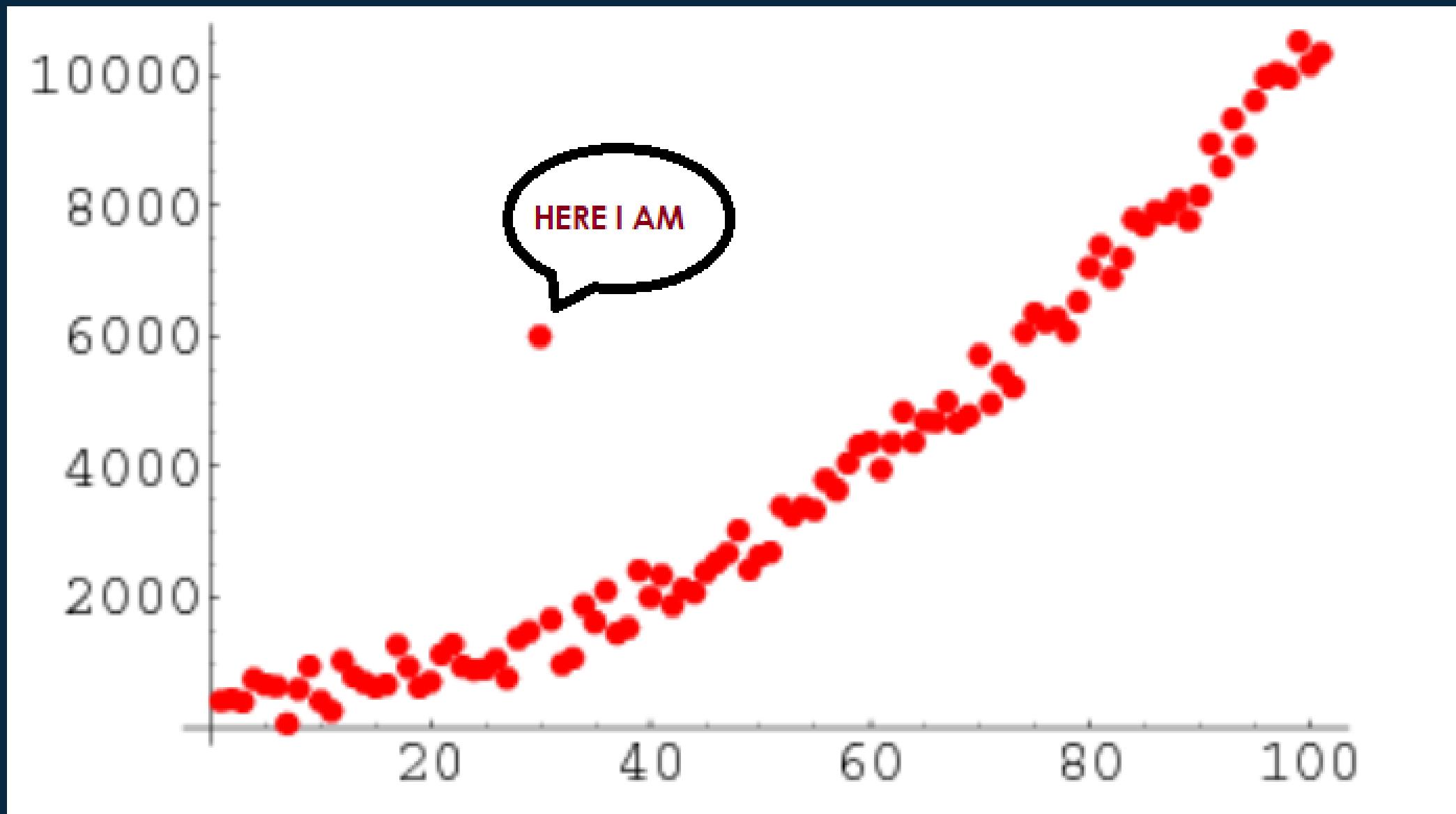
- In machine learning, a dataset is a collection of data used to train and evaluate a machine learning model. It typically consists of a set of observations, each of which contains **one or more features or attributes**.
- Dataset forms in machine learning include **tabular data** (e.g., CSV), **image data** (e.g., JPEG, PNG), **audio data** (e.g., WAV, MP3), **text data** (e.g., plain text, HTML), and **hierarchical data** (e.g., JSON, XML).
- CSV (Comma Separated Values) is a file format commonly used for storing and exchanging tabular data, where each row **represents an instance or observation**, and each column represents a **feature or attribute**.

# Preprocessing steps

- Data loading
  - Reading the data
- Data wrangling
  - Checking for null,duplicate values and dropping coloumns
- Data visualization
  - Graphical understanding of the data bar graphs, histograms,piecharts,scatterplots etc..
- Data imputation
  - Replacing missing values with mean values
- Data scaling
  - Removing bias towards a particular field(Normalization or standardization)

# Handling Outliers

Handling the data that is very different from the others



# Data partitioning

- In machine learning, the dataset is often partitioned into training and testing sets.
- The model is trained on the training dataset and then tested on the testing dataset.
- The testing dataset thus acts as the unseen dataset, which can be used to estimate a generalization error (the error expected when the model is applied to a real-world dataset after the model has been deployed).

# Data partitioning

Subject	t	Feature 1	Feature 2	Target		Subject	t	Feature 1	Feature 2	Target
Paul	1	1000	male	0		Paul	1	1000	male	0
Paul	2	1100	male	0		Paulina	1	10000	female	0
Paul	3	1200	male	1		George	1	50000	male	1
Paul	4	1300	male	1		Paul	2	1100	male	0
Crista	4	20	female	0	↑	Paulina	2	100000	female	1
Crista	5	100	female	0	Train	George	2	50000	male	1
Paulina	1	10000	female	0	↓	Paul	3	1200	male	1
Paulina	2	100000	female	1		Paulina	3	95000	female	1
Paulina	3	95000	female	1		George	3	50000	male	1
Paulina	4	97000	female	1		Paul	4	1300	male	1
Paulina	5	99000	female	1		Crista	4	20	female	0
Paulina	6	101000	female	1		Paulina	4	97000	female	1
George	1	50000	male	1	↑	George	4	50000	male	1
George	2	50000	male	1	Test	Crista	5	100	female	0
George	3	50000	male	1	↓	Paulina	5	99000	female	1
George	4	50000	male	1		George	5	50000	male	1
George	5	50000	male	1		Paulina	6	101000	female	1
George	6	50000	male	1		George	6	50000	male	1

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	