Applied Data Science
Trainer: Prashant Nair

Dataset / Data that is
used for performing analysis
or intelligence extraction

POPULATION
[universal set]

eg: XYZ company
- 4 branches
DEL, BOM, CHN, BLR
- each branch 100 emps.

∴ given dataset can be considered as population if it contains all 400 records

* maths *
(accurate)

SAMPLE
[SUBSET OF THE POPULATION that represent the entire population]

eg: XYZ company
- 4 branches
DEL, BOM, CHN, BLR
- each branch 100 emps.

∴ given dataset can be called as sample if it contains balanced records of each branch

data.csv → 12 records

DEL → 3
BOM → 3
CHN → 3
BLR → 3
} → 12

* STATISTICS *
(approximate)

Dealing with Samples

STATISTICS → a way to get information out of the data.

Outcome/Result of stats is always APPROXIMATE ← STATISTICS

DESCRIPTIVE STATISTICS

70% EDA [Data exploration]

EDA → Exploratory Data Analysis

- help you to organize and summarize the data.
- Identify the quality of the data w.r.t the given problem statement.
- To make your data compatible for intelligence extraction.

INFERENTIAL STATISTICS

30%

- Create Predictive models
- Technique for drawing and measuring the reliability of the conclusions/predictions of the population based on given sample

Types of data colums you deal with in a typical dataset

(data engineers)

Numerical Data (ND)

Data will be numeric

- Continuous ND: any entity that change w.r.t time
- Discrete ND: any entity that remains FIXED irrespective of time.

Categorical Data

data that helps GROUPING -or- CLASSIFYING

eg: gender, states in countries etc.

Ordinal Data

categorical data with Ranking.

eg: feedback rating

→ ★ ★ ★ ★ ★

→ ★ ★ ★ ★

→ ★ ★ ★

→ ★ ★

→ ★

Pure String Data.

data column contain text data

eg: weight (human)   eg: SAT

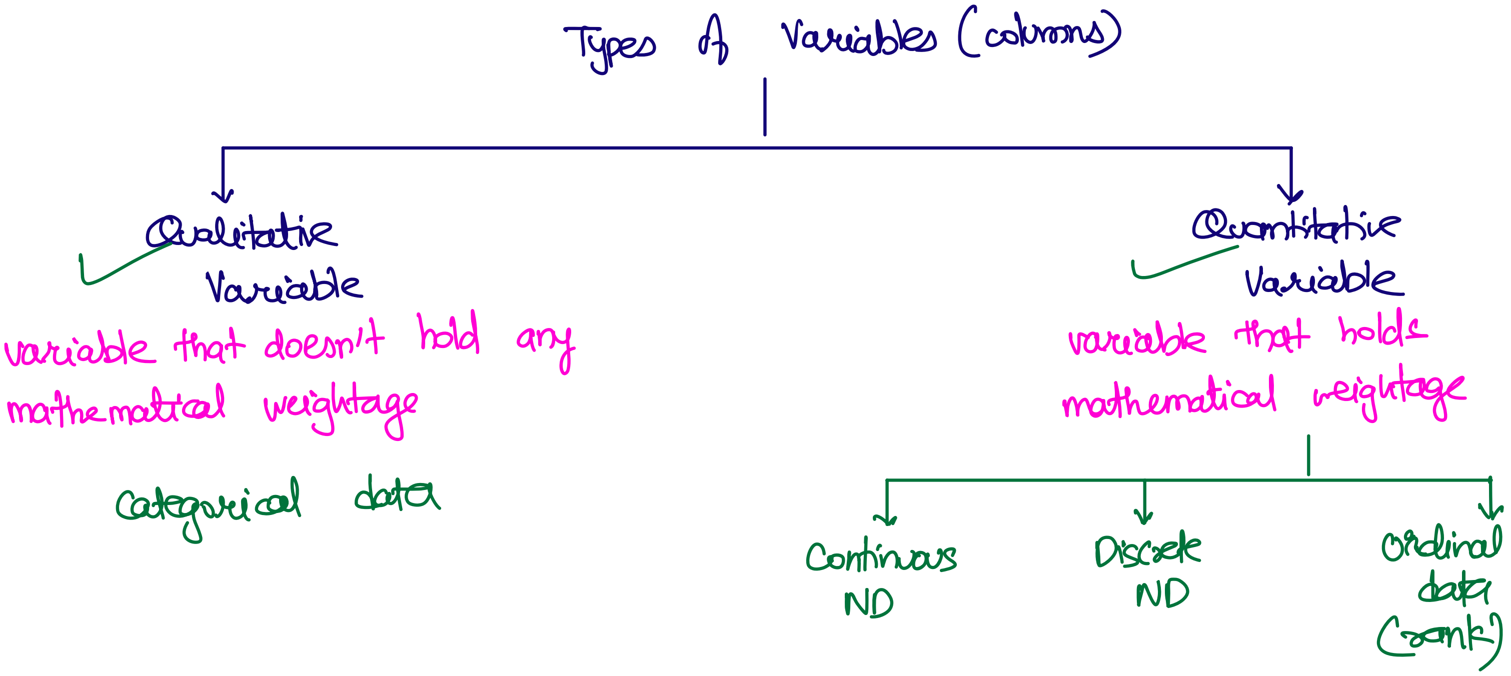
Practical: [Historical Data]

if range is defined by the domain for the column:

Discrete ND

else

CONTINUOUS ND



Types of Variables (columns)

Qualitative Variable

variable that doesn't hold any mathematical weightage

categorical data

Quantitative Variable

variable that holds mathematical weightage

Continuous ND

Discrete ND

Ordinal data (rank)

Descriptive Stats → Data Exploration
→ EDA (Exploratory Data Analysis)

Descriptive stats is the first step to deal with SAMPLE. Typical operations that a data analyst / data scientist / business analyst perfor include (not limited to)

① Check the type of DISTRIBUTION for each NUMERICAL column.

② Identify and deal with INAPPROPRIATE data

③ Identify and deal with OUTLIERS

④ Identify and deal with MISSING VALUES

⑤ Identify and deal with CATEGORICAL DATA

⑥ DISCOVER any

ⓐ ASSOCIATION

ⓑ RELATIONSHIP

ⓒ PATTERN

between two or more variables/columns in the given dataset.

WORKING DATASET

INTERNAL DATASET

data is generated, managed and maintained by same organization

EXTERNAL DATASET

data that sourced from third party or from open sources like social media, surveys etc