Mastering Pandas Library and EDA (Part-2)

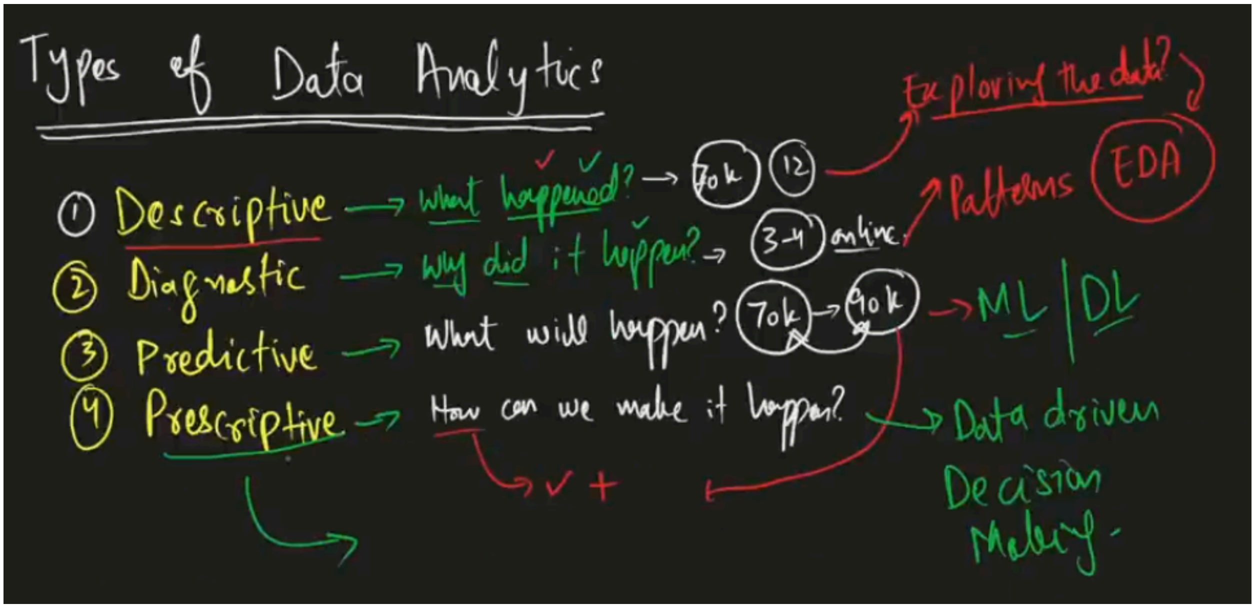
Data Scientist vs Analyst:

Data Scientist:-
→ Understand Business
→ Data Acquisition/Understanding
→ Data Preparation
→ Data Modeling (ML/DL)
→ Data/Model Evaluation
→ Monitor and Optimize
→ Model Deployment
→ Communicate Technical Insights

Data Analyst:-
→ Required Information about Data (Meta-data)
→ Data Collection
→ Has Assigned Goals by Company
→ Data Cleaning
→ Exploratory Data Analysis (EDA)
→ Generate Inference
→ Create Simple Models
→ Deploy and Interpret Models
→ Visualize Data
→ Reporting and Dashboarding

Type of Data Analytics:

1- Descriptive → What happened? (Our Sales increased by 20% in last 30 days) (exploring data)
2- Diagnostic → Why did it happen? (ads spend increased and CVR goes high) (finding patterns)
3- Predictive → What will happen? (Our sales can be increased 20% - 30%) (ML/DL Algo)
4- Prescriptive → How can we make it happen? (by increasing ad spend on high ROI campaigns) (Decision Making)



Types of Data Analytics

① Descriptive → what happened? → 70k 12 → Exploring the data?

② Diagnostic → why did it happen? → 3-4 online → Patterns EDA

③ Predictive → What will happen? 70k → 90k → ML | DL

④ Prescriptive → How can we make it happen? → Data driven Decision Making.

✓ +

Data Life Cycle:

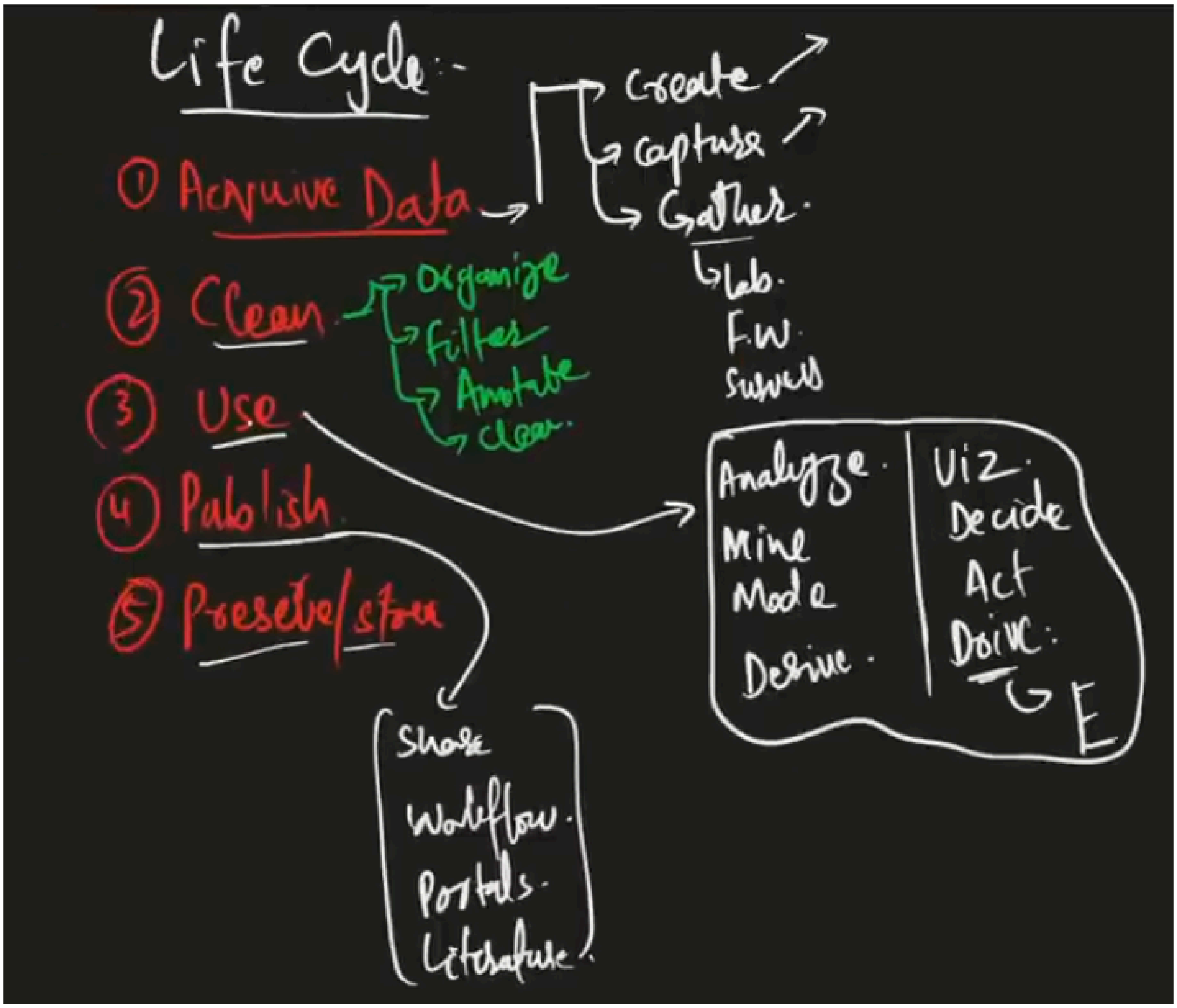
1- Acquire:
→ Create, capture, and gather data

2 - Clean
→ Organize, filter, annotate, and clean the data

3- Use
→ Analyze, mine, visualize, decide how to use for model, modeling, act etc

4- Publish
→ Share, workflow, portals, make part of literature etc

5-Preserve/Store
→ Safely saving the data



Rows and Columns:

Rows:-
→ Horizontal lines, going from left to right
→ Also called Records, Observations, Instances, Entries, and Data Points

Columns:-
→ Columns are the vertical lines, going from top to bottom
→ Also called Attributes, Dimensions, Variables, Properties, Features, and Fields

DataFrame:
→ two-dimensional, labeled data structure that organizes data into rows and columns

Structure/Unstructured Data:
→ Structured Data is organized in a predefined format, like rows, columns and header in a table
→ Unstructured Data has no fixed format or structure, like images, videos etc

Wh? Questions:

→ We need to write some questions before data collection:
→ Why? > How? > Where? > Who? > When? Etc

Primary and Secondary Data:

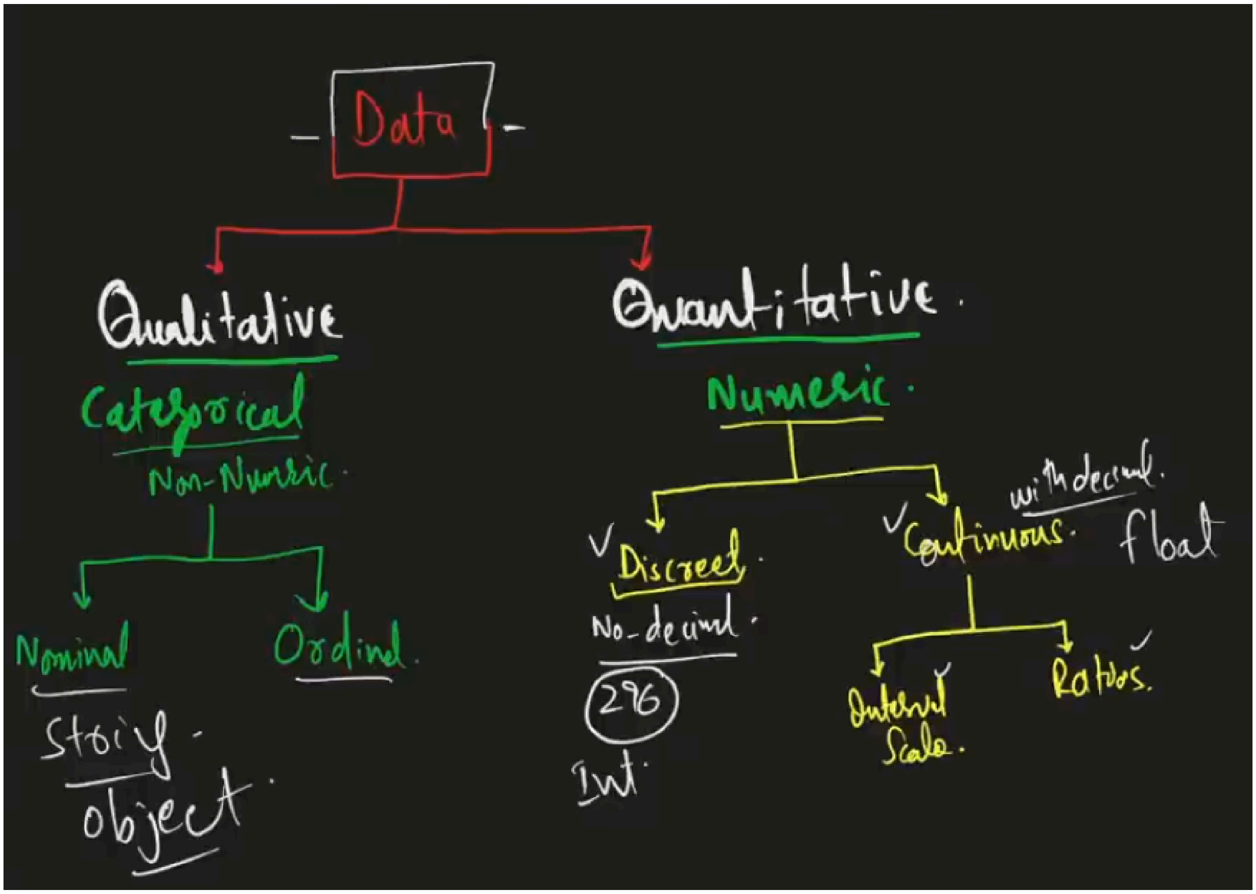
→ Data that is collected by yourself is Primary Data and it’s usually very expensive
→ Data by someone else is Secondary Data

Level of Measurement:

→ **Nominal (Str, Object, category):** data can only be categorized, no rank (name, color etc)
→ **Ordinal (category)**: data can be categorized and ranked (movie ratings: {poor, average, good})
→ **Interval (float):** data can be categorized, ranked, and evenly spaced (Temperature in °C or °F)
→ **Ratio (int):** data can be categorized, ranked, evenly spaced, and has a natural zero (weight, height etc)

Qualitative vs Quantitative:

→ Qualitative > Categorical/Non-Numeric (Nominal and Ordinal)
→ Quantitative > Numerical (Discrete, Continuous)
→ Discrete > No-Decimal
→ Continuous > Interval and Ratio


Data
Qualitative
Categorical
Non-Numeric
Nominal
Ordinal
String
object
Quantitative
Numeric
Discreet
No-decimal
296
Int
Continuous
with decimal
float
Interval Scale
Ratios