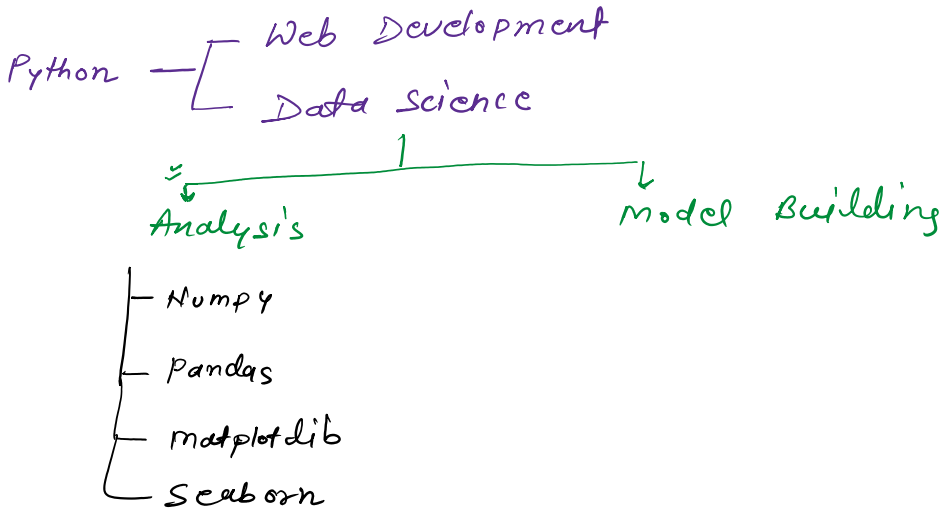
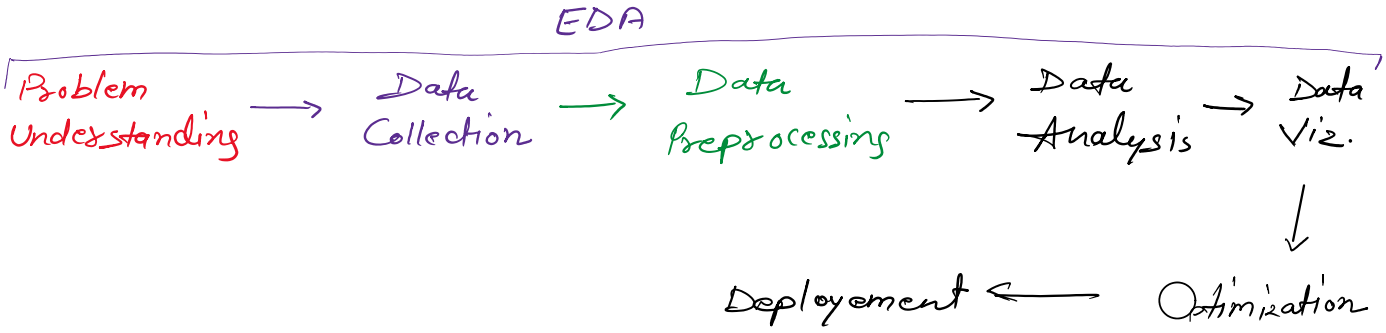
EDA

13 August 2025 15:45



Lifecycle



EDA : Exploratory Data Analysis

↳ Understanding Data

① Univariate Analysis

↓ ↓

Single Column

Name	Age	Gender

Analysis

Category ← unique category
count

Numerical ← max, min, avg, sum

② Bivariate

↓ ↓

Two Column

Cat + Cat → count

* Cat + Num → stats by category

* Num + Num → Relationship
[correlation]

column, bar, pie

scatter

③ Multivariate

↓ ↓

>2 Column

Correlation
-
bar, col

Categorical → count → scatter

Numerical → max, min, avg, sum
→ count, Group/intervals

countplot(), histogram, boxplot

Statistics

Collect, Organize, Analyze, Interpret Data.

Descriptive
- → Describing Data
- → Historical Event
- → Election Results
- → Entire Data

Inferential
- → Inferring from data (Conclude)
- → Forecasting
- → Election Exit Polls
- → Sample Data

→ Later ←

① Aggregate measure

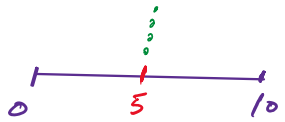
sum(), max(), min(), count(), average()

② Central Tendency measure : exactly or approx into two parts.

mean = $\frac{\text{Sum(value)}}{\text{Count(value)}}$

median = Sort data then middle value

mode = most repeated value



19, 20, 20, 21, 22, 20, 19, 35, 37

19, 19, 20, 20, 20, 21, 22, 35, 37

× Mean = 23.6

Median = 20 ✓ [Data contains Outlier]

Mode = 20

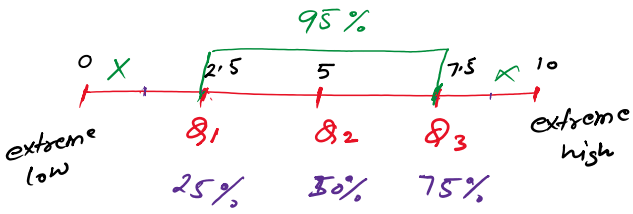
⇒ To select right CTM, we need to look at data distribution.

③ Dispersion / Data distribution measure

① Range

max, Min

② Interquartile Range [IQR]



Range = max - Min

IQR = $Q_3 - Q_1$

identify Outlier by defining lower & upper

Lower = $Q1 - (1.5 * IQR)$

Upper = $Q3 + (1.5 * IQR)$

③ Variance / Standard Deviation

↳ distance of any value from its average.

$$var = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2$$

$= (\ )^2$ → square unit

$std = \sqrt{var}$

$= (\ )$ ⇝ same unit

Higher Standard Deviation : Data more scattered