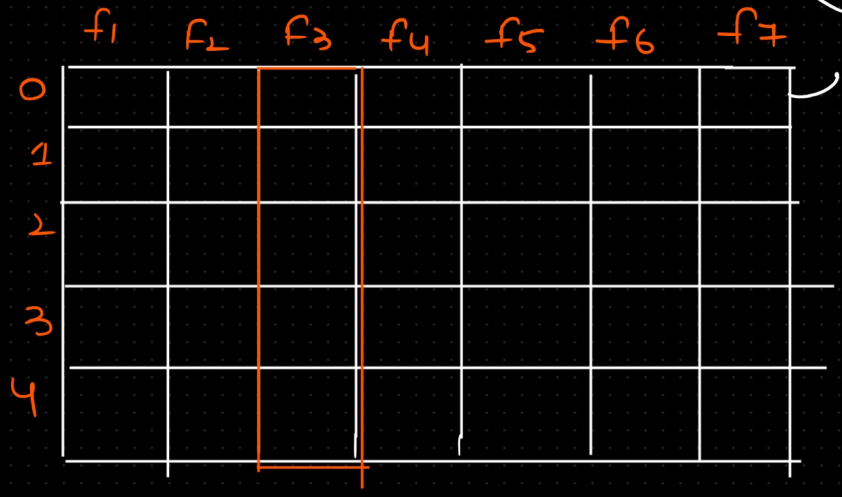
Pandas

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis/manipulation tool available in any language. It is already well on its way toward this goal.

[[ ]]

2D Data

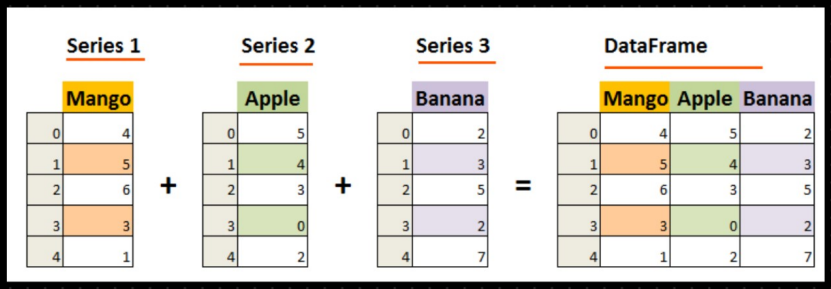
Data Structure



① Dataframe → tabular structure

② Series → [ ] 1D (column)

→ Single column from a list of multiple features



Statistics

Statistics is the study and manipulation of data, including ways to gather, review, analyze, and draw conclusions from data. The two major areas of statistics are descriptive and inferential statistics.

Description of a given dataset

Hypothesis

Conclusion or inference

Measure of dispersion

→ Range, IQR, Std, Variance

Measures of central tendency

→ Mean, Median, Mode

Data Distribution

Skewed distribution

normal/ Symmetric distribution

Central Limit Theorem

Application in a Visualization & Plots

→ Box Plot, Bar chart, scatterplot, histogram

Real time industry → Measures of Position

↳ Percentiles,

Quartiles,

Z-Score

Mean / Median → Numeric data

Mean → average

Median → middle value

{23, 32, 11, 15, 17, 47}

1 1 1 1 1 1

Sort (Ascending order)

{11, 15, 17, 23, 32, 47}

{23, 32, 11, 15, 17, 47, 32, 11, 32}

32 — top

3 — freq

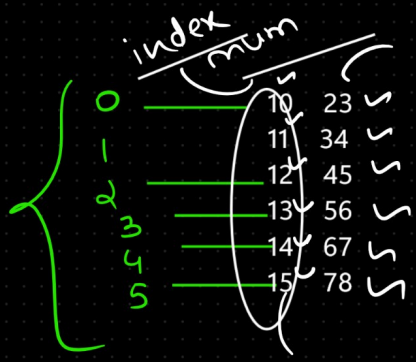
$\frac{(17+23)}{2} = \frac{41}{2}$

$\Rightarrow 20.5$

Mode → counts the frequency of all elements & returns the highest frequency

↓

Categorical data



index

get the access of the value

num of records = 6

loc | iloc

customized index | default index

loc [[12, 13]]

↳ iloc [2, 3]