# pandas Python data frames

## CONCEPTS

A DataFrame is a set of Series (columns) sharing one index (row labels). Most operations align on that index, not on position.

**Read** read_csv, read_parquet
↓ inspect
**Clean** select, astype, fillna
↓ transform
**Shape** groupby, merge, pivot
↓ export
**Write** to_parquet, to_csv

**Copy-on-Write**
pandas 3 uses Copy-on-Write: any derived frame or column acts as a copy. Chained assignment like df["a"][0] = 1 never changes df; use df.loc[0, "a"] = 1.

## READ AND WRITE

**pd.read_csv("*<file>*")**
Read a CSV file

**pd.read_parquet("*<file>*")**
Read a Parquet file

**df.to_csv("*<file>*", index=False)**
Write CSV without the index

## INSPECT

**df.head(*<n>*) / df.tail(*<n>*)**
First or last rows (default 5)

**df.info()**
Columns, dtypes, non-null counts

**df.describe()**
Summary statistics

**df["*<col>*"].value_counts()**
Count each distinct value

## SELECT

**df["*<col>*"]**
One column as a Series

## SELECT (cont.)

**df[["*<a>*", "*<b>*"]]**
Several columns as a DataFrame

**df.loc[*<rows>*, *<cols>*]**
Select by label

**df[df["*<col>*"] > 0]**
Filter rows with a mask

## CHANGE COLUMNS

**df["*<new>*"] = df["*<a>*"] * 2**
Add or overwrite a column

**df.rename(columns={"*<old>*": "*<new>*"})**
Rename columns

**df.drop(columns=["*<col>*"])**
Remove columns

## MISSING DATA

**df.isna().sum()**
Count missing values per column

**df.dropna()**
Drop rows with any missing value

**df.fillna(*<value>*)**
Fill missing values

## SORT AND DEDUPE

**df.sort_values("*<col>*", ascending=False)**
Sort rows by a column

## GROUP AND AGGREGATE

**df.groupby("*<key>*")["*<col>*"].sum()**
Total per group

## COMBINE

**df.merge(*<other>*, on="*<key>*")**
Join on a column (inner)

**pd.concat([*<df1>*, *<df2>*], ignore_index=True)**
Stack frames vertically