Swiss UMEF
UNIVERSITY OF APPLIED SCIENCES INSTITUTE

Big Data Analytics

07: In-Memory Analytics with Pandas. Grouping and Aggregating Data

Instructor: Oleh Tymchuk

#07: Agenda

- Introduction
- Grouping
- Aggregation
- Combination
- Data Visualization
- Practical cases

Introduction

What is Grouping and Aggregating?

Grouping:

- Splitting data into groups based on one or more criteria (e.g., categories, regions, or time periods).
- Example: Grouping sales data by region or product category.

Aggregating:

- Applying a function (e.g., sum, mean, count) to each group to summarize the data.
- Example: Calculating the total sales or average profit for each group.

Combination:

- Grouping and aggregating together to uncover patterns and insights within subgroups.

Why is Grouping and Aggregating Important?

Summarize Large Datasets:

- Break down complex data into manageable and meaningful chunks.

Analyze Patterns:

- Identify trends and relationships within subgroups (e.g., sales performance by region).

Prepare for Visualization:

- Create summarized data for effective charts and graphs.

Support Decision-Making:

- Provide actionable insights for businesses and data-driven strategies.

Real-World Applications

Business:

- Summarize sales by region, product, or time period
- Analyze customer behavior by demographic groups

Finance:

- Calculate average revenue or profit by category

Data Science:

- Feature engineering for machine learning models
- Preprocessing data for visualization or reporting

Grouping

Key Grouping Methods in Pandas

Method	How It Works	Example Use Case
groupby()	Splits data into groups based on one or more columns	Group sales data by region or product
pivot_table()	Creates a summary table by grouping and aggregating data across rows and columns	Summarize sales by region and product
resample()	Aggregates time-series data into fixed intervals (e.g., days, weeks)	Calculate weekly average sales
crosstab()	Computes frequency tables for combinations of categorical variables	Analyze frequency of products by region

Aggregation

Key Aggregation Methods in Pandas

Method	How It Works	Example Use Case
sum()	Calculates the total of numeric values	Total sales per region
mean()	Computes the average of numeric values	Average profit per product
count()	Counts the number of non-null values	Number of transactions per customer
min() / max()	Finds the minimum or maximum value	Identify the highest and lowest sales

Combination

Key Grouping Methods in Pandas

Method	How It Works	Example Use Case
groupby() + agg()	Groups data and applies multiple aggregation functions	Calculate total sales and average profit by region
pivot_table() + groupby()	Enhances pivot tables with extra calculations	Multi-dimensional product performance analysis
resample() + sum()	Aggregates time-series data into intervals and summarizes	Calculate monthly total revenue
crosstab() + normalize	Computes frequency tables with normalized values	Analyze percentage distribution of products by region

Data Visualization

Types of charts

Grouping Visualizations:

- Bar charts and heatmaps are ideal for showing summarized data by categories.

Aggregation Visualizations:

- Bar charts, line charts, and pie charts help visualize totals, averages, and distributions.

Combination Visualizations:

- Grouped bar charts and area charts are great for showing multiple aggregated metrics.

Practical cases

Q&A