Swiss UMEF

UNIVERSITY OF APPLIED SCIENCES INSTITUTE

Big Data Analytics
大数据分析

# 08: In-Memory Analytics with Pandas. Grouping and Aggregating Data
# 08: 使用 Pandas 进行内存分析 数据分组和聚合

Instructor: Oleh Tymchuk
授课教师：奥勒·特姆恰克

#08: Agenda 课程安排

- Introduction 介绍
- Grouping 分组
- Aggregation 聚合
- Combination 组合
- Data Visualization 数据可视化
- Practical cases 实际案例

Introduction 介绍

What is Grouping and Aggregating? 什么是分组和聚合？

Grouping 分组:

- Splitting data into groups based on one or more criteria (e.g., categories, regions, or time periods).
- **根据一个或多个**标准（**例如**，类别、地区或时间段）将数据分组。
- Example: Grouping sales data by region or product category.
  示例：**按地区或**产品类别对销售数据进行分组。

Aggregating 聚合:

- Applying a function (e.g., sum, mean, count) to each group to summarize the data.
  对每个组应用函数（例如 sum、mean、count）**来**汇总数据。
- Example: Calculating the total sales or average profit for each group.
  示例：计算每个组的总销售额或平均利润。

Combination 组合:

- Grouping and aggregating together to uncover patterns and insights within subgroups.
  将分组和聚合结合起来，以揭示子群体内的模式和洞见。

Why is Grouping and Aggregating Important? 为什么分组和聚合重要？

Summarize Large Datasets 汇总大型数据集：:

- Break down complex data into manageable and meaningful chunks.
  将复杂数据分解为易于管理且有意义的数据块。

Analyze Patterns 分析模式：:

- Identify trends and relationships within subgroups (e.g., sales performance by region). 识别子群体内的趋势和关系（例如，按地区划分的销售业绩）。

Prepare for Visualization 可视化准备:

- Create summarized data for effective charts and graphs.
  为有效的图表和图形创建汇总数据。

Support Decision-Making 支持决策:

- Provide actionable insights for businesses and data-driven strategies.
  为企业和数据驱动战略提供可行的见解。

Real-World Applications 实际应用

Business 商业:

- Summarize sales by region, product, or time period **按地区**、产品或时间段汇总销售额
- Analyze customer behavior by demographic groups **按人口**统计群体分析客户行为

Finance 金融:

- Calculate average revenue or profit by category **按**类别计算平均收入或利润

Data Science 数据科学:

- Feature engineering for machine learning models **机器学**习模型的特征工程
- Preprocessing data for visualization or reporting **用于可**视化或报告的数据预处理

Grouping 分组

Key Grouping Methods in Pandas Pandas 中的关键分组方法

Method 方法	How It Works 作用原理	Example Use Case 用例
groupby()	Splits data into groups based on one or more columns 根据一个或多个列将数据拆分为组	Group sales data by region or product 按地区或产品分组销售数据
pivot_table()	Creates a summary table by grouping and aggregating data across rows and columns 通过对行和列中的数据进行分组和聚合来创建汇总表	Summarize sales by region and product 按地区和产品汇总销售额
resample()	Aggregates time-series data into fixed intervals (e.g., days, weeks) 将时间序列数据聚合到固定间隔 (例如：天、周）	Calculate weekly average sales 计算每周平均销售额
crosstab()	Computes frequency tables for combinations of categorical variables 计算分类变量组合的频率表	Analyze frequency of products by region 按地区分析产品频率

Aggregation 聚合

Key Aggregation Methods in Pandas Pandas 中的关键聚合方法

Method 方法	How It Works 作用原理	Example Use Case 用例
sum()	Calculates the total of numeric values 计算数值的总和	Total sales per region 每个地区的总销售额
mean()	Computes the average of numeric values 计算数值的平均值	Average profit per product 每个产品的平均利润
count()	Counts the number of non-null values 统计非空值的数量	Number of transactions per customer 每位客户的交易次数
min() / max()	Finds the minimum or maximum value 查找最小值或最大值	Identify the highest and lowest sales 确定最高和最低销售额

Combination 组合

Key Grouping Methods in Pandas Pandas 中的关键分组方法

Method 方法	How It Works 作用原理	Example Use Case 用例
groupby() + agg()	Groups data and applies multiple aggregation functions 对数据进行分组并应用多个聚合函数	Calculate total sales and average profit by region 按地区计算总销售额和平均利润
pivot_table() + groupby()	Enhances pivot tables with extra calculations 通过额外计算增强数据透视表	Multi-dimensional product performance analysis 多维度产品性能分析
resample() + sum()	Aggregates time-series data into intervals and summarizes 将时间序列数据聚合到区间并进行汇总	Calculate monthly total revenue 计算月度总收入
crosstab() + normalize	Computes frequency tables with normalized values 计算具有归一化值的频率表	Analyze percentage distribution of products by region 分析各地区产品的百分比分布

Data Visualization 数据可视化

Types of charts 图表类型

Grouping Visualizations 分组可视化:

- Bar charts and heatmaps are ideal for showing summarized data by categories. 条形图和热力图非常适合按类别显示汇总数据。

Aggregation Visualizations 聚合可视化:

- Bar charts, line charts, and pie charts help visualize totals, averages, and distributions. 条形图、线图和饼图有助于直观地显示总数、平均值和分布。

Combination Visualizations 组合可视化:

- Grouped bar charts and area charts are great for showing multiple aggregated metrics. 分组条形图和面积图非常适合显示多个汇总指标。

Practical cases 实际案例

Q&A 问答