**💡 PHẦN 4 — THỰC HÀNH ỨNG DỤNG / PRACTICAL APPLICATIONS**

**📘 Tổng quan**

Các notebook này rèn kỹ năng **phân tích dữ liệu thực tế**:  
lọc, nhóm, tổng hợp, vẽ biểu đồ, so sánh giữa các biến (giới tính, năm, tiền tip, v.v.).

**🧩 1. Làm việc với DataFrame thực tế / Working with Real DataFrames**

| **Lệnh / Hàm** | **Giải thích (Việt)** | **English Explanation** | **Khi nào dùng / When to use** | **Tham số chính (Parameters)** |
| --- | --- | --- | --- | --- |
| pd.read\_csv('file.csv') | Đọc dữ liệu CSV | Read data from CSV file | Dùng cho dữ liệu dạng bảng | encoding, sep |
| df.head() | Xem vài dòng đầu | Display first rows | Kiểm tra cấu trúc ban đầu | n: số dòng |
| df.tail() | Xem vài dòng cuối | Display last rows | Kiểm tra cuối bảng | — |
| df.info() | Thông tin dữ liệu | Data info summary | Khi cần biết số dòng, kiểu dữ liệu | — |
| df.describe() | Thống kê mô tả | Summary statistics | Khi muốn tóm tắt dữ liệu số | include='all' |

**👩‍🍼 2. Phân tích dữ liệu Baby Names**

| **Lệnh** | **Giải thích (Việt)** | **English** | **Khi nào dùng** | **Tham số** |
| --- | --- | --- | --- | --- |
| df['year'].unique() | Danh sách các năm duy nhất | Unique years in data | Khi xác định giai đoạn dữ liệu | — |
| df[df['year']==2000] | Lọc theo năm 2000 | Filter rows by year | Khi phân tích một năm cụ thể | — |
| df.groupby('year')['births'].sum() | Tổng số trẻ sinh theo năm | Total births by year | Khi cần thống kê xu hướng thời gian | — |
| df.groupby('sex')['births'].sum() | Tổng số trẻ theo giới tính | Total births by gender | Khi so sánh giới | — |
| .sort\_values(by='births', ascending=False) | Sắp xếp theo số sinh | Sort by number of births | Tìm tên phổ biến nhất | — |
| .head(10) | Lấy top 10 tên phổ biến | Get top 10 names | — | — |
| .pivot\_table(values='births', index='year', columns='sex', aggfunc='sum') | Bảng tổng hợp theo năm–giới | Pivot table by year and sex | Khi so sánh 2 biến cùng lúc | — |
| .plot(kind='line') | Vẽ xu hướng sinh theo thời gian | Line plot of births | Thể hiện biến động qua năm | title, xlabel, ylabel |

**💡 Kỹ năng học được:**

* Sử dụng groupby để **tổng hợp dữ liệu theo năm hoặc giới tính**
* Dùng pivot\_table để tạo **bảng tổng hợp động (giống Excel)**
* Dùng plot(kind='line') để trực quan hóa xu hướng.

**🍽 3. Phân tích dữ liệu Tips (tiền tip nhà hàng)**

| **Lệnh** | **Giải thích (Việt)** | **English** | **Khi nào dùng** | **Tham số** |
| --- | --- | --- | --- | --- |
| df['tip\_pct'] = df['tip'] / df['total\_bill'] | Tạo cột tỷ lệ tip | Create tip percentage column | Khi muốn phân tích tỷ lệ | — |
| df.groupby('sex')['tip\_pct'].mean() | Trung bình tip theo giới | Mean tip % by gender | So sánh hành vi giữa nam & nữ | — |
| df.groupby(['sex','smoker'])['tip'].mean() | Tip trung bình theo giới & hút thuốc | Mean tip grouped by sex and smoker | Phân tích 2 yếu tố cùng lúc | — |
| df.groupby(['day'])['tip\_pct'].mean() | Tỷ lệ tip theo ngày | Mean tip by day | Xem ngày nào tip cao nhất | — |
| .agg({'tip':'mean','total\_bill':'sum'}) | Tổng hợp nhiều hàm cùng lúc | Apply multiple aggregations | Khi muốn nhiều kết quả trên cùng nhóm | — |
| .corr() | Tương quan giữa các biến | Correlation matrix | Xem biến nào liên hệ mạnh | — |
| plt.scatter(df['total\_bill'], df['tip']) | Biểu đồ phân tán bill–tip | Scatter plot of bill vs tip | Thấy mối quan hệ trực tiếp | — |
| plt.boxplot(df['tip']) | Biểu đồ hộp cho tip | Boxplot for tip distribution | Phát hiện giá trị ngoại lai | — |
| df['day'].value\_counts() | Đếm số giao dịch theo ngày | Count by day | Kiểm tra phân bố dữ liệu | normalize=True để lấy tỉ lệ |
| sns.violinplot(x='day', y='tip', data=df) *(nếu có seaborn)* | Biểu đồ violin tip theo ngày | Violin plot | Khi muốn xem phân bố chi tiết | — |

**💡 Kỹ năng học được:**

* Tính toán cột mới (derived columns)
* Gom nhóm & tính trung bình, tổng, tỉ lệ
* Dùng biểu đồ hộp hoặc scatter để phân tích ngoại lai và mối quan hệ.

**🧾 4. Phân tích dữ liệu PPXS (Phân phối xác suất mô phỏng)**

| **Lệnh** | **Giải thích (Việt)** | **English** | **Khi nào dùng** | **Tham số** |
| --- | --- | --- | --- | --- |
| np.random.normal(mu, sigma, size) | Sinh dữ liệu chuẩn | Generate normal distribution | Khi mô phỏng biến liên tục | — |
| np.mean(data) | Trung bình mẫu | Sample mean | Khi ước lượng trung bình quần thể | — |
| np.std(data, ddof=1) | Độ lệch chuẩn mẫu | Sample std | Khi cần độ phân tán | ddof=1 cho mẫu |
| plt.hist(data, bins=30, density=True) | Histogram dữ liệu | Data histogram | Khi so sánh với phân phối lý thuyết | bins, density |
| norm.pdf(x, loc, scale) | Hàm mật độ chuẩn | Normal PDF | Vẽ đường cong lý thuyết | loc, scale |
| plt.plot(x, pdf) | Vẽ đường cong PDF | Plot normal curve | Minh họa sự khớp giữa mô phỏng & lý thuyết | — |
| (data >= 180).sum()/len(data) | Tính tỷ lệ ≥ ngưỡng | Compute proportion ≥ threshold | Khi phân tích xác suất thực nghiệm | — |
| for i in range(100): ... | Lặp lại mô phỏng nhiều lần | Repeat simulation loop | Khi khảo sát biến động mẫu | — |
| .append() | Thêm giá trị vào danh sách | Append value to list | Khi lưu kết quả qua vòng lặp | — |

**🔄 5. Kết hợp Pandas + Matplotlib trong thực hành**

| **Lệnh** | **Giải thích (Việt)** | **English** | **Khi nào dùng** | **Tham số** |
| --- | --- | --- | --- | --- |
| df.plot(kind='bar') | Biểu đồ cột | Bar plot | So sánh nhóm | title, color, legend |
| df.plot(kind='line') | Biểu đồ đường | Line plot | Thể hiện xu hướng | — |
| df.plot(kind='hist') | Biểu đồ tần suất | Histogram | Phân bố dữ liệu | bins, alpha |
| df.plot(subplots=True) | Mỗi cột 1 biểu đồ con | Multiple subplots | Khi muốn so sánh nhiều biến | — |
| plt.figure(figsize=(8,5)) | Điều chỉnh kích thước | Figure size | Khi biểu đồ bị chật | — |
| plt.tight\_layout() | Căn lề biểu đồ con | Adjust subplot spacing | Khi bị đè nhãn | — |

**📊 6. Các hàm tổng hợp cao cấp đã xuất hiện**

| **Lệnh** | **Giải thích (Việt)** | **English** | **Khi nào dùng** | **Tham số** |
| --- | --- | --- | --- | --- |
| .apply(func) | Áp dụng hàm cho từng cột/dòng | Apply custom function | Khi cần xử lý đặc biệt | axis=0/1 |
| .map(func) | Biến đổi từng phần tử trong Series | Map function to series | Xử lý cột dạng chuỗi hoặc số | — |
| .isin(list) | Lọc giá trị có trong danh sách | Filter by list membership | Khi lọc theo nhiều giá trị | — |
| .between(a,b) | Lọc trong khoảng | Filter between range | Khi lọc dữ liệu theo điều kiện | — |
| .duplicated() | Kiểm tra dòng trùng | Check duplicates | Làm sạch dữ liệu | subset, keep |
| .drop\_duplicates() | Xóa dòng trùng | Drop duplicate rows | Khi dọn dữ liệu | — |

**🧠 7. Kỹ năng, công thức & quy trình rút ra**

| **Bước** | **Ý nghĩa (Việt)** | **English** |
| --- | --- | --- |
| 1️⃣ **Đọc & khám phá dữ liệu** | Hiểu cấu trúc, kiểu, số lượng, giá trị null | Data exploration |
| 2️⃣ **Làm sạch dữ liệu (cleaning)** | Xử lý NaN, trùng lặp, kiểu dữ liệu | Data cleaning |
| 3️⃣ **Phân tích thống kê mô tả** | Tính mean, median, var, std, value\_counts | Descriptive statistics |
| 4️⃣ **Phân nhóm & tổng hợp** | groupby, pivot\_table, agg | Aggregation & grouping |
| 5️⃣ **Trực quan hóa kết quả** | plot, hist, scatter, boxplot | Visualization |
| 6️⃣ **Đưa ra nhận xét / kết luận** | So sánh nhóm, xác định xu hướng | Interpretation |

**📘 Tổng kết Phần 4**

* **Bạn đã học cách phân tích dữ liệu thực tế bằng Pandas:**  
  đọc dữ liệu, tạo cột mới, nhóm, tính trung bình, đếm tần suất.
* **Kết hợp Matplotlib để trực quan hóa:**  
  biểu đồ đường, cột, phân tán, hộp.
* **Ứng dụng vào mô phỏng xác suất:**  
  dùng NumPy + SciPy để sinh dữ liệu và kiểm định.

👉 Tính đến đây, bạn đã có trọn bộ:

| **Phần** | **Chủ đề** | **Mục tiêu** |
| --- | --- | --- |
| **1** | Pandas | Làm việc với dữ liệu dạng bảng |
| **2** | Matplotlib | Trực quan hóa dữ liệu |
| **3** | Phân phối xác suất | Mô phỏng và thống kê xác suất |
| **4** | Ứng dụng (Tips / Baby Names / Ppxs) | Phân tích dữ liệu thực tế |