**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**--------\*\*\*--------**

A blue button with white text

AI-generated content may be incorrect.

BÁO CÁO THỰC HÀNH

MÔN PHÂN TÍCH DỮ LIỆU

**TÊN ĐỀ TÀI PHÂN TÍCH DỮ LIỆU THỰC TẾ VỚI PYTHON**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | **ThS. Đỗ Như Tài** |
| **Nhóm sinh viên thực hiện:** |  |
| Hồ Thị Thanh Thảo | 3122410389 |
| Nguyễn Thị Hồng Thắm | 3122410392 |
| Phan Văn Thảo | 3122410391 |
| Nguyễn Hoàng Thiên Bảo | 3122410019 |

***Thành phố Hồ Chí Minh, tháng 10 năm 2025***

# LỜI MỞ ĐẦU

Phân tích dữ liệu là một kỹ năng thiết yếu trong thời đại số hóa hiện nay. Với sự phát triển nhanh chóng của công nghệ, lượng dữ liệu được tạo ra mỗi ngày tăng theo cấp số nhân. Dữ liệu này tồn tại ở khắp nơi - từ các nền tảng mạng xã hội, thương mại điện tử, y tế cho đến chính phủ. Tuy nhiên, dữ liệu thô không có giá trị nếu chúng ta không biết cách khai thác và phân tích để rút ra những thông tin có ý nghĩa từ đó.

Python đã trở thành một trong những ngôn ngữ lập trình phổ biến nhất cho phân tích dữ liệu nhờ vào những thư viện mạnh mẽ như Pandas, NumPy và Matplotlib. Những công cụ này cung cấp cho chúng ta khả năng xử lý, trực quan hóa và phân tích dữ liệu một cách hiệu quả.

Báo cáo này được thực hiện dựa trên Chương 13 - "Data Analysis Examples" của cuốn sách "Python for Data Analysis" (O'Reilly 2022) của tác giả Wes McKinney. Chương này cung cấp những ví dụ thực tế về cách ứng dụng các kỹ thuật phân tích dữ liệu với Python vào các tập dữ liệu khác nhau từ các lĩnh vực khác nhau.

Trong báo cáo này, chúng em sẽ:

* Tìm hiểu về 5 tập dữ liệu thực tế: Bitly Data, MovieLens, US Baby Names, USDA Food Database, và Federal Election Commission Data
* Học cách tải, làm sạch, và xử lý dữ liệu từ các nguồn khác nhau (JSON, CSV)
* Áp dụng các kỹ thuật phân tích như groupby, pivot\_table, merge, explode
* Trực quan hóa dữ liệu bằng Matplotlib và Seaborn
* Rút ra những insights có giá trị từ dữ liệu

Mục đích của báo cáo này không chỉ là để hiểu rõ các kỹ thuật phân tích mà còn để thấy được sự ứng dụng thực tế của chúng trong các tình huống thực tế. Qua đó, chúng em hy vọng có thể nâng cao kỹ năng phân tích dữ liệu và góp phần áp dụng Python một cách hiệu quả hơn trong công việc sau này.

**I. GIỚI THIỆU CHƯƠNG (INTRODUCTION)**

1. Mục đích của chương
2. Các dataset được sử dụng
3. Các kỹ thuật chính được áp dụng
4. Cấu trúc học tập

**II. BITLY DATA FROM 1.USA.GOV (Phần 13.1)**

**2.1 Giới thiệu về dữ liệu**

* Nguồn gốc dữ liệu (URL shortening service)
* Định dạng JSON
* Cấu trúc các trường dữ liệu

**2.2 Đọc và xử lý dữ liệu JSON**

* Sử dụng module json trong Python
* Chuyển đổi JSON thành danh sách dictionary
* Xử lý lỗi KeyError

**2.3 Phân tích Time Zones (Múi giờ)**

* **Cách 1: Pure Python**
  + Sử dụng dict để đếm tần suất
  + Sử dụng defaultdict
  + Sử dụng collections.Counter
* **Cách 2: Với Pandas**
  + Tạo DataFrame từ danh sách records
  + Sử dụng value\_counts()
  + Xử lý missing values với fillna()

**2.4 Phân tích User Agent (Browser/Device)**

* Tách token đầu tiên từ chuỗi agent
* Đếm các browser phổ biến
* Phân tích Windows vs Non-Windows users

**2.5 Groupby và Visualization**

* Nhóm dữ liệu theo time zone và OS
* Sử dụng unstack() để tạo pivot table
* Vẽ biểu đồ với seaborn
* Chuẩn hóa dữ liệu (normalization)

**III. MOVIELENS 1M DATASET (Phần 13.2)**

**3.1 Giới thiệu MovieLens dataset**

* Cấu trúc dữ liệu (3 bảng)
* Kích thước dataset

**3.2 Tải và khám phá dữ liệu**

* Sử dụng pd.read\_table()
* Hiểu các trường dữ liệu
* Xem preview dữ liệu

**3.3 Merge/Join dữ liệu**

* Kết hợp các bảng lại với nhau
* Sử dụng pd.merge()
* Tạo DataFrame unified

**3.4 Phân tích Mean Ratings**

* Sử dụng pivot\_table()
* Tính điểm đánh giá trung bình theo giới tính
* Lọc dữ liệu (movies với 250+ ratings)

**3.5 Phân tích Disagreement**

* Tìm phim chia rẽ (divisive movies)
* Tính độ lệch chuẩn (standard deviation)
* So sánh ưa thích giữa nam và nữ

**3.6 Phân tích theo Genre**

* Split genres từ chuỗi pipe-separated
* Sử dụng explode() method
* Phân tích ratings theo genre và age group

**IV. US BABY NAMES 1880-2010 (Phần 13.3)**

**4.1 Giới thiệu dataset**

* Nguồn dữ liệu (US Social Security Administration)
* Định dạng file
* Số lượng dữ liệu

**4.2 Tải và kết hợp dữ liệu**

* Đọc các file CSV hàng năm
* Sử dụng pd.concat() để kết hợp
* Thêm cột year

**4.3 Tính toán tỷ lệ (Proportion)**

* Tính tỷ lệ sinh cho mỗi tên
* Sử dụng groupby().apply()
* Kiểm chứng (sanity check)

**4.4 Phân tích Naming Trends**

* Tạo top 1000 names
* Vẽ biểu đồ theo thời gian
* So sánh tên nam vs nữ

**4.5 Phân tích Diversity**

* Thử đo diversity của tên
* Tính cumulative sum
* Sử dụng searchsorted()

**4.6 The "Last Letter" Revolution**

* Phân tích chữ cái cuối cùng
* So sánh năm 1910, 1960, 2010
* Khám phá xu hướng tên nam gái

**4.7 Tên từ nam sang nữ (và ngược lại)**

* Phân tích ví dụ: Leslie/Lesley
* Tracking sự thay đổi theo giới tính qua thời gian

**V. USDA FOOD DATABASE (Phần 13.4)**

**5.1 Giới thiệu dataset**

* Cấu trúc dữ liệu JSON
* Thông tin dinh dưỡng (nutrients)
* Thông tin thực phẩm (food groups)

**5.2 Đọc và xử lý JSON**

* Load file JSON
* Hiểu cấu trúc nested data
* Extract fields

**5.3 Làm sạch dữ liệu**

* Rename cột để tránh trùng lặp
* Xử lý duplicates
* Merge info với nutrients

**5.4 Phân tích Nutrients**

* Nhóm theo nutrient và food group
* Tính median values
* Tìm food có nutrient cao nhất
* Visualize dữ liệu

**VI. 2012 FEDERAL ELECTION COMMISSION DATABASE (Phần 13.5)**

**6.1 Giới thiệu FEC dataset**

* Nguồn dữ liệu
* Kích thước dataset
* Các trường quan trọng

**6.2 Thêm thông tin chính trị**

* Mapping candidate names to parties
* Lọc dữ liệu (dương tính)
* Focus vào Obama vs Romney

**6.3 Phân tích Donations by Occupation**

* Đếm donations theo nghề nghiệp
* Làm sạch dữ liệu occupation/employer
* So sánh Democrats vs Republicans
* Visualize với biểu đồ

**6.4 Phân tích Top Donors**

* Tìm occupation top contributors
* Tìm employer top contributors
* So sánh Obama vs Romney

**6.5 Bucketing Donation Amounts**

* Sử dụng pd.cut() để binning
* Phân tích kích thước donations
* Tính toán percentages

**6.6 Phân tích Donations by State**

* Nhóm theo state
* Tính tổng donations
* So sánh percentage

**VII. PHẦN KẾT LUẬN (CONCLUSION)**

**7.1 Tóm tắt kỹ thuật chính**

* Data loading (JSON, CSV)
* Data cleaning & wrangling
* Groupby & aggregation
* Visualization

**7.2 Best practices**

* Xử lý missing data
* Tối ưu hóa performance
* Naming conventions

**7.3 Gợi ý học tập tiếp theo**

* Áp dụng vào projects riêng
* Khám phá datasets khác