|  |  |
| --- | --- |
| Tên | Công việc |
| Hồ Thị Thanh Thảo | Phân tích dataset MovieLens (đánh giá phim), làm báo cáo Code + báo cáo + slide |
| Nguyễn Thị Hồng Thắm | Phân tích dataset bitly (timezone) Code + báo cáo + slide |
| Nguyễn Hoàng Thiên Bảo | Phân tích dataset babynames (tên trẻ em) Phân tích dataset USDA (dinh dưỡng) Code + báo cáo + slide |
| Phan Văn Thảo | Phân tích dataset FEC (bầu cử), tổng hợp slide |
| Code + báo cáo + slide |

PHÂN CÔNG

**Lưu ý:**

Đặt tên file code, báo cáo và slide theo tên dataset của mình. VD: bitly, movielens, babynames, USDA, FEC.

File code nhớ trang trí

Yêu cầu tối thiểu cho mỗi dataset thầy có nhắn cho zalo gr, mọi người lên gr và làm theo.

Mục lục báo cáo sẽ như dưới đây, mọi người làm theo mẫu này (đây chỉ là gợi ý nên chưa đầy đủ hoặc có thể sai, mọi người có thể bổ sung thêm và chỉnh sửa cho dataset của mình).

**I. GIỚI THIỆU CHUNG**

1. Mục đích của chương
2. Công cụ, thư viện và kỹ thuật sử dụng (Python, Pandas, NumPy, Matplotlib, Seaborn)
3. Quy trình phân tích dữ liệu tổng quát (Data Loading → Cleaning → EDA → Visualization → Insights)
4. Danh sách các bộ dữ liệu được sử dụng trong chương

**II. PHÂN TÍCH DỮ LIỆU BITLY DATA FROM 1.USA.GOV**

2.1. Tổng quan dữ liệu  
2.2. Mô tả dữ liệu (Data Description)

* Cấu trúc dữ liệu JSON, các trường thông tin (tz, r, a, c, gr, cy, t, hc)
* Thống kê tổng quan: số lượng bản ghi, quốc gia, múi giờ, referrer
* Biểu đồ: Histogram tần suất truy cập, Pie chart quốc gia, Bar chart hệ điều hành

2.3. Câu hỏi nghiên cứu (Research Questions)  
2.4. Phân tích mô tả (Descriptive Analysis)  
2.5. Phân tích tương quan (Correlation Analysis)  
2.6. Phân tích mở rộng (Extended Analysis)  
2.7. Kết quả và nhận xét (Results & Discussion)

**III. PHÂN TÍCH DỮ LIỆU MOVIELENS 1M**

3.1. Tổng quan dữ liệu  
3.2. Mô tả dữ liệu (Data Description)

* Ba bảng: users, movies, ratings
* Số dòng, số cột, giá trị thiếu, kiểu dữ liệu
* Thống kê cơ bản: mean, median, std
* Biểu đồ: Histogram rating, Boxplot giới tính, Heatmap tương quan

3.3. Câu hỏi nghiên cứu  
3.4. Phân tích mô tả (Descriptive Analysis)  
3.5. Phân tích khám phá (Exploratory Analysis)  
3.6. Phân tích dự đoán (Predictive Analysis)  
3.7. Phân tích so sánh (Comparative Analysis)  
3.8. Phân tích mở rộng (Extended Analysis)  
3.9. Kết quả và nhận xét (Results & Discussion)

**IV. PHÂN TÍCH DỮ LIỆU US BABY NAMES (1880–2010)**

4.1. Tổng quan dữ liệu  
4.2. Mô tả dữ liệu (Data Description)

* Nguồn: SSA, định dạng CSV (mỗi năm 1 file)
* Cấu trúc cột: name, sex, births
* Thống kê cơ bản: số lượng tên, giới tính, tổng trẻ sinh mỗi năm
* Biểu đồ: Line chart số trẻ sinh, Histogram độ phổ biến tên, Boxplot giới tính

4.3. Câu hỏi nghiên cứu  
4.4. Phân tích mô tả (Descriptive Analysis)  
4.5. Phân tích khám phá (Exploratory Analysis)  
4.6. Phân tích so sánh (Comparative Analysis)  
4.7. Phân tích dự đoán (Predictive Analysis)  
4.8. Kết quả và nhận xét (Results & Discussion)

**V. PHÂN TÍCH DỮ LIỆU USDA FOOD DATABASE**

5.1. Tổng quan dữ liệu  
5.2. Mô tả dữ liệu (Data Description)

* Nguồn: Bộ Nông nghiệp Hoa Kỳ (USDA)
* Dạng JSON, cấu trúc: food groups, nutrients, description
* Thống kê: số nhóm thực phẩm, số chất dinh dưỡng, phạm vi giá trị
* Biểu đồ: Histogram năng lượng, Boxplot protein/fat, Heatmap tương quan

5.3. Câu hỏi nghiên cứu  
5.4. Phân tích mô tả (Descriptive Analysis)  
5.5. Phân tích so sánh (Comparative Analysis)  
5.6. Phân tích tương quan (Correlation Analysis)  
5.7. Phân tích dự đoán (Predictive Analysis)  
5.8. Kết quả và nhận xét (Results & Discussion)

**VI. PHÂN TÍCH DỮ LIỆU 2012 FEDERAL ELECTION COMMISSION (FEC)**

6.1. Tổng quan dữ liệu  
6.2. Mô tả dữ liệu (Data Description)

* Dạng CSV, thuộc tính: cand\_nm, contbr\_occupation, contbr\_employer, contb\_receipt\_amt, contbr\_st
* Thống kê: số lượng bản ghi, số người quyên góp, tổng tiền quyên góp, occupation phổ biến
* Biểu đồ: Histogram số tiền quyên góp, Boxplot nghề nghiệp, Bar chart theo bang

6.3. Câu hỏi nghiên cứu  
6.4. Phân tích mô tả (Descriptive Analysis)  
6.5. Phân tích so sánh (Comparative Analysis)  
6.6. Phân tích tương quan (Correlation Analysis)  
6.7. Phân tích mở rộng (Extended Analysis)  
6.8. Kết quả và nhận xét (Results & Discussion)

**VII. KẾT LUẬN CHUNG**

7.1. Tổng kết kỹ thuật phân tích đã áp dụng  
7.2. So sánh sự khác biệt giữa các bộ dữ liệu  
7.3. Bài học rút ra từ quá trình phân tích  
7.4. Hướng phát triển và mở rộng trong tương lai

**VIII. TÀI LIỆU THAM KHẢO (IEEE)**