1. Mục tiêu của Dự án

Dự án này nhằm trang bị cho sinh viên các kỹ năng nền tảng cần thiết để trực quan hóa dữ liệu một cách hiệu quả bằng Python. Thông qua việc học và tận dụng các thư viện Python như Matplotlib, Seaborn và Pandas để tạo ra nhiều loại hình trực quan hóa tĩnh, tương tác và hoạt hình, sinh viên sẽ khám phá các mối quan hệ ẩn trong tập dữ liệu mà họ đang tập trung phân tích trong một miền cụ thể.

---

2. Lưu ý và Ràng buộc

Bạn phải tuân theo các ràng buộc sau khi thực hiện dự án này:

- **Công việc không có báo cáo sẽ không được chấm điểm.**
- **Thành viên không đóng góp vào dự án sẽ không nhận được điểm.**
- **Nguồn tham khảo (nếu có)** phải được ghi đầy đủ trong phần Tham khảo của báo cáo. Lưu ý rằng cần phải phân biệt rõ ràng giữa việc tham khảo và đạo văn.
- **Cá nhân hoặc nhóm vi phạm đạo đức học thuật và gian lận sẽ bị phạt 0 điểm cho tất cả các môn học.**
- **Bài nộp phải được nén bằng định dạng .zip. KHÔNG CHẤP NHẬN các định dạng khác.** Nếu kích thước lớn hơn 20MB, hãy tải nó lên một dịch vụ lưu trữ bên ngoài như Google Drive hoặc OneDrive, sau đó gửi liên kết. Cuối cùng, vui lòng giữ liên kết ở chế độ công khai trong ít nhất 2 năm.

---

3. Nhiệm vụ

Trong bài thực hành này, bạn sẽ sử dụng Kaggle để xác định và thu thập một tập dữ liệu công khai phù hợp với một chủ đề mà nhóm sinh viên của bạn lựa chọn. Tập dữ liệu đã chọn phải được cấu trúc dưới dạng bảng, chứa tối thiểu năm cột dữ liệu riêng biệt và ít nhất 3.000 hàng để đảm bảo độ phức tạp và chiều sâu đủ để phân tích và trực quan hóa dữ liệu.

3.1 Nhiệm vụ 1: Thu thập dữ liệu

Nhóm của bạn sẽ thực hiện việc thu thập một tập dữ liệu và giải thích lý do lựa chọn này. Điều này bao gồm việc cung cấp một bản tường thuật chi tiết nêu rõ lý do bạn chọn tập dữ liệu đó và nêu bật sự liên quan của nó đối với các mục tiêu dự án của bạn. Hãy đề cập đến các điểm sau trong bài nộp của bạn:

- **Bối cảnh và Động cơ:** Mô tả chủ đề trung tâm hoặc câu chuyện bao quát đã thúc đẩy nhóm của bạn chọn chủ đề này. Vấn đề thực tế nào, câu hỏi, hoặc sự tò mò nào mà nó giải quyết, và tại sao nó lại hấp dẫn đối với việc trực quan hóa dữ liệu?
- **Chủ đề và Nguồn của Tập Dữ liệu:** Xác định rõ đối tượng chính của tập dữ liệu của bạn và xác định các nguồn đã được sử dụng để thu thập dữ liệu. Ví dụ, nó được thu thập từ các báo cáo của chính phủ, khảo sát, hay dữ liệu cảm biến?
- **Phương pháp Tạo Dữ liệu:** Giải thích cách mà tập dữ liệu được tạo ra. Những phương pháp hoặc quy trình nào (ví dụ: nhập liệu thủ công, thu thập tự động, thu thập qua API) được sử dụng để thu thập và tổ chức dữ liệu?
- **Khía cạnh Pháp lý và Đạo đức:** Xác nhận tính phù hợp của tập dữ liệu cho mục đích giáo dục. Nó có được công khai theo một giấy phép thích hợp (ví dụ: Creative Commons, Open Data)? Có bất kỳ hạn chế hoặc cân nhắc nào về mặt đạo đức cần lưu ý không?

3.2 Nhiệm vụ 2: Làm sạch và Tiền xử lý Dữ liệu

Dựa trên các kỹ năng đã học từ các khóa học trước như **Nhập môn Khoa học Dữ liệu** và **Lập trình cho Khoa học Dữ liệu**, bạn sẽ làm sạch và tiền xử lý tập dữ liệu của mình để chuẩn bị cho việc trực quan hóa. Bước này đảm bảo rằng dữ liệu là chính xác, nhất quán và được tối ưu hóa cho phân tích. Khám phá và ghi lại các khía cạnh sau đây:

- **Diễn giải Cấp Dòng (Row-Level Interpretation):** Mỗi dòng đại diện cho điều gì (ví dụ: một cá nhân, một sự kiện, một điểm thời gian)? Tất cả các dòng có nhất quán về ý nghĩa không, hay có những biến thể cần phải điều chỉnh để đồng nhất?
- **Định nghĩa Cột (Column Definitions):** Cung cấp lời giải thích rõ ràng về ý nghĩa của từng cột (ví dụ: tuổi, dấu thời gian, loại danh mục). Đảm bảo vai trò của các cột trong tập dữ liệu được xác định rõ ràng.
- **Loại Dữ liệu (Data Types):** Xác định kiểu dữ liệu của từng cột (ví dụ: số nguyên, số thực, chuỗi ký tự, kiểu ngày giờ). Làm nổi bật bất kỳ cột nào có kiểu dữ liệu không phù hợp cho việc trực quan hóa hoặc phân tích (ví dụ: số dưới dạng chuỗi) và đề xuất các phương pháp chuyển đổi.
- **Phân phối Giá trị (Value Distribution):** Phân tích sự phân phối của các giá trị trong mỗi cột. Đối với dữ liệu số, hãy mô tả giá trị trung bình, xu hướng trung tâm (trung bình cộng, trung vị), và độ phân tán (ví dụ: độ lệch chuẩn, khoảng giá trị). Đối với các cột dạng danh mục, hãy liệt kê các danh mục duy nhất và tần suất xuất hiện của chúng.

- **Nhu cầu Tiền xử lý:** Đánh giá liệu việc tiền xử lý có cần thiết hay không (ví dụ: xử lý giá trị thiếu, loại bỏ dữ liệu trùng lặp, chuẩn hóa dữ liệu). Nếu có, hãy nêu chi tiết các kỹ thuật cụ thể bạn sẽ áp dụng, chẳng hạn như bổ sung giá trị thiếu, loại bỏ các giá trị ngoại lai, hoặc mã hóa các biến phân loại.

3.3 Nhiệm vụ 3: Trực quan hóa Dữ liệu

Sau khi làm sạch tập dữ liệu và thực hiện phân tích dữ liệu khám phá (EDA) để phát hiện các mẫu và thông tin chi tiết, bạn sẽ tạo ra một loạt các hình ảnh trực quan để truyền đạt hiệu quả các thuộc tính và mối quan hệ của tập dữ liệu. Nhiệm vụ này nhấn mạnh việc sử dụng nhiều loại biểu đồ khác nhau để trình bày dữ liệu một cách sáng tạo và dễ hiểu, hỗ trợ đưa ra các kết luận dựa trên dữ liệu. Hãy giải quyết các yêu cầu sau:

- **Lựa chọn Thuộc tính (Attribute Selection):** Từ quá trình EDA của bạn, xác định các thuộc tính quan trọng nhất (ví dụ: biến hoặc đặc trưng) cần được trực quan hóa. Những thuộc tính này nên phản ánh các xu hướng chính, mẫu, hoặc mối quan hệ trong dữ liệu.
- **Lựa chọn Loại Biểu đồ (Chart Type Selection):** Chọn các kỹ thuật trực quan hóa phù hợp từ bộ công cụ của bạn — chẳng hạn như biểu đồ đường, biểu đồ cột, biểu đồ tán xạ, biểu đồ tròn, biểu đồ nhiệt, hoặc hộp biểu đồ — để thể hiện các thuộc tính này. Mỗi lựa chọn nên phù hợp với loại dữ liệu của thuộc tính đó và câu chuyện mà bạn muốn truyền tải.
- **Lý do Lựa chọn (Justification of Choices):** Đối với mỗi biểu đồ, cung cấp lời giải thích rõ ràng về lý do tại sao loại biểu đồ đó được chọn. Ví dụ: “Biểu đồ cột được chọn để so sánh tần suất của các loại danh mục do nó thể hiện rõ ràng sự khác biệt giữa các loại.” Nếu bạn sử dụng nhiều biểu đồ cho một thuộc tính (ví dụ: một biểu đồ cột và một biểu đồ tròn), hãy giải thích góc nhìn bổ sung mà mỗi biểu đồ cung cấp.
- **Tính Phức tạp Tăng Dần (Progressive Complexity):** Cấu trúc các hình ảnh trực quan của bạn theo thứ tự từ đơn giản đến phức tạp. Bắt đầu với các hình ảnh trực quan đơn giản (ví dụ: biểu đồ histogram của các thuộc tính đơn lẻ) và phát triển dần lên các màn hình đa biến (ví dụ: biểu đồ tán xạ hoặc ma trận tương quan) để tiết lộ các tương tác liên quan giữa các thuộc tính.
- **Khám phá Nguyên nhân và Hệ quả (Cause-and-Effect Exploration):** Điều tra các mối quan hệ nhân quả tiềm ẩn trong dữ liệu. Ví dụ: nếu trực quan hóa dữ liệu về COVID-19, hãy khám phá liệu có sự gia tăng tỷ lệ nhiễm bệnh theo thời gian hoặc theo vị trí địa lý hay không, và chọn các biểu đồ trực quan thích hợp (ví dụ: biểu đồ đường, biểu đồ phân tích hai trục) để minh họa hiệu quả mối quan hệ đó.

- **Trình bày Thông tin Chi tiết (Insight Presentation):** Sau khi tạo ra các biểu đồ trực quan, trình bày các thông tin chi tiết hoặc kết luận cụ thể được rút ra từ chúng. Ví dụ: "Biểu đồ tán xạ cho thấy mối tương quan mạnh giữa nhiệt độ và doanh số bán kem, gợi ý về một xu hướng theo mùa." Những thông tin này cần liên quan trực tiếp đến câu chuyện bạn muốn truyền tải và tăng cường sự hiểu biết về câu chuyện của dữ liệu.
- **Độ Phủ của Các Loại Biểu đồ (Breadth of Visualization Types):** Mặc dù không phải mọi mối quan hệ đều cần được trực quan hóa, hãy cố gắng kết hợp nhiều loại biểu đồ đã được học trong khóa học của bạn. Ưu tiên sự phù hợp và thông tin chi tiết thay vì việc thể hiện tràn lan, đảm bảo rằng mỗi biểu đồ đều đóng góp giá trị vào phân tích của bạn.
- **Sáng tạo và Tương tác (Innovation and Engagement):** Thử nghiệm với các kỹ thuật trực quan sáng tạo hoặc nâng cao — chẳng hạn như biểu đồ tương tác, biểu đồ cây, hoặc bản đồ không gian địa lý — để nâng cao sự tương tác và khám phá sâu sắc hơn. Những kỹ thuật này không chỉ cần đẹp mắt mà còn phải có ý nghĩa trong việc diễn giải dữ liệu.

3.4 Nhiệm vụ 4: Viết báo cáo và Trình bày kết quả

Trong nhiệm vụ cuối cùng này, bạn sẽ tổng hợp toàn bộ công việc từ Nhiệm vụ 1-3 thành một báo cáo chuyên nghiệp và tổ chức công việc tính toán của bạn trong các notebook Jupyter để đảm bảo tính rõ ràng và khả năng tái tạo. Mục tiêu là truyền đạt hiệu quả những phát hiện, biểu đồ trực quan và thông tin chi tiết của bạn tới người xem, đồng thời đảm bảo rằng quy trình của bạn được ghi chép đầy đủ và dễ tiếp cận. Thực hiện các yêu cầu sau đây:

Báo cáo Rõ ràng dưới dạng Tệp PDF (Clear Report in PDF Format):

- Biên soạn dự án của bạn thành một báo cáo trau chuốt, có cấu trúc tốt được nộp dưới dạng tệp PDF. Báo cáo cần bao gồm:
    - Giới thiệu bao quát chủ đề, động lực và mục tiêu.
    - Phần dành cho từng nhiệm vụ (thu thập dữ liệu, làm sạch dữ liệu, trực quan hóa dữ liệu), tóm tắt các phương pháp, phát hiện và thông tin chi tiết.
    - Kết luận phản ánh câu chuyện tổng thể mà dữ liệu kể lên và những điều rút ra từ các biểu đồ trực quan.
    - Định dạng phù hợp với các tiêu đề, các mục lục, và kiểu chữ nhất quán (ví dụ: kích thước, kiểu chữ) để đảm bảo tính dễ đọc và tính chuyên nghiệp.

Biểu đồ có Độ phân giải Cao trong Báo cáo (High-Resolution Charts in the Report):

- Nếu bạn nhúng các biểu đồ trực quan từ Nhiệm vụ 3 vào báo cáo, hãy đảm bảo rằng chúng có chất lượng cao.
- Các biểu đồ cần được xuất với độ phân giải cao (ví dụ: 300 DPI) để duy trì độ rõ nét và tính thẩm mỹ trong báo cáo của bạn.
- Mỗi biểu đồ cần kèm theo chú thích ngắn gọn mô tả mục đích của nó và giải thích ngắn về thông tin mà nó tiết lộ.
- Các biểu đồ cần được định cỡ phù hợp trong tài liệu (không bị nhòe hoặc quá lớn) và được gắn nhãn với tiêu đề, nhãn trục, và chú thích khi cần thiết để dễ dàng diễn giải.

Cấu trúc Rõ ràng của Notebook Jupyter (Clear Structure of Jupyter Notebooks):

Tổ chức mã và phân tích của bạn trong các notebook Jupyter, chia nhỏ công việc thành các notebook riêng biệt cho mỗi nhiệm vụ chính để nâng cao sự rõ ràng và tính mô-đun. Cụ thể:

- **Notebook Thu thập Dữ liệu (Data Collection Notebook):** Bao gồm mã để tìm kiếm, tải xuống và khám phá ban đầu tập dữ liệu từ Kaggle, cùng với các chú thích markdown giải thích lựa chọn của bạn và bối cảnh tập dữ liệu.
- **Notebook Làm sạch Dữ liệu (Data Cleaning Notebook):** Chứa tất cả các bước tiền xử lý (ví dụ: xử lý giá trị thiếu, chuyển đổi kiểu dữ liệu), với các chú thích markdown mô tả từng hành động và lý do thực hiện.
- **Notebook Trực quan hóa Dữ liệu (Data Visualization Notebook):** Chứa mã để tạo ra các biểu đồ trực quan, với các chú thích markdown giải thích lý do lựa chọn biểu đồ và trình bày các thông tin chi tiết rút ra từ mỗi biểu đồ.

Đảm bảo rằng mỗi notebook được tổ chức tốt với:

- Một tiêu đề và phần giới thiệu ngắn gọn ở đầu.
- Các phần hợp lý sử dụng tiêu đề markdown (ví dụ: **Section 1: Loading Data**).
- Các chú thích ngắn gọn trong các ô mã để giải thích chức năng.
- Các kết quả (ví dụ: biểu đồ, thống kê tóm tắt) được hiển thị trực tiếp để xem ngay lập tức.

---

4. Giới hạn (Limitations)

Phòng thí nghiệm này được thiết kế với các giới hạn cụ thể để đảm bảo trải nghiệm học tập tập trung và dễ tiếp cận. Chúng tôi yêu cầu bạn tuân theo các nguyên tắc sau:

Môi trường Lập trình (Programming Environment):

- Các bài tập dự kiến được hoàn thành trong một môi trường Python tiêu chuẩn, chẳng hạn như Jupyter Notebooks hoặc một IDE cơ bản (ví dụ: VS Code, PyCharm).
- Để duy trì sự nhất quán trong trải nghiệm học tập và các bộ kỹ năng cốt lõi, chúng tôi khuyến khích không sử dụng các công cụ trực quan hóa tiên tiến.
- Vui lòng tránh sử dụng các nền tảng nâng cao như Tableau, Power BI hoặc các công cụ giao diện người dùng khác không được yêu cầu trong phòng thí nghiệm này.

Thư viện được phép sử dụng (Permitted Libraries):

- Bạn được phép sử dụng các thư viện trực quan hóa dữ liệu Python đã được dạy trong khóa học này (ví dụ: Matplotlib, Seaborn) bao gồm NumPy (cho các hoạt động số học), Pandas (cho thao tác dữ liệu), và Plotly nếu được phép.
- **Tùy chọn Máy học (Machine Learning Option):**
    - Mặc dù tích hợp một số kỹ thuật máy học đơn giản (ví dụ: phân cụm bằng K-means hoặc hồi quy tuyến tính) có thể làm phong phú thêm quá trình khám phá dữ liệu và trực quan hóa, đây chỉ là một cải tiến tùy chọn chứ không phải là yêu cầu bắt buộc.
    - Mục tiêu chính vẫn là trực quan hóa dữ liệu hiệu quả, và việc sử dụng máy học chỉ nhằm hỗ trợ, không được làm lu mờ mục đích này.
- **Ngôn ngữ trong Báo cáo (Languages in Report):**
    - **Ngôn ngữ sử dụng:** Tiếng Việt.
- **Nội dung do AI tạo ra (AI-Generated Contents):**
    - **Giới hạn:** Không vượt quá 30%.

---

5. Tiêu chí Đánh giá (Evaluation Criteria)

Tiêu chí (Criteria)	Điểm (Mark)
Thu thập dữ liệu và tiền xử lý dữ liệu (Data collection & pre-process data)	5%
Lựa chọn, diễn giải và trực quan hóa các trường dữ liệu và mối quan hệ giữa chúng (Select, interpret, and visualize fields and their hidden relationships)	50%
Rút ra ý nghĩa logic từ mỗi biểu đồ trực quan (Derive logical meaning behind each visualized data)	20%
Cân nhắc nhiều mối quan hệ và nhiều quan điểm khác nhau (Consider many relationships and many different perspectives)	10%

Tiêu chí (Criteria)	Điểm (Mark)
Báo cáo trình bày một cấu trúc và định dạng hợp lý, rõ ràng (The report presents a logical and clear layout and format)	15%
Có phân tích, trực quan hóa với biểu đồ mới và rút ra thông tin hữu ích. Sử dụng các mô hình máy học cơ bản (There is analysis, visualization with novel charts and drawing of useful information. Use basic machine learning models)	5%
Tổng thể việc hiểu về mã nguồn đã nộp (Overall comprehension of the submitted source code)	5%
Tổng cộng (Total)	**110%**

---

6. Yêu cầu Nộp bài (Submission Requirements)

Để đảm bảo một bài nộp đầy đủ và có tổ chức, bạn phải cung cấp các thành phần sau, mỗi thành phần phải tuân theo các hướng dẫn cụ thể:

Docs Folder (Thư mục Tài liệu):

Thư mục này phải chứa báo cáo dự án của bạn, được nộp ở định dạng PDF để đảm bảo khả năng tương thích trên nhiều nền tảng và bảo toàn định dạng (ví dụ: biểu đồ, phông chữ). Báo cáo cần phải được trình bày toàn diện và có cấu trúc rõ ràng, bao gồm các phần sau:

- **Thông tin Nhóm (Group Information):** Liệt kê tên nhóm của bạn và mã số sinh viên của tất cả các thành viên để xác định rõ ràng nhóm của bạn.
- **Đáp ứng Yêu cầu (Requirement Fulfillment):** Đánh giá mức độ mà dự án của bạn đáp ứng từng yêu cầu của nhiệm vụ (ví dụ: thu thập dữ liệu, làm sạch, trực quan hóa). Cung cấp các ví dụ cụ thể hoặc bằng chứng, chẳng hạn như số lượng hàng/cột trong tập dữ liệu hoặc các loại trực quan hóa được tạo ra.
- **Đóng góp của Thành viên (Member Contributions):** Nêu rõ vai trò và mức độ đóng góp của từng thành viên trong nhóm (ví dụ: “Thành viên X xử lý tiền xử lý dữ liệu; Thành viên Y thiết kế các biểu đồ trực quan hóa”). Cần chi tiết và cụ thể để đảm bảo tính minh bạch và công bằng trong đánh giá.
- **Thuật toán và Triển khai (Algorithms and Implementation):** Giải thích các phương pháp kỹ thuật đã sử dụng, chẳng hạn như các kỹ thuật làm sạch dữ liệu (ví dụ: điền khuyết, chuẩn hóa) hoặc các phương pháp trực quan hóa (ví dụ: biểu đồ thanh bằng Matplotlib, bản đồ nhiệt bằng Seaborn). Bao gồm các đoạn mã hoặc chạy các ví dụ minh họa để chứng minh chức năng và hỗ trợ việc ra quyết định.

- **Phong cách Trình bày (Presentation Style):** Đảm bảo báo cáo của bạn rõ ràng, súc tích và trực quan. Sử dụng các biểu đồ, bảng hoặc sơ đồ có độ phân giải cao để hỗ trợ bài thuyết trình của bạn, và duy trì một định dạng hợp lý với cách trình bày thống nhất (ví dụ: tiêu đề, dấu đầu dòng).

Sources Folder (Thư mục Mã nguồn):

Thư mục này phải chứa toàn bộ mã nguồn được phát triển cho dự án, chủ yếu dưới dạng các notebook Jupyter (.ipynb) và các tập lệnh Python (.py). Để đảm bảo tính bền vững:

- **Tổ chức các Notebook theo nhiệm vụ (Organize notebooks by task):** Ví dụ, **data_collection.ipynb**, **data_cleaning.ipynb**, **data_visualization.ipynb** với tên tệp mô tả rõ ràng.
- **Bao gồm các ô Markdown hoặc các chú thích giải thích rõ ràng logic và mục đích của mã.**
- **Nếu bạn sử dụng các ngôn ngữ khác ngoài Python (ví dụ: R), cung cấp tệp README riêng với các hướng dẫn để tái tạo kết quả và đảm bảo môi trường thực thi.**

Datasets Folder (Thư mục Dữ liệu):

Thư mục này phải bao gồm tập dữ liệu (các tệp .csv) được sử dụng trong dự án, cùng với các tham chiếu rõ ràng đến nguồn dữ liệu của bạn.

- Đối với các tập dữ liệu dưới 100 MB, bao gồm các tệp trực tiếp (ví dụ: .csv, .xlsx). Đối với các tập dữ liệu lớn hơn, cung cấp một liên kết ổn định, có thể truy cập tới một nguồn bên ngoài (ví dụ: URL Kaggle) hoặc lưu trữ đám mây (ví dụ: Google Drive, OneDrive), đảm bảo quyền truy cập được thiết lập để bất kỳ ai có liên kết đều có thể xem được.
- Bao gồm một tệp README ngắn trong thư mục, mô tả tên tập dữ liệu, nguồn, kích thước và định dạng, kèm theo hướng dẫn truy cập nếu được lưu trữ bên ngoài.