**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**---****---**



**BÁO CÁO BÀI TẬP NHÓM 1**

**MÔN HỌC: TRỰC QUAN HÓA DỮ LIỆU**

**CHỦ ĐỀ: TRỰC QUAN HÓA VỚI PYTHON**

**Giảng viên hướng dẫn: Lê Nhựt Nam**

**Thành phố Hồ Chí Minh, tháng 10 năm 2024**

**MỤC LỤC**

[Thông tin nhóm 3](#_Toc181308177)

[I. Tổng quan 4](#_Toc181308178)

[1. Khái niệm trực quan hóa dữ liệu 4](#_Toc181308179)

[2. Tầm quan trọng của trực quan hóa dữ liệu 4](#_Toc181308180)

[3. Các thư viện hỗ trợ trực quan hóa của python 4](#_Toc181308181)

[4. Mục tiêu của bài lab 4](#_Toc181308182)

[II. Đánh giá mức độ hoàn thành 5](#_Toc181308183)

[III. Bảng phân công công việc 5](#_Toc181308184)

[IV. Nội dung 6](#_Toc181308185)

[1. Thu thập dữ liệu 6](#_Toc181308186)

[2. Khám phá và tiền xử lý dữ liệu 6](#_Toc181308187)

[3. Trực quan hóa dữ liệu 7](#_Toc181308188)

[Tài liệu tham khảo 9](#_Toc181308189)

Thông tin nhóm

|  |  |  |
| --- | --- | --- |
| **STT** | **MSSV** | **Họ và tên** |
| 1 | 22120339 | Nguyễn Thị Anh Thi |
| 2 | 22120389 | Dương Ngọc Kiều Trinh |
| 3 | 22120421 | Nguyễn Đoàn Minh Uyên |
| 4 | 22120422 | Nguyễn Phạm Tú Uyên |

1. Tổng quan
2. Khái niệm trực quan hóa dữ liệu

**Trực quan hóa dữ liệu (Data Visualization)** là quá trình biểu diễn dữ liệu dưới dạng hình ảnh như biểu đồ, đồ thị, bản đồ, và các hình thức trực quan khác. Mục tiêu của trực quan hóa là làm cho thông tin phức tạp trở nên dễ hiểu, dễ tiếp cận, và dễ phân tích.

1. Tầm quan trọng của trực quan hóa dữ liệu

Trực quan hóa dữ liệu đóng vai trò quan trọng trong nhiều lĩnh vực, từ kinh doanh, khoa học, giáo dục, đến y tế và các ngành công nghệ thông tin. Một số ứng dụng quan trọng của trực quan hóa:

**Tăng tính thuyết phục:** Khi dữ liệu được hiển thị một cách trực quan, nó không chỉ dễ hiểu mà còn có thể tăng tính thuyết phục cho các ý tưởng, giúp các nhà phân tích, nhà nghiên cứu, hoặc nhà quản lý thuyết phục người khác tốt hơn.

**Phát hiện bất thường:** Trực quan hóa bằng biểu đồ giúp các nhà nghiên cứu dễ dàng phát hiện các xu hướng tăng/giảm, các mẫu lặp lại, hoặc nhận diện các điểm bất thường trong dữ liệu.

**Hỗ trợ phân tích và dự đoán:** Trực quan hóa dữ liệu cung cấp cho nhà nghiên cứu thông tin cần thiết để đưa ra quyết định nhanh chóng và chính xác.

**Hỗ tợ ra quyết định:** Trực quan hóa dữ liệu cung cấp cho nhà quản lý thông tin cần thiết để đưa ra quyết định nhanh chóng và chính xác.

1. Các thư viện hỗ trợ trực quan hóa của python

Python là một ngôn ngữ phổ biến trong phân tích và trực quan hóa dữ liệu, nhờ vào thư viện phong phú và khả năng tích hợp tốt. Trong phạm vi bài lab, nhóm sử dụng các thư viện sau:

* Thư viện pandas: **Đọc file CSV, tạo DataFrame và quản lý kiểu dữ liệu Categorical**
* Thư viện **matplotlib** và seaborn: **hỗ trợ vẽ biểu đồ, trực quan hóa dữ liệu**

1. Mục tiêu của bài lab

Bài lab hướng tới 2 mục tiêu chính:

* Nắm bắt các thư viện phổ biến trong Python như Matplotlib, Seaborn,… và cách sử dụng chúng trong việc tạo biểu đồ và đồ thị. Hiểu được khi nào nên dùng thư viện nào, và làm thế nào để tùy chỉnh các biểu đồ theo nhu cầu.
* Phân tích các mối quan hệ giữa các biến, tìm ra các mối quan hệ tiềm ẩn trong dữ liệu hay nói cách khác là ý nghĩa đằng sau dữ liệu.

1. Đánh giá mức độ hoàn thành

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Công việc** | **Chi tiết** | **Mức độ hoàn thành** |
| **1** | Thu thập dữ liệu | Trả lời câu hỏi | **100%** |
| **2** | Khám phá dữ liệu | Thông tin dữ liệu | **100%** |
| Tiền xử lý | **100%** |
| **3** | Tìm mối quan hệ ẩn trong dữ liệu | Đặt câu hỏi | **100%** |
| Trực quan hóa dữ liệu | **100%** |
| **4** | Báo cáo |  | **100%** |

1. Bảng phân công công việc

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Họ và tên** | **Phân công** | **Mức độ hoàn thiện** |
| 1 | Nguyễn Thị Anh Thi | * Thu thập dữ liêu * Khám phá dữ liệu * Đặt câu hỏi và trực quan * Viết báo cáo | 100% |
| 2 | Dương Ngọc Kiều Trinh | * Tiền xử lý dữ liệu * Khám phá dữ liệu * Đặt câu hỏi và trực quan * Viết báo cáo | 100% |
| 3 | Nguyễn Đoàn Minh Uyên | * Tiền xử lý dữ liệu * Khám phá dữ liệu * Đặt câu hỏi và trực quan * Viết báo cáo | 100% |
| 4 | Nguyễn Phạm Tú Uyên | * Thu thập dữ liệu * Khám phá dữ liệu * Đặt câu hỏi và trực quan * Viết báo cáo | 100% |

1. Nội dung
2. Thu thập dữ liệu

Lý do chọn chủ đề: hiểu rõ các yếu tố có thể gây ảnh hưởng đến học sinh giúp nhà trường và giáo viên điều chỉnh phương pháp giảng dạy và hỗ trợ học sinh hiệu quả hơn. Hơn nữa, các yếu tố khác có thể mang lại nhiều thông tin giá trị cho việc cải thiện chất lượng giáo dục. Bản thân nhóm cũng rút ra được bài học kinh nghiệm, thấy những vấn đề ảnh hưởng đến việc học tập của mình và biết cách khắc phục tốt hơn.

Chủ đề chính của dữ liệu: Các yếu tố ảnh hưởng đến kết quả học tập của học sinh

Xây dựng dữ liệu: Dữ liệu được xây dựng dựa trên thông tin từ các nguồn giáo dục là trường học và học sinh.

Phương pháp: khảo sát và thu thập thông tin. Khảo sát để thu thập dữ liệu về hoàn cảnh cá nhân, gia đình, thái độ học tập của học sinh. Sau đó các yếu tố này được liên kết với việc thu thập thông tin từ trường học là kết quả học tập,…

Sử dụng tập dữ liệu: Tập dữ liệu này có thể được sử dụng để phân tích thống kê, xây dựng mô hình dự đoán giúp hiểu rõ mối quan hệ giữa các yếu tố với kết quả học tập, từ đó đưa ra khuyến nghị cải thiện cho học sinh hoặc các chương trình giáo dục.

Tính hợp pháp: Nếu dữ liệu không chứa thông tin nhạy cảm hoặc đã được ẩn danh thì việc sử dụng nó cho mục đích giáo dục và nghiên cứu là hợp pháp miễn là tuân thủ đúng các quy định về bảo mật, quyền riêng tư và không tiết lộ thông tin cá nhân của học sinh.

1. Khám phá và tiền xử lý dữ liệu

Trong phần này, chúng em thực hiện kiểm tra và trực quan hóa sự phân bố dữ liệu của các cột của dataset bằng các hàm có sẵn của thư viện pandas và matplotlib.

**Kích thước dữ liệu:** Dữ liệu gốc có 6607 dòng và 20 cột.

**Mô tả dữ liệu:** Dữ liệu là các yếu tố cần khảo sát xem mức độ ảnh hưởng kết quả học tập.

**Phân loại dữ liệu kiểu định lượng và định tính:**

* Các cột có thông tin kiểu định lượng: Hours\_Studied, Attendance, Sleep\_Hours, Previous\_ Scores, Tutoring\_Sessions, Physical\_Activity và xam\_Score.
* Các cột có kiểu dữ liệu định tính: Parental\_Involvement, Access\_to\_Resources, Extracurricular\_Activities, Motivation\_Level, Internet\_Access, Family\_Income, Teacher\_Quality, School\_Type, Peer\_Influence, Learning\_Disabilities, Parental\_Education\_Level, Distance\_from\_Home và Gender.

**Kiểm tra dữ liệu trùng lắp:** dữ liệu không có dòng nào bị trùng lắp.

**Kiểm tra dữ liệu bị thiếu:** Qua kiểm tra, các cột Teacher\_Quality, Parental\_Education\_Level và Distance\_from\_Home bị thiếu với tỉ lệ xấp xỉ 1%. Vì tỉ lệ này khá nhỏ nên nhóm tiến hành xóa các hàng chứa giá trị bị thiếu đó.

**Phân bố giá trị của các cột:**

* Với các cột có kiểu định lượng: Trong các cột có kiểu định lượng, Hours\_Studied, Attendance, Previous\_Scores và Exam\_Score có phân bố liên tục và 3 cột còn lại có phân bố rời rạc nên nhóm chọn biều đồ histogram cho các phân bố liên tục và bar chart cho các phân bố rời rạc.
* Với các cột có kiểu định lượng: Nhìn chung, các cột có kiểu định lượng đều có khá ít giá trị nên nhóm sử dụng pie chart để thể hiện phân bố giá trị của các cột này.

**Lưu dữ liệu:**

Lưu dữ liệu đã được tiền xử lý vào file *StudentPerformanceFactors\_clean.csv*

1. Trực quan hóa dữ liệu

Trong phần này, chúng em thực hiện trực quan hóa dữ liệu bằng thư viện matplotlib và seaborn. Mục đích là muốn biết được sự ảnh hưởng của các yếu tố khác như động lực học, gia đình, nhà trường, hoạt động ngoại khóa, lên thành tích học tập. Sau đây là lý do chọn biểu đồ để trực quan hóa cho từng câu hỏi.

Câu 1: Ảnh hưởng của hoạt động ngoại khóa lên thành tích học tập của học sinh.

Câu 2: Ảnh hưởng từ các nhân tố trong môi trường học tập lên động lực học của học sinh.

Câu 3: Mối tương quan giữa thời gian học và kết quả thi của học sinh.

Câu 4: Ảnh hưởng từ mức thu nhập gia đình đến điểm thi của học sinh.

Câu 5: Khả năng tiếp cận tới nguồn lực và chất lượng giáo viên ảnh hưởng thế nào đến việc học của học sinh.

Các biểu đồ được chọn cho phép thể hiện rõ ràng mối quan hệ giữa từng yếu tố (khả năng tiếp cận tài nguyên, chất lượng giáo viên) với kết quả học tập, cũng như tác động kết hợp của hai yếu tố này.

+ Biểu đồ biểu đồ phân tán cho khả năng tiếp cận tài nguyên và kết quả học tập:

* Lý do chọn biểu đồ: Scatter plot là biểu đồ phù hợp để thể hiện mối quan hệ giữa hai biến liên tục hoặc xếp hạng. Trong trường hợp này, biểu đồ scatter giúp minh họa sự phân bố kết quả học tập dựa trên mức độ tiếp cận tài nguyên (Low, Medium, High). Mỗi điểm trên biểu đồ đại diện cho một học sinh và vị trí của điểm đó biểu thị mối quan hệ giữa khả năng tiếp cận tài nguyên và kết quả học tập.
* Tính phù hợp: Scatter plot cung cấp cái nhìn tổng quan rõ ràng về xu hướng hoặc sự phân tán của dữ liệu, giúp dễ dàng quan sát mối quan hệ giữa khả năng tiếp cận tài nguyên và kết quả học tập. Nếu kết quả học tập tăng khi khả năng tiếp cận tài nguyên cao hơn, điều này sẽ được thể hiện bằng sự tập trung của các điểm ở mức độ cao hơn.

+ Box plot (biểu đồ hộp) cho chất lượng giáo viên và kết quả học tập:

* Lý do chọn biểu đồ: Box plot là biểu đồ lý tưởng để so sánh phân phối dữ liệu giữa các nhóm khác nhau. Nó giúp thể hiện phạm vi dữ liệu, giá trị trung vị, cũng như sự phân tán của kết quả học tập dựa trên chất lượng giáo viên.
* Tính phù hợp: Box plot thể hiện rõ sự khác biệt về điểm số giữa các nhóm chất lượng giáo viên (Low, Medium, High). Đặc biệt, biểu đồ này dễ dàng cho thấy các giá trị ngoại lệ, nếu có, giúp phân tích chính xác hơn sự ảnh hưởng của chất lượng giáo viên đến kết quả học tập của học sinh.

+ Line plot (biểu đồ đường) cho kết hợp giữa khả năng tiếp cận tài nguyên và chất lượng giáo viên:

* Lý do chọn biểu đồ: Line plot phù hợp để hiển thị xu hướng khi các biến có mối quan hệ theo thứ tự. Trong trường hợp này, biểu đồ đường giúp thể hiện sự thay đổi điểm số trung bình khi khả năng tiếp cận tài nguyên và chất lượng giáo viên thay đổi.
* Tính phù hợp: Biểu đồ đường dễ đọc và làm nổi bật xu hướng thay đổi điểm số khi cả hai yếu tố (khả năng tiếp cận tài nguyên và chất lượng giáo viên) được cải thiện. Điều này giúp nhấn mạnh sự kết hợp tác động của hai yếu tố này đến kết quả học tập của học sinh.

Câu 6: Mức độ động lực học tập có ảnh hưởng đến sự tiến bộ của học sinh không?

Các biểu đồ được chọn cung cấp một cái nhìn rõ ràng về cách mức độ động lực học tập ảnh hưởng đến điểm số và sự tiến bộ của học sinh. Heatmap cho phép phân tích sâu về sự khác biệt điểm số giữa các nhóm, trong khi line plot giúp nhận diện xu hướng cải thiện điểm số cuối kỳ khi động lực học tập tăng lên.

+ Heatmap (biểu đồ nhiệt) cho điểm số trước đây và điểm thi theo mức độ động lực học tập:

* Lý do chọn biểu đồ: Heatmap rất phù hợp để hiển thị sự phân bố dữ liệu giữa các nhóm, đặc biệt khi so sánh các giá trị trung bình của hai hoặc nhiều thuộc tính trong các nhóm phân loại (như mức độ động lực học tập).
* Tính phù hợp: Heatmap giúp thể hiện một cách rõ ràng mức độ khác biệt về điểm số giữa các nhóm động lực học tập. Màu sắc khác nhau thể hiện các mức độ khác nhau của điểm số, giúp dễ dàng so sánh và nhận diện sự chênh lệch giữa các nhóm.

+ Line plot (biểu đồ đường) cho mối quan hệ giữa động lực học tập và điểm số cuối kỳ:

* Lý do chọn biểu đồ: Line plot thích hợp để biểu diễn xu hướng hoặc mối quan hệ giữa hai biến khi có một sự sắp xếp thứ tự nhất định, trong trường hợp này là mức độ động lực học tập từ thấp đến cao.
* Tính phù hợp: Biểu đồ đường không chỉ thể hiện mối quan hệ giữa mức độ động lực và điểm số cuối kỳ mà còn cung cấp thông tin về sự tiến bộ khi mức độ động lực tăng lên. Vùng bóng xung quanh đường biểu diễn (confidence interval) giúp minh họa độ tin cậy của dữ liệu, cung cấp thông tin trực quan về sự biến động của điểm số.

**Câu 7: Liệu có mối tương quan nào giữa giới tính và thành tích học tập?**

**Câu 8: Giấc ngủ và các hoạt động thể chất ảnh hưởng như thế nào đến thành tích học tập?**

Tài liệu tham khảo