**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KỲ**

**MÔN HỌC: LẬP TRÌNH PYTHON**

**PHÂN TÍCH MỐI QUAN HỆ GIỮA TRÌNH ĐỘ HỌC VẤN, LOẠI NGÀNH NGHỀ, KINH NGHIỆM VÀ MỨC LƯƠNG CỦA NAM THANH NIÊN**

**Mã lớp học phần: IPPA233277\_04**

**Học kỳ 1 – Năm học 2024-2025**

**Giảng viên hướng dẫn: ThS. Trần Quang Khải**

**Danh sách sinh viên thực hiện:**

|  |  |
| --- | --- |
| **MSSV** | **Họ tên** |
| 23110334 | Trần Hữu Thoại |
| 23110212 | Nguyễn Trung Hậu |
| 23110247 | Nguyễn Hoàng Anh Kiệt |
| 23110354 | Đoàn Quân Tuấn |
| 23110351 | Trần Thành Trung |

***Thành phố Hồ Chí Minh, tháng 11 năm 2024***

*Nhận xét của giảng viên*

*TP. Hồ Chí Minh, ngày … tháng… năm 2024*

*Giảng viên ký tên*

**MỤC LỤC**

[**PHẦN 1: MỞ ĐẦU** 1](#_Toc14341)

[1. Lý do chọn đề tài 1](#_Toc675)

[2. Mục tiêu đề tài 1](#_Toc12201)

[3. Đối tượng và phạm vi nghiên cứu 2](#_Toc2476)

[**PHẦN 2. NỘI DUNG** 3](#_Toc337)

[**CHƯƠNG 1: TÓM TẮT CƠ SỞ LÝ THUYẾT** 3](#_Toc31946)

[**CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU** 4](#_Toc31946)

2.1. Nguồn gốc của dữ liệu........................................................................4

2.2. Cấu trúc của dữ liệu............................................................................4

2.3. Mục đích của dữ liệu..........................................................................6

2.3. Hạn chế, khiếm khuyết của dữ liệu....................................................6

2.4. Thống kê mô tả dữ liệu……...............................................................7

[**CHƯƠNG 3: LÀM SẠCH VÀ CHUẨN HÓA DỮ LIỆU** 9](#_Toc16207)

3.1. Phân tích những vấn đề gặp phải với dữ liệu......................................9

3.2. Thực hiện ý tưởng giải quyết vấn đề................................................10

3.3. Xây dựng thêm thuộc tính…………................................................12

[**CHƯƠNG 4: XÂY DỰNG CƠ CHẾ CRUD TRÊN TẬP DỮ LIỆU** .........14](#_Toc25002)

4.1. Create.............................................. ..........................................14

4.2. Read.............. ...................................................................................14

4.3. Update...............................................................................................21

4.4. Delete................................................................................................24

[**CHƯƠNG 5: VẼ CÁC LOẠI BIỂU ĐỒ BIỂU DIỄN DỮ LIỆU**  26](#_Toc25002)

5.1. Khái quát trực quan hóa dữ liệu…………........................................26

5.2. Phân tích và lựa chọn loại biểu đồ phù hợp......................................26

[**CHƯƠNG 6: XÂY DỰNG GIAO DIỆN CHƯƠNG TRÌNH**.....................32](#_Toc25002)

6.1. Giới thiệu về Tkinter…………………….........................................32

6.2. Mục đích sử dụng Tkinter trong phân tích dữ liệu.................................32

6.3. Các thành phần cơ bản của Tkinter………......................................32

6.4. Xây dựng giao diện tương tác phân tích dữ liệu sử dụng Tkinter............32

[**PHẦN KẾT LUẬN** 35](#_Toc434)

[**PHẦN TÀI LIỆU THAM KHẢO** 36](#_Toc16131)

**DANH MỤC HÌNH ẢNH**

[Hình 2.1: Các số liệu được thống kê dữ liệu 8](#_Toc183467095)

[Hình 3.1: Giá trị tệp dữ liệu sau khi tách dữ liệu 9](#_Toc183467096)

[Hình 3.2: Giá trị tệp dữ liệu trước khi tách dữ liệu 9](#_Toc183467097)

[Hình 3.3: Kiểm tra và tổng hợp số lượng giá trị bị thiếu (NaN) 9](#_Toc183467098)

[Hình 3.4: Code tách dữ liệu trong các hàng chưa hợp lệ 10](#_Toc183467099)

[Hình 3.5: Code điền giá trị cho các ô bị thiếu giá trị 11](#_Toc183467100)

[Hình 3.6: Code chuẩn hóa dữ liệu phân loại 11](#_Toc183467101)

[Hình 3.7: Code thay đổi dấu gạch dưới thành khoảng trắng 12](#_Toc183467102)

[Hình 4.1: Cửa sổ nhập dữ liệu cho bảng ghi mới 15](#_Toc183467103)

[Hình 4.2: Cửa sổ lựa chọn thao tác READ 16](#_Toc183467104)

[Hình 4.3: Cửa sổ chọn bản ghi 16](#_Toc183467105)

[Hình 4.4: Cửa sổ hiển thi thông tin bản ghi 17](#_Toc183467106)

[Hình 4.5: Kết quả hiển thị phân trang dữ liệu 18](#_Toc183467107)

[Hình 4.6: Cửa số thực hiện lựa chọn phương pháp lọc 19](#_Toc183467108)

[Hình 4.7: Cửa sổ thực hiện chức năng theo điều kiện cột 19](#_Toc183467109)

[Hình 4.8: Kết quả hiển thị lọc dữ liệu theo điều kiện của cột 19](#_Toc183467110)

[Hình 4.9: Cửa sổ thực hiện chức năng lọc các cột xuất hiện 20](#_Toc183467111)

[Hình 4.10: Kết quả hiển thị sau khi lọc tên cột 20](#_Toc183467112)

[Hình 4.11: Cửa sổ hiển thị yêu cầu hàng để lọc 21](#_Toc183467113)

[Hình 4.12: Kết quả hiển thị các hàng được lọc 21](#_Toc183467114)

[Hình 4.13: Cửa sổ thực hiện thao tác sắp xếp 21](#_Toc183467115)

[Hình 4.14: Hiển thị kết quả sắp xếp 22](#_Toc183467116)

[Hình 4.15: Cửa sổ lựa chọn thao tác UPDATE 22](#_Toc183467117)

[Hình 4.19: Hướng dẫn thực hiện xóa một bản ghi 26](#_Toc183467118)

[Hình 5.2: Biểu đồ thể hiện sự liên hệ giữa “wage” và “industry” 28](#_Toc183467119)

[Hình 5.3: Biểu đồ thống kê logarit của lương theo giờ 29](#_Toc183467120)

[Hình 5.4: Biểu đồ thống kê số người ở các dân tộc khác nhau 29](#_Toc183467121)

[Hình 5.5: Biểu đồ thống kê trình trạng hôn nhân 30](#_Toc183467122)

[Hình 5.6: Biểu đồ thống kê tình trạng sức khỏe 30](#_Toc183467123)

[Hình 5.7: Biểu đồ thống kê số lượng trong các lĩnh vực lao động khác nhau 31](#_Toc183467124)

[Hình 5.8: Biểu đồ thống kê số lượng lao động trong các nhóm nghề khác nhau 31](#_Toc183467125)

[Hình 5.9: Biểu đồ thống kê số lượng lao động ở các khu vực cư trú khác nhau 32](#_Toc183467126)

[Hình 6.1: Code xây dựng giao diện 34](#_Toc183467127)

[Hình 6.2: Giao diện người dùng 35](#_Toc183467128)

**DANH MỤC BẢNG**

B[ảng 2.1 . Bảng mô tả các cột của tập Males.csv 5](#_Toc150)

**KẾ HOẠCH PHÂN CÔNG NHIỆM VỤ THỰC HIỆN ĐỀ TÀI**

**CUỐI KỲ MÔN LẬP TRÌNH PYTHON**

**HỌC KỲ I NĂM HỌC 2024-2025**

**1. Mã lớp môn học: IPPA233277\_04**

**2. Giảng viên hướng dẫn: ThS. Trần Quang Khải**

**3. Tên đề tài: PHÂN TÍCH MỐI QUAN HỆ GIỮA TRÌNH ĐỘ HỌC VẤN, LOẠI NGÀNH NGHỀ, KINH NGHIỆM VÀ MỨC LƯƠNG CỦA NAM THANH NIÊN**

**4. Bảng phân công nhiệm vụ:**

|  |  |
| --- | --- |
| **Sinh viên thực hiện** | **Nội dung thực hiện** |
| Trần Hữu Thoại | Tìm hiểu Tkinter, code giao diện tương tác |
| Đoàn Quân Tuấn | Code thao tác READ và UPDATE dữ liệu |
| Nguyễn Trung Hậu | Mô tả dữ liệu, làm sạch và chuẩn hóa dữ liệu |
| Nguyễn Hoàng Anh Kiệt | Code thao tác CREATE và DELETE dữ liệu |
| Trần Thành Trung | Trực quan hóa dữ liệu bằng biểu đồ |

**LỜI CẢM ƠN**

Lời đầu tiên nhóm em xin gửi lời cảm ơn chân thành đến thầy Trần Quang Khải - giảng viên dạy học và hướng dẫn bộ môn Lập Trình Python của lớp IPPA233277\_04 vì đã hết mình hướng dẫn và giảng dạy chúng em nhưng kiến thức hay và bổ ích về bộ môn Lập Trình Python. Đồng thời tạo điều kiện cho chúng em có thể thực hiện đồ án thú vị này.

Vì thời gian thực hiện đồ án có hạn, nhưng nhóm em đã nỗ hết sức để hoàn thành đồ án đúng theo những yêu cầu mà thầy đề ra.Trong quá trình thực hiện, không tránh khỏi việc xuất hiện những sai sót chủ quan, nhóm em mong thầy có thể đưa ra góp ý cho kết qua làm việc của nhóm để đồ án của nhóm em được hoàn thiện nhất có thể.

Một lần nữa, chúng em xin trân trọng cảm ơn sự quan tâm giúp đỡ của thầy trong suốt một kì vừa qua. Hi vọng trong tương lai chúng em lại có dịp gặp lại thầy trong các môn học khác.

# PHẦN 1: MỞ ĐẦU

## Lý do chọn đề tài

Trong môi trường kinh doanh ngày nay, nguồn nhân lực được coi là một trong những yếu tố quyết định thành công của một doanh nghiệp. Không chỉ đơn thuần là những con số trong bảng lương, nguồn nhân lực đóng vai trò trọng yếu trong việc thúc đẩy sự phát triển, tạo ra lợi thế cạnh tranh và xác định hướng đi của tổ chức.

Bằng cách thu thập và phân tích các thông tin liên quan như tuổi tác, giới tính, trình độ học vấn, kinh nghiệm làm việc, và các yếu tố môi trường làm việc, các nhà nghiên cứu có thể phát hiện các xu hướng trong lực lượng lao động, cũng như các vấn đề cần giải quyết. Chẳng hạn, phân tích dữ liệu có thể giúp xác định những nhóm công nhân có mức thu nhập thấp, điều kiện làm việc kém, hoặc những yếu tố tác động tiêu cực đến năng suất lao động. Đồng thời, thông qua việc phân tích, các chính sách hỗ trợ người lao động có thể được xây dựng để cải thiện chất lượng công việc và nâng cao đời sống của họ.

Vậy, điều cần xem xét là làm thế nào để xây dựng những mô hình dự đoán đáng tin cậy. Để giải đáp vấn đề này, nhóm của chúng tôi đã quyết định thực hiện đề tài này dựa trên kiến thức mà chúng tôi đã học được từ môn Lập Trình Python, nhằm biểu diễn tập dữ liệu một cách tổng quát và dễ hiểu nhất.

## Mục tiêu đề tài

Mục tiêu cho đề tài “PHÂN TÍCH MỐI QUAN HỆ GIỮA TRÌNH ĐỘ HỌC VẤN, LOẠI NGÀNH NGHỀ, KINH NGHIỆM VÀ MỨC LƯƠNG CỦA NAM THANH NIÊN” là đưa ra những phân tích trên tập dữ liệu nguồn về các yếu tố ảnh hưởng đến tiền lương, nghiên cứu tình trạng sức khỏe ảnh hưởng tới việc làm,…Để từ đó giúp hiểu rõ hơn về các đặc điểm của nguồn lao động hiện tại mà còn cung cấp cơ sở để dự báo những thay đổi trong tương lai, từ đó hỗ trợ việc xây dựng các chiến lược phát triển nguồn nhân lực bền vững và hiệu quả.

## Đối tượng và phạm vi nghiên cứu

Nghiên cứu tập trung vào dữ liệu từ một nhóm gồm 4.360 quan sát được thu thập trong giai đoạn từ năm 1980 đến năm 1987. Dữ liệu bao gồm thông tin chi tiết về loại công việc, khu vực sinh sống, mức lương theo giờ của người lao động, cùng các yếu tố khác có liên quan. Phạm vi nghiên cứu bao quát các yếu tố kinh tế - xã hội trong giai đoạn cụ thể, nhằm phân tích mối quan hệ giữa đặc điểm công việc và thu nhập của người lao động trong thời kỳ này.

# PHẦN 2. NỘI DUNG

## CHƯƠNG 1: TÓM TẮT CƠ SỞ LÝ THUYẾT

Trong quá trình thực hiện đồ án phân tích dữ liệu, người thực hiện báo cáo cần sử dụng một số thư viện phổ biến trong Python để hỗ trợ xử lý, phân tích, trực quan hóa dữ liệu và xây dựng giao diện người dùng. Dưới đây là giới thiệu về các thư viện chính:

Numpy (Numerical Python)

 Xử lý dữ liệu dạng mảng một chiều và hai chiều.

 Thực hiện các phép tính thống kê cơ bản như trung bình, phương sai, độ lệch chuẩn.

 Hỗ trợ chuẩn hóa dữ liệu trước khi phân tích.

Pandas

 Đọc dữ liệu từ các tệp CSV, Excel, hoặc cơ sở dữ liệu.

 Tiền xử lý dữ liệu: xóa dữ liệu bị thiếu, thay thế giá trị, lọc dữ liệu theo điều kiện.

 Tính toán và tóm tắt dữ liệu, tạo bảng thống kê.

Matplotlib

 Trực quan hóa các xu hướng và mối quan hệ trong dữ liệu.

 So sánh dữ liệu qua các biểu đồ dạng thanh (bar), biểu đồ phân tán (scatter), hoặc histogram.

 Tích hợp biểu đồ vào các ứng dụng GUI.

Seaborn

* Là một thư viện xây dựng trên Matplotlib, cung cấp các biểu đồ nâng cao và dễ sử dụng hơn như heatmap, pairplot, và boxplot

 Trực quan hóa dữ liệu một cách đẹp mắt và nhanh chóng.

 Tạo biểu đồ thống kê (boxplot, violin plot) hoặc các mối quan hệ (scatterplot, lineplot).

Tkinter

* Hỗ trợ tạo các ứng dụng với giao diện thân thiện, cho phép người dùng tương tác với dữ liệu.

 Tích hợp các widget như Button, Entry, Canvas để dễ dàng thao tác với dữ liệu.

## CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU

* 1. **Nguồn gốc của dữ liệu**
  + *Lịch sử và nguồn gốc:* Bộ dữ liệu được lấy từ trang Kaggle.com bao gồm thông tin về loại công việc của người lao động, khu vực cư trú, tiền lương theo giờ, học vấn, kinh nghiệm làm việc,... được thu thập từ một cuộc khảo sát kéo dài từ năm 1980 đến năm 1987. Số lượng cuộc phỏng vấn mỗi năm là không đổi trong giai đoạn này, với 545 người nam tham gia mỗi năm.
  + *Công dụng:* Bộ dữ liệu dùng để phân tích các yếu tố ảnh hưởng đến tiền lương, nghiên cứu tình trạng sức khỏe liên quan đến việc làm, khảo sát mức độ ảnh hưởng của học vấn, kinh nghiệm và các đặc trưng cá nhân khác đến thu nhập và chất lượng cuộc sống.
  + *Lĩnh vực áp dụng****:*** Bộ dữ liệu này thường được dùng trong các lĩnh vực kinh tế lao động, xã hội học, y tế công cộng, hoặc nghiên cứu về bình đẳng thu nhập và cơ hội.
  1. **Cấu trúc của dữ liệu**

*Danh sách các đặc trưng:*

**Bảng 2.1:** Bảng mô tả các cột của tập Males.csv

|  |  |  |
| --- | --- | --- |
| Đặc trưng (cột) | Mô tả | Kiểu dữ liệu |
| rownames | Tên hàng được đánh theo số thứ tự từ nhỏ tới lớn | Int64 |
| nr | Mã số định danh duy nhất cho mỗi cá nhân được khảo sát, đảm bảo tính riêng biệt và tránh trùng lặp khi phân tích. | Int64 |
| year | Năm tiến hành cuộc khảo sát, nằm trong khoảng từ 1980 đến 1987. | Int64 |
| school | Số năm học tập của cá nhân, biểu thị mức độ học vấn hoặc trình độ giáo dục. | Int64 |
| exper | Số năm kinh nghiệm làm việc, được tính bằng công thức: *exper = age - 6 - school* (tuổi trừ đi 6 năm đầu đời và số năm học). Đặc trưng này phản ánh mức độ tích lũy kinh nghiệm làm việc. | Int64 |
| union | Tham gia công đoàn và được thương lượng lương theo nhóm hay không, có 2 trạng thái: ‘yes’ hoặc ‘no’. | Object |
| ethn | Dân tộc hoặc sắc tộc của cá nhân, gồm ba nhóm chính: black (người da đen), hisp (người gốc Tây Ban Nha), và other (các nhóm khác). | Object |
| married | Tình trạng hôn nhân, có 2 trạng thái: ‘yes’ hoặc ‘no’. | Object |
| health | Tình trạng sức khỏe của cá nhân, cho biết có vấn đề sức khỏe hay không, có 2 trạng thái: ‘yes’ hoặc ‘no’. | Object |
| wage | Logarit của mức lương theo giờ, là biến mục tiêu (target variable). Việc sử dụng log lương giúp giảm thiểu ảnh hưởng của các giá trị ngoại biên và làm dữ liệu phân phối đều hơn. | Float64 |
| industry | Ngành nghề hoặc lĩnh vực làm việc, là một biến phân loại (factor) gồm các ngành như Business\_and\_Repair\_Service, Personal\_Service, v.v | Object |
| occupation | Nghề nghiệp hoặc chức danh, là một biến phân loại như Service\_Workers, Craftsmen, Foremen\_and\_kindred,… | Object |
| residence | Nơi cư trú bao gồm: rural area, north east, northern central và south. | Object |

* 1. **Mục đích của dữ liệu**

*- Mục đích của việc thu thập dữ liệu:*Việc giữ nguyên số lượng khảo sát này qua các năm giúp bộ dữ liệu ổn định về cỡ mẫu và phản ánh liên tục các yếu tố lao động, kinh tế và nhân khẩu học trong suốt giai đoạn khảo sát. Điều này có thể mang lại độ tin cậy cao khi phân tích xu hướng theo thời gian, đồng thời cũng giúp xác định những thay đổi về kinh tế và xã hội qua từng năm.

+ Phân tích tiền lương: Tìm hiểu các yếu tố ảnh hưởng đến tiền lương, như học vấn, kinh nghiệm, tình trạng công đoàn, v.v.

+ Nghiên cứu sức khỏe và công việc: Đánh giá mối quan hệ giữa tình trạng sức khỏe và các điều kiện làm việc, như nghề nghiệp, ngành nghề, hoặc điều kiện lao động.

+ Đánh giá cơ hội nghề nghiệp: Nghiên cứu sự bình đẳng trong cơ hội việc làm và thu nhập theo sắc tộc, giới tính, và các yếu tố nhân khẩu học khác.

+ Hỗ trợ chính sách lao động: Thông tin này hữu ích để xây dựng chính sách về thu nhập, phúc lợi xã hội và bảo vệ người lao động.

* 1. **Hạn chế, khiếm khuyết của dữ liệu**
  + Hạn chế về thời gian:

+ Cũ và không cập nhật: Dữ liệu chỉ bao gồm giai đoạn từ năm 1980 đến 1987, nên không phản ánh được các thay đổi hiện tại về điều kiện lao động, mức lương, hoặc xu hướng kinh tế.

+ Không liên tục: Dữ liệu không bao gồm các giai đoạn sau 1987, gây khó khăn cho các phân tích xu hướng dài hạn và làm hạn chế tính ứng dụng của dữ liệu trong bối cảnh hiện tại.

* + Thiếu thông tin cá nhân cụ thể:

+ Giới hạn về thông tin nhân khẩu học: Chỉ có một số đặc trưng như dân tộc và tình trạng hôn nhân. Không có các thông tin khác như giới tính, tuổi tác chính xác (tuổi chỉ được suy ra từ kinh nghiệm và năm học), hoặc thông tin gia đình, có thể làm giảm độ sâu của các phân tích về xã hội học.

+ Hạn chế về khu vực cư trú: Chỉ có bốn loại khu vực cư trú chung chung (rural, northeast, north central, south), mà không chi tiết tới cấp thành phố hoặc quận huyện.

* + Giới hạn về đặc trưng sức khỏe

+ Thông tin sức khỏe giới hạn: Chỉ có một cột đơn lẻ về tình trạng sức khỏe (có vấn đề hay không), thiếu các chi tiết cụ thể về loại và mức độ nghiêm trọng của vấn đề sức khỏe. Điều này có thể hạn chế khả năng đánh giá sự ảnh hưởng của các vấn đề sức khỏe đến tiền lương và năng suất lao động.

* + Các biến phân loại không đầy đủ:

+ Mức độ chi tiết ngành nghề: Mặc dù có phân loại ngành nghề và chức danh công việc, nhưng chỉ bao gồm một số ngành và nghề lớn, không phản ánh đủ các nhóm nghề khác hoặc các ngành phụ. Việc này có thể gây sai lệch nếu nhóm công việc quá lớn và không đồng nhất.

+ Dân tộc không chi tiết: Đặc trưng dân tộc chỉ gồm ba nhóm black, hisp, và other, có thể không đại diện đầy đủ cho sự đa dạng về sắc tộc, đặc biệt trong những khu vực có dân số đa sắc tộc.

* + Dữ liệu có thể bị thiếu và sai lệch:

+ Thiếu dữ liệu: Có khả năng một số hàng bị thiếu giá trị ở cột residence, dẫn đến khó khăn khi xử lý hoặc yêu cầu phải có các kỹ thuật thay thế dữ liệu.

+ Sai lệch do tự báo cáo: Các thông tin như sức khỏe, tình trạng hôn nhân, và kinh nghiệm có thể phụ thuộc vào tự báo cáo, dẫn đến khả năng sai lệch nếu người tham gia không cung cấp thông tin chính xác hoặc có thiên kiến cá nhân.

* + Không có các biến kinh tế và xã hội khác:

+ Thiếu biến kinh tế và thị trường: Bộ dữ liệu không bao gồm các yếu tố kinh tế vĩ mô (ví dụ: tỷ lệ thất nghiệp, lạm phát), có thể ảnh hưởng lớn đến tiền lương và điều kiện lao động.

+ Thiếu thông tin về điều kiện làm việc: Không có dữ liệu về giờ làm, tính chất công việc (nguy hiểm hay không), hoặc các yếu tố phúc lợi như bảo hiểm, chế độ nghỉ phép, làm giảm khả năng phân tích sâu về ảnh hưởng của các yếu tố này đến sức khỏe và tiền lương.

* + Giới hạn trong sử dụng log của lương

+ Độ phức tạp trong xử lý log của lương: Việc sử dụng log của lương có thể khó diễn giải và yêu cầu tính toán bổ sung khi muốn chuyển đổi ngược lại về lương thực tế để dễ hiểu hơn.

* 1. **Thống kê mô tả dữ liệu**

Thống kê mô tả là bước đầu tiên và cơ bản trong quá trình phân tích dữ liệu, dùng các thuật toán khác nhau để thông kê và mô tả dữ liệu giúp chúng ta hiểu rõ hơn về bản chất và cấu trúc của dữ liệu thông qua các số liệu thống kê đơn giản.

Lệnh *df.describe()* trong pandas được sử dụng để tính toán và hiển thị các số liệu thống kê tóm tắt về dữ liệu trong DataFrame.

Khi gọi *df.describe()* trênn các cột số, nó trả về các thông tin sau:

* count: Số lượng giá trị không rỗng (non – null).
* mean: Giá trị trung bình
* std: Độ lệch chuẩn
* min: Giá trị nhỏ nhất
* 25%: Phân vị thứ nhất (Q1) (25% giá trị nhỏ hơn hoặc bằng giá trị này)
* 50%: Phân vị thứ hai (Q2) (Median – giá trị trung vị)
* 75%: Phân vị thứ ba (Q3) (75% giá trị nhỏ hơn hoặc bằng giá trị này)
* A screenshot of a computer

  Description automatically generatedmax: Giá trị lớn nhất

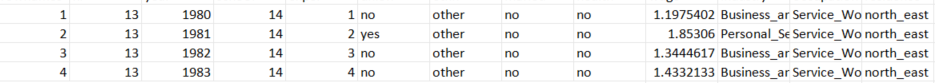
Hình 2.1: Các số liệu được thống kê dữ liệu

**CHƯƠNG 3: LÀM SẠCH VÀ CHUẨN HÓA DỮ LIỆU**

* 1. **Phân tích những vấn đề gặp phải với dữ liệu**

- Nhận định ban đầu dataframe đọc được gặp một vài vấn đề về dữ liệu :

+ (4360,13) : 4360 hàng và 13 cột (df.shape) và có thể chia làm hai dạng như sau:



# Hình 3.1: Giá trị tệp dữ liệu sau khi tách dữ liệu

Đây là các hàng bình thường, các giá trị được phân tách nhau trong trong các cột riêng biệt

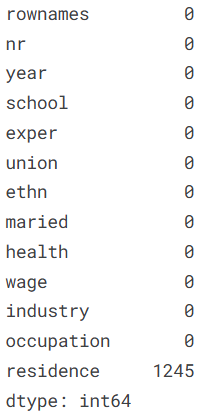


# Hình 3.2: Giá trị tệp dữ liệu trước khi tách dữ liệu

Đây là các hàng chưa đúng định dạng mong muốn, các giá trị của các cột viết liên tiếp và ngăn cách nhau bởi dấu ‘,’ trong cùng 1 ô dữ liệu.

Nên việc đầu tiên là chuyển các hàng chưa đúng định dạng về đúng định dạng mong muốn.

+Sau khi đã hoàn thành công việc đầu tiên. Kiểm tra các giá trị NULL ở các cột bằng(df.isnull().sum())



# Hình 3.3: Kiểm tra và tổng hợp số lượng giá trị bị thiếu (NaN)

Nhận thấy, đa số các cột đều không có gía trị NULL ngoại trừ cột “residence” có 1245 giá trị NULL trên 4360 giá trị của cột này.Vậy dataframe xảy ra việc mất dữ liệu, nó rơi vào khoảng 30% tổng giá trị của cột.Vậy công việc thứ hai là lấp đầy các ô NULL này bằng một giá trị nào đó. Có hai ý tưởng có công việc này là lấp đầy các ô NULL bằng giá trị có tần số xuất hiện nhiều nhất trong cột “residence” hoặc bằng giá trị ‘N/A’ (Not Available). Nhóm em chọn chọn ý tưởng đầu là lấp đầy các ô NULL bằng giá trị “south” có tần số xuất hiện nhiều nhất.

+ **Chuẩn hóa dữ liệu dạng danh mục**: đảm bảo các cột như union, ethn, maried, health có giá trị đồng nhất các giá trị viết hoa/thường.Trong trường hợp này nhóm em sẽ chuyển về chữ thường. Đồng thời bỏ dấu “\_” trong dữ liệu chuỗi và thay bằng dấu cách.

* 1. **Thực hiện ý tưởng giải quyết vấn đề**

A computer screen shot of a program

Description automatically generated+ Tách dữ liệu trong các hàng chưa hợp lệ

Hình 3.4: Code tách dữ liệu trong các hàng chưa hợp lệ

Hàm loadAndCleanCSV thực hiện việc đọc dữ liệu từ tệp CSV và làm sạch các dòng không hợp lệ. Cụ thể, hàm này bắt đầu bằng cách mở tệp CSV và sử dụng csv.reader để đọc từng dòng dữ liệu. Với mỗi dòng, nếu dòng có số cột bằng với số cột mong đợi (cols), nó sẽ được thêm vào danh sách all\_rows. Nếu không, hàm sẽ tách các phần tử trong dòng thành các giá trị riêng lẻ và chuẩn hóa độ dài dòng cho đúng bằng cách chèn thêm giá trị None vào các cột thiếu, cho đến khi dòng đủ số cột. Sau khi tất cả các dòng được xử lý, danh sách all\_rows sẽ được chuyển thành một DataFrame với các tên cột đã định trước. Tiếp theo, hàm clean\_data sẽ được gọi để làm sạch thêm dữ liệu này, bao gồm các thao tác như chuẩn hóa giá trị phân loại, điền giá trị cho các ô thiếu, và chuẩn hóa tên cột. Cuối cùng, dữ liệu đã làm sạch được lưu vào tệp CSV đầu ra tại đường dẫn outPath.

+ Lấp đầy các ô giá trị NULL

A computer screen shot of text

Description automatically generated

Hình 3.5: Code điền giá trị cho các ô bị thiếu giá trị

Hàm lapDayORong thực hiện việc điền giá trị mặc định vào các ô bị thiếu dữ liệu trong cột residence của DataFrame data. Đầu tiên, hàm tìm và thay thế các ô trống hoặc chứa khoảng trắng trong cột residence bằng giá trị NaN (dùng np.nan từ thư viện numpy) để đánh dấu chúng là ô thiếu dữ liệu. Điều này đảm bảo rằng mọi ô rỗng hoặc chứa khoảng trắng đều được coi là dữ liệu thiếu. Sau đó, hàm sử dụng phương thức fillna để thay thế các giá trị NaN bằng giá trị mặc định được chỉ định qua tham số fill\_value, với giá trị mặc định là 'south'. Kết quả là cột residence trong DataFrame được hoàn thiện với các ô thiếu được điền tự động, giúp chuẩn hóa dữ liệu trong cột này trước khi phân tích tiếp.

A close-up of a computer code

Description automatically generated+ Chuẩn hóa dữ liệu phân loại

Hình 3.6: Code chuẩn hóa dữ liệu phân loại

Hàm chuanHoaDuLieuPhanLoai chuẩn hóa các cột phân loại trong DataFrame data để đảm bảo tính nhất quán về định dạng văn bản. Hàm nhận một danh sách categorical\_columns chứa tên các cột cần chuẩn hóa. Với mỗi cột trong danh sách, hàm chuyển tất cả các giá trị văn bản trong cột đó thành chữ thường bằng cách sử dụng phương thức .str.lower(). Điều này giúp loại bỏ sự không nhất quán về chữ hoa và chữ thường, chẳng hạn như "Male" và "male" sẽ được đồng nhất thành "male", thuận tiện cho việc phân tích dữ liệu và so sánh giá trị giữa các bản ghi. Sau khi hoàn tất, hàm trả về DataFrame data với các cột phân loại đã được chuẩn hóa.

+ Hàm xử lí chuỗi (thay dấu gạch dưới bằng dấu cách)

A close-up of a computer code

Description automatically generated

Hình 3.7: Code thay đổi dấu gạch dưới thành khoảng trắng

Dữ liệu dễ đọc hơn khi sử dụng dấu cách thay vì dấu gạch dưới. Thay thế dấu gạch dưới bằng dấu cách giúp đồng bộ cách biểu diễn các giá trị, đặc biệt khi kết hợp hoặc so sánh dữ liệu từ nhiều nguồn khác nhau. Khi hiển thị dữ liệu trên biểu đồ, báo cáo hoặc giao diện người dùng, dấu cách thường được ưu tiên hơn để tạo cảm giác chuyên nghiệp và thân thiện. Nhiều công cụ phân tích dữ liệu hoặc trực quan hóa (như biểu đồ, báo cáo) xử lý văn bản tốt hơn nếu các giá trị không chứa các ký tự không cần thiết như dấu gạch dưới.

* 1. **Xây dựng thêm thuộc tính**

Nhằm cải thiện khả năng phân tích và nâng cao chất lượng dữ liệu, cần xây dựng thêm các thuộc tính cho cơ sở dữ liệu.

* + Mở rộng chiều thông tin: Các thuộc tính bổ sung cung cấp nhiều dữ liệu hơn, giúp phân tích sâu hơn và có cái nhìn toàn diện hơn về vấn đề.
  + Phân nhóm tốt hơn: Các thuộc tính mới giúp phân nhóm dữ liệu theo các tiêu chí khác nhau, như độ tuổi, cấp độ công việc, khu vực địa lý, v.v.
  + Dự đoán chính xác hơn: Trong các mô hình học máy hoặc thống kê, các thuộc tính mới có thể cải thiện độ chính xác và tính khả thi của dự đoán.

Đối với cơ sở dữ liệu hiện có, tác giả quyết định thêm 4 thuộc tính:

* + *age* (tuổi): được tính từ exper, school và 6 năm đầu đời (age = 6 + school + exper). Dữ liệu này giúp xác định độ tuổi và phân tích theo độ tuổi.

Code: df['age'] = 6 + df['school'] + df['exper']

* + *exper level:* Dựa trên *exper,* phân chia thành cấp dộ như sau:
* Beginner (kinh nghiệm dưới 3 năm)
* Intermediate (kinh nghiệm 3 – 8 năm)
* Advanced (kinh nghiệm 8 – 12 năm)
* Expert (kinh nghiệm trên 12 năm)

Code: df['exper level'] = pd.cut(

df['exper'],

bins=[ -float('inf'), 3, 8, 12, float('inf')],

labels=['beginner', 'intermediate', 'advanced', 'expert'])

* + *School level* (trình độ học vấn): Dựa trên *school,* phân chia thành các cấp độ như sau:
* Very low (học vấn dưới 5 năm)
* Low (học vấn từ 5 – 9 năm)
* Intermediate (học vấn 9 – 12 năm)
* High (học vấn 12 – 16 năm)

Code: df['school level'] = pd.cut(

df['school'],

bins=[0, 5, 9, 12, 16],

labels=['very low', 'low', 'intermediate', 'high'],

right=True)

* + *Wage level:* Dựa trên *wage,* phân chia thành các cấp độ như sau:
* Very low (log of hourly wage dưới -2)
* Low (từ -2 tới 0)
* Medium (từ 0 tới 2)
* High (từ 2 tới 3.5)
* Very high (trên 3.5)

Code : df['wage level'] = pd.cut(

df['wage'],

bins=[-float('inf'), -2, 0, 2, 3.5, float('inf')],

labels=['very low', 'low', 'medium', 'high', 'very high'])

## CHƯƠNG 4: XÂY DỰNG CƠ CHẾ CRUD TRÊN TẬP DỮ LIỆU

**CRUD** là viết tắt của **Create, Read, Update, Delete** - các hoạt động cơ bản thường được thực hiện trên dữ liệu trong hệ thống phần mềm, đặc biệt là cơ sở dữ liệu (database).

* 1. **Create**

Việc tạo bản ghi giúp thêm dữ liệu mới vào cơ sở dữ liệu để mở rộng thông tin và tăng cường khả năng quản lý. Tạo bản ghi thường được thực hiện khi có dữ liệu mới.

Quy trình thực hiện:

* Nhấn nút Create
* A screenshot of a computer

  Description automatically generatedNhập thông tin cho bản ghi.

Hình 4.1: Cửa sổ nhập dữ liệu cho bảng ghi mới

* 1. **Read**

Trong lĩnh vực phân tích dữ liệu, việc đọc và xử lý dữ liệu là một trong những bước quan trọng nhất để chuẩn bị cho quá trình phân tích và ra quyết định. Việc này không chỉ giúp chúng ta hiểu rõ dữ liệu mà còn tối ưu hóa hiệu quả sử dụng thông tin. Các thao tác như đọc một bản ghi, phân trang dữ liệu và sắp xếp dữ liệu đóng vai trò quan trọng trong việc xử lý dữ liệu quy mô lớn.

A screenshot of a computer

Description automatically generated Trong bài báo cáo này, tác giả trình bày 3 thao tác: đọc 1 bản ghi cụ thể thông qua rowname, phân trang dữ liệu, tìm tiếm và lọc dữ liệu.

Hình 4.2: Cửa sổ lựa chọn thao tác READ

1. Đọc một bản ghi

A screenshot of a computer

Description automatically generatedĐọc một bản ghi cụ thể trong cơ sở dữ liệu hoặc tập dữ liệu là một thao tác cơ bản và thường xuyên trong xử lý dữ liệu. Thao tác này được sử dụng khi cần truy xuất thông tin chi tiết của một đối tượng hoặc hàng dữ liệu cụ thể thông qua rowname.

Hình 4.3: Cửa sổ chọn bản ghi

A screenshot of a computer

Description automatically generatedGiao diện hiển thị cửa sổ yêu cầu nhập rowname, sau đó nhấn “Show Record Details”. Nếu tìm thấy, cửa sổ “Record Details” sẽ hiện lên và cho xem thông tin về đối tượng đã chọn.

Hình 4.4: Cửa sổ hiển thi thông tin bản ghi

1. Phân trang dữ liệu

Phân trang là kỹ thuật chia dữ liệu lớn thành các phần nhỏ hơn, thường sử dụng trong việc hiển thị dữ liệu trên giao diện người dùng hoặc tối ưu hóa xử lý dữ liệu.

A screenshot of a computer

Description automatically generatedMỗi trang chứa 32 bản ghi tương ứng với thông tin khảo sát của của 4 người trong vòng 8 năm. Để di chuyển sang các trang khác nhau, nhấn vào nút “Next” để đi đến trang kế tiếp, “Previous” để quay lại trang trước đó.

Hình 4.5: Kết quả hiển thị phân trang dữ liệu

1. Tìm kiếm và lọc

Tìm kiếm và lọc dữ liệu là hai thao tác quan trọng trong phân tích và xử lý dữ liệu. Khi làm việc với các tập dữ liệu lớn, không phải tất cả thông tin đều cần thiết. Việc tìm kiếm cho phép truy vấn thông tin cụ thể dựa trên điều kiện nhất định, trong khi lọc giúp loại bỏ dữ liệu không liên quan và giữ lại những phần cần thiết.

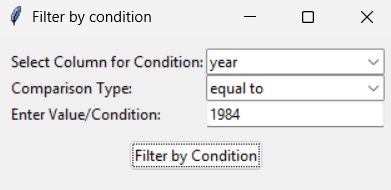
Trong bài báo cáo này tác giả thực hiện 4 thao tác tìm kiếm và lọc, bao gồm: filter by condition (lọc theo điều kiện của một cột bất kì), filter by colums (lọc theo các cột được chọn để xuất hiện), fliter by rows (lọc theo hàng), sort by column (sắp xếp cột tăng dần hay giảm dần).

A screenshot of a computer

Description automatically generated

Hình 4.6: Cửa số thực hiện lựa chọn phương pháp lọc

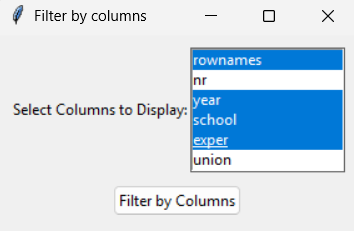
Trong cửa sổ Search and Filter, lựa chọn phương phương thức lọc

+ Filter by condition: lựa chọn cột điều kiện, lựa chọn phương thức so sánh (>, >=, <, <=, ==, !=), nhập giá trị so sánh và nhấn “Filter by Condition” để xuất hiện cửa sổ chứa dữ liệu được lọc.

A screenshot of a computer

Description automatically generatedHình 4.7: Cửa sổ thực hiện chức năng theo điều kiện cột

Hình 4.8: Kết quả hiển thị lọc dữ liệu theo điều kiện của cột

+ Filter by columns: lựa chọn các cột để đọc dữ liệu, nhấn “Filter by Columns” để xuất hiện cửa sổ chứa dữ liệu được lọc.

Hình 4.9: Cửa sổ thực hiện chức năng lọc các cột xuất hiện

A screenshot of a computer

Description automatically generated

Hình 4.10: Kết quả hiển thị sau khi lọc tên cột

+ Filter by rows: lọc ra các bản ghi từ from\_row đến to\_row

A screenshot of a computer

Description automatically generatedresult = df.iloc[from\_row:to\_row + 1]

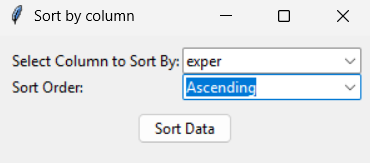
Hình 4.11: Cửa sổ hiển thị yêu cầu hàng để lọc

A screenshot of a computer

Description automatically generated

Hình 4.12: Kết quả hiển thị các hàng được lọc

+ Sort by columns: chọn 1 cột trong danh sách để s

ắp xếp (Select Column to Sort), có 2 cách sắp xếp: tăng dần (ascending) và giảm dần (descending)

Hình 4.13: Cửa sổ thực hiện thao tác sắp xếp

**A screenshot of a computer

Description automatically generated**

Hình 4.14: Hiển thị kết quả sắp xếp

### 3.3. Update

A screenshot of a computer

Description automatically generatedCập nhật dữ liệu (Update) là một trong những chức năng quan trọng của quản lý dữ liệu, cho phép thay đổi hoặc bổ sung giá trị của các bản ghi trong cơ sở dữ liệu.

Hình 4.15: Cửa sổ lựa chọn thao tác UPDATE

1. Cập nhật một bản ghi

Xác Định Bản Ghi Cần Cập Nhật:

* Xác định chính xác dòng dữ liệu (record) hoặc tập hợp bản ghi cần thay đổi thông qua một cột rowname.
* Đảm bảo việc xác định không bị nhầm lẫn hoặc ảnh hưởng đến các bản ghi không liên quan.

Cập Nhật Dữ Liệu:

* Tiến hành thay đổi giá trị dựa trên điều kiện và thông tin xác định.

Kiểm Tra Sau Cập Nhật:

* Kiểm tra lại tính chính xác của dữ liệu đã được cập nhật.
* So sánh giá trị trước và sau khi cập nhật để đảm bảo không bị lỗi

**A screenshot of a computer

Description automatically generated**

Hình 4.16: Cửa sổ chỉnh sửa lại thông tin bản ghi

1. Cập nhật lương dựa trên kinh nghiệm

Ý nghĩa của việc cập nhật dữ liệu lương dựa trên kinh nghiệm:

* Thúc đẩy động lực làm việc: Người lao động có kinh nghiệm cao sẽ được công nhận và hưởng mức lương phù hợp.
* Đảm bảo tính công bằng: Các mức lương được điều chỉnh dựa trên tiêu chí rõ ràng, không cảm tính.
* Hỗ trợ quản lý dữ liệu lớn: Việc tự động hóa giúp tránh sai sót khi xử lý thủ công và tiết kiệm thời gian.

Quy trình cập nhật dữ liệu:

* Đọc dữ liệu từ cơ sở dữ liệu
* Xác định tỷ lệ (%) tăng lương tương ứng với từng cấp độ (beginner, intermediate, advanced, expert)
* Áp dụng mức tăng lương vào dữ liệu hiện tại
* Tự động cập nhật lại các thuộc tính liên quan

Code:

df["wage"] = df.apply(

lambda row: row["wage"] + math.log(1 + raise\_map.get(row["exper level"], 0)),

axis=1

**A screenshot of a computer

Description automatically generated**)

Hình 4.17: Cửa sổ cập nhật tăng lương theo kinh nghiệm việc làm

1. Cập nhật lương dựa trên trình độ học vấn

Ý nghĩa của trình độ học vấn trong mức lương:

* Nâng cao giá trị cá nhân: Người có trình độ học vấn cao hơn thường mang lại giá trị lớn hơn cho tổ chức.
* Khuyến khích học tập và phát triển: Chính sách lương phù hợp với trình độ học vấn thúc đẩy sự học hỏi và phấn đấu trong công việc.
* Tạo sự công bằng: Mức lương được xây dựng dựa trên tiêu chí minh bạch, giúp tránh tình trạng chênh lệch không hợp lý.

Quy trình cập nhật dữ liệu

* Đọc dữ liệu từ cơ sở dữ liệu
* Xác định tỷ lệ (%) tăng lương tương ứng với từng cấp độ (vẻy low, low, intermediate, high)
* Áp dụng mức tăng lương vào dữ liệu hiện tại
* Tự động cập nhật lại các thuộc tính liên quan

Code:

df["wage"] = df.apply(

lambda row: row["wage"] + math.log(1 + raise\_map.get(row["school level"], 0)),

axis=1

**A screenshot of a computer

Description automatically generated**)

Hình 4.18: Cửa sổ tăng lương theo trình độ học vấn

### 3.4. Delete

Xóa bản ghi giúp loại bỏ những thông tin không còn phù hợp, sai lệch, hoặc không cần thiết. Điều này làm tăng tính chính xác và hiệu quả của cơ sở dữ liệu.

Quy tình thực hiện:

* Xác định bản muốn xóa và nhấp chuột vào
* Trên thanh chức năng bên phải màn hình nhấp chuột vào nút “Delete” để xóa đi bản ghi đó.

A screenshot of a computer

Description automatically generated

Hình 4.19: Hướng dẫn thực hiện xóa một bản ghi

## CHƯƠNG 5: VẼ CÁC LOẠI BIỂU ĐỒ BIỂU DIỄN DỮ LIỆU

* 1. **Khái quát trực quan hóa dữ liệu**

Trực quan hóa dữ liệu (Data Visualization) là quá trình biến đổi dữ liệu thô thành các biểu diễn trực quan, dễ hiểu, như biểu đồ, đồ thị, bảng, và bản đồ. Mục tiêu của trực quan hóa là giúp người dùng:

* Hiểu rõ cấu trúc dữ liệu.
* Nhận diện xu hướng, mẫu, và mối quan hệ trong dữ liệu.
* Hỗ trợ quá trình ra quyết định dựa trên dữ liệu

Vai trò của Trực quan hóa Dữ liệu

* Cải thiện khả năng hiểu dữ liệu: Giúp người dùng không cần phân tích dữ liệu thô phức tạp.
* Tăng tốc phân tích: Cung cấp cái nhìn tổng quan về dữ liệu ngay lập tức.
* Hỗ trợ giao tiếp: Truyền đạt thông tin hiệu quả hơn thông qua các biểu đồ, đồ thị trực quan.
* Phát hiện vấn đề: Giúp nhận diện các bất thường hoặc mẫu ẩn trong dữ liệu
  1. **Phân tích và lựa chọn loại biểu đồ phù hợp**

a) Biểu đồ quan hệ giữa mức lương (‘wage’) và số năm học tập (‘school’)

Phân tích mối quan hệ giữa mức lương (Wage) và số năm học tập (Education) thường giúp đánh giá tác động của giáo dục đến thu nhập. Để trực quan hóa mối quan hệ này, biểu đồ phân tán (scatter plot).

**A graph of a graph with blue dots

Description automatically generated**Hình 5.1: Biểu đồ thể hiện sự liên hệ giũa “wage” và “school”

b) Biểu đồ quan hệ giữa lương (‘wage’) và nghành nghề (‘industry’)

**Biểu đồ boxplot**: Để phân tích sự phân bố của lương trong các ngành nghề, giúp dễ dàng nhận thấy các điểm ngoại lệ và sự chênh lệch giữa các nhóm ngành nghề, so sánh phân bố của lương trong các ngành nghề khác nhau, giúp nhận diện ngành nào có mức lương cao hơn, hoặc có sự phân tán lớn hơn về mức lương

A diagram with blue rectangles and black text

Description automatically generated

Hình 5.2: Biểu đồ thể hiện sự liên hệ giữa “wage” và “industry”

c)Biểu đồ thống kê mức lương

Biểu đồ **histplot** thể hiện **lương theo đầu người** (tức là phân phối lương cho mỗi cá nhân trong dữ liệu), bạn có thể sử dụng **seaborn.histplot()** với trục **x** là **lương (wage)** và **y** là tần suất xuất hiện của các mức lương.

Nhận xét biểu đồ: hầu hết các điểm dữ liệu đều tập trung trong khoảng từ 1 đến 2, điều này cho thấy mức lương trung bình trong nguồn thu nhập của dữ liệu.

A graph of a distribution of wage

Description automatically generated

Hình 5.3: Biểu đồ thống kê logarit của lương theo giờ

1. A graph with blue squares

   Description automatically generatedCác biểu đồ khác

Hình 5.4: Biểu đồ thống kê số người ở các dân tộc khác nhau

Biểu đồ ethn (sns.countplot()): ngoài 2 dân tộc là black và hisp, thì còn nhiều dân tộc khác và những nhóm dân tộc khác có tần suất ít hơn 2 dân tộc chính

A blue rectangular object with white text

Description automatically generatedBiểu đồ maried (sns.countplot()): cho tháy tình trạng hôn nhân, chiếm số đông hơn là chưa kết hôn.

Hình 5.5: Biểu đồ thống kê trình trạng hôn nhân

Biểu đồ health(sns.countplot()): thể hiện trình trạng sức khỏe (có vấn đề về sức khỏe hay không?). Từ biểu đồ cho thấy mọi người đa số không gặp vấn đề gì về sức khỏe

A blue rectangle with white text

Description automatically generated

Hình 5.6: Biểu đồ thống kê tình trạng sức khỏe

Biểu đồ industry(sns.countplot()): thể hiện số lượng làm việc trong các lĩnh vực, nhóm ngành khác nhau. Dựa vào biểu đồ ta thấy 2 lĩnh vực chiếm ưu thế bỏ xa các nhóm ngành khác là manufacturing (sản xuất) và trade (thương mại). Biểu đồ giúp ta dễ dàng nhận diện các ngành lớn và ngành nhỏ, hỗ trợ trong việc phân tích cấu trúc thị trường lao động.

A bar graph with blue and white bars

Description automatically generated

Hình 5.7: Biểu đồ thống kê số lượng trong các lĩnh vực lao động khác nhau

Biểu đồ occupation (sns. countplot()): thể hiện số lượng lao động trong các nghề, trong đó nhóm nghề craftsmen, foremen, and kindred và operatives and kindred chiếm số đông. Các nghề khác sẽ có tần suất thấp hơn là các nghề ít phổ biến hơn. So sánh với các nghề khác có số lượng lao động ít hơn hỗ trợ trong việc phân tích thị trường lao động và chiến lược phát triển các nhóm nghề.

A graph of numbers and letters

Description automatically generated with medium confidence

Hình 5.8: Biểu đồ thống kê số lượng lao động trong các nhóm nghề khác nhau

Biểu đồ residence (sns. countplot()): cho thấy số lượng lao động phân bố theo các khu vực địa lý, trong đó south có thể sẽ có thanh cao nhất, thể hiện rằng lao động đang đổ về khu vực này, trong khi rural area sẽ có thanh thấp hơn, cho thấy ít lao động tại các khu vực nông thôn. Điều này phù hợp với nhận định của bạn về quá trình đô thị hóa, khi nguồn lao động di chuyển từ các khu vực nông thôn đến các khu vực đô thị (đặc biệt là A graph with blue squares

Description automatically generatedkhu vực south).

Hình 5.9: Biểu đồ thống kê số lượng lao động ở các khu vực cư trú khác nhau

## CHƯƠNG 6: XÂY DỰNG GIAO DIỆN CHƯƠNG TRÌNH

* 1. **Giới thiệu về Tkinter**

Tkinter là một thư viện tiêu chuẩn trong Python, được sử dụng để xây dựng các ứng dụng đồ họa người dùng (GUI - Graphical User Interface). Đây là thư viện đơn giản, dễ sử dụng và phổ biến cho những ai mới bắt đầu học lập trình GUI trong Python. Tkinter giúp tạo ra các cửa sổ, nút bấm, trường văn bản, menu, hộp thoại, và các thành phần đồ họa khác, giúp người dùng có thể tương tác với chương trình một cách trực quan.

Tkinter là một giao diện Python đối với Tk, một thư viện GUI cũ nhưng mạnh mẽ viết bằng ngôn ngữ C. Mặc dù có thể sử dụng các thư viện GUI khác như PyQt, wxPython hay Kivy, Tkinter vẫn là lựa chọn phổ biến và phù hợp với những ứng dụng nhỏ, đơn giản.

* 1. **Mục đích sử dụng Tkinter trong phân tích dữ liệu**
* Tạo giao diện người dùng (GUI) đơn giản: Tkinter có thể giúp tạo các ứng dụng phân tích dữ liệu với giao diện người dùng dễ hiểu, hỗ trợ các thao tác như chọn tệp dữ liệu, lọc dữ liệu, hiển thị đồ họa trực quan, v.v.
* Tương tác với dữ liệu: Tkinter có thể tích hợp các widget như Entry, Button, Listbox, và Canvas để người dùng nhập vào, điều chỉnh hoặc hiển thị dữ liệu trực tiếp trên giao diện.
* Trực quan hóa dữ liệu: Với khả năng tích hợp với các thư viện đồ họa như Matplotlib, Tkinter có thể được sử dụng để xây dựng các ứng dụng giúp người dùng trực quan hóa và phân tích dữ liệu.
  1. **Các thành phần cơ bản của Tkinter**

Tkinter cung cấp một bộ công cụ rất cơ bản để xây dựng giao diện đồ họa, bao gồm:

* Tk(): Đây là đối tượng chính để khởi tạo một cửa sổ ứng dụng GUI.
* Label: Được sử dụng để hiển thị văn bản, hình ảnh hoặc thông tin trong cửa sổ.
* Button: Nút bấm dùng để thực hiện các hành động khi người dùng nhấn vào.
* Entry: Trường văn bản cho phép người dùng nhập liệu.
* Canvas: Khu vực vẽ đồ họa, có thể vẽ hình, đường, văn bản và nhiều đối tượng khác.
* Frame: Một vùng chứa các widget con khác.
* Menu: Tạo thanh menu cho ứng dụng GUI.
* Scale, Scrollbar, Listbox, Text: Các widget khác phục vụ cho các tác vụ nhập liệu, hiển thị danh sách, cuộn văn bản..
  1. **Xây dựng giao diện tương tác phân tích dữ liệu sử dụng Tkinter**

**A screenshot of a computer code

Description automatically generated**

Hình 6.1: Code xây dựng giao diện

Kết quả giao diện:

**A screenshot of a computer

Description automatically generated**

Hình 6.2: Giao diện người dùng

# PHẦN KẾT LUẬN

* + 1. **Kết quả đạt được**
* Về kiến thức

+ Đựa trên tập dữ liệu:

* Tìm hiểu, đánh giá về tập dữ liệu ban đầu, phân tích những ưu, khuyết điểm, nắm bắt được những lỗi trong dữ liệu và trong quá trình thu thập. Từ đó đưa ra những hướng xử lí phù hợp.
* Kiến thức về phân tích, xử lí, chuẩn hóa và trực quan hóa dữ liệu lớn, thông qua những thư viện đã học để viết chương trình thực hiện các yêu cầu thêm, xóa, sửa,... trên tập dữ liệu.
* Hiểu cách tổ chức chương trình theo modules – packages, theo từng nhóm chức năng công dụng, mục đích.
* Nắm rõ kiến thức về xây dựng đồ họa tương tác, bước đầu tìm hiểu hiểu về xây dựng phần mềm.
* Về sản phẩm
* Phần mềm cho phép người dùng tương tác trực tiếp trên tập dữ liệu, các chức năng hoạt tốt và khá đa dạng.
* Tính ứng dụng thực tiễn tốt, đáp ứng các nhu cầu từ cơ bản đến nâng cao.
* Ưu và nhược điểm:
* Ưu điểm: Người dùng dễ dàng thực hiện các thao tác trên giao diện, thực hiện nay, chính xác.
* Nhược điểm: Nguồn dữ liệu còn những thiếu sót dẫn tới có sự sai sót trong các kết quả lọc dữ liệu, các đặc trưng thuộc tính chưa đa dạng.

1. **Hướng phát triển**

* Thêm các yếu tố môi trường cho sản phẩm.

# PHẦN TÀI LIỆU THAM KHẢO

[1]. Kaggle , *Wages and Education of Young Males Dataset*, truy cập ngày 27/11/2024, đường dẫn: https://www.kaggle.com/datasets/jacopoferretti/wages-and-education-of-young-males-dataset/data