Trường Đại học Kinh tế Đà Nẵng

Khoa Thống kê – Tin học



**BÁO CÁO THỰC TẬP NGHỀ NGHIỆP**

**Tên đề tài:**

**PHÂN TÍCH DỮ LIỆU KHÁCH HÀNG TẠI NGÂN HÀNG**

Sinh viên thực hiện: Trương Hoàng Long

Lớp: 44K21.1

Giảng viên hướng dẫn: Ths. Cao Thị Nhâm

Đà Nẵng, ngày 17 tháng 7 năm 2021

Đà Nẵng

LỜI CAM ĐOAN

Em xin cam đoan bài báo cáo thực tập nghề nghiệp là sự nghiên cứu độc lập của em dựa trên sự nỗ lực, cố gắng, tìm hiểu dưới sự hướng dẫn của giáo viên hướng dẫn: Ths. Cao Thị Nhâm. Bài báo cáo không có bất cứ sự sao chép của người khác. Nếu có phát hiện điều gì không trung thực em xin hoàn toàn chịu trách nhiệm.

LỜI CẢM ƠN

Em xin cảm ơn Khoa Thống kê – Tin học đã tạo điều kiện để em có kỳ thực tập nghề nghiệp này. Kỳ thực tập nghề nghiệp này đã giúp em hiểu hơn về cơ hội nghề nghiệp cũng như các kỹ năng cần thiết để làm việc sau này.

Em xin cảm ơn cô Cao Thị Nhâm đã hướng dẫn em trong kỳ thực tập này. Cô đã tận tâm hướng dẫn để em có thể hoàn thành kỳ thực tập và hỗ trợ em để hoàn thiện bài báo cáo.

MỤC LỤC

[LỜI CAM ĐOAN 2](#_Toc80390481)

[LỜI CẢM ƠN 3](#_Toc80390482)

[MỤC LỤC 4](#_Toc80390483)

[DANH SÁCH HÌNH ẢNH 7](#_Toc80390484)

[DANH SÁCH BẢNG BIỂU 10](#_Toc80390485)

[LỜI MỞ ĐẦU 11](#_Toc80390486)

[CHƯƠNG 1: TỔNG QUAN VỀ CÔNG NGHỆ VÀ CÔNG CỤ SỬ DỤNG 11](#_Toc80390487)

[1.1. Giới thiệu về phân tích dữ liệu 11](#_Toc80390488)

[Phân tích dữ liệu là gì? 11](#_Toc80390489)

[1.2. Công cụ sử dụng 12](#_Toc80390490)

[1.2.1. Công nghệ sử dụng 12](#_Toc80390491)

[1.2.2. Ngôn ngữ Python: 13](#_Toc80390492)

[1.2.3. Thư viện Pandas 13](#_Toc80390493)

[1.2.4. Thư viện Numpy 13](#_Toc80390494)

[1.2.5. Thư viện Matplotlib 13](#_Toc80390495)

[1.3. Cơ hội nghề nghiệp 13](#_Toc80390496)

[1.3.1. Vị trí việc làm của phân tích dữ liệu 13](#_Toc80390497)

[1.3.2. Mô tả vị trí việc làm 14](#_Toc80390498)

[1.3.3. Các kĩ năng cần có của chuyên viên phân tích dữ liệu 14](#_Toc80390499)

[1.3.4. Mức lương 15](#_Toc80390500)

[CHƯƠNG 2: TỔNG QUAN CƠ SỞ LÝ THUYẾT 15](#_Toc80390501)

[2.1. Python cơ bản 15](#_Toc80390502)

[2.1.1. Số học Python 15](#_Toc80390503)

[2.1.2. Các kiểu dữ liệu cơ bản 16](#_Toc80390504)

[2.1.3. Biến trong Python 18](#_Toc80390505)

[2.1.4. Danh sách trong Python 18](#_Toc80390506)

[2.1.5. Mảng 18](#_Toc80390507)

[2.1.6. Câu điều kiện IF 20](#_Toc80390508)

[2.1.7. Vòng lặp 20](#_Toc80390509)

[2.1.8. Hàm 22](#_Toc80390510)

[2.2. Pandas 22](#_Toc80390511)

[2.3. Numpy 26](#_Toc80390512)

[2.4. Matplotlib 27](#_Toc80390513)

[2.4.1. Biểu đồ cột (histogram) 27](#_Toc80390514)

[2.4.2. Biểu đồ Boxplot 27](#_Toc80390515)

[2.4.3. Biểu đồ Density 28](#_Toc80390516)

[2.4.4. Biểu đồ Scatter 28](#_Toc80390517)

[2.4.5. Biểu đồ Bar 29](#_Toc80390518)

[CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU 30](#_Toc80390519)

[3.1. Nhiệm vụ về đề tài 30](#_Toc80390520)

[3.2. Quy trình thực hiện 30](#_Toc80390521)

[3.3. Dữ liệu 30](#_Toc80390522)

[3.3.1. Giới thiệu dữ liệu 30](#_Toc80390523)

[3.3.2. Giải thích các trường của dữ liệu 31](#_Toc80390524)

[3.3.3. Nhiệm vụ 32](#_Toc80390525)

[3.3.4. Xử lý và làm sạch dữ liệu 32](#_Toc80390526)

[3.4. Trực quan hóa dữ liệu 37](#_Toc80390527)

[Tổng kết 46](#_Toc80390528)

[Tài liệu tham khảo 47](#_Toc80390529)

DANH SÁCH HÌNH ẢNH

[Hình 1.1: Logo Python 9](#_Toc80389768)

[Hình 1.2.1: Logo Google Colaboratory 11](#_Toc80389769)

[Hình 2.1.1.1: Code lấy căn bậc 2 14](#_Toc80389770)

[Hình 2.1.1.2: Code làm tròn số thập phân 14](#_Toc80389771)

[Hình 2.1.1.3: Code lấy giá trị tuyệt đối 14](#_Toc80389772)

[Hình 2.1.2.1: Code kiểm tra kiểu dữ liệu 15](#_Toc80389773)

[Hình 2.1.2.2: Code kiểm tra kiểu Float 15](#_Toc80389774)

[Hình 2.1.2.3: Code kiểm tra kiểu dữ liệu Bool 15](#_Toc80389775)

[Hình 2.1.2.4: Code kiểm tra kiểu dữ liệu strings 16](#_Toc80389776)

[Hình 2.1.2.5: Code kiểm tra giá trị bị thiếu 16](#_Toc80389777)

[Hình 2.1.3.1: Khai báo biến 16](#_Toc80389778)

[Hình 2.1.4.1: Lists 16](#_Toc80389779)

[Hình 2.1.5.1: Mảng 17](#_Toc80389780)

[Hình 2.1.5.2: Thực hiện phép toán cộng với mảng 17](#_Toc80389781)

[Hình 2.1.5.3: Sắp xếp lại mảng 17](#_Toc80389782)

[Hình 2.1.6.1: Lệnh IF 18](#_Toc80389783)

[Hình 2.1.7.1: Vòng lặp For 19](#_Toc80389784)

[Hình 2.1.7.2: Vòng lặp While 20](#_Toc80389785)

[Hình 2.1.8.1: Hàm 20](#_Toc80389786)

[Hình 2.2.1: Đọc dữ liệu 20](#_Toc80389787)

[Hình 2.2.2: Code truy xuất dữ liệu 21](#_Toc80389788)

[Hình 2.2.3: Code truy xuất dữ liệu theo điều kiện 21](#_Toc80389789)

[Hình 2.2.4: Code kiểu dữ liệu từng cột 22](#_Toc80389790)

[Hình 2.2.5: Code lấy dữ liệu theo điều kiện 22](#_Toc80389791)

[Hình 2.2.6: Code lấy 100 dòng dữ liệu đầu tiên 23](#_Toc80389792)

[Hình 2.2.7: Code xóa cột 23](#_Toc80389793)

[Hình 2.2.8: Code sửa giá trị cột 24](#_Toc80389794)

[Hình 2.3.1: Code khởi tạo mảng ngẫu nhiên 24](#_Toc80389795)

[Hình 2.3.2: Code tạo mảng cụ thể 25](#_Toc80389796)

[Hình 2.4.1: Biểu đồ Histogram 25](#_Toc80389797)

[Hình 2.4.2: Biểu đồ Boxplot 26](#_Toc80389798)

[Hình 2.4.3: Biểu đồ Density 26](#_Toc80389799)

[Hình 2.4.4: Biểu đồ Sactter 27](#_Toc80389800)

[Hình 2.4.5: Biểu đồ Bar 28](#_Toc80389801)

[Hình 3.2: Quy trình thực hiện 28](#_Toc80389802)

[Hình 3.3.1: Dữ liệu 29](#_Toc80389803)

[Hình 3.3.4.1: Import các thư viện cần thiết 30](#_Toc80389804)

[Hình 3.3.4.2: Đọc dữ liệu vào Google Colaboratory 31](#_Toc80389805)

[Hình 3.3.4.3: Code xem thống kê cơ bản các trường dữ liệu 31](#_Toc80389806)

[Hình 3.3.4.4: Code kiểm tra dữ liệu 32](#_Toc80389807)

[Hình 3.3.4.5: Code xóa dữ liệu 32](#_Toc80389808)

[Hình 3.3.4.6: Code kiểm tra tần suất dữ liệu 33](#_Toc80389809)

[Hình 3.3.4.7: Code tìm giá trị unknown 33](#_Toc80389810)

[Hình 3.3.4.8: Code xóa dữ liệu unknown 34](#_Toc80389811)

[Hình 3.3.4.9: Code xóa kiểm tra lại dữ liệu 34](#_Toc80389812)

[Hình 3.3.4.9: Code mã hóa các trường cần thiết 35](#_Toc80389813)

[Hình 3.4.1: Code kiểm tra độ tương quan giữa các cột 35](#_Toc80389814)

[Hình 3.4.2: Biểu đồ scatter hiển thị độ tương quan giữa các cột 36](#_Toc80389815)

[Hình 3.4.3: Lọc dữ liệu theo đăng ký tiền gửi có kỳ hạn 37](#_Toc80389816)

[Hình 3.4.4: Biểu đồ cột về Duration 37](#_Toc80389817)

[Hình 3.4.5: Biểu đồ cột về Balance 38](#_Toc80389818)

[Hình 3.4.6: Biểu đồ Bar hiển thị độ tuổi người đăng ký tiền gửi có kỳ hạn 39](#_Toc80389819)

[Hình 3.4.7: Biểu đồ cột hiển thị số liên hệ trong chiến dịch này 39](#_Toc80389820)

[Hình 3.4.7: Biểu đồ Bar hiển thị trình độ giáo dục 40](#_Toc80389821)

[Hình 3.4.7: Biểu đồ Bar hiển thị nghề nghiệp 41](#_Toc80389822)

[Hình 3.4.8: Biểu đồ Bar hiển thị khoản vay mua nhà 42](#_Toc80389823)

[Hình 3.4.9: Biểu đồ Bar hiển thị khoản vay cá nhân 43](#_Toc80389824)

[Hình 3.4.10: Biểu đồ Bar hiển thị tình trạng hôn nhân 44](#_Toc80389825)

DANH SÁCH BẢNG BIỂU

[Bảng 3.3.2: Giải thích các trường của dữ liệu 30](#_Toc80390087)

LỜI MỞ ĐẦU

1. Mục tiêu đề tài
   * Nghiên cứu ngôn ngữ lập trình Python.
   * Phân tích dữ liệu khách hàng tại ngân hàng.
2. Nhiệm vụ của đề tài

* Nghiên cứu ngôn ngữ lập trình Python về phân tích dữ liệu.
* Dùng ngôn ngữ Python để trực quan hóa dữ liệu và phân tích dữ liệu.
* Phân tích dữ liệu về khách hàng tại ngân hàng.

1. Kết cấu của đề tài

Đề tài được tổ chức gồm 3 phần nội dung và phần kết luận

* Chương 1: Tổng quan về công nghệ và công cụ sử dụng
* Chương 2: Phân tích dữ liệu
* Chương 3: Kết quả
* Kết luận

CHƯƠNG 1: TỔNG QUAN VỀ CÔNG NGHỆ VÀ CÔNG CỤ SỬ DỤNG

* 1. Giới thiệu về phân tích dữ liệu

Phân tích dữ liệu là gì?

Phân tích dữ liệu là quá trình đánh giá, khám phá dữ liệu sau đó đưa ra các quyết định dự đoán bằng các công cụ phân tích như: ngôn ngữ lập trình Python hoặc R, Microsoft Excel, …



Hình 1.1: Logo Python

**Quy trình phân tích dữ liệu:**

* Data Requirement Gathering: xác định mục tiêu nghiên cứu.
* Data Collection: thu thập dữ liệu.
* Data Cleaning: làm sạch dữ liệu.
* Data Analysis: phân tích dữ liệu.
* Data Interpretation: giải thích dữ liệu.
* Data Visualization: Trực quan hoá dữ liệu
* **Data Requirement Gathering:** xác định mục tiêu nghiên cứu

Xác định được mục tiêu của việc phân tích dữ liệu.

Xác định cách đo và công nghệ sử dụng

* **Data Collection:** thu thập dữ liệu

Dữ liệu thu thập phải được sắp xếp và xử lý để phân tích.

* **Data Cleaning:** làm sạch dữ liệu

Dữ liệu thu thập được có thể sẽ không hữu ích hoặc không liên quan đến mục tiêu phân tích.

Dữ liệu có thể chứa các bản ghi trùng lặp hoặc khoảng trống.

Cần phải được làm sạch trước khi phân tích để kết quả được gần nhất với mong đợi.

* **Data Analysis:** phân tích dữ liệu

Sử dụng công cụ và phần mềm để phân tích, giải thích và rút ra kết luận.

* **Data Interpretation:** giải thích dữ liệu

Chọn cách diễn đạt kết quả bằng biểu đồ hoặc từ ngữ

Sau đó dựa vào kết quả để quyết định hướng hành động

* **Data Visualization:** Trực quan hoá dữ liệu

Dữ liệu được hiển thị dưới dạng đồ thị hoặc biểu đồ.

Thường được sử dụng để thể hiện các xu hướng chưa biết bằng cách quan sát so sánh các bộ dữ liệu để tìm ra thông tin có ý nghĩa.

* 1. Công cụ sử dụng
     1. Công nghệ sử dụng

Google Colaboratory



Hình 1.2.1: Logo Google Colaboratory

* + 1. Ngôn ngữ Python:

Python là một ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng, do Guido van Rossum tạo ra và lần đầu ra mắt vào năm 1991. Python được thiết kế với ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ.

* + 1. Thư viện Pandas

Là một thư viện mã nguồn mở, hỗ trợ đắc lực trong thao tác dữ liệu. Đây cũng là bộ công cụ phân tích và xử lý dữ liệu mạnh mẽ của ngôn ngữ lập trình python. Thư viện này được sử dụng rộng rãi trong cả nghiên cứu lẫn phát triển các ứng dụng về khoa học dữ liệu. Thư viện này sử dụng một cấu trúc dữ liệu riêng là Dataframe. Pandas cung cấp rất nhiều chức năng xử lý và làm việc trên cấu trúc dữ liệu.

* + 1. Thư viện Numpy

Là một thư viện phục vụ cho khoa học máy tính của Python, hỗ trợ cho việc tính toán các mảng nhiều chiều, có kích thước lớn với các hàm đã được tối ưu.

* + 1. Thư viện Matplotlib

Là một thư viện vẽ đồ thị để trực quan hóa dữ liệu rất mạnh mẽ hữu ích cho người sử dụng Python. Dùng để tạo các đồ thị 2D từ dữ liệu trong mảng. Module được sử dụng nhiều nhất của Matplotlib là Pylot cung cấp giao diện như Matlab nhưng thay vào đó sử dụng Python và là mã nguồn mở.

* 1. Cơ hội nghề nghiệp
     1. Vị trí việc làm của phân tích dữ liệu

Chuyên viên phân tích dữ liệu

* + 1. Mô tả vị trí việc làm
* Vị trí công việc phổ biến là chuyên viên phân tích dữ liệu trong các lĩnh vực như tài chính, tiếp thị, bảo hiểm, truyền thông, chăm sóc sức khỏe, ...
* Công việc có thể làm sau khi ra trường như: thu thập, quản lý, trích xuất, phân tích hay lọc dữ liệu theo yêu cầu của các bên liên quan.
* Các ngành nghề tuyển tụng Data Analyts: ngân hàng, tài chính, viễn thông, mạng xã hội, giáo dục, y tế, sản xuất.
  + 1. Các kĩ năng cần có của chuyên viên phân tích dữ liệu

Có 8 kỹ năng chính:

1. **Làm sạch và chuẩn bị dữ liệu**

Đây là công việc chiếm 80% công việc của các chuyên gia dữ liệu. Đây là một kỹ năng rất quan trọng của chuyên viên phân tích dữ liệu.

1. **Phân tích và thăm dò dữ liệu**

Phân tích là lấy câu hỏi hoặc nhu cầu kinh doanh sau đó cần chuyển đổi và phân tích dữ liệu để rút ra câu trả lời đó.

Thăm dò là tìm kiếm các xu hướng hoặc mối quan hệ trong dữ liệu.

1. **Kiến thức thống kê**

Kiến thức thống kê sẽ giúp nhà phân tích dữ liệu định hướng phân tích, nâng cao khả năng khám phá và hiểu sâu hơn về dữ liệu đang làm việc.

1. **Tạo trực quan hóa dữ liệu**

Làm cho các xu hướng và mẫu trong dữ liệu dễ hiểu hơn đối với người nghe. Cần có khả năng tạo các cốt truyện, biểu đồ để giúp truyền đạt dữ liệu và phát hiện của mình trực quan hơn.

1. **Tạo Dashboards hoặc báo cáo**

Xây dựng các Dashboards, báo cáo sẽ giúp cho những người trong công ty (hoặc tổ chức) dễ hiểu về dữ liệu hơn để thuận tiện đưa ra các quyết định quan trọng.

1. **Kỹ năng truyền tải thông tin**

Các kỹ năng quan trọng: nghe, nói, viết, giải thích. Kỹ năng giao tiếp tốt sẽ giúp bạn làm việc hiểu quả với đồng nghiệp của mình. Cũng như dễ dàng truyền tải những kết quả mà bạn đã làm ra đến với mọi người.

1. **Kiến thức nghiệp vụ**

Đây là kiến thức về ngành mà bạn cần phân tích dữ liệu. Ví dụ như muốn phân tích về bệnh tiểu đường thì nhà phân tích dữ liệu cần có những kiến thức cơ bản về bệnh đó, hay về kinh doanh quần áo thì cần có kiến thức về thời trang, xu thế hiện nay. Có kiến thức về ngành đang phân tích sẽ giúp làm việc hiểu quả hơn với dữ liệu.

1. **Giải quyết vấn đề**

Xảy ra lỗi hay rào cản là vấn đề thường xuyên với phân tích dữ liệu. Vì vậy đây là kỹ năng rất quan trọng đối với ngành này.

* + 1. Mức lương

Ít kinh nghiệm hoặc mới ra trường: 6 – 10 triệu

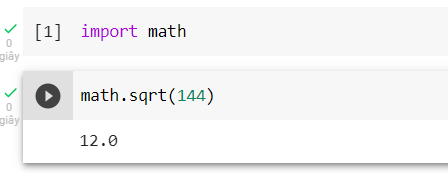
Nhiều kinh nghiệm, lâu năm: 15 – 24 triệu

CHƯƠNG 2: TỔNG QUAN CƠ SỞ LÝ THUYẾT

2.1. Python cơ bản

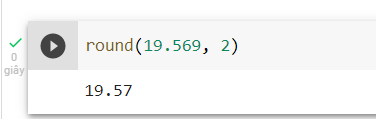
2.1.1. Số học Python

Sử dụng math. sqrt () để lấy căn bậc 2 của một số



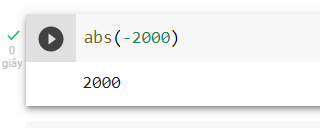
Hình 2.1.1.1: Code lấy căn bậc 2

Sử dụng round () để làm tròn số thập phân



Hình 2.1.1.2: Code làm tròn số thập phân

Sử dụng abs () để lấy giá trị tuyệt đối

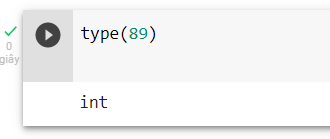


Hình 2.1.1.3: Code lấy giá trị tuyệt đối

2.1.2. Các kiểu dữ liệu cơ bản

**Integers (số nguyên)**

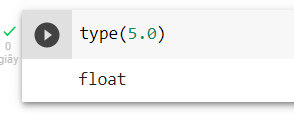
Ví dụ: 1; -1; 89; -8



Hình 2.1.2.1: Code kiểm tra kiểu dữ liệu

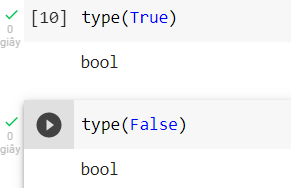
**Float**

Ví dụ: 1.0; 0.56; 5.0



Hình 2.1.2.2: Code kiểm tra kiểu Float

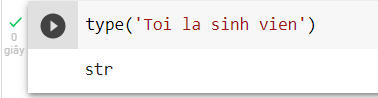
**Booleans: True or False**



Hình 2.1.2.3: Code kiểm tra kiểu dữ liệu Bool

**Strings: Dữ liệu văn bản được đặt trong dấu ngoặc kép hoặc đơn**

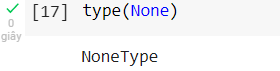
Ví dụ: ‘Toi la sinh vien’



Hình 2.1.2.4: Code kiểm tra kiểu dữ liệu strings

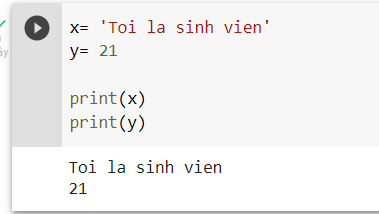
**None: biểu thị giá trị bị thiếu**

Ví dụ:



Hình 2.1.2.5: Code kiểm tra giá trị bị thiếu

2.1.3. Biến trong Python



Hình 2.1.3.1: Khai báo biến

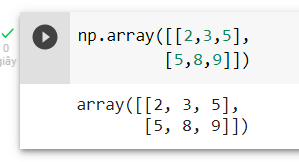
2.1.4. Danh sách trong Python



Hình 2.1.4.1: Lists

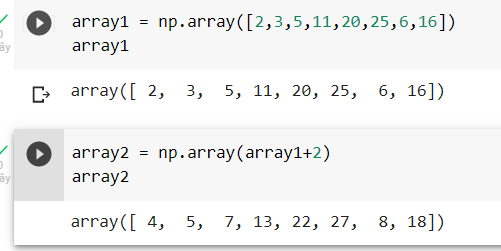
2.1.5. Mảng

Ví dụ:



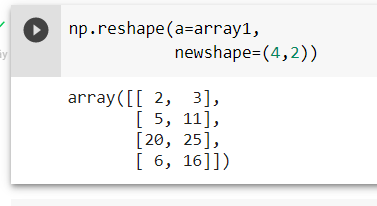
Hình 2.1.5.1: Mảng

Thực hiện phép toán cộng vào mảng



Hình 2.1.5.2: Thực hiện phép toán cộng với mảng

Sắp xếp lại mảng dùng reshape () (4 hàng và 2 cột)



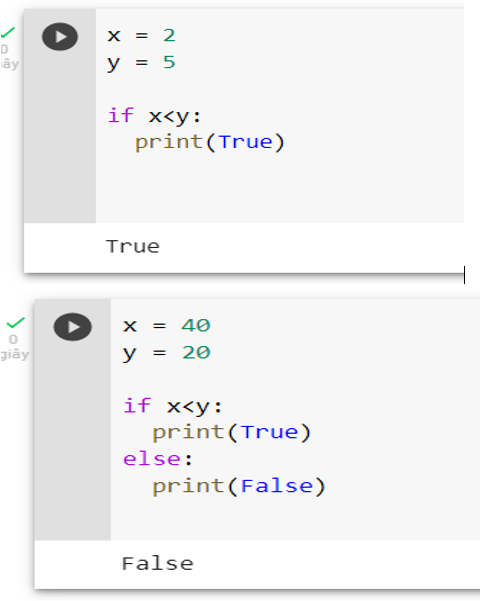
Hình 2.1.5.3: Sắp xếp lại mảng

2.1.6. Câu điều kiện IF

If<biểu thức>:

Lệnh

Ví dụ

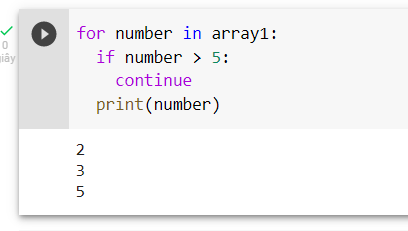


Hình 2.1.6.1: Lệnh IF

2.1.7. Vòng lặp

Vòng lặp FOR

Ví dụ: Loại những số nào lớn hơn 5



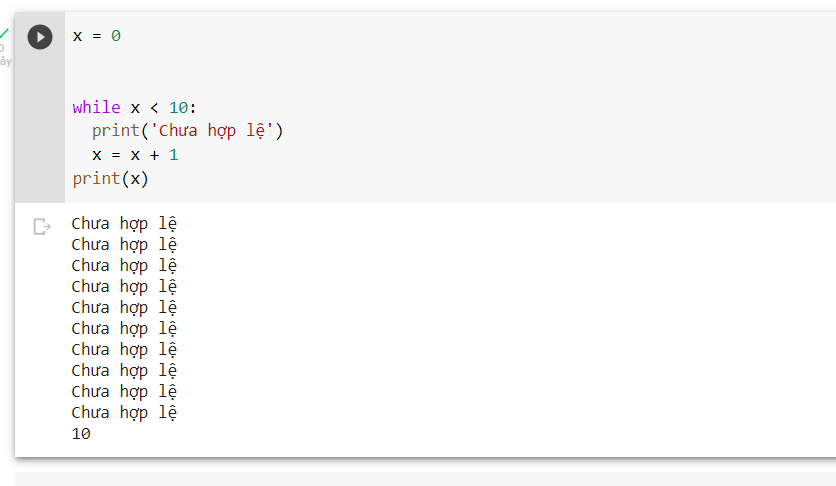
Hình 2.1.7.1: Vòng lặp For

Vòng lặp While

While <Biểu thức>:

Lệnh

Ví dụ:



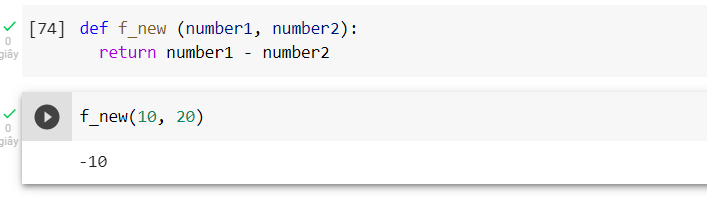
Hình 2.1.7.2: Vòng lặp While

2.1.8. Hàm

Def <tên hàm> (các tham số/ đối số)

Các câu lệnh

Ví dụ:



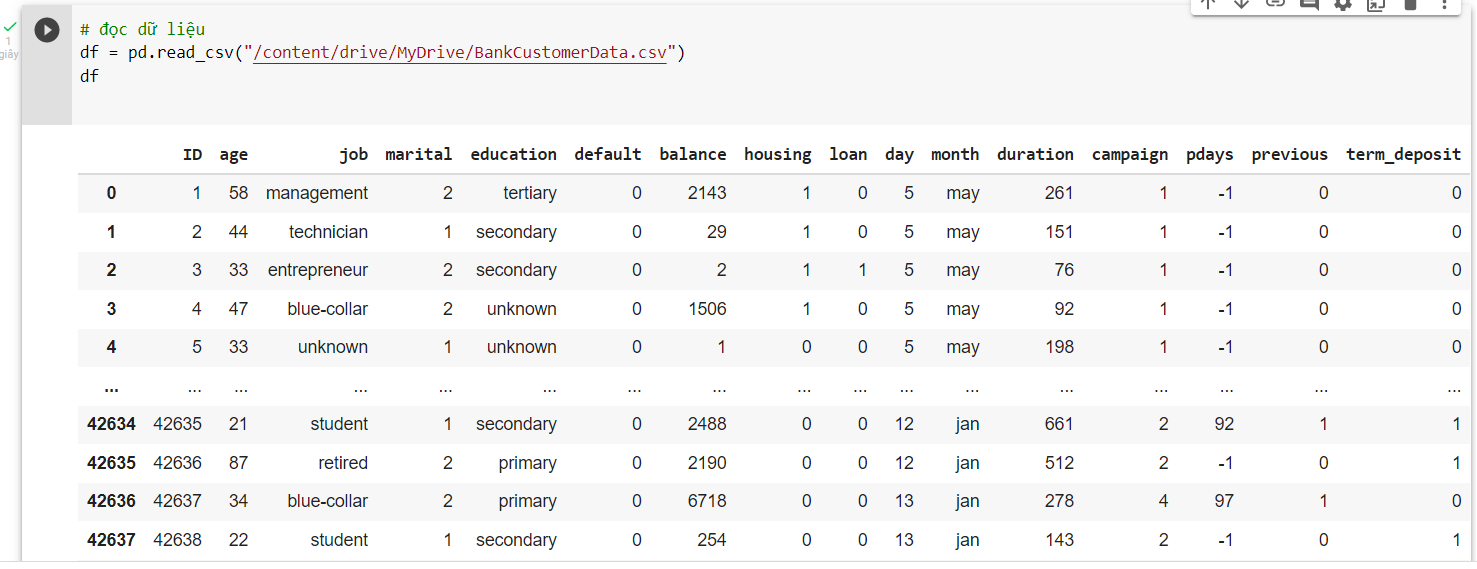
Hình 2.1.8.1: Hàm

2.2. Pandas

* **Đọc dữ liệu với Pandas**

Cho phép đọc/ ghi dữ liệu giữa bộ nhớ với nhiều định dạng file: csv, text, excel … bằng cách sử dụng hàm (pd. read\_csv ())

Ví dụ đọc dữ liệu trên Google Driver:



Hình 2.2.1: Đọc dữ liệu

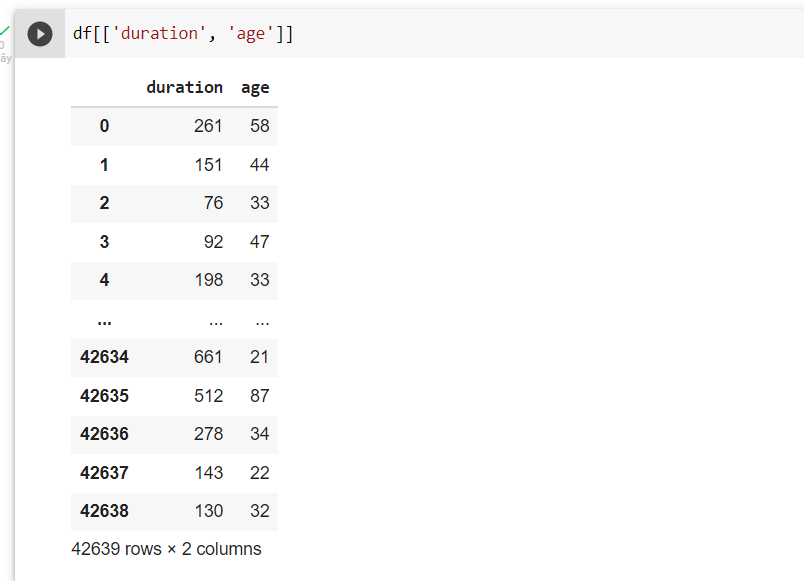
Truy xuất dữ liệu

* **Truy xuất dữ liệu theo tên cột: df [‘tên cột’]**



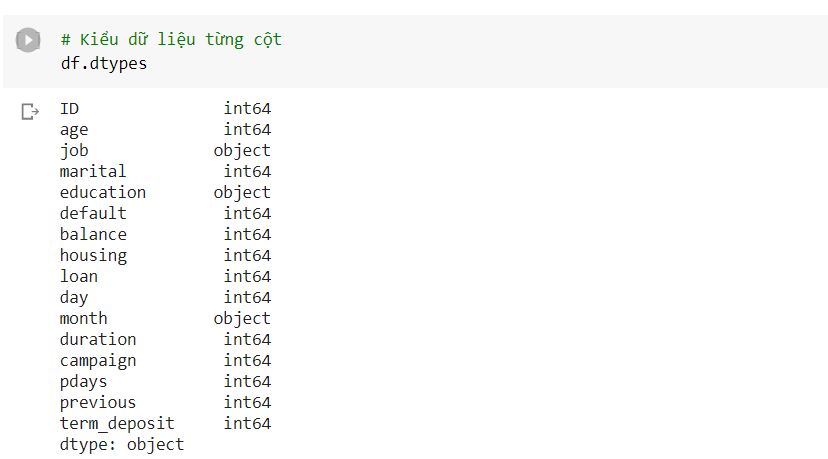
Hình 2.2.2: Code truy xuất dữ liệu

* **Truy xuất theo nhiều cột: df. [[‘tên cột 1’, ‘tên cột 2’]]**



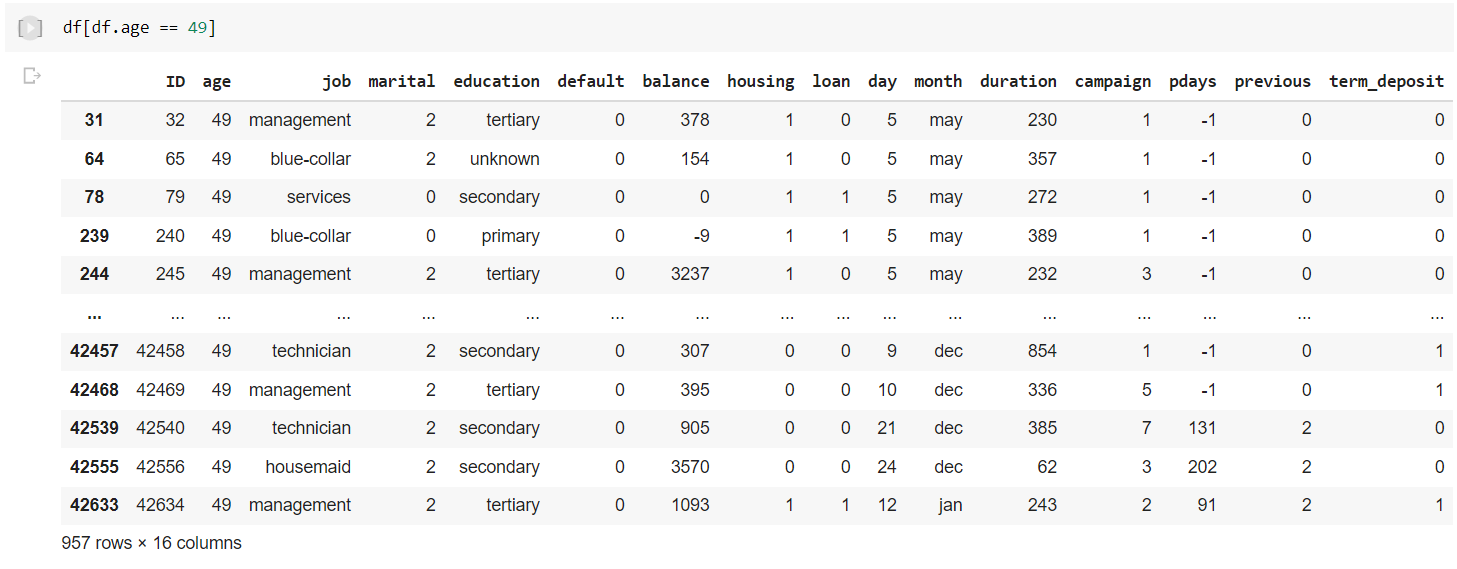
Hình 2.2.3: Code truy xuất dữ liệu theo điều kiện

* **Xem kiểu dữ liệu từng cột: df. dtype ()**



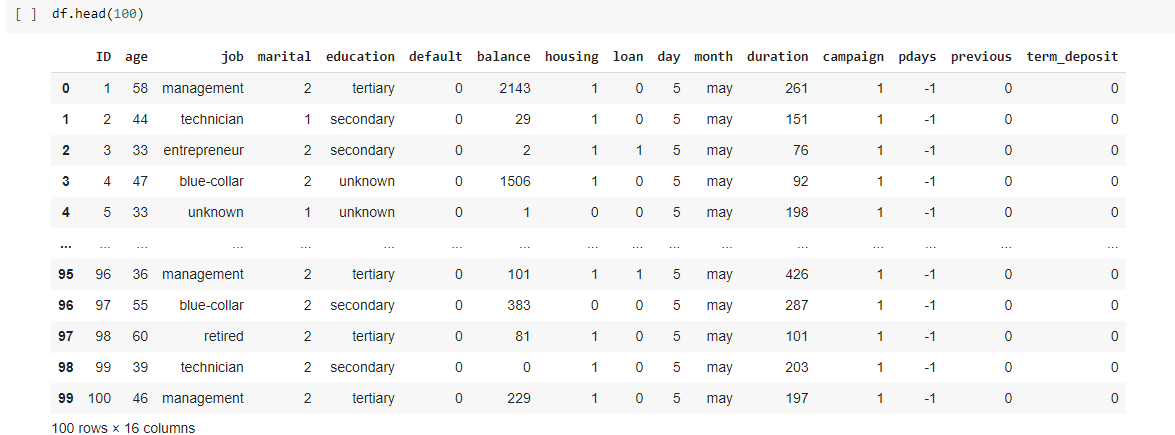
Hình 2.2.4: Code kiểu dữ liệu từng cột

* **Lấy dữ liệu theo điều kiện (Khách hàng có tuổi bằng 49)**



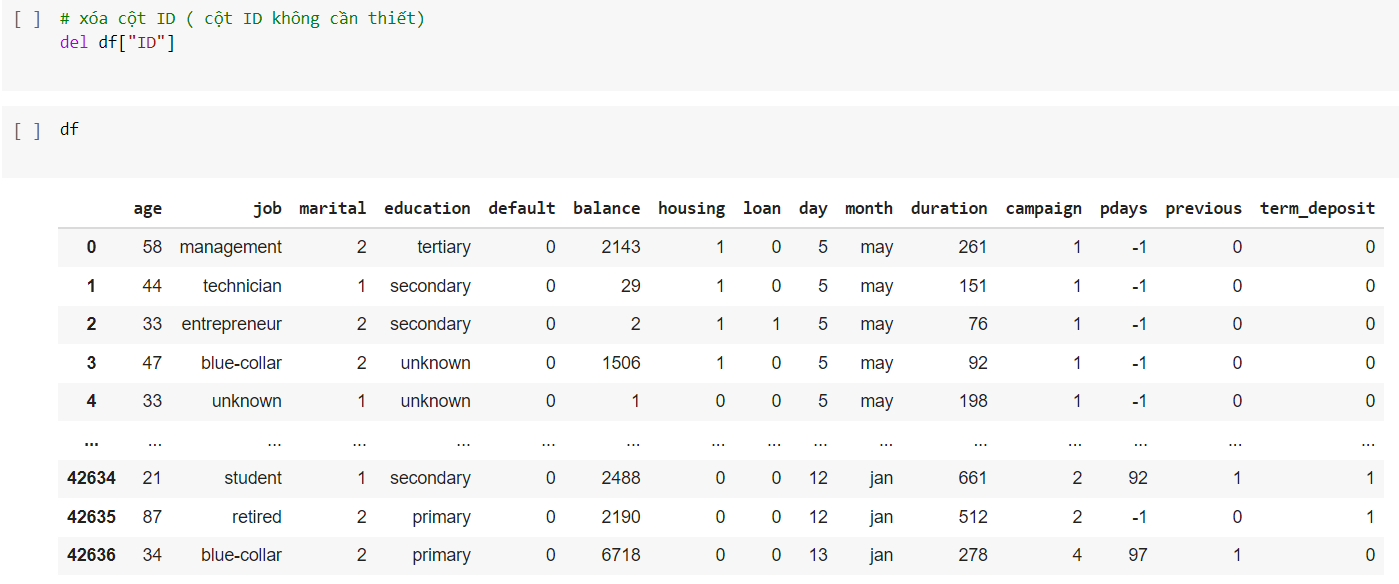
Hình 2.2.5: Code lấy dữ liệu theo điều kiện

* **Lấy 100 dòng đầu tiên của dữ liệu**



Hình 2.2.6: Code lấy 100 dòng dữ liệu đầu tiên

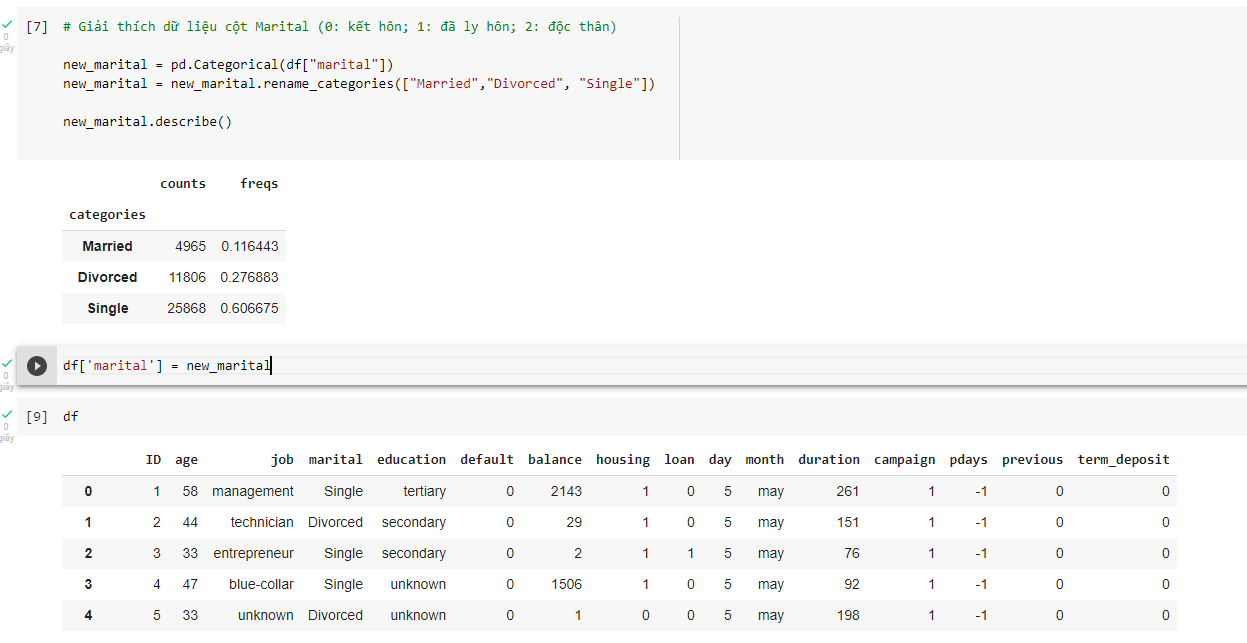
* **Xóa cột**



Hình 2.2.7: Code xóa cột

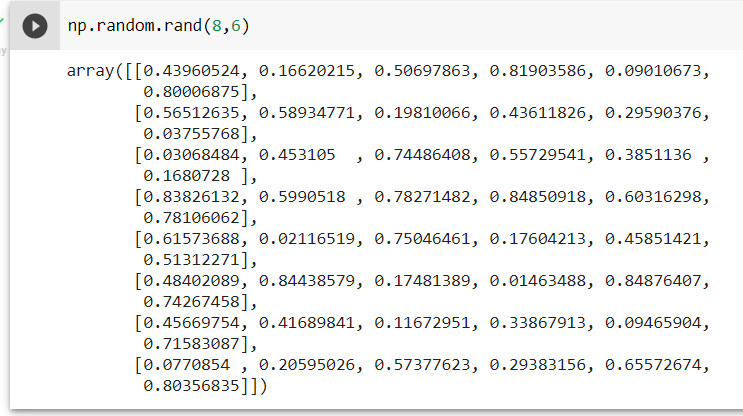
* **Sửa giá trị cột**

**Df [‘tên cột’] = tên cột mới**



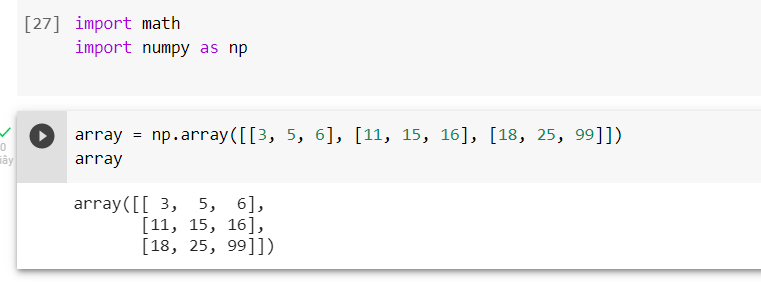
Hình 2.2.8: Code sửa giá trị cột

2.3. Numpy

**Khởi tạo 1 mảng ngẫu nhiên**

Hình 2.3.1: Code khởi tạo mảng ngẫu nhiên

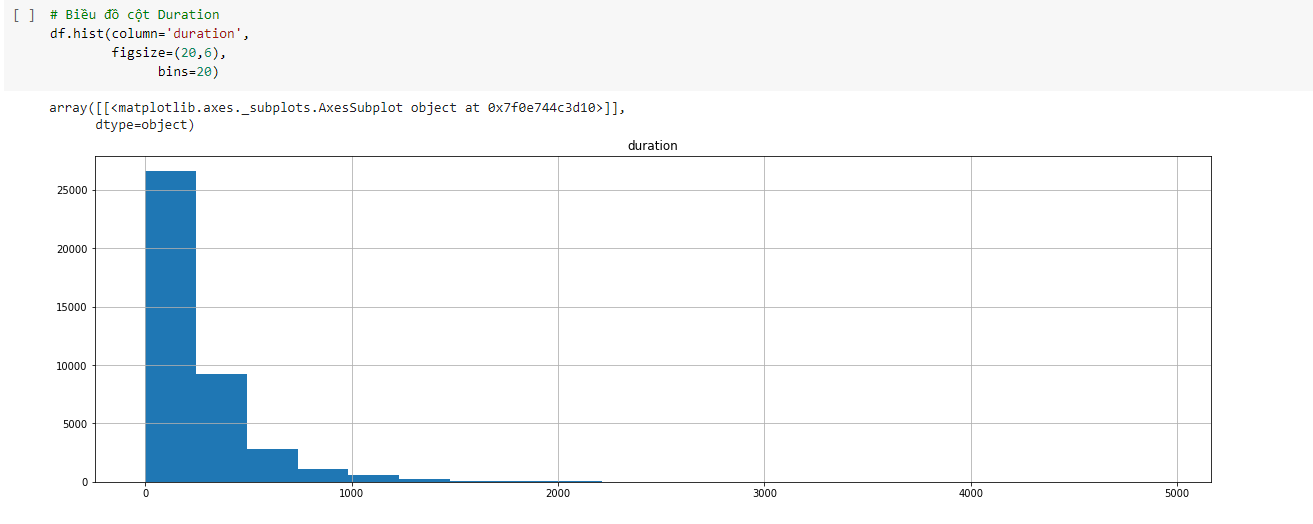
**Tạo một mảng cụ thể**



Hình 2.3.2: Code tạo mảng cụ thể

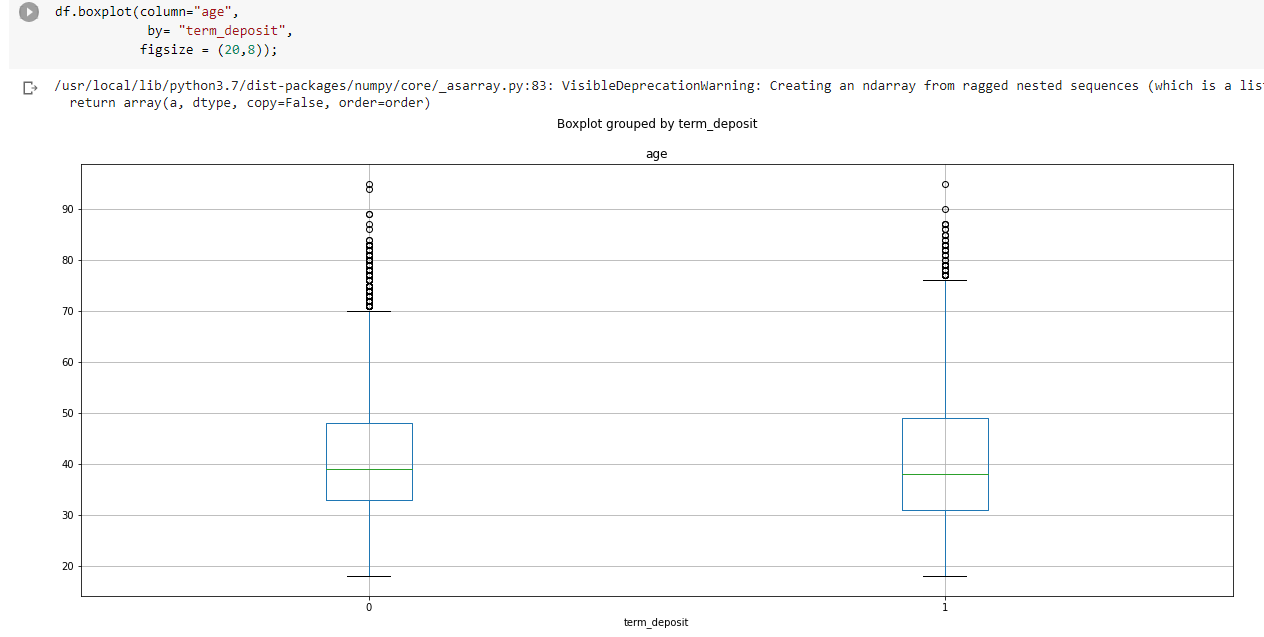
2.4. Matplotlib

2.4.1. Biểu đồ cột (histogram)



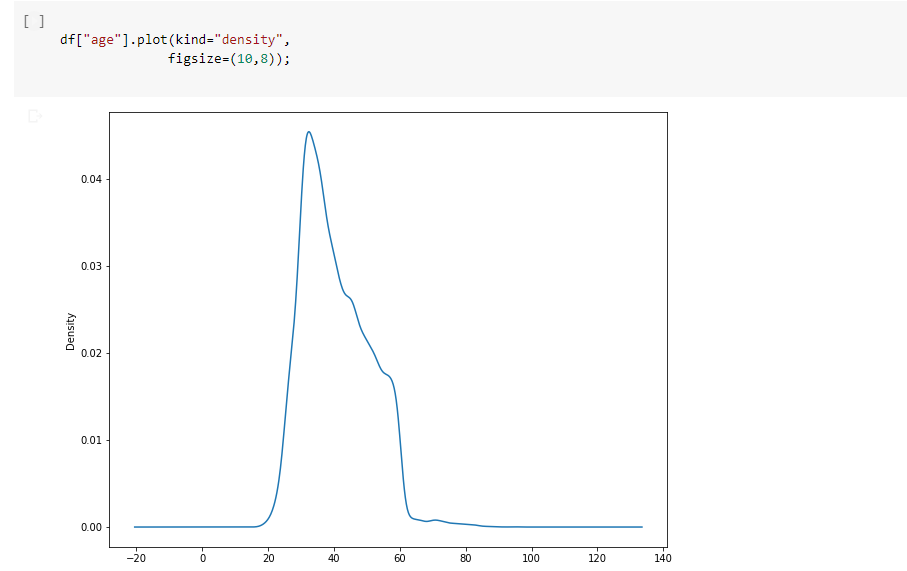
Hình 2.4.1: Biểu đồ Histogram

2.4.2. Biểu đồ Boxplot



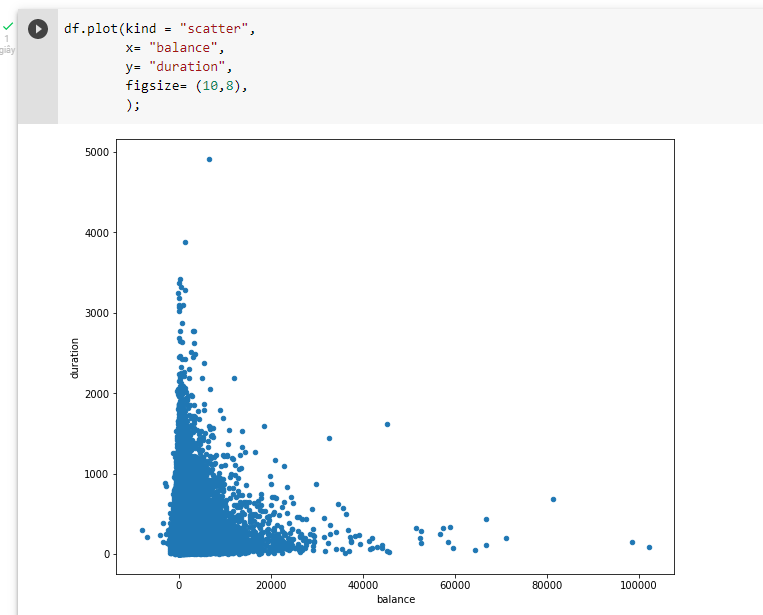
Hình 2.4.2: Biểu đồ Boxplot

2.4.3. Biểu đồ Density



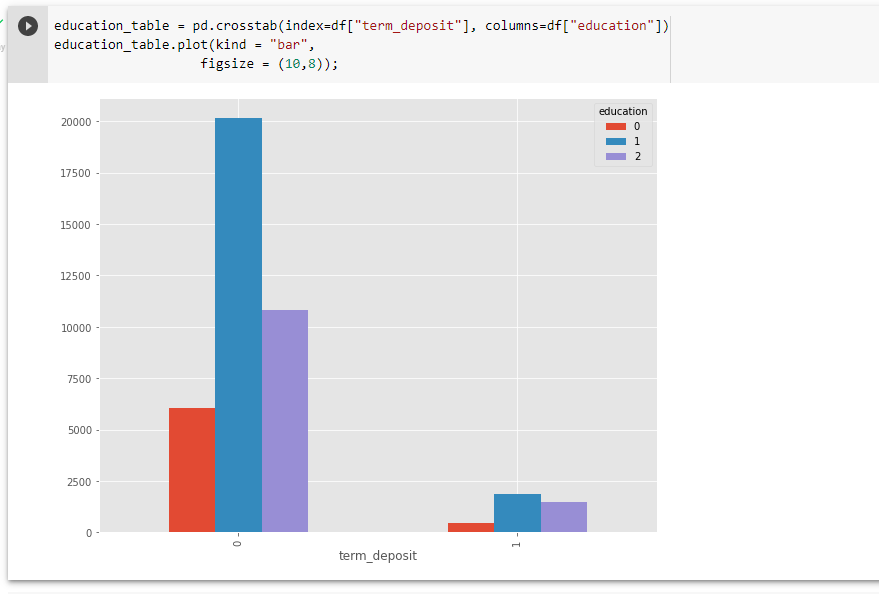
Hình 2.4.3: Biểu đồ Density

2.4.4. Biểu đồ Scatter



Hình 2.4.4: Biểu đồ Sactter

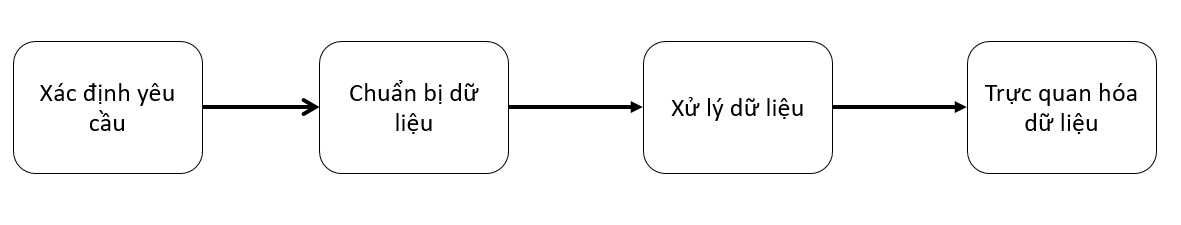
2.4.5. Biểu đồ Bar



Hình 2.4.5: Biểu đồ Bar

CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU

* 1. Nhiệm vụ về đề tài
  + Tiến hành xử lý và làm sạch dữ liệu.
  + Trực quan hóa dữ liệu.
  1. Quy trình thực hiện



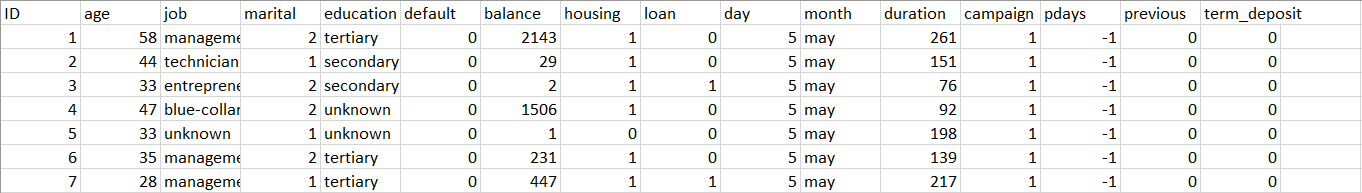
Hình 3.2: Quy trình thực hiện

* 1. Dữ liệu
     1. Giới thiệu dữ liệu

Dữ liệu về khách hàng tại ngân hàng gồm có 16 cột, 42639 dòng và có kích thước 2,555 KB. Dữ liệu được lưu trữ dưới dạng CSV.

Dữ liệu có liên quan đến các chiến dịch tiếp thị trực tiếp của một tổ chức ngân hàng. Các chiến dịch tiếp thị dựa trên các cuộc gọi điện thoại, để truy cập xem sản phẩm (tiền gửi có kỳ hạn ngân hàng) sẽ được (hoặc không) đăng ký.

Các trường của dữ liệu:



Hình 3.3.1: Dữ liệu

* + 1. Giải thích các trường của dữ liệu

|  |  |
| --- | --- |
| ID | Số thứ tự của khách hàng |
| Age | Độ tuổi của khách hàng |
| Job | Nghề nghiệp |
| Marital | Tình trạng hôn nhân (0: đã kết hôn, 1: đã ly hôn, 2: độc thân |
| Education | Trình độ giáo dục |
| Default | Vỡ nợ tín dụng (0: không có; 1: có) |
| Balance | Số dư trung bình hằng năm (EUR) |
| Housing | Khoản vay mua nhà (0: không có, 1: có) |
| Loan | Khoản vay cá nhân (0: không có, 1: có) |
| Day | Ngày liên hệ cuối cùng của ngân hàng đến khách hàng (liên hệ bằng điện thoại) |
| Month | Tháng liên hệ cuối cùng của ngân hàng đến khách hàng (liên hệ bằng điện thoại) |
| Duration | Thời lượng liên hệ cuối cùng, tính bằng giây |
| Campaign | Số liên hệ được thực hiện trong chiến dịch này |
| Pdays | Số ngày trôi qua kể từ lần liên hệ cuối cùng của chiến dịch trước (-1: khách hàng chưa được liên hệ trước đó) |
| Previous | Số liên hệ thực hiện trước chiến dịch này |
| Term\_deposit | Đăng ký tiền gửi có kỳ hạn (0: không có, 1: có) |

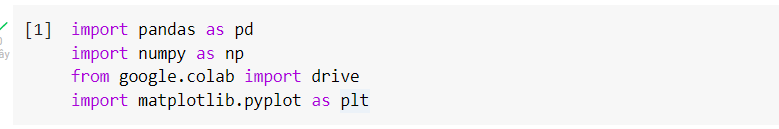
Bảng 3.3.2: Giải thích các trường của dữ liệu

* + 1. Nhiệm vụ

Dựa vào các trường dữ liệu hãy tiến hành trực quan hóa dữ liệu. Từ đó đưa ra nhận xét trường dữ liệu Term\_deposit (đăng ký tiền gửi có kỳ hạn) có mối tương quan như thế nào đối với các trường còn lại.

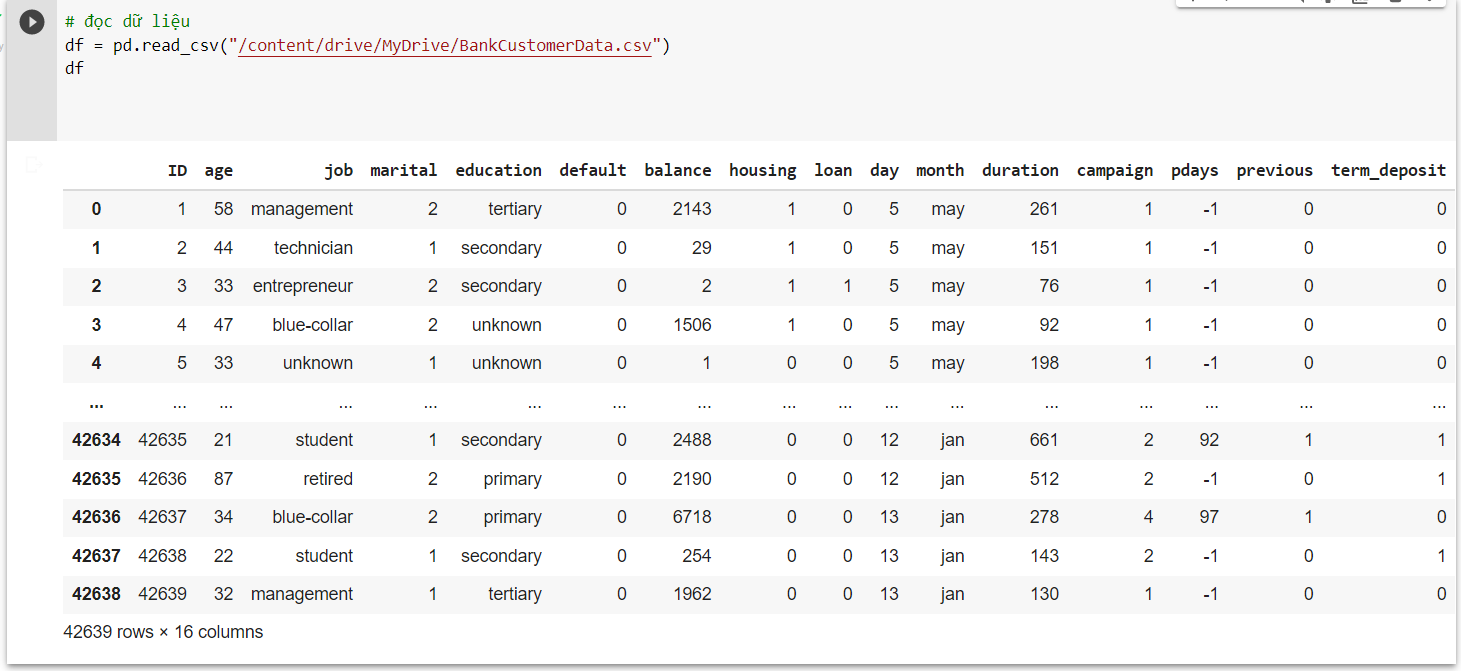
* + 1. Xử lý và làm sạch dữ liệu

**Thêm các thư viện cần thiết vào Google Colaboratory**



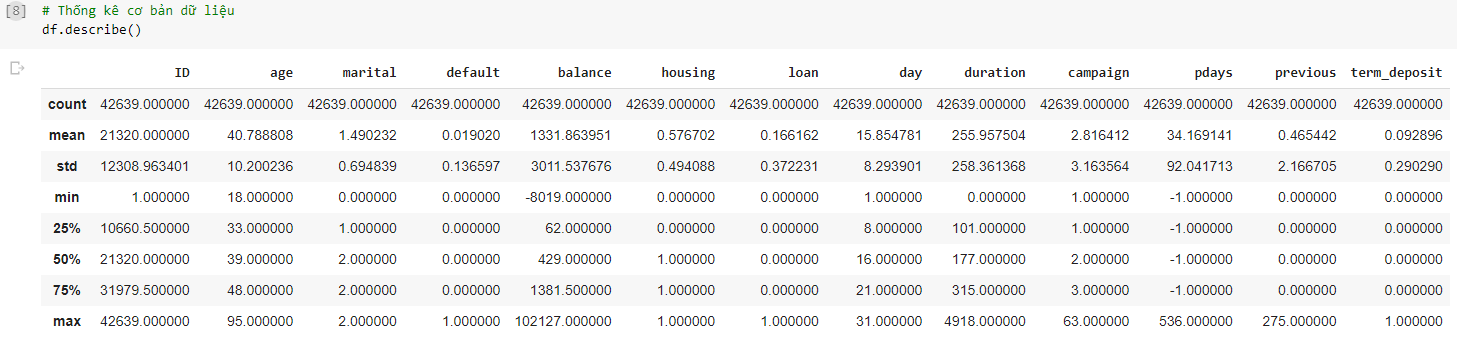
Hình 3.3.4.1: Import các thư viện cần thiết

**Đọc dữ liệu vào Google Colaboratory**



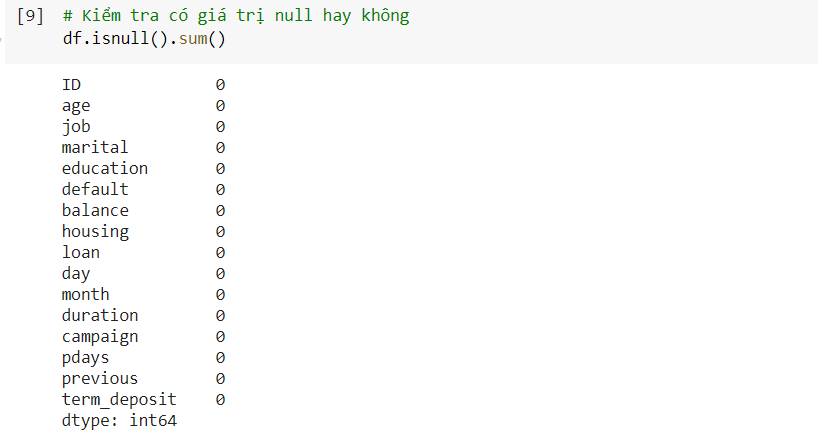
Hình 3.3.4.2: Đọc dữ liệu vào Google Colaboratory

**Xem thống kê cơ bản của các trường dữ liệu (có thể xem được số dòng, số cột, giá trị trung bình, độ lệch chuẩn, phân vị)**



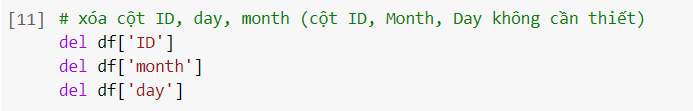
Hình 3.3.4.3: Code xem thống kê cơ bản các trường dữ liệu

**Kiểm tra dữ liệu có giá trị null hay không**



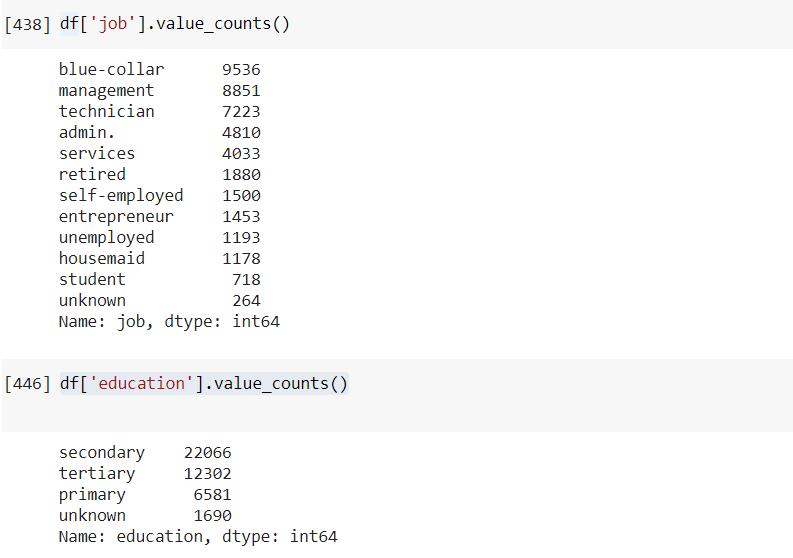
Hình 3.3.4.4: Code kiểm tra dữ liệu

**Xóa các cột không cần thiết**



Hình 3.3.4.5: Code xóa dữ liệu

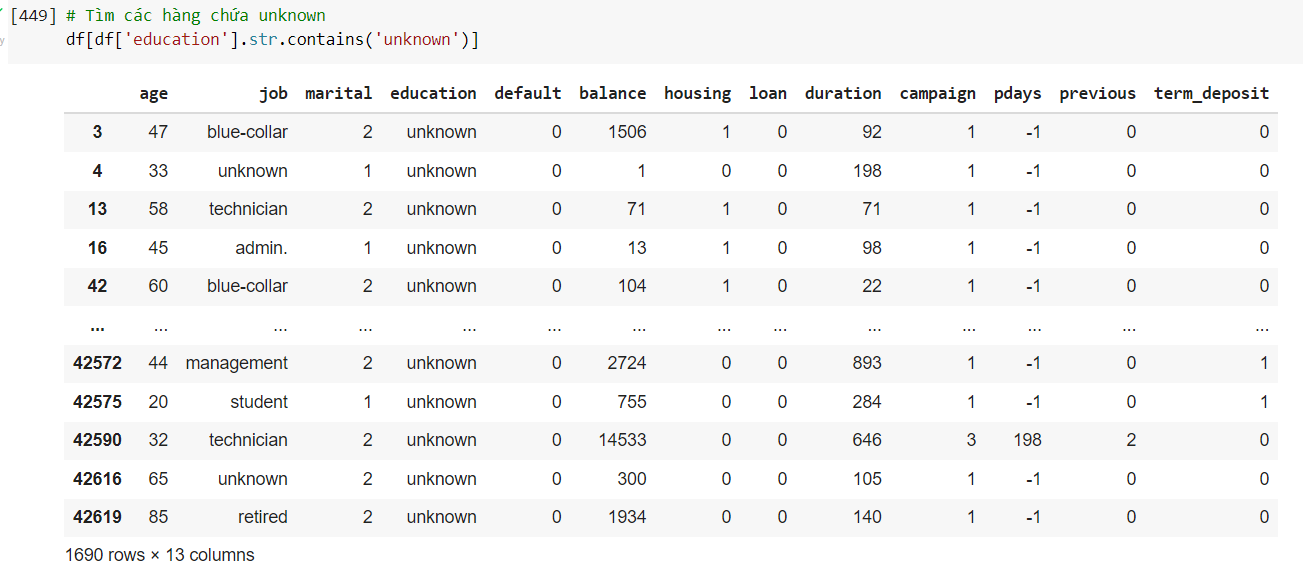
**Sử dụng values\_count () để xem tần suất của từng trường dữ liệu**



Hình 3.3.4.6: Code kiểm tra tần suất dữ liệu

Sau khi kiểm tra phát hiện trường Job và Education chứa giá trị unknown.

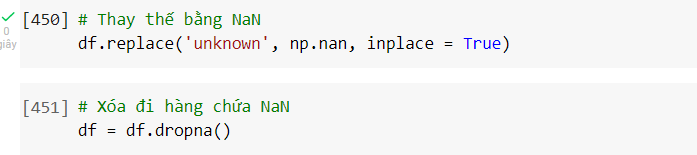
**Tìm các hàng chứa giá trị ‘unknown’ của trường Education**



Hình 3.3.4.7: Code tìm giá trị unknown

Tỉ lệ unknown của trường Education chiếm phần trăm ít so với toàn bộ dữ liệu nên tiến hành xóa để không làm nhiễu thông tin khi trực qua hóa dữ liệu.

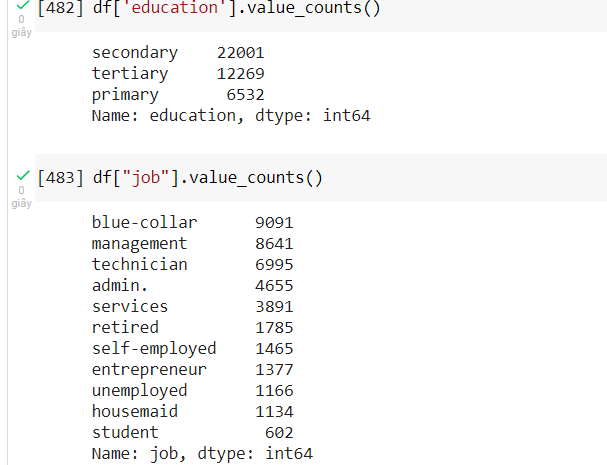
**Xóa dữ liệu chứa unknown của trường ‘Education’**



Hình 3.3.4.8: Code xóa dữ liệu unknown

Thay thế giá trị unknown bằng giá trị NaN, sau đó dùng dropna () để tiến hành xóa

**Kiểm tra lại sau khi xóa**



Hình 3.3.4.9: Code xóa kiểm tra lại dữ liệu

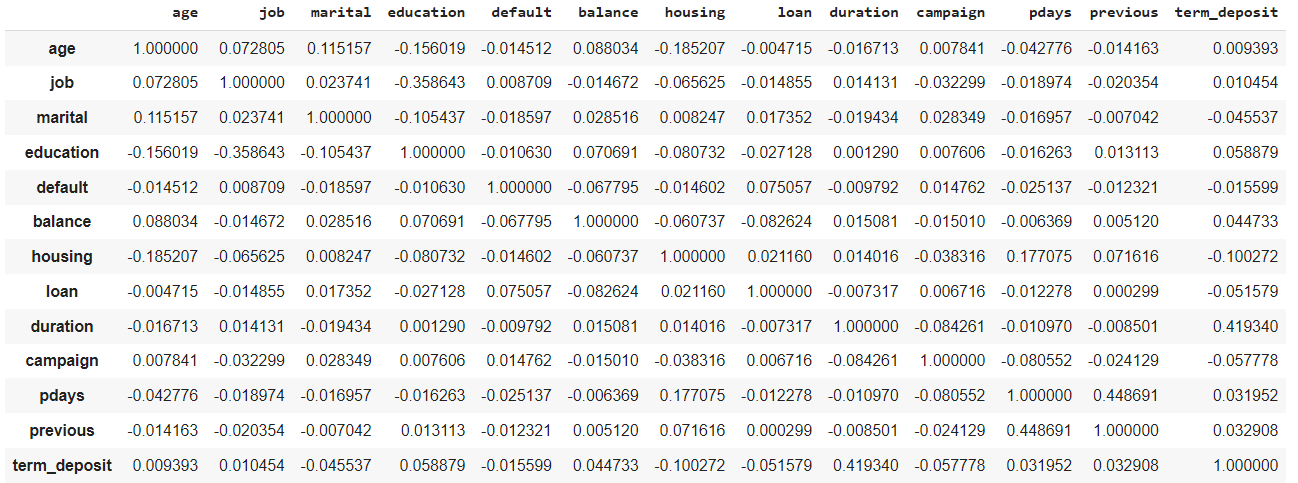
Sau khi xóa giá trị ‘unknown’ trường ‘Education’ thì giá trị ‘unknown’ của cột Job cũng mất. Như vậy chúng ta không cần xóa giá trị ‘unknown’ của trường ‘Job’

**Mã hóa trường ‘Education’ và ‘Job’ để thuận tiện cho việc tính toán**



Hình 3.3.4.9: Code mã hóa các trường cần thiết

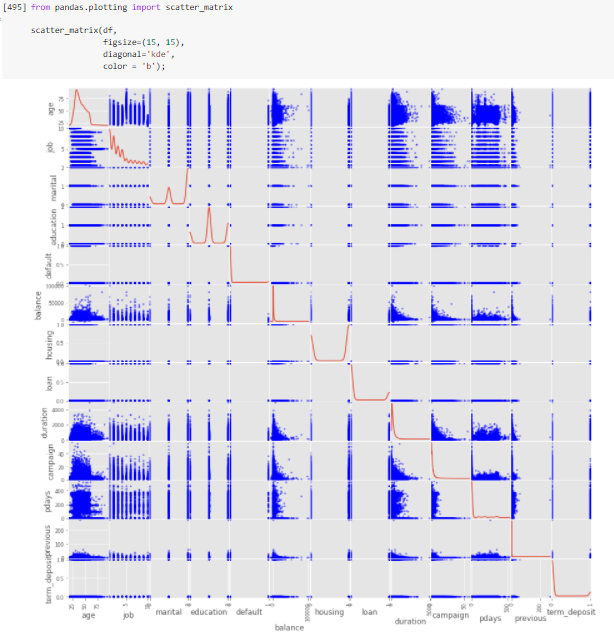
* 1. Trực quan hóa dữ liệu
* **Sử dụng df. corr () dể kiểm tra độ tương quan giữa các cột**



Hình 3.4.1: Code kiểm tra độ tương quan giữa các cột

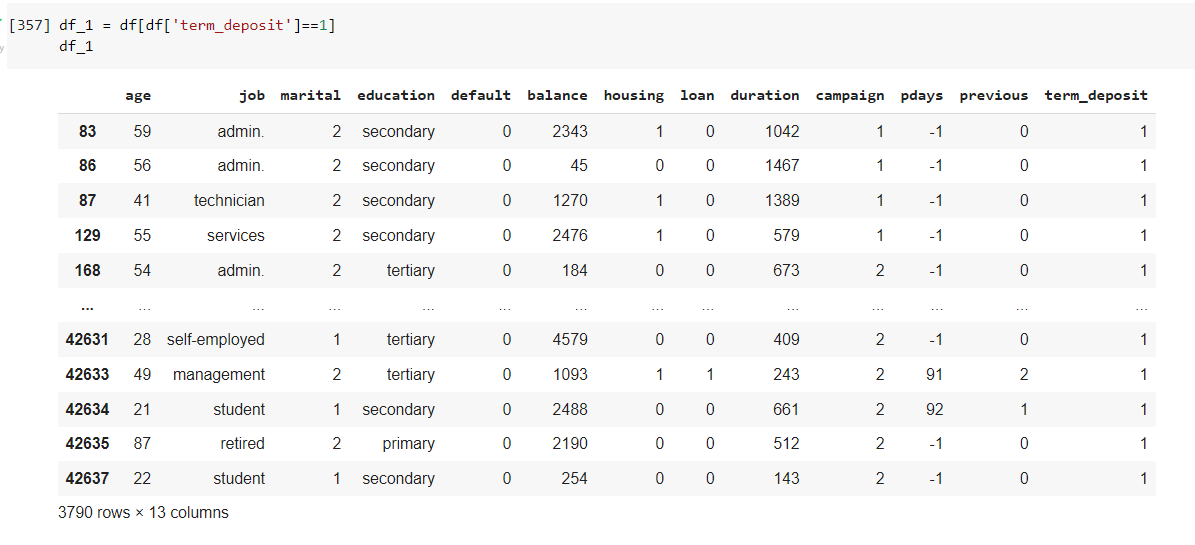
Nhìn vào cột ‘Term\_deposit’ có thể nhìn thấy được độ tương quan giữa các cột với nó. Tương quan dương cao nhất là cột ‘Duration’ với 0.419 sau đó lần lượt đến ‘Education’, ‘Balance’, ‘Previous’, ‘Pdays’, ‘Job’, ‘Age’ và các cột còn lại tương quan âm.

* **Sử dụng biểu đồ Scatter để biểu diễn độ tương quan giữa các cột**



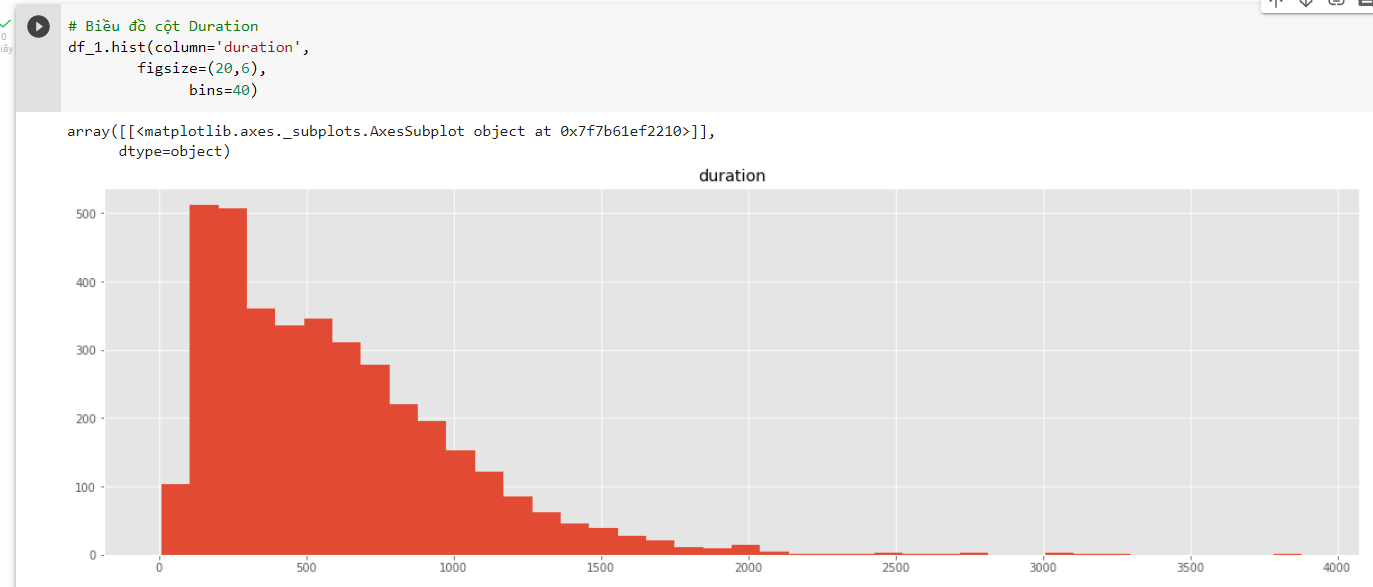
Hình 3.4.2: Biểu đồ scatter hiển thị độ tương quan giữa các cột

* **Lọc dữ liệu theo số người đăng ký tiền gửi có kỳ hạn: df\_1**



Hình 3.4.3: Lọc dữ liệu theo đăng ký tiền gửi có kỳ hạn

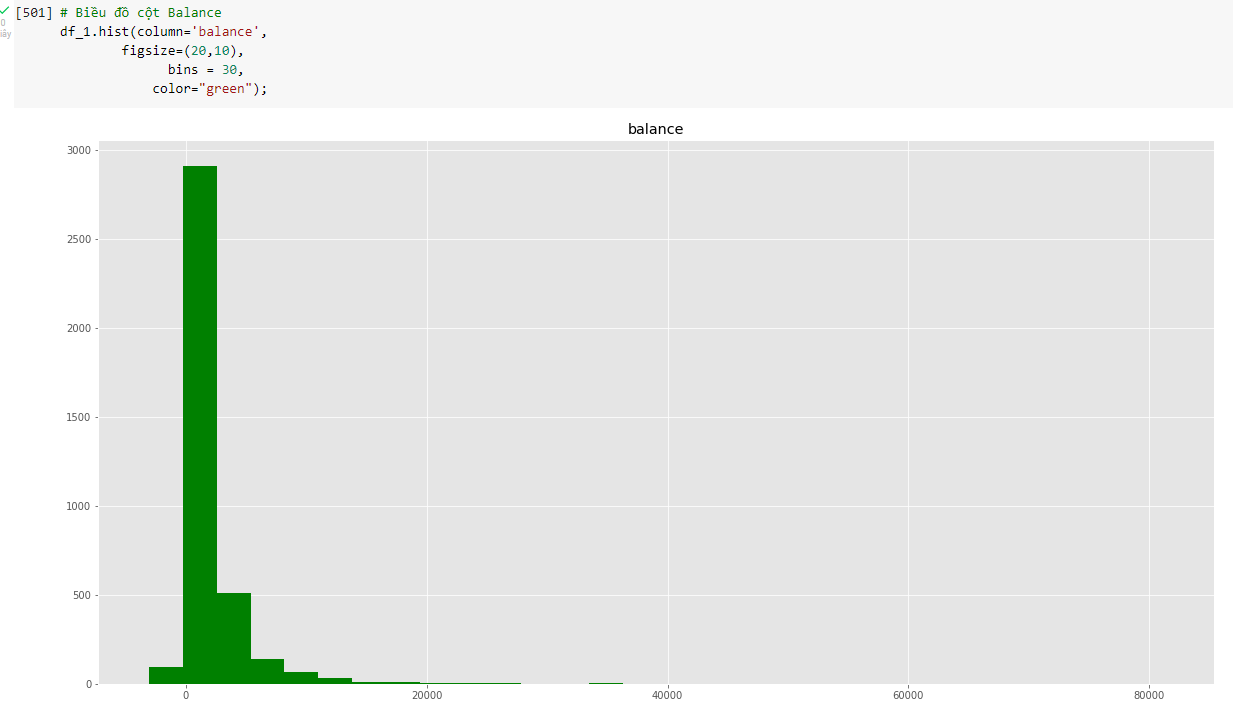
* **Biểu đồ cột hiển thị thời lượng liên hệ cuối cùng**



Hình 3.4.4: Biểu đồ cột về Duration

Giải thích: dựa vào biểu ta có thể thấy thời lượng liên hệ trong khoảng 200-300s thì có tỉ lệ đăng ký tiền gửi ngân hàng cao nhất tiếp đến 300-500s rồi sau đó giảm dần. Như vậy, thời lượng cuộc gọi hiệu quả nhất để gọi cho khách hàng là từ 200-300s

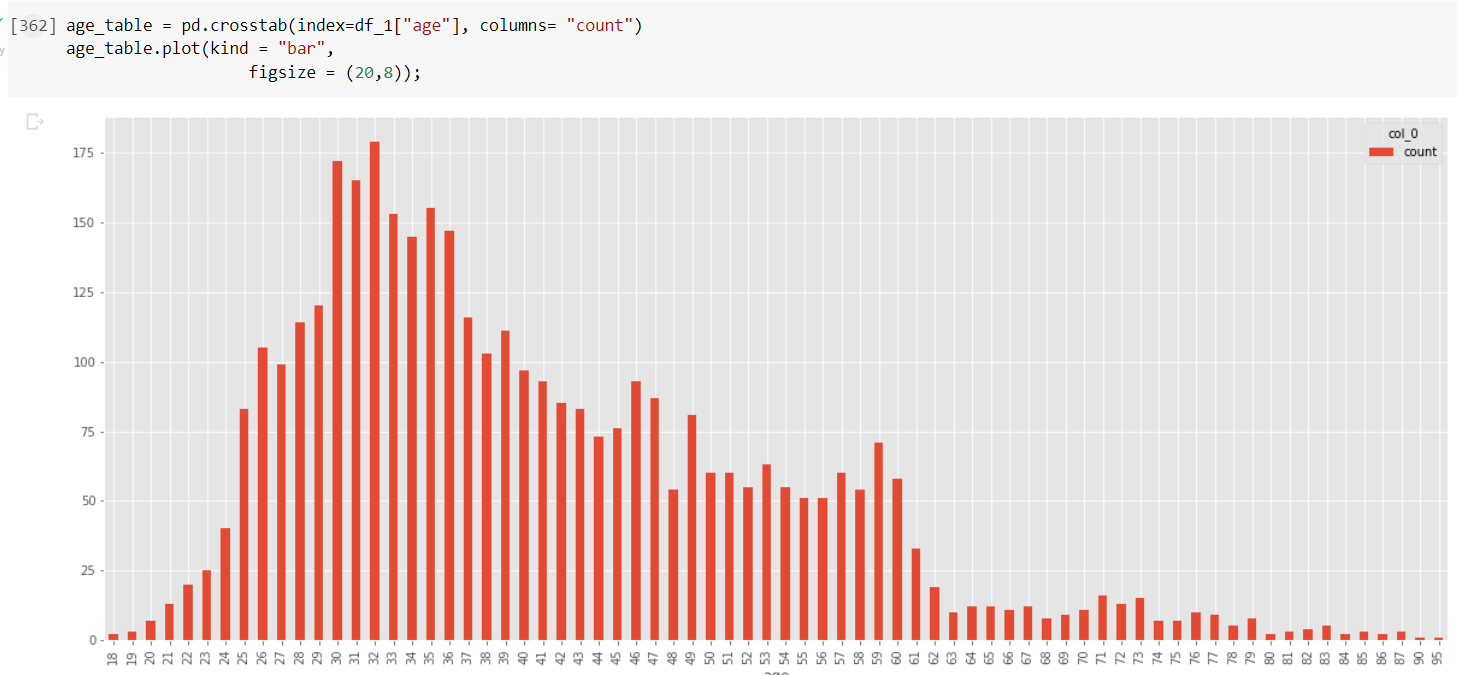
* **Biểu đồ cột hiển thị số dư trung bình hằng năm**



Hình 3.4.5: Biểu đồ cột về Balance

Giải thích: nhìn vào biểu đồ ta có thể thấy khoảng tiền chiếm tỉ lệ cao nhất là 0-2900 EUR, sau đó đến 2900-5800 EUR. Như vậy, khách hàng có số dư trung bình trong khoảng 0-2900 EUR sẽ có tỉ lệ cao đăng ký tiền gửi ngân hàng.

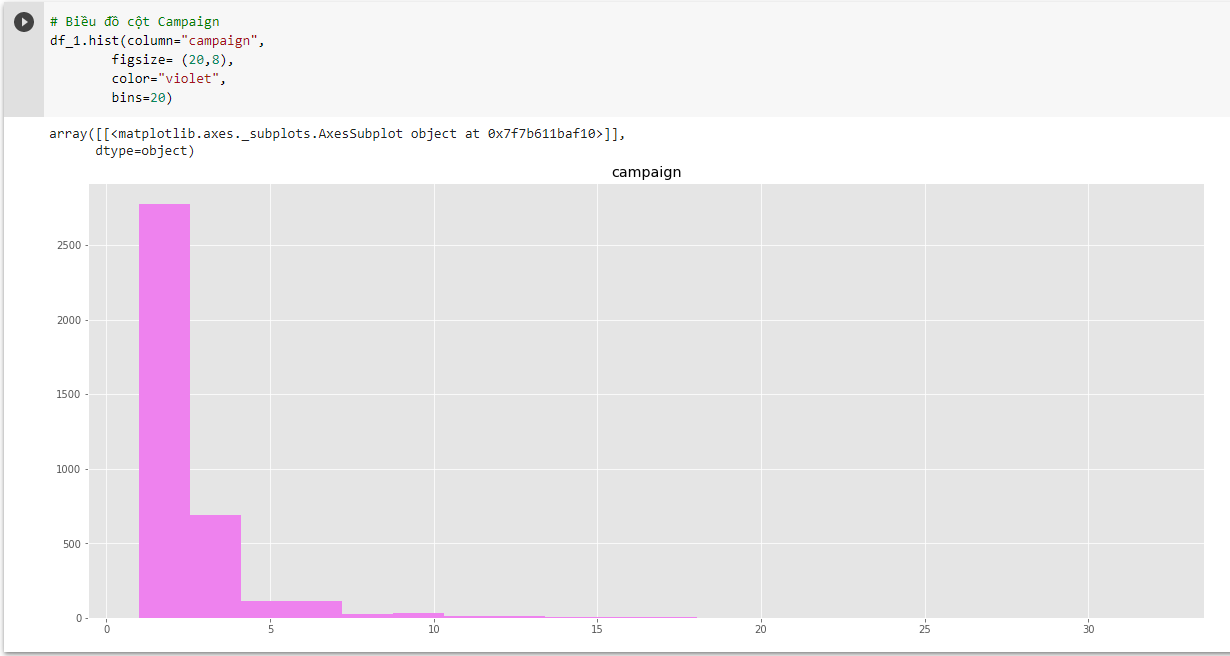
* **Biểu đồ Bar hiển thị độ tuổi**



Hình 3.4.6: Biểu đồ Bar hiển thị độ tuổi người đăng ký tiền gửi có kỳ hạn

Giải thích: nhìn vào biểu đồ ta có thể thấy độ tuổi có tỉ lệ cao là từ 30-36, rồi sau đó đến 37-46. Như vậy độ tuổi có tỉ lệ cao đăng ký tiền gửi có kỳ hạn là từ 30-36.

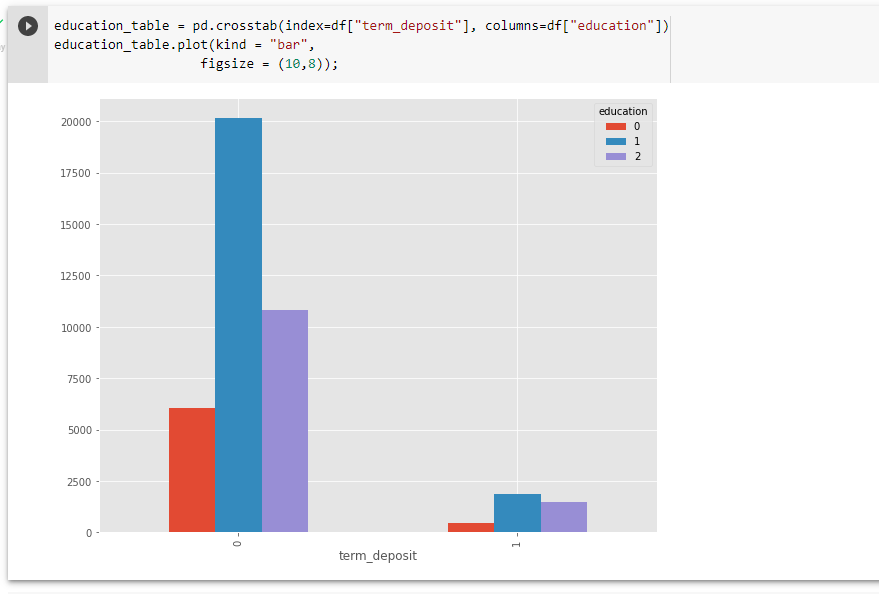
* **Biểu đồ cột hiển thị số liên hệ trong chiến dịch này**



Hình 3.4.7: Biểu đồ cột hiển thị số liên hệ trong chiến dịch này

Giải thích: số lượng liên hệ trong chiến dịch này từ 2-3 lần. Như vậy, với mỗi người ngân hàng liên hệ từ 2-3 lần thì sẽ có tỉ lệ cao là đăng ký tiền gửi có kỳ hạn.

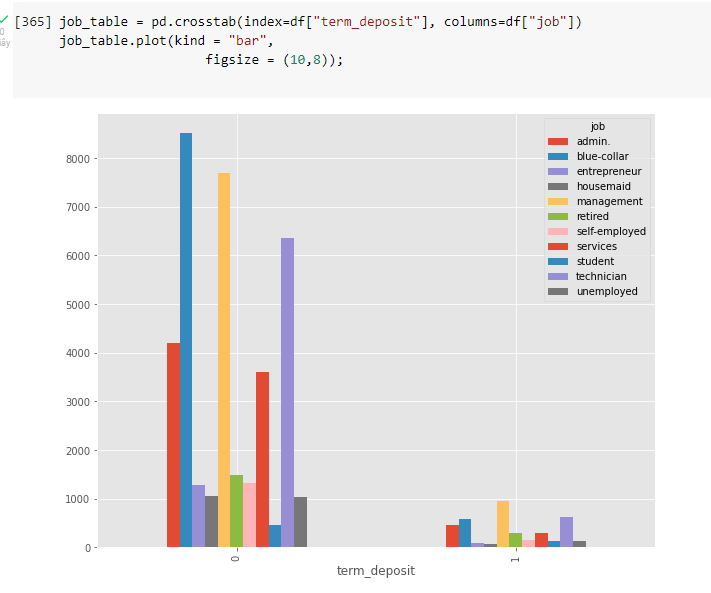
* **Biểu đồ Bar hiển thị trình độ giáo dục**



Hình 3.4.7: Biểu đồ Bar hiển thị trình độ giáo dục

Giải thích: nhìn vào biểu đồ ta có thể thấy trình độ ‘Secondary’ chiếm tỉ lệ cao nhất số lượng người đăng ký tiền gửi có kỳ hạn. Tuy nhiên số lượng người không đăng ký thì tỉ lệ ‘Secondary’ cũng cao nhất nên dựa vào trình độ giáo dục chúng ta khó có thể chọn được trình độ nào sẽ có tỉ lệ cao nhất.

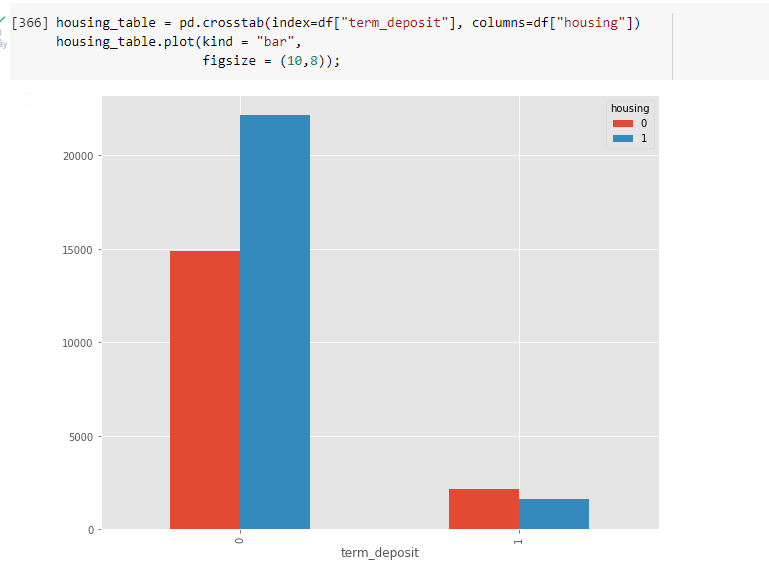
* **Biểu đồ Bar hiển thị nghề nghiệp**



Hình 3.4.7: Biểu đồ Bar hiển thị nghề nghiệp

Giải thích: dựa vào biểu đồ ta có thể thấy nghề chiếm tỉ lệ cáo của đăng ký tiền gửi có kỳ hạn là ‘Management’, sau đó tới ‘Technician’, ‘Blue-collar (công nhân)’. Như vậy, nghề nghiệp có tỉ lệ cao sẽ đăng ký tiền gửi có kỳ hạn là ‘Management’.

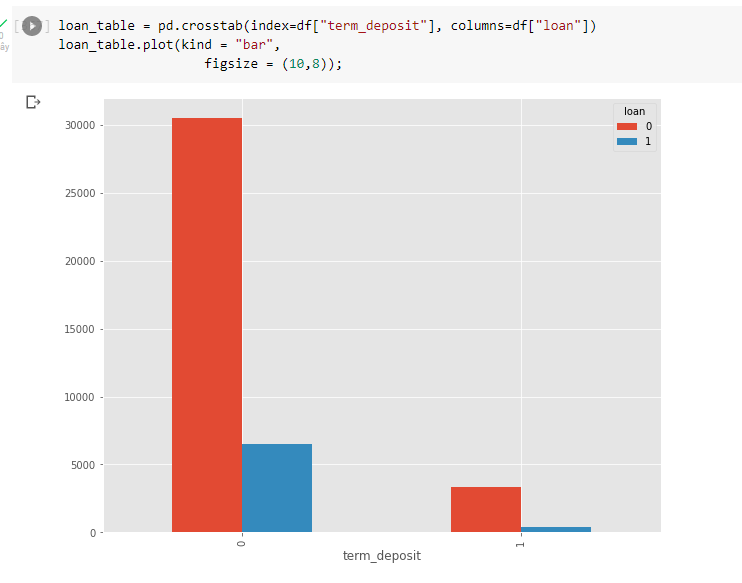
* **Biểu đồ hiển thị khoản vay để mua nhà**



Hình 3.4.8: Biểu đồ Bar hiển thị khoản vay mua nhà

Giải thích: tỉ lệ người không có khoản vay để mua nhà chiếm tỉ lệ cao trong tổng số người đăng ký tiền gửi có kỳ hạn. Như vậy, tỉ lệ cao người không có khoản vay mua nhà sẽ đăng ký tiền gửi có kỳ hạn.

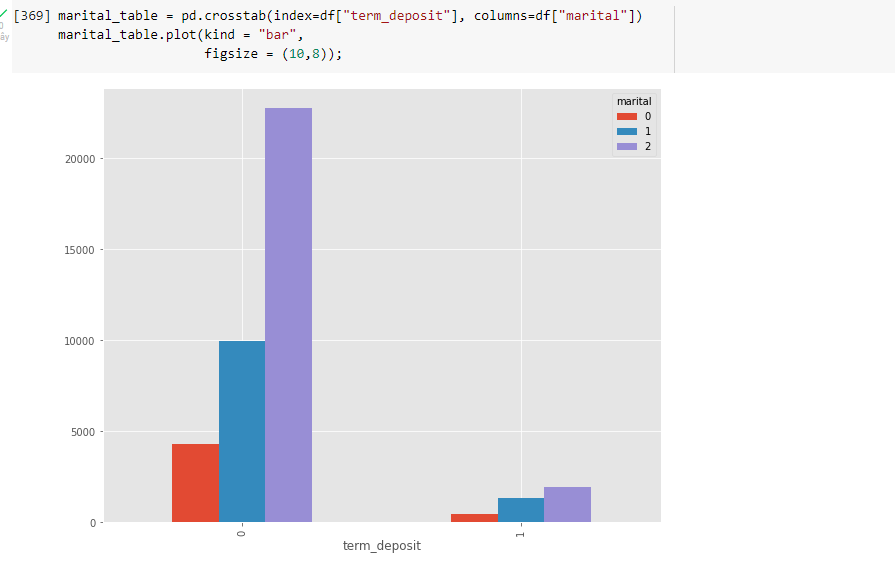
**Biểu đồ Bar hiển thị khoản vay cá nhân**



Hình 3.4.9: Biểu đồ Bar hiển thị khoản vay cá nhân

Giải thích: nhìn vào biểu đồ ta có thể thấy người không có khoản vay cá nhân chiếm tỉ lệ cao nhất số lượng người đăng ký tiền gửi có kỳ hạn. Tuy nhiên số lượng người không có khoản vay cá nhân mà không đăng ký cũng cao nhất nên dựa vào khoản vay cá nhân chúng ta khó chọn được ai sẽ có tỉ lệ cao hơn.

* **Biểu đồ Bar hiển thị tình trạng hôn nhân**



Hình 3.4.10: Biểu đồ Bar hiển thị tình trạng hôn nhân

Giải thích: nhìn vào biểu đồ ta có thể thấy người độc thân sẽ có tỉ lệ đăng ký cao nhất. Tuy nhiên số lượng người không đăng ký mà độc thân cũng cao nhất nên dựa vào tình trạng hôn nhân chúng ta khó chọn được ai sẽ có tỉ lệ cao hơn.

Tổng kết

Sau khi hoàn thành kỳ thực tập, em đã có được những kết quả:

* + - Biết sử dụng ngôn ngữ Python cơ bản, các thư viện để xử lý dữ liệu và trực quan hóa dữ liệu.
    - Tìm hiểu được các kỹ năng quan trọng về ngành.
    - Tìm hiểu được mức lương cũng như nhu cầu tuyển dụng của các doanh nghiệp.
    - Biết được những hạn chế hay khuyết điểm hiện tại của bản thân mà khắc phục dể chuẩn bị cho kỳ thực tập tốt nghiệp và cách làm việc sau này.

Tài liệu tham khảo

https://www.youtube.com/watch?v=RYbITgaJYX0&list=PLiC1doDIe9rCYWmH9wIEYEXXaJ4KAi3jc&index=10