

**UỶ BAN NHÂN DÂN THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

—🙐🕮🙐—



**MÔN HỌC: CÁC CÔNG NGHỆ LẬP TRÌNH HIỆN ĐẠI**

**BÀI BÁO CÁO**

**ĐỀ TÀI: TÌM HIỂU VÀ ỨNG DỤNG SCIKIT-LEARN TRONG PYTHON**

**THÀNH VIÊN**

LƯƠNG THANH TUẤN - 31222410447

TRẦN TRUNG VIỆT - 3122410470

TRẦN MINH TRƯỜNG - 3122410443

**GIẢNG VIÊN HƯỚNG DẪN: PHẠM THI VƯƠNG**

**TP. HỒ CHÍ MINH, 11/2025**

**MỤC LỤC**

# LỜI CẢM ƠN

Nhóm em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đến thầy Đỗ Ngọc Tài, giảng viên khoa Công nghệ Thông tin – Trường Đại học Sài Gòn, vì đã tận tình giảng dạy và hướng dẫn nhóm em trong suốt quá trình học tập cũng như hoàn thành bài báo cáo cuối kỳ này. Những kiến thức chuyên môn cùng với những kinh nghiệm thực tế mà thầy đã truyền đạt cho chúng em trong suốt khóa học không chỉ là nền tảng vững chắc giúp nhóm em hoàn thành tốt bài báo cáo, mà còn là nguồn động lực để chúng em tiếp tục rèn luyện, phát triển trên con đường học tập và nghề nghiệp sau này.

Trong quá trình thực hiện bài báo cáo, thầy đã không chỉ giúp nhóm em định hướng đúng đắn, giải quyết những khó khăn vướng mắc, mà còn truyền đạt những kỹ năng quý báu trong cách tiếp cận vấn đề và phương pháp làm việc hiệu quả. Nhờ có sự chỉ bảo tận tình và sự quan tâm sát sao của thầy, nhóm em đã có cơ hội học hỏi thêm nhiều kiến thức mới, từ lý thuyết cho đến thực hành. Những gì nhóm em học được không chỉ nằm trong khuôn khổ của bài báo cáo này, mà còn là hành trang cần thiết để ứng dụng vào các dự án thực tế và trong suốt quá trình làm việc sau này.

Dù đã cố gắng hết sức, nhưng với sự hạn chế về kiến thức chuyên môn và kinh nghiệm thực tiễn, nhóm em khó tránh khỏi những thiếu sót trong bài báo cáo. Nhóm em xin kính mong thầy thông cảm và bỏ qua cho những sai sót đó. Nhóm em rất mong nhận được những góp ý, nhận xét từ thầy, không chỉ để hoàn thiện tốt hơn bài báo cáo này mà còn giúp nhóm em rút ra những bài học quý báu, bổ sung và củng cố thêm những kỹ năng, kiến thức cần thiết để chuẩn bị tốt hơn cho những dự án, bài tập lớn và đặc biệt là bài báo cáo tốt nghiệp sắp tới.

Một lần nữa, nhóm em xin gửi lời cảm ơn sâu sắc đến thầy Đỗ Ngọc Tài. Kính chúc thầy luôn dồi dào sức khỏe, hạnh phúc và thành công trong sự nghiệp giảng dạy, tiếp tục dẫn dắt nhiều thế hệ sinh viên trên con đường chinh phục tri thức và vươn tới những thành công mới. Chúng em luôn trân trọng và biết ơn sự tận tâm của thầy.

*Nhóm em trân thành cảm ơn thầy!*

1. **LỜI MỞ ĐẦU**
2. **Lý do chọn đề tài**

Skicit-learn là thư viện máy học chuẩn mực trong hệ sinh thái Python, cung cấp các bộ thuật toán nền tảng (hồi quy, phân loại,…) cùng cơ chế Transformer/Pipeline giúp chuẩn hóa toàn bộ quy trình, tiền xử lý, huấn luyện, đánh giá đến suy luận. [Cùng với cộng đồng lớn, tài liệu đầy đủ và cập nhật thường xuyên](https://github.com/scikit-learn/scikit-learn), kết quả nghiên cứu/triển khai có tính phổ biến và chuyển biến cao giữa các dự án học thuật lẫn thực tế.

1. **Mục tiêu của báo cáo**

Báo cáo nhằm hiểu sâu thư viện scikit-learn ở mức thực hành, từ kiến trúc và các API cốt lõi (Transformer, Pipeline, Model Selection) đến cách thiết kế pipeline end-to-end cho các nghiên cứu tình huống điển hình (phân loại, hồi quy…), so sánh mô hình và đánh giá định lượng bằng thước đo phù hợp (AUC/F1, MAE/RMSE), rút ra insight có thể hành động và lựa chọn ngưỡng/quyết định theo bối cảnh, đồng thời triển khai suy luận mô hình dưới dạng API để ứng dụng di động/web có thể gọi sử dụng, và chuẩn hóa tính tái lập thông qua quản lý mã nguồn, phiên bản môi trường cùng hướng dẫn chạy để người khác có thể lặp lại kết quả.

## Bảng phân công công việc

### Bảng tiến độ làm việc

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **MÃ SV** | **HỌ TÊN** | **Email** | **TỈ LỆ PHẦN TRĂM** |
| 1 | 3122410447 | Lương Thanh Tuấn | aquattda@gmail.com | 100% |

Bảng 1. Bảng tiến độ làm việc

### Bảng phân công chi tiết

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **HỌ TÊN** | **CÔNG VIỆC** | **TRẠNG THÁI** |
| 1 | Lương Thanh Tuấn | Xây dựng quy trình khai phá dữ liệu | Hoàn thành |
| Viết báo cáo | Hoàn thành |
| Tìm hiểu và đặt câu hỏi | Hoàn thành |
| Xác định vấn đề | Hoàn thành |

Bảng 2. Bảng phân công chi tiết

# PHÂN TÍCH DỮ LIỆU

1. Lấy dữ liệu
2. Khám phá dữ liệu

Thông tin dữ liệu:

Mô tả chi tiết dữ liệu:

1. **Fixed acidity (axit cố định)**

**Ý nghĩa:**

* Tổng các axit không bay hơi trong rượu (tartaric, malic, lactic,…)
* Phạm vi ~ 4-12 g/L (cao: vị chua gắt, thấp: rượu nhẹ nhưng nhạt)

**Phương pháp:** Tách phần bay hơi (chưng cất), sau đó lấy cần cặn đem chuẩn độ kiềm (NaOH ~ 0.10 N) tới pH ~ 8.2

1. **Volatile acidity (axit bay hơi)**

**Ý nghĩa:**

* Tổng các axit bay hơi trong rượu, thành phần chính là acetic acid. Ngoài ra có thể có một ít formic, propionic,…
* Phạm vi ~0.1-1.5 g/L (cao: có mùi giấm giảm chất lượng, thấp: ổn định)

**Phương pháp:** Thực hiện phương pháp chưng cất giống Fixed ecidity.

1. **Citric acid**

**Ý nghĩa:**

* Là một axit hữu cơ yếu, không màu (phổ biến ở họ cam quýt).
* Giúp tạo độ chua tươi (“freshness”) cho rượu, giảm độ pH và cân bằng cấu trúc cồn/đường.
* Phạm vi ~0–1.0 g/L. (cao vừa: tươi; quá cao: bị gắt)

Phương pháp:

Sử dụng Enzymatic assay (UV–Vis 340 nm):

1. Cho enzyme phản ứng với citric acid trong mẫu.
2. Trộn mẫu với dung dịch NADH + enzyme trong cuvet.
3. Đo ánh sáng ở 340 nm trước khi cho enzyme tác động (A₁).
4. Thêm enzyme (khởi động phản ứng), chờ vài phút, đo lại (A₂).
5. Tính ΔA = A₁ − A₂ → tra/bấm công thức hoặc đường chuẩn để ra g/L citric acid (nhân hệ số pha loãng nếu có).

**ΔA càng lớn thì citric acid càng nhiều**

1. Residual sugar (glucose + fructose)

Ý nghĩa:

Là lượng đường dư còn lại chưa lên men trong rượu vang thành phẩm.

Tạo cảm giác ngọt cho rượu.

Phạm vi ~0.5-8 g/L

Phương pháp:

Sử dụng Enzymatic assay (UV–Vis 340 nm):

1. Cho enzyme phản ứng với citric acid trong mẫu.
2. Trộn mẫu với dung dịch NADH + enzyme trong cuvet.
3. Đo ánh sáng ở 340 nm với glucose (A₁).
4. Đo ánh sáng ở 340 nm với fructose (A2).
5. Tính RS = A₁ + A₂ (nhân hệ số pha loãng nếu có).
6. Chlorides (clorua)

Ý nghĩa:

Lượng ion clorua trong rượu

Phạm vi ~0.005–0.30 g/L (Clorua cao do quá trình sản xuất rượu vang đỏ là do các ion chiết xuất từ vỏ nho trong quá trình lên men.)

Phương pháp

Chuẩn độ bạc potentiometric (AgNO₃ + điện cực Ag/AgCl)

Nhỏ vài giọt AgNO₃ vào mẫu rượu, lợi dụng Cl có trong rượu để kết hợp với Ag tại ra kết tủa Agcl (trắng).

Khi Cl hết thì thêm một chút AgNO₃, do không còn Cl để tạo kết tủa nên khi đo điện thế sẽ nhảy mạnh (điểm tương đương)

Khi bắt được điểm tương đương sẽ ra được thể tích AgNO₃

Sử dụng công thức thể tính nồng độ Cl⁻:

Trong đó:

CAgNo3​​: nồng độ AgNO₃ chuẩn (mol/L)

Veq​: thể tích AgNO₃ (L)

Mcl=35.45 (mg/mmol)

Vmẫu: thể tích mẫu đem chuẩn độ (L)

1. Free sulfur dioxide (SO₂ tự do)

Là lượng SO2 chưa liên kết với các phân tử có trong rượu, nên có khả năng phản ứng để kháng khuẩn, chống oxy hóa.

Phạm vi ~15-35 mg/L, ngoài ra có thể đo phạm vi bằng chỉ số pH:

Phương pháp:

Aeration–Oxidation (AO)

Dùng Axit hóa để tách Free SO2 bay ra khỏi dung dịch và đẩy SO2 sang bình chứa H₂O₂ để thành H₂SO₄ và chuẩn độ để ra được Free SO2 ban đầu.

1. Total sulfur dioxide (SO₂ tổng)

Ý nghĩa: Là tổng của SO2 tự do và SO2 đã liên kết, phản ánh mức độ bảo quản tổng thể, siết chặt an toàn và chất lượng.

Phạm vi ~50-400 ml/L (Cao: Độ bảo quản tốt nhưng vi phạm quy định)

Phương pháp:

Aeration–Oxidation (AO)

Kiềm hoá/ủ để biến phần “bound” thành có thể giải phóng, rồi trong cùng phép đo dùng axit hoá + thổi khí để kéo tất cả SO₂ (free + bound đã giải phóng) sang bẫy và chuẩn độ.

1. Density (tỉ trọng)

Là khối lượng riêng của rượu đo tại nhiệt độ chuẩn 20 °C, phản ánh sự cân bằng cồn và đường dư. Khi tỉ trọng cao thì lượng đường dư tăng, khi tỉ trọng giảm thì cồn giảm.

Phạm vi ~0.987–1.040 g/cm³. (Density tăng thì residual tăng/ alcohol giảm và ngược lại).

Phương pháp:

Densimeter ống U dao động

Khử CO2 (lắc/siêu âm) để tránh sai số đo. Cho vào máy bơm mẫu cell ống U, trả về kết quả.

1. pH

Cho biết **mức độ axit/bazơ của rượu (thấp sẽ có vị chua**. Ảnh hưởng mạnh tới **ổn định vi sinh**, **màu**, **vị**, và **hiệu lực SO₂** (pH thấp → cùng một lượng free SO₂ cho **nhiều** SO₂ phân tử hơn → bảo quản tốt hơn).

Phạm vi ~**3.3 – 3.7**. pH < 3.3 ổn định vi sinh nhưng dễ “gắt” nếu acid cao, pH > 3.7 nguy cơ vi sinh cao nên cần lượng Free SO2 nhiều hơn và cảm giác mềm/”phẳng”

Phương pháp:

**Sử dụng máy HI84502 để pH-meter** với **điện cực thủy tinh**, đo ở **20 °C**.

**Hiệu chuẩn 2–3 điểm** bằng dung dịch đệm **pH 7.00 → 4.00 (→ 3.00)** trước khi đo.

**Khử CO₂** nhẹ và **khuấy đều** để số ổn định; nhúng điện cực, chờ **ổn**, đọc giá trị.

1. Sulphates (SO₄²⁻)

Ý nghĩa:

Là hàm lượng ion sunfat trong rượu, phụ gia liên quan hình thành SO₂.

Phạm vi ~0.3-0.8 g/L (Cao: gây đắng)

Phương pháp:

Turbidimetric/Gravimetric (BaSO₄):

Thêm BaCl₂ tạo kết tủa BaSO₄, đo độ đục (máy quan với bước sóng phù hợp) và so với đường chuẩn để ra nồng độ SO₄²⁻ ban đầu.

1. Alcohol (độ cồn)

Ý nghĩa: Là lượng ethanol theo thể tích trong rượu, ảnh hưởng đến hương vị.

Phạm vi ~12.0 – 14.5 % (tùy thuộc vào loại nho và cao sẽ gây nóng làm mất cân bằng).

Phương pháp:

Densimeter ống U dao động

Khử CO2 (lắc/siêu âm) để tránh sai số đo. Cho vào máy bơm mẫu cell ống U, trả về kết quả.

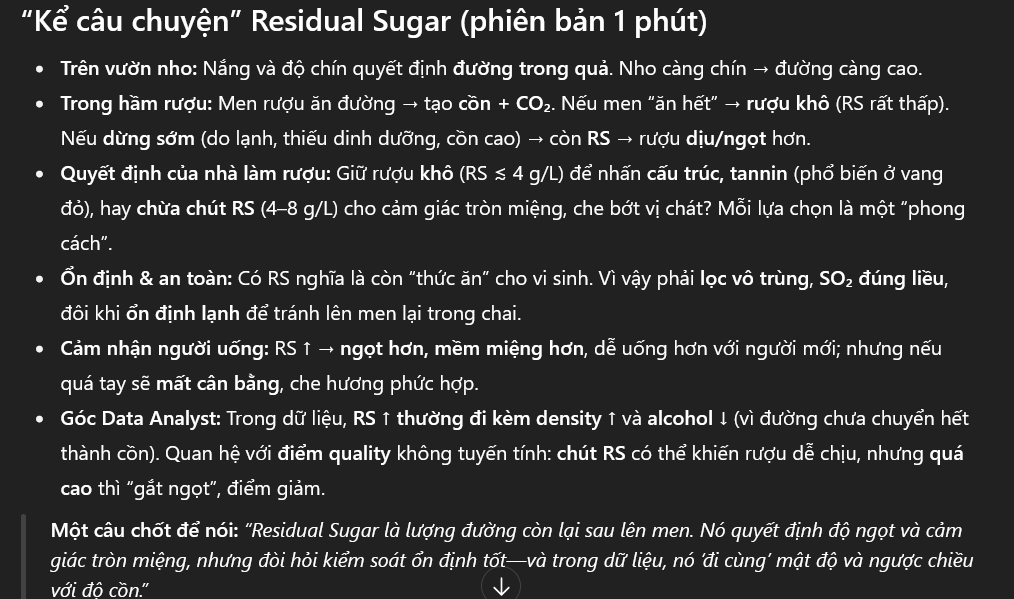
1. quality (điểm cảm quan)

Ý nghĩa: Là điểm đánh giá cảm quan tổng thể của rượu (mùi–vị–cấu trúc–cân bằng–hậu vị).

Phạm vi: ~3-8

Phương pháp:

Do các chuyên gia và giám khảo có kinh nghiệm lâu năm.



1. quality (điểm cảm quan, mục tiêu)

* Ý nghĩa: nhãn (3–8/9) dựa trên nếm thử.
* Lưu ý: không tuyến tính đơn giản với từng biến đơn lẻ; “chất lượng” là cân bằng tổng thể (acid–cồn–đường–đắng–mặn–hương).
* EDA: kiểm tra mất cân bằng lớp (class imbalance); xem phân bố theo loại vang (đỏ/trắng) nếu có.

A screenshot of a computer

AI-generated content may be incorrect.

|  |  |
| --- | --- |
| Tên cột | Ý nghĩa |
| Fixed acidity | Số lần mang thai |

## Chương 1: Giới thiệu chung

1. Lịch sử hình thành và phát triển

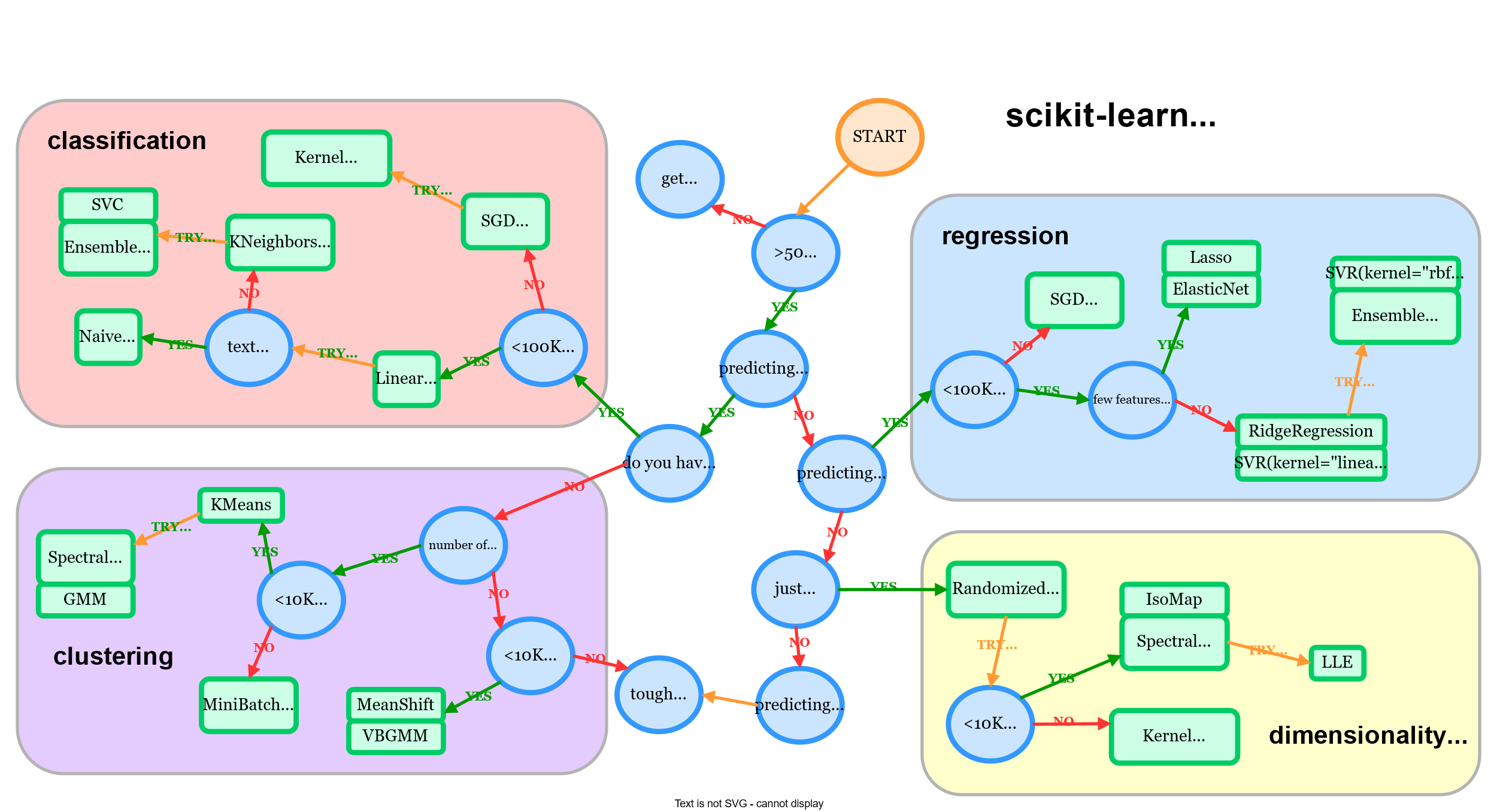
Theo wikipedia, scikit-learn bắt đầu năm 2007 do David Cournapeau trong khuôn khổ Google Summer of Code. Đến 2010, nhóm INRIA (Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel) tiếp quản và phát hành bản công khai đầu tiên (0.1 beta) vào 01/02/2010.

[A blue and orange circle with black text

AI-generated content may be incorrect.](https://commons.wikimedia.org/wiki/File:Scikit_learn_logo_small.svg)

Hình abc

Mục tiêu là cung cấp các công cụ học máy “đơn giản và hiệu quả” cho phân tích dữ liệu dự đoán, dễ dùng và tái sử dụng, xây trên NumPy, SciPy, Matplotlib, với API nhất quán để người dùng ráp pipeline đầu-cuối một cách chuẩn hoá.

[](https://scikit-learn.org/stable/machine_learning_map.html)

Hình 1. Abc

1. Sự thay đổi qua các phiên bản

[**Sau phiên bản 1.0 (9/2021)**](https://scikit-learn.org/stable/auto_examples/release_highlights/plot_release_highlights_1_0_0.html?utm_source=chatgpt.com), scikit-learn đánh dấu mốc ổn định hoá API và bổ sung các mảnh ghép hữu ích cho kỹ sư dữ liệu như SplineTransformer, QuantileRegressor, thuộc tính feature\_names\_in\_/get\_feature\_names\_out và chuẩn hoá plotting API; mô hình HistGradientBoosting cũng rời trạng thái “experimental” để sẵn sàng dùng rộng rãi.

[**Đến phiên bản 1.1 (2022)**](https://scikit-learn.org/stable/auto_examples/release_highlights/plot_release_highlights_1_1_0.html), dự án tập trung cải thiện hiệu năng, đặc biệt ở họ mô hình cây và tăng độ nhất quán của API, giúp việc truyền tên cột và thao tác EDA/tiền xử lý mượt hơn.

[**Phiên bản 1.2 (12/2022)**](scikit-learn.org/stable/auto_examples/release_highlights/plot_release_highlights_1_2_0.html?utm_source=chatgpt.com) mang lại set\_output(transform="pandas") để các transformer trả về DataFrame, cùng nhiều “display” trực quan hoá như PredictionErrorDisplay hay LearningCurveDisplay và nâng cấp cho Histogram-based Gradient Boosting.

[**Nhánh phiên bản 1.3 (2023)**](https://scikit-learn.org/stable/auto_examples/release_highlights/plot_release_highlights_1_3_0.html) tiếp tục các tinh chỉnh ổn định và hiệu năng, làm gọn trải nghiệm pipeline và đánh giá trong thực hành hằng ngày.

[**Bước sang phiên bản 1.4 (1/2024)**](https://scikit-learn.org/stable/auto_examples/release_highlights/plot_release_highlights_1_4_0.html), thư viện xử lý dữ liệu thực tế tốt hơn: HGB hiểu trực tiếp cột kiểu phân loại (categorical dtype), RandomForest có thể chia nhánh khi thiếu giá trị, mở rộng ràng buộc đơn điệu cho họ mô hình cây và hỗ trợ xuất Polars DataFrame.

[**Bản 1.5 (2024)**](https://scikit-learn.org/stable/auto_examples/release_highlights/plot_release_highlights_1_5_0.html) đưa ra FixedThresholdClassifier và TunedThresholdClassifierCV để tối ưu hoặc cố định ngưỡng quyết định theo metric và chi phí–lợi ích thay vì mặc định 0.5, đồng thời tối ưu PCA và cải thiện imputation, giúp mô hình gắn sát mục tiêu nghiệp vụ.

[**Ở phiên bản 1.6 (cuối 2024)**](https://scikit-learn.org/stable/auto_examples/release_highlights/plot_release_highlights_1_6_0.html), có FrozenEstimator, khả năng truyền thêm dữ liệu vào pipeline qua transform\_input, mở rộng hỗ trợ Array API, phát hành bánh xe thử nghiệm cho CPython free-threaded 3.13, ExtraTrees xử lý giá trị thiếu và tiện ích tải dữ liệu từ web tất cả nhằm mở đường cho hiệu năng và khả chuyển tốt hơn.

[**Cuối cùng, 1.7 (2025)**](https://scikit-learn.org/stable/auto_examples/release_highlights/plot_release_highlights_1_7_0.html) cải thiện trải nghiệm người dùng với biểu diễn HTML rõ ràng cho estimator, cho phép đặt validation set riêng cho HGB, hỗ trợ vẽ ROC trực tiếp từ kết quả cross-validation, làm nhất quán API của MLP và tiến thêm một bước trong việc dùng sparse arrays và Array API, khiến quy trình đánh giá và trình bày mô hình mạch lạc hơn.

1. **Hệ sinh thái**

Trọng tâm xoay quanh NumPy, SciPy, pandas, Matplotlib (nền tảng tính toán–dữ liệu–trực quan), joblib (lưu pipeline/mô hình), Optuna / scikit-optimize (tối ưu siêu tham số), Yellowbrick (trực quan đánh giá), FastAPI (đóng gói suy luận thành API) và Docker (đóng gói triển khai).

Dự án mã nguồn mở, có core maintainers và cộng tác viên quốc tế, tài liệu và ví dụ đầy đủ, quy trình review/CI chặt, phát hành 1.x đều đặn; mức độ hoạt động sôi nổi và ổn định, được dùng rộng rãi trong học thuật và doanh nghiệp.

## Chương 2: Cơ hội nghề nghiệp và thị trường

1. **Các vị trí công việc**

**Data Analyst**: khai thác dữ liệu doanh nghiệp, trực quan hoá, lập báo cáo KPI/insight; làm sạch/biến đổi dữ liệu với pandas/SQL; đôi khi áp dụng mô hình đơn giản (logistic/linear).

**Data Scientist**: xây dựng mô hình ML cho dự báo/phân loại/khuyến nghị; EDA, feature engineering, đánh giá (AUC/F1, RMSE…), triển khai thử nghiệm.

**Machine Learning Engineer**: sản xuất hoá mô hình tối ưu, đóng gói API (FastAPI), theo dõi mô hình; phối hợp với Data Engineer/MLOps; thường yêu cầu Python + scikit-learn + một số công cụ triển khai.

**Data Engineer** → xây dựng pipeline dữ liệu (ETL/ELT), kho dữ liệu/lakehouse; chuẩn bị dữ liệu cho DS/ML.

1. **Nhu cầu tuyển dụng**

**Hiện tại (VN):** thị trường CNTT 2024–2025 vẫn sôi động, nhiều vị trí data/AI xuất hiện cùng mặt bằng lương tăng ở vai trò chuyên môn mới; [ITviec tổng hợp 31 vị trí tiêu biểu và lộ trình mức lương cho giai đoạn 2024–2025](https://itviec.com/report/vietnam-it-salary-and-recruitment-market?utm_source=chatgpt.com).

**Xu hướng/động lực:** các báo cáo thị trường nhấn mạnh “AI surge” 2024–2025 và nhu cầu kỹ năng AI/ML; [VietnamWorks inTECH công bố báo cáo riêng về nhân lực IT trong bối cảnh AI](https://www.scribd.com/document/850046085/EN-IT-Report-2024-2025-VietnamWorks-inTECH?utm_source=chatgpt.com).

**Dự báo gần hạn (12–24 tháng):** kỹ năng AI/ML tiếp tục là “premium skill”, bài viết tổng hợp dựa trên [báo cáo AWS/LinkedIn cho biết lương có thể tăng đáng kể với kỹ năng AI](https://www.linkedin.com/pulse/preparing-future-how-ai-reshaping-vietnams-job-market-ly8bc/); kết hợp với báo cáo lương IT 2024–2025 cho thấy vai trò liên quan AI/ML vẫn nằm trong nhóm được chú ý.

1. **Mức lương tham khảo**

**Machine Learning Engineer (toàn VN):** trung bình ~**712 triệu VND/năm**; entry-level ~**506 triệu**, có thưởng; [theo SalaryExpert 2025](https://www.salaryexpert.com/salary/job/machine-learning-engineer/vietnam?utm_source=chatgpt.com).

**Machine Learning Engineer (TP.HCM):** khoảng **19–53 triệu VND/tháng** (khoảng tứ phân vị 25–75% [theo Glassdoor](https://www.glassdoor.com/Salaries/ho-chi-minh-vietnam-machine-learning-engineer-salary-SRCH_IL.0%2C19_IM1746_KO20%2C45.htm?utm_source=chatgpt.com)).

**Data Scientist (TP.HCM):** khoảng **20,5–36,7 triệu VND/tháng** (tứ phân vị 25–75% [theo Glassdoor](https://www.glassdoor.com/Salaries/ho-chi-minh-vietnam-data-scientist-salary-SRCH_IL.0%2C19_IM1746_KO20%2C34.htm?utm_source=chatgpt.com)).

1. **Các công ty/sản phẩm tiêu biểu đang sử dụng**

**J.P. Morgan** (ngân hàng đầu tư) — dùng rộng rãi cho phân loại, phân tích dự báo.

**Spotify** — dùng cho hệ thống gợi ý âm nhạc. [Scikit-learn](https://scikit-learn.org/stable/testimonials/testimonials.html)

**Booking.com, Evernote, Change.org, Zopa, Mars, BNP Paribas Cardif…** — nhiều testimonial chính thức trong tài liệu scikit-learn. [Scikit-learn](https://scikit-learn.org/stable/testimonials/testimonials.html)

# PHẦN 2: NỘI DUNG CỐT LÕI VÀ THỰC HÀNH

## Chương 3: Cơ hội nghề nghiệp và thị trường

1. Lý thuyết

Công cụ cần thiết bao gồm phiên bản python mới nhất (3.12.7), môi trường Anaconda/conda, sử dụng Jupyterlab/Notebook hoặc VS Code và các thư viện cần có bao gồm scikit-learn, pandas, numpy, matplotlib…

1. Hướng dẫn
   1. Sử dụng Anaconda

Đầu tiên cần tải trên trang chính thức:

<https://www.anaconda.com/download/success>

A screenshot of a computer

AI-generated content may be incorrect.

ấdsaas

Sẽ có 2 phiên bản Distribution (bao gồm nhiều package có sẵn nhưng chiếm nhiều bộ nhớ) và Miniconda (chỉ có conda, python và các thành phần cần thiết). Phù hợp với các hệ điều hành hiện có.

Đối với đề tài thì chỉ cần Miniconda (tự install các package cần thiết tránh lãng phí bộ nhớ).

A screenshot of a computer

AI-generated content may be incorrect.

Ấn ‘Next>’ để thực hiện tiến trình cài đặt.

A screenshot of a computer screen

AI-generated content may be incorrect.

Nhấn ‘I Agree’ để chấp thuận các điều khoản.

A screenshot of a computer

AI-generated content may be incorrect.

Just me (phù hợp với local) dễ dùng cho cá nhân.

All Users (phù hợp với cơ quan, tổ chức, công ty,…).

Với đồ án thì nên dùng ‘Just me’ là đủ.

A screenshot of a computer

AI-generated content may be incorrect.

Có thể giữ nguyên đường dẫn cài đặt, hoặc muốn thay đổi tùy vào mục đích cá nhân.

A screenshot of a computer

AI-generated content may be incorrect.

Không nên chọn ‘Add installation to my PATH env variable’ vì điều này có thể xung đột với python có trong máy từ trước.

Nếu sử dụng VSCode qua conda thì nên bật ‘Register Miniconad3…’, nhưng không khuyến khích.

Nếu máy của bạn không đủ bộ nhớ và luôn thường xuyên xóa các package trong env thì nên chọn ‘Clear the package cache…’ để mỗi lần tạo lại thì chỉ install các package cần thiết.

A screenshot of a computer

AI-generated content may be incorrect. Sau khi tải xong, tìm kiếm ‘Anaconda Prompt’ để cài môi trường và thư viện



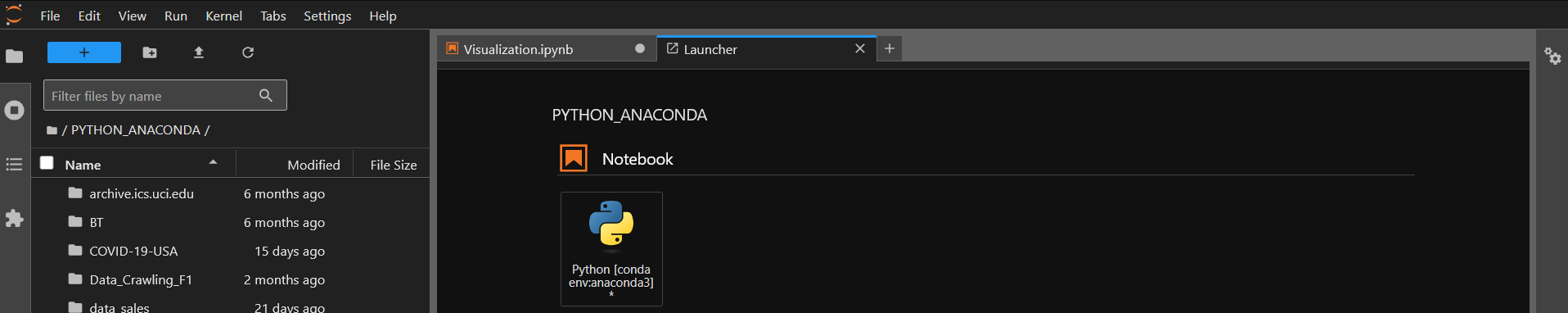
Sử dụng câu lệnh ‘conda create -n myenv python=<version> <package>’ để cài đặt môi trường tùy vào nhu cầu

Sau khi hoàn thành câu lệnh, hệ thống sẽ yêu cầu có accept với điều khoảng sử dụng (ToS) đó không (lần đầu), nếu có thì ghi ‘accept’, nếu muốn hủy nhấn ‘reject’ còn muốn xem chi tiết điều khoản nhấn ‘View’.

A screen shot of a computer

AI-generated content may be incorrect.

Sau khi nhấn ‘accept’, hệ thống tải và tiếp đến nhập ‘jupyter lab’ để vào localhost cá nhân.



Cuối cùng chọn biểu tượng python và code các dự án của bạn bằng python.

Ngoài ra có thể xem video hướng dẫn cài đặt:

<https://www.youtube.com/watch?v=mg6cMkz9Q0c>

* 1. Sử dụng Python (venv + pip)

Đầu tiên cần tải python mới nhất:

<https://www.python.org/downloads/>



Nhấn ‘Dowload Python 3.13.7’ đây là phiên bản mới nhất.

A screen shot of a computer

AI-generated content may be incorrect.

Nếu bạn đã cài python trước đó thì có thể chọn ‘Upgrade Now’ để cập nhật phiên bản mới nhất.

Nếu bạn chưa cài bao giờ thì chọn ‘Customize installation’

A screenshot of a computer

AI-generated content may be incorrect.

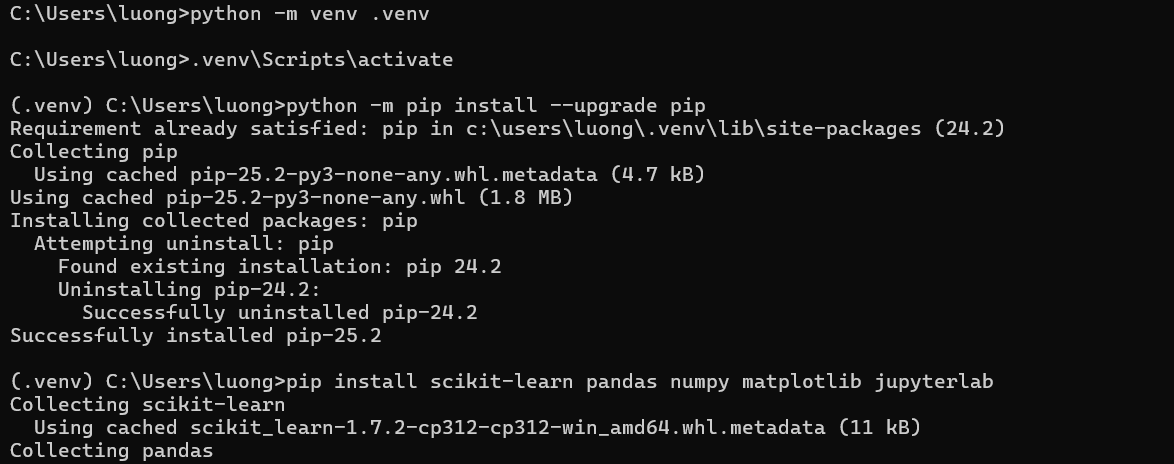
Đây là giao diện khi chọn ‘Customize installation’ sẽ tùy vào mục đích, tuy nhiên nên giữ nguyên.

A screenshot of a computer

AI-generated content may be incorrect.

Đây là phần chọn các chức năng nâng cao hơn, tùy vào mục đích. Nhưng vẫn nên giữ nguyên. Và chọn đường dẫn file hợp lý và ấn ‘Install’.

Đối với dùng JupyterLab/Notebook:

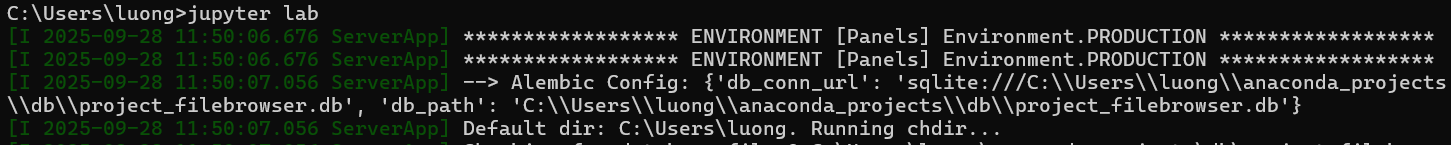


Thực hiện câu lệnh ‘python -m venv .venv’ để chuyển về .venv

Tiếp đến dùng ‘.venv\Scripts\activate’ để khởi động môi trường

Dùng ‘python -m pip install --upgrade pip’ để tải pip

Cuối cùng ‘pip install <package>’ cài các package cần thiết.



Sử dụng ‘jupyter lab’ để mở localhost giống Anaconda

Đối với dùng VS Code:

A screenshot of a computer

AI-generated content may be incorrect.

A logo on a black background

AI-generated content may be incorrect.

Vào Extensions (Ctrl + Shift + X) tải python và Jupter

A black and blue rectangle with white text

AI-generated content may be incorrect.

Nhấn ‘Ctrl + Shift + P’ và nhập ‘Python: Select Interperter’ để tạo môi trường

A screenshot of a computer program

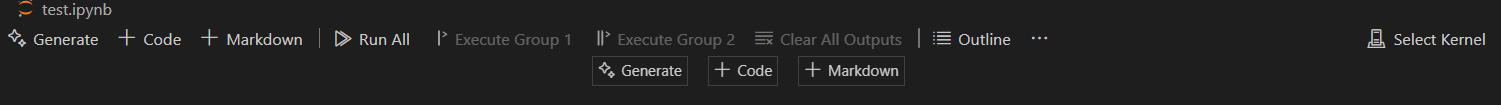
AI-generated content may be incorrect.

Chọn ‘Create Virtual Environment’ để tạo môi trường cho dự án của bạn

A screen shot of a computer

AI-generated content may be incorrect.

Hệ thống yêu cầu chọn Venv hay Conda tùy vào nhu cầu của bạn.



Ngoài ra bạn có thể tạo một file.ipynb và chọn ‘Slect Kernel’

A screen shot of a computer

AI-generated content may be incorrect.

Các kernel có sẵn trong VS Code hoặc bạn có thể chọn Kernel khác (nếu có). Cuối cùng thì bạn đã xong cài môi trường và có thể code được rồi!

1. Thực hành
2. Bảng phân công công việc
   1. Bảng tiến độ làm việc

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **MÃ SV** | **HỌ TÊN** | **Email** | **TỈ LỆ PHẦN TRĂM** |
| 1 | 3122410447 | Lương Thanh Tuấn | aquattda@gmail.com | 60% |
| 2 | 3122560072 | Đinh Phúc Thịnh | zaikaman123@gmail.com | 40% |

Bảng abc

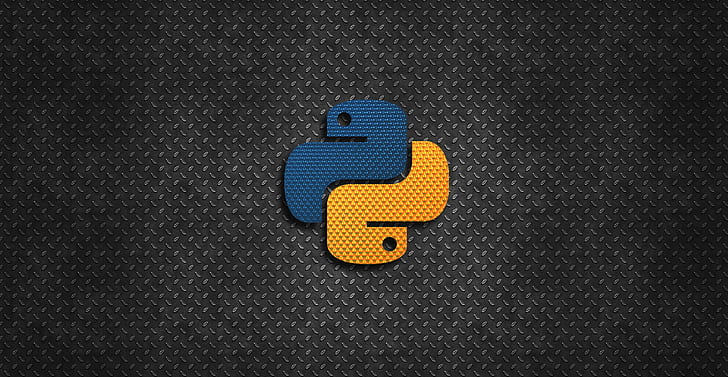
* 1. Bảng phân công chi tiết

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **HỌ TÊN** | **CÔNG VIỆC** | **TRẠNG THÁI** |
| 1 | Lương Thanh Tuấn | Xử lý dữ liệu | Hoàn thành |
| Viết báo cáo | Hoàn thành |
| Trực quan hóa dữ liệu | Hoàn thành |
| Xác định vấn đề | Hoàn thành |
| 2 | Đinh Phúc Thịnh | Trực quan hóa dữ liệu | Hoàn thành |
| Chuẩn bị dữ liệu | Hoàn thành |

Bảng abc

1. CƠ SỞ LÝ THUYẾT
2. **Tổng quan về Python**
   1. **Python là gì?**

Python là một trong những ngôn ngữ lập trình phổ biến nhất hiện nay, thường được sử dụng để xây dựng trang web và phần mềm, tự động hoá các tác vụ và tiến hành phân tích dữ liệu. Các nhà phát triển sử dụng Python vì nó hiệu quả, dễ học và có thể chạy trên nhiều nền tảng khác nhau. Phần mềm Python được tải xuống miễn phí, tích hợp tốt với tất cả các loại hệ thống và tăng tốc độ phát triển. Với sự phát triển của khoa học dữ liệu hiện nay, Python lại càng được ứng dụng rộng rãi hơn trong ngành Data Analyst.



ảnh abc

nguồn: [*https://s.net.vn/8wg0*](https://s.net.vn/8wg0)

* 1. **Ưu điểm**
* **Python khá dễ học**:

Python là một trong những ngôn ngữ lập trình dễ học nhất, do cú pháp rõ ràng, dễ đọc phù hợp với người mới bắt đầu. Python không có cú pháp phức tạp như các ngôn ngữ cấp cao khác như C hay C+. Chính vì vậy mà người học có thể nhanh chóng áp dụng vào các dự án thực tế khi phân tích dữ liệu với Python.

* **Khả năng ứng dụng rộng rãi, linh hoạt**:

Phân tích dữ liệu với Python trở nên phổ biến hơn bởi vì ngôn ngữ này được sử dụng trong nhiều lĩnh vực và dự án khác nhau. Các kỹ sư, nhà khoa học hay toán học đều có thể dùng ngôn ngữ này để hoàn thành các công việc của họ.

* **Nhiều loại thư viện**:

Các thư viện này đều được sử dụng miễn phí nên sẽ tiết kiệm được chi phí ngân sách phân tích dữ liệu. Thư viện này cũng đang được phát triển và nâng cấp thêm nhiều những tính năng cần thiết khác để giúp Data Analyst làm việc dễ dàng hơn với nguồn dữ liệu khổng lồ.

* **Python có cộng đồng lập trình viên vô cùng lớn**:

Học code không chỉ yêu cầu được đào tạo bài bản mà còn đòi hỏi quá trình tự học và tự nghiên cứu rất nhiều. Nhưng với những người mới bắt đầu, rất khó để tự tìm hiểu khi không có người hướng dẫn, giải đáp thắc mắc. Vì vậy, trong quá trình tự học, bất cứ lúc nào bạn cảm thấy khó khăn, muốn học hỏi thêm kinh nghiệm hoặc tìm tài liệu thì có thể tham gia vào các nhóm, cộng đồng ngôn ngữ Python để được giao lưu và học hỏi kinh nghiệm từ những người đi trước.



Bảng xếp hạng ngôn ngữ lập trình phổ biến nhất năm 2024

Nguồn: [*https://x.com/Python\_Dv*](https://x.com/Python_Dv)

* 1. **Nhược điểm**
* **Giới hạn tốc độ**:

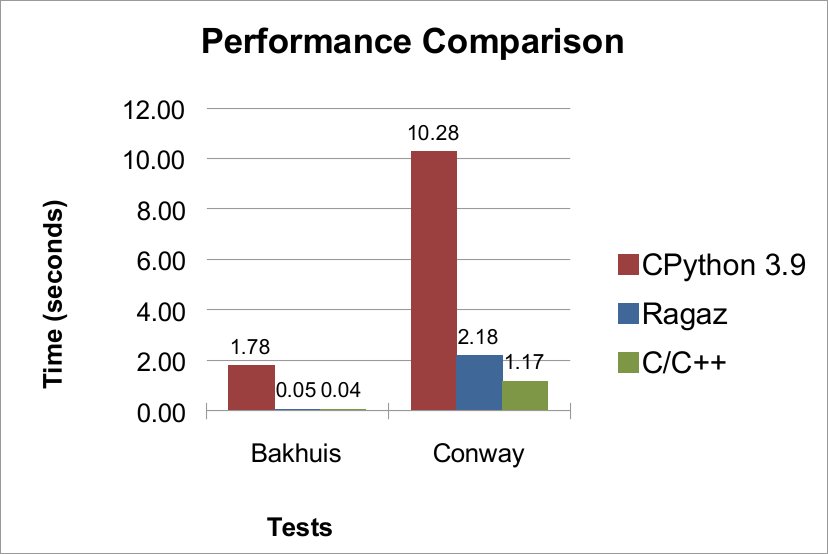
Vì là ngôn ngữ thông dịch, nên tốc độ của Python sẽ chậm hơn so với những ngôn ngữ khác. Mặc dù đã được cải thiện và tối ưu hóa trong vài năm trở lại đây nhưng so với C, C++ thì tốc độ vẫn là nhược điểm lớn nhất của Python.

* **Tiêu thụ bộ nhớ**:

Python có mức tiêu thụ bộ nhớ rất cao thế nên đây không phải là ngôn ngữ được lựa chọn hàng đầu cho những tác vụ đòi hỏi nhiều bộ nhớ.

* **Không có nguồn gốc từ môi trường di động**:

Python không có nguồn gốc từ di động và nó được một số lập trình viên coi là ngôn ngữ yếu cho điện thoại di động. Hai phần mềm Android và iOS không hỗ trợ Python như một ngôn ngữ lập trình chính thức. Tuy nhiên, Python có thể dễ dàng được sử dụng cho các mục đích di động, nhưng nó đòi hỏi một số tính năng bổ sung.



Hình ảnh so sánh hiệu suất của Python với Ragaz và C/C++

Nguồn: <https://s.net.vn/o77f>

1. **Vai trò của Python trong data Anaylyst**

* **Sử dụng Python để xử lý và làm sạch dữ liệu:**

Python cung cấp các thư viện mạnh mẽ như Pandas, NumPy và Dask, giúp bạn đọc, xử lý và làm sạch dữ liệu. Các thư viện này cho phép bạn thực hiện các thao tác như lọc, sắp xếp, ghép nối, biến đổi và tạo mẫu dữ liệu một cách dễ dàng và hiệu quả.

* **Sử dụng Python để mô hình hoá và dự đoán:**

Python cung cấp các thư viện mạnh mẽ như Scikit-learn, Statsmodels và TensorFlow, giúp bạn xây dựng và đào tạo các mô hình dự đoán. Bạn có thể áp dụng các thuật toán học máy, học sâu và thống kê để tạo ra các mô hình dự đoán cho các vấn đề phân loại, dự báo và gom cụm.

* **Sử dụng Python để phân tích số liệu và trực quan hóa**:

Python có các thư viện như Matplotlib, Seaborn và Plotly. Sử dụng Python cho Data Analysis để tạo các biểu đồ và đồ thị trực quan. Bằng cách sử dụng các thư viện này, bạn có thể tạo biểu đồ cột, biểu đồ đường, biểu đồ phân phối, biểu đồ tương quan và nhiều loại biểu đồ khác để khám phá và trình bày dữ liệu một cách hấp dẫn.

* **Sử dụng Python để phát triển tương tác:**

Python hỗ trợ các môi trường phát triển tương tác như Jupyter Notebook và JupyterLab. Điều này giúp bạn tạo ra các tệp notebook tương tác để thực hiện phân tích dữ liệu, viết mã và chia sẻ kết quả một cách trực quan. Môi trường notebook cũng hỗ trợ việc tài liệu hóa quy trình phân tích dữ liệu và giúp tăng tính tái sử dụng và khả năng chia sẻ.

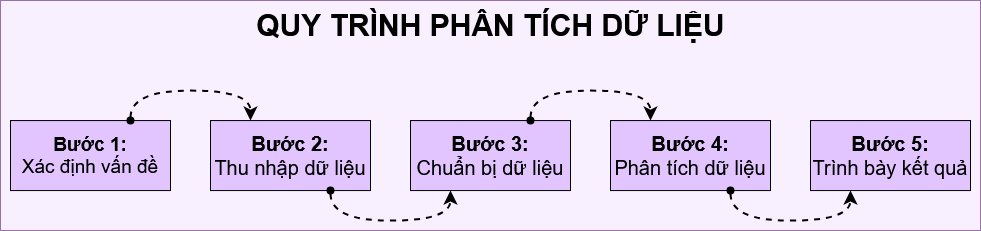
1. **Kết luận**

Python là một ngôn ngữ lập trình linh hoạt và mạnh mẽ, được ứng dụng rộng rãi trong phân tích dữ liệu nhờ vào cú pháp dễ học, khả năng tương thích cao, và sự hỗ trợ từ nhiều thư viện phong phú. Với các công cụ mạnh mẽ như Pandas, NumPy, Scikit-learn, và Matplotlib, Python không chỉ giúp các nhà phân tích xử lý dữ liệu một cách hiệu quả mà còn hỗ trợ xây dựng mô hình dự đoán, phân tích thống kê và trực quan hóa dữ liệu. Tuy nhiên, nhược điểm về tốc độ và bộ nhớ tiêu thụ của Python cần được cân nhắc khi xử lý dữ liệu lớn hoặc yêu cầu tối ưu hóa hiệu suất. Dù vậy, Python vẫn là một trong những lựa chọn hàng đầu trong lĩnh vực khoa học dữ liệu và phân tích dữ liệu, giúp các chuyên gia đưa ra quyết định dựa trên dữ liệu một cách nhanh chóng và chính xác.

1. **Tổng quan về Phân tích dữ liệu (Data Analyst)**
   1. **Phân tích dữ liệu là gì?**

Phân tích dữ liệu chuyển đổi dữ liệu thô thành thông tin chuyên sâu hữu ích. Quá trình này bao gồm đa dạng các công cụ, công nghệ và quy trình được áp dụng để tìm kiếm xu hướng và giải quyết vấn đề bằng dữ liệu. Phân tích dữ liệu có thể định hình các quy trình kinh doanh, cải thiện khả năng ra quyết định và thúc đẩy tăng trưởng doanh nghiệp.

* 1. **Quy trình phân tích dữ liệu**



Hinhanhr abc

**Mô tả quy trình**:

* **Bước 1: Xác định vấn đề**

Bước đầu tiên trong quy trình phân tích dữ liệu là xác định vấn đề cần nghiên cứu. Việc xác định rõ ràng vấn đề không chỉ giúp định hướng cho toàn bộ quá trình phân tích, mà còn tạo điều kiện thuận lợi cho việc phát triển các câu hỏi nghiên cứu cụ thể. Các câu hỏi này sẽ là kim chỉ nam cho toàn bộ quy trình và đảm bảo rằng mọi hoạt động phân tích đều hướng đến việc tìm kiếm câu trả lời cho những câu hỏi đó.

* **Bước 2: Thu nhập dữ liệu**

Sau khi vấn đề được xác định, bước tiếp theo là thu thập dữ liệu từ các nguồn đáng tin cậy. Dữ liệu cần được thu thập bằng các phương pháp phù hợp, chẳng hạn như khảo sát, phỏng vấn, hoặc sử dụng API từ các dịch vụ trực tuyến. Đảm bảo rằng dữ liệu thu thập được có độ chính xác và tính đại diện cao là điều cần thiết, vì chất lượng của dữ liệu sẽ ảnh hưởng trực tiếp đến kết quả phân tích.

* **Bước 3: Chuẩn bị dữ liệu**

Khi dữ liệu đã được thu thập, giai đoạn tiếp theo là chuẩn bị dữ liệu để đảm bảo rằng nó sẵn sàng cho phân tích. Quá trình này bao gồm làm sạch dữ liệu, xử lý các giá trị thiếu, và loại bỏ các dữ liệu ngoại lệ. Ngoài ra, việc chuyển đổi định dạng và tạo ra các biến mới cũng có thể cần thiết để nâng cao chất lượng và tính chính xác của dữ liệu. Giai đoạn chuẩn bị dữ liệu đóng vai trò quan trọng trong việc tạo ra một cơ sở dữ liệu đáng tin cậy cho phân tích.

* **Bước 4:Phân tích dữ liệu:**

Sau khi dữ liệu đã được chuẩn bị, bước phân tích được thực hiện. Trong giai đoạn này, các phương pháp phân tích thống kê và trực quan hóa dữ liệu sẽ được áp dụng để khám phá các mẫu và xu hướng trong dữ liệu. Việc sử dụng các công cụ phân tích như Python (với các thư viện pandas và matplotlib) hoặc Excel cho phép thực hiện các phân tích mô tả, phát hiện mối quan hệ giữa các biến, và kiểm tra các giả thuyết nghiên cứu. Trực quan hóa dữ liệu thông qua biểu đồ và bảng cũng rất quan trọng trong việc làm rõ thông tin và hỗ trợ việc ra quyết định.

* **Bước 5: Trình bày kết quả:**

Cuối cùng, kết quả phân tích cần được trình bày một cách rõ ràng và có hệ thống. Việc này không chỉ giúp người đọc hiểu rõ thông tin mà còn tạo điều kiện cho việc ra quyết định. Sử dụng báo cáo, biểu đồ, và bảng để thể hiện kết quả phân tích một cách trực quan là cần thiết, kèm theo các giải thích và nhận định cụ thể. Sự rõ ràng trong trình bày kết quả sẽ góp phần quan trọng vào việc truyền đạt thông tin và hỗ trợ các hành động tiếp theo dựa trên những phát hiện từ phân tích**.**

1. **Thư viện và môi trường phát triển**
   1. **Thư viện phát triển**

* **Thư viện Pandas:**

Pandas là một thư viện mã nguồn mở trong Python, được thiết kế đặc biệt để xử lý và phân tích dữ liệu một cách hiệu quả. Thư viện này cung cấp hai cấu trúc dữ liệu chính: **Series** (mảng một chiều có chỉ số) và **DataFrame** (mảng 2 chiều giống CSDL). Pandas nổi bật với các tính năng như làm sạch và xử lý dữ liệu, hỗ trợ trực quan hóa dữ liệu qua các thư viện như Matplotlib, và cung cấp các phương pháp phân tích thống kê mạnh mẽ. Thư viện này được sử dụng rộng rãi trong các lĩnh vực như tài chính, khoa học dữ liệu, và học máy để chuẩn bị và phân tích dữ liệu lớn.

* **Thư viện Numpy:**

NumPy (Numerical Python) là một thư viện mã nguồn mở trong Python, được sử dụng chủ yếu cho tính toán khoa học và xử lý dữ liệu. Thư viện này cung cấp một cấu trúc dữ liệu chính là **mảng N-dimensional** (ndarray), cho phép lưu trữ và thao tác với các mảng đa chiều một cách hiệu quả và nhanh chóng. NumPy hỗ trợ nhiều toán tử và hàm toán học, giúp thực hiện các phép toán số học, thống kê, và đại số tuyến tính. Một trong những ưu điểm lớn của NumPy là khả năng tối ưu hóa hiệu suất, nhờ vào việc thực hiện các phép toán ở mức thấp hơn, gần với mã máy. Điều này giúp NumPy trở thành nền tảng quan trọng cho nhiều thư viện khác trong Python, như Pandas, SciPy và scikit-learn. Sử dụng NumPy, người dùng có thể dễ dàng thực hiện các tác vụ tính toán phức tạp, từ việc xử lý dữ liệu đến việc phát triển các mô hình học máy.

* **Thư viện Matplotlib**:

Matplotlib là một thư viện mã nguồn mở trong Python, được sử dụng chủ yếu để tạo ra các biểu đồ và hình ảnh trực quan từ dữ liệu. Thư viện này cung cấp một giao diện linh hoạt cho việc vẽ biểu đồ 2D và 3D, cho phép người dùng dễ dàng tùy chỉnh các yếu tố như màu sắc, kiểu đường, nhãn trục và tiêu đề. Matplotlib rất phù hợp để trực quan hóa dữ liệu trong các lĩnh vực như khoa học, kỹ thuật và tài chính, nhờ vào khả năng tạo ra các biểu đồ như biểu đồ đường, biểu đồ cột, biểu đồ phân tán, và biểu đồ hình tròn. Một trong những điểm mạnh của Matplotlib là khả năng tích hợp tốt với các thư viện khác như NumPy và Pandas, giúp việc trực quan hóa dữ liệu trở nên dễ dàng hơn. Ngoài ra, Matplotlib còn hỗ trợ xuất hình ảnh ở nhiều định dạng khác nhau, như PNG, PDF và SVG, làm cho việc chia sẻ và báo cáo kết quả trở nên thuận tiện hơn.

* **Thư viện OS:**

Thư viện OS (Operating System) là một thư viện tiêu chuẩn trong Python, cung cấp các chức năng để tương tác với hệ điều hành. Thư viện này cho phép người dùng thực hiện các thao tác liên quan đến hệ thống tập tin, chẳng hạn như tạo, xóa, đổi tên và quản lý thư mục và tệp. Một trong những tính năng nổi bật của thư viện OS là khả năng truy cập thông tin hệ thống, như tên người dùng, đường dẫn tệp và biến môi trường. Bên cạnh đó, thư viện OS hỗ trợ các thao tác trên hệ điều hành, giúp người dùng thực hiện các lệnh shell và chạy các chương trình từ Python. Việc sử dụng thư viện OS giúp đơn giản hóa các tác vụ thường gặp trong lập trình và tăng tính linh hoạt khi phát triển ứng dụng, đặc biệt khi làm việc với dữ liệu và tệp trong các dự án phân tích dữ liệu.

* **Thư viện PDF2:**

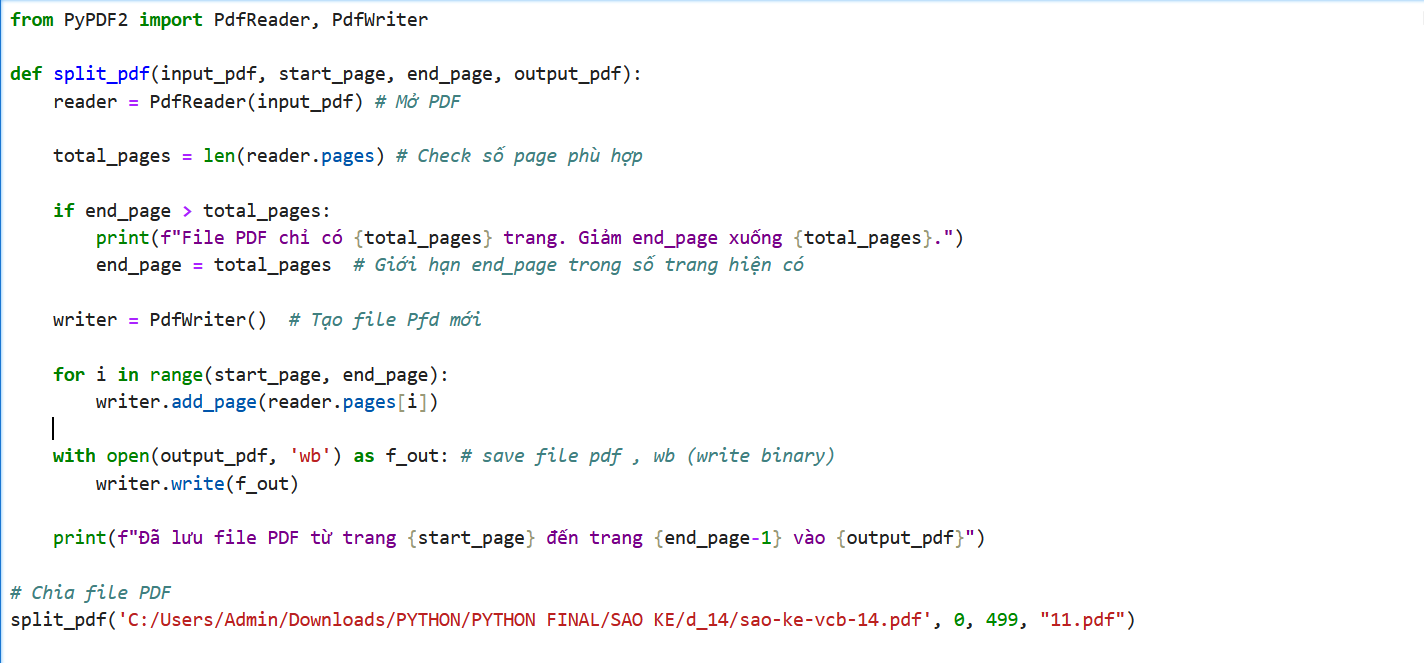
PyPDF2 là một thư viện mã nguồn mở trong Python, được sử dụng để thao tác với tệp PDF. Thư viện này cho phép người dùng thực hiện nhiều chức năng liên quan đến PDF, chẳng hạn như đọc, ghi, tách, và hợp nhất các tệp PDF. Với PyPDF2, người dùng có thể trích xuất văn bản từ các trang PDF, thêm hoặc xóa trang, và thậm chí mã hóa hoặc giải mã tệp PDF để bảo vệ nội dung. Thư viện này rất hữu ích trong các ứng dụng cần xử lý tài liệu điện tử, như phân tích dữ liệu từ báo cáo PDF hoặc tạo tài liệu PDF tự động. PyPDF2 cung cấp một giao diện dễ sử dụng, giúp lập trình viên nhanh chóng tích hợp các chức năng liên quan đến PDF vào trong dự án của mình mà không cần phải hiểu sâu về định dạng tệp PDF.

* 1. **Môi trường phát triển**

Jupyter Notebook là một môi trường phát triển tương tác mạnh mẽ, cho phép người dùng viết và chạy mã trong các ô (cell) riêng biệt, chủ yếu với ngôn ngữ Python nhưng cũng hỗ trợ nhiều ngôn ngữ khác. Với tính năng tương tác cao, người dùng có thể kiểm tra và điều chỉnh mã ngay lập tức, giúp quá trình phát triển và gỡ lỗi trở nên dễ dàng. Jupyter Notebook cho phép kết hợp mã, văn bản (Markdown), và hình ảnh trong cùng một tài liệu, tạo ra các báo cáo và tài liệu hướng dẫn trực quan và dễ đọc. Tất cả các ô mã và đầu ra được lưu lại, giúp theo dõi lịch sử làm việc và kết quả phân tích một cách thuận tiện. Ngoài ra, Jupyter có thể chạy trên các dịch vụ cloud như Google Colab, cho phép người dùng truy cập từ bất kỳ đâu mà không cần cài đặt phần mềm cục bộ. Nhờ vào những đặc điểm nổi bật này, Jupyter Notebook đã trở thành công cụ lý tưởng cho phân tích dữ liệu, học máy, và nghiên cứu khoa học.

1. PHÂN TÍCH DỮ LIỆU
2. Lấy dữ liệu

* Cách thu thập dữ liệu:
* Bước 1: Vào trang “Thông tin Chính phủ” lấy file PDF sao kê của ngân hàng Vietcombank từ ngày 7/9/2024 (ngày bão bắt đầu đổ bộ) đến 14/9/2024.
* Bước 2: Sau khi lấy được file PDF, tiếp đến chia dữ liệu ra làm nhiều file PDF nhỏ (kích thước 500 trang) sau đó sử dụng trang web <https://online2pdf.com> để chuyển file PDF thành file CSV.



Hinhanh abc

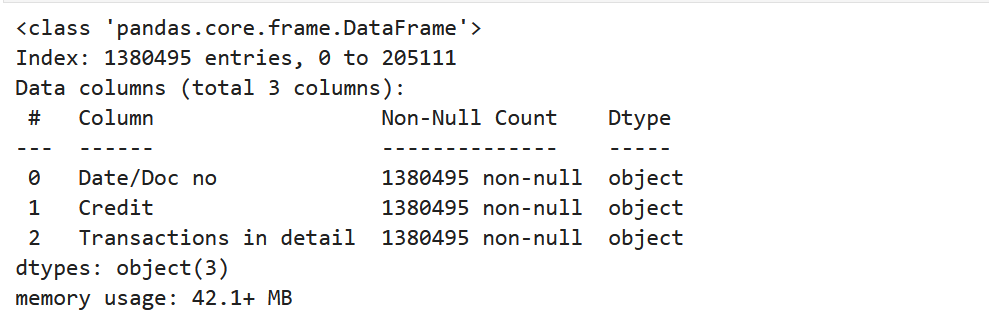
* Bước 3: Gộp hết vào thành 1 file CSV duy nhất.



Hinh anh abc

1. Khám phá dữ liệu

* Thông tin dữ liệu:

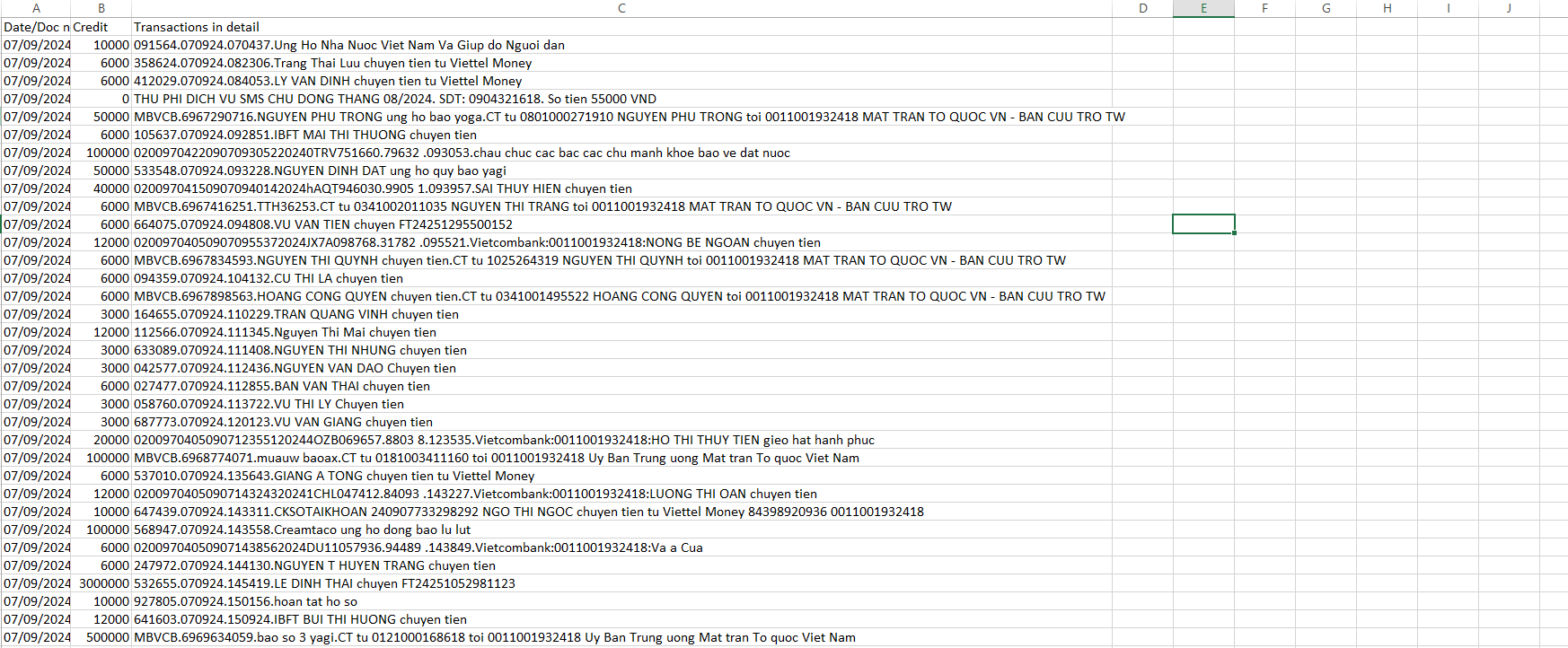


Hinh anh abc

* Vấn đề về dữ liệu CSV:
* Hiện tại, Dtype của cột Credit đang ở dạng object, điều này sẽ gây cản trở cho việc thực hiện các phép toán tài chính.
* Cột Date/Doc no không cần thiết vì chỉ cần ngày giao dịch để trả lời câu hỏi từ các bên liên quan.

1. Phân tích chi tiết

* Dữ liệu sau khi thu thập được bao gồm 1.380.495 dòng:



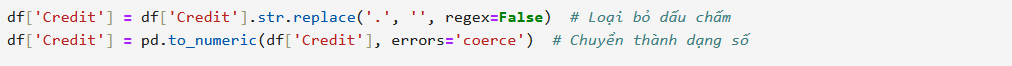
Hình ảnh abc

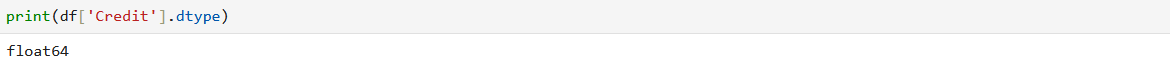
* **Mô tả chi tiết dữ liệu:**

|  |  |
| --- | --- |
| Tên cột | Ý nghĩa |
| Date/Doc no | Ngày giao dịch và mã giao dịch của ngân hàng. |
| Credit | Số tiền ghi có. |
| Transactions in detail | Nội dung chi tiết giao dịch. |

BẢNG ABC

* **Tiền xử lý dữ liệu:**
* Đổi Dtype của cột ‘Credit’ thành float và loại bỏ dấu phẩy:





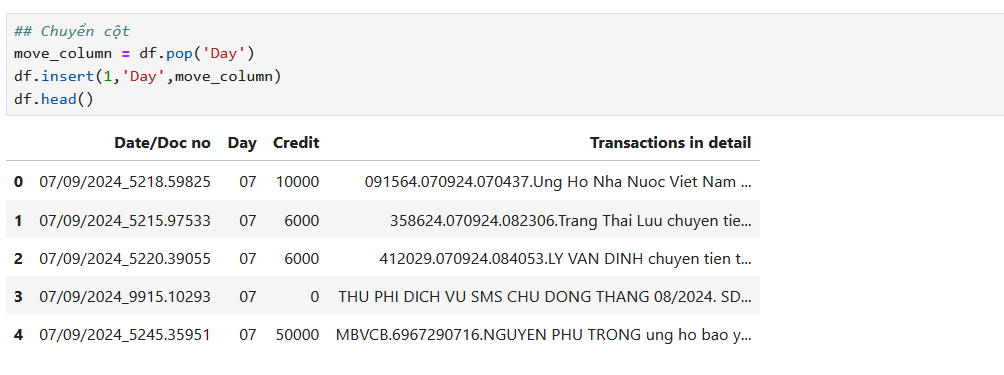
ảnh abc

* Thêm cột 'Day' từ cột 'Date/Doc no' để hiển thị các ngày giao dịch:



ảnh abc

* Chuyển cột ‘Day’ trước cột ‘Credit’:



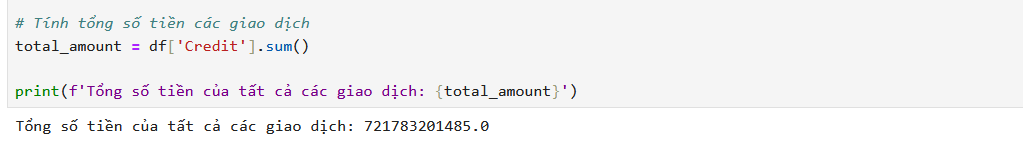
ảnh abc

* **Tổng quang về dữ liệu:**
* Số ngày giao dịch: 11 ngày.
* Tổng số lượng tất cả các giao dịch: 135.069.756.507 (135 tỷ)



ảnh abc

* Tổng số tiền giao dịch: 721.783.201.485 (721 tỷ)



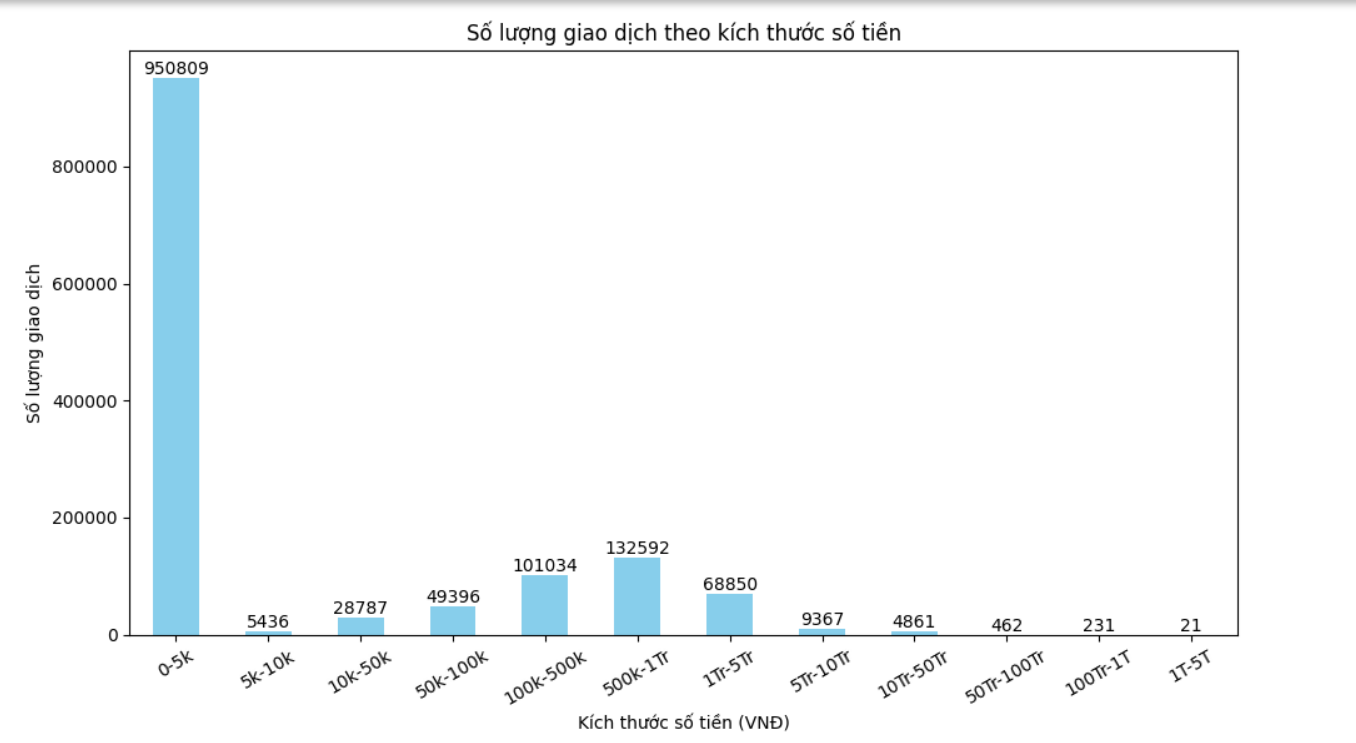
ảnh abc

* Tổng số tiền mỗi ngày:

|  |  |
| --- | --- |
| Ngày giao dịch | Số tiền |
| 07/09/2024 | 6.613.000 |
| 08/09/2024 | 885.410.201 |
| 09/09/2024 | 7.868.744.495 (7 tỷ) |
| 10/09/2024 | 126.308.988.811 ( 126 tỷ) |
| 11/09/2024 | 155.145.262.442 (155 tỷ) |
| 12/09/2024 | 154.998.779.363 (154 tỷ) |
| 13/09/2024 | 199.060.957.546 (199 tỷ) |
| 14/09/2024 | 77.509.647.294 (77 tỷ) |

Bảng abc

1. TRỰC QUANG HÓA DỮ LIỆU
2. Trong khoảng số tiền nào thì số lượng giao dịch đạt cao nhất hoặc thấp nhất, và số lượng giao dịch tương ứng cũng như tổng số tiền giao dịch là bao nhiêu?

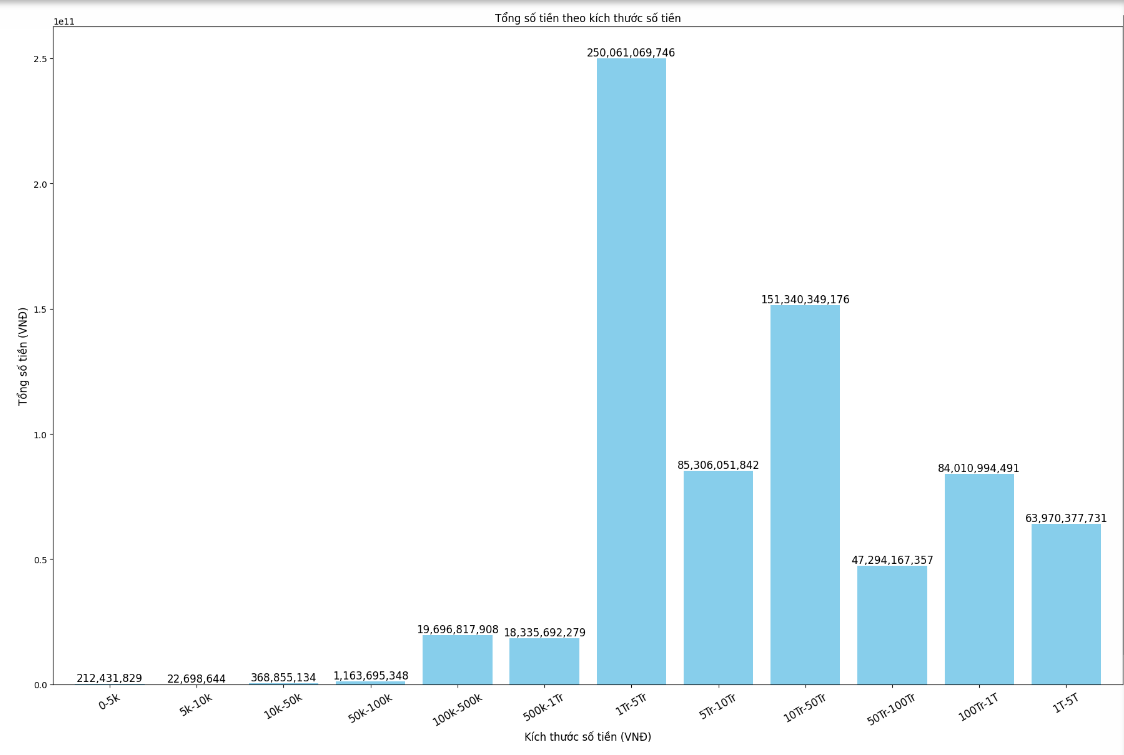


ảnh abc

**Mô tả biểu đồ:**

|  |  |
| --- | --- |
| Khoảng tiền giao dịch (VND) | Số lượng giao dịch |
| 0-5k | 950.809 |
| 5k-10k | 5.436 |
| 10k-50k | 28.787 |
| 50k-100k | 49.396 |
| 100k-500k | 101.034 |
| 500k-1Tr | 132.592 |
| 1Tr-5Tr | 68.850 |
| 5Tr-10Tr | 9.367 |
| 10Tr-50Tr | 4.861 |
| 50Tr-100Tr | 462 |
| 100Tr-1T | 231 |
| 1T-5T | 21 |

Bảng abc



**Mô tả biểu đồ:**

|  |  |
| --- | --- |
| Khoảng tiền giao dịch (VND) | Tổng số tiền (VND) |
| 0-5k | 212.431.829 |
| 5k-10k | 22.698.644 |
| 10k-50k | 368.855.134 |
| 50k-100k | 1.163.695.348 (1 tỷ) |
| 100k-500k | 19.696.817.908 (19 tỷ) |
| 500k-1Tr | 18.335.692.279 (18 tỷ) |
| 1Tr-5Tr | 250.061.069.746 (250 tỷ) |
| 5Tr-10Tr | 85.306.051.842 (85 tỷ) |
| 10Tr-50Tr | 151.340.349.176 (151 tỷ) |
| 50Tr-100Tr | 47.294.167.357 (47 tỷ) |
| 100Tr-1T | 84.010.994.491 (84 tỷ) |
| 1T-5T | 63.970.377.731 (63 tỷ) |

Bảng abc

* **Nhận xét :**
* Dữ liệu từ sao kê ngân hàng Vietcombank cho thấy xu hướng đóng góp của các cá nhân và tổ chức cho đồng bào miền Bắc có sự phân hóa rõ rệt giữa các khoảng tiền. Phần lớn số lượng giao dịch tập trung ở các mức đóng góp nhỏ, đặc biệt là khoảng **0-5k VND, với 950.809 giao dịch**, chiếm số lượng cao nhất. Điều này cho thấy có rất nhiều người tham gia đóng góp dù chỉ với các khoản tiền nhỏ. Tuy nhiên, **tổng giá trị của các giao dịch trong khoảng này chỉ đạt 212.431.829 VND**, đóng góp không đáng kể vào tổng quỹ.
* Các giao dịch ở mức trung bình (50k-500k VND) có số lượng cao thứ hai, đặc biệt ở các khoảng **50k-100k và 100k-500k VND với số lượng lần lượt là 49.396 và 101.034 giao dịch**. Đây có thể là các khoản đóng góp từ cá nhân, ở mức vừa phải nhưng vẫn thể hiện sự tham gia rộng rãi của cộng đồng.
* Đáng chú ý là các giao dịch trong khoảng **1 Tr-5 Tr VND**, tuy số lượng không nhiều nhưng **tổng giá trị giao dịch lại rất lớn, lên đến 250.061.069.746 VND**. Điều này cho thấy nhiều cá nhân và tổ chức sẵn sàng đóng góp với mức cao hơn, góp phần đáng kể vào tổng giá trị ủng hộ.
* Trong khi đó, các giao dịch ở mức rất cao, từ **50 Tr trở lên, có số lượng thấp nhất**, như ở khoảng **1 T-5 T VND chỉ có 21 giao dịch, nhưng tổng giá trị lại đạt 63.970.377.731 VND**, thể hiện vai trò quan trọng của những khoản đóng góp lớn từ các cá nhân hoặc tổ chức có khả năng tài chính mạnh.
* **Tóm lại, mặc dù các khoản đóng góp nhỏ chiếm số lượng lớn nhưng đóng góp chính vào tổng quỹ lại đến từ các giao dịch lớn**. Điều này cho thấy cả những khoản đóng góp nhỏ và lớn đều có ý nghĩa, thể hiện sự đoàn kết của cộng đồng trong việc hỗ trợ đồng bào miền Bắc.

Giá tiền nên thuần việt

* Tham khảo web báo cáo web ngân hang
* Với dòng tiền từ khắp đất nước chuyển vào thì có ảnh hưởng tới các vùng lân cận không?
* ( bắc lũ thì ảnh hưởng đến các miền khác…) => update ( đồ án thực tập)
* Sản phẩm thiết yếu/không thiết yếu nội thất (ngoài) so sánh giá trước và sau bão -> nhận xét nó ảnh hưởng ntn đến

( biểu đồ thực phẩm theo tháng/năm) CPI (2020) => Cục bão lũ có

Từ abc -> trả lời ( ảnh hưởng VN)

4 ngày

giao dịch: 200228

Tổng số tiền tất cả các giao dịch: 135.069.756.507(135 tỷ)

Tổng số tiền mỗi ngày:

2024-09-07 6.613.000

2024-09-08 885.410.201

2024-09-09 7.868.744.495 (7 tỷ)

2024-09-10 126.308.988.811 ( 126 tỷ)

Thống kê theo key\_word của nội dung:

Hoc sinh: Số lượng chuyển khoản: 348; Tổng tiền chuyển khoản: 51.141.396 ;Giá trị trung bình: 146958,0344827586

Hoc sinh (loại đi key 'tap the'): Số lượng chuyển khoản: 346; Tổng tiền chuyển khoản: 40.051.396 ;Giá trị trung bình: 115755,47976878613

Cua it long nhieu: Số lượng chuyển khoản: 1710; Tổng tiền chuyển khoản: ;352.066.087 Giá trị trung bình: 205886,60058479532

Xin loi: Số lượng chuyển khoản: 24; Tổng tiền chuyển khoản: 2.855.367 ;Giá trị trung bình: 118973,625

sinh vien: Số lượng chuyển khoản: 611; Tổng tiền chuyển khoản: 83.904.106 ;Giá trị trung bình: 137322,59574468085

Giao dịch nhỏ hơn 10.000:

Tổng số giao dịch: 1864

Tổng số tiền: 8.555.458

Giá trị trung bình: 4589,936765528819

15 cụm 2 từ xuất hiện nhiều nhất:

[('ung ho', 65612), ('ban cuu', 42101), ('tro tw', 42059), ('ct tu', 41251), ('dong bao', 38532), ('to quoc', 24248), ('mat tran', 24128), ('chuyen tien', 23438), ('toi 0011001932418', 23167), ('tran to', 18624), ('bao yagi', 18405), ('quoc vn', 18185), ('0011001932418 mat', 17432), ('khac phuc', 16780), ('ho dong', 14196)]

Khoảng tiền có số lượng giao dịch nhiều nhất: 100k-500k với 94 nghìn giao dịch với tổng số tiền gần 30 tỷ.

* Thêm mục tiêu nếu có câu hỏi về Lý do nhà nước lại công bố sao kê?

( tránh lạm phát, rửa tiến, minh bạch dòng tiền, cân bằng cung-cầu, liên quan vi mô vĩ mô )

**TÀI LIỆU THAM KHẢO**

[1] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2017.

[2] W. Shields, *SQL QuickStart Guide: The Simplified Beginner’s Guide to Managing, Analyzing, and Manipulating Data With SQL*, 1st ed. ClydeBank Media LLC, 2019.

[3] M. Shron, *Thinking with Data: How to Turn Information into Insights*, 1st ed. O'Reilly Media, 2014.

[4] C. N. Knaflic, *Storytelling with Data: A Data Visualization Guide for Business Professionals*, 1st ed. Wiley, 2015.

[5] A. Maheshwari, *Data Analytics Made Accessible*, 1st ed. DMM Direct, 2014.

[6] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*, O'Reilly Media, 2016.

[7] H. Chen, *Practical Data Science with Python*, Apress, 2019.

[8] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed., O'Reilly Media, 2019.

[9] N. C. Zakas, *Understanding ECMAScript 6: The Definitive Guide for JavaScript Developers*, No Starch Press, 2016.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.

[11] R. L. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu, *Data Mining for Scientific and Engineering Applications*, Springer, 2001.

[12] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.

[13] J. Heaton, *Deep Learning and Neural Networks: Python, Theano, and Keras*, Heaton Research, 2018.

[14] S. Raschka, *Python Machine Learning*, 3rd ed., Packt Publishing, 2019.

[15] E. Tufte, *The Visual Display of Quantitative Information*, 2nd ed., Graphics Press, 2001.