TRƯỜNG CAO ĐẲNG FPT POLYTECHNIC

BÁO CÁO DỰ ÁN TỐT NGHIỆP

- Chuyên ngành Xử Lý Dữ Liệu -

Phân Tích Hiệu Suất Làm Việc Từ Xa

GVHD: Thầy Văn Công Khanh

Danh sách thành viên :	Nhóm 3
Nguyễn Duy Lũy	– PS40096 (Nhóm trưởng)
Hà Thị Như Ý	– PS38021
Trịnh Thị Mỹ Huệ	– PS42117
Vi Ngọc Khánh Linh	– PS39736
Trần Thị Hưởng Thọ	– PS41852

TP.HCM 08 – 2025

LỜI CẢM ƠN

Nhóm chúng em xin gửi lời cảm ơn chân thành và sâu sắc đến Thầy **Văn Công Khanh** – giảng viên ngành *Xử lý dữ liệu* tại Trường Cao đẳng Thực hành FPT Polytechnic – người đã luôn tận tâm hướng dẫn, hỗ trợ và đồng hành cùng nhóm trong suốt quá trình thực hiện đề tài **"Phân tích hiệu suất làm việc từ xa."**

Bằng sự tận tâm, trách nhiệm và chuyên môn vững vàng, thầy đã giúp nhóm không chỉ hiểu rõ phương pháp phân tích dữ liệu mà còn tiếp cận được tư duy hệ thống, tư duy phản biện và tinh thần làm việc nghiêm túc. Những góp ý xác đáng, sự kiên nhẫn và động viên từ thầy là yếu tố then chốt giúp nhóm hoàn thiện dự án một cách khoa học và thực tiễn.

Chúng em cũng xin chân thành cảm ơn **quý thầy cô Trường Cao đẳng thực hành FPT Polytechnic**, đặc biệt là các giảng viên ngành **Xử lý dữ liệu**, đã trang bị cho chúng em nền tảng kiến thức vững chắc và kỹ năng cần thiết trong suốt quá trình học tập. Đây chính là hành trang quan trọng giúp chúng em tự tin bước vào các dự án thực tế như đề tài lần này.

Mặc dù đã nỗ lực hết mình, nhưng do giới hạn về thời gian và kinh nghiệm, nhóm không tránh khỏi những thiếu sót. Chúng em rất mong nhận được sự góp ý quý báu từ thầy cô để tiếp tục hoàn thiện trong các dự án sau.

Trân trọng cảm ơn!

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP.HCM, ngày ..... tháng ..... năm 2025

Giáo viên hướng dẫn

(Ký tên và ghi rõ họ tên)

LỜI MỞ ĐẦU

Làm việc từ xa đang trở thành một xu hướng phổ biến trong ngành công nghệ thông tin, mở ra nhiều cơ hội nhưng cũng đặt ra không ít thách thức về hiệu suất, khả năng phối hợp và mức độ hài lòng của nhân viên. Việc đánh giá hiệu quả của mô hình này cần dựa trên phân tích dữ liệu thực tế, khách quan và toàn diện.

Xuất phát từ thực tế đó, nhóm thực hiện đề tài **"Phân tích hiệu suất làm việc từ xa"** với mục tiêu khai thác và xử lý dữ liệu từ khảo sát **Stack Overflow Developer Survey 2024.** Qua đó, nhóm tìm hiểu các yếu tố ảnh hưởng đến hiệu suất làm việc, phân tích sự khác biệt giữa các nhóm đối tượng và đề xuất những góc nhìn hỗ trợ quá trình ra quyết định trong tổ chức.

Đây cũng là cơ hội để nhóm vận dụng kiến thức chuyên ngành **Xử lý dữ liệu**, phát triển kỹ năng phân tích, trực quan hóa và trình bày thông tin một cách có hệ thống và logic.

MỤC LỤC

DANH SÁCH BẢNG BIỂU:

DANH MỤC HÌNH VẼ:

1 TỔNG QUAN DỰ ÁN

1.1 GIỚI THIỆU DỰ ÁN

1.1.1 HIỆN TRẠNG

Trong kỷ nguyên công nghệ số, làm việc từ xa không còn là giải pháp tạm thời mà đang trở thành xu hướng phổ biến toàn cầu, đặc biệt trong lĩnh vực công nghệ thông tin. Mô hình này mang lại nhiều lợi ích như linh hoạt thời gian, tiết kiệm chi phí và mở rộng nguồn nhân lực.

Tuy nhiên, đi kèm với đó là không ít thách thức: giảm tương tác nhóm, khó kiểm soát tiến độ, nhân viên dễ cô lập và suy giảm tinh thần làm việc. Những yếu tố này có thể ảnh hưởng trực tiếp đến hiệu suất và sự hài lòng của người lao động.

Trước thực trạng đó, việc hiểu rõ hiệu quả thực tế của làm việc từ xa, đồng thời xác định các nguyên nhân làm suy giảm hiệu suất trở thành nhu cầu cấp thiết đối với cả doanh nghiệp lẫn người lao động trong quá trình chuyển đổi mô hình làm việc.

1.2 MỤC TIÊU, PHẠM VI

1.2.1 MỤC TIÊU DỰ ÁN:

Dự án hướng tới việc phân tích hiệu suất làm việc của lập trình viên theo từng mô hình làm việc (tại văn phòng, kết hợp, từ xa) nhằm hỗ trợ doanh nghiệp và sinh viên ngành CNTT đưa ra quyết định phù hợp với bối cảnh làm việc hiện đại.

Cụ thể:

- So sánh hiệu suất làm việc, mức độ hài lòng và thu nhập của lập trình viên theo ba mô hình làm việc:
  - Làm việc từ xa (remote)
  - Làm việc kết hợp (hybrid)
  - Làm việc tại văn phòng (in-person)
- Xác định các nguyên nhân chính ảnh hưởng đến hiệu suất làm việc từ xa, bao gồm:
  - Độ phức tạp công nghệ
  - Công nghệ & hệ thống
  - Quản lý công việc
  - Bảo mật và cập nhật
  - Khác
- Đưa ra khuyến nghị cải thiện mô hình làm việc từ xa, góp phần nâng cao hiệu quả cá nhân và tổ chức.
- Hỗ trợ ra quyết định cho hai nhóm đối tượng chính**:**
  - Doanh nghiệp công nghệ: lựa chọn mô hình làm việc tối ưu, xây dựng chính sách nhân sự hiệu quả, phù hợp với từng nhóm nhân viên.
  - Sinh viên năm cuối ngành CNTT: cần hiểu rõ sự khác biệt giữa các mô hình làm việc để lựa chọn hướng đi phù hợp, hiệu quả và bền vững..

1.2.2 PHẠM VI DỰ ÁN:

- Phạm vi không gian: nghiên cứu được thực hiện dựa trên dữ liệu khảo sát toàn cầu từ **Stack Overflow Developer Survey 2024**, tập trung phân tích nhóm lao động trong lĩnh vực công nghệ thông tin, đặc biệt là các lập trình viên và chuyên gia công nghệ đang làm việc theo hình thức từ xa (remote), kết hợp (hybrid) và tại văn phòng (in-person).
- Phạm vi thời gian:

- Thời gian nghiên cứu: Từ tháng 5/2025 đến tháng 8/2025.
- Thời gian lấy dữ liệu: Dữ liệu khảo sát được thu thập trong năm 2024.

1.3 SẢN PHẨM MỤC TIÊU

1.3.1 MÔ HÌNH PHÂN TÍCH DỮ LIỆU

- **Mô hình phân tích hiệu suất theo hình thức làm việc:** mô hình này so sánh hiệu suất (qua mức độ hài lòng, thu nhập) của lập trình viên giữa các hình thức làm việc: từ xa, kết hợp và tại văn phòng.
- **Mô hình xác định yếu tố ảnh hưởng đến hiệu suất làm việc từ xa:** sử dụng phương pháp phân nhóm dữ liệu và biểu đồ Pareto để xác định các nguyên nhân phổ biến khiến nhân viên làm việc kém hiệu quả khi làm việc từ xa.
- **Mô hình phân tích theo nhóm đối tượng:** tập trung phân tích hiệu suất làm việc theo vai trò (DevType), quy mô công ty (OrgSize), và khu vực địa lý (Country) để phát hiện sự khác biệt giữa các nhóm.

1.3.2 DASHBOARD & BIỂU ĐỒ

- Dashboard 1: hiển thị phân bố lập trình viên theo hình thức làm việc, vai trò, quốc gia, quy mô công ty.
- Dashboard 2*:* so sánh mức độ hài lòng và thu nhập trung bình giữa Remote, Hybrid, In-person
- Dashboard 3: phân tích mối quan hệ giữa sự hài lòng và mức lương theo từng nhóm đối tượng.
- Dashboard 4: xác định các nhóm làm việc từ xa hiệu quả hoặc cần hỗ trợ thêm.
- Dashboard 5: phân tích tác động tiêu cực của số lượng công cụ, độ phức tạp hệ thống đến hiệu suất remote.
- Biểu đồ Pareto*:* thống kê và sắp xếp các nguyên nhân phổ biến ảnh hưởng đến hiệu suất làm việc từ xa.

1.3.3 SCRIPT XỬ LÝ VÀ LÀM SẠCH DỮ LIỆU

Viết bằng Python (pandas, numpy,..) để thực hiện:

- Làm sạch dữ liệu: loại giá trị thiếu, chuẩn hóa định dạng, xử lý outliers
- Chuyển đổi dữ liệu: đổi đơn vị tiền tệ, mã hóa dữ liệu định tính, gom nhóm dữ liệu
- Tách bảng phân tích: xây dựng mô hình dữ liệu theo dạng Star Schema phục vụ trực quan hóa

1.3.4 BÁO CÁO VÀ SẢN PHẨM TRÌNH BÀY

- **Báo cáo Word:** tổng hợp toàn bộ quy trình và kết quả phân tích
- **Slide trình bày:** thiết kế bằng Canva, phục vụ báo cáo, trình bày kết quả trước hội đồng
- **Câu chuyện dữ liệu (data storytelling):** diễn giải theo 5 bước (Bối cảnh – Mâu thuẫn – Dữ liệu – Insight – Kết luận.), hướng đến đối tượng doanh nghiệp công nghệ và sinh viên CNTT.

1.4 LẬP KẾ HOẠCH DỰ ÁN

STT	HẠNG MỤC	BẮT ĐẦU	KẾT THÚC	KẾT QUẢ
1	Giới thiệu dự án	18/05/2025	20/05/2025	Xác định đề tài, hiện trạng nguồn dữ liệu
2	Phân tích yêu cầu	20/05/2025	01/06/2025	Hiểu rõ vấn đề, xây dựng định hướng cho dự án
2.1	Câu chuyện dữ liệu	16/06/2025	20/06/2025	Xây dựng câu chuyện giả định, giả thuyết phân tích, đối tượng tiếp nhận báo cáo
3	Tìm kiếm, nghiên cứu, khảo sát và thu thập dữ liệu	20/05/2025	06/06/2025	Hoàn thành
3.1	Kiểm thử và kiểm tra chất lượng dữ liệu	16/05/2025	20/06/2025	Đảm bảo chất lượng dữ liệu đạt độ tin cậy

4	Làm sạch và chuyển đổi dữ liệu	06/06/2025	16/06/2025	Dữ liệu tương đối sạch, sẵn sàng để xử lý
4.1	Xử lý dữ liệu	16/06/2025	20/06/2025	Hoàn thiện việc xử lý các giá trị ngoại lai, chuẩn hóa định dạng dữ liệu theo cấu trúc thống nhất
4.2	Mô hình hóa	20/6/2025	18/07/2025	Thiết kế cấu trúc dữ liệu theo mô hình Star Schema gồm 5 bảng
5	Xây dựng báo cáo	20/06/2025	10/07/2025	
5.1	Trực quan hóa dữ liệu	20/06/2025	22/07/2025	Xác định mục tiêu trực quan & chọn biểu đồ phù hợp với dữ liệu phân tích
5.2	Dashboard và Report	20/06/2025	25/07/2025	Thiết kế trực quan rõ ràng với màu sắc hài hòa, nhãn và tiêu đề dễ hiểu
6	Viết báo cáo	25/05/2025	10/08/2025	
6.1	Tạo Slide thuyết trình	27/06/2025	10/08/2025	Slide hoàn chỉnh
6.2	Viết document	25/05/2025	10/08/2025	Báo cáo hoàn chỉnh

Bảng 1

2 TỔNG QUAN DỮ LIỆU VÀ CÔNG NGHỆ

2.1 MÔ TẢ DỮ LIỆU:

2.1.1 THÔNG TIN BỘ DỮ LIỆU:

Dự án sử dụng bộ dữ liệu **Stack Overflow Developer Survey 2024**, một cuộc khảo sát thường niên quy mô lớn dành cho lập trình viên và chuyên gia công nghệ trên toàn thế giới. Đây là nguồn dữ liệu uy tín, được thu thập từ hơn 70.000 người tham gia thuộc nhiều quốc gia và lĩnh vực khác nhau trong ngành công nghệ.

Bộ dữ liệu cung cấp thông tin chi tiết về nhiều khía cạnh liên quan đến hiệu suất làm việc, bao gồm:

- Hình thức làm việc: tại văn phòng, kết hợp, từ xa hoàn toàn
- Mức độ hài lòng với công việc (JobSat)
- Thu nhập hàng năm (CompTotalUSD)
- Loại hình công việc (DevType)
- Quốc gia (Country)
- Kinh nghiệm làm việc (YearsCodePro)
- Các yếu tố khác: trình độ học vấn, ngôn ngữ lập trình, công nghệ sử dụng, kỹ năng kỹ thuật,...

Dữ liệu được cung cấp dưới định dạng csv gồm hai tệp:

- **survey_results_public.csv:** chứa dữ liệu khảo sát gốc
- **survey_results_schema.csv:** mô tả chi tiết các trường dữ liệu, giúp tra cứu và xử lý chính xác

Bộ dữ liệu này là cơ sở để nhóm tiến hành phân tích mối liên hệ giữa hình thức làm việc từ xa và hiệu suất lao động trong môi trường công nghệ hiện đại.

2.1.2 CÁC KHÁI NIỆM:

Để hiểu rõ hơn về dữ liệu và phân tích, dưới đây là các khái niệm chính:

- **Hình thức làm việc (RemoteWork)**
  - **Remote:** làm việc hoàn toàn từ xa
  - **Hybrid:** kết hợp từ xa và tại văn phòng
  - **In-Person:** làm việc trực tiếp toàn thời gian tại công ty
- **Mức độ hài lòng (Job Satisfaction)**: thể hiện cảm nhận cá nhân về công việc hiện tại.
- **Frustration:** các lý do gây khó chịu khi làm việc
- **Thu nhập hàng năm:** tổng thu nhập cá nhân trong một năm (CompTotal), được quy đổi về USD để thống nhất so sánh.
- **Kinh nghiệm làm việc (Years of Experience)**: số năm làm việc trong ngành công nghệ thông tin, bao gồm chính thức và học tập cá nhân (YearsCodePro).
- **Nhóm nghề nghiệp (DevType)**: chức danh hoặc lĩnh vực công việc chính của người tham gia: Backend Developer, Full-Stack, Data Scientist,...

2.1.3 CÁC TRƯỜNG DỮ LIỆU:

Bộ dữ liệu được sử dụng trong nghiên cứu gồm 32 trường thông tin, được trích xuất từ **Stack Overflow Developer Survey 2024**. Mỗi trường dữ liệu phản ánh một khía cạnh về thông tin cá nhân, môi trường làm việc, kỹ năng, mức độ hài lòng và trải nghiệm của người tham gia khảo sát. Chi tiết như sau:

STT	Tên trường dữ liệu	Mô tả	Kiểu dữ liệu
1	ResponseId	Mã định danh duy nhất cho mỗi người tham gia khảo sát	Số nguyên
2	MainBranch	Vai trò chính của người tham gia trong ngành CNTT	Chuỗi
3	Age	Tuổi của người tham gia	Số
4	Country	Quốc gia đang sinh sống	Chuỗi
5	EdLevel	Trình độ học vấn	Chuỗi
6	WorkExp	Số năm kinh nghiệm làm việc	Số
7	CompTotal	Tổng thu nhập hàng năm	Số
8	Currency	Đơn vị tiền tệ của thu nhập	Chuỗi

9	RemoteWork	Hình thức làm việc (remote, hybrid, in-person)	Chuỗi
10	Employment	Tình trạng việc làm	Chuỗi
11	OrgSize	Quy mô tổ chức	Chuỗi
12	JobSat	Mức độ hài lòng với công việc	Số
13	TimeSearching	Thời gian tìm kiếm thông tin	Số
14	TimeAnswering	Thời gian trả lời câu hỏi	Số
15	Frustration	Nguyên nhân gây khó khăn/kém hiệu suất	Chuỗi
16	SurveyLength	Cảm nhận về độ dài khảo sát	Chuỗi
17	SurveyEase	Cảm nhận về mức độ dễ trả lời khảo sát	Chuỗi
18	YearsCode	Số năm lập trình (bao gồm cả khi học)	Số
19	YearsCodePro	Số năm lập trình chuyên nghiệp	Số
20	DevType	Loại công việc phát triển phần mềm đảm nhiệm	Chuỗi

21	BuildvsBuy	Xu hướng tự xây dựng hay mua giải pháp	Chuỗi
22	TechEndorse	Công nghệ được đánh giá cao	Chuỗi
23	NEWCollabToolsHaveWorkedWith	Công cụ cộng tác đã từng sử dụng	Chuỗi
24	NEWCollabToolsWantToWorkWith	Công cụ cộng tác muốn sử dụng	Chuỗi
25	OpSysProfessional use	Hệ điều hành sử dụng cho công việc	Chuỗi
26	OfficeStackAsyncHaveWorkedWith	Bộ công cụ làm việc không đồng bộ đã sử dụng	Chuỗi
27	OfficeStackSyncHaveWorkedWith	Bộ công cụ làm việc đồng bộ đã sử dụng	Chuỗi
28	SOComm	Mức độ tham gia cộng đồng Stack Overflow	Chuỗi
29	AISelect	Công cụ AI đã lựa chọn	Chuỗi
30	AISearchDevHaveWorkedWith	Công cụ AI đã từng sử dụng cho phát triển phần mềm	Chuỗi
31	TBranch	Nhánh công việc trong công nghệ	Chuỗi

32	ICorPM	Vai trò cá nhân đóng góp (Individual Contributor) hoặc quản lý (Project Manager)	Chuỗi

Bảng 2

2.2 CÁC CÔNG CỤ, THƯ VIỆN, CÔNG NGHỆ SỬ DỤNG:

Trong quá trình thực hiện dự án "Phân tích hiệu suất làm việc từ xa", nhóm đã sử dụng các công cụ và công nghệ sau:

- Ngôn ngữ và công cụ xử lý dữ liệu:
  - Python với thư viện pandas (xử lý bảng dữ liệu, Excel), re (xử lý chuỗi), numpy (tính toán, NaN).
  - Microsoft Excel: kiểm tra, làm sạch dữ liệu ban đầu, thống kê mô tả.
- Công cụ trực quan hóa:
  - Tableau Public, Excel: tạo dashboard tương tác, biểu đồ phân tích theo nhiều chiều.
- Công cụ trình bày và báo cáo:
  - Canva: thiết kế slide hỗ trợ kể chuyện dữ liệu.
  - Microsoft Word: soạn thảo báo cáo.

2.3 TỔNG QUAN VỀ MÔ HÌNH VÀ THUẬT TOÁN LIÊN QUAN:

- Mô hình xử lý dữ liệu

  - Chuẩn hóa toàn bộ dữ liệu văn bản và số để đảm bảo đồng nhất, loại bỏ sai lệch định dạng.
  - Phát hiện và xử lý giá trị thiếu, ngoại lai (outlier) bằng **IQR** – phương pháp phù hợp với dữ liệu lệch phân phối, giúp loại bỏ ảnh hưởng của giá trị cực đoan; thay thế bằng giá trị trung bình để giữ tính liên tục.
  - Chuẩn hóa đơn vị tiền tệ về USD, tạo cơ sở so sánh nhất quán.

- Thuật toán và phương pháp phân tích

  - Thống kê mô tả: tỷ lệ, giá trị trung bình.
  - Phân tích so sánh: so sánh hiệu suất và mức độ hài lòng giữa các hình thức làm việc (Remote, Hybrid, In-person).
  - Biểu đồ Pareto: xác định các yếu tố chính ảnh hưởng đến hiệu suất.
  - Trực quan hóa dữ liệu đa chiều bằng Tableau để hỗ trợ ra quyết định.

2.4 CÂU CHUYỆN DỮ LIỆU:

2.4.1 BỐI CẢNH

Trong ngành công nghệ thông tin, xu hướng làm việc từ xa (remote work) đang ngày càng phát triển mạnh mẽ và trở thành một lựa chọn chiến lược trong cách vận hành của nhiều doanh nghiệp hiện đại. Song song với đó, các mô hình làm việc kết hợp (hybrid) và tại văn phòng (in-person) vẫn tiếp tục được duy trì, tạo nên sự đa dạng trong môi trường làm việc và cách tổ chức nhân sự. Mỗi

mô hình làm việc đều mang lại những lợi ích riêng:

- **Remote**: Linh hoạt thời gian, tiết kiệm chi phí, vượt rào cản địa lý.
- **In-person**: Tăng cường gắn kết, dễ quản lý trực tiếp.
- **Hybrid**: Kết hợp sự linh hoạt với kết nối đội nhóm.

Tuy nhiên, hiệu suất giữa các mô hình không giống nhau và bị chi phối bởi các yếu tố như **kinh nghiệm làm việc** hay **mức độ hỗ trợ từ công ty.** Điều này đặt ra hai câu hỏi lớn:

1. Mô hình nào mang lại hiệu quả cao nhất?
2. Nguyên nhân nào làm giảm hiệu suất khi làm việc từ xa?

Để trả lời, nhóm đã phân tích **Stack Overflow Developer Survey 2024** – hơn 70.000 phản hồi từ lập trình viên toàn cầu.

2.4.2 VẤN ĐỀ VÀ NHU CẦU RA QUYẾT ĐỊNH

Mỗi mô hình làm việc có ưu – nhược điểm riêng:

- **In-Person**: Quản lý thuận tiện, nhưng kém linh hoạt và tốn thời gian di chuyển.
- **Hybrid**: Cân bằng kết nối và tự do cá nhân, nhưng khó đồng bộ quy trình.
- **Remote**: Tự do, tiết kiệm chi phí, nhưng dễ cô lập, khó giám sát, nhất là với nhân sự mới.

Doanh nghiệp cần **dữ liệu thực tế** để chọn mô hình tối ưu, giữ chân nhân tài và xây dựng chính sách phù hợp. Sinh viên CNTT sắp ra trường cũng cần hiểu

rõ để chọn môi trường làm việc phù hợp.

Những rào cản lớn nhất khi làm việc từ xa gồm:

- Độ phức tạp công nghệ
- Công nghệ & hệ thống
- Quản lý công việc

2.4.3 KHAI PHÁ DỮ LIỆU

Bộ dữ liệu **Stack Overflow Developer Survey 2024** gồm hơn 70.000 phản hồi, chứa các trường quan trọng như:

- Hình thức làm việc (RemoteWork)
- Mức độ hài lòng (JobSat)
- Thu nhập (CompTotal)
- Loại công việc (DevType)
- Quốc gia (Country)
- Số năm kinh nghiệm (YearsCodePro)

Quy trình xử lý:

- **Python**: loại bỏ ngoại lai (IQR), chuẩn hóa dữ liệu, mã hóa biến.
- **Tableau**: Trực quan hóa phân tích, tạo biểu đồ tương tác.
- **Word & Canva**: Trình bày báo cáo rõ ràng, trực quan.

2.4.4 KHÁM PHÁ INSIGHT CHÍNH

Từ quá trình phân tích bộ dữ liệu **Stack Overflow Developer Survey 2024**, nhóm nghiên cứu đã rút ra một số phát hiện quan trọng liên quan đến hiệu quả của các mô hình làm việc trong ngành công nghệ thông tin:

- Làm việc từ xa mang lại lợi ích rõ rệt về thu nhập và mức độ hài lòng, đặc biệt đối với những người có kinh nghiệm lâu năm và kỹ năng tự quản lý tốt. Đây là nhóm có khả năng thích nghi cao với môi trường làm việc độc lập và công nghệ hỗ trợ.
- Nhân sự mới hoặc sinh viên mới tốt nghiệp thường gặp khó khăn khi làm việc từ xa, do thiếu kinh nghiệm trong giao tiếp nhóm, khả năng tự định hướng và thích nghi với quy trình làm việc không trực tiếp.
- Hiệu quả của từng mô hình làm việc chịu ảnh hưởng bởi nhiều yếu tố cá nhân và môi trường, bao gồm vị trí công việc, số năm kinh nghiệm chuyên môn và quốc gia làm việc. Những yếu tố này tạo ra sự khác biệt rõ rệt trong cách nhân viên phản ứng với từng mô hình.
- Các rào cản chính ảnh hưởng đến hiệu suất làm việc từ xa bao gồm:
  - Độ phức tạp của hệ thống công nghệ
  - Công cụ hỗ trợ chưa tối ưu
  - Cách quản lý công việc từ xa chưa hiệu quả
  - Quy trình cập nhật bảo mật chưa đồng bộ
- Mức độ hài lòng với môi trường làm việc có mối liên hệ chặt chẽ với hiệu suất công việc. Nhân viên cảm thấy hài lòng thường duy trì hiệu suất ổn định, bất kể họ làm việc từ xa, kết hợp hay tại văn phòng.

- Độ phức tạp của công nghệ là yếu tố ảnh hưởng mạnh nhất đến hiệu suất. Khi hệ thống công nghệ trở nên quá phức tạp, cả hiệu suất làm việc lẫn mức độ hài lòng đều có xu hướng giảm đáng kể.

2.4.5 THÔNG ĐIỆP VÀ HÀNH ĐỘNG

Thông điệp chính: làm việc từ xa có thể mang lại hiệu quả cao, nhưng không phù hợp với tất cả mọi người. Cần lựa chọn mô hình dựa trên kinh nghiệm, kỹ năng và đặc thù công việc cụ thể.

Ba nguyên nhân chính gây giảm hiệu suất làm việc từ xa gồm: độ phức tạp của công nghệ, hệ thống công nghệ và quản lý công việc. Chúng chiếm đến 76% tổng số các tác động tiêu cực. Điều này khẳng định việc ưu tiên giải quyết các vấn đề liên quan đến công nghệ và quy trình quản lý sẽ mang lại hiệu quả vượt trội, tạo ra sự thay đổi lớn nhất.

Khuyến nghị:

- **Doanh nghiệp**:
  - Đơn giản hóa hệ thống, giảm rào cản kỹ thuật.
  - Đào tạo & cung cấp tài liệu hướng dẫn.
  - Phân loại nhân sự để áp dụng mô hình phù hợp.
  - Cân nhắc hybrid như giải pháp cân bằng.
  - Cá nhân hóa chính sách, tránh áp dụng chung.

- **Sinh viên CNTT**:

  - Tìm hiểu kỹ đặc điểm của từng mô hình làm việc (remote, hybrid, in-person) để lựa chọn môi trường phù hợp với năng lực cá nhân và định hướng nghề nghiệp.
  - Nếu định hướng theo mô hình làm việc từ xa, cần chủ động rèn luyện các kỹ năng thiết yếu như:
    - Tự học và cập nhật kiến thức
    - Quản lý thời gian và công việc cá nhân
    - Giao tiếp và cộng tác hiệu quả qua môi trường trực tuyến

2.4.6 TRÌNH BÀY DỮ LIỆU

Để truyền tải câu chuyện dữ liệu hiệu quả, nhóm sử dụng các loại biểu đồ trực quan và dễ hiểu sau:

- **Biểu đồ cột (Bar chart)**: so sánh mức độ hài lòng, thu nhập và hiệu suất giữa các mô hình làm việc (remote – hybrid – In-person).
- **Biểu đồ tròn (Pie chart)**: thể hiện tỷ lệ lựa chọn mô hình làm việc hoặc mức độ hài lòng theo nhóm.
- **Biểu đồ phân tán (Scatter plot)**: phân tích mối liên hệ giữa kinh nghiệm, thu nhập và hình thức làm việc.
- **Biểu đồ đường (Line chart)**: theo dõi xu hướng làm việc từ xa theo năm, quốc gia hoặc độ tuổi.
- **Bản đồ nhiệt (Heatmap)**: minh họa mức độ hài lòng hoặc hiệu suất theo từng khu vực hoặc nhóm nghề nghiệp.
- **Biểu đồ Pareto:** các yếu tố ảnh hưởng đến hiệu suất làm việc từ xa.

Tất cả biểu đồ được xây dựng bằng Tableau, Excel đảm bảo sinh động, dễ tiếp cận và hỗ trợ truyền tải thông điệp nhanh chóng.

2.4.7 NHỮNG ĐIỀU CẦN LƯU Ý

- **Đảm bảo tính trung thực và khách quan** khi xử lý và trình bày dữ liệu, tránh thiên vị mô hình làm việc nào.
- **Lọc và làm sạch dữ liệu kỹ lưỡng** trước khi phân tích để loại bỏ giá trị thiếu, sai lệch hoặc nhiễu.
- **Chọn biểu đồ phù hợp với loại dữ liệu** để tránh gây hiểu lầm cho người xem.
- **Trực quan hóa cần rõ ràng, dễ hiểu**, hạn chế lạm dụng màu sắc hoặc hiệu ứng gây rối mắt.
- **Bảo vệ thông tin cá nhân**, không để lộ dữ liệu nhạy cảm từ bộ khảo sát gốc.
- **Truyền tải thông điệp đơn giản – rõ ràng**, tránh đưa quá nhiều thông tin gây loãng nội dung

3 QUY TRÌNH XỬ LÝ DỮ LIỆU:

3.1 KHÁM PHÁ VÀ PHÂN TÍCH DỮ LIỆU (EDA):

Giai đoạn khám phá dữ liệu (Exploratory Data Analysis – EDA) giúp nhóm hiểu rõ hơn về cấu trúc dữ liệu, phát hiện các điểm bất thường, xu hướng và mối quan hệ giữa các biến. Một số nội dung chính đã được thực hiện trong giai đoạn này gồm:

3.1.1 KHÁM PHÁ CẤU TRÚC DỮ LIỆU

- Tập dữ liệu gốc từ khảo sát **Stack Overflow Developer Survey 2024** có hơn **80 trường dữ liệu** và hàng chục nghìn dòng.
- Nhóm tập trung chọn lọc các trường liên quan trực tiếp đến đề tài, bao gồm:

  - RemoteWork: hình thức làm việc
  - JobSat: mức độ hài lòng với công việc
  - CompTotalUSD: mức lương quy đổi theo USD
  - Frustration: các yếu tố gây ảnh hưởng hiệu suất làm việc từ xa
  - DevType, OrgSize, Country: vai trò, quy mô tổ chức và quốc gia

3.1.2 THỐNG KÊ MÔ TẢ

- **Dữ liệu định lượng**: thực hiện thống kê như trung bình, trung vị, min, max, độ lệch chuẩn đối với trường CompTotalUSD.
- **Dữ liệu định tính**: đếm tần suất xuất hiện của các giá trị trong RemoteWork, JobSat, DevType,...

3.1.3 XỬ LÝ DỮ LIỆU THIẾU VÀ BẤT THƯỜNG

- Loại bỏ các dòng chứa giá trị thiếu ở các trường quan trọng
- Lọc bỏ các dòng có lương quá cao hoặc bằng 0 để tránh sai lệch (outliers).
- Chuẩn hóa tên các nhóm vai trò và hình thức làm việc.

3.1.4 PHÂN TÍCH SƠ BỘ MỐI QUAN HỆ GIỮA CÁC BIẾN

- So sánh mức độ hài lòng và mức thu nhập theo từng hình thức làm việc.
- Nhận diện mối tương quan giữa mức lương và mức độ hài lòng.
- Phân tích tỷ lệ xuất hiện các nguyên nhân gây ảnh hưởng hiệu suất từ xa từ trường Frustration.

3.1.5 TRỰC QUAN HÓA EDA

- Dùng biểu đồ cột, tròn, biểu đồ phân tán để hỗ trợ đánh giá sơ bộ xu hướng.
- Một số biểu đồ đã thực hiện trong quá trình EDA:

  - Phân bố mức lương trung bình theo hình thức làm việc
  - Tỷ lệ hài lòng theo hình thức làm việc
  - Phân bố vai trò lập trình viên quy mô tổ chức.

3.2 LÀM SẠCH VÀ CHUẨN HÓA DỮ LIỆU:

Quá trình xử lý dữ liệu trong dự án bao gồm hai bước chính:

- **Làm sạch dữ liệu**: xử lý các giá trị bị thiếu, lỗi định dạng và loại bỏ dữ liệu không hợp lệ.
- **Chuyển đổi dữ liệu**: chuẩn hóa định dạng, mã hóa giá trị, phân loại và chọn lọc trường dữ liệu phù hợp với mục tiêu phân tích.

3.2.1 CHUẨN BỊ DỮ LIỆU

Để tránh ảnh hưởng đến dữ liệu gốc khi xử lý, nhóm xây dựng lớp

DataCleaner với hàm khởi tạo sao chép toàn bộ **DataFrame** đầu vào. Việc này giúp đảm bảo dữ liệu gốc được giữ nguyên, hỗ trợ việc làm sạch và biến đổi dữ liệu một cách an toàn.

```
class DataCleaner:
    def __init__(self, df: pd.DataFrame):
        self.df = df.copy()
```

3.2.2 GIẢI PHÁP LƯU TRỮ DỮ LIỆU

Trong quá trình thực hiện dự án, nhóm đã cân nhắc giữa hai giải pháp lưu trữ phổ biến:

- **Nền tảng đám mây (Cloud-based):** mang lại tính linh hoạt cao, dễ dàng chia sẻ, đồng bộ dữ liệu nhanh giữa các thành viên. Bảo mật được đảm bảo bởi các nhà cung cấp lớn (Google Drive, OneDrive). Đặc biệt phù hợp với mô hình làm việc nhóm từ xa – xu hướng mà chính đề tài đang phân tích.
- **Ứng dụng tại chỗ (On-premise):** giúp kiểm soát toàn bộ dữ liệu nội bộ nhưng đòi hỏi chi phí triển khai cao, hạn chế trong việc mở rộng và khó truy cập từ xa – điều này không thuận tiện khi nhóm làm việc phân tán.

Nhóm quyết định sử dụng giải pháp lưu trữ đám mây (Google Drive và GitHub) nhằm:

- Đảm bảo khả năng truy cập dữ liệu mọi lúc, mọi nơi
- Hỗ trợ làm việc nhóm hiệu quả: các thành viên có thể truy cập dễ dàng,...
- Dễ dàng cập nhật, chia sẻ báo cáo và mã nguồn

- Quản lý phiên bản minh bạch và lưu trữ an toàn trên GitHub

Việc đồng bộ tài liệu giữa Google Drive (dữ liệu thô) và GitHub (dashboard, báo cáo, file dữ liệu sau xử lý) giúp nhóm duy trì hiệu suất cao trong toàn bộ vòng đời dự án.

3.2.3 GIẢI PHÁP PHÂN BỐ DỮ LIỆU

3.2.3.1 Ý nghĩa việc phân bố dữ liệu

Việc phân bố dữ liệu đóng vai trò then chốt trong quá trình phân tích và trình bày thông tin. Dữ liệu được tổ chức hợp lý giúp:

- Tăng tính trực quan: biểu đồ và bảng số liệu rõ ràng, dễ hiểu hơn nhờ dữ liệu được sắp xếp theo nhóm, đặc điểm hoặc thời gian.
- Phân loại xu hướng chính xác: dễ dàng nhận diện sự khác biệt giữa các nhóm đối tượng, từ đó phát hiện các mô hình ẩn hoặc mối tương quan.
- Nâng cao độ chính xác phân tích: tránh nhầm lẫn hoặc sai lệch khi xử lý và diễn giải dữ liệu.
- Hỗ trợ quyết định hiệu quả: giúp so sánh, đánh giá và đưa ra giải pháp một cách nhanh chóng, đặc biệt trong các báo cáo hiệu suất và dự báo.

3.2.3.2 Trình bày cách phân bố dữ liệu

Trong dự án, nhóm sử dụng cách phân bố dữ liệu rõ ràng và hiệu quả, đảm bảo truy cập linh hoạt và dễ phối hợp làm việc nhóm:

- **Dữ liệu thô:** lưu trữ trên Google Drive, phục vụ mục đích đối chiếu và truy xuất nguyên bản khi cần.

- **Dữ liệu đã xử lý, mã nguồn phân tích, biểu đồ và báo cáo:** được đồng bộ và lưu trữ tập trung trên GitHub để tiện theo dõi, cập nhật và kiểm soát phiên bản.
- **Phân quyền rõ ràng trên GitHub:** tất cả thành viên đều được tham gia vào kho lưu trữ và có quyền truy cập đầy đủ. Việc phân quyền và theo dõi tiến độ không dựa trên vai trò cố định, mà nhằm đảm bảo mọi thành viên đều nắm rõ tình hình chung, có thể kiểm tra, cập nhật nội dung ở mọi giai đoạn (xử lý dữ liệu, phân tích, trực quan hóa và viết báo cáo).
- **Trao đổi nhóm:** được thực hiện qua Zalo để phản hồi nhanh, còn mọi thay đổi chính thức đều được cập nhật trên GitHub nhằm đồng bộ hóa toàn bộ tiến độ và nội dung công việc.

3.2.4 LÀM SẠCH DỮ LIỆU

3.2.4.1 Các vấn đề ảnh hưởng tới dữ liệu

Trong quá trình xử lý khảo sát **Stack Overflow 2024**, nhóm gặp nhiều thách thức ảnh hưởng đến chất lượng dữ liệu như:

- **Thiếu dữ liệu**: nhiều trường bị bỏ trống (NA)
- **Không đồng nhất**: cùng nội dung nhưng khác định dạng, gây khó phân loại.
- **Giá trị bất thường**: dữ liệu sai lệch như lương cao bất hợp lý hoặc chọn nhiều hình thức làm việc không rõ ràng.
- **Dữ liệu tự do, rời rạc**: các câu trả lời văn bản như vị trí công việc, lý do chọn hình thức làm việc khó tổng hợp và trực quan.

Những vấn đề trên yêu cầu xử lý kỹ càng để đảm bảo phân tích chính xác và kể chuyện dữ liệu hiệu quả.

3.2.4.2 Các tiêu chí đánh giá chất lượng dữ liệu

Để đảm bảo dữ liệu phục vụ phân tích hiệu suất làm việc từ xa có chất lượng tốt, nhóm áp dụng các tiêu chí sau:

- Độ chính xác (Accuracy): Dữ liệu phải phản ánh đúng thực tế, tránh nhập sai hoặc thông tin không hợp lý.
- Tính đầy đủ (Completeness): Các trường dữ liệu cần thiết không bị thiếu (đặc biệt là về mô hình làm việc, mức độ hài lòng, thu nhập,…).
- Tính cập nhật (Timeliness): Dữ liệu phải phản ánh tình hình hiện tại, đặc biệt do môi trường làm việc thay đổi nhanh sau đại dịch.
- Tính nhất quán (Consistency): Tránh các biểu diễn khác nhau cho cùng một giá trị
- Tính độc nhất (Uniqueness): Không có bản ghi trùng lặp gây sai lệch khi thống kê.
- Tính liên quan và chi tiết (Relevance & Granularity): Chỉ giữ lại những dữ liệu phù hợp với mục tiêu phân tích, tránh nhiễu.

Trong phạm vi dự án này, nhóm đặc biệt chú trọng đến 4 tiêu chí chính:

- Tính đầy đủ: do nhiều trường dữ liệu bị bỏ trống.
- Tính nhất quán: do dữ liệu đầu vào đến từ nhiều biểu mẫu tự do.
- Độ chính xác: để loại bỏ các giá trị bất thường, không hợp lý.
- Tính liên quan và chi tiết: để lọc ra những thông tin thực sự hữu ích cho

mục tiêu phân tích hiệu suất làm việc từ xa.

3.2.5 CÁC BƯỚC LÀM SẠCH DỮ LIỆU

3.2.5.1 Trình bày các bước làm sạch trong phạm vi dự án

Gồm các bước loại bỏ lỗi, xử lý trống, chuẩn hóa văn bản:

- **clean_text:** loại bỏ khoảng trắng thừa, ép kiểu chuỗi.

```
def clean_text(self, text: str) -> str:
    if pd.isnull(text):
        return ""
    text = str(text).strip()
    text = re.sub(r'\s+', ' ', text)
    return text
```

- **clean_comptotal:** xử lý cột ComptotalUSD chỉ giữ lại giá trị số

```
def clean_comptotal(
    self,
    col: str = "CompTotalUSD",
    new_col: str | None = None,
    remove_outliers: bool = True,
    replace_with_mean: bool = True,
    min_value: float | None = None  # loại các giá trị quá nhỏ nếu muốn
) -> "DataCleaner":
    new_col = new_col or col
```

```
    def to_float(x):
        if pd.isnull(x):
            return np.nan
        cleaned = re.sub(r"[^0-9.\-]", "", str(x))
        try:
            return float(cleaned) if cleaned else np.nan
        except ValueError:
            return np.nan

    # Áp dụng hàm chuyển đổi
    self.df[new_col] = self.df[col].apply(to_float)

    s = self.df[new_col]
```

- Loại bỏ outlier bằng IQR, lọc giá trị nhỏ hơn `min_value`.
- Phần replace with mean để điền giá trị outlier bằng **mean**

```
    # Loại bỏ giá trị nhỏ hơn ngưỡng (nếu min_value được cung cấp)
    if min_value is not None:
        self.df.loc[s < min_value, new_col] = np.nan

    # loại bỏ outliers bằng IQR
    if remove_outliers:
        s = self.df[new_col]
        q1, q3 = s.quantile([0.25, 0.75])
        iqr = q3 - q1
```

```
            lower_bound = q1 - 1.5 * iqr
            upper_bound = q3 + 1.5 * iqr
            mask = (s < lower_bound) | (s > upper_bound)

            if replace_with_mean:
                mean_value = s[~mask].mean()
                self.df.loc[mask, new_col] = mean_value
            else:
                self.df.loc[mask, new_col] = np.nan

        return self
```

- **clean_semicolon_columns:** làm sạch cột chứa nhiều giá trị phân cách bằng dấu “;”

```
    def clean_semicolon_columns(self, cols):
        for col in cols:
            self.df[col] = self.df[col].apply(
                    lambda x: '-'.join([item.strip() for item in str(x).split(';') if item.strip()])
                if pd.notnull(x) else "")
        return self
```

- **clean_SOComm_text**: rút gọn phản hồi tự do từ người dùng **Stack Overflow.**

```
def clean_SOComm_text(self, col='SOComm'):
    mapping = {
        'Yes, definitely': 'Yes',
        'Yes, somewhat': 'Somewhat',
        'No, not really': 'No',
        'No, not at all': 'Strong No',
        'Neutral': 'Neutral'
    }
    self.df[col] = self.df[col].map(mapping)
    return self
```

- **clean_dev_type:** phân loại lập trình viên thành các nhóm tiêu chuẩn**.**

```
def clean_dev_type(self, col='DevType', new_col=None):
    new_col = new_col or col
    def map_type(text):
        if pd.isnull(text):
            return "Unknown"
        text = text.lower()
        if "full-stack" in text:
            return "Full-Stack Developer"
        elif "back-end" in text:
            return "Back-End Developer"
```

```
        elif "front-end" in text:
            return "Front-End Developer"
        elif "mobile" in text:
            return "Mobile Developer"
        elif "embedded" in text:
            return "Embedded Developer"
        elif "game" in text:
            return "Game Developer"
        elif "engineering manager" in text:
            return "Manager"
        elif "research" in text:
            return "Researcher"
        else:
            return "Other"
    self.df[new_col] = self.df[col].apply(map_type)
    return self
```

- **clean_buildvsbuy:** phân nhóm lựa chọn phát triển sản phẩm: Ready-to-go, Requires Customization, v.v.

```
def clean_buildvsbuy(self, col='BuildvsBuy', new_col=None):
    new_col = new_col or col
    def map_text(x):
        if pd.isnull(x):
            return "Unknown"
        x = x.strip().lower()
```

```
        if x.startswith("out-of-the-box"):
            return "Ready-to-go"
        elif "ready-to-go but also customizable" in x:
            return "Ready+Customizable"
        elif "customized and needs to be engineered" in x:
            return "Requires Customization"
        else:
            return "Other"
    self.df[new_col] = self.df[col].apply(map_text)
    return self
```

3.2.6 CHUYỂN ĐỔI DỮ LIỆU

3.2.6.1 Các kỹ thuật chuyển đổi

Dựa trên yêu cầu phân tích, nhóm đã áp dụng các kỹ thuật chuyển đổi dữ liệu phổ biến sau:

- Thay thế và ánh xạ giá trị: sử dụng replace() và map() trong Python để chuẩn hóa văn bản.
- Chuyển đổi kiểu dữ liệu: dùng astype(), pd.to_numeric() hoặc str.strip() để chuyển đổi giữa kiểu số và chuỗi một cách thống nhất.
- Gom nhóm (Binning): phân loại các giá trị liên tục như số năm kinh nghiệm thành các nhóm rời rạc dễ so sánh.
- Mã hóa thang điểm: chuyển các giá trị mang tính cảm nhận (như JobSat – mức độ hài lòng)
- Chuyển đổi đơn vị: đổi thời gian từ giờ sang phút và quy đổi tiền tệ về USD nhằm tăng tính đồng nhất khi phân tích.

Trong phạm vi dự án, nhóm ưu tiên sử dụng các kỹ thuật đơn giản nhưng hiệu quả như chuẩn hóa định dạng, ánh xạ giá trị, tạo biến mới và chuyển đổi đơn vị – giúp dữ liệu dễ xử lý và phân tích hơn.

3.2.6.2 Trình bày các phép chuyển đổi trong dự án

Chuyển đổi định dạng dữ liệu, đơn vị, hoặc dạng biểu diễn:

- **convert_currency_to_usd:** đổi đơn vị tiền tệ về USD.

```
        def convert_currency_to_usd(self, amount_col='CompTotal', currency_col='Currency', new_col='CompTotalUSD'):
    rates = {
    "USD": 1.00,
    "EUR": 1.10,
    "GBP": 1.30,
    "INR": 0.012,
    "VND": 0.000042,
    "JPY": 0.0068,
    "CNY": 0.14,
    "CAD": 0.73,
    "AUD": 0.66,
    "BRL": 0.19,
    "NOK": 0.093,
    "SEK": 0.094,
    "DKK": 0.15,
    "MXN": 0.058,
```

"PEN": 0.26,
"LKR": 0.0031,
"UAH": 0.025,
"ILS": 0.27,
"PLN": 0.26,
"ZAR": 0.0558,
"CHF": 1.12,
"SGD": 0.74,
"HKD": 0.128,
"MYR": 0.21,
"THB": 0.027,
"IDR": 0.000063,
"KRW": 0.00072,
"EGP": 0.021,
"NGN": 0.00067,
"PKR": 0.0036,
"BDT": 0.0086,
"CZK": 0.043,
"HUF": 0.0027,
"RON": 0.22,
"SAR": 0.27,
"AED": 0.27,
"KWD": 3.25,
"BHD": 2.65,
"QAR": 0.27,

"OMR": 2.60,
"CLP": 0.0011,
"COP": 0.00026,
"ARS": 0.0009,
"UYU": 0.025,
"BOB": 0.14,
"PYG": 0.00014,
"DOP": 0.017,
"GTQ": 0.13,
"HNL": 0.041,
"NIO": 0.027,
"CRC": 0.0019,
"JMD": 0.0064,
"XOF": 0.0017,
"XAF": 0.0017,
"CDF": 0.00037,
"GHS": 0.084,
"TZS": 0.00039,
"KES": 0.0078,
"UGX": 0.00027,
"ETB": 0.017,
"MWK": 0.00059,
"ZMW": 0.046,
"MZN": 0.016,
"BWP": 0.073,

```
    "MUR": 0.022,
    "MAD": 0.10,
    "TND": 0.32,
    "DZD": 0.0074,
    "LBP": 0.000009,  # biến động mạnh
    "SYP": 0.000079,
    "IRR": 0.000024,
    "IQD": 0.00076,
    "AFN": 0.012,
    "NPR": 0.0075,
    "MMK": 0.00047,
    "MNT": 0.00029,
    "KZT": 0.0022,
    "UZS": 0.000082,
    "AZN": 0.59,
    "GEL": 0.36,
    "AMD": 0.0025,
    "ALL": 0.011,
    "MKD": 0.017,
    "ISK": 0.0073,
    "BAM": 0.55,
    "RSD": 0.0093,
    "BYN": 0.31,
    "RUB": 0.011,
    "VES": 0.028,
```

"BGN": 0.56,
"NZD": 0.62,
"GBP": 1.30,
"BRL": 0.19,
"IRR": 0.000024,
"SEK": 0.095,
"PLN": 0.25,
"CHF": 1.12,
"AUD": 0.67,
"HKD": 0.13,
"TWD": 0.031,
"PHP": 0.018,
"AED": 0.27,
"TRY": 0.032,
"NZD": 0.62,
"IDR": 0.000065,
"MGA": 0.00022,
"BGN": 0.56,
"RWF": 0.00081,
"TTD": 0.15,
"JOD": 1.41,
"AOA": 0.0012,
"KGS": 0.011,
"MOP": 0.12,
"TJS": 0.091,

```
        "TMT": 0.29,
        "CUP": 0.041,
        "BTN": 0.012,
        "MVR": 0.065


    }
                                                self.df[currency_col]              =
self.df[currency_col].astype(str).str.strip().str.upper()
                    self.df[amount_col]    =    pd.to_numeric(self.df[amount_col],
errors='coerce')

        def convert(row):
            currency = row[currency_col]
            amount = row[amount_col]
            rate = rates.get(currency)
            if rate is None or pd.isnull(amount):
                return None
            return amount * rate

        self.df[new_col] = self.df.apply(convert, axis=1)
        return self
```

- **clean_time_columns**: chuyển đổi thời gian khảo sát từ giờ sang phút.

```
def clean_time_columns(self, cols=['TimeSearching', 'TimeAnswering']):
    def convert_to_minutes(value):
        if pd.isnull(value):
            return None
        value = value.lower()
        if "less than 15" in value:
            return 10
        elif "15-30" in value:
            return 22
        elif "30-60" in value:
            return 45
        elif "60-120" in value:
            return 90
        elif "over 120" in value:
            return 150
        else:
            return None

    for col in cols:
        self.df[col] = self.df[col].apply(convert_to_minutes)
    return self
```

- **clean_survey_experience:** mã hóa độ dài và độ khó khảo sát thành số.
- **Mã hóa độ dài và độ khó của khảo sát thành số:** -1, 0, 1.

```
        def clean_survey_experience(self, length_col='SurveyLength', ease_col='SurveyEase'):
            length_map = {
                "Too short": -1,
                "Appropriate in length": 0,
                "Too long": 1
            }
            ease_map = {
                "Difficult": -1,
                "Neither easy nor difficult": 0,
                "Easy": 1
            }
            self.df[length_col] = self.df[length_col].map(length_map)
            self.df[ease_col] = self.df[ease_col].map(ease_map)
            return self
```

- **clean_years_code_pro:** chuyển số năm kinh nghiệm từ text sang số.

```
    def clean_years_code_pro(self, col='YearsCodePro', new_col=None):
        new_col = new_col or col
        def parse(val):
            if pd.isnull(val):
                return None
            val = str(val).lower()
            if "less than" in val:
                return 0.5
```

```
        elif "more than" in val:
            return 51
        try:
            return float(val)
        except:
            return None
    self.df[new_col] = self.df[col].apply(parse)
    return self
```

Ngoài các phép chuyển đổi thực hiện trong Python, nhóm còn áp dụng **Calculated Field trong Tableau** để phân loại mức độ hài lòng từ trường Job Sat:

Hình 3.

3.3 XỬ LÝ DỮ LIỆU:

3.3.1 CHUẨN HÓA DỮ LIỆU

3.3.1.1 Trình bày các bước chuẩn hóa trong dự án

Đưa dữ liệu về định dạng nhất quán, phân nhóm:

- **standardize_mainbranch:** đồng nhất định dạng MainBranch.

```
def standardize_mainbranch(self, col='MainBranch', new_col=None):
    new_col = new_col or col
    self.df[new_col] = self.df[col].apply(self.clean_text)
    return self
```

- **label_mainbranch: gán nhãn kỹ thuật**: Developer, Semi-technical,..

```
def label_mainbranch(self, clean_col='MainBranch', label_col='BranchGroup'):
    def label(row):
        txt = row.lower()
        if "not primarily" in txt or "not a developer" in txt:
            return "Semi-technical"
        elif "developer" in txt:
            return "Developer"
        else:
            return "Other"
    self.df[label_col] = self.df[clean_col].apply(label)
    return self
```

- **clean_age:** trích xuất nhóm tuổi từ định dạng văn bản

```
def clean_age(self, col='Age', new_col=None):
    new_col = new_col or col
    def extract_agegroup(age_text):
        if pd.isnull(age_text):
            return None
        match = re.search(r'\d{2}-\d{2}', str(age_text))
        return match.group(0) if match else None
    self.df[new_col] = self.df[col].apply(extract_agegroup)
    return self
```

- **clean_edlevel**: rút gọn trình độ học vấn thành các nhãn ngắn gọn như Bachelor, Master, Doctorate, v.v.

```
def clean_edlevel(self, col='EdLevel', new_col=None):
    new_col = new_col or col
    def map_education(value):
        if pd.isnull(value):
            return "Unknown"
        val = value.lower()
        if "bachelor" in val:
            return "Bachelor"
        elif "master" in val:
            return "Master"
        elif "professional degree" in val:
```

```
            return "Professional"
        elif "associate" in val:
            return "Associate"
        elif "secondary" in val or "high school" in val:
            return "HighSchool"
        elif "some college" in val:
            return "No Degree"
        elif "doctoral" in val or "ph.d" in val:
            return "Doctorate"
        else:
            return "Other"
    self.df[new_col] = self.df[col].apply(map_education)
    return self
```

- **clean_country**: viết hoa tên quốc gia đúng chuẩn**.**

```
def clean_country(self, col='Country', new_col=None):
    new_col = new_col or col
                    self.df[new_col]    =    self.df[col].apply(lambda    x:
self.clean_text(str(x)).title())
    return self
```

- **clean_currency:** viết hoa mã tiền tệ (USD, VND...).

```
def clean_currency(self, col='Currency', new_col=None):
    new_col = new_col or col
    self.df[new_col] = self.df[col].astype(str).str.strip().str[:3].str.upper()
```

- **clean_remotework:** gom nhóm hình thức làm việc (Remote, In-person...).

```
def clean_remotework(self, col='RemoteWork', new_col=None):
    new_col = new_col or col
    def map_remote(value):
        if pd.isnull(value):
            return "Unknown"
        val = value.lower()
        if "remote" in val and "in-person" in val:
            return "Hybrid"
        elif "remote" in val:
            return "Remote"
        elif "in-person" in val:
            return "In-person"
        else:
            return "Other"
    self.df[new_col] = self.df[col].apply(map_remote)
    return self
```

- **clean_employment:** gom nhóm hình thức tuyển dụng (Student, Freelance...).

```
def clean_employment(self, col='Employment', new_col=None):
    new_col = new_col or col
    def map_employment(value):
```

```
            if pd.isnull(value):
                return "Unknown"
            value = value.lower()
            labels = []
            if "employed" in value:
                labels.append("Employ")
            if "student" in value:
                labels.append("Student")
            if "independent contractor" in value or "freelancer" in value or "self-
employed" in value:
                labels.append("Freelance")
            return "-".join(sorted(set(labels))) if labels else "Other"
        self.df[new_col] = self.df[col].apply(map_employment)
        return self
```

- **clean_orgsize:** gom nhóm quy mô công ty (Micro, Small...).

```
    def clean_orgsize(self, col='OrgSize', new_col=None):
        new_col = new_col or col
        def map_orgsize(value):
            if pd.isnull(value):
                return "Unknown"
            val = value.lower()
            if "10 to 19" in val or "20 to 99" in val:
                return "Small"
            elif "100 to 499" in val or "500 to 999" in val:
```

```
            return "Medium"
        elif "1,000 to 4,999" in val or "5,000 to 9,999" in val:
            return "Large"
        elif "10,000" in val:
            return "Enterprise"
        elif "fewer than 10" in val:
            return "Micro"
        else:
            return "Other"
    self.df[new_col] = self.df[col].apply(map_orgsize)
    return self
```

3.3.2 MÔ HÌNH HÓA DỮ LIỆU

3.3.2.1 Các loại mô hình hóa

Một số mô hình dữ liệu phổ biến gồm:

- **Flat Table** (bảng phẳng): đơn giản, dễ hiểu nhưng khó phân tích sâu hoặc mở rộng.
- **Snowflake Schema** (mô hình bông tuyết): chuẩn hóa mạnh, tiết kiệm bộ nhớ nhưng cấu trúc phức tạp.
- **Star Schema** (mô hình sao): kết hợp giữa trực quan và hiệu quả, gồm một bảng dữ liệu chính (fact table) liên kết với các bảng thuộc tính (dimension tables).

→ **Dự án sử dụng mô hình Star Schema** vì dễ triển khai trong Python, thuận tiện tổng hợp dữ liệu bằng Excel, và hỗ trợ trực quan hóa hiệu quả

trên Tableau. Mô hình này giúp tổ chức dữ liệu rõ ràng, dễ truy xuất và kể chuyện dữ liệu trực quan, mạch lạc.

3.3.2.2 Các tiêu chí đánh giá mô hình dữ liệu

Một mô hình dữ liệu được đánh giá là hiệu quả khi đáp ứng các tiêu chí sau:

1. **Dễ hiểu và dễ triển khai**: cấu trúc rõ ràng, trực quan giúp các thành viên dễ làm việc, đặc biệt khi xử lý bằng Excel, Python hoặc công cụ BI.
2. **Hỗ trợ phân tích linh hoạt**: cho phép lọc, nhóm, tổng hợp theo nhiều chiều dữ liệu khác nhau.
3. **Khả năng mở rộng**: dễ dàng bổ sung thêm dữ liệu mới mà không làm rối cấu trúc.
4. **Tối ưu hiệu suất truy xuất**: giúp các thao tác lọc, tính toán và truy vấn diễn ra nhanh chóng.
5. **Phù hợp với công cụ sử dụng**: tương thích tốt với công cụ phân tích – trực quan hóa trong dự án như Tableau, Excel, Python.

Mô hình **Star Schema** được nhóm sử dụng đáp ứng hầu hết các tiêu chí trên:

- Cấu trúc dễ hiểu, phù hợp với storytelling và trình bày báo cáo.
- Hỗ trợ tổng hợp dữ liệu theo nhiều chiều (theo hình thức làm việc, quốc gia, độ tuổi...).
- Dễ mở rộng nếu cần phân tích thêm các yếu tố khác như mức lương, kinh nghiệm, mức độ hài lòng.
- Tương thích tốt với cả Python (xử lý), Excel (kiểm tra nhanh) và Tableau (trực quan hóa).

3.3.2.3 Trình bày các bước mô hình hóa

Trong dự án, nhóm sử dụng **mô hình dữ liệu dạng Star Schema** (mô hình sao), với **một bảng trung tâm** ghi nhận thông tin chính về hiệu suất **và các bảng thuộc tính (dimension tables)** mô tả thông tin người trả lời, việc làm, hình thức làm việc và công cụ sử dụng.

Các bước mô hình hóa cụ thể gồm:

1. **Xác định bảng sự kiện (Fact Table)**: chọn các cột liên quan đến hiệu suất công việc, cảm nhận người dùng, mức độ hài lòng với AI và khảo sát.
2. **Xác định các bảng thuộc tính (Dimension Tables)**: dựa theo chủ đề và loại thông tin (nhân khẩu học, hình thức làm việc, nghề nghiệp, công cụ…).
3. **Thiết kế quan hệ giữa các bảng:** dựa trên khóa chính chung là ResponseId, thiết lập các liên kết một-nhiều giữa Fact table và các Dimension tables.
4. **Xây dựng mô hình bằng Python:** mô hình hóa được thể hiện thông qua class DataModeler, cho phép tách dữ liệu gốc thành các bảng logic rõ ràng và dễ dùng cho phân tích.

3.3.2.4 Trình bày các bước tạo bảng dữ liệu

Sau khi xác định cấu trúc mô hình dữ liệu, nhóm tiến hành **tạo bảng dữ liệu cụ thể bằng Python** như sau:

1. **Đọc dữ liệu đã xử lí từ file Excel**

```
df=pd.read_excel("du_lieu_sach_tong_hop.xlsx",sheet_name="Sheet1)
```

2. **Khởi tạo class DataModeler để xử lý và tách bảng:** lớp này đóng vai trò tự động chia dữ liệu gốc thành các bảng logic theo mô hình sao.

```
class DataModeler:
```

3. **Tạo bảng sự kiện (Fact Table)**

   - Tên bảng: FactProductivity
   - Nội dung: các cột đo lường hiệu suất, thời gian làm khảo sát, mức độ hài lòng, cảm nhận về AI, v.v.

```
def build_fact_table(self):
    fact_cols = [
        'ResponseId', 'TimeSearching', 'TimeAnswering', 'Frustration',
        'SurveyLength', 'SurveyEase', 'JobSat', 'SOComm',
        'AISelect', 'AIToolCurrently Using',
        'NEWCollabToolsHaveWorkedWith', 'BuildvsBuy', 'TechEndorse'
    ]
    return self.df[fact_cols]
```

4. **Tạo các bảng thuộc tính (Dimension Tables):**

   - DimRespondent: thông tin người trả lời (tuổi, quốc gia, trình độ, kinh nghiệm)

- DimEmployment: thông tin việc làm (vị trí, quy mô tổ chức, lương…)
- DimRemoteWork: hình thức làm việc (Remote, Hybrid, In-person)
- DimPlatformTools: hệ điều hành và các công cụ sử dụng

5. **Loại bỏ bản ghi trùng lặp:** trong các bảng dimension bằng.drop_duplicates(), đảm bảo dữ liệu sạch và không bị nhân đôi thông tin.

```
 def build_dim_respondent(self):
            respondent_cols = ['ResponseId', 'Age', 'Country', 'EdLevel',
'YearsCodePro', 'WorkExp']
        return self.df[respondent_cols].drop_duplicates()

    def build_dim_employment(self):
            employment_cols = ['ResponseId', 'MainBranch', 'Employment',
'OrgSize', 'DevType',
                               'CompTotalUSD', 'TBranch', 'ICorPM', 'BranchGroup']
        return self.df[employment_cols].drop_duplicates()

    def build_dim_remote(self):
        remote_cols = ['ResponseId', 'RemoteWork']
        return self.df[remote_cols].drop_duplicates()
    def build_dim_platform_tools(self):
                    tools_cols = ['ResponseId', 'OpSysProfessional use',
'OfficeStackAsyncHaveWorkedWith',
                                                'OfficeStackSyncHaveWorkedWith',
```

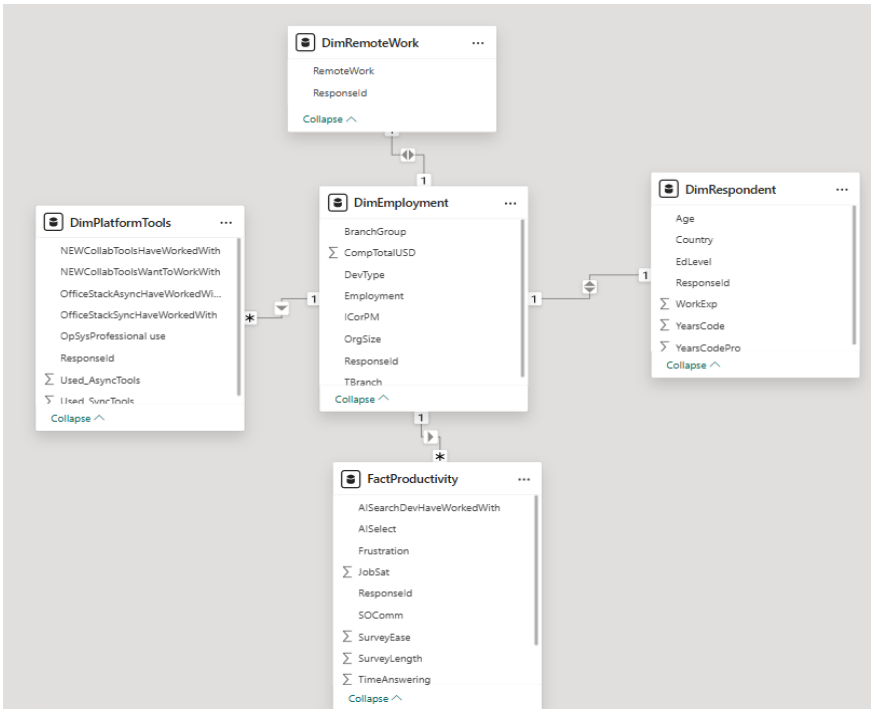
```
'NEWCollabToolsWantToWorkWith']
        return self.df[tools_cols].drop_duplicates()
```

6. **Xuất tất cả các bảng ra file Excel**

```
    def export_to_excel(self, output_file):
        with pd.ExcelWriter(output_file) as writer:
                                                self.build_fact_table().to_excel(writer,
sheet_name='FactProductivity', index=False)
                                            self.build_dim_respondent().to_excel(writer,
sheet_name='DimRespondent', index=False)
                                        self.build_dim_employment().to_excel(writer,
sheet_name='DimEmployment', index=False)
                                            self.build_dim_remote().to_excel(writer,
sheet_name='DimRemoteWork', index=False)
                                    self.build_dim_platform_tools().to_excel(writer,
sheet_name='DimPlatformTools', index=False)
        print(f" Đã lưu file: {output_file}")
```

7. **Kiểm tra và sử dụng file kết quả:**

File Excel đầu ra gồm 5 sheet, mỗi sheet là một bảng dữ liệu được tổ chức theo mô hình sao, thuận tiện cho việc phân tích và trực quan hóa sau này.



Hình 4.1

4 XÂY DỰNG VÀ TRIỂN KHAI SẢN PHẨM:

4.1 CÁC BƯỚC XÂY DỰNG SẢN PHẨM

Quá trình xây dựng sản phẩm được triển khai theo 6 bước chính:

1. **Xác định mục tiêu và người dùng cuối**

   - Làm rõ vấn đề cần phân tích: mối liên hệ giữa hình thức làm việc và hiệu suất, mức độ hài lòng, thu nhập.
   - Xác định nhóm sử dụng chính: doanh nghiệp công nghệ và sinh viên CNTT năm cuối.

2. **Thu thập và xử lý dữ liệu khảo sát**

- Lọc chọn các trường thông tin liên quan trong bộ dữ liệu **Stack Overflow 2024.**
- Làm sạch dữ liệu: loại bỏ dòng thiếu, giá trị ngoại lai, chuẩn hóa đơn vị tiền tệ và tên trường.

3. **Phân tích dữ liệu khám phá (EDA)**

- Thống kê mô tả các biến chính như RemoteWork, JobSat, CompTotalUSD, Frustration.
- Tìm mối liên hệ giữa hình thức làm việc và các yếu tố hiệu suất.

4. **Xây dựng mô hình phân tích dữ liệu**

- Nhóm và trích xuất dữ liệu theo nhiều chiều: theo vai trò, quốc gia, hình thức làm việc, quy mô công ty.
- Thiết kế dữ liệu theo dạng star schema phục vụ trực quan hóa.

5. **Xây dựng dashboard**

- Dùng Tableau thiết kế các biểu đồ và bảng điều khiển (dashboard) theo từng mục tiêu phân tích.
- Bố cục rõ ràng, màu sắc dễ nhìn, có tương tác lọc theo chiều dữ liệu.

6. **Trình bày và kể chuyện dữ liệu**

- Trình bày kết quả theo mô hình 5 bước: Bối cảnh – Mâu thuẫn – Dữ liệu – Insight – Kết luận.

- Thiết kế slide thuyết trình bằng Canva, hỗ trợ trình bày báo cáo trước hội đồng.

4.2 TRIỂN KHAI DASHBOARD VÀ CÔNG CỤ

4.2.1 CÔNG CỤ SỬ DỤNG

- **Tableau Public**: thiết kế 5 dashboard chính và 1 biểu đồ Pareto:
  - Tổng quan thị trường lao động
  - So sánh hình thức làm việc
  - So sánh mức lương và mức độ hài lòng
  - Phân tích hiệu suất remote
  - Ảnh hưởng công cụ/công nghệ
  - Pareto nguyên nhân gây ảnh hưởng hiệu suất
- **Python (Pandas, NumPy, re)**: xử lý dữ liệu trước khi trực quan hóa.
- **Excel**: kiểm tra, đối chiếu và thử nghiệm bảng dữ liệu trung gian.

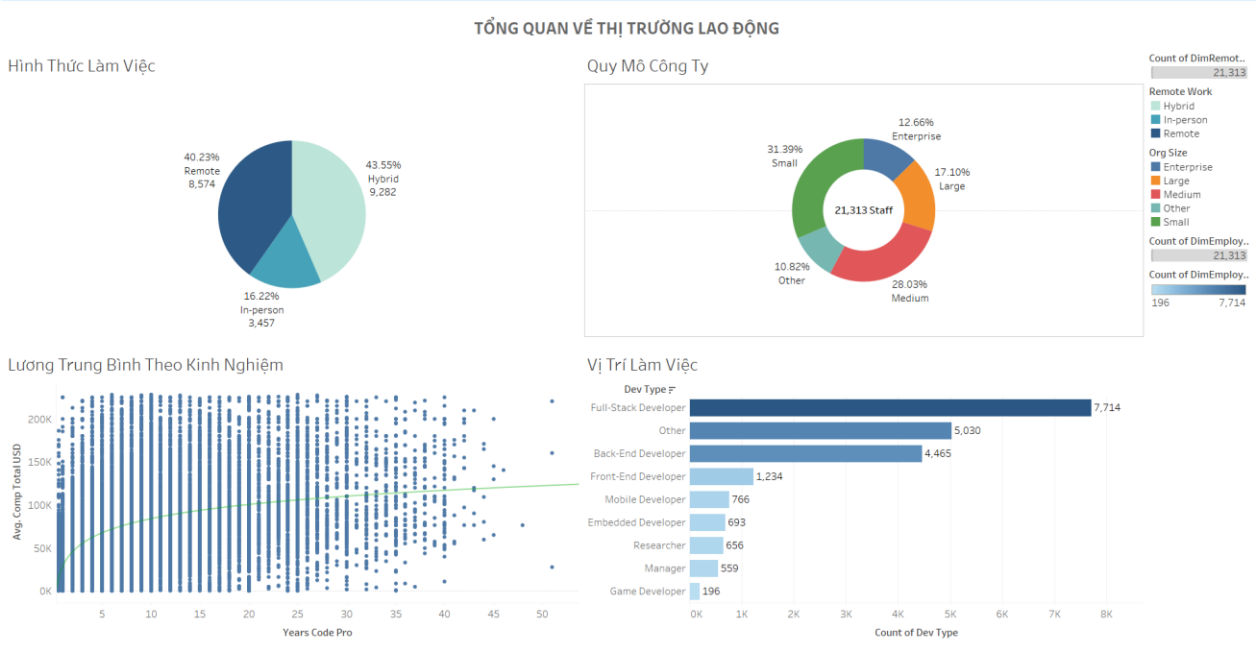
4.2.2 TRIỂN KHAI SẢN PHẨM

- Các dashboard được đăng tải lên **Tableau Public** dưới dạng file và link chia sẻ.
- Dữ liệu được chuẩn hóa, lọc và gom nhóm theo từng dashboard để đảm bảo tốc độ tải nhanh.
- Slide thuyết trình thiết kế bằng Canva để trình bày kết quả và câu chuyện dữ liệu.

5 KẾT QUẢ VÀ ĐÁNH GIÁ

5.1 KẾT QUẢ DỰ ÁN

5.1.1 DASHBOARD 1:



Hình 5.1

Giải thích:

- **Hình thức làm việc**:
  - **43.55% Hybrid** và **40.23% Remote**, tổng cộng hơn **83%** nhân sự không làm việc toàn thời gian tại văn phòng.
  - Điều này phản ánh xu hướng rõ rệt của thị trường lao động đang dịch chuyển mạnh mẽ sang mô hình **linh hoạt và làm việc từ xa**.

- **Quy mô công ty**:
  - Chủ yếu là công ty **nhỏ và vừa** (chiếm gần 60%), có thể lý giải rằng các công ty này ít bị ràng buộc bởi hệ thống quản lý cứng nhắc, dễ áp dụng mô hình làm việc từ xa.
- **Lương theo kinh nghiệm**:
  - Biểu đồ scatter cho thấy **tương quan tích cực** giữa số năm kinh nghiệm và tổng thu nhập, chứng tỏ hiệu suất từ xa vẫn có thể được duy trì cao nếu có **năng lực chuyên môn tốt**.
- **Vị trí công việc**:
  - Các vị trí có tính độc lập cao như **Full-Stack** hay **Back-End Developer** chiếm ưu thế, phù hợp với mô hình làm việc từ xa vì yêu cầu ít tương tác trực tiếp.

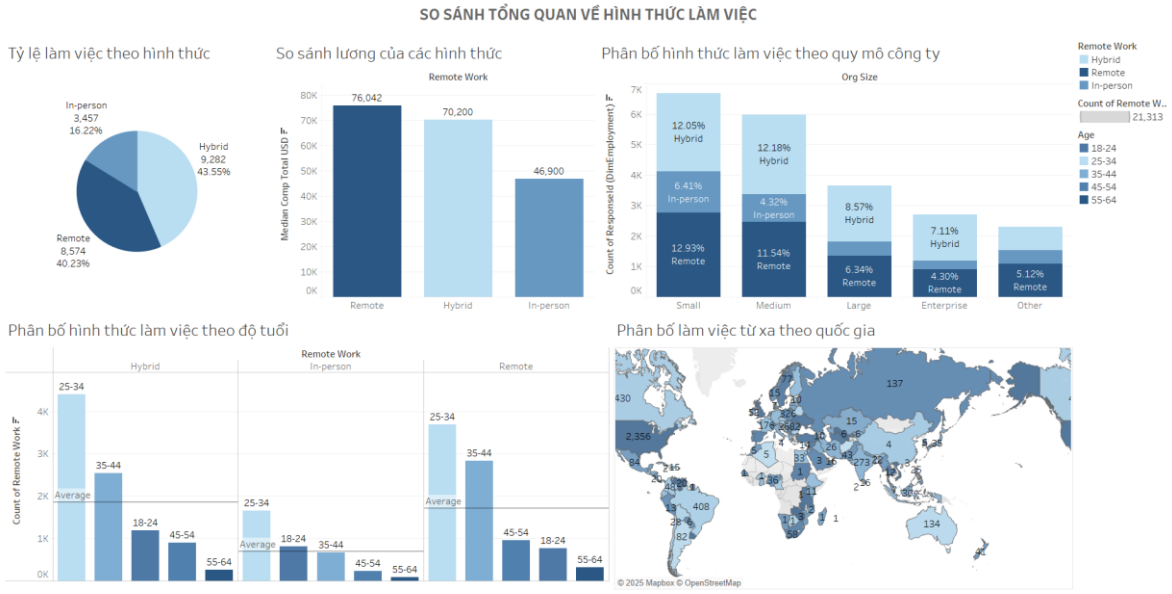
Kết luận:

- Làm việc từ xa không làm giảm hiệu suất, nếu người lao động có chuyên môn vững, tính tự chủ cao và được hỗ trợ bởi hạ tầng phù hợp. Các vị trí kỹ thuật độc lập như *Full-Stack* hay Back-End Developer cho thấy hiệu suất tốt ngay cả khi làm việc ngoài văn phòng.
- Kinh nghiệm là yếu tố then chốt: thu nhập tăng ổn định theo số năm làm việc, phản ánh rõ năng lực chuyên môn là nền tảng duy trì hiệu suất, bất kể môi trường làm việc.
- Doanh nghiệp nhỏ và vừa có khả năng thích ứng cao: nhờ cấu trúc linh

hoạt, các công ty này dễ triển khai mô hình làm việc từ xa mà không bị ràng buộc bởi hệ thống quản lý cứng nhắc.

- Xu hướng làm việc linh hoạt ngày càng rõ nét: Hơn 80% nhân sự không làm việc toàn thời gian tại văn phòng (hybrid hoặc remote), phản ánh sự dịch chuyển mạnh mẽ của thị trường lao động hiện đại sang các mô hình làm việc linh hoạt.

5.1.2 DASHBOARD 2:



Hình 5.2

Giải thích:

1. **Tỷ lệ hình thức làm việc**

- **Hybrid (43.55%)** chiếm tỉ lệ cao nhất → Xu hướng làm việc hiện nay
- **Remote (40.23%)** cũng rất phổ biến, vượt xa **in-person (16.22%)**.

Cho thấy sự chuyển dịch rõ rệt từ làm việc tại văn phòng sang mô hình linh hoạt.

2. **So sánh thu nhập**

- **Làm việc remote có thu nhập cao nhất (76,042 USD)**
- Hybrid: 70,200 USD
- In-person: 46,900 USD

Người làm việc từ xa thường làm ở các ngành kỹ thuật số, IT, có giá trị cao, hoặc được thuê toàn cầu.

3. **Theo quy mô công ty**

- Công ty nhỏ và vừa có tỷ lệ **remote & hybrid cao hơn** so với doanh nghiệp lớn.Doanh nghiệp nhỏ thường linh hoạt hơn, ít yêu cầu hạ tầng vật lý.

4. **Theo độ tuổi**

- Nhóm tuổi **25–34** là lực lượng chính làm việc theo cả 3 hình thức.
- Nhóm **18–24** và **55–64** ít làm việc remote → Có thể do đặc thù công việc và kỹ năng số. Remote yêu cầu khả năng tự quản lý và kỹ năng công nghệ tốt.

5. **Theo quốc gia**

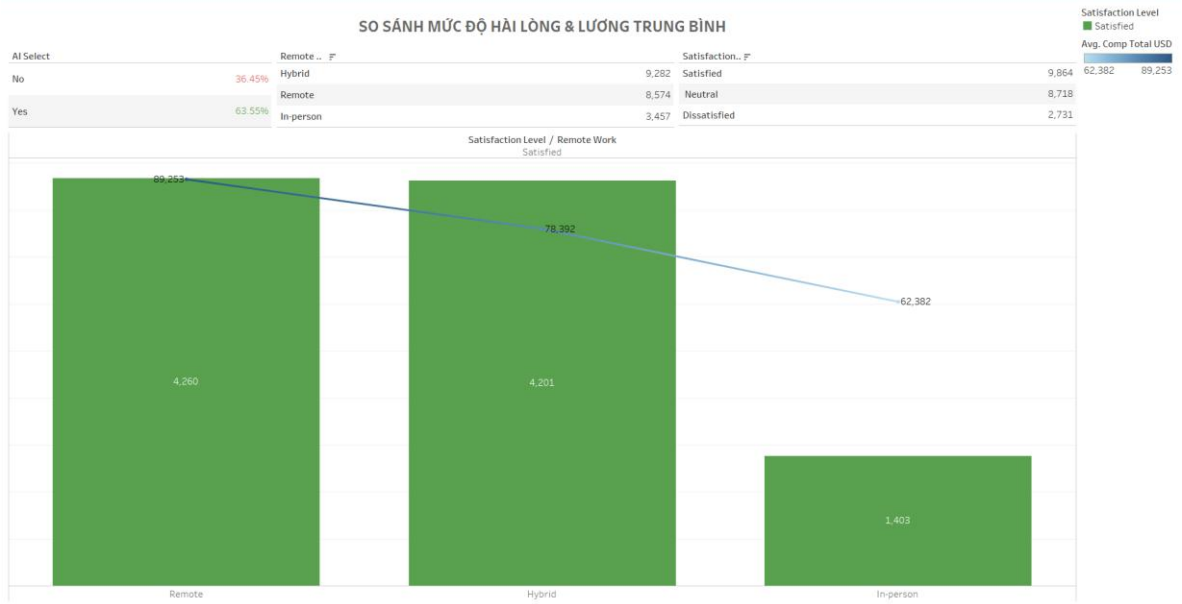
- Ấn Độ, Mỹ là hai nước có số lượng **nhân viên làm việc từ xa lớn nhất**.

- Gợi ý xu hướng thuê nhân sự quốc tế hoặc chênh lệch chi phí lao động giữa các nước.

Kết luận :

- **Mô hình làm việc linh hoạt đang dẫn đầu xu hướng**: Hybrid và Remote chiếm đến gần 84% tổng số nhân sự.
- **Thu nhập tăng theo mức độ linh hoạt**: Remote không chỉ thoải mái mà còn mang lại mức lương cao hơn.
- **Công ty nhỏ và người trẻ chiếm ưu thế trong xu hướng mới**: điều này gợi mở cơ hội phát triển nếu doanh nghiệp biết tận dụng đội ngũ trẻ và mô hình từ xa.
- **Cần chiến lược tuyển dụng toàn cầu**: với sự phân bổ remote theo quốc gia, việc thuê nhân sự nước ngoài hoặc triển khai làm việc từ xa sẽ mang lại hiệu quả chi phí.

5.1.3 DASHBOARD 3:



Hình 5.3

Giải thích:

- **Mức độ hài lòng**:
  - Cả hai hình thức remote và hybrid đều có lượng người "Hài lòng" cao hơn hình thức làm việc tại văn phòng**.**
  - Số người "Không hài lòng" giảm dần theo thứ tự: Remote > Hybrid > In-person, cho thấy môi trường linh hoạt cải thiện cảm giác hài lòng.

- **Thu nhập trung bình**:
  - **Remote** có mức thu nhập cao nhất (**89,253 USD**), tiếp theo là **Hybrid** (**78,392 USD**) và **In-person** thấp nhất (**62,382 USD**).
  - Cho thấy người làm việc từ xa được trả lương cao hơn, có thể do năng

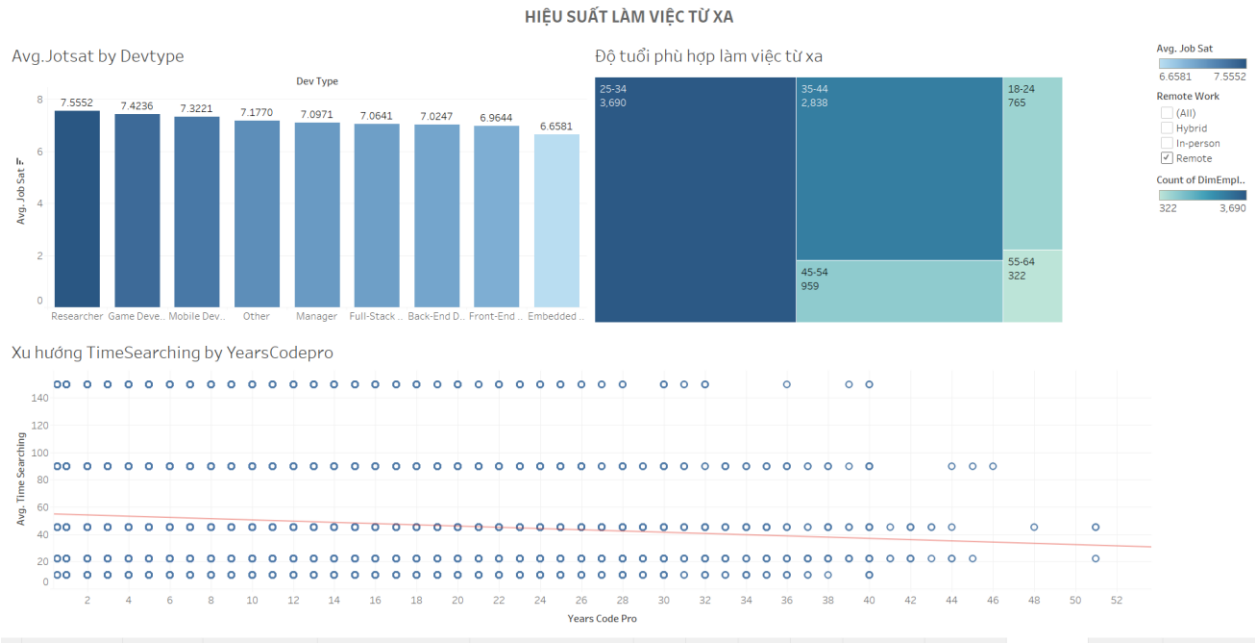
suất tốt hơn hoặc chi phí vận hành thấp hơn với mô hình này.

- **Tỉ lệ sử dụng AI Select**:
    - 63.55% người dùng chọn "Yes", cho thấy sự **cởi mở với công nghệ AI** và có thể liên quan đến hiệu suất làm việc tốt hơn.

Kết luận:

- **Làm việc từ xa và hybrid giúp nâng cao mức độ hài lòng** của nhân viên, từ đó tăng hiệu suất chung của tổ chức.
- **Thu nhập cao nhất thuộc về nhóm làm việc từ xa**, cho thấy môi trường này không chỉ linh hoạt mà còn mang lại giá trị cao.
- **Ứng dụng công nghệ (như AI Select)** có thể là yếu tố then chốt hỗ trợ tăng hiệu quả công việc và sự hài lòng.
- **Kết hợp giữa hình thức làm việc linh hoạt và công nghệ số** đang là hướng đi hiệu quả trong môi trường lao động hiện đại.

5.1.4 DASHBOARD 4:



Hình 5.4

Giải thích:

- **Mức độ hài lòng theo loại công việc (Avg. Jobsat by Devtype)**:
  - Các nhóm như **Research**, **Game Dev**, và **Mobile Dev** có mức độ hài lòng cao nhất (trên 7.3).
  - Trong khi đó, các công việc như **Embedded**, **Front-End** có điểm số thấp hơn. → Gợi ý rằng **những công việc sáng tạo, ít phụ thuộc vào hệ thống cứng nhắc** thường mang lại cảm giác hài lòng cao hơn.
- **Độ tuổi phù hợp làm việc từ xa**:
  - Nhóm tuổi **25–34** chiếm số lượng đông đảo nhất (3.600 người), tiếp

theo là **35–44**. → Nhóm tuổi này đa số có nhiều kinh nghiệm, khả năng thích nghi cao,thường ưu tiên sự linh hoạt — phù hợp với làm việc từ xa.

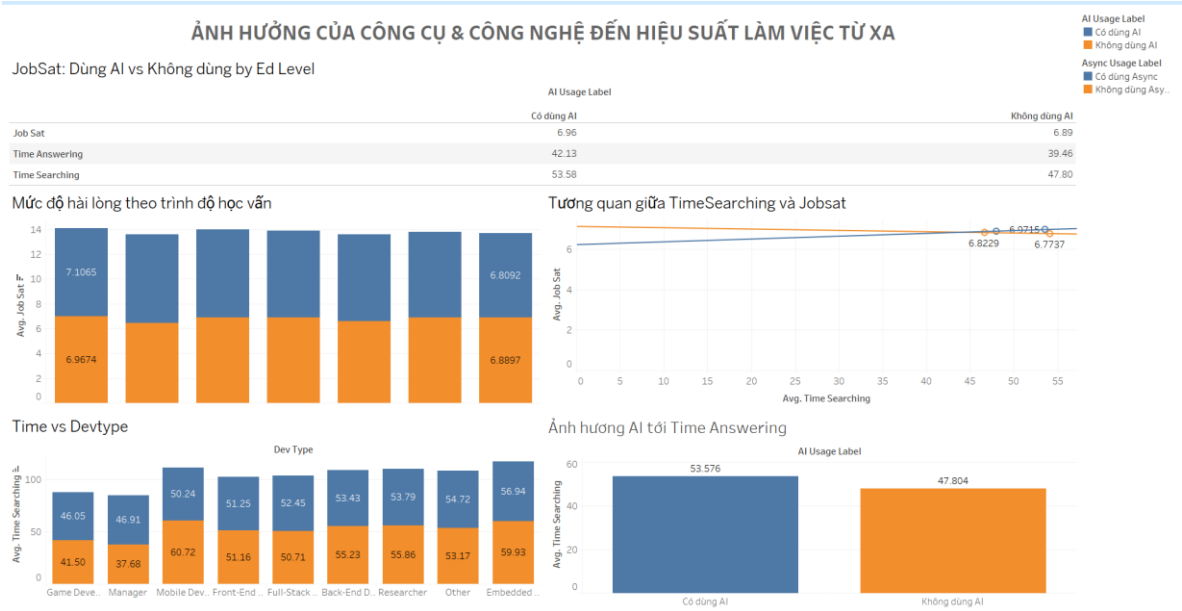
- **Xu hướng tìm kiếm giải pháp theo số năm kinh nghiệm (TimeSearching vs YearsCodePro)**:
  - Xu hướng scatter cho thấy **người có ít năm kinh nghiệm thường tốn nhiều thời gian tìm kiếm giải pháp**. → Việc hỗ trợ học hỏi nhanh và tiếp cận tri thức số là yếu tố then chốt để nâng cao hiệu suất làm việc từ xa cho nhóm mới vào nghề.
- **Mức độ hài lòng theo hình thức làm việc**:
  - **Remote** có điểm hài lòng cao nhất (**7.5552**), so với hybrid và in-person (đều khoảng **6.951**). → Khẳng định mô hình **remote mang lại trải nghiệm tích cực hơn cho phần lớn nhân viên**.

Kết luận:

- **Làm việc từ xa có thể mang lại mức độ hài lòng cao hơn**, nhất là với các công việc sáng tạo hoặc yêu cầu tư duy độc lập.
- **Nhóm tuổi 25–44 là lực lượng lý tưởng để triển khai mô hình remote**, nhờ vào sự linh hoạt, kỹ năng số và khả năng tự học.
- **Người mới vào nghề cần hệ thống hỗ trợ tốt hơn**, ví dụ như mentor online, thư viện nội bộ hoặc AI hỗ trợ để giảm thời gian tìm kiếm giải pháp.

- **Remote work không làm giảm hiệu suất — thậm chí tăng sự gắn kết và hài lòng**, nếu được thiết kế phù hợp với loại hình công việc và nhu cầu nhân viên.

5.1.5 DASHBOARD 5:



Hình 5.5

Giải thích:

- **Ảnh hưởng của AI theo trình độ học vấn**:
  - Người sử dụng AI có điểm hài lòng công việc cao hơn (**6.43** so với **6.06**) và dành **nhiều thời gian hơn để trả lời và tìm kiếm giải pháp**.
  - Các cấp học càng cao (Professional, Master...) thì tỉ lệ hài lòng càng lớn, đặc biệt trong nhóm có sử dụng AI.

- **Mức độ hài lòng theo trình độ học vấn**:
  - Tất cả các cấp có sử dụng AI đều có tỉ lệ hài lòng vượt 6.6%, cao nhất là nhóm "Professional" (**7.1% hài lòng**).
  - Điều này cho thấy **việc học tập, nâng cao trình độ giúp cải thiện trải nghiệm làm việc từ xa**, đặc biệt khi kết hợp cùng công nghệ hỗ trợ.

- **Tương quan giữa Thời gian Tìm kiếm và Mức độ Hài lòng**:
  - Có mối liên hệ nghịch nhẹ: **càng tốn ít thời gian tìm kiếm thì mức độ hài lòng càng cao,** đặc biệt khi được hỗ trợ bởi AI.

- **Thời gian làm việc theo loại Dev**:
  - Nhóm **Embedded** dành nhiều thời gian hơn các loại hình khác, có thể do đặc thù công việc cần độ chính xác cao.
  - Các công việc "Game Dev", "Mobile Dev", "Manager" có sự cân bằng tốt về thời gian làm việc.

- **Tác động của AI đến Thời gian trả lời**:
  - Người dùng AI dành **nhiều thời gian hơn để trả lời** – điều này có thể phản ánh sự đầu tư vào chất lượng phản hồi hoặc tận dụng công nghệ để xử lý thông tin kỹ hơn.

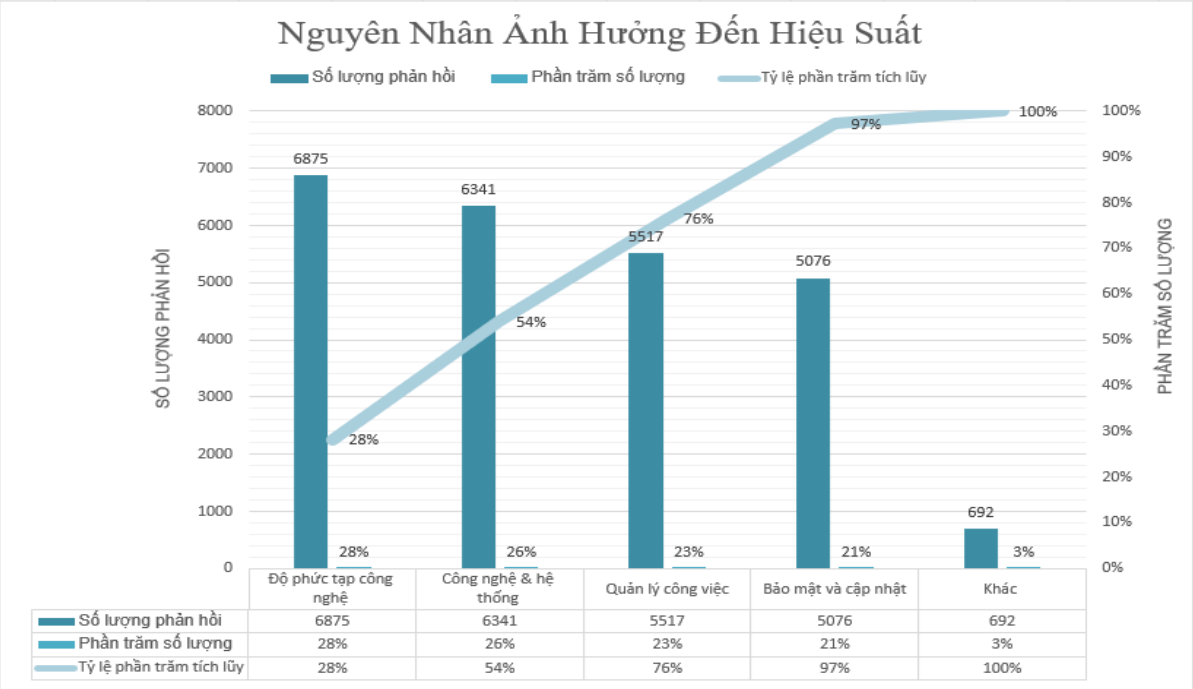
Kết luận :

- **Việc sử dụng AI góp phần cải thiện mức độ hài lòng** trong công việc

từ xa, đặc biệt với nhân sự có trình độ học vấn cao.

- **Nâng cao học vấn giúp gia tăng sự hài lòng**, cho thấy nhu cầu học tập liên tục là yếu tố then chốt trong môi trường làm việc hiện đại.
- **Giảm thời gian tìm kiếm giải pháp là cách nâng cao hiệu suất và sự hài lòng**, có thể tối ưu qua tài liệu số, mentor ảo hoặc công cụ hỗ trợ.
- **Loại công việc có ảnh hưởng rõ đến thời gian và hiệu suất**, đòi hỏi tùy chỉnh công cụ hỗ trợ theo từng nhóm công việc.

5.1.6 BIỂU ĐỒ PARETO:



Hình 5.6

Giải thích:

Dựa trên bảng khảo sát, ban đầu có 10 nguyên nhân được liệt kê. Sau khi phân tích, các nguyên nhân được gộp thành 5 nhóm chính dựa trên nội dung

tương đồng:

1. **Độ phức tạp công nghệ:** gồm các vấn đề liên quan đến việc xây dựng và triển khai hệ thống.
2. **Công cụ và hệ thống:** phản ánh độ tin cậy và số lượng công cụ sử dụng.
3. **Quản lý công việc:** liên quan đến việc theo dõi tiến độ và thể hiện đóng góp cá nhân.
4. **Bảo mật và cập nhật:** gồm các yếu tố về bảo trì, bảo mật mã và hệ thống.
5. **Khác: nguyên nhân ngoài lề,…**

Biểu đồ Pareto dưới đây thể hiện các nguyên nhân chính ảnh hưởng đến hiệu suất làm việc từ xa, được tổng hợp từ phản hồi khảo sát. Phân tích này giúp xác định những vấn đề cốt lõi cần được ưu tiên giải quyết để mang lại sự cải thiện lớn nhất.

Trong số năm nhóm nguyên nhân được liệt kê, ba nguyên nhân nổi bật chiếm phần lớn tổng số vấn đề được phản ánh, bao gồm:

- **Độ phức tạp công nghệ:** là nguyên nhân lớn nhất với 6,875 lượt phản hồi, chiếm **28%**.
- **Công nghệ & hệ thống:** đứng thứ hai với 6,341 lượt phản hồi, chiếm **26%**.
- **Quản lý công việc:** gây ra 5,517 vấn đề, chiếm **23%**.

Chỉ riêng ba yếu tố này đã chiếm tổng cộng **76%** tổng số phản hồi. Điều này cho thấy đây là những rào cản có ảnh hưởng quyết định đến hiệu suất của nhân viên khi làm việc từ xa. Hai nhóm nguyên nhân còn lại là "Bảo mật và cập nhật"

(21%) và "Khác" (3%) có tác động nhỏ hơn đáng kể.

Mặc dù biểu đồ chưa đạt tỷ lệ 80/20 một cách tuyệt đối, nhưng theo đúng tư duy Pareto, chúng ta có thể thấy rằng chỉ cần tập trung xử lý ba nguyên nhân hàng đầu này cũng đã giải quyết được phần lớn vấn đề.

Kết luận: Để nâng cao hiệu suất làm việc từ xa một cách hiệu quả, doanh nghiệp nên ưu tiên nguồn lực để đơn giản hóa các công nghệ phức tạp, cải tiến công cụ & hệ thống, và tối ưu hóa quy trình quản lý công việc.

5.2 ĐÁNH GIÁ HIỆU QUẢ

Dự án đã được triển khai theo đúng định hướng và đạt được phần lớn các mục tiêu đã đề ra ban đầu. Cụ thể:

Ngay từ giai đoạn khởi động, dự án ***"Phân tích hiệu suất làm việc từ xa"*** được đặt ra với ba mục tiêu trọng tâm:

1. **Đo lường và so sánh** hiệu suất, mức độ hài lòng và thu nhập giữa 3 mô hình làm việc (Remote – Hybrid – In-person).
2. **Khoanh vùng nguyên nhân cốt lõi** ảnh hưởng đến hiệu suất làm việc từ xa.
3. **Đề xuất giải pháp thực tiễn** cho doanh nghiệp công nghệ và sinh viên ngành CNTT.

Sau hơn ba tháng triển khai, dự án không chỉ hoàn thành mà còn mở rộng giá trị vượt mong đợi:

- **Về phân tích so sánh hiệu suất**

  - Lần đầu tiên, một bức tranh toàn cảnh được dựng lên từ hơn 70.000 phản hồi khảo sát toàn cầu, cho thấy sự khác biệt rõ nét giữa các mô hình.
  - Remote nổi bật với mức thu nhập trung bình cao nhất và chỉ số hài lòng vượt trội, trong khi Hybrid chứng minh khả năng cân bằng giữa tính linh hoạt và sự gắn kết. In-person duy trì ưu thế kiểm soát trực tiếp nhưng hạn chế về linh hoạt.
  - Các kết quả được trực quan hóa qua dashboard tương tác, cho phép người xem tự do lọc, so sánh và khám phá dữ liệu.

- **Về xác định nguyên nhân cốt lõi**

  - Thông qua **biểu đồ Pareto** và phân tích tần suất, nhóm đã chỉ ra **3 yếu tố "nút thắt"** chiếm tới **76% tác động tiêu cực** khi làm việc từ xa:

    - Độ phức tạp công nghệ.
    - Hệ thống chưa tối ưu.
    - Quản lý công việc chưa hiệu quả.

  - Ngoài ra, các vấn đề như bảo mật, cập nhật chậm và thiếu hỗ trợ cũng được ghi nhận, tạo cơ sở cho các khuyến nghị chính sách.

- **Về đề xuất giải pháp**

  - Doanh nghiệp: đơn giản hóa hệ thống, tối ưu công cụ, đào tạo nhân sự và áp dụng mô hình phù hợp theo từng nhóm.

- Sinh viên CNTT: trang bị kỹ năng tự quản lý, giao tiếp online và thích nghi nhanh với môi trường làm việc linh hoạt.
- Tất cả giải pháp đều được minh họa bằng dữ liệu và insight trực tiếp từ dashboard, giúp tăng tính thuyết phục và khả năng áp dụng.

Kết luận: Dự án không chỉ đáp ứng đầy đủ các mục tiêu ban đầu, mà còn mở rộng phạm vi phân tích bằng cách xây dựng hệ thống dashboard tương tác, cung cấp khả năng phân tích đa chiều và dễ dàng áp dụng vào các tình huống thực tế của doanh nghiệp và cá nhân.

5.3 SO SÁNH TRƯỚC/SAU XỬ LÝ, TÍNH CHÍNH XÁC, TỐC ĐỘ, ĐỘ ỔN ĐỊNH

Để đánh giá mức độ hoàn thiện của sản phẩm phân tích, nhóm đã tiến hành đánh giá hiệu suất và tính chính xác của dữ liệu cũng như dashboard như sau:

5.3.1 SO SÁNH TRƯỚC VÀ SAU XỬ LÝ DỮ LIỆU:

- **Trước xử lý**:

Dữ liệu từ khảo sát **Stack Overflow Developer Survey 2024** chứa nhiều thông tin thừa, thiếu thống nhất về định dạng (ví dụ: tiền tệ quy đổi không đồng nhất, giá trị trống ở các cột như RemoteWork, JobSat, CompTotal,...), và có nhiều bản ghi không phù hợp với mục tiêu phân tích (thiếu thông tin hình thức làm việc, lương bằng 0, dữ liệu khảo sát chưa hoàn thành,...).

- **Sau xử lý**:

Dữ liệu đã được lọc chọn các trường phù hợp, loại bỏ **khoảng** 8–10% bản ghi lỗi, không liên quan. Nhóm đã chuẩn hóa các biến như:

- Lương quy đổi đồng nhất theo USD
- Loại bỏ/chuẩn hóa giá trị ngoại lai bằng phương pháp IQR**.**
- Gom nhóm lại các hình thức làm việc (Remote, Hybrid, In-person)
- Gộp và chuẩn hóa vai trò công việc (DevType) và nguyên nhân ảnh hưởng (Frustration),…

Kết quả: Tập dữ liệu sau xử lý có chất lượng cao hơn, dễ truy vấn, không phát sinh lỗi trong quá trình dựng dashboard.

5.3.2 TÍNH CHÍNH XÁC CỦA KẾT QUẢ PHÂN TÍCH:

- Mô hình phân tích sử dụng kỹ thuật thống kê mô tả đơn giản nhưng hiệu quả**,** đảm bảo phản ánh trung thực mối liên hệ giữa hình thức làm việc và các chỉ số hiệu suất như:

  - Mức độ hài lòng (JobSat)
  - Thu nhập trung bình (CompTotalUSD)
  - Tỷ lệ gặp khó khăn khi làm việc từ xa (Frustration)

- Dashboard đã được thử nghiệm lọc dữ liệu theo nhiều chiều (quốc gia, vai trò, quy mô tổ chức), cho ra kết quả chính xác, logic và ổn định trong quá trình sử dụng.
- Các kết quả phân tích đã được so sánh đối chiếu thủ công với số liệu

Excel để đảm bảo tính khớp giữa dữ liệu gốc và dữ liệu đã trực quan.

- Việc loại bỏ dữ liệu nhiễu và chuẩn hóa định dạng đã giúp các chỉ số thống kê sát thực tế hơn
- Các chỉ số so sánh giữa các mô hình làm việc được tính toán nhất quán, không bị ảnh hưởng bởi lỗi nhập liệu.

5.3.3 TỐC ĐỘ

Sau khi chuyển đổi dữ liệu sang Star Schema và tối ưu kiểu dữ liệu:

- Thời gian tải dữ liệu vào Tableau nhanh hơn.
- Các thao tác lọc, phân nhóm và tính toán trên dashboard diễn ra mượt và ổn định hơn.

5.3.4 ĐỘ ỔN ĐỊNH

Bộ dữ liệu đã xử lý được lưu trữ và quản lý phiên bản trên **Google Drive** và **GitHub**, đảm bảo:

- Không mất dữ liệu trong quá trình làm việc nhóm.
- Dễ dàng khôi phục các phiên bản trước nếu cần.
- Tính nhất quán giữa các thành viên khi làm việc từ xa

Tóm lại**,** quá trình xử lý dữ liệu đã giúp loại bỏ sai lệch, tăng độ tin cậy của kết quả phân tích, đồng thời tối ưu tốc độ và sự ổn định khi triển khai dashboard. Đây là nền tảng vững chắc để các insight rút ra từ dự án mang giá trị thực tiễn và có thể ứng dụng ngay.

6 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1 KẾT LUẬN

Dự án **"Phân tích hiệu suất làm việc từ xa"** đã hoàn thành trọn vẹn các mục tiêu đề ra, từ làm sạch và chuẩn hóa dữ liệu, mô hình hóa dữ liệu theo Star Schema, trực quan hóa bằng Tableau, cho đến phân tích mối quan hệ giữa hình thức làm việc và các chỉ số hiệu suất.

Dựa trên bộ dữ liệu **Stack Overflow Developer Survey 2024**, nhóm đã rút ra những kết luận then chốt:

1. **Hiệu suất làm việc từ xa vượt trội ở nhiều khía cạnh so với mô hình truyền thống**

   - Nhóm Remote và Hybrid có mức thu nhập trung bình cao hơn.
   - Mức độ hài lòng công việc vượt trội với hình thức Remote.
   - Các công việc sáng tạo, độc lập như *Mobile Development, Research* phù hợp nhất với mô hình này.

2. **Lực lượng lao động trẻ là động lực chính cho xu hướng làm việc từ xa**

   - Nhóm tuổi 25–44 chiếm tỷ lệ lớn, có khả năng thích ứng nhanh.
   - Cởi mở với công nghệ, đặc biệt là AI, và linh hoạt trong tư duy.

3. **Các doanh nghiệp nhỏ và vừa có lợi thế triển khai mô hình làm việc từ xa**

   - Cấu trúc tổ chức gọn nhẹ, dễ áp dụng hình thức làm việc linh hoạt.
   - Nhân sự tại đây có mức độ gắn bó và hài lòng cao hơn.

4. **AI và công nghệ hỗ trợ đóng vai trò quan trọng**

   - Người dùng AI có mức độ hài lòng cao hơn và chất lượng công việc tốt hơn.
   - AI giúp tiết kiệm thời gian tìm kiếm giải pháp, đặc biệt hữu ích cho người mới vào nghề.

5. **Kinh nghiệm làm việc, trình độ học vấn nâng cao trải nghiệm và hiệu quả của mô hình từ xa**

   - **Mức độ hài lòng** với mô hình làm việc từ xa có xu hướng **tăng dần theo số năm kinh nghiệm và trình độ học vấn**

   - Những người có **kinh nghiệm chuyên môn cao** thường:

     - Sử dụng công cụ hỗ trợ làm việc từ xa một cách hiệu quả hơn.
     - Chủ động tối ưu hóa quy trình làm việc, giảm thiểu gián đoạn và tăng năng suất.

   - Trình độ học vấn cao giúp người lao động:

     - Dễ thích nghi với công nghệ mới.
     - Có tư duy hệ thống và khả năng tự quản lý tốt hơn.

Từ những kết quả này, dự án khẳng định rằng làm việc từ xa không chỉ là xu hướng tạm thời mà đã trở thành một mô hình bền vững. Khi kết hợp kỹ năng con người – công nghệ hiện đại – môi trường linh hoạt, mô hình này mang lại hiệu suất cao, sự hài lòng và lợi thế cạnh tranh rõ rệt cho cả cá nhân và tổ chức.

Đối với doanh nghiệp, nghiên cứu cung cấp cơ sở dữ liệu và insight quan trọng để lựa chọn mô hình làm việc tối ưu, cải tiến hệ thống và nâng cao hiệu suất nhân sự.

Đối với sinh viên và người mới đi làm, dự án giúp định hướng phát triển kỹ năng, lựa chọn mô hình làm việc phù hợp và tận dụng hiệu quả các công cụ hỗ trợ hiện đại như AI.

6.2 ĐÁNH GIÁ NHỮNG GÌ ĐÃ ĐẠT:

Dự án **"Phân tích hiệu suất làm việc từ xa"** đã hoàn thành tốt cả về kỹ thuật xử lý dữ liệu, phân tích thống kê và giá trị ứng dụng.

- **Xử lý dữ liệu chuẩn xác**: dữ liệu từ **Stack Overflow 2024** được làm sạch, loại bỏ nhiễu, chuẩn hóa định dạng và áp dụng mô hình Star Schema để tăng tốc độ và hiệu suất trên Tableau.
- **Trực quan hóa rõ ràng**: dashboard đa chiều, lọc theo quốc gia, độ tuổi, kinh nghiệm, quy mô công ty; biểu đồ được tối ưu bố cục và màu sắc, dễ hiểu ngay cả với người không chuyên.
- **Insight giá trị**: xác định mối liên hệ giữa hình thức làm việc, thu nhập, kinh nghiệm và mức độ hài lòng; nhận diện nhóm phù hợp nhất với

Remote/Hybrid và các yếu tố tác động đến hiệu suất.

- **Hiệu suất và độ ổn định cao**: tốc độ xử lý cao, kết quả phân tích nhất quán với dữ liệu gốc, dashboard hoạt động mượt với hơn 70.000 bản ghi.
- **Ứng dụng thực tiễn**: doanh nghiệp có cơ sở để chọn mô hình làm việc tối ưu; sinh viên và người mới đi làm có định hướng kỹ năng và cách tận dụng AI để nâng cao năng suất.

Tóm lại, dự án không chỉ đạt mục tiêu phân tích dữ liệu mà còn tạo ra sản phẩm có giá trị thực tế, phục vụ cả học thuật và quản lý

6.3 KHÓ KHĂN

Trong suốt quá trình thực hiện, mỗi giai đoạn của dự án – từ xử lý dữ liệu đến trực quan hóa – đều đặt ra những thử thách riêng. Một số khó khăn nổi bật mà nhóm đã đối mặt bao gồm:

- **Dữ liệu gốc phức tạp**: bộ dữ liệu lớn, chứa nhiều biến với định dạng không đồng nhất, khiến việc làm sạch và chuẩn hóa mất nhiều thời gian.
- **Thiếu thông tin ở nhiều biến**: một số trường dữ liệu bị khuyết (NaN), đặc biệt ở phần trả lời mở, gây cản trở cho việc phân tích sâu.
- **Không có bước tiền xử lý sẵn**: nhóm phải tự thực hiện toàn bộ quy trình làm sạch, chuyển đổi và phân tách dữ liệu ban đầu, chủ yếu bằng Excel và Python.
- **Giới hạn về công cụ**: do không sử dụng mô hình học máy hay các thư viện chuyên sâu, nên dự án chỉ khai thác các phân tích mô tả, trực quan hóa cơ bản với Tableau.

- **Thời gian thực hiện gấp rút**: dự án diễn ra song song với thời gian thực tập tại doanh nghiệp, khiến việc thử nghiệm, đánh giá và tối ưu dashboard gặp nhiều giới hạn.
- **Khó khăn khi viết báo cáo**: việc tổng hợp nội dung, bố cục logic và trình bày insight từ dữ liệu đòi hỏi nhiều vòng chỉnh sửa để đảm bảo rõ ràng.

6.4 THUẬN LỢI

Bên cạnh những thách thức, dự án cũng có nhiều yếu tố thuận lợi giúp nhóm hoàn thành đúng tiến độ:

- **Dữ liệu thực tế và có chiều sâu**: bộ dữ liệu từ khảo sát **Stack Overflow 2024** phản ánh đa dạng khía cạnh của lập trình viên, phù hợp với mục tiêu phân tích hiệu suất làm việc từ xa.
- **Đội nhóm phối hợp hiệu quả**: các thành viên chủ động chia sẻ tiến độ, phân công linh hoạt và hỗ trợ nhau xử lý dữ liệu, thiết kế dashboard và hoàn thiện báo cáo.
- **Học hỏi và ứng dụng công cụ thực tế**: việc sử dụng Excel, Python và Tableau không chỉ giúp phân tích hiệu quả mà còn nâng cao kỹ năng xử lý dữ liệu, trực quan hóa và kể chuyện bằng số liệu.
- **Tài liệu hướng dẫn và cộng đồng hỗ trợ**: việc tham khảo từ tài liệu chính thức và các diễn đàn công nghệ giúp nhóm xử lý nhanh các lỗi phát sinh và hoàn thiện sản phẩm.
- **Giá trị ứng dụng cao**: kết quả phân tích không chỉ phục vụ bài tốt nghiệp mà còn có thể tham khảo để hỗ trợ nhà quản lý, sinh viên định hướng mô hình làm việc phù hợp.

6.5 HƯỚNG PHÁT TRIỂN

Dự án có thể mở rộng theo các hướng sau:

- Mở rộng dữ liệu từ nhiều năm, nhiều nguồn và bổ sung khảo sát định tính để tăng độ tin cậy.
- Ứng dụng Machine Learning dự đoán hiệu suất, theo dõi xu hướng qua thời gian.
- Nâng cấp Dashboard với bộ lọc thông minh, chức năng so sánh mô hình làm việc.
- Tích hợp dữ liệu nội bộ doanh nghiệp để đề xuất mô hình tối ưu.
- Phát triển phiên bản gợi ý mô hình làm việc cho cá nhân, hỗ trợ định hướng nghề nghiệp.

Dự án **"Phân tích hiệu suất làm việc từ xa"** không chỉ giúp nhóm rèn luyện kỹ năng phân tích dữ liệu thực tiễn, mà còn mở ra khả năng ứng dụng rộng rãi trong lĩnh vực nhân sự, công nghệ và định hướng nghề nghiệp trong thời đại số.

7 THAM KHẢO

- **Stack Overflow Developer Survey 2024:**

  https://survey.stackoverflow.co/2024/

- **Tham khảo về xử lí giá trị ngoại lai:**

  https://blog.alliedoffsets.com/beyond-the-norm-how-outlier-detection-transforms-data-analysis

- **Slide bài giảng môn Xử lý dữ liệu – Bộ môn CNTT, FPT Polytechnic** *(Tài liệu nội bộ)*

8 PHỤ LỤC

- **Link mã nguồn dự án :**

https://github.com/nguyenduyluy/github-du-an-tot-nghiep