**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**



Lớp/Kỳ: DP19303/ Summer 2025

Các 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

GVHD: Thầy Văn Công Khanh

TP.HCM 08 – 2025

MỤC LỤC

[1 Giới thiệu dự án 4](#_Toc204439806)

[1.1 Giới thiệu 4](#_Toc204439807)

[1.2 Yêu cầu của công ty 6](#_Toc204439808)

[1.3 Lập kế hoạch dự án 8](#_Toc204439809)

[2 Phân tích yêu cầu khách hàng 10](#_Toc204439810)

[2.1 Phân tích yêu cầu 10](#_Toc204439811)

[2.2 Câu chuyện dữ liệu 14](#_Toc204439812)

[2.2.1 Đặt vấn đề 14](#_Toc204439813)

[2.2.2 Xác định câu chuyện 17](#_Toc204439814)

[2.2.3 Xác định rõ đối tượng 19](#_Toc204439815)

[2.2.4 Xác định câu chuyện chi tiết 20](#_Toc204439816)

[2.2.5 Trình bày dữ liệu 22](#_Toc204439817)

[2.2.6 Những điều cần lưu ý 23](#_Toc204439818)

[2.3 Kiến trúc hệ thống 24](#_Toc204439819)

[2.3.1 Kiến trúc 24](#_Toc204439820)

[2.3.2 Giải thích 24](#_Toc204439821)

[2.4 Giải thích về bộ dữ liệu khách hàng 25](#_Toc204439822)

[2.4.1 Các khái niệm 25](#_Toc204439823)

[2.4.2 Các trường dữ liệu 26](#_Toc204439824)

[3 Làm sạch và chuyển đổi dữ liệu 29](#_Toc204439825)

[3.1 Chuẩn bị dữ liệu 30](#_Toc204439826)

[3.1.1 Giải pháp lưu trữ dữ liệu 30](#_Toc204439827)

[3.1.2 Giải pháp phân bố dữ liệu 31](#_Toc204439828)

[3.2 Làm sạch dữ liệu 32](#_Toc204439829)

[3.2.1 Các vấn đề ảnh hưởng tới dữ liệu 32](#_Toc204439830)

[3.2.2 Các tiêu chí đánh giá chất lượng dữ liệu 33](#_Toc204439831)

[3.2.3 Các bước làm sạch dữ liệu 34](#_Toc204439832)

[3.3 Chuyển đổi dữ liệu 40](#_Toc204439833)

[3.3.1 Các trường hợp cần chuyển đổi 40](#_Toc204439834)

[3.3.2 Các kỹ thuật chuyển đổi 40](#_Toc204439835)

[3.3.3 Trình bày các phép chuyển đổi trong dự án 41](#_Toc204439836)

[4 Xử lý dữ liệu 50](#_Toc204439837)

[4.1 Chuẩn hóa dữ liệu 50](#_Toc204439838)

[4.1.1 Trình bày các bước chuẩn hóa trong dự án 50](#_Toc204439839)

[4.2 Mô hình hóa dữ liệu 56](#_Toc204439840)

[4.2.1 Các loại mô hình hóa 56](#_Toc204439841)

[4.2.2 Các tiêu chí đánh giá mô hình dữ liệu 56](#_Toc204439842)

[4.2.3 Trình bày các bước mô hình hóa 57](#_Toc204439843)

[4.2.4 Trình bày các bước tạo bảng dữ liệu 58](#_Toc204439844)

[5 Trực quan hóa dữ liệu 61](#_Toc204439845)

[5.1 Các kỹ thuật trực quan hóa 61](#_Toc204439846)

[5.2 Các nguyên tắc trực quan hóa 63](#_Toc204439847)

[6 Xây dựng báo cáo 64](#_Toc204439848)

[6.1 Dashboard và report 64](#_Toc204439849)

[6.2 Xây dựng báo cáo 65](#_Toc204439850)

[6.2.1 Dashboard vs Report 65](#_Toc204439851)

[6.2.2 Dashboard 66](#_Toc204439852)

[7 KẾT LUẬN 72](#_Toc204439853)

[7.1 Báo cáo 72](#_Toc204439854)

[7.1.1 Các bước viết báo cáo 72](#_Toc204439855)

[7.1.2 Tổng hợp 73](#_Toc204439856)

[7.2 Khó khăn 75](#_Toc204439857)

[7.3 Thuận lợi 76](#_Toc204439858)

[7.4 Hướng phát triển 76](#_Toc204439859)

[8 THAM KHẢO 78](#_Toc204439860)

# Giới thiệu dự án

## Giới thiệu

* **Hiện trạng**

Trong bối cảnh công nghệ ngày càng phát triển, làm việc từ xa không còn là giải pháp tạm thời mà đang trở thành một xu hướng phổ biến trên toàn thế giới. Nhiều doanh nghiệp, đặc biệt trong lĩnh vực công nghệ thông tin, đã mạnh dạn áp dụng mô hình làm việc từ xa toàn phần hoặc kết hợp làm việc tại nhà và tại văn phòng.

Sự thay đổi này làm thay đổi đáng kể cách nhân viên tổ chức công việc, giao tiếp và quản lý thời gian cá nhân. Đáng chú ý hơn, nó cũng làm thay đổi cách doanh nghiệp đánh giá hiệu suất lao động và mức độ hài lòng trong công việc.

Làm việc từ xa mang lại những lợi ích rõ ràng như linh hoạt thời gian, giảm chi phí vận hành và tiếp cận nhân lực toàn cầu. Tuy nhiên, đi cùng với đó là những thách thức không thể bỏ qua: nhân viên dễ bị cô lập, giảm tương tác nhóm, khó kiểm soát tiến độ công việc và ảnh hưởng đến tinh thần làm việc.

Thực tế này đặt ra một câu hỏi quan trọng cho các nhà quản lý và cả người lao động:

**Làm việc từ xa có thật sự giúp nâng cao hiệu suất như kỳ vọng trong thời đại công nghệ số, hay chỉ phù hợp với một nhóm người nhất định?**

Đây chính là vấn đề cốt lõi mà dự án hướng tới nhằm tìm ra câu trả lời dựa trên phân tích khách quan và dữ liệu thực tế.

* **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)
* Số giờ làm việc mỗi tuần (WorkWeekHrs)
* Thu nhập hàng năm (CompTotal)
* 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 Excel (.xlsx), gồm hai tệp:

* **survey\_results\_public.xlsx:** chứa dữ liệu khảo sát gốc
* **survey\_results\_schema.xlsx:** 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.

## Yêu cầu của công ty

* **Về mặt dữ liệu:**

Công ty mong muốn phân tích hiệu suất làm việc của nhân sự trong lĩnh vực công nghệ thông tin, từ đó đánh giá tác động của hình thức làm việc (tại văn phòng, hybrid, làm việc từ xa) đến hiệu suất và sự hài lòng của nhân viên thông các chỉ số thời gian làm việc trung bình, thu nhập hàng năm, hình thức làm việc, trình độ chuyên môn và mức độ hài lòng.

* **Về quản lý và lưu trữ:**

Để đảm bảo tính hiệu quả và chuyên nghiệp trong quá trình phân tích, công ty yêu cầu dữ liệu phải được **xử lý, lưu trữ và trình bày bằng các công cụ phù hợp với chuyên ngành Xử lý Dữ liệu**.

Cụ thể:

* **Sử dụng Python** để làm sạch và tiền xử lý dữ liệu đầu vào.
* **Trực quan hóa dữ liệu bằng Tableau**, thiết kế các báo cáo phân tích dưới dạng **dashboard có tính tương tác cao**, giúp người dùng dễ dàng khám phá thông tin và so sánh các nhóm dữ liệu.
* **Hệ thống hóa việc lưu trữ**: toàn bộ dữ liệu gốc, dữ liệu đã xử lý, và dashboard phân tích phải được lưu trữ có cấu trúc, đảm bảo dễ dàng truy xuất, chia sẻ và phục vụ cho công tác báo cáo định kỳ hoặc đột xuất.

Thông qua yêu cầu này, công ty mong muốn thiết lập một quy trình xử lý dữ liệu **chặt chẽ – minh bạch – dễ mở rộng**, phù hợp với định hướng phát triển và ứng dụng dữ liệu trong quản trị nhân sự.

* **Mục tiêu:**

Dự án hướng đến việc **phân tích và đánh giá hiệu suất làm việc theo từng hình thức** (tại văn phòng, kết hợp, từ xa), nhằm giúp công ty **xác định mô hình làm việc nào đang mang lại hiệu quả và mức độ hài lòng cao nhất cho nhân viên**.

Từ kết quả phân tích, công ty có thể:

* Đưa ra **quyết định chiến lược** trong việc mở rộng hoặc điều chỉnh mô hình làm việc linh hoạt
* **Tối ưu hóa chính sách đãi ngộ và quản lý nhân sự** dựa trên dữ liệu thực tế
* Xây dựng kế hoạch **phát triển nguồn nhân lực** phù hợp với xu hướng làm việc hiện đại và bối cảnh chuyển đổi số

Mục tiêu cuối cùng là giúp doanh nghiệp **nâng cao hiệu quả vận hành**, đồng thời tạo ra môi trường làm việc phù hợp, bền vững và thu hút nhân tài trong lĩnh vực công nghệ.

* **Đánh giá tính khả thi:**
* **Năng lực hiện có của nhóm:**

Nhóm đã được trang bị nền tảng kiến thức vững chắc về xử lý dữ liệu, bao gồm:

* Hiểu cấu trúc dữ liệu dạng bảng và các kỹ thuật làm sạch dữ liệu cơ bản
* Sử dụng thành thạo các công cụ như **Python** và **Tableau** trong phân tích và trực quan hóa dữ liệu
* Có kinh nghiệm làm việc nhóm, phân tích yêu cầu và khai thác thông tin từ dữ liệu thực tế
* **Kỹ năng học thêm trong quá trình làm dự án:**

Thông qua quá trình làm dự án, nhóm kỳ vọng nâng cao thêm các kỹ năng sau:

* Xử lý dữ liệu thực tế có quy mô lớn, bao gồm làm quen với cấu trúc phức tạp, xử lý giá trị khuyết (missing values), phát hiện và xử lý ngoại lai (outliers), cũng như chuẩn hóa định dạng để đảm bảo dữ liệu nhất quán và có thể phân tích được.
* Thiết kế dashboard tương tác linh hoạt trên Tableau, cho phép phân tích theo từng nhóm đối tượng như vị trí địa lý, hình thức làm việc, chức danh công việc,... giúp người dùng cuối dễ dàng khai thác và so sánh thông tin.
* Phát triển kỹ năng kể chuyện bằng dữ liệu (data storytelling) nhằm truyền tải kết quả phân tích một cách trực quan, thuyết phục và dễ hiểu, hỗ trợ hiệu quả cho việc ra quyết định của nhà quản lý.

## Lập kế hoạch dự án

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TT** | **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 | Phân tích yêu cầu KH | 22/05/2025 | 02/06/2025 | Xác định được mục tiêu trọng tâm và phân tích |
| 2.2 | 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 | Dữ liệu rời rạc và còn thiếu nhiều yếu tố quan trọng |
| 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 | 04/08/2025 |  |
| 6.1 | Tạo Slide thuyết trình | 27/06/2025 | 04/08/2025 | Slide hoàn chỉnh |
| 6.2 | Viết báo cáo document | 25/05/2025 | 04/08/2025 | Báo cáo hoàn chỉnh |

# Phân tích yêu cầu khách hàng

## Phân tích yêu cầu

* **Dữ liệu:**

Dữ liệu đầu vào của dự án là tệp “survey\_results\_public.xlsx”, được trích xuất từ cuộc khảo sát thường niên Stack Overflow Developer Survey 2024 – một trong những khảo sát uy tín và quy mô lớn nhất dành cho lập trình viên và chuyên gia công nghệ trên toàn thế giới, với hơn 70.000 phản hồi từ nhiều quốc gia.

Bộ dữ liệu phản ánh chân thực thực trạng làm việc trong ngành công nghệ thông tin, với nhiều trường dữ liệu quan trọng như:

* RemoteWork: Hình thức làm việc (tại văn phòng, kết hợp, từ xa)
* Employment: Tình trạng việc làm
* DevType: Loại công việc đảm nhiệm
* WorkWeekHrs: Số giờ làm việc trung bình mỗi tuần
* JobSat: Mức độ hài lòng với công việc
* CompTotal: Tổng thu nhập hằng năm
* Country, YearsCodePro, EdLevel, cùng nhiều trường bổ trợ khác

Đây là nguồn dữ liệu có độ tin cậy cao và phạm vi rộng, cho phép nhóm thực hiện phân tích chuyên sâu về mối liên hệ giữa hình thức làm việc và hiệu suất, đồng thời làm rõ mức độ ảnh hưởng của làm việc từ xa đến sự hài lòng của nhân viên trong môi trường công nghệ hiện đại.

* **Quản lý và lưu trữ:**

Để tạo ra những phân tích dễ hiểu và phục vụ tốt cho việc ra quyết định, dữ liệu cần được xử lý và tổ chức một cách khoa học. Nhóm dự án sử dụng các công cụ phổ biến và hiệu quả trong ngành để quản lý và làm sạch dữ liệu.

* Excel được dùng để xem và kiểm tra sơ bộ dữ liệu ban đầu, giúp phát hiện các lỗi như giá trị bị thiếu, định dạng sai hoặc thông tin chưa hợp lý.
* Python là công cụ chính để xử lý dữ liệu. Với các thư viện như pandas và numpy, nhóm thực hiện:
  + Dọn dẹp dữ liệu: xóa hoặc xử lý các ô trống, giá trị bất thường hoặc sai lệch
  + Đổi định dạng: làm cho dữ liệu đồng nhất, dễ hiểu
  + Tổng hợp thông tin cần thiết: ví dụ như nhóm dữ liệu theo quốc gia, hình thức làm việc hay chức danh công việc
* Sau khi dữ liệu được xử lý xong, nhóm lưu lại dưới dạng .csv hoặc .xlsx, là các định dạng quen thuộc, dễ sử dụng trong các công cụ báo cáo.
* Cuối cùng, dữ liệu được đưa vào phần mềm Tableau để tạo thành các biểu đồ và dashboard có thể tương tác, giúp người xem dễ dàng lọc và khám phá thông tin theo nhu cầu.

Việc xử lý và trình bày dữ liệu theo cách này giúp đảm bảo rằng ngay cả người không có kiến thức chuyên sâu về dữ liệu cũng có thể hiểu được nội dung phân tích, từ đó đưa ra các quyết định chính xác và kịp thời.

* **Công nghệ:**

**Dựa trên các yêu cầu thực tế từ doanh nghiệp và mục tiêu phân tích đề tài, nhóm đã cân nhắc các yếu tố sau:**

* Cần thực hiện phân tích định lượng về hiệu quả làm việc từ xa dựa trên nhiều chỉ số như giờ làm việc, thu nhập, mức độ hài lòng, loại công việc, v.v.
* Phải đảm bảo khả năng trực quan hóa kết quả một cách rõ ràng, hỗ trợ nhà quản lý dễ dàng hiểu và ra quyết định.
* Bộ dữ liệu có quy mô lớn, nhiều trường thông tin, yêu cầu xử lý linh hoạt và phân tích theo từng nhóm đối tượng (quốc gia, vị trí công việc, trình độ,...).
* Dữ liệu đầu vào cần có tính khách quan, toàn cầu và đáng tin cậy.
* Kết quả cần được trình bày sinh động, dễ hiểu, phục vụ cả người có chuyên môn và không chuyên môn.

**Từ những yêu cầu trên, nhóm đã đưa ra quyết định lựa chọn công nghệ như sau:**

* **Nguồn dữ liệu: Stack Overflow Developer Survey 2024**
* Được chọn nhờ tính uy tín, quy mô toàn cầu và mức độ cập nhật với thực tế ngành công nghệ thông tin.
* Bao quát nhiều khía cạnh như hình thức làm việc, mức thu nhập, thời gian làm việc, kỹ năng, quốc gia, v.v.
* Phù hợp để phân tích xu hướng làm việc từ xa trong bối cảnh hiện đại.
* **Quản lý và xử lý dữ liệu: Excel + Python**
* Excel: hỗ trợ thao tác nhanh với dữ liệu ban đầu, đánh giá sơ bộ và chuẩn bị cho bước xử lý chuyên sâu.
* Python (pandas, numpy,…): đảm bảo hiệu suất xử lý dữ liệu lớn, hỗ trợ làm sạch, chuẩn hóa, xử lý ngoại lai, và tổng hợp chỉ số phân tích linh hoạt theo nhiều tiêu chí.
* **Trực quan hóa dữ liệu: Tableau Public / Desktop**
* Tableau được lựa chọn nhờ khả năng thiết kế các dashboard tương tác mạnh mẽ, hỗ trợ lọc dữ liệu theo quốc gia, loại hình công việc, kinh nghiệm,...
* Giao diện trực quan, dễ sử dụng, phù hợp cho cả người dùng chuyên môn và không chuyên môn.
* **Trình bày kết quả: Canva + Microsoft Word**
* Canva: dùng để thiết kế slide trình bày chuyên nghiệp với hình ảnh minh họa, sơ đồ và biểu tượng trực quan. Giúp buổi thuyết trình sinh động, thu hút và dễ tiếp cận với nhiều đối tượng người xem.
* Microsoft Word: dùng để biên soạn báo cáo chi tiết, đầy đủ các phần: giới thiệu, phân tích dữ liệu, biểu đồ, nhận xét và kết luận. Đảm bảo định dạng chuẩn, rõ ràng và dễ theo dõi.

Nhờ sự phối hợp hợp lý giữa các công cụ, nhóm có thể đảm bảo tính khoa học trong xử lý dữ liệu, tính trực quan trong trình bày kết quả và tính thuyết phục trong việc hỗ trợ ra quyết định đúng với mục tiêu ban đầu mà doanh nghiệp đặt ra.

## Câu chuyện dữ liệu

### Đặt vấn đề

* **Mô tả thực trạng:**

Trong kỷ nguyên số, làm việc từ xa không còn là phương án tạm thời mà đã trở thành chuẩn mực mới, đặc biệt trong ngành công nghệ thông tin. Sau đại dịch, hàng loạt doanh nghiệp chuyển sang các mô hình linh hoạt như làm từ xa, kết hợp (hybrid) hoặc làm tại văn phòng tùy theo mục tiêu vận hành và chiến lược nhân sự.

Các mô hình này đem lại nhiều lợi ích:

* Giảm thời gian di chuyển
* Mở rộng cơ hội nghề nghiệp không giới hạn địa lý
* Tăng tính linh hoạt và sự chủ động

Nhưng cũng đi kèm thách thức:

* Giảm kết nối nhóm và giao tiếp trực tiếp
* Khó kiểm soát hiệu suất công việc
* Tăng cảm giác cô lập, đặc biệt với người mới

Hai nhóm đối tượng cần lời giải từ dữ liệu:

* **Doanh nghiệp công nghệ**: Đâu là mô hình tối ưu vừa đảm bảo hiệu suất, vừa duy trì sự gắn kết trong đội ngũ?
* **Sinh viên mới tốt nghiệp**: Nên chọn hình thức làm việc nào để dễ thích nghi và phát triển sự nghiệp trong môi trường hiện đại?

Từ những câu hỏi này, dự án được thực hiện nhằm khai thác dữ liệu thực tế để phân tích:

**“Liệu làm việc từ xa có thực sự hiệu quả trong ngành công nghệ, và ai là người phù hợp nhất với mô hình này?”**

* **Dữ liệu liên quan:**

Nguồn dữ liệu của dự án được trích xuất từ Stack Overflow Developer Survey 2024 – một khảo sát quy mô toàn cầu với hàng chục nghìn lập trình viên và chuyên gia công nghệ tham gia. Mỗi phản hồi là một lát cắt giá trị về xu hướng làm việc hiện nay: từ hình thức công tác (tại văn phòng, kết hợp, từ xa), mức độ hài lòng, thời gian làm việc trung bình, thu nhập hằng năm, đến năng lực chuyên môn, kinh nghiệm và trình độ học vấn.

Dẫu phong phú và đa chiều, dữ liệu thô chỉ thực sự phát huy giá trị khi được xử lý và chuyển hóa thành thông tin có ý nghĩa. Đó là lý do nhóm thực hiện các bước:

* Làm sạch, tổng hợp và phân tích dữ liệu bằng Python
* Trực quan hóa thông tin bằng Tableau
* Trình bày báo cáo qua Word và Canva, đảm bảo tính dễ hiểu, ứng dụng cao và phù hợp từng đối tượng người đọc.
* **Truyền tải mục tiêu dự án rõ ràng là chìa khóa kết nối hiệu quả với từng đối tượng**
* **Doanh nghiệp công nghệ:**
* Sử dụng biểu đồ cột nhóm và biểu đồ tròn có nhãn rõ ràng để so sánh dữ liệu.
* Chú thích các label và trendline rõ ràng, dễ nhận biết để làm nổi bật xu hướng vận hành và định hướng phát triển của dự án
* Đưa ra giải pháp dựa trên phân tích dữ liệu thực tế, hỗ trợ ra quyết định.
* **Sinh viên mới ra trường:**
* Thuyết trình bằng ngôn ngữ đơn giản: sử dụng từ ngữ thân thiện, dễ hiểu, tránh thuật ngữ chuyên ngành để sinh viên dễ tiếp thu.
* Biểu đồ dễ hiểu và dễ nhận biết: sử dụng các loại biểu đồ thông dụng như biểu đồ cột, biểu đồ nhóm, scatter chart kèm nhãn (label) rõ ràng, giúp người chưa có kiến thức chuyên môn cũng dễ nắm bắt thông tin và kết luận.
* **Mục tiêu:**

Câu chuyện dữ liệu hướng đến các mục tiêu sau:

* **Truyền đạt thông điệp chính một cách rõ ràng**: phân tích mối liên hệ giữa mô hình làm việc (từ xa, kết hợp, tại văn phòng) với hiệu suất, mức độ hài lòng và thu nhập của lập trình viên.
* **Hỗ trợ ra quyết định** cho hai nhóm đối tượng:
  + **Doanh nghiệp**: chọn mô hình làm việc tối ưu cho nhân viên.
  + **Sinh viên mới ra trường**: dịnh hướng chọn môi trường phù hợp để phát triển sự nghiệp.
* **Trình bày trực quan, dễ hiểu**: sử dụng biểu đồ đơn giản, ngôn ngữ thân thiện để tăng tính tiếp cận và ứng dụng của dữ liệu.

### Xác định câu chuyện

Dựa trên dữ liệu từ khảo sát Stack Overflow 2024, nhóm xây dựng câu chuyện xoay quanh hiệu suất làm việc từ xa qua các câu hỏi chính:

* Làm việc từ xa ảnh hưởng thế nào đến hiệu suất (giờ làm việc)?
* Hình thức làm việc có tác động đến mức độ hài lòng?
* Thu nhập thay đổi ra sao giữa các mô hình?
* Ai là người phù hợp nhất với làm việc từ xa?

Mục tiêu là giúp doanh nghiệp đưa ra chính sách linh hoạt, và hỗ trợ sinh viên hiểu môi trường làm việc thực tế trong ngành CNTT.

* **Mục tiêu cụ thể của nhóm:**
* Xác định mô hình làm việc tối ưu về hiệu suất và hài lòng.
* Phân nhóm đối tượng phù hợp theo vị trí, kinh nghiệm, mức lương, quốc gia,…
* Chuyển dữ liệu thành thông tin dễ hiểu, dễ áp dụng.
* Thúc đẩy tư duy ra quyết định dựa trên dữ liệu.
* Trình bày sinh động qua Python, Tableau, Canva.
* **Giải pháp triển khai:**
* **Xây dựng dashboard phân tích**: trực quan hóa hiệu suất, thu nhập và mức độ hài lòng theo từng mô hình làm việc bằng Tableau, giúp dễ dàng theo dõi và so sánh.
* **Phân nhóm đối tượng phù hợp**: dựa vào độ tuổi, kinh nghiệm, vai trò, quốc gia để xác định nhóm phù hợp với từng mô hình làm việc (remote, hybrid, In-Person).
* **Đề xuất chính sách linh hoạt**: gợi ý doanh nghiệp điều chỉnh mô hình làm việc theo đặc điểm nhân sự để tối ưu hiệu suất và giữ chân nhân tài.
* **Tư vấn hướng nghiệp**: trình bày thông tin bằng infographic, biểu đồ dễ hiểu giúp sinh viên lựa chọn môi trường làm việc phù hợp với bản thân.
* **Thiết lập báo cáo định kỳ**: hỗ trợ doanh nghiệp cập nhật xu hướng làm việc và hài lòng nhân sự để kịp thời điều chỉnh chiến lược.
* **Kỳ vọng đạt được:**
* Hiểu rõ xu hướng và nhu cầu làm việc.
* Xây dựng chính sách nhân sự linh hoạt.
* Tăng hiệu suất, giữ chân nhân tài.
* Đẩy mạnh ứng dụng dữ liệu vào quản lý.
* **Cách tiếp cận dữ liệu:**
* **Phân tích tương quan** giữa mô hình làm việc với hài lòng, thu nhập, giờ làm.
* **Phân tích xu hướng** ngành nghề, trình độ.
* **So sánh nhóm** làm việc remote – hybrid – In-Person
* **Phân khúc dữ liệu** theo vai trò, kỹ năng.
* **Định lượng bằng thống kê mô tả** (tỉ lệ)
* **Trực quan hóa dữ liệu** qua biểu đồ cột, scatter, biểu đồ tròn,… để kể chuyện hấp dẫn, dễ hiểu

### Xác định rõ đối tượng

* **Đối tượng mục tiêu của báo cáo:**
* **Doanh nghiệp công nghệ thông tin**: là đơn vị trực tiếp lựa chọn mô hình làm việc, doanh nghiệp cần dữ liệu để đánh giá hiệu quả, tối ưu hiệu suất và xây dựng chính sách nhân sự linh hoạt.
* **Sinh viên sắp ra trường ngành CNTT**: là lực lượng lao động tương lai, cần hiểu rõ đặc điểm từng hình thức làm việc để chọn môi trường phù hợp với năng lực và định hướng nghề nghiệp.
* **Đặc điểm của nhóm đối tượng:**
* **Doanh nghiệp công nghệ thông tin:**
* Đã có hiểu biết chuyên môn và kinh nghiệm vận hành, quan tâm đến hiệu suất, sự hài lòng và mức độ phù hợp của nhân viên với từng mô hình làm việc.
* Cần báo cáo có dữ liệu đáng tin cậy, trình bày trực quan, dễ tra cứu và hỗ trợ ra quyết định nhân sự một cách nhanh chóng.
* **Sinh viên sắp tốt nghiệp ngành CNTT:**
  + Có nền tảng kiến thức chuyên ngành nhưng ít trải nghiệm thực tế trong môi trường làm việc.
  + Quan tâm đến xu hướng làm việc, mức thu nhập và khả năng phát triển theo từng mô hình. Ưu tiên nội dung dễ hiểu, sinh động, có hình ảnh minh họa, biểu đồ trực quan thay vì bảng số liệu khô khan.

### Xác định câu chuyện chi tiết

* **Bối cảnh:**

Làm việc từ xa, kết hợp và tại văn phòng đang trở thành ba lựa chọn phổ biến trong ngành công nghệ. Tuy nhiên, mỗi mô hình mang đến hiệu quả khác nhau tùy vào đối tượng và hoàn cảnh.

Doanh nghiệp cần dữ liệu để chọn mô hình phù hợp, sinh viên sắp ra trường cần hiểu rõ môi trường làm việc tương lai. Việc kể một câu chuyện bằng dữ liệu là cách hiệu quả để kết nối nhu cầu của cả hai nhóm và đưa ra quyết định đúng đắn.

* **Đối tượng sử dụng báo cáo:**

Báo cáo hướng đến hai nhóm đối tượng chính:

* **Nhà quản lý và chuyên viên nhân sự trong lĩnh vực công nghệ:** sử dụng kết quả phân tích để đánh giá mô hình làm việc hiệu quả, hỗ trợ xây dựng chính sách nhân sự linh hoạt và phù hợp với từng nhóm nhân viên.
* **Sinh viên năm cuối ngành CNTT**: tham khảo để hiểu rõ đặc điểm, ưu – nhược điểm của các hình thức làm việc, từ đó đưa ra lựa chọn phù hợp với năng lực và định hướng nghề nghiệp cá nhân.
* **Cấu trúc kể chuyện dữ liệu:**

1. **Đặt vấn đề – Tại sao cần phân tích?**

Nêu thực trạng các mô hình làm việc trong ngành CNTT. Đặt câu hỏi trung tâm: “Làm việc từ xa có thực sự hiệu quả? Và ai phù hợp với mô hình này?”

1. **Giới thiệu dữ liệu**

Trình bày nguồn dữ liệu từ Stack Overflow Survey 2024, tập trung vào các trường: hình thức làm việc, thu nhập, mức độ hài lòng, giờ làm việc, kinh nghiệm,…

1. **Phân tích dữ liệu để kể chuyện**

So sánh hiệu suất và hài lòng giữa remote – hybrid – In-person

1. **Phân tích theo độ tuổi, nghề nghiệp, quốc gia**

* Tìm mối liên hệ giữa mô hình làm việc với thu nhập và kinh nghiệm
* Trực quan hóa – Truyền tải thông điệp
* Sử dụng biểu đồ, dashboard, infographic giúp người xem dễ hiểu, dễ nhớ và dễ hành động.

1. **Thông điệp & ứng dụng thực tiễn**

* Doanh nghiệp: điều chỉnh chính sách làm việc và nhân sự dựa trên dữ liệu
* Sinh viên: xác định mô hình phù hợp để định hướng nghề nghiệp hiệu quả hơn

### 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:** ưu tiên các hình thức làm việc phổ biến

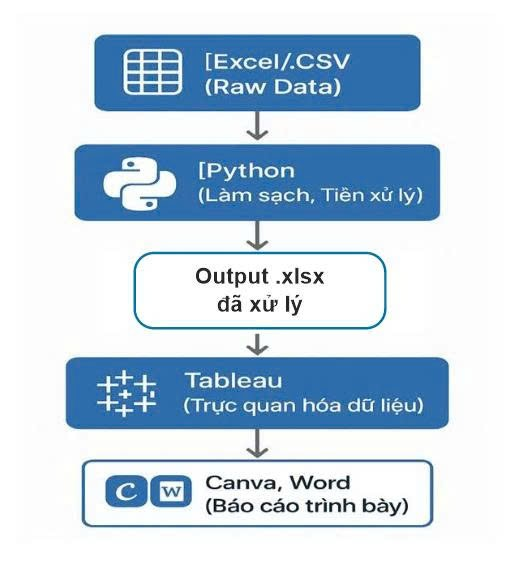
Tất cả biểu đồ được xây dựng bằng Tableau, Excel đảm bảo sinh động, dễ tiếp cận và hỗ trợ truyền tải thông điệp nhanh chóng.

### 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.

## Kiến trúc hệ thống

### Kiến trúc



### Giải thích

Sơ đồ thể hiện hành trình "biến dữ liệu thô thành câu chuyện trực quan" một cách logic, mạch lạc:

1. **Bắt đầu từ dữ liệu gốc (.CSV/.Excel)**  
   Dữ liệu khảo sát từ Stack Overflow 2024 là nguyên liệu đầu vào quan trọng, chưa qua xử lý.
2. **Tiền xử lý bằng Python**  
   Tận dụng sức mạnh của Python để làm sạch, chuẩn hóa và lọc dữ liệu – giúp đảm bảo đầu ra chính xác và đáng tin cậy.
3. **Xuất dữ liệu đã xử lý (.xlsx)**  
   Đây là bản dữ liệu đã được làm sạch và chuẩn hóa hoàn chỉnh, sẵn sàng để phân tích và trực quan hóa.
4. **Trực quan hóa bằng Tableau**  
   Dữ liệu được chuyển hóa thành biểu đồ, dashboard sinh động, dễ hiểu – hỗ trợ kể chuyện bằng hình ảnh.
5. **Trình bày chuyên nghiệp với Canva & Word**  
   Kết quả cuối cùng được đóng gói trong báo cáo đẹp mắt, dễ tiếp cận, phù hợp với cả doanh nghiệp và sinh viên.

## Giải thích về bộ dữ liệu khách hàng

### 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
* **Hiệu suất làm việc (Work Efficiency)**  
  Được đo lường gián tiếp qua số giờ làm việc mỗi tuần (WorkWeekHrs). Giá trị cao hay thấp phản ánh cường độ và khả năng duy trì hiệu suất.
* **Mức độ hài lòng (Job Satisfaction)**  
  Thể hiện cảm nhận cá nhân về công việc hiện tại. Dữ liệu khảo sát theo dạng lựa chọn: “Very satisfied”, “Somewhat dissatisfied”,...
* **Thu nhập hàng năm (Annual Compensation)**  
  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,...

### Các trường dữ liệu

Bộ dữ liệu sử dụng trong dự án được trích xuất từ khảo sát **Stack Overflow Developer Survey 2024**, một nguồn dữ liệu uy tín toàn cầu về cộng đồng lập trình viên. Sau quá trình làm sạch và chuẩn hóa, bộ dữ liệu gồm 36 trường, phản ánh các yếu tố về đặc điểm cá nhân, công việc và mức độ hài lòng, từ đó phục vụ trực tiếp cho việc phân tích hiệu suất làm việc từ xa.

|  |  |
| --- | --- |
| **Tên trường** | **Giải thích** |
| ResponseId | Mã định danh duy nhất cho mỗi người tham gia khảo sát |
| MainBranch | Vai trò chính của người tham gia trong ngành (ví dụ: lập trình viên, quản lý) |
| Age | Tuổi của người trả lời khảo sát |
| Country | Quốc gia nơi người tham gia đang sinh sống/làm việc |
| EdLevel | Trình độ học vấn |
| WorkExp | Số năm kinh nghiệm làm việc thực tế |
| CompTotal | Tổng thu nhập hàng năm (theo tiền tệ gốc) |
| Currency | Loại tiền tệ được sử dụng để khai báo thu nhập |
| RemoteWork | Hình thức làm việc (In-person, Remote, Hybrid) |
| Employment | Loại hình công việc (toàn thời gian, bán thời gian, tự do...) |
| OrgSize | Quy mô tổ chức/doanh nghiệp mà người tham gia đang làm việc |
| JobSat | Mức độ hài lòng với công việc hiện tại |
| TimeSearching | Thời gian trung bình mỗi ngày dành để tìm kiếm thông tin kỹ thuật |
| TimeAnswering | Thời gian trung bình mỗi ngày dành để giúp đỡ người khác |
| Frustration | Mức độ khó chịu khi gặp vấn đề trong công việc |
| SurveyLength | Đánh giá độ dài khảo sát (ngắn, trung bình, dài) |
| SurveyEase | Đánh giá mức độ dễ trả lời khảo sát |
| YearsCodePro | Số năm làm việc chuyên nghiệp với lập trình |
| DevType | Loại công việc kỹ thuật đang làm (backend, frontend, fullstack...) |
| BuildvsBuy | Xu hướng ưu tiên tự phát triển hay mua công nghệ có sẵn |
| TechEndorse | Kinh nghiệm đề xuất hoặc phản hồi về công nghệ cho người khác |
| NEWCollabToolsHaveWorkedWith | Các công cụ cộng tác từng sử dụng |
| NEWCollabToolsWantToWorkWith | Các công cụ cộng tác muốn được sử dụng trong tương lai |
| OpSysProfessional use | Hệ điều hành chính dùng cho công việc |
| OfficeStackAsyncHaveWorkedWith | Các công cụ làm việc không đồng bộ từng sử dụng |
| OfficeStackSyncHaveWorkedWith | Các công cụ làm việc đồng bộ từng sử dụng |
| SOComm | Mức độ tham gia cộng đồng Stack Overflow |
| AISelect, AISent, AIBen, AIAcc, AIToolCurrently Using | Các trường liên quan đến việc sử dụng AI trong công việc |
| TBranch | Lĩnh vực công nghệ chính đang làm |
| ICorPM | Vị trí: cá nhân phát triển hay quản lý dự án |
| BranchGroup | Nhóm phân loại lĩnh vực làm việc (gộp từ TBranch) |
| CompTotalUSD | Thu nhập quy đổi sang USD để dễ so sánh giữa các quốc gia |

# Làm sạch và chuyển đổi 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.

## 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()

### 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.

### Giải pháp phân bố dữ liệu

#### 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.

#### 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.

## Làm sạch dữ liệu

### 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), nhất là ở câu hỏi mở về thu nhập, hài lòng, công việc.
* **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.

### 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.

### Các bước làm sạch dữ liệu

#### Trình bày các bước làm sạch

Quá trình làm sạch dữ liệu trong dự án gồm 6 bước chính:

1. **Xử lý dữ liệu thiếu**
   * Xác định các ô trống (NaN)
   * Loại bỏ dòng thiếu toàn phần hoặc điền giá trị bằng trung bình/trung vị
   * Áp dụng mô hình dự đoán nếu dữ liệu đủ điều kiện
2. **Loại bỏ dữ liệu trùng lặp**
   * Sử dụng công cụ (Excel, Python) để tìm và xóa các bản ghi lặp lại
3. **Chuẩn hóa định dạng**
   * Đồng nhất chữ hoa/thường
   * Chuẩn hóa định dạng ngày tháng (YYYY-MM-DD)
   * Định dạng email, số điện thoại theo chuẩn thống nhất
4. **Xử lý giá trị ngoại lai (Outliers)**
   * Dùng IQR, Z-score để phát hiện
   * Loại bỏ hoặc thay thế nếu ảnh hưởng tiêu cực đến phân tích
5. **Mã hóa dữ liệu** *(nếu cần)*
   * Sử dụng Label encoding, One-hot encoding hoặc phân nhóm (binning)
   * Đảm bảo dữ liệu định tính có thể sử dụng cho phân tích định lượng
6. **Kiểm tra tính logic & nhất quán**
   * Loại bỏ dữ liệu sai logic
   * Đảm bảo sự thống nhất giữa các cột và giữa các nguồn dữ liệu

#### 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

        # Nếu cần 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

## Chuyển đổi dữ liệu

### Các trường hợp cần chuyển đổi

Trong quá trình xử lý và phân tích dữ liệu từ khảo sát Stack Overflow 2024, nhóm nhận thấy nhiều trường dữ liệu chưa ở định dạng phù hợp để phân tích hoặc trực quan hóa. Việc chuyển đổi là cần thiết trong các trường hợp sau:

* Chuẩn hóa định dạng: các trường chữ có sự không đồng nhất về viết hoa/thường, ký hiệu hoặc định dạng ngày tháng.
* Gom nhóm và rút trích: một số cột cần chia nhóm hoặc trích lọc thông tin chính để đơn giản hóa việc xử lý.
* Tạo biến mới: tạo thêm các trường biến đổi từ dữ liệu ban đầu để phục vụ phân tích
* Chuyển đổi đơn vị đo lường: như chuyển đổi tiền tệ, thời gian sang đơn vị thống nhất.
* Mã hóa dữ liệu định tính: một số cột văn bản cần mã hóa về dạng số để thuận tiện trong thao tác, tính toán.

### 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.

### 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 3.1***

# Xử lý dữ liệu

## Chuẩn hóa dữ liệu

### 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, Other.

    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 vấnthành 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()

        return self

* **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

## Mô hình hóa dữ liệu

### 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.

### 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).

### 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.

### 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")

1. **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:

1. **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', 'AISent', 'AIBen', 'AIAcc', 'AIToolCurrently Using',

            'NEWCollabToolsHaveWorkedWith', 'BuildvsBuy', 'TechEndorse'

        ]

        return self.df[fact\_cols]

1. **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
2. **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()

1. **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}")

1. **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***

# Trực quan hóa dữ liệu

## Các kỹ thuật trực quan hóa

* **Một số kỹ thuật trực quan hóa phổ biến**
* **Biểu đồ cột (Bar/Column Plot):** so sánh các nhóm dữ liệu rõ ràng, trực giác (ví dụ: hiệu suất theo hình thức làm việc).
* **Histogram:** phân tích phân phối dữ liệu, hữu ích cho biến liên tục như thu nhập, thời gian làm việc.
* **Biểu đồ đường (Line Plot):** theo dõi xu hướng theo thời gian hoặc nhóm liên tục.
* **Scatter Plot:** hiển thị mối liên hệ giữa hai biến (ví dụ: kinh nghiệm vs thu nhập).
* **Heatmap:** trực quan hóa mối tương quan giữa các biến định tính và định lượng.
* **Box Plot:** hiển thị phân phối dữ liệu và phát hiện các giá trị ngoại lai.
* **Treemap :** dùng để hiển thị dữ liệu phân cấp hoặc dữ liệu theo vùng địa lý.

Trong dự án, nhóm linh hoạt áp dụng nhiều kỹ thuật trực quan hóa phù hợp với đặc điểm của từng loại dữ liệu và mục tiêu phân tích:

* **TreeMap:** trực quan độ tuổi phù hợp với làm việc từ xa theo cấp bậc.
* **Biểu đồ tròn:** minh họa cơ cấu công ty và tỷ lệ các hình thức làm việc.
* **Scatter Plot:** phân tích mối quan hệ giữa số năm kinh nghiệm và thu nhập.
* **Biểu đồ Pareto:** làm nổi bật xu hướng lựa chọn hình thức làm việc phổ biến, hỗ trợ ra quyết định chiến lược.
* **Bảng dữ liệu:** tổng hợp số lượng người và mức độ hài lòng theo từng hình thức làm việc.
* **Biểu đồ kết hợp (cột + đường):** so sánh mức độ hài lòng và thu nhập trung bình giữa các hình thức làm việc.
* **Biểu đồ cột:** thể hiện mức độ hài lòng trung bình theo từng vị trí công việc,…

Ngoài ra, dự án còn sử dụng:

* **Card thông tin:** hiển thị nhanh các chỉ số quan trọng như tổng số người khảo sát, tỷ lệ làm việc từ xa, thu nhập trung bình,...
* **Biểu đồ phụ trợ:** gắn kèm để làm rõ xu hướng và bổ sung bối cảnh cho phân tích chính.

## Các nguyên tắc trực quan hóa

* **Nguyên tắc trực quan hóa**

1. **Chọn đúng loại biểu đồ**  
   Ưu tiên mục tiêu phân tích, không chọn biểu đồ theo cảm tính. “Hình thức phải phục vụ chức năng”.
2. **Tập trung vào dữ liệu quan trọng**  
   Không đưa toàn bộ dataset vào biểu đồ, chỉ chọn thông tin nổi bật và dễ hiểu.
3. **Thể hiện đúng bản chất số liệu**  
   Không bóp méo tỷ lệ hay làm sai lệch tương quan dữ liệu thực tế.
4. **Sử dụng màu sắc thông minh**  
   Màu dùng để nhấn mạnh, không dùng quá nhiều gây rối mắt hay khó hiểu.
5. **Giản lược và hiệu quả**  
   Tránh làm biểu đồ phức tạp quá mức, hãy để dữ liệu “nói” bằng cách trình bày gọn gàng, rõ ràng.

Trong dự án “***Phân tích hiệu suất làm việc từ xa”***, nhóm làm việc với nhiều loại dữ liệu khác nhau như độ tuổi, hình thức làm việc, mức độ hài lòng và thu nhập. Vì vậy, **việc lựa chọn đúng biểu đồ** đóng vai trò then chốt giúp:

* Truyền tải thông tin **rõ ràng, dễ hiểu**
* **Làm nổi bật insight** cần thiết
* Tránh gây hiểu nhầm cho người xem

Bên cạnh đó, dự án còn **kết hợp các filter, animation và hiệu ứng dashboard trong Tableau** để tăng tính tương tác. Nhờ đó, người dùng có thể linh hoạt lọc dữ liệu theo các tiêu chí cụ thể, thấy rõ sự thay đổi và hiểu được dữ liệu đang kể câu chuyện gì.

# Xây dựng báo cáo

## Dashboard và report

* **Cách tối ưu hóa Reports và Dashboards:**

1. **Dùng bộ lọc hợp lý**  
   Giới hạn dữ liệu hiển thị giúp tăng tốc độ tải và cải thiện trải nghiệm người xem.
2. **Giảm số lượng biểu đồ trên một trang**  
   Tập trung vào biểu đồ cần thiết để tránh rối mắt và đảm bảo hiệu năng.
3. **Tối ưu hình ảnh trực quan**  
   Đảm bảo biểu đồ rõ ràng, chất lượng tốt, không quá tải tài nguyên.
4. **Tận dụng bộ nhớ đệm (cache)**  
   Truy xuất từ cache giúp dashboard chạy mượt và ổn định hơn so với truy vấn trực tiếp.
5. **Tối ưu môi trường hoạt động**  
   Cấu hình hợp lý dữ liệu đầu vào, giảm độ trễ mạng và phân bổ tài nguyên hiệu quả.

## Xây dựng báo cáo

### Dashboard vs Report

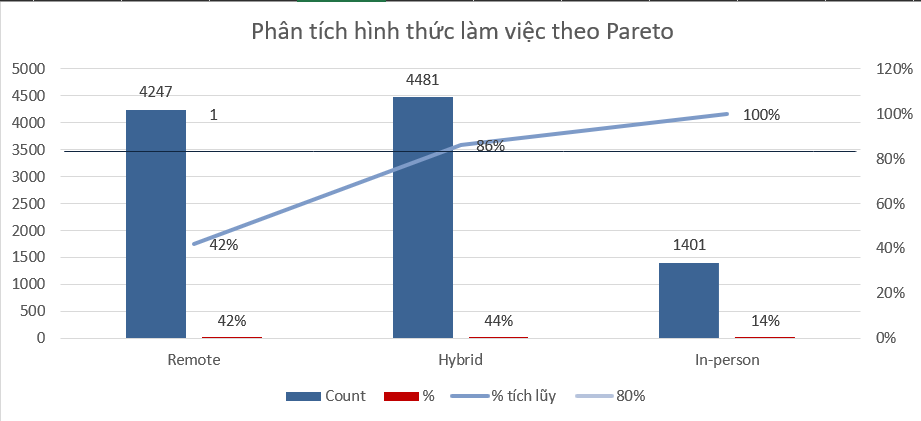
* ***So sánh Dashboard và Report***

|  |  |  |
| --- | --- | --- |
| **Tiêu chí** | **Dashboard** | **Report** |
| **Số trang** | Chỉ một trang | Nhiều trang |
| **Nguồn dữ liệu** | Từ nhiều dataset hoặc report | Từ một dataset duy nhất |
| **Trực quan hóa** | Tập trung trình bày thông tin cô đọng, sinh động | Trình bày tổng thể, hỗ trợ phân tích chuyên sâu |
| **Tạo trên Power BI Desktop** | Không thể tạo | Có thể tạo và chỉnh sửa |
| **Bộ lọc – Slicer** | Giới hạn, ít tương tác | Linh hoạt với nhiều loại filter, highlight |
| **Tương tác người dùng** | Cho phép ghim visuals từ nhiều nguồn vào một nơi | Áp dụng biến đổi và trực quan theo từng tập dữ liệu |
| **Cảnh báo & Đăng ký** | Có thể thiết lập cảnh báo và đăng ký nhận báo cáo | Không hỗ trợ cảnh báo, nhưng có thể đăng ký |
| **Dữ liệu chi tiết** | Không xem được dataset, chỉ có thể xuất | Xem và thao tác trực tiếp với dữ liệu |
| **Mục đích** | Giám sát nhanh hiệu suất, cấp quản lý | Phân tích chuyên sâu, trả lời câu hỏi kinh doanh |

* **Trong phạm vi dự án này, Dashboard được lựa chọn là công cụ trình bày chính vì các lý do sau:**
* **Tổng quan dễ hiểu:** Dashboard giúp trình bày toàn cảnh hiệu suất làm việc theo nhiều tiêu chí như hình thức làm việc, mức độ hài lòng, thu nhập,... trên một giao diện duy nhất, dễ theo dõi.
* **Tương tác linh hoạt:** nhờ các bộ lọc (slicer), animation và hiệu ứng tương tác có sẵn trong Tableau, người xem có thể tự điều chỉnh góc nhìn dữ liệu theo nhu cầu — ví dụ: lọc theo độ tuổi, vị trí công việc, số năm kinh nghiệm.
* **Ra quyết định nhanh:** các biểu đồ như Pareto, kết hợp cột – đường, card chỉ số,... cho phép người quản lý nhanh chóng nắm bắt insight và đưa ra quyết định chiến lược về mô hình làm việc hiệu quả.
* **Thân thiện với người dùng cuối:** không yêu cầu người xem phải hiểu sâu về kỹ thuật hay mô hình dữ liệu — chỉ cần quan sát trực quan để hiểu thông điệp dữ liệu.

### Dashboard

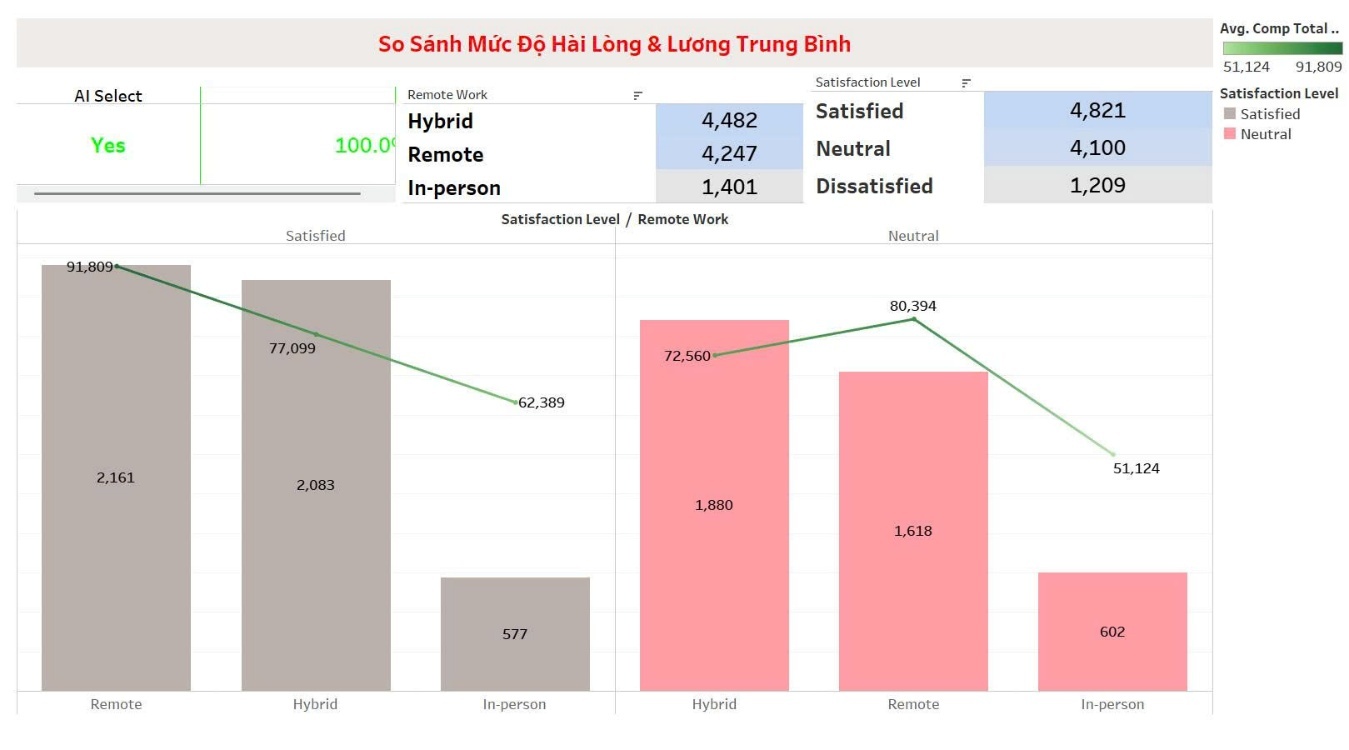
* **Biểu đồ Pareto:**



***Hình 6.1***

*Giải thích:*

* **Remote và Hybrid chiếm ưu thế vượt trội:** với tổng cộng 86% nhân sự làm việc theo hai hình thức này, xu hướng làm việc linh hoạt đang là chuẩn mực mới.
* **Hybrid slightly dẫn đầu Remote:** Hybrid có tỷ lệ 44%, cao hơn Remote (42%) → cho thấy nhu cầu kết hợp giữa làm việc từ xa và tại văn phòng đang được ưa chuộng.
* **In-person thấp nhất:** Chỉ 14% chọn hình thức này → làm việc hoàn toàn tại văn phòng không còn phổ biến.
* **Ngưỡng Pareto 80% được vượt qua bởi nhóm Remote + Hybrid:** điều này cho thấy chỉ cần tập trung cải tiến 2 nhóm này là đã tối ưu trải nghiệm cho phần lớn nhân sự.
* **Dashboard 1:**



***Hình 6.2***

*Giải thích:*

* **Mức độ hài lòng & lương cao nhất ở nhóm làm việc từ xa:**

Nhân viên làm việc **Remote** có mức hài lòng cao nhất (91,809₫), cao hơn Hybrid (77,099₫) và In-person (62,389₫).

**→** Khuyến khích mô hình làm việc từ xa bằng các chính sách hỗ trợ: phụ cấp công nghệ, văn hóa làm việc linh hoạt.

* **Nhóm Hybrid có tỷ lệ trung lập cao nhất về mức hài lòng:**

Nhiều nhân viên Hybrid có cảm nhận trung lập → chưa thực sự gắn bó.

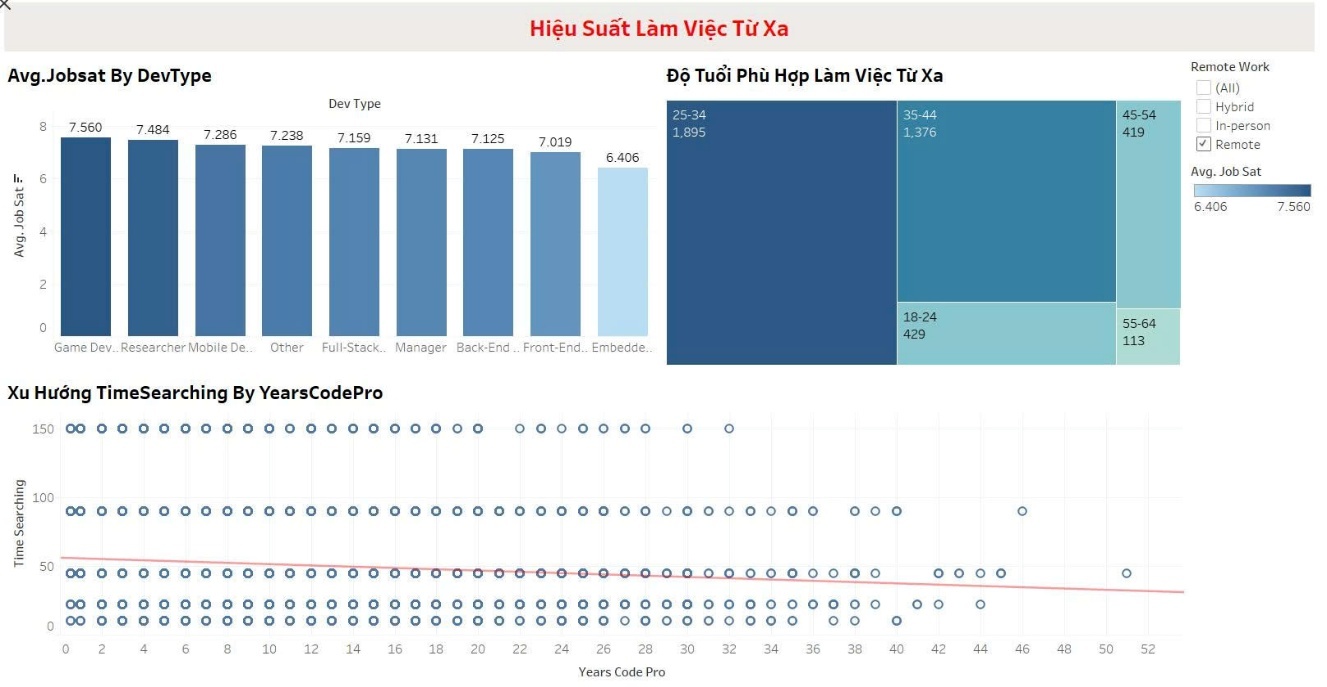
**→** Nâng cao trải nghiệm cho nhóm Hybrid qua hoạt động gắn kết định kỳ, phản hồi cá nhân hóa và môi trường làm việc linh hoạt.

* **Nhân viên In-person có mức lương & hài lòng thấp nhất:**

Mức hài lòng thấp nhất, lương thấp nhất → có nguy cơ nghỉ việc hoặc hiệu suất kém.

**→** Đánh giá lại chính sách lương & môi trường làm việc trực tiếp. Có thể thử mô hình luân chuyển công việc/hybrid để cải thiện.

* **Dashboard 2:**



***Hình 6.3***

*Giải thích:*

* **Mức độ hài lòng công việc:**

Game Dev & Researcher có mức hài lòng cao nhất.

Embedded Dev thấp nhất → cần cải thiện môi trường làm việc.

**→** Xem xét mô hình quản lý nhóm hài lòng cao để áp dụng cho nhóm còn lại.

* **Độ tuổi làm việc từ xa:**

Nhóm tuổi 25–34 chiếm tỷ lệ làm việc từ xa cao nhất.

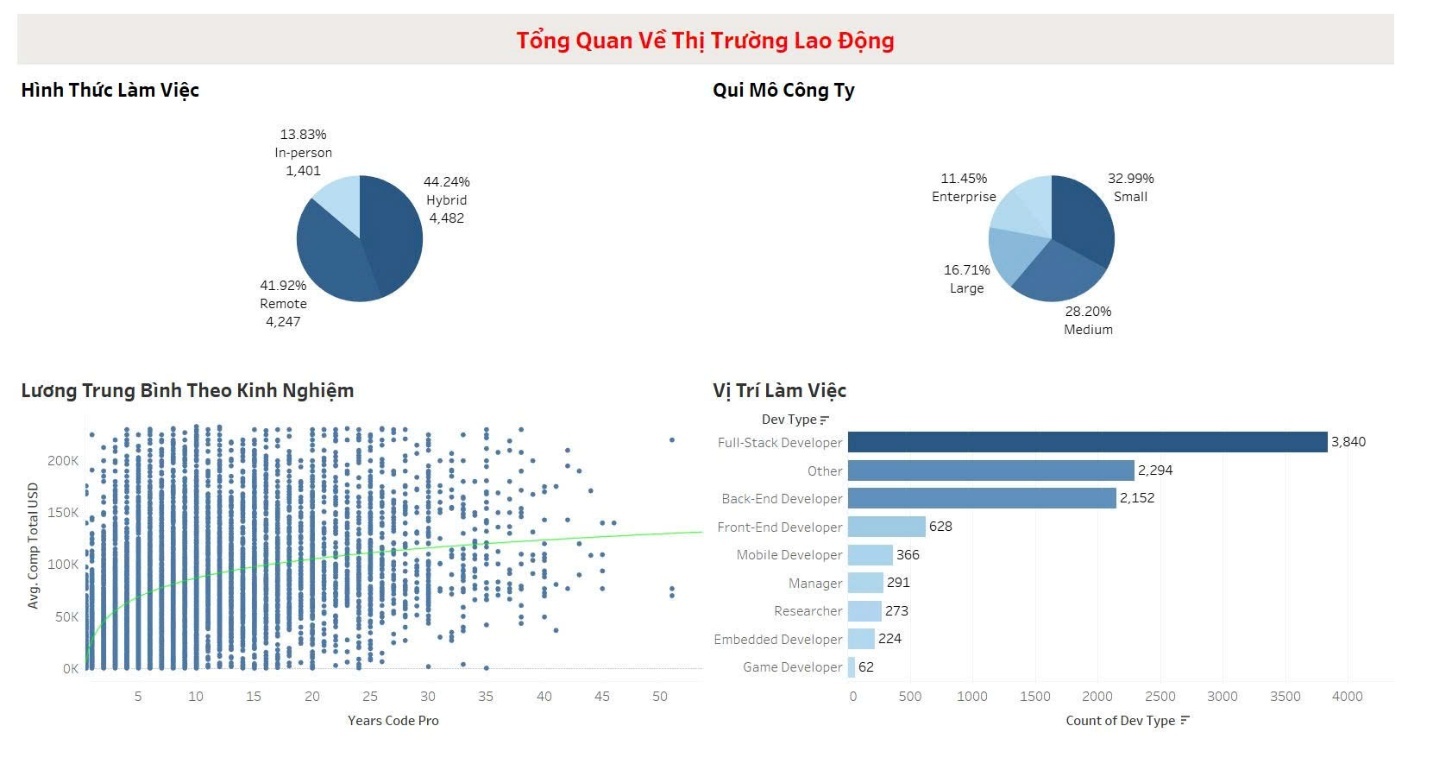
**→** Tăng hỗ trợ công cụ làm việc từ xa, xây dựng lộ trình phát triển năng lực cho nhóm này.

* **Kinh nghiệm & năng suất:**

Càng nhiều kinh nghiệm → càng ít phải tìm kiếm thông tin.

**→** Xây dựng chương trình mentorship & hệ thống chia sẻ tri thức nội bộ.

* **Dashboard 3:**



***Hình 6.4***

*Giải thích:*

* **Phân bổ hình thức làm việc:** Hybrid 44.24%, Remote 41.92%, In-person 13.83% → Hơn 86% làm việc từ xa/hybrid
* **Quy mô công ty:** chủ yếu là Small & Medium (>61%), doanh nghiệp lớn chiếm tỷ lệ nhỏ
* **Lương theo thâm niên:** lương tăng tuyến tính theo năm kinh nghiệm (0–52 năm), từ <$50K đến ~$200K
* **Tối ưu mô hình Hybrid & Remote:** đầu tư vào công cụ làm việc từ xa, tăng cường văn hóa kết nối và quản lý linh hoạt.
* **Phát triển lộ trình nghề nghiệp:** xây dựng chương trình tăng trưởng chuyên môn & lương theo thâm niên rõ ràng.
* **Hỗ trợ doanh nghiệp nhỏ & vừa:** cung cấp hạ tầng số, chính sách giữ chân nhân tài và nâng cao hiệu suất nhân sự.

# KẾT LUẬN

## Báo cáo

### Các bước viết báo cáo

1. **Xác định đối tượng người đọc**
   * Biết ai sẽ đọc (nhân viên, quản lý...), mức độ hiểu của họ và dùng ngôn ngữ phù hợp.
   * Sử dụng tiếng Việt, có thể kèm vài thuật ngữ chuyên ngành (chứng khoán, công nghệ...).
   * Tập trung truyền đạt rõ ràng, dễ hiểu.
2. **Phác thảo bố cục**
   * Giới thiệu: tóm tắt ngắn về chủ đề phân tích.
   * Nội dung chính: Dữ liệu – Phương pháp – Phân tích – Kết quả.
   * Kết luận: tóm tắt phát hiện quan trọng, đề xuất hoặc định hướng.
3. **Phát triển nội dung**
   * Dữ liệu: nguồn, cách thu thập, cấu trúc và tổ chức dữ liệu (CSV, Excel...).
   * Phương pháp: mô hình hóa, xử lý, thêm biến tính toán, công thức nghiệp vụ.
   * Phân tích: tách nhóm dữ liệu theo tiêu chí (sàn, ngành, mã...), mô tả logic rõ ràng.
   * Kết quả: trình bày phát hiện bằng biểu đồ và bảng số liệu.
4. **Viết phần mở đầu sau cùng**
   * Giới thiệu tổng quan vấn đề, bối cảnh, câu hỏi nghiên cứu và hướng triển khai.
   * Trình bày ngắn gọn nhưng đủ dẫn dắt người đọc hiểu mục tiêu của báo cáo.
5. **Kết luận ngắn gọn**
   * Tóm tắt phát hiện chính bằng ngôn ngữ dễ hiểu.
   * Nêu ý nghĩa thực tiễn (giá trị với người đọc/doanh nghiệp).
6. **Trực quan hóa dữ liệu**
   * Dùng biểu đồ phù hợp: Line chart, table, treemap, candlestick...
   * Card thông tin: tổng quan nhanh (tổng giao dịch, khối lượng...).
   * Bộ lọc (Slicer): cho phép người dùng lọc theo thời gian, ngành, mã CK.
7. **Đọc lại & chỉnh sửa**
   * Đọc như người lần đầu xem báo cáo.
   * Tinh gọn nội dung, sửa lỗi chính tả/ngữ pháp, đảm bảo mạch logic rõ ràng.
   * Cân nhắc bố cục, từ ngữ và tính thuyết phục.

### Tổng hợp

Dự án được triển khai nhằm phân tích hiệu suất làm việc từ xa trong ngành công nghệ, xoay quanh câu hỏi trung tâm:

**“Liệu làm việc từ xa có thực sự hiệu quả trong ngành công nghệ, và ai là người phù hợp nhất với mô hình này?”**

Hai nhóm đối tượng chính cần lời giải từ dữ liệu bao gồm:

* **Doanh nghiệp công nghệ**: mong muốn xác định mô hình làm việc tối ưu – giữa làm từ xa, tại văn phòng hoặc kết hợp – sao cho vừa đảm bảo hiệu suất cá nhân, vừa duy trì sự gắn kết nội bộ và văn hóa công ty.
* **Sinh viên mới tốt nghiệp**: cần định hướng lựa chọn hình thức làm việc phù hợp với năng lực bản thân, xu hướng ngành và cơ hội phát triển sự nghiệp trong bối cảnh hiện đại.

Để trả lời các câu hỏi trên, nhóm đã thực hiện toàn bộ quy trình phân tích dữ liệu: từ làm sạch, chuẩn hóa và chuyển đổi dữ liệu; mô hình hóa theo cấu trúc sao (star schema); trực quan hóa bằng nhiều kỹ thuật biểu đồ kết hợp với dashboard tương tác.

Một số insight nổi bật được khai thác bao gồm:

* Nhóm độ tuổi và vị trí công việc nào hài lòng nhất với mô hình làm việc từ xa.
* Mối quan hệ giữa kinh nghiệm, thu nhập và hình thức làm việc.
* Tỷ lệ hài lòng của nhân sự theo từng mô hình làm việc.

**Kết quả cuối cùng** là một báo cáo phân tích dễ hiểu, trực quan, được trình bày thông qua dashboard Tableau – giúp nhà quản lý nhân sự có cơ sở đưa ra quyết định tuyển dụng và tổ chức công việc, đồng thời cung cấp cái nhìn hữu ích cho sinh viên công nghệ trong việc lựa chọn hình thức làm việc phù hợp với bản thân.

## 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, mạch lạc.

## 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.

## Hướng phát triển

Dự án có thể tiếp tục mở rộng theo các định hướng sau:

* **Mở rộng phân tích chuyên sâu**: áp dụng thêm mô hình học máy để dự đoán hiệu suất làm việc từ các yếu tố cá nhân và môi trường.
* **Phát triển câu chuyện dữ liệu**: xây dựng chuỗi data storytelling mạch lạc, nhấn mạnh insight quan trọng qua từng biểu đồ, kết hợp narrative (lời dẫn) với trực quan để tăng tính thuyết phục và dễ tiếp cận.
* **Tăng cường kỹ thuật trực quan hóa**: tận dụng sâu hơn các tính năng nâng cao của Tableau như animation, parameter control, dynamic dashboard, drill-down,... để tăng tương tác và làm nổi bật xu hướng.
* **Tích hợp thêm nguồn dữ liệu**: mở rộng từ khảo sát Stack Overflow sang các nền tảng tuyển dụng, báo cáo nhân sự hoặc dữ liệu nội bộ doanh nghiệp.
* **Xây dựng dashboard động**: thiết kế hệ thống có thể cập nhật dữ liệu theo thời gian thực, phục vụ phân tích liên tục.
* **Tự động hóa quy trình xử lý**: chuẩn hóa pipeline Python từ tiền xử lý đến xuất dữ liệu, giúp tái sử dụng, tiết kiệm thời gian và dễ bảo trì.

**“Phân tích hiệu suất làm việc từ xa”** là dự án tâm huyết giúp nhóm rèn kỹ năng phân tích thực tiễn, đồng thời mở ra tiềm năng phát triển thành công cụ hỗ trợ ra quyết định trong thời đại số.

# THAM KHẢO

* **Stack Overflow Developer Survey 2024**  
  <https://survey.stackoverflow.co/2024/>
* **Danh mục các loại biểu đồ trong Data Visualization – 200Lab**  
  <https://200lab.io/blog/danh-muc-cac-loai-bieu-do-trong-data-visualization>
* **Làm sạch dữ liệu là gì? – MindX**  
  <https://mindx.edu.vn/blog/lam-sach-du-lieu>
* **6 Trụ cột của chất lượng dữ liệu – Mastering DA**  
  <https://mastering-da.com/6-tru-cot-cua-chat-luong-du-lieu>
* **Slide bài giảng môn Xử lý dữ liệu – Bộ môn CNTT, FPT Polytechnic** (Tài liệu nội bộ)