**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

**GV HD: Th. Văn Công Khanh**

TP.HCM 08 – 2025

MỤC LỤC

[1 Giới thiệu dự án 4](#_Toc165991907)

[1.1 Giới thiệu 4](#_Toc165991908)

[1.2 Yêu cầu của công ty 4](#_Toc165991909)

[1.3 Lập kế hoạch dự án 4](#_Toc165991910)

[2 Phân tích yêu cầu khách hàng 5](#_Toc165991911)

[2.1 Phân tích yêu cầu 5](#_Toc165991912)

[2.2 Câu chuyện dữ liệu 5](#_Toc165991913)

[2.2.1 Đặt vấn đề 5](#_Toc165991914)

[2.2.2 Xác định câu chuyện 6](#_Toc165991915)

[2.2.3 Xác định rõ đối tượng 6](#_Toc165991916)

[2.2.4 Xác định câu chuyện chi tiết 6](#_Toc165991917)

[2.2.5 Trình bày dữ liệu 6](#_Toc165991918)

[2.2.6 Những điều cần lưu ý 6](#_Toc165991919)

[2.3 Kiến trúc hệ thống 7](#_Toc165991920)

[2.3.1 Kiến trúc 7](#_Toc165991921)

[2.3.2 Giải thích 7](#_Toc165991922)

[2.4 Giải thích về bộ dữ liệu khách hàng 7](#_Toc165991923)

[2.4.1 Các khái niệm 7](#_Toc165991924)

[2.4.2 Các trường dữ liệu 7](#_Toc165991925)

[3 Làm sạch và chuyển đổi dữ liệu 8](#_Toc165991926)

[3.1 Chuẩn bị dữ liệu 8](#_Toc165991927)

[3.1.1 Giải pháp lưu trữ dữ liệu 8](#_Toc165991928)

[3.1.2 Giải pháp phân bố dữ liệu 8](#_Toc165991929)

[3.2 Làm sạch dữ liệu 8](#_Toc165991930)

[3.2.1 Các vấn đề ảnh hưởng tới dữ liệu 8](#_Toc165991931)

[3.2.2 Các tiêu chí đánh giá chất lượng dữ liệu 9](#_Toc165991932)

[3.2.3 Các bước làm sạch dữ liệu 9](#_Toc165991933)

[3.3 Chuyển đổi dữ liệu 9](#_Toc165991934)

[3.3.1 Các trường hợp cần chuyển đổi 9](#_Toc165991935)

[3.3.2 Các kỹ thuật chuyển đổi 9](#_Toc165991936)

[3.3.3 Trình bày các phép chuyển đổi trong dự án 9](#_Toc165991937)

[4 Xử lý dữ liệu 10](#_Toc165991938)

[4.1 Chuẩn hóa dữ liệu 10](#_Toc165991939)

[4.1.1 Trình bày các bước chuẩn hóa trong dự án 10](#_Toc165991940)

[4.2 Mô hình hóa dữ liệu 10](#_Toc165991941)

[4.2.1 Các loại mô hình hóa 10](#_Toc165991942)

[4.2.2 Các tiêu chí đánh giá mô hình dữ liệu 10](#_Toc165991943)

[4.2.3 Trình bày các bước mô hình hóa 10](#_Toc165991944)

[4.2.4 Trình bày các bước tạo bảng dữ liệu 11](#_Toc165991945)

[5 Trực quan hóa dữ liệu 11](#_Toc165991946)

[5.1 Các kỹ thuật trực quan hóa 11](#_Toc165991947)

[5.2 Các nguyên tắc trực quan hóa 11](#_Toc165991948)

[5.3 Trình bày tạo các report cho dự án 11](#_Toc165991949)

[5.3.1 Tạo visual thống kê chi tiết 11](#_Toc165991950)

[5.3.2 Tạo visual thống kê tổng thể 12](#_Toc165991951)

[6 Xây dựng báo cáo 12](#_Toc165991952)

[6.1 Dashboard và report 12](#_Toc165991953)

[6.2 Xây dựng báo cáo 12](#_Toc165991954)

[6.2.1 Dashboard vs Report 12](#_Toc165991955)

[6.2.2 Dashboard 13](#_Toc165991956)

[6.2.3 Report 13](#_Toc165991957)

[6.2.4 Bookmark 13](#_Toc165991958)

[7 KẾT LUẬN 14](#_Toc165991959)

[7.1 Báo cáo 14](#_Toc165991960)

[7.1.1 Các bước viết báo cáo 14](#_Toc165991961)

[7.1.2 Tổng hợp 14](#_Toc165991962)

[7.2 Khó khăn 14](#_Toc165991963)

[7.3 Thuận lợi 14](#_Toc165991964)

[7.4 Hướng phát triển 14](#_Toc165991965)

# Giới thiệu dự án

## Giới thiệu

* **Hiện trạng**

Trong thời đại công nghệ phát triển ngày nay, hình thức làm việc từ xa (remote work) đã trở thành một xu hướng phát triển mạnh mẽ trên toàn cầu. Nhiều doanh nghiệp – đặc biệt trong lĩnh vực công nghệ thông tin – đã chuyển sang mô hình làm việc từ xa hoàn toàn hoặc kết hợp (hybrid). Sự chuyển đổi này không chỉ thay đổi cách nhân viên làm việc, giao tiếp và quản lý thời gian mà còn ảnh hưởng đến cách đo lường hiệu suất và sự hài long với công việc.

Bên cạnh những lợi ích rõ rệt như linh hoạt về thời gian, tiết kiệm chi phí vận hành và mở rộng nguồn nhân lực toàn cầu, làm việc từ xa cũng đặt ra không ít thách thức như cảm giác cô lập, giảm tương tác giữa các thành viên trong nhóm, khó khăn trong việc theo dõi tiến độ công việc và lo ngại về sức khỏe tinh thần ngày càng được quan tâm.

Điều này dẫn đến một câu hỏi lớn đối với các nhà quản lý và doanh nghiệp: *“Liệu làm việc từ xa có thực sự mang lại hiệu quả như kỳ vọng, hay chỉ phù hợp với một số nhóm nhân viên cụ thể?”*

Để tìm câu trả lời một cách khách quan, cần phân tích dữ liệu thực tế từ một khảo sát quy mô lớn, nhằm xác định mối liên hệ giữa hình thức làm việc và các yếu tố như: số giờ làm việc mỗi tuần, mức độ hài lòng với công việc, mức thu nhập, kỹ năng chuyên môn và vị trí địa lý.

Từ đó, dự án đặt ra mục tiêu tìm hiểu:

*“Liệu làm việc từ xa có thực sự mang lại hiệu quả như kỳ vọng trong thời đại công nghệ số hiện nay?”*

* **Thông tin bộ dữ liệu**

Dự án sử dụng bộ dữ liệu Stack Overflow Developer Survey 2023, là khảo sát thường niên lớn nhất thế giới dành cho các lập trình viên và chuyên gia công nghệ. Khảo sát này thu thập thông tin từ hơn 70.000 người tham gia trên khắp thế giới, bao gồm nhiều khía cạnh như:

* Hình thức làm việc: tại văn phòng, hybrid, 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)
* Trình độ học vấn, ngôn ngữ lập trình, công nghệ sử dụng, và các kỹ năng kỹ thuật khác.

Dữ liệu được cung cấp ở định dạng Excel (.xlsx) bao gồm:

* File survey\_results\_public.xlsx: dữ liệu khảo sát gốc
* File survey\_results\_schema.xlsx: bảng chú thích các trường, giúp hiểu rõ ý nghĩa và cách xử lý từng cột.

## 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 hang năm, hình thức làm việc, trình độ chuyên môn và mức độ hài lòng.

* **Về quản lý và lưu trữ:**

Dữ liệu cần được xử lý, lưu trữ và trực quan hóa bằng các công cụ phù hợp với ngành Xử lý Dữ liệu. Công ty yêu cầu sử dụng Tableau để thực hiện các báo cáo phân tích dưới dạng dashboard có tương tác kết hợp Python để làm sạch dữ liệu. Toàn bộ dữ liệu và dashboard phải được lưu trữ một cách có hệ thống để dễ dàng truy xuất và báo cáo.

* **Mục tiêu:**

**Mục tiêu cuối cùng của đề tài** là cung cấp một cái nhìn toàn diện và dựa trên dữ liệu thực tế để **đánh giá mức độ hiệu quả của mô hình làm việc từ xa trong ngành công nghệ**. Dự án khai thác bộ dữ liệu từ Stack Overflow Developer Survey 2023 – một trong những khảo sát uy tín nhất trong cộng đồng lập trình viên toàn cầu – để phân tích sâu các yếu tố liên quan đến hiệu suất làm việc từ xa.

**Cụ thể, dự án hướng tới:**

* **Hiểu rõ sự khác biệt giữa các hình thức làm việc** (Remote – Hybrid – Onsite) dựa trên các khía cạnh: thu nhập, giờ làm việc, mức độ hài lòng, vai trò công việc,...
* **Tìm ra những xu hướng phổ biến trong giới công nghệ**, đặc biệt là cách lập trình viên lựa chọn môi trường làm việc sau đại dịch.
* **Hỗ trợ nhà quản lý và các tổ chức** ra quyết định chính sách phù hợp: có nên cho nhân viên làm từ xa? Nhóm ngành/vai trò nào phù hợp với mô hình hybrid?
* **Tăng khả năng thích ứng cá nhân**: giúp người lao động hiểu rõ lợi – hại của từng hình thức làm việc để chọn lựa môi trường phù hợp, từ đó **tối ưu hóa năng suất và chất lượng công việc**
* **Đánh giá tính khả thi:**
* **Năng lực hiện có của nhóm:**

Thành viên nhóm đã có kiến thức nền tảng về xử lý dữ liệu, hiểu cấu trúc dữ liệu dạng bảng và kỹ thuật làm sạch dữ liệu cơ bản.

Biết sử dụng Tableau, Python và có khả năng xây dựng các biểu đồ trực quan hóa dữ liệu.

Có kinh nghiệm làm việc nhóm và phân tích yêu cầu dựa trên dữ liệu thực tế.

* **Kỹ năng học thêm trong quá trình làm dự án:**

Làm quen với bộ dữ liệu lớn thực tế quy mô lớn, xử lý missing values, chuẩn hóa định dạng

Phát triển kỹ năng xây dựng dashboard có tính tương tác, phân tầng theo các đối tượng (vị trí, quốc gia, hình thức làm việc).

Kỹ năng kể chuyện bằng dữ liệu (data storytelling) để truyền tải kết quả phân tích một cách trực quan, dễ hiểu.

## Lập kế hoạch dự án

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TT** | **Hạng mục** | **Bắt đầu** | **Kết thúc** | **Kết quả** |
| 1 | Giới thiệu dự án | 25/05/2025 | 31/05/2025 | Xác định đề tài, hiện trạng nguồn dữ liệu |
| 1.1 | Giới thiệu | 25/05/2025 | 31/05/2025 | Trình bày ý tưởng của nhóm, dữ liệu đã khảo sát như thế nào? |
| 1.2 | Yêu cầu | 25/05/2025 | 31/05/2025 | Làm rõ yêu cầu của ý tưởng, dữ liệu và công nghệ sử dụng |
| 2 | Phân tích yêu cầu | 25/05/2025 | 01/06/2025 | Hiểu rõ vấn đề, xây dựng định hướng cho dự án |
| 2.1 | Phân tích yêu cầu KH | 16/05/2025 | 02/06/2025 | Xác định được mục tiêu trọng tâm và phân tích |
| 2.2 | Câu chuyện dữ liệu | 16/05/2025 | 31/05/2025 | Xây dựng câu chuyện giả định, giả thuyết phân tích, đối tượng tiếp nhận báo cáo |
| 3 | Làm sạch và chuyển đổi dữ liệu | 06/06/2025 | 16/06/2025 | Dữ liệu NaN nhiều, văn bản không đồng nhất, khó xác định giá trị chuẩn và đúng. |
| 3.1 | 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 |
| 3.2 | Kiểm thử và kiểm tra chất lượng dữ liệu | 16/06/2025 | 20/06/2025 |  |
| 4 | Xử lý dữ liệu |  |  |  |
| 4.1 | Chuẩn hóa dữ liệu |  |  |  |
|  | … |  |  |  |
| 5 | Trực quan hóa dữ liệu |  |  |  |
| 5.1 | Các kĩ thuật trực quan hóa |  |  |  |
|  | … |  |  |  |
| 6 | Xây dựng báo cáo |  |  |  |
| 6.1 | Dashboard và Report |  |  |  |
|  | Viết báo cáo | 25/05/2025 | 20/07/2025 |  |

# Phân tích yêu cầu khách hàng

## Phân tích yêu cầu

* **Dữ liệu:**

Bộ dữ liệu gốc được sử dụng là “*survey\_results\_public.xlsx*”, trích xuất từ Stack Overflow Developer Survey 2023 – khảo sát với hơn 70.000 phản hồi từ các lập trình viên, chuyên gia CNTT trên toàn thế giới.

Dữ liệu phản ánh chân thực tình trạng làm việc trong ngành công nghệ, bao gồm các trường:

* RemoteWork (Hình thức làm việc)
* Employment (Tình trạng công việc)
* DevType (Loại công việc)
* WorkWeekHrs (Số giờ làm việc mỗi tuần)
* JobSat (Mức độ hài lòng với công việc)
* CompTotal (Tổng thu nhập hằng năm)
* Country, YearsCodePro, EdLevel, v.v...

=> Đây là nguồn dữ liệu phong phú, phù hợp để thực hiện phân tích chuyêm sâu về tác động của làm việc từ xa đến hiệu suất và sự hài lòng trong công việc.

* **Quản lý và lưu trữ:**

Excel được sử dụng để đọc, kiểm tra sơ bộ dữ liệu thô.

Python (pandas, numpy,…) dùng để:

* Làm sạch dữ liệu (lọc missing values, xử lý outliers)
* Chuyển đổi định dạng dữ liệu
* Tổng hợp các chỉ số cần thiết (nhóm theo quốc gia, loại hình làm việc,...)

Sau khi xử lý, dữ liệu được lưu dưới định dạng .csv hoặc .xlsx, sẵn sàng import vào Tableau để trực quan hóa.

* **Công nghệ:**

**Microsoft Excel:**

* Loại bỏ các cột không cần thiết
* Kiểm tra sơ bộ dữ liệu
* Đánh giá nhanh các cột thiếu, trùng lặp hay lỗi định dạng

**Python (Jupyter Notebook):**

* Thư viện sử dụng: pandas, numpy, matplotlib, seaborn, plotly.
* Làm sạch dữ liệu, phân tích thống kê mô tả

**Tableau Public / Tableau Desktop:**

* Tạo dashboard tương tác để phân tích dữ liệu theo chiều sâu
* Sử dụng filter, parameter, map, bar chart, line chart, heatmap,... để minh họa câu chuyện dữ liệu.
* Trực quan hóa dữ liệu để dễ dàng khai thác và so sánh dữ liệu theo quốc gia, loại hình công việc, mức lương, thời gian làm việc,…

**Canva:**

* Thiết kế các slide trình bày kết quả trực quan
* Minh họa dữ liệu bằng hình ảnh, biểu tượng, sơ đồ giúp dễ hiểu và thu hút người xem

**Microsoft Word:**

* Dùng để viết báo cáo chi tiết: giới thiệu, phân tích, bảng biểu, biểu đồ và kết luận. Đảm bảo định dạng chuẩn, dễ đọc và đây là tài liệu chính thức được nộp cuối cùng.
* **Quyết định dùng công nghệ**

Từ yêu cầu của doanh nghiệp và mục tiêu phân tích, nhóm đã xác định rõ ràng là:

* Phân tích định lượng về hiệu quả làm việc từ xa dựa trên nhiều yếu tố như: giờ làm việc, mức độ hài lòng, thu nhập, loại công việc,…
* Khả năng trực quan hóa kết quả nhằm hỗ trợ các nhà quản lý đưa ra quyết định dễ dàng và nhanh chóng.
* Xử lý linh hoạt một bộ dữ liệu lớn, đa dạng trường thông tin, có tính phân tầng theo quốc gia, vị trí, trình độ,…
* Dữ liệu đầu vào phải có độ bao phủ toàn cầu, khách quan và đáng tin cậy.
* Kết quả cần được trình bày sinh động, rõ rang, dễ hiểu và có tính thuyết phục cao đối với người chuyên môn và không chuyên môn.
* **Từ những yêu cầu trên, nhóm quyết định lựa chọn công nghệ như sau:**

**Dữ liệu:**

Bộ dữ liệu từ *Stack Overflow Developer Survey 2023* được chọn vì mức độ uy tín, quy mô toàn cầu và nội dung sát với thực tế của ngành CNTT. Đây là nền tảng vững chắc cho phân tích xu hướng làm việc từ xa trong môi trường công nghệ hiện nay.

**Quản lý và xử lý dữ liệu:**

Kết hợp Excel và Python giúp linh hoạt từ thao tác đơn giản đến phân tích chuyên sâu. Python đảm bảo hiệu suất xử lý và khả năng tùy biến cao trong thao tác dữ liệu lớn.

**Công nghệ trực quan hóa:**

Tableau là công cụ được lựa chọn để xây dựng hệ thống dashboard trực quan, sinh động:

* Hỗ trợ tạo các biểu đồ tương tác
* Cho phép người dùng lọc theo nhiều tiêu chí như quốc gia, vai trò, kinh nghiệm,… giúp nhà quản lý theo dõi xu hướng và đưa ra quyết định chính xác.

**Trình bày báo cáo:**

Canva: được sử dụng để thiết kế các slide thuyết trình với hình ảnh minh họa trực quan, màu sắc sinh động và bố cục logic, giúp thu hút người xem và truyền tải thông tin dễ hiểu hơn trong các buổi thuyết trình.

Microsoft Word: dùng để biên soạn báo cáo học thuật chi tiết, bao gồm: phần mở đầu, phân tích dữ liệu, bảng biểu, biểu đồ, nhận xét và kết luận. Word đảm bảo định dạng chuẩn, dễ theo dõi và phù hợp để nộp cho giảng viên hay đối tác.

## Câu chuyện dữ liệu

### Đặt vấn đề

* **Mô tả thực trạng:**

Hiện nay, làm việc từ xa (remote work) không còn là xu hướng nhất thời mà đã trở thành một lựa chọn phổ biến trong ngành công nghệ thông tin. Nhiều doanh nghiệp áp dụng hình thức này để giảm chi phí vận hành, mở rộng nguồn nhân lực và tăng tính linh hoạt cho nhân viên.

Tuy nhiên, bên cạnh những lợi ích dễ thấy như tiết kiệm thời gian di chuyển, chủ động sắp xếp công việc, thì hình thức này cũng đặt ra hàng loạt thách thức:

* Giảm tương tác trực tiếp, dễ gây cảm giác cô lập
* Khó kiểm soát tiến độ và đánh giá hiệu suất
* Mất cân bằng giữa công việc và cuộc sống cá nhân

Điều này khiến các nhà quản lý đặt ra một câu hỏi quan trọng:

*Làm việc từ xa có thực sự hiệu quả như kỳ vọng không? Hay chỉ phù hợp với một số nhóm người, lĩnh vực hoặc môi trường nhất định?*

* **Dữ liệu liên quan:**

Dữ liệu sau khi thu thập và xử lý từ Stack Overflow Developer Survey 2023 trở thành một tập hợp khổng lồ các con số – hàng chục nghìn dòng thể hiện hình thức làm việc, mức độ hài lòng, thời gian làm việc, thu nhập, kỹ năng,...

Tuy nhiên, những con số này chỉ thực sự có giá trị khi được:

Diễn giải đúng cách: chỉ ra mối quan hệ giữa các yếu tố (ví dụ: làm việc từ xa có thực sự giúp tăng mức độ hài lòng không?)

Trình bày phù hợp với người đọc: là nhà quản lý, họ cần những kết luận dễ hiểu, trực quan, đi kèm biểu đồ, hình ảnh thay vì bảng số liệu khô khan.

Vì vậy, câu chuyện dữ liệu cần phải được kể một cách mạch lạc, dễ tiếp cận và có tính ứng dụng cao – đó là lý do nhóm không chỉ xử lý dữ liệu bằng Python mà còn sử dụng các công cụ trực quan hóa như Tableau, Canva và Word để chuyển hóa dữ liệu thành thông tin có ý nghĩa cho người dùng cuối.

* **Mục tiêu:**

Truyền đạt thông điệp dữ liệu một cách rõ ràng và hấp dẫn:

Giúp người đọc hoặc người nghe nhanh chóng nắm bắt được vấn đề cốt lõi từ những con số khô khan trong dữ liệu, mà không cần phải có chuyên môn sâu về phân tích dữ liệu.

Diễn giải kết quả một cách sinh động, dễ hiểu và thuyết phục:

Thay vì chỉ trình bày số liệu thuần túy, báo cáo sử dụng hình ảnh trực quan, biểu đồ, dashboard để hỗ trợ giải thích, tạo nên một câu chuyện logic và thuyết phục.

Hỗ trợ ra quyết định thực tiễn:

Giúp nhà quản lý hoặc người ra quyết định nhận biết được xu hướng, mối quan hệ giữa hình thức làm việc và hiệu suất/cảm nhận công việc, từ đó đề xuất chính sách phù hợp với từng tổ chức, từng nhóm nhân sự.

### Xác định câu chuyện

Từ bộ dữ liệu thu thập được, nhóm hướng tới việc khai thác những câu chuyện mang tính thực tiễn cao và có ảnh hưởng đến quyết định của doanh nghiệp trong bối cảnh hành trình chuyển đổi số. Các đặt vấn chính gồm:

* Làm việc từ xa ảnh hưởng đến hiệu suất làm việc như thế nào?
* Có mối liên hệ nào giữa hình thức làm việc (Remote/Hybrid/On-site) và mức độ hài lòng với công việc?
* Ai là nhóm đối tượng phù hợp nhất với làm việc từ xa? (theo độ tuổi, kinh nghiệm, DevType, quốc gia)

Hiệu suất làm việc thể hiện qua số giờ làm việc mỗi tuần (WorkWeekHrs):

→ Người làm việc từ xa có thực sự làm việc ít hơn hay nhiều hơn so với người làm việc tại văn phòng?

Mức độ hài lòng với công việc (JobSat):

→ Làm việc từ xa có mang lại sự thoải mái, chủ động hơn, dẫn đến hài lòng cao hơn? Hay lại gây ra cảm giác cô lập, giảm gắn kết và hài lòng?

Thu nhập hằng năm (CompTotal):

→ Có sự chênh lệch rõ rệt về thu nhập giữa các hình thức làm việc không? Nếu có, thì ở nhóm nghề nghiệp hoặc quốc gia nào?

Sự khác biệt theo đặc điểm cá nhân và nghề nghiệp:

→ Tác động của hình thức làm việc có giống nhau với mọi người không? Hay phụ thuộc vào các yếu tố như:

Quốc gia (Country)

Kinh nghiệm lập trình (YearsCodePro)

Trình độ học vấn (EdLevel)

Loại công việc đảm nhiệm (DevType)

* **Mục tiêu cụ thể của nhóm:**

**Làm rõ mối quan hệ giữa hình thức làm việc và hiệu suất công việc:**

Phân tích dữ liệu thực tế để kiểm chứng liệu làm việc từ xa (remote), kết hợp (hybrid), hay tại văn phòng (on-site) mang lại hiệu suất và sự hài lòng cao hơn cho nhân viên trong ngành công nghệ.

**Xác định nhóm đối tượng phù hợp nhất với từng hình thức làm việc:**

Dựa vào các biến như vị trí công việc (DevType), kinh nghiệm (YearsCodePro), quốc gia (Country), mức lương (CompTotal), mức độ hài lòng (JobSat), từ đó đưa ra những khuyến nghị linh hoạt theo từng nhóm đối tượng cụ thể thay vì áp dụng một mô hình chung cho toàn bộ tổ chức.

**Biến dữ liệu thành thông tin có ý nghĩa và giá trị thực tiễn:**

Không chỉ trình bày số liệu thuần túy, mà còn kể một câu chuyện có logic và chiều sâu, giúp người đọc (hoặc người nghe) hiểu nhanh – nhớ lâu – áp dụng được.

**Thúc đẩy tư duy ra quyết định dựa trên dữ liệu (data-driven thinking):**

Thông qua cách tiếp cận phân tích và trực quan hóa, mục tiêu là giúp các nhà quản lý, nhóm nhân sự, hoặc người học có thể:

* Ra quyết định tuyển dụng, điều phối nhân sự hợp lý hơn
* Xây dựng môi trường làm việc hiệu quả và phù hợp với từng bối cản
* Cân nhắc giữa hiệu suất và sự gắn kết trong tổ chức

**Gây hứng thú và dễ tiếp cận qua hình ảnh trực quan:**

Sử dụng công cụ như Tableau, Canva và biểu đồ Python để kể câu chuyện một cách sinh động, dễ hiểu, đặc biệt với người không chuyên phân tích dữ liệu. Giúp người đọc không bị “ngợp” vì số liệu, mà thay vào đó, hiểu rõ ý nghĩa và kết luận đằng sau từng biểu đồ.

* **Các giải pháp thực tiễn từ phân tích:**

1. **Phát triển bảng điều khiển dữ liệu (Data Dashboard):**

Thiết kế **dashboard trực quan** thể hiện các chỉ số chính như: hình thức làm việc phổ biến theo quốc gia, thời gian làm việc, mức lương trung bình, mức độ hài lòng,...

Cho phép **lọc và phân tích dữ liệu** theo các tiêu chí như loại công việc, quốc gia, kinh nghiệm, hình thức làm việc,...

Dashboard được chia sẻ nội bộ, giúp nhân sự, quản lý và lãnh đạo có cái nhìn **tổng quan và cập nhật liên tục** để ra quyết định phù hợp.

1. **Cung cấp báo cáo dữ liệu định kỳ:**

Thiết lập báo cáo hàng tháng/quý về các xu hướng làm việc từ xa, sự hài lòng và hiệu suất theo từng nhóm ngành/lãnh thổ.

Trình bày ngắn gọn, dễ hiểu, đi kèm biểu đồ và đề xuất thực tế, phục vụ **quyết định quản trị nhân sự, chính sách làm việc, tuyển dụng.**

Kết hợp báo cáo qua **Microsoft Word** và **Canva** để đảm bảo đồng thời yếu tố chuyên môn (viết học thuật) và trình bày hấp dẫn (thiết kế trực quan).

1. **Xây dựng mô hình dự đoán sự hài lòng hoặc hiệu suất:**

Sử dụng mô hình học máy đơn giản (linear regression, decision tree) để thử dự đoán **JobSat hoặc WorkWeekHrs** dựa trên các biến đầu vào như hình thức làm việc, loại công việc, kinh nghiệm,...

Mục tiêu: Giúp quản lý hiểu được **điều gì ảnh hưởng nhiều nhất đến sự hài lòng và hiệu suất**, từ đó chủ động cải thiện.

1. **Phát triển công cụ phân nhóm nhân sự phù hợp với từng mô hình làm việc:**

Dựa vào dữ liệu, tạo bộ tiêu chí gợi ý: ai phù hợp làm việc từ xa, ai nên hybrid hoặc onsite.

Kết hợp kỹ năng, loại công việc, độ gắn kết và năng suất lịch sử để đưa ra đề xuất cá nhân hóa.

1. **Thiết lập chương trình tư vấn nội bộ và đào tạo:**

Tổ chức buổi chia sẻ và đào tạo để:

* Giúp nhân viên hiểu rõ lợi - hại của từng mô hình làm việc.
* Nâng cao kỹ năng làm việc từ xa: quản lý thời gian, giao tiếp online, tránh kiệt sức.

Đào tạo quản lý kỹ năng **lãnh đạo nhóm phân tán**, sử dụng dữ liệu để giám sát hiệu quả nhưng vẫn tạo gắn kết.

1. **Chia sẻ phân tích dữ liệu với cộng đồng hoặc tổ chức đối tác:**

Đưa các insight phân tích lên website công ty, blog, hội thảo chuyên môn (dưới dạng infographic, dashboard hoặc video).

Góp phần xây dựng hình ảnh doanh nghiệp hiện đại, **ứng dụng dữ liệu vào ra quyết định nhân sự hiệu quả,** từ đó thu hút thêm nhân tài hoặc nhà đầu tư.

* **Kỳ vọng từ việc triển khai giải pháp của nhóm:**

Tăng cường sự hiểu biết của doanh nghiệp về xu hướng làm việc và nhu cầu của nhân viên.

Xây dựng chính sách linh hoạt hơn cho từng nhóm nhân sự.

Gia tăng hiệu suất và mức độ hài lòng, đồng thời giữ chân nhân tài.

Đẩy mạnh ứng dụng dữ liệu trong quản lý và ra quyết định nhân sự.

* **Cách tiếp cận dữ liệu**

Để khám phá và kể câu chuyện dữ liệu một cách hiệu quả, nhóm đã áp dụng nhiều cách tiếp cận phân tích khác nhau nhằm tìm ra mối quan hệ giữa hình thức làm việc và các yếu tố liên quan như hiệu suất, mức độ hài lòng, thu nhập,...

**Tìm kiếm mối tương quan (correlation):**

Xem xét mối liên hệ giữa hình thức làm việc (Remote/Hybrid/On-site) và các biến như:

* Mức độ hài lòng với công việc (JobSat)
* Tổng thu nhập (CompTotal)
* Số giờ làm việc mỗi tuần (WorkWeekHrs
* Trình độ học vấn và kinh nghiệm làm việc

**Phân tích xu hướng (trend analysis):**

Phân tích cách tỷ lệ làm việc từ xa thay đổi theo:

* Quốc gia, nhóm ngành nghề (DevType)
* Số năm kinh nghiệm (YearsCodePro)
* Trình độ học vấn (EdLevel)

**Phân tích so sánh (comparative analysis):**

So sánh mức độ hài lòng, hiệu suất và thu nhập giữa:

* Người làm việc tại nhà hoàn toàn và người làm việc tại văn phòng
* Nhóm làm việc hybrid so với các nhóm còn lại
* Các quốc gia khác nhau hoặc loại hình công việc khác nhau

**Phân khúc (segmentation):**

Phân chia dữ liệu thành các nhóm đối tượng (ví dụ theo vị trí địa lý, loại công việc, kỹ năng) để phân tích chuyên sâu hành vi và đặc điểm từng nhóm.

**Phân tích định lượng (quantitative):**

Thống kê mô tả và tính toán các giá trị trung bình, tỉ lệ phần trăm, phương sai,… để hiểu rõ mức độ ảnh hưởng của từng biến số.

**Trực quan hóa để tìm mẫu ẩn (visual storytelling):**

Sử dụng các biểu đồ cột, bản đồ nhiệt, biểu đồ phân tán để tìm ra quy luật, điểm bất thường, truyền tải câu chuyện một cách trực quan và sinh động.

### Xác định rõ đối tượng

**Đối tượng mục tiêu của báo cáo bao gồm:**

Ban lãnh đạo cấp cao trong doanh nghiệp công nghệ:

* Những người chịu trách nhiệm hoạch định chiến lược phát triển và mô hình vận hành tổ chức. Họ cần các bằng chứng dữ liệu cụ thể, rõ ràng để đưa ra quyết định về chính sách làm việc linh hoạt, định hướng nhân sự dài hạn và đầu tư hạ tầng số phù hợp.

Phòng nhân sự, đội ngũ phân tích dữ liệu nhân sự:

* Có nhu cầu hiểu sâu về mối quan hệ giữa hiệu suất làm việc, sự hài lòng và mô hình làm việc. Họ cần thông tin để xây dựng công cụ dự báo, hỗ trợ tuyển dụng, đào tạo và phát triển nhân lực.

Nhóm tư vấn chuyển đổi số, nghiên cứu hành vi tổ chức, và nhà đầu tư lĩnh vực công nghệ:

* Quan tâm đến xu hướng lao động, mô hình làm việc tối ưu và hiệu quả đầu tư nhân lực. Cần insight để đánh giá tiềm năng tổ chức và mức độ thích ứng với thị trường lao động toàn cầu.

**Đặc điểm của nhóm đối tượng:**

* Có năng lực phân tích dữ liệu hoặc có nhu cầu sử dụng thông tin để phục vụ chiến lược.
* Quan tâm đến giải pháp thực tiễn, khả thi, được minh chứng bằng dữ liệu cụ thể và trực quan.
* Ưu tiên kết quả có thể chuyển hóa thành hành động (actionable insights), không chỉ dừng ở mô tả.

### Xác định câu chuyện chi tiết

**Bối cảnh:** Khi thị trường lao động ngày càng cạnh tranh, việc tối ưu mô hình làm việc để vừa giữ chân nhân sự, vừa đảm bảo hiệu suất là một bài toán cấp thiết. Dữ liệu khảo sát từ cộng đồng lập trình viên toàn cầu cung cấp một nền tảng đáng tin cậy để tìm kiếm những mối quan hệ ẩn giữa hình thức làm việc và các chỉ số đo lường hiệu quả.

**Đối tượng sử dụng báo cáo:**

* Ban quản lý công ty công nghệ muốn xây dựng chính sách làm việc phù hợp.
* Các chuyên gia nhân sự đang thiết kế mô hình làm việc hiệu quả cho từng nhóm nhân viên.
* Giảng viên, sinh viên hoặc nhà nghiên cứu đang tìm hiểu về tác động của hình thức làm việc đến hiệu suất.

**Cấu trúc kể chuyện dữ liệu:**

1. Đặt vấn đề: Tại sao cần xem xét mô hình làm việc?
2. Trình bày hiện trạng: Phân bố hình thức làm việc hiện nay trong ngành CNTT.
3. Phân tích chi tiết:

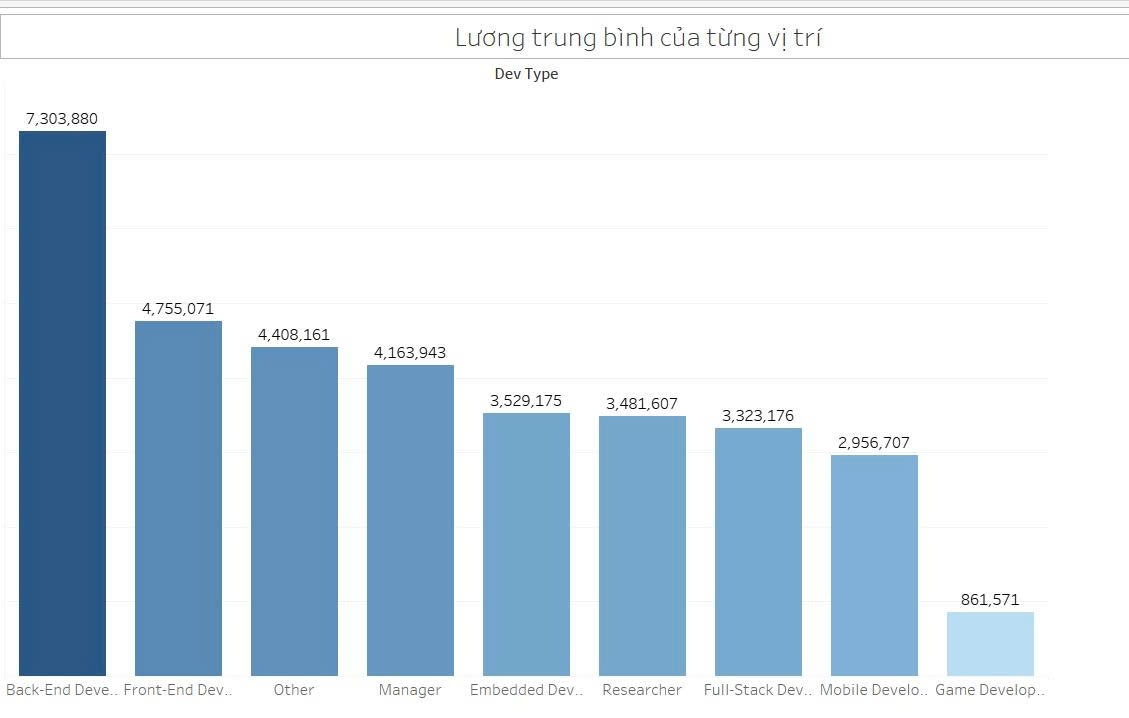
* Tác động đến hiệu suất (WorkWeekHrs)
* Tác động đến sự hài lòng (JobSat)
* Tác động đến thu nhập (CompTotal)

1. So sánh theo nhóm đặc trưng: DevType, EdLevel, Country, YearsCodePro.
2. Tổng kết bằng các insight hành động và đề xuất mô hình phù hợp.

### Trình bày dữ liệu

**Các hình thức biểu đồ và trình bày dữ liệu được sử dụng:**

**Bar Chart (Biểu đồ cột):**



**Boxplot (Biểu đồ hộp):**

**Scatter Plot (Biểu đồ phân tán):**

**Heatmap (Bản đồ nhiệt):**

**Map Chart (Bản đồ):**

**Treemap (Bản đồ cây):**

**Dashboard tổng hợp (Tableau):**

**Nguyên tắc thiết kế biểu đồ:**

* Biểu đồ cần rõ ràng, có tiêu đề, trục, đơn vị và ghi chú.
* Ưu tiên tương tác qua dashboard thay vì trình chiếu tĩnh.
* Sử dụng màu sắc có chủ đích (hài hòa, nổi bật theo nhóm dữ liệu).
* Hạn chế tối đa bảng số liệu dày đặc, chuyển sang biểu đồ hoặc infographic.

### Những điều cần lưu ý

Không nên suy diễn quan hệ nhân quả chỉ từ mối tương quan.

Cần kiểm định thống kê trước khi đưa ra kết luận áp dụng thực tế.

Không sử dụng dữ liệu để tạo thành kiến phân biệt (theo quốc gia, giới tính, tuổi,...).

Mọi biểu đồ, dashboard đều phải kèm giải thích rõ ràng, dễ truy xuất nguồn dữ liệu.

Đảm bảo báo cáo có thể được hiểu bởi cả người chuyên phân tích và người không chuyên.

Nếu mở rộng trong môi trường doanh nghiệp thực, phải tuân thủ quy định bảo mật dữ liệu.

## Kiến trúc hệ thống

### Kiến trúc

A diagram of a computer program

AI-generated content may be incorrect.

### Giải thích

**1. Excel -** Dữ liệu đầu vào (Raw Data)

Nguồn dữ liệu: Dữ liệu được lấy từ khảo sát Stack Overflow Developer Survey 2023, định dạng gốc là .csv

Ý nghĩa: Đây là nguồn dữ liệu gốc chứa hàng chục nghìn dòng phản ánh trung thực các đặc điểm như:

* Hình thức làm việc (RemoteWork)
* Mức độ hài lòng (JobSat)
* Số giờ làm việc mỗi tuần (WorkWeekHrs)
* Mức thu nhập (CompTotal)
* Loại công việc (DevType), quốc gia (Country), v.v.

Vai trò trong hệ thống: Là nguyên liệu đầu vào cần thiết cho toàn bộ hệ thống phân tích.

**2. Python** – Làm sạch và Tiền xử lý (Preprocessing)

Công cụ sử dụng: Jupyter Notebook với các thư viện:

* pandas, numpy: xử lý bảng dữ liệu, thống kê
* matplotlib, seaborn: trực quan hóa sơ bộ

Các thao tác chính:

* Xử lý missing values: Loại bỏ hoặc thay thế các giá trị bị thiếu (NaN) trong các cột quan trọng như JobSat, CompTotal, RemoteWork,…
* Chuẩn hóa định dạng dữ liệu: Chuyển đổi về định dạng phù
* Mã hóa dữ liệu định tính: Ví dụ, RemoteWork có thể được mã hóa từ text sang số để dễ phân tích.
* Trích xuất cột cần thiết: Loại bỏ các cột không liên quan đến mục tiêu dự án để tối ưu hiệu năng xử lý.

Kết quả: Bộ dữ liệu đã sạch, đồng nhất và sẵn sàng phân tích sâu.

**3. Xuất dữ liệu xử lý**

Lưu dữ liệu đã xử lý về dạng .csv trung gian để:

* Đảm bảo tính tái sử dụng và dễ chia sẻ.
* Làm đầu vào cho phần mềm trực quan hóa Tableau.

Vai trò kỹ thuật:

* Tách biệt rõ giai đoạn xử lý Python và trực quan hóa Tableau.
* Giúp kiểm thử và chỉnh sửa nếu phát hiện bất thường trong dashboard.

**4. Tableau –** Trực quan hóa dữ liệu

Mục tiêu: Biến các con số khô khan thành biểu đồ dễ nhìn – hiểu được xu hướng và mối quan hệ ẩn trong dữ liệu.

Chức năng Tableau áp dụng:

* Filter: lọc theo quốc gia, hình thức làm việc, loại công việc, thu nhập,...
* Các loại biểu đồ sử dụng:
* Biểu đồ cột so sánh thu nhập giữa các hình thức làm việc
* Biểu đồ phân tán giữa WorkWeekHrs và CompTotal
* Bản đồ nhiệt thể hiện mức độ hài lòng theo khu vực
* TreeMap và BarChart cho DevType

Kết quả: Xây dựng một dashboard trực quan, tương tác cao và phân tích được xu hướng công việc trong tương lai

**5. Canva, Word –** Trình bày và báo cáo

Canva:

* Thiết kế slide trình chiếu sinh động với biểu tượng, màu sắc, hình ảnh minh họa dễ hiểu.
* Là công cụ hỗ trợ kể chuyện bằng dữ liệu (data storytelling) – giúp người đọc nhanh chóng nắm bắt nội dung.

Microsoft Word:

* Biên soạn báo cáo chính thức: bao gồm các phần phân tích, bảng số liệu, biểu đồ, giải thích và kết luận.
* Đảm bảo tài liệu đáp ứng tiêu chuẩn học thuật và dễ theo dõi khi nộp cho giảng viên hoặc trình bày nội bộ.

## Giải thích về bộ dữ liệu khách hàng

### Các khái niệm

Bộ dữ liệu sử dụng trong dự án được lấy từ khảo sát Stack Overflow Developer Survey 2023. Dữ liệu phản ánh thực trạng nghề nghiệp, hình thức làm việc, trình độ học vấn, thu nhập và mức độ hài lòng của lập trình viên toàn cầu. Dưới đây là các khái niệm và nghiệp vụ chính:

* **RemoteWork**: Hình thức làm việc chính, gồm ba loại: Làm tại văn phòng (In-person), Kết hợp (Hybrid), và Làm việc từ xa (Remote).
* **DevType**: Vai trò công việc của người tham gia, có thể bao gồm nhiều vai trò như Data Scientist, Web Developer, DevOps Engineer, v.v.
* **JobSat**: Mức độ hài lòng với công việc, được chia theo thang đo từ "Rất không hài lòng" đến "Rất hài lòng".
* **CompTotal**: Tổng thu nhập hàng năm (USD), bao gồm lương, tiền thưởng và các phụ cấp.
* **WorkWeekHrs**: Số giờ làm việc trung bình mỗi tuần.
* **EdLevel**: Trình độ học vấn cao nhất đạt được như Cử nhân, Thạc sĩ, Tiến sĩ.
* **YearsCodePro**: Số năm kinh nghiệm lập trình chuyên nghiệp. Dữ liệu ở dạng văn bản như "Less than 1 year", "More than 50 years" cần được chuyển đổi sang dạng số.
* **Country**: Quốc gia làm việc chính của lập trình viên.
* **OrgSize**: Quy mô tổ chức/doanh nghiệp nơi làm việc, chia thành Micro, Small, Medium, Large.
* **Age**: Nhóm tuổi người tham gia khảo sát như 18–24, 25–34, 35–44, v.v.
* **Employment**: Hình thức làm việc, gồm toàn thời gian, bán thời gian, freelancer hoặc sinh viên.

### Các trường dữ liệu

Dữ liệu đầu vào của dự án gồm nhiều cột, mỗi cột phản ánh một khía cạnh của người tham gia khảo sát. Để đảm bảo tính chính xác và khả năng phân tích, mỗi trường dữ liệu được xử lý như sau:

* **RemoteWork** (Hình thức làm việc): Dữ liệu được gom lại thành ba nhóm chính để đơn giản hóa phân tích.
* **DevType** (Vai trò công việc): Là trường dạng multi-label, mỗi dòng có thể chứa nhiều vai trò, cần tách riêng và gom nhóm theo 10 nhóm phổ biến nhất để dễ phân tích.
* **JobSat** (Mức độ hài lòng): Dữ liệu thứ bậc được mã hóa thành số từ 0 (Không hài lòng) đến 2 (Hài lòng) để dễ phân tích định lượng.
* **CompTotal** (Tổng thu nhập): Các giá trị thiếu và ngoại lệ (rất nhỏ hoặc rất lớn) được xử lý hoặc loại bỏ để tránh ảnh hưởng đến mô hình.
* **WorkWeekHrs** (Giờ làm việc mỗi tuần): Loại bỏ các giá trị không hợp lý như dưới 5 giờ hoặc trên 100 giờ mỗi tuần.
* **EdLevel** (Trình độ học vấn): Gom nhóm lại thành 3 nhóm chính là Đại học, Sau đại học, và Khác.
* **YearsCodePro** (Kinh nghiệm lập trình): Các giá trị văn bản như “Less than 1 year” được chuyển thành số thực (ví dụ: 0.5), giúp phân tích định lượng.
* **Country** (Quốc gia): Chuẩn hóa tên quốc gia để tránh trùng lặp (ví dụ: "USA", "United States", "US").
* **OrgSize** (Quy mô tổ chức): Gom lại thành 4 nhóm Micro (<10 người), Small (10–99), Medium (100–999), Large (≥1000).
* **Age** (Nhóm tuổi): Gom lại thành các nhóm tuổi chính để dễ phân tích.
* **Employment** (Hình thức làm việc): Gom thành ba nhóm chính: Employed, Freelancer, Student để phục vụ mô hình hóa.

# Làm sạch và chuyển đổi dữ liệu

## Chuẩn bị dữ liệu

### Giải pháp lưu trữ dữ liệu

Nhóm đã lựa chọn lưu trữ dữ liệu và các tệp xử lý trên nền tảng **Google Drive**, vì lý do:

* **Tiện lợi chia sẻ nội bộ** giữa các thành viên nhóm.
* **Tự động đồng bộ hóa**, tránh mất dữ liệu khi thay đổi thiết bị làm việc.
* **Tương thích tốt với các công cụ văn phòng** như Google Sheets, Google Docs, đồng thời hỗ trợ import/export với Excel, Python.
* **Quản lý phiên bản (version control)** giúp dễ dàng kiểm tra lại quá trình thay đổi file.
* **Miễn phí và dung lượng đủ dùng** cho quy mô dự án.

Việc sử dụng Google Drive góp phần nâng cao hiệu quả làm việc nhóm, đồng thời đảm bảo an toàn và tiện lợi trong truy xuất dữ liệu khi làm việc từ xa.

### Giải pháp phân bố dữ liệu

#### Ý nghĩa việc phân bố dữ liệu

Dữ liệu được phân phối nội bộ theo cơ chế thủ công nhưng có tổ chức:

* **Phân chia thư mục rõ ràng** theo từng giai đoạn: Dữ liệu gốc, Dữ liệu đã làm sạch, File Tableau, Hình ảnh biểu đồ, Slide Canva,...
* **Phân quyền truy cập hợp lý**: thành viên xử lý Python truy cập thư mục "Làm sạch", thành viên thiết kế dashboard truy cập thư mục "Visual hóa",…
* **Tệp được đặt tên nhất quán**: TenBuoc\_TenNguoiThucHien\_NgayThang.xlsx.
* **Lưu vết chỉnh sửa** bằng file log hoặc ghi chú trong notebook Python

#### Trình bày cách phân bố dữ liệu

## Làm sạch dữ liệu

### Các vấn đề ảnh hưởng tới dữ liệu

Bộ dữ liệu gốc tồn tại nhiều vấn đề sau:

* Thiếu dữ liệu (NaN) ở nhiều trường quan trọng: JobSat, CompTotal, RemoteWork.
* Không thống nhất định dạng (ví dụ: Country có cả “US”, “USA”, “United States”).
* Kiểu dữ liệu không đúng (số nhưng lưu dưới dạng chuỗi như ở CompTotal, YearsCodePro).
* Dữ liệu dạng multi-label trong DevType, LanguageWorkedWith gây khó khăn khi phân nhóm

### Các tiêu chí đánh giá chất lượng dữ liệu

Để đảm bảo tính chính xác và giá trị sử dụng của dữ liệu trong quá trình phân tích, cần đánh giá chất lượng dữ liệu dựa trên một số tiêu chí sau:

**1. Tính đầy đủ (Completeness)**

Dữ liệu cần có đủ các trường thông tin quan trọng.

Không bị thiếu nhiều dòng ở các cột then chốt.

Ví dụ: Nếu phân tích JobSat mà >30% dữ liệu bị thiếu thì kết quả dễ sai lệch.

**2. Tính chính xác (Accuracy)**

Dữ liệu phản ánh đúng thực tế, không có lỗi nhập liệu hoặc định dạng.

Ví dụ: WorkWeekHrs không thể là 200 giờ/tuần, hoặc CompTotal không được là số âm.

**3. Tính nhất quán (Consistency)**

Dữ liệu được chuẩn hóa đồng nhất về cách viết, cách đặt tên, định dạng kiểu dữ liệu.

Ví dụ: Các giá trị trong Country phải được viết theo một quy ước duy nhất (viết hoa/thường, tên đầy đủ, không dấu…).

**4. Tính hợp lệ (Validity)**

Dữ liệu tuân thủ các quy tắc logic hoặc quy định của mô hình.

Ví dụ: YearsCodePro chỉ nhận giá trị số từ 0 đến 51 (từ “Less than 1 year” đến “More than 50 years”).

**5. Tính kịp thời (Timeliness)**

Dữ liệu cần được cập nhật và sử dụng đúng thời điểm.

Mặc dù không phải là dữ liệu thời gian thực, nhưng bộ dữ liệu từ Stack Overflow 2023 vẫn mang tính thời sự và đại diện cho xu hướng hiện tại.

**6. Tính dễ truy cập (Accessibility)**

Dữ liệu cần dễ lưu trữ, chia sẻ, truy xuất và thao tác cho tất cả thành viên trong nhóm.

Trong phạm vi dự án, nhóm đã áp dụng các tiêu chí đánh giá chất lượng dữ liệu như sau:

* **Tính đầy đủ**: Nhóm loại bỏ các dòng thiếu dữ liệu ở những cột trọng yếu như JobSat, RemoteWork để đảm bảo phân tích không bị sai lệch.
* **Tính chính xác**: Các giá trị bất thường (outliers) như WorkWeekHrs > 100 hoặc CompTotal < 1.000 được xử lý hoặc loại bỏ nhằm duy trì độ tin cậy.
* **Tính nhất quán**: Các trường văn bản như Country, DevType, EdLevel được chuẩn hóa (viết hoa/thường đồng nhất, xóa khoảng trắng, quy về tên chuẩn) để tránh trùng lặp giá trị do sai khác về cách ghi.
* **Tính hợp lệ**: Kiểm tra kiểu dữ liệu (số, chuỗi) và giới hạn hợp lý (ví dụ: YearsCodePro phải nằm trong khoảng từ 0.5 đến 51) nhằm đảm bảo dữ liệu tuân thủ logic nghiệp vụ.
* **Tính dễ truy cập**: Dữ liệu được lưu trữ theo từng giai đoạn (raw, cleaned, converted) trên Google Drive, chia quyền cụ thể cho từng thành viên, giúp thuận tiện thao tác và đảm bảo an toàn khi làm việc nhóm

### Các bước làm sạch dữ liệu

#### Trình bày các bước làm sạch

Làm sạch dữ liệu (Data Cleaning) là quá trình phát hiện, xử lý và loại bỏ các vấn đề không chính xác, thiếu sót hoặc không nhất quán trong dữ liệu nhằm chuẩn bị cho các bước phân tích tiếp theo. Dưới đây là các bước làm sạch dữ liệu được áp dụng trong hầu hết các dự án xử lý dữ liệu:

**Bước 1: Kiểm tra và loại bỏ giá trị thiếu (Missing Values)**

Xác định các cột có chứa giá trị NaN, null hoặc rỗng.

Quyết định giữ lại hay loại bỏ các dòng/cột tùy theo mức độ quan trọng và tỷ lệ thiếu.

Nếu tỷ lệ thiếu thấp và thuộc cột quan trọng → loại bỏ dòng.

Nếu tỷ lệ thiếu cao → cân nhắc thay thế bằng trung bình, trung vị hoặc mode.

**Bước 2: Chuẩn hóa định dạng (Standardize Formatting)**

Xóa khoảng trắng dư thừa, đồng bộ cách viết hoa/thường.

Chuẩn hóa định dạng ngày tháng, đơn vị tiền tệ hoặc giờ làm việc về cùng một chuẩn.

Ví dụ: viết lại các giá trị "united states", "USA", "us" thành "United States".

**Bước 3: Xử lý giá trị ngoại lệ (Outliers)**

Dùng mô tả thống kê hoặc biểu đồ boxplot để phát hiện giá trị quá lớn/nhỏ bất thường.

Quy tắc dùng: loại bỏ điểm nằm ngoài khoảng [Q1 – 1.5×IQR, Q3 + 1.5×IQR].

Trong dự án, loại bỏ WorkWeekHrs > 100 và CompTotal < 1.000 hoặc > 1.000.000.

**Bước 4: Mã hóa dữ liệu định tính (Encode Categorical Data)**

Chuyển dữ liệu dạng văn bản (string) sang dạng số để phục vụ phân tích định lượng.

Sử dụng kỹ thuật Label Encoding hoặc Ordinal Encoding với các trường như: JobSat, RemoteWork, EdLevel.

**Bước 5: Gom nhóm và phân loại (Grouping)**

Nhóm các giá trị rải rác thành nhóm chính để giảm độ phức tạp và tăng khả năng so sánh.

Ví dụ: gom EdLevel thành 3 nhóm: Đại học – Sau đại học – Khác; Country lấy top 10 quốc gia.

**Bước 6: Xử lý dữ liệu dạng nhiều nhãn (Multi-label Fields)**

Với các cột như DevType chứa nhiều nhãn trong 1 dòng (dạng phân cách bằng dấu ;), thực hiện:

Tách giá trị thành danh sách (.str.split(';'))

Dùng .explode() hoặc kỹ thuật pivot để phân tích riêng từng giá trị.

**Bước 7: Xác định kiểu dữ liệu (Data Type Checking)**

Kiểm tra các cột được định nghĩa đúng kiểu (int, float, string, datetime,...).

Chuyển đổi lại nếu bị sai kiểu do lỗi khi nhập/xuất dữ liệu.

**Bước 8: Lưu dữ liệu đã làm sạch**

Dữ liệu sau khi làm sạch được lưu thành file mới (.csv hoặc .xlsx), đặt tên rõ ràng theo giai đoạn (Cleaned\_StackOverflow2023.csv).

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

Code load dữ liệu và xử lý dữ liệu NaN. Lưu lại thành file dulieu.xlsx

A computer screen shot of text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Kiểm tra giá trị null, chuyển đổi về chuỗi, loại bỏ khoảng trắng đầu/cuối

A computer screen shot of text

AI-generated content may be incorrect.

- Chuẩn hóa cột MainBranch, Country: xóa khoảng trắng, chuẩn hóa chữ hoa/thường.

- Cột MainBranch được phân loại thành 3 nhóm Developer, Semi-technical và other

A screen shot of a computer program

AI-generated content may be incorrect.

A computer screen shot of text

AI-generated content may be incorrect.

* Cắt dữ liệu số nhóm tuổi của cột Age

A screen shot of a computer code

AI-generated content may be incorrect.

* Chuyển đổi giá trị cột SOComm để ngắn gọn và đồng nhất

A computer screen with text

AI-generated content may be incorrect.

- Chuẩn hóa EdLevel, Employment, RemoteWork: gom nhóm thành các nhãn đơn giản để dễ phân tích.

A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

Phần Biến đổi dữ liệu về dạng số

- CompTotal, YearsCodePro: chuyển từ chuỗi sang số để phân tích định lượng.

A black screen with white text

AI-generated content may be incorrect.

* Với 0.5 là "Less than 1 year", 51 là "More than 50 years"

A screen shot of a computer program

AI-generated content may be incorrect.

- SurveyLength, SurveyEase: chuyển về điểm số -1, 0, 1 để thuận tiện trực quan và tính toán.

A screen shot of a computer code

AI-generated content may be incorrect.

Phần Gộp nhóm và phân loại

- OrgSize: từ các giá trị chi tiết (100 to 499, 10,000+) thành nhóm như Small, Medium, Large, Enterprise, Micro, Other

A screen shot of a computer program

AI-generated content may be incorrect.

- DevType: phân loại thành Full-stack Developer, Back-End Developer, Front-End Developer, Mobile Developer, Embedded Developer, Game Developer, Researcher, Manager,…

A screen shot of a computer program

AI-generated content may be incorrect.

Phần Xử lý dữ liệu liệt kê

- Với các cột có dấu ';' như TechEndorse, AIToolCurrently Using: chuyển về định dạng dễ xử lý hơn.

A computer screen with colorful text

AI-generated content may be incorrect.

A computer screen shot of text

AI-generated content may be incorrect.

## Chuyển đổi dữ liệu

### Các trường hợp cần chuyển đổi

 **Chuỗi → Số**: YearsCodePro, CompTotal, SurveyLength, SurveyEase.

 **Gộp nhóm giá trị**: EdLevel, Employment, OrgSize, DevType, Age.

 **Mã hóa phân loại**: JobSat, RemoteWork.

### Các kỹ thuật chuyển đổi

**Phần Biến đổi dữ liệu về dạng số**

- CompTotal, YearsCodePro: chuyển từ chuỗi sang số để phân tích định lượng.

A black screen with white text

AI-generated content may be incorrect.

* Với 0.5 là "Less than 1 year", 51 là "More than 50 years"

A screen shot of a computer program

AI-generated content may be incorrect.

- SurveyLength, SurveyEase: chuyển về điểm số -1, 0, 1 để thuận tiện trực quan và tính toán.

A screen shot of a computer code

AI-generated content may be incorrect.

### Trình bày các phép chuyển đổi trong dự án

- Cột MainBranch được phân loại thành 3 nhóm Developer, Semi-technical và other

A screen shot of a computer program

AI-generated content may be incorrect.

A computer screen shot of text

AI-generated content may be incorrect.

* Cắt dữ liệu số nhóm tuổi của cột Age

A screen shot of a computer code

AI-generated content may be incorrect.

* Chuyển đổi giá trị cột SOComm để ngắn gọn và đồng nhất

A computer screen with text

AI-generated content may be incorrect.

- Chuẩn hóa EdLevel, Employment, RemoteWork: gom nhóm thành các nhãn đơn giản để dễ phân tích.

A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

# Xử lý dữ liệu

## Chuẩn hóa dữ liệu

### Trình bày các bước chuẩn hóa trong dự án

* *Trình bày cụ thể các bước chuẩn hóa dữ liệu bằng Power Query tác động cụ thể đến bộ dữ liệu dự án (từng cột, từng dòng).*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

**…**

## Mô hình hóa dữ liệu

* *Mô hình hóa dữ liệu:*
  + *Data model*
  + *Data modeling*

### Các loại mô hình hóa

* *Trình bày các loại mô hình*
* *Mô hình nào đang được áp dụng cho bộ dữ liệu dự án ?*

…

### Các tiêu chí đánh giá mô hình dữ liệu

* *Trình bày các tiêu chí đánh giá mô hình dữ liệu tốt ?*
* *Mô hình đang dùng trong dự án đáp ứng các tiêu chí nào ?*

…

### Trình bày các bước mô hình hóa

* *Trình bày cụ thể các bước mô hình hóa dữ liệu bằng Power BI trên bộ dữ liệu của dự án.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

…

### Trình bày các bước tạo bảng dữ liệu

* *Trình bày cụ thể các bước tạo bảng dữ liệu bằng Power BI trên bộ dữ liệu của dự án.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

…

# Trực quan hóa dữ liệu

## Các kỹ thuật trực quan hóa

Trong lĩnh vực phân tích dữ liệu, trực quan hóa là công cụ quan trọng giúp truyền tải thông tin nhanh chóng và dễ hiểu. Dưới đây là các kỹ thuật trực quan hóa thường được sử dụng:

**1. Biểu đồ cột (Bar chart / Column chart)**

Thể hiện sự so sánh giữa các nhóm.

Ví dụ: So sánh thu nhập trung bình giữa các hình thức làm việc (Remote, Hybrid, Onsite).

**2. Biểu đồ đường (Line chart)**

Thể hiện xu hướng theo thời gian.

Dù ít dùng trong dự án này (do không phải dữ liệu chuỗi thời gian), nhưng có thể áp dụng nếu phân tích theo năm/kỳ.

**3. Biểu đồ phân tán (Scatter plot)**

Thể hiện mối quan hệ giữa hai biến số liên tục.

Ví dụ: Mối liên hệ giữa số giờ làm việc (WorkWeekHrs) và thu nhập (CompTotal).

**4. Bản đồ nhiệt (Heatmap)**

Dùng để biểu diễn mức độ cao thấp bằng màu sắc.

Ví dụ: Mức độ hài lòng trung bình theo quốc gia hoặc nhóm công việc.

**5. Biểu đồ hộp (Boxplot)**

Dùng để phát hiện outlier và hiểu phân bố dữ liệu.

Áp dụng cho các trường như CompTotal, WorkWeekHrs.

6. Bản đồ địa lý (Map)

Trực quan dữ liệu theo vị trí địa lý.

Ví dụ: Phân bố hình thức làm việc theo quốc gia.

**7. TreeMap / Donut / Pie chart**

TreeMap: thể hiện tỷ trọng từng thành phần trong tổng thể.

Donut/Pie chart: chỉ nên dùng khi có ít hơn 5 nhóm để tránh rối.

**8. Dashboard tương tác (Interactive Dashboard)**

Kết hợp nhiều biểu đồ vào một trang

Dùng filter, parameter, tooltip giúp người xem thay đổi góc nhìn theo nhu cầu.

**Trong dự án, nhóm đã áp dụng các kỹ thuật trực quan hóa cụ thể như sau:**

Bar chart: Dùng để so sánh giữa các nhóm như hình thức làm việc (Remote, Hybrid, Onsite), vai trò (DevType), hoặc quốc gia (Country).

Scatter plot: Phân tích mối liên hệ giữa hai biến số định lượng là số giờ làm việc mỗi tuần (WorkWeekHrs) và tổng thu nhập hàng năm (CompTotal).

Heatmap: Thể hiện mức độ hài lòng trung bình (JobSat) theo từng quốc gia, bằng cách sử dụng sắc độ màu tăng dần.

Map (bản đồ): Trực quan hóa tỷ lệ làm việc từ xa theo quốc gia, giúp dễ nhận ra xu hướng theo khu vực địa lý.

TreeMap: Dùng để thống kê tỷ trọng của các loại vai trò công việc (DevType), cho thấy nhóm nghề nghiệp nào phổ biến nhất.

Dashboard tương tác: Kết hợp nhiều biểu đồ trên cùng một giao diện, cho phép người dùng lọc theo tiêu chí như quốc gia, hình thức làm việc hoặc nhóm nghề nghiệp để xem thông tin chi tiết phù hợp

## Các nguyên tắc trực quan hóa

Để đảm bảo thông tin từ dữ liệu được truyền tải một cách hiệu quả, trực quan hóa cần tuân theo một số nguyên tắc cơ bản như sau:

**1. Rõ ràng và dễ hiểu (Clarity)**

* Tránh làm rối biểu đồ bằng quá nhiều yếu tố.
* Sử dụng tiêu đề rõ ràng, đơn vị đo cụ thể, chú thích cần thiết.
* Tránh sử dụng quá nhiều màu sắc hoặc ký hiệu khó hiểu.

**2. Tối giản (Simplicity)**

* Chỉ hiển thị thông tin quan trọng, không nhồi nhét quá nhiều chi tiết.
* Ưu tiên biểu đồ đơn giản như cột, đường, tròn, bảng thay vì biểu đồ phức tạp khi không cần thiết.

**3. Nhất quán (Consistency)**

* Dùng cùng một màu cho cùng một nhóm dữ liệu trên các biểu đồ khác nhau.
* Cùng một khung phân tích thì trục tung/trục hoành nên thống nhất.

**4. Tập trung vào Insight (Insight-Driven)**

* Biểu đồ phải trả lời được câu hỏi cụ thể hoặc truyền tải một thông điệp nhất định.
* Tránh biểu đồ chỉ dừng ở mô tả mà không đưa ra gợi ý phân tích.

**5. Tương tác và linh hoạt (Interactivity – nếu có dùng dashboard)**

* Khi dùng Tableau, nên dùng **filter, parameter, tooltip, dashboard actions** để người xem có thể tương tác, khám phá dữ liệu theo góc nhìn riêng.

**6. Khả năng tiếp cận (Accessibility)**

* Màu sắc nên tương phản tốt, dễ nhìn.
* Tránh sử dụng màu quá rực hoặc khó phân biệt đối với người bị mù màu.
* Kích thước chữ đủ lớn để trình chiếu hoặc in ấn.

**Nguyên tắc quan trọng nhất trong dự án này: “Rõ ràng và dễ hiểu”**

**Lý do lựa chọn:**

Vì mục tiêu chính của dự án là **giúp nhà quản lý, nhân sự và người không chuyên phân tích dữ liệu** hiểu được xu hướng làm việc từ xa và các yếu tố ảnh hưởng đến hiệu suất – **do đó cần truyền tải bằng biểu đồ dễ đọc, trực tiếp, không rối mắt**.

Dữ liệu trong dự án rất phong phú (nhiều chiều: quốc gia, vai trò, kinh nghiệm, lương, mức độ hài lòng,...), nếu không trình bày rõ ràng thì:

* Người xem sẽ bị “ngợp” vì quá nhiều thông tin.
* Mất đi thông điệp chính của từng dashboard.

→ Vì vậy, **sự rõ ràng và dễ hiểu là nguyên tắc cốt lõi**, tạo tiền đề cho mọi quyết định dựa trên dữ liệu.

## Trình bày tạo các report cho dự án

### Tạo visual thống kê chi tiết

#### Tạo visual filter theo

…

#### Tạo visual filter theo

…

#### Tạo visual filter theo

…

#### Tạo visual tiếp theo ….

…

### Tạo visual thống kê tổng thể

* *Tạo các visual dùng trong trang thống kê tổng thể.*
* *Việc này giúp lưu vết các chỉnh sửa thay đổi trên.*

#### Tạo visual filter theo ngày giao dịch

…

#### Tạo visual thống kê Tổng GTGD

…

#### Tạo visual thống kê Tổng KLGD

…

#### Tạo visual tiếp theo ….

…

# Xây dựng báo cáo

## Dashboard và report

* *Cách tối ưu hóa:*
  + *Dashboard*
  + *Report*

…

## Xây dựng báo cáo

* *Xây dựng báo cáo:*
  + *Dashboard*
  + *Report*

### Dashboard vs Report

* *Trình bày sự khác nhau giữa Dashboard và Report.*
* *Giải thích vì sao làm dashboard, report trong dự án này ?*

…

*Giải thích:* …

### Dashboard

* *Trình bày các loại dashboard có trong dự án.*
* *Giải thích vì sao làm các dashboard này ?*

…

*Giải thích:* …

### Report

* *Trình bày các loại report có trong dự án.*
* *Giải thích vì sao làm các report này ?*

#### Tạo report Detail

*Giải thích:* …

…

#### Tạo report Overview

*Giải thích:* …

…

### Bookmark

* *Trình bày các loại bookmark có trong dự án.*
* *Giải thích vì sao làm các bookmark này ?*

#### Tạo bookmark slicer

*Giải thích:* ….

#### Tạo bookmark slicer chọn

…

*Giải thích:* …

#### Tạo bookmark slicer chọn

…

*Giải thích:* …

#### Tạo visual tiếp theo ….

…

*Giải thích:* …

# KẾT LUẬN

## Báo cáo

### Các bước viết báo cáo

* *Trình bày các bước viết báo cáo phân tích dữ liệu*

…

### Tổng hợp

* *Tổng hợp các câu trả lời trong Câu Chuyện Dữ Liệu ở Phân Tích Khách Hàng, kết hợp với các bước viết báo cáo phân tích, để tạo thành bản báo cáo hoàn chỉnh.*

*Gợi ý:*

…

## Khó khăn

Trong quá trình thực hiện dự án, nhóm gặp một số khó khăn chính như sau:

* Dữ liệu thiếu và không đồng nhất (NaN, lỗi định dạng): Một số cột quan trọng như JobSat, RemoteWork, CompTotal có tỷ lệ giá trị bị thiếu khá cao, gây khó khăn khi phân tích hoặc trực quan hóa.
* Văn bản không thống nhất: Các giá trị trong cột như Country, DevType, EdLevel được nhập theo nhiều cách khác nhau (viết hoa/thường, viết tắt, tên dài), cần xử lý thủ công để chuẩn hóa.
* Dữ liệu rời rạc, thiếu chiều thời gian: Dữ liệu chỉ thu thập một lần/năm, không có yếu tố thời gian (ngày, tháng), khiến việc phân tích xu hướng hoặc thay đổi theo thời gian bị hạn chế.
* Thiếu các biến định lượng phụ trợ: Một số yếu tố quan trọng như KPIs công việc, mức độ hoàn thành mục tiêu, kỹ năng mềm,... không được khảo sát nên khó phân tích toàn diện hiệu suất làm việc từ xa

## Thuận lợi

Mặc dù có nhiều khó khăn, nhóm cũng nhận thấy những thuận lợi sau trong quá trình triển khai:

* Nguồn dữ liệu uy tín và phong phú: Bộ dữ liệu từ Stack Overflow Developer Survey 2023 có quy mô lớn, được thu thập từ hàng chục nghìn lập trình viên toàn cầu, đảm bảo độ tin cậy cao.
* Dữ liệu bao phủ đa dạng yếu tố: Từ kỹ năng, vị trí công việc, mức lương, môi trường làm việc cho tới mức độ hài lòng – phù hợp với mục tiêu nghiên cứu về hiệu suất làm việc từ xa.
* Công cụ xử lý mạnh mẽ, miễn phí: Nhóm sử dụng các phần mềm như Tableau Public, Python (Pandas), Excel – thuận tiện thao tác và học tập.
* Phân công nhóm hợp lý: Mỗi thành viên đảm nhiệm một phần rõ ràng (xử lý, phân tích, trực quan), giúp tăng hiệu quả và tránh chồng chéo công việc

## Hướng phát triển

Để nâng cao giá trị ứng dụng và mở rộng phạm vi phân tích, dự án có thể được phát triển theo các hướng sau:

* Mở rộng dữ liệu theo chiều thời gian: Thu thập dữ liệu nhiều năm liên tiếp để phân tích xu hướng thay đổi trong hành vi làm việc từ xa theo thời gian (trước – trong – sau đại dịch).
* Tích hợp thêm dữ liệu doanh nghiệp nội bộ: Kết hợp dữ liệu từ hệ thống quản lý hiệu suất (HRM, KPI) của công ty để phân tích sâu hơn hiệu quả làm việc theo từng bộ phận hoặc vai trò cụ thể.
* Kết hợp phân tích định lượng và định tính: Phỏng vấn, khảo sát trực tiếp nhân viên hoặc quản lý để làm rõ các nguyên nhân đứng sau sự hài lòng hoặc không hài lòng với mô hình làm việc từ xa.
* Ứng dụng trí tuệ nhân tạo (AI/ML): Xây dựng mô hình học máy dự đoán mức độ hài lòng hoặc năng suất dựa trên đặc điểm cá nhân và môi trường làm việc – từ đó hỗ trợ ra quyết định điều phối nhân sự phù hợp hơn.
* Xây dựng hệ thống dashboard đa nền tảng và tương tác cao: Tối ưu hóa trực quan hóa bằng Tableau với filter theo quốc gia, hình thức làm việc, độ tuổi,… giúp nhà quản lý dễ dàng sử dụng trong thực tế.
* Phổ biến kết quả nghiên cứu đến đối tượng thực tế: Có thể chia sẻ dashboard cho các doanh nghiệp trong ngành CNTT, nhân sự, startup,... để hỗ trợ họ hoạch định chiến lược làm việc linh hoạt