**Bước 1: Problem Definition (Định nghĩa vấn đề):**

Bước đầu tiên trong phân tích dữ liệu là xác định rõ ràng vấn đề hoặc câu hỏi bạn muốn phân tích. Điều này giúp định hướng cho toàn bộ quá trình thu thập và phân tích dữ liệu.

**Bước 2: Data Collection (Thu thập dữ liệu):**

1. Dữ liệu sơ cấp:
   * Đây là loại dữ liệu không có sẵn và cần được thu thập thông qua các phương pháp như phỏng vấn, bảng câu hỏi, khảo sát, v.v.
   * Dữ liệu sơ cấp thường đa dạng và phong phú vì nó được thu thập trực tiếp từ nguồn gốc ban đầu theo yêu cầu cụ thể của nghiên cứu.
2. Dữ liệu thứ cấp:
   * Đây là loại dữ liệu đã có sẵn, được thu thập và cung cấp bởi bên thứ ba như công ty, doanh nghiệp, tổ chức chính phủ, v.v.
   * Dữ liệu thứ cấp thường đã qua xử lý và dễ tiếp cận nhưng có thể không đa dạng và phong phú bằng dữ liệu sơ cấp do hạn chế trong việc điều chỉnh theo nhu cầu cụ thể của nghiên cứu.

**Bước 3: Data Understanding (Hiểu biết về dữ liệu):**

Sau khi thu thập dữ liệu, bước tiếp theo là hiểu rõ về cấu trúc và ý nghĩa của dữ liệu đó. Dữ liệu thường được trình bày dưới dạng bảng với các hàng và cột.

* Hàng (rows): Đại diện cho từng "thể hiện" (instance) hoặc đối tượng trong bộ dữ liệu.
* Cột (columns): Đại diện cho các "biến số" (variables) hoặc thuộc tính của các thể hiện.

Các loại biến số:

1. Biến định tính (Qualitative Variables):
   * Lưu trữ thông tin về tính chất hoặc đặc điểm không thể định lượng bằng số.
   * Ví dụ: màu sắc, loại sản phẩm, ý kiến.
2. Biến định lượng (Quantitative Variables):
   * Lưu trữ thông tin có thể định lượng và thể hiện bằng số.
   * Ví dụ: tuổi tác, trọng lượng, số lượng sản phẩm.

Các loại biến định tính:

1. Biến định danh (Nominal Variables):
   * Biến định danh dùng để phân loại đối tượng mà không có thứ tự hay mức độ khác nhau giữa các giá trị.
   * Ví dụ: giới tính (nam, nữ), loại xe (xe máy, ô tô).
2. Biến phân loại (Categorical Variables):
   * Biến phân loại thường được sử dụng để nhóm các đối tượng thành các danh mục hoặc lớp khác nhau.
   * Ví dụ: loại sản phẩm (điện thoại, máy tính bảng).
3. Biến thứ bậc (Ordinal Variables):
   * Biến thứ bậc thể hiện thứ tự hoặc mức độ giữa các giá trị, nhưng không cho biết sự khác biệt cụ thể giữa chúng.
   * Ví dụ: mức độ hài lòng (rất hài lòng, hài lòng, không hài lòng).

Các loại biến định lượng:

1. Biến khoảng (Interval Variables):
   * Đặc điểm:
     + Biến khoảng có giá trị số và khoảng cách giữa các giá trị có ý nghĩa.
     + Không có điểm gốc tuyệt đối (zero point), nghĩa là số 0 không biểu thị sự vắng mặt hoàn toàn của thuộc tính.
     + Cho phép thực hiện các phép toán cộng và trừ.
   * Ví dụ:
     + Nhiệt độ đo bằng độ C hoặc độ F. Sự khác biệt giữa 20°C và 30°C có ý nghĩa, nhưng 0°C không có nghĩa là không có nhiệt độ.
2. Biến tỉ lệ (Ratio Variables):
   * Đặc điểm:
     + Biến tỉ lệ có giá trị số và khoảng cách giữa các giá trị có ý nghĩa.
     + Có điểm gốc tuyệt đối (zero point), nghĩa là số 0 biểu thị sự vắng mặt hoàn toàn của thuộc tính.
     + Cho phép thực hiện các phép toán như cộng, trừ, nhân, chia.
   * Ví dụ:
     + Chiều cao, cân nặng, thu nhập. Một người nặng 0 kg có nghĩa là không có khối lượng, và bạn có thể nói rằng một người nặng 60 kg gấp đôi người nặng 30 kg.

Xác định kiểu dữ liệu của biến số. Xác định có bao nhiêu mẫu dữ liệu (bao nhiêu cột, bao nhiêu dòng).

**Bước 4: Xử lý dữ liệu (Data Processing):**

Xử lý lỗi dữ liệu bao gồm việc sửa các giá trị dữ liệu không đúng, định dạng không thống nhất, dữ liệu bị thiếu (missing data), dữ liệu bị trùng (duplicate), dữ liệu rỗng (null), và dữ liệu bị nhiễu (noise).

**Bước 5: Khám phá insight data (Discovery Insight Data):**

Kỹ thuật EDA (Exploratory Data Analysis) bao gồm các kỹ năng như tổng hợp (summarization), thống kê (statistics), và trực quan hóa (visualization).