Exploratory Data Analysis (EDA) Framework for Data Projects

Exploratory Data Analysis (EDA) là bước quan trọng giúp hiểu rõ dữ liệu trước khi xây dựng mô hình hoặc đưa ra quyết định. Dưới đây là các bước EDA chuyên nghiệp, có thể áp dụng cho hầu hết các bộ dữ liệu:

0. Problem Understanding & Analytical Questions

- Xác định rõ **vấn đề kinh doanh (Business Problem)** hoặc mục tiêu dự án.
    - Ví dụ: Dự đoán khách hàng rời đi, phát hiện gian lận, dự báo doanh số,...
- Đặt câu hỏi phân tích:
    - Những yếu tố nào ảnh hưởng đến kết quả (target variable)?
    - Các nhóm đối tượng/khách hàng/nền tảng nào có xu hướng đặc biệt?
- Xác định rõ **biến mục tiêu (Target Variable)** và các yếu tố liên quan.

---

1. Data Overview - Loading & Basic Exploration

- Đọc dữ liệu từ các nguồn (CSV, Database, API, etc.).
- Kiểm tra kích thước dữ liệu, kiểu dữ liệu, thông tin cơ bản.
- Hiểu rõ **Data Dictionary** (ý nghĩa từng cột/feature).

a. Data Quality Assessment

- **Missing Values Analysis**:
    - Xác định các cột có giá trị thiếu (null, NaN).
    - Phân tích tỷ lệ thiếu dữ liệu theo cột.
- **Outlier Detection**:
    - Tìm giá trị ngoại lệ ở numerical features.

    - Phân tích các giá trị bất thường (dùng Boxplot, Z-score, IQR).
- **Data Consistency & Duplicates**:
    - Kiểm tra dữ liệu trùng lặp.
    - Kiểm tra lỗi nhập liệu, giá trị không hợp lệ.

b. Feature Classification (Identify Feature Types)

- **Categorical Variables**: Nhóm biến phân loại (nominal, ordinal).
- **Numerical Variables**: Phân loại continuous (liên tục) và discrete (rời rạc).
- **Date/Time Variables**: Thời gian, ngày tháng.
- **Mixed-Type Variables**: Các cột dạng alphanumeric, cần xử lý đặc biệt.
- Định nghĩa rõ **Target Variable** cho bài toán.

2. EDA by Visualization

a. Univariate Analysis (Phân tích từng biến độc lập)

- **Numerical Features**:
    - Phân tích phân phối dữ liệu (histogram, boxplot, density plot).
    - Kiểm tra outliers, skewness.
- **Categorical Features**:
    - Tần suất xuất hiện các giá trị (barplot, countplot).
    - Phân tích tỷ lệ xuất hiện của các nhóm.
- **Date/Time Features**:
    - Phân tích xu hướng thời gian (time series decomposition).

b. Bivariate Analysis (Mối liên hệ với Target Variable)

- Phân tích mối quan hệ giữa feature và target:
    - Categorical vs Target: Barplot, Grouped Stats.

    - Numerical vs Target: Boxplot, Violinplot, Correlation.
- Tìm hiểu các yếu tố ảnh hưởng trực tiếp đến target.
- So sánh tỷ lệ, sự khác biệt giữa các nhóm dữ liệu.

c. Multivariate Analysis (Tương tác giữa nhiều feature)

- Phân tích mối quan hệ nhiều chiều:
    - Crosstab, Pivot Table.
    - Heatmap thể hiện mức độ tương quan.
    - Pairplot cho numerical features.
- Phân tích tương tác giữa các feature và tác động đến target.

3. Data Cleaning & Feature Engineering (Dựa trên EDA)

- **Handling Missing Values**:
    - Xóa, thay thế (impute), hoặc xây dựng mô hình dự đoán giá trị thiếu.
- **Outlier Treatment**:
    - Cắt ngưỡng (clipping), winsorization, hoặc loại bỏ.
- **New Feature Creation**:
    - Feature Transformation (log, scaling).
    - Binning continuous features.
    - Tạo các feature tương tác.
    - Extract thông tin từ text, date, id (ex: extracting titles, time features).

4. Key Insights Summary

- Tóm tắt các phát hiện quan trọng:
    - Các yếu tố ảnh hưởng mạnh đến target.
    - Mối quan hệ đáng chú ý giữa các nhóm dữ liệu.
    - Phân bố dữ liệu tổng thể, các bất thường quan trọng.

- Kết nối các insight với bài toán kinh doanh.
- Đề xuất hướng xử lý tiếp theo.