MODULE 3 :

EDA là quá trình phân tích dữ liệu để tóm tắt đặc điểm chính, thường thông qua thống kê và trực quan hóa.

EDA quan trọng vì giống như “làm quen” với dữ liệu trước khi xây dựng mô hình giúp phát hiện dữ liệu sai, cần làm sạch thêm, hoặc cần thêm dữ liệu.

các công cụ: pandas, matplotlib / seaborn

Các kỹ thuật EDA:

-Thống kê tóm tắt (summary statistics):Trung bình (mean), trung vị (median), min, max, mode.

-Trực quan hóa dữ liệu (visualization) Histogram xem phân phối. Scatter plot xem quan hệ giữa hai biến. Box plot xem phân phối và phát hiện outlier.

-Lấy mẫu dữ liệu (sampling):

Khi dataset lớn dùng .sample() trong pandas để lấy ngẫu nhiên một số hàng

Stratified sampling (lấy mẫu phân tầng) giữ đúng tỷ lệ của các lớp (ví dụ: tỷ lệ bệnh nhân có bệnh trong mẫu phản ánh đúng tỷ lệ thật trong dân số).

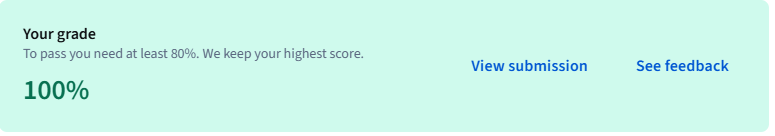
Matplotlib = mạnh mẽ, cơ bản.

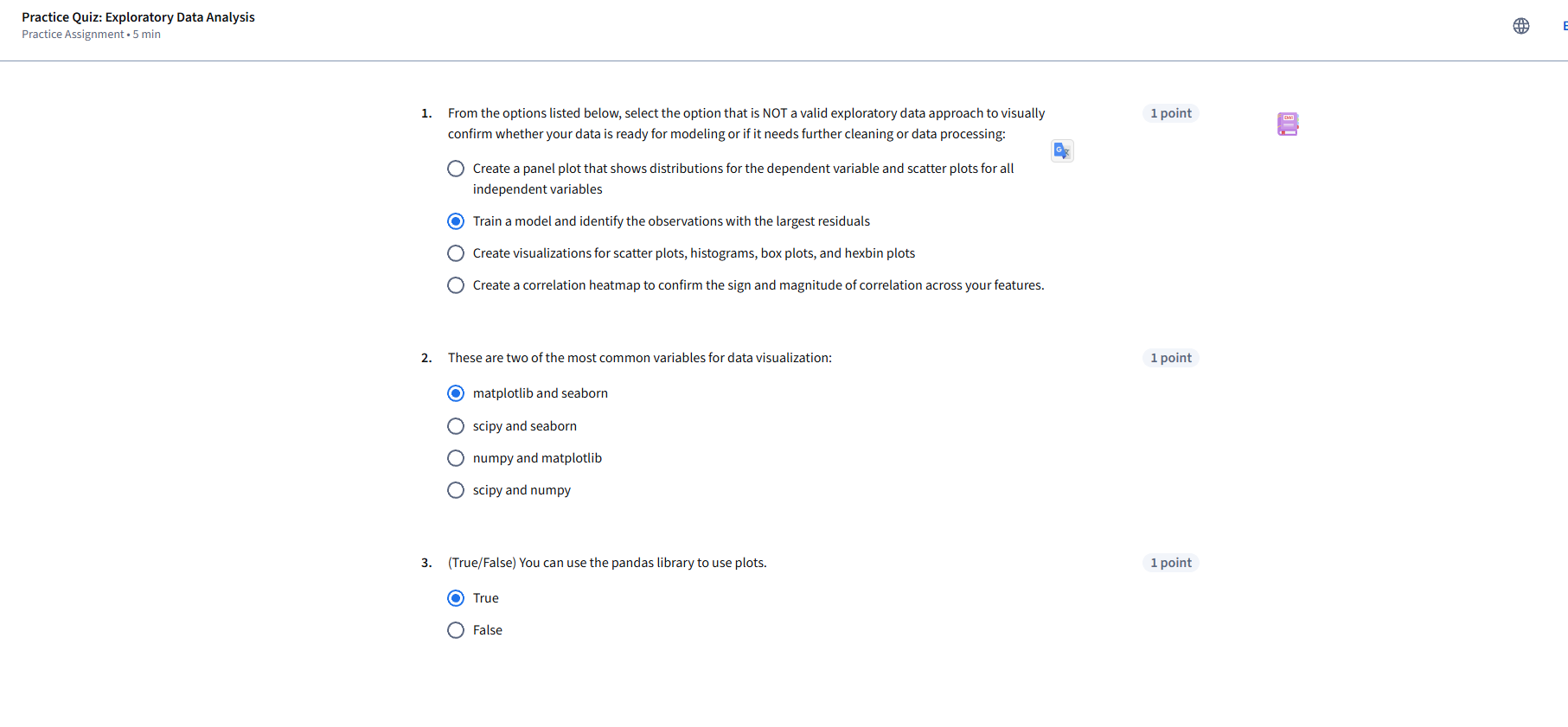
Pandas .plot() = tiện lợi, nhanh.

Seaborn = đẹp, nhiều tính năng thống kê sẵn.

Dùng EDA để tạo scatter plot, histogram, box plot, heatmap, pairplot,.. nhằm hiểu rõ hơn về dữ liệu, phân phối, outlier và mối quan hệ giữa biến.

Quiz : Practice Quiz: Exploratory Data Analysis





* Feature Engineering và Transformation: Dữ liệu thô thường không phù hợp trực tiếp cho mô hình

Mã hóa đặc trưng (Feature Encoding) Các mô hình ML yêu cầu dữ liệu số. Với dữ liệu phân loại (categorical), cần biến đổi thành số (VD: one-hot encoding).

Chuẩn hóa / Scale dữ liệu (Feature Scaling)

Hồi quy tuyến tính và giả địn

Vấn đề phân phối dữ liệu

Các phép biến đổi dữ liệu

Feature Expansion (Mở rộng đặc trưng) thêm đặc trưng bậc cao (polynomial features): x², x³, … và giúp mô hình tuyến tính mô phỏng quan hệ phi tuyến (non-linear) mà vẫn giữ bản chất là "linear regression".

Sklearn PolynomialFeatures

Lựa chọn biến (Feature Selection) không phải tất cả biến đều hữu ích cho mô hình cần chọn ra tập biến phù hợp nhất. Các biến này thường phải biến đổi (transform) trước khi đưa vào mô hình.

Mã hóa biến phân loại (Encoding) Vì mô hình ML cần **dữ liệu số**, dữ liệu phân loại (categorical) phải chuyển sang số. có 3 cách chính Mã hóa nhị phân (Binary encoding), Mã hóa One-Hot (One-Hot encoding), và Mã hóa thứ tự (Ordinal encoding)

Mở rộng và tỉ lệ hóa biến (Feature Scaling / Normalization) Các biến số thường có thang đo khác nhau (VD: giá = 10 → 10.000, số cửa hàng = 1 → 100).

Nếu để nguyên, biến lớn sẽ lấn át biến nhỏ thì mô hình học sai

Các phương pháp chuẩn hóa phổ biến:

1 Standard Scaling (Z-score normalization)

2Min-Max Scaling

3 Robust Scaling

Biến số liên tục (continuous numerical features)

-Biến đổi cần dùng:

Standard Scaling (chuẩn hóa Z-score: trừ trung bình, chia độ lệch chuẩn)

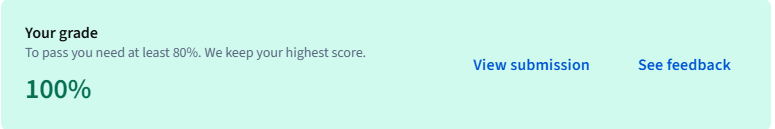
Min-Max Scaling (thu nhỏ về [0,1])

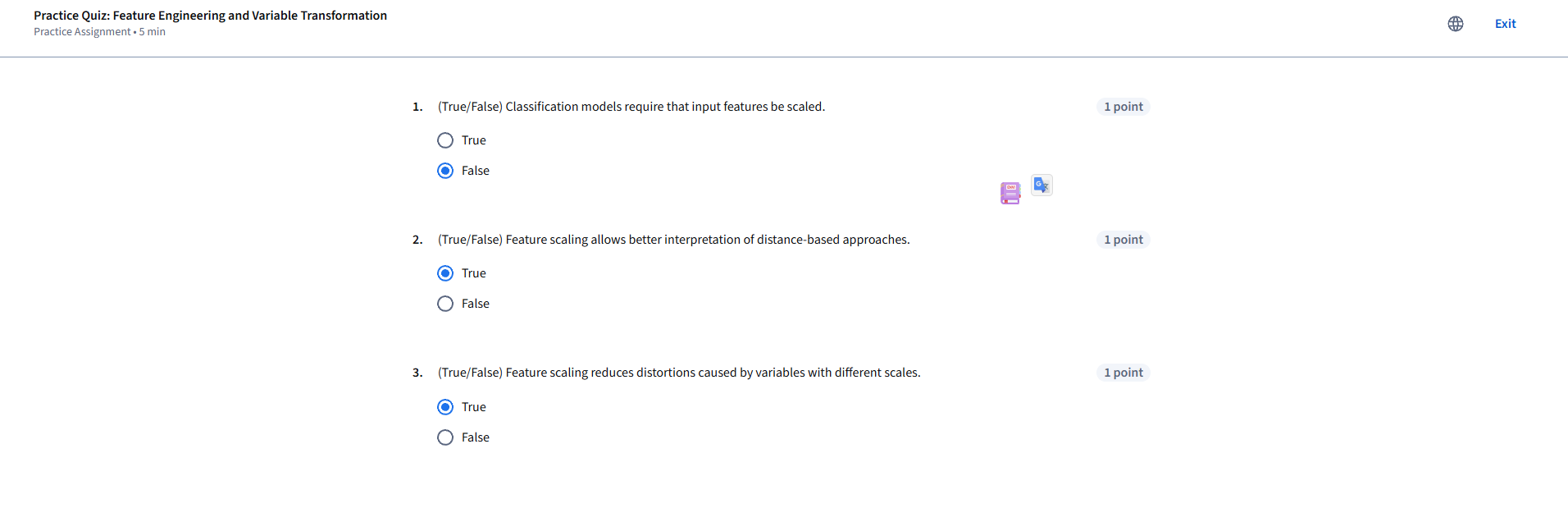
Robust Scaling (dùng median và IQR, chống nhiễu/outlier)

Biến danh nghĩa (categorical nominal features – không có thứ tự, ví dụ: màu sắc, giới tính)

Biến thứ tự (categorical ordinal features – có thứ tự, ví dụ: thấp, trung bình, cao)

Quiz : Practice Quiz: Feature Engineering and Variable Transformation





Graded Quiz: Module 3 - Exploratory Data Analysis and Feature Engineering

