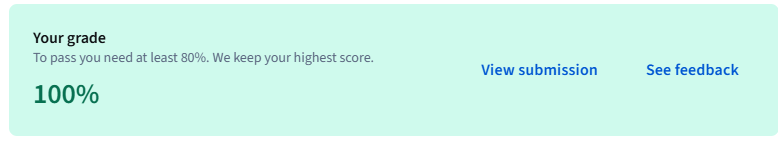
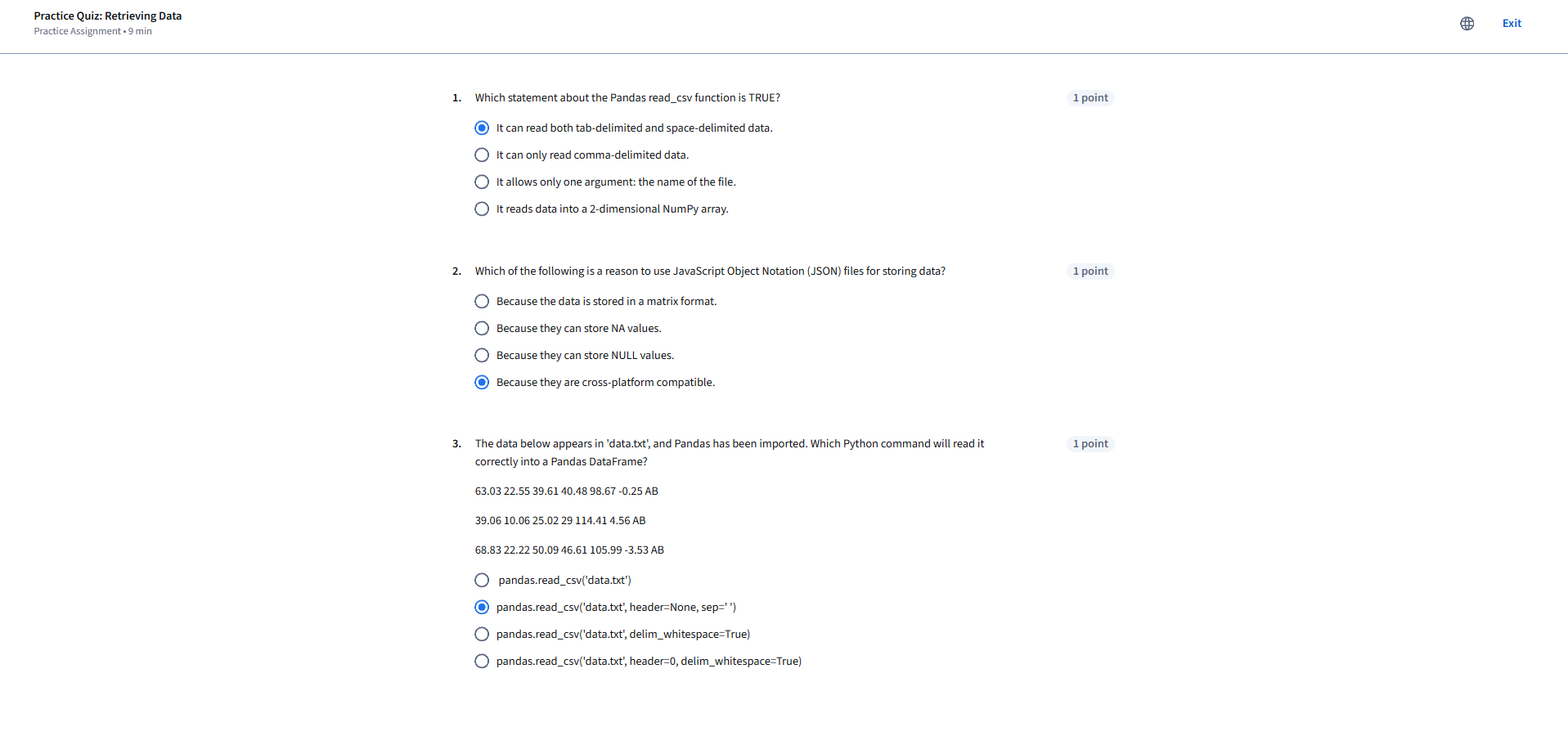
MODULE 2 : LẤY DỮ LIỆU VÀ CLEAN DATA

Bước đầu tiên của ML là lấy dữ liệu có nhiều nguồn khác nhau: CSV, JSON, SQL, NoSQL, API, cloud…

CSV: dữ liệu cách nhau bằng dấu phẩy. JSON: chuẩn phổ biến, nhìn giống dictionary trong Python. SQL: dữ liệu quan hệ, có schema rõ ràng. NoSQL: phi quan hệ, hay lưu ở dạng JSON. Phổ biến là MongoDB. API & cloud: nhiều dữ liệu được public qua API

Quiz: Practice Quiz: Retrieving Data



Clead data:

Dọn dẹp dữ liệu là bước đầu tiên quan trọng: nếu dữ liệu bẩn thì mô hình học sai kết quả vô dụng. Kiểu “rác vào thì rác ra”.

Quan sát sai hoặc nhãn sai mô hình học sai mối quan hệ giữa feature và target. Ví dụ trong ImageNet nếu gán nhãn nhầm 1 tấm hình thì model cũng sẽ bị đánh lừa.

Các vấn đề thường gặp với dữ liệu bẩn: Thiếu dữ liệu, quá nhiều dữ liệu nằm rải rác ở nhiều nơi, dữ liệu trùng lặp, dữ liệu không nhất quán, outlier,, Vấn đề nguồn dữ liệu lấy từ nhiều hệ thống khác nhau (on-premise vs cloud, DB khác loại) thì dễ bị mismatch.

Xử lý dữ liệu trùng lặp: Nếu dữ liệu thật sự có thể trùng (vd: 2 bông hoa giống hệt nhau trong dataset Iris) thì giữ lại.

Nếu chỉ là duplicate file (vd: ảnh giống y hệt trong bài toán gán nhãn) thì bỏ đi.

Luôn nên lọc dữ liệu, nhưng cũng cần giữ bản gốc để tham chiếu sau này.

Xử lý giá trị thiếu (Missing values): Xóa hàng/cột nhanh, dễ, không phải đoán giá trị thay thế nhưng có thể mất nhiều dữ liệu quan trọng, Thường xóa **hàng**, ít khi xóa **cột** vì cột nào cũng hữu ích

Gán giá trị (Imputation) thay thế giá trị null bằng mean / median / ước lượng khác.

chọn cách nào tùy dataset + context (ít missing thì xóa, nhiều thì nên gán/masking)

Outliers là điểm dữ liệu khác biệt hẳn so với phần lớn quan sát. đánh giá trước khi loại bỏ, vì có outlier thật sự mang ý nghĩa.

Dư lượng (Residuals) là Giá trị thực tế – Giá trị dự đoán. Kiểu cho thấy mức độ mô hình sai ở mỗi điểm dữ liệu.

Các loại dư lượng dùng để phát hiện ngoại lệ:

dư lượng chuẩn hóa (Standardized Residuals), Công thức: residual / standard error. Giúp so sánh công bằng khi dữ liệu ở các thang đo khác nhau.

Dư lượng đã xóa (Deleted Residuals) Cách làm: xóa một quan sát huấn luyện lại mô hình xem sự khác biệt so với mô hình ban đầu.

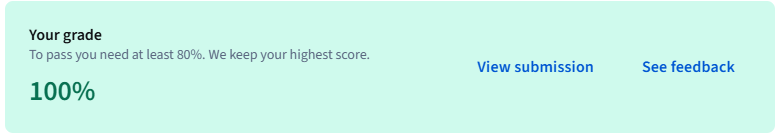
Nếu khác biệt lớn điểm đó có ảnh hưởng mạnh (có thể là ngoại lệ).

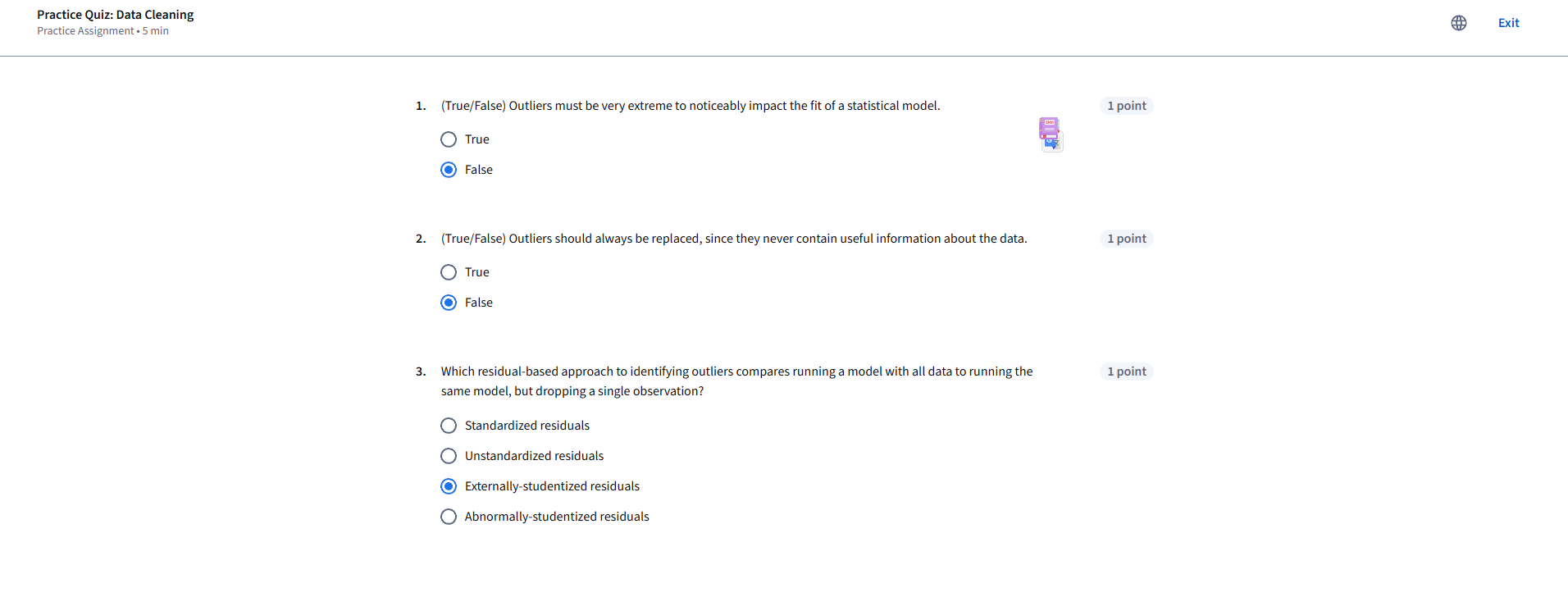
Dư lượng sinh viên (Studentized Residuals) Chính là dư lượng đã xóa nhưng được chuẩn hóa (giống loại 1 nhưng trên mô hình sau khi bỏ đi điểm đó).

Được dùng phổ biến nhất để xác định outlier trong hồi quy.

Cách xử lý ngoại lệ: Xóa hẳn, thay thế, biến đổi dữ liệu (Transformation),dự đoán lại giá trị, giữ nguyên.

Quiz : Practice Quiz: Data Cleaning





Graded Quiz: Module 2 - Retrieving Data and Cleaning Data

