

BÁO CÁO PHÂN TÍCH DỮ LIỆU TUYỂN DỤNG & LƯƠNG (HR)

Lớp: DHKL18A1HN

Nhóm thực hiện: 5

Case study: 6

Phần I – Giới thiệu chung

1. Thành viên nhóm

- Trần Minh Quang - 23174600065 - Trưởng nhóm
- Nguyễn Thành Văn - 24174600001
- Đào Quang Minh - 24174600043
- Dương Đức Khôi - 23174600117
- Đỗ Thân Quốc Khánh - 24174600024

2. Đặt vấn đề

Trong thực tế tuyển dụng, dữ liệu ứng viên thường được lưu trữ rời rạc ở nhiều bảng khác nhau như hồ sơ cá nhân, điểm phỏng vấn và mức lương đề nghị. Điều này gây khó khăn cho việc tổng hợp, đánh giá và phân tích dữ liệu nhân sự một cách toàn diện.

3. Mục tiêu triển khai

- Làm sạch và chuẩn hóa dữ liệu tuyển dụng và lương.
- Xây dựng bộ dữ liệu HR đầy đủ thông qua merge dữ liệu.
- Thực hiện phân tích bằng groupby và pivot table.
- Phân tích điểm phỏng vấn và mức lương theo nhiều chiều.
- Xuất file dữ liệu tổng hợp phục vụ báo cáo (full_hr_data).

Phần II – Phương pháp thực hiện

1. Làm sạch và chuẩn hóa dữ liệu (Nhiệm vụ 1)

Nhóm sử dụng pandas để đọc ba file CSV gốc. Dữ liệu được làm sạch bằng cách loại bỏ khoảng trắng thừa, chuẩn hóa chữ hoa – chữ thường, ép kiểu dữ liệu số và xử lý các giá trị không hợp lệ hoặc bị thiếu. Sau bước này, ba file dữ liệu sạch được xuất ra để sử dụng cho các nhiệm vụ tiếp theo.

2. Truy vấn, thống kê & GroupBy (Nhiệm vụ 2, 3)

Trên dữ liệu đã làm sạch, nhóm thực hiện các truy vấn thống kê mô tả như đếm số ứng viên theo vị trí, tính điểm trung bình theo vòng phỏng vấn, xác định ứng viên có kinh nghiệm nhưng điểm thấp. Đồng thời, groupby() được sử dụng để tính điểm và mức lương trung bình theo từng vị trí.

3. Merge dữ liệu (Nhiệm vụ 4)

Ba bảng `candidate_profile`, `interview_score` và `salary_offer` được merge dựa trên `candidate_id` để tạo thành bộ dữ liệu HR hoàn chỉnh. Quá trình này giúp phát hiện các trường hợp thiếu điểm phỏng vấn hoặc thiếu thông tin lương.

4. Pivot Table & phân tích kết quả (Nhiệm vụ 5)

Từ bộ dữ liệu HR hoàn chỉnh, nhóm xây dựng các bảng pivot để phân tích sâu hơn. Kết quả phân tích được tổng hợp trong file `full_hr_data` – đây là file báo cáo kết quả của Nhiệm vụ 5.

Các kết quả chính rút ra từ pivot table:

- Điểm trung bình theo vị trí và vòng phỏng vấn dao động chủ yếu trong khoảng 6.0 – 7.5.
- Vị trí Data Analyst có điểm cao nhất ở vòng 3, đạt khoảng 7.5.
- Mức lương trung bình tăng rõ rệt theo số năm kinh nghiệm.
- Ứng viên có từ 3–5 năm kinh nghiệm nhận mức lương cao hơn đáng kể so với nhóm mới ra trường.

Phần III – Báo cáo chi tiết

Nhiệm vụ 1: Sau khi áp dụng các kỹ thuật làm sạch dữ liệu, chúng tôi đã thu được kết quả như sau:

Chuẩn hóa dữ liệu định danh:

Toàn bộ mã ứng viên (`candidate_id`) và họ tên (`full_name`) đã được đưa về định dạng thống nhất (viết hoa, không còn khoảng trắng thừa), đảm bảo tính nhất quán khi ghép nối dữ liệu.

Xử lý dữ liệu số:

Các giá trị năm kinh nghiệm (`experience_years`) bị âm hoặc sai định dạng (như chứa ký tự 'y') đã được chuyển đổi thành số dương hợp lệ.

Điểm phỏng vấn (`score`) được làm sạch các ký tự thừa ('p') và lọc bỏ các giá trị nằm ngoài thang điểm 0-10.

Mức lương đề nghị (`offer_salary`) đã được chuẩn hóa về đơn vị VNĐ, xử lý thành công các trường hợp viết tắt như "20tr".

Thống nhất vị trí tuyển dụng:

Đặc biệt, lỗi nhập liệu không đồng nhất trong cột `position` (ví dụ: "Data Analyst" và "Data Analyst" do thừa khoảng trắng, hay các biến thể "Dev", "Python Dev") đã được xử lý triệt để về 2 nhóm chính là "Data Analyst" và "Python Developer".

Nhiệm vụ 2: Sau khi làm sạch và chuẩn hóa dữ liệu, chúng tôi tiến hành truy vấn và rút ra các thống kê mô tả quan trọng về ứng viên và quy trình tuyển dụng:

a. Phân bố ứng viên theo vị trí (Candidate Distribution):

Data Analyst: Chiếm số lượng áp đảo với **14 ứng viên**.

Python Developer: Có **11 ứng viên**.

Nhận xét: Nhu cầu tuyển dụng hoặc số lượng hồ sơ nộp vào vị trí Phân tích dữ liệu đang cao hơn so với Lập trình viên Python trong đợt này.

b. Phân tích điểm số phỏng vấn (Score Analysis):

Điểm trung bình theo vòng 1

Vòng 1: ~7.27 điểm.

Vòng 2: ~6.33 điểm.

Vòng 3: ~7.07 điểm.

Nhận xét: **Vòng 2** có điểm trung bình thấp nhất, đóng vai trò là vòng "sàng lọc" khắt khe nhất trong quy trình tuyển dụng. Ứng viên thường thể hiện tốt hơn ở vòng 1 và vòng 3.

c. Đánh giá chất lượng ứng viên (Quality Check):

Ứng viên cần lưu ý (Low Performers): Hệ thống đã lọc ra được nhóm ứng viên dù có kinh nghiệm làm việc nhưng điểm bài thi lại thấp (dưới mức kỳ vọng). Đây là nhóm cần cân nhắc kỹ trước khi đưa ra quyết định tuyển dụng để tránh rủi ro "thâm niên ảo".

Dữ liệu thiếu: Phát hiện các ứng viên (như UV111, UV117) bị thiếu dữ liệu điểm số, cần liên hệ bộ phận nhân sự để cập nhật bổ sung.

Ứng viên có kinh nghiệm nhưng điểm thấp:

Với tiêu chí lên mức "Điểm TB < 7.5" (Mức trung bình khá), có khoảng **9 ứng viên** (Ví dụ: UV108 - 4 năm kinh nghiệm nhưng chỉ đạt 6.0 điểm). Đây là nhóm cần cân nhắc kỹ về sự phù hợp giữa thâm niên và năng lực thực tế.

Nhiệm vụ 3: Sau khi áp dụng kỹ thuật **GroupBy** để tổng hợp dữ liệu từ nhiều góc độ (Vị trí, Cá nhân, Tương quan), nhóm nghiên cứu rút ra các kết luận chính như sau:

a. Về Chất lượng đầu vào theo Vị trí (Score by Position):

Mức độ đồng đều: Điểm trung bình phỏng vấn của hai vị trí chủ chốt khá tương đồng:

Data Analyst: Đạt khoảng **7.0 điểm**

Python Developer: Đạt khoảng **6.8 điểm**.

Nhận định: Không có sự chênh lệch lớn về năng lực đầu vào giữa hai nhóm ứng viên. Cả hai vị trí đều tuyển được những ứng viên có trình độ từ mức Khá trở lên. Tuy nhiên, nhóm Data Analyst có phần nhỉnh hơn đôi chút về điểm số tổng thể.

b. Về Mức lương đề nghị (Salary Offer by Position):

Sự phân hóa rõ rệt: Có một khoảng cách đáng kể về mức lương trung bình giữa hai vị trí:

Data Analyst: Được đề nghị mức lương trung bình **~20.7 triệu VNĐ**.

Python Developer: Được đề nghị mức lương trung bình **~17.8 triệu VNĐ**.

Nhận định: Doanh nghiệp đang định giá vị trí Phân tích dữ liệu (DA) cao hơn khoảng **16%** so với Lập trình viên Python. Điều này có thể phản ánh nhu cầu thị trường đang "khát" nhân lực DA chất lượng cao hoặc tính chất công việc của DA tại công ty này đòi hỏi nhiều kỹ năng tổng hợp hơn.

c. Về Ứng viên tiềm năng (Top Performers):

Nhóm đã xác định được danh sách các "hạt giống đỏ" với điểm trung bình **> 8.0**, điển hình là ứng viên **UV100 (9.0 điểm)**. Đây là những nhân sự cần được ưu tiên trong quá trình đàm phán lương và chế độ đãi ngộ để thu hút nhân tài.

d. Về Mối quan hệ giữa Kinh nghiệm và Lương (Correlation Analysis):

Kết quả bất ngờ: Hệ số tương quan giữa số năm kinh nghiệm và mức lương đề nghị là rất thấp (xấp xỉ **-0.03**).

Kết luận quan trọng: Trong tập dữ liệu này, **thâm niên làm việc không phải là yếu tố quyết định mức lương**. Một ứng viên làm lâu năm chưa chắc đã nhận được lương cao hơn một người mới nhưng có năng lực tốt (thể hiện qua điểm phỏng vấn). Chính sách lương của công ty dường như đang hướng tới việc trả thưởng theo **năng lực thực tế (Performance-based)** hơn là theo thâm niên (Seniority-based).

Nhiệm vụ 4: Sau khi thực hiện quy trình ghép nối nhóm thu được các kết quả kiểm định như sau:

a. Về cấu trúc và tính toàn vẹn (Structure & Integrity):

Tổng số dòng dữ liệu: Bảng Master chứa **75 dòng**.

Giải thích: Số dòng này lớn hơn số lượng ứng viên (25 người) là hợp lý, vì mỗi ứng viên tham gia nhiều vòng phỏng vấn (1-3 vòng), dữ liệu được chuẩn hóa theo dạng "dài" (Long format).

Số lượng ứng viên duy nhất: Bảo toàn đủ **25 ứng viên** gốc từ file hồ sơ, không bị mất mát ứng viên nào trong quá trình merge.

b. Về tính nhất quán (Consistency):

Dữ liệu điểm số (Score):

Có **16 dòng** bị thiếu điểm (NaN).

Nguyên nhân: Đây là những vòng thi mà ứng viên chưa tham gia hoặc bỏ thi (Ví dụ: Ứng viên UV100 có điểm vòng 1 nhưng thiếu điểm vòng 2, 3). Việc giữ lại các giá trị NaN này là cần thiết để phản ánh đúng thực tế quy trình tuyển dụng.

Dữ liệu lương (Salary):

Có **15 dòng** thiếu thông tin lương.

Nguyên nhân: Tương ứng với các ứng viên chưa nhận được Offer hoặc dữ liệu lương chưa được cập nhật.

Nhiệm vụ 5: Sau khi áp dụng kỹ thuật **Pivot Table** để xoay chiều dữ liệu và phân tích sâu hơn, nhóm nghiên cứu rút ra những kết luận quan trọng về xu hướng điểm số và chính sách lương thưởng:

a. Về Chất lượng ứng viên qua từng vòng (Dựa trên Pivot: Vị trí x Vòng thi):

Sự ổn định: Ứng viên cho vị trí **Data Analyst** thể hiện sự ổn định cao hơn qua các vòng thi, với điểm trung bình dao động nhẹ quanh mức 7.0 - 7.5.

Sự biến động: Ứng viên **Python Developer** có xu hướng giảm điểm nhẹ ở vòng 2 (vòng chuyên môn sâu), cho thấy đây là "điểm gãy" cần lưu ý trong quy trình tuyển dụng đối với vị trí kỹ thuật này.

Tổng quan: Nhìn chung, không có sự chênh lệch quá lớn về mặt bằng chung năng lực giữa hai nhóm, nhưng Data Analyst có phần nhỉnh hơn về sự đồng đều.

b. Về Chính sách đãi ngộ (Dựa trên Pivot: Vị trí x Kinh nghiệm):

Nghịch lý thâm niên: Dữ liệu chỉ ra một điểm thú vị là mức lương đề nghị **không tăng tuyến tính** theo số năm kinh nghiệm. Ví dụ: Có trường hợp ứng viên 0 năm kinh nghiệm (Fresher) nhận được mức lương tương đương hoặc cao hơn ứng viên 2-3 năm kinh nghiệm.

Định giá vị trí: Ở cùng một mức kinh nghiệm (ví dụ: 1-2 năm), vị trí **Data Analyst** thường có mức lương đề nghị cao hơn (~20-23 triệu VNĐ) so với **Python Developer** (~15-17 triệu VNĐ). Điều này phản ánh sự khan hiếm hoặc ưu tiên chiến lược của công ty cho mảng dữ liệu trong thời điểm hiện tại.

Phần IV – Kết luận

Case Study 06 giúp nhóm áp dụng toàn diện các kỹ thuật xử lý và phân tích dữ liệu HR bằng pandas. Đặc biệt, Nhiệm vụ 5 với pivot table đã cung cấp cái nhìn tổng quan về chất lượng ứng viên và mối quan hệ giữa kinh nghiệm và mức lương. File full_hr_data đóng vai trò là báo cáo kết quả phân tích cuối cùng của bài.

Phần V – Nhật ký đóng góp

Nguyễn Thành Văn - task1_cleaning

Đỗ Thân Quốc Khánh - task2_query

Trần Minh Quang - task3_groupby

Đào Quang Minh - task4_merge

Dương Đức Khôi - task5_pivot