[DA_Project TIMA] Task 2_Xử lý và làm sạch dữ liệu với Power BI & Python

1. Kiểm tra và xử lý giá trị thiếu

- **Power BI**: Dùng chức năng "Replace Values" trong Power Query để thay thế giá trị thiếu bằng một giá trị cụ thể hoặc một giá trị tính toán.
- **Python**: Dùng hàm fillna() để điền giá trị thiếu trong các cột.

2. Loại bỏ các bản sao

- **Power BI**: Dùng chức năng "Remove Duplicates" trong Power Query để loại bỏ các bản sao dữ liệu trong bảng.
- **Python**: Dùng hàm drop_duplicates() để loại bỏ các bản sao trong DataFrame.

3. Chuyển đổi kiểu dữ liệu

- **Power BI**: Sử dụng tính năng "Change Type" để chuyển đổi kiểu dữ liệu của các cột (ví dụ: từ chuỗi thành số hay ngày tháng).
- **Python**: Dùng hàm astype() để chuyển đổi kiểu dữ liệu của cột trong DataFrame.

4. Loại bỏ dòng có lỗi

- **Power BI**: Sử dụng "Remove Errors" trong Power Query để loại bỏ các dòng có giá trị lỗi (null hoặc sai kiểu dữ liệu).
- **Python**: Dùng hàm dropna() để loại bỏ các dòng có giá trị thiếu hoặc không hợp lệ.

5. Loại bỏ khoảng trắng thừa trong văn bản

- **Power BI**: Dùng chức năng "Trim" trong Power Query để loại bỏ các khoảng trắng thừa trong các cột chuỗi văn bản.
- **Python**: Dùng str.strip() để loại bỏ khoảng trắng ở đầu và cuối chuỗi trong DataFrame.

6. Điền giá trị thiếu

- **Power BI**: Sử dụng tính năng "Fill Down" hoặc "Fill Up" để điền các giá trị thiếu dựa trên giá trị có sẵn ở các dòng trên hoặc dưới.
- **Python**: Dùng hàm fillna() để điền giá trị thiếu bằng các giá trị có sẵn từ các dòng trên hoặc dưới.

7. Thay thế giá trị lỗi

- **Power BI**: Sử dụng "Replace Errors" trong Power Query để thay thế giá trị lỗi bằng một giá trị cụ thể.
- **Python**: Dùng hàm replace() để thay thế các giá trị lỗi trong DataFrame.

8. Thêm cột điều kiện

- **Power BI**: Dùng "Add Conditional Column" trong Power Query để tạo cột mới theo điều kiện cụ thể (ví dụ: nếu giá trị cột A lớn hơn 10, thì cột B bằng "True").
- **Python**: Tạo cột mới với np.where() hoặc apply() và một hàm điều kiện.

9. Tự động phát hiện kiểu dữ liệu

- **Power BI**: Sử dụng tính năng "Detect Data Type" trong Power Query để tự động phát hiện và điều chỉnh kiểu dữ liệu.
- **Python**: Dùng các hàm pd.to_datetime() hoặc pd.to_numeric() để chuyển đổi và chuẩn hóa kiểu dữ liệu.

10. Tạo cột sao chép

- **Power BI**: Dùng chức năng "Duplicate Column" để tạo một bản sao của cột dữ liệu ban đầu.

- **Python**: Sử dụng copy() để sao chép dữ liệu trong một cột của DataFrame.

11. Kết hợp bảng

- **Power BI**: Dùng "Merge Queries" trong Power Query để kết hợp dữ liệu từ các bảng khác nhau vào một bảng chung.
- **Python**: Sử dụng hàm merge() để kết hợp các DataFrame theo các cột chung.

12. Tách cột

- **Power BI**: Dùng "Split Column" trong Power Query để tách một cột thành nhiều cột theo một dấu phân cách (ví dụ: dấu phẩy, dấu cách).
- **Python**: Dùng str.split() để tách chuỗi trong cột thành các phần tử.

13. Nhóm dữ liệu

- **Power BI**: Sử dụng tính năng "Group By" trong Power Query để nhóm các dòng dữ liệu theo một hoặc nhiều cột.
- **Python**: Dùng hàm groupby() để nhóm dữ liệu trong DataFrame theo các cột được chỉ định.

14. Chuyển cột thành hàng

- **Power BI**: Dùng "Unpivot Columns" để chuyển các cột dữ liệu thành hàng trong Power Query.
- **Python**: Dùng hàm melt() để chuyển các cột thành hàng trong DataFrame.

15. Thêm cột tính toán

- **Power BI**: Tạo các cột tính toán mới bằng cách sử dụng "Add Custom Column" trong Power Query.
- **Python**: Dùng apply() hoặc lambda để tạo các cột tính toán trong DataFrame.

16. Loại bỏ cột không cần thiết

- **Power BI**: Dùng "Remove Columns" để loại bỏ những cột không cần thiết trong dữ liệu.
- **Python**: Dùng drop() để loại bỏ các cột không cần thiết từ DataFrame.

17. Kiểm tra dữ liệu thiếu theo từng nhóm

- **Power BI**: Dùng "Group By" và "Remove Duplicates" để kiểm tra các nhóm có dữ liệu thiếu.
- **Python**: Dùng groupby() và isnull() để kiểm tra dữ liệu thiếu theo nhóm trong DataFrame.

18. Thay đổi cấu trúc dữ liệu

- **Power BI**: Dùng "Pivot Columns" để thay đổi cấu trúc dữ liệu từ dạng hàng thành cột.
- **Python**: Sử dụng pivot_table() để chuyển dữ liệu từ dạng hàng sang cột trong DataFrame.

19. Điền giá trị trống

- **Power BI**: Dùng "Fill Down" hoặc "Fill Up" để điền giá trị trống trong các cột.
- **Python**: Dùng fillna() để điền giá trị trống vào các ô trống trong cột.

20. Kiểm tra lỗi trong dữ liệu

- **Power BI**: Kiểm tra và loại bỏ lỗi trong dữ liệu bằng "Remove Errors" trong Power Query.
- **Python**: Dùng dropna() hoặc replace() để loại bỏ hoặc thay thế các giá trị lỗi trong DataFrame.

21. Loại bỏ giá trị ngoại lai

- **Power BI**: Dùng "Remove Outliers" để loại bỏ các giá trị ngoại lai trong dữ liệu.
- **Python**: Dùng clip() để giới hạn các giá trị trong phạm vi xác định.

22. Chuẩn hóa dữ liệu

- **Power BI**: Sử dụng tính năng "Normalize" để chuẩn hóa các giá trị trong dữ liệu.

- **Python**: Dùng StandardScaler hoặc MinMaxScaler từ thư viện sklearn.preprocessing để chuẩn hóa dữ liệu.

23. Chuyển dữ liệu thành dạng đồng nhất

- **Power BI**: Dùng "Unpivot" để chuyển tất cả dữ liệu thành một định dạng chuẩn.
- **Python**: Dùng melt() hoặc stack() để chuyển các dữ liệu thành dạng chuẩn hóa.

24. Xử lý các giá trị bất thường

- **Power BI**: Dùng "Replace Values" để thay thế các giá trị bất thường bằng giá trị hợp lý.
- **Python**: Dùng replace() hoặc clip() để thay thế hoặc loại bỏ các giá trị bất thường.

25. Làm sạch dữ liệu không hợp lệ

- **Power BI**: Dùng "Remove Errors" và "Replace Values" để làm sạch các dữ liệu không hợp lệ.
- **Python**: Dùng dropna() và replace() để loại bỏ hoặc thay thế dữ liệu không hợp lệ.

26. Kiểm tra tính duy nhất của dữ liệu

- **Power BI**: Dùng "Remove Duplicates" để kiểm tra và loại bỏ các bản sao.
- **Python**: Dùng drop_duplicates() để kiểm tra và loại bỏ các bản sao trong DataFrame.

27. Điều chỉnh độ chính xác của dữ liệu số

- **Power BI**: Sử dụng "Round" trong Power Query để điều chỉnh độ chính xác của dữ liệu số.
- **Python**: Dùng round() để làm tròn các giá trị số trong DataFrame.

28. Chuyển đổi thời gian

- **Power BI**: Dùng "Date/Time" để chuyển đổi dữ liệu thành kiểu ngày/tháng/năm.
- **Python**: Dùng pd.to_datetime() để chuyển đổi cột thành kiểu datetime.

29. Xử lý dữ liệu không hợp lệ theo dạng chuỗi

- **Power BI**: Dùng "Replace Values" để thay thế các giá trị chuỗi không hợp lệ.
- **Python**: Dùng str.replace() để thay thế các chuỗi không hợp lệ trong cột.

30. Xử lý dữ liệu số không hợp lệ

- **Power BI**: Dùng "Replace Values" để thay thế các giá trị không hợp lệ trong cột số.

Python: Dùng replace() hoặc clip() để xử lý các giá trị số không hợp lệ.