Họ tên: Nguyễn Xuân Thuận

ĐỀ BÀI:

Phần I: Xử lý dữ liệu

1. Dữ liệu mẫu cung cấp có vấn đề gì không? Nếu có hãy mô tả lại.

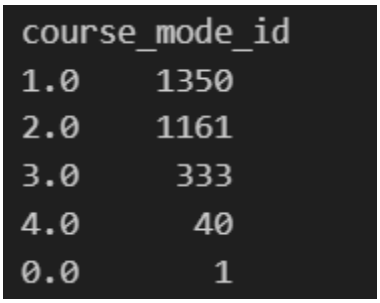
Lời giải:

- Dữ liệu cung cấp gồm 3 file excel là:

- Location: File này chứa danh sách các địa điểm, thường được sử dụng để xác định vị trí của khách hàng hoặc nơi tổ chức khóa học.
- Mode: File này chứa danh sách các hình thức học tập của khóa học, giúp phân loại khóa học theo cách thức tổ chức (trực tiếp, trực tuyến, hoặc kết hợp).
- Order: File này có số bản ghi là 4304, gồm 18 thuộc tính. Nó chứa thông tin chi tiết về các đơn hàng hoặc đăng ký khóa học, bao gồm thông tin khách hàng, thanh toán, và liên kết với các file location và mode. Order liên kết với file **location** thông qua trường **cus_location_id** và liên kết với file **mode** thông qua trường **course_mode_id**. Dữ liệu ở file location và mode thì không có vấn đề gì.

- Còn ở file order có những vấn đề sau:

- Thuộc tính **course_mode_id** đang có những giá trị như **0, 1, 2, 3, 4, null**. Ở bảng mode thì chỉ có **1, 2, 3, 4** điều này cho thấy dữ liệu đang gặp một số vấn đề về tính toàn vẹn và nhất quán. Lỗi này xảy ra có thể do dữ liệu nhập thủ công nên bị nhập sai, khóa học không được gán chế độ (mode) cụ thể, lỗi trong quá trình nhập liệu hoặc tích hợp dữ liệu.



course_mode_id	
1.0	1350
2.0	1161
3.0	333
4.0	40
0.0	1

Đây là số lượng của giá trị khác null. Có thể thấy được có 1 giá trị 0 duy nhất. Ta có thể xử lý theo cách là loại bỏ bản ghi này.

Số bản ghi bị null thuộc tính **course_mode_id** là 1419. Với số lượng bản ghi lớn thế này thì việc loại bỏ có thể gây ảnh hưởng đến kết quả phân tích. Ta sẽ tạm thời giữ lại để kiểm tra thêm các thuộc tính khác sau đó sẽ đưa ra cách giải quyết.

- Thuộc tính **course_schedule** cũng có vấn đề. Đây là thuộc tính chỉ lịch học của khóa học. Do chưa có description data nên theo suy đoán thì giá trị này gồm 6 số gồm 2 số đầu là

tháng và 4 số cuối là năm. Thuộc tính này có 269 bản ghi bị null, có 13 bản ghi gồm 2 giá trị (VD: 052023,062023 hay 112022,072023 …), có 44 bản ghi là số 999999. Đây là những trường hợp có sự bất thường trong trường dữ liệu **course_schedule.**

- Các thuộc tính **payment_amount, payment_status, total_fee, actually_received, original_fee** có sự liên kết với nhau như bên dưới:

Mối quan hệ giữa các thuộc tính:

- `original_fee` là giá gốc → `total_fee` là giá sau chiết khấu.
- `payment_amount` là số tiền khách hàng trả trong một lần thanh toán.
- `actually_received` là số tiền công ty thực sự nhận được sau khi trừ các khoản phí.
- `payment_status` cho biết trạng thái của quá trình thanh toán.

- Giá trị của payment_amount là từ 0 đến 68000000. Có 39 bản ghi với giá trị là 0, có 7 bản ghi với giá trị là 1000. Trong số 39 bản ghi có giá trị 0 thì có có những bản ghi bị null ở thuộc tính **payment_status, total_fee, actually_received**. Những thuộc tính này có giá trị null, bản ghi sẽ không có ý nghĩa hoặc gây lỗi khi xử lý dữ liệu. Với 7 bản ghi có **payment_amount** là 1000 thì cũng tương tự như trên bị null thuộc tính quan trọng và có 1 bản ghi với **original_fee** là 60000000 nhưng **total_fee** chỉ 1000. Đây là điểm bất thường trong dữ liệu. Còn với những bản ghi còn lại thì chưa phát hiện bất thường.
  Thuộc tính **lead_id** và **original_fee** bị null khá nhiều. Thuộc tính **lead_id** với 1467 bản ghi bị null.
  Thuộc tính **original_fee** với 1861 số bản ghi bị null.
  Thuộc tính **cus_location_id** có 13 bản ghi bị null. Các giá trị còn lại hợp lệ.

2. Trình bày quá trình xử lý dữ liệu của bạn để chuẩn bị cho phần phân tích.

Các bước để tiền xử lý dữ liệu:

- Xử lý các bản ghi trùng lặp.
- Xử lý kiểu dữ liệu (Data type).
- Xử lý dữ liệu thiếu.

Đầu tiên là đi kiểm tra xem các bản ghi có bị trùng lặp hay không. Sử dụng python để kiểm tra. Đọc dữ liệu bằng pandas. Sau đó kiểm tra trùng lặp theo id có kết quả như dưới:

```
duplicate_ids = df[df.duplicated(subset=['id'])]
print(f"Số bản ghi trùng ID: {duplicate_ids.shape[0]}")
✓ 0.0s
Số bản ghi trùng ID: 0
```

Có thể thấy dữ liệu không bị trùng lặp.

Bước tiếp theo là biến đổi kiểu dữ liệu về đúng định dạng của nó. Vì bước trực quan sử dụng PowerBI để trực quan nên sẽ sử dụng Power Query để biến đổi dữ liệu. Với thuộc tính

- id, course_mode_id, cus_pic_id, lead_id, cus_location_id: Whole Number (số nguyên).
- created, payment_date: Date/Time (Ngày và giờ).
- payment_amount, total_fee, actually_received, original_fee: Decimal Number (Số thực)
- cus_mail, cus_name, cus_mobi, course_schedule, payment_status, lead_source, payment_method: Text (Chuỗi).

Bước tiếp theo sẽ đi xử lý dữ liệu bị thiếu. Ở trên đã trình bày những dữ liệu bị thiếu nhưng em sẽ đi xử lý dữ liệu bị thiếu ở những cột quan trọng như **payment_amount, payment_status, total_fee, actually_received** bằng cách loại bỏ bản ghi.

Với cột **course_mode_id** loại bỏ những giá trị là 0. Tạm thời giữ lại giá trị null để phân tích.

Với **payment_status** loại bỏ những bản ghi bị null.

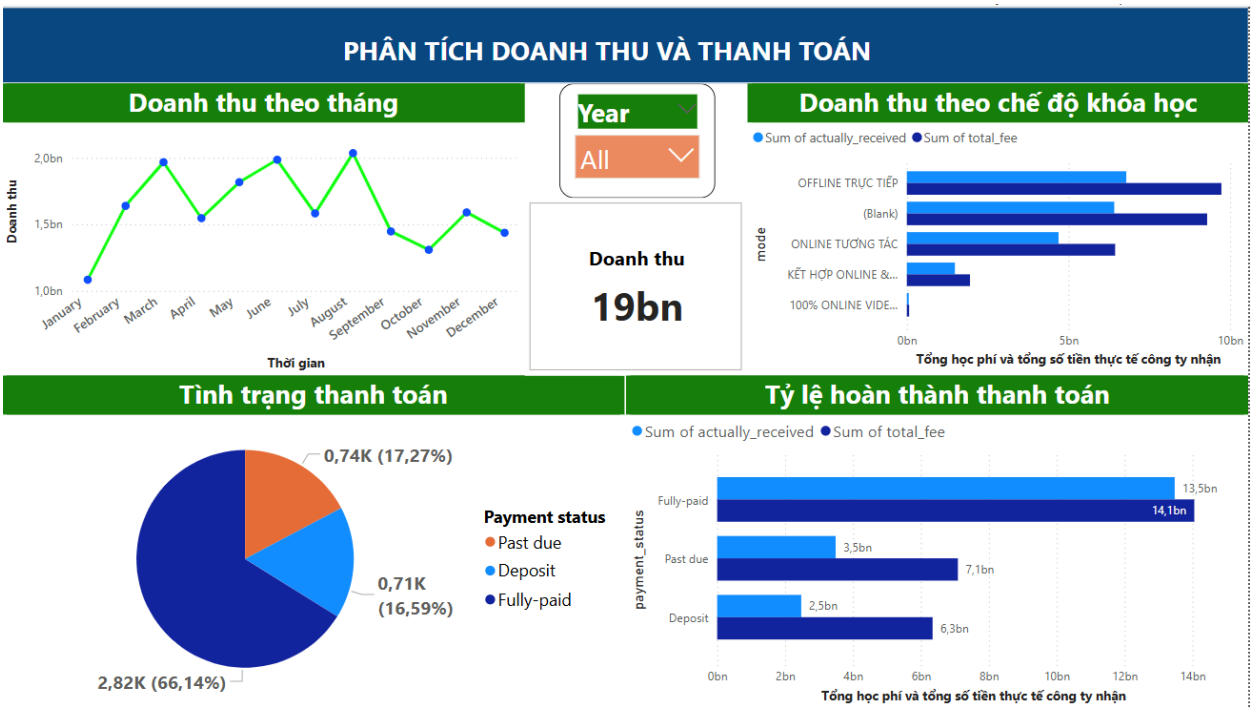
Loại bỏ những bản ghi null của thuộc tính **actually_received.**

Phần II: Phân tích dữ liệu

1. **Dựa vào các trường dữ liệu đã cho, hãy xây dựng dashboard và đưa ra 1 số nhận xét, phân tích về tình hình kinh doanh của công ty.**

   Để phân tích về tình hình kinh doanh của công ty. Em sẽ đi phân tích 2 vấn đề chính là phân tích doanh thu, thanh toán và phân tích khách hang.

Dưới đây là dashboard xây dựng để phân tích doanh thu và thanh toán:



Biểu đồ doanh thu theo tháng giúp công ty theo dõi xu hướng doanh thu theo các tháng trong năm. Có thể thấy được doanh thu ở các tháng 3, 6, 9 là những tháng có doanh thu cao. Có thể thấy được chu kỳ thường là 3 tháng thì doanh thu sẽ đạt đỉnh điểm. Tháng 12 là tháng cuối năm, đây là thời điểm mọi người dành thời gian cho việc khác nên doanh thu ở tháng này chưa được cao.

Biểu đồ tình trạng thanh toán giúp công ty theo dõi chất lượng khóa học. Khi trạng thái thanh toán là Fully-paid cao nghĩa là khách hàng hài lòng với dịch vụ của công ty. Với tỷ lệ 66,14% cho thấy chất lượng khóa học cũng khá tốt. Duy trì chất lượng dịch vụ để giữ chân khách hàng. Có thể khuyến khích khách hàng quay lại bằng các chương trình ưu đãi hoặc giới thiệu bạn bè.

Tỷ lệ đặt cọc khoảng 16,59% cho thấy một phần khách hàng đang trong quá trình thanh toán nhưng chưa hoàn tất. Cần theo dõi và hỗ trợ khách hàng để họ hoàn tất thanh toán. Công ty có thể gửi thông báo nhắc nhở khách hàng hoàn tất thanh toán hoặc cung cấp các phương thức thanh toán linh hoạt (ví dụ: trả góp, giảm giá nếu thanh toán sớm).

Có khoảng 17,27% khách hàng không thanh toán đúng hạn. Khách hàng đã không thanh toán đúng hạn, có thể do quên, gặp vấn đề tài chính, hoặc không hài lòng với dịch vụ. Tỷ lệ này cho thấy cần có biện pháp nhắc nhở hoặc hỗ trợ khách hàng để họ hoàn tất thanh toán (ví dụ: gia hạn thời gian thanh toán, giảm phí trễ hạn). Phân tích nhóm khách hàng quá hạn để tìm ra đặc điểm chung (ví dụ: cùng một nguồn khách hàng, cùng một loại khóa học).

Biểu đồ tỷ lệ hoàn thành thanh toán giúp công ty theo dõi được tỷ lệ khách hàng thanh toán với % bao nhiêu ở từng trạng thái thanh toán. Với trường hợp thanh toán đủ (đáng ra dữ liệu **total_fee** và **actually_received** phải bằng nhau). Có thể do một vài yếu tố nào đó như giảm giá nên 2 cột này chưa bằng nhau. Dưới đây là những bản ghi đã lọc ra với **total_fee** và **actually_received** khác nhau (18 bản ghi):

course_mode_id	course_schedule	payment_amount	payment_status	cus_pic_id	lead_source	payment_date	payment_method	total_fee	lead_id	actually_received
2.0	012021	1850000	Fully-paid	919	GL	2022-01-17	FB	3700000.0	NaN	1850000.0
2.0	012021	1850000	Fully-paid	55	GL	2022-01-17	FB	3700000.0	NaN	1850000.0
3.0	022021	3625000	Fully-paid	1593	FB	2022-01-18	FB	7250000.0	NaN	3625000.0
3.0	022021	3625000	Fully-paid	919	FB	2022-01-19	FB	7250000.0	NaN	3625000.0
3.0	022021	4000000	Fully-paid	919	OT	2022-01-21	TL	8000000.0	NaN	4000000.0
...	...	...	...	...	...	...	...	...	...	...
2.0	012024	2800000	Fully-paid	13256	UP	2023-12-23	TL	2800000.0	34698.0	2738400.0
4.0	012024	2000000	Fully-paid	13256	FB	2023-12-23	TL	2000000.0	34983.0	1956000.0
2.0	012024	3220000	Fully-paid	10421	TDF	2023-12-27	TL	3220000.0	35088.0	3149160.0
2.0	012024	2700000	Fully-paid	1659	TDF	2023-12-28	TL	2700000.0	35127.0	2640600.0
2.0	012024	9000000	Fully-paid	15870	TDF	2023-12-29	TL	9000000.0	35102.0	8206000.0

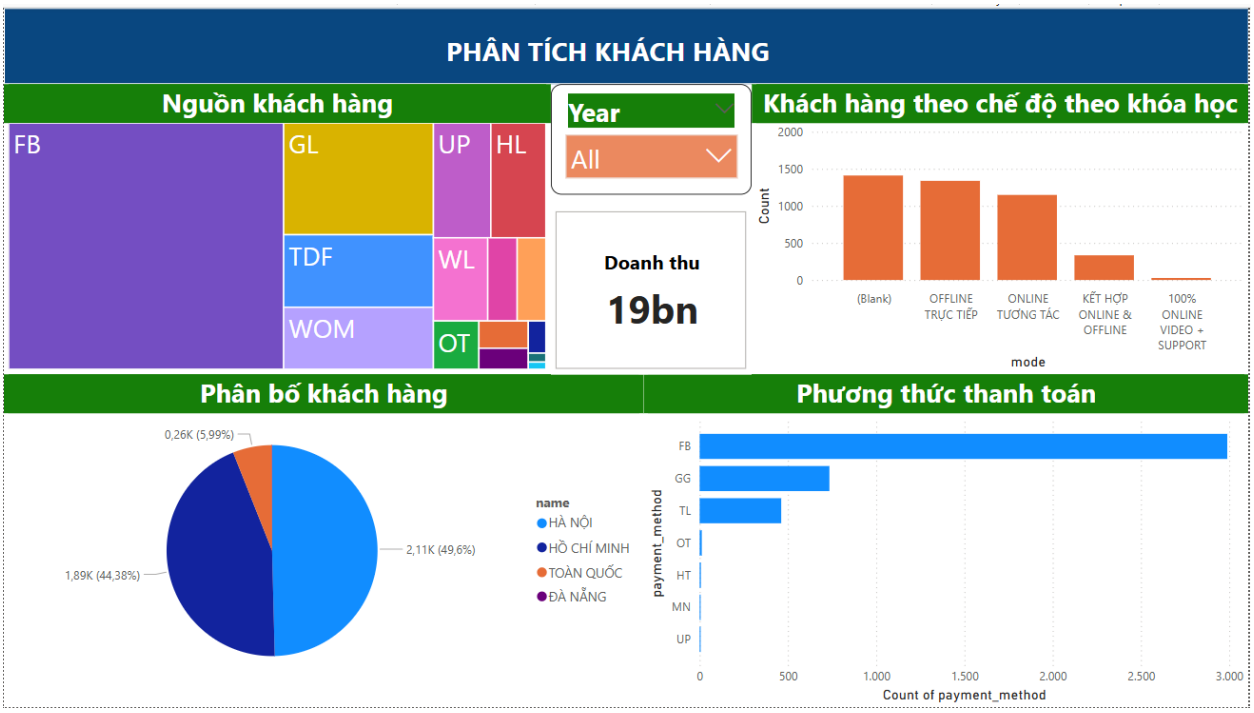
Biểu đồ này giúp theo dõi được có thể xảy ra lỗi nếu như trạng thái thanh toán là **fully-paid** mà tổng **total_fee** khác tổng **actually_received**.

Với trạng thái thanh toán là **past due** ở biểu đồ này giúp ta theo dõi được tỷ lệ những khách hàng thanh toán quá hạn đã thanh toán được khoảng bao nhiêu % với số tiền của khóa học. Có thể quan sát được khách hàng đã thanh toán khoảng 50% số tiền của khóa học. Với một lý do nào đó mà khách hàng đã không thanh toán và để quá hạn. Có thể do chất lượng khóa học chưa tốt hoặc lý do nào đó. Đây cũng là một cách để công ty theo dõi để kịp thời tìm ra nguyên nhân và tìm phương án xử lý.

Với trạng thái thanh toán là **deposit** giúp công ty quan sát được tỷ lệ số tiền khách hàng đặt cọc. Theo như quan sát thì thấy được khách hàng cọc khoảng 40% số tiền của khóa học. Đây là một dấu hiệu tốt. Có thể thấy được khách hàng khá tin tưởng vào khóa học của công ty. Nếu như tỷ lệ đặt cọc thấp thì sẽ là một dấu hiệu để biết đang có vấn đề để công ty tìm và giải quyết. Công ty cũng có thể dựa vào tỷ lệ này để gửi thông báo cho khách hàng để khách hàng hoàn thành thanh toán. Công ty cũng có thể đưa ra một số chiến lược thúc đẩy khách hàng thanh toán nếu như thấy tỷ lệ này thấp.

Biểu đồ doanh thu theo chế độ khóa học cho biết tỷ lệ doanh thu ở những chế độ nào cao. Quan sát biểu đồ có thể thấy được với chế độ offline trực tiếp và online tương tác đang cho ra doanh thu cao nhất. Công ty có thể dựa vào biểu đồ này để đưa ra những phương án phát triển theo 2 chế độ trên hoặc tìm ra nguyên nhân khiến 2 chế độ còn lại đang chưa làm tốt.

Dưới đây là dashboard phân tích khách hàng:



Biểu đồ nguồn khách hàng giúp công ty theo dõi được kênh marketing nào đang hiệu quả để công ty phát triển. Quan sát biểu đồ có thể thấy được nguồn khách hàng từ FB, GL, TDF, WOM đang làm tốt việc marketing. FB là kênh chính để công ty tìm được khách hàng. Qua đây công ty có thể điều chỉnh chiến lược marketing để không tốn tiền vào những kênh không mang lại hiệu quả cao.

Biểu đồ phân bố khách hàng giúp công ty xem được khu vực nào đang phát triển mạnh để có thể đầu tư tìm ra được nhiều khách hàng hơn. Có thể thấy Hà Nội và HCM đang là 2 nơi thu hút lượng khách hàng chính cho công ty.

Biểu đồ khách hàng theo chế độ khóa học giúp công ty quan sát được số lượng khách hàng theo chế độ học nào nhiều.

Biểu đồ phương thức thanh toán góp phần đánh giá được phương thức thanh toán nào đang được khách hàng ưu tiên sử dụng. Biểu đồ giúp công ty hiểu rõ hơn về hành vi của khách hàng và tối ưu hóa quy trình thanh toán. Có thể thấy FB, GG và TL đang được khách hàng chủ yếu sử dụng để thanh toán. Công ty sẽ biết được phương thức thanh toán nào được ưa chuộng nhất, từ đó điều chỉnh chiến lược tiếp thị và hỗ trợ khách hàng. Khuyến khích khách hàng sử dụng các phương thức này bằng cách giảm phí hoặc cung cấp ưu đãi.

2. Sử dụng mô hình RFM để phân khúc khách hàng và đưa ra một số gợi ý về phương hướng tiếp cận đố với từng phân khúc.

Mô hình **RFM (Recency - Frequency - Monetary)** giúp phân khúc khách hàng dựa trên:

- **Recency (R)** – Khoảng thời gian kể từ lần mua hàng gần nhất.
- **Frequency (F)** – Số lần mua hàng trong một khoảng thời gian.
- **Monetary (M)** – Tổng số tiền khách hàng đã chi tiêu.

Dựa vào 3 chỉ số này, khách hàng được chia thành các nhóm khác nhau để có chiến lược tiếp cận phù hợp.

Nhóm khách hàng	Điểm RFM	Chiến lược tiếp cận
VIP (Trung thành & có giá trị cao)	333, 323, 332	Ưu đãi đặc biệt, chương trình khách hàng thân thiết, tri ân.
Khách hàng tiềm năng	321, 322, 312	Giảm giá để khuyến khích mua hàng thường xuyên hơn.
Khách hàng mới	313, 311	Chăm sóc tốt, giới thiệu sản phẩm/dịch vụ.
Khách hàng ngủ quên	211, 221, 222	Gửi email/sms khuyến mãi, nhắc nhở quay lại mua hàng.
Khách hàng cần chăm sóc đặc biệt	111, 121	Tìm hiểu lý do rời bỏ, ưu đãi hấp dẫn để giữ chân.