



BÁO CÁO DỰ ÁN 1

Chuyên Ngành Xử Lý Dữ Liệu

Phân tích xu hướng và hành vi người tiêu dùng trên trang thương mại điện tử Olist tại Brazil

Lớp: DP19302

Nhóm: DP04

Các thành viên:

1. Vũ Hồng Vị - PH55312 (Nhóm trưởng)
2. Trịnh Thùy Linh - PH56123
3. Lê Thị Hồng Vân - PH55526
4. Phạm Ánh Hồng - PH55100
5. Phạm Văn Sâm - PH53301

GVHD: Chu Thị Ngân

MỤC LỤC

1 giới thiệu dự án	9
1.1 giới thiệu	9
1.2 yêu cầu của công ty	10
1.3 lập kế hoạch dự án	14
2 phân tích yêu cầu khách hàng	16
2.1 phân tích yêu cầu	16
2.2 câu chuyện dữ liệu	17
2.2.1 đặt vấn đề	17
2.2.2 xác định câu chuyện	18
2.2.3 xác định rõ đối tượng	18
2.2.4 xác định câu chuyện chi tiết	20
2.2.5 trình bày dữ liệu	20
2.2.6 những điều cần lưu ý	22
2.3 kiến trúc hệ thống	22
2.3.1 kiến trúc	22
2.3.2 giải thích	23
2.4 giải thích về bộ dữ liệu khách hàng	25
2.4.1 các khái niệm	25
2.4.2 các trường dữ liệu	25
3 làm sạch và chuyên đổi dữ liệu	31
3.1 chuẩn bị dữ liệu	31
3.1.1 giải pháp lưu trữ dữ liệu	31
3.1.2 giải pháp phân bố dữ liệu	33
3.2 làm sạch dữ liệu	38
3.2.1 các vấn đề ảnh hưởng tới dữ liệu	38
3.2.2 các tiêu chí đánh giá chất lượng dữ liệu	39
3.2.3 các bước làm sạch dữ liệu	41
3.3 chuyển đổi dữ liệu	54
3.3.1 các trường hợp cần chuyển đổi	54
3.3.2 các kỹ thuật chuyển đổi	56
3.3.3 trình bày các phép chuyển đổi trong dự án	58
4 xử lý dữ liệu	58
4.1 chuẩn hóa dữ liệu	58
4.1.1 trình bày các bước chuẩn hóa trong dự án	58
4.2 mô hình hóa dữ liệu	59
4.2.1 các loại mô hình hóa	59
4.2.2 các tiêu chí đánh giá mô hình dữ liệu	60
4.2.3 trình bày các bước mô hình hóa	61
4.2.4 trình bày các bước tạo bảng dữ liệu	65
4.3 xử lý dữ liệu dax	70

4.3.1 measure	70
4.3.2 calculated column	72
4.3.3 filter	78
5 trực quan hóa dữ liệu	79
5.1 các kỹ thuật trực quan hóa	79
5.2 các nguyên tắc trực quan hóa	80
5.3 trình bày cách thêm visual mới	82
5.4 trình bày tạo các report cho dự án	83
5.4.1 tạo visual thống kê chi tiết	83
5.4.2 tạo visual thống kê tổng thể	104
6 xây dựng báo cáo	107
6.1 dashboard và report	107
6.2 xây dựng báo cáo	108
6.2.1 dashboard và report	108
6.2.2 dashboard	108
6.2.3 bookmark	119
7 Dự báo, dự đoán	121
8 kết luận	125
8.1 báo cáo	125
8.1.1 các bước viết báo cáo	125
8.1.2 tổng hợp	127
8.2 khó khăn	128
8.3 thuận lợi	128
8.4 hướng phát triển	128
9 tổng kết	131

MỤC LỤC BẢNG

Bảng 1.1: Kế hoạch chi tiết	15
Bảng 2.1: Trình bày dữ liệu	21
Bảng 2.2: Bảng Customers	25
Bảng 2.3: Bảng Orders	26
Bảng 2.4: Bảng Order Items	26
Bảng 2.5: Bảng Sellers	27
Bảng 2.6: Bảng Products	27
Bảng 2.7: Bảng Geolocation	28
Bảng 2.8: Product Category Name Translation	28
Bảng 2.9: Bảng Order Payments	28
Bảng 2.10: Bảng Order Reviews	29

Bảng 2.11: Bảng Marketing Qualified Leads	29
Bảng 2.12: Bảng Closed Deals	30

MỤC LỤC HÌNH ẢNH

Hình 3.1 Tạo cơ sở dữ liệu phân phối	35
Hình 3.2 Thiết lập máy chủ phân phối	35
Hình 3.3 Cấu hình máy chủ xuất bản	35
Hình 3.4 Tạo án phám	35
Hình 3.5 Thêm các bài viết vào án phám	36
Hình 3.6 Tạo đăng ký	36
Hình 3.7: Dữ liệu các bảng	43
Hình 3.8: Kiểm tra sự không nhất quán trong tên thành phố theo vị trí địa lý	43
Hình 3.9: Thay thế các giá trị bị lỗi	44
Hình 3.10: Cập nhật dữ liệu sau khi bỏ dấu	44
Hình 3.11: Loại bỏ số ở giữa các thành phố	44
Hình 3.12: Xóa các ký tự và khoảng trắng thừa	45
Hình 3.13: Xóa phần . . . ở đầu	45
Hình 3.14: Xóa * ở đầu	45
Hình 3.15: Xóa phần 40. ở đầu	45
Hình 3.16: Chỉ trích xuất một lần xuất hiện rio de janeiro	46
Hình 3.17: Xóa bỏ chữ z ở cuối	46
Hình 3.18: Xóa các ký hiệu mã hóa %26apos%3B và %26 trong cột geography_city	46
Hình 3.19: Xóa khoảng trắng trong cột geography_city bằng hàm REPLACE	46
Hình 3.20: Viết hoa chữ cái đầu tiên của từ trong cột geography_city	47
Hình 3.21: Đổi tên geography_state thành tên chưa viết tắt	47
Hình 3.22: Kiểm tra xem có sự không nhất quán trong customer_city không	47
Hình 3.23: Thay thế customer_state viết tắt thành tên đầy đủ	48
Hình 3.24: Thay đổi customer_city thành một trường hợp thích hợp	48
Hình 3.25: Làm tròn price và freight_value lên 2 chữ số thập phân	48
Hình 3.26: freight_value:	49
Hình 3.27: Tạo cột mới	49

Hình 3.28: Đienia giá trị cho các cột mới	49
Hình 3.29: Xóa cột gốc	50
Hình 3.30: Chuyển đổi payment_value thành 2 chữ số thập phân	50
Hình 3.31: Đổi tên cột tên nhóm sản phẩm (tiếng Tây Ban Nha, tiếng Anh)	50
Hình 3.32: Xóa tiêu đề cột khỏi hàng đầu tiên	51
Hình 3.33: Thay thế dấu gạch dưới bằng dấu cách	51
Hình 3.34: Kiểm tra xem đã thay đổi đúng chưa	51
Hình 3.35: Join 2 bảng để lấy ra tên tiếng Anh	52
Hình 3.36: Viết hoa product_category_eng_name	52
Hình 3.37: Viết hoa seller_city	52
Hình 3.38: Đổi thành tên đầy đủ	53
Hình 3.39: Bảng customer trước làm sạch	53
Hình 3.40: Bảng customer sau làm sạch	53
Hình 3.41: Bảng geolocation trước làm sạch	53
Hình 3.42: Bảng geolocation sau làm sạch	54
Hình 3.43: Bảng order_items trước làm sạch	54
Hình 3.44: Bảng order_items sau làm sạch	54
Hình 3.45: Bảng order_reviews trước làm sạch	55
Hình 3.46: Bảng order_reviews sau làm sạch	55
Hình 3.47: Bảng order_payment trước làm sạch	55
Hình 3.48: Bảng order_payment sau làm sạch	55
Hình 3.49: Bảng product_category trước làm sạch	55
Hình 3.50: Bảng product_category sau làm sạch	55
Hình 3.51: Kiểm tra sự nhất quán trong tên thành phố theo vị trí địa lý	59
Hình 4.1: Mối quan hệ	64
Hình 4.2: Tạo database mới	66
Hình 4.3: Bảng product_category_name_translation	66
Hình 4.4: Bảng sellers	67
Hình 4.5: Bảng customers	67
Hình 4.6: Bảng geolocation	67
Hình 4.7: Bảng order_items	68
Hình 4.8: Bảng order_payments	68

Hình 4.9: Bảng order_reviews	69
Hình 4.10: Bảng orders	69
Hình 4.11: Bảng products	69
Hình 4.12: Bảng leads_qualified	69
Hình 4.13: Bảng leads_closed	70
Hình 4.14: Danh sách các measure được sử dụng trong báo cáo	73
Hình 4.15: Measure tỷ lệ khách hàng quay lại	73
Hình 4.16: Measure tính số lượng khách hàng tiềm năng	74
Hình 4.17: Công thức tính tỷ lệ đánh giá 5 sao	74
Hình 4.18: Công thức xếp hạng thành phố	74
Hình 4.19: Measure xếp hạng danh mục sản phẩm theo tổng giá trị	75
Hình 4.20: Measure tạo tiêu đề động cho biểu đồ top N sản phẩm bán chạy	75
Hình 4.21: Measure tạo tiêu đề động cho biểu đồ top N thành phố bán chạy	75
Hình 4.22: Bảng Customers	75
Hình 4.23: Bảng linear_data	76
Hình 4.24: Bảng order_items	76
Hình 4.25: Tạo cột tính tổng giá trị đơn hàng	76
Hình 4.26: Tạo cột kiểm tra giao hàng trễ của đơn vị vận chuyển	77
Hình 4.27: Tạo cột kết hợp ngày giờ giới hạn giao hàng	77
Hình 4.28: Tạo cột tính độ trễ giao hàng theo giờ	77
Hình 4.29: order_reviews	77
Hình 4.30: Tạo cột kiểm tra sự tồn tại của bình luận	77
Hình 4.31: Tạo cột tính thời gian hoàn thành đánh giá (theo giờ)	77
Hình 4.32: Tạo cột tính thời gian hoàn thành đánh giá (theo ngày)	77
Hình 4.33: Tạo cột tính thời gian hoàn thành đánh giá sau khi giao hàng	78
Hình 4.34: Tạo cột tính thời gian hoàn thành đánh giá trước khi giao hàng	78
Hình 4.35: Tạo cột kiểm tra thời điểm tạo đánh giá	78
Hình 4.36: Tạo cột phân tích số lượng khách hàng dựa trên điểm đánh giá	78
Hình 3.37: orders	78
Hình 4.38: Tạo cột tính thời gian giao hàng thực tế (theo ngày)	79
Hình 4.39: Tạo cột tính thời gian giao hàng (theo giờ)	79
Hình 4.40: Tạo cột tính thời gian vận chuyển (theo ngày)	79

Hình 4.41: Theo dõi thời gian từ đơn vị vận chuyển đến khách hàng	79
Hình 4.42: Tính toán thời gian giao hàng dự kiến	79
Hình 4.43: Đánh giá hiệu suất giao hàng	79
Hình 4.44: Xác định ngày mua trong tuần	79
Hình 4.45: Chuyển đổi ngày trong tuần	79
Hình 4.46: Bảng product_category_name	80
Hình 4.47: Phân tích doanh thu theo danh mục sản phẩm	80
Hình 4.48: Điểm đánh giá trung bình theo danh mục	81
Hình 4.49: Đếm số lượng đơn hàng theo danh mục sản phẩm	81
Hình 4.50: Bộ lọc chọn năm, bang, thành phố	81
Hình 4.51: Bộ lọc chọn ngày	81
Hình 5.1: Get more visual	85
Hình 5.2: Tìm visual mới cần thêm	86
Hình 5.3: Add visual	86
Hình 5.4: Hoàn thành	86
Hình 5.5: Card tổng khách hàng, doanh thu, tỷ lệ khách hàng quay trở lại	87
Hình 5.7: Biểu đồ phân bố khách hàng theo địa lý	88
Hình 5.8: Biểu đồ số lượng khách hàng theo thời gian	88
Hình 5.9: Biểu đồ top khách hàng theo doanh thu và số lượng khách hàng theo bang	89
Hình 5.10: Card thống kê về sản phẩm	89
Hình 5.11: Biểu đồ giá 5 loại sản phẩm (có giá cao nhất) theo thời gian	90
Hình 5.12: Biểu đồ số lượng 5 loại sản phẩm (có số lượng cao nhất) theo thời gian	90
Hình 5.13: Biểu đồ điểm đánh giá top 5 loại sản phẩm được yêu thích nhất	91
Hình 5.14: Biểu đồ trung bình giá của từng loại sản phẩm (top 5)	91
Hình 5.15: Top 3 sản phẩm bán chạy nhất (theo số lượng và doanh thu)	92
Hình 5.16: Card thống kê hành vi khách hàng	92
Hình 5.17: Biểu đồ số lượng đơn hàng theo thời gian (trong ngày)	93
Hình 5.18: Biểu đồ số lượng đơn hàng theo thời gian (ngày trong tuần)	93
Hình 5.19: Biểu đồ khách hàng trả góp theo số lần trả góp	94
Hình 5.20: Biểu đồ số lượng đơn hàng theo thời gian (ngày trong tháng)	94
Hình 5.21: Biểu đồ tỷ lệ hình thức thanh toán	95
Hình 5.22: Biểu đồ số lượng đơn hàng hủy, nhận xét của khách hàng	95

Hình 5.24: Biểu đồ số lượng real customer theo nguồn tiếp cận và loại hình kinh doanh	96
Hình 5.25: Biểu đồ thể hiện tỷ lệ khách hàng thực tế theo ngành hàng	96
Hình 5.26: Biểu đồ tỷ lệ khách hàng theo nguồn tiếp cận	97
Hình 5.27: Biểu đồ xu hướng số lượng real customer theo thời gian	97
Hình 5.28: Biểu đồ tỷ lệ real customers có hay không có công ty	98
Hình 5.29: Slicer theo năm, bang, thành phố	98
Hình 5.30: Card thống kê số giờ (ngày) trung bình delay	98
Hình 5.31: Biểu đồ tỷ lệ đơn hàng theo trạng thái đơn hàng	99
Hình 5.32: Biểu đồ gauge thời gian giao hàng	99
Hình 5.33: Biểu đồ tỷ lệ giao hàng trễ	100
Hình 5.34: Biểu đồ phân tích số lượng đơn hàng giao trễ theo thời gian	100
Hình 5.37: Bảng thông tin chi tiết khách hàng	102
Hình 5.38: Bảng thông tin chi tiết nhà phân phối	102
Hình 5.39: Biểu đồ tỷ lệ khách hàng sau phân loại	103
Hình 5.53: Trực quan tổng quan đánh giá của khách hàng	110
Hình 6.5: Phân tích chuyển đổi khách hàng tiềm năng	116
Hình 7.1: Xác định 10 đặc trưng có mối tương quan cao với sự hài lòng	124
Hình 7.2: Kết quả	124
Hình 7.3: Kiểu dữ liệu các đặc trưng có tương quan cao	125
Hình 7.4: Chọn 4 đặc trưng và import thư viện	125
Hình 7.5: Chuẩn bị dữ liệu	125
Hình 7.6: Khởi tạo các mô hình	126
Hình 7.7: Đánh giá từng mô hình	126
Hình 7.8: Chọn mô hình phù hợp nhất, khởi tạo mô hình	126
Hình 7.9: Huấn luyện và tìm tham số tốt nhất	126
Hình 7.10: Xây dựng pipeline với mô hình tốt nhất và huấn luyện với tập dữ liệu	127
Hình 7.11: Lưu và load pipeline	127
Hình 7.12: Xây dựng ứng dụng dự đoán sự hài lòng của khách hàng bằng Streamlit	128
Hình 7.13: Xây dựng mô hình dự đoán sự hài lòng trên jupyter notebook	128

1 GIỚI THIỆU DỰ ÁN

1.1 GIỚI THIỆU

Bộ dữ liệu này được cung cấp bởi Olist, cửa hàng bách hóa trực tuyến lớn nhất tại Brazil. Với sứ mệnh kết nối các doanh nghiệp nhỏ trên toàn quốc với các kênh bán hàng trực tuyến, Olist tạo điều kiện cho các doanh nghiệp tiếp cận thị trường thương mại điện tử mà không cần phải ký kết nhiều hợp đồng phức tạp với các nền tảng bán hàng khác nhau. Chỉ với một hợp đồng duy nhất, các doanh nghiệp có thể bán sản phẩm của mình trên Cửa hàng Olist, nơi mọi khía cạnh từ tiếp thị, xử lý đơn hàng, đến vận chuyển đều được Olist quản lý. Các sản phẩm sau đó sẽ được vận chuyển trực tiếp đến tay khách hàng thông qua mạng lưới đối tác hậu cần rộng lớn của Olist. Chi tiết hơn về dịch vụ có thể tham khảo tại www.olist.com.

Trong bối cảnh thị trường thương mại điện tử phát triển mạnh mẽ, việc thu thập hiểu khách hàng, tối ưu hóa quy trình bán hàng và nâng cao trải nghiệm người dùng trở thành yếu tố quyết định giúp các doanh nghiệp cạnh tranh và phát triển bền vững. Olist nhận thấy vai trò quan trọng của dữ liệu trong việc cung cấp những góc nhìn chuyên sâu, giúp đưa ra các quyết định chiến lược tối ưu. Do đó, công ty đã xây dựng và thu thập một bộ dữ liệu toàn diện về khách hàng, các đơn hàng, quy trình xử lý đơn hàng, và thông tin chi tiết về các nhà bán hàng tiềm năng. Dự án phân tích dữ liệu này nhằm khai thác những giá trị hữu ích từ bộ dữ liệu Marketing Funnel mà Olist cung cấp, từ đó tối ưu hóa chiến lược tiếp thị và cải thiện hiệu quả bán hàng. Bộ dữ liệu này không chỉ giúp Olist tối ưu hóa hoạt động kinh doanh của mình mà còn hỗ trợ các nhà bán hàng hiểu rõ hơn về thị trường và hành vi của khách hàng.

Bộ dữ liệu Marketing Funnel by Olist là một tập hợp phong phú, cung cấp cái nhìn toàn diện về quy trình bán hàng và marketing của các nhà bán hàng đã đăng ký trên nền tảng. Dữ liệu bao gồm thông tin từ khoảng 8,000 Marketing Qualified Leads (MQLs), đại diện cho những nhà bán hàng đã yêu cầu được liên hệ để bán sản phẩm trên nền tảng Olist trong giai đoạn từ ngày 1 tháng 6 năm 2017 đến ngày 1 tháng 6 năm 2018. Đây là một mẫu ngẫu nhiên, đại diện và phản ánh chân thực về hoạt động marketing của Olist, giúp cung cấp nhiều góc nhìn khác nhau về quy trình bán hàng, từ loại hình khách hàng tiềm năng, kích thước danh mục sản phẩm đến hồ sơ hành vi của khách hàng. Những thông tin này giúp Olist và các nhà bán hàng hiểu rõ hơn về thị trường, từ đó đưa ra các quyết định chiến lược, tối ưu hóa chiến dịch marketing và cải thiện trải nghiệm người dùng.

Dataset về Customers chứa thông tin chi tiết về khách hàng và vị trí địa lý của họ. Trong hệ thống của Olist, mỗi đơn hàng được gán với một mã customer_id duy nhất. Điều này có nghĩa là mỗi lần một khách hàng thực hiện đơn hàng mới, họ sẽ nhận được một mã customer_id mới. Tuy nhiên, để theo dõi khách hàng có hoạt động mua hàng lặp lại, Olist sử dụng customer_unique_id nhằm xác định những khách hàng có nhiều lần mua sắm, qua đó giúp nhận diện các khách hàng trung thành. Dữ liệu này còn bao gồm thông tin địa lý như mã vùng (zip code), thành phố và bang, giúp Olist tối ưu hóa chiến lược marketing địa phương hóa và cải thiện quy trình giao hàng.

Dataset về Orders lưu trữ thông tin về từng đơn hàng, bao gồm trạng thái đơn hàng, thời gian phê duyệt, thời gian vận chuyển và thời gian giao hàng dự kiến. Đây là nguồn thông

tin quan trọng để phân tích hiệu suất vận hành và trải nghiệm khách hàng. Thông qua việc phân tích dữ liệu về thời gian xử lý đơn hàng, Olist có thể đánh giá hiệu quả của quy trình giao hàng, phát hiện các điểm yếu trong chuỗi cung ứng và đưa ra các cải tiến nhằm tăng cường hiệu quả và sự hài lòng của khách hàng. Ví dụ, nếu dữ liệu cho thấy có sự chậm trễ đáng kể trong việc giao hàng ở một khu vực cụ thể, công ty có thể xem xét mở rộng hoặc tối ưu hóa các đối tác vận chuyển tại khu vực đó.

Marketing Funnel Dataset ghi nhận thông tin về các nhà bán hàng tiềm năng đã gửi yêu cầu liên hệ để bán sản phẩm qua Olist. Mỗi bản ghi đại diện cho một MQL và chứa các thông tin như loại sản phẩm, phân khúc kinh doanh, hồ sơ hành vi và các yếu tố khác liên quan đến khách hàng tiềm năng. Dataset này hỗ trợ Olist phân tích từng giai đoạn trong quy trình marketing, xác định các yếu tố giúp tăng tỷ lệ chuyển đổi từ khách hàng tiềm năng thành khách hàng thực sự. Ví dụ, qua phân tích, Olist có thể xác định liệu loại hình sản phẩm, kích thước danh mục hay phân khúc kinh doanh có ảnh hưởng đến quyết định trở thành nhà bán hàng hay không. Điều này giúp công ty tối ưu hóa chiến lược marketing cho từng nhóm khách hàng khác nhau, tập trung tài nguyên vào các yếu tố mang lại hiệu quả cao nhất và đạt được lợi thế cạnh tranh trong thị trường thương mại điện tử tại Brazil.

Dự án này tập trung vào việc phân tích hành vi mua hàng, nhận diện khách hàng trung thành và tối ưu hóa quy trình giao hàng. Ngoài ra, thông qua thông tin về MQLs, dự án sẽ giúp Olist tối ưu hóa tỷ lệ chuyển đổi từ khách hàng tiềm năng thành khách hàng thực sự, đồng thời cung cấp insights về thị trường và nhu cầu khách hàng. Những dữ liệu này không chỉ phục vụ cho mục tiêu phát triển của Olist mà còn giúp các doanh nghiệp nhỏ cộng tác hiểu rõ hơn về khách hàng và tối ưu hóa hoạt động kinh doanh của mình.

1.2 YÊU CẦU CỦA CÔNG TY

Về mặt dữ liệu

Công ty Olist mong muốn tận dụng dữ liệu chi tiết về khách hàng, đơn hàng, nhà bán hàng và các yếu tố marketing để phân tích hiệu quả bán hàng và tối ưu hóa chiến lược marketing. Bộ dữ liệu hiện tại bao gồm nhiều bảng dữ liệu liên kết với nhau, mỗi bảng mang thông tin riêng về khách hàng, đơn hàng, sản phẩm và nhà bán hàng. Cụ thể:

- Bảng Customers: Chứa thông tin chi tiết về khách hàng bao gồm customer_id, customer_unique_id, customer_zip_code_prefix, customer_city, và customer_state. Đây là dữ liệu quan trọng để xác định các khách hàng trung thành, phân tích phân bố địa lý và tìm hiểu hành vi mua sắm theo khu vực.
- Bảng Orders: Cung cấp dữ liệu về các đơn hàng với các cột như order_id, customer_id, order_status, order_purchase_timestamp, order_approved_at, order_delivered_carrier_date, order_delivered_customer_date. Dữ liệu này giúp phân tích quy trình vận hành từ lúc đặt hàng đến lúc giao hàng, đánh giá hiệu quả giao hàng và phát hiện các vấn đề tiềm ẩn trong chuỗi cung ứng.
- Bảng Order Items: Chứa thông tin chi tiết về các sản phẩm trong từng đơn hàng, bao gồm order_id, order_item_id, product_id, seller_id, shipping_limit_date, price, và freight_value. Bảng này cho phép tính toán doanh thu từ mỗi đơn hàng, phân tích chi phí vận chuyển và đánh giá mức độ phổ biến của các sản phẩm.

- Bảng Products: Cung cấp thông tin sản phẩm với các cột product_id, product_category_name, product_name_length, product_description_length, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, và product_width_cm. Dữ liệu này rất quan trọng để xác định các danh mục sản phẩm được ưa chuộng, phân tích các yếu tố ảnh hưởng đến quyết định mua hàng và giúp công ty tối ưu hóa danh mục sản phẩm.
- Bảng Sellers: Lưu trữ thông tin về nhà bán hàng với các trường như seller_id, seller_zip_code_prefix, seller_city, và seller_state. Thông tin này hỗ trợ phân tích hiệu quả của các nhà bán hàng trên nền tảng, phân bổ theo khu vực và đánh giá mức độ đóng góp của từng nhà bán hàng.
- Bảng Order Reviews: Chứa thông tin phản hồi của khách hàng, bao gồm review_id, order_id, review_score, review_comment_title, review_comment_message, review_creation_date, và review_answer_timestamp. Dữ liệu này giúp đánh giá mức độ hài lòng của khách hàng đối với sản phẩm và dịch vụ, từ đó cải thiện trải nghiệm khách hàng.
- Bảng Closed Deals: Chứa thông tin về các giao dịch thành công, bao gồm mql_id, seller_id, sdr_id, sr_id, won_date, business_segment, lead_type, lead_behaviour_profile, has_company, has_gtin, average_stock, business_type, declared_product_catalog_size, declared_monthly_revenue. Dữ liệu này cung cấp cái nhìn chi tiết về các giao dịch thành công, giúp doanh nghiệp hiểu rõ hơn về hiệu quả bán hàng và tập trung vào các khách hàng tiềm năng có khả năng chuyển đổi cao.
- Bảng Marketing Qualified Leads: Chứa thông tin về các khách hàng tiềm năng đã được xác định, bao gồm mql_id, first_contact_date, landing_page_id, và origin. Dữ liệu này rất quan trọng để đánh giá hiệu quả của các kênh marketing, từ đó tối ưu hóa chiến lược thu hút khách hàng.
- Bảng Product Category Name Translation: Chứa thông tin về tên danh mục sản phẩm bằng tiếng Anh, bao gồm product_category_name và product_category_name_english. Dữ liệu này hữu ích cho việc hiểu rõ hơn về danh mục sản phẩm.
- Bảng Geolocation: Chứa thông tin về tọa độ địa lý, bao gồm geolocation_zip_code_prefix, geolocation_lat, geolocation_lng, geolocation_city và geolocation_state. Thông tin này có thể được sử dụng để tạo bản đồ trực quan, phân tích bán hàng theo khu vực địa lý chi tiết, tối ưu hóa việc vận chuyển, kho bãi, và các quyết định kinh doanh khác dựa trên vị trí.
- Bảng Order Payments: Ghi lại thông tin về các khoản thanh toán của đơn hàng, bao gồm order_id, payment_sequential, payment_type, payment_installments, và payment_value. Dữ liệu này cung cấp cái nhìn chi tiết về phương thức thanh toán ưa thích của khách hàng, giúp doanh nghiệp hiểu rõ hơn về tình hình tài chính và đưa ra các quyết định kinh doanh phù hợp.
- Bộ dữ liệu này bao gồm hàng trăm nghìn bản ghi và đòi hỏi phải đảm bảo chất lượng để có thể khai thác tối đa. Dữ liệu cần phải được làm sạch, chuẩn hóa và mô hình hóa

trước khi phân tích. Quá trình làm sạch sẽ bao gồm loại bỏ các bản ghi trùng lặp, xử lý các giá trị thiếu và xác định các ngoại lệ để đảm bảo độ chính xác. Việc chuẩn hóa dữ liệu giúp duy trì sự nhất quán trong toàn bộ hệ thống, từ định dạng thời gian đến chuẩn hóa tên cột. Cuối cùng, dữ liệu sẽ được mô hình hóa để phù hợp với các phân tích cụ thể, giúp tối ưu hóa hiệu suất truy vấn và giảm thiểu thời gian xử lý.

Quản lý và lưu trữ

- Cần xây dựng hệ thống quản lý và lưu trữ dữ liệu hiệu quả để dễ dàng truy cập và sử dụng khi cần thiết.
- Xây dựng hệ thống quản lý dữ liệu: Hệ thống này phải dễ dàng sử dụng và truy cập để đảm bảo rằng dữ liệu luôn sẵn sàng khi cần thiết. Việc quản lý dữ liệu cần bao gồm cả việc làm sạch và chuẩn hóa dữ liệu để đảm bảo tính chính xác và nhất quán.
- Sử dụng công nghệ phù hợp: Cần lựa chọn các công cụ và công nghệ phù hợp để quản lý và lưu trữ dữ liệu. Các công nghệ này cần hỗ trợ việc phân tích dữ liệu và tạo ra các báo cáo, biểu đồ trực quan để giúp dễ dàng hiểu và sử dụng dữ liệu, đặc biệt giá thành, chi phí vận hành ở mức tối thiểu nhưng vẫn mang lại hiệu quả cao.
- Dữ liệu sẽ được quản lý và lưu trữ trên SQL Server nhằm đảm bảo tính bảo mật và dễ dàng truy cập, đồng thời cho phép thực hiện các truy vấn phân tích nhanh chóng. SQL Server cung cấp một môi trường lưu trữ dữ liệu ổn định, phù hợp cho các tập dữ liệu lớn với nhiều bảng và hàng trăm nghìn bản ghi. Việc lưu trữ dữ liệu trên SQL Server giúp tối ưu hóa khả năng truy xuất và đảm bảo an toàn dữ liệu. Hệ thống sẽ được thiết lập các quyền truy cập rõ ràng, chỉ những nhân viên có quyền mới có thể thao tác trên dữ liệu, điều này giúp bảo vệ tính bảo mật của dữ liệu.
- Dữ liệu cần được phân loại và chuẩn hóa theo từng bảng dữ liệu, cùng với việc thiết lập các khóa ngoại để đảm bảo tính toàn vẹn và liên kết giữa các bảng. Để quản lý dữ liệu hiệu quả, mỗi bảng sẽ có các chỉ mục phù hợp nhằm tăng tốc độ truy vấn. SQL Server cũng sẽ được cấu hình để tự động sao lưu định kỳ, tránh mất dữ liệu khi có sự cố xảy ra. Tất cả dữ liệu sẽ được kiểm soát chặt chẽ qua các quyền truy cập và cơ chế bảo mật của SQL Server.

Mục tiêu

- Dự án phân tích dữ liệu này có các mục tiêu cụ thể để tối ưu hóa hoạt động kinh doanh và tăng cường trải nghiệm khách hàng của công ty. Đầu tiên, dự án tập trung vào phân tích hành vi khách hàng và xác định những khách hàng trung thành dựa trên các đơn hàng lặp lại. Điều này giúp công ty xây dựng các chương trình khuyến mãi và chăm sóc khách hàng hiệu quả hơn nhằm tăng cường lòng trung thành của khách hàng.
- Tiếp theo, dự án sẽ phân tích dữ liệu về hiệu suất giao hàng và thời gian xử lý đơn hàng từ lúc đặt hàng đến khi giao hàng hoàn tất. Thông qua các dữ liệu về thời gian giao hàng và phản hồi từ khách hàng, công ty có thể xác định các điểm yếu trong quy trình vận hành và thực hiện các cải tiến cần thiết để giảm thiểu thời gian giao hàng và tăng cường trải nghiệm người dùng.

- Một mục tiêu quan trọng khác là tối ưu hóa các chiến dịch marketing thông qua dữ liệu về khách hàng tiềm năng (MQLs) và các thông tin từ bảng Order Reviews. Dự án sẽ giúp xác định các yếu tố có tác động tích cực đến quyết định mua hàng, từ đó cải thiện nội dung tiếp thị và tăng tỷ lệ chuyển đổi. Dữ liệu về phản hồi của khách hàng cũng sẽ được phân tích để hiểu rõ hơn về sự hài lòng của họ, giúp công ty điều chỉnh các yếu tố sản phẩm hoặc dịch vụ nhằm đáp ứng nhu cầu của khách hàng một cách tốt nhất.
- Cuối cùng, dự án hướng đến dự báo xu hướng và xác định các yếu tố có khả năng tác động đến doanh thu trong tương lai. Công ty mong muốn xây dựng các mô hình dự báo doanh thu dựa trên các yếu tố lịch sử và xu hướng tiêu dùng, qua đó đưa ra các chiến lược kinh doanh và marketing dài hạn hiệu quả. Các báo cáo sẽ giúp ban lãnh đạo đưa ra quyết định chiến lược dựa trên các số liệu và dự đoán được xây dựng từ mô hình.

Đánh giá tính khả thi

- Năng lực (skill hiện có): Với kinh nghiệm và kiến thức sẵn có, tôi đã có thể thực hiện các công việc cơ bản trong dự án này. Hiện tại, tôi có các kỹ năng mạnh về SQL, Python, NumPy, Excel, Power BI, và kiến thức vững về toán xác suất thống kê. Những kỹ năng này giúp tôi dễ dàng trong việc truy vấn, xử lý dữ liệu lớn, phân tích định lượng và xây dựng báo cáo chi tiết về các xu hướng và hành vi khách hàng. Bên cạnh đó, khả năng kể chuyện qua dữ liệu là một thế mạnh, giúp tôi truyền tải các kết quả phân tích một cách rõ ràng, dễ hiểu qua các báo cáo trực quan. Power BI là công cụ tôi sử dụng để tạo các biểu đồ và dashboard chuyên nghiệp, giúp các bên liên quan có cái nhìn trực quan và toàn diện về dữ liệu.

- Năng lực (skill sẽ học thêm cho dự án): Để đạt được mục tiêu dài hạn của dự án và tối ưu hóa quá trình phân tích, tôi sẽ bổ sung kiến thức và kỹ năng trong một số lĩnh vực mới. Tôi dự định học thêm về Pandas và Seaborn để xử lý dữ liệu nhanh chóng và trực quan hóa dữ liệu nâng cao trên Python. Bên cạnh đó, để thực hiện các yêu cầu dự báo và dự đoán, tôi sẽ nghiên cứu các thuật toán hồi quy nhằm xây dựng mô hình phân tích dữ liệu dự báo doanh thu và hành vi khách hàng trong tương lai. Matplotlib cũng là công cụ tôi sẽ học để tạo ra các biểu đồ chuyên sâu, linh hoạt và trực quan hơn, phù hợp cho việc trình bày dữ liệu trong các báo cáo chi tiết. Với việc kết hợp thêm các kỹ năng này, tôi tin tưởng rằng mình có đủ khả năng để đáp ứng yêu cầu và hoàn thành dự án một cách hiệu quả.

Tóm lại, với nền tảng kỹ năng hiện có và kế hoạch nâng cao kỹ năng cần thiết, tôi có đầy đủ năng lực để triển khai dự án này. Tôi sẽ phối hợp chặt chẽ với các bên liên quan để đảm bảo rằng các phân tích đáp ứng đúng mục tiêu của công ty và mang lại giá trị cao nhất từ dữ liệu.

1.3 LẬP KẾ HOẠCH DỰ ÁN

TT	HẠNG MỤC	BẮT ĐẦU	KẾT THÚC	THÀNH VIÊN	KẾT QUẢ
1	Giới thiệu dự án	4/11/2024	4/11/2024	Vị	100%
1.1	Giới thiệu	4/11/2024	4/11/2024	Vị	100%
1.2	Yêu cầu	4/11/2024	4/11/2024	Vị	100%
1.3	Lập kế hoạch dự án	9/11/2024	12/11/2024	Linh	100%
2	Phân tích yêu cầu KH	9/11/2024	11/11/2024	Cả nhóm	100%
2.1	Phân tích yêu cầu	9/11/2024	11/11/2024	Linh	100%
2.2	Câu chuyện dữ liệu	9/11/2024	11/11/2024	Linh	100%
2.3	Kiến trúc hệ thống	12/11/2024	13/11/2024	Vị	100%
2.4	Giải thích về bộ dữ liệu khách hàng	12/11/2024	13/11/2024	Vị, Linh	100%
3	Làm sạch và chuyển đổi dữ liệu	13/11/2024	13/11/2024	Vị	100%
3.1	Chuẩn bị dữ liệu	13/11/2024	13/11/2024	Linh, Sâm	100%
3.2	Làm sạch dữ liệu	13/11/2024	13/11/2024	Vị	100%
3.3	Chuyển đổi dữ liệu	13/11/2024	13/11/2024	Vị	100%
4	Xử lý dữ liệu	13/11/2024	14/11/2024	Vị	100%

4.1	Chuẩn hóa dữ liệu	13/11/2024	14/11/2024	Vị	100%
4.2	Mô hình hóa dữ liệu	13/11/2024	14/11/2024	Vị	100%
4.3	Xử lý dữ liệu DAX	13/11/2024	14/11/2024	Cả nhóm	100%
5	Trực quan hóa dữ liệu	15/11/2024	16/11/2024	Cả nhóm	100%
5.1	Các kỹ thuật trực quan hóa	16/11/2024	16/11/2024	Linh	100%
5.2	Các nguyên tắc trực quan hóa	16/11/2024	17/11/2024	Linh	100%
5.3	Trình bày cách thêm visual mới	17/11/2024	18/11/2024	Vị	100%
5.4	Trình bày tạo các report cho dự án	17/11/2024	19/11/2024	Vị	100%
6	Xây dựng báo cáo	20/11/2024	24/11/2024	Cả nhóm	100%
6.1	Dashboard và Report	20/11/2024	24/11/2024	Vân	100%
6.2	Xây dựng báo cáo	21/11/2024	24/11/2024	Vân	100%
7	Kết luận	25/11/2024	26/11/2024	Cả nhóm	100%
7.1	Báo cáo	25/11/2024	26/11/2024	Linh	100%
7.2	Khó khăn	26/11/2024	27/11/2024	Vân	100%
7.3	Thuận lợi	26/11/2024	28/11/2024	Vân	100%

7.4	Hướng phát triển	27/11/2024	29/11/2024	Linh	100%
8	Tổng kết	29/11/2024	29/11/2024	Hồng	100%

Bảng 1.1: Kế hoạch chi tiết

2 PHÂN TÍCH YÊU CẦU KHÁCH HÀNG

2.1 PHÂN TÍCH YÊU CẦU

Dữ liệu:

- Tổng quan:** Số lượng khách hàng, số lượng sản phẩm, số lượng nhà bán hàng, số lượng giao dịch, giá trị trung bình mỗi đơn hàng, doanh thu theo thời gian, sản phẩm có doanh thu cao nhất, danh mục sản phẩm phổ biến nhất, tỷ lệ phương thức thanh toán, tỷ lệ hoàn thành đơn hàng, số lượng review, trung bình review_score, tổng số rating 5 sao và tỷ lệ phần trăm.
- Phân tích khách hàng:** Phân bổ khách hàng dựa trên đánh giá, khu vực giao hàng nhiều nhất, nguồn khách hàng tiềm năng, doanh thu trung bình theo khu vực, khách hàng có doanh thu cao nhất.
- Phân tích sản phẩm:** Danh mục sản phẩm phổ biến nhất, xu hướng sản phẩm theo thời gian, giá trị trung bình của từng danh mục sản phẩm, sản phẩm được đánh giá cao nhất và thấp nhất, sản phẩm thường được mua cùng nhau, số lượng mua và đánh giá của sản phẩm mới.
- Phân tích hành vi khách hàng:** Tần suất mua sắm của khách hàng, tỷ lệ xem và mua hàng, ảnh hưởng của mã giảm giá, giờ cao điểm mua sắm, tỷ lệ chuyển đổi theo nguồn, thời gian trung bình trên trang và hành vi mua hàng.
- Phân tích chuyển đổi khách hàng tiềm năng:** Nguồn gốc của MQLs, tỷ lệ chuyển đổi MQL thành khách hàng, thời gian từ MQL đến khách hàng, hồ sơ hành vi của MQLs, hiệu quả của các phân khúc kinh doanh, kích thước danh mục sản phẩm và tỷ lệ chuyển đổi.

Quản lý và lưu trữ:

- Dữ liệu được lưu trữ trong các tệp CSV.
- Cần một hệ thống quản lý cơ sở dữ liệu (SQL) để lưu trữ và truy vấn dữ liệu hiệu quả.

Công nghệ:

- Sử dụng Power BI để trực quan hóa dữ liệu và tạo báo cáo.
- Sử dụng các kỹ thuật thống kê và học máy (ví dụ: hồi quy) để phân tích chuyên sâu và dự đoán xu hướng.
- Sử dụng ngôn ngữ lập trình Python và các thư viện như Pandas, NumPy để xử lý và phân tích dữ liệu.

- Sử dụng các công cụ AI hỗ trợ lên kế hoạch, phân tích, tạo ảnh, video, ...

Quyết định sử dụng các công nghệ:

- **Dữ liệu:** Cơ sở dữ liệu

- **Quản lý và lưu trữ:**

Ngôn ngữ truy vấn dữ liệu SQL Server.

Ngôn ngữ lập trình Python và các thư viện Pandas, NumPy.

- **Công nghệ:**

Công cụ trực quan hóa dữ liệu và tạo báo cáo Power BI.

Các kỹ thuật thống kê và học máy (ví dụ: hồi quy).

- **AI:** ChatGPT, Gemini, invideo, Aicolor...

Giải thích:

- **Cơ sở dữ liệu quan hệ:** Phù hợp để lưu trữ dữ liệu có cấu trúc và quan hệ giữa các bảng.
- **SQL Server:** Ngôn ngữ truy vấn dữ liệu mạnh mẽ, cho phép truy vấn và xử lý dữ liệu hiệu quả.
- **Python:** Ngôn ngữ lập trình phổ biến trong lĩnh vực khoa học dữ liệu, với các thư viện mạnh mẽ hỗ trợ phân tích dữ liệu.
- **Power BI:** Công cụ mạnh mẽ để trực quan hóa dữ liệu và tạo báo cáo tương tác.
- **Thống kê và học máy:** Giúp phân tích chuyên sâu và dự đoán xu hướng.

2.2 CÂU CHUYỆN DỮ LIỆU

2.2.1 ĐẶT VẤN ĐỀ

MÔ TẢ THỰC TRẠNG:

Olist là cửa hàng bách hóa trực tuyến lớn nhất Brazil, với sứ mệnh kết nối các doanh nghiệp nhỏ trên toàn quốc với kênh bán hàng trực tuyến. Olist tạo điều kiện cho các doanh nghiệp tiếp cận thị trường thương mại điện tử. Chỉ với một hợp đồng duy nhất, các doanh nghiệp có thể bán sản phẩm của mình trên Cửa hàng Olist, nơi mọi khía cạnh từ tiếp thị, xử lý đơn hàng, đến vận chuyển đều được Olist quản lý. Các sản phẩm sau đó sẽ được vận chuyển trực tiếp đến tay khách hàng thông qua mạng lưới đối tác hậu cần rộng lớn của Olist.

Trong bối cảnh thị trường thương mại điện tử phát triển mạnh mẽ, việc thu hút khách hàng, tối ưu hóa quy trình bán hàng và nâng cao trải nghiệm người dùng trở thành yếu tố quyết định giúp các doanh nghiệp cạnh tranh và phát triển bền vững. Olist nhận thấy vai trò quan trọng của dữ liệu trong việc cung cấp những góc nhìn chuyên sâu, giúp đưa ra các quyết định chiến lược tối ưu. Do đó, công ty đã xây dựng và thu thập một bộ dữ liệu toàn diện về khách hàng, các đơn hàng, quy trình xử lý đơn hàng, và thông tin chi tiết về các nhà bán hàng tiềm năng. Dự án phân tích dữ liệu này nhằm khai thác những giá trị hữu ích từ bộ dữ liệu Marketing Funnel mà Olist cung cấp, từ đó tối ưu hóa chiến lược tiếp thị và cải thiện

hiệu quả bán hàng. Bộ dữ liệu này không chỉ giúp Olist tối ưu hóa hoạt động kinh doanh của mình mà còn hỗ trợ các nhà bán hàng hiểu rõ hơn về thị trường và hành vi của khách hàng.

DỮ LIỆU LIÊN QUAN:

Các tập dữ liệu được cung cấp bao gồm thông tin về khách hàng, đơn hàng, sản phẩm, nhà bán hàng, vị trí địa lý, đánh giá, thanh toán và các hoạt động tiếp thị, 11 bảng dữ liệu được cung cấp bởi Olist bao gồm "customer", "geolocation", "order_items", "order_payments", "order_reviews", "orders", "product", "sellers", "Closed Deals", "Marketing Qualified Leads" và "Product Category Name Translation"

MỤC TIÊU:

Xây dựng câu chuyện dữ liệu hấp dẫn, trực quan và dễ hiểu, giúp người đọc nắm bắt thông tin nhanh chóng và chính xác. Câu chuyện cần làm rõ vấn đề, chỉ ra các insight quan trọng và đề xuất giải pháp khả thi, góp phần tối ưu hóa hoạt động kinh doanh của Olist và các đối tác.

2.2.2 XÁC ĐỊNH CÂU CHUYỆN

Thông điệp từ dữ liệu:

Câu chuyện sẽ tập trung vào hành trình của khách hàng, từ lúc tiếp cận Olist đến khi hoàn thành đơn hàng và đưa ra đánh giá. Đồng thời, câu chuyện sẽ khai thác hiệu quả hoạt động kinh doanh của Olist và các nhà bán hàng trên nền tảng.

Mục tiêu:

- Thúc đẩy lòng trung thành của khách hàng.
- Nâng cao trải nghiệm mua sắm.
- Tối ưu hóa hiệu quả tiếp thị và bán hàng.
- Phát triển kinh doanh bền vững.

Giải pháp:

- Cải thiện dịch vụ khách hàng.
- Tối ưu hóa quy trình vận chuyển.
- Cá nhân hóa trải nghiệm mua sắm.
- Nâng cao chất lượng sản phẩm.
- Xây dựng chương trình khách hàng thân thiết.

Cách tiếp cận:

- Tìm kiếm mối tương quan giữa các yếu tố (ví dụ: đánh giá và hành vi mua hàng).
- Xác định xu hướng mua sắm theo thời gian.
- Rút ra so sánh giữa các kênh tiếp thị, các nhóm khách hàng, các nhà bán hàng.

2.2.3 XÁC ĐỊNH RỘI ĐỒI TƯỢNG

Đối tượng:

Ban lãnh đạo:

- **Nhu cầu:** Họ cần hiểu rõ bức tranh tổng quan về hiệu quả kinh doanh, các chỉ số tăng trưởng, xu hướng thị trường, và các yếu tố ảnh hưởng đến lợi nhuận để đưa ra quyết định chiến lược.
- **Lợi ích:** Câu chuyện dữ liệu sẽ giúp họ nắm bắt được các thông tin này một cách nhanh chóng và hiệu quả, từ đó đưa ra các quyết định kinh doanh sáng suốt, tối ưu hóa hoạt động và nâng cao lợi nhuận.

Các nhà bán hàng:

- **Nhu cầu:** Họ cần thông tin chi tiết về hiệu quả bán hàng của từng sản phẩm, kênh bán hàng, khu vực thị trường, và hành vi của khách hàng để cải thiện hoạt động kinh doanh của mình.
- **Lợi ích:** Câu chuyện dữ liệu sẽ cung cấp cho họ những thông tin chi tiết và thực tiễn, giúp họ hiểu rõ hơn về thị trường, khách hàng và hiệu quả của các chiến lược bán hàng, từ đó điều chỉnh và cải thiện hoạt động kinh doanh của mình.

Phòng ban Marketing:

- Nhu cầu: Họ cần thông tin chi tiết về hiệu quả của các chiến dịch marketing, hành vi khách hàng, và xu hướng thị trường để tối ưu hóa các chiến lược tiếp thị.
- Lợi ích: Câu chuyện dữ liệu sẽ giúp họ hiểu rõ hơn về hiệu quả của các kênh tiếp thị, phân khúc khách hàng tiềm năng, và hành vi mua sắm của khách hàng, từ đó điều chỉnh chiến lược tiếp thị, phân bổ ngân sách hiệu quả và tạo ra các chiến dịch tiếp thị thành công hơn.

Kho:

- Nhu cầu: Họ cần thông tin về số lượng hàng tồn kho, xu hướng tiêu thụ sản phẩm, và dự báo nhu cầu trong tương lai để quản lý kho bãi hiệu quả.
- Lợi ích: Câu chuyện dữ liệu sẽ cung cấp cho họ cái nhìn tổng quan về tình hình hàng tồn kho, giúp họ dự báo nhu cầu, tối ưu hóa việc nhập hàng và xuất hàng, giảm thiểu chi phí lưu kho và tránh tình trạng hết hàng hoặc tồn kho quá mức.

Phân tích đối tượng:**Ban lãnh đạo:**

- Am hiểu về kinh doanh, có kiến thức chuyên sâu về các chỉ số kinh doanh, thị trường và chiến lược.
- Cần dữ liệu để hỗ trợ ra quyết định, đánh giá hiệu quả hoạt động và định hướng phát triển.

Các nhà bán hàng:

- Am hiểu về sản phẩm, quy trình bán hàng và thị trường.
- Cần dữ liệu để phân tích hiệu quả bán hàng, hiểu rõ hơn về khách hàng và cải thiện hoạt động kinh doanh.

Phòng ban Marketing:

- Am hiểu về thị trường, khách hàng, và các kênh tiếp thị.
- Cần dữ liệu để phân tích hiệu quả chiến dịch marketing, hành vi khách hàng, và xu hướng thị trường.
- Cần dữ liệu để hỗ trợ ra quyết định về chiến lược tiếp thị, phân bổ ngân sách và thiết kế chiến dịch.

Kho:

- Am hiểu về quy trình quản lý kho, luân chuyển hàng hóa và dự báo nhu cầu.
- Cần dữ liệu về số lượng hàng tồn kho, xu hướng tiêu thụ sản phẩm, và dự báo nhu cầu.
- Cần dữ liệu để hỗ trợ ra quyết định về nhập hàng, xuất hàng, và quản lý kho bãi.

2.2.4 XÁC ĐỊNH CÂU CHUYỆN CHI TIẾT

Bối cảnh: Thị trường thương mại điện tử cạnh tranh, Olist cần tối ưu hóa để phát triển.

Các bên liên quan: Olist, nhà bán hàng, khách hàng.

Mạch truyện:

- Khách hàng tiếp cận Olist qua kênh nào? (Nguồn gốc, hiệu quả kênh tiếp thị)
- Hành vi mua sắm của khách hàng? (Sản phẩm, danh mục, phương thức thanh toán)
- Trải nghiệm mua sắm? (Giao hàng, đánh giá)
- Các yếu tố ảnh hưởng đến quyết định mua hàng? (Giá, khuyến mãi, đánh giá)
- Olist hỗ trợ nhà bán hàng như thế nào? (Tỷ lệ chuyển đổi MQL, hiệu quả phân khúc)
- Xu hướng và dự báo? (Doanh thu, sản phẩm)

2.2.5 TRÌNH BÀY DỮ LIỆU

Câu chuyện chi tiết	Biểu đồ	Mô tả
Khách hàng tiếp cận Olist qua kênh nào?	Biểu đồ Bar, Biểu đồ Funnel	Phân tích hiệu quả của các kênh tiếp thị (mạng xã hội, email, tìm kiếm, v.v.) trong việc thu hút khách hàng tiềm năng. Giúp Olist xác định kênh hiệu quả nhất để tối ưu chiến lược tiếp thị và tăng tỷ lệ chuyển đổi.
Hành vi mua sắm của khách hàng?	Biểu đồ Scatter	Phân tích chi tiết hành vi mua sắm của khách hàng dựa trên các yếu tố như loại sản phẩm, danh mục hàng hóa, và phương thức

		than toán. Điều này giúp xác định các sản phẩm phổ biến, các lựa chọn thanh toán ưu tiên và cách mà hành vi mua sắm thay đổi theo các yếu tố này.
Trải nghiệm mua sắm?	Biểu đồ Heatmap, Biểu đồ Pie	Đánh giá và trực quan hóa trải nghiệm mua sắm của khách hàng, bao gồm thời gian giao hàng, chất lượng dịch vụ giao hàng và các đánh giá sau khi nhận sản phẩm. Các yếu tố này cho thấy mức độ hài lòng và những cải tiến tiềm năng trong quy trình giao hàng nhằm tối ưu hóa trải nghiệm mua sắm.
Các yếu tố ảnh hưởng đến quyết định mua hàng?	Biểu đồ đường, Biểu đồ Scatter	Phân tích ảnh hưởng của các yếu tố như giá cả, chương trình khuyến mãi và đánh giá từ người dùng khác. Việc trực quan hóa này giúp nhận biết các yếu tố quan trọng nhất thúc đẩy khách hàng ra quyết định, từ đó cung cấp thông tin để xây dựng chiến lược marketing và điều chỉnh giá cả hiệu quả.
Olist hỗ trợ nhà bán hàng như thế nào?	Biểu đồ Funnel, Biểu đồ đường	Đánh giá hiệu quả của các kênh hỗ trợ nhà bán hàng từ phía Olist, bao gồm tỷ lệ chuyển đổi từ khách hàng tiềm năng thành khách hàng thực sự, hiệu quả các phân khúc kinh doanh và hỗ trợ các nhà bán hàng tối ưu hóa bán hàng trên nền tảng. Điều này cung cấp thông tin để cải tiến

		các dịch vụ hỗ trợ nhà bán hàng.
Xu hướng và dự báo?	Biểu đồ đường, Biểu đồ Bar	Trực quan hóa xu hướng doanh thu và dự báo về các danh mục sản phẩm bán chạy nhất theo thời gian, giúp Olist nhận diện các cơ hội tăng trưởng và đưa ra quyết định chiến lược dài hạn. Các dự báo này hỗ trợ việc lập kế hoạch sản xuất và phân bổ nguồn lực hợp lý.

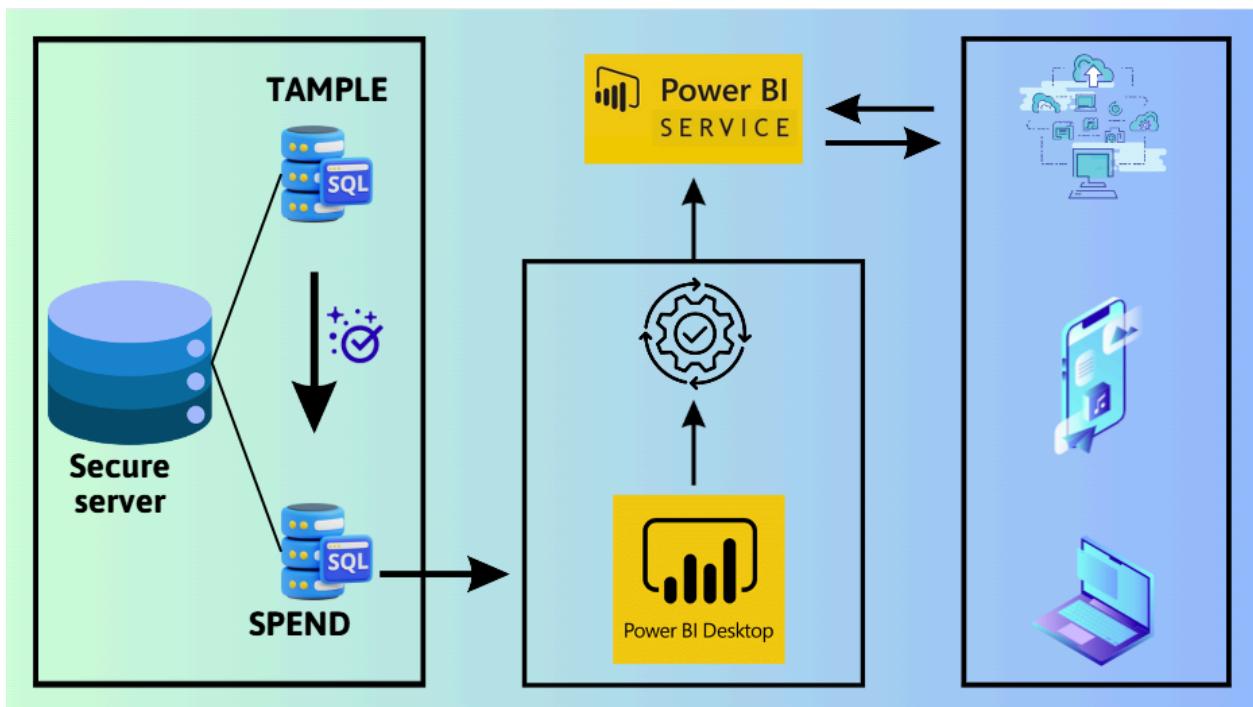
Bảng 2.1: Trình bày dữ liệu

2.2.6 NHỮNG ĐIỀU CẦN LUÔN Ý

- **Rõ ràng:** Dữ liệu và thông tin cần dễ hiểu.
- **Hấp dẫn:** Sử dụng hình ảnh, màu sắc, bố cục hợp lý.
- **Súc tích:** Tránh quá tải thông tin.
- **Kết nối:** Liên kết các thông tin, tạo thành mạch truyện.
- **Hành động:** Đưa ra khuyến nghị, giải pháp.

2.3 KIẾN TRÚC HỆ THỐNG

2.3.1 KIẾN TRÚC



Hình 2.1: Mô hình kiến trúc hệ thống

Máy chủ SQL:

- Sample: Cơ sở dữ liệu SQL Server này lưu trữ dữ liệu gốc.
- Spend : Cơ sở dữ liệu SQL Server này chứa dữ liệu chi tiêu cuối cùng khi tiền xử lý.

Secure Server: Dữ liệu đi từ Secure Server đến cơ sở dữ liệu SAMPLE và SPEND trong môi trường SQL Server.

Power BI Desktop:

- Power BI Desktop kết nối với cơ sở dữ liệu SPEND trong SQL Server để lấy dữ liệu nhằm mục đích phân tích và trực quan hóa.
- Dữ liệu này sau đó được sử dụng để tạo báo cáo và bảng thông tin.

Power BI Service:

- Các báo cáo và bảng thông tin được tạo trong Power BI Desktop sẽ được xuất bản lên Power BI Service.
- Dịch vụ Power BI cho phép làm mới dữ liệu, chia sẻ và cộng tác trên các báo cáo và bảng thông tin.

Người sử dụng/ Nhân viên: Nhân viên truy cập Dịch vụ Power BI bằng nhiều thiết bị khác nhau:

- Máy tính để bàn
- Máy tính bảng
- Điện thoại di động

Luồng dữ liệu có thể được tóm tắt như sau:

1. Dữ liệu di chuyển từ Secure Server sang SQL Server (cơ sở dữ liệu SAMPLE).
2. Sau đó, dữ liệu được xử lý và lưu trữ trong cơ sở dữ liệu SPEND.
3. Power BI Desktop truy cập cơ sở dữ liệu SPEND để tạo hình ảnh trực quan và báo cáo.
4. Những hình ảnh trực quan và báo cáo này được xuất bản trên Dịch vụ Power BI.
5. Nhân viên truy cập dịch vụ Power BI bằng nhiều thiết bị khác nhau để xem và tương tác với các báo cáo và bảng thông tin.

2.3.2 GIẢI THÍCH

Tổng quan thiết kế

Nguồn dữ liệu (Secure Server): Đây là nơi dữ liệu gốc được lưu trữ và bảo mật. Dữ liệu từ đây sẽ được chuyển vào hệ thống SQL Server để xử lý.

Hệ thống SQL Server:

- SAMPLE Database: Đây là bước đầu tiên trong quy trình xử lý dữ liệu, nơi dữ liệu ban đầu từ Secure Server được lưu trữ.

- SPEND Database: Sau khi dữ liệu được làm sạch và xử lý ở đây, dữ liệu sẽ được lưu trữ và chuẩn bị hoàn toàn cho việc phân tích.

Công cụ phân tích và trực quan hóa:

- **Power BI Desktop:** Công cụ này kết nối với SPEND Database để lấy dữ liệu đã được xử lý. Power BI Desktop được sử dụng để tạo các biểu đồ, báo cáo và bảng điều khiển.
- **Power BI Service:** Sau khi tạo ra các báo cáo và bảng điều khiển trong Power BI Desktop, chúng được xuất bản lên Power BI Service. Power BI Service cung cấp một nền tảng trực tuyến để lưu trữ, quản lý và chia sẻ các báo cáo này.

Người dùng cuối (Employees): Nhân viên truy cập Power BI Service qua các thiết bị như máy tính bàn, máy tính bảng và điện thoại di động. Họ sử dụng các báo cáo và bảng điều khiển để thu thập thông tin và đưa ra các quyết định dựa trên dữ liệu.

Quy trình và luồng dữ liệu

Thu thập dữ liệu: Dữ liệu từ Secure Server được chuyển vào SAMPLE Database trên SQL Server.

Xử lý và chuẩn bị dữ liệu: Dữ liệu từ SAMPLE Database được xử lý và làm sạch trong SPEND Database.

Lưu trữ dữ liệu đã xử lý: Dữ liệu đã được xử lý trong SPEND Database, nơi nó sẵn sàng cho việc phân tích.

Phân tích và tạo báo cáo: Power BI Desktop kết nối với SPEND Database để lấy dữ liệu và tạo các biểu đồ, báo cáo, bảng điều khiển. Các báo cáo này sau đó được xuất bản lên Power BI Service.

Truy cập và sử dụng dữ liệu: Nhân viên sử dụng các thiết bị khác nhau để truy cập Power BI Service, nơi họ có thể xem và tương tác với các báo cáo và bảng điều khiển.

Lợi ích của thiết kế hệ thống

Bảo mật dữ liệu: Việc sử dụng Secure Server và quy trình xử lý qua các bước trung gian đảm bảo rằng dữ liệu được bảo mật và xử lý chính xác trước khi được phân tích.

Quản lý dữ liệu hiệu quả: Sử dụng các cơ sở dữ liệu riêng biệt cho từng giai đoạn xử lý giúp quản lý dữ liệu một cách hiệu quả và có tổ chức.

Trực quan hóa dữ liệu mạnh mẽ: Power BI Desktop và Power BI Service cung cấp các công cụ mạnh mẽ để tạo ra các biểu đồ, báo cáo và bảng điều khiển, giúp người dùng dễ dàng hiểu và sử dụng dữ liệu.

Truy cập linh hoạt: Nhân viên có thể truy cập dữ liệu từ nhiều thiết bị khác nhau, giúp họ có thể làm việc mọi lúc, mọi nơi và ra quyết định nhanh chóng.

2.4 GIẢI THÍCH VỀ BỘ DỮ LIỆU KHÁCH HÀNG

2.4.1 CÁC KHÁI NIỆM

Dữ liệu thô (Raw Data): Dữ liệu chưa qua xử lý, được nhập trực tiếp từ các nguồn như file CSV, hệ thống giao dịch, v.v.

Làm sạch dữ liệu (Data Cleaning): Quá trình loại bỏ hoặc sửa chữa các lỗi, dữ liệu không hợp lệ, và các giá trị thiếu trong dữ liệu.

Chuẩn hóa dữ liệu (Data Normalization): Quá trình tổ chức dữ liệu để giảm sự dư thừa và cải thiện tính toàn vẹn của dữ liệu, thường thông qua việc tạo các bảng chuẩn hóa theo chuẩn 3NF (Third Normal Form).

Marketing Qualified Leads (MQLs): Khách hàng tiềm năng đáp ứng các tiêu chí nhất định, thể hiện sự quan tâm đến sản phẩm/dịch vụ và có khả năng cao trở thành khách hàng trả tiền. Ví dụ, họ có thể đã đăng ký nhận bản tin, tải xuống tài liệu hoặc tương tác với nội dung quảng cáo.

Closed Deals: Giao dịch thành công, khách hàng đã mua sản phẩm/dịch vụ. Điều này thường liên quan đến việc hoàn tất đơn đặt hàng, ký kết hợp đồng hoặc một hành động tương tự xác nhận việc mua bán.

Customer Unique ID: ID duy nhất để theo dõi khách hàng có hoạt động mua hàng lặp lại, giúp phân biệt với khách hàng mới. Điều này cho phép doanh nghiệp hiểu rõ hơn về hành vi của khách hàng trung thành và đưa ra các chiến lược tiếp thị phù hợp.

SDR (Sales Development Representative): Đại diện phát triển kinh doanh, chịu trách nhiệm tìm kiếm và tiếp cận khách hàng tiềm năng mới.

SR (Sales Representative): Đại diện bán hàng, chịu trách nhiệm tư vấn và chốt sales với khách hàng tiềm năng.

GTIN (Global Trade Item Number): Mã số nhận dạng sản phẩm duy nhất trên toàn cầu, được sử dụng để theo dõi và quản lý hàng hóa trong chuỗi cung ứng.

2.4.2 CÁC TRƯỜNG DỮ LIỆU

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	customer_id	text	ID của khách hàng
2	customer_unique_id	text	ID duy nhất của khách hàng
3	customer_zip_code_prefix	integer	Mã vùng của khách hàng
4	customer_city	text	Thành phố của khách hàng

5	customer_state	text	Bang của khách hàng
---	----------------	------	---------------------

Bảng 2.2: Bảng Customers

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	order_id	text	ID của đơn hàng
2	customer_id	text	ID của khách hàng
3	order_status	text	Trạng thái của đơn hàng (ví dụ: đã đặt, đã giao, đã hủy)
4	order_purchase_timestamp	timestamp	Thời gian khách hàng đặt hàng
5	order_approved_at	timestamp	Thời gian đơn hàng được phê duyệt
6	order_delivered_carrier_date	timestamp	Thời gian đơn hàng được giao cho đơn vị vận chuyển
7	order_delivered_customer_date	timestamp	Thời gian khách hàng nhận được hàng
8	order_estimated_delivery_date	timestamp	Thời gian dự kiến giao hàng

Bảng 2.3: Bảng Orders

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	order_id	text	ID của đơn hàng
2	order_item_id	integer	ID của sản phẩm trong đơn hàng (một đơn hàng có thể có nhiều sản phẩm)

3	product_id	text	ID của sản phẩm
4	seller_id	text	ID của người bán
5	shipping_limit_date	timestamp	Hạn vận chuyển cho người bán
6	price	float	Giá của sản phẩm
7	freight_value	float	Giá trị vận chuyển

Bảng 2.4: Bảng Order Items

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	seller_id	text	ID của người bán
2	seller_zip_code_prefix	integer	Mã vùng của người bán
3	seller_city	text	Thành phố của người bán
4	seller_state	text	bang của người bán

Bảng 2.5: Bảng Sellers

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	product_id	text	ID của sản phẩm
2	product_category_name	text	Tên danh mục sản phẩm
3	product_name_length	integer	Độ dài tên sản phẩm
4	product_description_length	integer	Độ dài mô tả sản phẩm
5	product_photos_qty	integer	Số lượng ảnh sản phẩm

6	product_weight_g	integer	Trọng lượng sản phẩm (gram)
7	product_length_cm	integer	Chiều dài sản phẩm (cm)
8	product_height_cm	integer	Chiều cao sản phẩm (cm)
9	product_width_cm	integer	Chiều rộng sản phẩm (cm)

Bảng 2.6: Bảng Products

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	geolocation_zip_code_prefix	integer	Mã vùng
2	geolocation_lat	float	Vĩ độ
3	geolocation_lng	float	Kinh độ

Bảng 2.7: Bảng Geolocation

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	product_category_name	text	Tên danh mục sản phẩm
2	product_category_name_english	text	Tên danh mục sản phẩm bằng tiếng Anh

Bảng 2.8: Product Category Name Translation

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	order_id	text	ID của đơn hàng
2	payment_sequential	integer	Số thứ tự thanh toán (một đơn hàng có thể có nhiều lần thanh toán)

3	payment_type	text	Loại thanh toán (ví dụ: thẻ tín dụng, boleto)
4	payment_installments	integer	Số lần trả góp
5	payment_value	float	Giá trị thanh toán

Bảng 2.9: Bảng Order Payments

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	review_id	text	ID của đánh giá
2	order_id	text	ID của đơn hàng
3	review_score	integer	Điểm đánh giá
4	review_comment_title	text	Tiêu đề bình luận
5	review_comment_message	text	Nội dung bình luận
6	review_creation_date	timestamp	Ngày tạo đánh giá
7	review_answer_timestamp	timestamp	Thời gian trả lời đánh giá

Bảng 2.10: Bảng Order Reviews

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	mql_id	text	ID của MQL
2	first_contact_date	timestamp	Ngày liên hệ đầu tiên
3	landing_page_id	text	ID của trang đích

4	origin	text	Nguồn gốc
---	--------	------	-----------

Bảng 2.11: Bảng Marketing Qualified Leads

STT	Trường dữ liệu	Kiểu dữ liệu	Mô tả
1	mql_id	text	ID của MQL
2	seller_id	text	ID của người bán
3	sdr_id	text	ID của SDR
4	sr_id	text	ID của SR
5	won_date	timestamp	Ngày chốt giao dịch
6	business_segment	text	Phân khúc kinh doanh
7	lead_type	text	Loại khách hàng tiềm năng
8	lead_behaviour_profile	text	Hồ sơ hành vi
9	has_company	boolean	Có công ty hay không
10	has_gtin	boolean	Có GTIN hay không
11	average_stock	integer	Lượng hàng trung bình
12	business_type	text	Loại hình kinh doanh
13	declared_product_catalog_size	integer	Kích thước danh mục sản phẩm
14	declared_monthly_revenue	float	Doanh thu hàng tháng

Bảng 2.12: Bảng Closed Deals

3 LÀM SẠCH VÀ CHUYỂN ĐỔI DỮ LIỆU

3.1 CHUẨN BỊ DỮ LIỆU

3.1.1 GIẢI PHÁP LUU TRỮ DỮ LIỆU

Lưu trữ dữ liệu là một phần quan trọng của bất kỳ hệ thống quản lý dữ liệu nào. Việc chọn lựa giải pháp lưu trữ phù hợp phụ thuộc vào nhiều yếu tố như quy mô dữ liệu, yêu cầu về bảo mật, khả năng mở rộng, chi phí và khả năng truy cập. Hiện nay, có hai giải pháp lưu trữ chính: giải pháp nền tảng đám mây (Cloud-based solution) và các ứng dụng tại chỗ (On-premise).

So sánh cloud-based solution and on-premise

Giải pháp nền tảng đám mây (Cloud-based solution)

Ưu điểm:

- Khả năng mở rộng: Dễ dàng mở rộng hoặc thu nhỏ tài nguyên dựa trên nhu cầu thực tế mà không cần đầu tư cơ sở hạ tầng mới.
- Chi phí: Trả phí theo mức sử dụng (pay-as-you-go), không cần chi phí đầu tư ban đầu lớn cho phần cứng và phần mềm.
- Truy cập từ xa: Có thể truy cập dữ liệu từ bất kỳ đâu có kết nối internet.
- Sao lưu và khôi phục: Các nhà cung cấp dịch vụ đám mây thường cung cấp các giải pháp sao lưu và khôi phục dữ liệu tự động.
- Bảo mật và tuân thủ: Các nhà cung cấp dịch vụ đám mây thường có các tiêu chuẩn bảo mật và tuân thủ nghiêm ngặt.

Nhược điểm:

- Chi phí dài hạn: Chi phí có thể tăng lên đáng kể theo thời gian nếu nhu cầu sử dụng tài nguyên lớn.
- Quản lý dữ liệu: Phụ thuộc vào nhà cung cấp dịch vụ trong việc quản lý và bảo mật dữ liệu.
- Tốc độ truy cập: Tốc độ truy cập dữ liệu có thể bị ảnh hưởng bởi chất lượng kết nối internet.

Các ứng dụng tại chỗ (on-premise)

Ưu điểm:

- Kiểm soát hoàn toàn: Doanh nghiệp có toàn quyền kiểm soát cơ sở hạ tầng và dữ liệu của mình.
- Bảo mật: Dữ liệu được lưu trữ tại chỗ, giảm thiểu rủi ro bị truy cập trái phép từ bên ngoài.

- Tốc độ truy cập: Tốc độ truy cập dữ liệu có thể nhanh hơn do không phụ thuộc vào kết nối internet.

Nhược điểm:

- Chi phí đầu tư ban đầu: Chi phí đầu tư ban đầu cho phần cứng, phần mềm, và cơ sở hạ tầng là rất lớn.
- Khả năng mở rộng: Việc mở rộng cơ sở hạ tầng đòi hỏi thêm chi phí và thời gian.
- Bảo trì và nâng cấp: Doanh nghiệp phải tự chịu trách nhiệm về bảo trì, nâng cấp và quản lý cơ sở hạ tầng.
- Sao lưu và khôi phục: Cần có kế hoạch sao lưu và khôi phục dữ liệu chi tiết để tránh mất mát dữ liệu.

Quyết định và lý do

Lựa chọn: giải pháp ứng dụng tại chỗ (on-premise) sử dụng SQL Server

Lý do:

- Kiểm soát hoàn toàn: Doanh nghiệp có toàn quyền kiểm soát cơ sở hạ tầng và dữ liệu của mình.
- Bảo mật: Dữ liệu được lưu trữ tại chỗ, giảm thiểu rủi ro bị truy cập trái phép từ bên ngoài.
- Tốc độ truy cập: Tốc độ truy cập dữ liệu có thể nhanh hơn do không phụ thuộc vào kết nối internet.
- Phù hợp với dữ liệu có cấu trúc: SQL Server là một hệ quản trị cơ sở dữ liệu quan hệ, phù hợp để lưu trữ dữ liệu có cấu trúc và quan hệ giữa các bảng, như trong trường hợp của Olist.
- Ngôn ngữ truy vấn mạnh mẽ: SQL là một ngôn ngữ truy vấn dữ liệu mạnh mẽ, cho phép truy vấn và xử lý dữ liệu hiệu quả.
- Khả năng mở rộng: Mặc dù việc mở rộng cơ sở hạ tầng on-premise có thể tốn kém, SQL Server cung cấp khả năng mở rộng tốt để đáp ứng nhu cầu của Olist trong tương lai.
- Ổn định và đáng tin cậy: SQL Server là một hệ thống ổn định và đáng tin cậy, phù hợp cho các tập dữ liệu lớn với nhiều bảng và hàng trăm nghìn bản ghi.
- Sao lưu và phục hồi: SQL Server cung cấp các tính năng sao lưu và phục hồi dữ liệu để đảm bảo an toàn dữ liệu.
- Cơ chế bảo mật: SQL Server có các cơ chế bảo mật mạnh mẽ để bảo vệ dữ liệu, bao gồm kiểm soát truy cập và mã hóa.
- Chi phí: Mặc dù chi phí đầu tư ban đầu cho cơ sở hạ tầng on-premise có thể cao, nhưng về lâu dài, nó có thể tiết kiệm hơn so với việc sử dụng các giải pháp đám mây, đặc biệt là khi nhu cầu sử dụng tài nguyên lớn.

Tóm lại, việc lựa chọn giải pháp ứng dụng tại chỗ sử dụng SQL Server mang lại cho Olist sự kiểm soát, bảo mật, tốc độ truy cập và khả năng mở rộng tốt, đồng thời đáp ứng các yêu cầu về chi phí và hiệu quả.

3.1.2 GIẢI PHÁP PHÂN BỐ DỮ LIỆU

Phân bố dữ liệu là một phần quan trọng trong việc thiết kế và quản lý cơ sở dữ liệu, giúp tăng hiệu suất, tính sẵn sàng và khả năng mở rộng của hệ thống. Nhân bản (Replication) là một kỹ thuật quan trọng trong việc phân bố cơ sở dữ liệu (CSDL) và thực thi các stored procedure. Nhân bản cho phép sao chép và duy trì các bản sao của cơ sở dữ liệu trên nhiều máy chủ khác nhau, giúp cải thiện khả năng truy cập và bảo mật dữ liệu.

Nhân bản trong csdl

Nhân bản là quá trình sao chép dữ liệu từ một cơ sở dữ liệu chính (primary database) sang một hoặc nhiều cơ sở dữ liệu phụ (secondary databases). Có ba loại nhân bản chính:

Nhân bản giao dịch (Transactional Replication):

- Mô tả: Sao chép các giao dịch từ cơ sở dữ liệu chính sang các cơ sở dữ liệu phụ. Các thay đổi được gửi ngay lập tức đến các bản sao, đảm bảo dữ liệu luôn được cập nhật.
- Ưu điểm: Đảm bảo dữ liệu nhất quán và cập nhật liên tục.
- Nhược điểm: Có thể tăng tải trên hệ thống mạng và máy chủ.

Nhân bản Snapshot (Snapshot Replication):

- Mô tả: Sao chép toàn bộ dữ liệu từ cơ sở dữ liệu chính vào các khoảng thời gian định trước.
- Ưu điểm: Đơn giản và dễ triển khai.
- Nhược điểm: Dữ liệu có thể không được cập nhật liên tục giữa các lần sao chép.

Nhân bản hợp nhất (Merge Replication):

- Mô tả: Cho phép các thay đổi từ cơ sở dữ liệu chính và cơ sở dữ liệu phụ được kết hợp lại với nhau.
- Ưu điểm: Phù hợp với các ứng dụng phân tán nơi mà cả hai bên có thể thực hiện các thay đổi độc lập.
- Nhược điểm: Phức tạp hơn trong việc xử lý xung đột dữ liệu.

Giải pháp nhân bản trong phạm vi dự án

Lựa chọn: Nhân bản giao dịch (Transactional Replication)

Lý do:

- Tính nhất quán cao: Đảm bảo dữ liệu nhất quán và cập nhật liên tục trên tất cả các bản sao. Điều này rất quan trọng đối với các ứng dụng yêu cầu dữ liệu chính xác và thời gian thực.

- Hiệu suất cao: Giảm tải cho cơ sở dữ liệu chính bằng cách phân tán các yêu cầu đọc đến các cơ sở dữ liệu phụ. Các stored procedure có thể được thực thi trên các bản sao để giảm tải trên hệ thống chính.
- Tính sẵn sàng cao: Cải thiện khả năng sẵn sàng và khôi phục sau sự cố bằng cách có nhiều bản sao dữ liệu trên các máy chủ khác nhau. Nếu một máy chủ gặp sự cố, các máy chủ khác vẫn có thể phục vụ yêu cầu truy cập dữ liệu.
- Dễ dàng quản lý: Dễ dàng triển khai và quản lý trong SQL Server, với các công cụ tích hợp hỗ trợ nhân bản giao dịch.

Cách triển khai nhân bản giao dịch trong sql server

Cấu hình máy chủ nhân bản: Thiết lập máy chủ phân phối (Distributor), máy chủ xuất bản (Publisher) và máy chủ đăng ký (Subscriber).

Tạo và cấu hình ấn phẩm: Tạo các ấn phẩm (Publication) trên máy chủ xuất bản chứa các bảng và stored procedure cần nhân bản.

Thiết lập đăng ký: Tạo các đăng ký (Subscription) trên máy chủ đăng ký để nhận dữ liệu từ máy chủ xuất bản.

Theo dõi và quản lý nhân bản: Sử dụng các công cụ quản lý của SQL Server để theo dõi trạng thái và hiệu suất của quá trình nhân bản.

Ví dụ:

```
--Tạo cơ sở dữ liệu phân phối
EXEC sp_adddistributiondb @distributor = 'DistributorServer',@password ='password'
```

Hình 3.1 Tạo cơ sở dữ liệu phân phối

```
--Thiết lập máy chủ phân phối
EXEC sp_adddistpublisher @database = 'distribution',
@data_folder = 'D:\Demo\Đự Án 1\Olist-E-commerce-Dataset',
@log_folder = 'D:\Demo\Đự Án 1\Olist-E-commerce-Dataset';
```

Hình 3.2 Thiết lập máy chủ phân phối

```
--Cấu hình máy chủ xuất bản
EXEC sp_adddistpublisher @publisher = 'PublisherServer',
@distribution_db = 'distribution',
@security_mode = 1;
```

Hình 3.3 Cấu hình máy chủ xuất bản

```
--Tạo ấn phẩm
EXEC sp_addpublication @publication = 'Ecommerce data',
@description ='Ecommerce data',
@sync_method = 'concurrent c',
@retention = 0,
@allow_push = N'true',
@allow_pull= N'true',
@allow_anonymous = N'true',
@enabled_for_internet = N' false';
```

Hình 3.4 Tạo ấn phẩm

```
--Thêm các bài viết vào các ấn phẩm
EXEC sp_addarticle @publication = 'Ecommerce data',
@article = 'DataRaw',
@source_object = 'DataRaw',
@type = N'logbased',
@description = 'Article description',
@creation_script = null,
@schema_option = 0x000000000803509F;
```

Hình 3.5 Thêm các bài viết vào ấn phẩm

```
--Tạo đăng ký
EXEC sp_addsubscription @publication = 'Ecommerce data',
@subscriber = 'SubscriberServer',
@destination_db = 'SubscriberDB',
@subscription_type = N'Push',
@sync_type = N'Automatic',
@article = N'all',
@update_mode = N'read only',
@subscriber_type = 0;
```

Hình 3.6 Tạo đăng ký

3.1.2.1 Ý nghĩa việc phân bố dữ liệu

Phân bố dữ liệu là quá trình tổ chức và quản lý dữ liệu sao cho dữ liệu được lưu trữ trên nhiều vị trí khác nhau, có thể là trên các máy chủ khác nhau trong cùng một hệ thống

hoặc trên các hệ thống khác nhau. Việc phân bố dữ liệu có ý nghĩa quan trọng trong nhiều khía cạnh của quản lý dữ liệu, từ việc cải thiện hiệu suất truy cập, tăng cường tính sẵn sàng và độ tin cậy, đến việc đảm bảo an toàn và bảo mật dữ liệu.

Các lợi ích của việc phân bố dữ liệu:

1. Cải thiện hiệu suất truy cập dữ liệu:

- Giảm tải trên máy chủ: Phân bố dữ liệu giúp phân tán tải công việc giữa các máy chủ khác nhau, tránh tình trạng quá tải trên một máy chủ đơn lẻ. Điều này cải thiện hiệu suất truy cập dữ liệu và đảm bảo rằng hệ thống có thể phục vụ nhiều người dùng đồng thời mà không bị chậm trễ.
- Tăng tốc độ truy cập: Dữ liệu được phân bố trên nhiều vị trí có thể giúp giảm độ trễ truy cập, đặc biệt là khi các máy chủ được đặt gần với người dùng cuối hoặc các ứng dụng truy cập dữ liệu.

2. Tăng cường tính sẵn sàng và độ tin cậy:

- Khả năng dự phòng: Khi dữ liệu được sao chép và lưu trữ trên nhiều máy chủ, hệ thống có khả năng tiếp tục hoạt động ngay cả khi một hoặc nhiều máy chủ gặp sự cố. Điều này làm tăng tính sẵn sàng của hệ thống và giảm thiểu thời gian ngừng hoạt động.
- Khả năng khôi phục: Dữ liệu phân bố giúp dễ dàng khôi phục lại dữ liệu trong trường hợp có sự cố hoặc mất mát dữ liệu. Các bản sao dự phòng có thể được sử dụng để khôi phục hệ thống một cách nhanh chóng và hiệu quả.

3. Cải thiện khả năng mở rộng:

- Mở rộng tài nguyên dễ dàng: Phân bố dữ liệu cho phép hệ thống mở rộng dễ dàng bằng cách thêm các máy chủ mới mà không cần thay đổi cấu trúc cơ bản của hệ thống. Điều này giúp hệ thống có thể phát triển và đáp ứng nhu cầu ngày càng tăng của người dùng.
- Phân tán dữ liệu theo khu vực: Hệ thống có thể phân bố dữ liệu theo khu vực địa lý để phục vụ người dùng ở các vị trí khác nhau một cách hiệu quả hơn. Ví dụ, dữ liệu có thể được lưu trữ ở các máy chủ gần với người dùng khu vực đó để giảm độ trễ và tăng tốc độ truy cập.

4. Đảm bảo an toàn và bảo mật dữ liệu:

- Bảo vệ dữ liệu trước các mối đe dọa: Phân bố dữ liệu trên nhiều máy chủ giúp bảo vệ dữ liệu trước các mối đe dọa như tấn công mạng, thiên tai, hoặc lỗi phần cứng. Dữ liệu có thể được sao lưu và bảo mật tại nhiều vị trí khác nhau.
- Tuân thủ quy định bảo mật: Phân bố dữ liệu theo khu vực địa lý cũng giúp tuân thủ các quy định bảo mật và bảo vệ dữ liệu của từng quốc gia hoặc khu vực.

5. Tối ưu hóa chi phí:

- Sử dụng tài nguyên hiệu quả: Phân bố dữ liệu cho phép sử dụng hiệu quả các tài nguyên máy chủ hiện có, giúp tối ưu hóa chi phí vận hành và bảo trì hệ thống.

- Giảm chi phí đầu tư: Bằng cách sử dụng các giải pháp đám mây và dịch vụ nhân bản dữ liệu, doanh nghiệp có thể giảm chi phí đầu tư vào phần cứng và cơ sở hạ tầng, chỉ trả phí dựa trên mức sử dụng thực tế.

Ứng dụng trong thực tế:

1. Hệ thống thương mại điện tử

Các hệ thống thương mại điện tử thường phải xử lý lượng lớn dữ liệu giao dịch và truy cập từ nhiều người dùng trên toàn thế giới. Việc phân bố dữ liệu giúp cải thiện tốc độ truy cập, đảm bảo tính sẵn sàng của hệ thống, và bảo vệ dữ liệu người dùng.

2. Ứng dụng tài chính

Các ứng dụng tài chính cần đảm bảo tính nhất quán và an toàn của dữ liệu giao dịch. Phân bố dữ liệu giúp cải thiện hiệu suất và đảm bảo rằng dữ liệu luôn được bảo vệ và có thể khôi phục nhanh chóng trong trường hợp xảy ra sự cố.

3. Dịch vụ trực tuyến

Các dịch vụ trực tuyến như mạng xã hội, dịch vụ truyền thông, và các nền tảng nội dung số cần xử lý và lưu trữ lượng lớn dữ liệu người dùng. Phân bố dữ liệu giúp hệ thống mở rộng dễ dàng và đáp ứng nhu cầu truy cập cao từ người dùng.

3.1.2.2 Trình bày cách phân bố dữ liệu

Việc phân bố dữ liệu trong dự án mang lại nhiều lợi ích như cải thiện hiệu suất, tăng cường tính sẵn sàng, khả năng mở rộng, và bảo mật dữ liệu. Bằng cách thiết lập hệ thống nhân bản, phân bố dữ liệu theo khu vực địa lý, phân bố tải công việc và thiết lập sao lưu định kỳ, dự án có thể đảm bảo rằng dữ liệu luôn sẵn sàng và an toàn, đáp ứng tốt nhu cầu của người dùng cuối và giúp doanh nghiệp đưa ra các quyết định dựa trên dữ liệu một cách hiệu quả.

1. Thiết lập hệ thống nhân bản

Nhân bản giao dịch (Transactional Replication): Sao chép các giao dịch từ cơ sở dữ liệu chính sang các cơ sở dữ liệu phụ để đảm bảo dữ liệu luôn được cập nhật và nhất quán.

Các bước thực hiện:

- Thiết lập máy chủ phân phối (Distributor Server): Thiết lập máy chủ phân phối để quản lý và theo dõi quá trình nhân bản.
- Tạo ấn phẩm (Publication): Tạo các ấn phẩm chứa các bảng và stored procedure cần nhân bản.
- Thiết lập đăng ký (Subscription): Tạo các đăng ký trên máy chủ đăng ký để nhận dữ liệu từ máy chủ xuất bản.

2. Phân bố dữ liệu theo các khu vực địa lý

Phân bố dữ liệu theo khu vực địa lý giúp giảm độ trễ truy cập và cải thiện hiệu suất.

Các bước thực hiện:

- Đặt máy chủ ở các khu vực địa lý khác nhau: Đặt các máy chủ cơ sở dữ liệu phụ ở các khu vực địa lý khác nhau để phục vụ người dùng khu vực đó.
- Nhân bản dữ liệu địa lý (Geographical Replication): Thiết lập nhân bản dữ liệu giữa các máy chủ ở các khu vực khác nhau để đảm bảo dữ liệu luôn được cập nhật.

3. Phân bố tải công việc

Phân bố tải công việc giúp giảm tải trên cơ sở dữ liệu chính và phân phối yêu cầu đọc/ghi giữa các máy chủ khác nhau.

Các bước thực hiện:

- Thiết lập cơ sở dữ liệu phụ chuyên biệt (Read Replica): Thiết lập các cơ sở dữ liệu phụ chuyên biệt cho các tác vụ đọc dữ liệu để giảm tải cho cơ sở dữ liệu chính.
- ### 4. Sao lưu và khôi phục dữ liệu

Sao lưu dữ liệu định kỳ và khả năng khôi phục dữ liệu nhanh chóng giúp đảm bảo tính sẵn sàng và an toàn của hệ thống.

Các bước thực hiện:

- Thiết lập kế hoạch sao lưu định kỳ: Lên lịch sao lưu cơ sở dữ liệu định kỳ (hàng ngày, hàng tuần) để đảm bảo có thể khôi phục dữ liệu khi cần.
- Kiểm tra và khôi phục dữ liệu: Thực hiện kiểm tra định kỳ các bản sao lưu và quy trình khôi phục để đảm bảo dữ liệu có thể được khôi phục nhanh chóng khi cần.

3.2 LÀM SẠCH DỮ LIỆU

3.2.1 CÁC VẤN ĐỀ ẢNH HƯỞNG TỚI DỮ LIỆU

3.2.1.1 Các vấn đề ảnh hưởng

Dữ liệu là tài sản quan trọng của bất kỳ hệ thống nào. Tuy nhiên, có nhiều vấn đề có thể ảnh hưởng tới chất lượng và tính toàn vẹn của dữ liệu. Các vấn đề này có thể đến từ nhiều nguồn khác nhau và nếu không được xử lý kịp thời, chúng có thể gây ra hậu quả nghiêm trọng.

Dữ liệu thiếu hoặc không chính xác: Dữ liệu thiếu hoặc không chính xác có thể gây ra sai lệch trong phân tích và đưa ra quyết định.

Ví dụ: Trường "order_approved_at" bị bỏ trống trong một số bản ghi của bảng Orders, hoặc giá trị "customer_city" chứa ký tự đặc biệt không hợp lệ.

Dữ liệu trùng lặp: Dữ liệu trùng lặp làm tăng kích thước cơ sở dữ liệu không cần thiết và có thể dẫn đến các phân tích không chính xác.

Ví dụ: Một khách hàng có nhiều bản ghi trong bảng Customers với các "customer_id" khác nhau nhưng cùng "customer_unique_id".

Dữ liệu ngoại lai (Outliers): Các giá trị ngoại lai có thể làm méo mó các kết quả phân tích.

Ví dụ: Một đơn hàng trong bảng Order Items có giá trị "price" cao bất thường so với các đơn hàng khác cùng loại sản phẩm.

Vấn đề về bảo mật dữ liệu: Dữ liệu nhạy cảm có thể bị truy cập trái phép nếu không được bảo mật đúng cách.

Ví dụ: Thông tin khách hàng bị rò rỉ do thiếu các biện pháp bảo mật.

Mất mát dữ liệu: Dữ liệu có thể bị mất mát do lỗi phần cứng, lỗi phần mềm, hoặc do thảm họa tự nhiên.

Ví dụ: Do lỗi phần cứng hoặc phần mềm, một phần dữ liệu trong bảng Order Reviews bị mất, bao gồm cả các đánh giá quan trọng của khách hàng.

Đồng bộ hóa dữ liệu: Dữ liệu không được đồng bộ hóa đúng cách giữa các hệ thống có thể dẫn đến sự không nhất quán.

Ví dụ: Dữ liệu về trạng thái đơn hàng ("order_status") trong bảng Orders không được cập nhật kịp thời, dẫn đến thông tin không chính xác trên hệ thống theo dõi đơn hàng của khách hàng.

Vấn đề về hiệu suất: Hiệu suất truy vấn chậm có thể gây ra sự chậm trễ trong việc truy cập và phân tích dữ liệu.

Ví dụ: Do số lượng bản ghi lớn trong bảng Order Items, truy vấn để tính toán tổng doanh thu theo từng danh mục sản phẩm mất nhiều thời gian, ảnh hưởng đến hiệu suất của hệ thống báo cáo.

3.2.1.2 Vấn đề đang tồn tại trong dự án

Việc nhận diện và xử lý các vấn đề ảnh hưởng đến dữ liệu là rất quan trọng để đảm bảo chất lượng và tính toàn vẹn của dữ liệu. Trong phạm vi dự án này, việc làm sạch và chuẩn hóa dữ liệu, xử lý dữ liệu trùng lặp và ngoại lai, đồng bộ hóa dữ liệu đúng cách, và tối ưu hóa hiệu suất là các yếu tố then chốt để thành công.

3.2.2 CÁC TIÊU CHÍ ĐÁNH GIÁ CHẤT LƯỢNG DỮ LIỆU

3.2.2.1 Các tiêu chí

Độ chính xác (Accuracy): Dữ liệu phải phản ánh đúng thực tế mà nó đại diện. Các giá trị dữ liệu phải chính xác và không có sai sót.

Ví dụ: Giá trị "freight_value" (giá trị vận chuyển) trong bảng Order Items phải khớp với chi phí vận chuyển thực tế mà Olist phải trả cho đơn vị vận chuyển.

Tính toàn vẹn (Integrity): Dữ liệu phải nhất quán và không bị hỏng. Tính toàn vẹn đảm bảo rằng các quan hệ giữa các thực thể trong cơ sở dữ liệu được duy trì.

Ví dụ: Mỗi đơn hàng trong bảng Orders phải có một mã "order_id" duy nhất và không trùng lặp. Các quan hệ giữa bảng Orders và bảng Order Items (qua "order_id") phải được duy trì để đảm bảo tính nhất quán của dữ liệu.

Tính đầy đủ (Completeness): Dữ liệu không được thiếu các giá trị quan trọng. Tất cả các trường dữ liệu cần thiết phải có giá trị.

Ví dụ: Mỗi bản ghi trong bảng Products phải có thông tin về tên danh mục sản phẩm ("product_category_name"), trọng lượng ("product_weight_g"), kích thước ("product_length_cm", "product_height_cm", "product_width_cm"), v.v.

Tính nhất quán (Consistency): Dữ liệu phải nhất quán trong toàn bộ cơ sở dữ liệu. Các giá trị tương tự phải được biểu diễn một cách thống nhất.

Ví dụ: Định dạng thời gian trong các trường "order_purchase_timestamp", "order_approved_at", "order_delivered_carrier_date" trong bảng Orders phải nhất quán (ví dụ: YYYY-MM-DD HH:MM:SS), và các giá trị "payment_value" trong bảng Order Payments phải sử dụng cùng một đơn vị tiền tệ (ví dụ: BRL).

Tính kịp thời (Timeliness): Dữ liệu phải được cập nhật kịp thời và phản ánh đúng thời điểm hiện tại. Dữ liệu cũ hoặc không được cập nhật kịp thời có thể không còn giá trị.

Ví dụ: Thông tin về trạng thái đơn hàng ("order_status") trong bảng Orders phải được cập nhật ngay khi có thay đổi (ví dụ: khi đơn hàng được giao cho đơn vị vận chuyển, khi đơn hàng được giao đến khách hàng).

Tính dễ hiểu (Understandability): Dữ liệu phải dễ hiểu và dễ sử dụng. Các trường dữ liệu và giá trị phải rõ ràng và có ý nghĩa.

Ví dụ: Các tên cột trong bảng Order Reviews (như "review_score", "review_comment_title", "review_comment_message") phải rõ ràng, dễ hiểu và không gây nhầm lẫn.

Tính truy xuất (Accessibility): Dữ liệu phải dễ dàng truy cập và sử dụng bởi các bên liên quan. Điều này bao gồm cả việc đảm bảo dữ liệu có thể được truy cập một cách an toàn và bảo mật.

Ví dụ: Dữ liệu từ các bảng Customers, Orders, Order Items, Products, Sellers, v.v. phải có thể truy cập từ Power BI và các công cụ phân tích khác một cách nhanh chóng và bảo mật.

3.2.2.2 Tiêu chí áp dụng trong dự án

Độ chính xác (Accuracy):

- Lý do: Đảm bảo các giá trị dữ liệu như giá sản phẩm ("price"), giá trị vận chuyển ("freight_value") trong bảng Order Items, điểm đánh giá ("review_score") trong bảng Order Reviews, và các thông tin khác là chính xác để phân tích và đưa ra quyết định đúng đắn.
- Cách thực hiện: Kiểm tra và xác minh dữ liệu với các nguồn đáng tin cậy, sử dụng các quy tắc và ràng buộc trong SQL Server để đảm bảo dữ liệu chính xác, ví dụ như kiểm tra định dạng dữ liệu, kiểm tra giới hạn giá trị, v.v.

Tính toàn vẹn (Integrity):

- Lý do: Duy trì tính toàn vẹn của các quan hệ giữa các bảng dữ liệu (ví dụ: giữa bảng Orders và Order Items, giữa bảng Customers và Orders) để tránh lỗi và mâu thuẫn.
- Cách thực hiện: Sử dụng các khóa chính và khóa ngoại trong SQL Server, và các ràng buộc toàn vẹn để đảm bảo dữ liệu không bị trùng lặp hoặc bị hỏng. Ví dụ, đảm bảo mỗi đơn hàng ("order_id") chỉ liên kết với một khách hàng ("customer_id") duy nhất.

Tính đầy đủ (Completeness):

- Lý do: Đảm bảo rằng tất cả các trường dữ liệu cần thiết đều có giá trị, giúp phân tích dữ liệu toàn diện và chính xác. Ví dụ, các trường "order_approved_at" trong bảng Orders, "product_category_name" trong bảng Products không được phép null.
- Cách thực hiện: Xử lý các giá trị NULL và các trường bị bỏ trống bằng cách sử dụng các kỹ thuật như điền giá trị trung bình, giá trị phổ biến nhất, hoặc sử dụng các thuật toán dự đoán.

Tính nhất quán (Consistency):

- Lý do: Dữ liệu nhất quán giúp đảm bảo rằng các phân tích và báo cáo dựa trên dữ liệu là đáng tin cậy.
- Cách thực hiện: Định dạng các trường dữ liệu một cách nhất quán trong toàn bộ cơ sở dữ liệu. Ví dụ, định dạng ngày tháng phải nhất quán (ví dụ: YYYY-MM-DD HH:MM:SS), và các giá trị tiền tệ phải sử dụng cùng một đơn vị (ví dụ: BRL).

Tính kịp thời (Timeliness):

- Lý do: Dữ liệu kịp thời giúp đưa ra các quyết định dựa trên thông tin mới nhất và chính xác, đặc biệt là các thông tin về trạng thái đơn hàng ("order_status").
- Cách thực hiện: Thiết lập các quy trình cập nhật dữ liệu định kỳ (ví dụ: hàng ngày) hoặc theo thời gian thực nếu có thể.

Tính dễ hiểu (Understandability):

- Lý do: Dữ liệu phải dễ hiểu và dễ sử dụng để các bên liên quan có thể sử dụng dữ liệu một cách hiệu quả.
- Cách thực hiện: Sử dụng các tên cột và các giá trị dữ liệu rõ ràng, dễ hiểu và không gây nhầm lẫn. Ví dụ, sử dụng tên cột đầy đủ và có ý nghĩa, sử dụng các giá trị phân loại rõ ràng và nhất quán.

Tính truy xuất (Accessibility):

- Lý do: Dữ liệu phải dễ dàng truy cập và sử dụng bởi các bên liên quan, bao gồm cả việc đảm bảo dữ liệu có thể được truy cập một cách an toàn và bảo mật.
- Cách thực hiện: Cung cấp quyền truy cập dữ liệu cho các bên liên quan thông qua các công cụ phù hợp như Power BI, SQL Server Management Studio, đồng thời thiết lập các cơ chế bảo mật để kiểm soát truy cập và ngăn chặn truy cập trái phép.

3.2.3 CÁC BƯỚC LÀM SẠCH DỮ LIỆU

3.2.3.1 Trình bày các bước làm sạch

Làm sạch dữ liệu là một bước quan trọng trong quá trình phân tích dữ liệu, giúp đảm bảo tính chính xác và đáng tin cậy của dữ liệu. Các bước làm sạch dữ liệu thường bao gồm:

- Xác định và xóa bỏ các giá trị thiếu: Kiểm tra và xử lý các giá trị bị thiếu bằng cách xóa các hàng chứa giá trị thiếu hoặc thay thế bằng giá trị trung bình, giá trị phổ biến, hoặc các kỹ thuật khác.

- Loại bỏ các giá trị ngoại lai: Xác định các giá trị ngoại lai hoặc bất thường có thể ảnh hưởng đến kết quả phân tích và loại bỏ hoặc điều chỉnh chúng.
- Chuẩn hóa dữ liệu: Đảm bảo tất cả các dữ liệu đều tuân theo một định dạng thống nhất, chẳng hạn như chuẩn hóa định dạng ngày tháng, chuyển đổi chữ hoa thành chữ thường, và loại bỏ các ký tự không cần thiết.
- Kiểm tra tính hợp lệ của dữ liệu: Đảm bảo dữ liệu tuân thủ các quy tắc và điều kiện nhất định, ví dụ như kiểm tra mã số định danh có đúng định dạng hay không, kiểm tra giá trị trong một phạm vi hợp lý.
- Xử lý dữ liệu trùng lặp: Xác định và loại bỏ các hàng dữ liệu trùng lặp để tránh tình trạng lặp lại thông tin.
- Chuyển đổi dữ liệu: Thực hiện các thao tác chuyển đổi dữ liệu như mã hóa lại các giá trị phân loại thành số, tính toán các biến mới từ các biến hiện có.

Việc làm sạch dữ liệu giúp đảm bảo rằng dữ liệu sử dụng trong phân tích là chính xác, nhất quán và sẵn sàng cho các bước xử lý tiếp theo, từ đó cải thiện độ tin cậy và tính chính xác của các kết quả phân tích.

3.2.3.2 Trình bày các bước làm sạch trong phạm vi dự án

Xem qua dữ liệu của từng bảng

```

14
15 select top 5 * from closed_deals
16 select top 5 * from customers
17 select top 5 * from geolocation
18 select top 5 * from marketing_qualified_leads
19 select top 5 * from order_items
20 select top 5 * from order_payments
21 select top 5 * from order_reviews
22 select top 5 * from orders
23 select top 5 * from product_category_name
24 select top 5 * from products

```

Result	Messages																																																																																																		
<table border="1"> <thead> <tr> <th>mlq_id</th> <th>seller_id</th> <th>adr_id</th> <th>er_id</th> <th>won_date</th> <th>business_segment</th> <th>lead_type</th> <th>lead_behaviour_pr</th> </tr> </thead> <tbody> <tr> <td>1 5420aadfec3549a85876ba1c529bd84</td> <td>2c43fb513632d29b3b58df74816f1b06</td> <td>a8387c01a09e99ce014107505b9238bc</td> <td>4ef15efb4b2723d0f3d81e51ec7afe</td> <td>2018-02-26 19:58:00.0000000</td> <td>pet</td> <td>online_medium</td> <td>cat</td> </tr> <tr> <td>2 a559fb36e9368110ed0e0043dc3b9a</td> <td>bbbd77893a450660432ea6652310eb7</td> <td>09285259593c61296eef0c734121db5</td> <td>d3d1e91a157ea790548eebf2195e53</td> <td>2018-05-08 20:17:00.0000000</td> <td>car_accessories</td> <td>industry</td> <td>eagle</td> </tr> <tr> <td>3 327174d3d848a2d047e8940d7615204ca</td> <td>61210e3a97004b3ba37eaae81836b4c</td> <td>b9087164b5fb2ca5a5c8572834bcb3f</td> <td>6565aa9eae3178a5cafb171827fa9b6</td> <td>2018-06-05 17:27:00.0000000</td> <td>home_appliances</td> <td>online_big</td> <td>cat</td> </tr> <tr> <td>4 ffeef87da7448875bc5ae2bafb6dd</td> <td>21e1781a36fa92725dde4730a88ca0f</td> <td>56bf83d4bb5703a51c2baab501ba467</td> <td>d3d1e91a157ea790548eebf2195e53</td> <td>2018-01-17 13:51:00.0000000</td> <td>food_drink</td> <td>online_small</td> <td>NULL</td> </tr> <tr> <td>5 fe640179b54e295c167a2b6be523e0</td> <td>ed5cb7b190ce606722747e8e48c8d0e</td> <td>4b339f9567d00b0ceaf5136b9f5949e</td> <td>d3d1e91a157ea790548eebf2195e53</td> <td>2018-07-03 20:17:00.0000000</td> <td>home_appliances</td> <td>industry</td> <td>wolf</td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th>customer_id</th> <th>customer_unique_id</th> <th>customer_zip_code_prefix</th> <th>customer_city</th> <th>customer_state</th> </tr> </thead> <tbody> <tr> <td>1 00012a2ce688dcda20d059ce98491703</td> <td>248fe10d632beb4e47267ff14484c9</td> <td>6273</td> <td>osasco</td> <td>SP</td> </tr> <tr> <td>2 000161a05860d5901007fb4+27140</td> <td>b0015e09bb4b6b647c52844fb6b633</td> <td>35550</td> <td>itapecerica</td> <td>MG</td> </tr> <tr> <td>3 000166190adaf884bcfa3d349edf079</td> <td>94b11d37cd61cb2994a194d11b98682b</td> <td>29830</td> <td>nova venecia</td> <td>ES</td> </tr> <tr> <td>4 0002414f95344307404faece7a261d5</td> <td>4093ad4ea282b5b3dd4f82e79db9e6</td> <td>39664</td> <td>mrendonca</td> <td>MG</td> </tr> <tr> <td>5 000379dec6252490c315e70c9a9b</td> <td>063873b19c2019e182dd52e048a22c</td> <td>4841</td> <td>sao paulo</td> <td>SP</td> </tr> </tbody> </table> <table border="1"> <thead> <tr> <th>geolocation_zip_code_prefix</th> <th>geolocation_lat</th> <th>geolocation_lng</th> <th>geolocation_city</th> <th>geolocation_state</th> </tr> </thead> <tbody> <tr> <td>1 64150</td> <td>-26.6844749450684</td> <td>-50.2986068725586</td> <td>são joão do triunfo</td> <td>PR</td> </tr> <tr> <td>2 84130</td> <td>-26.4287779947021</td> <td>-50.0048179626465</td> <td>palmeira</td> <td>PR</td> </tr> <tr> <td>3 84168</td> <td>-24.7711791992188</td> <td>-50.015127563477</td> <td>castro</td> <td>PR</td> </tr> </tbody> </table>	mlq_id	seller_id	adr_id	er_id	won_date	business_segment	lead_type	lead_behaviour_pr	1 5420aadfec3549a85876ba1c529bd84	2c43fb513632d29b3b58df74816f1b06	a8387c01a09e99ce014107505b9238bc	4ef15efb4b2723d0f3d81e51ec7afe	2018-02-26 19:58:00.0000000	pet	online_medium	cat	2 a559fb36e9368110ed0e0043dc3b9a	bbbd77893a450660432ea6652310eb7	09285259593c61296eef0c734121db5	d3d1e91a157ea790548eebf2195e53	2018-05-08 20:17:00.0000000	car_accessories	industry	eagle	3 327174d3d848a2d047e8940d7615204ca	61210e3a97004b3ba37eaae81836b4c	b9087164b5fb2ca5a5c8572834bcb3f	6565aa9eae3178a5cafb171827fa9b6	2018-06-05 17:27:00.0000000	home_appliances	online_big	cat	4 ffeef87da7448875bc5ae2bafb6dd	21e1781a36fa92725dde4730a88ca0f	56bf83d4bb5703a51c2baab501ba467	d3d1e91a157ea790548eebf2195e53	2018-01-17 13:51:00.0000000	food_drink	online_small	NULL	5 fe640179b54e295c167a2b6be523e0	ed5cb7b190ce606722747e8e48c8d0e	4b339f9567d00b0ceaf5136b9f5949e	d3d1e91a157ea790548eebf2195e53	2018-07-03 20:17:00.0000000	home_appliances	industry	wolf	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state	1 00012a2ce688dcda20d059ce98491703	248fe10d632beb4e47267ff14484c9	6273	osasco	SP	2 000161a05860d5901007fb4+27140	b0015e09bb4b6b647c52844fb6b633	35550	itapecerica	MG	3 000166190adaf884bcfa3d349edf079	94b11d37cd61cb2994a194d11b98682b	29830	nova venecia	ES	4 0002414f95344307404faece7a261d5	4093ad4ea282b5b3dd4f82e79db9e6	39664	mrendonca	MG	5 000379dec6252490c315e70c9a9b	063873b19c2019e182dd52e048a22c	4841	sao paulo	SP	geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state	1 64150	-26.6844749450684	-50.2986068725586	são joão do triunfo	PR	2 84130	-26.4287779947021	-50.0048179626465	palmeira	PR	3 84168	-24.7711791992188	-50.015127563477	castro	PR	✔ Query executed successfully.
mlq_id	seller_id	adr_id	er_id	won_date	business_segment	lead_type	lead_behaviour_pr																																																																																												
1 5420aadfec3549a85876ba1c529bd84	2c43fb513632d29b3b58df74816f1b06	a8387c01a09e99ce014107505b9238bc	4ef15efb4b2723d0f3d81e51ec7afe	2018-02-26 19:58:00.0000000	pet	online_medium	cat																																																																																												
2 a559fb36e9368110ed0e0043dc3b9a	bbbd77893a450660432ea6652310eb7	09285259593c61296eef0c734121db5	d3d1e91a157ea790548eebf2195e53	2018-05-08 20:17:00.0000000	car_accessories	industry	eagle																																																																																												
3 327174d3d848a2d047e8940d7615204ca	61210e3a97004b3ba37eaae81836b4c	b9087164b5fb2ca5a5c8572834bcb3f	6565aa9eae3178a5cafb171827fa9b6	2018-06-05 17:27:00.0000000	home_appliances	online_big	cat																																																																																												
4 ffeef87da7448875bc5ae2bafb6dd	21e1781a36fa92725dde4730a88ca0f	56bf83d4bb5703a51c2baab501ba467	d3d1e91a157ea790548eebf2195e53	2018-01-17 13:51:00.0000000	food_drink	online_small	NULL																																																																																												
5 fe640179b54e295c167a2b6be523e0	ed5cb7b190ce606722747e8e48c8d0e	4b339f9567d00b0ceaf5136b9f5949e	d3d1e91a157ea790548eebf2195e53	2018-07-03 20:17:00.0000000	home_appliances	industry	wolf																																																																																												
customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state																																																																																															
1 00012a2ce688dcda20d059ce98491703	248fe10d632beb4e47267ff14484c9	6273	osasco	SP																																																																																															
2 000161a05860d5901007fb4+27140	b0015e09bb4b6b647c52844fb6b633	35550	itapecerica	MG																																																																																															
3 000166190adaf884bcfa3d349edf079	94b11d37cd61cb2994a194d11b98682b	29830	nova venecia	ES																																																																																															
4 0002414f95344307404faece7a261d5	4093ad4ea282b5b3dd4f82e79db9e6	39664	mrendonca	MG																																																																																															
5 000379dec6252490c315e70c9a9b	063873b19c2019e182dd52e048a22c	4841	sao paulo	SP																																																																																															
geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state																																																																																															
1 64150	-26.6844749450684	-50.2986068725586	são joão do triunfo	PR																																																																																															
2 84130	-26.4287779947021	-50.0048179626465	palmeira	PR																																																																																															
3 84168	-24.7711791992188	-50.015127563477	castro	PR																																																																																															

Hình 3.7: Dữ liệu các bảng

QUÁ TRÌNH LÀM SẠCH TỪNG BẢNG

Bảng geolocation

```

33 -- Kiểm tra sự không nhất quán trong tên thành phố theo vị trí địa lý
34 select distinct geolocation_city as unique_geo, geolocation_state
35 from geolocation
36 group by geolocation_state, geolocation_city
37
38 -- Có nhiều dấu trong cột unique_geo (geolocation_city) --> thay thế chúng
39 select TRANSLATE(geolocation_city, 'áâçéíóúôâêõ', 'aaceiouaoe') as geolocation_city_re

```

Results Messages

unique_geo	geolocation_state
ibiam	SC
guatambú	SC
guaruja do sul	SC
guarulhos	SP
oliveira fortes	MG
conceição de jacareí	RJ
chacara	MG
astolfo dutra	MG
antônio gonçalves	BA
encruzilhada	BA
wenceslau guimaraes	BA
ubatã	BA
eunapolis	BA
lagoa real	BA
piaçá	BA
ituaçu	BA
cabaceiras	BA

Query executed successfully. VU-HONG-VI\MAVAYO (16.0 RTM) | VU-HONG-VI\DELL (61) | Olist_ECommerce_Raw

Hình 3.8: Kiểm tra sự không nhất quán trong tên thành phố theo vị trí địa lý

```

38 -- Có nhiều dấu trong cột unique_geo (geolocation_city) --> thay thế chúng
39 select TRANSLATE(geolocation_city, 'áâçéíóúôâêõ', 'aaceiouaoe') as geolocation_city_replaced
40 from geolocation
41
42 -- Cập nhật dữ liệu sau khi bỏ dấu vào bảng gốc

```

Results Messages

geolocation_city_replaced
sao joao do triunfo
palmeira
castro
castro
palmeira
castro
castro
castro
palmeira
castro
castro
sao joao do triunfo
castro
castro
palmeira
palmeira
castro

Executing query... VU-HONG-VI\MAVAYO (16.0 RTM) | VU-HONG-VI\DELL (61) | Olist_ECommerce_Raw_new | 00:00:01

Hình 3.9: Thay thế các giá trị bị lỗi

```

41
42 -- Cập nhật dữ liệu sau khi bỏ dấu vào bảng gốc
43 update geolocation
44 set geolocation_city = TRANSLATE(geolocation_city, 'áâçéíóúôâêõ', 'aaceiouaoe')
45

```

Messages

(1000163 rows affected)

Completion time: 2024-11-30T02:09:26.8849708+07:00

Hình 3.10: Cập nhật dữ liệu sau khi bỏ dấu

```

46 -- Có 1 số chữ số ở giữa tên các thành phố (geolocation_city) --> loại bỏ chúng
47 select TRANSLATE(geolocation_city, '0123456789', REPLICATE(' ', 10)) as geo_translated
48 from geolocation
49 where geolocation_city like '%[0-9]%'
50
51 -- Xóa các ký hiệu và khoảng trắng thừa

```

Results

	geo_translated
1	colonia z-
2	colonia z-
3	colonia z-
4	riacho fundo
5	quilometro do mutum
6	quilometro do mutum
7	lambani d' apos% boeste
8	o. centenario
9	o centenario
10	o centenario
11	maceia
12	sao joao do pau d' apos% balho

Hình 3.11: Loại bỏ số ở giữa các thành phố

```

50
51 -- Xóa các ký hiệu và khoảng trắng thừa
52 update geolocation
53 set geolocation_city = REPLACE(geolocation_city, ' ', '')
54 where geolocation_city like '%''%

```

(1579 rows affected)

Completion time: 2024-11-30T02:09:55.7821046+07:00

Hình 3.12: Xóa các ký tự và khoảng trắng thừa

```

56 -- Xóa phần ... ở đầu
57 update geolocation
58 set geolocation_city = SUBSTRING(geolocation_city, 4, LEN(geolocation_city))
59 where left(geolocation_city, 3) = '...'
60

```

(1 row affected)

Completion time: 2024-11-30T02:10:16.4415093+07:00

Hình 3.13: Xóa phần ... ở đầu

```

61 -- Xóa * ở đầu
62 update geolocation
63 set geolocation_city = SUBSTRING(geolocation_city, 2, LEN(geolocation_city))
64 where left(geolocation_city, 1) = '*'
65
66 -- -- Xóa phần 4o. ở đầu

```

52 %

Messages

(1 row affected)

Completion time: 2024-11-30T02:10:27.5027939+07:00

*Hình 3.14: Xóa * ở đầu*

```

66 -- -- Xóa phần 4o. ở đầu
67 update geolocation
68 set geolocation_city = STUFF(geolocation_city, 1, 4, '')
69 where left(geolocation_city, 4) = '4o. '

```

152 %

Messages

(1 row affected)

Completion time: 2024-11-30T02:10:41.5574333+07:00

Hình 3.15: Xóa phần 4o. ở đầu

```

73
74 -- Chỉ trích xuất một lần xuất hiện của rio de janeiro
75 update geolocation
76 set geolocation_city = SUBSTRING(geolocation_city, 1, CHARINDEX('rio de janeiro', geolocation_city))
77 where CHARINDEX('rio de janeiro', geolocation_city) > 0
78

```

52 %

Messages

(62152 rows affected)

Completion time: 2024-11-30T02:11:04.9544268+07:00

Hình 3.16: Chỉ trích xuất một lần xuất hiện rio de janeiro

```

79 -- Xóa bỏ chữ z ở cuối-
80 update geolocation
81 set geolocation_city = left(geolocation_city, LEN(geolocation_city) - 3)
82 where geolocation_city like '%z-3'
83
84 -- Xóa các ký hiệu mã hóa %26apos%3B và %26 trong cột geography_city

```

2 %

Messages

(3 rows affected)

Completion time: 2024-11-30T02:11:16.9298522+07:00

Hình 3.17: Xóa bỏ chữ z ở cuối

```

83
84 -- Xóa các ký hiệu mã hóa %26apos%3B và %26 trong cột geography_city
85 update geolocation
86 set geolocation_city = REPLACE(REPLACE(geolocation_city, '%26apos%3B', ''), '%26', '&')
87 where geolocation_city like '%sao joao do pau d%'
88
89 -- Xóa khoảng trắng trong cột geography_city bằng hàm REPLACE
90 update geolocation
%
```

(16 rows affected)

Completion time: 2024-11-30T02:11:26.3081862+07:00

Hình 3.18: Xóa các ký hiệu mã hóa %26apos%3B và %26 trong cột geography_city

```

-- Xóa khoảng trắng trong cột geography_city bằng hàm REPLACE
90 update geolocation
91 set geolocation_city = REPLACE(geolocation_city, ' ', '')
92 where geolocation_city like '%d alho%'
93
%
```

(3 rows affected)

Completion time: 2024-11-30T02:11:36.2753653+07:00

Hình 3.19: Xóa khoảng trắng trong cột geography_city bằng hàm REPLACE

```

-- Viết hoa chữ cái đầu tiên của mỗi từ trong cột geography_city
95 update geolocation
96 set geolocation_city = UPPER(SUBSTRING(geolocation_city, 1, 1)) + LOWER(SUBSTRING(geolocation_city, 2, 1))
%
```

(1000163 rows affected)

Completion time: 2024-11-30T02:11:51.3639851+07:00

Hình 3.20: Viết hoa chữ cái đầu tiên của từ trong cột geography_city

```

-- Đổi tên geography_state thành tên chưa viết tắt
99 UPDATE geolocation
100 SET geolocation_state =
101 CASE geolocation_state
102 WHEN 'AC' THEN 'Acre'
103 WHEN 'AL' THEN 'Alagoas'
%
```

(1000163 rows affected)

Completion time: 2024-11-30T02:12:23.1713364+07:00

Hình 3.21: Đổi tên geography_state thành tên chưa viết tắt

Bảng customers

```

134
135 -- Kiểm tra xem có sự không nhất quán trong customer_city không
136 select distinct customer_city, customer_state
137 from customers
138
139 -- Thay thế customer_state viết tắt thành tên đầy đủ

```

Results

customer_city	customer_state
boa esperanca	PR
joaquim nabuco	PE
turmalina	MG
sao jose da boa vista	PR
gurupi	TO
sao jose da laje	AL
ceilandia	DF
santo antonio do amparo	MG
soledade	PB
pedreiras	MA
	MG

Hình 3.22: Kiểm tra xem có sự không nhất quán trong customer_city không

```

-- Thay thế customer_state viết tắt thành tên đầy đủ
139 UPDATE customers
140 SET customer_state =
141 CASE customer_state
142 WHEN 'AC' THEN 'Acre'
143 WHEN 'AL' THEN 'Alagoas'
144

```

Messages

(99441 rows affected)

Completion time: 2024-11-30T02:15:45.6501716+07:00

Hình 3.23: Thay thế customer_state viết tắt thành tên đầy đủ

```

-- Thay đổi customer_city thành một trường hợp thích hợp
172 update customers
173 set customer_city = UPPER(LEFT(customer_city, 1)) + LOWER(SUBSTRING(customer_city, 2, LEN(customer_c
174

```

Messages

(99441 rows affected)

Completion time: 2024-11-30T02:16:00.4506803+07:00

Hình 3.24: Thay đổi customer_city thành một trường hợp thích hợp

Bảng order_items

```

180   -- Làm tròn price và freight_value lên 2 chữ số thập phân
181   -- price:
182   update order_items
183   set price = ROUND(price, 2)

% < Messages
(112650 rows affected)

```

Completion time: 2024-11-30T02:16:26.3412956+07:00

Hình 3.25: Làm tròn price và freight_value lên 2 chữ số thập phân

```

184
185   -- freight_value:
186   update order_items
187   set freight_value = ROUND(freight_value, 2)
188
189   -- Tạo cột mới
52 % < Messages
(112650 rows affected)

```

Completion time: 2024-11-30T02:16:40.8705903+07:00

Hình 3.26: freight_value:

```

--- 
189   -- Tạo cột mới
190   alter table order_items
191   add shipping_limit_date date,
192       shipping_limit_time time
2 % < Messages
Commands completed successfully.

```

Completion time: 2024-11-30T02:16:50.0874816+07:00

Hình 3.27: Tạo cột mới

```

194 -- Điền giá trị cho các cột mới
195 update order_items
196 set shipping_limit_dateee = CAST(shipping_limit_date AS DATE),
197     shipping_limit_time = CAST(shipping_limit_date AS TIME)
198
2 % < Messages

```

(112650 rows affected)

Completion time: 2024-11-30T02:17:05.0194588+07:00

Hình 3.28: Điền giá trị cho các cột mới

```

199 -- Xóa cột gốc
200 alter table order_items
201 drop column shipping_limit_date
% < Messages

```

Commands completed successfully.

Completion time: 2024-11-30T02:17:15.1495018+07:00

*Hình 3.29: Xóa cột gốc***Bảng order_payments**

```

210 -- Chuyển đổi payment_value thành 2 chữ số thập phân
211 update order_payments
212 set payment_value = ROUND(payment_value, 2)
213
% < Messages

```

(103886 rows affected)

Completion time: 2024-11-30T02:19:02.1041600+07:00

*Hình 3.30: Chuyển đổi payment_value thành 2 chữ số thập phân***Bảng products và product_category_name**

```

222 -- Đổi tên cột tên nhóm sản phẩm (tiếng Tây Ban Nha, tiếng Anh)
223 EXEC sp_rename 'dbo.product_category_name.column1', 'product_category_name', 'COLUMN';
224 EXEC sp_rename 'dbo.product_category_name.column2', 'product_category_name_english', 'COLUMN';
225
%
```

Messages

Caution: Changing any part of an object name could break scripts and stored procedures.
Caution: Changing any part of an object name could break scripts and stored procedures.

Completion time: 2024-11-30T02:19:25.5224551+07:00

Hình 3.31: Đổi tên cột tên nhóm sản phẩm (tiếng Tây Ban Nha, tiếng Anh)

```

226 -- Xóa tiêu đề cột khỏi hàng đầu tiên
227 DELETE FROM product_category_name
228 WHERE product_category_name = 'product_category_name'
229
230 DELETE FROM product_category_name
231 WHERE product_category_name_english = 'product_category_name_english'

```

2 %

Messages

(1 row affected)

(0 rows affected)

Completion time: 2024-11-30T02:19:40.8091464+07:00

Hình 3.32: Xóa tiêu đề cột khỏi hàng đầu tiên

```

233 -- Thay thế dấu gạch dưới bằng dấu cách
234 update product_category_name
235 set product_category_name_english = REPLACE(product_category_name_english, '_', ' ')
236
237 -- Kiểm tra xem đã thay đổi đúng chưa

```

Messages

(71 rows affected)

Completion time: 2024-11-30T02:19:53.7375185+07:00

Hình 3.33: Thay thế dấu gạch dưới bằng dấu cách

```

236
237 -- Kiểm tra xem đã thay đổi đúng chưa
238 select * from product_category_name
239
240 -- Join 2 bảng để lấy ra tên tiếng Anh

```

152 %

Results Messages

	product_category_name	product_category_name_english
1	beleza_saude	health beauty
2	informatica_acessorios	computers accessories
3	automotivo	auto
4	cama_mesa_banho	bed bath table

Hình 3.34: Kiểm tra xem đã thay đổi đúng chưa

```

240 -- Join 2 bảng để lấy ra tên tiếng Anh
241 -- Thêm 1 cột tên mới
242 alter table products
243 add product_category_eng_name nvarchar(50)
244
245 update products
246 set product_category_eng_name = p2.product_category_name_english
247 from products p1
248 join product_category_name p2 ON p1.product_category_name = p2.product_category_name
249

```

Messages

(32328 rows affected)

Completion time: 2024-11-30T02:20:37.2895932+07:00

Hình 3.35: Join 2 bảng để lấy ra tên tiếng Anh

```

250 -- Viết hoa product_category_eng_name
251 update products
252 set product_category_eng_name = UPPER(SUBSTRING(product_category_eng_name, 1, 1)) + LOWER(SUBSTRING(

```

Messages

(32951 rows affected)

Completion time: 2024-11-30T02:20:52.6999004+07:00

*Hình 3.36: Viết hoa product_category_eng_name***Bảng sellers**

```

261 -- Viết hoa seller_city
262 update sellers
263 set seller_city= UPPER(SUBSTRING(seller_city, 1, 1)) + LOWER(SUBSTRING(seller_city, 2, LEN(seller_cit
264

```

Messages

(3095 rows affected)

Completion time: 2024-11-30T02:21:36.7766559+07:00

Hình 3.37: Viết hoa seller_city

```

265 -- Đổi thành tên đầy đủ
266 UPDATE sellers
267 SET seller_state =
268 CASE seller_state
269 WHEN 'AC' THEN 'Acre'
270 WHEN 'AL' THEN 'Alagoas'

```

messages

(3095 rows affected)

Completion time: 2024-11-30T02:21:54.7662605+07:00

Hình 3.38: Đổi thành tên đầy đủ

Dữ liệu trước và sau làm sạch

Bảng customers

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
1	06b8999e2fba1a1fb88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409	franca	SP
2	18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	9790	sao bernardo do campo	SP
3	4e7b3e00288586ebd08712fd0374a03	060e732b5b29e8181a18229c7b0b2b5e	1151	sao paulo	SP
4	b2b6027bc5c5109e529d4dc6358b12c3	259dac757896d24d7702b9acbbff3f3c	8775	mogi das cruzes	SP
5	4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d066	13056	campinas	SP

Hình 3.39: Bảng customer trước làm sạch

	customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
1	00012a2ce6f3dca20d059ce98491703	248fe10d632bebe4f7267f1f44844c9	6273	Osasco	Sao Paulo
2	000161a058600d5901f007ab4c27140	b0015e09b4b6e47c52844fab5fb6638	35550	Itapecerica	Minas Gerais
3	0001fd6190edaaf884bcacf3d49edf079	94b11d37cd61cb2994a194d11f89682b	29830	Nova venecia	Espirito Santo
4	0002414f95344307404f0ace7a26f1d5	4893ad4ea28b2c5b3ddf4e82e79db9e6	39664	Mendonca	Minas Gerais
5	000379cdec625522490c315e70c7a9fb	0b83f73b19c2019e182fd552c048a22c	4841	Sao paulo	Sao Paulo

Hình 3.40: Bảng customer sau làm sạch

Bảng geolocation

	geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state
1	1037	-23.5456212812	-46.6392920480	sao paulo	SP
2	1046	-23.5460811270	-46.6448202984	sao paulo	SP
3	1046	-23.5461289664	-46.6429514836	sao paulo	SP
4	1041	-23.5443921649	-46.6394993063	sao paulo	SP
5	1035	-23.5415779617	-46.6416072233	sao paulo	SP

Hình 3.41: Bảng geolocation trước làm sạch

	geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state
57	64900	-9.07080936431885	-44.356819152832	Bom jesus	Piaui
58	64910	-8.94847393035889	-44.1308288574219	Santa luz	Piaui
59	64900	-9.07709980010986	-44.3620872497559	Bom jesus	Piaui
60	64965	-10.1382436752319	-43.9450912475586	Avelino lopes	Piaui
61	64900	-9.06859493255615	-44.3638954162598	Bom jesus	Piaui
62	64980	-10.4415760040283	-45.1667404174805	Corrente	Piaui
63	64900	-9.07069492340088	-44.3597297668457	Bom jesus	Piaui
64	64940	-9.74747562408447	-45.3046875	Monte alegre do p...	Piaui
65	64945	-9.1123685836792	-45.9205856323242	Santa filomena	Piaui
66	64965	-10.1377735137939	-43.9495620727539	Avelino lopes	Piaui
67	64930	-9.83220481872559	-45.3354835510254	Gilbues	Piaui
68	64965	-10.1308546066284	-43.9486045837402	Avelino lopes	Piaui
69	64925	-8.72849178314209	-44.2398948669434	Palmeira do piaui	Piaui
70	64900	-9.06809139251709	-44.3650360107422	Bom jesus	Piaui
71	64980	-10.2814598083496	-45.1762161254883	Corrente	Piaui
72	64900	-9.07259941101074	-44.3656120300293	Bom jesus	Piaui
73	64965	-10.1308546066284	-43.9486045837402	Avelino lopes	Piaui
74	64980	-10.4401197433472	-45.1714897155762	Corrente	Piaui
75	65020	-2.52442717552185	-44.2904777526855	Sao luis	Maranhao
76	65031	-2.55006194114685	-44.2878379821777	Sao luis	Maranhao

Hình 3.42: Bảng geolocation sau làm sạch

Bảng order item

	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date
1	6481f151bcd54678b7cc49139f2d6a7	9ef432eb625129730476186b10a928d	delivered	2017-10-02 10:56:33.000000	2017-10-02 11:17:15.000000	2017-10-04 19:55:00.000000	2017-10-10 21:25:13.000000	2017-10-18 00:00:00
2	53cd2c2c8bc7dc0e6741e2150273451	b0830b4747a6c6d20dea0b8c802d7ef	delivered	2018-07-24 20:41:37.000000	2018-07-26 03:24:27.000000	2018-07-26 14:31:00.000000	2018-08-07 15:27:45.000000	2018-08-13 00:00:00
3	47770eb5100c2df044946d9c07ec65d	41ce2a54c0b03fc3443c3s931a367089	delivered	2018-08-08 08:38:49.000000	2018-08-08 08:55:23.000000	2018-08-08 13:50:00.000000	2018-08-17 18:06:29.000000	2018-09-04 00:00:00
4	949d5d44dbf5de918fe9c1897b4f98a	f88197465ea7920a0ddcbe7375364d82	delivered	2017-11-18 19:28:06.000000	2017-11-18 19:45:59.000000	2017-11-22 13:39:59.000000	2017-12-02 00:28:42.000000	2017-12-15 00:00:00
5	ad21c59c0840e6ba83a9ce5573f8159	8ab97904e6daea8866dbdbc4fb7aad2c	delivered	2018-02-13 21:18:39.000000	2018-02-13 22:20:29.000000	2018-02-14 19:46:34.000000	2018-02-16 18:17:02.000000	2018-02-26 00:00:00

Hình 3.43: Bảng order items trước làm sạch

	order_id	order_item_id	product_id	seller_id	price	freight_value	shipping_limit_dateee	shipping_limit_time
1	00010242f8c5a6d1ba2dd792cb16214	1	4244733e06e7ecb4970a6e2683c13e61	48436dade18ac8b2bce089ec2a041202	58.9	13.29	2017-09-19	09:45:35.000000
2	0001877120320c557190d7a144bd3	1	e5f2d5b2082189ee658865ca93d83a8f	dd7ddc04e1b6c2c614352b383efe2d36	239.9	19.93	2017-05-03	11:05:13.000000
3	000229e3c398224ef6ca0657da4fc703e	1	c777355a18b7267abbee9df44f0fd	5b51032eddd242ad84c83cab8f98f23d	199	17.87	2018-01-18	14:48:30.000000
4	00024acbcdf0a6da1e931b038114c75	1	7634da152a4610f1595efa32f14722fc	9d7a1d34a5052409006425275ba1c2b4	12.99	12.79	2018-08-15	10:10:18.000000
5	00042b26cf59d7ce69dfabb4e55b4f9	1	ac6c3623069f30de03045865e4e10089	df56039f3a51e74553ab94004ba5c87	199.9	18.14	2017-02-13	13:57:51.000000

Hình 3.44: Bảng order_items sau làm sạch

Bảng Order reviews

	review_id	order_id	review_score	review_comment_title	review_comment_message	review_creation_date	review_answer_timestamp
1	7bc2406110b92639aa5980a40eba40	73c7af0711493712e6da79b0a377eb	4	NULL	NULL	2018-01-18 00:00:00.000000	2018-01-18 21:46:59.000000
2	80e641a11e59f04c1ad469d5645fdfe	a548910a1c6147796b98df73dbeba33	5	NULL	NULL	2018-03-10 00:00:00.000000	2018-03-11 03:05:13.000000
3	228ce5500dc1d8e020d813228746f0	f9e4b658b201a9f2ecdecbb34bed034b	5	NULL	NULL	2018-02-17 00:00:00.000000	2018-02-18 14:36:24.000000
4	e64bf393e7b32834bb789ff8bb30750e	658677c97b3059be17073785d3511b	5	NULL	Recebi bem antes do prazo estipulado.	2017-04-21 00:00:00.000000	2017-04-21 22:02:06.000000
5	f7c4243c7e1938f181bec41a392deb	8e6fb81e283fa7e4f11123a3b894f1	5	NULL	Parabéns lojas Iannister adorei comprar pela Int...	2018-03-01 00:00:00.000000	2018-03-02 10:26:53.000000

Hình 3.45: Bảng order reviews trước làm sạch

Results Messages										
order_status	order_purchase_date	order_purchase_time	order_approved_date	order_approved_time	order_delivered_to_carrier_date	order_delivered_to_carrier_time	order_delivered_to_customer_date	order_delivered_to_customer_time	order_delivered_to_customer	order_delivered_to_customer^
1 delivered	2017-10-02	10:56:33.000000	2017-10-02	11:07:15.000000	2017-10-04	19:55:00.000000	2017-10-10	21:25:13.000000		
2 delivered	2018-07-24	20:41:37.000000	2018-07-26	03:24:27.000000	2018-07-26	14:31:00.000000	2018-08-07	15:27:45.000000		
3 delivered	2018-08-08	08:38:49.000000	2018-08-08	08:55:23.000000	2018-08-08	13:50:00.000000	2018-08-17	18:06:29.000000		
4 delivered	2017-11-18	19:28:06.000000	2017-11-18	19:45:59.000000	2017-11-22	13:39:59.000000	2017-12-02	00:28:42.000000		
5 delivered	2018-02-13	21:18:39.000000	2018-02-13	22:20:29.000000	2018-02-14	19:46:34.000000	2018-02-16	18:17:02.000000		

Hình 3.46: Bảng order reviews sau làm sạch

Bảng order payments

Results Messages				
order_id	payment_sequential	payment_type	payment_installments	payment_value
b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.3300018310547
a9810da82917af2d9aef1278f1dcfa0	1	credit_card	1	24.3899993896484
25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	1	65.7099990844727
ba78997921bbcdc1373bb41e913ab953	1	credit_card	8	107.779998779297
42fdf880ba16b47b59251dd489d4441a	1	credit_card	2	128.449996948242

Hình 3.47: Bảng order payment trước làm sạch

Results Messages					
order_id	payment_sequential	payment_type	payment_installments	payment_value	payment_value^
b81ef226f3fe1789b1e8b2acac839d17	1	credit_card	8	99.33	99.3300018310547
a9810da82917af2d9aef1278f1dcfa0	1	credit_card	1	24.39	24.3899993896484
25e8ea4e93396b6fa0d3dd708e76c1bd	1	credit_card	1	65.71	65.7099990844727
ba78997921bbcdc1373bb41e913ab953	1	credit_card	8	107.78	107.779998779297
42fdf880ba16b47b59251dd489d4441a	1	credit_card	2	128.45	128.449996948242

Hình 3.48: Bảng order payment sau làm sạch

Bảng product category

Results Messages									
product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm	product_weight_cm
1 0006f942aeeb9f3007548bb9d3f33c38	perfumaria	53	596	6	300	20	16	16	
2 0008930e925c41fd95ebfe699fd2655	automotivo	56	752	4	1225	55	10	26	
3 000940fd7479715e4bef61dd91f2462	cama_mesa_banho	50	266	2	300	45	15	35	
4 00b8f95fc9e0096488278317764d19	utilidades_domesticas	25	364	3	550	19	24	12	
5 000d9be29b5207b54e86aa1b1ac54872	relogios_presentes	48	613	4	250	22	11	15	

Hình 3.49: Bảng product category trước làm sạch

Results Messages									
product_id	product_category_name	product_category_eng_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_length_cm	product_height_cm	product_width_cm
1 0006f942aeeb9f3007548bb9d3f33c38	perfumaria	Perfumery	53	596	6	300	20	16	16
2 0008930e925c41fd95ebfe699fd2655	automotivo	Auto	56	752	4	1225	55	10	26
3 000940fd7479715e4bef61dd91f2462	cama_mesa_banho	Bed bath table	50	266	2	300	45	15	35
4 00b8f95fc9e0096488278317764d19	utilidades_domesticas	Housewares	25	364	3	550	19	24	12
5 000d9be29b5207b54e86aa1b1ac54872	relogios_presentes	Watches gifts	48	613	4	250	22	11	15

Hình 3.50: Bảng product category sau làm sạch

3.3 CHUYỂN ĐỔI DỮ LIỆU

3.3.1 CÁC TRƯỜNG HỢP CẦN CHUYỂN ĐỔI

Chuyển đổi dữ liệu là một bước quan trọng trong quá trình làm sạch và chuẩn bị dữ liệu. Các trường hợp cần chuyển đổi dữ liệu thường gặp bao gồm:

Chuyển đổi kiểu dữ liệu:

- Chuyển đổi cột "order_purchase_timestamp" trong bảng Orders từ dạng văn bản (string) sang kiểu dữ liệu ngày tháng (datetime) để thuận tiện cho việc phân tích theo thời gian.
- Chuyển đổi cột "review_score" trong bảng Order Reviews từ dạng số thực (float) sang dạng số nguyên (integer) vì điểm đánh giá chỉ nhận các giá trị nguyên.

Mã hóa lại các giá trị phân loại:

- Chuyển đổi các giá trị trong cột "order_status" (trạng thái đơn hàng) trong bảng Orders thành các giá trị số để dễ dàng phân tích và trực quan hóa. Ví dụ: "delivered" = 1, "shipped" = 2, "canceled" = 3,...
- Mã hóa các giá trị trong cột "payment_type" (loại hình thanh toán) trong bảng Order Payments thành các giá trị số tương ứng.

Chuẩn hóa dữ liệu:

- Chuẩn hóa các giá trị trong cột "price" và "freight_value" trong bảng Order Items về cùng một đơn vị tiền tệ (ví dụ: BRL) để thuận tiện cho việc tính toán và so sánh.
- Chuẩn hóa các giá trị thời gian trong các bảng Orders và Order Reviews về cùng một định dạng (ví dụ: YYYY-MM-DD HH:MM:SS).

Xử lý dữ liệu thời gian:

- Tách các thành phần của "order_purchase_timestamp" trong bảng Orders thành các cột riêng biệt: năm, tháng, ngày, giờ, phút, giây để dễ dàng phân tích xu hướng theo thời gian.
- Tạo ra các biến thời gian mới, chẳng hạn như "thời gian giao hàng" (tính bằng số ngày) bằng cách lấy hiệu giữa "order_delivered_customer_date" và "order_purchase_timestamp".

Tạo các biến mới từ biến hiện có:

- Tính toán tổng giá trị đơn hàng bằng cách cộng "price" và "freight_value" trong bảng Order Items.
- Tính toán doanh thu theo từng danh mục sản phẩm ("product_category_name") dựa trên dữ liệu từ bảng Products và Order Items.

Xử lý giá trị trống hoặc giá trị mặc định:

- Thay thế các giá trị trống trong cột "review_comment_title" và "review_comment_message" trong bảng Order Reviews bằng giá trị "Không có đánh giá".
- Điều chỉnh các giá trị trống trong cột "product_weight_g" và "product_length_cm" trong bảng Products bằng giá trị trung bình của các sản phẩm cùng loại.

Gộp hoặc chia cột dữ liệu:

- Gộp các cột "customer_city" và "customer_state" trong bảng Customers thành một cột "customer_location".

- Chia cột "product_photos_qty" trong bảng Products thành các cột riêng biệt cho từng ảnh sản phẩm.

Loại bỏ hoặc chuyển đổi các ký tự không mong muốn:

- Loại bỏ các ký tự đặc biệt hoặc khoảng trắng thừa trong các cột chứa thông tin văn bản, chẳng hạn như "product_category_name" và "customer_city".
- Chuyển đổi tất cả các ký tự trong các cột văn bản thành chữ thường hoặc chữ hoa để đảm bảo tính nhất quán.

3.3.2 CÁC KỸ THUẬT CHUYỂN ĐỔI

Kỹ thuật chuyển đổi kiểu dữ liệu (Data Type Conversion) là một quy trình quan trọng trong quá trình xử lý và chuẩn bị dữ liệu để đảm bảo tính nhất quán và chính xác của dữ liệu. Trong phân tích dữ liệu và khoa học dữ liệu, dữ liệu thường đến từ nhiều nguồn khác nhau với các định dạng và kiểu dữ liệu khác nhau, đòi hỏi việc chuyển đổi để có thể xử lý và phân tích một cách hiệu quả.

Chuyển đổi kiểu dữ liệu bao gồm việc chuyển đổi giá trị của một biến từ kiểu dữ liệu này sang kiểu dữ liệu khác. Ví dụ, dữ liệu về giá cả có thể được lưu trữ dưới dạng chuỗi (string) nhưng cần chuyển đổi sang số thập phân (float) để thực hiện các phép tính toán học. Tương tự, dữ liệu ngày tháng có thể được lưu dưới dạng chuỗi và cần chuyển đổi sang kiểu ngày tháng (datetime) để phân tích theo thời gian.

Một số kỹ thuật chuyển đổi kiểu dữ liệu phổ biến bao gồm:

Chuyển đổi kiểu số (Numeric Conversion):

- Chuyển đổi cột "payment_installments" (Số lần trả góp) trong bảng Order Payments từ kiểu số thực (float) sang số nguyên (integer) vì số lần trả góp luôn là số nguyên.
- Chuyển đổi cột "review_score" (Điểm đánh giá) trong bảng Order Reviews từ kiểu số thực (float) sang số nguyên (integer) vì điểm đánh giá chỉ nhận giá trị nguyên từ 1 đến 5.

Chuyển đổi kiểu chuỗi (String Conversion):

- Chuyển đổi cột "customer_id" và "order_id" trong các bảng Customers, Orders, Order Items, Order Payments và Order Reviews từ kiểu số nguyên (integer) sang chuỗi (string) để thuận tiện cho việc nối chuỗi và tra cứu.
- Chuyển đổi các giá trị trong cột "order_status" (Trạng thái đơn hàng) trong bảng Orders sang dạng chuỗi để dễ dàng đọc và hiểu, ví dụ: "delivered", "shipped", "canceled",...

Chuyển đổi kiểu ngày tháng (Date Conversion):

- Chuyển đổi các cột thời gian trong bảng Orders (như "order_purchase_timestamp", "order_approved_at", "order_delivered_carrier_date", "order_delivered_customer_date", "order_estimated_delivery_date") từ dạng chuỗi (string) sang kiểu dữ liệu ngày tháng (datetime) để thực hiện các phép tính và phân tích theo thời gian.

- Chuyển đổi các cột thời gian trong bảng Order Reviews (như "review_creation_date", "review_answer_timestamp") từ dạng chuỗi sang kiểu dữ liệu ngày tháng.

Chuyển đổi kiểu phân loại (Categorical Conversion):

- Chuyển đổi các giá trị trong cột "payment_type" (Loại hình thanh toán) trong bảng Order Payments thành các giá trị số tương ứng, ví dụ: "credit_card" = 1, "boleto" = 2, "voucher" = 3,...
- Chuyển đổi các giá trị trong cột "lead_type" (Loại khách hàng tiềm năng) và "lead_behaviour_profile" (Hồ sơ hành vi) trong bảng Marketing Qualified Leads thành các giá trị số để thuận tiện cho việc phân tích.

Kỹ thuật chuyển đổi kiểu dữ liệu không chỉ giúp đảm bảo tính nhất quán của dữ liệu mà còn giúp tăng cường khả năng phân tích và trực quan hóa dữ liệu, từ đó cung cấp những thông tin chính xác và có giá trị cho việc ra quyết định.

Trong phạm vi dự án này, các kỹ thuật chuyển đổi sau sẽ được áp dụng:

Chuyển Đổi Kiểu Dữ Liệu

- Chuyển đổi ngày tháng từ chuỗi ký tự sang kiểu datetime.
- Chuyển đổi các giá trị số từ chuỗi ký tự sang kiểu số nguyên hoặc số thực.

Chuyển Đổi Định Dạng Dữ Liệu

- Định dạng ngày tháng để đảm bảo thống nhất.
- Định dạng số tiền để đảm bảo thống nhất.

Chuyển Đổi Giá Trị Phân Loại

- Chuyển đổi tất cả các giá trị phân loại về chữ thường hoặc chữ hoa.
- Thay thế các giá trị phân loại không nhất quán hoặc không chính xác.

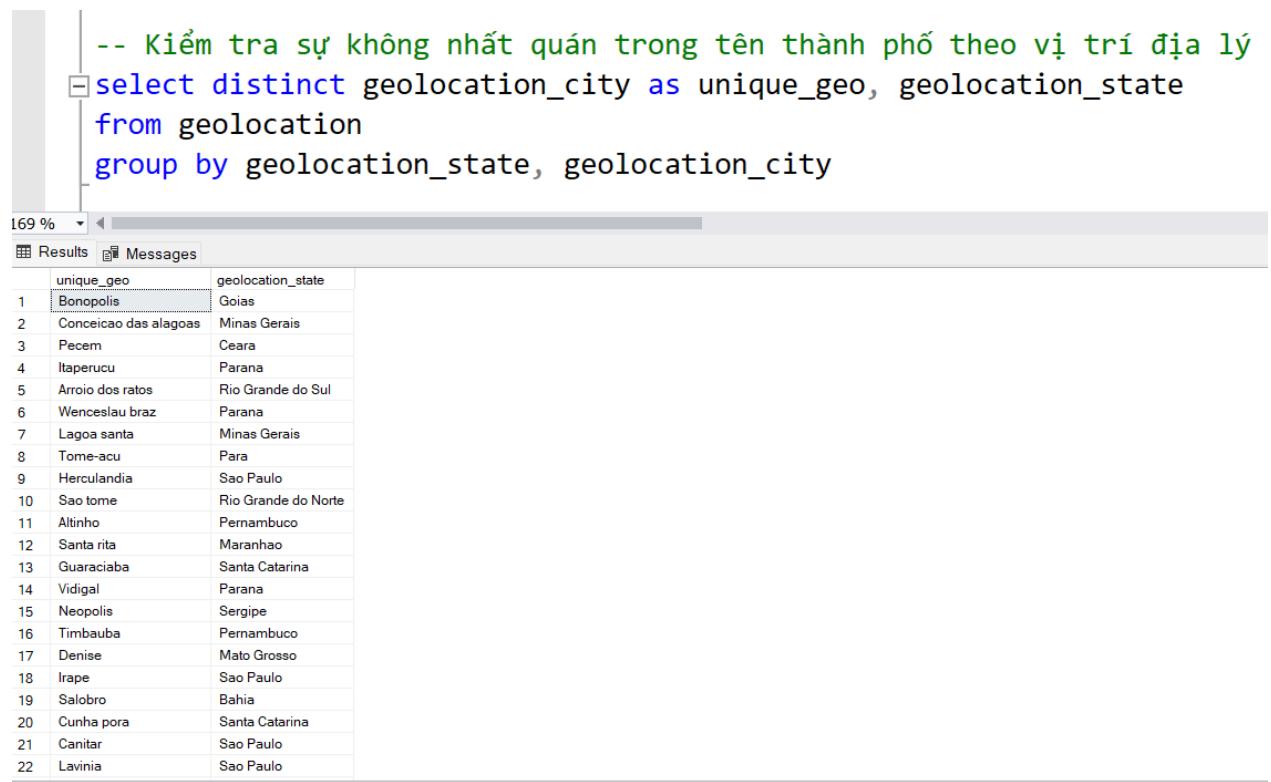
Tính Toán và Chuyển Đổi Các Chỉ Số

- Tính toán các chỉ số mới từ các cột hiện có.
- Chuyển đổi các chỉ số sang các đơn vị đo lường khác nhau nếu cần thiết.

Xử lý dữ liệu bị thiếu

- Xử lý các giá trị bị thiếu trong dữ liệu bằng cách thay thế bằng giá trị trung bình, giá trị phổ biến nhất, hoặc sử dụng các phương pháp nội suy.

3.3.3 TRÌNH BÀY CÁC PHÉP CHUYỂN ĐỔI TRONG DỰ ÁN



```
-- Kiểm tra sự không nhất quán trong tên thành phố theo vị trí địa lý
select distinct geolocation_city as unique_geo, geolocation_state
from geolocation
group by geolocation_state, geolocation_city
```

The screenshot shows a SQL query being run in SSMS. The results are displayed in a table titled 'Results'.

unique_geo	geolocation_state
1 Bonopoli	Goias
2 Conceicao das alagoas	Minas Gerais
3 Pecem	Ceara
4 Itaperucu	Parana
5 Arroio dos ratos	Rio Grande do Sul
6 Wenceslau braz	Parana
7 Lagoa santa	Minas Gerais
8 Tome-acu	Para
9 Herculandia	Sao Paulo
10 Sao tome	Rio Grande do Norte
11 Altinho	Pernambuco
12 Santa rita	Maranhao
13 Guaraciaba	Santa Catarina
14 Vidigal	Parana
15 Neopolis	Sergipe
16 Timbauba	Pernambuco
17 Denise	Mato Grosso
18 Irape	Sao Paulo
19 Salobro	Bahia
20 Cunha pora	Santa Catarina
21 Canitar	Sao Paulo
22 Lavinia	Sao Paulo

Hình 3.51: Kiểm tra sự nhát quán trong tên thành phố theo vị trí địa lý

4 XỬ LÝ DỮ LIỆU

4.1 CHUẨN HÓA DỮ LIỆU

4.1.1 TRÌNH BÀY CÁC BƯỚC CHUẨN HÓA TRONG DỰ ÁN

Trong dự án này, quá trình chuẩn hóa dữ liệu được thực hiện bằng Power Query để đảm bảo dữ liệu có tính nhất quán, chính xác và sẵn sàng cho việc phân tích dữ liệu bán hàng trên nền tảng thương mại điện tử Brazil. Đầu tiên, dữ liệu được tải vào Power Query từ SQL Server và mở trong Power Query Editor. Sau đó, các giá trị thiếu trong các cột quan trọng như order_date (Ngày đặt hàng), product_category_name (Danh mục sản phẩm) được xác định và thay thế bằng giá trị mặc định hoặc giá trị ước lượng. Tiếp theo, các giá trị không hợp lệ trong cột customer_state (Tình của khách hàng) và customer_city (Thành phố của khách hàng) được kiểm tra và sửa chữa để đảm bảo tính chính xác về địa lý. Để loại bỏ các giá trị ngoại lai, các giá trị quá lớn hoặc quá nhỏ so với khoảng giá trị trung bình trong các cột như payment_value (Giá trị thanh toán) được loại bỏ. Ngoài ra, các sản phẩm có giá trị âm hoặc giá trị quá thấp so với mặt bằng chung cũng được xem xét kỹ lưỡng. Sau khi làm sạch, các cột dữ liệu số được chuyển đổi về kiểu dữ liệu số (numeric) và các cột ngày tháng được chuyển đổi về kiểu dữ liệu ngày (date) để thuận tiện cho việc phân tích. Sau khi hoàn tất các bước chuẩn hóa, dữ liệu được lưu và tải lại vào Power BI để tiếp tục phân tích.

4.2 MÔ HÌNH HÓA DỮ LIỆU

4.2.1 CÁC LOẠI MÔ HÌNH HÓA

Mô hình hóa dữ liệu là quá trình tạo ra các biểu đồ và cấu trúc dữ liệu để thể hiện mối quan hệ giữa các thành phần dữ liệu trong hệ thống. Có nhiều loại mô hình hóa dữ liệu, mỗi loại phục vụ các mục đích khác nhau trong việc thiết kế và quản lý cơ sở dữ liệu. Các loại mô hình hóa dữ liệu chính bao gồm mô hình khái niệm, mô hình logic và mô hình vật lý.

- **Mô hình khái niệm (Conceptual Data Model):** Đây là mô hình cấp cao nhất, tập trung vào việc định nghĩa các thực thể và mối quan hệ giữa chúng mà không quan tâm đến chi tiết kỹ thuật. Nó thường được sử dụng trong giai đoạn đầu của dự án để phác thảo cấu trúc dữ liệu và xác định các yêu cầu nghiệp vụ.
- **Mô hình logic (Logical Data Model):** Mô hình này chi tiết hơn mô hình khái niệm và tập trung vào cách các thực thể và mối quan hệ được biểu diễn dưới dạng bảng và cột mà không phụ thuộc vào hệ quản trị cơ sở dữ liệu cụ thể. Nó bao gồm các ràng buộc, khóa chính và khóa ngoại để đảm bảo tính toàn vẹn dữ liệu.
- **Mô hình vật lý (Physical Data Model):** Đây là mô hình chi tiết nhất, bao gồm cấu trúc lưu trữ thực tế của dữ liệu trong hệ quản trị cơ sở dữ liệu cụ thể. Mô hình này định nghĩa chi tiết về cách dữ liệu được lưu trữ, chỉ mục, phân vùng, và các yếu tố kỹ thuật khác.

Trong dự án phân tích dữ liệu của Olist, cả mô hình logic và mô hình vật lý đều đóng vai trò quan trọng trong việc thiết kế và triển khai cơ sở dữ liệu. Mô hình logic tập trung vào việc biểu diễn các thực thể kinh doanh và mối quan hệ giữa chúng mà không phụ thuộc vào hệ quản trị cơ sở dữ liệu cụ thể. Ví dụ, mô hình logic sẽ xác định các thực thể như "Khách hàng" (Customers), "Đơn hàng" (Orders), "Sản phẩm" (Products), và mối quan hệ giữa chúng (một khách hàng có thể có nhiều đơn hàng, một đơn hàng có thể chứa nhiều sản phẩm).

Mô hình vật lý kế thừa từ mô hình logic và được triển khai trên hệ quản trị cơ sở dữ liệu SQL Server. Mô hình vật lý tập trung vào việc tổ chức lưu trữ dữ liệu trên ổ đĩa, tối ưu hóa hiệu suất truy vấn và đảm bảo tính toàn vẹn dữ liệu. Ví dụ, mô hình vật lý sẽ xác định kiểu dữ liệu của từng cột (như cột "order_purchase_timestamp" trong bảng Orders sẽ có kiểu dữ liệu datetime), các ràng buộc trên các cột (cột "order_id" là khóa chính), và các chỉ mục để tăng tốc độ truy vấn.

Sử dụng kết hợp mô hình logic và vật lý mang lại nhiều lợi ích cho dự án, bao gồm:

- **Chuẩn hóa dữ liệu:** Giúp giảm thiểu dữ liệu trùng lặp và đảm bảo tính nhất quán.
- **Tối ưu hóa hiệu suất:** Cải thiện tốc độ truy vấn và truy xuất dữ liệu.
- **Dễ dàng bảo trì:** Đơn giản hóa việc bảo trì và cập nhật cơ sở dữ liệu.
- **Tăng khả năng mở rộng:** Giúp cơ sở dữ liệu dễ dàng mở rộng khi dữ liệu tăng lên.

Tóm lại, mô hình logic và mô hình vật lý là những bước thiết yếu trong việc xây dựng một hệ thống cơ sở dữ liệu hiệu quả cho Olist, đảm bảo dữ liệu được tổ chức một cách logic và tối ưu hóa cho việc phân tích và ra quyết định kinh doanh.

4.2.2 CÁC TIÊU CHÍ ĐÁNH GIÁ MÔ HÌNH DỮ LIỆU

Các tiêu chí đánh giá mô hình dữ liệu tốt

Một mô hình dữ liệu tốt cần đáp ứng nhiều tiêu chí khác nhau để đảm bảo tính hiệu quả, toàn vẹn, và khả năng mở rộng. Dưới đây là các tiêu chí đánh giá một mô hình dữ liệu tốt:

Tính chính xác (Accuracy)

- Mô tả: Mô hình phải phản ánh đúng thực tế và yêu cầu nghiệp vụ mà nó được thiết kế để hỗ trợ.
- Các bảng như Customers, Orders, Products, Sellers phải đại diện chính xác cho các khía cạnh kinh doanh của Olist và mối quan hệ giữa chúng (khóa chính, khóa ngoại) phải thể hiện chính xác các quy trình và luồng công việc thực tế.

Tính toàn vẹn (Integrity)

- Mô tả: Dữ liệu phải được bảo vệ khỏi các lỗi và sự thiếu nhất quán. Điều này bao gồm việc sử dụng các ràng buộc toàn vẹn, khóa chính và khóa ngoại.
- Ví dụ: Sử dụng các ràng buộc toàn vẹn trong SQL Server để đảm bảo dữ liệu không bị trùng lặp hoặc mất mát. Mỗi đơn hàng trong bảng Orders phải có order_id duy nhất, và các quan hệ giữa bảng Orders và Order Items (qua order_id) phải được duy trì.

Tính nhất quán (Consistency)

- Mô tả: Dữ liệu phải nhất quán trong toàn bộ hệ thống, không có sự mâu thuẫn giữa các bảng và các trường dữ liệu.
- Ví dụ: Định dạng ngày tháng trong các cột order_purchase_timestamp, order_delivered_customer_date phải nhất quán, đơn vị tiền tệ trong các cột price, freight_value phải giống nhau (BRL).

Tính đầy đủ (Completeness)

- Mô tả: Mô hình phải bao gồm tất cả các thực thể và mối quan hệ cần thiết để đáp ứng yêu cầu nghiệp vụ.
- Ví dụ: Thông tin về khách hàng, đơn hàng, sản phẩm, người bán phải được lưu trữ đầy đủ trong các bảng tương ứng.

Tính mở rộng (Scalability)

- Mô tả: Mô hình phải có khả năng mở rộng để đáp ứng nhu cầu tăng trưởng của dữ liệu và hệ thống mà không cần thay đổi cấu trúc cơ bản.
- Ví dụ: Dễ dàng thêm các bảng mới hoặc mở rộng các bảng hiện có mà không làm gián đoạn hệ thống.

Tính hiệu quả (Efficiency)

- Mô tả: Mô hình phải hỗ trợ truy vấn và xử lý dữ liệu hiệu quả, giảm thiểu thời gian truy xuất và sử dụng tài nguyên hệ thống.
- Ví dụ: Tạo chỉ mục trên các cột customer_id, product_id, seller_id để tăng tốc độ truy vấn.

Tính dễ hiểu (Understandability)

- Mô tả: Mô hình phải dễ hiểu và dễ sử dụng cho các nhà phát triển, nhà quản lý dữ liệu và người dùng cuối.
- Ví dụ: Các tên bảng và cột phải rõ ràng và dễ hiểu.

Mô hình đang dùng trong dự án đáp ứng các tiêu chí nào?

Mô hình dữ liệu hiện tại trong dự án đã đáp ứng được hầu hết các tiêu chí quan trọng như tính chính xác, tính toàn vẹn, tính nhất quán, tính đầy đủ, tính mở rộng và tính hiệu quả. Điều này đảm bảo rằng cơ sở dữ liệu không chỉ chính xác và nhất quán mà còn có khả năng mở rộng và hiệu quả trong việc xử lý dữ liệu, hỗ trợ tốt cho các yêu cầu phân tích và báo cáo của dự án.

Tính chính xác (Accuracy): Mô hình phản ánh chính xác các thực thể và mối quan hệ trong dữ liệu.

Tính toàn vẹn (Integrity): Mô hình sử dụng các khóa chính và khóa ngoại để đảm bảo tính toàn vẹn dữ liệu. Các ràng buộc toàn vẹn đảm bảo rằng không có dữ liệu trùng lặp hoặc mất mát trong quá trình nhập và xử lý dữ liệu.

Tính nhất quán (Consistency): Định dạng và kiểu dữ liệu được chuẩn hóa trong toàn bộ cơ sở dữ liệu. Ví dụ, các cột ngày tháng và số liệu được định dạng nhất quán, đảm bảo tính nhất quán dữ liệu.

Tính đầy đủ (Completeness): Mô hình bao gồm tất cả các thông tin cần thiết.

Tính mở rộng (Scalability): Cấu trúc bảng và mối quan hệ giữa chúng cho phép mở rộng dễ dàng. Ví dụ, có thể thêm các bảng mới hoặc mở rộng các bảng hiện có mà không làm gián đoạn hệ thống.

Tính hiệu quả (Efficiency): Mô hình được thiết kế để hỗ trợ truy vấn và xử lý dữ liệu hiệu quả. Các chỉ mục và thiết kế bảng tối ưu giúp giảm thiểu thời gian truy xuất và sử dụng tài nguyên hệ thống.

4.2.3 TRÌNH BÀY CÁC BƯỚC MÔ HÌNH HÓA

Quá trình phân tích và xác định các bảng cần thiết giúp chuẩn hóa cơ sở dữ liệu, loại bỏ dữ liệu dư thừa, tăng tính toàn vẹn và hiệu quả trong quản lý dữ liệu. Việc chia nhỏ dữ liệu thành các bảng và thiết lập mối quan hệ giữa chúng giúp cơ sở dữ liệu dễ bảo trì và mở rộng hơn trong tương lai. Triển khai mô hình vật lý cụ thể với các lệnh SQL giúp đảm bảo rằng cơ sở dữ liệu được cấu trúc một cách chính xác và sẵn sàng cho việc sử dụng trong các ứng dụng thực tế.

Bước 1: Xác định các thực thể và thuộc tính

Khách hàng (Customers):

- Thực thể đại diện cho thông tin khách hàng.
- Thuộc tính chính: customer_id.
- Thuộc tính bổ sung: customer_unique_id, customer_zip_code_prefix, customer_city, customer_state

Đơn hàng (Orders):

- Lưu trữ thông tin về các đơn hàng.
- Thuộc tính chính: order_id.
- Thuộc tính bổ sung: customer_id, order_status, order_purchase_timestamp, order_delivered_customer_date.

Sản phẩm (Products):

- Thông tin chi tiết về sản phẩm.
- Thuộc tính chính: product_id.
- Thuộc tính bổ sung: product_category_name, product_name_length, product_photos_qty.

Danh mục sản phẩm (Product Category Name Translation):

- Tách biệt phần dịch danh mục sản phẩm.
- Thuộc tính chính: product_category_name.
- Thuộc tính bổ sung: product_category_name_english.

Mục hàng trong đơn hàng (Order Items):

- Liệt kê các sản phẩm trong từng đơn hàng.
- Thuộc tính chính: (order_id, order_item_id).
- Thuộc tính bổ sung: product_id, seller_id, price, freight_value

Thanh toán đơn hàng (Order Payments):

- Chi tiết thanh toán cho mỗi đơn hàng.
- Thuộc tính chính: order_id.
- Thuộc tính bổ sung: payment_type, payment_installments, payment_value.

Nhà bán hàng (Sellers):

- Thông tin về người bán hàng.
- Thuộc tính chính: seller_id.
- Thuộc tính bổ sung: seller_zip_code_prefix, seller_city, seller_state.

Khách hàng tiềm năng đủ điều kiện (Leads Qualified):

- Ghi nhận các leads tiềm năng trong hệ thống.
- Thuộc tính chính: mql_id.

- Thuộc tính bổ sung: first_contact_date, origin, landing_page_id.

Khách hàng tiềm năng đã chuyển đổi (Leads Closed):

- Dữ liệu về các leads đã chuyển đổi thành khách hàng.
- Thuộc tính chính: mql_id.
- Thuộc tính bổ sung: seller_id, won_date, business_type, declared_monthly_revenue.

Đánh giá đơn hàng (Order Reviews):

- Chi tiết đánh giá của khách hàng.
- Thuộc tính chính: review_id.
- Thuộc tính bổ sung: order_id, review_score, review_comment_message.

Địa lý (Geolocation):

- Thông tin mã bưu chính và vị trí địa lý.
- Thuộc tính chính: geolocation_zip_code_prefix.
- Thuộc tính bổ sung: geolocation_lat, geolocation_lng.

Bước 2: Xác định các mối quan hệ

<input type="checkbox"/> From: table (column) ↑	Relationship	To: table (column)	Status	
<input type="checkbox"/> closed_deals (mql_id)	1 —>—> 1	marketing_qualified_leads (mq...)	Active	...
<input type="checkbox"/> customers (customer_id)	1 —>—> 1	orders (customer_id)	Active	...
<input type="checkbox"/> customers (customer_zip_code...)	* —>—> *	geolocation (geolocation_zip...)	Active	...
<input type="checkbox"/> order_items (order_id)	* —>—> 1	orders (order_id)	Active	...
<input type="checkbox"/> order_items (product_id)	* —>—> 1	products (product_id)	Active	...
<input type="checkbox"/> order_items (seller_id)	* —>—> 1	sellers (seller_id)	Active	...
<input type="checkbox"/> order_payments (order_id)	* —>—> 1	orders (order_id)	Active	...
<input type="checkbox"/> order_reviews (order_id)	* —>—> 1	orders (order_id)	Active	...
<input type="checkbox"/> products (product_category_n...)	* —>—> 1	product_category_name (prod...)	Active	...
<input type="checkbox"/> sellers (seller_id)	1 —>—> 1	closed_deals (seller_id)	Active	...

Hình 4.1: Mối quan hệ

Bảng orders và bảng customers

- **Mối quan hệ:** 1:N (Một khách hàng có thể có nhiều đơn hàng).
- **Khóa ngoại:** customer_id trong bảng orders tham chiếu đến customer_id trong bảng customers.

Bảng order_items và bảng orders

- **Mối quan hệ:** 1:N (Một đơn hàng có thể chứa nhiều mục hàng).

- **Khóa ngoại:** order_id trong bảng order_items tham chiếu đến order_id trong bảng orders.

Bảng order_items và bảng products

- **Mối quan hệ:** 1:N (Một sản phẩm có thể xuất hiện trong nhiều mục hàng).
- **Khóa ngoại:** product_id trong bảng order_items tham chiếu đến product_id trong bảng products.

Bảng order_items và bảng sellers

- **Mối quan hệ:** 1:N (Một nhà bán hàng có thể bán nhiều sản phẩm qua các mục hàng).
- **Khóa ngoại:** seller_id trong bảng order_items tham chiếu đến seller_id trong bảng sellers.

Bảng order_reviews và bảng orders

- **Mối quan hệ:** 1:1 (Một đơn hàng chỉ có một bài đánh giá).
- **Khóa ngoại:** order_id trong bảng order_reviews tham chiếu đến order_id trong bảng orders.

Bảng order_payments và bảng orders

- **Mối quan hệ:** 1:N (Một đơn hàng có thể được thanh toán qua nhiều phương thức).
- **Khóa ngoại:** order_id trong bảng order_payments tham chiếu đến order_id trong bảng orders.

Bảng product_category_name_translation và bảng products

- **Mối quan hệ:** 1:1 (Một danh mục sản phẩm chỉ ứng với một bản dịch danh mục tiếng Anh).
- **Khóa ngoại:** product_category_name trong bảng products tham chiếu đến product_category_name trong bảng product_category_name_translation.

Bảng geolocation và bảng customers

- **Mối quan hệ:** 1:N (Một mã bưu chính có thể được liên kết với nhiều khách hàng).
- **Khóa ngoại:** customer_zip_code_prefix trong bảng customers tham chiếu đến geolocation_zip_code_prefix trong bảng geolocation.

Bảng leads_closed và bảng leads_qualified

- **Mối quan hệ:** 1:1 (Một khách hàng tiềm năng chỉ có thể được chuyển đổi một lần).
- **Khóa ngoại:** mql_id trong bảng leads_closed tham chiếu đến mql_id trong bảng leads_qualified.

Bảng leads_closed và bảng sellers

- **Mối quan hệ:** 1:N (Một seller có thể chuyển đổi nhiều khách hàng tiềm năng thành khách hàng).

- **Khóa ngoại:** seller_id trong bảng leads_closed tham chiếu đến seller_id trong bảng sellers

Bước 3: Kiểm tra và tối ưu hóa thiết kế

Kiểm tra toàn vẹn dữ liệu: Đảm bảo rằng thiết kế bảng và các mối quan hệ của bạn đảm bảo tính toàn vẹn và chính xác của dữ liệu.

Tối ưu hóa: Đảm bảo rằng cơ sở dữ liệu của bạn có thể xử lý khối lượng dữ liệu lớn một cách hiệu quả và các truy vấn sẽ được thực thi nhanh chóng.

Bước 4: Đảm bảo tính mở rộng và bảo trì

Thiết kế cơ sở dữ liệu cần đảm bảo tính linh hoạt và dễ dàng mở rộng trong tương lai. Các thay đổi trong yêu cầu kinh doanh có thể dẫn đến việc điều chỉnh cấu trúc cơ sở dữ liệu

4.2.4 TRÌNH BÀY CÁC BƯỚC TẠO BẢNG DỮ LIỆU

Bước 1: Tạo cơ sở dữ liệu trên SQL Server

```

1 | CREATE DATABASE ECommerceDB
2 | GO
3 | USE ECommerceDB

```

Hình 4.2: Tạo database mới

Bước 2: Tạo bảng product_category_name_translation

```

5 |
6 -- Bảng product_category_name_translation
7 [-]CREATE TABLE product_category_name_translation (
8     product_category_name VARCHAR(255) PRIMARY KEY,
9     product_category_name_english VARCHAR(255)
10    );
11

```

Hình 4.3: Bảng product_category_name_translation

Bước 3: Tạo bảng sellers

```

12  -- Bảng sellers
13  CREATE TABLE sellers (
14      seller_id VARCHAR(255) PRIMARY KEY,
15      seller_zip_code_prefix INTEGER,
16      seller_city VARCHAR(255),
17      seller_state VARCHAR(255)
18 );
19

```

Hình 4.4: Bảng sellers

Bước 4: Tạo bảng customers

```

20  -- Bảng customers
21  CREATE TABLE customers (
22      customer_id VARCHAR(255) PRIMARY KEY,
23      customer_unique_id VARCHAR(255),
24      customer_zip_code_prefix INTEGER,
25      customer_city VARCHAR(255),
26      customer_state VARCHAR(255)
27 );

```

Hình 4.5: Bảng customers

Bước 5: Tạo bảng geolocation

```

29  -- Bảng geolocation
30  CREATE TABLE geolocation (
31      geolocation_zip_code_prefix INTEGER PRIMARY KEY,
32      geolocation_lat REAL,
33      geolocation_lng REAL,
34      geolocation_city VARCHAR(255),
35      geolocation_state VARCHAR(255)
36 );
-- 

```

Hình 4.6: Bảng geolocation

Bước 6: Tạo bảng order_items

```

38  -- Bảng order_items
39  CREATE TABLE order_items (
40      order_id VARCHAR(255),
41      order_item_id INTEGER,
42      product_id VARCHAR(255),
43      seller_id VARCHAR(255),
44      shipping_limit_date TEXT,
45      price REAL,
46      freight_value REAL,
47      PRIMARY KEY (order_id, order_item_id),
48      FOREIGN KEY (order_id) REFERENCES orders(order_id),
49      FOREIGN KEY (product_id) REFERENCES products(product_id),
50      FOREIGN KEY (seller_id) REFERENCES sellers(seller_id)
51 );

```

Hình 4.7: Bảng order_items

Bước 7: Tạo bảng order_payments

```

53  -- Bảng order_payments
54  CREATE TABLE order_payments (
55      order_id VARCHAR(255),
56      payment_sequential INTEGER,
57      payment_type VARCHAR(255),
58      payment_installments INTEGER,
59      payment_value REAL,
60      PRIMARY KEY (order_id, payment_sequential),
61      FOREIGN KEY (order_id) REFERENCES orders(order_id)
62 );
-- 

```

Hình 4.8: Bảng order_payments

Bước 8: Tạo bảng order_reviews

```

64  -- Bảng order_reviews
65  CREATE TABLE order_reviews (
66      review_id VARCHAR(255) PRIMARY KEY,
67      order_id VARCHAR(255),
68      review_score INTEGER,
69      review_comment_title VARCHAR(255),
70      review_comment_message TEXT,
71      review_creation_date TEXT,
72      review_answer_timestamp TEXT,
73      FOREIGN KEY (order_id) REFERENCES orders(order_id)
74 );
-- 

```

Hình 4.9: *Bảng order_reviews*

Bước 9: Tạo bảng orders

```

76  -- Bảng orders
77  CREATE TABLE orders (
78      order_id VARCHAR(255) PRIMARY KEY,
79      customer_id VARCHAR(255),
80      order_status VARCHAR(255),
81      order_purchase_timestamp TEXT,
82      order_approved_at TEXT,
83      order_delivered_carrier_date TEXT,
84      order_delivered_customer_date TEXT,
85      order_estimated_delivery_date TEXT,
86      FOREIGN KEY (customer_id) REFERENCES customers(customer_id)
87 );

```

Hình 4.10: *Bảng orders*

Bước 10: Tạo bảng products

```

89  -- Bảng products
90  CREATE TABLE products (
91      product_id VARCHAR(255) PRIMARY KEY,
92      product_category_name VARCHAR(255),
93      product_name_lenght REAL,
94      product_description_lenght REAL,
95      product_photos_qty REAL,
96      product_weight_g REAL,
97      product_length_cm REAL,
98      product_height_cm REAL,
99      product_width_cm REAL,
100     FOREIGN KEY (product_category_name) REFERENCES product_category_name_translation(product_catego
101 );

```

Hình 4.11: *Bảng products*

Bước 11: Tạo bảng leads_qualified

```

103  -- Bảng leads_qualified
104  CREATE TABLE leads_qualified (
105      mql_id VARCHAR(255) PRIMARY KEY,
106      first_contact_date TEXT,
107      landing_page_id VARCHAR(255),
108      origin VARCHAR(255)
109 );

```

Hình 4.12: *Bảng leads_qualified*

Bước 12: Tạo bảng leads_closed

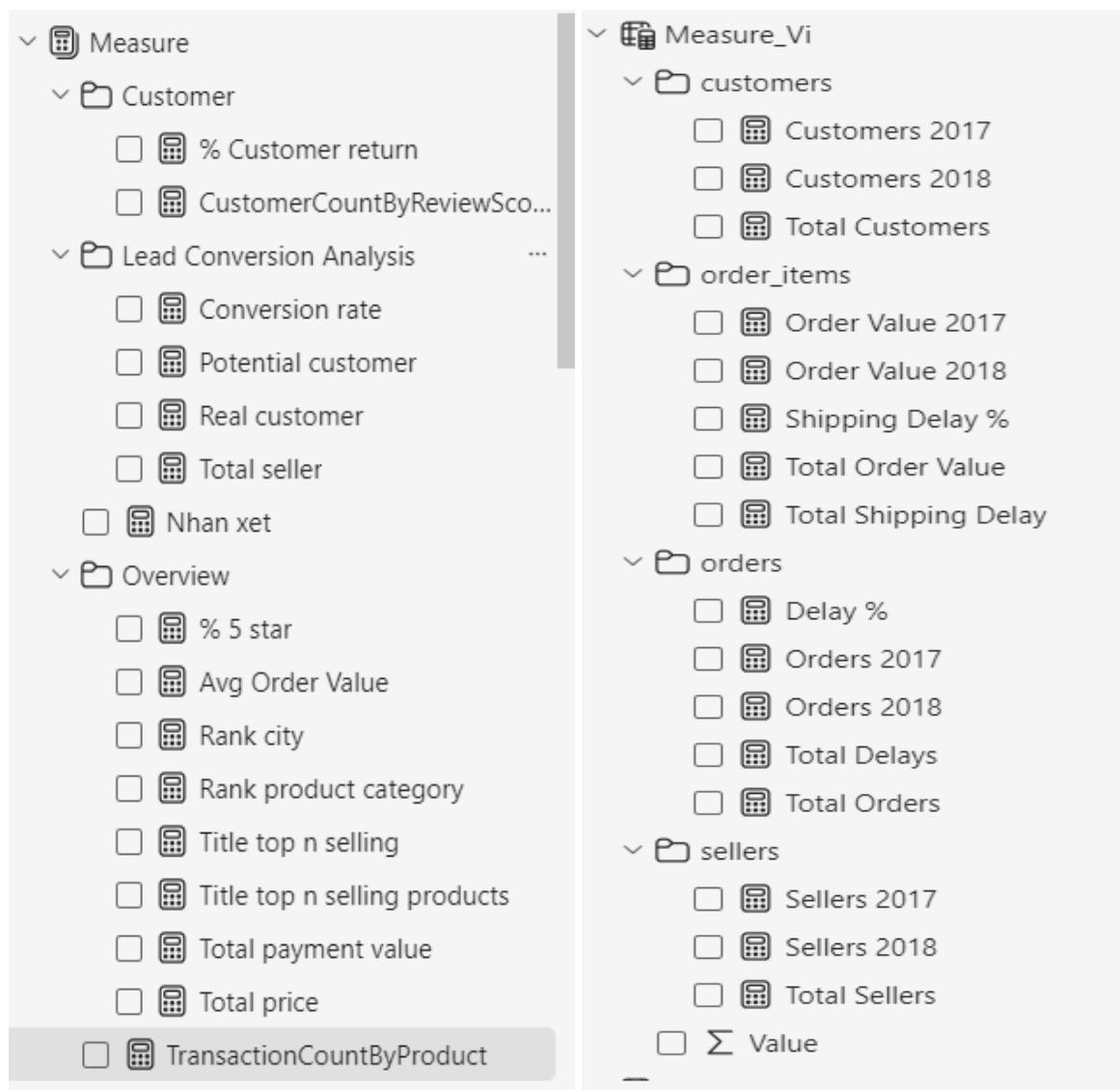
```
111  -- Bảng leads_closed
112  CREATE TABLE leads_closed (
113      mql_id VARCHAR(255),
114      seller_id VARCHAR(255),
115      sdr_id VARCHAR(255),
116      sr_id VARCHAR(255),
117      won_date TEXT,
118      business_segment VARCHAR(255),
119      lead_type VARCHAR(255),
120      lead_behaviour_profile VARCHAR(255),
121      has_company INTEGER,
122      has_gtin INTEGER,
123      average_stock VARCHAR(255),
124      business_type VARCHAR(255),
125      declared_product_catalog_size REAL,
126      declared_monthly_revenue REAL,
127      PRIMARY KEY (mql_id, seller_id),
128      FOREIGN KEY (mql_id) REFERENCES leads_qualified(mql_id),
129      FOREIGN KEY (seller_id) REFERENCES sellers(seller_id)
130  );
```

Hình 4.13: Bảng leads_closed

4.3 XỬ LÝ DỮ LIỆU DAX

4.3.1 MEASURE

4.3.1.1 Tạo calendar



Hình 4.14: Danh sách các measure được sử dụng trong báo cáo

```
1 % Customer return = 1 - DIVIDE(DISTINCTCOUNT(customers
    [customer_unique_id]), DISTINCTCOUNT(customers
    [customer_id]))
```

Hình 4.15: Measure tỷ lệ khách hàng quay lại

```
1 Potential customer = COUNTROWS(marketing_qualified_leads)
```

Hình 4.16: Measure tính số lượng khách hàng tiềm năng

```

1 % 5 star = DIVIDE(
2 COUNTROWS(
3     FILTER(
4         order_reviews,
5         order_reviews[review_score] = 5
6     )
7 ), COUNTROWS(order_reviews))

```

Hình 4.17: Công thức tính tỷ lệ đánh giá 5 sao

```

1 Rank city =
2     var _TopCity =
3         RANKX(
4             ALL(geolocation[geolocation_city]), [Total payment value],
5             ,DESC
6         )
7
8     RETURN
9         IF(
10             _TopCity<=Parameter[Parameter Value],
11             [Total payment value]
12         )

```

Hình 4.18: Công thức xếp hạng thành phố

```

1 Rank product category =
2     var _TopProductId =
3         RANKX(
4             ALL(product_category_name[product_category_name_english]), [Total price],
5             ,DESC
6         )
7
8     RETURN
9         IF(
10            _TopProductId<=Parameter1[Parameter1 Value],
11            [Total price])

```

Hình 4.19: Measure xếp hạng danh mục sản phẩm theo tổng giá trị

1 Title top n selling = "Top " & Parameter1[Parameter1 Value] & " best-selling products category by amount"

Hình 4.20: Measure tạo tiêu đề động cho biểu đồ top N sản phẩm bán chạy

1 Title top n selling products = "Top " & Parameter[Parameter Value] & " best-selling cities by amount"

Hình 4.21: Measure tạo tiêu đề động cho biểu đồ top N thành phố bán chạy

4.3.2 CALCULATED COLUMN

4.3.2.1 Bảng Customers

customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state	Customer_Clusters
000379cdec625522490c315e70c7a9fb	0b83f73b19c2019e182fd552c048a22c	4841	Sao paulo	Sao Paulo	Cluster1
0005aefbb69d34b3424dccc0a0e9fd0	616309b2eeb7bd9c05b0fdfbab28e6c6	3052	Sao paulo	Sao Paulo	Cluster1
00062b33cb9fffe976afdcff967ea74d	f90f55ee274a4ae21510b386134b09cd	2306	Sao paulo	Sao Paulo	Cluster1
001028b78fd413e19704b3867c369d3a	e579a935f49ffd73b93c18eaaa04efaf84	5387	Sao paulo	Sao Paulo	Cluster1
001051abfcfdbed9f87b4266213a5df1	4ea5df5937187bd3176aee08d4782104	2251	Sao paulo	Sao Paulo	Cluster1
0013280441d86a4f7a8006fdaf1b0fe	06caeba6db23a17bf9bc3846b768387	5409	Sao paulo	Sao Paulo	Cluster1
0013cd8e350a7cc76873441e431dd5ee	334fed5abcee3aa96c13f1432703e1fd	3585	Sao paulo	Sao Paulo	Cluster1
001574cd5824c0b1ea90dd4f4ba6d5b8	8141dd1e051afe7d72079570fe72d5f1	8248	Sao paulo	Sao Paulo	Cluster2
0019c9aaad15b043c48f0a1180f22ce8	4662682dade3cc1bfa04996c5225a849	4141	Sao paulo	Sao Paulo	Cluster1
001a5704156400917a187dd74e6cbc1	163b27a06a32c2fa565927170b59b5d4	2512	Sao paulo	Sao Paulo	Cluster1
001df1ee5c36767aa607001ab1a13a06	46b44ab325f78e5bb3dc0bbef1082082	1030	Sao paulo	Sao Paulo	Cluster1
0026955706fd4e2fa997f3f4c18d485a	47b6bc410befb9fa30a4c029dba944e5	2926	Sao paulo	Sao Paulo	Cluster1
0029cdf064769ca0df3186b54d068c99	74c286ab8cc161142213780711964b06	3183	Sao paulo	Sao Paulo	Cluster1
002ce108ccf0356ef5c8b1dce3c0ae29	e6ad2d9078cb28e4ac6ece8432e74c7	2728	Sao paulo	Sao Paulo	Cluster1
002d358e2462f87678443706cbf2eb21	5a77d3e9231351d3a4ea54c3fe824e79	5171	Sao paulo	Sao Paulo	Cluster1
002f90a6eb386bc43bc9ba200db31a89	bd3001c423ef9a9092b0a5158f1ccdf6	4929	Sao paulo	Sao Paulo	Cluster1

Hình 4.22: Bảng Customers

4.3.2.2 Bảng linear_data

Day of Year	Sum of Total price & freight	product_category_eng_name
2	469.04	Health beauty
3	3341.6	Health beauty
4	1358.92	Health beauty
5	2834.72	Health beauty
7	103.15	Health beauty
8	2613.12	Health beauty
9	2802.3	Health beauty
10	1866.15	Health beauty
11	4685.37	Health beauty
12	2177.14	Health beauty
15	2038.03	Health beauty

Hình 4.23: Bảng linear_data

4.3.2.3 Bảng order_items

shipping_limit_date	shipping_limit_time	Total price & freight	freight%	carrier delay	shipping_limit_datetime	delay (HOUR)	Day of Year
Thursday, February 15, 2018	1:48:42 AM	35	43.14%	0	2/15/2018 1:48:42 AM		46
Thursday, December 14, 2017	2:49:20 PM	35	43.14%	0	12/14/2017 2:49:20 PM		348
Tuesday, November 7, 2017	4:10:20 AM	35	43.14%	0	11/7/2017 4:10:20 AM		311
Thursday, August 31, 2017	3:05:52 AM	35	43.14%	1	8/31/2017 3:05:52 AM	43	243
Tuesday, June 27, 2017	8:30:29 AM	35	43.14%	0	6/27/2017 8:30:29 AM		178
Wednesday, February 28, 2018	2:08:47 AM	35	43.14%	0	2/28/2018 2:08:47 AM		59
Monday, June 26, 2017	4:25:14 PM	35	43.14%	0	6/26/2017 4:25:14 PM		177
Friday, December 1, 2017	7:31:47 PM	35	43.14%	0	12/1/2017 7:31:47 PM		335
Wednesday, October 18, 2017	5:35:44 PM	35	43.14%	0	10/18/2017 5:35:44 PM		291
Thursday, May 25, 2017	6:15:28 PM	35	43.14%	0	5/25/2017 6:15:28 PM		145
Thursday, October 5, 2017	9:20:06 PM	35	43.14%	0	10/5/2017 9:20:06 PM		278
Tuesday, August 22, 2017	9:25:21 AM	35	43.14%	0	8/22/2017 9:25:21 AM		234
Monday, February 5, 2018	1:51:29 PM	35	43.14%	1	2/5/2018 1:51:29 PM	65	36
Friday, January 5, 2018	2:09:32 AM	35	43.14%	0	1/5/2018 2:09:32 AM		5
Monday, February 26, 2018	8:35:34 PM	35	43.14%	0	2/26/2018 8:35:34 PM		57
Friday, September 1, 2017	1:15:18 PM	35	43.14%	0	9/1/2017 1:15:18 PM		244
Friday, October 6, 2017	11:49:24 AM	35	43.14%	1	10/6/2017 11:49:24 AM	18	279
Thursday, July 27, 2017	2:45:51 AM	35	43.14%	0	7/27/2017 2:45:51 AM		208
Monday, July 3, 2017	3:43:24 AM	35	43.14%	0	7/3/2017 3:43:24 AM		184
Monday, October 30, 2017	3:15:07 AM	35	43.14%	0	10/30/2017 3:15:07 AM		303
Thursday, August 24, 2017	10:15:15 PM	35	43.14%	0	8/24/2017 10:15:15 PM		236
Thursday, September 7, 2017	10:25:44 PM	35	43.14%	0	9/7/2017 10:25:44 PM		250
Monday, December 18, 2017	6:10:52 PM	35	43.14%	0	12/18/2017 6:10:52 PM		352
Thursday, September 7, 2017	10:33:55 PM	35	43.14%	0	9/7/2017 10:33:55 PM		250

Hình 4.24: Bảng order_items

ice	freight_value	shipping_limit_date	shipping_limit_time	Total price & freight	fr
R\$19.90	R\$15.10	Thursday, February 15, 2018	1:48:42 AM	35	

Hình 4.25: Tạo cột tính tổng giá trị đơn hàng

<code>1 carrier delay = IF(RELATED(orders[order_delivered_carrier_date]) > order_items[shipping_limit_datetime],1,0)</code>	<code>ce</code>	<code>freight_value</code>	<code>shipping_limit_date</code>	<code>shipping_limit_time</code>	<code>Total price & freight</code>	<code>freight%</code>	<code>carrier delay</code>
\$19.90	R\$15.10	Thursday, February 15, 2018	1:48:42 AM		35	43.14%	0

Hình 4.26: Tạo cột kiểm tra giao hàng trễ của đơn vị vận chuyển

<code>1 shipping_limit_datetime = order_items[shipping_limit_date] & " " & order_items[shipping_limit_time]</code>	<code>ice</code>	<code>freight_value</code>	<code>shipping_limit_date</code>	<code>shipping_limit_time</code>	<code>Total price & freight</code>	<code>freight%</code>	<code>carrier delay</code>	<code>shipping_limit_datetime</code>
\$19.90	R\$15.10	Thursday, February 15, 2018	1:48:42 AM		35	43.14%	0	2/15/2018 1:48:42 AM

Hình 4.27: Tạo cột kết hợp ngày giờ giới hạn giao hàng

<code>1 delay (HOUR) = IF(order_items[carrier delay] = 1,DATEDIFF(order_items[shipping_limit_date],RELATED(orders[order_delivered_carrier_date]),HOUR))</code>	<code>ce</code>	<code>freight_value</code>	<code>shipping_limit_date</code>	<code>shipping_limit_time</code>	<code>Total price & freight</code>	<code>freight%</code>	<code>carrier delay</code>	<code>shipping_limit_datetime</code>	<code>delay (HOUR)</code>
\$19.90	R\$15.10	Thursday, February 15, 2018	1:48:42 AM		35	43.14%	0	2/15/2018 1:48:42 AM	0

Hình 4.28: Tạo cột tính độ trễ giao hàng theo giờ

4.3.2.4 Bảng order_reviews

comments	review completion (hr)	review completion (days)	review completion post delivery	review completion prior delivery	review created prior de	CustomerCountByRev
1 No	68	2	1	0	No	57076
1 No	64	2	1	0	No	57076
1 No	60	2	1	0	No	57076
1 No	73	3	0	0	Yes	57076
1 No	66	2	1	0	No	57076
1 No	354	14	1	0	No	57076
1 No	63	2	1	0	No	57076
1 No	86	3	1	0	No	57076
1 No	63	2	1	0	No	57076
1 No	65	2	1	0	No	57076
1 No	16	0	1	0	No	57076
1 No	185	7	1	0	No	57076
1 No	216	9	1	0	No	57076
1 No	62	2	1	0	No	57076
1 No	64	2	1	0	No	57076

Hình 4.29: order_reviews

Structure	Formatting	Properties	Sort	Groups	Relationships
<code>1 comments = IF(OR(ISBLANK(order_reviews[review_comment_message]),order_reviews[review_comment_message] = ""), "No", "Yes")</code>	<code>comments</code>	<code>review completion (hr)</code>	<code>review completion (days)</code>	<code>review completion post delivery</code>	<code>review completion prior delivery</code>

Hình 4.30: Tạo cột kiểm tra sự tồn tại của bình luận

Structure	Formatting	Properties	Sort	Groups	Relationships
<code>1 review completion (hr) = DATEDIFF(order_reviews[review_creation_date],order_reviews[review_answer_timestamp],HOUR)</code>	<code>comments</code>	<code>review completion (hr)</code>	<code>review completion (days)</code>	<code>review completion post delivery</code>	<code>review completion prior de</code>

Hình 4.31: Tạo cột tính thời gian hoàn thành đánh giá (theo giờ)

Structure	Formatting	Properties	Sort	Groups	Relationships
<code>1 review completion (days) = DATEDIFF(order_reviews[review_creation_date],order_reviews[review_answer_timestamp],DAY)</code>	<code>comments</code>	<code>review completion (hr)</code>	<code>review completion (days)</code>	<code>review completion post delivery</code>	<code>review completion prior del</code>

Hình 4.32: Tạo cột tính thời gian hoàn thành đánh giá (theo ngày)

1 review completion post delivery = IF(order_reviews[review_creation_date] > RELATED(orders[order_delivered_customer_date]), DATEDIFF(RELATED(orders[order_delivered_customer_date]), order_reviews[review_creation_date], DAY), 0)
comments ▾ review completion (hr) ▾ review completion (days) ▾ review completion post delivery ▾ review completion prior delivery ▾ review created prior de

Hình 4.33: Tạo cột tính thời gian hoàn thành đánh giá sau khi giao hàng

1 review completion prior delivery = IF(order_reviews[review_creation_date] < RELATED(orders[order_delivered_customer_date]), DATEDIFF(order_reviews[review_creation_date], RELATED(orders[order_delivered_customer_date]), HOUR), 0)
comments ▾ review completion (hr) ▾ review completion (days) ▾ review completion post delivery ▾ review completion prior delivery ▾ review created prior de

Hình 4.34: Tạo cột tính thời gian hoàn thành đánh giá trước khi giao hàng

1 review created prior delivery? = IF(order_reviews[review_creation_date] < RELATED(orders[order_delivered_customer_date]), "Yes", "No")
PM No 68 2 1 0 No

Hình 4.35: Tạo cột kiểm tra thời điểm tạo đánh giá

1 CustomerCountByReviewScore =
2 CALCULATE(
3 DISTINCTCOUNT(orders[customer_id]),
4 FILTER(
5 order_reviews,
6 order_reviews[review_score] = EARLIER(order_reviews[review_score])
7)
8)

comments ▾ review completion (hr) ▾ review completion (days) ▾ review completion post delivery

Hình 4.36: Tạo cột phân tích số lượng khách hàng dựa trên điểm đánh giá

4.3.2.5 Bảng orders

The screenshot shows the Power BI Data View with the 'orders' table selected. Several columns are highlighted with red arrows pointing upwards from the table header:

- actual delivery time (days)
- approval time (hours)
- carrier delivery time (days)
- carrier to customer time (days)
- estimated delivery time (days)
- delivery delay
- Purchase_Day
- order_day

The right side of the screen shows the Power BI Field Explorer, which lists various measures and dimensions. A red box highlights the 'orders' table entry, and another red box highlights the 'actual delivery time (days)' measure under the 'orders' table.

Hình 3.37: orders

Structure	Formatting	Properties	Sort	Groups	Relationships
X ✓ 1 actual delivery time (days) = DATEDIFF(orders[order_purchase_timestamp],orders[order_delivered_customer_date],DAY)	actual delivery time (days) approval time (hours) carrier delivery time (days) carrier to customer time (days) estimated delivery time	26 2 6 20			

Hình 4.38: Tạo cột tính thời gian giao hàng thực tế (theo ngày)

Structure	Formatting	Properties	Sort	Groups
✓ 1 approval time (hours) = DATEDIFF(orders[order_purchase_timestamp],orders[order_approved_at],HOUR)	approval time (hours) carrier delivery time (days) carrier to customer time (days) estimated delivery time	26 2 6 20		

Hình 4.39: Tạo cột tính thời gian giao hàng (theo giờ)

Structure	Formatting	Properties	Sort	Groups	Relationships
1 carrier delivery time (days) = DATEDIFF(orders[order_approved_at],orders[order_delivered_carrier_date],DAY)	carrier delivery time (days) carrier to customer time (days) estimated delivery time	26 6 20			

Hình 4.40: Tạo cột tính thời gian vận chuyển (theo ngày)

Structure	Formatting	Properties	Sort	Groups	Relationships	Calculations
1 carrier to customer time (days) = DATEDIFF(orders[order_delivered_carrier_date],orders[order_delivered_customer_date],DAY)	carrier to customer time (days) estimated delivery time	26 20				

Hình 4.41: Theo dõi thời gian từ đơn vị vận chuyển đến khách hàng

Structure	Formatting	Properties	Sort	Groups	Relationships	Calculations
1 estimated delivery time (days) = DATEDIFF(orders[order_purchase_timestamp],orders[order_estimated_delivery_date],DAY)	estimated delivery time	26				

Hình 4.42: Tính toán thời gian giao hàng dự kiến

Structure	Formatting	Properties	Sort	Groups	Relationships	Calculations
✓ 1 delivery delay = IF(orders[actual delivery time (days)] > orders[estimated delivery time (days)],1,0)	delivery delay	26 0				

Hình 4.43: Đánh giá hiệu suất giao hàng

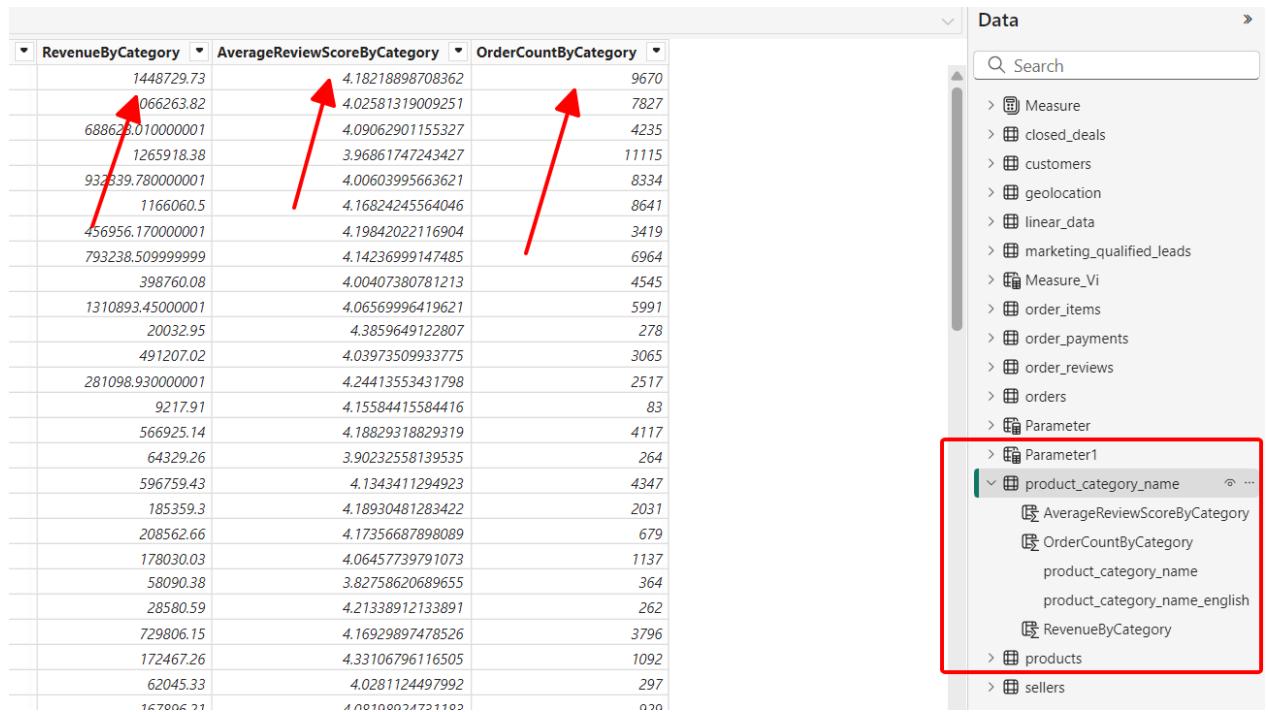
Structure	Formatting	Properties	Sort	Groups	Relationships	Calculations
✓ 1 Purchase_Day = WEEKDAY(orders[order_purchase_timestamp])						

Hình 4.44: Xác định ngày mua trong tuần

Structure	Formatting	Properties	Sort	Groups	Relationships	Calculations
✓ 1 order_day = IF(orders[Purchase_Day] = 1,"Mon",IF(orders[Purchase_Day] = 2,"Tue",IF(orders[Purchase_Day] = 3,"Wed",IF(orders[Purchase_Day] = 4,"Thu",IF(orders[Purchase_Day] = 5,"Fri",IF(orders[Purchase_Day] = 6,"Sat","Sun")))))	order day	26 1				

Hình 4.45: Chuyển đổi ngày trong tuần

4.3.2.6 Bảng product_category_name



RevenueByCategory	AverageReviewScoreByCategory	OrderCountByCategory
1448729.73	4.18218898708362	9670
066263.82	4.02581319009251	7827
688623.010000001	4.09062901155327	4235
1265918.38	3.96861747243427	11115
932339.780000001	4.0060395663621	8334
1166060.5	4.16824245564046	8641
456956.170000001	4.19842022116904	3419
793238.509999999	4.14236999147485	6964
398760.08	4.00407380781213	4545
1310893.45000001	4.06569996419621	5991
20032.95	4.3859649122807	278
491207.02	4.03973509933775	3065
281098.930000001	4.24413553431798	2517
9217.91	4.15584415584416	83
566925.14	4.18829318829319	4117
64329.26	3.90232558139535	264
596759.43	4.1343411294923	4347
185359.3	4.18930481283422	2031
208562.66	4.17356687898089	679
178030.03	4.06457739791073	1137
58090.38	3.82758620689655	364
28580.59	4.21338912133891	262
729806.15	4.16929897478526	3796
172467.26	4.33106796116505	1092
62045.33	4.0281124497992	297
157006.31	4.170110034731103	020

Hình 4.46: Bảng product_category_name

```

1 RevenueByCategory =
2 CALCULATE(
3     SUM(order_payments[payment_value]),
4     FILTER(
5         ALL(order_items),
6         order_items[product_id] IN
7             SELECTCOLUMNS(
8                 FILTER(
9                     products,
10                    products[product_category_name] = product_category_name[product_category_name]
11                ),
12                    "product_id",
13                    products[product_id]
14            )
15        )
16    )

```

Hình 4.47: Phân tích doanh thu theo danh mục sản phẩm

```

1 AverageReviewScoreByCategory =
2 CALCULATE(
3     AVERAGE(order_reviews[review_score]),
4     FILTER(
5         order_items,
6         order_items[product_id] IN
7         SELECTCOLUMNS(
8             FILTER(
9                 products,
10                products[product_category_name] = EARLIER(product_category_name[product_category_name])
11            ),
12            "product_id", products[product_id]
13        )
14    )
15 )

```

Hình 4.48: Điểm đánh giá trung bình theo danh mục

```

1 OrderCountByCategory =
2 CALCULATE(
3     COUNT(order_items[order_id]),
4     FILTER(
5         order_items,
6         order_items[product_id] IN
7         SELECTCOLUMNS(
8             FILTER(
9                 products,
10                products[product_category_name] = EARLIER(product_category_name[product_category_name])
11            ),
12            "product_id", products[product_id]
13        )
14    )
15 )

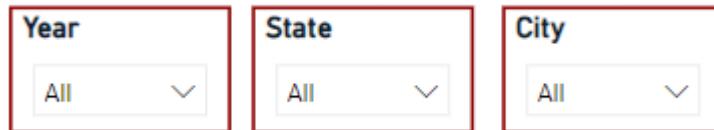
```



Hình 4.49: Đếm số lượng đơn hàng theo danh mục sản phẩm

4.3.3 FILTER

4.3.3.1 Tạo filter chọn năm, bang, thành phố



Hình 4.50: Bộ lọc chọn năm, bang, thành phố

4.3.3.2 Tạo filter chọn Ngày



Hình 4.51: Bộ lọc chọn ngày

5 TRỰC QUAN HÓA DỮ LIỆU

5.1 CÁC KỸ THUẬT TRỰC QUAN HÓA

Trực quan hóa dữ liệu là quá trình sử dụng các biểu đồ và đồ thị để truyền đạt thông tin một cách trực quan, dễ hiểu. Có nhiều kỹ thuật trực quan hóa dữ liệu khác nhau, mỗi kỹ thuật phù hợp với các loại dữ liệu và mục đích phân tích cụ thể. Dưới đây là một số kỹ thuật trực quan hóa phổ biến và kỹ thuật được áp dụng cho dự án này.

Temporal data visualization (Trực quan hóa dữ liệu chuỗi thời gian)

Trực quan hóa dữ liệu chuỗi thời gian (Temporal Data Visualization) là kỹ thuật trình bày các đối tượng dữ liệu theo chiều thời gian một cách trực quan và dễ hiểu. Kỹ thuật này thường sử dụng các loại biểu đồ như biểu đồ đường, biểu đồ cột và dòng thời gian để minh họa các thay đổi, xu hướng, và sự kiện xảy ra liên tục trong một khoảng thời gian nhất định. Ví dụ, biểu đồ đường có thể được sử dụng để biểu thị sự thay đổi của giá bán xe ô tô theo từng tháng hoặc từng năm, giúp người xem dễ dàng nhận thấy các xu hướng tăng giảm, các điểm biến động, và các chu kỳ thời gian cụ thể. Trực quan hóa dữ liệu chuỗi thời gian giúp người dùng theo dõi các biến động trong dữ liệu theo thời gian, từ đó đưa ra các quyết định dựa trên những hiểu biết sâu sắc và có căn cứ. Đây là một công cụ mạnh mẽ trong phân tích dữ liệu, đặc biệt hữu ích cho các lĩnh vực yêu cầu theo dõi xu hướng và dự báo như kinh doanh, tài chính và quản lý.

Hierarchical data visualization (Trực quan hóa dữ liệu phân cấp)

Trực quan hóa dữ liệu phân cấp (Hierarchical Data Visualization) là kỹ thuật dùng để trình bày các nhóm hoặc tập hợp các mục có liên kết chung với một mục cha, giúp hiển thị cấu trúc phân cấp và mối quan hệ giữa các phần tử dữ liệu. Các dạng trực quan phổ biến cho dữ liệu phân cấp bao gồm cây phân cấp, sơ đồ cây, và biểu đồ cây. Ví dụ, cây dữ liệu có thể được sử dụng để biểu thị lượng dữ liệu về hàng tồn kho, trong đó có nút cha đại diện cho danh mục lớn như "quần áo" và các nút con đại diện cho các mục nhỏ hơn như "áo sơ mi", "quần dài", và "tắt". Kỹ thuật này giúp người dùng dễ dàng hiểu được cấu trúc phân cấp của dữ liệu, nhìn thấy mối quan hệ và sự phụ thuộc giữa các phần tử, từ đó hỗ trợ việc phân tích, ra quyết định và quản lý dữ liệu hiệu quả hơn. Trực quan hóa dữ liệu phân cấp đặc biệt hữu ích trong các lĩnh vực như quản lý dự án, tổ chức dữ liệu, và phân tích hệ thống phức tạp.

Network data visualization (Trực quan hóa dữ liệu mạng)

Trực quan hóa dữ liệu mạng (Network Data Visualization) là kỹ thuật biểu diễn dữ liệu dưới dạng các điểm và mối liên kết giữa chúng trên một đồ thị, giúp hiển thị rõ ràng các mối quan hệ và tương tác phức tạp trong mạng lưới dữ liệu. Các dạng biểu đồ phổ biến cho dữ liệu mạng bao gồm biểu đồ phân tán, biểu đồ bong bóng, và đám mây từ. Ví dụ, biểu đồ phân tán có thể hiển thị mối quan hệ giữa hai biến, trong khi biểu đồ bong bóng thêm một yếu tố dữ liệu thứ ba thông qua kích thước của bong bóng. Đám mây từ trình bày tần suất xuất hiện của các từ bằng cách sử dụng các từ có kích cỡ khác nhau, giúp nhận diện nhanh chóng các từ quan trọng. Trực quan hóa dữ liệu mạng giúp người dùng hiểu được cấu trúc và động lực của các mạng phức tạp, chẳng hạn như mạng xã hội, mạng giao thông, hoặc mối liên kết giữa các yếu tố trong một hệ thống. Kỹ thuật này đặc biệt hữu ích trong việc phân

tích các mối quan hệ và tương tác, phát hiện các mẫu ẩn và xác định các yếu tố quan trọng trong mạng lưới dữ liệu.

Multidimensional data visualization (Trực quan hóa dữ liệu đa chiều)

Trực quan hóa dữ liệu đa chiều (Multidimensional Data Visualization) là kỹ thuật dùng để biểu diễn và phân tích dữ liệu có nhiều biến hoặc chiều, giúp người dùng dễ dàng so sánh và nhận diện các mối quan hệ giữa các yếu tố dữ liệu phức tạp. Các dạng biểu đồ phổ biến cho dữ liệu đa chiều bao gồm biểu đồ cột, biểu đồ tròn và đồ thị cột. Ví dụ, biểu đồ cột có thể so sánh các yếu tố dữ liệu khác nhau như doanh số bán hàng theo các hãng xe trong các khoảng thời gian khác nhau, biểu đồ tròn trực quan hóa tỷ lệ phần trăm của từng danh mục trong tổng thể như thị phần của các hãng xe. Kỹ thuật này cho phép người dùng theo dõi và phân tích sự thay đổi của một hoặc nhiều biến qua thời gian hoặc giữa các danh mục khác nhau, từ đó phát hiện ra các xu hướng, mẫu, và mối quan hệ quan trọng. Trực quan hóa dữ liệu đa chiều là công cụ mạnh mẽ trong các lĩnh vực như kinh doanh, tài chính, và nghiên cứu khoa học, nơi mà việc phân tích các yếu tố đa chiều là cần thiết để đưa ra các quyết định chính xác và có căn cứ.

Geospatial data visualization (Trực quan hóa dữ liệu không gian địa lý)

Trực quan hóa dữ liệu không gian địa lý (Geospatial Data Visualization) là kỹ thuật sử dụng các bản đồ và biểu đồ để trình bày dữ liệu liên quan đến các vị trí địa lý trong thế giới thực. Kỹ thuật này giúp biểu diễn thông tin không gian một cách trực quan, giúp người dùng dễ dàng nhận diện các mẫu, xu hướng và mối quan hệ trong dữ liệu địa lý. Các dạng trực quan phổ biến bao gồm bản đồ nhiệt, bản đồ mật độ, và bản đồ địa hình. Ví dụ, bản đồ nhiệt có thể được sử dụng để hiển thị lượng khách hàng ghé thăm các chi nhánh bán lẻ khác nhau, với màu sắc đậm nhạt biểu thị mật độ khách hàng. Bản đồ địa hình có thể minh họa các đặc điểm địa lý và các hiện tượng tự nhiên. Trực quan hóa dữ liệu không gian địa lý không chỉ giúp hiểu rõ hơn về thông tin không gian mà còn hỗ trợ việc ra quyết định dựa trên vị trí, tối ưu hóa các chiến lược kinh doanh và quản lý tài nguyên. Kỹ thuật này đặc biệt hữu ích trong các lĩnh vực như quản lý đô thị, logistics, marketing địa phương, và nghiên cứu môi trường.

Kỹ thuật đang được áp dụng cho dự án

Trong dự án này, chúng ta tập trung vào hai kỹ thuật chính: trực quan hóa dữ liệu chuỗi thời gian để theo dõi xu hướng và sự thay đổi theo thời gian (ví dụ: doanh thu theo tháng) và trực quan hóa dữ liệu đa chiều để khám phá mối quan hệ giữa các yếu tố khác nhau (ví dụ: doanh thu theo danh mục sản phẩm).

Ngoài ra, chúng ta có thể sử dụng bộ trợ các kỹ thuật như trực quan hóa dữ liệu phân cấp (hiển thị cấu trúc danh mục sản phẩm) và trực quan hóa dữ liệu không gian địa lý (phân tích dữ liệu theo vị trí địa lý).

Việc lựa chọn các kỹ thuật này dựa trên đặc điểm dữ liệu của Olist, vừa mang tính thời gian vừa có nhiều chiều thông tin, giúp tối ưu hóa việc phân tích và trình bày dữ liệu, từ đó hỗ trợ ra quyết định kinh doanh hiệu quả.

5.2 CÁC NGUYÊN TẮC TRỰC QUAN HÓA

5 nguyên tắc trực quan hóa dữ liệu

Chọn đúng loại biểu đồ:

- Sử dụng biểu đồ đường để hiển thị xu hướng theo thời gian, ví dụ như doanh thu theo tháng, số lượng đơn hàng theo quý.
- Sử dụng biểu đồ cột để so sánh giá trị giữa các danh mục, ví dụ như so sánh doanh thu giữa các danh mục sản phẩm, số lượng đơn hàng theo từng vùng miền.
- Sử dụng biểu đồ tròn để hiển thị tỷ lệ phần trăm, ví dụ như tỷ lệ đơn hàng thành công, tỷ lệ hủy đơn.

Không phải tất cả dữ liệu đều quan trọng:

- Tập trung vào những dữ liệu cốt lõi, liên quan trực tiếp đến mục tiêu phân tích.
- Loại bỏ những thông tin không cần thiết, tránh làm biểu đồ trở nên rối rắm.
- Ví dụ, khi phân tích doanh thu, chỉ nên tập trung vào các thông tin về doanh thu, số lượng đơn hàng, thời gian, thay vì hiển thị thêm các chi tiết như mã sản phẩm, thông tin chi tiết về khách hàng.

Biểu đồ thể hiện đúng tương quan số liệu thực tế:

- Đảm bảo biểu đồ phản ánh chính xác mối quan hệ giữa các dữ liệu.
- Ví dụ, khi sử dụng biểu đồ cột, trực tung nên bắt đầu từ 0 để tránh gây hiểu nhầm về tỷ lệ.
- Sử dụng tỷ lệ và khoảng cách phù hợp trong biểu đồ để thể hiện chính xác dữ liệu.

Sử dụng màu sắc hợp lý:

- Chọn màu sắc dễ phân biệt, có ý nghĩa và tránh sử dụng quá nhiều màu sắc.
- Sử dụng tương phản màu để làm nổi bật các thông tin quan trọng.
- Ví dụ, sử dụng màu xanh lá cây để thể hiện sự tăng trưởng, màu đỏ để thể hiện sự sụt giảm.

Trình bày dữ liệu đơn giản và hiệu quả:

- Sử dụng biểu đồ đơn giản, dễ hiểu, tránh các chi tiết thừa thãi.
- Đảm bảo các nhãn, chú thích, tiêu đề rõ ràng và dễ đọc.
- Ví dụ, sử dụng biểu đồ đường đơn giản để hiển thị xu hướng doanh thu theo thời gian, thay vì sử dụng biểu đồ kết hợp phức tạp.

Nguyên tắc quan trọng nhất: chọn đúng loại biểu đồ

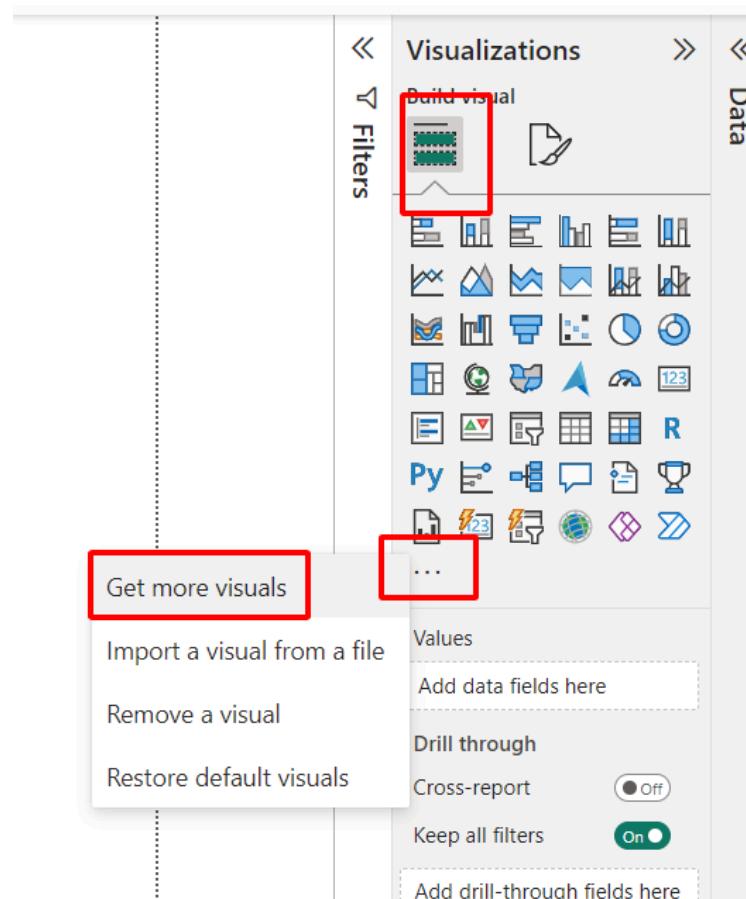
Việc chọn đúng loại biểu đồ là yếu tố then chốt để truyền tải thông tin hiệu quả. Biểu đồ phù hợp sẽ giúp người xem dễ dàng nắm bắt được ý nghĩa của dữ liệu và đưa ra quyết định chính xác.

Ví dụ, để phân tích xu hướng doanh thu theo thời gian, biểu đồ đường là lựa chọn tốt nhất. Để so sánh doanh thu giữa các danh mục sản phẩm, biểu đồ cột sẽ phù hợp hơn.

Tóm lại, việc áp dụng 5 nguyên tắc trên sẽ giúp cho việc trực quan hóa dữ liệu của Olist trở nên hiệu quả, hỗ trợ đắc lực cho quá trình phân tích và ra quyết định kinh doanh.

5.3 TRÌNH BÀY CÁCH THÊM VISUAL MỚI

Bước 1:



Hình 5.1: Get more visual

Bước 2:

The screenshot shows the 'Power BI visuals' marketplace search results for 'histogram'. A red box highlights the search bar at the top right containing the text 'histogram'. Another red box highlights the first result card for 'Histogram Chart (S...)' by PBIVizEdit.com, which has a 5-star rating of '(2)'. The results also include other options like 'Histogram Chart (P...)' by PBIVizEdit.com, 'Histogram with poi...' by MAQ Software, and 'Histogram by PQ S...' by PQ Systems.

Hình 5.2: Tìm visual mới cần thêm

Bước 3:

AppSource | Apps for Power BI visuals

< Apps

Histogram Chart (Standard)

PBIVizEdit.com

★★★★★ 5.0 (2)

Overview Plans + Pricing Ratings + reviews

Plot the distribution with bin size based on your need

This visual is not certified by Microsoft Power BI team. If you want the certified version of this visual, please check [Histogram Chart \(Pro\)](#). Please note all our Pro visuals can be exported to PPT/PDF.

All of our visuals, including those not certified by Microsoft, do not access any external services or resources and this can be verified with browser (Edge, Chrome, Firefox) developer tools.

Add | Download Sample | Sample Instructions

Starts at Add for free or buy an upgraded plan

Hình 5.3: Add visual

Bước 4:

Import successful

The visual was successfully imported into this report.

OK

Values

Add data fields here

Hình 5.4: Hoàn thành

Làm tương tự với các visual khác: Bullet Chart, Sankey Chart, Heatmap, Word Cloud

5.4 TRÌNH BÀY TẠO CÁC REPORT CHO DỰ ÁN

5.4.1 TẠO VISUAL THỐNG KÊ CHI TIẾT

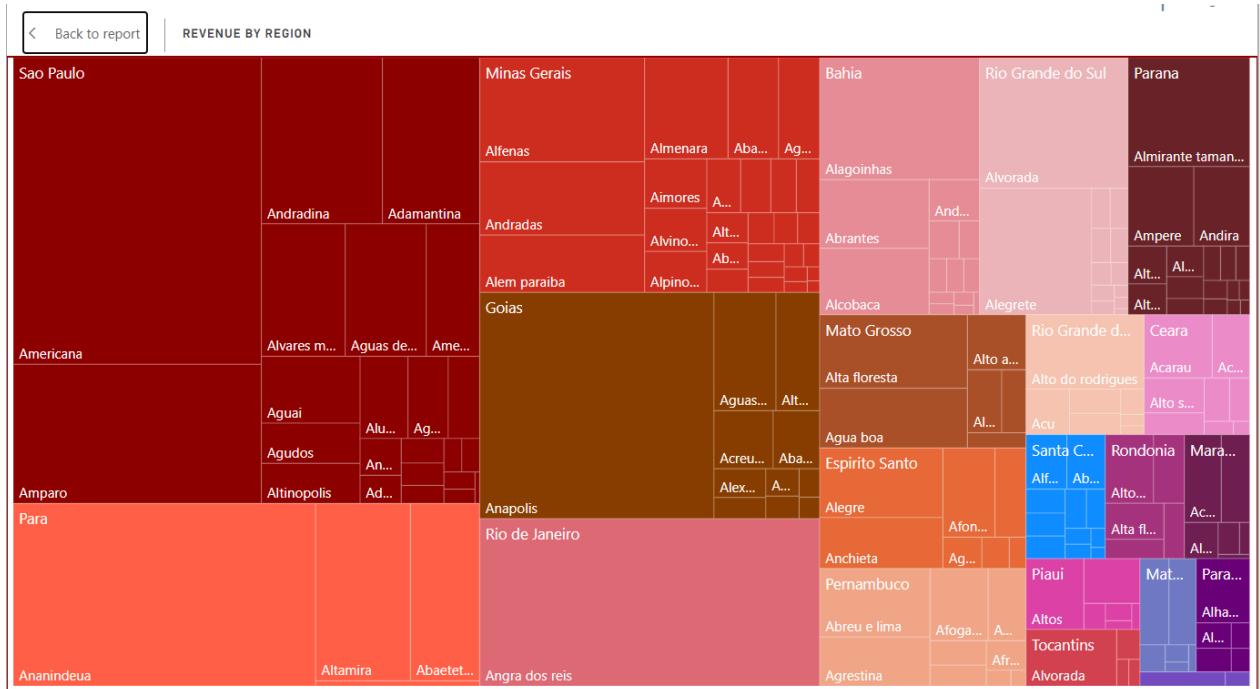
5.4.1.1 Trang Customer Analysis

5.4.1.1.1 Tạo visual thống kê



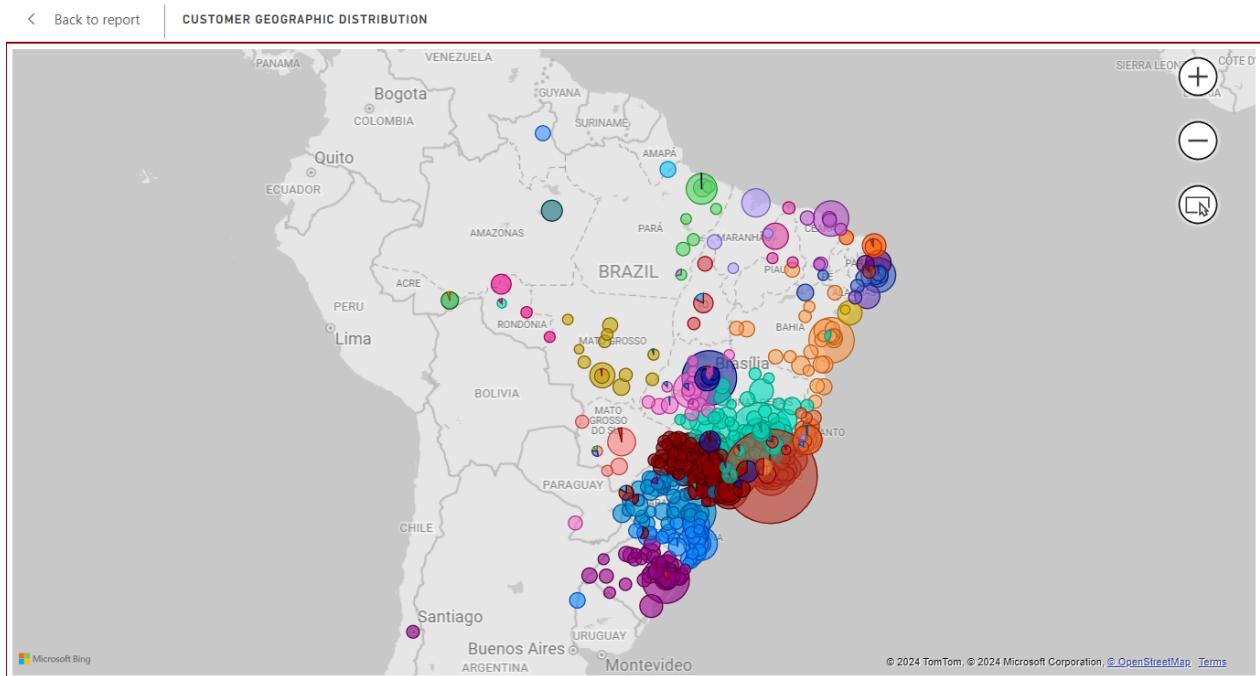
Hình 5.5: Card tóm tắt khách hàng, doanh thu, tỷ lệ khách hàng quay trở lại

5.4.1.1.2 Tạo visual doanh thu theo khu vực



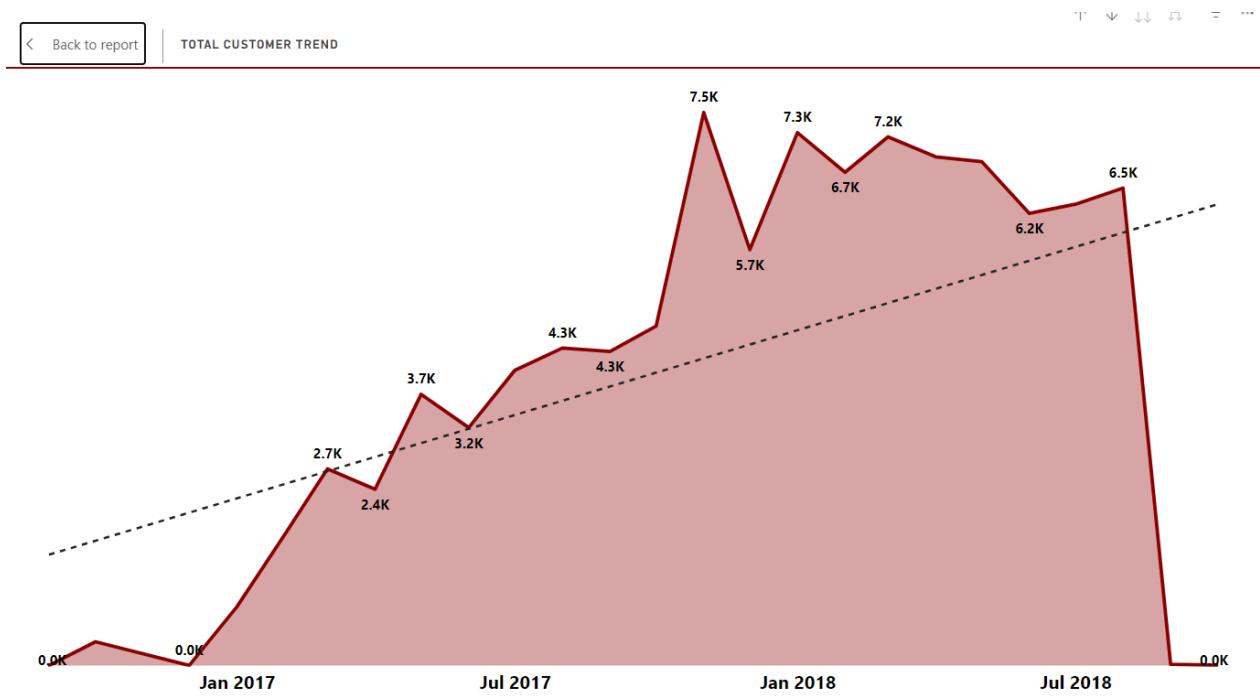
Hình 5.6: Biểu đồ doanh thu theo khu vực

5.4.1.1.3 Tạo visual thể hiện phân bố khách hàng theo địa lý



Hình 5.7: Biểu đồ phân bố khách hàng theo địa lý

5.4.1.1.4 Tạo visual xu hướng số lượng khách hàng theo thời gian



Hình 5.8: Biểu đồ số lượng khách hàng theo thời gian

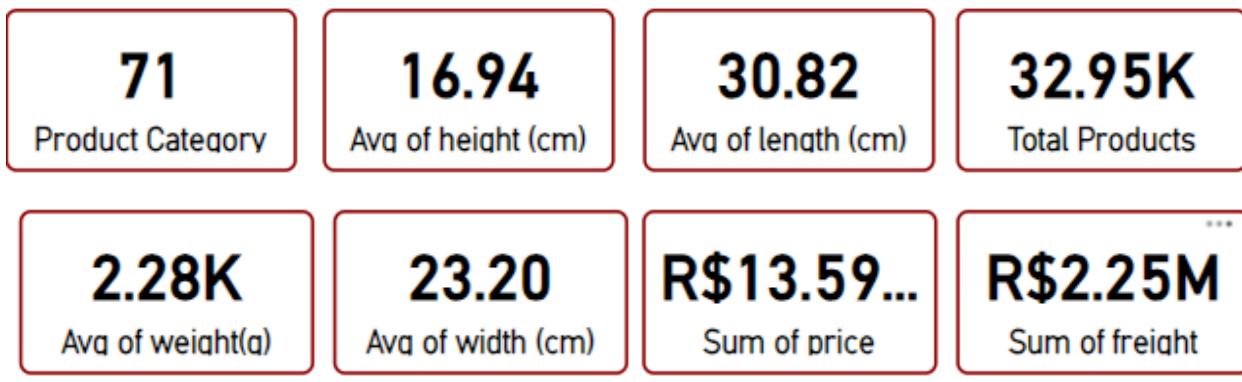
5.4.1.1.5 Tạo visual top khách hàng theo doanh thu và số lượng khách hàng theo bang



Hình 5.9: Biểu đồ top khách hàng theo doanh thu và số lượng khách hàng theo bang

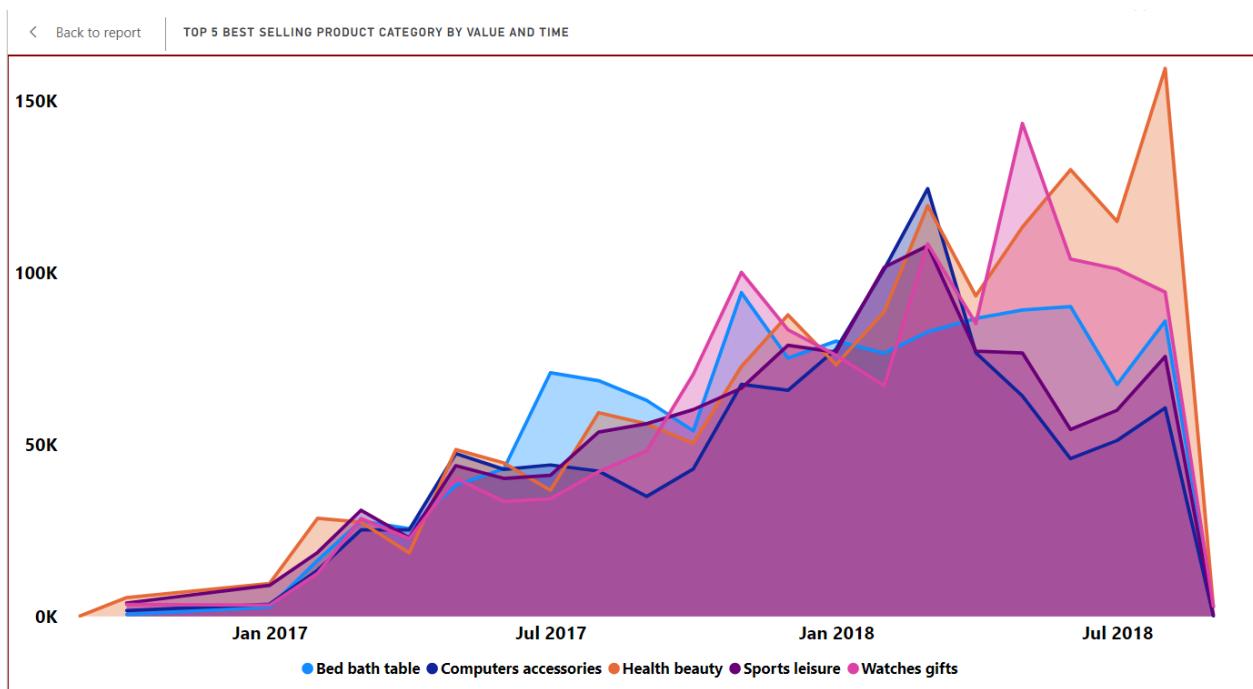
5.4.1.2 Trang Product & trend analysis

5.4.1.2.1 Tạo visual thống kê



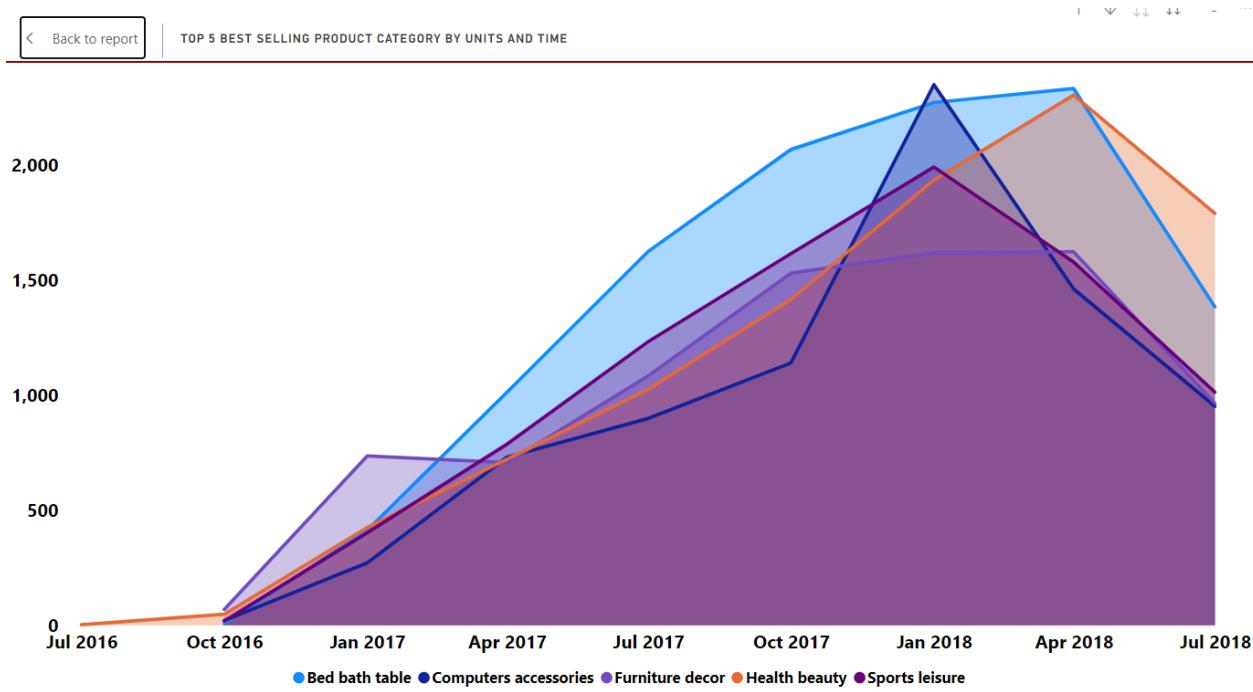
Hình 5.10: Card thống kê về sản phẩm

5.4.1.2.2 Tạo visual thể hiện 5 sản phẩm có giá cao nhất theo thời gian



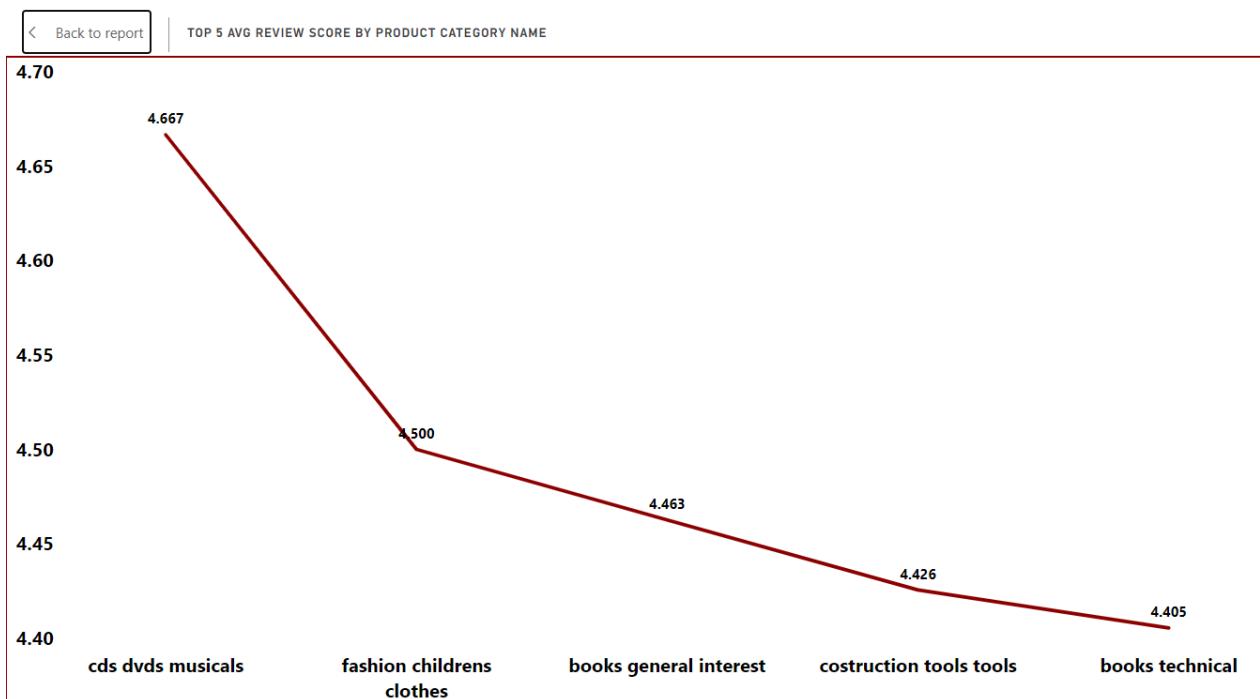
Hình 5.11: Biểu đồ giá 5 loại sản phẩm (có giá cao nhất) theo thời gian

5.4.1.2.3 Tạo visual số lượng 5 loại sản phẩm (có số lượng cao nhất) theo thời gian



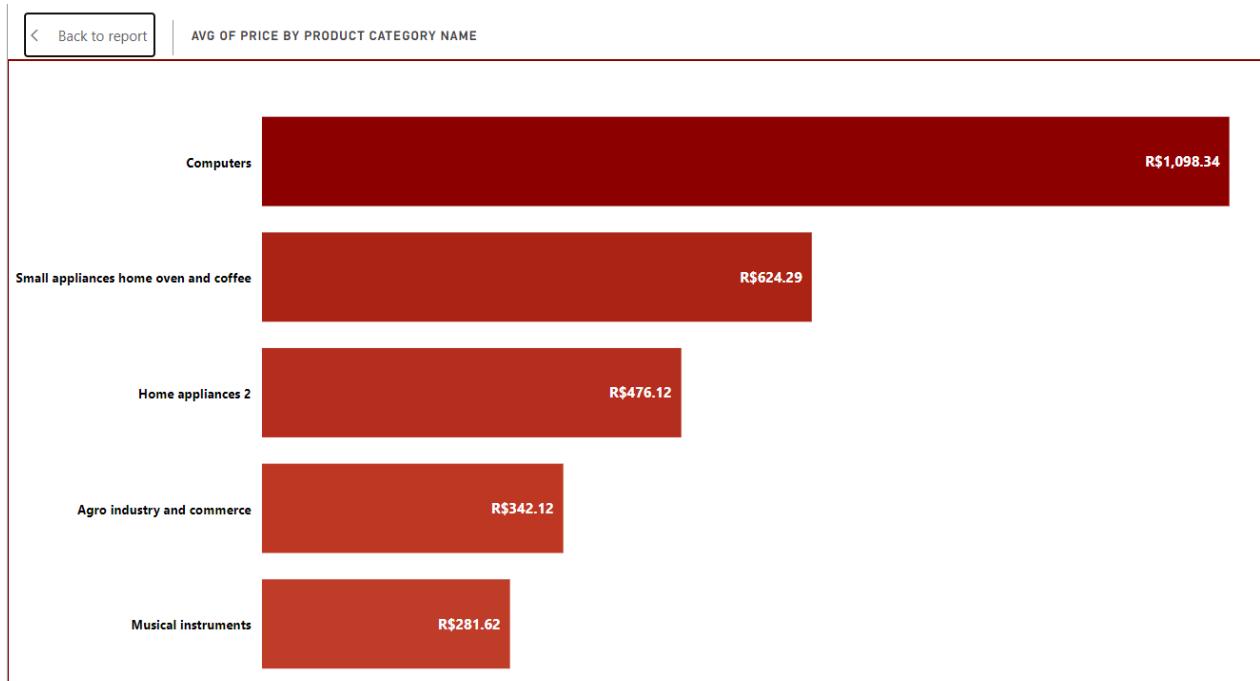
Hình 5.12: Biểu đồ số lượng 5 loại sản phẩm (có số lượng cao nhất) theo thời gian

5.4.1.2.4 Tạo visual top 5 loại sản phẩm được yêu thích nhất



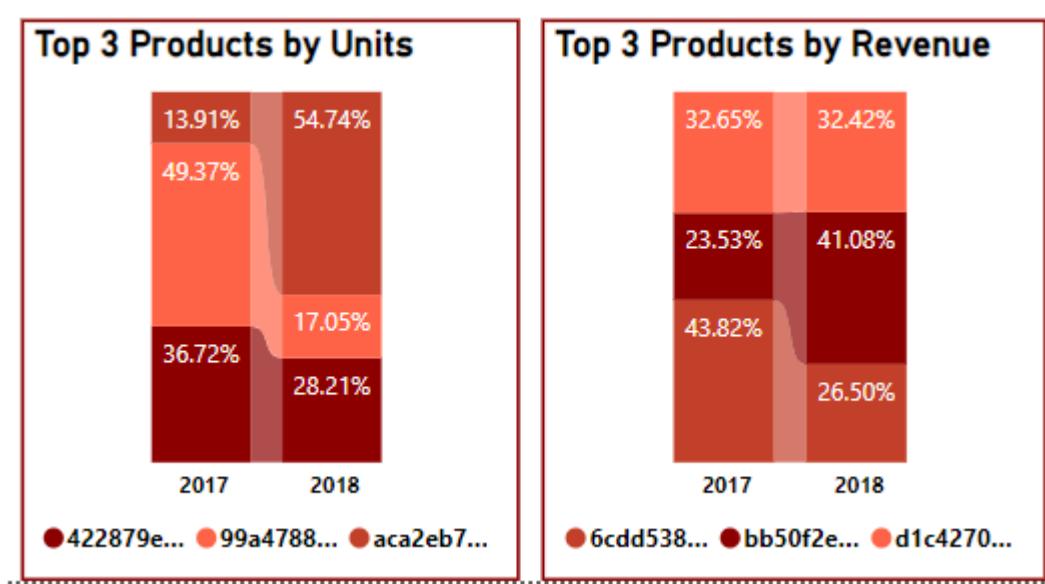
Hình 5.13: Biểu đồ điểm đánh giá top 5 loại sản phẩm được yêu thích nhất

5.4.1.2.5 Tạo visual trung bình giá của từng loại sản phẩm (top 5)



Hình 5.14: Biểu đồ trung bình giá của từng loại sản phẩm (top 5)

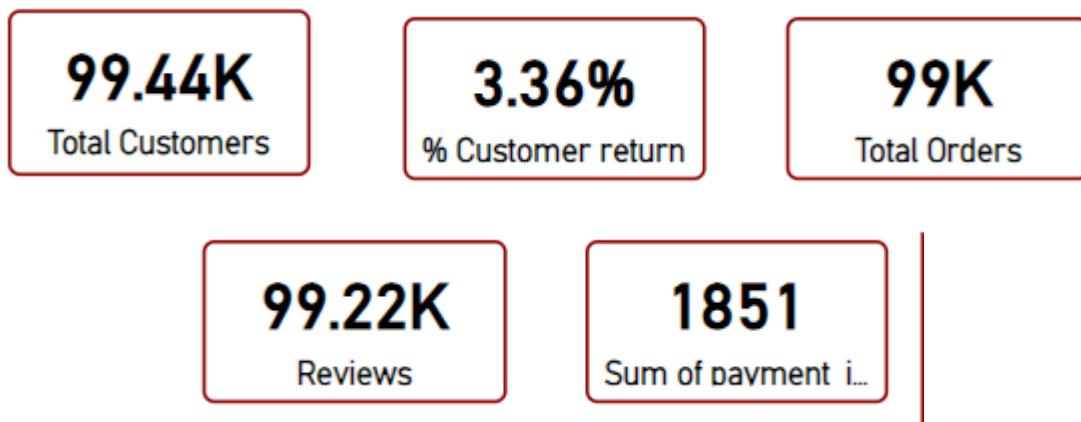
5.4.1.2.6 Tạo visual những sản phẩm bán được nhiều nhất (theo số lượng và doanh thu)



Hình 5.15: Top 3 sản phẩm bán chạy nhất (theo số lượng và doanh thu)

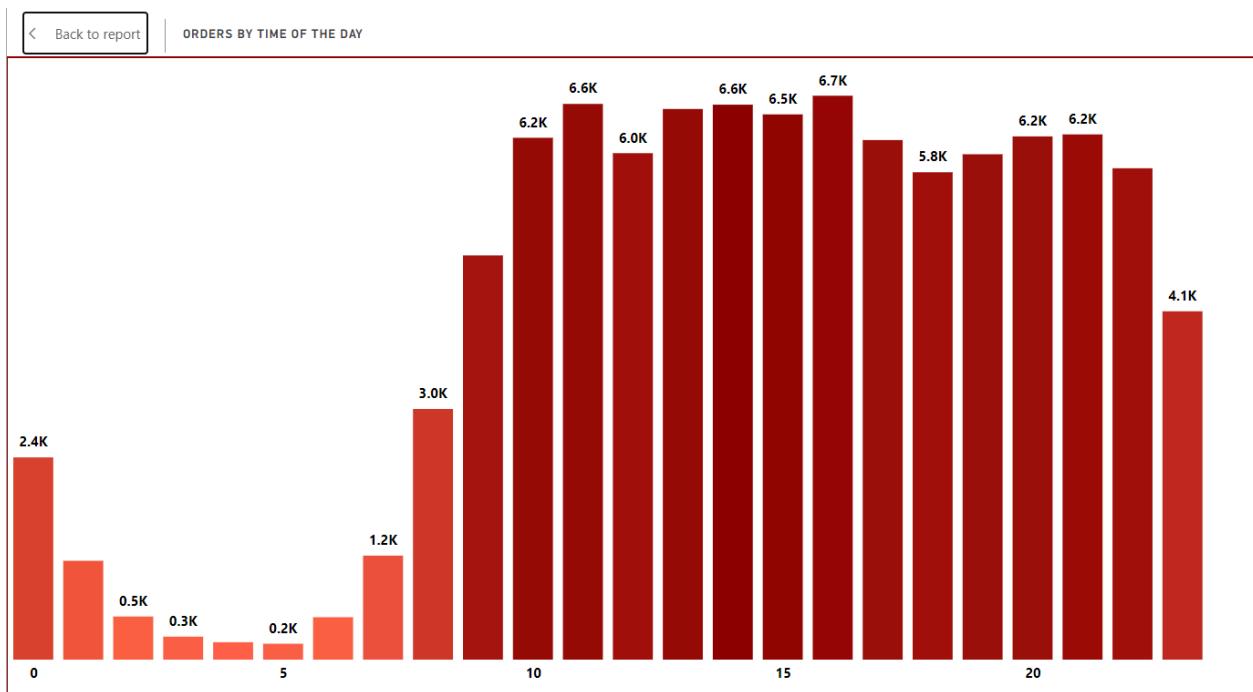
5.4.1.3 Trang Customer behavior

5.4.1.3.1 Tạo visual thống kê



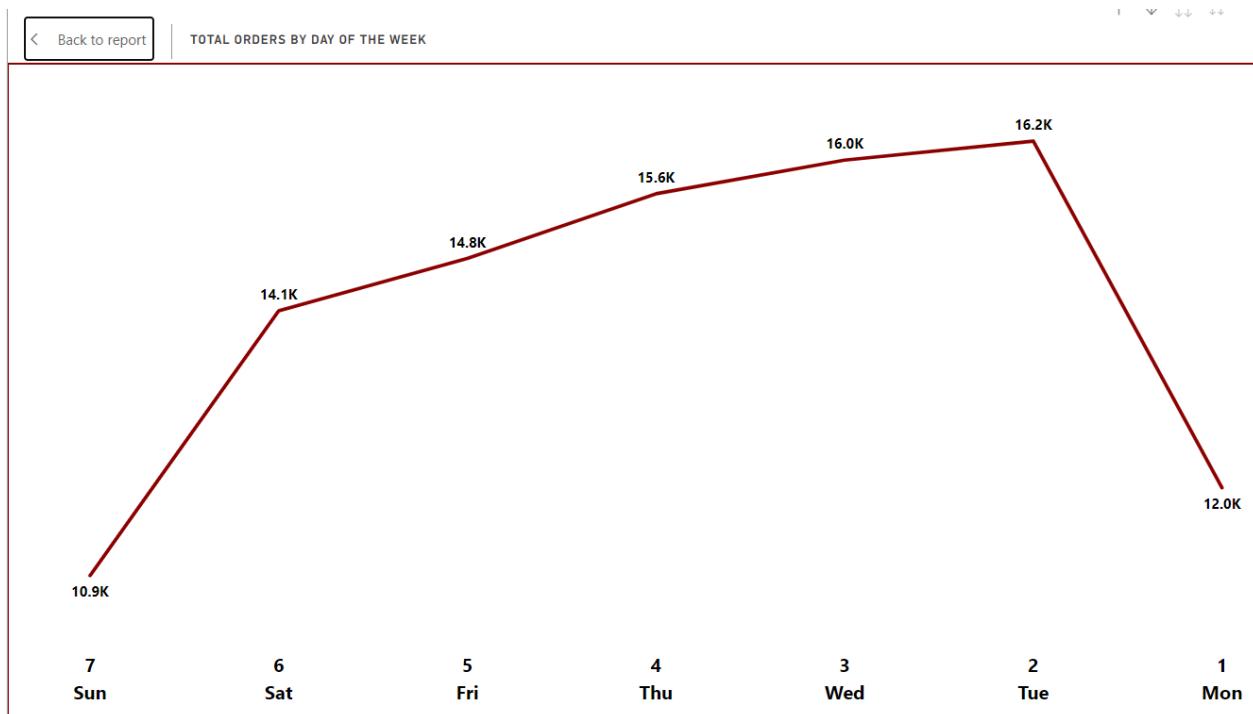
Hình 5.16: Card thống kê hành vi khách hàng

5.4.1.3.2 Tạo visual số lượng đơn hàng theo thời gian (giờ trong ngày)



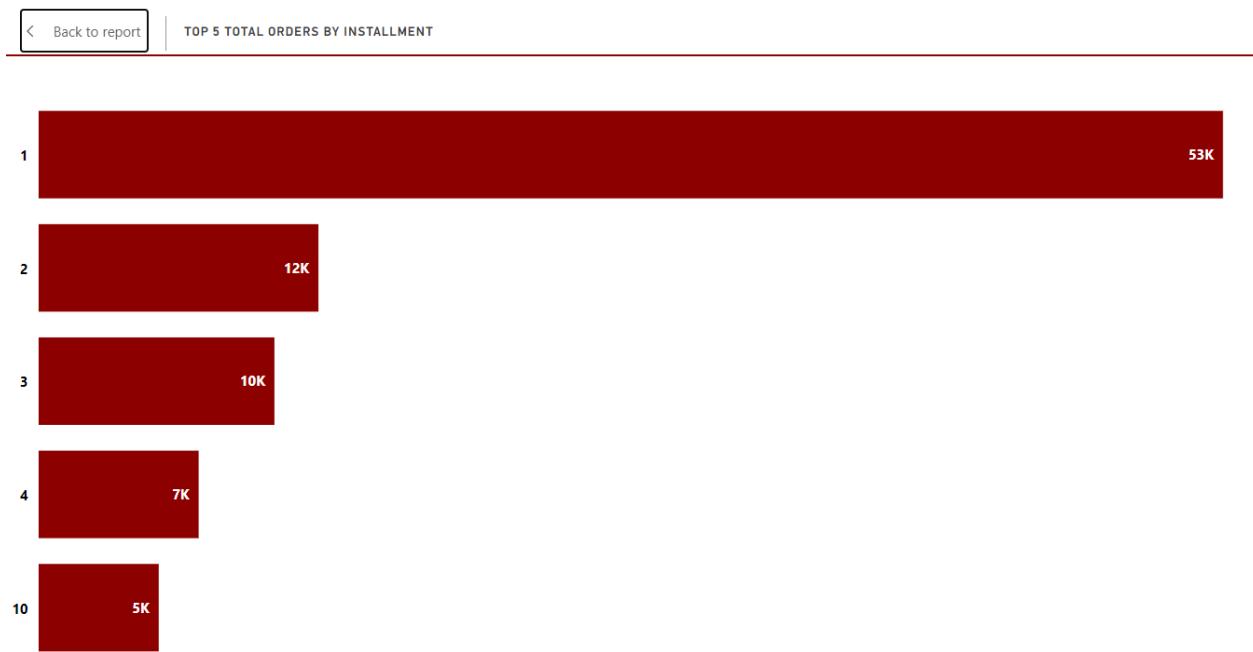
Hình 5.17: Biểu đồ số lượng đơn hàng theo thời gian (trong ngày)

5.4.1.3.3 Tạo visual số lượng đơn hàng theo thời gian (ngày trong tuần)



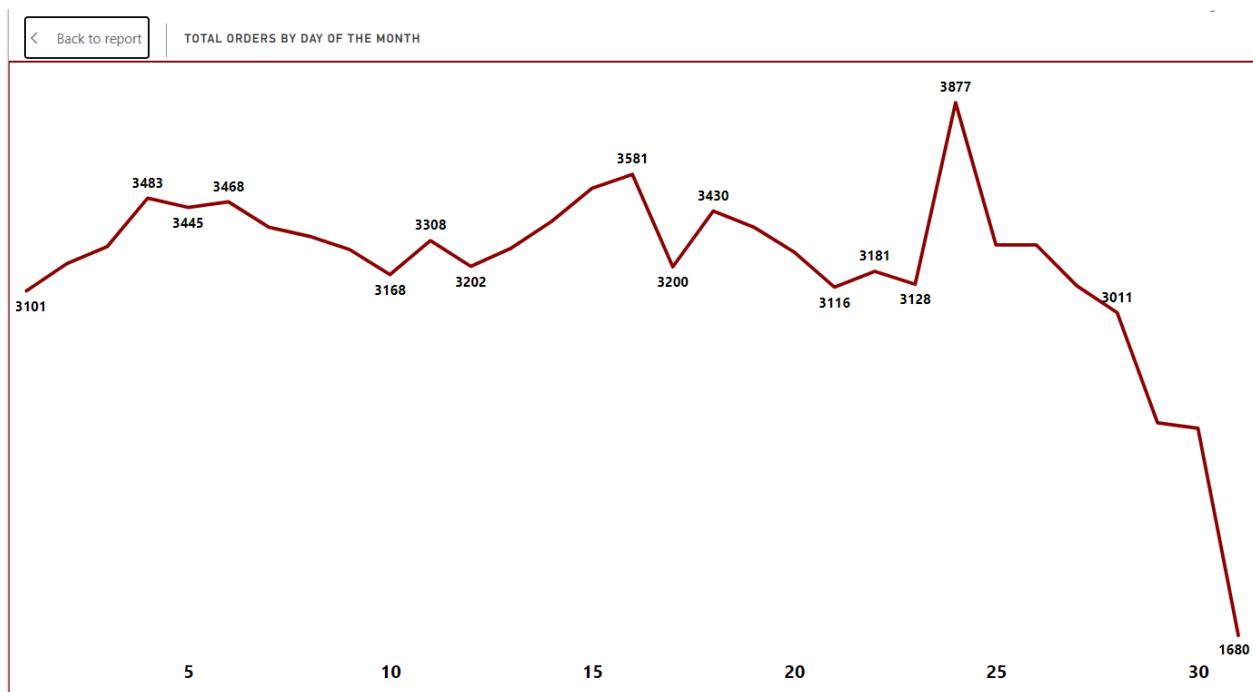
Hình 5.18: Biểu đồ số lượng đơn hàng theo thời gian (ngày trong tuần)

5.4.1.3.4 Tạo visual khách hàng trả góp theo số lần trả góp



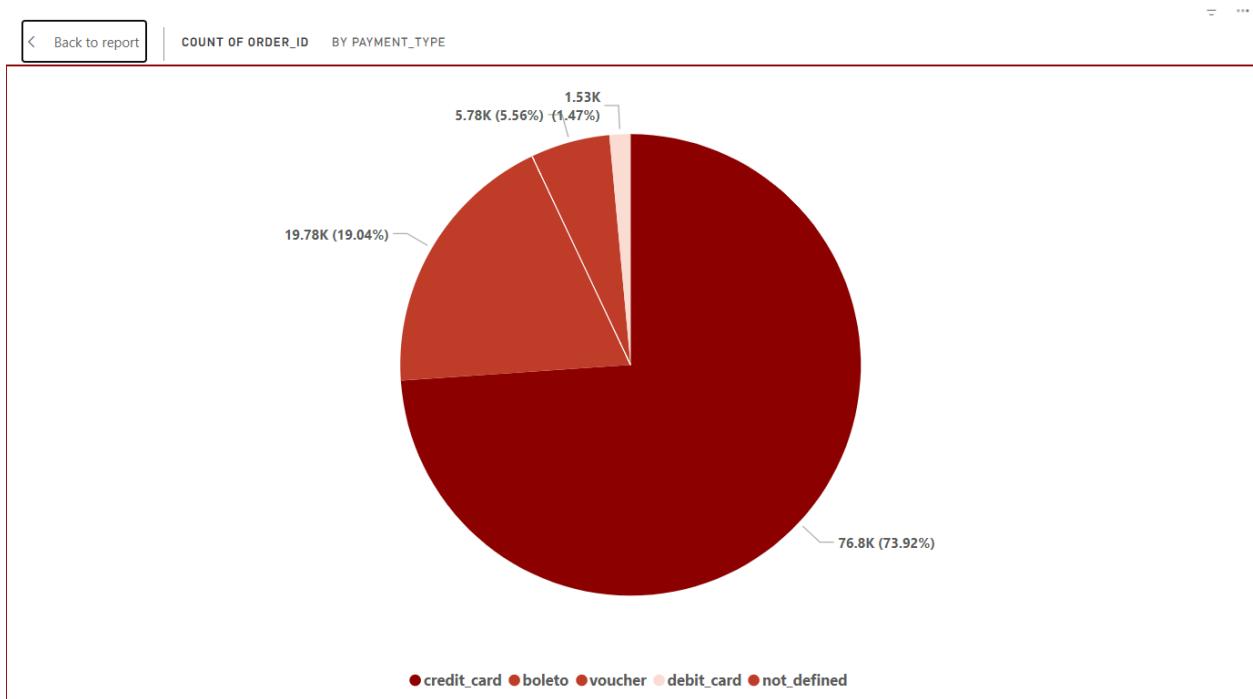
Hình 5.19: Biểu đồ khách hàng trả góp theo số lần trả góp

5.4.1.3.5 Tạo visual số lượng đơn hàng theo thời gian (ngày trong tháng)



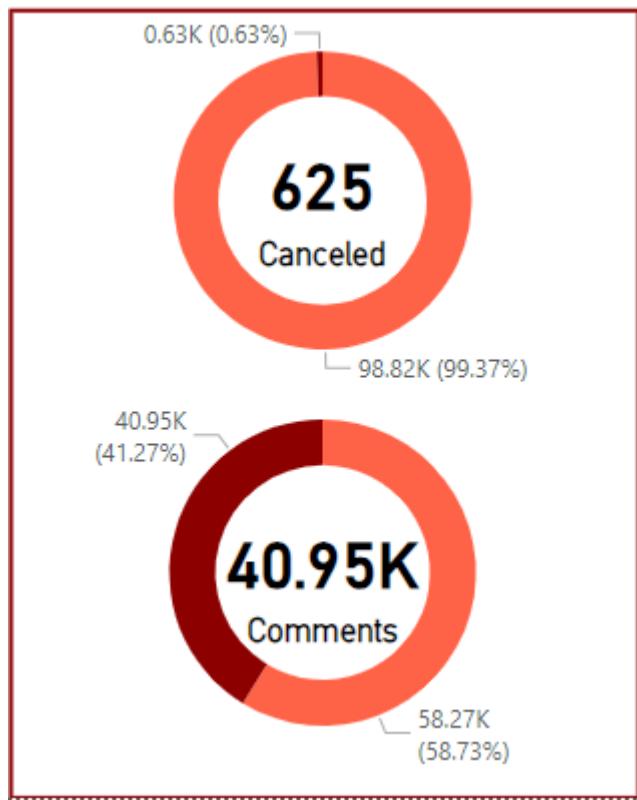
Hình 5.20: Biểu đồ số lượng đơn hàng theo thời gian (ngày trong tháng)

5.4.1.3.6 Tạo visual tỷ lệ hình thức thanh toán



Hình 5.21: Biểu đồ tỷ lệ hình thức thanh toán

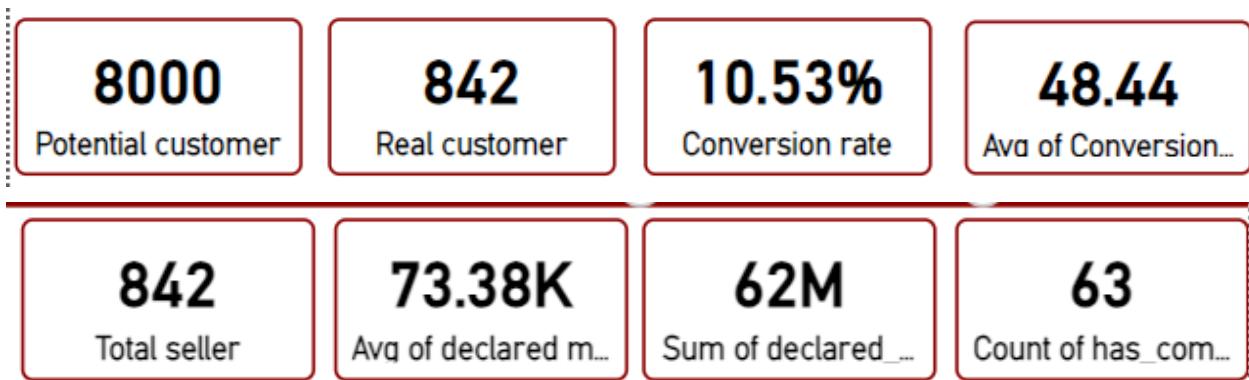
5.4.1.3.7 Tạo visual tỷ lệ đơn hàng hủy, khách hàng để lại nhận xét



Hình 5.22: Biểu đồ số lượng đơn hàng hủy, nhận xét của khách hàng

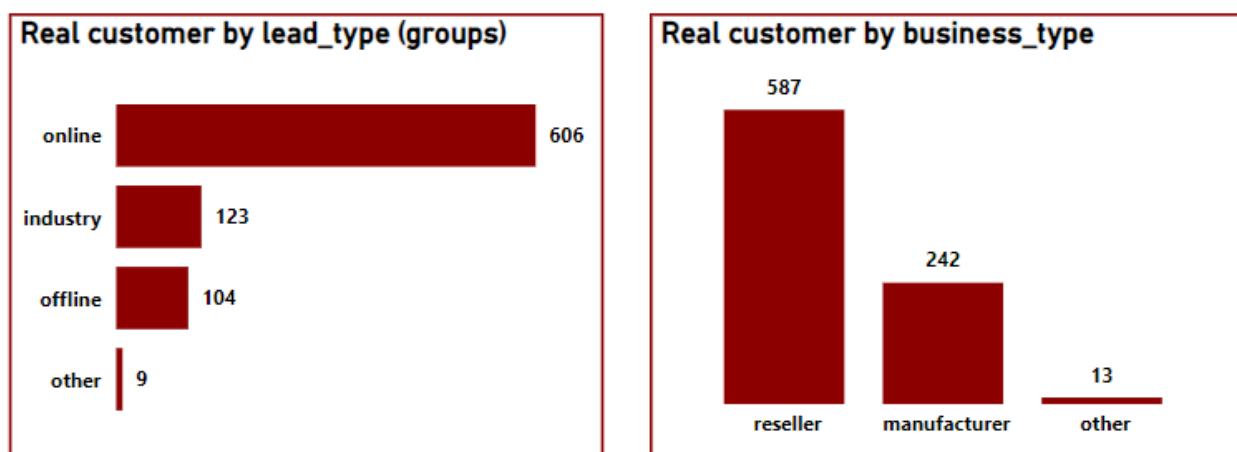
5.4.1.4 Trang Lead conversion analysis

5.4.1.4.1 Tạo visual thống kê



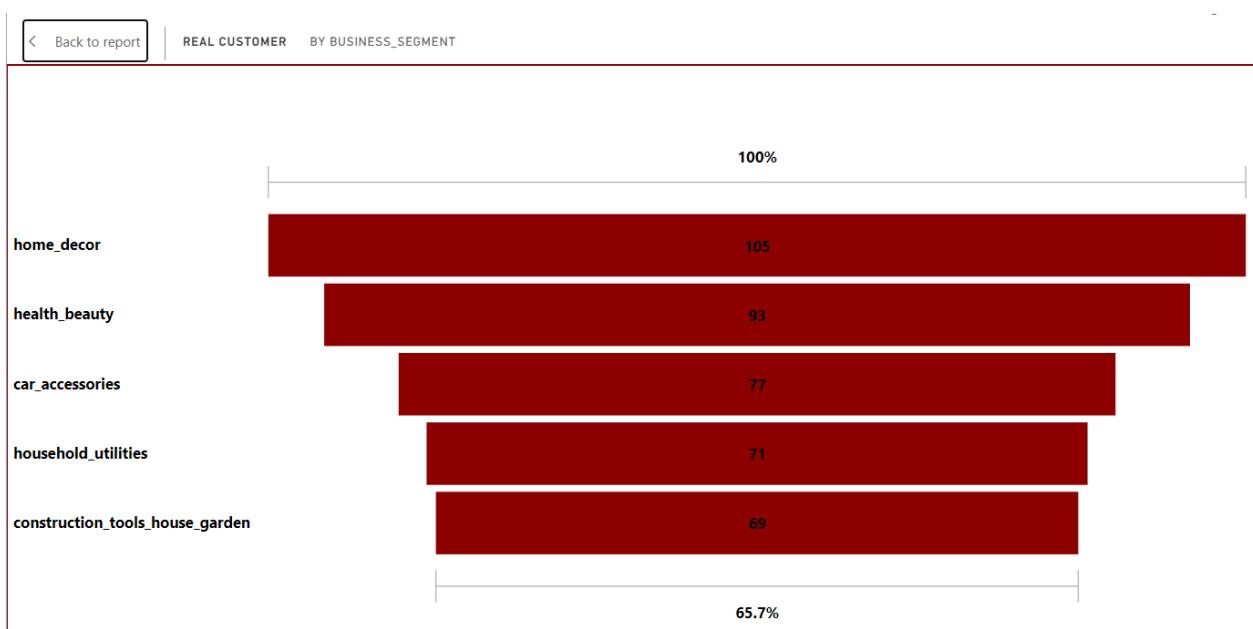
Hình 5.23: Card thống kê chuyển đổi khách hàng

5.4.1.4.2 Tạo visual



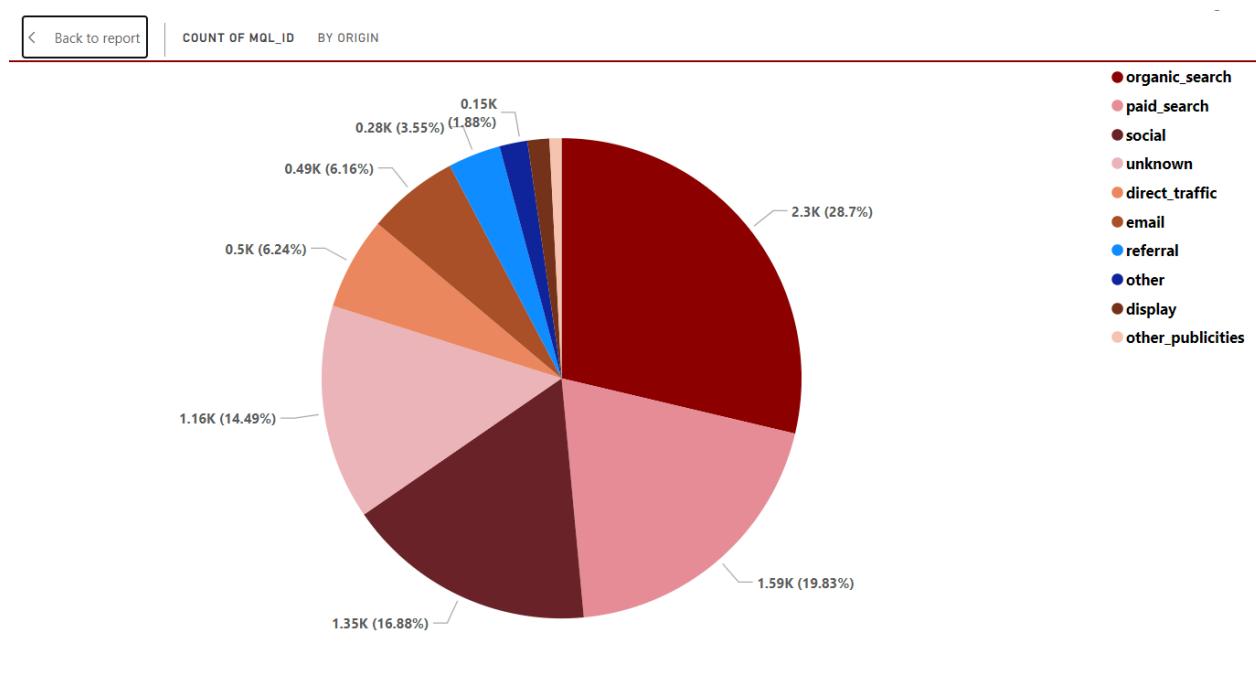
Hình 5.24: Biểu đồ số lượng real customer theo nguồn tiếp cận và loại hình kinh doanh

5.4.1.4.3 Tạo visual thể hiện tỷ lệ khách hàng thực tế theo ngành hàng



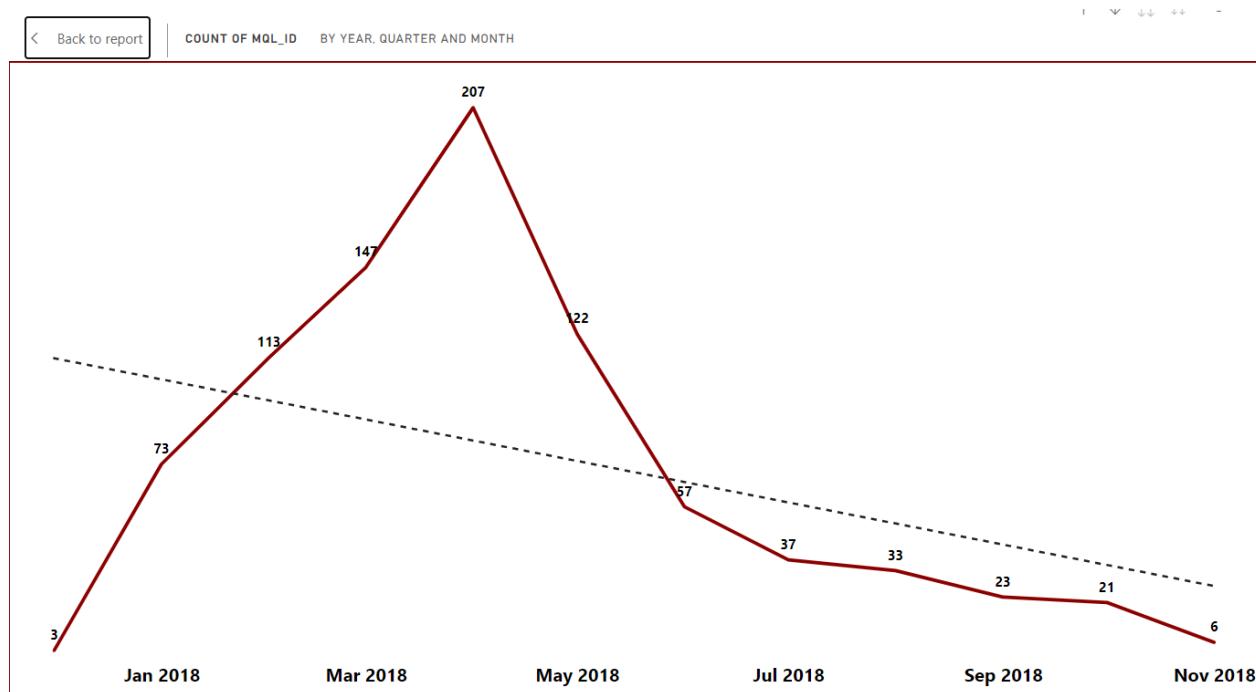
Hình 5.25: Biểu đồ thể hiện tỷ lệ khách hàng thực tế theo ngành hàng

5.4.1.4.4 Tạo visual tỷ lệ khách hàng theo nguồn tiếp cận



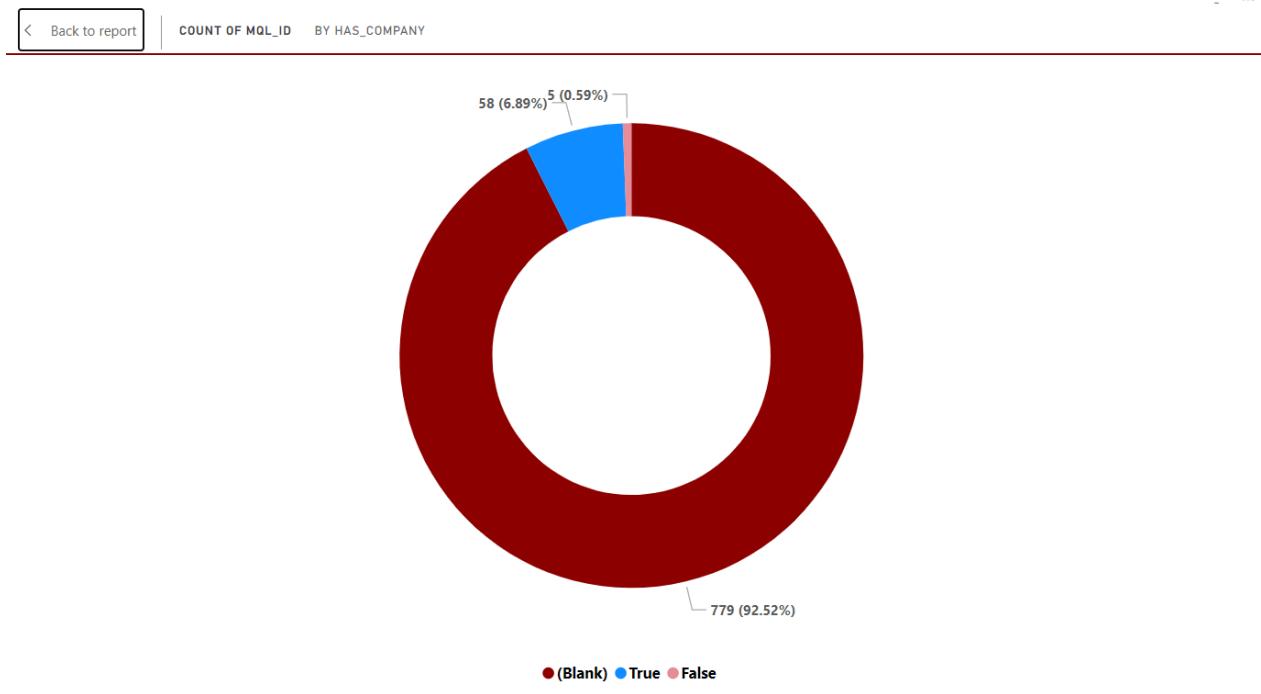
Hình 5.26: Biểu đồ tỷ lệ khách hàng theo nguồn tiếp cận

5.4.1.4.5 Tạo visual xu hướng số lượng real customer theo thời gian



Hình 5.27: Biểu đồ xu hướng số lượng real customer theo thời gian

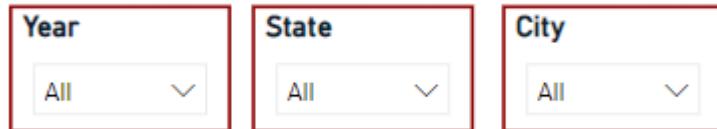
5.4.1.4.6 Tạo visual tỷ lệ real customers có hay không có công ty



Hình 5.28: Biểu đồ tỷ lệ real customers có hay không có công ty

5.4.1.5 Trang Order delivery

5.4.1.5.1 Tạo Slicer theo năm, bang, thành phố



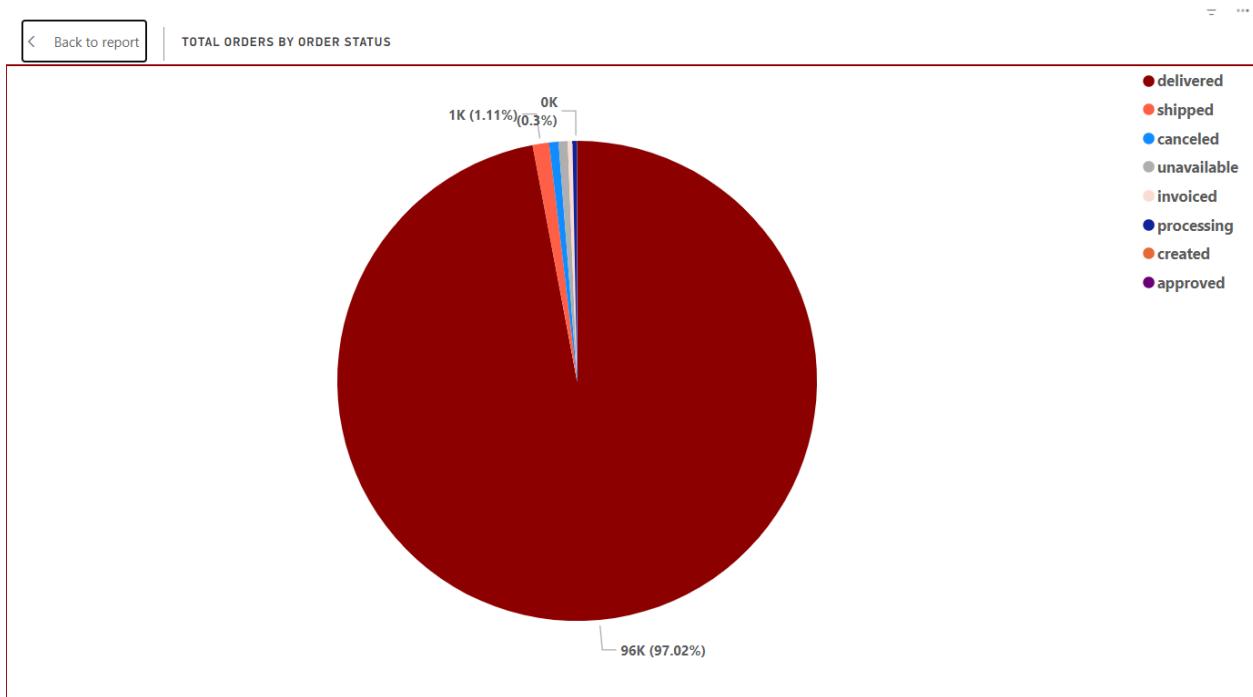
Hình 5.29: Slicer theo năm, bang, thành phố

5.4.1.5.2 Tạo card thống kê số giờ (ngày) trung bình delay



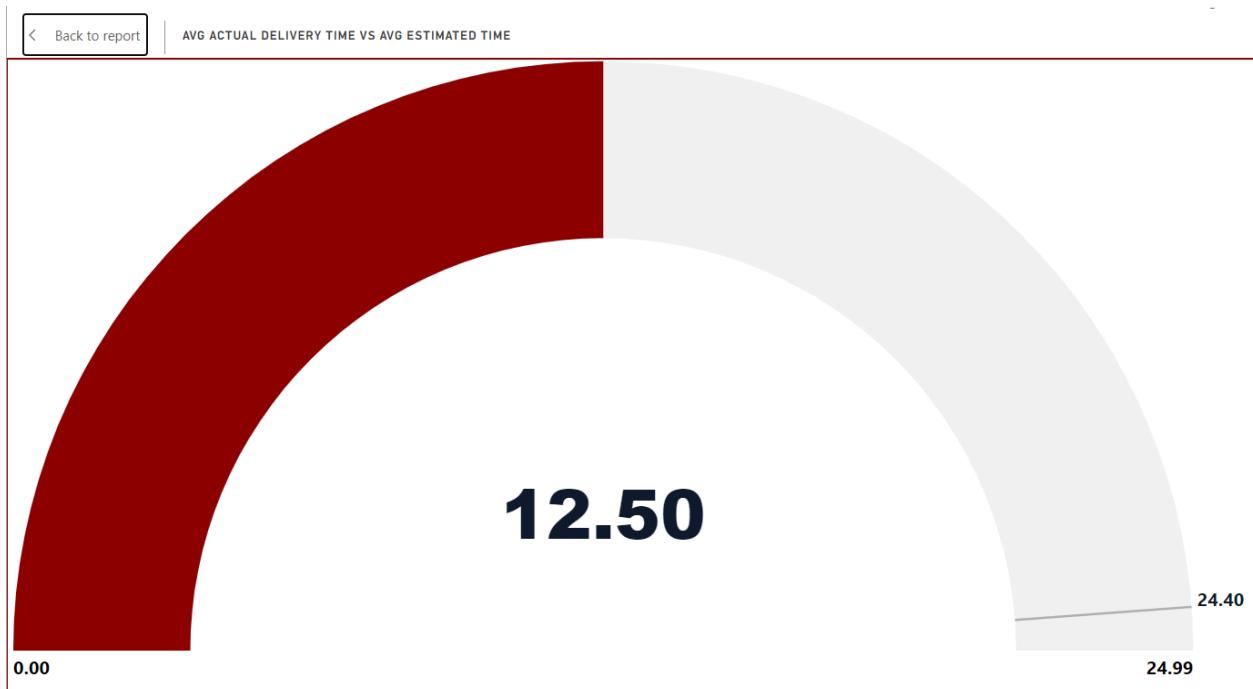
Hình 5.30: Card thống kê số giờ (ngày) trung bình delay

5.4.1.5.3 Tạo visual tỷ lệ đơn hàng theo trạng thái đơn hàng



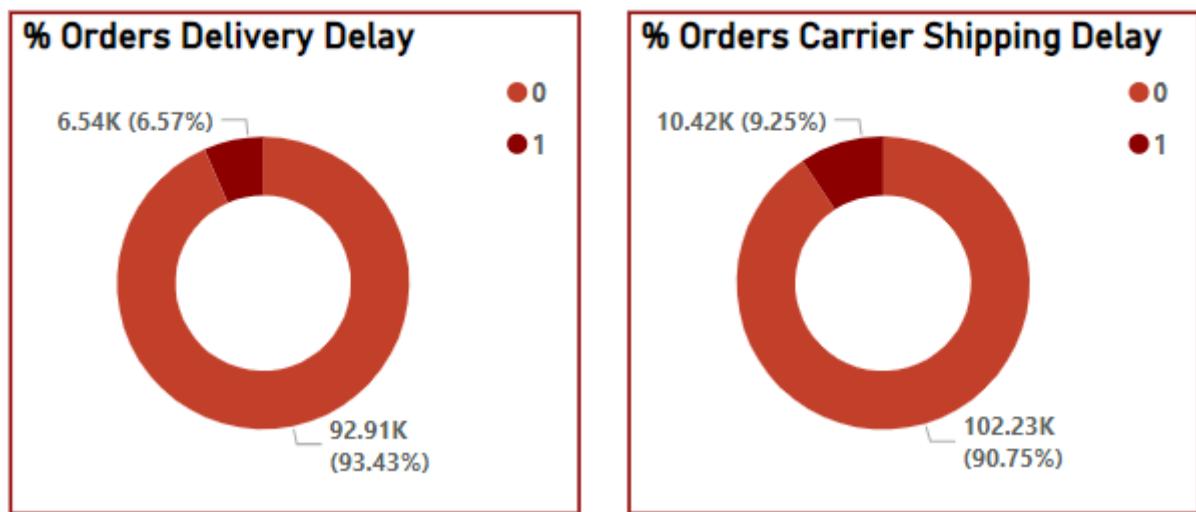
Hình 5.31: Biểu đồ tỷ lệ đơn hàng theo trạng thái đơn hàng

5.4.1.5.4 Tạo visual thời gian giao hàng



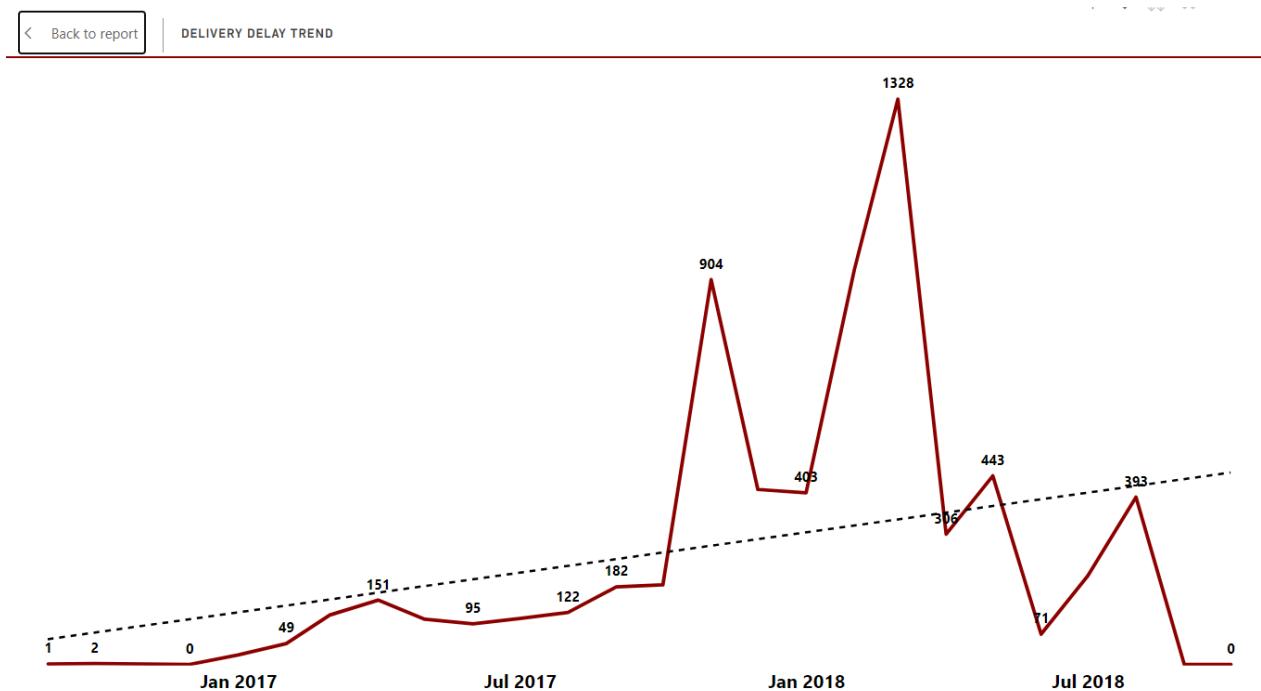
Hình 5.32: Biểu đồ gauge thời gian giao hàng

5.4.1.5.5 Tạo visual tỷ lệ giao hàng trễ



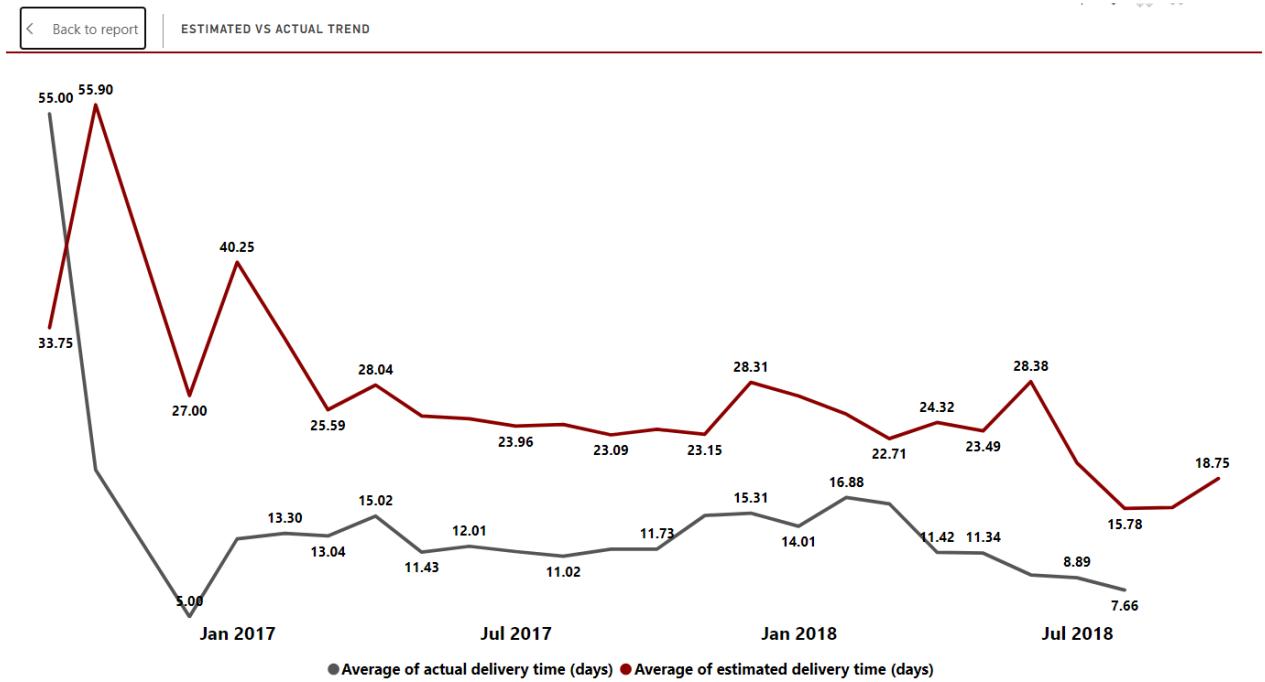
Hình 5.33: Biểu đồ tỷ lệ giao hàng trễ

5.4.1.5.6 Tạo visual phân tích số lượng đơn hàng giao trễ theo thời gian



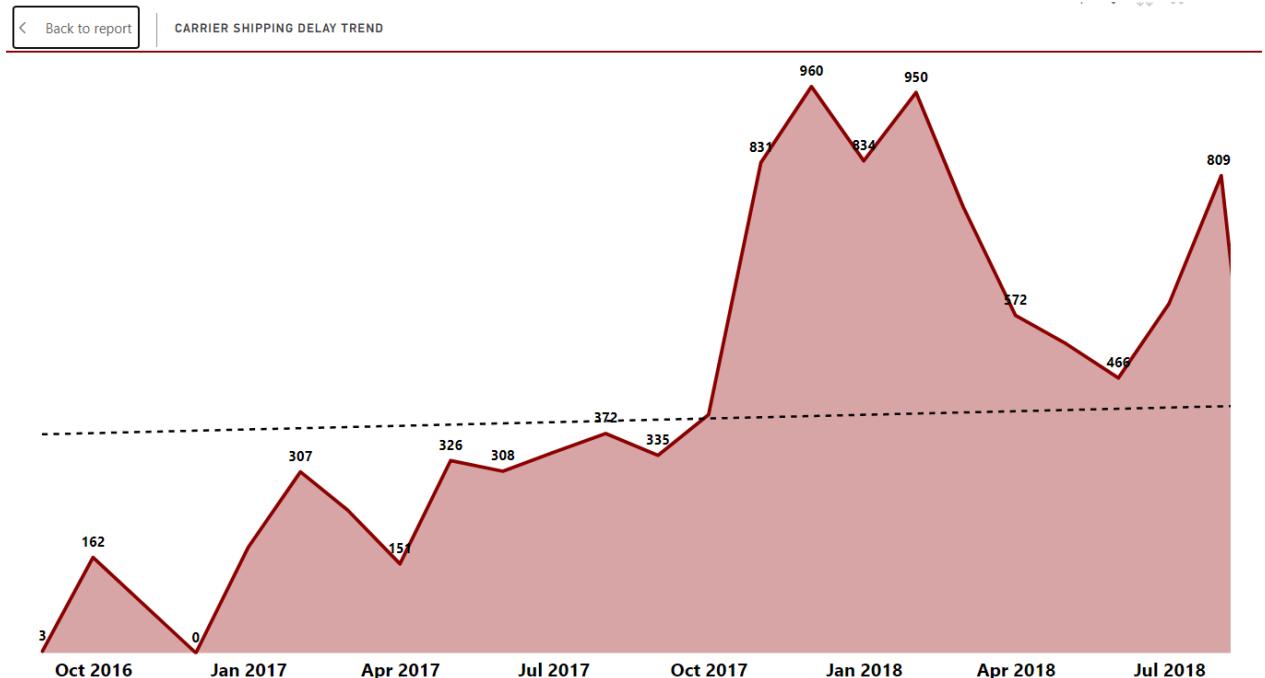
Hình 5.34: Biểu đồ phân tích số lượng đơn hàng giao trễ theo thời gian

5.4.1.5.7 Tạo visual so sánh xu hướng thời gian giao hàng thực tế và dự kiến



Hình 5.35: Biểu đồ so sánh xu hướng thời gian giao hàng thực tế và dự kiến

5.4.1.5.8 Tạo visual xu hướng giao hàng trễ của đơn vị vận chuyển



Hình 5.36: Biểu đồ xu hướng giao hàng trễ của đơn vị vận chuyển

5.4.1.6 Trang Segmentation

5.4.1.6.1 Tạo bảng thông tin khách hàng

[Back to report](#)

customer_unique_id	Customer_Clusters	Orders	Price	Freight	Payment value	Installments
fff5eb4918b2bf4b2da476788d42051c	Cluster2	1	R\$1,050.00	R\$1,794.96	2844.96	1
066ee6b9c6fc284260ff9a1274a82ca7	Cluster1	1	R\$419.40	R\$1,002.29	1421.69	1
eae0a83d752b1dd32697e0e7b4221656	Cluster2	2	R\$1,821.73	R\$961.28	2783.01	16
ef7361e14a64f77990f58e9c571e2f9a	Cluster2	1	R\$1,380.00	R\$711.33	2091.33	10
5a494c648fde2d1ec4eb614274ea7159	Cluster2	2	R\$1,065.00	R\$638.38	1703.38	2
c8460e4251689ba205045f3ea17884a1	Cluster2	4	R\$4,080.00	R\$575.88	4655.91	24
fffcf5a5ff07b0908bd4e2dbc735a684	Cluster2	1	R\$1,570.00	R\$497.42	2067.42	10
527f7f3237fb1397c459701bc765b6f0	Cluster1	1	R\$1,520.00	R\$497.08	2017.08	1
6d394722d5fc5e721aee6875a218d8db	Cluster1	1	R\$1,559.92	R\$479.28	2039.20	1
6411590d91c48640cb07e72fbb4a359e	Cluster2	1	R\$1,161.00	R\$458.73	1619.73	10
f9172a6495d46451776be8bc8e46032d	Cluster2	1	R\$859.66	R\$456.47	1316.14	10
3895f60f6e6a89e5cfb7b72ffdcdf7e0	Cluster2	1	R\$1,767.80	R\$436.24	2204.04	6
97734fdca127fddcb5f92f841690c3f2	Cluster2	1	R\$979.00	R\$409.68	1388.68	12
09ed9b91e77dcb56c1c1e7fdf7fc100c	Cluster1	1	R\$479.94	R\$401.58	881.52	1
cb26f33c1b5055d997cc798dc94790d5	Cluster2	1	R\$1,038.00	R\$389.25	1427.25	7

Hình 5.37: Bảng thông tin chi tiết khách hàng

5.4.1.6.2 Tạo bảng thông tin nhà phân phối

[Back to report](#)

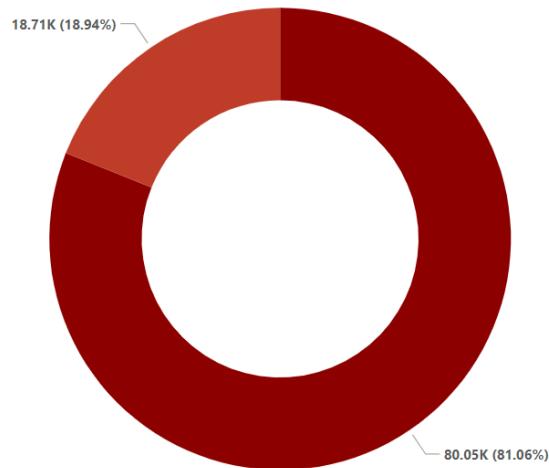
seller_id	Products	Price	Freight	Seller_Clusters
0015a82c2db000af6aaaf3ae2ecb0532	3	R\$2,685.00	R\$63.06	Cluster1
001cca7ae9ae17fb1caed9dfb1094831	239	R\$25,080.03	R\$8,854.14	Cluster3
001e6ad469a905060d959994f1b41e4f	1	R\$250.00	R\$17.94	Cluster1
002100f778ceb8431b7a1020ff7ab48f	55	R\$1,234.50	R\$793.66	Cluster1
003554e2dce176b5555353e4f3555ac8	1	R\$120.00	R\$19.38	Cluster1
004c9cd9d87a3c30c522c48c4fc07416	170	R\$19,712.71	R\$3,551.23	Cluster3
00720abe85ba0859807595bbf045a33b	26	R\$1,007.50	R\$315.98	Cluster1
00ab3eff1b5192e5f1a63bcecfee11c8	1	R\$98.00	R\$12.08	Cluster1
00d8b143d12632bad99c0ad66ad52825	1	R\$86.00	R\$51.10	Cluster1
00ee68308b45bc5e2660cd833c3f81cc	172	R\$20,260.00	R\$3,180.66	Cluster3
00fc707aaaad2d31347cf883cd2dfe10	135	R\$12,684.90	R\$2,285.09	Cluster1
010543a62bd80aa422851e79a3bc7540	2	R\$1,416.00	R\$31.95	Cluster1

Hình 5.38: Bảng thông tin chi tiết nhà phân phối

5.4.1.6.3 Tạo visual tỷ lệ khách hàng sau phân loại

[Back to report](#)

CUSTOMER CLUSTERS



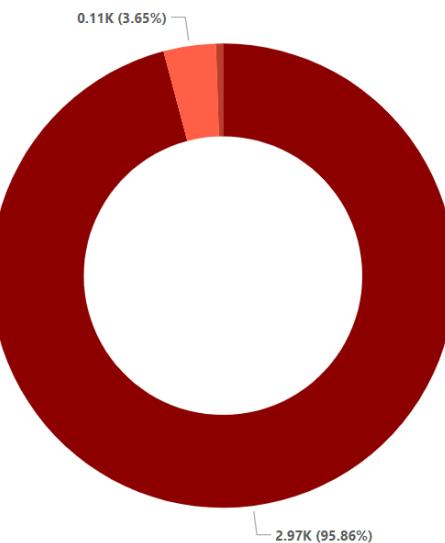
● Cluster1 ● Cluster2

Hình 5.39: Biểu đồ tỷ lệ khách hàng sau phân loại

5.4.1.6.4 Tạo visual tỷ lệ nhà phân phối sau phân loại

[Back to report](#)

SELLER CLUSTERS

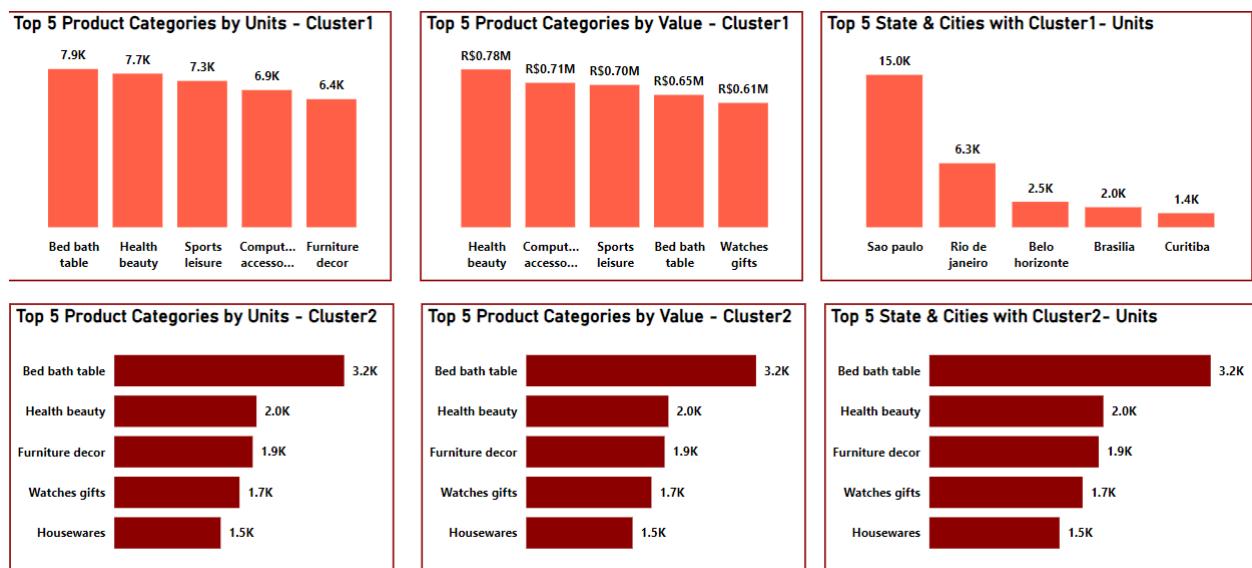


● Cluster1 ● Cluster3 ● Cluster2

Hình 5.40: Biểu đồ tỷ lệ nhà phân phối sau phân loại

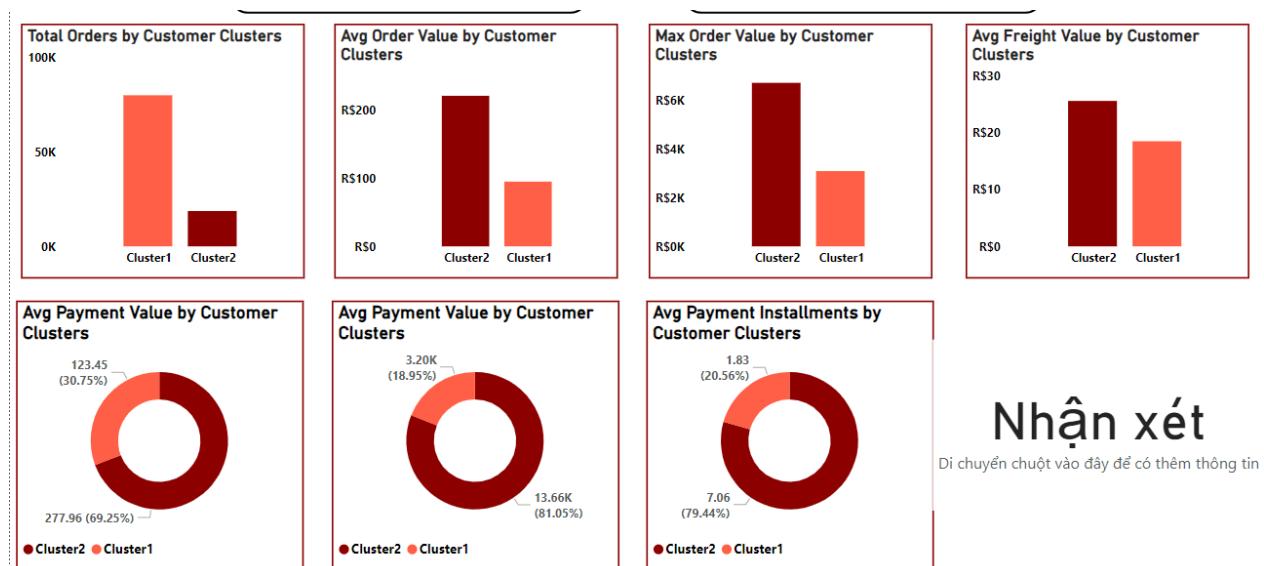
5.4.1.7 Trang Customer segmentation

5.4.1.7.1 Tạo visual phân tích từng loại khách hàng (trang 1)



Hình 5.41: Các biểu đồ phân tích từng loại khách hàng (trang 1)

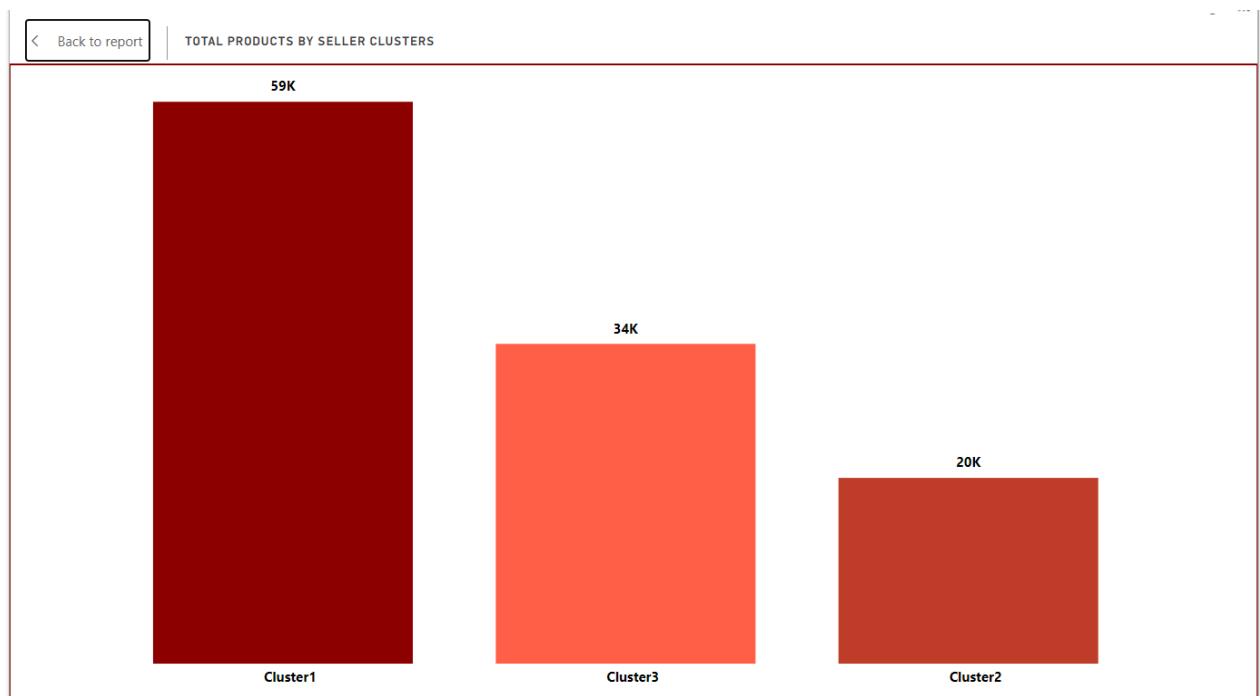
5.4.1.7.2 Tạo visual phân tích từng loại khách hàng (trang 2)



Hình 5.42: Các biểu đồ phân tích từng loại khách hàng (trang 2)

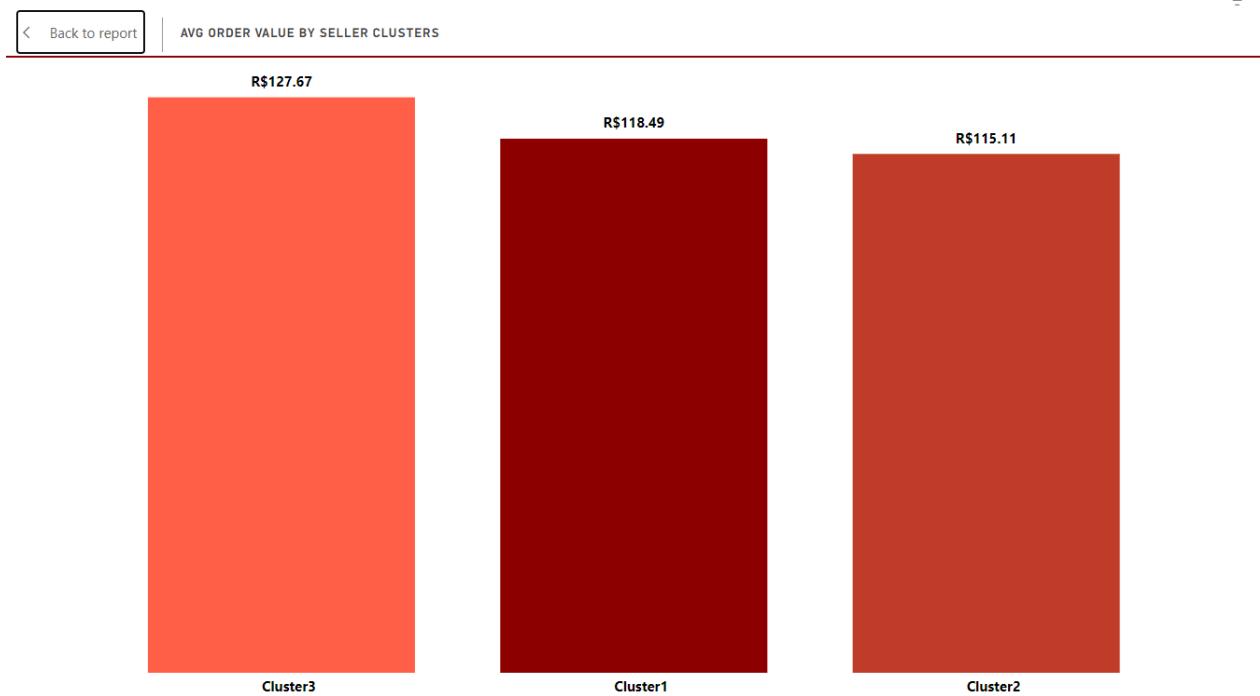
5.4.1.8 Trang Seller Segmentation

5.4.1.8.1 Tạo visual số lượng sản phẩm bán được của từng loại nhà phân phối



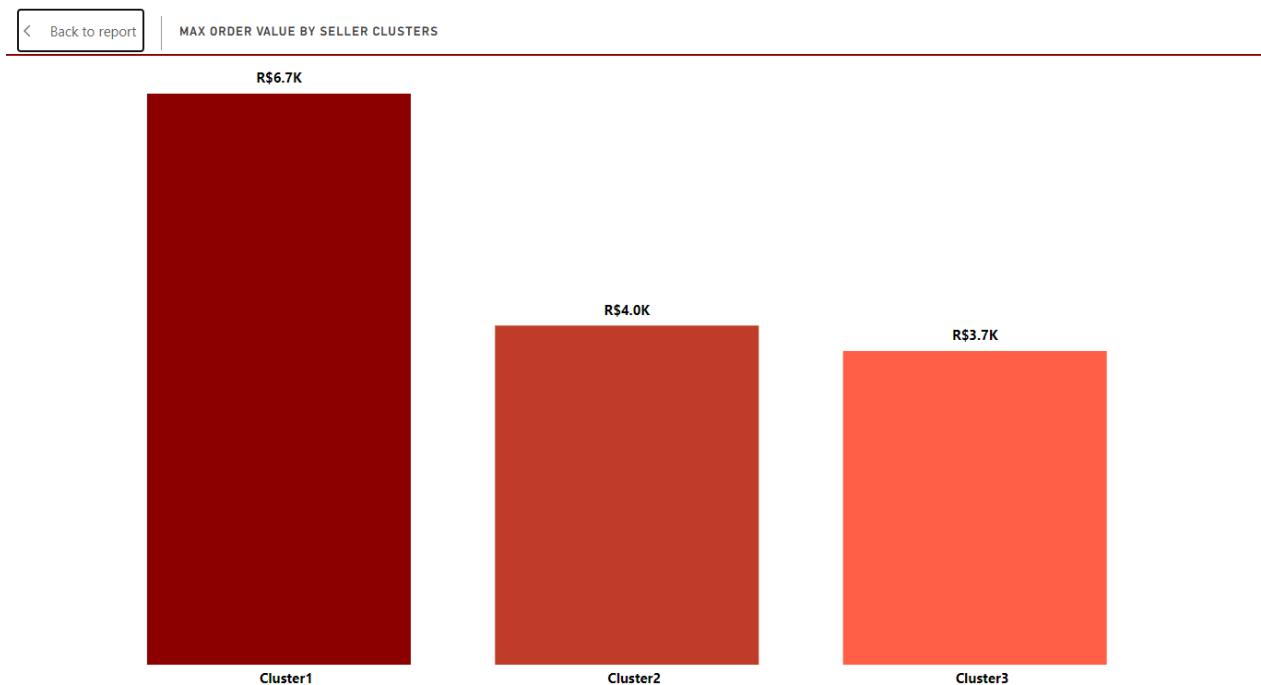
Hình 5.43: Biểu đồ số lượng sản phẩm bán được của từng loại nhà phân phối

5.4.1.8.2 Tạo visual giá trị trung bình mỗi đơn hàng của từng loại nhà phân phối



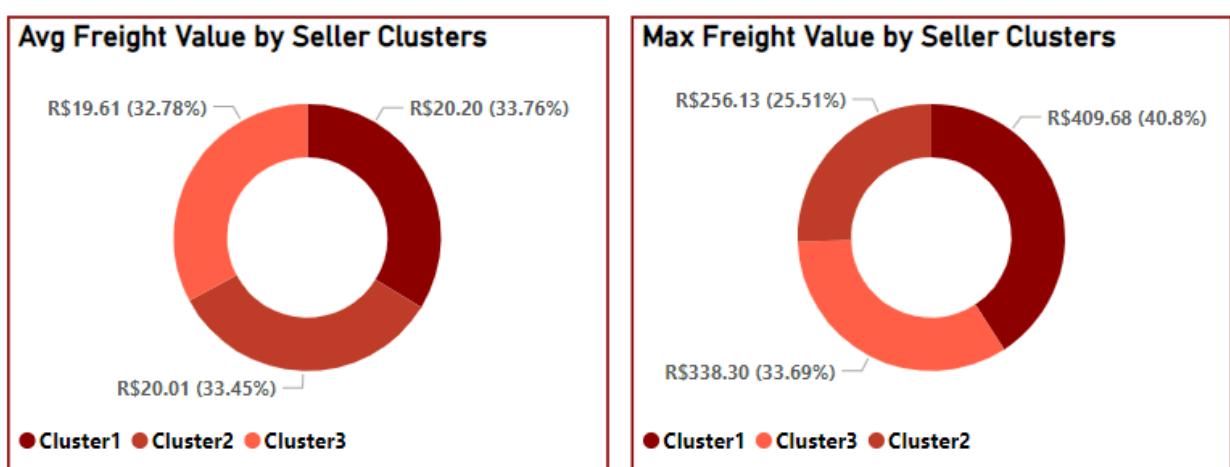
Hình 5.44: Biểu đồ giá trị trung bình mỗi đơn hàng của từng loại nhà phân phối

5.4.1.8.3 Tạo visual giá đơn hàng cao nhất của từng loại nhà phân phối



Hình 5.45: Biểu đồ giá đơn hàng cao nhất của từng loại nhà phân phối

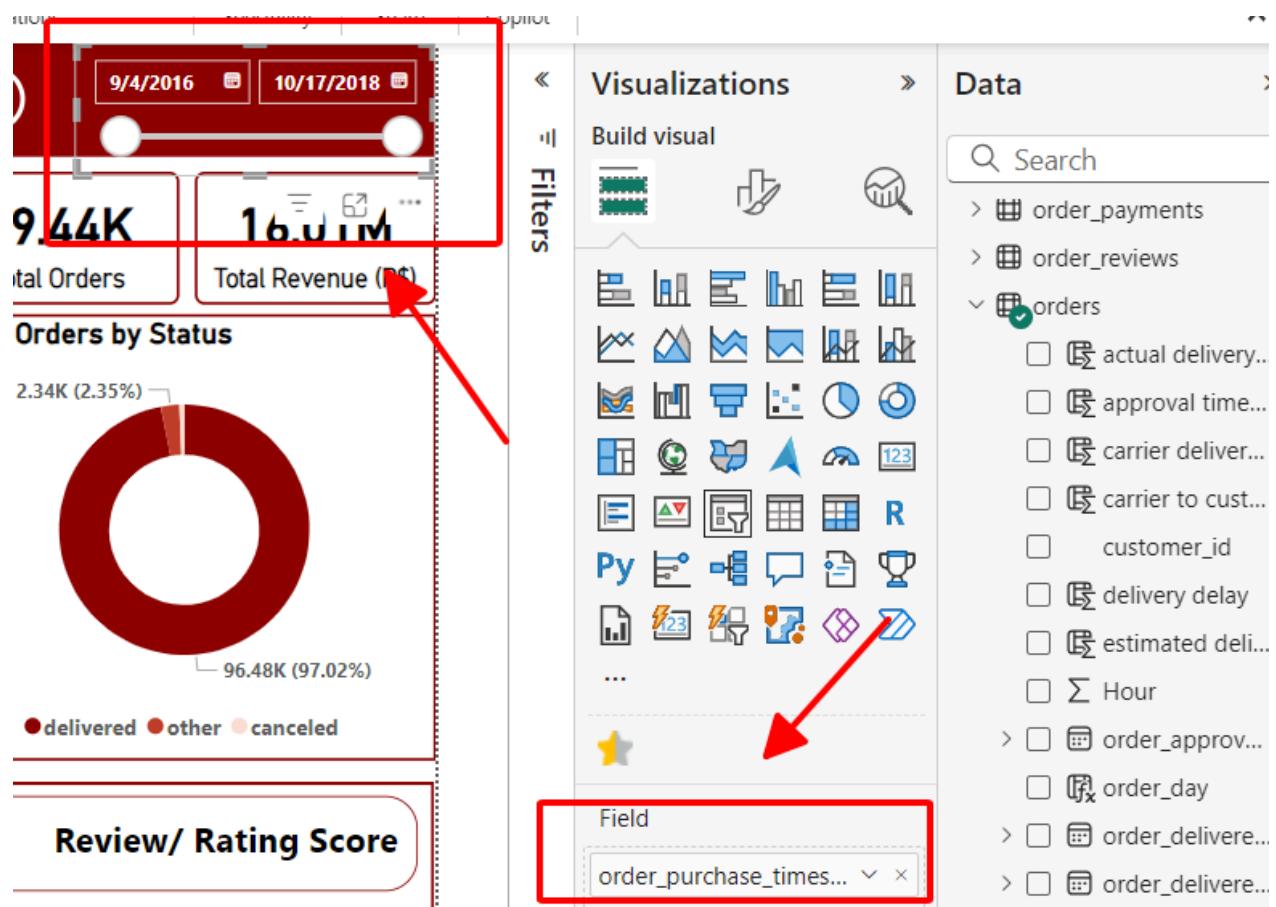
5.4.1.8.4 Tạo visual tỷ lệ giá trung bình, giá lớn nhất của từng nhà phân phối



Hình 5.46: Biểu đồ tỷ lệ giá trung bình, giá lớn nhất của từng nhà phân phối

5.4.2 TẠO VISUAL THỐNG KÊ TỔNG THỂ

5.4.2.1 Tạo visual filter theo ngày giao dịch



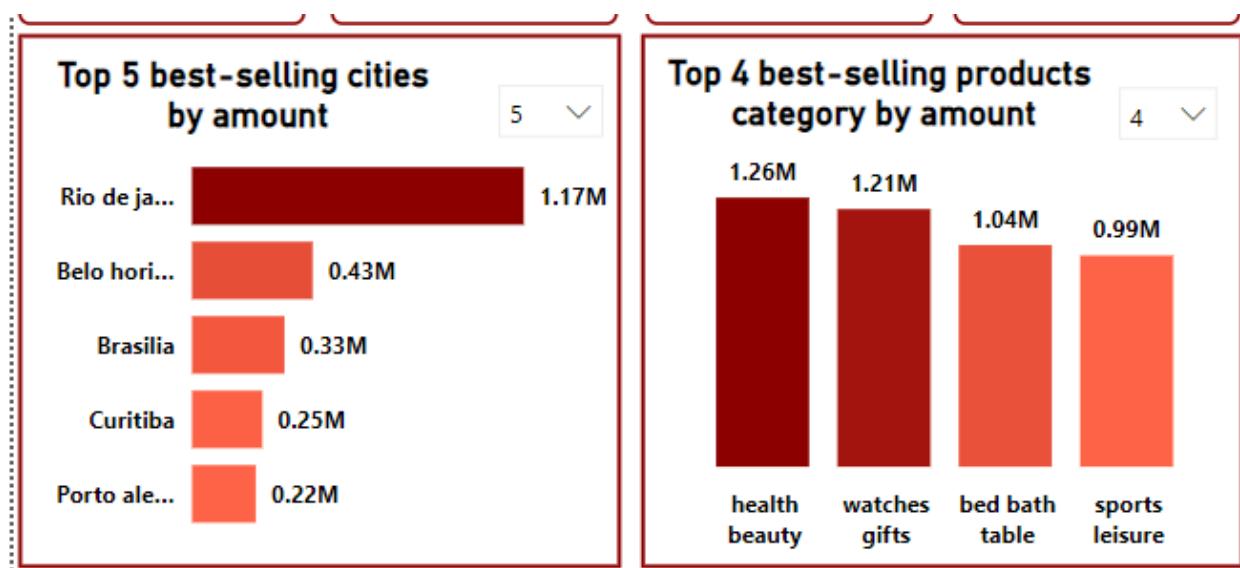
Hình 5.47: Slicer theo thời gian (Ngày giao dịch - Order purchase time)

5.4.2.2 Tạo visual thống kê (tính tổng theo từng đối tượng)



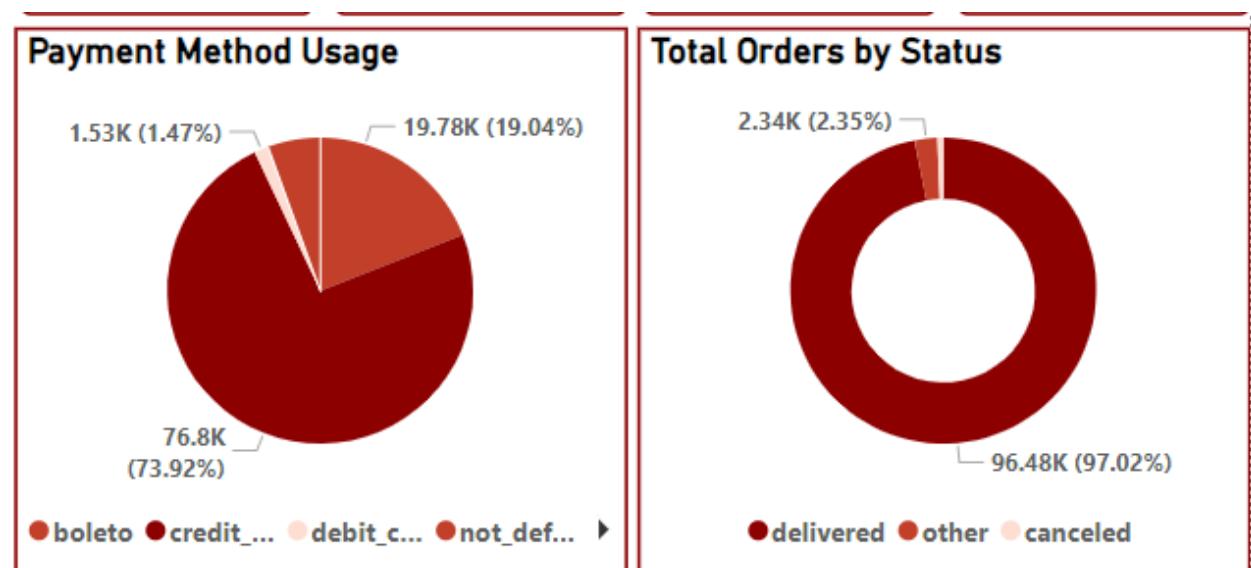
Hình 5.48: Card tổng quan về số liệu kinh doanh

5.4.2.3 Tạo visual thống kê top sản phẩm/ thành phố có doanh thu cao nhất



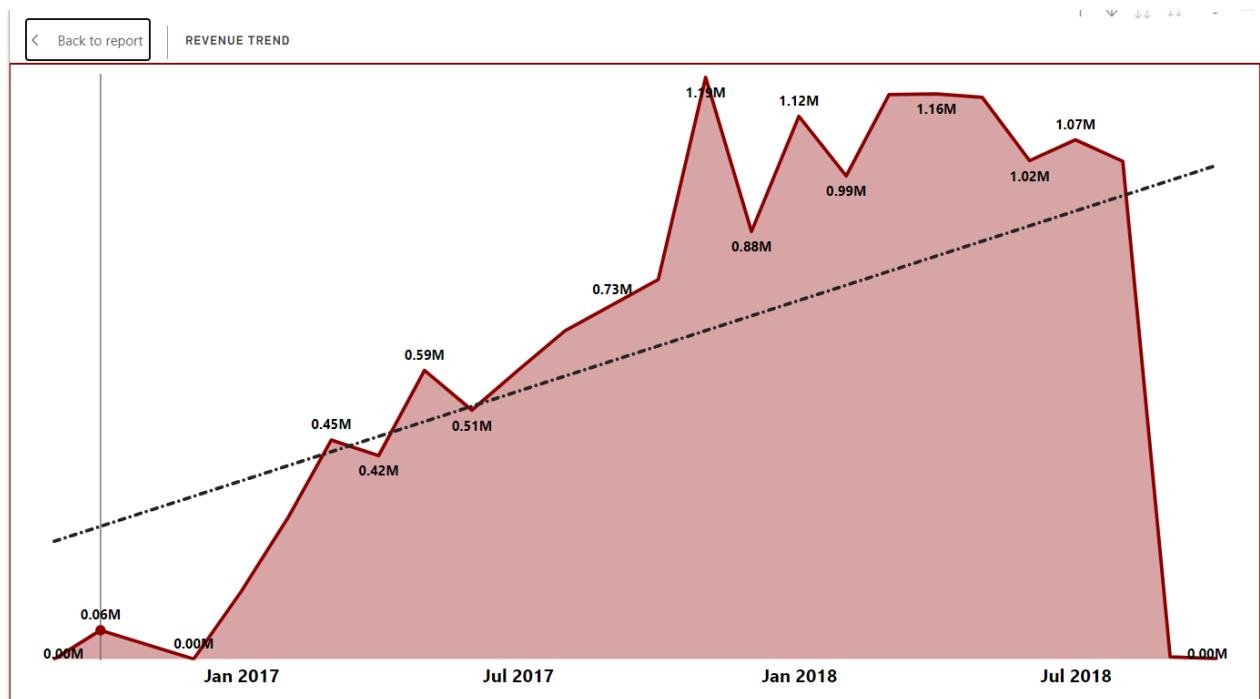
Hình 5.49: Biểu đồ top doanh thu theo thành phố và sản phẩm

5.4.2.4 Tạo visual tỷ lệ phương thức thanh toán và trạng thái đơn hàng



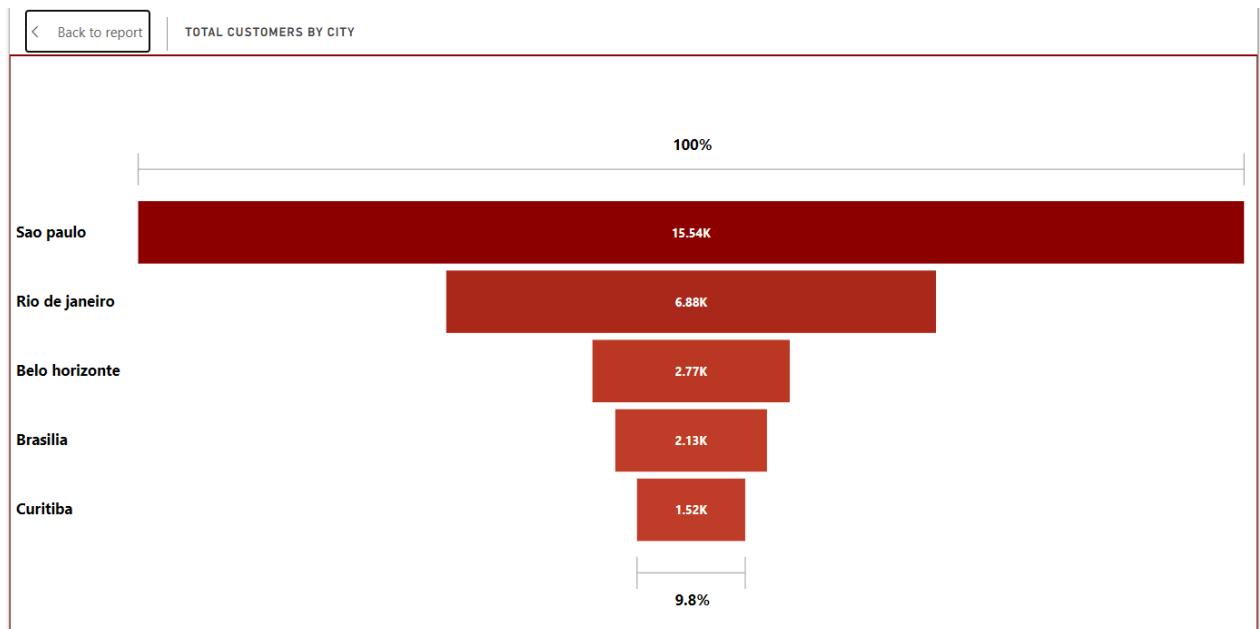
Hình 5.50: Biểu đồ tỷ lệ phương thức thanh toán và trạng thái đơn hàng

5.4.2.5 Tạo visual xu hướng doanh thu theo thời gian



Hình 5.51: Biểu đồ xu hướng doanh thu theo thời gian

5.4.2.6 Tạo visual tổng khách hàng theo thành phố



Hình 5.52: Biểu đồ tổng khách hàng theo thành phố (top 5)

5.4.2.7 Tạo visual thể hiện tổng quan sự đánh giá, nhận xét của khách hàng



Hình 5.53: Trực quan tổng quan đánh giá của khách hàng

6 XÂY DỰNG BÁO CÁO

6.1 DASHBOARD VÀ REPORT

Tối ưu hóa dashboard và report là bước quan trọng để đảm bảo thông tin được truyền tải một cách rõ ràng, dễ hiểu và hiệu quả nhất đến người sử dụng. Tối ưu hóa dashboard và report không chỉ giúp cải thiện hiệu suất mà còn đảm bảo thông tin được trình bày một cách rõ ràng, dễ hiểu và hấp dẫn. Bằng cách tập trung vào việc hiển thị các thông tin quan trọng, sử dụng các biểu đồ phù hợp và cung cấp các yếu tố tương tác, bạn có thể tạo ra các dashboard và report hiệu quả, hỗ trợ tốt cho việc ra quyết định dựa trên dữ liệu.

Tối ưu hóa dashboard

Giao diện trực quan và hấp dẫn

- **Sử dụng bố cục đơn giản:** Tránh quá nhiều chi tiết và yếu tố không cần thiết.
- **Sử dụng màu sắc hợp lý:** Chọn màu sắc tương phản tốt để làm nổi bật các thông tin quan trọng nhưng không gây rối mắt.

Tập trung vào thông tin quan trọng

- **Hiển thị các KPI chính:** Đặt các chỉ số quan trọng nhất lên phía trên và trung tâm của dashboard.
- **Sử dụng biểu đồ phù hợp:** Chọn đúng loại biểu đồ cho từng loại dữ liệu (ví dụ: biểu đồ đường cho xu hướng, biểu đồ cột cho so sánh).

Tương tác và bộ lọc

- **Thêm các bộ lọc:** Cho phép người dùng lọc dữ liệu theo các tiêu chí khác nhau (ví dụ: theo thời gian, theo hãng xe).

- **Sử dụng các yếu tố tương tác:** Cho phép người dùng tương tác với các biểu đồ để xem chi tiết hơn (ví dụ: drill-down, hover để xem thêm thông tin).

Hiệu suất

- **Tối ưu hóa dữ liệu nguồn:** Chỉ lấy dữ liệu cần thiết để tránh làm chậm dashboard.
- **Sử dụng các biện pháp tối ưu hóa:** Như tính toán trước các measure phức tạp và lưu vào bộ nhớ đệm.

6.2 XÂY DỰNG BÁO CÁO

6.2.1 DASHBOARD VÀ REPORT

Cách tối ưu hóa Dashboard:

- **Sử dụng trực quan phù hợp:** Chọn biểu đồ và đồ thị phù hợp để truyền đạt thông tin một cách hiệu quả. Ví dụ, sử dụng biểu đồ đường để theo dõi xu hướng theo thời gian, biểu đồ cột để so sánh các danh mục và biểu đồ tròn để hiển thị tỷ lệ phần trăm.
- **Sắp xếp hợp lý:** Sắp xếp các biểu đồ và đồ thị một cách logic và dễ hiểu. Đặt các thông tin quan trọng nhất ở vị trí nổi bật và sử dụng tiêu đề, chú thích rõ ràng để giải thích nội dung.
- **Thiết kế đơn giản:** Tránh sử dụng quá nhiều chi tiết và đồ họa không cần thiết. Thiết kế dashboard đơn giản, dễ nhìn và tập trung vào thông tin quan trọng.

6.2.2 DASHBOARD

Trong dự án này, chúng tôi sẽ tạo ra các loại bảng điều khiển khác nhau nhằm mục đích phục vụ các mục đích cụ thể và đáp ứng nhu cầu của các bên liên quan. Việc tạo ra các loại bảng điều khiển khác nhau giúp công ty Olist có cái nhìn toàn diện và chi tiết về các khía cạnh khác nhau của hoạt động kinh doanh. Mỗi bảng điều khiển phục vụ một mục tiêu cụ thể, từ tổng doanh thu, hiệu suất đơn hàng, hành vi khách hàng, hiệu quả nhà bán hàng đến phân tích tiếp thị. Điều này không chỉ giúp cải thiện hiệu quả quản lý và quyết định mà còn tăng cường khả năng cạnh tranh và phát triển sự bền vững của công ty.

Dưới đây là các loại dashboard:

Dashboard tổng quan (overview dashboard)



Hình 6.1: Overview

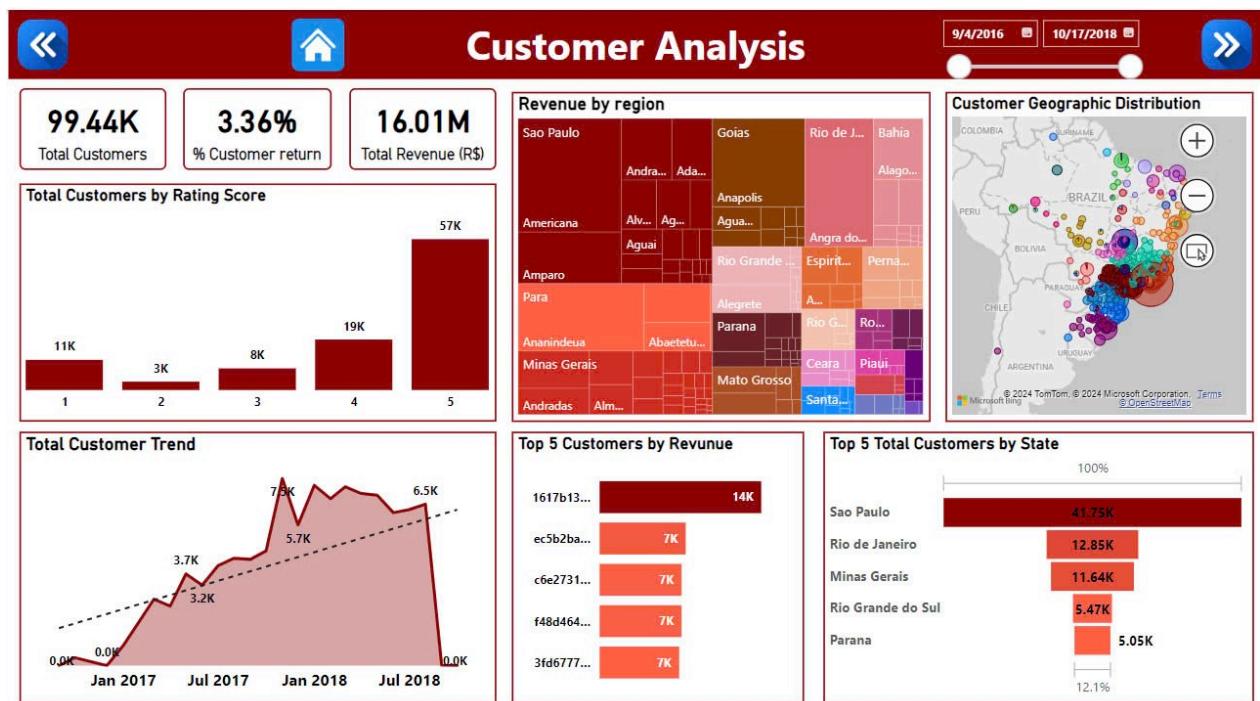
Nội dung:

- Biểu đồ cột ngang thể hiện top 5 thành phố có doanh thu cao nhất.
- Biểu đồ cột dọc thể hiện top 4 danh mục sản phẩm bán chạy nhất.
- Biểu đồ tròn thể hiện tỷ lệ phần trăm các phương thức thanh toán được sử dụng
- Biểu đồ tròn thể hiện tỷ lệ các trạng thái đơn hàng
- Biểu đồ đường thể hiện xu hướng thay đổi doanh thu theo thời gian.
- Biểu đồ cột ngang thể hiện tổng số khách hàng tại các thành phố lớn.
- Biểu tượng sao thể hiện tổng số đánh giá và tỷ lệ phần trăm các đánh giá tích cực.

Mục đích:

Dashboard này giúp theo dõi hiệu suất kinh doanh trên nền tảng Olist, cung cấp cái nhìn sâu sắc về hành vi tiêu dùng và xu hướng bán hàng. Nhờ đó, Olist có thể đánh giá hiệu quả chiến lược kinh doanh, mức độ ưu tiên hoạt động và xác định các khu vực hoặc danh mục sản phẩm tiềm năng. Đồng thời, thông tin này giúp cải thiện dịch vụ khách hàng, tăng khả năng cạnh tranh và hỗ trợ các quyết định về thời hạn chiến lược.

Dashboard phân tích khách hàng (customer demographics dashboard)



Hình 6.2: Phân tích khách hàng

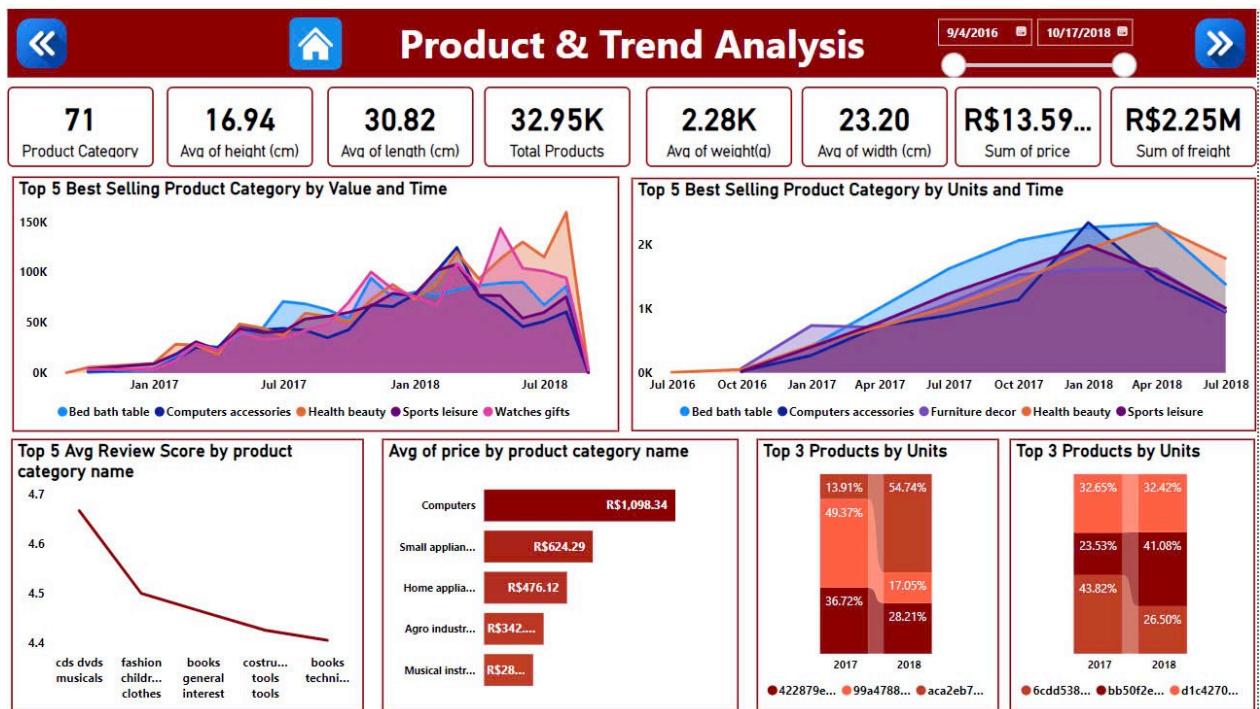
Nội dung:

- Biểu đồ cột hiển thị phân bố khách hàng theo đánh giá
- Biểu đồ đường thể hiện số lượng khách hàng theo thời gian
- Biểu đồ treemap hiển thị doanh thu phân bố theo các khu vực
- Bản đồ thể hiện phân bố khách hàng trên toàn Brazil
- Biểu đồ cột ngang hiển thị 5 khách hàng đóng góp doanh thu cao nhất và hiển thị số lượng khách hàng tại 5 bang có số khách hàng lớn nhất.

Mục đích:

Dashboard này hỗ trợ Olist theo dõi hiệu quả kinh doanh, phân tích hành vi tiêu dùng, xác định khu vực tiềm năng và khách hàng giá trị cao để ưu tiên chiến lược, đồng thời hỗ trợ cải thiện dịch vụ và tăng trưởng bền vững.

Dashboard phân tích sản phẩm và xu hướng (product and trend analysis dashboard)



Hình 6.3: Phân tích sản phẩm và xu hướng

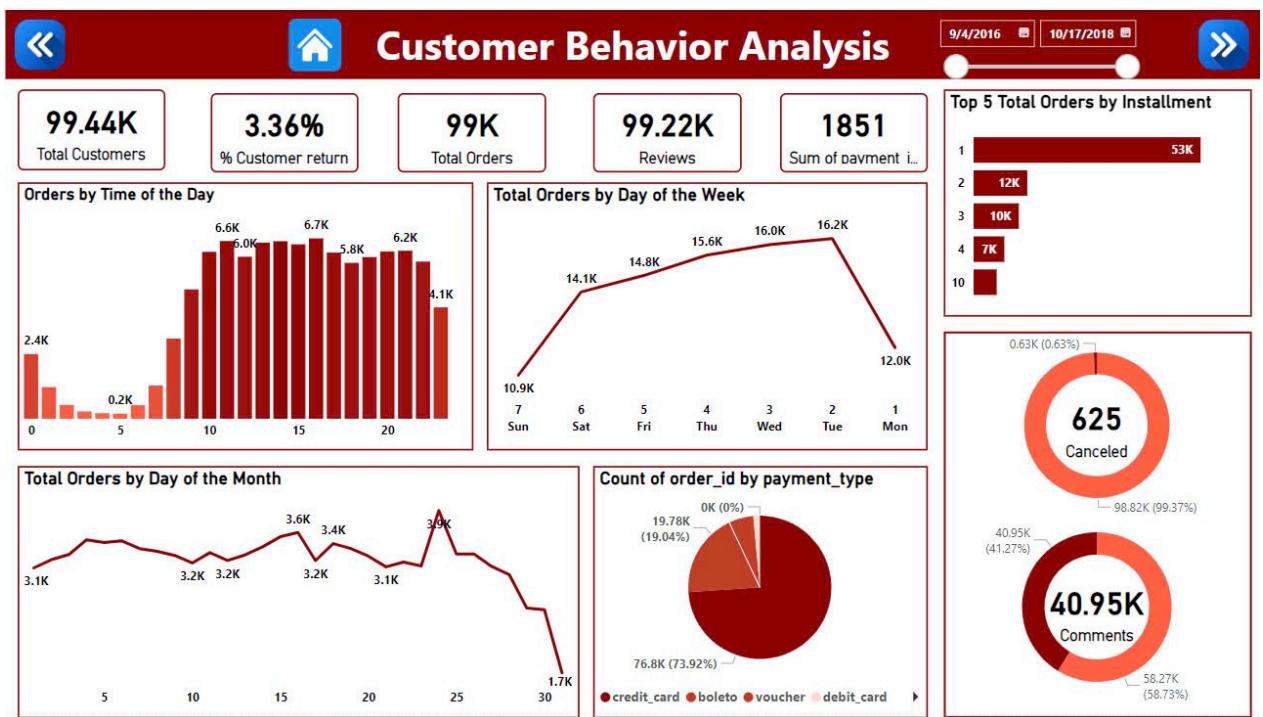
Nội dung:

- Biểu đồ đường hiển thị giá trị bán hàng theo thời gian của 5 danh mục sản phẩm bán chạy nhất.
- Biểu đồ đường hiển thị số lượng sản phẩm bán ra theo thời gian của 5 danh mục bán chạy nhất.
- Biểu đồ đường hiển thị điểm đánh giá trung bình của 5 danh mục sản phẩm có đánh giá cao nhất.
- Biểu đồ cột ngang hiển thị giá trung bình của các danh mục sản phẩm.
- Biểu đồ thanh xếp chòng hiển thị phần trăm đóng góp của các sản phẩm trong số lượng bán ra qua từng năm.

Mục đích:

Dashboard này giúp theo dõi xu hướng bán hàng của các danh mục sản phẩm theo thời gian, phân tích hành vi tiêu dùng qua điểm hài lòng của khách hàng và mức giá trung bình của sản phẩm. Nó cũng hỗ trợ tối ưu hóa chiến lược kinh doanh bằng cách xác định các danh mục và sản phẩm bán chạy nhất, từ đó ưu tiên phát triển và phân bổ nguồn lực hiệu quả. Cuối cùng, dashboard này giúp cải thiện dịch vụ dựa trên phản hồi và đánh giá của khách hàng, nâng cao chất lượng sản phẩm và trải nghiệm người dùng.

Dashboard phân tích hành vi khách hàng (customer behavior analysis dashboard)



Hình 6.4: Phân tích hành vi khách hàng

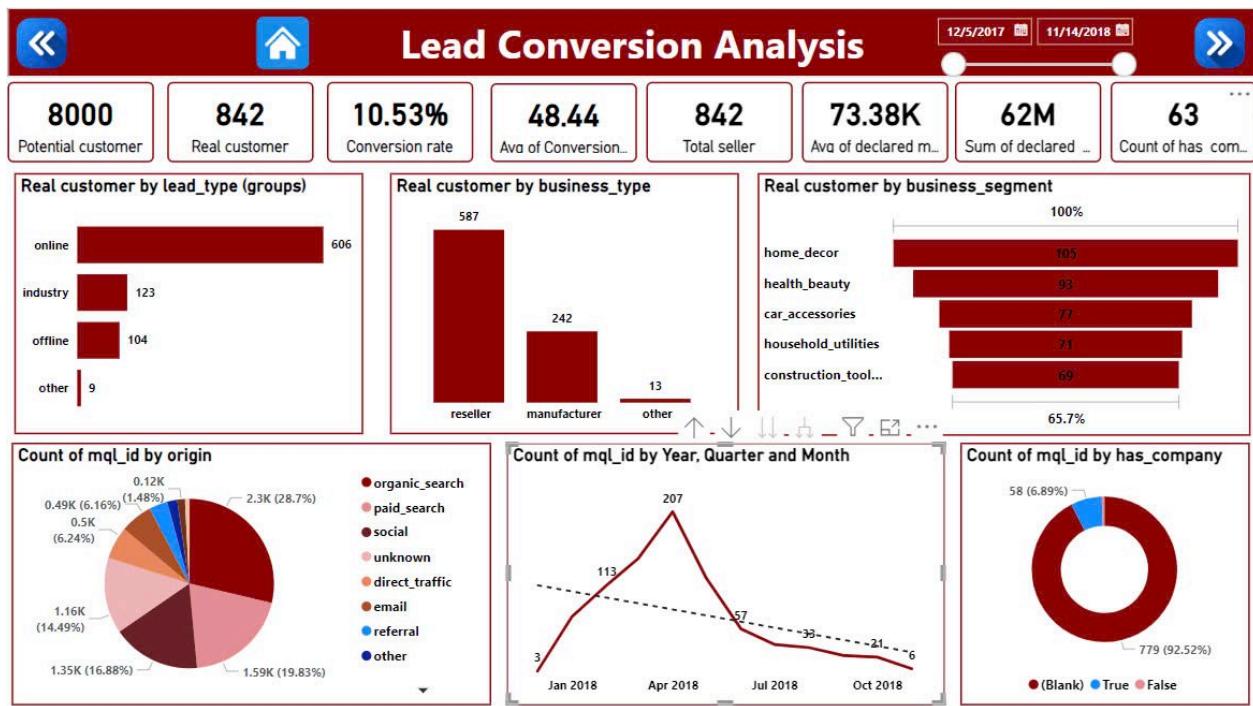
Nội dung:

- Biểu đồ cột thể hiện số lượng đơn hàng được đặt theo từng giờ trong ngày.
- Biểu đồ đường thể hiện tổng số đơn hàng theo từng ngày trong tuần.
- Biểu đồ cột thể hiện số lượng đơn hàng theo từng ngày trong tháng.
- Biểu đồ cột thể hiện 5 phương thức thanh toán trả góp phổ biến nhất.
- Biểu đồ hình tròn thể hiện tỷ lệ các phương thức thanh toán
- Biểu đồ tròn hiển thị tổng đơn hàng theo phương thức trả góp.
- Biểu đồ hình tròn thể hiện tỷ lệ đơn hàng bị hủy.
- Biểu đồ thể hiện số lượng bình luận (đánh giá) về các đơn hàng.

Mục đích:

Mục đích chung của dashboard này là phân tích hành vi khách hàng và tối ưu hóa chiến lược bán hàng. Nó giúp doanh nghiệp hiểu rõ thói quen mua sắm của khách hàng, bao gồm thời gian mua sắm, phương thức thanh toán, và xu hướng theo ngày trong tuần và tháng, từ đó cải thiện dịch vụ, giảm tỷ lệ hủy đơn hàng, và tối ưu hóa quy trình bán hàng và tiếp thị.

Dashboard phân tích chuyển đổi khách hàng tiềm năng (lead conversion analysis dashboard)



Hình 6.5: Phân tích chuyển đổi khách hàng tiềm năng

Nội dung:

- Biểu đồ cột ngang hiển thị số lượng khách hàng thực tế phân theo loại khách hàng tiềm năng
- Biểu đồ cột hiển thị khách hàng thực tế phân theo loại hình kinh doanh.
- Biểu đồ cột ngang phân chia khách hàng theo các phân khúc kinh doanh.
- Biểu đồ tròn hiển thị nguồn gốc của các khách hàng tiềm năng.
- Biểu đồ đường thể hiện số lượng chỉ theo thời gian.
- Biểu đồ tròn phân chia dựa trên thông tin công ty có công hoặc không.

Mục đích:

Bảng điều khiển này giúp doanh nghiệp phân tích hiệu quả chiến lược chuyển đổi tiềm năng khách hàng, bao gồm việc đánh giá nguồn gốc và chất lượng của khách hàng tiềm năng, xác định phân khúc và loại khách hàng có giá trị cao, theo dõi chuyển đổi hướng chuyển đổi theo thời gian để điều chỉnh chiến lược kinh doanh và tối ưu hóa kênh tiếp theo để tăng tỷ lệ chuyển đổi và thu nhập từ tiềm năng khách hàng.

Dashboard phân tích giao nhận hàng (delivery analytics dashboard)



Hình 6.6: Phân tích giao nhận hàng

Nội dung:

Biểu đồ tròn:

- Hiển thị số lượng đơn hàng theo trạng thái giao nhận (Order Status).
- Hiển thị thời gian giao hàng trung bình so với thời gian dự kiến (Avg Actual Delivery Time vs Estimated Time).
- Phân tích tỷ lệ trễ giao hàng (Order Delivery Delay).

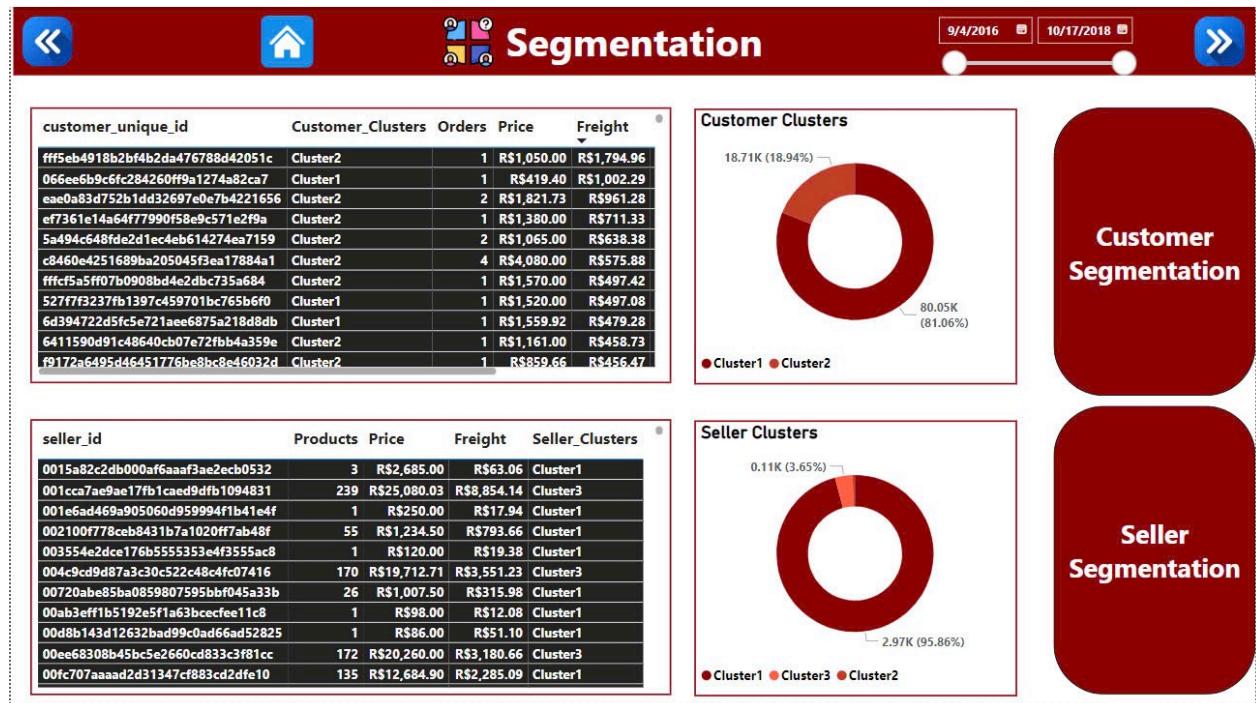
Biểu đồ đường:

- Biểu đồ xu hướng trễ giao hàng theo thời gian (Delivery Delay Trend).
- So sánh thời gian giao hàng ước tính và thực tế (Estimated vs Actual Time).
- Xu hướng trễ giao hàng của nhà vận chuyển (Carrier Shipping Delay Trend).

Mục đích:

Mục đích của dashboard là hỗ trợ doanh nghiệp giám sát hiệu quả quy trình logistics, xác định các vấn đề gây trễ giao hàng, và tối ưu hóa trải nghiệm khách hàng bằng cách cải thiện thời gian giao nhận và giảm thiểu sự sai lệch so với kế hoạch.

Dashboard phân loại khách hàng và nhà phân phối (Customer and Distributor Classification Dashboard)



Hình 6.7: Phân loại khách hàng và nhà phân phối

Nội dung:

Bảng dữ liệu:

- Customer Table: Hiển thị danh sách khách hàng, phân cụm (Cluster), số lượng đơn hàng (Orders), giá trị đơn hàng (Price), và phí vận chuyển (Freight).
- Seller Table: Hiển thị danh sách nhà phân phối, sản phẩm (Products), giá trị sản phẩm (Price), phí vận chuyển (Freight), và phân cụm nhà phân phối (Seller Clusters).

Biểu đồ tròn (Donut Chart):

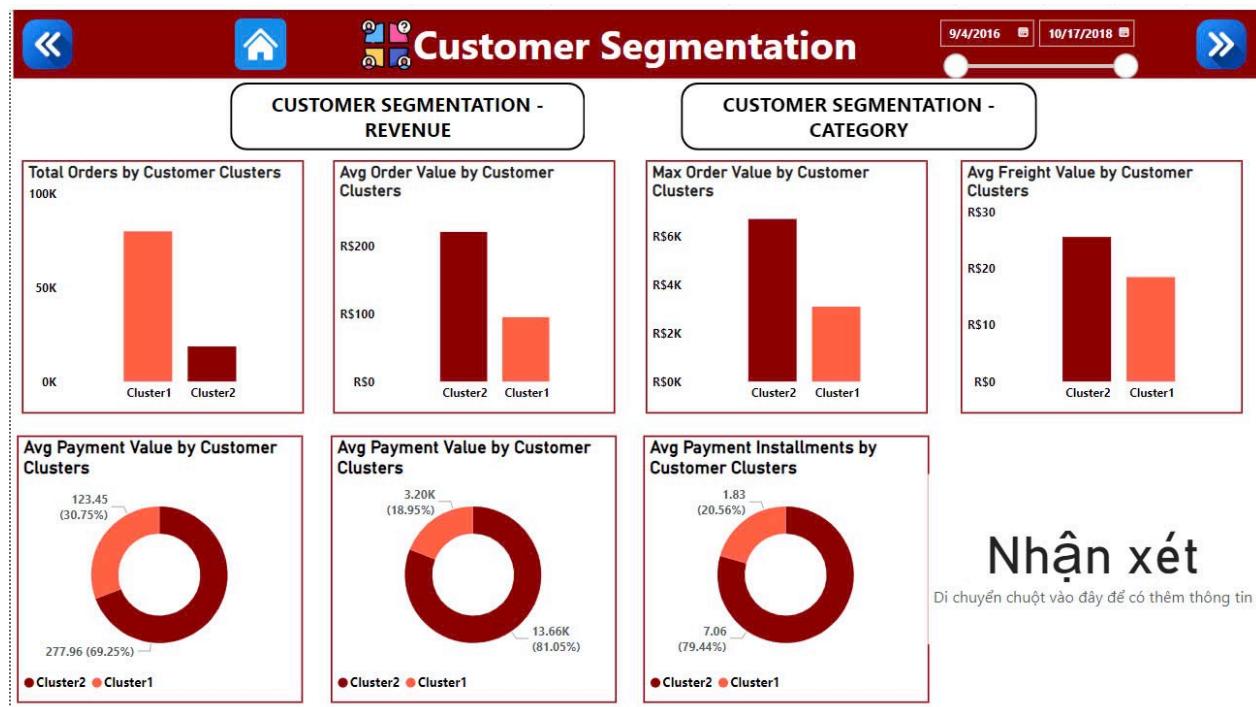
- Customer Clusters: Minh họa tỷ lệ phần trăm khách hàng thuộc các cụm (Clusters).
- Seller Clusters: Minh họa tỷ lệ phần trăm nhà phân phối thuộc các cụm (Clusters).

Thanh điều chỉnh thời gian (Time Filter): Cho phép lọc dữ liệu theo khoảng thời gian cụ thể để phân tích chi tiết hơn.

Mục đích:

Dashboard phân loại khách hàng và nhà phân phối giúp doanh nghiệp nhóm các khách hàng và nhà phân phối thành các cụm dựa trên hành vi mua hàng, giá trị đơn hàng, và các yếu tố liên quan. Điều này hỗ trợ việc xác định nhóm khách hàng hoặc nhà phân phối tiềm năng, thiết kế các chiến lược marketing phù hợp, tối ưu hóa mối quan hệ khách hàng và cải thiện hiệu quả phân phối. Hơn nữa, dashboard còn cung cấp khả năng lọc dữ liệu theo thời gian để phân tích xu hướng và thay đổi hành vi theo từng giai đoạn cụ thể.

Dashboard phân tích phân loại theo khách hàng (Dashboard classified by customer)



Hình 6.8: Phân loại theo khách hàng

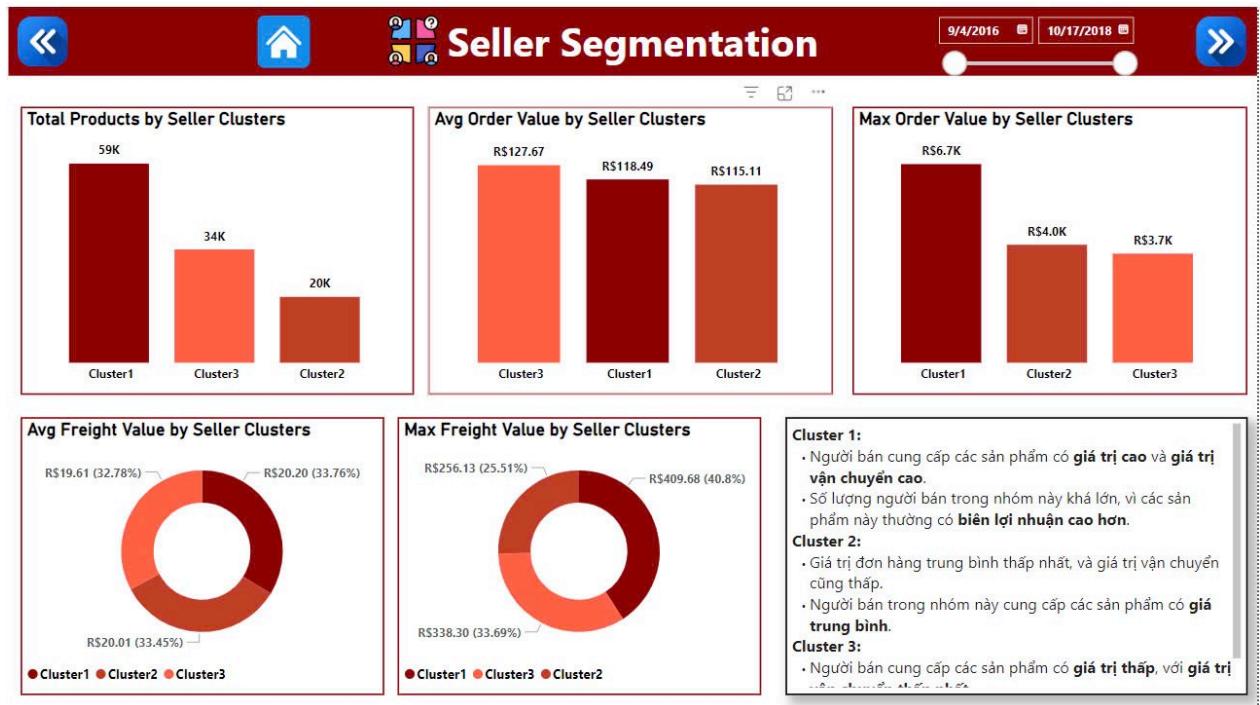
Nội dung:

- Biểu đồ cột: So sánh tổng số đơn hàng, giá trị đơn hàng trung bình, giá trị đơn hàng tối đa và giá trị vận chuyển trung bình giữa hai nhóm khách hàng (Cluster 1 và Cluster 2).
- Biểu đồ tròn: Thể hiện tỷ lệ giá trị thanh toán trung bình và số lần trả góp trung bình của mỗi nhóm khách hàng.

Mục đích:

Dashboard này giúp doanh nghiệp hiểu rõ khách hàng để cá nhân hóa trải nghiệm, tối ưu hóa chiến lược kinh doanh và nâng cao hiệu quả hoạt động.

Dashboard phân tích phân loại theo nhà phân phối (Dashboard classified by distributor)



Hình 6.9: Phân loại theo nhà phân phối

Nội dung:

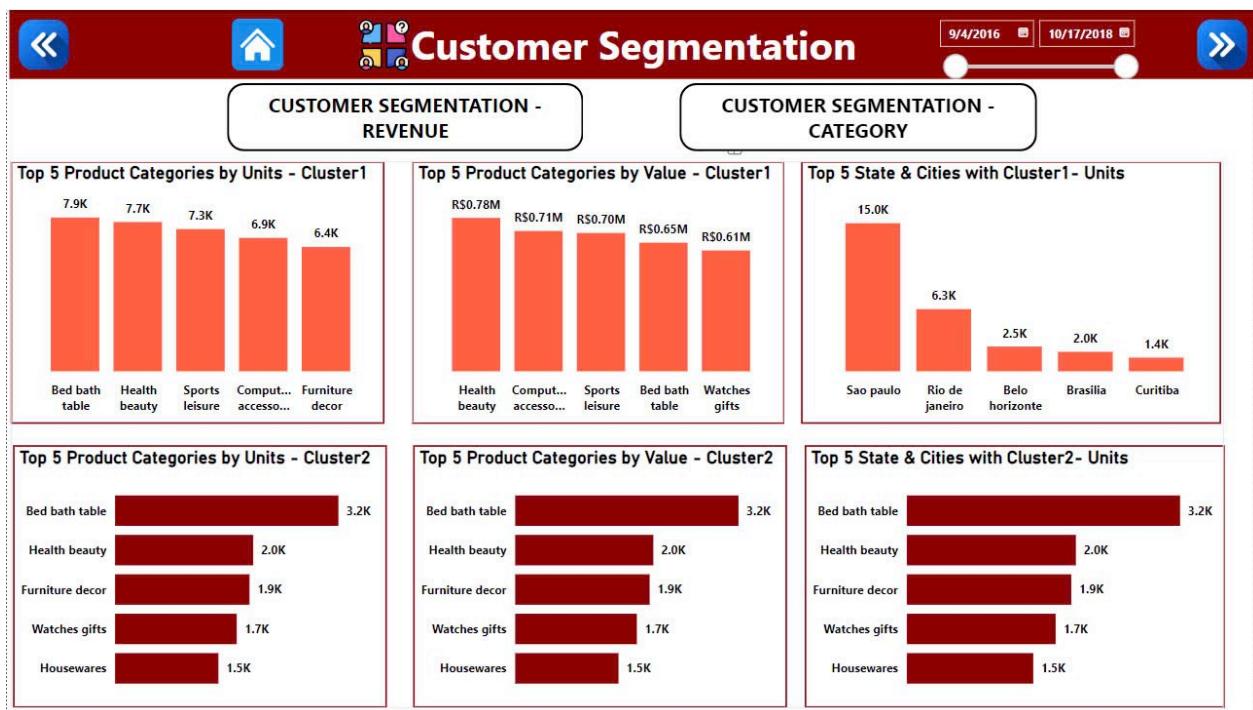
Biểu đồ cột: So sánh tổng số sản phẩm, giá trị đơn hàng trung bình, giá trị đơn hàng tối đa giữa ba nhóm nhà phân phối (Cluster 1, Cluster 2 và Cluster 3).

Biểu đồ tròn: Thể hiện tỷ lệ giá trị vận chuyển trung bình và giá trị vận chuyển tối đa của mỗi nhóm nhà phân phối.

Mục đích:

Dashboard này giúp doanh nghiệp hiểu rõ các nhóm nhà phân phối để phân bổ nguồn lực, đàm phán hợp đồng và quản lý rủi ro hiệu quả, từ đó tối ưu hóa hoạt động kinh doanh.

Dashboard phân tích loại khách hàng theo doanh thu và danh mục (Dashboard customers by revenue and category)



Hình 6.10: Phân loại khách hàng theo doanh thu và danh mục

Nội dung:

- Top 5 loại sản phẩm theo số lượng đơn vị (Units) của Cluster 1 và Cluster 2: Cho biết những sản phẩm nào được mỗi nhóm khách hàng mua nhiều nhất.
- Top 5 loại sản phẩm theo giá trị (Value) của Cluster 1 và Cluster 2: Cho biết những sản phẩm nào mang lại doanh thu cao nhất cho mỗi nhóm khách hàng.
- Top 5 bang và thành phố có số lượng đơn vị (Units) của Cluster 1 và Cluster 2: Xác định vị trí địa lý tập trung nhiều khách hàng của mỗi nhóm.

Mục đích:

Dashboard này cung cấp cho doanh nghiệp bức tranh chi tiết về hành vi và sở thích mua sắm của từng nhóm khách hàng. Thông qua việc hiển thị các sản phẩm phổ biến, giá trị mua hàng và vị trí địa lý, dashboard giúp doanh nghiệp hiểu rõ hơn về nhu cầu và mong muốn của từng phân khúc khách hàng. Từ đó, doanh nghiệp có thể điều chỉnh chiến lược kinh doanh, tập trung phát triển sản phẩm phù hợp và cá nhân hóa trải nghiệm mua sắm để gia tăng sự hài lòng và thúc đẩy doanh số.

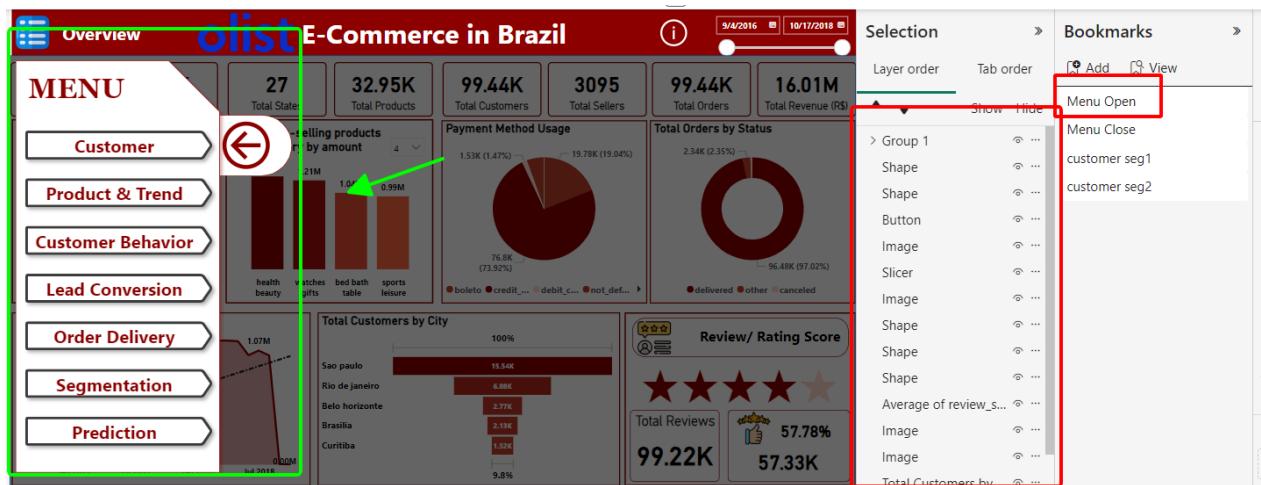
Giải thích:

Việc tạo các dashboard trong dự án của Olist đóng vai trò trò chơi cực kỳ quan trọng, mang lại cách tiếp cận trực quan và hệ thống để phân tích và hiểu được thùng dữ liệu phức tạp. Olist sử dụng các dashboard này để hỗ trợ quản lý và các bên liên quan nắm bắt thông tin nhanh chóng và chính xác nhằm đưa ra các kết quả chiến lược hiệu quả được quyết định. Các dashboard phân tích số liệu trợ giúp Olist theo dõi xu hướng bán hàng và hiệu suất kinh

doanh, nhận biết các mùa cao điểm, từ đó điều chỉnh chiến lược phù hợp để tối ưu hóa doanh thu. Dashboard phân tích sản phẩm hỗ trợ đánh giá tình trạng sản phẩm, giúp xác định các yếu tố ảnh hưởng đến giá trị và chất lượng, tạo cơ sở để cải thiện dịch vụ và nâng cao lòng tin của khách hàng. Dashboard phân tích người bán được phép Olist đo lường hiệu suất bán hàng của từng đối tác hoặc đội ngũ, từ đó xây dựng các chính sách phạt hợp lý, nâng cao năng lực làm việc và hiệu quả kinh doanh. Dashboard phân tích giá bán giúp xác định các yếu tố ảnh hưởng đến giá cả, hỗ trợ xây dựng chiến lược giá cạnh tranh và hợp lý hơn. Cuối cùng, Dashboard phân tích địa lý cung cấp thông tin chi tiết theo khu vực, tối ưu hóa công việc phân tích nguồn lực và chiến lược tiếp theo, giúp công cụ tiếp theo đạt được hiệu quả mục tiêu khách hàng tốt hơn. Tổng hợp lại, các dashboard không chỉ giúp Olist quản lý thùng rác một cách hiệu quả mà còn tăng cường khả năng quyết định dựa trên dữ liệu, góp phần cải thiện hiệu suất kinh doanh, nâng cao lợi thế cạnh tranh và hướng tới sự phát triển bền vững.

6.2.3 BOOKMARK

6.2.3.1 Tạo bookmark Menu Open



Hình 6.11: Bookmark Menu open

Giải thích: Dùng để mở menu chứa các button dẫn đến các trang dashboard khác

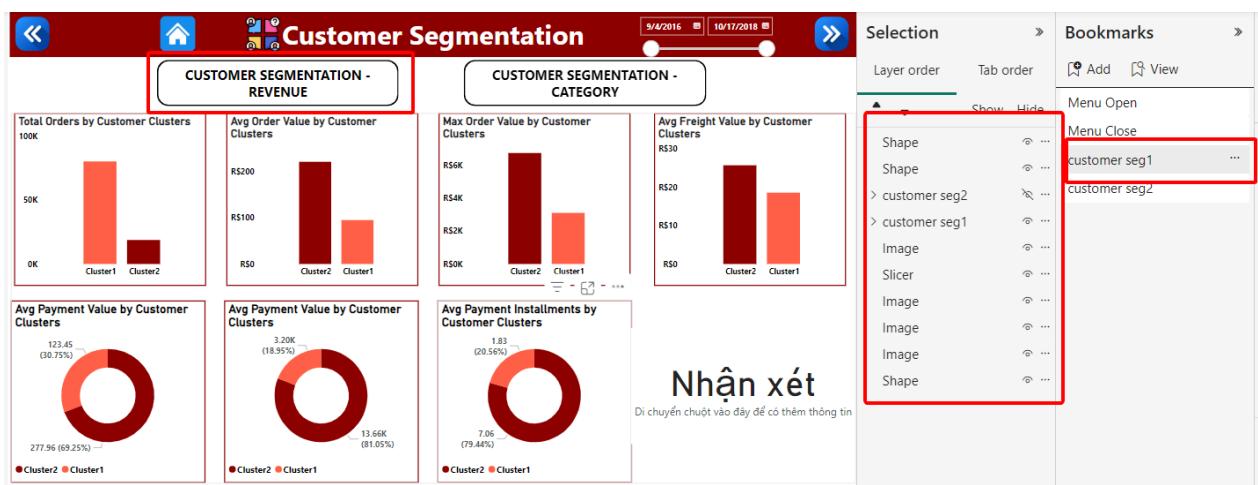
6.2.3.2 Tạo bookmark Menu Close



Hình 6.12: Bookmark Menu close

Giải thích: Dùng để đóng menu chứa các button dẫn đến các trang dashboard khác

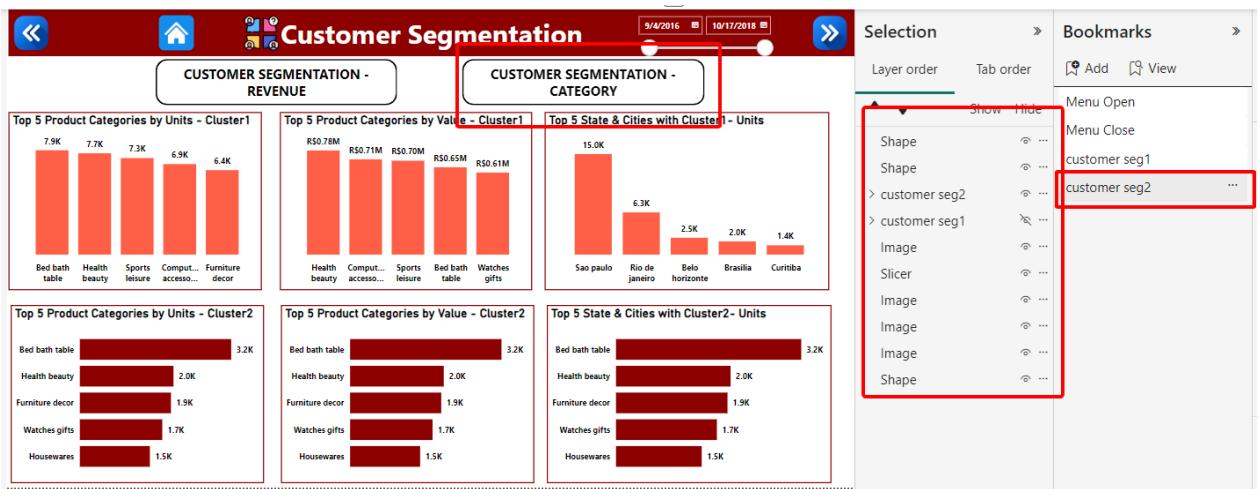
6.2.3.3 Tạo bookmark customer seg1



Hình 6.13: Bookmark customer seg1

Giải thích: Dùng để hiển thị phân tích phân loại khách hàng 1

6.2.3.4 Tạo bookmark customer seg2



Hình 6.14: Bookmark customer seg2

Giai thích: Dùng để hiển thị phân tích phân loại khách hàng 2

7 DỰ BÁO, DỰ ĐOÁN

```

1 # Xác định 10 đặc trưng hàng đầu có mối tương quan cao nhất với 'sự hài lòng'
2 # Chọn các cột dữ liệu số để tính toán ma trận tương quan
3 numeric_cols = merged_df.select_dtypes(include=[np.number])
4
5 # Tính toán ma trận tương quan chỉ cho các cột dữ liệu dạng số
6 corr_matrix = numeric_cols.corr()
7
8 # In ra 10 cột có tương quan cao nhất với satisfaction
9 print(corr_matrix['satisfaction'].sort_values(ascending=False)[1:-1])
10 ...
11 ...
12 MỤC ĐÍCH:
13 Xác định các đặc trưng quan trọng: Giúp chọn ra những yếu tố ảnh hưởng mạnh nhất đến mức độ hài lòng của khách hàng, phục vụ cho việc xây
14 Loại bỏ đặc trưng không liên quan: Loại bỏ các cột có tương quan thấp để giảm độ phức tạp của mô hình.
15 Hiểu rõ dữ liệu: Giúp phân tích và rút ra thông tin từ dữ liệu thực tế, hỗ trợ cải thiện dịch vụ khách hàng.
16 ...

```

Hình 7.1: Xác định 10 đặc trưng có mối tương quan cao với sự hài lòng

estimated_vs_actual_shipping	0.200293
order_month	0.027864
order_hour	0.009535
price	0.008271
payment_sequential	0.007372
order_value	0.005543
product_width_cm	-0.012533
order_day	-0.015242
product_length_cm	-0.017906
order_processing_time	-0.018413
product_height_cm	-0.020917
customer_zip_code_prefix	-0.022217
product_volume_m3	-0.022307
product_weight_g	-0.023465
freight_value	-0.028962
payment_installments	-0.042748
payment_value	-0.068124
order_item_id	-0.121016
time_to_delivery	-0.267583
Name: satisfaction, dtype: float64	

Hình 7.2: Kết quả

```

1 # Đặt ngưỡng tương quan
2 threshold = 0.05
3
4 # Lọc các đặc trưng có tương quan cao hoặc thấp đáng kể
5 # Lấy các đặc trưng có tương quan lớn hơn 7% hoặc nhỏ hơn -7% với 'satisfaction'
6 high_corr_features = corr_matrix.index[(corr_matrix['satisfaction'].abs() > threshold) & (corr_matrix.index != 'satisfaction')].tolist()
7
8 # In danh sách các đặc trưng có tương quan cao
9 print(high_corr_features)

[ 'order_item_id', 'payment_value', 'time_to_delivery', 'estimated_vs_actual_shipping', 'late_delivery']

1 # Kiểm tra kiểu dữ liệu cho các đặc trưng có tương quan cao
2 merged_df[high_corr_features].dtypes

[   ]
   order_item_id      float64
   payment_value      float64
   time_to_delivery    int64
   estimated_vs_actual_shipping    int64
   late_delivery      int64

dtype: object

```

Hình 7.3: Kiểu dữ liệu các đặc trưng có tương quan cao

```

[ ] 1 # Chọn các đặc trưng quan trọng nhất để sử dụng
2 top_4_features = ['payment_value', 'time_to_delivery', 'estimated_vs_actual_shipping', 'late_delivery']

[   ]
1 from sklearn.model_selection import train_test_split
2 from sklearn.preprocessing import StandardScaler
3 from sklearn.compose import ColumnTransformer
4 from sklearn.pipeline import Pipeline
5 from sklearn.linear_model import LogisticRegression
6 from sklearn.tree import DecisionTreeClassifier
7 from sklearn.ensemble import RandomForestClassifier
8 import xgboost as xgb
9 from sklearn.metrics import classification_report, confusion_matrix
10 import matplotlib.pyplot as plt
11 import seaborn as sns
12

```

Hình 7.4: Chọn 4 đặc trưng và import thư viện

```

12 # Chuẩn bị dữ liệu
13 top_6_features = ['estimated_vs_actual_shipping', 'order_month', 'order_hour', 'price', 'payment_sequential', 'order_value', 'payment_instalment']
14 x = merged_df[top_4_features]
15 y = merged_df['satisfaction']
16
17 #chia dữ liệu thành tập huấn luyện và kiểm tra
18 X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
19
20 #Tiền xử lý dữ liệu bằng Pipeline và ColumnTransformer (data processing)
21 ...
22 StandardScaler: Chuẩn hóa dữ liệu số để đưa các giá trị về cùng thang đo (mean=0, std=1).
23 ColumnTransformer: Chỉ áp dụng việc chuẩn hóa cho các cột số được chọn (top_4_features).
24 fit_transform và transform: Huấn luyện trên tập huấn luyện và áp dụng cho cả tập huấn luyện lẫn kiểm tra.
25 ...
26
27 #Pipeline for numerical features
28 numerical_transformer = Pipeline(steps=[
29     ('scaler', StandardScaler())
30 ])
31
32 # Applying ColumnTransformer to preprocess the data
33 preprocessor = ColumnTransformer(
34     transformers=[('num', numerical_transformer, top_4_features)])
35
36
37 # Preprocessing the data
38 X_train_preprocessed = preprocessor.fit_transform(X_train)
39 X_test_preprocessed = preprocessor.transform(X_test)
40

```

Hình 7.5: Chuẩn bị dữ liệu

```

41 # Khởi tạo các mô hình
42 ...
43 Logistic Regression: Mô hình tuyến tính dự đoán xác suất dựa trên hồi quy logistic.
44 Decision Tree: Cây quyết định với độ sâu tối đa là 10 và yêu cầu tối thiểu 50 mẫu để chia nhánh.
45 Random Forest: Tập hợp 100 cây quyết định với các giới hạn về độ sâu, số mẫu chia nhánh và số mẫu lá.
46 XGBoost: Mô hình tăng cường độ chính xác dựa trên thuật toán gradient.
47 ...
48 models = {
49     'Logistic Regression': LogisticRegression(random_state=42),
50     'Decision Tree': DecisionTreeClassifier(random_state=42, max_depth=10, min_samples_split=50),
51     'Random Forest': RandomForestClassifier(random_state=42, n_estimators=100, max_depth=10, min_samples_split=10, min_samples_leaf=4),
52     'XGBoost': xgb.XGBClassifier(random_state=42)
53 }
54

```

Hình 7.6: Khởi tạo các mô hình

```

79
80 # Đánh giá từng mô hình
81 for model_name, model in models.items():
82     print(f"Evaluating {model_name}")
83     evaluate_model(model, X_train_preprocessed, y_train, X_test_preprocessed, y_test, model_name)

```

Hình 7.7: Đánh giá từng mô hình

```

1 from sklearn.model_selection import GridSearchCV
2
3 # Khởi tạo mô hình XGBoost
4 xgb_model = xgb.XGBClassifier(random_state=42) #Mô hình XGBoost dùng cho bài toán phân loại.
5
6 # Xác định lưới tham số (Parameter Grid) để tìm kiếm
7 param_grid = {
8     'n_estimators': [50, 100, 150], # Ít cây hơn để giữ cho mô hình đơn giản hơn
9     'max_depth': [3, 4, 5], # giá trị nhỏ để tránh overfitting
10    'learning_rate': [0.1, 0.01, 0.05],
11    'subsample': [0.8, 1.0],
12    'colsample_bytree': [0.8, 1.0]
13 }
14
15 # Khởi tạo GridSearchCV
16 grid_search = GridSearchCV(
17     estimator=xgb_model,
18     param_grid=param_grid,
19     scoring='accuracy', # or another scoring metric
20     cv=3,
21     verbose=1,
22     n_jobs=-1
23 )

```

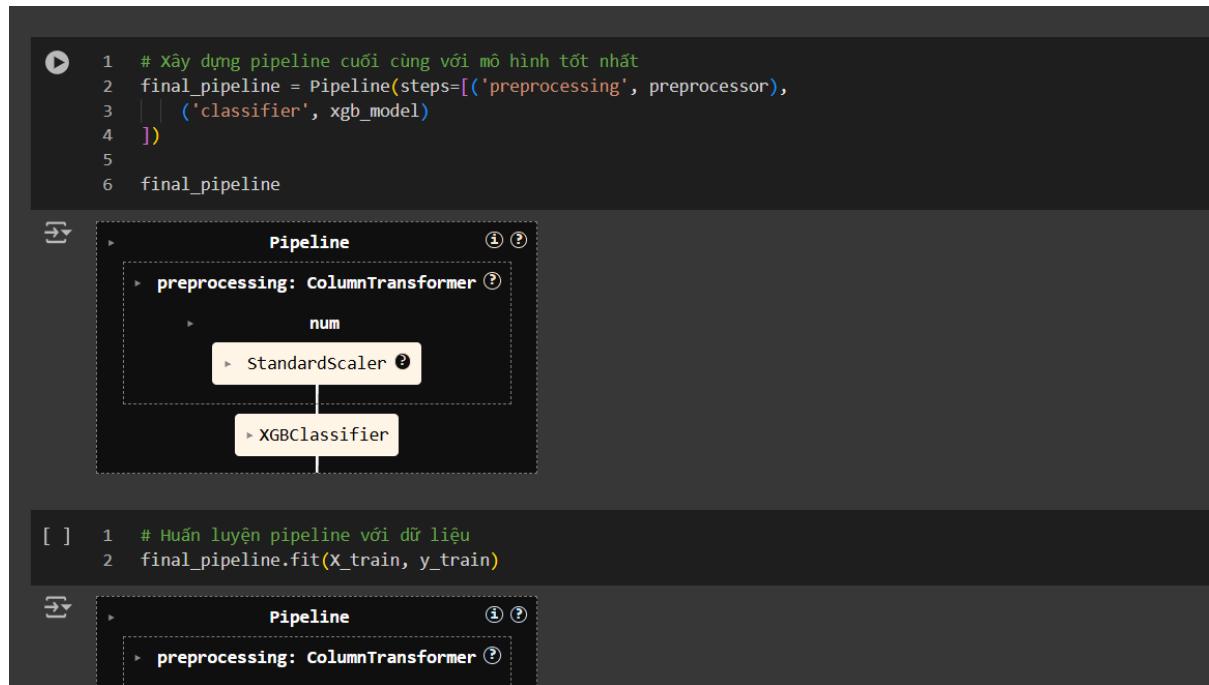
Hình 7.8: Chọn mô hình phù hợp nhất, khởi tạo mô hình

```

46
49 # Huấn luyện và tìm tham số tốt nhất
50 # GridSearchCV sẽ thử nghiệm tất cả các kết hợp của lưới tham số (108 trường hợp) và chọn tham số có độ chính xác cao nhất.
51 grid_search.fit(X_train_preprocessed, y_train)
52
53 # Lấy mô hình tốt nhất
54 best_model = grid_search.best_estimator_
55
56 # Dự đoán trên tập huấn luyện và kiểm tra
57 train_preds = best_model.predict(X_train_preprocessed)
58 test_preds = best_model.predict(X_test_preprocessed)

```

Hình 7.9: Huấn luyện và tìm tham số tốt nhất



The screenshot shows a Jupyter Notebook cell with Python code to build a pipeline:

```

1 # Xây dựng pipeline cuối cùng với mô hình tốt nhất
2 final_pipeline = Pipeline(steps=[('preprocessing', preprocessing),
3                                 ('classifier', xgb_model)])
4
5 final_pipeline

```

Below the code is a visual representation of the pipeline structure:

```

graph TD
    subgraph Pipeline [Pipeline]
        preprocessing[preprocessing: ColumnTransformer]
        preprocessing --> num[num]
        num --> StandardScaler[StandardScaler]
        StandardScaler --> XGBClassifier[XGBClassifier]
    end

```

Another cell shows the pipeline being fitted:

```

[ ] 1 # Huấn luyện pipeline với dữ liệu
2 final_pipeline.fit(x_train, y_train)

```

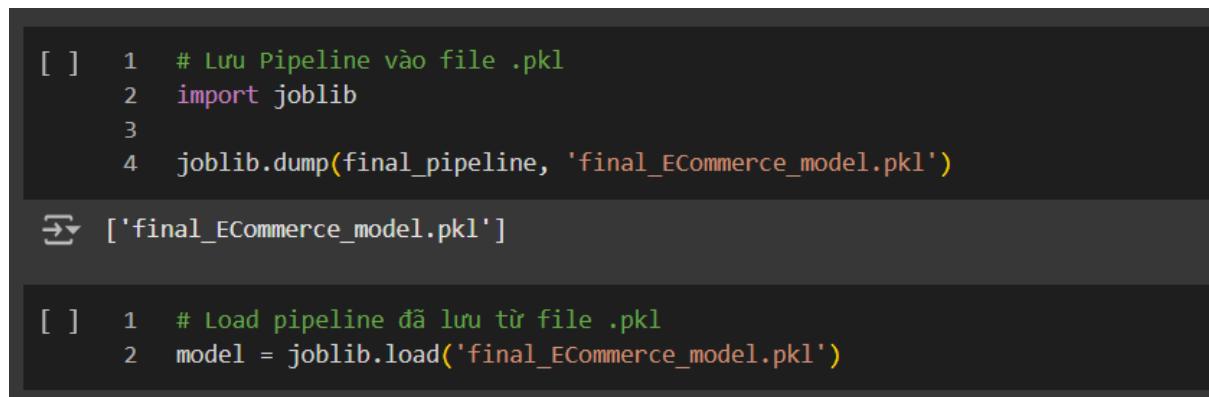
And finally, another cell shows the pipeline object:

```

[ ] Pipeline
      preprocessing: ColumnTransformer

```

Hình 7.10: Xây dựng pipeline với mô hình tốt nhất và huấn luyện với tập dữ liệu



The screenshot shows a Jupyter Notebook cell with Python code to save the pipeline:

```

[ ] 1 # Lưu Pipeline vào file .pkl
2 import joblib
3
4 joblib.dump(final_pipeline, 'final_ECommerce_model.pkl')

```

Below the code is the output showing the saved file:

```

[ ] ['final_ECommerce_model.pkl']

```

Another cell shows the pipeline being loaded:

```

[ ] 1 # Load pipeline đã lưu từ file .pkl
2 model = joblib.load('final_ECommerce_model.pkl')

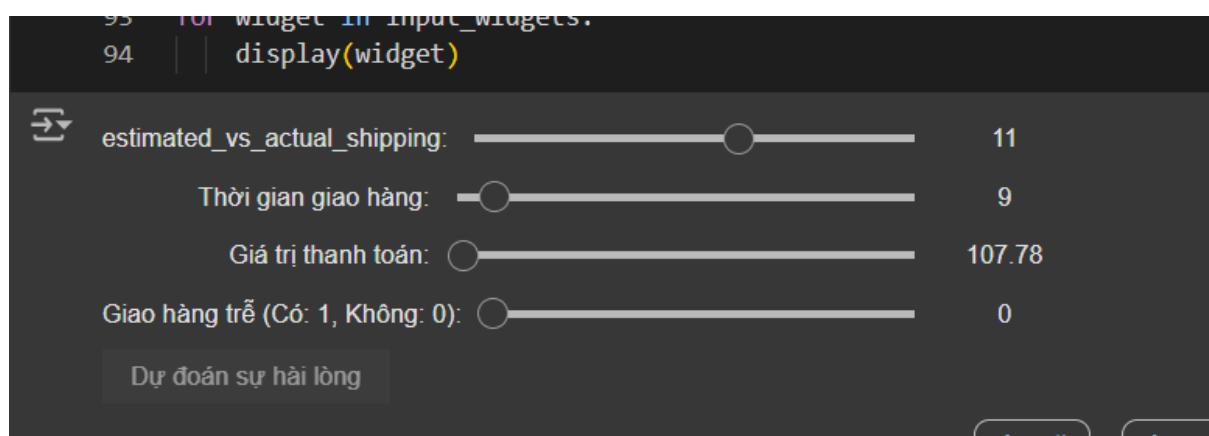
```

Hình 7.11: Lưu và load pipeline

```
Xây dựng ứng dụng dự đoán sự hài lòng của khách hàng bằng Streamlit

1 %%writefile ECB.py
2 # Ghi nội dung bên dưới vào tệp ECB.py
3
4 import streamlit as st
5 import joblib
6 import numpy as np
7
8 # Tải pipeline đã lưu
9 model = joblib.load('final_ECommerce_model.pkl')
10
11 # Định nghĩa giao diện ứng dụng
12 def main():
13     st.title('Customer Satisfaction Prediction App')
14
15     # Nhập dữ liệu đầu vào
16     # Xác định các đầu vào với phạm vi phù hợp và các giá trị mặc định dựa trên dữ liệu của bạn
17     estimated_vs_actual_shipping = st.number_input('Estimated vs Actual Shipping Days', min_value=-189, max_value=146, value=11)
18     time_to_delivery = st.number_input('Time to Delivery', min_value=-7, max_value=208, value=9)
19     payment_value = st.number_input('Payment Value', min_value=0.0, max_value=13664.08, value=107.78)
20     late_delivery = st.number_input('Late Delivery', min_value=0, max_value=1, value=0)
21
22     # Dự đoán khi nhấn nút
23     if st.button('Predict Satisfaction'): # st.button: Hiển thị nút bấm để kích hoạt dự đoán.
24         # Tạo một mảng với dữ liệu đầu vào
25         # Đảm bảo tất cả các đầu vào được đưa vào mảng theo đúng thứ tự
26         input_data = np.array([[estimated_vs_actual_shipping, time_to_delivery, payment_value, late_delivery]])
27
28         # Get the prediction
29         prediction = model.predict(input_data) # Thực hiện dự đoán dựa trên dữ liệu đầu vào.
30
31         # Trả kết quả
32         if prediction[0] == 1:
33             st.success('The customer is satisfied.')
34         else:
35             st.error('The customer is not satisfied')
36
37 if __name__ == '__main__':
38     main()
```

Hình 7.12: Xây dựng ứng dụng dự đoán sự hài lòng của khách hàng bằng Streamlit



Hình 7.13: Xây dựng mô hình dự đoán sự hài lòng trên jupyter notebook

8 KẾT LUẬN

8.1 BÁO CÁO

8.1.1 CÁC BƯỚC VIẾT BÁO CÁO

Viết báo cáo phân tích dữ liệu đòi hỏi một quy trình có cấu trúc để đảm bảo rằng báo cáo được trình bày rõ ràng, mạch lạc và có giá trị đối với người đọc. Việc viết báo cáo phân tích dữ liệu đòi hỏi một quy trình có cấu trúc và cẩn thận. Bằng cách tuân theo các bước, bạn có thể tạo ra một báo cáo phân tích dữ liệu chất lượng, giúp truyền đạt các phát hiện quan trọng một cách rõ ràng và hiệu quả, từ đó hỗ trợ việc ra quyết định và cải thiện hiệu suất kinh doanh. Dưới đây là các bước cụ thể để viết một báo cáo phân tích dữ liệu:

Bước 1: Xác định mục tiêu báo cáo

Xác định câu hỏi nghiên cứu: Hiểu rõ mục tiêu của báo cáo và các câu hỏi chính cần trả lời.

Đối tượng đọc báo cáo: Xác định ai sẽ đọc báo cáo và điều chỉnh nội dung cho phù hợp với nhu cầu và mức độ hiểu biết của họ.

Bước 2: Thu thập dữ liệu

Chọn nguồn dữ liệu: Xác định các nguồn dữ liệu cần thiết để trả lời các câu hỏi nghiên cứu.

Thu thập dữ liệu: Thu thập dữ liệu từ các nguồn đã xác định.

Bước 3: Chuẩn bị dữ liệu

Làm sạch dữ liệu: Kiểm tra và xử lý các lỗi, giá trị thiếu, và giá trị ngoại lai trong dữ liệu.

Chuẩn hóa dữ liệu: Đảm bảo rằng dữ liệu được định dạng đúng cách và sẵn sàng cho phân tích.

Bước 4: Phân tích dữ liệu

Lựa chọn phương pháp phân tích: Chọn các phương pháp phân tích phù hợp với mục tiêu báo cáo (ví dụ: phân tích mô tả, phân tích hồi quy, phân tích xu hướng).

Thực hiện phân tích: Sử dụng các công cụ và kỹ thuật phù hợp để phân tích dữ liệu.

Bước 5: Trực quan hóa dữ liệu

Chọn loại biểu đồ phù hợp: Chọn các loại biểu đồ và đồ thị phù hợp để trực quan hóa dữ liệu (ví dụ: biểu đồ đường, biểu đồ cột, biểu đồ tròn).

Tạo biểu đồ: Sử dụng các công cụ trực quan hóa dữ liệu để tạo biểu đồ minh họa cho các phát hiện quan trọng.

Bước 6: Viết báo cáo

Giới thiệu:

- Trình bày mục tiêu của báo cáo.
- Giới thiệu ngắn gọn về dữ liệu và phương pháp phân tích.

Phương pháp: Mô tả chi tiết các phương pháp và công cụ sử dụng trong quá trình phân tích.

Kết quả:

- Trình bày các phát hiện chính từ phân tích dữ liệu.
- Sử dụng biểu đồ và đồ thị để minh họa cho các phát hiện.

Thảo luận:

- Giải thích ý nghĩa của các phát hiện.
- Đề xuất các hành động hoặc quyết định dựa trên kết quả phân tích.

Kết luận:

- Tóm tắt lại các điểm chính của báo cáo.

- Nêu rõ các kết luận chính và các bước tiếp theo nếu có.

Bước 7: Xem xét và chỉnh sửa

Đọc lại báo cáo: Kiểm tra lại báo cáo để đảm bảo rằng nội dung rõ ràng, chính xác và không có lỗi.

Chỉnh sửa: Sửa các lỗi ngữ pháp, chính tả và đảm bảo rằng báo cáo có cấu trúc logic.

Bước 8: Trình bày báo cáo

Định dạng: Đảm bảo rằng báo cáo được định dạng chuyên nghiệp và dễ đọc.

Trình bày: Chuẩn bị sẵn sàng để trình bày báo cáo trước các bên liên quan nếu cần thiết.

8.1.2 TỔNG HỢP

Sau khi phân tích bộ dữ liệu của Olist, chúng tôi đã kiểm chứng các giả thuyết và thu được những kết quả quan trọng sau:

- **Thời gian mua hàng ảnh hưởng đến giá trị đơn hàng:** Các đơn hàng được thực hiện gần đây có xu hướng có giá trị cao hơn, cho thấy xu hướng tăng giá trị đơn hàng theo thời gian.
- **Danh mục sản phẩm ảnh hưởng đến giá trị đơn hàng:** Một số danh mục sản phẩm có giá trị trung bình cao hơn so với các danh mục khác, ví dụ như các sản phẩm điện tử, đồ gia dụng có giá trị cao hơn các sản phẩm thời trang, sách vở.
- **Vị trí địa lý ảnh hưởng đến giá trị đơn hàng:** Các đơn hàng từ các thành phố lớn hoặc khu vực phát triển có xu hướng có giá trị cao hơn so với các khu vực khác.
- **Phương thức thanh toán ảnh hưởng đến giá trị đơn hàng:** Các đơn hàng sử dụng thẻ tín dụng có xu hướng có giá trị cao hơn so với các phương thức thanh toán khác như boleto.
- **Mức độ hài lòng của khách hàng ảnh hưởng đến giá trị đơn hàng:** Những khách hàng hài lòng với sản phẩm và dịch vụ (thể hiện qua điểm đánh giá cao) có xu hướng mua hàng với giá trị cao hơn trong các lần mua hàng tiếp theo.

Các phân tích cũng chỉ ra rằng:

- Thời gian xử lý đơn hàng (từ khi đặt hàng đến khi giao hàng) có ảnh hưởng đến mức độ hài lòng của khách hàng. Thời gian xử lý càng ngắn, khách hàng càng hài lòng.
- Các chương trình khuyến mãi và giảm giá có tác động tích cực đến quyết định mua hàng và giá trị đơn hàng.
- Olist có thể cải thiện hiệu quả kinh doanh bằng cách tập trung vào các kênh tiếp thị hiệu quả, tối ưu hóa quy trình xử lý đơn hàng, và nâng cao trải nghiệm khách hàng.

Những phát hiện này cung cấp cho Olist cái nhìn sâu sắc về các yếu tố ảnh hưởng đến hoạt động kinh doanh, từ đó hỗ trợ việc ra quyết định chiến lược, tối ưu hóa hoạt động và nâng cao hiệu quả kinh doanh.

8.2 KHÓ KHĂN

Trong quá trình phân tích hành vi và xu hướng người tiêu dùng trên trang thương mại điện tử Olist tại Brazil, chúng tôi đã gặp phải nhiều khó khăn đáng kể. Đầu tiên, việc thu thập và làm sạch dữ liệu gặp nhiều thách thức do dữ liệu bị thiếu, lỗi, hoặc không nhất quán, đòi hỏi nhiều thời gian và công sức để xử lý. Thứ hai, xác định và lựa chọn các phương pháp phân tích phù hợp để kiểm chứng các giả thuyết cũng không hề đơn giản, đòi hỏi sự tỉ mỉ và kiến thức chuyên sâu về phân tích dữ liệu. Thứ ba, việc trực quan hóa dữ liệu để trình bày kết quả một cách rõ ràng và hấp dẫn cũng gặp nhiều khó khăn, đặc biệt là khi phải lựa chọn và sử dụng các loại biểu đồ phù hợp. Ngoài ra, cơ sở vật chất kém và thiếu thốn về kinh phí đã làm hạn chế khả năng tiếp cận các công cụ phân tích tiên tiến, khiến quá trình làm việc trở nên khó khăn hơn. Hơn nữa, nhóm thực hiện dự án là một nhóm mới thành lập, do đó kỹ năng làm việc nhóm chưa được phát huy hiệu quả tối đa, dẫn đến sự thiếu phối hợp và một số lỗi giao tiếp nội bộ. Những khó khăn này đã thử thách đội ngũ thực hiện dự án, nhưng cũng giúp chúng tôi học hỏi và nâng cao kỹ năng trong quá trình làm việc.

8.3 THUẬN LỢI

Mặc dù gặp nhiều khó khăn nhưng việc phân tích hành vi và xu hướng người tiêu dùng trên trang thương mại điện tử Olist tại Brazil, cũng có nhiều thuận lợi đáng kể. Đầu tiên, đội ngũ thực hiện dự án, dù mới thành lập, bao gồm những thành viên có năng lực và nhiệt huyết, sẵn sàng học hỏi và đóng góp ý tưởng sáng tạo. Sự đa dạng trong kinh nghiệm và kiến thức của các thành viên đã giúp chúng tôi nhanh chóng thích nghi và giải quyết các vấn đề phát sinh.. Thứ hai , về mặt dữ liệu thì Olist đã cung cấp sẵn hai bộ dữ liệu khá đầy đủ về hoạt động kinh doanh, bao gồm thông tin về khách hàng, đơn hàng, sản phẩm, tiếp thị...Có nhiều công cụ phân tích dữ liệu mạnh mẽ và miễn phí có sẵn như Python với các thư viện Pandas và Numpy. Ngoài ra, sự phối hợp tốt giữa các thành viên trong nhóm, cùng với sự hướng dẫn tận tình từ giảng viên hướng dẫn, đã giúp dự án triển triển thuận lợi và đạt được các mục tiêu đề ra. Những thuận lợi này không chỉ giúp chúng tôi vượt qua các thách thức mà còn nâng cao chất lượng và hiệu quả của dự án, góp phần quan trọng vào việc cung cấp các thông tin giá trị cho việc phân tích hành vi và xu hướng người dùng .

8.4 HƯỚNG PHÁT TRIỂN

Dựa trên kết quả phân tích dữ liệu, dưới đây là một số hướng phát triển cho dự án phân tích dữ liệu của Olist, bên cạnh việc tập trung vào sự phát triển của Olist:

Mở rộng phạm vi phân tích:

- **Phân tích chi tiết hơn về hành vi khách hàng:** Phân khúc khách hàng, xây dựng mô hình dự đoán hành vi mua hàng.
- **Phân tích hiệu quả bán hàng của từng nhà bán hàng:** Xếp hạng nhà bán hàng, phân tích sản phẩm bán chạy của từng nhà bán hàng, đưa ra đề xuất cải thiện hiệu quả kinh doanh cho từng nhà bán hàng.
- **Phân tích chiến dịch marketing:** Đánh giá hiệu quả chiến dịch quảng cáo, tối ưu hóa chi phí marketing, xác định kênh quảng cáo hiệu quả.

- **Phân tích dữ liệu không gian địa lý:** Phân tích xu hướng mua sắm theo khu vực, tối ưu hóa vị trí kho bãi và dịch vụ giao hàng.

Nâng cao chất lượng dữ liệu:

- **Thu thập thêm dữ liệu:** Bổ sung thông tin về nhân khẩu học khách hàng, sở thích, hành vi trên mạng xã hội.
- **Cải thiện chất lượng dữ liệu:** Xử lý dữ liệu bị thiếu, xác định và loại bỏ dữ liệu ngoại lai, đảm bảo tính nhất quán của dữ liệu.
- **Tự động hóa quy trình làm sạch và chuẩn hóa dữ liệu.**

Phát triển mô hình dự đoán:

- **Dự đoán doanh thu:** Xây dựng mô hình dự đoán doanh thu trong tương lai, hỗ trợ lập kế hoạch kinh doanh.
- **Dự đoán nhu cầu sản phẩm:** Hỗ trợ quản lý kho bãi và chuỗi cung ứng.

Cải thiện trực quan hóa dữ liệu:

- **Sử dụng nhiều kỹ thuật trực quan hóa khác nhau:** Tạo ra dashboard tương tác, sử dụng bản đồ và biểu đồ đa dạng.
- **Thiết kế giao diện báo cáo thân thiện với người dùng:** Dễ dàng tìm kiếm, lọc và phân tích thông tin.
- **Phát triển ứng dụng trực quan hóa dữ liệu trên thiết bị di động.**

Nâng cao trải nghiệm khách hàng:

- **Tối ưu hóa thời gian giao hàng:** Rút ngắn thời gian xử lý đơn hàng, hợp tác với các đối tác vận chuyển uy tín, cung cấp nhiều lựa chọn vận chuyển cho khách hàng.
- **Cải thiện dịch vụ khách hàng:** Đầu tư vào hệ thống hỗ trợ khách hàng, cung cấp thông tin đầy đủ và chính xác về sản phẩm và dịch vụ, xử lý khiếu nại nhanh chóng và hiệu quả.
- **Cá nhân hóa trải nghiệm mua sắm:** Phân tích hành vi và sở thích của khách hàng để đưa ra các đề xuất sản phẩm phù hợp, cá nhân hóa giao diện website và ứng dụng.

Tối ưu hóa hoạt động tiếp thị:

- **Tập trung vào các kênh tiếp thị hiệu quả:** Phân tích hiệu quả của từng kênh tiếp thị, đầu tư vào các kênh mang lại tỷ lệ chuyển đổi cao.
- **Phát triển các chiến dịch tiếp thị sáng tạo:** Thu hút khách hàng mới, kích thích nhu cầu mua sắm, tăng nhận diện thương hiệu.
- **Xây dựng chương trình khách hàng thân thiết:** Khuyến khích khách hàng mua hàng lặp lại, tăng giá trị vòng đời khách hàng.

Mở rộng thị trường và danh mục sản phẩm:

- **Mở rộng sang các khu vực địa lý mới:** Phân tích thị trường tiềm năng, phát triển mạng lưới vận chuyển và kho bãi.

- **Mở rộng danh mục sản phẩm:** Cung cấp đa dạng sản phẩm, đáp ứng nhu cầu của nhiều phân khúc khách hàng.
- **Hợp tác với các nhà bán hàng mới:** Mở rộng mạng lưới nhà bán hàng, tăng sự lựa chọn cho khách hàng.

Nâng cao hiệu quả hoạt động:

- **Tối ưu hóa quy trình vận hành:** Sử dụng công nghệ để tự động hóa các quy trình, giảm thiểu lỗi và nâng cao năng suất.
- **Quản lý kho bãi hiệu quả:** Dự báo nhu cầu, tối ưu hóa lượng hàng tồn kho, giảm thiểu chi phí lưu kho.
- **Phân tích dữ liệu thường xuyên:** Theo dõi các chỉ số hiệu quả kinh doanh, nhận diện các vấn đề và cơ hội.

Bằng cách tập trung vào các hướng phát triển này, Olist có thể tiếp tục củng cố vị thế dẫn đầu trên thị trường thương mại điện tử, mang lại lợi ích cho cả doanh nghiệp và khách hàng.

9 TỔNG KẾT

Dự án: Phân tích xu hướng và hành vi người tiêu dùng trên trang thương mại điện tử Olist tại Brazil

Olist, nền tảng thương mại điện tử hàng đầu tại Brazil, đã thực hiện một dự án phân tích dữ liệu sâu rộng nhằm khai thác tối đa tiềm năng của bộ dữ liệu Brazilian E-Commerce Public Dataset và Marketing Funnel. Mục tiêu chính của dự án là:

- **Hiểu rõ hành vi khách hàng:** Phân tích chi tiết hành vi mua sắm, sở thích và thói quen của khách hàng để xây dựng các chiến lược tiếp thị hiệu quả.
- **Tối ưu hóa quy trình bán hàng:** Nhận diện các điểm hạn chế trong quy trình bán hàng và đề xuất các giải pháp cải tiến để tăng tỷ lệ chuyển đổi.
- **Cải thiện hiệu quả tiếp thị:** Đánh giá hiệu quả của các chiến dịch marketing và tối ưu hóa ngân sách quảng cáo.

Phương pháp thực hiện

Dự án đã áp dụng một quy trình phân tích dữ liệu toàn diện, bao gồm:

1. **Thu thập và làm sạch dữ liệu:** Sưu tầm dữ liệu từ nhiều nguồn khác nhau và tiến hành làm sạch, chuẩn hóa để đảm bảo tính chính xác và nhất quán.
2. **Mô hình hóa dữ liệu:** Xây dựng các mô hình thống kê và máy học để phát hiện các mối quan hệ phức tạp giữa các yếu tố ảnh hưởng đến doanh số.
3. **Trực quan hóa dữ liệu:** Sử dụng Power BI để tạo ra các báo cáo trực quan, dễ hiểu, giúp người dùng nắm bắt nhanh chóng các thông tin quan trọng.

Kết quả đạt được

Qua quá trình phân tích, dự án đã đạt được những kết quả đáng kể:

- **Xác định các yếu tố ảnh hưởng đến doanh số:** Thời gian mua hàng, danh mục sản phẩm, vị trí địa lý, phương thức thanh toán và mức độ hài lòng của khách hàng đều có tác động đáng kể đến giá trị đơn hàng.
- **Đánh giá hiệu quả các chương trình khuyến mãi:** Các chương trình khuyến mãi đã được chứng minh là có hiệu quả trong việc thúc đẩy doanh số và cải thiện trải nghiệm khách hàng.
- **Nhận diện các điểm hạn chế trong quy trình bán hàng:** Sự chậm trễ trong việc xử lý đơn hàng quá lâu hay dịch vụ giao hàng kéo dài là một trong những nguyên nhân khiến khách hàng không hài lòng và rời bỏ.

Các gợi ý phát triển

Dựa trên kết quả phân tích, Olist có thể tập trung vào các hướng phát triển sau:

- **Mở rộng phạm vi phân tích:** Áp dụng các kỹ thuật phân tích tiên tiến hơn để khám phá sâu hơn vào dữ liệu.

- **Nâng cao chất lượng dữ liệu:** Đầu tư vào việc thu thập và quản lý dữ liệu một cách hiệu quả.
- **Phát triển mô hình dự đoán:** Xây dựng các mô hình dự đoán để dự báo doanh số, hành vi khách hàng và xu hướng thị trường.
- **Cải thiện trực quan hóa dữ liệu:** Tạo ra các dashboard tương tác và trực quan hơn để hỗ trợ ra quyết định.
- **Nâng cao trải nghiệm khách hàng:** Tối ưu hóa quy trình bán hàng, dịch vụ khách hàng và các chương trình khuyến mãi.
- **Tối ưu hóa hoạt động tiếp thị:** Tập trung vào các kênh tiếp thị hiệu quả và cá nhân hóa trải nghiệm khách hàng.
- **Mở rộng thị trường và danh mục sản phẩm:** Xâm nhập vào các thị trường mới và đa dạng hóa sản phẩm.
- **Nâng cao hiệu quả hoạt động:** Tối ưu hóa các quy trình nội bộ và giảm chi phí.

Nguồn:

<https://www.kaggle.com/datasets/datascientist97/e-commerce-sales-data-2024>

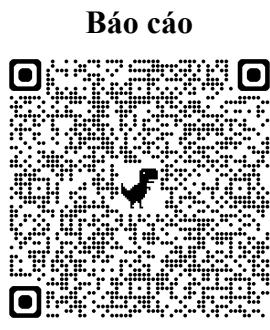
<https://www.kaggle.com/datasets/terencicp/e-commerce-dataset-by-olist-as-an-sqlite-database>



Slide



Trello



Báo cáo



Power BI