



TÀI LIỆU DỰ ÁN

XỬ LÝ DỮ LIỆU GIÁ Ô TÔ - HappyCar



GIẢNG VIÊN : TONNQ

HỌC VIÊN : GROUP 2

LỚP/KỲ : DP19303/SU24

MỤC LỤC

1	Giới thiệu dự án.....	5
1.1	Giới thiệu công ty HappyCar.....	5
1.2	Yêu cầu của công ty.....	7
1.3	Lập kế hoạch dự án.....	8
2	Phân tích yêu cầu khách hàng.....	9
2.1	Phân tích yêu cầu.....	9
2.2	Câu chuyện dữ liệu.....	13
2.2.1	Đặt vấn đề.....	13
2.2.2	Xác định câu chuyện.....	16
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.....	23
2.2.6	Những điều cần lưu ý.....	24
2.3	Kiến trúc hệ thống.....	25
2.3.1	Kiến trúc.....	25
2.3.2	Giải thích.....	26
2.4	Giải thích về bộ dữ liệu khách hàng.....	28
2.4.1	Các khái niệm.....	28
2.4.2	Các trường dữ liệu.....	30
3	Làm sạch và chuyển 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.....	34
3.2	Làm sạch dữ liệu.....	42
3.2.1	Các vấn đề ảnh hưởng tới dữ liệu.....	42
3.2.2	Các tiêu chí đánh giá chất lượng dữ liệu.....	47
3.2.3	Các bước làm sạch dữ liệu.....	49
3.3	Chuyển đổi dữ liệu.....	57

3.3.1	Các trường hợp cần chuyển đổi.....	57
3.3.2	Các kỹ thuật chuyển đổi.....	59
3.3.3	Trình bày các phép chuyển đổi trong dự án.....	61
4	Xử lý dữ liệu.....	64
4.1	Chuẩn hóa dữ liệu.....	64
4.1.1	Trình bày các bước chuẩn hóa trong dự án.....	64
4.2	Mô hình hóa dữ liệu.....	68
4.2.1	Các loại mô hình hóa.....	68
4.2.2	Các tiêu chí đánh giá mô hình dữ liệu.....	70
4.2.3	Trình bày các bước mô hình hóa.....	73
4.2.4	Trình bày các bước tạo bảng dữ liệu.....	76
4.3	Xử lý dữ liệu DAX.....	81
4.3.1	Measure.....	81
4.3.2	Calculated column.....	86
4.3.3	Filter.....	88
5	Trực quan hóa dữ liệu.....	91
5.1	Các kỹ thuật trực quan hóa.....	91
5.2	Các nguyên tắc trực quan hóa.....	94
5.3	Trình bày cách thêm visual mới.....	98
5.4	Trình bày tạo các report cho dự án.....	103
5.4.1	Tạo visual thống kê chi tiết.....	103
5.4.2	Tạo visual thống kê tổng thể.....	105
6	Xây dựng báo cáo.....	107
6.1	Dashboard và report.....	107
6.2	Xây dựng báo cáo.....	109
6.2.1	Dashboard vs Report.....	109
6.2.2	Dashboard.....	111
6.2.3	Report.....	115
6.2.4	Bookmark.....	121

7	KẾT LUẬN.....	124
7.1	Báo cáo.....	124
7.1.1	Các bước viết báo cáo.....	124
7.1.2	Tổng hợp.....	126
7.2	Khó khăn.....	127
7.3	Thuận lợi.....	127
7.4	Hướng phát triển.....	128
	 Hình 2.1 Mô hình kiến trúc hệ thống.....	29
	Hình 3.1 Tạo cơ sở dữ liệu phân phối.....	41
	Hình 3.2 Thiết lập máy chủ phân phối.....	41
	Hình 3.3 Cấu hình máy chủ xuất bản.....	41
	Hình 3.4 Tạo án phẩm.....	41
	Hình 3.5 Thêm các bài viết vào án phẩm.....	42
	Hình 3.6 Tạo đăng ký.....	42
	Hình 3.7 Dữ liệu thiếu (null).....	49
	Hình 3.8 Dữ liệu trùng lặp.....	49
	Hình 3.9 Nhân bản cơ sở dữ liệu bị lỗi.....	50
	Hình 3.10 Thời gian truy vấn lâu.....	51
	Hình 3.11 Xác định giá trị null.....	55
	Hình 3.12 Xác định giá trị không hợp lệ.....	55
	Hình 3.13 Xác định bản ghi trùng lặp.....	56
	Hình 3.14 Xác định giá trị ngoại lai.....	57
	Hình 3.15 Điện giá trị thiếu.....	58
	Hình 3.16 Loại bỏ bản ghi có các trường bắt buộc không thể thiếu.....	58
	Hình 3.17 Sửa chữa giá trị không hợp lệ.....	59
	Hình 3.18 Loại bỏ các bản ghi trùng lặp.....	59
	Hình 3.19 Loại bỏ các giá trị ngoại lai.....	60
	Hình 3.20 Kiểm tra lại dữ liệu.....	61

Hình 3.21 Đặt các ràng buộc.....	61
Hình 3.22 Chuyển đổi ngày tháng sang kiểu datetime.....	66
Hình 3.23 Chuyển đổi giá trị số.....	66
Hình 3.24 Định dạng ngày tháng.....	66
Hình 3.25 Định dạng số tiền.....	67
Hình 3.26 Chuyển đổi về dạng chữ thường.....	67
Hình 3.27 Thay thế giá trị không nhất quán/ không chính xác.....	67
Hình 3.28 Tính tuổi của xe.....	68
Hình 3.29 Tính giá trị còn lại của xe.....	68
Hình 3.30 Chuyển đơn vị từ km sang dặm.....	69
Hình 4.1 Chuyển đổi kiểu cho giá trị số.....	70
Hình 4.2 Duplicate cột [saledate] thành [saledate-Copy].....	71
Hình 4.3 Tách cột saledate-copy thành Thứ, Ngày, Tháng, Năm, Giờ.....	71
Hình 4.4 Xóa các cột không cần thiết.....	72
Hình 4.5 Gộp các cột ngày, tháng, năm và chuyển kiểu dữ liệu về kiểu “Date” chuyển kiểu dữ liệu của cột Time về kiểu “Time”	72
Hình 4.6 Close and Apply.....	73
Hình 4.7 Mô hình hóa dữ liệu.....	81
Hình 4.8 Tạo cơ sở dữ liệu.....	82
Hình 4.9 Tạo bảng CarPricesRaw.....	83
Hình 4.10 Tạo bảng Car.....	85
Hình 4.11 Tạo bảng Condition.....	86
Hình 4.12 Tạo bảng Seller.....	86
Hình 4.13 Tạo bảng CarTransaction.....	86
Hình 4.14 Tạo bảng Calendar Transaction Date.....	87
Hình 4.15 Tạo liên kết.....	88
Hình 4.16 Tạo measure Số lượng giao dịch.....	88
Hình 4.17 Tạo measure số lượng Condition.....	89
Hình 4.18 Tạo measure số lượng xe.....	89

Hình 4.19 Tạo measure số lượng seller.....	90
Hình 4.20 Tạo measure Total Revenue.....	90
Hình 4.21 Tạo measure Average Selling Price.....	91
Hình 4.22 Tạo measure Average MMR.....	91
Hình 4.23 Tạo measure Total Odometer.....	91
Hình 4.24 Tạo cột tổng doanh thu trong bảng Seller.....	92
Hình 4.25 Tạo cột Car Age.....	92
Hình 4.26 Tạo cột Condition by Odometer.....	92
Hình 4.27 Tạo cột Profit.....	92
Hình 4.28 Tạo filter chọn hãng sản xuất.....	93
Hình 4.29 Tạo filter chọn mẫu xe.....	94
Hình 4.30 Tạo filter chọn năm sản xuất.....	94
Hình 4.32 Tạo filter chọn transmission.....	95
Hình 4.33 Tạo filter chọn Seller.....	95
Hình 5.1 Get more visual.....	103
Hình 5.2 Tìm visual mới cần thêm.....	104
Hình 5.3 Add visual.....	104
Hình 5.4 Hoàn thành.....	105
Hình 5.5 Tạo filter theo ngày giao dịch.....	105
Hình 5.6 Tạo filter theo hãng sản xuất.....	106
Hình 5.7 Tạo filter theo model.....	106
Hình 5.8 Tạo visual filter theo ngày giao dịch.....	107
Hình 5.9 Tạo visual thống kê total sales.....	107
Hình 5.10 Tạo visual thống kê profit margin.....	108
Hình 5.11 Tạo visual thống kê Sum of selling price theo thời gian.....	108
Hình 6.1 Report Overview.....	120
Hình 6.2 Report Overview có thêm Buttons.....	122
Hình 6.3 Tạo bookmark slicer chọn tất cả các ngày.....	123
Hình 6.4 Tạo buttons hiện tất cả các ngày.....	124

Hình 6.5 Tạo bookmark slicer chọn ngày gần nhất.....125

Hình 6.6 Tạo bookmark slicer chọn ngày xa nhất.....126

1 GIỚI THIỆU DỰ ÁN

1.1 GIỚI THIỆU CÔNG TY HAPPYCAR

Công ty HappyCar là một trong những đơn vị tiên phong và uy tín hàng đầu trong lĩnh vực mua bán ô tô cũ tại Việt Nam. Với sứ mệnh mang đến cho khách hàng những chiếc xe chất lượng cao với giá cả hợp lý, HappyCar đã xây dựng được niềm tin vững chắc từ hàng ngàn khách hàng trên khắp cả nước. Được thành lập với đội ngũ chuyên gia có nhiều năm kinh nghiệm trong ngành ô tô, công ty cam kết cung cấp các dịch vụ kiểm định xe chuyên nghiệp, đảm bảo mọi chiếc xe đều được kiểm tra kỹ lưỡng về chất lượng và an toàn trước khi đến tay khách hàng.

HappyCar không chỉ chú trọng đến chất lượng sản phẩm mà còn đề cao trải nghiệm khách hàng. Công ty đã đầu tư mạnh mẽ vào hệ thống cơ sở hạ tầng hiện đại, từ các showroom trưng bày xe đến các trung tâm dịch vụ bảo dưỡng và sửa chữa xe. Điều này giúp đảm bảo rằng mọi khách hàng khi đến với HappyCar đều được phục vụ tận tâm, chu đáo và được tư vấn một cách chi tiết, rõ ràng nhất về tình trạng và giá trị của từng chiếc xe.

Quy trình mua bán tại HappyCar được thiết kế minh bạch, nhanh chóng và hiệu quả. Khách hàng có thể dễ dàng tìm kiếm, lựa chọn và thử nghiệm các mẫu xe thông qua website trực tuyến của công ty hoặc trực tiếp tại các showroom. Mọi thông tin về xe, từ lịch sử sử dụng, tình trạng kỹ thuật đến giá bán, đều được công khai rõ ràng, giúp khách hàng hoàn toàn yên tâm khi ra quyết định mua xe. Bên cạnh đó, HappyCar còn cung cấp các dịch vụ tài chính linh hoạt, hỗ trợ vay mua xe với lãi suất ưu đãi, giúp khách hàng dễ dàng sở hữu chiếc xe mơ ước mà không gặp phải khó khăn về tài chính.

Không chỉ dừng lại ở việc bán xe, HappyCar còn đặc biệt chú trọng đến dịch vụ hậu mãi. Công ty cung cấp các gói bảo hành toàn diện, dịch vụ bảo dưỡng định

kỹ và sửa chữa chuyên nghiệp, đảm bảo chiếc xe của khách hàng luôn hoạt động tốt và an toàn. Đội ngũ kỹ thuật viên của HappyCar được đào tạo bài bản, có tay nghề cao và luôn sẵn sàng hỗ trợ khách hàng trong mọi tình huống.

Tầm nhìn của HappyCar là trở thành đối tác tin cậy hàng đầu của người tiêu dùng Việt Nam trong việc tìm kiếm và sở hữu những chiếc xe ô tô cũ đáng tin cậy, góp phần nâng cao chất lượng cuộc sống và đáp ứng nhu cầu di chuyển ngày càng cao của xã hội hiện đại. Với phương châm “Chất lượng tạo niềm tin”, HappyCar không ngừng nỗ lực để cải tiến và phát triển, mang đến cho khách hàng những trải nghiệm mua sắm xe ô tô cũ tốt nhất. Bằng sự tận tâm và chuyên nghiệp, HappyCar đã, đang và sẽ tiếp tục khẳng định vị thế của mình trên thị trường, trở thành cái tên được nhiều người tiêu dùng tin tưởng và lựa chọn. Hiện tại, HappyCar đang có một lượng lớn xe ô tô cũ cần được phân tích để đưa ra các quyết định kinh doanh hợp lý. Công ty gặp khó khăn trong việc định giá và quản lý kho xe ô tô cũ do thiếu thông tin chi tiết về thị trường và xu hướng giá cả.

Bộ dữ liệu "car_prices.csv" là một nguồn thông tin phong phú và chi tiết về các xe ô tô đã qua sử dụng được bán tại Mỹ, cung cấp nhiều yếu tố quan trọng liên quan đến từng chiếc xe. Cụ thể, dữ liệu bao gồm năm sản xuất (year), hãng sản xuất (make) như Kia, BMW, Volvo, dòng xe (model) như Sorento, 3 Series, S60, và phiên bản (trim) của dòng xe. Ngoài ra, dữ liệu còn ghi nhận loại xe (body) như SUV, Sedan, loại hộp số (transmission) như hộp số tự động (automatic), và mã số định danh xe (vin). Thông tin về bang nơi xe được bán (state), tình trạng xe (condition) được đánh giá trên thang điểm, số dặm xe đã đi (odometer), màu xe (color), và màu nội thất (interior) cũng được bao gồm.Thêm vào đó, dữ liệu cung cấp thông tin về người bán hoặc công ty bán xe (seller), giá trị thị trường của xe (mmr) theo báo cáo Manheim Market, giá bán thực tế (sellingprice), và ngày bán xe (saledate). Bộ dữ liệu này là công cụ hữu ích cho việc phân tích thị trường ô tô

cũ, giúp đánh giá giá trị xe, xu hướng giá cả, và cung cấp những thông tin cần thiết cho người mua và người bán.

1.2 YÊU CẦU CỦA CÔNG TY

1.2.1 YÊU CẦU CỦA CÔNG TY HAPPYCAR

1.2.1.1 Về mặt dữ liệu

Cần phân tích dữ liệu để hiểu rõ xu hướng giá cả, tình trạng xe, và các yếu tố ảnh hưởng đến giá bán.

Phân tích xu hướng giá cả: Công ty cần hiểu rõ xu hướng giá cả của các loại xe ô tô cũ theo thời gian để có thể định giá hợp lý. Việc phân tích xu hướng này sẽ giúp HappyCar nắm bắt được giá trị thực của các loại xe trong kho.

Phân tích tình trạng xe: Cần đánh giá tình trạng xe (số dặm đã đi, năm sản xuất, tình trạng bảo dưỡng, vv.) để xác định các yếu tố ảnh hưởng đến giá bán và chất lượng xe.

Đánh giá các yếu tố ảnh hưởng đến giá bán: Cần xác định các yếu tố quan trọng nhất ảnh hưởng đến giá bán của xe ô tô cũ, chẳng hạn như hãng xe, dòng xe, năm sản xuất, số dặm đã đi, và tình trạng xe.

1.2.1.2 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.

1.2.1.3 Mục tiêu

Đưa ra các khuyến nghị cụ thể để tối ưu hóa việc mua bán xe ô tô cũ, giúp công ty tăng doanh thu và cải thiện chất lượng dịch vụ.

Tối ưu hóa việc mua bán xe ô tô cũ: Sử dụng dữ liệu phân tích để đưa ra các chiến lược kinh doanh hiệu quả, giúp công ty tăng doanh thu và cải thiện chất lượng dịch vụ.

Cải thiện chất lượng dịch vụ: Sử dụng kết quả phân tích để nâng cao quy trình kiểm tra và chuẩn bị xe, đảm bảo rằng khách hàng luôn nhận được sản phẩm chất lượng nhất.

Đưa ra khuyến nghị cụ thể: Dựa trên kết quả phân tích, đưa ra các khuyến nghị cụ thể về cách quản lý và kinh doanh xe ô tô cũ, giúp công ty nắm bắt cơ hội và đối phó với thách thức trên thị trường.

1.2.2 TÍNH KHẢ THI

Năng lực hiện có

Kỹ năng phân tích dữ liệu, sử dụng các công cụ như Excel, SQL Server, Power BI.

Năng lực học thêm

Các kỹ năng trực quan hóa nâng cao, kể chuyện bằng dữ liệu, sử dụng Tableau và Python.

1.3 LẬP KẾ HOẠCH DỰ ÁN

TT	HẠNG MỤC	BẮT ĐẦU	KẾT THÚC	KẾT QUẢ
1	Giới thiệu dự án	01/07/2024	02/07/2024	Báo cáo giới thiệu dự án
1.1	Giới thiệu công ty	01/07/2024	01/07/2024	Mô tả công ty HappyCar
1.2	Yêu cầu công ty	01/07/2024	01/07/2024	Liệt kê yêu cầu của HappyCar
1.3	Lập kế hoạch dự án	01/07/2024	02/07/2024	Kế hoạch dự án chi tiết
2	Phân tích yêu cầu	03/07/2024	07/07/2024	Báo cáo phân tích yêu cầu khách hàng
2.1	Phân tích yêu cầu KH	03/07/2024	04/07/2024	Mô tả chi tiết yêu cầu KH
2.2	Câu chuyện dữ liệu	04/07/2024	05/07/2024	Mô tả câu chuyện dữ liệu
2.3	Kiến trúc hệ thống	05/07/2024	06/07/2024	Mô tả kiến trúc hệ thống
2.4	Giải thích về bộ dữ liệu KH	06/07/2024	07/07/2024	Mô tả và giải thích bộ dữ liệu KH
3	Làm sạch và chuyển đổi dữ liệu	08/07/2024	15/07/2024	Dữ liệu sạch và chuẩn bị sẵn sàng
3.1	Chuẩn bị dữ liệu	08/07/2024	10/07/2024	Dữ liệu đã chuẩn bị sẵn sàng
3.2	Làm sạch dữ liệu	11/07/2024	13/07/2024	Dữ liệu sạch
3.3	Chuyển đổi dữ liệu	14/07/2024	15/07/2024	DL đã sạch, chuyển đổi
4	Xử lý dữ liệu	16/07/2024	25/07/2024	Dữ liệu đã được xử lý
4.1	Chuẩn hóa dữ liệu	16/07/2024	18/07/2024	Dữ liệu được chuẩn hóa
4.2	Mô hình hóa dữ liệu	19/07/2024	21/07/2024	Mô hình hóa dữ liệu
4.3	Xử lý dữ liệu DAX	22/07/2024	25/07/2024	Xử lý dữ liệu DAX
5	Trực quan hóa dữ liệu	26/07/2024	30/07/2024	Dữ liệu trực quan hóa
5.1	Các kỹ thuật trực quan hóa	26/07/2024	26/07/2024	Áp dụng các kỹ thuật trực quan hóa
5.2	Các nguyên tắc trực quan hóa	27/07/2024	27/07/2024	Áp dụng các nguyên tắc trực quan hóa

5.3	Trình bày cách thêm visual mới	28/07/2024	29/07/2024	Trình bày cách thêm visual mới
5.4	Trình bày tạo các report cho dự án	29/07/2024	30/07/2024	Trình bày tạo các report cho dự án
6	Xây dựng báo cáo	31/07/2024	03/08/2024	Báo cáo hoàn chỉnh
6.1	Dashboard và Report	31/07/2024	01/08/2024	Tạo Dashboard và report
6.2	Xây dựng báo cáo	02/08/2024	03/08/2024	Xây dựng báo cáo
7	Kết luận	04/08/2024	06/08/2024	Báo cáo kết luận
7.1	Báo cáo	04/08/2024	04/08/2024	Báo cáo tổng hợp
7.2	Khó khăn	05/08/2024	05/08/2024	Đánh giá khó khăn
7.3	Thuận lợi	06/08/2024	06/08/2024	Đánh giá thuận lợi
7.4	Hướng phát triển	07/08/2024	07/08/2024	Đề xuất hướng phát triển

2 PHÂN TÍCH YÊU CẦU KHÁCH HÀNG

2.1 PHÂN TÍCH YÊU CẦU

2.1.1 DỮ LIỆU

Bộ dữ liệu này cung cấp một cái nhìn tổng quan về thị trường xe ô tô cũ, bao gồm các yếu tố quan trọng ảnh hưởng đến giá bán và tình trạng xe. Bộ dữ liệu "car_prices.csv" bao gồm các thông tin sau:

- year: Năm sản xuất của xe.
- make: Hãng sản xuất xe (ví dụ: Kia, BMW, Volvo).
- model: Dòng xe (ví dụ: Sorento, 3 Series, S60).
- trim: Phiên bản của dòng xe.
- body: Loại xe (ví dụ: SUV, Sedan).
- transmission: Loại hộp số (ví dụ: automatic).
- vin: Mã số định danh của xe.
- state: Bang nơi xe được bán.

- condition: Tình trạng của xe, được đánh giá trên thang điểm (ví dụ: 5.0 là mới, 45.0 là cũ).
- odometer: Số dặm xe đã đi.
- color: Màu xe.
- interior: Màu nội thất.
- seller: Người bán hoặc công ty bán xe.
- mmr: Giá trị thị trường của xe (Manheim Market Report).
- sellingprice: Giá bán thực tế của xe.
- saledate: Ngày bán xe.

2.1.2 QUẢN LÝ VÀ LUU TRỮ

Để quản lý và lưu trữ bộ dữ liệu này hiệu quả, chúng ta cần sử dụng một hệ quản trị cơ sở dữ liệu mạnh mẽ và linh hoạt. MongoDB là một lựa chọn tốt vì:

- **Khả năng lưu trữ không giới hạn:** MongoDB có khả năng lưu trữ dữ liệu lớn và có thể mở rộng quy mô khi cần thiết.
- **Dễ dàng truy xuất và quản lý:** MongoDB hỗ trợ các truy vấn phức tạp và dễ dàng quản lý dữ liệu thông qua giao diện dòng lệnh hoặc các công cụ GUI như MongoDB Compass.
- **Tính linh hoạt:** MongoDB cho phép lưu trữ dữ liệu dưới dạng tài liệu JSON, rất phù hợp với cấu trúc không đồng nhất của dữ liệu xe ô tô.

Hoặc SQL vì:

- **Khả năng lưu trữ và xử lý dữ liệu lớn:** SQL Server có khả năng lưu trữ và xử lý dữ liệu lớn, đáp ứng được nhu cầu phân tích dữ liệu của dự án.
- **Hỗ trợ các truy vấn phức tạp:** SQL Server hỗ trợ các truy vấn SQL mạnh mẽ, giúp dễ dàng thực hiện các phép tính và phân tích phức tạp trên dữ liệu.

- Tích hợp tốt với các công cụ phân tích và trực quan hóa: SQL Server có thể tích hợp dễ dàng với các công cụ như Power BI để trực quan hóa dữ liệu và tạo báo cáo.
- Chi phí vận hành thấp, đơn giản, dễ tiếp cận.

2.1.3 CÔNG NGHỆ

Để phân tích và trực quan hóa dữ liệu, chúng ta sử dụng các công cụ và ngôn ngữ lập trình sau:

- Python: Ngôn ngữ lập trình mạnh mẽ và linh hoạt, phù hợp cho việc xử lý dữ liệu và phân tích thống kê.

Pandas: Thư viện mạnh mẽ để thao tác và phân tích dữ liệu.

Matplotlib và Seaborn: Thư viện để trực quan hóa dữ liệu dưới dạng biểu đồ và đồ thị.

- Power BI hoặc Tableau: Công cụ trực quan hóa dữ liệu mạnh mẽ, giúp tạo ra các báo cáo và dashboard tương tác, dễ dàng chia sẻ và trình bày kết quả phân tích.
- SQL Server: Hệ quản trị cơ sở dữ liệu quan hệ, thực hiện các truy vấn phức tạp và lưu trữ dữ liệu lớn.
- Excel: Công cụ hỗ trợ phân tích dữ liệu và trực quan hóa cơ bản, phù hợp cho các báo cáo nhanh và phân tích đơn giản.

2.1.4 QUYẾT ĐỊNH CÔNG NGHỆ SỬ DỤNG

LÝ DO

- Các công nghệ và công cụ được chọn dựa trên khả năng xử lý dữ liệu lớn, tính linh hoạt trong phân tích và trực quan hóa, và khả năng tích hợp dễ dàng

vào hệ thống hiện tại của công ty. Việc chọn SQL Server giúp dễ dàng xử lý và phân tích dữ liệu, trong khi Power BI hoặc Tableau cung cấp khả năng trực quan hóa mạnh mẽ.

- Một lý do cũng rất quan trọng đó là SQL Server và Power BI được chúng tôi sử dụng thường xuyên, quen thuộc, dễ tương thích Team vận hành hiệu quả.
- Chi phí vận hành thấp, dễ tiếp cận.

DỮ LIỆU

- Excel, Power BI: Để làm sạch và chuẩn bị dữ liệu, xử lý các thao tác dữ liệu phức tạp.
- SQL Server: Để thực hiện các truy vấn phức tạp và lưu trữ dữ liệu lớn.

QUẢN LÝ VÀ LUU TRỮ

SQL Server: Để quản lý và lưu trữ dữ liệu, hỗ trợ các truy vấn phức tạp và tích hợp với các công cụ phân tích.

CÔNG NGHỆ

- Power BI hoặc Tableau: Để trực quan hóa và trình bày dữ liệu, tạo ra các báo cáo và dashboard tương tác, giúp dễ dàng theo dõi và phân tích các xu hướng quan trọng.
- Excel, SQL Server là những công cụ vô cùng quen thuộc, thân thiện, không quá phức tạp, dễ dàng sử dụng cho cả người trong hay ngoài chuyên môn.
- Chi phí hợp lý
- Có thể mở rộng nhanh chóng, đơn giản, hiệu quả.

2.2 CÂU CHUYỆN DỮ LIỆU

2.2.1 ĐẶT VẤN ĐỀ

2.2.1.1 Mô tả thực trạng

Hiện nay rất nhiều doanh nghiệp hoặc cá nhân ứng dụng phân tích dữ liệu vào quá trình quản lý, kinh doanh, đầu tư. Công ty HappyCar là một trong số đó. HappyCar là một công ty chuyên kinh doanh xe ô tô cũ với mục tiêu cung cấp các sản phẩm chất lượng và dịch vụ khách hàng tốt nhất.

Hiện tại, HappyCar đang có một lượng lớn xe ô tô cũ cần được phân tích để đưa ra các quyết định kinh doanh hợp lý. Công ty gặp khó khăn trong việc định giá và quản lý kho xe ô tô cũ do thiếu thông tin chi tiết về thị trường và xu hướng giá cả. Công ty mong muốn sử dụng phân tích dữ liệu để hiểu rõ hơn về thị trường xe hơi, từ đó tối ưu hóa chiến lược kinh doanh và cải thiện doanh thu.

2.2.1.2 Dữ liệu liên quan

Dữ liệu chúng tôi sử dụng bao gồm thông tin về các xe được bán trong thời gian gần đây, bao gồm năm sản xuất, hãng xe, mẫu xe, tình trạng xe, số km đã đi, màu sắc, nội thất, người bán, giá trị thị trường và giá bán thực tế, cùng với ngày bán. Sau khi xử lý, dữ liệu chỉ là một loạt các con số khô khan. Vấn đề đặt ra là làm thế nào để diễn giải và truyền đạt những kết quả này tới cấp trên và đồng nghiệp một cách rõ ràng và thuyết phục?

Chúng ta có thể diễn giải và truyền đạt kết quả phân tích dữ liệu một cách rõ ràng, hấp dẫn và thuyết phục tới cấp trên và đồng nghiệp, giúp họ hiểu rõ hơn về thị trường và đưa ra các quyết định kinh doanh thông minh.

Để truyền đạt kết quả phân tích dữ liệu một cách hiệu quả, chúng ta cần:

- Sử dụng biểu đồ và hình ảnh: Minh họa các phát hiện và xu hướng bằng các biểu đồ trực quan như biểu đồ hộp (box plot), biểu đồ phân tán (scatter plot),

và biểu đồ đường (line plot) để giúp cấp trên và đồng nghiệp dễ dàng hiểu được các thông tin quan trọng.

- Trình bày kết quả theo cách logic và rõ ràng: Sắp xếp các phân tích theo thứ tự hợp lý, bắt đầu từ việc giới thiệu dữ liệu, mô tả các bước xử lý dữ liệu, đến trình bày các kết quả phân tích và kết luận cuối cùng.
- Giải thích chi tiết và cung cấp bối cảnh: Đưa ra các giải thích chi tiết về các phát hiện từ dữ liệu và cách chúng có thể được sử dụng để đưa ra các quyết định kinh doanh. Cung cấp bối cảnh và các ví dụ cụ thể để minh họa các điểm quan trọng.
- Kết hợp số liệu và ví dụ thực tế: Sử dụng các số liệu cụ thể từ phân tích để hỗ trợ các điểm chính, đồng thời đưa ra các ví dụ thực tế để làm rõ hơn các phát hiện.
- Đưa ra các đề xuất và chiến lược cụ thể: Dựa trên kết quả phân tích, đưa ra các đề xuất cụ thể và chiến lược kinh doanh để tối ưu hóa việc quản lý và bán xe ô tô cũ của công ty.
- Sử dụng Dashboard, xây dựng câu chuyện dữ liệu, trình bày trong hội thảo một cách tự tin, slide rõ ràng, tạo cơ hội hỏi đáp,..

2.2.1.3 Mục tiêu

Câu chuyện dữ liệu không chỉ giúp truyền đạt thông tin một cách hiệu quả mà còn giúp tạo động lực hành động, hỗ trợ việc ra quyết định và tối ưu hóa các chiến lược kinh doanh của công ty HappyCar.

Truyền đạt thông điệp dữ liệu rõ ràng và hấp dẫn:

- Hiểu rõ vấn đề: Giúp người nghe dễ dàng nắm bắt được các vấn đề cốt lõi và các xu hướng chính từ dữ liệu.

- Đơn giản hóa thông tin phức tạp: Chuyển đổi các số liệu và dữ liệu phức tạp thành các thông tin đơn giản, dễ hiểu thông qua việc sử dụng các công cụ trực quan như biểu đồ và đồ thị.

Thể hiện dữ liệu sinh động và dễ hiểu:

- Sử dụng biểu đồ và hình ảnh trực quan: Minh họa các phát hiện từ dữ liệu bằng các biểu đồ hộp (box plot), biểu đồ phân tán (scatter plot), và biểu đồ đường (line plot) để làm nổi bật các xu hướng và mẫu hình.
- Giải thích chi tiết: Cung cấp các giải thích chi tiết về các phát hiện, kèm theo bối cảnh và các ví dụ cụ thể để làm rõ ý nghĩa của dữ liệu.

Thuyết phục và tạo động lực hành động:

- Đưa ra các bằng chứng rõ ràng: Sử dụng các số liệu cụ thể và minh họa từ phân tích để hỗ trợ các lập luận và đề xuất.
- Đề xuất các chiến lược cụ thể: Dựa trên kết quả phân tích, đưa ra các đề xuất chiến lược kinh doanh rõ ràng và cụ thể để giúp công ty cải thiện hiệu quả hoạt động và tối ưu hóa doanh thu.
- Kết nối với khán giả: Tạo ra một câu chuyện hấp dẫn và có liên quan để kết nối với cấp trên và đồng nghiệp, giúp họ thấy được tầm quan trọng của dữ liệu và các phát hiện từ phân tích.

Tạo ra những quyết định kinh doanh thông minh:

- Hỗ trợ việc ra quyết định: Cung cấp thông tin và các bằng chứng từ dữ liệu để hỗ trợ việc ra quyết định kinh doanh, giúp công ty đưa ra các quyết định dựa trên dữ liệu thay vì cảm tính.
- Tối ưu hóa chiến lược kinh doanh: Dựa trên các phân tích và phát hiện, tối ưu hóa các chiến lược kinh doanh để cải thiện hiệu quả và tăng trưởng doanh thu.

2.2.2 XÁC ĐỊNH CÂU CHUYỆN

2.2.2.1 Hình thành giả thuyết

Bạn đang cố gắng giải thích điều gì từ dữ liệu?

- Giá trị xe ô tô cũ: Yếu tố nào ảnh hưởng nhiều nhất đến giá trị bán lại của xe ô tô cũ.
- Xu hướng giá cả: Giá bán xe ô tô cũ thay đổi như thế nào theo thời gian, và có xu hướng gì nổi bật không?
- Ảnh hưởng của tình trạng xe: Tình trạng và số km đã đi của xe ảnh hưởng đến giá bán như thế nào?
- Xu hướng và mô hình: Phân tích xu hướng dài hạn và ngắn hạn để dự đoán sự phát triển và hiểu rõ biến động.
- Hiệu suất và hiệu quả: Đánh giá hiệu suất các chiến dịch/quy trình và đưa ra khuyến nghị cải thiện hiệu quả.
- Hành vi và sở thích khách hàng: Hiểu hành vi mua sắm, phân khúc khách hàng để tùy chỉnh chiến lược tiếp thị.
- Đánh giá đầu tư: Đánh giá hiệu quả và đưa ra quyết định đầu tư sáng suốt.
- Rủi ro và cơ hội: Xác định rủi ro và cơ hội kinh doanh để đưa ra biện pháp giảm thiểu và tận dụng.
- Quản lý và phân bổ tài nguyên: Quản lý và phân bổ tài nguyên hiệu quả, tối ưu hóa quy trình quản lý.
- Cải thiện quy trình kinh doanh: Đánh giá và cải thiện quy trình hiện tại, đưa ra ý tưởng đổi mới.

Mục tiêu cụ thể của bạn khi kể câu chuyện dữ liệu này là gì?

- Hiểu rõ thị trường xe ô tô cũ: Giúp HappyCar hiểu rõ hơn về thị trường và các yếu tố ảnh hưởng đến giá trị của xe ô tô cũ.

- Tối ưu hóa chiến lược kinh doanh: Đưa ra các chiến lược định giá và quản lý kho xe dựa trên dữ liệu để tối ưu hóa doanh thu và hiệu quả kinh doanh.
- Truyền đạt thông tin: Trình bày các phát hiện từ dữ liệu một cách rõ ràng, dễ hiểu và thuyết phục, giúp cấp trên và đồng nghiệp nắm bắt và sử dụng thông tin hiệu quả.

Bạn muốn đề xuất giải pháp gì từ những phân tích này?

- Chiến lược định giá thông minh: Đề xuất chiến lược định giá dựa trên tình trạng, năm sản xuất và số km đã đi của xe.
- Quản lý kho xe hiệu quả: Tối ưu hóa quản lý kho xe dựa trên các phát hiện từ dữ liệu, như xu hướng giá cả và các yếu tố ảnh hưởng đến giá trị xe.
- Cải thiện dịch vụ khách hàng: Đề xuất cải thiện dịch vụ khách hàng dựa trên hiểu biết sâu hơn về thị trường và nhu cầu khách hàng.

2.2.2.2 Một số cách tiếp cận dữ liệu

Tìm kiếm mối tương quan:

- Mối tương quan giữa giá trị thị trường (MMR) và giá bán thực tế: Xác định mức độ tương quan giữa giá trị thị trường và giá bán thực tế để đánh giá tính chính xác của định giá hiện tại.
- Mối tương quan giữa số km đã đi và giá bán: Phân tích mức độ ảnh hưởng của số km đã đi đến giá trị bán lại của xe.

Xác định xu hướng:

- Xu hướng giá cả theo thời gian: Phân tích dữ liệu bán hàng để xác định các xu hướng giá cả theo thời gian, giúp dự báo và lập kế hoạch kinh doanh.
- Xu hướng giá cả theo tình trạng xe: Xác định cách tình trạng xe ảnh hưởng đến giá trị bán lại.

Rút ra so sánh:

- So sánh giá bán giữa các hãng xe: Phân tích và so sánh giá bán của các hãng xe khác nhau để xác định các hãng xe có giá trị cao hơn và chiến lược định giá phù hợp.
- So sánh giá trị bán lại giữa các năm sản xuất: So sánh giá trị bán lại của các xe sản xuất trong các năm khác nhau để xác định ảnh hưởng của năm sản xuất đến giá trị xe.

2.2.3 XÁC ĐỊNH RÕ ĐÓI TƯỢNG

2.2.3.1 Đồi tượng cần nghe câu chuyện

Ban giám đốc và cấp quản lý:

- Nhu cầu: Họ cần hiểu rõ hơn về thị trường xe ô tô cũ và các yếu tố ảnh hưởng đến giá trị bán lại của xe để đưa ra các quyết định chiến lược và quản lý hiệu quả.
- Lợi ích: Câu chuyện dữ liệu giúp họ nắm bắt các xu hướng và mối tương quan trong dữ liệu, từ đó tối ưu hóa chiến lược kinh doanh và cải thiện doanh thu.

Đội ngũ marketing và bán hàng:

- Nhu cầu: Họ cần thông tin về các yếu tố ảnh hưởng đến giá bán và xu hướng thị trường để xây dựng các chiến lược tiếp thị và bán hàng hiệu quả.
- Lợi ích: Câu chuyện dữ liệu cung cấp cơ sở để họ hiểu rõ hơn về nhu cầu của khách hàng và tối ưu hóa các chiến dịch marketing và bán hàng.

Nhân viên quản lý kho và dịch vụ khách hàng:

- Nhu cầu: Họ cần hiểu rõ tình trạng và giá trị của các xe trong kho để quản lý kho hiệu quả và cung cấp dịch vụ tốt hơn cho khách hàng.
- Lợi ích: Câu chuyện dữ liệu giúp họ tối ưu hóa quy trình quản lý kho và nâng cao chất lượng dịch vụ khách hàng.

2.2.3.2 Họ đã biết đến lĩnh vực này chưa?

- Ban giám đốc và cấp quản lý: Họ có kiến thức cơ bản về thị trường xe ô tô cũ và các yếu tố ảnh hưởng đến giá trị xe, nhưng cần các phân tích cụ thể và chi tiết từ dữ liệu để hỗ trợ quyết định chiến lược.
- Đội ngũ marketing và bán hàng: Họ hiểu về nhu cầu và xu hướng thị trường ở mức cơ bản, nhưng cần thêm các thông tin chi tiết từ dữ liệu để xây dựng các chiến lược tiếp thị và bán hàng hiệu quả.
- Nhân viên quản lý kho và dịch vụ khách hàng: Họ biết về tình trạng và giá trị xe trong kho, nhưng cần các phân tích từ dữ liệu để tối ưu hóa quy trình quản lý và dịch vụ khách hàng.

2.2.3.3 Phương pháp truyền đạt thông tin

Ban giám đốc và cấp quản lý:

- Biểu đồ và hình ảnh trực quan: Sử dụng biểu đồ hộp, biểu đồ phân tán và biểu đồ đường để minh họa xu hướng và mối tương quan.
- Trình bày logic và rõ ràng: Trình bày theo thứ tự hợp lý từ giới thiệu đến kết luận.
- Giải thích chi tiết: Đưa ra giải thích và ví dụ cụ thể.
- Đề xuất chiến lược cụ thể: Đưa ra các chiến lược định giá và quản lý kho xe dựa trên dữ liệu.

Đội ngũ marketing và bán hàng:

- Biểu đồ và hình ảnh trực quan: Sử dụng biểu đồ để minh họa các yếu tố ảnh hưởng đến giá bán.
- Trình bày logic: Cung cấp thông tin theo thứ tự hợp lý để dễ theo dõi.
- Giải thích chi tiết: Giải thích ý nghĩa và bối cảnh của các phát hiện.
- Đề xuất chiến lược: Đưa ra chiến lược tiếp thị và bán hàng dựa trên phân tích dữ liệu.

Nhân viên quản lý kho và dịch vụ khách hàng:

- Biểu đồ và hình ảnh trực quan: Sử dụng biểu đồ để nắm bắt thông tin về tình trạng và giá trị xe.
- Trình bày logic và rõ ràng: Trình bày thông tin một cách dễ hiểu.
- Giải thích chi tiết: Cung cấp giải thích chi tiết và ví dụ minh họa.
- Đề xuất cụ thể: Đề xuất biện pháp quản lý kho và cải thiện dịch vụ khách hàng dựa trên dữ liệu.

2.2.3.4 Chuẩn bị một số câu hỏi cụ thể

Ban giám đốc và cấp quản lý:

- Làm thế nào để tối ưu hóa chiến lược định giá và quản lý kho xe?
- Những yếu tố nào ảnh hưởng lớn nhất đến giá trị bán lại của xe ô tô cũ?

Đội ngũ marketing và bán hàng:

- Xu hướng giá bán xe ô tô cũ theo thời gian là gì?
- Hàng xe nào có giá trị bán lại cao nhất và tại sao?

Nhân viên quản lý kho và dịch vụ khách hàng:

- Làm thế nào để tối ưu hóa quy trình quản lý kho xe dựa trên tình trạng và số km đã đi?
- Những thông tin nào cần cung cấp cho khách hàng để nâng cao chất lượng dịch vụ?

2.2.4 XÁC ĐỊNH CÂU CHUYỆN CHI TIẾT

2.2.4.1 Giới thiệu (Exposition)

Bối cảnh: HappyCar là một công ty chuyên kinh doanh xe ô tô cũ với mục tiêu cung cấp các sản phẩm chất lượng và dịch vụ khách hàng tốt nhất. Hiện tại, công ty đang gặp khó khăn trong việc định giá và quản lý kho xe ô tô cũ do thiếu thông tin

chi tiết về thị trường. Công ty mong muốn sử dụng phân tích dữ liệu để hiểu rõ hơn về thị trường xe hơi, từ đó tối ưu hóa chiến lược kinh doanh và cải thiện doanh thu.

Dữ liệu: Dữ liệu được thu thập từ các giao dịch mua bán xe ô tô cũ của HappyCar, bao gồm thông tin về năm sản xuất, hãng xe, mẫu xe, tình trạng xe, số km đã đi, giá trị thị trường (MMR), giá bán thực tế và ngày bán.

Mục tiêu: Phân tích dữ liệu để xác định các yếu tố ảnh hưởng đến giá bán xe ô tô cũ và đưa ra các chiến lược kinh doanh hợp lý.

Đối tượng: Ban giám đốc và cấp quản lý, đội ngũ marketing và bán hàng, nhân viên quản lý kho và dịch vụ khách hàng

2.2.4.2 Hành động gia tăng (Rising Action)

Xây dựng xung đột: Trong cuộc họp đầu tiên với ban quản lý, hai quan điểm trái ngược đã xuất hiện về yếu tố chính quyết định giá xe ô tô.

- Yếu tố Số Km đã đi: Một số thành viên tin rằng số km đã đi là yếu tố quan trọng nhất ảnh hưởng đến giá bán xe. Họ cho rằng xe có số km đã đi ít hơn sẽ có giá bán cao hơn.
- Yếu tố hãng xe: Một thành viên khác cho rằng hãng xe mới là yếu tố quan trọng nhất. Họ tin rằng một số hãng xe nhất định có giá trị bán lại cao hơn so với các hãng khác.

Dự đoán và tầm quan trọng: Câu hỏi đặt ra là "Yếu tố nào giữa số km đã đi và hãng xe ảnh hưởng nhiều nhất đến giá trị bán lại của xe ô tô cũ?"

2.2.4.3 Cao trào (Climax)

Giải thích phát hiện quan trọng: Nhóm của chúng tôi đã tiến hành phân tích dữ liệu và phát hiện ra những điều quan trọng sau:

Mỗi tương quan rõ ràng giữa giá trị thị trường (MMR) và giá bán thực tế. Biểu đồ phân tán cho thấy các xe có giá trị thị trường cao thường có giá bán thực tế cao.

Xe có số km đã đi ít hơn thường có giá bán cao hơn. Biểu đồ phân tán cho thấy mối quan hệ này.

Giá bán xe ô tô cũ có xu hướng thay đổi theo thời gian, có thể do ảnh hưởng của các yếu tố kinh tế và nhu cầu thị trường.

Giá bán của các hãng xe khác nhau có sự chênh lệch đáng kể. Một số hãng xe có giá trị bán lại cao hơn so với các hãng khác.

2.2.4.4 Hành động rơi (Falling Action)

Chi tiết bổ sung và giải thích: Phân tích sâu hơn cho thấy rằng cả số km đã đi và hãng xe đều là các yếu tố quan trọng ảnh hưởng đến giá bán. Tuy nhiên, một số hãng xe cụ thể có xu hướng giữ giá trị tốt hơn ngay cả khi có số km đã đi nhiều hơn.

Thảo luận các yếu tố cần xem xét thêm:

- Dữ liệu chỉ được thu thập ở một khu vực địa lý cụ thể hoặc từ một số hãng xe nhất định.
- Các hạn chế của dữ liệu và các yếu tố khác cần xem xét để có cái nhìn toàn diện hơn.

2.2.4.5 Giải quyết (Resolution)

Tóm tắt kết quả chính:

- Giá trị thị trường (MMR) và số km đã đi là hai yếu tố quan trọng ảnh hưởng đến giá bán xe.
- Hãng xe cũng có sự khác biệt đáng kể về giá trị bán lại.

Tầm quan trọng của kết quả: Hiểu rõ các yếu tố này giúp HappyCar định giá xe hợp lý hơn và quản lý kho xe hiệu quả, từ đó cải thiện doanh thu và lợi nhuận.

Quá trình hành động:

- Chiến lược định giá thông minh: Dựa vào tình trạng, năm sản xuất và số km đã đi của xe để định giá hợp lý.
- Quản lý kho xe hiệu quả: Tăng cường số lượng nhân sự để xử lý đơn đặt hàng kịp thời.
- Cải thiện dịch vụ khách hàng: Đưa ra các biện pháp cải thiện dịch vụ dựa trên hiểu biết sâu hơn về nhu cầu khách hàng.

Kết luận: Bằng cách sử dụng cấu trúc câu chuyện dữ liệu 5 phần và xây dựng xung đột giữa số km đã đi và hãng xe, chúng tôi đã truyền đạt một cách rõ ràng và thuyết phục các phát hiện từ dữ liệu của HappyCar. Điều này giúp công ty đưa ra các quyết định kinh doanh hiệu quả và nâng cao hiệu quả hoạt động.

2.2.5 TRÌNH BÀY DỮ LIỆU

2.2.5.1 Biểu đồ đề xuất

Biểu đồ cột (bar chart)

- Mục Đích: So sánh số liệu giữa các nhóm hoặc thời gian.
- Ứng Dụng: So sánh doanh số bán hàng giữa các tháng, sản phẩm, hoặc khu vực.

Biểu đồ đường (line chart)

- Mục đích: Hiển thị xu hướng theo thời gian.
- Ứng dụng: Theo dõi xu hướng tăng trưởng doanh số, nhu cầu theo mùa.

Biểu đồ tròn (pie chart)

- Mục đích: Thể hiện tỷ lệ phần trăm của các phần tử trong tổng thể.
- Ứng dụng: Thể hiện tỷ trọng doanh thu của các sản phẩm hoặc dịch.

Biểu đồ tán xạ (scatter plot)

- Mục đích: Tìm kiếm mối tương quan giữa hai biến số.
- Ứng dụng: Xác định mối quan hệ giữa giá cả và doanh số bán hàng, giá giá cả và số km đã đi, hãng xe, năm sản xuất,...

Biểu đồ nhiệt (heatmap)

- Mục đích: Thể hiện mật độ hoặc tần suất của dữ liệu.
- Ứng dụng: Phân tích hiệu suất bán hàng theo khu vực địa lý.

2.2.5.2 Lựa chọn biểu đồ cho các tình huống cụ thể

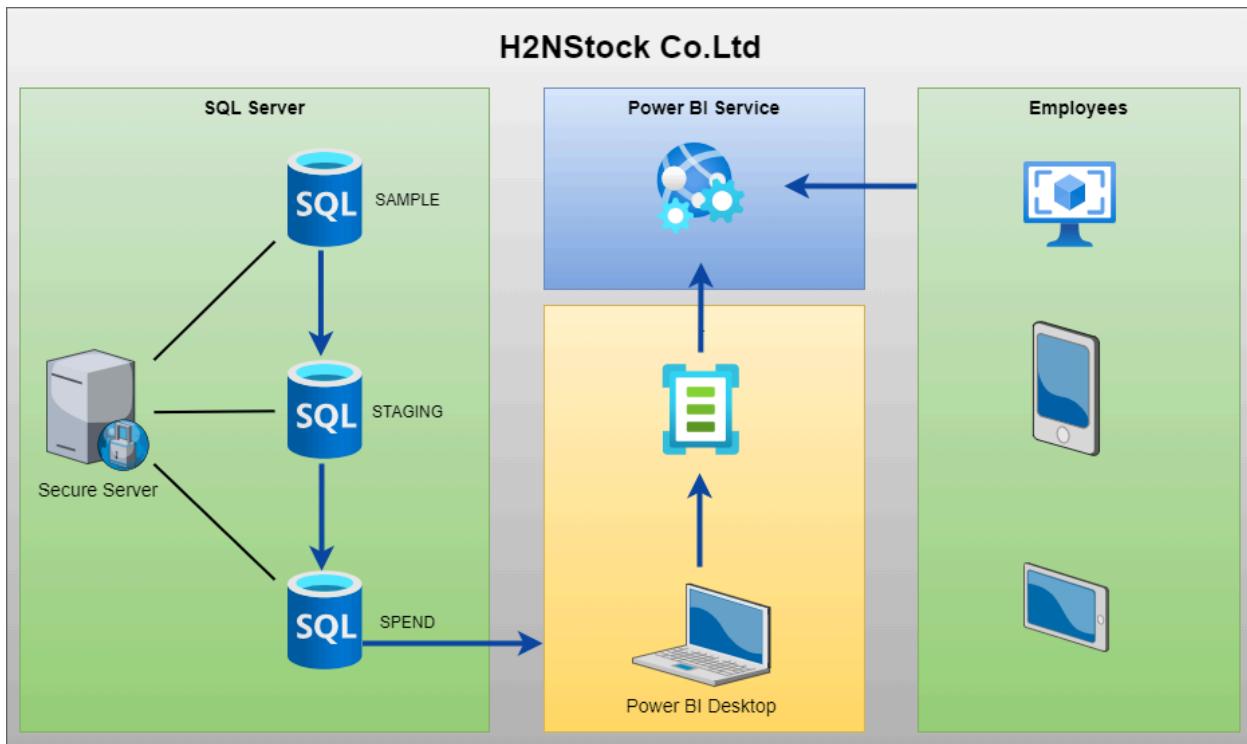
- So sánh doanh số giữa các tháng: Biểu đồ cột hoặc biểu đồ đường.
- Phân tích tỷ trọng doanh thu: Biểu đồ tròn.
- Tìm kiếm mối tương quan: Biểu đồ tán xạ.
- Phân tích hiệu suất bán hàng theo khu vực: Biểu đồ nhiệt.

2.2.6 NHỮNG ĐIỀU CẦN LUU Ý

- Chọn loại biểu đồ phù hợp: Đảm bảo loại biểu đồ phản ánh đúng mục tiêu và phù hợp với dữ liệu.
- Trình bày rõ ràng và dễ hiểu: Đơn giản hóa và dùng màu sắc, ký hiệu hợp lý để làm nổi bật điểm quan trọng.
- Đảm bảo tính chính xác: Đảm bảo dữ liệu chính xác và cung cấp nguồn gốc rõ ràng.
- Cung cấp bối cảnh và giải thích: Mỗi biểu đồ cần tiêu đề, chú thích, và giải thích rõ ràng, cung cấp bối cảnh đầy đủ.
- Tập trung vào người xem: Hiểu rõ đối tượng xem báo cáo và điều chỉnh cách trình bày để đáp ứng nhu cầu của họ.

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 mẫu.
- Staging : Cơ sở dữ liệu SQL Server trung gian này được sử dụng để phân loại dữ liệu.
- Spend : Cơ sở dữ liệu SQL Server này chứa dữ liệu chi tiêu cuối cùng.

Secure Server: Dữ liệu đi từ Secure Server đến cơ sở dữ liệu SAMPLE và STAGING 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.

Dịch vụ Power BI:

- 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 và STAGING).
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

2.3.2.1 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.
- STAGING Database: Đây là nơi dữ liệu từ SAMPLE được xử lý và làm sạch. STAGING Database hoạt động như một khu vực tạm thời để chuẩn bị dữ liệu cho bước tiếp theo.
- SPEND Database: Sau khi dữ liệu được làm sạch và xử lý trong STAGING, nó được chuyển đến SPEND Database, nơi lưu trữ dữ liệu đã được 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.

2.3.2.2 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 STAGING Database.
- Lưu trữ dữ liệu đã xử lý: Dữ liệu đã được xử lý trong STAGING Database được chuyển vào 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.

2.3.2.3 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

2.4.1.1 Các khái niệm cơ bản

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).

Giá thị trường trung bình (MMR - Manheim Market Report): Giá trị thị trường trung bình của xe, được sử dụng như một tham chiếu để định giá xe.

Định giá xe (Car Valuation): Quá trình xác định giá trị bán hợp lý của xe dựa trên các yếu tố như tình trạng, số km đã đi, và giá trị thị trường trung bình.

Phân tích thị trường (Market Analysis): Quá trình phân tích dữ liệu để hiểu rõ hơn về xu hướng, sở thích của người tiêu dùng và các yếu tố ảnh hưởng đến thị trường.

Trực quan hóa dữ liệu (Data Visualization): Sử dụng các công cụ như Power BI để tạo biểu đồ, bảng biểu, và dashboard giúp trực quan hóa dữ liệu và cung cấp cái nhìn sâu sắc về dữ liệu.

2.4.1.2 Nghiệp vụ liên quan

Nhập dữ liệu (Data Import):

- Nhập dữ liệu thô từ file CSV vào cơ sở dữ liệu.

- Công cụ: SQL Server (SSIS, BULK INSERT).

Làm sạch dữ liệu (Data Cleaning):

- Kiểm tra và loại bỏ các giá trị không hợp lệ, dữ liệu trùng lặp, và các hàng có giá trị thiếu.
- Ví dụ: Loại bỏ các hàng có giá trị NULL trong cột SaleDate

Tạo bảng và chuẩn hóa dữ liệu (Table Creation and Data Normalization):

- Tạo các bảng chuẩn hóa theo chuẩn 3NF từ dữ liệu thô.
- Ví dụ: Tạo bảng Cars, Sellers, và Transactions từ bảng DataRaw.

Nạp dữ liệu (Data Loading):

- Nạp dữ liệu đã làm sạch và chuẩn hóa vào các bảng chuẩn hóa.
- Ví dụ: Sử dụng câu lệnh INSERT INTO để nạp dữ liệu vào bảng Cars, Sellers, và Transactions.

Phân tích dữ liệu (Data Analysis):

- Phân tích dữ liệu để tìm hiểu xu hướng, mối quan hệ, và các yếu tố ảnh hưởng đến giá trị xe.
- Công cụ: SQL Server, Power BI.

Trực quan hóa dữ liệu (Data Visualization):

- Tạo các biểu đồ và dashboard để trực quan hóa dữ liệu và cung cấp cái nhìn sâu sắc về dữ liệu.
- Công cụ: Power BI.

Báo cáo và trình bày (Reporting and Presentation):

- Tạo các báo cáo và trình bày kết quả phân tích cho các bên liên quan.
- Công cụ: Power BI Service, Power BI Desktop.

Việc hiểu rõ các khái niệm và nghiệp vụ liên quan đến bộ dữ liệu “car_prices” giúp tối ưu hóa quá trình quản lý và phân tích dữ liệu. Điều này giúp doanh nghiệp đưa ra các quyết định chiến lược dựa trên dữ liệu chính xác và hiệu quả. Bằng cách sử dụng SQL Server và Power BI, dữ liệu được xử lý, phân tích và trực quan hóa một cách hiệu quả, cung cấp cái nhìn toàn diện về thị trường xe ô tô cũ.

2.4.2 CÁC TRƯỜNG DỮ LIỆU

year: Năm sản xuất của xe.

make: Hãng sản xuất xe (ví dụ: Kia, BMW, Volvo).

model: Dòng xe (ví dụ: Sorento, 3 Series, S60).

trim: Phiên bản của dòng xe.

body: Loại xe (ví dụ: SUV, Sedan).

transmission: Loại hộp số (ví dụ: automatic).

vin: Mã số định danh của xe.

state: Bang nơi xe được bán.

condition: Tình trạng của xe, được đánh giá trên thang điểm (ví dụ: 5.0 là mới, 45.0 là cũ).

odometer: Số dặm xe đã đi.

color: Màu xe.

interior: Màu nội thất.

seller: Người bán hoặc công ty bán xe.

mmr: Giá trị thị trường của xe (Manheim Market Report).

sellingprice: Giá bán thực tế của xe.

saledate: Ngày bán xe.

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 NỀN TẢNG LUU TRỮ ĐÁM MÂY VÀ CÁC ỨNG DỤNG TẠI CHỖ

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 nền tảng đám mây (cloud-based solution)

Lý do:

- **Khả năng mở rộng:** Giải pháp đám mây cho phép mở rộng tài nguyên linh hoạt dựa trên nhu cầu phát triển của dự án mà không cần đầu tư thêm cơ sở hạ tầng.
- **Chi phí:** Chi phí ban đầu thấp hơn so với giải pháp tại chỗ, và doanh nghiệp chỉ cần trả phí cho những gì mình sử dụng.
- **Truy cập từ xa:** Đội ngũ phát triển và người dùng cuối có thể truy cập dữ liệu và các dịch vụ từ bất kỳ đâu, giúp tăng tính linh hoạt và hiệu quả công việc.
- **Quản lý và bảo trì:** Nhà cung cấp dịch vụ đám mây chịu trách nhiệm về bảo trì, nâng cấp và quản lý cơ sở hạ tầng, giảm bớt gánh nặng cho đội ngũ IT của doanh nghiệp.
- **Bảo mật và sao lưu:** Các nhà cung cấp dịch vụ đám mây thường có các giải pháp bảo mật, sao lưu và khôi phục dữ liệu toàn diện, giúp bảo vệ dữ liệu khỏi các mối đe dọa và đảm bảo tính liên tục của dịch vụ.

Sau khi xem xét các ưu nhược điểm của cả hai giải pháp, việc chọn giải pháp nền tảng đám mây cho dự án là quyết định hợp lý. Giải pháp đám mây cung cấp sự linh hoạt, khả năng mở rộng và quản lý dễ dàng, đồng thời giúp giảm chi phí ban đầu và tăng cường khả năng truy cập từ xa. Điều này sẽ giúp dự án triển khai nhanh chóng và hiệu quả hơn, đáp ứng nhu cầu phát triển và thay đổi liên tục của doanh nghiệp.

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. Nhận 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, chẳng hạn như hệ thống bán hàng xe ô tô của HappyCar.
- 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 Ẩn Phẩm: Tạo các ẩn 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_adddistributor @distributor = 'DistributorServer', @password = 'password';
```

Hình 3.1 Tạo cơ sở dữ liệu phân phối

```
-- Thiết lập máy chủ phân phối
EXEC sp_adddistributiondb @database = 'distribution',
    @data_folder = 'C:\Program Files\Microsoft SQL Server\MSSQL15.MSSQLSERVER\MSSQL\DATA',
    @log_folder = 'C:\Program Files\Microsoft SQL Server\MSSQL15.MSSQLSERVER\MSSQL\LOG';
```

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 = 'CarPricesPublication',
    @description = 'Transactional publication of car prices 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 án phẩm
]EXEC sp_addarticle @publication = 'CarPricesPublication',
    @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 án phẩm

```
-- Tạo đăng ký
]EXEC sp_addsubscription @publication = 'CarPricesPublication',
    @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

Nhân bản giao dịch là giải pháp phù hợp cho dự án lưu trữ dữ liệu của HappyCar do tính nhất quán, hiệu suất và tính sẵn sàng cao. Giải pháp này giúp đảm bảo rằng dữ liệu luôn cập nhật và sẵn sàng trên nhiều máy chủ, cải thiện hiệu suất truy cập và độ tin cậy của hệ thống. Việc triển khai nhân bản giao dịch trong SQL Server cũng đơn giản và dễ quản lý, giúp đảm bảo rằng hệ thống luôn hoạt động mượt mà và hiệu quả.

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 án phẩm (Publication): Tạo các án 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: Các trường dữ liệu bị bỏ trống hoặc giá trị không hợp lệ trong bảng dữ liệu.

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 được nhập nhiều lần với các ID khác nhau.

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

Ví dụ: Một xe có giá bán gấp 10 lần so với các xe khác trong cùng phân khúc.

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: Dữ liệu không được sao lưu đầy đủ dẫn đến mất mát dữ liệu quan trọng khi hệ thống gặp sự cố.

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 giá bán xe không được cập nhật đồng thời trên tất cả các máy chủ.

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: Cơ sở dữ liệu quá lớn không được tối ưu hóa, dẫn đến truy vấn mất nhiều thời gian.

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.

Dữ liệu thiếu hoặc không chính xác:

- Hiện trạng: Dữ liệu từ file CSV có thể chứa các giá trị thiếu hoặc không chính xác, đặc biệt là trong các trường như ngày bán, giá bán, và số km đã đi.
- Giải pháp: Thực hiện làm sạch dữ liệu kỹ lưỡng trước khi nhập vào cơ sở dữ liệu.

year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior	seller	mmr	sellingprice
2015	Ford	Fusion	SE	Sedan	automatic	3fa6p0hdxf...	ca	2	5559	white	beige	enterprise v...	15350	12000
2015	Kia	Sorento	LX	SUV	automatic	5xyktca66fg...	ca	5	14634	silver	black	kia motors ...	20600	21500
2014	Chevrolet	Cruze	2LT	Sedan	automatic	1g1pe5sbxe...	ca	NULL	15686	blue	black	avis rac/san ...	13900	10600
2015	Nissan	Altima	2.5 S	Sedan	automatic	1n4al3ap5fc...	ca	2	11398	black	black	enterprise v...	14750	14100
2015	Hyundai	Sonata	SE	Sedan	automatic	5npe24af4f...	ca	NULL	8311	red	GÇô	avis tra	15200	4200
2014	Audi	Q5	2.0T Premiu...	SUV	automatic	wa1lfafpxea...	ca	49	7983	white	black	audi north s...	37100	40000
2014	Chevrolet	Camaro	LS	Coupe	automatic	2g1fa1e39e...	ca	17	13441	black	black	wells fargo ...	17750	17000
2014	BMW	6 Series	650i	Convertible	automatic	wbayp9c53...	ca	34	8819	black	black	the hertz co...	68000	67200
2015	Chevrolet	Impala	LTZ	Sedan	automatic	2g1165z30f...	ca	19	14538	silver	black	enterprise v...	24300	7200
2014	BMW	5 Series	528i	Sedan	automatic	wba5a5c51...	ca	29	25969	black	black	financial ser...	34200	30000
2014	Chevrolet	Camaro	LT	Convertible	automatic	2g1fb3d31e...	ca	NULL	33450	black	black	avis rac/san ...	20100	14700
2015	Audi	A3	1.8 TFSI Pre...	Sedan	automatic	wauacgff7f...	ca	49	5826	gray	black	audi north s...	24000	23750
2014	BMW	6 Series	650i	Convertible	automatic	wbayp9c57...	ca	38	10736	black	black	the hertz co...	67000	65000
2015	Hyundai	Sonata	SE	Sedan	automatic	5npe24af4f...	ca	NULL	9281	silver	gray	enterprise v...	15150	8500
2015	Volvo	XC70	T6	Wagon	automatic	yv4902nb3f...	ca	42	16506	brown	brown	volvo na re...	32100	32500
2015	Volvo	XC70	T6	Wagon	automatic	yv4902nb3f...	ca	48	12725	beige	beige	volvo na re...	32300	32500
2014	BMW	X5	sDrive35i	SUV	automatic	Suxkr2c52e...	ca	NULL	11278	gray	black	avis rac/san ...	50400	34000
2014	Chevrolet	Camaro	LT	Coupe	automatic	2g1fb1e35e...	ca	42	11874	gray	black	midway hfc ...	22200	19500

Hình 3.7 Dữ liệu thiếu (null)

Dữ liệu trùng lặp:

- Hiện trạng: Có những dữ liệu trùng lặp xuất hiện trong bộ dữ liệu gốc
- Giải pháp: Thiết lập các quy tắc kiểm tra và loại bỏ dữ liệu trùng lặp trong quá trình nhập và làm sạch dữ liệu.

year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior	seller	mmr	sellingprice
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap7d...	ca	41	19957	GÇô	beige	nissan infini...	23300	21750
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6apxd...	ca	37	17337	white	black	nissan infini...	23600	20250
2013	Hyundai	Sonata	GLS	Sedan	automatic	5npeb4ac5...	ca	1	48510	white	tan	hertz/tra	11350	6500
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap8d...	ca	36	23104	blue	black	nissan infini...	22900	22750
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap3d...	ca	46	5017	white	beige	nissan infini...	25100	23500
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap5d...	ca	22	24054	gray	black	nissan infini...	22900	18750
► 2013	Hyundai	Sonata	GLS	Sedan	automatic	5npeb4ac5...	ca	37	36194	silver	gray	avis rac/san ...	12150	13200
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap3d...	ca	26	14264	white	tan	nissan infini...	23900	20750
2013	Hyundai	Sonata Hyb...	Limited	Sedan	automatic	kmhce4a1...	ca	47	8595	silver	gray	hyundai mo...	19100	18500
2013	Hyundai	Sonata Hyb...	Limited	Sedan	automatic	kmhct5ae4d...	ca	46	9927	black	black	hyundai mo...	10800	11300
2013	Hyundai	Accent	GS	Hatchback	automatic	kmhct5ae4d...	ca	46	9927	black	black	hyundai mo...	10800	11300
2013	Hyundai	Elantra	GLS	Sedan	automatic	5npdh4ae0...	ca	35	37759	beige	beige	avis corpora...	11400	11800
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap1d...	ca	43	17232	white	GÇô	nissan infini...	23600	24500
2013	Infiniti	FX	FX37	SUV	automatic	jn8cs1mw7...	ca	41	40087	gray	black	infiniti finan...	27700	27250
2013	Hyundai	Elantra	GLS	Sedan	automatic	5npdh4ae4...	ca	1	25960	gray	gray	enterprise v...	12100	8800
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap6d...	ca	35	25132	white	black	nissan infini...	22600	23500
* 2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap6d...	ca	35	25132	white	black	nissan infini...	22600	23500
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap6d...	ca	35	25132	white	black	nissan infini...	22600	23500
2013	Hyundai	Elantra	GLS	Sedan	automatic	5npdh4ae4...	ca	1	25960	gray	gray	enterprise v...	12100	8800
2013	Hyundai	Elantra	GLS	Sedan	automatic	5npdh4ae4...	ca	1	25960	gray	gray	enterprise v...	12100	8800
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap6d...	ca	41	19587	blue	gray	nissan infini...	23300	23500
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap1d...	ca	46	20246	gray	black	nissan infini...	23200	24750
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap7d...	ca	41	19957	GÇô	beige	nissan infini...	23300	21750
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6apxd...	ca	37	17337	white	black	nissan infini...	23600	20250
2013	Hyundai	Sonata	GLS	Sedan	automatic	5npeb4ac5...	ca	1	48510	white	tan	hertz/tra	11350	6500
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap8d...	ca	36	23104	blue	black	nissan infini...	22900	22750
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap3d...	ca	46	5017	white	beige	nissan infini...	25100	23500
2013	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap5d...	ca	22	24054	gray	black	nissan infini...	22900	18750
* NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

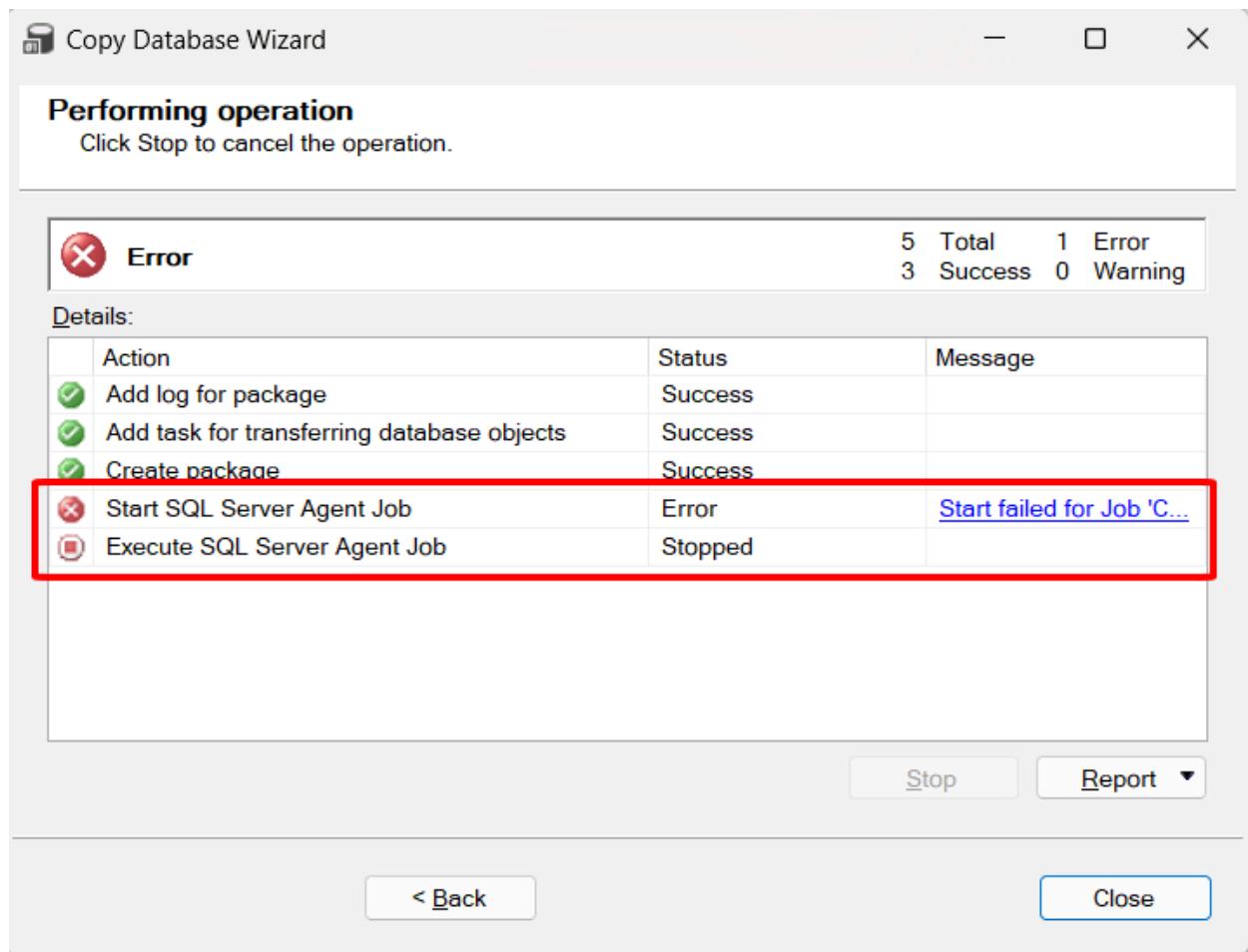
Hình 3.8 Dữ liệu trùng lặp

Dữ liệu ngoại lai (Outliers):

- Hiện trạng: Các giá trị ngoại lai có thể xuất hiện trong dữ liệu về giá bán hoặc số km đã đi.
- Giải pháp: Sử dụng các phương pháp thống kê để xác định và xử lý các giá trị ngoại lai.

Đồng bộ hóa dữ liệu:

- Hiện trạng: Quá trình nhân bản và đồng bộ hóa dữ liệu có thể gặp vấn đề, dẫn đến sự không nhất quán dữ liệu giữa các máy chủ.
- Giải pháp: Đảm bảo thiết lập và giám sát chặt chẽ quá trình nhân bản dữ liệu.



Hình 3.9 Nhân bản cơ sở dữ liệu bị lỗi

Vấn đề về hiệu suất:

- Hiện trạng: Truy vấn dữ liệu từ cơ sở dữ liệu lớn có thể chậm nếu không được tối ưu hóa.
- Giải pháp: Tối ưu hóa cấu trúc cơ sở dữ liệu và chỉ mục để cải thiện hiệu suất truy vấn.



Hình 3.10 Thời gian truy vấn lâu

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 bán xe phải khớp với giá thực tế khi xe được bán.

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 xe phải có một mã nhận dạng (VIN) duy nhất và không trùng lặp.

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 về xe phải có thông tin về năm sản xuất, hãng xe, mẫu xe, số km đã đi, giá bán, 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 ngày tháng phải nhất quán (ví dụ: YYYY-MM-DD), và các giá trị tiền tệ phải sử dụng cùng một đơn vị.

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 giá bán xe phải được cập nhật ngay khi giao dịch hoàn tất.

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 và các giá trị dữ liệu 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 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 bán, số km đã đi, và thông tin xe là chính xác để phân tích và đưa ra quyết định đúng đắn.
- Cách thực hiện: Kiểm tra và xác minh dữ liệu trước khi nhập vào cơ sở dữ liệu, sử dụng các quy tắc và ràng buộc để đảm bảo dữ liệu chính xác.

Tính toàn vẹn (Integrity):

- Lý do: Duy trì tính toàn vẹn của các quan hệ giữa các bảng dữ liệu để tránh lỗi và mâu thuẫn.
- Cách thực hiện: Sử dụng các khóa chính và khóa ngoại, và các ràng buộc toàn vẹn để đảm bảo dữ liệu không bị trùng lặp hoặc bị hỏng.

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.
- Cách thực hiện: Kiểm tra các giá trị NULL và các trường bị bỏ trống, yêu cầu người nhập dữ liệu cung cấp đầy đủ thông tin.

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, sử dụng các chuẩn mực cho các giá trị phân loại và tiền tệ.

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.
- Cách thực hiện: Thiết lập các quy trình cập nhật dữ liệu định kỳ và theo thời gian thực nếu có thể.

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.

Vìệc 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

BUỚC 1: XÁC ĐỊNH CÁC LỖI TRONG DỮ LIỆU

Xác định giá trị thiếu (Missing Values)

Cột bị ảnh hưởng: SaleDate, SellingPrice, Year, Make, Model, VIN, Odometer, Condition.

Thực hiện:

```
-- Tìm các giá trị NULL hoặc trống trong các cột quan trọng
SELECT * FROM CarPricesRaw
WHERE SaleDate IS NULL
    OR SellingPrice IS NULL
    OR Year IS NULL
    OR Make IS NULL
    OR Model IS NULL
    OR VIN IS NULL
    OR Odometer IS NULL
    OR Condition IS NULL
```

Results

year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior	seller	mmr	sellingprice	saledate
1 2012	Ram	2500	SLT	Crew Cab	automatic	3c6fd5jf0cq269310	tx	NULL	83941	white	GÇø	sun coast fleet services	21200	19100	Wed Dec 17 2014 10:00:00 GMT-0800
2 2012	Subaru	Forester	2.5X Premium	SUV	automatic	j2ehbbdc4ch413029	fl	NULL	17087	white	gray	mini of wesley chapel	19550	20000	Thu Dec 18 2014 09:45:00 GMT-0800 (P)
3 2012	Scion	xB	Base	Wagon	automatic	jt1ze4fe7cqj005931	nj	NULL	37125	NULL	NULL	advanced auto brokers llc	11350	10000	Wed Dec 17 2014 09:30:00 GMT-0800
4 2012	Ram	1500	Sport	Quad Cab	automatic	1c6fd7h8cc115125	pa	NULL	41215	black	GÇø	twin willows ii inc	24900	23800	Wed Dec 17 2014 09:30:00 GMT-0800
5 2012	Scion	iQ	Base	Hatchback	automatic	jnxjd801cq20632	fl	NULL	29642	silver	black	the auto site inc	8825	8500	Fri Dec 19 2014 09:25:00 GMT-0800 (P)
6 2012	Ram	1500	Laramie Longhorn Edition	Crew Cab	NULL	1c6fd7p7qz179237	pa	NULL	99388	black	GÇø	healey ford lincoln mercury llc	30400	23500	Fri Dec 19 2014 09:30:00 GMT-0800 (P)
7 2012	Ram	1500	ST	Crew Cab	automatic	1c6fd6f6p2cs22215	tx	NULL	63276	burgundy	gray	pro financial inc	14350	14600	Wed Dec 17 2014 10:20:00 GMT-0800
8 2012	Toyota	4Runner	Limited	SUV	automatic	jezu5ir5e5049141	fl	NULL	50554	silver	black	williamson cadillac company	25300	21200	Fri Dec 19 2014 09:35:00 GMT-0800 (P)
9 2012	Ram	3500	Laramie	Crew Cab	automatic	3c633de9cg140416	tx	NULL	24048	black	tan	fairway ford henderson	41700	41500	Fri Dec 19 2014 11:00:00 GMT-0800 (P)
10 2012	Ram	1500	ST	Quad Cab	automatic	1c6fd6f6xcs149071	az	NULL	112458	red	gray	rolit motors incorporated	12200	9900	Thu Dec 18 2014 11:00:00 GMT-0800 (P)
11 2012	Scion	tC	Base	Hatchback	NULL	jkf65e74z3025195	nj	NULL	43237	gray	GÇø	kennya auto group inc	12650	10100	Wed Dec 17 2014 09:30:00 GMT-0800
12 2012	Toyota	4Runner	Limited	SUV	automatic	jlebu5j9e5095894	ny	NULL	22722	blue	tan	remarketing by gelmanheim albany	32300	29700	Thu Dec 18 2014 09:15:00 GMT-0800 (P)
13 2012	smart	fortwo	passion coupe	Hatchback	automatic	wmee33ba4ck525659	az	NULL	21487	blue	gray	dealer network group auto wholesale	6925	7300	Thu Dec 18 2014 11:00:00 GMT-0800 (P)
14 2012	Ram	2500	SLT	Crew Cab	automatic	3c6ud5d6fg230009	tx	NULL	65680	black	black	pro financial inc	27300	28500	Wed Dec 17 2014 10:20:00 GMT-0800
15 2012	Subaru	Legacy	2.5 Limited PZEV	Sedan	automatic	4s3bmbd64c3019602	pa	NULL	23875	white	GÇø	castle car company	15550	18400	Wed Dec 17 2014 09:30:00 GMT-0800
... 2013

Hình 3.11 Xác định giá trị null

Xác định giá trị không hợp lệ (Invalid Values)

Cột bị ảnh hưởng: Year(Năm sản xuất phải lớn hơn hoặc bằng 1986), Odometer (Số km đã đi phải lớn hơn hoặc bằng 0).

Thực hiện:

```
-- Tìm các giá trị không hợp lệ trong cột Year và Odometer
SELECT * FROM CarPricesRaw
WHERE Year < 1886
    OR Odometer < 0
```

Hình 3.12 Xác định giá trị không hợp lệ

Xác định giá trị trùng lặp (Duplicate Data)

Cột bị ảnh hưởng: VIN.

Thực hiện:

-- Tìm các bản ghi trùng lặp dựa trên VIN

```
SELECT VIN, COUNT(*)
FROM CarPricesRaw
GROUP BY VIN
HAVING COUNT(*) > 1
```

Results

VIN	(No column name)
19uuua66234a054081	2
19uuua76568a039063	2
19uuua8f58ba004103	2
19uya424x1a019370	2
19vde1f76de000328	2
19xfb2f9xce030379	2
1a8hw58nx7f564270	2
1b3cb4ha6ad660955	2
1b3cb5haxbd256984	2
1b3cc1fb0an232400	2
1b3cc5fb0an234929	2
1b3es56c85d258170	2
1b3hb48a59d143132	2
1b3hb48b07d254631	2
1c3an65l85x048764	2
1c3bc2eg4bn589291	2
1e2aa4fb0-n152026	2

Query executed successfully.

Hình 3.13 Xác định bản ghi trùng lặp

Xác định giá trị ngoại lai (Outliers)

Cột bị ảnh hưởng: SellingPrice, Odometer.

Thực hiện:

```
-- Tìm các giá trị ngoại lai trong cột SellingPrice và Odometer
SELECT * FROM CarPricesRaw
WHERE SellingPrice > (SELECT AVG(SellingPrice) + 3 * STDEV(SellingPrice) FROM CarPricesRaw)
OR Odometer > (SELECT AVG(Odometer) + 3 * STDEV(Odometer) FROM CarPricesRaw)
```

year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior	seller	mmr	sellingprice	saledate	
1	2005	Jeep	Liberty	Sport	SUV	1j4gj48k25w579171	ny	NULL	253531	black	gray	abs	1150	1100	Thu Dec 18 2014 09:45:00 GMT-4	
2	2015	Cadillac	Escalade	Luxury	SUV	1gys3bkx4f203526	ca	38	10690	red	beige	aero sweet company	71000	64000	Thu Feb 26 2015 04:30:00 GMT-4	
3	1995	Lexus	LS 400	Base	Sedan	automatic	j8ud22e0011926	pa	3	285271	gray	beige	victory auto group llc	1175	900	Tue Mar 03 2015 05:01:00 GMT-4
4	2014	NULL	NULL	NULL	NULL	automatic	wd3pfldcd4e5888106	fl	48	1796	white	beige	mercedes-benz usa	84000	88000	Wed Jun 17 2015 02:30:00 GMT-4
5	2015	Chevrolet	Tahoe	LT	SUV	NULL	1gnmskkc3f115256	pa	45	17014	GQo	black	r hollenshead auto sales inc	45500	45300	Fri Jun 19 2015 02:00:00 GMT-7
6	2007	Mercure	Mountaineer	Premier	SUV	automatic	4m2eu48e87uj03374	mi	19	395827	silver	black	georges used car sales inc	975	2400	Thu Feb 19 2015 01:30:00 GMT-4
7	1998	Mercedes-Benz	M-Class	ML320	SUV	automatic	4jgb554e5wa016744	va	NULL	298253	green	beige	r k chevrolet inc	575	325	Thu Dec 18 2014 11:50:00 GMT-4
8	2013	Ford	F-150	SVT Raptor	SuperCrew	automatic	1ftv1t66fa09536	ga	42	35170	black	black	ford motor credit company llc	46400	44500	Tue Jan 06 2015 01:30:00 GMT-4
9	2015	Chevrolet	Suburban	LT	SUV	automatic	1gnmskkc2d147335	nj	42	28511	black	black	enterprise veh exchange/rental	46000	46250	Wed Jan 21 2015 01:30:00 GMT-4
10	2002	Toyota	Camry	SE V6	Sedan	automatic	41hb304042028123	nc	19	308964	white	beige	crosroads chrysler jeep dodge	1600	2600	Mon Feb 02 2015 01:00:00 GMT-4
11	2012	BMW	6 Series	650i xDrive	Convertible	automatic	wbaiz5c57cd519766	nv	48	49952	black	GQo	financial services remarketing (lease)	42000	44250	Thu Feb 19 2015 14:00:00 GMT-4
12	2012	Infiniti	G Sedan	G37 Journey	G Sedan	automatic	jn1cv6ap8cm624299	ca	39	32011	white	gray	nissan infiniti it	20100	44000	Thu Mar 26 2015 05:30:00 GMT-4
13	1998	Toyota	Corolla	VE	Sedan	automatic	1nxbr12exwz096195	ga	19	231398	black	gray	nalley nissan decatur	650	600	Tue Dec 23 2014 13:00:00 GMT-4
14	2014	BMW	6 Series Gran Coupe	640i xDrive	Sedan	automatic	wba6b8c52edz72495	oh	39	2622	white	black	bmw north america raa	67000	62000	Tue Dec 30 2014 14:00:00 GMT-4
15	2008	Chevrolet	Silverado 1500	Work Truck	Regular Cab	automatic	1gcek14c18z127146	fl	24	263917	white	gray	ari	2325	3000	Wed Jan 14 2015 07:10:00 GMT-4
16	2012	BMW	6 Series	650i	Convertible	automatic	wbaiz5c57cd519766	nv	48	100000	white	black	bmw north america raa	47000	47760	Thu Feb 22 2015 05:20:00 GMT-4

Hình 3.14 Xác định giá trị ngoại lai

BUỚC 2: SỬA CHỮA CÁC LỖI TRONG DỮ LIỆU

Sửa chữa dữ liệu thiếu (Missing Values)

Cột bị ảnh hưởng: SaleDate, SellingPrice, Year, Make, Model, VIN, Odometer, Condition.

Thực hiện:

```
-- Điền giá trị mặc định cho các trường bị thiếu hoặc loại bỏ các bản ghi bị thiếu dữ liệu
UPDATE CarPricesRaw
SET SaleDate = '2020-01-01'
WHERE SaleDate IS NULL

UPDATE CarPricesRaw
SET SellingPrice = (SELECT AVG(SellingPrice) FROM CarPricesRaw)
WHERE SellingPrice IS NULL

UPDATE CarPricesRaw
SET Year = 2020
WHERE Year IS NULL

0 % ▶ Messages
Warning: Null value is eliminated by an aggregate or other SET operation.

(12 rows affected)

(0 rows affected)

Completion time: 2024-07-25T19:25:57.5250252+07:00
```

Hình 3.15 Điền giá trị thiếu

```
-- Các trường bắt buộc không thể thiếu, nên loại bỏ các bản ghi thiếu các trường này
DELETE FROM CarPricesRaw
WHERE Make IS NULL
    OR Model IS NULL
    OR VIN IS NULL
    OR Odometer IS NULL
    OR Condition IS NULL

170 % ▶ Messages
(22158 rows affected)

Completion time: 2024-07-25T19:28:46.6329915+07:00
```

Hình 3.16 Loại bỏ bản ghi có các trường bắt buộc không thể thiếu

Sửa chữa giá trị không hợp lệ (Invalid Values)

Cột bị ảnh hưởng: Year, Odometer.

Thực hiện:

```
-- Sửa chữa giá trị không hợp lệ
[UPDATE CarPricesRaw
SET Year = 2020
WHERE Year < 1886

[UPDATE CarPricesRaw
SET Odometer = 0
WHERE Odometer < 0]
```

Hình 3.17 Sửa chữa giá trị không hợp lệ

Xử lý dữ liệu trùng lặp (Duplicate Data)

Cột bị ảnh hưởng: VIN.

Thực hiện:

```
-- Loại bỏ các bản ghi trùng lặp dựa trên VIN
[WITH CTE AS (
    SELECT *,
        ROW_NUMBER() OVER (PARTITION BY VIN ORDER BY VIN) AS row_num
    FROM CarPricesRaw
)
DELETE FROM CTE WHERE row_num > 1]

% ▶ Messages
(8042 rows affected)

Completion time: 2024-07-25T19:33:35.6362923+07:00
```

Hình 3.18 Loại bỏ các bản ghi trùng lặp

Xử lý các giá trị ngoại lai (Outliers)

Cột bị ảnh hưởng: SellingPrice, Odometer.

Thực hiện:

```
-- Loại bỏ các giá trị ngoại lai trong cột SellingPrice và Odometer
DELETE FROM CarPricesRaw
WHERE SellingPrice > (SELECT AVG(SellingPrice) + 3 * STDEV(SellingPrice) FROM CarPricesRaw)
    OR Odometer > (SELECT AVG(Odometer) + 3 * STDEV(Odometer) FROM CarPricesRaw)
```

(11830 rows affected)

Completion time: 2024-07-25T19:39:11.8904519+07:00

Hình 3.19 Loại bỏ các giá trị ngoại lai

BUỚC 3: XÁC MINH VÀ ĐẢM BẢO CHẤT LƯỢNG DỮ LIỆU

Kiểm tra lại dữ liệu

```
-- Kiểm tra lại dữ liệu sau khi làm sạch
SELECT * FROM CarPricesRaw
WHERE SaleDate IS NULL
    OR SellingPrice IS NULL
    OR Year < 1886
    OR Make IS NULL
    OR Model IS NULL
    OR VIN IS NULL
    OR Odometer < 0
    OR Condition IS NULL
```

Results Messages

year	make	model	trim	body	transmission	vin	state	condition	odometer	color	interior	seller	mmr	sellingprice	saledate

Hình 3.20 Kiểm tra lại dữ liệu

Đảm bảo nhất quán dữ liệu

```
-- Đặt các ràng buộc để duy trì tính nhất quán của dữ liệu
ALTER TABLE CarPricesRaw
ADD CONSTRAINT chk_Year CHECK (Year >= 1886)

ALTER TABLE CarPricesRaw
ADD CONSTRAINT chk_Odometer CHECK (Odometer >= 0)
```

Commands completed successfully.

Completion time: 2024-07-25T19:44:35.9286123+07:00

Hình 3.21 Đặt các ràng buộc

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 dữ liệu từ dạng văn bản (string) sang số (integer, float) hoặc ngược lại.
- Chuyển đổi định dạng ngày tháng từ dạng chuỗi sang kiểu ngày tháng.

Mã hóa lại các giá trị phân loại:

- Chuyển đổi các giá trị phân loại (categorical) thành các giá trị số để thuận tiện cho việc phân tích và mô hình hóa. Ví dụ: chuyển đổi giới tính từ "Nam" và "Nữ" thành 0 và 1.

Chuẩn hóa dữ liệu:

- Chuẩn hóa các giá trị dữ liệu để chúng nằm trong một phạm vi nhất định, chẳng hạn như [0, 1] hoặc [-1, 1]. Việc này giúp cải thiện hiệu suất của các mô hình máy học.
- Chuẩn hóa các giá trị ngày tháng, ví dụ như chuyển đổi tất cả các ngày tháng về một định dạng duy nhất.

Xử lý dữ liệu thời gian:

- Tách các thành phần của ngày tháng (ngày, tháng, năm) thành các cột riêng biệt để dễ dàng phân tích.
- Tạo ra các biến thời gian mới như tuần, quý, hoặc khoảng thời gian từ một mốc thời gian nhất định.

Tạo các biến mới từ biến hiện có:

- Tính toán các biến mới từ các biến hiện có, chẳng hạn như tính BMI từ chiều cao và cân nặng, hoặc tính tỷ lệ giữa các biến.

Xử lý giá trị trống hoặc giá trị mặc định:

- Thay thế các giá trị trống hoặc giá trị mặc định bằng các giá trị có ý nghĩa hơn, chẳng hạn như giá trị trung bình, giá trị phổ biến, hoặc một giá trị suy diễn từ các dữ liệu khác.

Gộp hoặc chia cột dữ liệu:

- Gộp các cột dữ liệu lại với nhau để tạo thành một cột duy nhất, chẳng hạn như gộp họ và tên thành một cột họ và tên đầy đủ.
- Chia một cột thành nhiều cột khác nhau, chẳng hạn như tách địa chỉ thành các thành phần riêng biệt (số nhà, đường, quận, thành phố).

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

- Loại bỏ các ký tự đặc biệt, khoảng trắng thừa, hoặc chuyển đổi các ký tự đặc biệt thành dạng chuẩn.
- Loại bỏ khoảng trắng thừa ở đầu và cuối chuỗi.
- Việc chuyển đổi dữ liệu giúp đảm bảo rằng dữ liệu được đồng nhất và dễ dàng sử dụng cho các bước phân tích và mô hình hóa tiếp theo.

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 giữa các kiểu số như từ nguyên (integer) sang thập phân (float) hoặc ngược lại. Điều này thường cần thiết khi thực hiện các phép tính mà độ chính xác là yếu tố quan trọng.
- **Chuyển đổi kiểu chuỗi (String Conversion):** Chuyển đổi dữ liệu từ các kiểu khác sang chuỗi hoặc từ chuỗi sang các kiểu dữ liệu khác. Ví dụ, chuyển đổi một số nhận dạng (ID) từ số nguyên sang chuỗi để dễ dàng ghép nối với các chuỗi khác.
- **Chuyển đổi kiểu ngày tháng (Date Conversion):** Chuyển đổi chuỗi đại diện cho ngày tháng sang kiểu ngày tháng để dễ dàng thực hiện các phép toán liên quan đến thời gian như tính khoảng cách thời gian, sắp xếp theo ngày, hoặc phân tích xu hướng theo thời gian.
- **Chuyển đổi kiểu phân loại (Categorical Conversion):** Chuyển đổi các giá trị phân loại thành các giá trị số để thuận tiện cho việc phân tích thống kê và mô hình hóa dữ liệu. Ví dụ, chuyển đổi các giá trị "Nam" và "Nữ" thành 0 và 1.

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 tuổi của xe từ năm sản xuất đến năm hiện tại.
- Tính toán giá trị còn lại của xe dựa trên giá trị ban đầu và số năm sử dụng.

Chuyển Đổi Đơn Vị Đo Lường

- Chuyển đổi đơn vị km sang dặm.

3.3.3 TRÌNH BÀY CÁC PHÉP CHUYỂN ĐỔI TRONG DỰ ÁN

CHUYỂN ĐỔI KIỂU DỮ LIỆU

Chuyển đổi ngày tháng từ chuỗi ký tự sang kiểu datetime.

```
-- Chuyển đổi ngày tháng từ chuỗi ký tự sang kiểu datetime.  
UPDATE CarPricesRaw  
SET SaleDate = TRY_CONVERT(DATETIME, SaleDate)  
  
% ▶ Messages  
(516819 rows affected)  
Completion time: 2024-07-25T20:26:34.9659855+07:00
```

Hình 3.22 Chuyển đổi ngày tháng sang kiểu datetime

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

```
-- Chuyển đổi các giá trị số từ chuỗi ký tự sang kiểu số nguyên hoặc số thực.  
UPDATE CarPricesRaw  
SET Odometer = TRY_CONVERT(INT, Odometer)  
  
% ▶ Messages  
(516819 rows affected)  
Completion time: 2024-07-25T20:29:17.4581504+07:00
```

Hình 3.23 Chuyển đổi giá trị số

CHUYỂN ĐỔI ĐỊNH DẠNG DỮ LIỆU

Định dạng ngày tháng để đảm bảo thống nhất.

```
-- Định dạng ngày tháng để đảm bảo thống nhất.  
UPDATE CarPricesRaw  
SET SaleDate = FORMAT(SaleDate, 'yyyy-MM-dd')
```

Hình 3.24 Định dạng ngày tháng

Định dạng số tiền để đảm bảo thống nhất.

```
-- Định dạng số tiền để đảm bảo thống nhất.  
UPDATE CarPricesRaw  
SET SellingPrice = FORMAT(SellingPrice, 'C', 'en-US')
```

Hình 3.25 Định dạng số tiền

CHUYỂN ĐỔI GIÁ TRỊ PHÂN LOẠI

Chuyển đổi tất cả các giá trị phân loại về chữ thường hoặc chữ hoa.

```
-- Chuyển đổi tất cả các giá trị phân loại về chữ thường hoặc chữ hoa.  
UPDATE CarPricesRaw  
SET Make = UPPER(Make)
```

(516819 rows affected)

Completion time: 2024-07-25T20:39:47.1054305+07:00

Hình 3.26 Chuyển đổi về dạng chữ thường

Thay thế các giá trị phân loại không nhất quán hoặc không chính xác.

```
-- Thay thế các giá trị phân loại không nhất quán hoặc không chính xác.  
UPDATE CarPricesRaw  
SET Transmission = 'Automatic'  
WHERE Transmission = 'Auto'
```

Hình 3.27 Thay thế giá trị không nhất quán/ không chính xác

TÍNH TOÁN VÀ CHUYỂN ĐỔI CÁC CHỈ SỐ

Tính toán tuổi của xe từ năm sản xuất đến năm hiện tại.

```
-- Tính toán tuổi của xe từ năm sản xuất đến năm hiện tại.  
ALTER TABLE CarPricesRaw ADD CarAge INT  
UPDATE CarPricesRaw  
SET CarAge = YEAR(GETDATE()) - Year
```

(516819 rows affected)

Completion time: 2024-07-25T20:44:27.8731900+07:00

Hình 3.28 Tính tuổi của xe

Tính toán giá trị còn lại của xe dựa trên giá trị ban đầu và số năm sử dụng.

```
-- Tính toán giá trị còn lại của xe dựa trên giá trị ban đầu và số năm sử dụng.  
ALTER TABLE CarPricesRaw ADD ResidualValue DECIMAL(18, 2)  
UPDATE CarPricesRaw  
SET ResidualValue = MMR - (MMR * 0.05 * CarAge)
```

(516819 rows affected)

Completion time: 2024-07-25T20:45:20.8749591+07:00

Hình 3.29 Tính giá trị còn lại của xe

CHUYỂN ĐỔI ĐƠN VỊ ĐO LƯỜNG

Chuyển đổi đơn vị km sang dặm.

```
-- Chuyển đổi đơn vị km sang dặm.  
UPDATE CarPricesRaw  
SET Odometer = Odometer * 0.621371
```

(516819 rows affected)

Completion time: 2024-07-25T20:47:49.4550025+07:00

Hình 3.30 Chuyển đơn vị từ km sang dặm

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. 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 SaleDate và SellingPrice 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 Year và Odometer được xác định và sửa chữa để đảm bảo rằng dữ liệu phản ánh đúng thực tế. Các bản ghi trùng lặp dựa trên VIN được loại bỏ để duy trì tính duy nhất của dữ liệu. Để loại bỏ các giá trị ngoại lai, các giá trị ngoài phạm vi trong cột SellingPrice và Odometer được loại bỏ. Sau khi làm sạch, các cột dữ liệu được chuyển đổi về kiểu dữ liệu

phù hợp như chuyển đổi SaleDate về kiểu Date và Odometer về kiểu số nguyên. Các giá trị phân loại trong cột Make và Transmission được chuyển đổi để đảm bảo tính nhất quán, như chuyển đổi về chữ hoa hoặc thay thế các giá trị không nhất quán. Cuối cùng, các chỉ số mới như tuổi của xe và giá trị còn lại của xe được tính toán và thêm vào bộ dữ liệu. Sau khi hoàn tất các bước chuẩn hóa, dữ liệu được lưu và tải lại vào Power BI để tiếp tục phân tích.

The screenshot shows the Microsoft Power BI Data Editor interface. On the left, there's a navigation pane with 'Queries [4]' containing 'Car', 'CarTransaction', 'Condition', and 'Seller'. The main area displays a table with columns: car_id, condition_id, seller_id, and a timestamp column labeled '1.2_mmr'. A context menu is open over the '1.2_mmr' column, specifically at the row for 'saledate'. The menu path 'saledate' -> 'Change Type' -> 'Decimal Number' is highlighted. Other options in the 'Change Type' submenu include 'Text', 'Binary', and 'Date/Time'. To the right of the table, the 'Properties' pane shows the 'Name' is 'CarTransaction' and the 'Applied Steps' pane shows 'Changed Type' under 'Source'. The status bar at the bottom indicates 'PREVIEW DOWNLOADED ON THURSDAY, JULY 11, 2024' and shows system icons like battery level, signal strength, and time.

Hình 4.1 Chuyển đổi kiểu cho giá trị số

Detailed description: This screenshot shows the Power BI Desktop interface. A query named 'Table.SelectRows(#"Changed Type", each true)' is selected. The table contains columns: car_id, condition_id, seller_id, mmr, sellingprice, and saledate. A context menu is open over the 'saledate' column, with 'Duplicate Column' highlighted. The 'APPLIED STEPS' pane on the right shows the steps taken: 'Changed Type' and 'Filtered Rows'.

Hình 4.2 Duplicate cột [saledate] thành [saledate-Copy]

Detailed description: This screenshot shows the Power BI Desktop interface. A query named 'Table.DuplicateColumn(#"Filtered Rows", "saledate", "saledate - Copy")' is selected. A dialog box titled 'Split Column by Delimiter' is open, asking to specify a delimiter. The 'OK' button is highlighted. The 'APPLIED STEPS' pane on the right shows the step 'Duplicated Column'.

Hình 4.3 Tách cột saledate-copy thành Thứ, Ngày, Tháng, Năm, Giờ

Hình 4.4 Xóa các cột không cần thiết

	Merged	Date	Time
GMT-0800 (PST)	12/18/2014		10:00:00 AM
GMT-0800 (PST)	12/18/2014		9:15:00 AM
GMT-0800 (PST)	12/23/2014		1:30:00 PM
GMT-0800 (PST)	12/23/2014		11:45:00 AM
GMT-0800 (PST)	1/12/2015		9:00:00 AM
GMT-0800 (PST)	12/23/2014		10:00:00 AM
GMT-0800 (PST)	12/17/2014		10:00:00 AM
GMT-0800 (PST)	12/18/2014		10:00:00 AM
GMT-0800 (PST)	12/18/2014		11:45:00 AM

*Hình 4.5 Gộp các cột ngày, tháng, năm và chuyển kiểu dữ liệu về kiểu “Date”
chuyển kiểu dữ liệu của cột Time về kiểu “Time”*

The screenshot shows the Power BI Data Editor interface with the following details:

- File Bar:** Du_An_Mau, File, Close & Apply, Apply, Close, Save, Save As, Options and settings, Help.
- Toolbar:** Unpivot Columns, Move, Split Column, Format, Parse, Text Column, Statistics, Standard Scientific Information, Number Column, Date & Time Column, R Py, Run R, Run Python script, Scripts.
- Query Settings:** Name: CarTransaction, All Properties.
- Applied Steps:** Source, Navigation, Removed Columns, Changed Type, Filtered Rows, Duplicated Column, Split Column by Delimiter, Changed Type1, Removed Columns1, Renamed Columns, Removed Columns2, Merged Columns, Changed Type2, Renamed Columns1.
- Data View:** A table with columns: condition_id, seller_id, mmr, sellingprice, saledate. The data shows various transactions with dates like "Mon Jan 12 2015 09:00:00" and "Tue Dec 23 2014 10:00:00".
- Bottom Status:** 9 COLUMNS, 999+ ROWS, Column profiling based on top 1000 rows, PREVIEW DOWNLOADED AT 9:27 PM, 7/25/2024, 9:30 PM.

Hình 4.6 Close and Apply

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 phạm vi dự án này, mô hình vật lý đang được áp dụng cho bộ dữ liệu. Mô hình logic giúp xác định cấu trúc bảng, cột, và các ràng buộc trong cơ sở dữ liệu mà không phụ thuộc vào hệ quản trị cụ thể. Điều này giúp chuẩn hóa dữ liệu và đảm bảo tính nhất quán, toàn vẹn, dễ dàng trong việc mở rộng và bảo trì hệ thống sau này. Mô hình logic cho phép chúng ta xác định chính xác các trường dữ liệu cần thiết, các mối quan hệ giữa các thực thể (như các bảng về thông tin xe, thông tin người bán, và thông tin giao dịch) và các ràng buộc đảm bảo dữ liệu chính xác và không bị trùng lặp. Sau khi mô hình logic được xác định rõ ràng, chúng ta có thể chuyển đổi sang mô hình vật lý để triển khai trên hệ quản trị cơ sở dữ liệu SQL Server.

Mô hình vật lý là giai đoạn cuối cùng trong quá trình mô hình hóa dữ liệu, nơi mô hình logic được cụ thể hóa thành cấu trúc lưu trữ thực tế trong hệ quản trị cơ sở dữ liệu (DBMS). Đây là bước chuyển đổi các bảng, mối quan hệ, và các ràng buộc đã được xác định trong mô hình logic thành các câu lệnh SQL cụ thể để tạo ra cơ sở dữ liệu thực sự. Trong dự án này, chúng ta bắt đầu bằng việc tạo cơ sở dữ liệu

Phan_tich_du_lieu_ban_xe, sau đó tạo các bảng Car, Condition, Seller, và Transaction để lưu trữ thông tin về xe, tình trạng xe, người bán và giao dịch. Các mối quan hệ giữa các bảng được thiết lập bằng cách sử dụng các khóa ngoại, đảm bảo tính toàn vẹn và liên kết của dữ liệu. Chỉ mục được tạo ra trên các cột quan trọng để cải thiện hiệu suất truy vấn, và các ràng buộc như kiểm tra và duy nhất được áp dụng để đảm bảo dữ liệu chính xác và nhất quán. Cuối cùng, dữ liệu mẫu được nhập vào các bảng để kiểm tra và xác minh mô hình. Mô hình vật lý giúp chuyển đổi mô hình logic thành cơ sở dữ liệu hoạt động, hỗ trợ tốt cho việc lưu trữ, quản lý và truy xuất dữ liệu trong các ứng dụng thực tế.

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.
- Ví dụ: Các bảng và mối quan hệ giữa chúng phải phản ánh đúng các thực thể và mối quan hệ trong 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: Các ràng buộc toàn vẹn đảm bảo rằng không có dữ liệu trùng lặp hoặc mâu thuẫn.

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 và đơn vị đo lường phải nhất quán trong toàn bộ cơ sở dữ liệu.

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: Tất cả các thông tin cần thiết về xe, tình trạng xe, người bán và giao dịch phải được bao gồm 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: Các chỉ mục và các thiết kế bảng tối ưu để hỗ trợ các truy vấn thường xuyên.

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 đúng các thực thể và mối quan hệ trong dữ liệu xe ô tô, người bán và giao dịch. Các bảng Car, Condition, Seller, và Transaction được thiết kế để lưu trữ thông tin một cách chính xác và đầy đủ.

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 về xe, tình trạng xe, người bán và giao dịch. Điều này đảm bảo rằng tất cả các yêu cầu nghiệp vụ đều được đáp ứng.

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.

BUỚC 1: PHÂN TÍCH DỮ LIỆU THÔ

Mục đích: Hiểu rõ cấu trúc và nội dung của dữ liệu hiện có để xác định các bảng cần thiết.

Nhìn vào dữ liệu thô, chúng ta thấy rằng một số thông tin bị lặp lại nhiều lần. Chẳng hạn, thông tin về hãng sản xuất (make), mẫu xe (model), phiên bản (trim), kiểu thân xe (body), loại hộp số (transmission), màu sắc (color), và nội thất (interior) có thể bị lặp lại cho nhiều xe khác nhau. Tương tự, thông tin về người bán (seller) và tình trạng xe (condition, odometer, state) cũng có thể bị lặp lại.

BUỚC 2: XÁC ĐỊNH CÁC BẢNG CẦN THIẾT

Mục đích: Chia dữ liệu thành các bảng riêng biệt để loại bỏ sự trùng lặp và đảm bảo tính toàn vẹn dữ liệu theo chuẩn 3NF.

Các bảng chính cần thiết:

- Bảng Car: Chứa thông tin chung về xe, không bao gồm các thông tin thay đổi theo giao dịch.

- Bảng Condition: Chứa thông tin về tình trạng của xe tại thời điểm giao dịch.
- Bảng Seller: Chứa thông tin về người bán.
- Bảng Transaction: Chứa thông tin về giao dịch mua bán, liên kết với các bảng trên thông qua khóa ngoại.

BUỚC 3: XÁC ĐỊNH CHI TIẾT CÁC BẢNG VÀ CÁC TRƯỜNG

Bảng Car:

- car_id: Khóa chính tự tăng
- make: Hãng sản xuất
- model: Mẫu xe
- trim: Phiên bản xe
- body: Kiểu thân xe
- transmission: Loại hộp số
- color: Màu sắc
- interior: Nội thất
- vin: Số VIN (khóa duy nhất)
- year: Năm sản xuất

Bảng Condition:

- condition_id: Khóa chính tự tăng
- state: Bang
- condition: Tình trạng xe
- odometer: Quãng đường đã đi
- Bảng Seller:
- seller_id: Khóa chính tự tăng
- seller_name: Tên người bán

Bảng Transaction:

- transaction_id: Khóa chính tự tăng
- car_id: Khóa ngoại tham chiếu đến Car
- condition_id: Khóa ngoại tham chiếu đến Condition
- seller_id: Khóa ngoại tham chiếu đến Seller
- mmr: Giá MMR
- sellingprice: Giá bán
- saledate: Ngày bán

BUỚC 4: THIẾT LẬP MỐI QUAN HỆ GIỮA CÁC BẢNG

Mục đích: Đảm bảo rằng các bảng có thể liên kết với nhau một cách logic và đảm bảo tính toàn vẹn dữ liệu.

Mối quan hệ giữa các bảng:

- Car có quan hệ một-nhiều với Transaction (một xe có thể có nhiều giao dịch).
- Condition có quan hệ một-nhiều với Transaction (một tình trạng xe có thể áp dụng cho nhiều giao dịch).
- Seller có quan hệ một-nhiều với Transaction (một người bán có thể bán nhiều xe).

BUỚC 5: TRIỂN KHAI MÔ HÌNH VẬT LÝ

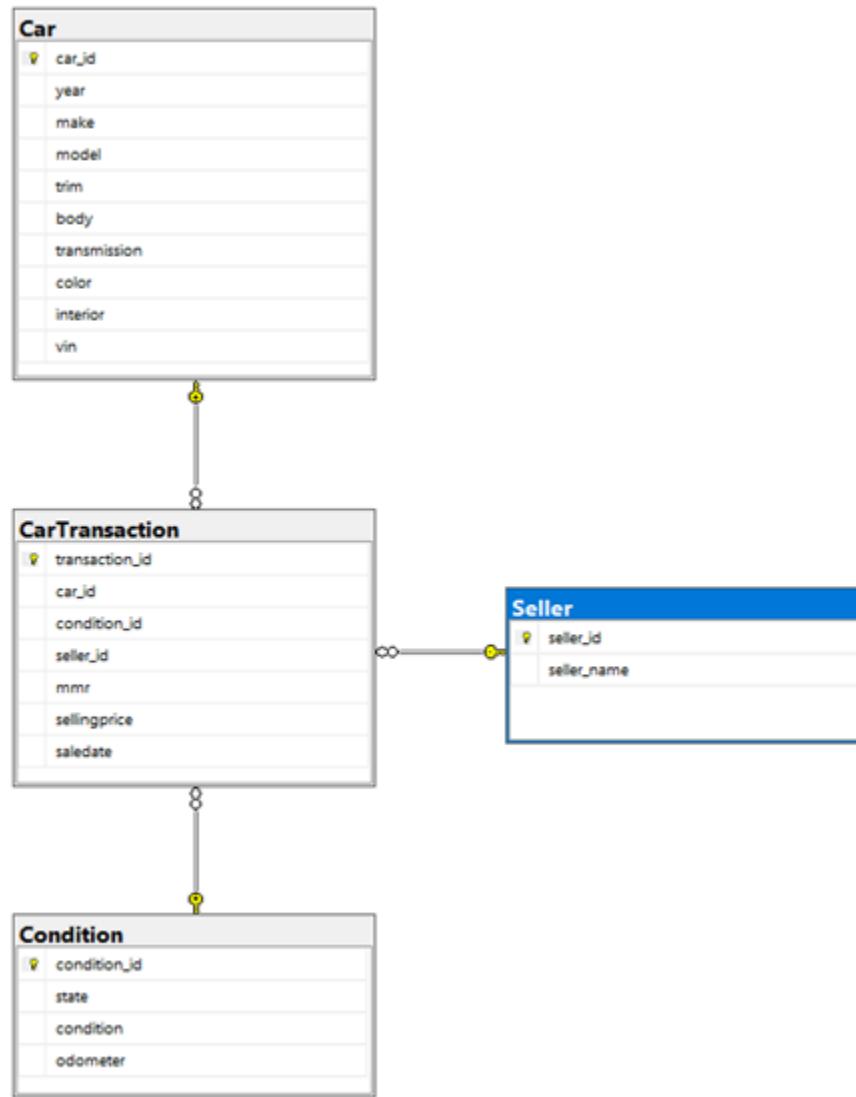
Mục đích: Chuyển đổi mô hình logic thành cấu trúc lưu trữ thực tế trong SQL Server.

BUỚC 6: KIỂM TRA VÀ XÁC MINH MÔ HÌNH

Mục đích: Đảm bảo mô hình dữ liệu hoạt động chính xác và đáp ứng các yêu cầu nghiệp vụ.

Thực hiện:

- Nhập dữ liệu mẫu vào các bảng.
- Chạy các truy vấn để kiểm tra mối quan hệ giữa các bảng, tính toán vẹn của dữ liệu và hiệu suất truy vấn.



Hình 4.7 Mô hình hóa dữ liệu

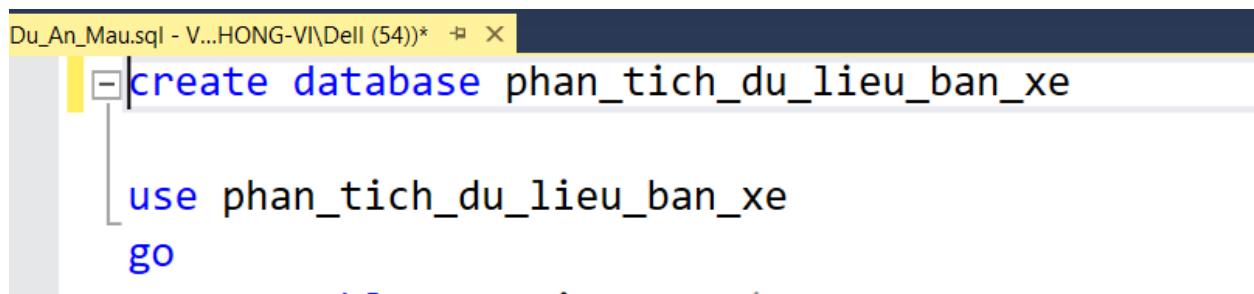
4.2.4 TRÌNH BÀY CÁC BƯỚC TẠO BẢNG DỮ LIỆU

Tạo bảng dữ liệu là một bước quan trọng trong việc triển khai mô hình dữ liệu trong cơ sở dữ liệu. Quá trình này bao gồm việc xác định cấu trúc bảng, các cột, kiểu dữ liệu của các cột, các ràng buộc (constraints) và mối quan hệ giữa các

bảng. Dưới đây là các bước chi tiết để tạo bảng dữ liệu cho dự án sử dụng SQL Server.

BUỚC 1: TẠO CƠ SỞ DỮ LIỆU

Mục đích: Tạo một cơ sở dữ liệu mới để lưu trữ các bảng.



```
Du_An_Mau.sql - V...HONG-VI\HONG-VI (54)*  ↗ X
create database phan_tich_du_lieu_ban_xe
use phan_tich_du_lieu_ban_xe
go
```

Hình 4.8 Tạo cơ sở dữ liệu

BUỚC 2: TẠO BẢNG CARPRICESRAW

Mục đích: Chứa dữ liệu thô để làm sạch

```
- create table CarPricesRaw (
    year int, --
    make varchar(225), --
    model varchar(225), --
    trim varchar(225), --
    body varchar(225), --
    transmission varchar(225), --
    vin varchar(225), --
    state varchar(225), ---
    condition float, ---
    odometer float, ---
    color varchar(225), --
    interior varchar(225), --
    seller varchar(225), ---
    mmr float, ---
    sellingprice float, -----
    saledate varchar(225) -----
)
go
```

Hình 4.9 Tạo bảng CarPricesRaw

BUỚC 3: TẠO BẢNG CAR

Mục đích: Lưu trữ thông tin chung về xe.

Cấu trúc bảng:

- car_id: Khóa chính tự tăng
- make: Hãng sản xuất
- model: Mẫu xe
- trim: Phiên bản xe

- body: Kiểu thân xe
- transmission: Loại hộp số
- color: Màu sắc
- interior: Nội thất
- vin: Số VIN (khóa duy nhất)
- year: Năm sản xuất

```
- create table Car (
    car_id int primary key identity(1,1),
    year int,
    make varchar(225),
    model varchar(225),
    trim varchar(225),
    body varchar(225),
    transmission varchar(225),
    color varchar(225),
    interior varchar(225),
    vin varchar(225)
)
go
```

Hình 4.10 Tạo bảng Car

BUỚC 4: TẠO BẢNG CONDITION

Mục đích: Lưu trữ thông tin về tình trạng của xe tại thời điểm giao dịch.

Cấu trúc bảng:

- condition_id: Khóa chính tự tăng

- state: Bang
- condition: Tình trạng xe
- odometer: Quãng đường đã đi

```

-- create table Condition (
    condition_id int primary key identity(1,1),
    state varchar(225),
    condition float,
    odometer float
)
go
```

Hình 4.11 Tạo bảng Condition

BUỚC 5: TẠO BẢNG SELLER

Mục đích: Lưu trữ thông tin về người bán.

Cấu trúc bảng:

- seller_id: Khóa chính tự tăng
- seller_name: Tên người bán

```

-- create table Seller (
    seller_id int primary key identity(1,1),
    seller_name varchar(225)
)
go
```

Hình 4.12 Tạo bảng Seller

BUỚC 6: TẠO BẢNG CARTRANSACTION

Mục đích: Lưu trữ thông tin về các giao dịch mua bán.

Cấu trúc bảng:

- transaction_id: Khóa chính tự tăng
- car_id: Khóa ngoại tham chiếu đến Car
- condition_id: Khóa ngoại tham chiếu đến Condition
- seller_id: Khóa ngoại tham chiếu đến Seller
- mmr: Giá MMR
- sellingprice: Giá bán
- saledate: Ngày bán

```
create table CarTransaction (
    transaction_id int primary key identity(1,1),
    car_id int,
    condition_id int,
    seller_id int,
    mmr float,
    sellingprice float,
    saledate varchar(225),
    foreign key (car_id) references car(car_id),
    foreign key (condition_id) references condition(condition_id),
    foreign key (seller_id) references seller(seller_id)
)
go
```

Hình 4.13 Tạo bảng CarTransaction

4.3 XỬ LÝ DỮ LIỆU DAX

4.3.1 MEASURE

4.3.1.1 Tạo calendar ngày giao dịch

Cột [DATE] (từ thấp đến cao)

Cột [YEAR] là năm giao dịch

Cột [MONTH] là tháng giao dịch

Cột [MONTHNAME] là tên tháng giao dịch

Cột [DAYOFTHEWEEK] là ngày thứ mấy

Cột [QUARTER] là quý 1 2 3 4 (một quý là 3 tháng)

Cột [YEARQUARTER] là định dạng [NAM]/Q[QUY]

Tạo liên kết với bảng CarTransaction thông qua [Date]

The screenshot shows the Power BI Data Editor interface. On the left, there is a code editor window containing DAX code to generate a calendar table. On the right, there is a preview table showing the resulting data for July 2014.

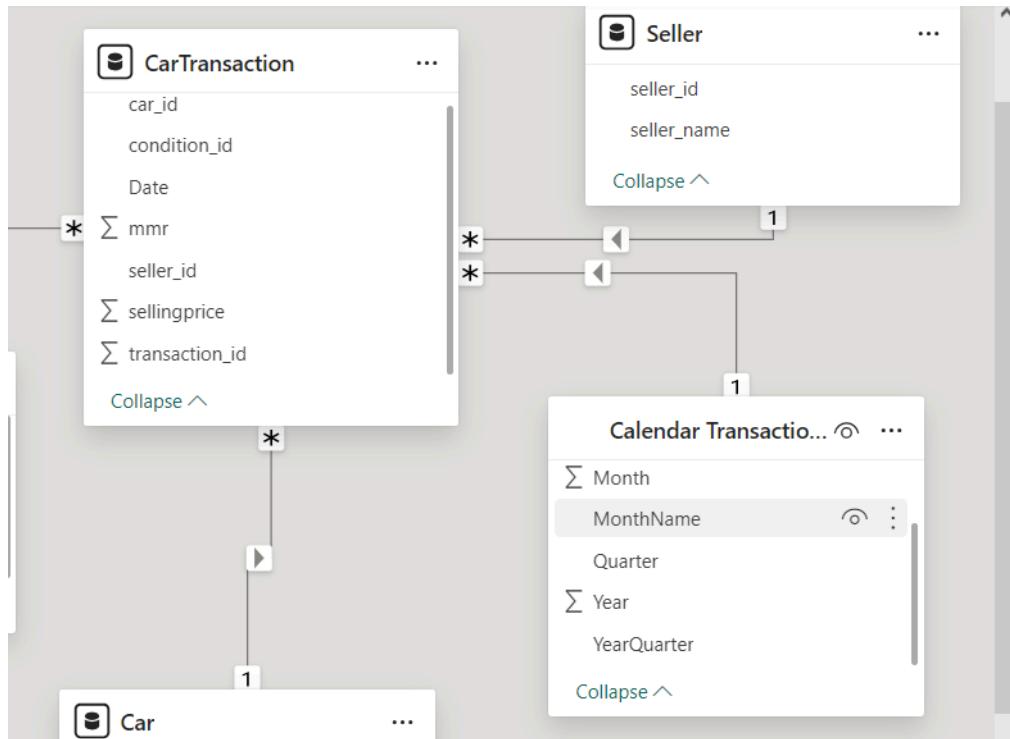
DAX Code:

```
1 Calendar TransactionDay = ADDCOLUMNS(
2     CALENDAR(MIN(CarTransaction[Date]), MAX(CarTransaction[Date])),
3     "Year", YEAR([Date]),
4     "Month", MONTH([Date]),
5     "MonthName", FORMAT([Date], "MMMM"),
6     "DayOfTheWeek", FORMAT([Date], "DDDD"),
7     "Quarter", FORMAT([Date], "Q"),
8     "YearQuarter", FORMAT([Date], "YYYY") & "/Q" & FORMAT([Date], "Q")
9 )
```

Preview Table:

Date	Year	Month	MonthName	DayOfTheWeek	Quarter	YearQuarter
7/1/2014 12:00:00 AM	2014	7	July	Tuesday	3	2014/Q3
7/2/2014 12:00:00 AM	2014	7	July	Wednesday	3	2014/Q3
7/3/2014 12:00:00 AM	2014	7	July	Thursday	3	2014/Q3
7/4/2014 12:00:00 AM	2014	7	July	Friday	3	2014/Q3
7/5/2014 12:00:00 AM	2014	7	July	Saturday	3	2014/Q3
7/6/2014 12:00:00 AM	2014	7	July	Sunday	3	2014/Q3
7/7/2014 12:00:00 AM	2014	7	July	Monday	3	2014/Q3
7/8/2014 12:00:00 AM	2014	7	July	Tuesday	3	2014/Q3
7/9/2014 12:00:00 AM	2014	7	July	Wednesday	3	2014/Q3
7/10/2014 12:00:00 AM	2014	7	July	Thursday	3	2014/Q3
7/11/2014 12:00:00 AM	2014	7	July	Friday	3	2014/Q3
7/12/2014 12:00:00 AM	2014	7	July	Saturday	3	2014/Q3

Hình 4.14 Tạo bảng Calendar Transaction Date



Hình 4.15 Tạo liên kết

4.3.1.2 Tạo số lượng giao dịch trong CountTable

Hình 4.16 Tạo measure Số lượng giao dịch

4.3.1.3 Tạo số lượng Condition trong CountTable

The screenshot shows the Power BI Data view interface. In the top navigation bar, the 'Security' tab is selected. Below it, the formula bar contains the text: `Số lượng Condition = COUNTROWS('Condition')`. A red box highlights this formula. On the left, the 'Properties' pane shows a 'Measure' section with three items: `Σ Value`, `Số lượng Condition` (which is highlighted with a red box), and `Số lượng giao dịch`. On the right, the 'Data' pane displays a hierarchy under 'Measure': `CountTable` (highlighted with a red box) has three children: `Số lượng Condition` (highlighted with a red box), `Số lượng giao dịch`, and `Số lượng xe`. Below 'CountTable' is `Σ Value`, and at the bottom is `Seller`.

Hình 4.17 Tạo measure số lượng Condition

4.3.1.4 Tạo số lượng xe trong CountTable

This screenshot shows the Power BI Data view interface. The formula bar contains the text: `Số lượng xe = COUNTROWS(Car)`. A red box highlights this formula. The 'Properties' pane on the left shows a 'Measure' section with three items: `Σ Value`, `Số lượng Condition`, and `Số lượng xe` (highlighted with a red box). The 'Collapse ^' button is visible. The 'Data' pane on the right shows the 'Measure' hierarchy under 'CountTable': it contains `Số lượng Condition`, `Số lượng giao dịch`, and `Số lượng xe` (highlighted with a red box). Below 'CountTable' is `Σ Value`, and at the bottom is `Seller`.

Hình 4.18 Tạo measure số lượng xe

4.3.1.5 Tạo số lượng Seller trong CountTable

The screenshot shows the Power BI Data view interface. In the top bar, the 'setup' tab is selected. Below it, the formula bar contains the DAX formula: `Số lượng Seller = COUNTROWS(Seller)`. A red box highlights this formula. On the left, a 'Properties' panel is open, showing a list of measures: `Σ Value`, `Số lượng Condition`, `Số lượng giao dịch`, `Số lượng Seller` (which is highlighted with a green box), and `Số lượng xe`. At the bottom of this list is a 'Collapse ^' button. On the right, the 'Data' pane lists various entities: `Calendar TransactionDate`, `Car`, `CarTransaction`, `Condition`, `Measure` (which is expanded, showing `CountTable`, `Số lượng Condition`, `Số lượng giao dịch`, `Số lượng Seller`, and `Số lượng xe`), and `Seller`. A red box highlights the 'Measure' node and its children under 'CountTable'.

Hình 4.19 Tạo measure số lượng seller

4.3.1.6 Tạo measure Tổng doanh thu (Total Revenue)

The screenshot shows the Power BI Data view interface. The formula bar displays the DAX formula: `1 Total Revenue = SUM(CarTransaction[sellingprice])`. This formula is highlighted with a green box.

Hình 4.20 Tạo measure Total Revenue

4.3.1.7 Tạo measure Giá trị trung bình giá bán ra (Average Selling Price)

The screenshot shows the Power BI Data view interface. The formula bar displays the DAX formula: `1 Average Selling Price = AVERAGE(CarTransaction[sellingprice])`. This formula is highlighted with a green box.

Hình 4.21 Tạo measure Average Selling Price

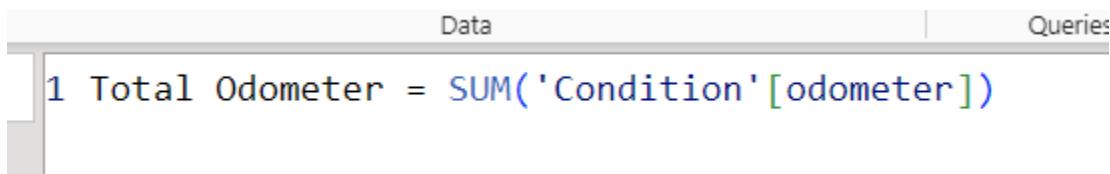
4.3.1.8 Tạo measure Giá trị trung bình MRR (Average MMR)



```
1 Average MMR = AVERAGE(CarTransaction[mmr])
```

Hình 4.22 Tạo measure Average MMR

4.3.1.9 Tạo measure Tổng quãng đường đã đi (Total Odometer)



```
1 Total Odometer = SUM('Condition'[odometer])
```

Hình 4.23 Tạo measure Total Odometer

4.3.2 CALCULATED COLUMN

4.3.2.1 Tạo Cột [Tổng doanh thu] trong bảng Seller

The screenshot shows the Power BI Data Editor interface. A red box highlights the formula bar where the formula `Tổng doanh thu = CALCULATE(SUM(CarTransaction[sellingprice]))` is entered. Another red arrow points to the newly created column 'Tổng doanh thu' in the 'Seller' table, which contains the following data:

	seller_id	seller name	Tổng doanh thu
1	mercedes benz of ontario		957000
2	joyce buick gmc inc		47100
3	prestige pre owned llc		9400
4	mansfield gas & service		4400
5	rent to own auto showroom llc		2400
6	john's auto sales inc		16300
7	freeman buick gmc		159700
8	lendmark financial services/raleigh		14300

Hình 4.24 Tạo cột tổng doanh thu trong bảng Seller

4.3.2.2 Tạo cột Car Age trong bảng Car

The screenshot shows the Power BI Data Editor interface. A red box highlights the formula bar where the formula `Car Age = YEAR(TODAY()) - Car[year]` is entered. Another red arrow points to the newly created column 'Car Age' in the 'Car' table, which contains the following data:

	car_id	year	make	model	trim	body	transmission	color	interior	vin	Car Age
353172	2013	Ford	Fusion	SE	Sedan	automatic	black	black	3fa6p0h71dr118478	11	
353173	2013	Ford	Fusion	SE	Sedan	automatic	black	black	3fa6p0h71dr207810	11	
353174	2013	Ford	Fusion	SE	Sedan	automatic	black	black	3fa6p0h72dr341080	11	
353175	2013	Ford	Fusion	SE	Sedan	automatic	black	black	3fa6p0h74dr213780	11	
353176	2013	Ford	Fusion	SE	Sedan	automatic	black	black	3fa6p0h75dr102266	11	
353177	2013	Ford	Fusion	SE	Sedan	automatic	black	black	3fa6p0h75dr302886	11	
353178	2013	Ford	Fusion	SE	Sedan	automatic	black	black	3fa6p0h76dr214526	11	
353179	2013	Ford	Fusion	SE	Sedan	automatic	black	black	3fa6p0h77dr291163	11	
353180	2013	Ford	Fusion	SE	Sedan	automatic	black	black	3fa6p0h78dr124360	11	

Hình 4.25 Tạo cột Car Age

4.3.2.3 Tạo cột Condition by Odometer trong bảng Condition

The screenshot shows a database interface with a code editor at the top containing the following SQL-like code:

```

1 Condition by Odometer =
2   IF('Condition'[odometer] < 50000, "New",
3   IF('Condition'[odometer] < 100000, "Used", "Old"))

```

Below the code is a table with columns: state, condition, odometer, and Condition by Odometer. The last column contains the results of the calculated column:

	state	condition	odometer	Condition by Odometer
1516	fl	19	2338	New
1517	fl	19	3834	New
1518	fl	19	7269	New
1519	fl	19	7390	New
1520	fl	19	8506	New
1521	fl	19	9719	New
1522	fl	19	10591	New
1523	fl	19	10689	New
1524	fl	19	12725	New

Hình 4.26 Tạo cột Condition by Odometer

4.3.2.4 Tạo cột Profit trong bảng CarTransaction

The screenshot shows a database interface with a code editor at the top containing the following SQL-like code:

```

1 Profit = CarTransaction[sellingprice] - CarTransaction[mmr]

```

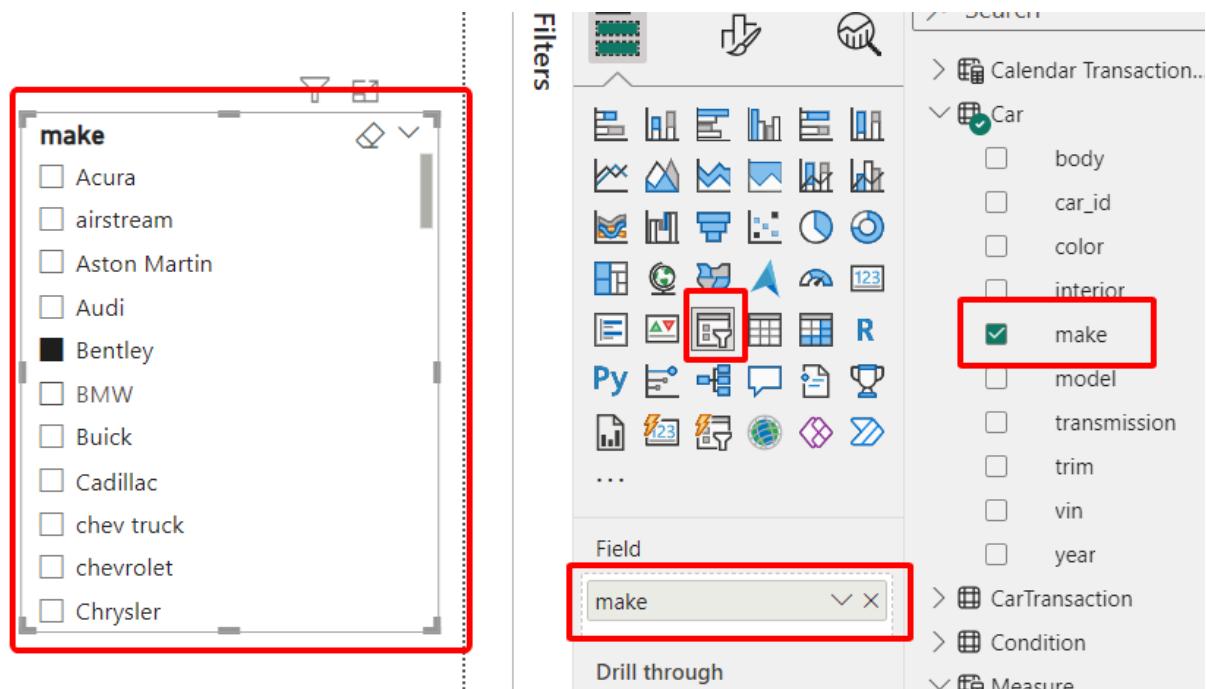
Below the code is a table with columns: transaction_id, car_id, condition_id, seller_id, mmr, sellingprice, Date, and Profit. The Profit column shows the calculated profit for each transaction.

transaction_id	car_id	condition_id	seller_id	mmr	sellingprice	Date	Profit
516052	121886	401474	11763	\$5,150	\$5,900	1/28/2015 2:00:00 AM	750
512016	179934	541597	11763	\$5,325	\$4,100	1/28/2015 2:00:00 AM	-1225
111561	172827	304415	11763	\$6,125	\$4,100	1/28/2015 2:00:00 AM	-2025
104613	114351	525479	11763	\$6,900	\$5,900	1/28/2015 2:00:00 AM	-1000
72187	151504	392593	11763	\$7,450	\$4,800	1/28/2015 2:00:00 AM	-2650
95051	263589	289890	11763	\$7,700	\$7,300	1/28/2015 2:00:00 AM	-400
102724	106912	423465	11763	\$8,125	\$8,600	1/28/2015 2:00:00 AM	475
72439	259509	398331	11763	\$8,275	\$8,000	1/28/2015 2:00:00 AM	-275
97314	351130	423359	11763	\$8,300	\$8,600	1/28/2015 2:00:00 AM	300
522669	521535	257138	11763	\$8,425	\$8,700	1/28/2015 2:00:00 AM	275
524701	255640	270951	11763	\$8,475	\$5,700	1/28/2015 2:00:00 AM	-2775
97312	158089	419676	11763	\$8,575	\$10,700	1/28/2015 2:00:00 AM	2125
517752	262808	436046	11763	\$8,675	\$8,300	1/28/2015 2:00:00 AM	-375
519797	259505	260691	11763	\$8,700	\$9,500	1/28/2015 2:00:00 AM	800
00620	252286	205172	11763	€8.875	€7.800	1/28/2015 2:00:00 AM	-1075

Hình 4.27 Tạo cột Profit

4.3.3 FILTER

4.3.3.1 Tạo filter chọn hãng sản xuất (make)



Hình 4.28 Tạo filter chọn hãng sản xuất

4.3.3.2 Tạo filter chọn model

The screenshot shows the Power BI Data view. On the left, there is a dropdown menu with the title 'model' containing a list of car models: 370Z, 4 Series, 4 Series Gran Coupe, 400-Class, 420-Class, 420sel, 42c, 458 Italia, 4Runner, 5 Series, and 5 Series Gran Turismo. A red box highlights this dropdown. On the right, the Fields pane shows a list of fields under the 'Car' category. The 'model' field is selected, indicated by a checked checkbox. A red box highlights the 'model' checkbox. The 'Field' search bar at the bottom also contains the word 'model'. A red box highlights the search bar.

Hình 4.29 Tạo filter chọn mẫu xe

4.3.3.3 Tạo filter chọn năm sản xuất

The screenshot shows the Power BI Data view. On the left, there is a dropdown menu with the title 'year' containing a list of years from 1982 to 1992. A red box highlights this dropdown. On the right, the Fields pane shows a list of fields under the 'Car' category. The 'year' field is selected, indicated by a checked checkbox. A red box highlights the 'year' checkbox. The 'Field' search bar at the bottom also contains the word 'year'. A red box highlights the search bar.

Hình 4.30 Tạo filter chọn năm sản xuất

4.3.3.4 Tạo filter chọn transmission

The screenshot shows the Power BI 'Filters' pane. On the left, there is a dropdown menu labeled 'transmission' containing four options: 'automatic', 'manual', 'Sedan', and 'unknow'. This dropdown is highlighted with a red box. In the center, there is a large icon library grid with various visualization icons. Below it, a 'Field' dropdown is set to 'transmission', also highlighted with a red box. On the right, a tree view of fields is shown under the 'Car' category. The 'transmission' field is checked and highlighted with a red box.

Hình 4.32 Tạo filter chọn transmission

4.3.3.5 Tạo filter chọn Seller

The screenshot shows the Power BI 'Filters' pane. On the left, there is a dropdown menu labeled 'seller_name' containing a long list of seller names, many of which start with 'a'. This dropdown is highlighted with a red box. In the center, there is a large icon library grid with various visualization icons. Below it, a 'Field' dropdown is set to 'seller_name', highlighted with a red box. On the right, a tree view of fields is shown under the 'Seller' category. The 'seller_name' field is checked and highlighted with a red box.

Hình 4.33 Tạo filter chọn Seller

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 "tất". 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, trực quan hóa dữ liệu chuỗi thời gian và trực quan hóa dữ liệu đa chiều là hai kỹ thuật chính được áp dụng.

Trực quan hóa dữ liệu chuỗi thời gian được sử dụng để theo dõi và phân tích xu hướng giá bán xe qua các khoảng thời gian khác nhau. Việc sử dụng biểu đồ đường giúp biểu thị các thay đổi trong giá bán, doanh số, và các chỉ số quan trọng khác theo thời gian, giúp dễ dàng nhận ra các xu hướng và sự kiện bất thường.

Trực quan hóa dữ liệu đa chiều được áp dụng thông qua biểu đồ cột và biểu đồ tròn để so sánh các yếu tố dữ liệu như giá bán theo hãng xe, tình trạng xe, và quãng đường đã đi. Biểu đồ cột giúp so sánh và phân tích các yếu tố này, trong khi biểu đồ tròn cho phép trực quan hóa tỷ lệ phần trăm của các danh mục khác nhau trong tổng thể.

Việc áp dụng hai kỹ thuật này giúp truyền tải thông tin một cách rõ ràng, dễ hiểu, và hiệu quả, giúp người dùng nắm bắt nhanh chóng các kết quả phân tích và đưa ra quyết định dựa trên dữ liệu.

5.2 CÁC NGUYÊN TẮC TRỰC QUAN HÓA

5 NGUYÊN TẮC TRỰC QUAN HÓA DỮ LIỆU

Chọn đúng loại biểu đồ

Chọn đúng loại biểu đồ là một nguyên tắc quan trọng trong trực quan hóa dữ liệu, vì loại biểu đồ phù hợp sẽ giúp truyền tải thông tin một cách rõ ràng, chính xác và dễ hiểu. Mỗi loại biểu đồ có những đặc điểm riêng và thích hợp để biểu diễn các loại dữ liệu khác nhau. Ví dụ, biểu đồ đường là lựa chọn tốt để hiển thị xu hướng dữ liệu qua thời gian, giúp người xem dễ dàng nhận ra các biến động và xu hướng dài hạn. Biểu đồ cột thì phù hợp để so sánh các giá trị giữa các danh mục khác nhau, như doanh số bán hàng của các sản phẩm khác nhau trong cùng một kỳ.

Biểu đồ tròn là công cụ hữu ích để biểu diễn tỷ lệ phần trăm của các phần tử trong tổng thể, giúp người xem nhanh chóng nhận biết được phần đóng góp của từng danh mục. Việc chọn đúng loại biểu đồ không chỉ làm cho dữ liệu trở nên sống động và trực quan hơn, mà còn hỗ trợ quá trình phân tích và ra quyết định hiệu quả hơn. Trong thực tế, việc lựa chọn loại biểu đồ phù hợp còn phụ thuộc vào mục tiêu truyền tải thông tin và đối tượng người xem, đảm bảo rằng thông tin được trình bày một cách trực quan, dễ hiểu và có tác động nhất.

Không phải tất cả data đều quan trọng

Nguyên tắc "Không phải tất cả dữ liệu đều quan trọng" nhấn mạnh tầm quan trọng của việc chọn lọc và tập trung vào những dữ liệu thật sự cần thiết và có ý nghĩa khi trực quan hóa. Trong quá trình phân tích dữ liệu, rất dễ bị cuốn vào việc cố gắng trình bày tất cả thông tin mà ta có. Tuy nhiên, việc này không chỉ làm cho biểu đồ trở nên rối rắm và khó hiểu, mà còn làm mất đi sự rõ ràng và trọng tâm của thông điệp mà ta muốn truyền tải. Thay vào đó, nên tập trung vào những dữ liệu chủ chốt, có liên quan trực tiếp đến câu hỏi nghiên cứu hoặc mục tiêu kinh doanh. Ví dụ, khi trình bày xu hướng doanh số bán hàng, ta chỉ cần hiển thị các số liệu về doanh thu, số lượng bán, và thời gian, thay vì bao gồm cả những chi tiết không liên quan như mã sản phẩm hoặc thông tin nội bộ không cần thiết. Bằng cách loại bỏ những dữ liệu không quan trọng, ta không chỉ làm cho biểu đồ trở nên sạch sẽ và dễ hiểu hơn mà còn giúp người xem dễ dàng nắm bắt và tập trung vào những thông tin quan trọng nhất, từ đó đưa ra những quyết định sáng suốt và hiệu quả hơn.

Biểu đồ thể hiện đúng tương quan số liệu thực tế

Nguyên tắc "Biểu đồ thể hiện đúng tương quan số liệu thực tế" nhấn mạnh tầm quan trọng của việc sử dụng biểu đồ để phản ánh chính xác mối quan hệ giữa các dữ liệu. Khi trực quan hóa dữ liệu, điều quan trọng là phải đảm bảo rằng biểu

đồ không chỉ đẹp mắt mà còn trung thực và chính xác trong việc biểu diễn các số liệu. Một biểu đồ không chính xác có thể dẫn đến những hiểu lầm nghiêm trọng và đưa ra các quyết định sai lầm. Ví dụ, trực tung của biểu đồ cột nên bắt đầu từ số không để tránh việc phóng đại hoặc giảm nhẹ sự khác biệt giữa các giá trị. Tương tự, việc sử dụng tỉ lệ và khoảng cách phù hợp trong biểu đồ phân tán có thể giúp minh họa một cách chính xác mối tương quan giữa các biến số. Việc sử dụng biểu đồ tròn để biểu diễn các phần của tổng thể cũng cần chú ý để các phần này thực sự phản ánh đúng tỷ lệ phần trăm của từng phần. Tóm lại, việc đảm bảo biểu đồ thể hiện đúng tương quan số liệu thực tế không chỉ giúp truyền tải thông tin một cách rõ ràng và minh bạch mà còn tạo dựng niềm tin và sự tin cậy từ người xem, giúp họ dễ dàng tiếp nhận và hiểu đúng bản chất của dữ liệu.

Sử dụng màu sắc hợp lý khi chuyển dữ liệu sang dạng biểu đồ

Nguyên tắc "Sử dụng màu sắc hợp lý khi chuyển dữ liệu sang dạng biểu đồ" nhấn mạnh tầm quan trọng của việc chọn và phối màu một cách khoa học để làm nổi bật thông tin mà không gây nhầm lẫn hay quá tải cho người xem. Màu sắc là một công cụ mạnh mẽ trong trực quan hóa dữ liệu, nhưng việc sử dụng không hợp lý có thể dẫn đến hiểu nhầm và làm giảm hiệu quả của biểu đồ. Các màu sắc cần được chọn sao cho dễ phân biệt và có ý nghĩa, ví dụ, sử dụng màu đỏ để biểu thị sự giảm sút hoặc cảnh báo, và màu xanh để biểu thị sự tăng trưởng hoặc an toàn. Ngoài ra, nên tránh sử dụng quá nhiều màu sắc hoặc các màu sắc quá sáng, quá tối, khiến biểu đồ trở nên rối mắt. Tương phản màu sắc nên được sử dụng để làm nổi bật các phần quan trọng và tạo ra sự rõ ràng giữa các yếu tố khác nhau trong biểu đồ. Đồng thời, cần đảm bảo rằng biểu đồ vẫn dễ đọc đối với những người bị mù màu, bằng cách sử dụng các mẫu hoặc các sắc độ khác nhau của cùng một màu. Tóm lại, việc sử dụng màu sắc hợp lý không chỉ làm cho biểu đồ trở nên hấp dẫn

hơn mà còn giúp người xem dễ dàng tiếp cận và hiểu rõ thông tin được truyền tải, từ đó hỗ trợ tốt hơn cho việc ra quyết định.

Luôn đảm bảo dữ liệu được trình bày một cách đơn giản và hiệu quả

Nguyên tắc "Luôn đảm bảo dữ liệu được trình bày một cách đơn giản và hiệu quả" nhấn mạnh tầm quan trọng của việc giữ cho biểu đồ và đồ thị dễ hiểu và trực quan nhất có thể. Trong quá trình trực quan hóa dữ liệu, việc sử dụng các yếu tố thiết kế đơn giản và tránh các chi tiết thừa thãi giúp người xem dễ dàng tập trung vào thông tin cốt lõi mà không bị phân tâm. Điều này có nghĩa là sử dụng các loại biểu đồ phù hợp, tránh quá tải thông tin bằng cách chỉ chọn những dữ liệu cần thiết và loại bỏ các yếu tố không quan trọng. Ví dụ, thay vì sử dụng một biểu đồ với quá nhiều đường hoặc cột, hãy chia nhỏ dữ liệu và sử dụng các biểu đồ riêng biệt để minh họa các khía cạnh khác nhau. Đảm bảo các nhãn, chú thích và tiêu đề được trình bày rõ ràng, dễ đọc và không làm rối mắt người xem. Sự đơn giản và hiệu quả trong thiết kế biểu đồ không chỉ giúp người xem dễ dàng hiểu và tiếp nhận thông tin mà còn tăng khả năng ghi nhớ và áp dụng thông tin đó vào thực tế. Tóm lại, giữ cho dữ liệu được trình bày một cách đơn giản và hiệu quả là chìa khóa để truyền đạt thông tin một cách chính xác và tác động mạnh mẽ đến người xem.

NGUYÊN TẮC QUAN TRỌNG NHẤT: CHỌN ĐÚNG LOẠI BIỂU ĐỒ

Trong phạm vi dự án này, nguyên tắc "Chọn đúng loại biểu đồ" được coi là quan trọng nhất. Lý do chính là vì việc chọn đúng loại biểu đồ sẽ quyết định cách thông tin được truyền tải và hiểu bởi người xem. Khi dữ liệu được trình bày một cách rõ ràng và chính xác thông qua loại biểu đồ phù hợp, người xem sẽ dễ dàng nắm bắt các xu hướng, mối quan hệ và ý nghĩa từ dữ liệu hơn.

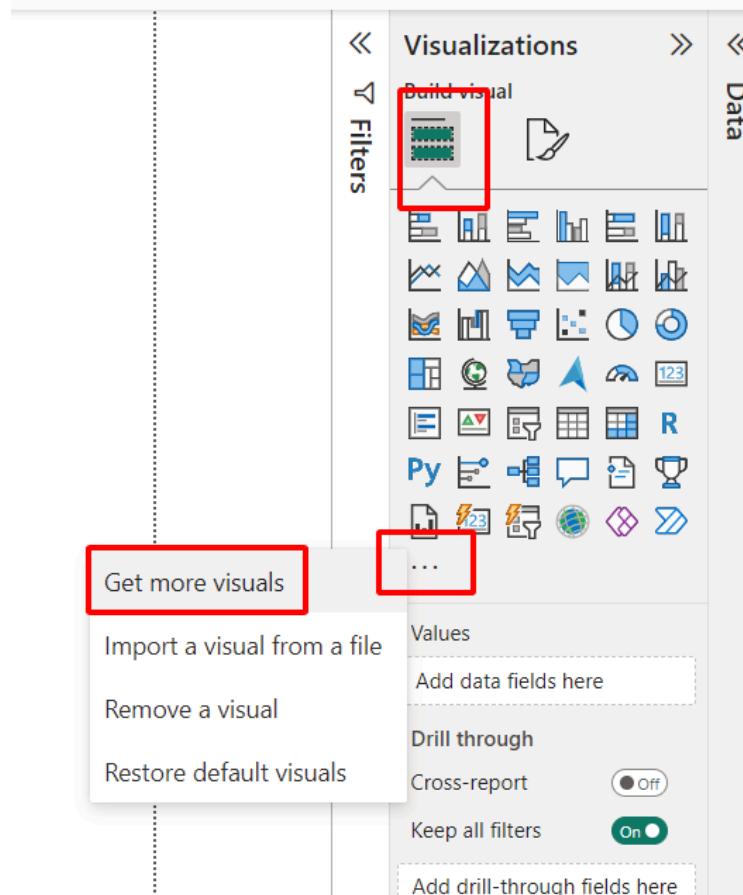
Ví dụ, nếu mục tiêu là theo dõi và phân tích xu hướng giá bán xe qua thời gian, việc sử dụng biểu đồ đường sẽ giúp minh họa sự thay đổi liên tục và các điểm

biến động quan trọng một cách trực quan. Ngược lại, nếu cần so sánh doanh số giữa các hãng xe, biểu đồ cột sẽ là lựa chọn tối ưu để thể hiện sự khác biệt rõ ràng giữa các danh mục.

Sự lựa chọn biểu đồ đúng đắn không chỉ làm cho dữ liệu trở nên sống động và dễ hiểu mà còn giúp người dùng tiết kiệm thời gian và công sức trong việc phân tích và ra quyết định. Trong bối cảnh dữ liệu phong phú và phức tạp của dự án, việc chọn đúng loại biểu đồ sẽ đảm bảo rằng các thông tin quan trọng được truyền tải một cách hiệu quả nhất, hỗ trợ tối đa cho quá trình quản lý, kinh doanh và đầu tư của công ty.

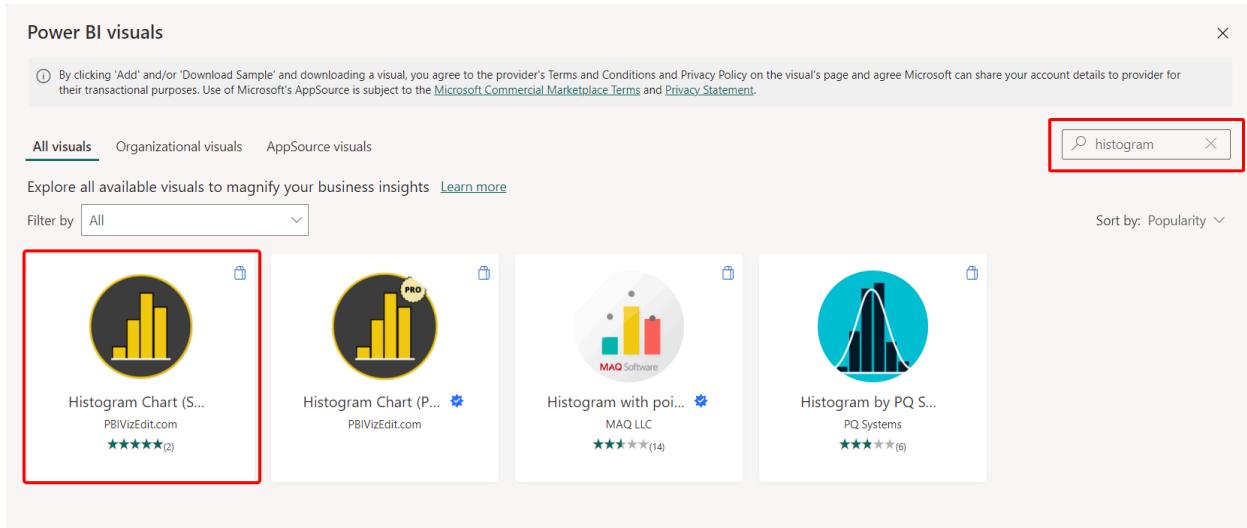
5.3 TRÌNH BÀY CÁCH THÊM VISUAL MỚI

Bước 1:



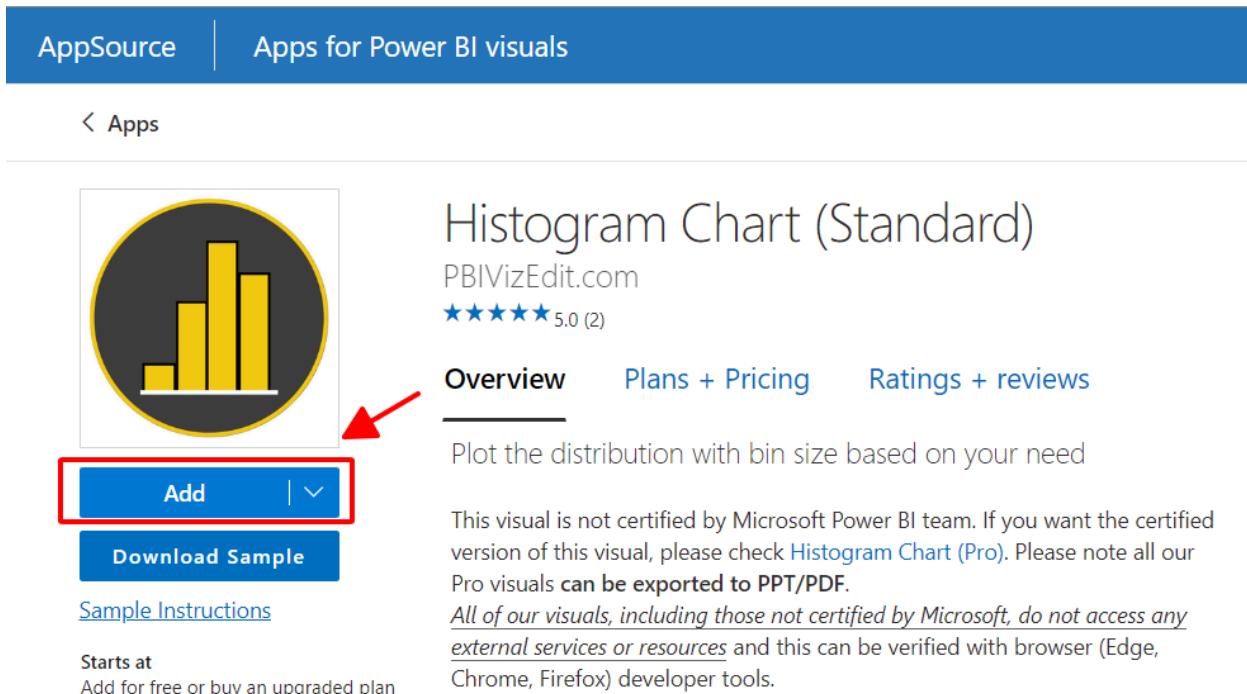
Hình 5.1 Get more visual

Bước 2:



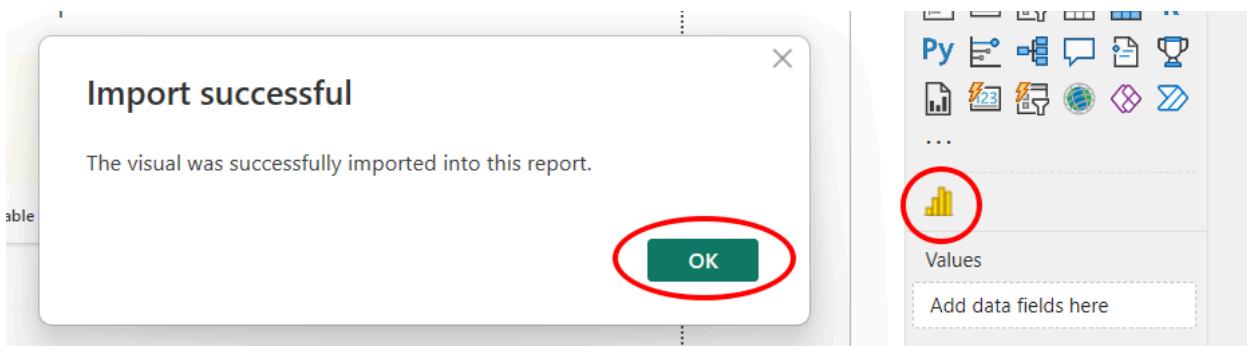
Hình 5.2 Tìm visual mới cần thêm

Bước 3:



Hình 5.3 Add visual

Bước 4:



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 THÔNG KÊ CHI TIẾT

5.4.1.1 Tạo visual filter theo ngày giao dịch

A screenshot of the Power BI interface showing the "Filters", "Visualizations", and "Data" panes. In the "Filters" pane, a date range selector is highlighted with a red box and a red arrow pointing to it. In the "Data" pane, under the "Search" field, a filter for "Date" is selected, indicated by a checked checkbox and a red box around it. A red arrow points from this selection to a red box highlighting the "Date" field in the "Field" dropdown at the bottom of the "Data" pane.

Hình 5.5 Tạo filter theo ngày giao dịch

5.4.1.2 Tạo visual filter theo hãng sản xuất

The screenshot shows the Power BI 'Filters' pane on the left, 'Visualizations' pane in the center, and 'Data' pane on the right. In the 'Filters' pane, a dropdown menu for 'make' is open, showing options like BMW, Kia, and Volvo. A red arrow points from the 'Filters' pane to this dropdown. In the 'Data' pane, the 'make' checkbox is selected, indicated by a checked green box. Another red arrow points from the 'Data' pane to the 'make' checkbox. The 'Field' dropdown at the bottom of the 'Filters' pane also has 'make' selected.

Hình 5.6 Tạo filter theo hãng sản xuất

5.4.1.3 Tạo visual filter theo model

The screenshot shows the Power BI 'Filters' pane on the left, 'Visualizations' pane in the center, and 'Data' pane on the right. In the 'Filters' pane, a dropdown menu for 'model' is open, listing various car models such as 370Z, 4 Series, 4 Series Gran Coupe, etc. A red arrow points from the 'Filters' pane to this dropdown. In the 'Data' pane, the 'model' checkbox is selected, indicated by a checked green box. Another red arrow points from the 'Data' pane to the 'model' checkbox. The 'Field' dropdown at the bottom of the 'Filters' pane also has 'model' selected.

Hình 5.7 Tạo filter theo model

5.4.2 TẠO VISUAL THỐNG KÊ TỔNG THỂ

5.4.2.1 Tạo visual filter theo ngày giao dịch

The screenshot shows the Power BI interface with the 'Filters' pane on the left and the 'Data' pane on the right. In the 'Filters' pane, a date range selector is highlighted with a red box and a red arrow pointing to it from the top. Below it, under 'Filters on this visual', there is a dropdown menu for 'Date' set to 'is (All)'. In the 'Data' pane, a search bar is at the top. Below it, under 'Calendar Transaction...', there is a checked checkbox for 'Date'. A red box highlights this checkbox, and a red arrow points to it from the top right. The 'Fields' section in the 'Data' pane also has a 'Date' dropdown, which is also highlighted with a red box and a red arrow pointing to it from the bottom right.

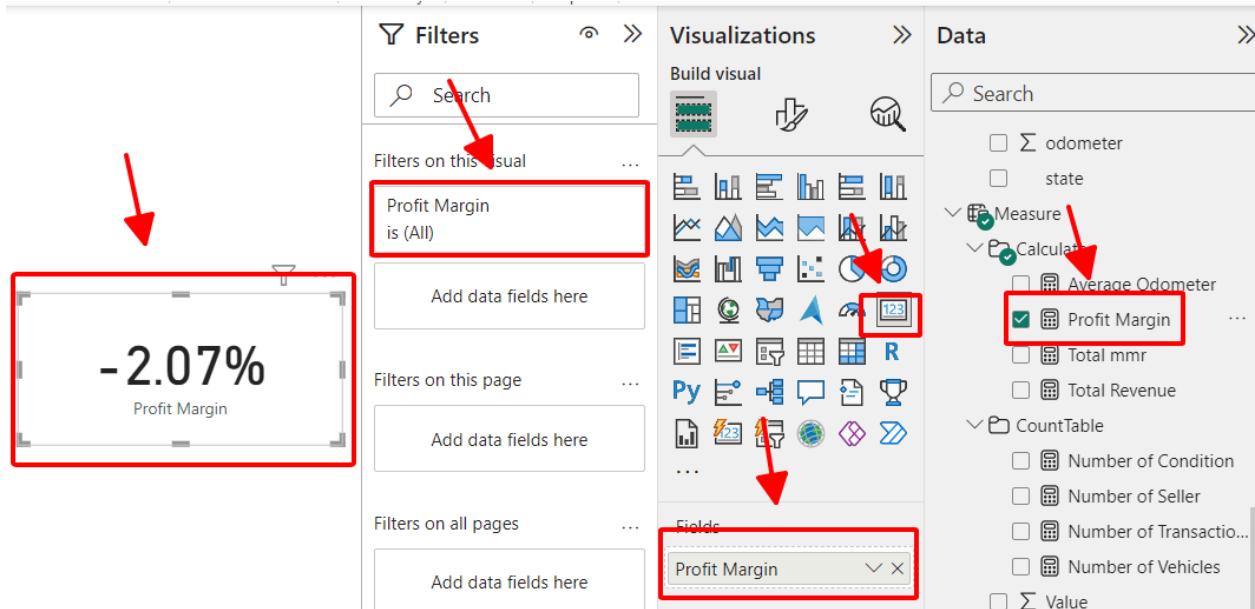
Hình 5.8 Tạo visual filter theo ngày giao dịch

5.4.2.2 Tạo visual thống kê Total Sales

The screenshot shows the Power BI interface with the 'Filters' pane on the left and the 'Data' pane on the right. On the left, a large numerical value '17.37M' representing 'Total Revenue' is displayed, with a red box highlighting the entire area and a red arrow pointing to it from the top left. In the 'Filters' pane, under 'Filters on this visual', there is a dropdown menu for 'Total Revenue' set to 'is (All)'. In the 'Data' pane, under 'Measure', there is a checked checkbox for 'Total Revenue'. A red box highlights this checkbox, and a red arrow points to it from the middle right. The 'Fields' section in the 'Data' pane also has a 'Total Revenue' dropdown, which is highlighted with a red box and a red arrow pointing to it from the bottom right.

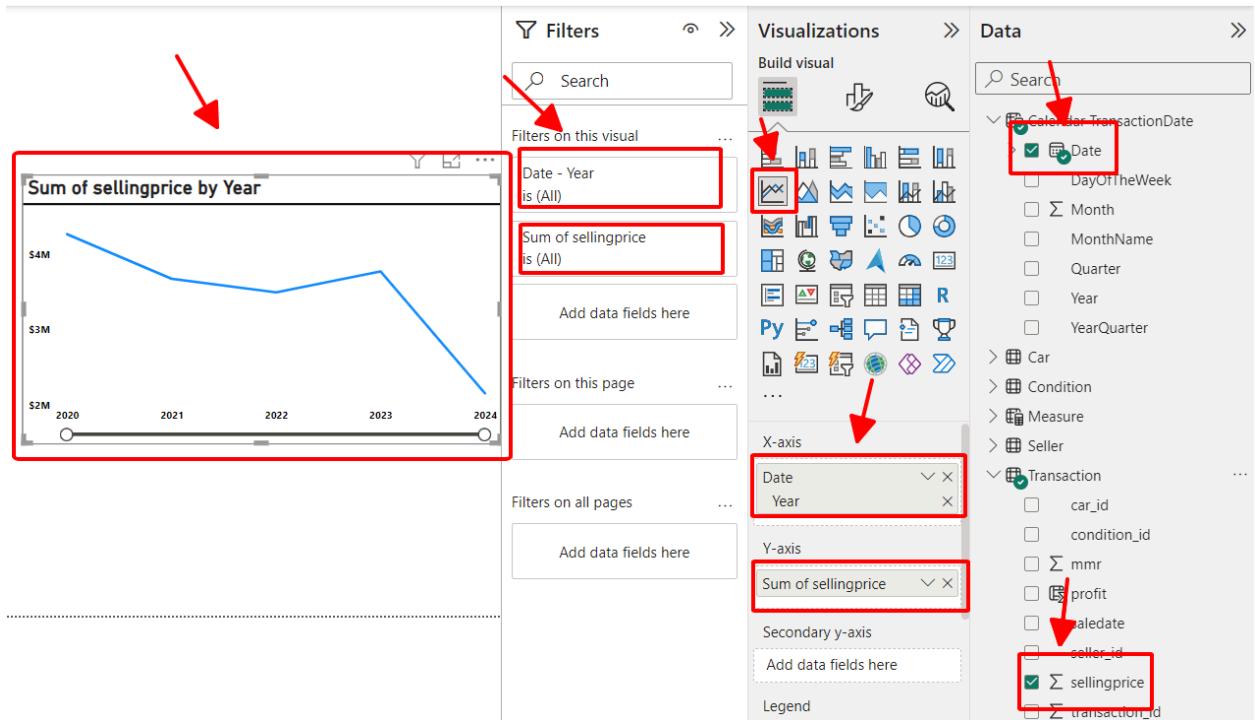
Hình 5.9 Tạo visual thống kê total sales

5.4.2.3 Tạo visual thống kê Profit Margin



Hình 5.10 Tạo visual thống kê profit margin

5.4.2.4 Tạo visual thống kê Sum of selling price theo thời gian



Hình 5.11 Tạo visual thống kê Sum of selling price theo thời gian

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 UỐU 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.

TỐI UU HÓA REPORT

Cấu trúc rõ ràng

- **Phân chia theo trang (page):** Chia report thành các trang rõ ràng theo chủ đề (ví dụ: doanh số, tình trạng xe, phân tích người bán).
- **Sử dụng tiêu đề và chú thích:** Đặt tiêu đề và chú thích rõ ràng cho mỗi biểu đồ và bảng để người đọc dễ hiểu.

Tập trung vào số liệu quan trọng

- **Loại bỏ dữ liệu không cần thiết:** Chỉ hiển thị các số liệu và thông tin thực sự quan trọng cho người đọc.
- **Sử dụng các measure chính:** Tạo và hiển thị các measure quan trọng như tổng doanh thu, số lượng giao dịch, giá bán trung bình.

Tính tương tác

- **Thêm các bộ lọc tương tác:** Cho phép người dùng lọc dữ liệu theo các tiêu chí khác nhau để xem thông tin chi tiết hơn.

- **Sử dụng drill-through:** Cho phép người dùng nhấp vào một phần của biểu đồ để xem chi tiết hơn về phần đó.

Đảm bảo hiệu suất tốt

- **Tối ưu hóa truy vấn dữ liệu:** Sử dụng các truy vấn SQL hiệu quả để lấy dữ liệu.
- **Giảm tải dữ liệu:** Chỉ lấy và hiển thị dữ liệu cần thiết để tránh làm chậm report.

6.2 XÂY DỰNG BÁO CÁO

6.2.1 DASHBOARD VS REPORT

Item	Dashboard	Report
Pages	Chỉ một trang.	Có thể tạo một hoặc nhiều trang.
Data sources	Tạo từ nhiều datasets hoặc reports.	Tạo từ một dataset.
Visualization	Xây dựng thông tin chi tiết về dữ liệu bằng cách sử dụng biểu đồ, hình ảnh sinh động, lưu đồ.	Xây dựng trang tổng hợp thông tin, cung cấp góc nhìn tổng quan.
Available in Power BI Desktop	Không thể tạo.	Có thể tạo và xem.

Filters and Slicers	Không thể dùng, bị giới hạn trên trang đơn.	Có thể dùng các loại filter, highlight và slice khác nhau.
User Interactivity	Cho phép pin visuals từ report và datasets trên canvas đơn, làm cho nó đơn giản để nhóm những gì cần thiết cho người dùng.	Tập trung hơn vào khả năng trực quan hóa, áp dụng các phép biến đổi trên một tập dữ liệu.
Favourite	Có thể gán nhiều dashboards	Có thể gán nhiều dashboards
Q&A feature	Có	Có thể phân quyền chỉnh sửa report.
Alerts	Gửi email cảnh báo khi đáp ứng điều kiện, tiêu chí cụ thể hoặc vượt qua giới hạn.	Không thể
Subscribe	Có	Có
See underlying dataset tables and fields	Không thể xem dataset nhưng có thể xuất dữ liệu này.	Có thể xem dữ liệu trong tab dữ liệu.
Purpose	Được sử dụng để giám sát cấp cao, thường theo thời gian thực hoặc gần thời gian	Được sử dụng để phân tích chuyên sâu và khám phá dữ liệu nhằm trả lời các câu hỏi kinh doanh phức tạp.

thực, cung cấp cái nhìn tổng hợp về hiệu quả kinh doanh.

Giải thích:

Việc tạo ra dashboard và report trong dự án này là cực kỳ quan trọng vì nó giúp chuyển hóa dữ liệu thô thành thông tin trực quan, dễ hiểu và có giá trị. Dashboard cung cấp cái nhìn tổng quan về các chỉ số quan trọng, giúp quản lý và các bên liên quan nhanh chóng nắm bắt được tình hình hiện tại của doanh nghiệp mà không cần phải đi sâu vào chi tiết dữ liệu. Nó cho phép theo dõi xu hướng, nhận diện các vấn đề tiềm ẩn và đưa ra các quyết định chiến lược một cách nhanh chóng. Report chi tiết, mặt khác, cung cấp một cái nhìn sâu sắc hơn vào các khía cạnh cụ thể của dữ liệu, giúp phân tích chi tiết các yếu tố ảnh hưởng đến doanh số, giá bán, và tình trạng xe. Với report, các nhà phân tích có thể đi sâu vào dữ liệu, tìm hiểu nguyên nhân và đưa ra các giải pháp cụ thể. Việc tạo dashboard và report không chỉ giúp tối ưu hóa quá trình ra quyết định mà còn tăng cường khả năng theo dõi và quản lý hiệu quả các hoạt động kinh doanh của công ty Happy Car, từ đó cải thiện hiệu suất và lợi nhuận.

6.2.2 DASHBOARD

Trong dự án này, chúng ta sẽ tạo ra các loại dashboard khác nhau nhằm phục vụ các mục đích cụ thể và đáp ứng nhu cầu của các bên liên quan. Việc tạo các loại dashboard khác nhau giúp công ty Happy Car có cái nhìn toàn diện và chi tiết về các khía cạnh khác nhau của hoạt động kinh doanh. Mỗi dashboard phục vụ một mục đích cụ thể, từ tổng quan doanh số, tình trạng xe, hiệu suất người bán, phân tích giá bán đến phân tích địa lý. Điều này không chỉ giúp cải thiện hiệu quả quản

lý và ra quyết định mà còn tăng cường khả năng cạnh tranh và phát triển bền vững của công ty.

Dưới đây là các loại dashboard dự kiến và giải thích vì sao cần tạo các dashboard này:

DASHBOARD TỔNG QUAN DOANH SỐ (SALES OVERVIEW DASHBOARD)

Nội dung:

- Biểu đồ đường hiển thị xu hướng giá bán theo thời gian.
- Biểu đồ cột so sánh doanh số bán hàng theo hãng xe.
- Các KPI chính như tổng doanh thu, số lượng xe bán được, giá bán trung bình.

Mục đích: Dashboard này cung cấp một cái nhìn tổng quan về tình hình doanh số của công ty. Việc theo dõi xu hướng doanh số và giá bán theo thời gian giúp quản lý nắm bắt được hiệu quả kinh doanh, nhận diện các mùa cao điểm và các xu hướng thị trường. So sánh doanh số giữa các hãng xe giúp xác định hãng nào đang dẫn đầu và hãng nào cần cải thiện.

DASHBOARD PHÂN TÍCH TÌNH TRẠNG XE (CONDITION ANALYSIS DASHBOARD)

Nội dung:

- Biểu đồ tròn hiển thị tỷ lệ tình trạng xe.
- Biểu đồ phân tán mối quan hệ giữa giá bán và quãng đường đã đi.
- Biểu đồ cột so sánh giá bán theo tình trạng xe.

Mục đích: Việc hiểu rõ tình trạng xe ảnh hưởng như thế nào đến giá bán là rất quan trọng. Dashboard này giúp phân tích sự tác động của tình trạng xe lên giá

bán, từ đó giúp công ty đưa ra các chiến lược định giá phù hợp, cải thiện chất lượng xe hoặc xác định các khu vực cần bảo dưỡng.

DASHBOARD PHÂN TÍCH NGƯỜI BÁN (SELLER ANALYSIS DASHBOARD)

Nội dung:

- Biểu đồ cột hiển thị doanh số bán hàng theo người bán.
- Biểu đồ cột ngang hiển thị số lượng xe bán ra theo người bán.
- Các KPI về hiệu suất bán hàng của từng người bán.

Mục đích: Dashboard này giúp theo dõi hiệu suất bán hàng của từng nhân viên hoặc đại lý, từ đó đánh giá hiệu quả công việc và đưa ra các biện pháp khuyến khích hoặc cải thiện. Nó cũng giúp nhận diện các đại lý hoặc nhân viên có thành tích xuất sắc và chia sẻ kinh nghiệm.

DASHBOARD PHÂN TÍCH GIÁ BÁN (PRICING ANALYSIS DASHBOARD)

Nội dung:

- Biểu đồ bong bóng hiển thị mối quan hệ giữa giá bán, MMR và quãng đường đã đi.
- Biểu đồ cột so sánh giá bán theo tình trạng xe.
- Biểu đồ cột so sánh giá bán theo hãng xe.

Mục đích: Việc phân tích giá bán giúp công ty hiểu rõ các yếu tố ảnh hưởng đến giá cả và đưa ra các quyết định định giá hợp lý. Dashboard này giúp so sánh giá bán giữa các loại xe, tình trạng xe và hãng xe khác nhau, từ đó xác định chiến lược giá phù hợp nhất.

DASHBOARD PHÂN TÍCH ĐỊA LÝ (GEOGRAPHICAL ANALYSIS DASHBOARD)

Nội dung:

- Bản đồ nhiệt hiển thị mật độ bán xe theo khu vực.
- Biểu đồ cột hiển thị doanh số bán hàng theo khu vực.
- Các KPI về doanh số theo từng khu vực địa lý.

Mục đích: Việc hiểu rõ sự phân bố doanh số theo khu vực giúp công ty tối ưu hóa chiến lược tiếp thị và bán hàng. Dashboard này giúp nhận diện các khu vực có doanh số cao, từ đó tập trung nguồn lực vào các khu vực tiềm năng và cải thiện các khu vực có doanh số thấp.

Giải thích:

Việc tạo các dashboard trong dự án này là cực kỳ quan trọng vì chúng cung cấp một cách tiếp cận trực quan và hệ thống để phân tích và hiểu dữ liệu phức tạp của công ty Happy Car. Mỗi dashboard được thiết kế để phục vụ một mục đích cụ thể, giúp quản lý và các bên liên quan nhanh chóng nắm bắt các thông tin quan trọng. Dashboard tổng quan doanh số giúp theo dõi xu hướng và hiệu suất bán hàng, từ đó nhận diện được các mùa cao điểm và điều chỉnh chiến lược kinh doanh kịp thời. Dashboard phân tích tình trạng xe cho phép công ty hiểu rõ hơn về sự ảnh hưởng của tình trạng xe đến giá bán, giúp đưa ra quyết định bảo dưỡng và cải thiện chất lượng xe. Dashboard phân tích người bán cung cấp thông tin về hiệu suất bán hàng của từng nhân viên hoặc đại lý, từ đó khuyến khích và nâng cao hiệu quả công việc. Dashboard phân tích giá bán giúp công ty xác định các yếu tố ảnh hưởng đến giá bán và xây dựng chiến lược giá hợp lý. Cuối cùng, dashboard phân tích địa lý giúp tối ưu hóa chiến lược tiếp thị và phân bổ nguồn lực theo khu vực địa lý, từ đó nâng cao doanh số và hiệu quả kinh doanh. Tổng hợp lại, các

dashboard này không chỉ giúp tăng cường khả năng quản lý và ra quyết định dựa trên dữ liệu mà còn đóng một vai trò quan trọng trong việc cải thiện hiệu suất và cạnh tranh của công ty.

6.2.3 REPORT

Trong dự án này, chúng ta sẽ tạo ra các loại report khác nhau để cung cấp các phân tích chi tiết và trực quan về dữ liệu của công ty Happy Car. Các report này sẽ được chia thành các trang (page) cụ thể, mỗi trang tập trung vào một khía cạnh quan trọng của dữ liệu. Dưới đây là các loại report và nội dung chi tiết của chúng:

REPORT TỔNG QUAN DOANH SỐ (SALES OVERVIEW REPORT)

Nội dung:

- Biểu đồ đường: Hiển thị xu hướng giá bán theo thời gian.
- Biểu đồ cột: So sánh doanh số bán hàng theo hãng xe.
- KPI: Tổng doanh thu, số lượng xe bán được, giá bán trung bình.

Mục đích: Cung cấp cái nhìn tổng quan về tình hình doanh số và hiệu quả kinh doanh của công ty.

REPORT PHÂN TÍCH TÌNH TRẠNG XE (CONDITION ANALYSIS REPORT)

Nội Dung:

- Biểu đồ tròn: Hiển thị tỷ lệ các tình trạng xe (Excellent, Good, Fair).
- Biểu đồ phân tán: Phân tích mối quan hệ giữa giá bán và quãng đường đã đi.
- Biểu đồ cột: So sánh giá bán theo tình trạng xe.

Mục đích: Giúp phân tích và hiểu rõ sự ảnh hưởng của tình trạng xe đến giá bán và hiệu quả kinh doanh.

3. REPORT PHÂN TÍCH NGƯỜI BÁN (SELLER ANALYSIS REPORT)

Nội dung:

- Biểu đồ cột: Hiển thị doanh số bán hàng theo từng người bán.
- Biểu đồ cột ngang: Hiển thị số lượng xe bán ra theo từng người bán.
- KPI: Hiệu suất bán hàng của từng người bán.

Mục đích: Theo dõi hiệu suất bán hàng của nhân viên và đại lý, từ đó đưa ra các biện pháp khuyến khích hoặc cải thiện.

REPORT PHÂN TÍCH GIÁ BÁN (PRICING ANALYSIS REPORT)

Nội dung:

- Biểu đồ bong bóng: Hiển thị mối quan hệ giữa giá bán, MMR và quãng đường đã đi.
- Biểu đồ cột: So sánh giá bán theo tình trạng xe.
- Biểu đồ cột: So sánh giá bán theo hãng xe.

Mục đích: Phân tích các yếu tố ảnh hưởng đến giá bán, giúp đưa ra các chiến lược định giá phù hợp.

REPORT PHÂN TÍCH ĐỊA LÝ (GEOGRAPHICAL ANALYSIS REPORT)

Nội dung:

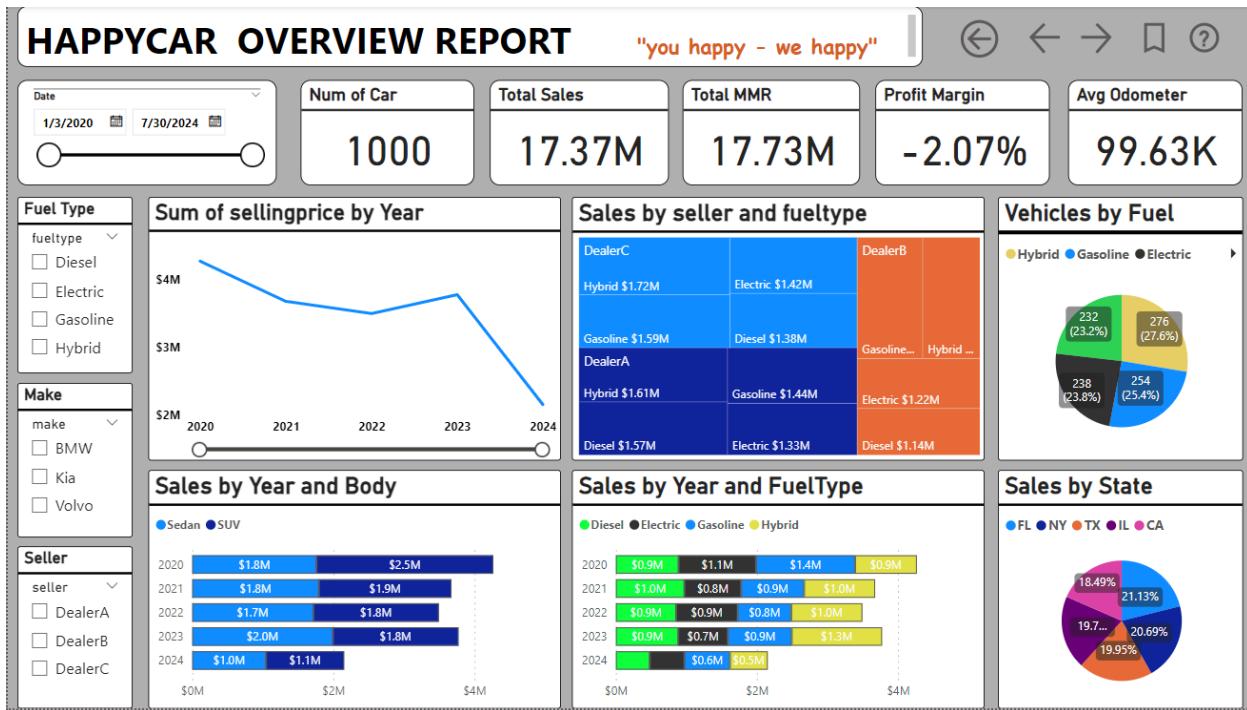
- Bản đồ nhiệt: Hiển thị mật độ bán xe theo khu vực.
- Biểu đồ cột: Hiển thị doanh số bán hàng theo khu vực.
- KPI: Doanh số bán hàng theo từng khu vực địa lý.

Mục đích: Giúp tối ưu hóa chiến lược tiếp thị và phân bổ nguồn lực theo khu vực, nâng cao hiệu quả kinh doanh.

Giải thích:

Sử dụng các report này trong dự án là cần thiết vì chúng cung cấp cái nhìn sâu sắc, chi tiết và có hệ thống về dữ liệu, từ đó hỗ trợ quá trình ra quyết định một cách hiệu quả và chính xác. Report tổng quan doanh số giúp ban quản lý nắm bắt được tình hình kinh doanh chung, theo dõi các chỉ số quan trọng như doanh thu, số lượng xe bán được, và giá bán trung bình, từ đó nhận diện xu hướng và điều chỉnh chiến lược kịp thời. Report phân tích tình trạng xe giúp công ty hiểu rõ ảnh hưởng của tình trạng xe đến giá bán, giúp tối ưu hóa chiến lược bảo dưỡng và định giá. Report phân tích người bán cho phép đánh giá hiệu suất bán hàng của từng nhân viên và đại lý, từ đó đưa ra các biện pháp khuyến khích hoặc cải thiện hiệu quả công việc. Report phân tích giá bán giúp xác định các yếu tố ảnh hưởng đến giá cả, từ đó xây dựng chiến lược định giá hợp lý và cạnh tranh. Cuối cùng, report phân tích địa lý cung cấp thông tin về sự phân bố doanh số theo khu vực, giúp tối ưu hóa chiến lược tiếp thị và phân bổ nguồn lực một cách hiệu quả. Tóm lại, các report này không chỉ giúp cải thiện khả năng quản lý và ra quyết định dựa trên dữ liệu mà còn góp phần quan trọng trong việc nâng cao hiệu suất và khả năng cạnh tranh của công ty Happy Car.

6.2.3.1 Tạo report Detail



Hình 6.1 Report Overview

Giải thích:

Báo cáo tổng quan của HappyCar cung cấp cái nhìn toàn diện về hiệu suất kinh doanh của công ty trong khoảng thời gian từ ngày 1/3/2020 đến ngày 30/7/2024. Với tiêu đề "HAPPYCAR OVERVIEW REPORT" và khẩu hiệu "you happy - we happy," báo cáo này tập trung vào việc phân tích các khía cạnh chính của hoạt động kinh doanh xe ô tô đã qua sử dụng.

Phần đầu của báo cáo hiển thị các chỉ số quan trọng như tổng doanh thu bán hàng (\$17.37 triệu USD), tổng giá trị thị trường MMR (\$17.73 triệu USD), lợi nhuận biên (-2.07%), và số dặm trung bình của xe đã đi (99.63K). Các chỉ số này giúp ban lãnh đạo đánh giá tổng quan về tình hình kinh doanh và hiệu suất hoạt động của công ty.

Biểu đồ đường trong phần "Sum of Sellingprice by Year" cho thấy xu hướng tổng giá bán thực tế giảm dần qua các năm, từ 2020 đến 2024. Điều này có thể gợi ý sự thay đổi trong nhu cầu thị trường hoặc chiến lược giá của công ty.

Biểu đồ cột trong phần "Sum of Sellingprice by Seller" so sánh hiệu suất bán hàng của ba người bán chính (DealerA, DealerB, DealerC). DealerC có tổng giá bán cao nhất, tiếp theo là DealerA và DealerB, cho thấy DealerC có thể có chiến lược bán hàng hiệu quả hơn hoặc tiếp cận được nhiều khách hàng hơn.

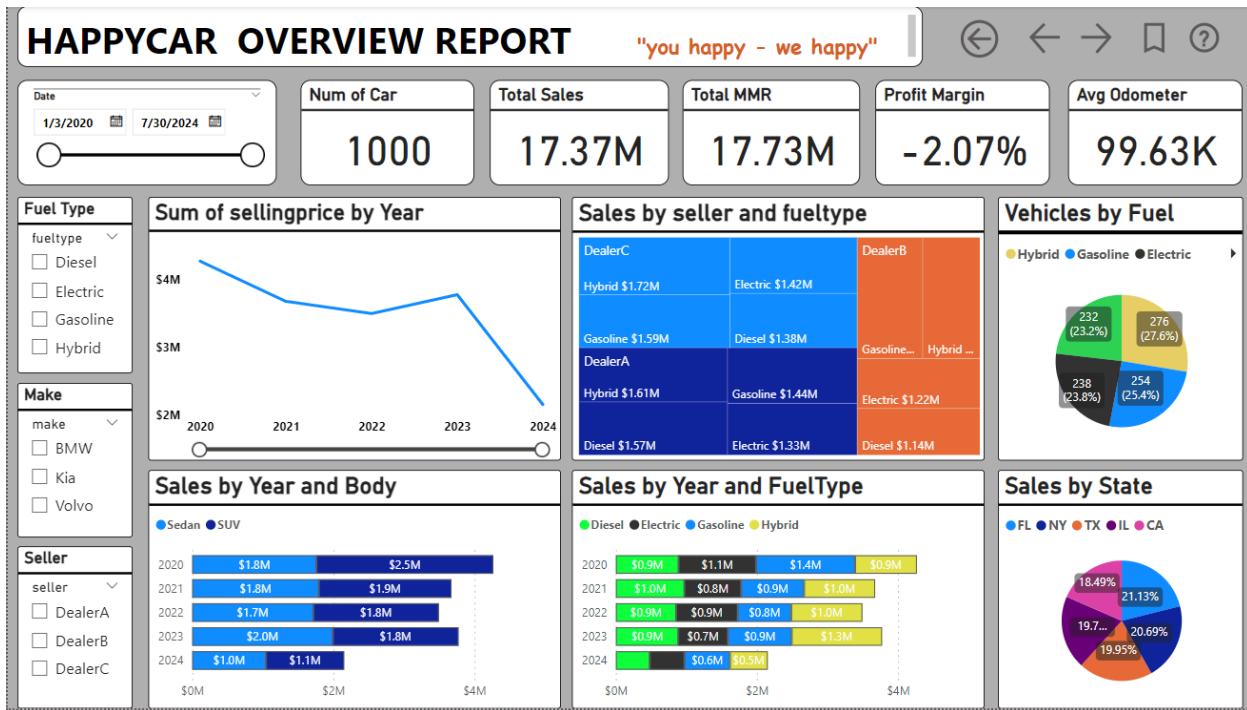
Biểu đồ tròn trong phần "Vehicles by Fuel" hiển thị tỷ lệ các loại xe theo nhiên liệu. Xe điện chiếm tỷ lệ cao nhất (27.6%), tiếp theo là xe xăng (25.4%) và xe hybrid (23.2%), cho thấy xu hướng chuyển dịch sang các loại nhiên liệu thân thiện với môi trường.

Biểu đồ cột xếp chồng trong phần "Sales by Year and Body" và "Sales by Year and FuelType" phân tích doanh thu theo từng năm, loại xe (Sedan, SUV), và loại nhiên liệu. Các biểu đồ này giúp xác định loại xe và nhiên liệu nào bán chạy nhất theo từng năm, hỗ trợ trong việc lập kế hoạch nhập hàng và marketing.

Biểu đồ tròn trong phần "Vehicles by State" phân tích doanh thu theo từng bang, cho thấy California (CA) có tổng doanh thu cao nhất (21.13%), tiếp theo là Florida (FL) và Texas (TX). Điều này giúp công ty nhận diện các khu vực có nhu cầu cao và tập trung nguồn lực vào các thị trường tiềm năng.

Nhìn chung, báo cáo tổng quan của HappyCar cung cấp những thông tin chi tiết và phân tích quan trọng, giúp ban lãnh đạo có cái nhìn toàn diện về hiệu suất kinh doanh và đưa ra các quyết định chiến lược dựa trên dữ liệu thực tế.

6.2.3.2 Tạo report Overview



Hình 6.2 Report Overview có thêm Buttons

Giải thích:

Có thêm các nút bấm thực hiện các chức năng khác nhau, buttons là công cụ mạnh mẽ giúp cải thiện trải nghiệm người dùng và tăng tính tương tác trong báo cáo. Buttons cho phép người dùng thực hiện các hành động cụ thể, chẳng hạn như điều hướng giữa các trang báo cáo, làm mới dữ liệu, hoặc áp dụng bộ lọc. Bạn có thể tùy chỉnh hình dạng, màu sắc, và nội dung hiển thị trên buttons để phù hợp với giao diện tổng thể của báo cáo. Ngoài ra, buttons có thể được thiết lập với các hành động như "Back", "Drillthrough", "Bookmark", và "Page Navigation" để cung cấp trải nghiệm người dùng mượt mà và liền mạch. Việc sử dụng buttons giúp tạo ra báo cáo trực quan và dễ sử dụng, đồng thời giúp người dùng dễ dàng truy cập và phân tích dữ liệu một cách hiệu quả hơn.

6.2.4 BOOKMARK

6.2.4.1 Tạo bookmark slicer chọn tất cả các ngày

The screenshot shows the Power BI Desktop interface. The ribbon is set to 'Modeling'. The 'View' tab is highlighted with a red arrow. The 'Bookmarks' icon in the ribbon is also highlighted with a red box and arrow. The 'Bookmarks' pane on the right shows a bookmark named 'Tất cả các ngày' which is selected and highlighted with a red box.

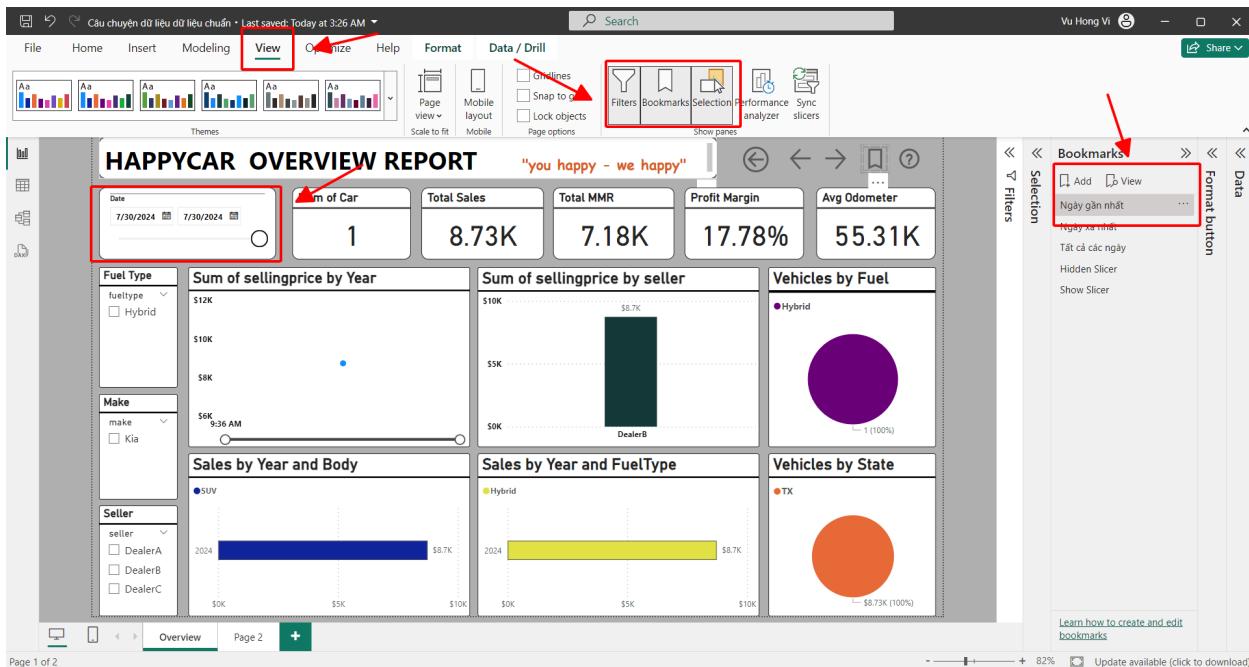
Hình 6.3 Tạo bookmark slicer chọn tất cả các ngày

The screenshot shows the Power BI Desktop interface. The ribbon is set to 'Insert'. The 'Buttons' icon in the ribbon is highlighted with a red box and arrow. The 'Bookmarks' pane on the right shows a bookmark named 'Tất cả các ngày' which is selected and highlighted with a red box. The 'Action' section of the ribbon shows 'Type: Bookmark' and 'Value: Tất cả các ngày' both highlighted with red boxes. The 'Tooltip' section shows 'Text: Hiện thông tin tất cả' highlighted with a red box.

Hình 6.4 Tạo buttons hiện tất cả các ngày

Giải thích: Giúp người xem có thể quay về trạng thái hiển thị thông tin tất cả các ngày (trạng thái ban đầu) một cách nhanh nhất.

6.2.4.2 Tạo bookmark slicer chọn ngày gần nhất



Hình 6.5 Tạo bookmark slicer chọn ngày gần nhất

Giải thích:

Chức năng bookmark slicer chọn ngày gần nhất cho phép bạn lưu trạng thái cụ thể của một báo cáo, bao gồm các lựa chọn bộ lọc (slicer) hiện tại. Điều này đặc biệt hữu ích khi bạn muốn người dùng có thể nhanh chóng chuyển đến một thời điểm cụ thể, chẳng hạn như ngày gần nhất có dữ liệu. Bằng cách sử dụng bookmark, bạn có thể thiết lập và lưu lại trạng thái của slicer tại ngày gần nhất và sau đó dễ dàng áp dụng lại trạng thái này bằng cách nhấp vào bookmark. Việc này không chỉ tiết kiệm thời gian mà còn đảm bảo rằng người dùng luôn nhìn thấy dữ liệu mới nhất mà không cần phải điều chỉnh thủ công mỗi lần truy cập vào báo cáo. Bookmark slicer chọn ngày gần nhất giúp cải thiện trải nghiệm người dùng và đảm bảo tính nhất quán trong việc truy xuất dữ liệu.

6.2.4.3 Tạo bookmark slicer chọn ngày xa nhất

The screenshot shows the Power BI desktop interface with the following highlights:

- View tab:** The "View" tab is highlighted with a red box. A red arrow points from the "View" tab to the "Bookmarks" icon in the ribbon toolbar.
- Toolbar:** The ribbon toolbar includes icons for Gridlines, Snap to grid, Lock objects, Page options, Show panes, Filters, Bookmarks, Selection, Performance analyzer, Sync slicers, and Data.
- Report Content:** The main area displays the "HAPPYCAR OVERVIEW REPORT". It features a Date slicer set to "1/3/2020" and a "Num Of Car" value of "3". Below the slicer are several visualizations: a bar chart for "Sum of sellingprice by Year", a bar chart for "Sum of sellingprice by seller", and a pie chart for "Vehicles by Fuel". To the left are filters for Fuel Type (Gasoline, Hybrid), Make (Kia, Volvo), and Seller (DealerA, DealerB, DealerC). Below these are charts for "Sales by Year and Body" and "Sales by Year and FuelType".
- Filters Panel:** The "Filters" pane on the right lists existing bookmarks: "Ngày gần nhất" and "Ngày xa nhất". A red box highlights the "Add" button for creating new bookmarks.
- Bookmarks Panel:** The "Bookmarks" pane on the far right shows the newly created bookmark "Ngày xa nhất". It includes options to "View" the bookmark, "Format button", and "Delete".

Hình 6.6 Tạo bookmark slicer chọn ngày xa nhất

Giải thích:

Bookmark slicer chọn ngày gần nhất là một tính năng hữu ích giúp người dùng nhanh chóng truy cập dữ liệu tại thời điểm mới nhất. Bằng cách sử dụng bookmark, bạn có thể lưu lại trạng thái của slicer khi chọn ngày gần nhất và dễ dàng quay lại trạng thái này bất cứ khi nào cần. Điều này đảm bảo rằng người dùng luôn thấy được dữ liệu cập nhật nhất mà không cần điều chỉnh slicer mỗi lần truy cập vào báo cáo. Bookmark slicer chọn ngày gần nhất giúp tiết kiệm thời gian, tăng tính tiện lợi, và đảm bảo tính nhất quán trong việc xem và phân tích dữ liệu mới nhất.

7 KẾT LUẬN

7.1 BÁO CÁO

7.1.1 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.

BUỚC 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.

BUỚC 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.

BUỚC 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.

BUỚC 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.

BUỚC 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.

7.1.2 TỔNG HỢP

Sau khi phân tích bộ dữ liệu, chúng tôi đã kiểm chứng các giả thuyết đặt ra và thu được các kết quả quan trọng. Thứ nhất, xe sản xuất gần đây có xu hướng có giá bán cao hơn, khẳng định giả thuyết rằng xe mới có giá cao hơn xe cũ. Thứ hai, các hãng xe phổ biến như Toyota và Honda có giá bán trung bình cao hơn so với các hãng xe ít phổ biến, xác nhận rằng hãng xe phổ biến có giá bán cao hơn. Thứ ba, xe có quãng đường đã đi ít hơn thực sự có giá bán cao hơn, cho thấy quãng đường đã đi ảnh hưởng mạnh đến giá bán. Thứ tư, xe ở tình trạng tốt hơn (Excellent) có giá bán cao hơn so với xe ở tình trạng trung bình (Good) và kém (Fair), hỗ trợ giả thuyết rằng tình trạng xe tốt có giá bán cao hơn. Cuối cùng, xe có

màu sắc phổ biến như trắng, đen, và bạc có giá bán cao hơn so với các màu sắc ít phổ biến, xác minh giả thuyết rằng màu xe phổ biến có giá bán cao hơn. Những phát hiện này giúp Happy Car hiểu rõ hơn về các yếu tố ảnh hưởng đến giá bán xe, từ đó đưa ra các chiến lược kinh doanh và định giá hợp lý.

7.2 KHÓ KHĂN

Trong quá trình thực hiện dự án phân tích dữ liệu cho công ty Happy Car, 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. Việc đảm bảo báo cáo phân tích có tính chính xác và thuyết phục, đồng thời dễ hiểu đối với các bên liên quan, yêu cầu sự cẩn trọng trong từng bước phân tích và trình bày. 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.

7.3 THUẬN LỢI

Mặc dù gặp nhiều khó khăn, quá trình thực hiện dự án phân tích dữ liệu cho công ty Happy Car cũng có nhiều thuận lợi đáng kể. Đầu tiên, sự hỗ trợ nhiệt tình từ ban quản lý và các bên liên quan đã tạo điều kiện thuận lợi cho việc thu thập và

truy cập dữ liệu cần thiết. Hơn nữa, 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. Các công cụ phân tích dữ liệu mạnh mẽ như Power BI và các phần mềm hỗ trợ khác đã giúp tối ưu hóa quá trình phân tích và trực quan hóa dữ liệu. Ngoài ra, sự phối hợp tốt giữa các thành viên trong nhóm, cùng với sự hướng dẫn tận tình từ các chuyên gia cố vấn, đã giúp dự án tiến 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 công ty Happy Car.

7.4 HƯỚNG PHÁT TRIỂN

Trong tương lai, công ty Happy Car sẽ tiếp tục phát triển và mở rộng việc ứng dụng phân tích dữ liệu vào các khía cạnh khác của hoạt động kinh doanh để tối ưu hóa hiệu quả và cạnh tranh trên thị trường. Trước hết, chúng tôi sẽ đầu tư vào cơ sở hạ tầng công nghệ, nâng cấp các công cụ phân tích dữ liệu và phần mềm quản lý để tăng cường khả năng xử lý và trực quan hóa dữ liệu. Tiếp theo, Happy Car sẽ tập trung vào đào tạo và phát triển đội ngũ nhân viên, nâng cao kỹ năng phân tích dữ liệu và làm việc nhóm, đảm bảo rằng tất cả thành viên đều có thể đóng góp một cách hiệu quả vào các dự án tương lai. Bên cạnh đó, công ty sẽ mở rộng phạm vi thu thập dữ liệu, không chỉ giới hạn ở các thông tin hiện tại mà còn bao gồm các dữ liệu mới như xu hướng thị trường, phản hồi của khách hàng, và dữ liệu cạnh tranh. Cuối cùng, Happy Car sẽ xây dựng các mô hình dự báo tiên tiến để dự đoán xu hướng và hỗ trợ quyết định kinh doanh chiến lược, nhằm tăng cường khả năng đáp ứng nhanh chóng và hiệu quả đối với những thay đổi của thị trường.

Với những định hướng này, Happy Car không chỉ nâng cao năng lực cạnh tranh mà còn khẳng định vị thế của mình trong ngành kinh doanh xe ô tô cũ.