



**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG VIỆT - HÀN
KHOA KHOA HỌC MÁY TÍNH**

Xây dựng Kho dữ liệu: Phân tích và Dự đoán Kết quả Xổ số

Báo cáo Đồ án môn học Kho Dữ Liệu

Sinh viên thực hiện:

Lê Thế Dũng (22IT049)
Nguyễn Thị Tố Uyên (22NS082)
Phạm Minh Nhật (23IT.B152)
Nguyễn Thị Bình Minh (23IT.B133)

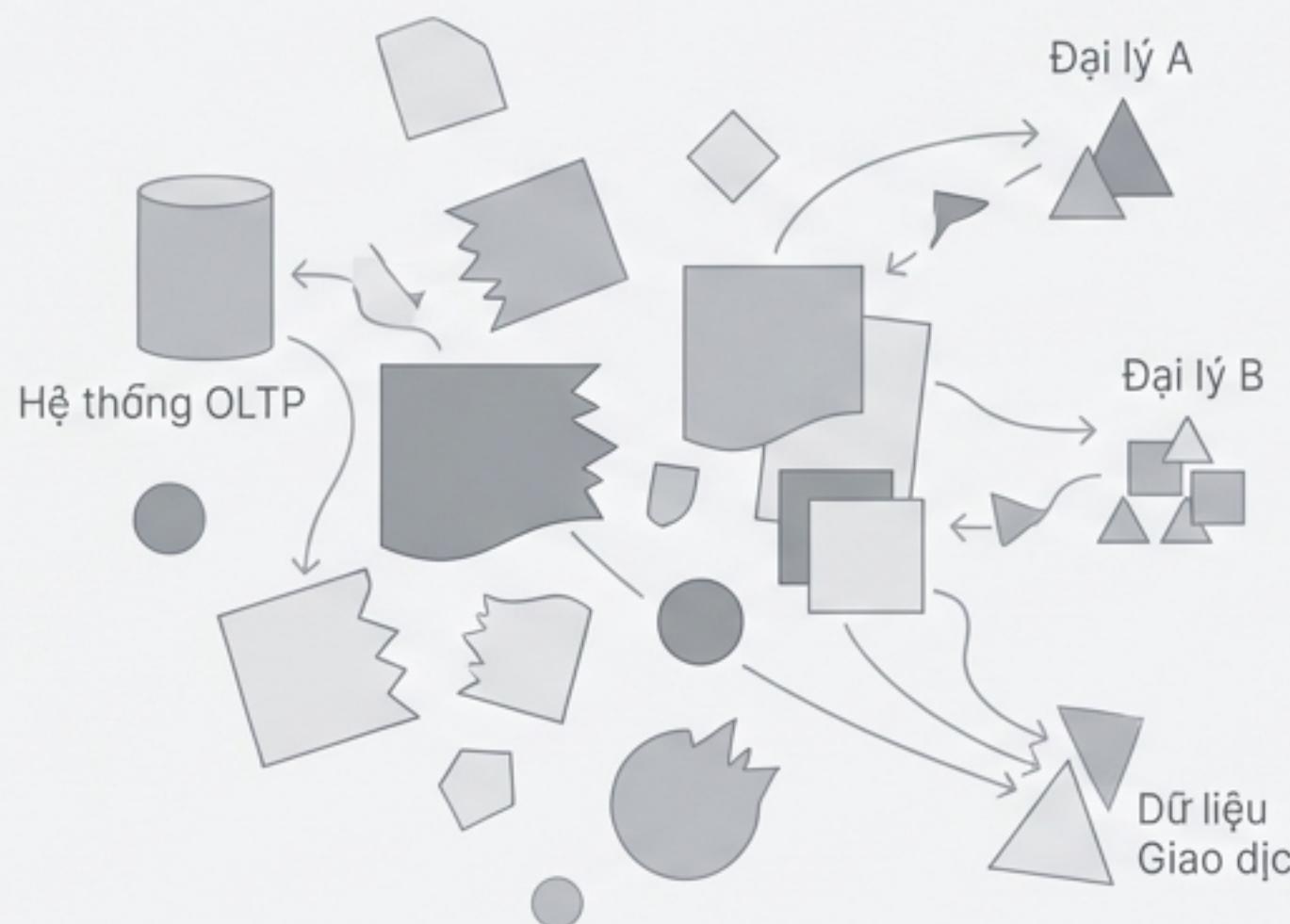
Giảng viên hướng dẫn:

ThS. Trần Thanh Liêm

Đà Nẵng, tháng 12 năm 2025

Bối cảnh: Tiềm năng ẩn giấu trong Dữ liệu Xổ số

Ngành xổ số Việt Nam tạo ra khối lượng dữ liệu khổng lồ hàng ngày, nhưng tiềm năng của nó chưa được khai thác triệt để.



Key Challenges

- Dữ liệu Phân mảnh:** Dữ liệu phát sinh từ nhiều đài và đại lý, được lưu trữ trong các hệ thống OLTP truyền thống, khó khăn cho việc phân tích tổng thể.
- Thiếu Phân tích Sâu:** Các hệ thống hiện tại chủ yếu phục vụ ghi nhận giao dịch, không được thiết kế cho việc phân tích xu hướng lịch sử, hành vi người chơi, hay dự báo.
- Bỏ lỡ Cơ hội Kinh doanh:** Không có công cụ để trả lời các câu hỏi chiến lược như:
 - Xu hướng doanh thu biến động theo mùa vụ ra sao?
 - Các con số 'nóng' và 'lạnh' ảnh hưởng đến hành vi mua vé như thế nào?
 - Làm thế nào để dự báo doanh thu và tối ưu hóa kế hoạch kinh doanh?

The Opportunity

Xây dựng một kho dữ liệu chuyên biệt để biến dữ liệu thô thành tài sản chiến lược, hỗ trợ ra quyết định thông minh.

Sứ mệnh Dự án: Xây dựng Nền tảng Phân tích và Dự báo Thông minh

Thiết kế và triển khai một hệ thống kho dữ liệu hoàn chỉnh, có khả năng phân tích đa chiều và tích hợp học máy để dự đoán xu hướng kinh doanh trong ngành xổ số.



1. Thiết kế Kho dữ liệu Chuẩn mực

- Xây dựng kho dữ liệu theo mô hình Lược đồ Sao (Star Schema), tối ưu cho các truy vấn phân tích (OLAP).



2. Tự động hóa Dòng chảy Dữ liệu

- Phát triển quy trình ETL (Extract, Transform, Load) hoàn chỉnh để tích hợp, làm sạch và nạp dữ liệu lịch sử 10 năm.



3. Khai phá Dữ liệu Đa chiều

- Triển khai hơn 21 phương thức truy vấn chuyên sâu để phân tích doanh thu và kết quả xổ số từ nhiều góc độ.



4. Tích hợp Trí tuệ Dự báo

- Ứng dụng mô hình Machine Learning (Hồi quy Tuyến tính) để dự báo doanh thu 12 tháng tới.



5. Phân tích Tương quan Chéo

- Nghiên cứu mối liên hệ giữa kết quả xổ số ngày T và doanh thu bán vé ngày T+1 để tìm ra insights kinh doanh.

Bản thiết kế Hệ thống: Kiến trúc Tổng quan



1. Nguồn dữ liệu

Dữ liệu đầu vào từ các file CSV, mô phỏng dữ liệu kết quả xổ số và doanh thu.



2. Tầng ETL

Quy trình trích xuất, biến đổi và nạp dữ liệu được thực thi bằng Python và Pandas để làm sạch và chuẩn hóa.



3. Kho dữ liệu

Trái tim của hệ thống, được xây dựng trên SQLite với Lược đồ Sao, chứa các bảng Dimension và Fact.



4. Tầng Phân tích

Lớp logic chứa hơn 21 phương thức truy vấn SQL để thực hiện các phân tích OLAP.



5. Tầng Dự báo ML

Tích hợp mô hình Hồi quy Tuyến tính (Scikit-learn) để dự đoán doanh thu.



6. Tầng Giao diện

Dashboard tương tác được xây dựng bằng Streamlit để trực quan hóa dữ liệu và kết quả.

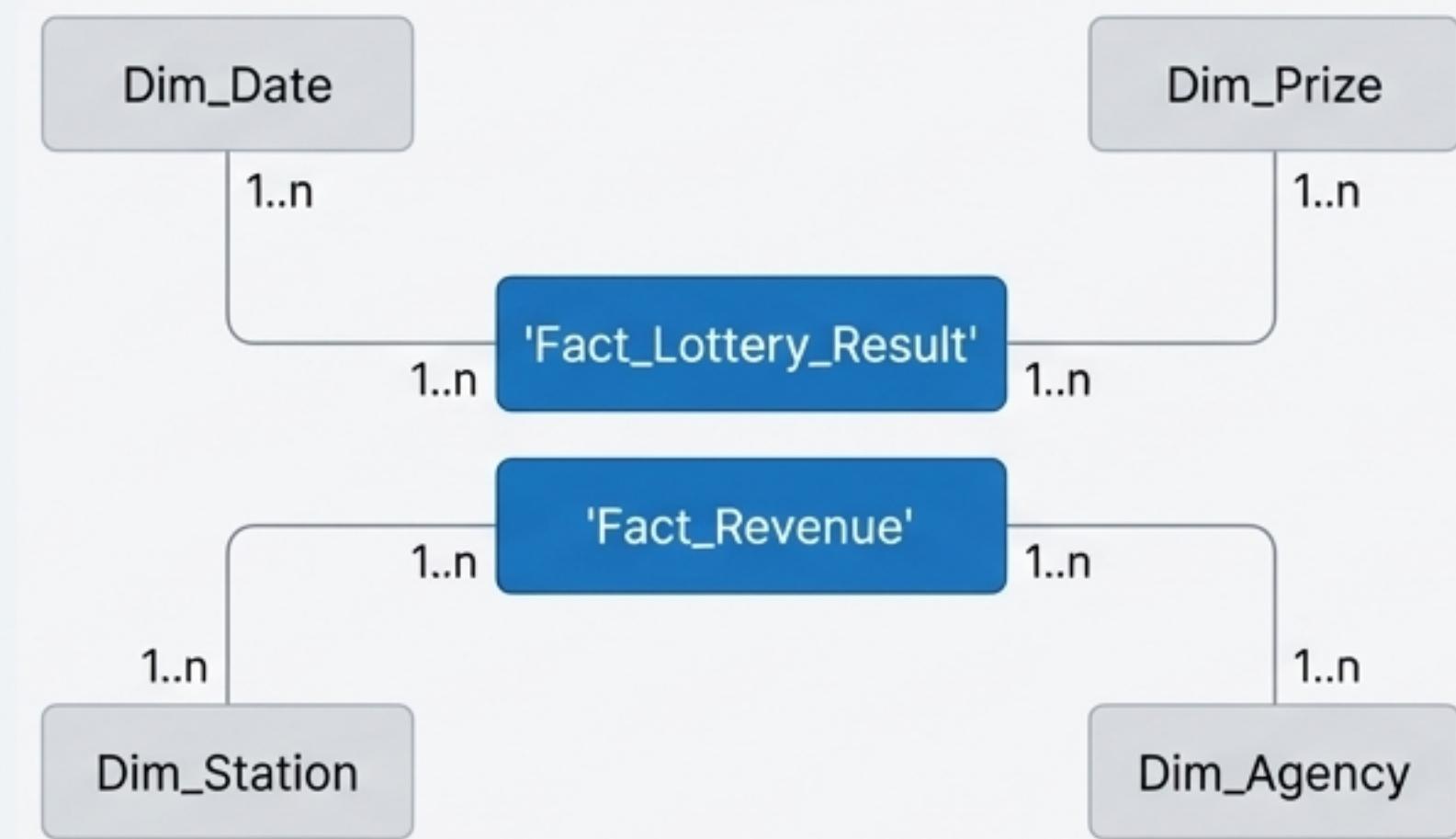
Trái tim Kho dữ liệu: Tại sao chọn Lược đồ Sao?

The Rationale (The "Why")

Lược đồ Sao được lựa chọn vì những ưu điểm vượt trội cho mục đích phân tích:

- Hiệu năng Truy vấn Cao:** Cấu trúc phi chuẩn hóa (denormalized) giúp giảm số lượng phép nối (join), tăng tốc độ truy vấn đáng kể.
- Dễ hiểu và Trực quan:** Mô hình logic rõ ràng, giúp người dùng nghiệp vụ dễ dàng hiểu và tự viết các câu truy vấn phân tích.
- Tối ưu cho OLAP:** Hỗ trợ xuất sắc các thao tác phân tích đa chiều như Roll-up, Drill-down, Slice, và Dice.
- Linh hoạt Mở rộng:** Dễ dàng thêm các chiều phân tích (Dimension) hoặc chỉ số đo lường (Measure) mới trong tương lai.

The Blueprint (The "What")



Sơ đồ thể hiện một bảng Fact trung tâm (Fact_Lottery_Result, Fact_Revenue) được bao quanh bởi các bảng Dimension (Dim_Date, Dim_Station, Dim_Prize, Dim_Agency).

Xây dựng Ngũ cảnh: Các Chiều Dữ liệu (Dimensions)

Các bảng Dimension cung cấp “ai, cái gì, khi nào, ở đâu” cho dữ liệu, cho phép phân tích từ nhiều góc độ.



1. Dim_Date (Chiều Thời gian)

- Mục đích:** Phân tích theo ngày, tuần, tháng, quý, năm và các thuộc tính thời gian (cuối tuần, đầu tháng).
- Thuộc tính chính:** `date_id` (PK), `full_date`, `day_of_week`, `month`, `year`, `is_weekend`



2. Dim_Station (Chiều Đài phát hành)

- Mục đích:** Phân tích theo khu vực địa lý (Bắc, Trung, Nam) và so sánh hiệu suất giữa các đài.
- Thuộc tính chính:** `station_id` (PK), `station_name`, `region`



3. Dim_Prize (Chiều Giải thưởng)

- Mục đích:** Phân tích dữ liệu theo loại giải, giá trị, và số lượng giải.
- Thuộc tính chính:** `prize_id` (PK), `prize_name`, `quantity`, `prize_value`



4. Dim_Agency (Chiều Đại lý)

- Mục đích:** Phân tích hiệu quả kinh doanh của các kênh phân phối và loại đại lý (Cấp 1, Cấp 2, Cấp 3).
- Thuộc tính chính:** `agency_id` (PK), `agency_name`, `agency_type`

Ghi nhận Sự kiện: Các Bảng Dữ kiện (Facts)

Các bảng Fact chứa các chỉ số đo lường (measures) định lượng và là nơi các sự kiện kinh doanh được ghi lại.

1. Fact_Lottery_Result (Kết quả Xổ số)

- **Mục đích:** Lưu trữ chi tiết từng con số trúng thưởng trong mỗi kỳ quay.
- **Độ chi tiết (Granularity):** Mỗi bản ghi đại diện cho **một con số** của **một giải thưởng cụ thể** tại **một đài vào một ngày**.
- **Chỉ số chính (Measures):** `result_number`, `prize_sequence`.
- **Khóa ngoại (Foreign Keys):** `date_id`, `station_id`, `prize_id`.

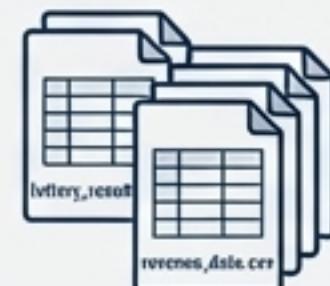
2. Fact_Revenue (Doanh thu Bán vé)

- **Mục đích:** Ghi nhận hoạt động kinh doanh, bao gồm doanh thu, lợi nhuận và số vé bán ra.
- **Độ chi tiết (Granularity):** Mỗi bản ghi đại diện cho **tổng giao dịch** của **một đại lý** tại **một đài** trong **một ngày**.
- **Chỉ số chính (Measures):** `tickets_sold`, `total_revenue`, `net_profit`, `commission`.
- **Khóa ngoại (Foreign Keys):** `date_id`, `station_id`, `agency_id`.

Dòng chảy Dữ liệu Thông minh: Quy trình ETL

Quy trình ETL là xương sống của việc nạp dữ liệu, **đảm bảo dữ liệu trong kho luôn chính xác**, nhất quán và sẵn sàng để phân tích.

Extract (Trích xuất)



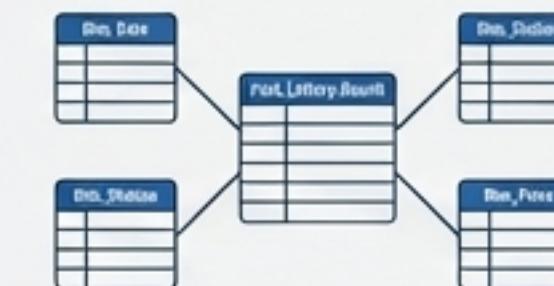
- Nguồn:** Đọc dữ liệu từ hai file nguồn `lottery_results.csv` và `revenue_data.csv`.
- Công cụ:** Sử dụng thư viện `Pandas` trong Python.
- Hành động:** Tải dữ liệu vào bộ nhớ, kiểm tra định dạng và tính toàn vẹn cơ bản của file.

Transform (Biến đổi)



- Đây là bước quan trọng nhất.
- Làm sạch:** Xử lý giá trị thiếu, loại bỏ bản ghi trùng lặp.
 - Chuẩn hóa:** Chuyển đổi kiểu dữ liệu (vd: string sang datetime), tạo các thuộc tính phái sinh cho Dim_Date (vd: `is_weekend`).
 - Ánh xạ:** Tạo các bảng Dimension và ánh xạ các giá trị nghiệp vụ sang khóa thay thế (surrogate key).

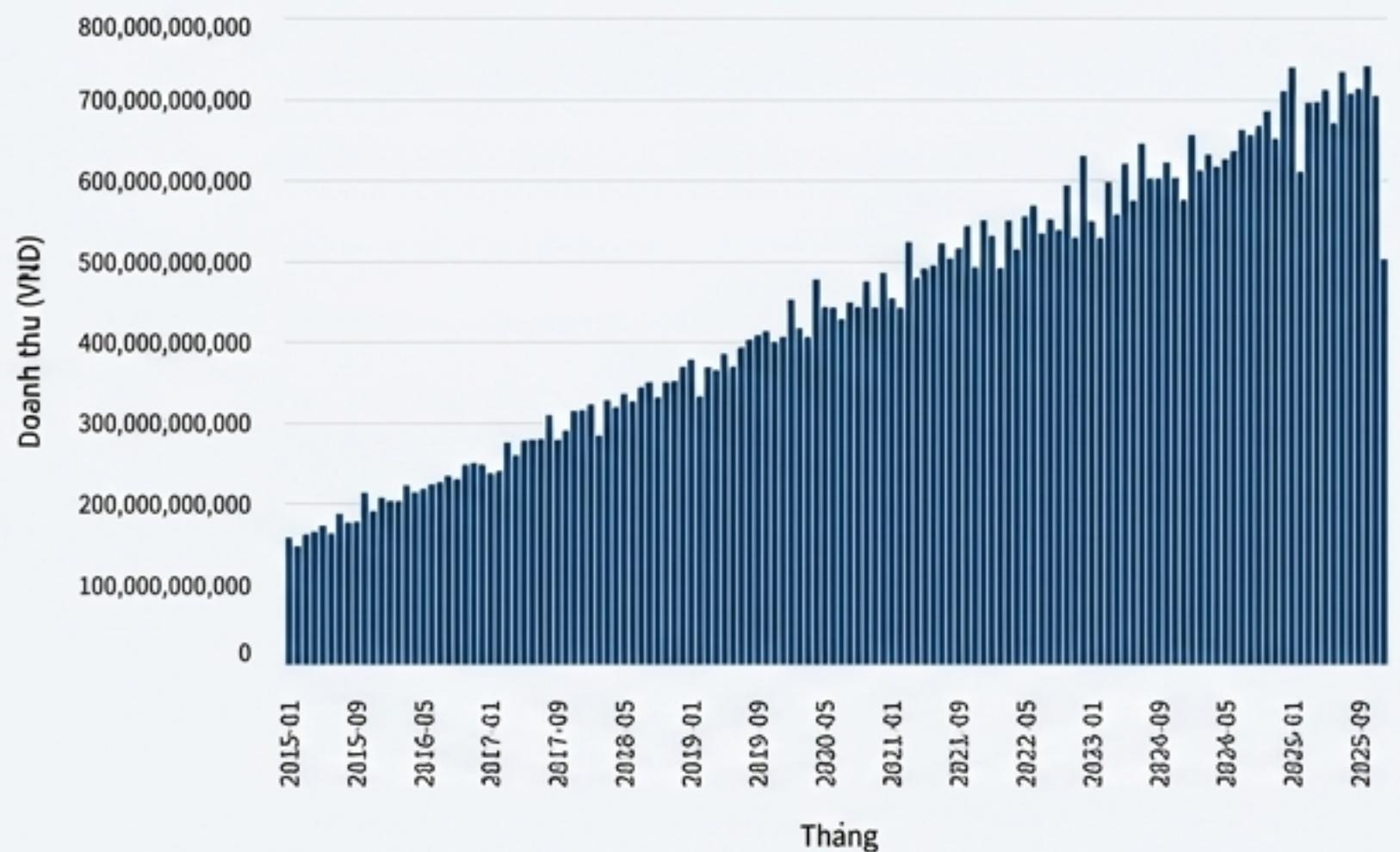
Load (Nạp)



- Chiến lược:** "Load Dimensions First, Then Facts" để đảm bảo tính toàn vẹn tham chiếu.
- Cơ chế:** Áp dụng cơ chế nạp gia tăng (incremental load), chỉ nạp dữ liệu mới để tối ưu hiệu năng.
- Đích:** Nạp các DataFrame đã được xử lý vào kho dữ liệu SQLite.

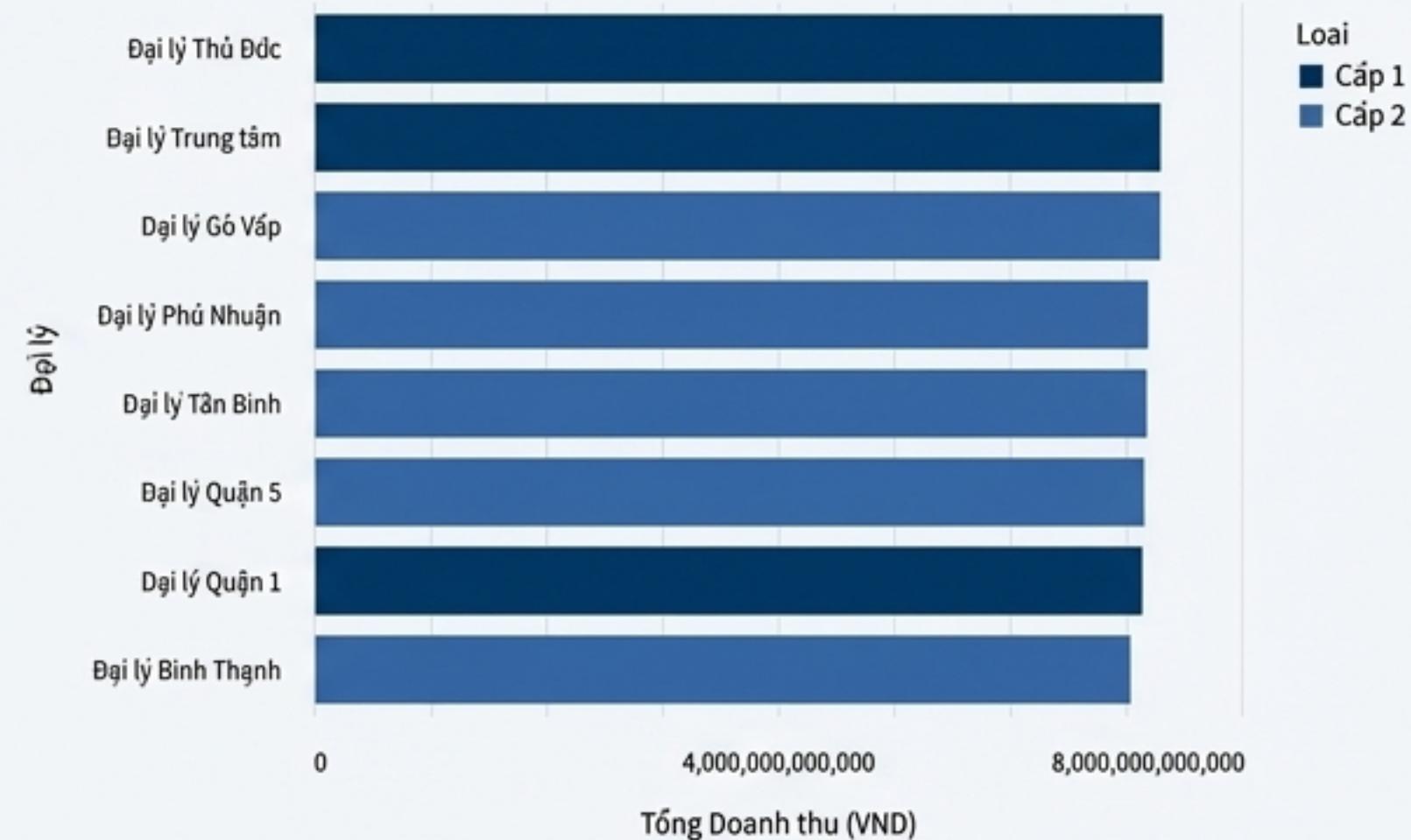
Khai phá Insights: Phân tích Doanh thu Trực quan

Tăng trưởng Doanh thu Bền vững qua Từng tháng



Biểu đồ cho thấy một xu hướng tăng trưởng rõ rệt và ổn định qua các năm, với các biến động theo mùa vụ thể hiện qua các đỉnh và đáy trong từng năm.

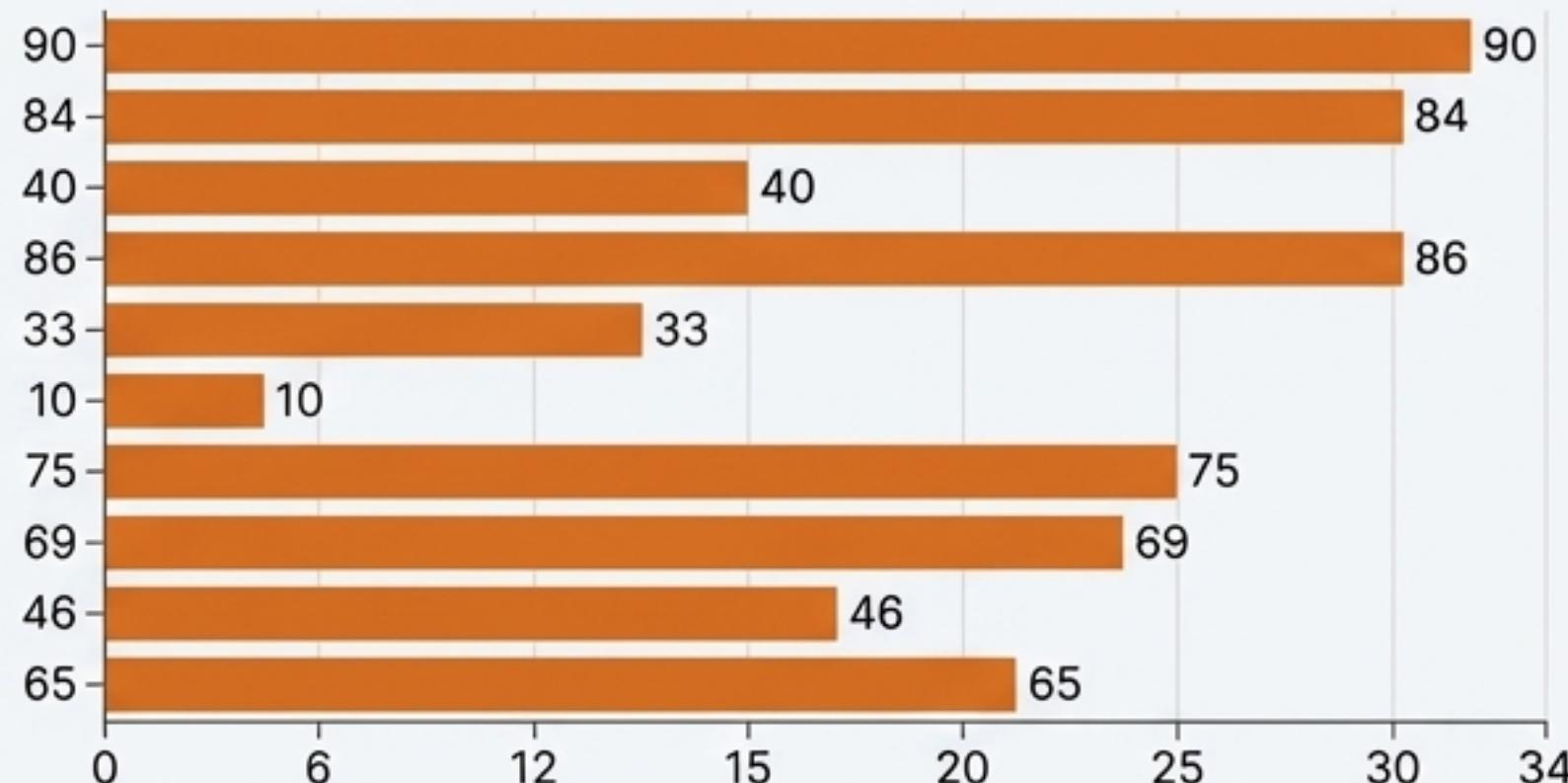
Xác định các Đại lý Chủ lực Thúc đẩy Doanh thu



Các đại lý Cấp 1 (màu xanh đậm) chiếm phần lớn top đầu, cho thấy vai trò quan trọng của kênh phân phối quy mô lớn. Dữ liệu này giúp nhận diện đối tác chiến lược.

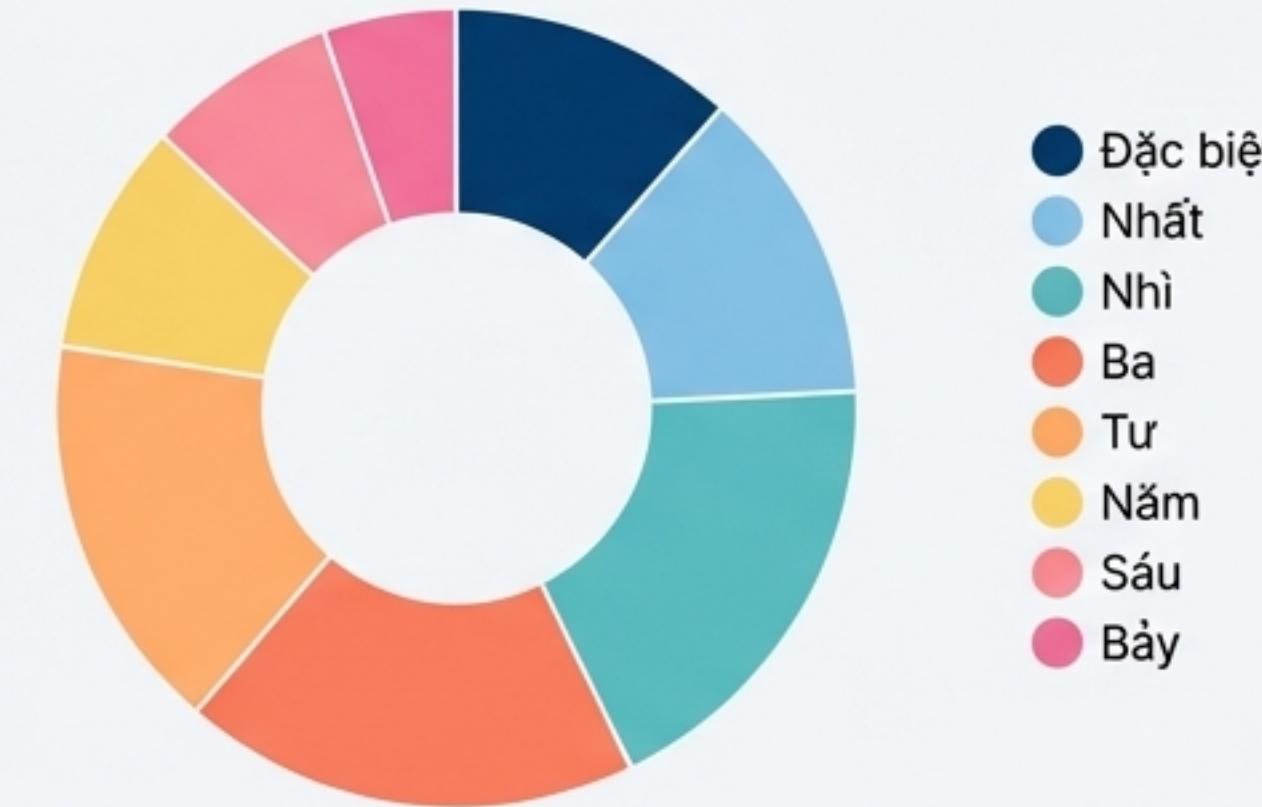
Giải mã Mẫu số: Phân tích Kết quả Xổ số

Khám phá các Con số "Nóng" và "Lạnh"



Biểu đồ xác định các con số xuất hiện thường xuyên ("nóng", ví dụ: 90, 84) và các con số ít xuất hiện ("lạnh"). Thông tin này cung cấp góc nhìn thống kê thú vị về xu hướng gần đây.

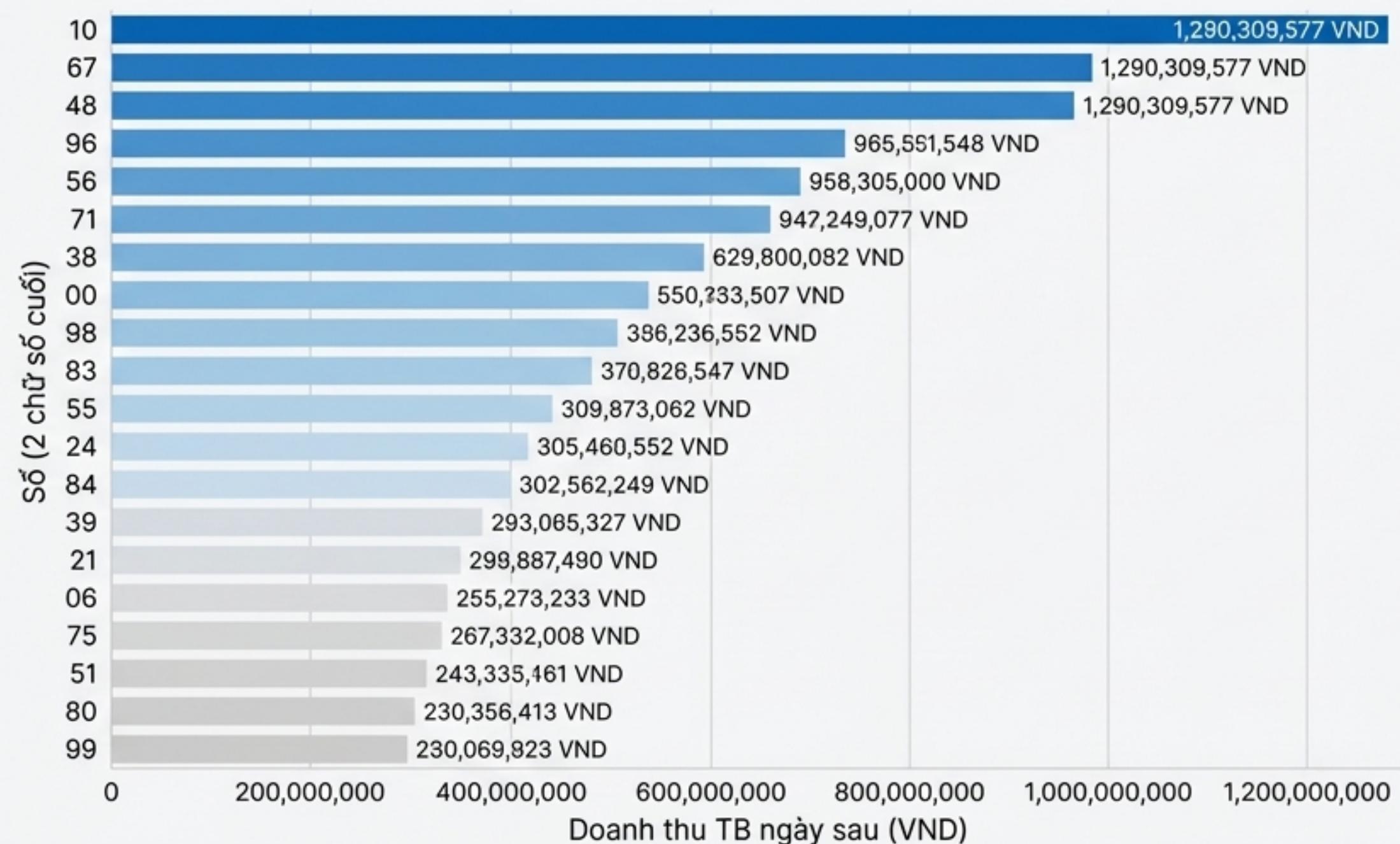
Phân bổ Giải thưởng - Đảm bảo Tính Cân bằng và Hấp dẫn



Cấu trúc giải thưởng được phân bổ hợp lý, với các giải có giá trị thấp chiếm số lượng lớn, trong khi giải đặc biệt là hiếm nhất. Điều này duy trì sự hấp dẫn và tính công bằng của trò chơi.

Tương quan Chéo: Kết quả Xổ số hôm nay Ảnh hưởng Doanh thu ngày mai?

Hệ thống cho phép phân tích một câu hỏi kinh doanh quan trọng: Liệu kết quả xổ số của ngày T có ảnh hưởng đến doanh thu bán vé vào ngày T+1 không?



Có sự ảnh hưởng

Một số con số có tương quan dương mạnh với doanh thu ngày hôm sau.

Số có Impact cao nhất

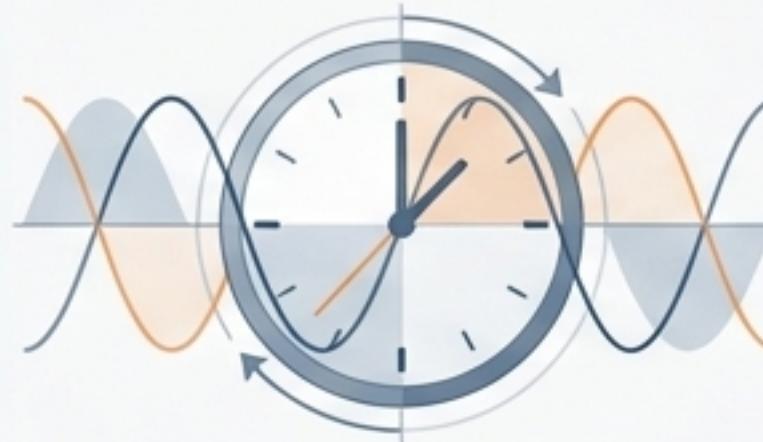
Các số như **10, 67, 48** khi xuất hiện có liên quan đến doanh thu trung bình ngày hôm sau cao hơn đáng kể (lên tới ~1,290 tỷ VND).

Ứng dụng

Insight này có thể được sử dụng để tối ưu hóa chiến lược marketing và quản lý lượng vé bán ra sau khi có kết quả của một số con số nhất định.

Kiến tạo Dữ liệu cho Tương lai: Feature Engineering cho Mô hình Dự báo

Để mô hình học máy có thể 'hiểu' và dự đoán doanh thu, chúng ta cần tạo ra các đặc trưng (features) có ý nghĩa từ dữ liệu thời gian thô.



1. Temporal Features (Đặc trưng Thời gian)

- Mục đích:** Giúp mô hình nhận diện xu hướng dài hạn và tính mùa vụ.
- Ví dụ:** `month`, `year`, `quarter`. Đặc biệt, `sin_month` & `cos_month` được sử dụng để mã hóa tính chu kỳ của tháng, giúp mô hình hiểu rằng tháng 12 gần với tháng 1.

2. Lag Features (Đặc trưng Trễ)

- Mục đích:** Giúp mô hình 'nhìn' vào quá khứ để dự đoán tương lai. Doanh thu tháng này thường có liên quan mật thiết đến các tháng trước.
- Ví dụ:** `revenue_lag1` (doanh thu tháng trước), `revenue_lag3` (doanh thu 3 tháng trước).

3. Moving Averages (Trung bình Trượt)

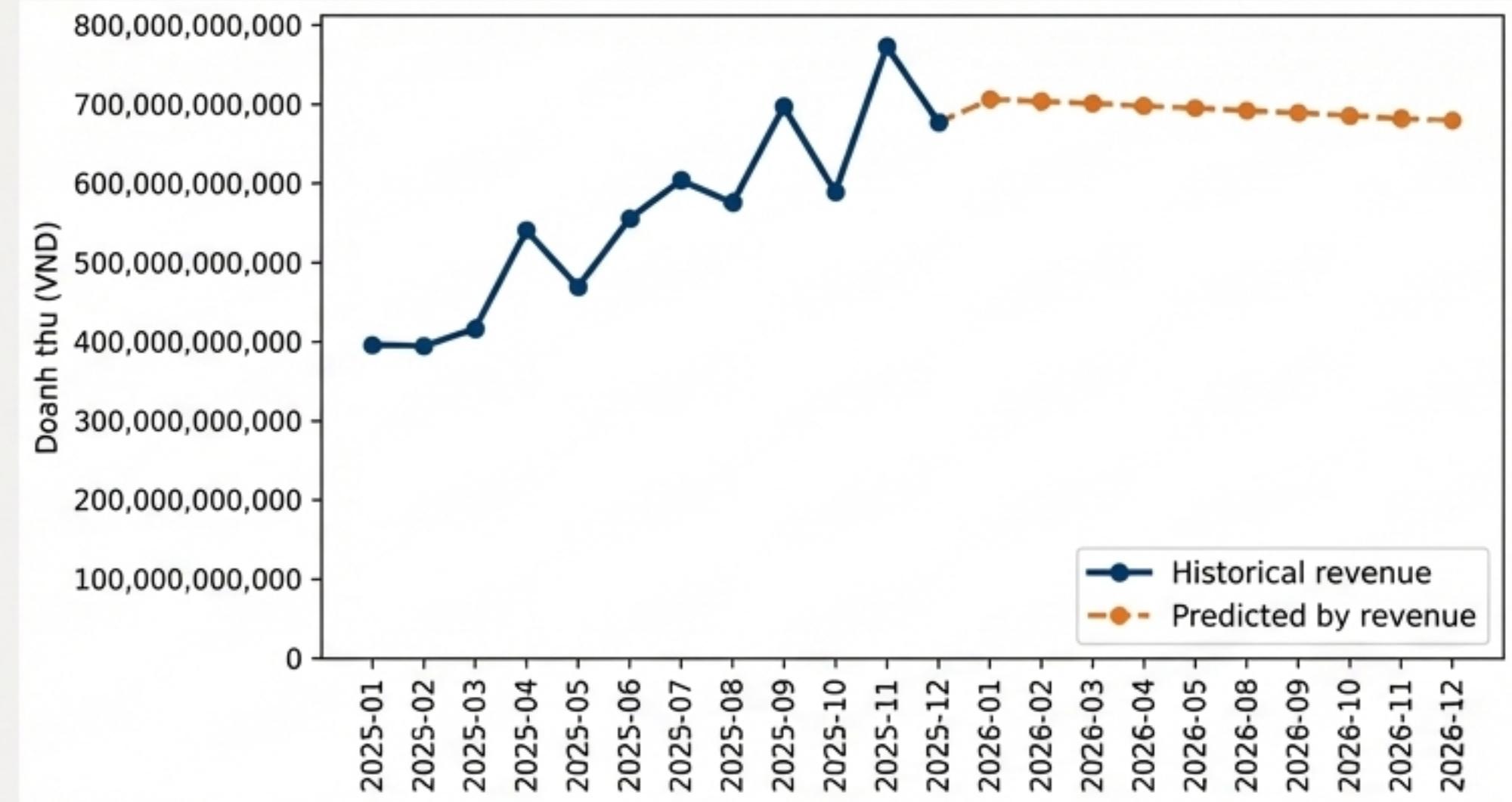
- Mục đích:** Làm mượt dữ liệu, loại bỏ nhiễu ngắn hạn và làm nổi bật xu hướng tổng thể.
- Ví dụ:** `revenue_ma3` (trung bình doanh thu 3 tháng gần nhất).

Cỗ máy Thời gian: Dự báo Doanh thu 12 tháng tới bằng Hồi quy Tuyến tính

Các chỉ số đánh giá độ chính xác của mô hình trên tập kiểm tra:

MAE (Mean Absolute Error):
18,276,597,996 VND

RMSE (Root Mean Square
Error): 29,080,200,633 VND



- Mô hình Hồi quy Tuyến tính đã nắm bắt thành công xu hướng tăng trưởng chung của doanh thu.
- Đường dự đoán (màu cam) cho thấy một xu hướng ổn định, cung cấp một cơ sở định lượng để lập kế hoạch kinh doanh và phân bổ nguồn lực cho năm tới.

Tổng kết Thành tựu Đạt được

Dự án đã xây dựng thành công một hệ thống kho dữ liệu toàn diện, biến dữ liệu xổ số từ dạng thô thành các insights phân tích và khả năng dự báo chiến lược.

Key Achievements

- ✓ Thiết kế & Triển khai Kho dữ liệu: Hoàn thành mô hình Lược đồ Sao với 4 bảng Dimension và 2 bảng Fact, tối ưu cho phân tích đa chiều.
- ✓ Xây dựng ETL Pipeline Tự động: Quy trình xử lý hiệu quả dữ liệu lịch sử 10 năm (~180,000 bản ghi kết quả và ~36,500 bản ghi doanh thu).
- ✓ Phát triển Năng lực Phân tích Sâu: Triển khai hơn 21 phương thức truy vấn, khám phá các xu hướng doanh thu, mẫu số và mối tương quan chéo.
- ✓ Tích hợp Thành công Machine Learning: Xây dựng mô hình dự báo doanh thu với Hồi quy Tuyến tính, cung cấp dự báo định lượng cho 12 tháng tới.
- ✓ Trực quan hóa Dữ liệu Hiệu quả: Xây dựng Dashboard tương tác bằng Streamlit, giúp người dùng dễ dàng tiếp cận và khai thác thông tin.

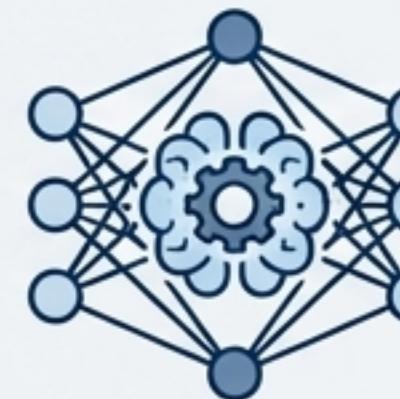
Chương tiếp theo: Hướng phát triển trong Tương lai

Nền tảng hiện tại là một khởi đầu vững chắc. Các cải tiến trong tương lai có thể nâng cao đáng kể năng lực của hệ thống.



1. Nâng cấp Hạ tầng (Infrastructure)

- Di chuyển từ SQLite sang hệ quản trị CSDL mạnh mẽ hơn như **PostgreSQL** hoặc **MySQL** để tăng khả năng mở rộng và chịu tải.
- Triển khai xử lý dữ liệu thời gian thực (real-time processing) thay vì xử lý theo lô (batch).



2. Tối ưu Mô hình Học máy (Advanced ML)

- Áp dụng các thuật toán chuỗi thời gian phức tạp hơn như **ARIMA**, **LSTM** để cải thiện độ chính xác dự báo.
- Phát triển các mô hình phân khúc khách hàng hoặc hệ thống gợi ý.



3. Mở rộng Giao diện và Tính năng (Enhanced UI/UX)

- Tích hợp các tính năng tương tác sâu hơn như **drill-down** và bộ lọc động trên dashboard.
- Xây dựng hệ thống cảnh báo tự động khi phát hiện các xu hướng bất thường.



4. Tích hợp Nguồn dữ liệu Mới (Data Enrichment)

- Kết hợp thêm dữ liệu từ các nguồn bên ngoài như mạng xã hội hoặc dữ liệu kinh tế vĩ mô để làm giàu thêm bối cảnh phân tích.