

# “Trí tuệ tập thể” là chìa khóa thiết kế mô hình học sâu hiện đại cho diễn giải dữ liệu chuỗi thời gian

Nguyễn Viết Tuấn Kiệt<sup>12</sup>  
kiet.nvt220032@sis.hust.edu.vn

Tăng Trần Mạnh Hưng<sup>12</sup>  
hung.ttm230037@sis.hust.edu.vn

Nguyễn Công Hùng<sup>13</sup>  
hung.nc224858@sis.hust.edu.vn

Mai Lê Phú Quang<sup>12</sup>  
quang.mlp230058@sis.hust.edu.vn

## Tóm tắt nội dung

Trong bối cảnh phân tích dữ liệu chuỗi thời gian, khi mà hàng loạt các mô hình học sâu với hiệu suất và khả năng trích xuất đặc trưng phức tạp đã được phát triển, việc sử dụng một mô hình đơn lẻ thường dẫn đến hiện tượng “bão hòa kiến thức” – nơi mô hình đã khai thác hết thông tin từ dữ liệu và không thể vượt qua giới hạn riêng của nó. Bài báo này đề xuất phương pháp học tập tập thể (ensemble learning) như một chìa khóa trong thiết kế các mô hình học sâu hiện đại, nhằm kết hợp ưu điểm của các mô hình khác nhau đồng thời khắc phục hạn chế riêng lẻ của từng mô hình. Phương pháp học tập tập thể cho phép khai thác thông tin đa chiều từ dữ liệu chuỗi thời gian, từ đó cải thiện đáng kể hiệu quả dự báo và diễn giải. Qua các thí nghiệm so sánh, kết quả cho thấy phương pháp đề xuất không chỉ tăng cường độ chính xác mà còn cải thiện độ tin cậy so với các phương pháp truyền thống, mở ra hướng đi mới cho việc ứng dụng trí tuệ tập thể trong lĩnh vực học sâu hiện đại.

**Mã nguồn:** [github.com/HaiAu2501/DataFlow-2025](https://github.com/HaiAu2501/DataFlow-2025)

## Mục lục

### 1 Giới thiệu

2

### 2 Quy trình phân tích

2

- 2.1 Kiểm chứng giả thuyết . . . . . 2  
2.2 Phát hiện tính chất . . . . . 2

### 3 Quy trình dự báo

3

- 3.1 Thông hiểu dữ liệu . . . . . 3

3.2 Tiền xử lý dữ liệu . . . . . 3

3.3 Thiết kế mô hình . . . . . 3

3.4 Huấn luyện mô hình . . . . . 4

### 4 Kết quả thực nghiệm

4.1 Dánh giá kết quả . . . . . 5

4.2 Thủ nghiệm so sánh . . . . . 5

4.3 Tổng kết . . . . . 5

### 5 Phụ lục

7

<sup>1</sup>Phòng thí nghiệm nghiên cứu Mô hình hóa, Mô phỏng và Tối ưu hóa; Trung tâm Nghiên cứu Quốc tế về Trí tuệ nhân tạo BKAI; SoICT

<sup>2</sup>Chương trình tài năng - Khoa học máy tính; Khoa Khoa học máy tính; Trường Công nghệ Thông tin và Truyền thông; Đại học Bách khoa Hà Nội

<sup>3</sup>Khoa học máy tính; Khoa Khoa học máy tính; Trường Công nghệ Thông tin và Truyền thông; Đại học Bách khoa Hà Nội

# 1 Giới thiệu

Trong thời đại công nghệ phát triển, dữ liệu chuỗi thời gian ngày càng quan trọng trong dự báo và ra quyết định [1]. Dù học sâu mang lại nhiều giải pháp, một mô hình đơn lẻ thường gặp hiện tượng *bão hòa kiến thức* (saturation) [2, 3, 4], hạn chế khả năng khai thác thông tin và khám phá tri thức mới.

Phương pháp *học tập tập thể* (ensemble learning) là hướng đi tiềm năng, kết hợp nhiều mô

hình để xử lý dữ liệu phức tạp, cải thiện độ chính xác và tăng cường ổn định [5, 6, 7]. Bài báo này khám phá hiệu quả của chiến lược này trong nâng cao hiệu suất hệ thống học sâu cho dữ liệu chuỗi thời gian.

Chúng tôi sử dụng tập dữ liệu từ cuộc thi **Data Flow** để thiết kế mô hình và đánh giá, chi tiết tại [Phụ lục](#).

## 2 Quy trình phân tích

### Kết quả phân tích dữ liệu lịch sử và khám phá các thông tin hữu ích

#### 2.1 Kiểm chứng giả thuyết

*Kiểm chứng giả thuyết* (hypothesis testing) là bước quan trọng trong diễn giải dữ liệu, giúp xác định thông tin hữu ích và mối quan hệ giữa các thuộc tính. Quá trình này đảm bảo hiểu biết có chọn lọc về dữ liệu và hỗ trợ trích chọn đặc trưng quan trọng.

Chúng tôi áp dụng các kỹ thuật kiểm chứng như ANOVA [8], T-Test [9], hệ số tương quan Pearson [10], xu hướng Mann-Kendall [11, 12], Kruskal-Wallis [13], Chi-square [14] để phân tích mối quan hệ giữa các thuộc tính. Kết quả được trình bày trong [Bảng 5](#).

#### 2.2 Phát hiện tính chất

*Phát hiện tính chất* bao gồm *trích xuất đặc trưng* (feature extraction), *lựa chọn đặc trưng* (feature selection) và *phát hiện mẫu* (pattern recognition) nhằm hiểu dữ liệu sâu hơn.

Chúng tôi trực quan hóa dữ liệu qua đồ thị, biểu đồ, bảng và áp dụng các phương pháp *khai phá dữ liệu* (data mining), *học máy* (machine learning) cùng các độ đo thống kê để khai thác thông tin.

a. *Tính chất tổng thể*: Phân tích dữ liệu chuỗi thời gian nhằm đánh giá tình hình kinh doanh, xem xét các đặc trưng như *tính dừng*, *tính mùa vụ*, *tính tự tương quan* (ACF & PACF), *tính khả phân* (PCA) [15], *tính nhân quả* (Granger causality) [16]. Các mô hình VAR [17], ARIMA [18], Prophet [19], XGBoost [20], BSTS/UCM [21] được áp dụng. Kết quả trong [Phụ lục 5.II.a](#).

b. *Tính chất nhóm*: Tổ chức dữ liệu thành nhóm dựa trên mức độ tương đồng để khám phá đặc tính ẩn. Sử dụng K-Means Clustering [22], *biên Pareto* (Pareto front), *độ đo Silhouettes* [23] để đánh giá. Chi tiết trong [Phụ lục 5.II.b](#).

c. *Phân tích theo khu vực*: Giúp doanh nghiệp hiểu nhu cầu và hành vi tiêu dùng theo từng địa điểm để điều chỉnh chiến lược giá, tiếp thị, phân phối sản phẩm. Xem chi tiết tại [Phụ lục 5.II.c](#).

d. *Phân tích theo ngành hàng*: Đánh giá doanh thu, chi phí, vòng đời sản phẩm để tối ưu ngân sách, phát hiện xu hướng tiêu dùng, cải thiện danh mục sản phẩm và chiến lược tiếp thị. Chi tiết tại [Phụ lục 5.II.d](#).

### 3 Quy trình dự báo

#### Chi tiết mô hình học sâu được sử dụng để dự báo

##### 3.1 Thông hiểu dữ liệu

Trong các thuộc tính của tập huấn luyện (**train**) và tập kiểm thử (**test**), chúng tôi xem xét các yếu tố có thể dự đoán bao gồm:

- (i) Tổng doanh thu (**Revenue**) và tổng sản phẩm bán ra (**Units**) theo từng ngày trên toàn quốc.
- (ii) Tổng doanh thu (**Revenue**) và tổng sản phẩm bán ra (**Units**) theo từng ngày và từng khu vực (**Region**), bao gồm: miền Tây, miền Trung, miền Đông.

Chúng tôi nhấn mạnh *các dự đoán (i) và (ii) mang ý nghĩa thống kê đối với doanh nghiệp* bởi các nguyên nhân sau:

Thứ nhất, các dự đoán (i) và (ii) giúp giảm thiểu tác động của biến động địa phương và tăng tính ổn định dữ liệu. Dự đoán ở cấp độ tổng thể giúp trung hòa các biến động bất thường tại từng cửa hàng riêng lẻ, như sự kiện địa phương, thời tiết, hay các yếu tố ngoại nhiên khác (điều mà không thể kiểm soát hay ghi chép chính xác được). Điều này tạo ra bức tranh tổng quan ổn định hơn về hiệu suất kinh doanh.

Thứ hai, bằng việc quan sát xu hướng chung của thị trường trên quy mô lớn, doanh nghiệp có cơ sở để xây dựng chiến lược cấp độ tổng thể chính xác hơn. Việc quan sát hàng loạt biến động đơn lẻ ở cấp độ cửa hàng không giúp doanh nghiệp điều chỉnh chiến lược tiếp thị, cơ cấu ngành hàng vì biến đổi đó là không lường trước được, không đủ dữ liệu để phân tích.

Cuối cùng, phân tích này tiết kiệm thời gian và nguồn lực. Thay vì thu thập và phân tích dữ liệu từ từng cửa hàng, dự báo tổng thể giảm bớt khối lượng công việc, tập trung vào

các chỉ số quan trọng ở cấp độ cao hơn.

##### 3.2 Tiền xử lý dữ liệu

Về đặc điểm của dữ liệu, chúng tôi nhận thấy, có 2 vấn đề lớn: (1) dữ liệu không bao gồm tất cả các ngày và; (2) dữ liệu có nhiều giá trị ngoại lai.

Để xử lý (1), chúng tôi điền dữ liệu cho các ngày thiếu thông tin thống kê trong tập huấn luyện bằng phương pháp nội suy. Với (2), chúng tôi không lọc bỏ dữ liệu như vậy mà coi đó như một tiêu chí đánh giá mô hình, mô hình buộc phải học cả những ngoại lai và biến động bất thường. Nhận thấy các mô hình trong Phụ lục 5.II.a tỏ ra yếu kém nếu chỉ dựa vào biến thời gian, kết hợp cùng kết quả của các kiểm chứng trong Mục 2.1 (Bảng 5), chúng tôi tin rằng, doanh thu bị ảnh hưởng bởi những yếu tố sau:

- Thông tin lịch sử: doanh thu trong chuỗi lịch sử, bởi nó phản ánh chính sách chung của doanh nghiệp.
- Thông tin ngoại lai, tri thức về lịch (static/calendar features): thứ trong tuần, số thứ tự của tháng trong năm, ngày cuối tuần/ngày thường, ngày lễ,... đều suy ra từ ngày.

Công thức xem tại Phụ lục 5.III.Đọc thêm.

##### 3.3 Thiết kế mô hình

Chúng tôi thiết kế hệ thống các mô hình đơn lẻ, mỗi mô hình giữa trên *kiến trúc hỗn hợp* (hybrid model) bằng việc tổ hợp các nhánh (branch) mạng với kiến trúc khác nhau. Vì vậy, mỗi nhánh trích xuất được một đặc trưng thông tin và mang thế mạnh riêng biệt. Mỗi mô hình sau đó được huấn luyện đến mức bao

hỏa, tức không thể cải thiện thêm nữa dù có tăng tham số, tăng số vòng lặp tối ưu. Điều này làm cho mỗi mô hình đều thực sự mạnh mẽ, để tham gia vào quá trình học tập tập thể như đã đề cập.

Chúng tôi lấy tên của mô hình xương sống để định danh mô hình, nhưng về mặt thiết kế, mô hình của chúng tôi đã cải tiến nhiều so với bản gốc. Các mô hình của chúng tôi bao gồm:

*a. Hybird Temporal Fusion Transformer:* Tận dụng sự độc đáo của cơ chế Positional Encoding [24] trong kiến trúc tân tiến Temporal Transformer [25], đồng thời bổ sung lớp Multi-scale Feature Extractor [26] nhằm trích xuất thông tin chuỗi thời gian ở quy mô khác nhau. Chi tiết về kiến trúc xem tại [Phụ lục 5.III.a](#).

*b. Hybrid Temporal Convolutional Network:* Kết hợp nhánh dựa trên toán tử tích chập nhằm trích xuất thông tin cục bộ của Temporal Convolutional Network [27], cùng các nhánh dựa trên Neural Ordinary Differential Equations [28] để mô hình hóa thông tin liên tục. Chi tiết xem tại [Phụ lục 5.III.b](#).

*c. Hybrid Forecasting Model:* Tổ hợp thông tin độc đáo từ các nhánh theo kiến trúc N-HiTS [29], iTransformer (Inverted Transformer) [30] và nhánh mang thông tin tự hồi quy dựa trên Long-Short Term Memory (LSTM) [31]. Đọc thêm tại [Phụ lục 5.III.c](#).

*d. Conditional Variational AutoEncoder:* Tự cải tiến từ mô hình VAE [32], kết hợp đặc trưng động và tĩnh rồi ánh xạ vào không gian ẩn (latent space) qua bộ mã hóa. Tái tạo đầu ra bằng phương pháp lấy mẫu khả vi (reparameterization) và bộ giải mã. Đọc thêm tại [Phụ lục 5.III.d](#).

*e. Distributional Conditional Forecast:* Tự cải tiến tốt nhất từ mô hình VAE. Khác với (d), mô hình trả về các tham số của một phân phối

chuẩn thay vì một giá trị cố định. Đọc thêm tại [Phụ lục 5.III.e](#).

## 3.4 Huấn luyện mô hình

### a. Hàm mất mát

Các mô hình (a), (b), (c) sử dụng *Huber Loss* [33] nhằm giảm ảnh hưởng của ngoại lai và ổn định quá trình học.

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^n \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2, & |y_i - \hat{y}_i| < \delta \\ \delta \cdot (|y_i - \hat{y}_i| - \frac{1}{2}\delta), & |y_i - \hat{y}_i| \geq \delta \end{cases}$$

trong đó  $\mathbf{y}, \hat{\mathbf{y}} \in \mathbb{R}^n$  là giá trị thực và giá trị dự báo.

Mô hình (d) sử dụng tổng trọng số của mất mát tái tạo (reconstruction loss, hay  $\mathcal{L}_{\text{recon}}$ ) dựa trên *Huber Loss* và mất mát giữa phân phối của không gian ẩn  $q(z|x) \sim \mathcal{N}(\mu, \sigma^2)$  với phân phối chuẩn  $p(z) \sim \mathcal{N}(0, I)$  dựa trên *Kullback-Leibler Divergence* [34].

$$\begin{aligned} \mathcal{L}_{\text{VAE}}(\mathbf{y}, \hat{\mathbf{y}}) &= \mathcal{L}_{\text{recon}}(\mathbf{y}, \hat{\mathbf{y}}) + \lambda \cdot \mathcal{L}_{\text{KL}} \\ \mathcal{L}_{\text{KL}} &= -\frac{1}{2} \mathbb{E} [1 + \log \sigma^2 - \mu^2 - \sigma^2] \end{aligned}$$

Mô hình (e) sử dụng tổng trọng số của *Negative Log-Likelihood Loss* (NLL) [35, 36, 37] và *Kullback-Leibler Divergence Loss*.

$$\begin{aligned} \mathcal{L}_{\text{DCF}} &= \mathcal{L}_{\text{NLL}} + \lambda \cdot \mathcal{L}_{\text{KL}} \\ \mathcal{L}_{\text{NLL}} &= \frac{1}{2} \mathbb{E} \left[ \log(2\pi) + \log \sigma^2 + \frac{(\mathbf{y} - \mu)^2}{\sigma^2} \right] \end{aligned}$$

### b. Chiến lược huấn luyện

Tất cả các mô hình đều sử dụng thuật toán tối ưu Adam [38], kết hợp chiến lược giảm tốc độ học (learning rate) theo cấp số nhân nếu mắc kẹt. Riêng mô hình (d) và (e) sử dụng thêm chiến lược *luyễn kim tuyến tính* (linear annealing) [39] để giải quyết vấn đề *posterior collapse*, giúp mô hình VAE học tốt hơn không gian ẩn.

## 4 Kết quả thực nghiệm

## 4.1 Dánh giá kết quả

Sau khi huấn luyện các mô hình (a), (b), (c), (d), (e) đến mức bão hòa, thu được kết quả như [Bảng 1](#) đối với dự đoán (i). Chúng tôi thiết kế *siêu mô hình* (meta-model) bằng cách tổ hợp tuyến tính các mô hình trên, với kết quả vượt trội hơn tất cả mô hình trước đó.

**Meta (1)** là tổ hợp tuyến tính của các mô hình đề xuất, được huấn luyện lại trên tập xác thực (validation, 20% của tập huấn luyện) để giảm sai khác bình phương (OLS).

**Meta (2)** là tổ hợp tuyến tính, có kẽ thêm dữ liệu đặc trưng, thay vì chỉ tham khảo dự đoán của các mô hình đề xuất. Mô hình này cũng tận dụng cận trên và cận dưới trong khoảng tin cậy 95% của phân phối do (e) trả về.

Sau đó, chúng tôi đánh giá mô hình dự đoán cho riêng kết quả **Units** và **Revenue**, hiệu quả mô hình được trình bày trong [Bảng 3](#). Mô hình **Meta (2)** do chúng tôi đề xuất đứng đầu và cho kết quả vượt trội đáng kể so với các mô hình khác.

Bảng 1: Kết quả của các mô hình đề xuất tính trên 2 kết quả trả về: **Units** và **Revenue**

Mô hình	R-squared	MAPE	RMSE
(a)	0.9268	0.5517	196779
(b)	0.9507	0.4439	168876
(c)	0.9387	0.5783	171857
(d)	0.9557	0.3846	150960
(e)	0.9650	0.2691	127062
Meta (1)	0.9708	0.2563	119236
<b>Meta (2)</b>	<b>0.9740</b>	<b>0.2433</b>	<b>118243</b>
(f)	0.9682	0.2565	110593
Meta (3)	0.9745	0.2415	112202
<b>Meta(4)</b>	<b>0.9786</b>	<b>0.2283</b>	<b>104450</b>

Ngoài ra, chúng tôi sử dụng mô hình (e) (hay DCF) để dự đoán tổng sản phẩm bán ra và tổng doanh thu cho từng vùng như đã đề cập

ở (ii), kết quả như trong [Bảng 2](#). Chi tiết hơn có thể xem tại [Bảng 4](#).

Bảng 2: Kết quả dự đoán theo vùng

Vùng	R-squared	MAPE	RMSE
Central	0.9784	0.6205	56552
East	0.9762	0.6781	38682
West	0.9814	0.1886	47860

## 4.2 Thủ nghiệm so sánh

Vấn đề mà chúng tôi, cũng như các bạn hẳn sẽ thắc mắc: *liệu việc tiếp tục bổ sung mô hình vào hệ thống học tập tập thể có tiếp tục mang đến cải thiện nào không?* Để giải quyết câu hỏi này, chúng tôi bổ sung mô hình (f) vào hệ thống học tập.

*f. Probabilistic Forecast Transformer:* Nhận thấy sự ưu việt của mô hình trả về phân phối xác suất lấy cảm hứng từ VAE (như (e)), chúng tôi tiếp tục cải tiến kiến trúc của (e) bằng việc bổ sung cơ chế Positional Encoding. Hầm mắt mát và chiến lược huấn luyện giống như (e). Xem tại [Phụ lục 5.III.f](#).

**Meta (3)** là tổ hợp tuyến tính từ kết quả trả về của tất cả 6 mô hình, vượt trội hơn hẳn **Meta (2)**. Cuối cùng, **Meta (4)** cân nhắc thêm cận trên và cận dưới trong khoảng tin cậy 95% từ kết quả trả về của (f), là mô hình tốt nhất trong tất cả tiêu chí.

## 4.3 Tổng kết

Kết quả thực nghiệm và so sánh đã cho thấy, phương pháp học tập tập thể của chúng tôi tỏ ra hữu hiệu trong ngữ cảnh của bài toán. Nghiên cứu cũng đóng góp một hướng đi mới cho phân tích và xử lý dữ liệu chuỗi thời gian, đồng thời tận dụng hiệu quả hệ thống mô hình tiên tiến trong lĩnh vực này.

Bảng 3: Kết quả của các mô hình đề xuất  
trên mỗi kết quả trả về: Units và Revenue

Mô hình	Units			Revenue		
	R-squared	MAPE	RMSE	R-squared	MAPE	RMSE
(a)	0.9283	0.6857	34	0.9254	0.4178	278288
(b)	0.9564	0.5260	27	0.9451	0.3617	238827
(c)	0.9343	0.6141	33	0.9431	0.4225	243043
(d)	0.9553	0.4436	27	0.9561	0.3256	213490
(e)	0.9610	0.3068	25	0.9689	<b>0.2313</b>	179693
Meta (1)	0.9691	0.2714	23	0.9726	0.2412	168625
<b>Meta (2)</b>	<b>0.9750</b>	<b>0.2430</b>	<b>20</b>	<b>0.9731</b>	0.2436	<b>167221</b>
(f)	0.9599	0.3017	26	0.9764	0.2112	156402
Meta (3)	0.9733	0.2472	21	0.9757	0.2358	158678
<b>Meta (4)</b>	<b>0.9782</b>	<b>0.2337</b>	<b>19</b>	<b>0.9790</b>	<b>0.2228</b>	<b>147715</b>

Bảng 4: Kết quả của các mô hình (e) đối với từng vùng

Vùng	Units			Revenue		
	R-squared	MAPE	RMSE	R-squared	MAPE	RMSE
Central	0.9738	0.6291	12	0.9831	0.6120	79977
East	0.9836	0.2140	8	0.9792	0.1631	67685
West	0.9690	0.3880	9	0.9834	0.9682	54704

## 5 Phụ lục

### Mô tả dữ liệu

Dữ liệu được cung cấp bao gồm 4 bảng: **train** (tập huấn luyện), **test** (tập kiểm tra), **product** (thông tin sản phẩm), **geography** (thông tin địa lý). Chi tiết các cột trong từng bảng như sau:

Bảng **train** và **test**:

- **ProductID** (khóa ngoại): Mã sản phẩm duy nhất, để xác định sản phẩm từ bảng **product**.
- **Date**: Ngày giao dịch.
- **Zip** (khóa ngoại): Mã bưu điện duy nhất, để xác định vị trí cửa hàng từ bảng **geography**.
- **Units**: Số lượng bán trong giao dịch.
- **Revenue**: Tổng doanh thu trong giao dịch.
- **COGS** (Cost of Goods Sold): Chi phí vốn.

Bảng **product**:

- **ProductID** (khóa chính).
- **Product**: Tên sản phẩm.
- **Category**: Danh mục sản phẩm, thuộc (Urban, Rural, Youth, Mix).
- **Segment**: Phân khúc sản phẩm.

Bảng **geography**:

- **Zip** (khóa chính).
- **District**: Quận, trong địa chỉ cửa hàng.
- **City**: Thành phố.
- **State**: Bang.
- **Region**: Vùng/Miền, trong 3 miền (East, Central, West) của Mỹ.

Tập **train** chứa dữ liệu các giao dịch từ ngày 04/07/2010 đến ngày 31/12/2020, tập **test** bắt đầu từ 01/01/2021 đến 31/12/2021.

## I. Kiểm chứng giả thuyết

Bảng 5: Kiểm định giả thuyết thống kê với ngưỡng ý nghĩa  $\alpha = 0.05$   
và kết quả: thừa nhận (✓) hoặc bác bỏ (✗)

Giả thuyết thống kê	Thuật toán	Kết quả
Doanh thu trung bình có khác biệt giữa các khu vực (Region)?	ANOVA	✓
Số lượng sản phẩm bán ra có khác nhau giữa các danh mục (Category)?	ANOVA	✓
Doanh thu trung bình trước và sau năm 2016 có khác biệt đáng kể?	T-Test	✓
Doanh thu và số lượng bán có tương quan dương?	Pearson	✓
COGS có khác nhau giữa các phân khúc sản phẩm?	ANOVA	✓
Lợi nhuận trung bình có khác biệt giữa các khu vực?	ANOVA	✓
Doanh thu trung bình có xu hướng tăng qua các năm không?	Mann-Kendall	✗
Giá bán trung bình có khác biệt giữa các danh mục sản phẩm?	ANOVA	✓
Doanh thu vào cuối tuần và ngày thường có khác biệt không?	T-Test	✓
Tỷ suất lợi nhuận gộp có khác nhau giữa các khu vực?	ANOVA	✓
Tỷ suất lợi nhuận trên chi phí có khác nhau giữa các danh mục sản phẩm?	Kruskall-Wallis	✓
Các sản phẩm có đóng góp khác nhau vào tổng số lượng bán?	Chi-square	✓
Biến động doanh thu giữa các tuần có ảnh hưởng đến lợi nhuận?	Pearson	✓
Doanh thu có bị ảnh hưởng bởi mùa không?	ANOVA	✓
Tỷ suất lợi nhuận theo năm có thay đổi không?	Chi-square	✓
Hệ số biến động doanh thu có khác nhau giữa các khu vực?	Kruskall-Wallis	✓
Doanh thu vào các ngày nghỉ lễ có cao hơn ngày thường không?	T-Test	✓

## II. Phát hiện tính chất

### a. Tính chất tổng thể

#### 1. Trực quan hóa

- Trực quan hóa doanh thu (Revenue), doanh số (Units) và giá vốn (COGS) theo tuần, tháng, quý năm: [Hình 1](#).
- Top 10 & Bottom 10 theo sản phẩm, bang, thành phố dựa trên một hoặc cặp tiêu chí trong {doanh thu, doanh số, giá vốn} theo phương pháp rank sum: [Hình 2](#), [Hình 3](#), [Hình 4](#).
- Thứ hạng phân khúc, danh mục và vùng: [Hình 5](#), [Hình 6](#), [Hình 7](#).
- Sự thay đổi thứ hạng của doanh thu các tháng, quý theo năm: [Hình 8](#).

### Phân chia giai đoạn

Từ đây, ta có thể quan sát rõ ràng 3 giai đoạn trong lịch sử kinh doanh, gồm:

#### (i) Giai đoạn 2010 - 2016: Tăng trưởng mạnh

- Doanh thu, doanh số và giá vốn đều tăng, đạt đỉnh vào khoảng 2015 - 2016.
- Có xu hướng tăng trưởng theo mùa, phản ánh chiến lược bán hàng hiệu quả.
- Chi phí giá vốn tăng nhưng không ảnh hưởng tiêu cực đến lợi nhuận.

### (ii) Giai đoạn 2017 - 2019: Suy giảm đáng kể

- Doanh thu và doanh số giảm mạnh, kéo theo sự sụt giảm trong hiệu suất kinh doanh.
- Giá vốn cũng giảm, nhưng không đủ để bù đắp sự suy giảm doanh số.
- Có thể do cạnh tranh gia tăng, chiến lược bán hàng kém hiệu quả hoặc thay đổi trong hành vi mua sắm.

### (iii) Giai đoạn 2020 - nay: Dấu hiệu phục hồi nhẹ

- Doanh thu, giá vốn có dấu hiệu phục hồi, nhưng chưa đạt mức tăng trưởng trước đây.
- Doanh số vẫn thấp nhưng có xu hướng nhích lên từ năm 2020.
- Có thể do chiến lược mới bắt đầu có hiệu quả hoặc thị trường dần phục hồi.

## 2. Kiểm chứng tính chất chuỗi thời gian

Bảng 6: Kết quả kiểm định ADF cho chuỗi gốc và chuỗi sai phân bậc 1

Biến số	Tần suất	Chuỗi gốc		Sai phân bậc 1	
		ADF Statistic	p-value	ADF Statistic	p-value
Revenue	Tuần	-5.2091	<b>0.0000</b>	-6.3732	0.0000
	Tháng	-0.6036	0.8702	-4.1177	<b>0.0009</b>
	Quý	-0.9633	0.7664	-2.9259	<b>0.0424</b>
	Năm	-1.0192	0.7462	-3.9495	<b>0.0017</b>
Units	Tuần	-3.4712	<b>0.0088</b>	-5.8792	0.0000
	Tháng	-0.4834	0.8952	-3.5885	<b>0.0060</b>
	Quý	-0.6592	0.8570	-3.1441	<b>0.0235</b>
	Năm	-1.2567	0.6488	-4.9255	<b>0.0000</b>
COGS	Tuần	-5.5841	<b>0.0000</b>	-6.3883	0.0000
	Tháng	-0.6909	0.8491	-4.0735	<b>0.0011</b>
	Quý	-1.0329	0.7411	-2.9227	<b>0.0428</b>
	Năm	-1.1051	0.7130	-3.8786	<b>0.0022</b>

- Chi tiết trực quan hóa xem tại [Hình 9](#).
- Chi tiết Sesonal Decomposition theo chu kỳ 3, 6, 12 tháng: [Hình 10](#).

---

### Phân rã dữ liệu

---

Các kết luận rút ra từ hình ảnh trực quan bao gồm:

#### (i) Phân tích Doanh thu (Revenue)

*Xu hướng (Trend):*

- Doanh thu tăng mạnh từ 2010 - 2016, sau đó giảm nhanh từ 2017 - 2019.

- Xu hướng giảm rõ rệt hơn trong chu kỳ 12 tháng, phản ánh sự suy thoái kéo dài.
- Giai đoạn 2020 có dấu hiệu phục hồi nhẹ nhưng chưa rõ ràng.

*Thành phần thời vụ (Seasonal):*

- Ở chu kỳ 3 tháng, có sự dao động cao, cho thấy ảnh hưởng của các chương trình khuyến mãi ngắn hạn.
- Ở chu kỳ 6 tháng, mô hình thời vụ lặp lại đều đặn hơn.
- Ở chu kỳ 12 tháng, tác động của các mùa cao điểm và kỳ nghỉ lễ được thể hiện rõ.

*Phần dư (Residual):*

- Trước 2016, phần dư có biên độ dao động lớn, phản ánh sự tăng trưởng nhanh.
- Sau 2017, phần dư giảm dần, cho thấy sự ổn định hơn nhưng kém tích cực.
- Giai đoạn sau 2020, phần dư có xu hướng giảm nhẹ nhưng chưa thể kết luận sự phục hồi mạnh mẽ.

## (ii) Phân tích Doanh số (Units)

*Xu hướng (Trend):*

- Doanh số tăng dần từ 2010 - 2016, sau đó giảm mạnh từ 2017.
- Ở chu kỳ 12 tháng, có sự sụt giảm rõ ràng hơn từ năm 2017.
- Giai đoạn 2020 có dấu hiệu đi ngang, nhưng chưa có sự phục hồi mạnh.

*Thành phần thời vụ (Seasonal):*

- Ở chu kỳ 3 tháng, dao động ngắn hạn cho thấy các đợt giảm giá tác động đáng kể.
- Ở chu kỳ 6 tháng, xu hướng thời vụ có sự lặp lại tương đối ổn định.
- Ở chu kỳ 12 tháng, mùa cao điểm và các dịp lễ hội có ảnh hưởng rõ rệt.

*Phần dư (Residual):*

- Trước 2016, phần dư có độ biến động cao, cho thấy hoạt động kinh doanh sôi động.
- Sau 2017, phần dư giảm mạnh, thể hiện sự chững lại của thị trường.
- Giai đoạn sau 2020, phần dư có xu hướng dao động nhẹ, phản ánh sự điều chỉnh trong kinh doanh.

## (iii) Phân tích Giá vốn (COGS)

*Xu hướng (Trend):*

- Giá vốn tăng cùng doanh thu giai đoạn 2010-2016, nhưng giảm sau 2017.
- Chu kỳ 12 tháng cho thấy sự suy giảm mạnh trong chi phí từ 2017.
- Giai đoạn 2020 có dấu hiệu phục hồi nhẹ, nhưng vẫn thấp hơn so với giai đoạn đỉnh cao.

*Thành phần thời vụ (Seasonal):*

- Chu kỳ 3 tháng cho thấy các dao động mạnh trong từng quý.
- Chu kỳ 6 tháng phản ánh sự thay đổi rõ ràng giữa các giai đoạn bán hàng.
- Chu kỳ 12 tháng thể hiện sự ảnh hưởng của các mùa mua sắm lớn đến chi phí.

*Phần dư (Residual):*

- Trước 2016, phần dư có mức biến động cao, phản ánh thị trường sôi động.
- Sau 2017, phần dư giảm mạnh, cho thấy sự ổn định nhưng với quy mô nhỏ hơn.
- Giai đoạn sau 2020, phần dư có xu hướng ổn định trở lại nhưng chưa có tín hiệu tăng trưởng rõ rệt.

### Tự tương quan & Tự tương quan riêng phần

Tiếp theo, đối với kết quả của biểu đồ đã trực quan tại [Hình 11](#) thể hiện:

- **Autocorrelation Function (ACF):** Đo lường mức độ tương quan của một chuỗi thời gian với chính nó ở các độ trễ khác nhau. Phân tích này cho thấy mức độ phụ thuộc của một giá trị vào các giá trị trước đó.
- **Partial Autocorrelation Function (PACF):** Đo lường mức độ tương quan riêng phần giữa một điểm trong chuỗi và độ trễ của nó, sau khi loại bỏ ảnh hưởng của các độ trễ trung gian. PACF giúp xác định số lượng độ trễ cần sử dụng ( $p$ ) trong mô hình ARIMA.

Từ quan sát có thể rút ra các phỏng đoán:

- Revenue, Units, và COGS đều có tính tự tương quan mạnh, đặc biệt ở các độ trễ ngắn (1-2) và cao nhất với độ trễ 12.
- PACF cho thấy AR(1) hoặc AR(2) có thể là lựa chọn tốt để mô hình hóa dữ liệu.
- Dữ liệu có thể phù hợp với mô hình ARIMA( $p, 1, q$ ) với  $p = 1$  hoặc 2, do dữ liệu đã có tính dừng sau sai phân bậc 1.

### Trung bình & Độ lệch chuẩn trong phân phối dữ liệu

Quan sát [Hình 12](#) đề cập 2 đại lượng thống kê cơ bản nhất, bao gồm:

- **Trung bình (Mean):** Nếu mean tăng thì dữ liệu có xu hướng tăng trưởng. Nếu mean giảm hoặc dao động nhiều thì dữ liệu không ổn định, có thể bị ảnh hưởng bởi thị trường hoặc chiến lược kinh doanh.
- **Độ lệch chuẩn (Standard Deviation):** Std thấp thì chuỗi ổn định, ít biến động. Std cao nghĩa là chuỗi dao động lớn, có nhiều rủi ro hoặc cơ hội bất thường.

Một số tính chất có thể quan sát được, chẳng hạn:

- **Trung bình (Mean)**

*Revenue (Doanh thu):*

- Doanh thu tăng ổn định theo thời gian ở tất cả các mức (tháng, quý, năm).
- Xu hướng tăng rõ ràng nhất ở dữ liệu theo năm, cho thấy sự mở rộng hoặc tăng trưởng chung của doanh nghiệp.
- Tuy nhiên, có các giai đoạn sụt giảm theo chu kỳ, có thể liên quan đến yếu tố mùa vụ hoặc suy thoái kinh tế.

*Units (Doanh số):*

- Biến động nhiều hơn so với doanh thu.
- Có sự ổn định tương đối từ năm 2014 - 2018 nhưng giảm nhẹ sau đó.
- Biến động mạnh ở dữ liệu theo tháng và quý, có thể do thay đổi chiến lược bán hàng hoặc sự thay đổi trong nhu cầu thị trường.

*COGS (Giá vốn):*

- Tăng trưởng đều theo thời gian, đặc biệt là xu hướng mạnh ở dữ liệu theo năm.
- Có sự gia tăng song song với doanh thu, phản ánh chi phí tăng theo quy mô bán hàng.

#### • Độ lệch chuẩn (Std)

*Revenue (Doanh thu):*

- Std tăng theo thời gian, đặc biệt từ 2015 trở đi, cho thấy biến động doanh thu ngày càng lớn.
- Có những giai đoạn dao động mạnh (2017 - 2020), có thể do thị trường hoặc các chính sách giá cả thay đổi.

*Units (Doanh số):*

- Std có mức độ biến động khá cao theo tháng và quý, phản ánh nhu cầu khách hàng không ổn định.
- Std theo năm cũng có xu hướng tăng nhẹ, cho thấy rủi ro trong việc dự báo doanh số.

*COGS (Giá vốn):*

- Std tăng dần từ 2015, nghĩa là chi phí trở nên biến động hơn, có thể do biến động giá nguyên vật liệu hoặc thay đổi chiến lược cung ứng.
- Tăng mạnh nhất vào 2019-2020, cho thấy sự bất ổn cao trong giá vốn.

### Trung bình trượt & Độ lệch chuẩn trượt

Trực quan tại [Hình 13](#), với xu hướng tổng thể tương tự.

- **Trung bình trượt (Moving Average - MA):** Trung bình trượt là một phương pháp làm mịn dữ liệu chuỗi thời gian bằng cách tính giá trị trung bình của một số điểm dữ liệu gần nhất trong một cửa sổ thời gian nhất định. Nếu trung bình trượt tăng, có thể cho thấy xu hướng tăng trong doanh thu/doanh số. Ngược lại, nếu giảm, có thể phản ánh xu

hướng suy giảm. Ngoài ra, trung bình trượt giúp loại bỏ biến động ngẫu nhiên trong dữ liệu, giúp dễ dàng quan sát xu hướng tổng thể.

- **Độ lệch chuẩn trượt (Rolling Standard Deviation):** Độ lệch chuẩn trượt là một chỉ số đo lường sự biến động của dữ liệu trong một khoảng thời gian di động. Sự gia tăng độ lệch chuẩn trượt có thể báo hiệu thời kỳ bất ổn trong doanh thu hoặc giá vốn. Nếu độ lệch chuẩn trượt đột ngột tăng, có thể có một sự kiện quan trọng (ví dụ: thay đổi chiến lược kinh doanh, tác động của yếu tố mùa vụ, cú sốc kinh tế).

### Độ nhọn & Độ lệch trong phân phối dữ liệu

Kết hợp với phân tích được đề cập ở [Hình 14](#), tập trung vào 2 đại lượng thống kê:

- **Kurtosis (Độ nhọn):** Đo lường mức độ tập trung của dữ liệu xung quanh giá trị trung bình.
  - Kurtosis cao ( $>3$ ): Dữ liệu có nhiều ngoại lệ, phân phối có đuôi dài.
  - Kurtosis thấp ( $<3$ ): Phân phối dữ liệu ít tập trung vào trung tâm, phân tán đều hơn.
- **Skewness (Độ lệch):** Đo lường mức độ bất đối称 của phân phối dữ liệu.
  - Skewness  $> 0$ : Phân phối lệch phải (có nhiều giá trị lớn).
  - Skewness  $< 0$ : Phân phối lệch trái (có nhiều giá trị nhỏ).

Chúng tôi rút ra một số đặc trưng sau:

- **Kurtosis (Độ nhọn)**

*Revenue (Doanh thu):*

- Kurtosis dao động mạnh theo thời gian, với các đỉnh cao vào các giai đoạn 2011-2012, 2016-2018 và 2020.
- Sau 2018, độ nhọn có xu hướng giảm dần, cho thấy dữ liệu ít bị ảnh hưởng bởi các ngoại lệ hơn.
- Các giai đoạn có kurtosis cao có thể phản ánh sự biến động lớn trong doanh thu, liên quan đến các sự kiện kinh tế hoặc chiến lược kinh doanh.

*Units (Doanh số):*

- Có mức độ kurtosis cao hơn so với Revenue và COGS, đặc biệt trong giai đoạn 2010-2015.
- Xu hướng giảm sau 2016, nhưng vẫn có những đỉnh đáng kể vào năm 2019-2020.
- Điều này cho thấy doanh số có nhiều thời điểm đột biến hoặc sụt giảm mạnh, có thể do yếu tố mùa vụ hoặc biến động thị trường.

*COGS (Giá vốn):*

- Giá trị kurtosis cao vào các giai đoạn 2012, 2016 và 2019, phản ánh sự thay đổi lớn trong chi phí.

- Sau 2017, độ nhọn có xu hướng ổn định hơn, cho thấy dữ liệu ít bị ảnh hưởng bởi ngoại lệ hơn.
- Điều này có thể liên quan đến sự thay đổi trong chiến lược cung ứng hoặc kiểm soát chi phí tốt hơn.

### • Skewness (Độ lệch)

*Revenue (Doanh thu):*

- Skewness chủ yếu dương, cho thấy dữ liệu bị lệch phải, có nhiều giá trị cao bất thường.
- Các giai đoạn 2011-2012, 2016-2018 và 2020 có độ lệch cao, đồng bộ với những giai đoạn có kurtosis cao.
- Sau 2014, có xu hướng giảm nhẹ, phản ánh sự cân bằng dần trong phân phối doanh thu.

*Units (Doanh số):*

- Skewness dao động liên tục, với các đỉnh lớn trong các giai đoạn 2012, 2016 và 2019.
- Điều này cho thấy doanh số có nhiều khoảng thời gian tăng giảm mạnh, không ổn định.
- Sự bất đối xứng có thể gây ảnh hưởng đến các mô hình dự báo, cần xử lý dữ liệu phù hợp.

*COGS (Giá vốn):*

- Skewness dương trong hầu hết các giai đoạn, cho thấy chi phí có nhiều giá trị cao bất thường.
- Xu hướng tăng vào năm 2016-2018, cho thấy có thể có sự biến động lớn trong chi phí sản xuất hoặc nhập hàng.
- Sau 2019, độ lệch có xu hướng giảm nhẹ, phản ánh sự kiểm soát tốt hơn về giá vốn.

## Ma trận tương quan

Biểu đồ trực quan tại [Hình 15](#) thể hiện ma trận tương quan Pearson giữa Doanh thu (Revenue), Doanh số (Units) và Giá vốn (COGS) theo các mức tháng, quý, năm. Ý nghĩa của Pearson Correlation:

- Gần 1: Mối quan hệ tuyến tính mạnh theo chiều thuận (cùng tăng/cùng giảm).
- Gần -1: Mối quan hệ tuyến tính mạnh theo chiều nghịch.
- Gần 0: Không có mối quan hệ tuyến tính rõ ràng.

Các nhận xét rút ra, bao gồm:

*Revenue và COGS:*

- Hệ số tương quan gần bằng 1 ở tất cả các mức (tháng, quý, năm), cho thấy chi phí giá vốn tăng cùng với doanh thu.

- Điều này phản ánh rằng phần lớn doanh thu bị chi phối bởi giá vốn hàng bán.
- Có thể cần kiểm soát giá vốn để tối ưu hóa lợi nhuận.

*Revenue và Units:*

- Tương quan cao (0.92 - 0.93), cho thấy doanh thu phụ thuộc nhiều vào doanh số.
- Ở mức quý và năm, tương quan tăng nhẹ, thể hiện sự ổn định hơn trong dài hạn.
- Tuy nhiên, không phải mọi thay đổi trong Units đều dẫn đến thay đổi tương đương trong Revenue, có thể do biến động giá.

*COGS và Units:*

- Tương quan thấp hơn một chút ( $\approx 0.92$ ), cho thấy giá vốn có thể bị ảnh hưởng bởi các yếu tố khác ngoài doanh số.
- Nếu chi phí đầu vào hoặc chiến lược giá thay đổi, mối quan hệ này có thể bị xáo trộn.
- Cần phân tích thêm về chiến lược định giá và kiểm soát chi phí để tối ưu lợi nhuận.

### Biểu đồ đáp ứng xung

Phân tích [Hình 16](#) sử dụng mô hình VAR để xem xét ảnh hưởng của các cú sốc (shocks) đến Revenue (Doanh thu) và Units (Doanh số) theo thời gian. Đây là phương pháp phổ biến trong kinh tế lượng để phân tích mối quan hệ động giữa các biến.

Giải thích các biểu đồ:

- **Phản ứng của Revenue**

*Cú sốc từ Revenue:*

- Doanh thu có phản ứng mạnh ngay lập tức khi có cú sốc và giảm dần về 0 sau khoảng 10 bước.
- Điều này cho thấy Revenue có tính tự tương quan cao, ảnh hưởng mạnh từ quá khứ.
- Xu hướng giảm nhanh sau một số bước cho thấy cú sốc không có tác động kéo dài.

*Cú sốc từ Units:*

- Cú sốc trong Units gây ra biến động lớn trong Revenue với cả tác động dương và âm.
- Điều này có thể phản ánh ảnh hưởng của doanh số đến doanh thu phụ thuộc vào yếu tố giá và chiến lược bán hàng.
- Sau khoảng 10 bước, tác động của cú sốc giảm dần về 0, cho thấy ảnh hưởng không kéo dài.

- **Phản ứng của Units**

*Cú sốc từ Revenue:*

- Một cú sốc Revenue có tác động nhỏ lên Units nhưng gây biến động mạnh trong những bước đầu tiên.

- Sau khoảng 10 bước, tác động này gần như biến mất, cho thấy Revenue không có ảnh hưởng dài hạn đến Units.
- Điều này có thể chỉ ra rằng doanh thu và doanh số có mối quan hệ ngắn hạn hơn là dài hạn.

*Cú sốc từ Units:*

- Units có phản ứng mạnh ngay lập tức khi có cú sốc nhưng giảm nhanh theo thời gian.
- Điều này cho thấy Units có tính tự tương quan cao, nhưng ảnh hưởng không kéo dài.
- Cú sốc ban đầu có tác động mạnh nhưng không kéo dài quá 10 bước, phản ánh sự ngắn hạn của biến động doanh số.

## Tính chất nhân quả

Bảng 7 thể hiện kiểm định nhân quả Granger giữa các biến Doanh thu (Revenue), Doanh số (Units), Vốn (COGS) và Lợi nhuận (Profit).

- Kiểm định Granger xác định liệu một biến có thể dự báo một biến khác hay không.
- Nếu p-value < 0.05, có thể kết luận rằng biến ở cột bên trái Granger gây ra biến ở cột bên phải.

Chúng tôi rút ra những kết luận tiêu biểu sau:

- **Mối quan hệ giữa Doanh thu và Doanh số**

*Doanh thu → Doanh số:*

- Với lags = 2, 3, 4, p-value < 0.05, cho thấy Doanh thu có thể dự báo Doanh số.
- Ảnh hưởng mạnh nhất ở lag = 4 (F-stat = 35.08, p-value = 0.0000).
- Điều này có nghĩa là doanh thu trong quá khứ có tác động đến doanh số hiện tại.

*Doanh số → Doanh thu:*

- Từ lag = 2 trở đi, p-value < 0.05, chứng tỏ Doanh số có thể dự báo Doanh thu.
- Ảnh hưởng mạnh nhất ở lag = 4 (F-stat = 26.41, p-value = 0.0000).
- Cho thấy mối quan hệ hai chiều giữa Doanh số và Doanh thu, có thể phản ánh chiến lược giá và khuyến mãi.

- **Mối quan hệ giữa Doanh thu và Vốn**

*Doanh thu → Vốn:*

- Tất cả p-value > 0.05, chứng tỏ không có mối quan hệ nhân quả rõ ràng.
- Điều này có nghĩa là sự thay đổi trong doanh thu không thể dùng để dự báo sự thay đổi trong vốn.

*Doanh số → Vốn:*

- Khi lag = 2, 3, 4, p-value < 0.05, cho thấy Doanh số có thể dự báo Vốn.
- Ảnh hưởng mạnh nhất ở lag = 4 (F-stat = 24.92, p-value = 0.0000).
- Điều này có thể chỉ ra rằng việc thay đổi số lượng hàng bán ảnh hưởng đến mức vốn cần sử dụng.

- **Mối quan hệ giữa Doanh thu, Doanh số và Lợi nhuận**

*Doanh thu → Lợi nhuận:*

- Tất cả p-value > 0.05, chứng tỏ không có mối quan hệ nhân quả giữa Doanh thu và Lợi nhuận.
- Điều này có thể xảy ra nếu chi phí và biên lợi nhuận thay đổi không đồng nhất với doanh thu.

*Doanh số → Lợi nhuận:*

- Khi lag = 2, 3, 4, p-value < 0.05, cho thấy Doanh số có thể dự báo Lợi nhuận.
- Ảnh hưởng mạnh nhất ở lag = 4 (F-stat = 26.11, p-value = 0.0000).
- Điều này chứng tỏ rằng sự thay đổi trong số lượng hàng bán có tác động mạnh đến lợi nhuận, có thể do chiến lược giá hoặc chi phí biến đổi.

### Thay đổi ngoại lai

Biểu đồ trên thể hiện số lượng outliers trong các biến Revenue (Doanh thu), Units (Doanh số), COGS (Giá vốn) và Profit (Lợi nhuận) theo tháng, quý, năm. Outliers (Điểm ngoại lai) là những giá trị nằm ngoài phạm vi bình thường của dữ liệu, có thể do:

- Sự kiện bất thường (khuyến mãi lớn, cú sốc kinh tế).
- Sai sót dữ liệu hoặc sự thay đổi mô hình kinh doanh.
- Biến động mạnh trong thị trường.

Hệ thống lại, chúng ta nhận thấy:

- **Tổng quan về số lượng outliers**

*Theo tháng:*

- Số lượng outliers có xu hướng giảm dần sau năm 2014, nhưng tăng mạnh trở lại vào 2019-2020.
- Biến Revenue và Profit có nhiều điểm ngoại lai hơn so với Units và COGS.

*Theo quý:*

- Số lượng outliers tương đối ổn định, nhưng có các giai đoạn tăng đột biến vào 2013-2015 và 2019.
- Điều này cho thấy có những biến động định kỳ trong dữ liệu, có thể do yếu tố mùa vụ hoặc chiến lược kinh doanh.

*Theo năm:*

- Số lượng outliers có xu hướng tăng dần theo thời gian, đặc biệt sau năm 2018.
- Điều này cho thấy sự biến động mạnh hơn trong các biến tài chính, có thể do môi trường kinh doanh thay đổi hoặc tác động từ thị trường.
- Profit có số lượng outliers cao nhất, cho thấy lợi nhuận không ổn định theo thời gian.

Bảng 7: Kết quả kiểm định nhân quả Granger

Mối quan hệ	Lags	F test		Chi-squared test	
		F-stat	p-value	Chi2	p-value
Doanh thu → Doanh số	1	0.8810	0.3498	0.9027	0.3421
	2	4.3562	<u>0.0149</u>	9.0784	<u>0.0107</u>
	3	17.9131	0.0000	56.9821	0.0000
	4	35.0827	0.0000	151.5075	0.0000
Doanh thu → Vốn	1	0.0630	0.8023	0.0645	0.7995
	2	0.7100	0.4937	1.4796	0.4772
	3	0.7495	0.5248	2.3842	0.4966
	4	1.4326	0.2278	6.1868	0.1856
Doanh thu → Lợi nhuận	1	0.0198	0.8883	0.0203	0.8867
	2	1.1402	0.3232	2.3762	0.3048
	3	1.5944	0.1945	5.0717	0.1666
	4	2.1137	0.0837	9.1280	0.0580
Doanh số → Doanh thu	1	0.1071	0.7440	0.1098	0.7404
	2	3.6360	<u>0.0293</u>	7.5775	<u>0.0226</u>
	3	12.5488	0.0000	39.9182	0.0000
	4	26.4478	0.0000	114.2172	0.0000
Doanh số → Vốn	1	0.1056	0.7458	0.1082	0.7422
	2	3.5450	<u>0.0320</u>	7.3880	<u>0.0249</u>
	3	11.9613	0.0000	38.0493	0.0000
	4	24.9293	0.0000	107.6594	0.0000
Doanh số → Lợi nhuận	1	0.1248	0.7244	0.1279	0.7206
	2	3.4504	<u>0.0349</u>	7.1907	<u>0.0275</u>
	3	12.5736	0.0000	39.9971	0.0000
	4	26.1165	0.0000	112.7864	0.0000

- **Mối quan hệ giữa các biến và outliers**

*Revenue* và *Profit*:

- Cả hai biến có nhiều outliers, đặc biệt vào các năm 2013-2015 và 2019-2020.

- Điều này có thể phản ánh những thay đổi đột ngột trong doanh thu và lợi nhuận do chính sách giá cả hoặc biến động thị trường.

*Units và COGS:*

- Số lượng outliers trong Units có xu hướng ổn định hơn so với Revenue và Profit.
- COGS có số lượng outliers thấp hơn, nhưng vẫn có sự gia tăng trong các năm 2019-2020.
- Điều này có thể chỉ ra rằng chi phí sản xuất ít biến động hơn so với doanh thu và lợi nhuận.

### Các mô hình chuỗi thời gian

Vì dữ liệu khó đoán, nhiều ngoại lai và nhiễu, chúng tôi thực hiện dự báo theo tổng doanh thu tháng. Kết quả trình bày tại [Bảng 8](#), đều không tốt và kém hơn mô hình học sâu đề xuất.

Bảng 8: Kết quả dự báo của các mô hình

Mô hình	R-squared	MAPE	RMSE
ARIMA	0.5589	0.3625	8,094,308.66
Prophet	0.5249	0.4041	8,400,599.31
XGBoost	-0.1005	0.5081	12,785,256.11
BSTS/UCM	0.5972	0.3875	7,735,224.28

### 3. Kết luận tổng thể

#### Đặc điểm kinh doanh và thị trường

- *Tăng trưởng doanh thu nhưng có xu hướng biến động mạnh*
  - Doanh thu có xu hướng tăng dài hạn, nhưng mức độ biến động cao, đặc biệt từ năm 2016-2020.
  - Biến động doanh thu có thể phản ánh sự cạnh tranh gay gắt, xu hướng tiêu dùng thay đổi hoặc tác động của thương mại điện tử.
- *Doanh số không ổn định và tác động đến lợi nhuận*
  - Doanh số không tăng trưởng đồng đều như doanh thu, có thể do thay đổi chiến lược giá hoặc dịch chuyển phân khúc khách hàng.
  - Doanh số có quan hệ nhân quả mạnh với lợi nhuận, cho thấy số lượng hàng bán ra ảnh hưởng đáng kể đến khả năng sinh lời.
- *Chi phí giá vốn có tương quan mạnh với doanh thu*
  - Tương quan gần bằng 1 giữa Revenue và COGS cho thấy công ty phụ thuộc nhiều vào chi phí đầu vào.

- Điều này có thể gây rủi ro tài chính nếu giá nguyên liệu biến động hoặc nếu công ty chưa tối ưu hóa biên lợi nhuận.
  - *Lợi nhuận không bị ảnh hưởng trực tiếp bởi doanh thu*
    - Kiểm định Granger cho thấy doanh thu không dự báo được lợi nhuận, nhưng doanh số có tác động mạnh.
    - Công ty có thể đang chi tiêu quá nhiều vào quảng cáo hoặc quản lý hàng tồn kho mà chưa tối ưu hóa lợi nhuận.
- 

### Phân tích rủi ro và cơ hội kinh doanh

- *Rủi ro kinh doanh*
    - Biến động giá nguyên vật liệu và chuỗi cung ứng ảnh hưởng đến chi phí sản xuất.
    - Ngành thời trang có tính chu kỳ cao, dễ bị tác động bởi xu hướng và thay đổi hành vi tiêu dùng.
    - Mức độ cạnh tranh cao từ các thương hiệu lớn và sự bùng nổ của thương mại điện tử làm giảm biên lợi nhuận.
  - *Cơ hội phát triển*
    - Công ty có thể tận dụng phân khúc cao cấp hoặc chiến lược giá linh hoạt để tối ưu lợi nhuận.
    - Thương mại điện tử là xu hướng tất yếu, cần đầu tư mạnh vào nền tảng bán hàng online và tối ưu trải nghiệm khách hàng.
    - Tối ưu hóa chi phí logistics và chuỗi cung ứng có thể giúp cải thiện biên lợi nhuận mà không cần tăng giá bán.
- 

### Chiến lược kinh doanh khuyến nghị

- *Tối ưu hóa doanh số và chiến lược giá*
  - Ứng dụng mô hình định giá động để tối đa hóa doanh thu theo nhu cầu thị trường.
  - Tăng doanh số bằng cách mở rộng danh mục sản phẩm hoặc triển khai chiến lược khuyến mãi hiệu quả hơn.
- *Kiểm soát chi phí và tối ưu chuỗi cung ứng*
  - Áp dụng mô hình just-in-time inventory để giảm chi phí lưu kho và cải thiện dòng tiền.
  - Da dạng hóa nguồn cung ứng để giảm rủi ro biến động giá nguyên liệu.
- *Chuyển dịch sang thương mại điện tử và DTC (Direct-to-Consumer)*
  - Đầu tư mạnh vào digital marketing, sử dụng AI để cá nhân hóa trải nghiệm khách hàng.

- Xây dựng hệ sinh thái bán lẻ online, tối ưu hóa kênh bán hàng đa nền tảng (website, Amazon, Shopify).
- *Tận dụng xu hướng thời trang bền vững và thị trường ngách*
  - Phát triển dòng sản phẩm thời trang bền vững để thu hút khách hàng quan tâm đến môi trường.
  - Tận dụng mô hình resale (bán lại) như một phần của chiến lược kinh doanh dài hạn.

## b. Tính chất nhóm

Phân tích tính chất nhóm trong dữ liệu, đặc biệt trong lĩnh vực kinh tế, có ý nghĩa quan trọng trong việc hiểu sâu hơn về cấu trúc của dữ liệu, nhận diện các quy luật tiềm ẩn và hỗ trợ ra quyết định.

### 1. *Tham số cụm*

Bảng 9: Kết quả phân cụm với nhóm 4 thuộc tính: doanh thu, doanh số, giá bán trung bình và lợi nhuận biên

Số cụm ( $k$ )	WCSS	Silhouette Score
2	5285.69	0.2807
3	3480.30	0.3229
4	2723.99	0.3405
5	2461.05	0.3230
6	1953.50	0.3388
<b>7</b>	<b>1768.24</b>	<b>0.3426</b>
8	1615.72	0.2928
9	1466.61	0.3005
<b>10</b>	<b>1265.51</b>	<b>0.3270</b>

Để tìm ra số  $k$  cụm tối ưu, chúng tôi kỳ vọng tối thiểu hóa độ đo WCSS và tối đa hóa độ đo Silhouette. Biên Pareto theo 2 độ đo này trình bày tại [Hình 18](#). Kết quả tối ưu tại  $k = 7$  và  $k = 10$ . Giảm thiểu tối đa số lượng cụm có thể để tránh phức tạp quan sát, chúng tôi đề xuất sử dụng  $k = 7$ . Dẫn đến kết quả như [Hình 19](#).

- *Cụm 0 - Sản phẩm có biên lợi nhuận cao nhưng doanh thu thấp*
  - Sản phẩm thuộc cụm này có mức lợi nhuận (Profit Margin) rất cao, nhưng doanh thu (Revenue) và số lượng bán (Units) lại thấp.
  - Giá trung bình (AvgPrice) thuộc mức trung bình đến thấp.
  - **Chiến lược:** Đây có thể là các sản phẩm ngách hoặc sản phẩm cao cấp nhưng thị trường hạn chế. Công ty có thể tận dụng kênh marketing tập trung hơn để tăng nhận diện thương hiệu và tìm kiếm khách hàng tiềm năng. Ngoài ra, có thể thử nghiệm tăng giá để tối ưu hóa lợi nhuận.
- *Cụm 1 - Sản phẩm giá thấp, doanh thu thấp, lợi nhuận biên trung bình*

- Đây là nhóm sản phẩm có giá rẻ, doanh thu thấp và lợi nhuận biên ở mức trung bình.
- Số lượng bán ra cũng thấp, cho thấy đây không phải sản phẩm phổ biến.
- **Chiến lược:** Nên xem xét lại danh mục sản phẩm này. Có thể thực hiện cải tiến hoặc loại bỏ các sản phẩm không hiệu quả. Nếu giữ lại, có thể áp dụng chiến lược khuyến mãi, giảm giá hoặc bán theo gói để tăng số lượng tiêu thụ.

- *Cụm 2 - Sản phẩm cao cấp, giá cao, lợi nhuận biên tốt*

- Sản phẩm trong cụm này có giá trung bình cao nhất, doanh thu tương đối ổn định và lợi nhuận biên tốt.
- Mặc dù số lượng bán ra không quá cao, nhưng giá cao giúp duy trì doanh thu lớn.
- **Chiến lược:** Đây có thể là các dòng sản phẩm cao cấp, công ty nên tập trung vào chiến lược thương hiệu, nhấn mạnh vào chất lượng và giá trị sản phẩm thay vì chạy đua giảm giá. Hợp tác với KOLs và làm chiến dịch truyền thông sẽ giúp tăng giá trị thương hiệu.

- *Cụm 3 - Sản phẩm chủ lực, doanh thu cao, số lượng tiêu thụ lớn*

- Đây là cụm có doanh thu cao nhất trong số các cụm, đi kèm với số lượng tiêu thụ rất lớn.
- Lợi nhuận biên duy trì ở mức khá, cho thấy sự ổn định của nhóm sản phẩm này.
- **Chiến lược:** Nên tiếp tục tập trung vào sản phẩm này, tối ưu hóa quy trình sản xuất để giảm chi phí và tăng biên lợi nhuận. Đồng thời, mở rộng phân phối để tiếp cận thêm khách hàng mới.

- *Cụm 4 - Sản phẩm giá rẻ, số lượng bán khá, nhưng lợi nhuận biên thấp*

- Đây là nhóm sản phẩm có giá bán thấp, số lượng tiêu thụ ở mức trung bình, nhưng biên lợi nhuận không cao.
- Doanh thu không quá nổi bật, cho thấy khó khăn trong việc tạo ra giá trị lớn từ nhóm này.
- **Chiến lược:** Có thể tập trung vào các chương trình giảm chi phí sản xuất hoặc đẩy mạnh bán hàng số lượng lớn để tạo ra doanh thu đáng kể hơn.

- *Cụm 5 - Sản phẩm doanh thu cực cao, lợi nhuận cao, số lượng tiêu thụ khổng lồ*

- Đây là nhóm có doanh thu và số lượng bán ra lớn nhất, đồng thời biên lợi nhuận vẫn duy trì ở mức tốt.
- Đây có thể là dòng sản phẩm chiến lược của công ty.
- **Chiến lược:** Tiếp tục đẩy mạnh sản xuất và tối ưu hóa chuỗi cung ứng. Cần tập trung giữ vững thị phần bằng các chiến lược duy trì khách hàng trung thành và cải tiến sản phẩm để gia tăng giá trị cho khách hàng.

- *Cụm 6 - Sản phẩm trung bình, lợi nhuận biên biến động*

- Nhóm sản phẩm này có doanh thu, số lượng tiêu thụ và lợi nhuận biên biến động khá lớn.
- Có thể đây là nhóm sản phẩm không có định vị rõ ràng hoặc có sự biến động theo mùa vụ.
- **Chiến lược:** Cần phân tích chi tiết hơn để xác định lý do biến động. Nếu sản phẩm có tính thời vụ, có thể tận dụng bằng cách đẩy mạnh bán hàng vào mùa cao điểm. Nếu biến động do chiến lược giá chưa ổn định, cần điều chỉnh để tối ưu hóa lợi nhuận.

## 2. Phân cụm toàn thời gian

### Phân cụm sản phẩm

- Phân cụm theo doanh thu và doanh số: [Hình 20](#).
- Phân cụm theo doanh thu và biên lợi nhuận: [Hình 21](#).
- Phân cụm theo trung bình và độ lệch chuẩn doanh thu: [Hình 22](#).
- Phân cụm theo trung bình và độ lệch chuẩn doanh số: [Hình 23](#).
- Phân cụm theo trung bình và độ lệch chuẩn lợi nhuận: [Hình 24](#).

### Phân tích thị trường

- *Sự tập trung dày đặc ở vùng doanh thu thấp*
  - Cả hai biểu đồ đều cho thấy phần lớn sản phẩm có doanh thu và doanh số thấp.
  - Phần lớn danh mục sản phẩm có thể thuộc nhóm tiêu chuẩn, có mức tiêu thụ vừa phải nhưng chưa đạt đến nhóm có hiệu suất vượt trội.
- *Sự tách biệt của một nhóm nhỏ sản phẩm có doanh thu rất cao*
  - Một số ít sản phẩm có doanh thu vượt trội, tạo thành một nhóm biệt lập trên trực x. Những sản phẩm này là những sản phẩm chủ lực đóng góp phần lớn doanh thu cho doanh nghiệp.
- *Sự phân bố chẽ về lợi nhuận biên ở nhóm sản phẩm thấp*
  - Biểu đồ doanh thu & lợi nhuận biên cho thấy một dải dày đặc ở vùng doanh thu thấp, nơi biên lợi nhuận của các sản phẩm này khá đồng đều. Điều này có thể gợi ý rằng nhiều sản phẩm có cấu trúc chi phí và lợi nhuận khá tương đồng.
- *Các nhóm sản phẩm không lan tỏa đồng đều mà có những khoảng trống rõ rệt*
  - Cụm sản phẩm có xu hướng hình thành theo từng mức độ tăng trưởng, chứ không phải phân bố ngẫu nhiên.
  - Một số sản phẩm có thể dễ dàng tăng trưởng mạnh và vượt trội hẳn, một số bị kẹt trong vùng trung bình và không thể bứt phá.

- Có thể có rào cản về giá cả, thị trường hoặc chiến lược marketing làm cho dữ liệu phân tầng thay vì liên tục.
- *Sự giãn cách lớn giữa các nhóm doanh thu cao và doanh thu thấp*
  - Biểu đồ cho thấy một sự cắt đứt rõ rệt giữa nhóm sản phẩm doanh thu thấp và nhóm doanh thu cao. Điều này gợi ý rằng việc chuyển đổi một sản phẩm từ doanh thu trung bình sang doanh thu cao có thể không đơn giản, mà cần một chiến lược cụ thể.

### Rủi ro và chiến lược kinh doanh

- *Rủi ro đầu tư*

- Phần lớn sản phẩm có doanh thu, doanh số hay lợi nhuận biên thấp hoặc trung bình nhưng ổn định. Ngược lại sản phẩm có đặc trưng trên cao thường biến động mạnh.
- Nhóm có độ lệch chuẩn doanh thu hay doanh số cao có thể bao gồm các sản phẩm bán theo mùa, sản phẩm bị ảnh hưởng bởi xu hướng ngắn hạn, chiến dịch marketing, xu hướng thị trường hoặc chương trình khuyến mãi.
- Nhóm có lợi nhuận biên cao nhưng biến động mạnh có thể là sản phẩm cao cấp hoặc có mức giá thay đổi theo thời gian.

- *Chiến lược kinh doanh*

- Nếu muốn ổn định dòng tiền, doanh nghiệp nên tập trung vào nhóm sản phẩm có doanh thu thấp nhưng ổn định. Ngoài ra nên uy trì nguồn cung và chiến lược marketing ổn định.
- Nếu chấp nhận rủi ro cao hơn, có thể đầu tư mạnh vào nhóm có doanh thu cao nhưng biến động lớn bằng cách tối ưu chiến lược bán hàng: Cân nhắc điều chỉnh giá, tối ưu chương trình khuyến mãi để kéo dài chu kỳ bán hàng.
- Tận dụng xu hướng thị trường: Xây dựng chiến lược marketing mạnh mẽ cho nhóm sản phẩm doanh thu và doanh số biến động cao.

### 3. Biên Pareto toàn sản phẩm

#### Trực quan hóa

- Biên Pareto theo doanh thu và doanh số: [Hình 25](#).
- Biên Pareto theo doanh thu và biên lợi nhuận: [Hình 26](#).
- Biến động biên Pareto theo doanh thu và doanh số qua các năm: [Hình 27](#).
- Biến động biên Pareto theo doanh thu và lợi nhuận biên qua các năm: [Hình 28](#).

#### Đặc điểm thị trường

- *Phân tầng thị trường rõ ràng*

- Có sự tách biệt rõ ràng giữa các nhóm sản phẩm, với một số sản phẩm vượt trội nằm trên biên Pareto, trong khi phần lớn các sản phẩm còn lại bị phân tán bên dưới.
- Thị trường có tính cạnh tranh cao, chỉ một số sản phẩm thực sự chiếm ưu thế, còn lại có hiệu suất thấp.

- *Sự đánh đổi giữa doanh số, doanh thu và lợi nhuận*

- Các sản phẩm bán chạy nhất không phải lúc nào cũng có lợi nhuận cao.
- Có sự xuất hiện của những sản phẩm lợi nhuận biên rất tốt nhưng chưa đạt doanh thu, doanh cao, cho thấy chúng có thể chưa được khai thác hết tiềm năng thị trường.
- Nhóm sản phẩm có doanh thu lớn nhưng lợi nhuận biên giảm dần có thể là do chi phí vận hành cao hoặc bị áp lực giá cạnh tranh.

- *Sự thay đổi về mối quan hệ giữa doanh số và doanh thu*

- Trong giai đoạn đầu (2010 - 2013), có nhiều sản phẩm có doanh số cao nhưng doanh thu chưa tối ưu, thể hiện thị trường có nhiều sản phẩm giá thấp được tiêu thụ với số lượng lớn.
- Từ 2014 trở đi, xu hướng chuyển dịch sang các sản phẩm có doanh thu cao hơn nhưng doanh số giảm dần, cho thấy có sự chuyển đổi từ việc bán số lượng lớn sang tối ưu hóa giá trị sản phẩm.
- Sản phẩm giá thấp dần bị thay thế hoặc mất ưu thế, có thể do chi phí sản xuất tăng hoặc người tiêu dùng thay đổi nhu cầu.

- *Xu hướng thay đổi về lợi nhuận biên*

- Giai đoạn đầu (2010 - 2012): Lợi nhuận biên khá cao nhưng có xu hướng giảm mạnh khi doanh thu tăng.
- Giai đoạn giữa (2013 - 2016): Lợi nhuận biên trở nên ổn định hơn, phản ánh rằng các công ty đã tối ưu được chi phí và định giá sản phẩm hiệu quả hơn.
- Giai đoạn sau (2017 - 2020): Lợi nhuận biên tiếp tục giảm với tốc độ nhanh hơn khi doanh thu tăng, có thể do thị trường trở nên cạnh tranh hơn, buộc các doanh nghiệp phải giảm giá hoặc tăng chi phí marketing, chi phí sản xuất tăng, làm ảnh hưởng đến lợi nhuận biên.

### **Kết luận tổng thể**

- *Đặc trưng của thị trường*

- Thị trường có xu hướng bị chi phối bởi một số ít sản phẩm mạnh, trong khi phần lớn các sản phẩm còn lại có mức đóng góp không đáng kể.
- Thị trường đang dần chuyển dịch từ bán số lượng lớn sang bán sản phẩm có giá trị cao hơn.
- Có sự đánh đổi giữa doanh thu, doanh số và lợi nhuận, cần một chiến lược cân bằng thay vì chỉ tập trung vào một yếu tố duy nhất.
- Lợi nhuận biên ngày càng giảm, cho thấy áp lực cạnh tranh và chi phí gia tăng.

- *Chiến lược tối ưu theo xu hướng thị trường*

- Duy trì và mở rộng nhóm sản phẩm hiệu suất cao, xây dựng chiến lược marketing mạnh mẽ để khai thác triệt để nhóm sản phẩm có tiềm năng doanh thu hoặc lợi nhuận biên cao.
- Định vị lại giá trị sản phẩm, hướng đến khách hàng sẵn sàng chi trả cao hơn thay vì chạy theo số lượng bán ra.
- Xây dựng chiến lược giá linh hoạt thay vì giảm giá liên tục để duy trì doanh thu. Tập trung vào công nghệ tối ưu hóa chuỗi cung ứng để giảm chi phí sản xuất và vận hành.

### c. Phân tích theo khu vực

Phân tích theo khu vực giúp hiểu rõ sự khác biệt về phát triển, tiêu dùng và chính sách giữa các vùng địa lý. Điều này hỗ trợ doanh nghiệp đưa ra quyết định phù hợp, tối ưu hóa phân bổ nguồn lực và thúc đẩy tăng trưởng bền vững. Ngoài ra, doanh nghiệp có thể sử dụng phân tích này để định vị thị trường, cải thiện chiến lược tiếp thị và mở rộng hoạt động.

Các kết luận có thể đưa ra dựa vào [Hình 7](#):

- Doanh thu theo vùng:

- East có doanh thu cao nhất, vượt xa so với Central và West.
- Central và West có doanh thu gần tương đương, nhưng vẫn thấp hơn East đáng kể.

- Doanh số bán hàng theo vùng:

- East cũng dẫn đầu về số lượng đơn hàng bán ra, điều này khẳng định rằng đây là thị trường lớn nhất.
- Central và West có doanh số thấp hơn nhiều so với East, cho thấy nhu cầu thấp hơn hoặc mạng lưới bán hàng chưa mở rộng bằng.

### 1. Trực quan hóa

- Trực quan hóa doanh thu (**Revenue**), doanh số (**Units**) và giá vốn (**COGS**) theo tháng, quý, năm theo từng khu vực: [Hình 29](#), [Hình 30](#), [Hình 31](#).
- Top 10 & Bottom 10 của từng khu vực dựa trên doanh thu theo phương pháp rank sum: [Hình 32](#).
- Ảnh hưởng của danh mục sản phẩm (**Category**) và phân khúc sản phẩm (**Segment**) theo số lượng sản phẩm bán ra (**Units**) và doanh thu (**Revenue**) của từng vùng: [Hình 33](#), [Hình 34](#).
- Ảnh hưởng của các thời điểm trong năm đến doanh thu: [Hình 35](#).

### 2. Kiểm chứng tính chất chuỗi thời gian

#### Ma trận tương quan

Biểu đồ trực quan tại [Hình 36](#) thể hiện ma trận tương quan Pearson giữa doanh thu (Revenue), doanh số (Units), giá vốn (COGS) và lợi nhuận (Profit) giữa các cặp vùng.

Các nhận xét rút ra, bao gồm:

- Tất cả các hệ số tương quan đều dương và khá cao (thường trên 0.8), cho thấy sự biến động của các chỉ số (doanh thu, doanh số, giá vốn, lợi nhuận) ở các vùng có xu hướng đồng nhất. Tức là khi một vùng có xu hướng tăng/giảm ở chỉ số nào đó, các vùng khác cũng có xu hướng tăng/giảm tương tự. Điều này gợi ý rằng các vùng có thể chịu tác động chung từ các yếu tố thị trường hoặc yếu tố thời vụ.
- Trong cả 4 biểu đồ, cặp vùng East - West (ví dụ) thường có tương quan cao nhất (xấp xỉ trên 0.9), cho thấy hai vùng này có sự dao động khá tương đồng.

### Dánh giá sự thay đổi phuơng sai

Trước hết, ta trực quan hóa tính tỷ lệ doanh thu trên số sản phẩm theo từng khu vực qua các năm tại [Hình 37](#). Qua đó, có nhận xét rút ra như sau:

- *Xu hướng tổng quan*

**Central** duy trì mức tỷ lệ doanh thu/sản phẩm cao nhất so với các vùng khác trong hầu hết các năm.

**East** có mức tỷ lệ doanh thu/sản phẩm khá cao nhưng có xu hướng giảm nhẹ từ năm 2012 đến 2017, sau đó ổn định.

**West** có tỷ lệ thấp nhất trong ba vùng nhưng có xu hướng tăng trưởng đều từ 2011 đến 2016 trước khi chững lại.

- *Phân tích theo từng vùng*

**Central:**

- Mặc dù có sự giảm nhẹ từ 2011 đến 2016, vùng này đã phục hồi mạnh vào năm 2017, có thể do điều chỉnh chiến lược giá hoặc sản phẩm.
- Sau 2017, tỷ lệ này ổn định và duy trì ở mức cao nhất so với các vùng khác.

**East:**

- Đạt đỉnh vào năm 2011, nhưng sau đó có xu hướng giảm dần cho đến 2017, phản ánh khả năng cạnh tranh giá có thể suy giảm.
- Từ 2018 trở đi, tỷ lệ này dần ổn định nhưng vẫn thấp hơn so với Central.

**West:**

- Có mức tỷ lệ doanh thu/sản phẩm thấp nhất trong ba vùng, nhưng có xu hướng tăng nhẹ từ 2011 đến 2016.
- Từ 2017 trở đi, tỷ lệ này không có nhiều thay đổi, cho thấy có thể vùng này chưa khai thác tốt giá trị trên mỗi sản phẩm bán ra.

- *Chiến lược kinh doanh theo vùng*

**Central:** Nên tập trung vào chiến lược duy trì lợi thế về giá trị sản phẩm, có thể thông qua tiếp thị hoặc gia tăng trải nghiệm khách hàng.

**East:** Cần cải thiện chiến lược giá hoặc chất lượng sản phẩm để ngăn chặn xu hướng giảm, có thể thông qua chiến dịch định giá linh hoạt hoặc ra mắt sản phẩm mới.

**West:** Có cơ hội cải thiện tỷ lệ doanh thu/sản phẩm bằng cách nâng cao giá trị cảm nhận hoặc đẩy mạnh chiến lược tiếp thị sản phẩm cao cấp hơn.

Tiếp theo, ta trực quan sự thay đổi phuơng sai của tỷ lệ doanh thu/số sản phẩm qua các năm tại [Hình 38](#). Từ đó ta thu được các đánh giá sau:

Đánh giá theo giai đoạn:

- Giai đoạn 2010-2012: Thị trường chưa ổn định, các vùng có thể áp dụng chiến lược giá khác nhau.
- Giai đoạn 2012-2016: Sự khác biệt giá giữa các vùng giảm dần, có thể do cạnh tranh mạnh hơn hoặc sự đồng bộ trong chính sách giá.
- Giai đoạn 2016-2020: Thị trường trở nên ổn định, không có sự chênh lệch lớn giữa các vùng.

Ý nghĩa kinh doanh:

- Sự suy giảm phuơng sai cho thấy chiến lược giá đang trở nên đồng nhất giữa các vùng.
- Nếu muốn khai thác thị trường, có thể cần tập trung vào các yếu tố khác ngoài giá cả (chất lượng, dịch vụ,...).
- Nếu muốn tăng lợi nhuận, có thể xem xét phân khúc khách hàng thay vì áp dụng mức giá đồng đều cho tất cả các vùng.

### Phân tích yếu tố mùa vụ theo khu vực

Trực quan hóa yếu tố mùa vụ tại [Hình 39](#). Các khía cạnh thú vị ta có thể xem xét:

Phân tích dữ liệu gốc (**Observed**):

- Các đường đều có dạng “gợn sóng” lặp lại, biểu hiện của tính mùa vụ (các đỉnh/đáy lặp lại theo chu kỳ).
- Ở **Revenue** lẫn **Units**, có thể thấy biên độ dao động khá đều, cho thấy rằng mùa vụ là yếu tố chính, chưa kể xu hướng tăng/giảm nhẹ.
- So sánh 3 vùng:
  - Khu vực **Central**: Đường Observed có dao động rõ, nhưng nhìn chung vẫn ổn định.
  - Khu vực **West**: Dao động cũng rõ nét, tùy giai đoạn có thể mức trung bình cao/thấp hơn Central.
  - Khu vực **East**: Tương tự, có các đỉnh/đáy đều đặn; độ cao của đỉnh (Revenue/Units) có thể khác so với 2 vùng còn lại.

Phân tích xu hướng (**Trend**):

- Xu hướng (Trend) thường thể hiện dài hạn: tăng/giảm hoặc đi ngang.
- Quan sát:
  - Ở cả Revenue và Units, đường Trend đều giảm dần về cuối. Tức là sau một giai đoạn ổn định/nhỉnh lên, xu hướng chung là đi xuống.
  - Mức độ giảm của Revenue thường rõ hơn so với Units. Điều này có thể gợi ý rằng giá bán trung bình (hoặc giá trị sản phẩm) giảm, vì số lượng bán ra (Units) không giảm quá mạnh nhưng doanh thu (Revenue) lại giảm đáng kể.
- So sánh 3 vùng:
  - **Central:** Trend giảm dần, nhưng không quá đột ngột.
  - **West:** Có vẻ giảm khá mạnh (đặc biệt với Revenue).
  - **East:** Cũng giảm về cuối chuỗi, có thời điểm xuống thấp hơn cả Central.
- Nhìn chung, Trend ở cả 3 vùng đều suy giảm, phản ánh có thể là giai đoạn thị trường chững lại hoặc giá bán giảm.

Thành phần thời vụ (**Seasonal**):

- Ở cả **Revenue** và **Units**, **Seasonal** có biên độ dao động khá đều, thể hiện rằng các tháng cao điểm/thấp điểm trong năm gần như lặp lại giống nhau qua các năm.
- Hình dạng **Seasonal** giữa 3 vùng tương đối giống nhau (các tháng đỉnh/dáy thường trùng nhau), cho thấy yếu tố mùa vụ có thể do đặc trưng chung (ví dụ: dịp lễ, Tết, mùa du lịch...).
- Biên độ **Seasonal** của **Revenue** có thể lớn hơn **Units** (vì doanh thu chịu ảnh hưởng cả giá bán, chương trình khuyến mãi, v.v.).

### Xây dựng mô hình VAR liên khu vực

Áp dụng mô hình VAR cho doanh thu theo tháng của 3 khu vực, với độ trễ (lag) tối đa là 12 (tức 1 năm). Kết quả tốt nhất theo độ đo AIC có độ trễ 12. Biểu đồ đáp ứng xung được trình bày tại [Hình 40](#). Quan sát biểu đồ này, chúng tôi đề xuất với doanh nghiệp những góc nhìn về tình trạng kinh doanh như sau:

- *Ảnh hưởng của cú sốc trong cùng một vùng*

**Central:** Cú sốc ban đầu tạo ảnh hưởng mạnh, sau đó giảm dần và dao động xung quanh mức cân bằng.

**East:** Ban đầu phản ứng tích cực nhưng suy yếu nhanh chóng, sau đó dao động không ổn định.

**West:** Có xu hướng giảm dần về mức cân bằng sau cú sốc, nhưng với biên độ dao động thấp hơn các vùng khác.

- *Ảnh hưởng của cú sốc từ một vùng lên các vùng khác*

**Cú sốc từ Central:**

- Ảnh hưởng mạnh nhất lên East, tạo ra phản ứng ban đầu dương nhưng nhanh chóng dao động mạnh.
- Ảnh hưởng lên West ít ổn định hơn, phản ứng biến động với chu kỳ rõ rệt.

### Cú sốc từ East:

- Tác động mạnh mẽ lên chính nó, có sự dao động lớn kéo dài.
- Ảnh hưởng lên Central và West khá rõ, nhưng phản ứng giảm dần theo thời gian.

### Cú sốc từ West:

- Ảnh hưởng lên Central và East có xu hướng tích cực ban đầu, nhưng sau đó biến động không ổn định.
- West có sự điều chỉnh chậm hơn so với các vùng khác.

- *Chiến lược kinh doanh theo vùng dựa trên tác động cú sốc*

**Central:** Do có ảnh hưởng mạnh đến các vùng khác, nên cần duy trì ổn định giá cả và chiến lược phân phối để tránh biến động tiêu cực lan rộng.

**East:** Cần điều chỉnh giảm thiểu sự biến động sau cú sốc, có thể bằng cách đa dạng hóa danh mục sản phẩm hoặc ổn định nguồn cung.

**West:** Ảnh hưởng từ cú sốc thường chậm và có chu kỳ, cần cải thiện khả năng thích ứng nhanh hơn để giảm độ trễ trong điều chỉnh chiến lược.

### Kiểm chứng nhân quả

Bảng 10: Kết quả kiểm định nhân quả Granger giữa các vùng (dữ liệu theo tháng)

Mối quan hệ	Lags	F Test		Chi-squared Test	
		F-stat	p-value	Chi2	p-value
Central → East	1	0.6383	0.4259	0.6540	0.4187
	2	1.6629	0.1940	3.4654	0.1768
	3	4.0618	0.0087	12.9207	0.0048
	4	5.6142	0.0004	24.2454	0.0001
	5	7.9002	0.0000	43.4513	0.0000
Central → West	1	1.5205	0.2199	1.5579	0.2120
	2	8.1337	0.0005	16.9509	0.0002
	3	20.2334	0.0000	64.3632	0.0000
	4	18.3715	0.0000	79.3388	0.0000
	5	23.6492	0.0000	130.0706	0.0000
East → Central	1	1.3861	0.2414	1.4202	0.2334
	2	0.9441	0.3919	1.9674	0.3739
	3	1.4591	0.2294	4.6414	0.2000
	4	2.5040	0.0462	10.8135	0.0287

Tiếp tục trang sau

*Tiếp tục từ trang trước*

Mối quan hệ	Lags	F Test		Chi-squared Test	
		F-stat	p-value	Chi2	p-value
East → West	5	3.7403	0.0036	20.5719	0.0010
	1	0.2061	0.6506	0.2112	0.6458
	2	5.4991	0.0052	11.4602	0.0032
	3	17.4755	0.0000	55.5903	0.0000
	4	13.9055	0.0000	60.0519	0.0000
	5	20.3367	0.0000	111.8516	0.0000
West → Central	1	0.9399	0.3342	0.9630	0.3264
	2	2.8797	0.0601	6.0014	0.0498
	3	4.6669	0.0041	14.8455	0.0020
	4	6.8004	0.0001	29.3680	0.0000
	5	10.9561	0.0000	60.2584	0.0000
	1	1.6440	0.2022	1.6845	0.1943
West → East	2	3.3118	0.0398	6.9018	0.0317
	3	7.1079	0.0002	22.6104	0.0000
	4	6.5187	0.0001	28.1514	0.0000
	5	15.5958	0.0000	85.7768	0.0000

Từ bảng trên cho thấy:

- **Kiểm định nhân quả Granger giữa các vùng**

*Tầm quan trọng của khu vực Central*

- Khu vực Central có ảnh hưởng mạnh mẽ đến East và West trong trung và dài hạn. Điều này gợi ý rằng Central có thể đóng vai trò là trung tâm kinh tế hoặc thị trường cốt lõi của công ty.
- Khi Central có biến động về doanh thu hoặc chiến lược, các khu vực khác sẽ chịu tác động đáng kể, đặc biệt là sau một vài giai đoạn (khoảng 3-5 tháng).
- Điều này có ý nghĩa chiến lược: công ty nên tập trung vào việc tối ưu hóa hoạt động tại Central để gián tiếp tác động tích cực lên các vùng khác.

*Sự phụ thuộc của West vào Central và East*

- West dường như phụ thuộc nhiều vào Central và East, khi kết quả kiểm định cho thấy có mối quan hệ nhân quả mạnh từ hai khu vực này đến West.
- Điều này có thể phản ánh đặc điểm chuỗi cung ứng hoặc hệ thống phân phối: hàng hóa, xu hướng thị trường hoặc quyết định kinh doanh tại Central và East có thể quyết định đến doanh thu tại West.
- Nếu công ty muốn nâng cao hiệu quả hoạt động ở West, họ cần điều chỉnh chiến lược tại Central và East, thay vì chỉ tối ưu riêng khu vực West.

*Vai trò hạn chế của East trong việc ảnh hưởng đến Central*

- Mặc dù Central ảnh hưởng mạnh đến East, chiều ngược lại không mạnh bằng. Điều này cho thấy khu vực Central có thể có vai trò dẫn dắt, trong khi East chủ yếu phản ứng với những biến động từ Central.
- Công ty không nên kỳ vọng East sẽ là động lực tăng trưởng chính để thúc đẩy Central, thay vào đó nên tận dụng Central như một động lực tác động đến toàn bộ hệ thống.

#### *Ứng dụng chiến lược kinh doanh từ kết quả kiểm định*

- Công ty nên tập trung tối ưu hóa doanh thu tại Central, vì khu vực này có khả năng tác động mạnh mẽ đến các khu vực khác.
- Cần xây dựng chiến lược điều chỉnh tại Central trước khi mở rộng các sáng kiến kinh doanh sang East hoặc West.
- West có thể được xem như một khu vực chịu ảnh hưởng nhiều từ quyết định kinh doanh tại Central và East, do đó cần đảm bảo sự phối hợp giữa các vùng này để tránh mất cân bằng trong hệ thống phân phối và doanh thu.

### **d. Phân tích theo ngành hàng**

Trong phần này, chúng tôi thực hiện phân tích dữ liệu theo từng danh mục sản phẩm nhằm rút ra các đặc điểm quan trọng về xu hướng tiêu dùng, chiến lược giá và phân khúc thị trường.

#### Danh mục sản phẩm

- **Urban:** Sản phẩm phục vụ nhu cầu tiêu dùng ở các khu vực đô thị với cơ sở hạ tầng phát triển.
- **Rural:** Sản phẩm phục vụ khách hàng ở khu vực nông thôn với nhu cầu cơ bản và giá cả hợp lý.
- **Mix:** Sản phẩm phù hợp cho cả thành thị và nông thôn, hoặc sự kết hợp giữa các yếu tố khác nhau.
- **Youth:** Sản phẩm nhắm đến giới trẻ, yêu thích sự sáng tạo, thời trang và xu hướng mới.

#### Phân khúc thị trường

- **Productivity:** Hướng đến khách hàng tìm kiếm sản phẩm tăng năng suất công việc hoặc cuộc sống.
- **Convenience:** Dành cho khách hàng ưu tiên sự tiện lợi, dễ sử dụng và tiết kiệm thời gian.
- **Moderation:** Nhắm đến khách hàng có nhu cầu tiêu dùng trung bình, cân bằng giữa giá cả và chất lượng.
- **Extreme:** Dành cho khách hàng có nhu cầu cao, sẵn sàng chi trả cho các sản phẩm cao cấp.
- **Youth:** Nhắm đến khách hàng trẻ tuổi, thích các sản phẩm thời trang, sáng tạo và theo xu hướng.
- **Select:** Nhắm đến khách hàng chọn lọc, tìm kiếm sản phẩm đặc biệt nhưng không quá xa xỉ.

- **All Season:** Các sản phẩm sử dụng quanh năm, không phụ thuộc vào mùa vụ.
- **Regular:** Dành cho khách hàng tìm kiếm sản phẩm ổn định, bền bỉ và lâu dài.

### 1. Phân tích theo danh mục (category)

#### Trực quan hóa

- Trực quan hóa doanh thu (Revenue), doanh số (Units) theo Category: [Hình 41](#).
- Trực quan hóa biên lợi nhuận và tổng lợi nhuận theo Category: [Hình 42](#).

#### Xu hướng theo thời gian

- Trực quan hóa doanh thu (Revenue), doanh số (Units), lợi nhuận, biên lợi nhuận theo tuần, tháng, quý, năm: [Hình 43](#).

Chúng tôi đề xuất chiến lược đầu tư hợp lý theo từng nhóm như sau:

**Urban (Khu vực thành thị):** Đây từng là phân khúc mạnh nhất nhưng đang giảm mạnh từ 2016.

- Ngắn hạn: Nếu có cơ hội phục hồi, có thể đầu tư lướt sóng khi giá thấp.
- Dài hạn: Tránh đầu tư lớn vì xu hướng giảm kéo dài.
- Tái cơ cấu: Tìm cách đa dạng hóa sản phẩm hoặc dịch vụ để giữ khách hàng.

**Rural (Khu vực nông thôn):** Có sự ổn định nhưng đang suy giảm dần.

- Đầu tư vào các sản phẩm thiết yếu: Tập trung vào các mặt hàng có nhu cầu dài hạn.
- Tối ưu hóa chi phí: Doanh thu giảm nhưng biên lợi nhuận vẫn tốt, nên cần cắt giảm chi phí vận hành để duy trì lợi nhuận.
- Tiếp thị & mở rộng thị trường: Có thể mở rộng thị trường bằng các sản phẩm giá rẻ hơn để phù hợp với nhu cầu nông thôn.

**Youth (Phân khúc giới trẻ):** Biên lợi nhuận cao nhất và ổn định.

- Tăng cường đầu tư: Đây là phân khúc có tiềm năng nhất để phát triển bền vững.
- Đổi mới sản phẩm: Cập nhật xu hướng để phù hợp với nhu cầu giới trẻ.
- Tận dụng Marketing số: Dùng nền tảng kỹ thuật số để tiếp cận nhóm khách hàng trẻ, tối ưu lợi nhuận.

**Mix (Hỗn hợp):** Doanh thu và lợi nhuận thấp, biên lợi nhuận cũng không cao.

- Chỉ đầu tư ngắn hạn: Không nên đặt kỳ vọng lớn vào phân khúc này.
- Tìm kiếm cơ hội niche (thị trường ngách): Tập trung vào phân khúc đặc biệt, thay vì đầu tư dàn trải.
- Tái cấu trúc hoặc hợp tác: Cân nhắc sáp nhập hoặc hợp tác với các danh mục khác để cải thiện hiệu suất.

## Tần suất mua hàng

Tần suất mua hàng được xác định bằng số ngày trung bình giữa các lần mua lại của từng loại sản phẩm (Category), tính theo từng mã Zip, bang (State), thành phố (City) và khu vực (Region), trong bốn nhóm sản phẩm: Mix, Rural, Urban và Youth.

Bảng 11: Tần suất mua hàng theo Zip Code, Thành phố, Bang và Vùng

Địa lý	Khu vực	Tần suất mua hàng theo danh mục			
		Mix	Rural	Urban	Youth
Zip Code	1001	-	175.17	103.56	1225.0
	1002	-	339.25	369.22	-
	1005	-	207.57	226.19	-
	1007	656.5	182.23	79.58	-
	:	:	:	:	:
Thành phố	Abbeville, AL	-	-	631.40	-
	Abbeville, GA	-	0.00	418.00	-
	Abbeville, LA	-	128.81	90.35	1900.0
	Zwolle, LA	-	0.00	401.50	-
	:	:	:	:	:
Bang	AK	17.76	2.64	2.69	29.59
	AL	11.30	0.67	0.45	8.18
	AR	12.41	1.19	0.63	5.29
	TX	2.81	0.23	0.10	1.59
	:	:	:	:	:
Vùng	Central	0.540	0.049	0.025	0.259
	East	0.340	0.030	0.015	0.179
	West	0.317	0.033	0.038	0.365

Từ các bảng trên, có thể thấy:

- Các khu vực đô thị (Urban) có tần suất mua hàng cao hơn tại một số thành phố lớn.
- Phân khúc Youth có xu hướng tập trung vào một số ít địa điểm có lượng mua cao.
- Khu vực Rural có mức mua ẩn định hơn, trải rộng trên nhiều bang khác nhau.
- Khu vực West có tần suất mua hàng Youth cao hơn so với các vùng khác.
- Central có mức mua hàng cao hơn trong danh mục Mix, phản ánh xu hướng tiêu dùng đặc thù của khu vực này.
- Trực quan hóa độ tương quan tần suất mua hàng giữa các Category: [Hình 44](#).

## Phân tích rủi ro

Bảng 12: Thống kê lợi nhuận theo Category

Loại chỉ số	Category	Giá trị			
		Mix	Rural	Urban	Youth
Standard Deviation	Profit	1019.93	655.89	1700.44	539.26
	Profit Margin	0.169340	0.171602	0.173159	0.161742
Skewness	Profit	4.9405	8.9327	4.0126	14.4896
	Profit Margin	0.5387	-0.0797	0.1862	-0.3920
Kurtosis	Profit	161.6989	446.3366	111.7757	971.1284
	Profit Margin	-1.0519	-1.1469	-1.2138	-0.8390

- Trực quan hóa độ lệch chuẩn của biên lợi nhuận và tổng lợi nhuận theo Category: [Hình 45](#).
- Trực quan hóa độ lệch chuẩn của biên lợi nhuận và tổng lợi nhuận của các thành phần Category theo thời gian: [Hình 46](#).
- Trực quan hóa phân phối lợi nhuận và biên lợi nhuận theo Category: [Hình 47](#).

Các nhận xét rút ra từ [Bảng 12](#) như sau:

- **Phân tích Standard Deviation (Độ lệch chuẩn)**

*Profit:*

- Nhóm Urban có độ lệch chuẩn cao nhất (1700.44), cho thấy lợi nhuận tại khu vực đô thị có biến động mạnh nhất.
- Ngược lại, nhóm Rural có độ lệch chuẩn thấp nhất (655.89), phản ánh lợi nhuận ổn định hơn ở khu vực nông thôn.
- Nhóm Youth có độ lệch chuẩn thấp (539.26), cho thấy thị trường giới trẻ có ít biến động lợi nhuận hơn.

*Profit Margin:*

- Biên lợi nhuận dao động rất ít giữa các nhóm, với mức cao nhất ở Urban (0.173159) và thấp nhất ở Youth (0.161742).
- Điều này cho thấy công ty có chiến lược giá tương đối ổn định, không có sự khác biệt quá lớn về biên lợi nhuận giữa các nhóm khách hàng.

- **Phân tích Skewness (Độ lệch)**

*Profit:*

- Nhóm Youth có skewness cao nhất (14.4896), cho thấy lợi nhuận ở nhóm này bị lệch phải mạnh, có nhiều giá trị lợi nhuận cao bất thường.

- Ngược lại, nhóm Rural có skewness cao (8.9327), cũng phản ánh một số giá trị lợi nhuận lớn bất thường.
- Nhóm Urban có skewness thấp nhất (4.0126), cho thấy sự phân phối lợi nhuận ít bất đối xứng hơn.

*Profit Margin:*

- Các giá trị skewness nhỏ hơn nhiều so với Profit, nghĩa là phân phối biên lợi nhuận ít bị ảnh hưởng bởi outliers.
- Nhóm Mix có skewness cao nhất (0.5387), nhưng vẫn thấp so với Profit, cho thấy sự ổn định trong biên lợi nhuận.

#### • Phân tích Kurtosis (Độ nhọn)

*Profit:*

- Giá trị Kurtosis rất cao ở tất cả các nhóm, đặc biệt ở Youth (971.1284), cho thấy lợi nhuận có nhiều outliers cực lớn.
- Nhóm Rural cũng có Kurtosis cao (446.3366), nghĩa là có một số giá trị lợi nhuận đột biến tại khu vực này.
- Urban có Kurtosis thấp hơn (111.7757), cho thấy lợi nhuận tại đô thị có sự phân tán ít cực đoan hơn.

*Profit Margin:*

- Tất cả các nhóm đều có Kurtosis âm, phản ánh biên lợi nhuận có phân phối gập với chuẩn hơn và ít outliers cực đoan hơn.
- Điều này cho thấy công ty có chính sách giá tương đối ổn định giữa các nhóm khách hàng.

Dựa trên phân tích trên, chiến lược đầu tư có thể được đề xuất dựa theo mức độ chấp nhận rủi ro:

- **Chiến lược rủi ro cao - lợi nhuận cao:** Danh mục **Urban** có mức độ biến động lớn, tạo cơ hội thu lời cao nhưng cũng có rủi ro lớn. Đầu tư vào danh mục này đòi hỏi khả năng chịu rủi ro cao, cùng với chiến lược quản lý rủi ro như phân tích xu hướng và cắt lỗ kịp thời.
- **Chiến lược cân bằng rủi ro - lợi nhuận:** Danh mục **Mix, Rural** có rủi ro ở mức trung bình và lợi nhuận ổn định theo thời gian. Đây là lựa chọn phù hợp nếu muốn có sự ổn định nhưng vẫn chấp nhận một mức độ rủi ro vừa phải. Chiến lược đa dạng hóa có thể giúp giảm thiểu rủi ro.
- **Chiến lược an toàn - rủi ro thấp:** Danh mục **Youth** có lợi nhuận thấp nhất nhưng cũng ít biến động nhất. Đây là lựa chọn phù hợp cho các khoản đầu tư dài hạn và ít rủi ro, có thể kết hợp với chiến lược thu nhập thụ động.

Tóm lại, nếu chấp nhận rủi ro cao, có thể đầu tư vào **Urban** để tối đa hóa lợi nhuận. Nếu muốn sự cân bằng giữa rủi ro và lợi nhuận, **Mix** và **Rural** là lựa chọn tốt. Nếu muốn đầu tư an toàn, ít rủi ro, nên chọn **Youth**.

---

## Kiểm tra tương quan giữa các thành phần của Category

---

- Trực quan hóa tương quan giữa các thành phần của Category theo doanh thu [Hình 48](#) và doanh số [Hình 49](#).

Các kết luận chung tôi đưa ra, bao gồm:

- Tương quan giữa Category theo Doanh thu*

*Theo tháng:*

- **Mix** và **Youth** có tương quan cao (0.83), cho thấy hai phân khúc này có xu hướng doanh thu đồng biến.
- **Urban** cũng có mối liên hệ chặt chẽ với cả **Mix** (0.68) và **Youth** (0.81), phản ánh sự liên kết giữa đô thị và các sản phẩm đa phân khúc.
- **Rural** có mức tương quan yếu với các phân khúc còn lại, thậm chí có tương quan âm với **Mix** (-0.09), cho thấy Rural hoạt động độc lập với các khu vực khác.

*Theo quý:*

- **Mix** và **Youth** có tương quan cao nhất (0.86), phản ánh sự tăng trưởng doanh thu của hai phân khúc này diễn ra đồng thời theo quý.
- **Urban** cũng có mối liên hệ mạnh với **Mix** (0.67) và **Youth** (0.84).
- **Rural** tiếp tục có mức tương quan thấp với các phân khúc còn lại, chỉ đạt 0.11 với **Youth**.

*Theo năm:*

- **Mix** và **Youth** vẫn giữ mối liên kết mạnh (0.85), cho thấy chiến lược kinh doanh ở hai phân khúc này có sự gắn kết chặt chẽ trong dài hạn.
- **Urban** có tương quan cao với **Rural** (0.81), cho thấy sự chuyển dịch dòng tiền giữa đô thị và khu vực nông thôn.
- **Rural** vẫn có mức tương quan thấp với **Youth** (0.34), cho thấy hai phân khúc này không ảnh hưởng nhiều đến nhau.

- Tương quan giữa Category theo Doanh số*

*Theo tháng:*

- **Mix** và **Youth** có mức tương quan cao nhất (0.84), phản ánh hai phân khúc này có mô hình bán hàng tương tự nhau theo thời gian ngắn.
- **Urban** có liên kết mạnh với cả **Mix** (0.73) và **Youth** (0.83), cho thấy sự tương đồng về chiến lược bán hàng.
- **Rural** có mức tương quan rất thấp với các phân khúc còn lại, đặc biệt với **Mix** (-0.05), chứng tỏ chiến lược kinh doanh ở **Rural** có sự khác biệt rõ rệt.

*Theo quý:*

- Mỗi quan hệ giữa **Mix** và **Youth** tiếp tục mạnh (0.86), củng cố xu hướng bán hàng đồng bộ giữa hai phân khúc này.
- **Urban** có mối liên hệ mạnh với cả **Mix** (0.72) và **Youth** (0.86), phản ánh sự tương quan trong chiến lược tiêu thụ sản phẩm.
- **Rural** duy trì mức tương quan thấp với các phân khúc khác, chỉ đạt 0.17 với Youth.

*Theo năm:*

- **Mix** và **Youth** tiếp tục duy trì mức tương quan cao (0.85), thể hiện tính gắn kết chặt chẽ trong mô hình kinh doanh dài hạn.
- **Urban** có mối liên hệ đáng kể với **Rural** (0.77), cho thấy sự kết nối giữa doanh số bán hàng ở hai phân khúc này theo năm.
- **Rural** vẫn có mức tương quan thấp với Youth (0.38), thể hiện sự khác biệt trong mô hình tiêu thụ sản phẩm giữa hai phân khúc này.

### Kiểm tra yếu tố mùa vụ

- Trực quan hóa yếu tố mùa vụ: [Hình 50](#).

Giải nghĩa cho hình ảnh trực quan của từng thành phần như sau:

- **Mix**

- **Xu hướng:** Tăng mạnh đến 2016-2017, sau đó giảm nhẹ và hồi phục sau 2018.
- **Mùa vụ:** Dao động theo chu kỳ rõ ràng, biên độ lớn.
- **Phản dư:** Biến động cao, có một số điểm bất thường.

- **Rural**

- **Xu hướng:** Giảm dần từ 2015, cho thấy sự suy giảm dài hạn.
- **Mùa vụ:** Rõ ràng nhưng không quá mạnh, có tính lặp lại theo chu kỳ.
- **Phản dư:** Biến động lớn, có nhiều điểm lệch so với xu hướng.

- **Urban**

- **Xu hướng:** Tăng đến 2016-2017, sau đó giảm mạnh và phục hồi nhẹ.
- **Mùa vụ:** Tính chu kỳ mạnh, dao động đều.
- **Phản dư:** Biến động lớn, có những đợt tăng giảm đột ngột.

- **Youth**

- **Xu hướng:** Tăng mạnh đến 2016, sau đó giảm nhanh rồi ổn định.
- **Mùa vụ:** Chu kỳ rõ ràng, dao động mạnh.
- **Phản dư:** Có một số điểm đột biến lớn, đặc biệt giai đoạn 2016.

Các danh mục đều có tính mùa vụ rõ ràng, nhưng xu hướng dài hạn khác nhau. **Urban** và **Mix** có biến động mạnh nhất, trong khi **Rural** suy giảm dài hạn. **Youth** có nhiều biến động đột ngột, có thể do sự kiện đặc biệt.

### Hiệu suất của từng Category

Bảng 13: Tổng hợp cửa hàng, vùng có doanh thu và lợi nhuận cao nhất cũng như thời điểm bán chạy của các danh mục

Category	Cửa hàng		Vùng		Tháng	Quý	Năm
	Doanh thu	Lợi nhuận	Doanh thu	Lợi nhuận			
Urban	Houston, TX	Houston, TX	East	East	4	2014Q2	2014
Rural	Miami, FL	Miami, FL	West	West	12	2011Q4	2011
Mix	Miami, FL	Miami, FL	West	West	4	2016Q2	2016
Youth	New York, NY	New York, NY	East	East	6	2016Q2	2016

### Kiểm định

Bảng 14: Kiểm định ảnh hưởng của mùa và ngày lễ đến doanh thu và doanh số

Category	ANOVA - Mùa		T-Test - Ngày lễ	
	Doanh thu	Đơn vị bán	Doanh thu	Đơn vị bán
Urban	0.000000	$1.85 \times 10^{-7}$	$1.58 \times 10^{-10}$	0.0841
Rural	$2.27 \times 10^{-285}$	$1.95 \times 10^{-13}$	$3.24 \times 10^{-3}$	$1.3 \times 10^{-5}$
Mix	$4.78 \times 10^{-20}$	0.3645	$3.23 \times 10^{-4}$	0.1845
Youth	0.0230	0.0060	0.6853	0.9127

#### Urban:

- Mùa có ảnh hưởng đến cả doanh thu và doanh số ( $p < 0.05$ ).
- Ngày lễ có ảnh hưởng đến doanh thu nhưng không ảnh hưởng đến doanh số.

#### Rural:

- Mùa có ảnh hưởng đáng kể đến cả doanh thu và doanh số.
- Ngày lễ có ảnh hưởng đến cả doanh thu và doanh số.

#### Mix:

- Mùa có ảnh hưởng đến doanh thu nhưng không ảnh hưởng đáng kể đến doanh số.
- Ngày lễ có ảnh hưởng đến doanh thu nhưng không ảnh hưởng đến doanh số.

#### Youth:

- Mùa có ảnh hưởng đến cả doanh thu và doanh số.
- Ngày lễ không có ảnh hưởng đáng kể đến doanh thu và doanh số.

## 2. Phân tích theo phân khúc (segment)

### Trực quan hóa

- Trực quan hóa doanh thu (Revenue), doanh số (Units) theo Segment: [Hình 51](#).
- Trực quan hóa biên lợi nhuận và tổng lợi nhuận theo Segment: [Hình 52](#).

Những nhận định cơ bản:

- Convenience có doanh thu cao nhất, cho thấy nhu cầu mạnh mẽ đối với các sản phẩm tiện lợi, dễ sử dụng và phù hợp với lối sống bận rộn. Đây là phân khúc chiến lược quan trọng.
- Moderation và Extreme có doanh thu ổn định, phục vụ nhu cầu tiêu dùng trung bình và cao hơn. Các sản phẩm trong các phân khúc này có thể tập trung vào chất lượng và tính năng đặc biệt, thu hút khách hàng tìm kiếm sự khác biệt.
- Youth và Regular có doanh thu thấp nhất, có thể do chưa khai thác được hết tiềm năng của nhóm khách hàng này. Cần cải thiện chiến lược tiếp thị và sáng tạo sản phẩm để thu hút sự chú ý từ đối tượng trẻ tuổi và những khách hàng tìm kiếm sự ổn định.

### Xu hướng theo thời gian

- Trực quan hóa doanh thu (Revenue), doanh số (Units), lợi nhuận, biên lợi nhuận theo tuần, tháng, quý, năm: [Hình 53](#).

### Tần suất mua hàng

Tần suất mua hàng được xác định bằng số ngày trung bình giữa các lần mua lại của từng loại sản phẩm (Segment), tính theo từng mã Zip, bang (State), thành phố (City) và khu vực (Region), trong bốn nhóm sản phẩm: Mix, Rural, Urban và Youth.

Bảng 15: Tần suất mua hàng theo Zip Code, Thành phố, Bang và Vùng cho từng phân khúc

Địa lý	Khu vực	Tần suất mua hàng theo phân khúc							
		All Season	Convenience	Extreme	Moderation	Productivity	Regular	Select	Youth
Zip Code	1001	-	150.06	373.50	363.71	120.56	-	1051.0	1225.0
	1002	-	2168.00	429.60	-	542.80	-	308.0	-
	1005	-	244.54	-	2027.00	223.54	-	-	-
	1007	656.5	185.19	249.25	213.88	183.00	-	2544.0	-
	:	:	:	:	:	:	:	:	:
Thành phố	Abbeville, AL	-	-	746.00	1518.00	-	-	-	-
	Abbeville, GA	-	-	-	-	0.00	-	-	-
	Abbeville, LA	-	140.00	164.29	273.86	133.96	-	-	1900.0
	Zwolle, LA	-	535.33	-	-	0.00	-	-	-
	:	:	:	:	:	:	:	:	:
Bang	AK	17.76	5.46	11.20	11.04	3.09	66.47	17.86	29.59
	AL	11.79	0.93	1.96	1.69	0.73	27.28	8.40	8.18
	AR	12.66	1.08	2.96	3.21	1.31	38.50	12.92	5.29
	TX	2.94	0.20	0.42	0.38	0.26	4.29	1.94	1.59
	:	:	:	:	:	:	:	:	:
Vùng	Central	0.561	0.051	0.128	0.086	0.056	1.160	0.415	0.259
	East	0.357	0.033	0.062	0.061	0.034	0.613	0.257	0.179
	West	0.345	0.094	0.121	0.160	0.037	0.918	0.259	0.365

Bảng trên cho thấy:

- Các phân khúc có sự phân bố khác nhau theo địa lý, với sự khác biệt đáng kể trong tần suất mua hàng.
- Phân khúc Youth có xu hướng tập trung cao tại một số khu vực nhất định như các thành phố lớn và vùng West.
- Phân khúc Productivity và Convenience có sự trải rộng đều hơn trên các khu vực khác nhau.
- Các khu vực đô thị (Urban) và vùng Central có tần suất mua hàng cao hơn trong các phân khúc All Season và Select.
- Vùng East có xu hướng tiêu dùng thấp hơn so với các vùng khác trong đa số phân khúc.
- Trực quan hóa độ tương quan tần suất mua hàng giữa các Segment: [Hình 54](#).

### Phân tích rủi ro

Loại chỉ số	Segment	Giá trị							
		All Season	Convenience	Extreme	Moderation	Productivity	Regular	Select	Youth
Standard Deviation	Profit	943.50	1352.70	1190.63	2360.35	657.78	1079.63	739.20	539.26
	Profit Margin	0.1637	0.1650	0.1871	0.1737	0.1717	0.1514	0.1622	0.1617
Skewness	Profit	1.5437	1.1572	1.5311	4.4256	9.2878	7.0797	11.6385	14.4896
	Profit Margin	0.5778	-0.0259	0.5326	0.2113	-0.1852	0.0618	0.7178	-0.3920
Kurtosis	Profit	4.5907	6.6246	10.2079	106.0322	481.9063	207.8706	527.4488	971.1284
	Profit Margin	-1.0038	-1.1350	-1.1344	-1.3041	-1.1121	-0.7637	-0.5268	-0.8390

Bảng trên cho thấy:

- Phân khúc Moderation có độ lệch chuẩn tổng lợi nhuận cao nhất, cho thấy sự biến động lớn.
- Phân khúc Youth có độ lệch chuẩn thấp nhất trong cả lợi nhuận và biên lợi nhuận, phản ánh tính ổn định.
- Phân khúc Select và Productivity có skewness lợi nhuận cao, cho thấy một số điểm dữ liệu có giá trị lớn đáng kể.
- Phân khúc Convenience có skewness biên lợi nhuận gần 0, phản ánh phân phối cân đối.
- Phân khúc Extreme có độ lệch chuẩn biên lợi nhuận cao nhất, cho thấy sự chênh lệch lớn trong lợi nhuận biên.
- Trực quan hóa độ lệch chuẩn của biên lợi nhuận và tổng lợi nhuận theo Segment: [Hình 55](#).
- Trực quan hóa độ lệch chuẩn của biên lợi nhuận và tổng lợi nhuận của các thành phần Segment theo thời gian: [Hình 56](#).
- Trực quan hóa phân phối lợi nhuận và biên lợi nhuận theo Segment: [Hình 57](#).

Phân phối lợi nhuận cho thấy các danh mục có lợi nhuận cao nhưng biến động mạnh gồm **Moderation, Select, Youth, Regular, Productivity**, trong đó **Youth có mức biến động cao nhất**. Ngược lại, **Convenience và All Season có lợi nhuận thấp hơn nhưng ổn định hơn**.

Về biến lợi nhuận, **All Season và Convenience ít biến động nhất**, trong khi **Select, Extreme và Youth dao động lớn**, phản ánh mức rủi ro cao.

Xét độ lệch chuẩn lợi nhuận theo thời gian, **Moderation và Extreme có độ lệch chuẩn cao nhất**, thể hiện rủi ro lớn. **Youth có độ lệch chuẩn thấp nhưng skewness cao**, cho thấy khả năng xuất hiện lợi nhuận cực lớn hoặc cực thấp.

Dựa trên phân tích, có ba chiến lược đầu tư chính:

- **Rủi ro cao - lợi nhuận cao:** Youth, Select, Productivity, Extreme. Thích hợp cho nhà đầu tư chịu rủi ro cao, cần áp dụng các biện pháp quản lý rủi ro như VaR hoặc hedging.
- **Cân bằng rủi ro - lợi nhuận:** Convenience, Regular, Select. Phù hợp với nhà đầu tư mong muốn lợi nhuận ổn định nhưng vẫn có cơ hội sinh lời.
- **An toàn - rủi ro thấp:** All Season, Convenience. Dành cho nhà đầu tư ưu tiên sự ổn định và bảo toàn vốn.

Tóm lại, nếu bạn chấp nhận rủi ro cao, hãy đầu tư vào **Youth, Select, Productivity, Extreme**. Nếu muốn cân bằng, chọn **Convenience, Regular, Select**. Nếu ưu tiên an toàn, **All Season và Convenience** là lựa chọn tối ưu.

### Kiểm tra tương quan giữa các thành phần

- Trực quan hóa tương quan giữa các thành phần của Segment theo doanh thu [Hình 58](#) và doanh số [Hình 59](#).

Về phần này, quan điểm của chúng tôi như sau:

- *Tương quan giữa Segment theo Doanh thu*

*Theo tháng:*

- **Extreme** và **Convenience** có tương quan rất cao (0.94), cho thấy hai phân khúc này có xu hướng tăng trưởng doanh thu đồng thời.
- **All Season** có mức tương quan mạnh với **Moderation** (0.88) và **Youth** (0.87), phản ánh chiến lược định vị tương đồng.
- **Productivity** có mức tương quan thấp hoặc âm với hầu hết các phân khúc khác, cho thấy phân khúc này hoạt động độc lập hơn.

*Theo quý:*

- Mối quan hệ giữa **Extreme** và **Convenience** tiếp tục duy trì ở mức cao (0.94), phản ánh sự gắn kết chiến lược dài hạn giữa hai phân khúc này.
- **All Season** có mức tương quan rất cao với **Moderation** (0.91) và **Youth** (0.89), cho thấy các dòng sản phẩm có tính mùa vụ hoặc nhắm đến giới trẻ có xu hướng tăng trưởng doanh thu đồng bộ.
- **Productivity** vẫn có mức tương quan thấp với các phân khúc còn lại, phản ánh sự khác biệt trong chiến lược tiếp cận thị trường.

*Theo năm:*

- **Convenience** và **Extreme** tiếp tục có mức tương quan cao (0.91), cho thấy sự đồng bộ trong mô hình kinh doanh của hai phân khúc này trong dài hạn.
- **All Season** và **Moderation** duy trì mức tương quan cao (0.93), phản ánh sự phụ thuộc của doanh thu hai phân khúc này theo chu kỳ thời gian.
- **Productivity** có mức tương quan thấp với hầu hết các phân khúc khác, cho thấy chiến lược tăng trưởng khác biệt.

- *Tương quan giữa Segment theo Doanh số*

*Theo tháng:*

- **Extreme** và **Convenience** có mức tương quan cao nhất (0.95), phản ánh sự đồng bộ trong số lượng bán ra của hai phân khúc này.
- **All Season** có mối liên hệ mạnh với **Moderation** (0.87) và **Youth** (0.87), cho thấy doanh số của các phân khúc này thường dao động cùng nhau.
- **Productivity** tiếp tục có mức tương quan thấp với các phân khúc khác, phản ánh mô hình bán hàng riêng biệt.

*Theo quý:*

- Mối liên kết giữa **Extreme** và **Convenience** vẫn giữ ở mức rất cao (0.96), cho thấy sự phụ thuộc giữa hai phân khúc này trong chiến lược doanh số.
- **All Season** có mức tương quan mạnh với **Moderation** (0.90) và **Youth** (0.89), phản ánh sự ảnh hưởng qua lại giữa doanh số của các phân khúc này.

- **Productivity** vẫn có mức tương quan thấp với các phân khúc khác, cho thấy đặc điểm tăng trưởng doanh số riêng biệt.

*Theo năm:*

- **Extreme** và **Convenience** tiếp tục có mức tương quan cao (0.93), cho thấy chiến lược dài hạn trong quản lý doanh số của hai phân khúc này có sự liên kết mạnh.
- **All Season** và **Moderation** có mức tương quan cao (0.90), phản ánh xu hướng doanh số của hai phân khúc này đồng biến trong dài hạn.
- **Productivity** vẫn có mức tương quan thấp, cho thấy chiến lược bán hàng có sự tách biệt rõ ràng với các phân khúc còn lại.

### Kiểm tra yếu tố mùa vụ

- Trực quan hóa yếu tố mùa vụ: [Hình 60](#).

Các khía cạnh có thể xem xét, chẳng hạn:

- *Phân tích xu hướng (Trend)*

*Các phân khúc có xu hướng tăng trưởng mạnh trước 2017:*

- **All Season, Moderation, Youth** có xu hướng doanh thu tăng mạnh từ 2012 đến 2017, sau đó chững lại hoặc giảm nhẹ.
- **Select** có đỉnh tăng trưởng vào năm 2016, nhưng sau đó giảm mạnh.
- **Convenience** và **Extreme** cũng đạt đỉnh vào khoảng năm 2017, sau đó suy giảm đáng kể.

*Các phân khúc có xu hướng suy giảm liên tục:*

- **Productivity** và **Regular** có xu hướng giảm dần từ năm 2012 đến 2020, cho thấy doanh thu không có dấu hiệu phục hồi đáng kể.
- **Convenience** cũng có dấu hiệu suy giảm dài hạn, đặc biệt sau năm 2017.

- *Thành phần thời vụ (Seasonality)*

**Tính thời vụ được duy trì ở tất cả phân khúc:**

- Tất cả phân khúc đều có sự biến động theo mùa, phản ánh tác động của các yếu tố thời vụ như xu hướng mua sắm theo thời gian.
- **All Season, Moderation, Extreme**, và **Youth** có biên độ thời vụ lớn, cho thấy ảnh hưởng rõ ràng của các mùa cao điểm và giảm giá.
- **Productivity** và **Regular** có biên độ thời vụ thấp hơn, thể hiện ít chịu ảnh hưởng từ yếu tố mùa vụ.

- *Phần dư (Residual)*

**Giai đoạn tăng trưởng mạnh có phần dư cao:**

- **Select, Youth**, và **All Season** có phần dư dao động mạnh từ năm 2014 đến 2017, cho thấy mức tăng trưởng cao nhưng không ổn định.

- Sau 2017, phần dư có xu hướng giảm dần, phản ánh sự chững lại của doanh thu.

### Giai đoạn suy giảm có phần dư biến động lớn:

- **Convenience** và **Productivity** có phần dư dao động thất thường sau 2017, cho thấy sự không chắc chắn trong doanh thu.
- **Regular** và **Moderation** có mức biến động phần dư tương đối thấp, cho thấy mô hình kinh doanh ổn định hơn.

### Hiệu suất của từng Segment

Bảng 16: Tổng hợp cửa hàng, vùng có doanh thu và lợi nhuận cao nhất cũng như thời điểm bán chạy của các phân khúc

Segment	Cửa hàng		Vùng		Tháng	Quý	Năm
	Doanh thu	Lợi nhuận	Doanh thu	Lợi nhuận			
Convenience	San Antonio, TX	San Antonio, TX	East	East	6	2014Q2	2014
Productivity	Miami, FL	Miami, FL	West	West	12	2011Q4	2011
Select	Miami, FL	Houston, TX	West	West	12	2014Q2	2014
Moderation	Houston, TX	Las Vegas, NV	East	East	4	2015Q2	2015
Extreme	Miami, FL	San Diego, CA	East	East	4	2014Q2	2014
Youth	New York, NY	New York, NY	East	East	6	2016Q2	2016
Regular	San Francisco, CA	San Francisco, CA	East	East	6	2016Q2	2011
All Season	Miami, FL	Miami, FL	West	West	4	2016Q2	2016

### Kiểm định

Bảng 17: Kiểm định ảnh hưởng của mùa và ngày lễ đến doanh thu và doanh số theo từng phân khúc

Segment	ANOVA - Mùa		T-Test - Ngày lễ	
	Doanh thu	Đơn vị bán	Doanh thu	Đơn vị bán
Convenience	0.000000	$2.17 \times 10^{-2}$	$8.75 \times 10^{-8}$	0.3031
Productivity	$3.82 \times 10^{-240}$	$3.53 \times 10^{-13}$	$3.71 \times 10^{-3}$	$1.50 \times 10^{-5}$
Select	$2.75 \times 10^{-3}$	$5.77 \times 10^{-2}$	0.2886	0.2459
Moderation	$6.54 \times 10^{-42}$	$2.52 \times 10^{-3}$	0.0892	0.1537
Extreme	$6.15 \times 10^{-68}$	$1.36 \times 10^{-2}$	0.0289	0.5414
Youth	$2.30 \times 10^{-2}$	$6.00 \times 10^{-3}$	0.6853	0.9127
Regular	$4.97 \times 10^{-11}$	$2.57 \times 10^{-1}$	0.8232	$3.59 \times 10^{-4}$
All Season	$9.09 \times 10^{-19}$	$7.03 \times 10^{-1}$	$2.16 \times 10^{-4}$	0.1490

Bảng trên cho thấy:

- Mùa có ảnh hưởng đến doanh thu ở tất cả các phân khúc.
- Mùa có ảnh hưởng đến doanh số ở hầu hết các phân khúc, trừ Regular và All Season.
- Ngày lễ có ảnh hưởng đến doanh thu ở các phân khúc Convenience, Productivity, Extreme, và All Season.
- Ngày lễ có ảnh hưởng đến doanh số chỉ ở các phân khúc Productivity và Regular.
- Phân khúc Youth và Select không có bằng chứng rõ ràng về sự ảnh hưởng của ngày lễ đến doanh thu hoặc doanh số.

### 3. Liên hệ giữa Category và Segment

Bảng 18: Bảng phân bố Segment theo Category

Category	Regular	All Season	Convenience	Extreme	Moderation	Productivity	Select	Youth
Mix	0	28683	0	0	0	1770	0	0
Rural	0	0	0	0	0	283348	38940	0
Urban	13732	0	233783	123465	131162	0	0	0
Youth	0	0	0	0	0	0	0	46678

Mỗi phần tử của Segment thuộc đúng một Category (trừ Productivity), do đó có thể xem Category là tập lớn của Segment.

### Tỉ lệ đóng góp của Segment trong mỗi Category

Category	Metric	All Season	Convenience	Extreme	Moderation	Productivity	Regular	Select	Youth
Mix	Revenue (%)	92.78	0.00	0.00	0.00	7.22	0.00	0.00	0.00
	Units (%)	94.15	0.00	0.00	0.00	5.85	0.00	0.00	0.00
Rural	Revenue (%)	0.00	0.00	0.00	0.00	80.79	0.00	19.21	0.00
	Units (%)	0.00	0.00	0.00	0.00	87.89	0.00	12.11	0.00
Urban	Revenue (%)	0.00	40.82	18.09	39.30	0.00	1.79	0.00	0.00
	Units (%)	0.00	46.24	24.56	26.42	0.00	2.78	0.00	0.00
Youth	Revenue (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
	Units (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00

Bảng 19: Tỷ lệ đóng góp của Segment vào Category

Bảng 20: Phương sai của Units (%) và Revenue (%) theo Category

Category	Variance Units (%)	Variance Revenue (%)
Mix	1092.74	1058.73
Rural	945.89	806.56
Urban	313.84	327.36
Youth	1250.00	1250.00

- Trực quan hóa tỉ lệ phần trăm doanh số và doanh thu theo thời gian: [Hình 61](#).
- Trực quan hóa phương sai phần trăm của doanh số và doanh thu theo thời gian [Hình 62](#).

## II. Chi tiết kiến trúc mạng

### Đọc thêm. Tri thức lịch & Đầu vào của mô hình

#### 1. Cyclical Encoding

Mỗi ngày được xem xét là thứ bao nhiêu trong tuần (day of week, dow), từ 0: Thứ Hai đến 6: Chủ Nhật. Ngày thứ  $k \in [0, 6]$  trong tuần được thêm thuộc tính  $dow_{\sin}$  và  $dow_{\cos}$  như sau:

$$\begin{aligned} dow_{\sin} &= \sin\left(\frac{2\pi k}{7}\right) \\ dow_{\cos} &= \cos\left(\frac{2\pi k}{7}\right) \end{aligned}$$

Tương tự, mỗi ngày được xem là thuộc về tháng (month) bao nhiêu, từ 1: Tháng Một đến 12: Tháng Mười Hai. Ngày thuộc về tháng  $k \in [1, 12]$  được thêm thuộc tính  $month_{\sin}$  và  $month_{\cos}$  như sau:

$$\begin{aligned} month_{\sin} &= \sin\left(\frac{2\pi k}{12}\right) \\ month_{\cos} &= \cos\left(\frac{2\pi k}{12}\right) \end{aligned}$$

#### 2. Cờ nhị phân

Các cờ `is_weekend` (có phải ngày cuối tuần), `is_holiday` (có phải kỳ nghỉ lễ tại Mỹ), `is_month_start` (có phải ngày bắt đầu của tháng), `is_month_end` (có phải ngày kết thúc của tháng) được thêm vào, nhận giá trị nhị phân: 1 nếu có và 0 nếu không.

#### 3. Trung bình và độ lệch chuẩn trượt

Chúng tôi sử dụng *cửa sổ trượt* (rolling window) để tính *trung bình trượt* (rolling mean) và *độ lệch chuẩn trượt* (rolling standard deviation) cho từng ngày, với kích thước cửa sổ là 7 ngày và 30 ngày.

$$\begin{aligned} \text{rolling}_{\text{mean}}(t) &= \frac{1}{k} \sum_{t-k+1}^t X_i \\ \text{rolling}_{\text{std}}(t) &= \sqrt{\frac{1}{k} \sum_{t-k+1}^t (X_i - \bar{X})^2} \end{aligned}$$

trong đó  $t$  là thời điểm,  $X_i$  là các giá trị `Units` hoặc `Revenue` của các ngày trong cửa sổ trượt và  $k \in \{7, 30\}$ .

#### a. Hybrid Temporal Fusion Transformer

##### 1. Positional Encoding

Cho đầu vào là chuỗi các embedding  $\mathbf{x} \in \mathbb{R}^{T \times d_{\text{model}}}$  với  $T$  là số bước thời gian, mô-đun này cộng vào mỗi vector vị trí một giá trị phụ thuộc vào vị trí. Cụ thể, với vị trí  $\text{pos} \in \{0, 1, \dots, T-1\}$

và chỉ số kênh  $i \in \{0, 1, \dots, d_{\text{model}} - 1\}$ , ta định nghĩa

$$\text{PE}(\text{pos}, i) = \begin{cases} \sin\left(\text{pos} \cdot 10000^{-\frac{2i}{d_{\text{model}}}}\right), & \text{nếu } i \text{ chẵn,} \\ \cos\left(\text{pos} \cdot 10000^{-\frac{2(i-1)}{d_{\text{model}}}}\right), & \text{nếu } i \text{ lẻ.} \end{cases}$$

Giá trị  $\text{PE}(\text{pos}, i)$  được cộng vào vector embedding ban đầu để đưa thông tin vị trí vào mô hình.

## 2. Multi-Scale Feature Extraction

Cho chuỗi đầu vào  $\mathbf{x}_{\text{seq}} \in \mathbb{R}^{T \times F_{\text{in}}}$ , mô-đun này thực hiện các phép biến đổi cục bộ trên các “nhánh” với kích thước nhân tử (kernel) khác nhau, ví dụ  $k \in \{3, 5, 7\}$ . Mỗi nhánh tính toán

$$\mathbf{y}^{(k)} = \text{Dropout}\left(\text{BN}(\text{ReLU}(\text{Conv}_k(\mathbf{x}_{\text{seq}})))\right),$$

trong đó  $\text{Conv}_k$  là phép tích chập 1 chiều với kích thước nhân tử  $k$  và padding thích hợp để giữ kích thước chuỗi không đổi. Các đầu ra của các nhánh được nối theo kêt:

$$\mathbf{Y} = \text{Concat}\left(\mathbf{y}^{(3)}, \mathbf{y}^{(5)}, \mathbf{y}^{(7)}\right) \in \mathbb{R}^{T \times \tilde{F}},$$

với  $\tilde{F} = F_{\text{out}} \times 3$ . Sau đó, thực hiện phép gộp toàn cục theo chiều thời gian (global average pooling):

$$\mathbf{z} = \frac{1}{T} \sum_{t=1}^T \mathbf{Y}_t \in \mathbb{R}^{\tilde{F}},$$

và sau đó chuyển sang không gian đặc trưng có kích thước  $d_{\text{model}}$  qua một ánh xạ tuyến tính:

$$\mathbf{m} = \mathbf{W}_{\text{proj}} \mathbf{z} + \mathbf{b}_{\text{proj}} \in \mathbb{R}^{d_{\text{model}}}.$$

## 3. Variable Selection Network

Với đầu vào đặc trưng tĩnh  $\mathbf{x}_{\text{cal}} \in \mathbb{R}^{F_{\text{cal}}}$ , mô-đun này thực hiện hai nhiệm vụ: (i) chiều từng đặc trưng đơn lẻ vào không gian  $d_{\text{model}}$ , và (ii) học trọng số tổng hợp các đặc trưng. Cụ thể, với mỗi đặc trưng thứ  $i$  (với  $i = 1, \dots, F_{\text{cal}}$ ), ta có phép chiếu:

$$\mathbf{p}_i = \mathbf{W}_{\text{proj}}^{(i)} x_{\text{cal},i} + b_{\text{proj}}^{(i)} \in \mathbb{R}^{d_{\text{model}}}.$$

Các vector  $\mathbf{p}_i$  được tập hợp lại thành ma trận

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_{F_{\text{cal}}} \end{bmatrix} \in \mathbb{R}^{F_{\text{cal}} \times d_{\text{model}}}.$$

Song song, một mạng đa tầng (MLP) tính toán các trọng số cho từng đặc trưng dựa trên  $\mathbf{x}_{\text{cal}}$ :

$$\mathbf{w} = \text{Softmax}\left(\mathbf{W}_2 \text{Dropout}\left(\text{ReLU}(\mathbf{W}_1 \mathbf{x}_{\text{cal}} + \mathbf{b}_1)\right) + \mathbf{b}_2\right) \in \mathbb{R}^{F_{\text{cal}}}.$$

Cuối cùng, kết hợp các đặc trưng với trọng số:

$$\mathbf{s} = \sum_{i=1}^{F_{\text{cal}}} w_i \mathbf{p}_i \in \mathbb{R}^{d_{\text{model}}}.$$

#### 4. Temporal Fusion Transformer

Mô hình tổng hợp gồm nhiều thành phần chính như sau:

##### (i) Xử lý chuỗi thời gian:

Đầu tiên, chuỗi  $\mathbf{x}_{\text{seq}} \in \mathbb{R}^{T \times F_{\text{in}}}$  được chiếu qua một phép ánh xạ tuyến tính:

$$\tilde{\mathbf{x}} = \mathbf{W}_{\text{in}} \mathbf{x}_{\text{seq}} + \mathbf{b}_{\text{in}} \in \mathbb{R}^{T \times d_{\text{model}}}.$$

Sau đó, áp dụng Positional Encoding để đưa thông tin vị trí vào:

$$\tilde{\mathbf{x}} \leftarrow \tilde{\mathbf{x}} + \text{PE},$$

và xử lý qua Transformer Encoder, thu được đầu ra  $\mathbf{T} \in \mathbb{R}^{T \times d_{\text{model}}}$ . Lấy vector ở bước thời gian cuối cùng:

$$\mathbf{t} = \mathbf{T}_T \in \mathbb{R}^{d_{\text{model}}}.$$

##### (ii) Trích xuất đặc trưng cục bộ:

Như đã mô tả ở mục 2, ta có đặc trưng đa tỉ lệ:

$$\mathbf{m} \in \mathbb{R}^{d_{\text{model}}}.$$

Kết hợp các đặc trưng từ Transformer và Multi-Scale:

$$\mathbf{c} = \mathbf{t} + \mathbf{m} \in \mathbb{R}^{d_{\text{model}}}.$$

##### (iii) Kết hợp thông tin tĩnh:

Từ mô-đun Variable Selection Network (mục 3), ta thu được đặc trưng tĩnh  $\mathbf{s} \in \mathbb{R}^{d_{\text{model}}}$ . Sau đó, thực hiện phép nối:

$$\mathbf{u} = \text{Concat}(\mathbf{c}, \mathbf{s}) \in \mathbb{R}^{2d_{\text{model}}},$$

và tính hàm cửa (gating) qua một ánh xạ sigmoid:

$$\mathbf{g} = \sigma(\mathbf{W}_{\text{gate}} \mathbf{u} + \mathbf{b}_{\text{gate}}) \in \mathbb{R}^{d_{\text{model}}}.$$

Sau đó, thông tin được tổng hợp theo cơ chế gating:

$$\mathbf{f} = \mathbf{g} \odot \mathbf{c} + (1 - \mathbf{g}) \odot \mathbf{s} \in \mathbb{R}^{d_{\text{model}}},$$

trong đó  $\odot$  là phép nhân từng phần.

##### (iv) Fusion và kết nối dữ:

Thực hiện một ánh xạ phi tuyến (thông qua một mạng gồm lớp tuyến tính, ReLU và dropout) và thêm kết nối dữ:

$$\mathbf{f}_{\text{fusion}} = f(\mathbf{f}) + \mathbf{f} \in \mathbb{R}^{d_{\text{model}}}.$$

##### (v) Thành phần tự hồi quy:

Từ chuỗi đầu vào, lấy vector cuối cùng  $\mathbf{x}_{\text{last}} \in \mathbb{R}^{F_{\text{in}}}$  và ánh xạ sang không gian  $d_{\text{model}}$ :

$$\mathbf{a} = \mathbf{W}_{\text{AR}} \mathbf{x}_{\text{last}} + \mathbf{b}_{\text{AR}} \in \mathbb{R}^{d_{\text{model}}}.$$

### (vi) Kết hợp cuối cùng:

Nối kết quả từ fusion và autoregressive:

$$\mathbf{z} = \text{Concat}(\mathbf{f}_{\text{fusion}}, \mathbf{a}) \in \mathbb{R}^{2d_{\text{model}}},$$

sau đó thực hiện ánh xạ tuyến tính và phi tuyến để dự báo đầu ra:

$$\mathbf{y} = f_{\text{final}}(\mathbf{z}) \in \mathbb{R}^{F_{\text{out}}},$$

trong đó  $f_{\text{final}}$  là tổ hợp của các lớp tuyến tính, ReLU và dropout.

Như vậy, mô hình **Hybrid Temporal Fusion Transformer** tổng hợp thông tin thời gian qua Transformer và trích xuất đặc trưng cục bộ đa tỉ lệ, đồng thời kết hợp thông tin tĩnh qua mạng lựa chọn đặc trưng, cuối cùng kết hợp thêm thành phần tự hồi quy để dự báo đầu ra.

## b. Hybrid Temporal Convolutional Network

### 1. TCN Block và TCN Time Series Branch

Trong TCN (Temporal Convolutional Network), mỗi khối (block) thực hiện một phép tích chập nhân quả (causal convolution) với cơ chế kết nối dư (residual connection). Cụ thể, với đầu vào là  $\mathbf{x} \in \mathbb{R}^{C_{\text{in}} \times T}$ , khối TCN tính

$$\mathbf{y} = \text{Dropout}\left(\text{ReLU}(\text{Conv}(\mathbf{x}))\right),$$

trong đó phép tích chập sử dụng kernel kích thước  $k$  với hệ số giãn nở (dilation)  $\delta$  và padding được tính sao cho kích thước theo chiều thời gian không thay đổi. Nếu số kênh vào  $C_{\text{in}}$  khác với số kênh ra  $C_{\text{out}}$ , một phép chiếu tuyến tính được thực hiện:

$$\mathbf{r} = \begin{cases} \mathbf{x}, & \text{nếu } C_{\text{in}} = C_{\text{out}}, \\ \mathbf{W}_{\text{down}} \mathbf{x} + \mathbf{b}_{\text{down}}, & \text{nếu } C_{\text{in}} \neq C_{\text{out}}. \end{cases}$$

Sau đó, kết quả đầu ra của khối là

$$\mathbf{z} = \text{ReLU}(\mathbf{y}_{:T} + \mathbf{r}),$$

với chỉ số :  $T$  đảm bảo kích thước theo chiều thời gian bằng đầu vào. Chuỗi các khối TCN được xếp chồng trong **TCN Time Series Branch** để trích xuất đặc trưng từ chuỗi dữ liệu  $\mathbf{x}_{\text{seq}} \in \mathbb{R}^{T \times F_{\text{in}}}$ . Sau khi chuyển đổi dạng  $\mathbf{x}_{\text{seq}}$  thành ma trận  $\mathbb{R}^{F_{\text{in}} \times T}$ , các khối TCN được áp dụng tuần tự, và đầu ra của time step cuối cùng được chọn làm đặc trưng của nhánh:

$$\mathbf{h}_{\text{TCN}} \in \mathbb{R}^{C_{\text{last}}},$$

với  $C_{\text{last}}$  là số kênh của khối cuối cùng, và sau đó được chiếu sang không gian  $\mathbb{R}^{d_{\text{model}}}$  qua ánh xạ tuyến tính.

### 2. ODE-Transformer Branch

Nhánh ODE-Transformer kết hợp hai thành phần chính: Transformer Encoder và một khối giải tích ODE (ODEBlock). Đầu tiên, chuỗi đầu vào  $\mathbf{x}_{\text{seq}} \in \mathbb{R}^{T \times F_{\text{in}}}$  được chiếu sang không gian  $d_{\text{model}}$ :

$$\tilde{\mathbf{x}} = \mathbf{W}_{\text{in}} \mathbf{x}_{\text{seq}} + \mathbf{b}_{\text{in}} \in \mathbb{R}^{T \times d_{\text{model}}}.$$

Sau đó, Transformer Encoder xử lý chuỗi này để thu được đặc trưng theo thời gian:

$$\mathbf{T} = \text{TransformerEncoder}(\tilde{\mathbf{x}}) \in \mathbb{R}^{T \times d_{\text{model}}},$$

và vector đặc trưng tại bước thời gian cuối cùng được chọn:

$$\mathbf{t} = \mathbf{T}_T \in \mathbb{R}^{d_{\text{model}}}.$$

Đồng thời, để khởi tạo quá trình tích phân ODE, trung bình các vector embedding của chuỗi được tính:

$$\mathbf{x}_0 = \frac{1}{T} \sum_{t=1}^T \tilde{\mathbf{x}}_t \in \mathbb{R}^{d_{\text{model}}}.$$

Khối ODE dựa trên một hàm số ODEFunc biểu diễn bởi mạng MLP, định nghĩa động học của trạng thái  $\mathbf{h}(t)$  theo phương trình vi phân:

$$\frac{d\mathbf{h}(t)}{dt} = f(\mathbf{h}(t); \theta),$$

với  $f$  gồm các lớp tuyến tính và phi tuyến (ReLU). Qua tích phân trên khoảng thời gian từ  $t_0$  đến  $t_1$ , đầu ra của khối ODE là

$$\mathbf{o} = \text{ODEBlock}(\mathbf{x}_0) \in \mathbb{R}^{d_{\text{model}}}.$$

Cuối cùng, nhánh này cho ra cặp đặc trưng:

$$(\mathbf{t}, \mathbf{o}) \in \mathbb{R}^{d_{\text{model}}} \times \mathbb{R}^{d_{\text{model}}}.$$

### 3. Variable Selection Network

Tương tự như mô tả ở phần trước, với đầu vào đặc trưng tĩnh  $\mathbf{x}_{\text{cal}} \in \mathbb{R}^{F_{\text{cal}}}$ , mỗi đặc trưng  $x_{\text{cal},i}$  được chiếu riêng qua ánh xạ tuyến tính:

$$\mathbf{p}_i = \mathbf{W}_{\text{proj}}^{(i)} x_{\text{cal},i} + b_{\text{proj}}^{(i)} \in \mathbb{R}^{d_{\text{model}}},$$

với  $i = 1, \dots, F_{\text{cal}}$ . Các vector  $\mathbf{p}_i$  được ghép lại thành ma trận

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_{F_{\text{cal}}} \end{bmatrix}.$$

Một MLP được sử dụng để tính trọng số cho từng đặc trưng dựa trên  $\mathbf{x}_{\text{cal}}$ :

$$\mathbf{w} = \text{Softmax}\left(\mathbf{W}_2 \text{Dropout}\left(\text{ReLU}(\mathbf{W}_1 \mathbf{x}_{\text{cal}} + \mathbf{b}_1)\right) + \mathbf{b}_2\right) \in \mathbb{R}^{F_{\text{cal}}}.$$

Cuối cùng, các đặc trưng được kết hợp theo trọng số:

$$\mathbf{s} = \sum_{i=1}^{F_{\text{cal}}} w_i \mathbf{p}_i \in \mathbb{R}^{d_{\text{model}}}.$$

#### 4. Temporal Convolutional Network Tổng Hợp

Mô hình tổng hợp **Temporal Convolutional Network** kết hợp ba thành phần chính: nhánh TCN, nhánh ODE-Transformer và nhánh đặc trưng tĩnh, cùng với cơ chế kết hợp gating và kết nối dư. Cụ thể:

##### (i) Nhánh thời gian:

- **TCN branch:** Đầu ra từ TCN branch là đặc trưng  $\mathbf{h}_{\text{TCN}} \in \mathbb{R}^{d_{\text{model}}}$  sau khi chiếu qua ánh xạ tuyến tính.

- **ODE-Transformer branch:** Nhánh này cho ra cặp đặc trưng  $\mathbf{t}$  và  $\mathbf{o}$ , được cộng lại thành

$$\mathbf{h}_{\text{ODE}} = \mathbf{t} + \mathbf{o} \in \mathbb{R}^{d_{\text{model}}}.$$

Hai đặc trưng thời gian được nối với nhau:

$$\mathbf{c} = \text{Concat}(\mathbf{h}_{\text{TCN}}, \mathbf{h}_{\text{ODE}}) \in \mathbb{R}^{2d_{\text{model}}}.$$

Sau đó, một hàm cửa (gating) được áp dụng thông qua ánh xạ sigmoid:

$$\mathbf{g}_{\text{time}} = \sigma(\mathbf{W}_{\text{gate},\text{time}} \mathbf{c} + \mathbf{b}_{\text{gate},\text{time}}) \in \mathbb{R}^{d_{\text{model}}},$$

và đặc trưng thời gian được tổng hợp:

$$\mathbf{f}_{\text{time}} = \mathbf{g}_{\text{time}} \odot (\mathbf{h}_{\text{TCN}} + \mathbf{h}_{\text{ODE}}).$$

##### (ii) Nhánh tĩnh:

Đặc trưng tĩnh  $\mathbf{s} \in \mathbb{R}^{d_{\text{model}}}$  từ Variable Selection Network được nối với  $\mathbf{f}_{\text{time}}$ :

$$\mathbf{u} = \text{Concat}(\mathbf{f}_{\text{time}}, \mathbf{s}) \in \mathbb{R}^{2d_{\text{model}}},$$

và áp dụng hàm cửa:

$$\mathbf{g}_{\text{static}} = \sigma(\mathbf{W}_{\text{gate},\text{static}} \mathbf{u} + \mathbf{b}_{\text{gate},\text{static}}) \in \mathbb{R}^{d_{\text{model}}}.$$

Từ đó, thông tin được kết hợp:

$$\mathbf{f}_{\text{all}} = \mathbf{g}_{\text{static}} \odot \mathbf{f}_{\text{time}} + (1 - \mathbf{g}_{\text{static}}) \odot \mathbf{s}.$$

Sau khi qua một tầng fusion phi tuyến (bao gồm ánh xạ tuyến tính, ReLU và Dropout) với kết nối dư:

$$\mathbf{f}_{\text{fusion}} = f(\mathbf{f}_{\text{all}}) + \mathbf{f}_{\text{all}}.$$

##### (iii) Thành phần tự hồi quy (Autoregressive):

Từ đầu vào chuỗi, vector của time step cuối cùng  $\mathbf{x}_{\text{last}} \in \mathbb{R}^{F_{\text{in}}}$  được chiếu qua ánh xạ tuyến tính:

$$\mathbf{a} = \mathbf{W}_{\text{AR}} \mathbf{x}_{\text{last}} + \mathbf{b}_{\text{AR}} \in \mathbb{R}^{d_{\text{model}}}.$$

##### (iv) Kết hợp cuối cùng:

Đặc trưng tổng hợp và vector autoregressive được nối:

$$\mathbf{z} = \text{Concat}(\mathbf{f}_{\text{fusion}}, \mathbf{a}) \in \mathbb{R}^{2d_{\text{model}}},$$

và qua tầng cuối cùng gồm ánh xạ tuyến tính, ReLU, Dropout và ánh xạ tuyến tính để dự báo đầu ra:

$$\mathbf{y} = f_{\text{final}}(\mathbf{z}) \in \mathbb{R}^{F_{\text{out}}}.$$

Như vậy, mô hình **Temporal Convolutional Network** tổng hợp đặc trưng thời gian từ hai nhánh (TCN và ODE-Transformer) và đặc trưng tĩnh qua cơ chế lựa chọn, đồng thời bổ sung thông tin autoregressive để đưa ra dự báo cuối cùng.

### c. Hybrid Forecasting Model

#### 1. Variable Selection Network

Đối với đặc trưng tĩnh, với đầu vào là  $\mathbf{x}_{\text{cal}} \in \mathbb{R}^{F_{\text{cal}}}$ , mỗi biến  $x_{\text{cal},i}$  được chiếu qua một ánh xạ tuyến tính:

$$\mathbf{p}_i = \mathbf{W}_{\text{proj}}^{(i)} x_{\text{cal},i} + b_{\text{proj}}^{(i)} \in \mathbb{R}^{d_{\text{model}}}, \quad i = 1, \dots, F_{\text{cal}},$$

với các vector đặc trưng được tập hợp thành ma trận

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_{F_{\text{cal}}} \end{bmatrix}.$$

Một MLP được sử dụng để tính trọng số cho các biến:

$$\mathbf{w} = \text{Softmax}\left(\mathbf{W}_2 \text{Dropout}\left(\text{ReLU}(\mathbf{W}_1 \mathbf{x}_{\text{cal}} + \mathbf{b}_1)\right) + \mathbf{b}_2\right) \in \mathbb{R}^{F_{\text{cal}}},$$

với đó, đặc trưng tĩnh được tổng hợp qua:

$$\mathbf{s} = \sum_{i=1}^{F_{\text{cal}}} w_i \mathbf{p}_i \in \mathbb{R}^{d_{\text{model}}}.$$

#### 2. NHITS Block và NHITS Branch

Mỗi block trong mô hình NHITS nhận đầu vào là chuỗi thời gian phẳng  $\mathbf{x} \in \mathbb{R}^{D_{\text{in}}}$ , với  $D_{\text{in}} = \text{window\_size} \times \text{num\_series}$ . Block này được xây dựng qua một chuỗi các lớp tuyến tính và hàm kích hoạt ReLU, biểu diễn hàm số ẩn:

$$\mathbf{h} = f(\mathbf{x}; \theta) \in \mathbb{R}^{d_{\text{hidden}}},$$

sau đó, block chia ra hai đầu ra:

$$\text{backcast : } \hat{\mathbf{x}} = \mathbf{W}_{\text{back}} \mathbf{h} + \mathbf{b}_{\text{back}} \in \mathbb{R}^{D_{\text{in}}},$$

$$\text{forecast : } \hat{\mathbf{y}} = \mathbf{W}_{\text{fore}} \mathbf{h} + \mathbf{b}_{\text{fore}} \in \mathbb{R}^{F_{\text{out}}},$$

trong đó  $F_{\text{out}}$  tương ứng với số biến cần dự báo (ví dụ, 2 giá trị).

Trong NHITS Branch, các block được xếp chồng theo cơ chế hiệu chỉnh residual. Cho đầu vào ban đầu  $\mathbf{r}^{(0)} = \mathbf{x}_{\text{flat}}$ , với mỗi block  $j$  ( $j = 1, \dots, n_{\text{blocks}}$ ) ta tính:

$$\begin{aligned} \hat{\mathbf{x}}^{(j)}, \hat{\mathbf{y}}^{(j)} &= \text{NHITSBlock}_j(\mathbf{r}^{(j-1)}), \\ \mathbf{r}^{(j)} &= \mathbf{r}^{(j-1)} - \hat{\mathbf{x}}^{(j)}, \end{aligned}$$

và tổng các dự báo được cộng lại:

$$\hat{\mathbf{Y}} = \sum_{j=1}^{n_{\text{blocks}}} \hat{\mathbf{y}}^{(j)} \in \mathbb{R}^{F_{\text{out}}}.$$

Ở đây,  $\mathbf{x}_{\text{flat}} \in \mathbb{R}^{\text{window\_size} \times \text{num\_series}}$  được vector hoá trước khi đưa vào block.

### 3. DeepAR Branch

Mô hình DeepAR sử dụng kiến trúc LSTM để trích xuất đặc trưng thời gian từ chuỗi đầu vào  $\mathbf{x}_{\text{seq}} \in \mathbb{R}^{T \times F_{\text{in}}}$  (với  $F_{\text{in}} = \text{num\_series}$ ). Qua LSTM, ta có:

$$\{\mathbf{h}_t\}_{t=1}^T = \text{LSTM}(\mathbf{x}_{\text{seq}}),$$

và lấy hidden state của bước thời gian cuối cùng:

$$\mathbf{h}_{\text{DeepAR}} = \mathbf{h}_T \in \mathbb{R}^{d_{\text{model}}}.$$

### 4. Inverted Transformer Branch

Nhánh Inverted Transformer xử lý chuỗi đầu vào  $\mathbf{x}_{\text{seq}} \in \mathbb{R}^{T \times F_{\text{in}}}$  (với  $F_{\text{in}} = \text{num\_series}$ ) thông qua một phép chiếu ban đầu:

$$\tilde{\mathbf{x}} = \mathbf{W}_{\text{in}} \mathbf{x}_{\text{seq}} + \mathbf{b}_{\text{in}} \in \mathbb{R}^{T \times d_{\text{model}}}.$$

Sau đó, chuỗi được truyền qua  $L$  lớp Inverted Transformer Encoder, mỗi lớp bao gồm:

$$\begin{aligned} \mathbf{x}' &= \mathbf{W}_1 \tilde{\mathbf{x}} \in \mathbb{R}^{T \times (d_{\text{model}} \times 4)}, \\ \mathbf{a} &= \text{MultiHeadAttn}(\mathbf{x}', \mathbf{x}', \mathbf{x}') \in \mathbb{R}^{T \times (d_{\text{model}} \times 4)}, \\ \tilde{\mathbf{x}}_{\text{attn}} &= \mathbf{W}_2 \mathbf{a} \in \mathbb{R}^{T \times d_{\text{model}}}, \end{aligned}$$

kết hợp với kết nối dư và chuẩn hoá, cuối cùng chọn vector của time step cuối cùng:

$$\mathbf{h}_{\text{Trans}} = \tilde{\mathbf{x}}_{\text{attn}, T} \in \mathbb{R}^{d_{\text{model}}}.$$

### 5. Hybrid Forecasting Model

Mô hình tổng hợp kết hợp đầu ra của ba nhánh thời gian và nhánh đặc trưng tĩnh. Cụ thể, ta có:

$$\begin{aligned} \mathbf{y}_{\text{NHITS}} &\in \mathbb{R}^{F_{\text{out}}} \quad (\text{dự báo từ NHITS Branch}), \\ \mathbf{h}_{\text{Trans}} &\in \mathbb{R}^{d_{\text{model}}} \quad (\text{đặc trưng từ Inverted Transformer Branch}), \\ \mathbf{h}_{\text{DeepAR}} &\in \mathbb{R}^{d_{\text{model}}} \quad (\text{đặc trưng từ DeepAR Branch}). \end{aligned}$$

Các đầu ra này được nối lại:

$$\mathbf{c} = \text{Concat}(\mathbf{y}_{\text{NHITS}}, \mathbf{h}_{\text{Trans}}, \mathbf{h}_{\text{DeepAR}}) \in \mathbb{R}^{F_{\text{out}} + 2d_{\text{model}}},$$

với sau đó được chiếu qua một lớp tổng hợp:

$$\mathbf{f}_{\text{time}} = f_{\text{time}}(\mathbf{c}) \in \mathbb{R}^{d_{\text{model}}}.$$

Đồng thời, đặc trưng tĩnh  $\mathbf{s} \in \mathbb{R}^{d_{\text{model}}}$  được tính từ Variable Selection Network. Sau đó, hai đặc trưng được kết hợp qua cơ chế cửa gating:

$$\begin{aligned}\mathbf{u} &= \text{Concat}(\mathbf{f}_{\text{time}}, \mathbf{s}) \in \mathbb{R}^{2d_{\text{model}}}, \\ \mathbf{g} &= \sigma(\mathbf{W}_{\text{gate}} \mathbf{u} + \mathbf{b}_{\text{gate}}) \in \mathbb{R}^{d_{\text{model}}}, \\ \mathbf{f} &= \mathbf{g} \odot \mathbf{f}_{\text{time}} + (1 - \mathbf{g}) \odot \mathbf{s} \in \mathbb{R}^{d_{\text{model}}}.\end{aligned}$$

Một tầng fusion với kết nối dư được áp dụng:

$$\mathbf{f}_{\text{fusion}} = f_{\text{fusion}}(\mathbf{f}) + \mathbf{f} \in \mathbb{R}^{d_{\text{model}}}.$$

Cuối cùng, một thành phần autoregressive sử dụng giá trị cuối của chuỗi ban đầu  $\mathbf{x}_{\text{last}} \in \mathbb{R}^{F_{\text{in}}}$  được chiếu sang không gian  $d_{\text{model}}$ :

$$\mathbf{a} = \mathbf{W}_{\text{AR}} \mathbf{x}_{\text{last}} + \mathbf{b}_{\text{AR}} \in \mathbb{R}^{d_{\text{model}}},$$

và sau đó, hai thành phần được nối lại:

$$\mathbf{z} = \text{Concat}(\mathbf{f}_{\text{fusion}}, \mathbf{a}) \in \mathbb{R}^{2d_{\text{model}}},$$

để qua tầng cuối cùng đưa ra dự báo:

$$\hat{\mathbf{y}} = f_{\text{final}}(\mathbf{z}) \in \mathbb{R}^{F_{\text{out}}}.$$

Như vậy, mô hình **Hybrid Forecasting Model** tổng hợp đầu ra từ các nhánh NHITS, Inverted Transformer và DeepAR, kết hợp thêm thông tin từ đặc trưng tĩnh và thành phần autoregressive để đưa ra dự báo cuối cùng.

## d. Conditional Variational AutoEncoder

### 1. Mô hình VAE điều kiện

Mô hình **Conditional Forecast VAE** kết hợp thông tin chuỗi thời gian và đặc trưng tĩnh để học một biểu diễn tiềm ẩn có điều kiện, từ đó đưa ra dự báo. Đầu vào của encoder là sự nối của vector phẳng của chuỗi thời gian  $\mathbf{x}_{\text{seq}} \in \mathbb{R}^{\text{window\_size} \times \text{num\_series}}$  và đặc trưng tĩnh  $\mathbf{x}_{\text{cal}} \in \mathbb{R}^{\text{static\_dim}}$ , trong khi decoder nhận đầu vào là sự nối của vector tiềm ẩn và đặc trưng tĩnh, đồng thời bổ sung thông tin qua skip connection từ flatten( $\mathbf{x}_{\text{seq}}$ ).

### 2. Encoder

Dầu tiên, chuỗi thời gian được vector hoá thành:

$$\mathbf{x}_{\text{flat}} \in \mathbb{R}^{\text{window\_size} \times \text{num\_series}},$$

với kích thước sau khi vector hoá là  $\text{window\_size} \cdot \text{num\_series}$ . Sau đó, ta nối với đặc trưng tĩnh:

$$\mathbf{e} = \text{Concat}(\text{flatten}(\mathbf{x}_{\text{seq}}), \mathbf{x}_{\text{cal}}) \in \mathbb{R}^{\text{window\_size} \cdot \text{num\_series} + \text{static\_dim}}.$$

Hàm số  $f_{\text{enc}}$  được biểu diễn qua các lớp tuyến tính, ReLU và Dropout:

$$\mathbf{h} = f_{\text{enc}}(\mathbf{e}) \in \mathbb{R}^{\text{hidden\_dim}},$$

từ đó tính ra trung bình và log phương sai của phân phối tiềm ẩn:

$$\boldsymbol{\mu} = \mathbf{W}_\mu \mathbf{h} + \mathbf{b}_\mu \in \mathbb{R}^{\text{latent\_dim}},$$

$$\log \boldsymbol{\sigma}^2 = \mathbf{W}_{\log \text{var}} \mathbf{h} + \mathbf{b}_{\log \text{var}} \in \mathbb{R}^{\text{latent\_dim}}.$$

Kết quả  $\mathbf{x}_{\text{flat}}$  được lưu lại để sử dụng cho skip connection trong decoder.

### 3. Reparameterization

Để có thể thực hiện lan truyền ngược qua mẫu ngẫu nhiên, phương pháp reparameterization được áp dụng:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \odot \exp(0.5 \log \boldsymbol{\sigma}^2), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

trong đó  $\odot$  là phép nhân từng phần.

### 4. Decoder và Skip Connection

Đầu vào của decoder là sự nối của vector tiềm ẩn  $\mathbf{z}$  và đặc trưng tĩnh  $\mathbf{x}_{\text{cal}}$ :

$$\mathbf{d}_{\text{in}} = \text{Concat}(\mathbf{z}, \mathbf{x}_{\text{cal}}) \in \mathbb{R}^{\text{latent\_dim} + \text{static\_dim}},$$

đi qua hàm số  $f_{\text{dec}}$  với các lớp tuyến tính, ReLU và Dropout để thu được đặc trưng ẩn:

$$\mathbf{h}_{\text{dec}} = f_{\text{dec}}(\mathbf{d}_{\text{in}}) \in \mathbb{R}^{\text{hidden\_dim}}.$$

Song song, skip connection được tính qua ánh xạ tuyến tính của flatten( $\mathbf{x}_{\text{seq}}$ ):

$$\mathbf{h}_{\text{skip}} = \mathbf{W}_{\text{skip}} \text{flatten}(\mathbf{x}_{\text{seq}}) + \mathbf{b}_{\text{skip}} \in \mathbb{R}^{\text{hidden\_dim}}.$$

Hai đặc trưng này được nối lại:

$$\mathbf{c} = \text{Concat}(\mathbf{h}_{\text{dec}}, \mathbf{h}_{\text{skip}}) \in \mathbb{R}^{2 \cdot \text{hidden\_dim}},$$

và cuối cùng, qua tầng cuối cùng  $f_{\text{final}}$ , ta có dự báo:

$$\hat{\mathbf{y}} = f_{\text{final}}(\mathbf{c}) \in \mathbb{R}^{\text{output\_dim}},$$

trong đó output\_dim là số lượng biến đầu ra dự báo.

### 5. Hàm mất mát VAE

Hàm mất mát của mô hình VAE điều kiện gồm hai thành phần:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}},$$

trong đó

$$\mathcal{L}_{\text{recon}} = \text{SmoothL1}(\hat{\mathbf{y}}, \mathbf{y}_{\text{true}})$$

là hàm mất mát tái tạo, và

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \mathbb{E} \left[ 1 + \log \boldsymbol{\sigma}^2 - \boldsymbol{\mu}^2 - \exp(\log \boldsymbol{\sigma}^2) \right]$$

là hàm mất mát Kullback-Leibler, được cân bằng bởi hệ số  $\lambda_{\text{KL}}$ .

Như vậy, mô hình **Conditional Forecast VAE** xây dựng encoder để học biểu diễn tiềm ẩn có điều kiện từ dữ liệu chuỗi thời gian và đặc trưng tĩnh, sử dụng cơ chế reparameterization để cho phép học qua gradient, và decoder kết hợp thông tin từ vector tiềm ẩn và đặc trưng tĩnh, bổ sung thông qua skip connection từ dữ liệu chuỗi thời gian ban đầu để đưa ra dự báo.

## e. Distributional Conditional Forecast

### 1. Giới thiệu mô hình

Mô hình **Distributional Conditional Forecast** mở rộng mô hình VAE điều kiện bằng cách cho decoder xuất ra cặp tham số  $(\mu_y, \log \sigma_y^2)$  cho phân phối dự báo, thay vì chỉ một vector cố định. Điều này cho phép mô hình tự điều chỉnh độ bất định (đại diện bởi  $\sigma_y$ ) cho các điểm dữ liệu có dao động mạnh (spike).

### 2. Encoder

Encoder nhận đầu vào là sự nối của vector phẳng của chuỗi thời gian và các đặc trưng tĩnh. Cụ thể, với chuỗi thời gian ban đầu  $\mathbf{x}_{\text{seq}} \in \mathbb{R}^{\text{window\_size} \times \text{num\_series}}$ , ta vector hoá thành

$$\text{flatten}(\mathbf{x}_{\text{seq}}) \in \mathbb{R}^{\text{window\_size} \cdot \text{num\_series}},$$

và nối với các đặc trưng tĩnh  $\mathbf{x}_{\text{cal}} \in \mathbb{R}^{\text{static\_dim}}$ :

$$\mathbf{e} = \text{Concat}\left(\text{flatten}(\mathbf{x}_{\text{seq}}), \mathbf{x}_{\text{cal}}\right) \in \mathbb{R}^{\text{window\_size} \cdot \text{num\_series} + \text{static\_dim}}.$$

Sau đó, hàm số  $f_{\text{enc}}$  được biểu diễn qua các lớp tuyến tính, ReLU và Dropout:

$$\mathbf{h}_{\text{enc}} = f_{\text{enc}}(\mathbf{e}) \in \mathbb{R}^{\text{hidden\_dim}},$$

từ đó ta tính được trung bình và log phương sai của phân phối tiềm ẩn:

$$\begin{aligned} \boldsymbol{\mu}_z &= \mathbf{W}_{\mu_z} \mathbf{h}_{\text{enc}} + \mathbf{b}_{\mu_z} \in \mathbb{R}^{\text{latent\_dim}}, \\ \log \boldsymbol{\sigma}_z^2 &= \mathbf{W}_{\log \text{var}_z} \mathbf{h}_{\text{enc}} + \mathbf{b}_{\log \text{var}_z} \in \mathbb{R}^{\text{latent\_dim}}. \end{aligned}$$

Đồng thời, vector phẳng  $\text{flatten}(\mathbf{x}_{\text{seq}})$  được lưu lại để dùng cho skip connection trong decoder.

### 3. Reparameterization

Để có thể lan truyền gradient qua mẫu ngẫu nhiên, phương pháp reparameterization được áp dụng:

$$\mathbf{z} = \boldsymbol{\mu}_z + \boldsymbol{\epsilon} \odot \exp\left(0.5 \log \boldsymbol{\sigma}_z^2\right), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

trong đó  $\odot$  biểu diễn phép nhân từng phần.

### 4. Decoder và Skip Connection

Decoder nhận đầu vào là sự nối của vector tiềm ẩn  $\mathbf{z}$  và các đặc trưng tĩnh:

$$\mathbf{d}_{\text{in}} = \text{Concat}\left(\mathbf{z}, \mathbf{x}_{\text{cal}}\right) \in \mathbb{R}^{\text{latent\_dim} + \text{static\_dim}}.$$

Qua hàm số  $f_{\text{dec}}$  (được biểu diễn qua các lớp tuyến tính, ReLU và Dropout), ta thu được đặc trưng ẩn:

$$\mathbf{h}_{\text{dec}} = f_{\text{dec}}(\mathbf{d}_{\text{in}}) \in \mathbb{R}^{\text{hidden\_dim}}.$$

Song song, thông tin từ chuỗi thời gian ban đầu được giữ lại thông qua skip connection. Cụ thể, vector phẳng  $\text{flatten}(\mathbf{x}_{\text{seq}})$  được chiếu qua một ánh xạ tuyến tính:

$$\mathbf{h}_{\text{skip}} = \mathbf{W}_{\text{skip}} \text{flatten}(\mathbf{x}_{\text{seq}}) + \mathbf{b}_{\text{skip}} \in \mathbb{R}^{\text{hidden\_dim}}.$$

Hai đặc trưng  $\mathbf{h}_{\text{dec}}$  và  $\mathbf{h}_{\text{skip}}$  được nối lại:

$$\mathbf{c} = \text{Concat}(\mathbf{h}_{\text{dec}}, \mathbf{h}_{\text{skip}}) \in \mathbb{R}^{2 \cdot \text{hidden\_dim}},$$

và sau đó qua tầng cuối cùng, ta thu được đầu ra của decoder:

$$(\mu_y, \log \sigma_y^2) = f_{\text{final}}(\mathbf{c}) \in \mathbb{R}^{\text{output\_dim} \times 2},$$

trong đó đầu ra gồm output\_dim cặp tham số cho phân phối Gaussian dự báo.

### 5. Hàm mất mát

Hàm mất mát của mô hình bao gồm hai thành phần: hàm mất mát tái tạo (negative log-likelihood) và hàm mất mát KL cho latent vector. Cụ thể, giả sử dự báo  $y$  được mô hình hóa theo phân phối Gaussian với tham số  $(\mu_y, \sigma_y)$ , ta có:

$$\text{NLL} = \frac{1}{2} \sum_{j=1}^{\text{output\_dim}} \left[ \log(2\pi) + \log \sigma_{y,j}^2 + \frac{(y_j - \mu_{y,j})^2}{\sigma_{y,j}^2} \right],$$

trong đó tổng được thực hiện trên các biến đầu ra của mỗi mẫu và trung bình qua batch. Hàm mất mát KL cho latent vector được tính theo:

$$\text{KL} = -\frac{1}{2} \mathbb{E} \left[ 1 + \log \sigma_z^2 - \mu_z^2 - \exp(\log \sigma_z^2) \right].$$

Hàm mất mát tổng thể của mô hình là:

$$\mathcal{L} = \text{NLL} + \lambda_{\text{KL}} \text{KL},$$

trong đó  $\lambda_{\text{KL}}$  là hệ số cân bằng giữa hai thành phần mất mát.

Như vậy, mô hình **Distributional Conditional Forecast** mở rộng mô hình VAE điều kiện bằng cách cho phép dự báo không chỉ giá trị trung bình mà còn cả độ bất định của dự báo, qua đó có khả năng tự động điều chỉnh sigma cho các điểm dao động mạnh.

## f. Probabilistic Forecast Transformer

### 1. Positional Encoding

Để đưa thông tin vị trí vào biểu diễn của Transformer, mô-đun Positional Encoding được sử dụng. Với đầu vào là embedding  $\mathbf{x} \in \mathbb{R}^{T \times d_{\text{model}}}$  (với  $T$  là số bước thời gian,  $d_{\text{model}}$  là kích thước embedding), ta tính:

$$\text{PE(pos, } i) = \begin{cases} \sin\left(\text{pos} \cdot 10000^{-\frac{2i}{d_{\text{model}}}}\right), & \text{nếu } i \text{ chẵn,} \\ \cos\left(\text{pos} \cdot 10000^{-\frac{2(i-1)}{d_{\text{model}}}}\right), & \text{nếu } i \text{ lẻ,} \end{cases}$$

sau đó cộng vào  $\mathbf{x}$  và áp dụng Dropout.

### 2. Encoder

Encoder của mô hình nhận đầu vào là sự nối của chuỗi thời gian lịch sử và các đặc trưng tĩnh. Cụ thể, với chuỗi thời gian  $\mathbf{x}_{\text{seq}} \in \mathbb{R}^{\text{window\_size} \times \text{num\_series}}$ , ta thực hiện:

$$\text{flatten}(\mathbf{x}_{\text{seq}}) \in \mathbb{R}^{\text{window\_size} \cdot \text{num\_series}},$$

và sau đó nối với các đặc trưng tĩnh  $\mathbf{x}_{\text{cal}} \in \mathbb{R}^{\text{static\_dim}}$  để thu được đầu vào:

$$\mathbf{e} = \text{Concat}\left(\text{flatten}(\mathbf{x}_{\text{seq}}), \mathbf{x}_{\text{cal}}\right) \in \mathbb{R}^{\text{window\_size} \cdot \text{num\_series} + \text{static\_dim}}.$$

Đầu vào này được chiếu qua một lớp tuyến tính lên không gian  $d_{\text{model}}$ :

$$\tilde{\mathbf{e}} = \mathbf{W}_{\text{enc}} \mathbf{e} + \mathbf{b}_{\text{enc}} \in \mathbb{R}^{d_{\text{model}}}.$$

Với mục đích sử dụng Transformer, ta mở rộng chiều thời gian thành  $T = 1$ :

$$\tilde{\mathbf{e}} \rightarrow \tilde{\mathbf{E}} \in \mathbb{R}^{1 \times d_{\text{model}}},$$

sau đó cộng thêm thông tin vị trí thông qua Positional Encoding và truyền qua Transformer Encoder:

$$\mathbf{x}_{\text{trans}} = \text{TransformerEncoder}\left(\text{PosEnc}(\tilde{\mathbf{E}})\right) \in \mathbb{R}^{1 \times d_{\text{model}}}.$$

Kết quả được nén lại thành vector:

$$\mathbf{x}_{\text{enc}} = \text{Squeeze}\left(\mathbf{x}_{\text{trans}}\right) \in \mathbb{R}^{d_{\text{model}}}.$$

Sau đó, một tầng FC kết hợp ReLU và Dropout tạo biểu diễn ẩn:

$$\mathbf{h}_{\text{enc}} = f_{\text{enc}}(\mathbf{x}_{\text{enc}}) \in \mathbb{R}^{\text{hidden\_dim}},$$

từ đó ta tính trung bình và log phương sai của phân phối tiềm ẩn:

$$\begin{aligned} \boldsymbol{\mu}_z &= \mathbf{W}_{\mu_z} \mathbf{h}_{\text{enc}} + \mathbf{b}_{\mu_z} \in \mathbb{R}^{\text{latent\_dim}}, \\ \log \boldsymbol{\sigma}_z^2 &= \mathbf{W}_{\log \text{var}_z} \mathbf{h}_{\text{enc}} + \mathbf{b}_{\log \text{var}_z} \in \mathbb{R}^{\text{latent\_dim}}. \end{aligned}$$

Đồng thời, vector flatten( $\mathbf{x}_{\text{seq}}$ ) được lưu lại làm thông tin cho skip connection trong decoder.

### 3. Reparameterization

Để mẫu ngẫu nhiên có thể lan truyền gradient, phương pháp reparameterization được áp dụng:

$$\mathbf{z} = \boldsymbol{\mu}_z + \boldsymbol{\epsilon} \odot \exp\left(0.5 \log \boldsymbol{\sigma}_z^2\right), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

### 4. Decoder và Skip Connection

Decoder nhận đầu vào là sự nối của vector tiềm ẩn  $\mathbf{z} \in \mathbb{R}^{\text{latent\_dim}}$  và đặc trưng tĩnh  $\mathbf{x}_{\text{cal}} \in \mathbb{R}^{\text{static\_dim}}$ :

$$\mathbf{d}_{\text{in}} = \text{Concat}\left(\mathbf{z}, \mathbf{x}_{\text{cal}}\right) \in \mathbb{R}^{\text{latent\_dim} + \text{static\_dim}}.$$

Qua các tầng FC với ReLU và Dropout, ta thu được biểu diễn ẩn của decoder:

$$\mathbf{h}_{\text{dec}} = f_{\text{dec}}(\mathbf{d}_{\text{in}}) \in \mathbb{R}^{\text{hidden\_dim}}.$$

Bên cạnh đó, thông tin chi tiết từ chuỗi thời gian ban đầu được giữ lại qua skip connection, bằng cách chiếu vector phẳng

$$\text{flatten}(\mathbf{x}_{\text{seq}}) \in \mathbb{R}^{\text{window\_size} \cdot \text{num\_series}}$$

vào không gian ẩn:

$$\mathbf{h}_{\text{skip}} = \mathbf{W}_{\text{skip}} \text{flatten}(\mathbf{x}_{\text{seq}}) + \mathbf{b}_{\text{skip}} \in \mathbb{R}^{\text{hidden\_dim}}.$$

Sau đó, hai đặc trưng được nối lại:

$$\mathbf{c} = \text{Concat}(\mathbf{h}_{\text{dec}}, \mathbf{h}_{\text{skip}}) \in \mathbb{R}^{2 \cdot \text{hidden\_dim}},$$

và qua tầng cuối cùng ta thu được đầu ra của decoder, đó là cặp tham số  $(\mu_y, \log \sigma_y^2)$  cho mỗi biến dự báo:

$$(\mu_y, \log \sigma_y^2) = f_{\text{final}}(\mathbf{c}) \in \mathbb{R}^{\text{output\_dim} \times 2}.$$

### 5. Hàm mất mát

Hàm mất mát bao gồm hai thành phần: hàm mất mát tái tạo và hàm mất mát KL cho latent vector. Với giả sử  $y$  được mô hình hóa theo phân phối Gaussian với tham số  $(\mu_y, \sigma_y)$ , hàm mất mát tái tạo (negative log-likelihood) được tính:

$$\text{NLL} = \frac{1}{2} \sum_{j=1}^{\text{output\_dim}} \left[ \log(2\pi) + \log \sigma_{y,j}^2 + \frac{(y_j - \mu_{y,j})^2}{\sigma_{y,j}^2} \right],$$

trung bình qua batch. Hàm mất mát KL cho latent vector được tính theo:

$$\text{KL} = -\frac{1}{2} \mathbb{E} \left[ 1 + \log \sigma_z^2 - \mu_z^2 - \exp(\log \sigma_z^2) \right].$$

Hàm mất mát tổng thể là:

$$\mathcal{L} = \text{NLL} + \lambda_{\text{KL}} \text{KL},$$

trong đó  $\lambda_{\text{KL}}$  điều chỉnh tầm quan trọng của thành phần KL.

Như vậy, mô hình **ProbabilisticForecastTransformer** kết hợp các thành phần của VAE (encoder, reparameterization và decoder) với kiến trúc Transformer để xây dựng biểu diễn ẩn có điều kiện từ dữ liệu chuỗi thời gian và đặc trưng tĩnh, từ đó dự báo đầu ra dưới dạng phân phối Gaussian với tham số  $(\mu_y, \log \sigma_y^2)$ .

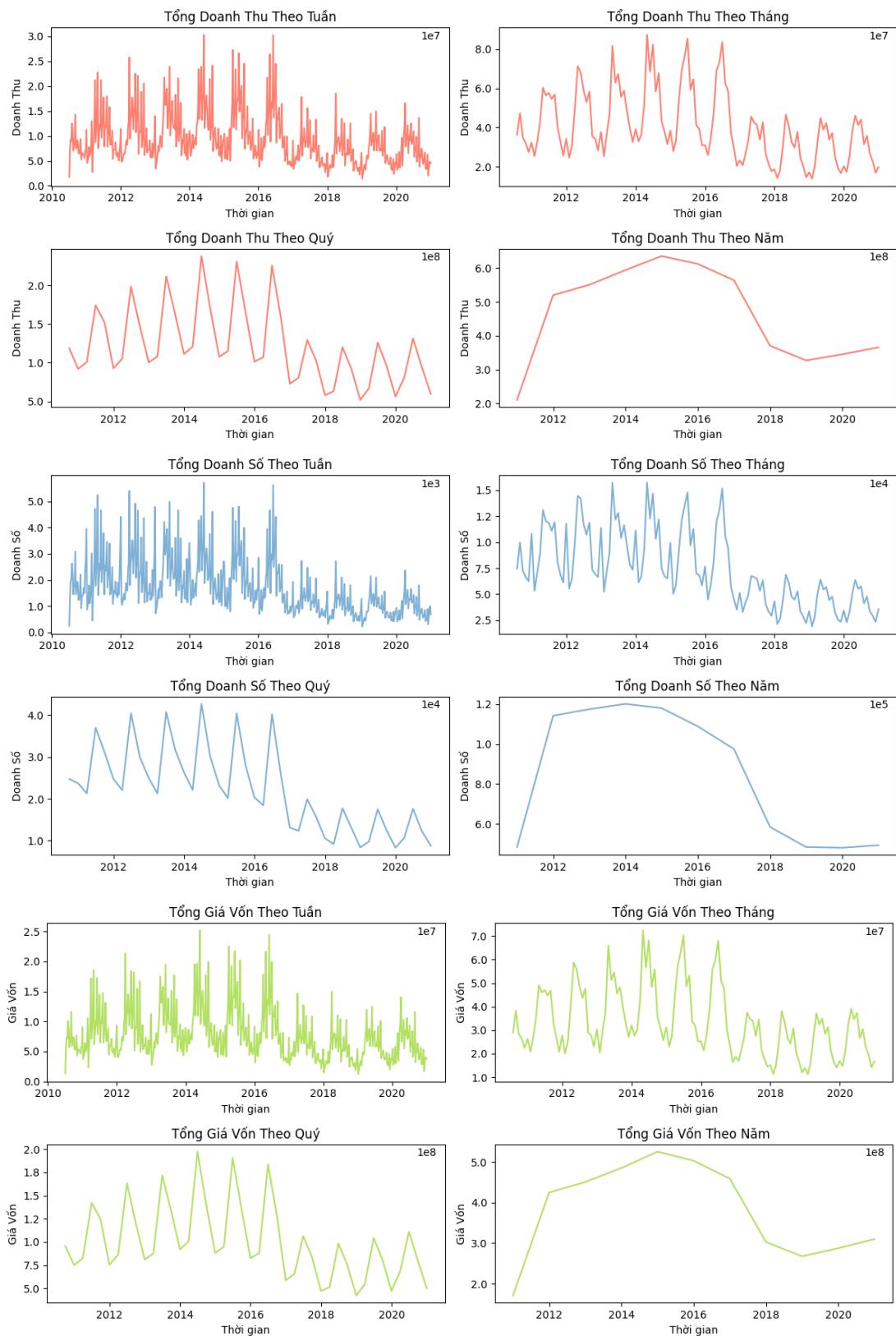
## Tài liệu

- [1] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018. <https://otexts.com/fpp3/>.
- [2] Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.
- [3] Preetum Nakkiran. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, 2019.
- [4] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- [5] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15. Springer, 2000.
- [6] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. CRC Press, 2012.
- [7] A. Khosravi, S. Nahavandi, D. Creighton, and G. Jones. Dynamic weighting methods for the prediction of time series data with an ensemble of neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 24(3):430–444, 2013.
- [8] Ronald A. Fisher. *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford University Press, 1992.
- [9] Student. *The probable error of a mean*, volume 6. Oxford University Press, 1908.
- [10] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.
- [11] Henry B. Mann. Nonparametric tests against trend. *Econometrica: Journal of the Econometric Society*, pages 245–259, 1945.
- [12] Maurice G. Kendall. Rank correlation methods. 1975.
- [13] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.
- [14] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.
- [15] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
- [16] Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.
- [17] Christopher A. Sims. Macroeconomics and reality. *Econometrica*, 48(1):1–48, 1980.
- [18] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 5th edition, 2015.

- [19] Sean J Taylor and Benjamin Letham. Forecasting at scale. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1515–1524. ACM, 2018.
- [20] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.
- [21] Steven L Scott and Hal R Varian. Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2):4–23, 2014.
- [22] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1979.
- [23] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
- [24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [25] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020.
- [26] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Deep convolutional neural networks for facial expression recognition. *2016 IEEE Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9, 2016.
- [27] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation, 2016.
- [28] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations, 2019.
- [29] Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza, Max Mergenthaler-Canseco, and Artur Dubrawski. N-hits: Neural hierarchical interpolation for time series forecasting, 2022.
- [30] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting, 2024.
- [31] Christian Bakke Vennerød, Adrian Kjærnan, and Erling Stray Bugge. Long short-term memory rnn, 2021.
- [32] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [33] Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964.
- [34] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

- [35] Donglai Zhu, Hengshuai Yao, Bei Jiang, and Peng Yu. Negative log likelihood ratio loss for deep neural network classification. *arXiv preprint arXiv:1804.10690*, 2018.
- [36] Remy Lau. Cross-entropy, negative log-likelihood, and all that jazz. *Towards Data Science*, 2022.
- [37] Lester James Miranda. Understanding softmax and the negative log-likelihood. *ljvmiranda921.github.io*, 2017.
- [38] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [39] Victor Prokhorov, Ehsan Shareghi, Yingzhen Li, Mohammad Taher Pilehvar, and Nigel Collier. On the importance of the Kullback-Leibler divergence term in variational autoencoders for text generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 118–127. Association for Computational Linguistics, 2019.

Hình 1: Trực quan hóa tổng doanh thu, doanh số và giá vốn theo thời gian



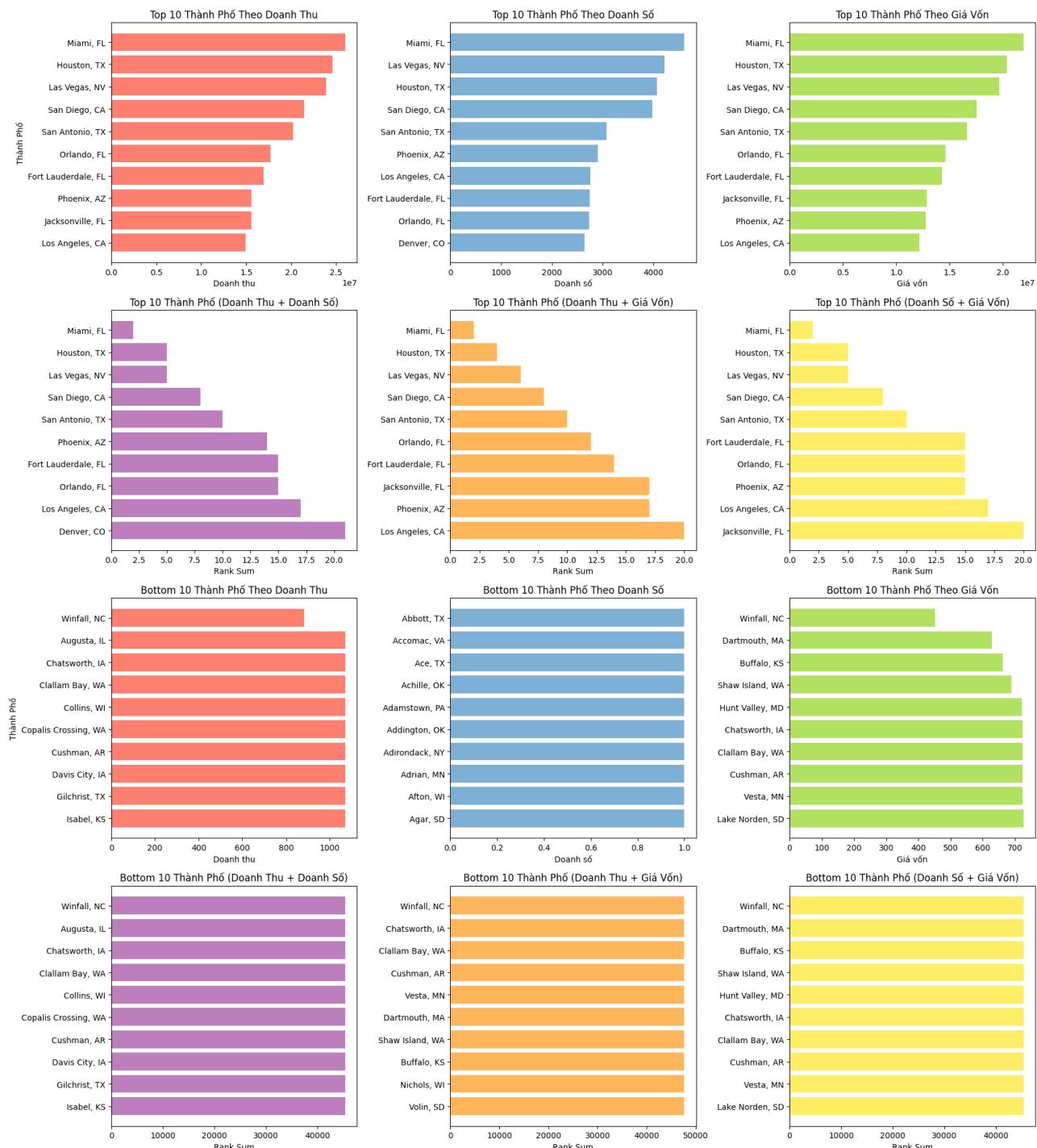
Hình 2: Top 10 và Bottom 10 sản phẩm theo doanh thu, doanh số, giá vốn và các cặp tiêu chí theo phương pháp Rank Sum trên toàn thời gian



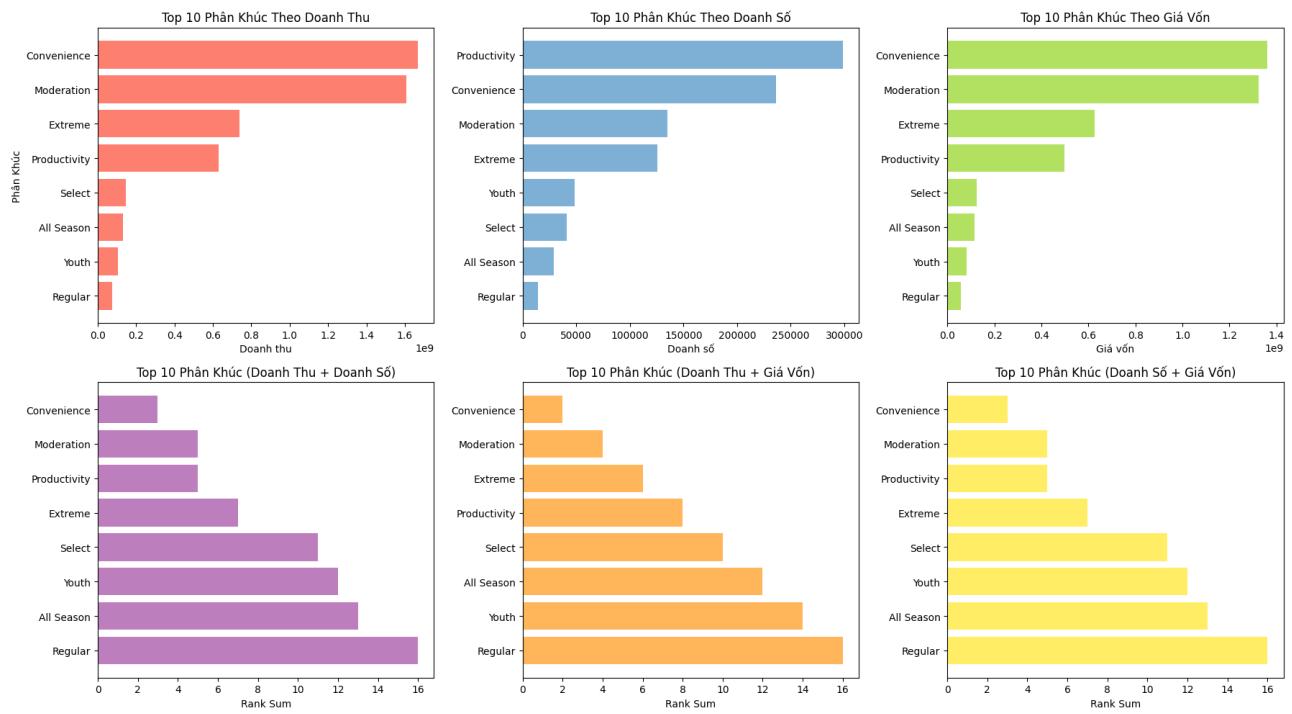
Hình 3: Top 10 và Bottom 10 bang theo doanh thu, doanh số, giá vốn và các cặp tiêu chí theo phương pháp Rank Sum trên toàn thời gian



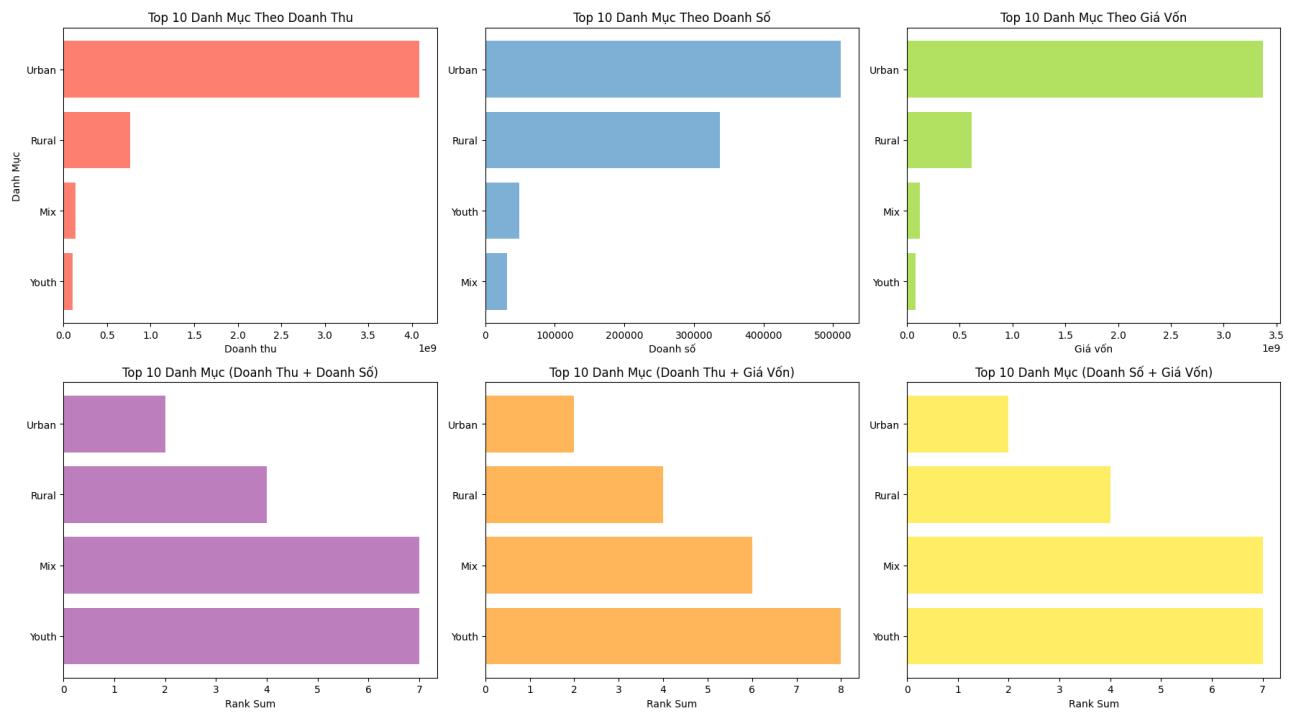
Hình 4: Top 10 và Bottom 10 thành phố theo doanh thu, doanh số, giá vốn và các cặp tiêu chí theo phương pháp Rank Sum trên toàn thời gian



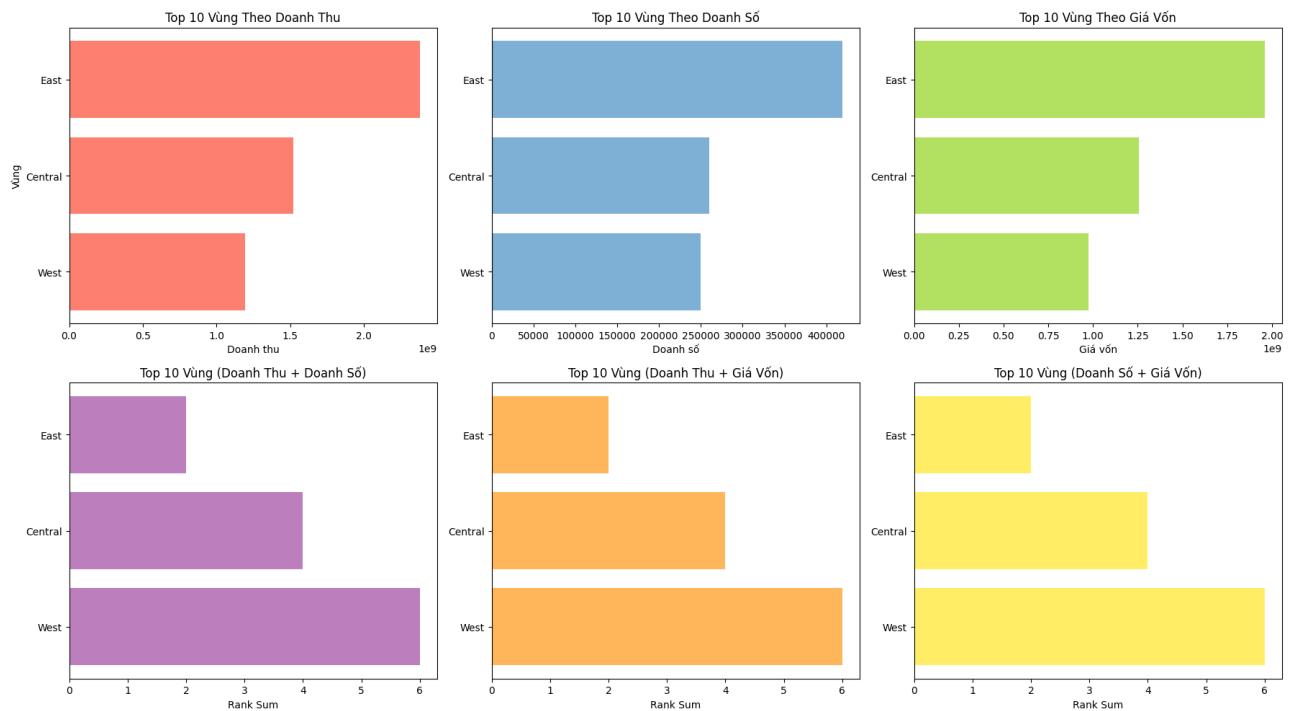
Hình 5: Thứ hạng phân khúc theo doanh thu, doanh số, giá vốn và các cặp tiêu chí theo phương pháp Rank Sum trên toàn thời gian



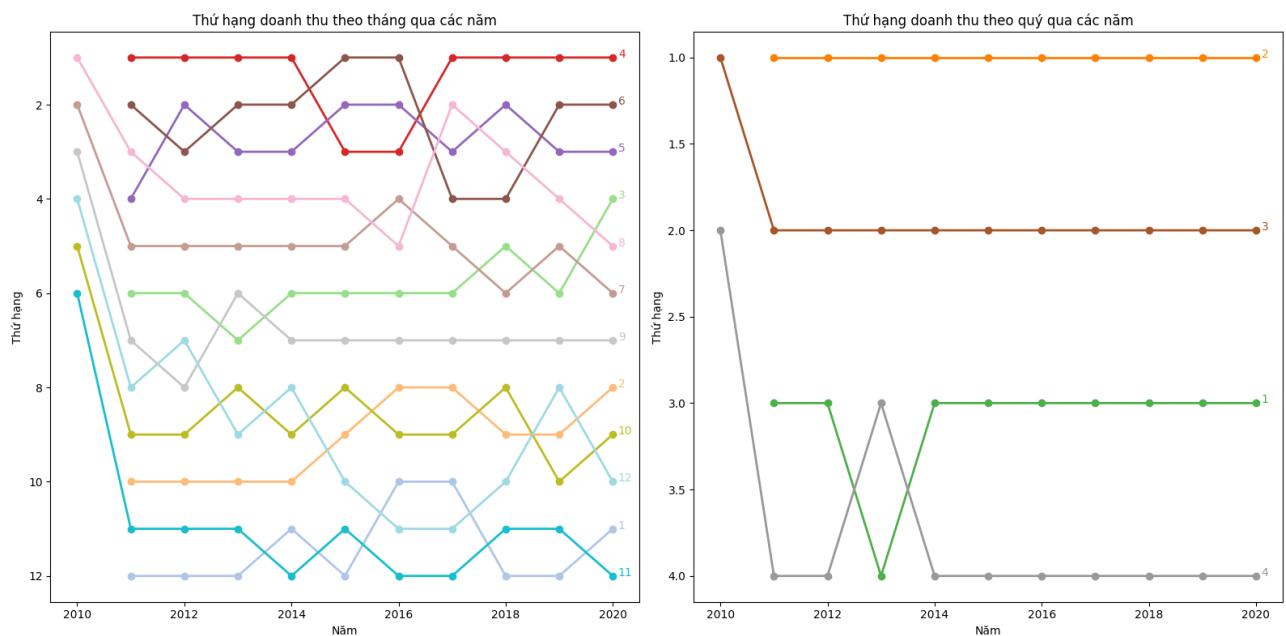
Hình 6: Thứ hạng danh mục theo doanh thu, doanh số, giá vốn và các cặp tiêu chí theo phương pháp Rank Sum trên toàn thời gian



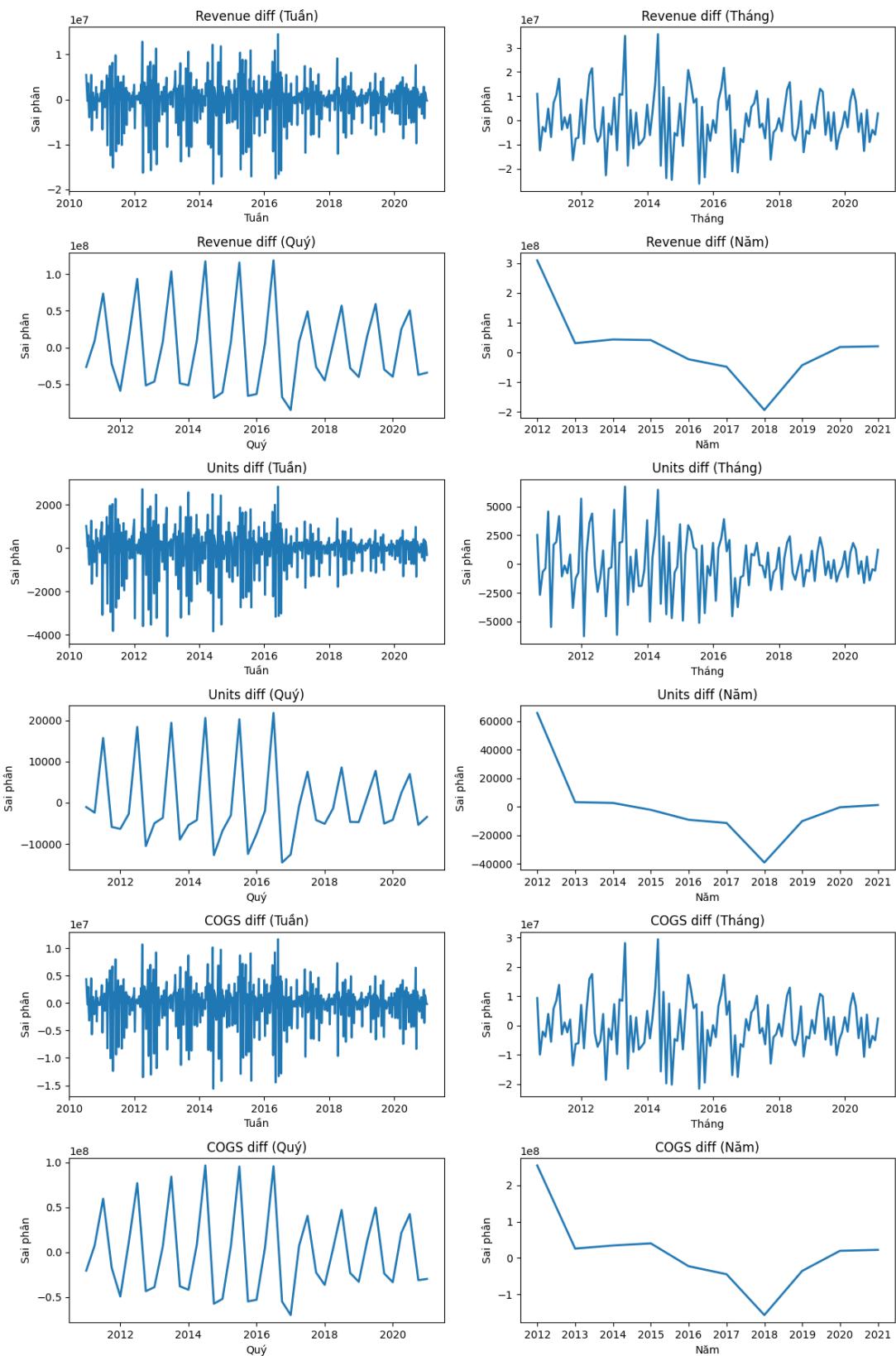
Hình 7: Thứ hạng vùng theo doanh thu, doanh số, giá vốn và các cặp tiêu chí theo phương pháp Rank Sum trên toàn thời gian



Hình 8: Sự thay đổi về thứ hạng doanh thu của các tháng và quý theo năm



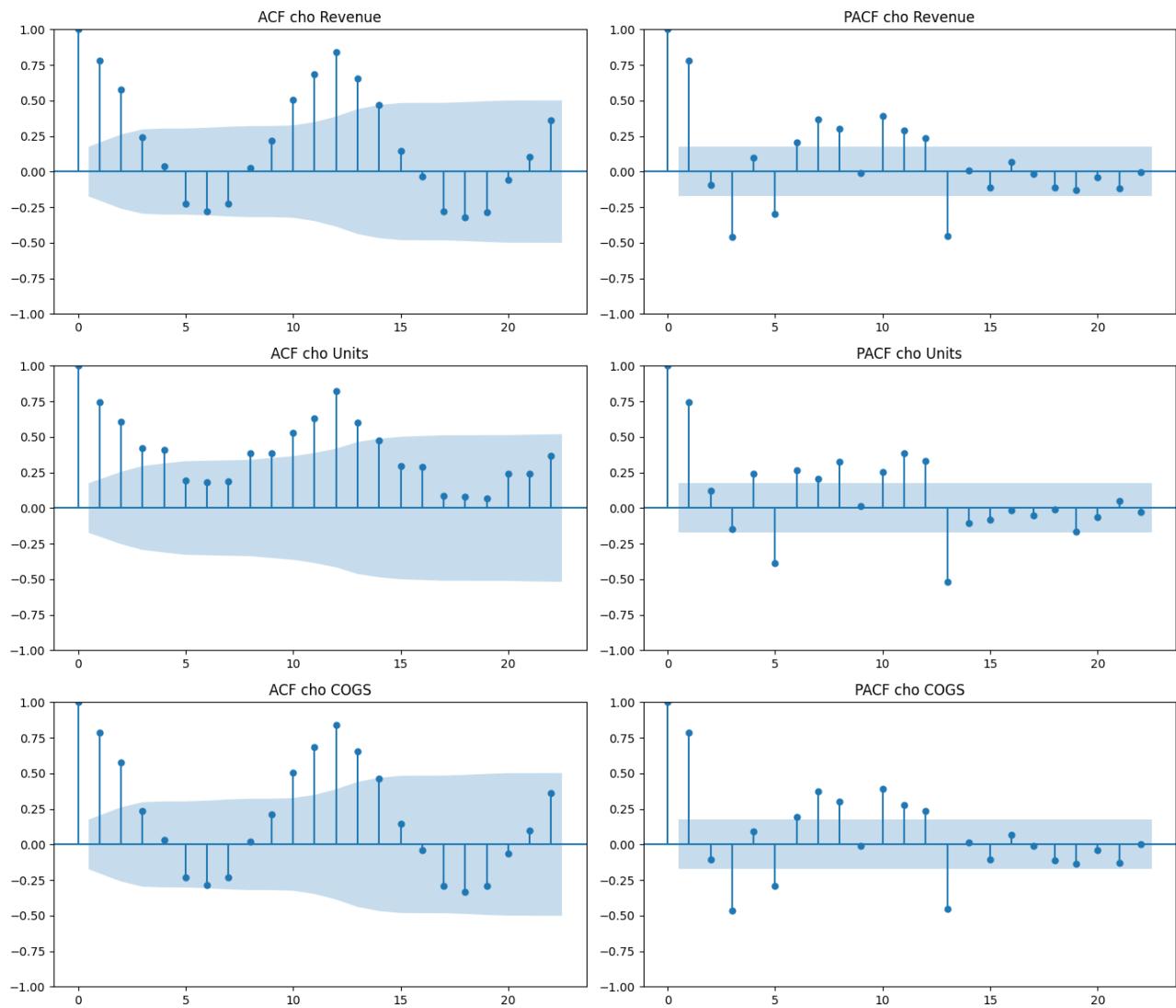
Hình 9: Các chuỗi thời gian sau khi lấy sai phân bậc 1



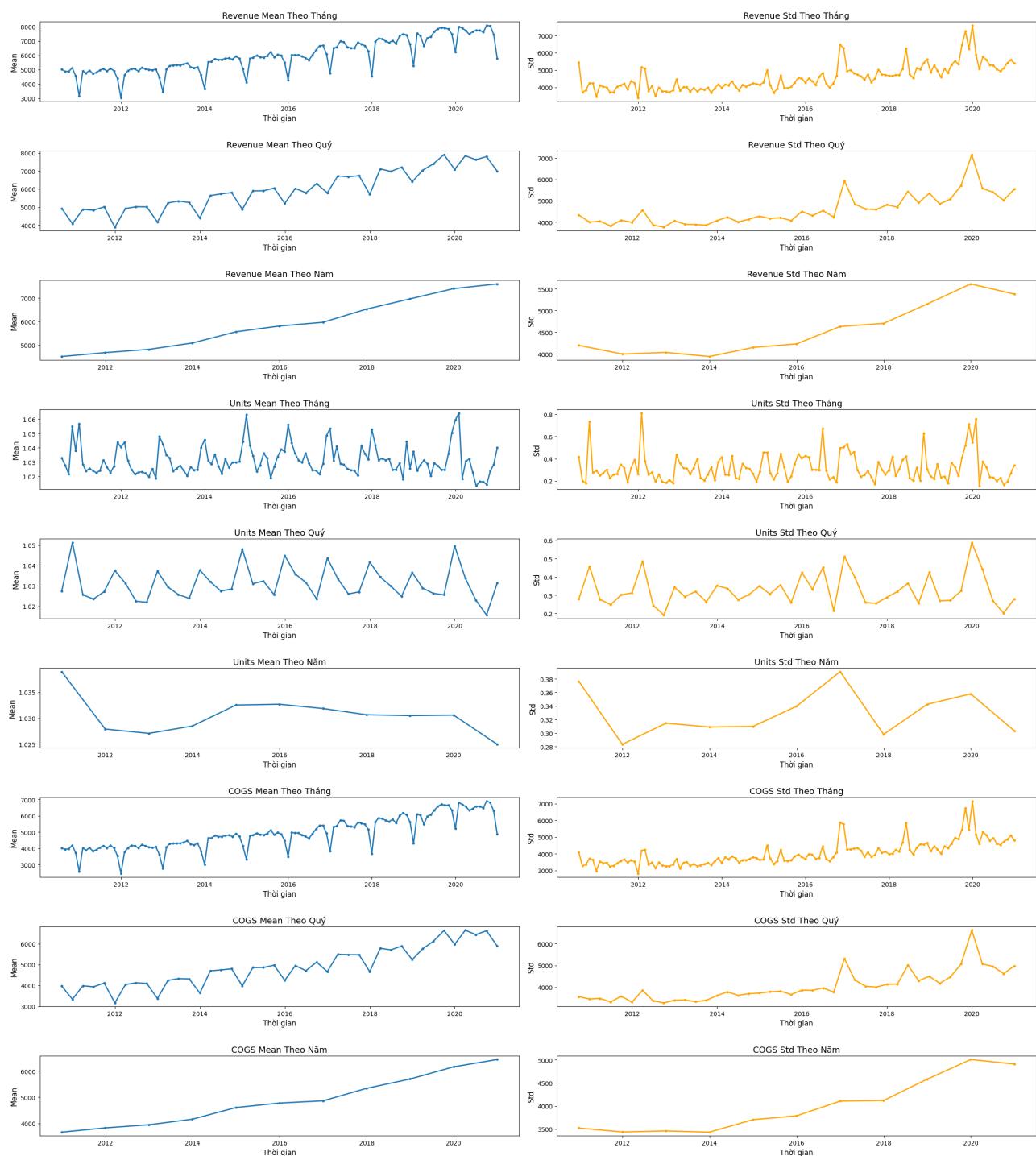
Hình 10: Sesonal Decomposition doanh thu, doanh số và giá vốn với các chu kỳ: 3, 6, 12 tháng



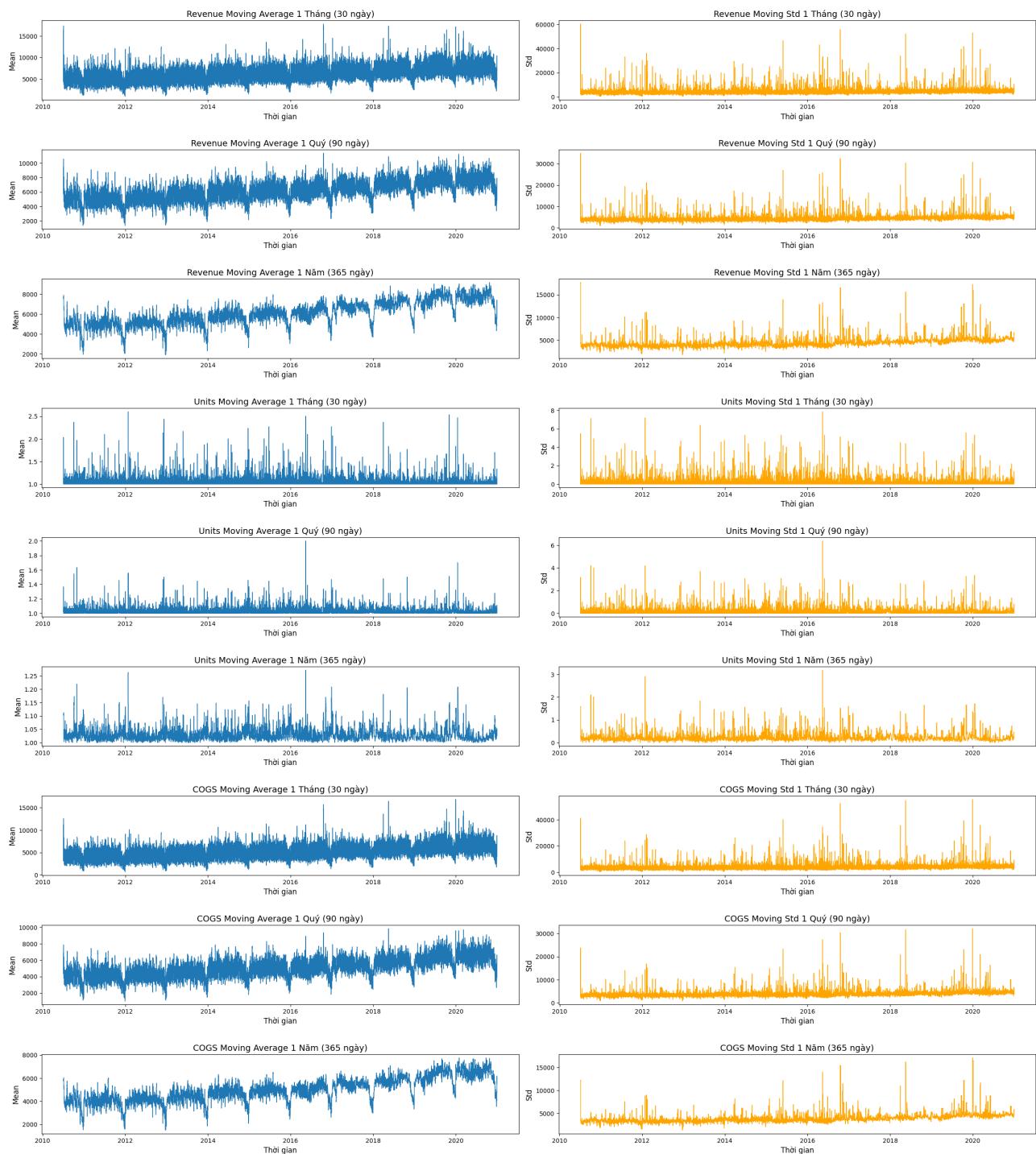
Hình 11: Phân tích ACF và PACF cho doanh thu, doanh số và giá vốn



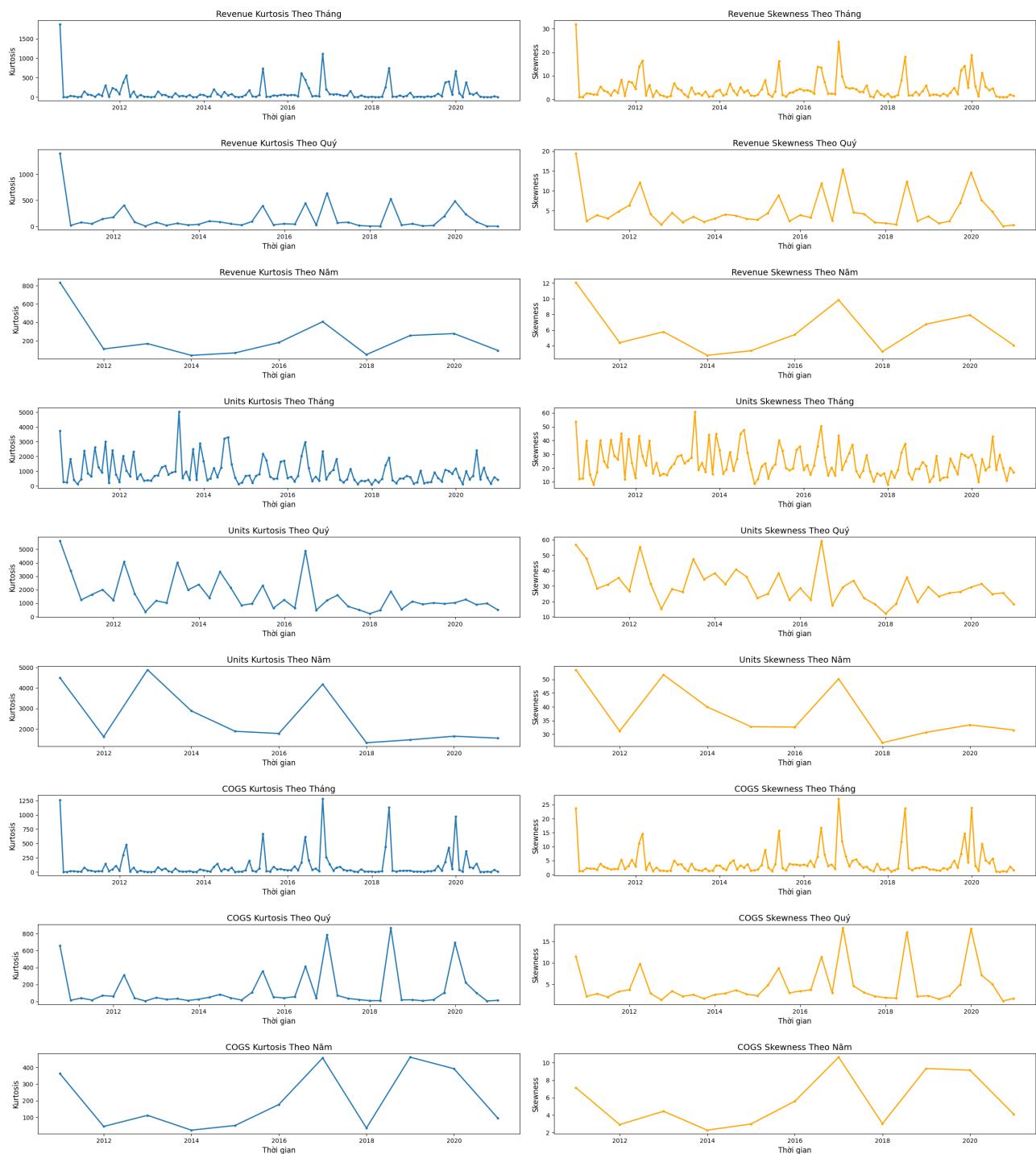
Hình 12: Phân tích trung bình và độ lệch chuẩn doanh thu, doanh số và giá vốn theo thời gian



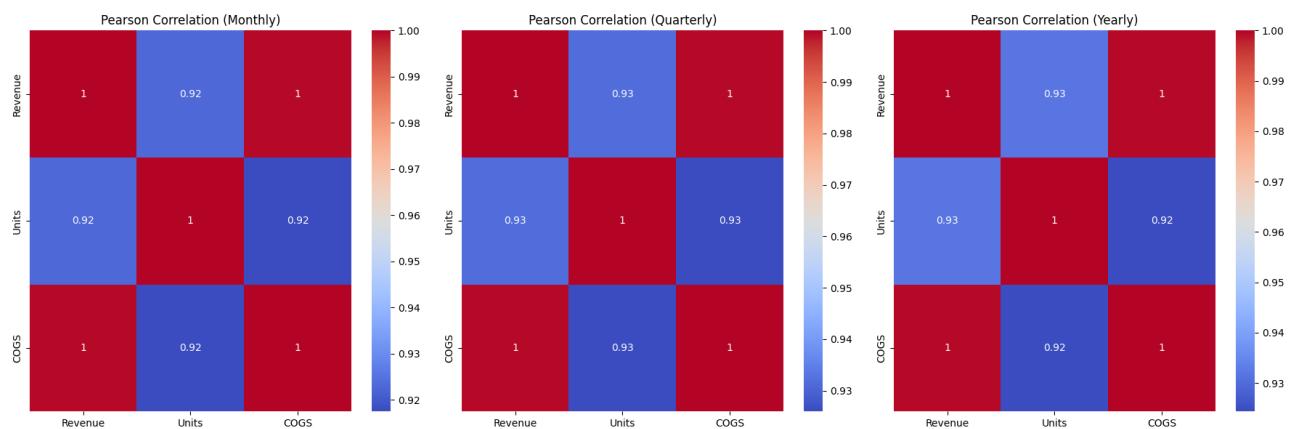
Hình 13: Phân tích trung bình trượt và độ lệch chuẩn trượt doanh thu, doanh số và giá vốn theo thời gian



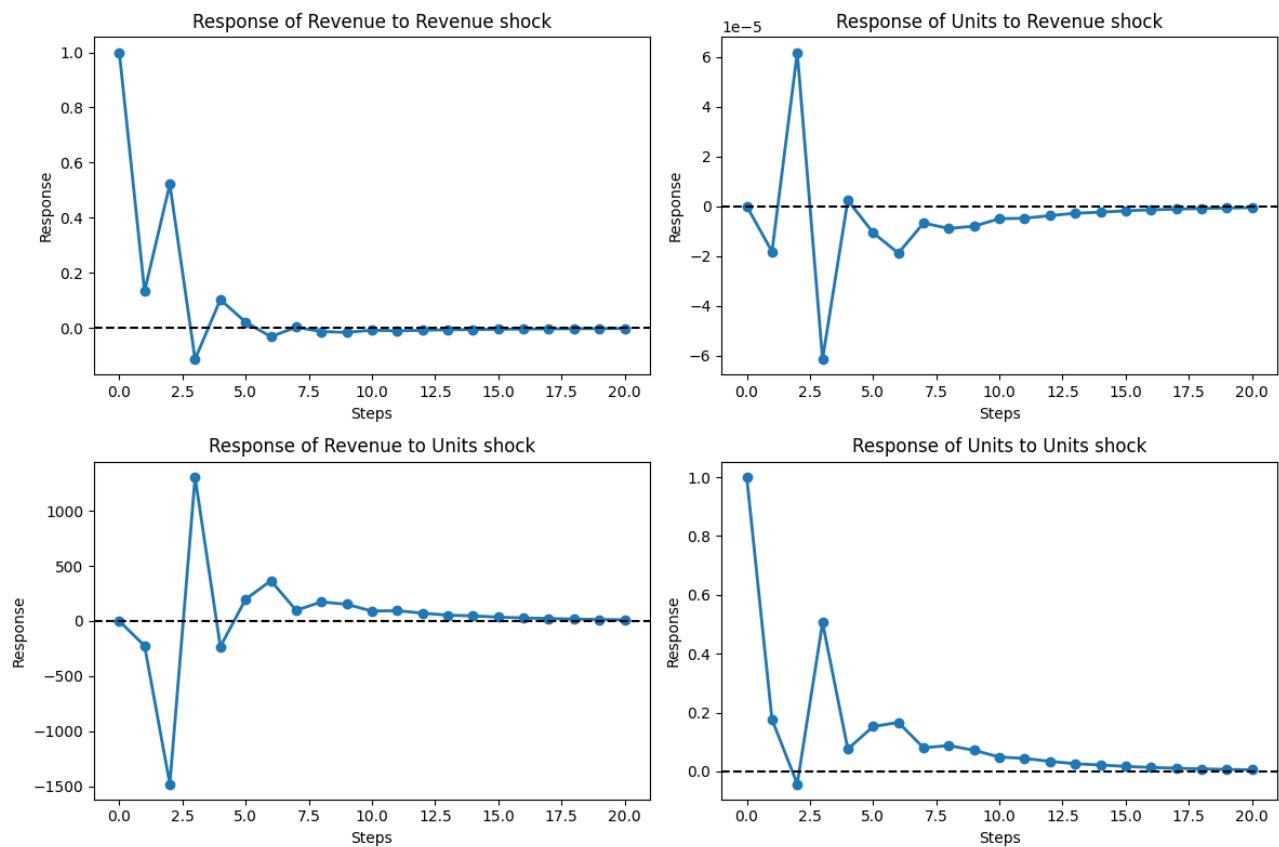
Hình 14: Phân tích Kurtosis và Skewness doanh thu, doanh số và giá vốn theo thời gian



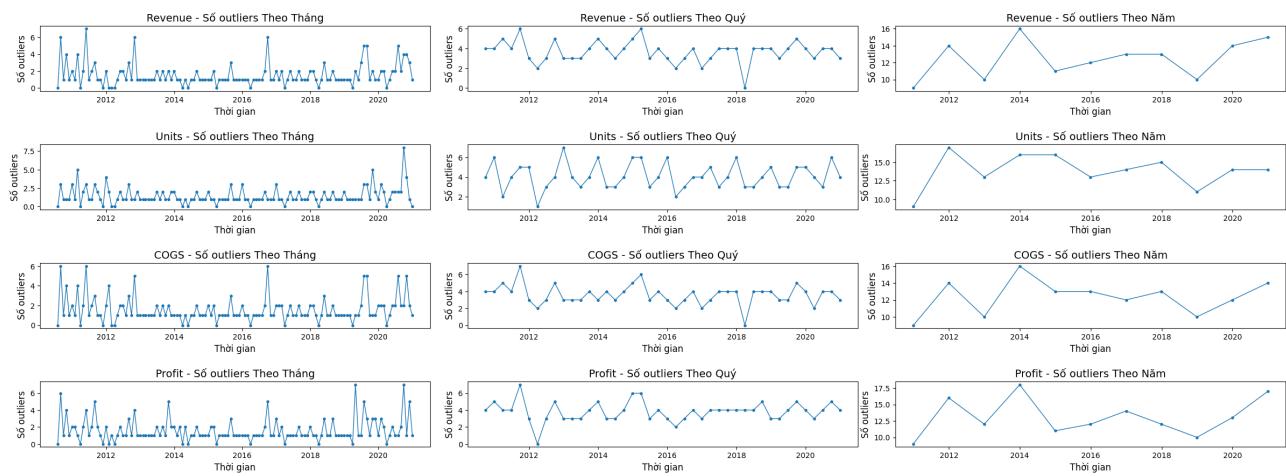
Hình 15: Ma trận tương quan giữa doanh thu, doanh số và giá vốn theo thời gian



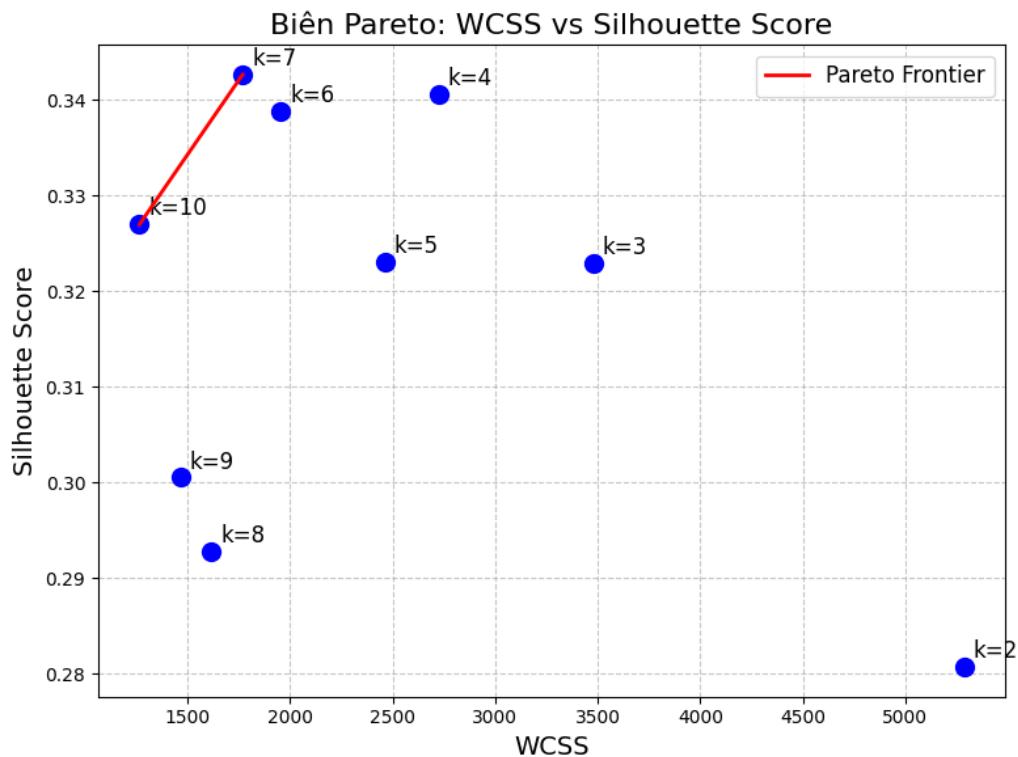
Hình 16: Biểu đồ đáp ứng xung giữa doanh thu và doanh số theo mô hình VAR



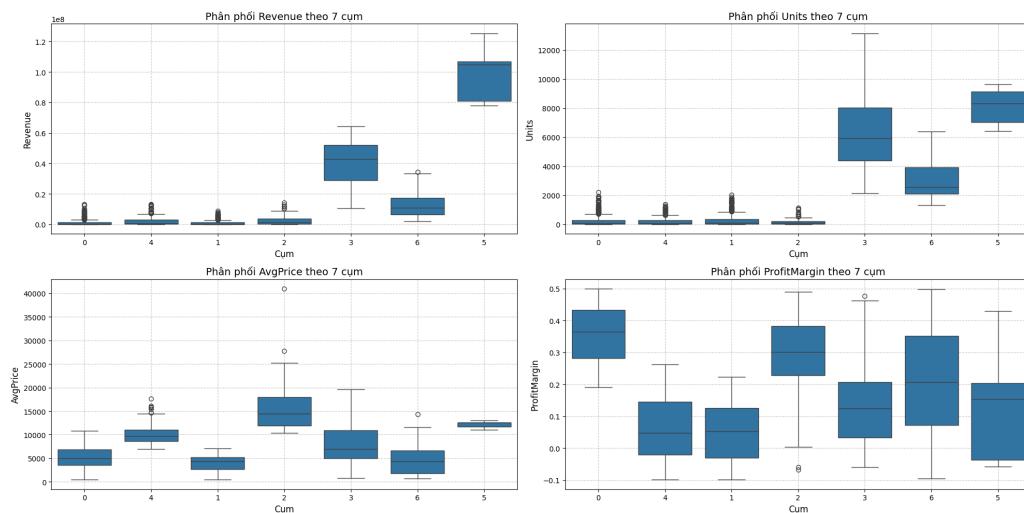
Hình 17: Biểu đồ thay đổi số lượng ngoại lai theo thời gian



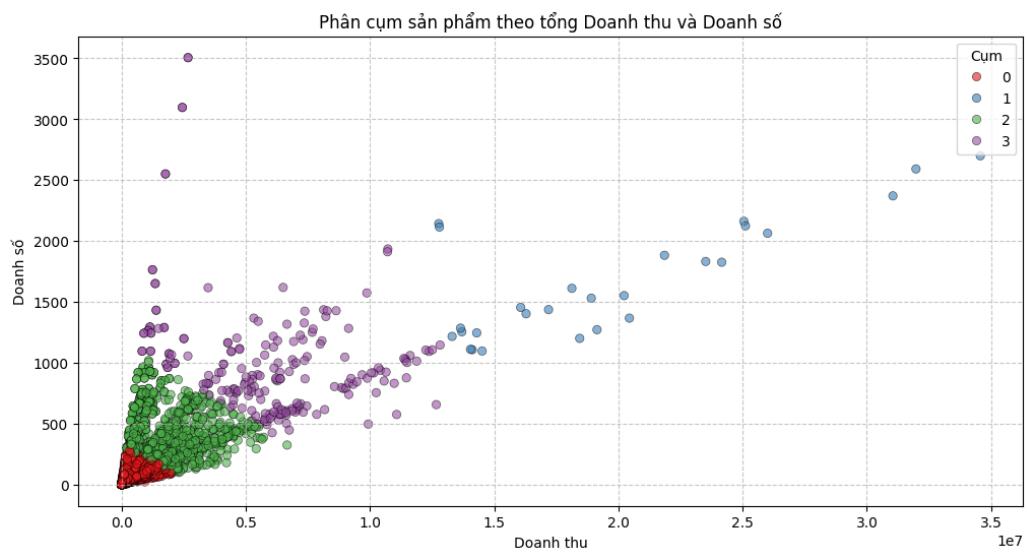
Hình 18: Biên Pareto theo chỉ số WCSS và Silhouette



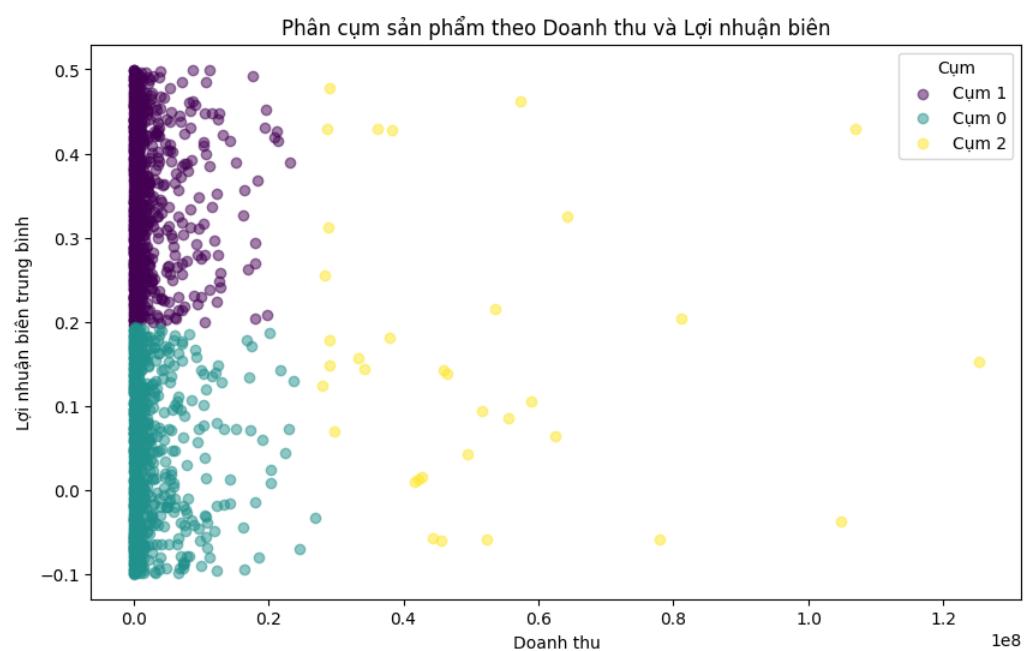
Hình 19: Biên Pareto theo chỉ số WCSS và Silhouette



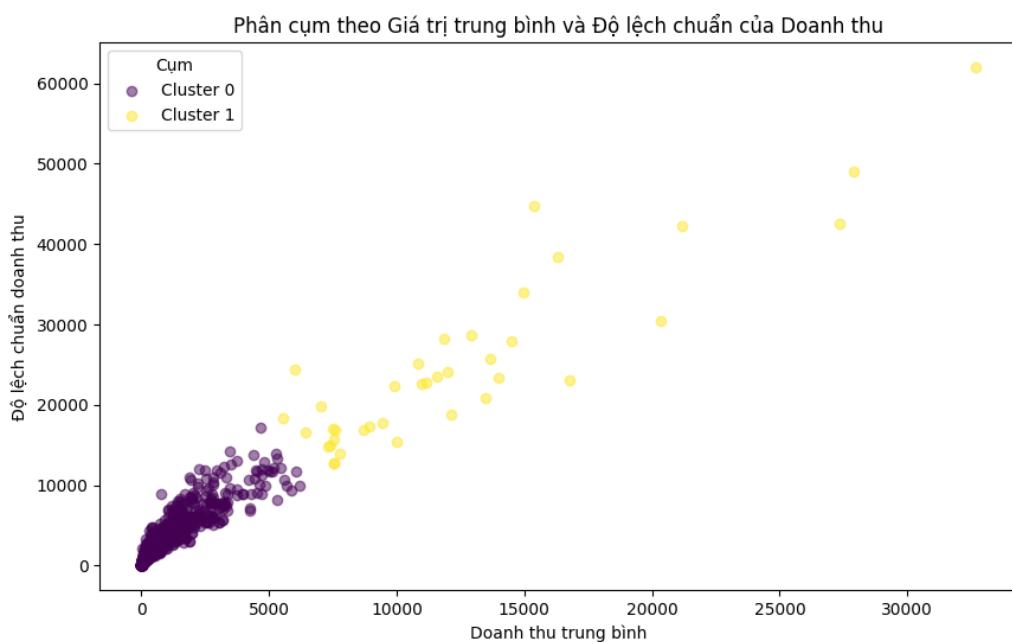
Hình 20: Biểu đồ phân cụm sản phẩm theo Doanh thu và Doanh số



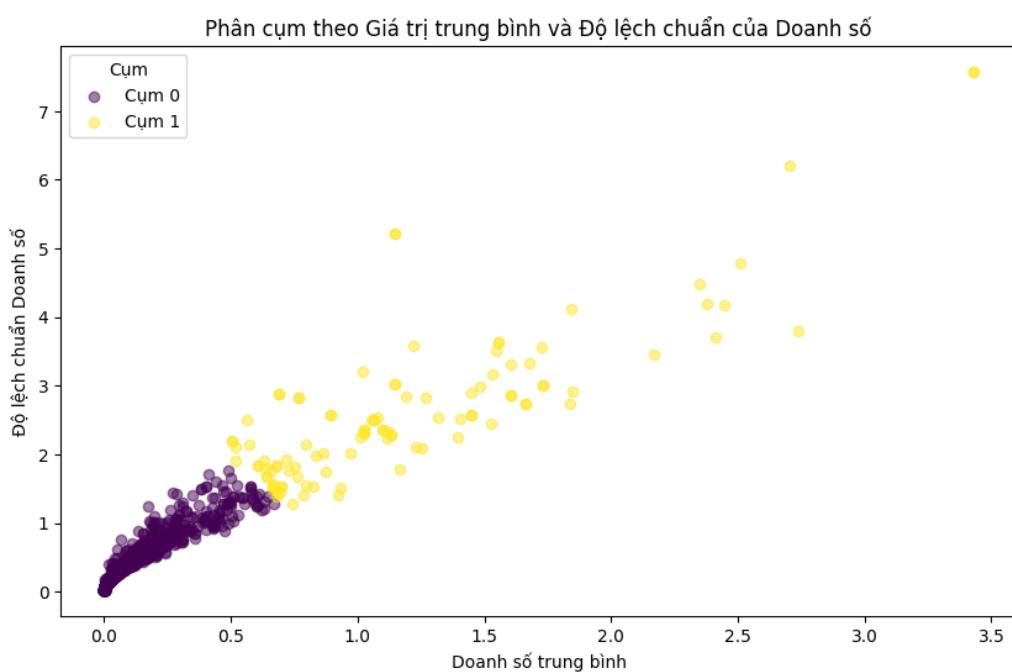
Hình 21: Biểu đồ phân cụm sản phẩm theo Doanh thu và Biên lợi nhuận



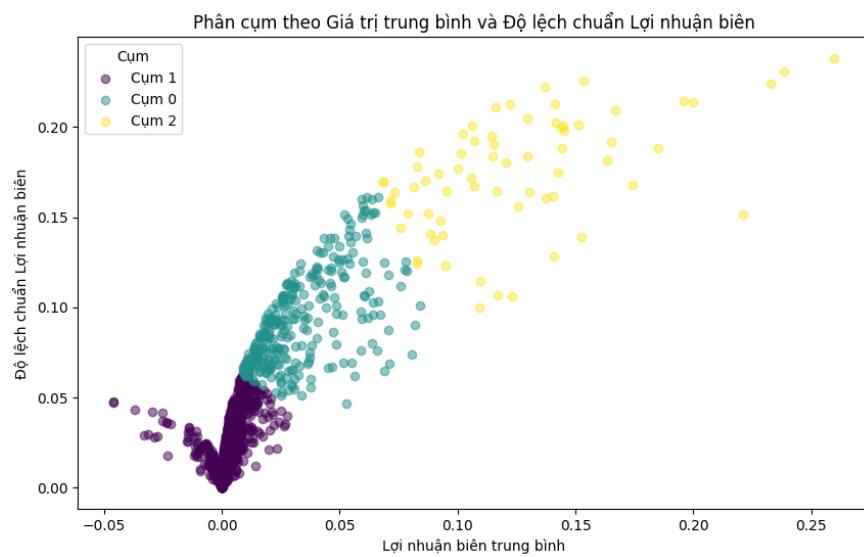
Hình 22: Biểu đồ phân cụm sản phẩm theo Trung bình và Độ lệch chuẩn Doanh thu



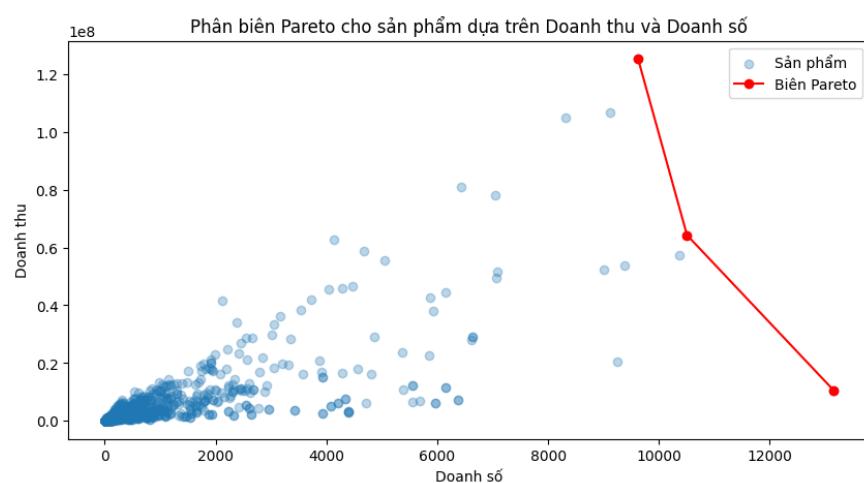
Hình 23: Biểu đồ phân cụm sản phẩm theo Trung bình và Độ lệch chuẩn Doanh số



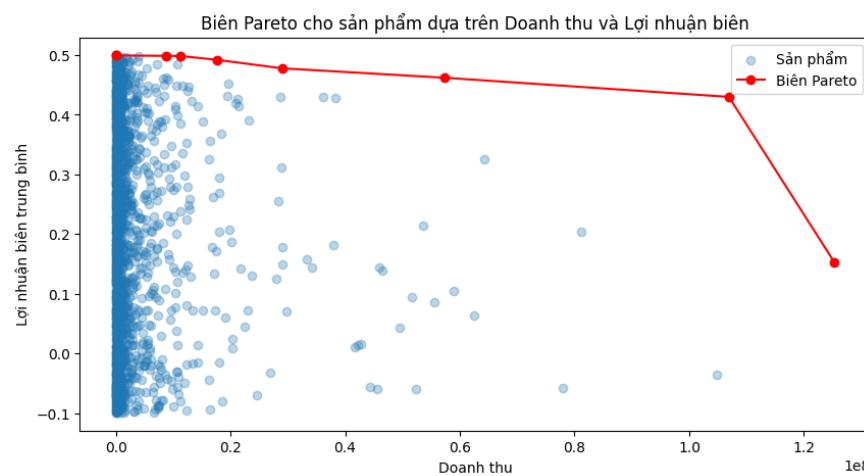
Hình 24: Biểu đồ phân cụm sản phẩm theo Trung bình và Độ lệch chuẩn Lợi nhuận biên



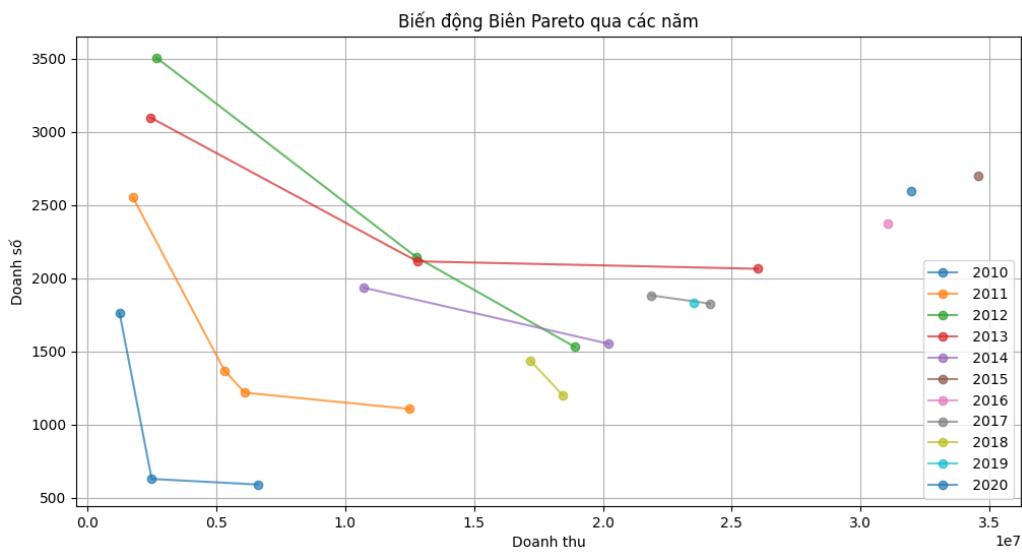
Hình 25: Biểu đồ biên Pareto theo Doanh thu và Doanh số



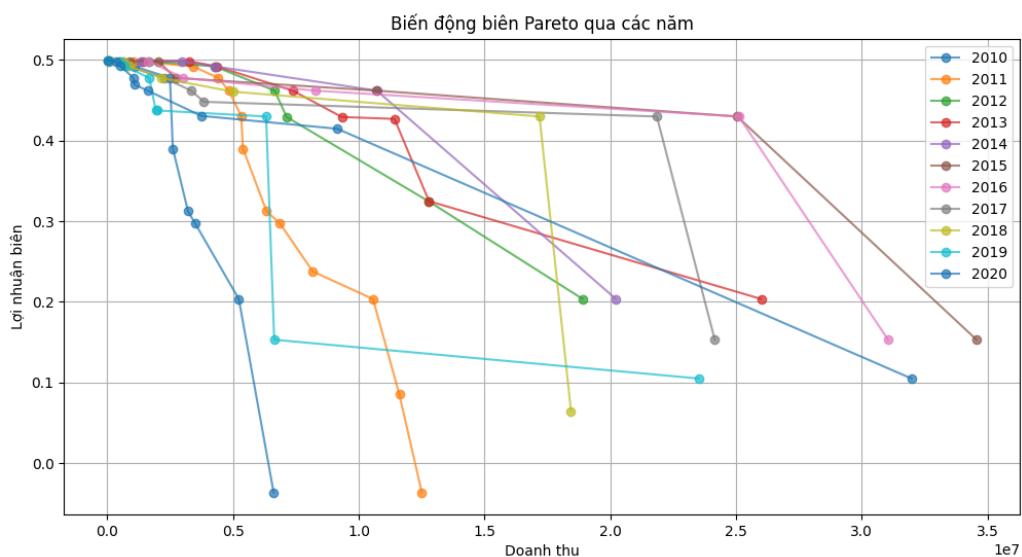
Hình 26: Biểu đồ biên Pareto theo Doanh thu và Biên lợi nhuận



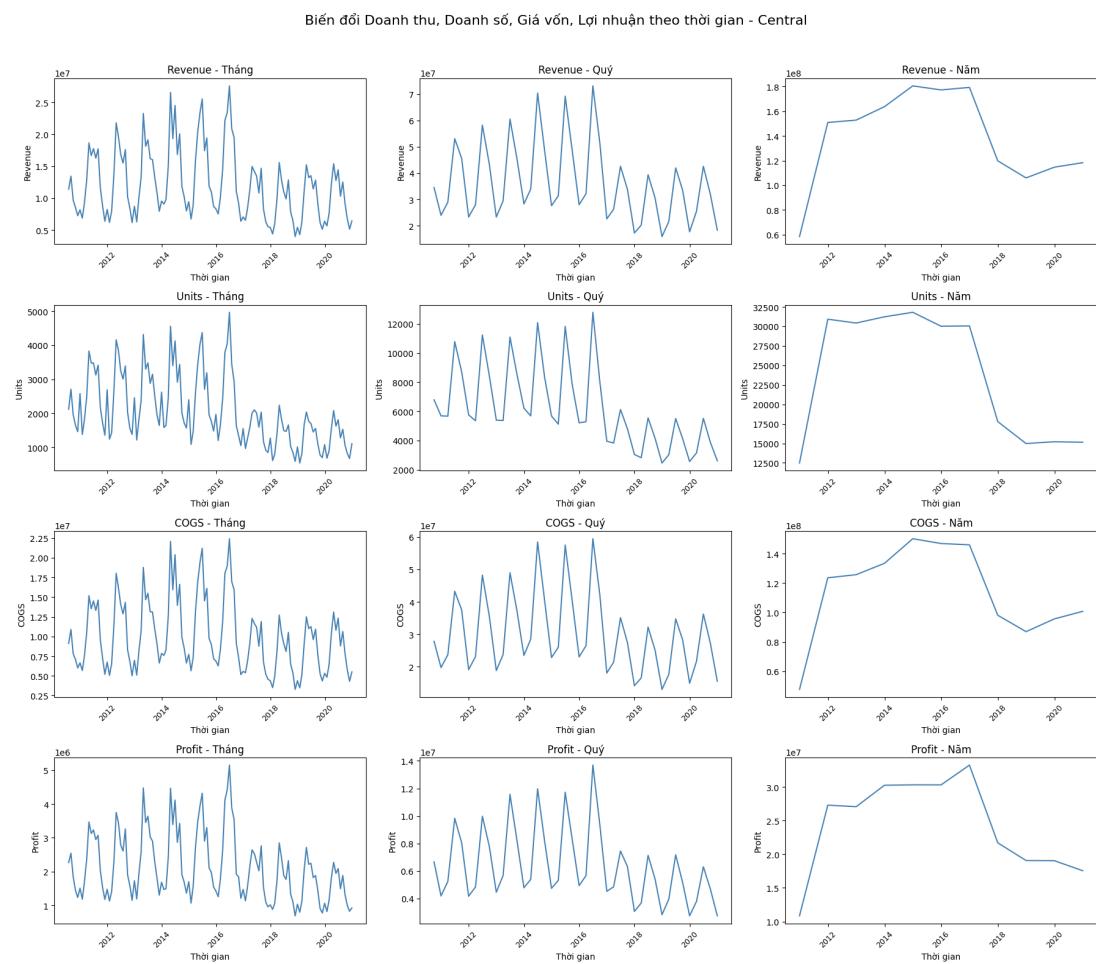
Hình 27: Biểu đồ biến động biên Pareto theo Doanh thu và Doanh số qua các năm



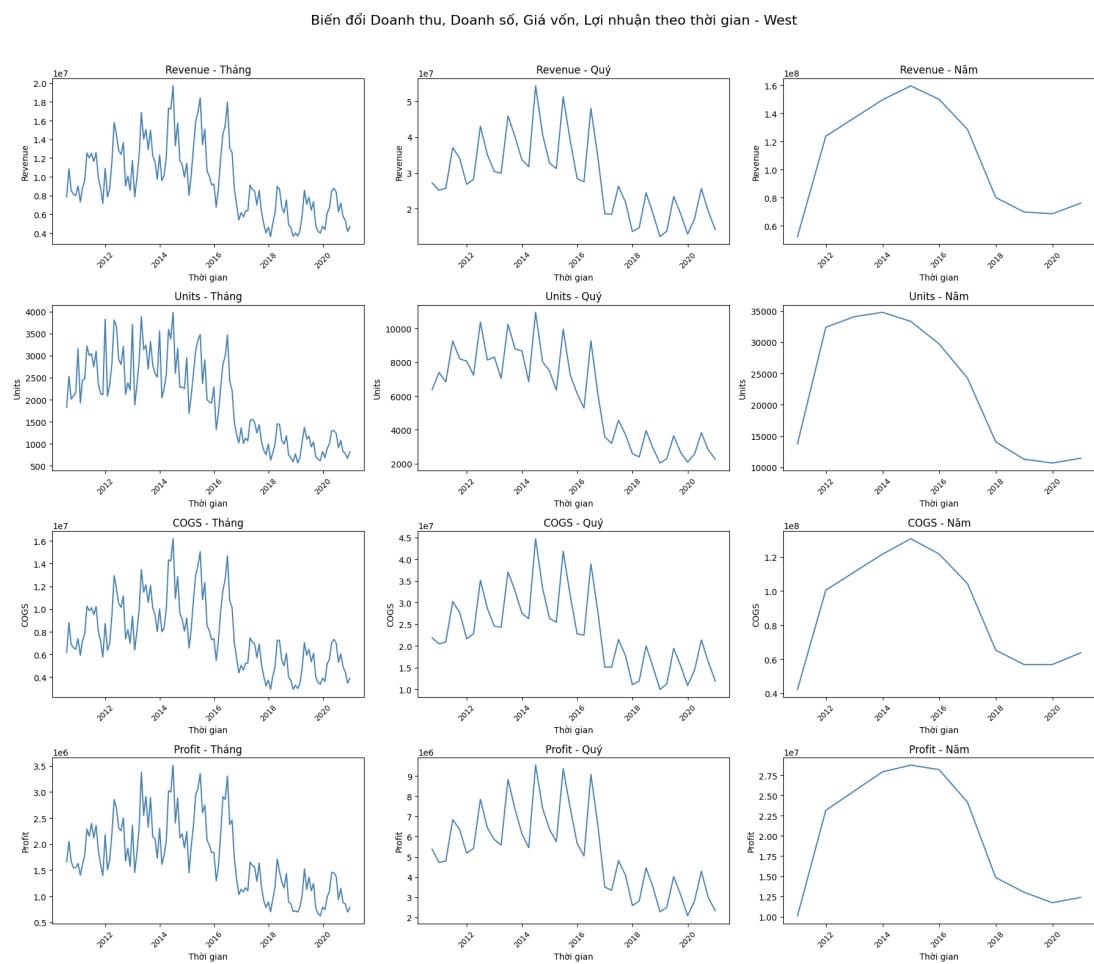
Hình 28: Biểu đồ biến động biên Pareto theo Doanh thu và Biên lợi nhuận qua các năm



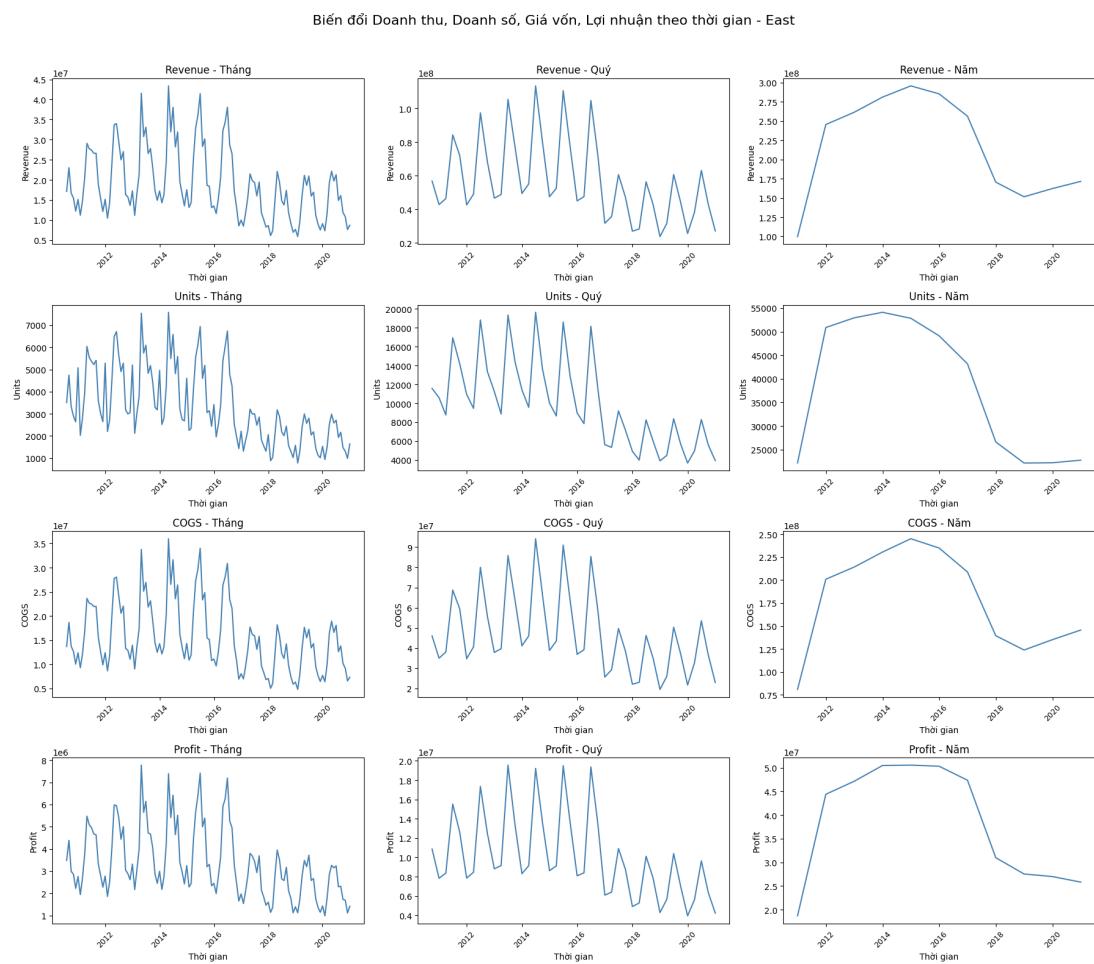
Hình 29: Biến đổi Doanh thu, Doanh số, Giá vốn, Lợi nhuận theo thời gian - Central



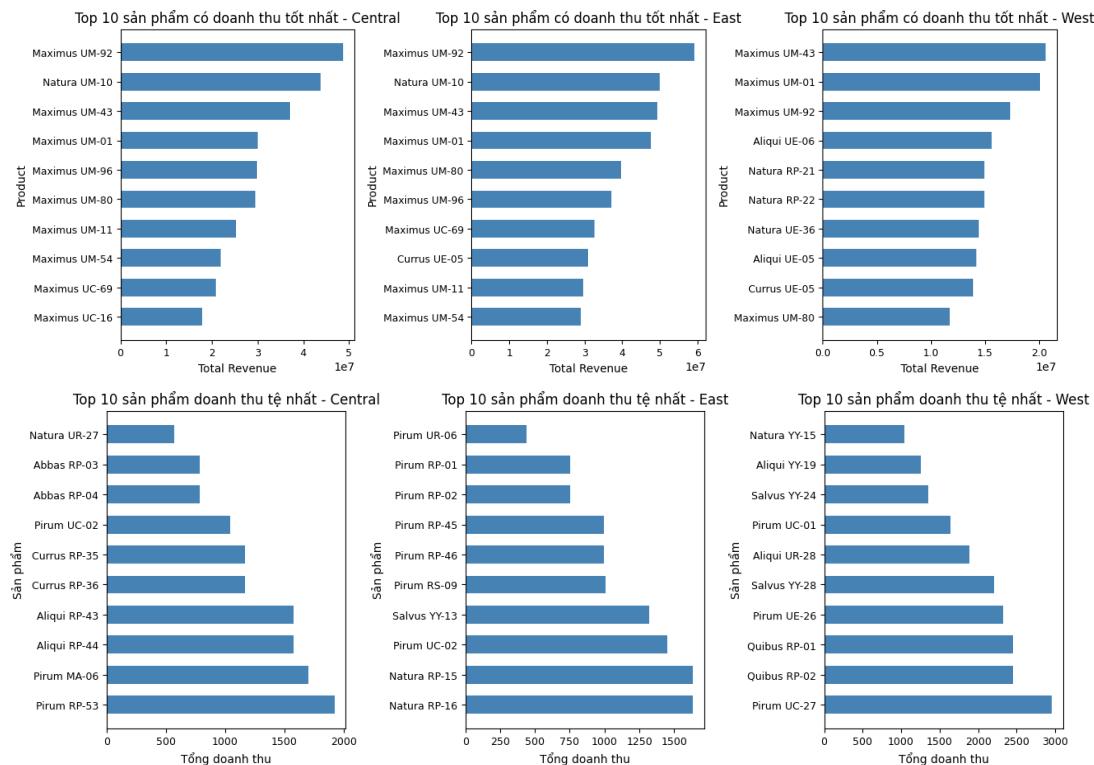
Hình 30: Biến đổi Doanh thu, Doanh số, Giá vốn, Lợi nhuận theo thời gian - West



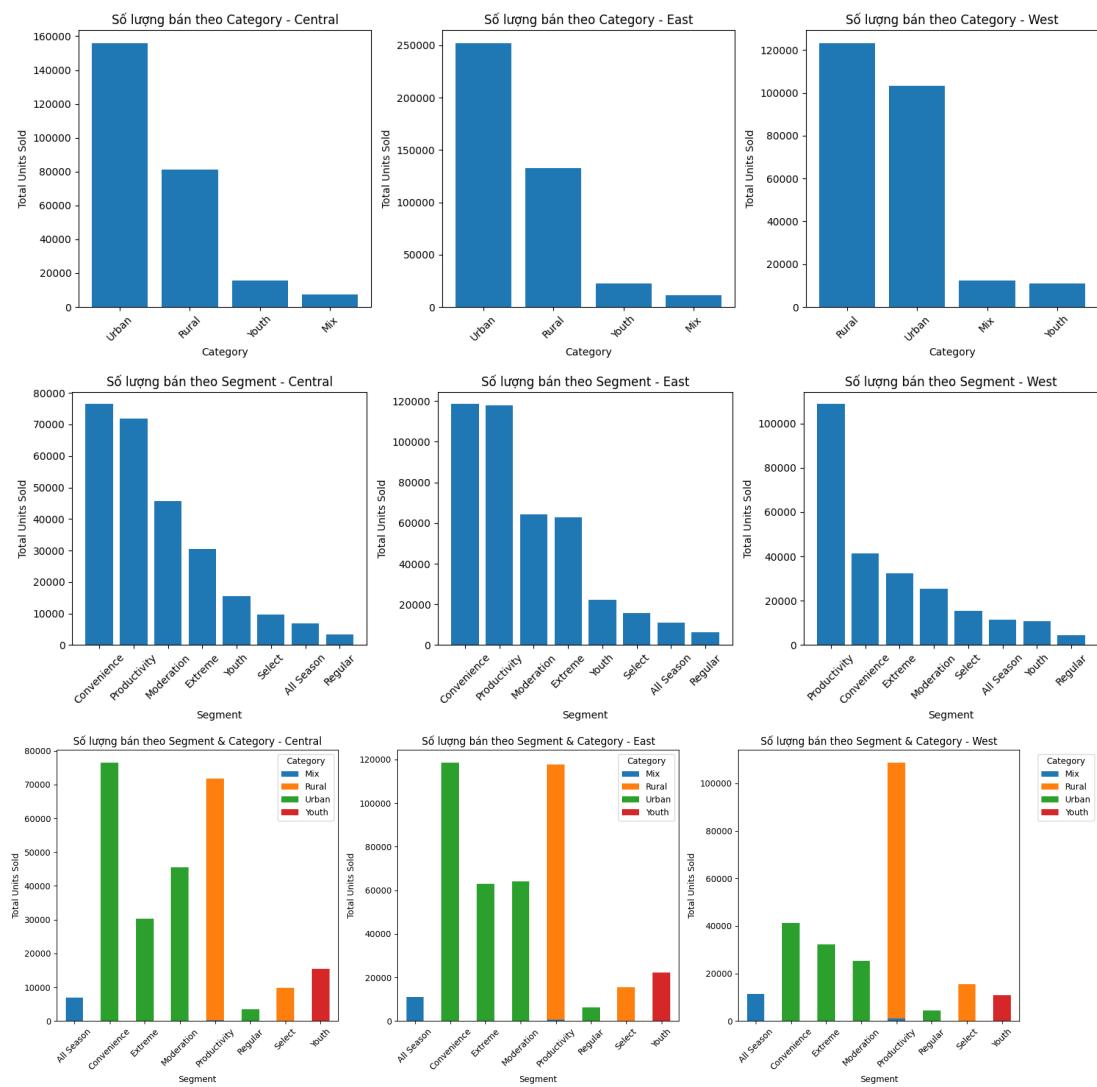
Hình 31: Biến đổi Doanh thu, Doanh số, Giá vốn, Lợi nhuận theo thời gian - East



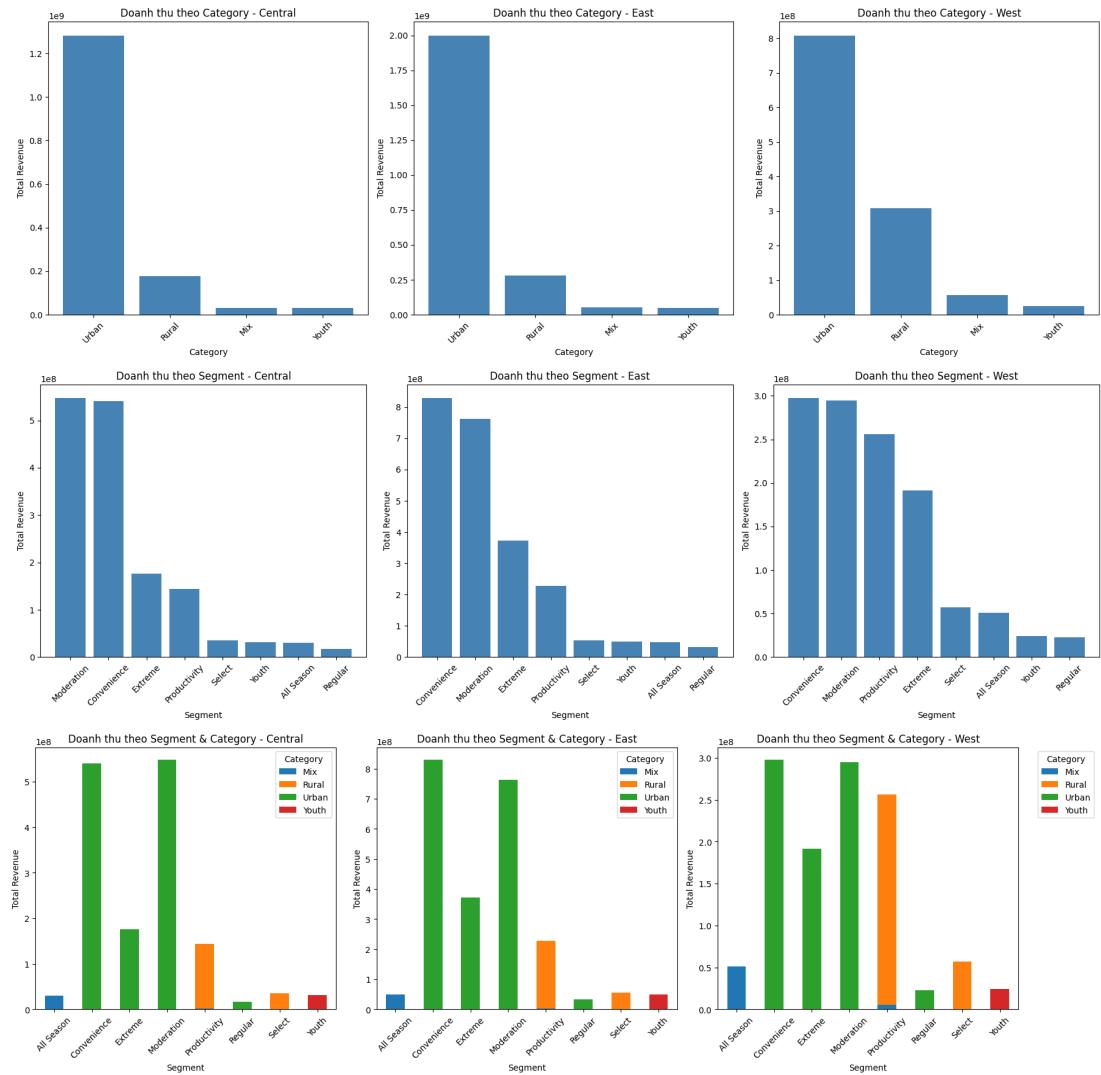
Hình 32: Top 10 &amp; Bottom 10 của từng khu vực dựa trên doanh thu



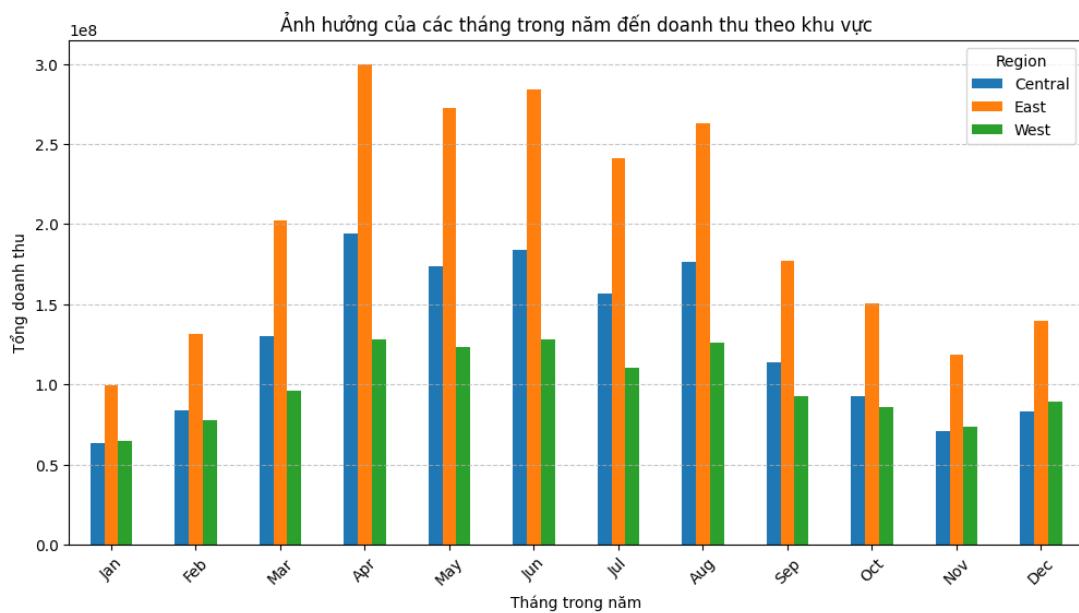
Hình 33: Thứ hạng về doanh số của các danh mục và phân khúc sản phẩm tại từng vùng



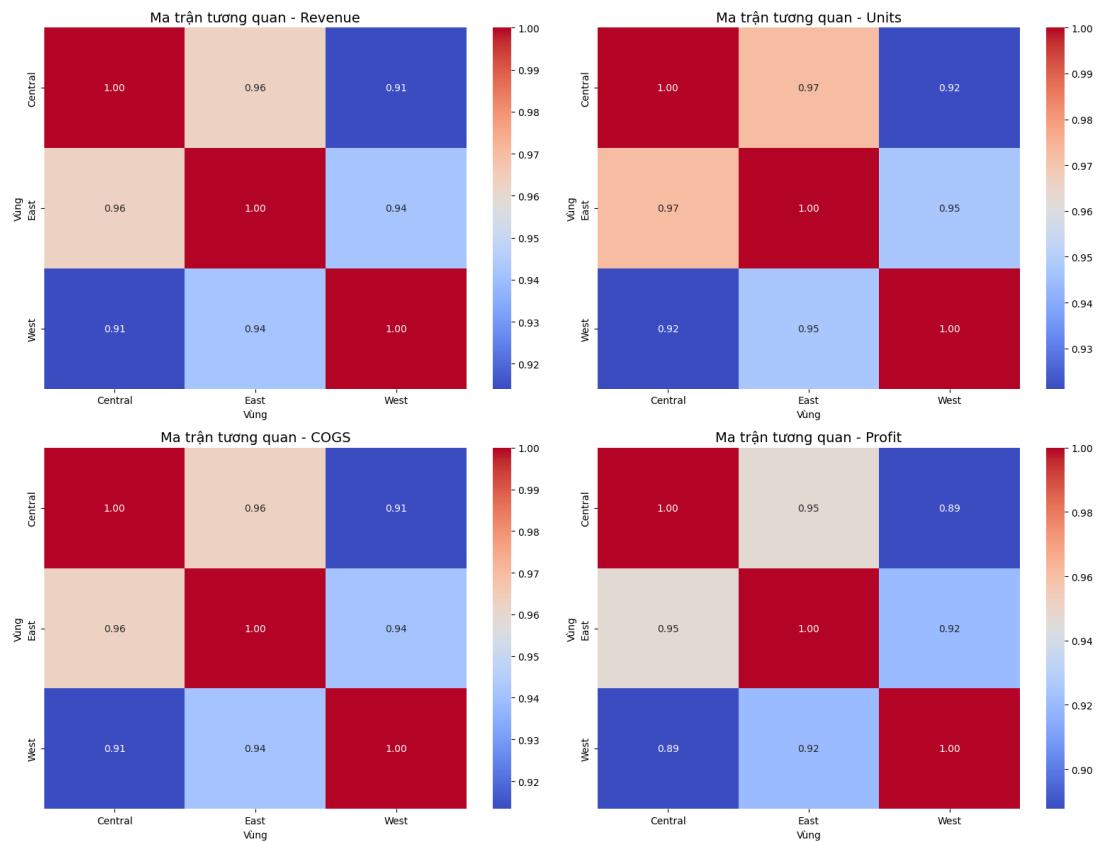
Hình 34: Thứ hạng về doanh thu của các danh mục và phân khúc sản phẩm tại từng vùng



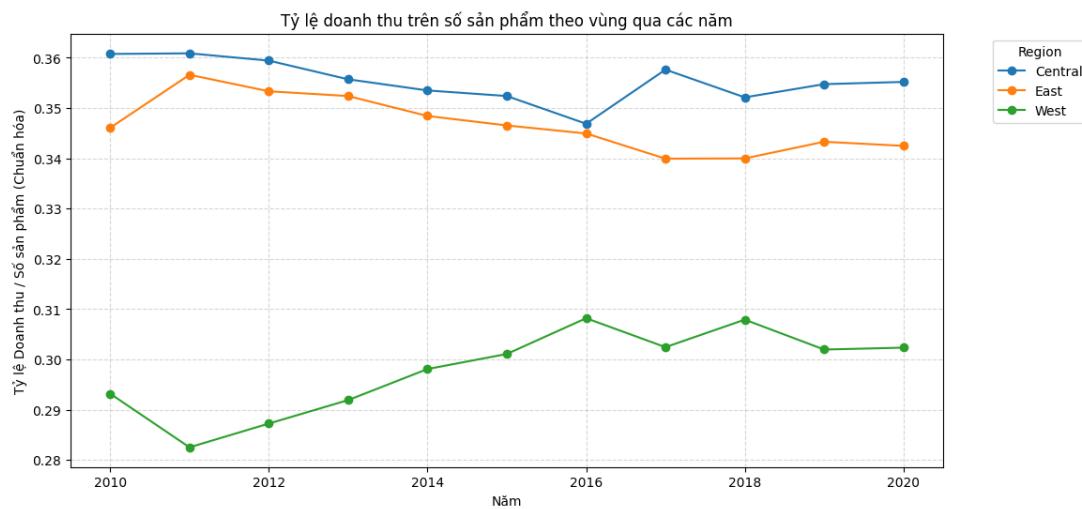
Hình 35: Ảnh hưởng của các thời điểm trong năm đến doanh thu



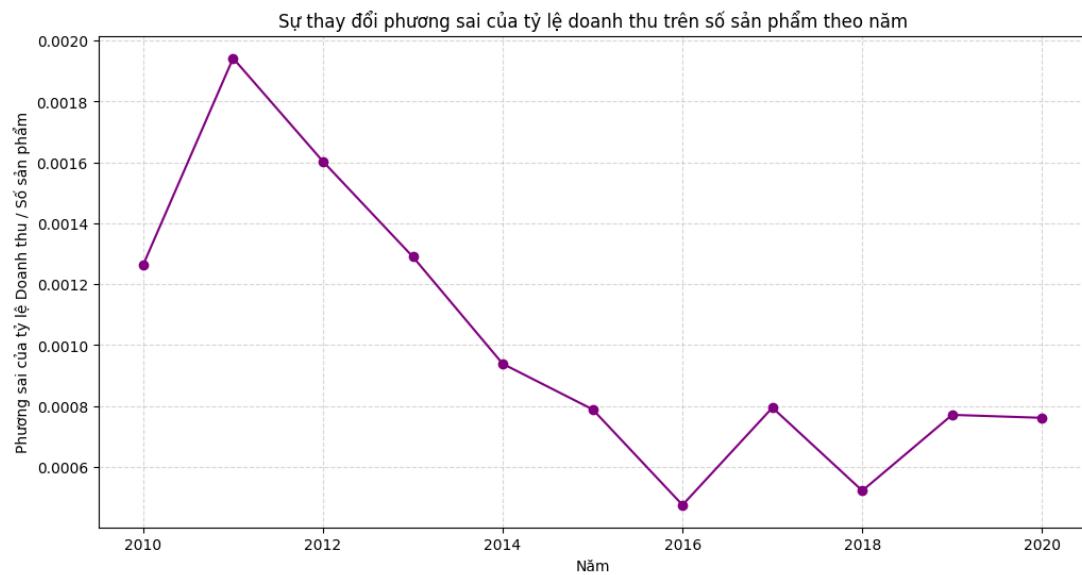
Hình 36: Ma trận tương quan về doanh thu, doanh số, giá vốn, lợi nhuận giữa các cặp vùng



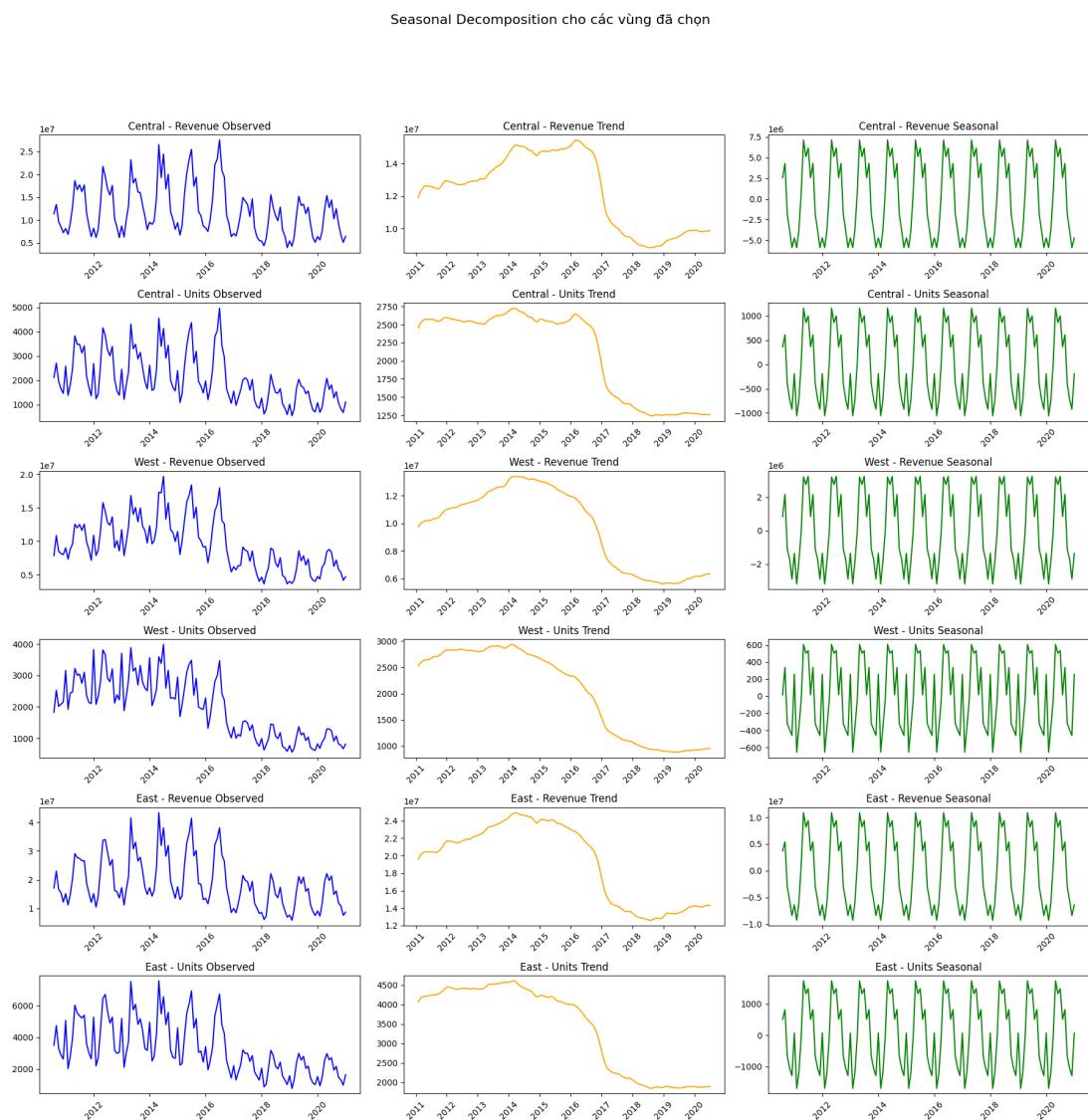
Hình 37: Tỷ lệ doanh thu trên số sản phẩm theo khu vực qua các năm



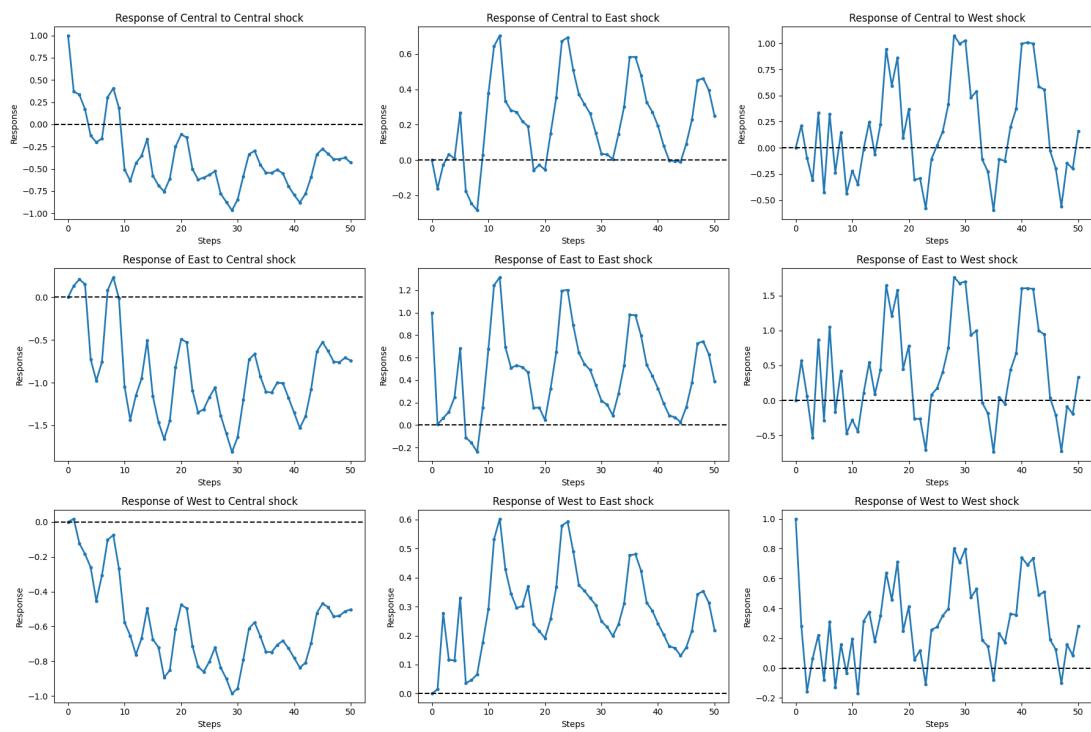
Hình 38: Sự thay đổi phương sai của tỷ lệ doanh thu trên số sản phẩm theo năm



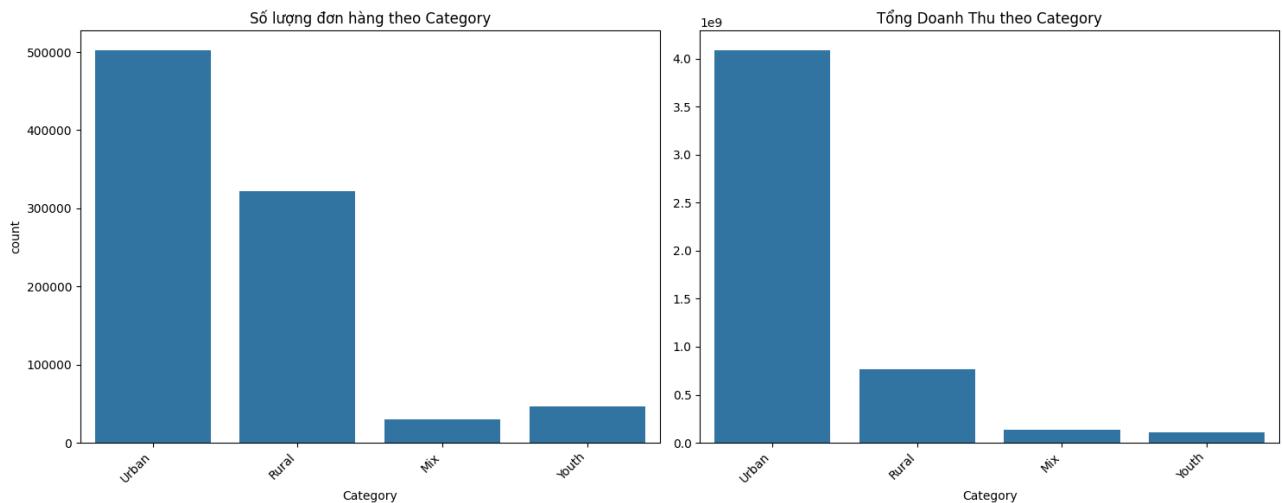
Hình 39: Yếu tố mùa vụ của từng khu vực



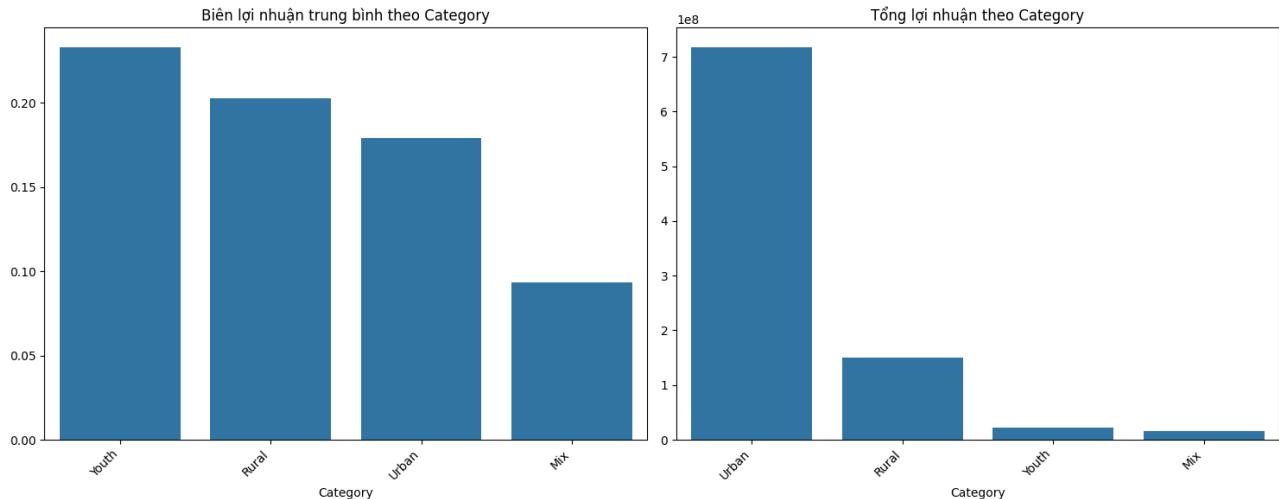
Hình 40: Biểu đồ đáp ứng xung cho 3 khu vực theo mô hình VAR



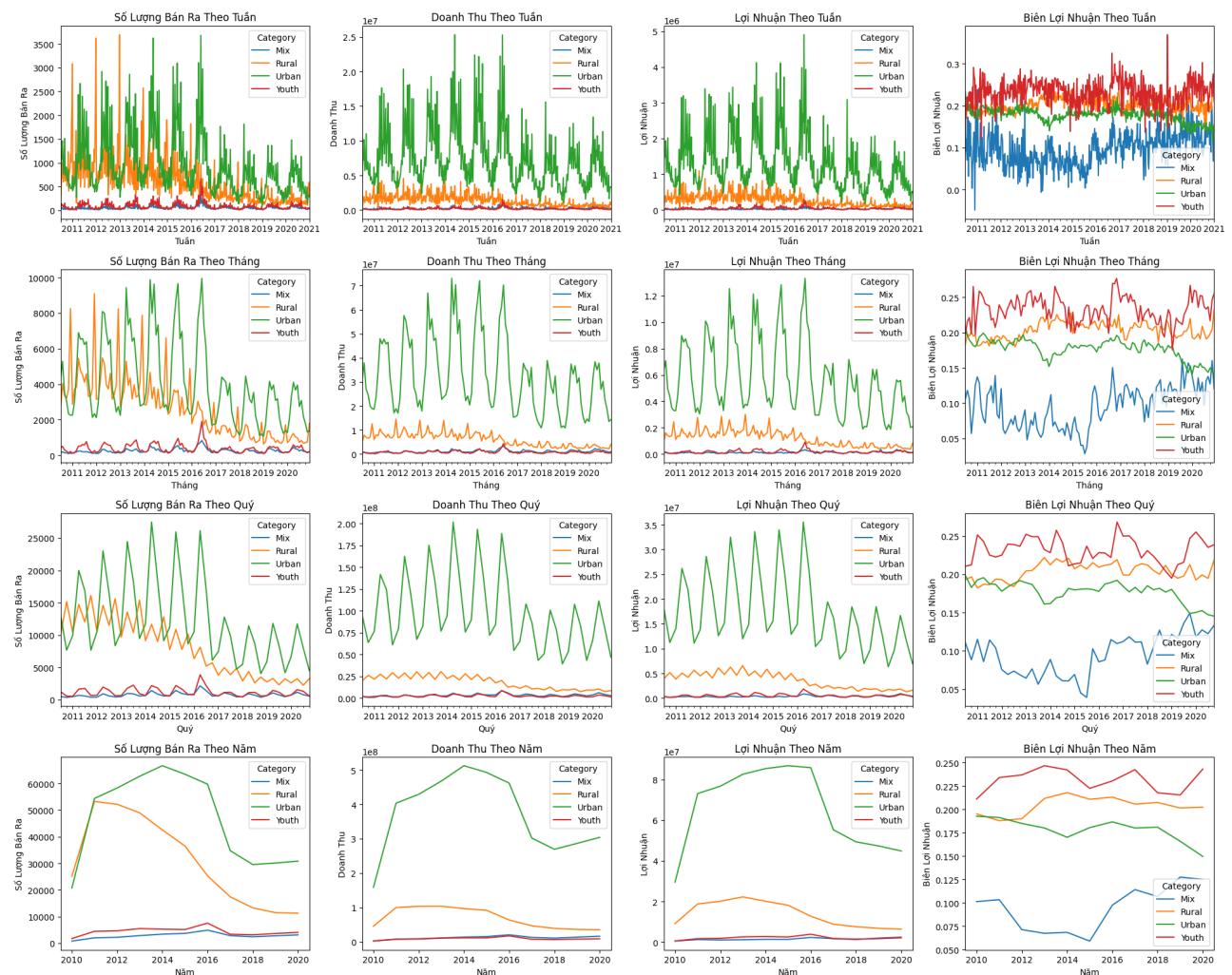
Hình 41: Biểu đồ số lượng đơn hàng và tổng doanh thu của từng danh mục sản phẩm



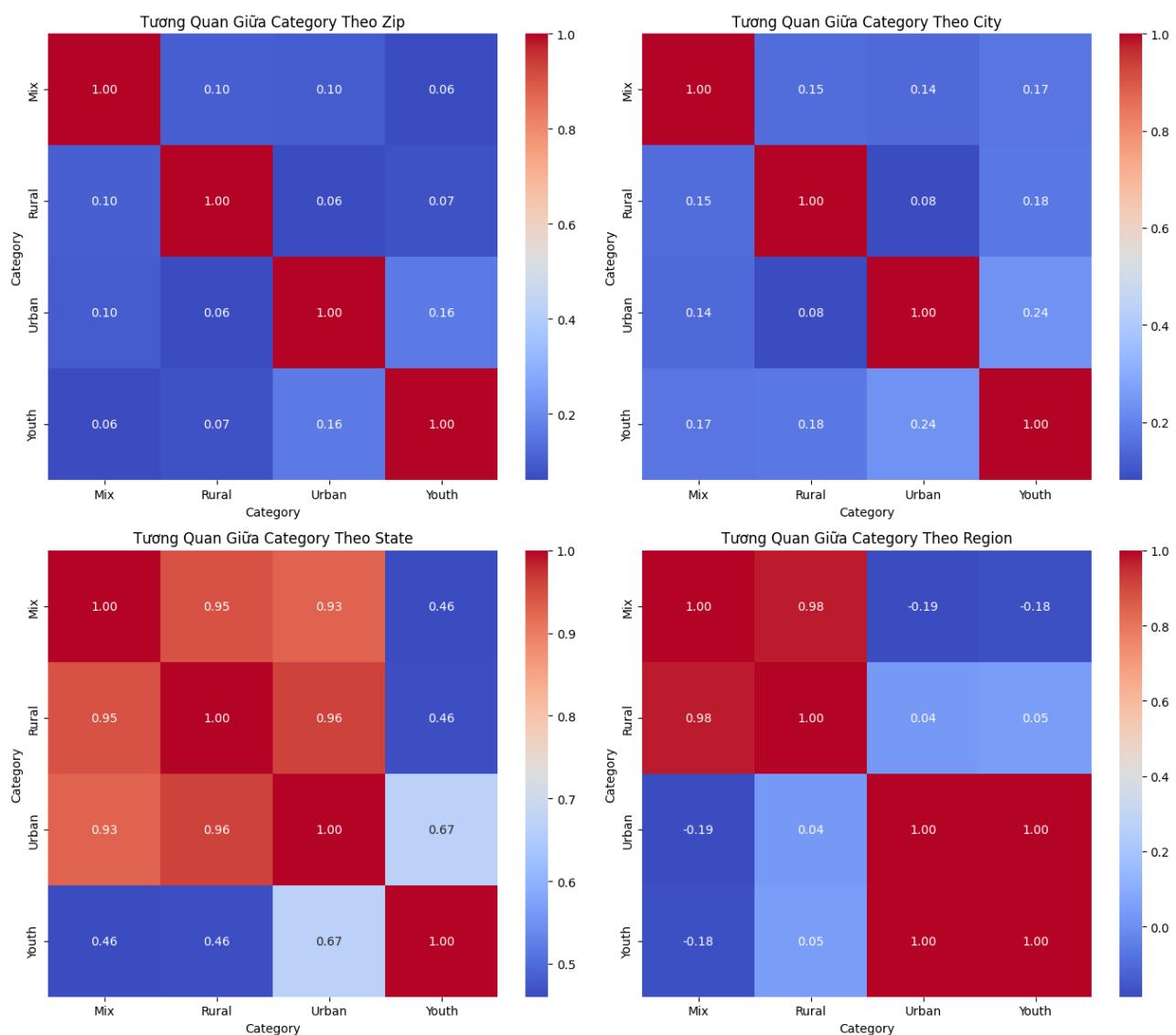
Hình 42: Biểu đồ mối quan hệ giữa biên lợi nhuận và tổng lợi nhuận theo danh mục sản phẩm



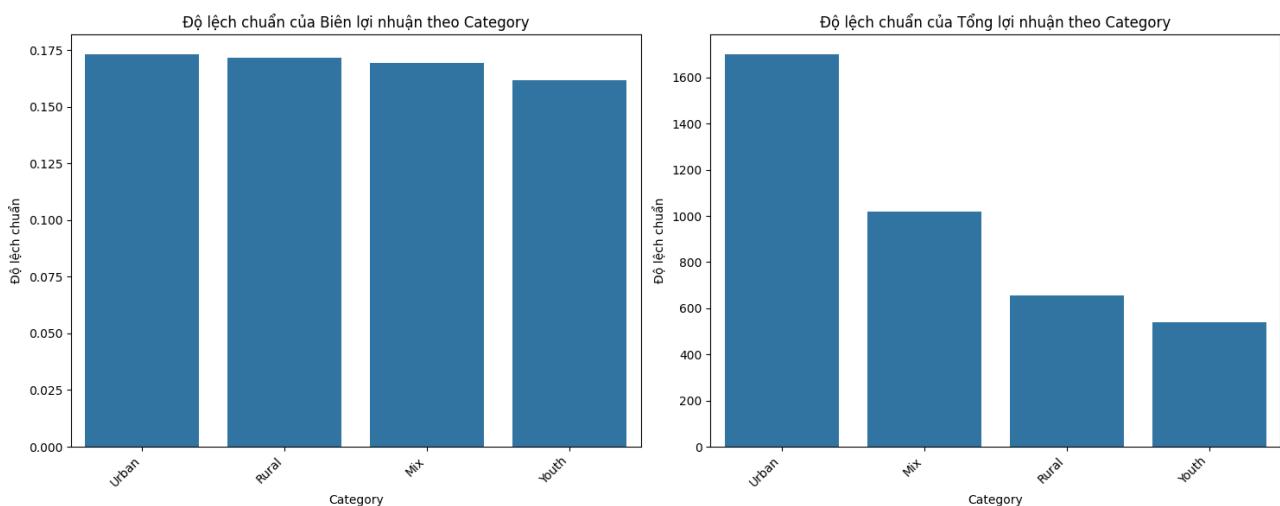
Hình 43: Biểu đồ biến động doanh số, doanh thu, lợi nhuận và biên lợi nhuận của từng danh mục theo thời gian



Hình 44: Biểu đồ mối quan hệ giữa các danh mục sản phẩm theo khu vực, thành phố và mã vùng



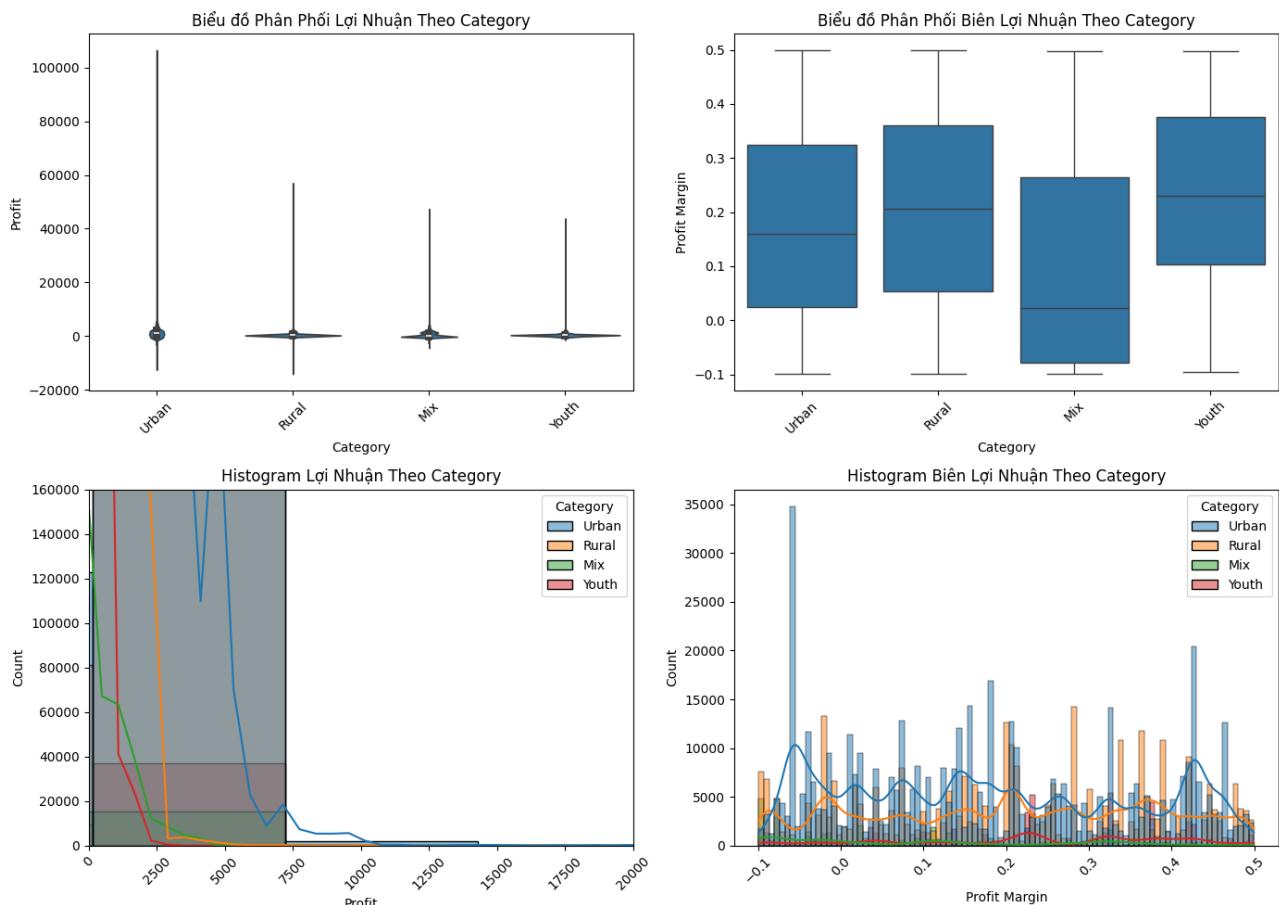
Hình 45: Biểu đồ độ lệch chuẩn của biên lợi nhuận và lợi nhuận theo danh mục sản phẩm



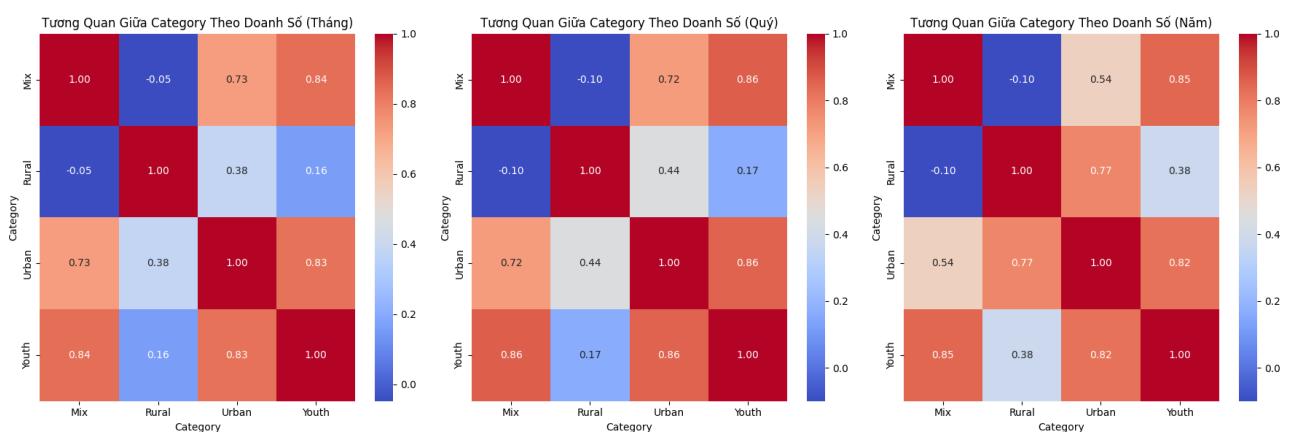
Hình 46: Biểu đồ biến động độ lệch chuẩn của biên lợi nhuận và lợi nhuận theo thời gian



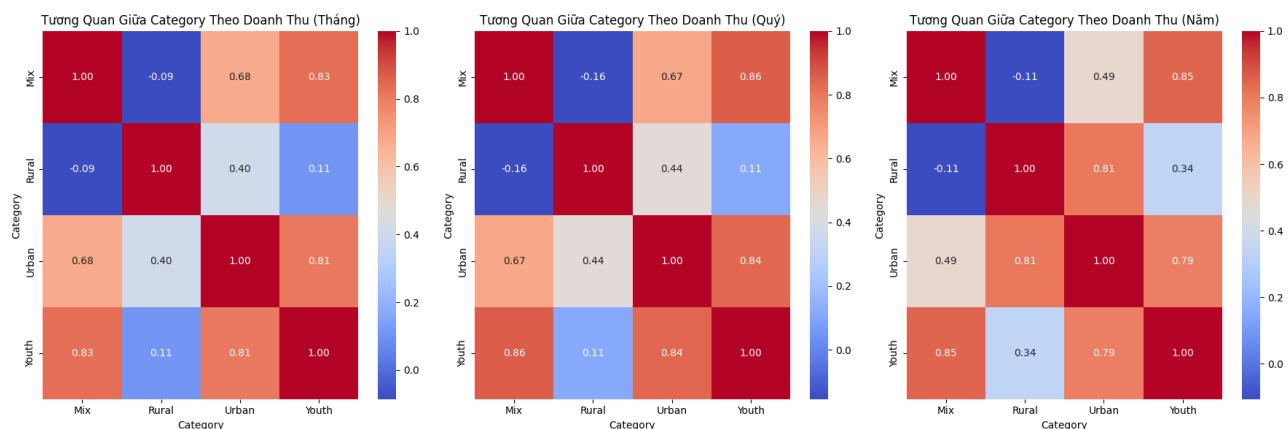
Hình 47: Biểu đồ phân phối và histogram của biên lợi nhuận và lợi nhuận theo danh mục sản phẩm



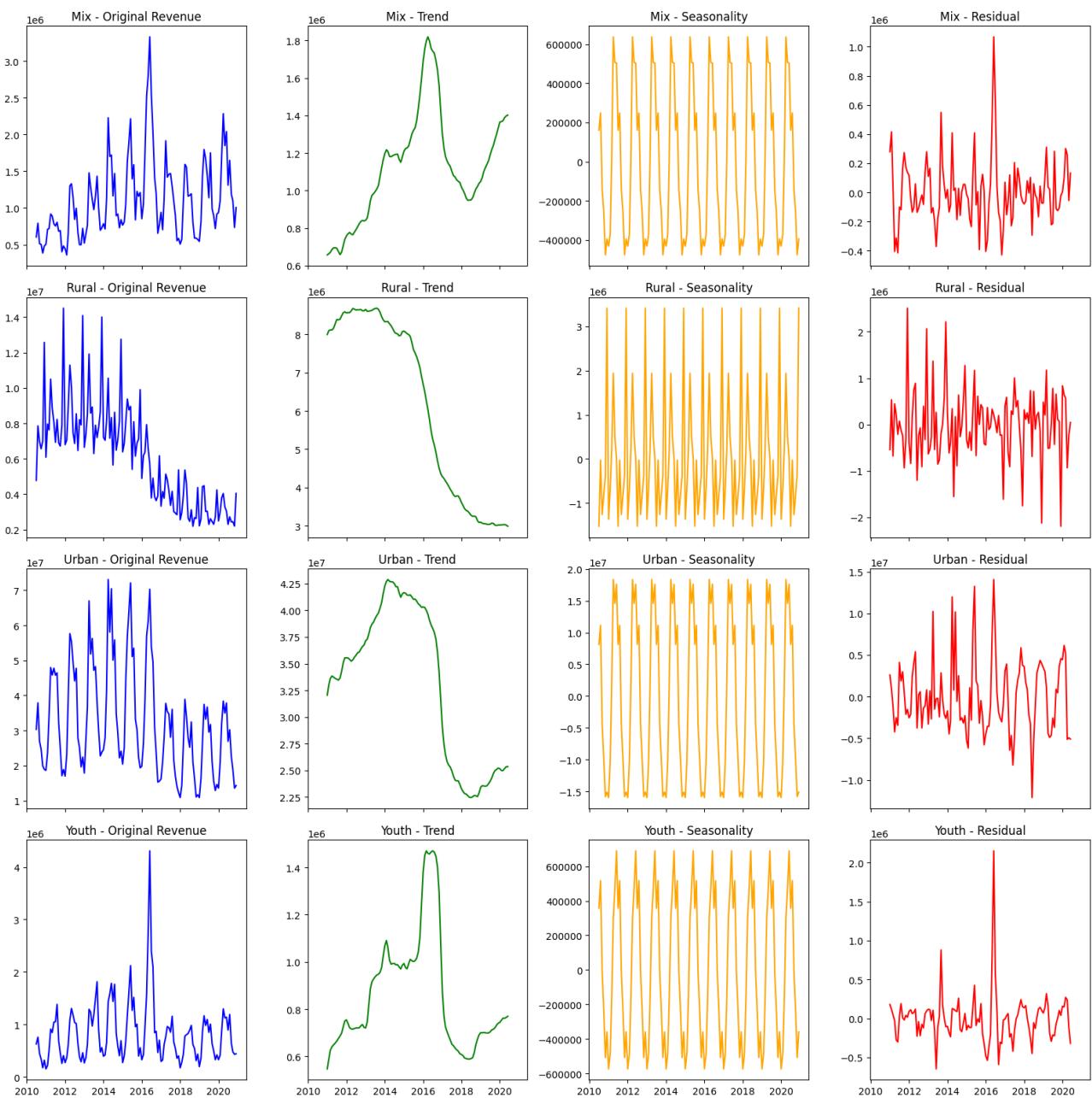
Hình 48: Biểu đồ mối quan hệ giữa doanh số của các danh mục sản phẩm theo thời gian



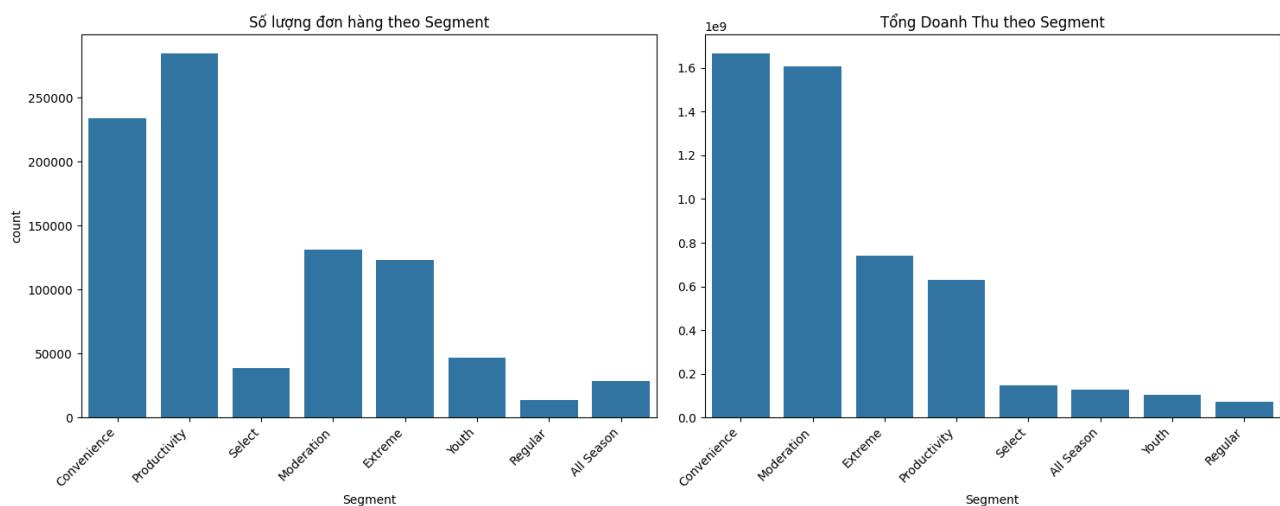
Hình 49: Biểu đồ mối quan hệ giữa doanh thu của các danh mục sản phẩm theo thời gian



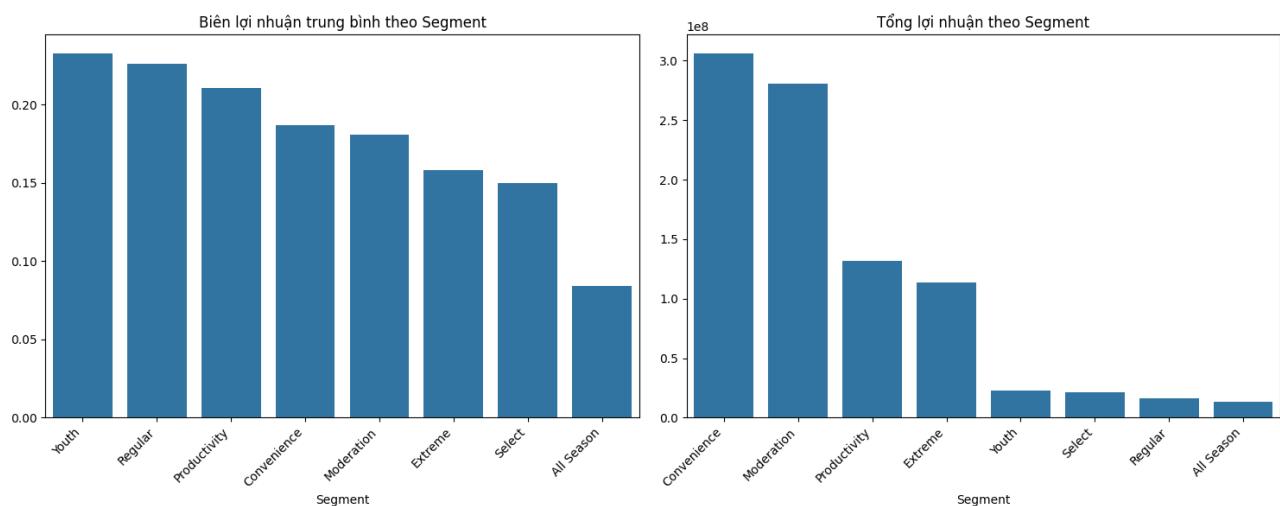
Hình 50: Biểu đồ xu hướng và tính mùa vụ của doanh thu và doanh số theo thời gian



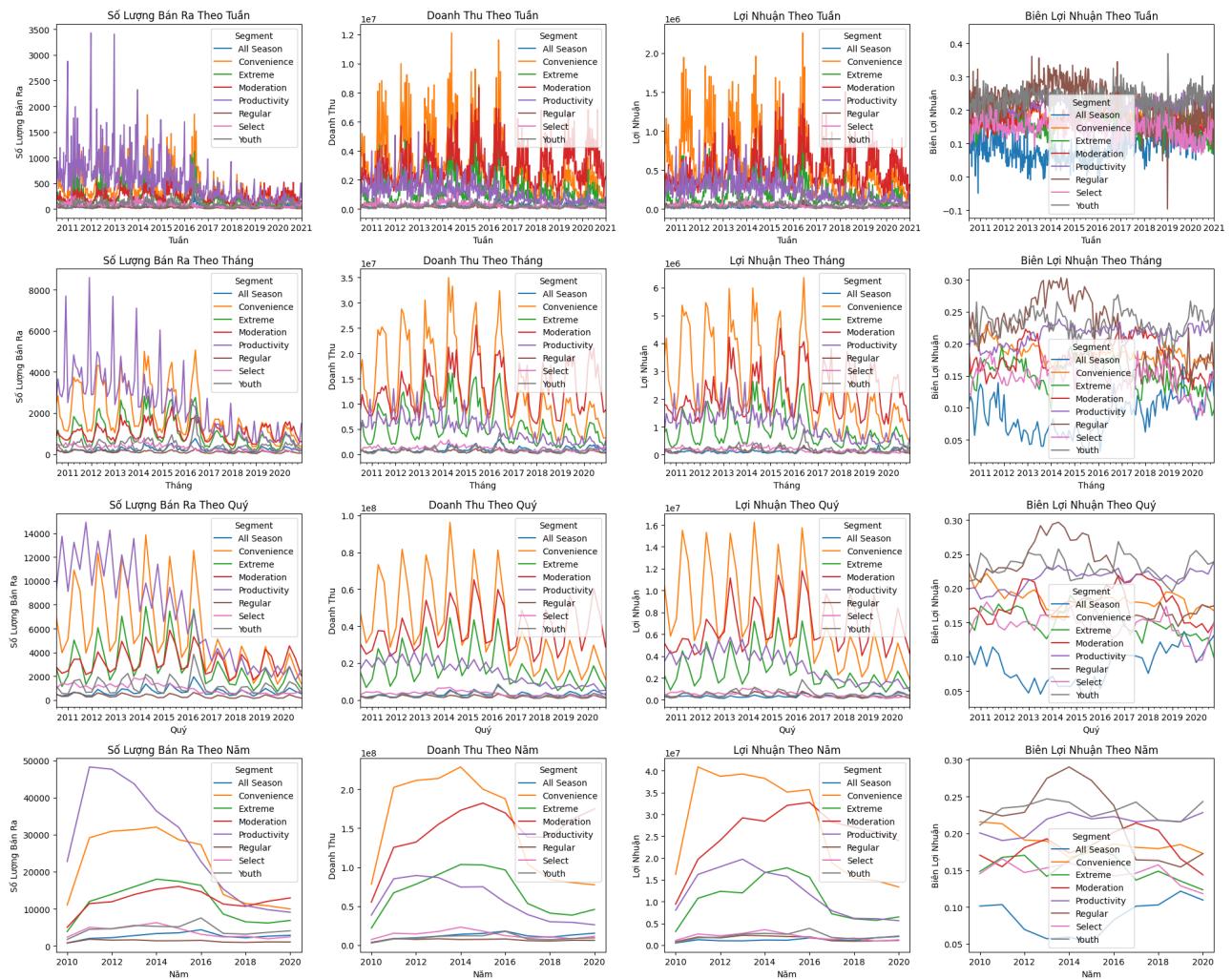
Hình 51: Biểu đồ số lượng đơn hàng và tổng doanh thu của từng phân khúc khách hàng



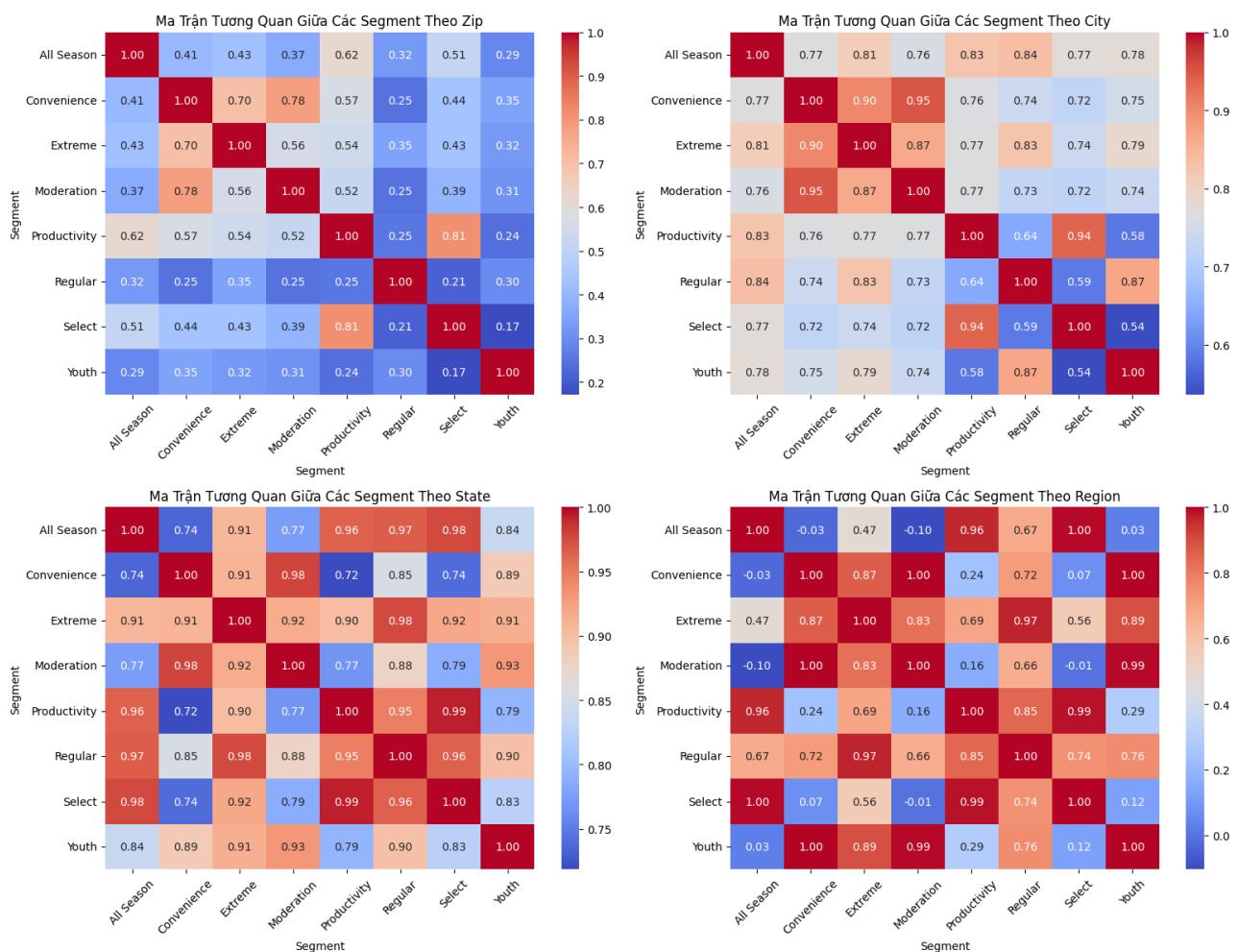
Hình 52: Biểu đồ mối quan hệ giữa biên lợi nhuận và tổng lợi nhuận theo phân khúc khách hàng



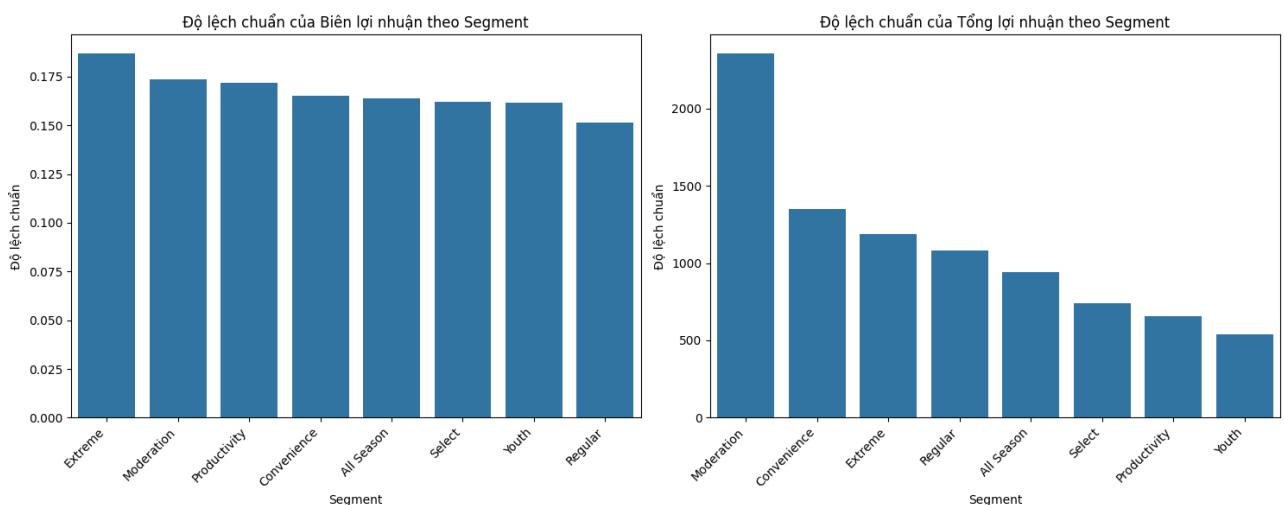
Hình 53: Biểu đồ biến động doanh số, doanh thu, lợi nhuận và biên lợi nhuận của từng phân khúc theo thời gian



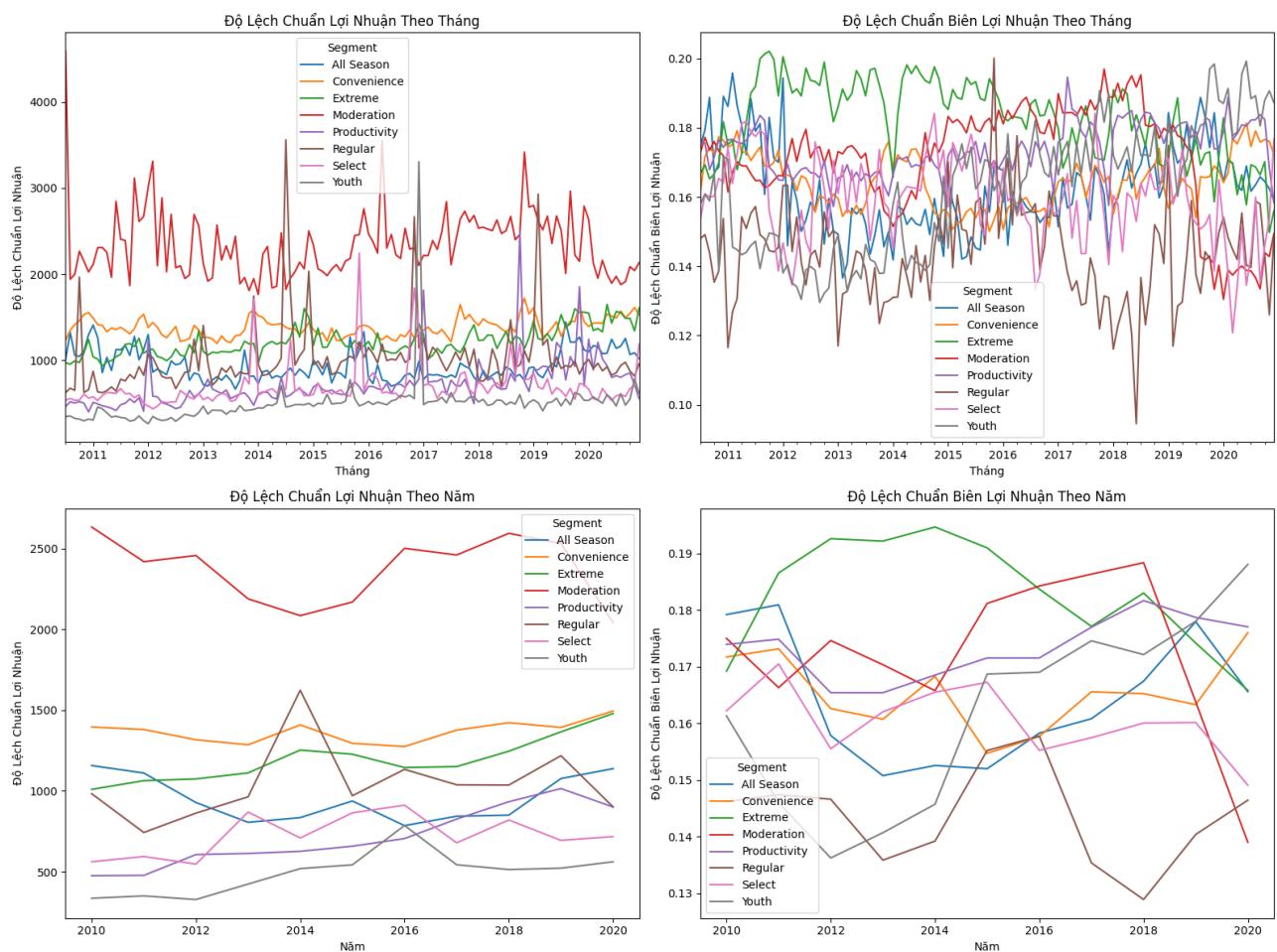
Hình 54: Biểu đồ mối quan hệ giữa các phân khúc khách hàng theo khu vực, thành phố và mã vùng



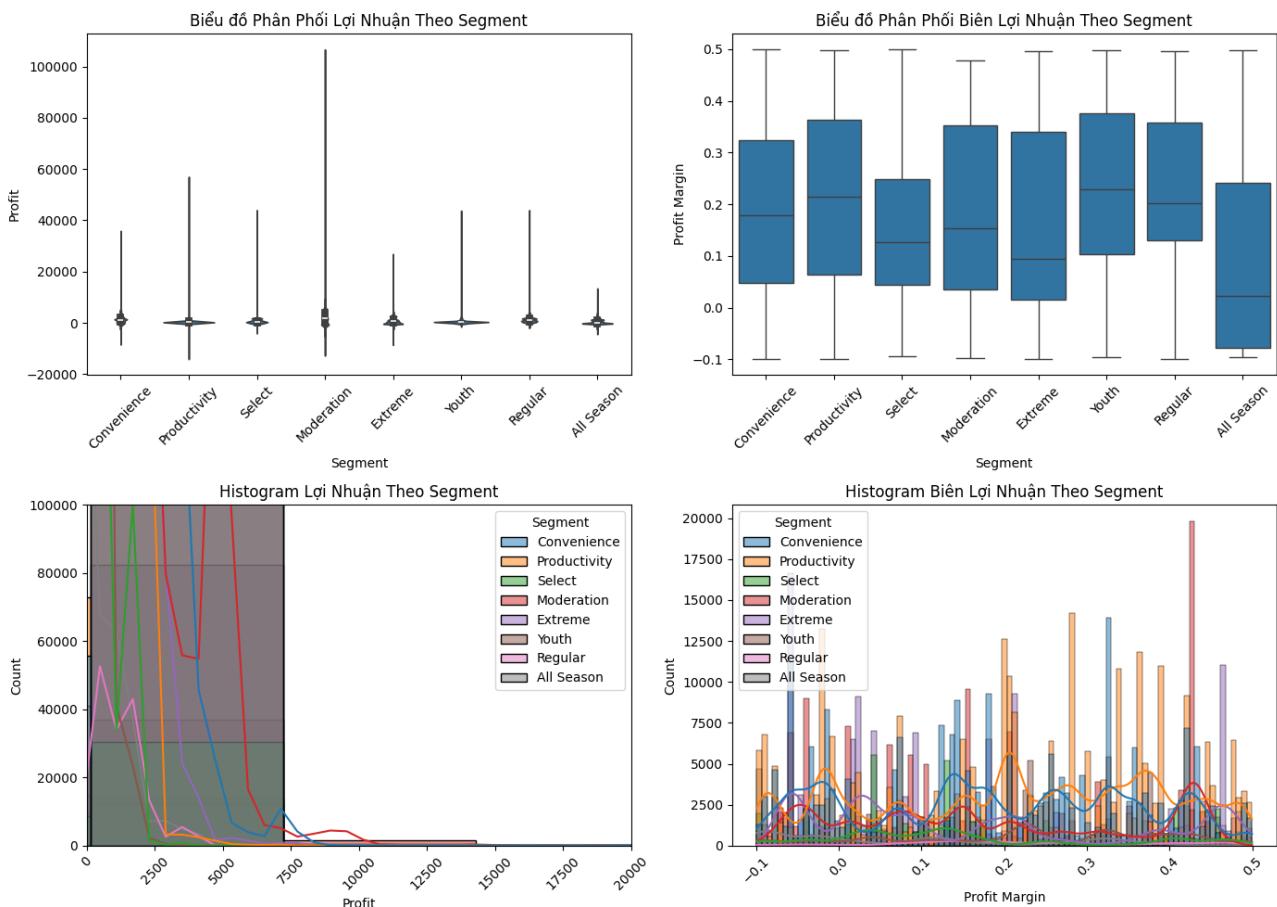
Hình 55: Biểu đồ độ lệch chuẩn của biên lợi nhuận và lợi nhuận theo phân khúc khách hàng



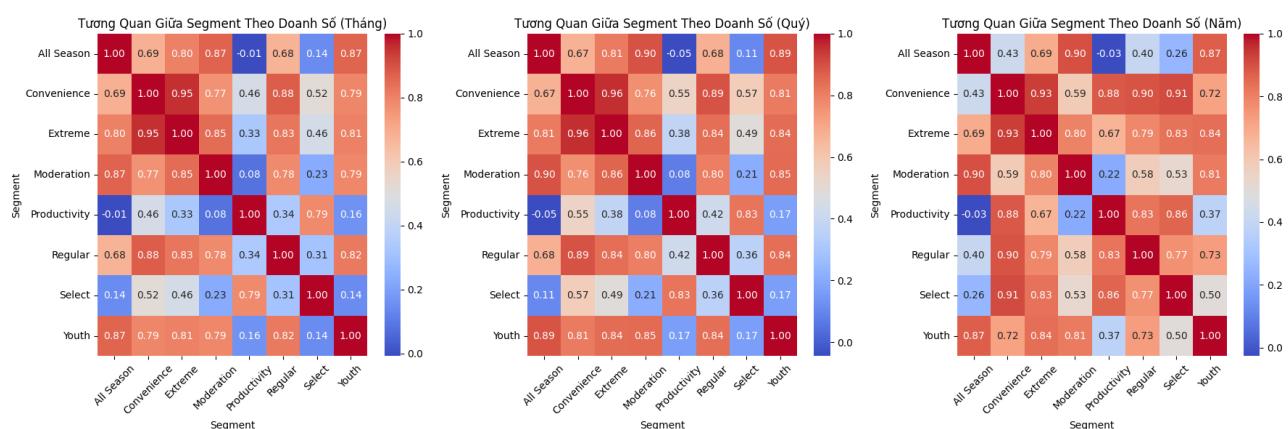
Hình 56: Biểu đồ biến động độ lệch chuẩn của biên lợi nhuận và lợi nhuận theo thời gian



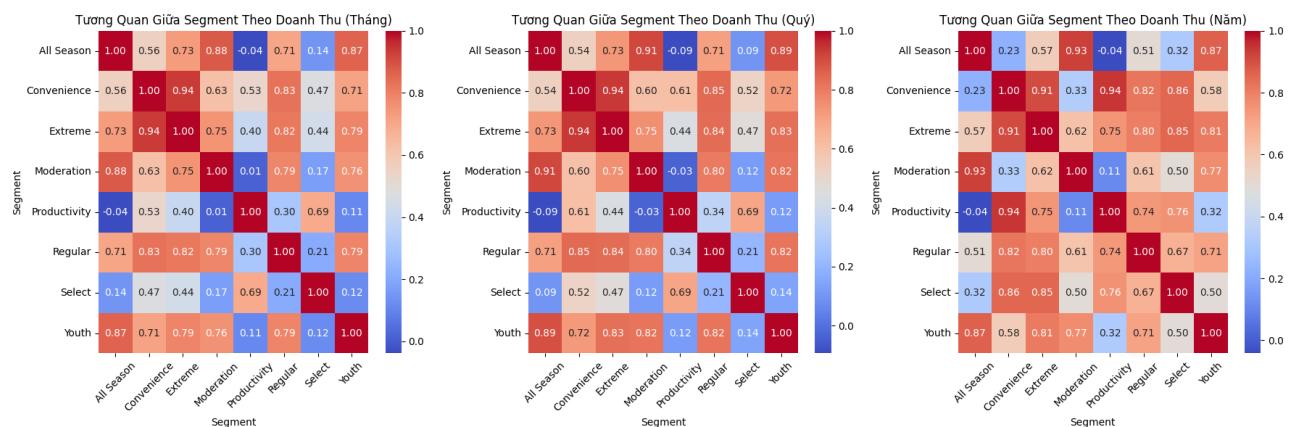
Hình 57: Biểu đồ phân phối và histogram của biên lợi nhuận và lợi nhuận theo phân khúc khách hàng



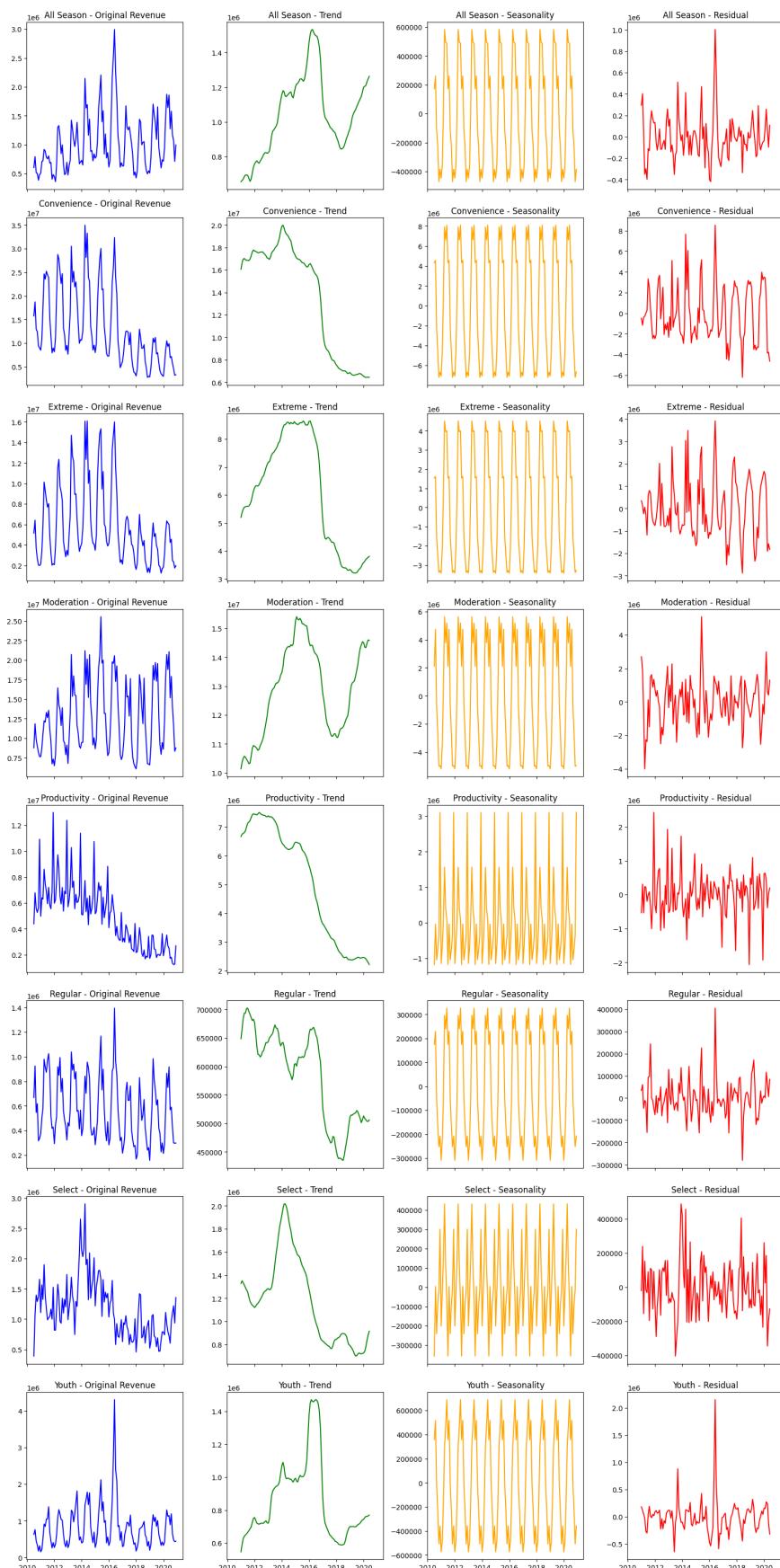
Hình 58: Biểu đồ mối quan hệ giữa doanh số của các phân khúc khách hàng theo thời gian



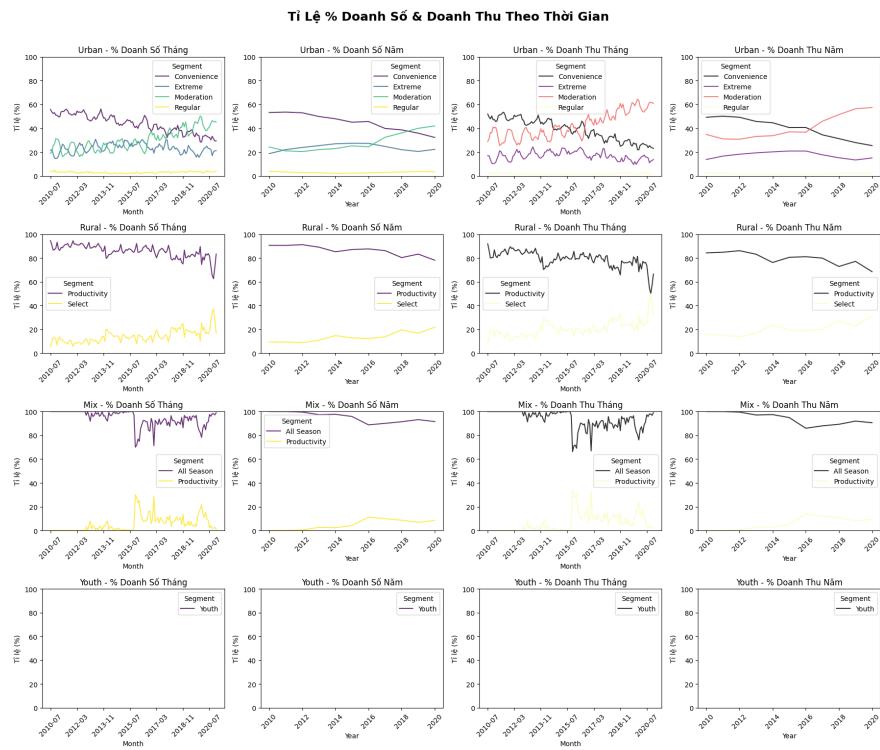
Hình 59: Biểu đồ mối quan hệ giữa doanh thu của các phân khúc khách hàng theo thời gian



Hình 60: Biểu đồ xu hướng và tính mùa vụ của doanh thu và doanh số theo thời gian



Hình 61: Biểu đồ tỉ lệ phần trăm doanh số và doanh thu theo thời gian



Hình 62: Biểu đồ phương sai phần trăm doanh số và doanh thu theo thời gian

