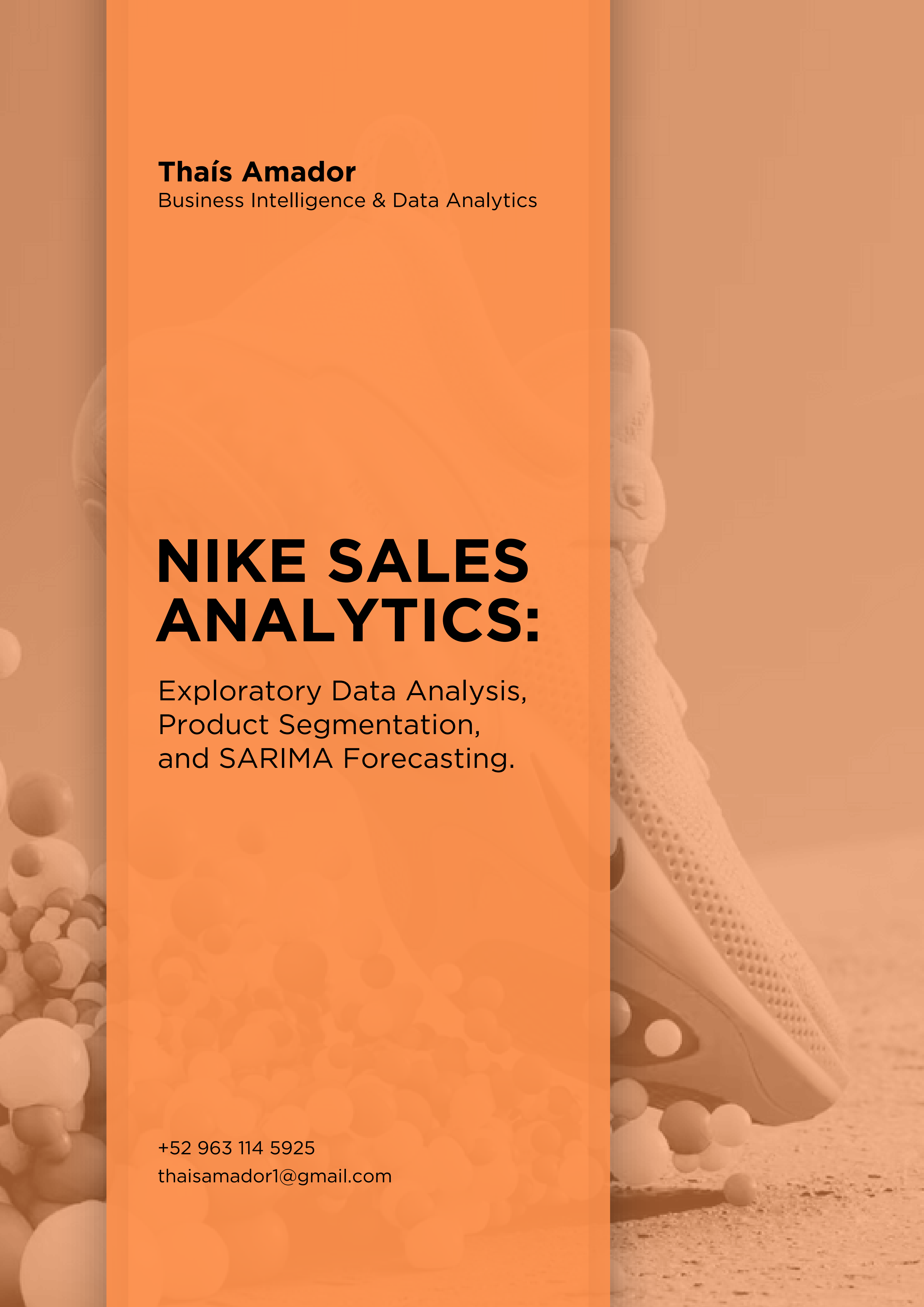


Thaís Amador
Business Intelligence & Data Analytics

NIKE SALES ANALYTICS:

Exploratory Data Analysis,
Product Segmentation,
and SARIMA Forecasting.

+52 963 114 5925
thaisamador1@gmail.com

Introduction

Context

Nike is one of the world’s leading sportswear and footwear brands, operating in a highly competitive market where sales performance is influenced by factors such as product category, seasonality, marketing campaigns, and regional demand.

This project focuses on analyzing Nike’s historical sales data to uncover key trends, segment products based on performance, and forecast revenue using time series modeling techniques.

Dataset Description

The dataset used comes from Kaggle – Nike Sales Uncleaned Dataset, containing transactional-level information on Nike sales, including:

- Order details: order date, order ID, shipping date.
- Product information: category, sub-category, model.
- Sales metrics: units sold, unit price, total revenue, discounts, profit.
- Customer and shipping data: region, country, ship mode.

Project Objective

The main goals of this analysis are to:

- Clean and preprocess the raw sales dataset.
- Perform an exploratory data analysis (EDA) to identify sales trends, seasonality, and category performance.
- Apply K-Means clustering to segment products based on sales and profitability metrics.
- Build a SARIMA time series model to forecast monthly revenue.
- Develop an interactive Tableau dashboard for visualization and decision-making.

Tools and Technical Approach

All data processing, cleaning, and analysis were performed in Python. The workflow included:

- Data cleaning: handling missing values, correcting data types, standardizing categories.
- EDA: identifying trends, correlations, and outliers.
- Clustering: applying K-Means to create product performance segments.
- Time series forecasting: building and validating a SARIMA model for revenue prediction.
- Visualization: designing an interactive dashboard in Tableau.

Methodology

- Imported and processed the Nike Sales dataset in Python using Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and Statsmodels.
- Cleaned and transformed the data: handled missing values, fixed data types, standardized product/category fields, and created calculated metrics (Revenue = Unit Price × Units, Profit, Margin).
- Conducted EDA to identify seasonality, top-performing categories/models, channel/region differences, and discount–profit relationships.
- Built K-Means to segment products by performance (revenue, units, margin, return rate if available).
- Modeled monthly revenue with SARIMA (train/validation split, diagnostics, forecast).
- Built an interactive Tableau dashboard with KPIs, trends, geography, and cluster filters.

Python Analysis and Results

Data Exploration

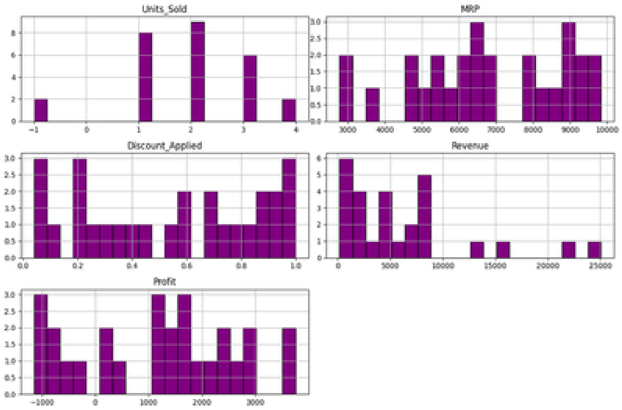
The dataset was loaded into Python using Pandas and explored to understand its structure, completeness, and key variables.

Key Findings:

- The dataset contained 2,500 records and 13 columns.
- Notable missingness before cleaning: Discount_Applied ~66.7%, MRP ~50.2%, Units_Sold ~49.4%, Order_Date ~24.6%, Size ~20.4%.
- Order_Date was standardized to datetime to enable time-based analysis.
- Calculated additional metrics such as Revenue and Profit Margin to support analysis.
- Detected data-quality issues to address (negative Units_Sold values present).

Exploratory Data Analysis (EDA)

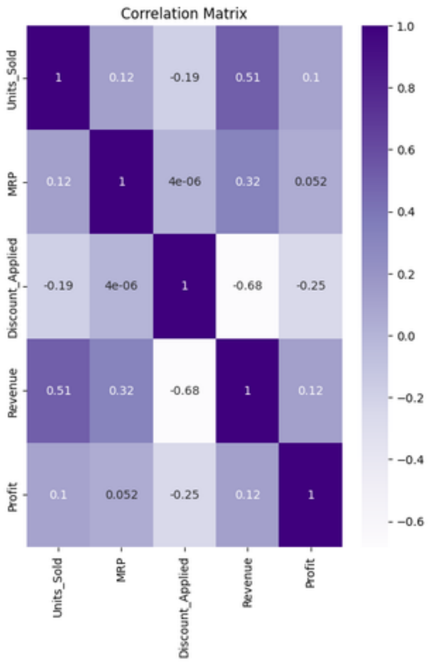
Distribution of Numerical Features



- Units_Sold: Most values are concentrated around 1 to 3 units.
- MRP: Prices are spread between 3,000 and 10,000, with clusters around 6,000-7,000 and 9,000.
- Discount_Applied: Discounts vary significantly, with peaks near 0, 0.2 and 1.0.
- Profit: Profit values range from highly negative to moderately high positive.

Correlation Matrix

- Units_Sold and Revenue show a moderate positive correlation of 0.51, since selling more units generally increases revenue.
- Discount_Applied has a strong negative correlation with Revenue of -0.68, indicating that higher discounts significantly reduce overall revenue.
- MRP is weakly correlated with Revenue, 0.32, and Units_Sold, 0.12, suggesting that price alone does not strongly drive either.
- Profit has weak correlations with all variables, including Discount_Applied with -0.25, implying that other unobserved factors may influence profit variability.

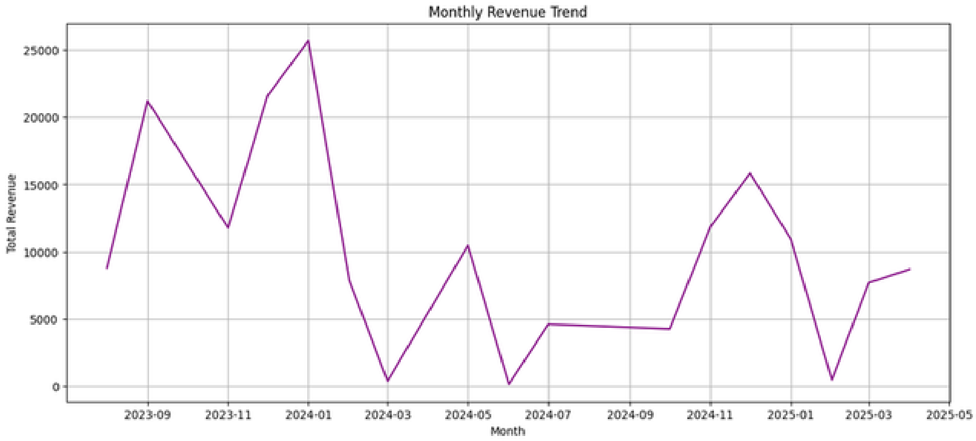


Log Transformation & Differencing

Revenue showed a clear upward trend and seasonal spikes. SARIMA requires a stationary series.

Initial Augmented Dickey–Fuller (ADF) test:

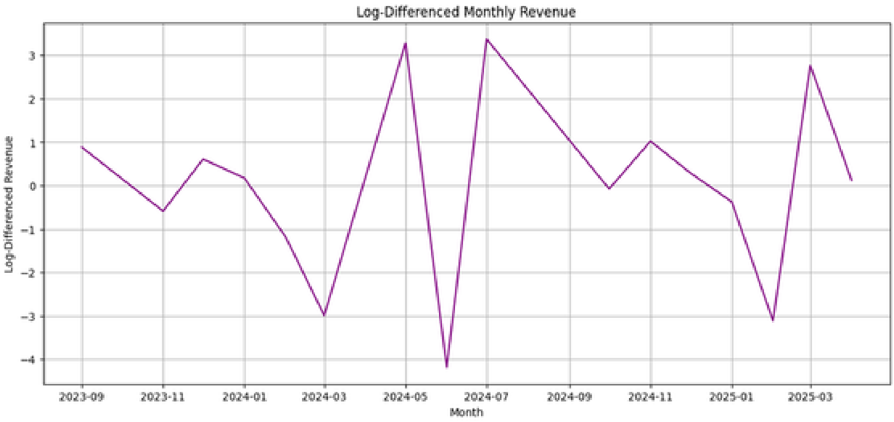
- Original series: p-value = 0.078 → non-stationary.



Transformations applied:

- Log Transformation: Reduced variance, p-value (0.003) < 0.05, still with trend.
- Differencing: Alone did not achieve stationarity, p-value (0.152) > 0.05.
- Log + Differencing (Log-Diff): p-value 1.177-10) < 0.05 → achieved strong stationarity.

The Log-Diff transformation stabilized both mean and variance, preparing the data for SARIMA modeling.



SARIMA Forecasting

A Seasonal ARIMA (SARIMA) model was applied to the log-transformed and differenced monthly revenue series, using parameters (1, 1, 1) × (1, 0, 1, 12).

Model Summary:

- AR(1): Coefficient of 0.0834, p = 0.939, indicates that short-term autoregressive effects are weak and not statistically significant.
- MA(1): Coefficient of -0.9083, p = 0.701, suggests a negative moving average component, but also not statistically significant.
- Seasonal AR(12): Coefficient of 0.2181, p ≈ 1.000, implies minimal yearly autoregressive impact.
- Seasonal MA(12): Coefficient of 0.3217, p ≈ 1.000, indicates a weak yearly moving average effect.
- σ^2: The estimated variance of the residuals is 1.9369.

Diagnostic Tests:

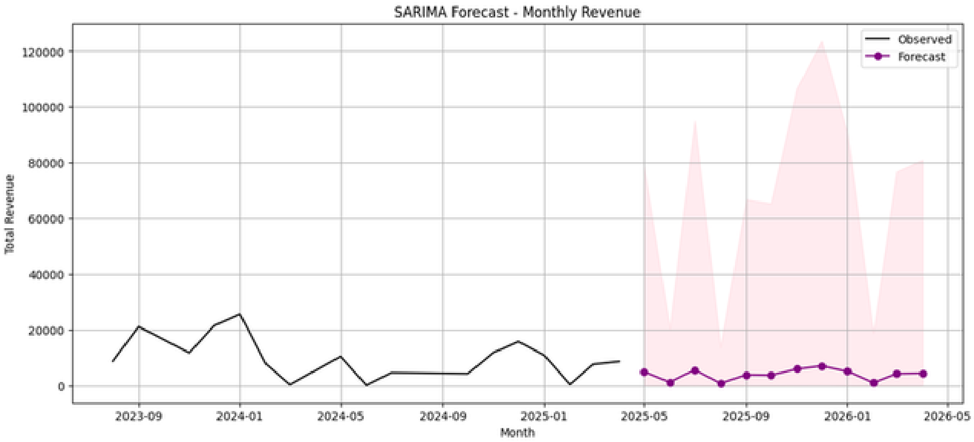
- Ljung-Box Q-Test, p = 0.48: No significant autocorrelation remains in the residuals, meaning the model captures most of the temporal structure.
- Jarque-Bera, p = 0.08: Residuals are close to a normal distribution.
- Heteroskedasticity Test, p = 0.05: Slight signs of variance instability over time.
- Skew = -1.36: Residuals are moderately left-skewed.
- Kurtosis = 3.30: Residual distribution is close to mesokurtic.

```
                                   SARIMAX Results
==========================================================================================
                 Dep. Variable:                  Revenue   No. Observations:          17
                Model:           SARIMAX(1, 1, 1)x(1, 0, 1, 12)   Log Likelihood          -30.274
                  Date:                   Sat, 02 Aug 2025   AIC                   70.548
                     Time:                        15:36:01   BIC                74.411
                  Sample:                               0   HQIC               70.745
                                                     - 17
                    Covariance Type:                        opg
=================================================================================
                          coef    std err        z      P>|z|      [0.025      0.975]
               -------------------------------------------------------------------
                 ar.L1        0.0834      1.088      0.077      0.939      -2.050      2.217
                ma.L1        -0.9083      2.369     -0.383      0.701      -5.552      3.735
             ar.S.L12      0.2181   1.16e+06   1.88e-07      1.000   -2.28e+06   2.28e+06
             ma.S.L12       0.3217    1.6e+06   2.01e-07      1.000   -3.14e+06   3.14e+06
            sigma2         1.9369   4.94e+05   3.92e-06      1.000   -9.69e+05   9.69e+05
==================================================================================
                 Ljung-Box (L1) (Q):                0.51   Jarque-Bera (JB):              5.03
                    Prob(Q):                        0.48   Prob(JB):                  0.08
                  Heteroskedasticity (H):                7.21   Skew:                     -1.36
                 Prob(H) (two-sided):               0.05   Kurtosis:                 3.30
==================================================================================

                                        Warnings:
          [1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Despite individual coefficients not being statistically significant, likely due to the limited amount of historical data available, the SARIMA model effectively removes autocorrelation from the residuals and captures the underlying seasonal patterns in monthly revenue. The model is adequate for forecasting purposes, though future refinements and the inclusion of a longer time series could improve coefficient significance and variance stability.

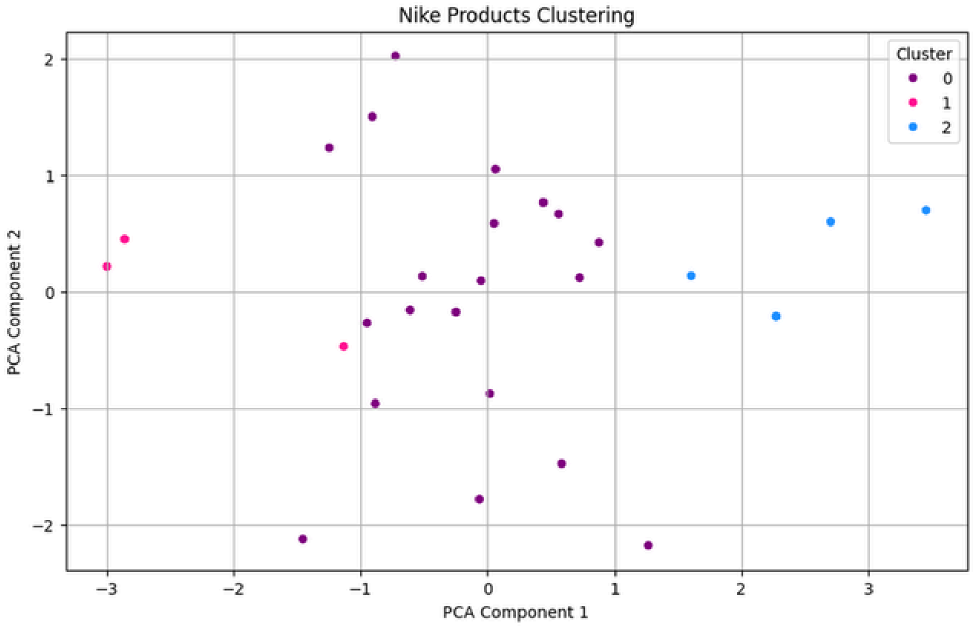
Forecasting Results



The plot shows observed revenue in black and forecasted values in purple, with the shaded area representing the 95% confidence interval.

Clustering Analysis

K-Means clustering was applied to segment Nike products based on their sales and pricing characteristics. Principal Component Analysis (PCA) was used to reduce dimensionality for visualization.



Key Findings:

- **Cluster 0 (Purple):** Products with high units sold, moderate price (MRP), moderate discounts, and solid profitability.
  - Avg. Units Sold: 2.00
  - Avg. MRP: ₹6,691.06
  - Avg. Discount: 58%
  - Avg. Revenue: ₹4,662.70
  - Avg. Profit: ₹1,387.12

- **Cluster 1 (Pink):** Low-performing products with high discounts and negative average profit.
  - Avg. Units Sold: -0.33
  - Avg. MRP: ₹5,279.29
  - Avg. Discount: 82%
  - Avg. Revenue: ₹1,066.68
  - Avg. Profit: -₹840.03

- **Cluster 2 (Blue)**: Premium-priced products with the highest revenue and profit per sale, and lower discounts.
  - Avg. Units Sold: 2.75
  - Avg. MRP: ₹8,350.22
  - Avg. Discount: 16%
  - Avg. Revenue: ₹18,947.38
  - Avg. Profit: ₹1,597.92

This segmentation allows Nike to target different marketing and pricing strategies for each cluster:

- Cluster 0: Maintain balanced discounting and inventory.
- Cluster 1: Evaluate product viability or adjust pricing/discount strategy.
- Cluster 2: Focus on premium branding and maintaining quality perception.

Dashboard Overview

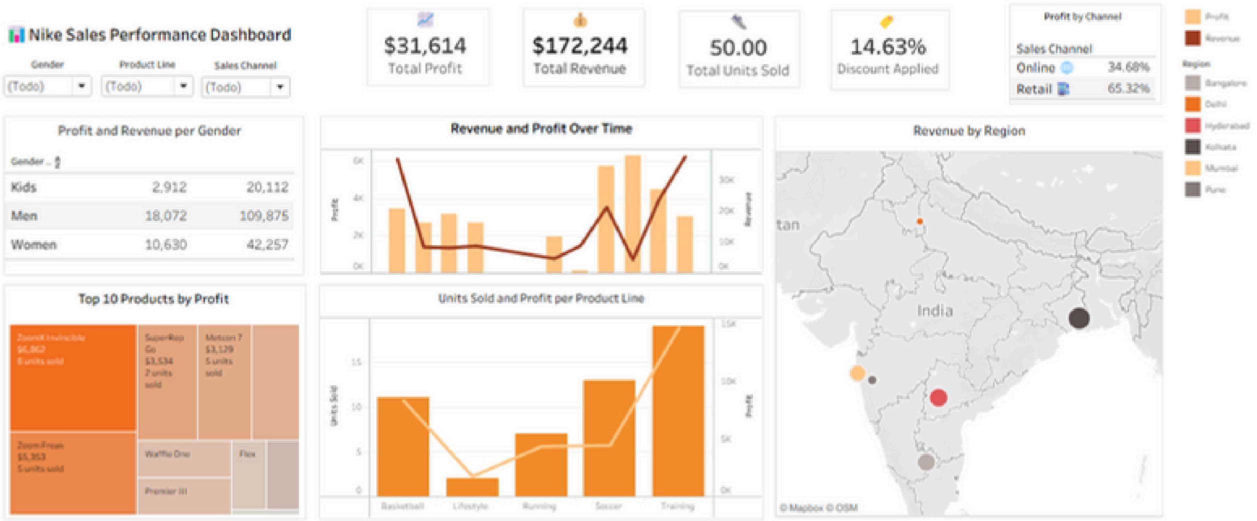
The Nike Sales Performance Dashboard was developed in Tableau to provide an interactive view of key business metrics and insights.

Dashboard Link: Nike Sales Performance Dashboard

Main KPIs

- **Total Profit:** ₹31,614
- **Total Revenue:** ₹172,244
- **Total Units Sold:** 50
- **Average Discount Applied:** 14.63%
- **Profit by Channel:** Retail (65.32%) dominates over Online (34.68%).

Nike Sales Performance Dashboard



Nike Sales Performance Dashboard

Gender: (Todo)
Product Line: (Todo)
Sales Channel: (Todo)

$31,614 Total Profit

$172,244 Total Revenue

50.00 Total Units Sold

14.63% Discount Applied

Profit by Channel

Sales Channel	
Online	34.68%
Retail	65.32%

Profit and Revenue per Gender

Gender		
Kids	2,912	20,112
Men	18,072	109,875
Women	10,630	42,257

- Profit and Revenue by Gender: Men generate the highest revenue ₹109,875 and profit ₹18,072.
- Kids have the lowest revenue share.
- Revenue and Profit Over Time: Profit and revenue fluctuate across months, with certain peaks linked to specific promotions or product launches.
- Top 10 Products by Profit: Premium shoes like ZoomX Invincible and SuperRep Go lead in profitability despite limited units sold.
- Units Sold and Profit per Product Line: Training products have the highest units sold and profit, while Lifestyle products underperform.
- Revenue by Region: Revenue concentration is high in key cities such as Mumbai, Delhi, and Bangalore.

Conclusions & Recommendations

Conclusions

The analysis of the Nike sales dataset revealed clear patterns in product performance, customer preferences, and revenue distribution. Men’s products dominate in both revenue and profit, with high-value premium footwear driving overall profitability despite relatively low sales volumes. Training products stand out as the top-performing product line in terms of both units sold and profit. Regional sales are concentrated in major metropolitan areas such as Mumbai, Delhi, and Bangalore, highlighting strong brand presence in urban markets.

The SARIMA model successfully identified seasonal patterns in monthly revenue despite data limitations, providing a foundation for future forecasting. Clustering analysis segmented products into distinct performance groups, enabling more targeted marketing and inventory strategies.

Recommendations

- Focus on High-Profit Product Lines – Allocate more marketing and promotional resources to premium footwear and training products, which yield higher margins.
- Expand Presence in Underperforming Regions – Develop regional marketing campaigns to increase sales in low-revenue cities while maintaining strong presence in major urban hubs.
- Leverage Seasonal Forecasting – Use SARIMA-based forecasts to anticipate demand fluctuations, plan promotions, and optimize inventory levels during peak seasons.
- Product Portfolio Optimization – Consider reducing focus on low-performing lifestyle products and redirecting resources to categories with higher sales and profitability potential.
- Channel Strategy Enhancement – Since retail channels contribute significantly to revenue, enhance in-store experience while expanding online presence to capture emerging digital shoppers.