ACENET
Microcredential in Advanced Computing
ISP Report

Project title: Sales Forecasting and Trend Analysis - a data-driven approach with a retail store dataset

Participant name: Jordan Pardy

Date: July 30, 2024

Abstract:

The project analyzes a retail superstore dataset using Python libraries to identify key sales trends and patterns. Advanced forecasting techniques, including the SARIMA machine learning model, are employed to predict future sales and provide insight into business strategy.

1. Introduction

Sales forecasting is more important than ever in increasingly competitive retail environments. Businesses are leveraging the data they have on their sales, customers, products, product categories, markets, etc. every day to make decisions on every aspect of their operations and remain competitive and responsive to their clients' needs. This project takes a fictional retail orders dataset and leverages data analysis techniques, high-performance computing (HPC) and machine learning to process a large amount of data efficiently and draw useful predictions and conclusions from it.

2. Background

In today's data-driven world, having all the information you need to stay ahead of competition is critical. Sales forecasting can be used to inform budgeting and financial planning, effectively allocating limited resources, allowing for the development of targeted or specialized marketing campaigns to maximize revenue and profitability, optimizing supply-chain and ensuring customer satisfaction and loyalty by meeting and exceeding their needs.

3. Analysis

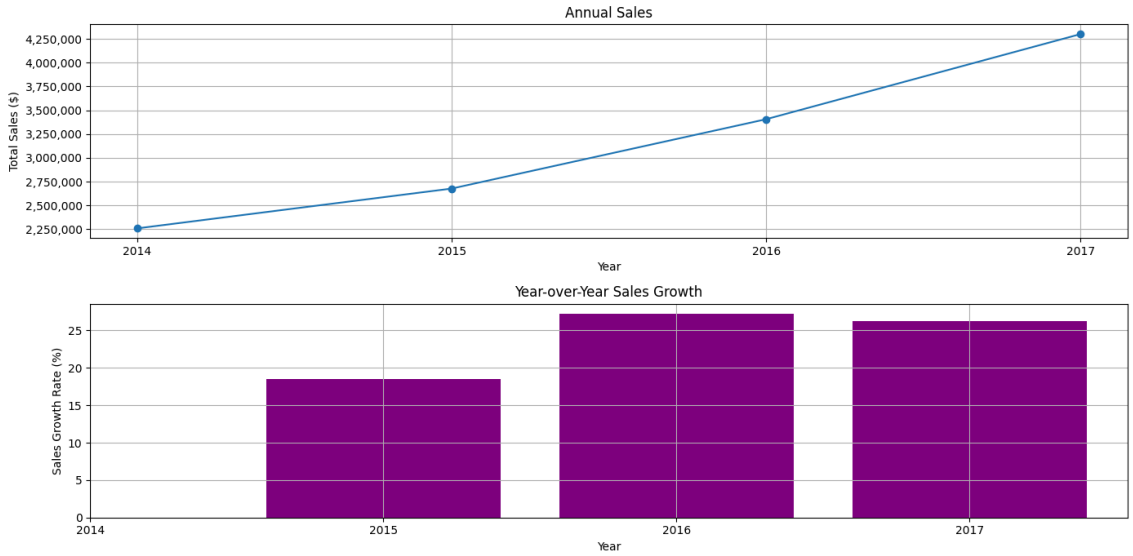
The dataset is composed of historical sales data for a retail store, and includes information such as order date, shipping date, shipping mode, customer IDs and names, segment (consumer, corporate, home office), location/region, product category, sales, quantity, discounts given, and profit. The dataset is from the Tableau website as a sample offering.

The data preparation process for this project includes loading the Excel dataset into a pandas Dataframe, inspecting the data by looking at its basic structure and data types, identifying if there are any missing values, and then performing feature engineering by creating new features for date extraction and aggregate features to sum data.

I chose the SARIMA (Seasonal AutoRegressive Integrated Moving Average) model, as it is well-suited for analysis and forecasting with time-series data. Additionally, SARIMA is designed to handle seasonal patterns in time-series data and to explore trends over time.

High-performance computing (HPC) is incorporated into the project through the use of Dask, which can help scale data analysis or machine learning workflows in larger datasets by distributing the workload across multiple cores in a cluster.

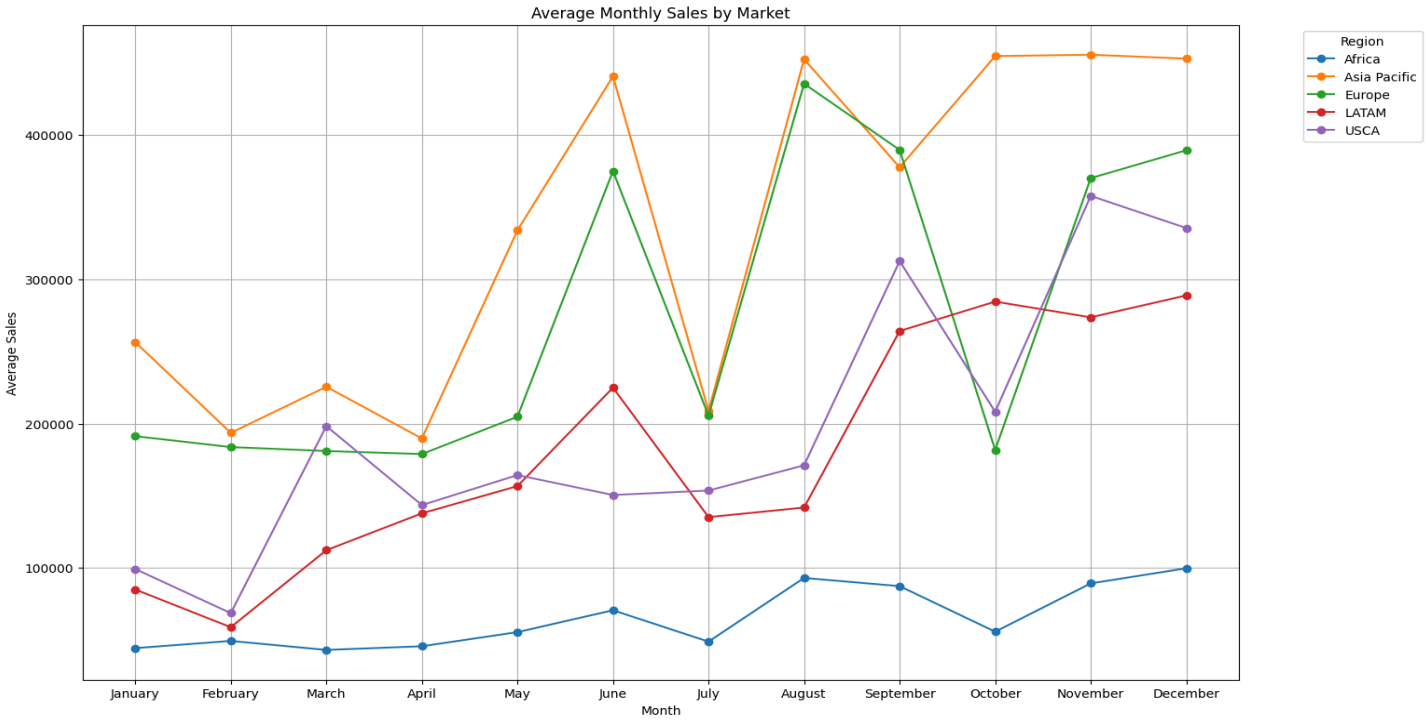
4. Results



Sales increase year-over-year in the data - 18.5% in the first year, 27.2% in the second, and 26.25% in the third, reaching approximately $4.3 million in the final year of the dataset. There is significant seasonality in the data, with peak revenue occurring in the fall for each product category:

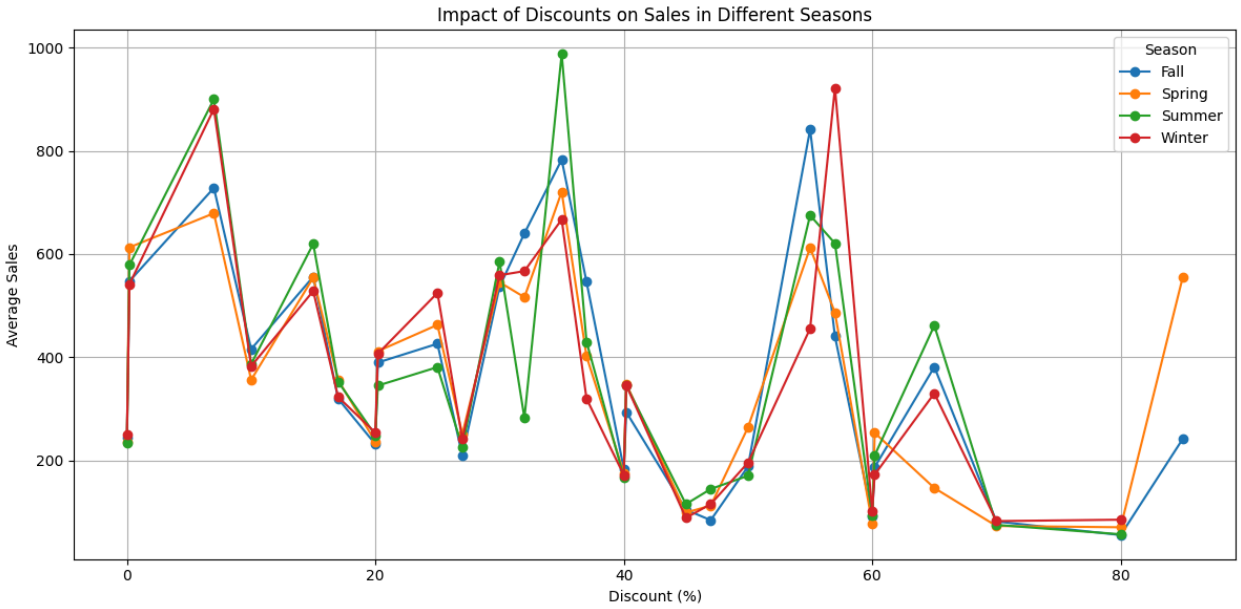
Category	Top Month	Revenue in Top Month
Furniture	September 2017	$175776.64
Office Supplies	November 2017	$163851.32
Technology	November 2017	$222579.04

Seasonal patterns across markets

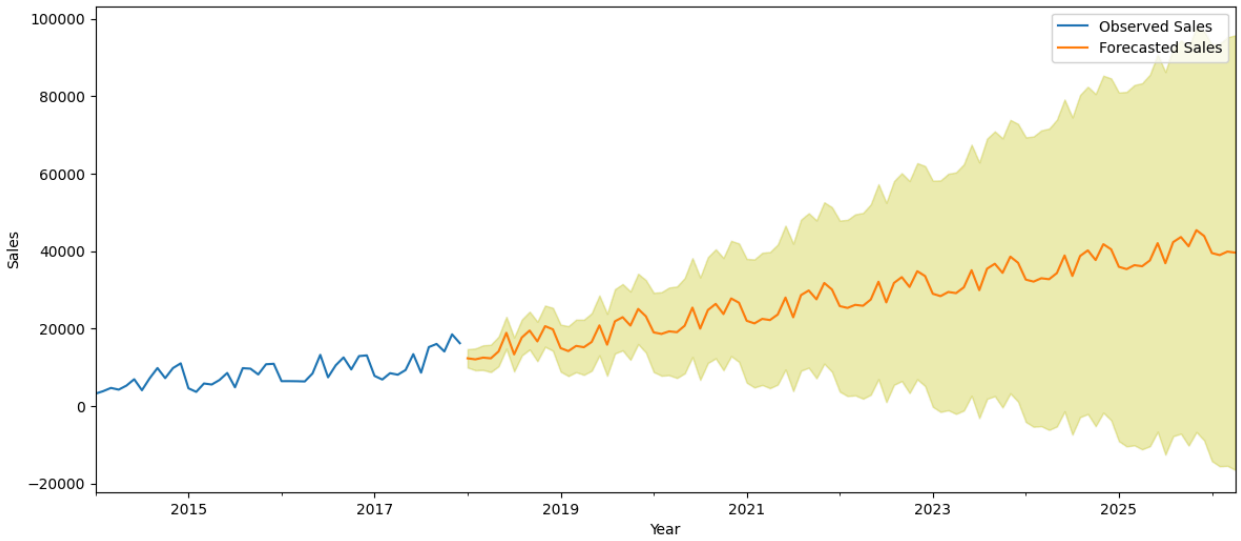


There is lower seasonality in the African market, with larger spikes in the fall and holiday seasons in American markets (LATAM and USCA). The European and Asia-Pacific markets see a significant jump in sales in the spring (from April to June), with a large dip in July, followed by another significant jump in the back-to-school season of August to September.

Impact of Discounts on Sales Revenue in Different Seasons



Offering either a modest discount of approximately 10%, or a more significant discount in the 30% range correlates with a significant jump in sales revenue across all seasons. There is a notable increase in the summer season in particular with a 35% discount. Significant discounting beyond 55% has little effect on revenue, and lowers profitability to the point where it likely would not be feasible for the business.



The SARIMA model predicts a modest upward trend in sales over time, from the end of the dataset timeframe onwards. As shown in the figure above, the confidence interval gets wider as more time passes. The seasonality shown throughout the period is predicted to continue as well.

5. Discussion

SARIMA is a good choice for time-series forecasting in the short term, with the limitation of the past predicting the future. In this particular dataset, and much of retail, seasonality and promotional strategies are critical considerations.

There is lots of opportunity to build from the basics outlined in this project and add more advanced techniques to evaluate examples based on a particular business and its needs, as outlined in the section below.

Conclusion

For businesses today, having a competitive edge is critical. Data analysis and forecasting plays a major role in staying ahead of competitors and knowing your customers fully so that you can maximize retention, revenue, and profitability. This project reiterates the significance of seasonality in retail and taking advantage of promotional and pricing strategies to maximize profitability.

Future research could break down the project into more specific areas:

- External factors such as economic indicators, market trends, weather patterns, and social media sentiment could be integrated to enhance analysis.
- A closer analysis of customer segments could be performed to identify high-value customers and tailor marketing strategies.
- Dynamic pricing models could be developed to adjust prices in real-time based on demand forecasts, competitor's pricing, or inventory levels.
- Natural language processing (NLP) could be used to extract insights from customer reviews or feedback for analysis.

References

Everything Retailers Should Know About Demand Forecasting - Shopify Canada

The Advantages of Data-Driven Decision Making - Harvard Business School

Supplementary Materials

GitHub: https://github.com/pardyj/isp

Dataset: https://github.com/pardyj/isp/blob/master/dataset.xlsx