Suraj Kumar Rai Sunder, 2026, 14:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Sales Forecasting and Business Intelligence Dashboard Using Data Analytics and Power BI

Suraj Kumar Rai[1]; Adarsh Tiwari[2]; Aditya Sunder[3] Mr. Sachin Chauhan[4]
[1,2,3,]Dept. Of Computer Science & Engineering Hmr Institute Of Technology & Management Delhi, India;
[4]**Faculty** Supervisor, Dept. Of Computer Science & Engineering
Hmr Institute Of Technology & Management
Delhi, India

Abstract—Accurate sales forecasting supports inventory planning, budgeting and strategic decision-making in retail. This paper presents an end-to-end data analytics pipeline that combines data cleaning, exploratory data analysis (EDA), SQL-based KPI computation, time-series analysis, forecasting models and an interactive Power BI dashboard. The study uses the Sample Superstore dataset (9,994 transactions, January 2014 to December 2017), aggregated into a 48-point monthly sales series. Four models are compared on a 12-month hold-out set: Moving Average, Linear Regression, ARIMA and Seasonal ARIMA (SARIMA). SARIMA(0,1,1)x(0,1,1,12) performs best with MAE = 13,006.71, RMSE = 15,495.15 and MAPE = 25.73%, improving on the moving-average baseline by 40.27 percentage points of MAPE. The final model, refit on all data, forecasts six months ahead (January–June 2018) with 95% confidence intervals, and the results are integrated into a Power BI dashboard with KPI cards, trend charts, regional maps, slicers and a forecast page. The findings show that modelling seasonality is the key factor for forecasting this retail series and that a BI layer makes the forecast directly usable by business stakeholders.

Index Terms—sales forecasting, SARIMA, ARIMA, time series, Power BI, business intelligence, data analytics, SQL.

I. INTRODUCTION

Sales forecasting is a core activity in retail and supply-chain management. Over-estimating demand leads to excess inventory and tied-up cash, while under-estimating it causes stock-outs and lost revenue. Accurate projections are especially valuable for small and medium enterprises because they simplify short-term demand and supply planning [1]. Forecasts also form the foundation for financial management and help decision-makers decide where to invest based on expected sales trends.

A time series is a sequence of observations recorded at regular intervals, and time-series analysis aims to identify underlying patterns and forecast future values from history [2]. Classical models such as ARIMA are interpretable and effective when series show stable, linear dynamics [3], while machine learning and deep learning models can capture nonlinear behaviour [4]. However, a forecast is only useful if decision-makers can see and interact with it, which is where Business Intelligence (BI) tools such as Power BI become important.

This project therefore has two connected goals: (i) build and compare forecasting models for retail sales, and (ii) deliver the results through an interactive BI dashboard. The main contributions are:

- A documented data-cleaning and EDA workflow on a retail transaction dataset.
- SQL-based computation of business KPIs, cross-checked against Python results.
- A comparison of four forecasting models using MAE, RMSE and MAPE on a chronological hold-out set.
- A six-month forecast with 95% confidence intervals for business planning.
- A Power BI dashboard with DAX measures integrating historical analysis and the forecast.

The rest of the paper is organized as follows: Section II reviews related work, Section III describes the methodology, Section IV presents results, Section V discusses findings and limitations, and Section VI concludes.

© 2026 Suraj Kumar Rai, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

II. LITERATURE REVIEW

Traditional sales forecasting relies on statistical time-series models, particularly the ARIMA framework, which captures temporal dependence, trend and autocorrelation for short- to medium-term forecasting. Hyndman and Athanasopoulos [3] note that ARIMA performs well when sales series have stable patterns and linear dynamics. Fattah et al. [5] reported strong accuracy for ARIMA in food-demand prediction, Dar et al. [6] showed reliable ARIMA performance on volatile financial data using the Box–Jenkins methodology, and Nithya et al. [7] applied ARIMA to sales data.

ARIMA has limits when data show nonlinear relationships, complex seasonality or structural change, so machine learning and deep learning models have gained prominence. Siami-Namini and Namin [8] reported that LSTM models outperformed ARIMA on financial time series, and Bandara et al. [9] used LSTM networks for e-commerce demand forecasting. Taylor and Letham [10] proposed an additive open-source framework (Prophet) for forecasting at scale.

A recent comparative study [11] evaluated ARIMAX against Random Forest, XGBoost and SVR for weekly retail sales in the Indian market, using TV, Radio and Online advertising spend as exogenous variables. ARIMAX achieved the lowest errors (MAE = 14.20, RMSE = 16.60, MAPE = 11.56%), SVR

was comparable (MAPE = 11.66%), and the tree ensembles had higher errors. A Diebold–Mariano test found no significant difference between ARIMAX and SVR ($p = 0.3734$). The authors concluded that no single model is universally best and recommended benchmarking classical against ML approaches; their stated limitations include restricted predictors, static exogenous inputs and a single train–test split.

Research gap.

Existing studies concentrate on model accuracy and often depend on exogenous variables such as advertising spend, which are not always available. Few works connect the forecast to an interactive BI dashboard. This work

- forecasts from sales history alone using seasonal structure,
- validates KPIs independently with SQL, and (c) delivers the forecast through Power BI.

III. METHODOLOGY

The Workflow Is: Data Cleaning → Eda → Sql Kpi Analysis→ time-series analysis and feature engineering → model building → evaluation and final forecast → Power BI dashboard. The tools used were Excel, Python (Pandas, Matplotlib, Seaborn, statsmodels), SQL (SQLite) and Power BI Desktop.

1. *Dataset And Cleaning*

The Sample Superstore dataset has 9,994 rows and 21 columns, including Order Date, Ship Mode, Segment, Region, Category, Sub-Category, Product Name, Sales, Quantity, Discount and Profit. Orders span 3 January 2014 to 30 December 2017. Cleaning checks found no duplicate rows, no missing values and no unparseable dates, so no rows were removed. Data types and date/category formats were standardized, and total sales before and after cleaning were identical (2,297,200.86), confirming that no information was lost.

2. *Exploratory Data Analysis*

Summary statistics were computed for Sales, Quantity, Discount and Profit, followed by breakdowns by category, sub-category, region and segment, and correlation analysis. Charts were produced with Matplotlib and Seaborn.

3. *Sql Kpi Analysis*

The cleaned data was loaded into a SQLite database with proper data types and keys, and the row count (9,994) was verified against the cleaned file. Queries using GROUP BY, JOIN and aggregate functions computed total sales, total profit, top products, regional sales and monthly revenue. Python was connected to the database to fetch results into Pandas.

4. *Time-Series Preparation*

Sales were aggregated into daily, weekly and monthly series; the monthly series (48 points) was used for modelling. The series was decomposed into trend, seasonal and residual components. Stationarity was tested with the Augmented Dickey–Fuller (ADF) test,

with first differencing applied where required. Month, quarter, year, lag and rolling-mean features were engineered for the regression model.

5. *Forecasting Models*

The series was split chronologically: the first 36 months (2014–2016) for training and the last 12 months (2017) for testing. Four models were built:

Moving Average (k = 3): baseline using the mean of the last three observations.

- **Linear Regression:** regression on engineered time and calendar features.
- **ARIMA:** order selected by lowest AIC with ACF/PACF support, giving ARIMA(0,1,1).
- **SARIMA:** ARIMA with seasonal terms (period 12) to capture yearly seasonality; the selected model was SARIMA(0,1,1)x(0,1,1,12).

6. *Evaluation Metrics*

Models were compared on the test set using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE), and the model with the lowest MAPE was selected. It was then refit on all 48 months to produce a six-month forecast with 95% confidence intervals, and its residuals were checked with the Ljung–Box test.

7. *Dashboard Design*

The cleaned data and forecast output (exported as CSV) were imported into Power BI and prepared in Power Query. A data model with a date table and relationships was created, and DAX measures were written for total sales, profit, profit margin, growth percentage and average order value. The dashboard has four pages (Sales Overview, Product and Category, Region, and Sales Forecast) with KPI cards, line and bar charts, a map, tables, tooltips and slicers for year, region, category and segment. Dashboard values were validated against the SQL results.

IV. RESULTS AND ANALYSIS

1. Descriptive Statistics

The average transaction had sales of 229.86 (SD = 623.25), quantity of 3.79, discount of 0.16 and profit of 28.66 (Table I). Sales ranged from 0.44 to 22,638.48 and profit from −6,599.98 to 8,399.98, indicating strong outliers and some heavily loss-making orders. The median sale (54.49) is far below the mean, showing a right-skewed distribution.

TABLE I SUMMARY STATISTICS (n = 9,994)

Variable	Mean	Std	Min	Max
Sales	229.86	623.25	0.44	22,638.48
Quantity	3.79	2.23	1	14
Discount	0.16	0.21	0.00	0.80
Profit	28.66	234.26	−6,599.98	8,399.98

2. Business Insights from EDA and SQL

SQL queries returned total sales of 2,297,200.86 and total profit of 286,397.02, an overall profit margin of 12.47%, and the SQL row count matched the cleaned file. Table II summarizes performance by category, region and segment.

TABLE II
Sales And Profit By Category, Region And Segment

Group	Sales	Profit
Category		
Technology	836,154.03	145,454.95
Furniture	741,999.80	18,451.27
Office Supplies	719,047.03	122,490.80
Region		
West	725,457.82	108,418.45
East	678,781.24	91,522.78
Central	501,239.89	39,706.36
South	391,721.90	46,749.43
Segment		
Consumer	1,161,401.34	134,119.21
Corporate	706,146.37	91,979.13
Home Office	429,653.15	60,298.68

Technology leads in both sales and profit. Furniture has almost the same sales as Office Supplies but far lower profit (18,451.27 vs 122,490.80). The top five sub-categories by sales were Phones (330,007.05), Chairs (328,449.10), Storage(223,843.61), Tables (206,965.53) and Binders (203,412.73).

Three sub-categories were loss-making: Tables (−17,725.48), Bookcases (−3,472.56) and Supplies (−1,189.10); Tables are thus the fourth-largest sub-category by sales yet the largest source of loss. The West region leads in sales and profit, while Central has higher sales than South but lower profit (39,706.36 vs

46,749.43). Discount and profit are negatively correlated (−0.219), whereas sales and profit are moderately positively correlated (0.479). The best-selling products were the Canon imageCLASS 2200 Advanced Copier (61,599.82), the Fellowes PB500 binding machine (27,453.38) and the Cisco TelePresence System EX90 (22,638.48).

[Insert figure here]

Fig. 1. Category-wise and region-wise sales and profit.

4. Time-Series Characteristics

The monthly series has 48 points and a linear trend slope of per month, indicating a clear upward trend. Seasonality is strong: the peak months are December, November and September, and the lowest are February, January and August. The ADF test gave $p = 0.0002$ on the original series and $p < 0.0001$ after first differencing, so both were reported as stationary at the 5% level. Because the series shows a visible upward trend and strong seasonality, first and seasonal differencing were still retained in the ARIMA/SARIMA models, supported by ACF/PACF diagnostics.

[Insert figure here]

Fig. 2. Monthly sales and seasonal decomposition (trend, seasonal, residual).

5. *Model Comparison*

The selected orders were ARIMA(0,1,1) (AIC = 807.70) and SARIMA(0,1,1)x(0,1,1,12) (AIC = 199.14). AIC values are not

compared across these models because they use different differencing; test-set accuracy is used instead (Table III).

TABLE III FORECAST ACCURACY ON THE 12-MONTH TEST SET

Model	MAE	RMSE	MAPE (%)
Moving Average (k=3)	26,605.82	31,208.55	66.00
Linear Regression	13,939.09	17,212.50	30.59
ARIMA(0,1,1)	24,905.10	28,847.35	59.96
SARIMA(0,1,1)x(0,1,1,12)	13,006.71	15,495.15	25.73

SARIMA achieved the lowest error on all three metrics. Relative to the moving-average baseline, MAPE improved from 66.00% to 25.73%, a reduction of 40.27 percentage points. Non-seasonal ARIMA (MAPE 59.96%) barely improved on the baseline, whereas Linear Regression with calendar features (30.59%) captured much of the seasonal pattern. The gain therefore comes mainly from modelling seasonality, which ARIMA(0,1,1) cannot represent.

Fig. 3. Actual versus predicted sales for all models on the test set.

[Insert figure here]

6. *Final Forecast*

SARIMA(0,1,1)x(0,1,1,12) was refit on all 48 months to forecast January–June 2018 (Table IV). The Ljung–Box test on the residuals (lag 12) gave $p = 0.4992$, so no significant autocorrelation remains, supporting model adequacy.

TABLE IV SIX-MONTH SALES FORECAST WITH 95% CONFIDENCE INTERVAL

Month	Forecast	Lower 95%	Upper 95%
Jan 2018	52,477.99	21,855.75	83,100.23
Feb 2018	38,167.15	7,483.29	68,851.01
Mar 2018	72,866.00	42,121.96	103,610.05
Apr 2018	54,758.65	23,954.49	85,562.81
May 2018	63,273.42	32,409.22	94,137.62
Jun 2018	64,527.96	33,603.78	95,452.14

The forecast is lower than the last actual month (December 2017: 83,829.32), consistent with the seasonal pattern, since December is among the strongest months and January and February among the weakest. March 2018 is the highest forecast month (72,866.00). The intervals are wide (about

±30,600 around each forecast), reflecting the uncertainty of modelling from only four years of data.

7. *Power Bi Dashboard*

The **Sales Overview** page shows KPI cards (total sales, profit, profit margin, growth %, average order value) and a monthly trend. The **Product and Category** page exposes loss-making items such as Tables and Bookcases. The **Region** page uses a map and bar charts. The **Sales Forecast** page compares actual and predicted sales from the exported forecast. Slicers, drill-through and tooltips make the dashboard interactive, and testing confirmed that its figures match the SQL outputs.

[Insert figure here]

Fig. 4. Power BI dashboard pages (Overview, Product, Region, Forecast).

8. Discussion

The results agree with the observation in the literature that no single model is universally best and that performance depends on data characteristics [11]. In our data the dominant structure is yearly seasonality, so the seasonal model won. In the base study [11] the series was weekly without pronounced seasonality, and ARIMAX with advertising spend as an exogenous input was best (MAPE 11.56%). The two MAPE values cannot be compared directly because the datasets, frequencies, horizons and inputs differ; what the studies share is that a well-specified, interpretable statistical model is highly competitive.

9. Business Implications.

- Stock and budget planning should follow the strong Q4 peak (September, November, December) and the low periods (January, February, August).
- Discount policy needs review, given the negative discount–profit correlation (−0.219).
- Tables and Bookcases generate sales but losses; pricing, discounting and supplier costs should be examined.
- The Central region is less profitable than South despite higher sales and deserves investigation.
- Technology is the largest profit contributor and should be prioritized in inventory and promotion.

Limitations. (1) A MAPE of 25.73% is moderate; with 36 training months the seasonal model sees only three yearly cycles. (2) A single chronological split was used; rolling-origin cross-validation would be more robust. (3) Only sales history is modelled, without promotions, festivals, pricing or macroeconomic variables. (4) The confidence intervals are wide, so forecasts should be treated as planning ranges. (5) Superstore is a sample dataset from a US-style retail setting, so results may not generalize without retraining. (6) Machine learning and deep learning models (Random Forest, XGBoost, SVR, LSTM, Prophet) were not evaluated.

VI. CONCLUSION AND FUTURE SCOPE

This paper presented an end-to-end sales analytics solution covering data cleaning, EDA, SQL KPIs, time-series analysis, forecasting and a Power BI dashboard. On 9,994 Superstore transactions, SARIMA(0,1,1)x(0,1,1,12) was the best of four models with MAE = 13,006.71, RMSE = 15,495.15 and MAPE = 25.73%, a 40.27 percentage-point improvement over the moving-average baseline, with no remaining residual autocorrelation. The analysis revealed a strong seasonal pattern, a negative effect of discounts on profit, and loss-making sub-categories. Integrating the forecast into an interactive dashboard turns the statistical output into a practical decision-support tool.

Future work includes adding exogenous variables (promotions, festivals, pricing), comparing ARIMAX, Random Forest, XGBoost, SVR, LSTM and Prophet, using rolling-origin cross-validation and Diebold–Mariano tests, forecasting at category or region level, and automating data refresh in Power BI. memory neural network methodology," in Proc. 26th Int. Conf. Neural Inf. Process. (ICONIP), Sydney, NSW, Australia, 2019, pp. 462–474.

REFERENCES

1. S. Cheriyan, S. Ibrahim, S. Mohanan, and S. Treesa, "Intelligent sales prediction using machine learning techniques," in Proc. Int. Conf. Comput., Electron. Commun. Eng. (iCCECE), Southend, U.K., 2018, pp. 53–58.
2. G. Box, "Box and Jenkins: Time series analysis, forecasting and control," in A Very British Affair,

Palgrave Advanced Texts in Econometrics, 2013, pp. 161–215.
3. R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, Australia: OTexts, 2021.
4. A. K. Tyagi and P. Chahal, "Artificial intelligence and machine learning algorithms," in Research Anthology on Machine Learning Techniques, Methods, and Applications, IGI Global, 2022.
5. J. Fattah, L. Ezzine, Z. Aman, H. El Moussami, and A. Lachhab, "Forecasting of demand using ARIMA model," Int. J. Eng. Bus. Manag., vol. 10, 2018.
6. A. A. Dar, A. Jain, M. Malhotra, A. R. Farooqi, O. Albalawi, M. S. Khan, and Hiba, "Time series analysis with ARIMA for historical stock data and future projections," Soft Comput., vol. 28, pp. 12531–12542, 2024.
7. V. Nithya, S. Rajesh, and S. Nithyasri, "Sales data forecasting using ARIMA model," in Proc. 5th Int. Conf. Artif. Intell. Smart Energy (ICAIS), 2025, pp. 128–138.
8. S. Siami-Namini and A. S. Namin, "Forecasting economics and financial time series: ARIMA vs. LSTM," arXiv preprint arXiv:1803.06386, 2018.
9. K. Bandara, P. Shi, C. Bergmeir, H. Hewamalage, Q. Tran, and B. Seaman, "Sales demand forecast in e-commerce using a long short-term
10. S. J. Taylor and B. Letham, "Forecasting at scale," PeerJ Preprints, vol. 5, p. e3190v2, 2017.
11. A. A. Dar, S. Sanyal, K. Khan, M. S. Khan, A. Smerat, and V. Arora, "Forecasting retail sales in the Indian market: Integrating time-series and machine learning approaches," IEEE, base paper.