Pharmaceutical Sales Forecasting and Planning

Ahmed M. Alaa

December 2025

Contents

1 - Introduction

1.1 - Project overview

Accurate demand forecasting is a critical task in pharmaceutical operations, as it directly impacts inventory management, procurement planning, and service availability. Pharmacies must balance the risk of stockouts against excess inventory while accounting for strong seasonal patterns, weekly effects, and closure days. This project aims to analyze historical pharmaceutical sales data and develop reliable forecasting models to support short-term and long-term sales planning.

The analysis is based on daily sales data aggregated across eight drug categories, with a primary focus on modeling and forecasting **total daily pharmaceutical sales**. The workflow includes data cleaning, exploratory data analysis, seasonality assessment, feature engineering, and the application of multiple predictive models. Several modeling approaches are evaluated, ranging from machine-learning models with lagged features to time-series models designed to capture temporal dependencies and calendar effects. Model performance is assessed using standard accuracy metrics, and an ensemble forecast is constructed to improve robustness.

1.2 - About the Data source

The dataset used in this study originates from the **Pharmaceutical Sales Data** dataset publicly available on **Kaggle**, compiled by Milan Zdravković. The specific file utilized, `salesdaily.csv`, contains daily sales observations for eight pharmaceutical drug categories over multiple years, along with calendar information such as dates and weekdays. This real-world dataset enables a detailed examination of sales behavior, including weekly and monthly seasonality, holiday effects, and periods of zero sales corresponding to pharmacy closures. The dataset serves as the foundation for both the exploratory analysis and the forecasting models developed in this report. *(Zdravković ,2018)*

Drug Classification

The dataset includes daily sales records for **57 individual pharmaceutical products**, which are aggregated into eight therapeutic groups based on the **Anatomical Therapeutic Chemical (ATC) Classification System**. Each group represents a clinically meaningful drug category, allowing sales patterns to be analyzed at a higher and more interpretable level. The eight ATC categories included in the dataset are:

- **M01AB** – Anti-inflammatory and antirheumatic products (non-steroids), acetic acid derivatives and related substances
- **M01AE** – Anti-inflammatory and antirheumatic products (non-steroids), propionic acid derivatives
- **N02BA** – Other analgesics and antipyretics, salicylic acid and derivatives
- **N02BE/B** – Other analgesics and antipyretics, pyrazolones and anilides
- **N05B** – Psycholeptics, anxiolytic drugs
- **N05C** – Psycholeptics, hypnotics and sedatives

- **R03** – Drugs for obstructive airway diseases
- **R06** – Antihistamines for systemic use

For the primary forecasting task, sales across these eight categories are aggregated into a single **total daily sales** variable. This approach captures overall pharmacy demand while preserving the underlying therapeutic structure for exploratory analysis and interpretation.

A more granular, category-level analysis and forecasting of individual drug groups is intentionally deferred and will be addressed separately in **Part 2** of this study.

2 - Methods and Analysis

2.1- Data Preparation and Cleaning

Data cleaning and exploratory analysis were performed using the tidyverse ecosystem in R (Wickham et al., 2019).

The analysis was conducted using the daily pharmaceutical sales dataset. Initial data preparation focused on ensuring consistency, correctness, and suitability for time series modeling.

load data and libraries

Key preprocessing steps included:

- **Removal of irrelevant or incorrect columns** generated during data collection (e.g., redundant time fields).
- **Construction of a proper date variable** to enable time-based analysis.
- **Feature engineering**, including:
  - Calendar features: year, month, and weekday
  - A total daily sales variable (t__sales) calculated as the sum of sales across all drug categories
- **Identification of zero-sales days**, which were interpreted as pharmacy closure days and encoded using a binary indicator variable (is__closed).

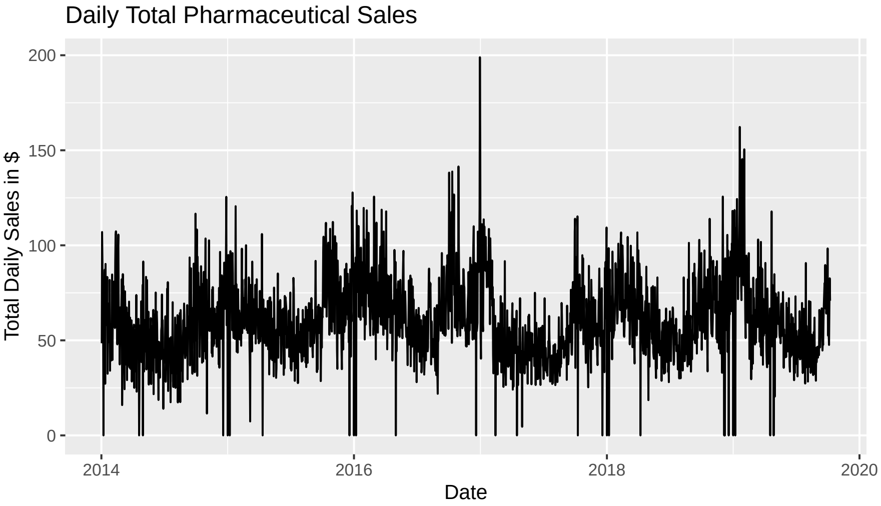
These steps ensured that the dataset accurately reflected operational behavior and was suitable for both statistical and machine learning models.

2.2 - Exploratory Data Analysis (EDA)

Exploratory analysis was performed to understand sales dynamics, seasonal patterns, and variability across time.

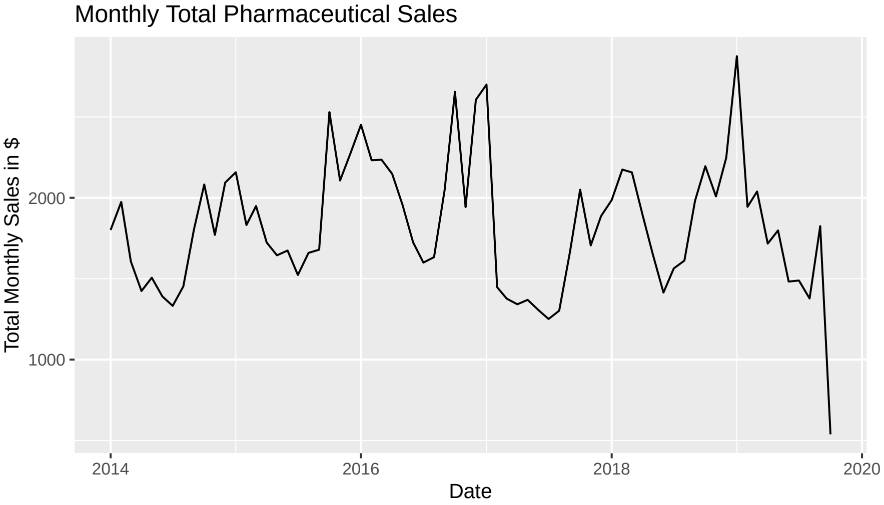
A-Sales Trends and Seasonality

1- Daily Total Pharmaceutical Sales



Daily sales plots revealed high short-term volatility, making it difficult to identify long-term patterns.

2 - Monthly Total Pharmaceutical Sales

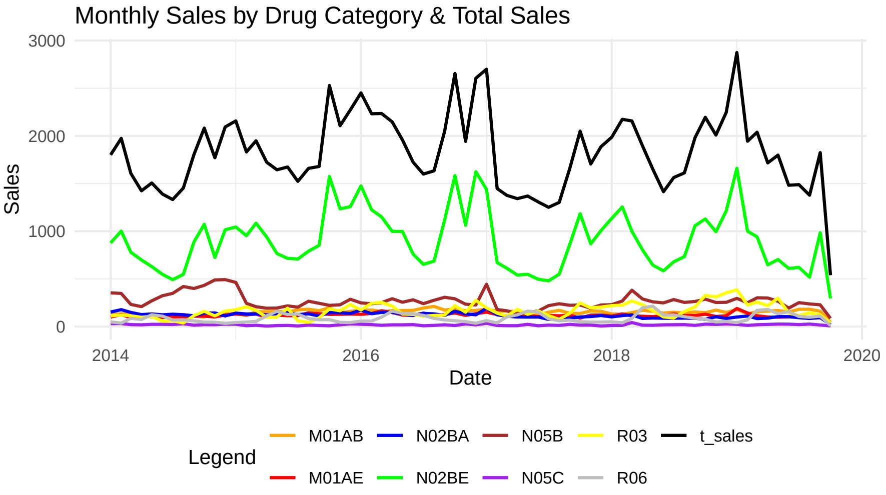


To improve interpretability, sales were **aggregated to a monthly level**, revealing clear seasonal effects:

- Sales tend to **decline during summer months (June–August)**
- Higher demand is observed in **late autumn and winter (October–January)**

B - Drug Category Behavior

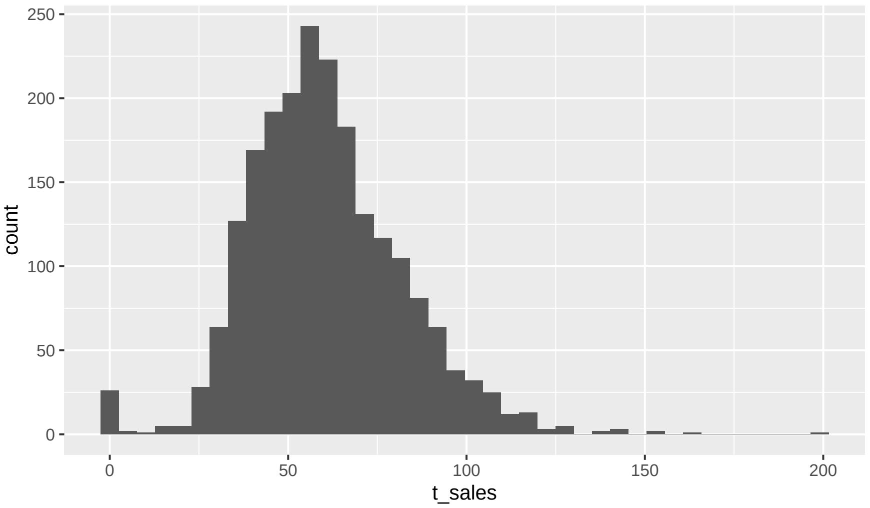
1-Monthly Sales by Drug Category



Monthly sales by drug category were visualized to assess whether individual categories exhibited distinct trends. While categories showed varying magnitudes and volatility, their **aggregate behavior closely followed total sales**, justifying the focus on total sales for forecasting.

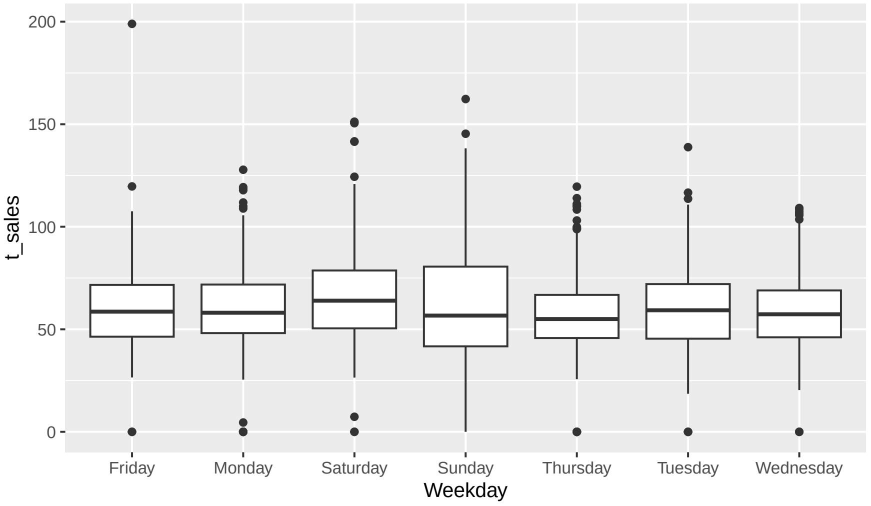
C -Distribution and Calendar Effects

1 - Distribution of daily sales



Daily total sales were **right-skewed**, indicating occasional high-demand days.

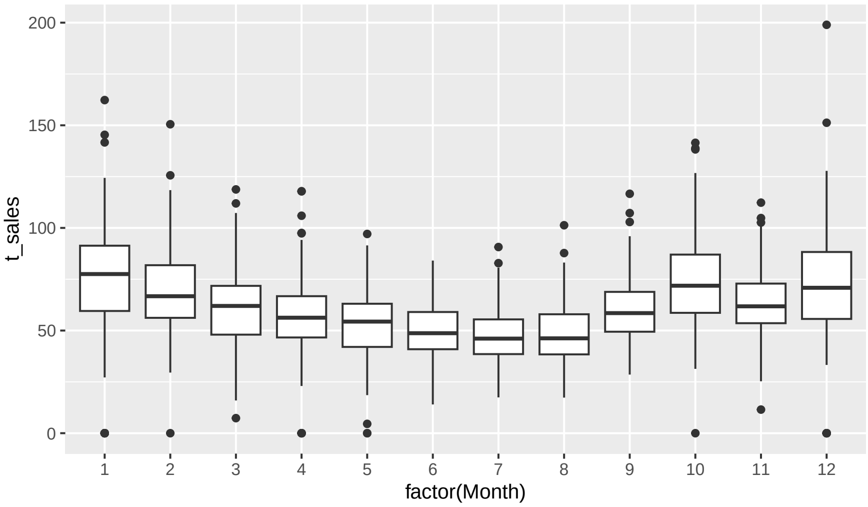
2 - Weekly seasonality



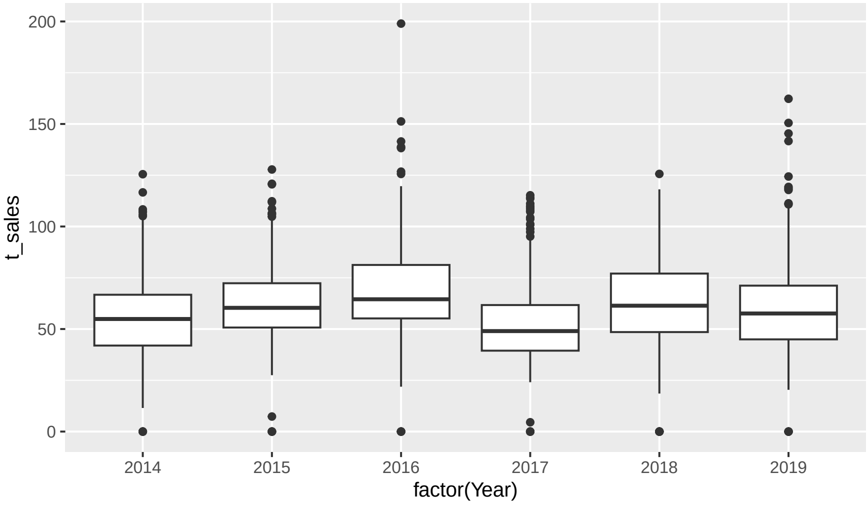
Weekly seasonality was observed through boxplots:

- Sales variability differed across weekdays
- **Sundays** showed **lower median** sales with higher variability, likely reflecting **reduced operating hours or demand patterns**

3- Monthly Seasonality



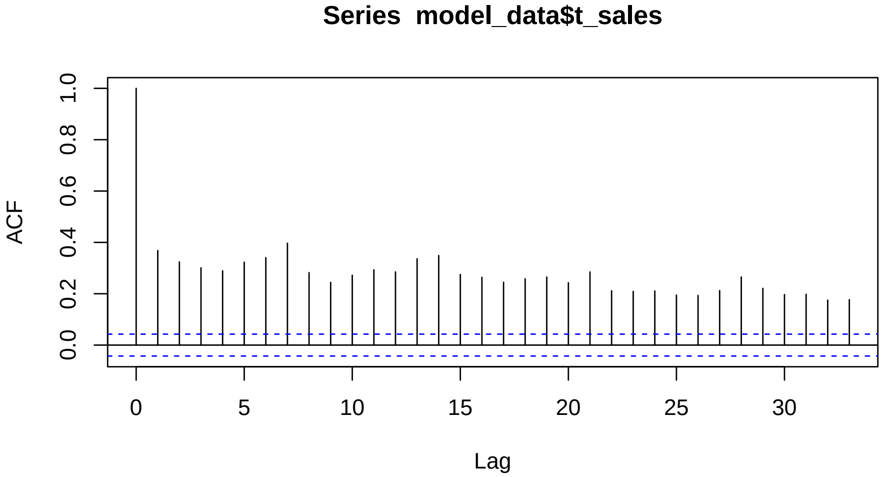
4- Annual Seasonality



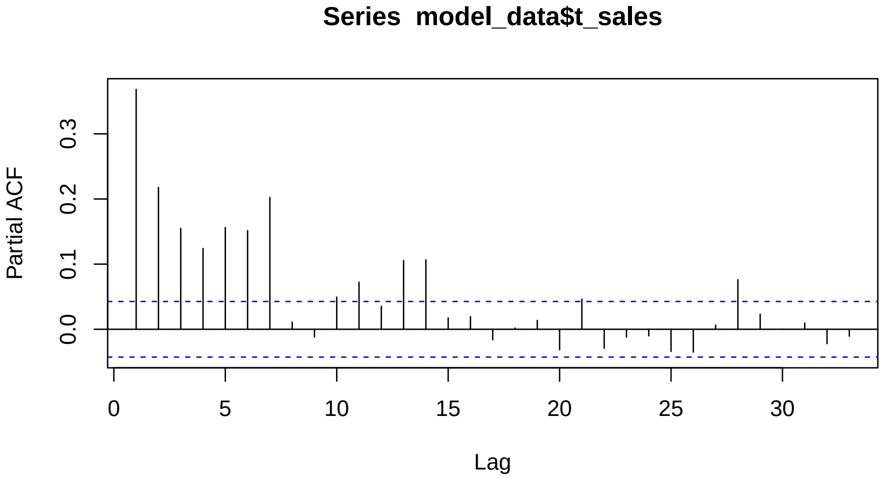
moderate **annual seasonality**

D - Autocorrelation Analysis

1- Autocorrelation (ACF)



2- partial autocorrelation (PACF)



Autocorrelation (ACF) and partial autocorrelation (PACF) plots indicated:

- Strong correlations at **weekly lags (7, 14, 21 days)**
- Evidence that current sales depend on recent past values

These findings motivated the use of lag-based features and time series models.

E - Stationarity Testing

Before fitting time series models, formal stationarity tests were applied:

1 - ADF test (Augmented Dickey–Fuller)

```
##
##  Augmented Dickey-Fuller Test
##
## data:  model_data$t_sales
## Dickey-Fuller = -5.2507, Lag order = 12, p-value = 0.01
## alternative hypothesis: stationary
```

2- KPSS test

```
##
##  KPSS Test for Level Stationarity
##
## data:  model_data$t_sales
## KPSS Level = 0.4363, Truncation lag parameter = 8, p-value = 0.06151
```

Both tests supported the conclusion that the sales series is **stationary**, allowing models to be fitted without differencing.

Autocorrelation and stationarity were assessed using ACF, PACF, and unit root tests as implemented in the forecast package *(Hyndman & Khandakar, 2008).*

2.3 - Modeling Approach

To capture different aspects of the data, multiple modeling techniques were applied. These models range from machine learning approaches to classical time series methods.

A- Elastic Net Regression with Lag Features

Elastic Net regression was used as a **machine learning baseline model** incorporating both autoregressive behavior and calendar effects.

Features included:

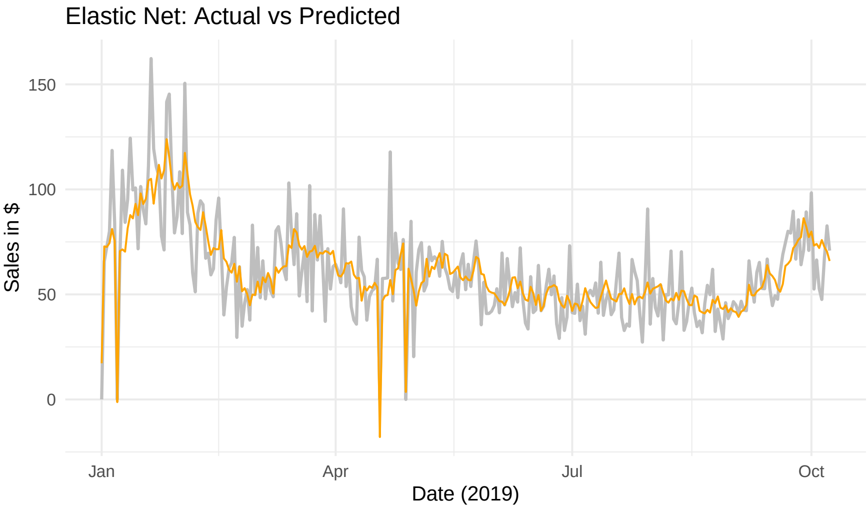
- Lagged sales values (1, 7, 14, 21, 30, and 60 days)
- Rolling averages (7-day and 30-day)
- Calendar variables (weekday, month)
- Closure indicator (`is_closed`)

Elastic Net combines **L1 and L2 regularization**, allowing it to:

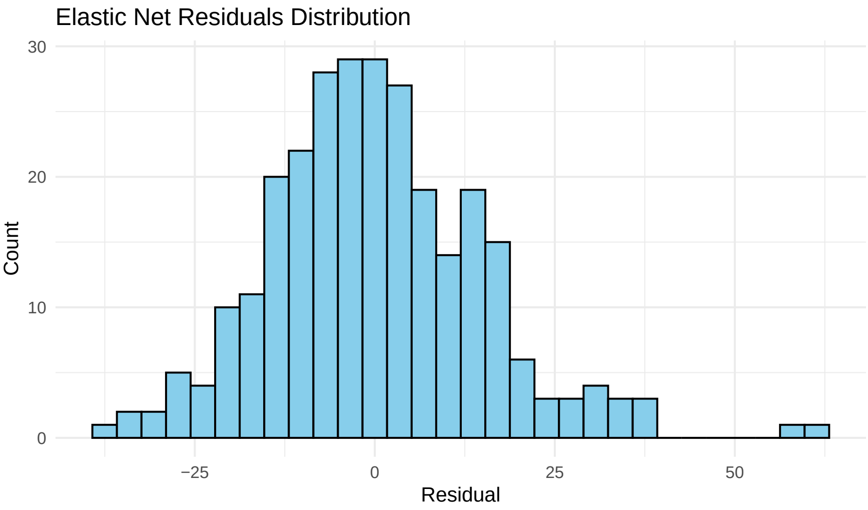
- Handle multicollinearity among lag features
- Perform implicit feature selection
- Maintain interpretability

This model performed well for short-term forecasting and **served as a strong benchmark** *(Friedman et al., 2010)*.

1- Elastic Net predictions VS Actual data



2- Elastic Net predictions Residuals

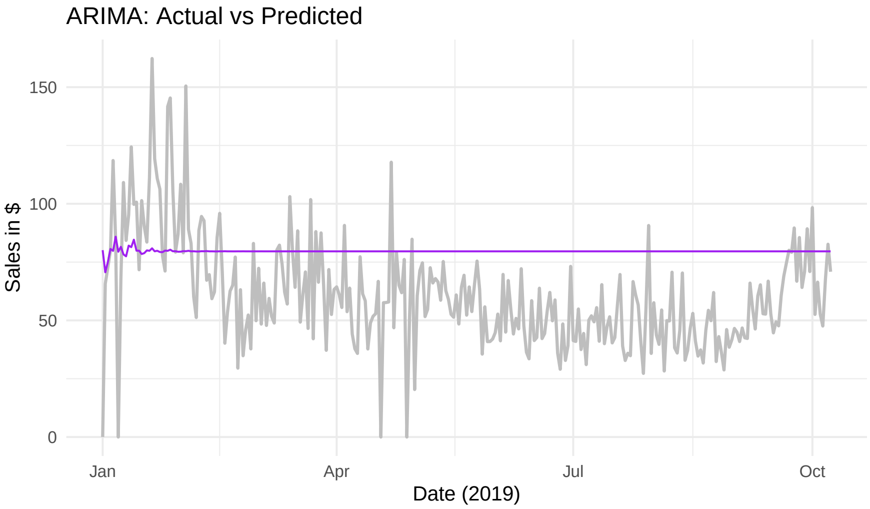


B - ARIMA (Autoregressive Integrated Moving Average)

A plain ARIMA model was fitted using only the historical sales series. While ARIMA captured temporal dependence, its performance was limited due to the absence of external explanatory variables.

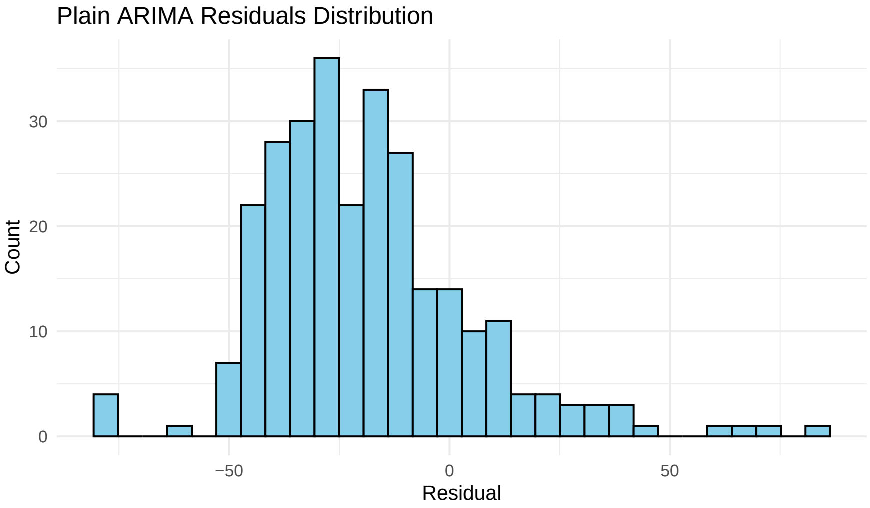
This model served primarily as a **baseline time series comparator**.

1 - ARIMA predictions VS Actual data



ARIMA alone without regressors Shows bad prediction of Sales

2- ARIMA predictions Residuals



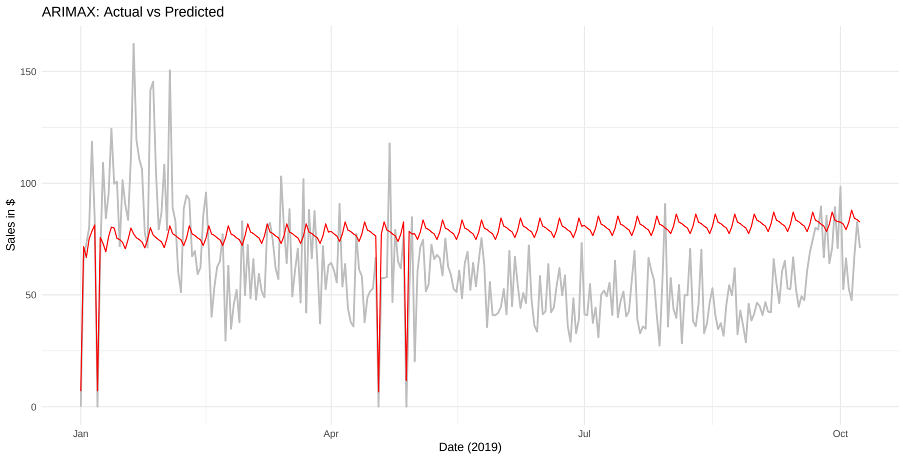
C - ARIMAX (ARIMA with Exogenous Variables)

To improve upon ARIMA, an ARIMAX model was fitted by incorporating:

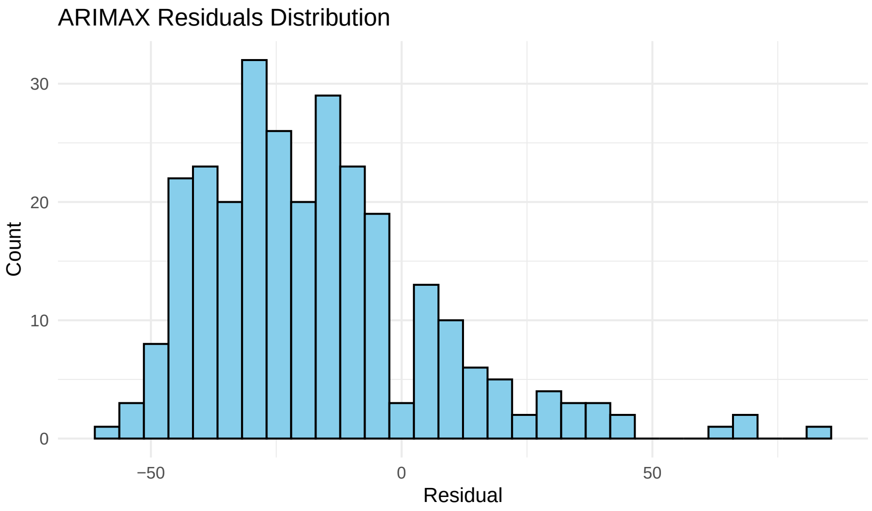
- Weekday
- Month
- Pharmacy closure indicator

Including these exogenous variables improved forecast accuracy by allowing the model to account for calendar-driven demand changes *(Hyndman & Khandakar, 2008).*

1 - ARIMAX predictions VS Actual data



2 - ARIMAX predictions Residuals



D - Prophet Model with Regressors

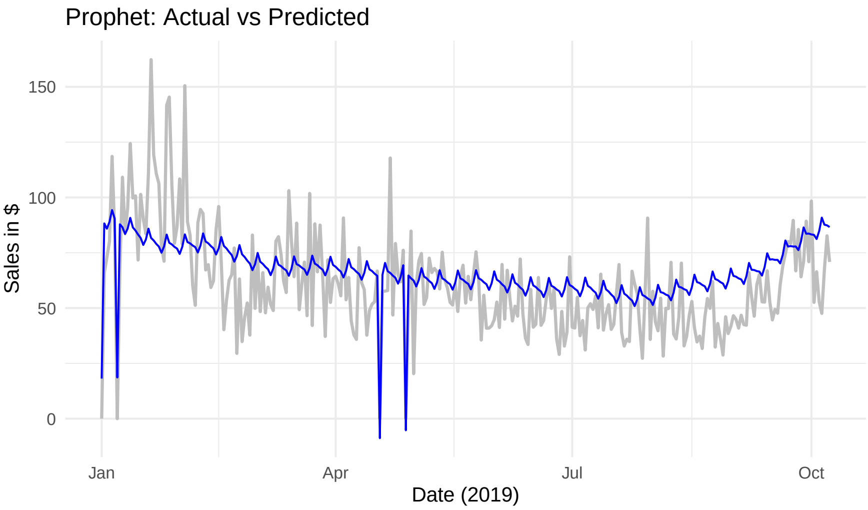
Facebook's Prophet model was used as a more advanced forecasting approach *(Taylor & Letham, 2018).*

Prophet is well-suited for business time series due to its ability to model:

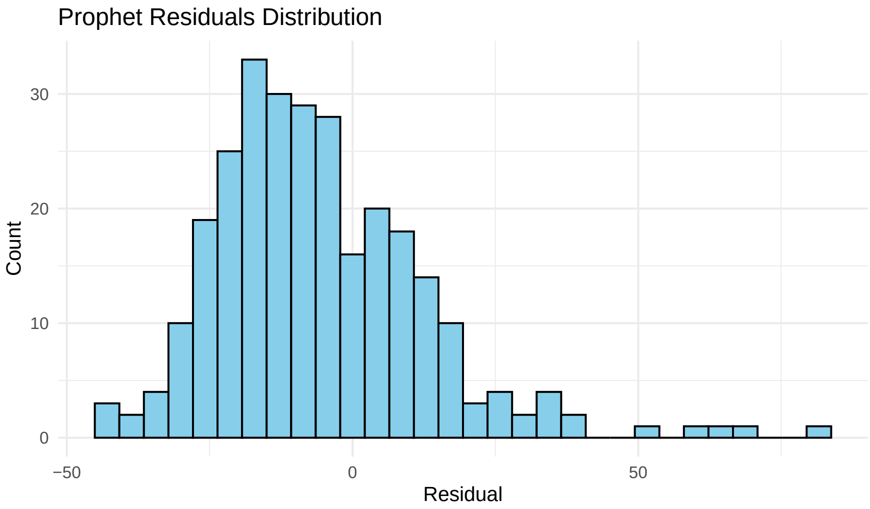
- Weekly seasonality
- Yearly seasonality
- Holiday and event effects

The `is_closed` variable was included as an external regressor, enabling the model to explicitly account for non-operational days.

1 - Prophet predictions VS Actual data



2 - Prophet predictions Residuals



E - Ensemble Forecasting

Finally, an **ensemble forecast** was created by averaging predictions from the ARIMAX and Prophet models (Hyndman & Khandakar, 2008).

The rationale behind this approach was to:

- Reduce model-specific bias
- Improve robustness for long-term forecasting
- Balance statistical and decomposition-based modeling strengths

2.4 - Evaluation Strategy

Model performance was evaluated using a **holdout test set (2019)** and compared using:

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)

Visual diagnostics, including predicted vs. actual plots and residual analysis, were also used to assess model behavior and stability.

3 - Results

This section presents the forecasting performance of the different models applied to daily pharmaceutical sales data. Model accuracy was evaluated using **Root Mean Squared Error (RMSE)**, **Mean Absolute Error (MAE)**, and **Mean Absolute Percentage Error (MAPE)** on the 2019 test period. Lower values indicate better predictive performance.

3.1 - Model Performance Comparison

Table 1: 2019 Predictions (head)

Date	Actual	ElasticNet	ARIMA	ARIMAX	Prophet	Ensemble
2019-01-01	0.000	17.22835	80.11785	6.945949	18.13519	12.54057
2019-01-02	65.677	72.87620	70.61954	71.554029	88.27721	79.91562
2019-01-03	72.816	72.60216	75.03043	66.783525	85.86147	76.32250
2019-01-04	80.070	74.43921	80.65052	75.076763	88.94326	82.01001
2019-01-05	118.550	81.14926	79.80763	78.459973	94.30835	86.38416
2019-01-06	83.990	75.34962	85.89129	81.314626	90.42202	85.86832

Metrics

Table 2: Metrics

Model	RMSE	MAE	MAPE
Elastic Net	14.86895	11.45963	20.13843
ARIMA	30.00150	25.50599	51.04266
ARIMAX	29.05929	25.05741	51.76689
Prophet	19.41464	15.58293	30.07354
Ensemble	23.48974	19.51082	39.78916

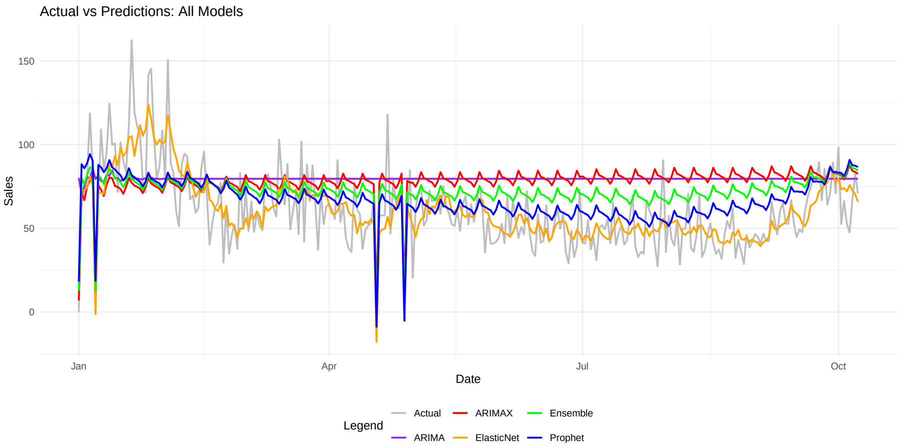
Table (Metrics)summarizes the performance of all evaluated models.

- **Elastic Net with lagged features** achieved the best overall accuracy across all metrics.
- **Prophet with holiday (closure) effects** provided robust performance, particularly in capturing seasonal patterns and zero-sales days.
- **ARIMAX**, despite incorporating calendar-based regressors, showed limited improvement over the plain ARIMA model.
- **Plain ARIMA** performed worst, highlighting the importance of external information and nonlinear patterns.
- The **ensemble model**, defined as the average of ARIMAX and Prophet forecasts, produced moderate performance but did not outperform the best individual models.

Key findings:

- Lag-based machine learning models outperform traditional time-series models for short-horizon forecasts.
- Calendar effects and closure information significantly improve forecast stability.
- Combining models does not always guarantee better accuracy, especially when individual model strengths overlap.

3.2 - Visual Evaluation of Forecasts



Visual inspection of forecast plots confirms the quantitative results:

- Elastic Net predictions closely track actual sales, with smaller deviations during high-volatility periods.
- Prophet captures seasonal trends effectively but smooths short-term fluctuations.
- ARIMA and ARIMAX models struggle to adapt to sudden changes and zero-sales days.
- The ensemble forecast reduces extreme errors but inherits weaknesses from both base models.
- Residual plots further indicate:
- No strong remaining trend in Elastic Net residuals
- Higher variance and clustering in ARIMA-based residuals

3.3 - Forecasting 2020 Sales

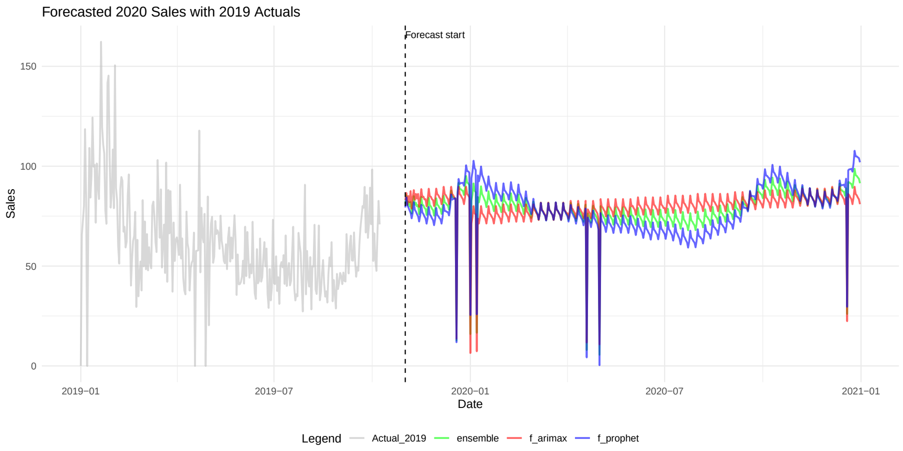
For forward-looking planning, **ARIMAX and Prophet models** were used to forecast sales from **November 2019 through December 2020**, incorporating known pharmacy closure days.

Daily Forecasts

Table 3: Daily Forecasts (head)

Date	Weekday	Month	is_closed	f_arimax	f_prophet	ensemble
2019-11-01	Friday	11	0	82.98705	79.30513	81.14609
2019-11-02	Saturday	11	0	86.60892	84.38627	85.49760
2019-11-03	Sunday	11	0	80.67047	80.34507	80.50777
2019-11-04	Monday	11	0	85.05038	79.29651	82.17345
2019-11-05	Tuesday	11	0	81.88303	77.84424	79.86363
2019-11-06	Wednesday	11	0	87.53134	76.67940	82.10537

Plotting the Forecasting Results



To support business decision-making:

- Daily forecasts were aggregated into **weekly and monthly sales plans**
- Monthly forecasts were rounded for operational usability

Ensemble forecasts were included to provide a balanced planning reference

Monthly sales Plan

Table 4: Monthly Sales Plans

Year	Month	Planned_Sales
2019	Nov	2408
2019	Dec	2555
2020	Jan	2446
2020	Feb	2336
2020	Mar	2388
2020	Apr	2207

Year	Month	Planned_Sales
2020	May	2252
2020	Jun	2212
2020	Jul	2248
2020	Aug	2322
2020	Sep	2442
2020	Oct	2738
2020	Nov	2527
2020	Dec	2684

Weekly sales Plan

Table 5: Weekly Sales Plans (head)

Week_Start	Planned_Sales
2019-10-28	247
2019-11-04	567
2019-11-11	563
2019-11-18	557
2019-11-25	555
2019-12-02	561

These aggregated forecasts enable:

- Inventory and procurement planning
- Seasonal demand anticipation
- Risk reduction through model diversification

3.4 Summary of Results

- **Best-performing model (short-term):** Elastic Net with lagged features
- **Best-performing model (long-term):** Prophet with closure effects
- **Weakest model:** Plain ARIMA
- **Recommended planning signal:** Prophet or ARIMAX–Prophet ensemble

4 - Conclusion

This project analyzed daily pharmaceutical sales data and developed forecasting models to support sales planning and inventory decision-making. Using historical sales from 2014–2019, multiple time-series and machine learning approaches were evaluated to understand sales behavior and predict future demand.

The exploratory analysis revealed **strong weekly and monthly seasonality**, with noticeable declines during summer months and increased sales toward the end of the year. Several zero-sales days were identified and treated as **store closure events**, which proved important for improving model accuracy. Aggregating data to the monthly level helped uncover clearer trends and supported better interpretation for business users.

From a modeling perspective, **Elastic Net with lagged and rolling features achieved the best short-term predictive accuracy**, demonstrating the importance of historical dependency in daily sales. However, its reliance on past sales makes it less suitable for long-term forecasting. **Prophet**, enhanced with a store-closure regressor, provided more robust performance for longer horizons, while **ARIMAX** showed limited improvement due to weak explanatory power of calendar-based regressors alone. An **ensemble forecast** combining Prophet and ARIMAX was used to balance stability and interpretability.

The final forecasts for **November 2019 through December 2020** were produced at daily, weekly, and monthly levels to support operational planning. Monthly and weekly aggregations were emphasized to assist purchasing and inventory management, where longer-term demand signals are more actionable than daily fluctuations.

4.1 - Limitations

- The models rely solely on historical sales and calendar effects, excluding external drivers such as promotions, pricing changes, epidemics, or economic conditions.
- Closure days in 2020 were inferred from historical patterns and public holidays, which may not fully capture unplanned closures.
- Forecasting was performed at the **total sales level**, potentially masking category-specific dynamics.

4.2 - Future Work

Future extensions of this project will include:

- **Category-level forecasting** to capture heterogeneous demand patterns across drug classes
- Incorporation of **exogenous variables** such as promotions, weather, or epidemiological data
- Probabilistic forecasting and uncertainty intervals for risk-aware planning

Overall, this analysis demonstrates that combining classical time-series methods with machine learning and domain-informed features can produce reliable forecasts that are both **accurate and operationally useful**.

5 - References

1. Zdravković, M. (2018). *Pharmaceutical Sales Data.* Kaggle.

   Available at: https://www.kaggle.com/datasets/milanzdravkovic/pharma-sales-data

2. Hyndman, R. J., & Khandakar, Y. (2008). *Automatic Time Series Forecasting: The forecast Package for R.* Journal of Statistical Software, 27(3), 1–22.

3. Friedman, J., Hastie, T., & Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 33(1), 1–22.

4. Taylor, S. J., & Letham, B. (2018). *Forecasting at Scale.* The American Statistician, 72(1), 37–45.

5. Wickham, H., Averick, M., Bryan, J., et al. (2019). *Welcome to the tidyverse.* Journal of Open Source Software, 4(43), 1686.