# Forecasting pharmaceutical sales, Teva Pharma

[**Forecasting pharmaceutical sales,   
Teva Pharma 1**](#_heading=h.gjdgxs)

[INTRODUCTION 1](#_heading=h.30j0zll)

[Data Description 1](#_heading=h.nozwnqk7qox0)

[Task Definition 2](#_heading=h.q4xkx45913tl)

[Insights about the data 3](#_heading=h.ie7ud4wmu19y)

[Data Preprocessing: 6](#_heading=h.4d34og8)

[Feature Selection/Engineering: 10](#_heading=h.gv7itxz1cgv2)

[SARIMAX Model Implementation 13](#_heading=h.gl8s1i5f6pqh)

[Evaluation: 13](#_heading=h.26in1rg)

[Prediction: 14](#_heading=h.lnxbz9)

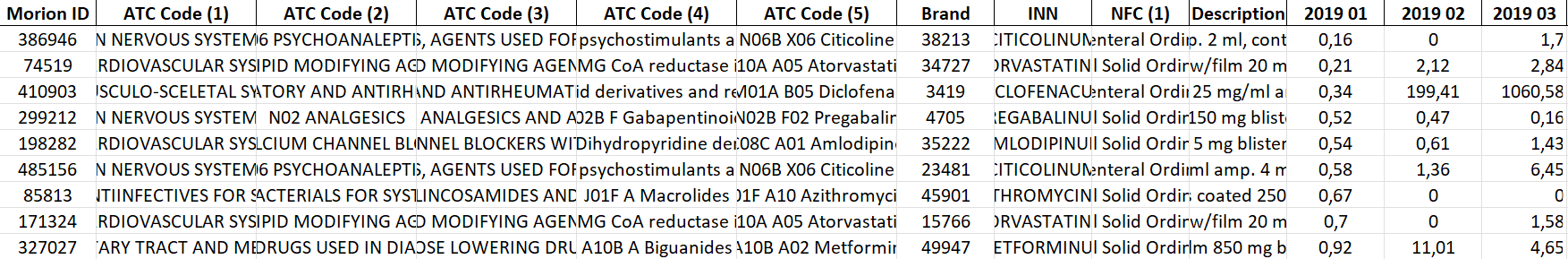
## **INTRODUCTION**

This document outlines the process for forecasting pharmaceutical sales for different medications produced and offered by Teva Pharmaceuticals. The objective is to predict future sales for various products using historical sales data and product characteristics. Accurately forecasting sales is crucial for inventory management, supply chain optimization, and strategic planning. The project leverages machine learning techniques to analyze past sales patterns and project future demand, providing actionable insights for Teva Pharmaceuticals.

### Data Description

The company provided two datasets for executing the forecast - a sampled dataset of 20 molecules (INNs), to be used for the development of a solution, and a full data set consisting of a much bigger range of medicines marketed by the corporation. Each dataset consists of the following columns, each providing essential information about the products and their sales:

* **Morion ID**: Unique identifier for the medical product.
* **ATC Code (1)**: First level of ATC classification (most general), representing the broad therapeutic area.
* **ATC Code (2)**: Second level of ATC classification, providing a more specific therapeutic sub-category.
* **ATC Code (3)**: Third level of ATC classification, offering further detail on the sub-category.
* **ATC Code (4)**: Fourth level of ATC classification, narrowing down to specific therapeutic properties.
* **ATC Code (5)**: Fifth level of ATC classification (most granular), detailing the exact type of drug.
  + ***\*ATC groups*** *refer to the therapeutical group of the product.*
* **Brand**: Anonymized code representing the brand of the product.
* **INN**: International Nonproprietary Name, indicating the active molecule or combination of molecules. This is the core concept of pharmaceutical products. It is the active substance, that is providing the main therapeutical effect.
* **NFC (1)**: Form or presentation of the product, such as tablet, gel, syrup, injection, etc. One active substance can take many forms: oral tablets, gels, syrups, injections, etc. Each product, independent of the form, is characterized by strength and pack size. Strength is how much active substance is in one pill/ tube/ ampoule, etc. Pack size refers to how many pills are in the package, or how large the tube is, etc. The quantities sold largely depend on these characteristics.
* **Description**: Detailed product description, including concentration/strength and pack size.
* **2019 01 - 2023 12**: Monthly sales quantities (number of packs) for each product ID over the specified period.



### Task Definition

Our primary task was to forecast sales for each product over a 12-month period. This forecasting is essential due to the interrelations between similar products. When one product is unavailable (indicated by zero sales), patients often opt for alternative replacements. To capture these dynamics, forecasting at the Molecule (INN) level or ATC levels 4 or 5 is ideal. A single model should ideally forecast for all products derived from one INN, and hierarchical models can be particularly useful. Normalization poses a challenge because variations in product presentation, strength, and package sizes affect sales quantities.

Our second assignment aimed to develop a model that links product characteristics such as strength, package size, and form of administration to sales quantity. For instance, does a package of 10 tablets sell twice as much as a package of 20 tablets?

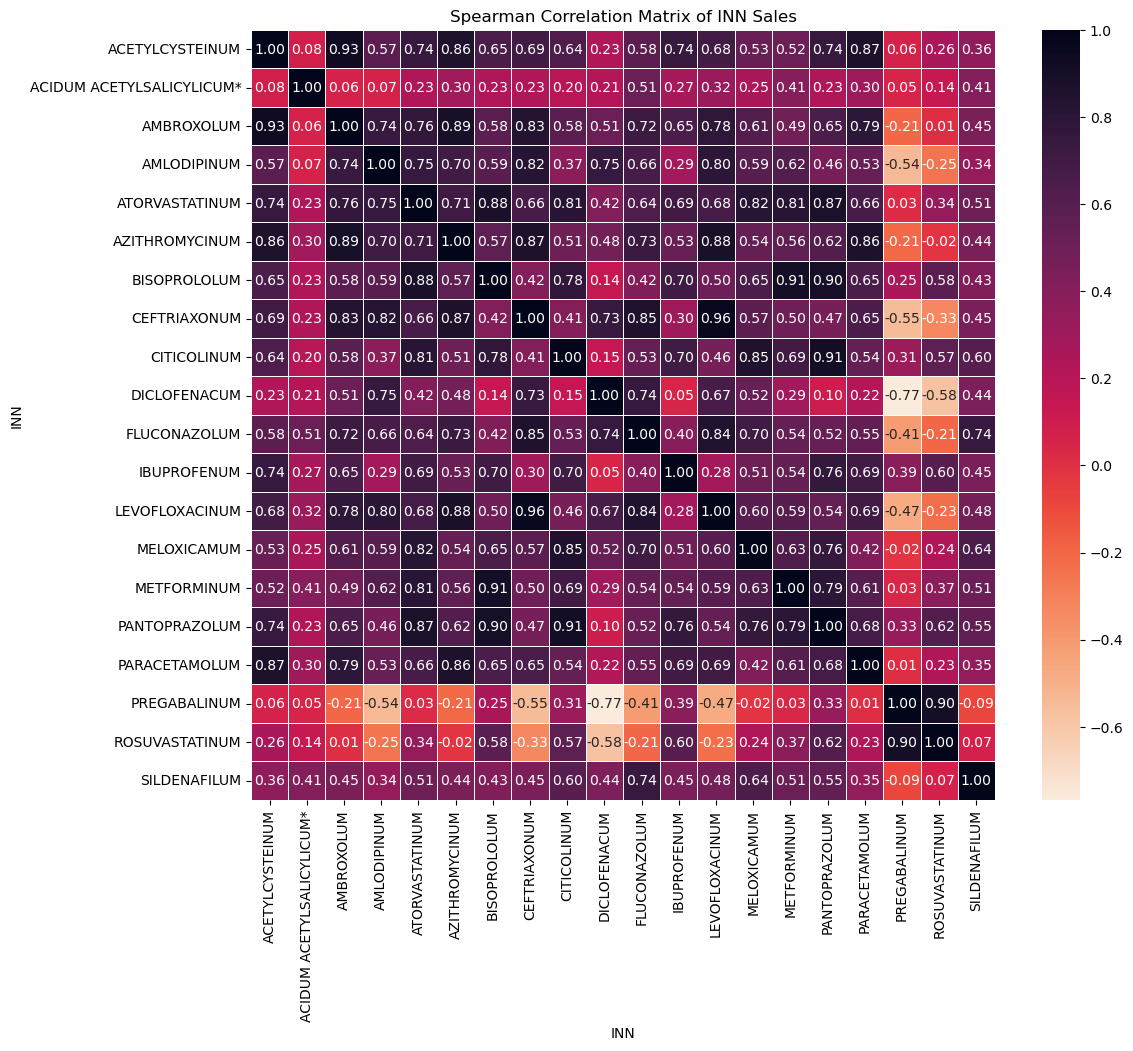
The final task was to conceptualize organizing the products into a network or graph based on their characteristics (e.g., strength, pack size, therapeutic class). The idea was to apply graph machine learning and graph forecasting techniques to potentially improve the sales predictions.

Due to time constraints the forecast described in the next few parts includes only the first two tasks set by the company. We focus on predicting the amount of sales for the best selling drug in each INN (molecule) and to evaluate which factors affect its sales.

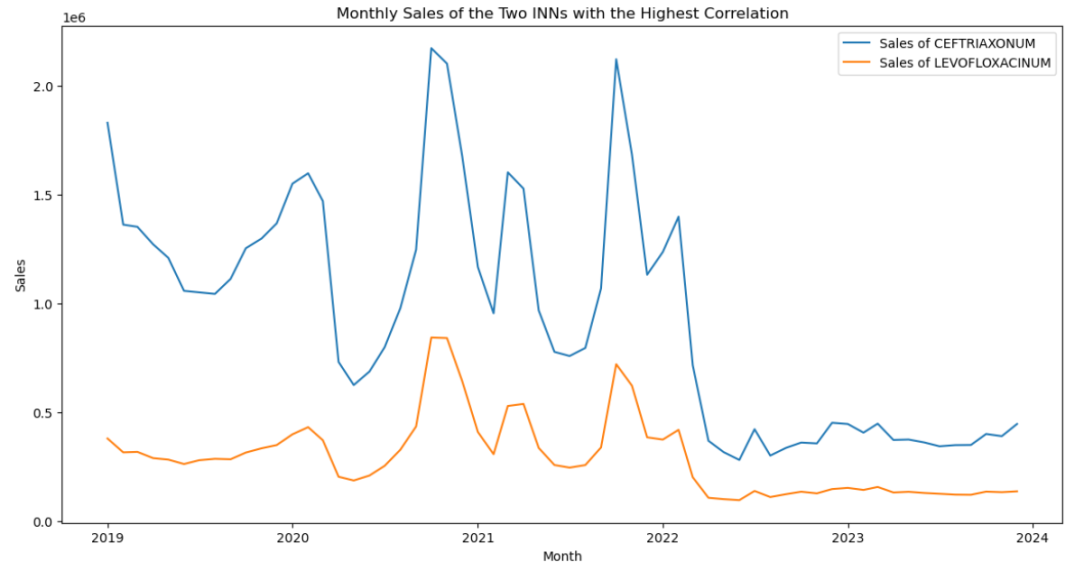
### Insights about the data

To get familiar with the data and understand the challenges ahead, we performed several analyses and visualizations.

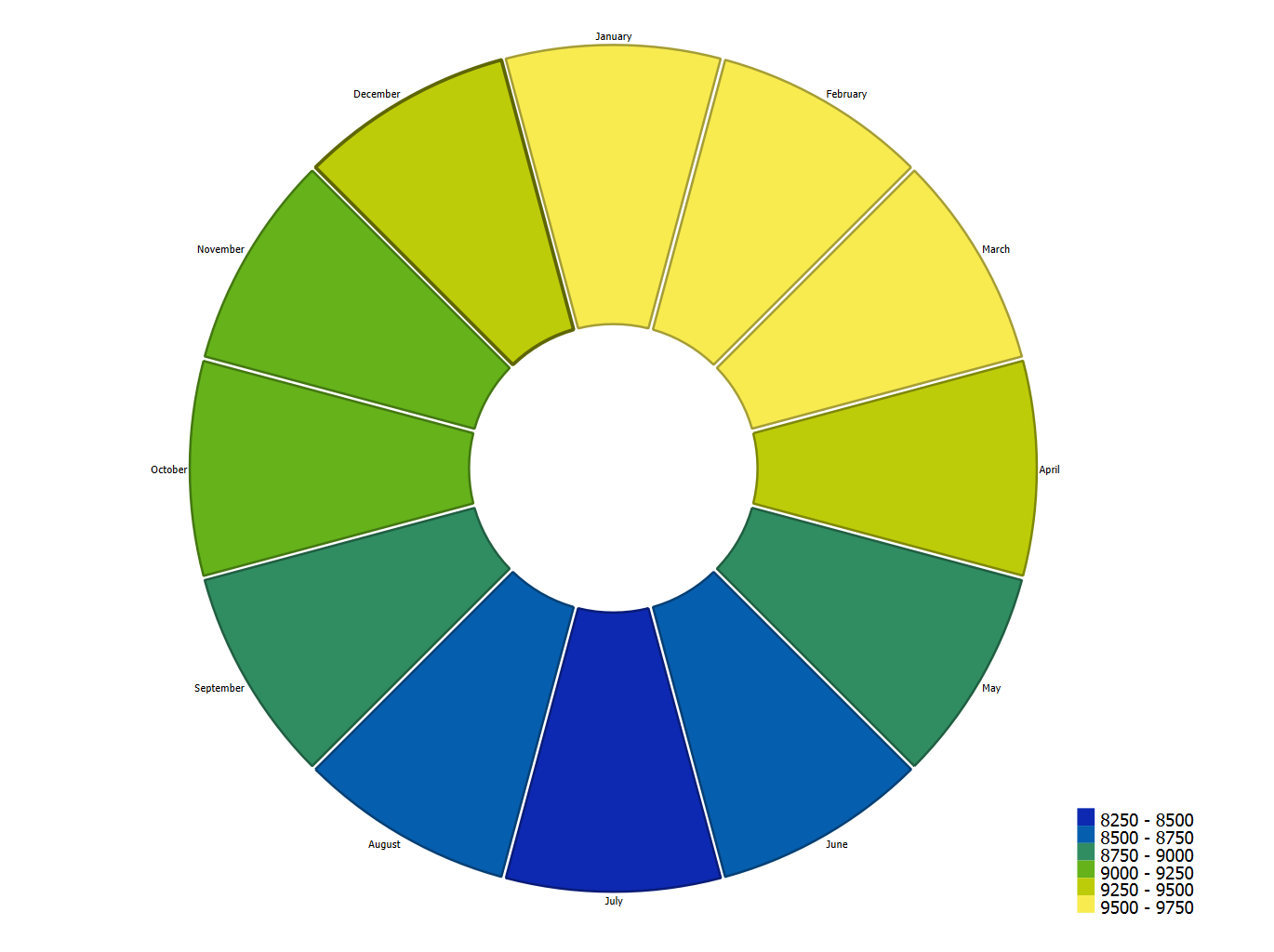
We created a correlation matrix between the sales of each molecule to identify the most correlated molecules.



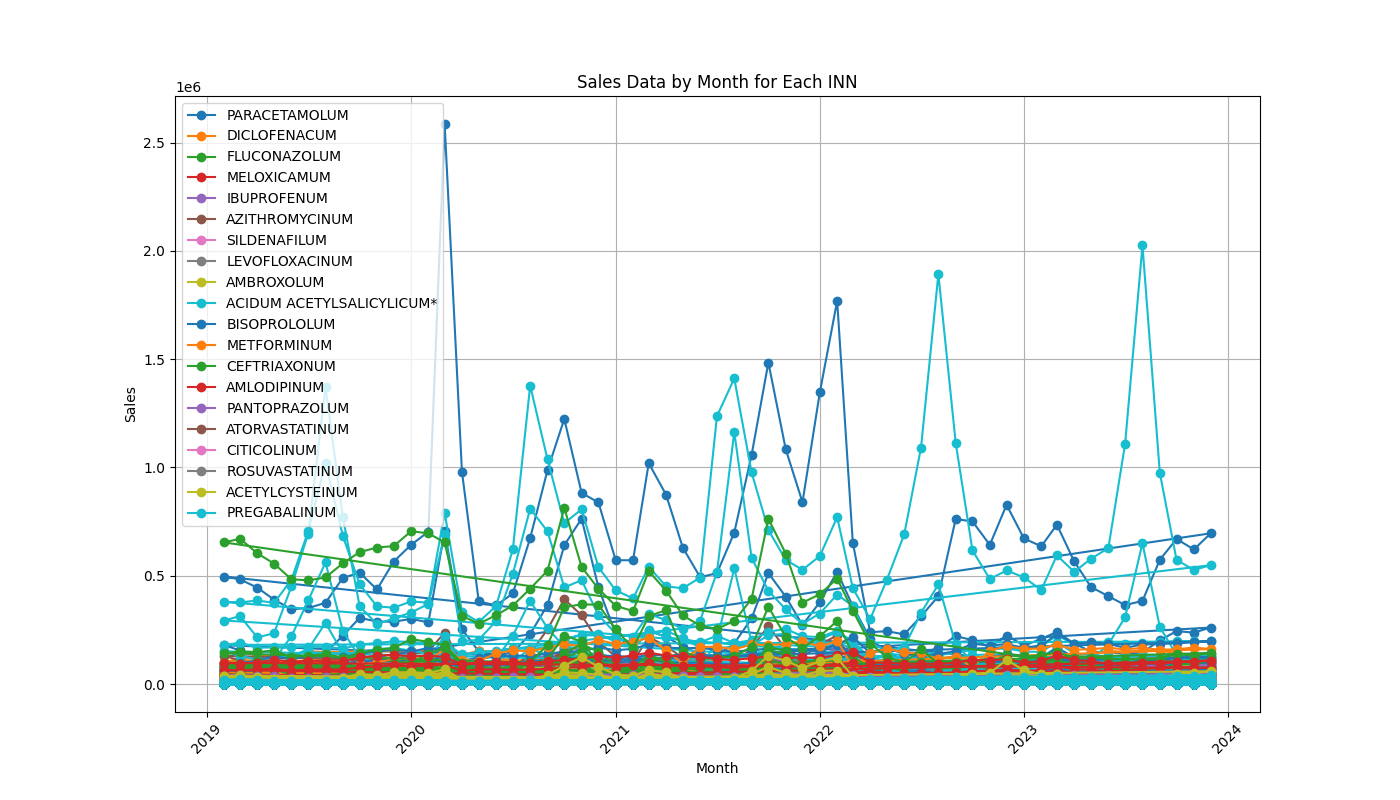
As observed from the correlation heatmap, the majority of the molecules have a correlation coefficient above 0.8. This indicates that their sales trends have been highly similar over the past five years. The graph below shows the monthly sales of the two molecules with the highest correlation: CEFTIRAXONUM and LEVOFLOXACINUM. Both of these molecules are antibiotics and often used to treat similar types of infections, which likely contributes to their similar sales patterns over time. Additionally, LEVOFLOXACINUM was widely used during the COVID-19 pandemic to treat secondary bacterial infections in COVID-19 patients. This, along with the increased need for antibiotics like CEFTIRAXONUM during the pandemic, has influenced the sales trends of both molecules during the year 2021 - 2022, during which the virus was actively circulating.



We also plotted the mean sales for each month to observe any seasonal trends or patterns in the data. The graph below shows the mean sales for each month, providing insights into how sales fluctuate throughout the year. Sales are highest during the months of January through March, which are typically when people experience the highest incidence of illness.



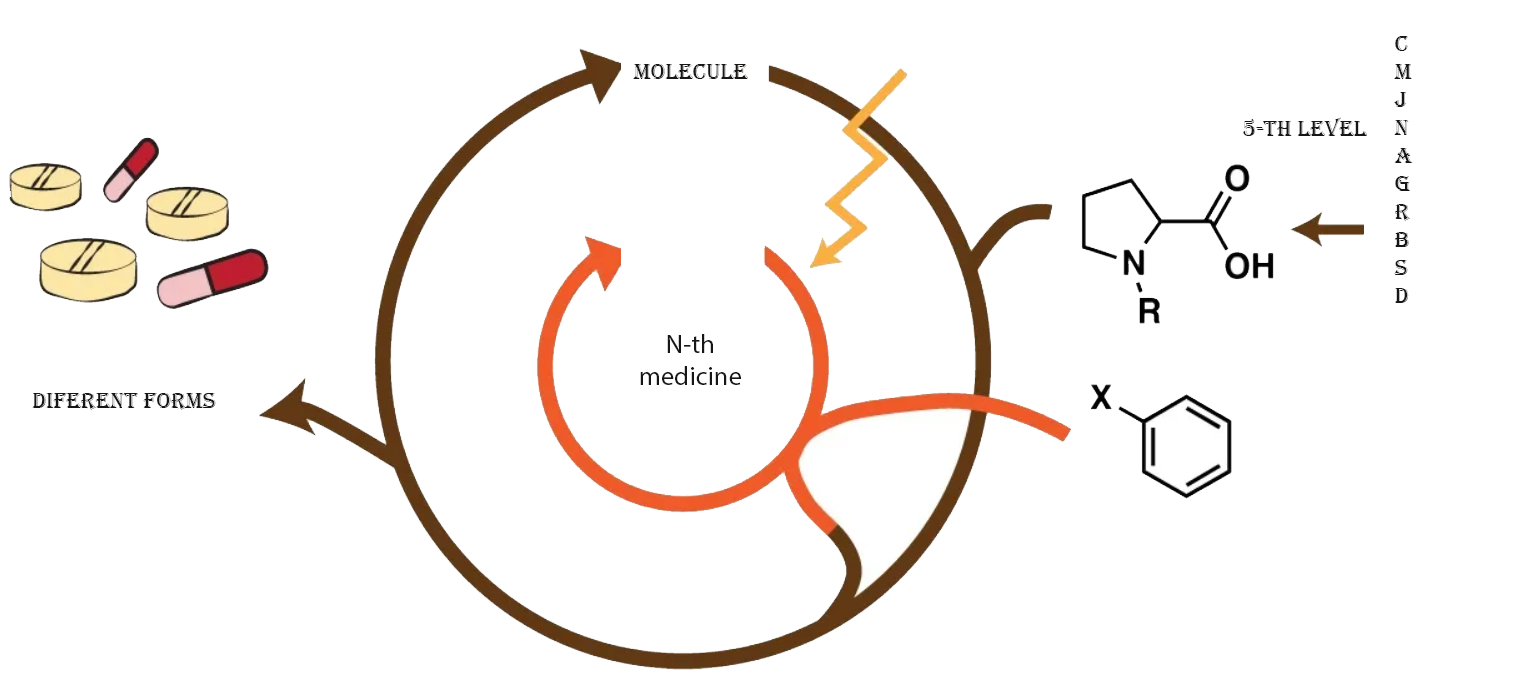
The graph below displays the monthly sales data for various molecules from 2019 to 2023. Certain molecules exhibit clear seasonal trends, with regular peaks in sales indicating higher demand during specific times of the year. Notably, there are significant sales spikes for molecules like PRACETAMOLUM and LEVOFLOXACINUM around 2020 and 2021, reflecting their increased use during the COVID-19 pandemic for treating symptoms and secondary infections.



*Figure 1 Sales for the period 02.2019-12.2023*

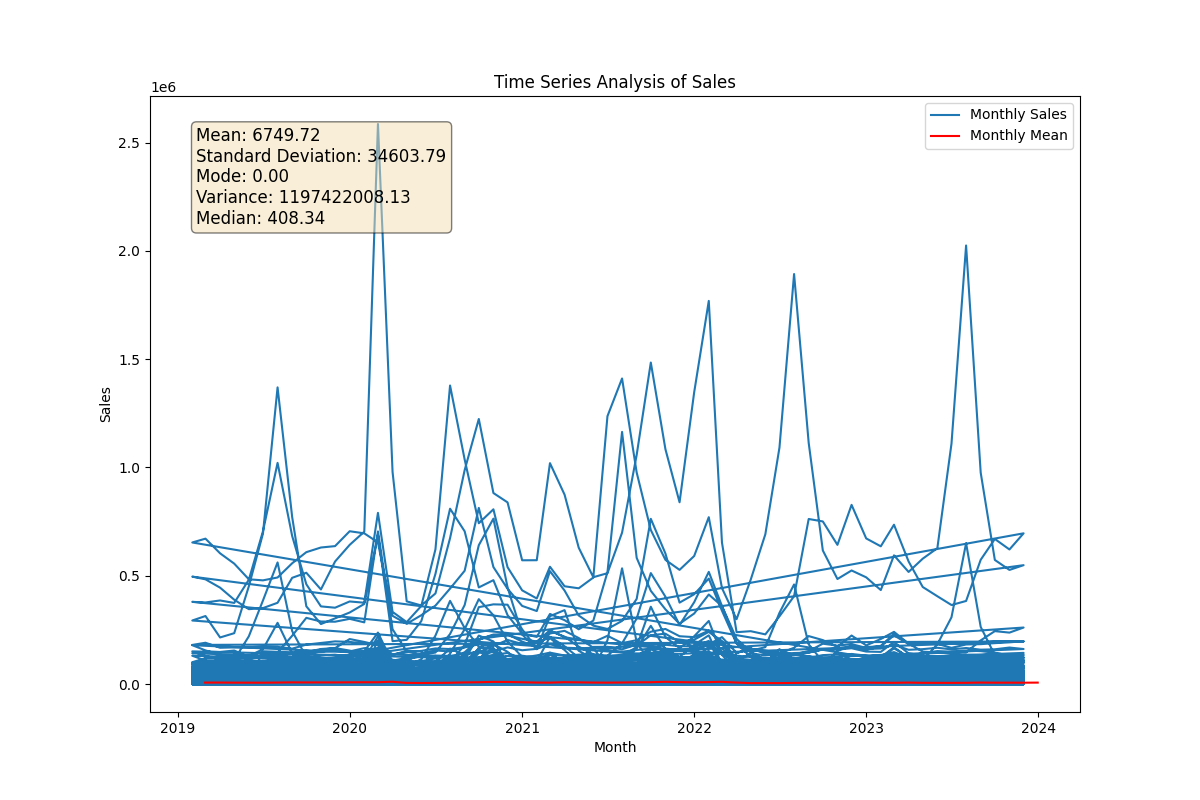
*Figure 2 Mean sales per month*

Sales by individual formulas for the same period can be seen in the charts below.



*Figure 12 Data for one molecule*

* Product specifics:
* A molecule that contains an active substance.
* A dosage form that contains the active substance of the molecule
* A description that has information about packaging, tablets, syrup, spray or other type of aggregate state. This data is not structured and the text string.



## Data Preprocessing:

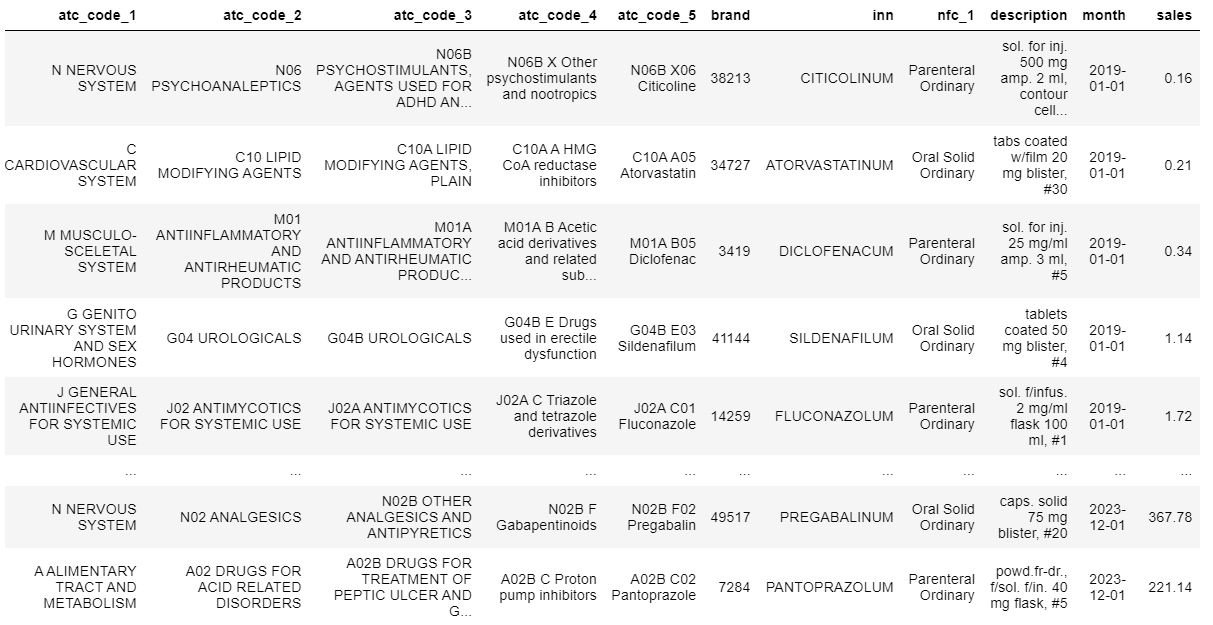
The data preparation process included several critical steps to ensure the dataset was cleaned, transformed, and normalized appropriately. All data preprocessing and machine learning techniques for forecasting were conducted using the Python programming language. The datasets were imported as data frames using the *pandas* library.

A significant portion of the sales data for each medication in the original dataset had zero values. After consulting with the Associate Director of the Data Science team at Teva Pharmaceuticals, we learned that these zeros often resulted from supply chain issues. Such problems typically take less than six months to resolve. Based on this insight, we removed all observations where sales amounts were zero for more than 12 consecutive months.

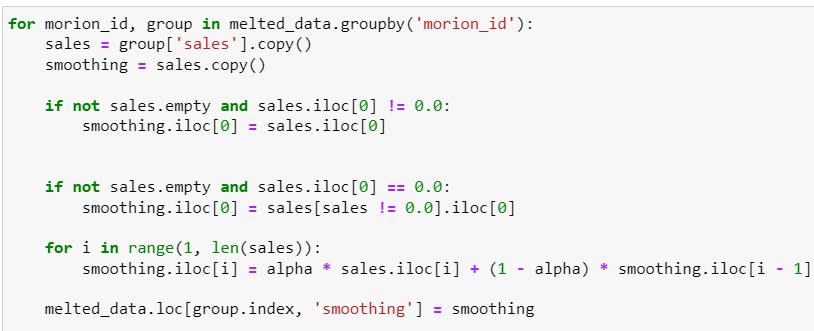
Each medication in the dataset is identified by a unique Morion ID, which includes every form, concentration, etc. Some of these IDs were recorded with a minus sign. This anomaly was discussed during a domain knowledge meeting, and since no particular reason was provided for this irregularity, those observations were also removed from the data frame.

To ensure consistency and ease of use, the columns are renamed using a predefined dictionary.



The dataset is transformed from a wide format to a long format for better handling of time series data.   


Exponential smoothing is applied to the sales data to smooth out fluctuations and highlight trends. This technique helps in better forecasting by giving more weight to recent observations.



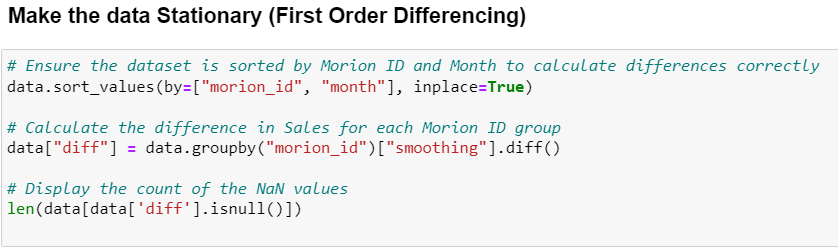
Various features such as concentration, drug size, pack size, and form are extracted from the description column using regular expressions. This step is crucial for understanding the product characteristics and their impact on sales.



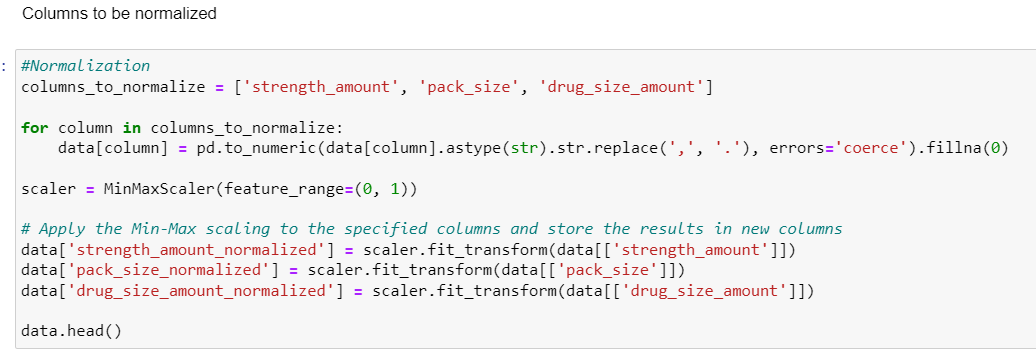
Categorical variables are encoded using LabelEncoder and normalized using MinMaxScaler. This step ensures that the categorical data is in a numerical format suitable for machine learning algorithms.



To make the data stationary, first-order differencing is applied to the smoothed sales data. Stationarity is a key requirement for many time series forecasting models.

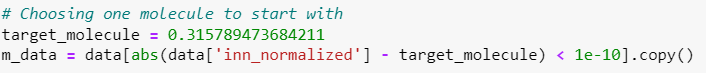


Numerical columns are normalized using MinMaxScaler. Normalization ensures that the data is scaled to a standard range, improving the performance of machine learning models.

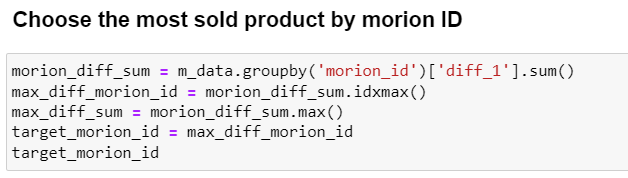


## Feature Selection/Engineering:

In order to test our workflow and identify potential areas for improvement or optimization, we integrated the entire process using a single molecule (INN - 0.315789473684211).

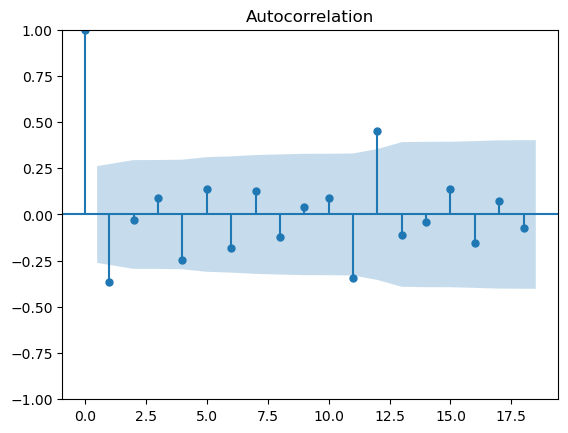


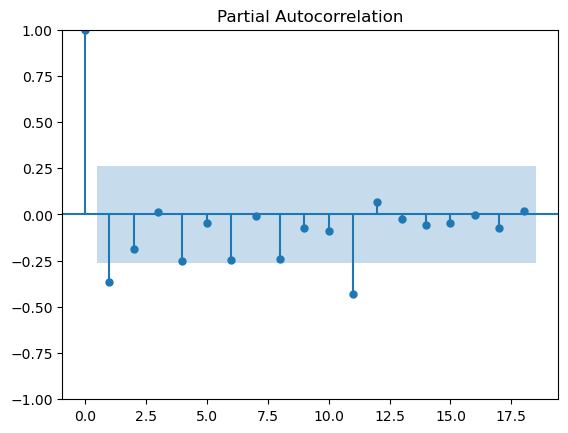
We identified the product with the highest sales, in terms of the differenced values, for further analysis. This involved grouping the data by morion\_id and summing up the differenced sales values (diff\_1). The Morion ID with the highest sum was chosen as the target:



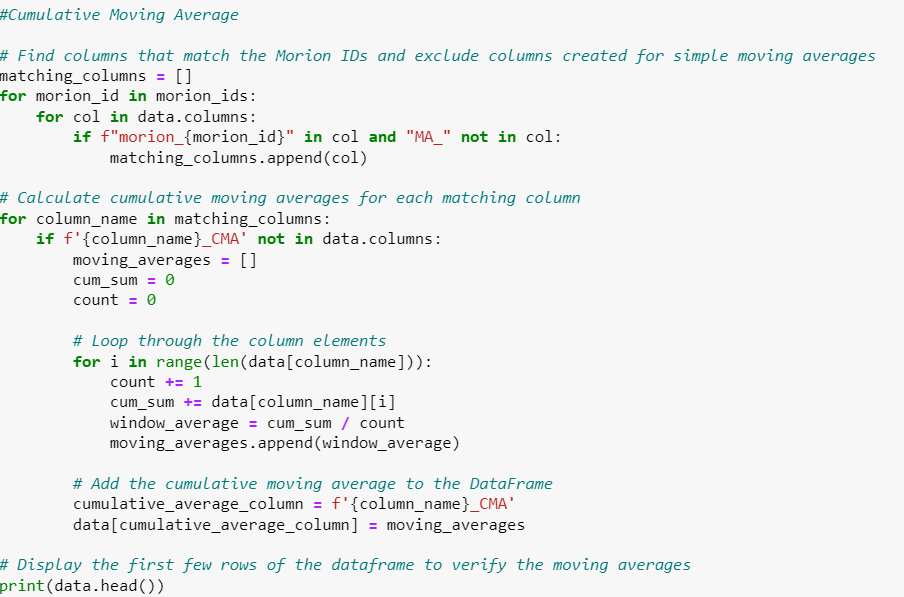
To facilitate the analysis, we transposed the data so that each Morion ID became a separate column. This structure allows for a more straightforward comparison and manipulation of sales data across different products.

We plotted the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) to identify potential lags for the SARIMA model. These plots help in determining the order of the AR (AutoRegressive) and MA (Moving Average) components of the model. The ACF plot shows the correlation of the time series with its own lagged values. The PACF plot shows the correlation of the time series with its own lagged values, controlling for the values of the time series at all shorter lags.





We calculated the Simple Moving Average (SMA) and Cumulative Moving Average (CMA) for the drug with Morion ID - 112504. SMA helps in smoothing out short-term fluctuations and highlighting longer-term trends or cycles. CMA is a type of moving average that calculates the average of all previous data points up to the current point, providing a long-term perspective on trends.



The final step in our analysis involves calculating the Variance Inflation Factor (VIF) to identify and remove multicollinear variables. Multicollinearity occurs when two or more predictors in the model are highly correlated, which can inflate the variance of the coefficient estimates and make the model unstable. VIF helps in detecting multicollinearity by quantifying how much the variance is inflated.We executed the VIF calculation function on the prepared dataset. The function iteratively removed columns with VIF above the threshold, which was set to 10.

The variables which remained in our dataset after the VIF exclusion are the predictors which we used in our model. Once again, we predicted the most sold drug from the molecule BISOPROLOLUM which are beta blockers for the cardiovascular system.

The internal factors fed in the model included the brand, atc\_code\_5, strength, size, form and pack size of the drug.

The model included several internal factors such as the brand, ATC code 5, strength, size, form, and pack size of the drug. Additionally, we considered external factors, specifically the sales of other drugs containing the same active molecules. This analysis helps us determine if the sales of these similar drugs impact the target drug (Morion ID = 112504) and assess their interchangeability. Essentially, we want to know if another drug can replace the target drug if it goes out of stock.

## SARIMAX Model Implementation

To solve the current time series case we used the SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors) model. It was chosen for its ability to handle seasonal data and incorporate exogenous variables that might influence the target variable. It fits our case problem accurately since our sales data has a clearly visible seasonality and our data includes many factors which are correlated to the sales.

In order to use the model, our data needs to be stationary. Stationarity refers to the statistical properties of a time series remaining constant over time, such as constant mean, constant variance, and constant autocovariance. You can use the Dickey-Fuller test for this.

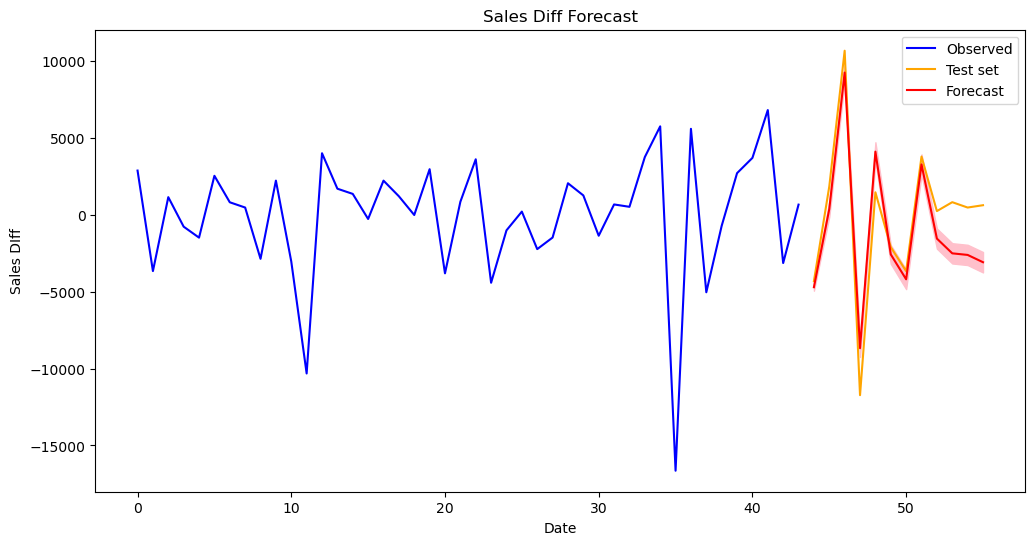
Perform the ADF Test: Inside the function, it calls the adfuller function on the timeseries. The autolag=’AIC’ parameter specifies that the lag order should be chosen based on the Akaike Information Criterion (AIC). Retrieve the p-Value: It extracts the p-value from the ADF test result, which is stored in the variable p\_value. Print the Results: It prints the ADF Statistic (result[0]), the p-value (p\_value), and a statement indicating whether the time series is stationary or non-stationary based on the p-value. If the p-value is less than 0.05, it considers the series as “Stationary”; otherwise, it’s labeled as “Non-Stationary.”

### Evaluation:

To evaluate the performance of our SARIMAX model, we used the Mean Absolute Percentage Error (MAPE) measures the average magnitude of error produced by a model, or how far off predictions are on average. The MAPE for our forecast was calculated as 2.25%, meaning that the average absolute percentage difference between our model's predictions and the actual sales values is 2.25%. This low MAPE value suggests that the model performs well in terms of accuracy, providing reliable sales forecasts for our target medicament.

Due to time constraints, we couldn’t run the model for drugs with other active molecules. There was a recommendation from the company, based on their previous analyses, that each molecule should have a separate predicting model to ensure accurate forecasts. This will be our next step: to extend the forecasting to include other molecules and develop a comprehensive, automated forecasting process for the supply chain of Teva Pharmaceuticals.

### Prediction:



The plot shows the observed sales differences (blue), test set (orange), and forecasted values (red) for a pharmaceutical product. The SARIMAX model captures the overall trend very accurately, but the confidence interval (pink shaded area) indicates some uncertainty in the forecast. While the model follows the general sales trend, high volatility in the data leads to prediction errors, suggesting a need for further refinement to improve accuracy.

In conclusion, while the SARIMAX model provides a useful forecast that captures the general trend of sales differences, there is room for improvement in accurately predicting the magnitude of sales fluctuations. The model's performance, reflected in the MAE and MSE, indicates that further refinement and potentially more sophisticated modeling techniques may be needed to improve forecast accuracy.

There is room for improving the forecast, which includes increasing the number of relevant factors or variables, such as accounting for the effects of COVID-19. Additionally, implementing rolling window validation of the model can provide more robust performance metrics. Exploring hierarchical models like HTS Prophet, Bayesian Hierarchical Models, and XGBoost could offer better accuracy and efficiency. Comparing these methods in terms of accuracy, efficiency, running time, and complexity of implementation would help identify the most suitable approach for improving the forecast. By incorporating these enhancements, the model could better handle volatility and provide more reliable sales predictions, ultimately supporting better decision-making for inventory management and strategic planning.