# Final Project

# Pharmaceutical Sales Forecasting

**California State University East Bay**

**BAN 673-02**

**Prof. Zinovy Radovilsky**

## Group Members:


**Sonali Akshay Pandey**

**Hanna Khan**

**Likhitha Thunam**

**Ishraque Shahriar**

**Amarjeet**

Summary

For this project, transactional data from a pharmacy's Point-of-Sale system, sourced from Kaggle, was used to forecast pharmaceutical sales for the upcoming year. The dataset spans six years (2014-2019) and includes detailed records of sales, specifically focusing on M01AB drugs classified under various Anatomical Therapeutic Chemical (ATC) categories. The time series analysis revealed significant trends and seasonal patterns, with significant fluctuations in drug sales across different periods—typically higher sales in the winter months and lower in the summer, reflecting seasonal health trends.

Various forecasting models were employed, including regression-based models, Two-level models - with regression and trailing MA for residuals and AR(1) model for residuals, advanced exponential smoothing, and Autoregressive Integrated Moving Average (ARIMA) models. Enhancements such as trailing moving averages for residuals and autoregressive models for residuals were applied to refine the accuracy of the regression model. The evaluation of these models was based on their Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), ensuring robustness in predictive performance.

The analysis clearly demonstrates that sales benefit significantly from the advanced forecasting techniques applied, with significant autocorrelation across all analyzed time lags. The best-performing model was a two-level model that combined regression with trend and seasonality and a trailing moving average for residuals. This model, particularly when applied to the finely segmented monthly data, provided the most accurate forecasts, showcasing its efficacy in handling the complex dynamics of pharmaceutical sales.

# Introduction

Pharmaceutical sales are a vital component of the healthcare system, significantly impacting both the economy and public health. The dataset used in this analysis spans over a period of six years (2014-2019) and detailing transactions of 57 specific pharmaceutical drugs. These drugs are classified under various categories according to the Anatomical Therapeutic Chemical (ATC) Classification System. This report will focus specifically on the category M01AB, which includes anti-inflammatory and antirheumatic products, non-steroids, Acetic acid derivatives and related substances, providing a detailed analysis of trends and forecasting in this segment.

The pharmaceutical industry is highly significant, contributing to various economic and healthcare sectors. It not only supports large-scale manufacturing and research and development activities but also has a profound impact on the healthcare services sector, affecting everything from healthcare providers to insurance services and patient care.

Pharmaceutical sales are influenced by a range of factors, including seasonal health trends, regulatory changes, advancements in drug development, and shifts in healthcare policies. For example, sales volumes for certain medications peak during flu season, reflecting higher demand. Moreover, as healthcare becomes more personalized and targeted, the pharmaceutical sales landscape continues to evolve, driven by innovation and consumer health awareness.

The goal of this project is to harness advanced time series forecasting techniques to predict future drug sales, aiding the pharmacy in inventory management, strategic planning, and operational adjustments. By analyzing historical sales data, the project seeks to uncover underlying patterns and trends that can inform more accurate sales forecasts. This analysis is crucial for the pharmacy's ability to respond proactively to market demands and ensure optimal stock levels, thereby minimizing waste and maximizing the availability of essential medications.

## Eight Steps of Forecasting

### Step 1: Define Goal

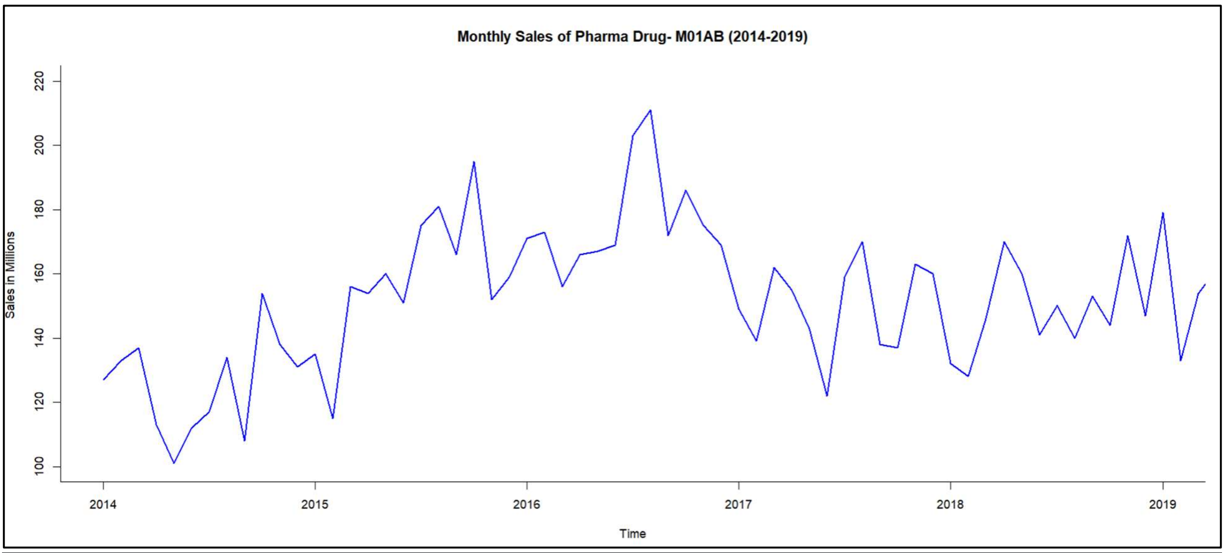
The objective of this project is to forecast monthly sales for the M01AB category of anti-inflammatory drugs, using historical data from 2014 to 2019. The goal is to identify a predictive model that accurately captures seasonal trends and overall sales patterns. The most effective model will be selected based on its accuracy and reevaluated biannually using the latest semi-annual data to ensure ongoing relevance.

Forecasts from this model will help optimize pharmacy inventory management by aligning drug supply with projected demand. The analysis was performed using R, which was chosen for its robust data processing and time series analysis capabilities.

### Step 2: Get data

This analysis utilizes a dataset obtained from a single pharmacy's Point-of-Sale system, detailing monthly sales from January 2014 to October 2019. The dataset focuses specifically on pharmaceutical drugs classified under the Anatomical Therapeutic Chemical (ATC) Classification System. For the purpose of this project, the emphasis will be on the M01AB category—anti-inflammatory and antirheumatic products, non-steroids, Acetic acid derivatives and related substances.

### Step 3: Explore and Visualize Series




The line chart above displays the monthly sales data of pharmaceutical drugs in the M01AB category (Anti-inflammatory and antirheumatic products, non-steroids, Acetic acid derivatives and related substances) from 2014 to 2019. This visualization captures both the trend and seasonality in the sales data over the six-year period. Several observations can be made from this plot:

**Trend:** Sales volumes fluctuate noticeably, but overall, the trend does not show a consistent upward or downward trajectory over the years. Instead, sales exhibit cyclical patterns that suggest a strong seasonal component.

**Seasonality:** Sales peaks and troughs appear regularly, indicating a clear seasonal pattern. Higher sales are typically observed during the colder months, likely due to the increased incidence of inflammatory conditions during these times, while lower sales occur during the warmer months.

**Cyclic Behavior:** Beyond the basic seasonality, the sales show multi-year cyclical trends, with broader peaks and valleys suggesting varying market conditions or external influences on drug demand.

From this analysis, it is evident that the sales data for M01AB category drugs are influenced by both seasonal changes and broader market dynamics. Understanding these patterns will be crucial for effective inventory management and forecasting future sales trends.

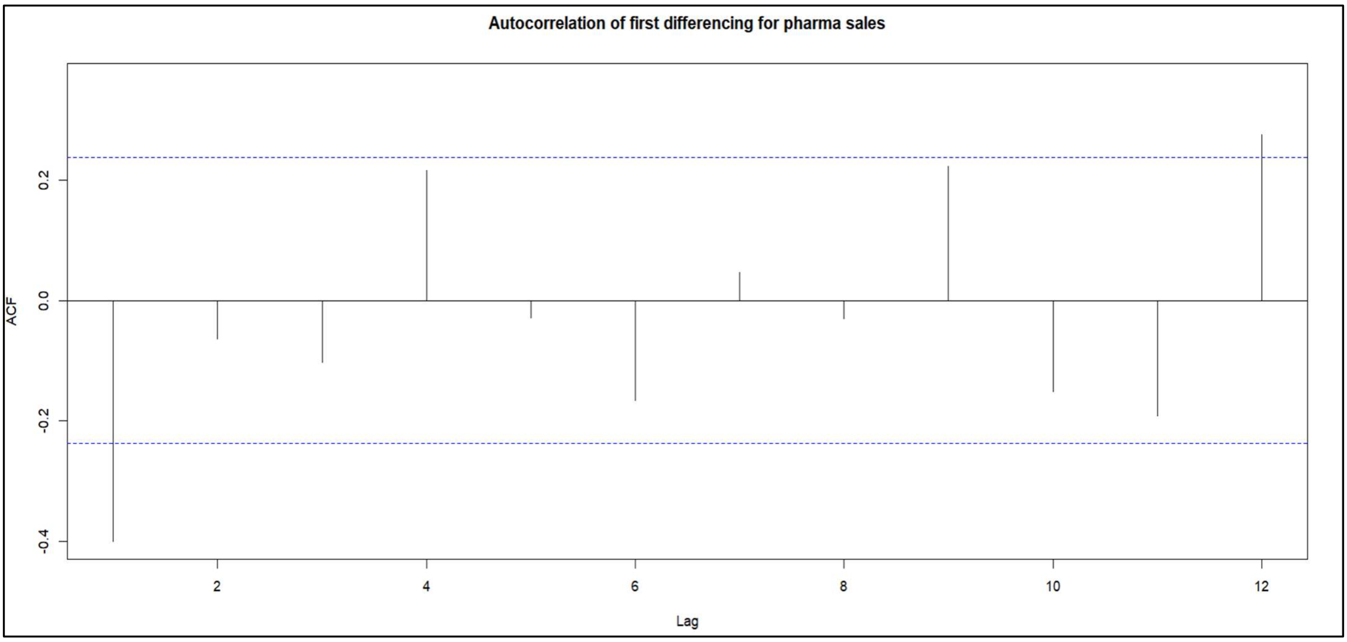
**Evaluating Predictability:**

Approach 1: Hypothesis Testing

```
> # Apply z-test to test the null hypothesis that beta
> # coefficient of AR(1) is equal to 1.
> ar1 <- 0.5807
> s.e. <- 0.0975
> null_mean <- 1
> alpha <- 0.05
> z.stat <- (ar1-null_mean)/s.e.
> z.stat
[1] -4.300513
> p.value <- pnorm(z.stat)
> p.value
[1] 8.520166e-06
> if (p.value<alpha) {
+   "Reject null hypothesis"
+ } else {
+   "Accept null hypothesis"
+ }
[1] "Reject null hypothesis"
```

Based on the p-value which is smaller than 0.05, we reject the null hypothesis that $\beta 1 = 1$. Therefore, the time series data for pharma M01AB sales, according to this test, is predictable and not a random walk.

Approach 2: Autocorrelation plot for first differenced data




The autocorrelation plot displayed above represents the first differencing of monthly pharmaceutical sales data in the M01AB category. Significant autocorrelations at lags 1 and 12, which cross the level of significance (horizontal blue threshold lines) indicating 95% confidence intervals, point to dependencies at these intervals. The negative autocorrelation at lag 1, crossing the lower confidence threshold, suggests the existence of a trend component in the original series. Conversely, a notable positive autocorrelation at lag 12, exceeding the upper threshold, indicates a strong seasonal pattern in the original series. This confirms that besides the level component, the data exhibits significant trend and seasonal components crucial for developing an accurate forecasting model.

**Step 4: Data Preprocessing**

For this project, several preprocessing steps were undertaken to prepare the pharmaceutical sales dataset for analysis:

**Column Selection:** Initially, the dataset included various drug sales categories. To focus our analysis on the M01AB category—anti-inflammatory and antirheumatic products, non-steroids, Acetic acid derivatives, and related substances—we deleted columns containing data for other drug categories from the dataset (.csv file). This step ensured that our dataset solely consisted of the relevant data needed for forecasting M01AB sales.

**Data Conversion:** The sales figures in the M01AB column were originally in a non-numeric format. To facilitate mathematical operations and statistical analysis, we converted these sales figures into numerical values.

**Handling Missing Data:** During the preprocessing, it was identified that sales data for one specific period, 1/31/2017, was missing. To address this gap, we implemented a method called imputation. Specifically, we calculated the average of the sales data points from January 2017 and applied this average value to the missing date. This approach helped maintain the integrity and continuity of the dataset, allowing for more accurate forecasting.

**Step 5: Partition Series**

For the time series analysis of pharmaceutical sales, the dataset was carefully split into two parts to allow for effective training and validation of forecasting models. These partitioned validation and training data sets are shown in figures 1 and 2 of the appendix.

**Training Data:** This set includes 59 records from the dataset, which will be used to build and calibrate the forecasting models. This portion represents the vast majority of the available data, providing a robust basis for model development.

**Validation Data:** Comprising the last 10 records of the dataset, this segment is reserved for testing the model's performance. This partition will help to assess the model's predictive accuracy with new records.

**Step 6 & 7: Apply Forecasting & Comparing Performance**

In our comprehensive analysis of pharmaceutical sales forecasting for the M01AB category, we employed a variety of statistical models to ensure robustness in our predictions. The models tested included:

| Models Utilized | RMSE | MAPE |
|---|---|---|
| Linear regression with trend and seasonality | 18.959 | 10.222 |
| Linear regression with trend | 20.928 | 11.483 |
| Linear regression with seasonality | 20.121 | 11.029 |
| Holt- Winter's Method | 17.247 | 9.348 |
| Linear regression with trend and seasonality + trailing MA for residuals | 9.803 | 5.403 |
| Linear regression with trend and seasonality + AR(1) model for residuals | 14.582 | 8.149 |
| Seasonal ARIMA (1,1,1)(1,1,1) Model | 14.487 | 6.571 |
| Auto ARIMA Model | 16.533 | 8.955 |

Each model was rigorously evaluated based on its Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), two critical metrics for assessing forecasting accuracy. After thorough evaluation, we identified that the four models demonstrating the highest accuracy and reliability were:

1. **Linear regression with trend and seasonality + trailing MA for residuals** (RMSE: 9.803, MAPE: 5.403) — This model yielded the best performance, showcasing strong predictive capabilities with the lowest error rates.
2. **Linear regression with trend and seasonality + AR(1) model for residuals** (RMSE: 14.582, MAPE: 8.149) — This method also showed good accuracy, particularly in capturing seasonal fluctuations.
3. **Seasonal ARIMA (1,1,1)(1,1,1) Model** (RMSE: 14.487, MAPE: 6.571) — This model was effective in modeling both seasonal patterns and non-seasonal dynamics in the data.
4. **Auto ARIMA** (RMSE: 16.533, MAPE: 8.955) — Providing a solid balance between both seasonal patterns and non-seasonal dynamics in the data.

These models were selected for detailed reporting due to their superior performance, offering the most reliable insights for strategic decision-making in pharmaceutical sales management. By implementing these models, we can enhance accuracy in predicting future sales, optimize inventory management, and better align supply with projected demand.

Model 1: Two-level forecasting - **Regression Model with trend and seasonality and Training MA for residuals**

For two-level forecasting, we applied trailing MA in data with trend and seasonality, a combination of two forecasting models is applied:

**Level 1** - Regression model with trend and/or seasonality. It is also used to remove trend (de-trending) and/or seasonality (de-seasonalizing) in historical data and identify residuals (errors) – differences between actual sales and regression forecast in respective time periods

**Level 2** - Trailing MA to forecast regression model's residuals (errors)

The total forecast used in predictions is a combination (sum) of regression model and trailing MA forecasts.

Now, for level 1: Model summary with linear trend and seasonality for validation partition is demonstrated below:

```
> summary(train.lin.season)

Call:
tslm(formula = train.ts ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.035 -15.730  -1.035  14.700  43.800 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 134.2946    11.0471  12.157 5.75e-16 ***
trend         0.3402     0.1761   1.932   0.0595 .  
season2      -5.5402    14.3304  -0.387   0.7008    
season3       7.9196    14.3336   0.553   0.5833    
season4       7.7793    14.3390   0.543   0.5901    
season5       2.0391    14.3466   0.142   0.8876    
season6      -5.5011    14.3563  -0.383   0.7033    
season7      15.9587    14.3682   1.111   0.2725    
season8      22.0185    14.3822   1.531   0.1326    
season9       1.8783    14.3983   0.130   0.8968    
season10     17.3380    14.4166   1.203   0.2353    
season11     13.7978    14.4370   0.956   0.3442    
season12     10.2489    15.2240   0.673   0.5042    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.66 on 46 degrees of freedom
Multiple R-squared:  0.229,	Adjusted R-squared:  0.02793 
F-statistic: 1.139 on 12 and 46 DF,  p-value: 0.3541
```

The regression model includes trend and 11 seasonal dummy variables for Season 2 in February through season 12 in December.

The regression equation will be:

$$y_t = 134.2946 + 0.3402\, t - 5.5402\, D2 + 7.9196\, D3 + \dots + 10.2489\, D12$$

Below is the point forecast for monthly sales in the validation period:

```
> train.lin.season.pred
         Point Forecast    Lo 0    Hi 0
Dec 2018       164.9565 164.9565 164.9565
Jan 2019       155.0478 155.0478 155.0478
Feb 2019       149.8478 149.8478 149.8478
Mar 2019       163.6478 163.6478 163.6478
Apr 2019       163.8478 163.8478 163.8478
May 2019       158.4478 158.4478 158.4478
Jun 2019       151.2478 151.2478 151.2478
Jul 2019       173.0478 173.0478 173.0478
Aug 2019       179.4478 179.4478 179.4478
Sep 2019       159.6478 159.6478 159.6478
```

The trailing MA forecast for residuals in the validation period is presented below:

```
> ma.trail.res
            Jan         Feb         Mar         Apr         May         Jun         Jul         Aug         Sep
2014                          -3.4347826 -11.0347826 -24.5681159 -28.7681159 -30.5014493 -26.5014493 -30.6347826
2015 -11.7260870 -13.2869565  -4.5173913  -1.1173913  11.0159420  13.4826087  17.4159420  17.4159420  19.6159420
2016  10.1913043  23.2971014  22.7333333  18.1333333  13.2666667  21.7333333  31.0000000  38.6666667  36.8666667
2017   9.7753623   3.8811594   1.9840580   1.0507246  -0.4826087  -9.6826087 -11.4159420  -9.4159420  -6.8826087
2018  -6.9739130 -12.5347826 -16.7652174  -7.0318841   0.7681159   3.2347826  -6.4985507 -20.1652174 -18.9652174
            Oct         Nov         Dec
2014 -19.1014493 -15.3681159 -10.8318841
2015  25.4826087  18.2159420  12.7521739
2016  30.4000000  20.8000000  16.6695652
2017 -15.0159420 -14.9492754 -10.7463768
2018 -21.7652174  -8.6985507
```

- Now, we first Fit the regression model with linear trend and seasonality for **entire data set**. Then we created regression forecast for future 12 periods.
- We identified regression residuals for entire data set and then used trailing MA to forecast residuals for entire data set. We then created forecast for trailing MA residuals for future 12 periods.
- After this we developed 2-level forecast for future 12 periods by combining regression forecast and trailing MA for residuals for future 12 periods.

Model summary with linear trend and seasonality for entire dataset is demonstrated below:

```
Call:
tslm(formula = sales.ts ~ trend + season)

Residuals:
   Min     1Q Median     3Q    Max 
-38.61 -13.82  -0.22  14.88  43.54 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 138.2727     9.4646  14.609   <2e-16 ***
trend         0.3407     0.1281   2.660   0.0102 *  
season2     -12.3407    12.1510  -1.016   0.3142    
season3       2.3187    12.1530   0.191   0.8494    
season4       3.3113    12.1564   0.272   0.7863    
season5      -0.3627    12.1611  -0.030   0.9763    
season6      -9.5367    12.1672  -0.784   0.4365    
season7      13.2893    12.1746   1.092   0.2797    
season8      18.2820    12.1834   1.501   0.1391    
season9      -1.8920    12.1935  -0.155   0.8772    
season10     13.3447    12.7492   1.047   0.2997    
season11      9.8040    12.7537   0.769   0.4453    
season12      2.6633    12.7595   0.209   0.8354    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.05 on 56 degrees of freedom
Multiple R-squared:  0.2652,	Adjusted R-squared:  0.1077 
F-statistic: 1.684 on 12 and 56 DF,  p-value: 0.0954
```

The regression model includes trend and 11 seasonal dummy variables for Season 2 in February through season 12 in December.

The regression equation will be:

$$y_t = 138.27 + 0.34\,t - 12.34\,D2 - 2.3187\,D3 + \ldots + 2.6633\,D12$$

Below is the point forecast for monthly sales in the entire dataset set:

```
> tot.trend.seas.pred
         Point Forecast     Lo 0     Hi 0
Oct 2019       175.4640 175.4640 175.4640
Nov 2019       172.2640 172.2640 172.2640
Dec 2019       165.4640 165.4640 165.4640
Jan 2020       163.1413 163.1413 163.1413
Feb 2020       151.1413 151.1413 151.1413
Mar 2020       166.1413 166.1413 166.1413
Apr 2020       167.4747 167.4747 167.4747
May 2020       164.1413 164.1413 164.1413
Jun 2020       155.3080 155.3080 155.3080
Jul 2020       178.4747 178.4747 178.4747
Aug 2020       183.8080 183.8080 183.8080
Sep 2020       163.9747 163.9747 163.9747
```

The trailing MA forecast for residuals in the training partition is presented below:

```
> tot.ma.trail.res.pred
         Point Forecast     Lo 0     Hi 0
Oct 2019       3.002178 3.002178 3.002178
Nov 2019       3.002178 3.002178 3.002178
Dec 2019       3.002178 3.002178 3.002178
Jan 2020       3.002178 3.002178 3.002178
Feb 2020       3.002178 3.002178 3.002178
Mar 2020       3.002178 3.002178 3.002178
Apr 2020       3.002178 3.002178 3.002178
May 2020       3.002178 3.002178 3.002178
Jun 2020       3.002178 3.002178 3.002178
Jul 2020       3.002178 3.002178 3.002178
Aug 2020       3.002178 3.002178 3.002178
Sep 2020       3.002178 3.002178 3.002178
```

Below table presents regression forecast, trailing MA for residuals, and total forecast for future 12 periods.

```
> future12.df
   Regression.Fst MA.Residuals.Fst Combined.Fst
1         175.464            3.002      178.466
2         172.264            3.002      175.266
3         165.464            3.002      168.466
4         163.141            3.002      166.144
5         151.141            3.002      154.144
6         166.141            3.002      169.144
7         167.475            3.002      170.477
8         164.141            3.002      167.144
9         155.308            3.002      158.310
10        178.475            3.002      181.477
11        183.808            3.002      186.810
12        163.975            3.002      166.977
```

Accuracy measures:

| RMSE | | MAPE | |
|---|---|---|---|
| **Validation Partition** | **Entire Dataset** | **Validation Partition** | **Entire Dataset** |
| 14.633 | 9.803 | 7.248 | 5.403 |

Plot for original pharma M01AB sales time series data, regression model and regression forecast for future 12 periods is figure 3 in the appendix.

Model 2: Two-level forecasting - **Regression Model with trend and seasonality and AR(1) Model for residuals**

For two-level forecasting, we applied Auto Regressive model AR1 in data with trend and seasonality, a combination of two forecasting models is applied:

**Level 1** - Regression model with trend and/or seasonality. It is also used to remove trend (de-trending) and/or seasonality (de-seasonalizing) in historical data and identify residuals (errors) – differences between actual sales and regression forecast in respective time periods

**Level 2** - AR1 to forecast regression model's residuals (errors)

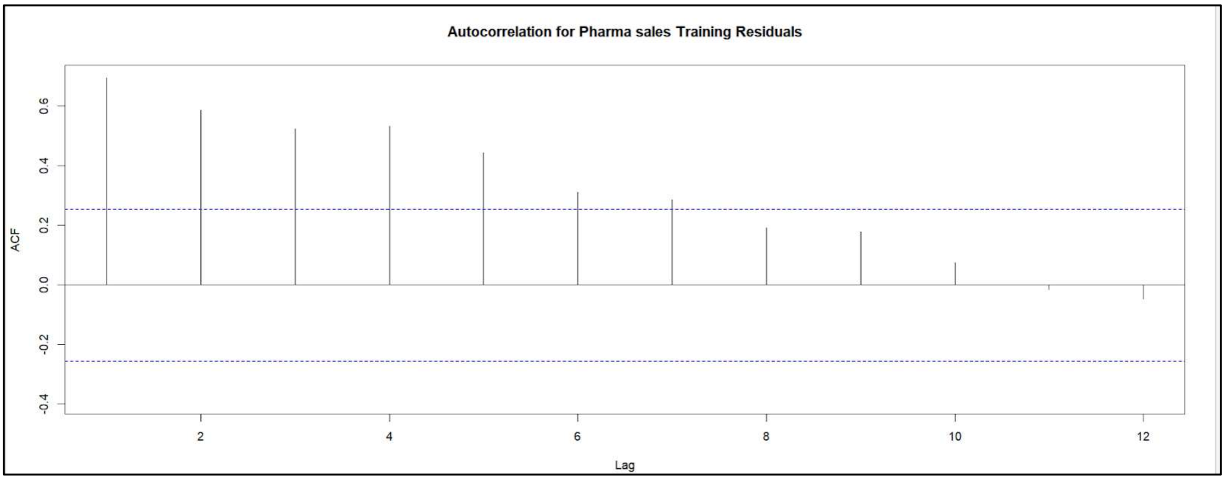
The total forecast used in predictions is a combination (sum) of the regression model and AR1 forecasts.

Now, for level 1: The model with linear trend and seasonality for Validation Partition is demonstrated below:

```
> summary(train.lin.season)

Call:
tslm(formula = train.ts ~ trend + season)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.035 -15.730  -1.035  14.700  43.800 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 134.2946    11.0471  12.157 5.75e-16 ***
trend         0.3402     0.1761   1.932   0.0595 .  
season2      -5.5402    14.3304  -0.387   0.7008    
season3       7.9196    14.3336   0.553   0.5833    
season4       7.7793    14.3390   0.543   0.5901    
season5       2.0391    14.3466   0.142   0.8876    
season6      -5.5011    14.3563  -0.383   0.7033    
season7      15.9587    14.3682   1.111   0.2725    
season8      22.0185    14.3822   1.531   0.1326    
season9       1.8783    14.3983   0.130   0.8968    
season10     17.3380    14.4166   1.203   0.2353    
season11     13.7978    14.4370   0.956   0.3442    
season12     10.2489    15.2240   0.673   0.5042    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.66 on 46 degrees of freedom
Multiple R-squared:  0.229,	Adjusted R-squared:  0.02793 
F-statistic: 1.139 on 12 and 46 DF,  p-value: 0.3541
```

Regression Model Forecast for pharma M01AB sales in the validation period:

```
> train.lin.season.pred
         Point Forecast    Lo 0     Hi 0
Dec 2018       164.9565 164.9565 164.9565
Jan 2019       155.0478 155.0478 155.0478
Feb 2019       149.8478 149.8478 149.8478
Mar 2019       163.6478 163.6478 163.6478
Apr 2019       163.8478 163.8478 163.8478
May 2019       158.4478 158.4478 158.4478
Jun 2019       151.2478 151.2478 151.2478
Jul 2019       173.0478 173.0478 173.0478
Aug 2019       179.4478 179.4478 179.4478
Sep 2019       159.6478 159.6478 159.6478
```



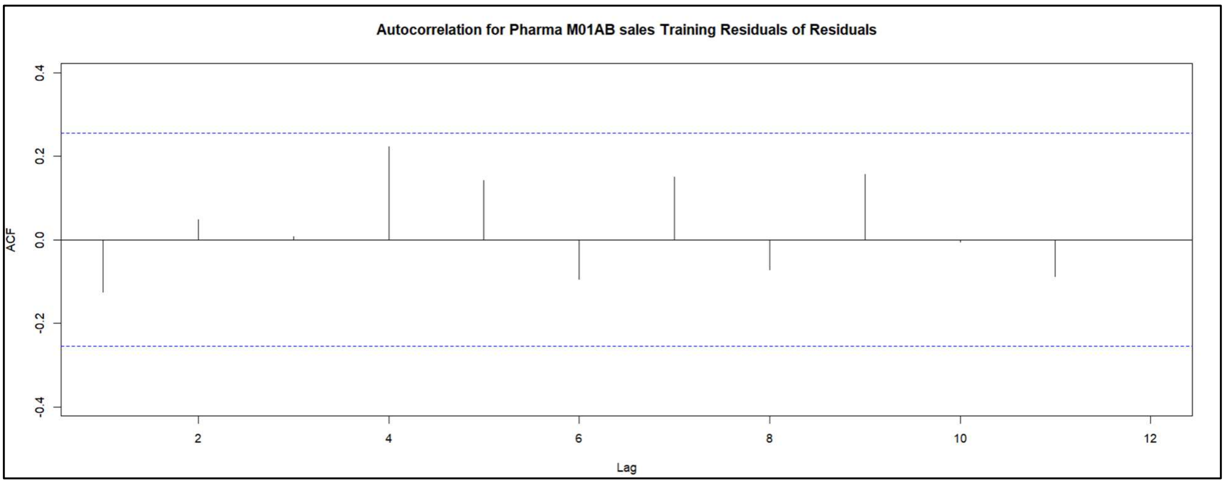

- After developing a *regression model with linear trend and seasonality*, we plot above autocorrelation plot for regression model's residuals with a maximum of 12 lags.
- From the above plot, we observe that for all lags from lag 1 to lag 7, the autocorrelation of residuals is **statistically significant** and these autocorrelations are not incorporated in the regression model.
- Thus, it would be a **good idea to add these autocorrelations of residuals with AR model** to enhance the overall forecast of the model.

Below summary specifies ARIMA(1,0,0) model which is an **AutoRegressive (AR) model of order 1**, with **no differencing (d=0)** and **no moving average components (q=0)**.

```
> summary(res.ar1)
Series: train.lin.season$residuals
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1     mean
      0.6847  -0.1303
s.e.  0.0916   5.7337

sigma^2 = 214.2:  log likelihood = -241.34
AIC=488.68   AICc=489.12   BIC=494.92

Training set error measures:
                    ME    RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set 0.1215911 14.3859 12.15734 -9.850859 194.4802 0.5064405 -0.1245495
```

The *AR(1)* model's equation is:

$$e_t = -0.1303 + 0.6847\, e_{t-1}$$




- Analyzing above plot, we see that all the autocorrelations of AR(1) (residuals of residuals) are **not statically significant** and are **random**.
- Thus, all the relationships(autocorrelation) existing in the historical data pharma sales data are incorporated in the AR(1) model for residuals.

Now, we develop a two-level forecast (regression model with *linear trend and seasonality* and *AR(1)* model for residuals) for the entire data set:

Summary regression model with *linear trend and seasonality* on **entire data set**:

```
> summary(lin.season)

Call:
tslm(formula = sales.ts ~ trend + season)

Residuals:
   Min     1Q Median     3Q    Max 
-38.61 -13.82  -0.22  14.88  43.54 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 138.2727     9.4646  14.609   <2e-16 ***
trend         0.3407     0.1281   2.660   0.0102 *  
season2     -12.3407    12.1510  -1.016   0.3142    
season3       2.3187    12.1530   0.191   0.8494    
season4       3.3113    12.1564   0.272   0.7863    
season5      -0.3627    12.1611  -0.030   0.9763    
season6      -9.5367    12.1672  -0.784   0.4365    
season7      13.2893    12.1746   1.092   0.2797    
season8      18.2820    12.1834   1.501   0.1391    
season9      -1.8920    12.1935  -0.155   0.8772    
season10     13.3447    12.7492   1.047   0.2997    
season11      9.8040    12.7537   0.769   0.4453    
season12      2.6633    12.7595   0.209   0.8354    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.05 on 56 degrees of freedom
Multiple R-squared:  0.2652,	Adjusted R-squared:  0.1077 
F-statistic: 1.684 on 12 and 56 DF,  p-value: 0.0954
```

Below is the summary for AR(1) model for regression residuals: The summary specifies that the model is an **AutoRegressive (AR) model of order 1**, with **no differencing (d=0)** and **no moving average components (q=0)**.

```
> summary(residual.ar1)
Series: lin.season$residuals 
ARIMA(1,0,0) with non-zero mean 

Coefficients:
         ar1     mean
      0.6318  -0.2488
s.e.  0.0913   4.6550

sigma^2 = 219:  log likelihood = -283.07
AIC=572.14   AICc=572.51   BIC=578.84

Training set error measures:
                    ME     RMSE      MAE      MPE     MAPE      MASE       ACF1
Training set 0.1411106 14.58232 12.08606 95.41579 248.1865 0.5366644 -0.1836873
```

The model equation is:

$$e_t = -0.2488 + 0.6318\, e_{t-1}$$

The below table presents regression forecast, AR1 for residuals, and the total forecast for future 12 periods.

```
> table.df
   Reg.Forecast AR(1)Forecast Combined.Forecast
1       175.464         0.612           176.076
2       172.264         0.295           172.559
3       165.464         0.095           165.559
4       163.141        -0.032           163.110
5       151.141        -0.112           151.030
6       166.141        -0.162           165.979
7       167.475        -0.194           167.281
8       164.141        -0.214           163.927
9       155.308        -0.227           155.081
10      178.475        -0.235           178.240
11      183.808        -0.240           183.568
12      163.975        -0.243           163.731
```

Accuracy measures for the model:

| RMSE | | MAPE | |
|---|---|---|---|
| **Validation Partition** | **Entire Dataset** | **Validation Partition** | **Entire Dataset** |
| 12.293 | 14.582 | 6.059 | 8.149 |

Plot for historical data, predictions for historical data and forecast for 12 future periods is presented in figure 4 in the appendix.

Model 3: Seasonal ARIMA Model:

Seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA*(p d, q)* model

*ARIMA (p, d, q) (P, D, Q)[m]* model is used to forecast data with level, trend, and seasonality components. In addition to the *(p, d, q)* parameters, it includes seasonal parameters:

- *P* = order of autoregressive seasonal model *AR(P)* – number of autocorrelation lags included
- *D* = order of differencing in *AR* seasonal model – indicates how many rounds of lag-1 differencing are performed to remove certain trend
- *Q* = order of moving average *MA(Q)* – number of residuals' autocorrelation lags included
- m = number of seasons

In R, *ARIMA(p, d, q)(P, D, Q)* model is defined by Arima() function as *order = c(p, d, q), seasonal = c(P, D, Q)*. Seasonality m is identified by the type of time series data used.

The output for this ARIMA (1,1,1)(1,1,1)[12] model for validation partition is presented below:

```
> summary(train.arima.seas)
Series: train.ts
ARIMA(1,1,1)(1,1,1)[12]

Coefficients:
         ar1      ma1     sar1     sma1
      0.2889  -0.6691   0.1154  -0.5207
s.e.  0.2837   0.2109   0.5285   0.5741

sigma^2 = 375.1:  log likelihood = -200.86
AIC=411.71   AICc=413.21   BIC=420.86

Training set error measures:
                     ME     RMSE      MAE        MPE     MAPE      MASE       ACF1
Training set -0.6648589 16.34032 11.63805 -0.8817577 7.531263 0.4695181 0.01321215
```

| Parameters | Values from summary | Explaination |
|---|---|---|
| p | 1 | Order 1 autoregressive model AR(1) for seasonality |
| d | 1 | Order 1 differencing to remove linear trend |
| q | 1 | Order 1 moving average MA(1) model for error lags |
| P | 1 | Order 1 autoregressive model AR(1) for seasonality |
| D | 1 | First differencing for the seasonal part |
| Q | 1 | Order 1 moving average MA(1) for the seasonal error lags |
| m | 12 | Monthly seasonality |

The model's equation is:

***yt - yt-1 = 0.2889(yt-1 -yt-2) - 0.6691et-1 + 0.1154(yt-1 -yt-13) - 0.5207pt-1***

Using the model's equation, see below the forecast for the validation period:

```
> train.arima.seas.pred
         Point Forecast     Lo 0     Hi 0
Dec 2018       162.3494 162.3494 162.3494
Jan 2019       141.3773 141.3773 141.3773
Feb 2019       135.7565 135.7565 135.7565
Mar 2019       151.6481 151.6481 151.6481
Apr 2019       164.3954 164.3954 164.3954
May 2019       155.9916 155.9916 155.9916
Jun 2019       140.7101 140.7101 140.7101
Jul 2019       158.5176 158.5176 158.5176
Aug 2019       156.6575 156.6575 156.6575
Sep 2019       152.0197 152.0197 152.0197
```

The output for this ARIMA (1,1,1)(1,1,1)[12] model for the **entire data set** is presented below:

```
> summary(arima.seas)
Series: sales.ts
ARIMA(1,1,1)(1,1,1)[12]

Coefficients:
         ar1      ma1    sar1     sma1
      0.0818  -0.6054  0.1872  -0.9876
s.e.  0.2355   0.1851  0.1885   1.5332

sigma^2 = 278.5:  log likelihood = -243.43
AIC=496.87   AICc=498.07   BIC=506.99

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE          ACF1
Training set -0.3154934 14.48734 10.18724 -0.6857992 6.571364 0.4317269 -0.006466028
```

ARIMA (1, 1, 1) (1, 1, 1)[12] means the following:

| Parameters | Values from summary | Explaination |
|---|---|---|
| p | 1 | Order 1 autoregressive model AR(1) for seasonality |
| d | 1 | Order 1 differencing to remove linear trend |
| q | 1 | Order 1 moving average MA(1) model for error lags |
| P | 1 | Order 1 autoregressive model AR(1) for seasonality |
| D | 1 | First differencing for the seasonal part |
| Q | 1 | Order 1 moving average MA(1) for the seasonal error lags |
| m | 12 | Monthly seasonality |

The model's equation is:

***yt - yt-1 = 0.0818(yt-1 -yt-2) - 0.6054et-1 + 0.1872(yt-1 -yt-13) - 0.9876ρt-1***

Below is the point forecast for monthly sales in the entire dataset set :

```
> arima.seas.pred
         Point Forecast    Lo 0     Hi 0
Oct 2019       174.8626 174.8626 174.8626
Nov 2019       178.3546 178.3546 178.3546
Dec 2019       167.5186 167.5186 167.5186
Jan 2020       174.6337 174.6337 174.6337
Feb 2020       155.7930 155.7930 155.7930
Mar 2020       171.7743 171.7743 171.7743
Apr 2020       173.6060 173.6060 173.6060
May 2020       172.2580 172.2580 172.2580
Jun 2020       162.2269 162.2269 162.2269
Jul 2020       186.3473 186.3473 186.3473
Aug 2020       190.8609 190.8609 190.8609
Sep 2020       170.8172 170.8172 170.8172
```

Model Accuracy for the model:

| RMSE | | MAPE | |
|---|---|---|---|
| **Validation Partition** | **Entire Dataset** | **Validation Partition** | **Entire Dataset** |
| 17.62 | 14.487 | 8.25 | 6.571 |

Plot for historical data, predictions for historical data and seasonal ARIMA forecast for 12 future periods is presented in figure 5 in the appendix.

Model 4: Auto ARIMA Model:

The Autoregressive Integrated Moving Average (ARIMA) model is a versatile tool suitable for forecasting data that exhibits level, trend, and seasonal patterns. Given that our dataset includes these three components, the ARIMA model is well-suited for our analysis. We developed an optimal ARIMA model by automatically determining the best (p, d, q)(P, D, Q) parameters through the auto.arima() function.

Below is the output for the auto ARIMA Model developed on the validation partition:

```
> summary(train.auto.arima)
Series: train.ts
ARIMA(0,1,1)(0,0,1)[12]

Coefficients:
          ma1    sma1
      -0.5528  0.3775
s.e.   0.1248  0.1548

sigma^2 = 274.6:  log likelihood = -245.23
AIC=496.46   AICc=496.9   BIC=502.64

Training set error measures:
                   ME     RMSE     MAE        MPE     MAPE     MASE       ACF1
Training set 0.990594 16.14432 13.0013 -0.1843154 8.679864 0.524516 0.07682238
```

As observed in the above summary, the model consists of a moving average component lagged 1 period, and an order 1 seasonal autoregressive component:

| Parameters | Values from summary | Explaination |
|---|---|---|
| p | 0 | No autoregressive componenet |
| d | 1 | Order 1 differencing to remove linear trend |
| q | 1 | Order 1 moving average MA(1) model for error lags |
| P | 1 | Order 1 autoregressive model AR(1) for seasonality |
| D | 0 | No differencing for the seasonal part |
| Q | 0 | No moving average MA(1) for the seasonal error lags |
| m | 12 | Monthly seasonality |

The model's equation will be:

$$yt - yt\text{-}1 = - 0.5528et\text{-}1 + 0.3775yt\text{-}1$$

Below is the point forecast for monthly sales in the validation partition:

```
> train.auto.arima.pred
         Point Forecast    Lo 0     Hi 0
Dec 2018       162.2318 162.2318 162.2318
Jan 2019       155.0735 155.0735 155.0735
Feb 2019       155.5328 155.5328 155.5328
Mar 2019       157.3497 157.3497 157.3497
Apr 2019       167.8252 167.8252 167.8252
May 2019       165.6151 165.6151 165.6151
Jun 2019       161.7931 161.7931 161.7931
Jul 2019       161.2988 161.2988 161.2988
Aug 2019       156.3784 156.3784 156.3784
Sep 2019       163.8798 163.8798 163.8798
```

Below is the output for the auto ARIMA Model developed on the **entire data set.**

```
> summary(auto.arima)
Series: sales.ts
ARIMA(0,1,1)(1,0,0)[12]

Coefficients:
          ma1    sar1
      -0.5950  0.2841
s.e.   0.1013  0.1251

sigma^2 = 285.8:  log likelihood = -288.47
AIC=582.94   AICc=583.32   BIC=589.6

Training set error measures:
                   ME     RMSE      MAE        MPE     MAPE     MASE       ACF1
Training set 1.194367 16.53278 13.55543 -0.1193916 8.955166 0.574468 0.02304395
```

As observed in the above summary, the model consists of a moving average component lagged 1 period, and an order 1 seasonal autoregressive component.

| Parameters | Values from summary | Explaination |
|---|---|---|
| p | 0 | No autoregressive componenet |
| d | 1 | Order 1 differencing to remove linear trend |
| q | 1 | Order 1 moving average MA(1) model for error lags |
| P | 1 | Order 1 autoregressive model AR(1) for seasonality |
| D | 0 | No differencing for the seasonal part |
| Q | 0 | No moving average MA(1) for the seasonal error lags |
| m | 12 | Monthly seasonality |

The model's equation will be:

$$yt - yt\text{-}1 = -\ 0.5950et\text{-}1 + 0.2841yt\text{-}1$$

Below is the point forecast for monthly sales in the entire dataset set:

```
> auto.arima.pred
         Point Forecast     Lo 0     Hi 0
Oct 2019       166.4468 166.4468 166.4468
Nov 2019       174.4028 174.4028 174.4028
Dec 2019       167.2992 167.2992 167.2992
Jan 2020       176.3918 176.3918 176.3918
Feb 2020       163.3212 163.3212 163.3212
Mar 2020       169.2882 169.2882 169.2882
Apr 2020       171.2772 171.2772 171.2772
May 2020       173.2662 173.2662 173.2662
Jun 2020       168.4358 168.4358 168.4358
Jul 2020       176.9600 176.9600 176.9600
Aug 2020       176.9600 176.9600 176.9600
Sep 2020       171.2772 171.2772 171.2772
```

Accuracy measures for the model:

| RMSE | | MAPE | |
|---|---|---|---|
| **Validation Partition** | **Entire Dataset** | **Validation Partition** | **Entire Dataset** |
| 15.795 | 16.533 | 8.193 | 8.955 |

Plot for historical data, predictions for historical data and seasonal ARIMA forecast for 12 future periods is presented in figure 5 in the appendix.

**Step 8: Implement Forecast**

| Methodology | RMSE | MAPE |
|---|---|---|
| Linear regression with trend and seasonality + trailing MA for residuals | 9.803 | 5.403 |
| Linear regression with trend and seasonality + AR(1) model for residuals | 14.582 | 8.149 |
| Seasonal ARIMA (1,1,1)(1,1,1) Model | 14.487 | 6.571 |
| Auto ARIMA Model | 16.533 | 8.955 |
| Seasonal Naïve | 28.689 | 14.842 |

The table above evaluates the performance of various forecasting methodologies by comparing their Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The results clearly indicate that the "Linear regression with trend and seasonality + trailing MA for residuals" outperforms the other models, with the lowest RMSE of 9.803 and MAPE of 5.403. This model effectively incorporates linear trends and seasonal variations, enhanced by a trailing moving average for residuals, making it the most accurate for forecasting M01AB

sales. Consequently, this approach is recommended as the optimal model for predicting future sales trends in the pharmaceutical drug category - M01AB.

## Conclusion

This case study has thoroughly examined the task of forecasting monthly pharmaceutical sales for the M01AB category—Anti-inflammatory and antirheumatic products, non-steroids, using advanced statistical models. Our analysis began with a detailed exploration of the dataset sourced from a pharmacy's Point-of-Sale system, which recorded transactions from 2014 to 2019. After rigorous preprocessing to clean and prepare the data, including handling missing values and focusing exclusively on the M01AB category, we moved forward with partitioning the data into training and validation sets. This allowed for a robust assessment of the models' predictive performances.

We implemented several forecasting models, including various configurations of linear regression and ARIMA models. The models were evaluated based on their RMSE and MAPE values, essential metrics for gauging forecast accuracy. Our findings reveal that the "Linear regression with trend and seasonality + trailing MA for residuals" model provided the most accurate predictions, achieving the lowest RMSE and MAPE scores among the tested methodologies.

The success of this model can be attributed to its ability to effectively capture and account for the underlying patterns in the data—namely the trends and seasonal fluctuations that characterize the sales of M01AB pharmaceuticals. The trailing moving average component of the model further refined the predictions by smoothing out residual errors, enhancing the model's overall forecast precision.

Based on our comprehensive analysis, it is recommended that the pharmacy utilize this model for ongoing forecasting efforts. This will enable more effective inventory management, ensuring that supply aligns with anticipated demand, thereby reducing both overstock and stockout situations. Moreover, the insights gained from this forecasting process can aid in strategic decision-making, helping the pharmacy to adapt to trends and potentially increase profitability.

In conclusion, this case study underscores the importance of tailored, data-driven forecasting models in the pharmaceutical retail industry. By continuing to refine these models and adapt them to emerging trends, the pharmacy can maintain an optimal balance of supply and demand, crucial for operational efficiency and customer satisfaction.

## Bibliography

Milan Zdravkovic. "Pharma Sales Data." Kaggle, 2024, www.kaggle.com/datasets/milanzdravkovic/pharma-sales-data

## Appendix

**Figure 1:** Training Partition - Jan 2014 to Nov 2018

```
> train.ts
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2014 127 133 137 113 101 112 117 134 108 154 138 131
2015 135 115 156 154 160 151 175 181 166 195 152 159
2016 171 173 156 166 167 169 203 211 172 186 175 169
2017 149 139 162 155 143 122 159 170 138 137 163 160
2018 132 128 146 170 160 141 150 140 153 144 172
```

**Figure 2:** Validation Partition - Dec 2018 to Sep 2019

```
> valid.ts
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2018                                             147
2019 179 133 154 161 168 151 181 181 161
```

**Figure 3:** Plot for original pharma M01AB sales time series data, regression model and regression forecast for future 12 periods.

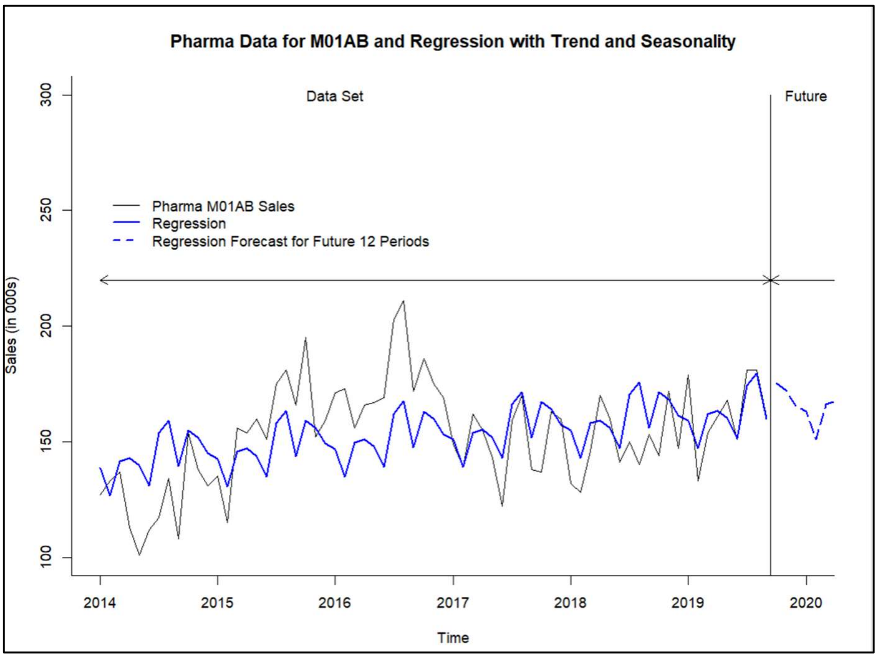



**Figure 4:** Plot historical data, predictions for historical data and forecast for 12 future periods.

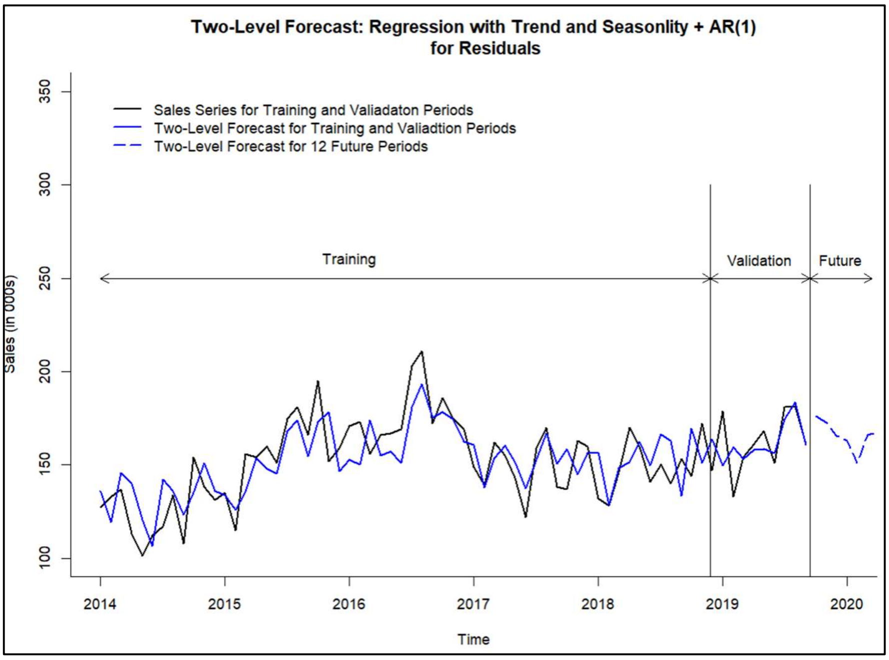



**Figure 5:** Plot historical data, predictions for historical data and seasonal ARIMA forecast for 12 future periods.

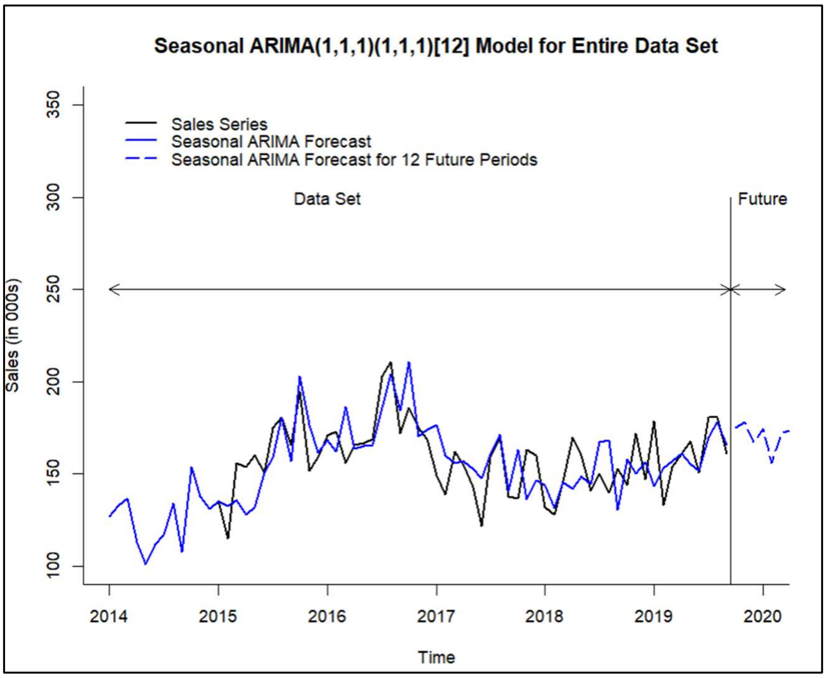



**Figure 6:** Plot historical data, predictions for historical data and Auto ARIMA forecast for 12 future periods.