Time Series Analysis Project Report

Post Graduate Programme in Artificial Intelligence and data Science

Group Member	
Girish Khule	23PGAI0057
Raj Rajeshwari Bajre	23PGAI0093
Darshana Nitin Mandhana	23PGAI0089
Anurag Pandey	23PGAI0059

Introduction

The importance of sales forecasting is well known. It helps a company to plan according to the predicted future ensuring maximum profitability and also helps in making informed decisions. It is crucial for budgeting and finance. A sound sales forecast assists the production planning team to ensure enough products are manufactured so that the demand is met, ensuring maximum utilisation of resources to maximise profit. It also opens the opportunity to think about expansion/shift in production.

Specifically in pharmaceutical industry, accurate sales forecasting is especially important because of the high costs and long lead times associated with drug development and commercialization. It can take many years and billions of dollars to bring a new drug to market, so it is crucial for pharmaceutical companies to have a good understanding of future demand for their products.

In addition, because of high cost of development and production, prices of pharmaceutical products are often very high. With accurate sales forecasting a company can accurate determine how much to charge for a product and still be profitable.
In short, accurate sales forecasting is essential for the long-term viability and success of a pharmaceutical company. Hence, this particular problem was chosen.
Also, the sales intuitively should have a trend as well as seasonality. Which will make it interesting to apply all the methods studied in the course.

About the dataset

The dataset is built from the initial dataset consisted of 600000 transactional data collected in 6 years (period 2014-2019), indicating date and time of sale, pharmaceutical drug brand name and sold quantity, exported from Point-of-Sale system in the individual pharmacy. Selected group of drugs from the dataset (57 drugs) is classified to the following Anatomical

Therapeutic Chemical (ATC) Classification System categories:

- 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 drugs, Anxiolytic drugs
- N05C - Psycholeptics drugs, Hypnotics and sedatives drugs
- R03 - Drugs for obstructive airway diseases
- R06 - Antihistamines for systemic use

Sales data are resampled to the hourly, daily, weekly and monthly periods. Data is already pre-processed, where processing included outlier detection and treatment and missing data imputation. Forecasting is being done for all 8 categories.

Methodology followed

The data was first split into training and testing sets where the data for the last year (out of 6 years) was kept in for testing.
The analysis was done in two phases – 1. Weekly and daily data analysis with the objective to make potentially useful conclusions and propositions for improving sales and marketing strategies. 2. Stationarity, autocorrelation and predictability analysis of the time series in individual groups to infer the initial set of parameters for implementing the forecasting methods.

The analysis was carried out in the following steps:

- Heatmap was plotted to check for correlation between the categories. No correlation was found.
- Boxplots were employed to check for the seasonality. Weekly as well as daily data was used.
- Augmented Dickery Fuller (ADF) test was carried on for each of the category for both daily and weekly data. This ruled out stationarity in the data.

 Auto-Correlation analysis was done to plot the ACF and PACF graphs to determine the value of p,d and q for ARIMA.

 Three forecasting methods were used:

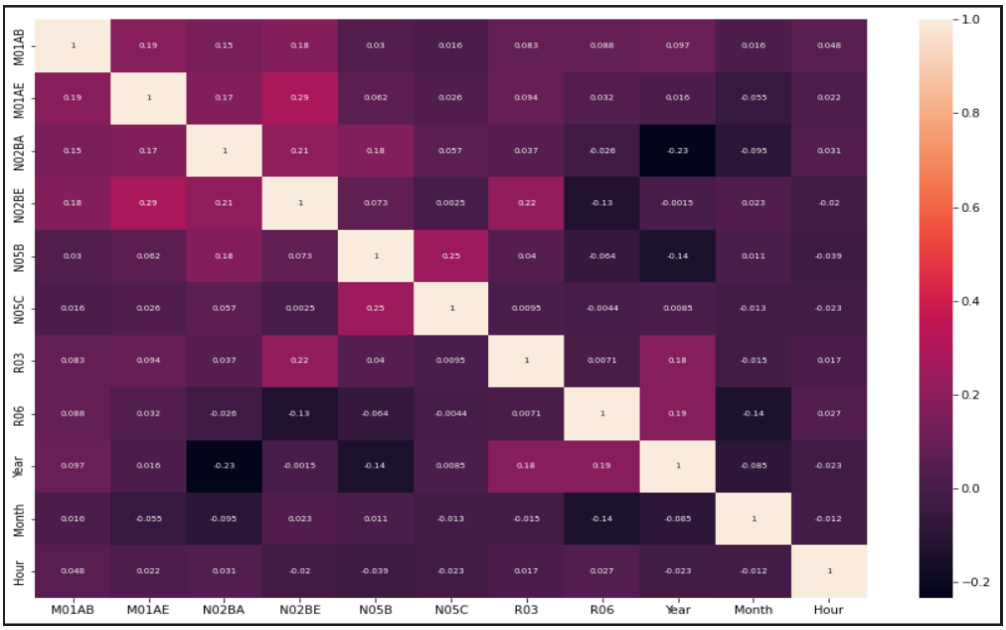
1. Naïve
2. Seasonal Naïve
3. ARIMA with various pdq values

Key performance indicator for forecasting accuracies in both approaches was Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) was also provided as illustration as data on different groups of pharmaceutical products were on significantly diverse scales.

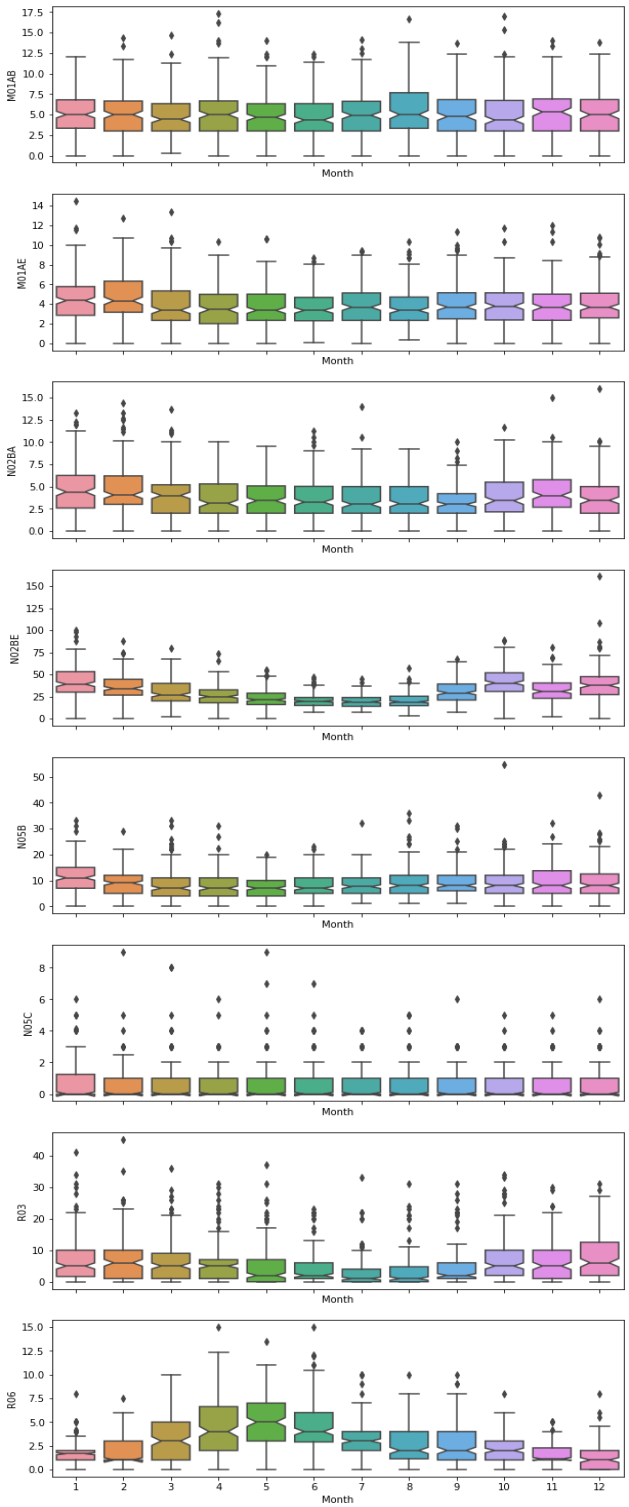
Time series analysis

Exploring seasonality using boxplots.
Seasonality is clearly confirmed for the categories of
- R03
- R06
- N02BE

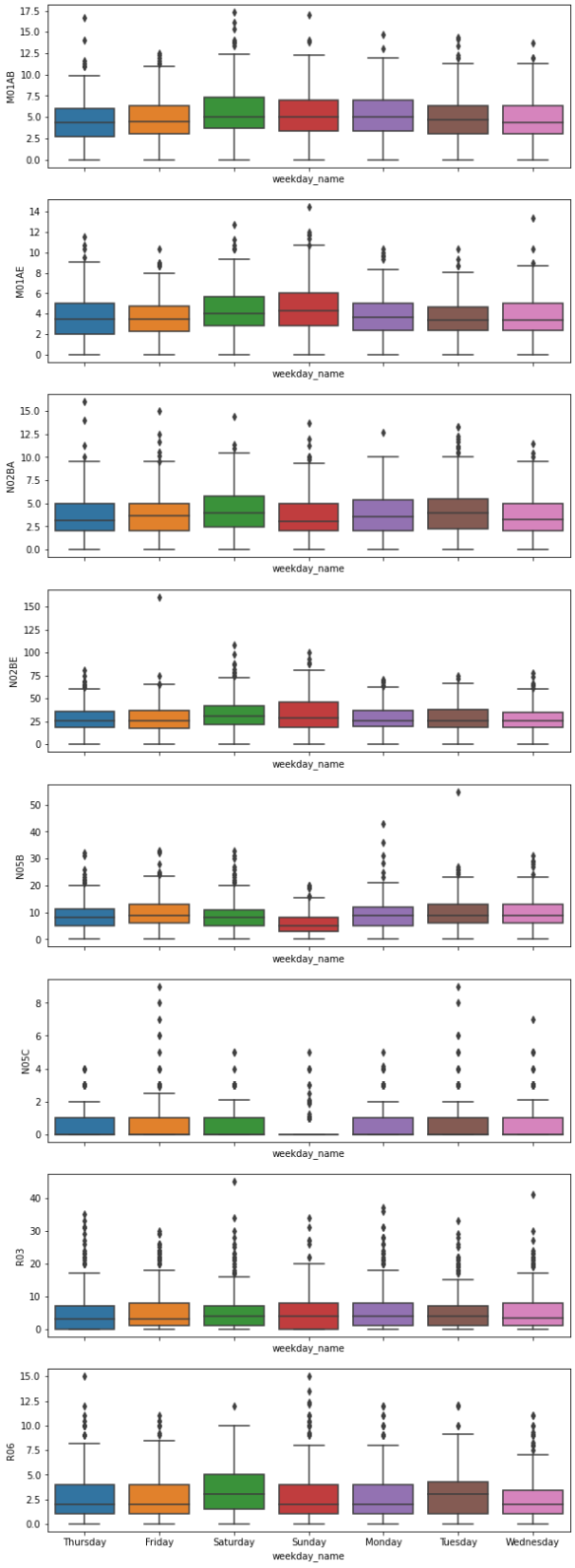
Sales would be difficult to predict as there are more outliers for the categories of
- R03
- N05C



Exploring monthly seasonality for the categories.



Now exploring weekly seasonality for the categories



Rolling window means to identify seasonality patterns

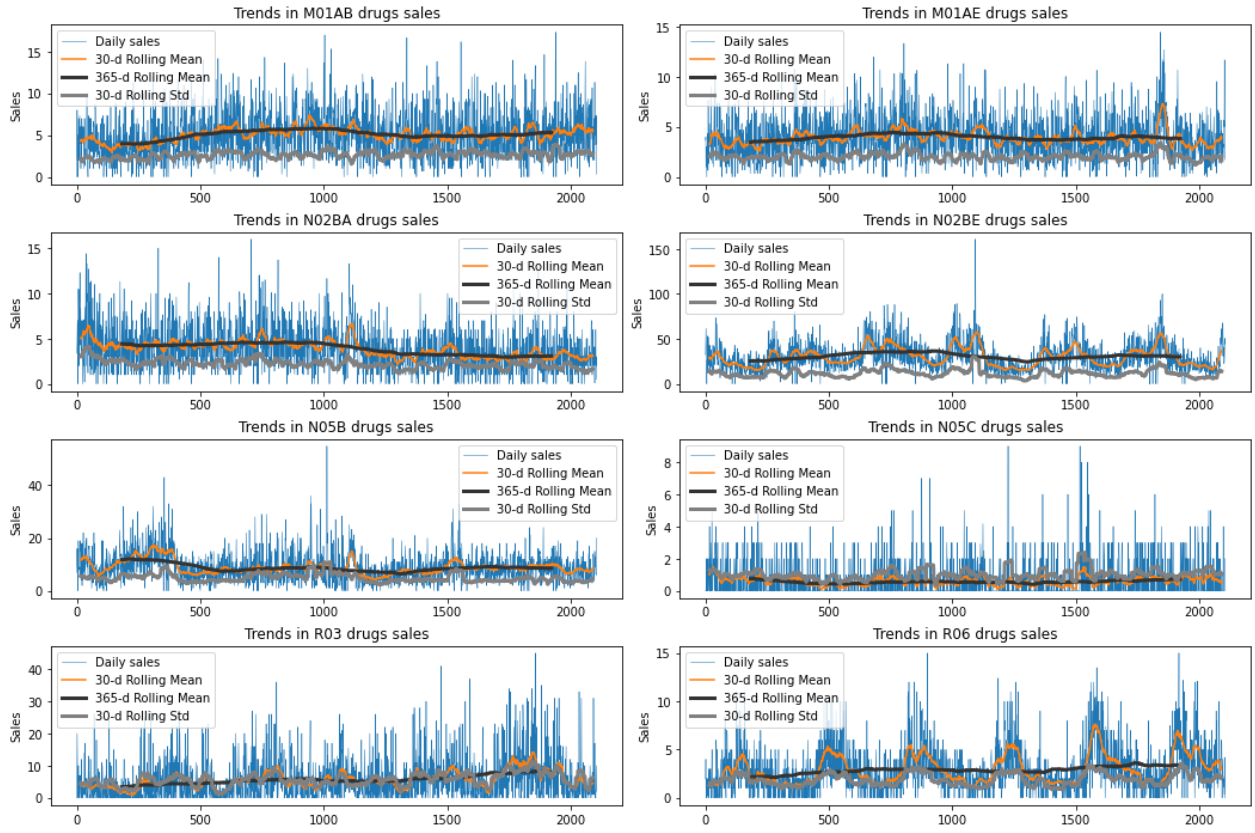
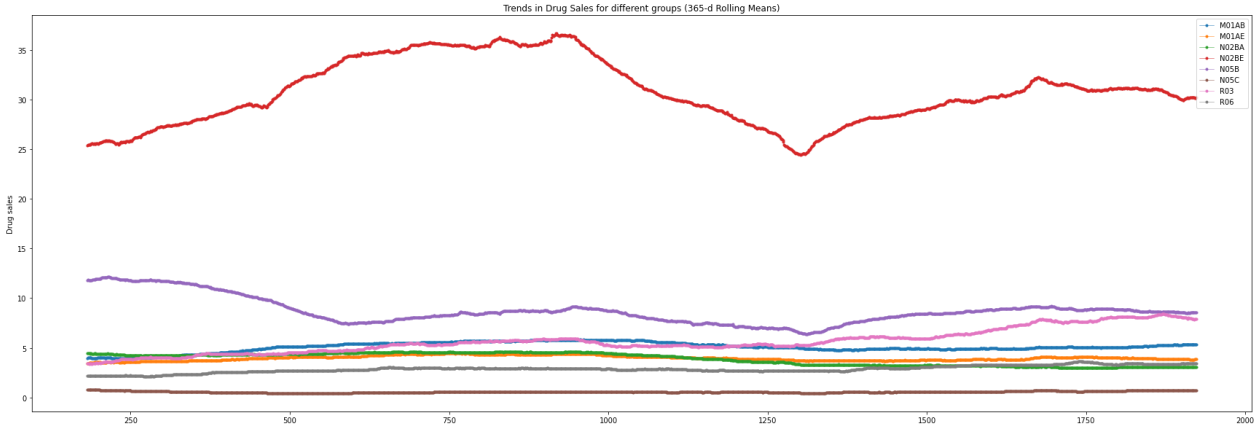
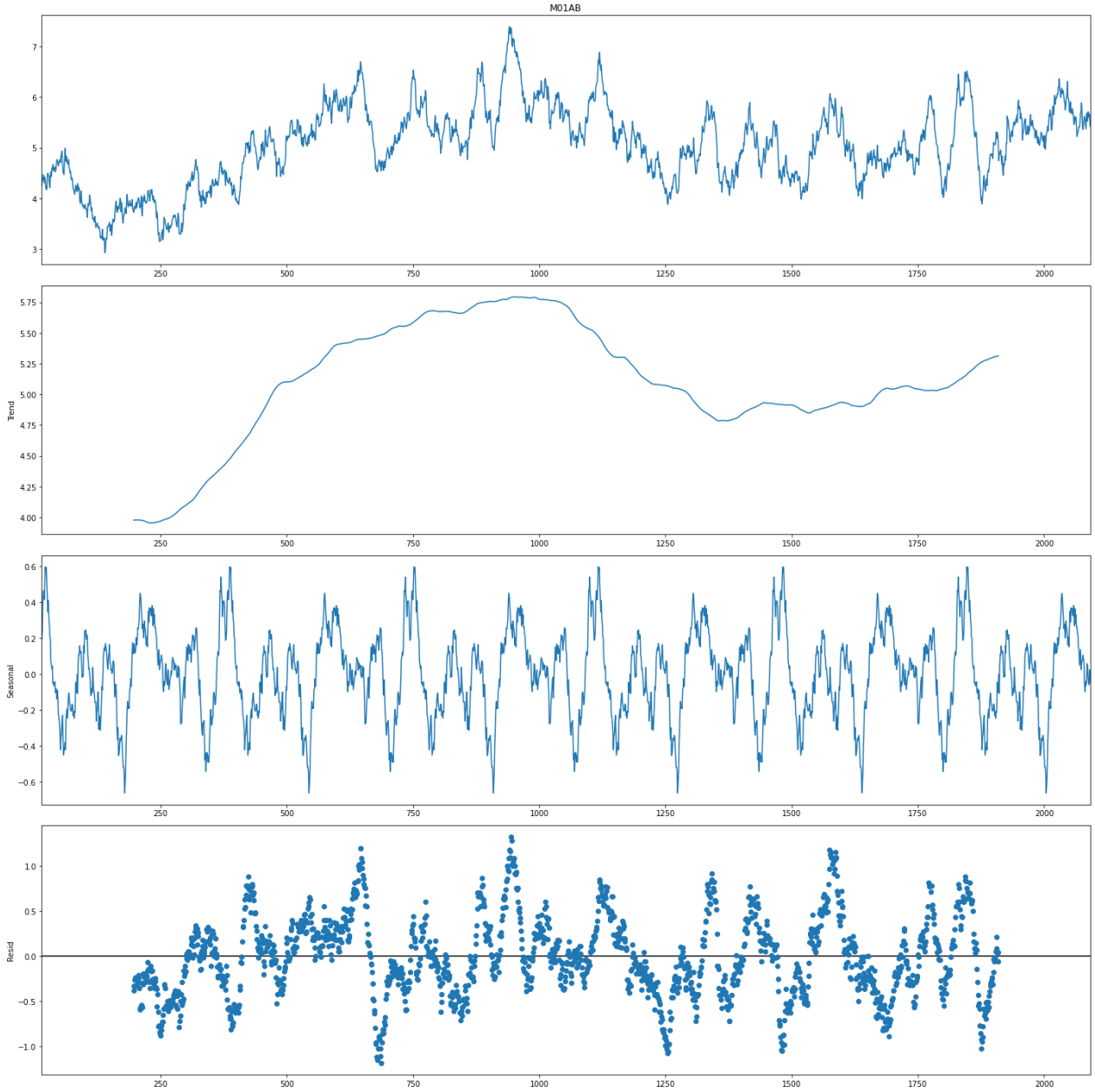


Image below shows trends for each of the drug categories, represented by the 365-d rolling means for each of those categories.



Trends and seasonality can be explored in time series decomposition view, based on 30d rolling means.



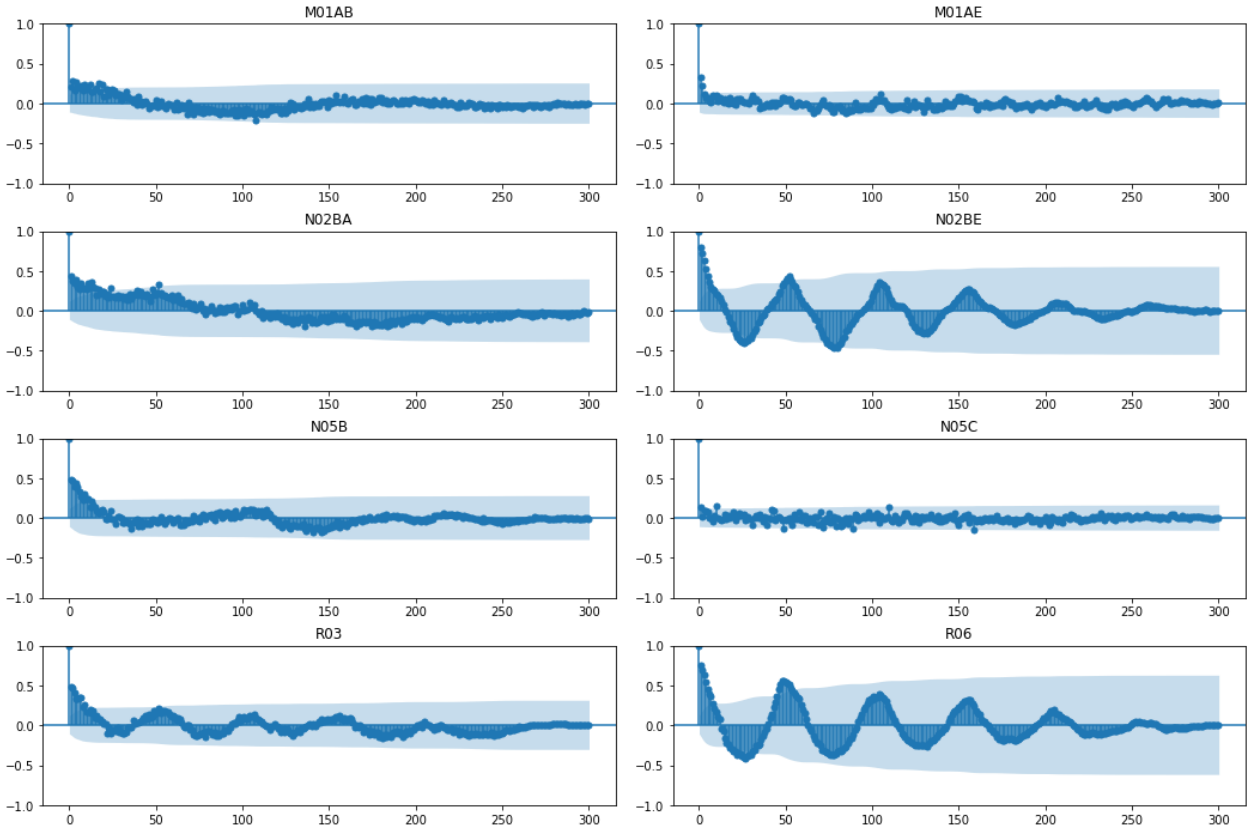
Stationarity analysis

We can use Augmented Dickey-Fuller (ADF) test to check stationarity of the data.

Augmented Dickey-Fuller (ADF) test have shown that all data, but N02BA (P-value=0.249) in the series were stationary, with maximum confidence.

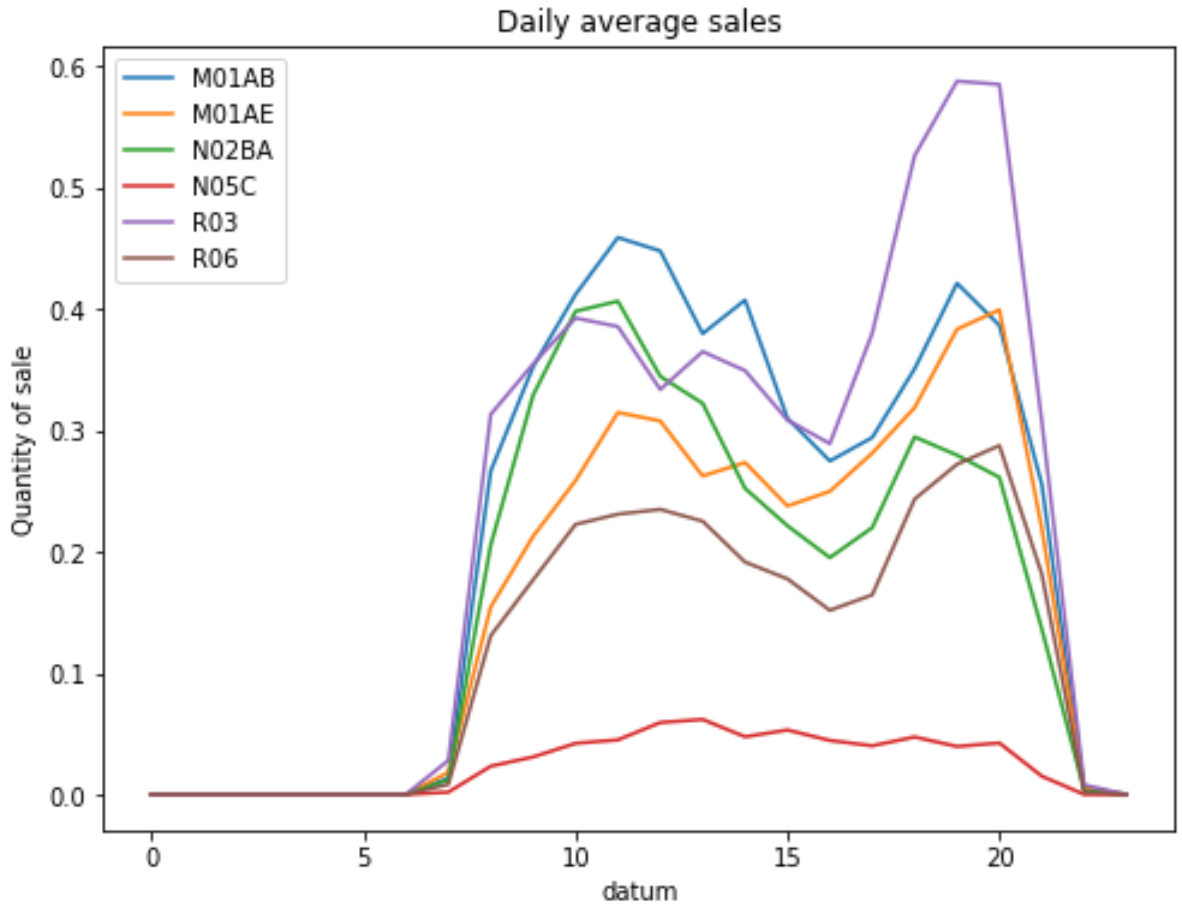
Autocorrelation analysis

Minor autocorrelation is observed at ACF (Auto-Correlation Function) and PACF (Partial Auto-Correlation Function) plots for all series, with exception of N05C sales. N02BE, R03 and R06 series were found to exhibit annual seasonality.



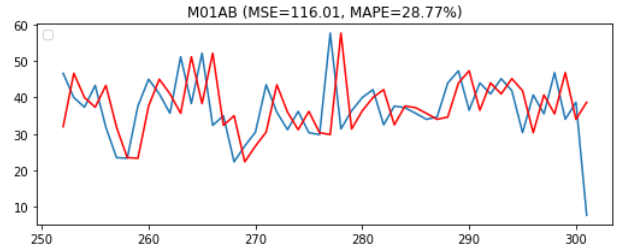
Data distribution analysis

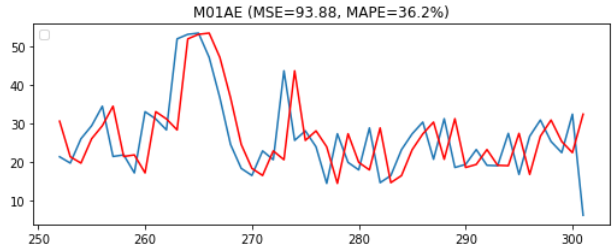
Chart with daily sales for different categories of interest is shown below. N02BE and N05B charts, though showing the similar trends, are suppressed because of the larger scale which makes the other illustrations less readable.

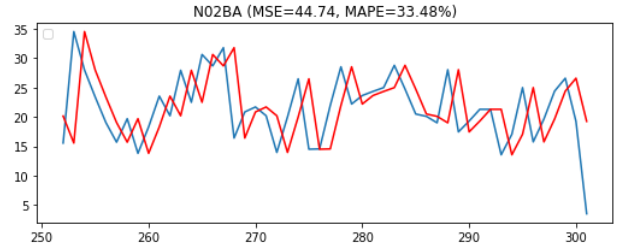


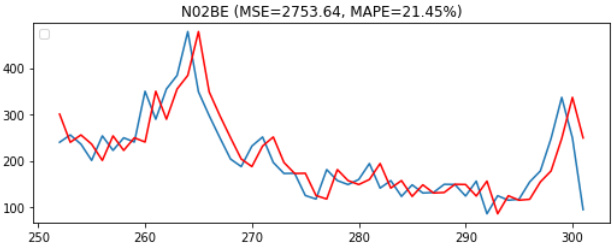
Time series forecasting

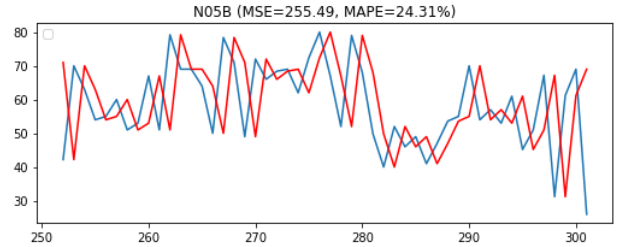
Naïve forecasting

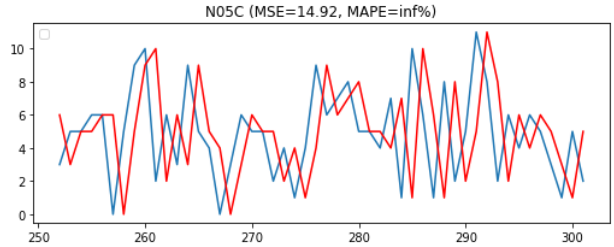


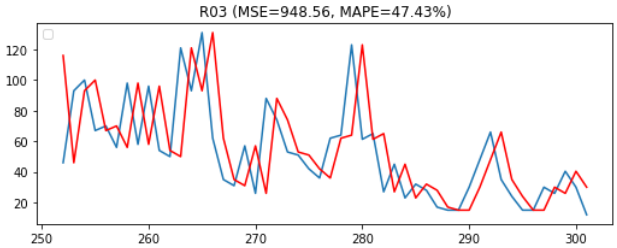


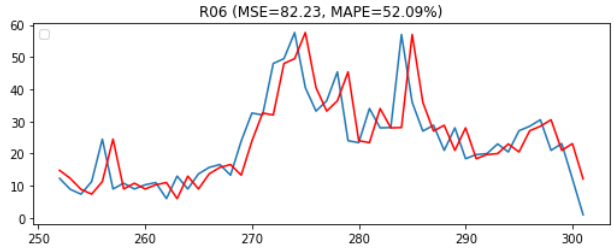




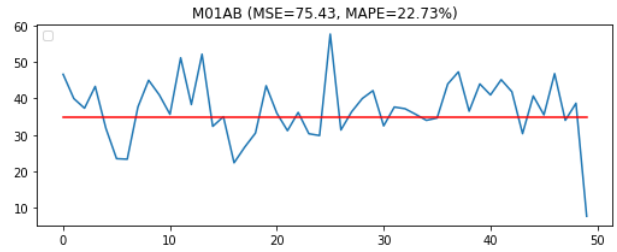


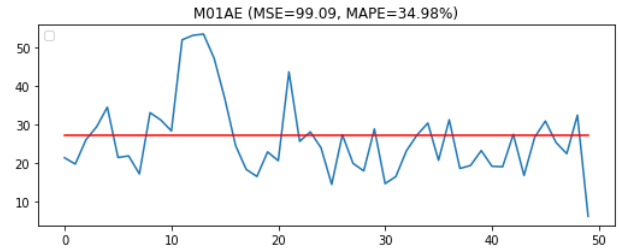


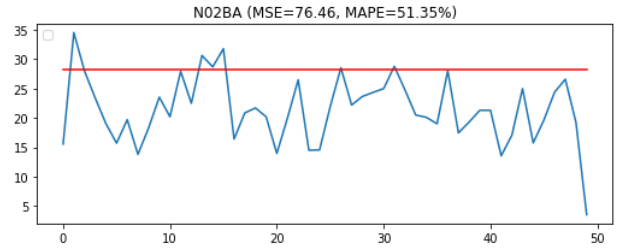


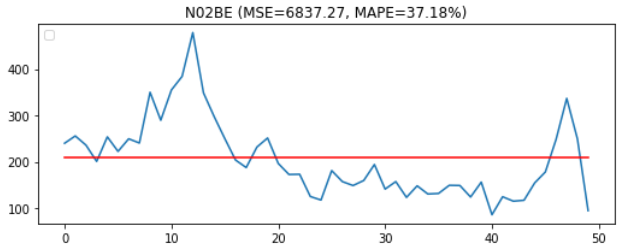


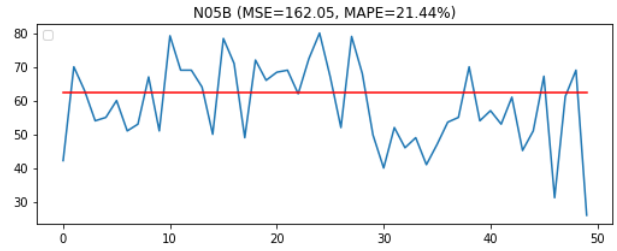
Average method forecasting

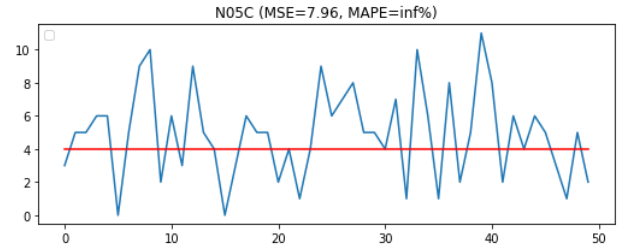


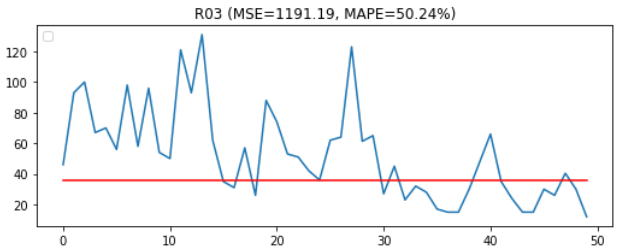


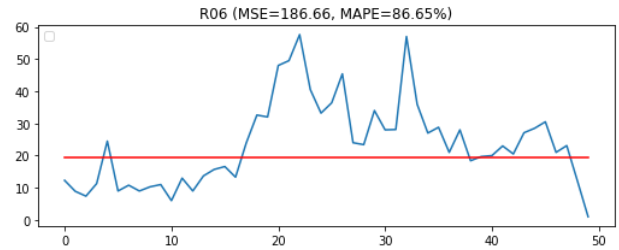






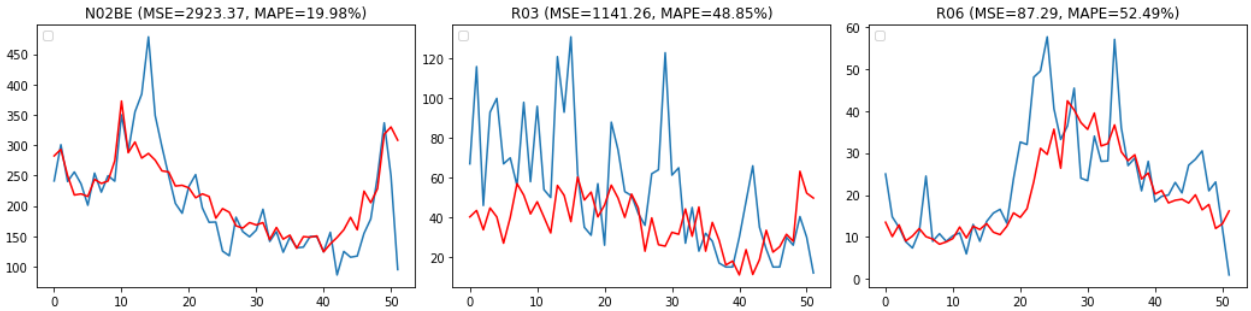






Seasonal Naïve forecasting

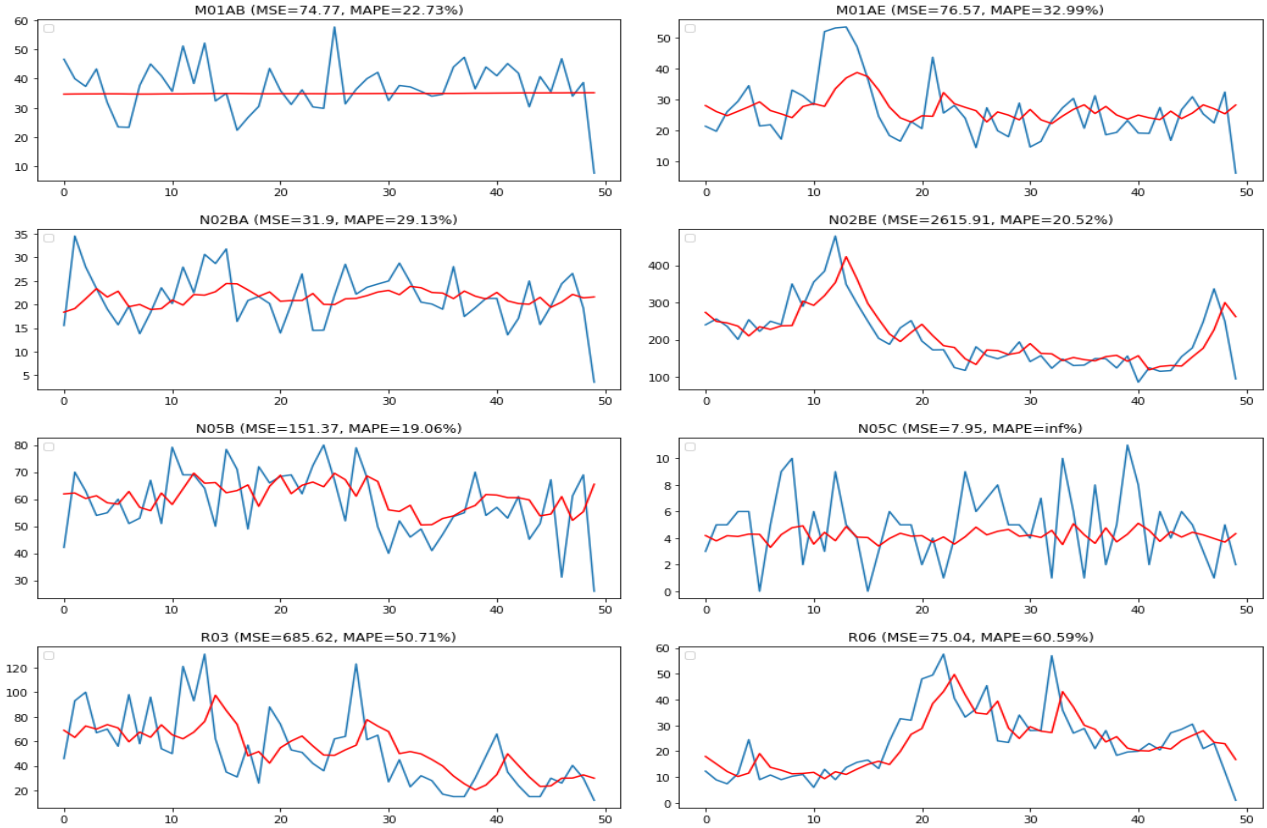
Seasonal Naïve forecast was carried out for the series that has been found as seasonal: N02BE, R03 and R06.



ARIMA Forecasting

Choosing parameters for ARIMA model

First, method arma_order_select_ic was used to determine initial p and q parameters. The method computes Akaike's Information Criterion (AIC) for many ARIMA models and chooses the best configuration.



Conclusion

ARIMA outperform reference benchmarks - Naive and Seasonal Naive forecasts.

- Mostly, naïve method is employed for pharma sales forecasting. But, with this analysis, it can be concluded for all categories, that ARIMA performed better than seasonal and Naïve method. This shows that seasonality analysis is useful for identifying the time of low/high.
- Sales to accordingly implement marketing campaigns to conclude, time-series analyses and forecasts have guided potentially useful conclusions and recommendations to the pharmacy. Daily, weekly and annual seasonality analysis were proven useful for identifying the periods in which special sales and marketing campaigns could be implemented, except for N05B and N05C categories of drugs which did not exhibit significant regularities. Forecasts have proven better than Naïve methods and in acceptable intervals for long-term planning. It is highly likely that the forecasts could be significantly improved by expanding the problem scope to multivariate time series forecasting and by including explanatory variables.
- Weather data. Sales of antirheumatic drugs in M01AB and M01AE categories could be affected by the changes of atmospheric pressure. Sudden declines in all categories could be explained by extreme weather conditions, such as heavy rain, thunderstorms and blizzards.
- Price of the drugs. Sales spikes may be explained by the discounts, applied in a short term. Introducing this feature may facilitate what-if forecasting analysis of sales performance during marketing campaigns involving price reductions.
- Dates of the pension payoff. Sales spikes are visible at the dates of state pensions payoff.
- National holidays, as non-working days with seasonal patterns similar to Sundays are expected to disrupt daily sales.
- Future work on univariate time series forecasting includes increasing the number of data, exploring different other accuracy metrics, optimization of hyper-parameters for LSTM models and testing other architectures, such as CNN LSTM and ConvLSTM. However, key improvements in sales forecasting are expected from reducing the uncertainty of the models by expanding to multivariate time series forecasting problem, as explained above.