PHARMACEUTICAL DRUG SALES VOLUMES
TIME SERIES ANALYSIS AND FORECASTING

Project Report by:
Joseph Gadbois

Analysis and Forecasting by:
Adam Castillo
Joseph Gadbois
Vichetra Lim
Hamzah Sami
Christian Ramirez

STAT 580 Time Series Analysis

Group Project

December 8, 2020

Introduction

The prominence of pharmaceutical drugs stems from the various ailments incurred by the countless number of people who use them. As time progresses, pharmacies and related companies may be interested in questions regarding the sales volume of these drugs, and which classes of drugs follow consistent sales volume patterns. The purpose of this study is to analyze the sales volume of pharmaceutical drugs and produce reliable forecasts allowing for optimal preparation of future weeks and months.

Data Description

The data used in this study come from Kaggle and consists of transactional data over a six-year period from 2014 to 2019. The raw data is weekly sales volume where the variables include the date, and pharmaceutical drug classes including M01AB, M01AE, N02BA, N02BE/B, N05B, N05C, R03, and R06 where each drug class is a univariate time series.

Due to similarities of drug classes, the data were further transformed such that there is a univariate time series for M01, N02, N05, and R0. The descriptions of the transformed data can be found in the table below.

Variable	Description
Date	*Date of weekly sales volume*
M01	*Anti-inflammatory and anti-rheumatic products*
N02	*Other analgesics and antipyretics*
N05	*Psycholeptic drugs*
R0	*Antihistamines and drugs for obstructive airway diseases*

Note that the initial drug classifications follow the Anatomical Therapeutic Chemical (ATC) Classification System.

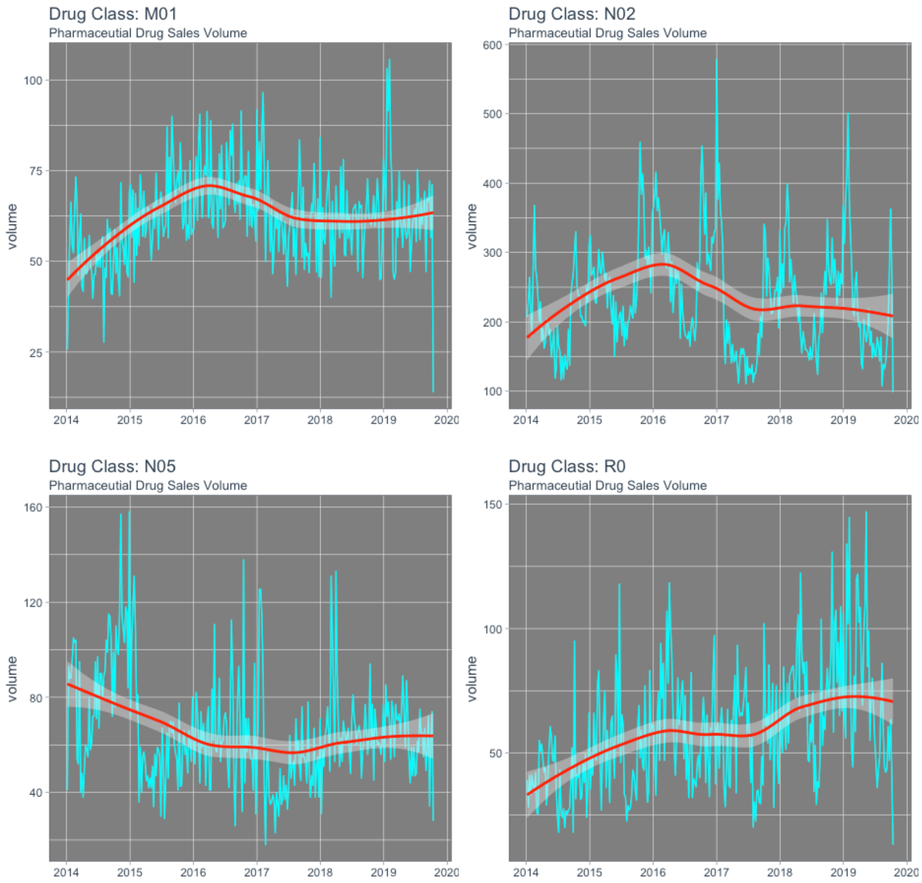
Research Questions

The following three research questions were used to guide this study.

1. How can we deal with outliers in a time series?
2. For each drug class, the last 11 weeks of 2019 will be forecasted. Are the forecasts reasonable?
3. For each drug class, are there any life events that could affect the volume of pharmaceutical drug sales?

Exploratory Data Analysis

The initial steps to any analysis are exploratory to obtain basic knowledge of the data. The time series plots for all four variables and a Loess regression curve are displayed in the image below.



The Loess regression curves were plotted with a decent sized smoothing parameter to give an idea of any potential trends present in the data. Based on the time series plots, the only variable that displays a stable trend is R0 with the exception of the year 2016. Regarding trends, the other drug classes display trends in a smaller partition of the series, but unstable across the date range of available observations. Note that Loess regression can be fit using a smaller smoothing parameter to model periodicity if needed.

Data Preparation

The first week present in the data is January 1st, 2014 and the last week is October 13th, 2019. The data were split such that the train data consists of observations up to December 30th, 2018 which is the last week in 2018. Observations starting from the week of January 6th, 2019 are then set aside for test data. Since the last week of sales is in October of 2019, the forecast horizon of interest consists of the last 11 weeks of the year.

Addressing outliers, STL decomposition, or Seasonal Trend decomposition using Loess, was used to identify and smooth the series for each variable. The use of Loess regression makes STL a robust method due to the locally estimated regressions, allowing it to handle any type of seasonality including a seasonal component that changes over time. STL is robust to outliers because of local estimation, and thus replaces them with the estimated value from Loess at a given time step.

Testing for Stationarity

Time series analysis and forecasting with ARIMA models generally requires the series to be stationary, meaning that the mean, variance, and autocovariance functions must be independent of the time index. To determine if the M01 series is stationary, the augmented Dickey-Fuller (ADF) test was performed testing the hypotheses

$$H_0: Y_t \sim random\ walk\ \ vs.\ H_1: Y_t \sim causal\ process.$$

The test provides a $p - val = 0.1724 > 0.05 = \alpha$ and thus, the series is not stationary. This was carried out again for the first-order difference of M01 to find that it is a causal process.

Testing is repeated for the N02, N05, and R0 drug classes. After conducting the ADF test for these variables, it was found that none of them were stationary at the five percent significance level. Since none of these series are stationary, the ADF test was performed on the first differences like that of the M01 series to find that N02 needs one differencing order for seasonal and non-seasonal components. N05 and R0 are stationary with the first difference of the series, but the seasonal difference is unneeded.

Model Selection

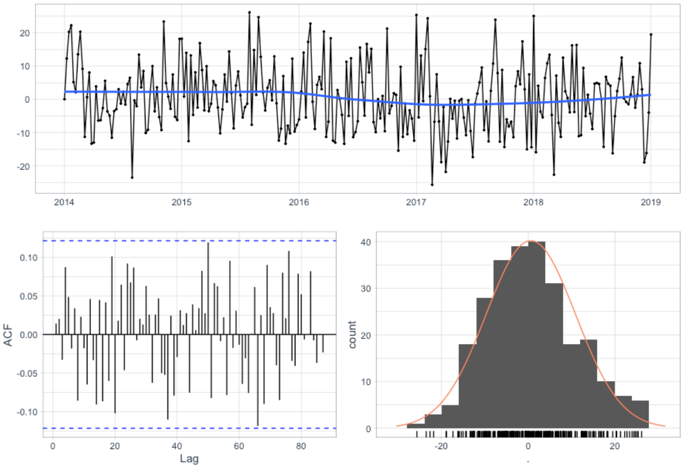
For each time series, three potential models are considered based on the ACF and PACF which can be found in the appendix. To determine the best model, the AIC, AICC, and BIC are used for comparison and the selected model is that with the minimum values for the information criteria. If no model has the minimum value for all three, then the selected model is that with two minimum criteria values.

M01 Model

The selected model for the M01 series is $ARIMA(2,1,1)$, given by

$$\phi_2(B)\nabla Y_t = \theta_1(B)w_t.$$

The model operators are estimated to be $\phi_2(B) = (1 - 0.018B - 0.040B^2)$ and $\theta_1(B) = (1 - 0.899B)$.



An *ARIMA* model assumes that the residuals are white noise and thus, model assumptions are assessed via plotting the series, ACF, and a histogram. Based on a visual assessment, the plots suggest that the residuals are white noise.

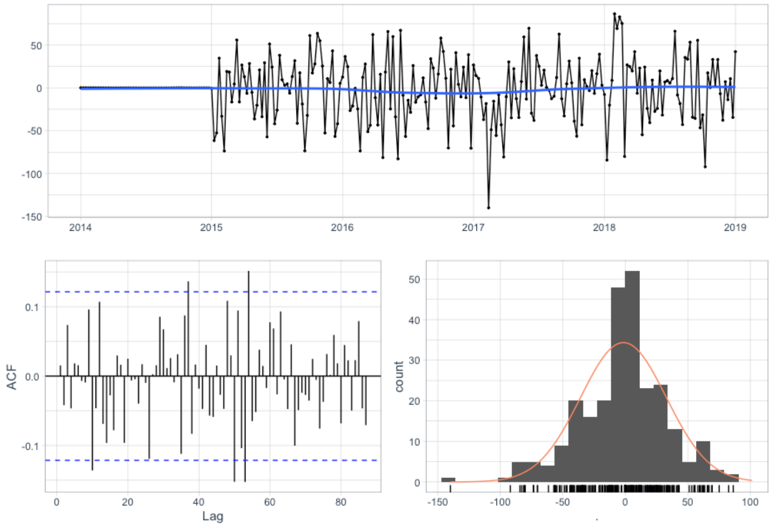
To further assess the model, the ADF test can be applied to residuals to determine whether they are stationary or not. After performing the ADF test, it can be concluded that the residuals are white noise, and the model can be deemed reliable to use for inference.

N02 Model

The best model for the N02 series is $ARIMA(2,1,3)(1,1,0)_{52}$ which is a seasonal model defined by

$$\Phi_1(B^{52})\phi_2(B)\nabla^{52}\nabla Y_t = \theta_3(B)w_t.$$

The non-seasonal operators are estimated as $\phi_2(B) = (1 - 1.306B + 0.672B^2)$ and $\theta_3(B) = (1 - 1.787B + 1.344B^2 - 0.440B^3)$. The seasonal autoregressive operator is estimated to be $\Phi_1(B^{52}) = (1 + 0.481B^{52})$. Plots to assess the model residuals are displayed in the image below.



The residual plot does not look exactly like white noise based on an initial visual assessment. Considering the ACF, there are a few significant correlations, but they are inconsistent and thus, cannot be implemented by adding more seasonal orders to the model. That said, the histogram shows that the residuals appear to have zero-mean which is required.

Visually assessing the residuals for this model does not allow for a confident conclusion that the model assumptions have been met. For a more confident assessment, the ADF test was performed to find that the residuals are stationary, and the model is okay to use.

N05 Model

The model selected for the N05 series is $ARIMA(3,1,2)$ and defined by

$$\phi_3(B)\nabla Y_t = \theta_2(B).$$

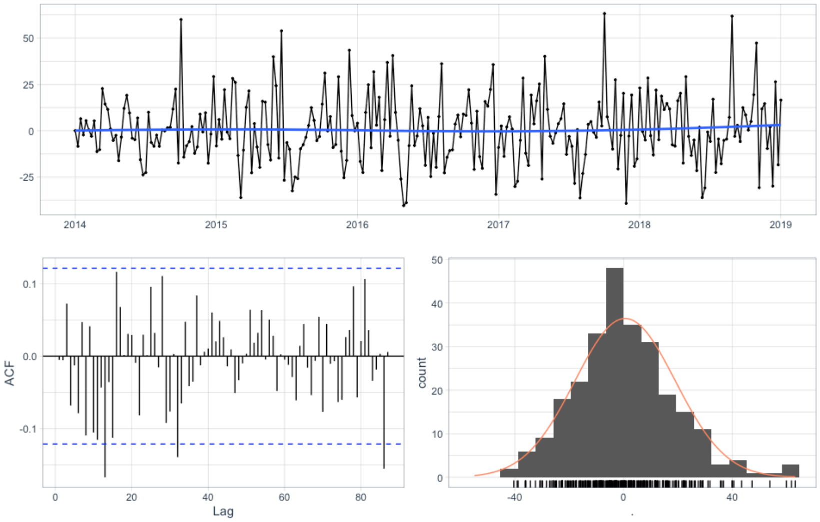
The autoregressive operator is estimated to be $\phi_3(B) = (1 - 0.069B + 0.017B^2 + 0.142B^3)$ and the moving average operator is $\theta_2(B) = (1 - 0.658B + 0.002B^2)$. A visual assessment of the residuals implies they are white noise, and the model is safe to use. These plots can be found in the appendix.

R0 Model

The best model for the R0 series is $ARIMA(0,1,2)(0,0,1)_{52}$ and defined by

$$\nabla Y_t = \theta_1(B)\Theta_1(B^{52})w_t.$$

The seasonal and non-seasonal moving average operators are estimated and given by $\theta_1(B) = (1 - 0.775B + 0.111B^2)$ and $\Theta_1(B^{52}) = (1 + 0.172B^{52})$. The model residuals are plotted below for residual analysis.



The ACF of the residuals has significant autocorrelations but at an inconsistent period and thus, additional seasonal orders will not benefit the model. The histogram also displays a bell-shaped curve implying the residuals are white noise.

The ADF test is performed and results in the conclusion that the residuals are stationary, and the model is reliable to use.

Forecasting

Now that a model has been selected for each time series and assumptions are satisfied, the next step is to predict forecasts for validation with test data. If the validation forecasts seem reasonable, the model can then be used to forecast future time steps.

Evaluation Metrics

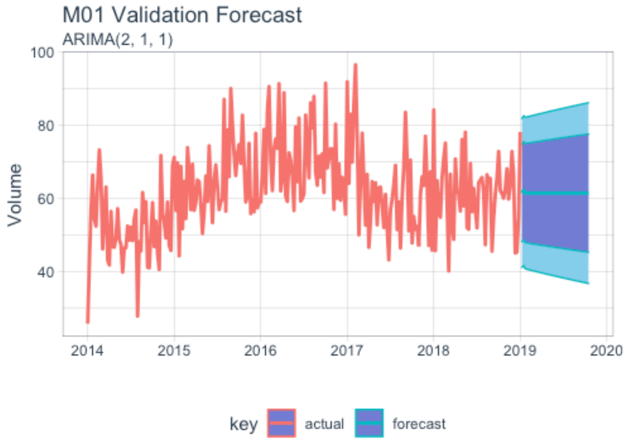
Part of model validation is ensuring that the model has not been overfit to the train data set. This is measured using RMSE and MAPE as evaluation metrics to gauge how the model performs predicting time steps that are known. If the metrics for the train data are not significantly better than for the test data, the model can be used to predict future time steps. The evaluation metrics for all models are summarized in the table below.

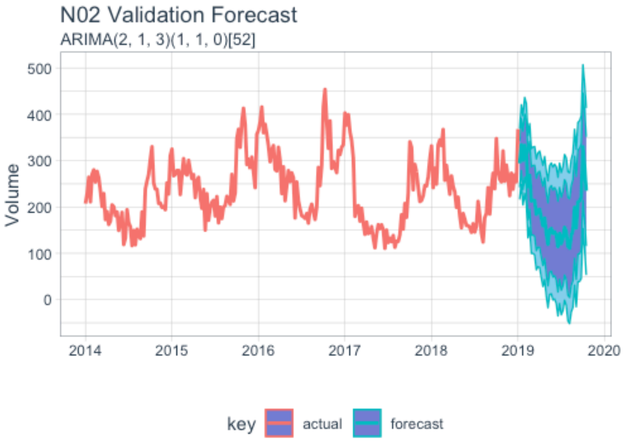
Metric	M01	N02	N05	R0
RMSE Train	10.323	34.292	18.376	18.450
RMSE Test	15.481	54.893	13.740	35.717
MAPE Train	13.584	10.742	23.902	28.255
MAPE Test	22.453	21.118	21.082	64.966

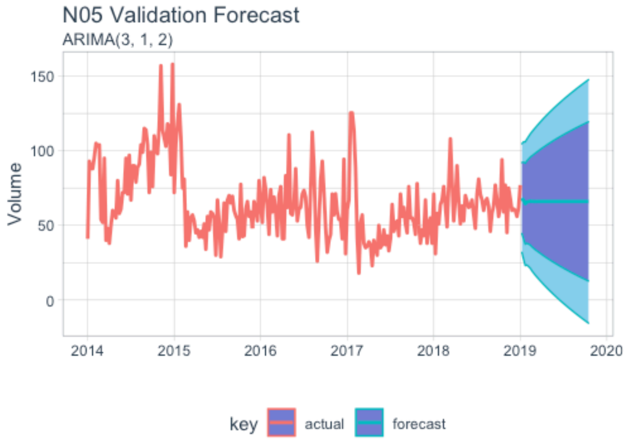
Considering the magnitude of the observations, the performance metrics for train and test data suggest that the models were not overfit to the train data.

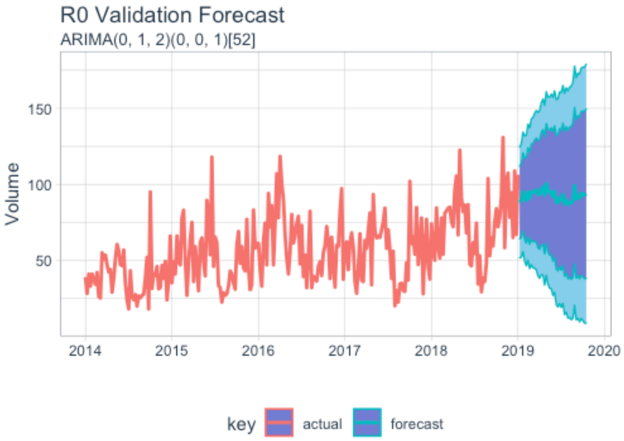
Validation Forecasts

The predicted forecasts for each time series are plotted below along with an 80 percent prediction interval in a darker shade of blue, and a 95 percent prediction interval in a lighter shade of blue. The validation forecasts can be found and observed in the image below.







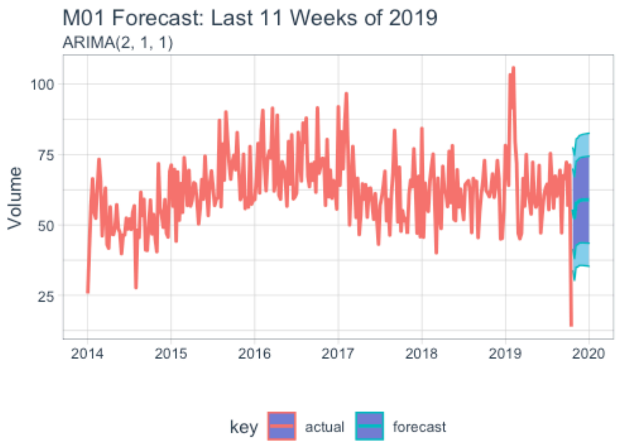


Observing the predicted forecasts, note that the non-seasonal model predictions are close to a straight line. This is because an optimal forecast minimizes the mean squared prediction error, but the prediction intervals differ from M01 to N05. The M01 series is very close to white noise when differenced and thus, the interval does not expand wider as time steps get larger. The prediction interval for N05 gets wider as time steps get larger because the series is far less consistent than the M01 series.

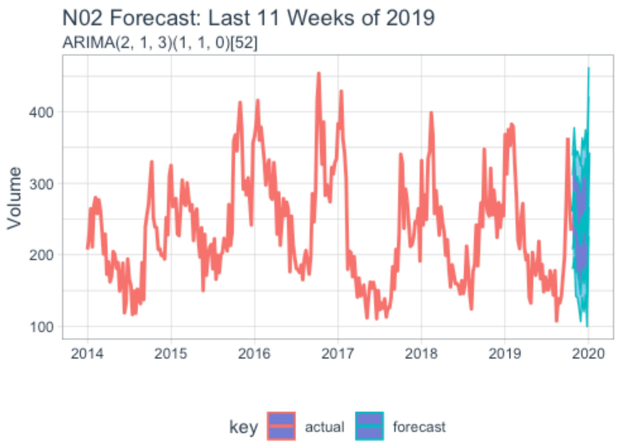
The seasonal forecasts take more shape to the actual predicted values. The forecast for N02 has a large swooping shape which is similar to the general shape of the entire series. For R0, the forecast does not display as excessive of a pattern, but this is also because the series does not evidently display seasonality. One thing to note is that the 95 percent prediction intervals for N02 and N05 include values less than zero and this is not possible. This is something that must be accounted for, but the majority of the intervals do not fall below zero.

Forecasting Future Values

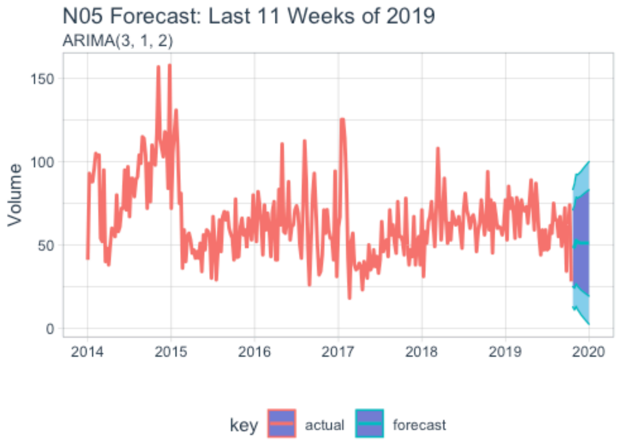
Since the evaluation metrics and validation forecasts are reasonable, the models can now be used to forecast future values of the time series. The forecasts for the last 11 weeks of 2019 have been plotted in the same format as above and displayed below.


M01 Forecast: Last 11 Weeks of 2019

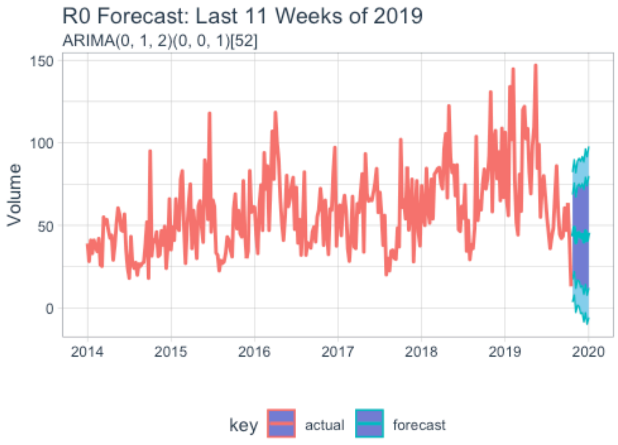
ARIMA(2, 1, 1)


N02 Forecast: Last 11 Weeks of 2019

ARIMA(2, 1, 3)(1, 1, 0)[52]


N05 Forecast: Last 11 Weeks of 2019

ARIMA(3, 1, 2)


R0 Forecast: Last 11 Weeks of 2019

ARIMA(0, 1, 2)(0, 0, 1)[52]

Conclusion

The time series analysis of pharmaceutical drug sales provides information allowing for inferences regarding trends and patterns over time. Since analysis has been completed, the research questions can now formally be answered.

Research Question 1

How can we deal with outliers in a time series?

In each time series, outliers were identified and replaced using a robust STL decomposition. STL was selected since it is robust to any type of seasonality due to local estimation. Outliers are identified and replaced with the predicted value from Loess regression which follows the general trend and periodicity of the series.

Research Question 2

For each drug class, the last 11 weeks of 2019 will be forecasted. Are the forecasts reasonable?

The predicted forecasts for each time series seem to capture the potential values the series might take. For the predictions that do not look like the actual series, the prediction intervals seem wide enough to capture values that would be expected in the future.

Research Question 3
For each drug class, are there any life events that could affect the volume of pharmaceutical drug sales?

Given the nature of this study, there are many alternative remedies to ailments people receive. The legalization of marijuana could potentially have a large effect on the N02 and N05 drug classes because it is known to reduce swelling and help with mental illnesses. Psychedelic drugs are also currently going through clinical trials to be used for mental illnesses especially those who suffer from PTSD. It is difficult to determine why pharmaceutical drug sales change over time inconsistently because of the abundance of factors that could affect drug usage.

Concluding Thoughts
This study provides endless potential for future research both related to time series analysis and pharmaceutical drugs. Since multiple drugs were considered, they can be modeled as a multivariate time series or with non-time-based predictor variables as well. It may also be interesting to analyze these data over a longer period of time to see if there are potentially multiple different seasonal components that could help improve model performance.

Aside from *ARIMA* models, there are many other time series models that may be suited for these data as well. Spectral analysis can be useful in this case as well because it can remove multiple frequencies of periodicity before the series is modeled. Then once analysis and predictions have been made, the periodic values can be added back to give the predictions values that align accurately with the true values.

Appendices

Appendix A: Software Information

This project was completed using the R statistical software and the following libraries loaded in the order they are listed.

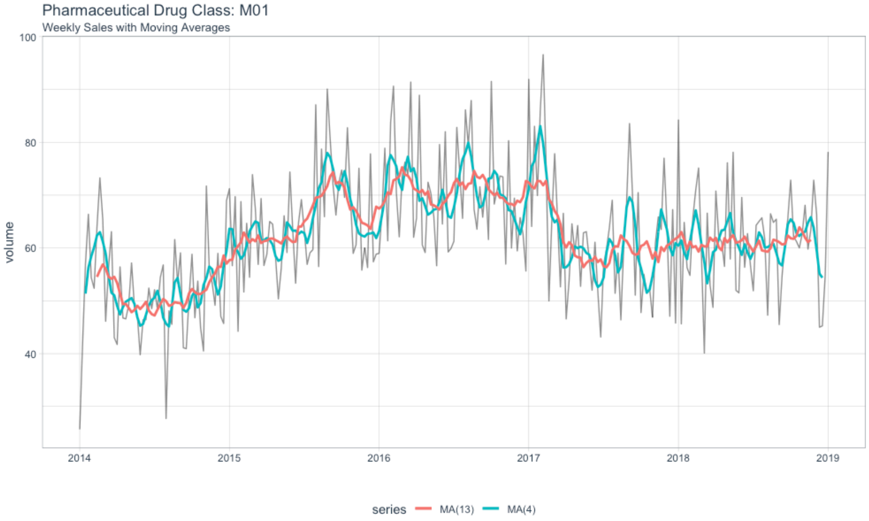
1. tidyverse
2. tidyquant
3. timetk
4. tseries
5. astsa
6. sweep
7. forecast
8. ggfortify

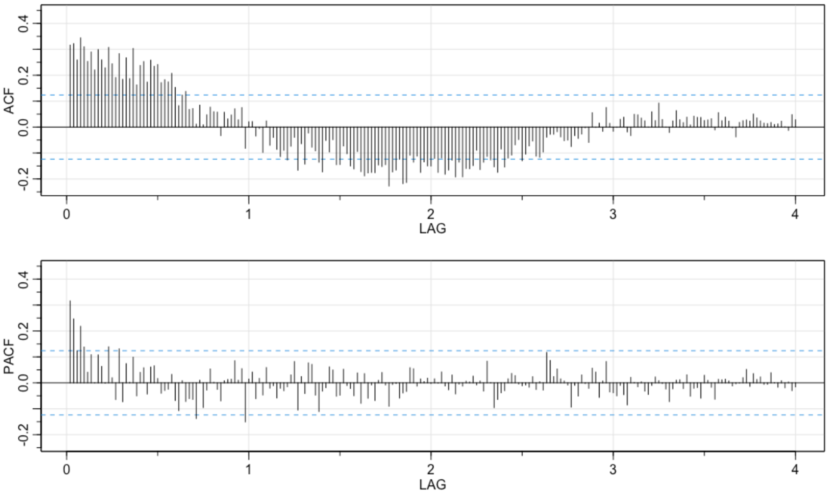
The libraries must be loaded in this order to avoid conflicting functions.

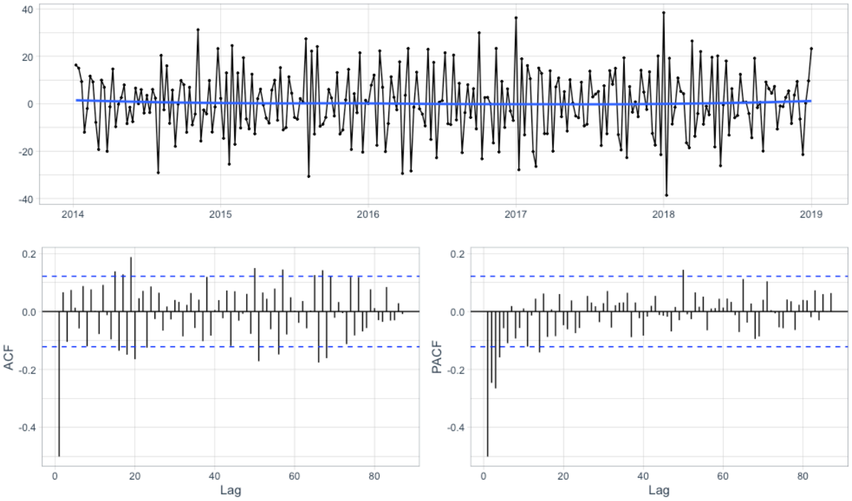
The code for this project can be found in GitHub repository of the link below.

github.com/Applied_Statistics_Grad_Projects

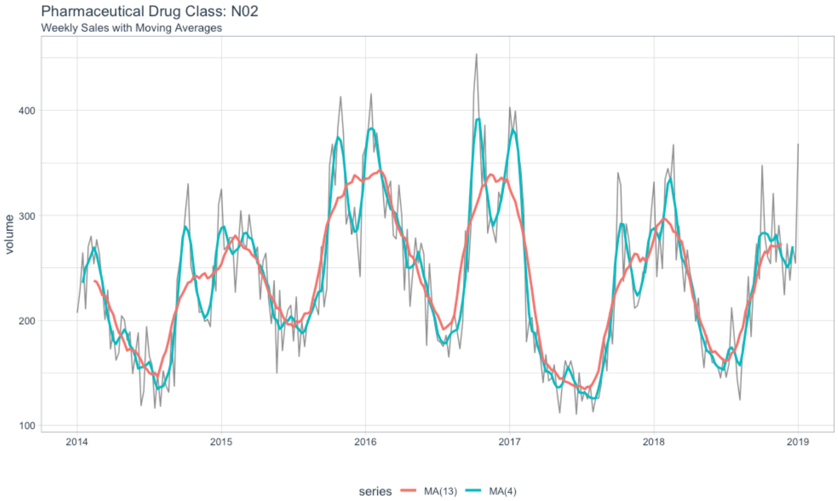
Appendix B: M01 Additional Outputs

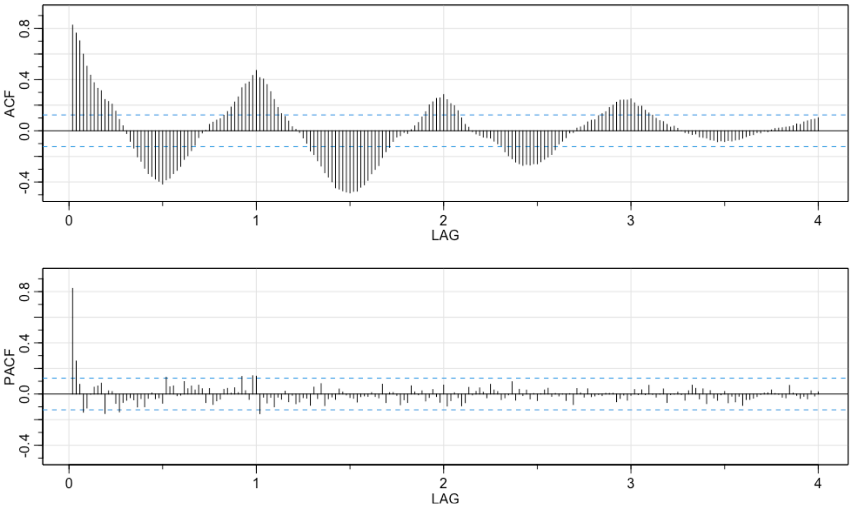


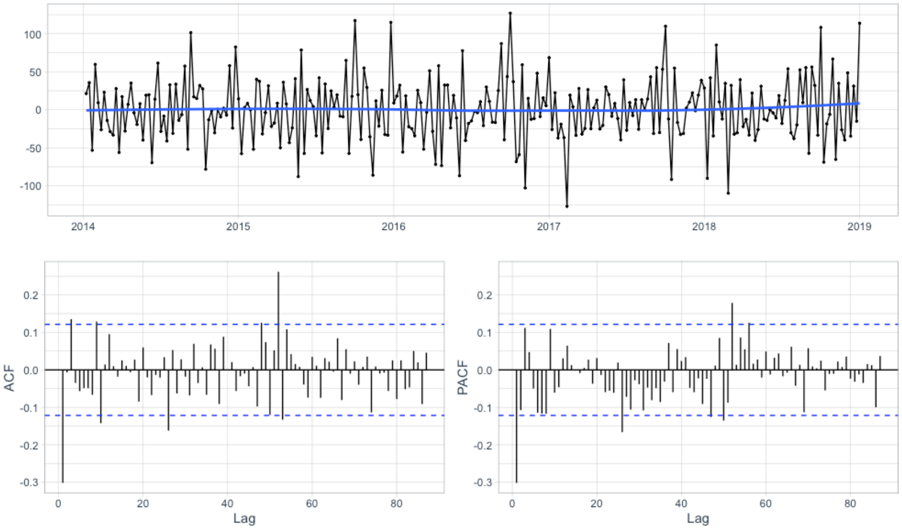




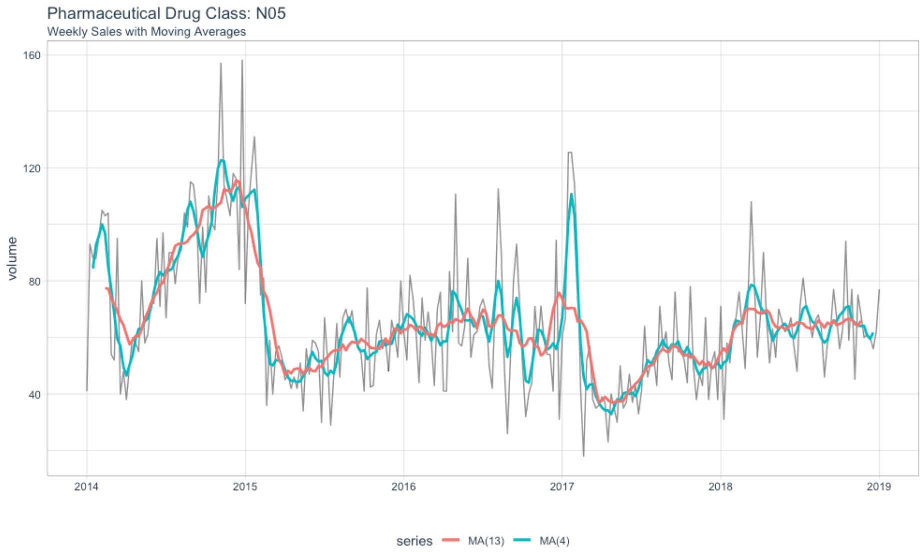
Appendix C: N02 Additional Outputs





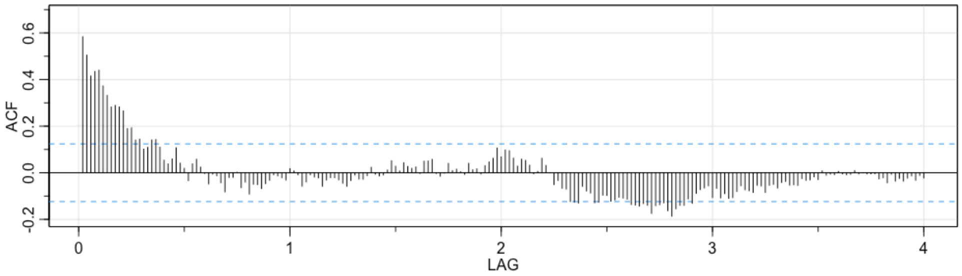


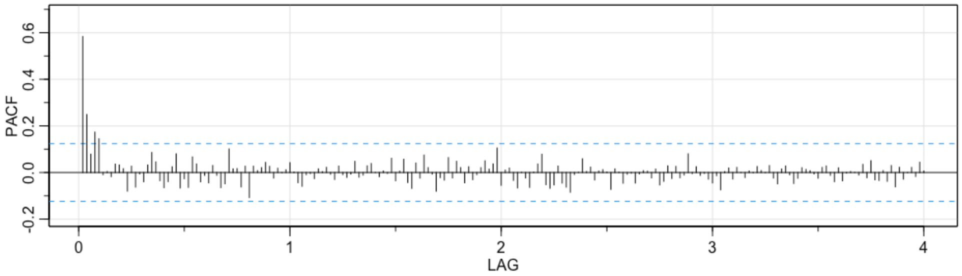
Appendix D: N05 Additional Outputs

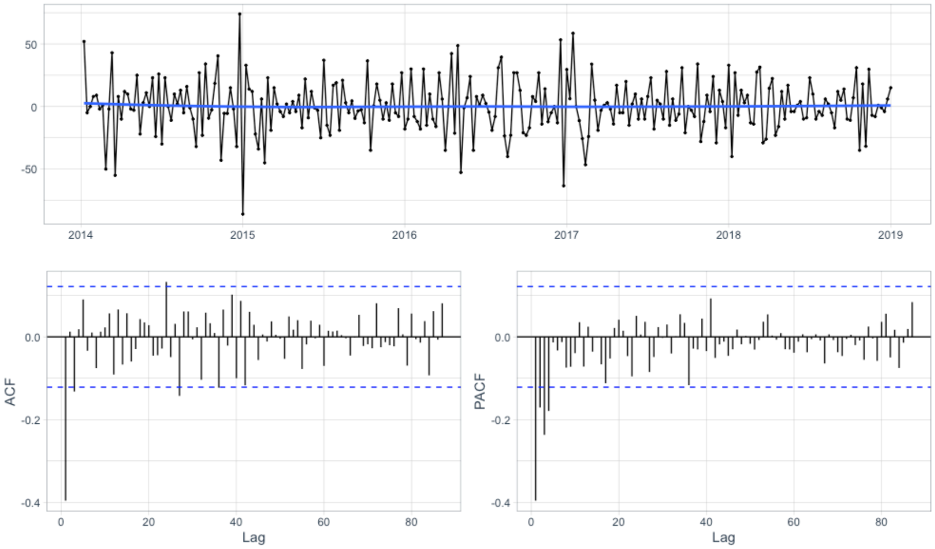


Pharmaceutical Drug Class: N05

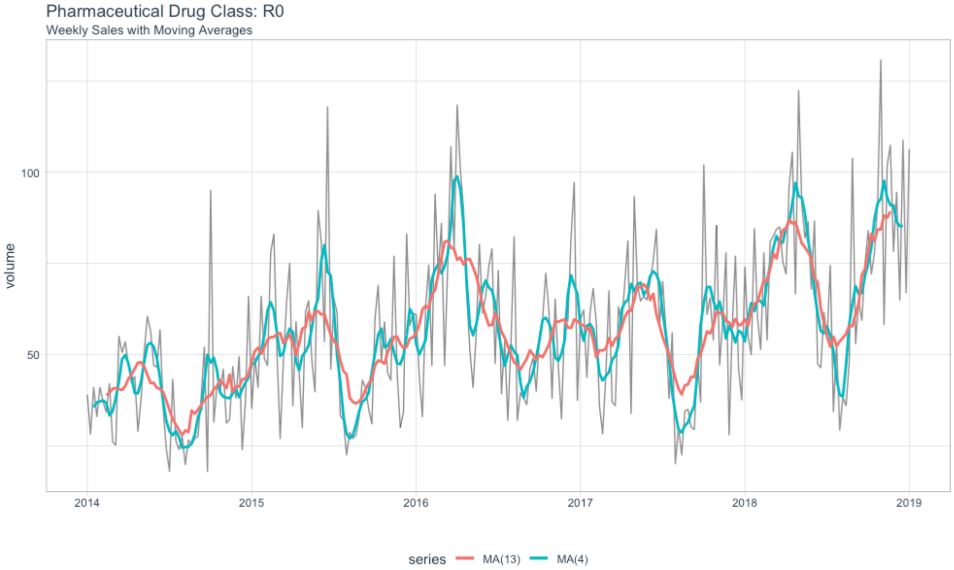
Weekly Sales with Moving Averages

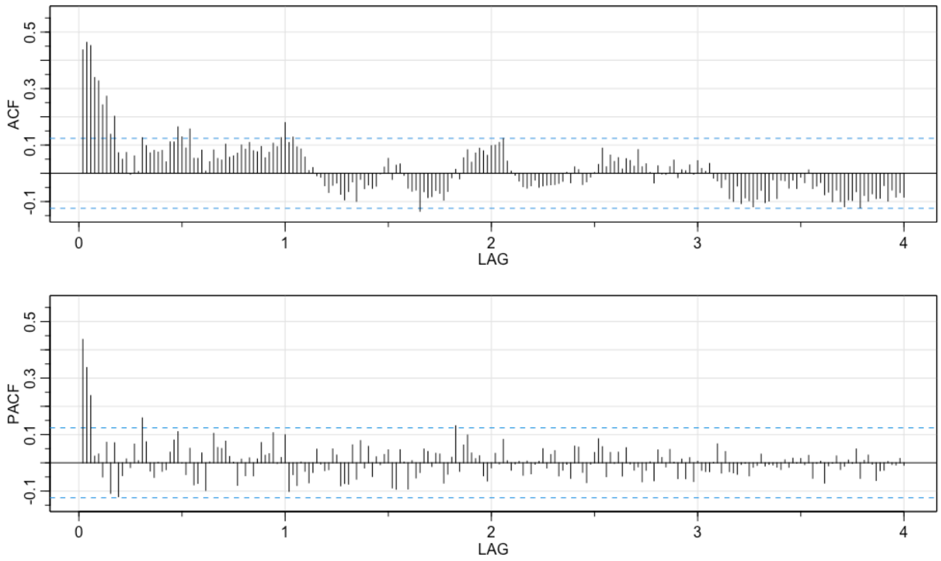


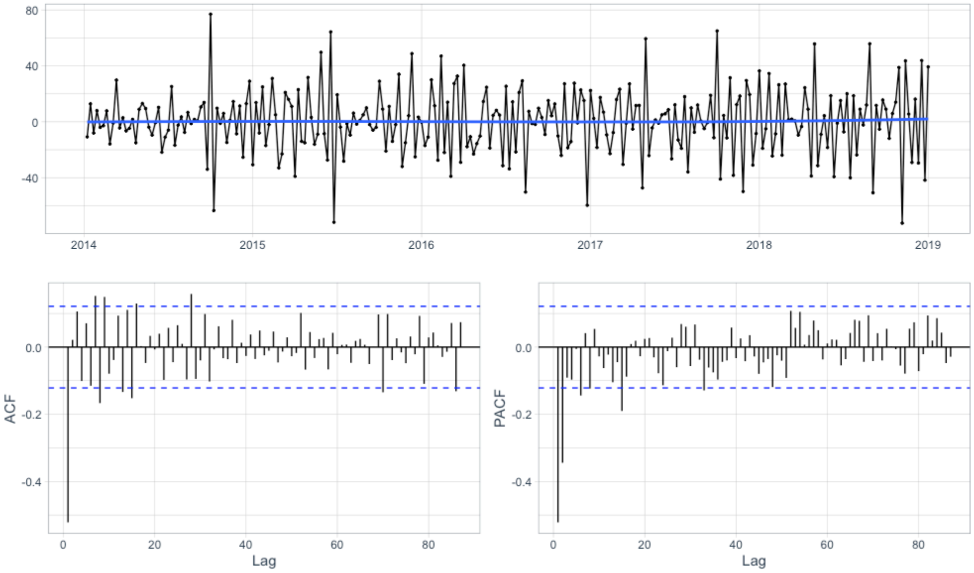




Appendix E: R0 Additional Outputs




ACF
0.5
0.3
0.1
-0.1
0
1
2
3
4
LAG
PACF
0.5
0.3
0.1
-0.1
0
1
2
3
4
LAG


80
40
0
-40
2014
2015
2016
2017
2018
2019
ACF
0.0
-0.2
-0.4
0
20
40
60
80
Lag
PACF
0.0
-0.2
-0.4
0
20
40
60
80
Lag