TIME SERIES OF PHARMACEUTICAL DRUG SALES

Hamzah Sami

Stat 580

Introduction

Industries often evolve as a result of the times and the consumers they are trying to appease. These changes often mean improvements, but the trend in progress can be hard to measure. This is especially true for the pharmaceutical industry. Pharma companies have a notorious reputation for selling drugs and medications that cause addiction and other issues. However, these accusations only apply to a select number of drugs and have become a recent issue subject to legislation. Pharma companies and drug use have become an issue stemming from the 1960s and even further. In the past 60 years, there have been a myriad of laws, new customer bases, and advanced medications. While we could study all these factors and their effect on the pharmaceutical industry, our goal for this project is to ultimately study pharmaceutical sales data from 2012 to 2019. While studying data for seven years may seem questionable, the reason for doing so is because significant legislation have been passed during this timeframe of interest that may have affected pharmaceutical sales. Our goal is to learn about these effects and to model trends on multiple classes of drugs that were being sold. In doing so, we hope to develop greater insight on the pharmaceutical market and to learn about the market behavior for certain types of drugs.

Dataset

The dataset being used for this report was obtained from Kaggle. The data discusses pharmaceutical sales over the past 8 years (2012-2020). Selected drugs in this particular dataset are grouped by class according to the Anatomical Therapeutic Chemical (ATC) Classification System. The variables in the dataset are as follows:

M01AB: anti-inflammatory and anti-rheumatic products, non-steroids, acetic acid derivatives and related substances
M01AE: anti-inflammatory and anti-rheumatic products, non-steroids, propionic acid derivatives
N02BA: other analgesics and antipyretics, salicylic acid and derivatives
N02BE: other analgesics and antipyretics, pyrazolones and anilides
N05B: psycholeptic drugs, anxiolytic drugs
N05C: psycholeptic drugs, hypnotics and sedative drugs
R03: drugs for obstructive airway diseases
R06: antihystamines for systematic use

From looking at the variables, we see that there is some overlap between variables. For example, N05B and N05C are both different classes of psycholeptic drugs. To account for this overlap, we decided to reclassify the variables into four classes: M01, N02, N05, and R0, and Total (a composite variable consisting of the four previously mentioned variables). These classes enable us to look for trends more easily in the data and to also ascertain any real-life events that played a part in affecting the behavior of these variables.

Research Questions

1)How can outliers be removed from a time series?
2)For each class of drugs, the last 11 weeks of 2019 have been forecasted. Do the forecasts make sense?
3)For each class of drugs, are there any life events that could affect how the volume of pharmaceutical drug sales?

Exploratory Data Analysis

When we plotted the time series plot for the four variables with loess regression for trend estimation, we note that M01 and R0 have upward trends but there is a decline from 2016 and 2017 for both. With regards to the N02 variable, the time series plot is very inconsistent which makes it difficult to ascertain a trend. Unlike the other three variables, the N05 variable has a downward trend from 2014-2016 and flattens out after the decline.
As a result, it was difficult to determine a trend for these four variables. The only variable of note that had a real trend we could analyze was R0.

When we plotted the ACF and PACF for each variable, we observe that the N02 variable is somewhat periodic. We attributed this periodicity to a seasonality effect which was reflected in our plot having a seasonality effect for 52 weeks.

From running the Dickey-Fuller test, we see that the R0 variable is the only stationary variable. It is of note that for the other three variables, we had to check the second and even third order differences to see whether the time series for M01, N02, and N05 were stationary.

Model Selection

For our project, we chose to make models for the four variables: M01, N02, N05, and R0. Initially we attempted to make a model consisting of all four variables but it became too difficult to model and analyze. The reason for this is because the N02 had the largest amount of data and behavior to track and this wound up dominating the other variables. We used the ARIMA model to track the behavior of these four variables and below, we'll explain how we found the model for each.

For all four models, we trained the data starting from January 1st, 2014 through December 30th, 2018 and then we tested the data from observations starting on January 6th, 2019 through October 13th 2019

MO1

When we ran the models for M01, we ended up getting three models: ARIMA(2,1,1), ARIMA(3,1,1), and ARIMA(4,1,1). Upon running the summary function for each fitted model,

we chose the ARIMA(2,1,1) to forecast our data because it had the lowest AIC, AICC, and BIC values of the three models.

N02

When we ran the models for N02, we ended up getting three models: ARIMA(2, 1, 0)(1, 1, 0), ARIMA(2, 1, 3)(1, 1, 0), and ARIMA(0, 1, 1)(1, 1, 0). Upon running the summary function for each fitted model, we chose the ARIMA(2,1,0)(1,1,0) to forecast our data because it had the lowest AIC, AICC, and BIC values of the three models.

N05

When we ran the models for N05, we ended up getting three models: ARIMA (0, 1, 3)(1, 0, 0), ARIMA(3, 1, 2), and ARIMA(3, 1, 2)(1, 0, 1). Upon running the summary function for each fitted model, we chose the ARIMA(0,1,3)(1,0,0) to forecast our data because it had the lowest AIC, AICC, and BIC values of the three models.

R0

When we ran the models for R0, we ended up getting three models: ARIMA(0, 1, 2)(0, 0, 1), ARIMA(0, 1, 3)(0,0,1), and ARIMA(3, 1, 0)(1, 0, 0). Upon running the summary function for each fitted model, we chose the ARIMA(0,1,2)(0,0,1) to forecast our data because it had the lowest AIC, AICC, and BIC values of the three models.

Residual Analysis and Forecasts

With regards to the residual analysis for all four variables, it was apparent that the histograms and ACF plots were all centered at zero for each of the variables. In such a situation, we can conclude that there was white noise error. In addition, N02 plot had a seasonal period of 52 weeks which we ended up seeing after observing periodicity in our residuals. Furthermore, it is of note that for each model corresponding to its own variable, the root mean squared error was minimal and close in value to the other types of error we checked for. This indicates that we had no overfitting in our model.

For all forecasts, we assessed the first 41 weeks of 2019 so that we could compare it to actual data and account for error. After doing so, we forecasted for 11 weeks into 2020 to make predictions.

M0:

From plotting the M0 forecast, we can see that the forecast is constant for 41 weeks and the prediction interval even contains negative values. In spite of this, the error is low and based off the prediction on the data for 11 weeks in 2020, we predict that there is a slight increase coming in 2020 based off the data provided.

N02:

From plotting the N02 forecast, we can see that the forecast is declining steeply for 41 weeks and the prediction interval contains the time series. When we go to make a prediction on the data for 11 weeks in 2020, we predict that there is a steep increase coming in 2020 based off the data provided.

N05:

From plotting the N05 forecast, we can see that the forecast is constant for 41 weeks and the prediction interval contains the time series. When we go to make a prediction on the data for 11 weeks in 2020, we predict that there is a minimal increase in 2020 but the forecast will be mostly constant based off the data provided.

R0:

From plotting the R0 forecast, we can see that the forecast follows the trend similar to what occurred in previous years for 41 weeks and the prediction interval contains the time series. When we go to make a prediction on the data for 11 weeks in 2020, we predict that there is a a mostly constant trend in 2020 that resembles the behavior modeled in the time series from 2012-2019.

Answers to Research Questions

1)In order to verify whether outliers were affecting our ability to make proper predictions, we ended up running the fitted model for all four variables with and without STL Decomposition. Recall STL Decomposition stands for Seasonal and Trend Decomposition, a robust method that uses Loess regression to remove outliers. This was especially handy as STL Decomposition can handle periodicity including seasonal components that change over time. By using such a method, we observed that there were no significant changes reflected in the ACF and PACF plots. As a result, we were able to conclude that outliers did not affect our ability to make forecasts on the data.

2)The forecasts were reasonable for the four variables. They are as follows:

M01: M01 had a rather constant forecast and this was reflected in having a small prediction interval which could have been explained. Since M01 consists of drugs that reduce swelling and ibuprofen, these drugs may have a small prediction interval because of other treatments becoming more popular. As of 2016, the legalization of marijuana has led to CBD and other natural maladies becoming popular options.

N02: Like MO1, N02 had a constant forecast with peaks in 2016, 2017, and 2019. However, since 2019, there has been a decrease in usage and so N02 has a small prediction interval. N02 are painkilling drugs and this reduced prediction interval can also be attributed to the legalization of marijuana. Painkillers have also gained bad press for the past four years because these types of

drugs allegedly promote addicting behaviors that lead to other harmful conditions. The bad press and increase of new options and treatments have possibly affected the prediction interval.

N05: The time series of N05 has been constant that initially had increasing behavior but over time, there is a decline from 2015 and onward. N05 refers to psycholeptic drugs such as antidepressants and SSRI medications. From analyzing the forecast of N05, we can see N05 has a wider prediction interval than M01 and N02. The wide prediction interval is an interesting result and could be explained by analyzing big events that impacted the lives of many. The fact that the prediction interval is wide could suggest that mental states of people are hard to measure on a large scale and as a result of that, it becomes difficult to ascertain sales.

R0: R0's time series appears to rise until 2018 where it declines sharply after 2018. The forecast also has the largest prediction interval of the four variables As stated previously, R0 refers to allergy medications and antihistamines and the size of the prediction interval indicated by the forecast could be explained by the fact that allergies are seasonal and depending on the season, a certain type of drug will be sold in the market place. Furthermore, other natural maladies have entered the marketplace which makes competition among these drugs a hindrance for companies trying to sell medications. It is of note that the prediction interval for R0 does contain negative values. While it doesn't affect our forecast and prediction interval, it is a worthwhile observation that could have been worse had it gone unnoticed.

Overall, the forecasts were reasonable for these four drugs. Real life events and perhaps the introductions of new drugs in the marketplace are responsible for affecting the trajectories of the time series for each.

3)M01 refers to antirheumatic and anti-inflammatory products so the reason for why there has been a decline in the sales of such products could be attributed to new products such as CBD entering the marketplace that are affecting the sales for current products. The legalization of marijuana in states beginning from 2016 does support the intuition behind the decline indicated in the time series.

N02, painkilling drugs, have also been impacted by the legalization of marijuana. The decline in sales from 2016 and onward does coincide with the legalization of marijuana in 2016.

N05, psycholeptic drugs, have also been affected by the introduction of cannabis and CBD products entering the marketplace. Furthermore, the increase in the sales of these products from 2015-2016 could be attributed to hostility stemming from the 2016 election. In addition, new treatments such as the micro dosing of psychedelics and therapy have become viable options for consumers to consider.

R0, drugs for obstructive airway disease, have increased sales from 2016 for a number of reasons. The rise of allergy medication as well as the increase in use of vaping products among young people may be the underlying reason behind all this.

While all these explanations are viable reasons for the behavior in sales of each drugs, it is important to consider significant events such as the covid-19 pandemic as well as the rise in

automation which have decimated jobs and might be forcing people into either not being able to afford medication or engaging in harmful behaviors.

Conclusion

Analyzing this dataset was a good experience for my group because it enabled us to consider the ramifications of legislation and other external factors on the selling of pharmaceutical drugs. Being able to analyze the trend behind painkillers and fever medication and linking that to the legalization of marijuana was an eye-opening insight we were able to make. We were also able to observe how much the conversation and depiction of mental health has affected drug sales for antidepressants and other mood-altering drugs. As seen from our data, there was a drop-off in sales from 2016 and onward and we believe that this was in part due to therapy becoming more of a factor in treatment. For additional work, we would want to observe if whether costs for fitness equipment and other options that enable a healthy lifestyle.

We would also want to consider data from the year 2020 as we would be interested in delving deeper into how covid-19 affected pharmaceutical sales this year. We believed as a group, there would be a decrease in sales since companies are currently focused on launching a vaccine for covid-19.

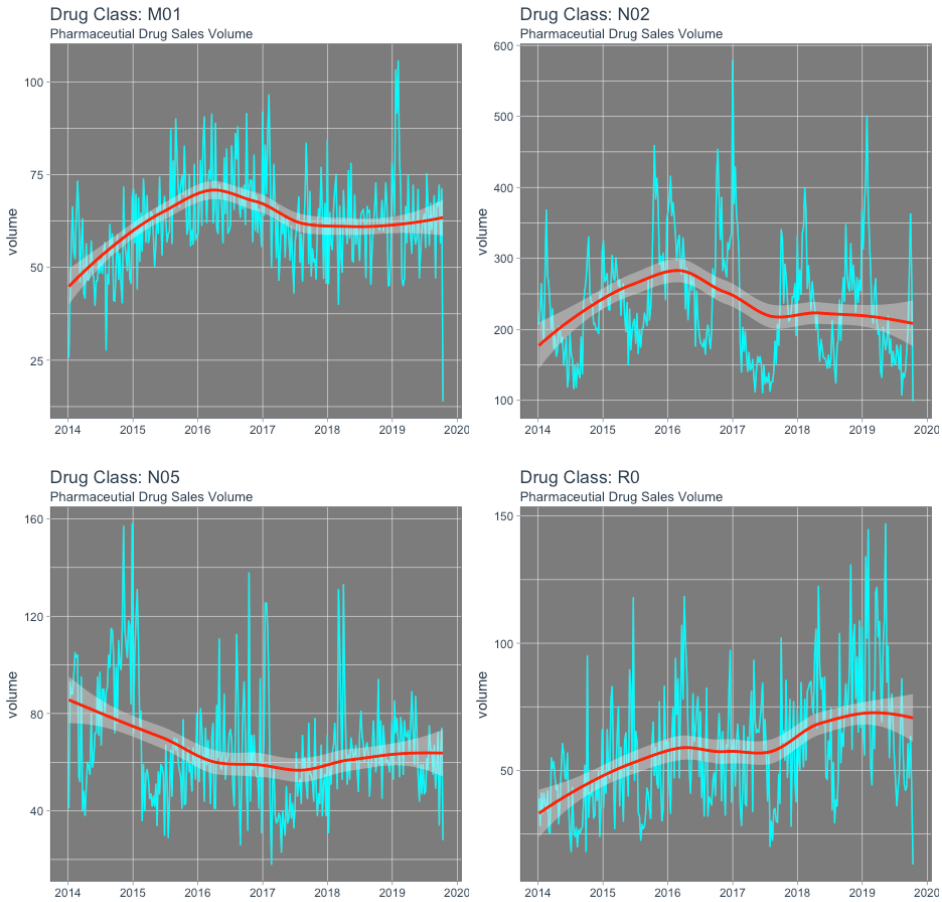
Overall, I found this project to be a good application of my learning time series. I believe it's allowed me to take steps forward in my learning time series that may not have been there initially. I hope that this project is something that I can continue to build upon as time progresses toward the end of 2020 because the changes in legislation and products will be significant factors to assess beyond 2020.

Appendix

```
#------------------------------------------------------------------------------

# STAT 580 TIME SERIES ANALYSIS GROUP PROJECT
# PRESENTATION OUTPUTS AND IMAGES

#------------------------------------------------------------------------------

# FUNCTIONS

#------------------------------------------------------------------------------

# load libraries
load_libraries = function(){
  library(tidyverse); library(tidyquant); library(timetk);
  library(tseries); library(astsa); library(sweep); library(forecast);
  library(ggfortify);
}

# data
load_weekly_data = function(path = '/Users/hamzah/documents/csulb statistics/stat
580/580 final project/salesweekly.csv', print = T){
  read_csv(path) %>%
    mutate(datum = mdy(datum)) %>%
    rename(date = datum) %>%
    mutate(total = M01AB+M01AE+N02BA+N02BE+N05B+N05C+R03+R06,
        M01 = M01AB+M01AE, N02 = N02BA+N02BE,
        N05 = N05B+N05C, R0 = R03+R06) %>%
    select(date, total, M01, N02, N05, R0) -> df
  if (print){ df %>% print() }
  return(df)
}

# variable descriptions
variable_descriptions = function(){
  cat('\nPharmaceutical Drug Sales Volumes',
    '\n-------------------------------------------------------------',
    '\ntotal: total drug sales volume',
    '\nM01  : anti-inflammatory and antirheumatic products',
    '\nN02  : other analgesics and antipyretics',
    '\nN05  : psycholeptics drugs',
    '\nR0   : antihistamines and drugs for obstructive airway diseases',
    '\n-------------------------------------------------------------')
}
```

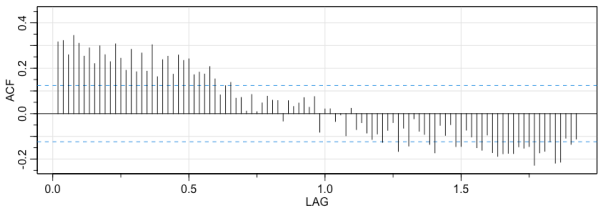
```
#------------------------------------------------------------------------------

# set up
load_libraries()
library(ggpubr)

# load data
load_weekly_data() -> df

# plot all time series
df %>%
  ggplot(aes(x = date, y = M01)) +
  geom_line(color = 'cyan') +
  geom_smooth(method = 'loess', color = 'red', fill = 'gray90') +
  ggtitle('Drug Class: M01', subtitle = 'Pharmaceutial Drug Sales Volume') +
  labs(x = '', y = 'volume') +
  scale_x_date(date_breaks = '1 year', date_labels = '%Y') +
  theme_tq_dark() -> p_m01
df %>%
  ggplot(aes(x = date, y = N02)) +
  geom_line(color = 'cyan') +
  geom_smooth(method = 'loess', color = 'red', fill = 'gray90') +
  ggtitle('Drug Class: N02', subtitle = 'Pharmaceutial Drug Sales Volume') +
  labs(x = '', y = 'volume') +
  scale_x_date(date_breaks = '1 year', date_labels = '%Y') +
  theme_tq_dark() -> p_n02
df %>%
  ggplot(aes(x = date, y = N05)) +
  geom_line(color = 'cyan') +
  geom_smooth(method = 'loess', color = 'red', fill = 'gray90') +
  ggtitle('Drug Class: N05', subtitle = 'Pharmaceutial Drug Sales Volume') +
  labs(x = '', y = 'volume') +
  scale_x_date(date_breaks = '1 year', date_labels = '%Y') +
  theme_tq_dark() -> p_n05
```

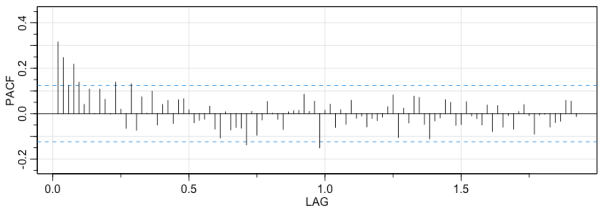
```
df %>%
   ggplot(aes(x = date, y = R0))
+
   geom_line(color = 'cyan') +
   geom_smooth(method =
'loess', color = 'red', fill =
'gray90') +
   ggtitle('Drug Class: R0',
subtitle = 'Pharmaceutial Drug
Sales Volume') +
   labs(x = '', y = 'volume') +
   scale_x_date(date_breaks =
'1 year', date_labels = '%Y') +
   theme_tq_dark() -> p_r0

# plot all series in a grid
ggarrange(p_m01, p_n02,
p_n05, p_r0)

# split data
df %>% filter(date < '2019-
01-06') -> train
df %>% filter(date >= '2019-
01-06') -> test
```




```
#-------------------------------------------------------------------------------

# M01 EDA and Model Selection

#-------------------------------------------------------------------------------

# create time series object
train %>%
   tk_ts(select = M01, start = 2014, frequency = 52, silent = T) %>%
   tsclean() -> x
#outlier check
```

```
train %>%
  tk_ts(select = M01, start = 2014, frequency = 52,
silent = T) -> x

# ACF and PACF
x %>% acf2(max.lag = 100, main = '')
```

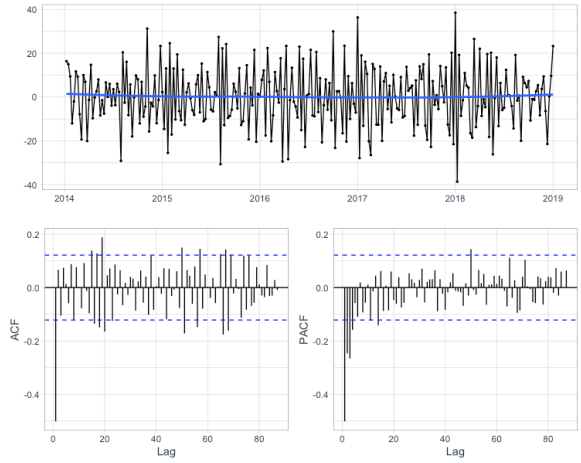





```
# ADF test
x %>% adf.test(alternative = 'stationary')

# estimating difference order
ndiffs(x)
nsdiffs(x)
```

```
        Augmented Dickey-Fuller Test

data:  .
Dickey-Fuller = -2.9582, Lag order = 6, p-value = 0.1724
alternative hypothesis: stationary
```

```
# plot time series, ACF, and PACF
diff(x) %>% ggtsdisplay(plot.type = 'partial', smooth
= T, theme = theme_tq())
```

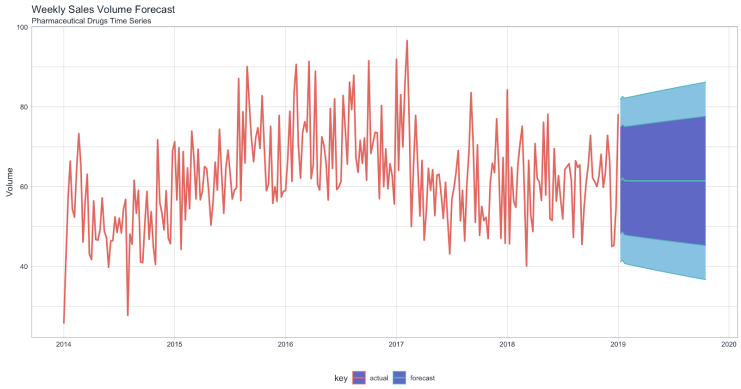



```
# model 1
x %>% Arima(order = c(2, 1, 1)) -> fit1
fit1 %>% residuals() %>%
   ggtsdisplay(plot.type = 'histogram', smooth = T,
theme = theme_tq())
fit1 %>% residuals() %>% adf.test(alternative =
'stationary')

# model 2
x %>% Arima(order = c(3, 1, 1)) -> fit2
fit2 %>% residuals() %>%
   ggtsdisplay(plot.type = 'histogram', smooth = T, theme = theme_tq())
fit2 %>% residuals() %>% adf.test(alternative = 'stationary')

# model 3
x %>% Arima(order = c(4, 1, 1)) -> fit3
fit3 %>% residuals() %>%
   ggtsdisplay(plot.type = 'histogram', smooth = T, theme = theme_tq())
fit3 %>% residuals() %>% adf.test(alternative = 'stationary')

# select best model by summary
fit1 %>% summary()
fit2 %>% summary()
fit3 %>% summary()

# select best model
```

```
x %>% Arima(order = c(2, 1, 1), seasonal = c(0,
0, 0)) -> fit
fit %>% summary()
```

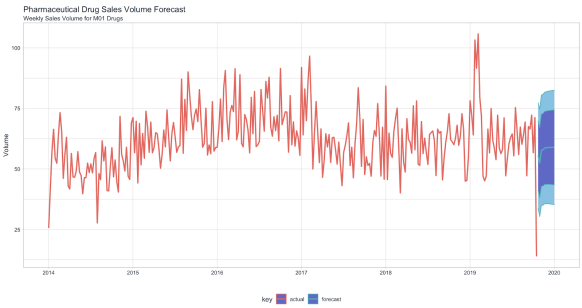
```
Series: .
ARIMA(2,1,1)

Coefficients:
          ar1     ar2      ma1
      0.0182  0.0397  -0.8989
s.e.  0.0716  0.0702   0.0320

sigma^2 estimated as 108.2:  log likelihood=-977.15
AIC=1962.3   AICc=1962.45   BIC=1976.54

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set 0.5922022 10.32341 8.218401 -1.494926 13.58374 0.6432362 0.01414144
```

```
# analyze residuals
residuals(fit) %>% adf.test(alternative =
'stationary')
residuals(fit) %>%
  ggtsdisplay(plot.type = 'histogram', smooth = T, theme = theme_tq())

# predict forecast
fit %>% forecast(h = 41) -> fc

# evaluate forecast
fc %>% accuracy(test$M01)

# plot forecast
sw_sweep(fc) %>%
  ggplot(aes(x = index, y = M01, color = key)) +
  geom_ribbon(aes(ymin = lo.95, ymax = hi.95), fill = 'skyblue') +
  geom_ribbon(aes(ymin = lo.80, ymax = hi.80), fill = 'slateblue', alpha = 0.7) +
  geom_line(size = 1) +
  labs(title = 'Weekly Sales Volume Forecast', x = '', y = 'Volume',

      subtitle = 'Pharmaceutical Drugs Time Series') +
  scale_x_yearmon(n = 6, format = '%Y') + theme_tq()

# forecasting the rest of 2019

df %>%
  tk_ts(select = M01, start = 2014, frequency =
52, silent = T) %>%
  tsclean() -> x
```

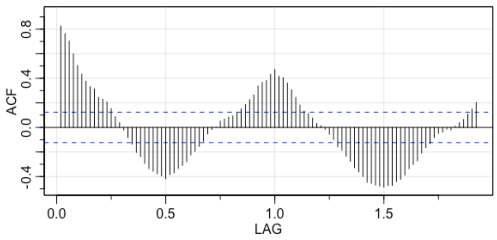


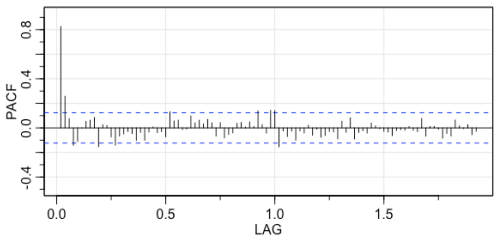

```
# define model
x %>% Arima(order = c(2, 1, 1), seasonal =
c(0, 0, 0)) -> fit
fit %>% summary()

# analyze residuals
residuals(fit) %>% adf.test(alternative = 'stationary')
residuals(fit) %>%
  ggtsdisplay(plot.type = 'histogram', smooth = T, theme = theme_tq())

# predict forecast
```

```
fit %>% forecast(h = 11) -> fc

# plot forecast
sw_sweep(fc) %>%
  ggplot(aes(x = index, y = M01, color = key)) +
  geom_ribbon(aes(ymin = lo.95, ymax = hi.95), fill =
'skyblue') +
  geom_ribbon(aes(ymin = lo.80, ymax = hi.80), fill =
'slateblue', alpha = 0.7) +
  geom_line(size = 1) +
  labs(title = 'Pharmaceutical Drug Sales Volume Forecast', x = '', y = 'Volume',
       subtitle = 'Weekly Sales Volume for M01 Drugs') +
  scale_x_yearmon(n = 6, format = '%Y') + theme_tq()
```



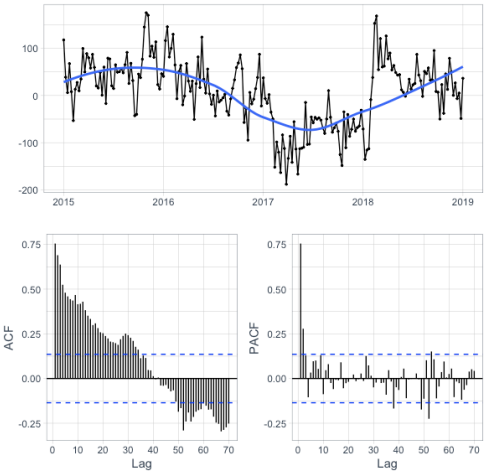

```
#--------------------------------------------------------------------------------


# N02 EDA and Model Selection

#--------------------------------------------------------------------------------

# create time series object
train %>%
  tk_ts(select = N02, start = 2014, frequency = 52, silent =
T) %>%
  tsclean() -> x

#outlier
train %>%

 tk_ts(select = N02, start = 2014, frequency = 52, silent =
T) -> x

# ACF and PACF
x %>% acf2(max.lag = 100, main = '')

# ADF test
x %>% adf.test(alternative = 'stationary')

# estimating difference order
ndiffs(x)
nsdiffs(x)

# plot time series, ACF, and PACF (???)
diff(x, lag = 52) %>% ggtsdisplay(plot.type = 'partial', smooth = T, theme = theme_tq())
```






```
        Augmented Dickey-Fuller Test

data:  .
Dickey-Fuller = -3.4055, Lag order = 6, p-value = 0.05403
alternative hypothesis: stationary
```

```
# model 1

x %>% Arima(order = c(2, 1, 0), seasonal = c(1, 1, 0)) -> fit1
fit1 %>% residuals() %>%
  ggtsdisplay(plot.type = 'histogram', smooth = T, theme =
theme_tq())
fit1 %>% residuals() %>% adf.test(alternative =
'stationary')
```

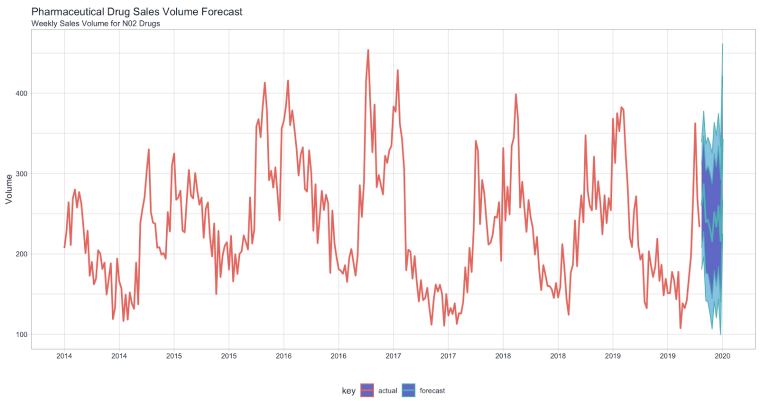



```
# model 2
x %>% Arima(order = c(2, 1, 3), seasonal = c(1, 1, 0)) -> fit2
fit2 %>% residuals() %>%
  ggtsdisplay(plot.type = 'histogram', smooth = T, theme =
theme_tq())
fit2 %>% residuals() %>% adf.test(alternative = 'stationary')
```

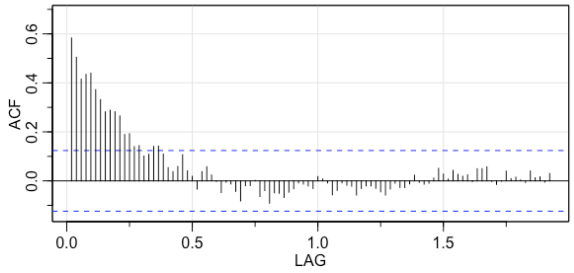
```
# model 3
x %>% Arima(order = c(0, 1, 1), seasonal = c(1, 1, 0)) ->
fit3
fit3 %>% residuals() %>%
  ggtsdisplay(plot.type = 'histogram', smooth = T, theme
= theme_tq())
fit3 %>% residuals() %>% adf.test(alternative
= 'stationary')
```

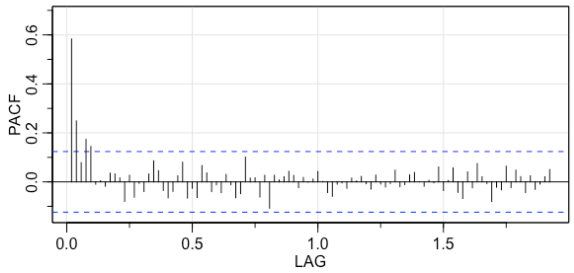
```
Series: .
ARIMA(2,1,0)(1,1,0)[52]

Coefficients:
         ar1      ar2     sar1
      -0.4046  -0.1586  -0.4743
s.e.   0.0692   0.0689   0.0687

sigma^2 estimated as 1605:  log likelihood=-1067.98
AIC=2143.96   AICc=2144.16   BIC=2157.31

Training set error measures:
                     ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set -0.9804172 35.50922 25.20443 -1.244968 10.84608 0.4397098 0.01058886
```


```
# select best model by summary
fit1 %>% summary()
fit2 %>% summary()
fit3 %>% summary()
```

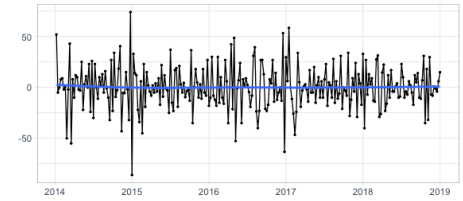


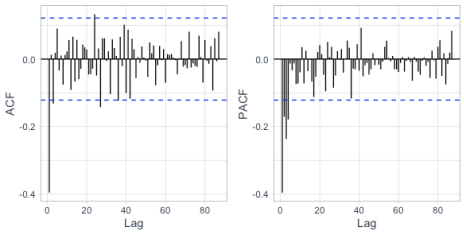
```
#------------------------------------------------------------------------------

# N05 EDA and Model Selection

#------------------------------------------------------------------------------

# create time series object
train %>%
   tk_ts(select = N05, start = 2014, frequency = 52, silent
= T) %>%

   tsclean() -> x

#uncleaned series
train %>%
 tk_ts(select = N05, start = 2014, frequency = 52, silent
= T) -> x

# ACF and PACF
x %>% acf2(max.lag = 100, main = '')

# ADF test
x %>% adf.test(alternative = 'stationary')

# estimating difference order
ndiffs(x)
nsdiffs(x)

# plot time series, ACF, and PACF
diff(x) %>% ggtsdisplay(plot.type = 'partial', smooth = T,
theme = theme_tq())

# model 1

x %>% Arima(order = c(0, 1, 3), seasonal = c(1, 0, 0)) ->
fit1
fit1 %>% residuals() %>%
   ggtsdisplay(plot.type = 'histogram', smooth = T, theme =
theme_tq())
```





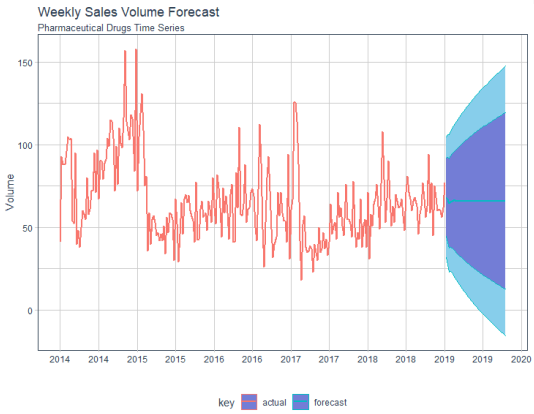

```
        Augmented Dickey-Fuller Test

data:  .
Dickey-Fuller = -3.3689, Lag order = 6, p-value = 0.06014
alternative hypothesis: stationary
```



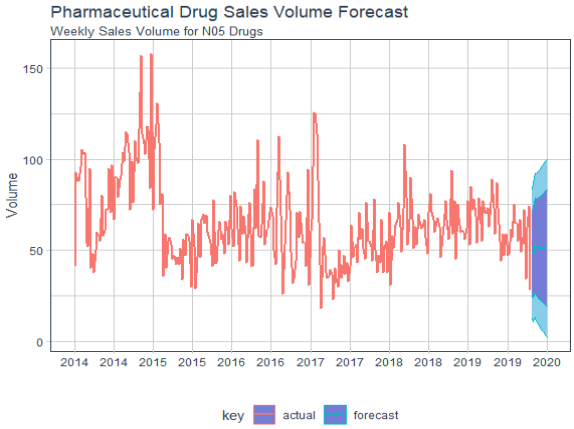


```
fit1 %>% residuals() %>% adf.test(alternative = 'stationary')

# model 2
x %>% Arima(order = c(3, 1, 2), seasonal = c(0, 0, 0))
-> fit2
fit2 %>% residuals() %>%
   ggtsdisplay(plot.type = 'histogram', smooth = T,
theme = theme_tq())
fit2 %>% residuals() %>% adf.test(alternative =
'stationary')

# model 3
x %>% Arima(order = c(3, 1, 2), seasonal = c(1, 0, 1)) ->
fit3
fit3 %>% residuals() %>%
   ggtsdisplay(plot.type = 'histogram', smooth = T, theme =
theme_tq())
fit3 %>% residuals() %>% adf.test(alternative =
'stationary')
```

```
Series: .
ARIMA(0,1,3)(1,0,0)[52]

Coefficients:
          ma1      ma2      ma3    sar1
      -0.5838  -0.0478  -0.1107  0.0893
s.e.   0.0635   0.0686   0.0715  0.0691

sigma^2 estimated as 342.8:  log likelihood=-1126.29
AIC=2262.59   AICc=2262.82   BIC=2280.39

Training set error measures:
                     ME      RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 0.09402082 18.33773 13.63872 -7.59249 23.83615 0.5281089 0.006760296
```

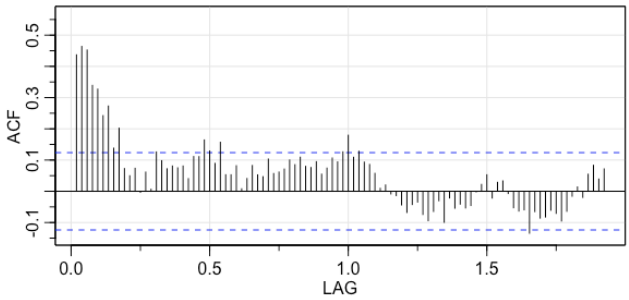


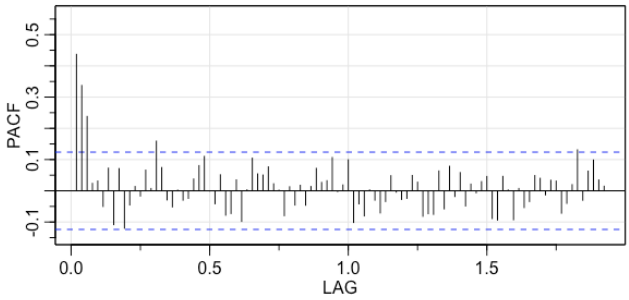

```
# select best model by summary
fit1 %>% summary()
fit2 %>% summary()
fit3 %>% summary()
```




```
#-------------------------------------------------------------------------

# R0 EDA and Model Selection
```

```
#------------------------------------------------------------------------

# create time series object
train %>%
   tk_ts(select = R0, start = 2014, frequency = 52,
silent = T) %>%
   tsclean() -> x

#outlier

train %>%
  tk_ts(select = R0, start = 2014, frequency = 52,
silent = T) -> x
```

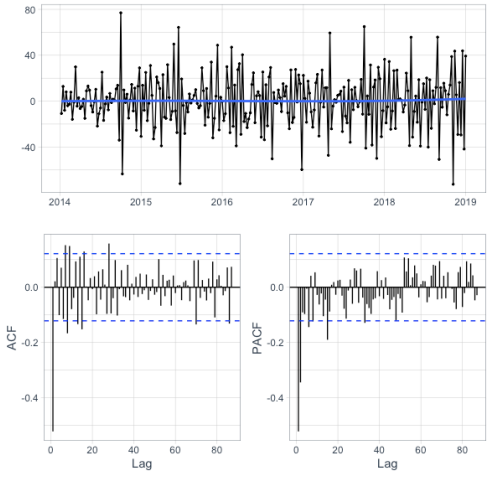





```
# ACF and PACF
x %>% acf2(max.lag = 100, main = '')

# ADF test
x %>% adf.test(alternative = 'stationary')
```

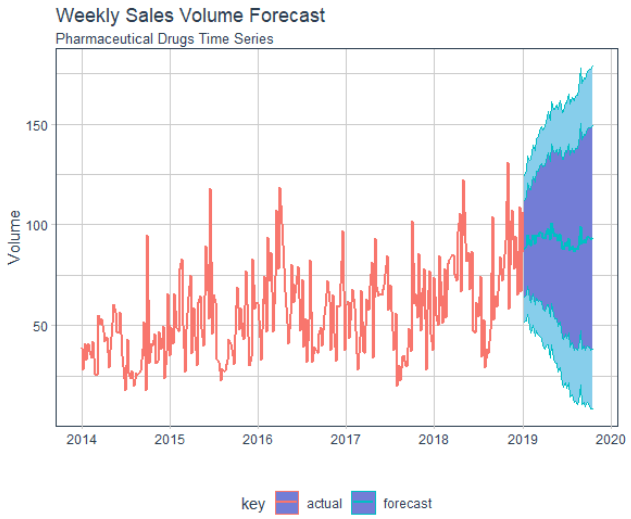
```
        Augmented Dickey-Fuller Test

data:  .
Dickey-Fuller = -4.149, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

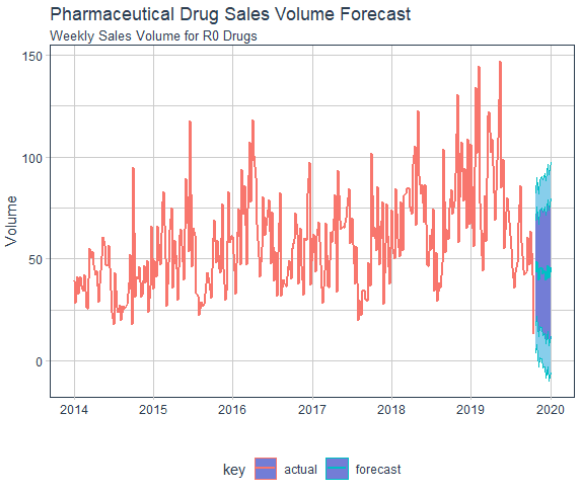
```
# estimating difference order
ndiffs(x)

nsdiffs(x)

# plot time series, ACF, and PACF
diff(x) %>% ggtsdisplay(plot.type = 'partial', smooth = T,
theme = theme_tq())

# model 1
x %>% Arima(order = c(0, 1, 2), seasonal = c(0, 0, 1)) ->
fit1
fit1 %>% residuals() %>%
   ggtsdisplay(plot.type = 'histogram', smooth = T, theme =
theme_tq())
fit1 %>% residuals() %>% adf.test(alternative =
'stationary')
```




```
# model 2
x %>% Arima(order = c(0, 1, 3), seasonal = c(0, 0, 1)) -> fit2
```

```
fit2 %>% residuals() %>%
   ggtsdisplay(plot.type = 'histogram', smooth = T,
theme = theme_tq())
fit2 %>% residuals() %>% adf.test(alternative =
'stationary')


# model 3
x %>% Arima(order = c(3, 1, 0), seasonal = c(1, 0, 0))
-> fit3
fit3 %>% residuals() %>%
   ggtsdisplay(plot.type = 'histogram', smooth = T,
theme = theme_tq())
fit3 %>% residuals() %>% adf.test(alternative =
'stationary')

# select best model by summary
fit1 %>% summary()
fit2 %>% summary()
fit3 %>% summary()
```






```
#-------------------------------------------------------------------------------
```