Project 1 Group 5

Aaron Zalki, Mia Siracusa, Lidiia Tronina, John Suh, Henry Vasquez

6/21/2020

In the first part of this project, we want to perform fundamental analysis of the data.

### Import and Clean Data

First step is to import the data from excel. When we imported the excel file, R was not reading the dates correctly, so we converted the first column in the date format.

## SeriesInd group Var01   
## Min. :2011-05-06 Length:10572 Min. : 9.03   
## 1st Qu.:2013-01-29 Class :character 1st Qu.: 23.10   
## Median :2014-11-03 Mode :character Median : 38.44   
## Mean :2014-11-01 Mean : 46.98   
## 3rd Qu.:2016-08-05 3rd Qu.: 66.78   
## Max. :2018-05-01 Max. :195.18   
## NA's :854   
## Var02 Var03 Var05 Var07   
## Min. : 1339900 Min. : 8.82 Min. : 8.99 Min. : 8.92   
## 1st Qu.: 12520675 1st Qu.: 22.59 1st Qu.: 22.91 1st Qu.: 22.88   
## Median : 21086550 Median : 37.66 Median : 38.05 Median : 38.05   
## Mean : 37035741 Mean : 46.12 Mean : 46.55 Mean : 46.56   
## 3rd Qu.: 42486700 3rd Qu.: 65.88 3rd Qu.: 66.38 3rd Qu.: 66.31   
## Max. :480879500 Max. :189.36 Max. :195.00 Max. :189.72   
## NA's :842 NA's :866 NA's :866 NA's :866

Next step is to examine the data before converting it to a time series to see if there is any missing data or other problems with the data. By doing this we discovered a few problems that needed to be dealt with:

1. All variables except date are NA after 10/13/17.
2. There are several NA that are in the middle of the data set.
3. There are outliers in data that are far above the normal.
4. The date field has only workdays (Monday through Friday).

We removed all blank observations after 10/13/17. Other NA’s data were imputed using the median since the number of missing values was so small. We used median for each variable and group separately.

## SeriesInd group Var01   
## Min. :2011-05-06 Length:9732 Min. : 9.03   
## 1st Qu.:2012-12-10 Class :character 1st Qu.: 23.16   
## Median :2014-07-25 Mode :character Median : 38.40   
## Mean :2014-07-23 Mean : 46.98   
## 3rd Qu.:2016-03-01 3rd Qu.: 66.80   
## Max. :2017-10-13 Max. :195.18   
## Var02 Var03 Var05 Var07   
## Min. : 1339900 Min. : 8.82 Min. : 8.99 Min. : 8.92   
## 1st Qu.: 12521025 1st Qu.: 22.63 1st Qu.: 22.93 1st Qu.: 22.92   
## Median : 21086550 Median : 37.62 Median : 38.01 Median : 37.98   
## Mean : 37031871 Mean : 46.12 Mean : 46.55 Mean : 46.56   
## 3rd Qu.: 42464900 3rd Qu.: 65.97 3rd Qu.: 66.43 3rd Qu.: 66.39   
## Max. :480879500 Max. :189.36 Max. :195.00 Max. :189.72

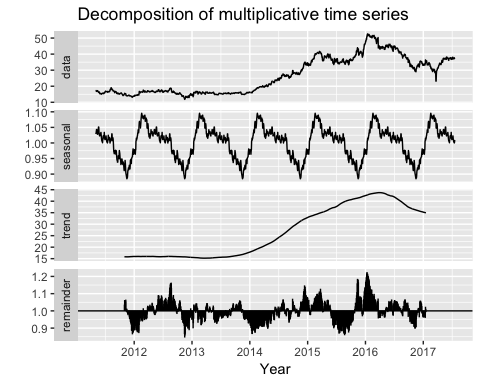
Time series objects were created for each group and variable separately. We used 261 days as our frequency, which is the approximate number of weekdays in a year.

### Forecast

For each group and variable, we want to run at least 2 models, and see which has the better performance. Before running any models we will check the ACF and PACF plots, seasonal plot, time series decomposition plot to see what it can recommend for what type of model they suggest might be most appropriate.

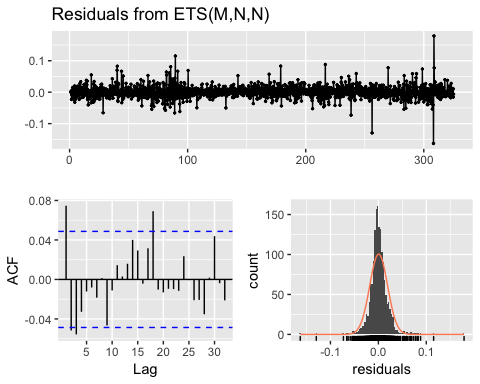
### S04 – Forecast Var01

For each of the variables in group 4-6, we want to try running ETS and ARIMA models. Before running any models for these 3 groups we checked the time plot, seasonal plot and decomposition to see what they recommend.

This data has a strong upward trend as you can see on the decomposition of multiplicative time series. 

#### ETS (S04 – Var01)

## ETS(M,N,N)   
##   
## Call:  
## ets(y = s04\_var01\_o)   
##   
## Smoothing parameters:  
## alpha = 0.9999   
##   
## Initial states:  
## l = 17.2061   
##   
## sigma: 0.0187  
##   
## AIC AICc BIC   
## 9409.319 9409.334 9425.494   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01214913 0.5278631 0.3286017 0.02970937 1.243549 0.398056  
## ACF1  
## Training set 0.04504415  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01214913 0.5278631 0.3286017 0.02970937 1.243549 0.398056  
## ACF1  
## Training set 0.04504415

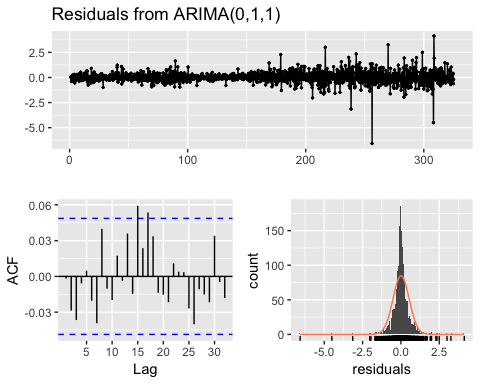


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,N,N)  
## Q\* = 24.847, df = 8, p-value = 0.00165  
##   
## Model df: 2. Total lags used: 10

The residuals plot looks not too bad, but our Ljung-Box test has an extremely small p-value indicating that there is some autocorrelation in our data.

#### ARIMA (S04 – Var01)

## Series: s04\_var01\_o   
## ARIMA(0,1,1)   
##   
## Coefficients:  
## ma1  
## 0.0482  
## s.e. 0.0254  
##   
## sigma^2 estimated as 0.2784: log likelihood=-1263.13  
## AIC=2530.25 AICc=2530.26 BIC=2541.03  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.0115981 0.5272811 0.3278779 0.02852278 1.240684 0.3971791  
## ACF1  
## Training set -0.001734489  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.0115981 0.5272811 0.3278779 0.02852278 1.240684 0.3971791  
## ACF1  
## Training set -0.001734489

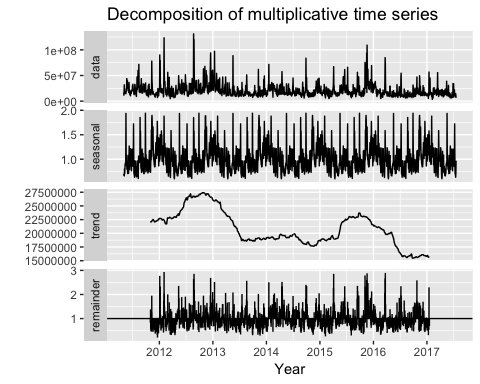


##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(0,1,1)  
## Q\* = 10.234, df = 9, p-value = 0.3319  
##   
## Model df: 1. Total lags used: 10

The ARIMA model gave us better AIC and Ljung-Box results.

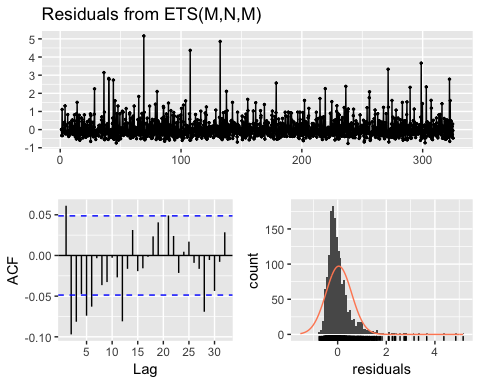
### S04 – Forecast Var02

There is an outlier in the second variable for Group 4 data that is far above the normal. The outlier was replaces with the median.



#### ETS (S04 – Var02)

## ETS(M,N,M)   
##   
## Call:  
## ets(y = s04\_var02\_o)   
##   
## Smoothing parameters:  
## alpha = 0.4519   
## gamma = 1e-04   
##   
## Initial states:  
## l = 17590156.2508   
## s = 1.0137 0.9936 1.0714 0.9871 0.9341  
##   
## sigma: 0.5297  
##   
## AIC AICc BIC   
## 64286.96 64287.05 64330.10   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set -39597.28 11266756 7024668 -13.05099 33.82442 0.7007043  
## ACF1  
## Training set 0.04595119  
## ME RMSE MAE MPE MAPE MASE  
## Training set -39597.28 11266756 7024668 -13.05099 33.82442 0.7007043  
## ACF1  
## Training set 0.04595119

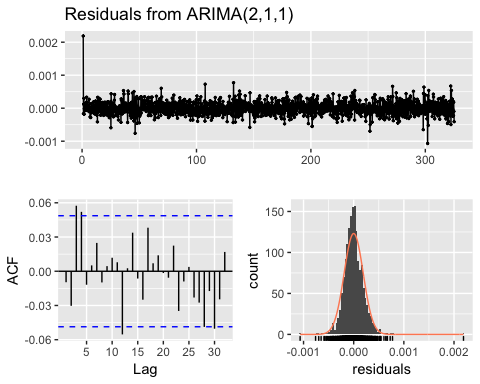


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,N,M)  
## Q\* = 55.143, df = 3, p-value = 6.4e-12  
##   
## Model df: 7. Total lags used: 10

#### ARIMA (S04 – Var02)

No seasonal differencing was recommended by auto.arima() but a box-cox transformation with λ = -0.4565625 was.

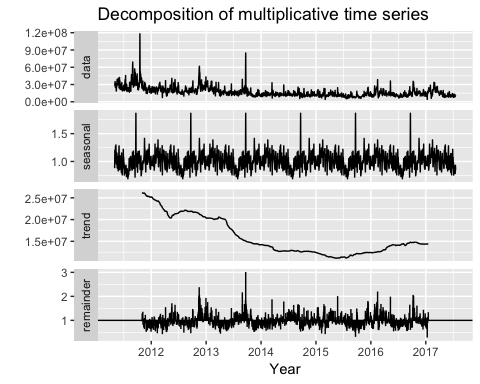
## Series: s04\_var02\_o   
## ARIMA(2,1,1)   
## Box Cox transformation: lambda= -0.4565625   
##   
## Coefficients:  
## ar1 ar2 ma1  
## 0.4430 0.0447 -0.9505  
## s.e. 0.0288 0.0279 0.0142  
##   
## sigma^2 estimated as 3.626e-08: log likelihood=11655.69  
## AIC=-23303.37 AICc=-23303.35 BIC=-23281.81  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 2143310 11039260 6200040 -4.052969 27.92882 0.6184484  
## ACF1  
## Training set 0.1276906  
## ME RMSE MAE MPE MAPE MASE  
## Training set 2143310 11039260 6200040 -4.052969 27.92882 0.6184484  
## ACF1  
## Training set 0.1276906



##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(2,1,1)  
## Q\* = 13.116, df = 7, p-value = 0.06933  
##   
## Model df: 3. Total lags used: 10

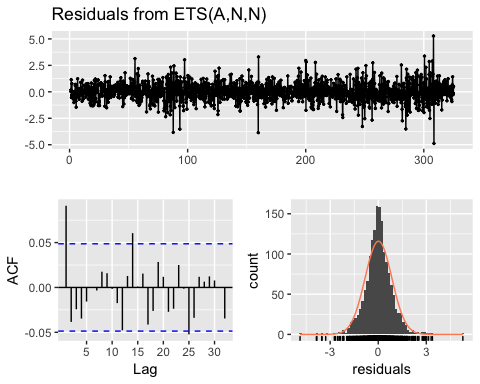
ARIMA is the preferred model for this variable forecasting.

### S05 – Forecast Var02



#### ETS (S05 – Var02)

## ETS(A,N,N)   
##   
## Call:  
## ets(y = s05\_var02\_o)   
##   
## Smoothing parameters:  
## alpha = 0.9999   
##   
## Initial states:  
## l = 69.2306   
##   
## sigma: 0.8349  
##   
## AIC AICc BIC   
## 11407.62 11407.64 11423.79   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01368779 0.834412 0.6030765 0.01199384 0.7286703 0.4064651  
## ACF1  
## Training set 0.09108834  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01368779 0.834412 0.6030765 0.01199384 0.7286703 0.4064651  
## ACF1  
## Training set 0.09108834

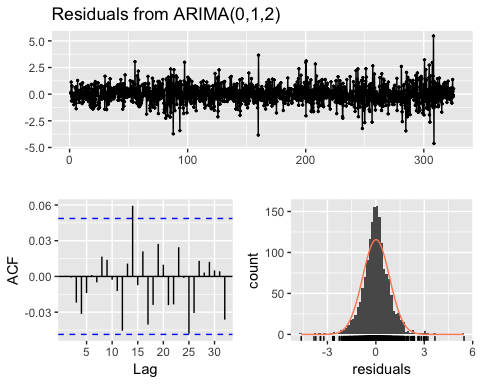


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(A,N,N)  
## Q\* = 20.144, df = 8, p-value = 0.009805  
##   
## Model df: 2. Total lags used: 10

#### ARIMA (S05 – Var02)

## Series: s05\_var02\_o   
## ARIMA(0,1,2)   
##   
## Coefficients:  
## ma1 ma2  
## 0.0942 -0.0371  
## s.e. 0.0249 0.0257  
##   
## sigma^2 estimated as 0.6903: log likelihood=-1998.72  
## AIC=4003.44 AICc=4003.46 BIC=4019.61  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01299072 0.8300814 0.6013961 0.01147491 0.7260563 0.4053325  
## ACF1  
## Training set 0.0005545699  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01299072 0.8300814 0.6013961 0.01147491 0.7260563 0.4053325  
## ACF1  
## Training set 0.0005545699

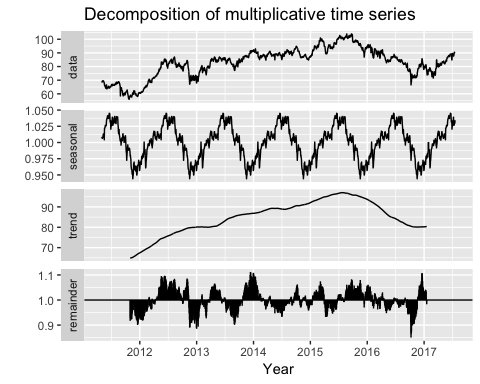
checkresiduals(s05\_var02\_arima)



##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(0,1,2)  
## Q\* = 3.536, df = 8, p-value = 0.8964  
##   
## Model df: 2. Total lags used: 10

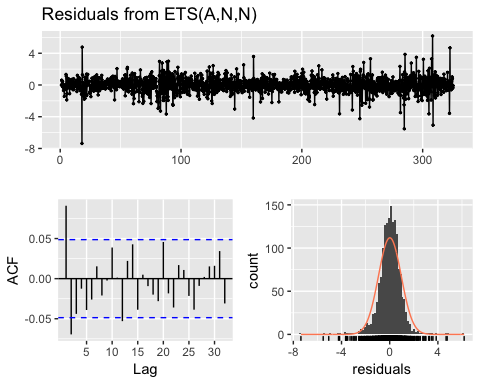
The ARIMA model resulted in the best fit with the best RMSE and a Ljung-Box p-value. The plot of the forecast also looks like a more reasonable estimate of what we can expect based on the historical data.

### S05 – Forecast Var03



#### ETS (S05 – Var03)

## ETS(A,N,N)   
##   
## Call:  
## ets(y = s05\_var03\_o)   
##   
## Smoothing parameters:  
## alpha = 0.9999   
##   
## Initial states:  
## l = 68.1973   
##   
## sigma: 0.9447  
##   
## AIC AICc BIC   
## 11808.35 11808.36 11824.52   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE  
## Training set 0.01325822 0.9441228 0.6650643 0.009688165 0.8230491  
## MASE ACF1  
## Training set 0.4171032 0.09092057  
## ME RMSE MAE MPE MAPE  
## Training set 0.01325822 0.9441228 0.6650643 0.009688165 0.8230491  
## MASE ACF1  
## Training set 0.4171032 0.09092057

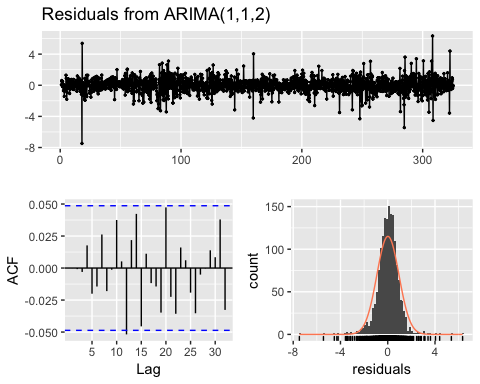


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(A,N,N)  
## Q\* = 31.866, df = 8, p-value = 9.842e-05  
##   
## Model df: 2. Total lags used: 10

#### ARIMA (S05 – Var03)

## Series: s05\_var03\_o   
## ARIMA(1,1,2)   
##   
## Coefficients:  
## ar1 ma1 ma2  
## 0.6437 -0.5500 -0.1288  
## s.e. 0.1652 0.1646 0.0248  
##   
## sigma^2 estimated as 0.8787: log likelihood=-2193.84  
## AIC=4395.68 AICc=4395.71 BIC=4417.25  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01471411 0.9362551 0.6613117 0.01099967 0.8178549 0.4147497  
## ACF1  
## Training set -6.261877e-05  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01471411 0.9362551 0.6613117 0.01099967 0.8178549 0.4147497  
## ACF1  
## Training set -6.261877e-05

checkresiduals(s05\_var03\_arima)

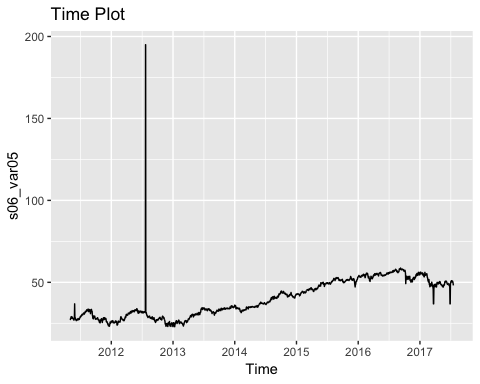


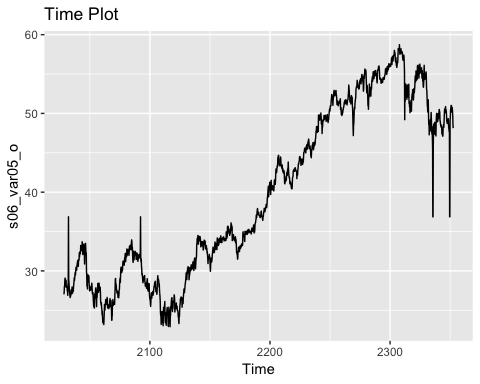
##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(1,1,2)  
## Q\* = 5.4726, df = 7, p-value = 0.6025  
##   
## Model df: 3. Total lags used: 10

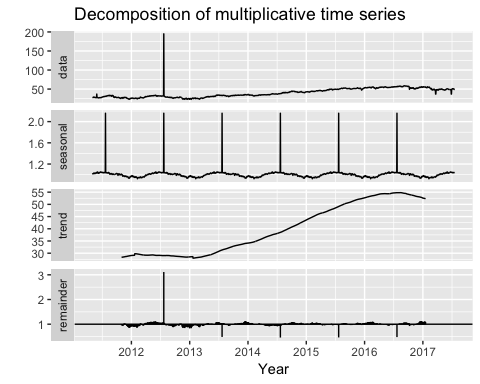
The ARIMA model gave us slightly better results than ETS model, based on AIC and sigma^2(standart deviation).

### S06 – Forecast Var05

Here again, we can clearly see an outlier that is most likely a data error so we imputed that point with the mean of the other data for the same variable and group. We did the same with another NA data point for next variable (Var07).

 Let’s plot the data again after these transformations are performed to see what impact they have.

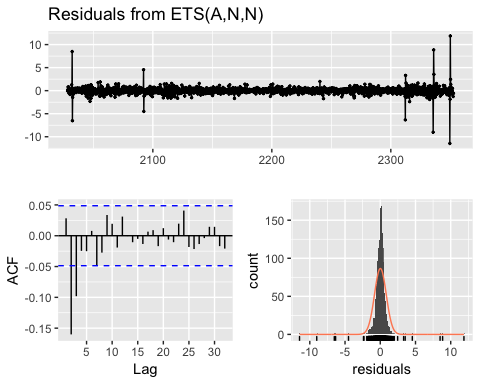




This group’s outliers were replaced with the median since it was so far above the norm, so it seemed likely to be an error.

#### ETS (S06 – Var05)

## ETS(A,N,N)   
##   
## Call:  
## ets(y = s06\_var05\_o)   
##   
## Smoothing parameters:  
## alpha = 0.8367   
##   
## Initial states:  
## l = 27.0856   
##   
## sigma: 0.8347  
##   
## AIC AICc BIC   
## 11406.85 11406.86 11423.02   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01557398 0.8342136 0.4583358 0.01358191 1.247755 0.5023548  
## ACF1  
## Training set 0.02825661  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01557398 0.8342136 0.4583358 0.01358191 1.247755 0.5023548  
## ACF1  
## Training set 0.02825661

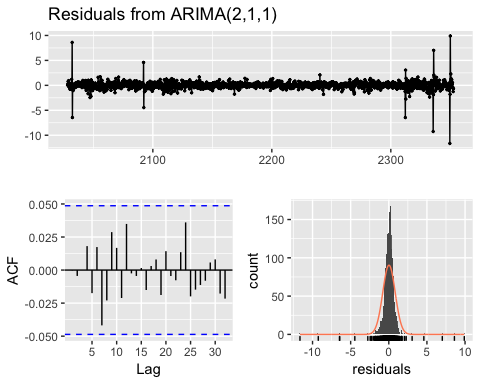


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(A,N,N)  
## Q\* = 68.223, df = 8, p-value = 1.109e-11  
##   
## Model df: 2. Total lags used: 10

#### ARIMA (S06 – Var05)

## Series: s06\_var05\_o   
## ARIMA(2,1,1)   
##   
## Coefficients:  
## ar1 ar2 ma1  
## 0.5527 -0.0864 -0.7094  
## s.e. 0.0629 0.0307 0.0594  
##   
## sigma^2 estimated as 0.6682: log likelihood=-1971.89  
## AIC=3951.79 AICc=3951.81 BIC=3973.35  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.02463791 0.8164195 0.4643094 0.0316663 1.266586 0.508902  
## ACF1  
## Training set -0.0002354859  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.02463791 0.8164195 0.4643094 0.0316663 1.266586 0.508902  
## ACF1  
## Training set -0.0002354859

checkresiduals(s06\_var05\_arima)

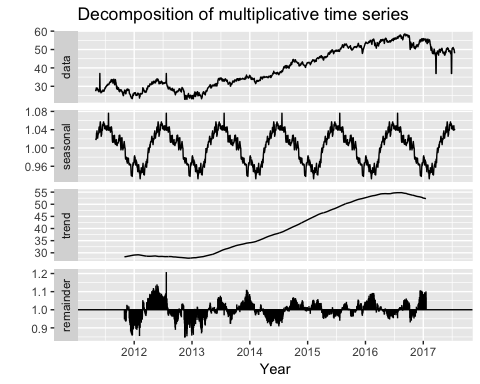


##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(2,1,1)  
## Q\* = 7.0771, df = 7, p-value = 0.4209  
##   
## Model df: 3. Total lags used: 10

The auto.arima function gave us the best results so that model will be used for predictions.

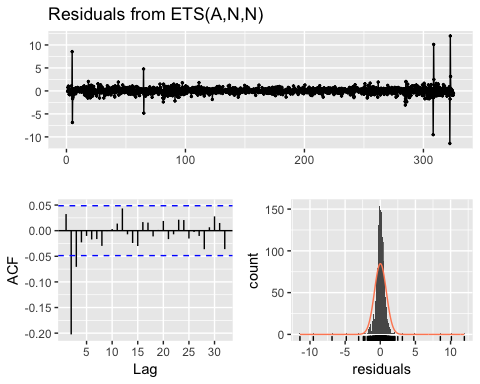
### S06 – Forecast Var07

This group’s outlier was also replaced with the median since it was so far above the norm, so it seemed likely to be an error.



#### ETS (S06 – Var07)

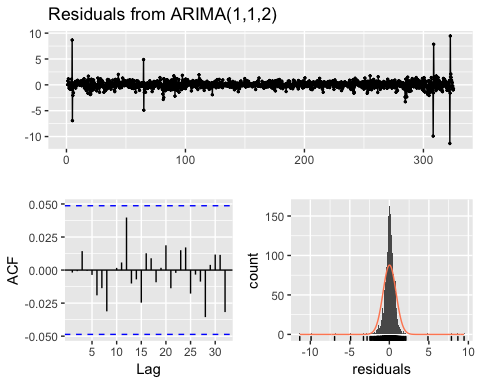
## ETS(A,N,N)   
##   
## Call:  
## ets(y = s06\_var07\_o)   
##   
## Smoothing parameters:  
## alpha = 0.8458   
##   
## Initial states:  
## l = 27.4391   
##   
## sigma: 0.8473  
##   
## AIC AICc BIC   
## 11455.32 11455.34 11471.50   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01502264 0.8467723 0.4624363 0.01090339 1.258436 0.4960822  
## ACF1  
## Training set 0.0324403  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.01502264 0.8467723 0.4624363 0.01090339 1.258436 0.4960822  
## ACF1  
## Training set 0.0324403



##   
## Ljung-Box test  
##   
## data: Residuals from ETS(A,N,N)  
## Q\* = 79.85, df = 8, p-value = 5.24e-14  
##   
## Model df: 2. Total lags used: 10

#### ARIMA (S06 – Var07)

## Series: s06\_var07\_o   
## ARIMA(1,1,2)   
##   
## Coefficients:  
## ar1 ma1 ma2  
## 0.2743 -0.4101 -0.1873  
## s.e. 0.1035 0.1027 0.0360  
##   
## sigma^2 estimated as 0.6801: log likelihood=-1986.23  
## AIC=3980.46 AICc=3980.48 BIC=4002.02  
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.02361582 0.8236674 0.4708819 0.02765548 1.285544 0.5051422  
## ACF1  
## Training set -0.001869352  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.02361582 0.8236674 0.4708819 0.02765548 1.285544 0.5051422  
## ACF1  
## Training set -0.001869352



##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(1,1,2)  
## Q\* = 2.8608, df = 7, p-value = 0.8976  
##   
## Model df: 3. Total lags used: 10

The ARIMA model has the lower AIC compare to the ETS model. The residuals plot and Ljung-Box looks good. The plot of the forecast also looks like a more reasonable estimate of what we can expect based on the historical data.