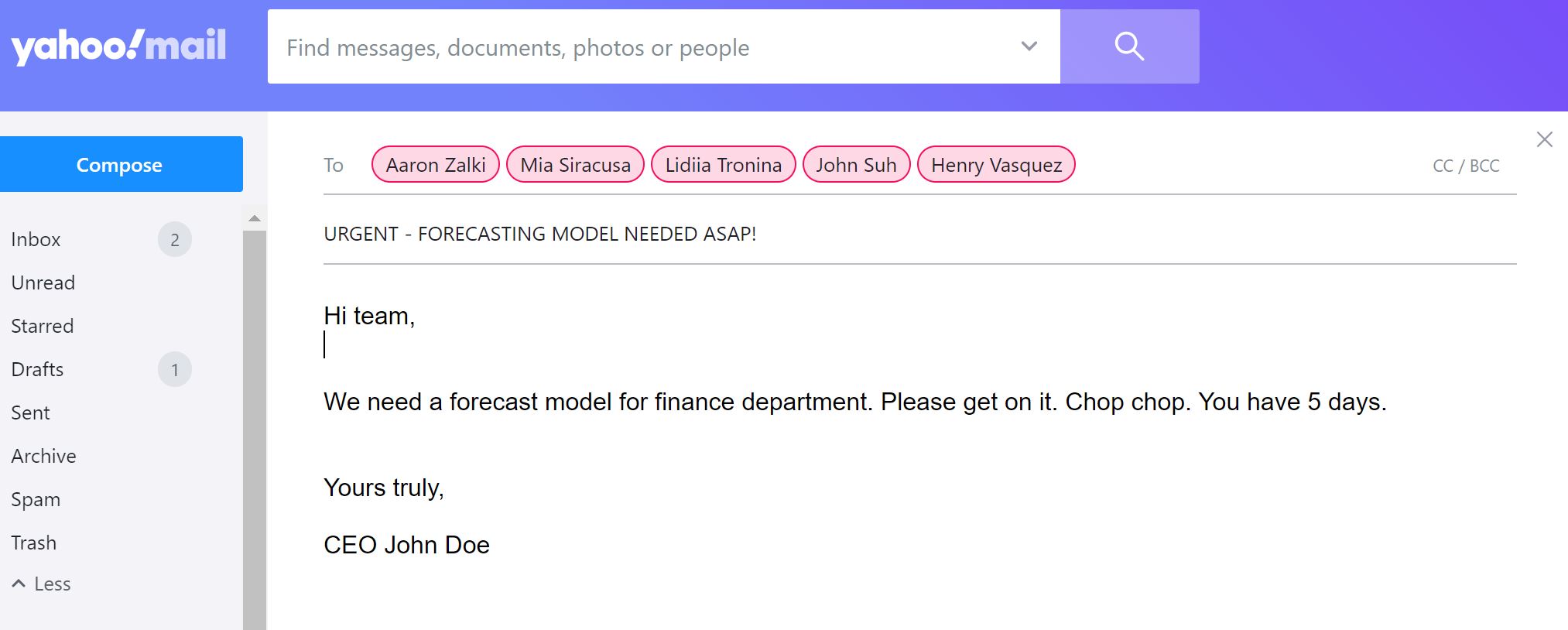
Project 1 Group 5

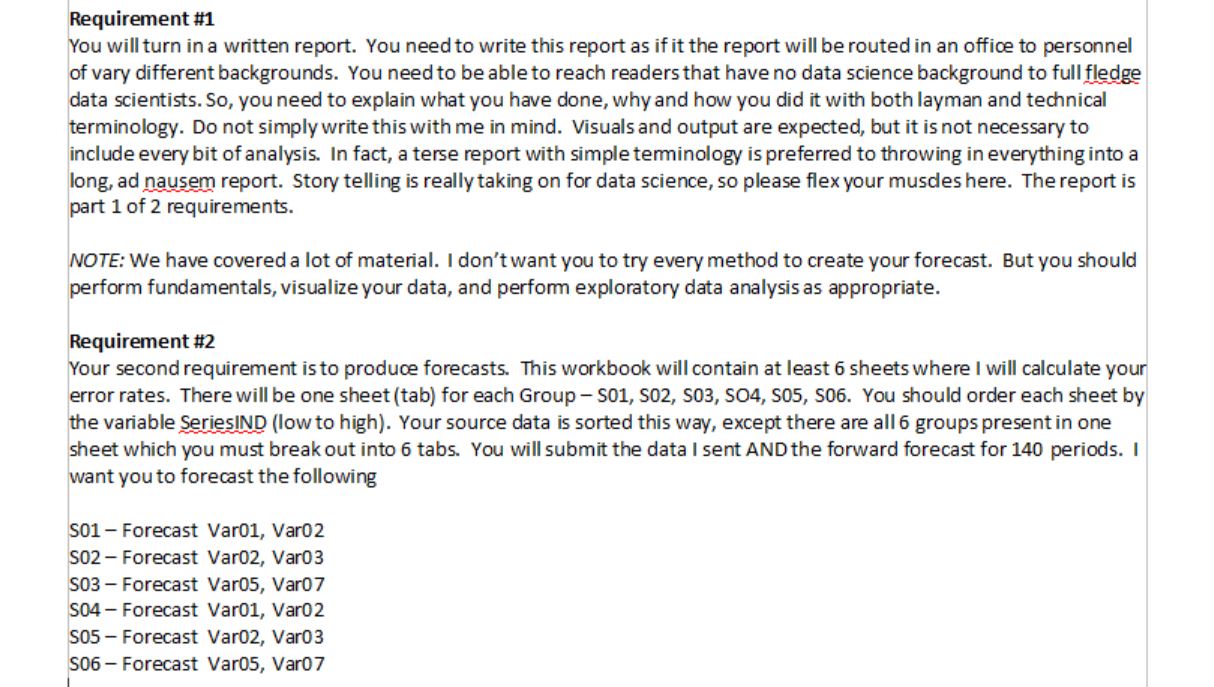
Aaron Zalki, Mia Siracusa, Lidiia Tronina, John Suh, Henry Vasquez

6/21/2020

On or about 6/20/2020, our team was tasked by the CEO of ACME Corp to research and study forecasting models that could be used by the company in order to produce reliable forecasts that would help with the annual planning and budgeting by the finance department. This directive was sent via email accompanied by the attachment of the document below. Along with the credentials to access a secret SQL database called DATA624 that had the data to model the forecast on. We have come up with a forecasting model for the company as instructed by our CEO which shall be shared in our conclusion. Our research and methodologies that brought us to our conclusion and what we ultimately recommended will also be shared in the report.



EmailFromCEO



EmailAttachment

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.

### Decompostion

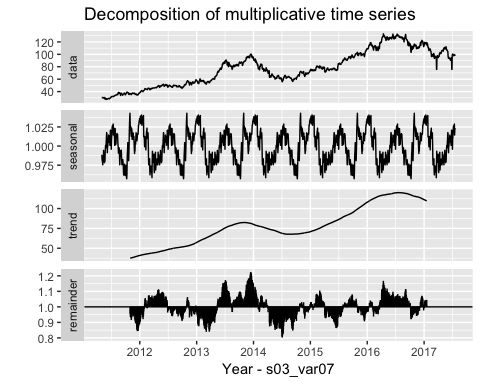
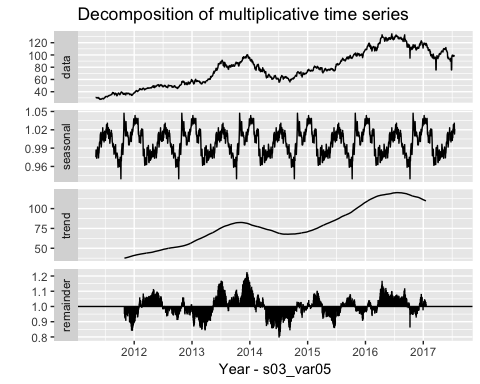
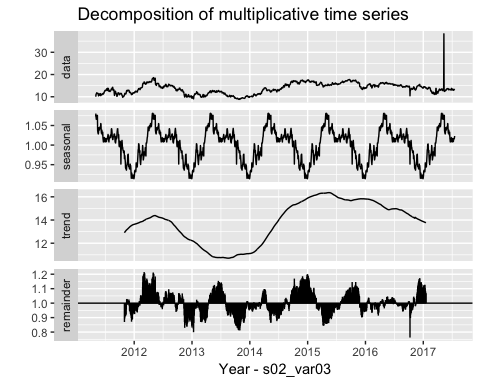
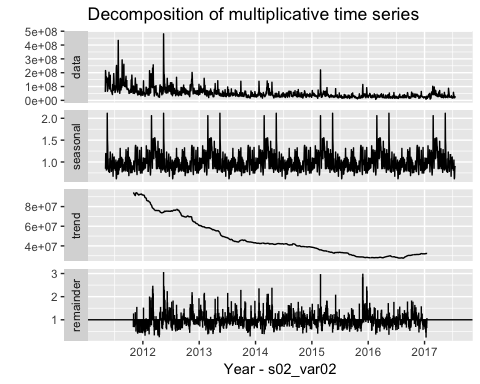
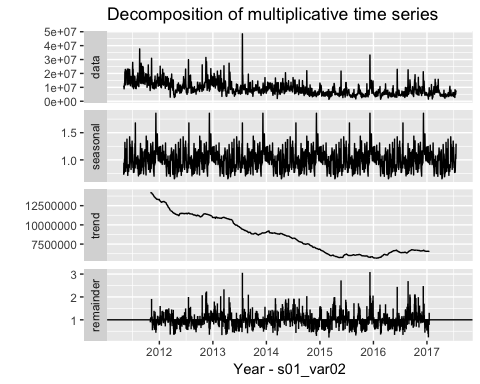
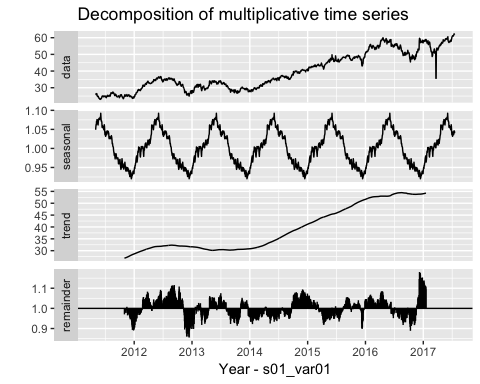
We will now run a decomposition function on all our series for each our groups and variables. Decomposition is a mathematical procedure that will produce different time series from the time series it is decomposing that will usually be split into several components such as trend and seasonality. We need this in order to see if the time series we want to model has those qualities which would factor in what models would be more appropriate to use.

REFERENCE LINK TO RESEARCH MATERIAL:

<https://anomaly.io/seasonal-trend-decomposition-in-r/index.html>

#### S01- S03

After running the decomposition on all the series, we saw that s01\_var01, s02\_var02, s03\_var05 and s03\_var07 had some trend and and seasonality while s01\_var02 and s02\_var02 did not exhibit as much seasonality as the others.



#### Forecast S01- S03

We will now pick two forecast methods to use with our data. One will be the NAIVE method which we will use as our benchmark method and the other will be ETS which will be an automatic method to compare to the benchmark. An automatic model such as ETS was chosen because it will automatically select the best parameters that account for seasonality, trend and other characteristics of the time series. Rob Hyndman showed in a slide at one of his seminars on forecasting that an automatic forecasting procedure can be beneficial because most users are not experts at fitting time series models. That sometimes even experts cannot beat the best automatic algorithms. And lastly that many businesses and industries need thousands of forecasts every week or month. So therefore an automatic method is valued for its ease of use and also efficiency.

Rob Hyndman seminar link:

<https://www.youtube.com/watch?v=1Lh1HlBUf8k&feature=emb_logo>

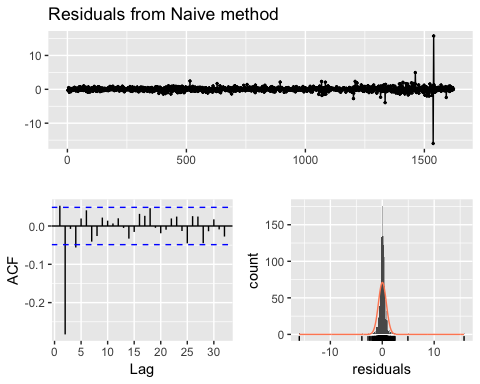
After running the forecast methods on each time series, we will record the residual sd and compare the two. A lower residual sd is better, so the method that produces a lower sd will be selected for that particular time series.

And then lastly, a checkresiduals function will be run on all the series to further evaluate the models.

Please note that for ETS a model passes the Ljung-Box test when the p-value is greater than 0.05 which means the residuals are independent and that is needed for the model to be correct. All the ETS models were greater than .05

#### NAIVE (S01-Var01) residual sd .7584

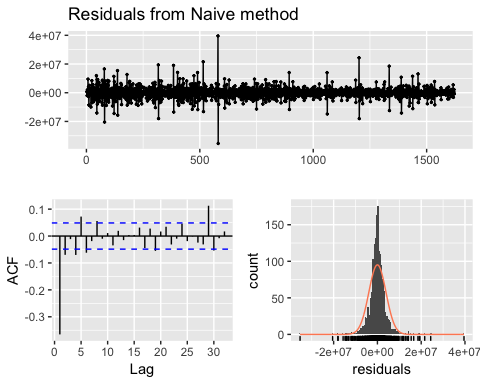
##   
## Forecast method: Naive method  
##   
## Model Information:  
## Call: naive(y = s01\_var01)   
##   
## Residual sd: 0.7584   
##   
## Error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.02202344 0.7584407 0.3694694 0.03558492 0.9615419 1  
## ACF1  
## Training set 0.0528916  
##   
## Forecasts:  
## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## 1623 62.31 61.33802 63.28198 60.82348 63.79652  
## 1624 62.31 60.93541 63.68459 60.20775 64.41225  
## 1625 62.31 60.62648 63.99352 59.73528 64.88472  
## 1626 62.31 60.36604 64.25396 59.33697 65.28303  
## 1627 62.31 60.13659 64.48342 58.98605 65.63395  
## 1628 62.31 59.92914 64.69086 58.66879 65.95121  
## 1629 62.31 59.73838 64.88162 58.37705 66.24295  
## 1630 62.31 59.56082 65.05918 58.10550 66.51450  
## 1631 62.31 59.39406 65.22594 57.85045 66.76955  
## 1632 62.31 59.23633 65.38367 57.60922 67.01078



##   
## Ljung-Box test  
##   
## data: Residuals from Naive method  
## Q\* = 148.18, df = 10, p-value < 2.2e-16  
##   
## Model df: 0. Total lags used: 10

#### NAIVE (S01-Var02) residual sd 3919214.7158

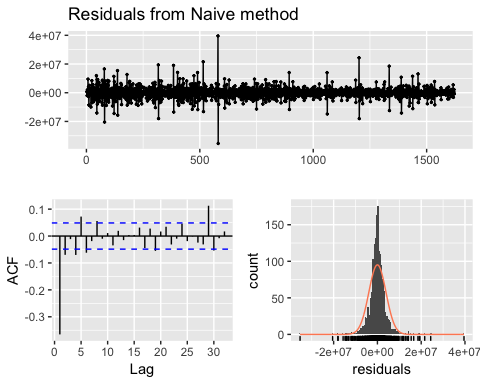
##   
## Forecast method: Naive method  
##   
## Model Information:  
## Call: naive(y = s01\_var02)   
##   
## Residual sd: 3919214.7158   
##   
## Error measures:  
## ME RMSE MAE MPE MAPE MASE ACF1  
## Training set -2265.268 3918006 2525392 -6.828774 28.97905 1 -0.3654519  
##   
## Forecasts:  
## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## 1623 6697300 1676173 11718427 -981851.2 14376451  
## 1624 6697300 -403646 13798246 -4162659.8 17557260  
## 1625 6697300 -1999547 15394147 -6603380.1 19997980  
## 1626 6697300 -3344954 16739554 -8661002.5 22055602  
## 1627 6697300 -4530282 17924882 -10473804.2 23868404  
## 1628 6697300 -5601899 18996499 -12112702.2 25507302  
## 1629 6697300 -6587354 19981954 -13619824.4 27014424  
## 1630 6697300 -7504592 20899192 -15022619.6 28417220  
## 1631 6697300 -8366081 21760681 -16340153.7 29734754  
## 1632 6697300 -9180898 22575498 -17586308.4 30980908



##   
## Ljung-Box test  
##   
## data: Residuals from Naive method  
## Q\* = 253.87, df = 10, p-value < 2.2e-16  
##   
## Model df: 0. Total lags used: 10

#### NAIVE (S02-Var02) residual sd 28643387.1185

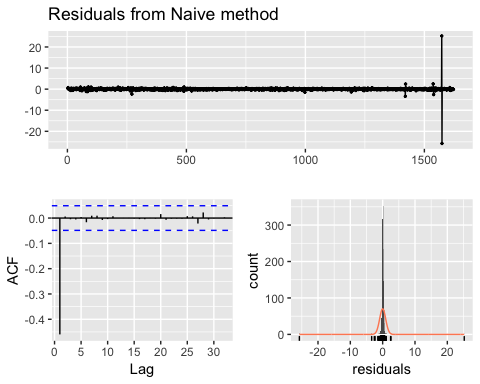
##   
## Forecast method: Naive method  
##   
## Model Information:  
## Call: naive(y = s02\_var02)   
##   
## Residual sd: 28643387.1185   
##   
## Error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set -24743.99 28634561 15651244 -6.737919 29.26663 1  
## ACF1  
## Training set -0.2413426  
##   
## Forecasts:  
## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## 1623 20745800 -15950867 57442467 -35376909 76868509  
## 1624 20745800 -31151124 72642724 -58623696 100115296  
## 1625 20745800 -42814692 84306292 -76461583 117953183  
## 1626 20745800 -52647534 94139134 -91499618 132991218  
## 1627 20745800 -61310442 102802042 -104748392 146239992  
## 1628 20745800 -69142309 110633909 -116726200 158217800  
## 1629 20745800 -76344455 117836055 -127740931 169232531  
## 1630 20745800 -83048048 124539648 -137993192 179484792  
## 1631 20745800 -89344201 130835801 -147622327 189113927  
## 1632 20745800 -95299250 136790850 -156729789 198221389



##   
## Ljung-Box test  
##   
## data: Residuals from Naive method  
## Q\* = 253.87, df = 10, p-value < 2.2e-16  
##   
## Model df: 0. Total lags used: 10

#### NAIVE (S02-Var03) residual sd 0.9396

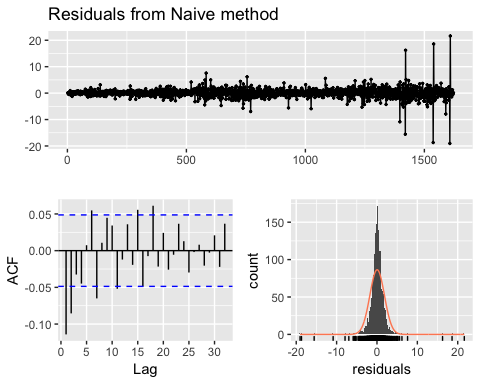
##   
## Forecast method: Naive method  
##   
## Model Information:  
## Call: naive(y = s02\_var03)   
##   
## Residual sd: 0.9396   
##   
## Error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.001795188 0.9392752 0.2136027 -0.09286091 1.562846 1  
## ACF1  
## Training set -0.459794  
##   
## Forecasts:  
## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## 1623 12.96 11.756270 14.16373 11.119054 14.80095  
## 1624 12.96 11.257669 14.66233 10.356510 15.56349  
## 1625 12.96 10.875079 15.04492 9.771389 16.14861  
## 1626 12.96 10.552541 15.36746 9.278109 16.64189  
## 1627 12.96 10.268379 15.65162 8.843520 17.07648  
## 1628 12.96 10.011477 15.90852 8.450623 17.46938  
## 1629 12.96 9.775231 16.14477 8.089316 17.83068  
## 1630 12.96 9.555338 16.36466 7.753019 18.16698  
## 1631 12.96 9.348811 16.57119 7.437163 18.48284  
## 1632 12.96 9.153473 16.76653 7.138419 18.78158



##   
## Ljung-Box test  
##   
## data: Residuals from Naive method  
## Q\* = 344.35, df = 10, p-value < 2.2e-16  
##   
## Model df: 0. Total lags used: 10

#### NAIVE (S03-Var05) residual sd 1.8016

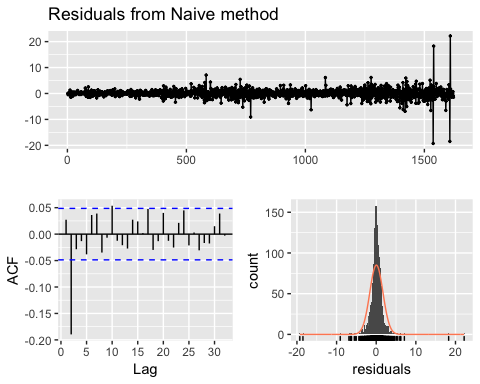
##   
## Forecast method: Naive method  
##   
## Model Information:  
## Call: naive(y = s03\_var05)   
##   
## Residual sd: 1.8016   
##   
## Error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.0420728 1.801582 1.058463 0.04872238 1.386291 1  
## ACF1  
## Training set -0.1140862  
##   
## Forecasts:  
## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## 1623 98.69 96.38118 100.9988 95.15897 102.2210  
## 1624 98.69 95.42484 101.9552 93.69636 103.6836  
## 1625 98.69 94.69101 102.6890 92.57407 104.8059  
## 1626 98.69 94.07236 103.3076 91.62793 105.7521  
## 1627 98.69 93.52732 103.8527 90.79437 106.5856  
## 1628 98.69 93.03457 104.3454 90.04077 107.3392  
## 1629 98.69 92.58144 104.7986 89.34776 108.0322  
## 1630 98.69 92.15967 105.2203 88.70272 108.6773  
## 1631 98.69 91.76354 105.6165 88.09690 109.2831  
## 1632 98.69 91.38887 105.9911 87.52389 109.8561



##   
## Ljung-Box test  
##   
## data: Residuals from Naive method  
## Q\* = 55.254, df = 10, p-value = 2.831e-08  
##   
## Model df: 0. Total lags used: 10

#### NAIVE (S03-Var07) residual sd 1.6587

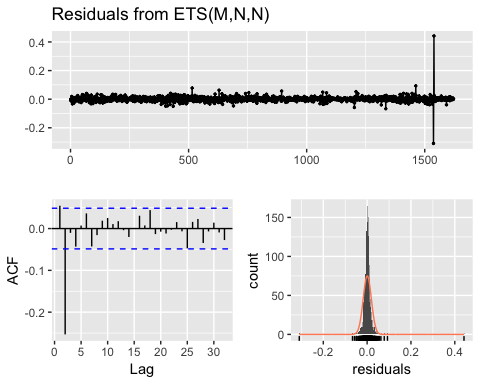
##   
## Forecast method: Naive method  
##   
## Model Information:  
## Call: naive(y = s03\_var07)   
##   
## Residual sd: 1.6587   
##   
## Error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.04118886 1.658734 0.9754798 0.05069813 1.279719 1  
## ACF1  
## Training set 0.02722712  
##   
## Forecasts:  
## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## 1623 97.34 95.21424 99.46575 94.08894 100.5911  
## 1624 97.34 94.33373 100.34627 92.74230 101.9377  
## 1625 97.34 93.65808 101.02191 91.70900 102.9710  
## 1626 97.34 93.08849 101.59150 90.83788 103.8421  
## 1627 97.34 92.58667 102.09332 90.07041 104.6096  
## 1628 97.34 92.13299 102.54701 89.37656 105.3034  
## 1629 97.34 91.71578 102.96421 88.73850 105.9415  
## 1630 97.34 91.32746 103.35253 88.14461 106.5354  
## 1631 97.34 90.96274 103.71726 87.58682 107.0932  
## 1632 97.34 90.61777 104.06222 87.05924 107.6207



##   
## Ljung-Box test  
##   
## data: Residuals from Naive method  
## Q\* = 74.99, df = 10, p-value = 4.779e-12  
##   
## Model df: 0. Total lags used: 10

#### ETS (S01-Var01) residual sd 0.0187

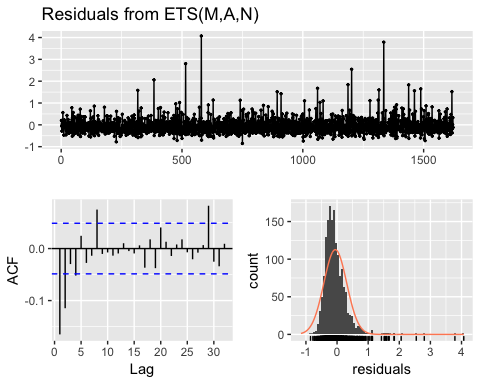
## ETS(M,N,N)   
##   
## Call:  
## ets(y = s01\_var01)   
##   
## Smoothing parameters:  
## alpha = 0.9999   
##   
## Initial states:  
## l = 26.6072   
##   
## sigma: 0.0187  
##   
## AIC AICc BIC   
## 10873.94 10873.96 10890.12   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.02201376 0.758211 0.3692509 0.03557276 0.9609744 0.9994087  
## ACF1  
## Training set 0.0529696



##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,N,N)  
## Q\* = 119.56, df = 8, p-value < 2.2e-16  
##   
## Model df: 2. Total lags used: 10

#### ETS (S01-Var02) residual sd 0.3724

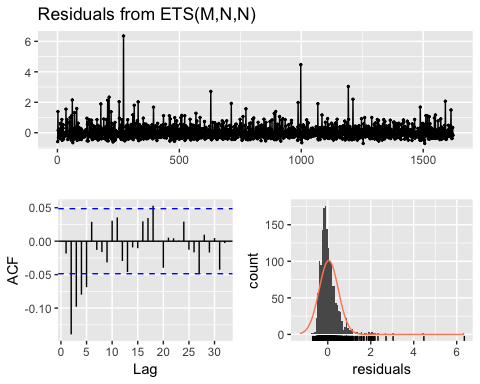
## ETS(M,A,N)   
##   
## Call:  
## ets(y = s01\_var02)   
##   
## Smoothing parameters:  
## alpha = 0.8669   
## beta = 8e-04   
##   
## Initial states:  
## l = 7092879.624   
## b = 1533385.7973   
##   
## sigma: 0.3724  
##   
## AIC AICc BIC   
## 60663.30 60663.34 60690.26   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set -934840.1 3885303 2614580 -19.06936 32.17793 1.035316  
## ACF1  
## Training set -0.262807



##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,A,N)  
## Q\* = 83.468, df = 6, p-value = 6.661e-16  
##   
## Model df: 4. Total lags used: 10

#### ETS (S02-Var02) residual sd 0.4495

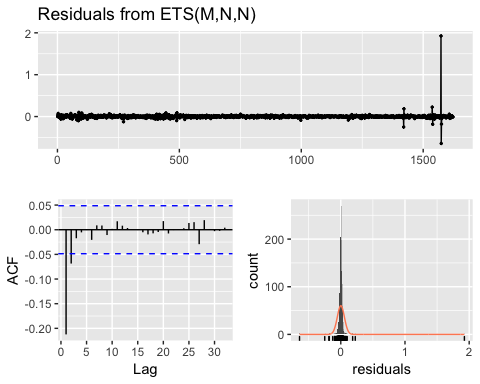
## ETS(M,N,N)   
##   
## Call:  
## ets(y = s02\_var02)   
##   
## Smoothing parameters:  
## alpha = 0.6537   
##   
## Initial states:  
## l = 145674004.3673   
##   
## sigma: 0.4495  
##   
## AIC AICc BIC   
## 66430.83 66430.84 66447.00   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set -116422.7 27122130 14979348 -8.359695 28.14981 0.9570708  
## ACF1  
## Training set -0.0001340659



##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,N,N)  
## Q\* = 71.102, df = 8, p-value = 2.964e-12  
##   
## Model df: 2. Total lags used: 10

#### ETS (S02-Var03) residual sd 0.0552

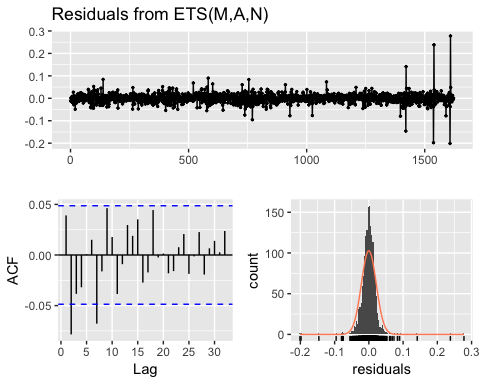
## ETS(M,N,N)   
##   
## Call:  
## ets(y = s02\_var03)   
##   
## Smoothing parameters:  
## alpha = 0.8713   
##   
## Initial states:  
## l = 10.0754   
##   
## sigma: 0.0552  
##   
## AIC AICc BIC   
## 11036.42 11036.43 11052.59   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.002050742 0.8891435 0.2175926 -0.09402523 1.592937 1.018679  
## ACF1  
## Training set -0.3832941



##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,N,N)  
## Q\* = 82.546, df = 8, p-value = 1.499e-14  
##   
## Model df: 2. Total lags used: 10

#### ETS (S03-Var05) residual sd 0.0218

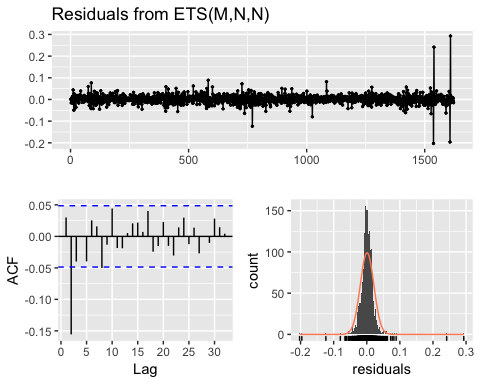
## ETS(M,A,N)   
##   
## Call:  
## ets(y = s03\_var05)   
##   
## Smoothing parameters:  
## alpha = 0.8506   
## beta = 1e-04   
##   
## Initial states:  
## l = 30.8088   
## b = 0.0659   
##   
## sigma: 0.0218  
##   
## AIC AICc BIC   
## 13425.73 13425.77 13452.69   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set -0.02729464 1.786195 1.055654 -0.05906346 1.38593 0.997346  
## ACF1  
## Training set 0.01957225



##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,A,N)  
## Q\* = 28.871, df = 6, p-value = 6.434e-05  
##   
## Model df: 4. Total lags used: 10

#### ETS (S03-Var07) residual sd 0.0205

## ETS(M,N,N)   
##   
## Call:  
## ets(y = s03\_var07)   
##   
## Smoothing parameters:  
## alpha = 0.9999   
##   
## Initial states:  
## l = 30.5536   
##   
## sigma: 0.0205  
##   
## AIC AICc BIC   
## 13232.50 13232.51 13248.67   
##   
## Training set error measures:  
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.04117966 1.658227 0.9748925 0.05071092 1.278974 0.999398  
## ACF1  
## Training set 0.02730971

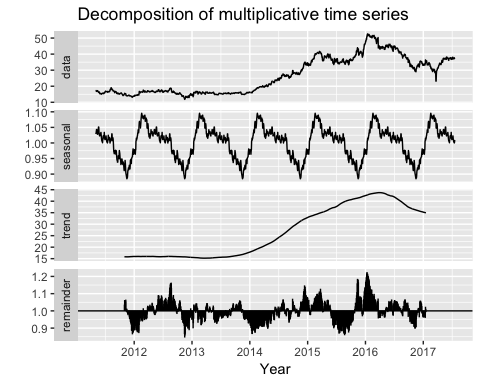


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,N,N)  
## Q\* = 55.084, df = 8, p-value = 4.252e-09  
##   
## Model df: 2. Total lags used: 10

The ETS model had the lower residual sd for all 3 groups. Therefore based on that criteria, ETS will be used to produce the forecasts for S01-S03.

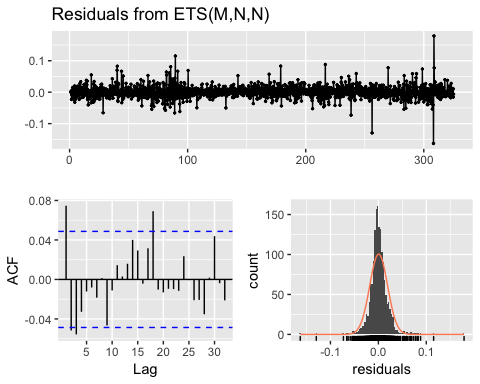
### S04 – Forecast Var01

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

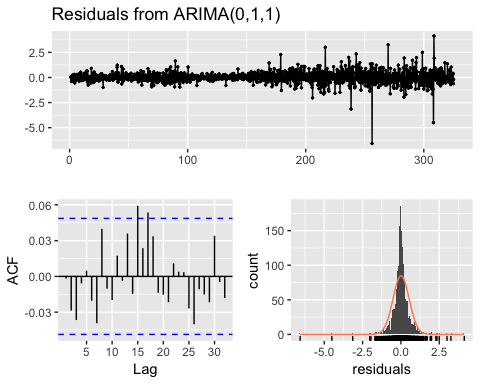


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,N,N)  
## Q\* = 24.847, df = 8, p-value = 0.00165  
##   
## Model df: 2. Total lags used: 10

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

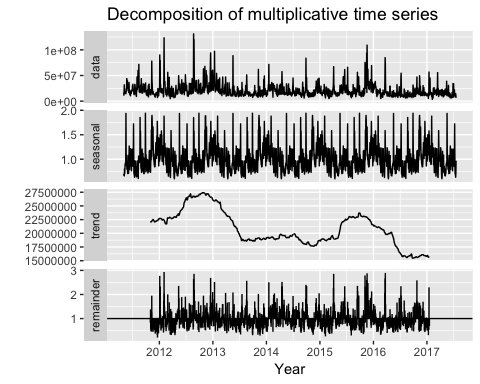


##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(0,1,1)  
## Q\* = 10.234, df = 9, p-value = 0.3319  
##   
## Model df: 1. Total lags used: 10

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

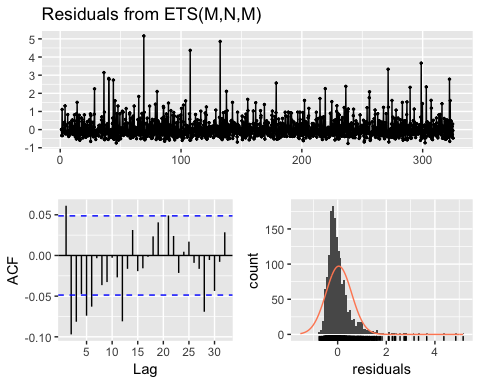
### S04 – Forecast Var02

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

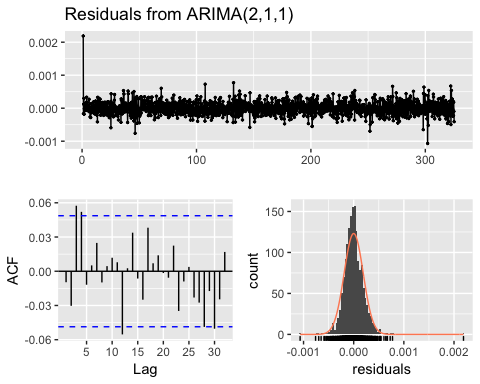


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(M,N,M)  
## Q\* = 55.143, df = 3, p-value = 6.4e-12  
##   
## Model df: 7. Total lags used: 10

#### ARIMA (S04 – Var02)

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

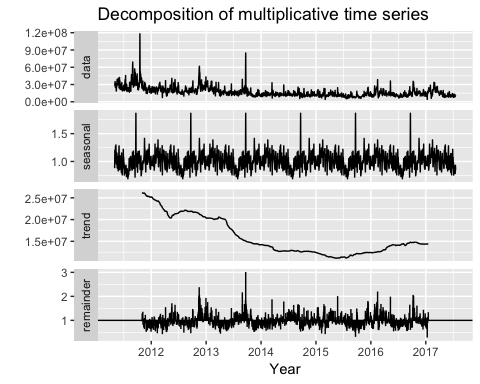
## Series: s04\_var02\_o   
## ARIMA(2,1,1)   
## Box Cox transformation: lambda= -0.4565625   
##   
## Coefficients:  
## ar1 ar2 ma1  
## 0.4430 0.0447 -0.9505  
## s.e. 0.0288 0.0279 0.0142  
##   
## sigma^2 estimated as 3.626e-08: log likelihood=11655.69  
## AIC=-23303.37 AICc=-23303.35 BIC=-23281.81  
##   
## 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

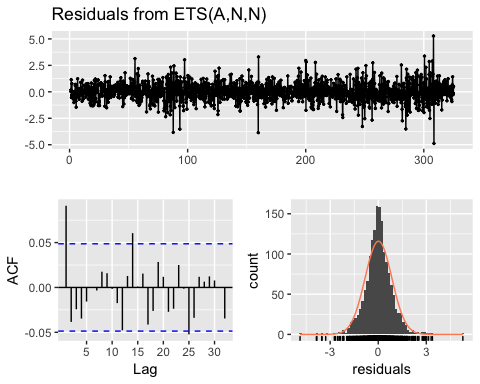
ARIMA is the preferred model for this variable forecasting.

### 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

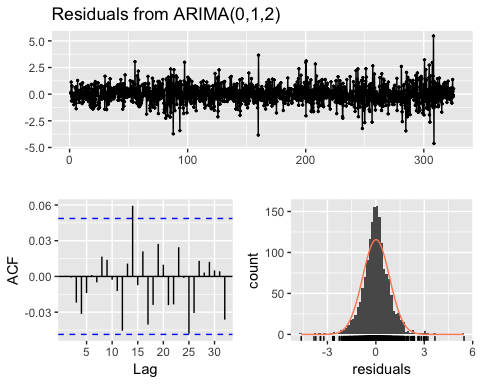


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(A,N,N)  
## Q\* = 20.144, df = 8, p-value = 0.009805  
##   
## Model df: 2. Total lags used: 10

#### 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

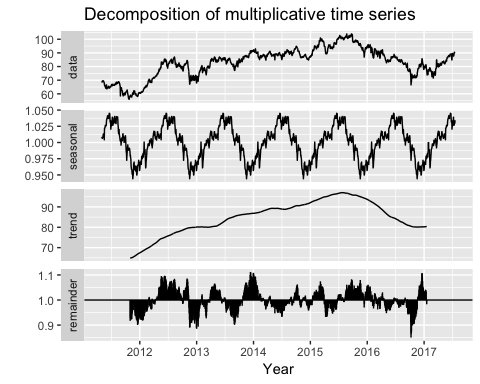
checkresiduals(s05\_var02\_arima)



##   
## 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

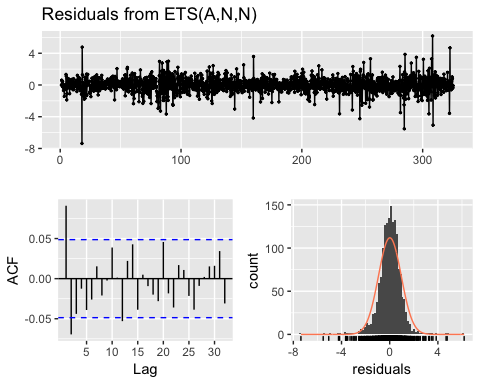
The ARIMA model resulted in the best fit with the best RMSE and a Ljung-Box p-value. The plot of the forecast also looks like a more reasonable estimate of what we can expect based on the historical data.

### 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

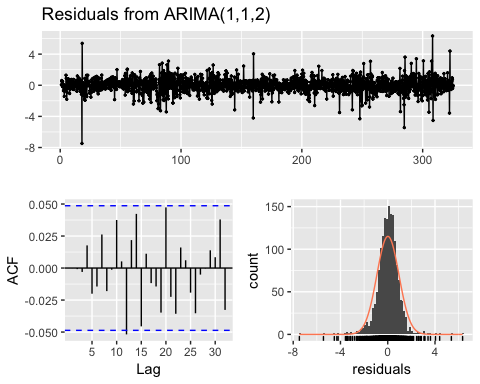


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(A,N,N)  
## Q\* = 31.866, df = 8, p-value = 9.842e-05  
##   
## Model df: 2. Total lags used: 10

#### 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)

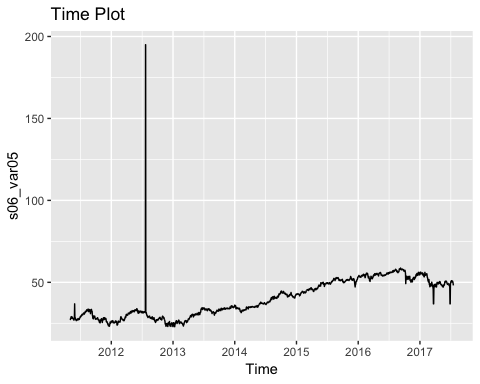


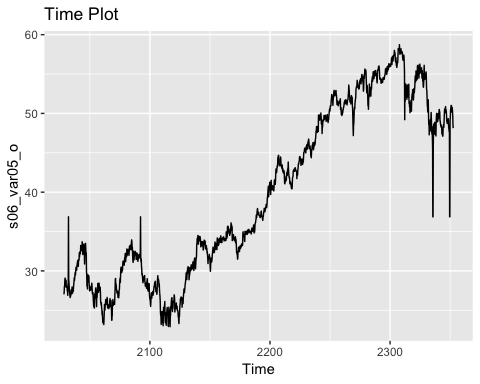
##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(1,1,2)  
## Q\* = 5.4726, df = 7, p-value = 0.6025  
##   
## Model df: 3. Total lags used: 10

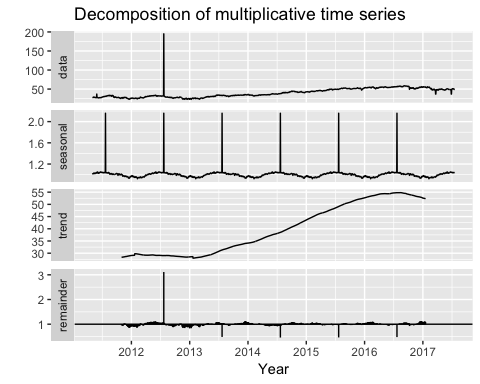
The ARIMA model gave us slightly better results than ETS model, based on AIC and sigma^2(standart deviation).

### 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.

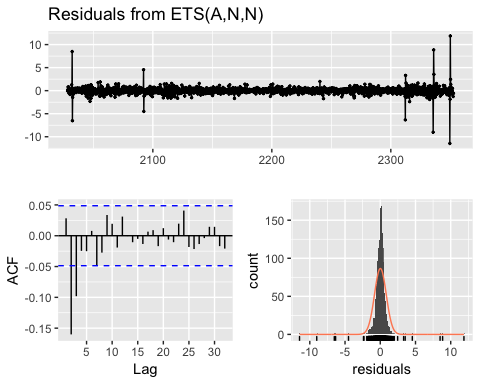




This group’s outliers were replaced with the median since it was so far above the norm, so it seemed likely to be an error.

#### 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

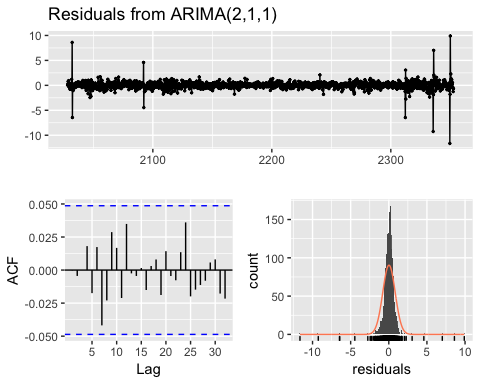


##   
## Ljung-Box test  
##   
## data: Residuals from ETS(A,N,N)  
## Q\* = 68.223, df = 8, p-value = 1.109e-11  
##   
## Model df: 2. Total lags used: 10

#### 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)

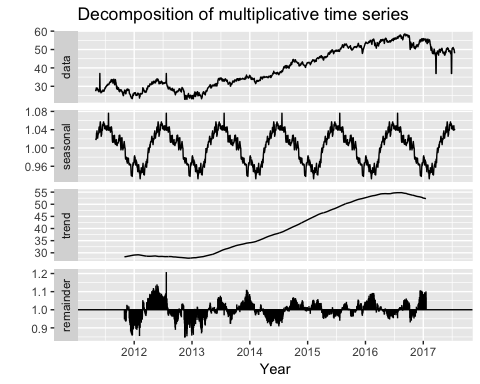


##   
## Ljung-Box test  
##   
## data: Residuals from ARIMA(2,1,1)  
## Q\* = 7.0771, df = 7, p-value = 0.4209  
##   
## Model df: 3. Total lags used: 10

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

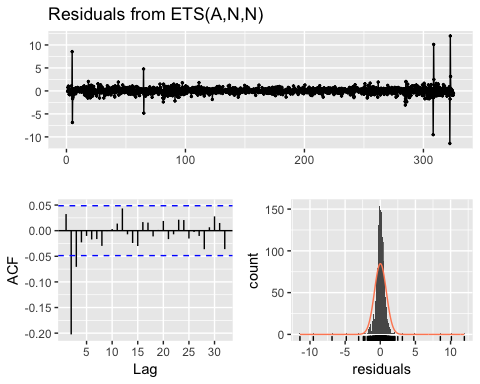
### S06 – Forecast Var07

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)

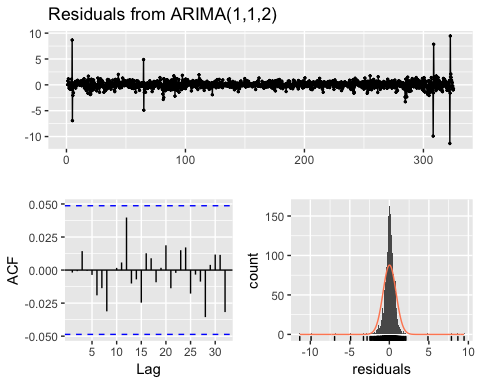
## ETS(A,N,N)   
##   
## Call:  
## ets(y = s06\_var07\_o)   
##   
## Smoothing parameters:  
## alpha = 0.8458   
##   
## Initial states:  
## l = 27.4391   
##   
## sigma: 0.8473  
##   
## AIC AICc BIC   
## 11455.32 11455.34 11471.50   
##   
## Training set error measures:  
## ME RMSE MAE MPE 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.

### Summary

After running all our models, we have come to the conclusion that the best model overall was the ARIMA model followed by the ETS model when ETS was evaluated against a more simpler Naive model which served as the baseline model. ARIMA had the best pure measurements when it came to the evaluations of the fit on each times series. However because we ran the models head to head for ETS and Naive on 3 separate groups for each of the variables within the group and then ran ARIMA and ETS head to head on the other 3 groups. We have decided to include ETS along with ARIMA as our recommendations for the forecasting model to be used by the finance department. We decided on these two models not just due to having a good fit as seen from the residual analysis, AIC and Ljung-Box test. But also because as previously mentioned, these are automated models that can be very valuable when many forecasts are needed and needed in a shorter period of time. Also the company may not have an in-house expert forecaster and using open source algorithms could be a viable alternative. And as such we have attached along with our report a forward forecast of 140 periods that were produced from both the ARIMA model and the ETS model. The Finance department will test these forecasts on the actuals and let us know the results along with their feedback.