Data ScienceTech Institute
Time Series Analysis Final Project Report

S19 Cohort Motoharu DEI

January 18th, 2020

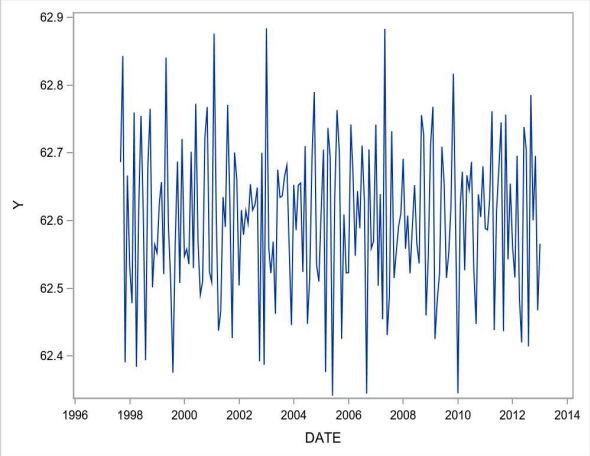
Contents

PART 1: Exercise

1.1. Quiz 1

1. You receive the SAS data set E1 from a colleague. Represent with a graph the timeseries and identify and estimate an appropriate model to fit the data. Justify your choice. (2 points)

Here is the graph of SAS data set E1.

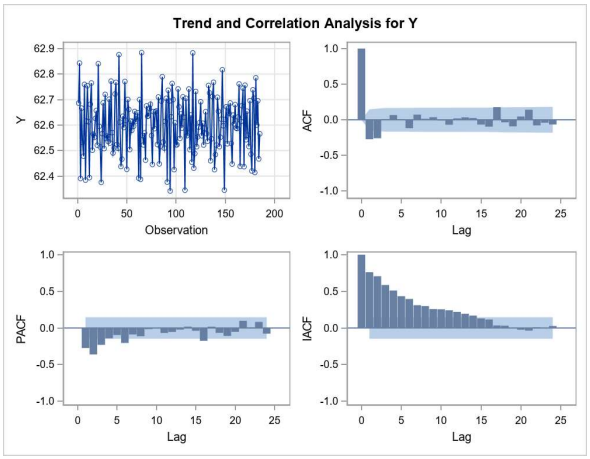


ADF test results are the followings.

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-0.0027	0.6814	-0.07	0.6595		
	1	-0.0044	0.6810	-0.19	0.6163		
	2	-0.0008	0.6818	-0.06	0.6624		
	3	-0.0010	0.6817	-0.11	0.6458		
Single Mean	0	-234.322	0.0001	-17.89	<.0001	159.96	0.0010
	1	-507.138	0.0001	-16.07	<.0001	129.24	0.0010
	2	-4755.64	0.0001	-12.93	<.0001	83.59	0.0010
	3	744.7233	0.9999	-10.85	<.0001	58.83	0.0010
Trend	0	-234.328	0.0001	-17.84	<.0001	159.10	0.0010
	1	-508.110	0.0001	-16.05	<.0001	128.88	0.0010
	2	-5029.33	0.0001	-12.92	<.0001	83.44	0.0010
	3	725.3341	0.9999	-10.87	<.0001	59.05	0.0010

Apparently, the data is not zero mean. Otherwise, most p-values are small, meaning non-stationarity hypothesis (H0) was rejected.

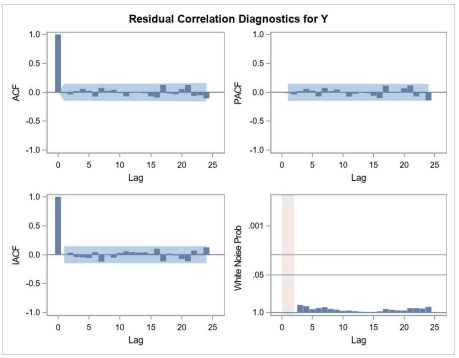
Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	30.32	6	<.0001	-0.273	-0.259	0.010	0.066	0.006	-0.118
12	32.69	12	0.0011	0.073	0.014	0.036	-0.010	-0.070	0.019
18	42.62	18	0.0009	0.034	0.024	-0.068	-0.096	0.178	-0.035
24	51.65	24	0.0009	-0.093	0.046	0.141	-0.079	-0.041	-0.066

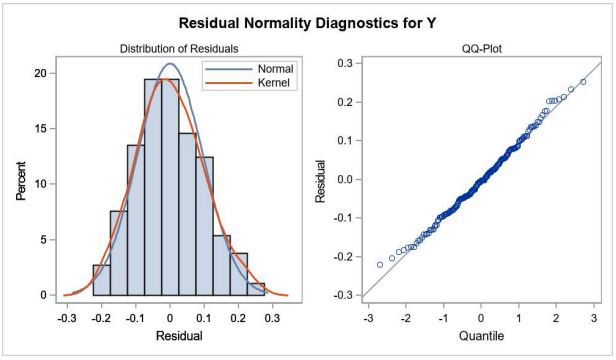


Here, the p-values in autocorrelation check for white noise (Ljung-Box test) is very low, meaning there exists auto correlation. PACF's trend is not evident but IACF has clearly exponential decreasing trend. ACF is zero after t=2. Therefore, this is assumed to be MA(2) model.

Here's the results of Ljung-Box test and autocorrelation plots with assigning ARMA(p=0, q=2), which represents the goodness of fit is high.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	2.18	4	0.7026	-0.013	-0.036	0.021	0.059	0.024	-0.074
12	4.81	10	0.9033	0.076	0.027	0.042	-0.013	-0.070	-0.008
18	10.97	16	0.8110	-0.001	-0.015	-0.071	-0.091	0.127	-0.020
24	19.07	22	0.6408	-0.034	0.055	0.127	-0.064	-0.052	-0.105
30	23.80	28	0.6922	0.036	-0.080	0.001	0.026	-0.108	0.036
36	27.64	34	0.7714	0.080	0.006	-0.009	-0.052	0.004	-0.087





Here's the final model parameters.

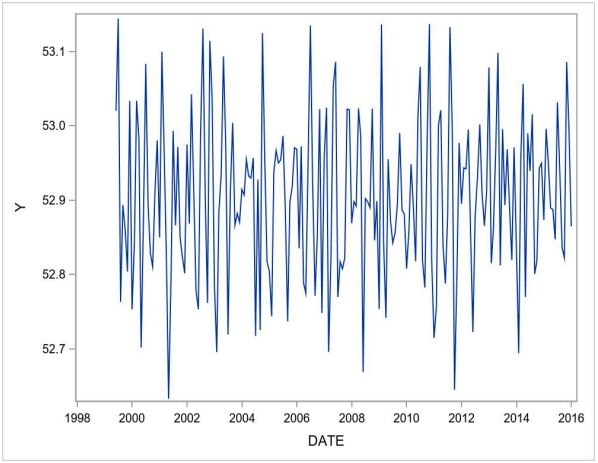
Model for variable Y	
Estimated Mean	62.60008

Moving Average Factors	
Factor 1:	1 - 0.56908 B**(1) - 0.30819 B**(2)

1.2. Quiz 2

> 2. Identify and estimate a relevant model for the variable Y in the SAS data set E2. You will use the Maximum Likelihood estimation method to obtain your model. Explain how you have decided which model to select. (2 points)

Here is the graph of SAS data set E2.

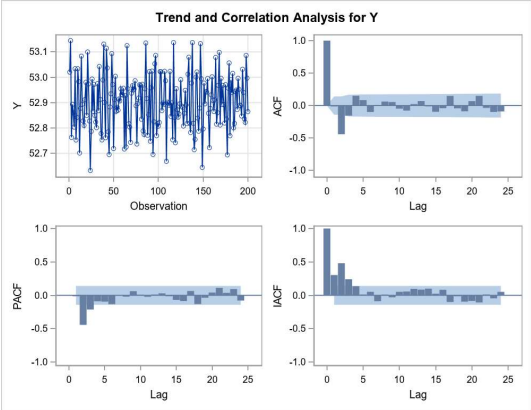


ADF test results are the followings.

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-0.0038	0.6812	-0.09	0.6516		
	1	-0.0047	0.6810	-0.15	0.6307		
	2	0.0005	0.6822	0.03	0.6916		
	3	-0.0001	0.6820	-0.01	0.6793		
Single Mean	0	-201.802	0.0001	-14.27	<.0001	101.85	0.0010
	1	-526.528	0.0001	-16.45	<.0001	135.28	0.0010
	2	-3398.64	0.0001	-13.17	<.0001	86.75	0.0010
	3	1392.561	0.9999	-10.57	<.0001	55.87	0.0010
Trend	0	-202.126	0.0001	-14.26	<.0001	101.72	0.0010
	1	-530.990	0.0001	-16.51	<.0001	136.40	0.0010
	2	-3942.38	0.0001	-13.27	<.0001	88.01	0.0010
	3	1194.449	0.9999	-10.74	<.0001	57.67	0.0010

Apparently, the data is not zero mean. Otherwise, most p-values are small, meaning non-stationarity hypothesis (H0) was rejected.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	53.09	6	<.0001	-0.014	-0.442	-0.156	0.150	0.084	-0.099
12	56.39	12	<.0001	0.006	0.061	0.054	-0.048	-0.078	0.022
18	65.08	18	<.0001	0.068	-0.012	-0.096	-0.041	0.148	-0.043
24	77.28	24	<.0001	-0.089	0.067	0.148	-0.037	-0.099	-0.091
30	84.81	30	<.0001	0.058	0.024	0.052	0.002	-0.155	0.035
36	91.64	36	<.0001	0.143	0.020	-0.049	-0.027	0.007	-0.067



Here, the p-values in autocorrelation check for white noise (Ljung-Box test) is very low, meaning there exists auto correlation. Now it looks to be the model is ARMA model with non-zero p and non-zero-q. Therefore, I will estimate parameters by ESACF, SCAN, and MINIC with p=0:12 and q=0:12.

ARMA(p+d,q) Tentative Order Selection Tests	
ESACF	
p+d	q
0	4
8	4
9	3
6	6
4	7
5	7
11	3
12	2

(5% Significance Level)

ARMA(p+d,q) Tentative Order Selection Tests	
SCAN	
p+d	q
2	1
3	0
0	3

(5% Significance Level)

Minimum Table Value: BIC(3,0) = -4.85277

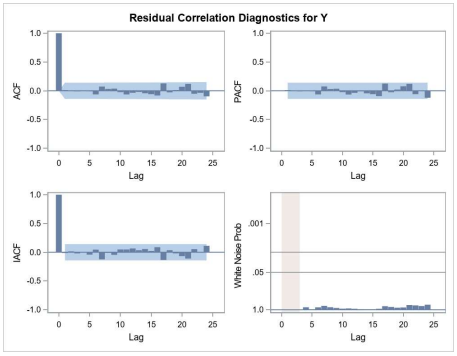
Through three estimations, I will check following four options: (p,q) = (0,4), (2,1), (3,0), (0,3).

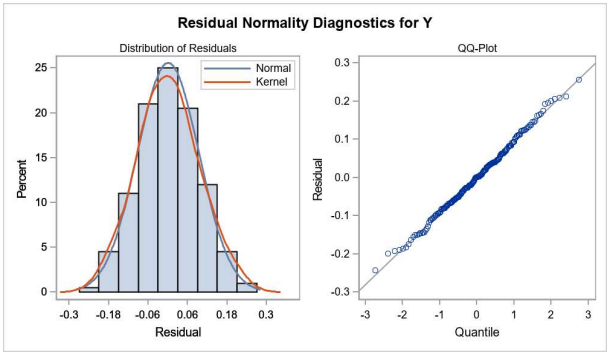
Based on AICs below, I choose (p,q)=(2,1).

(p,q)	AIC
(0,4)	-368.205
(2,1)	-372.378
(3,0)	-366.918
(0,3)	-365.058

Here're the Ljung-Box test results and autocorrelation plots, which represents the model is reasonable.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	0.91	3	0.8232	0.011	0.004	-0.010	0.005	-0.008	-0.064
12	3.78	9	0.9255	0.073	0.031	0.037	-0.024	-0.064	-0.035
18	10.34	15	0.7981	-0.020	-0.043	-0.060	-0.083	0.128	-0.028
24	17.73	21	0.6658	0.012	0.070	0.119	-0.052	-0.036	-0.097
30	23.11	27	0.6792	0.040	-0.050	0.023	0.064	-0.089	0.078
36	27.43	33	0.7407	0.110	0.026	-0.002	-0.024	0.020	-0.064





And the final model parameters are the followings.

Model for variable Y	
Estimated Mean	52.90186

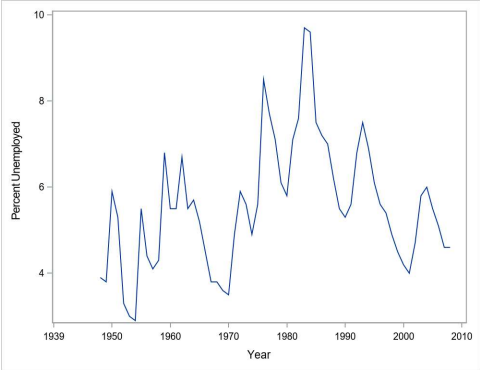
Autoregressive Factors	
Factor 1:	1 - 0.41283 B**(1) + 0.44116 B**(2)

Moving Average Factors	
Factor 1:	1 - 0.5939 B**(1)

1.3. Quiz 3

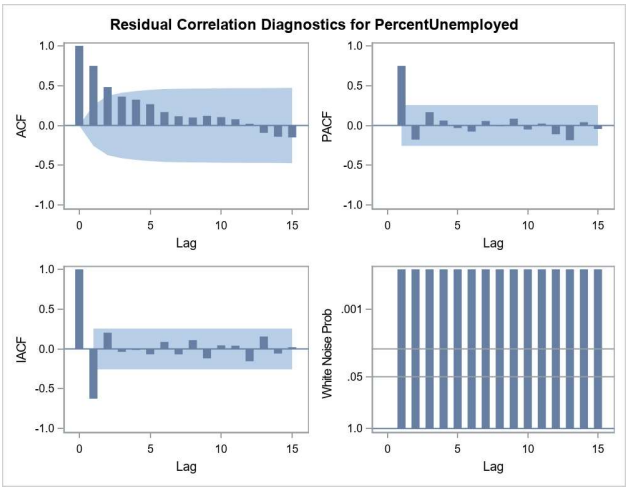
> 3. Perform the Ljung-Box White Noise Probability test on the variable PercentUnemployed in the SAS data set E3. You should give the null and alternative hypothesis. What can you conclude from this test? (2 points)

Here is the plot of the data.



Here is the results of the Ljung-Box White Noise Probability test.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	73.65	6	<.0001	0.748	0.482	0.362	0.324	0.267	0.168
12	77.71	12	<.0001	0.115	0.100	0.121	0.105	0.078	0.021
18	84.60	18	<.0001	-0.093	-0.142	-0.150	-0.109	-0.073	-0.116
24	105.00	24	<.0001	-0.115	-0.082	-0.027	-0.109	-0.261	-0.317



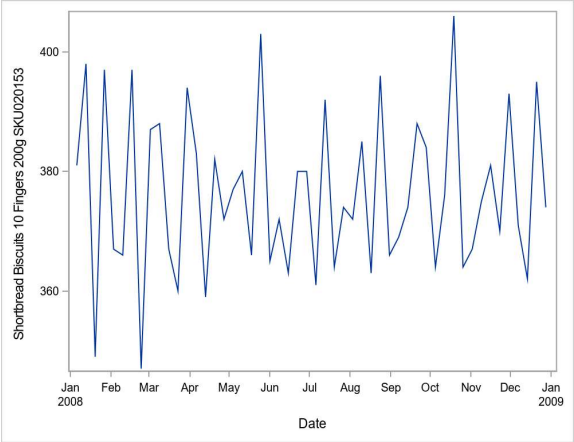
The null hypothesis and alternative hypothesis of Ljung-Box White Noise Probability test are:

- H0: The error term ε_t is white noise (meaning with zero mean, constant variance, and no autocorrelation) on $Y_t = \phi_0 + \varepsilon_t$.
- H1: Otherwise.

1.4. Quiz 4

> 4. Using the PROC ESM in SAS, generate a forecast for the next 12 periods for the variable Biscuits in the SAS data set E4 with the model of your choice. Justify your choice. (2 points)

This is the plot of the data.



The data has no trend nor yearly seasonality. Funnel effect was not observed. No outstanding outlier observed. However, the data has strong cyclic movement where cycle length=3 or 4.

Then, I will try Double ESM, Additive Holt-Winter with cycle = 3 and 4 weeks, and Multicative Holt-Winter with cycle = 3 and 4.

I also plan the following 6 train/validation splits and choose the model having minimum average MAE over the 6 validation sets.

- Fold1 - training: January to June, test: July to December
- Fold2 - training: January to July, test: August to December
- Fold3 - training: January to August, test: September to December
- Fold4 - training: January to September, test: October to December
- Fold5 - training: January to October, test: November to December
- Fold6 - training: January to November, test: December

Here is the summary of the average MAEs for all the models.

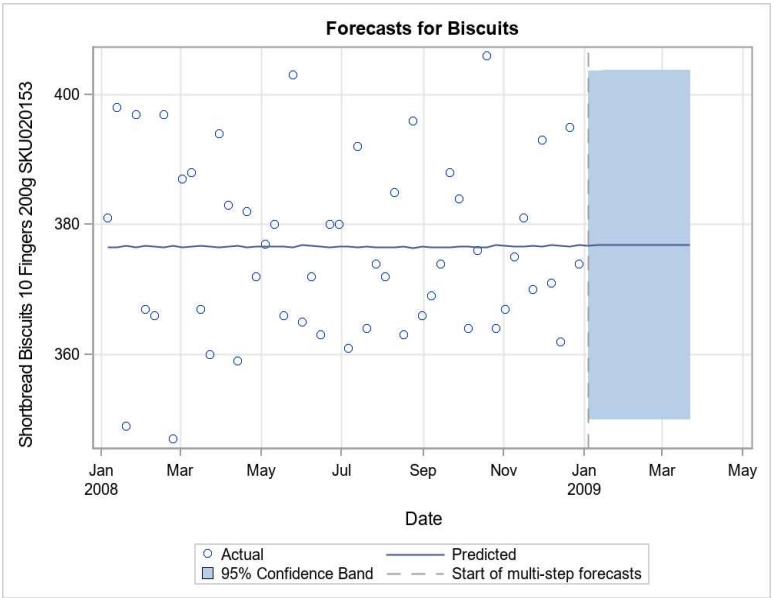
Variable	N	Mean	Std Dev	Minimum	Maximum
MAE_double1	6	9.8151022	0.7394115	8.8655802	10.8485332
MAE_addwinters3	6	11.7928508	1.9553108	8.5827825	14.2237050
MAE_addwinters4	6	10.7839279	1.7912023	9.2885852	14.2343501
MAE_winters3	6	12.2246935	2.3079573	8.6196000	14.6953726
MAE_winters4	6	10.7628090	1.7550603	9.5199402	14.2335079

, where

- *double1*: Double ESM,
- *addwinter3*: Additive Holt-Winter ESM with 3-week cycle,
- *addwinter4*: Additive Holt-Winter ESM with 4-week cycle,
- *winter3*: Multicative Holt-Winter ESM with 3-week cycle, and
- *winter4*: Multicative Holt-Winter ESM with 4-week cycle.

Double ESM has minimum average MAE and MAE SD, therefore I choose Double ESM as my final model. A possible reason why seasonal model was not chosen would be the cycle in the data is irregular and the seasonal models tend to mis-capture it.

Here is the forecast for the next 12 periods by Double ESM, the final ESM.



PART 2: Case Study

The Sales department asked you to provide a statistical forecast for 3 key products for the next 16 months (last forecast in December 2019). You managed to extract the relevant data in the file DSTI_SAS_ETS_Evaluation_Part2.csv.
Using all what you have learned in Times Series in SAS, generate a forecast for the 3 different products. You will explain all the steps you have followed to choose the models and you will write a quick report for the Sales department to understand the sales evolution of these products.

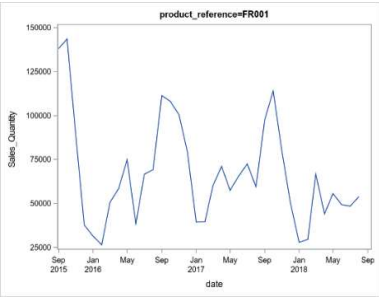
2.1. Technical explanation of the steps

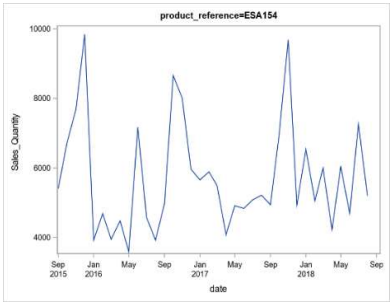
2.1.1. Preprocessing

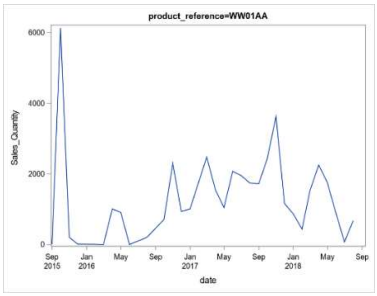
The very first step of preprocessing was after the import of .csv file, the conversion of month to date time format of the SAS. See the attached SAS code for more details.

The data of product_reference=WW01AA had three missing months in the data, which were January of 2016, June of 2016, and September of 2016. I imputed them with the averages of neighboring months, 8.5, 101, and 456, respectively.

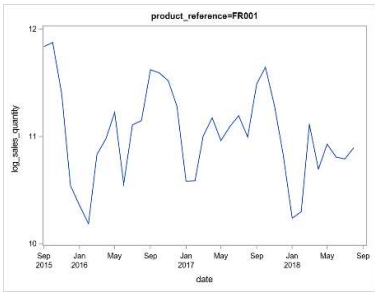
Here are the plots of preprocessed data by product_reference.

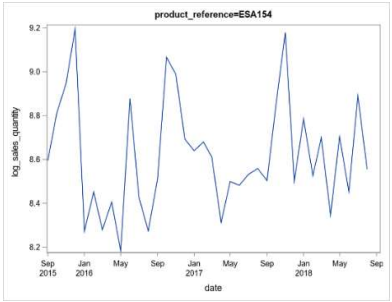


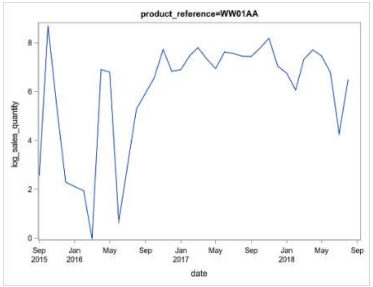




Also, here are the plots after log transformation.





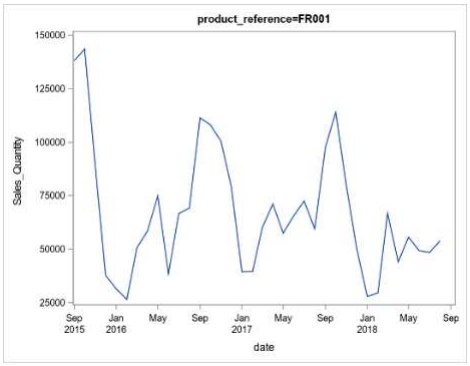


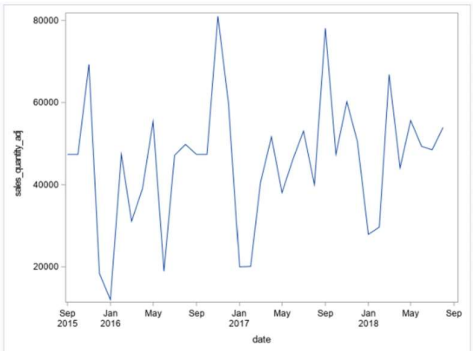
2.1.2. Modeling of "FR001" product_reference

Firstly, I detected the outliers using *proc arima* and *outlier* option with *maxnum=7*. There were 6 spikes and 1 level shift identified.

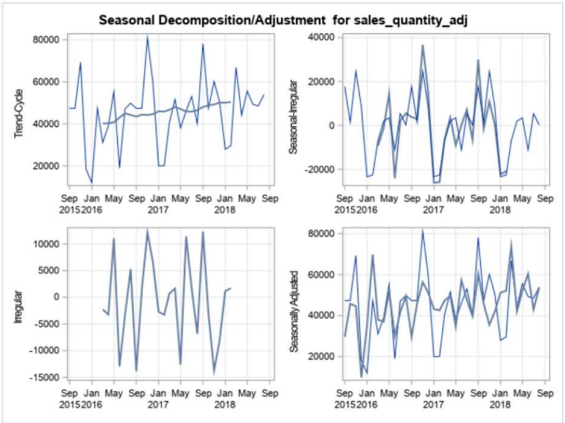
Outlier Details					
Obs	Time ID	Type	Estimate	Chi-Square	Approx Prob>ChiSq
2	01-OCT-2015	Additive	76611.0	7.85	0.0051
1	01-SEP-2015	Additive	71259.0	7.47	0.0063
28	01-DEC-2017	Shift	-19465.4	5.74	0.0166
26	01-OCT-2017	Additive	47134.0	6.56	0.0104
13	01-SEP-2016	Additive	44487.0	10.31	0.0013
14	01-OCT-2016	Additive	41251.0	11.08	0.0009
6	01-FEB-2016	Additive	-40327.0	10.72	0.0011

The following two charts are the comparison of before (left) and after (right) the adjustments of outliers.





Using *proc timeseries*, the outlier-adjusted data has the following decompositions: a somewhat strong trend and a strong seasonality.

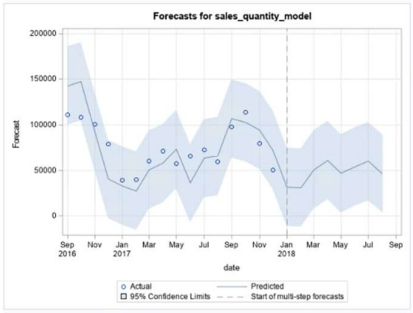


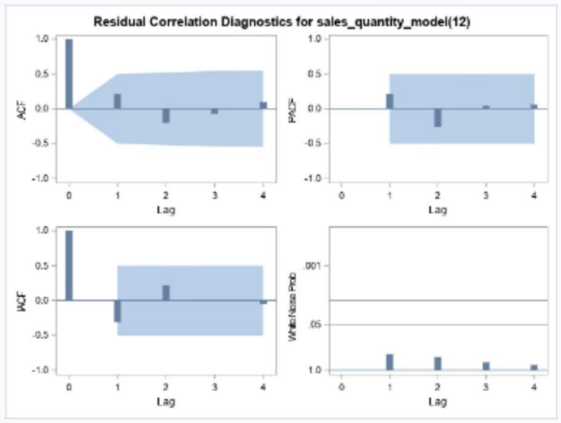
Based on these observations, the following four models are examined. The model fit was done on the data from September of 2015 to December of 2017 (28 months) and the model validation was don on the data from January of 2018 to August of 2018 (8 months). I chose MAE as the validation score.

Model #	Data	Explanatory Variable	Differencing	MAE on Validation Set
1	Raw	Dummy variables for three spikes and one level shift	None	17,535.00
2	Raw	Deterministic trend	12 months	8,878.19
3	Raw	None	1 month and 12 months	18,116.70
4	Outlier-adjusted	Exponential smoothing (Winters additive method)		8,510.06

Since model 2 and 4 are closely good, I will take a further look at the results.

Model 2:



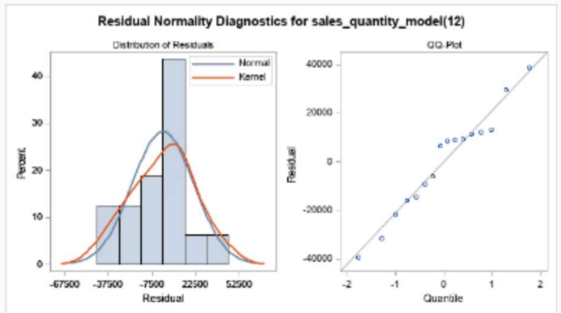


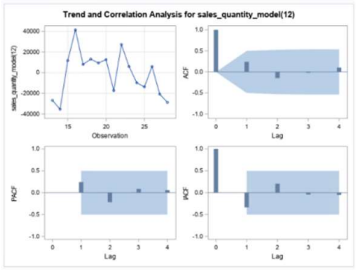
Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations				
6	4.32	6	0.6337	0.214	-0.207	-0.077	0.098	-0.289	-0.024
12	11.65	12	0.4741	0.135	-0.173	-0.298	-0.006	-0.009	-0.167

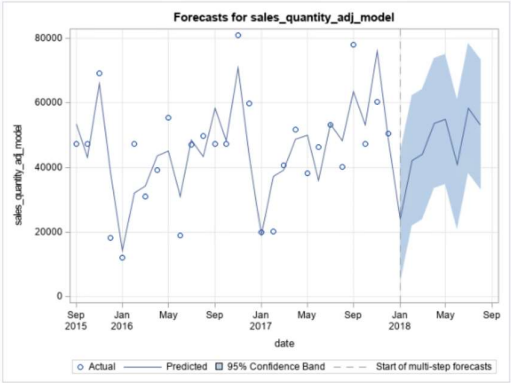
Augmented Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-10.9934	0.0098	-2.82	0.0080		
	1	-19.1220	<.0001	-3.09	0.0045		
	2	-12.1130	0.0048	-1.62	0.0945		
	3	-1.9828	0.3135	-0.57	0.4478		
Single Mean	0	-11.0053	0.0490	-2.72	0.0942	3.69	0.1981
	1	-20.1651	0.0003	-3.07	0.0530	4.72	0.0719
	2	-13.7448	0.0111	-1.54	0.4808	1.24	0.7632
	3	0.5510	0.9644	0.16	0.9566	0.70	0.8877
Trend	0	-11.1967	0.2087	-2.85	0.2032	4.83	0.2882
	1	-19.1868	0.0051	-4.36	0.0201	13.17	0.0010
	2	-37.0922	<.0001	-4.77	0.0120	14.29	0.0010
	3	-38.5259	<.0001	-2.29	0.4058	4.03	0.4292





Model 4:



Both look successfully captured the seasonality with lower January and February, and increasing trend toward the end of year as we observed after the outlier adjustments. For narrower CI, **I preferred exponential smoothing model (Model 4)**.

2.1.3. Modeling of "ESA154" product_reference

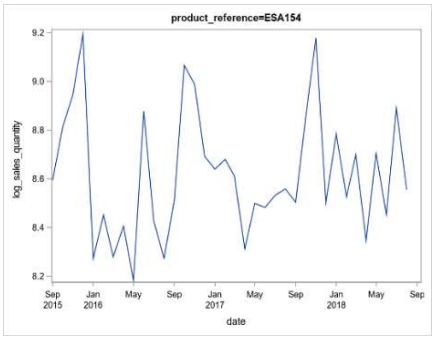
As we observed in '2.1.1. Preprocessing' section, this product has more unskewed distributed with logarithmic transformation. Therefore, **I will use the log transformed data for entire modeling and exponentialize back after prediction**.

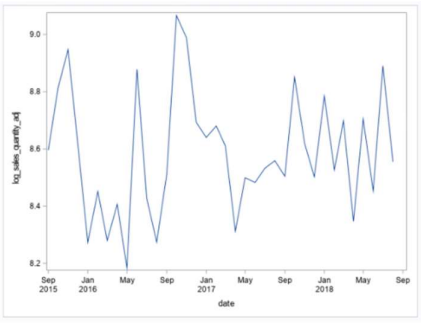
Firstly, I detected the outliers using *proc arima* and *outlier* option with *maxnum=5*. There were 2 spikes.

Outlier Details

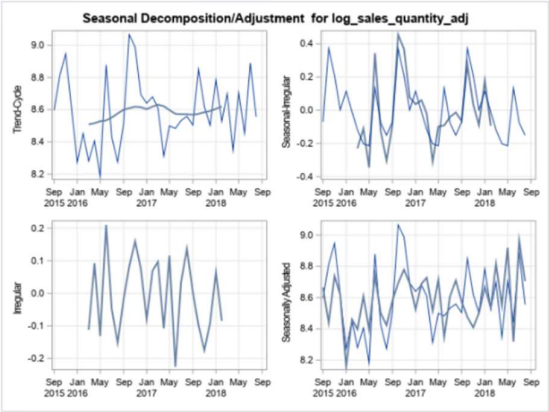
Obs	Time ID	Type	Estimate	Chi-Square	Approx Prob>Chi Sq
4	01-DEC-2015	Additive	0.57532	5.31	0.0212
27	01-NOV-2017	Additive	0.55924	5.11	0.0238

The following two charts are the comparison of before (left) and after (right) the adjustments of outliers.





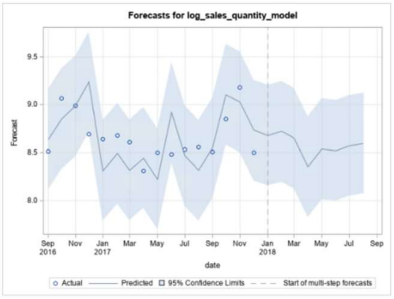
Using *proc timeseries*, the outlier-adjusted data has the following decompositions: a very weak trend and a strong seasonality.



Based on these observations, the following four models are examined. The model fit was done on the data from September of 2015 to December of 2017 (28 months) and the model validation was don on the data from January of 2018 to August of 2018 (8 months). I chose MAE as the validation score.

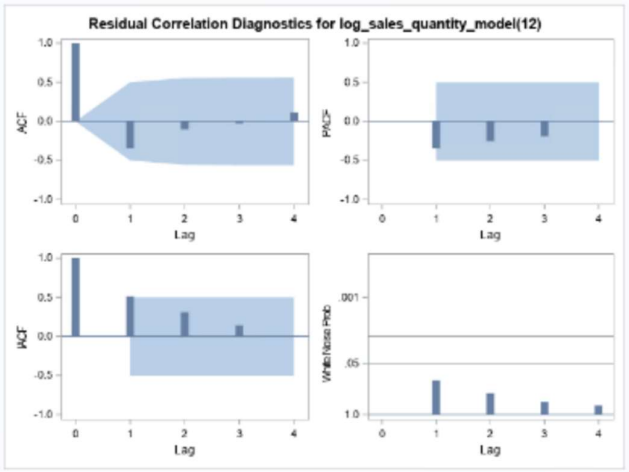
Model #	Data	Explanatory Variable	Differencing	MAE on Validation Set
1	Raw (log transformed)	None	12 months	0.1185
2	Raw (log transformed)	None	2 months and 12 months	0.1387
3	Outlier-adjusted	Exponential smoothing (Additive seasonal exponential smoothing)		0.2234

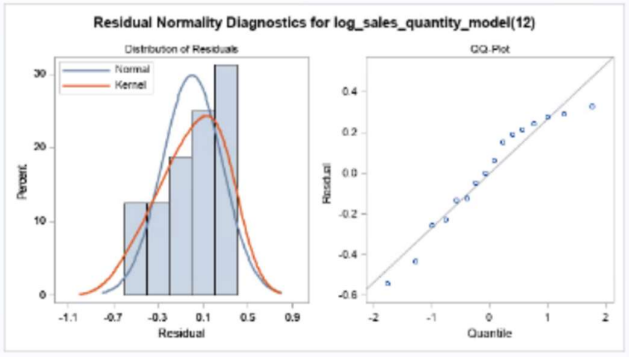
Model 1 is the best. A further look at the results (next page):



Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	9.07	6	0.1699	-0.343	-0.105	-0.026	0.111	-0.293	0.374
12	14.90	12	0.2469	-0.044	-0.232	-0.027	0.176	-0.184	0.070

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-19.6833	<.0001	-5.09	<.0001		
	1	-26.5139	<.0001	-3.42	0.0022		
	2	-40.2379	<.0001	-2.51	0.0162		
	3	-11.8185	0.0049	-1.49	0.1191		
Single Mean	0	-20.4598	0.0003	-5.16	0.0011	13.34	0.0010
	1	-31.6516	<.0001	-3.51	0.0247	6.18	0.0276
	2	-151.945	0.0001	-2.67	0.1052	3.62	0.2149
	3	167.1391	0.9999	-2.16	0.2295	2.35	0.5090
Trend	0	-20.4661	0.0031	-5.04	0.0060	12.92	0.0010
	1	-32.8631	<.0001	-3.41	0.0894	5.83	0.1142
	2	-171.140	0.0001	-2.56	0.2978	3.31	0.5538
	3	1230.013	0.9999	-2.39	0.3653	4.43	0.3581





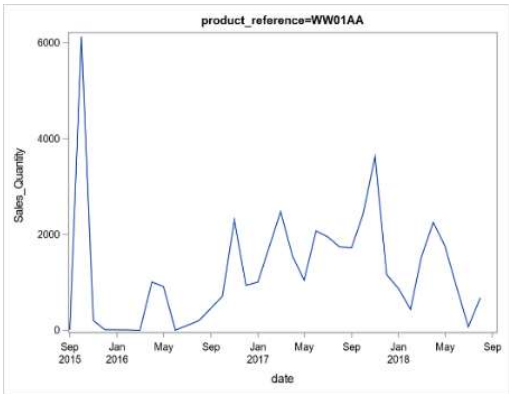
The diagnosis metrics look good.

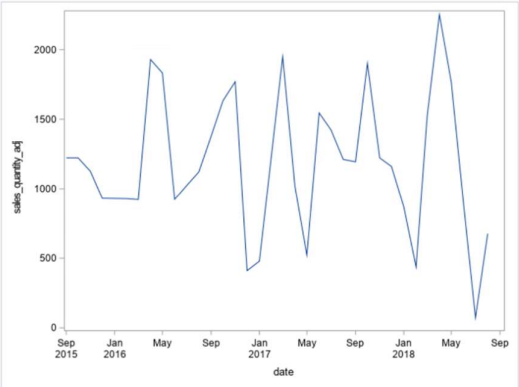
2.1.4. Modeling of "WW01AA" product_reference

Firstly, I detected the outliers using *proc arima* and *outlier* option with *maxnum=5*. There were 3 spikes and two temporary shifts.

Outlier Details					
Obs	Time ID	Type	Estimate	Chi-Square	Approx Prob>ChiSq
2	01-OCT-2015	Additive	4912.4	17.57	<.0001
3	01-NOV-2015	Temp(12)	-921.77778	7.80	0.0052
27	01-NOV-2017	Additive	2400.4	15.42	<.0001
15	01-NOV-2016	Temp(12)	526.76389	12.00	0.0005
1	01-SEP-2015	Additive	-1210.6	6.34	0.0118

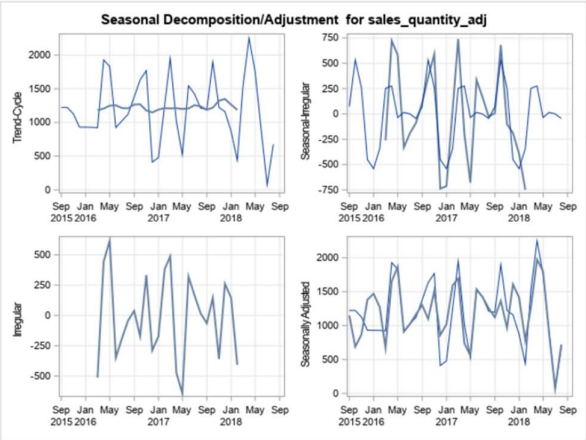
The following two charts are the comparison of before (left) and after (right) the adjustments of outliers.





The original data is highly volatile and finding cycles is not apparent. Therefore**, I will used the outlier-adjusted data for the entire modeling**.

Using *proc timeseries*, the outlier-adjusted data has the following decompositions: a very weak trend and a strong seasonality.

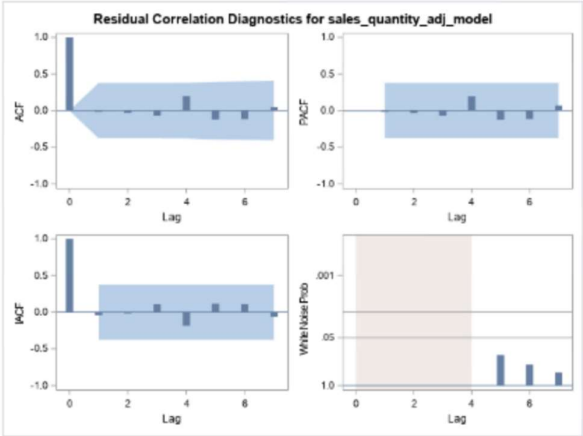


Based on these observations, the following four models are examined. The model fit was done on the data from September of 2015 to December of 2017 (28 months) and the model validation was don on the data from January of 2018 to August of 2018 (8 months). I chose MAE as the validation score.

Model #	Data	ARIMA(p,q)	Differencing	MAE on Validation Set
1	Outlier-adjusted	p=2, q=2	12 months	654.11
2	Outlier-adjusted	p=2	1 month and 12 months	884.16
3	Outlier-adjusted	p=2, q=2	None	556.30
4	Outlier-adjusted	Exponential smoothing (Additive seasonal exponential smoothing)		532.34

Since model 3 and 4 are closely good, I will take a further look at the results.

Model 3:

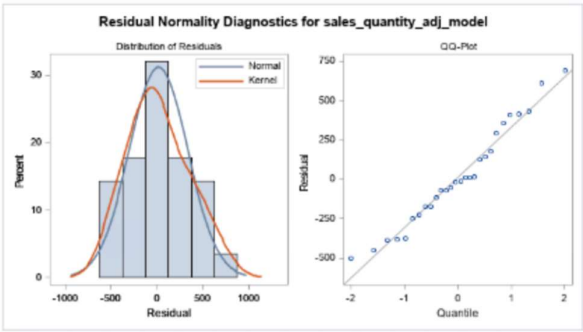


Autocorrelation Check of Residuals

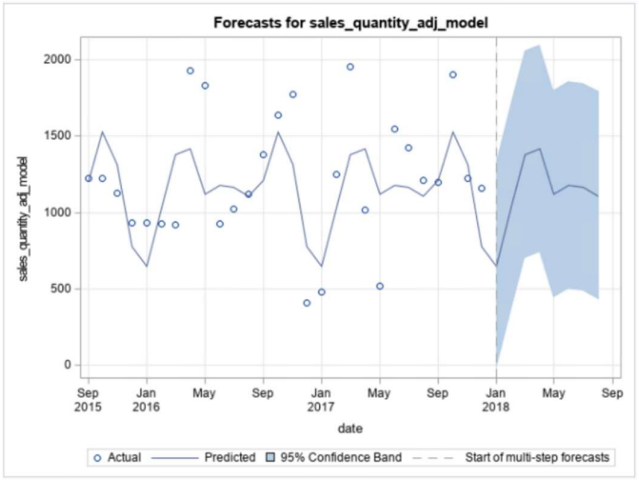
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	2.57	2	0.2765	-0.019	-0.032	-0.071	0.193	-0.124	-0.114
12	11.20	8	0.1906	0.050	-0.313	-0.061	-0.133	0.232	-0.120
18	15.11	14	0.3706	0.148	0.157	0.037	-0.011	0.091	0.077
24	20.48	20	0.4283	-0.089	-0.024	-0.130	-0.114	-0.055	-0.032

Warning: The model defined by the new estimates is unstable. The iteration process has been terminated.

Warning: Estimates may not have converged.



Model 4:



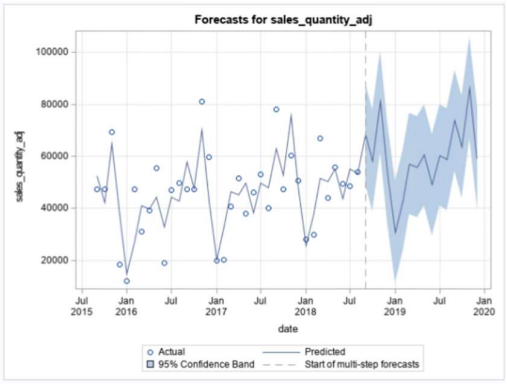
Though Model 2 diagnosis looks good but there appeared an error message notifying the estimate is unstable. Therefore, **I prefer Model 4**.

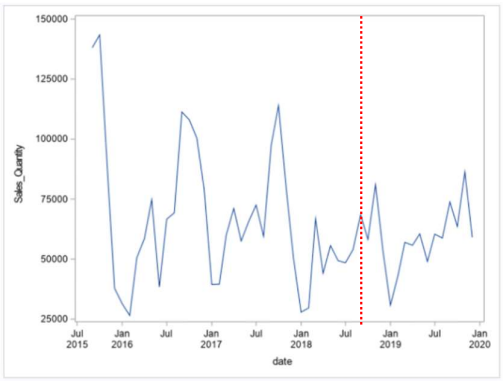
2.1.5. Final prediction of next 16 months

Data used in the training, prediction, and CI of prediction

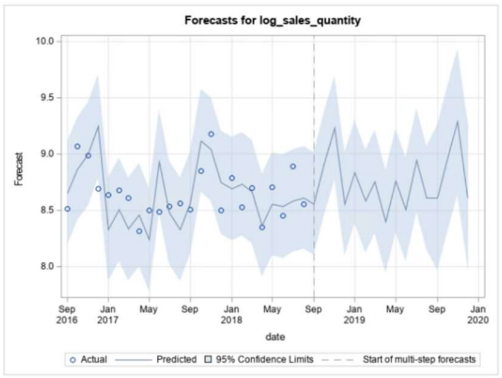
Entire period with original data and predicted values

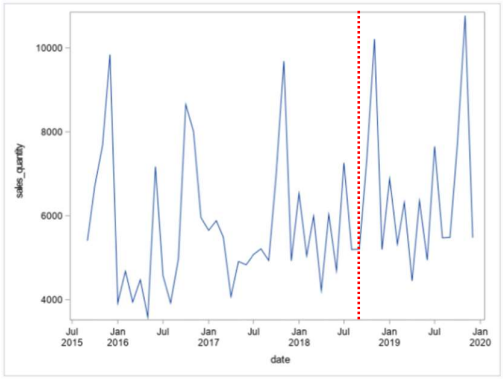
"product_reference"="FR001"



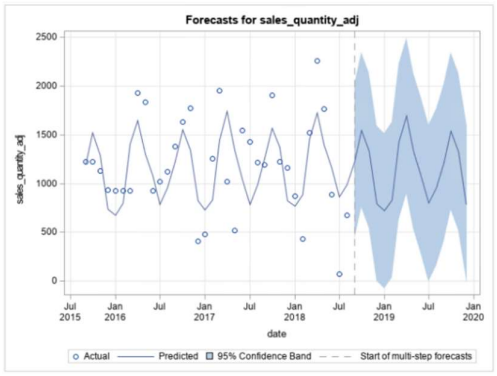


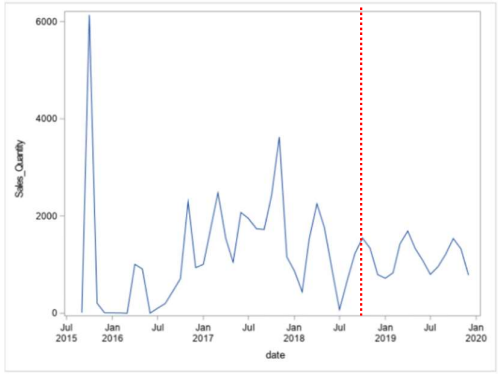
"product_reference"="ESA154"





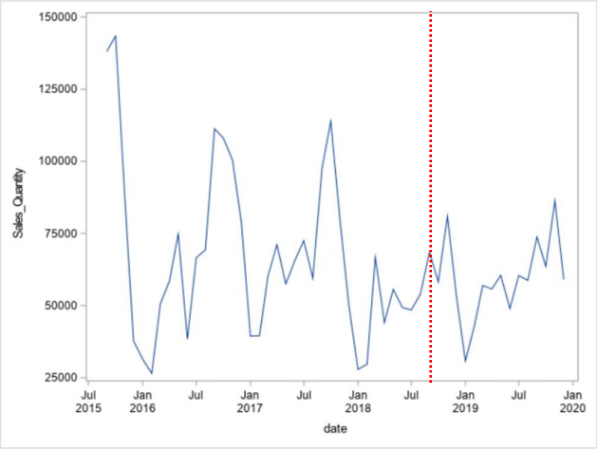
"product_reference"="WW01AA"





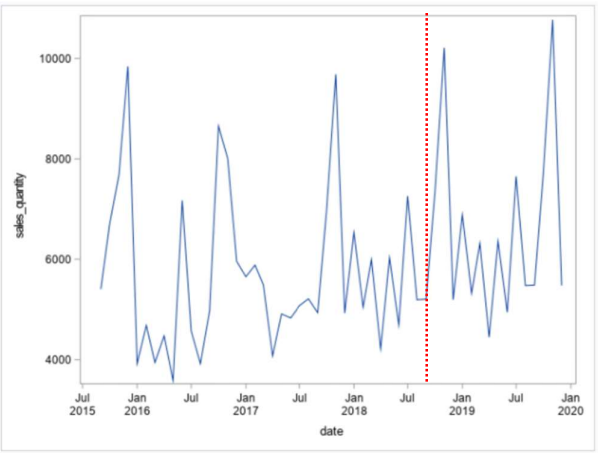
2.2. A quick report to sales department

We predict the future sales quantity of product “FR001” as below (right to the dotted red line).



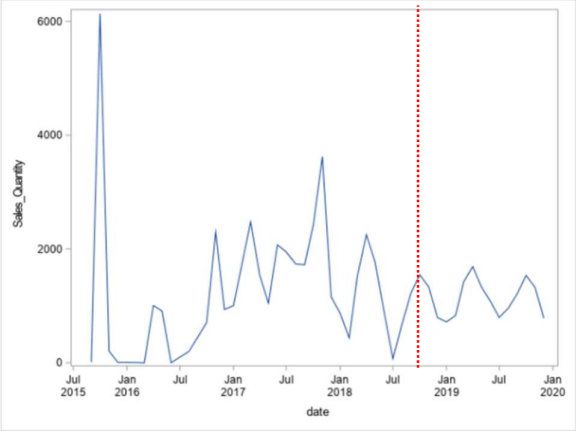
We predict the highest sales in November in a year, next two months, December and January, have sharp drop of sales which reaches at the lowest in a year on January. The sales from February bounce back and keep going up to the end of the year, with some monthly variance. Also, we predict the overall increasing trend.

We predict the future sales quantity of product “ESA154” as below.



We predict the sharp high sales in every November. Other months have up and down cycle within every two months, with lowest in April and highest in July. Also, we predict the overall increasing trend.

We predict the future sales quantity of product “WW01AA” as below.



We predict there will be a mid-term cycle in the sales for six months, with lowest on January and July, and highest on April and October. Please also note that due to high data volatility in the past, our prediction returns wide range of confidence interval, meaning the possible range of sales value is wide, as much as plus or minus 750. We would suggest to keep our eye wide open for the couple of next months to see how the actual sales turn out compared to our prediction.

[End of report]