Tutorial 4: ARIMA

Finally, a model where you can use SAS EG. Creating ARIMA is going to be a piece of cake for you!

Before you start: Create a new process flow

This tutorial assumes that you have already created a project. You need to open it. For this tutorial we will create a new Process Flow, which will let us separate tutorials from each other and keep our project more organized (described in tutorial 2).

Click on the icon that looks like a sheet of paper → Process Flow → Right click on the icon Process Flow once you see in the pane on the left Process Flow **→ Call it tutorial4.**

When you work with the materials for Tutorial3, make sure you left click on that Process Flow so that all the new datasets you open and create, get opened under that tutorial.

For this tutorial we are going to use QUERY_FOR_BIKESHARINGDAILY, that we created in tutorial 1.

Running a libname statement!

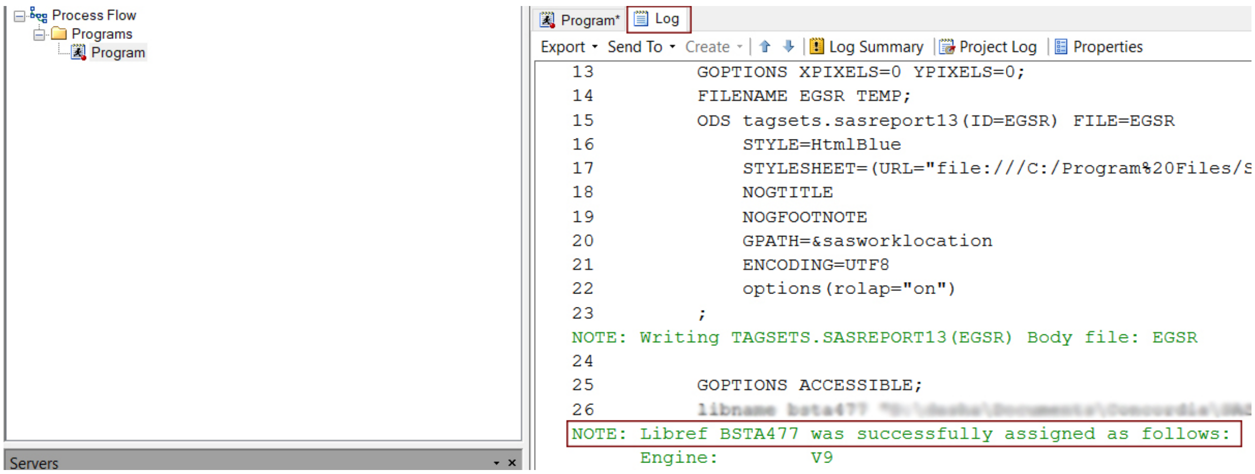
If you work from your own computer and you have assigned a permanent libname statement, you don't have to do anything.

If you are working from the school's computer, then just open a new project and run a libname statement in a SAS program. Make sure that you are in the TUTORIAL4 Process Workflow.

Click on the icon that looks like a sheet of paper → Choose program → Type the libname statement (as shown below)

```
libname bsta477 "C:\bsta477";
```

Remember to always check the log. It will show whether the library has been successfully assigned or not!



Autocorrelation and Partial Correlation Plots

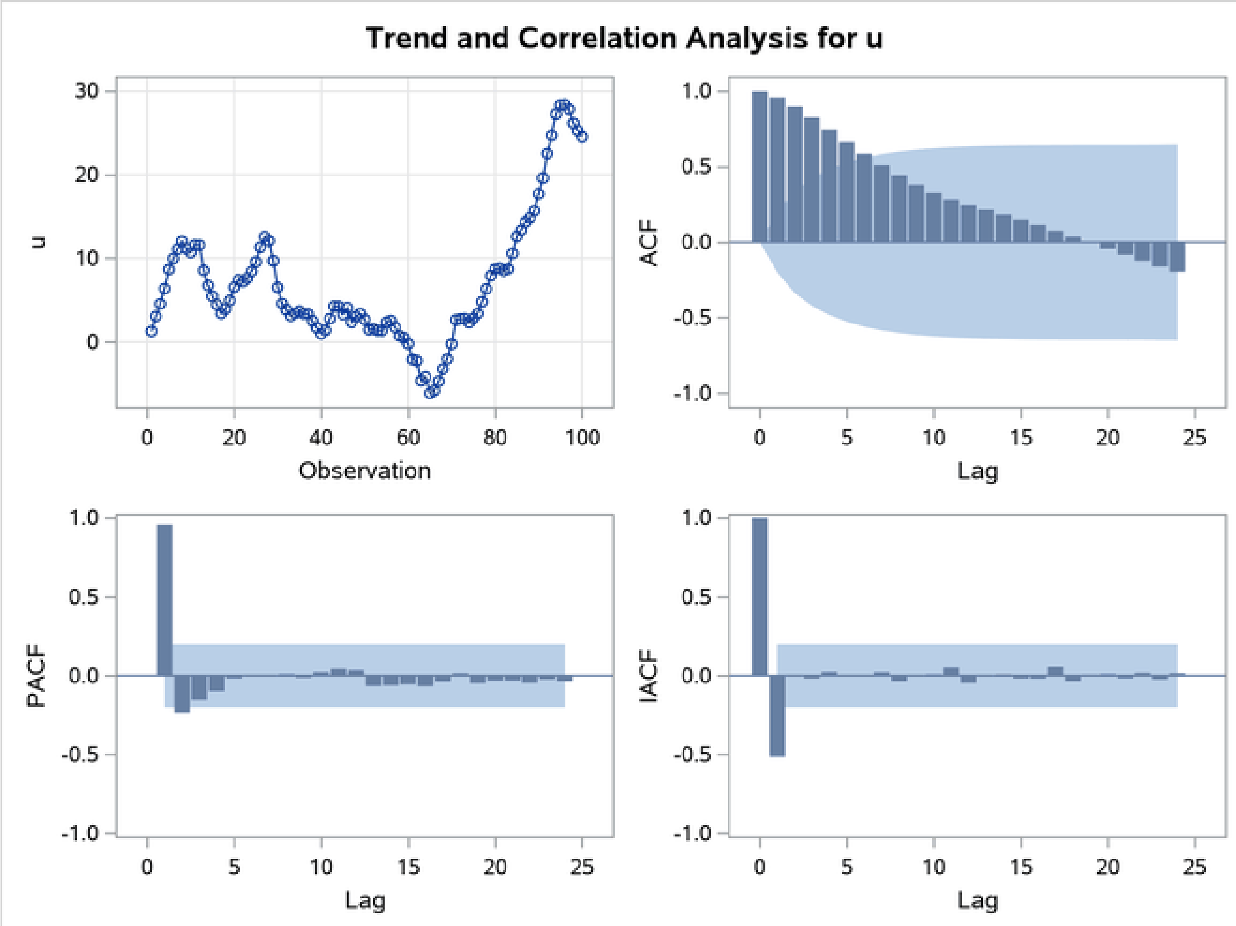
When applying ARIMA we should make sure that our data is stationary and we should also investigate what order of processes we should apply. Shall we use MA, AR or just difference the data?

A. Look at the autocorrelation (ACF) and partial (PACF) autocorrelation plots to estimate what order autocorrelations and what processes we have in the data.

Time series graph. You can visually examine the time series plot and assess whether there seems to be a trend and seasonality. In the case of the output presented here, we may see that the trend is present since the time series data points depart from the mean and don't randomly fluctuate around it. We cannot say much about seasonality, though, since it may be masked by the trend or, perhaps, we would need to zoom in and investigate a short period of time to see the data closer and estimate the presence of seasonality.

ACF and PACF plots. Other clues to look for would be ACF and PACF plots – Autocorrelation and Partial autocorrelation plots. How do we know if we are dealing with strong or moderate autocorrelations? By looking at the confidence interval of 95%, which is represented by the blue ribbon surrounding the mean of 0. All spikes within that blue ribbon would correspond to low or insignificant autocorrelation. The spikes that go over it correspond to moderate or strong autocorrelations. The higher the spikes, the stronger the level of autocorrelations.

When you do a preliminary analysis of your data, SAS will produce this output:



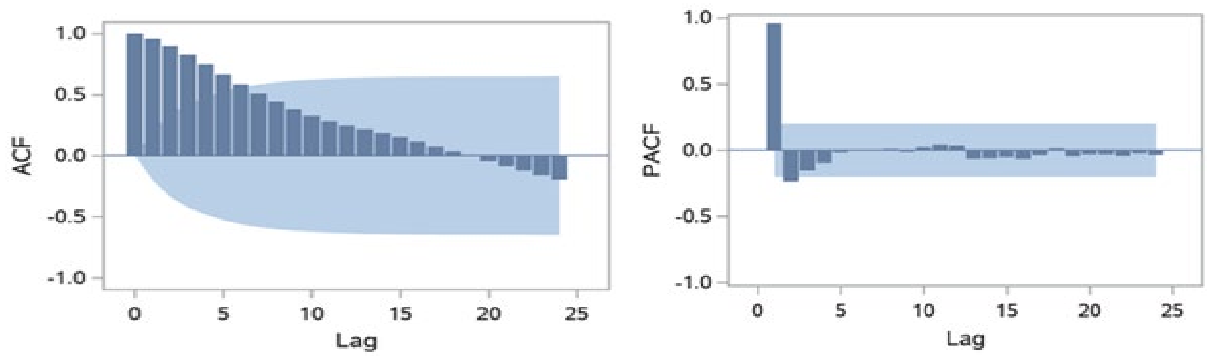
ACF. By looking at the autocorrelation plot we can assess whether there is a trend. A good indicator of a trend is a few autocorrelation coefficient spikes in row that go over the blue ribbon. We would expect strong autocorrelations at time lags of 0 and 1, because when there is no lag (lag0), the data point (say, January 2018) perfectly correlates with itself (with January 2018, which is represented by lag0). At lag1 we would still expect some autocorrelations present since February sales are related to January sales. However, if there is no trend present in the data, we would expect a sudden drop in autocorrelation coefficients for lag2, lag3 and so on. In this case we can see strong autocorrelations up to lag5. So, there is a trend present in the data, which we have already visually identified when we were examined the time series plot.

PACF: Sometimes the trend can mask seasonality, and but we can still look at the PACF (partial autocorrelation plot). We should investigate it for spikes exceeding the confidence interval level. We should be looking for the highest spike in a range of spikes. Once identified, that spike may indicate a seasonality interval, i.e. 6 months or 2 months etc. In this case it doesn't seem that there is any seasonality in the data present. But! It can also be masked by the trend. To continue the examination, we should have a look at this plot again after we difference the data, say, a lag1 for starters. The plot may look different then.

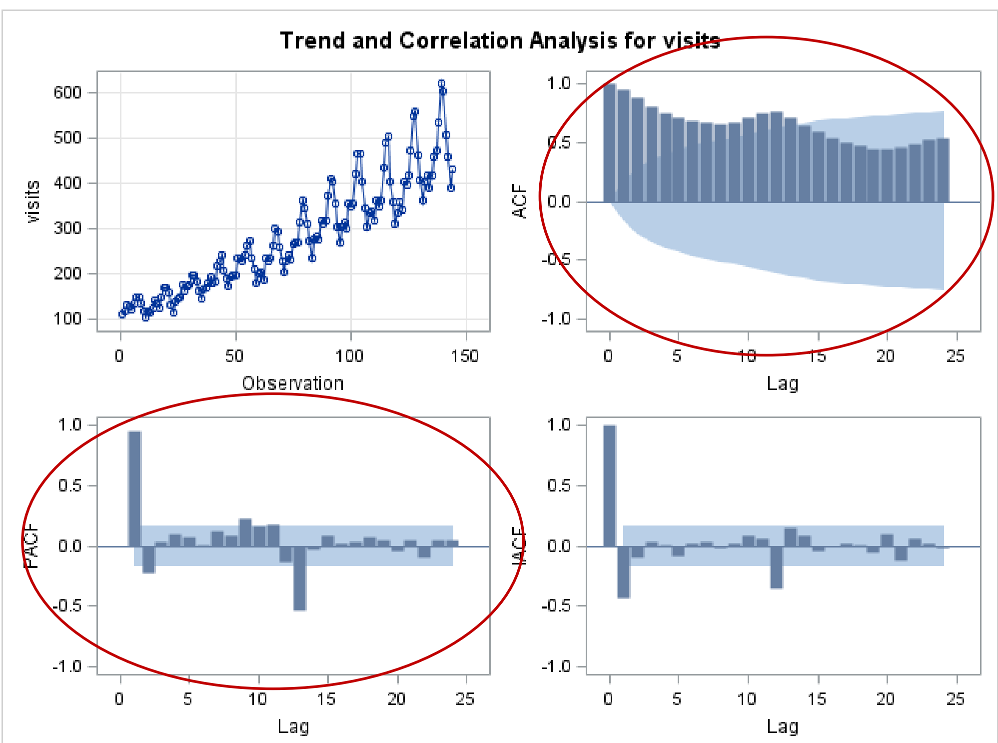
What else are ACF and PACF plots good for?

They are good for assessing which processes you are dealing with and at which order: AR or MA and is it AR(1) or MA(2) etc.? A way to find it out if to look at the ACF and PACF plots.

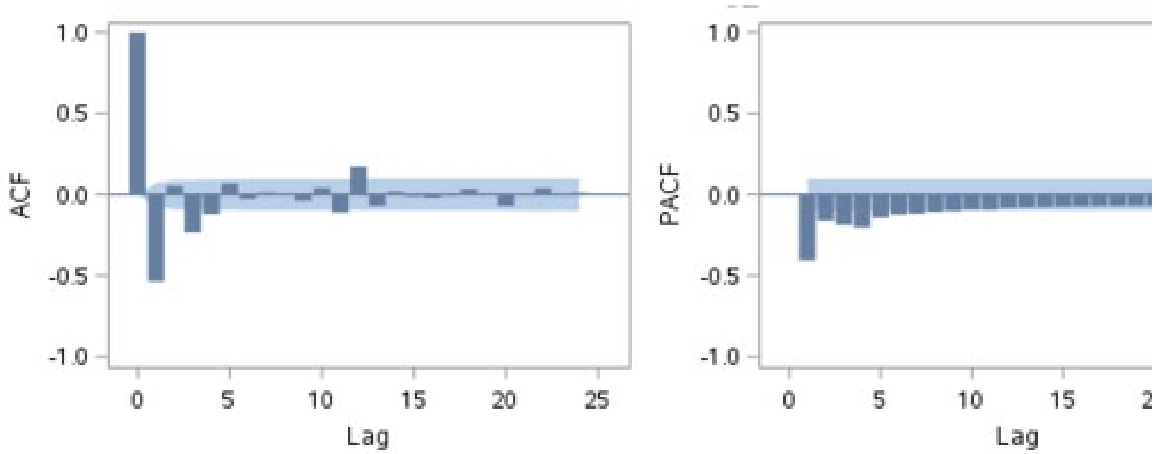
AR process. The plots presented below are typical for AR processes: Autocorrelations gradually fall to zero at the ACF (autocorrelation plot) and there are a few spikes at the Partial autocorrelation plot (PACF).



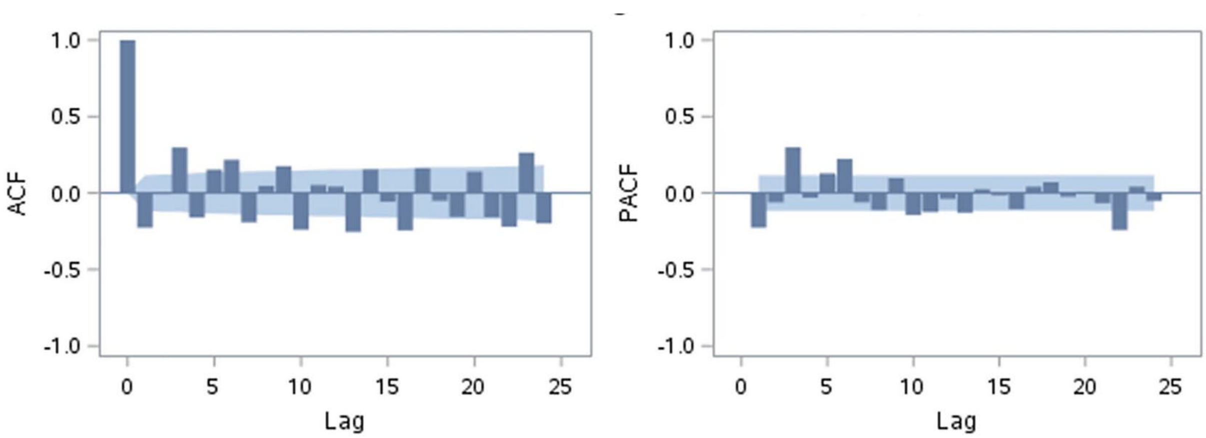
Or another example



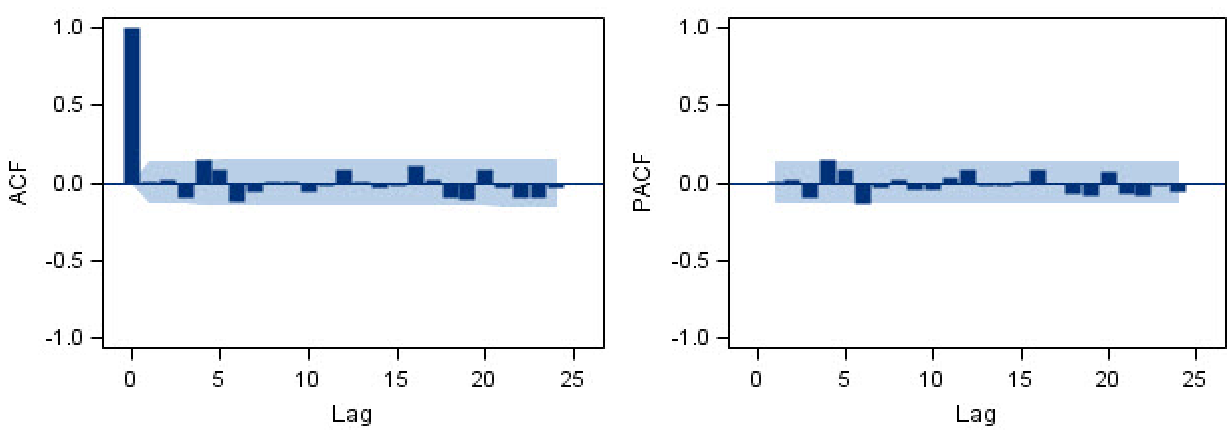
MA process. The plots behavior will be an inverse of what is observed for the AR processes: there are a few spikes at the Autocorrelation plot (ACF) and autocorrelations gradually fall to zero at the Partial autocorrelation plot (PACF).



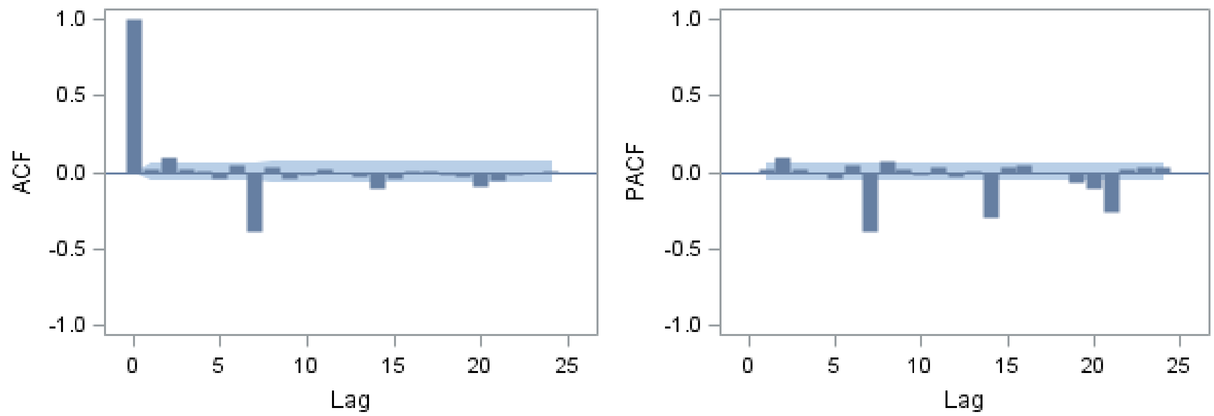
Both processes are present:



An example of stationary data:



An example of seasonality in the data:



How good is your model?

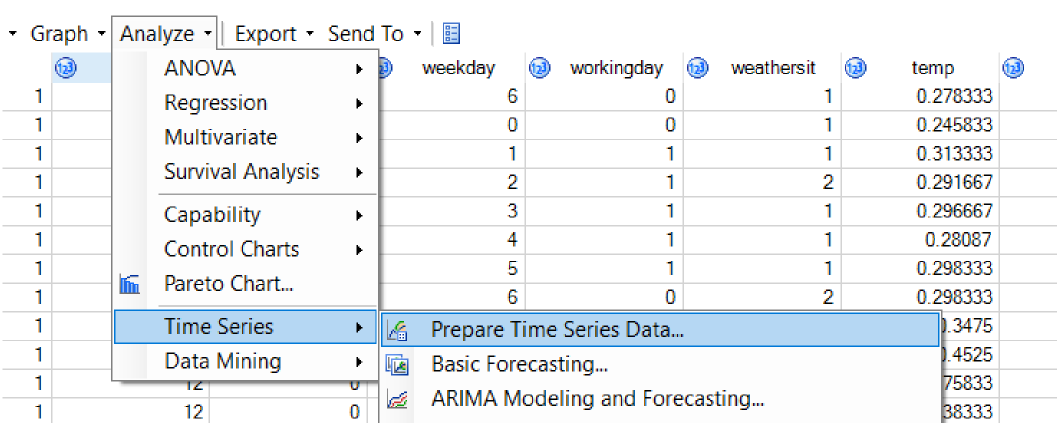
After you have done your forecasting, you should estimate the goodness of fit of your model and look at the residual autocorrelation and partial autocorrelation plots. They should help you identify whether your model is going to be robust and reliable. No autocorrelations or partial autocorrelations should be present.

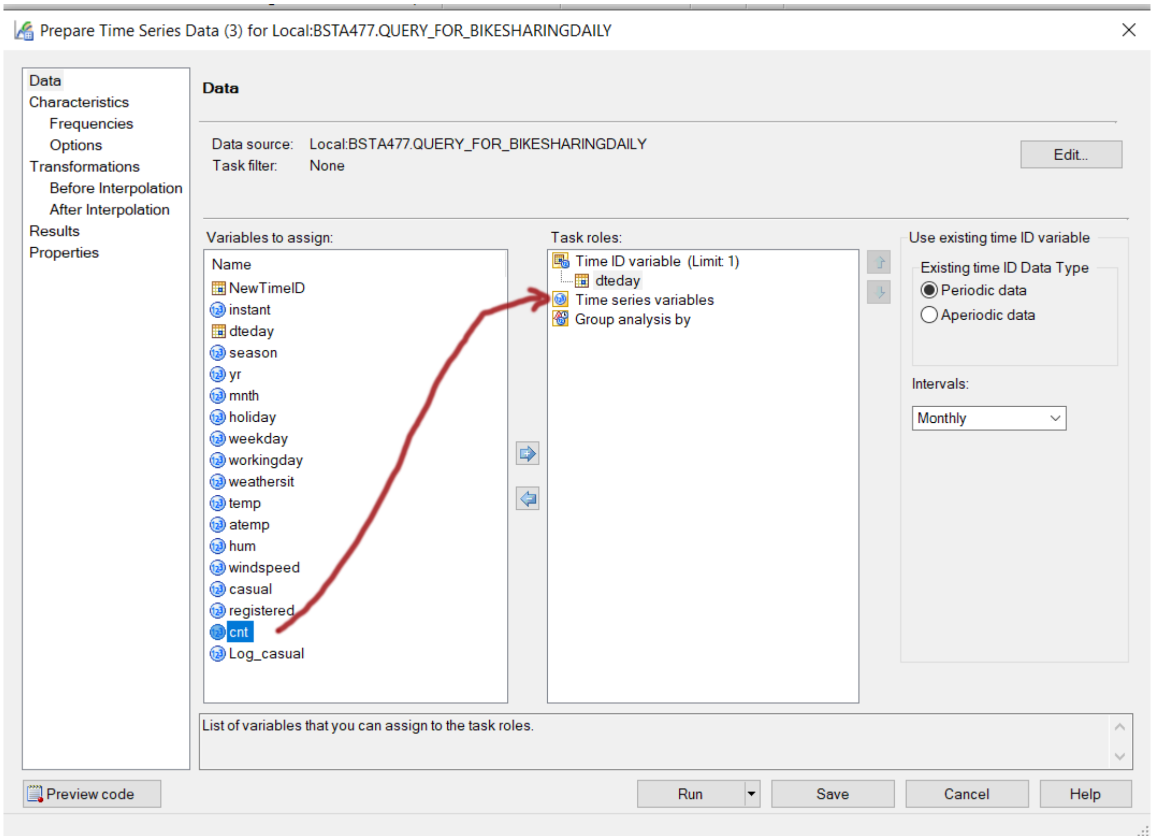
Also, check the Ljung-Box statistics (White Noise in SAS) to see whether any strong correlations are present.

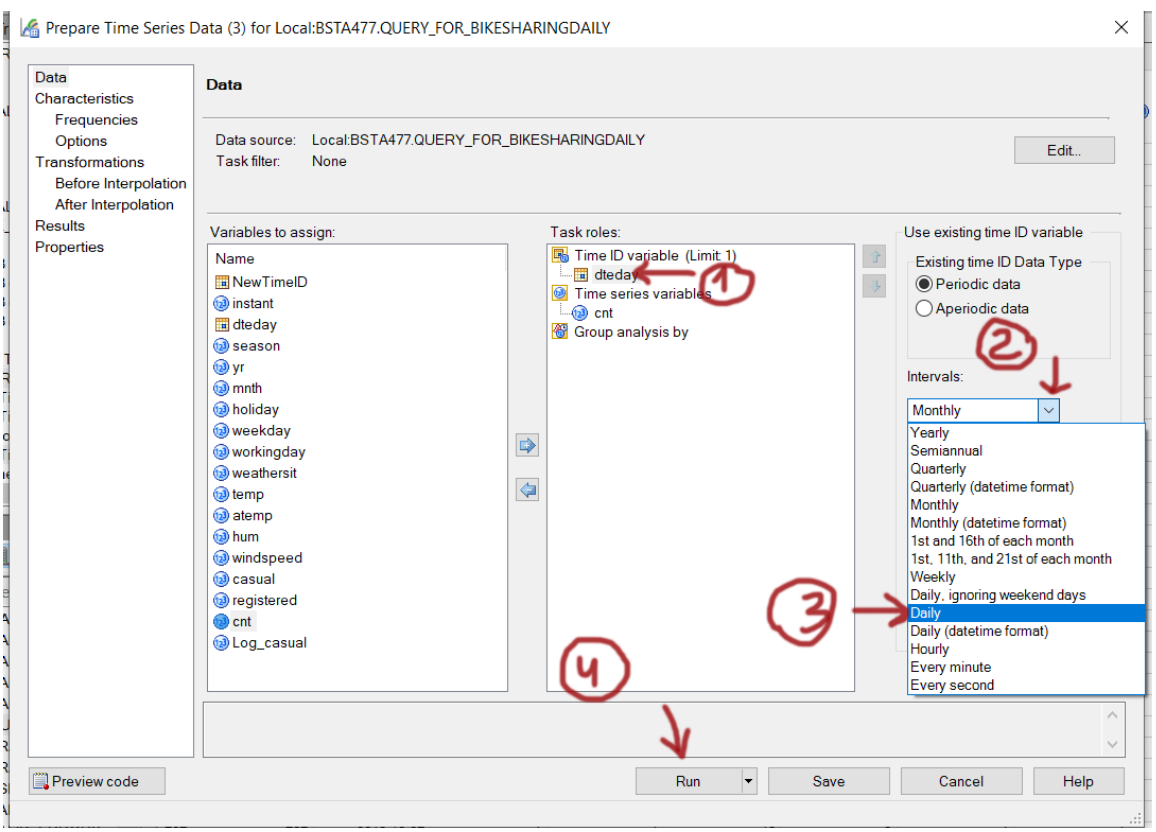
Prepare your time series data for analysis

Open the dataset → Analyze → Time series → Prepare Time Series Data → Drag cnt to the Time series variables → Click on the Time ID variable → Click on the arrow to the right under Intervals → Select Daily → Click Run to create a temporary dataset, which we are going to be using for the further analysis.

P.S. If you want to save it **permanently**!!!, before you click run, go to Results → Browse → Find the library you assigned before → Give it your name → Save.



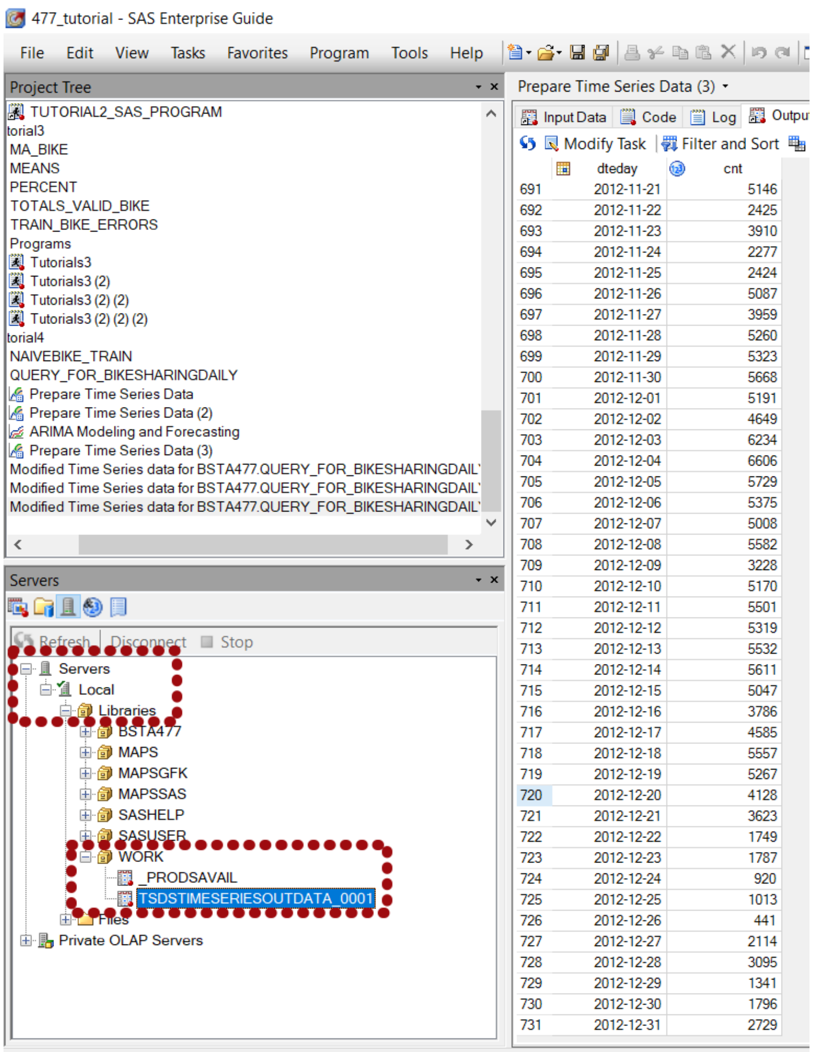

Prepare Time Series Data (3) for Local:BSTA477.QUERY_FOR_BIKESHARINGDAILY
Data
Characteristics
Frequencies
Options
Transformations
Before Interpolation
After Interpolation
Results
Properties
Data
Data source: Local:BSTA477.QUERY_FOR_BIKESHARINGDAILY
Task filter: None
Edit...
Variables to assign:
Name
NewTimeID
instant
dteday
season
yr
mnth
holiday
weekday
workingday
weathersit
temp
atemp
hum
windspeed
casual
registered
cnt
Log_casual
Task roles:
Time ID variable (Limit: 1)
dteday
Time series variables
Group analysis by
Use existing time ID variable
Existing time ID Data Type
Periodic data
Aperiodic data
Intervals:
Monthly
List of variables that you can assign to the task roles.
Preview code
Run
Save
Cancel
Help


Prepare Time Series Data (3) for Local:BSTA477.QUERY_FOR_BIKESHARINGDAILY
Data
Characteristics
Frequencies
Options
Transformations
Before Interpolation
After Interpolation
Results
Properties
Data
Data source: Local:BSTA477.QUERY_FOR_BIKESHARINGDAILY
Task filter: None
Edit...
Variables to assign:
Name
NewTimeID
instant
dteday
season
yr
mnth
holiday
weekday
workingday
weathersit
temp
atemp
hum
windspeed
casual
registered
cnt
Log_casual
Task roles:
Time ID variable (Limit: 1)
dteday
Time series variables
cnt
Group analysis by
1
Use existing time ID variable
Existing time ID Data Type
Periodic data
Aperiodic data
2
Intervals:
Monthly
Yearly
Semiannual
Quarterly
Quarterly (datetime format)
Monthly
Monthly (datetime format)
1st and 16th of each month
1st, 11th, and 21st of each month
Weekly
Daily, ignoring weekend days
Daily
Daily (datetime format)
Hourly
Every minute
Every second
3
4
Preview code
Run
Save
Cancel
Help

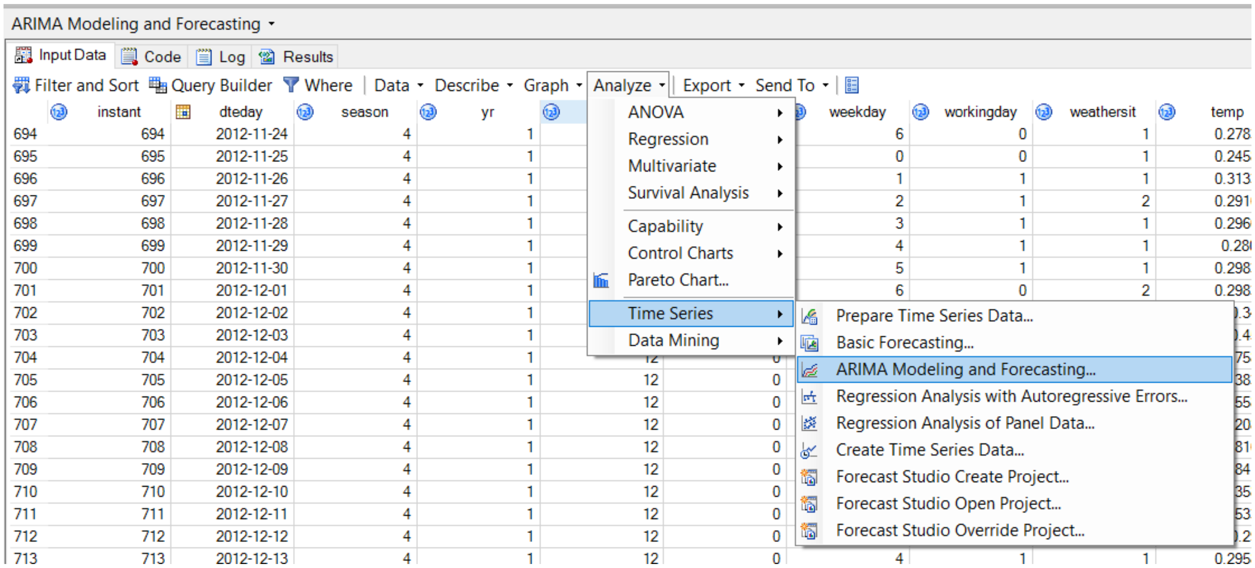
We are going to be using this dataset for further analysis!

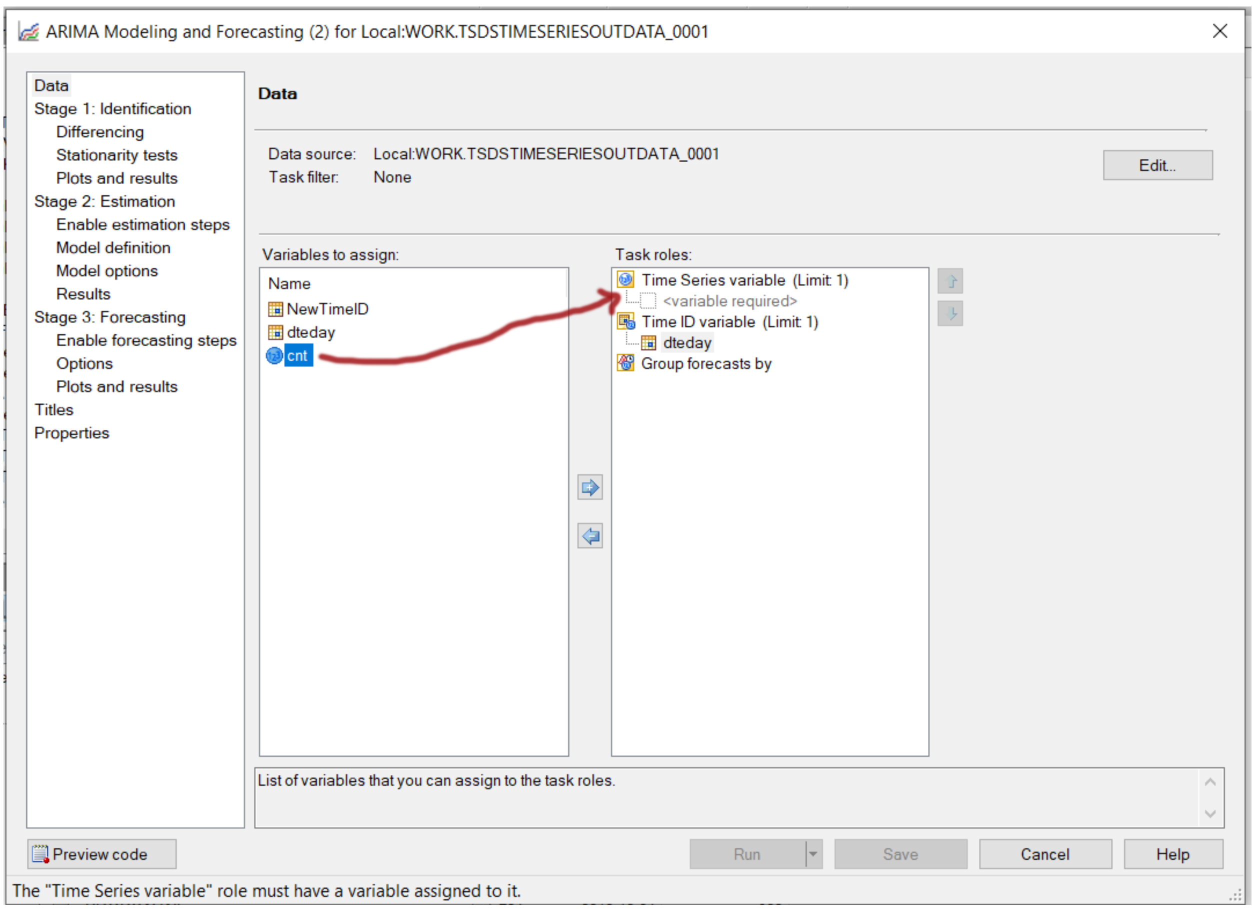
ARIMA: Step 1 – examine your data

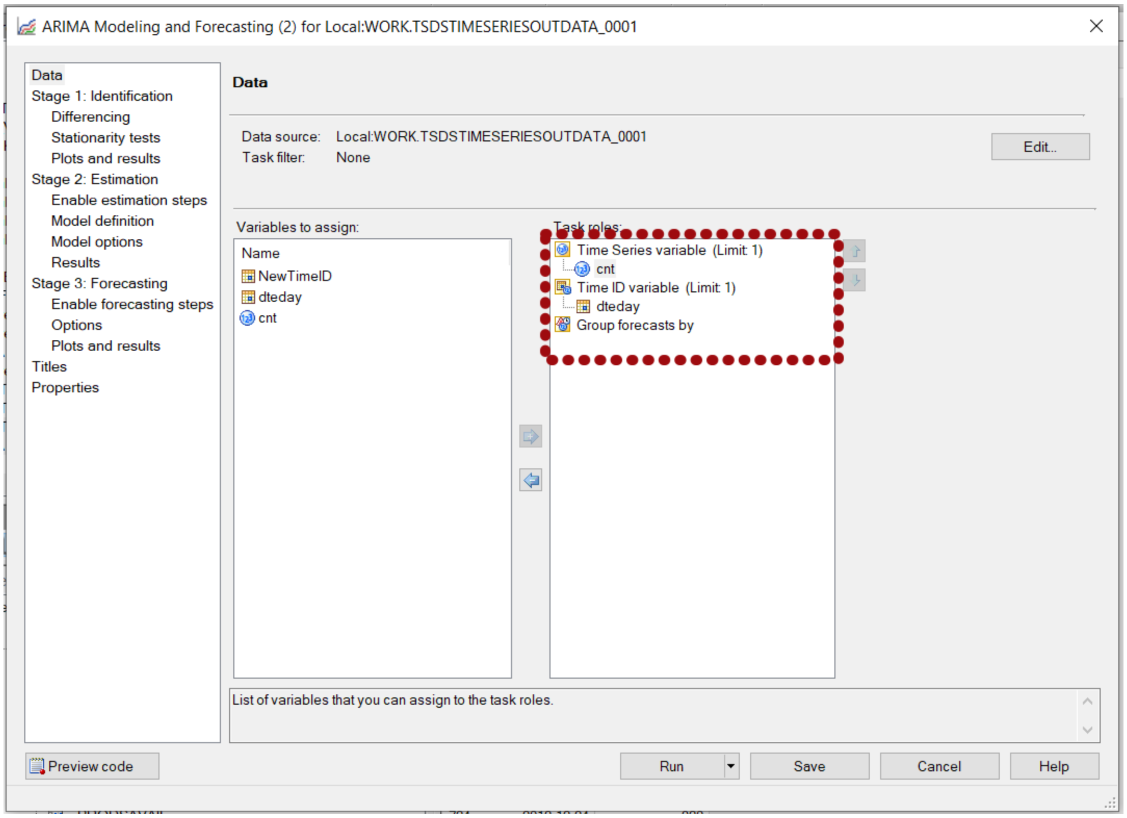
We are going to work with the dataset that we have just prepared for analysis in the previous step. It should be open. If not, you can easily locate it in the work library:

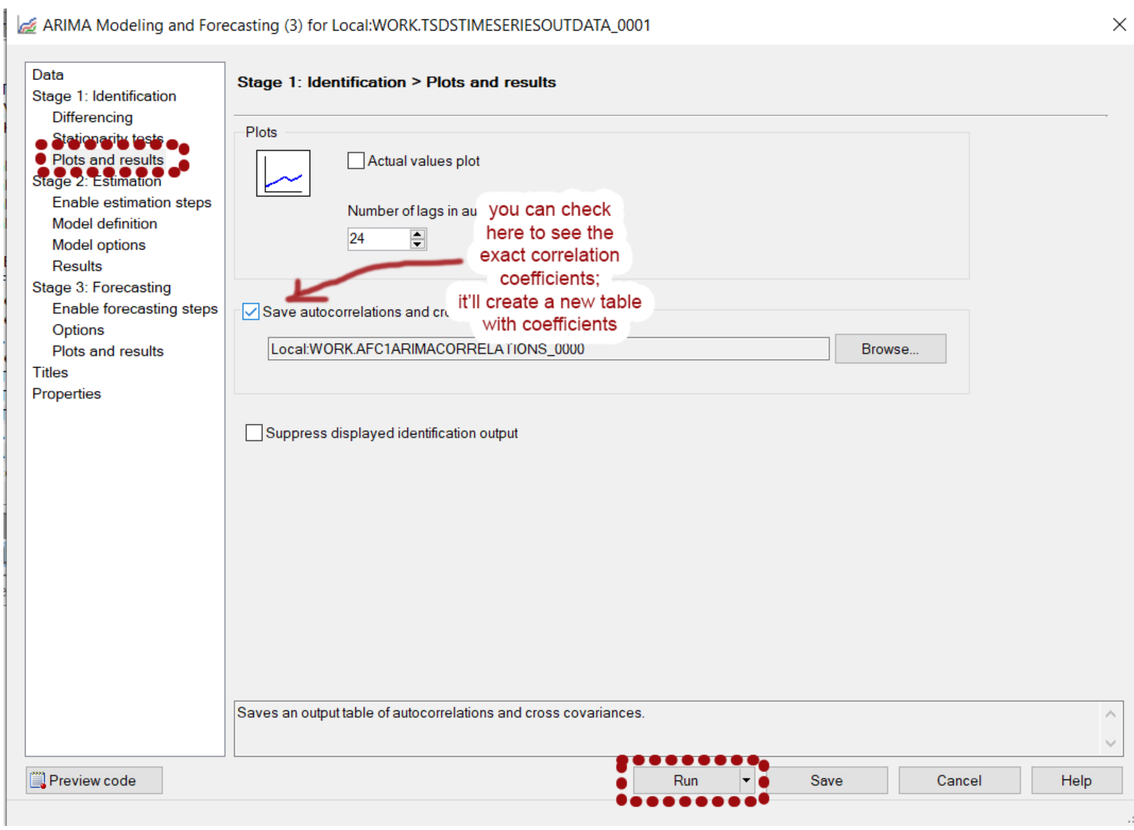


Open the dataset → Analyze → Time series → ARIMA Modelling and Forecasting → Data: Drag cnt to the right, under Time Series variable → Stage1: Identification → Plots and results → if you want to: Save acutoccorelations and cross covariances. It will give you a table with exact autocorrelations → Run → Study the autocorrelation and partial autocorrelation plots to decide if differencing is needed, whether there are AR or/and MA processes are present as well as what order.


ARIMA Modeling and Forecasting
Input Data
Code
Log
Results
Filter and Sort
Query Builder
Where
Data
Describe
Graph
Analyze
Export
Send To
ANOVA
Regression
Multivariate
Survival Analysis
Capability
Control Charts
Pareto Chart...
Time Series
Data Mining
Prepare Time Series Data...
Basic Forecasting...
ARIMA Modeling and Forecasting...
Regression Analysis with Autoregressive Errors...
Regression Analysis of Panel Data...
Create Time Series Data...
Forecast Studio Create Project...
Forecast Studio Open Project...
Forecast Studio Override Project...


ARIMA Modeling and Forecasting (2) for Local:WORK.TSDSTIMESERIESOUTDATA_0001
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Data source: Local:WORK.TSDSTIMESERIESOUTDATA_0001
Task filter: None
Edit...
Variables to assign:
Name
NewTimeID
dteday
cnt
Task roles:
Time Series variable (Limit: 1)
<variable required>
Time ID variable (Limit: 1)
dteday
Group forecasts by
List of variables that you can assign to the task roles.
Preview code
Run
Save
Cancel
Help
The "Time Series variable" role must have a variable assigned to it.


ARIMA Modeling and Forecasting (2) for Local:WORK.TSDSTIMESERIESOUTDATA_0001
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Data
Data source: Local:WORK.TSDSTIMESERIESOUTDATA_0001
Task filter: None
Edit...
Variables to assign:
Name
NewTimeID
dteday
cnt
Task roles:
Time Series variable (Limit: 1)
cnt
Time ID variable (Limit: 1)
dteday
Group forecasts by
List of variables that you can assign to the task roles.
Preview code
Run
Save
Cancel
Help


ARIMA Modeling and Forecasting (3) for Local:WORK.TSDSTIMESERIESOUTDATA_0001
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Stage 1: Identification > Plots and results
Plots
Actual values plot
Number of lags in au
24
you can check here to see the exact correlation coefficients;
it'll create a new table with coefficients
Save autocorrelations and cr
Local:WORK.AFC1ARIMACORRELATIONS_0000
Browse...
Suppress displayed identification output
Saves an output table of autocorrelations and cross covariances.
Preview code
Run
Save
Cancel
Help

ARIMA Modeling and Forecasting

Results

The ARIMA Procedure

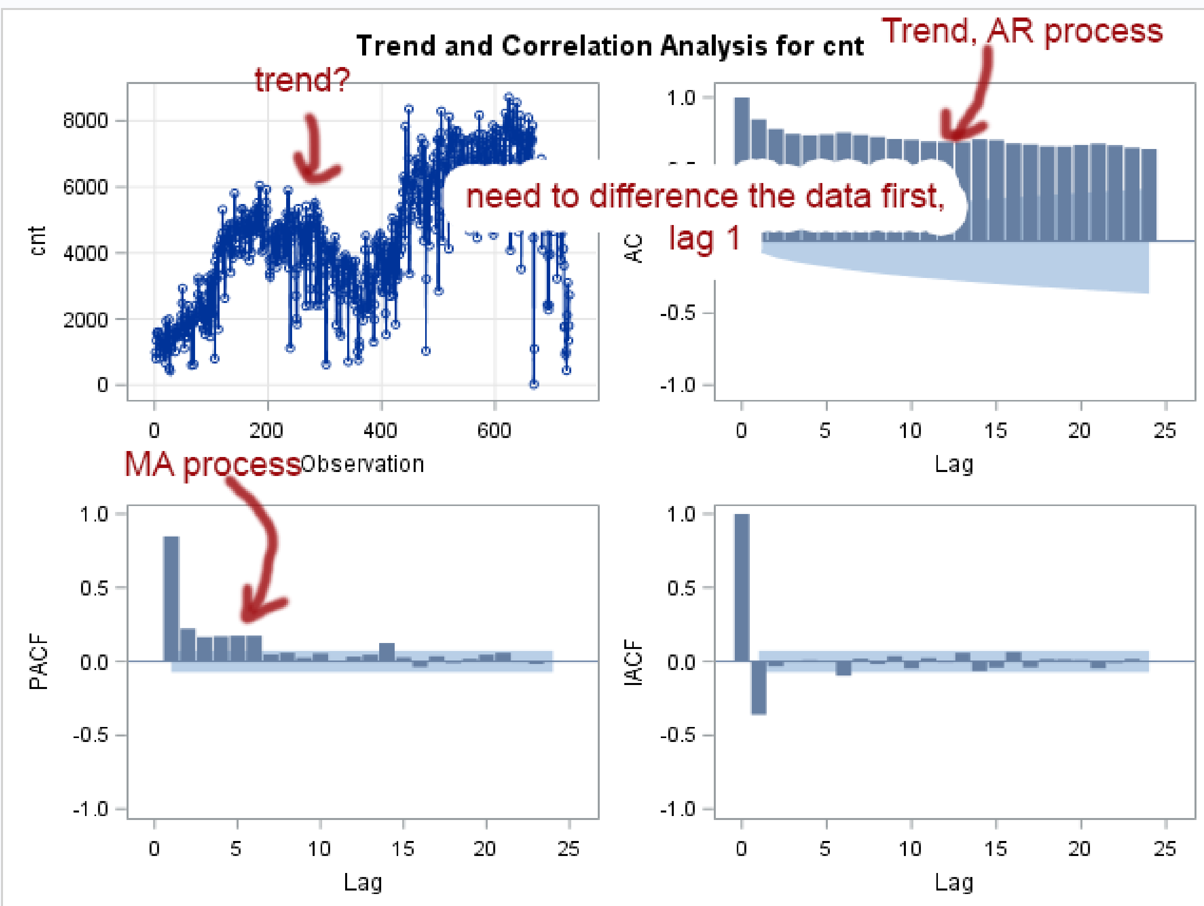
Name of Variable = cnt	
Mean of Working Series	4504.349
Standard Deviation	1935.886
Number of Observations	731

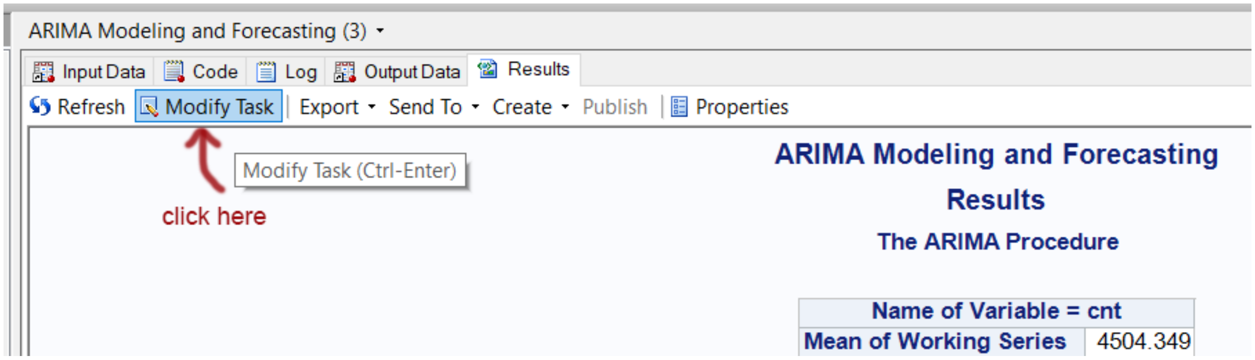
Ljung-Box

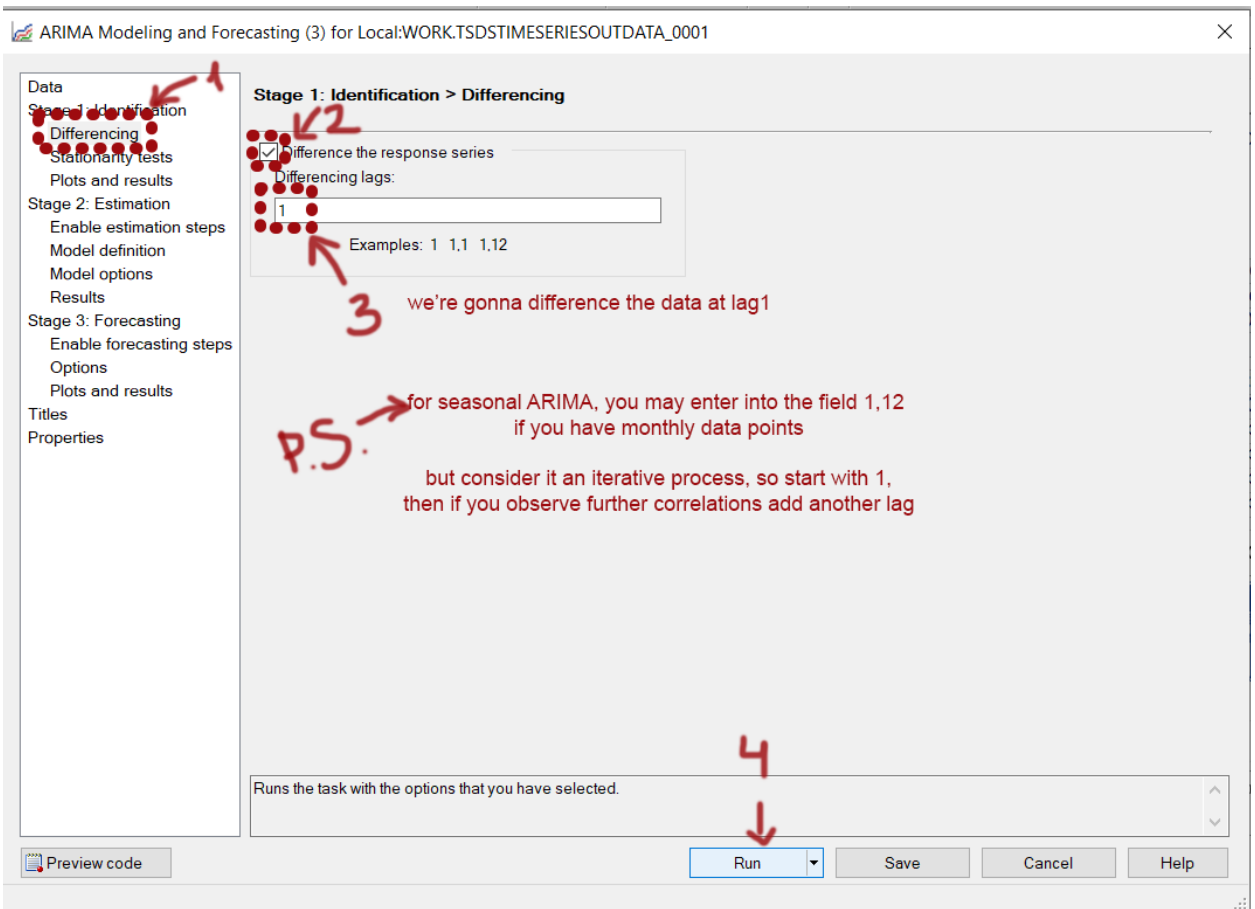
the autocorrelation coefficients, at alpha=0.05 statistically significant

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
				1	2	3	4	5	6
6	2608.34	6	<.0001	0.846	0.779	0.746	0.737	0.742	0.755
12	4855.42	12	<.0001	0.739	0.726	0.709	0.704	0.692	0.689
18	6966.52	18	<.0001	0.689	0.707	0.702	0.680	0.673	0.661
24	8948.32	24	<.0001	0.659	0.670	0.678	0.667	0.651	0.642




ARIMA Modeling and Forecasting (3)
Input Data
Code
Log
Output Data
Results
Refresh
Modify Task
Export
Send To
Create
Publish
Properties
Modify Task (Ctrl-Enter)
click here
ARIMA Modeling and Forecasting
Results
The ARIMA Procedure
Name of Variable = cnt
Mean of Working Series
4504.349


ARIMA Modeling and Forecasting (3) for Local:WORK.TSDSTIMESERIESOUTDATA_0001
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
1
2
3
4
Stage 1: Identification > Differencing
Difference the response series
Differencing lags:
1
Examples: 1 1,1 1,12
we're gonna difference the data at lag1
P.S.
for seasonal ARIMA, you may enter into the field 1,12
if you have monthly data points
but consider it an iterative process, so start with 1,
then if you observe further correlations add another lag
Runs the task with the options that you have selected.
Preview code
Run
Save
Cancel
Help

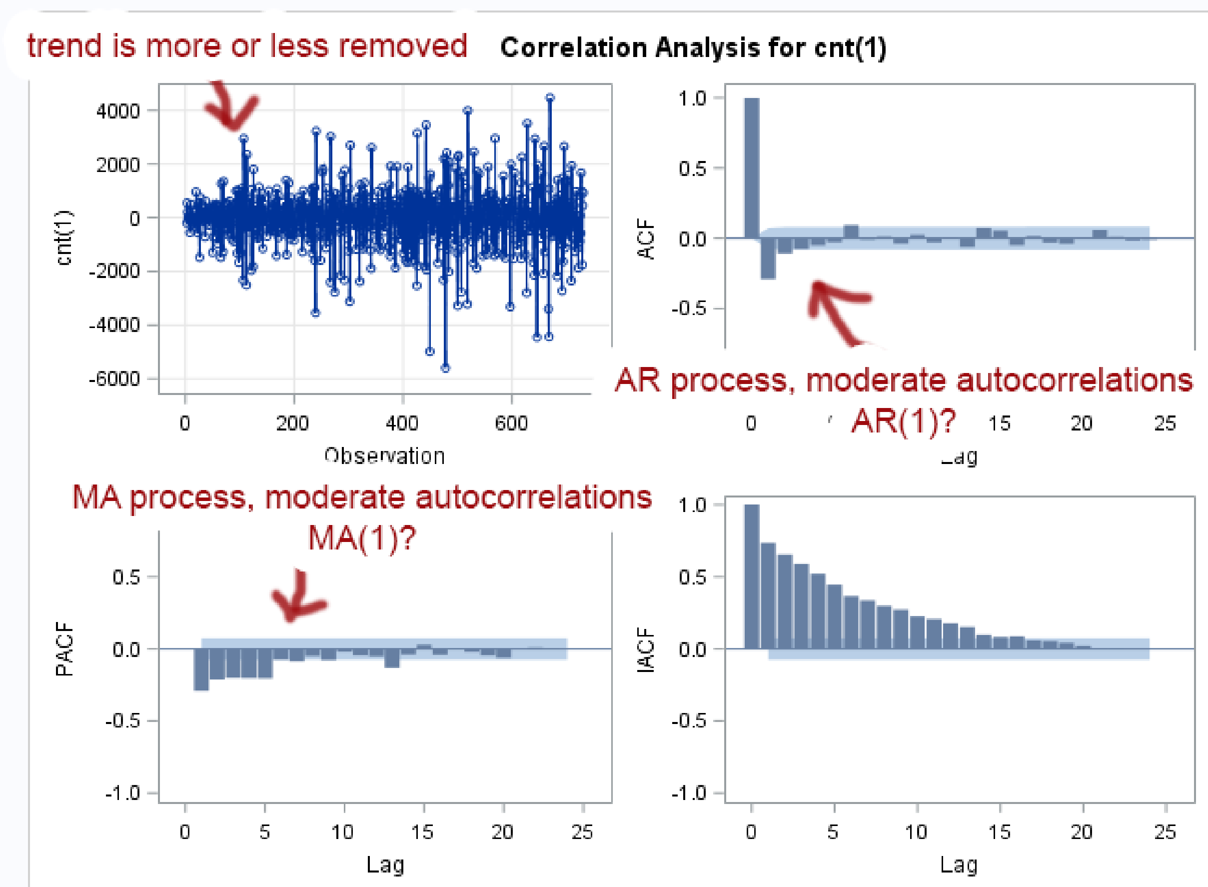
Results

The ARIMA Procedure

one datapoint removed because we used lag1 to perform differencing

Name of Variable = cnt	
Period(s) of Differencing	
Mean of Working Series	
Standard Deviation	1064.525
Number of Observations	730
Observation(s) eliminated by differencing	1

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	83.90	6	<.0001	-0.291	-0.109	-0.076	-0.051	-0.027	0.094
12	86.22	12	<.0001	-0.010	0.013	-0.037	0.026	-0.028	-0.004
18	98.28	18	<.0001	-0.062	0.075	0.055	-0.050	0.017	-0.030
24	102.56	24	<.0001	-0.039	0.005	0.061	0.011	-0.015	-0.010

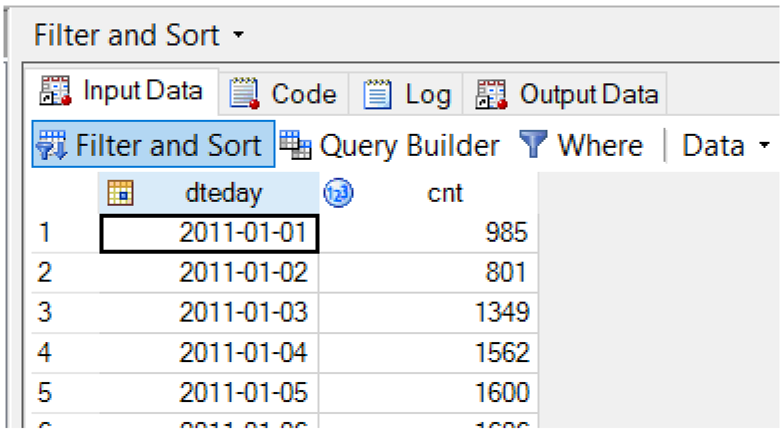


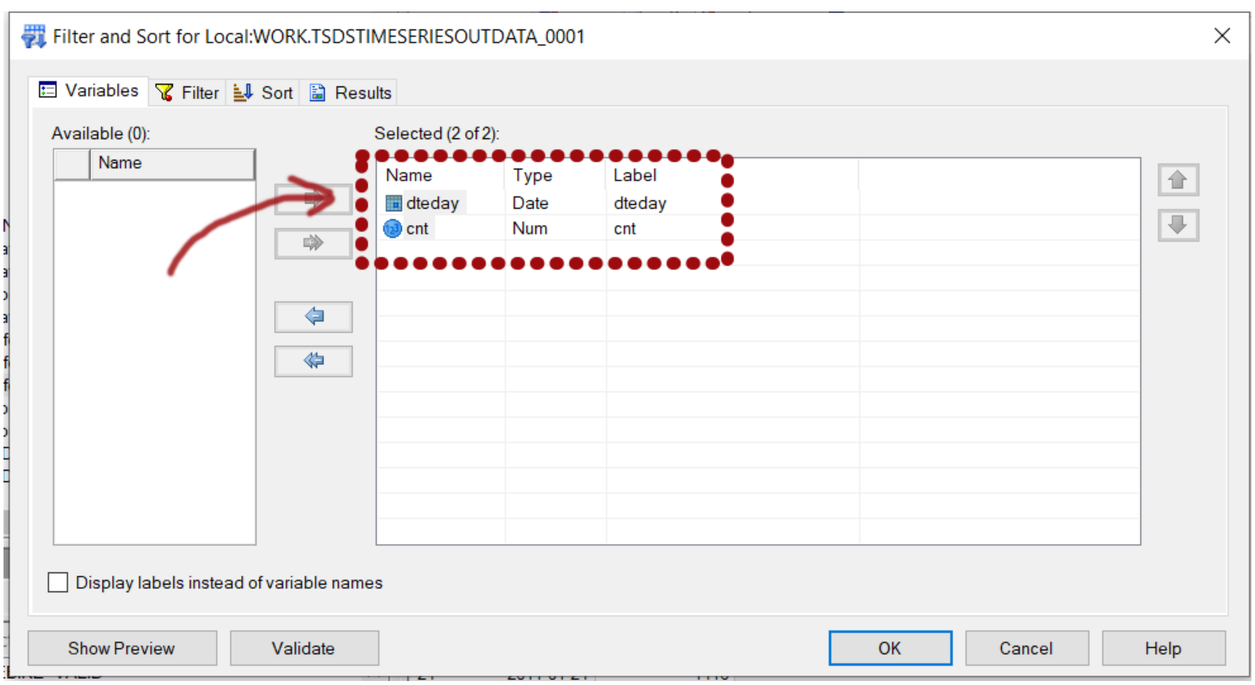
ARIMA: Step 2 – Build your Model

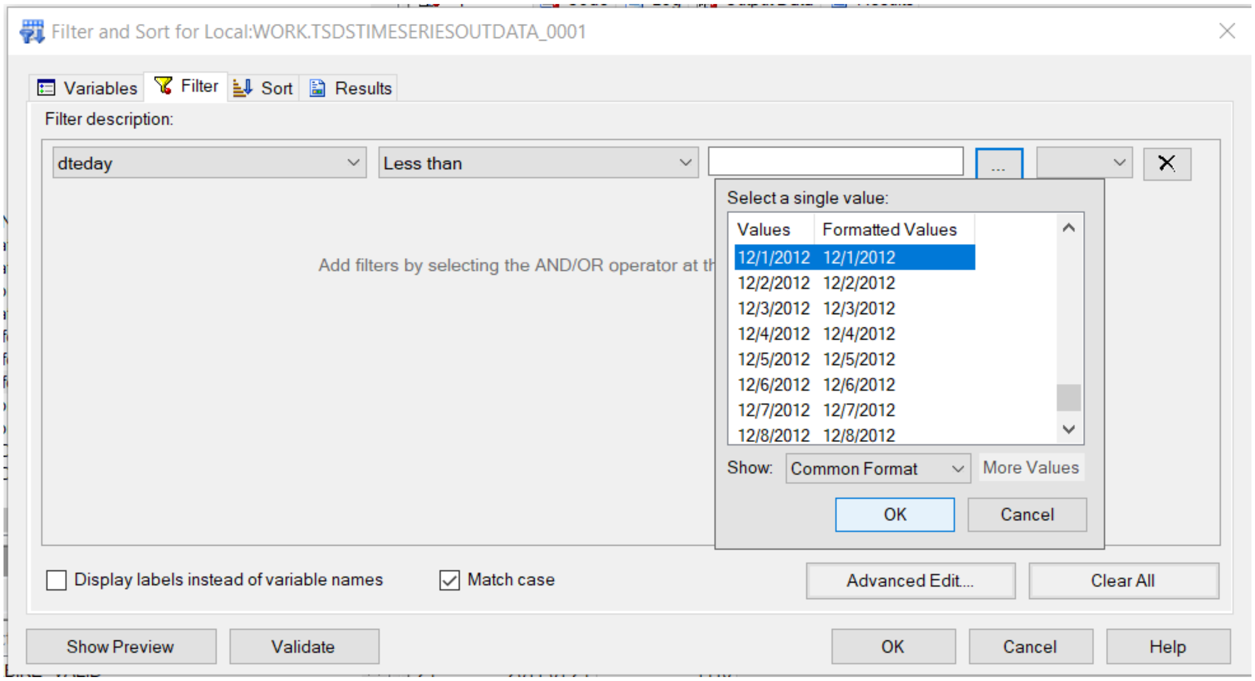
A. Create a Training Dataset

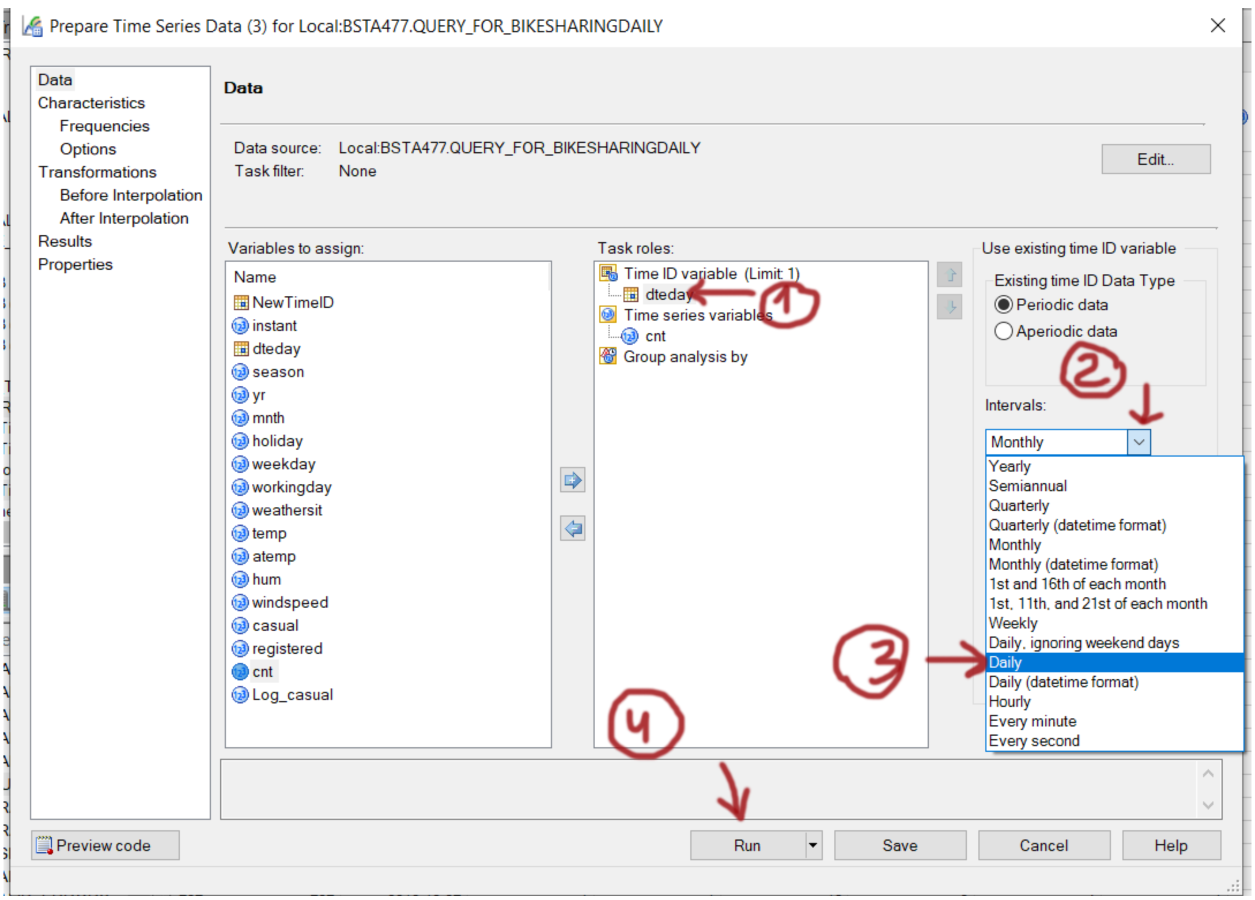
Discussed in Tutorial3: Filter and Sort, condition: dteday; less then 12/01/2012

Open the dataset → Filter and Sort → Variables: Drag the variables you will need for the forecast to the right (date and the variable you want to forecast) → Filter: Set the filtering values (dteday; less then; December 1, 2012) → Results: Permanently save the dataset: Replace the WORK library with the one you ran at the beginning of the session – BSTA477 → OK



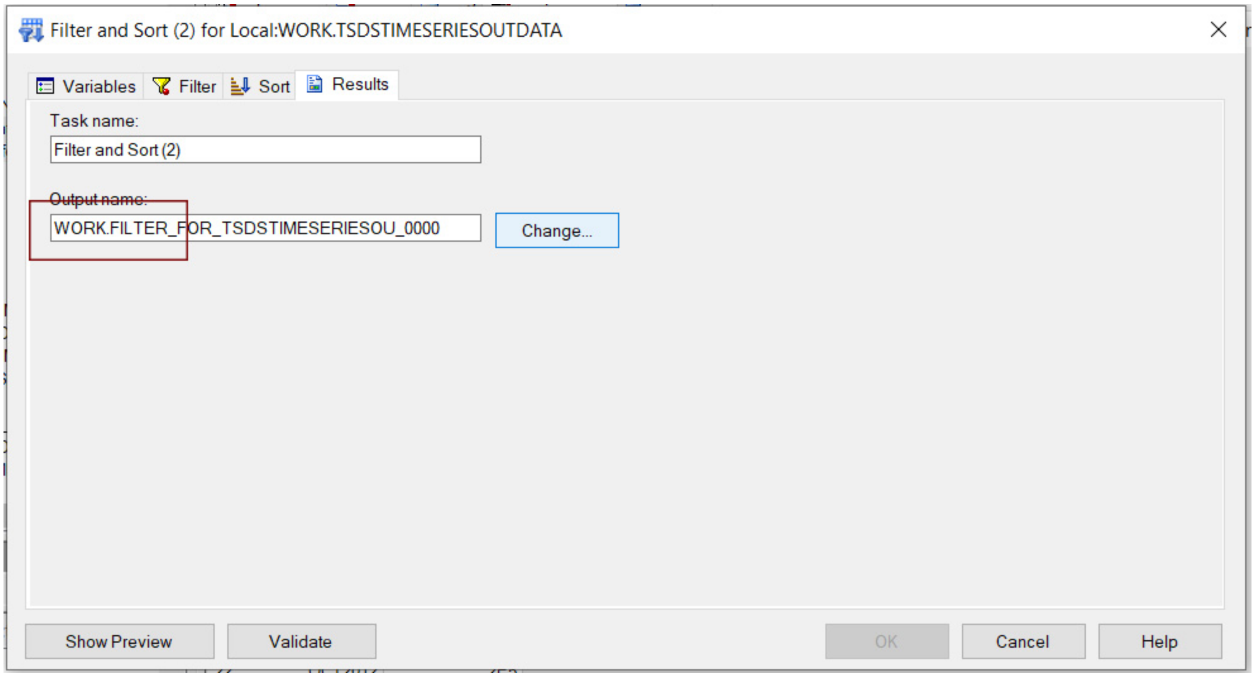


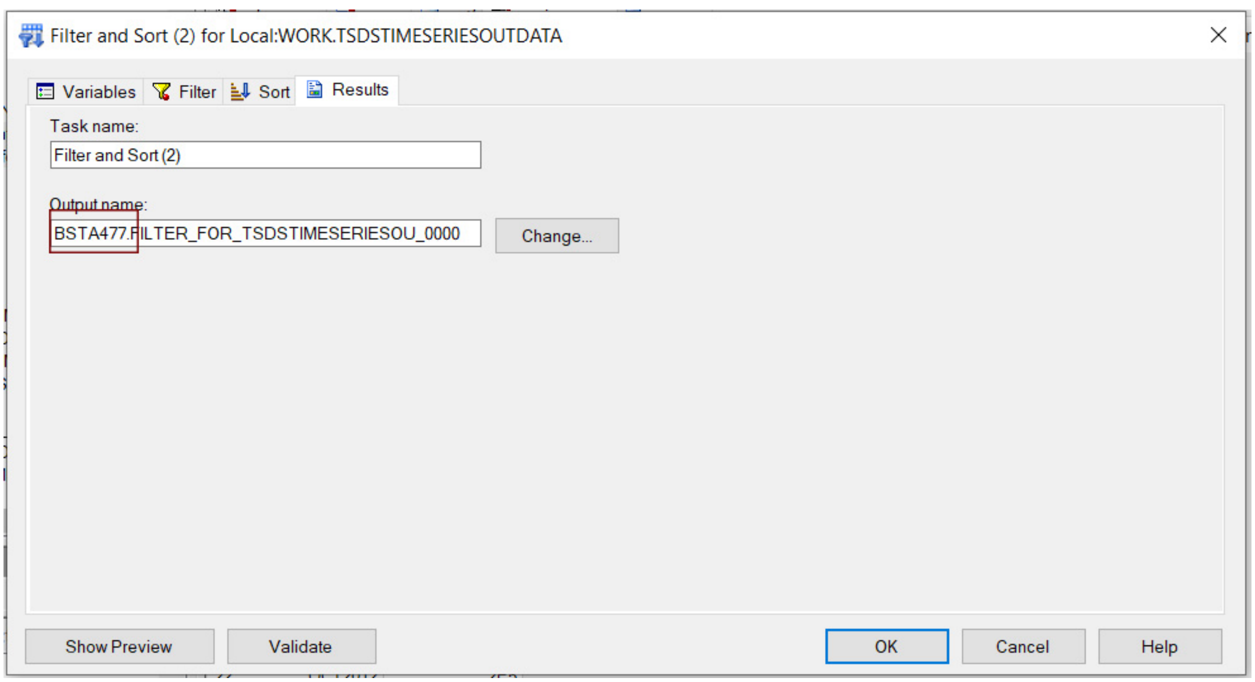




CREATES A TEMPORARY DATASET!

To create a permanent dataset, before you hit "RUN", go to Results and change the library name for bsta477 or click on change, find your library and give your dataset a name and store it there.





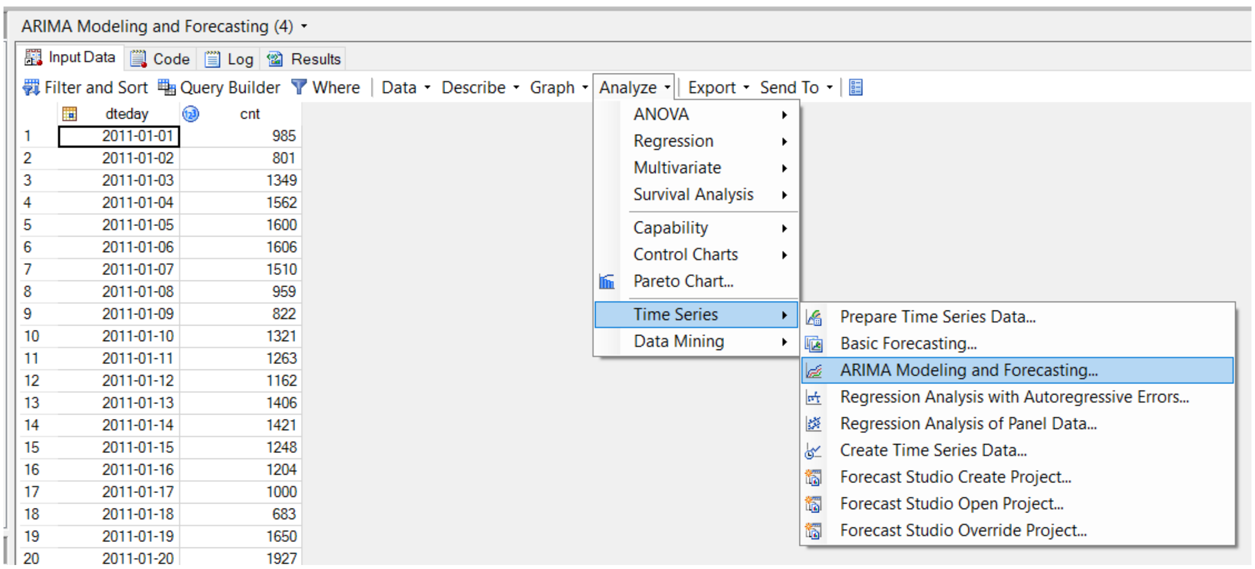
B. Difference the data, estimate AR and MA orders

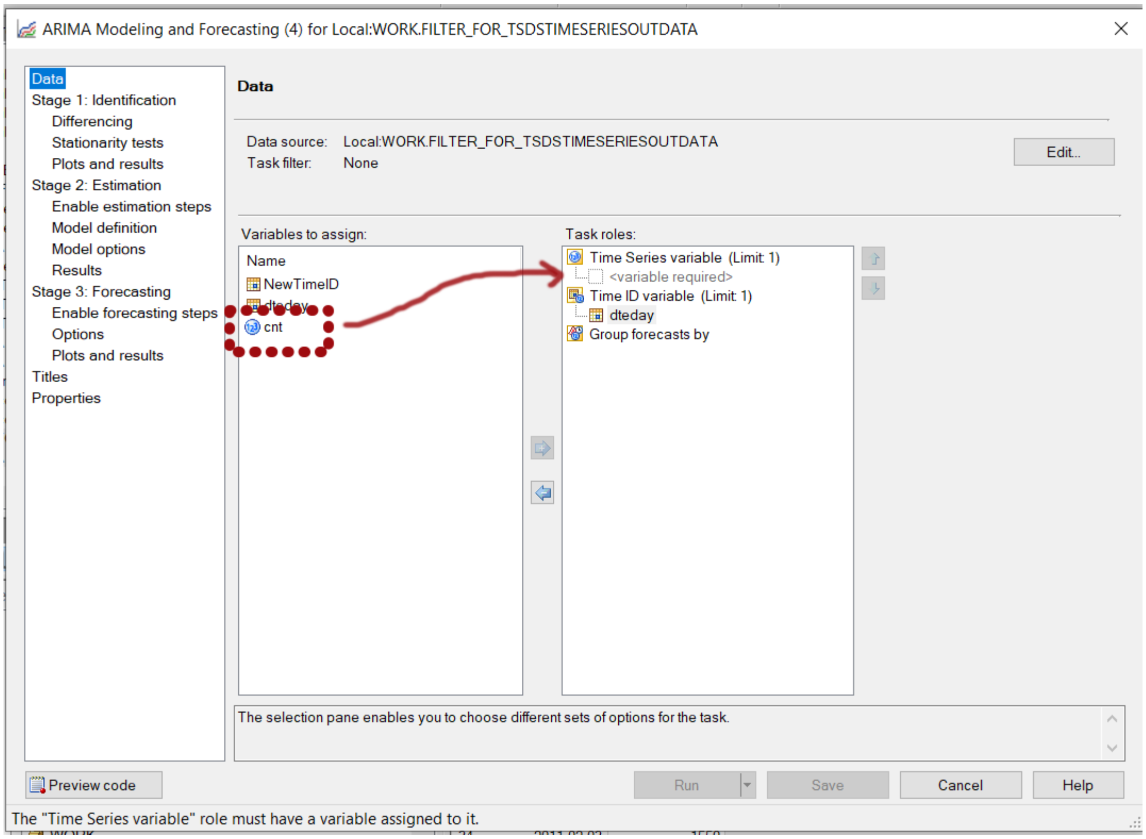
We are going to use the newly created **TRAIN dataset** with accumulated values of total member and non-member rides on a daily basis. As we can see from the table above, we now only have 700 points of data.

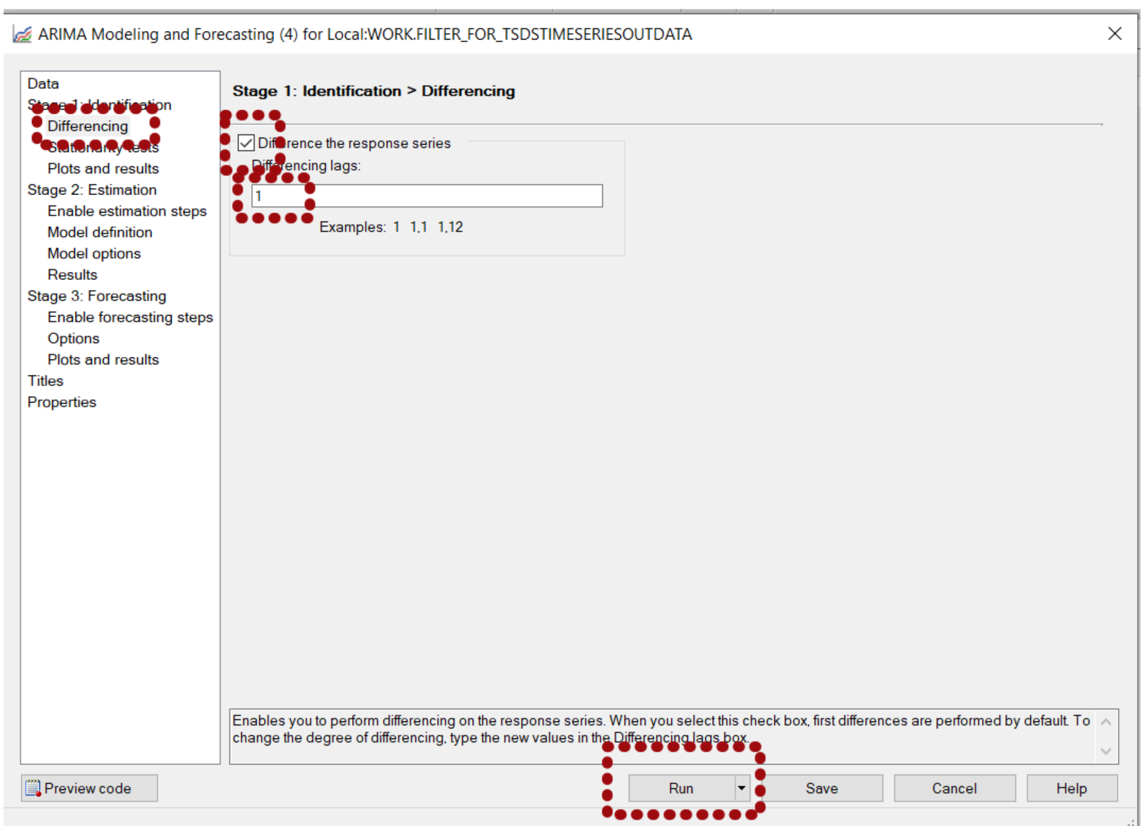
Difference the data.

Repeat Step 1 on the Train Dataset.

We have discussed already that we would need to difference the data due to the presence of the trend: Open the file you want to use for forecasting → Analyze → Time Series → ARIMA Modelling and Forecasting → Data: Drag the cnt to the right to the Time Series Variable → Stage 1: Identification → Differencing → Difference the response series → Leave at default → Run




ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Data
Data source: Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Task filter: None
Edit...
Variables to assign:
Name
NewTimeID
dteday
cnt
Task roles:
Time Series variable (Limit: 1)
<variable required>
Time ID variable (Limit: 1)
dteday
Group forecasts by
The selection pane enables you to choose different sets of options for the task.
Preview code
Run
Save
Cancel
Help
The "Time Series variable" role must have a variable assigned to it.

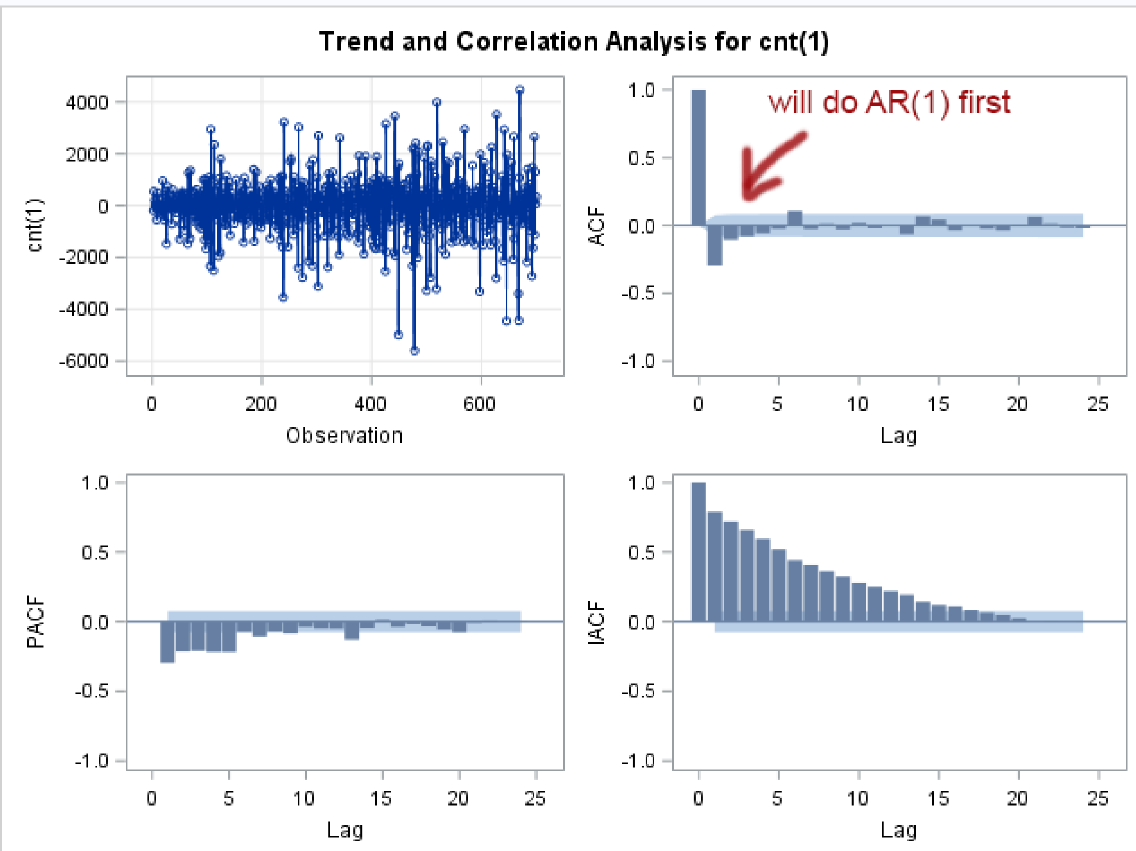

ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Stage 1: Identification > Differencing
Difference the response series
Differencing lags:
1
Examples: 1 1,1 1,12
Enables you to perform differencing on the response series. When you select this check box, first differences are performed by default. To change the degree of differencing, type the new values in the Differencing lags box.
Preview code
Run
Save
Cancel
Help

Results

The ARIMA Procedure

Name of Variable = cnt	
Period(s) of Differencing	1
Mean of Working Series	6.699571
Standard Deviation	1066.822
Number of Observations	699
Observation(s) eliminated by differencing	1

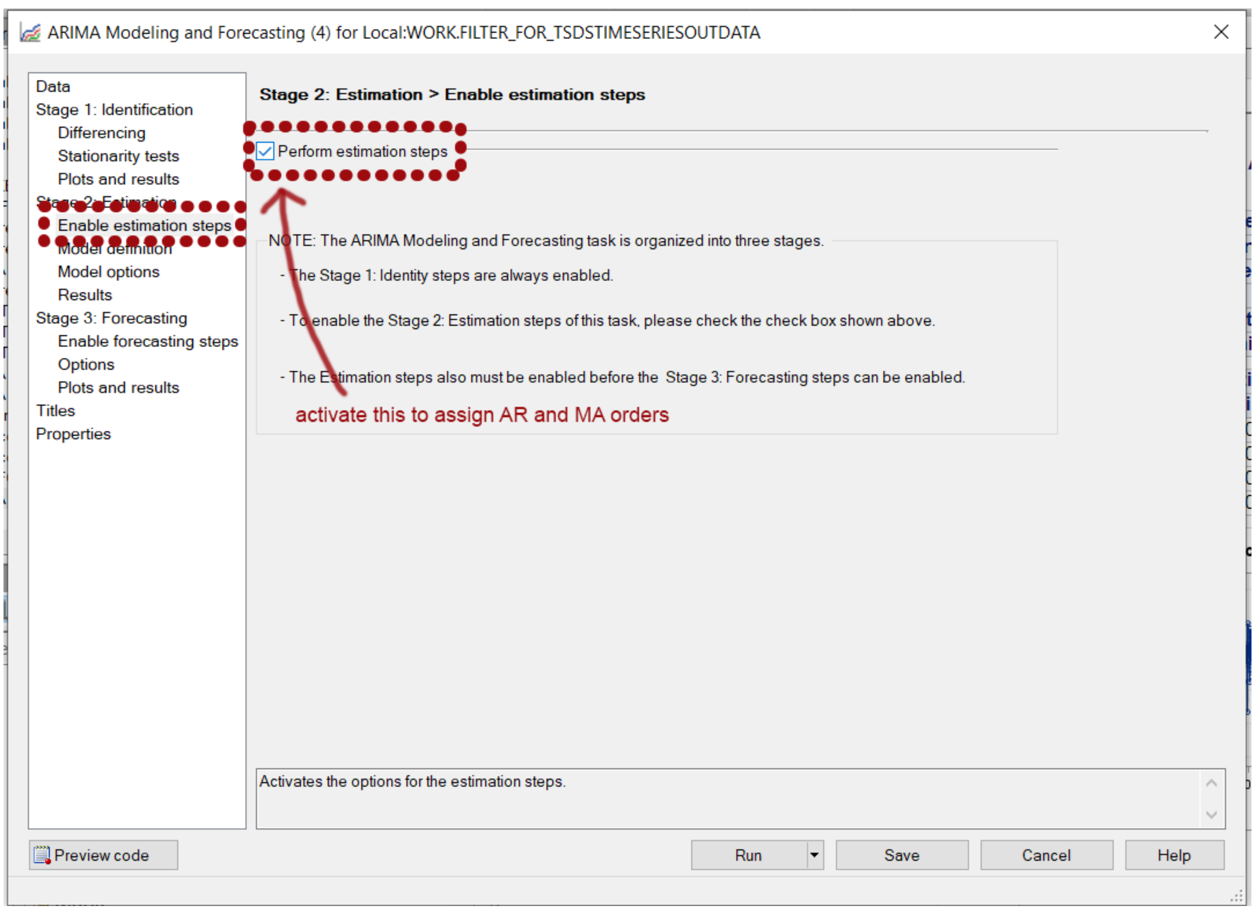
Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	84.85	6	<.0001	-0.296	-0.107	-0.080	-0.057	-0.024	0.108
12	86.49	12	<.0001	-0.026	0.012	-0.029	0.018	-0.017	0.002
18	95.37	18	<.0001	-0.064	0.067	0.045	-0.035	0.000	-0.023
24	99.73	24	<.0001	-0.036	0.007	0.064	0.013	-0.014	-0.018

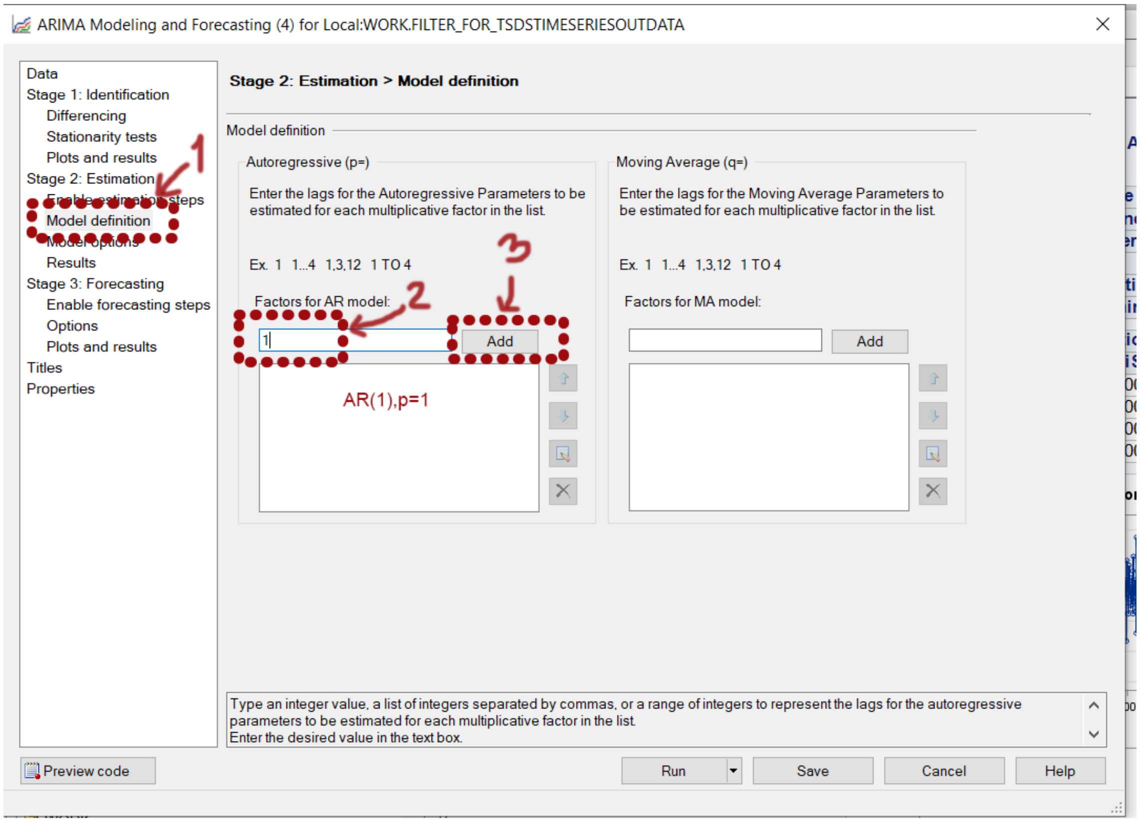


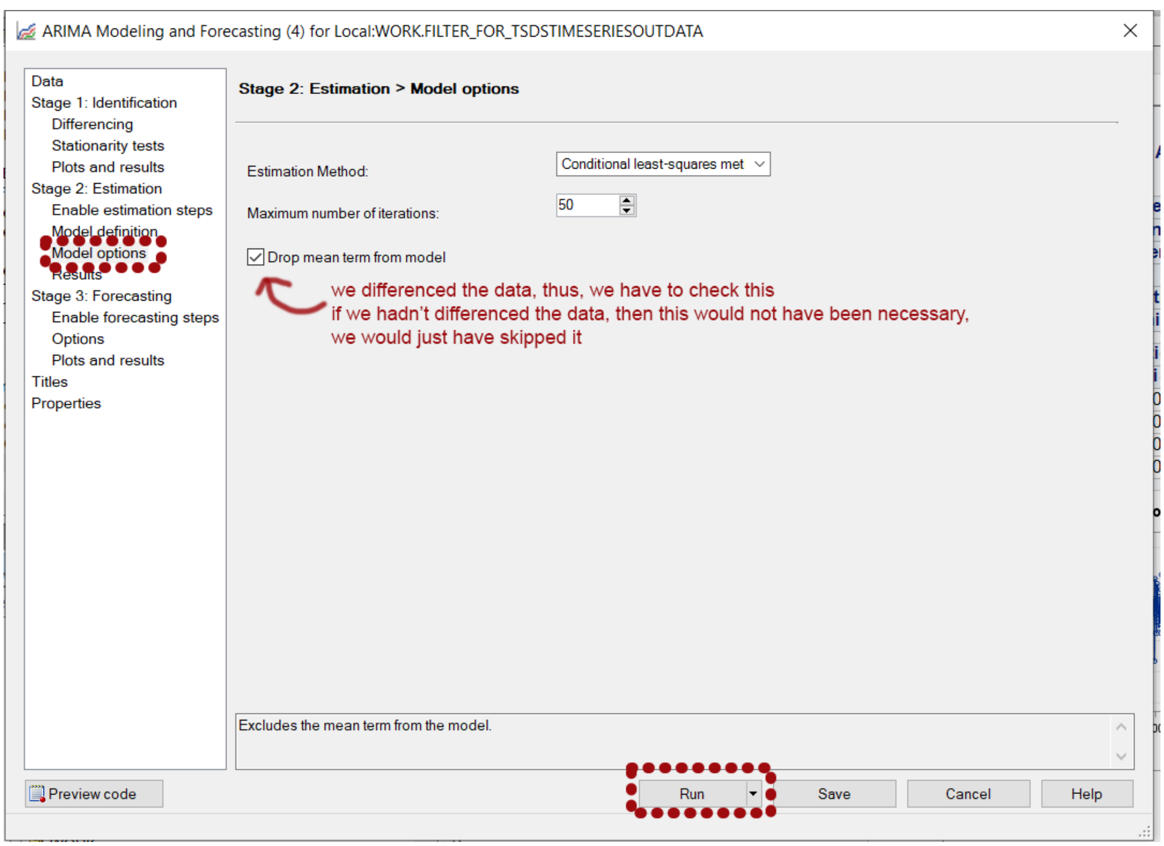
Estimate AR or/and MA parameters.

We are going to estimate AR parameters here:

With the results open, click: Modify task → Stage 2: Estimation → Enable estimation steps → Check: Perform estimation steps → Model definition: We're going to start with AR(1): under Factors for AR model enter 1 → Click Add → Model Options: Check Drop mean term from model (should be done only when you difference the data) → Run → Explore the results




ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Stage 2: Estimation > Model definition
Model definition
Autoregressive (p=)
Enter the lags for the Autoregressive Parameters to be estimated for each multiplicative factor in the list.
Ex. 1 1..4 1,3,12 1 TO 4
Factors for AR model:
Add
AR(1),p=1
Moving Average (q=)
Enter the lags for the Moving Average Parameters to be estimated for each multiplicative factor in the list.
Factors for MA model:
Type an integer value, a list of integers separated by commas, or a range of integers to represent the lags for the autoregressive parameters to be estimated for each multiplicative factor in the list.
Enter the desired value in the text box.
Preview code
Run
Save
Cancel
Help


ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Stage 2: Estimation > Model options
Estimation Method:
Conditional least-squares met
Maximum number of iterations:
50
Drop mean term from model
we differenced the data, thus, we have to check this
if we hadn't differenced the data, then this would not have been necessary,
we would just have skipped it
Excludes the mean term from the model.
Preview code
Run
Save
Cancel
Help

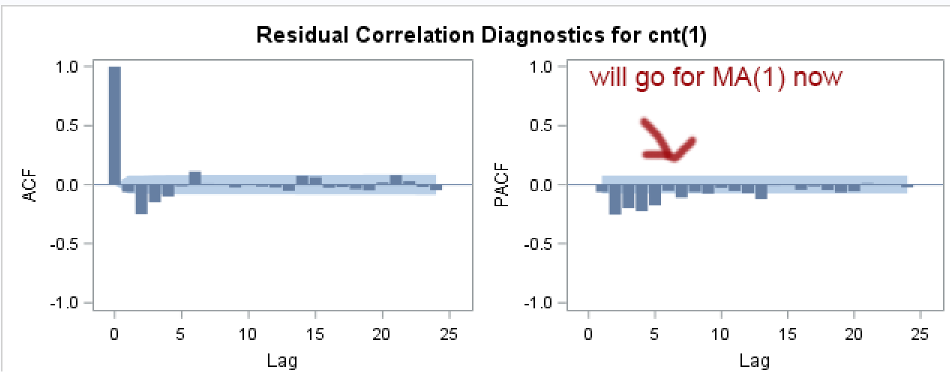
Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
AR1,1	-0.29596	0.03616	-8.19	<.0001	1

The AR coefficient to be used in a formula
It's statistically significant

Variance Estimate	1039959
Std Error Estimate	1019.784
AIC	11669.11
SBC	11673.65
Number of Residuals	699

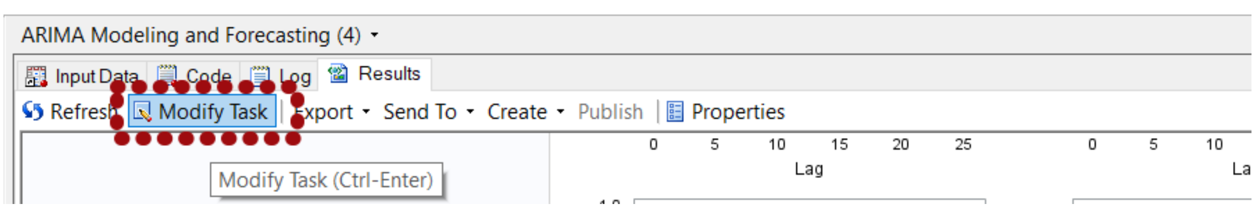
* AIC and SBC do not include log determinant.

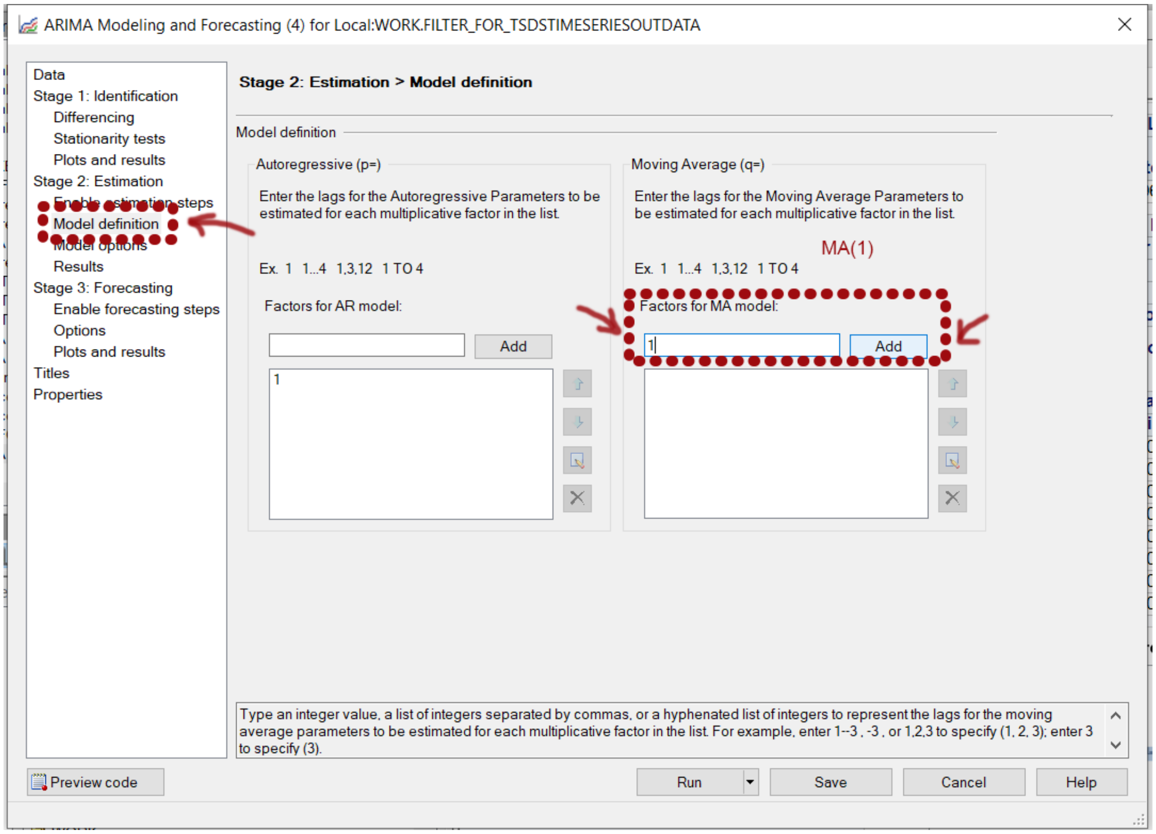
Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	78.03	5	<.0001	-0.063	-0.249	-0.148	-0.101	-0.012	0.113
12	79.09	11	<.0001	0.008	-0.003	-0.024	0.007	-0.014	-0.024
18	89.86	17	<.0001	-0.053	0.074	0.064	-0.027	-0.019	-0.038
24	99.04	23	<.0001	-0.047	0.017	0.082	0.031	-0.018	-0.046
30	119.04	29	<.0001	-0.064	0.038	0.024	0.049	0.121	-0.065
36	130.57	35	<.0001	-0.072	-0.086	0.002	0.019	0.047	0.025
42	139.93	41	<.0001	0.042	-0.085	0.009	-0.045	0.026	0.030
48	143.39	47	<.0001	0.036	-0.013	-0.055	0.006	0.009	-0.008

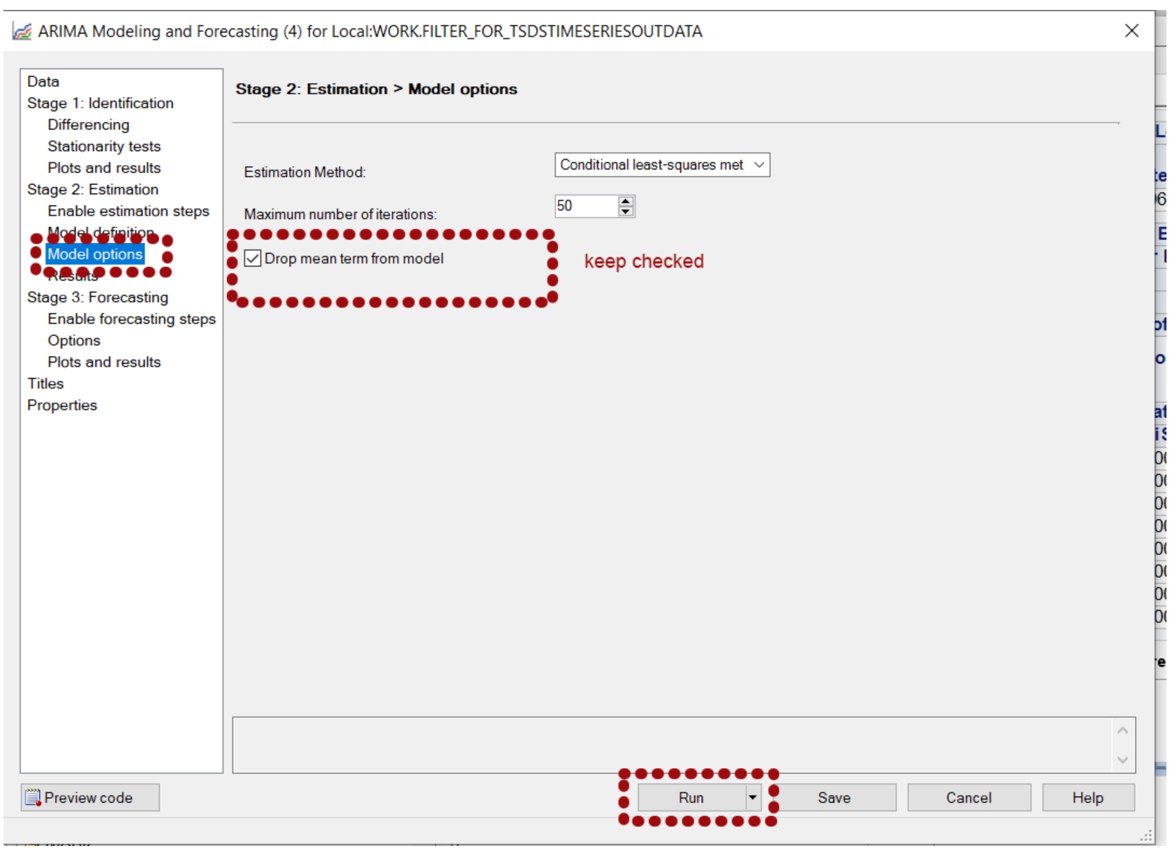


We can see that now we need to add an MA process to the ARIMA model:

With the results open, click: Modify task → Stage 2: Estimation → Enable estimation steps → Check: Perform estimation steps → Model definition: Don't change anything for the AR, we're going to add MA: under Factors for MA model enter 1 → Click Add → Model Options: Keep the Drop mean term from model checked (should be done only when you difference the data) → Run → Explore the results




ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Stage 2: Estimation > Model definition
Model definition
Autoregressive (p=)
Enter the lags for the Autoregressive Parameters to be estimated for each multiplicative factor in the list.
Ex. 1 1...4 1,3,12 1 TO 4
Factors for AR model:
Add
1
Moving Average (q=)
Enter the lags for the Moving Average Parameters to be estimated for each multiplicative factor in the list.
MA(1)
Ex. 1 1...4 1,3,12 1 TO 4
Factors for MA model:
1
Add
Type an integer value, a list of integers separated by commas, or a hyphenated list of integers to represent the lags for the moving average parameters to be estimated for each multiplicative factor in the list. For example, enter 1--3 , -3 , or 1,2,3 to specify (1, 2, 3); enter 3 to specify (3).
Preview code
Run
Save
Cancel
Help


ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Stage 2: Estimation > Model options
Estimation Method:
Conditional least-squares met
Maximum number of iterations:
50
Drop mean term from model
keep checked
Preview code
Run
Save
Cancel
Help

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.90074	0.02036	44.23	<.0001	1
AR1,1	0.33267	0.04419	7.53	<.0001	1

Variance Estimate	837882
Std Error Estimate	915.359
AIC	11519.08
SBC	11528.18
Number of Residuals	699

MA and AR coefficients to be used in the formula

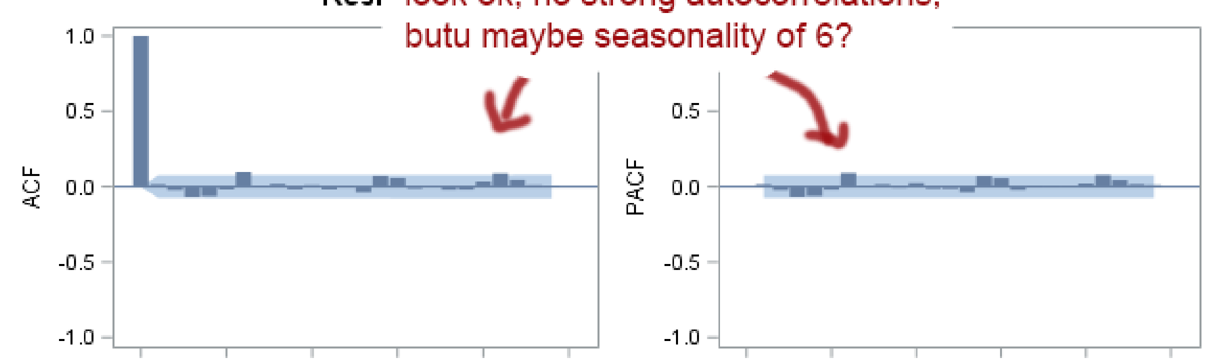
* AIC and SBC do not include log determinant.

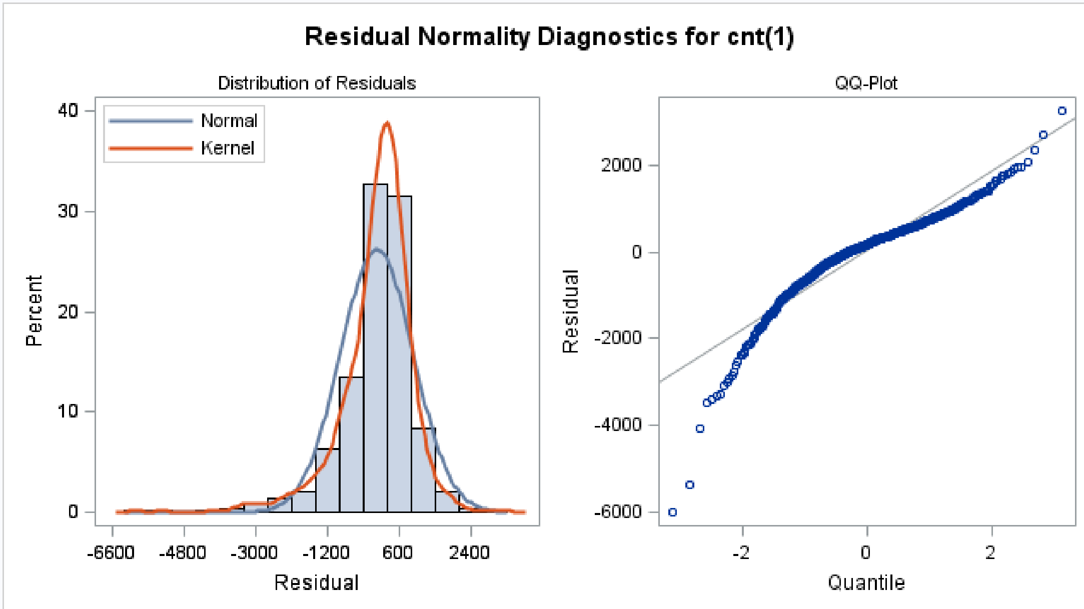
Correlations of Parameter Estimates		
Parameter	MA1,1	AR1,1
MA1,1	1.000	0.588
AR1,1	0.588	1.000

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	13.38	4	0.0096	0.016	-0.020	-0.068	-0.062	-0.016	0.098
12	14.13	10	0.1673	0.008	0.019	-0.015	0.012	-0.015	-0.004
18	21.91	16	0.1460	-0.035	0.074	0.061	-0.010	-0.000	-0.017
24	30.60	22	0.1046	-0.017	0.035	0.089	0.049	0.012	-0.003
30	53.25	28	0.0027	-0.026	0.068	0.056	0.062	0.122	-0.060
36	63.56	34	0.0016	-0.054	-0.092	-0.004	0.007	0.049	0.017
42	73.37	40	0.0010	0.053	-0.078	0.024	-0.045	0.035	0.025
48	75.79	46	0.0037	0.038	0.002	-0.035	0.014	0.019	-0.000

Resi

look ok, no strong autocorrelations, butu maybe seasonality of 6?





Model for variable cnt	
Period(s) of Differencing	1

No mean term in this model.

Autoregressive Factors	
Factor 1:	1 - 0.33267 B**(1)

Moving Average Factors	
Factor 1:	1 - 0.90074 B**(1)

Generated by the SAS System ('Local', X64_10HOME) on November 17, 2019 at 4:50:09 PM

The model is adequate, we can forecast now!

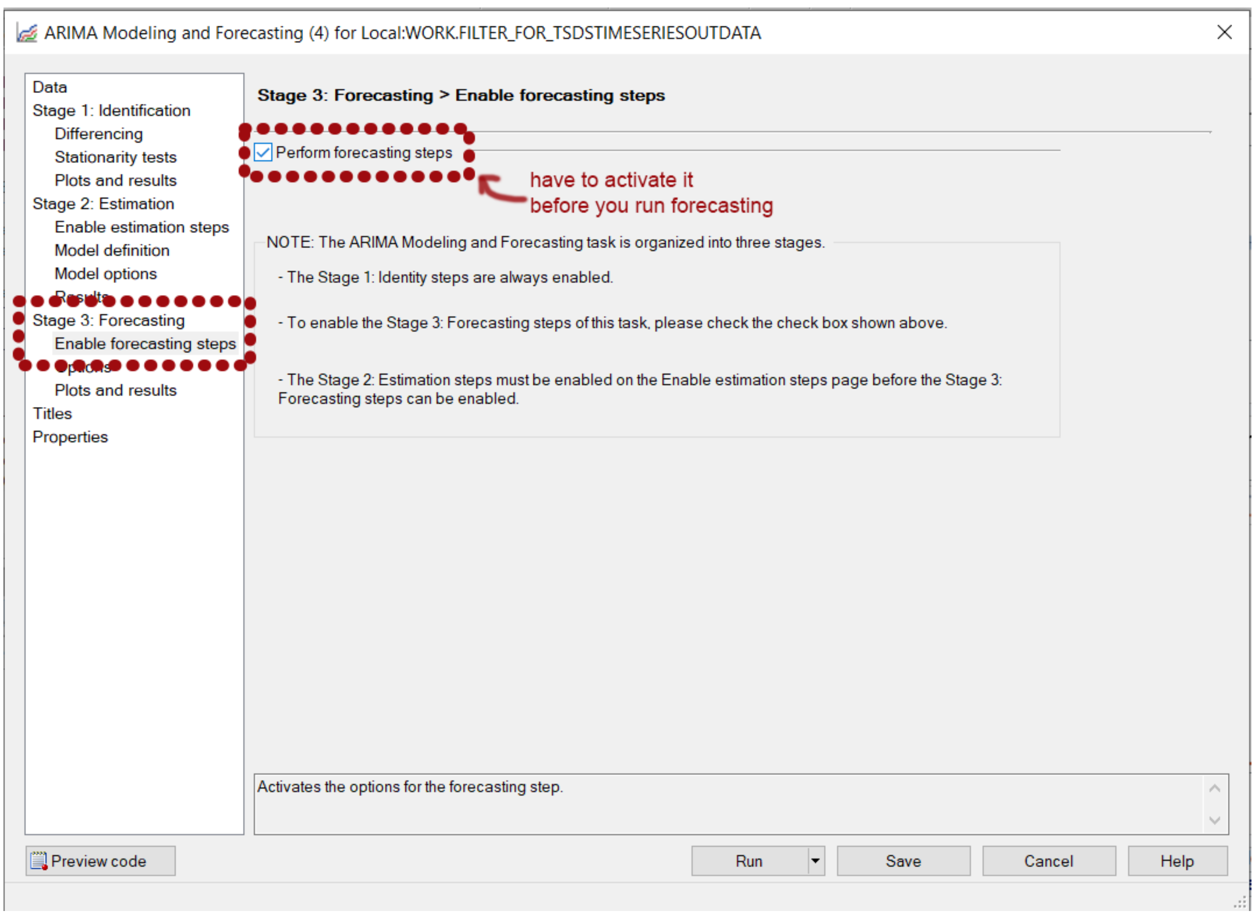
ARIMA: Step 3 - Forecast Total Rides for the Validation Period

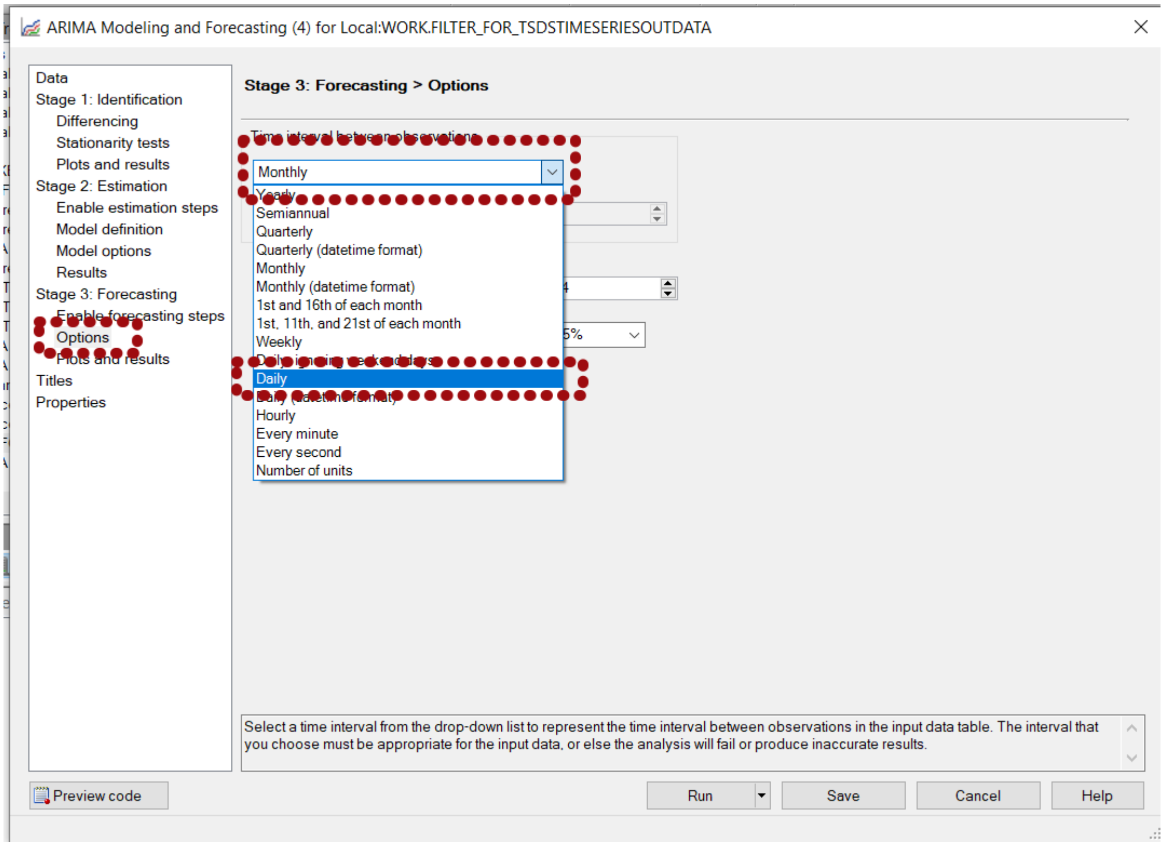
If you remember, we only used the training dataset to build our model. Now, it is time to use it by estimating how well it performs on new data, i.e. the validation data. For this we need to forecast those 31 days we have left out from the training dataset and then compare them to the actual datapoints and assess the accuracy of the model.

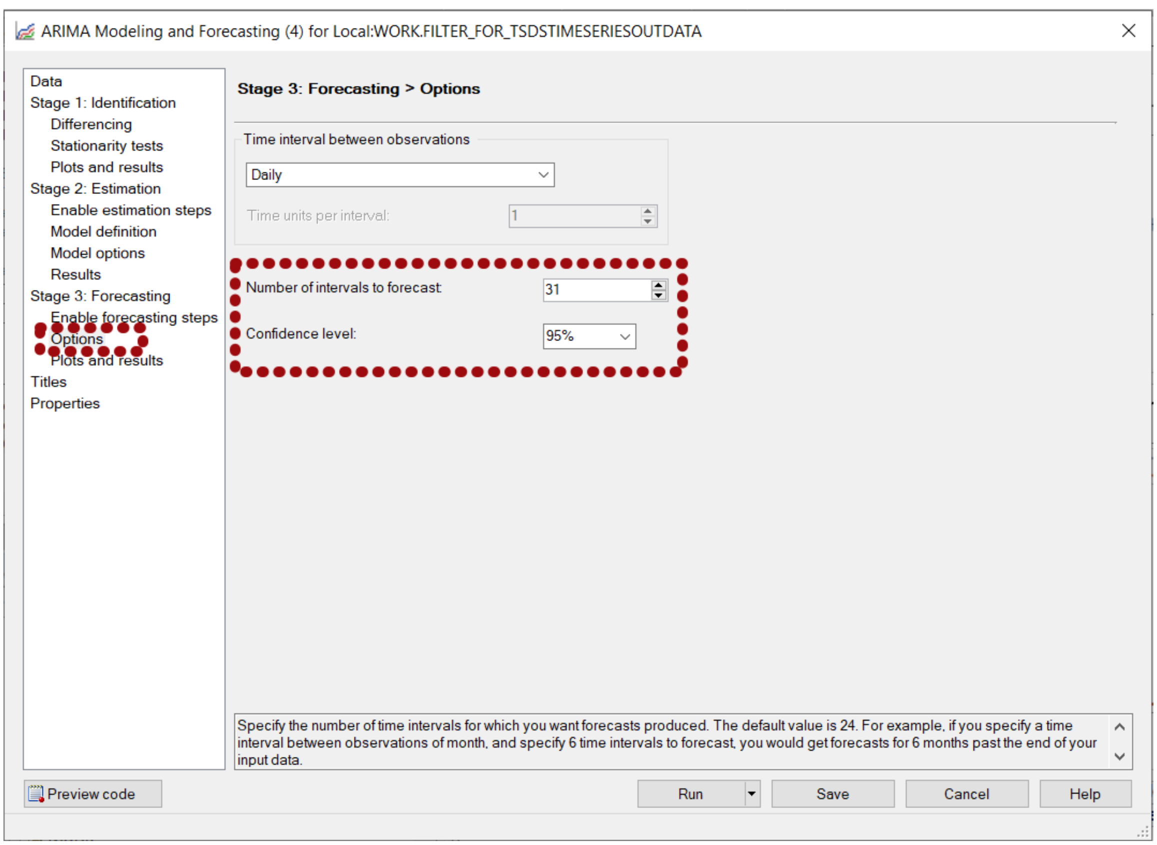
With the results open, click: Modify task → Stage 3: Forecasting→ Enable forecasting steps → Check: Perform forecasting steps → Options → Time interval between observations: Set it to Daily → Number of intervals to forecast: 31 (December of 2012 has 31 days) → Confidence level 95% → Plots and results → Forecasting plots options: Check forecasts, Residuals → You can export forecasts to calculate residuals for the validation dataset: Save forecasts → Run

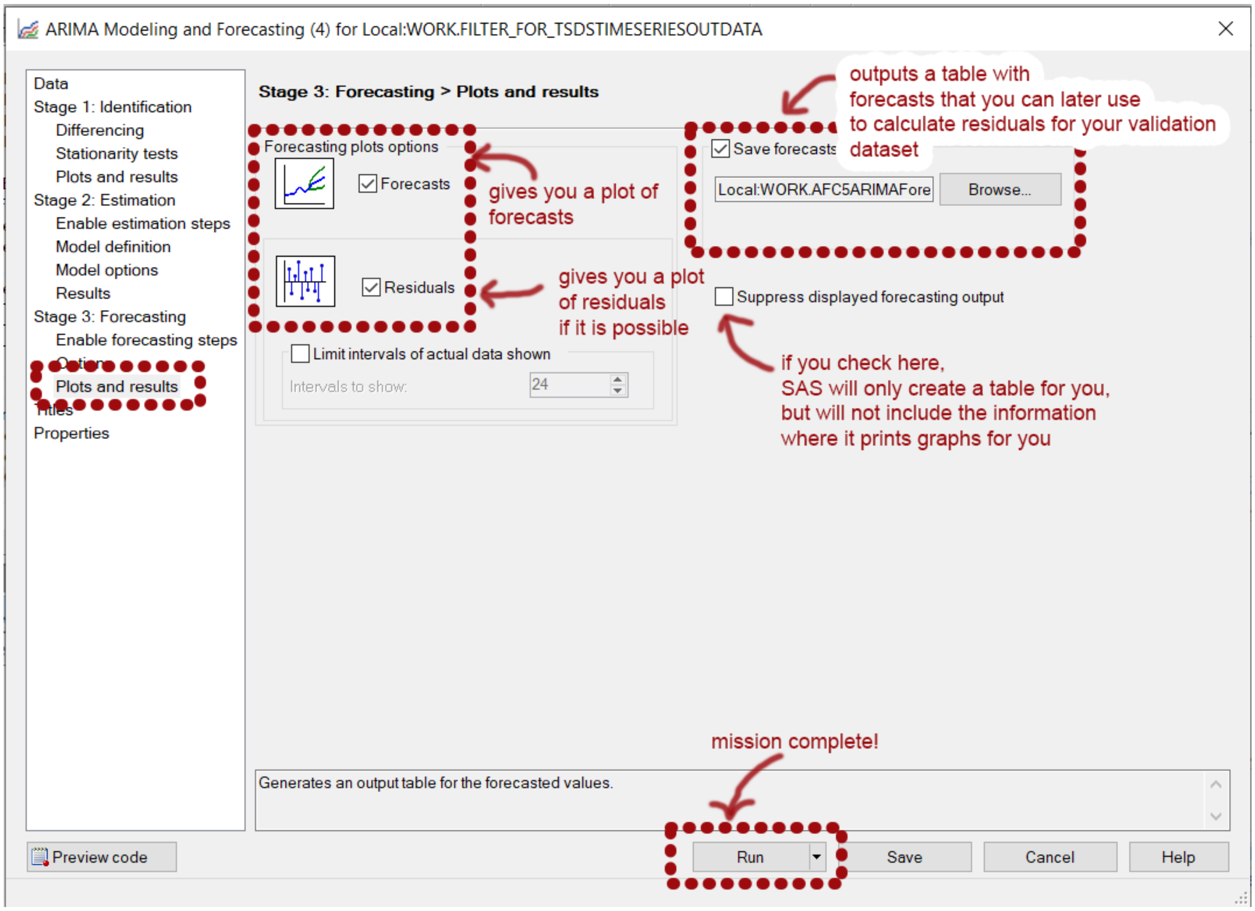
→ To the left of the Results tab, click on Output Data → Export → Select the extension as XLXS (remember to choose the location where you want to save your file; now you'll be able to use it to calculate residuals for the validation dataset by retrieving actual CNT values).

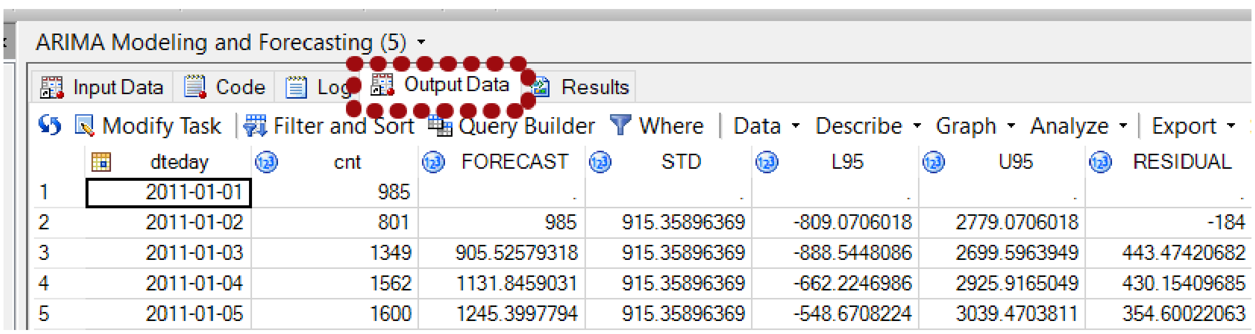
If you remember, we only used the training dataset to build our model. Now, it is time to use it by estimating how well it performs on new data, i.e. the validation data. For this we need to forecast those 31 days we have left out from the training dataset and then compare them to the actual datapoints and assess the accuracy of the model.

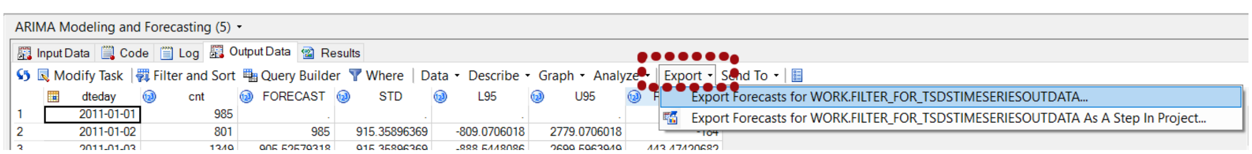


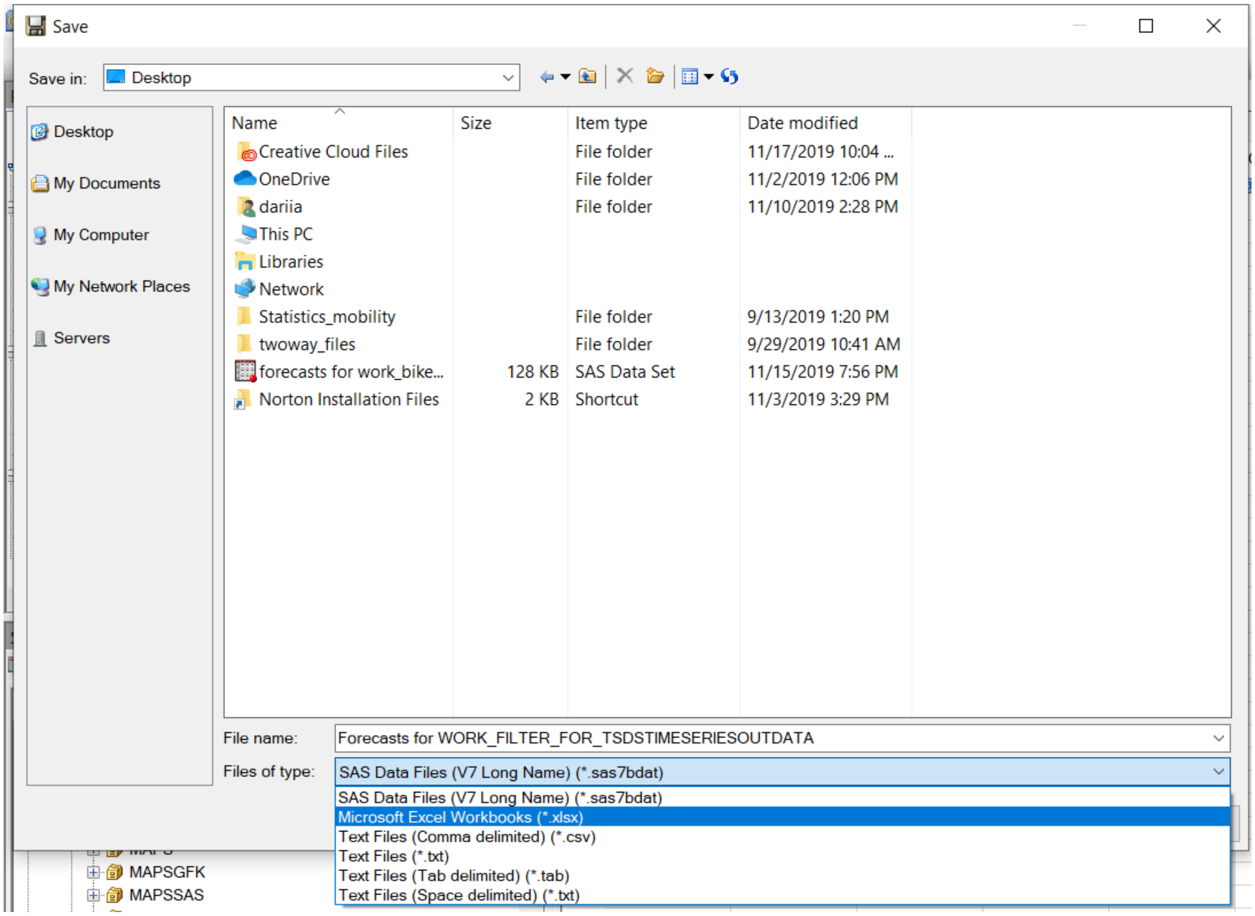

ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Stage 3: Forecasting > Options
Monthly
Semiannual
Quarterly
Quarterly (datetime format)
Monthly
Monthly (datetime format)
1st and 16th of each month
1st, 11th, and 21st of each month
Weekly
Daily
Hourly
Every minute
Every second
Number of units
Select a time interval from the drop-down list to represent the time interval between observations in the input data table. The interval that you choose must be appropriate for the input data, or else the analysis will fail or produce inaccurate results.
Preview code
Run
Save
Cancel
Help


ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Stage 3: Forecasting > Options
Time interval between observations
Daily
Time units per interval:
1
Number of intervals to forecast
31
Confidence level:
95%
Specify the number of time intervals for which you want forecasts produced. The default value is 24. For example, if you specify a time interval between observations of month, and specify 6 time intervals to forecast, you would get forecasts for 6 months past the end of your input data.
Preview code
Run
Save
Cancel
Help


ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Stage 3: Forecasting > Plots and results
Forecasting plots options
Forecasts
Residuals
gives you a plot of forecasts
gives you a plot of residuals if it is possible
Save forecasts
Local:WORK.AFC5ARIMAFore
Browse...
outputs a table with forecasts that you can later use to calculate residuals for your validation dataset
Suppress displayed forecasting output
if you check here, SAS will only create a table for you, but will not include the information where it prints graphs for you
Limit intervals of actual data shown
Intervals to show:
24
mission complete!
Generates an output table for the forecasted values.
Preview code
Run
Save
Cancel
Help


ARIMA Modeling and Forecasting (5)
Input Data
Code
Log
Output Data
Results
Modify Task
Filter and Sort
Query Builder
Where
Data
Describe
Graph
Analyze
Export
dteday
cnt
FORECAST
STD
L95
U95
RESIDUAL
1 2011-01-01 985 . . . . .
2 2011-01-02 801 985 915.35896369 -809.0706018 2779.0706018 -184
3 2011-01-03 1349 905.52579318 915.35896369 -888.5448086 2699.5963949 443.47420682
4 2011-01-04 1562 1131.8459031 915.35896369 -662.2246986 2925.9165049 430.15409685
5 2011-01-05 1600 1245.3997794 915.35896369 -548.6708224 3039.4703811 354.60022063

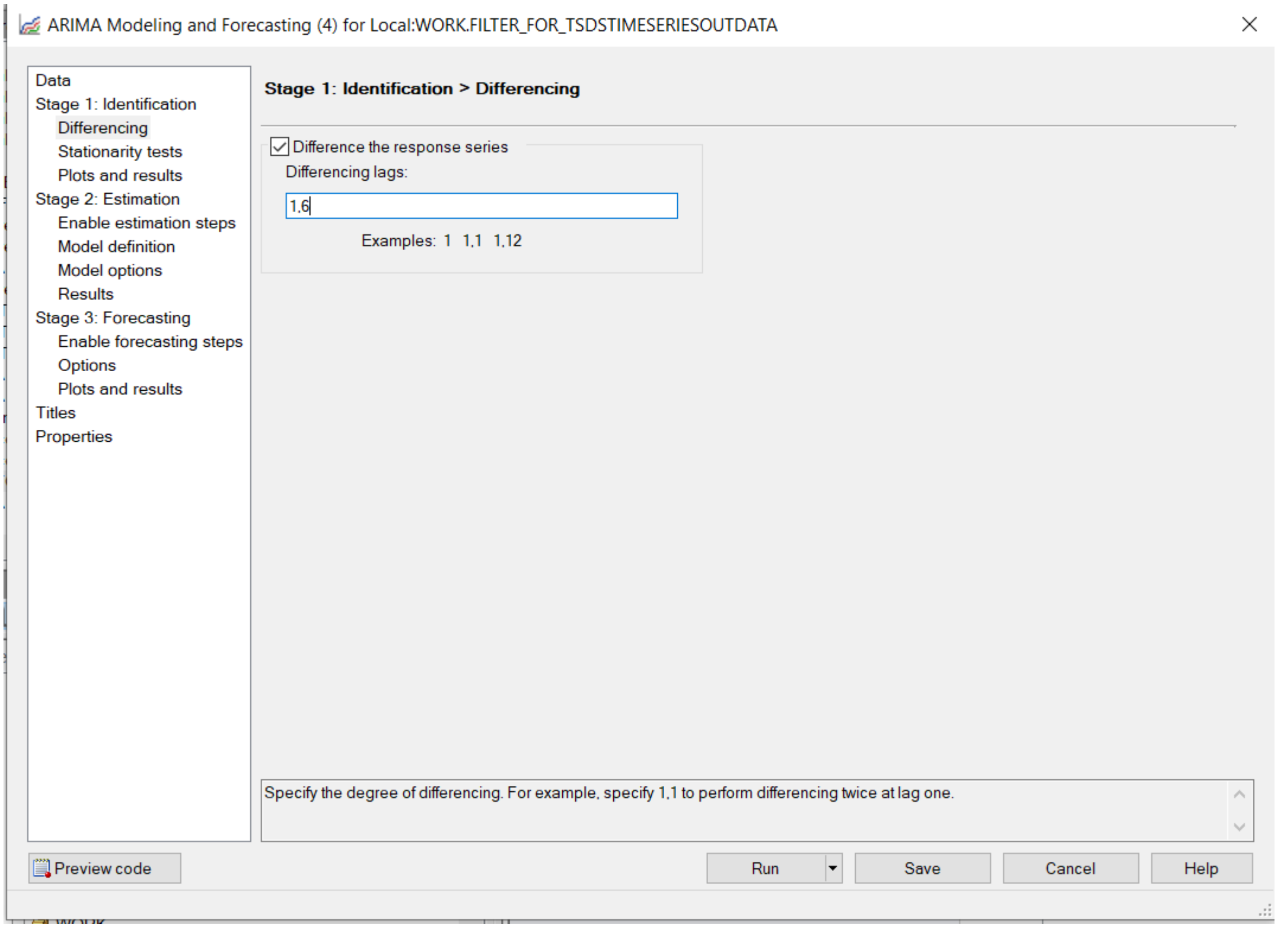

ARIMA Modeling and Forecasting (5)
Input Data
Code
Log
Output Data
Results
Modify Task
Filter and Sort
Query Builder
Where
Data
Describe
Graph
Analyze
Export
Send To
Export Forecasts for WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA...
Export Forecasts for WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA As A Step In Project...



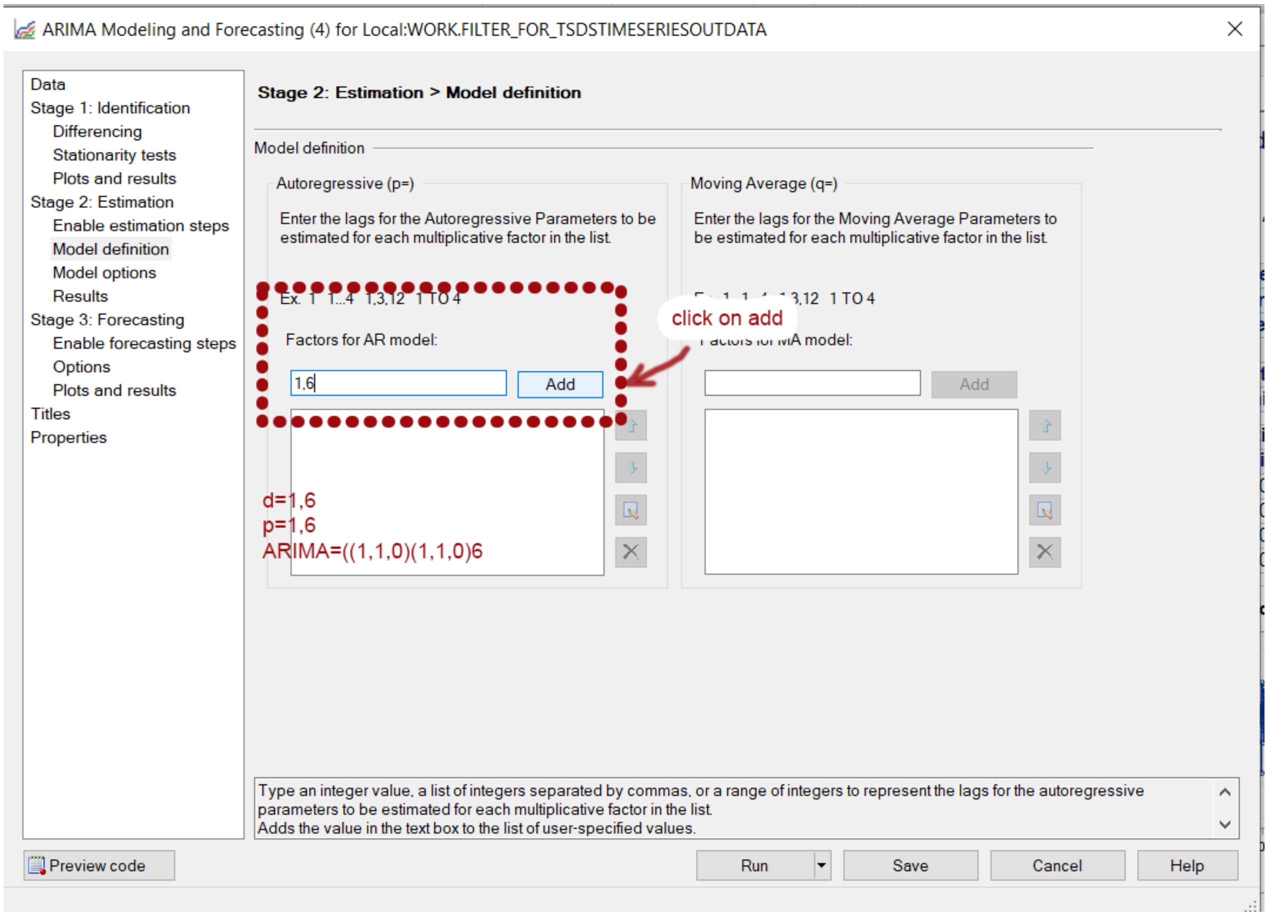
Seasonal ARIMA: How to Input Values

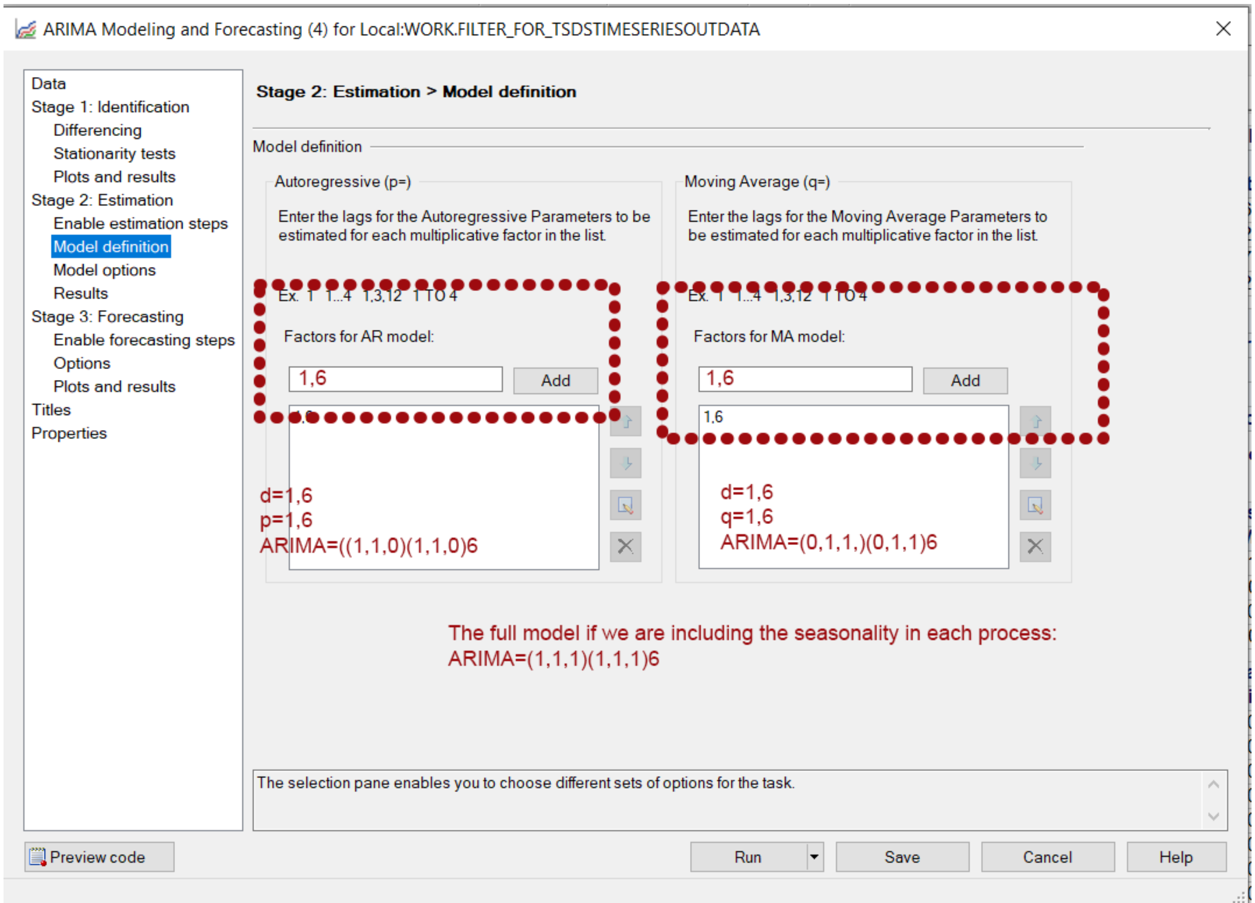
For seasonal ARIMA, you just have to enter two lags, the non-seasonal one and the seasonal one.

Double differencing at lag1 and at lag6



Setting AR at lag1 and adding a seasonal lag of 6


ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Stage 2: Estimation > Model definition
Model definition
Autoregressive (p=)
Enter the lags for the Autoregressive Parameters to be estimated for each multiplicative factor in the list.
Ex. 1 1...4 1,3,12 1 TO 4
Factors for AR model:
1,6
Add
Moving Average (q=)
Enter the lags for the Moving Average Parameters to be estimated for each multiplicative factor in the list.
click on add
d=1,6
p=1,6
ARIMA=((1,1,0)(1,1,0)6
Add
Type an integer value, a list of integers separated by commas, or a range of integers to represent the lags for the autoregressive parameters to be estimated for each multiplicative factor in the list.
Adds the value in the text box to the list of user-specified values.
Preview code
Run
Save
Cancel
Help


ARIMA Modeling and Forecasting (4) for Local:WORK.FILTER_FOR_TSDSTIMESERIESOUTDATA
Data
Stage 1: Identification
Differencing
Stationarity tests
Plots and results
Stage 2: Estimation
Enable estimation steps
Model definition
Model options
Results
Stage 3: Forecasting
Enable forecasting steps
Options
Plots and results
Titles
Properties
Stage 2: Estimation > Model definition
Model definition
Autoregressive (p=)
Enter the lags for the Autoregressive Parameters to be estimated for each multiplicative factor in the list.
Factors for AR model:
1,6
Add
d=1,6
p=1,6
ARIMA=((1,1,0)(1,1,0)6
Moving Average (q=)
Enter the lags for the Moving Average Parameters to be estimated for each multiplicative factor in the list.
Factors for MA model:
1,6
Add
1,6
d=1,6
q=1,6
ARIMA=(0,1,1,)(0,1,1)6
The full model if we are including the seasonality in each process:
ARIMA=(1,1,1)(1,1,1)6
The selection pane enables you to choose different sets of options for the task.
Preview code
Run
Save
Cancel
Help

Conditional Least Squares Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx Pr > |t|**	**Lag**
MA1,1	0.0013665	0.0093286	0.15	0.8836	1
MA1,2	0.97925	0.0092349	106.04	<.0001	6
AR1,1	-0.28875	0.03707	-7.79	<.0001	1
AR1,2	0.09525	0.03771	2.53	0.0118	6

statisticaly insignificant

coefficients

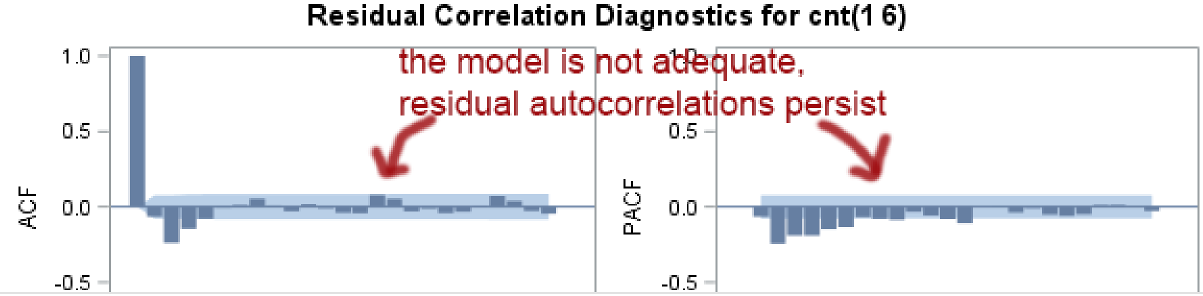
Variance Estimate	1056748
Std Error Estimate	1027.983
AIC	11583.04
SBC	11601.2
Number of Residuals	693

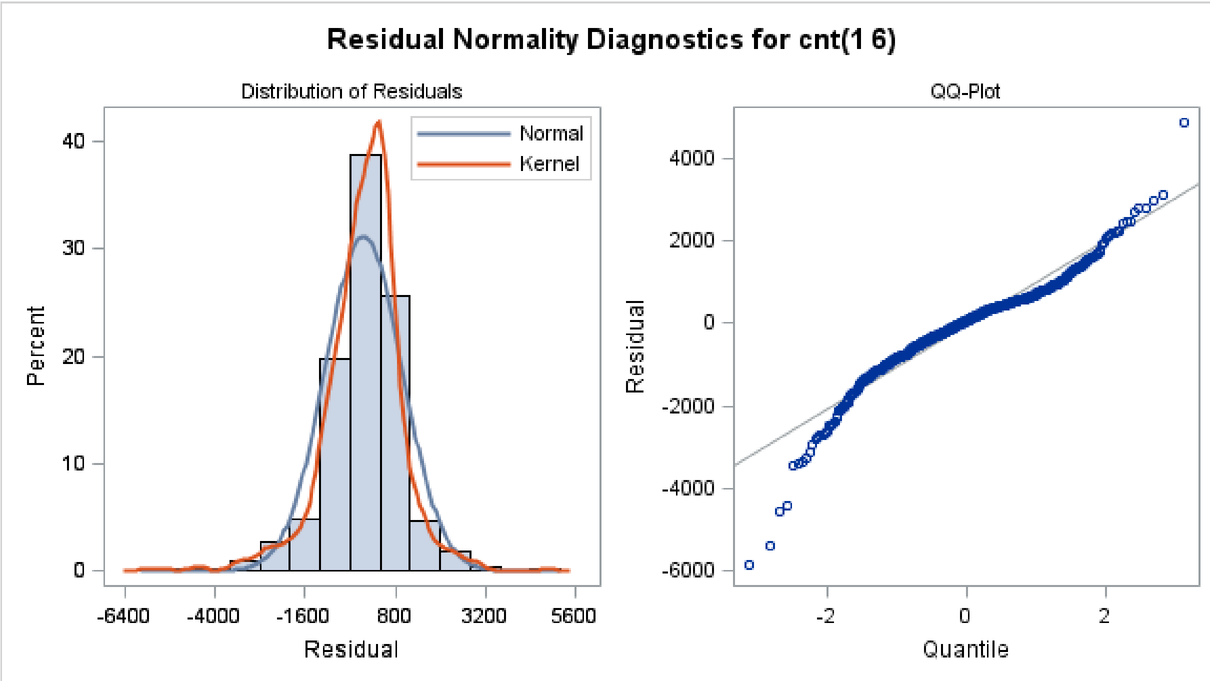
* AIC and SBC do not include log determinant.

Correlations of Parameter Estimates				
Parameter	**MA1,1**	**MA1,2**	**AR1,1**	**AR1,2**
MA1,1	1.000	-0.405	0.188	-0.152
MA1,2	-0.405	1.000	-0.003	0.244
AR1,1	0.188	-0.003	1.000	0.004
AR1,2	-0.152	0.244	0.004	1.000

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr > ChiSq**	**Autocorrelations**					
6	60.42	2	<.0001	-0.062	-0.238	-0.141	-0.077	-0.005	0.015
12	64.50	8	<.0001	0.056	0.010	-0.027	0.020	-0.009	-0.037
18	73.99	14	<.0001	-0.040	0.079	0.055	-0.028	-0.011	-0.039
24	81.58	20	<.0001	-0.031	0.009	0.073	0.041	-0.026	-0.043
30	102.39	26	<.0001	-0.052	0.048	0.012	0.054	0.126	-0.069
36	111.92	32	<.0001	-0.066	-0.078	-0.006	0.021	0.038	0.028
42	120.65	38	<.0001	0.052	-0.077	0.009	-0.044	0.023	0.027
48	126.36	44	<.0001	0.036	0.005	-0.073	0.025	0.008	-0.018

Residual Correlation Diagnostics for cnt(1 6)





Model for variable cnt	
Period(s) of Differencing	1,6

No mean term in this model.

Coefficients to be used in the formula

Autoregressive Factors	
Factor 1:	1 + 0.28875 B**(1) - 0.09525 B**(6)

Moving Average Factors	
Factor 1:	1 - 0.00137 B**(1) - 0.97925 B**(6)