

BSTA 477/677 – Winter 2021

Tutorial 1 - Jan 31st, 2021

[SAS Enterprise Guide - Basic](#)

[Basic data manipulation](#)

[Data import](#)

[Data preparation](#)

[Data summary](#)

[Data partitioning](#)

[Data merge](#)

[Other cleaning techniques](#)

[Time Series Graphics](#)

[Histograms](#)

[Line plot](#)

[Scatter plot](#)

[ACF, PACF](#)

[Seasonal plot](#)

[Time series decomposition](#)

[Classical Decomposition](#)

[Automatic SAS EG inferred classical decomposition.](#)

[SAS Seasonality User manually defined m-MA](#)

[X11 Decomposition](#)

[X13 Decomposition](#)

[Box-cox transformation](#)

[SAS Resources](#)

[SAS Student hub](#)

Data used: Bike sharing Washington D.C. dataset

Software used: SAS EG

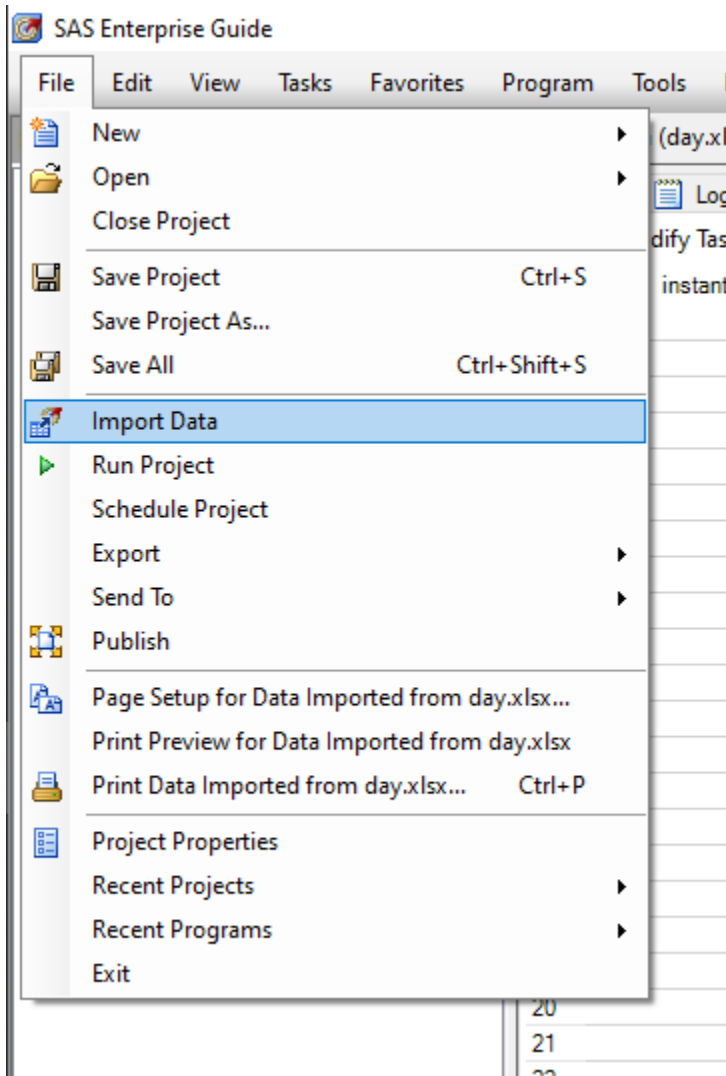
SAS Enterprise Guide - Basic

- Process flow
- Task bar
- Library bar - Data storage
- Note/Warning/Error bar


Basic data manipulation

Data import


File > Import Data > Open > (optional) Browse SAS output data set location > (optional) Adjust variable source format



Note: For importing excel files, there are options to select the sheet to import and/or select a range of cells for the sheet.

 Import Data from day.xlsx

2 of 4 Select Data Source



Select range

☒ Use a worksheet

day

☐ Use a specific range of cells within the worksheet

Top-left cell:

Lower-right cell:

☐ Expand row range as needed

Reset Range

☐ Use a predefined named range

☒ First row of range contains field names

☒ Rename columns to comply with SAS naming conventions.

<Back

Next>

Finish

Cancel

Help

Data preparation

Simple aggregation of time series data:

Output data tab > Analyze > Time series > Prepare Time Series data > (optional) Adjust Existing Time ID variable interval > (optional) Interpolation method: Simple aggregation

The screenshot shows the 'Prepare Time Series Data' dialog box. The table contains the following data:

	dteday	cnt
1	01JAN2011	985
2	02JAN2011	801
3	03JAN2011	1349
4	04JAN2011	1562
5	05JAN2011	1600
6	06JAN2011	1606
7	07JAN2011	1510
8	08JAN2011	959
9	09JAN2011	822
10	10JAN2011	1321
11	11JAN2011	1263
12	12JAN2011	1162
13	13JAN2011	1406
14	14JAN2011	1421
15	15JAN2011	1248
16	16JAN2011	1204
17	17JAN2011	1000
18	18JAN2011	683
19	19JAN2011	1650
20	20JAN2011	1927
21	21JAN2011	1543
22	22JAN2011	981
23	23JAN2011	986
24	24JAN2011	1416

The 'Analyze' menu is open, showing the following options:

- ANOVA
- Regression
- Multivariate
- Survival Analysis
- Capability
- Control Charts
- Pareto Chart...
- Time Series
- Data Mining

The 'Time Series' option is selected, and a sub-menu is open showing the following options:

- 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...

Data summary

- Code: Proc content

```
proc contents data=bsta477.bike_sharing_day_data;  
run;  
|
```

Information given:

The CONTENTS Procedure

Data Set Name	BSTA477.BIKE_SHARING_DAY_DATA	Observations	731
Member Type	DATA	Variables	16
Engine	V9	Indexes	0
Created	01/29/2021 16:47:02	Observation Length	128
Last Modified	01/29/2021 16:47:02	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information

Data Set Page Size	65536
Number of Data Set Pages	2
First Data Page	1
Max Obs per Page	511
Obs in First Data Page	490
Number of Data Set Repairs	0
ExtendObsCounter	YES
Filename	C:\Users\mathd\Documents\BSTA477\bike_sharing_day_data.sas7bdat
Release Created	9.0401M4
Host Created	X64_10PRO
Owner Name	MATHIEU-PC\mdugre
File Size	192KB
File Size (bytes)	196608

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
11	atemp	Num	8	BEST12.	BEST12.
14	casual	Num	8	BEST12.	BEST12.
16	cnt	Num	8	BEST12.	BEST12.
2	dteday	Num	8	DATE9.	DATE9.
6	holiday	Num	8	BEST12.	BEST12.
12	hum	Num	8	BEST12.	BEST12.
1	instant	Num	8	BEST12.	BEST12.
5	mnth	Num	8	BEST12.	BEST12.
15	registered	Num	8	BEST12.	BEST12.
3	season	Num	8	BEST12.	BEST12.
10	temp	Num	8	BEST12.	BEST12.
9	weathersit	Num	8	BEST12.	BEST12.
7	weekday	Num	8	BEST12.	BEST12.
13	windspeed	Num	8	BEST12.	BEST12.
8	workingday	Num	8	BEST12.	BEST12.
4	yr	Num	8	BEST12.	BEST12.

- Task:
 - Output data tab > Describe > Summary statistics (choose an analysis variable)

Data Imported from day.xlsx

	instant	dteday	season	holiday	weekday	workingday
1	1	01JAN2011	1	0	6	0
2	2	02JAN2011	1	0	0	0
3	3	03JAN2011	1	0	1	1
4	4	04JAN2011	1	0	2	1
5	5	05JAN2011	1	0	3	1
6	6	06JAN2011	1	0	4	1
7	7	07JAN2011	1	0	5	1
8	8	08JAN2011	1	0	6	0
9	9	09JAN2011	1	0	0	0
10	10	10JAN2011	1	0	1	1
11	11	11JAN2011	1	0	2	1
12	12	12JAN2011	1	0	3	1
13	13	13JAN2011	1	0	4	1
14	14	14JAN2011	1	0	5	1
15	15	15JAN2011	1	0	6	0
16	16	16JAN2011	1	0	0	0
17	17	17JAN2011	1	1	1	0
18	18	18JAN2011	1	0	2	1
19	19	19JAN2011	1	0	3	1
20	20	20JAN2011	1	0	4	1
21	21	21JAN2011	1	0	5	1
22	22	22JAN2011	1	0	6	0
23	23	23JAN2011	1	0	0	0

Options to explore variables:

Summary Statistics for Local:BSTA477.BIKE_SHARING_DAY_DATA

Data

- Statistics
 - Basic
 - Percentiles
 - Additional
- Plots
- Results
- Titles
- Properties

Data

Data source: Local:BSTA477.BIKE_SHARING_DAY_DATA

Task filter: None

Edit...

Variables to assign:

Name

- instant
- dteday
- season
- yr
- mnth
- holiday
- weekday
- workingday
- weathersit
- temp
- atemp
- hum
- windspeed
- casual
- registered

Task roles:

- Analysis variables
- cnt
- Classification variables
- weekday
- Frequency count (Limit: 1)
- Relative weight (Limit: 1)
- Copy variables
- Group analysis by

Class level weekday

Sort by: Unformatted values

Sort order: Ascending

Missing values: Exclude

☐ Allow multi-label formats

The selection pane enables you to choose different sets of options for the task.

Preview code Run Save Cancel Help

- Output data tab > Describe > Summary tables

Data Imported from day.xlsx ▾

	instant	dteday	season	holiday	weekday
1	1	01JAN2011	1	0	6
2	2	02JAN2011	1	0	0
3	3	03JAN2011	1	0	1
4	4	04JAN2011	1	0	2
5	5	05JAN2011	1	0	3
6	6	06JAN2011	1	0	4
7	7	07JAN2011	1	0	5
8	8	08JAN2011	1	0	6
9	9	09JAN2011	1	0	0
10	10	10JAN2011	1	0	1
11	11	11JAN2011	1	0	2
12	12	12JAN2011	1	0	3
13	13	13JAN2011	1	0	4
14	14	14JAN2011	1	0	5
15	15	15JAN2011	1	0	6
16	16	16JAN2011	1	0	0
17	17	17JAN2011	1	0	1
18	18	18JAN2011	1	0	2
19	19	19JAN2011	1	0	3

Describe ▾

- List Data...
- Summary Statistics Wizard...
- Summary Statistics...
- Summary Tables Wizard...
- Summary Tables...**
- List Report Wizard...
- Characterize Data...
- Distribution Analysis...
- One-Way Frequencies...
- Table Analysis...

Data partitioning

Goal of data partitioning is to partition the time series into a training set and a validation set.

Example: Data set with 731 observations. Goal: 80% observation in training set, 20% validation set. Because 80% of 731 is 585, time ID variable < 586 will be 80% of data set. Therefore, time ID variable > 585 will be 20% of the data set.

Training set:

Output data tab > Filter and Sort > Variable tab: Choose variables in the output set > Filter: Choose ID variable > Filter: Less than > Filter: ... Choose value at 586 (because 80% of 731)

Data Imported from day.xlsx ▾

Filter and Sort Query Builder Where Data ▾ Describe

	instant	dteday	season	
1	1	01JAN2011	1	
2	2	02JAN2011	1	
3	3	03JAN2011	1	
4	4	04JAN2011	1	
5	5	05JAN2011	1	
6	6	06JAN2011	1	

Filter and Sort for Local:BSTA477.BIKE_SHARING_DAY_DATA

Variables Filter Sort Results

Filter description:

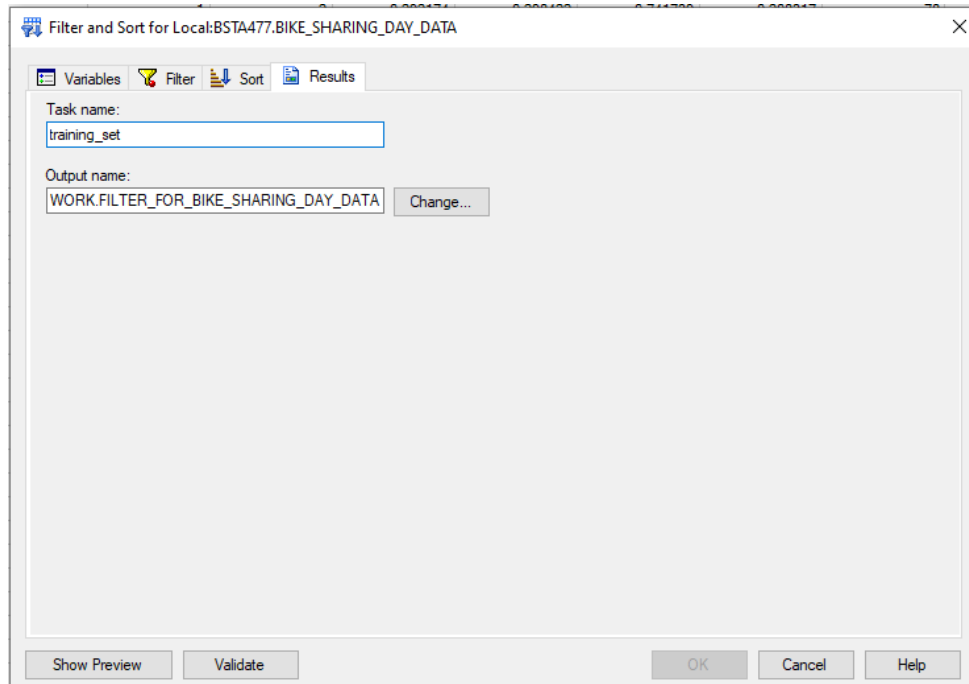
instant ▾ Less than ▾ 586 ... ▾ ✕

Add filters by selecting the AND/OR operator at the end of the expression

☐ Display labels instead of variable names ☒ Match case

Advanced Edit.... Clear All

Show Preview Validate OK Cancel Help



Validation set:

Output data tab > Filter and Sort > Variable tab: Choose variables in the output set > Filter:
Choose ID variable > Filter: More than > Filter: ... Choose value at 585

Data merge

Reference: [SAS Help Center MERGE statement](#)

Example: Data_1 (var: ID, name, job), Data_2 (var: ID, name, age). Merge two data sets by column name column.

Data merged;
Merge data_1 data_2;
By name;
Run;

Other cleaning techniques

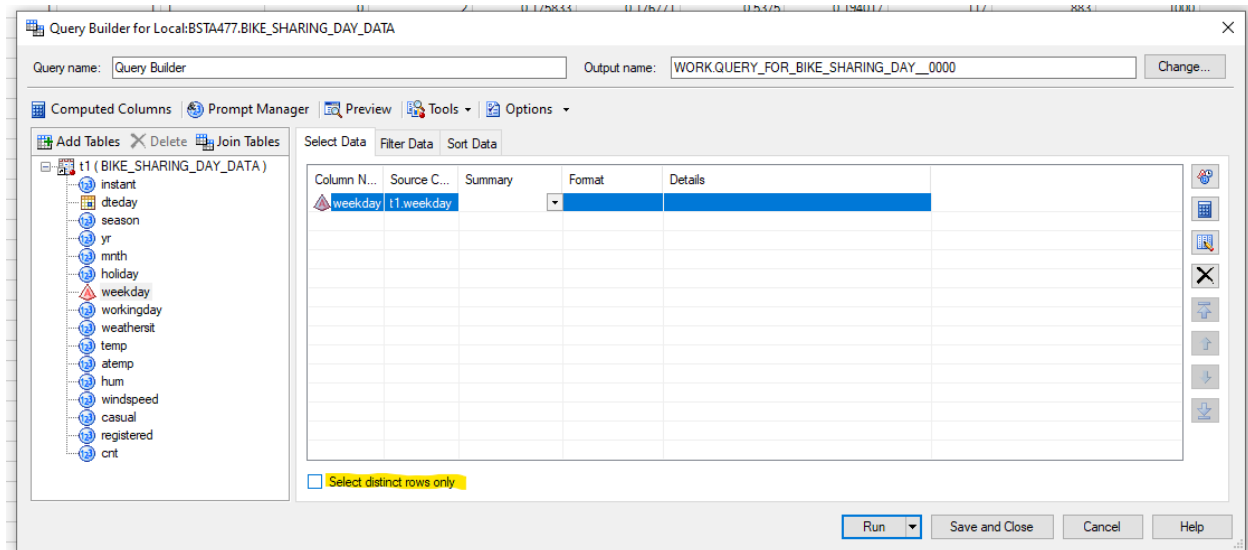
- Remove duplicates:

Output data tab > Query Builder > Drag columns to be in output data > Select only distinct rows

Data Imported from day.xlsx ▾

Filter and Sort Query Builder Where Data ▾ Describe ▾ Graph ▾

	instant	dteday	season	yr
1	1	01JAN2011	1	0
2	2	02JAN2011	1	0
3	3	03JAN2011	1	0
4	4	04JAN2011	1	0
5	5	05JAN2011	1	0



- Missing data: numerical and categorical: Proc Content

Reference: [Managing missing data using SAS Enterprise Guide](#)

Time Series Graphics

Histograms

Output data tab > Describe > Distribution analysis > Select analysis variable > Distribution
Summary: Normal selected > Plots Appearance: Histogram plot > Run

Data Imported from day.xlsx ▾

Filter and Sort Query Builder Where Data Describe Graph Analyze Export Send To

	instant	dteday	season	holiday	week
1	1	01JAN2011	1	0	6
2	2	02JAN2011	1	0	0
3	3	03JAN2011	1	0	1
4	4	04JAN2011	1	0	2
5	5	05JAN2011	1	0	3
6	6	06JAN2011	1	0	4
7	7	07JAN2011	1	0	5
8	8	08JAN2011	1	0	6
9	9	09JAN2011	1	0	0
10	10	10JAN2011	1	0	1
11	11	11JAN2011	1	0	2
12	12	12JAN2011	1	0	3
13	13	13JAN2011	1	0	4
14	14	14JAN2011	1	0	5

Describe ▾

- List Data...
- Summary Statistics Wizard...
- Summary Statistics...
- Summary Tables Wizard...
- Summary Tables...
- List Report Wizard...
- Characterize Data...
- Distribution Analysis...**
- One-Way Frequencies...
- Table Analysis...

Distribution Analysis (2) for Local:BSTA477.BIKE_SHARING_DAY_DATA

Data Distributions Summary Normal Lognormal Exponential Weibull Beta Gamma Kernel Plots Appearance Inset Tables Titles Properties

Plots > Appearance

Note: Insets are valid on histogram, probability and quantile-quantile plots only.

Axis color: Background color: Axis width:

☒ Histogram Plot

☐ Probability Plot

☐ Quantiles plot

☐ Box plot

☐ Text-based plots

Produces a stem and leaf plot or bar chart (depending on the number of observations), box plot and normal probability plot. Produces a side-by-side plot if there is a by variable.

Creates a histogram and optionally superimposes density curves for continuous theoretical distributions and for kernel density estimates.

Preview code Run Save Cancel Help

Line plot

Output data tab > Graph > Line plot

Data Imported from day.xlsx									
Filter and Sort	Query Builder	Where	Data	Describe	Graph	Analyze	Export	Send To	
instant	date	season	yr						
1	01JAN2011	1			Bar Chart Wizard...	weekday	workingday	weathersit	temp
2	02JAN2011	1			Bar Chart...				atemp
3	03JAN2011	1			Pie Chart Wizard...				hum
4	04JAN2011	1			Pie Chart...				windspeed
5	05JAN2011	1			Line Plot Wizard...				
6	06JAN2011	1			Line Plot...				
7	07JAN2011	1			Scatter Plot...				
8	08JAN2011	1			Scatter Plot Matrix...				
9	09JAN2011	1			Area Plot...				
10	10JAN2011	1			Bar-Line Chart...				
11	11JAN2011	1			Bubble Plot...				
12	12JAN2011	1			Donut Chart...				
13	13JAN2011	1			Contour Plot...				
14	14JAN2011	1			Box Plot...				
15	15JAN2011	1			Radar Chart...				
16	16JAN2011	1			Surface Plot...				
17	17JAN2011	1			Tile Chart...				
18	18JAN2011	1			Map Chart...				
19	19JAN2011	1			Open ODS Graphics Designer...				
20	20JAN2011	1			Show ODS Statistical Graph...				
21	21JAN2011	1							
22	22JAN2011	1							
23	23JAN2011	1							
24	24JAN2011	1							
25	25JAN2011	1							
26	26JAN2011	1							
27	27JAN2011	1							
28	28JAN2011	1							
29	29JAN2011	1							

Scatter plot

Output data tab > Graph > Scatter plot

Data Imported from day.xlsx									
Filter and Sort	Query Builder	Where	Data	Describe	Graph	Analyze	Export	Send To	
instant	date	season	yr						
1	01JAN2011	1			Bar Chart Wizard...	weekday	workingday	weathersit	
2	02JAN2011	1			Bar Chart...				
3	03JAN2011	1			Pie Chart Wizard...				
4	04JAN2011	1			Pie Chart...				
5	05JAN2011	1			Line Plot Wizard...				
6	06JAN2011	1			Line Plot...				
7	07JAN2011	1			Scatter Plot...				
8	08JAN2011	1			Scatter Plot Matrix...				
9	09JAN2011	1			Area Plot...				
10	10JAN2011	1			Bar-Line Chart...				
11	11JAN2011	1			Bubble Plot...				
12	12JAN2011	1			Donut Chart...				
13	13JAN2011	1			Contour Plot...				
14	14JAN2011	1			Box Plot...				
15	15JAN2011	1			Radar Chart...				
16	16JAN2011	1			Surface Plot...				
17	17JAN2011	1			Tile Chart...				
18	18JAN2011	1			Map Chart...				
19	19JAN2011	1			Open ODS Graphics Designer...				
20	20JAN2011	1			Show ODS Statistical Graph...				
21	21JAN2011	1							
22	22JAN2011	1							
23	23JAN2011	1							
24	24JAN2011	1							
25	25JAN2011	1							
26	26JAN2011	1							
27	27JAN2011	1							
28	28JAN2011	1							
29	29JAN2011	1							
30	30JAN2011	1							

ACF, PACF

Example code: Use this PROC TIMESERIES for general ACF, PACF plots

```
ods graphics on;
proc timeseries data=bsta477.bike_sharing_day_data plots=(acf pacf);
var cnt;
run;
```

Example code: Use PROC ARIMA below for specific ACF, PACF lags required.

```
proc arima data=bsta477.bike_sharing_day_data;
identify var=cnt nlag=20;
run;
```

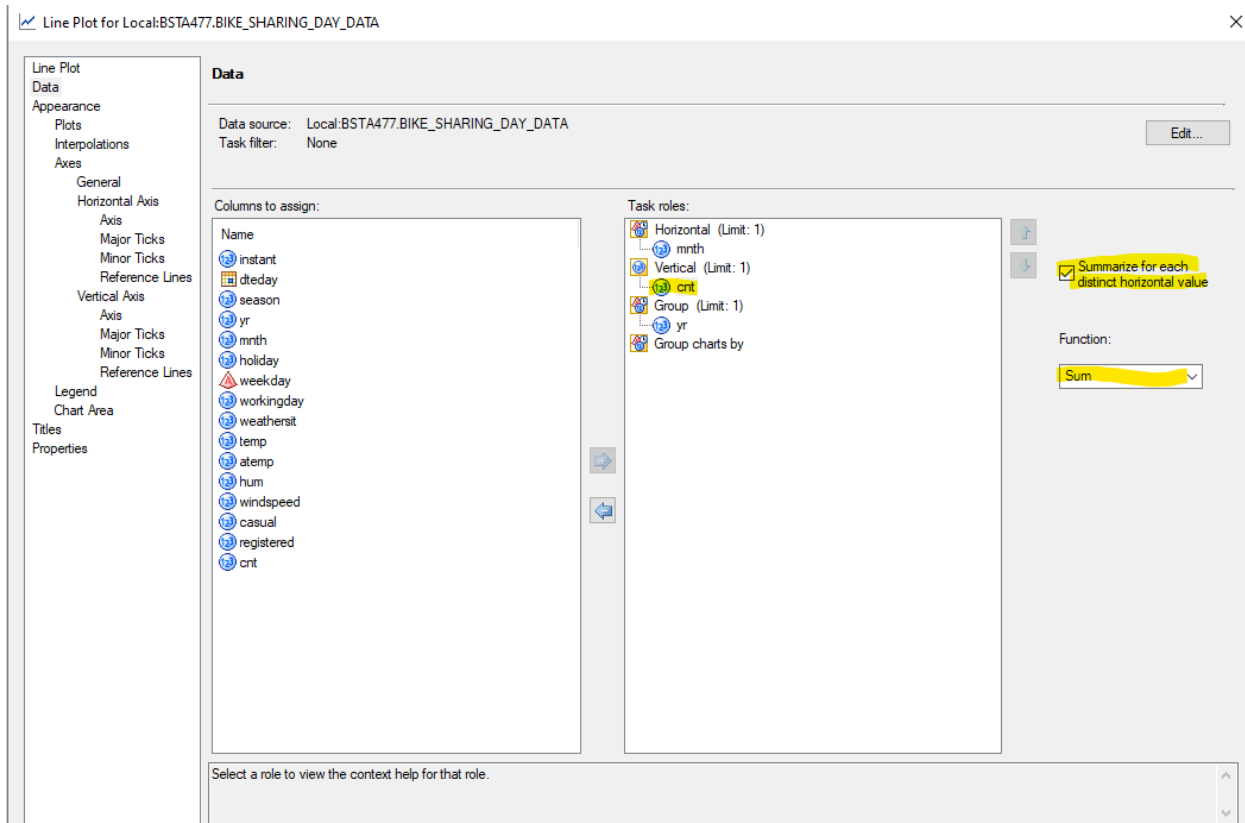
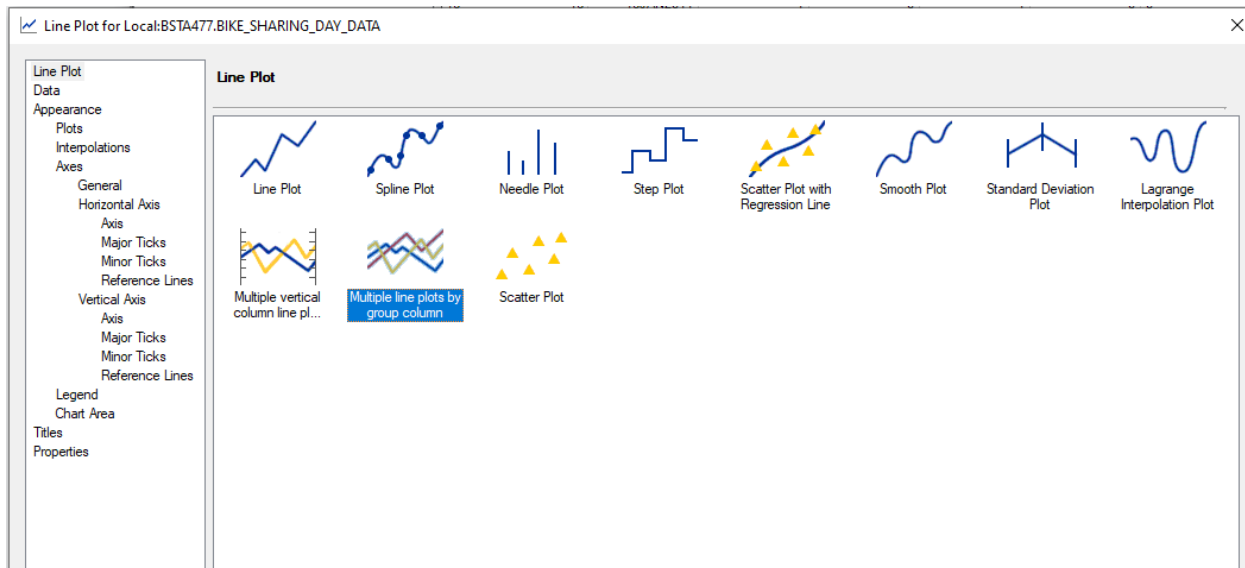
Seasonal plot

Seasonal plots can be created using line plots with information about months and years.

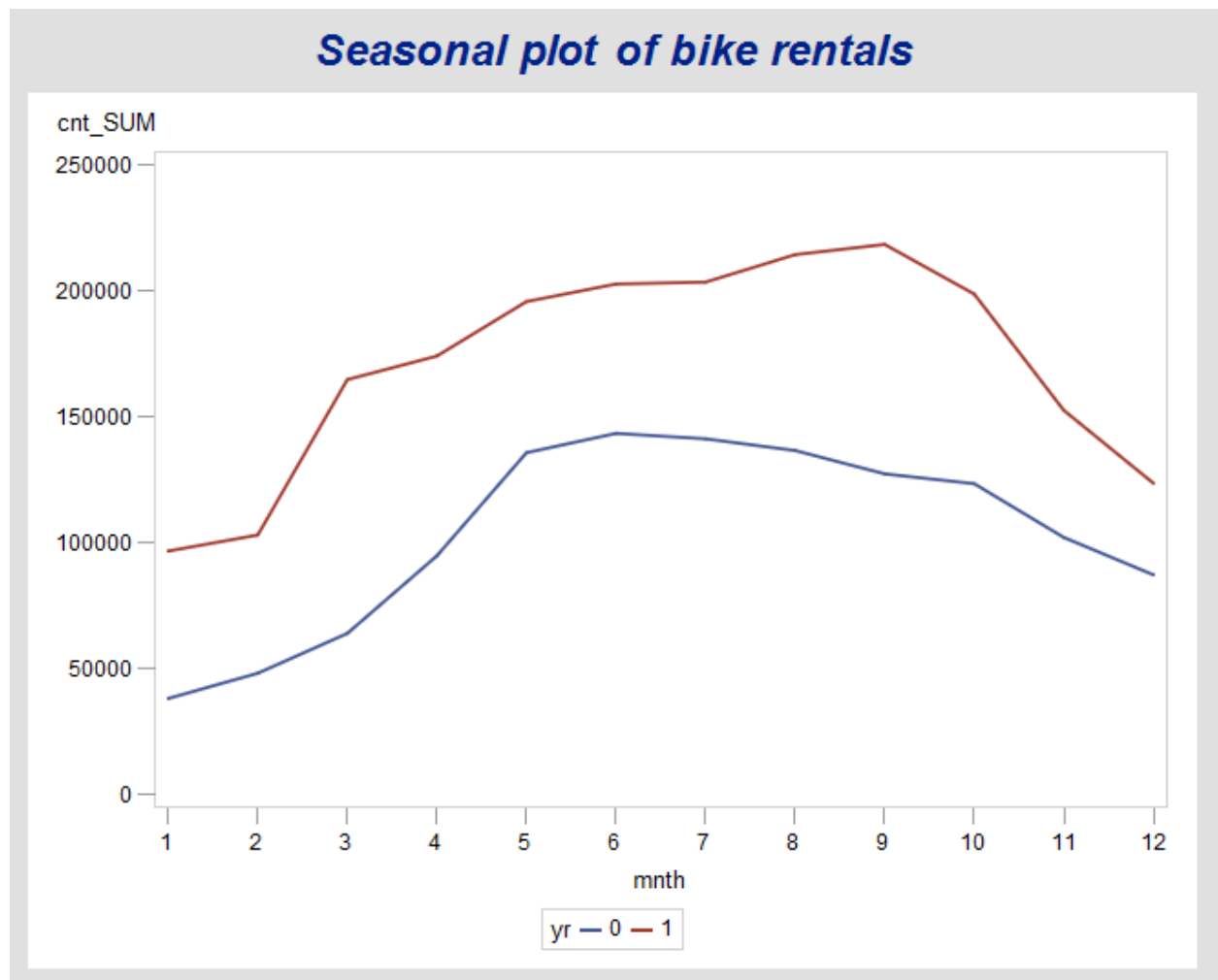
Example: Make seasonal plots to compare bike rentals by month and year.

The screenshot shows the SAS Enterprise Guide interface. At the top, a menu bar includes 'Filter and Sort', 'Query Builder', 'Where', 'Data', 'Describe', 'Graph', 'Analyze', 'Export', and 'Send To'. Below the menu bar is a data table titled 'Data Imported from day.xlsx'. The table has columns: 'instant', 'dteday', 'season', 'yr', and 'weekday'. The 'instant' column is highlighted with a black border. The 'Graph' menu is open, showing options: 'Bar Chart Wizard...', 'Bar Chart...', 'Pie Chart Wizard...', 'Pie Chart...', 'Line Plot Wizard...', 'Line Plot...', 'Scatter Plot...', 'Scatter Plot Matrix...', 'Area Plot...', 'Bar-Line Chart...', 'Bubble Plot...', 'Donut Chart...', 'Contour Plot...', 'Box Plot...', and 'Radar Chart'. The 'Line Plot...' option is selected and highlighted in blue.

	instant	dteday	season	yr	weekday
1	1	01JAN2011	1		
2	2	02JAN2011	1		
3	3	03JAN2011	1		
4	4	04JAN2011	1		
5	5	05JAN2011	1		
6	6	06JAN2011	1		
7	7	07JAN2011	1		
8	8	08JAN2011	1		
9	9	09JAN2011	1		
10	10	10JAN2011	1		
11	11	11JAN2011	1		
12	12	12JAN2011	1		
13	13	13JAN2011	1		
14	14	14JAN2011	1		
15	15	15JAN2011	1		
16	16	16JAN2011	1		
17	17	17JAN2011	1		
18	18	18JAN2011	1		



Here, we put the bike rentals variable (cnt) in vertical axis, month variable (mnth) which is the common time indicator for each year in the horizontal axis, and year variable (yr) as a group value. **Remember to click on the numerical variable to indicate the summarize options.** (Clicked on cnt variable and chose the summarize option)



Year 0 here is 2011, year 1 is 2012.

Time series decomposition

Classical Decomposition

Reference:

- [How to visualize time series decomposition](#)
- [The TIMESERIES procedure](#)

Automatic SAS EG inferred classical decomposition.

Attention:

- Interval variable is the seasonal length. Example: interval = qrt - quarter => m = 4

- The data either has to have the interval frequency or has use accumulate for the operation to work
- Interval options
 - Day => m = 7
 - Hour => m=24
 - Minute => m= 60
 - Second => m=60
 - Week => m=52
 - Month => m=12
 - Year => m=1
 - QRT => m=4

Code example: Decomposing time series with data accumulated in quarterly frequency with m-MA = 4.

```
proc timeseries data=bsta477.bike_sharing_day_data outdecomp=outdecomp plots=(series decomp sc tc cc residual);
  id dteday interval=qtr accumulate=total;
  var cnt;
  decomp orig tcs tcc sic tc sc cc ic / mode=add;
run;
```

SAS Seasonality User manually defined m-MA

Used this option to define the seasonal period m. This option is also used to predict one step ahead using the **Moving average** method. For **Centered moving average method**, one has to calculate manually (in Excel)

SEASONALITY= *number*

specifies the length of the seasonal cycle. For example, SEASONALITY=3 means that every group of three time periods forms a seasonal cycle. By default, the length of the seasonal cycle is one (no seasonality) or the length implied by the INTERVAL= option specified in the ID statement. For example, INTERVAL=MONTH implies that the length of the seasonal cycle is 12.

Notes:

- Do not use Seasonality syntax with Interval syntax above.
- Seasonality syntax does not require the data to have the same frequency (opposed to Interval syntax).

Code example: Use moving average operation to predict one step ahead with MA(4), simply enter seasonality = 4.


```
proc timeseries data=bsta477.bike_sharing_day_data outdecomp=outdecomp_odd plots=(series decomp sc tc cc residual) seasonality=4;
var cnt;
decomp orig tcs tcc sic tc sc cc ic / mode=add;
run;
```

=> The prediction one head ahead is in the Trend Cycle component of the result output. The residuals are the Irregular component of the result output.

X11 Decomposition

Reference: [X11 Decomposition SAS Documentation](#)

X13 Decomposition

Reference: [X13 Decomposition SAS Documentation](#)

Box-cox transformation

Please refer to your textbook for the theory on Box-cox transformation.

To use box-cox transformation, you have to test with all lambda possible to find the best one that creates the best normal distribution curve for the chosen variable.

For lambda selection, SAS automated the process of Box cox transformation. You need: a chosen variable, and a column of zeros.

Ex: Apply box cox transformation to transform bike rentals variable to have normal distribution.

1. Add a zero column

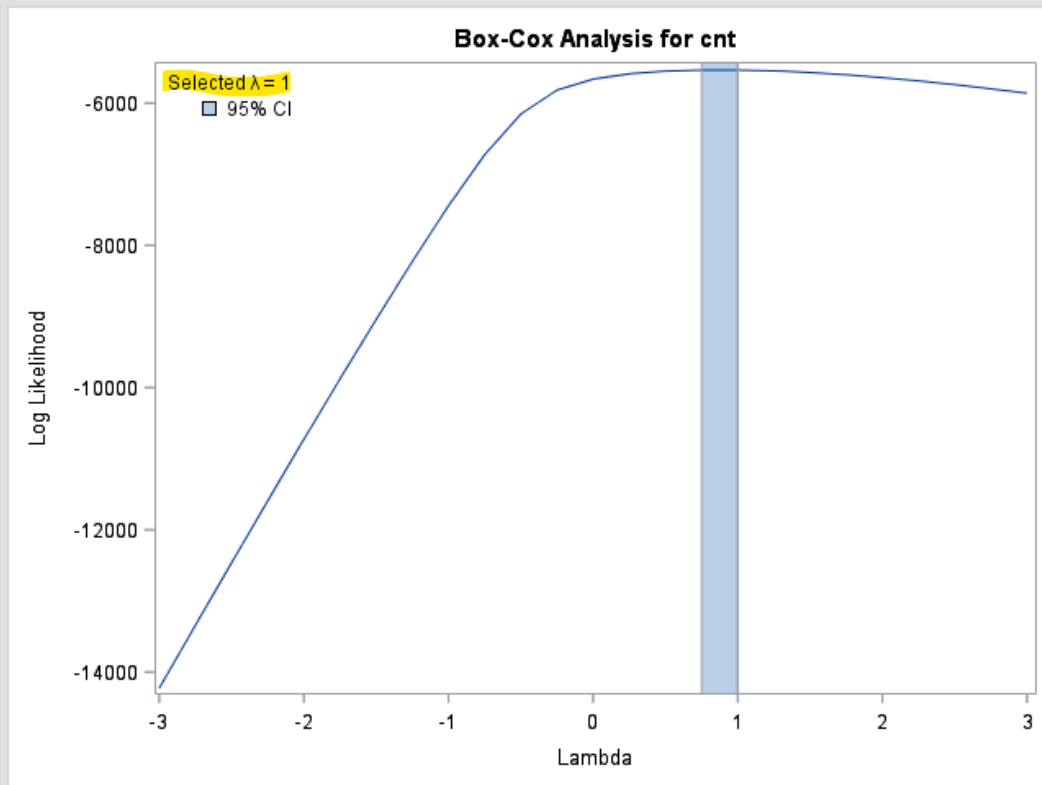
```
data bsta477.bike_sharing_data_day;
set bsta477.bike_sharing_data_day;
zero = 0;
run;
```

2. Use PROC TRANSREG

```
proc transreg data=bsta477.bike_sharing_data_day maxiter=0 nozeroconstant;
model BoxCox(cnt) = identity(zero);
output;
run;
```

3. Check for optimal lambda

The TRANSREG Procedure



SAS Resources

SAS Student hub

- Free SAS e-Learning for Academics for Students:
https://www.sas.com/en_ca/learn/academic-programs/resources/free-sas-e-learning.html#for-students
- SAS Enterprise Guide:
 - Learning:
<https://support.sas.com/en/software/enterprise-guide-support.html#tutorials>
 - Documentation:
<https://documentation.sas.com/?activeCdc=egdoccdc&cdcId=egcdc&cdcVersion=8.3&docsetId=egug&docsetTarget=titlepage.htm&locale=en>