IBM Software

Clustering Course
Lab 1: Analysis of financial market data

Contents

Hello and welcome to first lab!

Let me give you an overview of what we are going to practice in this lab.

First, we are going to download stock market data for S&P 500 Index.
We will then see summary statistics to have an overall idea about our data set.

We will examine return distribution of the index and we will finish the lab with discussing about volatility and have few words about crisis periods.

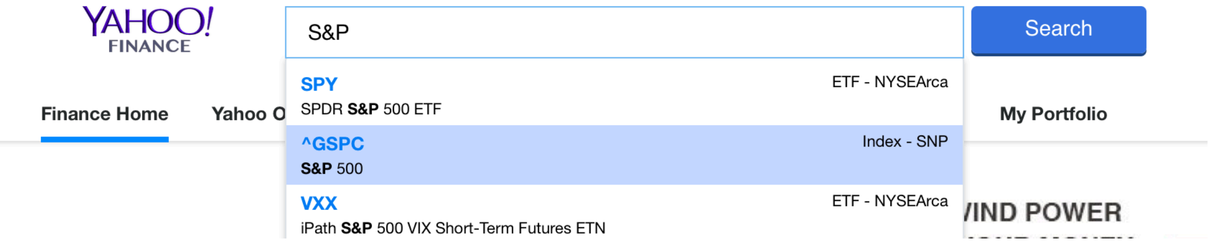
By the end of this lab, you will learn how to use SPSS Modeler for exploratory data analysis.

Let's get started.

1. Downloading daily data for S&P 500 Companies

S&P 500 stands for Standard and Poor's 500, and it's stock market index which tracks largest 500 companies based on their market capitalization, listed on the NYSE and NASDAQ which are American stock exchanges.

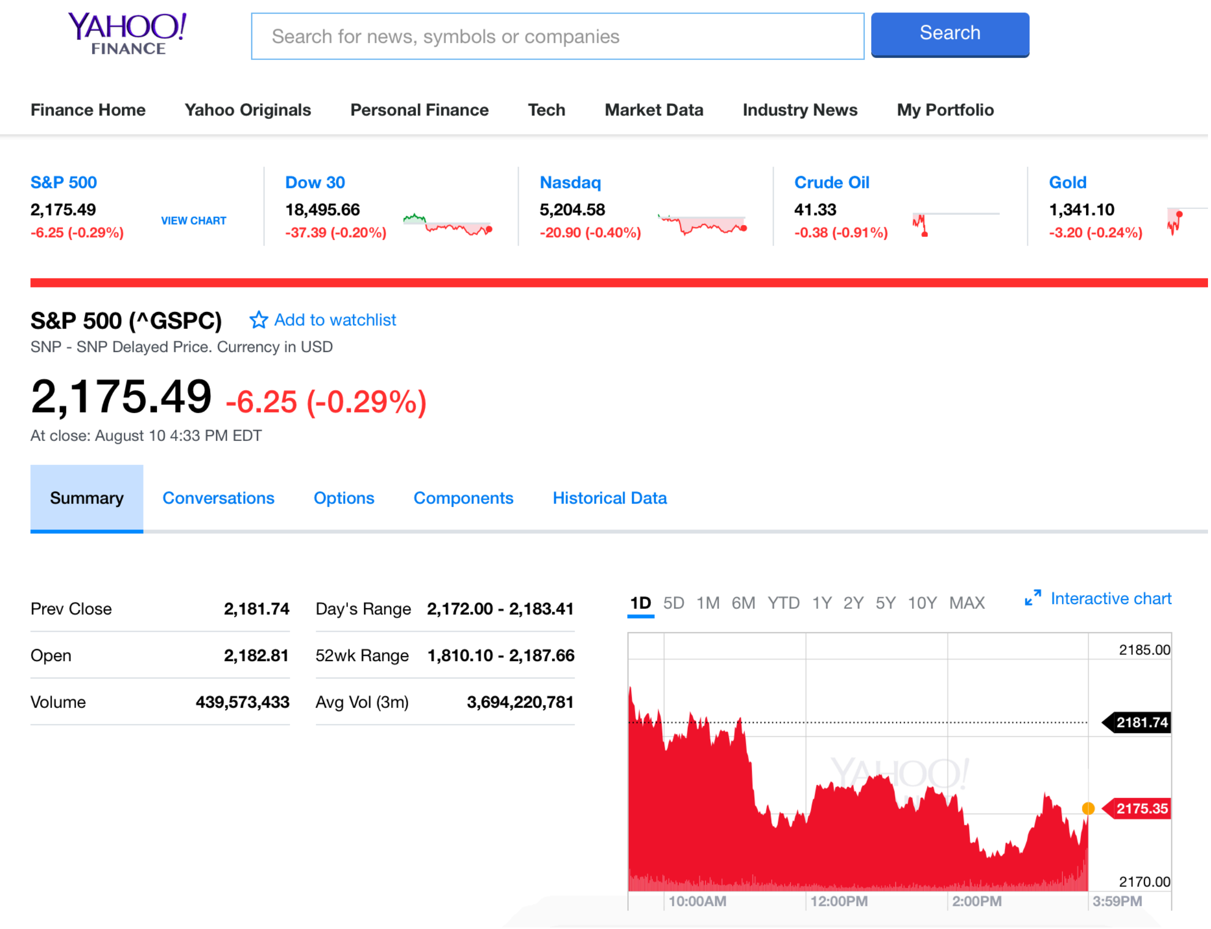
I will open my browser and navigate to finance.yahoo.com. I will type "S&P" to search box and will choose ^GSPC from available symbols.



S&P

SPY SPDR **S&P** 500 ETF	ETF - NYSEArca
^GSPC **S&P** 500	Index - SNP
VXX iPath **S&P** 500 VIX Short-Term Futures ETN	ETF - NYSEArca

I will be presented with following page which gives us S&P 500 Index specific information.



S&P 500	Dow 30	Nasdaq	Crude Oil	Gold
2,175.49	18,495.66	5,204.58	41.33	1,341.10
-6.25 (-0.29%)	-37.39 (-0.20%)	-20.90 (-0.40%)	-0.38 (-0.91%)	-3.20 (-0.24%)

S&P 500 (^GSPC)

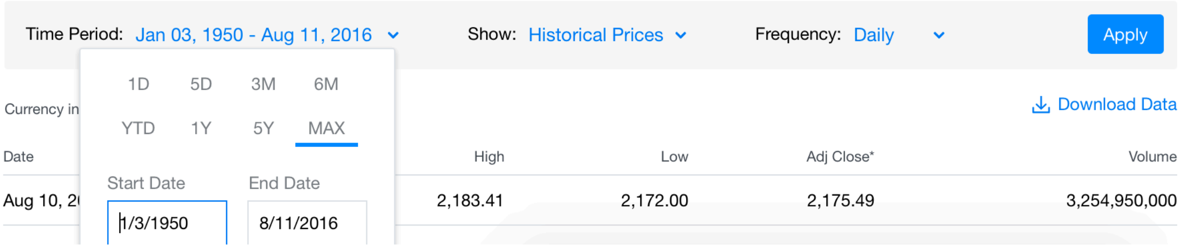
SNP - SNP Delayed Price. Currency in USD

2,175.49 -6.25 (-0.29%)

At close: August 10 4:33 PM EDT

Prev Close	2,181.74	Day's Range	2,172.00 - 2,183.41
Open	2,182.81	52wk Range	1,810.10 - 2,187.66
Volume	439,573,433	Avg Vol (3m)	3,694,220,781

I will click on "Historical Data", in this page, I will adjust time range of dataset by selecting "MAX" and I will click "Apply" and then "Download Data"



Time Period: Jan 03, 1950 - Aug 11, 2016

Show: Historical Prices

Frequency: Daily

Start Date: 1/3/1950

End Date: 8/11/2016

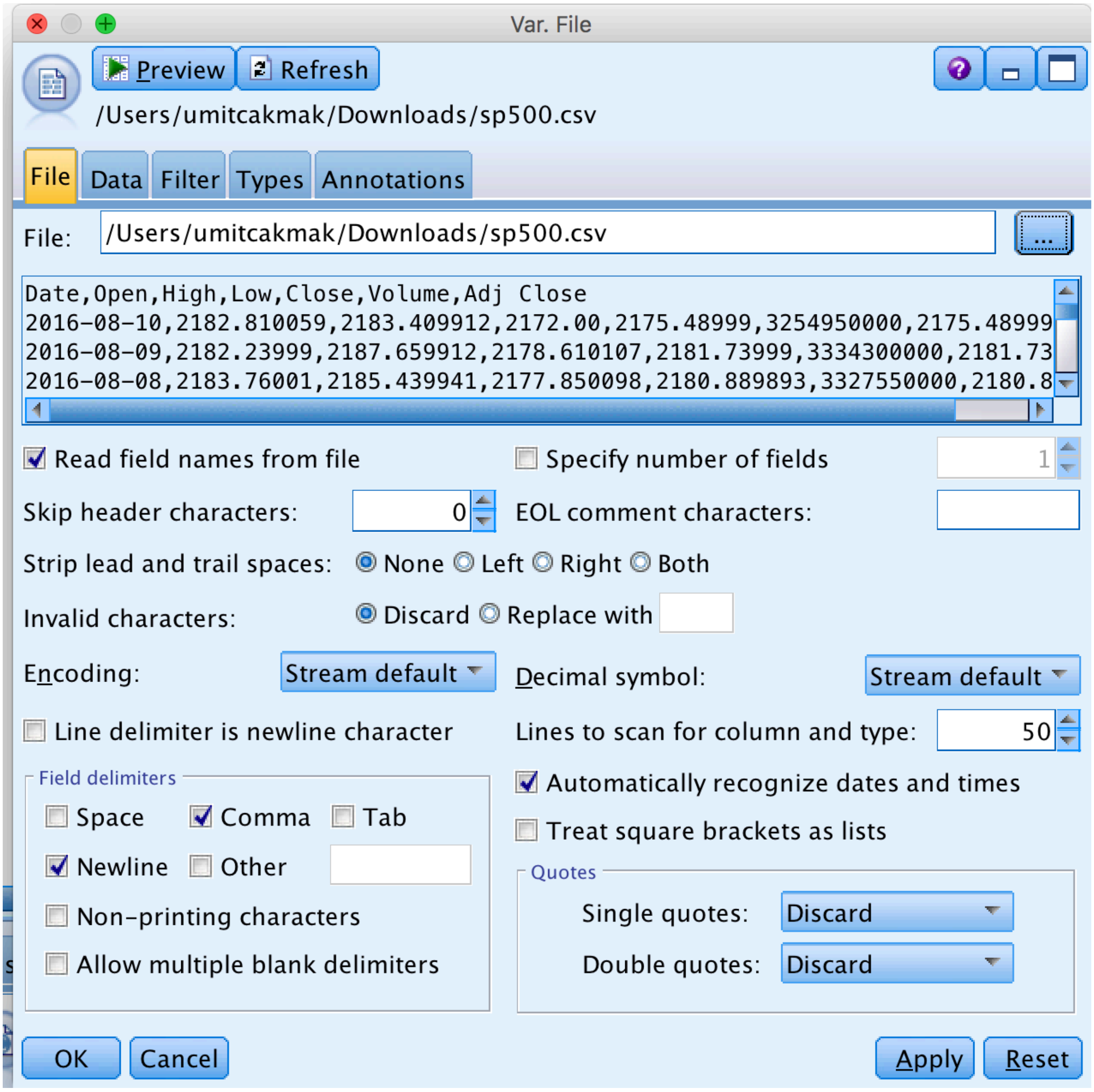
Date	High	Low	Adj Close*	Volume
Aug 10, 2	2,183.41	2,172.00	2,175.49	3,254,950,000

What I will have is comma separated file with S&P 500 Index data. I will rename file as “sp500.csv”

I will open SPSS Modeler and save stream as “Lab_1” into my lab folder.

From “Sources” palette, I will add “Var. File” node to import sp500.csv file.

My import settings looks like following and these are default settings when you add “Var. File” node into your stream.



/Users/umitcakmak/Downloads/sp500.csv

File: /Users/umitcakmak/Downloads/sp500.csv

```
Date,Open,High,Low,Close,Volume,Adj Close
2016-08-10,2182.810059,2183.409912,2172.00,2175.48999,3254950000,2175.48999
2016-08-09,2182.23999,2187.659912,2178.610107,2181.73999,3334300000,2181.73
2016-08-08,2183.76001,2185.439941,2177.850098,2180.889893,3327550000,2180.8
```

☑ Read field names from file

□ Specify number of fields: 1

Skip header characters: 0

EOL comment characters:

Strip lead and trail spaces: ● None ○ Left ○ Right ○ Both

Invalid characters: ● Discard ○ Replace with

Encoding: Stream default

Decimal symbol: Stream default

□ Line delimiter is newline character

Lines to scan for column and type: 50

Field delimiters

□ Space ☑ Comma □ Tab

☑ Newline □ Other

□ Non-printing characters

□ Allow multiple blank delimiters

☑ Automatically recognize dates and times

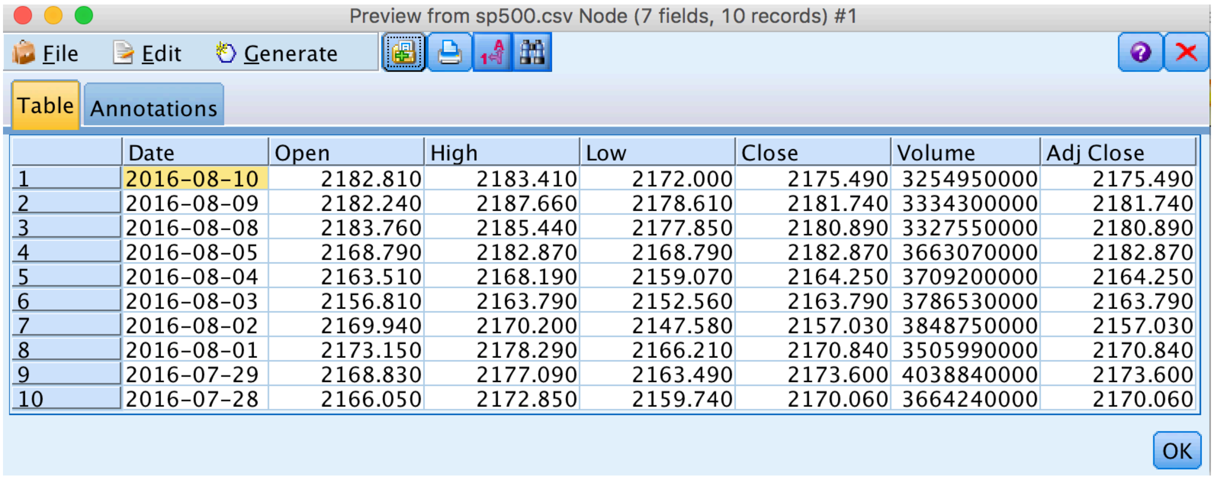
□ Treat square brackets as lists

Quotes

Single quotes: Discard

Double quotes: Discard

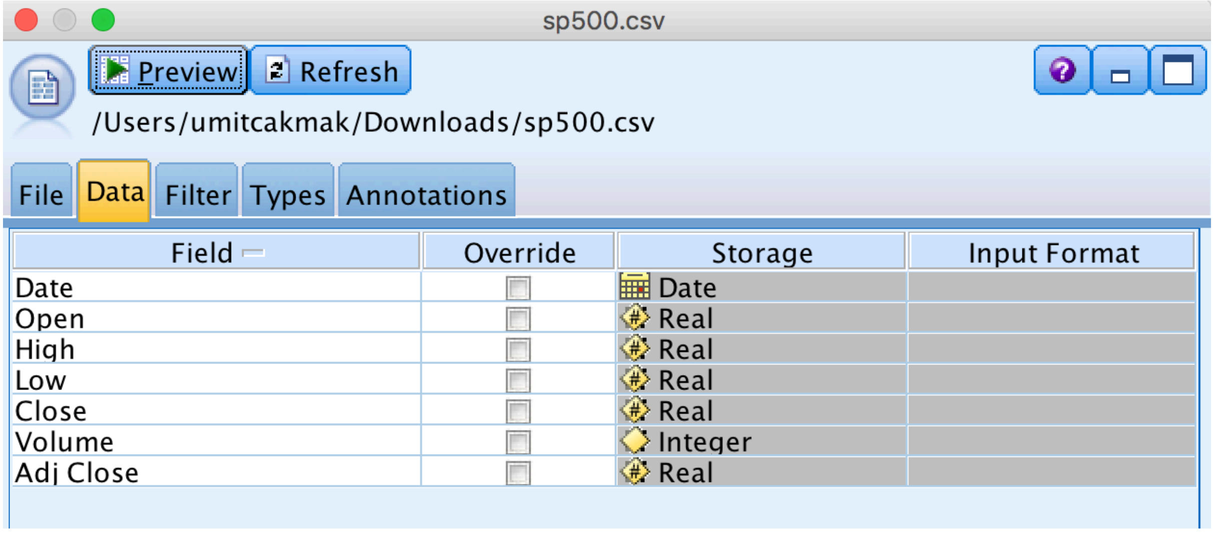
You can click “Preview” to see first 10 records of dataset.



Preview from sp500.csv Node (7 fields, 10 records) #1

	Date	Open	High	Low	Close	Volume	Adj Close
1	2016-08-10	2182.810	2183.410	2172.000	2175.490	3254950000	2175.490
2	2016-08-09	2182.240	2187.660	2178.610	2181.740	3334300000	2181.740
3	2016-08-08	2183.760	2185.440	2177.850	2180.890	3327550000	2180.890
4	2016-08-05	2168.790	2182.870	2168.790	2182.870	3663070000	2182.870
5	2016-08-04	2163.510	2168.190	2159.070	2164.250	3709200000	2164.250
6	2016-08-03	2156.810	2163.790	2152.560	2163.790	3786530000	2163.790
7	2016-08-02	2169.940	2170.200	2147.580	2157.030	3848750000	2157.030
8	2016-08-01	2173.150	2178.290	2166.210	2170.840	3505990000	2170.840
9	2016-07-29	2168.830	2177.090	2163.490	2173.600	4038840000	2173.600
10	2016-07-28	2166.050	2172.850	2159.740	2170.060	3664240000	2170.060

You can also see in “Data” tab, field types of the data set, and we can see that field types are correctly identified for all fields.

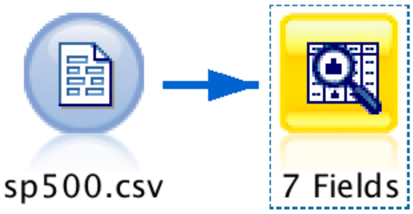


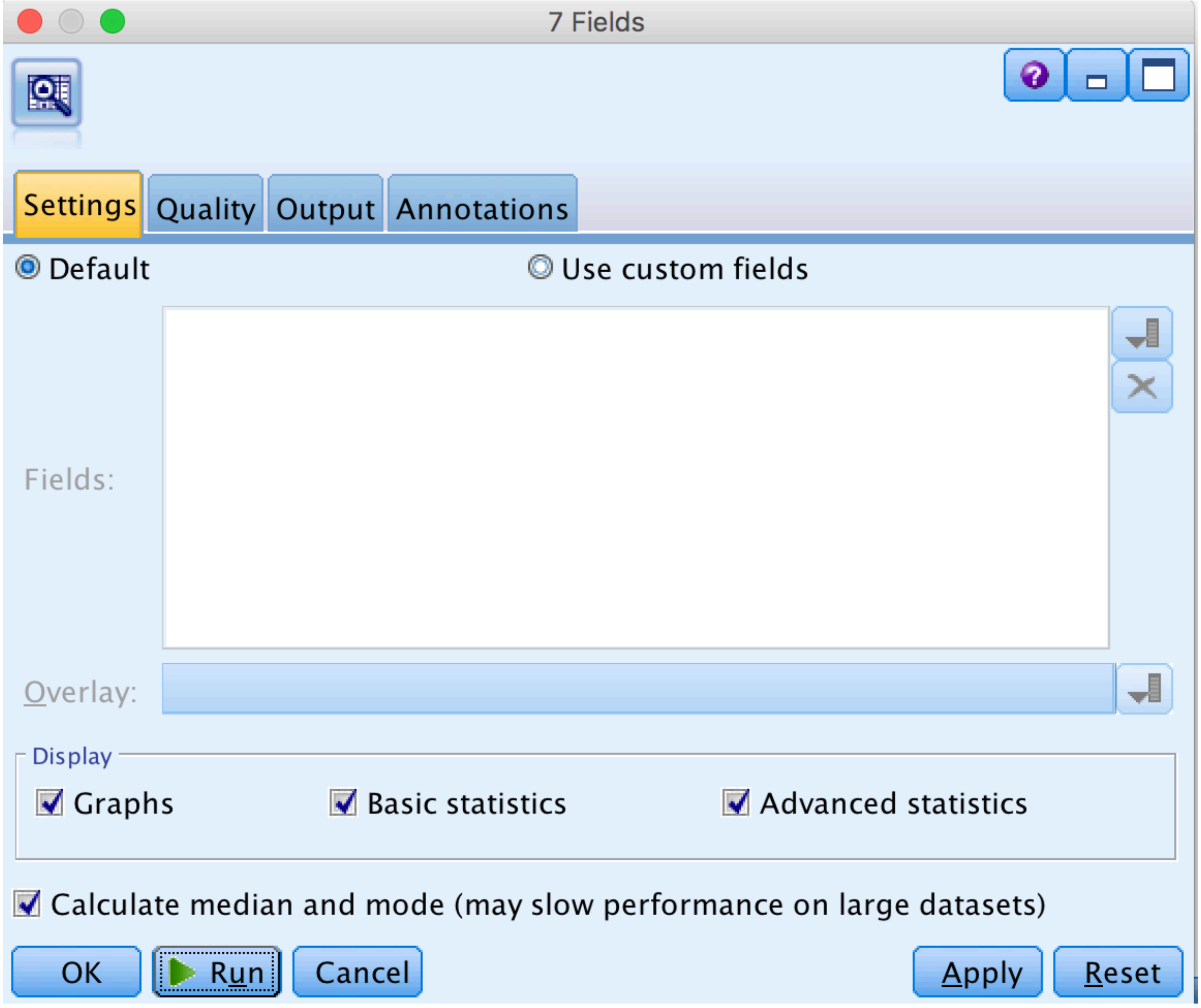
Field	Override	Storage	Input Format
Date	□	Date	
Open	□	Real	
High	□	Real	
Low	□	Real	
Close	□	Real	
Volume	□	Integer	
Adj Close	□	Real	

2. Summary Statistics

Before we can continu with summary statistics we can do a little “Data Audit” to have general feel about our dataset.

From “Output” palette, add a “Data Audit” node, double click to open it, check “Advanced Statistics” box and “Calculate Median and Mode” as well and and click “Run”





● Default ○ Use custom fields

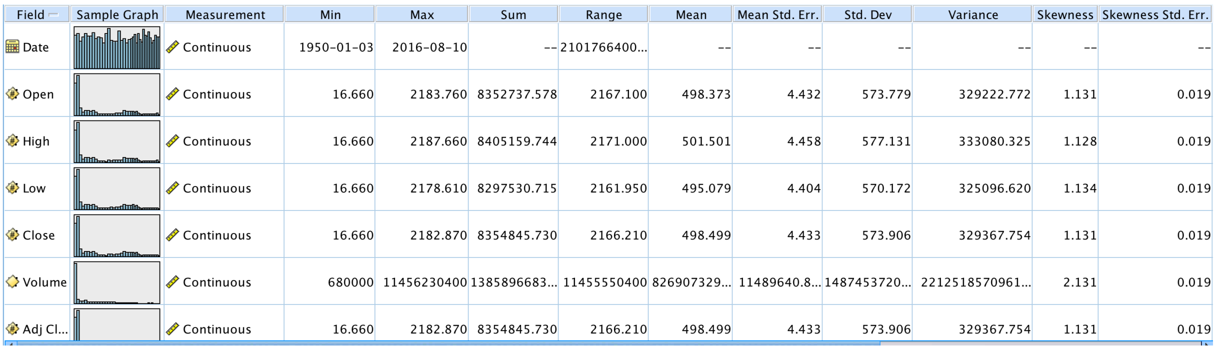
Fields:

Overlay:

Display

☑ Graphs ☑ Basic statistics ☑ Advanced statistics

☑ Calculate median and mode (may slow performance on large datasets)

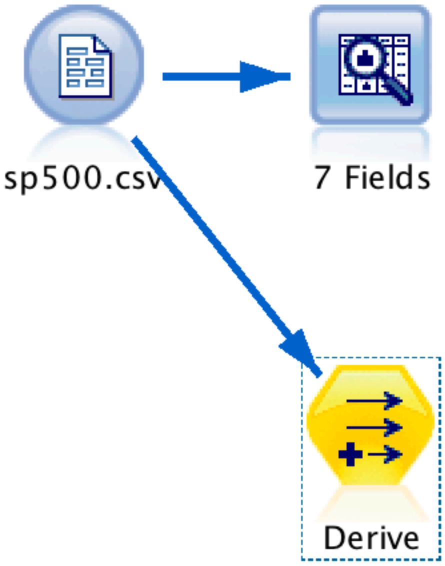


Field	Sample Graph	Measurement	Min	Max	Sum	Range	Mean	Mean Std. Err.	Std. Dev	Variance	Skewness	Skewness Std. Err.
Date		Continuous	1950-01-03	2016-08-10	--	2101766400...	--	--	--	--	--	--
Open		Continuous	16.660	2183.760	8352737.578	2167.100	498.373	4.432	573.779	329222.772	1.131	0.019
High		Continuous	16.660	2187.660	8405159.744	2171.000	501.501	4.458	577.131	333080.325	1.128	0.019
Low		Continuous	16.660	2178.610	8297530.715	2161.950	495.079	4.404	570.172	325096.620	1.134	0.019
Close		Continuous	16.660	2182.870	8354845.730	2166.210	498.499	4.433	573.906	329367.754	1.131	0.019
Volume		Continuous	680000	11456230400	1385896683...	11455550400	826907329...	11489640.8...	1487453720...	2212518570961...	2.131	0.019
Adj Cl...		Continuous	16.660	2182.870	8354845.730	2166.210	498.499	4.433	573.906	329367.754	1.131	0.019

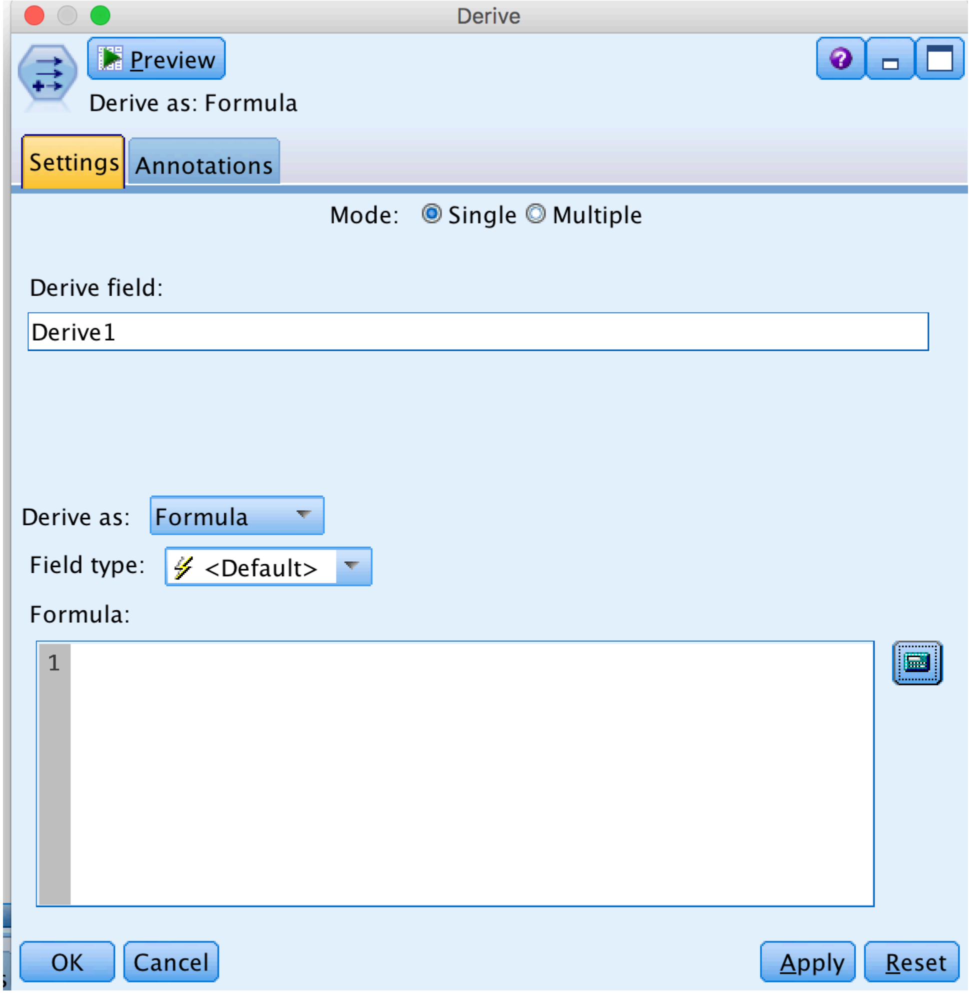
We can see main properties of dataset such as minimum and maximum values, mean, standard deviation, variance, skewness and kurtosis which will be important when we are examining return series.

3. Return distributions

In order to see return distributions, we can calculate daily return based on “Adj Close” field by using “Derive” node.



“Derive” node allows us to use CLEM expressions which are set of ready-to-use libraries which you can use to manipulate your data set.



Derive as: Formula

Mode: ● Single ○ Multiple

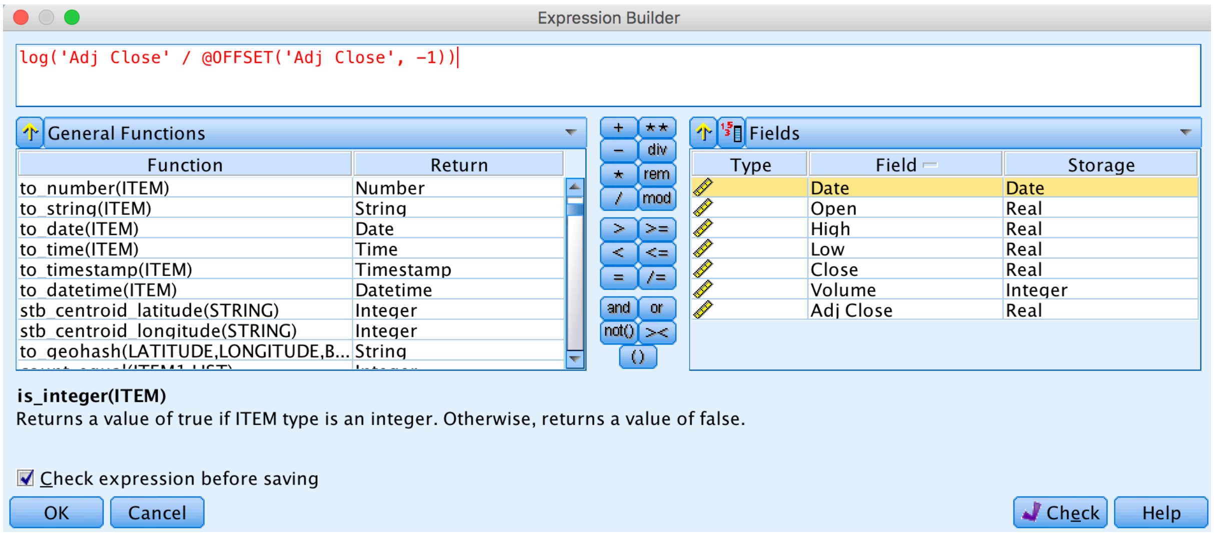
Derive field: Derive1

Derive as: Formula

Field type: <Default>

Formula:

You can click on this button to Launch Expression Builder and see available function also check syntax of your expressions before running them.



```
log('Adj Close' / @OFFSET('Adj Close', -1))
```

Function	Return
to_number(ITEM)	Number
to_string(ITEM)	String
to_date(ITEM)	Date
to_time(ITEM)	Time
to_timestamp(ITEM)	Timestamp
to_datetime(ITEM)	Datetime
stb_centroid_latitude(STRING)	Integer
stb_centroid_longitude(STRING)	Integer
to_geohash(LATITUDE,LONGITUDE,B...	String

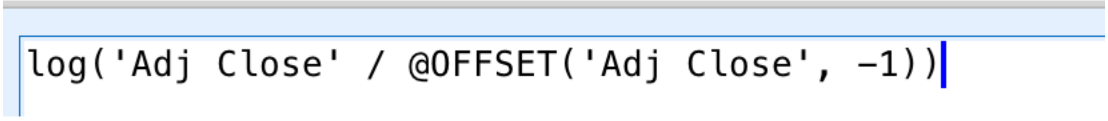
Type	Field	Storage
	Date	Date
	Open	Real
	High	Real
	Low	Real
	Close	Real
	Volume	Integer
	Adj Close	Real

is_integer(ITEM)
Returns a value of true if ITEM type is an integer. Otherwise, returns a value of false.

☑ Check expression before saving

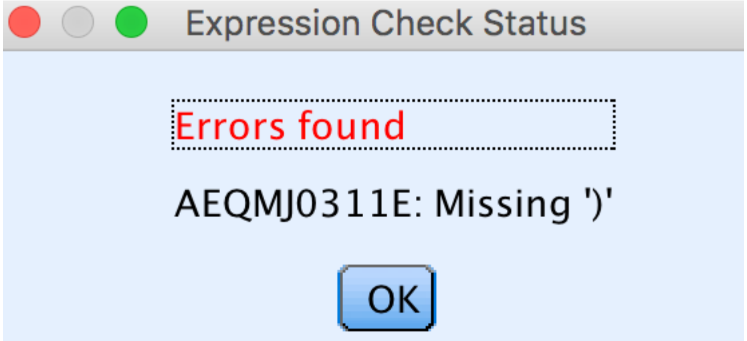
In this case I will use “log” and “OFFSET” function to calculate log returns based on “Adj Close” field.

In text box above, I write my expression and click "Check", if there are no errors, expression color will turn to black



```
log('Adj Close' / @OFFSET('Adj Close', -1))
```

if I delete last paranthesis, then click "Check" again, it will give me an error.

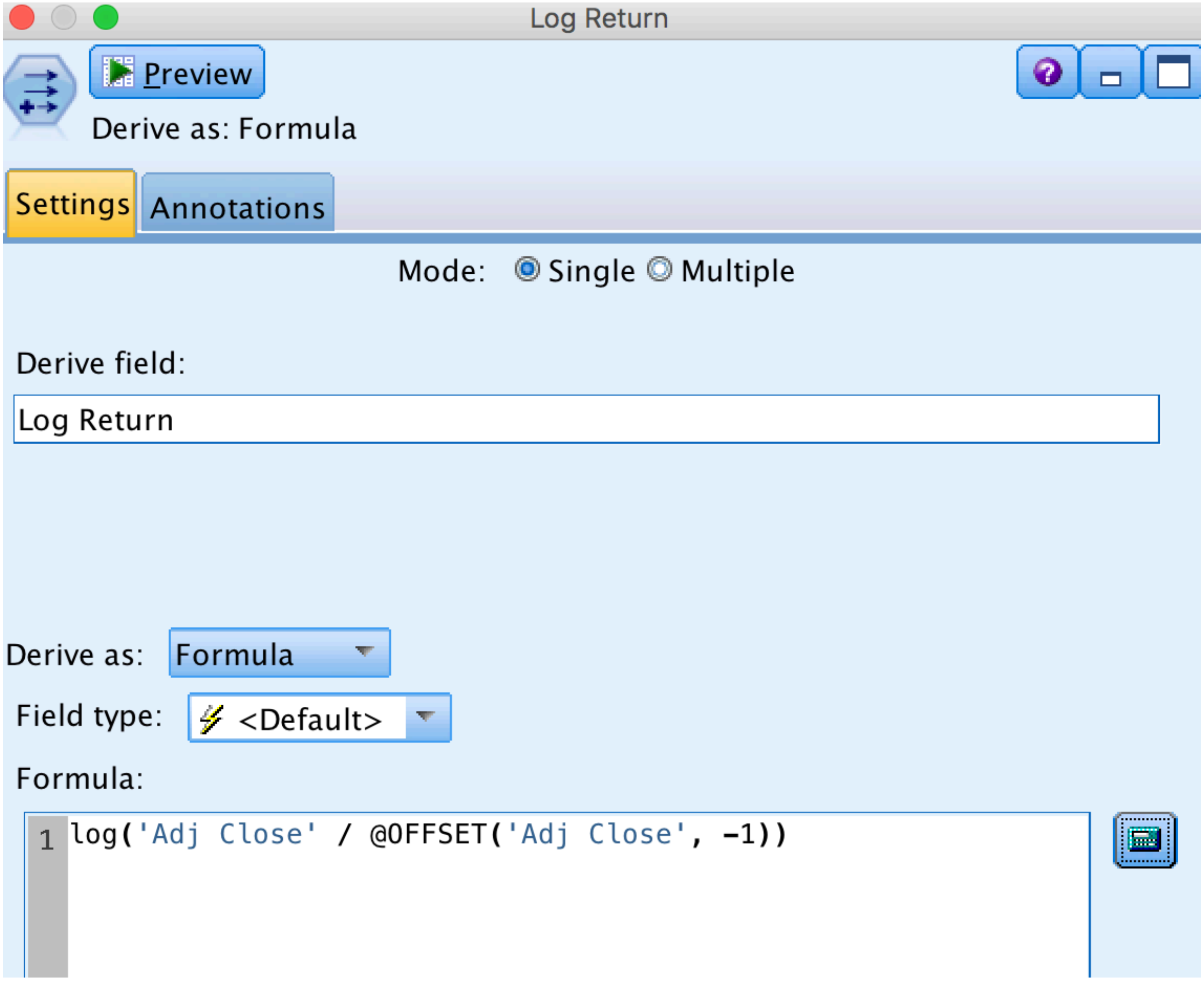


Expression Check Status

Errors found

AEQMJ0311E: Missing ')'

I can click "OK", and set name for this field in "Derive field:" text box and click "OK"



Log Return

Derive as: Formula

Mode: ● Single ○ Multiple

Derive field: Log Return

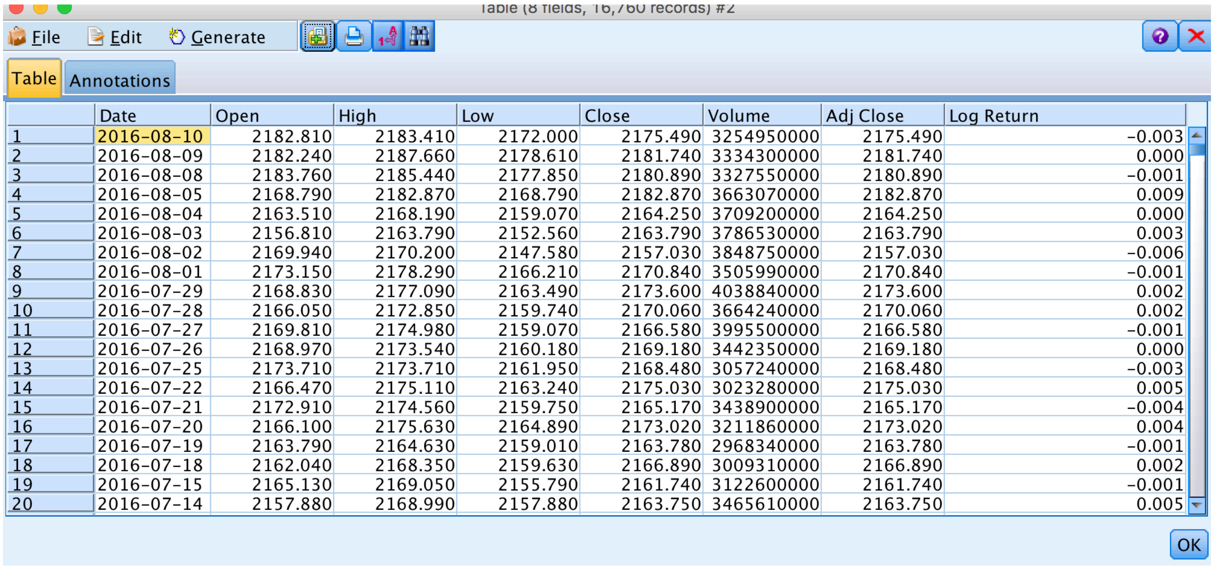
Derive as: Formula

Field type: <Default>

Formula:

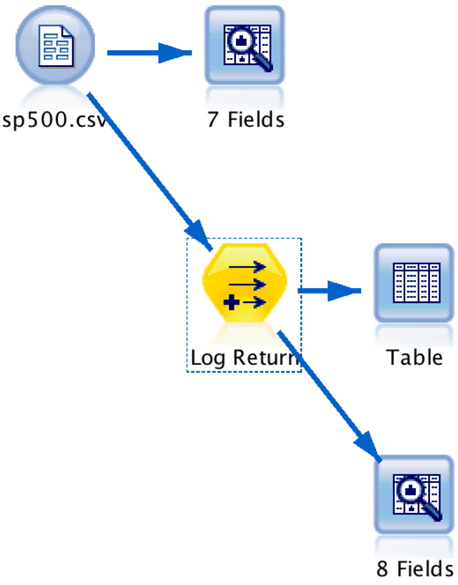
```
log('Adj Close' / @OFFSET('Adj Close', -1))
```

I can add “Table” node from “Output” palette to see my derived field “Log Return”

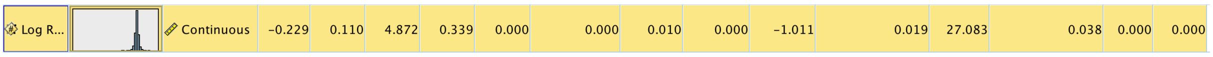


	Date	Open	High	Low	Close	Volume	Adj Close	Log Return
1	2016-08-10	2182.810	2183.410	2172.000	2175.490	3254950000	2175.490	-0.003
2	2016-08-09	2182.240	2187.660	2178.610	2181.740	3334300000	2181.740	0.000
3	2016-08-08	2183.760	2185.440	2177.850	2180.890	3327550000	2180.890	-0.001
4	2016-08-05	2168.790	2182.870	2168.790	2182.870	3663070000	2182.870	0.009
5	2016-08-04	2163.510	2168.190	2159.070	2164.250	3709200000	2164.250	0.000
6	2016-08-03	2156.810	2163.790	2152.560	2163.790	3786530000	2163.790	0.003
7	2016-08-02	2169.940	2170.200	2147.580	2157.030	3848750000	2157.030	-0.006
8	2016-08-01	2173.150	2178.290	2166.210	2170.840	3505990000	2170.840	-0.001
9	2016-07-29	2168.830	2177.090	2163.490	2173.600	4038840000	2173.600	0.002
10	2016-07-28	2166.050	2172.850	2159.740	2170.060	3664240000	2170.060	0.002
11	2016-07-27	2169.810	2174.980	2159.070	2166.580	3995500000	2166.580	-0.001
12	2016-07-26	2168.970	2173.540	2160.180	2169.180	3442350000	2169.180	0.000
13	2016-07-25	2173.710	2173.710	2161.950	2168.480	3057240000	2168.480	-0.003
14	2016-07-22	2166.470	2175.110	2163.240	2175.030	3023280000	2175.030	0.005
15	2016-07-21	2172.910	2174.560	2159.750	2165.170	3438900000	2165.170	-0.004
16	2016-07-20	2166.100	2175.630	2164.890	2173.020	3211860000	2173.020	0.004
17	2016-07-19	2163.790	2164.630	2159.010	2163.780	2968340000	2163.780	-0.001
18	2016-07-18	2162.040	2168.350	2159.630	2166.890	3009310000	2166.890	0.002
19	2016-07-15	2165.130	2169.050	2155.790	2161.740	3122600000	2161.740	-0.001
20	2016-07-14	2157.880	2168.990	2157.880	2163.750	3465610000	2163.750	0.005

I can add “Data Audit” Node here again to see main properties for “Log Return” field.

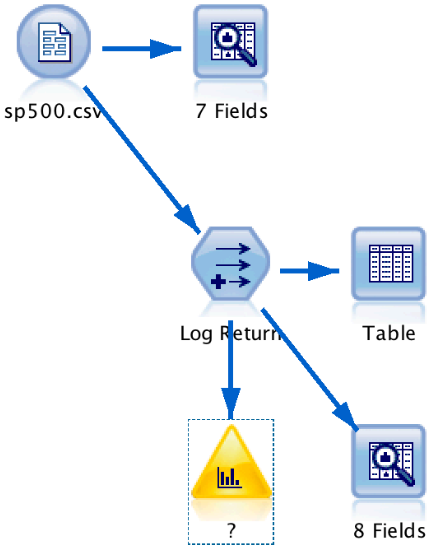


We can see here minimum return and maximum return. Notice how index lost almost 23% of it’s value in just one day.

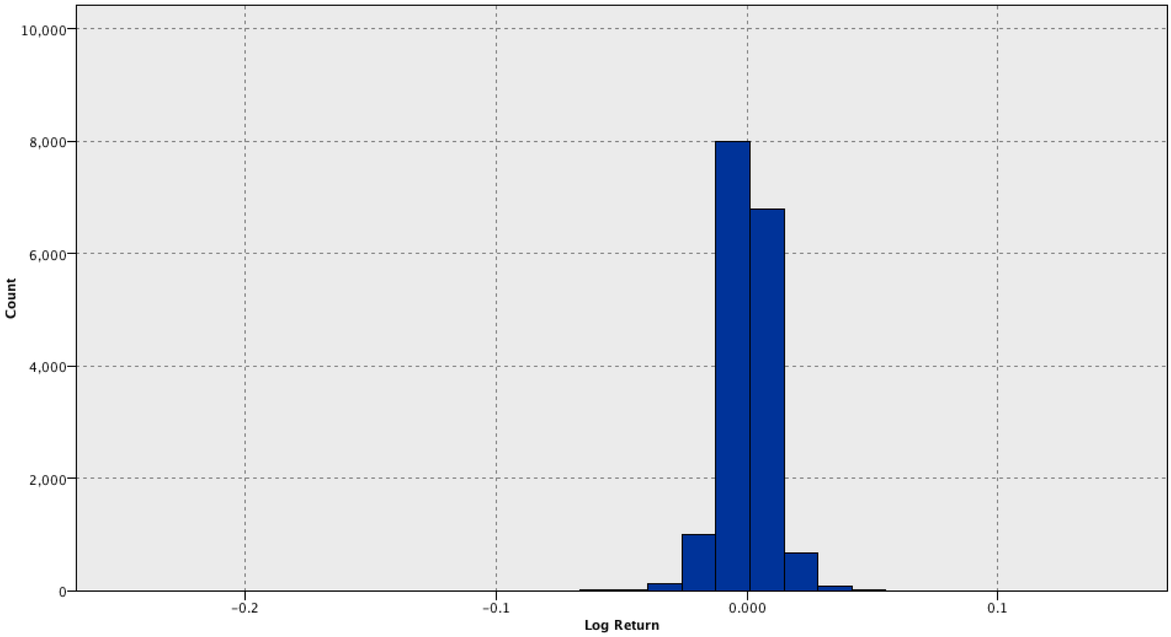


Log R...		Continuous	-0.229	0.110	4.872	0.339	0.000	0.000	0.010	0.000	-1.011	0.019	27.083	0.038	0.000	0.000

Now I can see return distribution of S&P 500 Index by adding “Histogram” node from “Graphs” palette.



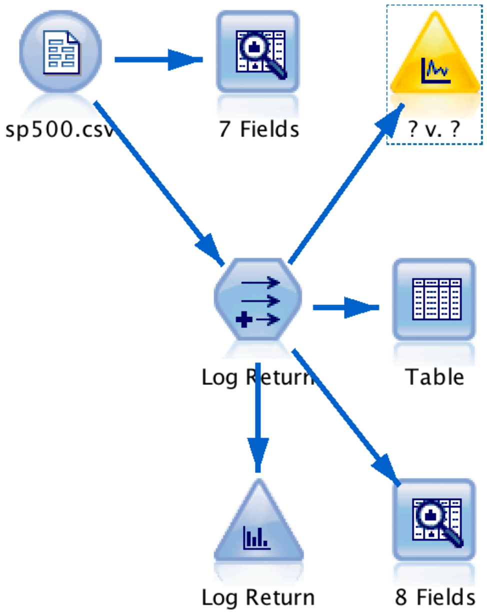
Double click on “Histogram” to open it and only thing you need to do here is selecting “Log Return” field.



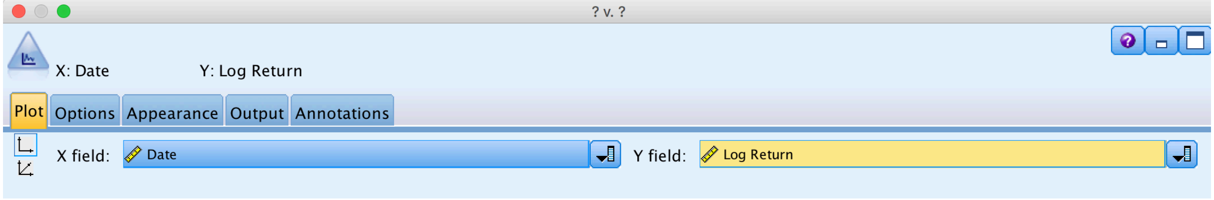
Histogram shows us distribution of daily log returns and notice extreme tails of the distribution.

4. Plotting return series

If we would like to see dates for those extreme events we can use “Plot” node.



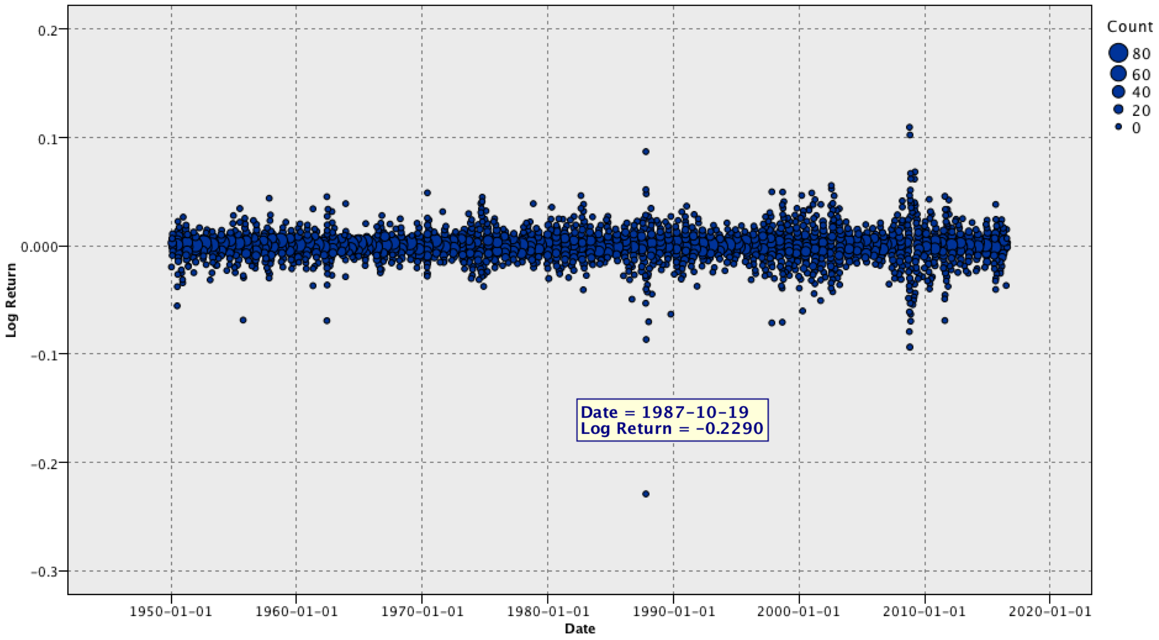
Set field selection as I do and click run



X: Date Y: Log Return

X field: Date Y field: Log Return

We can see that 23% daily loss happened on 19th October, 1987. This event is referred as “Black Monday” where markets recorded huge losses.



Summary

In this lab, you have learned basic functionalities of SPSS Modeler to import and work with your dataset.

Thank you and hope to see you next time.