

IBM Software

Clustering Course

Lab 2: Portfolio construction

Contents

Constructing risk efficient portfolios

1. Downloading historical data for 3 companies and S&P 500 Index
2. Data preparation
3. Calculating Log Return Series for All Datasets
4. Markowitz Portfolio Optimization

Summary

Hello everybody! Welcome to second lab of this series.

We are already on our way to construct our first investment portfolio. Our portfolio will consist of 3 stocks.

The question is, if we decided to invest in these 3 stocks, how we should allocate our capital?

Markowitz portfolio theory has an answer for the problem. We will see how we can implement markowitz portfolio theory to create efficient portfolios.

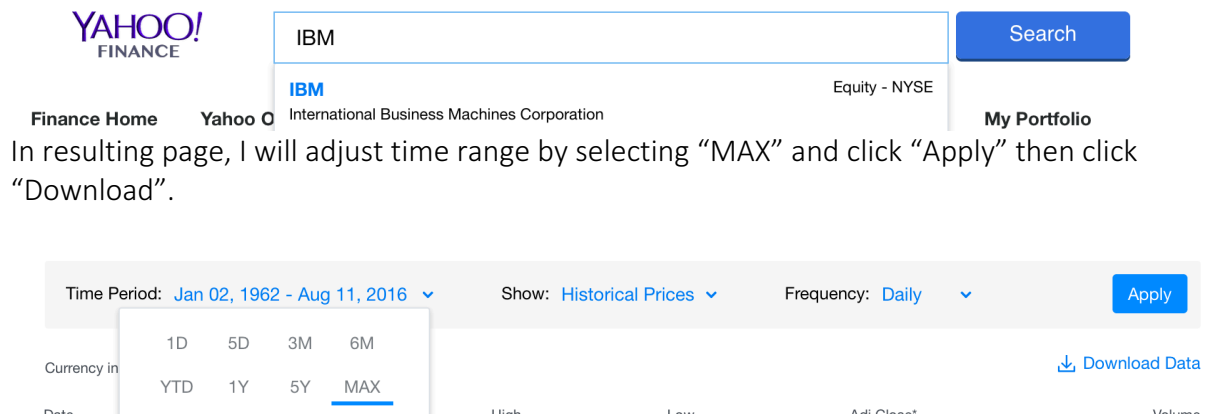
Let's get started.

1. Downloading historical data for 3 companies and S&P 500 Index

I will open my browser and navigate to finance.yahoo.com

I will download historical data for IBM, Procter & Gamble and Goldman Sachs.

I will type "IBM" in search box and will select IBM.



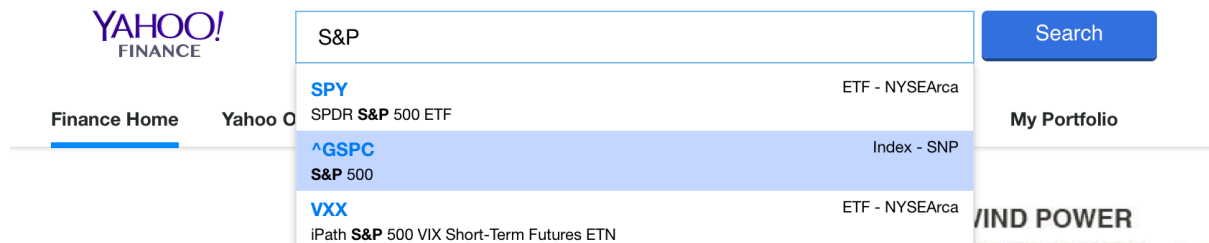
The screenshot shows the Yahoo Finance search interface. The search bar contains "IBM". Below the search bar, the results for "IBM" are displayed, including the company name "International Business Machines Corporation" and its ticker "IBM". The "Show" dropdown is set to "Historical Prices", and the "Frequency" dropdown is set to "Daily". The "Time Period" is set to "Jan 02, 1962 - Aug 11, 2016". The "Download Data" button is visible.

I will rename file as "IBM.csv"

I will go through same steps and download data for Procter & Gamble and Goldman Sachs as well. I will rename files as "PG.csv" and "GS.csv".

In CAPM, we use market return and for U.S. stock markets we can use S&P 500 Index for that purpose.

I will download historical data for S&P 500 Index and rename it as "sp500.csv"



The screenshot shows the Yahoo Finance search interface. The search bar contains "S&P". Below the search bar, the results for "S&P" are displayed, including the ticker "SPY" for "SPDR S&P 500 ETF", the ticker "^GSPC" for "S&P 500", and the ticker "VXX" for "iPath S&P 500 VIX Short-Term Futures ETN". The "Download Data" button is visible.

2. Data preparation

I will open SPSS Modeler and create a new stream, I will save stream as “Lab_2” to my labs folder.

I will add 4 “Var. File” node from “Sources” palette and import downloaded datasets to my stream.



Var. File



Var. File



Var. File



Var. File

First IBM,

Var. File

Preview Refresh

/Users/umitcakmak/Desktop/DataSets/IBM.csv

File Data Filter Types Annotations

File: /Users/umitcakmak/Desktop/DataSets/IBM.csv

Date,Open,High,Low,Close,Volume,Adj Close
2016-09-12,155.259995,158.529999,154.839996,158.289993,4318400,158.289993
2016-09-09,158.029999,158.399994,155.649994,155.690002,5186000,155.690002
2016-09-08,160.550003,161.210007,158.759995,159.00,3963200,159.00

☒ 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

OK Cancel Apply Reset

Second Apple,

Var. File

Preview Refresh

/Users/umitcakmak/Desktop/DataSets/AAPL.csv

File Data Filter Types Annotations

File: /Users/umitcakmak/Desktop/DataSets/AAPL.csv

Date,Open,High,Low,Close,Volume,Adj Close
2016-09-12,102.650002,105.720001,102.529999,105.440002,44802300,105.440002
2016-09-09,104.639999,105.720001,103.129997,103.129997,46557000,103.129997
2016-09-08,107.25,107.269997,105.239998,105.519997,53002000,105.519997

☒ 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

OK Cancel Apply Reset

Third Google,

Var. File

Preview

Refresh

?

/Users/umitcakmak/Desktop/DataSets/GOOG.csv

File

Data

Filter

Types

Annotations

File: /Users/umitcakmak/Desktop/DataSets/GOOG.csv

Date,Open,High,Low,Close,Volume,Adj Close

2016-09-12,755.130005,770.289978,754.00,769.02002,1286800,769.02002

2016-09-09,770.099976,773.244995,759.659973,759.659973,1812200,759.659973

2016-09-08,778.590027,780.349976,773.580017,775.320007,1260600,775.320007

☒ 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

OK

Cancel

Apply

Reset

And finally, S&P 500 Index

Var. File

Preview Refresh

/Users/umitcakmak/Desktop/DataSets/SP500.csv

File Data Filter Types Annotations

File: /Users/umitcakmak/Desktop/DataSets/SP500.csv

Date,Open,High,Low,Close,Volume,Adj Close
2016-09-12,2120.860107,2163.300049,2119.120117,2159.040039,4010480000,2159.
2016-09-09,2169.080078,2169.080078,2127.810059,2127.810059,4233960000,2127.
2016-09-08,2182.76001,2184.939941,2177.48999,2181.300049,3727840000,2181.30

☒ 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

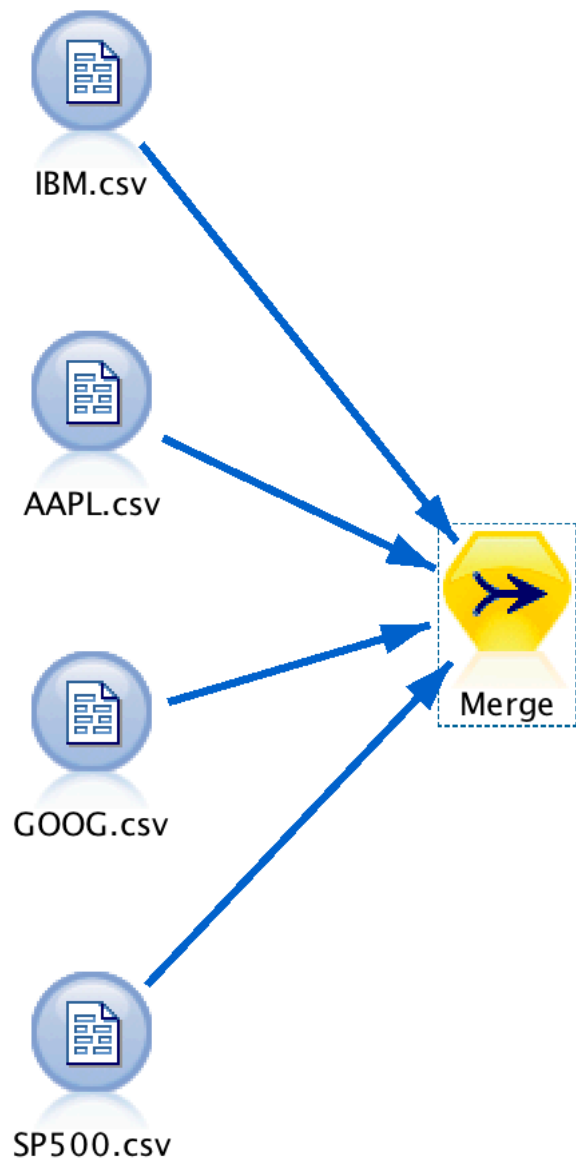
Quotes
Single quotes: Discard
Double quotes: Discard

OK Cancel Apply Reset

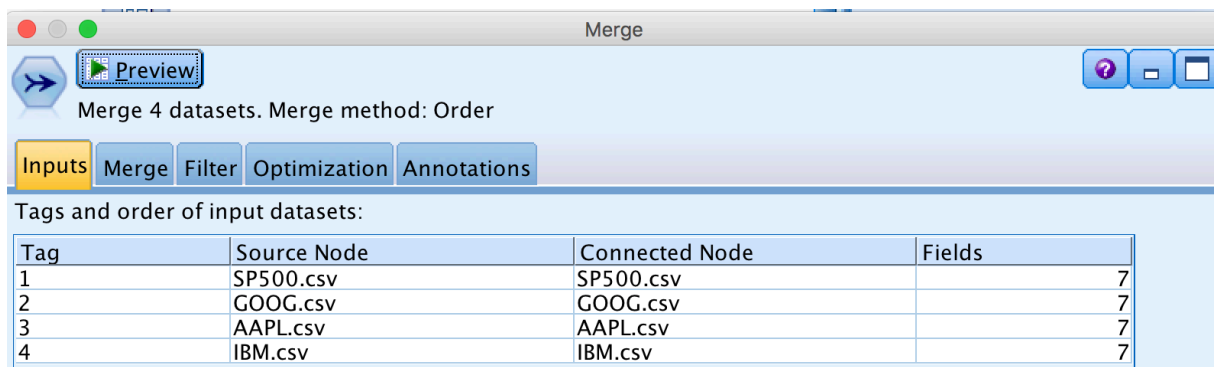
In each "Var. File" node in "Data" tab, you will see that field types are identified correctly

Field	Override	Storage	Input Format
Date	<input type="checkbox"/>	Date	
Open	<input type="checkbox"/>	Real	
High	<input type="checkbox"/>	Real	
Low	<input type="checkbox"/>	Real	
Close	<input type="checkbox"/>	Real	
Volume	<input type="checkbox"/>	Integer	
Adj Close	<input type="checkbox"/>	Real	

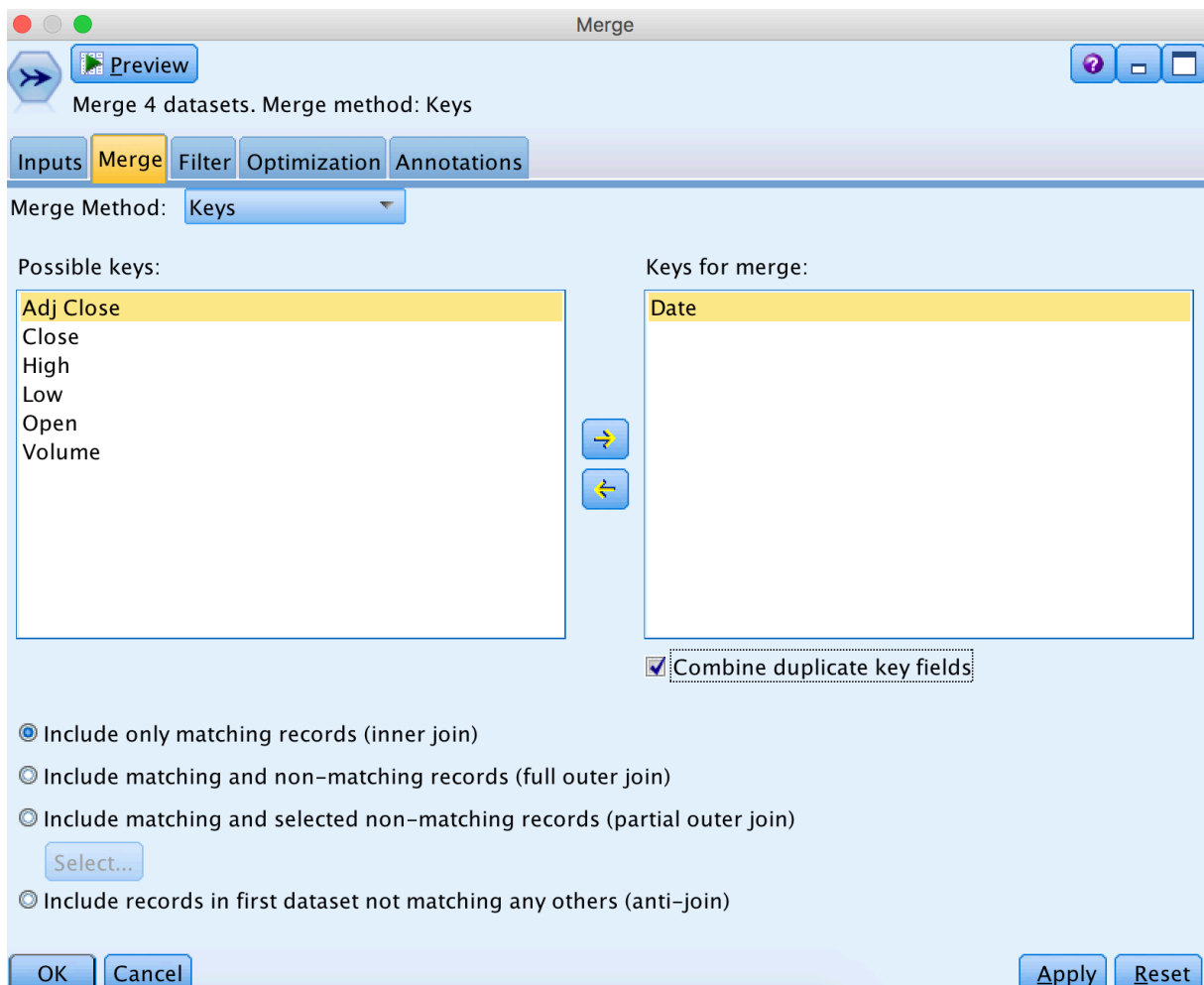
I can now merge these datasets into one by using “Merge” node from “Record Ops” palette.



Double click to open “Merge” node. In first tab, it shows you list of input files.



In “Merge” tab, we need to select “Key” for merging these datasets. “Key” should be common in all datasets and it will be used to match records and merge them. In our case, it’s “Date” column.



We will use default option “inner join” to merge datasets based on shorter dataset.

In “Filter” tab, we will filter out unnecessary columns and use suitable name, initials in this case, for “Adj Close” field coming from each dataset.

Merge

Preview

?

Merge 4 datasets. Merge method: Keys

Inputs

Merge

Filter

Optimization

Annotations

Fields: 25 in, 20 filtered, 4 renamed, 5 out

Field	Tag	Source Node	Connected N...	Filter	Field
Date				→	Date
Open	1	SP500.csv	SP500.csv	✗	Open
High	1	SP500.csv	SP500.csv	✗	High
Low	1	SP500.csv	SP500.csv	✗	Low
Close	1	SP500.csv	SP500.csv	✗	Close
Volume	1	SP500.csv	SP500.csv	✗	Volume
Adj Close	1	SP500.csv	SP500.csv	→	SP500 AC
Open	2	GOOG.csv	GOOG.csv	✗	Open
High	2	GOOG.csv	GOOG.csv	✗	High
Low	2	GOOG.csv	GOOG.csv	✗	Low
Close	2	GOOG.csv	GOOG.csv	✗	Close
Volume	2	GOOG.csv	GOOG.csv	✗	Volume
Adj Close	2	GOOG.csv	GOOG.csv	→	GOOG AC
Open	3	AAPL.csv	AAPL.csv	✗	Open
High	3	AAPL.csv	AAPL.csv	✗	High
Low	3	AAPL.csv	AAPL.csv	✗	Low
Close	3	AAPL.csv	AAPL.csv	✗	Close
Volume	3	AAPL.csv	AAPL.csv	✗	Volume
Adj Close	3	AAPL.csv	AAPL.csv	→	AAPL AC
Open	4	IBM.csv	IBM.csv	✗	Open
High	4	IBM.csv	IBM.csv	✗	High
Low	4	IBM.csv	IBM.csv	✗	Low
Close	4	IBM.csv	IBM.csv	✗	Close
Volume	4	IBM.csv	IBM.csv	✗	Volume
Adj Close	4	IBM.csv	IBM.csv	→	IBM AC

☒ View current fields
 ☐ View unused field settings

OK

Cancel

Apply

Reset

From “Output” palette, we can add “Table” node to see resulting dataset.

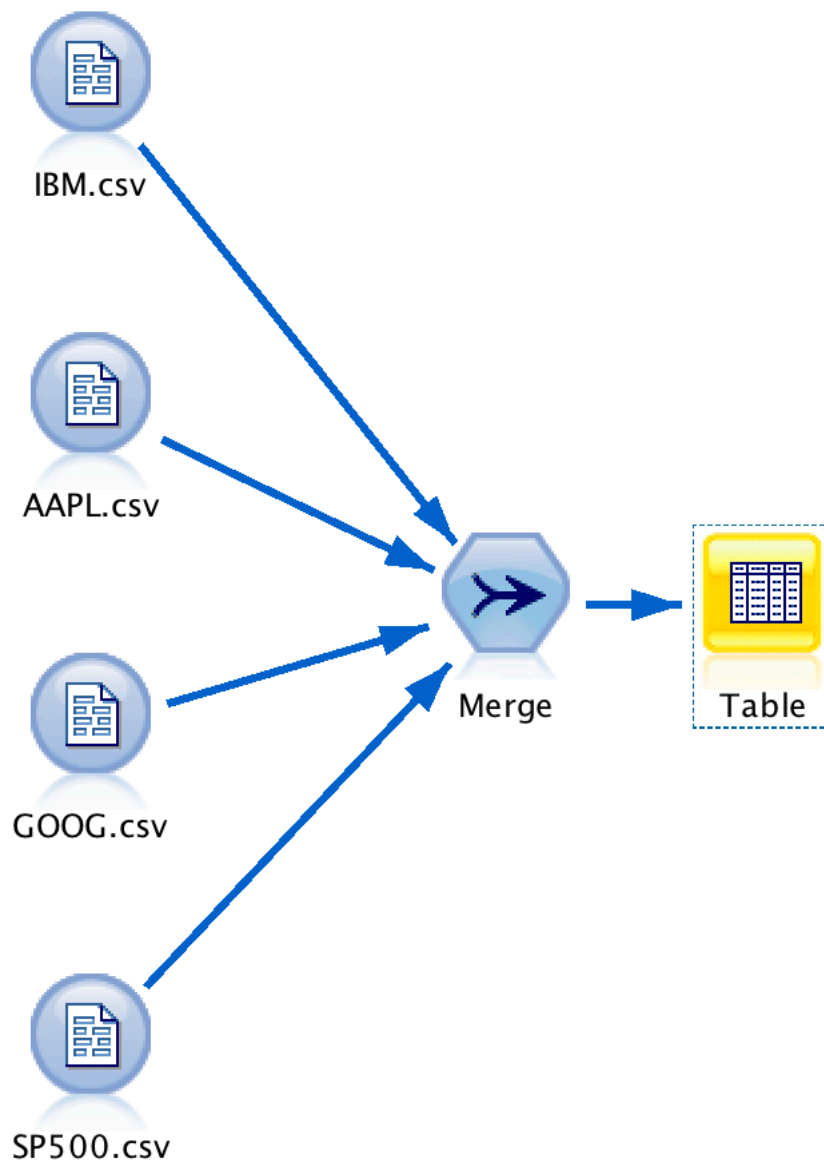


Table (5 fields, 3,038 records)					
<div> <div>File</div> <div>Edit</div> <div>Generate</div> <div></div> <div></div> <div></div> <div></div> </div> <div> <div>?</div> <div>X</div> </div>					
<div> <div>Table</div> <div>Annotations</div> </div>					
	Date	SP500_AC	GOOG_AC	AAPL_AC	IBM_AC
1	2004-08-19	1091.230	50.120	2.008	67.324
2	2004-08-20	1098.350	54.101	2.014	67.610
3	2004-08-23	1095.680	54.645	2.032	67.134
4	2004-08-24	1096.190	52.383	2.089	67.181
5	2004-08-25	1104.960	52.947	2.161	67.467
6	2004-08-26	1105.090	53.901	2.267	67.165
7	2004-08-27	1107.770	53.022	2.246	67.364
8	2004-08-30	1099.150	50.954	2.231	66.936
9	2004-08-31	1104.240	51.134	2.255	67.165
10	2004-09-01	1105.910	50.075	2.345	66.793
11	2004-09-02	1118.310	50.704	2.332	67.070
12	2004-09-03	1113.630	49.955	2.304	66.928
13	2004-09-07	1121.300	50.739	2.339	67.388
14	2004-09-08	1116.270	51.099	2.377	68.093
15	2004-09-09	1118.380	51.104	2.335	68.553
16	2004-09-10	1123.920	52.612	2.346	68.807
17	2004-09-13	1125.820	53.696	2.327	68.593
18	2004-09-14	1128.330	55.689	2.321	68.775
19	2004-09-15	1120.370	55.944	2.302	68.498
20	2004-09-16	1123.500	56.928	2.377	68.300

OK

3. Calculating Log Return Series for All Datasets

We will add “Derive” node from “Field Ops” palette

We use “Multiple” option to calculate log return series for all data sets at once.

Derive

Preview

Derive as: Formula

Settings Annotations

Mode: ☐ Single ☒ Multiple

Derive from:

Field name extension: Add as: ☒ Suffix ☐ Prefix

Derive as: TIP: Refer to selected fields by using @FIELD

Field type:

Formula:

1

OK Cancel Apply Reset

We will add all series to “Derive from:” section. We will use “log” and “OFFSET” formula to calculate log returns and notice how did we use “@FIELD” to refer to all added fields. We will rename this new field as original field name plus suffix “_Log_Return”

You can click preview to see resulting dataset.

Preview from _LR Node (9 fields, 10 records) #1

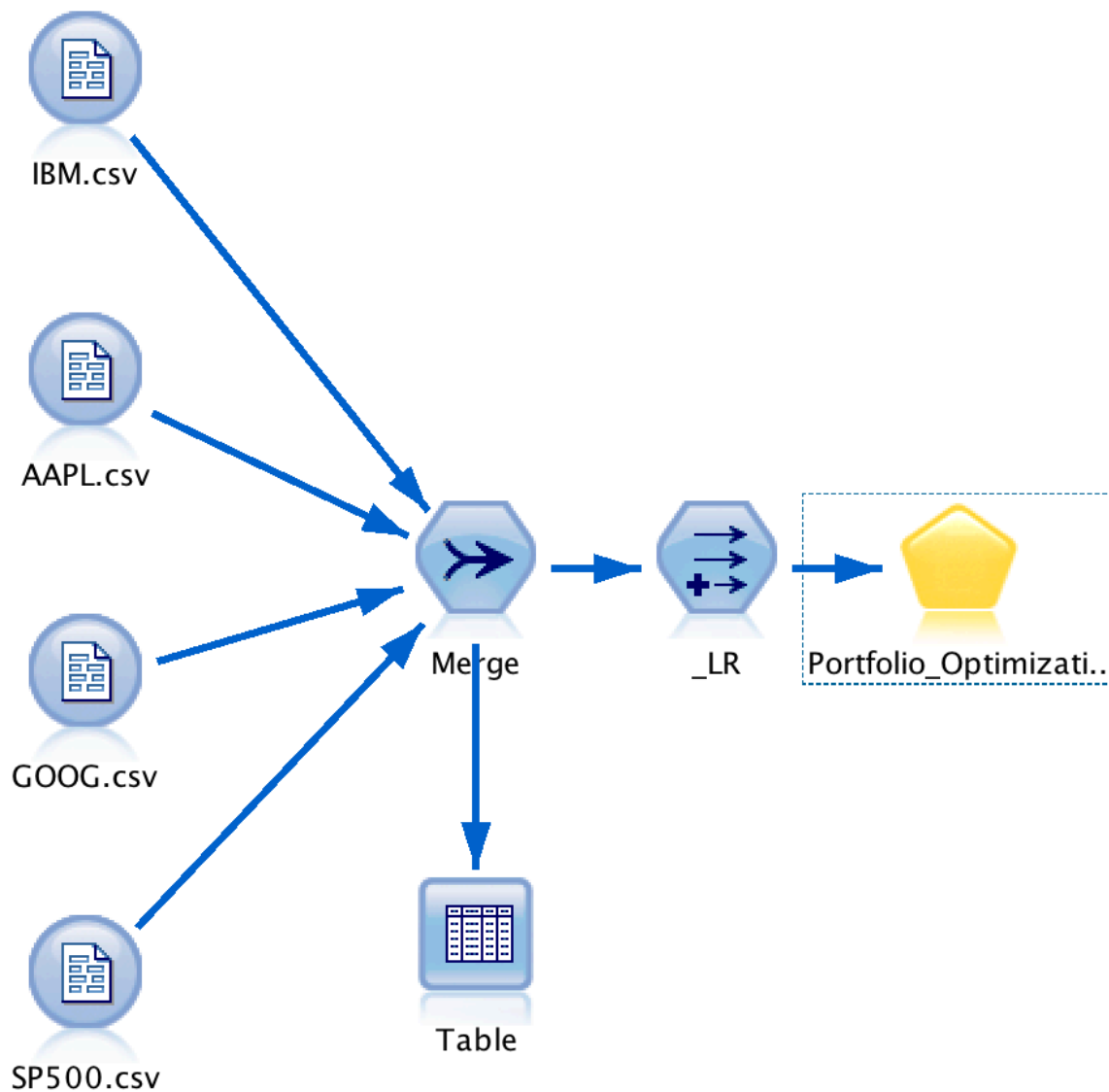
File Edit Generate

Table Annotations

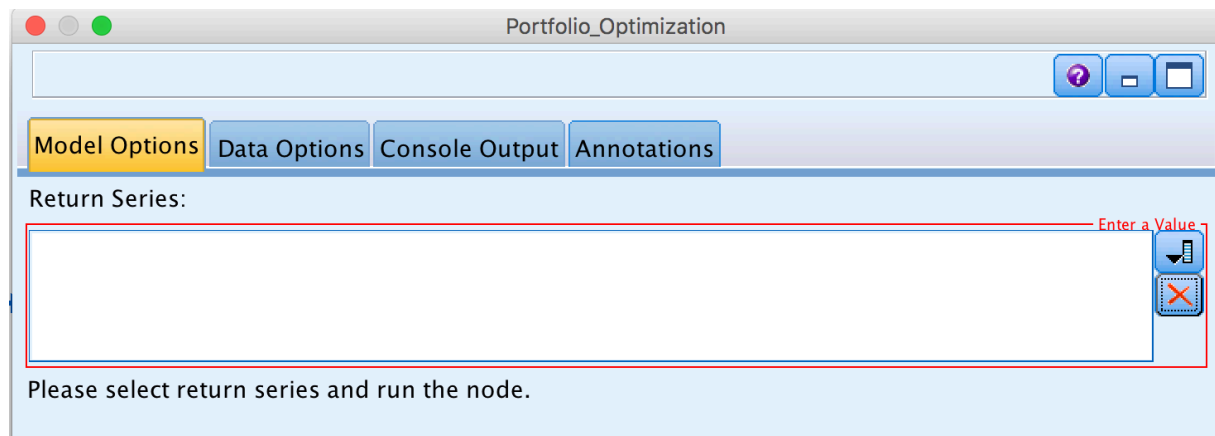
	Date	SP500_AC	GOOG_AC	AAPL_AC	IBM_AC	SP500_AC_LR	GOOG_AC_LR	AAPL_AC_LR
1	2004-08-19	1091.230	50.120	2.008	67.324	\$null\$	\$null\$	\$null\$
2	2004-08-20	1098.350	54.101	2.014	67.610	0.007	0.076	0.003
3	2004-08-23	1095.680	54.645	2.032	67.134	-0.002	0.010	0.009
4	2004-08-24	1096.190	52.383	2.089	67.181	0.000	-0.042	0.028
5	2004-08-25	1104.960	52.947	2.161	67.467	0.008	0.011	0.034
6	2004-08-26	1105.090	53.901	2.267	67.165	0.000	0.018	0.048
7	2004-08-27	1107.770	53.022	2.246	67.364	0.002	-0.016	-0.009
8	2004-08-30	1099.150	50.954	2.231	66.936	-0.008	-0.040	-0.007
9	2004-08-31	1104.240	51.134	2.255	67.165	0.005	0.004	0.011
10	2004-09-01	1105.910	50.075	2.345	66.793	0.002	-0.021	0.039

We will need to download following extension from this url (https://github.com/Umit-Mert/Portfolio_Optimization)

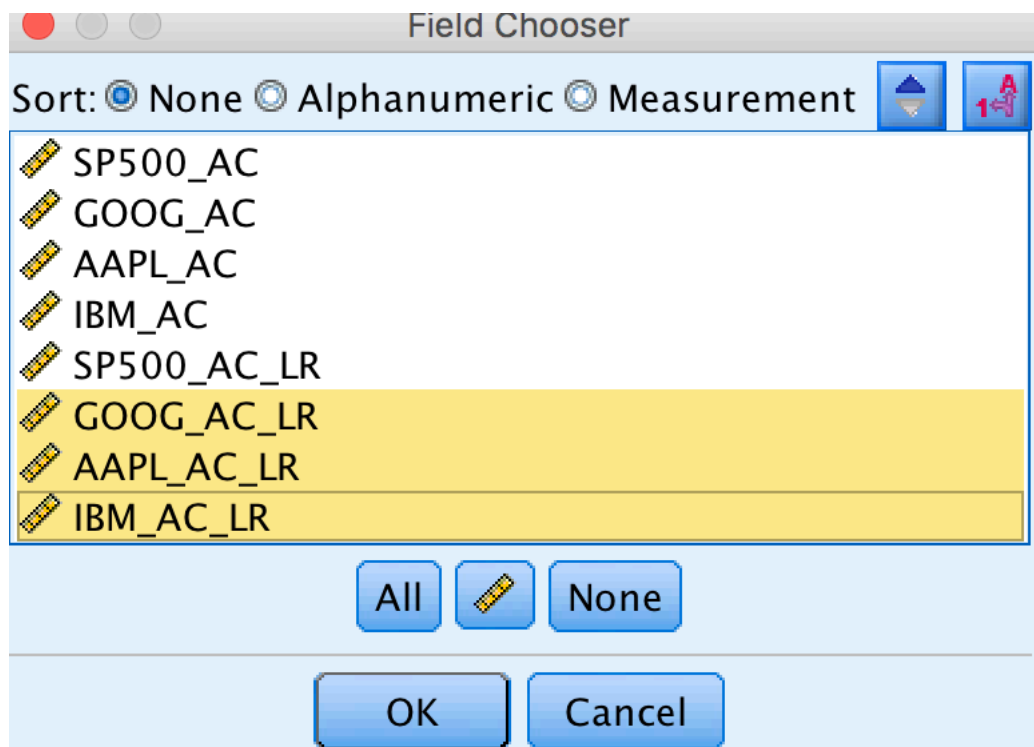
Once download and install finished. We need to add “Portfolio_Optimization” node from “Modeling” palette to our stream.



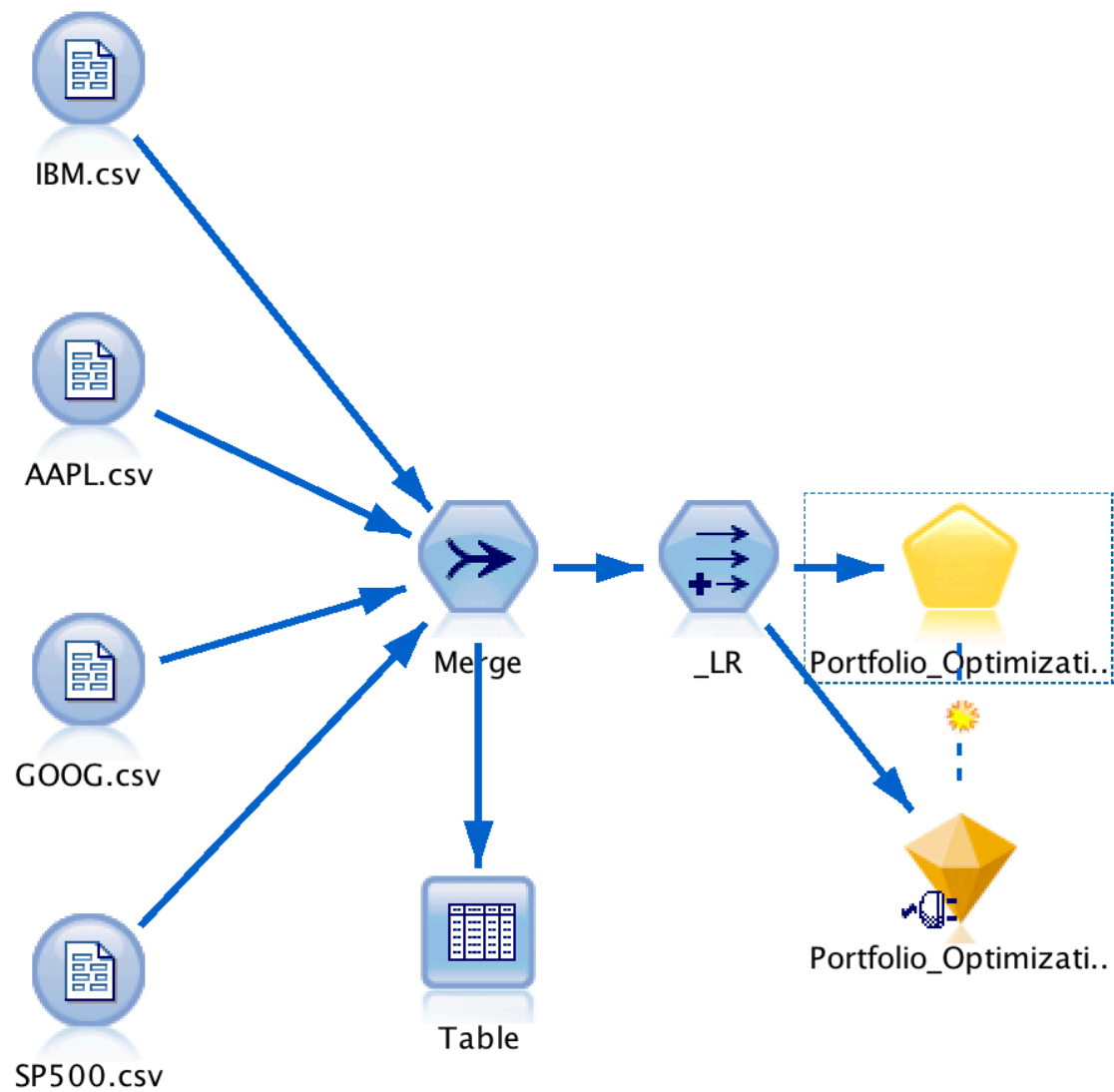
Double click to open that and select return series that we would like to work with.



In our case, return series are "GOOG", "AAPL" and "IBM".

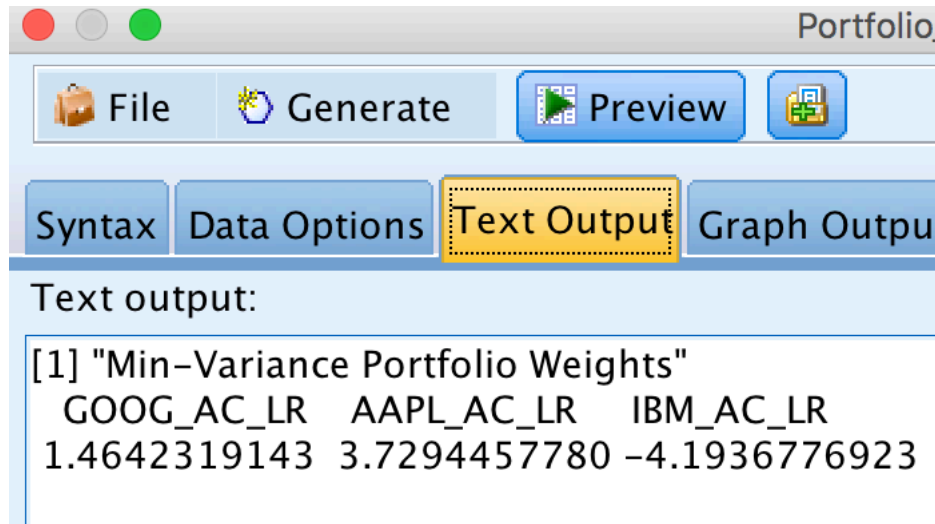


Let's click run and open the model nugget.

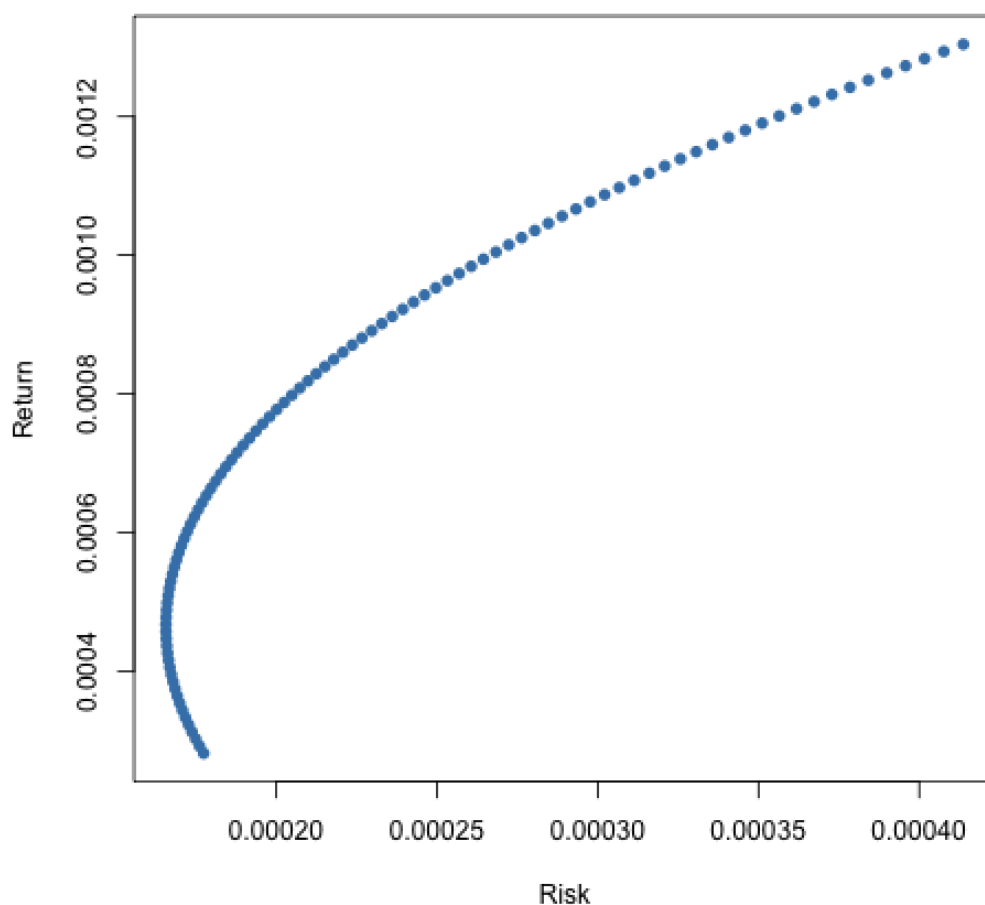


In model nugget, you should click “Text Output” tab to see the result.

We can see proportions in Portfolio weights section that we should invest in stocks “GOOG”, “AAPL” and “IBM”



You can also click “Graph Output” tab to see Markowitz bullet.



Our portfolio weights are referring to minimum variance portfolio which is far left point closest to “Return” axis.

Summary

In this lab, you learned how to use “MPT” node to construct risk efficient portfolios. Thank you and hope to see you next time.