#--------------------------------------

3. List down all the stocks in the industry that you picked. Collect the last one year stock price data for these listed stocks.
   Calculate the historical distance measure between all the possible pairs of stocks.
   From these find the pair with the smallest historical distance measure.

List down all the stocks in the industry that you picked.

#---

The main industry I selected was software.
The related industry I selected was Semiconductors.

The "stock universe" I am working in is the Standard and Poor's 500 (S&P 500.)
In order to know what ticker symbols are in S&P 500 sectors and industry,
I looked at the webpage https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

Using python I web data scrapped the data from the S&P 500 Component Stocks table on the webpage https://en.wikipedia.org/wiki/List_of_S%26P_500_companies. After cleaning the data, I saved a cache version to a .csv file and worked from that locally.
Webpages are dynamic and can change at anytime. Working from a .csv is stable for this
assignment. The code can dynamically get the data from the website should the companies
in the S&P 500 list change.

I constructed a Pandas DataFrame from the .csv file and extracted the information I needed. A list of ticker symbols for companies in the Software and Semiconductors industry.
The lists I got were:

```
# Software industry (Main Industry)
['ADBE',
 'ADP',
 'ADSK',
 'AKAM',
 'ANSS',
 'ATVI',
 'CA',
 'CDNS',
 'CRM',
 'CTXS',
 'EA',
 'EBAY',
 'FB',
 'FIS',
 'FISV',
 'GOOG',
 'GOOGL',
 'INTU',
 'MA',
 'MSFT',
 'NFLX',
 'NTAP',
 'ORCL',
 'PAYX',
 'RHT',
 'SNPS',
 'SYMC',
 'TSS',
```

```
 'V',
 'VRSN',
 'WU']

# Semiconductor industry (Related Industry
['ADI',
 'AMAT',
 'AMD',
 'AVGO',
 'INTC',
 'KLAC',
 'LRCX',
 'MCHP',
 'MU',
 'NVDA',
 'QCOM',
 'QRVO',
 'SWKS',
 'TXN',
 'XLNX']
```

Collect the last one year stock price data for these listed stocks.
#--

In order to the last one year stock price data, I pull down the data from https://www.quandl.com/,
via Pandas. Google Finance and Yahoo Finance have an API, but as of recent (maybe around september 2017,)
both of these API's no longer work correctly. They do not download the data. The quandl data still seems
to work via Pandas API.

With the stock price data downloaded, I got 252 trading days (1 year,) of stock data for all the ticker
symbols from the main industry (software) and related industry (semiconductors.) I saved the data
in a Pandas Panel. A note, most people use Pandas DataFrames and Series, but Panels seemed appropriate in
this case because the items in a Panel is like a worksheet in Excel. I stored each ticker symbol as an
item in a Panel and saved/cached the Panel as an Excel file. I could open the Excel file and see that
each worksheet is a ticker symbol. I could also load in the Excel file and build a Pandas Panel from it.
This workflow allowed to cached the dynamic data (changes every day) statically and locally. It also allowed
me to view the data in Excel, which is intuitive to debug since data is easily readible without doing any
special in code.

Calculate the historical distance measure between all the possible pairs of stocks.
#--

From my understanding, this is a measure for selecting the stocks and it is
subjective in a sense. Selecting a measure is common in other fields, such
as Machine Learning/Deep Learning, where you have to decide the loss function
measure. I've come across other literature that used other measures, including
correlation between stock prices, but I am presumed historical distance measure is
meant to use the same measured used in the paper
The 'Pairs Trading: Performance of a relative value arbitrage rule' paper by Gatev, Goetzmann and Rouwenhorst.

In section 2.1 Pairs Formation, page 10 of the paper, it goes on to explain,
"screen out all stocks from the CRSP daily files that have one or more days with no trade."
and
"choose a matching partner for each stock by finding the security that minimizes the sum of squared deviations
between the two normalized price series."

Another explanation can be found here https://quantpedia.com/Screener/Details/12, under the Simple trading strategy.

This is the Pair Formation Period.
I checked if any stocks had one or more days with no trade and removed them.
I used adjusted closing price because it takes into account stock split and dividends.
I normalized the price by dividing each trading by the starting price during the formation
period. This makes the starting price for all the stocks $1 (price normalized.)

Next was pair up the stocks and to minimizes the sum of squared deviations between the
two normalized price series. This was done by taking the normalized price of the first stock
and subtracting the normalized price of the second stock and squaring the difference per
trading day. At the end, sum up all the squared difference. This was all done in Pandas
and Numpy (which Pandas was built on top of.)

The pairs and there minimized sum of squared deviations values were then sorted in ascending
order according to their values.

Here is the sorted list for the Main industry Software paired within the same industry. (Only listing 10 because the list is long):
NOTE: There is a difference between GOOG and GOOGL
https://www.investopedia.com/articles/markets/052215/goog-or-googl-which-google-should-you-buy.asp

```
[(('GOOG', 'GOOGL'), 0.04045301814067839),
 (('CRM', 'V'), 0.2072181498473113),
 (('V', 'VRSN'), 0.27237885594731637),
 (('FIS', 'FISV'), 0.2857409476522602),
 (('INTU', 'MSFT'), 0.2961582963157633),
 (('CRM', 'MSFT'), 0.36001117965462753),
 (('MA', 'VRSN'), 0.36546272129588436),
 (('MA', 'V'), 0.3923483790166055),
 (('MA', 'MSFT'), 0.39764108964701733),
 (('MA', 'TSS'), 0.41429499358425526)]
```

Here is the sorted list for the Related industry Semiconductors paired within the same industry. (Only listing 10 because the list is long):

```
[(('ADI', 'XLNX'), 0.8472312767165687),
 (('AVGO', 'MCHP'), 1.234849025424745),
 (('KLAC', 'MCHP'), 1.59847575685719),
 (('AVGO', 'SWKS'), 1.822496536249247),
 (('KLAC', 'QRVO'), 1.8528020213566363),
 (('ADI', 'TXN'), 1.9859323681800614),
 (('ADI', 'QRVO'), 2.0794069574152445),
 (('ADI', 'KLAC'), 2.2691757219751896),
```

```
 (('TXN', 'XLNX'), 2.4415248210604705),
 (('MCHP', 'SWKS'), 2.457283560344462)]
```

Here is the sorted list for the Main industry Software paired with related Semiconductors industry. (Only listing 10 because the list is long):

```
[(('GOOGL', 'XLNX'), 0.5245130065173396),
 (('MSFT', 'TXN'), 0.5759653537096523),
 (('SNPS', 'MCHP'), 0.598657846501867),
 (('FB', 'MCHP'), 0.6189008071597084),
 (('FIS', 'XLNX'), 0.6465886000734158),
 (('EBAY', 'ADI'), 0.6669700061807349),
 (('GOOG', 'XLNX'), 0.6938522819778221),
 (('FISV', 'XLNX'), 0.7275039920196817),
 (('FISV', 'ADI'), 0.8126773010290321),
 (('V', 'MCHP'), 0.9116461786887379)]
```

From these find the pair with the smallest historical distance measure.

#----------

The Rubric did not say we were to follow the paper
The 'Pairs Trading: Performance of a relative value arbitrage rule' paper by Gatev, Goetzmann and Rouwenhorst.
exactly. If so, according to the paper section 2.2 Trading Period, page 11.

"the top 5 and 20 pairs with the smallest historical distance measure, in addition to the 20 pairs after the top 100
(pairs 101-120). This last set is valuable because most of the top pairs share certain characteristics,"

I believe the paper's decision on what to pick was subjective and not a rule of thumb. It was what they did to get
the results they presented. They mentioned in the paper that the Pairs Trading strategy is flexible. Stocks do not have
to be in the same industry, you can do this with more than 2 stocks in a pair. The measuring criteria for selecting stocks
can be something else besides historical distance measure.

The Rubric says "From these find the pair with the smallest historical distance measure."

Here is the sorted list for the Main industry Software paired within the same industry. (Only listing 10 because the list is long):
NOTE: There is a difference between GOOG and GOOGL
https://www.investopedia.com/articles/markets/052215/goog-or-googl-which-google-should-you-buy.asp

```
(('GOOG', 'GOOGL'), 0.04045301814067839)
```

Here is the sorted list for the Related industry Semiconductors paired within the same industry. (Only listing 10 because the list is long)

```
(('ADI', 'XLNX'), 0.8472312767165687)
```

Here is the sorted list for the Main industry Software paired with related Semiconductors industry. (Only listing 10 because the list is long):

```
(('GOOGL', 'XLNX'), 0.5245130065173396)
```

With (('GOOG', 'GOOGL'), 0.04045301814067839) being the smallest overall.

(('GOOG', 'GOOGL'), 0.04045301814067839) is also the smallest historical distance from the main industry that I picked, Software.

4. List down all the stocks in the related industry. Collect the last one year stock price data for these listed stocks.
Make all possible pairs with one stock from the main industry and another stock from the related industry.
If there are 3 stocks in main industry and 4 stocks in the related industry, then we will have a total of 3 * 4 pairs.
For these pairs, calculate the historical distance measure and pick the pair with the lowest historical distance measure.
#----------

The same procedure as the last question. I did Main industry, Software, and Related industry, Semiconductors, at the same time using Python.
A quick synopsis
- Web data scraped S&P 500 for ticker symbols from https://en.wikipedia.org/wiki/List_of_S%26P_500_companies
- Put data in Pandas DataFrame and saved .csv
- Got all the ticker symbols for the Main industry, Software, and Related industry, Semiconductors.
  For this question, the ticker symbols for the Related industry, Semiconductors are

```
['ADI',
 'AMAT',
 'AMD',
 'AVGO',
 'INTC',
 'KLAC',
 'LRCX',
 'MCHP',
 'MU',
 'NVDA',
 'QCOM',
 'QRVO',
 'SWKS',
 'TXN',
 'XLNX']
```

- Got one year (252 trading days) of stock price data for these ticker symbols. Got them from https://www.quandl.com/
- Made posible pairs between main industry and related industry.

```
# Software industry (Main Industry)
['ADBE',
 'ADP',
 'ADSK',
 'AKAM',
 'ANSS',
 'ATVI',
 'CA',
 'CDNS',
 'CRM',
 'CTXS',
 'EA',
 'EBAY',
 'FB',
 'FIS',
 'FISV',
 'GOOG',
 'GOOGL',
```

```
 'INTU',
 'MA',
 'MSFT',
 'NFLX',
 'NTAP',
 'ORCL',
 'PAYX',
 'RHT',
 'SNPS',
 'SYMC',
 'TSS',
 'V',
 'VRSN',
 'WU']

# Semiconductor industry (Related Industry)
['ADI',
 'AMAT',
 'AMD',
 'AVGO',
 'INTC',
 'KLAC',
 'LRCX',
 'MCHP',
 'MU',
 'NVDA',
 'QCOM',
 'QRVO',
 'SWKS',
 'TXN',
 'XLNX']
```

- Software industry (Main Industry) have 15 ticker symbols. Semiconductors industry (Related Industry,) have 31 tickers.
  15 x 31 = 465 pairs.
- I calculated the historical distance measure:

In section 2.1 Pairs Formation, page 10 of the paper, it goes on to explain,
"screen out all stocks from the CRSP daily files that have one or more days with no trade."
and
"choose a matching partner for each stock by finding the security that minimizes the sum of squared deviations
between the two normalized price series."

Another explanation can be found here https://quantpedia.com/Screener/Details/12, under the Simple trading strategy.

This is the Pair Formation Period.
I checked if any stocks had one or more days with no trade and removed them.
I used adjusted closing price because it takes into account stock split and dividends.
I normalized the price by dividing each trading by the starting price during the formation
period. This makes the starting price for all the stocks $1 (price normalized.)

Next was pair up the stocks and to minimizes the sum of squared deviations between the
two normalized price series. This was done by taking the normalized price of the first stock
and subtracting the normalized price of the second stock and squaring the difference per

trading day. At the end, sum up all the squared difference. This was all done in Pandas
and Numpy (which Pandas was built on top of.)

The pairs and there minimized sum of squared deviations values were then sorted in ascending
order according to their values.

Here is the sorted list for the Main industry Software paired with related Semiconductors industry. (Only listing 10 because the list is long, 465 pairs):

```
[(('GOOGL', 'XLNX'), 0.5245130065173396),
 (('MSFT', 'TXN'), 0.5759653537096523),
 (('SNPS', 'MCHP'), 0.5986578446501867),
 (('FB', 'MCHP'), 0.6189008071597084),
 (('FIS', 'XLNX'), 0.6465886000734158),
 (('EBAY', 'ADI'), 0.6669700061807349),
 (('GOOG', 'XLNX'), 0.6938522819778221),
 (('FISV', 'XLNX'), 0.7275039920196817),
 (('FISV', 'ADI'), 0.8126773010290321),
 (('V', 'MCHP'), 0.9116461786887379)]
```

The smallest historical distance for main industry (Software) and related industry (Semiconductors) pair is:

```
(('GOOGL', 'XLNX'), 0.5245130065173396)
```