# data_preparation

April 14, 2022

## 1 Data preparation

### 1.0.1 Install and importing libraries

[1]:
```
!pip install yfinance
```

```
Requirement already satisfied: yfinance in c:\users\agnie\anaconda3\lib\site-
packages (0.1.70)
Requirement already satisfied: lxml>=4.5.1 in c:\users\agnie\anaconda3\lib\site-
packages (from yfinance) (4.6.3)
Requirement already satisfied: requests>=2.26 in
c:\users\agnie\anaconda3\lib\site-packages (from yfinance) (2.27.1)
Requirement already satisfied: pandas>=0.24.0 in
c:\users\agnie\anaconda3\lib\site-packages (from yfinance) (1.2.4)
Requirement already satisfied: numpy>=1.15 in c:\users\agnie\anaconda3\lib\site-
packages (from yfinance) (1.18.5)
Requirement already satisfied: multitasking>=0.0.7 in
c:\users\agnie\anaconda3\lib\site-packages (from yfinance) (0.0.10)
Requirement already satisfied: python-dateutil>=2.7.3 in
c:\users\agnie\anaconda3\lib\site-packages (from pandas>=0.24.0->yfinance)
(2.8.1)
Requirement already satisfied: pytz>=2017.3 in
c:\users\agnie\anaconda3\lib\site-packages (from pandas>=0.24.0->yfinance)
(2021.1)
Requirement already satisfied: six>=1.5 in c:\users\agnie\anaconda3\lib\site-
packages (from python-dateutil>=2.7.3->pandas>=0.24.0->yfinance) (1.15.0)
Requirement already satisfied: charset-normalizer~=2.0.0 in
c:\users\agnie\anaconda3\lib\site-packages (from requests>=2.26->yfinance)
(2.0.12)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
c:\users\agnie\anaconda3\lib\site-packages (from requests>=2.26->yfinance)
(1.26.4)
Requirement already satisfied: idna<4,>=2.5 in
c:\users\agnie\anaconda3\lib\site-packages (from requests>=2.26->yfinance)
(2.10)
Requirement already satisfied: certifi>=2017.4.17 in
c:\users\agnie\anaconda3\lib\site-packages (from requests>=2.26->yfinance)
(2020.12.5)
```

[2]:
```
import pandas as pd
import numpy as np
import yfinance as yf
import datetime

import base64
from IPython.display import HTML
```

### 1.0.2 Uploading data

[3]:
```
url = "https://raw.githubusercontent.com/Agablue-red/Machine-Learning/master/
↪data/CONVICTIONLISTTOPN_BSLD-408.csv"
df = pd.read_csv(url, index_col=False, names=['info', 'date', 'symbol',␣
↪'symbol2', 'sector', 'number', 'score'])
df
```

[3]:

| | info | date | symbol | symbol2 | sector | number | score |
|---|---|---|---|---|---|---|---|
| 0 | 10:01:54.481 77425 [77425-thread-2] INFO a.s… | 2004-02-11 | SU | SU | Energy Minerals | GN63J3-R | 0.953727 |
| 1 | 10:01:54.481 77425 [77425-thread-2] INFO a.s… | 2004-02-11 | GGG | GGG | Producer Manufacturing | H5490W-R | 0.952753 |
| 2 | 10:01:54.481 77425 [77425-thread-2] INFO a.s… | 2004-02-11 | WGR | WGR | Energy Minerals | V0622Q-R | 0.947634 |
| 3 | 10:01:54.481 77425 [77425-thread-2] INFO a.s… | 2004-02-11 | CWT | CWT | Utilities | GSWXLY-R | 0.934181 |
| 4 | 10:01:54.481 77425 [77425-thread-2] INFO a.s… | 2004-02-11 | BLL | BLL | Process Industries | VFT0VQ-R | 0.922862 |
| … | … | … | … | … | … | … | … |
| 37355 | 10:27:03.049 77425 [77425-thread-2] INFO a.s… | 2022-02-09 | PEP | PEP | Consumer Non-Durables | PPCTFP-R | 0.701507 |
| 37356 | 10:27:03.049 77425 [77425-thread-2] INFO a.s… | 2022-02-09 | SSNC | SSNC | Technology Services | G92RX2-R | 0.701123 |
| 37357 | 10:27:03.049 77425 [77425-thread-2] INFO a.s… | 2022-02-09 | GEF | GEF | Process Industries | MPX0N4-R | 0.697954 |
| 37358 | 10:27:03.049 77425 [77425-thread-2] INFO a.s… | 2022-02-09 | DPZ | DPZ | Consumer Services | F05QG0-R | 0.697741 |
| 37359 | 10:27:03.049 77425 [77425-thread-2] INFO a.s… | 2022-02-09 | LIFZF | LIFZF | Non-Energy Minerals | Q404Y1-R | 0.695644 |

[37360 rows x 7 columns]

[4]:
```
# Removing column data
df.drop(['info','symbol2','number'], axis=1, inplace=True)
```

[5]:
```
# Convert argument to datetime
df['date'] = pd.to_datetime(df['date'])
df.set_index('date', inplace=True)
```

[6]:
```
df.head()
```

[6]:

| date | symbol | sector | score |
|---|---|---|---|
| 2004-02-11 | SU | Energy Minerals | 0.953727 |
| 2004-02-11 | GGG | Producer Manufacturing | 0.952753 |
| 2004-02-11 | WGR | Energy Minerals | 0.947634 |
| 2004-02-11 | CWT | Utilities | 0.934181 |
| 2004-02-11 | BLL | Process Industries | 0.922862 |

### 1.0.3 Information about dataset

[7]:
```
print('Shape of raw dataset: {}'.format(df.shape))
```

```
Shape of raw dataset: (37360, 3)
```

[8]:
```
# Return the data type of each column
df.dtypes
```

[8]:
```
symbol     object
sector     object
score     float64
dtype: object
```

[9]:
```
print('Number of unique dates: {}'.format(df.index.nunique()))
```

```
Number of unique dates: 467
```

[10]:
```
# Return the number of missing values
df.isnull().sum()
```

[10]:
```
symbol    0
sector    0
score     0
dtype: int64
```

[11]:
```
print('Number of duplicate rows: {}'.format(df.duplicated().sum()))
```

```
Number of duplicate rows: 0
```

[12]:
```
df.symbol.unique()
```

[12]:
```
array(['SU', 'GGG', 'WGR', …, 'DELL', 'BOOT', 'AGCO'], dtype=object)
```

[13]:
```
print('Number of unique symbols: {}'.format(df.symbol.nunique()))
```

```
Number of unique symbols: 1834
```

[14]:
```
df.sector.unique()
```

[14]:
```
array(['Energy Minerals', 'Producer Manufacturing', 'Utilities',
       'Process Industries', 'Consumer Services', 'Transportation',
       'Retail Trade', 'Finance', 'Health Technology', 'Miscellaneous',
       'Non-Energy Minerals', 'Distribution Services',
       'Consumer Non-Durables', 'Commercial Services',
       'Technology Services', 'Consumer Durables', 'Health Services',
       'Electronic Technology', 'Industrial Services', 'Communications'],
      dtype=object)
```

[15]:
```
print('Number of unique sectors: {}'.format(df.sector.nunique()))
```

```
Number of unique sectors: 20
```

[16]:
```
# basic statistics
df.score.describe()
```

[16]:
```
count    37360.000000
mean         0.731634
std          0.118071
min          0.413554
25%          0.655228
50%          0.743032
75%          0.813181
max          0.987225
Name: score, dtype: float64
```

### 1.0.4 Download Financial Data from Yahoo

[17]:
```
# delete an unnecessary part in the 'symbol' column
df['symbol'] = df['symbol'].str.replace(".", " ")
df['symbol'] = df['symbol'].str.split(' ')

xyz = []
for x in df["symbol"].to_numpy():
  xyz.append(x[0])
df["symbol"] = xyz
```

```
<ipython-input-17-e3e036e6b8ef>:2: FutureWarning: The default value of regex
will change from True to False in a future version. In addition, single
character regular expressions will*not* be treated as literal strings when
regex=True.
  df['symbol'] = df['symbol'].str.replace(".", " ")
```

[18]:
```
tickers = df.symbol.unique()
list_tickers = tickers.tolist()
Symbol = yf.Tickers(list_tickers)
```

[19]:
```
df_yahoo = yf.download(list_tickers, start='2004-02-10', end='2022-02-10',␣
↪interval="1wk")['Close']
```

```
[**********100%**********]  1804 of 1804 completed

397 Failed downloads:
- FDO: No data found for this date range, symbol may be delisted
- VAR: No data found, symbol may be delisted
- EGOV: No data found, symbol may be delisted
- HNR: No data found for this date range, symbol may be delisted
- NGLS: No data found for this date range, symbol may be delisted
- PALDF: No data found, symbol may be delisted
- ETFC: No data found, symbol may be delisted
- BJS: No data found for this date range, symbol may be delisted
- VARI: No data found for this date range, symbol may be delisted
- MTSC: No data found, symbol may be delisted
- KWD: No data found for this date range, symbol may be delisted
- BEAV: No data found for this date range, symbol may be delisted
- BLKIA: No data found for this date range, symbol may be delisted
- TFCF: No data found, symbol may be delisted
- CVA: No data found, symbol may be delisted
- TSS: No data found, symbol may be delisted
- CLGX: No data found, symbol may be delisted
- HDLM: No data found, symbol may be delisted
- GXDX: No data found for this date range, symbol may be delisted
- HDS: No data found, symbol may be delisted
- LBYYQ: No data found, symbol may be delisted
- IGTE: No data found for this date range, symbol may be delisted
- AIRM: No data found for this date range, symbol may be delisted
- JAH: No data found for this date range, symbol may be delisted
- RATE: No data found for this date range, symbol may be delisted
- BYI: No data found for this date range, symbol may be delisted
- ANH: No data found, symbol may be delisted
- PLMD: No data found for this date range, symbol may be delisted
- EFXFF: No data found, symbol may be delisted
- PSUNQ: No data found for this date range, symbol may be delisted
- ZQKSQ: No data found for this date range, symbol may be delisted
- MXIM: No data found, symbol may be delisted
- SPSS: No data found for this date range, symbol may be delisted
- TLM: No data found for this date range, symbol may be delisted
- XTO: No data found for this date range, symbol may be delisted
- BRSS: No data found, symbol may be delisted
- ASFI: No data found, symbol may be delisted
```

- DMND: No data found for this date range, symbol may be delisted
- CRRC: No data found for this date range, symbol may be delisted
- ININ: No data found for this date range, symbol may be delisted
- CKH: No data found, symbol may be delisted
- HW: No data found for this date range, symbol may be delisted
- GYI: No data found for this date range, symbol may be delisted
- SCLN: No data found for this date range, symbol may be delisted
- OCR: No data found for this date range, symbol may be delisted
- HYSL: No data found for this date range, symbol may be delisted
- GRA: No data found, symbol may be delisted
- AQNT: No data found for this date range, symbol may be delisted
- CGX: No data found for this date range, symbol may be delisted
- MWIV: No data found for this date range, symbol may be delisted
- CUB: No data found, symbol may be delisted
- WRI: No data found, symbol may be delisted
- KWKAQ: No data found for this date range, symbol may be delisted
- UNS: No data found for this date range, symbol may be delisted
- ACAS: No data found for this date range, symbol may be delisted
- KKD: No data found for this date range, symbol may be delisted
- WWIN: No data found, symbol may be delisted
- CEPH: No data found for this date range, symbol may be delisted
- PMZFF: No data found, symbol may be delisted
- STU: No data found for this date range, symbol may be delisted
- MLS: No data found for this date range, symbol may be delisted
- INCLF: No data found, symbol may be delisted
- KEM: No data found, symbol may be delisted
- PRX: No data found for this date range, symbol may be delisted
- NXG: No data found for this date range, symbol may be delisted
- CHKAQ: No data found, symbol may be delisted
- PPP: No data found for this date range, symbol may be delisted
- BRCM: No data found for this date range, symbol may be delisted
- HCR: No data found, symbol may be delisted
- THX: No data found for this date range, symbol may be delisted
- GR: No data found for this date range, symbol may be delisted
- ENSI: No data found for this date range, symbol may be delisted
- TNB: No data found for this date range, symbol may be delisted
- AGN: No data found, symbol may be delisted
- GAS: No data found for this date range, symbol may be delisted
- NTY: No data found for this date range, symbol may be delisted
- HRC: No data found, symbol may be delisted
- TRK: No data found, symbol may be delisted
- NBL: No data found, symbol may be delisted
- WFSI: No data found for this date range, symbol may be delisted
- HUB: No data found for this date range, symbol may be delisted
- GSF: No data found for this date range, symbol may be delisted
- LUFK: No data found for this date range, symbol may be delisted
- HOFD: No data found, symbol may be delisted
- BBX: No data found, symbol may be delisted

- LABL: No data found, symbol may be delisted
- LNCR: No data found for this date range, symbol may be delisted
- CORE: No data found, symbol may be delisted
- RSCR: No data found for this date range, symbol may be delisted
- SGK: No data found for this date range, symbol may be delisted
- CLE: No data found for this date range, symbol may be delisted
- CY: No data found, symbol may be delisted
- GCGMF: No data found, symbol may be delisted
- RSHCQ: No data found for this date range, symbol may be delisted
- AACB: No data found for this date range, symbol may be delisted
- CDX: Data doesn't exist for startDate = 1076367600, endDate = 1644447600
- MRD: No data found for this date range, symbol may be delisted
- ARTI: No data found for this date range, symbol may be delisted
- CMD: No data found, symbol may be delisted
- STRZA: No data found for this date range, symbol may be delisted
- EQY: No data found for this date range, symbol may be delisted
- HOTT: No data found for this date range, symbol may be delisted
- BPL: No data found, symbol may be delisted
- GTM: No data found for this date range, symbol may be delisted
- TE: No data found for this date range, symbol may be delisted
- GTRC: No data found for this date range, symbol may be delisted
- FTO: No data found for this date range, symbol may be delisted
- GWR: No data found, symbol may be delisted
- TIBX: No data found for this date range, symbol may be delisted
- JDAS: No data found for this date range, symbol may be delisted
- MPS: No data found for this date range, symbol may be delisted
- OVTI: No data found for this date range, symbol may be delisted
- ARDNA: No data found for this date range, symbol may be delisted
- HPOL: No data found for this date range, symbol may be delisted
- BKC: No data found for this date range, symbol may be delisted
- AIPC: No data found, symbol may be delisted
- HNH: No data found for this date range, symbol may be delisted
- CTX: No data found for this date range, symbol may be delisted
- GVHR: No data found for this date range, symbol may be delisted
- AYR: No data found, symbol may be delisted
- SRR: No data found for this date range, symbol may be delisted
- ACO: No data found for this date range, symbol may be delisted
- PCL: No data found for this date range, symbol may be delisted
- DPL: No data found for this date range, symbol may be delisted
- EPE: No data found, symbol may be delisted
- LO: No data found for this date range, symbol may be delisted
- CTWS: No data found, symbol may be delisted
- OAK: No data found, symbol may be delisted
- ODSY: No data found for this date range, symbol may be delisted
- CMO: No data found, symbol may be delisted
- NDN: No data found for this date range, symbol may be delisted
- NZ: No data found for this date range, symbol may be delisted
- AEA: No data found for this date range, symbol may be delisted

- DFODQ: No data found, symbol may be delisted
- CLP: No data found for this date range, symbol may be delisted
- UFS: No data found, symbol may be delisted
- KDN: No data found for this date range, symbol may be delisted
- TQNT: No data found for this date range, symbol may be delisted
- MCF: No data found, symbol may be delisted
- HNZ: No data found, symbol may be delisted
- BF: No data found for this date range, symbol may be delisted
- AVX: No data found, symbol may be delisted
- FRX: No data found, symbol may be delisted
- LEARQ: No data found, symbol may be delisted
- VRX: No data found for this date range, symbol may be delisted
- DTV: No data found, symbol may be delisted
- LAF: No data found for this date range, symbol may be delisted
- TIN: No data found for this date range, symbol may be delisted
- EE: Data doesn't exist for startDate = 1076367600, endDate = 1644447600
- MHM: No data found for this date range, symbol may be delisted
- MCRS: No data found for this date range, symbol may be delisted
- NEWCQ: No data found, symbol may be delisted
- ACAT: No data found for this date range, symbol may be delisted
- GISX: No data found for this date range, symbol may be delisted
- IM: No data found for this date range, symbol may be delisted
- ACS: No data found for this date range, symbol may be delisted
- LDR: No data found for this date range, symbol may be delisted
- PETM: No data found for this date range, symbol may be delisted
- WLSM: No data found, symbol may be delisted
- ALXN: No data found, symbol may be delisted
- COCOQ: No data found, symbol may be delisted
- VIAB: No data found, symbol may be delisted
- MATK: No data found for this date range, symbol may be delisted
- SIRO: No data found for this date range, symbol may be delisted
- BOBE: No data found for this date range, symbol may be delisted
- GDI: No data found, symbol may be delisted
- ASCA: Data doesn't exist for startDate = 1076367600, endDate = 1644447600
- CHG: No data found for this date range, symbol may be delisted
- TSY: No data found for this date range, symbol may be delisted
- VLTR: No data found for this date range, symbol may be delisted
- NTLS: No data found for this date range, symbol may be delisted
- BGPIQ: No data found, symbol may be delisted
- CMCSK: No data found for this date range, symbol may be delisted
- STRZB: No data found for this date range, symbol may be delisted
- NVLS: No data found for this date range, symbol may be delisted
- CTRX: No data found for this date range, symbol may be delisted
- GTK: No data found for this date range, symbol may be delisted
- MNT: No data found for this date range, symbol may be delisted
- FSYS: No data found, symbol may be delisted
- ARG: No data found for this date range, symbol may be delisted
- OMM: No data found for this date range, symbol may be delisted

- MDP: No data found for this date range, symbol may be delisted
- ZLC: No data found for this date range, symbol may be delisted
- MRX: No data found for this date range, symbol may be delisted
- KSWS: No data found for this date range, symbol may be delisted
- AZR: No data found for this date range, symbol may be delisted
- QCOR: No data found for this date range, symbol may be delisted
- BLUD: No data found for this date range, symbol may be delisted
- FLI: No data found for this date range, symbol may be delisted
- MER: No data found for this date range, symbol may be delisted
- PSSI: No data found for this date range, symbol may be delisted
- NST: No data found for this date range, symbol may be delisted
- KL: No data found, symbol may be delisted
- EPHC: No data found for this date range, symbol may be delisted
- DLM: No data found for this date range, symbol may be delisted
- RCRC: No data found for this date range, symbol may be delisted
- BPO: No data found for this date range, symbol may be delisted
- BDG: Data doesn't exist for startDate = 1076367600, endDate = 1644447600
- HUSKF: No data found, symbol may be delisted
- NVE: No data found for this date range, symbol may be delisted
- PDE: No data found for this date range, symbol may be delisted
- DUNDF: No data found, symbol may be delisted
- DTGF: No data found, symbol may be delisted
- MWP: No data found for this date range, symbol may be delisted
- HTS: No data found for this date range, symbol may be delisted
- KOG: No data found for this date range, symbol may be delisted
- PNRA: No data found for this date range, symbol may be delisted
- CHTT: No data found for this date range, symbol may be delisted
- PGL: No data found for this date range, symbol may be delisted
- DNKN: No data found, symbol may be delisted
- Q: No data found for this date range, symbol may be delisted
- WLP: No data found for this date range, symbol may be delisted
- LDL: No data found, symbol may be delisted
- POT: No data found for this date range, symbol may be delisted
- CBM: No data found, symbol may be delisted
- MHS: No data found for this date range, symbol may be delisted
- EDE: No data found for this date range, symbol may be delisted
- MFW: No data found for this date range, symbol may be delisted
- MFE: No data found for this date range, symbol may be delisted
- RTN: No data found, symbol may be delisted
- CNL: No data found for this date range, symbol may be delisted
- RAH: No data found for this date range, symbol may be delisted
- ISIL: No data found for this date range, symbol may be delisted
- PPS: No data found for this date range, symbol may be delisted
- ALSK: No data found, symbol may be delisted
- EAS: No data found for this date range, symbol may be delisted
- ANCUF: No data found, symbol may be delisted
- CNW: No data found for this date range, symbol may be delisted
- TIVO: No data found, symbol may be delisted

- IQNT: No data found for this date range, symbol may be delisted
- GGP: No data found for this date range, symbol may be delisted
- FSH: No data found for this date range, symbol may be delisted
- CEC: No data found for this date range, symbol may be delisted
- SAPE: No data found for this date range, symbol may be delisted
- PPDI: No data found for this date range, symbol may be delisted
- CHSI: No data found for this date range, symbol may be delisted
- STMP: No data found, symbol may be delisted
- PIRRQ: No data found, symbol may be delisted
- KPP: No data found for this date range, symbol may be delisted
- MOLX: No data found for this date range, symbol may be delisted
- AMMD: No data found for this date range, symbol may be delisted
- TKLC: No data found for this date range, symbol may be delisted
- SPLS: No data found for this date range, symbol may be delisted
- HTSI: No data found for this date range, symbol may be delisted
- TXU: No data found for this date range, symbol may be delisted
- EMC: No data found for this date range, symbol may be delisted
- KMP: No data found for this date range, symbol may be delisted
- PGN: No data found for this date range, symbol may be delisted
- CPN: No data found, symbol may be delisted
- APU: No data found, symbol may be delisted
- MVK: No data found for this date range, symbol may be delisted
- FNSR: No data found, symbol may be delisted
- WDR: No data found, symbol may be delisted
- LNY: No data found for this date range, symbol may be delisted
- ITC: No data found for this date range, symbol may be delisted
- CYSVF: No data found, symbol may be delisted
- PBG: No data found for this date range, symbol may be delisted
- DRC: No data found for this date range, symbol may be delisted
- BEC: No data found for this date range, symbol may be delisted
- ARD: No data found, symbol may be delisted
- XEC: No data found, symbol may be delisted
- VPHM: No data found for this date range, symbol may be delisted
- RAVN: No data found, symbol may be delisted
- RTEC: No data found, symbol may be delisted
- CNXM: No data found, symbol may be delisted
- RPAI: No data found, symbol may be delisted
- NTRI: No data found, symbol may be delisted
- ROGFF: No data found, symbol may be delisted
- WBMD: No data found for this date range, symbol may be delisted
- CMX: No data found for this date range, symbol may be delisted
- LTM: No data found, symbol may be delisted
- BGGSQ: No data found, symbol may be delisted
- TOUSQ: No data found, symbol may be delisted
- LM: No data found, symbol may be delisted
- MGG: No data found for this date range, symbol may be delisted
- ABI: No data found for this date range, symbol may be delisted
- ABVT: No data found for this date range, symbol may be delisted

- IDC: No data found for this date range, symbol may be delisted
- MOLXA: No data found for this date range, symbol may be delisted
- GMCR: No data found for this date range, symbol may be delisted
- PMC: No data found for this date range, symbol may be delisted
- WFM: No data found for this date range, symbol may be delisted
- HAR: No data found for this date range, symbol may be delisted
- HPY: No data found for this date range, symbol may be delisted
- AXE: No data found, symbol may be delisted
- THI: No data found for this date range, symbol may be delisted
- DNEX: No data found for this date range, symbol may be delisted
- JW: No data found, symbol may be delisted
- RAI: No data found for this date range, symbol may be delisted
- STB: No data found, symbol may be delisted
- COGN: No data found for this date range, symbol may be delisted
- TCP: No data found, symbol may be delisted
- MV: No data found for this date range, symbol may be delisted
- CKR: No data found for this date range, symbol may be delisted
- CATM: No data found, symbol may be delisted
- HSNI: No data found for this date range, symbol may be delisted
- TWC: No data found for this date range, symbol may be delisted
- WYE: No data found for this date range, symbol may be delisted
- MNIQQ: No data found, symbol may be delisted
- HMA: Data doesn't exist for startDate = 1076367600, endDate = 1644447600
- ARJ: No data found for this date range, symbol may be delisted
- TIF: No data found, symbol may be delisted
- HUG: No data found for this date range, symbol may be delisted
- HMSY: No data found, symbol may be delisted
- JOSB: No data found for this date range, symbol may be delisted
- DLLR: No data found for this date range, symbol may be delisted
- PTRY: No data found for this date range, symbol may be delisted
- BCR: No data found for this date range, symbol may be delisted
- SNDK: No data found for this date range, symbol may be delisted
- CNVR: No data found for this date range, symbol may be delisted
- BXG: No data found, symbol may be delisted
- ANST: No data found for this date range, symbol may be delisted
- HF: No data found, symbol may be delisted
- RIC: No data found for this date range, symbol may be delisted
- FRK: No data found for this date range, symbol may be delisted
- NXY: No data found for this date range, symbol may be delisted
- CEB: No data found for this date range, symbol may be delisted
- SYKE: No data found, symbol may be delisted
- LGCYQ: No data found, symbol may be delisted
- MAUXF: No data found for this date range, symbol may be delisted
- POPE: No data found, symbol may be delisted
- PNY: No data found for this date range, symbol may be delisted
- CRDN: No data found for this date range, symbol may be delisted
- ANN: No data found for this date range, symbol may be delisted
- PIXR: No data found for this date range, symbol may be delisted

- HSP: No data found for this date range, symbol may be delisted
- PAS: No data found for this date range, symbol may be delisted
- FUR: No data found for this date range, symbol may be delisted
- GLYT: No data found for this date range, symbol may be delisted
- MEDQ: No data found for this date range, symbol may be delisted
- BRL: No data found for this date range, symbol may be delisted
- WGR: No data found for this date range, symbol may be delisted
- ARRO: No data found for this date range, symbol may be delisted
- SIAL: No data found for this date range, symbol may be delisted
- JAS: No data found for this date range, symbol may be delisted
- EPB: No data found for this date range, symbol may be delisted
- LAACZ: No data found, symbol may be delisted
- UTIW: No data found for this date range, symbol may be delisted
- PNP: No data found for this date range, symbol may be delisted
- PIKE: No data found, symbol may be delisted
- AGU: No data found for this date range, symbol may be delisted
- ADVBQ: No data found for this date range, symbol may be delisted
- VMED: No data found for this date range, symbol may be delisted
- IWOV: No data found for this date range, symbol may be delisted
- CLC: No data found for this date range, symbol may be delisted
- TEG: No data found for this date range, symbol may be delisted
- PTV: No data found for this date range, symbol may be delisted
- GYMB: No data found for this date range, symbol may be delisted
- STJ: No data found for this date range, symbol may be delisted
- TLRDQ: No data found, symbol may be delisted
- HSH: No data found for this date range, symbol may be delisted
- DGIT: No data found for this date range, symbol may be delisted
- MIK: No data found, symbol may be delisted
- UIC: No data found for this date range, symbol may be delisted
- STR: No data found for this date range, symbol may be delisted
- ALD: No data found for this date range, symbol may be delisted
- BEZ: No data found for this date range, symbol may be delisted
- OUTR: No data found for this date range, symbol may be delisted
- PXP: No data found for this date range, symbol may be delisted
- MOG: No data found for this date range, symbol may be delisted
- FEIC: No data found for this date range, symbol may be delisted
- RSE: No data found for this date range, symbol may be delisted
- MCRL: No data found for this date range, symbol may be delisted
- ELNK: No data found for this date range, symbol may be delisted
- BNI: No data found for this date range, symbol may be delisted
- TECD: No data found, symbol may be delisted
- CBST: No data found for this date range, symbol may be delisted
- DFT: No data found for this date range, symbol may be delisted
- HITK: No data found for this date range, symbol may be delisted
- KFN: No data found for this date range, symbol may be delisted
- LGF: No data found for this date range, symbol may be delisted
- KTO: No data found for this date range, symbol may be delisted
- TLP: No data found, symbol may be delisted

```
- LDG: No data found for this date range, symbol may be delisted
- ISCA: No data found, symbol may be delisted
- EV: No data found, symbol may be delisted
- ROH: No data found for this date range, symbol may be delisted
- PRXL: No data found for this date range, symbol may be delisted
- TCO: No data found, symbol may be delisted
- ERT: No data found for this date range, symbol may be delisted
- ORB: No data found for this date range, symbol may be delisted
- RSTI: No data found for this date range, symbol may be delisted
- ASNAQ: No data found, symbol may be delisted
- IPHS: No data found, symbol may be delisted
- KNL: No data found, symbol may be delisted
- BTUUQ: No data found for this date range, symbol may be delisted
- VCI: No data found for this date range, symbol may be delisted
- POM: No data found for this date range, symbol may be delisted
- LCAV: No data found for this date range, symbol may be delisted
- QLGC: No data found for this date range, symbol may be delisted
- LXK: No data found for this date range, symbol may be delisted
- PVAHQ: No data found for this date range, symbol may be delisted
- ASEI: No data found for this date range, symbol may be delisted
- POG: No data found for this date range, symbol may be delisted
- OKS: No data found for this date range, symbol may be delisted
- TFCFA: No data found, symbol may be delisted
- WNR: No data found for this date range, symbol may be delisted
```

[20]:
```
df_yahoo.head(5)
```

[20]:

| Date | A | AACB | AAIC | AAP | AAPL | AAT | AAWW | … |
|---|---|---|---|---|---|---|---|---|
| 2004-02-09 00:00:00 | 26.523605 | NaN | 512.0 | 27.933332 | 0.410714 | NaN | NaN | … |
| 2004-02-10 00:00:00 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | … |
| 2004-02-11 00:00:00 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | … |
| 2004-02-12 00:00:00 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | … |
| 2004-02-13 00:00:00 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | … |

| Date | ABBV | ABC | ABCD | … | XTO | XYL | YELL | YLWDF | … |
|---|---|---|---|---|---|---|---|---|---|
| 2004-02-09 00:00:00 | NaN | 14.1525 | NaN | … | NaN | NaN | 245775.0 | NaN | … |
| 2004-02-10 00:00:00 | NaN | NaN | NaN | … | NaN | NaN | NaN | NaN | … |
| 2004-02-11 00:00:00 | NaN | NaN | NaN | … | NaN | NaN | NaN | NaN | … |
| 2004-02-12 00:00:00 | NaN | NaN | NaN | … | NaN | NaN | NaN | NaN | … |
| 2004-02-13 00:00:00 | NaN | NaN | NaN | … | NaN | NaN | NaN | NaN | … |

| Date | YUM | ZBRA | ZD | ZLC | ZQKSQ | ZTS |
|---|---|---|---|---|---|---|
| 2004-02-09 00:00:00 | 12.692308 | 46.453335 | 10.186957 | NaN | NaN | NaN |
| 2004-02-10 00:00:00 | NaN | NaN | NaN | NaN | NaN | NaN |

```
2004-02-11 00:00:00        NaN        NaN        NaN  NaN    NaN  NaN
2004-02-12 00:00:00        NaN        NaN        NaN  NaN    NaN  NaN
2004-02-13 00:00:00        NaN        NaN        NaN  NaN    NaN  NaN

[5 rows x 1804 columns]
```

[21]:
```
print('Shape of dataset from Yahoo: {}'.format(df_yahoo.shape))
```

```
Shape of dataset from Yahoo: (4605, 1804)
```

[22]:
```
columns_nan = df_yahoo.columns[df_yahoo.isna().all()].tolist()
print('Number of missing index: {}'.format(len(columns_nan)))
```

```
Number of missing index: 397
```

### 1.0.5 Preparation financial data

[23]:
```
# create copy DataFrame
data = df_yahoo.copy(deep=True)
```

[24]:
```
# remove missing values from columns
data.dropna(how='any', axis=1, thresh=1, inplace=True)
# remove missing values from rows
data.dropna(how='any', axis=0, thresh=3, inplace=True)
```

[25]:
```
data_ = data.reset_index()
```

[26]:
```
# unpivot a DataFrame
data2 = pd.melt(data_, id_vars='Date', value_vars=data.columns.to_list())
data2
```

[26]:
```
              Date variable       value
0       2004-02-09        A   26.523605
1       2004-02-16        A   25.071531
2       2004-02-23        A   24.456366
3       2004-03-01        A   24.899857
4       2004-03-08        A   22.639484
…              …      …          …
1703872 2022-01-10      ZTS  206.179993
1703873 2022-01-17      ZTS  200.330002
1703874 2022-01-24      ZTS  195.300003
1703875 2022-01-31      ZTS  199.539993
1703876 2022-02-07      ZTS  202.289993

[1703877 rows x 3 columns]
```

[27]:
```
# Return the number of missing values
data2.isnull().sum()
```

[27]:
```
Date             0
variable         0
value       561605
dtype: int64
```

[28]:
```
print('Number of data without missing: {}'.format(len(data2) - data2.value.
 ↪isnull().sum()))
```

```
Number of data without missing: 1142272
```

[29]:
```
# removing missing values
data2.dropna(inplace=True)
```

[30]:
```
data2.isnull().sum()
```

[30]:
```
Date        0
variable    0
value       0
dtype: int64
```

[31]:
```
print('Number of weekly: {}'.format(data2.Date.nunique()))
```

```
Number of weekly: 1211
```

[32]:
```
print('Shape of dataset from Yahoo without empty index: {}'.format(data2.shape))
```

```
Shape of dataset from Yahoo without empty index: (1142272, 3)
```

### 1.0.6 Calculation of the rate of return

[33]:
```
#create empty columns
data2["return_rate"] = np.nan

#create new DataFrame
df_rr = pd.DataFrame(columns=['Date', 'symbol', 'value', 'return_rate'])

#create symbol list
symbols = data2["variable"].unique().tolist()

for sym in symbols:

    data_symbol = data2.loc[data2["variable"] == sym]

    for i in range(0, len(data_symbol)):
        if i+1<len(data_symbol):
```

```
            data_symbol["return_rate"].iloc[i+1] = (data_symbol["value"].
↪iloc[i+1]/data_symbol["value"].iloc[i])-1

    df_rr = pd.concat([df_rr, data_symbol])

df_rr
```

```
C:\Users\agnie\anaconda3\lib\site-packages\pandas\core\indexing.py:1637:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_block(indexer, value, name)
C:\Users\agnie\anaconda3\lib\site-packages\pandas\core\indexing.py:692:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  iloc._setitem_with_indexer(indexer, value, self.name)
<ipython-input-33-8d70a402dcdf>:16: RuntimeWarning: divide by zero encountered
in double_scalars
  data_symbol["return_rate"].iloc[i+1] =
(data_symbol["value"].iloc[i+1]/data_symbol["value"].iloc[i])-1
```

[33]:

| | Date | symbol | value | return_rate | variable |
|---|---|---|---|---|---|
| 0 | 2004-02-09 | NaN | 26.523605 | NaN | A |
| 1 | 2004-02-16 | NaN | 25.071531 | -0.054746 | A |
| 2 | 2004-02-23 | NaN | 24.456366 | -0.024536 | A |
| 3 | 2004-03-01 | NaN | 24.899857 | 0.018134 | A |
| 4 | 2004-03-08 | NaN | 22.639484 | -0.090779 | A |
| … | … | … | … | … | … |
| 1703872 | 2022-01-10 | NaN | 206.179993 | -0.023260 | ZTS |
| 1703873 | 2022-01-17 | NaN | 200.330002 | -0.028373 | ZTS |
| 1703874 | 2022-01-24 | NaN | 195.300003 | -0.025109 | ZTS |
| 1703875 | 2022-01-31 | NaN | 199.539993 | 0.021710 | ZTS |
| 1703876 | 2022-02-07 | NaN | 202.289993 | 0.013782 | ZTS |

[1142272 rows x 5 columns]

[34]:
```
# test
df_rr.loc[df_rr['variable'] == "SU"]
```

[34]:

| | Date | symbol | value | return_rate | variable |
|---|---|---|---|---|---|
| 1441090 | 2004-02-09 | NaN | 12.830000 | NaN | SU |
| 1441091 | 2004-02-16 | NaN | 12.580000 | -0.019486 | SU |

```
1441092 2004-02-23    NaN  12.990000     0.032591       SU
1441093 2004-03-01    NaN  14.175000     0.091224       SU
1441094 2004-03-08    NaN  13.565000    -0.043034       SU
…              …    …         …           …       …
1442296 2022-01-10    NaN  28.230000     0.062877       SU
1442297 2022-01-17    NaN  27.070000    -0.041091       SU
1442298 2022-01-24    NaN  28.299999     0.045438       SU
1442299 2022-01-31    NaN  28.719999     0.014841       SU
1442300 2022-02-07    NaN  28.889999     0.005919       SU

[940 rows x 5 columns]
```

### 1.0.7 Preparation of the target dataset

[35]:
```
# create copy DataFrame
df_rr_ = df_rr[["Date", "variable", "value", "return_rate"]].copy(deep=True)
```

[36]:
```
# add 2 days
df_rr_['Date'] = df_rr_['Date'] + datetime.timedelta(days=2)
```

[37]:
```
# test
df_rr_.loc[df_rr_['variable'] == "SU"]
```

[37]:
```
              Date variable      value  return_rate
1441090 2004-02-11       SU  12.830000          NaN
1441091 2004-02-18       SU  12.580000    -0.019486
1441092 2004-02-25       SU  12.990000     0.032591
1441093 2004-03-03       SU  14.175000     0.091224
1441094 2004-03-10       SU  13.565000    -0.043034
…              …      …         …           …
1442296 2022-01-12       SU  28.230000     0.062877
1442297 2022-01-19       SU  27.070000    -0.041091
1442298 2022-01-26       SU  28.299999     0.045438
1442299 2022-02-02       SU  28.719999     0.014841
1442300 2022-02-09       SU  28.889999     0.005919

[940 rows x 4 columns]
```

[38]:
```
data3 = pd.merge(df, df_rr_,  how='left', left_on=['date','symbol'], right_on =␣
↪['Date','variable'])
```

[39]:
```
# Remove rows which contains null values.
dataset = data3.dropna(subset=['variable','value'])

# Remove column data
dataset.drop('variable', axis=1, inplace=True)
```

```
# Rename column value
dataset.rename(columns={'value': 'close'}, inplace=True)

dataset.set_index('Date', inplace=True)
```

```
C:\Users\agnie\anaconda3\lib\site-packages\pandas\core\frame.py:4308:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  return super().drop(
C:\Users\agnie\anaconda3\lib\site-packages\pandas\core\frame.py:4441:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  return super().rename(
```

[40]:
```
print("Old data frame length:", len(df),
      "\nNew data frame length:", len(dataset),
      "\nNumber of rows deleted: ",
       (len(df)-len(dataset)))
```

```
Old data frame length: 37360
New data frame length: 30466
Number of rows deleted:  6894
```

[42]:
```
print("Number of all unique symbols:", df.symbol.nunique(),
      "\nNumber of missing symbols:", len(columns_nan),
      "\nNumber of symbols in dataset: ",
       (df.symbol.nunique()-len(columns_nan)))
```

```
Number of all unique symbols: 1804
Number of missing symbols: 397
Number of symbols in dataset:  1407
```

[43]:
```
dataset.head(5)
```

[43]:

| Date | symbol | sector | score | close | return_rate |
|---|---|---|---|---|---|
| 2004-02-11 | SU | Energy Minerals | 0.953727 | 12.830000 | NaN |
| 2004-02-11 | GGG | Producer Manufacturing | 0.952753 | 9.322222 | NaN |
| 2004-02-11 | CWT | Utilities | 0.934181 | 14.245000 | NaN |
| 2004-02-11 | BLL | Process Industries | 0.922862 | 8.012500 | NaN |
| 2004-02-11 | APA | Energy Minerals | 0.912117 | 39.509998 | NaN |

### 1.0.8 Download CSV

[ ]:
```
def create_download_link( df, title = "Download CSV file", filename = "data.
↪csv"):
    csv = df.to_csv()
    b64 = base64.b64encode(csv.encode())
    payload = b64.decode()
    html = '<a download="{filename}" href="data:text/csv;base64,{payload}"␣
↪target="_blank">{title}</a>'
    html = html.format(payload=payload,title=title,filename=filename)
    return HTML(html)

create_download_link(dataset)
```