data_preparation

June 8, 2022

1 Data preparation

1.0.1 Install and importing libraries

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

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting yfinance
  Downloading yfinance-0.1.70-py2.py3-none-any.whl (26 kB)
Requirement already satisfied: multitasking>=0.0.7 in
/usr/local/lib/python3.7/dist-packages (from yfinance) (0.0.10)
Requirement already satisfied: pandas>=0.24.0 in /usr/local/lib/python3.7/dist-
packages (from yfinance) (1.3.5)
Collecting lxml>=4.5.1
  Downloading lxml-4.9.0-cp37-cp37m-manylinux_2_17_x86_64.manylinux2014_x86_64.m
anylinux_2_24_x86_64.whl (6.4 MB)
     |                 | 6.4 MB 6.4 MB/s
Collecting requests>=2.26
  Downloading requests-2.27.1-py2.py3-none-any.whl (63 kB)
     |                 | 63 kB 1.1 MB/s
Requirement already satisfied: numpy>=1.15 in
/usr/local/lib/python3.7/dist-packages (from yfinance) (1.21.6)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-
packages (from pandas>=0.24.0->yfinance) (2022.1)
Requirement already satisfied: python-dateutil>=2.7.3 in
/usr/local/lib/python3.7/dist-packages (from pandas>=0.24.0->yfinance) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-
packages (from python-dateutil>=2.7.3->pandas>=0.24.0->yfinance) (1.15.0)
Requirement already satisfied: charset-normalizer~=2.0.0 in
/usr/local/lib/python3.7/dist-packages (from requests>=2.26->yfinance) (2.0.12)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.7/dist-packages (from requests>=2.26->yfinance)
(2022.5.18.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.7/dist-
packages (from requests>=2.26->yfinance) (2.10)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/usr/local/lib/python3.7/dist-packages (from requests>=2.26->yfinance) (1.24.3)
```

```
Installing collected packages: requests, lxml, yfinance
  Attempting uninstall: requests
    Found existing installation: requests 2.23.0
    Uninstalling requests-2.23.0:
      Successfully uninstalled requests-2.23.0
  Attempting uninstall: lxml
    Found existing installation: lxml 4.2.6
    Uninstalling lxml-4.2.6:
      Successfully uninstalled lxml-4.2.6
ERROR: pip's dependency resolver does not currently take into account all
the packages that are installed. This behaviour is the source of the following
dependency conflicts.
google-colab 1.0.0 requires requests~=2.23.0, but you have requests 2.27.1 which
is incompatible.
datascience 0.10.6 requires folium==0.2.1, but you have folium 0.8.3 which is
incompatible.
Successfully installed lxml-4.9.0 requests-2.27.1 yfinance-0.1.70
```

[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  \
0      10:01:54.481 77425 [77425-thread-2] INFO  a.s…  2004-02-11     SU
1      10:01:54.481 77425 [77425-thread-2] INFO  a.s…  2004-02-11    GGG
2      10:01:54.481 77425 [77425-thread-2] INFO  a.s…  2004-02-11    WGR
3      10:01:54.481 77425 [77425-thread-2] INFO  a.s…  2004-02-11    CWT
4      10:01:54.481 77425 [77425-thread-2] INFO  a.s…  2004-02-11    BLL
…                                                    …          …      …
37355  10:27:03.049 77425 [77425-thread-2] INFO  a.s…  2022-02-09    PEP
37356  10:27:03.049 77425 [77425-thread-2] INFO  a.s…  2022-02-09   SSNC
37357  10:27:03.049 77425 [77425-thread-2] INFO  a.s…  2022-02-09    GEF
```

```
37358  10:27:03.049 77425 [77425-thread-2] INFO  a.s…  2022-02-09    DPZ
37359  10:27:03.049 77425 [77425-thread-2] INFO  a.s…  2022-02-09  LIFZF

      symbol2                  sector    number     score
0          SU         Energy Minerals  GN63J3-R  0.953727
1         GGG  Producer Manufacturing  H5490W-R  0.952753
2         WGR         Energy Minerals  V0622Q-R  0.947634
3         CWT               Utilities  GSWXLY-R  0.934181
4         BLL      Process Industries  VFT0VQ-R  0.922862
…         …                       …         …         …
37355     PEP   Consumer Non-Durables  PPCTFP-R  0.701507
37356    SSNC     Technology Services  G92RX2-R  0.701123
37357     GEF      Process Industries  MPX0N4-R  0.697954
37358     DPZ       Consumer Services  F05QG0-R  0.697741
37359   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]:

```
           symbol                  sector     score
date
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
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py: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.
```

[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="1d")['Close']
```

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

413 Failed downloads:
- NTLS: No data found for this date range, symbol may be delisted
- CHSI: No data found for this date range, symbol may be delisted
- EE: Data doesn't exist for startDate = 1076371200, endDate = 1644451200
- CATM: No data found, symbol may be delisted
- BBX: No data found, symbol may be delisted
- WYE: No data found for this date range, symbol may be delisted
- DFODQ: No data found, symbol may be delisted
- LAACZ: No data found, symbol may be delisted
- LEARQ: No data found, symbol may be delisted
- QLGC: No data found for this date range, symbol may be delisted
- ABI: No data found for this date range, symbol may be delisted
- PIKE: No data found for this date range, symbol may be delisted
- BPL: No data found, symbol may be delisted
- TLRDQ: No data found, symbol may be delisted
- NGLS: No data found for this date range, symbol may be delisted
- RAVN: No data found, symbol may be delisted
- KL: No data found, symbol may be delisted
- LCAV: No data found for this date range, symbol may be delisted
- WWIN: No data found, symbol may be delisted
- HPY: No data found for this date range, symbol may be delisted
- SPSS: No data found for this date range, symbol may be delisted
- CNW: No data found for this date range, symbol may be delisted
- KEM: No data found, symbol may be delisted
- STB: No data found, symbol may be delisted
- ENSI: No data found for this date range, symbol may be delisted
- MCRL: No data found for this date range, symbol may be delisted
- ORB: No data found for this date range, symbol may be delisted
- PIRRQ: No data found, symbol may be delisted
- OAK: No data found, symbol may be delisted
- WNR: No data found for this date range, symbol may be delisted
- TSS: No data found, symbol may be delisted
- MXIM: No data found, symbol may be delisted
- ZLC: No data found for this date range, symbol may be delisted
- CORE: No data found, symbol may be delisted
- CMLEF: No data found, symbol may be delisted
- RSTI: No data found for this date range, symbol may be delisted
- ISIL: No data found for this date range, symbol may be delisted
- DFT: No data found for this date range, symbol may be delisted
- BPMP: No data found, symbol may be delisted
- MVK: No data found for this date range, symbol may be delisted
- TFCFA: No data found, symbol may be delisted
- IPHS: No data found, symbol may be delisted
- BLUD: No data found for this date range, symbol may be delisted
- CMCSK: No data found for this date range, symbol may be delisted
- Q: No data found for this date range, symbol may be delisted
- GSF: No data found for this date range, symbol may be delisted
- HMSY: No data found, symbol may be delisted
- EFXFF: No data found, symbol may be delisted
- INCLF: No data found, symbol may be delisted
- VIAB: No data found, symbol may be delisted
- UFS: No data found, symbol may be delisted
- LNY: No data found for this date range, symbol may be delisted
- OKS: No data found for this date range, symbol may be delisted
- PGN: No data found for this date range, symbol may be delisted
- MCF: No data found, symbol may be delisted
- CBST: No data found for this date range, symbol may be delisted
- SRR: No data found for this date range, symbol may be delisted
- POG: No data found for this date range, symbol may be delisted
- BXG: No data found, symbol may be delisted
- AYE: No data found for this date range, symbol may be delisted
- POT: No data found for this date range, symbol may be delisted
- WLP: No data found for this date range, symbol may be delisted
- GYMB: No data found for this date range, symbol may be delisted
- MWIV: No data found for this date range, symbol may be delisted
- TECD: No data found, symbol may be delisted
- EPE: No data found, symbol may be delisted
- ANST: No data found for this date range, symbol may be delisted
- XLNX: No data found, symbol may be delisted
- WFSI: No data found for this date range, symbol may be delisted
- HUSKF: No data found, symbol may be delisted
- TRK: No data found, symbol may be delisted
- KSWS: No data found for this date range, symbol may be delisted
- FSH: No data found for this date range, symbol may be delisted
- LDL: No data found, symbol may be delisted
- CRDN: No data found for this date range, symbol may be delisted
- GLYT: No data found for this date range, symbol may be delisted
- BRSS: No data found, symbol may be delisted
- RPAI: No data found, symbol may be delisted
- BLKIA: No data found for this date range, symbol may be delisted
- MFW: No data found for this date range, symbol may be delisted
- KNL: No data found, symbol may be delisted
- LTM: No data found, symbol may be delisted
- ISCA: No data found, symbol may be delisted
- JW: No data found, symbol may be delisted
- NTRI: No data found, symbol may be delisted
- RATE: No data found for this date range, symbol may be delisted
- VCI: No data found for this date range, symbol may be delisted
- GTRC: No data found for this date range, symbol may be delisted
- VIAC: No data found, symbol may be delisted
- TIVO: No data found, symbol may be delisted
- NXY: No data found for this date range, symbol may be delisted
- KOG: No data found for this date range, symbol may be delisted
- HSNI: No data found for this date range, symbol may be delisted
- PCL: No data found for this date range, symbol may be delisted
- FRK: No data found for this date range, symbol may be delisted
- THI: No data found for this date range, symbol may be delisted
- DNKN: No data found, symbol may be delisted
- PCZ: No data found for this date range, symbol may be delisted
- GISX: No data found for this date range, symbol may be delisted
- LGCYQ: No data found, symbol may be delisted
- ADVBQ: No data found for this date range, symbol may be delisted
- QCOR: No data found for this date range, symbol may be delisted
- IQNT: No data found for this date range, symbol may be delisted
- ACAS: No data found for this date range, symbol may be delisted
- PSSI: No data found for this date range, symbol may be delisted
- HNH: No data found for this date range, symbol may be delisted
- DRC: No data found for this date range, symbol may be delisted
- NST: No data found for this date range, symbol may be delisted
- PSUNQ: No data found for this date range, symbol may be delisted
- FRX: No data found, symbol may be delisted
- LUFK: No data found for this date range, symbol may be delisted
- ASEI: No data found for this date range, symbol may be delisted
- ANN: No data found for this date range, symbol may be delisted
- HTS: No data found for this date range, symbol may be delisted
- TLM: No data found for this date range, symbol may be delisted
- PBG: No data found for this date range, symbol may be delisted
- BGPIQ: No data found, symbol may be delisted
- CLP: No data found for this date range, symbol may be delisted
- EMC: No data found for this date range, symbol may be delisted
- AIPC: No data found, symbol may be delisted
- AEA: No data found for this date range, symbol may be delisted
- PPDI: No data found for this date range, symbol may be delisted
- BYI: No data found for this date range, symbol may be delisted
- TOUSQ: No data found, symbol may be delisted
- RTN: No data found, symbol may be delisted
- BKC: No data found for this date range, symbol may be delisted
- NVLS: No data found for this date range, symbol may be delisted
- HRC: No data found, symbol may be delisted
- MNIQQ: No data found, symbol may be delisted
- OCR: No data found for this date range, symbol may be delisted
- CMD: No data found, symbol may be delisted
- TIN: No data found for this date range, symbol may be delisted
- HCR: No data found, symbol may be delisted
- GXDX: No data found for this date range, symbol may be delisted
- VARI: No data found for this date range, symbol may be delisted
- MOG: No data found for this date range, symbol may be delisted
- HFC: No data found, symbol may be delisted
- CKR: No data found for this date range, symbol may be delisted
- ZQKSQ: No data found for this date range, symbol may be delisted
- RTI: No data found for this date range, symbol may be delisted
- EPB: No data found for this date range, symbol may be delisted
- ANH: No data found, symbol may be delisted
- BRL: No data found for this date range, symbol may be delisted
- TWC: No data found for this date range, symbol may be delisted
- SAPE: No data found for this date range, symbol may be delisted
- WRI: No data found, symbol may be delisted
- WLSM: No data found, symbol may be delisted
- HDS: No data found, symbol may be delisted
- MER: No data found for this date range, symbol may be delisted
- ARRO: No data found for this date range, symbol may be delisted
- CVA: No data found, symbol may be delisted
- DPL: No data found for this date range, symbol may be delisted
- ADS: No data found, symbol may be delisted
- AACB: No data found for this date range, symbol may be delisted
- KDN: No data found for this date range, symbol may be delisted
- SCLN: No data found for this date range, symbol may be delisted
- FTO: No data found for this date range, symbol may be delisted
- CTRX: No data found for this date range, symbol may be delisted
- CHG: No data found for this date range, symbol may be delisted
- ARTI: No data found for this date range, symbol may be delisted
- AIRM: No data found for this date range, symbol may be delisted
- TCO: No data found, symbol may be delisted
- ASNAQ: No data found, symbol may be delisted
- PMZFF: No data found, symbol may be delisted
- PPP: No data found for this date range, symbol may be delisted
- ANCUF: No data found, symbol may be delisted
- TCP: No data found, symbol may be delisted
- PNP: No data found for this date range, symbol may be delisted
- SIRO: No data found for this date range, symbol may be delisted
- PXP: No data found for this date range, symbol may be delisted
- OVTI: No data found for this date range, symbol may be delisted
- HNR: No data found for this date range, symbol may be delisted
- EMS: No data found for this date range, symbol may be delisted
- TKLC: No data found for this date range, symbol may be delisted
- ARDNA: No data found for this date range, symbol may be delisted
- WBMD: No data found for this date range, symbol may be delisted
- HSH: No data found for this date range, symbol may be delisted
- APU: No data found, symbol may be delisted
- CUB: No data found, symbol may be delisted
- BOBE: No data found for this date range, symbol may be delisted
- TE: No data found for this date range, symbol may be delisted
- EQY: No data found for this date range, symbol may be delisted
- KKD: No data found for this date range, symbol may be delisted
- EDE: No data found for this date range, symbol may be delisted
- JOSB: No data found for this date range, symbol may be delisted
- FUR: No data found for this date range, symbol may be delisted
- PPS: No data found for this date range, symbol may be delisted
- DNEX: No data found for this date range, symbol may be delisted
- KWD: No data found for this date range, symbol may be delisted
- CBM: No data found, symbol may be delisted
- OMM: No data found for this date range, symbol may be delisted
- MCRS: No data found for this date range, symbol may be delisted
- MNR: No data found, symbol may be delisted
- FEIC: No data found for this date range, symbol may be delisted
- TFCF: No data found, symbol may be delisted
- MV: No data found for this date range, symbol may be delisted
- MRX: No data found for this date range, symbol may be delisted
- STRZB: No data found for this date range, symbol may be delisted
- ACAT: No data found for this date range, symbol may be delisted
- ITC: No data found for this date range, symbol may be delisted
- DTGF: No data found, symbol may be delisted
- BRCM: No data found for this date range, symbol may be delisted
- BGGSQ: No data found, symbol may be delisted
- FNSR: No data found, symbol may be delisted
- KWKAQ: No data found for this date range, symbol may be delisted
- MHS: No data found for this date range, symbol may be delisted
- ETFC: No data found, symbol may be delisted
- RTEC: No data found, symbol may be delisted
- MEDQ: No data found for this date range, symbol may be delisted
- ALSK: No data found, symbol may be delisted
- BMC: No data found for this date range, symbol may be delisted
- HMA: Data doesn't exist for startDate = 1076371200, endDate = 1644451200
- ASFI: No data found, symbol may be delisted
- CEB: No data found for this date range, symbol may be delisted
- HNZ: No data found, symbol may be delisted
- MHM: No data found for this date range, symbol may be delisted
- GMCR: No data found for this date range, symbol may be delisted
- GAS: No data found for this date range, symbol may be delisted
- ININ: No data found for this date range, symbol may be delisted
- PNY: No data found for this date range, symbol may be delisted
- TQNT: No data found for this date range, symbol may be delisted
- ABVT: No data found for this date range, symbol may be delisted
- DGIT: No data found for this date range, symbol may be delisted
- ARD: No data found, symbol may be delisted
- LO: No data found for this date range, symbol may be delisted
- KRA: No data found, symbol may be delisted
- FDO: No data found for this date range, symbol may be delisted
- CKH: No data found, symbol may be delisted
- RAI: No data found for this date range, symbol may be delisted
- AMMD: No data found for this date range, symbol may be delisted
- VPHM: No data found for this date range, symbol may be delisted
- PGL: No data found for this date range, symbol may be delisted
- TIBX: No data found for this date range, symbol may be delisted
- BEAV: No data found for this date range, symbol may be delisted
- GDI: No data found, symbol may be delisted
- IGTE: No data found for this date range, symbol may be delisted
- BTUUQ: No data found for this date range, symbol may be delisted
- HUG: No data found for this date range, symbol may be delisted
- MDP: No data found for this date range, symbol may be delisted
- ARG: No data found for this date range, symbol may be delisted
- BEC: No data found for this date range, symbol may be delisted
- LM: No data found, symbol may be delisted
- VRX: No data found for this date range, symbol may be delisted
- GCGMF: No data found, symbol may be delisted
- JAH: No data found for this date range, symbol may be delisted
- XTO: No data found for this date range, symbol may be delisted
- NTY: No data found for this date range, symbol may be delisted
- HUB: No data found for this date range, symbol may be delisted
- CSE: No data found for this date range, symbol may be delisted
- COGN: No data found for this date range, symbol may be delisted
- KMP: No data found for this date range, symbol may be delisted
- NVE: No data found for this date range, symbol may be delisted
- CRRC: No data found for this date range, symbol may be delisted
- HOFD: No data found, symbol may be delisted
- CHTT: No data found for this date range, symbol may be delisted
- LDR: No data found for this date range, symbol may be delisted
- CINR: No data found, symbol may be delisted
- IDC: No data found for this date range, symbol may be delisted
- KFN: No data found for this date range, symbol may be delisted
- STU: No data found for this date range, symbol may be delisted
- JH: No data found for this date range, symbol may be delisted
- CEPH: No data found for this date range, symbol may be delisted
- MATK: No data found for this date range, symbol may be delisted
- CTWS: No data found, symbol may be delisted
- CHKAQ: No data found, symbol may be delisted
- XEC: No data found, symbol may be delisted
- CYSVF: No data found, symbol may be delisted
- GTM: No data found for this date range, symbol may be delisted
- ALXN: No data found, symbol may be delisted
- PAS: No data found for this date range, symbol may be delisted
- CTX: No data found for this date range, symbol may be delisted
- HITK: No data found for this date range, symbol may be delisted
- CGX: No data found for this date range, symbol may be delisted
- TLP: No data found, symbol may be delisted
- OUTR: No data found for this date range, symbol may be delisted
- CNL: No data found for this date range, symbol may be delisted
- CY: No data found, symbol may be delisted
- IM: No data found for this date range, symbol may be delisted
- PRX: No data found for this date range, symbol may be delisted
- UIC: No data found for this date range, symbol may be delisted
- VAR: No data found, symbol may be delisted
- AZR: No data found for this date range, symbol may be delisted
- RRD: No data found, symbol may be delisted
- LABL: No data found, symbol may be delisted
- HET: No data found for this date range, symbol may be delisted
- BF: No data found for this date range, symbol may be delisted
- PMC: No data found for this date range, symbol may be delisted
- BDG: Data doesn't exist for startDate = 1076371200, endDate = 1644451200
- STJ: No data found for this date range, symbol may be delisted
- POPE: No data found, symbol may be delisted
- ROGFF: No data found, symbol may be delisted
- PVAHQ: No data found for this date range, symbol may be delisted
- GRP: No data found for this date range, symbol may be delisted
- HF: No data found, symbol may be delisted
- LXK: No data found for this date range, symbol may be delisted
- GR: No data found for this date range, symbol may be delisted
- SIAL: No data found for this date range, symbol may be delisted
- CSFFF: No data found, symbol may be delisted
- BDK: No data found for this date range, symbol may be delisted
- NEWCQ: No data found, symbol may be delisted
- EPHC: No data found for this date range, symbol may be delisted
- ROH: 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
- WGR: No data found for this date range, symbol may be delisted
- CDX: Data doesn't exist for startDate = 1076371200, endDate = 1644451200
- IWOV: No data found for this date range, symbol may be delisted
- MOLXA: No data found for this date range, symbol may be delisted
- AXE: No data found, symbol may be delisted
- NDN: No data found for this date range, symbol may be delisted
- MNT: No data found for this date range, symbol may be delisted
- TTEC: CircuitBreaker 'redis' is OPEN and does not permit further calls
- IEP: CircuitBreaker 'redis' is OPEN and does not permit further calls
- PLMD: No data found for this date range, symbol may be delisted
- PAHC: CircuitBreaker 'redis' is OPEN and does not permit further calls
- GRA: No data found, symbol may be delisted
- WDR: No data found, symbol may be delisted
- EGOV: No data found, 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
- HW: No data found for this date range, symbol may be delisted
- DTV: No data found, symbol may be delisted
- HDLM: No data found, symbol may be delisted
- HOTT: No data found for this date range, symbol may be delisted
- MTSC: No data found, symbol may be delisted
- PVG: No data found, symbol may be delisted
- CNXM: No data found, symbol may be delisted
- RILY: CircuitBreaker 'redis' is OPEN and does not permit further calls
- ACO: No data found for this date range, symbol may be delisted
- HPOL: No data found for this date range, symbol may be delisted
- PCP: No data found for this date range, symbol may be delisted
- NXG: No data found for this date range, symbol may be delisted
- MIK: No data found, symbol may be delisted
- LDG: No data found for this date range, symbol may be delisted
- TXU: CircuitBreaker 'redis' is OPEN and does not permit further calls
- CSX: CircuitBreaker 'redis' is OPEN and does not permit further calls
- FRTA: No data found, symbol may be delisted
- CLC: No data found for this date range, symbol may be delisted
- MRD: No data found for this date range, symbol may be delisted
- GTK: No data found for this date range, symbol may be delisted
- STRZA: CircuitBreaker 'redis' is OPEN and does not permit further calls
- COCOQ: No data found, symbol may be delisted
- ASCA: Data doesn't exist for startDate = 1076371200, endDate = 1644451200
- RSE: No data found for this date range, symbol may be delisted
- RAH: No data found for this date range, symbol may be delisted
- DUNDF: No data found, symbol may be delisted
- BLC: No data found for this date range, symbol may be delisted
- PALDF: No data found, symbol may be delisted
- VIACA: No data found, symbol may be delisted
- ALD: No data found for this date range, symbol may be delisted
- NHP: No data found for this date range, symbol may be delisted
- GGP: No data found for this date range, symbol may be delisted
- ERT: No data found for this date range, symbol may be delisted
- STMP: No data found, symbol may be delisted
- ODSY: No data found for this date range, symbol may be delisted
- FSYS: No data found, symbol may be delisted
- RSHCQ: No data found for this date range, symbol may be delisted
- KCI: No data found for this date range, symbol may be delisted
- LBYYQ: No data found, symbol may be delisted
- HYSL: No data found for this date range, symbol may be delisted
- PETM: No data found for this date range, symbol may be delisted
- SHS: No data found for this date range, symbol may be delisted
- AGN: No data found, symbol may be delisted
- NAFC: No data found for this date range, symbol may be delisted
- CLGX: No data found, symbol may be delisted
- TIF: No data found, symbol may be delisted
- UTIW: No data found for this date range, symbol may be delisted
- AGU: No data found for this date range, symbol may be delisted
- EV: No data found, symbol may be delisted
- BN: No data found for this date range, symbol may be delisted
- QSFT: No data found for this date range, symbol may be delisted
- BNI: No data found for this date range, symbol may be delisted
- PWER: No data found for this date range, symbol may be delisted
- HSP: No data found for this date range, symbol may be delisted
- WFM: No data found for this date range, symbol may be delisted
- UNS: No data found for this date range, symbol may be delisted
- BJS: No data found for this date range, symbol may be delisted
- GVHR: No data found for this date range, symbol may be delisted
- NZ: No data found for this date range, symbol may be delisted
- SYKE: No data found, symbol may be delisted
- POM: No data found for this date range, symbol may be delisted
- DLLR: No data found for this date range, symbol may be delisted
- JAS: No data found for this date range, symbol may be delisted
- AYR: No data found, symbol may be delisted
- SGK: No data found for this date range, symbol may be delisted
- LGF: No data found for this date range, symbol may be delisted
- KPP: No data found for this date range, symbol may be delisted
- PNRA: No data found for this date range, symbol may be delisted
- RSCR: No data found for this date range, symbol may be delisted
- AQNT: No data found for this date range, symbol may be delisted
- PDE: No data found for this date range, symbol may be delisted
- ELNK: No data found for this date range, symbol may be delisted
- GYI: No data found for this date range, symbol may be delisted
- BPO: No data found for this date range, symbol may be delisted
- PRXL: No data found for this date range, symbol may be delisted
- DMND: 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
- PIXR: No data found for this date range, symbol may be delisted
- ARJ: No data found for this date range, symbol may be delisted
- GWR: No data found, symbol may be delisted
- CEC: No data found for this date range, symbol may be delisted
- NBL: No data found, symbol may be delisted
- VLTR: No data found for this date range, symbol may be delisted
- CMO: No data found, symbol may be delisted
- PTV: No data found for this date range, symbol may be delisted
- TSY: No data found for this date range, symbol may be delisted
- RCRC: No data found for this date range, symbol may be delisted
- PTRY: No data found for this date range, symbol may be delisted
- AVX: No data found, symbol may be delisted
- MAUXF: No data found for this date range, symbol may be delisted
- MWP: No data found for this date range, symbol may be delisted
- CNVR: No data found for this date range, symbol may be delisted
```

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

[20]:
```
                    A  AACB        AAIC        AAP      AAPL  AAT  AAWW  ABBV  \
Date
2004-02-09        NaN   NaN         NaN        NaN       NaN  NaN   NaN   NaN
2004-02-10  26.731045   NaN  500.600006  27.280001  0.410357  NaN   NaN   NaN
2004-02-11  26.752504   NaN  514.799988  27.753332  0.425000  NaN   NaN   NaN
2004-02-12  26.409157   NaN  515.400024  28.046667  0.423750  NaN   NaN   NaN
2004-02-13  26.523605   NaN  512.000000  27.933332  0.410714  NaN   NaN   NaN

                ABC  ABCD  …  XTO  XYL      YELL  YLWDF        YUM  \
Date                       …
2004-02-09      NaN   NaN  …  NaN  NaN       NaN    NaN        NaN
2004-02-10  14.1675   NaN  …  NaN  NaN  239025.0    NaN  11.926671
2004-02-11  14.2500   NaN  …  NaN  NaN  244200.0    NaN  12.117182
2004-02-12  14.1450   NaN  …  NaN  NaN  247500.0    NaN  12.677930
2004-02-13  14.1525   NaN  …  NaN  NaN  245775.0    NaN  12.692308

                 ZBRA         ZD  ZLC  ZQKSQ  ZTS
Date
2004-02-09        NaN        NaN  NaN    NaN  NaN
2004-02-10  43.740002   9.769565  NaN    NaN  NaN
2004-02-11  47.119999  10.500000  NaN    NaN  NaN
2004-02-12  46.866669  10.326087  NaN    NaN  NaN
2004-02-13  46.453335  10.186957  NaN    NaN  NaN

[5 rows x 1804 columns]
```

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

```
Shape of dataset from Yahoo: (4552, 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: 413
```

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       2004-02-10        A    26.731045
1       2004-02-11        A    26.752504
2       2004-02-12        A    26.409157
3       2004-02-13        A    26.523605
4       2004-02-17        A    26.816881
…              …       …          …
6305398 2022-02-03      ZTS  200.919998
6305399 2022-02-04      ZTS  199.539993
6305400 2022-02-07      ZTS  200.320007
6305401 2022-02-08      ZTS  201.300003
6305402 2022-02-09      ZTS  202.289993

[6305403 rows x 3 columns]
```

[27]:

```
# Return the number of missing values
data2.isnull().sum()
```

[27]:

```
Date             0
variable         0
value       849112
dtype: int64
```

[28]:

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

```
Number of data without missing: 5456291
```

[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: 4533
```

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

```
Shape of dataset from Yahoo without empty index: (5456291, 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
```

```
/usr/local/lib/python3.7/dist-packages/pandas/core/indexing.py:1732:
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)
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:16: RuntimeWarning:
divide by zero encountered in double_scalars
  app.launch_new_instance()
```

[33]:

```
              Date symbol       value  return_rate variable
0       2004-02-10    NaN   26.731045          NaN        A
1       2004-02-11    NaN   26.752504     0.000803        A
2       2004-02-12    NaN   26.409157    -0.012834        A
3       2004-02-13    NaN   26.523605     0.004334        A
4       2004-02-17    NaN   26.816881     0.011057        A
…              …    …           …          …        …
6305398 2022-02-03    NaN  200.919998    -0.006183      ZTS
6305399 2022-02-04    NaN  199.539993    -0.006868      ZTS
6305400 2022-02-07    NaN  200.320007     0.003909      ZTS
6305401 2022-02-08    NaN  201.300003     0.004892      ZTS
6305402 2022-02-09    NaN  202.289993     0.004918      ZTS

[5456291 rows x 5 columns]
```

[34]:

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

[34]:

```
              Date symbol      value  return_rate variable
5321742 2004-02-10    NaN  13.175000          NaN       SU
5321743 2004-02-11    NaN  13.285000     0.008349       SU
5321744 2004-02-12    NaN  12.960000    -0.024464       SU
5321745 2004-02-13    NaN  12.830000    -0.010031       SU
5321746 2004-02-17    NaN  12.985000     0.012081       SU
…              …    …          …          …        …
5326270 2022-02-03    NaN  29.219999    -0.038816       SU
5326271 2022-02-04    NaN  28.719999    -0.017112       SU
5326272 2022-02-07    NaN  28.990000     0.009401       SU
5326273 2022-02-08    NaN  28.469999    -0.017937       SU
5326274 2022-02-09    NaN  28.889999     0.014752       SU

[4533 rows x 5 columns]
```

[35]:

```python
#Pivot table
df_width = df_rr.pivot(index='Date', columns='variable', values='return_rate')
```

[36]:

```python
#Converting Date column
df_width = df_width.reset_index()
df_width['Date'] =  pd.to_datetime(df_width['Date'])
```

[37]:

```python
df_width
```

```
[37]: variable       Date         A      AAIC       AAP      AAPL       AAT  \
      0        2004-02-10       NaN       NaN       NaN       NaN       NaN
      1        2004-02-11  0.000803  0.028366  0.017351  0.035684       NaN
      2        2004-02-12 -0.012834  0.001166  0.010569 -0.002941       NaN
      3        2004-02-13  0.004334 -0.006597 -0.004041 -0.030763       NaN
      4        2004-02-17  0.011057  0.027734  0.014320  0.006956       NaN
      …             …       …       …       …       …       …
      4528     2022-02-03 -0.016986 -0.002915 -0.017703 -0.016720 -0.021164
      4529     2022-02-04 -0.004725  0.000000 -0.023657 -0.002950 -0.000285
      4530     2022-02-07 -0.005315  0.014620 -0.004626 -0.004235  0.001138
      4531     2022-02-08  0.003135 -0.008646  0.018855  0.018467 -0.002274
      4532     2022-02-09  0.025779 -0.005814  0.008601  0.008294  0.014530

      variable      AAWW      ABBV       ABC  ABCD  …       XPO      XRAY  \
      0              NaN       NaN       NaN   NaN  …       NaN       NaN
      1              NaN       NaN  0.005823   NaN  …  0.000000  0.014908
      2              NaN       NaN -0.007368   NaN  … -0.071970 -0.007458
      3              NaN       NaN  0.000530   NaN  … -0.004082 -0.007741
      4              NaN       NaN  0.006889   NaN  … -0.004098  0.011703
      …            …       …       …   …  …       …       …
      4528     -0.011508  0.015221  0.013228   NaN  … -0.021798 -0.018840
      4529     -0.038933 -0.000568 -0.009211   NaN  … -0.028693 -0.005703
      4530      0.020321  0.013367  0.020496   NaN  … -0.000943 -0.003442
      4531      0.002809  0.006876  0.020300   NaN  …  0.011953  0.016884
      4532      0.023425 -0.002160  0.004781   NaN  …  0.082997  0.025849

      variable       XRX       XYL      YELL  YLWDF       YUM      ZBRA        ZD  \
      0              NaN       NaN       NaN    NaN       NaN       NaN       NaN
      1         0.002000       NaN  0.021650    NaN  0.015974  0.077275  0.074766
      2         0.015303       NaN  0.013514    NaN  0.046277 -0.005376 -0.016563
      3        -0.013761       NaN -0.006970    NaN  0.001134 -0.008819 -0.013474
      4         0.016611       NaN -0.005188    NaN  0.006514  0.000143  0.007256
      …            …       …       …    …       …       …       …
      4528     -0.006790 -0.120273 -0.079717    0.0  0.001278 -0.016847 -0.048428
      4529      0.004558 -0.012439 -0.054861    0.0 -0.000638 -0.010250  0.004990
      4530     -0.015427 -0.011601  0.019348    0.0 -0.001517 -0.009442  0.006951
      4531     -0.014286  0.021350  0.064935    0.0 -0.000879 -0.013324  0.012129
      4532      0.004208  0.041589  0.030019    0.0  0.021845  0.029123  0.026500

      variable       ZTS
      0              NaN
      1              NaN
      2              NaN
      3              NaN
      4              NaN
      …            …
      4528     -0.006183
```

```
4529    -0.006868
4530     0.003909
4531     0.004892
4532     0.004918

[4533 rows x 1392 columns]
```

[38]:

```
#Creating missing date
row_date = [pd.to_datetime('2018-12-05')]
row_nan = np.repeat(np.nan, len(df_width.columns)-1).tolist()
row_new = row_date + row_nan

#Adding missing row
df_width.loc[-1] = row_new
df_width = df_width.fillna(0)

#sorting by date
df_width = df_width.sort_values(by="Date")
```

[39]:

```
#Setting index
df_width.index = range(len(df_width))
```

[40]:

```
#Creating dataframe with dates
df_rates = pd.DataFrame()

var2 = 1
data_list = []
for i in range(len(df_width)):

    if df_width["Date"].iloc[i].dayofweek != 2:
        continue
    elif df_width["Date"].iloc[i].dayofweek == 2:
        var2 = var2+1
        if var2==2:
            var2=0
            data_list.append(df_width["Date"].iloc[i])

df_rates["Date"]=pd.Series(data_list)
```

[41]:

```
#Sum of rates for 2 weeks
for c in range(1, len(df_width.columns)):

    var = 0
    var_list = []
    table_var = df_width[df_width.columns[c]]
    var2 = 1
    for i in range(len(table_var)):
        if df_width["Date"].iloc[i].dayofweek != 2:
            var = var + table_var.iloc[i]
        elif df_width["Date"].iloc[i].dayofweek == 2:
            var2 = var2+1
            if var2==2:
                var2=0
                var = var + table_var.iloc[i]
                var_list.append(var)
                var = 0
        df_rates[df_width.columns[c]]=pd.Series(var_list)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:19:
DeprecationWarning: The default dtype for empty Series will be 'object' instead
of 'float64' in a future version. Specify a dtype explicitly to silence this
warning.
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:19:
PerformanceWarning: DataFrame is highly fragmented.  This is usually the result
of calling `frame.insert` many times, which has poor performance.  Consider
joining all columns at once using pd.concat(axis=1) instead.  To get a de-
fragmented frame, use `newframe = frame.copy()`
```

[41]:

```
```

[43]:

```
#unpivoting tables
df_long = pd.melt(df_rates, id_vars='Date', value_vars=df_rates.columns.
 ↪to_list())

#Merging tables
df_marged = pd.merge(df, df_long,  how='left', left_on=['date','symbol'],␣
 ↪right_on = ['Date','variable'])
```

[44]:

```
df_marged
```

[44]:

```
      symbol                   sector     score       Date variable     value
0         SU         Energy Minerals  0.953727 2004-02-11       SU  0.008349
1        GGG  Producer Manufacturing  0.952753 2004-02-11      GGG  0.011734
2        WGR         Energy Minerals  0.947634        NaT      NaN       NaN
3        CWT               Utilities  0.934181 2004-02-11      CWT  0.004778
4        BLL      Process Industries  0.922862 2004-02-11      BLL -0.004917
…      …                      …       …        …      …       …
37355    PEP   Consumer Non-Durables  0.701507 2022-02-09      PEP -0.003515
37356   SSNC     Technology Services  0.701123 2022-02-09     SSNC  0.058040
37357    GEF      Process Industries  0.697954 2022-02-09      GEF -0.023572
37358    DPZ       Consumer Services  0.697741 2022-02-09      DPZ  0.063856
37359  LIFZF     Non-Energy Minerals  0.695644 2022-02-09    LIFZF  0.132637
```

```
[37360 rows x 6 columns]
```

[45]:

```
df_yahoo_2w = pd.DataFrame()
df_yahoo_agr = df_yahoo
df_yahoo_agr=df_yahoo_agr.reset_index()
df_yahoo_agr["Date"] = pd.to_datetime(df_yahoo_agr["Date"])

#Creating missing date
row_date = [pd.to_datetime('2018-12-05')]
row_nan = np.repeat(np.nan, len(df_yahoo_agr.columns.values.tolist())-1).
 ↪tolist()
row_new = row_date + row_nan

#Adding missing row
#df_yahoo_agr.loc[-1] = row_new
df_yahoo_agr = df_yahoo_agr.fillna(0)

#sorting by date
df_yahoo_agr = df_yahoo_agr.sort_values(by="Date")

#Setting index
df_yahoo_agr.index = range(len(df_yahoo_agr))

#Creating dataframe with dates
df_yahoo_2w = pd.DataFrame()

var2 = 1
data_list = []
for i in range(len(df_yahoo_agr)):

    if df_yahoo_agr["Date"].iloc[i].dayofweek != 2:
        continue
    elif df_yahoo_agr["Date"].iloc[i].dayofweek == 2:
        var2 = var2+1
        if var2==2:
            var2=0
            data_list.append(df_yahoo_agr["Date"].iloc[i])

df_yahoo_2w["Date"]=pd.Series(data_list)

#Sum of rates for 2 weeks
for c in range(1, len(df_yahoo_agr.columns)):

    var = []
    var_list = []
    table_var = df_yahoo_agr[df_yahoo_agr.columns[c]]
    var2 = 1
    for i in range(len(table_var)):
        if df_yahoo_agr["Date"].iloc[i].dayofweek != 2:
            var.append(table_var.iloc[i])
        elif df_yahoo_agr["Date"].iloc[i].dayofweek == 2:
            var2 = var2+1
            if var2==2:
                var2=0
                var = var + table_var.iloc[i]
                var_list.append(var.mean())
                var = []
        df_yahoo_2w[df_yahoo_agr.columns[c]]=pd.Series(var_list)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:56:
DeprecationWarning: The default dtype for empty Series will be 'object' instead
of 'float64' in a future version. Specify a dtype explicitly to silence this
warning.
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:56:
PerformanceWarning: DataFrame is highly fragmented.  This is usually the result
of calling `frame.insert` many times, which has poor performance.  Consider
joining all columns at once using pd.concat(axis=1) instead.  To get a de-
fragmented frame, use `newframe = frame.copy()`
```

[46]:
```
df_yahoo_2w
```

[46]:
```
          Date           A  AACB         AAIC         AAP        AAPL  \
0   2004-02-11   40.118027   0.0   765.099991   41.393332    0.630179
1   2004-02-25   49.841611   0.0  1065.771441   53.142858    0.813724
2   2004-03-10   46.909871   0.0  1085.525032   54.147501    0.940246
3   2004-03-24   42.621603   0.0  1032.650002   52.200000    0.923348
4   2004-04-07   45.432761   0.0  1016.624992   54.086667    0.979866
..         …         …   …          …         …         …
462 2021-12-15  307.388748   0.0     7.141250  472.234997  350.072502
463 2021-12-29  315.267138   0.0     7.007143  474.844288  353.957147
464 2022-01-12  301.293743   0.0     7.150000  474.668749  351.644999
465 2022-01-26  273.509992   0.0     6.871429  459.019996  325.887146
466 2022-02-09  283.638750   0.0     6.825000  461.587494  347.619999

           AAT        AAWW        ABBV         ABC  …  XTO         XYL  \
0     0.000000    0.000000    0.000000   21.333750  …  0.0    0.000000
1     0.000000    0.000000    0.000000   28.454643  …  0.0    0.000000
2     0.000000    0.000000    0.000000   28.263438  …  0.0    0.000000
3     0.000000    0.000000    0.000000   26.795000  …  0.0    0.000000
4     0.000000    0.000000    0.000000   27.212188  …  0.0    0.000000
..         …         …         …         …  …  …         …
462  71.139999  174.806252  252.979997  243.972497  …  0.0  240.997499
463  72.711427  183.924287  267.361428  261.421425  …  0.0  236.472858
```

```
464   76.095001  177.021247  272.962492  268.872505  …  0.0  233.215000
465   72.191426  157.648568  268.005709  263.680004  …  0.0  211.255714
466   70.941251  158.871249  282.599995  280.606251  …  0.0  192.053748

               YELL      YLWDF         YUM         ZBRA          ZD  ZLC  \
0    363712.500000   0.000000   18.080517    68.990000   15.384782  0.0
1    474503.571429   0.000000   25.711719    89.077144   19.506832  0.0
2    458034.375000   0.000000   26.861072    94.012500   19.051087  0.0
3    487659.375000   0.000000   26.611252    90.915000   16.925544  0.0
4    515259.375000   0.000000   27.715224    95.436668   19.399457  0.0
..             …          …           …            …          …   …
462      25.945000  22.000000  262.751247  1195.444992  217.409999  0.0
463      25.092857  22.000000  273.201429  1189.350002  220.274287  0.0
464      22.602500  22.056875  265.078745  1103.969994  219.658754  0.0
465      20.824286  22.070000  245.639998   980.082864  210.169996  0.0
466      21.153749  22.070000  251.634996  1004.248741  207.602501  0.0

     ZQKSQ         ZTS
0      0.0    0.000000
1      0.0    0.000000
2      0.0    0.000000
3      0.0    0.000000
4      0.0    0.000000
..     …          …
462    0.0  459.233753
463    0.0  485.928567
464    0.0  438.091248
465    0.0  391.767140
466    0.0  400.252493

[467 rows x 1805 columns]
```

[47]:

```
#unpivoting tables
df_yahoo_long = pd.melt(df_yahoo_2w, id_vars='Date', value_vars=df_yahoo_2w.
 ↪columns.to_list())
df_yahoo_long
```

[47]:

```
             Date variable       value
0      2004-02-11        A   40.118027
1      2004-02-25        A   49.841611
2      2004-03-10        A   46.909871
3      2004-03-24        A   42.621603
4      2004-04-07        A   45.432761
…             …      …          …
842463 2021-12-15      ZTS  459.233753
842464 2021-12-29      ZTS  485.928567
842465 2022-01-12      ZTS  438.091248
```

```
842466 2022-01-26      ZTS  391.767140
842467 2022-02-09      ZTS  400.252493

[842468 rows x 3 columns]
```

[48]:

```
#Merging tables
df_marged2 = pd.merge(df_marged, df_yahoo_long,  how='left',␣
 ↪left_on=['symbol','Date'], right_on = ['variable','Date'])

df_marged2
```

[48]:

```
      symbol                  sector     score       Date variable_x  \
0         SU         Energy Minerals  0.953727 2004-02-11         SU
1        GGG  Producer Manufacturing  0.952753 2004-02-11        GGG
2        WGR         Energy Minerals  0.947634        NaT        NaN
3        CWT               Utilities  0.934181 2004-02-11        CWT
4        BLL      Process Industries  0.922862 2004-02-11        BLL
…        …                       …        …          …          …
37355    PEP    Consumer Non-Durables  0.701507 2022-02-09        PEP
37356   SSNC     Technology Services  0.701123 2022-02-09       SSNC
37357    GEF      Process Industries  0.697954 2022-02-09        GEF
37358    DPZ       Consumer Services  0.697741 2022-02-09        DPZ
37359  LIFZF     Non-Energy Minerals  0.695644 2022-02-09      LIFZF

        value_x variable_y     value_y
0      0.008349         SU   19.872500
1      0.011734        GGG   14.028889
2           NaN        NaN         NaN
3      0.004778        CWT   22.045000
4     -0.004917        BLL   12.162500
…           …         …          …
37355 -0.003515        PEP  344.388752
37356  0.058040       SSNC  161.817498
37357 -0.023572        GEF  115.113750
37358  0.063856        DPZ  887.071259
37359  0.132637      LIFZF   66.151249

[37360 rows x 8 columns]
```

1.0.7 Preparation of the target dataset

[49]:

```
#Dropping columnd
df_marged2.drop(["variable_y", "variable_x"], inplace=True, axis=1)

#Replacing 0 with NaN values
df_marged2["value_y"].replace(0,np.nan, inplace=True)
```

```
#Dropping missing values
df_marged2=df_marged2.dropna()

#Sorting by date and symbol
df_marged2 = df_marged2.sort_values(by=["Date","symbol"], ascending=True)

#Changing column names
df_marged2.rename(columns = {'value_y':'close','value_x':'return_rate'},␣
 ↪inplace = True)
```

```
[50]: df_marged2
```

```
[50]:       symbol                   sector     score       Date  return_rate  \
      65       AEE                Utilities  0.670127 2004-02-11     0.002350
      40       AOS  Producer Manufacturing  0.753176 2004-02-11     0.007533
      5        APA          Energy Minerals  0.912117 2004-02-11     0.005808
      66      ARLP          Energy Minerals  0.669621 2004-02-11    -0.011510
      63       ATO                Utilities  0.672410 2004-02-11     0.000765
      …        …                      …       …          …          …
      37315    WGO        Consumer Durables  0.778997 2022-02-09     0.078665
      37309     WM      Industrial Services  0.802717 2022-02-09     0.002500
      37280    WSO  Producer Manufacturing  0.948063 2022-02-09    -0.002957
      37352    WSO  Producer Manufacturing  0.710300 2022-02-09    -0.002957
      37350  YLWDF      Technology Services  0.718153 2022-02-09     0.000000

                  close
      65      70.309999
      40       8.005000
      5       59.630001
      66      13.578750
      63      39.230000
      …            …
      37315  133.997500
      37309  293.646246
      37280  549.996265
      37352  549.996265
      37350   22.070000

      [30173 rows x 6 columns]
```

1.0.8 Download CSV

```
[51]: def create_download_link( df, title = "Download CSV file", filename =␣
       ↪"data_rates.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(df_marged2)
```

```
[51]: <IPython.core.display.HTML object>
```