We get some stock data programmatically from Yahoo Finance. To start with, go to Google Colab and open a new notebook to be used for this exercise. o start with, install the yfinance library in your notebook:

!pip3 install yfinance

It is very useful to look at the PyPi website page of the library (<https://pypi.org/project/yfinance/>), which can tell you that the project is quite popular mid-July 2020: it has 165k installs per month.

Now, Get Historical S&P 500 Index's Data from Yahoo Finance using yfinance API

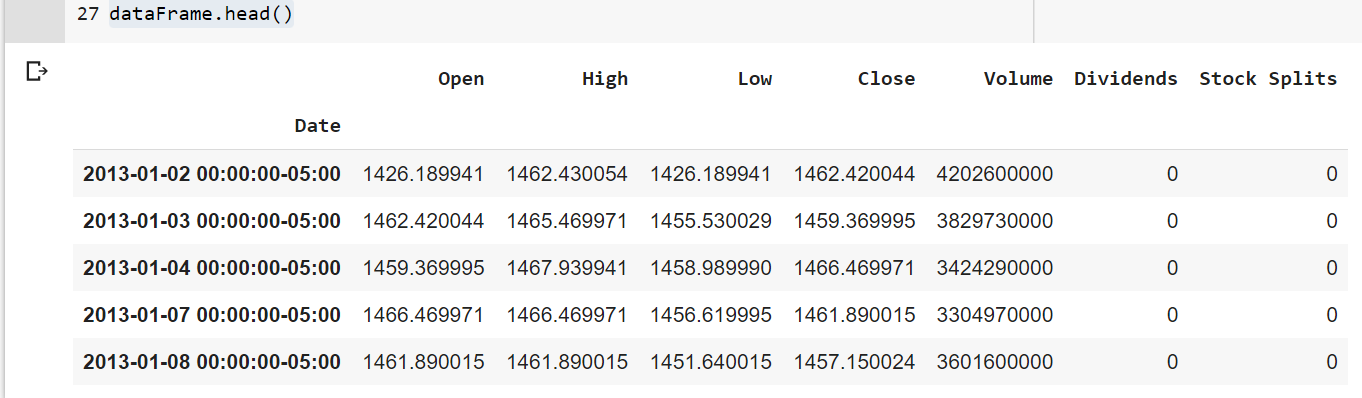
import yfinance as yahooFinance

Then Store Retrieved data from yfinance into a dataFrame

dataFrame = pd.DataFrame(historicalData.history(start=startDate,end=endDate))

dataFrame.head()

We can retrieve historical market prices too and display them:



We have the flexibility to get historical market data for the provided start and end dates too:

startDate = datetime.datetime(2013, 1, 1)

endDate = datetime.datetime(2018, 12, 31)

Descriptive statistics include those that summarize the central tendency, dispersion and shape of a dataset’s distribution, excluding NaN values.

|  |  |
| --- | --- |
| The number of non-NA/null observations and mean was as follows: |  |

**Open High Low Close Volume Dividends Stock Splits**

**count** 1509. 1509 1509. 1509. 1.509000e+03 1509.0 1509.0

**mean** 2153 2162 2143 2153 3.550010e+09 0.0 0.0

Standard deviation of the observations were as follows:

**Open High Low Close Volume Dividends Stock Splits**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **std** | 368.35 | 369.38 | 366.64 | 367.65 | 6.760519e+08 | 0.0 | 0.0 |

The lower, 50 and upper percentiles. By default the lower percentile is 25 and the upper percentile is 75. The 50 percentile is the same as the median.

**Open High Low Close Volume Dividends Stock Splits**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **25%** | 1905.65 | 1918.78 | 1885.77 | 1904.01 | 3.152890e+09 | 0.0 | 0.0 |
| **50%** | 2087.37 | 2095.60 | 2078.26 | 2087.79 | 3.455270e+09 | 0.0 | 0.0 |
| **75%** | 2433.75 | 2441.39 | 2427.70 | 2433.14 | 3.825410e+09 | 0.0 | 0.0 |

The Shapiro-Wilk test is a statistical test used to check if a continuous variable follows a normal distribution. The null hypothesis (H0) states that the variable is normally distributed, and the alternative hypothesis (H1) states that the variable is NOT normally distributed. So after running this test:

If p ≤ 0.05: then the null hypothesis can be rejected (i.e. the variable is NOT normally distributed).

If p > 0.05: then the null hypothesis cannot be rejected (i.e. the variable MAY BE normally distributed).

Since we had a small sample size n= 1509, determining the distribution of the variable X was important for choosing an appropriate statistical method. So a Shapiro-Wilk test was performed and showed that the distribution of X departed significantly from normality (0.381351113319397, p-value < 0.01), p=0,0000). Based on this outcome, a non-parametric test was used, and the median with the interquartile range were used to summarize the variable above.

A large enough sample size will make the Shapiro-Wilk test detect the smallest deviation from normality, in this case the p-value will be < 0.05 even if the variable is, in fact, normally distributed. Conversely, a very small sample size will reduce the statistical power of the Shapiro-Wilk test to reject the null hypothesis, in this case the p-value will be ≥ 0.05 even if the data clearly do not come from a normal distribution.