SZ50

A dataset contains day-level price and volume information of 34 stocks from Shanghai Stock Exchange with a total of 99212 data samples.

Dataset Snapshot

NATURE OF CONTENT

Date, trading price (open, high, low, close price), trading volume, stock ticker and day of the week (0, 1, 2, 3, 4 representing Monday to Friday).

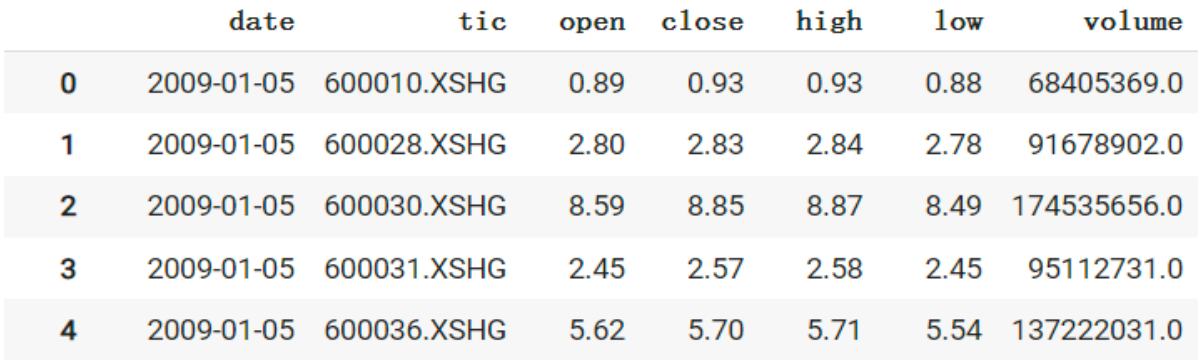
BREAKDOWN-BY INSTANCE

Total instances	99212
Training	79424
Validation	9894
Testing	9894
Total cryptocurrencies	34
Instances per stock	2918

NOTES

Stock data is collected daily from 2009-01-02 to 2021-01-01 on all trading days. The recommend split is [0.8,0.1,0.1] for training, validation and test respectively.

EXAMPLES OF ACTUAL DATA POINT



	date	tic	open	close	high	low	volume
0	2009-01-05	600010.XSHG	0.89	0.93	0.93	0.88	68405369.0
1	2009-01-05	600028.XSHG	2.80	2.83	2.84	2.78	91678902.0
2	2009-01-05	600030.XSHG	8.59	8.85	8.87	8.49	174535656.0
3	2009-01-05	600031.XSHG	2.45	2.57	2.58	2.45	95112731.0
4	2009-01-05	600036.XSHG	5.62	5.70	5.71	5.54	137222031.0

Publisher & Licenses & Access

PUBLISHER

JointQuant

DIRECT LINKS TO DATASET

GitHub link

LICENSE PERMISSIONS

You are free to share and adapt.
Attribution required.
You cannot apply any additional restrictions.

ACCESS

Open Access

ACCESS COST

Free

LICENSE TYPE(S)

CC-BY-4.0

Motivations & Use

DATASET PURPOSE

The dataset was created to provide representative data of stock trading for research in various quantitative trading tasks by selecting the most influential stocks in Shanghai Stock Exchange.

INTENDED USE CASES

- Algorithmic trading
- Portfolio Management

EXTENDED USE

- Intraday trading
- High frequency trading

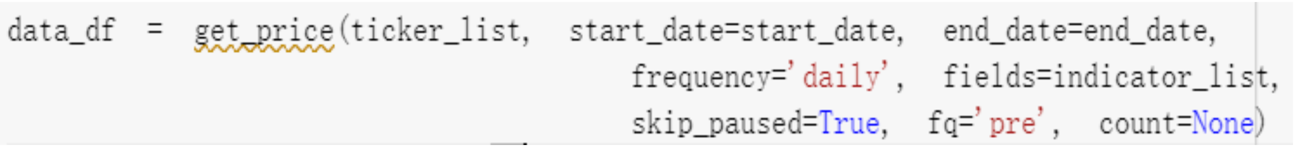
Collection

DATA SOURCE

Collected from open-source data of Shanghai Stock Exchange, retrieved from JoinQuant API

DATA COLLECTION

Retrieved from JoinQuant API (jqdatasdk) using the following code, where tic is the stock ticker of sz50 list:



```
data_df  =  get_price(ticker_list,  start_date=start_date,  end_date=end_date,
                                    frequency=’daily’,  fields=indicator_list,
                                    skip_paused=True,  fq=’pre’,  count=None)
```

Preprocessing

INDICATOR ADJUSTMENT

The money indicator is removed, and only volume is kept as a representation of trading quantity. Z-score normalization is applied on each feature to map them to a range of [-1, 1]

DATA CLEANING

Firstly, all the NaN terms are dropped, and it is observed that some of the stocks are lack of data (number of instances less than 2918). In order to maintain consistency, data corresponding to these stocks are filtered out.

FEATURE GENERATION

We generate 11 temporal features to describe the financial markets. z_{open}, z_{high}, z_{low} represent the relative values of the open, high, low prices compared with the close price at current time step, respectively. z_{close} represents the relative values of the closing prices compared with time step t-1. z_{dk} represents a long-term moving average of the adjusted close prices during the last *k* time steps compared to the current close prices. The detailed calculation formulas are as follow:

Features	**Calculation Formula**
$z_{open}, z_{high}, z_{low}$ z_{close}, z_{adj_close}	$z_{open} = open_t / close_t - 1$ $z_{close} = close_t / close_{t-1} - 1$
$z_{d\_5}, z_{d\_10}, z_{d\_15}$ $z_{d\_20}, z_{d\_25}, z_{d\_30}$	$z_{d\_5} = \frac{\sum_{i=0}^{4} adj\_close_{t-i}/5}{adj\_close_t} - 1$

Maintenance & Status

STATUS	FIRST RELEASE	CURRENT VERSION
Actively Maintained	08/2022	1.0