

## 第二讲：金融数据源及基本处理

Financial Data Source & Basic Operations



纪慧诚

金程教育资深培训讲师

CFA FRM RFP

# CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

## ▶ PART 1

金融数据概况分类

## ▶ PART 2

数据源

## ▶ PART 3

数据的导入和存储

专业来自101%的投入!



	结构化数据Structured	非结构化数据Unstructured
历史数据Historical	每日收盘价	财经新闻上的文章
实时数据Real time	股票的Bid/ask prices	推特上的发文、实时推送的新闻
原始数据Raw data	行情类：tick，合约 交易类：委托/成交，账号资金	舆情，公告，研报
处理后的数据 Well-behaved data after data handling	行情：分钟/小时/日/周/月/年 主力合约，财务，流通股，板块	关联度，热度

- 不同数据需要不同的数据分析处理方法
  - ✓ 非结构化数据：自然语言处理（NLP）
  - ✓ 时间序列分析、截面数据分析、面板数据分析

# CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

## ▶ PART 1

金融数据概况分类

## ▶ PART 2

数据源

## ▶ PART 3

数据的导入和存储

专业来自101%的投入!



## ➤ Quandl

- Quandl Python API documentation : <https://docs.quandl.com/docs/python>
- 如何安装quandl

conda install quandl 或 pip install quandl

- 如何调用quandl

```
import quandl as q  
data = q.get('EOD/KO')  
data.info()
```



## ➤ Quandl金融数据

- Quandl数据包括以下八大类Equity Prices、Equity Fundamentals、Equity Earnings、Estimates、Analyst Ratings、Futures、Options、Economics, FX and Rates。

Categories	Data Product Name	API	Free/Premium
Equity Prices	End of Day US Stock Prices	Tables	Free
Futures	CFTC Commitment of Traders Data	Time-series	Free
Options	Chicago Board Options Exchange	Time-series	Free
Economics, FX and Rates	Federal Reserve Economic Data	Time-series	Free



## ➤ Google Finance

- Package 'pandas\_datareader'中的 'web'
- 举例：
  - ✓ 从Google finance 读取苹果公司的股价并且做简单的数据处理

```
import pandas as pd
from pandas_datareader import data as web
data = web.DataReader('AAPL', data_source = 'google', end = '2017-08-01')['close']
data = pd.DataFrame(data)
data.rename(columns={'close': 'price'}, inplace = True)
```



## ➤ Yahoo ! Finance

- data\_reader()
- 2017年5月17日Yahoo ! Finance毫无预期地停止了平台上的金融数据API服务，并且宣称未来没有重启该项服务的计划。

```
from yahoo_finance import Share  
yahoo = Share('YHOO')  
print yahoo.get_open()  
print yahoo.get_price()
```

```
>>>none  
>>>none
```





## ➤ Sina

- Google国内金融数据的来源
- 可使用爬虫技术 ( urllib ) 抓取
- 哪些数据？
  - ✓ 单只股票
    - ◆ 获取某只股票的最新行情，比如宝钢股份 ( 600019 )  
<http://hq.sinajs.cn/list=sh600019>
    - ◆ 一次查询多只股的最新行情：宝钢股份 ( 600019 ) 和贵州茅台 ( 600519 )：  
<http://hq.sinajs.cn/list=sh600019,sh600519>

```
var hq_str_sh600019= "宝钢股份,6.750,6.710,6.990,7.000,6.720,6.990,7.000,  
169768963,1172077425.000,259920,6.990,364473,6.980,502606,6.970,260682,  
6.960,429000,6.950,8045451,7.000,1242671,7.010,1217692,7.020,673881,  
7.030,271372,7.040,2017-07-03,15:00:00,00";
```



## ➤ Sina

- 哪些数据？（续）

- ✓ 大盘指数

- ◆ 上证综合指数（000001）：[http://hq.sinajs.cn/list=s\\_sh000001](http://hq.sinajs.cn/list=s_sh000001)
    - ◆ 深圳成指：[http://hq.sinajs.cn/list=s\\_sz399001](http://hq.sinajs.cn/list=s_sz399001)

```
var hq_str_s_sh000001="上证指数,3195.9116,3.4847,0.11,1403057,15657578";
```

指数名称，当前点数，当前价格，涨跌率，成交量（手），成交额（万元）

- ✓ 绘图

- ◆ 对于股票的K线图等的获取可通过请求[http://image.sinajs.cn/.../.../\\*.gif](http://image.sinajs.cn/.../.../*.gif)此URL获取，其中\*代表股票代码，详见如下：
      - 分时线：<http://image.sinajs.cn/newchart/min/n/sh000001.gif>
      - 日K线查询：<http://image.sinajs.cn/newchart/daily/n/sh000001.gif>
      - 周K线查询：<http://image.sinajs.cn/newchart/weekly/n/sh000001.gif>
      - 月K线查询：<http://image.sinajs.cn/newchart/monthly/n/sh000001.gif>





## ➤ Tushare

- Tushare中文官网：<http://tushare.org/index.html>
- Tushare Python API documentation：<https://pypi.python.org/pypi/tushare/>
  - 一个免费、开源的python财经数据接口包。
  - *'A utility for crawling historical and Real-time Quotes data of China stocks.'*
- 如何安装Tushare

```
conda install tushare 或 pip install tushare
```

- 数据来源：新浪财经、腾讯财经、上交所和深交所



- Tushare有哪些数据？

数据分类	详情
交易数据	历史行情 复权数据 实时行情 历史分笔 实时分笔 当日历史分笔 大盘指数行情列表 大单交易数据
投资参考数据	分配预案 业绩预告 限售股解禁 基金持股新股数据 融资融券（沪市） 融资融券（深市）
股票分类数据	行业分类 概念分类 地域分类 中小板分类创业板分类 风险警示板分类 沪深300成份及权重 上证50成份股 中证500成份股 终止上市股票列表 暂停上市股票列表
基本面数据	股票列表 业绩报告（主表） 盈利能力 营运能力 成长能力 偿债能力 现金流量
宏观经济数据	存款利率 贷款利率 存款准备金率 货币供应量 货币供应量(年底余额) 国内生产总值(年度) 国内生产总值(季度) 三大需求对GDP贡献 三大产业对GDP拉动 三大产业贡献率 居民消费价格指数 工业品出厂价格指数
新闻事件数据	即时新闻 信息地雷 新浪股吧
龙虎榜数据	每日龙虎榜列表 个股上榜统计 营业部上榜统计 机构席位追踪 机构成交明细
银行间同业拆放利率	Shibor拆放利率 银行报价数据 Shibor均值数据 贷款基础利率（LPR） LPR均值数据



## ➤ 如何调用tushare

- 举例：
  - ✓ 获取历史行情（旧）

```
import tushare as ts  
ts.get_hist_data('600848',start='2015-01-05',end='2015-01-09')
```

- ✓ 获取历史行情（新）：**get\_k\_data**，集分钟数据、日周月数据，前后复权数据，揽括所有股票、指数和ETF的行情数据API

参数	释义	参数	释义
code	证券代码：支持沪深A、B股支持全部指数支持ETF基金	index	是否为指数：默认为False设定为True时认为code为指数代码
ktype	数据类型：默认为D日线数据D=日k线W=周M=月 5=5min 15=15min 30=30min 60=60min	start	开始日期format：YYYY-MM-DD为空时取当前日期
autype	复权类型：qfq-前复权hfq-后复权None-不复权，默认为qfq	end	结束日期：format：YYYY-MM-DD



- 举例（续）：

- ✓ **get\_k\_data应用举例**

- ◆ 获取宝钢股份近一年半的前复权日线行情：

```
ts.get_k_data('600019')
```

- ◆ 获取宝钢股份近6年后复权周线行情：

```
ts.get_k_data('600019', ktype='W', autype='hfq')
```

- ◆ 获取宝钢股份近期5分钟行情：

```
ts.get_k_data('600019', ktype='5')
```

- ◆ 获取沪深300指数10月份日线行情：

```
ts.get_k_data('399300', index=True, start='2016-10-01',  
end='2016-10-31')
```



## ➤ 优矿

- 官网：<https://uqer.io/home/>
- 十二类数据分类：热门数据、行情、沪深股票、新三板、债券、基金、港股、期货期权及衍生品、指数、宏观行业、新闻/公告/社交、研究报告



# CONTENTS

PROFESSIONAL · LEADING · VALUE-CREATING

## ▶ PART 1

金融数据概况分类

## ▶ PART 2

数据源

## ▶ PART 3

数据的导入和存储

专业来自101%的投入!



# 数据的导入和存储



金程教育  
GOLDEN FUTURE

## ➤ open()内建函数: pure python

- Python内建函数：<https://docs.python.org/2/library/functions.html>
- 举例
  - ✓ 在project文件夹下放入本地文件 'demo.csv'

```
f = open('demo.csv', 'r')  
data = f.read()  
f.close()  
print data
```



## ➤ Python标准库中的file format : csv

- Python标准库中的csv : <https://docs.python.org/2/library/csv.html>

### ✓ csv.reader()

```
import csv
csv_reader = csv.reader(open('demo.csv','r'))
data = [ | for | in csv_reader]
print data[:5]
```

### ✓ csv.DictReader()

- ◆ 读取部分数据 : nested list

```
import csv
csv_reader = csv.DictReader(open('demo.csv','r'))
data = [ | for | in csv_reader]
print data[:3]
```



## ➤ Pandas读取数据

- Pandas支持的数据读取的文件格式有：pickle table **csv** fwf magpack clipboard **excel** **json** html **hdf** feather **sas** **sql** **gbq**(google bigquery) **stata**

✓ csv

```
data = pd.read_csv('demo.csv', index_col = 0, parse_dates=True)
```

✓ excel

```
data = pd.read_excel('demo.xlsx', index_col = 0).sort_index()
```

✓ json

```
data = pd.read_json('demo.json')
```



## ➤ Pandas存储指定数据对象

- 函数：Series/DataFrame/Panel.to\_csv/excel/json/...

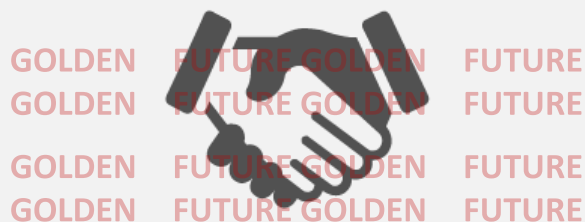
DataFrame.to\_excel ('demo.xlsx', sheet\_name = 'sheet 1')

	csv	excel	hdf	sql	clipboard	feather	stata	gbq
Series	√	√	√	√	√			
Dataframe	√	√	√	√	√	√	√	√
Panel		√	√		√			



金程教育  
GOLDEN FUTURE

# Thank you!



专业来自101%的投入!