

量化投资—Using R

前言

罗智超 Rokia.org

2017 年 6 月 20 日

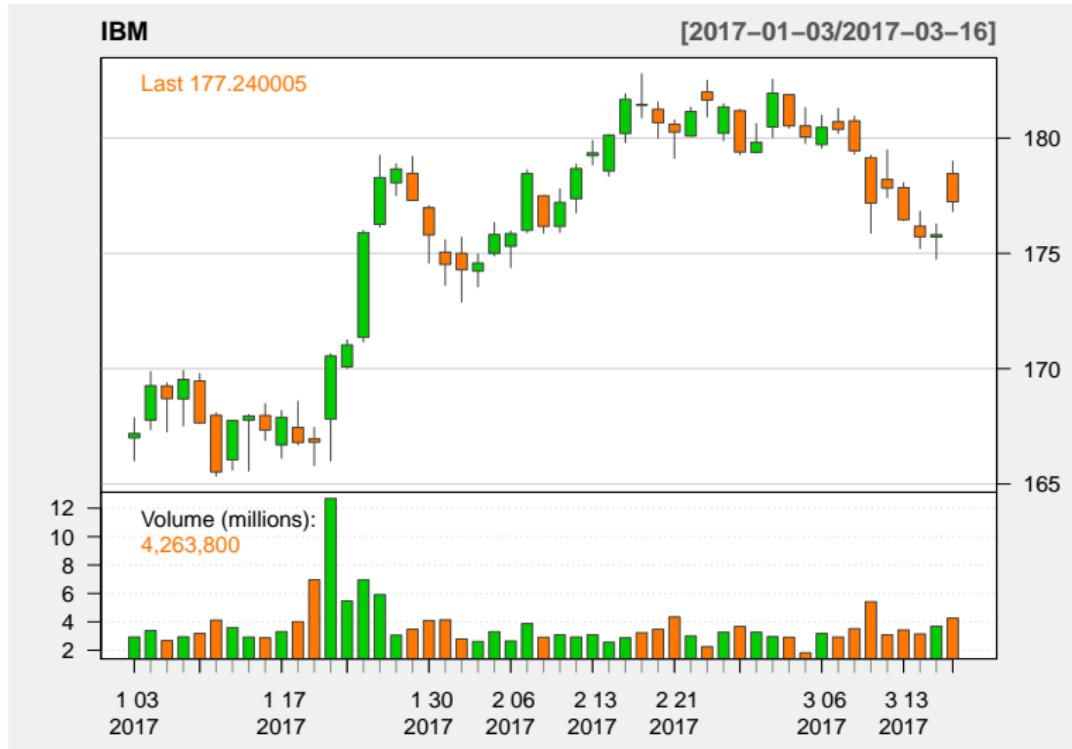
- ① 什么是 R 语言
- ② 安装 Rstudio & R
- ③ 为什么 R 适合量化投资分析
- ④ R 语言在量化投资分析中的应用框架
- ⑤ Level2 高频交易数据分析举例
- ⑥ R 语言 (数据分析) 学习路径

什么是 R 语言

数据分析领域的“小李飞刀”

- MATLAB, Eviews, Stata, Gauss, Fortran, SAS, R
- C, C#, C++, Java, Python, JavaScript, Nodejs
- R: 一个免费开源、能够自由有效地用于统计计算和绘图的语言和环境

两个小例子感受 R 语言的魅力 (绘图)



两个小例子感受 R 语言的魅力 (Cont'd)

- 上张 ppt 的图就是由下面两行代码生成的

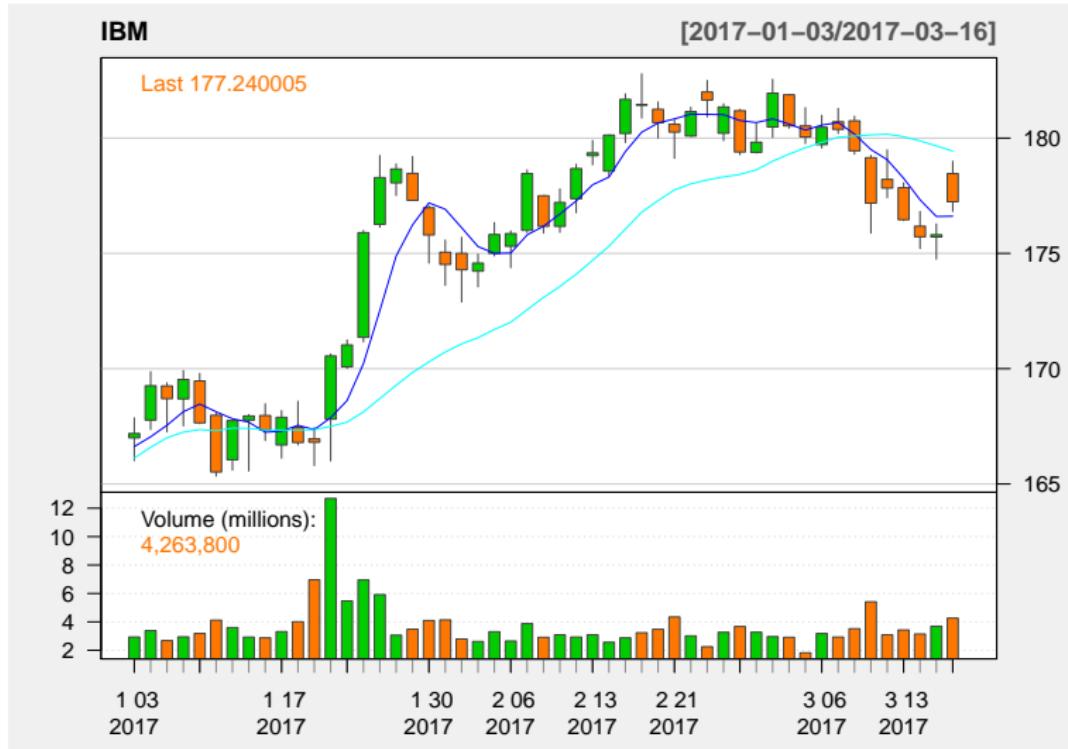
```
#getSymbols("IBM",src="yahoo")
chartSeries(IBM ,subset="2017",
            theme = "white",type = "candles")
```

两个小例子感受 R 语言的魅力 (Cont'd)

- 如果我想加两条移动平均如何？

```
addSMA(n=c(5,20))
```

两个小例子感受 R 语言的魅力 (Cont'd)



两个小例子感受 R 语言的魅力 (Rmd)

- 传统的分析报告 PPT:Copy+Paste
- 艺术化编程: 文字 + 分析代码
- 交互式探索分析 shiny::runApp("095-plot-interaction-advanced")

安装 Rstudio & R

安装 Rstudio

- www.rstudio.com

安装 R

- www.r-project.org 3.4.x

安装 Package

- tools-global options-packages- cran mirror
- packages-install
- command

```
#install.packages("quantmod")
#install.packages("quantstrat",
#repos = "http://R-Forge.R-project.org")
#library(devtools)
#install_github("hadley/ggplot2")
#https://github.com/hadley/
```

为什么 R 适合量化投资分析

R 语言有众多强大的时间处理函数

```
library(zoo)
library(xts)
library(lubridate)
library(chron)
library(timeDate)
```

R 语言的向量运算是很多编程语言无法比拟的

```
# 使用向量方法计算交易信号及累计收益率
# 计算交易信号
ds$ma1<-c(rep(NA,ma1-1),rollmean(ds$Close,ma1))
ds$ma2<-c(rep(NA,ma2-1),rollmean(ds$Close,ma2))
a<-sign(ds$ma1-ds$ma2)
ds$signal<-c(NA,diff(a))
# 计算累计收益率
ds$s<-ifelse(ds$signal== -2,-1,ifelse(ds$signal== 2,1,0))
ds$s<-cumsum(ds$s)
ds$r<-diff(ds$Close)/ds$Close
ds$rTrade<-cumprod(1+ds$s*ds$r)
```

R 语言有海量的统计、机器学习算法包

```
library(e1071)
library(rpart)
library(caret)
library(MXNetR)
library(darch)
library(deepnet)
library(h2o)
```

R 语言 +Rcpp 使 R 的应用无限可能

- 量化金融领域有大量算法是由 C++ 实现的
- Rcpp 提供了 R 语言与 C++ 的无缝连接
- CRAN 上面有超过 1 千个 PACKAGE 依赖 Rcpp
- 《Rcpp: seamless R and C++ integration》

R 语言金融数据库接口

```
#WindR
library(WindR)
w.start()
#GTA by JDBC

#Yahoo...
library(quantmod)
getSymbols("IBM",src="yahoo")
```

R 语言金融数据库接口 (Cont'd)

代码生成器

日期排列 多维序列 日内跳价 分钟序列 实时行情 数据集 经济数据 交易 资管 日期函数 TDays More Tools Help

```
w_wset_data<-w.wset('futureoir','startdate=2016-03-27;enddate=2017-03-27;varity=RB.SHF;order_by=long;ranks=all')
```

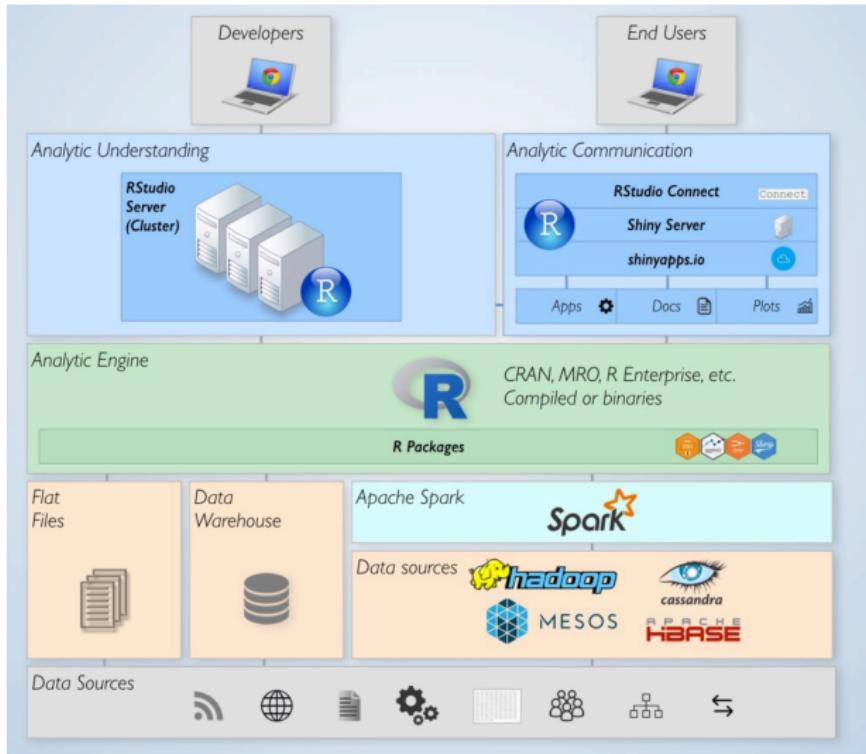
输出结果

Rank	Date	rank	member_name	long_positn	long_positn	long_positn	short_positn	short_positn	short_positn	net_long_positn	net_long_positn	net_short_positn	net_short_positn	vol	vol_increas	vol_rate	settle	
1	2017/3/27	NaF	斯百名合计	443620	-24530	25.359847	411001	-39765	23.454209	32619	11235	NaF	NaF	1.061430	1277474	22011	23.93495	3061
2	2017/3/27	NaF	前十名合计	713781	-3852	40.732811	612945	-42010	34.919358	100838	38158	NaF	NaF	5.154313	1982165	97428	37.138125	3061
3	2017/3/27	NaF	第二十名...	929594	901	52.889921	882963	-35396	50.367222	45991	36107	NaF	NaF	2.601699	2799166	107144	52.449572	3061
4	2017/3/27	1	永乐期货	124832	-12305	7.128971	108968	-25241	6.210231	15849	11538	NaF	NaF	0.30444	18394	3432	3.4397	3061
5	2017/3/27	2	中行期货	106735	-5913	6.03398	92992	-5024	5.306697	12743	-489	NaF	NaF	0.127193	280107	-26950	5.240125	3061
6	2017/3/27	3	海通期货	84743	-6222	4.855949	87169	-1105	5.006831	848	5028	1117	0.172682	504919	11074	9.460234	3061	
7	2017/3/27	4	华泰期货	65655	405	3.746672	62026	1644	3.606818	2449	-1239	NaF	NaF	0.139755	151775	-11848	2.943878	3061
8	2017/3/27	5	国泰君安	62655	805	3.515474	37343	-365	2.131018	25312	1170	NaF	NaF	1.444456	132104	-4425	2.475119	3061
9	2017/3/27	6	申银期货	71909	1429	3.304639	32825	729	1.973194	25084	150	NaF	NaF	1.431445	147508	13228	2.763731	3061
10	2017/3/27	7	银河期货	55098	-2536	3.246431	49336	-6503	2.032531	7253	3967	NaF	NaF	0.4139	157179	15383	2.944928	3061
11	2017/3/27	8	南华期货	55219	8493	3.151131	36804	3977	2.100059	18415	3116	NaF	NaF	1.050872	99664	14849	1.067319	3061
12	2017/3/27	9	方正中期	51810	-1146	2.967795	58051	-1739	3.312742	NaF	6133	-993	0.349966	118177	-9556	2.214181	3061	
13	2017/3/27	10	东航期货	48226	10430	2.752068	10955	-1666	0.625159	37211	12098	NaF	NaF	2.126909	73865	24426	1.383945	3061
14	2017/3/27	11	华信万达	28191	2717	1.659515	NaF	NaF	28191	2717	NaF	NaF	1.659515	NaF	NaF	NaF	3061	
15	2017/3/27	12	指南期货	28372	-2689	1.619078	35878	1308	2.036003	NaF	NaF	T308	3997	0.416925	26277	11946	0.46233	3061
16	2017/3/27	13	国元安信	27224	1049	1.553566	39102	-1258	2.231397	NaF	11878	-2307	0.677831	141501	14473	2.651183	3061	
17	2017/3/27	14	渤海期货	24629	3571	1.40548	NaF	NaF	24629	3571	NaF	NaF	1.40548	101038	-5173	1.4925	3061	
18	2017/3/27	15	指南期货	22229	973	1.256522	22997	1876	1.312348	NaF	768	903	0.043027	NaF	NaF	3061		
19	2017/3/27	16	光大期货	19512	-3804	1.113473	NaF	NaF	19512	-3804	NaF	NaF	1.113473	92249	-3126	1.72839	3061	
20	2017/3/27	17	中证期货	15599	1256	0.947424	39059	-1495	2.228943	NaF	NaF	22460	-2752	1.281704	91745	1621	1.719847	3061
21	2017/3/27	18	东证期货	16219	877	0.925554	NaF	NaF	16219	877	NaF	NaF	0.925554	149338	19035	2.807308	3061	
22	2017/3/27	19	国泰君安	15591	-203	0.867434	7382	-747	0.421262	8169	544	NaF	NaF	0.466173	NaF	NaF	3061	
23	2017/3/27	20	广州期货	15247	1006	0.870096	NaF	NaF	15247	1006	NaF	NaF	0.870096	59027	-34514	1.067202	3061	
24	2017/3/27	21	华泰期货	15168	-44	0.865578	NaF	NaF	15168	-44	NaF	NaF	0.865578	NaF	NaF	3061		
25	2017/3/27	22	瑞达期货	15034	-1015	0.857931	NaF	NaF	15034	-1015	NaF	NaF	0.857931	NaF	NaF	3061		
26	2017/3/27	23	浙商期货	13894	-2603	0.793278	NaF	NaF	13894	-2601	NaF	NaF	0.793278	NaF	NaF	3061		
27	2017/3/27	24	兴证期货	10795	-365	0.615457	17914	704	1.022201	NaF	NaF	T129	1069	0.406924	68108	-1398	1.276001	3061

大数据也有 R 语言的一席之地

```
library(rhadoop)
library(rhive)
library(SparkR)
library(sparklyr)
```

大数据也有 R 语言的一席之地 (Cont'd)



谁说业界不用 R

- 微软收购 Revolution, 将其产品整合至 MS SQL
- Oracle 在其 Advanced Analytics 产品中集成 R 语言
- 主流深度学习的产品都提供 R API

R 语言在量化投资分析中的应用框架

何为量化投资？

- 百度百科：量化投资是指通过数量化方式及计算机程序化发出买卖指令，以获取稳定收益为目的的交易方式。
- 个人理解：综合运用数据、模型和算法来计算、优化、验证投资逻辑有效性的概率水平。

量化投资分析的几点个人思考

- 价值投资、技术分析、专家直觉本质上都是不冲突的，都是从不同的角度论证自己的投资逻辑
- 量化投资比拼的是不对称信息处理的速度与能力
- “金融即数据”，比拼的还有数据的广度
- 开源社区带来的思考：单兵作战年代已经不再
- 家家都在制造“轮子”：基础分析数据平台的整合还有很长的路要走，但是有很多第三方平台。如：RiceQuant

R 语言量化工具包

	第三方独立R包	Rmetrics体系R包
数据管理	zoo, xts, RQuantLib, WindR RJDBC, rhadoop, SparkR, rhive, rredis, rmongodb	timeSeries, timeDate, fImport
指标计算	TTR, TSA, urca,	fArma, fAsianOptions, fBasics, fCopulae, fExoticOptions, fGarch, fNonlinear, fOptions, fRegression, fTrading, fUnitRoots
回测交易	FinancialInstrument, quantstrat, blotter	fTrading
投资组合	PortfolioAnalytics, stockPortfolio	fAssets
风险管理	PerformanceAnalytics	fPortfolio, fExtremes

Figure 3: R 语言量化工具包

R 语言量化工具包 (Cont'd)

- 数据管理：包括数据集抓取、存储、读取、时间序列、数据处理等，涉及 R 包有 zoo(时间序列对象), xts(时间序列处理), timeSeries(Rmetrics 系时间序列对象) timeDate(Rmetrics 系时间序列处理), data.table(数据处理), quantmod(数据下载和图形可视化), RQuantLib(QuantLib 数据接口), WindR(Wind 数据接口), RJDBC(数据库访问接口), rhadoop(Hadoop 访问接口), rhive(Hive 访问接口), rredis(Redis 访问接口), rmongodb(MongoDB 访问接口), SparkR(Spark 访问接口), fImport(Rmetrics 系数据访问接口) 等。

R 语言量化工具包 (Cont'd)

- 指标计算：包括金融市场的技术指标的各种计算方法，涉及 R 包有 TTR(技术指标), TSA(时间序列计算), urca(单位根检验), fArma(Rmetrics 系 ARMA 计算), fAsianOptions(Rmetrics 系亚洲期权定价), fBasics(Rmetrics 系计算工具), fCopulae(Rmetrics 系财务分析), fExoticOptions(Rmetrics 系期权计算), fGarch(Rmetrics 系 Garch 模型), fNonlinear(Rmetrics 系非线模型), fOptions(Rmetrics 系期权定价), fRegression(Rmetrics 系回归分析), fUnitRoots(Rmetrics 系单位根检验) 等。
- rmetrics.org

R 语言量化工具包 (Cont'd)

- 回测交易：包括金融数据建模，并验证用历史数据验证模型的可靠性，涉及 R 包有 FinancialInstrument(金融产品), quantstrat(策略模型和回测), blotter(账户管理), fTrading(Rmetrics 系交易分析)等。

R 语言量化工具包 (Cont'd)

- 投资组合：对多策略或多模型进行管理和优化，涉及 R 包有 PortfolioAnalytics(组合分析和优化), stockPortfolio(股票组合管理), fAssets(Rmetrics 系组合管理) 等

R 语言量化工具包 (Cont'd)

- 风险管理：对持仓进行风险指标的计算和风险提示，涉及 R 包有 PerformanceAnalytics(风险分析),fPortfolio(Rmetrics 系组合优化), fExtremes(Rmetrics 系数据处理) 等。

R 语言量化研究框架

Quantitative analysis package hierarchy

Application Area	R Package
Performance metrics and graphs	PerformanceAnalytics - Tools for performance and risk analysis
Portfolio optimization and quantitative trading strategies	PortfolioAnalytics - Portfolio analysis and optimization quantstrat – Rules-based trading system development blotter – Trading system accounting infrastructure
Data access and financial charting	quantmod - Quantitative financial modeling framework TTR - Technical trading rules
Time series objects	xts - Extensible time series zoo - Ordered observation

Level2 高频交易数据分析举例

Level2 数据库表内容

- 个股期权：个股期权静态信息、个股期权分笔数据、个股期权分时数据
- 股指期货：股指期货分笔数据、股指期货分时数据
- 国债期货：国债期货分笔交易数据、国债期货分时交易数据
- 商品期货：商品期货分笔数据、商品期货委托统计行情、商品期货实时结算行情、商品期货分价成交量行情、商品期货套利深度行情
- 沪深股票：指数行情、逐笔成交、十档行情、集合竞价、委托

Level2 数据库表内容 (Cont'd)

2.7 GTA_SZL2_TRADE (逐笔成交)

序号	字段名	中文标题	数据类型	单位	说明
1	SECURITYID	证券代码	NVARCHAR(20)		
2	TRDDATE	日期	NVARCHAR(8)		以 YYYYMMDD 表示
3	TRADETIME	成交时间	NVARCHAR(20)		
4	SETNO	证券集代号	INT		2012 年 8 月份前该字段为 NULL
5	RECNO	成交索引	INT		1、指定证券集上唯一记录标识，从 1 开始计数 2、2012 年 8 月份前该字段为 NULL
6	BUYORDERRECNO	买方委托索引	INT		1、从开始计数，0 表示无对应委托 2、2012 年 8 月份前该字段为 NULL
7	SELLORDERRECNO	卖方委托索引	INT		1、从开始计数，0 表示无对应委托 2、2012 年 8 月份前该字段为 NULL
8	TRADEPRICE	成交价格	DECIMAL(9, 3)		3 位小数
9	TRADEQTY	成交量数	INT		
10	ORDERKIND	成交类别	NVARCHAR(20)		1、2012 年 8 月份前该字段为 0 2、用法详情见下表

Level2 数据库表内容 (Cont'd)

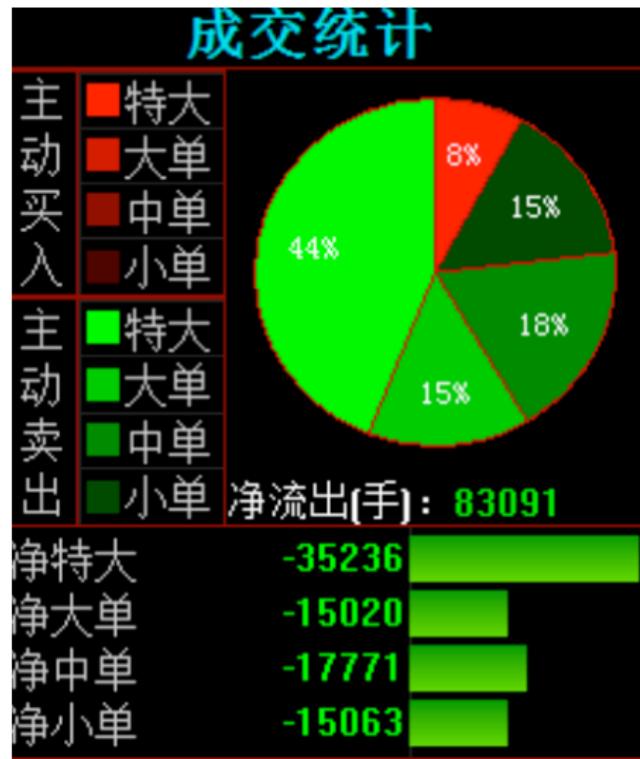
11	FUNCTIONCODE	成交代码	NVARCHAR(20)		1、2012年8月份前该字段为0 2、用法详情见下表
12	RECDATETIME	时间戳	DATETIME		2012年8月份前该字段为NULL
13	UNIX	UNIX	BIGINT		由日期和成交时间构成的时间对应的UNIX时间戳
14	MARKET	市场	NVARCHAR(4)		SZSE 表示深圳

ORDERKIND	FUNCTIONCODE	业务含义
0	0	交易业务成交记录
0	C	交易业务撤单回报记录
2	C	“即时成交剩余撤销委托”未能成交部分或其他原因的自动撤单回报记录
K	0	ETF基金申赎/赎回成功回报记录或 ETF基金赎回/赎回成功证券给付明细回报记录
K	C	ETF基金申赎/赎回撤单回报记录
V	C	“最优五档即使成交剩余撤销委托”未能成交部分的自动撤单或其他原因的自动撤单回报记录
W	C	“全额成交或撤销委托”未能成交时的自动撤单或其他原因自动撤单回报记录
X	C	本方最优价格委托的撤单回报记录
Y	C	对手方最优价格委托的撤单回报记录
Z	C	ETF基金申购/赎回成交允许/必须现金替代明细回报记录

Level2 数据库表数据容量

- 存储方式 : MSSQL
- 同步方式 : 收盘后推送数据文件, 解压后插入数据库库表
- 同步时间 : 收盘后更新, 凌晨 5 点前可以更新完全
- 数据量 : 1 年 10T, 1 天 50G 左右

沪深股票逐笔成交数据分析



沪深股票逐笔成交数据分析 (Cont'd)

- 问题来源: 单纯给我一天的“饼图”意义不大, 如同感受今天的天气要与前一天的进行比较。我需要一段时间的不同规模的资金净流入情况如何? 这几个月不同规模的买卖单资金占比分布变化如何? 另外, 不同的股票使用同一个大单划分标准是否合理?

沪深股票逐笔成交数据分析 (Cont'd)

```
qp<-seq(0.8,1,0.01)
colnames(ds)[names(ds) == side] <- "ORDERNO"
ds_sOrder <- ds %>%
  group_by(TRDDATE, ORDERNO) %>%
  summarise(
    tmSum=sum(TRDMONEY,na.rm=T)) %>%
    arrange(TRDDATE,tmSum) %>%
    group_by(TRDDATE)%>%
    mutate(tmCSum=cumsum(tmSum),
    sumP=round(tmCSum/sum(tmSum),4),
    quan=findInterval(tmSum,quantile(tmSum,qp),
    rightmost.closed = TRUE)
  )
```

沪深股票逐笔成交数据分析 (Cont'd)

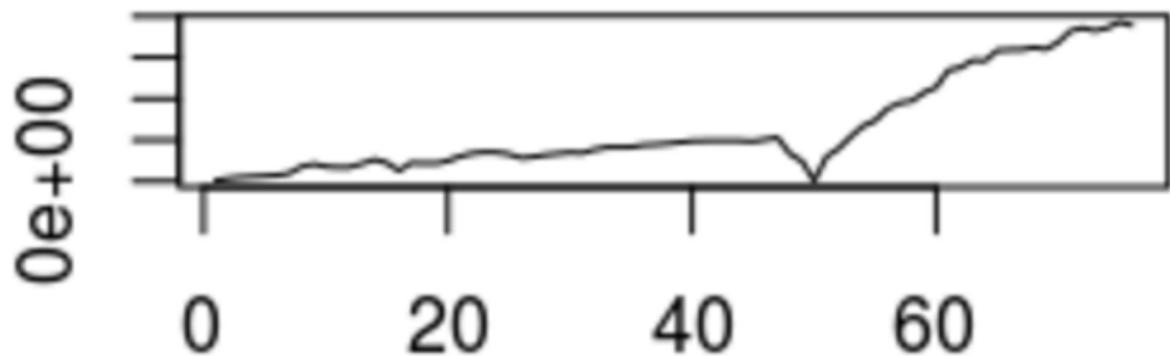


Figure 8: 5 万以下

沪深股票逐笔成交数据分析 (Cont'd)

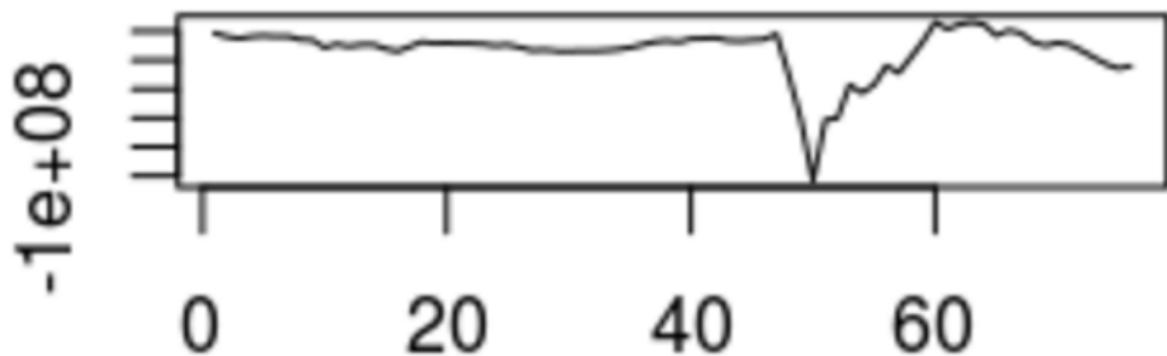


Figure 9: 5-10 万

沪深股票逐笔成交数据分析 (Cont'd)

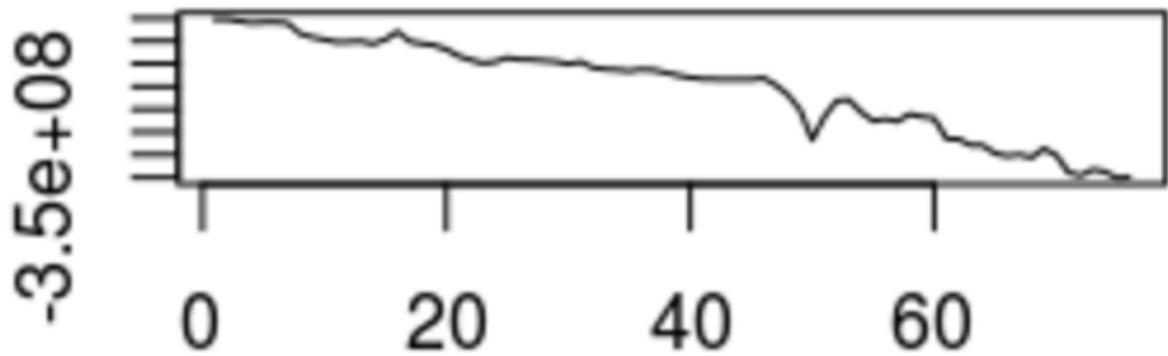
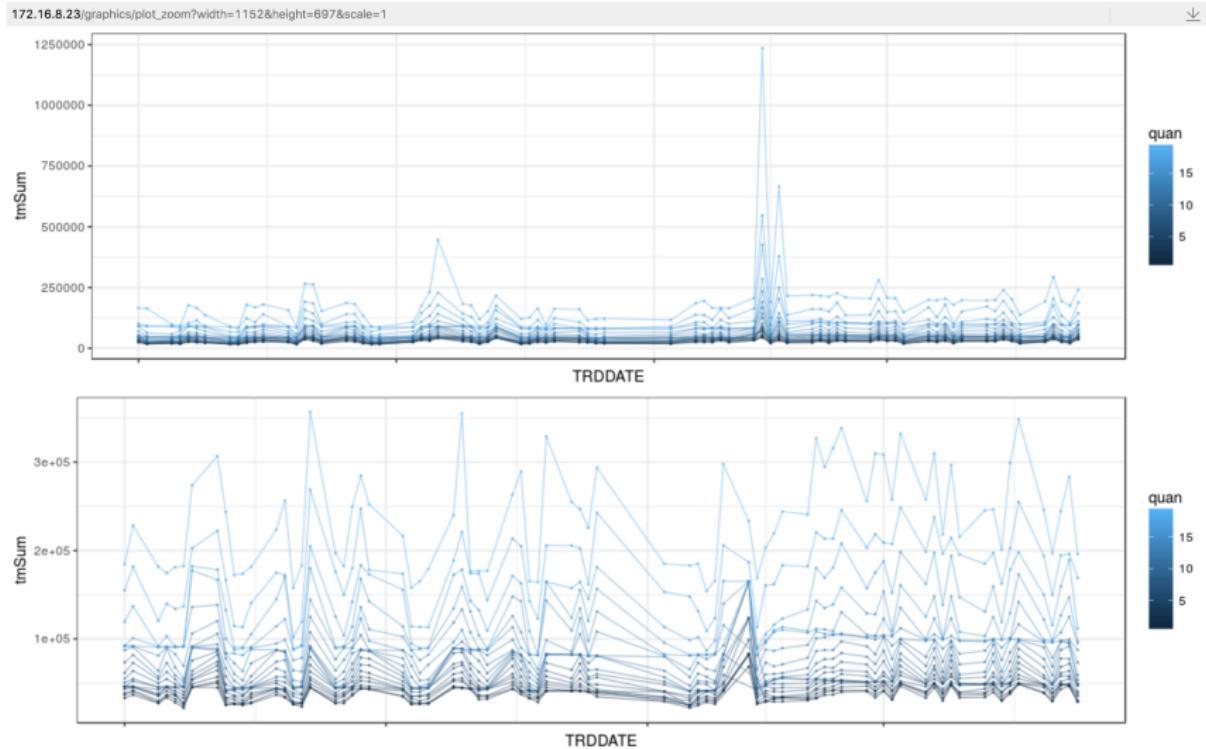


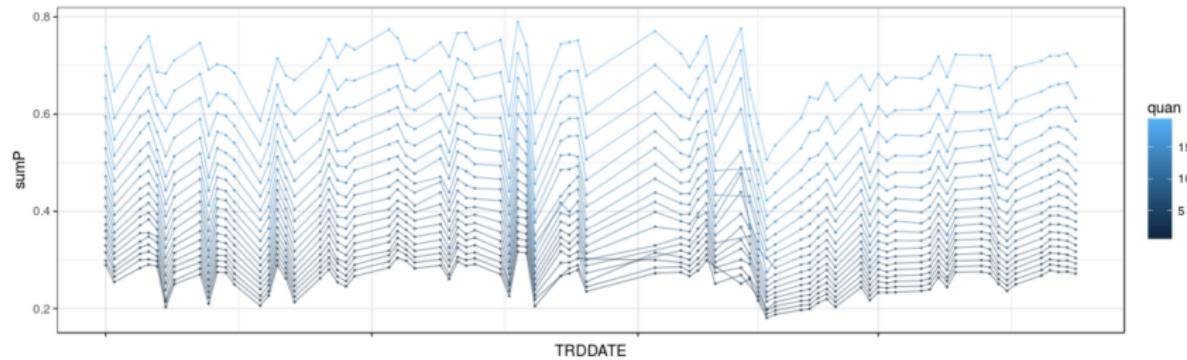
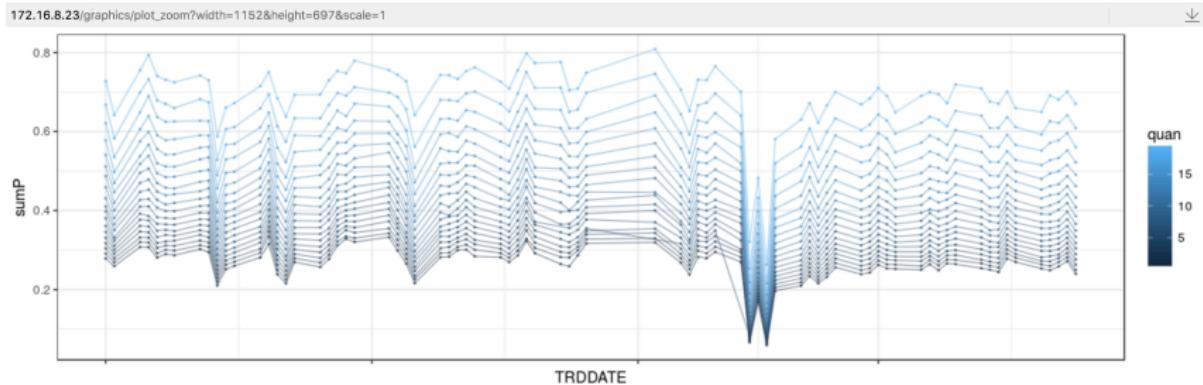
Figure 10: 10-50 万

沪深股票逐笔成交数据分析 (Cont'd)



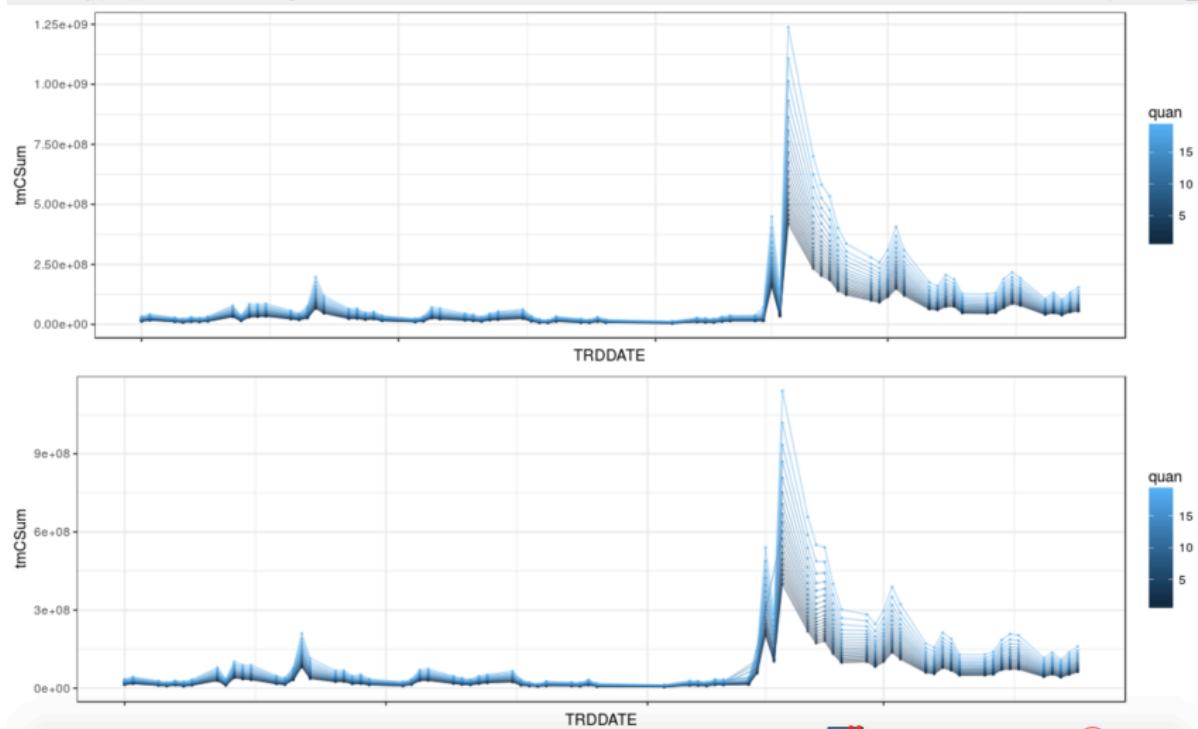
沪深股票逐笔成交数据分析 (Cont'd)

172.16.8.23/graphics/plot_zoom?width=1152&height=697&scale=1



沪深股票逐笔成交数据分析 (Cont'd)

172.16.8.23/graphics/plot_zoom?width=1152&height=697&scale=1



R 语言 (数据分析) 学习路径

R 语言 (数据分析) 学习工具篇

- 换 Mac(Option)
- 个人使用 Rstudio, 内网可以使用 Rstudio Server(Free)

R 语言 (数据分析) 学习教材篇

网盘

- 教材下载地址 (注：所有电子图书资料都收集自互联网)：
<https://pan.baidu.com/s/1kUZmrEf> 密码: ty8m

入门

- 《An Introduction to R》 by Bill Venables & David Smith

数据操纵

- 《Data Manipulation With R》 by Phil Spector

绘图

- 《ggplot2: Elegant Graphics for Data Analysis》 by Hadley Wickham

R 语言 (数据分析) 学习教材篇 (Cont'd)

高级

- 《Advanced R》 by Hadley Wickham

统计思维

- 《The Lady Tasting Tea show Statistics Revolutionized Science in the Twentieth Century》
- 《统计学：从数据到结论》 by 吴喜之

我的课件地址

- 《量化投资 Using R 》 <https://github.com/zhichaoluo/quantR>
- 《Data Analysis 课件》
<https://github.com/zhichaoluo/dataanalysis>
- 《Advanced R 讨论组课件》
<https://github.com/zhichaoluo/AdvancedR>

R 语言 (数据分析) 学习软件包篇

```
# 数据处理
library(dplyr)
library(data.table)
# 绘图
library(ggplot2)
# 时间序列
library(zoo)
library(xts)
```

数据分析 VS 烹小鲜

“哥哥， R 语言 (数据分析) 咋学呢？”



Figure 14: R 语言 (数据分析)

数据分析 VS 烹小鲜 (Cont'd)

“哥哥先带你去取数据”



数据分析 VS 烹小鲜 (Cont'd)

如果唐嫣神仙妹妹陪我获取数据，再多的数据俺也不觉得累！

数据分析 VS 烹小鲜 (Cont'd)

“妹纸， 哥哥这才是主数据”



数据分析 VS 烹小鲜 (Cont'd)

“妹纸，你说我用 R 还是 SAS 呢？”



数据分析 VS 烹小鲜 (Cont'd)

“妹纸，早知道不用 SAS 了，你早该推荐使用 R 啊”



数据分析 VS 烹小鲜 (Cont'd)

“妹纸，数据预处理，不仅是体力活，还是个精细活，学着点啊”



数据分析 VS 烹小鲜 (Cont'd)

“妹纸，我建模型的时候是不是很帅气啊！”



数据分析 VS 烹小鲜 (Cont'd)

“参数调整的时候也帅！”



数据分析 VS 烹小鲜 (Cont'd)

“建模完成撰写分析报告虽然烧脑但也难掩帅气！”



数据分析 VS 烹小鲜 (Cont'd)

“总结：只有掌握好 R 语言 (数据分析)，才能和 PPMM 分享甜蜜成果！”



”

数据分析 VS 烹小鲜 (Cont'd)

“童鞋，你想多了，你的厨房还在那里！”



结束

