Pure Mathematics 理论数学, 2023, 13(6), 1601-1609
Published Online June 2023 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2023.136163

Hans汉斯

基于R语言的数据抓取与可视化分析

刘栩同，刘　宇，徐宇航，吴林轩，孙德山

辽宁师范大学数学学院，辽宁　大连

收稿日期：2023年5月8日；录用日期：2023年6月9日；发布日期：2023年6月19日

摘　要

在大数据的背景下，对数据进行分析和处理变得越来越关键，而数据可视化技术由于可以将数据中隐藏的特征和隐藏的信息直接展现出来而备受重视。R语言的资料可视化，以图像为基础，以清楚而高效的方式传递和交流资讯，可以协助使用者快速辨识出图案，让决策人可以以视觉的方式来理解各个层面的详细资讯。通过对一系列个股的实例研究，完成了K线图的绘制和正态性的测试，从而为股市的研究奠定了一定的理论基础。

关键词

R语言，数据抓取，可视化

Data Capture and Visualization Analysis Based on R Language

Xutong Liu, Yu Liu, Yuhang Xu, Linxuan Wu, Deshan Sun

School of Mathematics, Liaoning Normal University, Dalian Liaoning

Received: May 8th, 2023; accepted: Jun. 9th, 2023; published: Jun. 19th, 2023

Abstract

Along with the arrival of the big data age, it is especially important to analyze and deal with the data. Because of its ability to show the features of the data and the underlying information, it has been widely concerned in the past years. The visual representation of R-language is based on

文章引用: 刘栩同, 刘宇, 徐宇航, 吴林轩, 孙德山. 基于 R 语言的数据抓取与可视化分析[J]. 理论数学, 2023, 13(6): 1601-1609. DOI: 10.12677/pm.2023.136163

graphic method, which makes it easier for the user to recognize the model more rapidly and make the decision-makers know the level of detail. Based on the case of a set of shares, this article has carried out the fundamental analysis of K-figure and normal test, which can be used to analyze the stock market.

Keywords

R Language, Data Capture, Visualization

1. 引言

在大数据的背景下，数据可视化在人们的日常工作和生活中发挥着重要作用。因此，如何更好地对这些数据作出合理地处理，是当前许多数据工作者所关心的问题。与常规的资料处理方式相比，可视化方式更为直接和高效。伴随着互联网技术的发展，互联网上产生了大量的互联网数据，包含了大量的用户行为、交易等信息，蕴含了极大的经济价值。当前常用的可视化技术包括 Excel、QT、窗体、PowerBI、谷歌分析等，它们是可视化的核心技术。但是，目前使用的各种可视化软件都有不同程度的缺陷，比如不能对 Excel 中的数据表进行直接的修改，必须通过其他的软件来完成；表单仅能将资料录入表单，不能进行互动分析；还有浏览 BI 存在的问题，例如无法直接处理这些数据；谷歌的这份报告只能给出几个可视化的例子。

在当今大数据环境下，R 语言提供了一种非常完善的数据分析与可视化工具。它能将文本、图像、视频等多种形式的数据以结构化的形式呈现出来，同时还能将复杂的数据进行可视化。

在金融领域中，许多应用都需要挖掘海量数据，对这些数据进行可视化分析就显得尤为重要。本文以财务数据为例，探讨了 R 语言在财务数据可视化分析中的应用。

2. R 语言简介

R 是一种具有数据处理、统计学、可视性、文法简洁、图形绘制功能强大、适用面广泛、数据库庞大、数据处理、图形绘制便捷等特点的统计学编程语言；拥有最好的视觉资料才是关键。其不足之处在于，其表现比较糟糕，并且难以在大段文字中工作。R 是一种很好的阅读资料的工具，可以阅读 EXCEL、CSS、SEC、WEB、Web、Web 等多种资料，也可以阅读网站、网站、网站等资料，并且可以使用 R 来进行资料的处理[1]。

R 是一个集数据处理，计算，绘图于一体的综合软件系统[2]。该系统的主要功能有：数据存储与处理系统；阵列操作工具(在矢量和矩阵操作中具有特别强大的功能)；完善的统计分析工具，出色的统计绘图功能；程序语言简单，功能强大，可以控制数据的输入输出，可以实现分支，循环，用户可以定制功能。

R 是一种统计程序语言，它能进行数据分析，统计分析，可视化，语法简单，有很强的绘图能力，应用范围很广，数据库很丰富，数据分析和绘图都很方便；最重要的是要有最佳的可视化数据。它的缺点是，它的性能相对较差，而且很难处理大的文本。R 具有很强的读数据功能，它能读到各种数据，如 EXCEL，CSS，SEC，WEB 等，也能读到网页等，还能用 R 对数据进行整理。

3. 基于 R 语言的数据抓取算法的实现

R 为用户提供了一系列访问数据文件和数据库的方式[3]，包括文件类型、数据流类型和数据库类型。在对这些信息进行处理前，首先要对这些信息进行提取，并将这些信息提取到存储空间中。

首先获取 Tushare 网页文件。打开 Tushare 网页，我们可以看到它的界面非常友好，用户界面，数据提取器，分析报告，股票信息都在这里显示。再获取股票信息。使用爬虫抓取 Tushare 网页的数据，把数据保存到本地文件。接下来对其进行分析。

4. 基于 R 语言的数据可视化分析

数据的可视化分析是对海量数据进行分析挖掘的重要手段[4]，而 R 语言在数据挖掘中的应用广泛，具有良好的可视化效果。分析结果如下：

对提取的数据进行分析，主要代码如代码一。

代码一 爬取股票数据代码

```
library('Tushare')
install.packages("Tushare")
library(xlsx)
bar<-Tushare::pro_bar(token='***')
api<-Tushare::pro_api(token='***')
stockname=api(api_name='stock_basic')
head(stockname)
data1=api(api_name='daily',ts_code="000001.SZ",start_date="20191001",end_date="20230331")
sh=api(api_name='index_basic',ts_code='000001.sh',start_date='20200101',
end_date='20230331')
```

接着将得到的 data1、sh、stockname 为变量的数据保存，保存代码如代码二。

代码二 数据保存代码

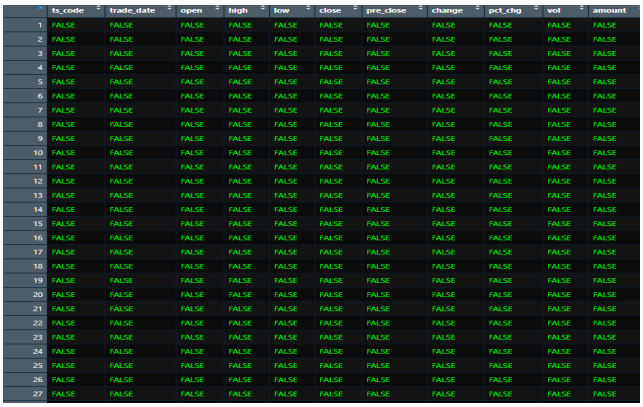
```
write.xlsx(data1,"../R/data.xlsx")
write.xlsx(sh,"../R/sh.xlsx")
write.xlsx(stockname,"../R/stockname.xlsx")
```

在对 data1 中的数据进行预处理，找出缺失值和空白值，调用 is.na()函数，is.na()作用于对象后，若相应的数值为缺失值则返回 TURE，若为 FALSE，通过求和函数 sum()可计算出缺失值总数。这段编码类似于 3 号编码。

代码三 找到缺失值代码

```
#数据预处理
a=is.na(data1)#缺失值则返回 TURE，否则返回 FALSE。
a_sum<-sum(a)#计算缺失值的总和
```

代码运行效果图，如图 1 所示，并得到 a_sum 的值为 0。



	ts_code	trade_date	open	high	low	close	pre_close	change	pct_chg	vol	amount
1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
7	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
8	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
10	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
11	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
12	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
13	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
14	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
15	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
16	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
17	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
18	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
19	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
20	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
21	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
22	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
23	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
24	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
25	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
26	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
27	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Figure 1. The result display of a
图 1. a 的结果展示图

调用 chartSeries 函数绘制 K 线图，绘制代码如代码四。

代码四 绘制 K 线图代码

```
#股票数据的 K 线图
library(TTR)
library(zoo)
library(xts)
library(quantmod)
df<-read.xlsx("2.xlsx",1 )
myvars <- c("open","high","low","close","volume")
data <- xts(df[myvars], order.by=as.Date(as.character(df[,1]),format="%Y-%m-%d"))
head(data)#取前 6 项
#没有增加布林线的 K 线图
stock <- as.xts(data, descr=df)
chartSeries(x=stock["2019-10-08/"], name='K 线图', line.type="l", bar.type="ohcl",theme="white", up.col='red', dn.col='green')
```

代码结果如图 2 所示

Figure 2. The graph without Bollinger bands
图 2. 不加布林线的图

图 2 的 K 线图表示股票日收盘价和成交量变化趋势，从图中可以看出收盘价最高出现在 2020 年年底和 2021 年 4 月份左右，收盘价最低出现在 2022 年年底。对股票数据进行典型图形绘制，主要增加了布林线(BBands)指标和趋向指标(ADX)等及技术分析指标，为什么要增加布林线指标和平均趋向指标呢?首先，根据其原理，一般来说股价一般是围绕如均线、成本线等价值中枢在一定的范围内波动，布林线指标在这一基础上认为股价信道的宽窄会随着股价的变化而变化，自动加以调整，具有变异性。因此，接着增加布林线技术分析指标。代码操作如代码五所示。

代码五 增加布林线的 K 线图代码

```
#增加布林线的 K 线图
stock <- as.xts(data, descr=df)
chartSeries(x=stock["2019-10-08/"], name='K 线图', line.type="l", bar.type="ohcl", theme="white", up.col='red', dn.col='green'.
TA="addVo();addSMA(5);addSMA(10);addMACD();")
```

代码结果展示如下图 3。

Figure 3. The graph with Bollinger bands
图 3. 加布林线的图

从图 3 可以看出，从 2021 年 6 月份开始，上轨，中轨和下轨线同时下行，显示出了股市的弱势特征，在此阶段出现了下行的走势，股票价格持续下滑。可能是受到了外在的影响。在 2021 的六月，上、中、下三个阶段将同时启动。利用 R 方法求出，并据此进行的分析。此时需要装载一个安装包来进行性能分析，然后装载一个程序包，并在对数据报酬率的计算中，需要呼叫 periodReturn 函数来对每个时间段内的报酬率分类进行估计。此实现代码在代码六中被展示。

代码六 计算收益代码

```
#分析日收益率
cou<-df[,5]
Profit<-diff(log(cou))
head(Profit)#看前 5 个
tail(Profit)#看后 5 个
my_Profit<-data.frame(Profit)#转化成数据类型
```

计算的结果如图 4 所示。

```
> head(Profit)#看前5个
[1]  0.011900180 -0.011900180  0.008740620 -0.003169575  0.017309638  0.006220860
> tail(Profit)#看后5个
[1]  0.022962445  0.002325582 -0.024097552 -0.034496613  0.000615574 -0.003081667
```

Figure 4. Calculation result graph
图 4. 计算结果图

接着画出股票的收益率 K 线图，代码如代码七所示。

代码七 绘制股票收益 K 线图代码

```
#绘制 K 线图
ddf<-read.xlsx("123.xlsx",1)
myvars_1<-c("close")
data_1<-xts(ddf[myvars_1], order.by=as.Date(as.character(ddf[,1]),format="%Y-%m-%d"))
head(data_1)#取前 6 项
#收益率的 K 线图
stock <- as.xts(data_1, descr=ddf)
chartSeries(x=stock["2019-10-08/"], name='K 线图', line.type="l", bar.type="ohcl",theme="white", up.col='red', dn.col='green')
```

结果如下图 5 所示。

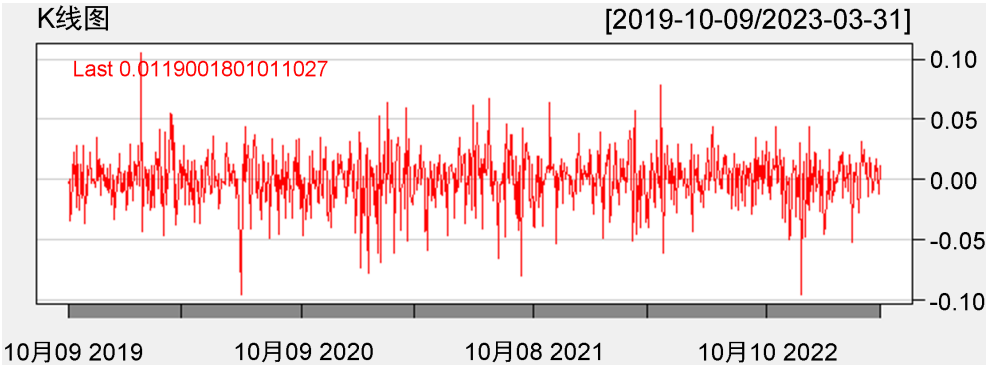


Figure 5. Earnings K-line chart
图 5. 收益 K 线图

绘制密度函数图之前需下载 fBasics 软件包，在 R 中编写代码载入相关程缉包，获取密度函数，查看数据的取值范围，根据这一范围，绘制密度函数即可。代码如代码八所示。

代码八 绘制密度函数代码

```
#导入密度函数的包
library(fBasics)
library(timeDate)
library(timeSeries)
de=density(Profit)     #获取密度函数
range(Profit)          #查看数据的取值范围，相当于 c(min(x),max(x))
x=seq(-.17,.17,.001)   #生成一个下界是 -0.17，上界是 0.17，时间间隔是 0.001 的数据，取值范围主要由 range 的结果决定
plot(de$x,de$y,xlab='x',ylab='density',type='l') #画密度函数图
ys=dnorm(x,mean(Profit),stdev(Profit))  #新建一个与 SINA.Pro 均值和标准差一致的正态分布函数
lines(x,ys,lty=2) #在密度函数图上增加正态分布曲线(图中虚线)
```

所得到的图表显示在图 6 中。

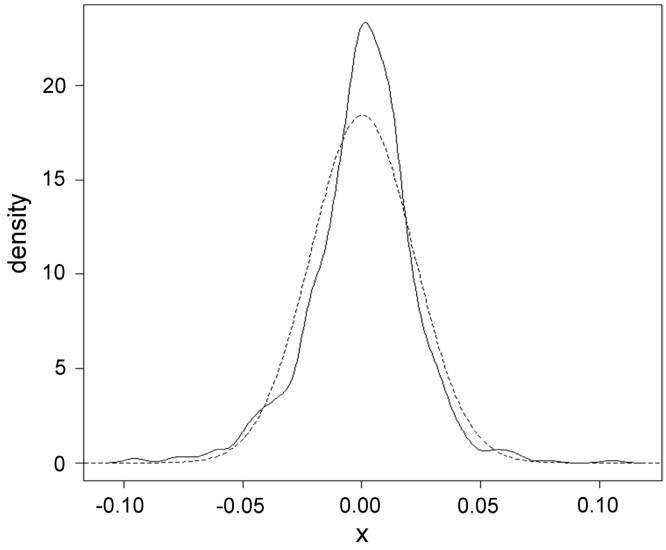


Figure 6. Stock return density map
图 6. 股票收益密度图

基于以上分析，我们得出了新浪股价的概率曲线偏离了正态分布，因而还需要对其进行正态性测试以进行进一步的验证。基本统计量和正规检验编码见代码九。

代码九 绘制股票收益 K 线图代码

```
#股票的基本统计量
basicStats(Profit)
#股票收益的正态检验
normalTest(Profit,method='jb')    # `jb`代表 JB 正态性检验
```

如图 7 和图 8 所示。

```
> basicStats(Profit)
                Profit
nobs        847.000000
NAs           0.000000
Minimum      -0.095629
Maximum       0.105075
1. Quartile  -0.010546
3. Quartile   0.012633
Mean          0.000303
Median        0.001338
Sum           0.256885
SE Mean       0.000744
LCL Mean     -0.001156
UCL Mean      0.001763
Variance      0.000468
Stdev         0.021641
Skewness     -0.325095
Kurtosis      2.221864
```

Figure 7. Basic statistics graph
图 7. 基本统计量图

图 7 中的计算结果来看，调整后的新浪收益率数据中，均值等于 0.000303，非常接近于 0，表示新浪股票收益率有显著向 0 集中的趋势；方差等 0.000704，接近于 0，表示这段时期内新浪股票收益率的离散程度比较小，也可说是不分散的；偏度为 0.000468，明显不等于 0，说明新浪股票收益率分布具有非对称性；峰度(Kurtosis)等于 2.221864，明显小于 3，说明了新浪股票收益率存在明显的高峰厚尾现象。这与之前绘制的密度函数相吻合，但仍需进一步进行正态性检验，正态性检验如图 8 所示。

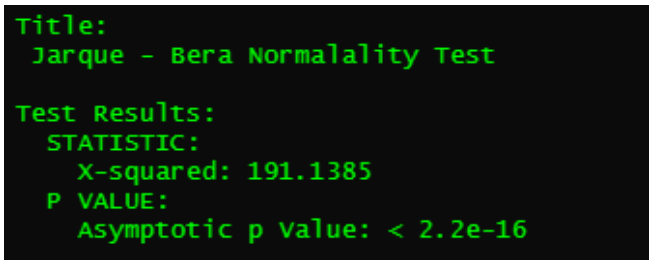


Figure 8. JB normality test chart
图 8. JB 正态检验图

由于利用 R 计算 JB 统计量十分便捷，故本文主要运用 JB 检验。调用 normalTest()函数，方法设置改为 JB 即可。根据图 8 的 JB 值为 191.1385，且 P = 2.2e−16 < 0.05，表明在 5%的显著性水平下应拒绝原假设，说明新浪股票收益率不服从正态分布。由于股市成交量是股票买卖双方完成交易的数量，这也是技术分析中经常使用的重要指标。应用 R 计算某时间段股票总成交量是十分简便的，只需输入函数命令 getSymbols，并分别调用 chartSeries()、summary()和 sum()三个函数便可得到股票成交量 K 线图和成交量数据汇总结果，如图 9。

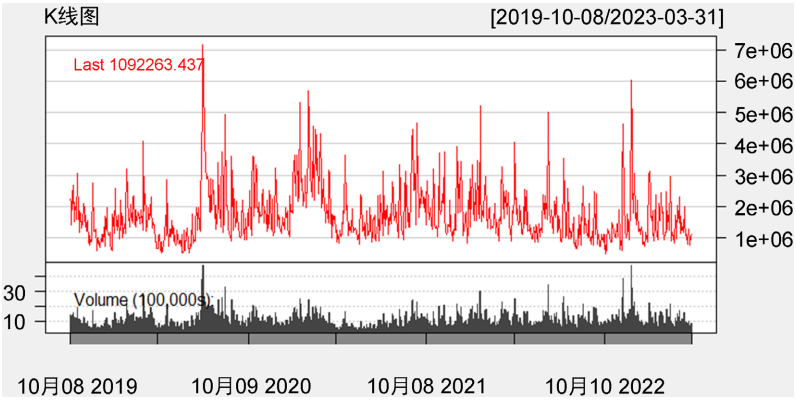


Figure 9. K-line diagram and volume data summary chart
图 9. K 线图和成交量数据汇总图

5. 小结

R 语言在数据采集与可视化方面的优势，在证券市场上得到了广泛的运用。因为资讯的品质与资讯的表现有著密切的关系，因此，透过资料的解析，将其呈现出来，有助于使用者了解资料中的资讯，进而发掘资料的价值。资料视觉化的实质就是视觉交谈。资料可视化是利用资料处理与绘图技术，以清楚而又高效的方式，来表达资料所呈现出来的资料。资料与视觉资料是相互补充的。资料是资料的依据，而资料的可视化则使资料更具弹性。利用可视化技术，可以使公司更好更高效地获取有用的信息。

比如，在一个电商网站上，一般都会把建议放在首页，商品详情页以及商品页面上。当使用者浏览

到这个网页，他们会被这个网页上所陈列的商品所吸引，并且想要看到这个网页上与商品有关的资讯。在这种情况下，我们就可以利用 R 语言中的一个脚本抓取功能来抓取这个数据。

参考文献

[1] 杨晓伟, 杨鸿鲜, 刘相国, 刘倩倩. 基于 R 语言的金融数据分析——以新浪股票数据为例[J]. 贵阳学院学报(自然科学版), 2020, 15(1): 43-49+62. https://doi.org/10.16856/j.cnki.52-1142/n.2020.01.012

[2] 孟诗琼, 孟诗瑶, 尹志. 基于 R 语言的汽车消费数据挖掘及可视化方法[J]. 宁波工程学院学报, 2015, 27(4): 17-23.

[3] 庄旭东, 王志坚. 基于 R 语言爬虫技术的网页信息抓取方法研究——以抓取二手房数据为例[J]. 科技风, 2019(6): 54+56. https://doi.org/10.19392/j.cnki.1671-7341.201906047

[4] 李天赐. 基于 R 语言的财务数据可视化方法应用研究[D]: [硕士学位论文]. 哈尔滨: 黑龙江大学, 2019. https://doi.org/10.27123/d.cnki.ghlju.2019.001156