数据集清洗上机

邱飞

万里学院

## smart move

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

smart\_move2 = filter(return,return >= -0.11)  
smart\_move = filter(smart\_move2,return <= 0.11)  
  
smart\_girl = filter(smart\_move,  
 code == "SHA000001" | code == "SHE399001")  
  
library(ggplot2)  
smart\_pic\_girl = ggplot(smart\_girl,   
 aes(x = date, y = return, color = code))+geom\_line()

## show smart pic



## 如果股市没有涨跌幅限制

shanghai = filter(stock, code == "SHA000001")  
shenzhen = filter(stock, code == "SHE399001")  
  
summary(shanghai$close)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.11 2863.00 3086.00 3142.00 3389.00 5166.00

summary(shenzhen$close)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6998 8171 9723 9897 10820 18100

## 找出异常数值点

shanghai\_sort = arrange(shanghai, close)  
  
library(knitr)  
kable(head(shanghai\_sort))

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| code | date | open | high | low | close | volume |
| SHA000001 | 2016-09-13 | 3.09 | 3.11 | 3.07 | 3.11 | 12286226 |
| SHA000001 | 2016-10-26 | 7.05 | 7.05 | 6.95 | 6.99 | 7384515 |
| SHA000001 | 2014-06-19 | 2054.62 | 2059.25 | 2017.65 | 2023.73 | 92914273 |
| SHA000001 | 2014-06-23 | 2026.23 | 2033.32 | 2022.92 | 2024.37 | 70485162 |
| SHA000001 | 2014-06-25 | 2030.43 | 2030.62 | 2018.36 | 2025.50 | 67539609 |
| SHA000001 | 2014-06-20 | 2013.41 | 2027.15 | 2010.53 | 2026.67 | 67167245 |

## 清除异常值点

hushen\_new = filter(stock,  
 code == "SHA000001" | code == "SHE399001")  
  
hushen\_smart = filter(hushen\_new, !(code == "SHA000001" & close < 2000))

## 计算收益率

hushen\_grouped = group\_by(hushen\_smart,code)  
  
hushen\_return = transmute(hushen\_grouped,  
 date = date,  
 return = close/lag(close) -1)

## Adding missing grouping variables: `code`

smart\_pic\_boy = ggplot(hushen\_return,   
 aes(x = date, y = return, color = code))+geom\_line()

## show

## Warning: Removed 2 rows containing missing values (geom\_path).



## 对全体数据清洗

library(dplyr)  
return\_clean = filter(return,return>=-0.1 & return <= 0.1)

## 看一下我们清洗的成果

#清洗前的记录数  
(number\_before = nrow(return))

## [1] 1014928

#清洗后的记录数  
(number\_after = nrow(return\_clean))

## [1] 989634

#被清洗掉的数据条数  
number\_kill = number\_before - number\_after  
  
number\_kill

## [1] 25294

数据清洗前，一共有1014928条数据，数据清洗后，一共有989634条数据，总共清洗了25294条数据。

## 用清洗过后的数据直接画图

* 看看与之前我们逐步清洗出来的结果是否一致

hushen = filter(return\_clean,  
 code == "SHA000001" | code == "SHE399001")  
library(ggplot2)  
hushen\_plot = ggplot(hushen,  
 aes(x = date, y = return,color = code))+  
 geom\_line()

## show plot



## 对股票数据集进行清洗

* 既然return数据集中有错误的数据
* 而return数据集又是从stock数据集计算得到的
* 那么着说明stock数据集中也有错误的数据存在

## 于是，以收益率数据集为条件，对股票数据集进行清洗

* stock和return都有code和date列
* return清洗的过程是去掉一些不符合条件的记录，也就是说它的date的范围会变小
* 于是我们可以用return里面的date的取值来替代stock里面的date，这样就起到了**近似**对stock清洗的效果。

## 清洗股票数据集

* 使用dplyr里的inner\_join函数来合并stock和清洗后的return数据集。
* inner\_join函数by后面的参数是合并的参照变量

stock\_grouped = group\_by(stock,code)  
  
stock\_clean = inner\_join(stock\_grouped,return\_clean,  
 by = c("code","date"))

## 保存清洗过后的股票数据集

save(stock\_clean,file = "stock\_clean.rda")

* 至此，本讲的数据清洗部分完成了。后续我们分析的数据集将是stock\_clean。