大数据高端人才专项计划

东北大学 Northeastern University

HORIZON 昊宸科技

R数据处理包进阶

- **数据处理包进阶：**

在学习了R的一些处理数据的方法后，建议大家去学习下其他数据处理的第三方包，如之前理论课介绍的data.table。

推荐大家去学习这些包： dplyr 、tidyr和lubridate

实训项目

数据介绍：

systation.csv 中包含的是沈阳地铁站相关线路的经纬度数据，字段介绍如下表：

变量名称	描述
station	站名
line	线路编号
line_id	站编号
gps_lon	站处于的经度
gps_lat	站处于的维度

SY-20150401.csv中提供的是一卡通刷卡数据，字段介绍如下表：

变量名称	描述
V1	卡编号
V2	刷卡日期
V3	刷卡时间
V4	刷卡站点
V5	刷卡类型
V6	票价
V7	是否优惠

注：输入数据后将字段名，重命名为c('card_id','date','time','station','vehicle','money','property')

实训项目

题目1：

1.数据处理【只用SY-20150401.csv】

统计每5分钟，各站点的进出站流量（说明：00:00:01在第一个5分钟内，00:10:13在第三个5分钟内），由于一天可能多次乘坐地铁，根据卡号和进站时间，查询最近出站的时间，作为本次出站时间。参考函数lubridate::hms，lubridate:: period_to_seconds。

处理的最终结果：dataframe(名称trade.metro.in.out)

字段	字段说明
card.id	卡id
time.in	进站时间
line.in	进站线路
station.in	进站站点
M5.in	第几个5分钟进站
time.out	出站时间
line.out	出站线路
station.out	出站站点
money	票价
M5.out	第几个5分钟出站
duration	乘坐时间（单位：秒）

提交的表格按照card.id排序。

Notebook中展示输出结果为

head(trade.metro.in.out,10)

并输出本地文件

shmetro_line_in_out.csv

实训项目

题目2：

1.统计进站与出站之间的流量

（1）通过题目1中的dataframe（trade.metro.in.out）进行统计，统计进站与出站之间的流量；然后选取流量最大的前10个站点，在Notebook中查看前6条。

提交内容：代码和处理后的shmetro_line_in_out.csv ，将所要提交的内容打成压缩包。

THANKS