大数据高端人才专项计划

東北大學 Northeastern University

HORIZON 昊宸科技

R数据处理知识点回顾

- **知识点1：**

变量创建、变量重编码、缺失值、日期值处理，数据类型转化，数据排序。

- **知识点2：**

数据集的合并，选取子集，使用SQL操作数据框，数据的整合与重构。

- **知识点3：**

控制流：条件与循环

- **知识点4：**

用户自编函数

实训项目

题目1：

1.将item_feature1.csv读入，存储到df中；并给df的列分别命名为： date 、item_id 、 cate_id 、 cate_level_id 、 brand_id 、 supplier_id 、 pv_ipv 、cart_uv 、 collect_uv 和cart_ipv 。

注：【日期、商品id、仓库id、仓库级别id、品牌id、供应商id、浏览次数、加购人次、收藏人次和被加购次数】

2.为df中的cart_uv 重新编码并将新变量命名为recode，将小于5000的归为less,将大于等于5000小于15000的归为common,其他的归为many；查看尾部的10条数据。

3.查看df中是否有缺失值；如有缺失值，删除df中所有含缺失值的行。

4.将df中的date字段转换成日期类型，如：2015-02-13。

5.将df按照date字段升序排列，另存为df_asc；并查看前10条数据。

6.将df按照date字段升序和item_id降序排序，另存在df1中；并查看前5条数据。

实训项目

题目2：

1.从df中选取date 、 item_id 、 cate_id 、 cart_uv 、 recode、 collect_uv 和cart_ipv字段另存为df1；剔除df1中的cart_ipv字段另存为df2；从df1中选取item_id 大于500的并且recode为less的数据另存为df3。

2.从df中选取date为2015-02-14，item_id为300，并保留date到supplier_id其间的所有列，另存为df_sub。

3.从df中无放回的随机抽取500条样本，另存为df4；查看样本的维度和数据的头部数据。

4.从df1中选取列从item_id到cate_id的数据，另存为df1_temp,然后与df按照item_id合并存为df5。

5.从df1中利用sql的方法选取item_id为300的数据，另存为df6中。【注：sqldf包】

6.从df2中有放回的随机取出与df6一样多的数据条数做为df_tem，然后与df6按列（横向）合并，另存为df7。

7.从df中选取 date 、 item_id 、 cate_id 和cart_ipv 另存为feature，并将feature 按日期升序排列，取出feature 中唯一的cate_id 【去重即可】。

注：提交2个题目的代码和运行结果，格式：R第2次实训+姓名+学号。

THANKS