数据仓库与数据挖掘

软硕161 徐毅 2016213585

# 一、数据清理

## 1.1数据清理原则

由于数据中存在噪声，在开始本次实验之前首先对每个数据文件进行了格式检查，主要看每个字段是否满足相应的格式要求，每张表的主键是否唯一。实验中采用了编程检测的方式。

## 1.2 account表清理

account清理时需要注意以下事项：

1. account\_id必须为数字且不能重复
2. district\_id为数字
3. date可以转化为标准的日期格式，且为合法日期
4. frequency字段必须在{ ‘POPLATEK MESICNE’, ‘POPLATEK TYDNE’, ‘POPLATEK PO OBRATU’ }中出现

该表中的异常数据点如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| account\_id | district\_id | frequency | date | 错误原因和处理方法 |
| 1675\_[typo] | 36 | POPLATEK MESICNE | 1993/2/1 0:00 | 主键有非法字符，可修复 |
| "1801\_[I'm typo, 2]" | 46 | POPLATEK MESICNE | 1993/2/13 0:00 | 主键有非法字符，可修复 |
| 640 | 32 | POPLATEKMESICNE | 1993/6/2 0:00 | Frequency字段错误，可修复 |
| 576 | 55 | POPLATEK MESICNE | 1993/1/1 0:00 | 主键重复，丢弃 |

## 1.3 card表清理

card清理时需要注意以下事项：

1. card\_id为数字且不能重复
2. disp\_id为数字
3. type字段在{ ‘junior’, ‘classic’, ‘gold’ }中出现
4. issued以转化为标准的日期格式，且为合法日期

该表中的异常数据点如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| card\_id | disp\_id | type | issued | 错误原因和处理方法 |
| 546 | 3465 | golden | 1997/8/19 0:00 | type字段错误，可修复 |
| 339 | 2154 | classic | 1998/10/19 0:00 | disp\_id不存在，丢弃 |

## 1.4 client表

client清理时需要注意以下事项：

1. client\_id为数字且不能为重复
2. birth\_number必须为六位数字，且是合法的日期
3. district\_id必须为数字

在清理该表的时候，顺便统计了所有人的性别和星座，生成了gender和star字段。

该表中的异常数据点如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| clinet\_id | birth\_number | district\_id | gender | 错误原因和处理方法 |
| 2 | 450229 | 1 | MALE | 日期格式错误，丢弃 |
| 697 | 740232 | 1 | MALE | 日期格式错误，丢弃 |
| 828 | 582210 | 10 | MALE | 日期格式错误，丢弃 |
| 2154 | 576628 | 60 | FEMALE | 日期格式错误，丢弃 |
| 4484 | 550011 | 37 | MALE | 日期格式错误，丢弃 |
| 4485 | 190-20 | 36 | MALE | 日期格式错误，丢弃 |

## 1.5 disp清理

disp清理时需要注意以下事项：

1. disp\_id为数字且不能重复
2. client\_id合法数字
3. account\_id合法数字
4. type只能在{ ‘owner’, ‘disponent’ }中取值

该表中的异常数据点如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| disp\_id | client\_id | account\_id | type | 错误原因及处理方法 |
| 2 | 2 | 2 | OWNER | client\_id不存在，丢弃 |
| 697 | 697 | 579 | OWNER | client\_id不存在，丢弃 |
| 828 | 828 | 690 | DISPONENT | client\_id不存在，丢弃 |
| 2154 | 2154 | 1777 | OWNER | client\_id不存在，丢弃 |
| 4484 | 4484 | 3719 | OWNER | client\_id不存在，丢弃 |
| 4485 | 4485 | 3720 | OWNER | client\_id不存在，丢弃 |

## 1.6 district清理

client清理时需要注意以下事项：

1. district\_id为数字且不能重复
2. Hab\_number为数字
3. City\_number为数字
4. Ave\_salary为数字
5. Umemploy\_rate为数字
6. Crime\_number为数字

清理的过程中没有发现异常数据

## 1.7 order清理

order清理时需要注意以下事项：

1. order\_id为数字且不能重复
2. account\_id为数字为account表的外键
3. bank\_to两个字母
4. account\_to为数字
5. Amount为数字
6. K\_symbol必须在{ ‘POJISTNE’, ‘SIPO’, ‘LEASING’, ‘UVER’}中出现

该表中的异常数据点如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| order\_id | account\_id | bank\_to | account\_to | amount | K\_symbol | 错误原因及处理方法 |
| 41351 | 8045 | AB | 43162244 | 4340 | UVER | account\_id不存在，丢弃 |
| 42946 | 9139 | CD | 64458685 | 9264 | SIPO | account\_id不存在，丢弃 |
| 44598 | 10245 | IJ | 7045454 | 3151 | UVER | account\_id不存在，丢弃 |
| 46339 | 11321 | CD | 79047823 | 8758 | SIPO | account\_id不存在，丢弃 |

## 1.8 trans清理

trans清理时需要注意以下事项：

1. trans\_id为数字且不能重复
2. account\_id为数字为account表的外键
3. date为合法的日期格式
4. type在{ ‘PRIJEM’, ’VYDAJ’ }中取值
5. operation在{ ‘VYBER KARTOU’, ’VKLAD’, ’PREVOD Z UCTU’, ’VYBER’, ’PREVOD NA UCET’ }中取值
6. amount为数字
7. balance为数字
8. k\_symbol在{ ‘POJISTNE’, ’SLUZBY’, ’UROK’, ’SANKC. UROK’, ’SIPO’, ’DUCHOD’, ’UVER’, ’ ’}中取值
9. bank两个字母

这部分错误较多，主要是bank字段为空，其他的还有一些balance和account\_id的格式不对的问题

## 1.9 loan清理

loan清理时需要注意以下事项：

1. loan\_id为数字且不能重复
2. account\_id为数字为account表的外键
3. date是合法的日期格式
4. amount为数字
5. duration书数字且是12的倍数
6. payments数字且payments \* duration = amount
7. status在{ ‘A’,’B’,’C’, ’D’ }中取值
8. payduration为数字且payduration < duration

该表中的异常数据点如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| loan\_id | account\_id | amount | duration | payments | 错误原因及处理方法 |
| 5997 | 4894 | 117024 | 22 | 4876 | duration不为12的倍数，丢弃 |
| 5495 | 2580 | 89340 | -60 | 1489 | duration为负数，取正数 |
| 7130 | 10411 | 123402 | 24 | 5142 | payments\*duration!=amount，丢弃 |

# 二、数据导入