股权数据分析--SAS

背景

现有原始股权表CCB_FIVE_INVESTMENT_DETAIL_S7如下（除P9_end_date数据类型为datetime，其他字段均为字符串类型）：

层级	集团名称	股东名称	被投资公司名称	营业状态	持股比例	统一社会信用代码	P9_end_date
1	A1	B1	C1	在营（开业）	99.999	12345678XT21	2019-11-01
1	A1	B2	C2	在营（开业）	81.000	12345678XT22	2019-11-01
2	A2	B3	C3	在营（开业）	10.000	12345678XT23	2019-11-01

任务

任务一

已知数据存在质量问题，请按照下面的步骤采用SAS的PROC SQL步（模板见附录）对表进行清洗，将清洗后的表新建为CCB_FIVE_INVESTMENT_DETAIL_S7：

1. 过滤企业名称中的特殊字符；
2. 股权表为拉链表，取P9_end_date>2019-10-31 and P9_start_date<2019-10-31 ，取 营业状态 = '在营（开业）'；
3. 对企业名称中全角字符、半角字符、空格和括号做去除操作；
4. 持股比例限制在（0,100%]；
5. 重复数据（以集团名称、股东名称、被投资公司名称三个字段作为主键）直接删除，保留一条；

任务二

任务一得到的表CCB_FIVE_INVESTMENT_DETAIL_S7，经分析发现存在如下逻辑问题，请采用SAS的PROC SQL步分别找出这些公司：

1. 自持股（自己投资自己的公司，建表为AA）；
2. AB交叉持股（A投资B，B投资A，建表为AB）；

3. 母公司之间持股（母公司：层级为1的股东，建表为MU）；

任务三

采用表CCB_FIVE_INVESTMENT_DETAIL_S7，统计如下信息，请写出相应的PROC SQL语句：

1. 总数据量、集团数量、股东数量、被投资公司数量；
2. 各层级的数据量、被投资公司数量；
3. 结合任务二新建的表，分别统计母公司之间持股、自持股、AB交叉持股、重复数据的数据量、被投资公司数量；

附录

PROC SQL模板

建表模板

```
proc sql;
connect to greenplm as gpconn (server=a123 port = 5432 database ="a" authdomain
= a1234);

execute
(
  drop table if exists schame1.table1;
  create table schame1.table1 with (appendonly=true, compresslevel=5) as
  (

    select * from schame1.table

  )
  with data
  distributed by (CRCRD_CARDNO)  /* 新表主键 */
)by gpconn;
execute (commit) by gpconn;

execute
(
  drop table if exists schame1.table2;
  create table schame1.table2 with (appendonly=true, compresslevel=5) as
  (

    select * from schame1.table1
```

```
  )
  with data
  distributed by (CRCRD_CARDNO)  /* 新表主键 */
)by gpconn;
execute (commit) by gpconn;

disconnect from gpconn;
quit;
```

查询模板

```
proc sql;
connect to greenplm as gpconn (server=a123 port = 5432 database ="a" authdomain
= a1234);

select * from connection to gpconn (


  select * from schame1.table1

  )

disconnect gpconn;
quit;
```