**8.0版问卷数据清洗规则**

**一、筛选出专业名称为“测试专业”的答卷，作为废卷处理。**

# IF departments\_name = testing\_departments then rinse student\_info

**二、根据报告类型清洗学生整条答卷信息。**

1、8.0**对外报告**分析“提交答卷（有提交时间）+未提交答卷（无提交时间但回答了A2）”的学生。“未提交答卷”的学生回答了问卷中毕业去向“**A2、(必答题)您目前的就业状态：**”题目，且对外报告分析涉及到各学院、各专业的指标较少，对外增量的一般都是客观数据增量。因此适用此规则，同时也提高了回收率。

即毕业去向题为“**A2、(必答题)您目前的就业状态：**”，如该题答题信息为空，作为废卷处理。

# IF A2 = NULL then rinse student\_info

2、8.0**对内报告**分析“提交答卷”学生。对内报告分析涉及到各学院、各专业的指标较多，避免各指标分析的人数不一致，故不分析“未提交答卷”的学生群体。当回收率低于总人数20%的情况下，与校方或销售沟通同意后，当部分学校要求保留未交卷数据时，分析“提交答卷+未提交答卷”的学生（使用第一条规则，同8.0对外报告处理）。

#IF submission\_time != NULL:

student\_info = student\_info\_complete

ELSE: student\_info = student\_info\_incomplete

IF recovery\_rate < 0.2 and Universities\_or\_salesmen = agree:

student\_info\_incomplete is in report\_internal

ELSE: student\_info\_incomplete is not in report\_internal

**三、跳答规则，是否导致无效数据：**

# IF A2 not in (在国内工作, 自由职业) then rinse B1, B2, B3, B4, B5, B6, B7, B8, B9-1, B9-2, B10-1, B10-2, D1, D2

# IF A2 = 自由职业 then rinse B1,B2,B3,B4, B10-1, B10-2

# IF B9-1 not in (比较不相关, 很不相关) then rinse B9-2

# IF B10-1 not in (1次, 2次, 3次及以上) then rinse B10-2

# IF A2 != 未就业 then rinse C1, C2

# IF A2 != 在国内升学 then rinse E1,E2,E3,E4

# IF E3 not in (比较不相关, 很不相关) then rinse E4

# IF A4 != 出国/出境 then rinse F1, F2, F3, F4

# IF F1 != 求学 then rinse F2, F3

# IF F3 not in (比较不相关, 很不相关) then rinse F4

# IF A4 != 自主创业 then rinse G1, G2, G3, G4, G5

**I2-1各选项跳答题：**

#IF choice1 in I2-1 = NULL then rinse I2-1-1, I2-1-2

#IF choice2 in I2-1 = NULL then rinse I2-2-3, I2-2-4-R

#IF choice3 in I2-1 = NULL then rinse I2-3-5, I2-3-6

#IF choice4 in I2-1 = NULL then rinse I2-4-7, I2-4-8

#IF choice5 in I2-1 = NULL then rinse I2-5-9, I2-5-10, I2-5-11, I2-5-12, I2-5-13

#IF choice6 in I2-1 = NULL then rinse I2-6-14, I2-6-15-R, I2-6-16, I2-6-17, I2-6-18

#IF choice7 in I2-1 = NULL then rinse I2-7-19, I2-7-20-R

#IF choice8 in I2-1 = NULL then rinse I2-8-21, I2-8-22-R

#IF choice9 in I2-1 = NULL then rinse I2-9-23, I2-9-24, I2-9-25

#IF choice10 in I2-1 = NULL then rinse I2-10-26, I2-10-27

#IF choice11 in I2-1 = NULL then rinse I2-11-28, I2-11-29

#IF choice12 in I2-1 = NULL then rinse I2-12-30, I2-12-31, I2-12-32

#IF choice13 in I2-1 = NULL then rinse I2-13-33-R, I2-13-34

#IF choice14 in I2-1 = NULL then rinse I2-14-35-R, I2-14-36

#IF choice15 in I2-1 = NULL then rinse I2-15-37, I2-15-38, I2-15-39

#IF choice16 in I2-1 = NULL then rinse I2-16-40, I2-16-41-R, I2-16-42, I2-16-43-R

#IF choice17 in I2-1 = NULL then rinse I2-17-44, I2-17-45, I2-17-46, I2-17-47-R

#IF choice18 in I2-1 = NULL then rinse I2-18-48, I2-18-49, I2-18-50, I2-18-51-R

#IF choice19 in I2-1 = NULL then rinse I2-19-52, I2-19-53-R, I2-19-54, I2-19-55, I2-19-56-R, I2-19-57

#IF choice20 in I2-1 = NULL then rinse I2-20-58, I2-20-59-R, I2-20-60, I2-20-61-R, I2-20-62, I2-20-63-R

#IF choice21 in I2-1 = NULL then rinse I2-21-64, I2-21-65

#IF choice22 in I2-1 = NULL then rinse I2-22-66, I2-22-67, I2-22-68-R

**五、筛选出选项为“无法评价”、****“以上均不需改进”的选项用空值代替。**

#IF choice in (无法评价，以上均不需改进) then choice = NULL

**六、问卷G1中清洗“国际组织”和“军队”，如下：**

**G1、(必答题)您自主创业的行业是**：（两个下拉列表进行选择）【编码题】

（行业分类同B5题选项。但后续在数据清洗时，将选择“国际组织”和“军队”这两项人员在该题的答案清洗掉。）

#IF G1 = 国际组织 or 军队 then G1 = NULL

**七、数字题：如B6题 月薪异常数据**

1、清洗较低薪酬：

#IF salary < 1000 then salary = NULL

2、清洗较高薪酬：

1. 将清洗较低薪酬后的“薪酬水平”数据列降序排列；

# sorted(salary, reverse = False)

2）确定“回答薪酬人数-清洗较低薪酬人数”的前0.3%人的较高薪酬。筛选并清洗掉这些人的薪酬。0.3%人四舍五入取整数，当第0.3%人此处的薪酬=第N人的薪酬，一直清洗到第N人的薪酬，用缺失值（空值）代替。

如全校调研毕业生人数为4000人，其中3500人回答薪酬，有200人的薪酬低于1000元，此时清洗较高薪酬是前（3500-200）\*0.3%=9.9人。四舍五入为10人。当第10人、第N人的薪酬为一个值时（并列第10位的薪酬），一直清洗到第N人的薪酬。

1. 清洗大于4倍标准差的薪酬：

1）求出剔除两端异常值之后“薪酬水平”数据列的均值和标准差；

均值(=AVERAGE(C:C)) 标准差(=STDEV(C:C))

2）根据命令，将“薪酬减去均值的绝对值”超过“四倍标准差”的薪酬值用缺失值（空值）代替。

=if(a5="","",if(abs(a5-均值）>4\*标准差，"",a5))

**八、就业结果中为了和往届（6.0问卷）保持一致，只分析A2题选择“在国内工作”的群体，不与“自由职业”群体同时分析。即就业结果不分析“自由职业”。**

# IF A2 = 自由职业 then rinse B5,B6,B7,B8, B9-1, B9-2