**8.0版问卷数据清洗规则**

**一、筛选出测试人员的答卷，作为废卷处理。**

即账号为JC（搜索）或专业名称为“测试专业”的答卷，作为废卷处理。

**二、根据报告类型清洗学生整条答卷信息。**

1、7.0**对外报告**分析“提交答卷+未提交答卷”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。

2、7.0**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。

**三、根据毕业去向那道题目筛选出空白选项，作为废卷处理。**

毕业去向题为“**A2、(必答题)您目前的就业状态：**”，如若是空白答卷，作为废卷处理。

**四、跳答规则，是否导致无效数据：**

根据跳转规则筛选不同的答卷，以“**A2、(必答题)您目前的就业状态：**”为主查看是否有多填的选项，对于多填的选项，将“多填的选项”清洗掉。

如“未就业”分析群体是【A2选择“6.未就业”的人群】，清除回答此题的就业人群。

**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。**

**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。）

**七、数字题：如月薪异常数据**

1、清洗较低薪酬：

先将＜1000的值筛选出并清洗掉，用缺失值（空值）代替；

=IF(A1="","",IF(A1<1000,"<1000",A1))

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；

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))

* 如果是985学校与报告组人员沟通一下，例如南开大学、苏州大学等。