数据清洗规则说明

1. 去掉选项前缀与冗余描述

- **适用列**： 工作意愿 、 全职或兼职 、 14.您希望它是线上的还是线下的？
- **规则**：去掉诸如 A.、 B. 前缀，仅保留末尾的答案。
- **示例**：
  - A.希望 → 希望
  - B.兼职 → 兼职
  - B.线下 → 线下

2. 地区字段标准化

- **适用列**： 常住地
- **规则**：去掉前缀与括号内省份列表，仅保留大区名。
- **映射示例**：
  - A.东部地区（…） → 东部地区
  - B.中部地区（…） → 中部地区
  - C.西部地区（…） → 西部地区
  - D.东北地区（…） → 东北地区
  - E.香港、澳门、台湾或国外 → 香港、澳门、台湾或国外

3. 孩子数量数值化

- **适用列**： 孩子数量
- **规则**：映射为整数。
- **映射表**：
  - A.没有 → 0
  - B.1个 → 1
  - C.2个 → 2
  - D.3个 → 3

4. 分值类题目提取分值

- **适用列**：
  25 — 27 、 28 — 29 、 31 — 33 、 35 — 39
- **规则**：提取“X分”前的整数。
- **示例**：
  - H.7分-能够独立完成C任务 → 7
  - C.2分-… → 2

5. 重要程度题（0–3 映射）

- **适用列**： 20. 、 21. 两个“重要吗？”问题
- **规则**：文本映射为 0–3 数字。
- **映射表**：

- A.非常重要 → 3
- B.相对重要 → 2
- C.不太重要 → 1
- D.完全不在意 → 0

6. 学历题（0–6 映射）

- **适用列**： 22.您的学历是？（0-6）
- **规则**：不同学历映射为 0–6 数字。
- **映射表**：
  - A.未上过学 → 0
  - B.小学 → 1
  - C.初中 → 2
  - D.高中/中专/职高 → 3
  - E.大学专科 → 4
  - F.大学本科 → 5
  - G.硕士及以上 → 6

7. 二元题（0/1 转换）

- **适用列**：
  - 23.您以前有参加过工作吗？
    - A.有 → 1
    - B.无 → 0
  - 34.您会使用手机、电脑等来上网吗？
    - A.会 → 1
    - B.不会 → 0

8. 数值型题清理

- **适用列**： 年龄 、 家务劳动时间 、 闲暇时间 、 17. 每周天数、 18. 每天小时、 19. 月收入、 24. 工作年限
- **规则**：全部转为整数；去掉空格、全角字符。

9. 通用处理

- 去除前后空格、统一标点（中英文括号均识别）。
- 保持原列名和顺序。
- 编号 字段保持原样，不做修改。

三个完整样本清洗前后对照

下表展示了编号 **315–313** 的三个样本在清洗前后的对比。

编号	常住地（前）	常住地（后）	孩子数量（前）	孩子数量（后）	工作意愿（前）	工作意愿（后）	全职兼职（前）	全职兼职（后）	线上/线下（前）	线上/线下（后）	社保重要性（前）	社保（
315	C.西部地区（内蒙古、广西…）	西部地区	B.1个	1	A.希望	希望	A.全职	全职	B.线下	线下	B.相对重要	2
314	B.中部地区（山西、安徽…）	中部地区	B.1个	1	B.不希望	不希望	B.兼职	兼职	A.线上	线上	D. 完全不在意	0
313	A.东部地区（北京、天津…）	东部地区	A.没有	0	A.希望	希望	A.全职	全职	B.线下	线下	B.相对重要	2