**数据预处理说明**

**1. 基本说明**

预处理后一共两份处理后数据，一份MiddleData.csv，一份QuantityData.csv。前者保留大量字符串，且列数较少，逻辑清晰，适合作图。后者将所有列全部数字化，适合进行模型构建等操作。

**2. 数据解析**

|  |  |  |  |
| --- | --- | --- | --- |
| **数据名** | **数据含义** | **MiddleData数据组织方式** | **QuantityData数据组织方式** |
| area | 大地区（北京上海等） | 列名area，字符串形式 | 独热码，列名area\_xxx，仅在正确地区显示1 |
| area\_detail | 具体地区（黄浦区、闵行区等） | 列名area\_detail，字符串形式 | 独热码，列名area\_detail\_xxx，仅在正确地区显示1 |
| company\_people | 公司现有人数（规模） | 列名company\_people，以字符串形式体现，约7类 | 已数字化，未归一化 |
| company\_service | 公司的业务 | 两列，company\_service\_1与company\_service\_2，分别是公司的两类业务，总种类数约60 | 多热码显示，每一列名company\_service\_xxx，若有该公司业务两类则两列有1 |
| company\_type | 公司种类（私企国企等） | 列名company\_type，共一列，直接字符串显示 | 独热码处理 |
| education | 要求的学历 | 列名education，共一列，直接字符串显示 | 数字化，["无", "初中及以下", "中技", "中专", "高中", "大专", "本科","硕士","博士"]分别对应[3, 0, 1, 1, 2, 2, 4, 6, 9] |
| hire\_num | 招聘人数 | 列名hire\_num，数字化，999代表不限人数 | 数字化，999（不限人数）被替换为中位数1 |
| Job\_work\_type | 工作性质 | 两列，job\_work\_type\_1与job\_work\_type\_2，分别代表两种工作性质 | 多热码编码，采用30个关键字符串进行关键字匹配，如：“销售”，列名job\_work\_type\_销售。若包含该子串，则该列为1（共30列，每一列代表一个关键词） |
| language | 语言要求 | 两列，language \_1与language \_2，分别是两类语言（因为有的工作需要两门语言） | 多热码，列名language\_xxx，如有两门语言则两列1 |
| work\_experience | 工作经历年限 | 列名work\_experience，字符串显示，共约10类 | 数字化，["无工作经验", "1年经验", "2年经验", "3-4年经验", "5-7年经验", "8-9年经验", "10年以上经验"]分别对应数字[0, 1, 2, 3.5, 6, 8.5, 11] |
| welfare | 公司福利列表（五险一金等） | 列名welfare，单一一个字符串，分一种福利间用逗号分隔 | 多热码编码，采用70个最常见福利进行关键字匹配，如：“销售”。若包含该子串，则该列为1. |
| salary\_max | 年薪最大值 | 数字化 | 数字化 |
| salary\_min | 年薪最小值 | 数字化 | 数字化 |
|  |  |  |  |