**《大数据分析与处理》实验报告**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **年级、专业、班级** | | **2019级计算机科学与技术卓越02班** | | | **姓名** | **李燕琴 李放** |
| **实验题目** | **实验一 数据探索与预处理** | | | | | |
| **实验时间** | **2022年4月7日** | | **实验地点** | **DS3 303** | | |
| **实验成绩** |  | | **实验性质** | **□验证性 □设计性 √综合性** | | |
| 教师评价：  □算法/实验过程正确； □源程序/实验内容提交 □程序结构/实验步骤合理；  □实验结果正确； □语法、语义正确； □报告规范；  其他：  评价教师签名： | | | | | | |
| 一、实验目的   * 熟悉如何利用Python语言来处理现实中大数据； * 通过对真实的数据集进行标准化、缺失值处理、数据概化等预处理操作，掌握数据预处理的方法； * 通过思考及实践自主发现数据间的关系，掌握如何利用可视化手段发现数据间的关系，为后续分析奠定基础。 | | | | | | |
| 二、实验项目内容  实验基于找到网人力资源数据集进行，数据中包括了用户姓名、部门、职务、工作教育经历以及所属公司的名称、注册资金额、年龄等字段。  本次实验分为两个部分，具体内容如下：  **1 数据清洗**  1.1 找出不符合规则的数据并进行修正  例如：字段内部多余的空格，字段为空或特殊字符的数据，字段“tel”不符合电话格式的数据等，选择合适的处理手段，例如删除、置默认值、置平均值/众数等，将数据处理为规则的数据。  1.2 根据字段含义将数据处理为合适的格式  例如：将“部门-职务”数据拆分为string类型的“部门”和“职务”两部分，将“注册资金”与“公司年龄”处理为数值类型，将“公司成立时间”处理为标准的yyyy-MM-dd日期格式，如2017-02-03，将“教育经历”与“工作经历”处理为合适的格式。  1.3 进行数据概化  例如：  职务等级划分为A级（关键词“董事/主席”等）、B级（关键词“总经理/总裁/副总经理”等）、C级（关键词“总监/副总监/经理/副经理/主任/主管”等）、D级（关键词“工程师/xx员/实习生”等）、E级（其他）；  工作类别划分为市场类（关键词“销售/市场/客户”等）、技术类（关键词“业务/技术/项目”等）、营销类（关键词“营销/宣传”等）、其他类（关键词“财务/运营/行政/人力”等）；  公司按照地区和领域进行划分，领域为科技类（关键词“科技/软件/信息技术”等），文化、传媒广告类（关键词“文化/传媒/广告”等）、咨询类（关键词“咨询”等）、管理类（关键词“管理”等）、贸易类（关键词“贸易/商贸/科贸/工贸”等）与其他类（关键词“机械/设备/建筑”等）；  注册资金等级划分为“1000万以下”、“1000万以上5000万以下”、“5000万以上1亿以下”、“1亿以上”。  **2 数据可视化**  2.1 按照数据类型进行图表展示  例如：按照地区、职务、工作类别、领域、公司年龄等字段制作合适的图表（饼图、直方图等），能够直观有重点地展示出基础数据的分布情况并得出适当结论。  2.2 相关分析  选择特定的字段，利用散点图等方法分析不同字段之间的关联关系，并得出相应的结论。例如：用户是否认证与公司是否认证之间的关联关系，公司注册资金与公司年龄的关系等。  2.3 对比分析  选择特定的字段，通过盒图、直方图、饼图等方式对比分析不同类型的用户或公司的差异，例如：分析认证用户和非认证用户职务分布情况的差异，分析认证公司和非认证公司注册资金、公司年龄的差异等。  **3 思考**  3.1 针对该数据中包含的具有一定隐私性的信息（如姓名），思考如何通过预处理方法对数据做隐私保护？ | | | | | | |
| 三、实验过程或算法（源程序）  本实验数据处理，主要基于pandas、numpy等python第三方库函数展开实验。具体步骤如下： 数据清洗初步读取数据 通过pandas读取的表格数据格式如下：    为了方便后续的数据处理，我们在读取数据时，即采用了默认值填补空缺数据,其中注册资金利用原数据的众数（元）填充，主要填补说明如：   |  |  |  | | --- | --- | --- | | 列 | 默认值 | 备注 | | 部门\_职务 | 无 |  | | 工作内容 | 无 |  | | 公司名称 | 无 |  | | qq | -1 | 读取时，为浮点格式，用-1替换方便后续分情况处理数值转字符串的情况 | | tel | -1 | 读取时，为浮点格式，用-1替换方便后续分情况处理数值转字符串的情况 | | 注册资金 | <资金众数>元 | 根据原数据格式填充 | | 公司年龄 | 0年 | 根据原数据格式填充 | | 教育经历 | 空字符串 | 避免解析为浮点值nan | | 工作经历 | 空字符串 | 避免解析为浮点值nan | | Email | 无 |  |  字段格式化 (1)qq和tel字符串化：其中空值默认为‘无’，电话正则表达式如下：  r"^1(?:3\d|4[4-9]|5[0-35-9]|6[67]|7[013-8]|8\d|9\d)\d{8}$"  具体处理如图 1所示:    图 1 qq和tel的处理  (2) 部门-职务的提取：  “部门-职务”的规范格式为“<部门> <职务>”，但因存在如图 2所示的格式数据，不能简单的用空格分割进行提取。    图 2 异常“部门-职务”格式  本实验我们采取如下的正则表达式：  r'\b\S+\b'  处理的逻辑如图 3所示：    图 3 部门-职务处理逻辑  (3) 注册资金和公司年龄的数字格式提取  注册资金和公司年龄的末尾汉字格式如图 4所示。    图 4 资金和年龄的主要汉字格式  故在转为数字格式时，需要单独提取出数字和汉字，并统一单位。  本实验中，注册资金统一为“元”，未填写资金的利用众数填补；公司年龄不足一年的视为1年，未填写的利用默认值0年填补。    图 5 注册资金和公司年龄的处理  (4) 公司的成立时间格式化  利用pandas提供的to\_datetime函数，直接对原始“公司成立时间”进行格式化，其中未填写时间的视为NaT（即空），我们对这个空值不做处理，保持pandas原有的格式，因为这样就可以利用pandas提供的分周、分季节的函数进行更高级的数据挖掘。  (5) 教育经历和工作经历的理解  本实验提供的数据，教育经历和工作经历的字段因比较复杂且固定的格式的数量不一，不易用与二维表表示，且在后续的数据挖掘中没有用到，故我们讨论决定对教育经历和工作经历不做处理。  (6) 替换多余的空格  处理好上述内容后，因“部门”和“职务”等字段尚且存在一些多余的空格，故对这些字段替换掉多余的空格，且如果是空，则替换为“无”。 数据概化 根据题目要求，我们利用如图 6所示的正则化公式对“职务”、“工作内容”、“公司领域”进行匹配，如果含有这些字段则将其分到该类中。需要注意的是，我们主观的认为一个“职务”只能属于某一个类别，并从上到下依次匹配，如果能匹配成功则范围对应的key作为其类别。  其中工作类别划分和题目涉及到的不太一样，如工作类别中，将“财务|运营|行政|人力”视为管理类，其他值视为其他类，以保证分类的主观合理性。  此外，对于公司年龄和注册资金，做了如图 7所示的划分。    图 6 等级划分正则化    图 7 注册资金和公司年龄的划分 数据可视化按照数据类型进行图表展示 根据1.3的概化职务等级、工作类别、公司领域、资金等级、公龄等级，绘制饼图展示数据分布。 相关分析 我们选择了用户是否认证与公司是否认证之间的关联关系，因均为0-1数据，故我们采用饼图和二维表格的分析方式观察其关联关系。  此外，因为原值的数据如图 8所示，呈现幂律分布的特征，故我们还对公司（注册资金+1）的log值与公司年龄绘制散点图，分析其关联分析。    图 8 注册资金（元）的分布示意图 对比分析 为了对比分析不同类型的用户或公司的差异，我们主要做了如下操作：  ①对于分析认证用户和非认证用户职务分布情况的差异分析，我们采取直方图进行对比呈现。  ②对于分析认证公司和非认证公司注册资金、公司年龄的差异等，我们分析了二者的盒图和直方图对比。 思考用户隐私保护的思考 针对该数据中包含的具有一定隐私性的信息（如姓名），思考如何通过预处理方法对数据做隐私保护？ 源程序 详见附件 ***BigData\_Lab1.ipynb*** 或 ***BigData\_Lab1.html***。 | | | | | | |
| 四、实验结果及分析和（或）源程序调试过程 数据清洗 根据三中提到的步骤，数据清洗后，数据详情如图 9所示，其中除了公司成立时间只有6971条外（原因详见第三部分），其他数据均为7483条，示例数据如图 10所示。详细数据详见附件 ***BigData-Lab1-数据清洗后的数据.csv*** 。    图 9 数据清洗后的详情数据    图 10 数据清洗后的示例数据 数据可视化数据分布分析 (1) 职务等级分布：主要以C（总监|副总监|经理|副经理|主任|主管）为主，A（董事|主席）最少，符合现实情况。    图 11 职务等级分布  (2)工作类别分布：除了其他类别，占比比较大的有市场类（销售|市场|客户）和技术类（业务|技术|项目），分别达到18.99%和6.70%。    图 12 工作类别分布  (3) 公司领域分布：除了其他类别，占比比较大的有科技类（科技|软件|信息技术），符合认知，且能证明(2)工作类别分布的合理性。    图 13 公司领域分布  (4) 资金等级分布：67.89%的公司的注册资金在1000万以下，但是异于常识的是，注册1亿以上的公司，占了10.34%。    图 14 资金等级分布  (5) 公司年龄等级分布：可以看到，约44.97%的公司都属于新起之秀，而开了超过20年的老公司，只占了4.37%。且5年老店和10年老店占比适中。    图 15 公司年龄等级分布 相关分析 (1)用户认证和公司认证的关系  如图 16所示，具体(1,1)占比最多，即用户认证且公司认证的占了75.49%。根据表 1的分析，可以知道，用户认证的数量与公司认证的数量呈正相关的关系。  表 1 用户认证和公司认证的关系   |  |  |  | | --- | --- | --- | |  | 公司认证 | 公司未认证 | | 用户认证 | 5649 | 24 | | 用户未认证 | 868 | 942 |     图 16 用户认证和公司认证的关系  通过卡方测定，否定用户认证和公司认证无关的概率为0.0.故用户认证和公司认证具有极强的相关性。  (2)公司年龄和公司注册资金的关系  如图 17所示，除了个别离群点之外，公司年龄和注册资金之间没有太明显的关联关系。    图 17 公司年龄（年）和注册资金（元）之间的关系  通过QQ图以及Pearson相关矩阵可以得到，公司年龄和注册资金仅存在弱正相关。 对比分析 ①认证用户和非认证用户职务分布情况的差异分析  如图 18所示，认证用户相对于非认证用户来说，职务分布更集中在C类，且其他类别的职务数量均比非认证用户多。    图 18 认证用户和非认证用户职务分布情况  ②分析认证公司和非认证公司注册资金差异  比①更明显的是，已认证的公司注册资金平均比未认证的公司高，且分布数量也远大于未认证的公司，离群值比未认证的公司更少。    图 19 认证公司和非认证公司注册资金（元）情况-直方图分析    图 20 认证公司和非认证公司注册资金（元）情况-盒图分析  ③分析认证公司和非认证公司年龄差异  已认证的公司平均年龄大于未认证的公司，离群值比未认证的公司更少，且诸如20年以上的老店，均为已认证公司。    图 21 认证公司和非认证公司公司年龄（年）情况-直方图分析    图 22 认证公司和非认证公司公司年龄（年）情况-盒图分析 思考用户隐私保护的思考 （1）对分析结果影响不大的数据（比如姓名）：可以直接采用md5加密（无法解密的）或者也可以选择hash以及字符映射等等方式。 （2）对分析结果有一定影响的数据（比如企业中个人的工资）：可以采用取数异或，或者用一定函数计算进行加密（能够解密的），这样既保护了数据，也对结果不会产生影响。 | | | | | | |

备注：

1、教师在布置需撰写实验报告的实验前，应先将报告书上的“实验题目”、“实验性质”、“实验目的”、“实验项目内容”等项目填写完成，然后再下发给学生。

2、教师在布置需撰写报告的实验项目时，应告知学生提交实验报告的最后期限。

3、学生应按照要求正确地撰写实验报告：

* 1. 在实验报告上正确地填写“实验时间”、“实验地点”等栏目。
  2. 将实验所涉及的源程序文件内容（实验操作步骤或者算法）填写在“实验过程或算法（源程序）”栏目中。
  3. 将实验所涉及源程序调试过程（输入数据和输出结果）或者实验的分析内容填写在“实验结果及分析和（或）源程序调试过程”栏目中。
  4. 在实验报告页脚的“报告创建时间：”处插入完成实验报告时的日期和时间。
  5. 学生将每个实验完成后，按实验要求的文件名通过网络提交（上载）到指定的服务器所规定的共享文件夹中。每个实验一个电子文档，如果实验中有多个电子文档（如源程序或图形等），则用WinRAR压缩成一个压缩包文档提交，压缩包文件名同实验报告文件名（见下条）。
  6. 提交的实验报告电子文档命名为：“年级（两位数字不要“级”字）专业（缩写：计算机科学与技术专业（计科）、网络工程专业（网络）、信息安全专业（信息）、物联网工程（物联网））班级（两位数字）学号（八位数字）姓名实验序号（一位数字）．doc。如学号为20115676、年级为2011级、专业为“计算机科学与技术”专业、班级为“02班”、姓名为“王宇”的学生，完成的第一次实验命名为： 11计科02班20115676王宇1．Doc，以后几次实验的报告名称以此类推。

4、教师（或助教）在评价学生实验时，应根据其提交的其他实验相关资料（例如源程序文件等）对实验报告进行仔细评价。评价后应完成的项目有:

1. 在“成绩”栏中填写实验成绩。每个项目的实验成绩按照五级制（优、良、中、及格、不及格）方式评分，实验总成绩则通过计算每个项目得分的平均值获得（平均值计算时需将五级制转换为百分制优=95、良=85、中=75、及格=65、不及格=55）。
2. 在“教师评价”栏中用符号标注评价项目结果（用√表示正确，用×表示错误，用≈表示 半对半错）。
3. 在“教师评价”栏中“评价教师签名”填写评价教师（或助教）姓名。将评价后的实验报告转换为PDF格式文件归档。
4. 课程实验环节结束后，任课教师将自己教学班的实验报告文件夹进行清理。在提交文件夹中，文件总数为实验次数×教学班学生人数（如，教学班人数为90人，实验项目为5，其文件数为：90×5=450）。任课教师一定要认真清理，总数相符，否则学生该实验项目不能得分。最后将学生提交的实验报告刻光盘连同实验成绩一起放入试卷袋存档。