****

**综 合 实 验 报 告**

课程名称： 数据采集及预处理训练

专 业： 数据科学与大数据技术

班 级： 20大数据B2班

姓 名： 黄华弢

学 号： 202010620216

指导教师： 容达

工学院

2022年10月

|  |  |
| --- | --- |
| **实验项目名称** | 数据清洗 |
| **一、实验目的**  1、掌握Kettle工具常用控件的应用。  2、能够根据任务要求创建转换或作业逻辑。  3、理解数据的走向及产生数据结果的构成。 | |
| **二、实验内容和要求**  **1、实验内容**  （1）缺失值填充、异常值处理的源数据如图1、图2所示。  **注意：图2的源数据可以是excel、文本、数据库等文件。**    图1：缺失值数据 图2：修补异常值数据  （2）处理后的数据如图3、图4所示。    **2、实验要求**  （1）提交ktr、源数据、处理后数据文件；  （2）编辑实训报告（图片要清晰），文字排版规范、操作步骤明了。  （3）完成实验总结并签名（纸质版需要手写签名）。 | |
| **三、实验环境及材料**  1、硬件环境要求：  微型计算机一台（Pentium或以上）。  2、软件环境要求：  JDK、Kettle、MySql等。 | |
| **四、实验方法和步骤**  图一：  1、通过使用Kettle工具，创建一个转换，并添加“文本文件输入”控件、“过滤记录”控件、“空操作（什么也不做）”控件、“替换NULL值”控件、“合并记录”控件、“字段选择”控件以及Hop跳连接线。    2、双击“文本文件输入”控件，进入“文本文件输入”配置界面。单击【浏览】按钮，选择要去除缺失值的文件；单击【增加】按钮，将要去除缺失值的文件添加到“文本文件输入”控件中。    3、单击“内容”选项卡；在清除分隔符处的默认分隔符“；”，单击【Insert TAB】按钮，在分隔符处插入一个制表符；取消勾选“头部”复选框。    4、单击“字段”选项卡；根据文件的内容添加对应的字段名称，并指定数据类型。    5、单击【预览记录】按钮，查看文件的数据是否成功抽取到文本文件输入流中。    6、双击“过滤记录”控件，进入“过滤记录”配置界面。在“条件”处设置过滤的条件，将过滤字段设置为workclass、过滤值设置为Private作为过滤条件。在“发送true数据给步骤：”下拉框中选择“空操作”，将workclass字段值为Private的数据放在“空操作”控件中；在“发送false数据给步骤：”下拉框中选择“空操作（什么也不做）2”，将workclass字段值不为Private的数据放在“空操作（什么也不做）2”控件中。    7、双击“替换NULL值”控件，进入“替换NULL值”配置界面。勾选“选择字段”处的复选框，并在“字段”框添加字段为hours\_per\_week，值替换为44    8、双击“合并记录”控件，进入“合并行（比较）”配置界面。在”旧数据源：”下拉框选择“替换NULL值”，“新数据源：”下拉框选择“空操作（什么也不做）2”；在“匹配的关键字：”部分，添加关键字段，即userid。    9、双击“替换NULL值2”控件，进入“替换NULL值”配置界面。勾选“选择字段”处的复选框，并在“字段”框添加字段为workclass，值替换为Private     10、双击“字段选择”控件，进入“选择/改名值”配置界面。双击“字段选择”控件，进入“选择/改名值”配置界面。    11.启动  图二：  1.通过使用[Kettle](https://so.csdn.net/so/search?q=Kettle&spm=1001.2101.3001.7020)工具，创建一个转换，并添加“文本文件输入”控件、“过滤记录”控件、“空操作（什么也不做）”控件、“设置值为NULL”控件、“合并记录”控件、“替换NULL值”控件、字段选择控件以及Hop跳连接线。    2.双击“文本文件输入”控件，进入“文本文件输入”配置界面。添加文件    3、单击“内容”选项卡；在清除分隔符处的默认分隔符“；”，单击【Insert TAB】按钮，在分隔符处插入一个制表符；取消勾选“头部”复选框。    4. 单击“字段”选项卡；根据文件的内容添加对应的字段名称，并指定数据类型。    5. 单击【预览记录】按钮，查看文件的数据是否成功抽取到文本文件输入流中。    6.双击“过滤记录”控件，进入“过滤记录”配置界面。在“条件”处设置过滤的条件，即设置Height字段的取值范围，从而判断数据表中的每个数据是否为异常值。若是在非异常值的取值范围内，则是非异常值，否则是异常值。    7. 双击“设置值为NULL”控件，进入“设置值为NULL”界面；在“字段”处添加要设为NULL值的字段名称和值。    8.双击“合并记录”控件，进入“合并行（比较）”界面。在“旧数据源：”处的下拉框选择“设置为NULL值”，“新数据源：”处的下拉框选择“空操作（什么也不做）2”；在“匹配的关键字：”处，添加关键字段，即id。    9. 双击“替换NULL值”控件，进入“替换NULL值”界面。勾选“选择字段”处的复选框，并在“字段”框添加字段为Hight，值替换为170    10. .双击“字段选择”控件，进入“选择/改名值”界面。在“移除”选项卡处添加要移除的字段名称，这里移除的是字段flagfield。    11.添加” 文本文件输出”控件，并填入输出目录。 | |
| **五、实验结果**  图一：  Ktr：      源数据：      处理后数据文件：      **图二：**  Ktr：      源数据：      处理后数据文件： | |
| **六、实验总结**  通过本次实验，我掌握Kettle工具常用控件的应用，熟练并能够根据任务要求创建转换或作业逻辑，清晰并准确的理解数据的走向及产生数据结果的构成。通过对数据缺失值的填充、数据异常值的处理，我对《数据清洗》这门课程有了更加深入的学习与理解。通过使用Kettle工具进行实验，我对“文本文件输入”控件、“过滤记录”控件、“空操作（什么也不做）”控件、“替换NULL值”控件、“合并记录”控件、“字段选择”控件以及Hop跳连接线、 “合并记录”控件等控件有了更加深入的学习，也对数据“缺失值的填充”和“数据异常值的处理  ”的运行流程有了深入的了解。  手写签名： | |
| **批阅教师评语：**  评定等级： 批阅教师签名： 年 月 日 | |