数据清洗的概念及实操

张宁
武汉大学图书馆

一个工具和案例

https://openrefine.org/

https://libguides.lib.whu.edu.cn/filesdown

数据分析步骤

4、美观，有意义
不能喧宾夺主

3、前人经验与自身经验的积累
创新性的核心环节

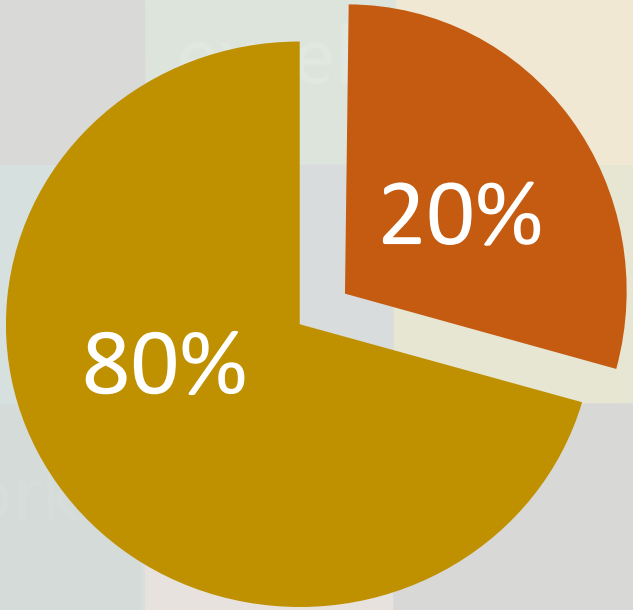
2、认清脏数据的危害
关注重复值、异常值与空缺值

1、可靠的数据源
灵活的获取方式

目录

无处不在的二八定律

今天的培训中：“废话”和基本概念。实际操作。



学习时间分配中：系统学习基础知识和技能。处理实际问题中的摸索。

数据处理中：前期预处理上花费的时间和精力。包括获取数据，理解数据，处理数据。后期探索分析，统计、预测

数据.....分析.....展示

原始数据.....干净数据.....分析.....展示

数据清洗的必要性

假设所有的数据都是不干净的。
Garbage in，garbage out.

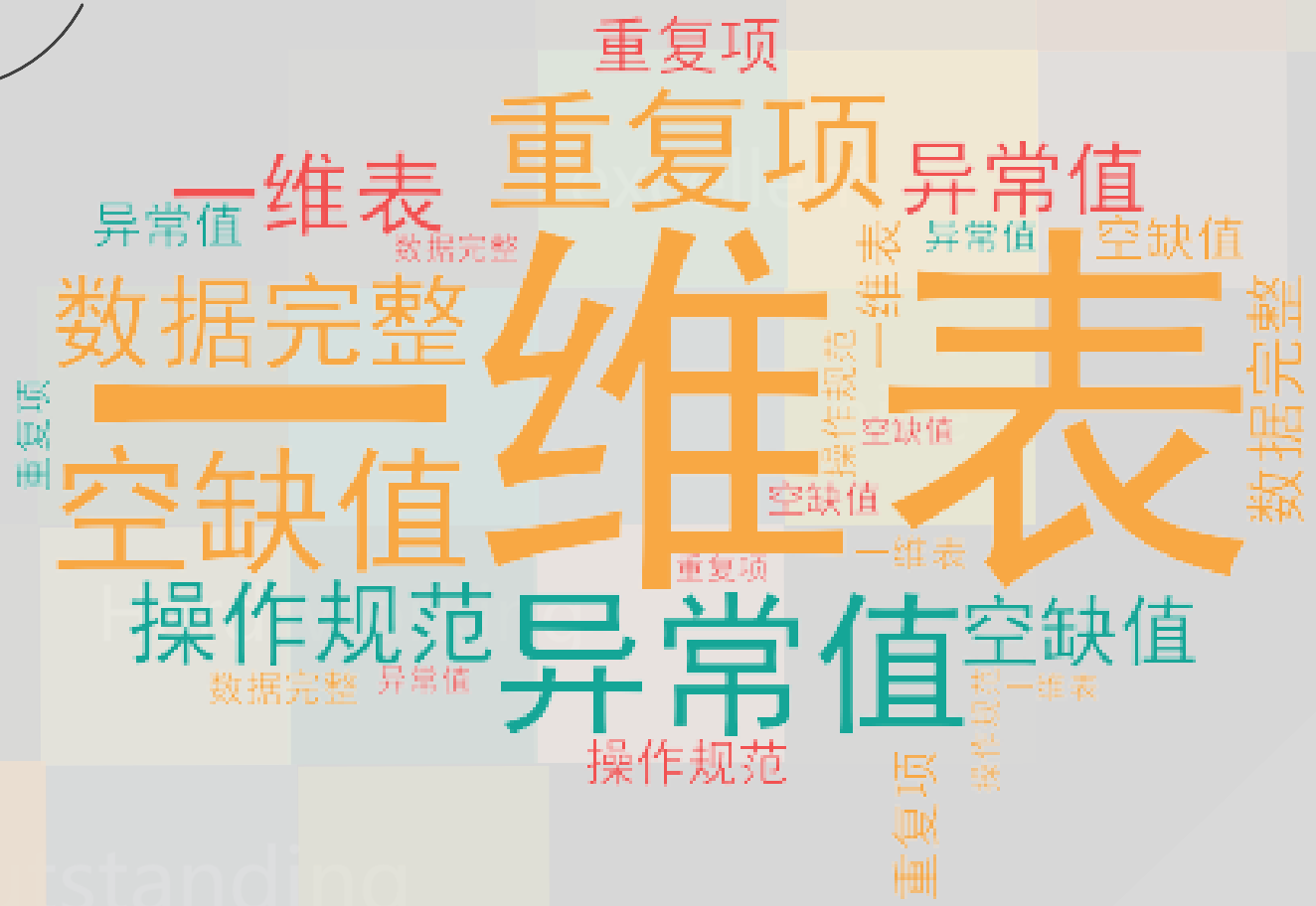
表结构的不合理，字段结构不合理。
人为操作：输入错误，缺乏审校。
前期清洗过程中不合理的操作：多数据源合并冲突，规则改变或者遗漏，理解偏差制定了不合理的规则。

错误的结果导致错误的决策。
数据客户声誉受损。
分析师的职业声誉受损。

数据质量的考量维度

准确性：描述是否与客观实体的特征相符。
合理性：字段结构设置，字段类型的设置等。
规范性：字段填写是否符合规范。
一致性：同一值对应不同字段是否一致。
重复性：是否可以出现重复的信息。
及时性：数据的产生和更新是否及时。
完整性：是否缺失记录或者字段。

数据处理的基础认知


重复项
重复项
一维表
异常值
异常值
数据完整
异常值
空缺值
数据完整
一维表
空缺值
空缺值
空缺值
操作规范
异常值
空缺值
操作规范

一维表

工作表数据区域的顶端行为字段名称（标题），以后各行为数据（记录），并且各列只包含一个类型数据的数据区域。每一列是否是一个独立的参数。

城市	性别	补助金额
武汉	女	120
武汉	男	108
广州	女	105
广州	男	100

城市	男	女
武汉	108	120
广州	100	105
天津	102	110

实操

1 数据属性

表名称:
来源:
采集时间:
多少字段（多少列）:
多少行:

备份原始数据！！！！

2 构造一维表，并做全局校验。

用google-refine跟着老师按步骤处理一个数据表。

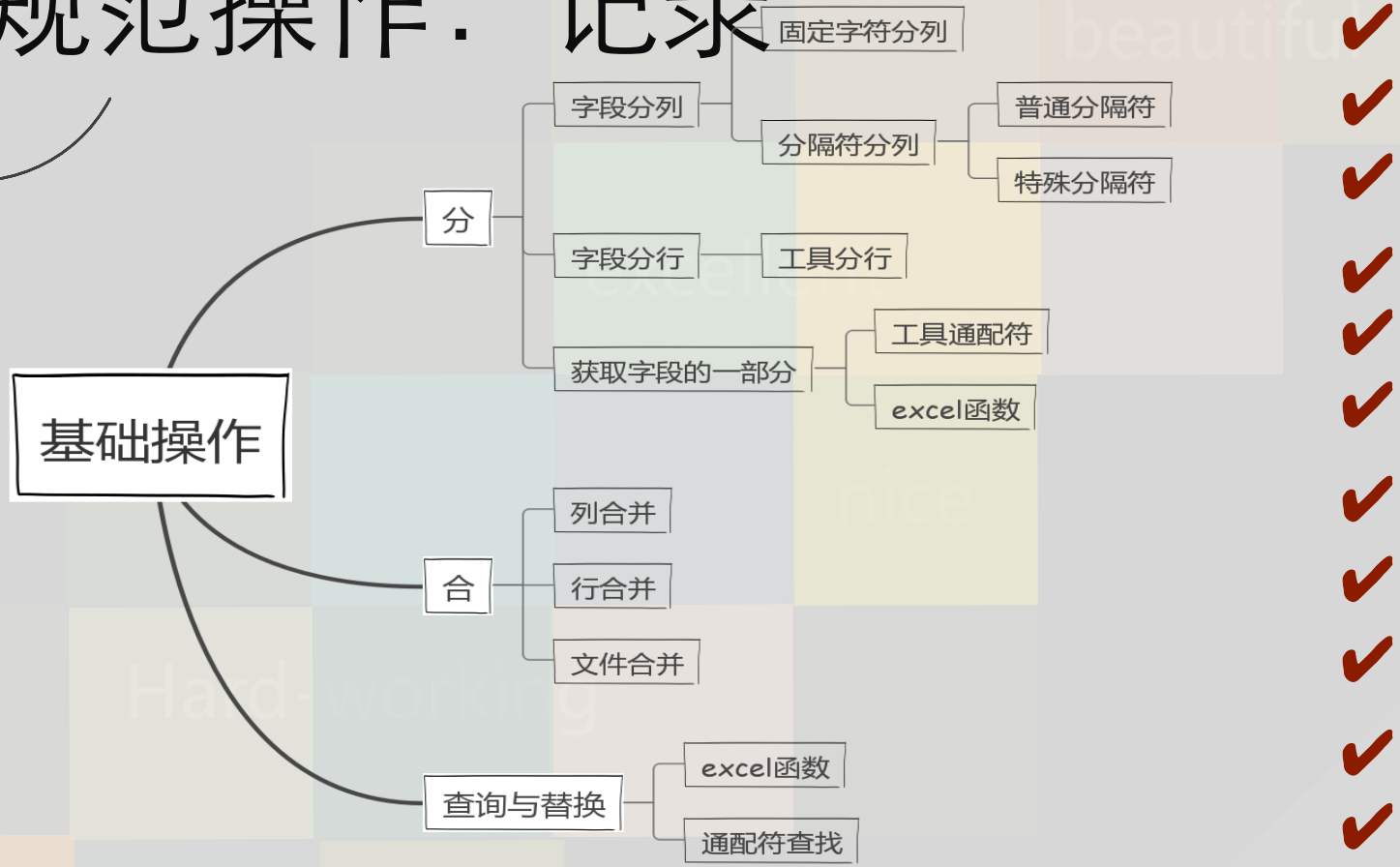
3 保留需要字段，研究其定义、类型及质量考量维度。

字段名称:
定义:
字段类型:
可否为空?
可否重复?
是否有变量范围?

4 对保留字段分别做细节校验。

值的范围及统计
空缺值
重复值
离群值

规范操作：记录


基础操作
分
字段分列
固定字符分列
分隔符分列
普通分隔符
特殊分隔符
字段分行
工具分行
获取字段的一部分
工具通配符
excel函数
合
列合并
行合并
文件合并
查询与替换
excel函数
通配符查找

案例

表1是三个省份三个城市一些家庭的信息，具体见表。
表2是电信登记的包含表1中家庭成员中个人身份证名下的手机号码信息。
三个城市分别要对自己的市民分男女发放某种补助，表3是补助金额表。
请根据三张表格数据回答下面问题：

家庭编号	省市	地址	家庭成员
1	湖北/武汉	西安外国语大学经济金融学院, Phone:	刘杰\女\37458419391118****谢
2	湖北/武汉	西安建筑科技大学环境与市政工程学院, 中国环境科学研究院湖泊环境	董艳慧\女\41300319540510****

身份证号码	手机
17701019960413****	13766405722
20657320090123****	
21015719340709****	13296745224
30863819560616****	13688391918

城市	男	女
武汉	108	120
广州	100	105
天津	102	110

说明：案例信息均为虚构， 没有实际意义。

案例解析

1、在表1的家庭信息表的“地址”字段中，包含了每个家庭的家庭电话信息。请统计出哪个城市的家庭电话数量最多?

城市的家庭电话数量 / 表1 / 字段中数据提取 / 家庭信息

2、武汉市补贴总额是多少?

城市的补贴金额-城市、人数、男女情况 / 表1&表2 / 段分离及表链接-字段匹配 / 个人信息

3、在家庭信息表中找出年龄最大的人的所有信息，包括姓名、性别、身份证号码、手机、家庭编号、所在省、所在城市、家庭地址、家庭电话及补贴金额。

个人信息 / 表1&表2 / 行分离 、表链接-字段匹配 / 个人信息

明确目的

1、构造一个家庭信息表格

2、构造一个个人信息表格

3、用数据透视表统计数据

熟悉数据

1、读表，理解字段，发现构造表格与现有表格的关系-**建立一维表**

2、重复数据/重复字段检测-**评估重复值的合理性及对结果的影响，决定处理**

3、空缺数据/空缺字段检测-**评估空缺值的合理性及对结果的影响，决定处理**

4、异常值检测-**是否需要关注？异常值是错误的还是关注点？结果可解释**

技术实现

1、复制数据

2、对表1省市字段分离
3、将地址字段中电话号码分离 ----家庭信息表构造成功

4、表1中家庭成员字段分行，并填充，构造新表
5、新表家庭成员字段分列，创建姓名、性别和身份证字段
6、新表用身份证链接表2，创建手机字段信息
7、新表中合并城市与性别成为新字段，表3相同步骤
8、新表与表3中新字段查询匹配， 创建对应补助金额字段----个人信息表构造成功

9、数据透视表完成统计信息
10、对个人信息表中生日信息排序，完成问题。

数据清洗步骤

推荐图书

赫夫. 统计数字会撒谎[M]. 北京 : 中国城市出版社, 2009

单册状态	应还日期/催还应还日期	应还时间	馆藏位置	馆藏地	索书号
保存本	在架上		总馆中文阅览区C1-C4		C8-49/H327
外借书	在架上		总馆中文图书借阅A2-A5		C8-49/H327
外借书	在架上		总馆中文图书借阅A2-A5		C8-49/H327

吕峻闽. 数据可视化分析 : Excel 2016+Tableau[M]. 北京 : 电子工业出版社, 2017

单册状态	应还日期/催还应还日期	应还时间	馆藏位置	馆藏地	索书号
保存本	在架上		信息馆中文阅览4楼西		TP391.13/L788
已借出	20220602	22:30	信息馆借阅区2楼东		TP391.13/L788

谢谢反馈