工作任务三：数据清洗

（一）识别数据

常见的数据问题有：

（1）**数据值问题**：数据值缺失、缺失数据值被0补位、明显的数据不完整、出现不合理数据值；（2）**格式问题**：拼写不一致、姓名顺序不一致、日期格式不一致、未标明数据值单位、文本被　转为数字、数字被储存为文本；

（3）**其他问题**：打字错误或乱码、分类不合理、字段命名含糊不清、数据表有65536行（数据溢出）等。

（一）数据清洗

1、常见的数据清洗方式有：

（1）**手动编辑**：只适合数据量很小的情况；

（2）**使用电子表格**：如使用Excel表格进行数值处理和函数计算；

（3）**使用工具修正格式**：如使用Open Refine等可轻松地修正数据格式问题；

（4）**使用计算机语言相关软件**：如使用Python、R语言等进行数据处理。

（一）数据清洗

2、使用open refine进行数据清洗

（1）安装open refine软件

（2）导入数据文件

（3）重命名标题栏

（4）统一数据格式

（5）修改不合理数据值

（6）对不不合理数据重新分类

（7）删除空值或缺失值

（8）导出数据文件

思政融入：

自2018年3月以来，《新京报》推出了“有理数”新栏目，进一步拓展了数据新闻报道的样式和种类。但是有观点认为，长图形式虽然直观，但受限于篇幅和图像质量，难以全面呈现所有背景文字，从而导致信息内容的部分缺失。请谈一谈在完成数据清洗时，我们应具备什么意识？

思政元素：问题意识、科学求实

谢谢观看