1. 数据清理任务：

* 填写空缺值
* 识别离群点和平滑噪音
* 纠正不一致的数据
* 解决数据集成造成的冗余

1. 处理空缺值：

* 忽略元组
* 人工填写空缺值
* 使用全局变量填充空缺值
* 使用属性的平均值填充空缺值
* 使用与给定元组属同一类的所有样本的平均值
* 使用最可能的值填充空缺值：使用像Bayesian公式或判定树这样的基于推断的方法

1. 处理噪声数据

* 分箱：分箱方法通过考察数据的近邻来光滑有序数据的值
* 聚类：监测并且去除孤立点
* 回归：通过让数据适应回归函数来平滑数据
* 计算机和人工检查结合：计算机检测可疑数据，然后对它们进行人工判断。

1. 数据清理过程：

* 偏差检测
* 数据变换
* 偏差检测和数据变换迭代执行

1. 处理数据集成中的冗余数据
   1. 相关性分析检测
   2. X2测试
2. 数据变换
3. 平滑：去除数据中的噪声
4. 聚集：汇总，数据立方体的构建
5. 数据概化：沿概念分层向上概化
6. 规范化：将数据按比例缩放，使之落入一定小的特定的区间。
7. 属性构造：通过现有属性构造新的属性，并添加到属性集中，以增加对高维数据的结构的理解和精确度。
8. 规范化：
9. 最小-最大规范化
10. Z-score规范化
11. 小数定标规范化