Data Cleaning

数据清理的作用是去除不属于该数据集的数据，因为

- 数据缺失
- Tables需要Merge或Join
- 格式不一致

数据类型错误思路：

- 对于时间日期相关，现成库pandas.to_datetime
- 类型错误，直接使用df[”col”].astype(someType)进行转换
- 更加复杂的问题，使用自定义的函数进行df[”col”].apply

数据合并思路：

- 确认所有类型是匹配的（每一列数据类型一致，列名一致）
- 解决数据冲突问题，并选择合适的Join方式（Inner Join, Outer Join）

标签更新：将一个标签类型数量增加（将good, bad变成very good, good, very bad, bad），或中途改变记录方式（之前数据会空缺），解决思路如下：

- 将数据集分成两份，保持数据一致性
- 根据已有数据推测旧数据

重复记录：

- 完全一样的重复行可以用df.drop_duplicates()
- 若有细微差别则需要找到真值并去除重复

离群值：

- **离群值检测**：通过Z-Scores来检测离群值，通常离群值的Z值小于-2或大于2，这意味这些值偏离均值较远
- **离群值处理**：根据goal考虑离群值是否能够移除，可以思考
  - 偶然性
  - 评估对模型的影响
  - 和数据核心内容的相关性

缺失数据：

- **Data Missing Completely at Random (MCAR)**：数据缺失没有任何的模式，所有内容有相同概率缺失（比如磁盘损坏）
- **Data Missing at Random (MAR)**：缺失与观察变量有关，而与缺失的变量本身无关（比如在健康问卷，男性有更大概率不回答问题，这和问题本身无关而是和性别有关）
- **Data Missing not at Random (MNAR)**：缺失和变量本身有关（比如低GPA学生不太可能在问卷中写自己的GPA）

解决数据缺失思路：

- MCAR/MAR：
  - 标签数据：可以将缺失变成一个新的标签
  - 对于数值数据，如果缺失值较少（1%），则可以直接Drop。**Listwise Deletion**将任何有缺失的行直接移除，**Pairwise Deletion**只将要分析的数据缺失的行移除。
  - **Imputation**：用估计值来替代缺失值，如均值（通常用于数值型，且无偏）、中位数（通常用于数值型，且有偏），Hot Imputation（从同一个数据集中选值填充），Cold Imputation（从不同数据集中选值填充），Mode Imputation（选择最常出现的值填充，通常用于字符），Hot-deck Imputation（找相近行，用该行的值填充），Bayesian Imputation（概率推断），Multiple Imputation（多次预测值，并填入多次的统计数据，如均值）
  - 注：当不超过5%数据缺失时可以使用Mean, Median, Mode Imputation
- MNAR：

    - 最糟糕的情况，因为缺失基于数据本身，因此存在系统性的偏差，存在一下几种类型：
        - 数据在连续的行或时间段内缺失：Imputation
        - 数据由于值本身而缺失：做一些Sensitivity Analysis（用不同假设解释缺失数据是什么）；构建不同的包含或不包含缺失数据的模型来理解缺失数据的影响
        - 数据因为超过了值域限制而缺失：去除值域外的值；推断缺失值的分布；收集更多数据

错误数据检查：

- 找到Attractors（异常尖峰）；Discontinuities（猛烈变化）值；不合理的模式；数据超过合理区间