# Part 1：

1、找到相对应的样本记录的方法： 首先把first\_name、middle\_name、last\_name三个字段合并为一个字段name然后根据相同ssn、name还有gender判断是不是用一个样本，如果这三列相同，则判定为是同一个records。

2、emp-edu 中有4075条记录是其独有的，medical中的ssn全部在emp-edu中出现，我们将这部分记录直接删除，避免合并数据后出现缺失值。

3、数据存存在1498条重复records，对于重复records需要删除，已经做了删除处理。

# Part 2：

1、存在缺失值，共有14个指标存在缺失数据，缺失数据的个数为49211个。因为此数据集为个人信息数据，无法使用0或者均值等数值方法来代替缺失值,所以使用“-1" 来填充缺失数据。

2、数据中的phone字段，存在一些非法的电话号码，比如邮箱等。对于这类数据，需要替换为缺失数据，并用“-1”代替。还有bmi的值存在负数，我们知道，bmi的值肯定是正数， 不可能是负数。所以需要把bmi的值为负数的部分替换为“-1”，数据类型为字符串，不可参与数学运算。

# Part3

1、汇总的数据里面包含phone\_med和phone两列，分别来源于emp-edu.csv和medical.csv， 对于同一个人的电话号码存在不一致的情况，我们把phone缺失的号码用phone\_med 来代替，把phone\_med中缺失的号码用phone来代替。如果phone\_med和phone都存在的情况，我们保留两个电话号码，并用竖线“|”分割开来。这样做的好处是保证了数据不丢失。另外还有street\_address\_med suburb\_med postcode\_med state\_med直接删除，因为这几个变量是具有实际意义的变量，一般不会有变化。

2、考虑教育和职业与健康的关系，我们建议删除无关变量。比如street\_address、suburb、postcode、credit\_card\_number。