数据清理总结

1、缺失值的产生原因是什么?

缺失值是指数据字段没有值，缺失值的产生主要是由于测量设备或人为因素所造成的数据遗漏。在 pandas 中，我们使用 NaN 表示缺失值。

2、如何处理数据表中的缺失值?

1）检查缺失值

通过 info 查看数据集的简明摘要信息来检查缺失值。例如练习中的代码：

data_cleaning.info()

2）缺失值处理

① 对于**整列缺失**的情况，直接在原表中删除缺失列，例如练习中的代码：

data_cleaning.drop('爱好',axis=1,inplace=True)

② 字段中有**个别缺失值**

从业务或建模的角度判断，如果这个字段没有加入建模的必要性，就删除数据，即直接删除缺失数据对应的所有字段（方法同①）。否则，可以选择删除缺失值所在的记录行，或者进行数据填补。

删除缺失记录，如练习中的代码：

data_cleaning.dropna(inplace=True)

缺失值填充，可以填充均值、常数或估计值。如“联系-指数年之差分析”中将'Assoc. Index Year'一列 16 个空值填充“0”值的代码：

politic_relation['Assoc. Index Year'].fillna(0)

3、数据重复产生的原因是什么?

在数据集中常常会出现重复行，这主要是由于系统或者人工记录的问题导致。

4、如何处理数据表中的重复数据?

1）检查重复情况

查看重复一般分为整行重复查看和部分字段重复查看，但都是使用 duplicated()和 sum()检查数据的重复情况，只不过部分重复查看需要在 duplicated 后的括号内放上需要查看的字段。例如练习中的代码：

data_cleaning.duplicated('学号').sum()

2）删除重复行

① 对于整行重复的直接使用 drop()进行删除字段，例如练习中的代码：

data_cleaning.drop_duplicates(inplace=True)

注意：在原表中删除需要设置参数 inplace=True

② 如果想删除部分字段重复的数据行，比如练习中'学号'重复的数据行，将'学号'添加进去：

data_cleaning.drop_duplicates(['学号'],inplace=True)