数据拆合总结

1、什么是数据拆解？

数据拆解是由于原始数据字段隐含信息量太大，计算机不易理解，可以从多维度、结构化和流程化的进行数据字段拆解，然后形成新增数据字段，如出生地拆解为区域和一二三线城市两个维度以便分析。

编号	出生地
1	北京
2	上海
3	济南
4	青岛
5	重庆
6	厦门
7	宁波

编号	区域	级别
1	华北	一线
2	华东	一线
3	华东	二线
4	华东	二线
5	西南	一线
6	华北	二线
7	华北	二线

2、如何通过数据合并生成案例中的官员政治关系数和亲属为官数的数据表？

1）获取官员政治关系数

①针对政治表中的“联系”字段，需要进行字段内的数据合并，对每个官员的 6 种不同类型的政治关系分别计数，产生 6 个对应的新字段

②行索引转换为列索引

上述结果中官员的政治关系为行索引，使用 unstack() 将其转化为列索引：

politic_relation['联系'].groupby(politic_relation['姓名']).value_counts().unstack()

③对这些数据中的空值进行填充“0”值

部分官员的政治关系为空值，空值代表没有，所以使用 fillna() 对上述结果进行填充“0”值，例如案例中的代码：

politic_relation['联系'].groupby(politic_relation['姓名']).value_counts().unstack().fillna(0)

2）获取每位官员的亲属为官数

①通过官员集合与亲戚集合相交，得到同时是官员的亲戚

②在亲戚表中筛选出这些亲戚所在的行，将亲戚按照官员姓名分组并计算其亲戚个数，从而获取每位官员的亲戚为官数

3）将两组数据进行合并

①在第一组数据中增加一列为“亲戚为官数”，将第二组数据与其进行合并

②对合并后结果中亲戚为官数一列填充“0”值