**二、基本数据情况统计（构成以及数据分类）**

**（一）数据库中重要的要素：**

1、**任职地、任职职位、任职职位简写**

注：任职地、任职职位简写都是根据“任职职位”提炼出来的，有的提炼不出来，所以是空的，这个情况只会在第二部分履历信息中会出现。对于空值就只能空着了。

2、官职相关：**官职类别、地区、品级、官职名称**（“官职类别”和“官职名称”不同）

3、基本信息：**姓名、民族、旗分、科举**

4、分类标准

（1）科举级别，分类依据如下：

进士（进士、翻译进士、武进士）；

举人（举人、翻译举人、武举人）；

举人以下（其他非空的类型，如贡生、监生等等）

（2）科举性质，分类依据如下：

普通科举：进士、举人、武进士

翻译科举：翻译进士、翻译举人、 翻译生员

其他：监生、贡生、官学生等等

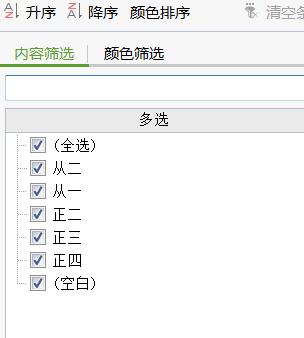
**（二）基本数据情况统计的目的：**

目的主要是搞清楚数据库各个重要要素的基本构成情况，例如，某一列的数据包括几个类别，所有类的名称需要列出来；某些列的数据总数、空值、非空值有多少。

**1、数据类别的统计：**

（1）目的：这一部分就是对数据本身的一个统计。即每一列下共有多少类数据，类似于excel表中“筛选”的功能。

例如，“品级”这一项目，通过筛选如下图：则需要统计和列出的“类别数”为6，分别为“从一、正二、从二、正三、正四、空白”共6类，“类别”就是将具体的类别列出。（因为excel表中的“筛选”功能只能将各个类别的名称列出，不能统计出共有几类，所以需要用算法将类别数和名称分别列出）



最终统计的结果为：（“类别数”只是举例用）

|  |  |  |  |
| --- | --- | --- | --- |
| 项目 | 地区 | 官职类别 | 品级 |
| 类别数 | 10 | 9 | 6 |
| 类别 | 盛京 | 将军 | 从一 |
|  | 吉林 | 副都统 | 正二 |
|  | 黑龙江 | 府尹 | 从二 |
|  | 齐齐哈尔 | 五部侍郎 | …… |
|  | …… | …… |  |

（2）所需要的结果：

需要分别统计“官职名称、地区、官职类别、品级、科举”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 项目 | 官职名称 | 地区 | 官职类别 | 品级 | 科举 |
| （需要统计的内容1） | 类别数 | 类别数 | 类别数 | 类别数 | 类别数 |
| （需要统计的内容2） | 类别 | 类别 | 类别 | 类别 | 类别 |

**2、官职、姓名重复的情况：**

注：官职、姓名重复的情况指的是同一人，而非重名者。

（1）目的：需要统计出“姓名”以及在不同限制条件下出现的次数，目的在于考虑到整个数据库中姓名重复的情况、同一“官职名称”内姓名重复的情况。

（2）需要的结果：“条件”中各项出现1次、2次……的数量和比例。如下图：

|  |  |  |  |
| --- | --- | --- | --- |
| 条件 | ①“姓名” | **②“官职名称”+“姓名”** | ③姓名相同，“官职名称”不相同（去掉官职名称内相同的情况） |
| 1次（唯一） |  |  |  |
| 2次 |  |  |  |
| 3次 |  |  |  |
| 4次 |  |  |  |

①“姓名”：即一个人在此数据库中出现的次数统计

②“官职名称”+“姓名”：即同一官职内，一个人出现的次数统计

③ 姓名相同，“官职名称”不相同：即官员共任几个职位，要考虑到“官职名称”内重复的情况。在同一“官职名称”中出现多次的情况算作1，如果某人在职位1中出现2次，在职位2中出现1次，那么此项姓名相同、“官职名称”不相同为2次，即去掉“官职名称”中重复出现的次数。

**“官职名称”+“姓名”唯一是接下来三中（一）（二）分析的前提。**

**3、履历情况统计**

（1）目的：统计官员的履历情况

（2）需要的结果：各种情况出现的数量和比例

大前提：**“官职名称”+“姓名”唯一，即去掉重复的情况**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 条件 | 有履历 | | 无履历 | | 履历不正确 | |
| 数量 | 比例 | 数量 | 比例 | 数量 | 比例 |
| 官职类别1 |  |  |  |  |  |  |
| 官职类别2 |  |  |  |  |  |  |
| …… |  |  |  |  |  |  |

“有履历”是指“姓名内编号”=“1”

“无履历”是指“姓名内编号”=“0”

“履历不正确”是指“任职时间”-“生年”小于0或者大于90的情况