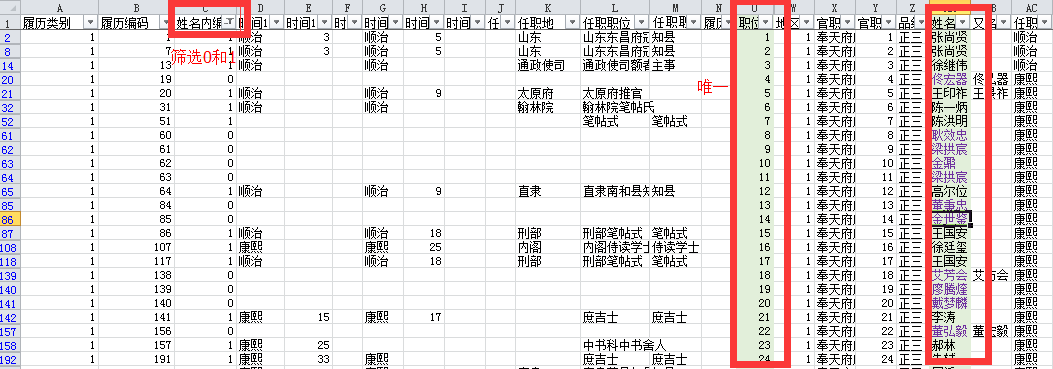
**数据处理方案**

1. **数据处理前的加工**
2. 核对各项数据，是否符合数据库分析处理的格式，然后集中修改。
3. 将其中的纪年方式改为公元纪年，如“乾隆四十（年）”改为“1775（年）”。注：可以提供清朝纪年对应表，编写程序进行替代。
4. 数据库说明：
5. 数据库的数据分为履历信息和基本信息两部分。列O“职位总序号”之前的为官员的履历信息。列O之后为官员的基本信息。



1. 列C“姓名内编号”，如果官员无履历信息则编号为0，如果有履历信息为1-n,如上图中1、2……6则表示某人共有6条履历信息，但是列O “职位总序号”都为1，说明这是同一人。所以只要筛选列C“姓名内编号”=0和1，就保留了所有人唯一的基本信息，就能将履历信息隐藏掉。如下图。



（3）其他各列的解释见附件“ 履历数据库数据说明”。

1. **数据处理方案**

对比要素：

1. 履历信息中：任职地、任职职位
2. 官职：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
4. 分类前提：

（1）科举级别

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

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

**第一部分：官员的基本信息。**

由于第一部分的统计只为基本信息中数量和比例的统计，所以要考虑同一官员多次出现的情况。需要排除同一官职内多次任职的情况（“官职类别”+“姓名” 唯一）。所以，在进行（一）（二）统计前，需要对数据本身进行特征分析。

1. 基本数据情况。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 项目 | 履历信息 | 官职名称 | 地区 | 官职类别 | 品级 | 科举 |
| （需要统计的内容1） | 数据量 | 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 |

例如“品级”，类别数为6，类别为从一、正二、从二、正三、正四。

输出的结果出了类别的数量，还有类别的名称，类似于excel的“筛选”功能。

1. 不考虑履历信息的职官表。

不需要考虑履历信息。所以列C“姓名内编码”为0和1，即只需要筛选出职官表就可以了。筛选后一共1980条

1. 重复的情况。

|  |  |  |  |
| --- | --- | --- | --- |
| 项目 | “姓名” | **“官职名称”+“姓名”** | 姓名相同，“官职名称”不相同 |
| 1次（唯一） |  |  |  |
| 2次 |  |  |  |
| 3次 |  |  |  |
| 4次 |  |  |  |

其中，第3项 姓名相同，“官职名称”不相同 ，意思是**官员共任几个职位**，在同一职位中任职多次的情况算作1个职位，如果某人在职位1中出现2次，在职位2中出现1次，那么第3项姓名相同、“官职名称”不相同为2次，即去掉同一职位中重复出现的次数。

**“官职名称”+“姓名”唯一是接下来（一）（二）分析的前提。**

1. 履历情况

根据“官职类别”分别统计履历情况。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 项目 | 有履历 | | 无履历 | | 履历不正确 | |
| 官职类别 | 数量 | 比例 | 数量 | 比例 | 数量 | 比例 |
| 官职类别1 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

“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

1. **科举与入仕**

需要分析官员的基本信息，不需要考虑履历信息。所以列C“姓名内编码”为0和1，即只需要筛选出职官表就可以了。筛选后一共1980条。前提是**“官职名称”+“姓名”唯一。**

1、“官职名称”2、“地区”3、“官职类别”4、“民族”下普通科举与翻译科举的差异。如下图：

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 官职名称 | 进士 | | 举人 | | 武进士 | | 翻译进士 | | 翻译举人 | | 翻译生员 | | 其他 | | 空白 | |
| 数量 | 比例 | 数量 | 比例 | 数量 | 比例 | 数量 | 比例 |  |  |  |  |  |  |  |  |
| 官职1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 官职2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| …… |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

分类前提：

进士、举人、武进士；翻译进士、翻译举人、翻译生员；其他（其他类别非空值）、空白。

1. **科举与职位的静态分布**

同上，职位的静态分布需要分析官员的基本信息，不需要考虑履历信息。所以列C“姓名内编码”为0和1，即只需要筛选出职官表就可以了。筛选后一共1980条。前提是**“官职名称”+“姓名”唯一**

需要统计的内容有不同1、“官职名称”2、“地区”3、“官职类别”4、“民族”的“科举”分类下所具有的数量和比例。例如：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 官职名称 | 进士 | | 举人 | | 举人以下 | | 空白 | |
| 数量 | 比例 | 数量 | 比例 | 数量 | 比例 | 数量 | 比例 |
| 官职1 |  |  |  |  |  |  |  |  |
| 官职2 |  |  |  |  |  |  |  |  |
| …… |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

需要输出4种这样的结果。

分类前提：进士、举人、举人以下、空白。

**第二部分 官员的履历信息**

1. **科举与官员的动态流动（暂定）**

注：由于这部分动态分析的思路还有待完善，故分析的条件只是暂定的内容，但是各部分分析的主要内容是固定的。

动态流动情况需要结合官员的履历情况分析，即所用的数据是全部数据大约20000多条。

又因为数据库中履历信息是在“人物传记库”中检索“姓名”的结果，所以有可能实际履历信息与官员是不对应的。排除的依据可以用“任职时间”减去“生年”，如果数值为负或者大于90岁，可以认为其履历信息是错误的，因此分析中可以排除，**即排除履历不正确的数据。**

1. 最初任职职位

在不同“官职类别”下，不同“科举级别”的初次任职职位，即官员的起点。

条件：“姓名内编码”=1或2（即时间最早的几条履历所对应的职位），统计“任职职位”中出现的各种职位的频次。

结果：输出结果按照不同的“官职类别”（一级文件夹）下“科举级别”（二级文件夹）输出统计结果（“任职职位”或“任职职位简写”+次数）。

1. 最初任职的地点

在不同“官职类别”下，不同“科举级别”的初次任职的地点。

条件：“姓名内编码”=1或2（即时间最早的几条履历所对应的职位），统计“任职地”中出现的各种地方或机构的频次。

结果：输出结果按照不同的“官职类别”（一级文件夹）下“科举级别”（二级文件夹）输出统计结果（“任职地”+次数）。

1. 上升空间

以进士（进士、武进士和翻译进士共325人）群体为例。

去掉“履历错误”的人。

官员履历中，所能达到的最高职位。即履历中最后几条，对应的“姓名内编号”最大。

结果：输出结果按照不同的“官职类别”（一级文件夹）下“科举级别”（二级文件夹）输出统计结果（“任职职位”或“任职职位简写”+次数）。

1. 任职年限

在不同“官职类别”下，不同“科举级别”的任职年限。筛选出职官表，列C“姓名内编码”为0和1，计算任职年限。

条件：“任职年限”去掉为空的数据，计算后一个时间减去前一个时间的差。

结果：输出结果按照不同的“官职类别”（一级文件夹）下“科举级别”（二级文件夹）输出统计结果。时间差为0、1、2、3、4……，输出的结果为时间差和对应的次数。（时间差+次数）。

1. 区域间的流转

基本信息中的“官职类别”和“地区”为官员目前所在地区，履历信息中的“任职地”是官员一生中流动的地区。此项统计目的就是为了统计官员从目前所在地到全国其他地方流动的特点。

条件：只需要统计不同官职类别和科举级别下，官员履历中“任职地”和出现的次数。

结果：输出结果按照不同的“官职类别”（一级文件夹）下“科举级别”（二级文件夹）输出统计结果。输出的结果为地点和对应的次数。（“任职地”+次数）。

注：1、2、3项的统计条件待定，待定原因：

1. 不好确定条件中“任职地”“任职职位”“任职职位简写”是否选择“姓名内编号”的1或2或3？即最初任职地和职位的选择。
2. 涉及履历中“职位”的统计，应该选择“任职职位”还是“任职职位简写”。因为有的“任职职位”内容没办法简写。