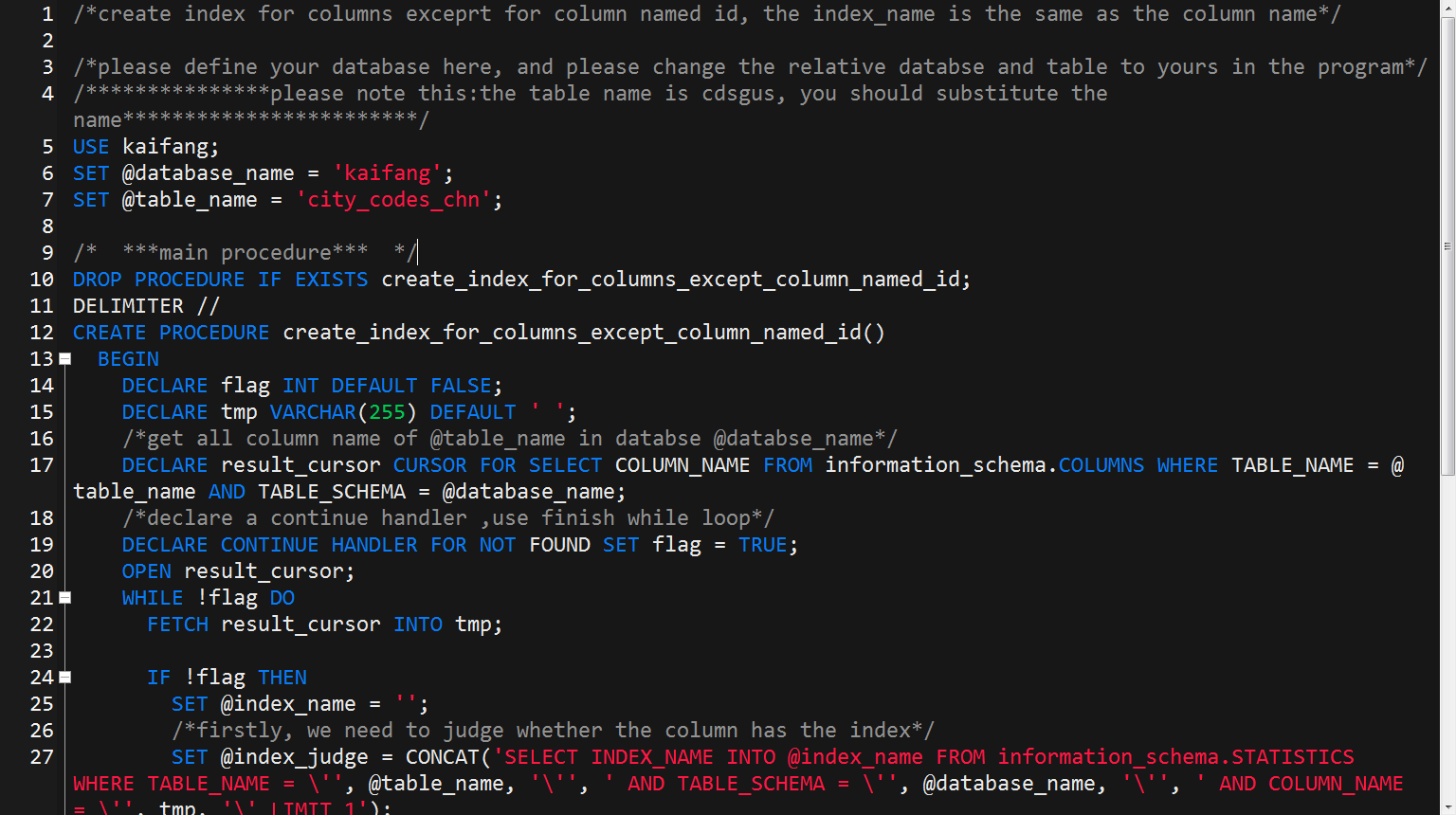
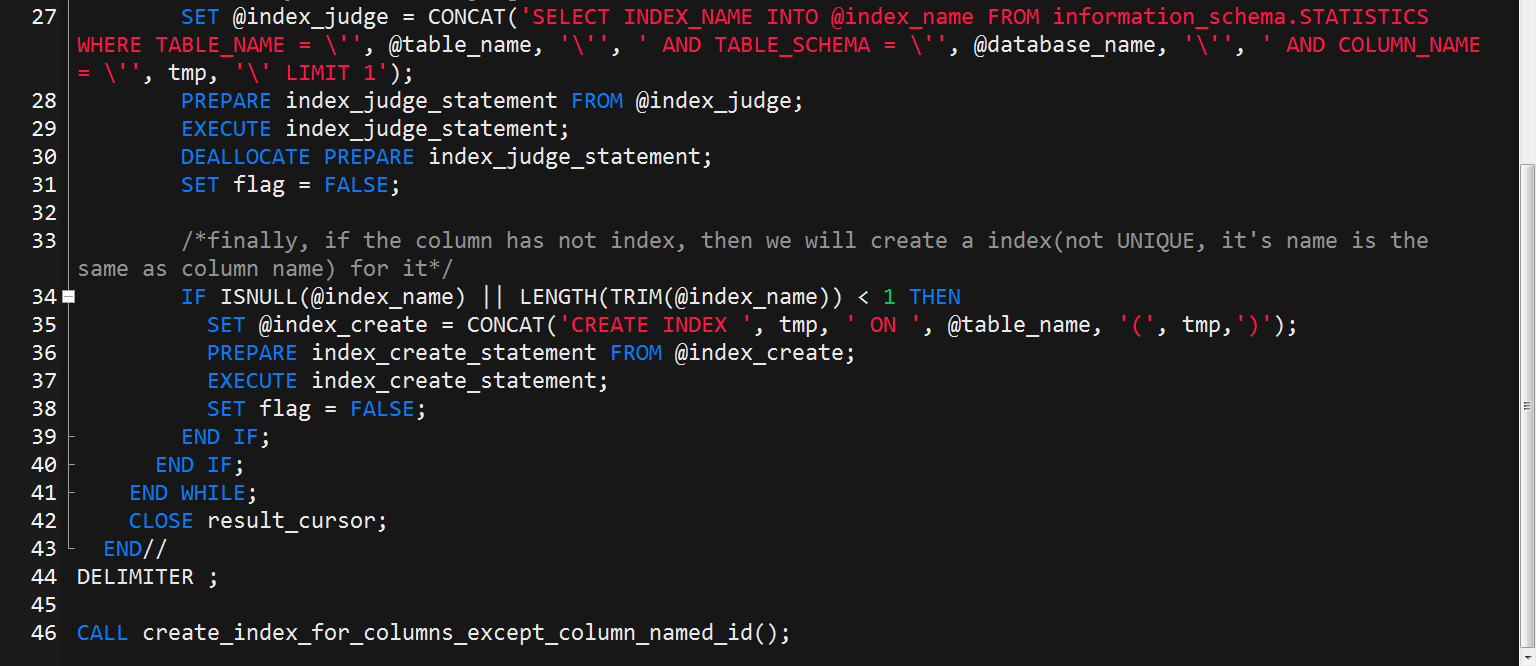
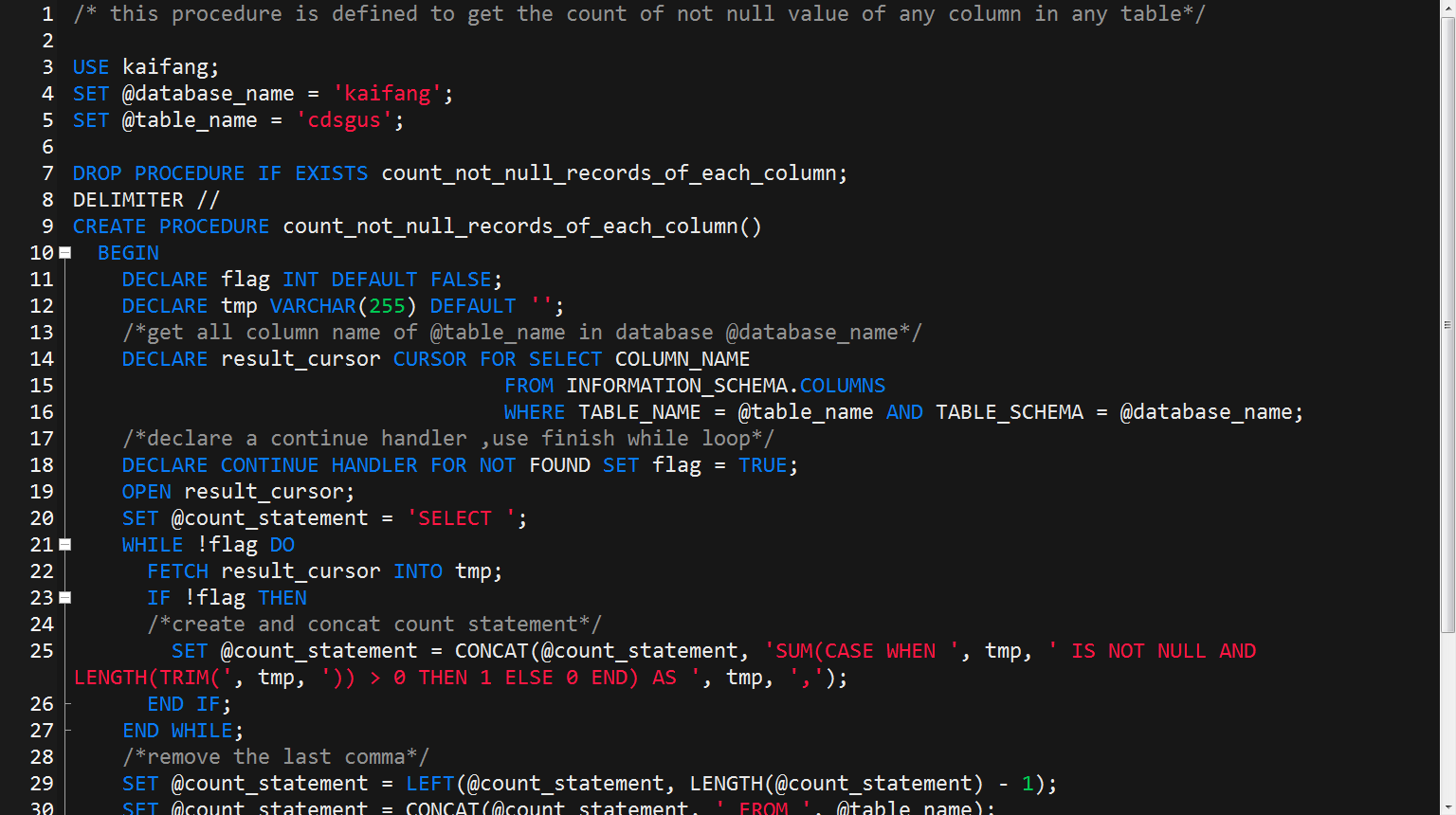
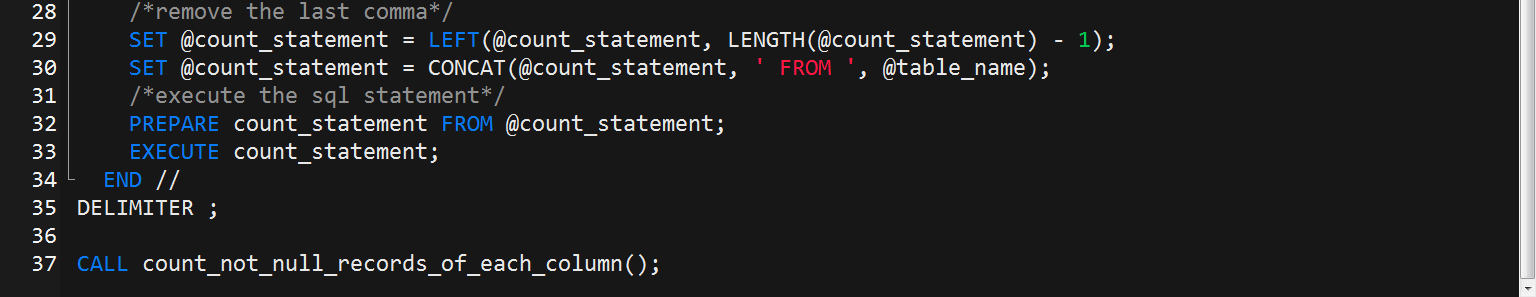
**2KW数据分析报告**

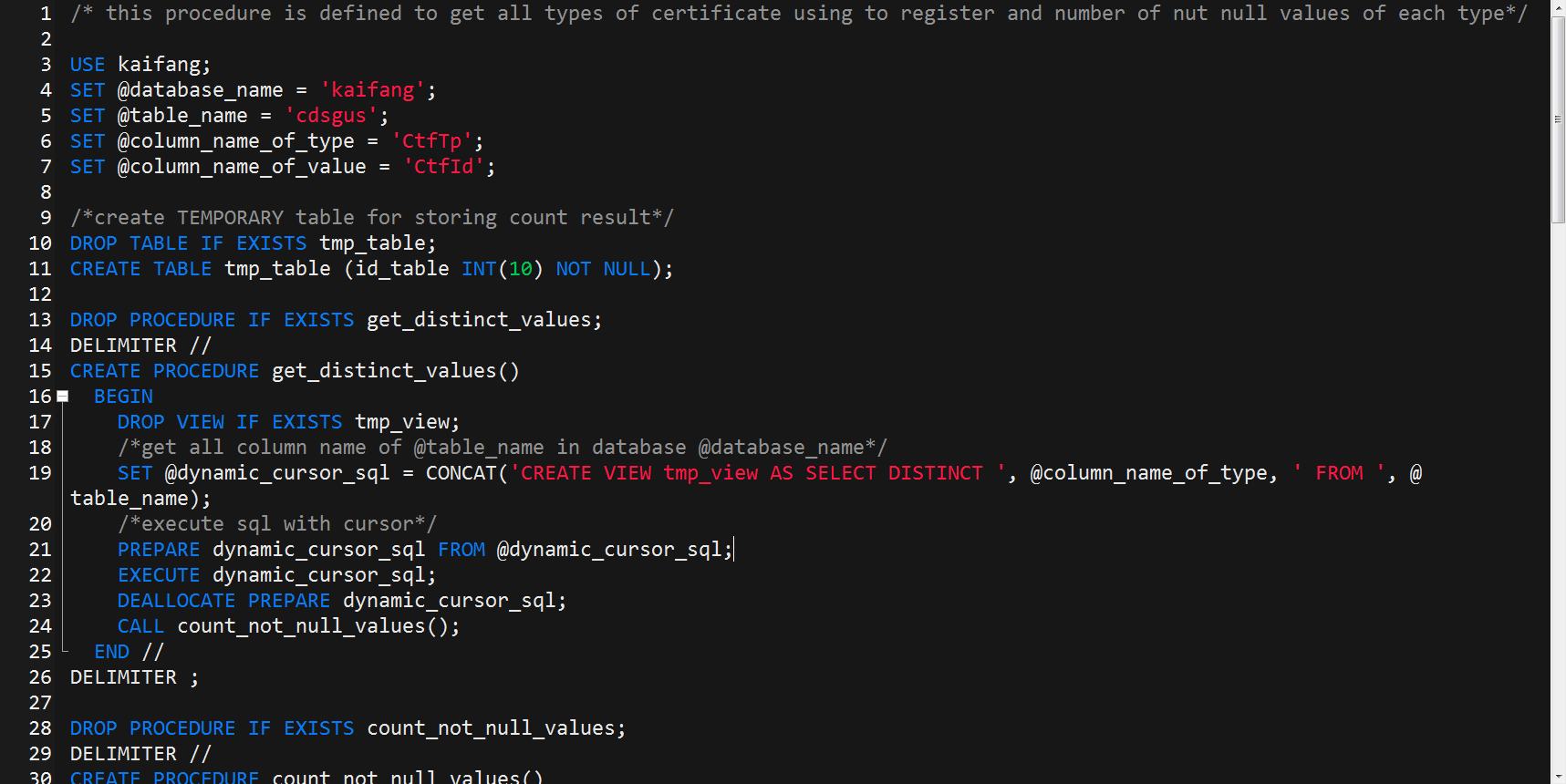
1. **结论**
2. **数据清洗**
3. **创建索引**

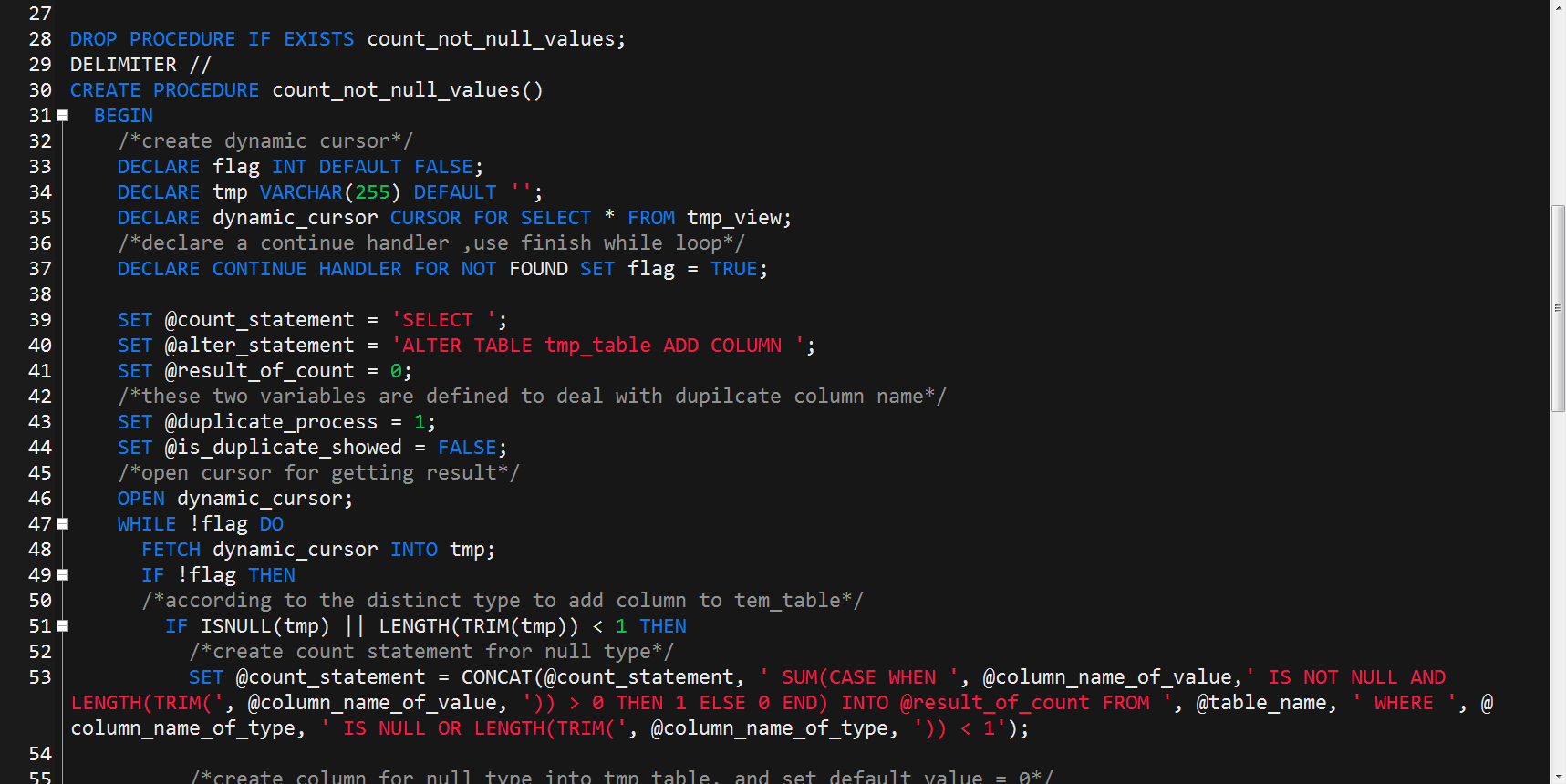
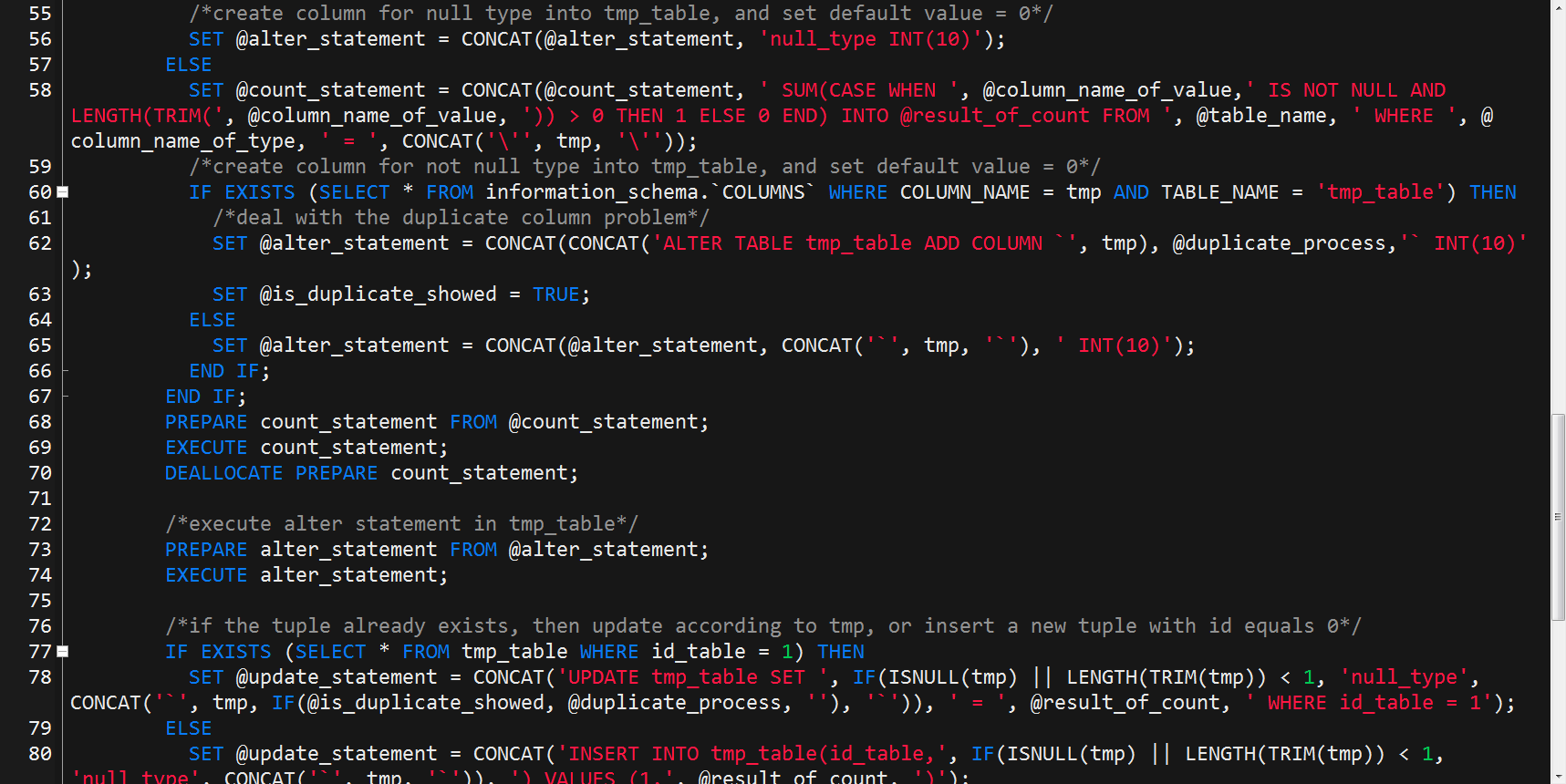
本部分对初始数据表中的除id外的其他字段创建普通索引，暂且不考虑创建主键和组合索引。目的是为提高之后数据清洗工作中sql处理的效率，sql代码如下图所示：

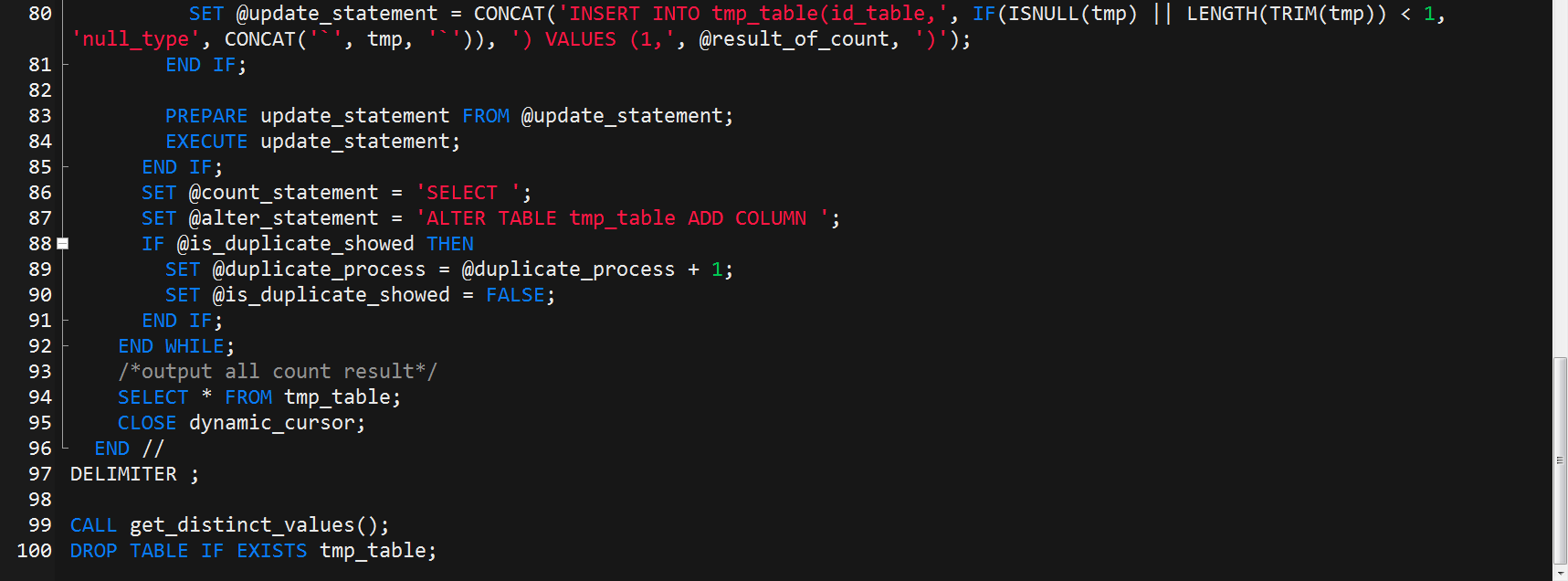


1. **删除与修正**
2. **统计表中每个属性列所有非空值的个数**
3. **定义非空值**：值非NULL并且去除空格后的长度不为0的值。
4. **sql代码**：
5. **结果(根据数量进行倒序排列)**：

|  |  |
| --- | --- |
| **字段** | **数量** |
| id | 20,050,144 |
| CtfId | 20,050,133 |
| Name | 20,049,875 |
| CtfTp | 20,049,096 |
| Birthday | 19,896,482 |
| Version | 19,287,789 |
| Gender | 19,258,580 |
| Family | 19,042,893 |
| District2 | 17,557,317 |
| Dirty | 17,298,128 |
| District3 | 17,258,331 |
| District4 | 16,972,167 |
| Address | 16,588,392 |
| Nation | 13,367,440 |
| Mobile | 7,716,206 |
| Tel | 5,380,056 |
| EMail | 2,909,070 |
| Zip | 1,703,521 |
| Fax | 1,019,377 |
| LastNm | 994,430 |
| District1 | 763,488 |
| Company | 694,928 |
| FirstNm | 342,895 |
| Duty | 338,012 |
| Taste | 317,709 |
| Education | 309,813 |
| CZip | 44,452 |
| Descriot | 20,588 |
| CardNo | 10,115 |
| CAddress | 2,034 |
| CTel | 993 |
| District5 | 701 |
| District6 | 52 |

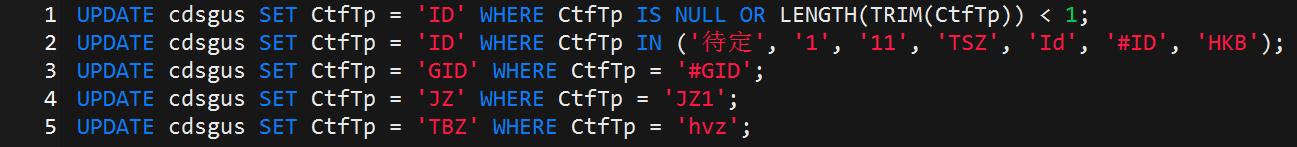
1. **修正关键字段-证件号(CtfId)和证件类型(CtfTp)**
2. **原因**：证件号包含着个人的极其重要的属性，因此应当首先对其以及其所关联的重要字段证件类型进行修正。
3. **修正过程**：
4. **统计证件类型及每个证件类型非空值的数量**，sql代码如下：





1. **统计结果(根据数量倒序排列)**

|  |  |  |
| --- | --- | --- |
| **证件类型** | **字段释义** | **非空值数** |
| ID | 身份证(默认选中) | 18,888,953 |
| VSA | 签证 | 378,608 |
| OTH | 其他 | 243,486 |
| JID | 军人证 | 120,464 |
| TXZ | 通行证(待定) | 116,962 |
| TBZ | 台胞证 | 114,167 |
| JZ | 驾照(待定) | 85,855 |
| GID | Genaral ID，即一般证件(待定) | 46,743 |
| HXZ | 回乡证 | 43,968 |
| JLZ | 居留证(待定) | 5,025 |
| hvz | 台胞证 | 2,607 |
| SBZ | 士兵证 | 1,706 |
| 无值 |  | 1,047 |
| JGZ | 军官证 | 340 |
| 待定 |  | 76 |
| HKB | 户口本 | 70 |
| 11 |  | 28 |
| Id | 同ID | 11 |
| TSZ | 类似通行证 | 11 |
| #ID | 同ID | 3 |
| JZ1 | 驾照 | 1 |
| #GID | 同GID | 1 |
| 1 |  | 1 |

1. **处理**：
2. 将#GID修正类型为GID；
3. 将JZ1修正为JZ；
4. 对#ID、Id、HKB、待定、TSZ、1、11、无值全部修改为ID；
5. 将hvz修改为TBZ(大部分与台胞证格式相似)。
6. **sql代码**：
7. **更改后证件类型及数量**

|  |  |  |
| --- | --- | --- |
| **证件类型** | **字段释义** | **非空值数** |
| ID | 身份证(默认选中) | 18,890,200 |
| VSA | 签证 | 378,608 |
| OTH | 其他 | 243,486 |
| JID | 军人证 | 120,464 |
| TXZ | 通行证(待定) | 116,962 |
| TBZ | 台胞证 | 116,774 |
| JZ | 驾照(待定) | 85,856 |
| GID | Genaral ID，即一般证件(待定) | 46,744 |
| HXZ | 回乡证 | 43,968 |
| JLZ | 居留证(待定) | 5,025 |
| SBZ | 士兵证 | 1,706 |
| JGZ | 军官证 | 340 |

1. **修正中国人15、18位身份证号并水平分表**
2. **分表原因**：由于该表中在中国，因此中国人巨多，从而决定将中国人的信息单独建表存放。
3. **新表字段说明**：

|  |  |  |
| --- | --- | --- |
| **字段** | **释义** | **其他说明** |
| id | 本表id | 自增，主键 |
| name | 用户填写的姓名 | 来自原表Name |
| identity\_number | 身份证号 | **经身份证号有效性验证** |
| gender | 性别 | **由有效身份证号进行修正** |
| birthday | 出生日期 | **由有效身份证号进行修正** |
| contact\_address | 联系地址 | 来自原表Address |
| zip | 邮编 | 来自原表 |
| mobile | 手机号 |  |
| tel\_number | 固话号 |  |
| email | 邮箱 |  |
| nation | 民族 |  |
| register\_date | 注册时间 |  |
| id\_cdsgus | 原表id | 关联原表的id |

1. **处理说明**：对每个CtfId的值作如下处理过程
2. 首先判断该值是否符合18位ID正则，如果符合，则根据18位身份证验算规则验证是否有效。如果有效，则该值所在元组添加入“拥有中国ID的用户表”，反之修正其身份证号后再添加入“拥有中国ID的用户表”中。
3. 如果不符合18位ID正则，则判断是否符合15位ID正则，如果符合，则按照置换规则将15位ID转为18位添加入“拥有中国ID的用户表”中。
4. 如果18位和15位都匹配失败，那么我们认为这很有可能是一个非中国ID的数据，但考虑到
5. **sql代码**
6. **删除类型1**：根据统计结果，我们决定以数据量2000W的1%即20W作为阈值，删除字段CZip、Descriot、CardNo、CAddress、CTel、District5、District6。
7. **删除类型2**：
8. **sql代码**：
9. **修正­**