|  |  |
| --- | --- |
|  |  |
| 项目报告1 |  |
|  |  |
|  | 课程：数据挖掘姓名：姜其升 |
|  | 学号：1752058指导教师：饶卫雄教授 |

[项目报告1 1](#_Toc3485)

[1 数据预处理 3](#_Toc2192)

[1.1主要思路 3](#_Toc23102)

[1.2关键代码 3](#_Toc6276)

[1.3处理结果 3](#_Toc7663)

## 1 数据预处理

### 1.1主要思路

首先分别计算出每条交易记录的四级品类结构、星期、距离最早记录的周数、月数、天数等信息。处理缺失的品牌信息（bndno），如果商品的bndno不存在，则将该商品作为一个单独的品牌。

之后，按照时间和要求统计的信息进行分组求和。形成时间序列数据，导出为csv文件，位于P1文件夹下。

### 1.2关键代码

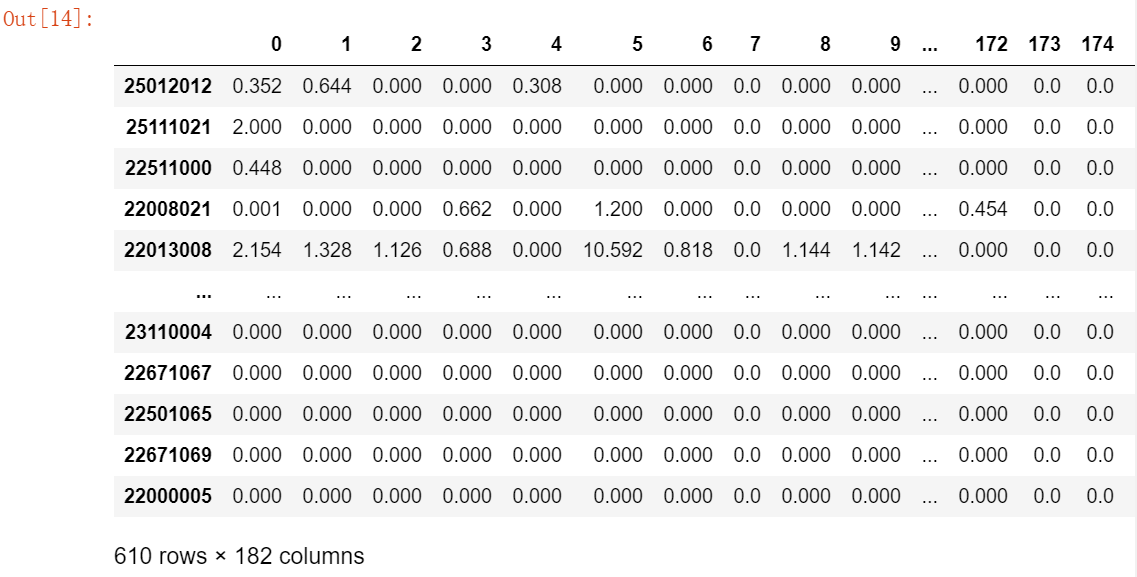
1. **def** getTimeInfo(time, item):
2. col = csv\_data[time].drop\_duplicates().values
3. row = csv\_data[item].drop\_duplicates().values
4. g\_dp = csv\_data["qty"].groupby([csv\_data[time], csv\_data[item]]).sum()
5. new\_dp = pd.DataFrame(0.0, index = row, columns = col)
6. **for** i **in** g\_dp.index.levels[0]:
7. **for** j **in** g\_dp[i].index:
8. new\_dp[i][j] = g\_dp[i][j]
9. **return** new\_dp
11. **def** getTime2Info(time):
12. col2 = csv\_data[time].drop\_duplicates().values
13. g\_dp2 = csv\_data["qty"].groupby([csv\_data[time], csv\_data["pluno"], csv\_data["bndno"], csv\_data["type5"], csv\_data["type4"], csv\_data["type3"], csv\_data["type2"]]).sum()
14. new\_dp2 = pd.DataFrame(g\_dp2)
15. new\_dp2.reset\_index(inplace=True)
16. new\_dp3 = csv\_data[["pluno", "bndno", "type5", "type4", "type3", "type2"]].drop\_duplicates().sort\_values(["pluno", "bndno"])
17. **for** i **in** col2:
18. new\_dp3[i] = 0.0
19. **for** i, row **in** new\_dp2.iterrows():
20. new\_dp3.loc[(new\_dp3["pluno"]==row["pluno"])&(new\_dp3["bndno"]==row["bndno"]), int(row[time])] = row["qty"]
21. **return** new\_dp3

### 1.3处理结果

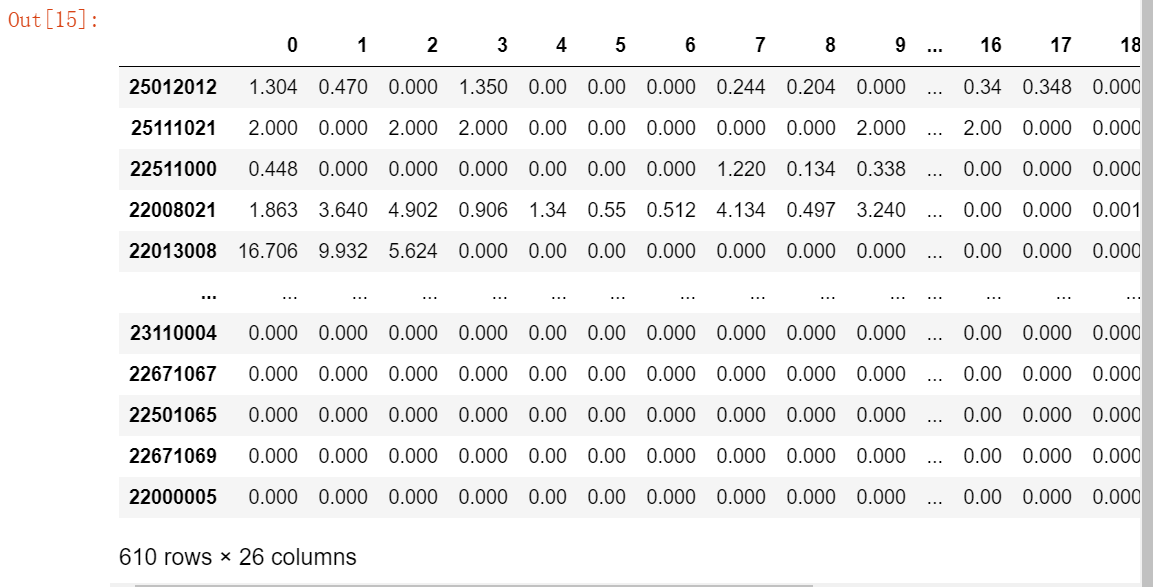
横行为时间，列为相应的汇总依据，下表为对应的csv文件名称。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 每天 | 每周 | 每月 |
| 商品编号 | day\_pluno | week\_pluno | month\_pluno |
| 品牌编号 | day\_bndno | week\_bndno | month\_bndno |
| 一级品类 | day\_type2 | week\_type2 | month\_type2 |
| 二级品类 | day\_type3 | week\_type3 | month\_type3 |
| 三级品类 | day\_type4 | week\_type4 | month\_type4 |
| 四级品类 | day\_type5 | week\_type5 | month\_type5 |
| 商品编号、品牌编号、一级品类、二级品类、三级品类、四级品类 | day\_pluno\_bndno\_type | week\_pluno\_bndno\_type | month\_pluno\_bndno\_type |

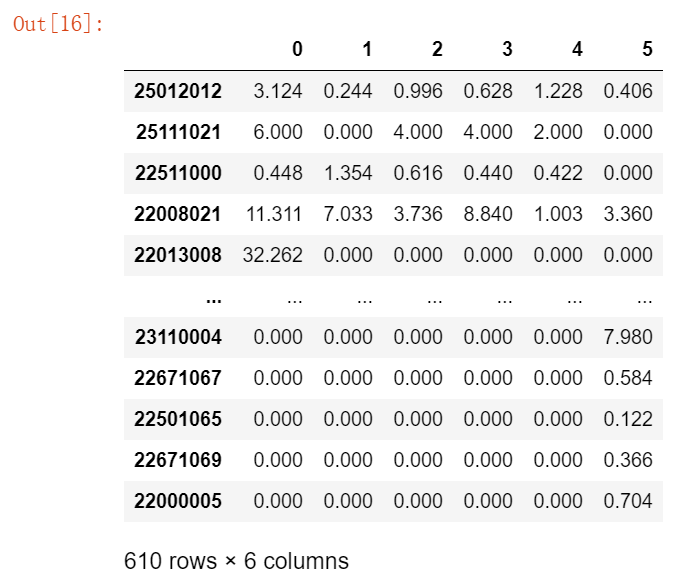
day\_pluno



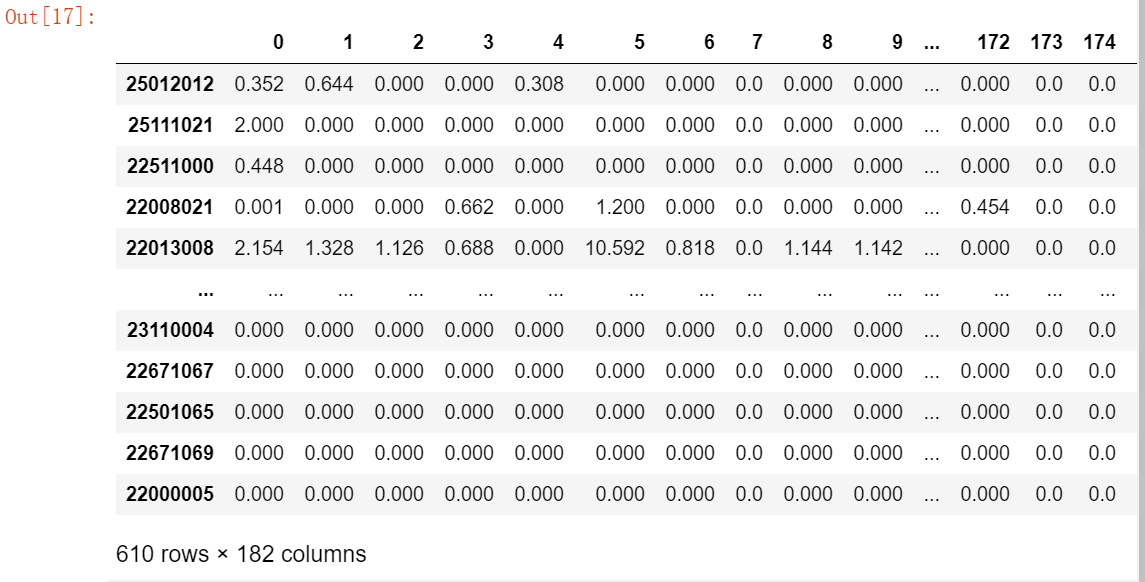
week\_pluno



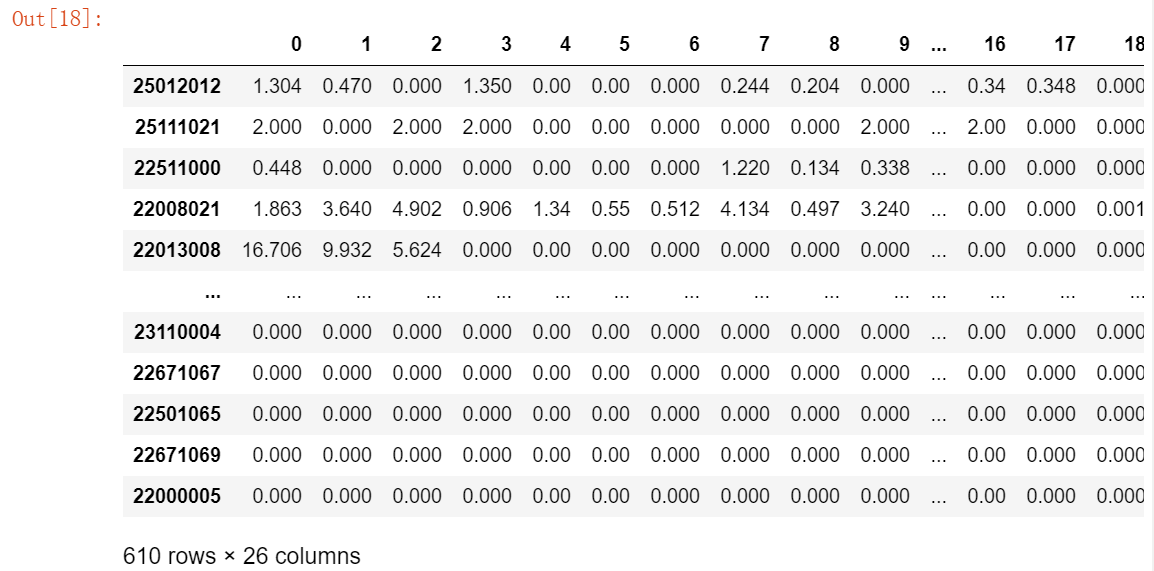
month\_pluno



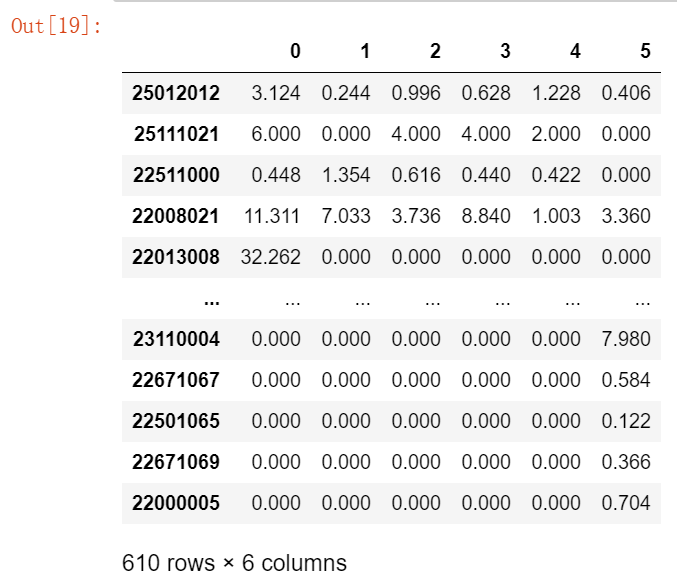
day\_bndno



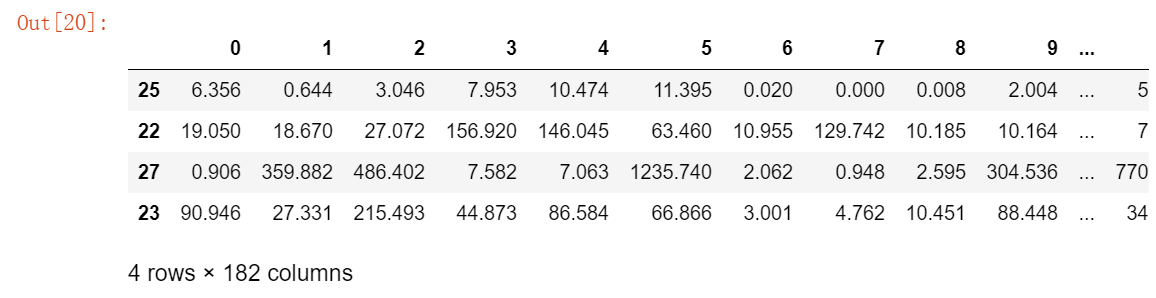
week\_bndno



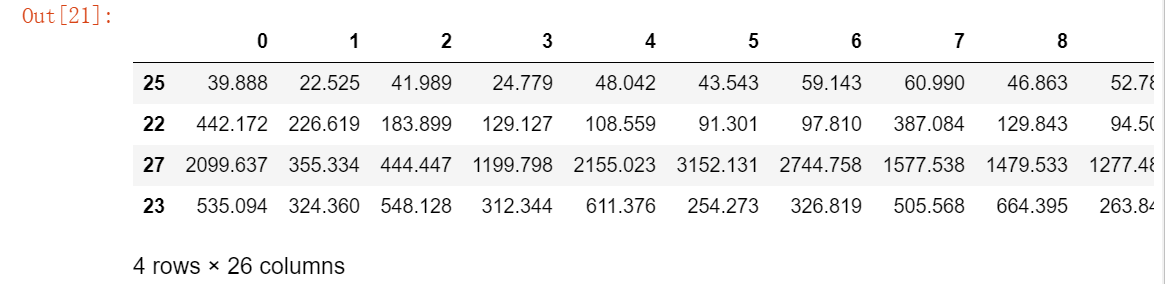
month\_bndno



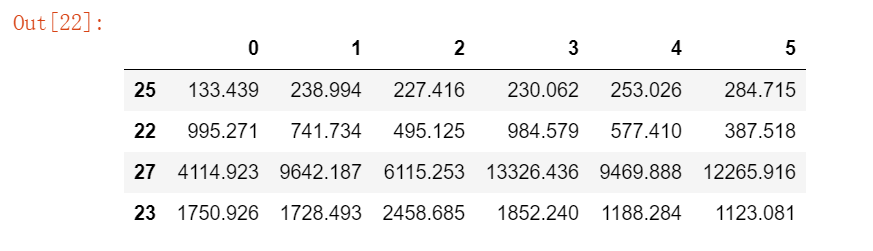
day\_type2



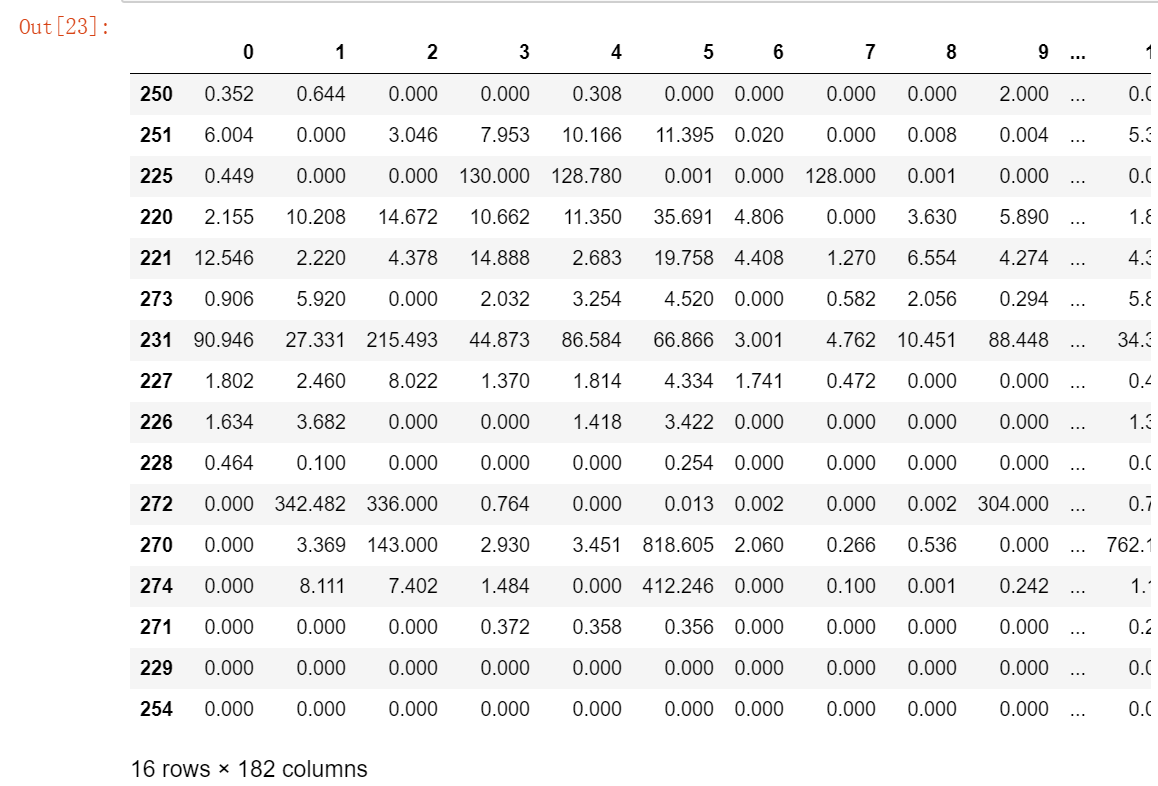
week\_type2



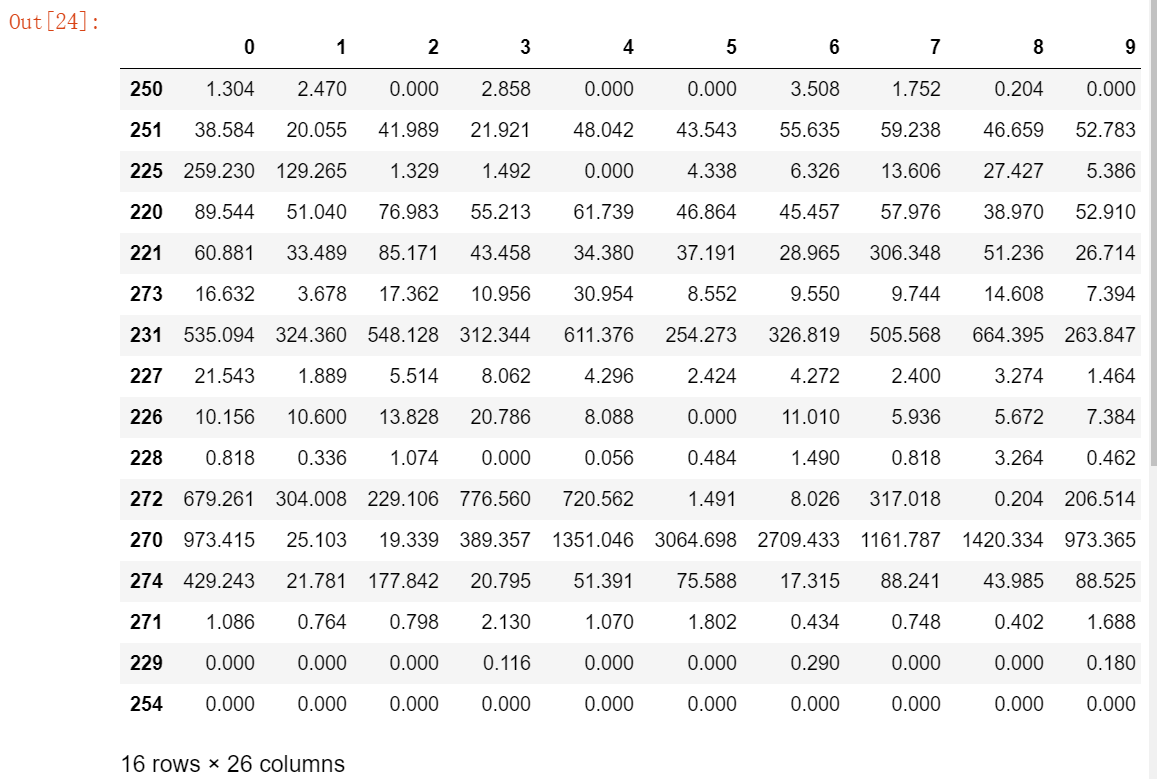
month\_type2



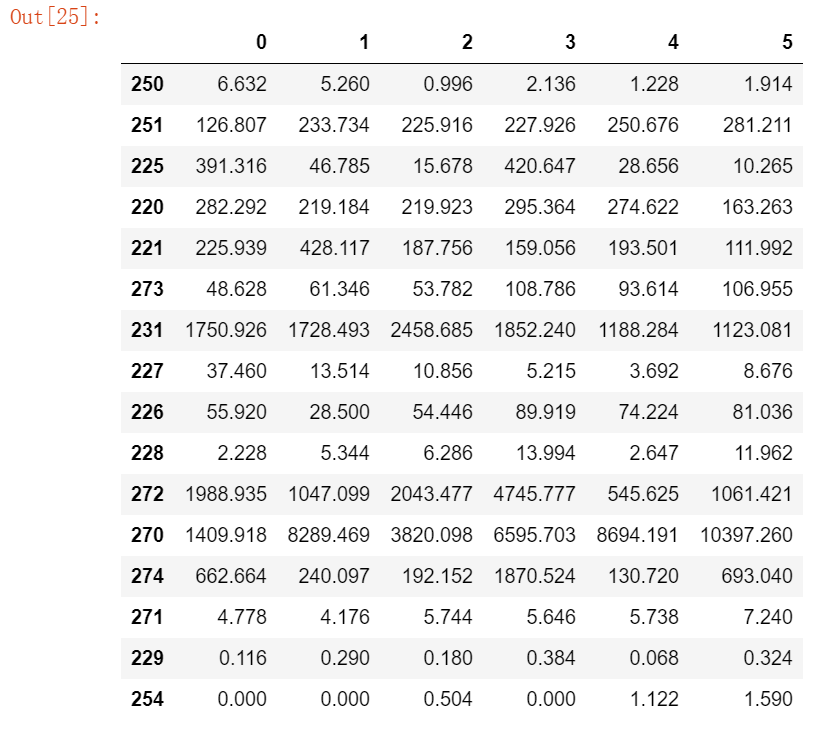
day\_type3

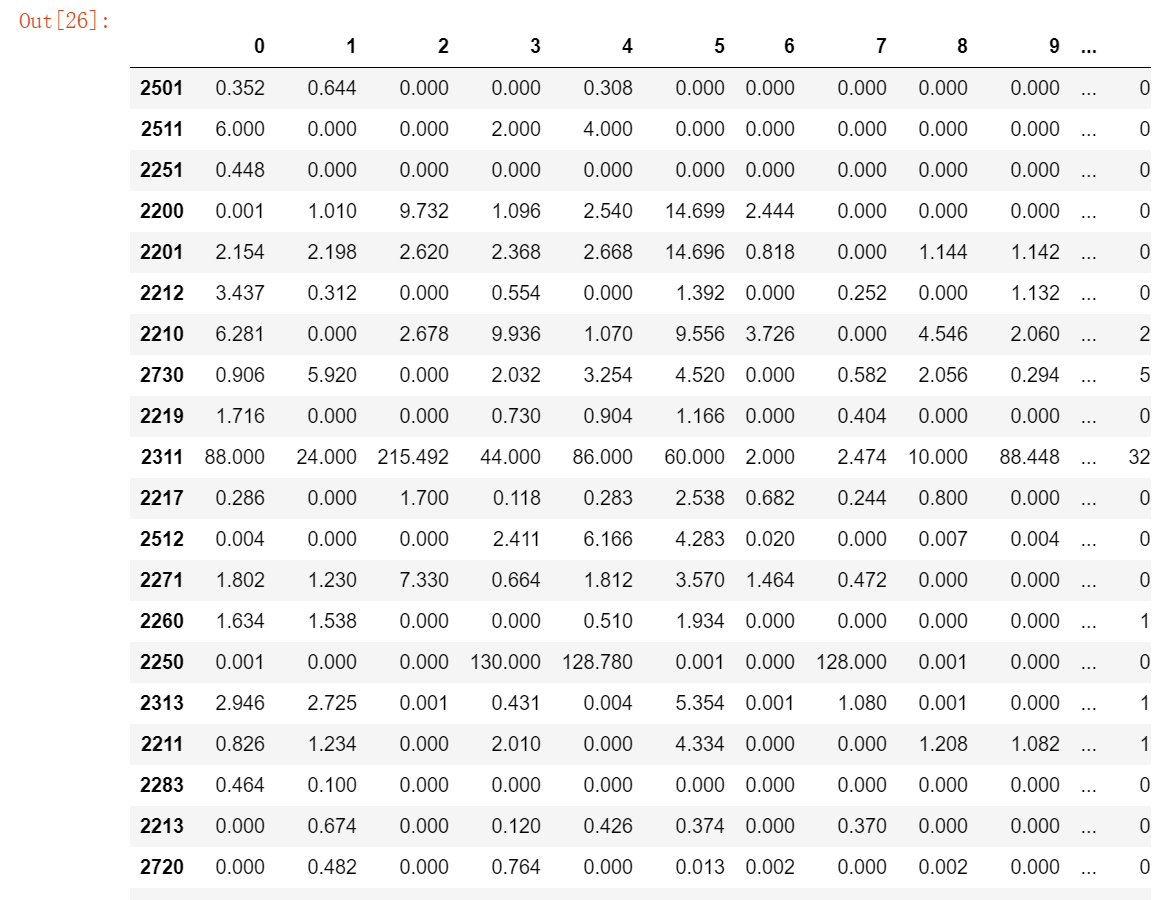


week\_type3

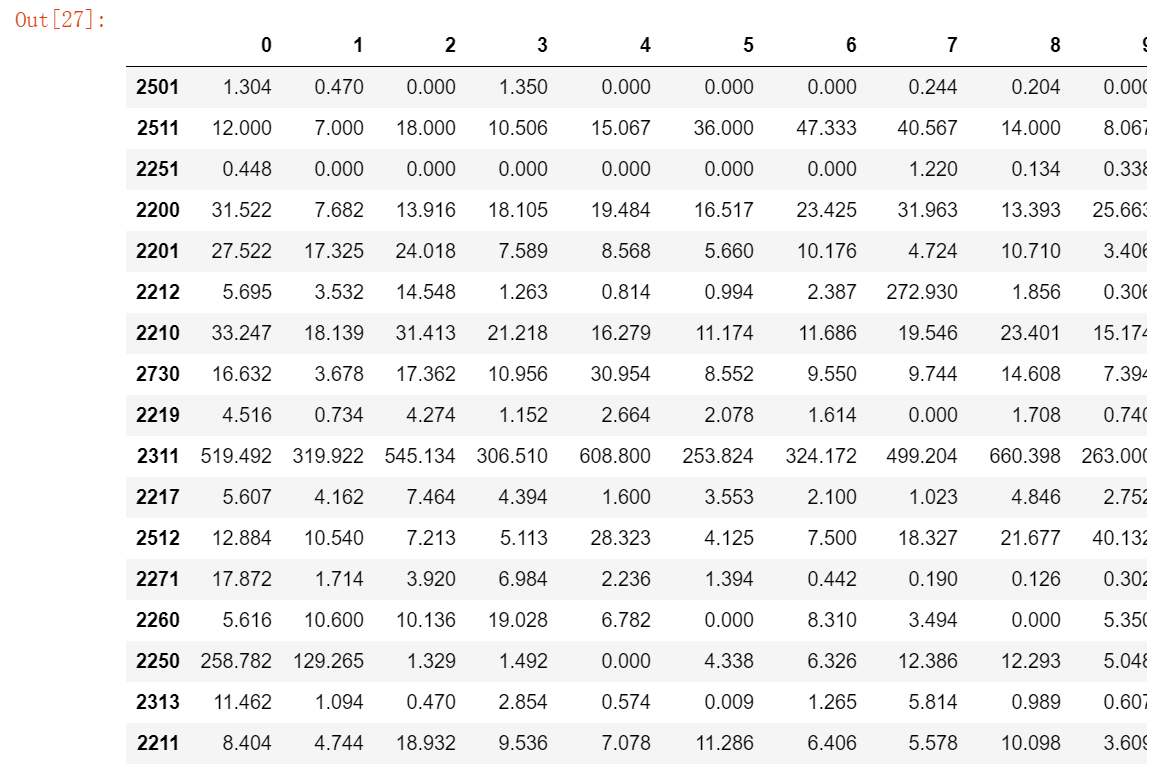


month\_type3

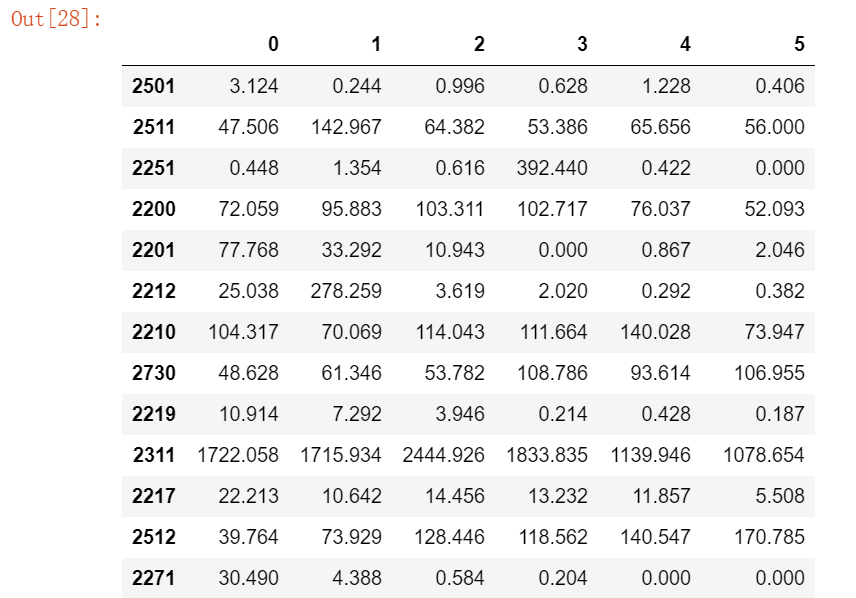


day\_type4

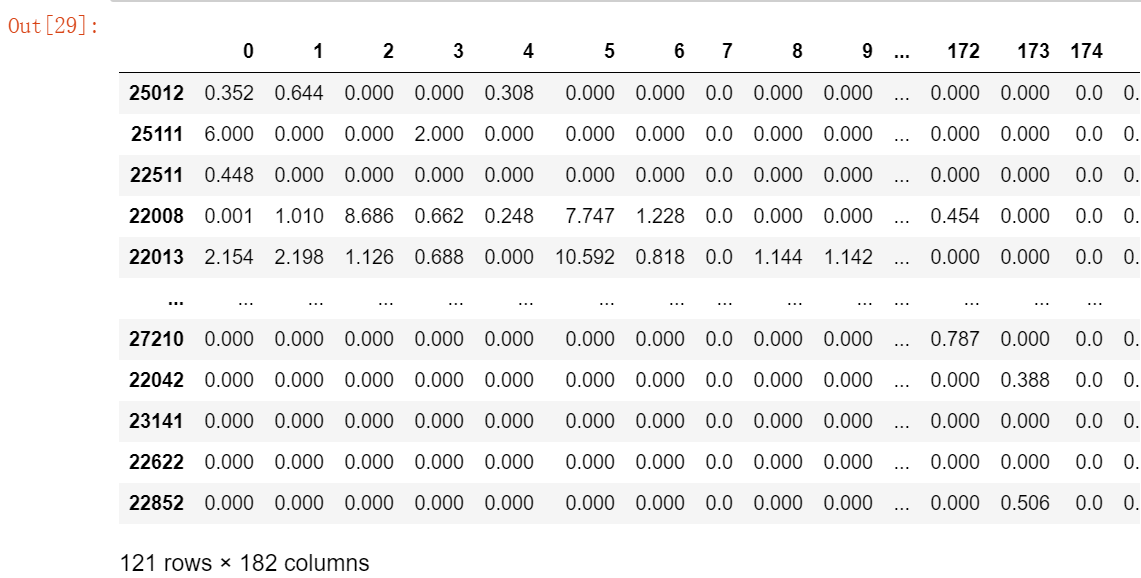
week\_type4



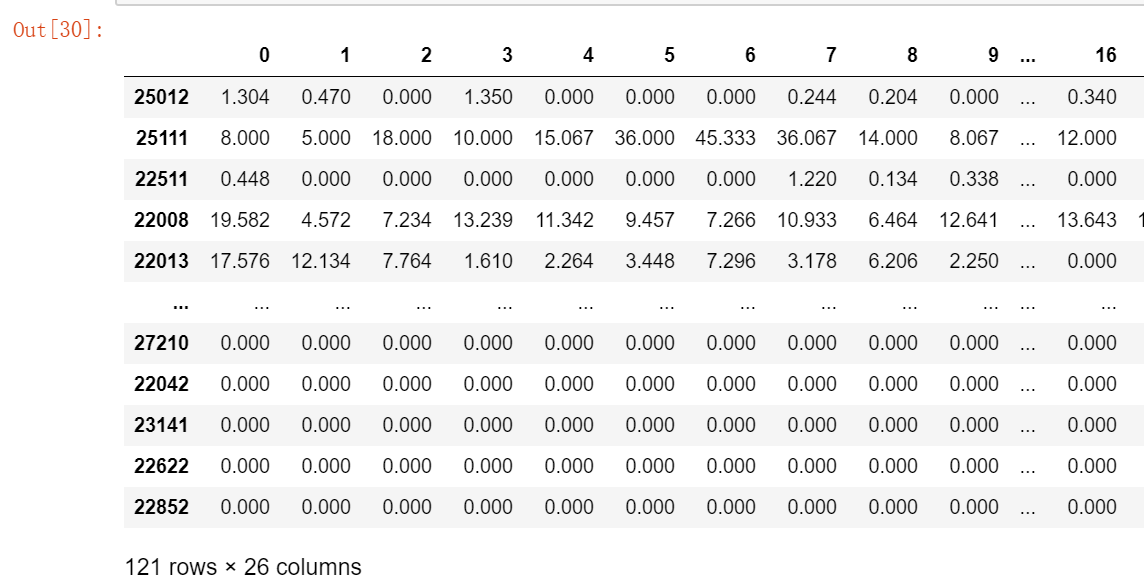
month\_type4



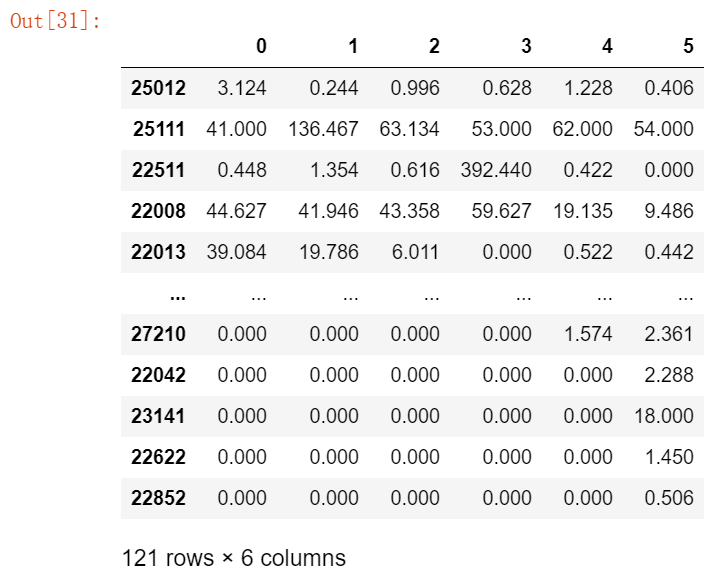
day\_type5



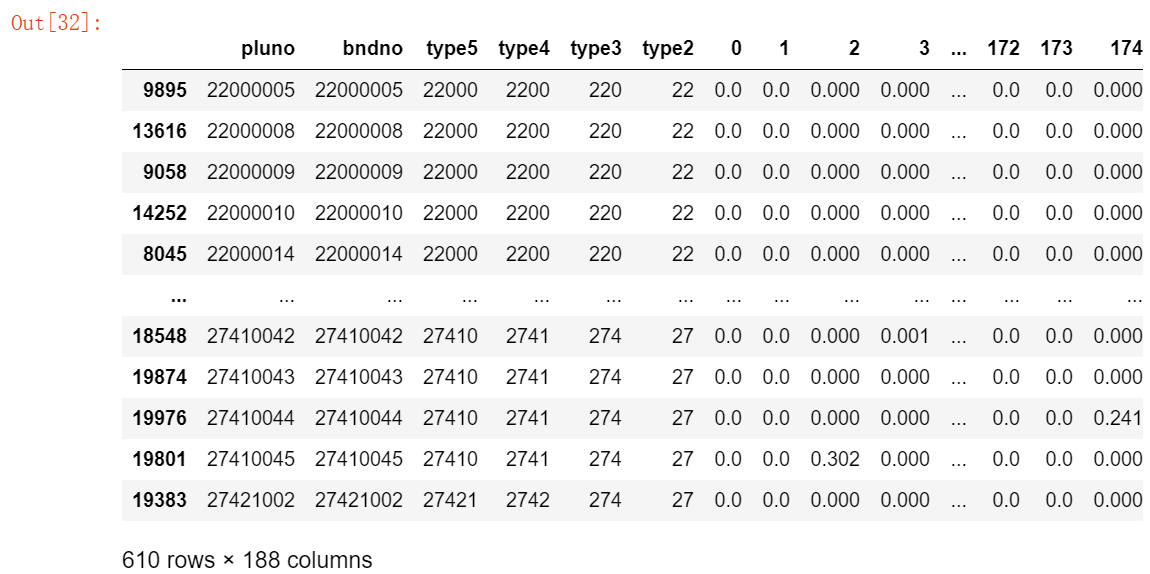
week\_type5



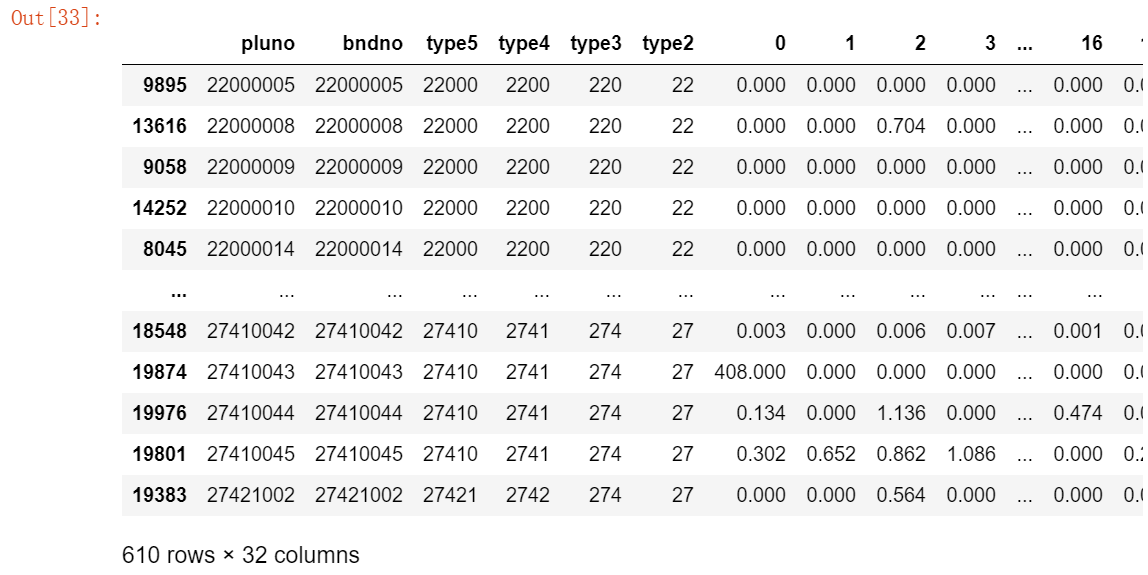
month\_type5



day\_pluno\_bndno\_type



week\_pluno\_bndno\_type



month\_pluno\_bndno\_type

