|  |  |
| --- | --- |
|  |  |
| 项目报告2 |  |
|  |  |
|  | 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2 1](#_Toc16320)

[1特征工程 3](#_Toc19692)

[1.1基本思路 3](#_Toc2642)

[1.2代码举例 3](#_Toc20514)

[1.3处理结果 3](#_Toc8603)

## 1特征工程

### 1.1基本思路

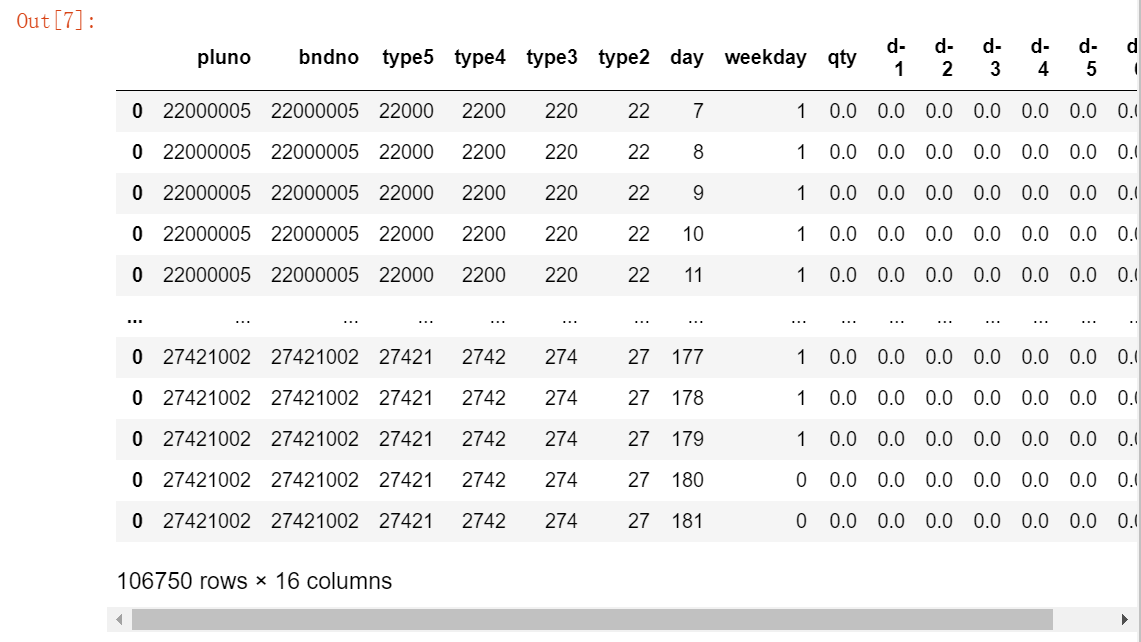
分别按要求对销量进行统计，对于最大值、最小值、均值等分别使用max、min和np.mean进行求解，之后分别导出csv文件，位于P2文件夹下。

### 1.2代码举例

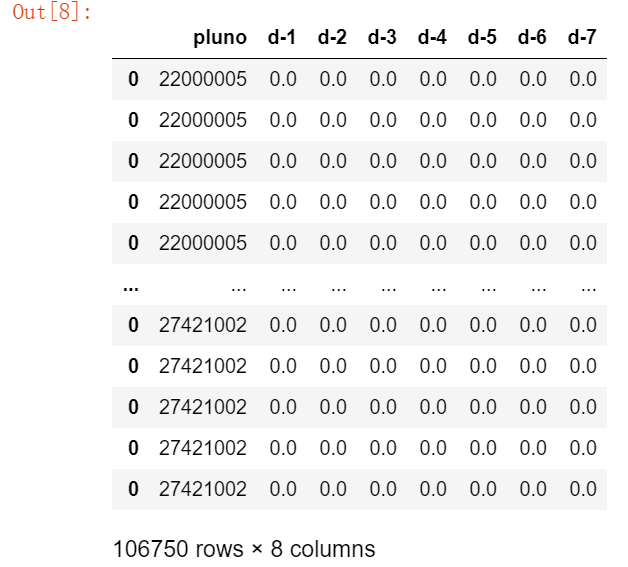
1. feature\_i\_header = ["pluno", "bndno", "type5", "type4", "type3", "type2", "day", "weekday", "qty", "d-1", "d-2", "d-3", "d-4", "d-5", "d-6", "d-7"]
2. feature\_i = pd.DataFrame(columns=feature\_i\_header)
3. **for** i, row **in** day\_pluno\_bndno\_type.iterrows():
4. **for** j **in** range(7, max(days)+1):
5. weekday = 1
6. **if** j%7 == 5 **or** j%7 == 6:
7. weekday = 0
8. feature\_i = feature\_i.append([{"pluno":int(row["pluno"])
9. , "bndno":int(row["bndno"])
10. , "type5":int(row["type5"])
11. , "type4":int(row["type4"])
12. , "type3":int(row["type3"])
13. , "type2":int(row["type2"])
14. , "day":j
15. , "weekday":weekday
16. , "qty":row[j]
17. , "d-1":row[j-1]
18. , "d-2":row[j-2]
19. , "d-3":row[j-3]
20. , "d-4":row[j-4]
21. , "d-5":row[j-5]
22. , "d-6":row[j-6]
23. , "d-7":row[j-7]}])

### 1.3处理结果

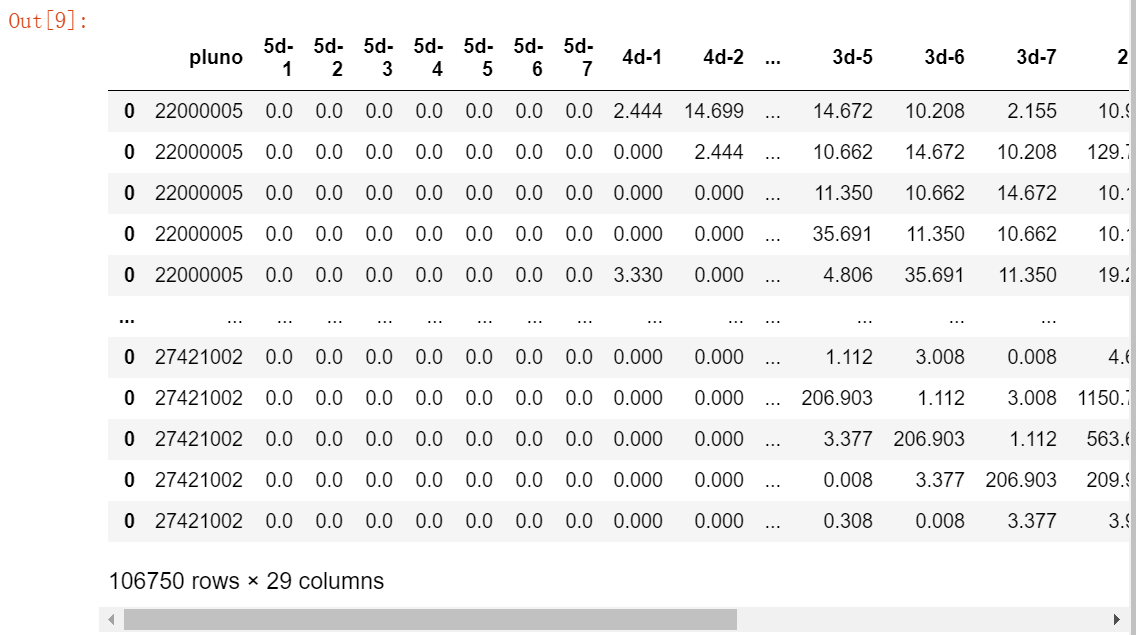
特征i，文件名为feature\_i。



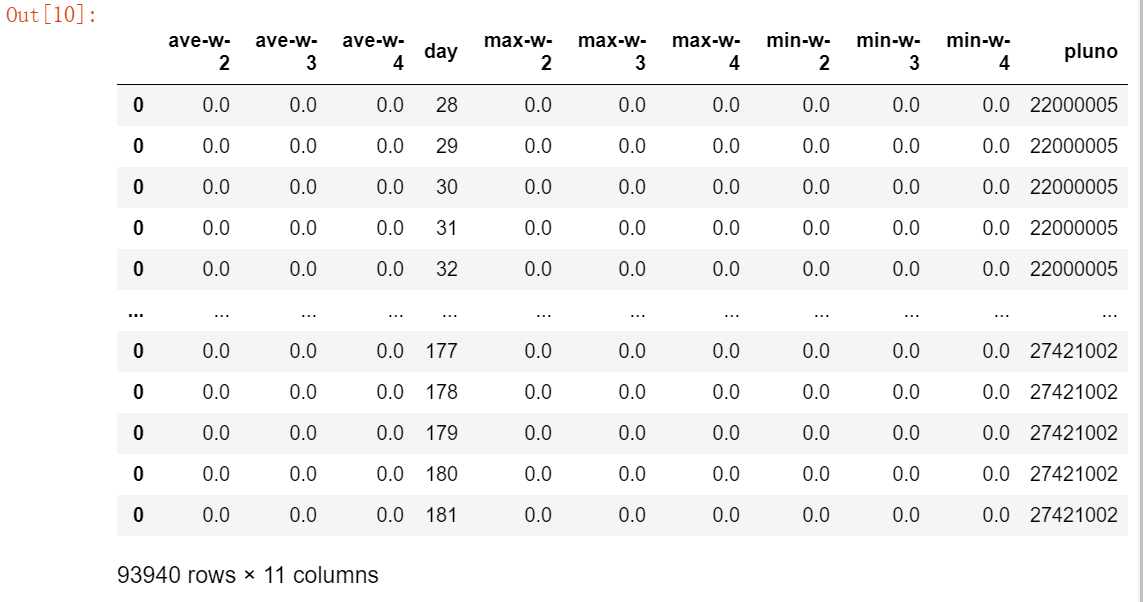
特征ii，文件名为feature\_ii。



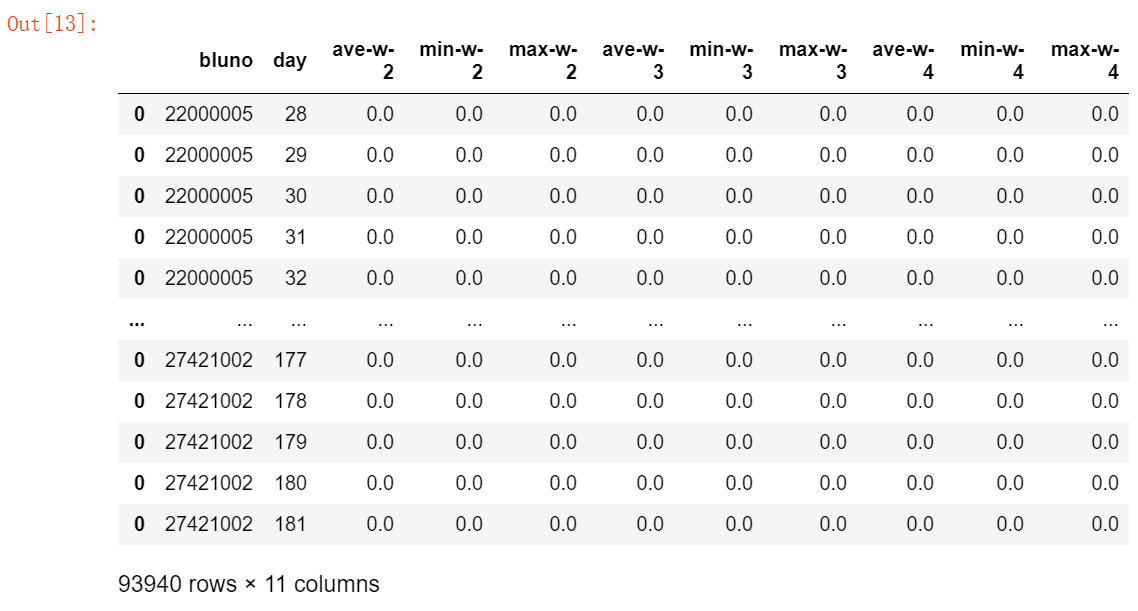
特征iii，文件名为feature\_iii。其中5d-1~5d-7分别代表第四级品类前1日至前7日当日销量；4d-1~4d-7分别代表第三级品类前1日至前7日当日销量；3d-1~3d-7分别代表第二级品类前1日至前7日当日销量；2d-1~2d-7分别代表第一级品类前1日至前7日当日销量。



特征iv，文件名为feature\_iv。其中ave-w-2、max-w-2、min-w-2分别代表前第2周中该周的每日销量平均值、最大值、最小值；ave-w-3、max-w-3、min-w-3分别代表前第3周中该周的每日销量平均值、最大值、最小值；ave-w-4、max-w-4、min-w-4分别代表前第4周中该周的每日销量平均值、最大值、最小值。



特征v，文件名为feature\_v。



特征vi，文件名为feature\_vi。其中2ave-w-2、2max-w-2、2min-w-2分别代表第一级品类前第2周中该周的每日销量平均值、最大值、最小值；2ave-w-3、2max-w-3、2min-w-3分别代表第一级品类前第3周中该周的每日销量平均值、最大值、最小值；3ave-w-4、3max-w-4、3min-w-4分别代表第二级品类前第4周中该周的每日销量平均值、最大值、最小值；3ave-w-2、3max-w-2、3min-w-2分别代表第二级品类前第2周中该周的每日销量平均值、最大值、最小值；3ave-w-3、3max-w-3、3min-w-3分别代表第二级品类前第3周中该周的每日销量平均值、最大值、最小值；3ave-w-4、3max-w-4、3min-w-4分别代表第二级品类前第4周中该周的每日销量平均值、最大值、最小值；4ave-w-2、4max-w-2、4min-w-2分别代表第三级品类前第2周中该周的每日销量平均值、最大值、最小值；4ave-w-3、4max-w-3、4min-w-3分别代表第三级品类前第3周中该周的每日销量平均值、最大值、最小值；4ave-w-4、4max-w-4、4min-w-4分别代表第三级品类前第4周中该周的每日销量平均值、最大值、最小值；5ave-w-2、5max-w-2、5min-w-2分别代表第四级品类前第2周中该周的每日销量平均值、最大值、最小值；5ave-w-3、5max-w-3、5min-w-3分别代表第四级品类前第3周中该周的每日销量平均值、最大值、最小值；5ave-w-4、5max-w-4、5min-w-4分别代表第四级品类前第4周中该周的每日销量平均值、最大值、最小值。

