数据结果分析

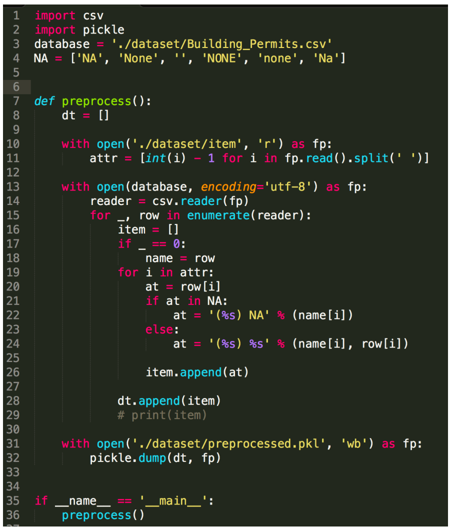
数据预处理

为了便于进行数据挖掘，需要对数据集进行处理，转换成适合关联规则挖掘的形式。由于并不是所有的属性都有进行关联规则挖掘的价值，因此在这里选了 9 个标称属性进行关联规则挖掘。这 9 个属性分别是：

1. Permit Type
2. Street Number
3. Current Status
4. Structural Notification
5. Existing Use
6. Proposed Use
7. Existing Construction Type(Definition)
8. Proposed Construction Type(Definition)
9. Site Permit

这 9 个属性分别对应上图中的 2,7,14,20,28,30,34,36,38

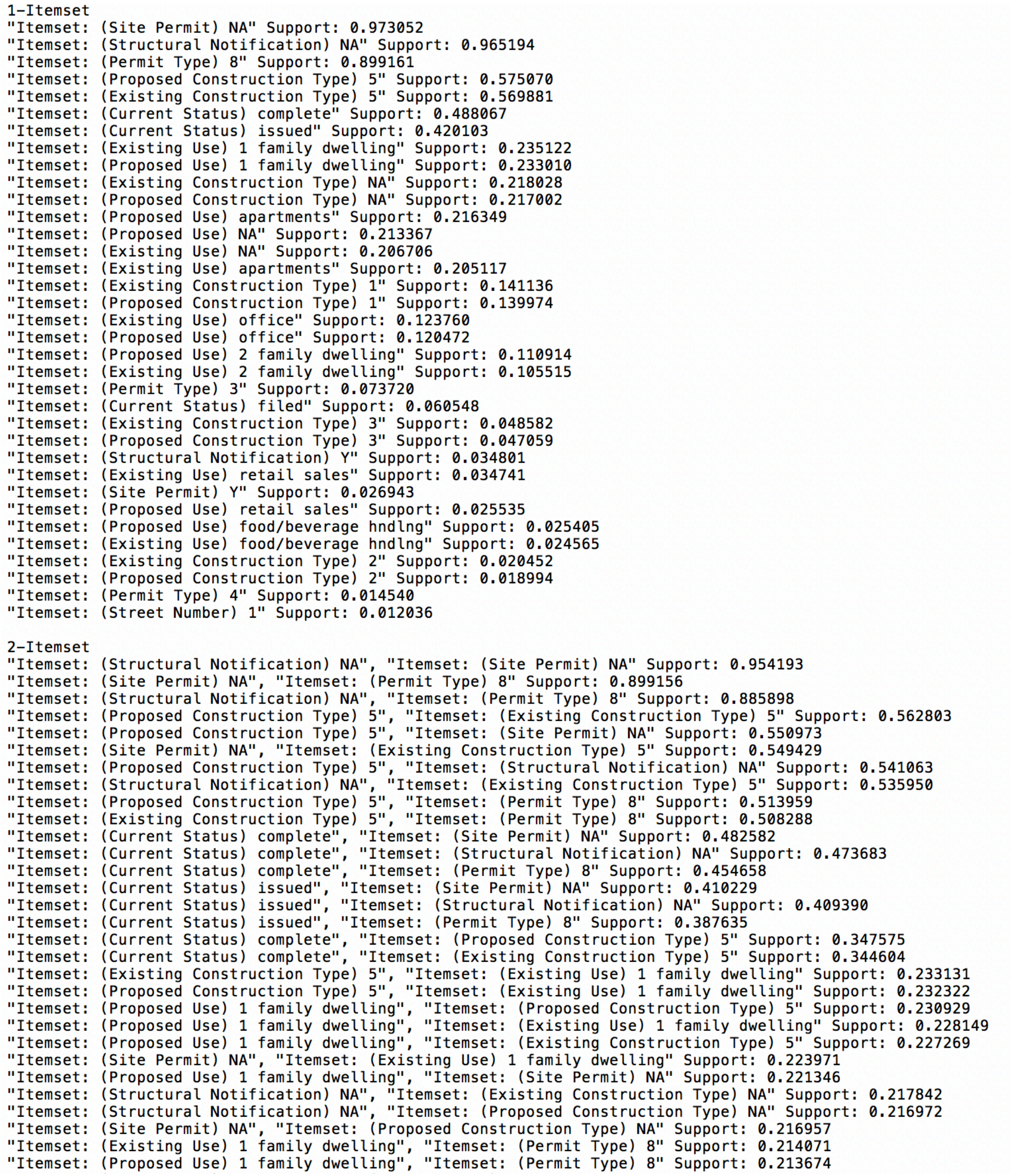
选出这 9 个属性后，由于数据集中存在很多数据的缺失，我们这里首先对缺失数据进行填充，然后剔除。数据预处理代码段如下：



```python
import csv
import pickle
database = './dataset/Building_Permits.csv'
NA = ['NA', 'None', '', 'NONE', 'none', 'Na']


def preprocess():
    dt = []

    with open('./dataset/item', 'r') as fp:
        attr = [int(i) - 1 for i in fp.read().split(' ')]

    with open(database, encoding='utf-8') as fp:
        reader = csv.reader(fp)
        for _, row in enumerate(reader):
            item = []
            if _ == 0:
                name = row
            for i in attr:
                at = row[i]
                if at in NA:
                    at = '(%s) NA' % (name[i])
                else:
                    at = '(%s) %s' % (name[i], row[i])

                item.append(at)

            dt.append(item)
            # print(item)

    with open('./dataset/preprocessed.pkl', 'wb') as fp:
        pickle.dump(dt, fp)


if __name__ == '__main__':
    preprocess()
```

频繁项集

我们使用 Apriori 算法来计算频繁项集，这里我们把支持度阈值设置为 0.01。我们把计算出的频繁项集根据项集的元素个数分别按支持度降序排序，结果存储在 txt 文件中，部分结果如下图所示，最终结果存储在 frequent_items.txt 中



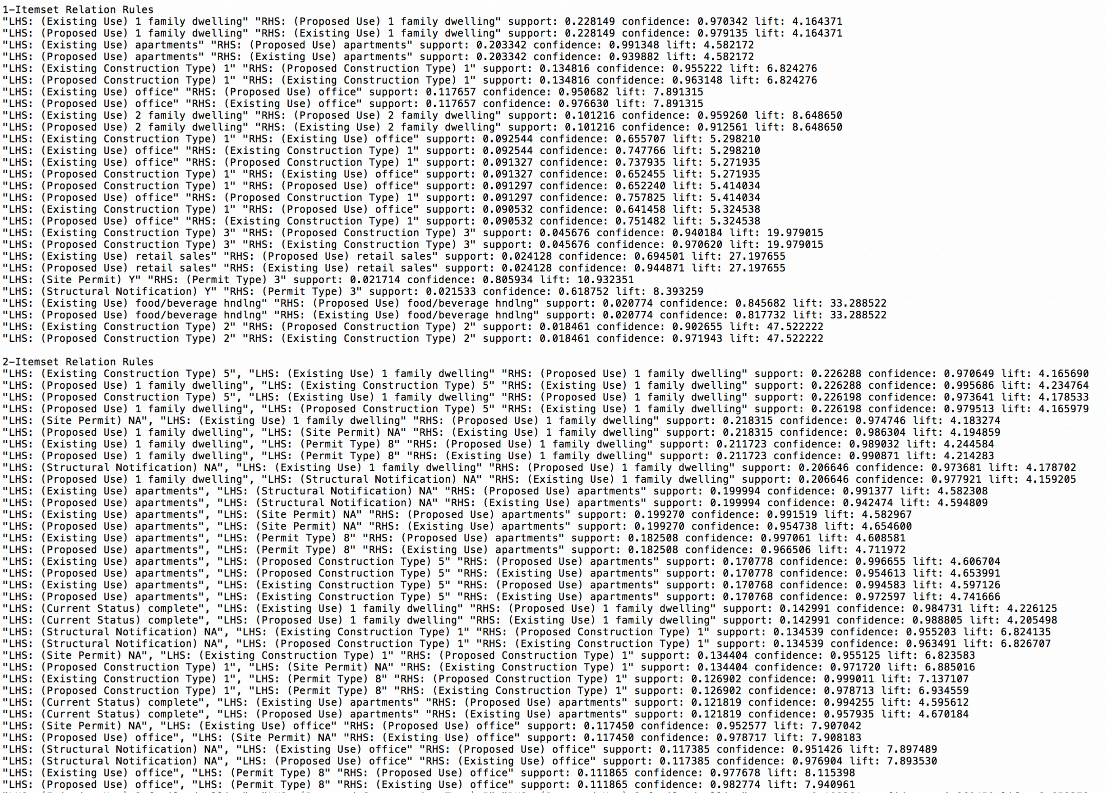
```
1-Itemset
"Itemset: (Site Permit) NA" Support: 0.973052
"Itemset: (Structural Notification) NA" Support: 0.965194
"Itemset: (Permit Type) 8" Support: 0.899161
"Itemset: (Proposed Construction Type) 5" Support: 0.575070
"Itemset: (Existing Construction Type) 5" Support: 0.569881
"Itemset: (Current Status) complete" Support: 0.488067
"Itemset: (Current Status) issued" Support: 0.420103
"Itemset: (Existing Use) 1 family dwelling" Support: 0.235122
"Itemset: (Proposed Use) 1 family dwelling" Support: 0.233010
"Itemset: (Existing Construction Type) NA" Support: 0.218028
"Itemset: (Proposed Construction Type) NA" Support: 0.217002
"Itemset: (Proposed Use) apartments" Support: 0.216349
"Itemset: (Proposed Use) NA" Support: 0.213367
"Itemset: (Existing Use) NA" Support: 0.206706
"Itemset: (Existing Use) apartments" Support: 0.205117
"Itemset: (Existing Construction Type) 1" Support: 0.141136
"Itemset: (Proposed Construction Type) 1" Support: 0.139974
"Itemset: (Existing Use) office" Support: 0.123760
"Itemset: (Proposed Use) office" Support: 0.120472
"Itemset: (Proposed Use) 2 family dwelling" Support: 0.110914
"Itemset: (Existing Use) 2 family dwelling" Support: 0.105515
"Itemset: (Permit Type) 3" Support: 0.073720
"Itemset: (Current Status) filed" Support: 0.060548
"Itemset: (Existing Construction Type) 3" Support: 0.048582
"Itemset: (Proposed Construction Type) 3" Support: 0.047059
"Itemset: (Structural Notification) Y" Support: 0.034801
"Itemset: (Existing Use) retail sales" Support: 0.034741
"Itemset: (Site Permit) Y" Support: 0.026943
"Itemset: (Proposed Use) retail sales" Support: 0.025535
"Itemset: (Proposed Use) food/beverage hndlng" Support: 0.025405
"Itemset: (Existing Use) food/beverage hndlng" Support: 0.024565
"Itemset: (Existing Construction Type) 2" Support: 0.020452
"Itemset: (Proposed Construction Type) 2" Support: 0.018994
"Itemset: (Permit Type) 4" Support: 0.014540
"Itemset: (Street Number) 1" Support: 0.012036

2-Itemset
"Itemset: (Structural Notification) NA", "Itemset: (Site Permit) NA" Support: 0.954193
"Itemset: (Site Permit) NA", "Itemset: (Permit Type) 8" Support: 0.899156
"Itemset: (Structural Notification) NA", "Itemset: (Permit Type) 8" Support: 0.885898
"Itemset: (Proposed Construction Type) 5", "Itemset: (Existing Construction Type) 5" Support: 0.562803
"Itemset: (Proposed Construction Type) 5", "Itemset: (Site Permit) NA" Support: 0.550973
"Itemset: (Site Permit) NA", "Itemset: (Existing Construction Type) 5" Support: 0.549429
"Itemset: (Proposed Construction Type) 5", "Itemset: (Structural Notification) NA" Support: 0.541063
"Itemset: (Structural Notification) NA", "Itemset: (Existing Construction Type) 5" Support: 0.535950
"Itemset: (Proposed Construction Type) 5", "Itemset: (Permit Type) 8" Support: 0.513959
"Itemset: (Existing Construction Type) 5", "Itemset: (Permit Type) 8" Support: 0.508288
"Itemset: (Current Status) complete", "Itemset: (Site Permit) NA" Support: 0.482582
"Itemset: (Current Status) complete", "Itemset: (Structural Notification) NA" Support: 0.473683
"Itemset: (Current Status) complete", "Itemset: (Permit Type) 8" Support: 0.454658
"Itemset: (Current Status) issued", "Itemset: (Site Permit) NA" Support: 0.410229
"Itemset: (Current Status) issued", "Itemset: (Structural Notification) NA" Support: 0.409390
"Itemset: (Current Status) issued", "Itemset: (Permit Type) 8" Support: 0.387635
"Itemset: (Current Status) complete", "Itemset: (Proposed Construction Type) 5" Support: 0.347575
"Itemset: (Current Status) complete", "Itemset: (Existing Construction Type) 5" Support: 0.344604
"Itemset: (Existing Construction Type) 5", "Itemset: (Existing Use) 1 family dwelling" Support: 0.233131
"Itemset: (Proposed Construction Type) 5", "Itemset: (Existing Use) 1 family dwelling" Support: 0.232322
"Itemset: (Proposed Use) 1 family dwelling", "Itemset: (Proposed Construction Type) 5" Support: 0.230929
"Itemset: (Proposed Use) 1 family dwelling", "Itemset: (Existing Use) 1 family dwelling" Support: 0.228149
"Itemset: (Proposed Use) 1 family dwelling", "Itemset: (Existing Construction Type) 5" Support: 0.227269
"Itemset: (Site Permit) NA", "Itemset: (Existing Use) 1 family dwelling" Support: 0.223971
"Itemset: (Proposed Use) 1 family dwelling", "Itemset: (Site Permit) NA" Support: 0.221346
"Itemset: (Structural Notification) NA", "Itemset: (Existing Construction Type) NA" Support: 0.217842
"Itemset: (Structural Notification) NA", "Itemset: (Proposed Construction Type) NA" Support: 0.216972
"Itemset: (Site Permit) NA", "Itemset: (Proposed Construction Type) NA" Support: 0.216957
"Itemset: (Existing Use) 1 family dwelling", "Itemset: (Permit Type) 8" Support: 0.214071
"Itemset: (Proposed Use) 1 family dwelling", "Itemset: (Permit Type) 8" Support: 0.213674
```

关联规则

我们使用支持度、置信度和提升度来度量关联规则。

为了挖掘出强关联信息，我们将置信度阈值设为 0.6，将支持度阈值设为 3。

得到的部分关联规则如下所示，最终结果存储在 relation_rules.txt 中



```
1-Itemset Relation Rules
"LHS: (Existing Use) 1 family dwelling" "RHS: (Proposed Use) 1 family dwelling" support: 0.228149 confidence: 0.970342 lift: 4.164371
"LHS: (Proposed Use) 1 family dwelling" "RHS: (Existing Use) 1 family dwelling" support: 0.228149 confidence: 0.979135 lift: 4.164371
"LHS: (Existing Use) apartments" "RHS: (Proposed Use) apartments" support: 0.203342 confidence: 0.991348 lift: 4.582172
"LHS: (Proposed Use) apartments" "RHS: (Existing Use) apartments" support: 0.203342 confidence: 0.939882 lift: 4.582172
"LHS: (Existing Construction Type) 1" "RHS: (Proposed Construction Type) 1" support: 0.134816 confidence: 0.955222 lift: 6.824276
"LHS: (Proposed Construction Type) 1" "RHS: (Existing Construction Type) 1" support: 0.134816 confidence: 0.963148 lift: 6.824276
"LHS: (Existing Use) office" "RHS: (Proposed Use) office" support: 0.117657 confidence: 0.950682 lift: 7.891315
"LHS: (Proposed Use) office" "RHS: (Existing Use) office" support: 0.117657 confidence: 0.976630 lift: 7.891315
"LHS: (Existing Use) 2 family dwelling" "RHS: (Proposed Use) 2 family dwelling" support: 0.101216 confidence: 0.959260 lift: 8.648650
"LHS: (Proposed Use) 2 family dwelling" "RHS: (Existing Use) 2 family dwelling" support: 0.101216 confidence: 0.912561 lift: 8.648650
"LHS: (Existing Construction Type) 1" "RHS: (Existing Use) office" support: 0.092544 confidence: 0.655707 lift: 5.298210
"LHS: (Existing Use) office" "RHS: (Existing Construction Type) 1" support: 0.092544 confidence: 0.747766 lift: 5.298210
"LHS: (Existing Use) office" "RHS: (Proposed Construction Type) 1" support: 0.091327 confidence: 0.737935 lift: 5.271935
"LHS: (Proposed Construction Type) 1" "RHS: (Existing Use) office" support: 0.091327 confidence: 0.652455 lift: 5.271935
"LHS: (Proposed Construction Type) 1" "RHS: (Proposed Use) office" support: 0.091297 confidence: 0.652240 lift: 5.414034
"LHS: (Proposed Use) office" "RHS: (Proposed Construction Type) 1" support: 0.091297 confidence: 0.757825 lift: 5.414034
"LHS: (Existing Construction Type) 1" "RHS: (Proposed Use) office" support: 0.090532 confidence: 0.641458 lift: 5.324538
"LHS: (Proposed Use) office" "RHS: (Existing Construction Type) 1" support: 0.090532 confidence: 0.751482 lift: 5.324538
"LHS: (Existing Construction Type) 3" "RHS: (Proposed Construction Type) 3" support: 0.045676 confidence: 0.940184 lift: 19.979015
"LHS: (Proposed Construction Type) 3" "RHS: (Existing Construction Type) 3" support: 0.045676 confidence: 0.970620 lift: 19.979015
"LHS: (Existing Use) retail sales" "RHS: (Proposed Use) retail sales" support: 0.024128 confidence: 0.694501 lift: 27.197655
"LHS: (Proposed Use) retail sales" "RHS: (Existing Use) retail sales" support: 0.024128 confidence: 0.944871 lift: 27.197655
"LHS: (Site Permit) Y" "RHS: (Permit Type) 3" support: 0.021714 confidence: 0.805934 lift: 10.932351
"LHS: (Structural Notification) Y" "RHS: (Permit Type) 3" support: 0.021533 confidence: 0.618752 lift: 8.393259
"LHS: (Existing Use) food/beverage hndlng" "RHS: (Proposed Use) food/beverage hndlng" support: 0.020774 confidence: 0.845682 lift: 33.288522
"LHS: (Proposed Use) food/beverage hndlng" "RHS: (Existing Use) food/beverage hndlng" support: 0.020774 confidence: 0.817732 lift: 33.288522
"LHS: (Existing Construction Type) 2" "RHS: (Proposed Construction Type) 2" support: 0.018461 confidence: 0.902655 lift: 47.522222
"LHS: (Proposed Construction Type) 2" "RHS: (Existing Construction Type) 2" support: 0.018461 confidence: 0.971943 lift: 47.522222

2-Itemset Relation Rules
"LHS: (Existing Construction Type) 5", "LHS: (Existing Use) 1 family dwelling" "RHS: (Proposed Use) 1 family dwelling" support: 0.226288 confidence: 0.970649 lift: 4.165690
"LHS: (Proposed Use) 1 family dwelling", "LHS: (Existing Construction Type) 5" "RHS: (Existing Use) 1 family dwelling" support: 0.226288 confidence: 0.995686 lift: 4.234764
"LHS: (Proposed Construction Type) 5", "LHS: (Existing Use) 1 family dwelling" "RHS: (Proposed Use) 1 family dwelling" support: 0.226198 confidence: 0.973641 lift: 4.178533
"LHS: (Proposed Use) 1 family dwelling", "LHS: (Proposed Construction Type) 5" "RHS: (Existing Use) 1 family dwelling" support: 0.226198 confidence: 0.979513 lift: 4.165979
"LHS: (Site Permit) NA", "LHS: (Existing Use) 1 family dwelling" "RHS: (Proposed Use) 1 family dwelling" support: 0.218315 confidence: 0.974746 lift: 4.183274
"LHS: (Proposed Use) 1 family dwelling", "LHS: (Site Permit) NA" "RHS: (Existing Use) 1 family dwelling" support: 0.218315 confidence: 0.986304 lift: 4.194859
"LHS: (Existing Use) 1 family dwelling", "LHS: (Permit Type) 8" "RHS: (Proposed Use) 1 family dwelling" support: 0.211723 confidence: 0.989032 lift: 4.244584
"LHS: (Proposed Use) 1 family dwelling", "LHS: (Permit Type) 8" "RHS: (Existing Use) 1 family dwelling" support: 0.211723 confidence: 0.990871 lift: 4.214283
"LHS: (Structural Notification) NA", "LHS: (Existing Use) 1 family dwelling" "RHS: (Proposed Use) 1 family dwelling" support: 0.206646 confidence: 0.973681 lift: 4.178702
"LHS: (Proposed Use) 1 family dwelling", "LHS: (Structural Notification) NA" "RHS: (Existing Use) 1 family dwelling" support: 0.206646 confidence: 0.977921 lift: 4.159205
"LHS: (Existing Use) apartments", "LHS: (Structural Notification) NA" "RHS: (Proposed Use) apartments" support: 0.199994 confidence: 0.991377 lift: 4.582308
"LHS: (Proposed Use) apartments", "LHS: (Structural Notification) NA" "RHS: (Existing Use) apartments" support: 0.199994 confidence: 0.942474 lift: 4.594809
"LHS: (Existing Use) apartments", "LHS: (Site Permit) NA" "RHS: (Proposed Use) apartments" support: 0.199270 confidence: 0.991519 lift: 4.582967
"LHS: (Proposed Use) apartments", "LHS: (Site Permit) NA" "RHS: (Existing Use) apartments" support: 0.199270 confidence: 0.954738 lift: 4.654600
"LHS: (Existing Use) apartments", "LHS: (Permit Type) 8" "RHS: (Proposed Use) apartments" support: 0.182508 confidence: 0.997061 lift: 4.608581
"LHS: (Proposed Use) apartments", "LHS: (Permit Type) 8" "RHS: (Existing Use) apartments" support: 0.182508 confidence: 0.966506 lift: 4.711972
"LHS: (Existing Use) apartments", "LHS: (Proposed Construction Type) 5" "RHS: (Proposed Use) apartments" support: 0.170778 confidence: 0.996655 lift: 4.606704
"LHS: (Proposed Use) apartments", "LHS: (Proposed Construction Type) 5" "RHS: (Existing Use) apartments" support: 0.170778 confidence: 0.954613 lift: 4.653991
"LHS: (Existing Use) apartments", "LHS: (Existing Construction Type) 5" "RHS: (Proposed Use) apartments" support: 0.170768 confidence: 0.994583 lift: 4.597126
"LHS: (Proposed Use) apartments", "LHS: (Existing Construction Type) 5" "RHS: (Existing Use) apartments" support: 0.170768 confidence: 0.972597 lift: 4.741666
"LHS: (Current Status) complete", "LHS: (Existing Use) 1 family dwelling" "RHS: (Proposed Use) 1 family dwelling" support: 0.142991 confidence: 0.984731 lift: 4.226125
"LHS: (Current Status) complete", "LHS: (Proposed Use) 1 family dwelling" "RHS: (Existing Use) 1 family dwelling" support: 0.142991 confidence: 0.988805 lift: 4.205498
"LHS: (Structural Notification) NA", "LHS: (Existing Construction Type) 1" "RHS: (Proposed Construction Type) 1" support: 0.134539 confidence: 0.955203 lift: 6.824135
"LHS: (Structural Notification) NA", "LHS: (Proposed Construction Type) 1" "RHS: (Existing Construction Type) 1" support: 0.134539 confidence: 0.963491 lift: 6.826707
"LHS: (Site Permit) NA", "LHS: (Existing Construction Type) 1" "RHS: (Proposed Construction Type) 1" support: 0.134404 confidence: 0.955125 lift: 6.823583
"LHS: (Proposed Construction Type) 1", "LHS: (Site Permit) NA" "RHS: (Existing Construction Type) 1" support: 0.134404 confidence: 0.971720 lift: 6.885016
"LHS: (Existing Construction Type) 1", "LHS: (Permit Type) 8" "RHS: (Proposed Construction Type) 1" support: 0.126902 confidence: 0.999011 lift: 7.137107
"LHS: (Proposed Construction Type) 1", "LHS: (Permit Type) 8" "RHS: (Existing Construction Type) 1" support: 0.126902 confidence: 0.978713 lift: 6.934559
"LHS: (Current Status) complete", "LHS: (Existing Use) apartments" "RHS: (Proposed Use) apartments" support: 0.121819 confidence: 0.994255 lift: 4.595612
"LHS: (Current Status) complete", "LHS: (Proposed Use) apartments" "RHS: (Existing Use) apartments" support: 0.121819 confidence: 0.957935 lift: 4.670184
"LHS: (Site Permit) NA", "LHS: (Existing Use) office" "RHS: (Proposed Use) office" support: 0.117450 confidence: 0.952577 lift: 7.907042
"LHS: (Proposed Use) office", "LHS: (Site Permit) NA" "RHS: (Existing Use) office" support: 0.117450 confidence: 0.978717 lift: 7.908183
"LHS: (Structural Notification) NA", "LHS: (Existing Use) office" "RHS: (Proposed Use) office" support: 0.117385 confidence: 0.951426 lift: 7.897489
"LHS: (Structural Notification) NA", "LHS: (Proposed Use) office" "RHS: (Existing Use) office" support: 0.117385 confidence: 0.976904 lift: 7.893530
"LHS: (Existing Use) office", "LHS: (Permit Type) 8" "RHS: (Proposed Use) office" support: 0.111865 confidence: 0.977678 lift: 8.115398
"LHS: (Proposed Use) office", "LHS: (Permit Type) 8" "RHS: (Existing Use) office" support: 0.111865 confidence: 0.982774 lift: 7.940961
```

结果分析

通过对关联规则分析，得到以下结论：

1. 建筑物的 Existing Use 与 Proposed Use 基本保持一致
2. 建筑物的 Existing Construction Type 与 Proposed Construction Type 基本保持一致
3. permit type 8 出现次数最多，且对应的建筑物种类最多
4. current status 为 issued 时，对应的 Proposed Use 多为 Office 或者 family dwelling
5. 属性 2，4，9 不存在明显规则