# 数据挖掘：关联规则挖掘

姓名：相铮 学号：2620170056

一、数据集预处理

数据集：Building\_Permits.csv

数据集预处理：首先对数据集进行预处理，对数据进行部分处理，提高数据处理速度，且有利于关联规则挖掘。

预处理代码

# coding=utf-8

import pandas as pd

obj = pd.read\_csv("e:/Building\_Permits.csv",low\_memory=False)

attr=[]

for item in obj.columns:

n = obj[item].value\_counts()

if n.count() < 100:

attr.append(item)

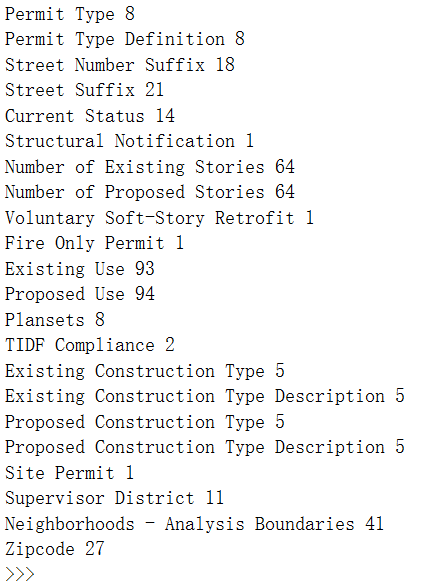
print(item,n.count())

df = obj[attr]

df.to\_csv("./test.csv",index=False,header=False)

处理结果：

保留属性如下：



生成测试集见文件夹：test.csv

1. 找出频繁项集

读取在第一步中处理好的文件，然后使用apriori算法找出频繁项集，设置的最小支持度为0.6。

发现关联规则要求项集必须满足的最小支持阈值，称为项集的最小支持度，记为minSupport。支持度大于或等于minSupport的项集称为频繁项集，简称频繁集，反之则称为非频繁集。通常k-项集如果满足minSupport，称为k-频繁集，记作Lk。关联规则的最小置信度记为minConfidence，它表示关联规则需要满足的最低可靠性。

1. 导出关联规则，计算其支持度和置信度

1.支持度（Support）

支持度表示项集{X,Y}在总项集里出现的概率。公式为：

Support(X→Y) = P(X,Y) / P(I) = P(X∪Y) / P(I) = num(XUY) / num(I)

其中，I表示总事务集。num()表示求事务集里特定项集出现的次数。

比如，num(I)表示总事务集的个数num(X∪Y)表示含有{X,Y}的事务集的个数（个数也叫次数）。

2.置信度 （Confidence）

置信度表示在先决条件X发生的情况下，由关联规则”X→Y“推出Y的概率。即在含有X的项集中，含有Y的可能性，公式为：

Confidence(X→Y) = P(Y|X) = P(X,Y) / P(X) = P(XUY) / P(X)

3.Apriori算法：使用候选项集找频繁项集

Apriori算法是一种最有影响的挖掘布尔关联规则频繁项集的算法。其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则。在这里，所有支持度大于最小支持度的项集称为频繁项集，简称频集。

该算法的基本思想是：首先找出所有的频集，这些项集出现的频繁性至少和预定义的最小支持度一样。然后由频集产生强关联规则，这些规则必须满足最小支持度和最小可信度。然后使用第1步找到的频集产生期望的规则，产生只包含集合的项的所有规则，其中每一条规则的右部只有一项，这里采用的是中规则的定义。一旦这些规则被生成，那么只有那些大于用户给定的最小可信度的规则才被留下来。为了生成所有频集，使用了递推的方法。

1. 关联规则评价及分析

1.提升度（Lift）

提升度表示含有X的条件下，同时含有Y的概率，与Y总体发生的概率之比。

Lift(X→Y) = P(Y|X) / P(Y)

实验中设置的最小支持度是0.15，最小置信度0.6。

结果见result.txt。