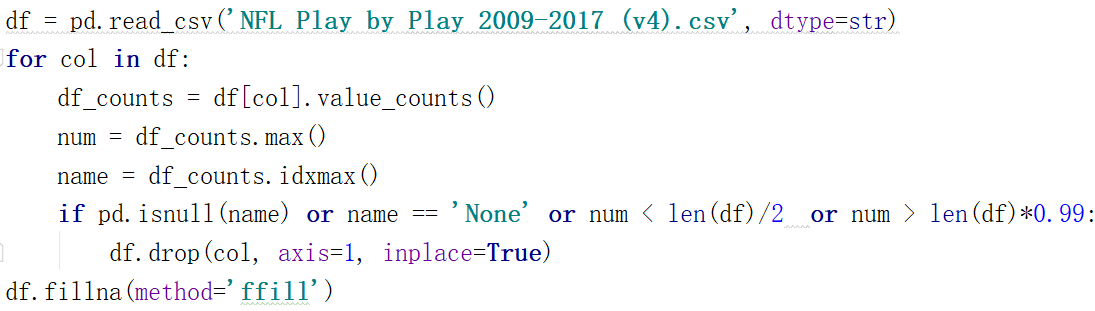
**关联规则挖掘报告**

王鹏 2120171069

一、对数据集进行处理

本次作业选用数据集NFL Play by Play 2009-2017 (v4)，利用Python中的pandas库进行csv数据文件的读取，对数据集进行处理，转换成适合关联规则挖掘的形式：



对于数据集中存在过多“NA”和“None”的字段进行剔除，众数小于文件长度一半的字段进行剔除，并且，对于大多数值都为同一个的字段（例如，对于PlayAttempted属性，它的值全为1），我们认为其没有研究意义，应将其剔除。然后用ffill方法对缺失值进行简单的填充处理

二、找出频繁项集

Apriori算法的两个输入参数分别是最小支持度和数据集。该算法首先会生成所有单个数据字段值的项集列表。接着扫描每行数据来查看哪些项集满足最小支持度要求，那些不满足最小支持度的集合会被去掉。然后，对剩下来的集合进行组合以生成包含两个元素的项集。接下来，再重新扫描每行记录，去掉不满足最小支持度的项集。该过程重复进行直到所有项集都被去掉。

2.1 生成候选项集



2.2 生成频繁项集

生成频繁项级算法的伪代码如下：

当集合中项的个数大于0时

构建一个k个项组成的候选项集的列表

检查数据以确认每个项集都是频繁的

保留频繁项集并构建k+1项组成的候选项集的列表

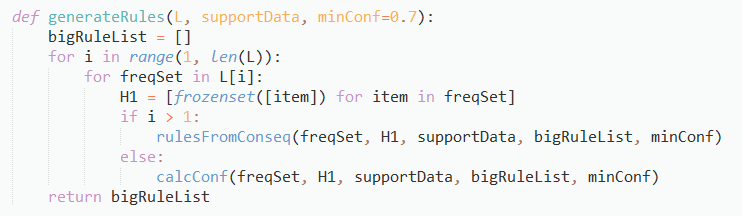
具体相关方法如下：

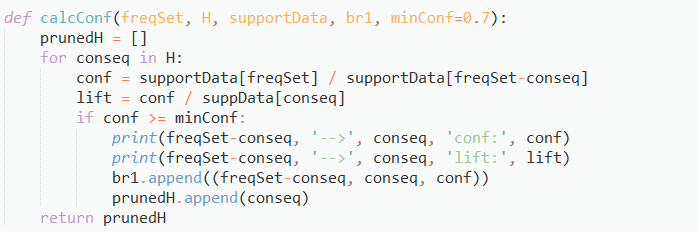


函数最后返回的是频繁项集列表以及每个频繁项的支持度。

三、导出关联规则，计算置信度和提升度

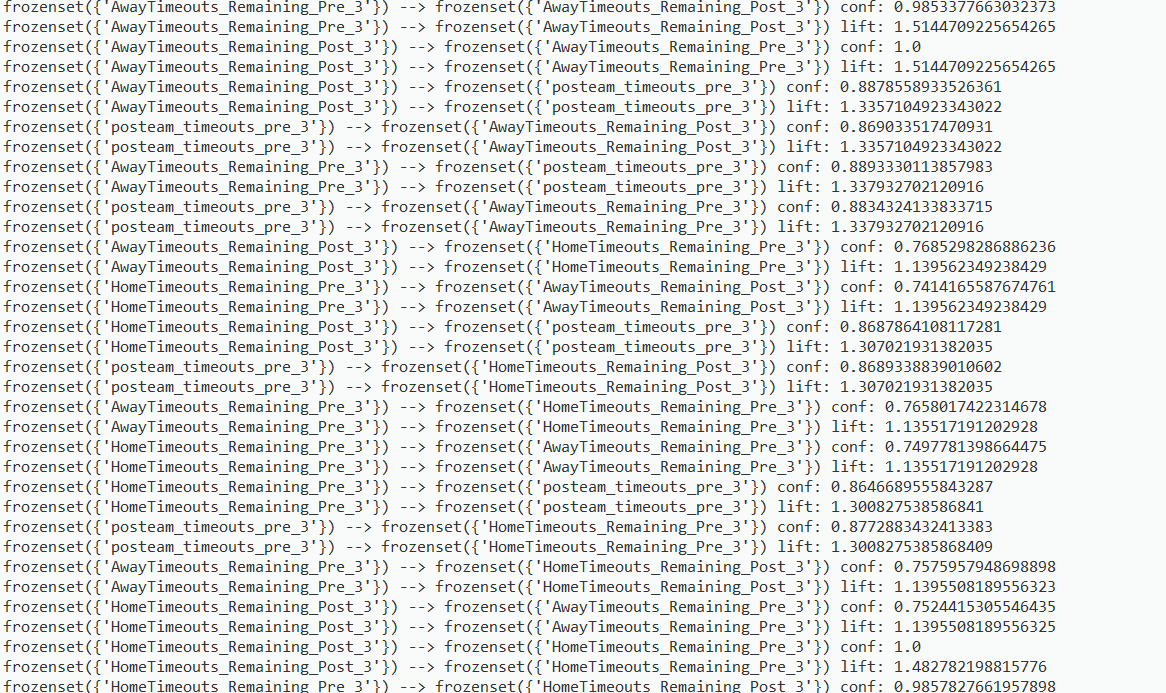
根据频繁集学习关联规则，针对规则右部的元素个数进行分级，导出关联规则，并计算关联规则的提升度和置信度：





四、规则挖掘结果及评价

下图为部分规则挖掘结果及使用Lift评价的结果：



根据这些挖掘到的规则的置信度结果和提升度（Lift）结果，可知置信度较高，且提升度均大于1，则可认为这些规则有用。