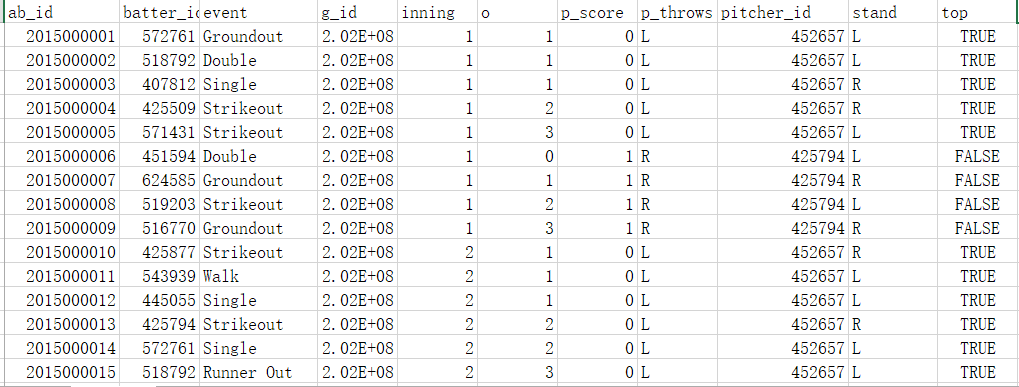
## 关联规则挖掘

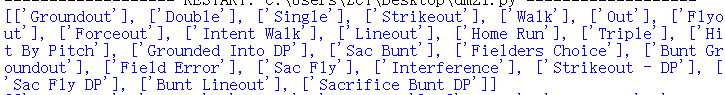
数据集：MLB Pitch Data 2015-2018（atbats.csv）

**1.数据预处理**

原始数据如下，共740390条，可以看出其中并没有项集：



将数据中每位投手的一轮操作中的事件作为一个集合，即event作为项，每次pitcher\_id变换则进入下一个项集。由于event种类较多，导致计算机运行过于缓慢，将部分相关event合并,如‘Catcher Interference’与‘Batter Interference’合并为‘Interference’，共整合为24种如下：

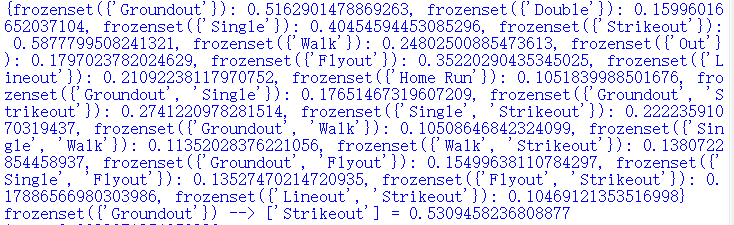


将数据整理为列表格式并将前50项展示如下：

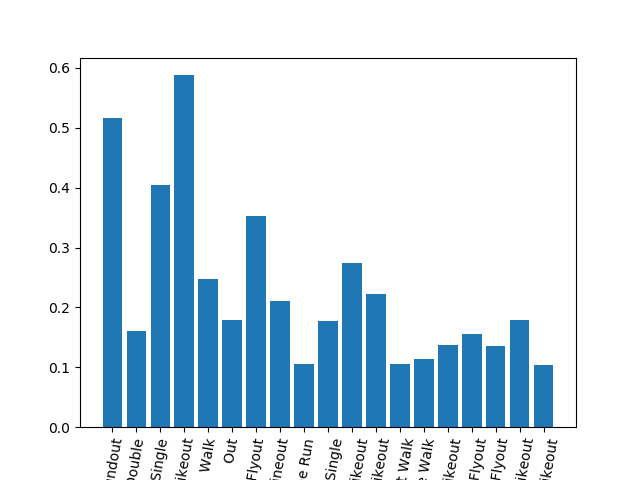


**2.关联规则挖掘**

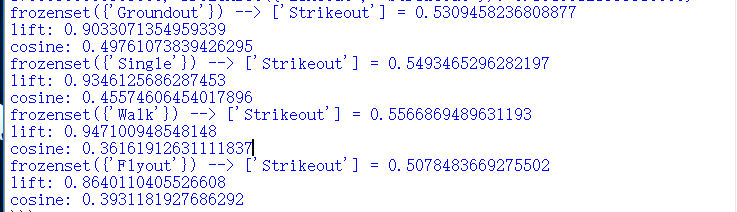
**使用aprior算法，由于数据量过大导致较为稀疏，设定支持度为0.1，置信度为0.5，得到频繁项集及其支持度如下：**

****

**作出满足要求的频繁项集的支持度的直方图如下：**

****

**对频繁二项及以上项集计算提升度与余弦度量，结果如下：**

****