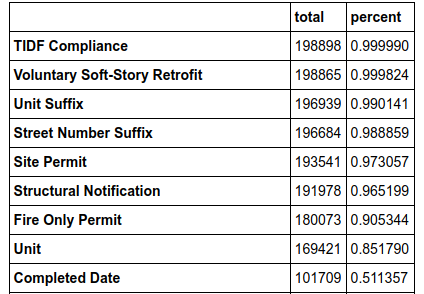
**关联规则挖掘**

数据源：San Francisco Building Permits

1. **数据处理**

为了便于从数据集中挖掘出关联规则，需要对数据集进行清洗，将其转换为适合挖掘的形式，重点处理缺失值。

首先统计每列特征包含缺失值的数量和比例，按照缺失情况从高至低排序，下表截取缺失比例最高的部分属性。将缺失比例大于80%的属性直接删除（“Unit”等8个属性）。



根据数据之间的相关性，来填补部分属性的部分缺失值。例如，在某行数据中，“Number of Existing Stories”属性缺失，并且“Number of Existing Stories”属性没有缺失，可以使用“Number of Existing Stories”的值来填补“Number of Existing Stories”的缺失值。下表中同一行的两个属性之间可以相互填补缺失值。

|  |  |
| --- | --- |
| Number of Existing Stories | Number of Existing Stories |
| Existing Use | Proposed Use |
| Existing Units | Proposed Units |
| Revised Cost | Estimated Cost |

对于下表中的属性，使用众数来填补属性的缺失值。

|  |
| --- |
| Existing Construction Type |
| Existing Construction Type Description |
| Proposed Construction Type |
| Proposed Construction Type Description |
| Number of Proposed Stories |
| Number of Existing Stories |
| Existing Use |
| Proposed Use |
| Existing Units |
| Proposed Units |

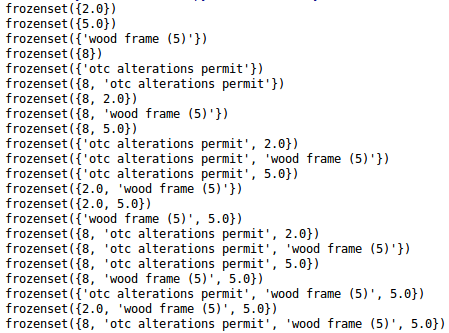
对于下表中的属性，将剩余缺失值填补为设置好空值，或者零值。

|  |  |
| --- | --- |
| Permit Expiration Date | NoDate |
| First Construction Document Date | NoDate |
| Completed Date | NoDate |
| Issued Date | NoDate |
| Street Suffix | NoSuffix |
| Description | NoDescrip |
| Estimated Cost | 0 |
| Revised Cost | 0 |
| Plansets | 0 |

经过以上处理，最后再删除仍旧包含NAN值的数据。最终处理完毕的数据集大小为197174 × 35，保存清洗完毕的数据。

1. **频繁项集**

设置频繁项集的最小置信度阈值为0.7、最小支持度阈值为0.7，挖掘出的频繁项集结果为：



1. **关联规则**

设置关联规则的最小置信度阈值为0.9，导出的关联规则为：

