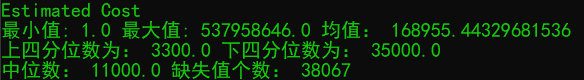
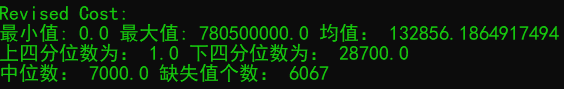
数据挖掘第一次作业报告

数据集1

数据集2 Building\_permit.csv\

数值属性分析





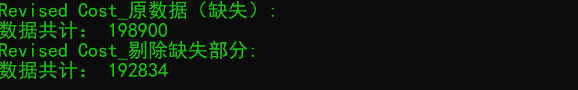


3.2 数据缺失的处理

观察数据集中缺失数据，分析其缺失的原因。

缺失的数据属性中的：Street Number Suffix, Unit, Unit suffix, Completed Date, Structural Notification, Voluntary Soft-Story, Fire Only Permit, TIDF Compliance。这些数据几乎全部缺失，可以推断这些数据不是每条数据中的重要属性，或者说是这些属性数据与其他属性所包含的信息有关，是作为补充说明的数据，可以不记录与数据集中。

其中，数据属性中的Revised Cost 数据是有一部分缺失值，缺失的原因有可能是在数据复制备份的时候缺失，或者是数据处理和访问错误，或者是人为的操作失误。而我认为这个属性包含的是重要的数据，所以我会用以下方法来处理Revised Cost属性的缺失值。



分别使用下列四种策略对缺失值进行处理:

将缺失部分剔除

用最高频率值来填补缺失值

通过属性的相关关系来填补缺失值

通过数据对象之间的相似性来填补缺失值