**数据挖掘**

**--作业一实验报告**

学 院：自动化学院

专 业：系统工程

年 级：2017 级

姓 名：顾炎极

学 号：2620170030

提交日期：2018/4/14

**一、实验目的**

### 根据现有的数据集，获得数据的统计摘要和可视化结果。

### 用不同的策略对数据中的缺失值进行处理，并可视化地对比新旧数据集。

**二、运行环境**

* Windows 10
* Matlab2017a

**三、实验步骤及内容**

1. 准备工作

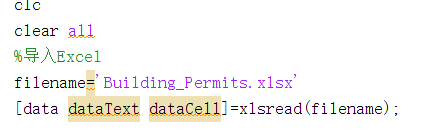
所需数据集共有两个，数据集1为NFL Play-by-Play 2009-2017，数据集2为San Francisco Building Permits。共368个样本，27个特征。

下文实验步骤的描述将以数据集1为例，数据集2类似类似。

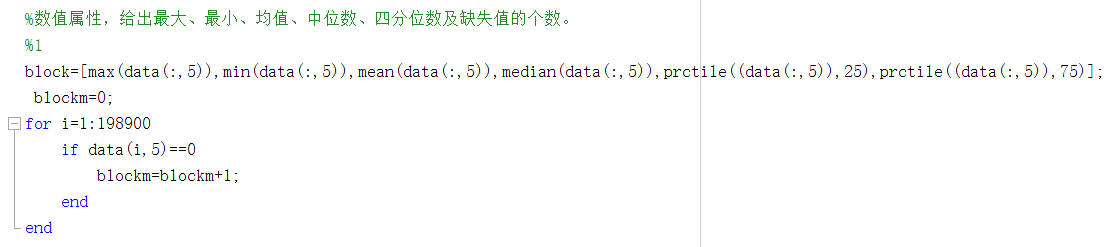
1. 数据可视化和摘要

（1）数据摘要

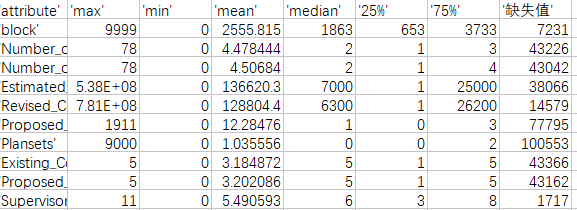
* 加载数据集



* 利用matlab获取数据的数值属性

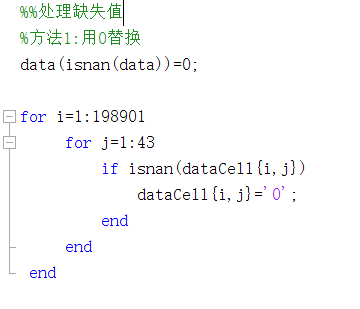


结果如下图，内容包括最大（Max）、最小（Min）、均值（Mean）、中位数（Median）、四分位数（25%.和75%.）及缺失值的个数。

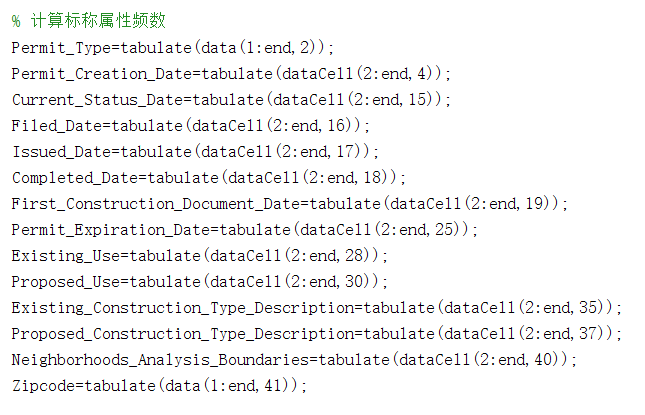


* 利用matlab获取数据的标称属性频数

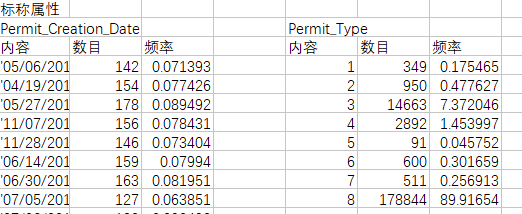
由于标称属性中含有较多的缺失值，为方便统计，首先将缺失值替换为0



再选择出数据的标称属性，如下图所示



最终由matlab的cell函数得到频数统计结果，详细结果在附件中

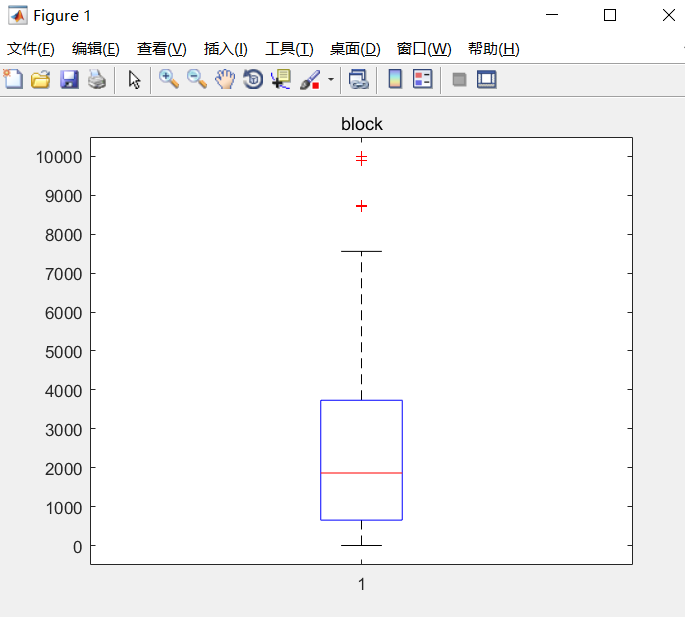
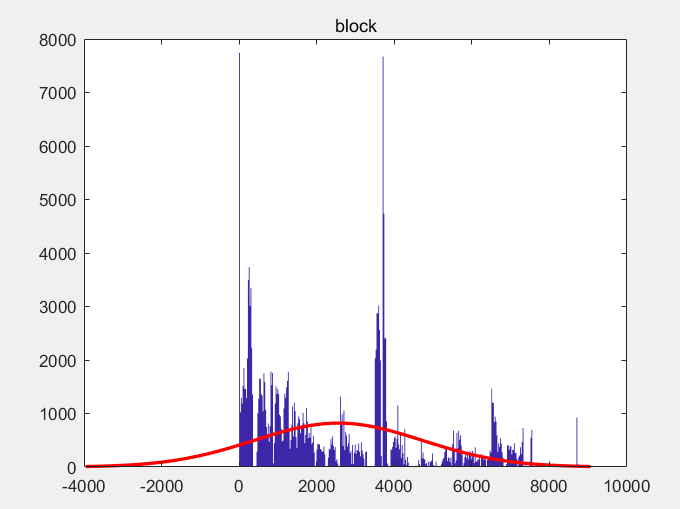


1. 数据可视化

* 绘制直方图利用hisfit语句绘制直方图，首先我们需要选出数值属性，共有10个，分别为：'block'，'Number\_of\_Existing\_Stories'，'Number\_of\_Proposed\_Stories'，'Estimated\_Cost'，'Revised\_Cost'，'Proposed\_Units'，'Plansets'，'Existing\_Construction\_Type'，'Proposed\_Construction\_Type'，'Supervisor\_District'。
* 以绘制block，Number\_of\_Existing\_Stories，Number\_of\_Proposed\_Stories属性的直方图和柱形图为例，语句如下

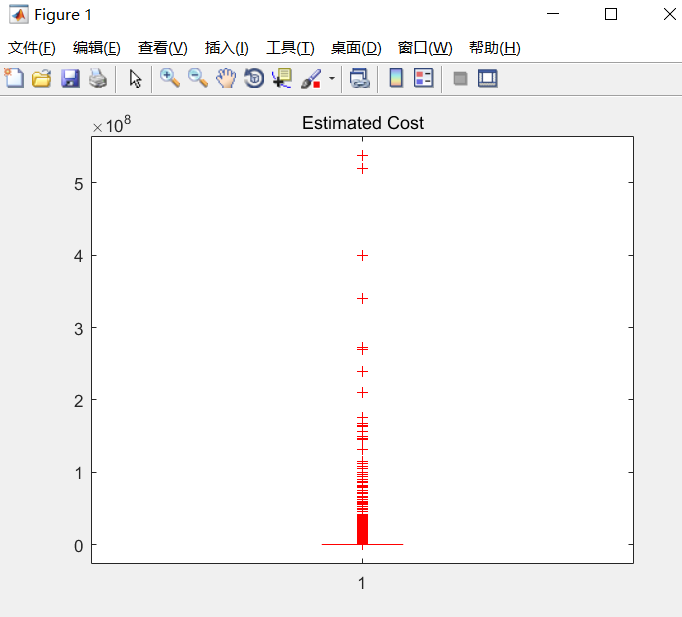
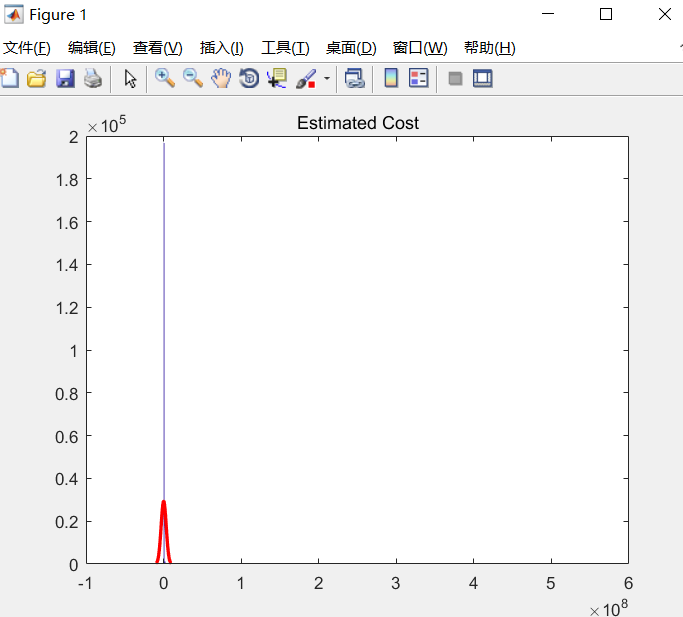
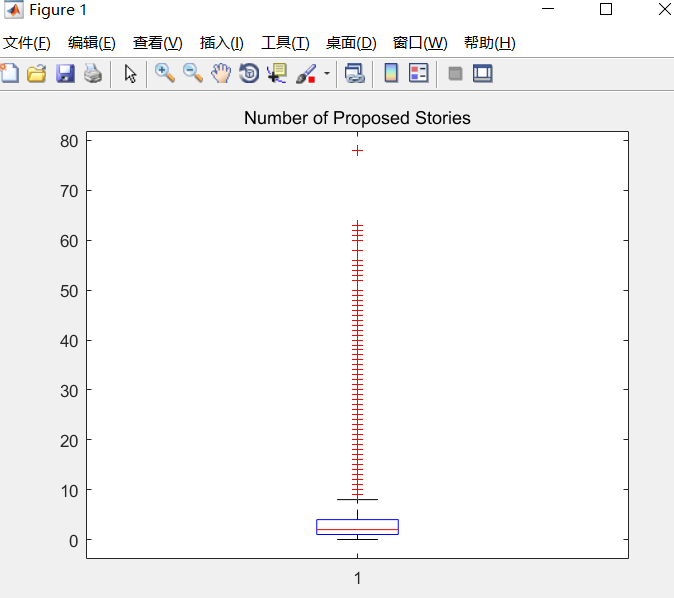
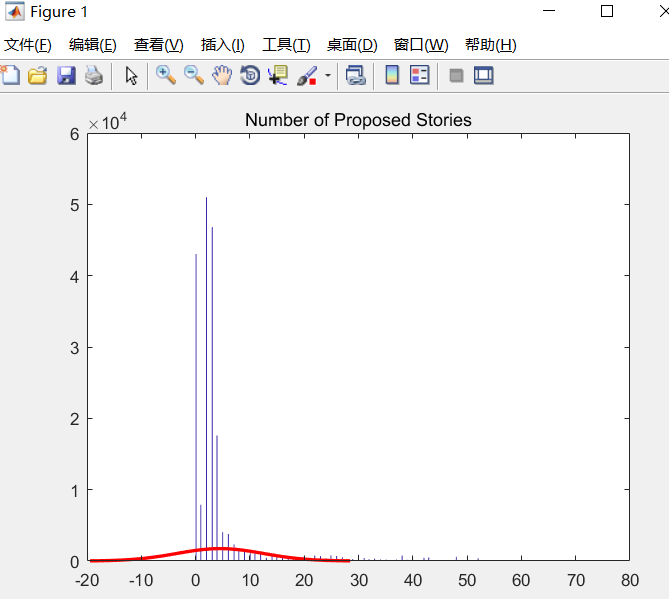
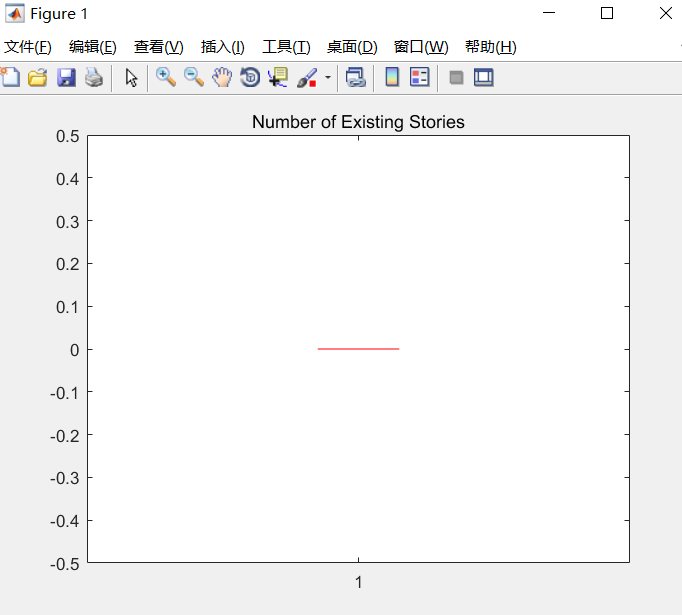
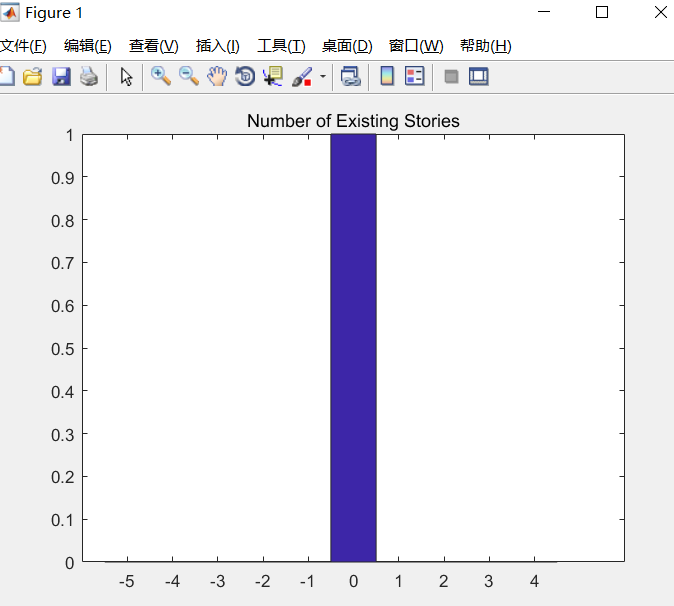


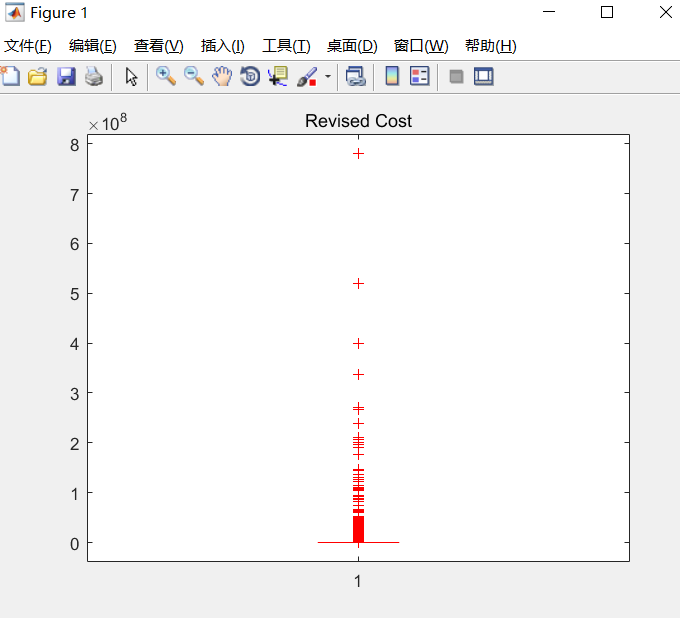
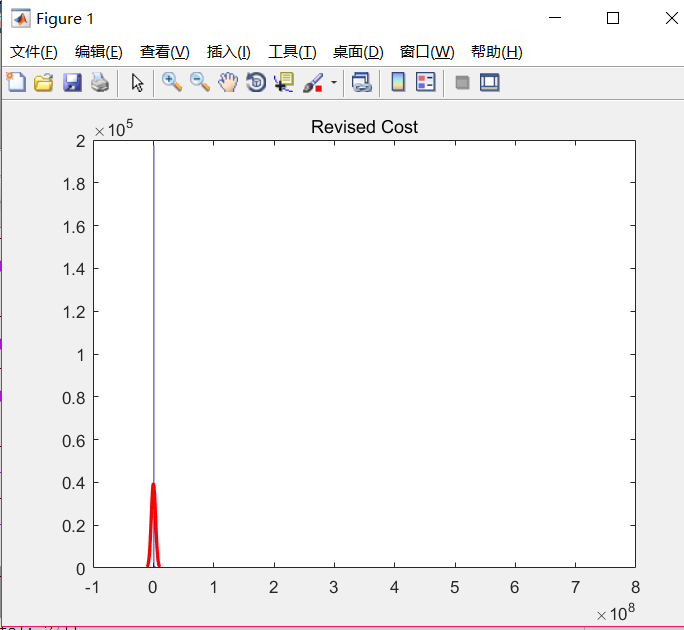
block属性的直方图有和盒图结果如下

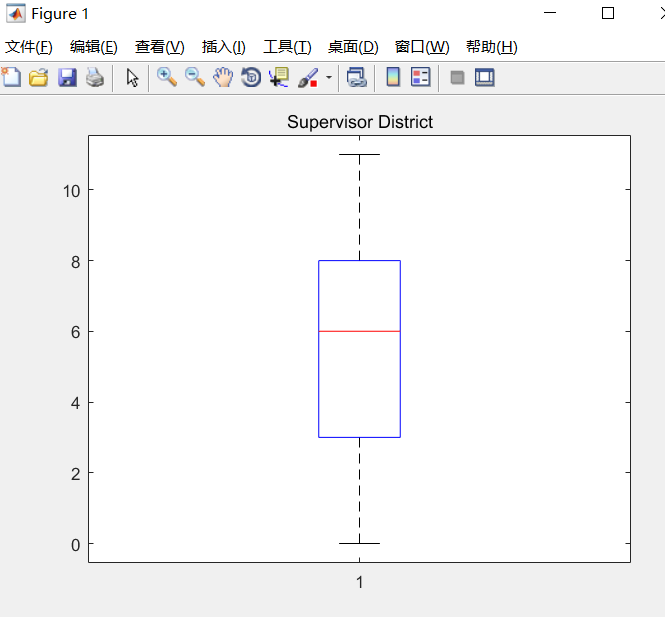
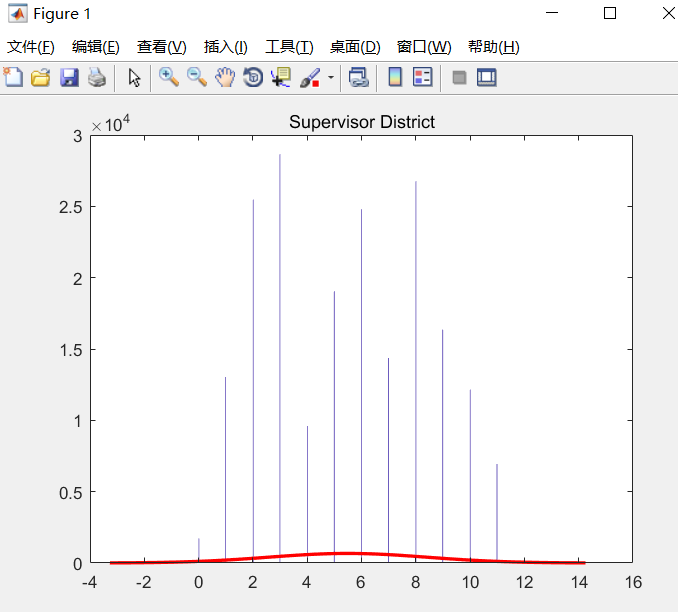
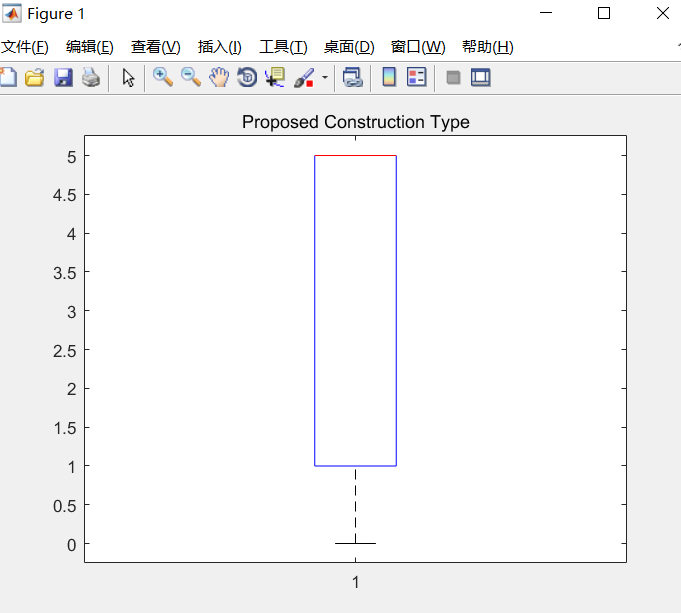
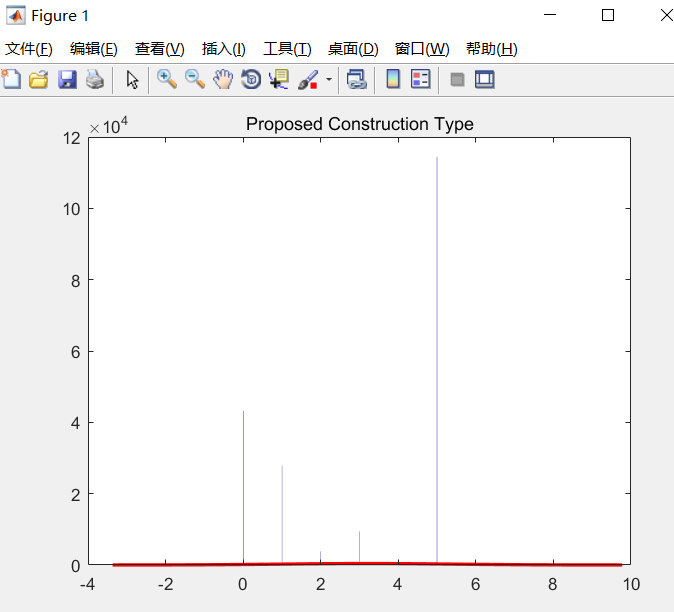
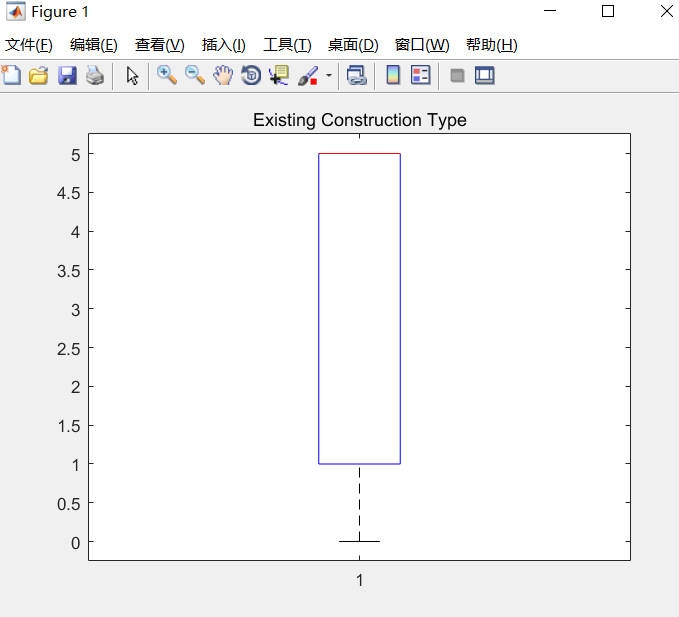
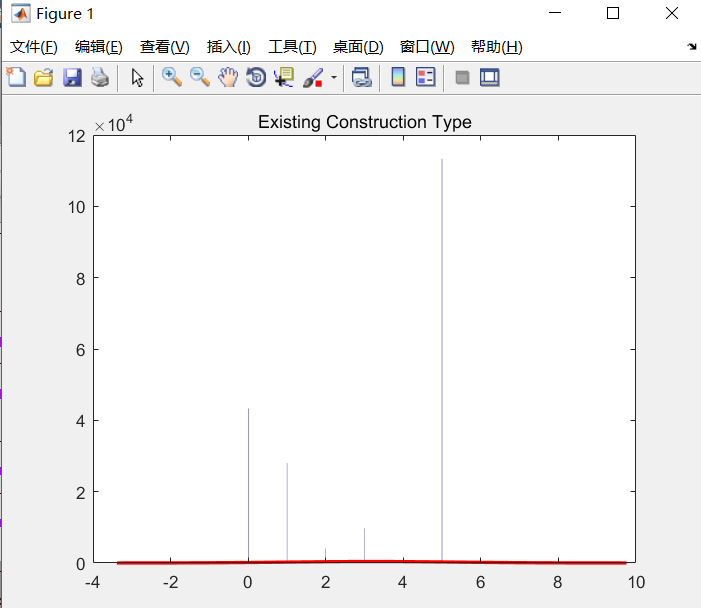
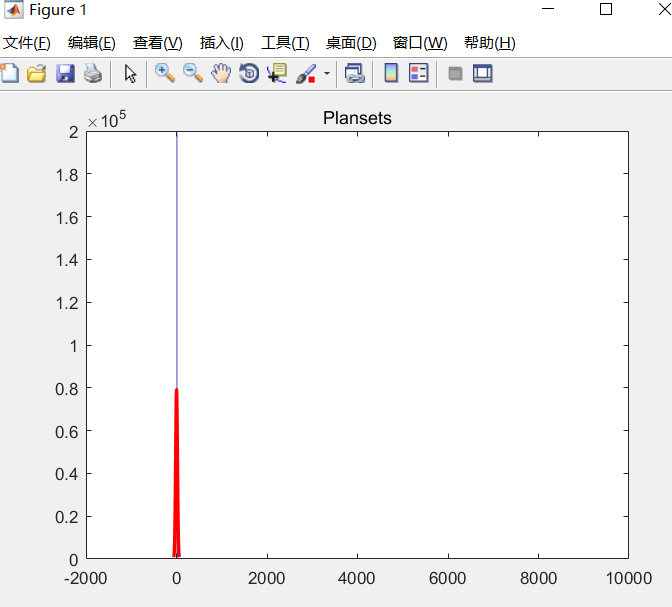
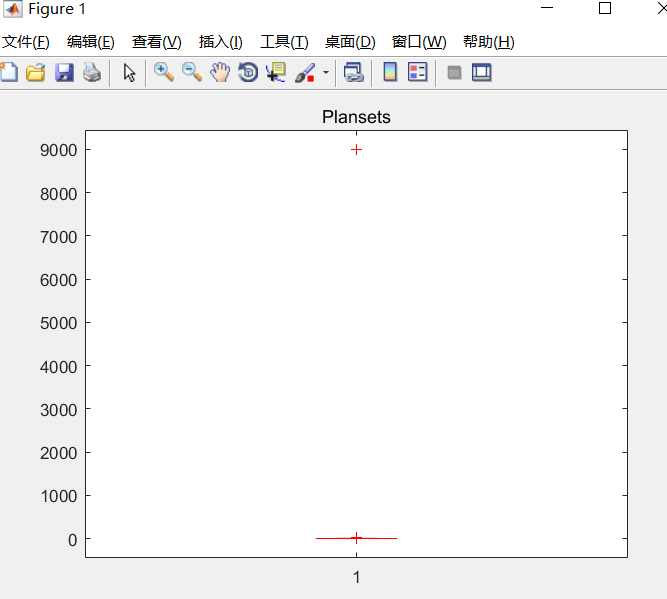
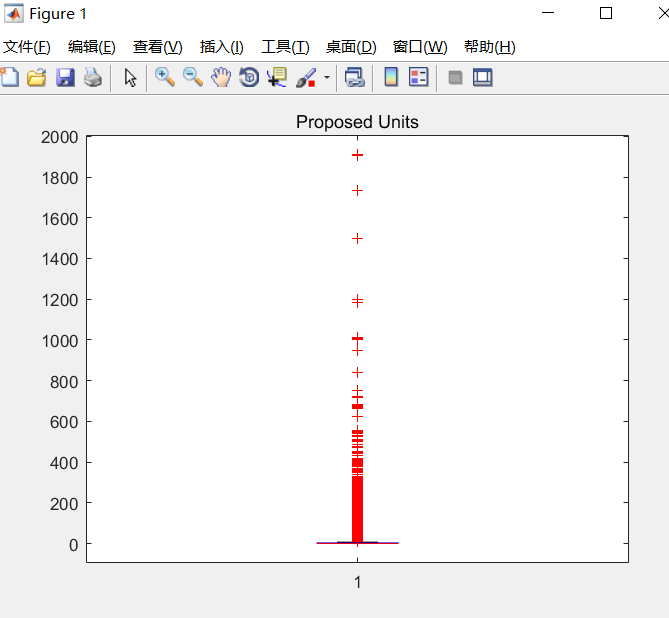
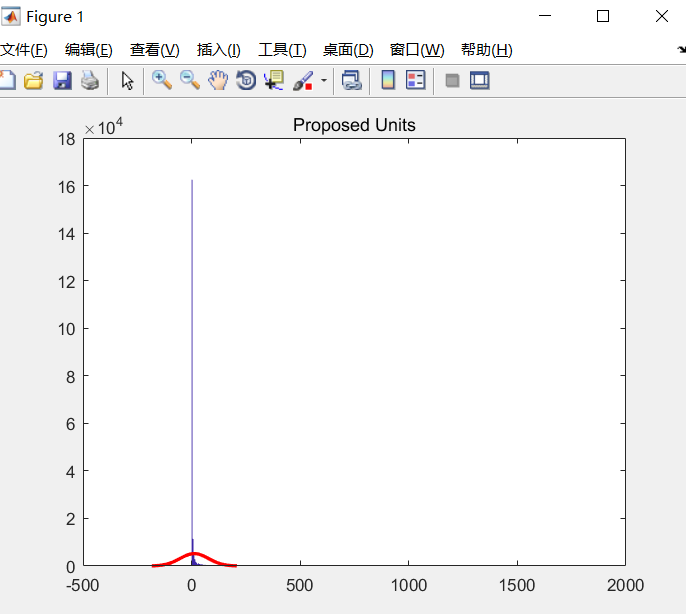


在直方图中可以看出，该属性的分布基本为正态分布，从盒图中我们可以清楚地看到该属性数据的最大值、最小值、均值、中位数、四分位数等信息。

同样，可以得到其他九个属性的直方图和柱形图







**四、实验感想**

通过本次实验，我掌握了数据探索性分析与预处理的基本步骤与方法，同时巩固了matlab，收获很大。但是，由于知识、能力有限，通过属性的相关关系和通过数据对象之间的相似性来填补缺失值这两部分的工作还是不太会，没能完成。希望在之后的课程中能解决，并不断学习、不断进步。