**大作业一：数据探索性分析与数据预处理**

**姓名：夏铭泽**

**学号：2620170053**

**1. 问题描述**

本次作业中，将对2个数据集进行探索性分析与预处理。

**2. 数据说明**

数据集1: NFL Play-by-Play 2009-2017

数据集2: San Francisco Building Permits

**3. 数据分析要求**

**3.1 数据可视化和摘要**

数据摘要

对标称属性，给出每个可能取值的频数，

数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。

数据的可视化

针对数值属性，

绘制直方图，用qq图检验其分布是否为正态分布。

绘制盒图，对离群值进行识别

**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。

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

处理后，可视化地对比新旧数据集。

**4.实验环境及语言**

**语言及环境依赖**

语言： python

依赖的包：pandas, pylab, matplotlib, scipy, numpy

**5.实现方法**

**5.1 数据摘要**

**5.1.1 概述**

数据摘要部分主要进行的处理有以下两个方面：

1. 对标称属性，给出每个可能取值的频数

2. 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数

**5.1.2 实现方法**

语言：python

结果：json，图

使用pandas.read\_csv读取源文件

对于标称属性NominalAttribute计算出每个取值的频数，结果保存在1（2）.json中

对于数值属性NumericAttribute计算出最大、最小、均值、中位数、四分位数及缺失值的个数，结果保存在1（2）.txt中

**5.2 数据可视化**

**5.2.1 概述**

数据可视部分主要进行的处理是把数值部分的每一个属性包含的数据分别可视化（绘制成直方图，qq图，盒图）：

**5.2.2 实现方法**

语言：python

结果：直方图，qq图，盒图

**（1）. 绘制直方图，qq图，盒图**

直方图的绘制方法：主要用到的是pylab.hist()方法

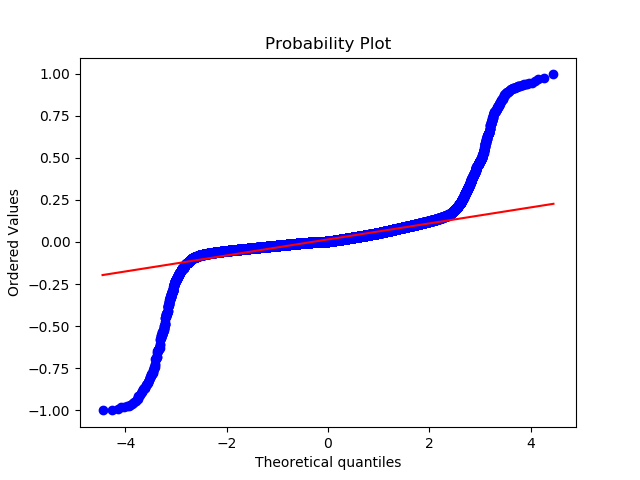
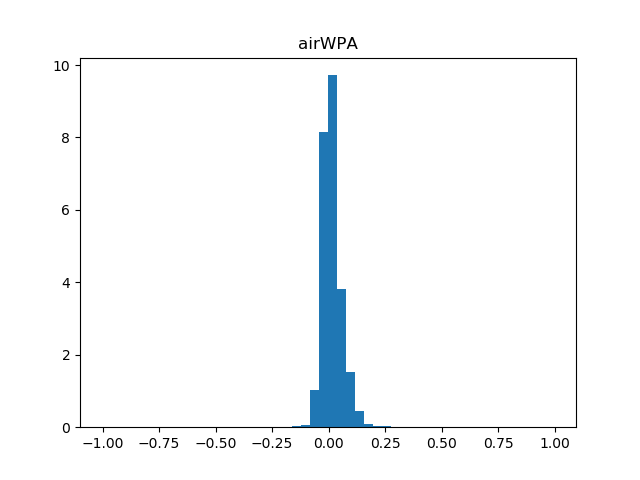
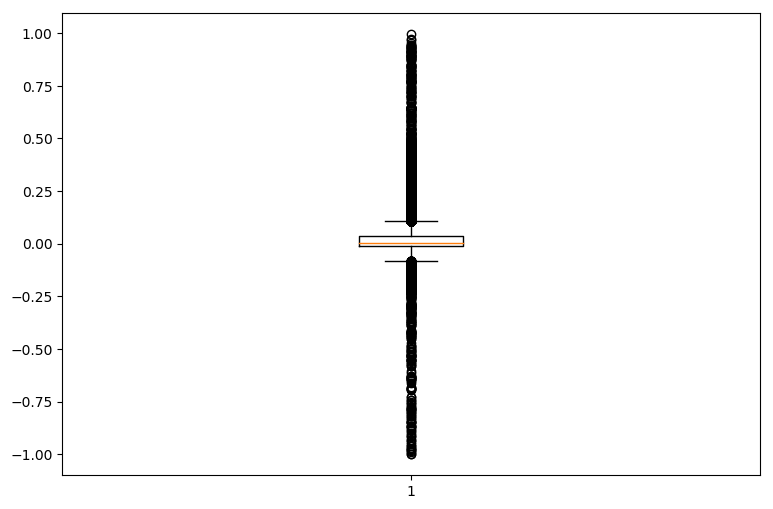
qq图的绘制方法：主要用到的是scipy.stats.probplot方法

盒图的给制方法：主要用到的是boxplot()方法

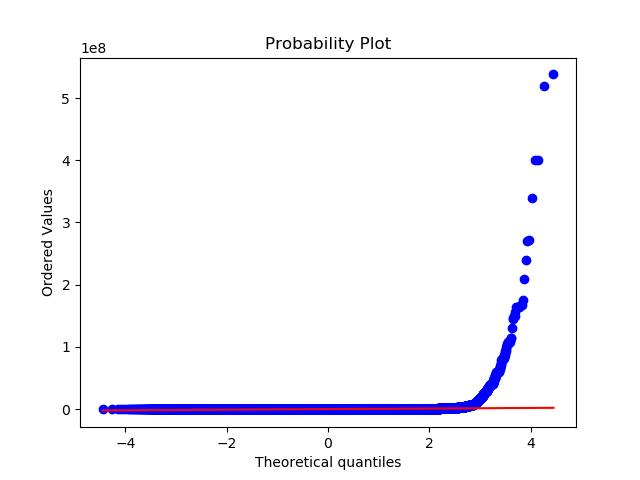
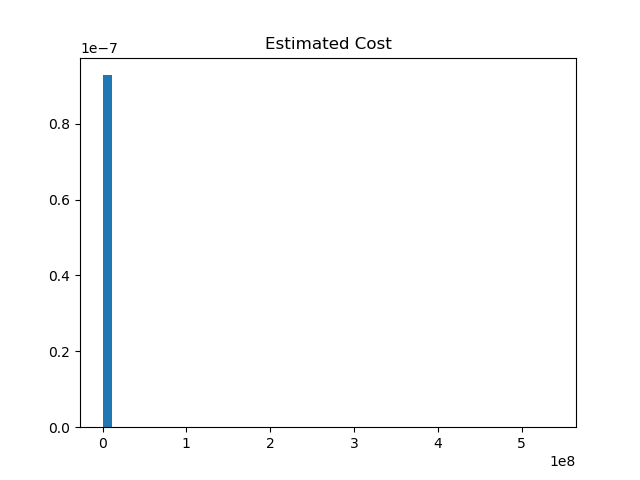
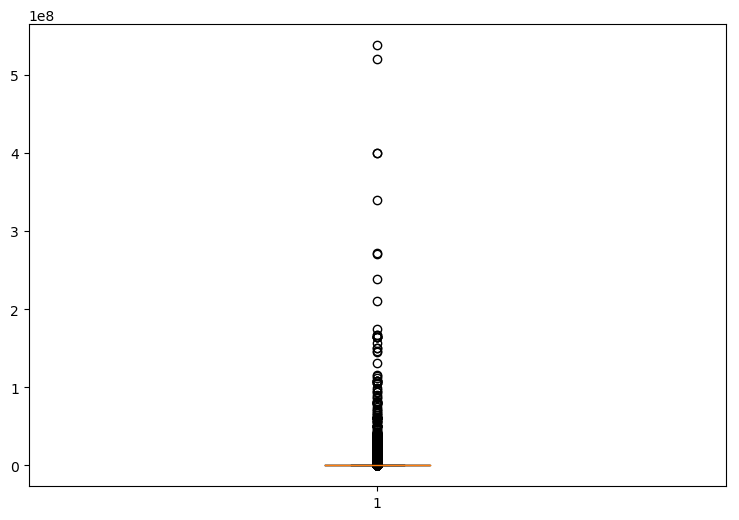
结果分别保存成图片

**（2）. 结果展示**

**详细图片保存在1（Building\_Permits），2（NFL Play by Play 2009-2017 (v4)）文件夹中**



airWPA盒图，直方图，QQ图



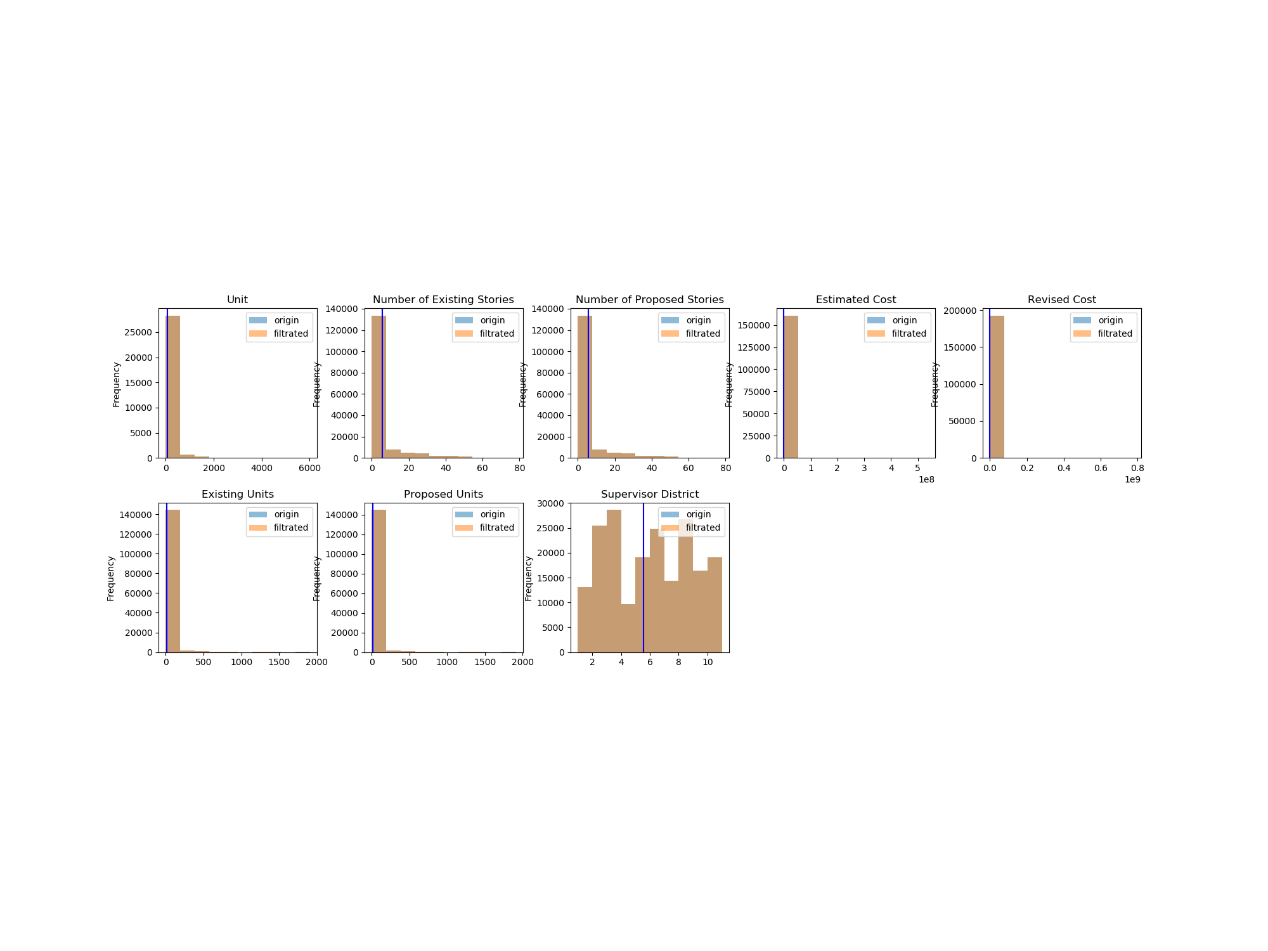
Estimated Cost盒图，直方图，QQ图

**5.3. 数据缺失的处理**

**1） 将缺失部分剔除**

**使用dropna函数删除缺失数据**

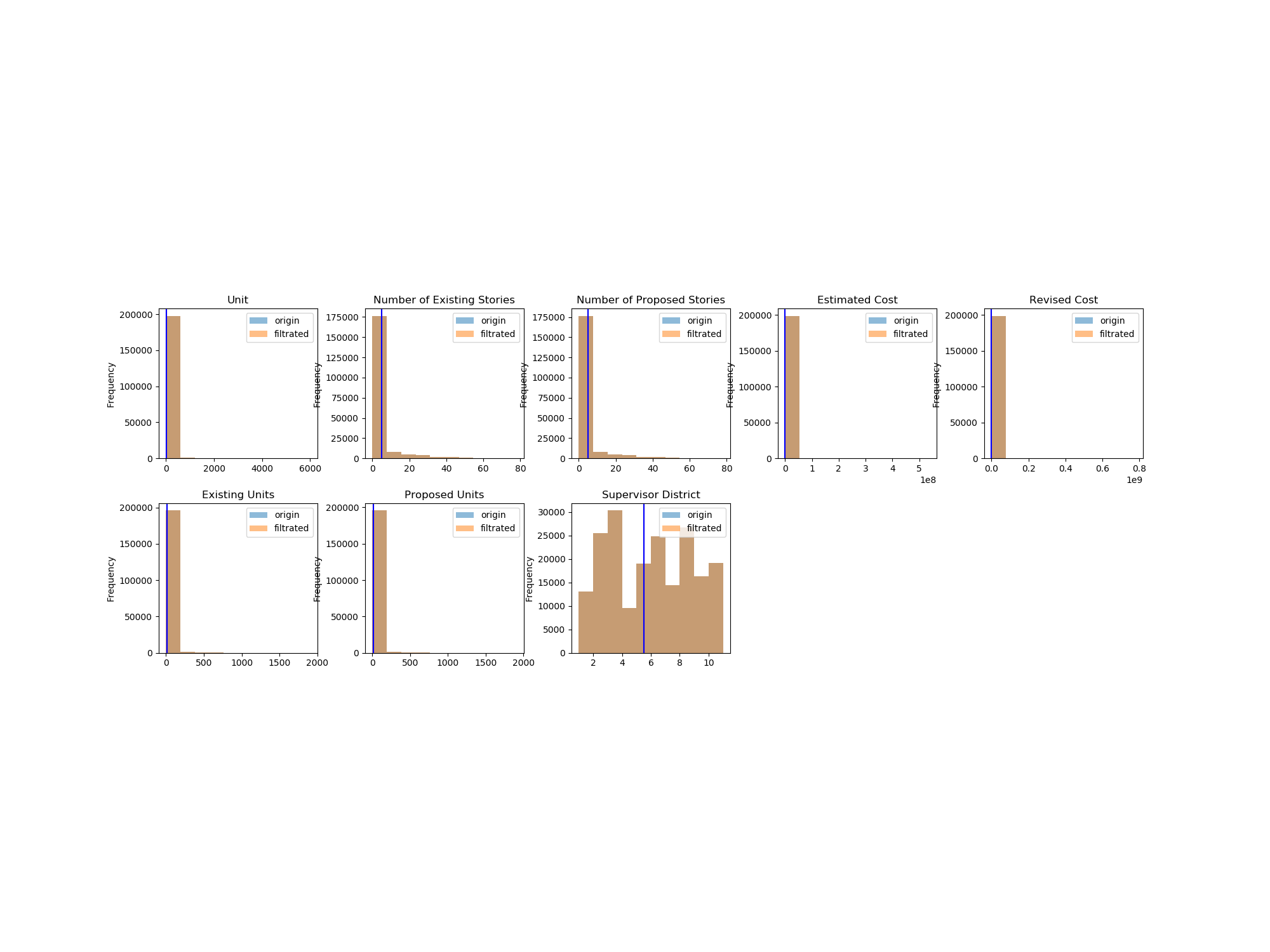
通过对比数值属性前后直方图来可视化处理前后效果

****

**2) 使用高频数值来填补缺失值**

**通过计算当前属性的众数然后对缺失值赋值**

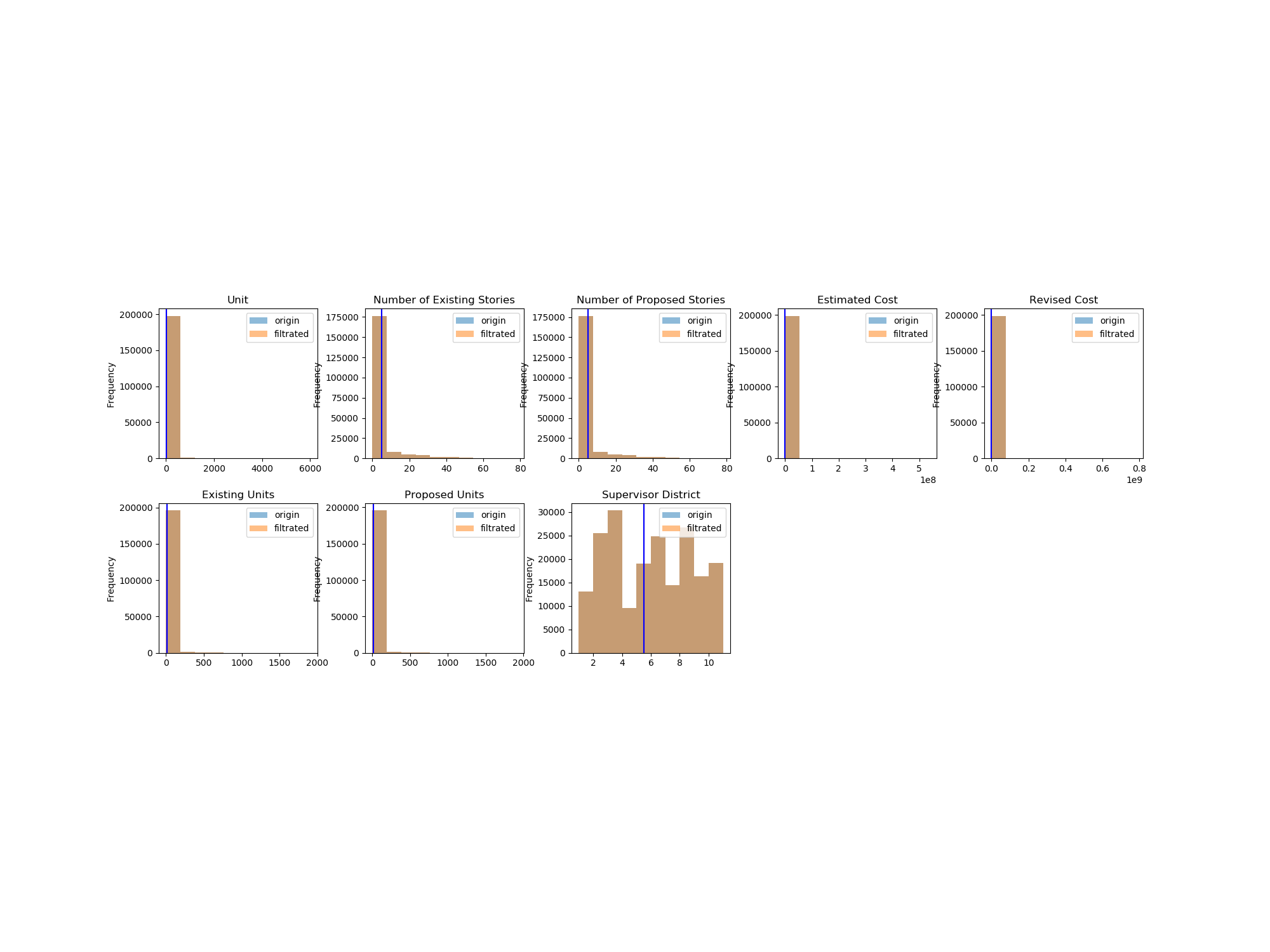
通过对比数值属性前后直方图来可视化处理前后效果

****

**3) 通过属性的相关关系来填补缺失值**

**通过interpolate进行插值来填补缺失值**

通过对比数值属性前后直方图来可视化处理前后效果



**4） 使用数据对象之间的相似型填补缺失值**