数据挖掘大作业一：数据探索性分析与数据预处理

姓名:康杨

学号:2120171024

1. 问题描述

对数据集进行探索性分析与预处理。

1. 数据说明

数据集: San Francisco Building Permits

1. 数据分析

3.1数据可视化和摘要

3.1.1数据摘要

* 标称属性,给出其每个可能取值的频数

由于数据量较大,数据结果保存在/Building\_Permits/result\_Build\_nominal.txt中

* 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。

Permit Type Street Number Unit Number of Existing Stories \

count 198900.000000 198900.000000 29479.000000 156116.000000

mean 7.522323 1121.728944 78.517182 5.705773

std 1.457451 1135.768948 326.981324 8.613455

min 1.000000 0.000000 0.000000 0.000000

25% 8.000000 235.000000 0.000000 2.000000

50% 8.000000 710.000000 0.000000 3.000000

75% 8.000000 1700.000000 1.000000 4.000000

max 8.000000 8400.000000 6004.000000 78.000000

Number of Proposed Stories Estimated Cost Revised Cost \

count 156032.000000 1.608340e+05 1.928340e+05

mean 5.745043 1.689554e+05 1.328562e+05

std 8.613284 3.630386e+06 3.584903e+06

min 0.000000 1.000000e+00 0.000000e+00

25% 2.000000 3.300000e+03 1.000000e+00

50% 3.000000 1.100000e+04 7.000000e+03

75% 4.000000 3.500000e+04 2.870750e+04

max 78.000000 5.379586e+08 7.805000e+08

Existing Units Proposed Units Plansets \

count 147362.000000 147989.000000 161591.000000

mean 15.666164 16.510950 1.274650

std 74.476321 75.220444 22.407345

min 0.000000 0.000000 0.000000

25% 1.000000 1.000000 0.000000

50% 1.000000 2.000000 2.000000

75% 4.000000 4.000000 2.000000

max 1907.000000 1911.000000 9000.000000

Existing Construction Type Proposed Construction Type \

count 155534.000000 155738.000000

mean 4.072878 4.089529

std 1.585756 1.578766

min 1.000000 1.000000

25% 3.000000 3.000000

50% 5.000000 5.000000

75% 5.000000 5.000000

max 5.000000 5.000000

Supervisor District Zipcode Record ID

count 197183.000000 197184.000000 1.989000e+05

mean 5.538403 94115.500558 1.162048e+12

std 2.887041 9.270131 4.918215e+11

min 1.000000 94102.000000 1.293532e+10

25% 3.000000 94109.000000 1.308567e+12

50% 6.000000 94114.000000 1.371840e+12

75% 8.000000 94122.000000 1.435000e+12

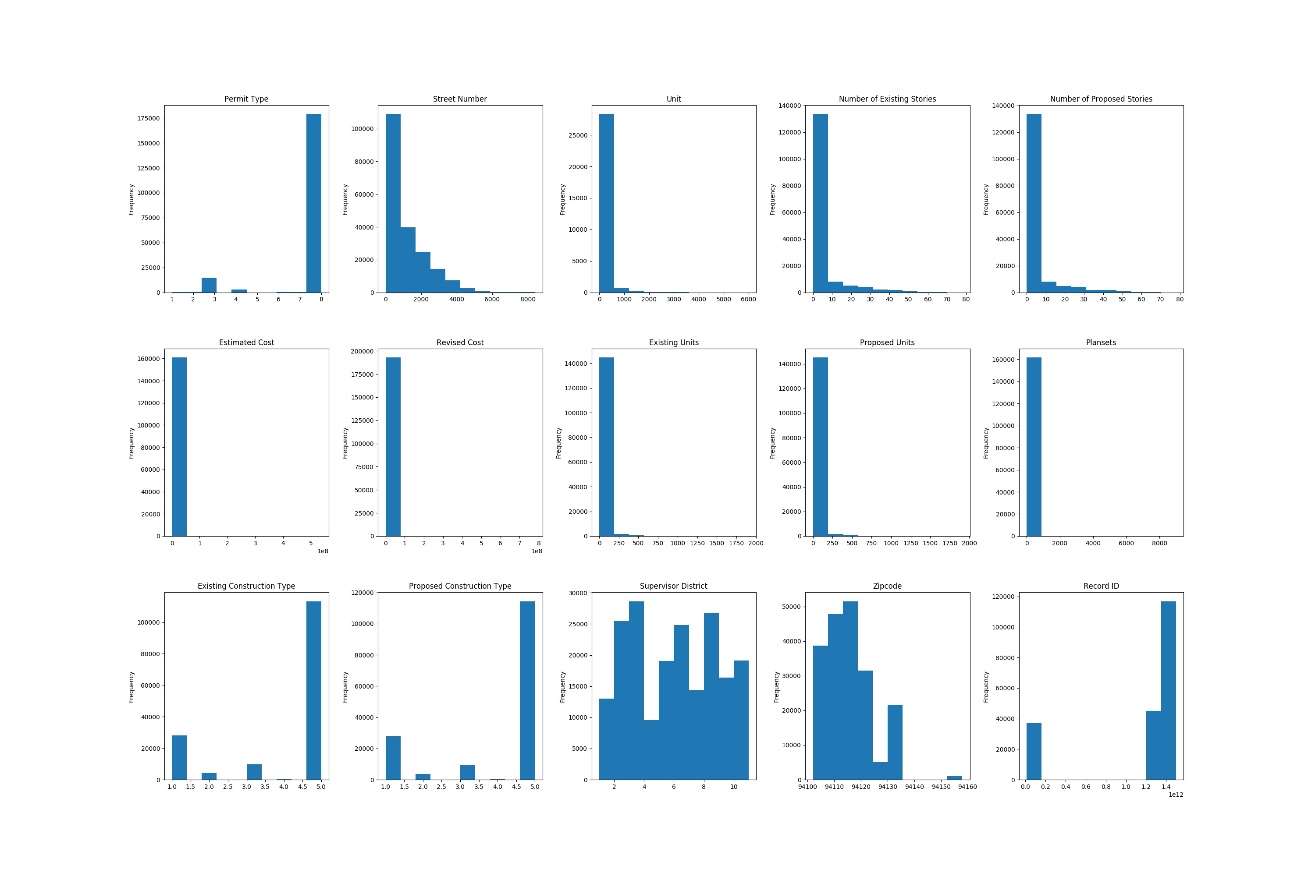
max 11.000000 94158.000000 1.498342e+12

3.1.2数据可视化

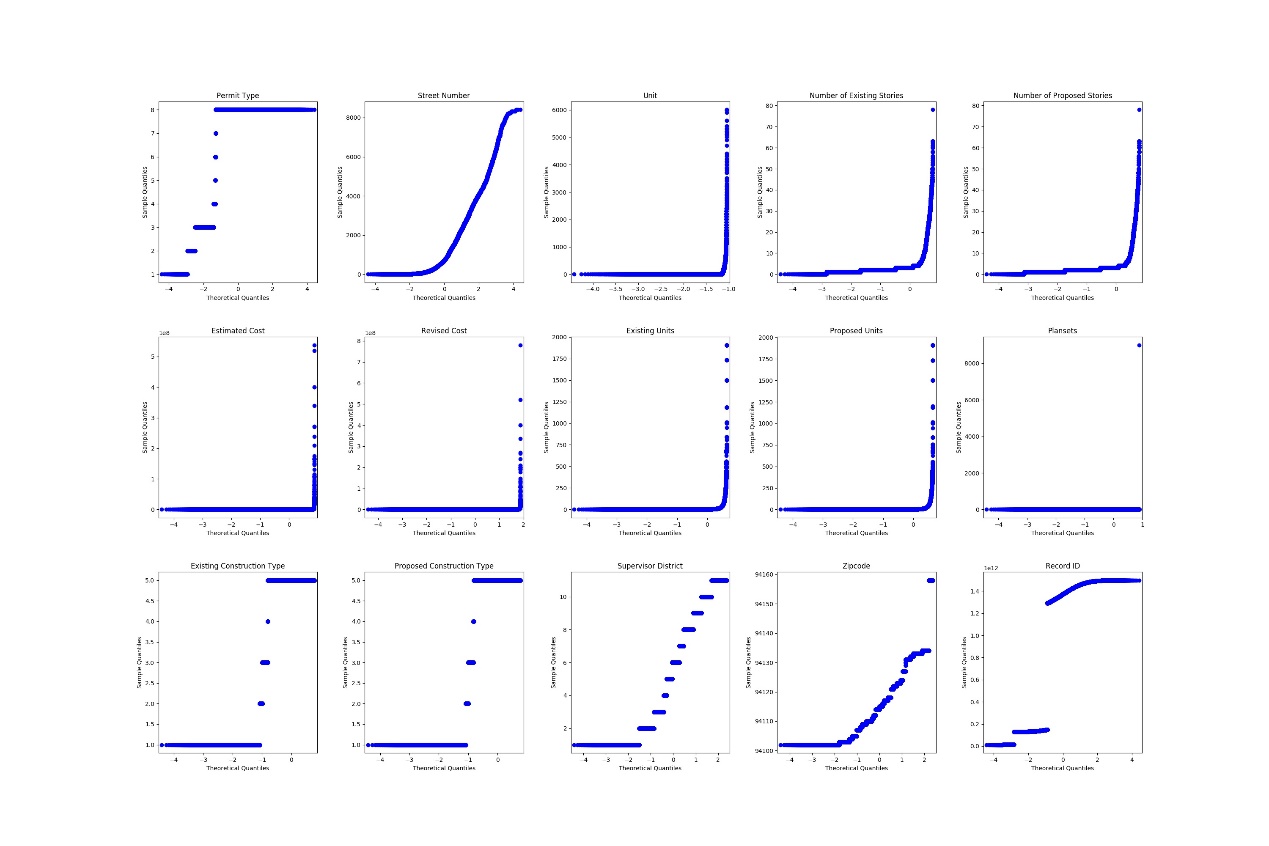
针对数值属性，

* 绘制直方图，用qq图检验其分布是否为正态分布。

直方图如下所示:



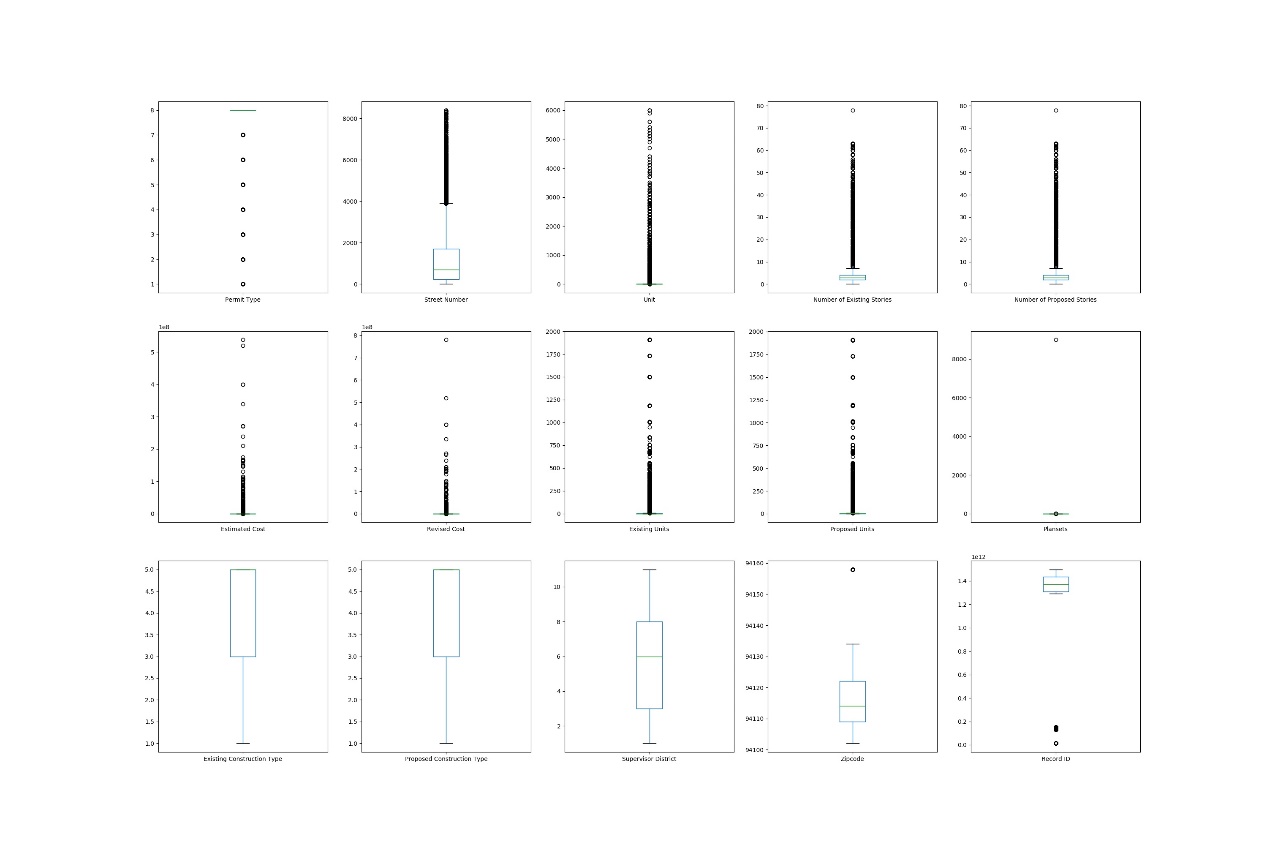
qq图如下所示:



由各个属性的qq图可以看出,无属性满足正态分布

* 绘制盒图，对离群值进行识别

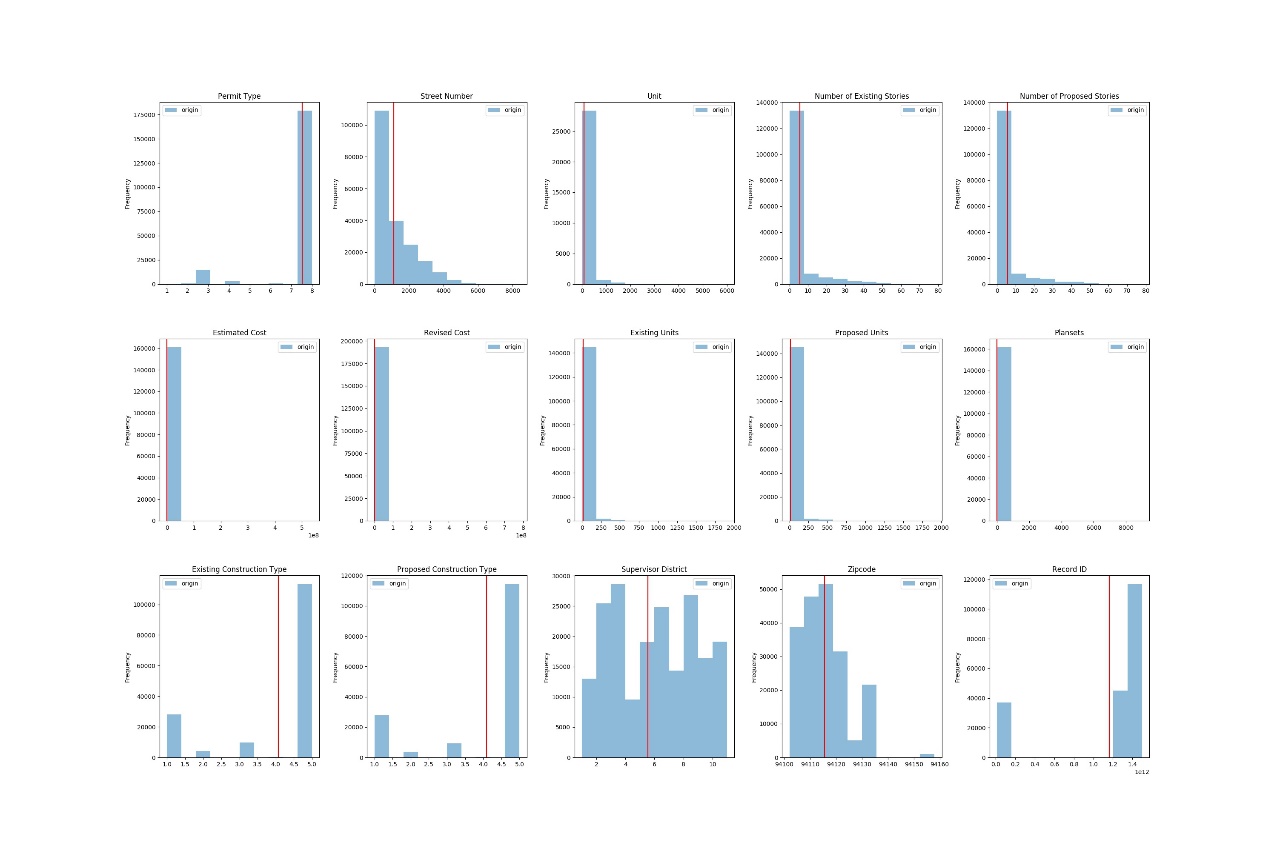
盒图如下所示:



从各个属性的盒图观察可得,属性Permit Type、Street Number、Unit、Number of Existing Stories、Number of Proposed Stories、Estimated Cost、Revised Cost、Existing Units、Proposed Units、Plansets、Zipcode、Record ID存在离群值

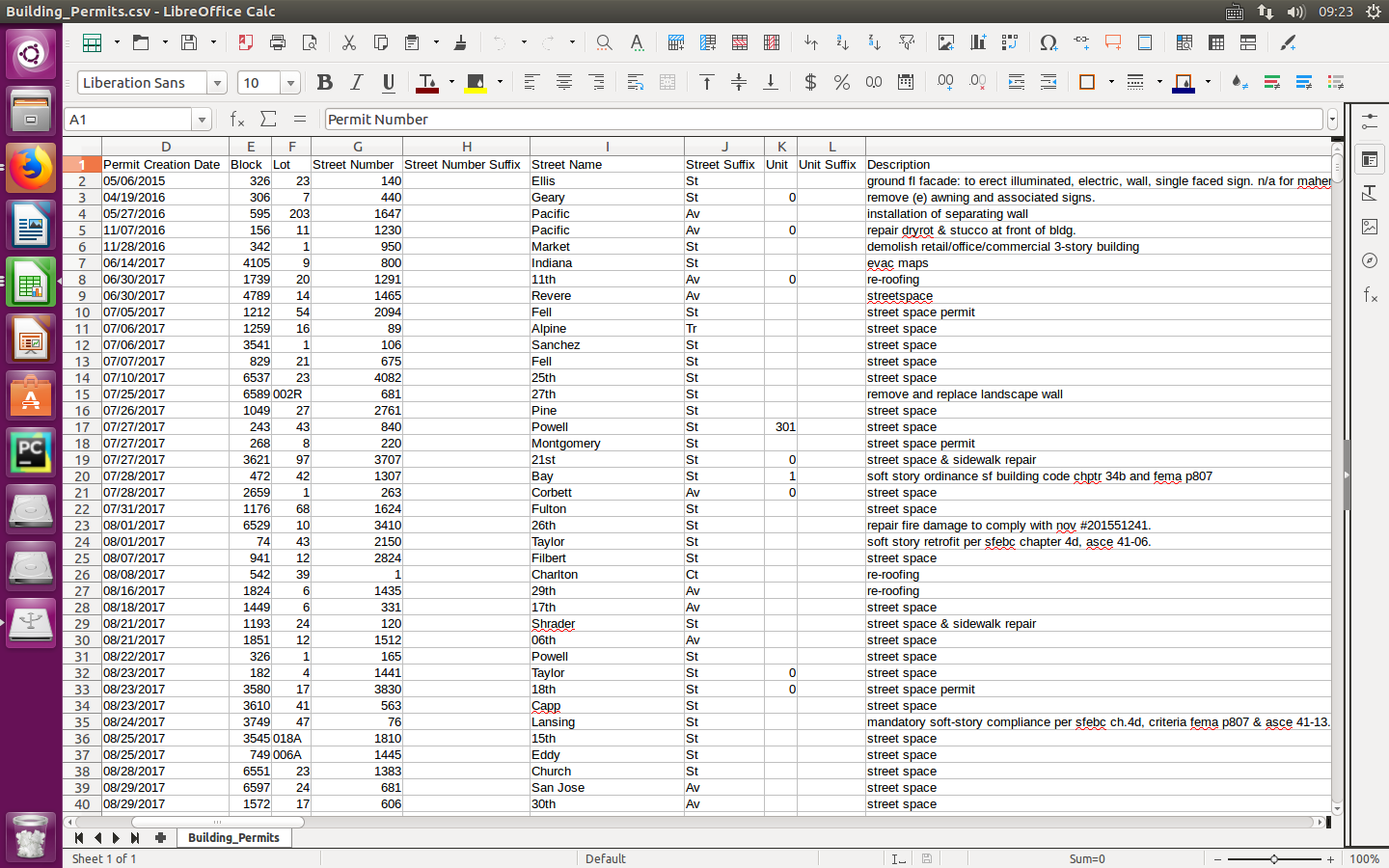
3.2数据缺失处理

* 将缺失部分剔除

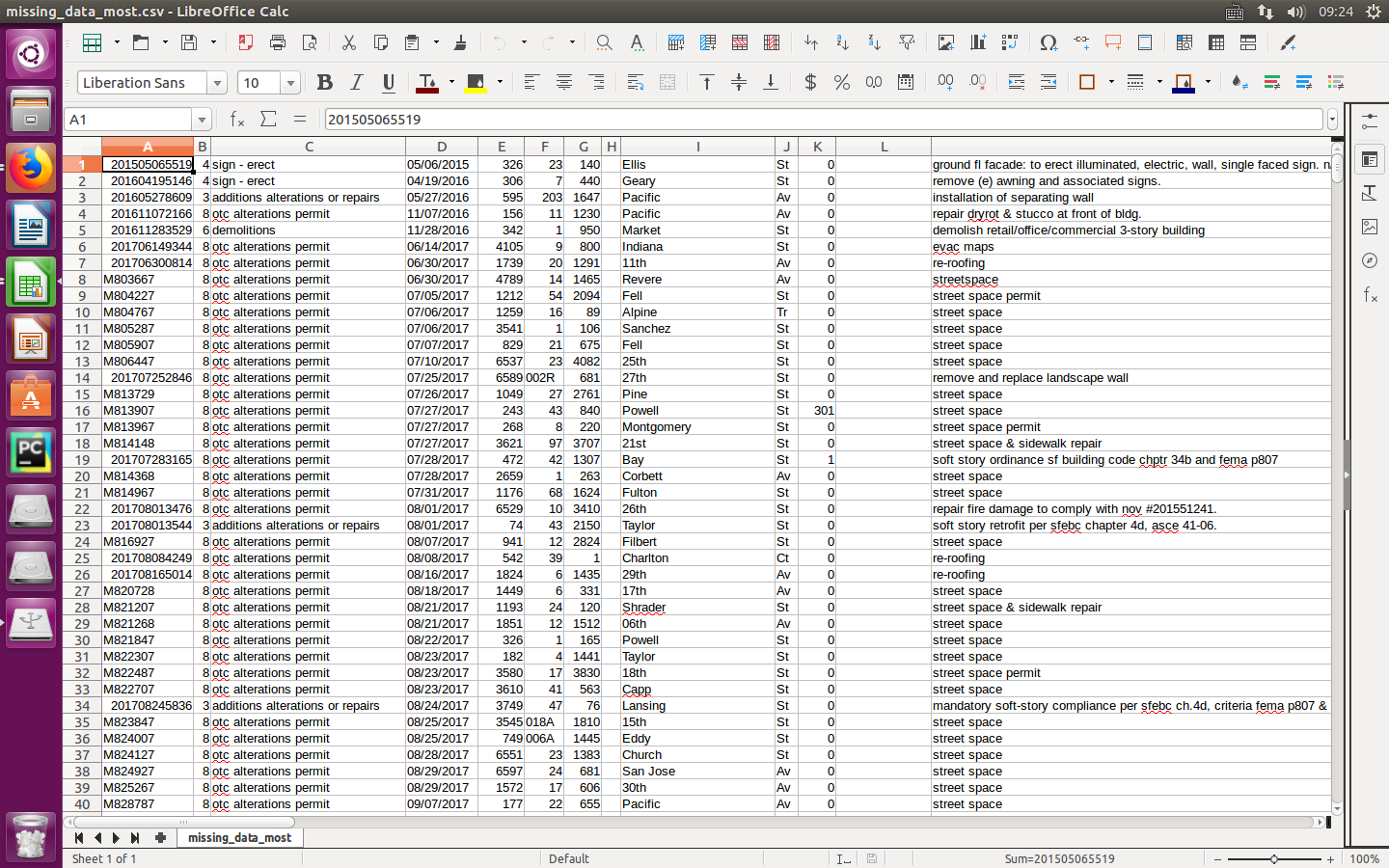


* 用最高频率值来填补缺失值

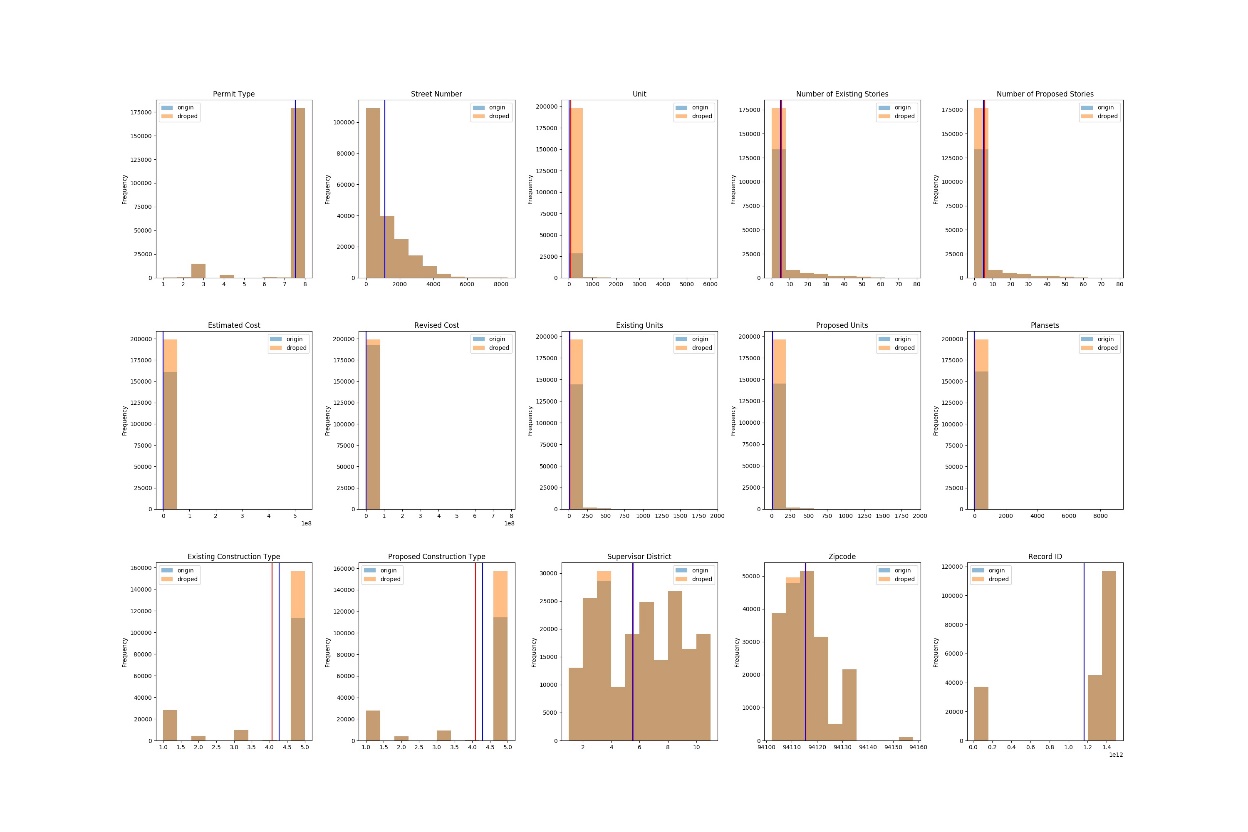
填补前,csv数据文件部分情况如下所示:



填补后,csv数据文件部分情况如下所示:

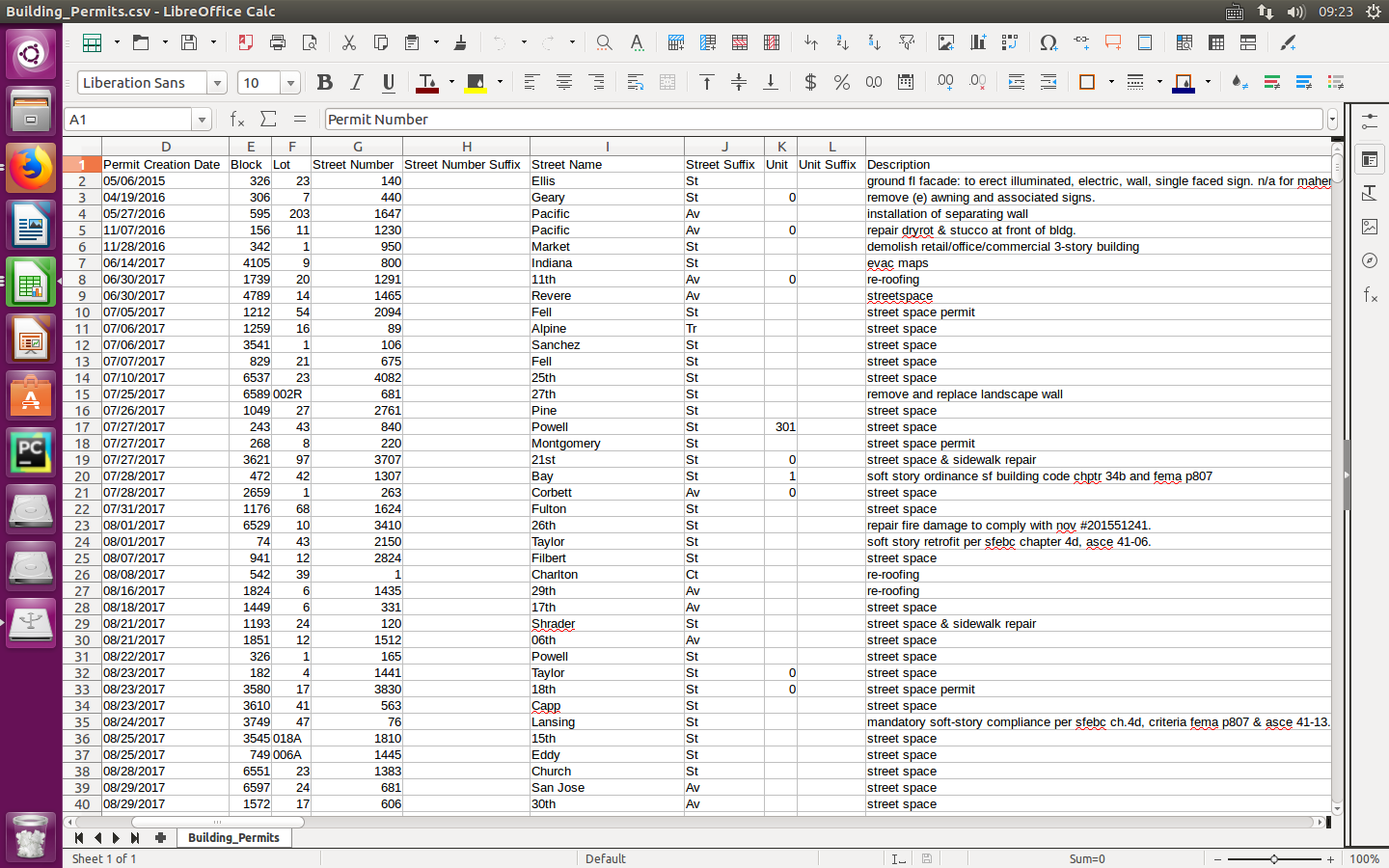


可视化对比填补前后数据,直方图如下所示:

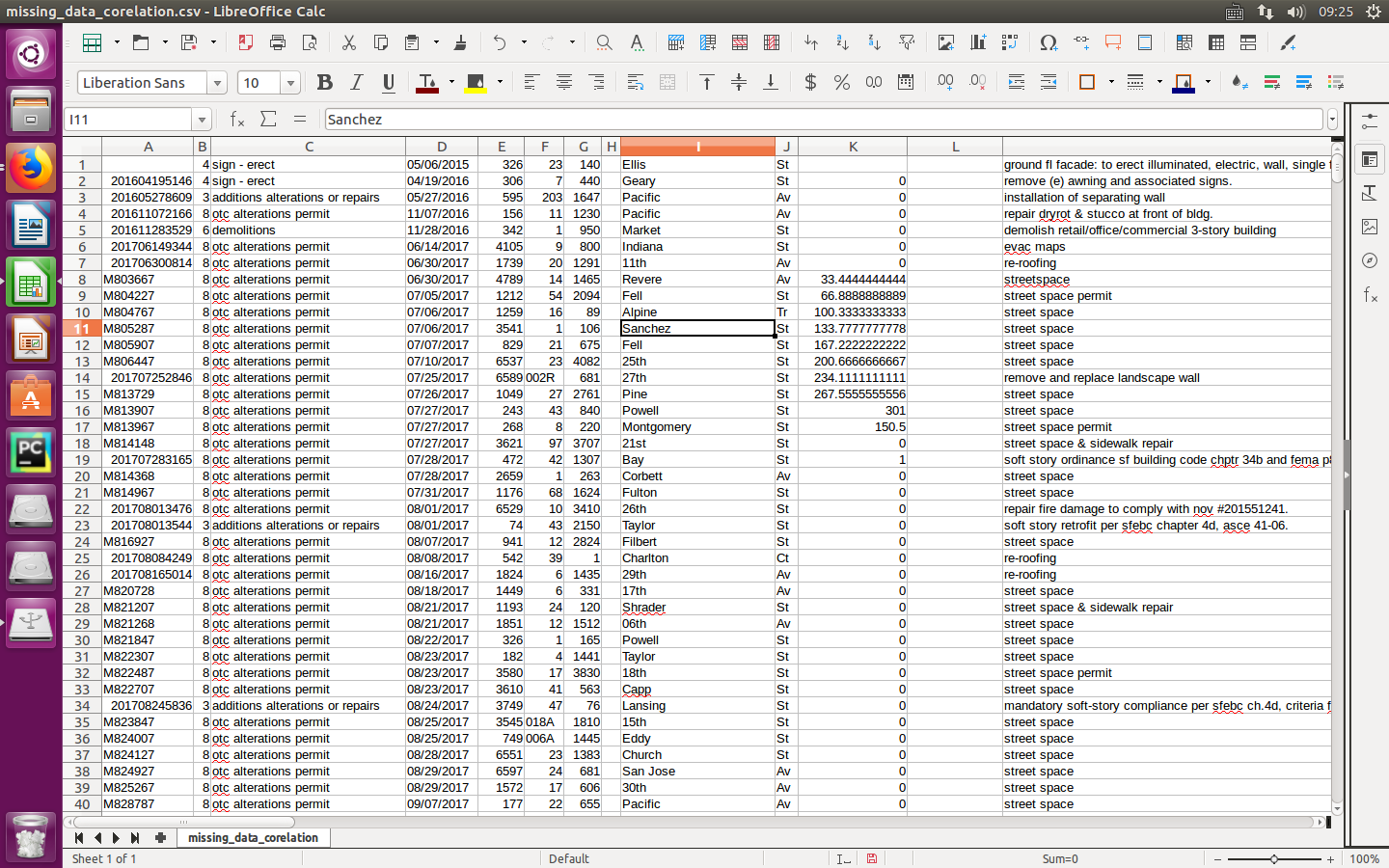


* 通过属性的相关关系来填补缺失值

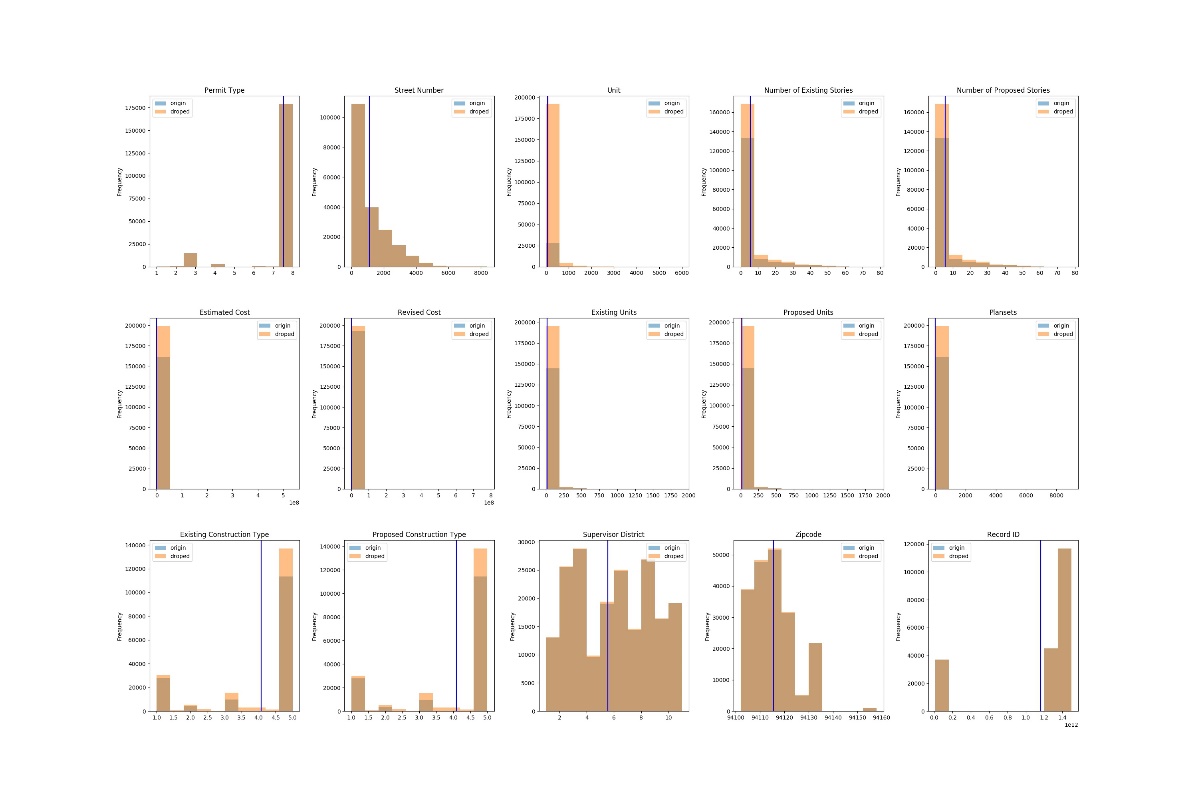
填补前,csv数据文件部分情况如下所示:



填补后,csv数据文件部分情况如下所示:



可视化对比填补前后数据,直方图如下所示:



* 通过数据对象之间的相似性来填补缺失值