数据挖掘大作业一：数据探索性分析与数据预处理

姓名：孙澈

学号：2120171054

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

下载数据: 地址

数据集中属性解释：

- 数据集 1: 参考
- 数据集 2: 见下载地址中 `DataDictionaryBuildingPermit.xlsx`

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. 提交内容

4.1 分析过程

数据集 2: San Francisco Building Permits

1）数据摘要

San Francisco Building Permits 数据集（下简称为 permits 数据集）本身保存为 csv 格式，根据数据集的属性解释得知该数据集有 43 个属性，如下表所示：

Sl No	Column name	Attribute type
1	Permit Number	Nominal attribute
2	Permit Type	Nominal attribute
3	Permit Type Definition	Nominal attribute
4	Permit Creation Date	Nominal attribute
5	Block	Nominal attribute
6	Lot	Nominal attribute
7	Street Number	Nominal attribute
8	Street Number Suffix	Nominal attribute
9	Street Name	Nominal attribute
10	Street Name Suffix	Nominal attribute
11	Unit	Nominal attribute

12	Unit suffix	Nominal attribute
13	Description	Nominal attribute
14	Current Status	Nominal attribute
15	Current Status Date	Nominal attribute
16	Filed Date	Nominal attribute
17	Issued Date	Nominal attribute
18	Completed Date	Nominal attribute
19	First Construction Document Date	Nominal attribute
20	Structural Notification	Nominal attribute
21	Number of Existing Stories	Numerical attribute
22	Number of Proposed Stories	Numerical attribute
23	Voluntary Soft-Story Retrofit	Nominal attribute
24	Fire Only Permit	Nominal attribute
25	Permit Expiration Date	Nominal attribute
26	Estimated Cost	Numerical attribute
27	Revised Cost	Numerical attribute
28	Existing Use	Nominal attribute
29	Existing Units	Numerical attribute
30	Proposed Use	Nominal attribute
31	Proposed Units	Numerical attribute
32	Plansets	Nominal attribute
33	TIDF Compliance	Nominal attribute
34	Existing Construction Type	Nominal attribute
35	Existing Construction Type Description	Nominal attribute
36	Proposed Construction Type	Nominal attribute
37	Proposed Construction Type Description	Nominal attribute
38	Site Permit	Nominal attribute
39	Supervisor District	Nominal attribute
40	Neighborhoods - Analysis Boundaries	Nominal attribute
41	Zipcode	Nominal attribute
42	Location	Nominal attribute
43	Record ID	Nominal attribute

可以看出，只有ATTRIBUTES_Number = {'Number of Existing Stories'; 'Number of Proposed Stories'; 'Estimated Cost'; 'Revised Cost'; 'Existing Units'; 'Proposed Units'}这六个属性是数值属性，其他的是标称属性。

标称属性的频数：因为标称属性过多，且个别标称属性的类别也过多，在实验中，只统计了类别数小于1000的标称属性的频数，根据是否给出数字的类别分别存储在permits\result\ATTRIBUTES_Nominal\ATTRIBUTES_Nominal_without_Category_Label文件夹和permits\result\ATTRIBUTES_Nominal、ATTRIBUTES_Nominal_with_Category_Label文件夹下，在报告中只列举部分标称属性的统计量。

frequence of Permit Type attribute

Value	Type Description	Count	Percent
1	new construction	349	0.18%
2	new construction wood frame	950	0.48%
3	additions alterations or repairs	14663	7.37%
4	sign - erect	2892	1.45%
5	grade or quarry or fill or excavate	91	0.05%
6	demolitions	600	0.30%
7	wall or painted sign	511	0.26%
8	otc alterations permit	178844	89.92%

frequence of Current Status attribute

Type Description	Count	Percent
appeal	2	0.00%
approved	733	0.37%
cancelled	1536	0.77%
complete	97077	48.81%
disapproved	2	0.00%
expired	1370	0.69%
filed	12043	6.05%
incomplete	2	0.00%
issued	83559	42.01%
plancheck	16	0.01%
reinstated	563	0.28%
revoked	50	0.03%
suspend	193	0.10%
withdrawn	1754	0.88%

frequence of Supervisor District attribute

Type Description	Count	Percent
1	13038	6.61%
10	12152	6.16%
11	6940	3.52%
2	25483	12.92%
3	28649	14.53%
4	9592	4.86%
5	19045	9.66%
6	24797	12.58%
7	14365	7.29%
8	26760	13.57%
9	16362	8.30%

frequence of Zipcode attribute

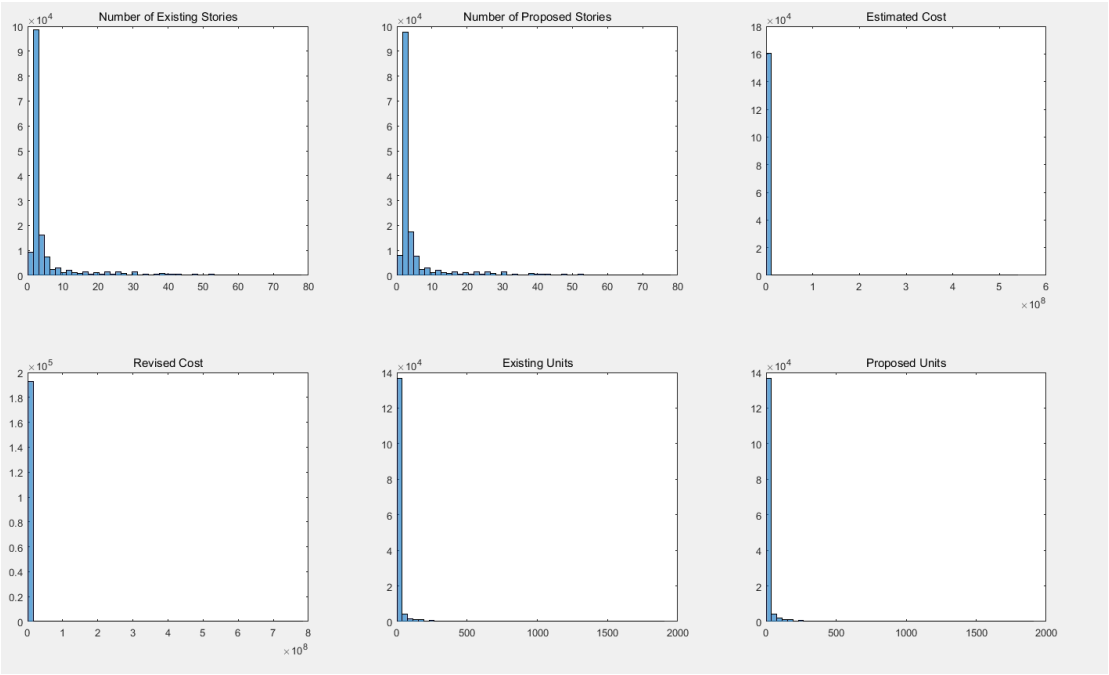
Type	Description	Count	Percent
	94102	7164	3.63%
	94103	10986	5.57%
	94104	4229	2.14%
	94105	8628	4.38%
	94107	7706	3.91%
	94108	5320	2.70%
	94109	11348	5.76%
	94110	17837	9.05%
	94111	5385	2.73%
	94112	7897	4.00%
	94114	13404	6.80%
	94115	10095	5.12%
	94116	6421	3.26%
	94117	11780	5.97%
	94118	9812	4.98%
	94121	7773	3.94%
	94122	8886	4.51%
	94123	9515	4.83%
	94124	5265	2.67%
	94127	4993	2.53%
	94129	23	0.01%
	94130	81	0.04%
	94131	7664	3.89%
	94132	3507	1.78%
	94133	7424	3.77%
	94134	2983	1.51%
	94158	1058	0.54%

数值属性的最大值、最小值、均值、中位数、四分位数、及缺失值的个数：

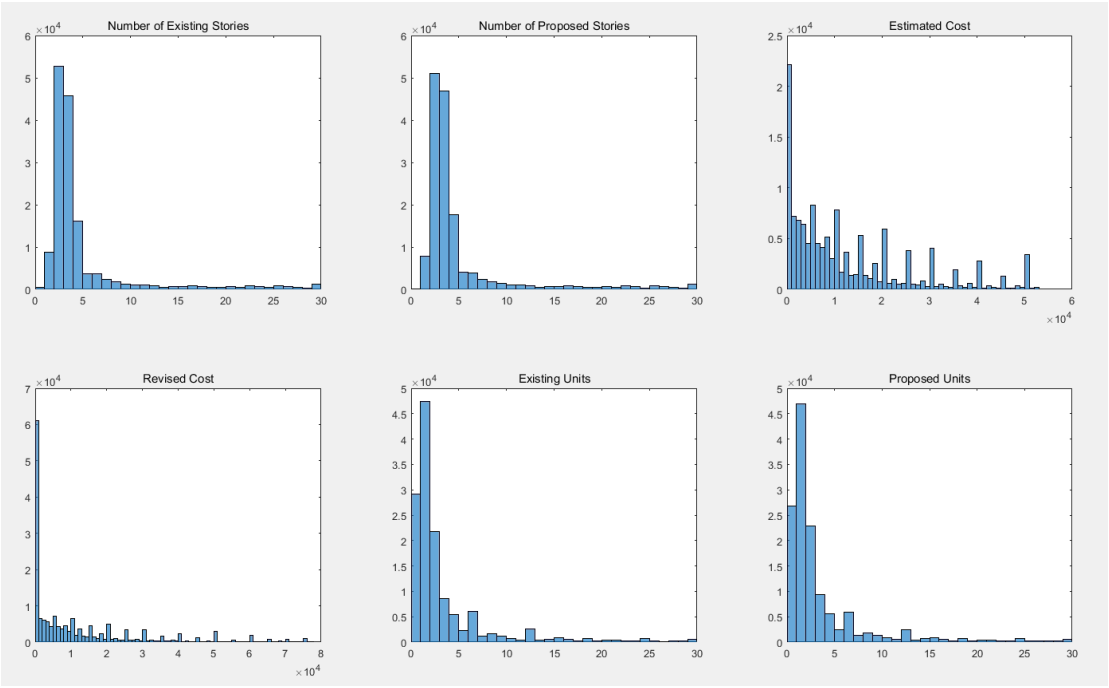
```
Data abstract of attribute Number of Existing Stories:
  Maximium:     78
  Minimium:     0
  Average:      5.7058
  Median:       3
  Quartile:     2, 4
  Missing data: 42784

Data abstract of attribute Number of Proposed Stories:
  Maximium:     78
  Minimium:     0
  Average:      5.745
  Median:       3
  Quartile:     2, 4
  Missing data: 42868

Data abstract of attribute Estimated Cost:
  Maximium:     537958646
  Minimium:     1
  Average:      168955.4433
  Median:       11000
  Quartile:     3300, 35000
  Missing data: 38066

Data abstract of attribute Revised Cost:
  Maximium:     780500000
  Minimium:     0
  Average:      132856.1865
  Median:       7000
  Quartile:     1, 28710
  Missing data: 6066

Data abstract of attribute Existing Units:
  Maximium:     1907
  Minimium:     0
  Average:      15.6662
  Median:       1
  Quartile:     1, 4
  Missing data: 51538

Data abstract of attribute Proposed Units:
  Maximium:     1911
  Minimium:     0
  Average:      16.511
  Median:       2
  Quartile:     1, 4
  Missing data: 50911
```

2）数据可视化

在数值属性本身存在的缺失情况下的直方图（存储在\permits\result\figure）：

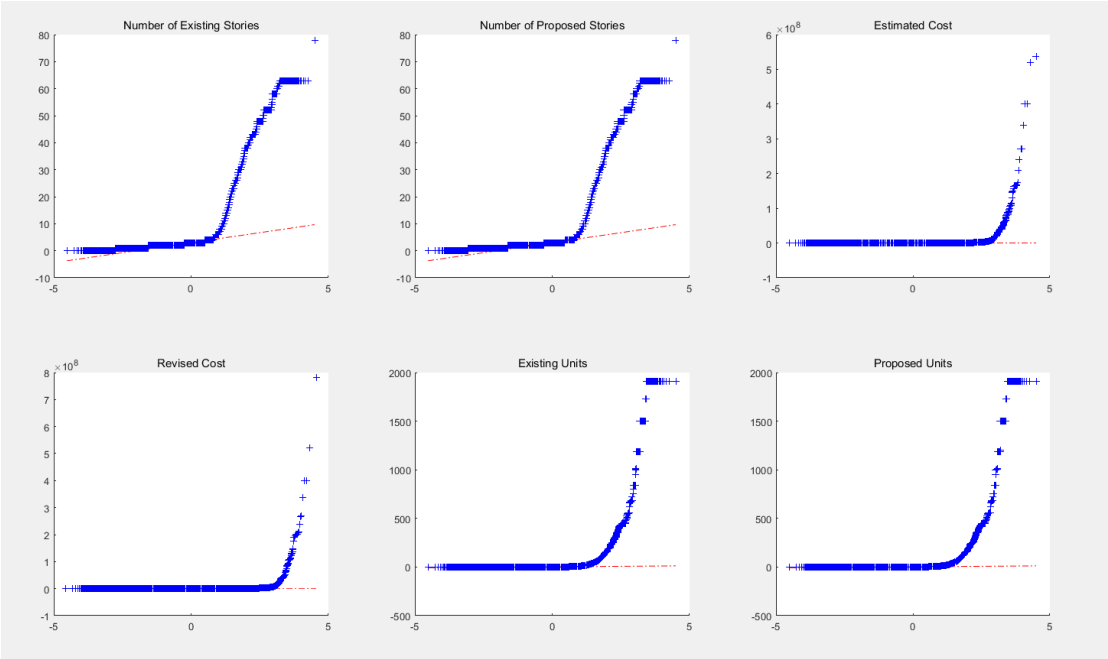


从上图可以看出 Number of Existing Stories 和 Number of Proposed Stories 分布类似，在数值较小的情况下接近正态分布，但是仍然存在数量不少的数值较大数据，而 Estimate Cost 和 Reviewed Cost 分布类似，均在数值较小的分布较为频繁，但是存在较大的值，而 Existing Units 和 Proposed Units 类似。因为大部分数值属性的数据存在较小的数值区域内，为了更好地观察它们，下图展示了在数值较小的情况下的数据分布情况（忽视了数值比较大的数据，关注较多的但值比较小的情况的分布）。

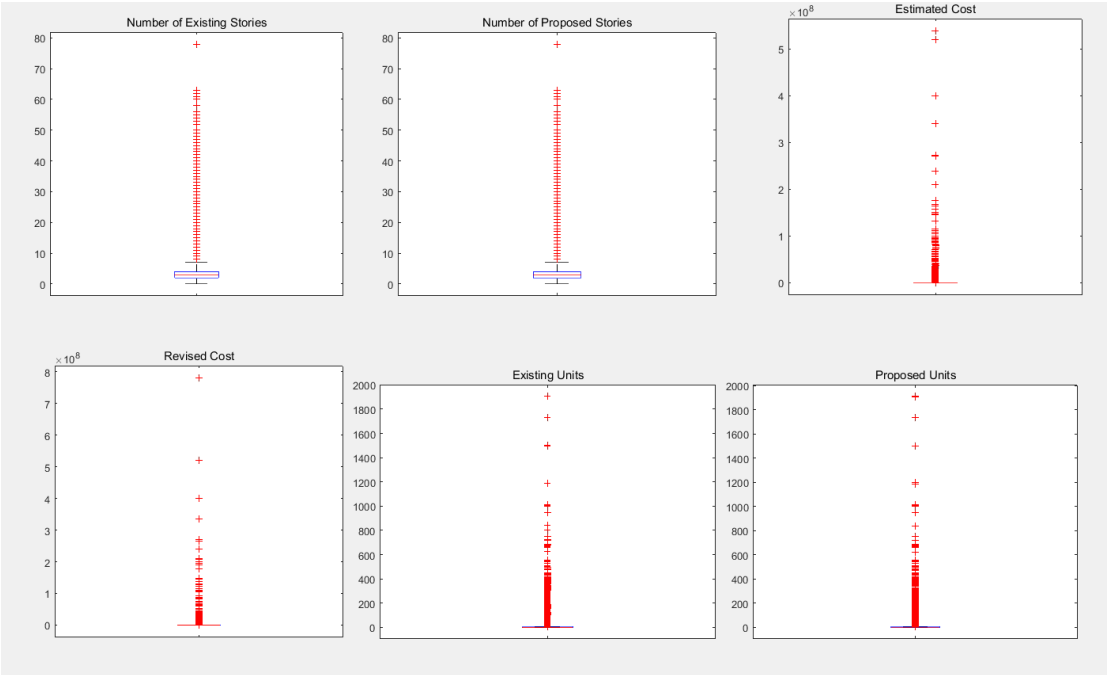


可以进一步看出，前两个属性 Number of Existing Stories 和 Number of Proposed Stories 在值较小的情况下是正态分布，而剩下的四个属性均集中在较小的值处。从 QQ 图也可以判定其和正态分布的相似度。

QQ 图：可以看出前两个属性 Number of Existing Stories 和 Number of Proposed Stories 在值较小的情况下更符合正态分布。



盒图：由盒图可以看出，所有的属性均有较多的值比较大离群值，而后四个属性特别明显。



3）数据集预处理

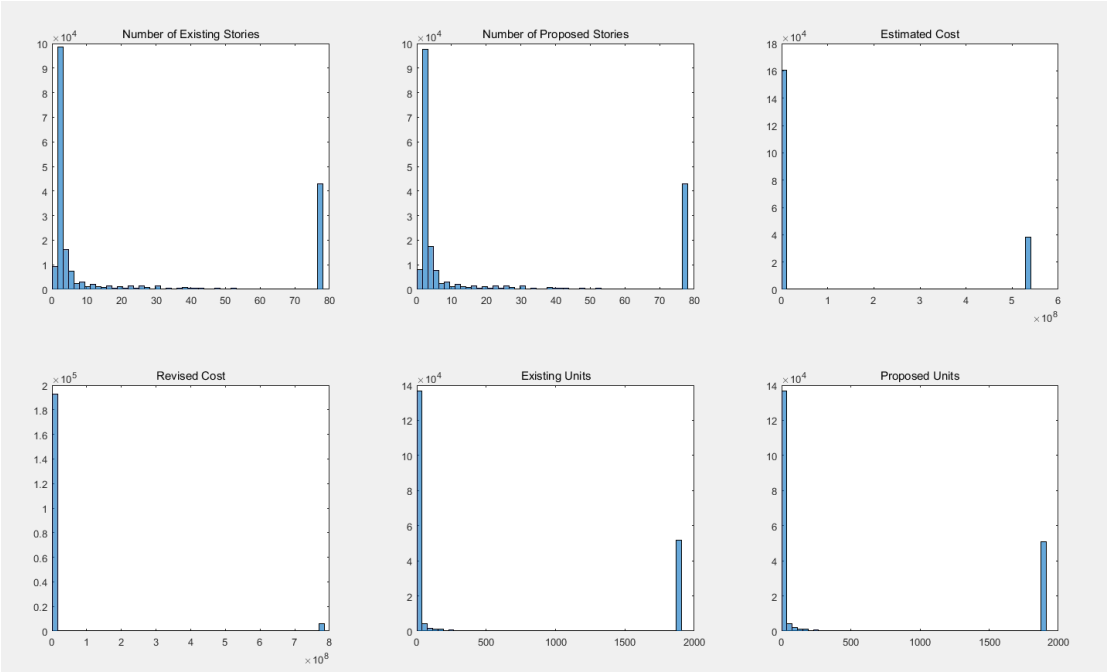
属性 Number of Existing Stories 和 Number of Proposed Stories 缺失除了丢失的原因外，还包括不适用某些 permit types，故该项不填。Estimated Cost 和 Revised Cost 缺失除了丢失的原因外，还包括写入不及时等原因，Proposed Use 和 Proposed Units 的缺失可能由于不太好统计。

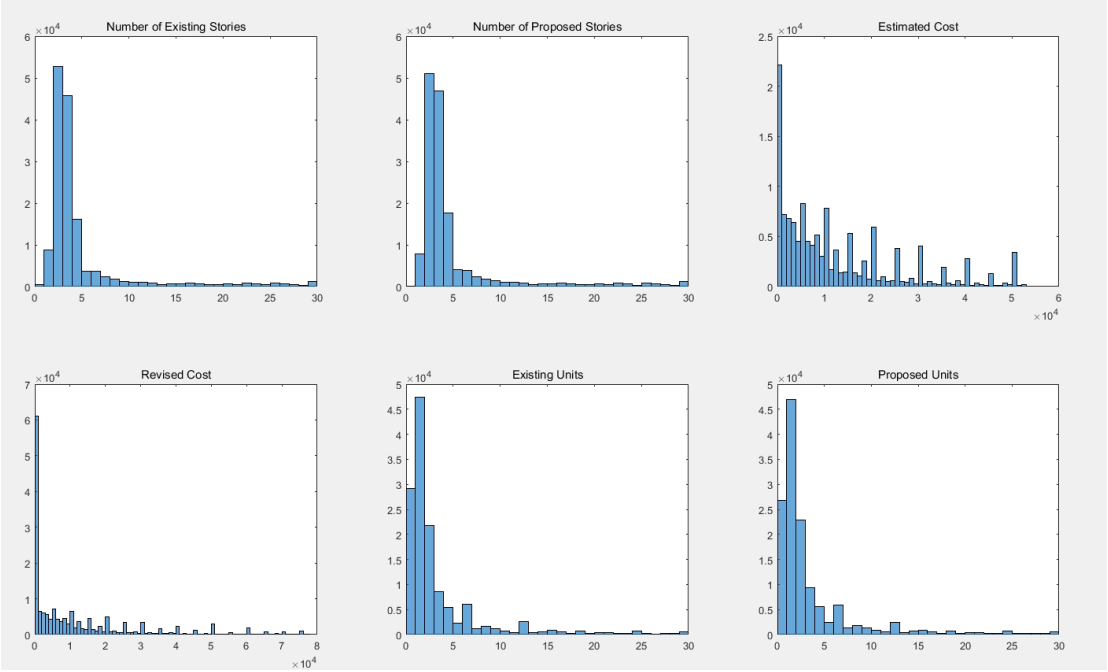
除了辨别出标称属性和数值属性之外，数据清洗也十分重要，最直观的原因是某些属性的缺失值过多，为了保证填充缺失值的质量，需要数据清洗。我根据

每个属性的缺失值的百分比确定是否保留该属性。造成数据缺失的原因有几个。有些值可能会丢失，因为它们不存在，有些值可能由于不正确的数据收集或不良的数据输入而丢失。例如，在数据中， Street Number Suffix 并不总是存在，因此填写它不太理想。但是，locations, zipcodes and permit expiration 可以填写，同时 TDIF compliance 和 Voluntary Soft-Story Retrofit 有近 100％的缺失值，所以放弃它们是有道理的。因此本文中决定删除丢失值至少为数据 60％的属性。

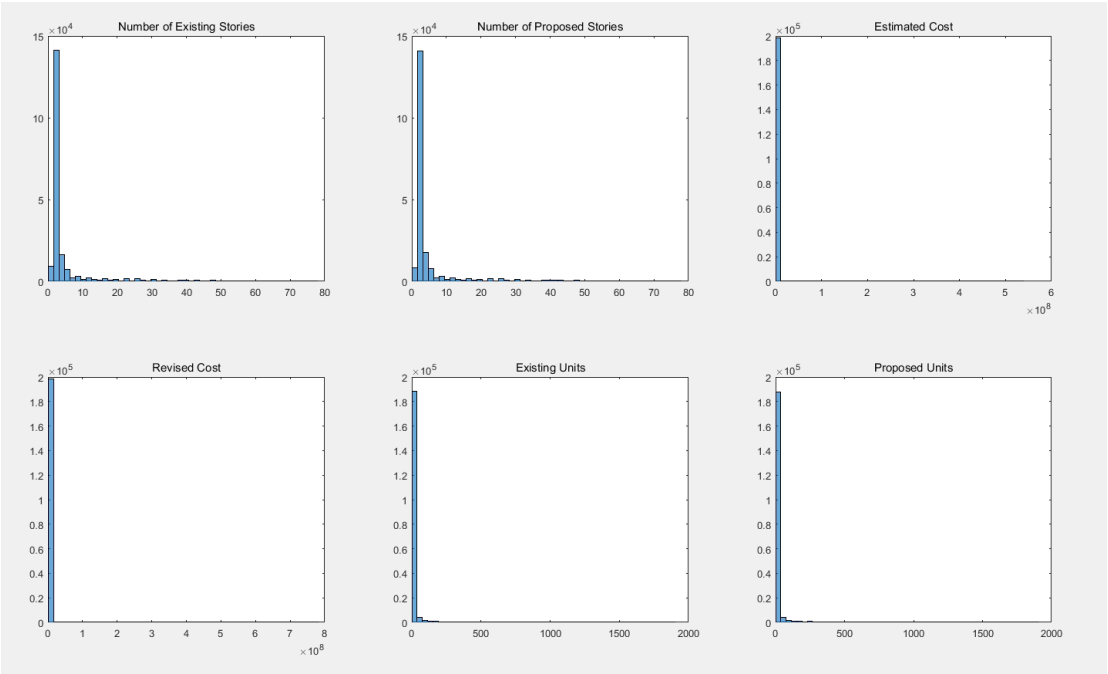
1. 剔除缺省值的操作已经在之前做好， 在此不再赘述。

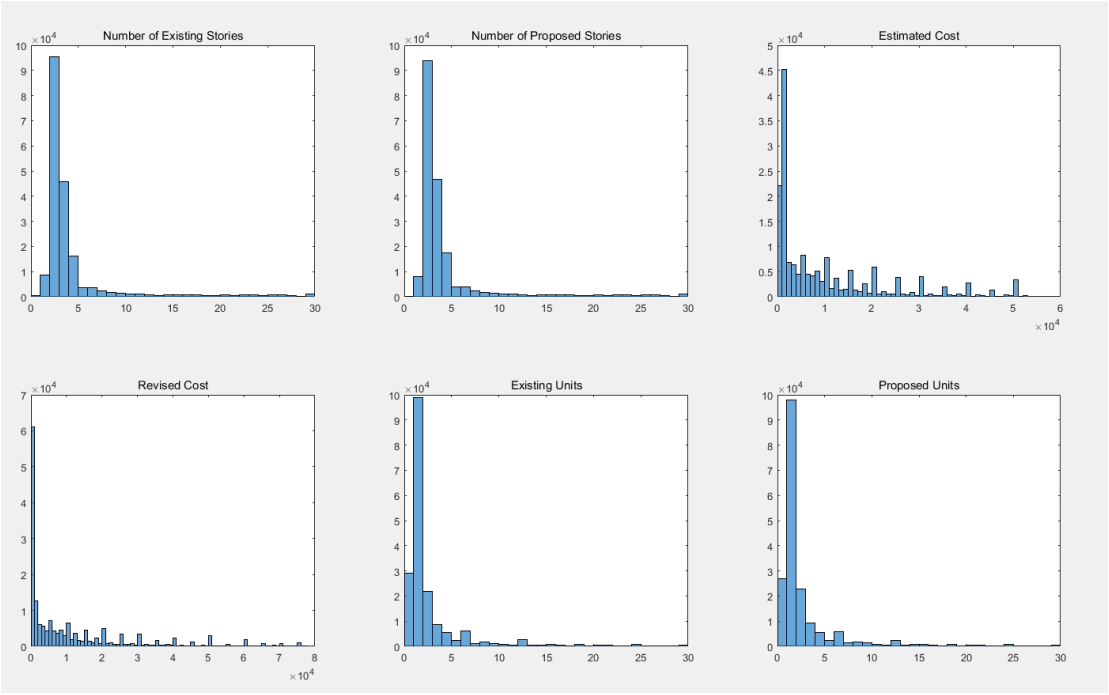
2. 用最高频率值来填补缺失值： 在此属性中的缺失值用此属性中所计算出的最高频率值填补。原始直方图以及较小值的直方图如下，可以看出，虽然各个属性均分布在较小的区域内，但是仍有为数不少的数据处在较大值区域内（众数的值比较大），因此在较小值的直方图中变化不明显。



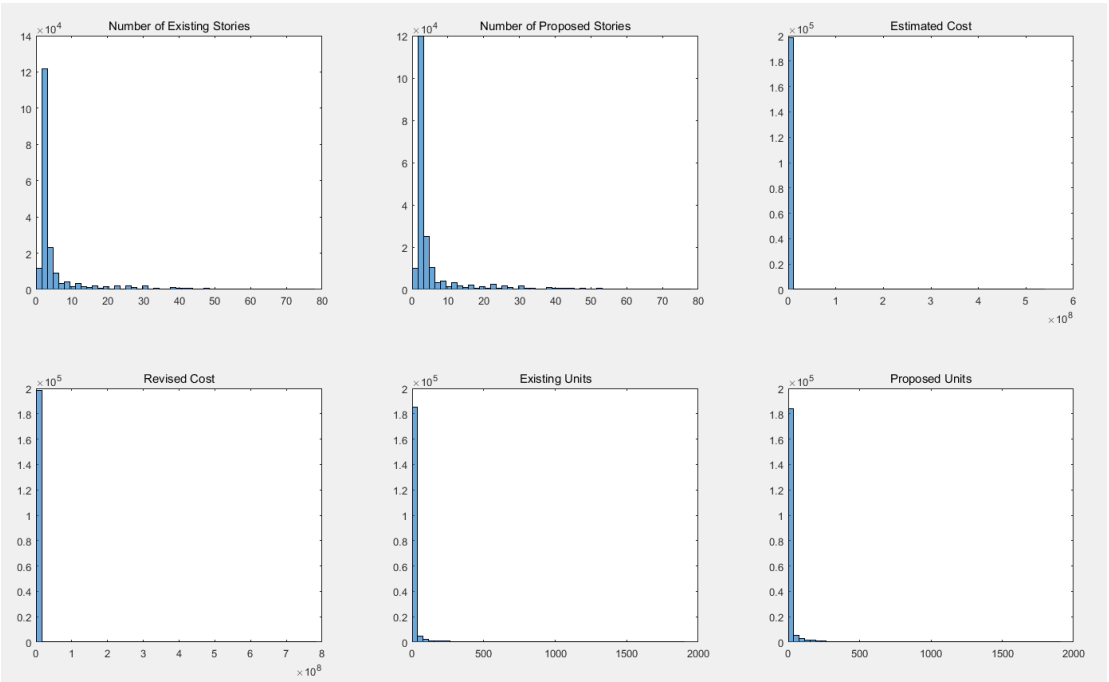


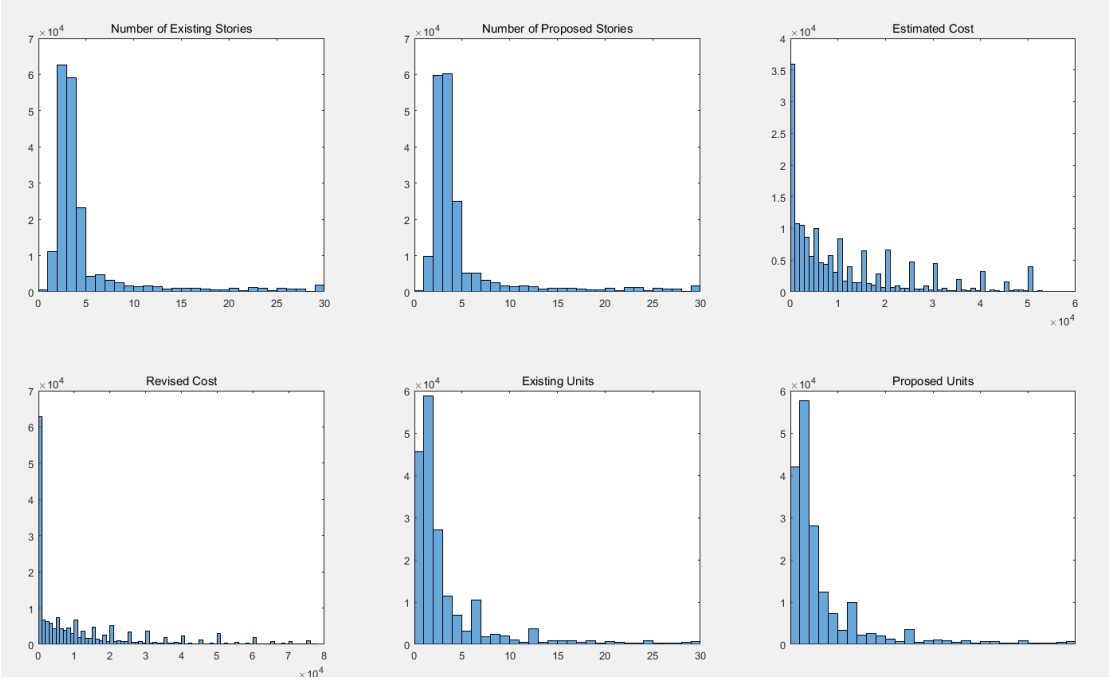
3. 通过属性的相关关系来填补缺失值： 计算两个属性的相关性， 相关性越大越可以根据另一个属性推断缺失属性的值。 通过另一属性的回归分析， 计算当前的缺失值。 直方图以及较小分布的直方图如下：





4. 通过数据对象之间的相似性来填补缺失值： 计算两个样本的相似程度， 越相似证明越可以使用该样本推断当前含缺失。直方图及较小分布的直方图如下：





由以上四幅图比较可知，按相似性填补的结果和处理前更相似，从值较小的分布直方图可以看出，相似性填补尤其是第三种填补方法，增大了属性值比较小的频数，符合原来的分布情况，从原始的直方图也可以看出来按用最高频率值来填补缺失值的结果和处理前差别最大。

4.2 分析程序

程序由三部分组成, permits\main.m、permits\mypreprocessing.m、permits\myvisualization.m。下面分别简单介绍它们的功能（具体见代码文件）。

1. main.m
   (1) 读入数据并保存成字符串元胞格式。
   (2) 标记数值属性和标称属性。
   (3) 处理无数字类别的标称数据（数据处理如格式变换、去除缺失值、统计标称属性有哪些取值 unique_current_attribute、根据 unique_current_attribute 标识标称属性的字符串表示以及与数字类别的对应关系，存在 permits\data\Nominal_Label\Nominal2Category_Label 文件夹下、输出统计量）。
   (4) 处理有数字类别的标称数据，与（3）类似，只是标称属性的字符串表示以及与数字类别的对应关系由属性字段给出。
   (5) 处理数值属性（数据处理如格式变换、去除缺失值、输出统计量），存储在\permits\result\ATTRBUTES_Number
   (6) myvisualization(total_attribute_number,attribute); 可视化原始数据。\permits\result\figure
   (7) 填补缺失值并存储及可视化，输入 1：最大频率填补（存储在 building_permits_filled_by_maximium.txt），2：属性的相关关系来填补（存储在 building_permits_filled_by_attribute.txt），3：数据对象之间的相似性来填补（存储在 building_permits_filled_by_similarity.txt），其他：退出。
2．myvisualization.m 绘制 6 个属性的直方图、QQ 图、盒图\permits\result\figure
3. mypreprocessing ,PREPROCESSING 对数据进行预处理，支持三种种方式（method 取值为 1~3）。三种方式分别为：最高频值替代，属性相关关系，对象相似性。将缺失部分剔除

是缺省方式，不在预处理函数中。