Building_Permits 数据分析与预处理

唐正 2120171060

1 数据可视化和摘要

1.1 数据摘要

1.1.1 标称数据

每个可能取值的频数（由于内容过多，选取四个显示部分）：

标称属性 <Permit Type> 频数统计

value	count
8	178844
3	14663
4	2892
2	950
6	600
7	511
1	349
5	91

标称属性 <Block> 频数统计

value	count
3708	1195
3735	750
7331	680
0289	640
3709	584
3717	578
3707	576
3721	567
3706	561
0259	554

标称属性 <Lot> 频数统计

value	count
001	10114
007	5317
002	5183
003	5042
006	4835
008	4773
009	4590
005	4549
004	4384

标称属性 <Street Number> 频数统计

value	count
1	2394
101	1153
100	1143
50	1103
201	1026
555	994
2	814
55	734

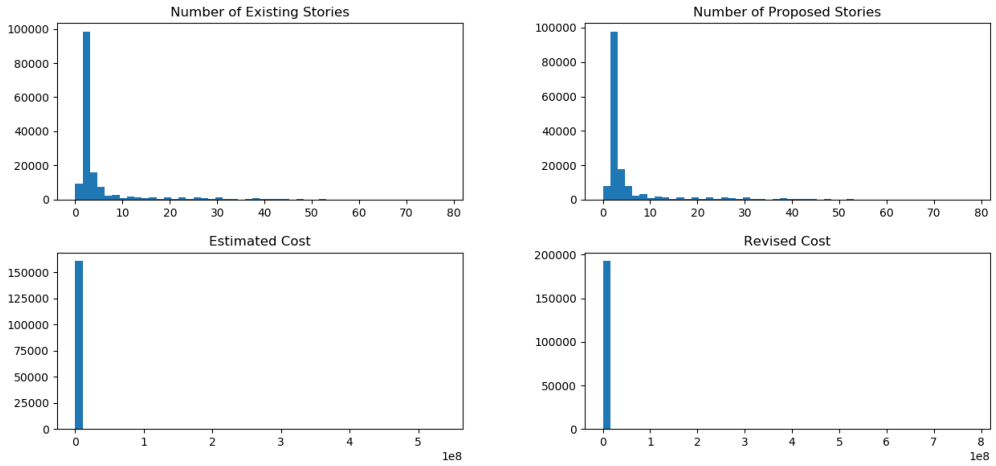
1.1.2 数值属性

数值属性的最大、最小、均值、中位数、四分位数及缺失值的个数：

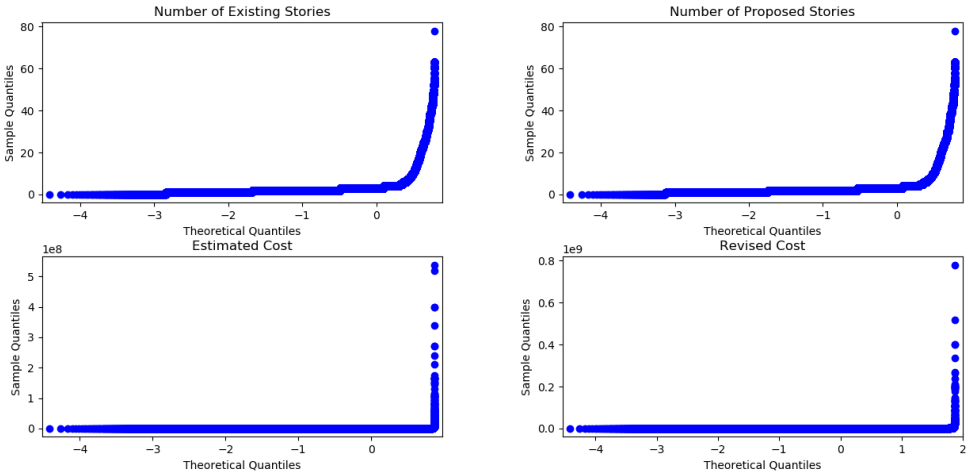
	max	min	mean	50%	25%	75%	NaN
Number of Existing Stories	78.0	0.0	5.705773	3.0	2.0	4.0	42784
Number of Proposed Stories	78.0	0.0	5.745043	3.0	2.0	4.0	42868
Estimated Cost	537958646.0	1.0	168955.443297	11000.0	3300.0	35000.0	38066
Revised Cost	780500000.0	0.0	132856.186492	7000.0	1.0	28707.5	6066
Existing Units	1907.0	0.0	15.666164	1.0	1.0	4.0	51538
Proposed Units	1911.0	0.0	16.510950	2.0	1.0	4.0	50911

1.2 数据可视化

1.2.1 直方图

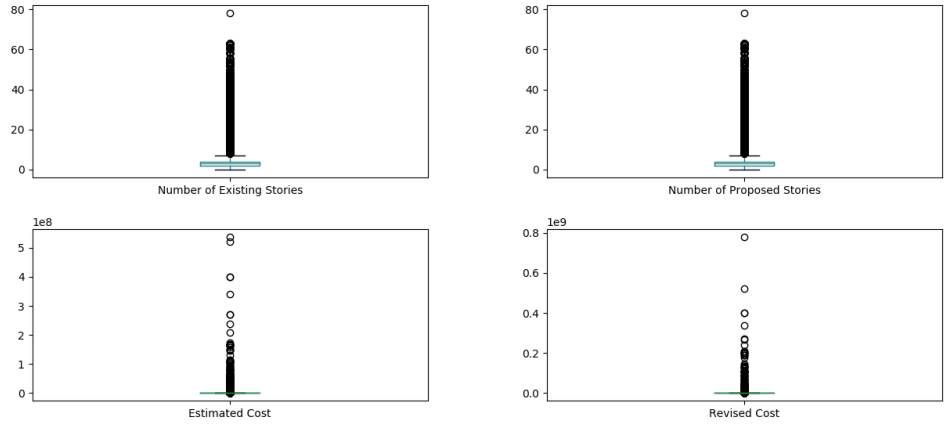


1.2.2 qq 图



由 qq 图可知图像若是近似直线的，其对应属性为正态分布态。

1.2.3 盒图



2 数据缺失处理

2.1 将缺失部分剔除

无缺失的字段：Permit Number，Permit Type，Permit Type Definition，Permit Creation Date，Block，Lot，Street Number，Street Name，Current Status，Current Status Date，Filed Date，Record ID 。

无填充意义的字段：Unit，Unit suffix，Description，Issued Date，Completed Date，First Construction Document Date，Permit Expiration Date，Existing Construction Type Description，Proposed Construction Type Description，Zipcode，Location，Street Number Suffix，Street Name Suffix，Existing Use，Existing Units，Proposed Use，Proposed Units，Plansets，Existing Construction Type，Proposed Construction Type，Supervisor District，Neighborhoods - Analysis Boundaries，Number of Existing Stories，Number of Proposed Stories，Estimated Cost，Revised Cost。

可填充的属性字段：Structural Notification，Voluntary Soft-Story Retrofit，Fire Only Permit，TIDF Compliance，Site Permit

可填充的属性字段中除了 TIDF Compliance 均为布尔型，空表示否，可用 N

填充；TIDF Compliance 字段只有两条记录不为空，空表示否，可用 N 填充，对其进行剔除。

2.2 用最高频率值来填补缺失值

无

2.3 通过属性的相关关系来填补缺失值

```
标称属性 <Structural Notification> 频数统计
value                          count
- - - - - - - - - - - - - - - - - - - - - - - - -
N                              191978
Y                              6922
-----------------------------------------------

标称属性 <Voluntary Soft-Story Retrofit> 频数统计
value                          count
- - - - - - - - - - - - - - - - - - - - - - - - -
N                              198865
Y                              35
-----------------------------------------------

标称属性 <Fire Only Permit> 频数统计
value                          count
- - - - - - - - - - - - - - - - - - - - - - - - -
N                              180073
Y                              18827
-----------------------------------------------

标称属性 <TIDF Compliance> 频数统计
value                          count
- - - - - - - - - - - - - - - - - - - - - - - - -
N                              198898
P                              1
Y                              1
-----------------------------------------------
```

2.4 通过数据对象之间的相似性来填补缺失值

无