## 作业一：数据探索性分析与数据预处理

## 数据可视化和摘要

### 1.1数据属性分析

#### 1.1.1 Building\_Permits.csv

#### 1.1.1.1标称属性

|  |  |
| --- | --- |
| Permit Type | count |
| 8 | 178844 |
| 3 | 14663 |
| 4 | 2892 |
| 2 | 950 |
| 6 | 600 |
| 7 | 511 |
| 1 | 349 |
| 5 | 91 |

|  |  |
| --- | --- |
| Street Suffix |  |
| St | 138358 |
| Av | 43219 |
| Bl | 3555 |
| Wy | 3540 |
| Dr | 3267 |
| Tr | 1466 |
| Ct | 667 |
| Pl | 538 |
| Rd | 389 |
| Ln | 354 |
| Hy | 240 |
| Pz | 210 |
| Pk | 128 |
| Cr | 97 |
| Al | 83 |
| Wk | 9 |
| Rw | 5 |
| So | 2 |
| Sw | 2 |
| No | 2 |
| Hl | 1 |

|  |  |
| --- | --- |
| Existing Construction Type |  |
| 5 | 113350 |
| 1 | 28072 |
| 3 | 9663 |
| 2 | 4068 |
| 4 | 381 |

|  |  |
| --- | --- |
| Structural Notification |  |
| Y | 6922 |

|  |  |
| --- | --- |
| Voluntary Soft-Story Retrofit |  |
| Y | 35 |

|  |  |
| --- | --- |
| Fire Only Permit |  |
| Y | 18827 |

|  |  |
| --- | --- |
| Current Status |  |
| complete | 97077 |
| issued | 83559 |
| filed | 12043 |
| withdrawn | 1754 |
| cancelled | 1536 |
| expired | 1370 |
| approved | 733 |
| reinstated | 563 |
| suspend | 193 |
| revoked | 50 |
| plancheck | 16 |
| appeal | 2 |
| incomplete | 2 |
| disapproved | 2 |

|  |  |
| --- | --- |
| Proposed Construction Type |  |
| 5 | 114382 |
| 1 | 27841 |
| 3 | 9360 |
| 2 | 3778 |
| 4 | 377 |

|  |  |
| --- | --- |
| Site Permit |  |
| Y | 5359 |

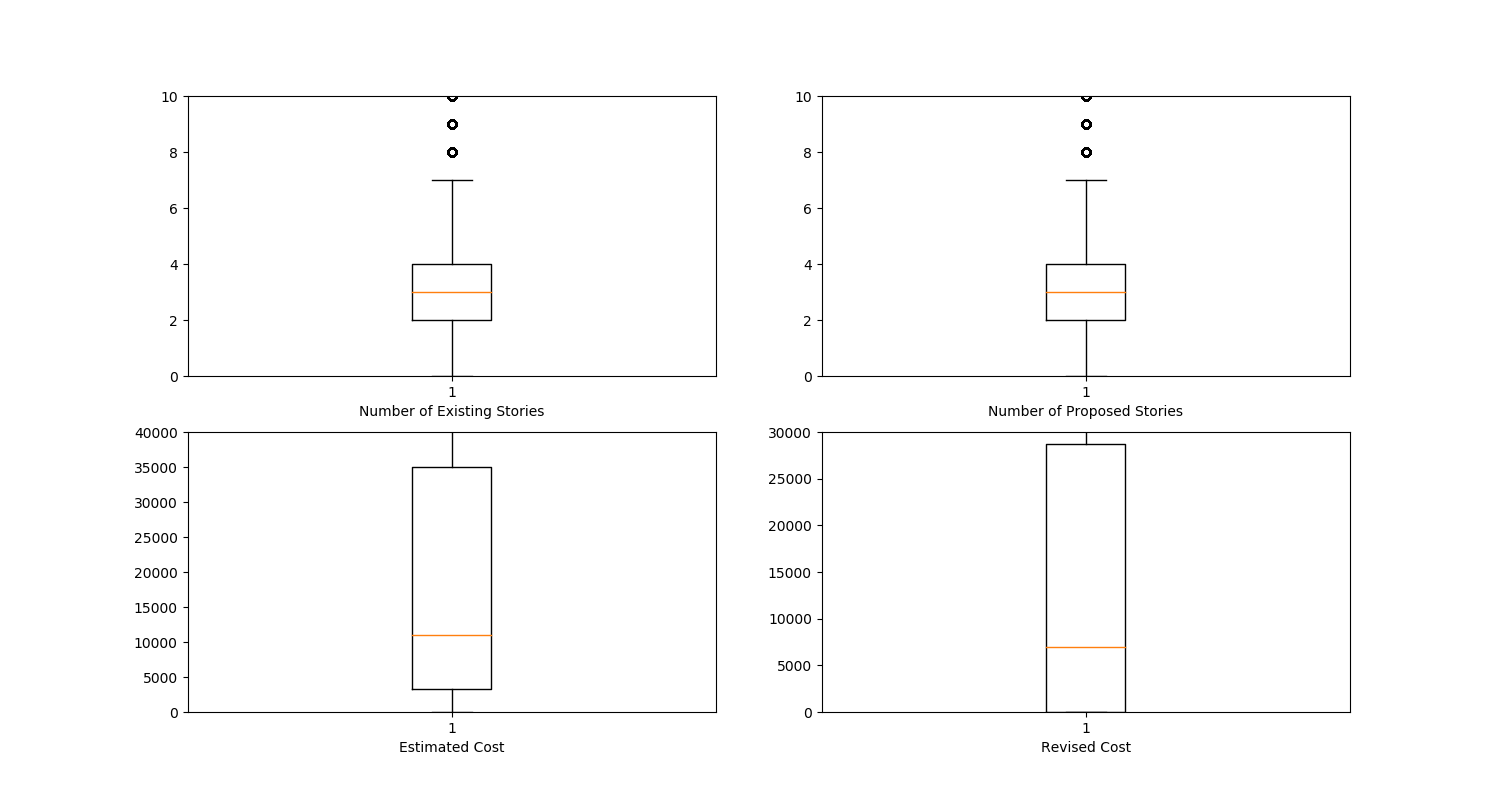
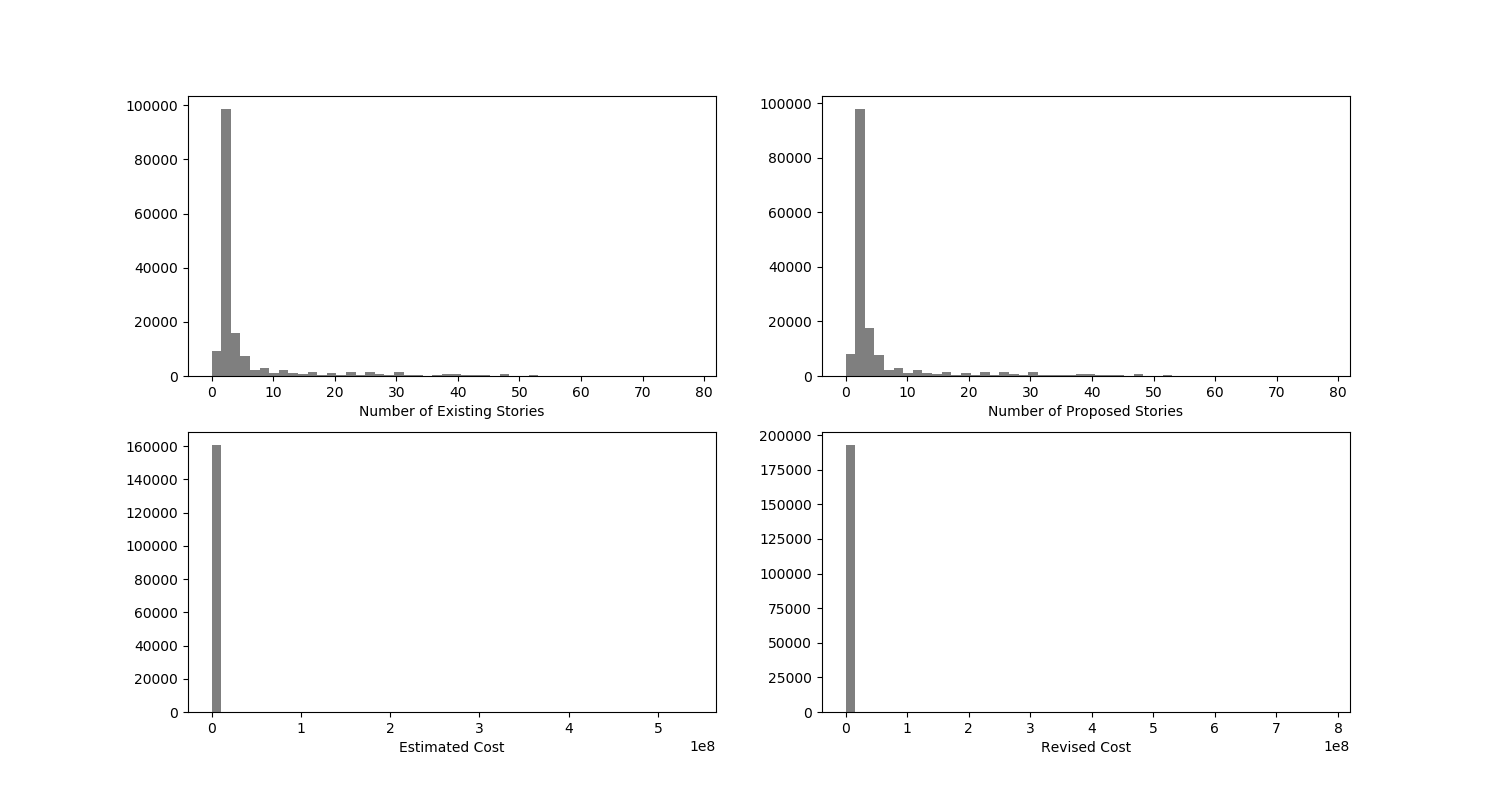
|  |  |
| --- | --- |
| Supervisor District |  |
| 3 | 28649 |
| 8 | 26760 |
| 2 | 25483 |
| 6 | 24797 |
| 5 | 19045 |
| 9 | 16362 |
| 7 | 14365 |
| 1 | 13038 |
| 10 | 12152 |
| 4 | 9592 |
| 11 | 6940 |

|  |  |
| --- | --- |
| Neighborhoods - Analysis Boundaries |  |
| Financial District/South Beach | 21816 |
| Mission | 14681 |
| Sunset/Parkside | 10207 |
| West of Twin Peaks | 8739 |
| Castro/Upper Market | 8527 |
| Pacific Heights | 8508 |
| Marina | 8244 |
| Outer Richmond | 7854 |
| Noe Valley | 7844 |
| South of Market | 7572 |
| Bernal Heights | 6067 |
| Nob Hill | 6009 |
| Haight Ashbury | 5798 |
| Inner Sunset | 5776 |
| Bayview Hunters Point | 5669 |
| Russian Hill | 5495 |
| Hayes Valley | 5489 |
| Tenderloin | 4783 |
| Inner Richmond | 4458 |
| Potrero Hill | 4293 |
| Presidio Heights | 4084 |
| North Beach | 4054 |
| Western Addition | 3867 |
| Chinatown | 3765 |
| Lone Mountain/USF | 3358 |
| Excelsior | 3072 |
| Oceanview/Merced/Ingleside | 2654 |
| Glen Park | 2637 |
| Mission Bay | 2287 |
| Outer Mission | 2242 |
| Twin Peaks | 1702 |
| Portola | 1433 |
| Lakeshore | 1308 |
| Seacliff | 992 |
| Visitacion Valley | 900 |
| Japantown | 700 |
| Treasure Island | 81 |
| Golden Gate Park | 64 |
| Presidio | 51 |
| Lincoln Park | 49 |
| McLaren Park | 46 |

#### 1.1.2数值属性

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Number of Existing Stories | Number of Proposed Stories | Estimated Cost | Revised Cost |
| count | 156116 | 156032 | 160834 | 192834 |
| mean | 5.705773271 | 5.745042684 | 168955.4433 | 132856.1865 |
| std | 8.613454685 | 8.613283687 | 3630385.954 | 3584902.592 |
| min | 0 | 0 | 1 | 0 |
| 25% | 2 | 2 | 3300 | 1 |
| 50% | 3 | 3 | 11000 | 7000 |
| 75% | 4 | 4 | 35000 | 28707.5 |
| max | 78 | 78 | 537958646 | 780500000 |
| mode | 2 | 2 | 1 | 1 |
| loss\_count | 42783 | 42867 | 38065 | 6065 |

可视化图



#### 1.1.2 NFL Play By Play 2009-2017.csv

1.1.2.1 标称属性

nominal\_property = ['Drive', 'qtr', 'down', 'SideofField', 'ydstogo', 'GoalToGo', 'FirstDown', 'posteam', 'DefensiveTeam', 'PlayAttempted', 'sp', 'Touchdown', 'ExPointResult', 'TwoPointConv', 'DefTwoPoint', 'Safety', 'Onsidekick', 'PuntResult', 'PlayType', 'Passer', 'Passer\_ID', 'PassAttempt', 'PassOutcome', 'PassLength', 'QBHit', 'PassLocation', 'InterceptionThrown', 'Interceptor', 'Rusher', 'Rusher\_ID', 'RushAttempt', 'RunLocation', 'RunGap', 'Receiver', 'Receiver\_ID', 'Reception', 'ReturnResult', 'Returner', 'BlockingPlayer', 'Tackler1', 'Tackler2', 'FieldGoalResult', 'Fumble', 'RecFumbTeam', 'RecFumbPlayer', 'Sack', 'Challenge.Replay', 'ChalReplayResult', 'Accepted.Penalty', 'PenalizedTeam', 'PenaltyType', 'PenalizedPlayer', 'HomeTeam', 'AwayTeam', 'Timeout\_Indicator', 'Timeout\_Team', 'Season']

1.1.2.2 数值属性

numercial\_property = ['TimeUnder', 'TimeSecs', 'PlayTimeDiff', 'yrdln', 'yrdline100', 'ydsnet', 'Yards.Gained', 'AirYards', 'YardsAfterCatch', 'FieldGoalDistance', 'Penalty.Yards', 'PosTeamScore', 'DefTeamScore', 'ScoreDiff', 'AbsScoreDiff', 'posteam\_timeouts\_pre', 'HomeTimeouts\_Remaining\_Pre', 'AwayTimeouts\_Remaining\_Pre', 'HomeTimeouts\_Remaining\_Post', 'AwayTimeouts\_Remaining\_Post', 'No\_Score\_Prob', 'Opp\_Field\_Goal\_Prob', 'Opp\_Safety\_Prob', 'Opp\_Touchdown\_Prob', 'Field\_Goal\_Prob', 'Safety\_Prob', 'Touchdown\_Prob', 'ExPoint\_Prob', 'TwoPoint\_Prob', 'ExpPts', 'EPA', 'airEPA', 'yacEPA', 'Home\_WP\_pre', 'Away\_WP\_pre', 'Home\_WP\_post', 'Away\_WP\_post', 'Win\_Prob', 'WPA', 'airWPA', 'yacWPA']

1.1.1.3分析结果

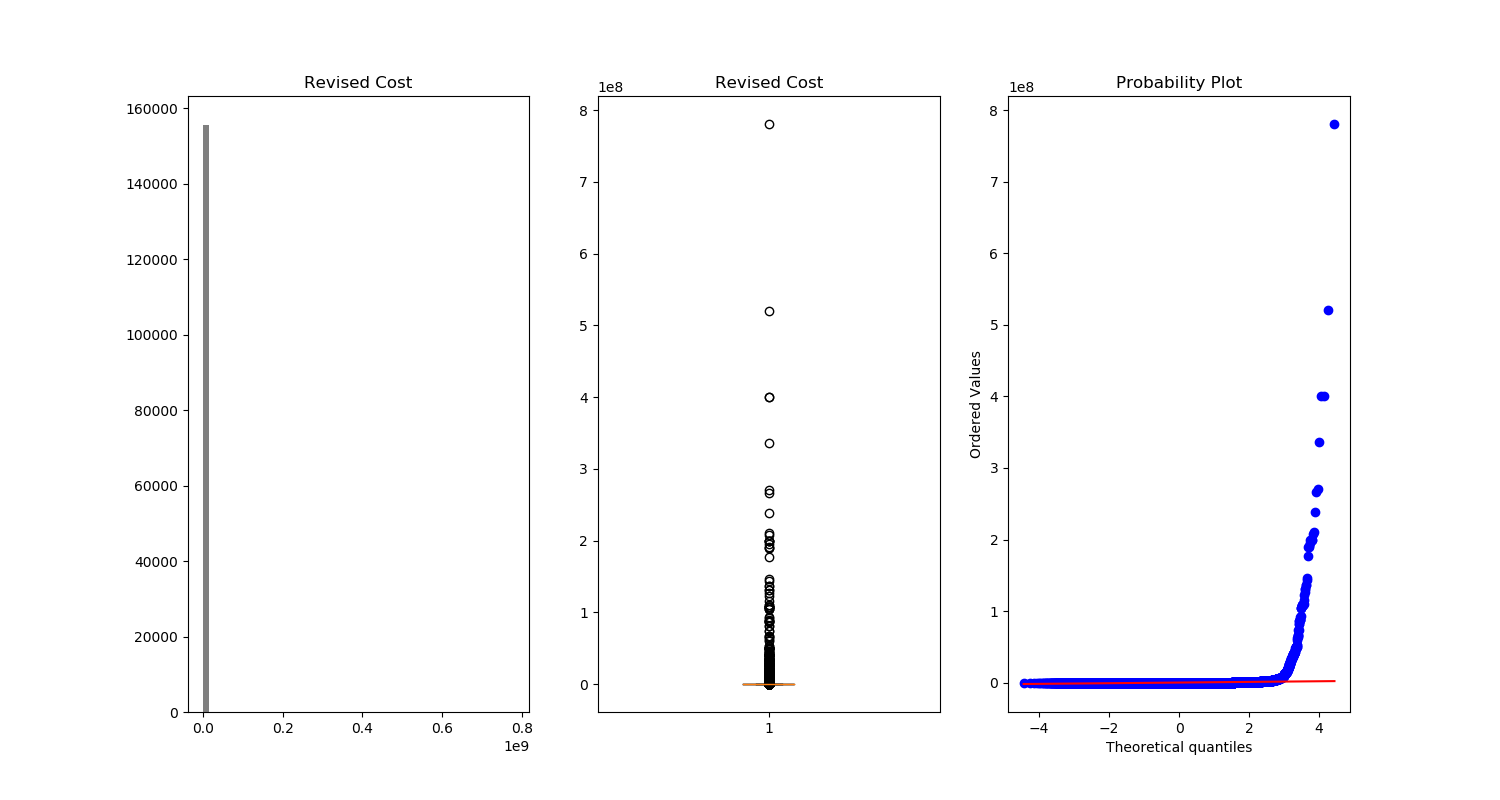
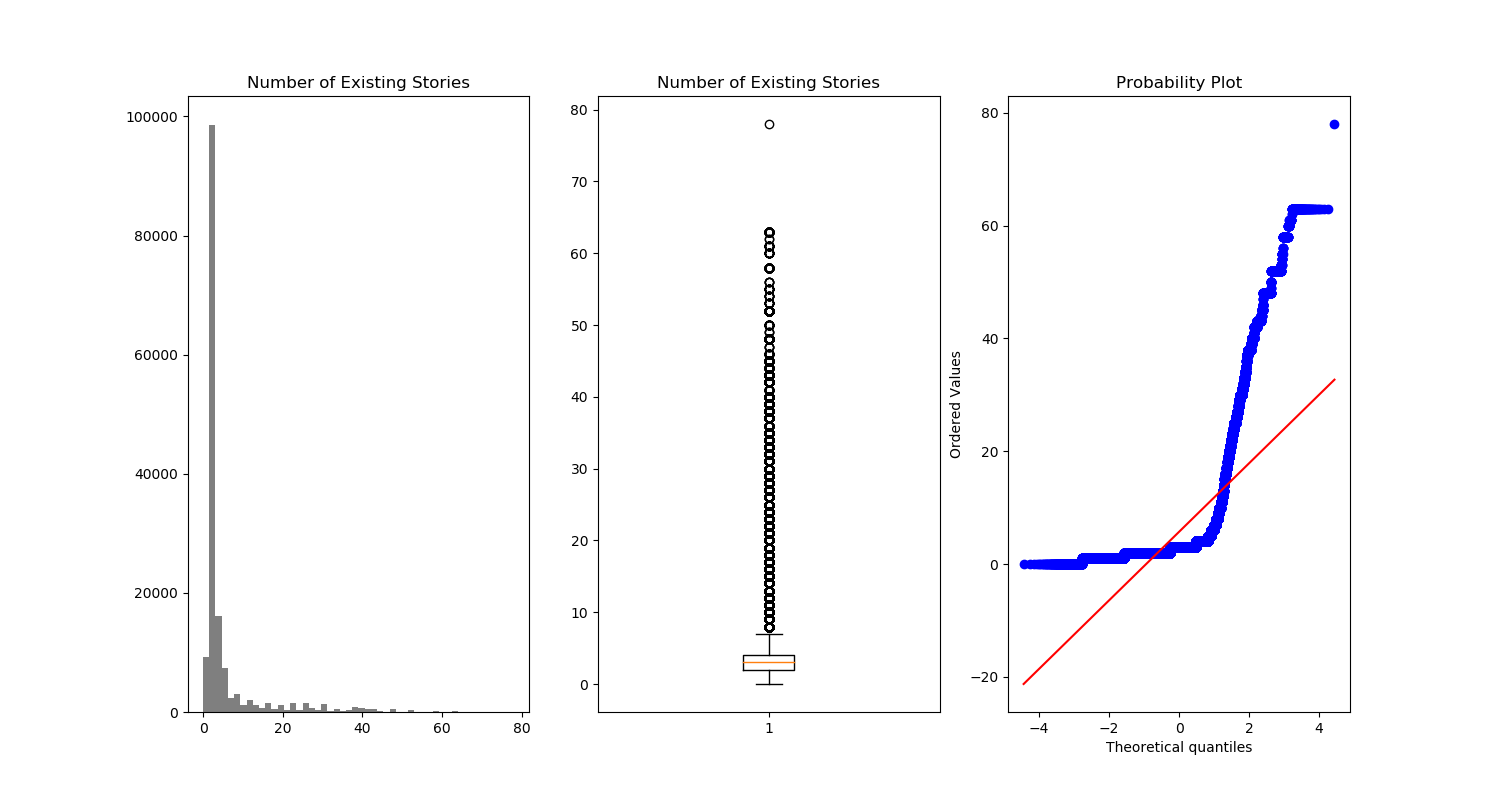
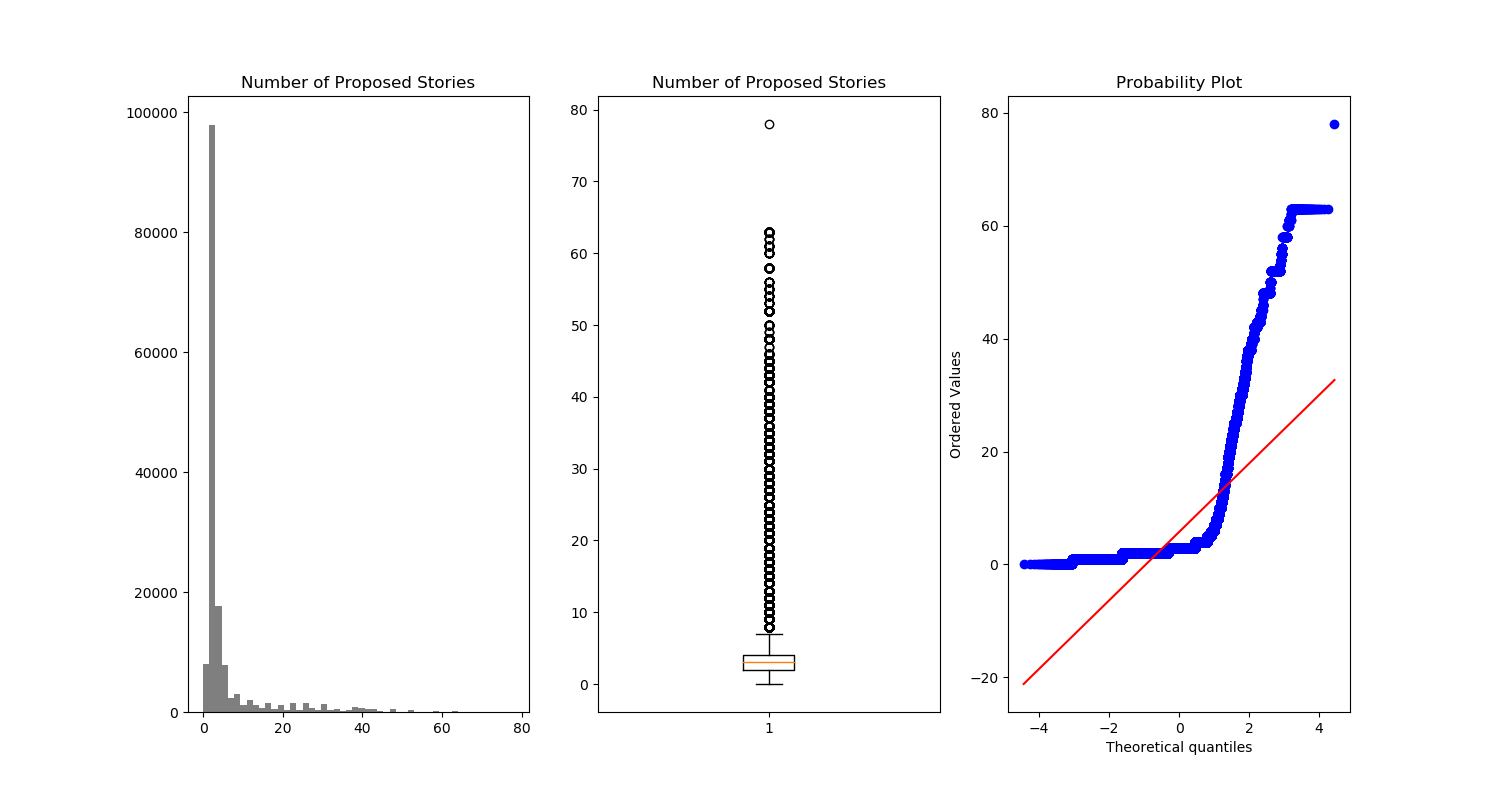
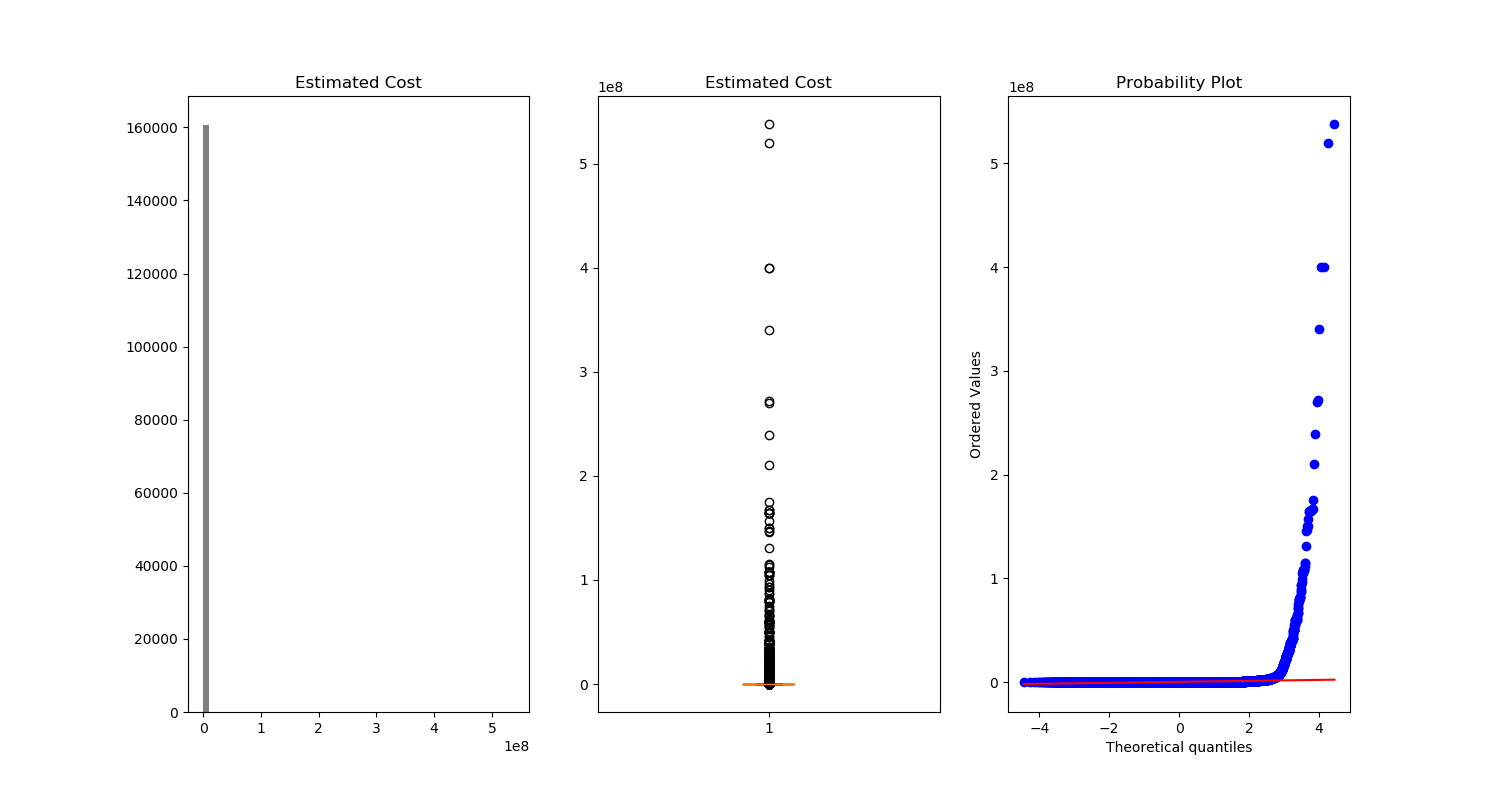
见附件

## 数据缺失处理

### 2.1将缺失部分剔除

数据分析部分即通过删除缺失数据进行分析的

关键代码：data[i].dropna()



### 2.2用最高频率值来填补缺失值

使用value\_counts函数统计每列不同数值出现的频率

该函数回自动按频次进行排序，因此取第一行值即可，然后按照该值为缺失位赋值。

def highest\_freq\_fill(part\_column):

top\_data = {}

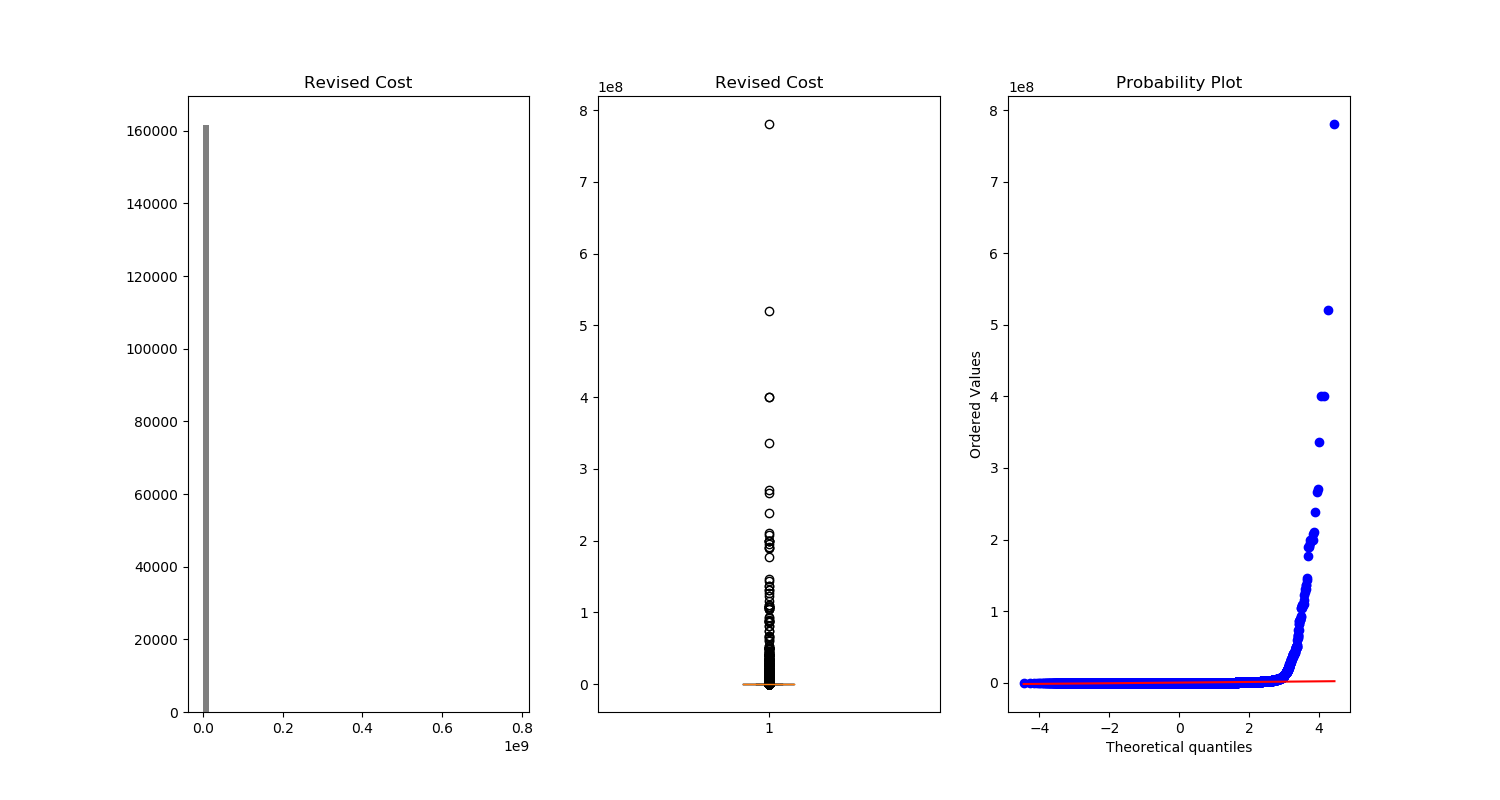
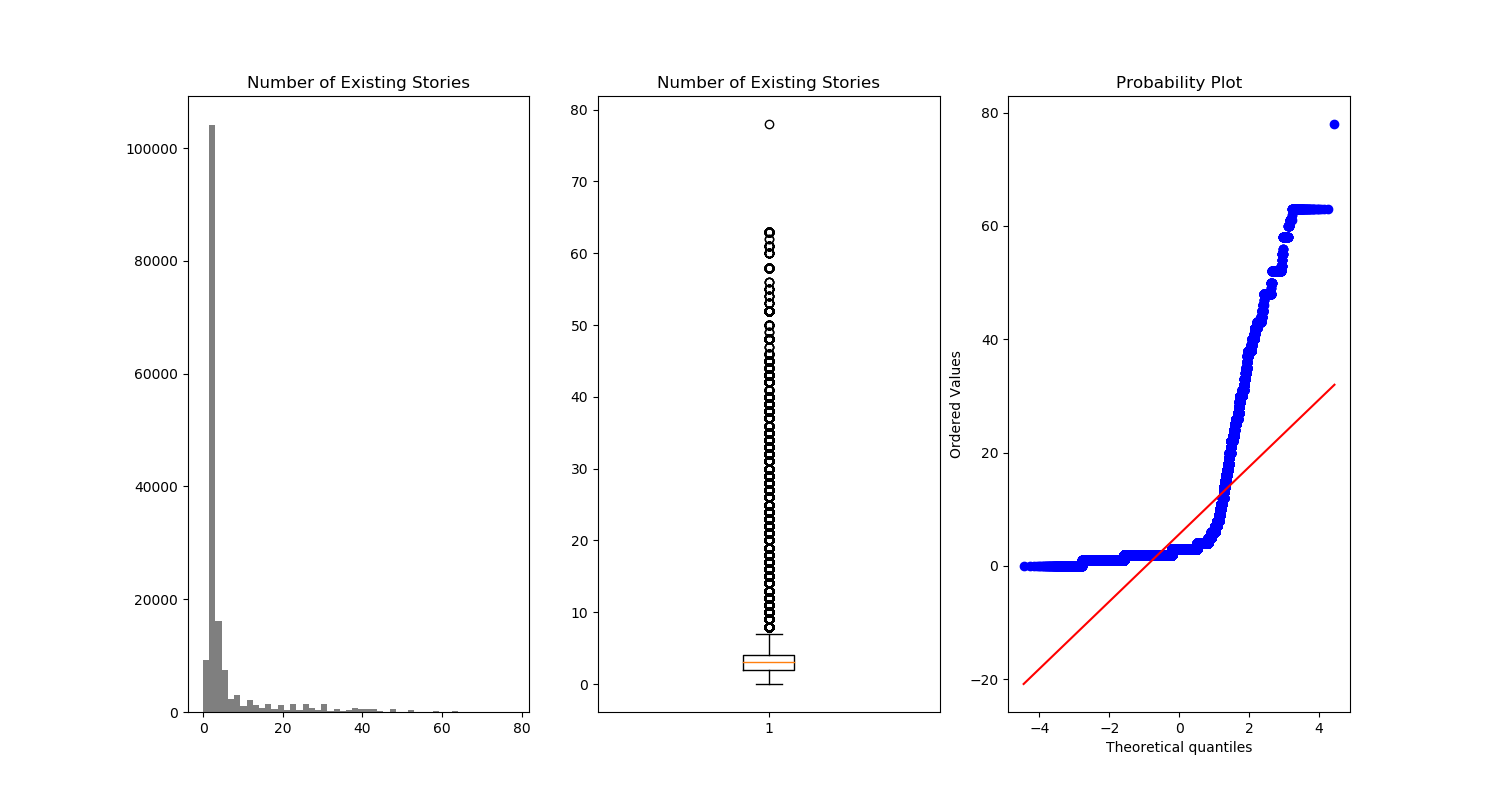
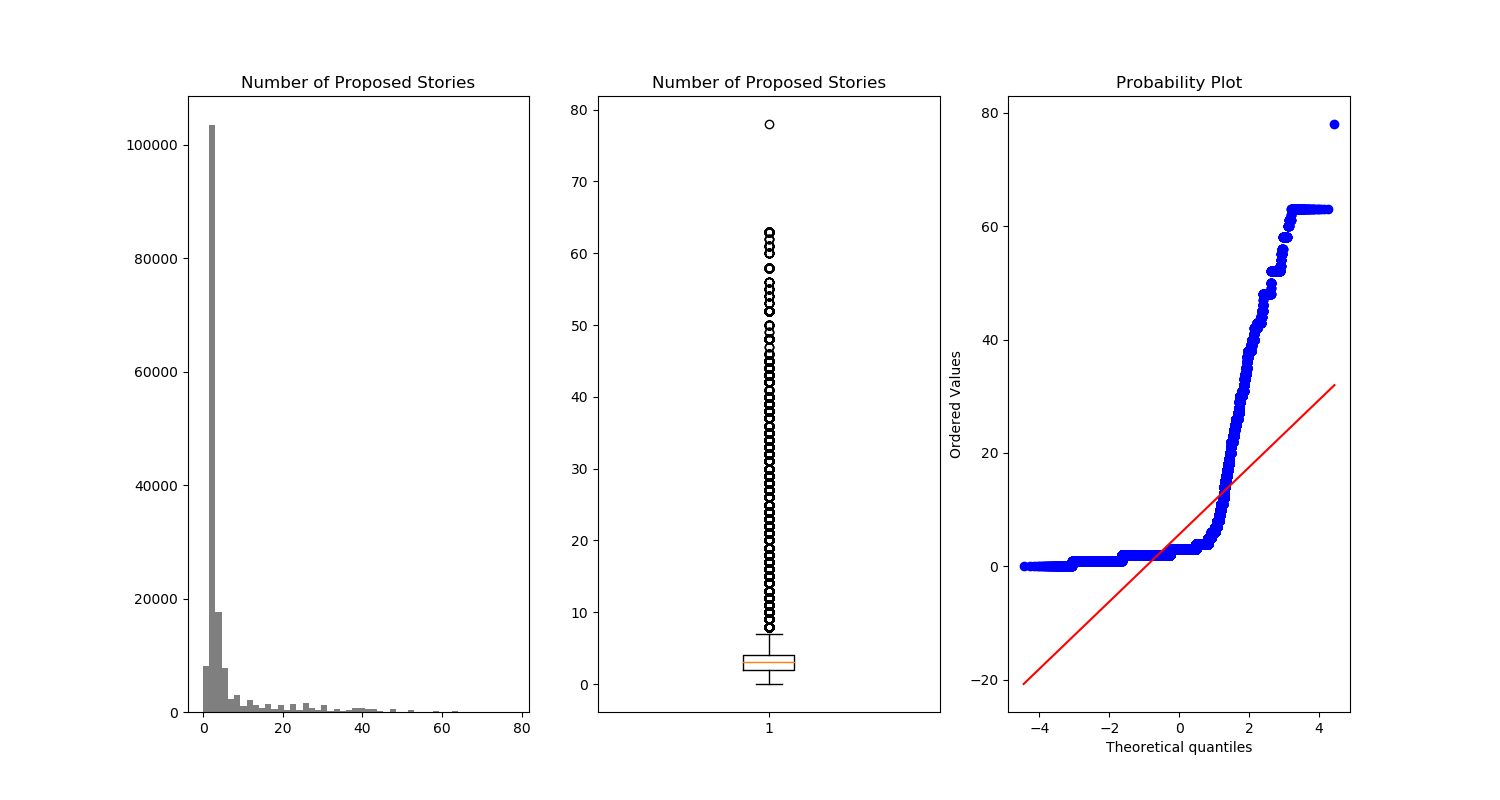
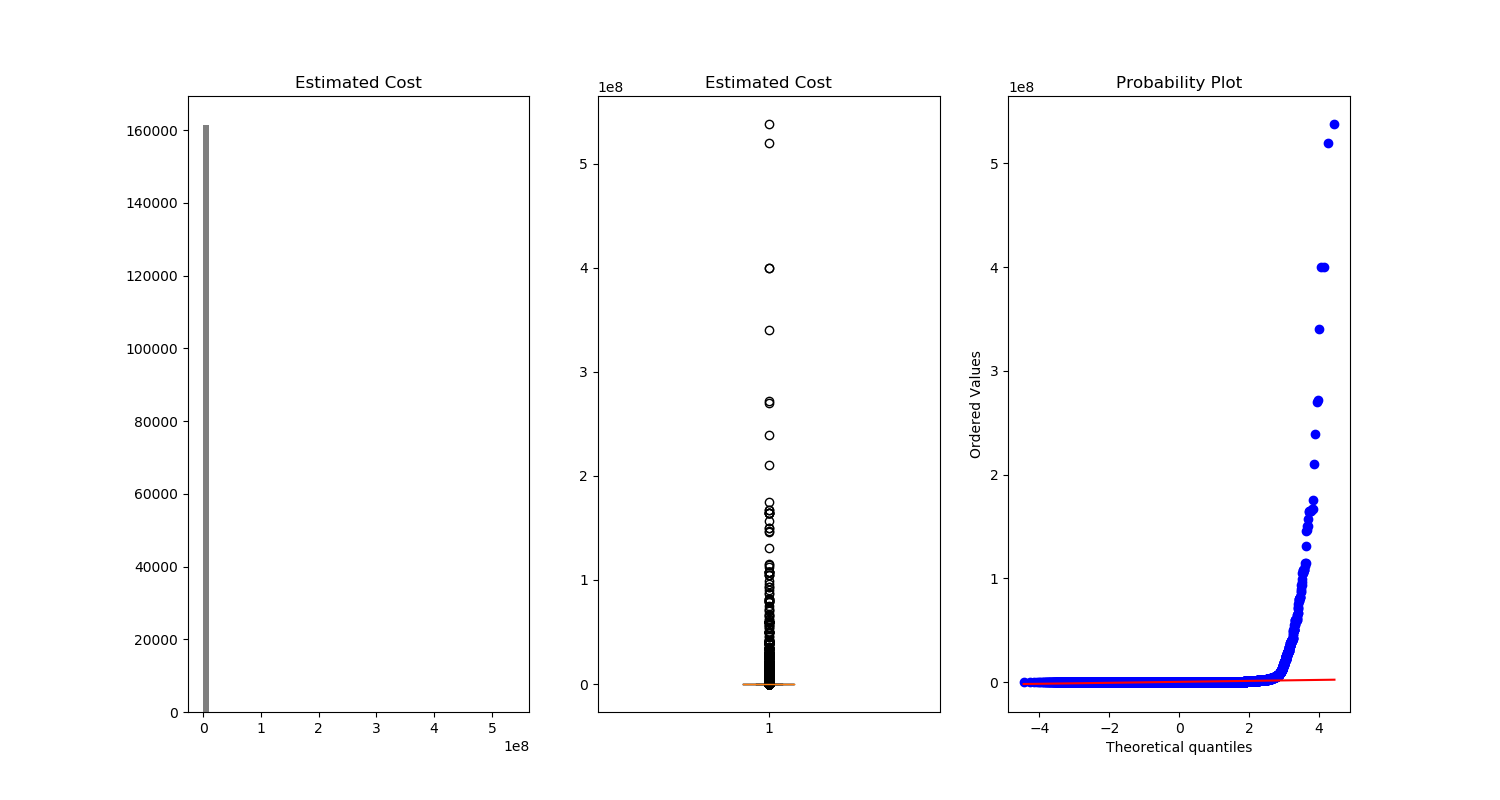
for i in part\_column:

top\_data[i] = sf\_permits[[i]].apply(pd.value\_counts).index[0] # class df

print(top\_data)

sf\_permits.fillna(top\_data).drop(["Unnamed: 0"],axis=1).to\_csv("data/Part\_Building\_Permits\_2.csv")

结果：



### 2.3通过属性的相关关系来填补缺失值

Number of Existing Stories和Number of Proposed Stories相关度较高，Estimated Cost 和Revised Cost相关度较高，对于这几个属性的缺失值用相关属性对应数值填充。

def property\_relation\_fill(data):

x\_1 = 'Number of Existing Stories'

y\_1 = 'Number of Proposed Stories'

x\_2 = 'Estimated Cost'

y\_2 = 'Revised Cost'

for j in range(len(data)):

if (data[y\_1].isnull())[j] and (data[x\_1].notnull())[j] : # 如果为空即插值。

data[y\_1][j] = data[x\_1][j]

for j in range(len(data)):

if (data[x\_1].isnull())[j] and (data[y\_1].notnull())[j] : # 如果为空即插值。

data[x\_1][j] = data[y\_1][j]

for j in range(len(data)):

if (data[y\_2].isnull())[j] and (data[x\_2].notnull())[j] : # 如果为空即插值。

data[y\_2][j] = data[x\_2][j]

for j in range(len(data)):

if (data[x\_2].isnull())[j] and (data[y\_2].notnull())[j] : # 如果为空即插值。

data[x\_2][j] = data[y\_2][j]

print(data)

data.drop(["Unnamed: 0"],axis=1).to\_csv("data/Part\_Building\_Permits\_3.csv")

结果：

### 2.4通过数据对象之间的相似性来填补缺失值

不会做，尝试失败

**3.总体分析**

1. 数据缺失的原因

通过观察数据，发现属性“Permits Number”中以“M”开头的项，从第21列开始（即数据文件中属性“Number of Existing Stories”）基本为空，也就是说这部分属性下记录不存在，而非未被记录导致的缺失。其次，还有少量数据可能是记录丢失导致的。

1. 缺失处理后数据统计

通过处理后数据的统计信息对比新旧数据集。

1. 去除缺失值

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Number of Existing Stories | Number of Proposed Stories | Estimated Cost | Revised Cost |
| count | 156116 | 156032 | 160834 | 155634 |
| mean | 5.705773271 | 5.745042684 | 168955.4433 | 164607.1916 |
| std | 8.613454685 | 8.613283687 | 3630385.954 | 3989752.146 |
| min | 0 | 0 | 1 | 0 |
| 25% | 2 | 2 | 3300 | 3000 |
| 50% | 3 | 3 | 11000 | 12000 |
| 75% | 4 | 4 | 35000 | 40000 |
| max | 78 | 78 | 537958646 | 780500000 |
| mode | 2 | 2 | 1 | 1 |
| loss\_count | 5521 | 5605 | 803 | 6003 |

1. 按频率最高

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Number of Existing Stories | Number of Proposed Stories | Estimated Cost | Revised Cost |
| count | 161637 | 161637 | 161637 | 161637 |
| mean | 5.579195976 | 5.615177837 | 168116.0908 | 158493.9194 |
| std | 8.4917894 | 8.490319602 | 3621376.416 | 3915087.215 |
| min | 0 | 0 | 1 | 0 |
| 25% | 2 | 2 | 3000 | 2500 |
| 50% | 3 | 3 | 11000 | 10420 |
| 75% | 4 | 4 | 35000 | 37000 |
| max | 78 | 78 | 537958646 | 780500000 |
| mode | 2 | 2 | 1 | 1 |
| loss\_count | 0 | 0 | 0 | 0 |

（3）按属性相关性

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Number of Existing Stories | Number of Proposed Stories | Estimated Cost | Revised Cost |
| count | 159526 | 159526 | 161095 | 161095 |
| mean | 5.714657799 | 5.748451663 | 168854.7768 | 188237.7308 |
| std | 8.606952109 | 8.611304675 | 3627977.104 | 4058311.802 |
| min | 0 | 0 | 0 | 0 |
| 25% | 2 | 2 | 3200 | 3000 |
| 50% | 3 | 3 | 11000 | 12000 |
| 75% | 4 | 4 | 35000 | 40000 |
| max | 78 | 78 | 537958646 | 780500000 |
| mode | 2 | 2 | 1 | 1 |
| loss\_count | 2111 | 2111 | 542 | 542 |

通过观察可以发现数据总体数值差别不大，缺失数据相对总体数据量较小，如果不对数据做进一步清洗和筛选，结果不会有太大变化，正如以上几幅图所示，基本看不出区别。

1. 总结

若想得到更好的结果，需要对不合理的数据进行清洗，比如，去掉首尾端点部分值，观察中间集中部分值的分布情况。其次，需要对数据有更深层次的理解，需要人为的主观判断，而非依赖机器。

1. **程序及文件描述**

build\_data\_clean.py：删除数据空行空列及第一列以“M”开始的行。

data\_analysis\_1\_1.py：主要分析程序，包括标称数据的词频统计，数值属性的统计量和直方图、盒图、qq图的绘制。

missing\_data\_process\_1\_2.py：缺失数据处理生成新数据集

data1和data2文件夹保存两个数据集生成的文件，figure为绘制的图表，norminal为标称属性生成结果，每个属性对应一个文件，desc.csv为数值属性统计数据结果。

data：缺失处理后的数据集文件

