分析报告—数据挖掘作业一

姓名：王 钰

学号：2120171076

**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

**4、实验环境及语言**

* **语言及环境依赖**
* 语言： python
* 环境：windows
* 依赖的包：xlrd, pylab, matplotlib, scipy, numpy
* xlrd: 数据摘要处理时用到
* pylab, matplotlib, scipy, numpy：数据可视化时用于生成图

**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

依赖包：xlrd

**（1）. 把原始数据导入excel表格**

经过对数据集的初步整理，去除了没有意义的属性。分离出标称属性和数值属性，保存为excl文档，作为新的数据集形式，命名为dataset1、dataset2。

**（2）. 利用python脚本实现数据摘要的获得(以数据集一为例，数据集二同理)**

脚本为：dataset1.py ，该脚本涉及到的关键函数如下：

init() #从excel获取数据，分离出每一个属性的数据集

nominalDatacost #对标称属性，给出每个可能取值的频数，

statistic(chemicalParameters,frequency) #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

**（3）. 统计结果保存为json数据(以数据集一为例，数据集二同理)**

nominalDatacost.json 是标称属性每个可能取值的频数统计结果

其格式如下：

{

"Permit Type": {

"4.0": 2892,

"3.0": 14663,

"8.0": 178844,

"6.0": 600,

"7.0": 511,

"1.0": 349,

"5.0": 91,

"2.0": 950

},

**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

依赖包：pylab, matplotlib, scipy, numpy

**（1）. 提取出数值数据，以json格式保存**

利用“数据摘要”的方法，把原数据中的数值数据提取出来，以json格式保存，结果是data.json。此部分用到的脚本为cleaning.py。

**（2）. 绘制直方图，qq图，盒图**

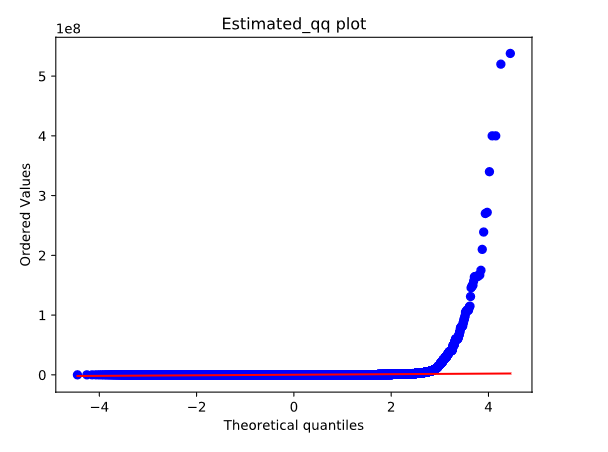
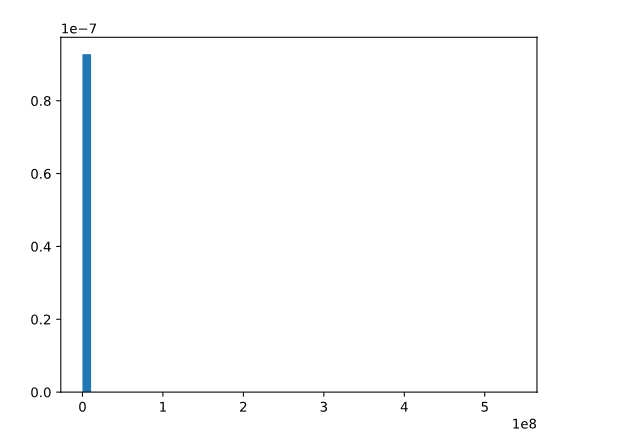
+ 直方图的绘制方法：主要用到的是pylab.hist()方法

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

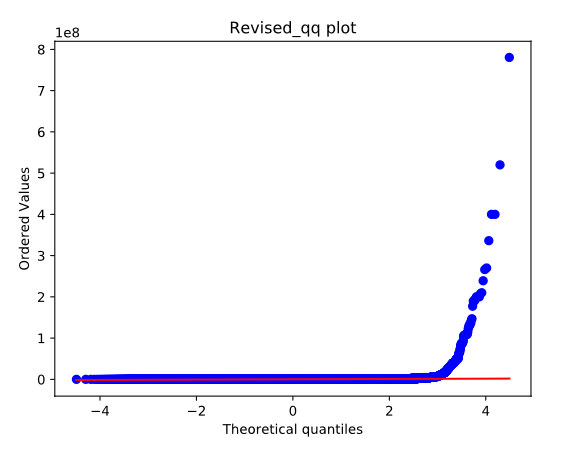
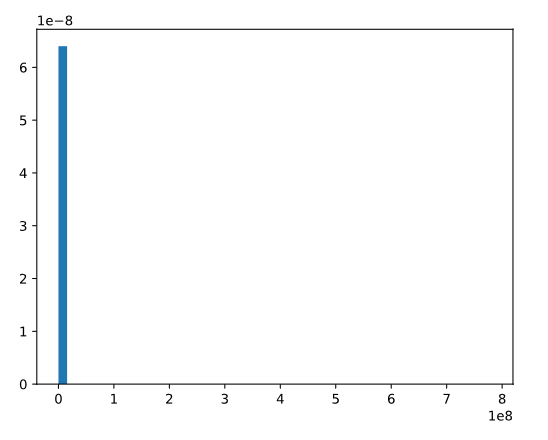
+ 盒图的给制方法：主要用到的是boxplot()方法

**（3）. 结果展示**结果图分别以svg和png两种格式保存：

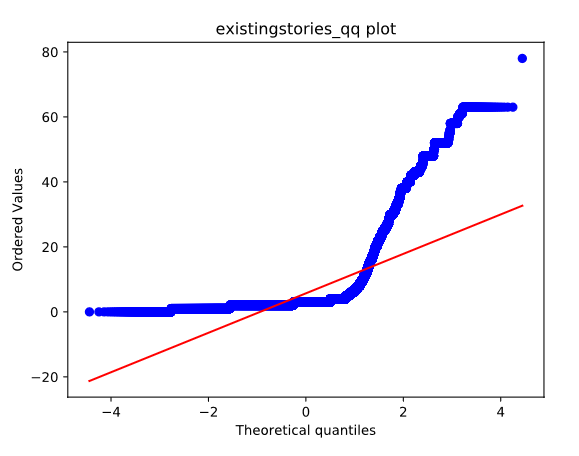
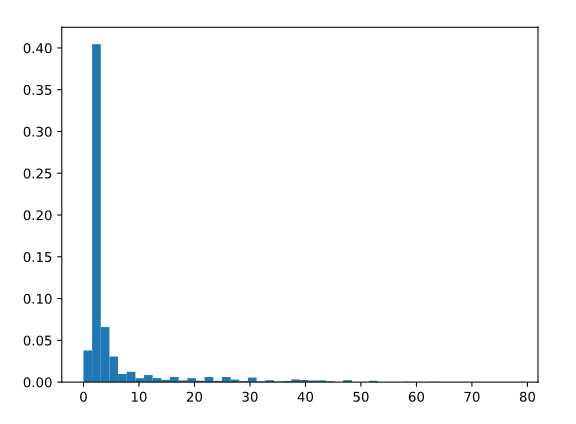
Dataset1:



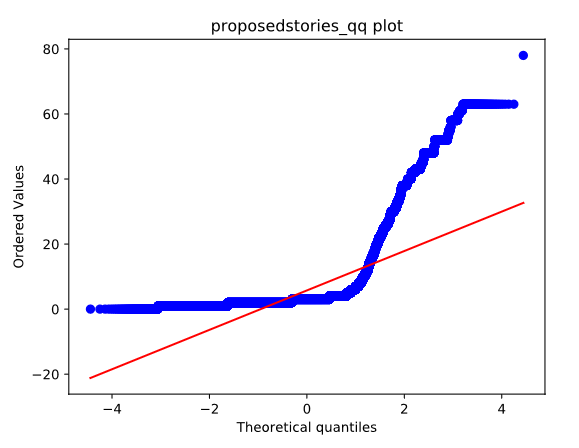
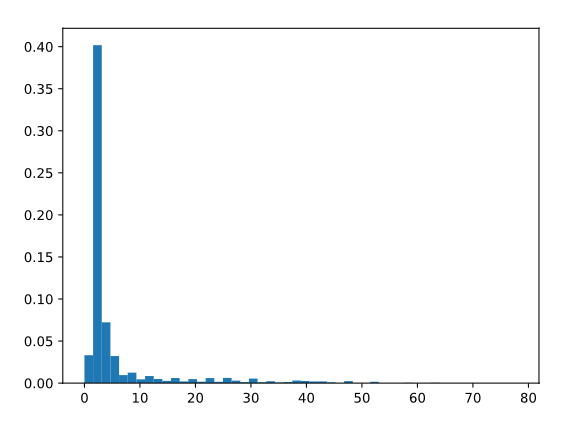
Estimated Cost属性的直方图和qq图



Rivised Cost属性的直方图和qq图

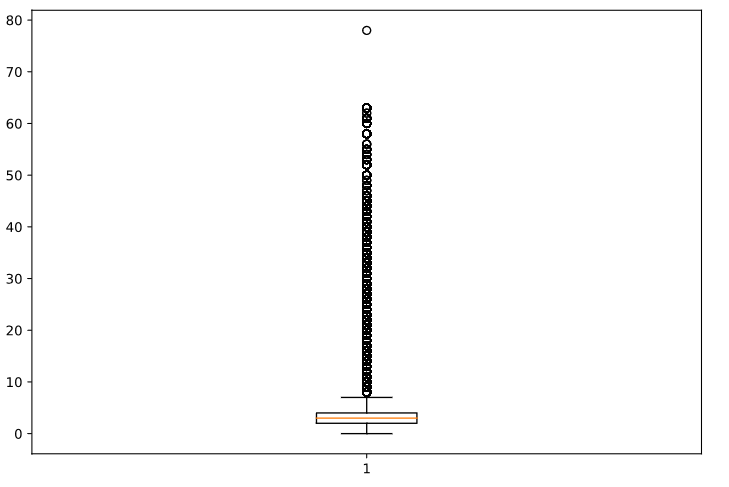


existingstories属性的直方图和qq图

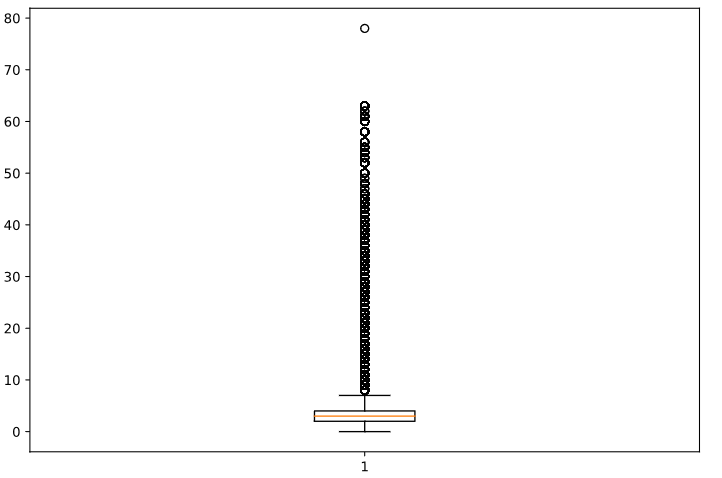


proposedstories属性的直方图和qq图

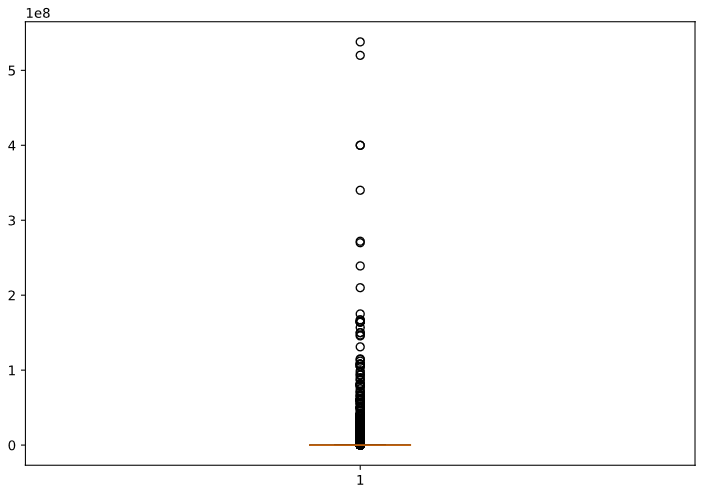
盒图：



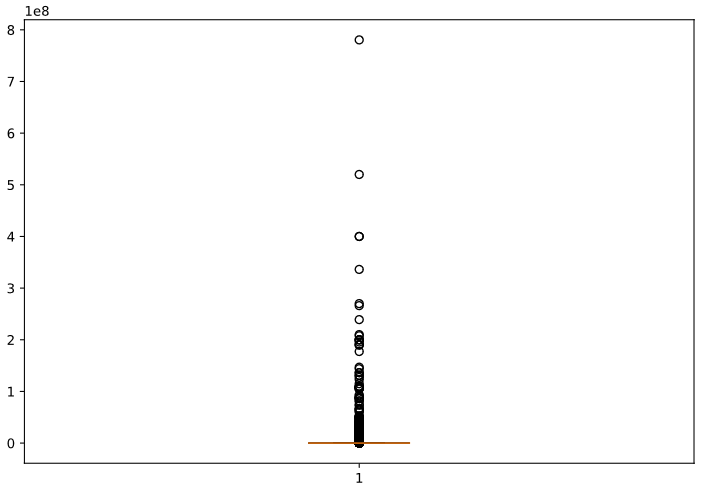
Estimated Cost盒图



Rivised Cost盒图



Existingstories盒图



Proposedstories盒图