**数据挖掘第一次作业报告**

计算机学院 2120171017 黄焯恒

**数据集1 NFL Play by Play 2009-2017 (v4).csv**

对标称属性，给出每个可能取值的频数，部分分析结果如下，完整结果存储在result\_1.txt中：

PlayType

Pass:159353

Run:120831

Kickoff:23403

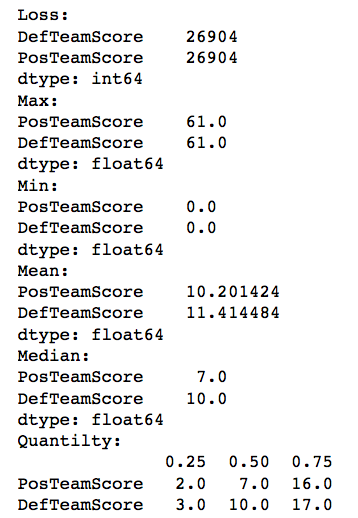
Punt:22003

No Play:21414

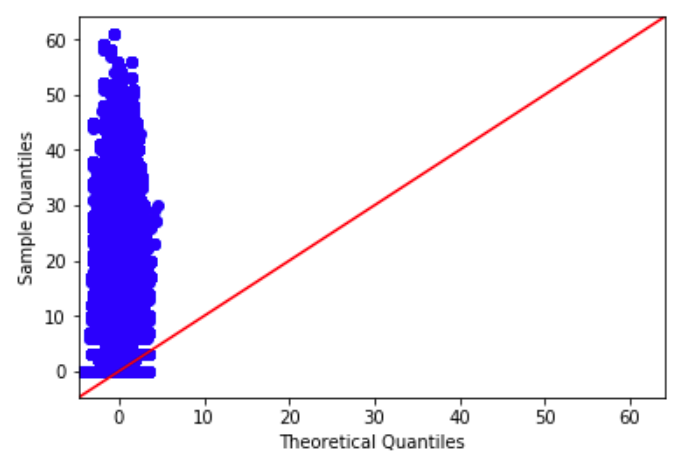
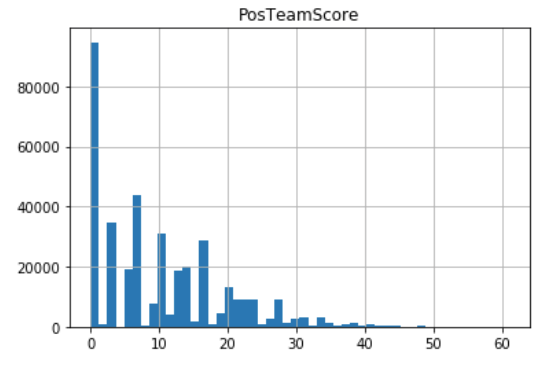
Timeout:16206

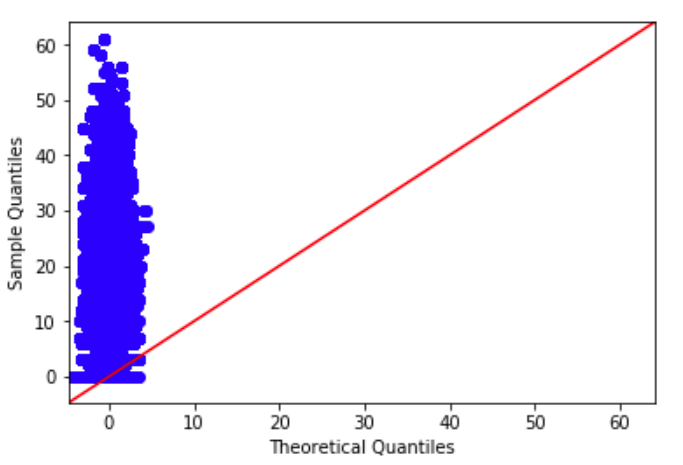
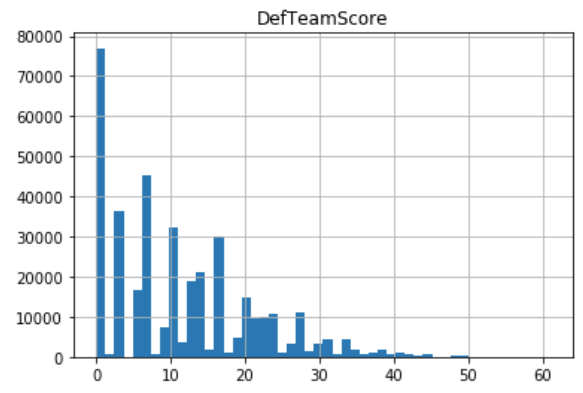
……

数值属性分析结果：

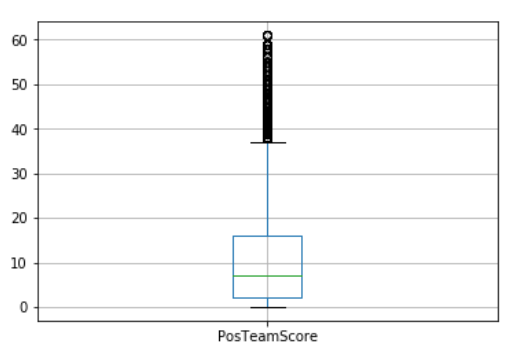
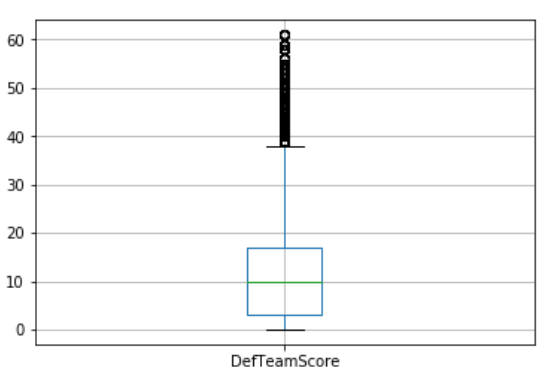


针对数值属性，绘制直方图，用qq图检验其分布是否为正态分布。





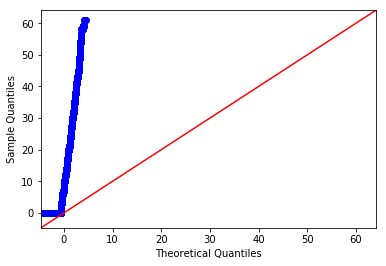
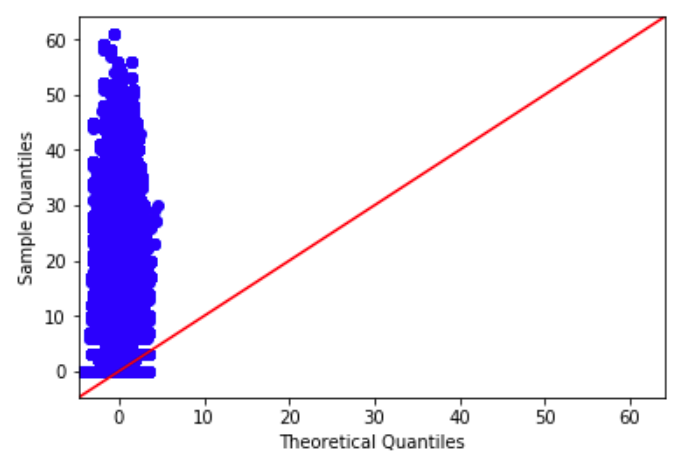
绘制出DefTeamScore和PosTeamScore的直方图和q-q图，两个数值属性不符合正态分布。分别绘制出的盒图如下：

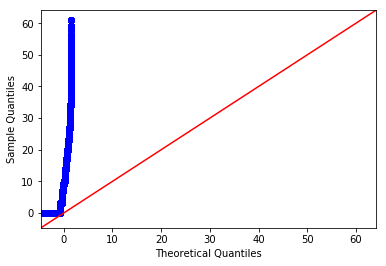
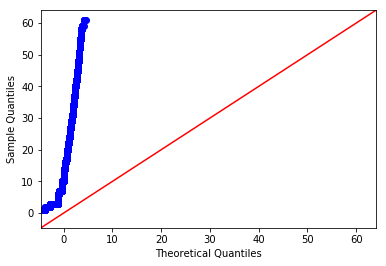


**缺失值处理**

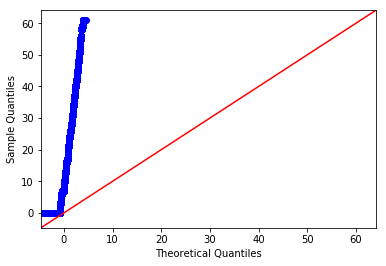
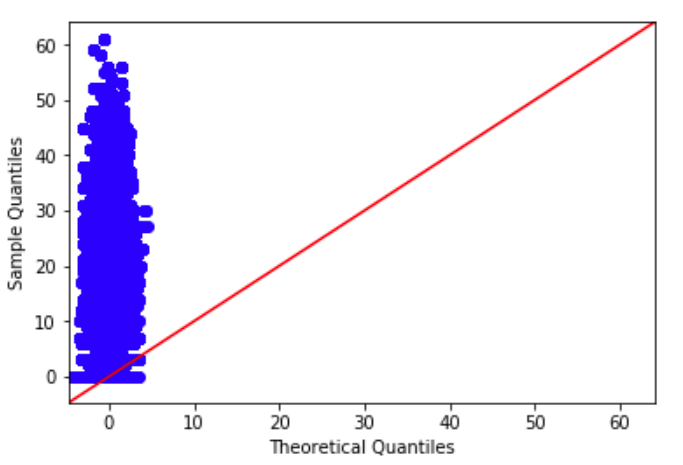
对两个数值属性进行缺失值处理：（1）将缺失部分剔除；（2）用最高频率值来填补缺失值；（3）通过属性的相关关系来填补缺失值；（4）通过数据对象之间的相似性来填补缺失值。处理之后绘制q-q图来观察,

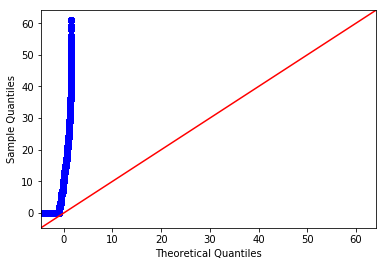
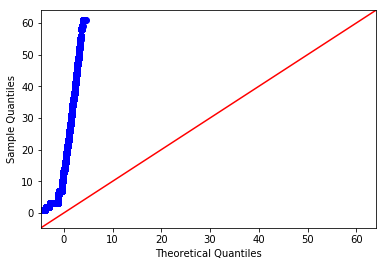
PosTeamScore:





DefTeamScore:

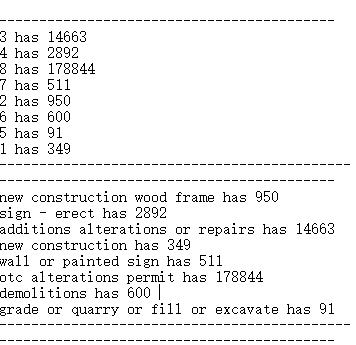




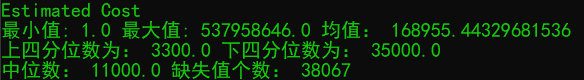
**数据集2 Building\_permit.csv**

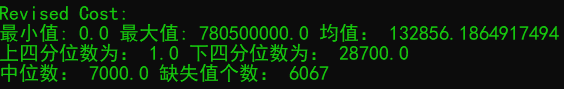
**标称属性分析：**

选取了标称属性包括： Permit Number, Permit Type'. Permit Type Definition , Street Name, Current Status, Number of Existing Stories，Existing Use，Existing Use，Proposed Use，Supervisor District, Neighborhoods - Analysis Boundaries, Zipcode, Location，Record ID。对这些标称属性分析各个元素的频数。结果如下，其余结果存在result\_2.txt文件中。



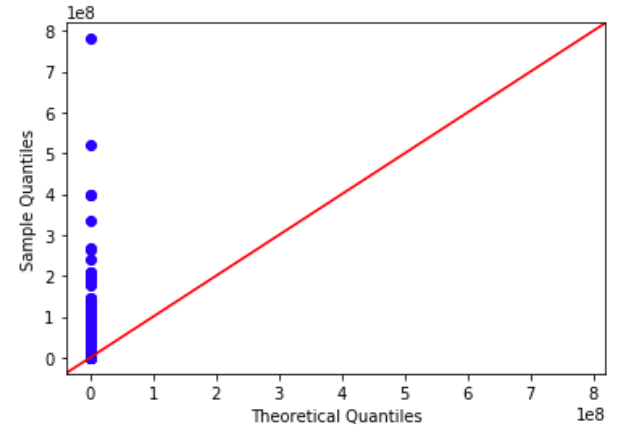
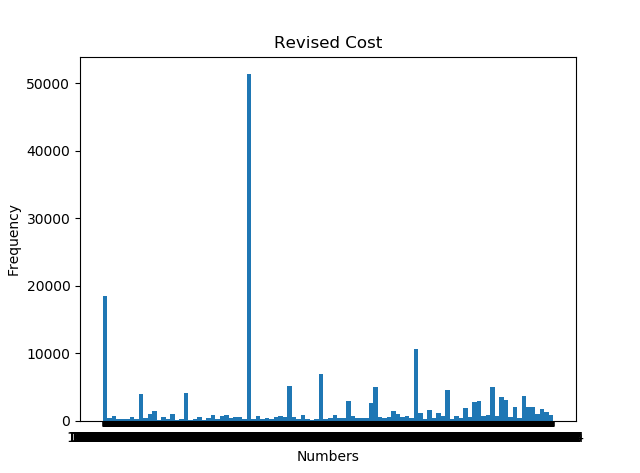
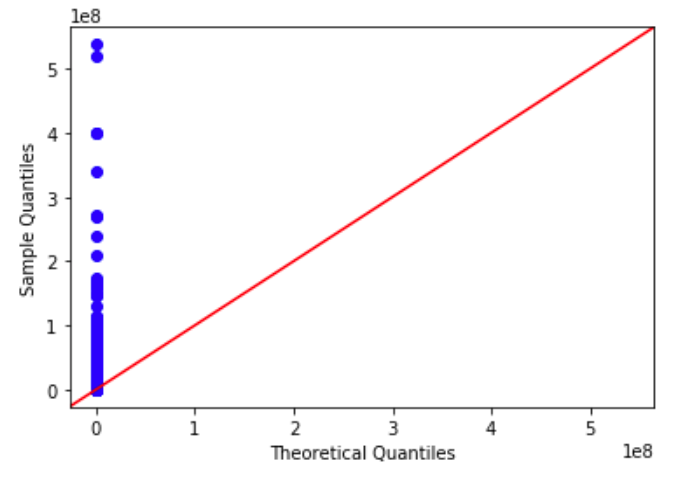
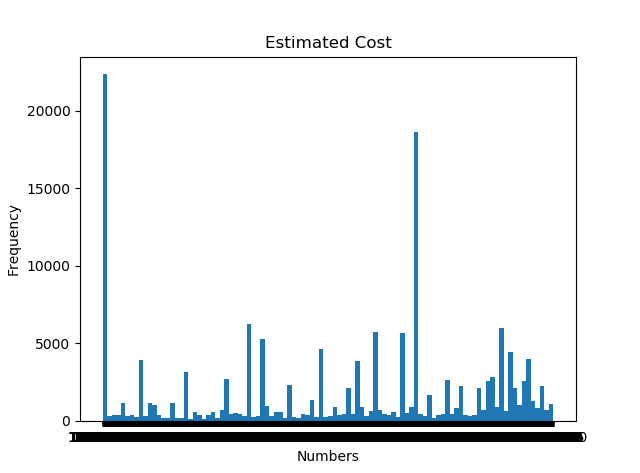
**数值属性分析**

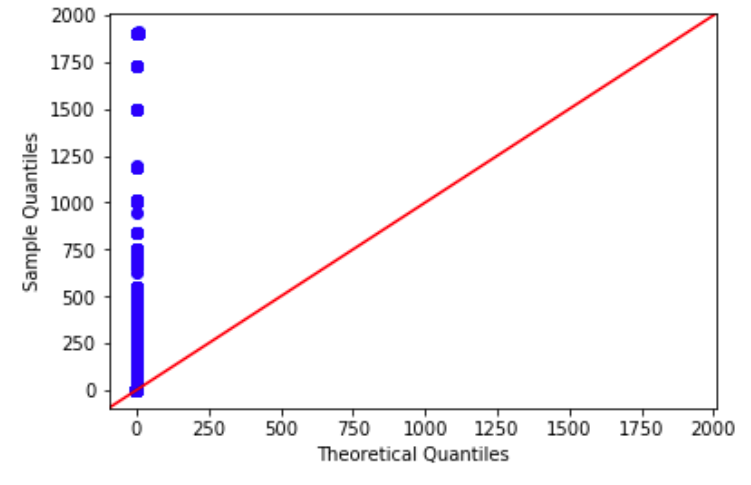
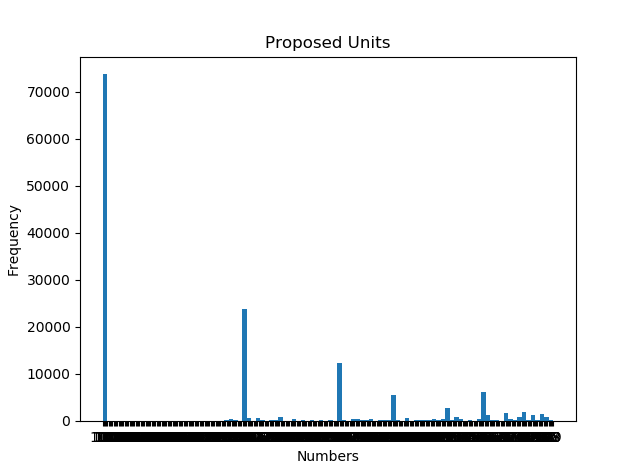




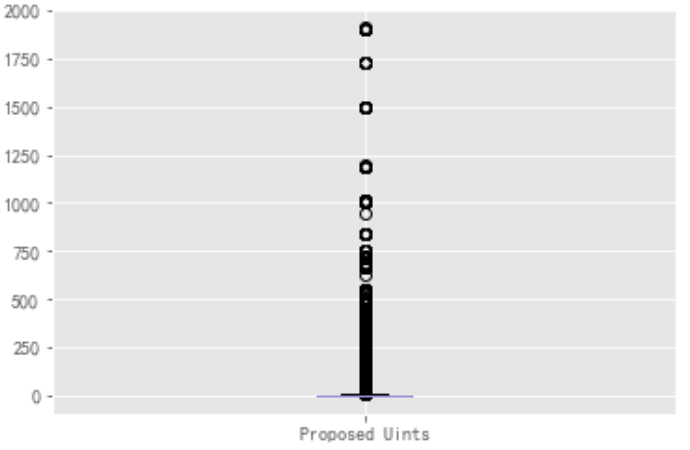
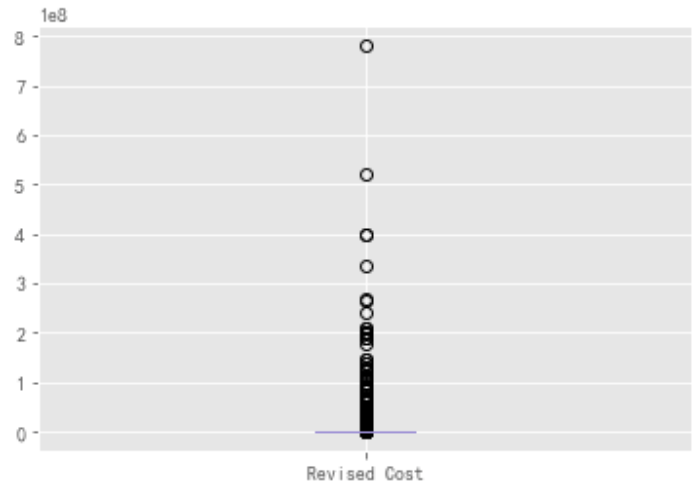
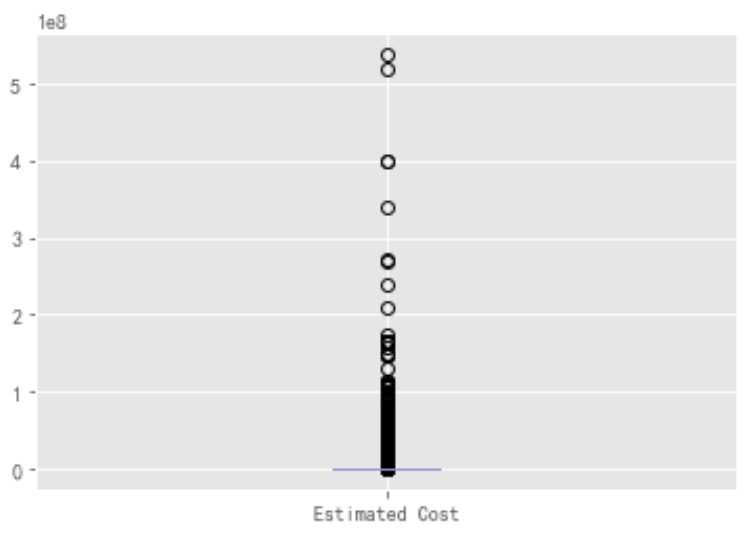


其中为数值属性绘制的直方图和q-q图如下





绘制的盒图如下

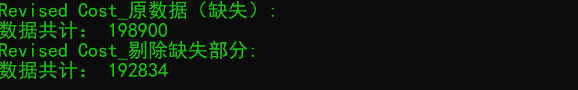


**数据缺失的处理**

观察数据集中缺失数据，分析其缺失的原因。

缺失的数据属性中的：Street Number Suffix, Unit, Unit suffix, Completed Date, Structural Notification, Voluntary Soft-Story, Fire Only Permit, TIDF Compliance。这些数据几乎全部缺失，可以推断这些数据不是每条数据中的重要属性，或者说是这些属性数据与其他属性所包含的信息有关，是作为补充说明的数据，可以不记录与数据集中。

其中，数据属性中的Revised Cost 数据是有一部分缺失值，缺失的原因有可能是在数据复制备份的时候缺失，或者是数据处理和访问错误，或者是人为的操作失误。而我认为这个属性包含的是重要的数据。



Revised Cost属性的缺失值处理的方法为：（1）将缺失部分剔除；（2）用最高频率值来填补缺失值；（3）通过Estimated Cost属性的相关关系来填补缺失值，通过观察可以知道Estimated Cost 和Revised Cost之间的值的大小有相关关系，在大概率下同一条数据的两个属性值相等，所以当Revised Cost缺失时，用同一条数据的Estimated Cost值来填充

