# 作业一 数据探索性分析与数据预处理

**姓名：李晓波**

**学号：2620170054**

## 数据集一： San Francisco Building Permits

### 一、实验环境及语言

**语言及环境依赖**

语言： python

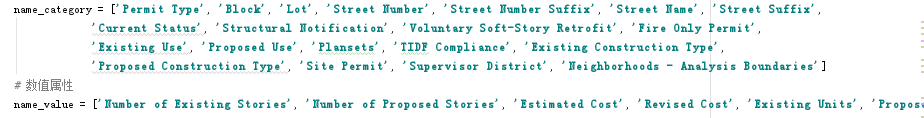
依赖的包：numpy,pandas,matplotlib,statsmodels

### 二、实现方法

首先进行读取数据，利用pandas库中的read\_csv函数读取

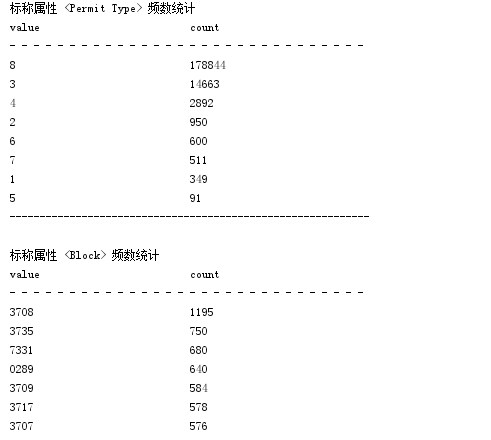
#### 2.1 数据摘要

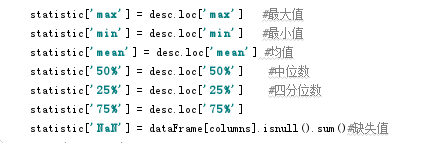
要求标称属性的频数和数值属性的中位数等不同的要求，所以先要把不同属性分开，得知



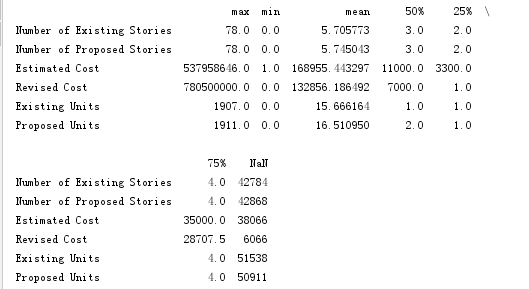
其中name\_category为标称属性，name\_value为数值属性

计算标称属性每个取值的频数，就要利用pandas库中的value\_counts方法进行统计，下图为一部分统计结果



* 而计算数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数也是利用pandas库如下图：
* 

其结果为：



#### 2.2数据的可视化

数据的可视化利用了matplotlib API函数，通常的引入约定为



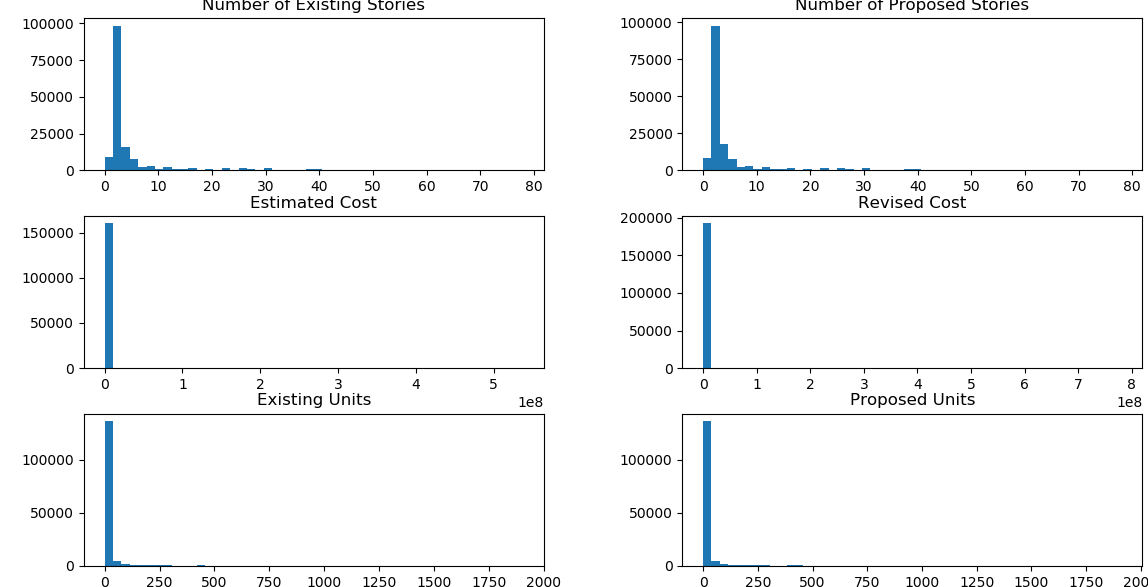
利用figure函数创建窗口，然后用add\_suplob创建图，十分方便

其中直方图的绘制方法：主要用到的是pylab.hist()方法

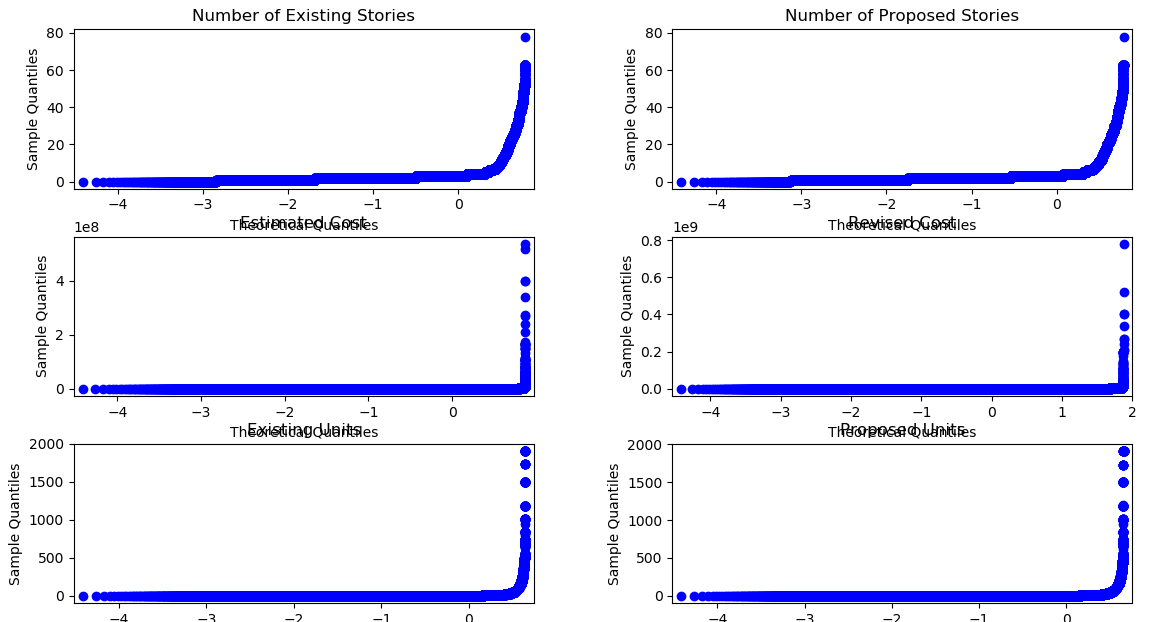
qq图的绘制方法：主要用到的是scipy.stats.probplot方法

盒图的给制方法：主要用到的是boxplot()方法

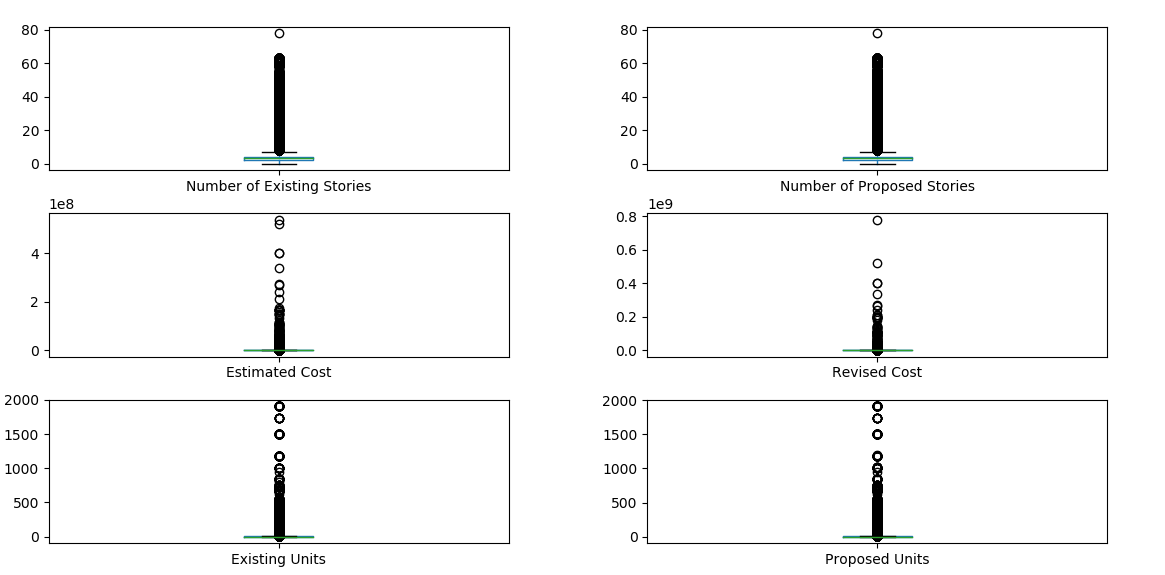
其结果分别为：



直方图



qq图



盒图

由图知基本大部分值都集中在特定的范围，近似满足正太分布

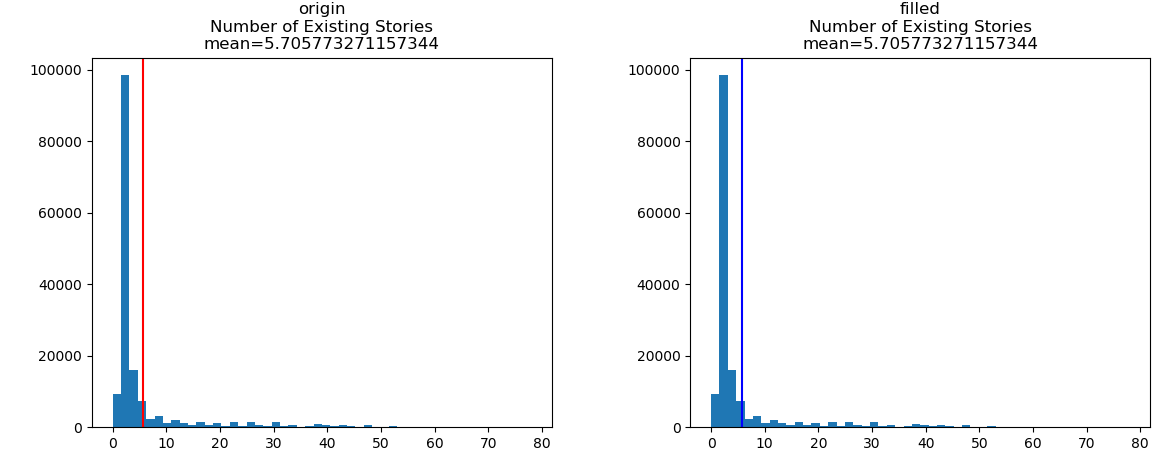
#### 2.3数据缺失的处理

##### 2.3.1剔除部分缺失值

可以直接调用



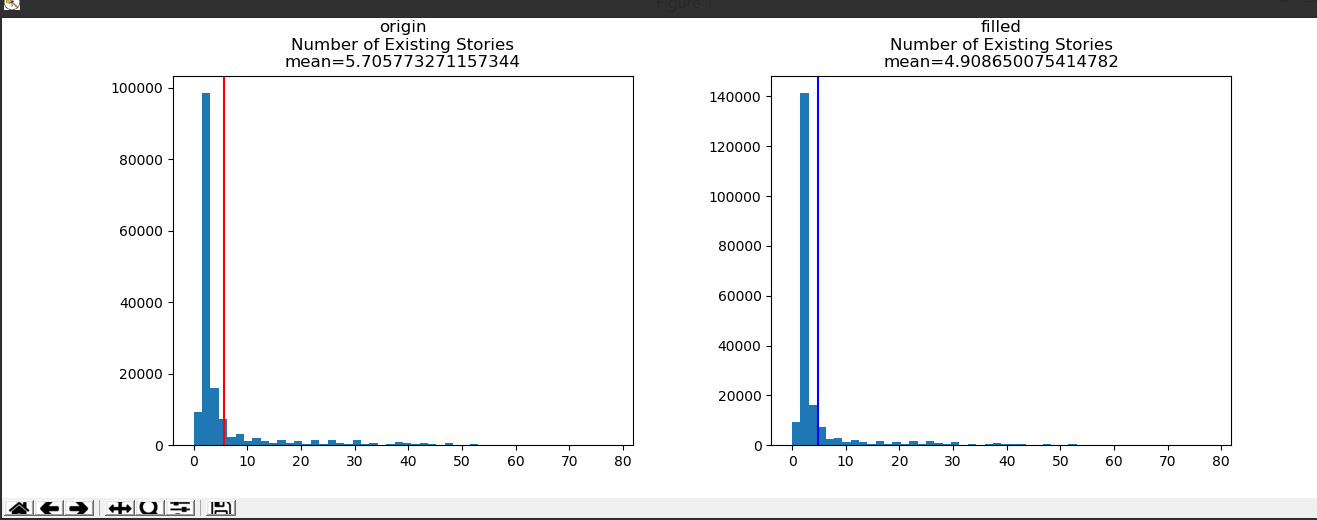
将全为空的行删除，然后用直方图对比，结果为：



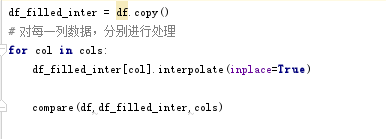
这是其中一个数值属性列的对比，由于全为缺失值的行较少，基本没发生变化，其余都差不多不一一列举，可知这种方法一般很少用

##### 2.3.2用最高频率值来填补缺失值

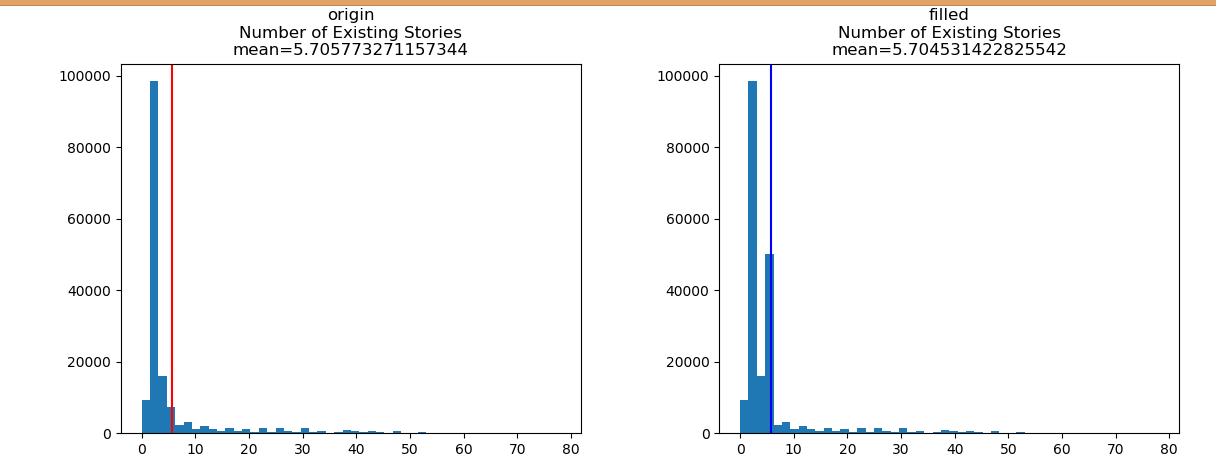
填充缺失值可以直接用fillna()函数直接填充，如图其中结果用直方图表示为



##### 2.3.3通过属性的相关关系来填补缺失值



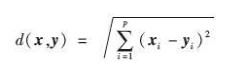
利用interplolate函数对每列进行处理，然后通过直方图比较，其结果为：



emmmm

##### 2.3.4通过数据对象之间的相似性来填补缺失值

探索数据集列（变量）之间的相关性，利用欧式距离，公式为：



可以通过添加包中的函数 knnImputation()来实现

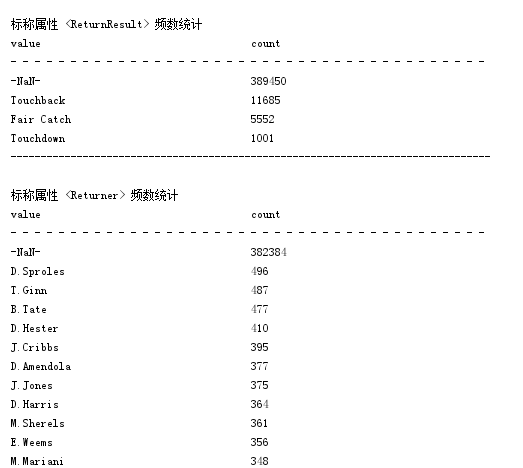
其结果为：

## 数据集二：NFL Play-by-Play 2009-2017

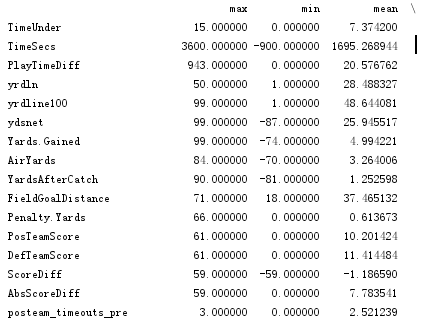
方法与数据集一相同，直接展现结果

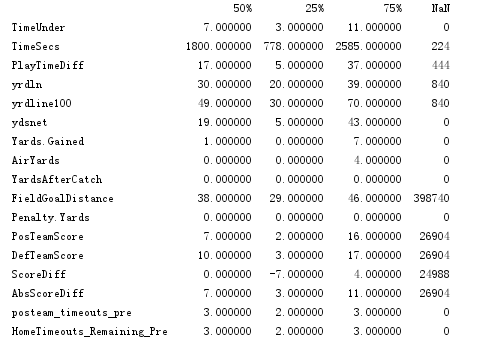
### 数据摘要

#### 对标称属性，给出每个可能取值的频数，（一部分）



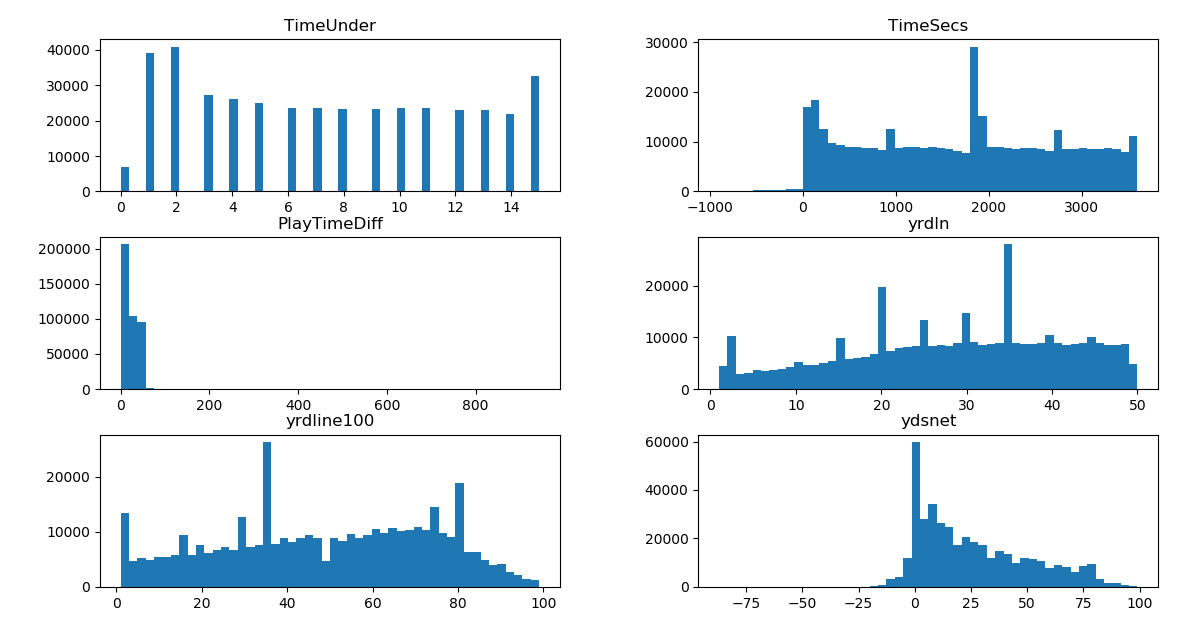
#### **数值属性，给出最大、最小、均值、中位数、四分位数**

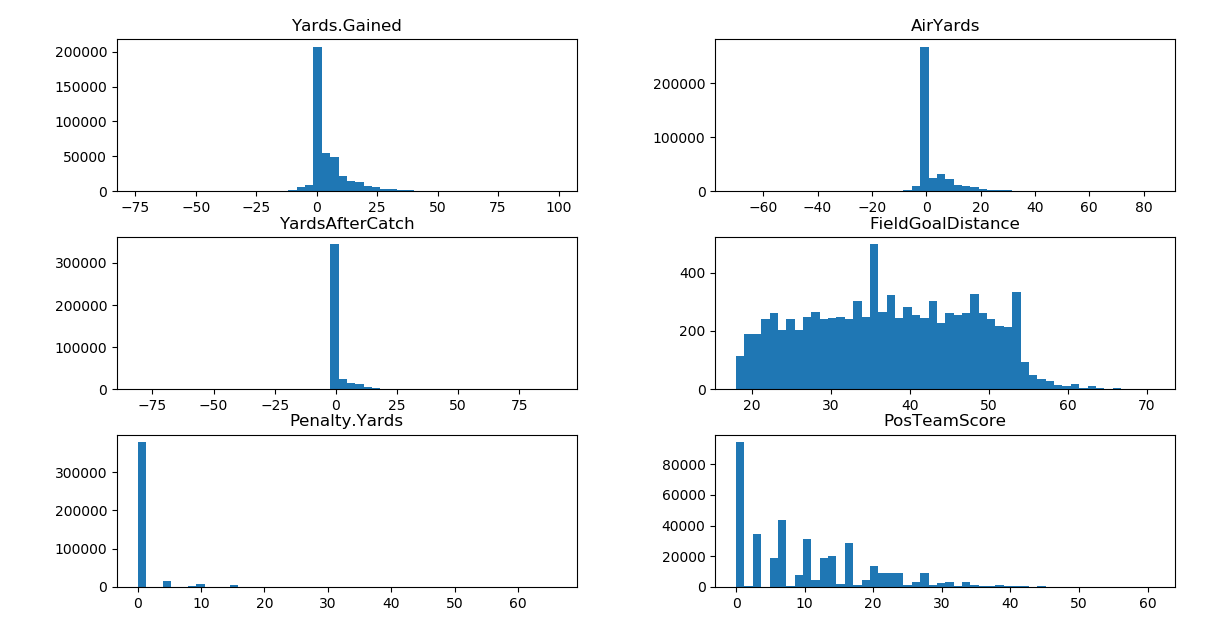


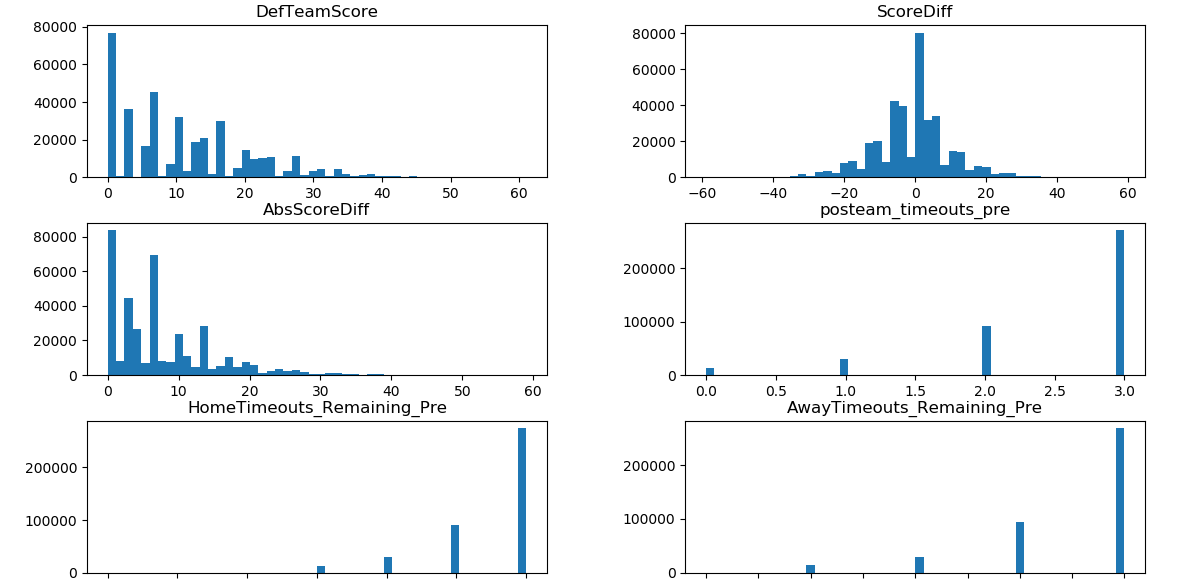


### 数据的可视化

#### 直方图

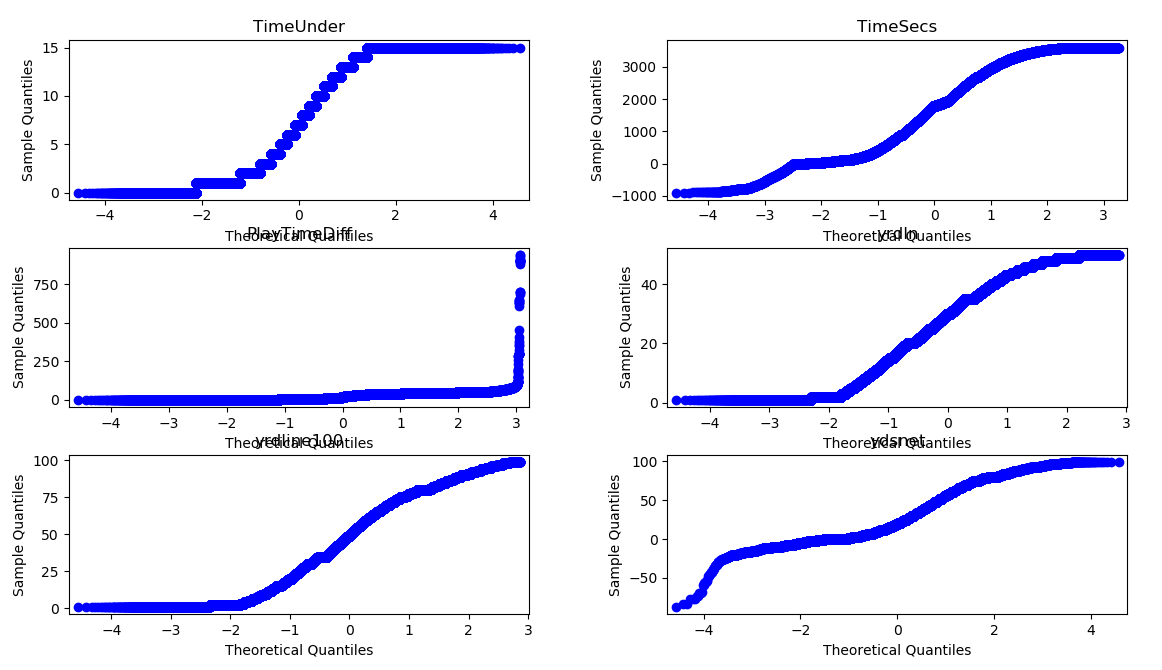


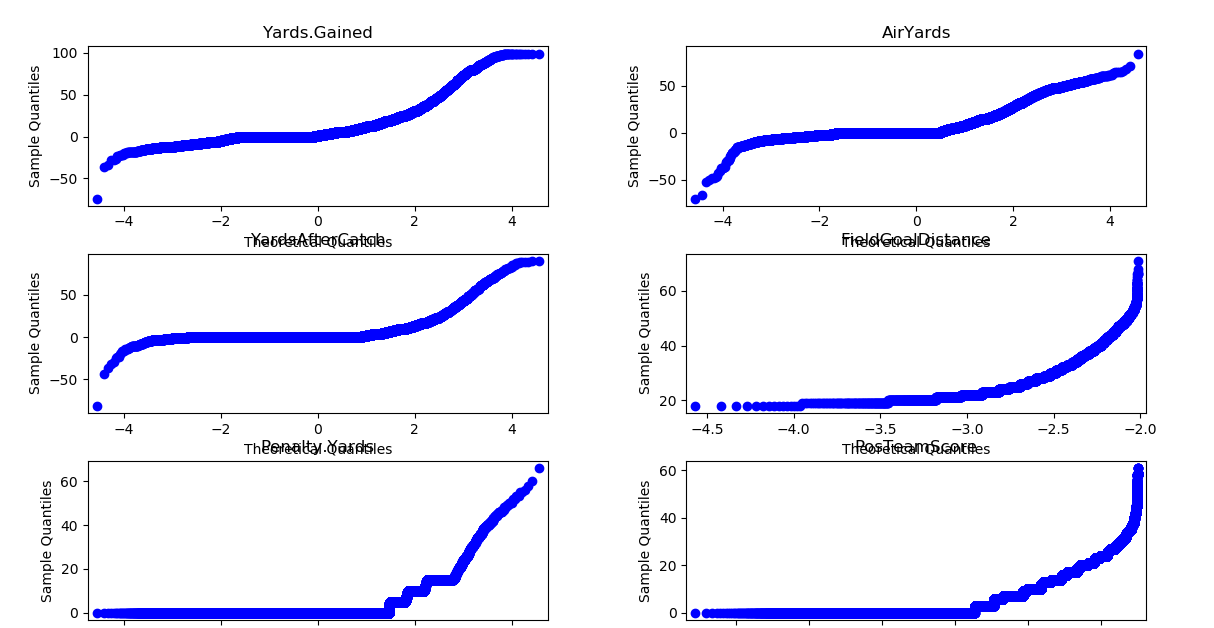




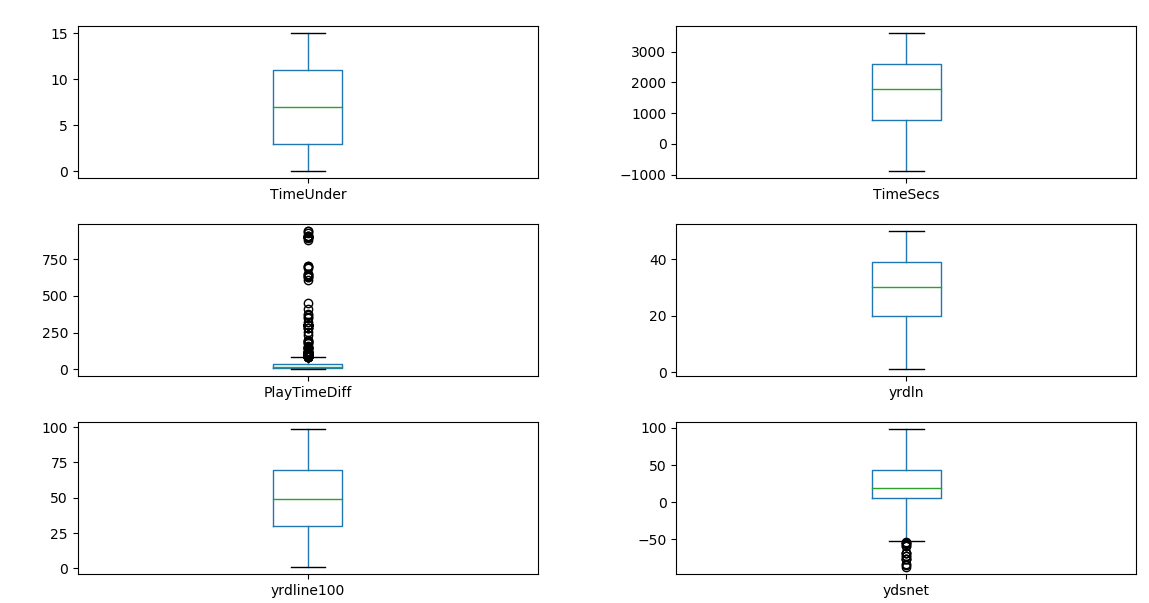
太多就不一一列举，可看出大部分标称属性都集中在个别上，成正态分布

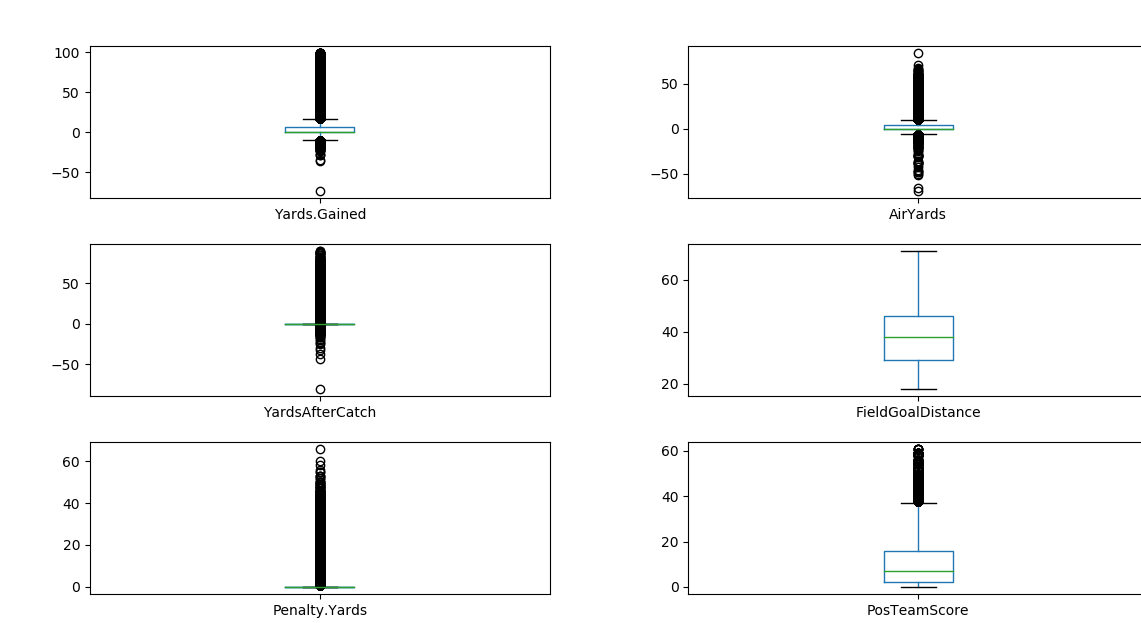
#### qq图





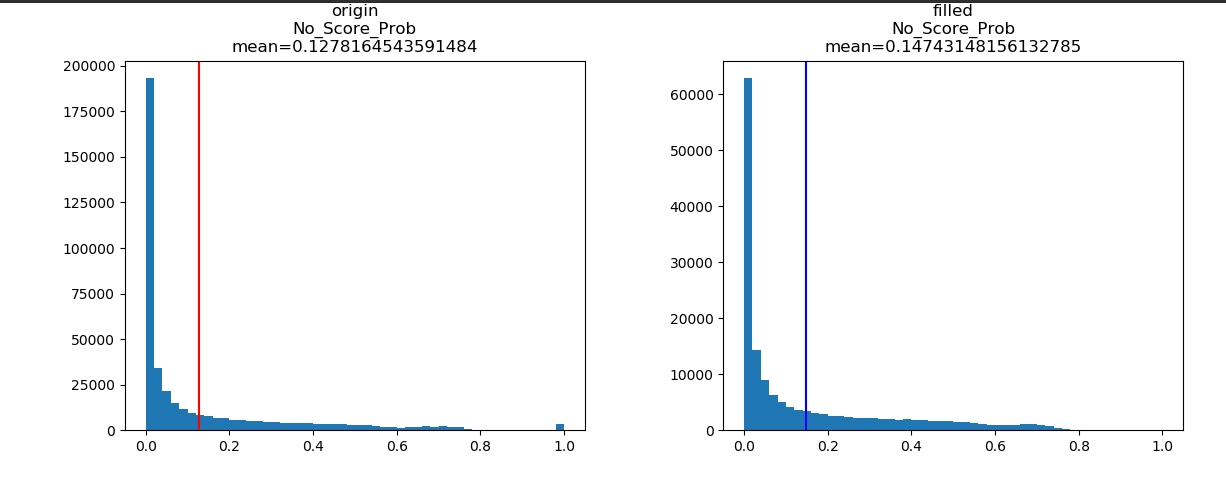
#### 盒图

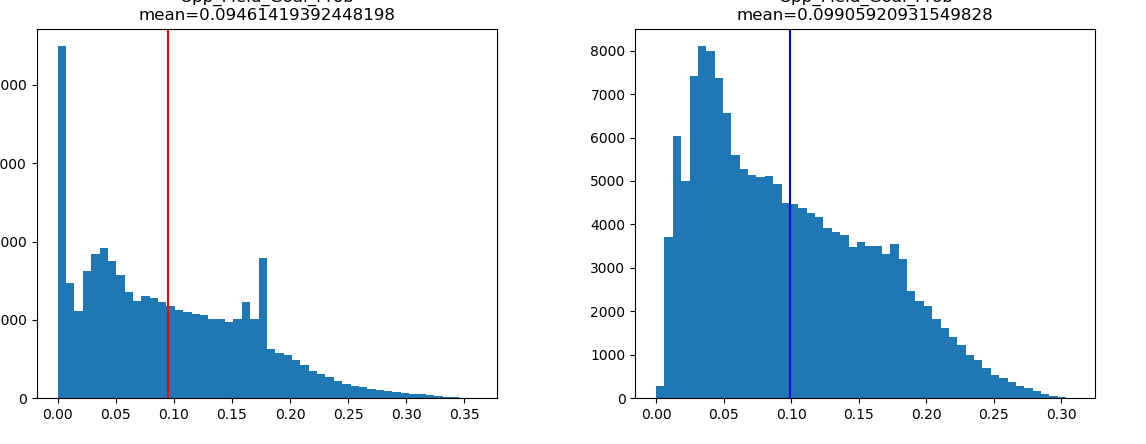




### 数据缺失值的处理

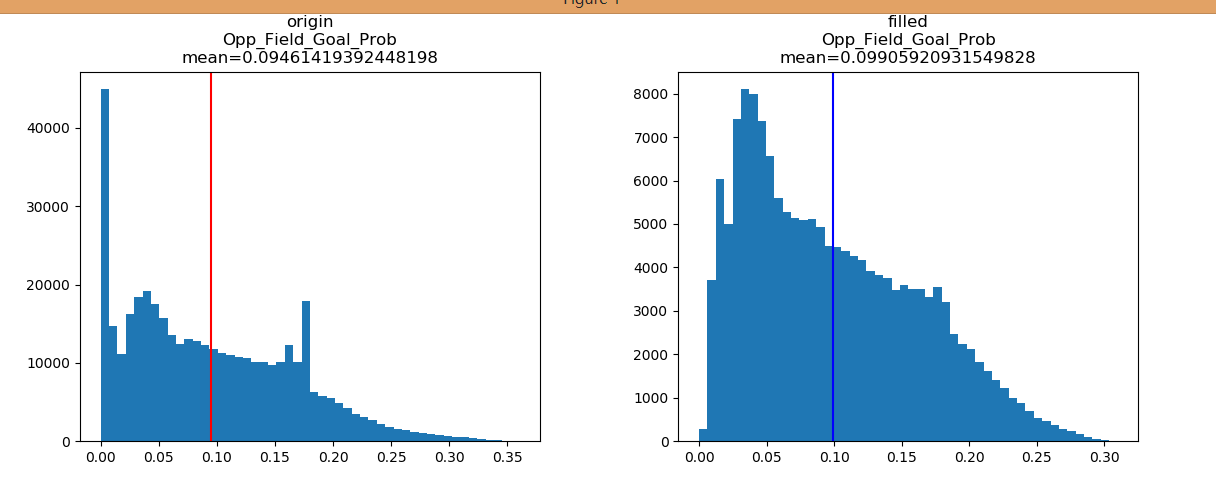
#### 剔除部分缺失值

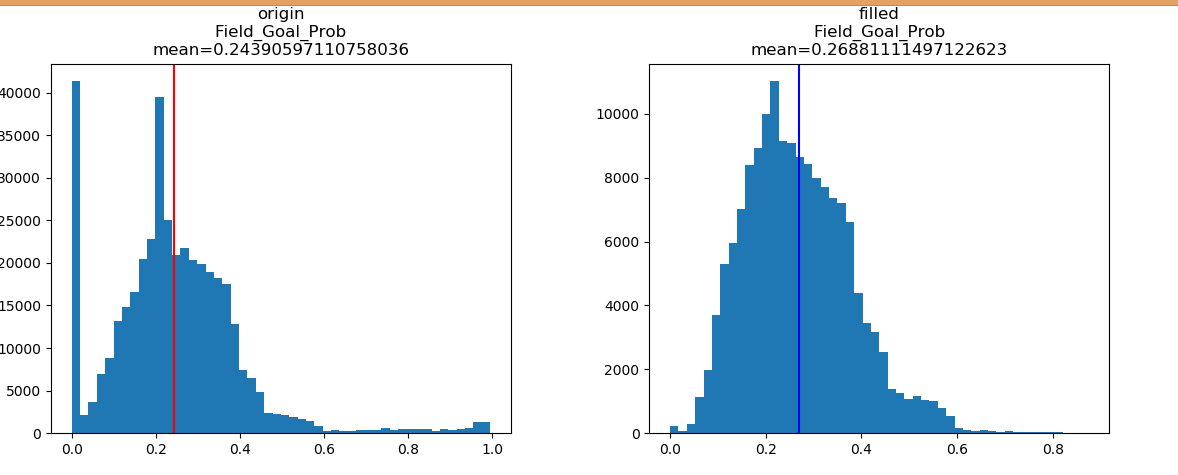




这个数据比较大相对于数据集一，可以看出比数据1变化变大，而且使其更接近正态分布

#### 用最高频率值来填补缺失值





#### 通过属性相关关系填补缺失值

