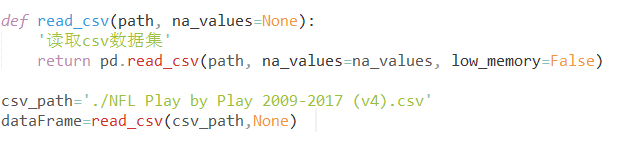
**NFL Play by Play 2009-2017 (v4)数据分析与预处理**

1 数据读取和属性分类

利用Python中的pandas库进行csv数据文件的读取：



对属性进行分类，分为标称属性和数值属性：



2 数据可视化和摘要

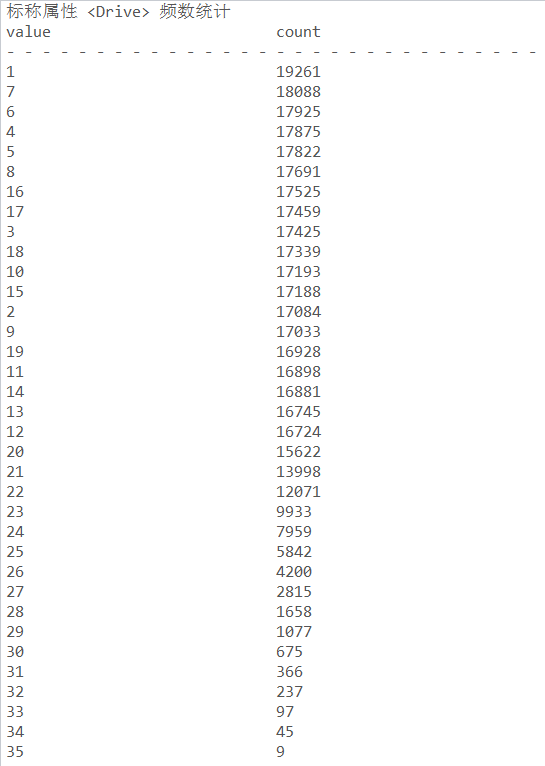
2.1 数据摘要

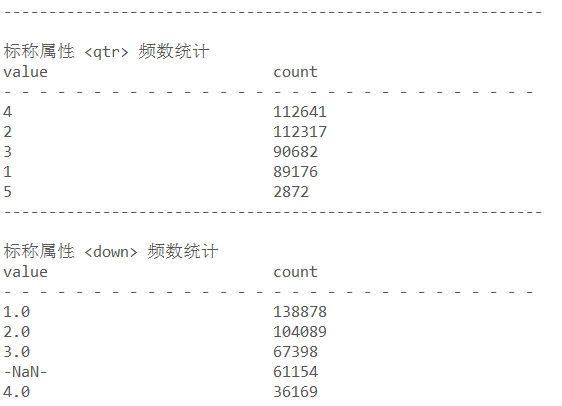
**标称数据：**

对于标称属性，使用pandas中的value\_counts函数统计每个标称属性的所有可能取值的频数。



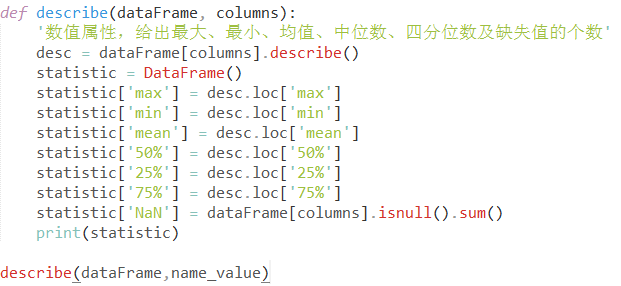
部分结果如下：



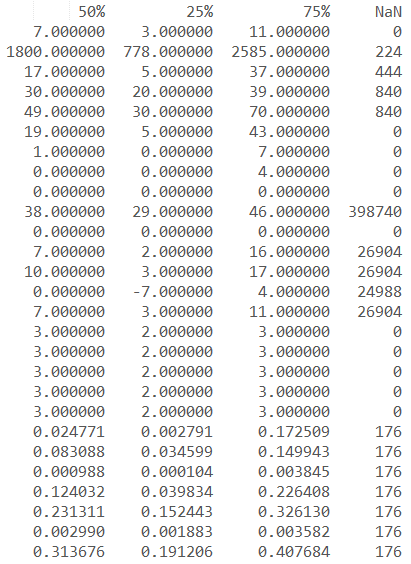
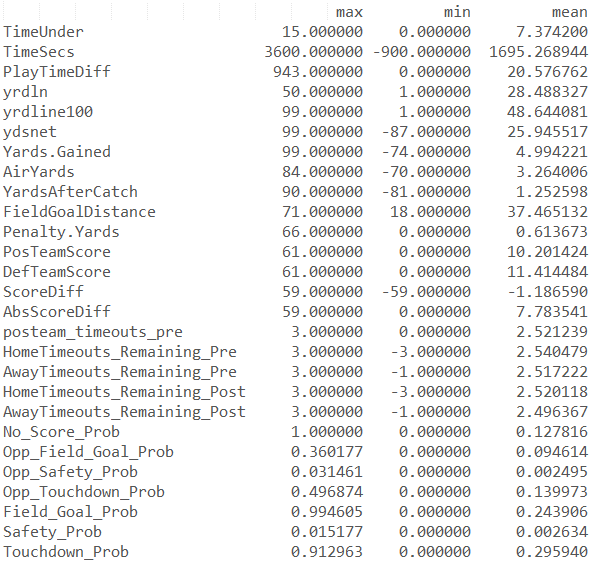


**数值属性：**

对于数值属性，使用pandas中describe()函数给出其最小、最大、均值、中位数、四分位数及缺失值个数：



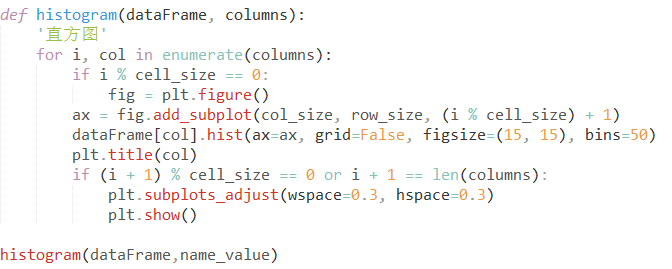
部分结果如下：



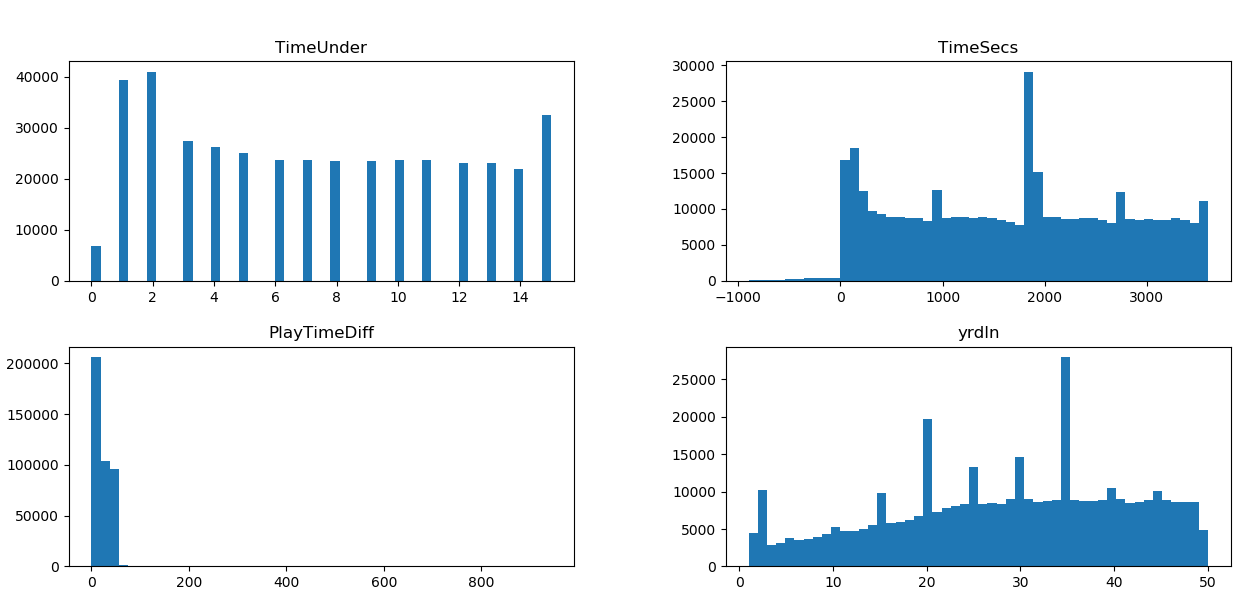
2.2 数据可视化

**直方图**

使用matplotlib绘制直方图

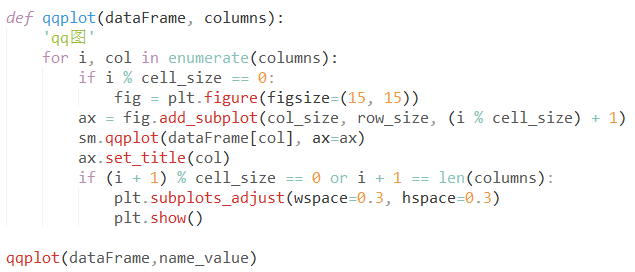


部分结果如下：

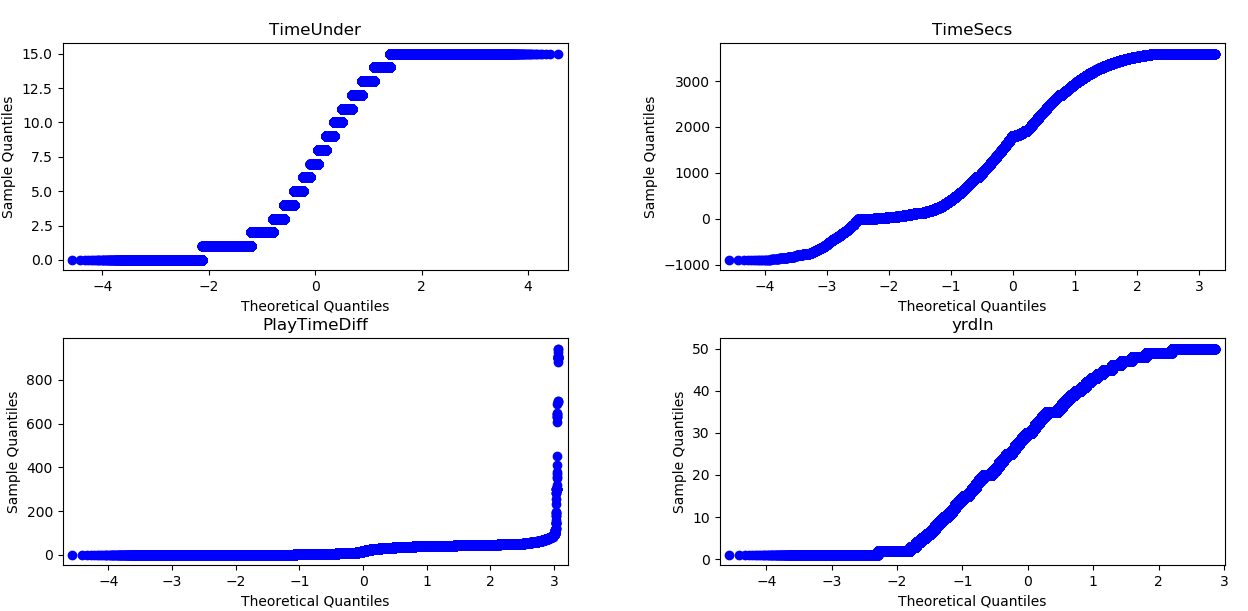


**qq图：**

使用matplotlib绘制qq图



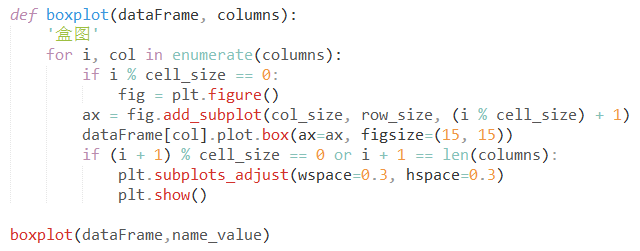
部分结果如下：



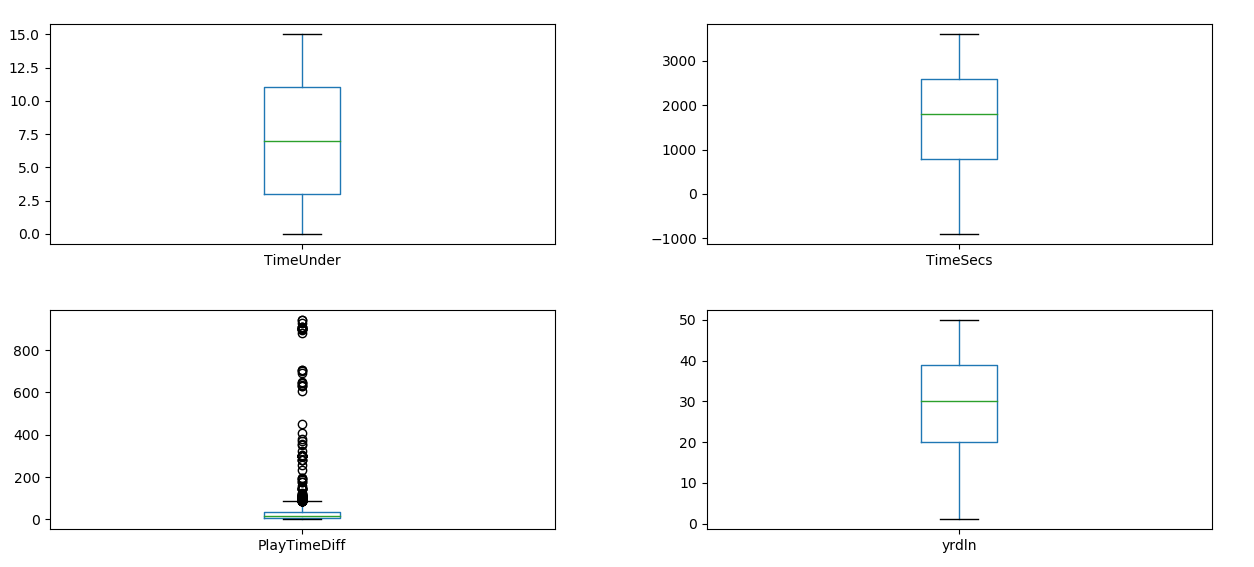
根据qq图可知图像1、2和4是近似直线的，其对应属性（TimeUnder、TimeSecs、yrdln）为正态分布态。

**盒图：**

使用matplotlib绘制盒图，对离群值进行识别：



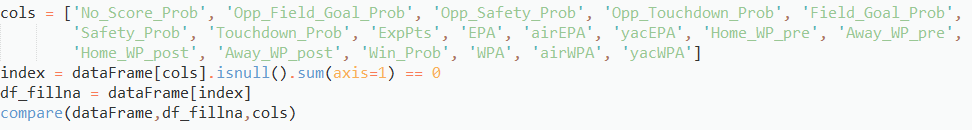
部分结果如下：



3 数据缺失处理

3.1 将缺失部分剔除

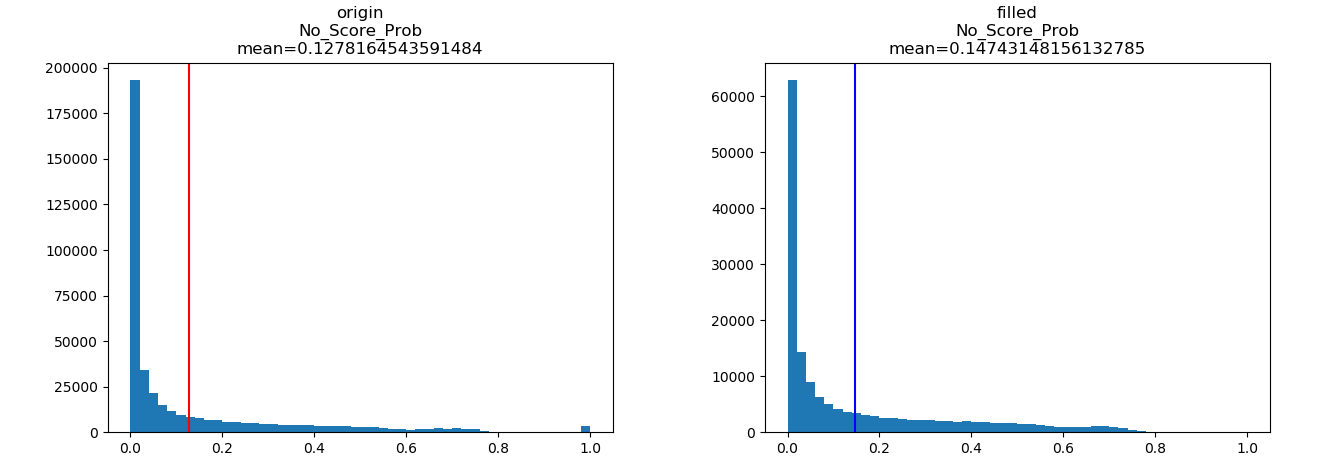
根据分析，可填充的数值属性字段有：No\_Score\_Prob, Opp\_Field\_Goal\_Prob, Opp\_Safety\_Prob, Opp\_Touchdown\_Prob, Field\_Goal\_Prob, 'Safety\_Prob', Touchdown\_Prob, ExpPts, EPA, airEPA, yacEPA, Home\_WP\_pre, Away\_WP\_pre, Home\_WP\_post, Away\_WP\_post, Win\_Prob, WPA, airWPA, yacWPA。对缺失部分进行剔除



通过直方图比较新旧数据集的数值属性：



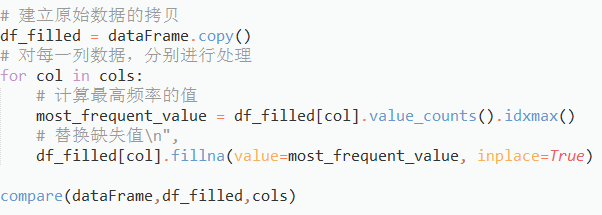
部分结果如下：



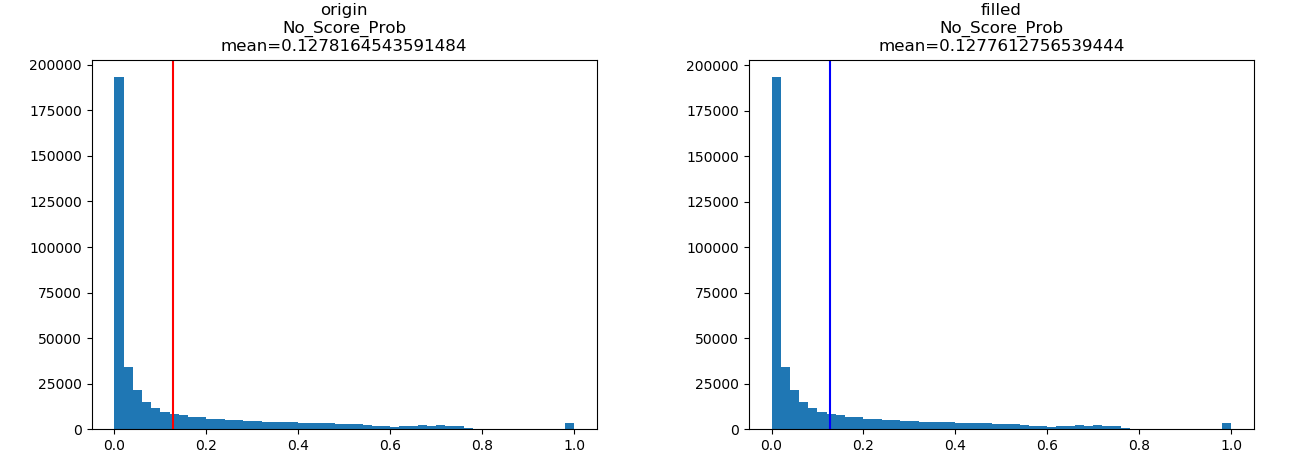
在直方图中，左边的红色垂线表示旧数据集的均值，右边的蓝色垂线表示剔除有缺失的数据得到的新数据集的均值。

3.2 用最高频率值来填补缺失值

找到每个属性中出现次数最多的值，用这个值填充这个属性中所有的缺失值：



在直方图中，左边的红色垂线表示旧数据集的均值，右边的蓝色垂线表示剔除有缺失的数据得到的新数据集的均值。

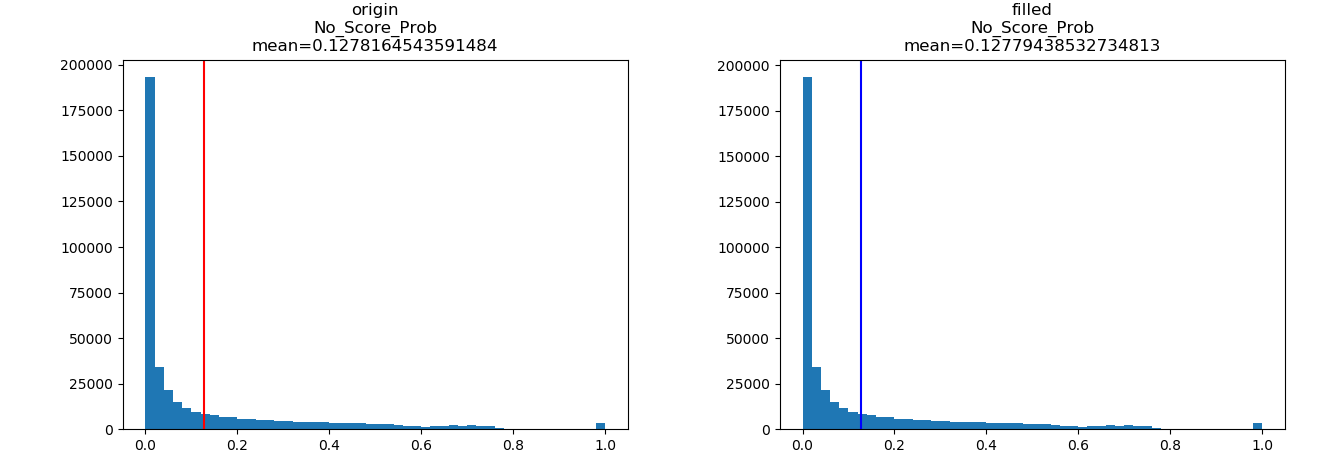


3.3 通过属性的相关关系来填补缺失值

使用pandas中的interpolate()函数，对于每个数值属性进行插值计算，利用得到的插值填充缺失值：



在直方图中，左边的红色垂线表示旧数据集的均值，右边的蓝色垂线表示剔除有缺失的数据得到的新数据集的均值。



3.4 通过数据对象之间的相似性来填补缺失值

无