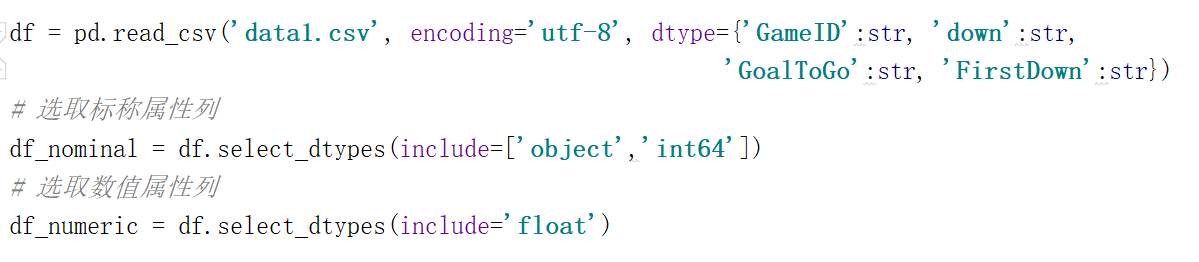
**NFL Play by Play 2009-2017 (v4)数据分析与预处理**

王鹏 2120171069

一、数据读取和属性分类

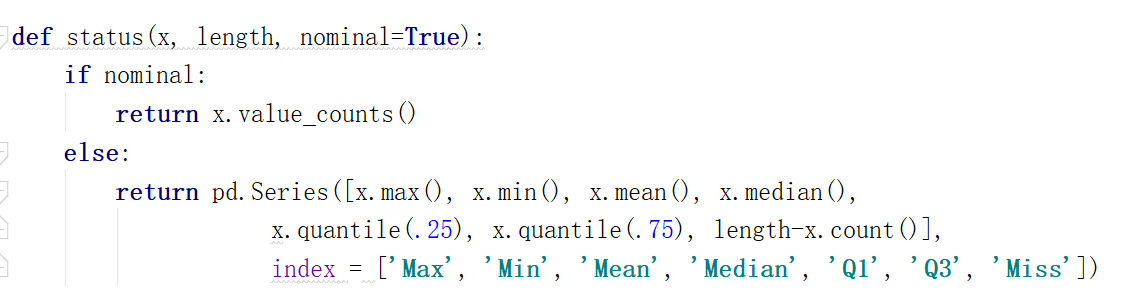
利用Python中的pandas库进行csv数据文件的读取，并选取标称属性和数值属性：



二、数据可视化和摘要

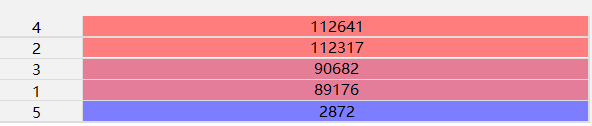
2.1 数据摘要

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。对于数值属性，调用pandas的统计方法进行给出其最小、最大、均值、中位数、四分位数及缺失值个数，具体实现函数如下：

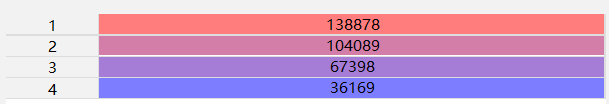


**标称属性部分结果如下：**

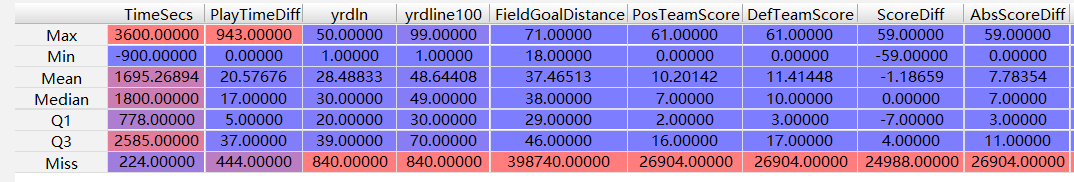
①属性“qtr”的频数统计



②属性“down”的频数统计



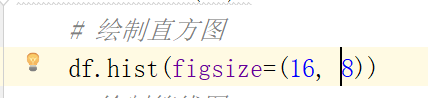
**数值属性部分结果如下：**



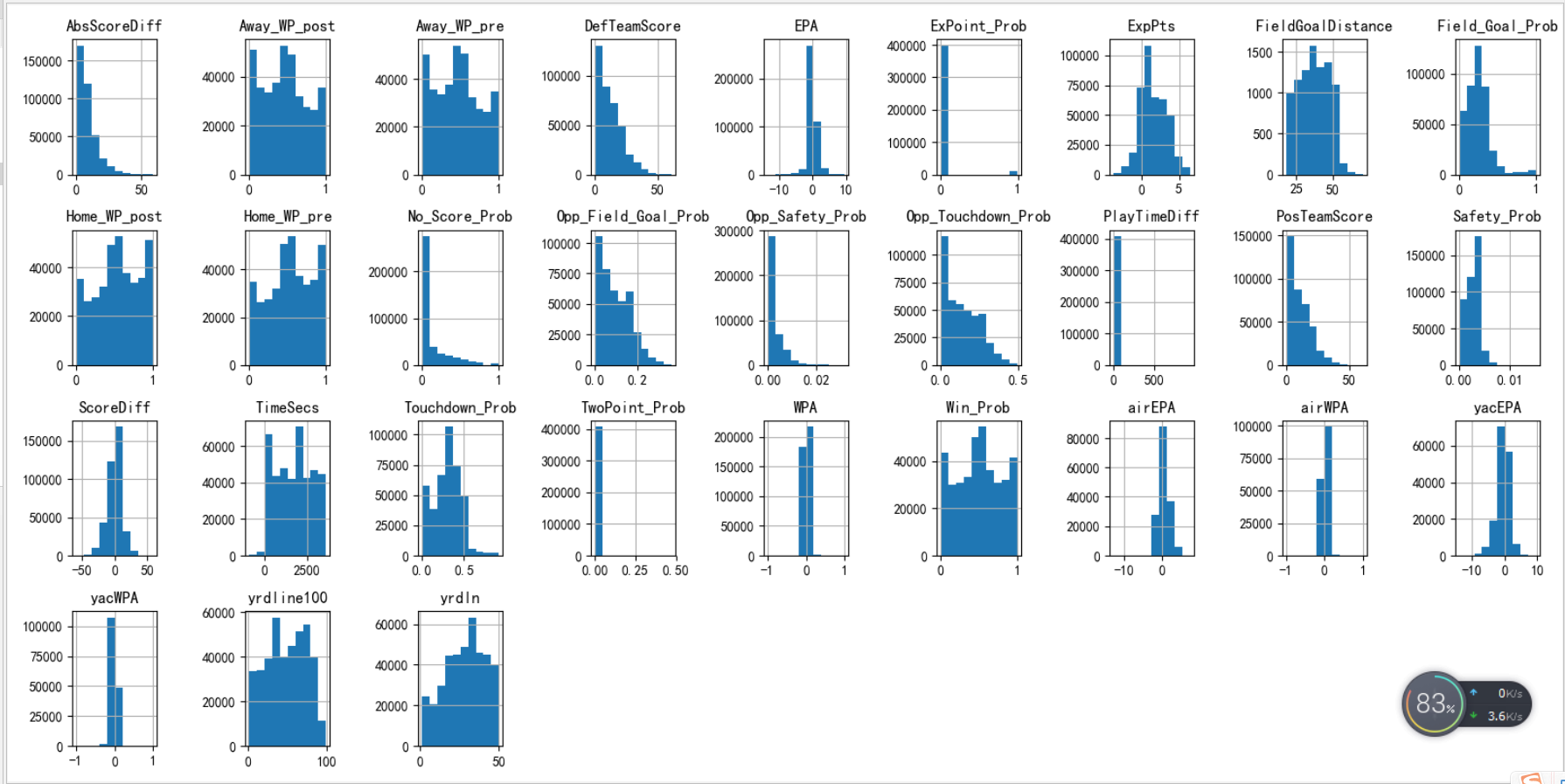
2.2 数据可视化

**直方图**

直接使用pandas自带函数绘制直方图

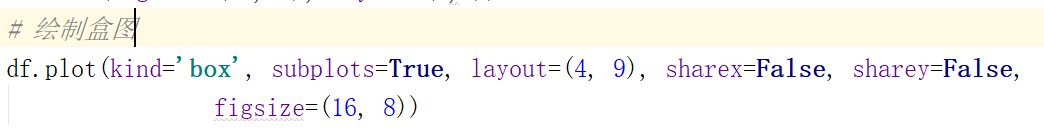


结果如下：

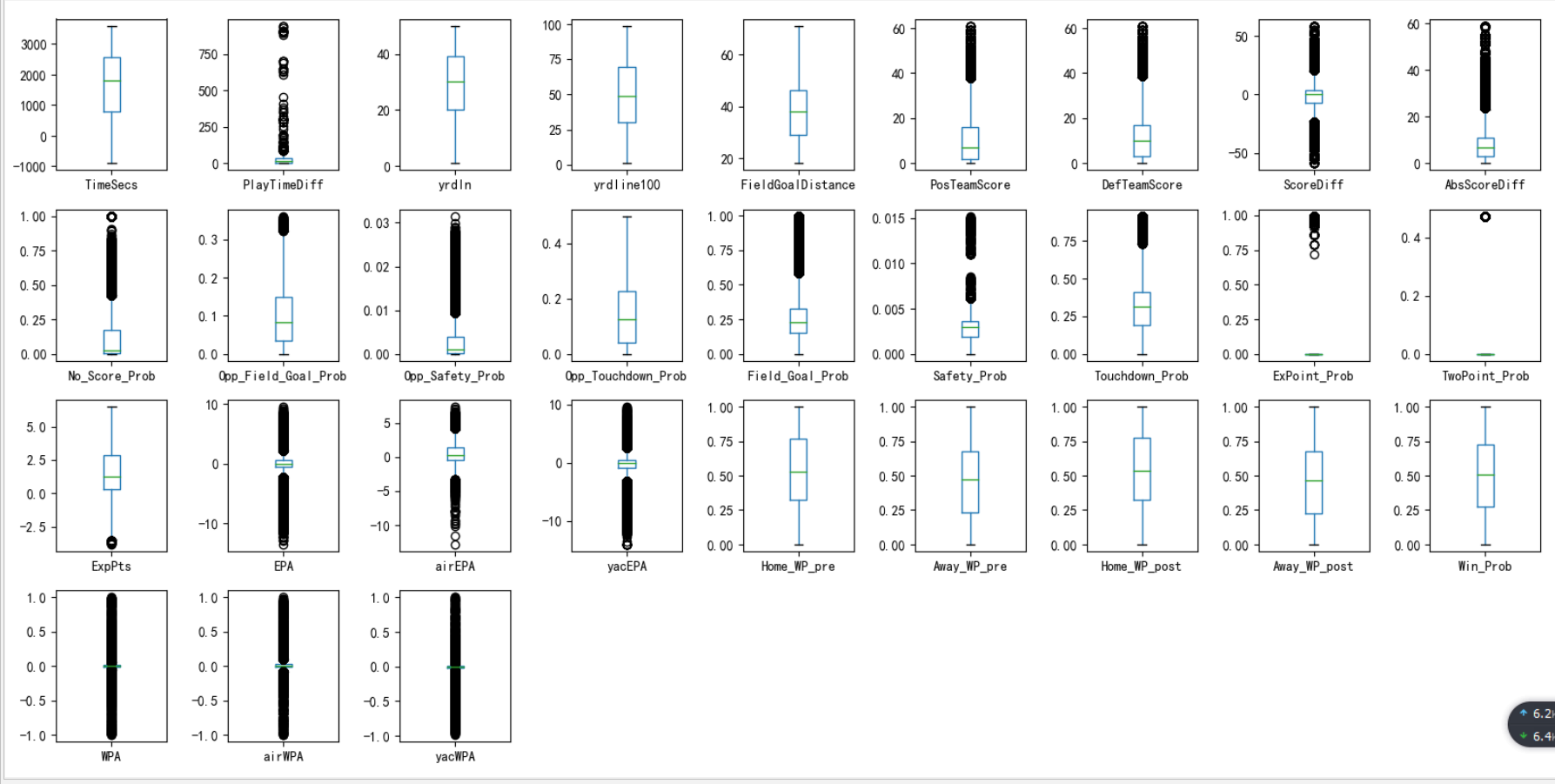


**盒图：**

直接使用pandas自带函数绘制盒图

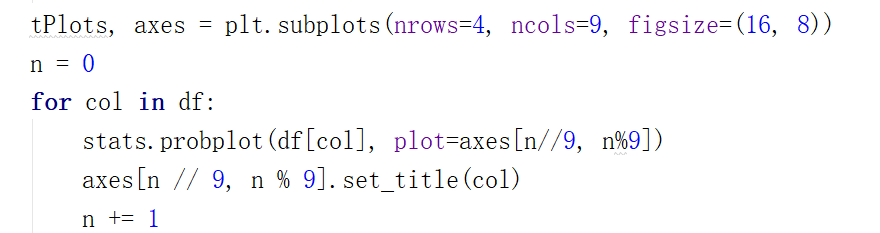


结果如下：

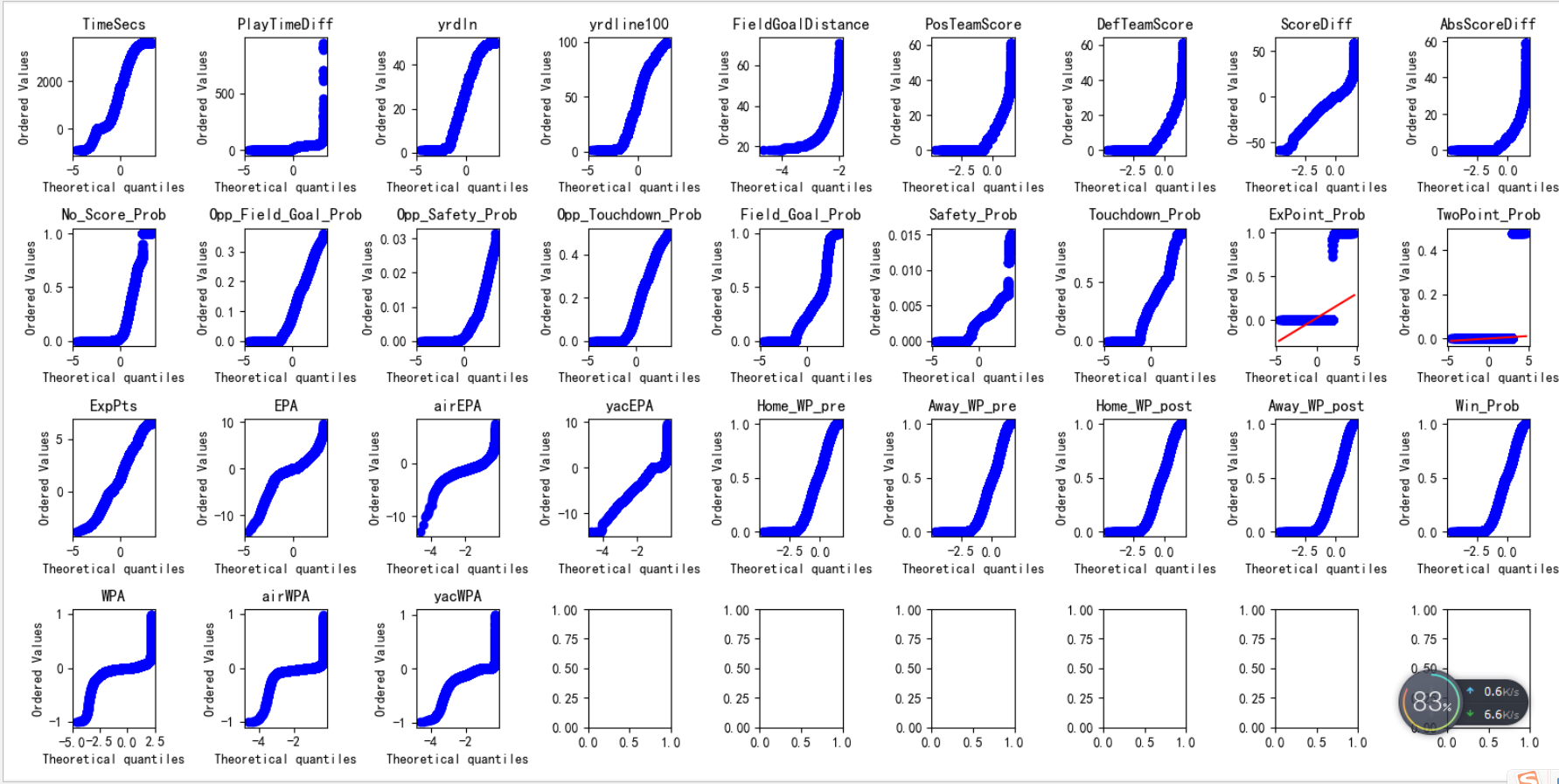


**Q-Q图：**

使用scipy库进行绘图，如下所示：



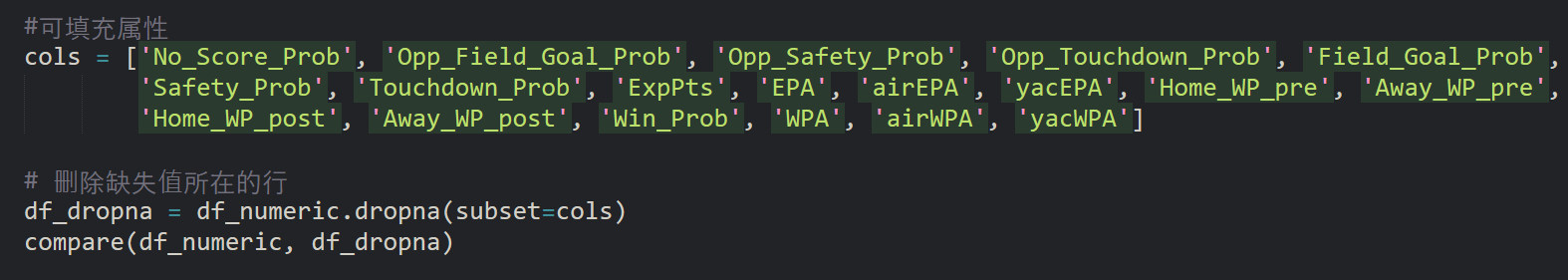
部分结果如下：



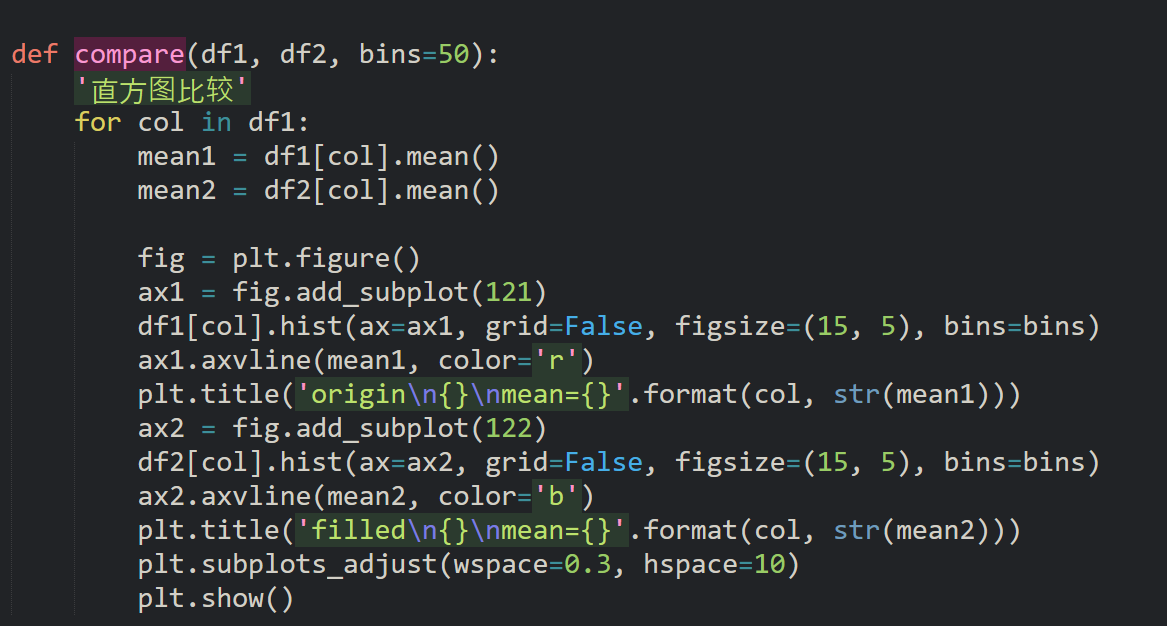
三、数据缺失处理

3.1 将缺失部分剔除

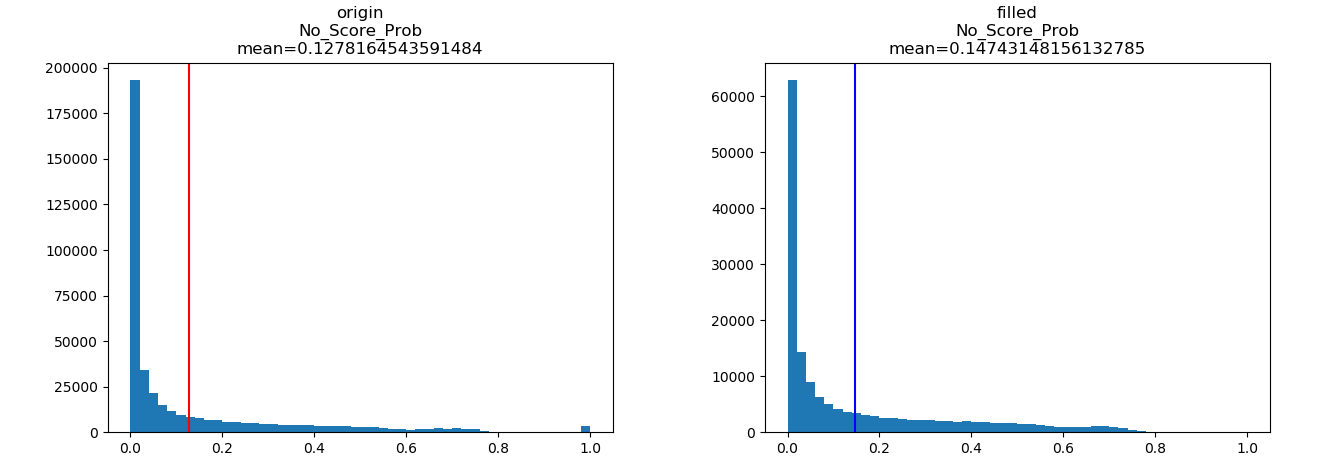
根据分析，可填充的数值属性字段有：No\_Score\_Prob, Opp\_Field\_Goal\_Prob, Opp\_Safety\_Prob, Opp\_Touchdown\_Prob, Field\_Goal\_Prob, 'Safety\_Prob', Touchdown\_Prob, ExpPts, EPA, airEPA, yacEPA, Home\_WP\_pre, Away\_WP\_pre, Home\_WP\_post, Away\_WP\_post, Win\_Prob, WPA, airWPA, yacWPA。对缺失部分进行剔除



通过直方图比较新旧数据集的数值属性：



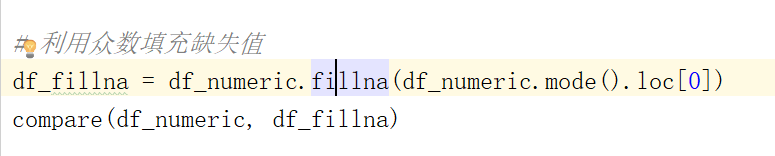
部分结果如下：



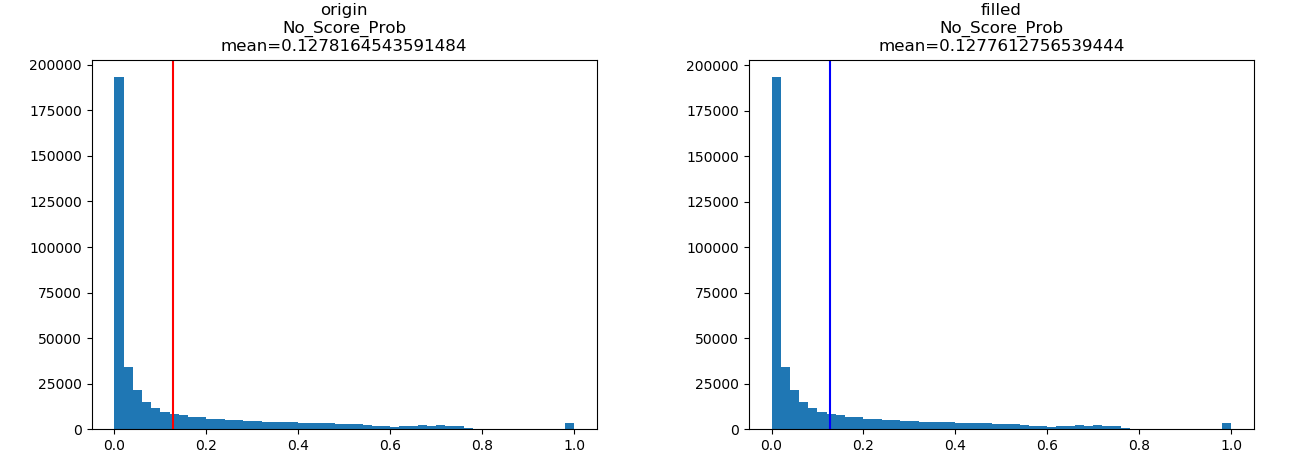
在直方图中，左边的红色垂线表示旧数据集的均值，右边的蓝色垂线表示剔除有缺失的数据得到的新数据集的均值。

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：

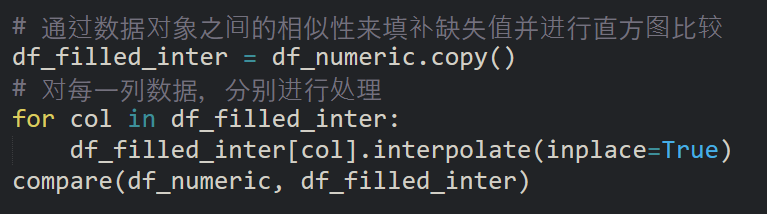


部分结果如下：

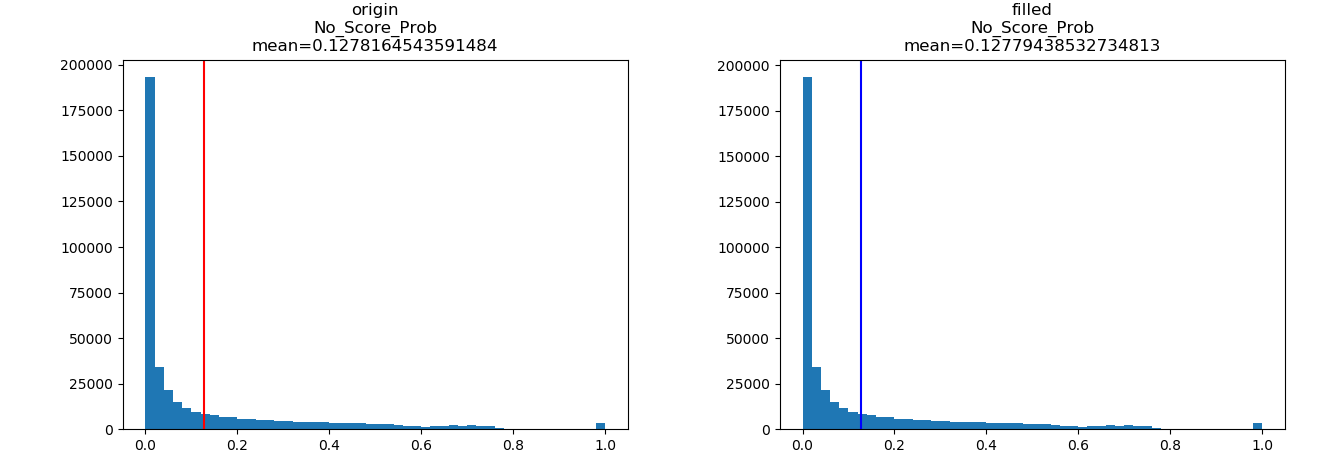


3.3 通过数据对象之间的相似性来填补缺失值

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：

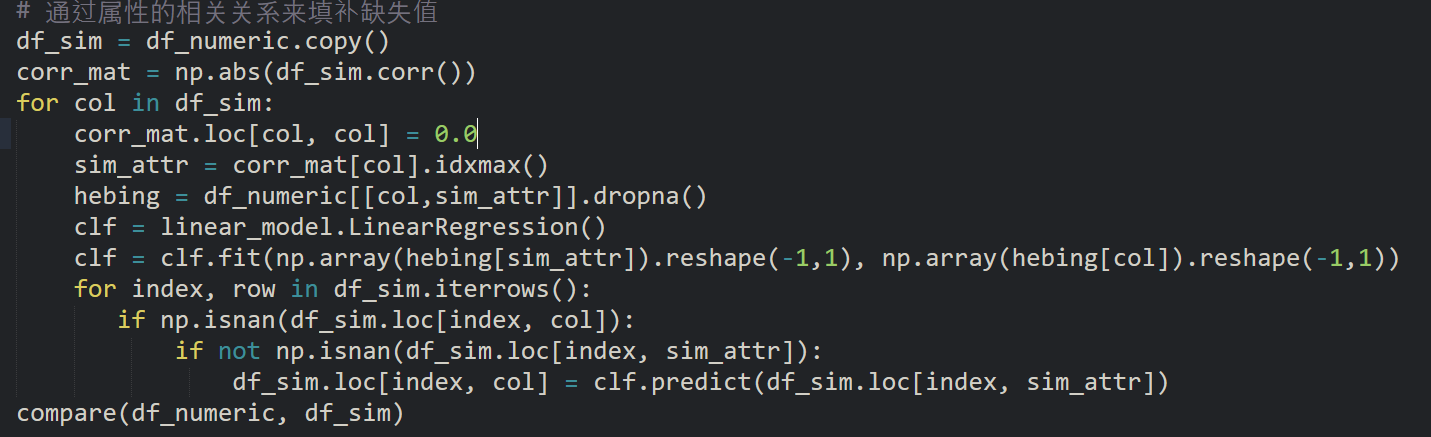


部分结果如下：

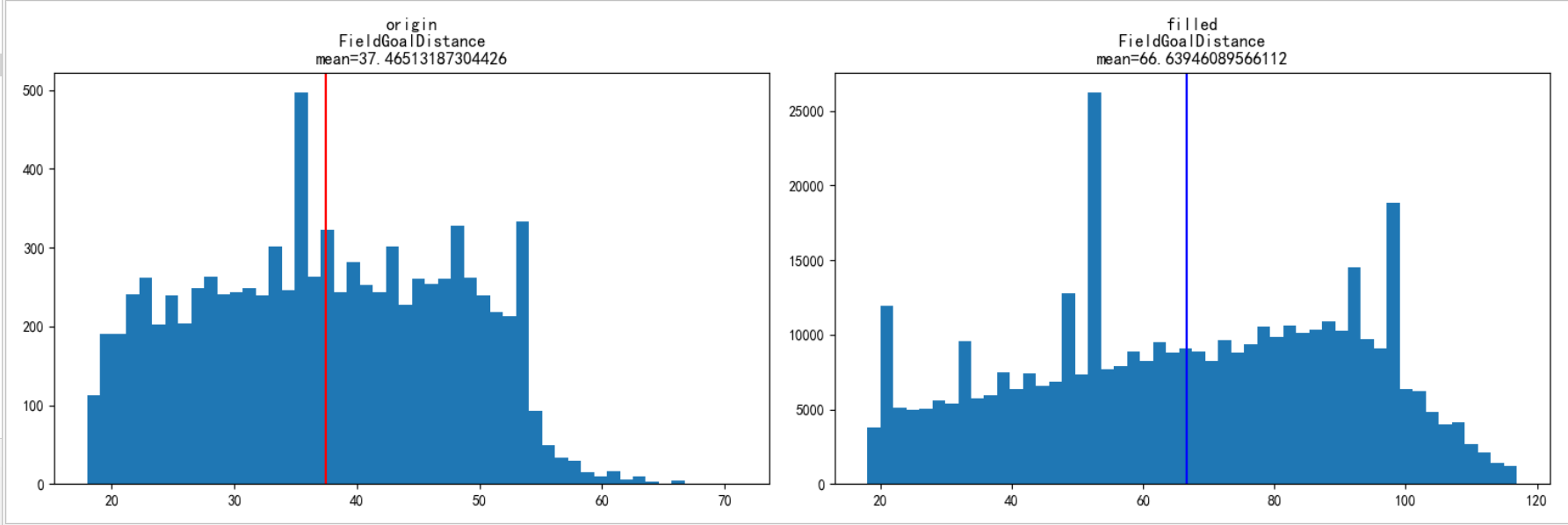


3.4 通过属性的相关关系来填补缺失值

利用pandas的corr()方法计算相似性矩阵，然后得到每个属性与之相关系数最大的属性，然后利用sklearn库中的线性回归方法得到两个属性间的线性函数关系，在基于此线性函数关系预测属性的缺失值，具体如下所示：



选取的一个结果如下：



在结果中，只有该属性出现了较大变化，其它属性基本上不变，通过检查原数据，发现这个这个属性的缺失值较多，所以经过相关系数填充缺失值后，新数据与原始数据显著不同。