数据探索性分析与数据预处理

针对给定的两个数据集，NFL Play-by-Play 2009-2017和San Francisco Building Permits，初步观察发现第一个数据集的属性个数明显比第二个数据集大。对于数据属性，一般分为四个类：标称属性，二元属性，序数属性，数值属性。要求对数据集中的标称属性和数值属性做预处理，因此首先需要对所有属性进行人工分类。

结合属性的实际含义和实例，两个数据集中的标称属性和数值属性可以比较容易划分出来，划分结果见实验代码。针对数据可视化和摘要这部分实验，数据摘要的结果见相应的out.txt文件（在文件夹output中），文件中包含了对标称属性的每个取值的频数的统计，也包含了对数值属性的统计分析。数据可视化的结果见output文件夹中的相应png图片。在数据可视化的过程中，能很轻易通过直方图和QQ图发现数据的分布特征，直观看出哪些数据属性是正态分布，那些不是。例如，在数据集NFL Play-by-Play 2009-2017中，很明显可以看出WPA属性，ScoreDiff属性就是符合正态分布的。

针对缺失的数据可以采用四种策略对缺失进行处理：

1. 将缺失部分剔除，我选择对数据中凡是出现了NaN的行都进行删除，这样处理可以保证剩下的数据在每一个属性上都是有效的，但是也会造成很大的信息损失。
2. 用高频率值来填补缺失值，对于每一列，找出除了NaN之外出现频率最高的值对NaN值进行填充。
3. 通过属性相关关系来填补缺失值，这种方法需要根据属性间的线性关系来进行评估，选择一个线性关系最高的属性来评估另外一个属性缺失的部分。
4. 通过对数据对象之间的相似性来填补数据，需要对每一行数据进行对比，找到比较相似的数据，进行填充。

在对数据填充或者删除之后，用直方图的方式表示了处理前后的数据差异。见output文件夹中的名称中含有’compare’的图片