数据集NFLPlaybyPlay2009-2017 (v4)

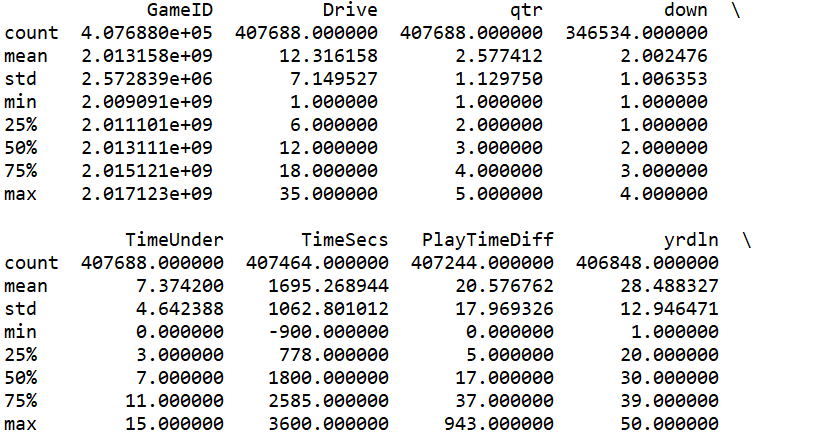
**一、数据可视化和摘要**

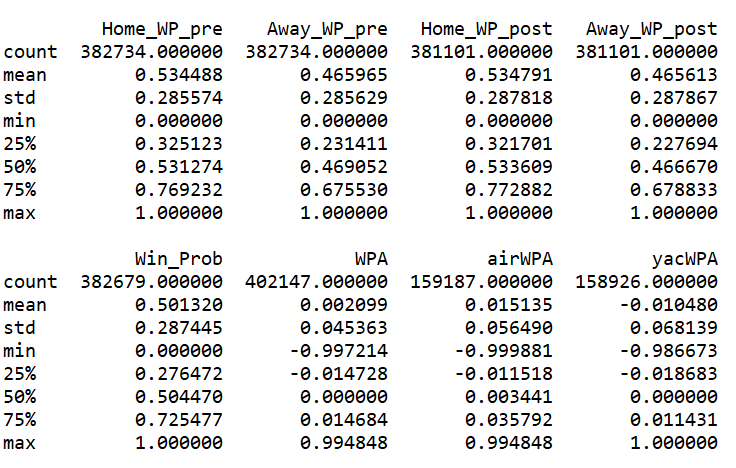
1. 数据摘要：

• 对标称属性，给出每个可能取值的频数，

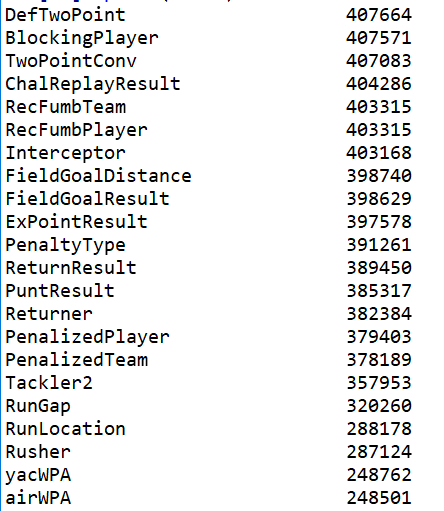
|  |  |
| --- | --- |
| **ExPointResult**  Aborted 7  Blocked 81  Made 9850  Missed 172  **TwoPointConv**  Failure 322  Success 283 | **PlayType**  End of Game 1973  Extra Point 10063  Field Goal 8928  Half End 40  Kickoff 23403  No Play 21414  Pass 159353  Punt 22003  QB Kneel 3530  Quarter End 4914  Run 120831  Sack 10649  Spike 640  Timeout 16206  Two Minute Warning 3741 |

• 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。

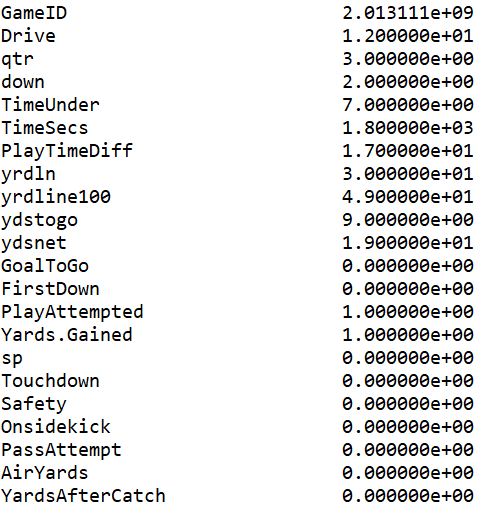




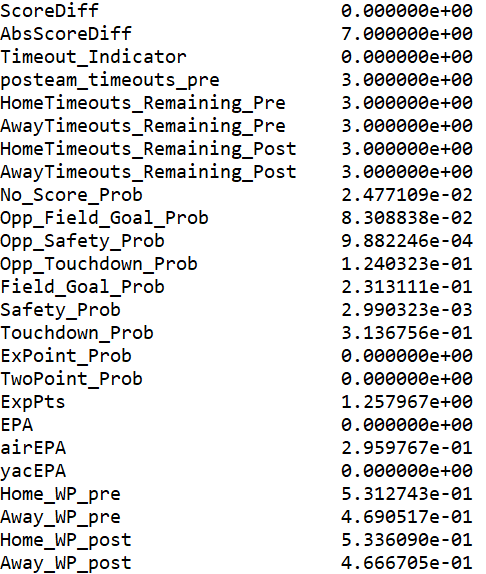
缺失数：



中位数：



。。。

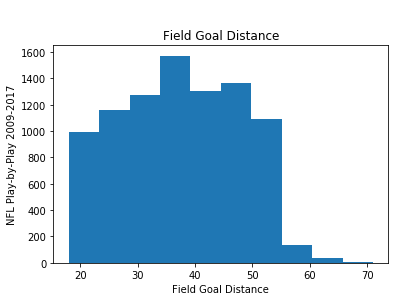


1. 数据的可视化

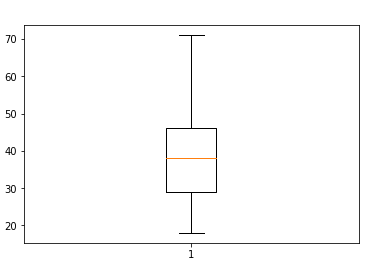
针对数值属性

（1）绘制直方图，用qq图检验其分布是否为正态分布

#直方图



（2）绘制盒图，对离群值进行识别



二、数据缺失的处理

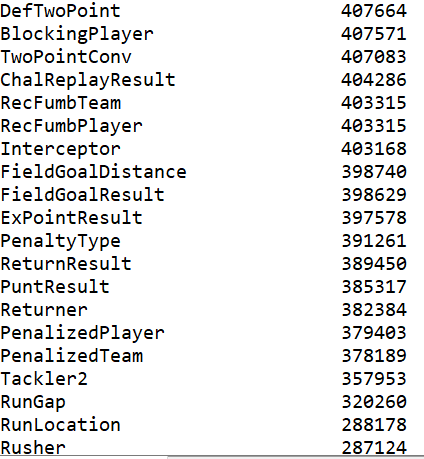
• 将缺失部分剔除

• 用最高频率值来填补缺失值

• 通过属性的相关关系来填补缺失值

• 通过数据对象之间的相似性来填补缺失值

#统计缺失数据



。。。

#去掉一列

df = df.drop(['TIDF Compliance'], axis = 1)

#填-1

df['Existing Units'] = df['Existing Units'].fillna(-1)

#使用平均数填补

df[' Revised Cost'] = df[' Revised Cost'].fillna(df[' Revised Cost'].mean())

#使用出现次数最多的值填补

df['Existing Construction Type'] = df['Existing Construction Type'].fillna('5.0')

df.product\_type[df.product\_type.isnull()]=df.product\_type.dropna().mode().values

#返回已经去掉重复行的数据集

df.drop\_duplicates()

#去掉这个特征为空的行

df\_new = df.drop(df[df['Description'].isnull()].index)