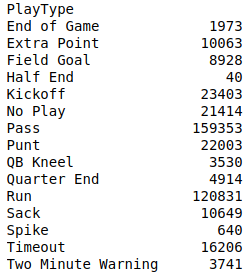
## 数据探索性分析与数据预处理

数据集一：NFL Play-by-Play 2009-2017

**1. 数据摘要**

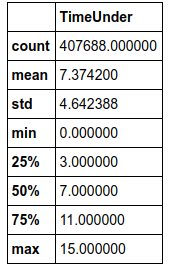
i. 标称属性

以“PlayType”属性为例，列举出了所有可能的取值，以及对应的频数：



ii. 数值属性

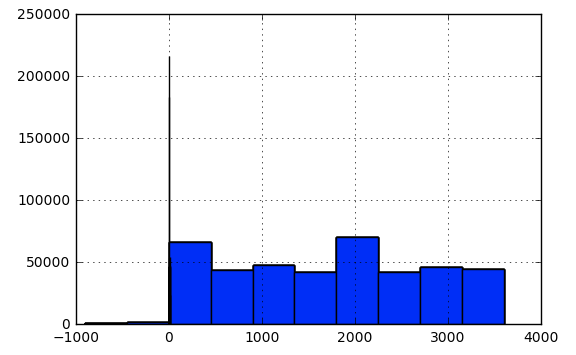
以“TimeUnder”属性为例，分别给出了非空值数据的个数（count），平均值（mean），方差（std），最小值（min），四分位数（min，25%，50%，75%，max）以及最大值（max）。



1. **数据可视化**

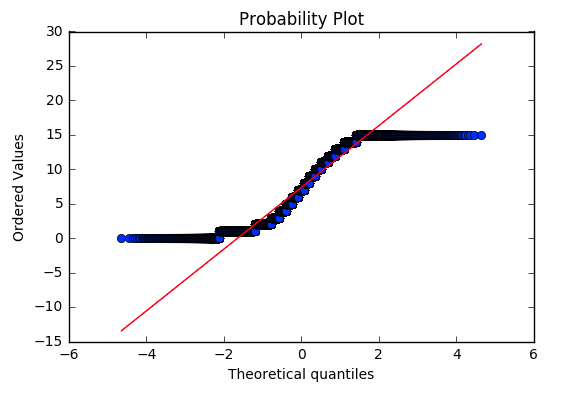
针对数值属性，均以“TimeUnder”属性为例：

i. 直方图



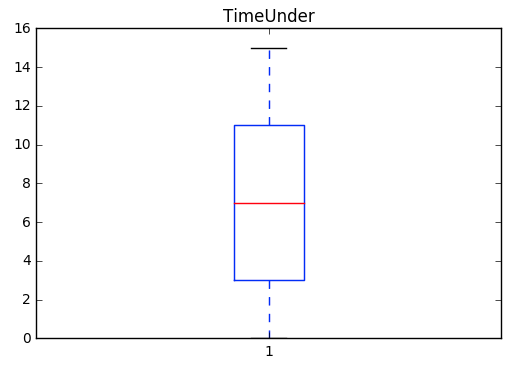
ii. Q-Q图

观察可以看出，Q-Q图上的点趋近于落在一条直线附近，因此该属性值的分布满足正态分布。



iii. 盒图

由盒图观察可得离群点。



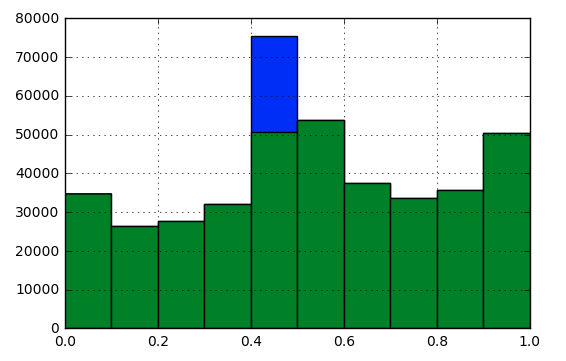
1. **数据缺失**

i. 数据缺失原因

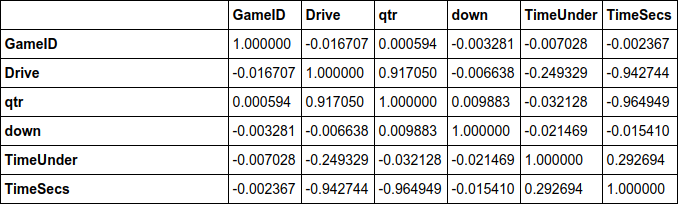
观察数据集中缺失的数据，原因主要可能是不存在这个值，或者值丢失。

ii. 处理缺失数据

剔除缺失部分（绿色） vs 用最高频率值来填补缺失值（蓝色）



对于数值属性，可以通过计算协方差矩阵，来判断数据之间的相似度，利用属性的相关关系来填补缺失值。下图截取部分协方差矩阵值，观察可以发现，”Drive“属性和”qtr“属性相关系数为0.91，二者之间的正相关性很高，因此当其中一个数据缺失时，可以使用另一个数据值进行填充。同理，”TimeSecs“属性和”Drive“、”qtr“属性之间的负相关性很高，它们之间也可以相互填补缺失值。

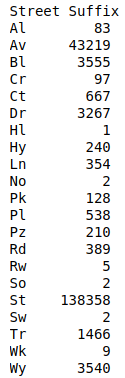


数据集二：San Francisco Building Permits

**1. 数据摘要**

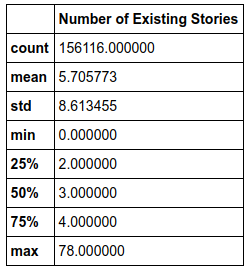
i. 标称属性

以“Street Suffix”属性为例，列举出了所有可能的取值，以及对应的频数：



ii. 数值属性

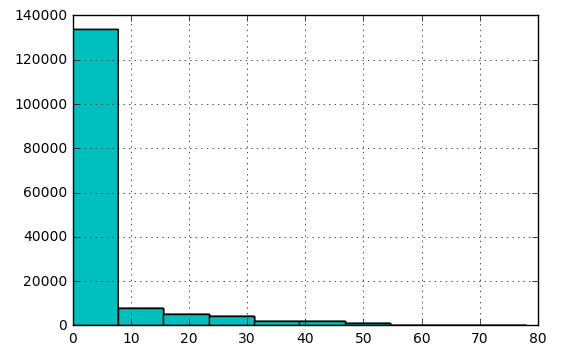
以“Number of Existing Stories”属性为例，分别给出了非空值数据的个数（count），平均值（mean），方差（std），最小值（min），四分位数（min，25%，50%，75%，max）以及最大值（max）。



1. **数据可视化**

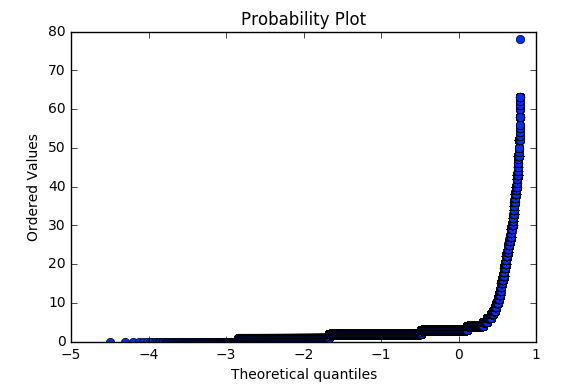
针对数值属性，均以“Number of Existing Stories”属性为例：

i. 直方图



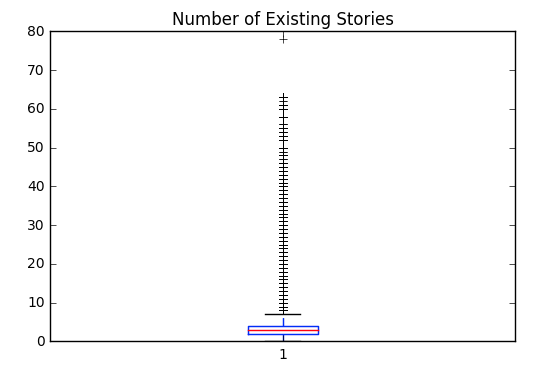
ii. Q-Q图

观察可以看出，Q-Q图上的点并不趋近于落在一条直线附近，因此该属性值的分布不满足正态分布。



iii. 盒图

由盒图观察可得离群点。



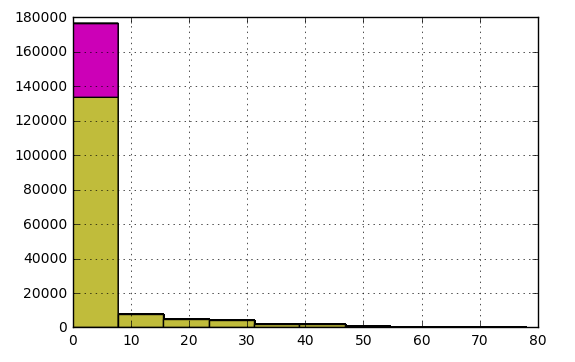
1. **数据缺失**

i. 数据缺失原因

观察数据集中缺失的数据，发现“Street Number Suffix”下有大量缺失值，由于它对地址来说并不是一个普遍存在的数据，所以我猜想它的缺失值根本不存在；有少量的“Zipcode”数据缺失，由于每个地址的邮编一定存在，所以它应该是没有被记录。

ii. 处理缺失数据

剔除缺失部分（黄色） vs 用最高频率值来填补缺失值（粉色）



对于数值属性，可以通过计算协方差矩阵，来判断数据之间的相似度，利用属性的相关关系来填补缺失值。下图截取部分协方差矩阵值，观察可以发现，”Number of Existing Stories“属性和”Number of Proposed Stories“属性相关系数为0.99，二者之间的相关性很高，因此当其中一个数据缺失时，可以使用另一个数据值进行填充。同理，”Estimated Cost“属性和”Revised Cost“属性之间的相关系数为0.97，也可以相互填补缺失值。

