数据探索性分析与数据预处理

2120171107

朱佳琪

计算机科学与技术

# 1. 问题描述

本次作业中，将对2个数据集进行探索性分析与预处理。

* 数据集1: NFL Play-by-Play 2009-2017
* 数据集2: San Francisco Building Permits

数据集中属性解释：

* 数据集1中数值属性有：'Number of Existing Stories', 'Number of Proposed Stories', 'Estimated Cost', 'Revised Cost','Existing Units', 'Proposed Units'
* 数据集2中数值属性有：'Drive', 'qtr', 'down', 'TimeUnder', 'TimeSecs', 'PlayTimeDiff', 'yrdln', 'yrdline100', 'ydstogo', 'ydsnet', 'Yards.Gained', 'AirYards', 'YardsAfterCatch', 'FieldGoalDistance', 'Penalty.Yards', 'PosTeamScore', 'DefTeamScore', 'ScoreDiff', 'AbsScoreDiff', 'posteam\_timeouts\_pre', 'HomeTimeouts\_Remaining\_Pre', 'AwayTimeouts\_Remaining\_Pre', 'HomeTimeouts\_Remaining\_Post', 'AwayTimeouts\_Remaining\_Post', 'No\_Score\_Prob', 'Opp\_Field\_Goal\_Prob', 'Opp\_Safety\_Prob', 'Opp\_Touchdown\_Prob', 'Field\_Goal\_Prob', 'Safety\_Prob', 'Touchdown\_Prob', 'ExPoint\_Prob', 'TwoPoint\_Prob', 'ExpPts', 'EPA', 'airEPA', 'yacEPA', 'Home\_WP\_pre', 'Away\_WP\_pre', 'Home\_WP\_post', 'Away\_WP\_post', 'Win\_Prob', 'WPA', 'airWPA', 'yacWPA'

# 2. 实验环境

* 语言及环境依赖

语言： python

依赖的包：pandas,numpy,scipy,json,matplotlib

# 3. 数据分析

## 3.1 数据可视化和摘要

##### 3.1.1数据摘要

* **对标称属性，给出每个可能取值的频数**

使用python字典统计其各个对标称属性的频数，并将字典存入json文件中。其JSON结构如下：

{‘对标陈属性1’：频数；‘对标称属性2’：频数；……}

以数据集San Francisco Building Permits为例，对标称属性数据摘要信息如下：

"{'Permit Number': {'201602179765': 101,…,…,…,…}, 'Permit Type Definition': {'otc alterations permit': 178844, …,…,…,…},…,…,}

* **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。**

对数值属性，通过遍历其每个属性的取值，其JSON结构如下：

{‘数值属性1’：[最大数，最小数，均值，中位数，[Q1, Q2, Q3]，缺失值的个数], …,…,…}

以数据集San Francisco Building Permits为例，数值属性的数据摘要如下：

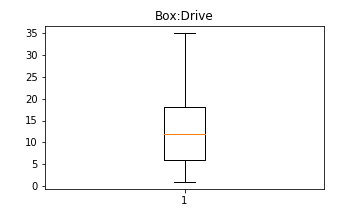
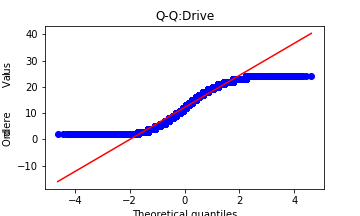
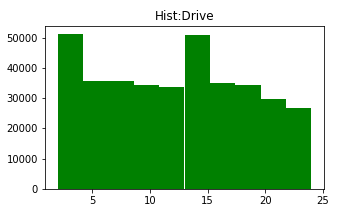
{"Number of Existing Stories": [78.0, 0.0, 5.705773271157344, 3.0, [2.0, 3.0, 4.0], 42784], "Number of Proposed Stories": [78.0, 0.0, 5.745042683552092, 3.0, [2.0, 3.0, 4.0], 42868], "Estimated Cost": [537958646.0, 1.0, 168955.44329681536, 11000.0, [3300.0, 11000.0, 35000.0], 38066], "Revised Cost": [780500000.0, 0.0, 132856.1864917494, 7000.0, [1.0, 7000.0, 28710.0], 6066], "Existing Units": [1907.0, 0.0, 15.666164275729155, 1.0, [1.0, 1.0, 4.0], 51538], "Proposed Units": [1911.0, 0.0, 16.510950138185947, 2.0, [1.0, 2.0, 4.0], 50911]}

##### 3.1.2数据的可视化

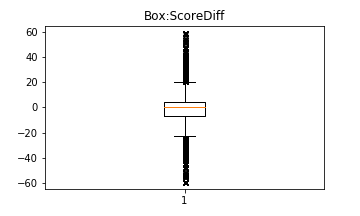
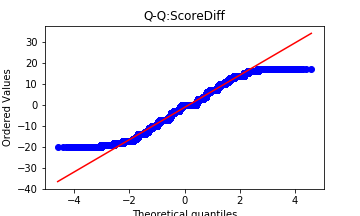
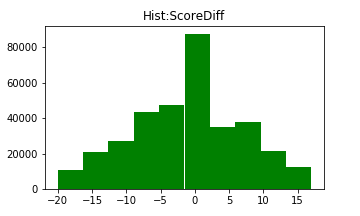
针对数值属性绘制直方图，用qq图检验其分布是否为正态分布。绘制盒图，对离群值进行识别。

从数据集NFL Play-by-Play 2009-2017和San Francisco Building Permits中选取几组较有特征的数值属性分析如下：

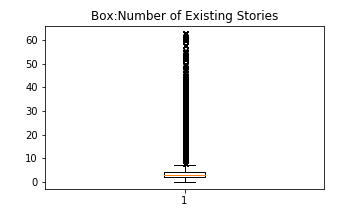
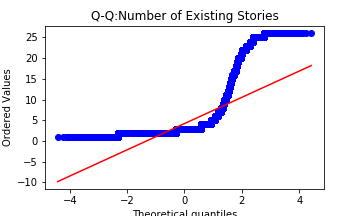
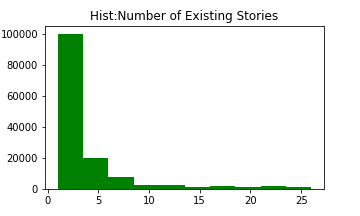
1. Drive是数据集NFL Play-by-Play 2009-2017的数值属性，由下图所示直方图、QQ图可以看出Drive属性不满足正态分布，由盒图可以看出，该属性所对应的数据分布较为集中，几乎不存在离群点。



1. ScoreDiff是数据集NFL Play-by-Play 2009-2017的数值属性，由下图所示直方图、QQ图可以看出ScoreDiff属性满足正态分布，由盒图可以看出，该属性所对应的数据分布不够集中，两端存在大量离群点。



3. Number of Existing Stories 是数据集San Francisco Building Permits的数值属性，由下图所示直方图、QQ图可以看出ScoreDiff属性不满足正态分布，由盒图可以看出，该属性所对应的数据分布不够集中也不够均匀，顶端存在大量离群点。

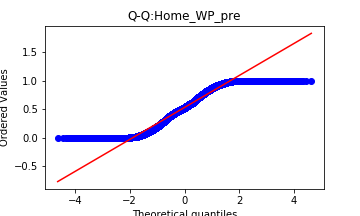
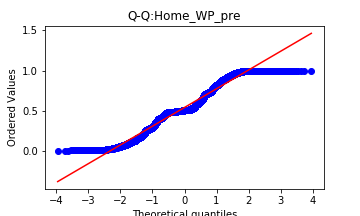
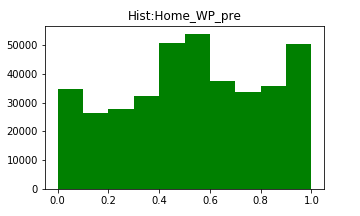
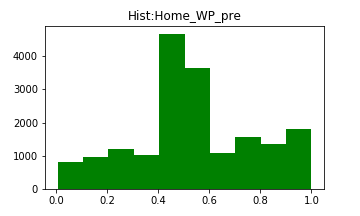
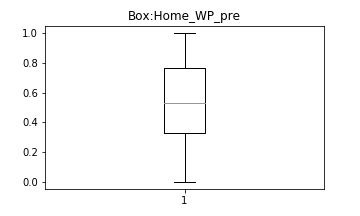
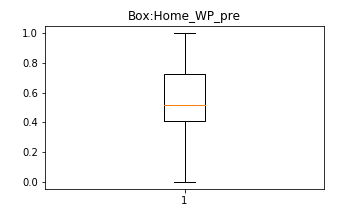


## 3.2 数据缺失的处理

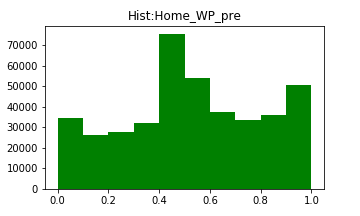
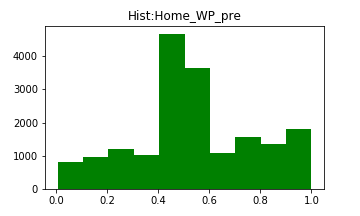
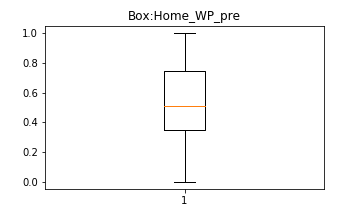
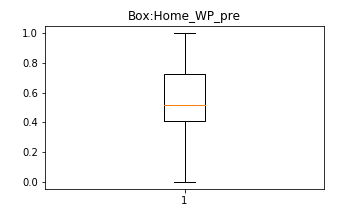
观察数据集中缺失数据，分析其缺失的原因。

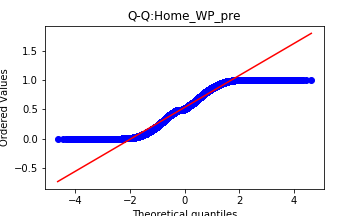
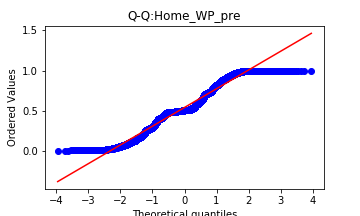
分别使用下列四种策略对缺失值进行处理:

* 将缺失部分剔除，以Home\_WP\_pre数值属性为例，缺失数据剔除前后效果对比如下：

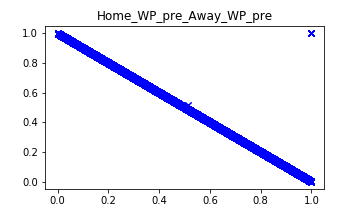


* 用最高频率值来填补缺失值，以Home\_WP\_pre数值属性为例，最高频率值来填补缺失值前后效果对比如下：

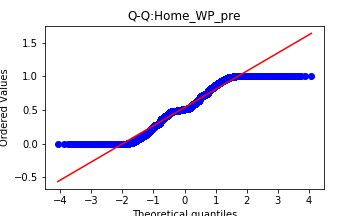
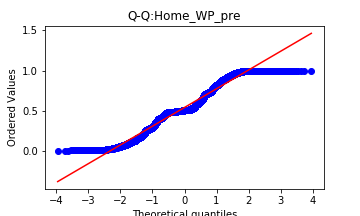
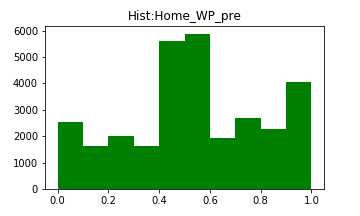
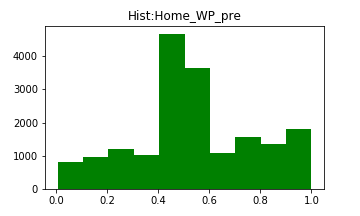
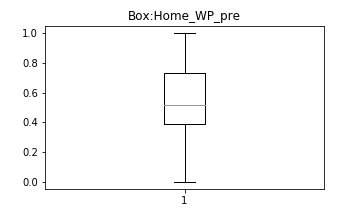
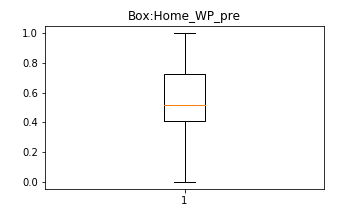




* 通过属性的相关关系来填补缺失值，其中数据集NFL Play-by-Play 2009-2017中，Away\_WP\_pre属性和Home\_WP\_pre属性满足Away\_WP\_pre+ Home\_WP\_pre = 1。如下图所示;

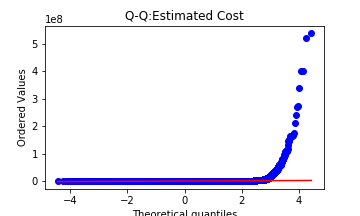
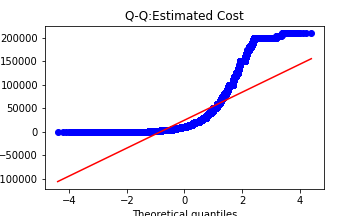
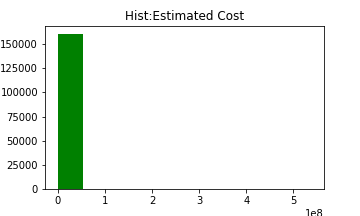
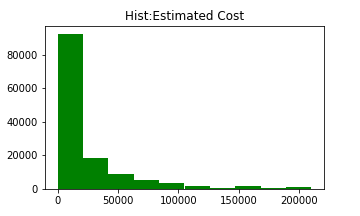
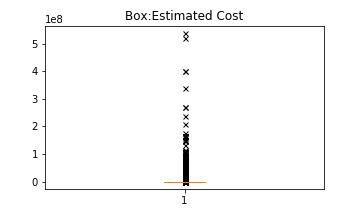
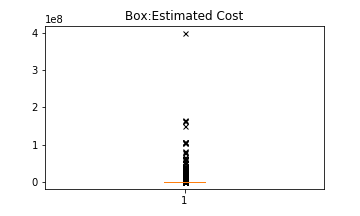


将属性Home\_WP\_pre按属性相关关系填充之后，前后对比如下：

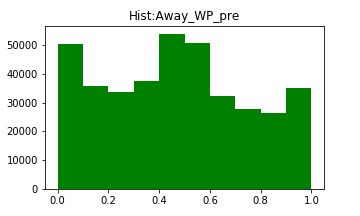
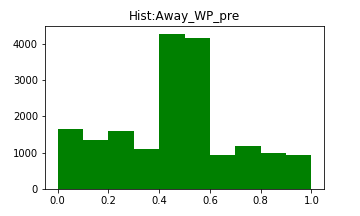
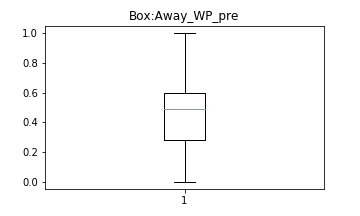
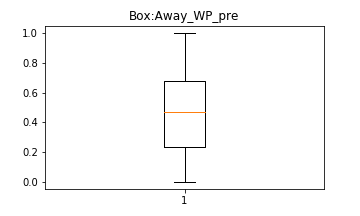


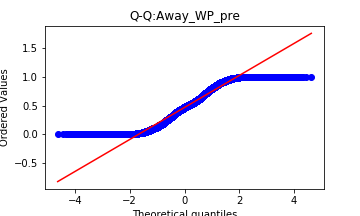
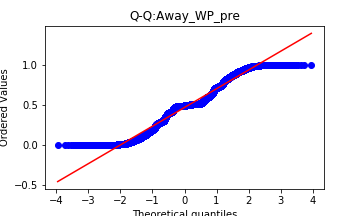
* 通过数据对象之间的相似性来填补缺失值，通过计算两列属性间的距离，选取最近的属性进行填充，填充结果如下，（总觉的好像哪里不对，但是两个数据集的两个属性填充效果都没什么规律可循）

Estimated Cost属性填充对比效果如下：



Away\_WP\_pre属性填充对比效果如下：





# 4. 实验结果

* 数据摘要所得json信息在results文件夹中。
* 实验过程中所产生的所有图片信息在result\_pic文件夹中，实验报告中只选取部分有特征的属性进行分析。