数据挖掘 ---- 作业一

程伟 2120160984 第10组

# 要求：

1、数据可视化和摘要

数据摘要

* 对标称属性，给出每个可能取值的频数，
* 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。

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

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
* 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

处理后，可视化地对比新旧数据集。

# 实验环境

本实验程序基于python开发，使用的第三方库如下：

pandas、numpy、scipy -- 分析数据

matplotlib -- 绘图

json – 存储格式化数据

# 实验结果

## 数据集Building\_Permits

### 数据摘要：

存放标称属性统计结果的json文件结构设计：{‘属性1’：频数；‘属性2’：频数；……}

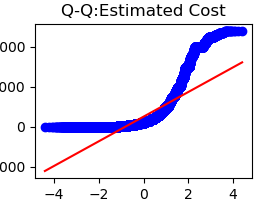
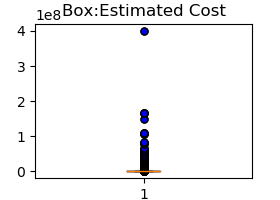
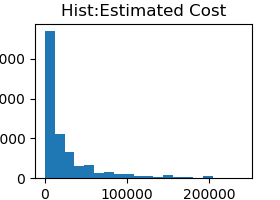
存放数值属性统计结果的json文件结构设计：{‘属性1’：[最大数，最小数，均值，中位数，[Q1, Q2, Q3]，缺失值的个数]；‘属性2’：[…….]；……}

标称属性的统计结果存放在Building\_Permits\_nominal\_attr.json文件中，数值属性的统计结果则存放在Building\_Permits\_numeric\_attr.json文件中。

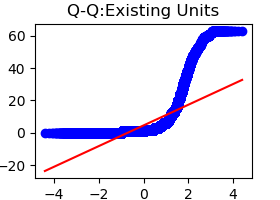
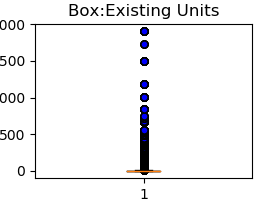
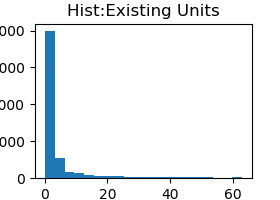
数据集Building\_Permits标称属性和数值属性是通过人工来区分的。

### 数据可视化：

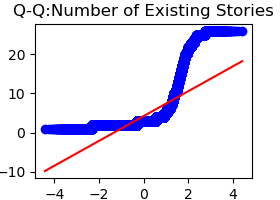
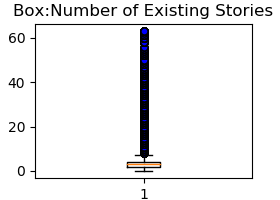
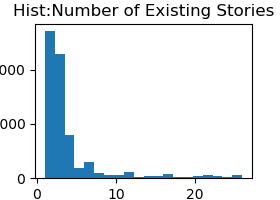
Estimated Cost属性直方图、盒图以及qq图，由图可知该属性离散点较多，不满足正态分布；



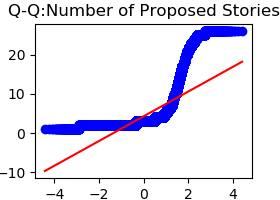
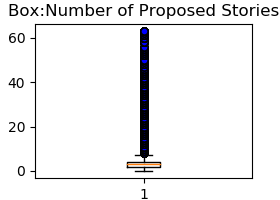
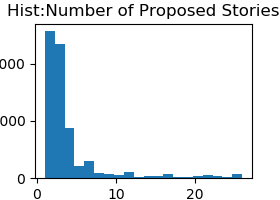
Existing Units属性直方图、盒图以及qq图，由图可知该属性离散点较多，不满足正态分布；



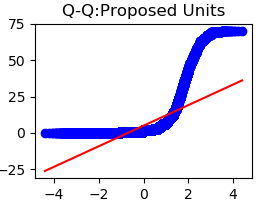
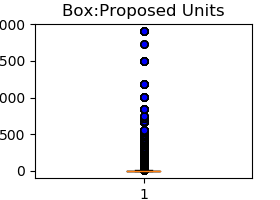
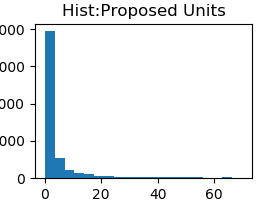
Number of Exiting Stories 属性直方图、盒图以及qq图，由图可知该属性离散点较多，不满足正态分布；



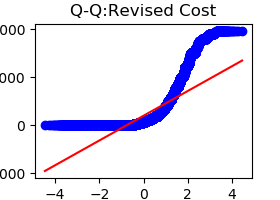
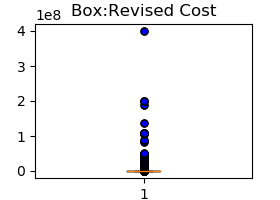
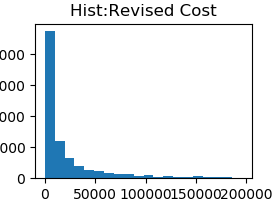
Number of Proposed Stories 属性直方图、盒图以及qq图，由图可知该属性离散点较多，不满足正态分布；



Proposed Units属性直方图、盒图以及qq图，由图可知该属性离散点较多，不满足正态分布；



Revised Cost属性直方图、盒图以及qq图，由图可知该属性离散点较多，不满足正态分布；

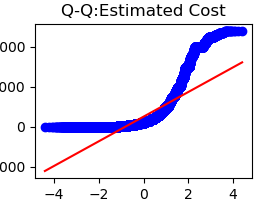
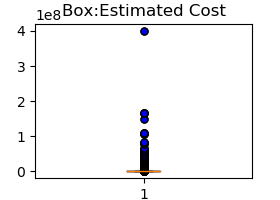
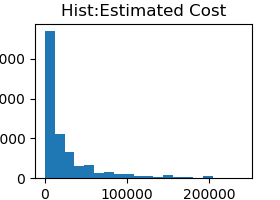


### 数据缺失处理：

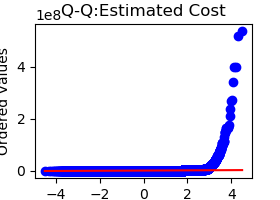
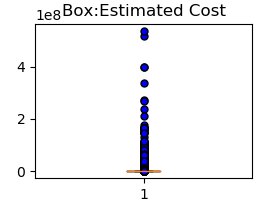
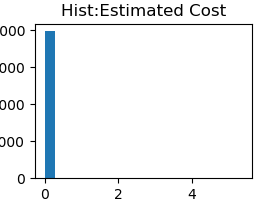
因为属性数量很多，且缺失值也很多，在文档中难以将所有属性的缺失处理结果都展示出来，故只展示Estimated Cost这个属性的缺失处理结果。

分别使用下列四种策略对缺失值进行处理，旧数据图示结果在上一部分已给出，此处不再重复展示:

1将缺失部分剔除，Estimated Cost属性直方图、盒图、qq图如下所示：



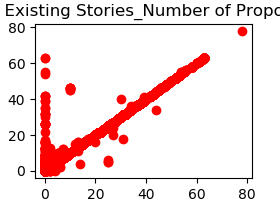
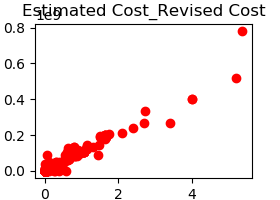
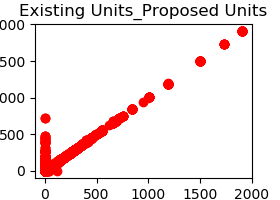
2用最高频率值来填补缺失值，Estimated Cost属性直方图、盒图、qq图如下所示：



因为缺失值很多，全部填补之后会使得众数更多，使得直方图中其他数值显得很小。

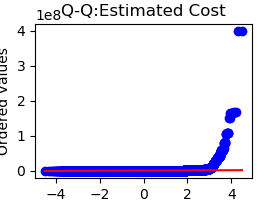
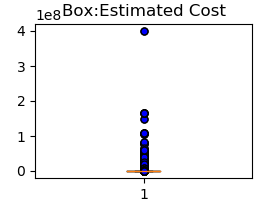
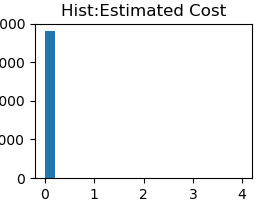
3通过属性的相关关系来填补缺失值：

通过分析属性间的相关性，可以得知'Existing Units'属性和'Proposed Units'属性呈正相关，'Estimated Cost'属性和'Revised Cost'属性呈正相关，'Number of Existing Stories'属性和 'Number of Proposed Stories'属性呈正相关。具体相关性如下图所示：

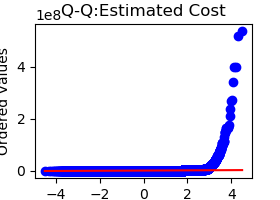
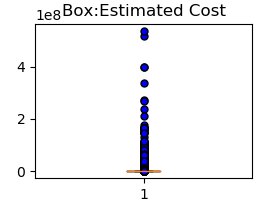
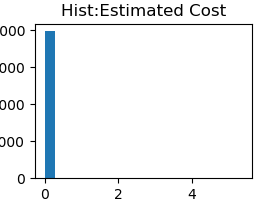


其中'Estimated Cost'属性和'Revised Cost'属性成1:1正比，即value(Estimated Cost) = value(Revised Cost)。则可以按此比例来填补Estimated Cost属性的缺失值。

Estimated Cost属性直方图、盒图、qq图如下所示:



4通过数据对象之间的相似性来填补缺失值，此处是利用混合属性的相异性度量方法来等效计算相似性的。Estimated Cost属性直方图、盒图、qq图如下所示：



## 数据集 NFL Play by Play 2009-2017 (v4)

### 数据摘要：

存放标称属性统计结果的json文件结构设计：{‘属性1’：频数；‘属性2’：频数；……}

存放数值属性统计结果的json文件结构设计：{‘属性1’：[最大数，最小数，均值，中位数，[Q1, Q2, Q3]，缺失值的个数]；‘属性2’：[…….]；……}

标称属性的统计结果存放在NFL Play by Play 2009-2017 (v4)\_nominal\_attr.json文件中，数值属性的统计结果则存放在NFL Play by Play 2009-2017 (v4)\_numeric\_attr.json文件中。

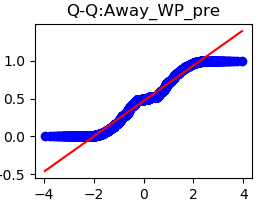
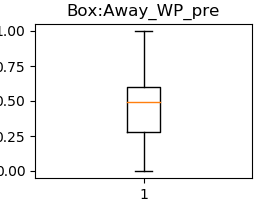
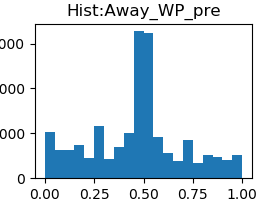
数据集NFL Play by Play 2009-2017 (v4)因为属性太多，便通过一些条件来区分标称属性和数值属性。

最后通过条件筛选出的条件有'AbsScoreDiff', 'AirYards', 'Away\_WP\_post', 'Away\_WP\_pre', 'DefTeamScore', 'Drive', 'EPA', 'ExPoint\_Prob', 'ExpPts', 'FieldGoalDistance', 'Field\_Goal\_Prob', 'Home\_WP\_post', 'Home\_WP\_pre', 'No\_Score\_Prob', 'Opp\_Field\_Goal\_Prob', 'Opp\_Safety\_Prob', 'Opp\_Touchdown\_Prob', 'Penalty.Yards', 'PlayTimeDiff', 'PosTeamScore', 'Safety\_Prob', 'ScoreDiff', 'TimeSecs', 'TimeUnder', 'Touchdown\_Prob', 'WPA', 'Win\_Prob', 'Yards.Gained', 'YardsAfterCatch', 'airEPA', 'airWPA', 'yacEPA', 'yacWPA', 'ydsnet', 'ydstogo', 'yrdline100', 'yrdln'。

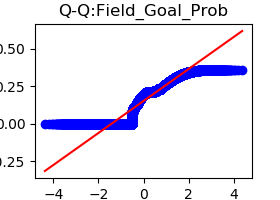
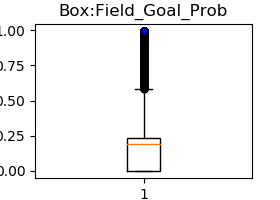
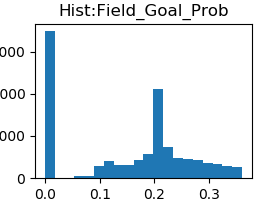
由于篇幅有限，报告中只展示其中'Away\_WP\_pre','FieldGoalDistance','Field\_Goal\_Prob', 'Home\_WP\_pre','Opp\_Field\_Goal\_Prob', 'Touchdown\_Prob','yrdline100', 'yrdln'等属性。

### 数据可视化：

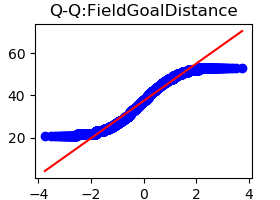
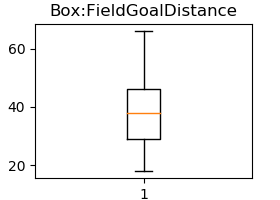
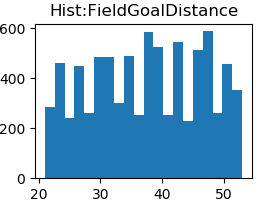
Away\_WP\_pre属性直方图、盒图以及qq图，由图可知该属性离散点少，不满足正态分布；



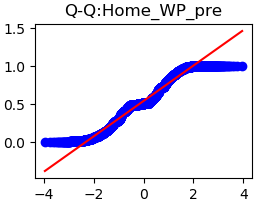
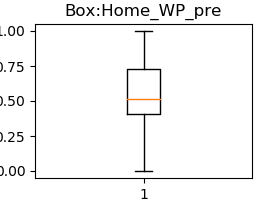
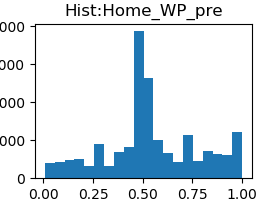
Field\_Goal\_Prob属性直方图、盒图以及qq图，由图可知该属性离散点多，不满足正态分布；



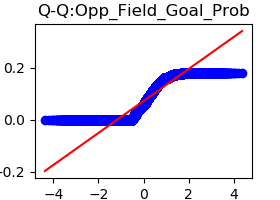
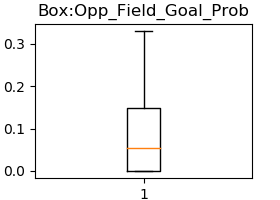
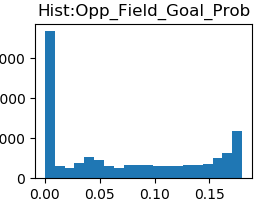
FieldGoalDistance属性直方图、盒图以及qq图，由图可知该属性离散点少，不满足正态分布；



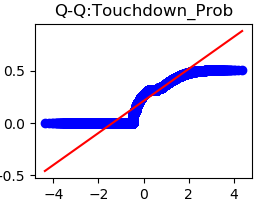
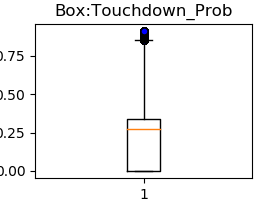
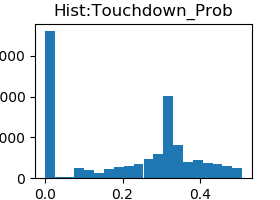
Home\_WP\_pre属性直方图、盒图以及qq图，由图可知该属性离散点少，近似一个正态分布；



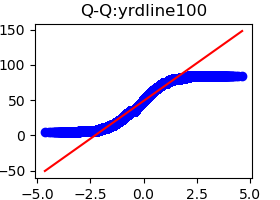
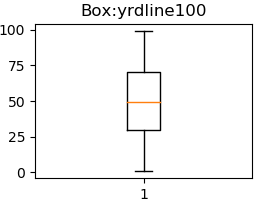
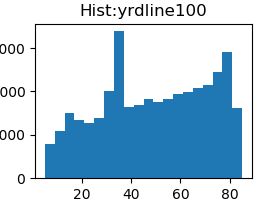
属性直方图、盒图以及qq图，由图可知该属性离散点少，不满足正态分布；



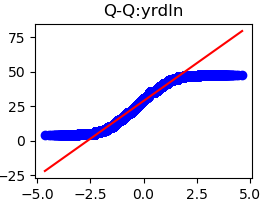
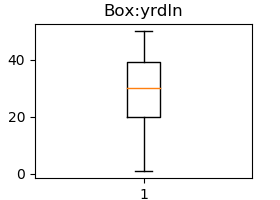
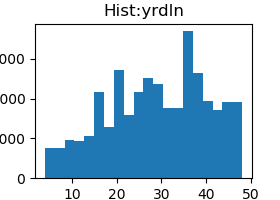
Opp\_Field\_Goal\_Prob属性直方图、盒图以及qq图，由图可知该属性离散点少，不满足正态分布；



yrdline100属性直方图、盒图以及qq图，由图可知该属性离散点少，不满足正态分布；



'yrdln'属性直方图、盒图以及qq图，由图可知该属性离散点少，不满足正态分布；

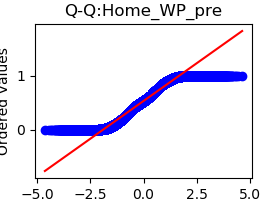
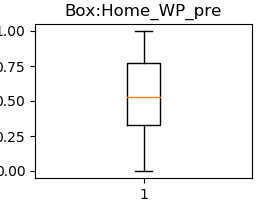
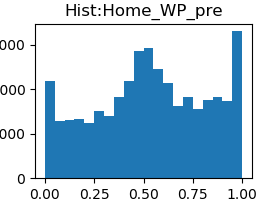


### 数据缺失处理：

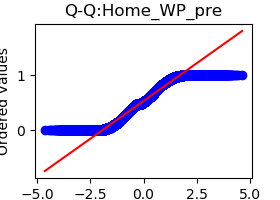
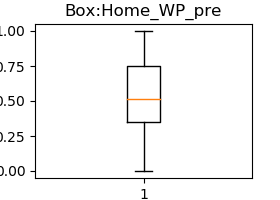
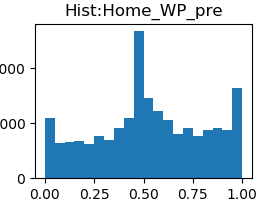
因为属性数量很多，且缺失值也很多，在文档中难以将所有属性的缺失处理结果都展示出来，故只展示Home\_WP\_pre这个属性的缺失处理结果。

分别使用下列四种策略对缺失值进行处理，旧数据图示结果在上一部分已给出，此处不再重复展示:

1 将缺失部分剔除，Home\_WP\_pre属性直方图、盒图、qq图如下所示:

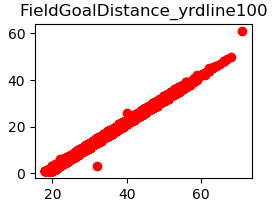
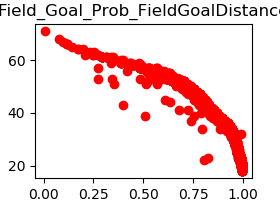


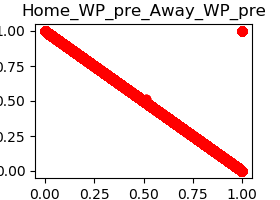
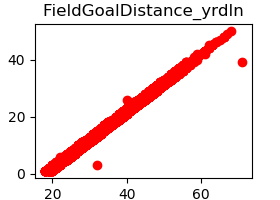
2 用最高频率值来填补缺失值，Home\_WP\_pre属性直方图、盒图、qq图如下所示:

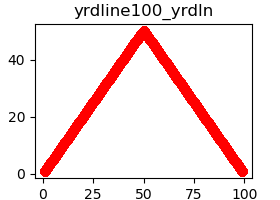
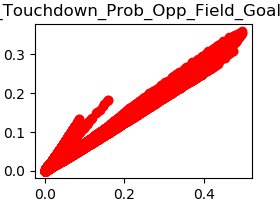


3通过属性的相关关系来填补缺失值：

通过分析属性间的相关性，可以得知'FieldGoalDistance'属性和'yrdline100'属性呈正相关，'FieldGoalDistance'属性和'yrdln'属性呈正相关，'Number of Existing Stories'属性和 'Number of Proposed Stories'属性呈正相关, 'Away\_WP\_pre'属性和'Home\_WP\_pre'属性呈负相关，'Touchdown\_Prob'属性'Opp\_Field\_Goal\_Prob'属性呈正相关和。具体相关性如下图所示：

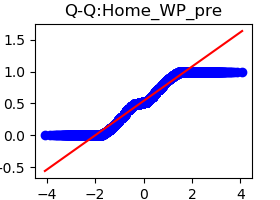
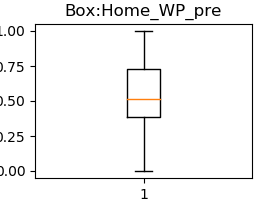
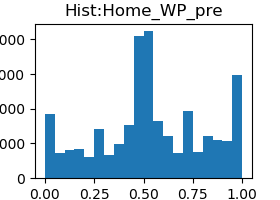






其中'Away\_WP\_pre'属性和'Home\_WP\_pre'属性满足value(Away\_WP\_pre) = 1-value(Home\_WP\_pre)

Home\_WP\_pre属性直方图、盒图、qq图如下所示:



4 通过数据对象之间的相似性来填补缺失值，此处是利用混合属性的相异性度量方法来等效计算相似性的。Home\_WP\_pre属性直方图、盒图、qq图如下所示:

