数据探索性分析与数据预处理

分析报告

赵文天 2120171105

1. **数据可视化和摘要**
2. 数据摘要

数据集中的所有属性分为标称属性和数值属性。标称属性的值可以用一组符号的有限或无限集合表示，并且所有值之间没有大小关系。数值属性的值可以用整数或实数集合表示。在要处理的数据集中，有些属性的值的形式是数值，但不能以数值属性看待。如用0和1表示真或假的属性或用数字表示编号的属性，对其求均值没有意义，应当视为标称属性。因此，在对数据进行处理前，需要根据数据集的的每个属性的意义区分数值属性和标称属性。

数据中的1.csv和2.csv分别是对NFL Play to Play数据集和Building Permits数据集中的数值属性的标记。文件中第一行为数据集中的属性名称，第二行的值为对应列的属性是否为数值属性，数值属性被标记为1。

1. 数值属性的可视化

对于每个数值属性，绘制了直方图，分位图（Q-Q图）以及盒图。

绘制直方图时，在属性的取值范围内划分了30个bin，统计每个bin内的样本数并绘制直方图。绘制Q-Q图时，分别取属性的取值集合中的分位数和正态分布的分位数，并分别以两组分位数为横坐标和纵坐标在直角坐标系中描点。在Q-Q图中，直线y=x作为标记在图中画出，用于确定两组数据的分布是否接近，即判断某个属性的取值是否符合正态分布。盒图中标记出了四分位数以及离群值。

1. **数据缺失的处理**
2. 数据集中的缺失数据

在NFL play by play数据集中共有约41万个样本，根据1.1中的方法对NFL Play by Play数据集中的数值属性和标称属性进行区分，共得到42个数值属性。其中部分属性的缺失值较多，这些属性的缺失情况以及对缺失值的分析如下：

属性FieldGoalDistance：共有398740个缺失值。

属性airEPA：共有248394个缺失值。

属性yacEPA：共有248498个缺失值。

属性airWPA：共有248501个缺失值。

属性yacWPA：共有248762个缺失值。

在Building Permits数据集中共有约19万个样本，对其中的数值属性和标称属性进行区分后得到6个数值属性。对这些属性中的缺失值的分析如下：

属性Number of Existing Stories：共有42784个缺失值。此属性的含义是建筑中现有的层数，数据字典中说明该属性对特定种类的许可不适用，这可能是该属性中有缺失值的原因。

属性Number of Proposed Stories：共有42868个缺失值。此属性的含义是申请建造或改建的层数。

属性Estimated Cost：共有38066个缺失值，该属性表示工程的初始预算。

属性Revised Cost：共有6066个缺失值，该属性表示工程修订后的预算。

属性Existing Units：共有51538个缺失值。

属性Proposed Units：共有50911个缺失值。

1. 对缺失值进行填补

在对缺失值进行填补的过程中，使用了四种处理方法，分别是剔除缺失部分，用最高频率值（即众数）填补缺失值，通过属性的相关关系填补缺失值以及通过数据对象之间的相似性填补缺失值。对于直接剔除缺失值和用众数填补缺失值，其方法是显然的。

对于通过数据对象之间的相似性填补缺失值的方法，首先考虑某个有缺失值的样本。使用k nearest neighbor方法得到与其最接近的若干个样本，并用这些样本中的属性填补缺失值。样本之间的相似程度用欧氏距离度量，在计算某两个有缺失值的样本时，忽略两个样本中缺失的属性。

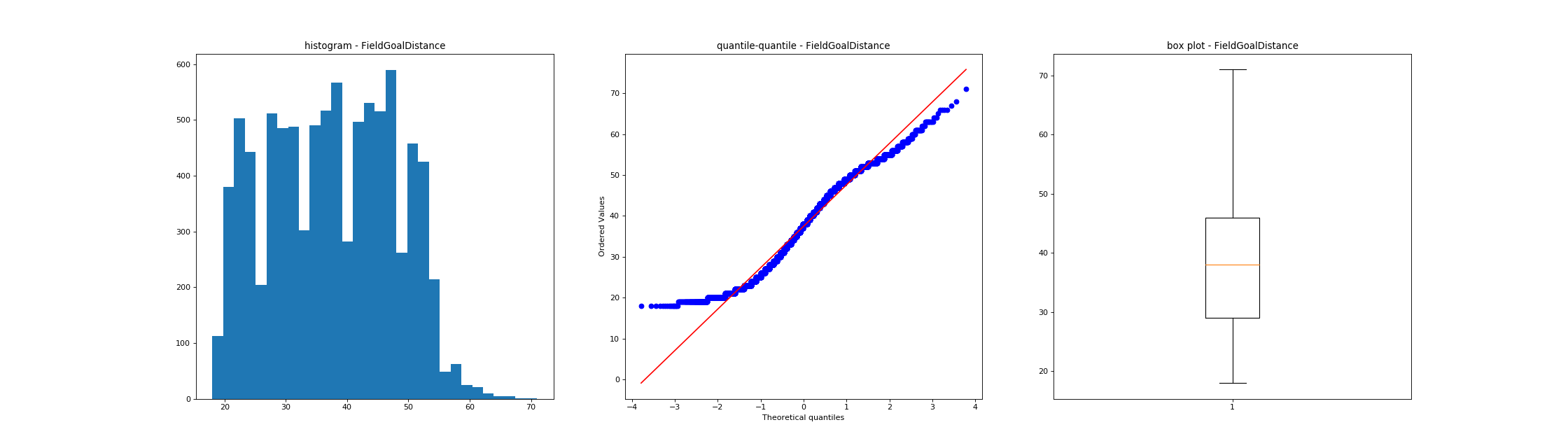
在通过属性的相关关系填补缺失值时，使用皮尔逊相关系数度量两个属性之间的相关程度。对于随机变量X和Y，其计算公式为

其中，表示X与Y的协方差，D(X)和D(Y)分别表示X和Y的方差。对数据集中的m个数值属性求两两之间的相关系数，得到m\*m的相关矩阵。在填补某个属性的缺失值时，根据相关矩阵中的相关系数确定该属性与其他属性之间的线性相关关系，并确定要填补的值。

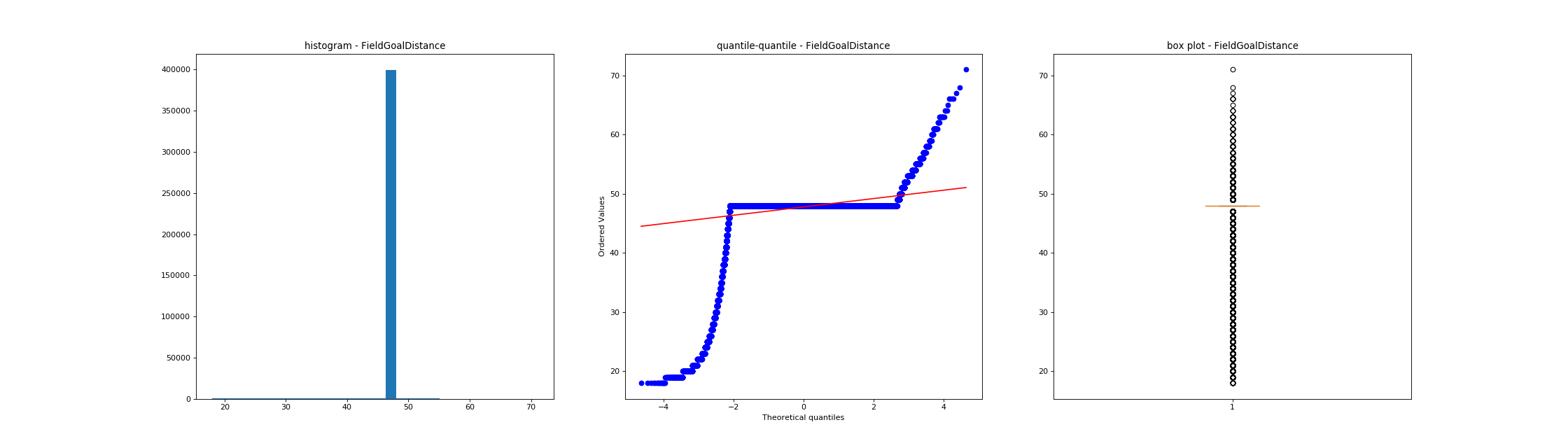
1. 新旧数据集的对比

对NFL Play by Play数据集进行缺失值填补后，对原数据集中缺失值较多的属性与填补后的属性进行对比。

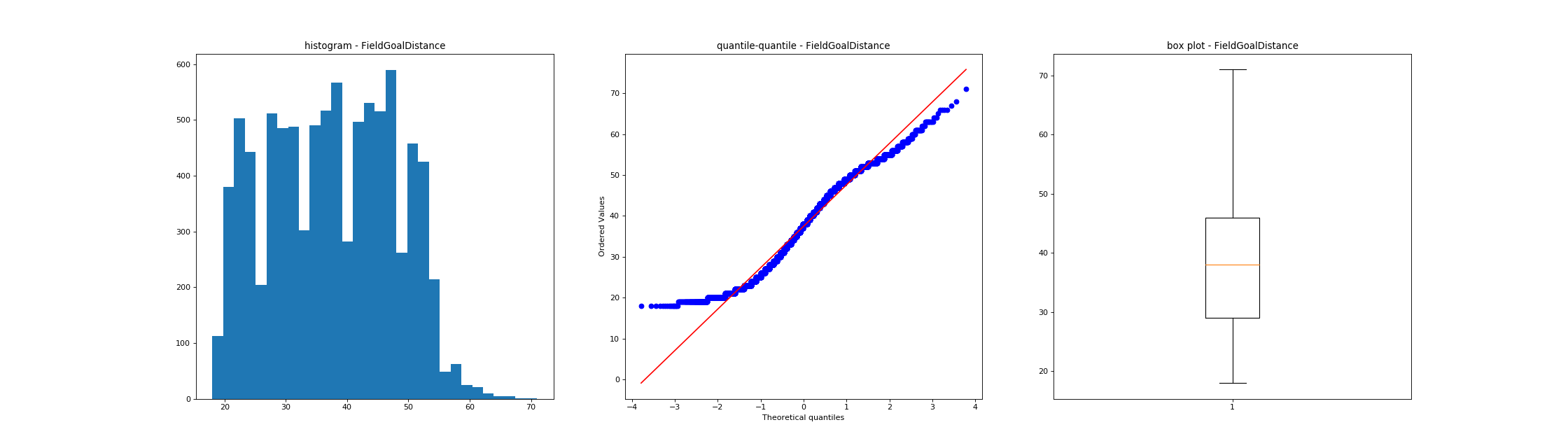
属性FieldGoalDistance：



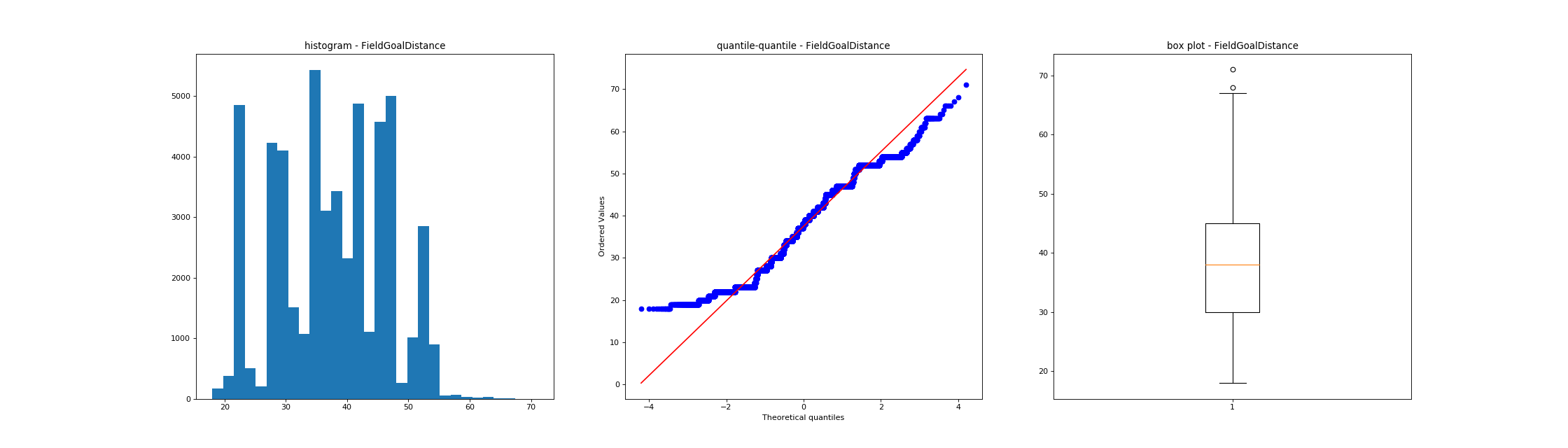
(a)



(b)



(c)



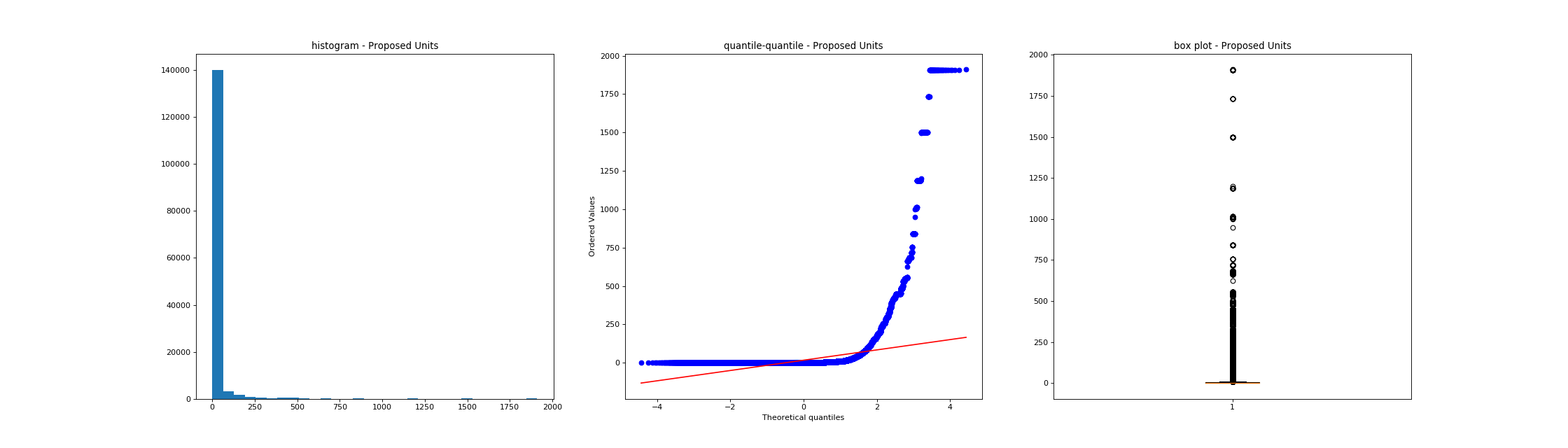
(d)

上图(a)(b)(c)(d)分别表示剔除缺失值，使用众数填补缺失值，通过属性的相关关系填补缺失值以及通过样本之间的相关性填补缺失值处理后的属性。左侧为直方图，中间为Q-Q图，右侧为盒图。

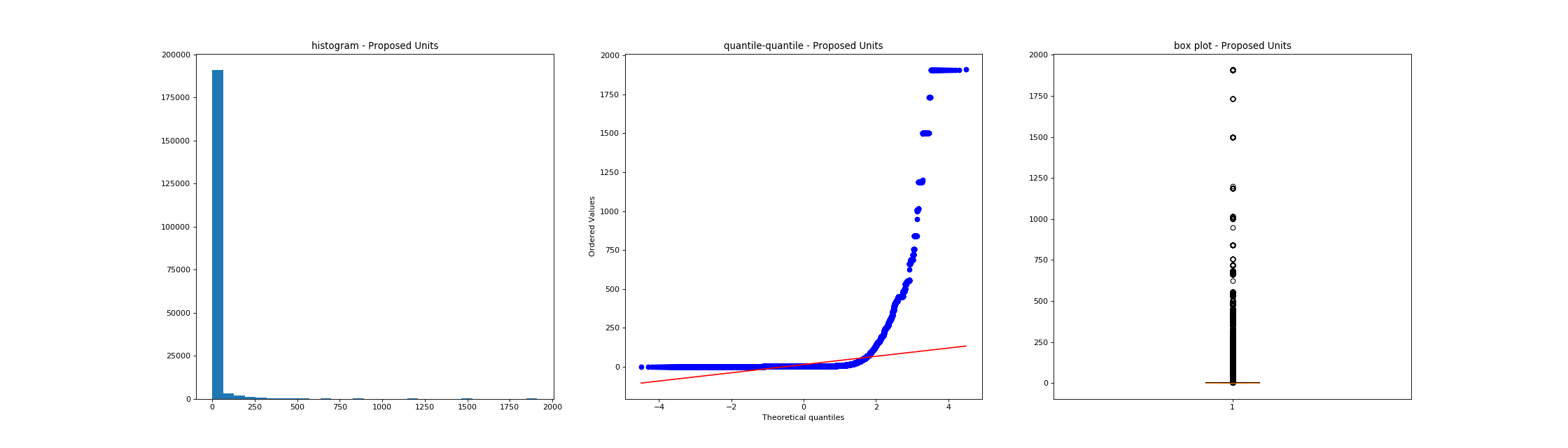
由于该属性有约39万个缺失值，即该属性的绝大多数值都缺失。由直方图可知，用最高频数的值填补后，绝大多数值都集中在一处。除最高频数值之外的其他值都变为离群值。Q-Q图显示原始数据在20-60的范围内与正态分布较为接近，但用最高频数值填补后数据的分布与正态分布相差更大。用属性的相关关系填补缺失值以及通过样本之间的相关关系填补缺失值的方法填充后，数据的分布与原始数据的分布较为接近。这说明对于绝大多数值缺失的属性，使用最高频数的值填充会造成较大的偏差。

对Building Permits数据集进行缺失值填补后，对原数据集中有缺失的属性与填补后的数据进行对比。

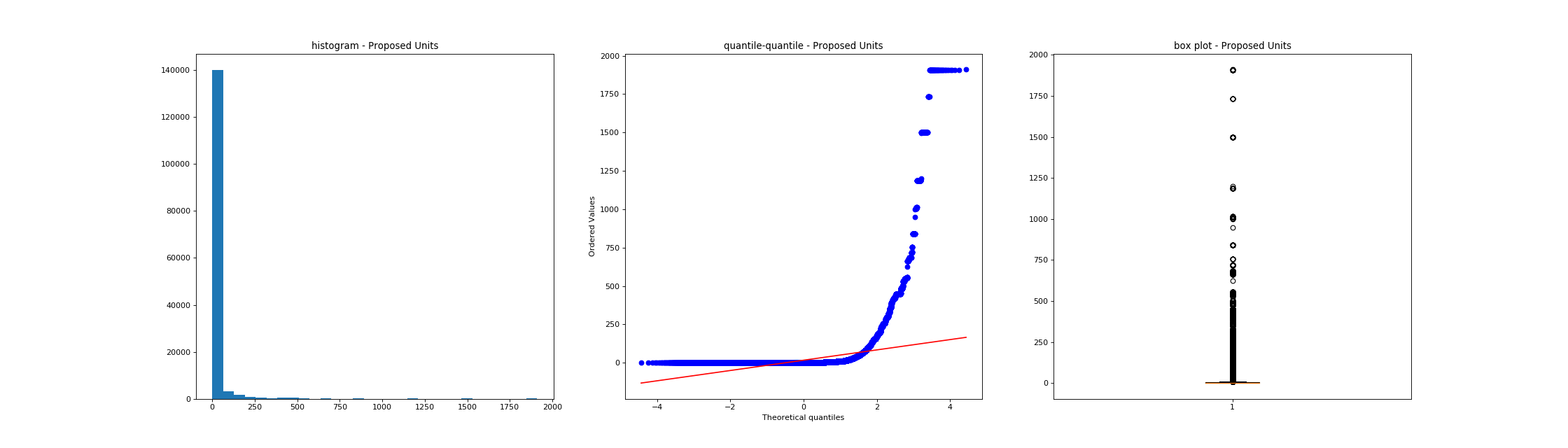
属性Estimated Cost:



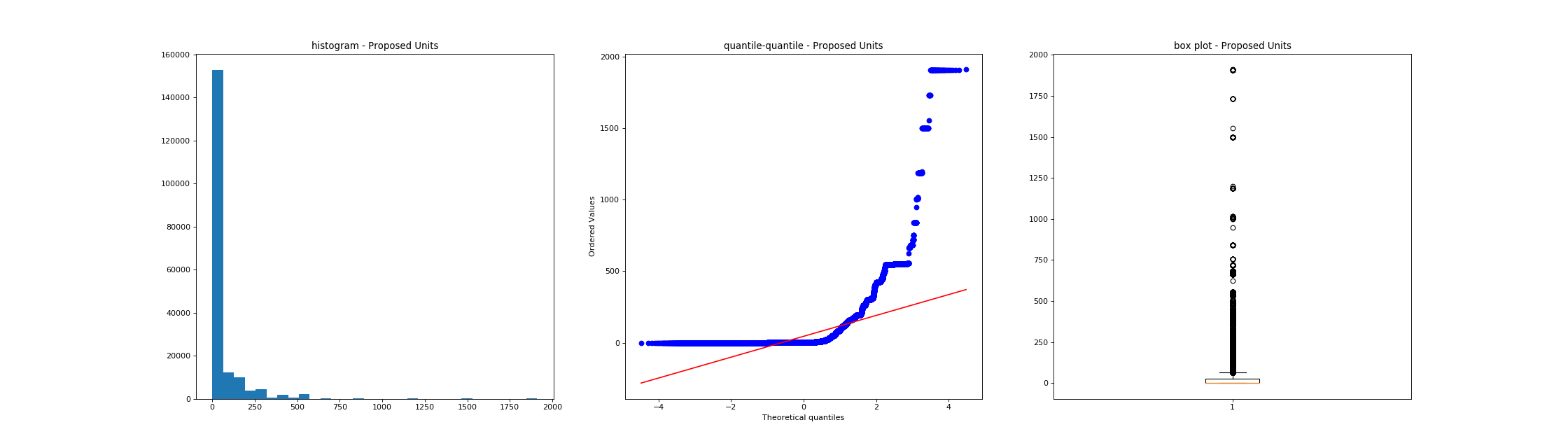
(a)



(b)



(c)



(d)

上图(a)(b)(c)(d)分别表示剔除缺失值，使用众数填补缺失值，通过属性的相关关系填补缺失值以及通过样本之间的相关性填补缺失值处理后的属性。由于原数据集中，该属性的值的分布属于长尾分布。数据集中共有198900个样本，其中17090个样本的Estimated Cost属性的值在0至60直接，其他样本分布在60至537958646之间。在对缺失值进行填补后，若对整个数据集中该属性的值进行分析，发现该属性的值的分布没有发生明显变化。