数据挖掘作业一

作业一使用R语言完成。

1. 数据可视化和摘要

为简便起见，报告中只用几个典型例子说明。

* 1. 数据摘要

R语言中有非常完善的数据摘要工具，这部分使用dim()函数查看数据集的维度，names()函数查看列的名称。

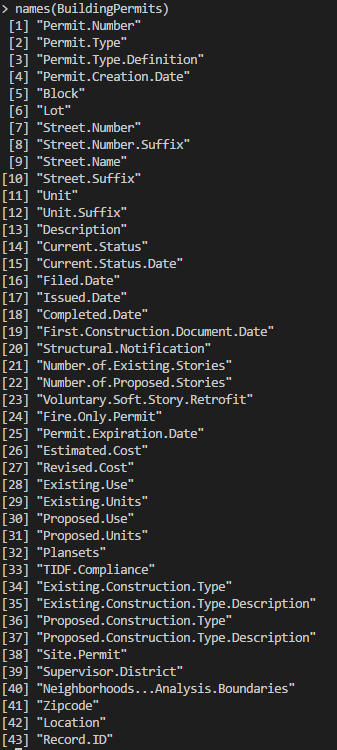
* 首先使用read.csv()函数导入数据集。



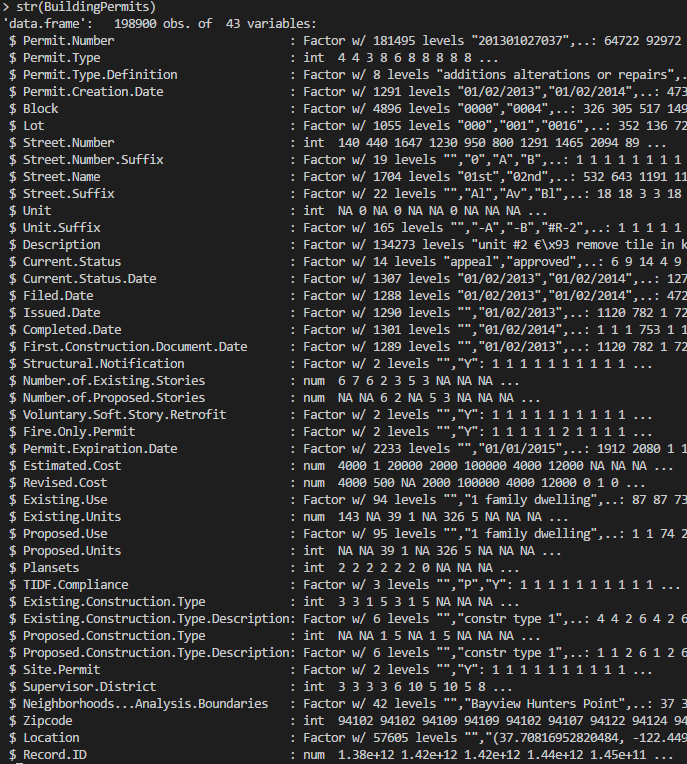
* 使用dim()函数查看数据集的维度



* 使用names()函数查看列的名称

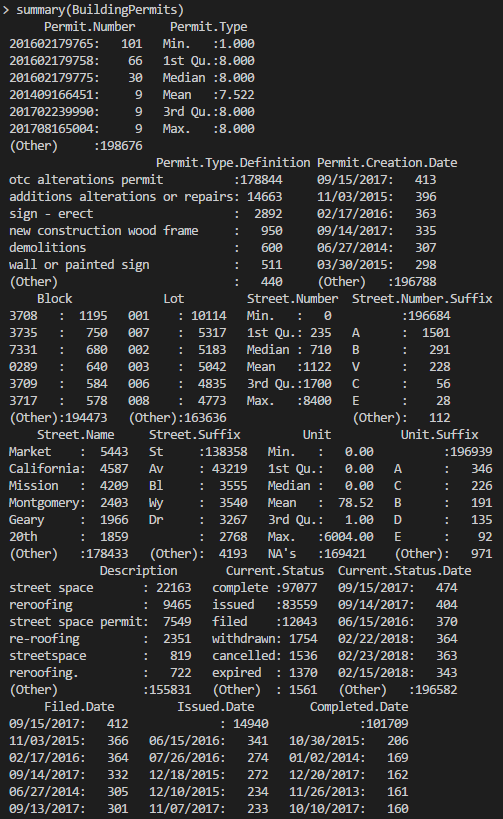


* 使用str（）函数查看数据类型



可以看到BuildingPermits数据集中的数据类型有factor，int和num这三种

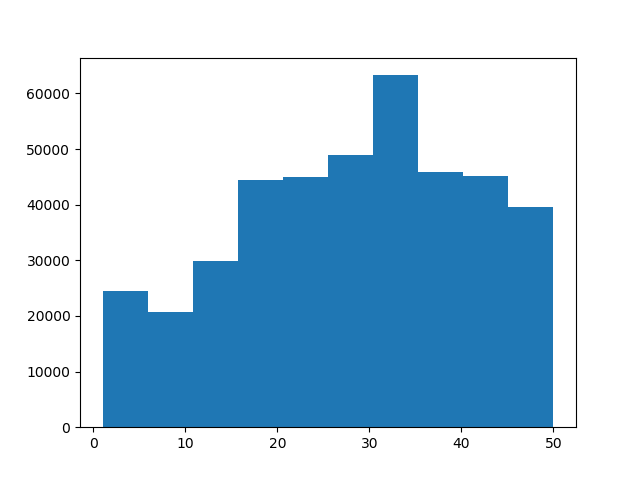
* summary（）函数返回数据的分布情况，对于数值类型，summary（）函数返回最大值、最小值和四个分位数，对于标称数据，summary（）函数返回每个可能取值的频数



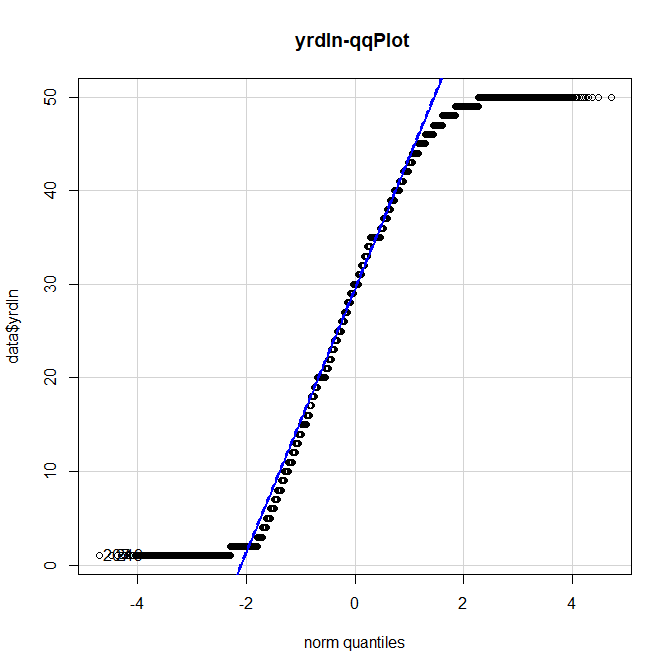
1.2数据可视化

以NFL Play by Play 2009-2017 (v4).csv中的yrdln为例

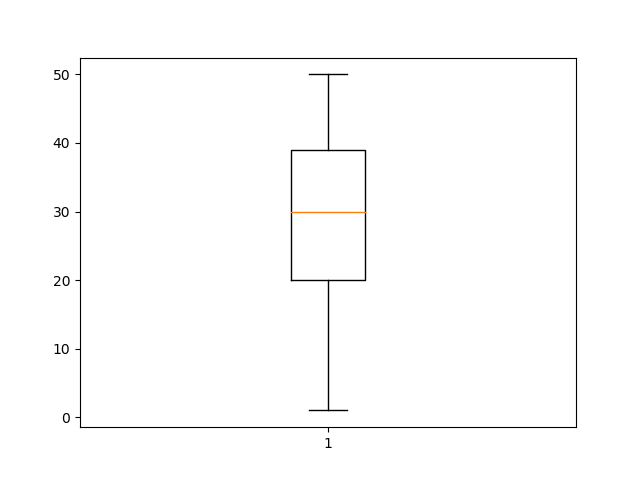
（1）直方图为：



（2）QQ图为：



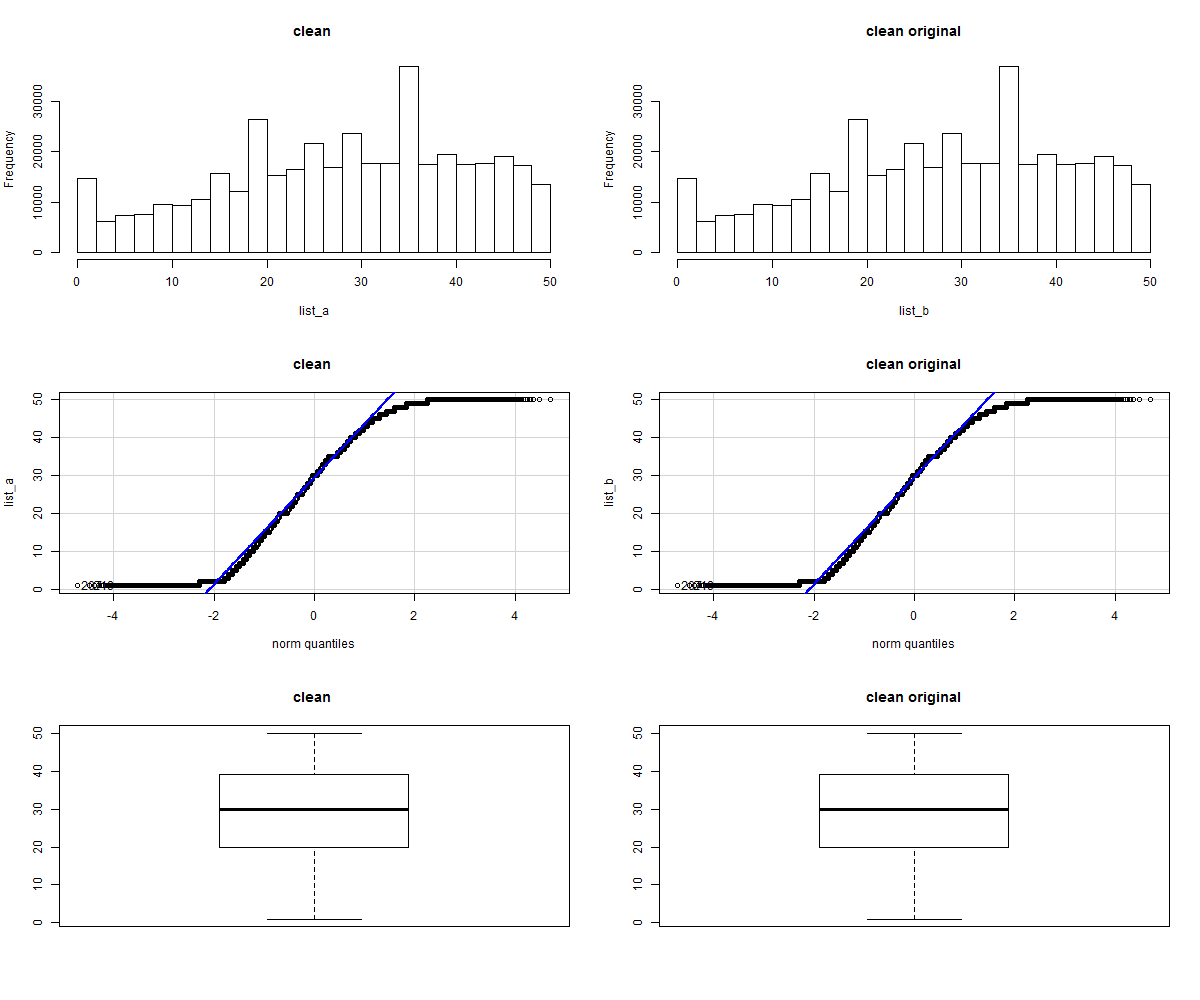
（3）盒图为：



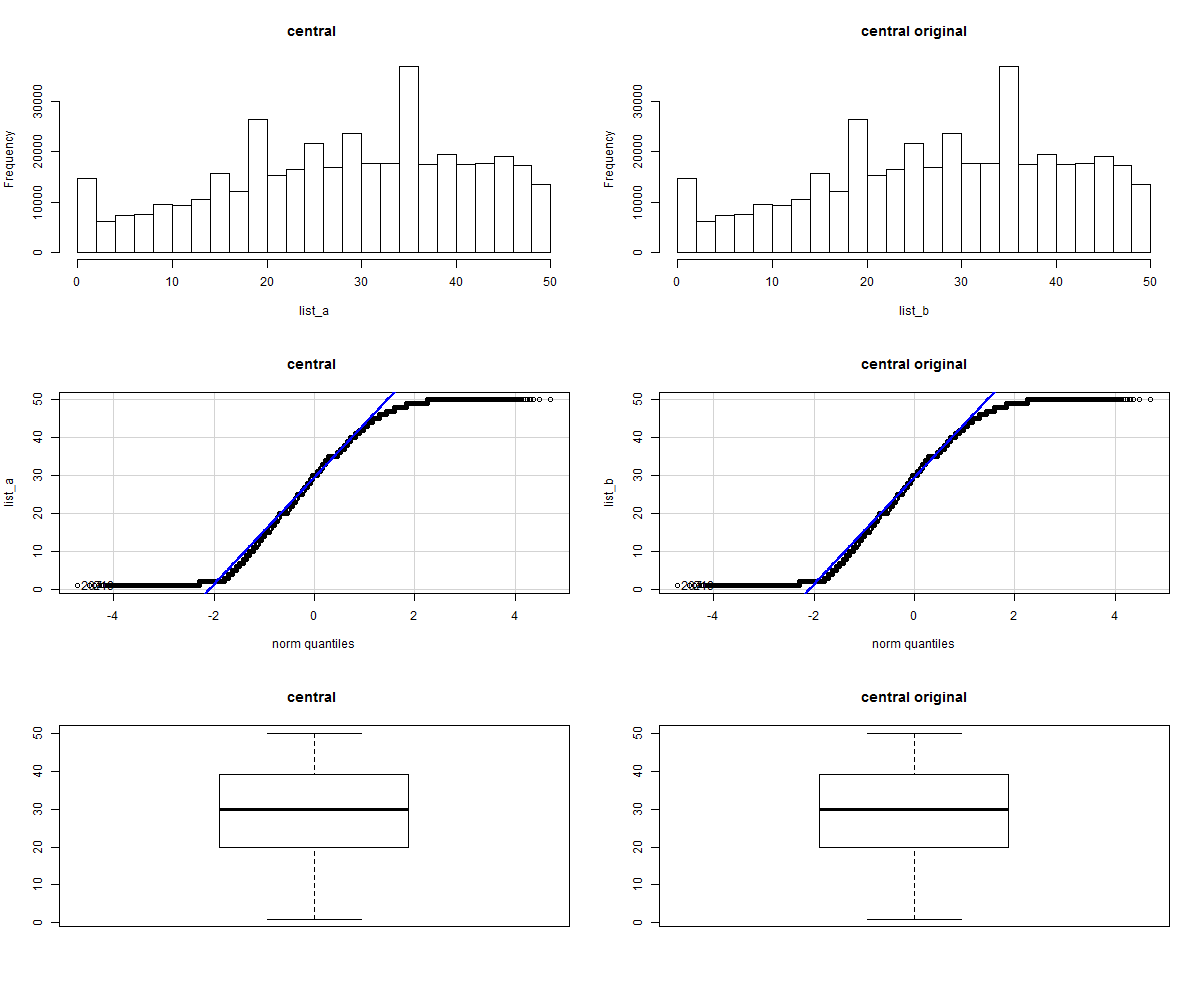
1. 数据缺失的处理

为简便起见，文档以NFL Play by Play 2009-2017 (v4).csv中的yrdln为例，使用4中缺失值处理策略对数据集进行处理，并绘制前后对比图。

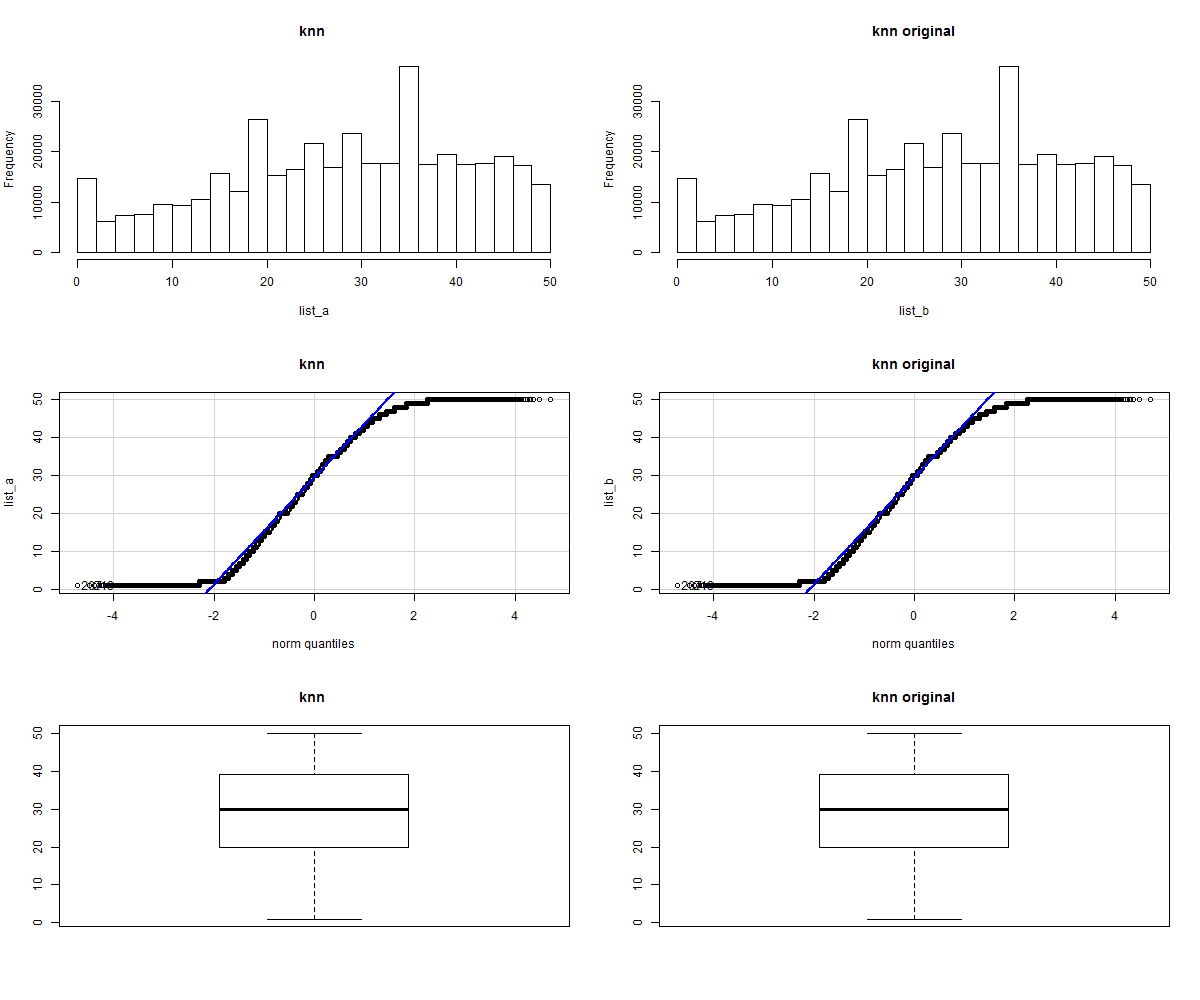
* 将缺失部分剔除



* 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

