数据探索性分析与数据预处理

在本次作业中，将使用R语言对2个数据集进行探索性分析与预处理,由于数据量庞大，现选取其中个别数据量进行分析和处理。选取数据属性如下：

数据集1：NFL Play-by-Play 2009-2017

"Drive" - Drive number

"sp" - Whether the play resulted in a score (any kind of score)

"qtr" - Quarter of Game

"down" - Down of the given play

"time" - Time at start of play

"yrdln" - Between 0 and 50

"ydstogo" - Yards to go for a first down

"ydsnet" - Total yards gained on a given drive

"posteam" - The team on offense

"AirYards" - Number of yards the ball was thrown in the air for both complete and incomplete pass attempts (negative means behind line of scrimmage)

"YardsAfterCatch" - Number of yards receiver gained after catch

"QBHit" - Binary: 1 if the QB was knocked to the ground else 0

"desc" - A detailed description of what occured during the play

数据集2： San Francisco Building Permits

Permit Type Definition - "Description of the Permit type, for example new construction, alterations"

Current Status - Current status of the permit application.

Structural Notification - Notification to meet some legal need, given or not

Number of Existing Stories - "Number of existing stories in the building. Not applicable for certain permit types"

Number of Proposed Stories - Number of proposed stories for the construction/alteration

"Voluntary Soft-Story Retrofit" - Soft story to meet earth quake regulations

Fire Only Permit - Fire hazard prevention related permit

Estimated Cost - Initial estimation of the cost of the project

Revised Cost - Revised estimation of the cost of the project

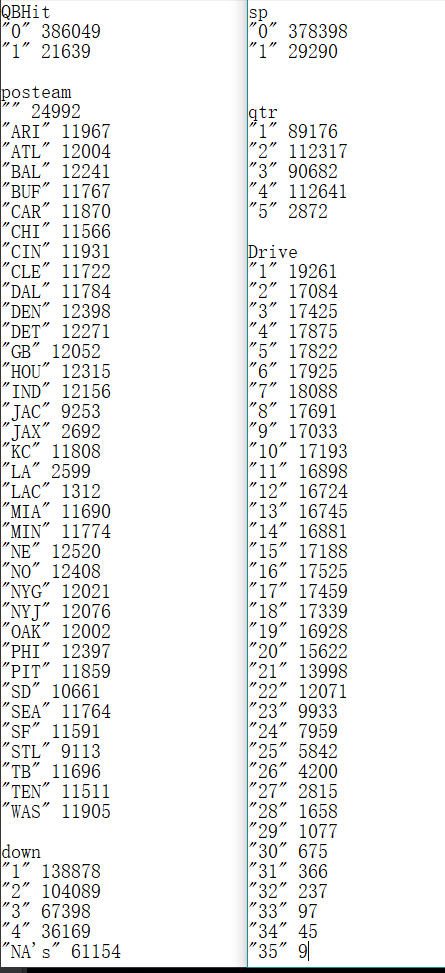
Proposed Use - Proposed use of the building

1. 数据可视化和摘要

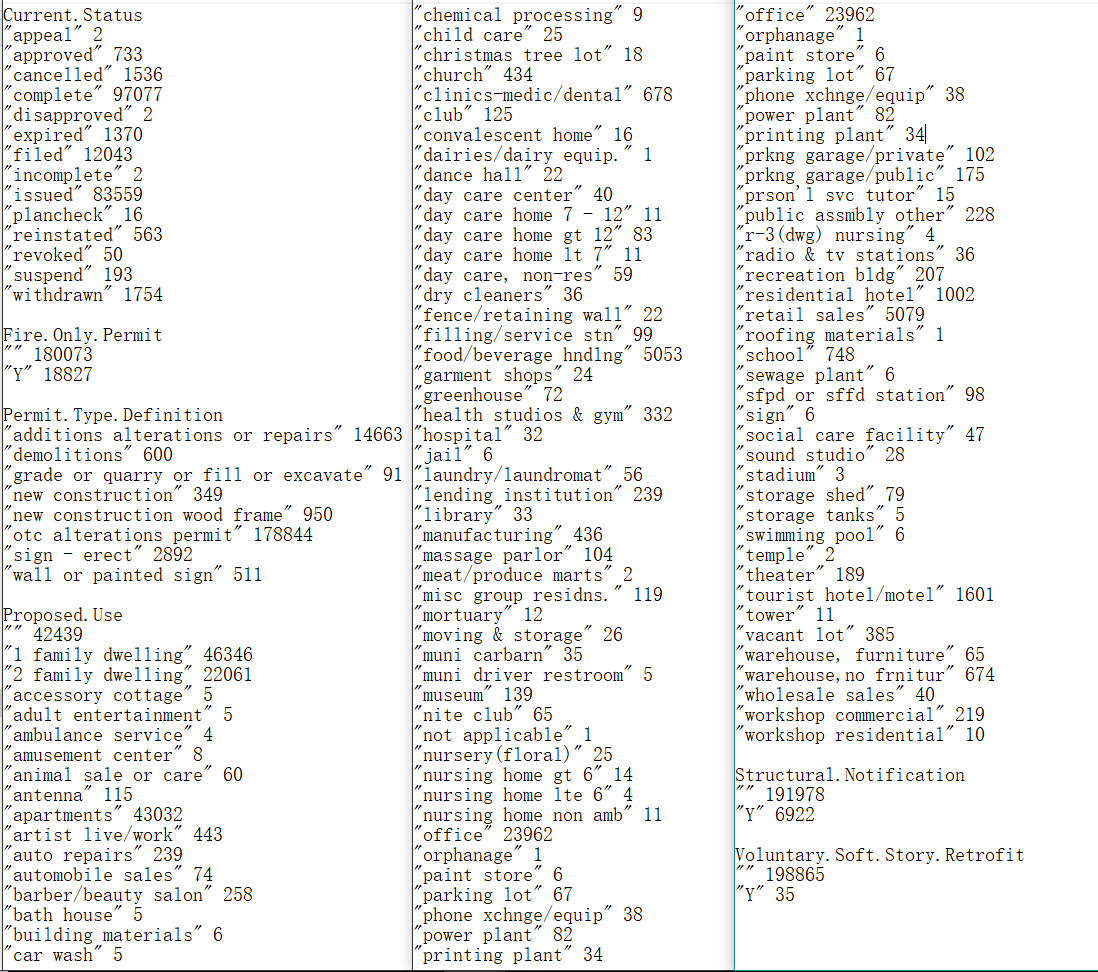
数据摘要：对标称属性给出每个可能取值的频数，对数值属性给出最大、最小、均值、中位数、四分位数及缺失值个数。

程序函数：summary(x, maxsum = );

结果：标称属性结果写入txt文件，数值属性结果整理为excel形式，如下分别为两个数据集的结果。



|  |  |
| --- | --- |
|  | Min. 1st Qu. Median Mean 3rd Qu. Max. NA's |
| yrdln | 1.00 20.00 30.00 28.49 39.00 50.00 840 |
| ydstogo | 0.000 3.000 9.000 7.309 10.000 50.000 |
| ydsnet | -87.00 5.00 19.00 25.95 43.00 99.00 |
| AirYards | -70.000 0.000 0.000 3.264 4.000 84.000 |

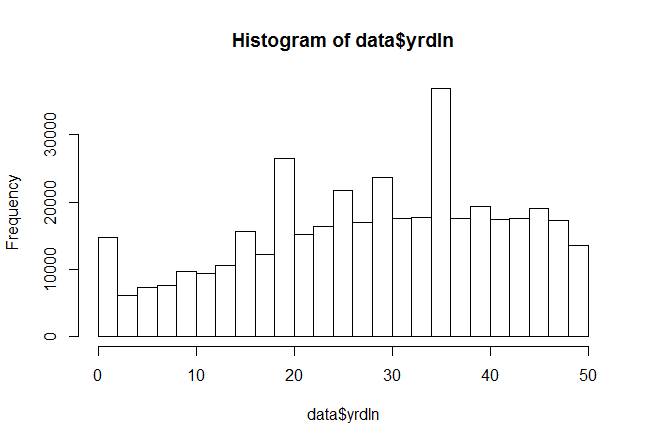


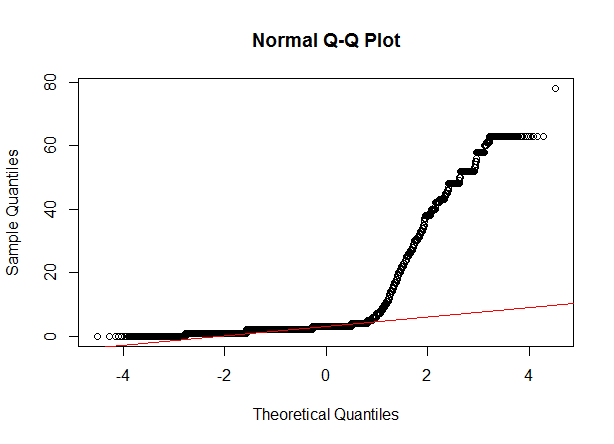
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Number of Existing Stories | Number of Proposed Stories | Estimated Cost | Revised Cost |
| 最大值 | 78 | 78 | 538000000 | 780500000 |
| 最小值 | 0 | 0 | 1 | 0 |
| 均值 | 5.71 | 5.75 | 169000 | 132856 |
| 中位数 | 3 | 3 | 11000 | 7000 |
| 四分位数 | 2 | 2 | 3300 | 1 |
| 4 | 4 | 35000 | 28708 |
| 缺失值  个数 | 42784 | 42868 | 38066 | 6066 |

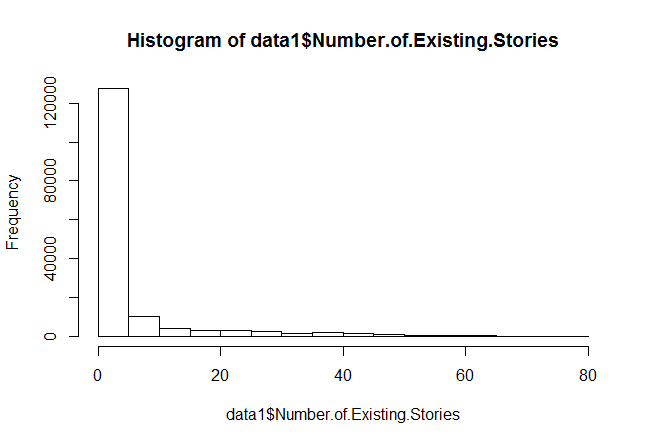
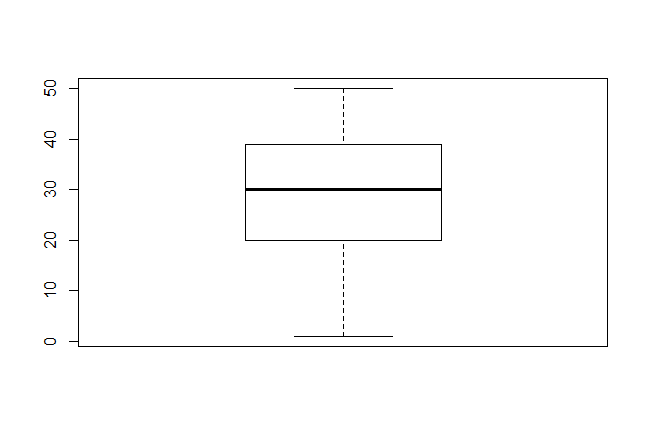
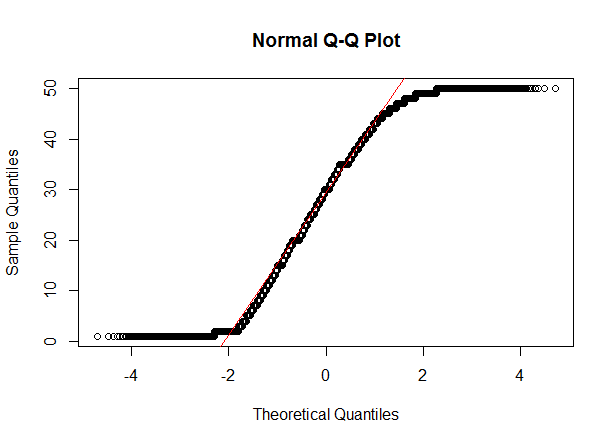
数据的可视化：针对数值属性，绘制直方图并用qq图检验其分布是否为正太分布；绘制盒图对离群值进行识别。

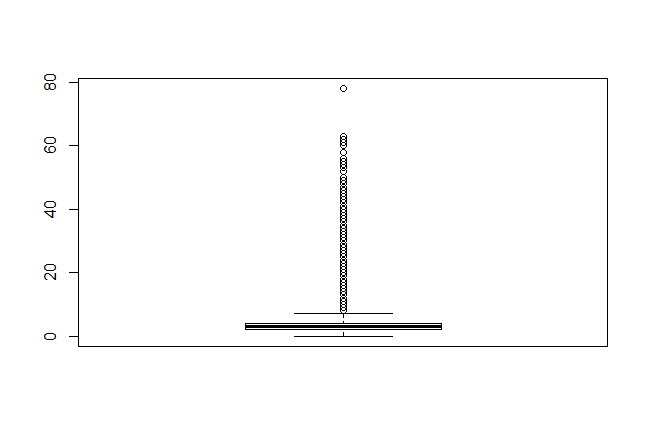
程序函数：hist(x);qqnorm(x);qqline(x,col = );boxplot(x);

结果显示如下（数据集1以属性yrdln为例，数据集2以属性Number of Existing Stories为例，其他结果见作业一文件夹）：





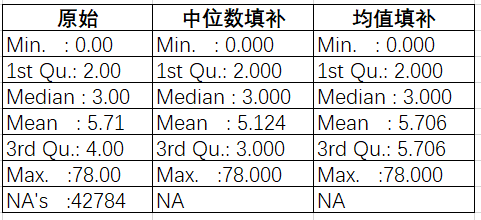


如上述结果所示，直方图能够很直观的反应属性的分布情况，盒图也能将离群值很好地进行识别。直观地从qq图上来看，数据集1的yrdln属性并非来自正态分布的数据，数据集2的Numbers Of Existing Stories属性可能服从正太分布。

1. 数据缺失的处理

将缺失部分剔除，对数据本身无影响，数据分析结果于剔除前一致；

除此之外，分别采取用均值和中位数填补缺失值，数据分析结果如下所示（以数据集2的Number Of Existing Stories属性为例，分别为填补之前、中位数填补和均值填补）：



上图为填补缺失值前后的新旧数据集的可视化对比，可见无论是中位数填补还是均值填补，对最值和四分之一位数没有影响，四分之三位数有相应的改变，根据你填补数据的不同，均值和中位数其中一个会有所改变。