数据挖掘作业一

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。

1.1 数据摘要

（1）标称属性：以Building\_Permits中的Permit Type为例，每个可能取值的频数如下：

|  |  |
| --- | --- |
| 取值 | 频数 |
| 8 | 178844 |
| 3 | 14663 |
| 4 | 2892 |
| 2 | 950 |
| 6 | 600 |
| 7 | 511 |
| 1 | 349 |
| 5 | 91 |

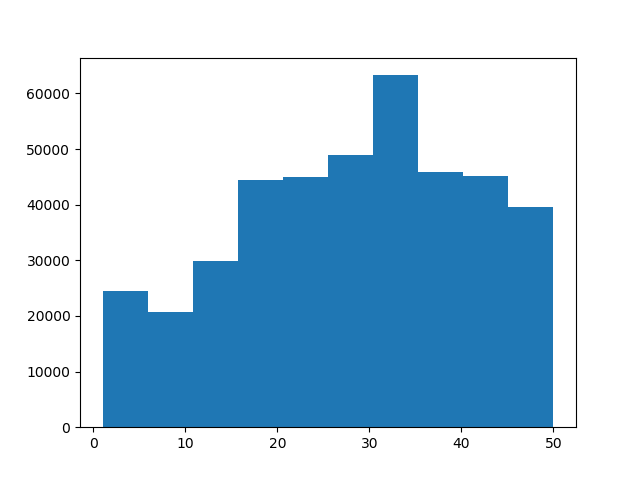
（2）数值属性：以NFL Play by Play 2009-2017 (v4).csv中的yrdln为例

|  |  |
| --- | --- |
| 最大 | 50 |
| 最小 | 1 |
| 均值 | 28.488327 |
| 中位数 | 30 |
| 25% | 20 |
| 75% | 39 |
| 缺失值 | 840 |

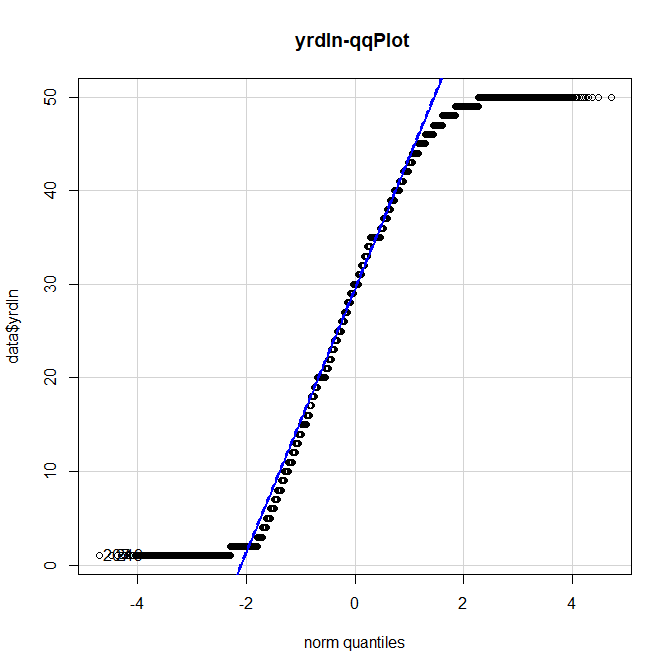
1.2数据可视化

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

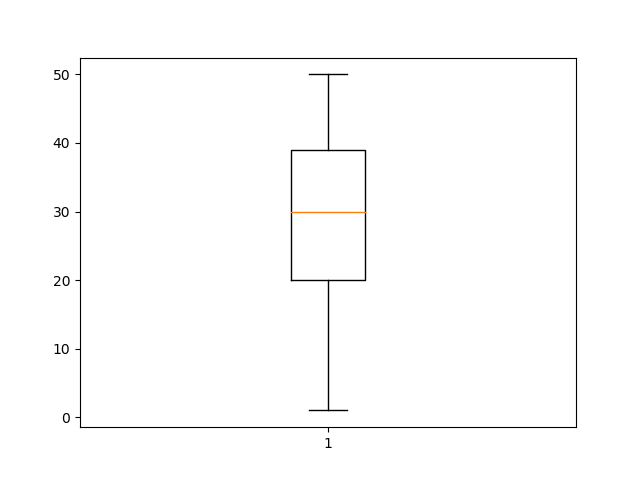
（1）直方图为：



（2）QQ图为：



（3）盒图为：



1. 数据缺失的处理

2.1将缺失部分剔除

2.2用最高频率值来填补缺失值

2.3通过属性的相关关系来填补缺失值

2.4通过数据对象之间的相似性来填补缺失值