# 用电量数据清洗

数据源：15021740588.xls

数据总数：889575 条（处理前） 1033811条（处理后）

数据清洗步骤：

**1.删除时间跨度超过2小时（不含）的数据**

|  |  |  |  |
| --- | --- | --- | --- |
| 0-1小时之间 | 1-2小时之间 | 2-10小时之间 | 超过10小时 |
| 773508 | 108625 | 6346 | 1071 |
| 86.9% | 12.2% | 0.7% | 0.1% |

剔除超过两个小时以上的数据，剔除数据占总数比为0.8%。

**2. 将时间进行规整：**

规整处理方法

A 时间跨度没有超过一个小时的：无需规整

B 时间跨度在两个整点时间段，按照时间跨度比例均匀分配用电量

用电量按照时间长度作为权重分摊到两个时间段内。

算作9点的

算作8点的

10：00

8：00

9：00

需要规整的数据标准：floor(start)-ceiling(end)=2 并且时间持续长度大于1.1小时[[1]](#footnote-1)，小于等于2小时。这部分需要处理的数据条目有101399条，占12.2%。进过上述处理后，数据条目增多至207316条，有部分数据横跨三个时间段。

数据时间段辨识：

经过上述处理，总数据条目增多至1033811条，每条数据时间持续长度小于1.1小时。

开始时间如果在8:50至8:59该段用电量算作9点时段。

10：00

8：00

9：00

算作9点的

由于电表截取时间截点不处于整点，因此经过处理之后，有部分数据在同一个时间段内有两组数据。通过简单的聚集函数，用电量求和。

**select** ID,date,hour,sum(Consumption)as Consumption

**from** mydata

**group by** ID,date,hour

最终总数据条数为837597条。

**2 描述性统计与异常值处理**

2.1电表日期统计：

总共有**177**块电表，每块电表的记录的起止时间不尽相同。

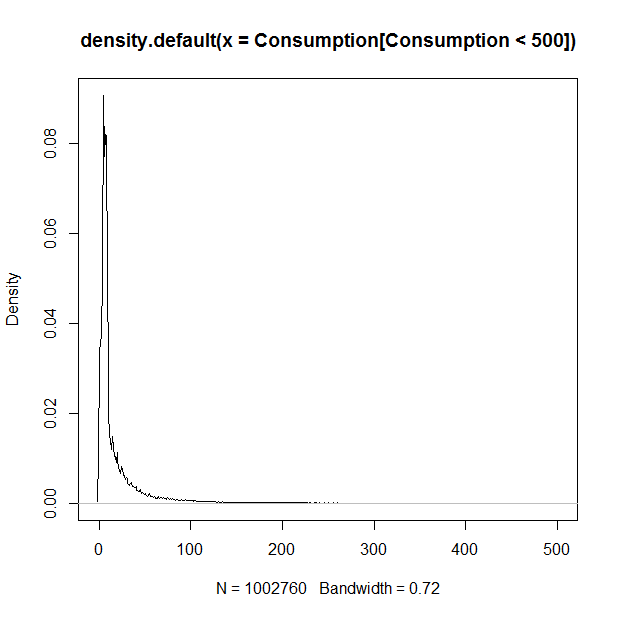
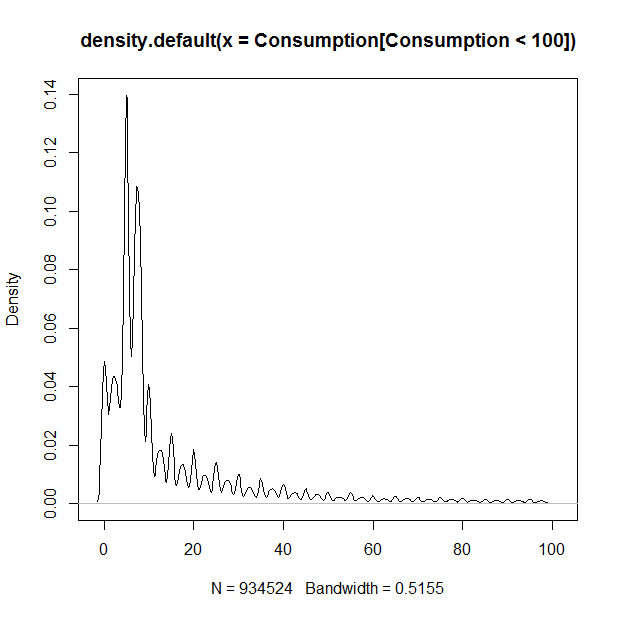
大部分电表的起止时间是从2016年10月30日至2017年6月9日

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Min | 1st Q | Median | Mean | 3rdQ | Max |
| Start | 2016-10-03 | 2016-10-05 | 2016-10-20 | 2016-10-30 | **2016-10-30** | 2017-05-26 |
| End | 2016-10-21 | 2017-06-09 | 2017-06-09 | 2017-05-28 | **2017-06-09** | 2017-06-09 |

有**124**块电表满足在2016-10-30至2017-6-9之间都有完整记录的。（在此暂不处理）

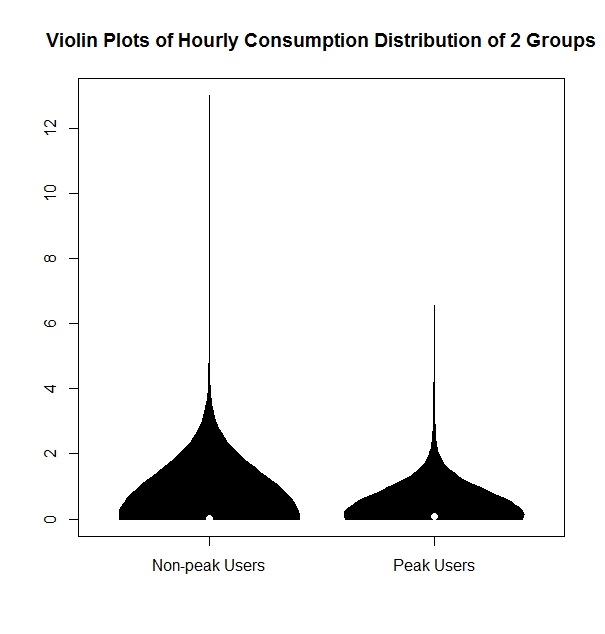
**3.删除用电量异常值**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 负值 | 0-10（不含） | 10-50 | 50-100 | 超过100 | 超过300 | 超过2000 |
| 0 | 578029 | 296274 | 60221 | 51764 | 47116 | 407 |
| 0% | 55.9% | 28.65% | 5.82% | 5.01% | 4.55% | 0.039% |



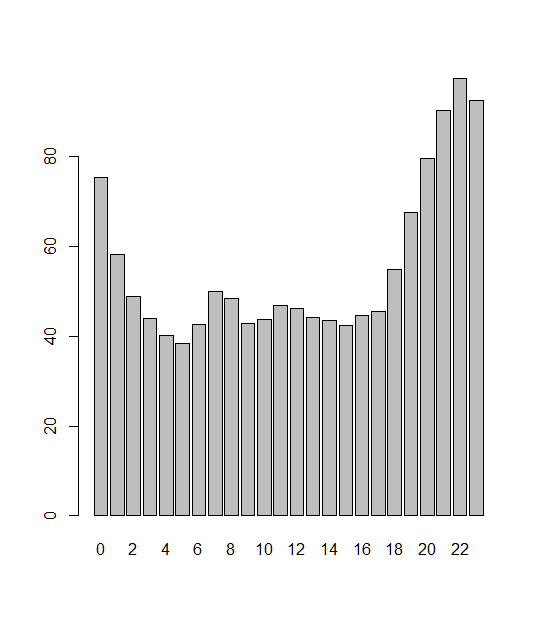
经观测，一小时用电量的核密度图形 似 对数型正态分布（lognormal），最大值为13260。

经查家庭式中央空调10P为24000W 耗电量在10 kWh左右。再考虑到一般家庭用电额定6KW,长时间供电可以负载8.5KW,间歇工作最大可以负载15KW。



**4. 时间段统计**

经观察，用电集中时段为晚上的7：00至次日的凌晨2点。



非峰谷用户

峰谷用户

谷时段，价格便宜

# 天气清洗数据报告

数据源：天气实测数据.csv

数据总数：229705 条

1. 极值处理

原始数据中有13行空行，直接删除。

同时将温度，湿度，降雨量等数据中的9999值改为NA值

处理完后的数据缺失值个数：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tem | High\_tem | Low\_tem | Rain | Wind\_d | Wind\_v | Pressure | Humidity |
| 38 | 无 | 无 | 6 | 9055 | 9055 | 83 | 42 |

1. 时间规整

数据总数为229705条，由于天气数据采集的频率不同，前8863条数据时间间隔为1小时，而后221043条数据的时间间隔为10分钟。

经过处理，按照每小时间隔聚集后的数据量为37197条，加上原先无需处理的，共45859条数据。

每个小时有6个数据点 00分 10分 20分 30分 40分 50分

数据 处理方式：

温度 平均值

最高温度 最大值

最低温度 最小值

降水量 求和

风向 平均

风速 平均

气压 平均

湿度 平均

代码如下：

need\_aggregate$floor\_date<-floor\_date(need\_aggregate$date\_time,unit="hours")//将所需要的数据时间段取整。

aggregate\_complete<-sqldf("select station\_code,

floor\_date as date\_time,

avg(tem) as tem,

max(high\_tem) as high\_tem,

min(low\_tem) as low\_tem,

sum(rain) as rain,

avg(wind\_direction) as wind\_direction,

avg(wind\_velocity)as wind\_velocity,

avg(pressure) as pressure,

avg(humidity) as humidity

from need\_aggregate group by floor\_date"

天气数据与电量数据整合

用电量数据格式

No ID hour Consumption date\_time

1 2719 80000089 7 323 2017-02-13 07:00:00

2 12580 80000088 8 5 2017-02-11 08:00:00

3 12596 80000088 21 7 2017-02-12 21:00:00

4 12604 80000088 7 7 2017-02-13 07:00:00

5 16997 80000068 2 5 2017-01-16 02:00:00

6 26866 80000066 15 5 2017-02-01 15:00:00

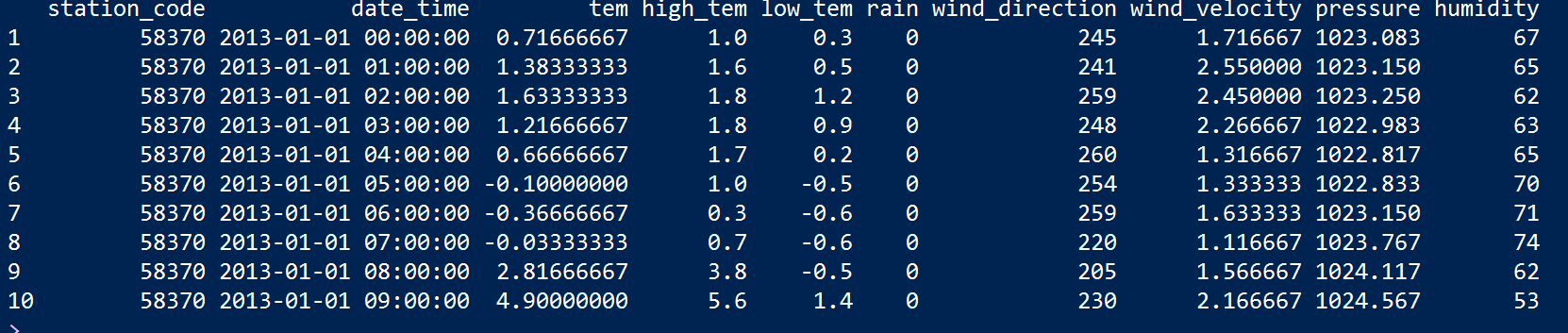
7 29678 80000085 16 20 2016-10-03 16:00:00

8 39830 80000065 16 10 2016-10-28 16:00:00

9 41805 80000065 9 3 2017-01-27 09:00:00

10 46844 80000084 13 25 2017-01-19 13:00:00

天气数据格式



以date\_time 作为两表连接的桥梁值，同时要包含全部的用电量数据。因此采用左外联结的形式。

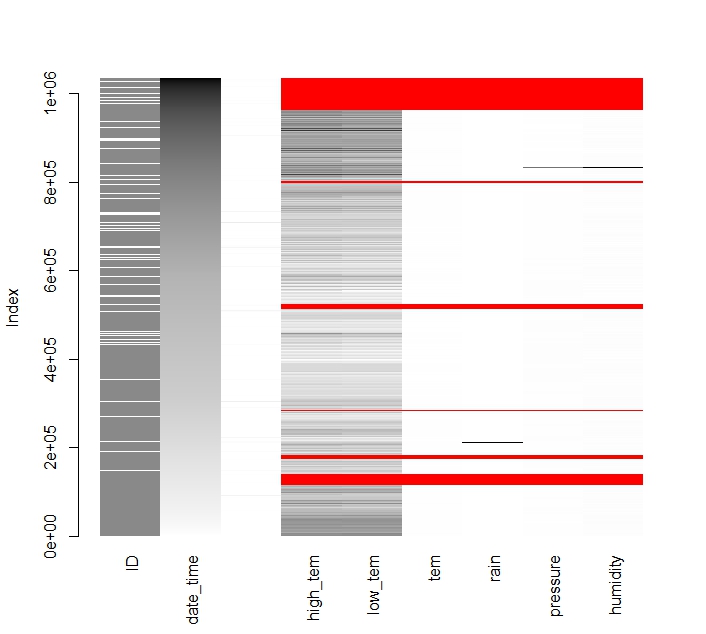
SQL语句：

**select** ID, sampleele.date\_time, Consumption, high\_tem,low\_tem,tem,rain,pressure,humidity **from** sampleele **left outer join** sampleweather

**on** sampleele.date\_time **=** sampleweather.date\_time

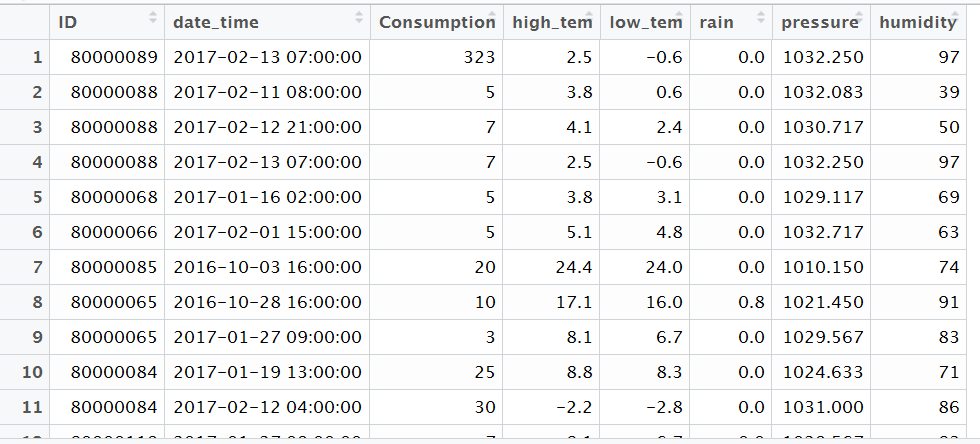
链接后的两表行数 1,033,811 行

其中天气缺失值NA 有124937条 占总数据比为12.08%



图中红色部分为缺失值，每行按照日期从小到大排序，我们可以看到有关天气数据的缺失是扎堆出现，是两张表链接后的结构性数据缺失。对于此部分缺失值，我们除了删除行以外，别无选择。

清洗后样式如下：



1. 有不少数据是7：59-8：59，虽然也横跨两个时间段，但看作是8点时间段处理。而该部分时间总长在1.1小时以内。 [↑](#footnote-ref-1)