数据科学基础 I (Matlab)

东北大学

数据探索性分析案例

东北大学

案例

已知某家庭用户每分钟的电能表记录数据，尝试对数据进行探索性分析。

数据项	说明	数据类型
date	日期	文本
time	时间	文本
global_active_power	全屋有功功率(kW)	数值
global_reactive_power	全屋无功功率(kW)	数值
voltage	供电电压(V)	数值
global_indensity	电流(A)	数值
sub_metering_1	电表1（厨房，含洗碗机、烤箱、微波炉）	数值
sub_metering_2	电表2（洗衣房，含洗衣机、烘干机、冰箱）	数值
sub_metering_3	电表3（含热水器和空调）	数值

data: http://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

案例

数据基本信息及简单处理

- 时间跨度：2006年12月16日-2010年11月26日
- 记录间隔：1分钟
- 有效数据记录数：2075259条
- 简单处理：为简化起见，将同一天的记录合并累加，变为记录间隔为1天的数据集。

案例

2075259x9 table

	1 Date	2 Time	3 Global_ac	4 Global_rea	5 Voltage	6 Global_int	7 Sub_mete	8 Sub_mete	9 Sub_meter
	16/12/2006	17:24:00	4.2160	0.4180	234.8400	18.4000	0	1	17
	16/12/2006	17:25:00	5.3600	0.4360	233.6300	23	0	1	16
	16/12/2006	17:26:00	5.3740	0.4980	233.2900	23	0	2	17
	16/12/2006	17:27:00	5.3880	0.5020	233.7400	23	0	1	17
	16/12/2006	17:28:00	3.6660	0.5280	235.6800	15.8000	0	1	17
	16/12/2006	17:29:00	3.5200	0.5220	235.0200	15	0	2	17
	16/12/2006	17:30:00	3.7020	0.5200	235.0900	15.8000	0	1	17
	16/12/2006	17:31:00	3.7000	0.5200	235.2200	15.8000	0	1	17
	16/12/2006	17:32:00	3.6680	0.5100	233.9900	15.8000	0	1	17
	16/12/2006	17:33:00	3.6620	0.5100	233.8600	15.8000	0	2	16

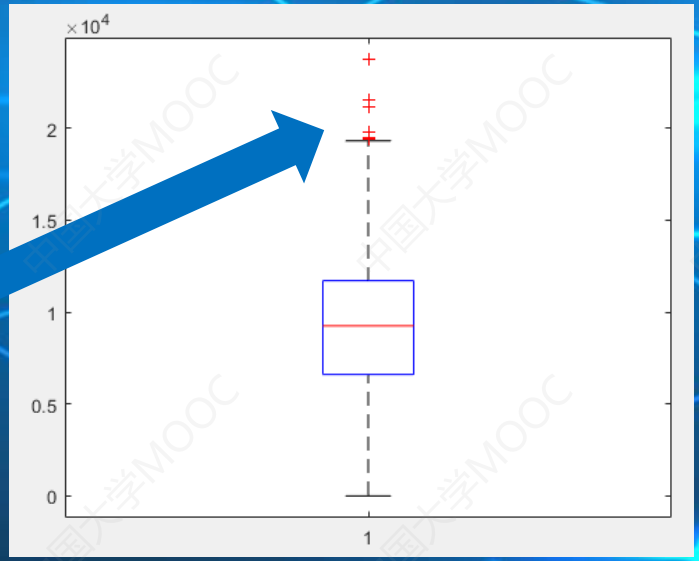
1440x9 table

	1 dt	2 cnt	3 act_power	4 react_powe	5 Voltage	6 intensity	7 m1	8 m2	9 m3
1	2006-12-17 00:00:00	1440	3.3905e+03	226.0060	3.4573e+05	1.4399e+04	2033	4187	13341
2	2006-12-18 00:00:00	1440	2.2038e+03	161.7920	3.4737e+05	9.2472e+03	1063	2621	14018
3	2006-12-19 00:00:00	1440	1.6662e+03	150.9420	3.4848e+05	7.0940e+03	839	7602	6197
4	2006-12-20 00:00:00	1440	2.2257e+03	160.9980	3.4892e+05	9.3130e+03	0	2648	14063
5	2006-12-21 00:00:00	1440	1.7166e+03	144.1660	3.4662e+05	7.2386e+03	1765	2623	10421
6	2006-12-22 00:00:00	1440	2.3413e+03	186.9060	3.4731e+05	9.8970e+03	3151	350	11131
7	2006-12-23 00:00:00	1440	4.7734e+03	221.4700	3.4580e+05	2.0200e+04	2669	425	14726
8	2006-12-24 00:00:00	1440	2.5500e+03	149.9000	3.4803e+05	1.1002e+04	1703	5082	6891

案例

利用boxplot分析集中趋势和离中趋势（以m3电能表为例）

```
>> boxplot(gt.m3)
```

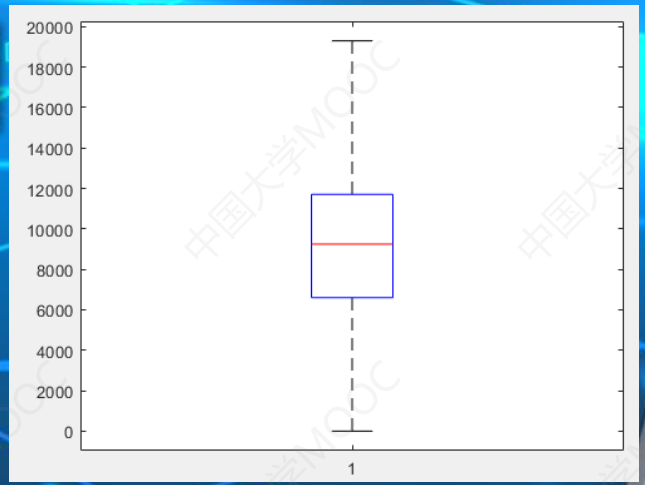


少量读数过大的离群点

案例

去除异常（高于上界的取上界）

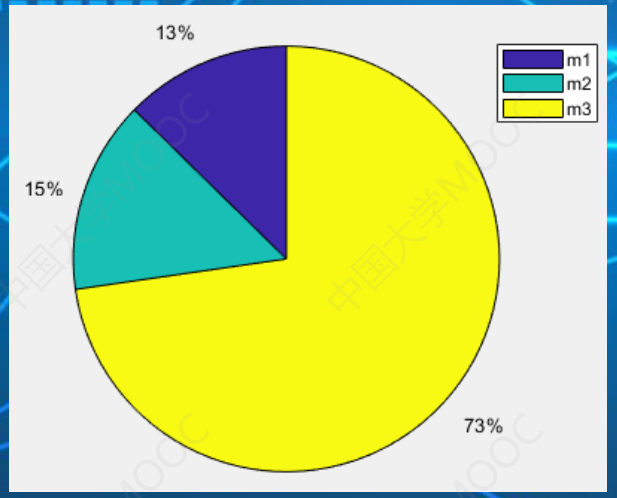
```
>>gt.m3(find(gt.m3>19300))=19300;
```



其他电能表的异常值，同样方法处理

案例

以饼图分析三个电能表的占比

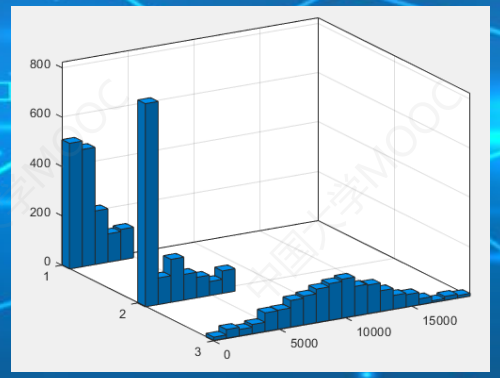



```
>> pie([sum(gt.m1) sum(gt.m2) sum(gt.m3)])
```

案例

按电能表编号分组，以三维直方图分析分布特征

- 可以发现，m1和m2表读数分布在低值段，m3表分布在高段




```
>> [m n]=size(gt.m1)
>> mt=[ones(m,n) gt.m1; 2.*ones(m,n) gt.m2; 3.*ones(m,n) gt.m3]
>> histogram2(m(:,1),m(:,2))
```

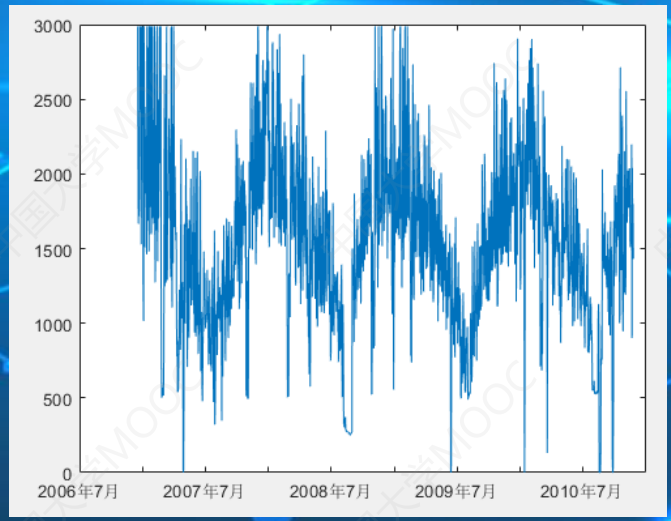
案例

探索有功功率是否存在周期性变化

- 用电存在明显的年度周期，
  每年7-8月份用电低谷，
  每年12-1月份用电处于峰值

思考：可能原因？

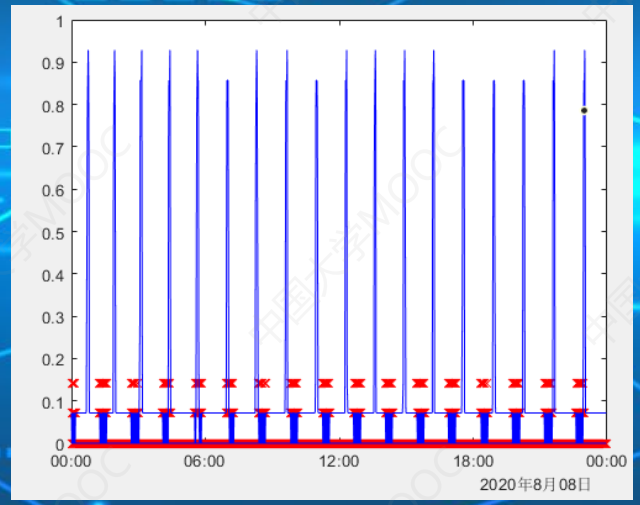
```
>> plot(gt.dt,gt.act_power)
```



案例

- 选取一天对比分析m2和m3电能表在总电能表和中的比值变化情况

```
>> plot(gt2.tm, gt2.m2/gt2.msum,’rx’)
>> hold on
>> plot(gt2.tm, gt2.m3/gt2.msum,’b’)
```



案例

探索家庭用电数据中有功功率与其他指标的相关性

```
>> for i=2:1:6
>> cor = corrcoef(data(:,i),data(:,1));
>> corr(i) = cor(1,2);
>> disp(['corrcoef of active power and data index ' num2str(i) ‘ is ' num2str(corr(i))]);
>> end
>>[m,index]=max(corr);
>> disp(['the max corr is ' num2str(index) ' ' num2str(m)]);
```

案例

探索家庭用电数据中有功功率与其他指标的相关性

```
corrcoef of active power and data index 2 is 0.12778
corrcoef of active power and data index 3 is 0.99922
corrcoef of active power and data index 4 is 0.54903
corrcoef of active power and data index 5 is 0.48506
corrcoef of active power and data index 6 is 0.75121
the max corr is 3 0.99922
```

思考：通过进一步分析，你能为该家庭提供一个以节能为优化目标的家电升级计划吗?