17-基于数据挖掘的电力负荷预测 中期报告

成员：蔡静轩(3220180783), 王峤(3220180867)

此中期报告主要记录在项目的数据预处理过程中遇到的问题及处理方法，描述了数据探索和清洗过程。考虑到调试及可视化等问题，首先利用jupyter notebook进行处理，最后整理保存数据预处理代码。

数据预处理流程分如下几个阶段：

一、 数据探索

二、 分析数据构成

三、 查看数据描述

四、 删除无效列

五、 拆分时间数据

六、 填补缺失数据

七、 分析负荷曲线

八、 查看数据分布

九、 增加特征

十、 填补缺失值

十一、 处理离群点

一、数据探索阶段

首先，我们尝试了区分电机、区分消纳类型这两种方式，分别对不同的电机或不同的消纳类型以半小时或一天为刻度建模，但是效果都很不理想，并且时间损耗极大。后来将 6 台消纳的负荷合并至每半小时只有一行数据或每天只有一行数据，重新进行数据分析和建模过程。

	区分 6 台电机	区分 3 类消纳	合并所有消纳
半小时为刻度	不理想	不理想	不理想
一天为刻度	不理想	不理想	

二、分析数据构成

数据格式如下图所示，其中compsid，Expertsid，areaenergysid列为空，后续将删除这三列。

```
In [27]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200241 entries, 0 to 200240
Data columns (total 13 columns):
RecNo           200241 non-null int64
Name            200241 non-null object
Data_Time       200241 non-null object
Value           200241 non-null float64
Type            200241 non-null int64
Equsid          200241 non-null int64
Energysid       200241 non-null int64
Expertsid       0 non-null float64
compsid         0 non-null float64
posisid         200241 non-null int64
buildingsid     200241 non-null int64
Load_Time       200241 non-null object
areaenergysid   0 non-null float64
dtypes: float64(4), int64(6), object(3)
memory usage: 19.9+ MB
```

三、查看数据描述

此处注意到Value的最大值居然高达 35W，疑似异常数据；同时数据的总行数为 200241 行，简单计算可知若每半个小时收集 6 行数据，理论上应有 210528 行数据，因此数据集应该存在缺失行数据的现象。Type和Energysid列的标准差为 0，说明无二值，也应删除。

```
In [24]: data.describe()
```

Out[24]:

	RecNo	Value	Type	Equsid	Energysid	Expertsid	compsid	posisid	buildingsid	areaenergysid
count	2.002410e+05	200241.000000	200241.0	2.002410e+05	200241.0	0.0	0.0	200241.000000	200241.000000	0.0
mean	6.234372e+08	51.047068	1.0	1.503701e+11	32.0	NaN	NaN	4700.505890	1503.700506	NaN
std	1.929797e+08	809.827533	0.0	9.541004e+07	0.0	NaN	NaN	954.100532	0.954101	NaN
min	3.736157e+08	0.000000	1.0	1.503000e+11	32.0	NaN	NaN	4000.000000	1503.000000	NaN
25%	4.273284e+08	0.000000	1.0	1.503000e+11	32.0	NaN	NaN	4000.000000	1503.000000	NaN
50%	6.212504e+08	0.000000	1.0	1.503000e+11	32.0	NaN	NaN	4000.000000	1503.000000	NaN
75%	8.085939e+08	42.000000	1.0	1.505000e+11	32.0	NaN	NaN	6000.000000	1505.000000	NaN
max	9.470235e+08	358825.984000	1.0	1.505000e+11	32.0	NaN	NaN	6000.000000	1505.000000	NaN

从排序结果可以肯定，35W为异常数据（2016.12 月）。

```
In [25]: data.sort_values(by='Value', ascending=False)
```

Out[25]:

	RecNo	Name	Data_Time	Value	Type	Equsid	Energysid	Expertsid	compsid	posisid	buildingsid	Load_Time	areaer
92325	591866171	JF_2ATC_SC701109_EC	9/12/2016 03:30:00	358825.984	1	150300000109	32	NaN	NaN	4000	1503	9/12/2016 04:22:04	
148880	797960682	JF_2AES_SC100481_EC	1/7/2017 14:00:00	4532.200	1	150500000081	32	NaN	NaN	6000	1505	4/7/2017 10:17:50	
135222	748556316	JF_2ATC_SC701109_EC	13/5/2017 07:00:00	1332.992	1	150300000109	32	NaN	NaN	4000	1503	13/5/2017 07:37:06	
128076	723519801	JF_2ATC_SC701109_EC	18/4/2017 11:30:00	1241.088	1	150300000109	32	NaN	NaN	4000	1503	18/4/2017 12:15:59	
132936	740575719	JF_2ATC_SC701109_EC	5/5/2017 08:30:00	1239.296	1	150300000109	32	NaN	NaN	4000	1503	5/5/2017 09:09:35	
103390	632639366	JF_2ATC_SC701109_EC	19/1/2017 10:30:00	1230.336	1	150300000109	32	NaN	NaN	4000	1503	19/1/2017 11:08:43	
132948	740617833	JF_2ATC_SC701109_EC	5/5/2017 09:30:00	1226.752	1	150300000109	32	NaN	NaN	4000	1503	5/5/2017 10:10:01	

四、删除无效列

从上述结果来看，RecNo列为标记类型数据，Name为电机标记，posisid与buildingsid为建筑标记，均为无效数据。只需要保留Equsid列即可。Load_Time为生成数据的时间，对预测Value没有帮助，也可以删除。Equsid列数值过大，将其重新标记为 1 到 6，方便后续处理。

```
In [32]: print(len(set(data.RecNo)))
         print(set(data.Name))
         print(set(data.Type))
         print(set(data.Energysid))
         print(set(data.posisid))
         print(set(data.buildingsid))
         print(set(data.Equsid))
```

```
200241
{'JF_2AES_SC100481_EC', 'JF_2AES_SC100478_EC', 'JF_2ATC_SC701110_EC', 'JF_2ATC_SC701112_EC', 'JF_2ATC_SC701109_EC', 'JF_2ATC_SC701111_EC'}
{1}
{32}
{4000, 6000}
{1505, 1503}
{150300000109, 150500000078, 150300000110, 150300000112, 150300000111, 150500000081}
```

代码文件：createDataSet.py

五、拆分时间数据

Data_Time为时间序列，间隔为半小时，将其拆解为Year, Month, Day, Hour, Half共 5 列，其中Half标记是否为半小时，整点为 0，半点为 1。

拆分时间数据之后，数据文件更新为DataSet1.0.csv。

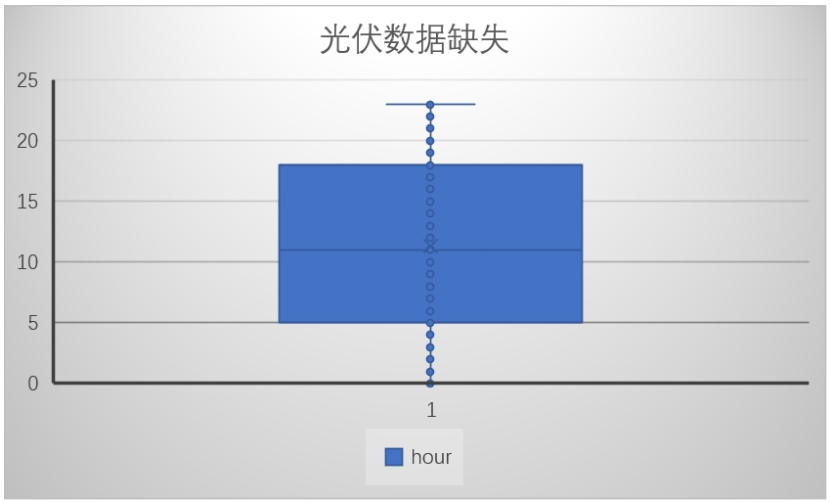
代码文件：createDataSet.py

六、填补缺失数据

利用dataAnalysis.py文件中的detect_missing_value()函数收集缺失的数据，保存在lose_data.csv文件中。结果显示共有 125 天缺失数据，有 7 个小时是完全没有数据的（2016.5.1 20:30~23:30，2017.5.14 14:00~15:00， 2017.10.23 00:00~03:00），有 3385 个半点总共 10203 行数据缺失。

将数据扩充为标准的 210528 行，每半个小时均有 6 行数据，所有补齐的Value列暂时空缺。这个过程中注意到，数据集存在连续几个小时都缺失数据的现象，疑似停电或检修。缺失数据中大部分为某半点时刻光伏数据缺失，通常情况为 3 台机组同时缺失数据，缺失的时间点几乎全天各个时段都有。

填补缺失数据之后数据文件更新为DataSet2.0.csv。



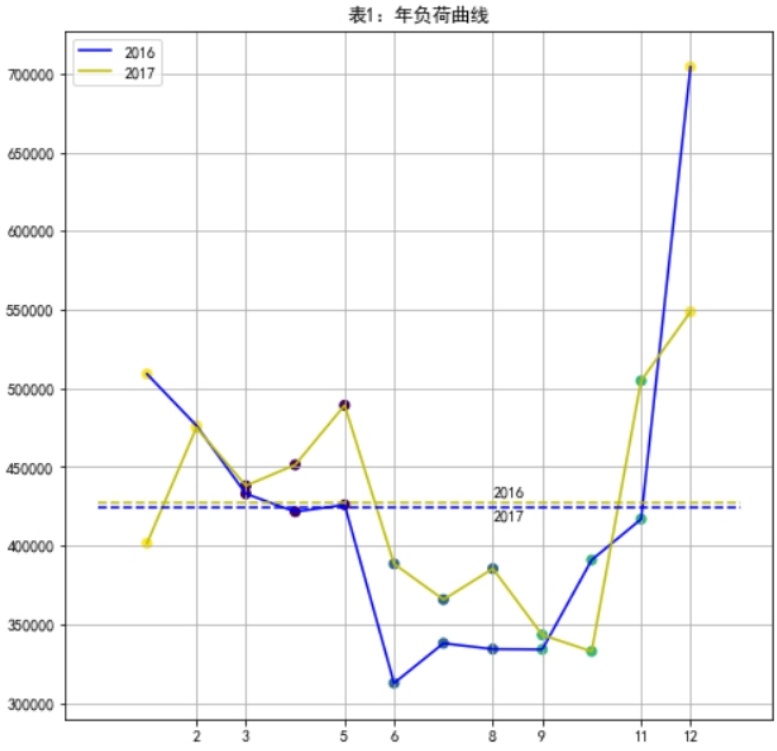
代码文件：fillRowData.py

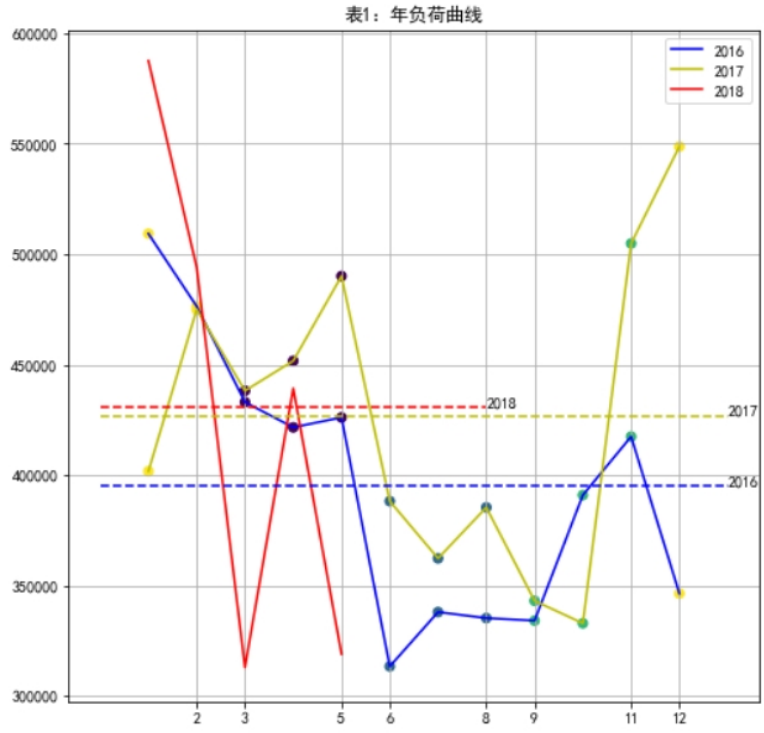
七、分析负荷曲线

对DataSet2.0.csv分析负荷曲线，负荷统计为 6 台消纳的总数据。

1）年负荷曲线

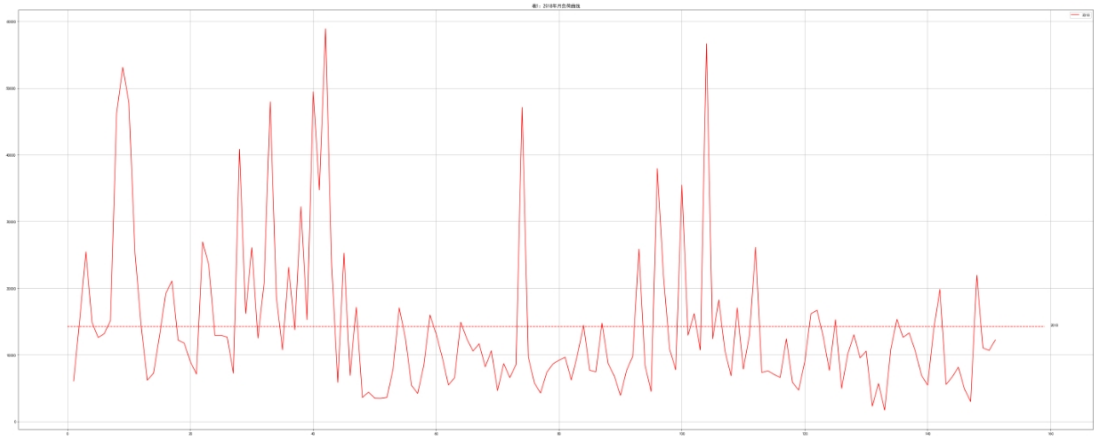
暂不考虑 12 月的异常数据，负荷曲线没有明显的年周期性。第一张表未处理任何异常值，第二张表处理了异常值并加入了 18 年数据。夏季负荷为全年最低，冬季负荷为全年最高。2 年中 6~9 月的负荷都明显低于全年均值。





2）月负荷曲线

红线为后续添加的 2018.1~5 月用电量均值。这里将 12 月的异常值 35W数据归 0，月负荷曲线并未显示出明显的周期性。



3）日负荷曲线

日负荷曲线较为复杂，但是也没有表现出明显的周期性。数据中并未给出检修/大功率作业的时间标记。

4）节假日负荷曲线

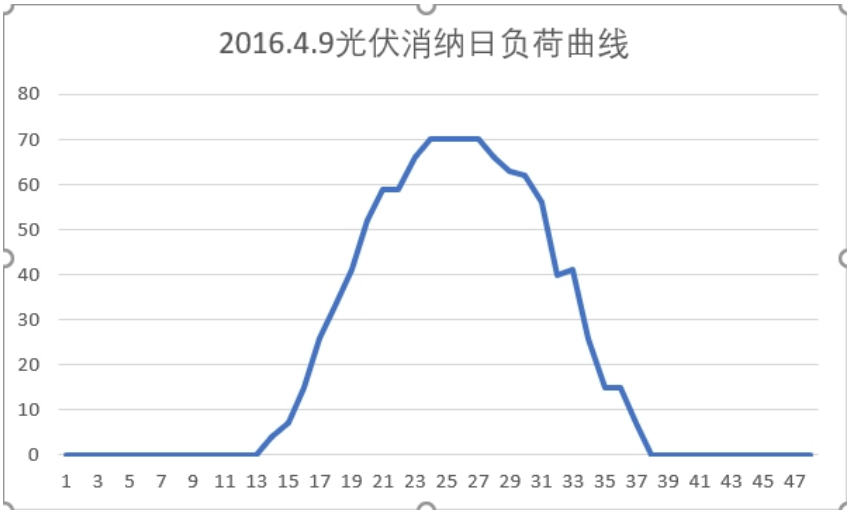
节假日负荷曲线普遍偏低，但是也有一些异常点如下。下图中的几个点的负荷值是明显高于周围几个点的。

2016	2017	2018
初一、初六	除夕、初二、初五	初一
清明1	清明2	清明2、清明3
劳动1、劳动3	劳动1	劳动3
端午1	端午1、端午2	
国庆3、国庆4,、国庆7	国庆1、国庆4	
	元旦1	

5）结论

原始数据中，负荷没有表现出明显的时间规律。时间序列法无法使用。考虑根据不同的消纳类型绘制负荷曲线，观察是否有周期性。

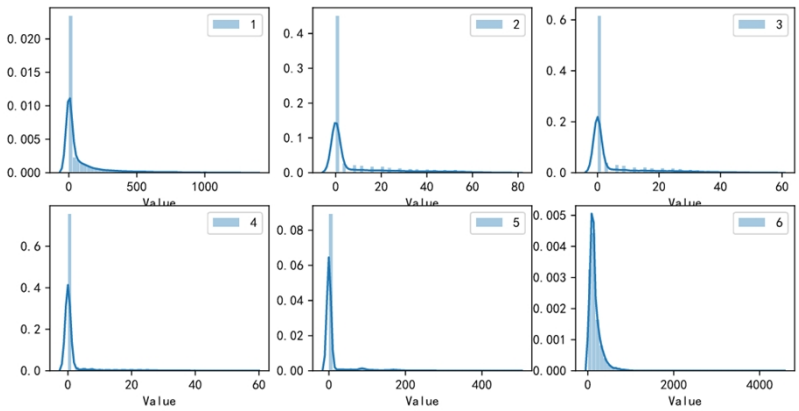
通过 3 种不同的消纳类型的负荷曲线可知，光伏消纳有明显的日周期性，全年每天的 0~5 点，19~24 点光伏消纳皆为 0，日负荷曲线呈现明显的二次函数形状。但是单独预测光伏消纳意义不大，其用电量只占总体的极小比例。



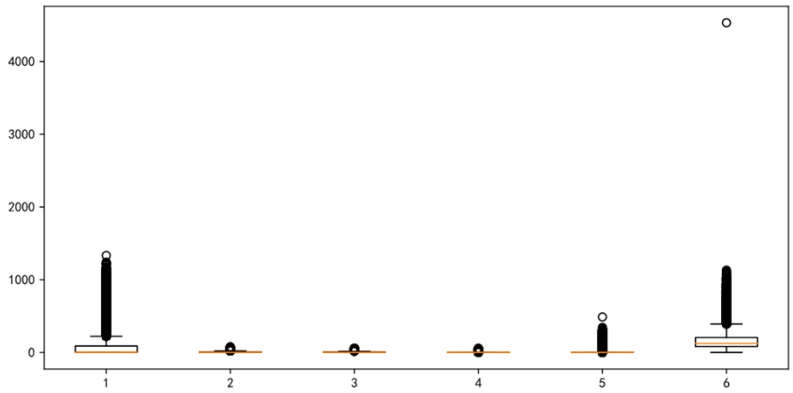
代码文件：dataAnalysis.py

八、查看数据分布

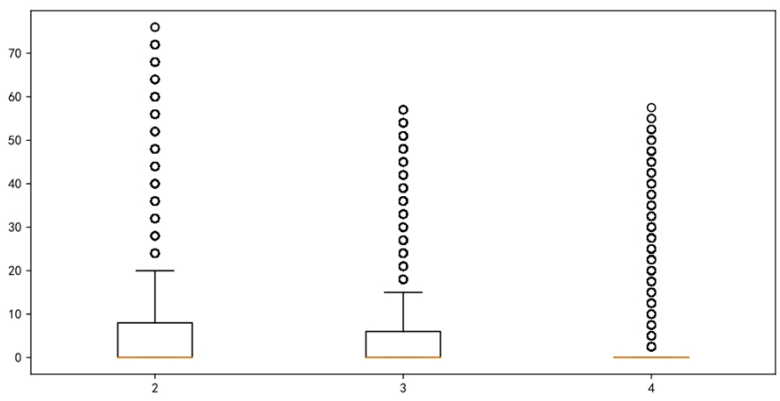
6 台消纳的数据分布密度曲线如下所示，统计了 2 年内以半小时为间隔的消纳数据。可以看出 6 台消纳的值基本集中在某个范围之内。



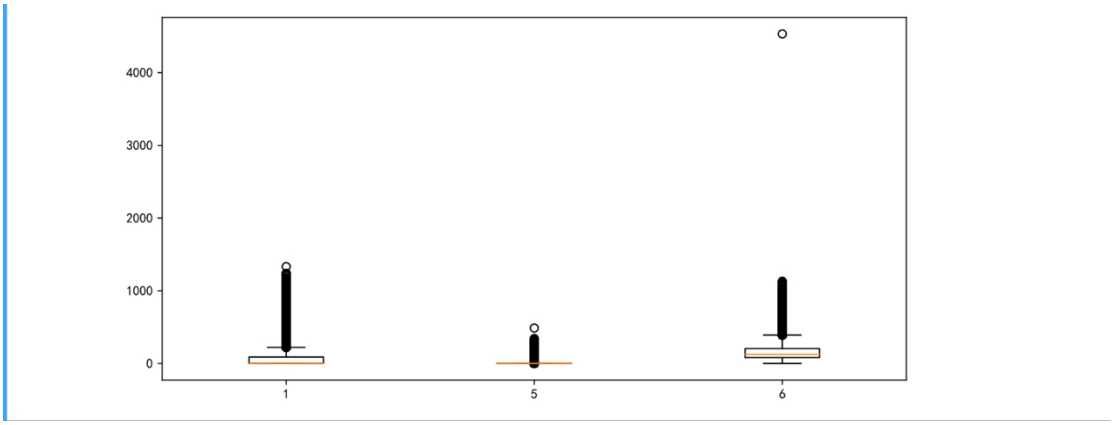
箱型图显示 5 号和 6 号大电网消纳有明显的 2 个离群点。



3 台光伏消纳也有一些离群点。



风力消纳和大电网消纳的离群点较多。



结论：数据中存在较少的离群点，考虑到无冲击负荷且负荷的变化应较为平稳，因此可以采用平滑处理等方式校正离群点。

九、增加特征

将气候特征（降水量、湿度）、日期特征加入分析，日期特征为星期几，是否为工作日，是否为节假日，季节。注意到气候信息中有相当大一部分半点时刻为空值，用前后时间点的平均值填充。在DataSet3.0.csv中添加了气候特征，并在DataSet4.0.csv中填补了气候数据的缺失值，加入了日期特征。

根据任务描述，负荷曲线与其之前时间的负荷可能存在某些关系，因此在按半小时预测的数据集中可以加入同一台消纳半小时之前的负荷、同一台消纳同一时间前一天的负荷、同一台消纳同一时间前一周的负荷这 3 列特征；在按天预测的数据集中可以加入前一天、前两天、前三天、前一周的负荷这 4 列特征。

代码文件：DataAnalysis.py

1）相关性分析

这里运用了pearson相关系数、spearman秩相关、Kendall Tau相关系数。下表中F_half、F_day、F_week分别表示同一台消纳半小时之前的负荷、同一台消纳同一时间前一天的负荷、同一台消纳同一时间前一周的负荷。

pearson correlation coefficient							结论
Equsid	F_half	F_day	F_week	rh_10	t_10		1、光伏消纳有明显的时序关系.
1	强+	弱+	弱+	弱-	弱-		2、风力和大电网消纳也有一定的时序关系.
2	极强+	强+	中+	中-	弱+		
3	极强+	强+	中+	中-	弱+		
4	极强+	强+	中+	弱-	弱+		
5	强+	中+	中+	弱-	弱+		
6	强+	中+	中+	弱-	——		
spearman correlation coefficient							结论
Equsid	F_half	F_day	F_week	rh_10	t_10		1、光伏消纳有明显的时序关系.
1	强+	弱+	弱+	弱-	——		2、风力和大电网消纳也有一定的时序关系.
2	极强+	强+	中+	中-	弱+		
3	极强+	强+	中+	中-	弱+		
4	极强+	强+	中+	弱-	弱+		
5	强+	中+	中+	弱-	弱+		
6	强+	中+	中+	弱-	弱+		
kendall correlation coefficient							结论
Equsid	dayOfWee	isWorkday	isHoliday	Season	Hour		1、因季节的标签十分均匀, 因此系数其实体现不出季节的影响.
1	——	——	——	——	弱-		2、周末与平时的大电网用电有微弱区别.
2	——	——	——	——	——		3、节假日主要影响大电网供电.
3	——	——	——	——	——		
4	——	——	——	——	——		
5	弱-	弱+	弱-	——	弱+		
6	弱-	弱+	弱-	——	弱-		

相关性	负	正
无	-0.09~0.0	0.0~0.09
弱	-0.3~-0.1	0.1~0.3
中	-0.5~-0.3	0.3~0.5
强	-0.9~-0.5	0.5~0.9
极强	-1.0~-0.9	0.9~0.1

下图中Value为日负荷。

```
### pearson correlation coefficient of Equsid all ###
          Value  Tem max  Tem min  RH max  RH min
Value      1.00    -0.22    -0.23   -0.23   -0.24
Tem_max   -0.22     1.00     0.95    0.08    0.02
Tem_min   -0.23     0.95     1.00    0.23    0.27
RH_max    -0.23     0.08     0.23    1.00    0.73
RH_min    -0.24     0.02     0.27    0.73    1.00
### spearman correlation coefficient of Equsid all ###
          Value  Tem_max  Tem_min  RH_max  RH_min
Value      1.00    -0.05    -0.09   -0.25   -0.28
Tem_max   -0.05     1.00     0.93    0.13   -0.03
Tem_min   -0.09     0.93     1.00    0.30    0.22
RH_max    -0.25     0.13     0.30    1.00    0.76
RH_min    -0.28    -0.03     0.22    0.76    1.00
### kendall correlation coefficient of Equsid all ###
            Value   Day  dayOfWeek  isWorkday  isHoliday
Value        1.00  0.03      -0.20       0.32      -0.20
Day          0.03  1.00      -0.00       0.00      -0.12
dayOfWeek   -0.20 -0.00       1.00      -0.69       0.05
isWorkday    0.32  0.00      -0.69       1.00      -0.09
isHoliday   -0.20 -0.12       0.05      -0.09       1.00
Season      -0.05 -0.01       0.01      -0.01       0.08
```

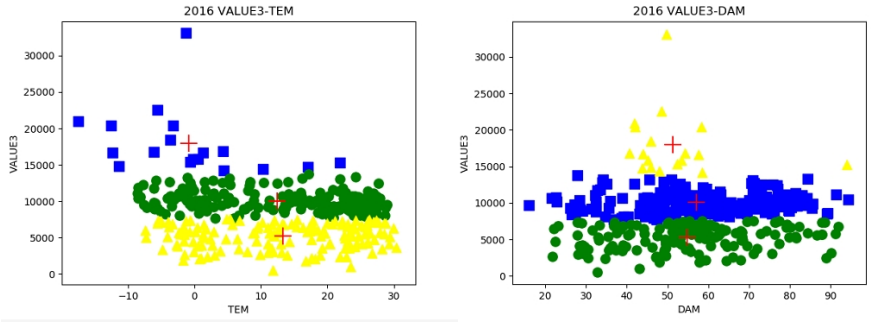
相关性分析的结果显示，光伏消纳有明显的时序关系，用电量与气候弱相关，日用电量在周末与节假日时较低。

我们也对特征之间的相关性进行了分析，结果显示除了 3 列负荷特征之外，负荷的各个特征之间都是几乎没有相关性的。

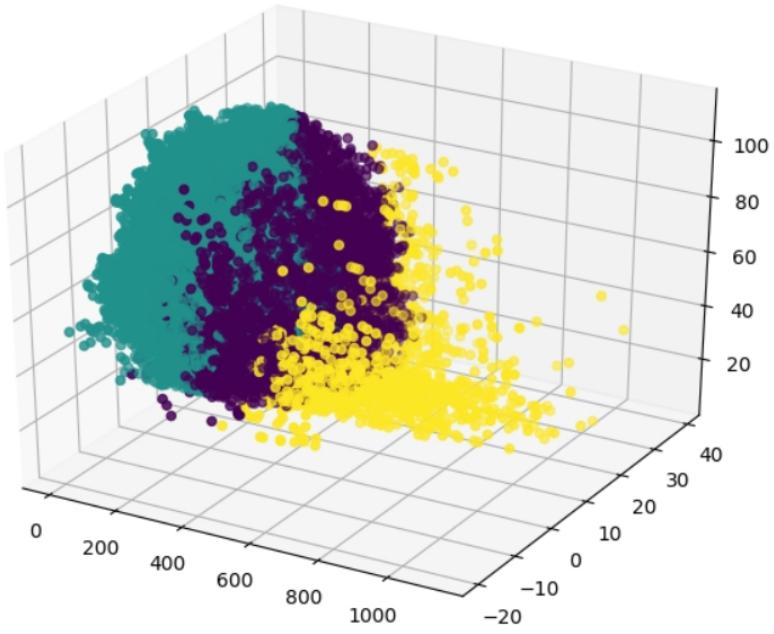
单纯的时间序列用不到特征，所以不予考虑，特征列中可以尝试增加上述 3 列特征，但是这样在预测的时候只能按样本逐个预测，无法并行计算，时间损耗极大。

2）聚类分析

考虑到负荷与气候特征可能存在某种聚类关系，聚类分析的结果如下所示。TEM表示温度，DAM表示湿度。



聚类结果显示，负荷Value与气候特征之间几乎没有聚类关系。如果考虑 3 维空间，聚类结果如下所示。结果表明同样负荷Value与气候特征之间几乎没有聚类关系。



十、填补缺失值

DataSet2.0 版本有非常多的缺失值需要填充，观察lose_data.csv文件缺失数据分布可知，大电网为主要负荷，辅以风力消纳与光伏消纳。消纳顺序大致为 6 5 1 2 3 4。这与任务描述的 3 种消纳类型不区分负荷类型供电，大电网为主要供电方式契合。且当大电网或风电数据缺失时，通常为连续数小数的缺失，疑似

为停电检修。因此我们检测半小时数据，发现一条基本规律，当 6 的Value为 0 时， 12345 必为 0. 当 5 为 0 时， 1234 必为 0. 当 1 为 0 时， 234 必为 0。DataSet5.0.csv完成了填补缺失值的工作。

因此填补策略为：

第一步, 156 全部填 0, 234 每天的 0~6 点、19~24 点全部填 0。

第二步, 234 剩下时段采用平滑处理。

测试数据集（2018 年）所有缺失值均填 0。

代码文件：fillLoseData.py

十一、处理离群点

离群点的 2 种类型——波峰异常、冲击负荷。DataSet6.0.csv完成了处理离群点的操作。

代码文件：handleOutliers.py

1）波峰异常

波峰异常的判定依据如下, 超过阈值则改为阈值。这里其实总共只修改了不到 5 行数据的值。

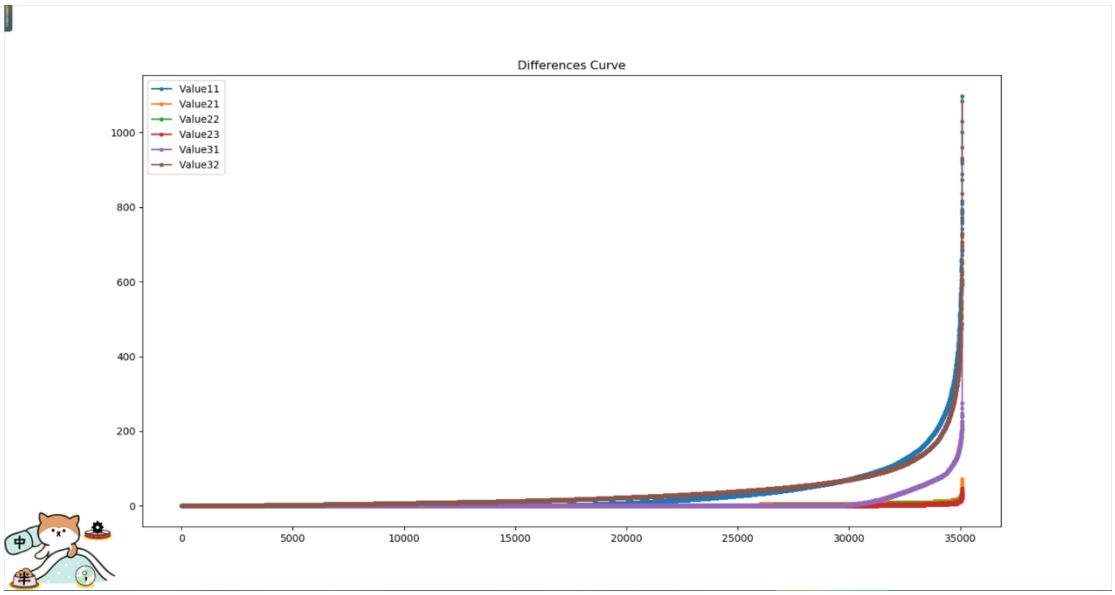
---消纳 1 > 1333

---消纳 2 > 78　　消纳 3 > 78　　消纳 4 > 78

---消纳 5 > 444　　消纳 6 > 1133

2）冲击负荷

相同消纳前后半小时负荷值之差应该处于一个合理的范围之内。6 台电机的前后时刻误差曲线如下所示。



我们取了 0.05%的差值结点，以此为阈值校正负荷值。

冲击负荷的判定依据如下, 超过阈值则调整为相邻值加/减阈值。

---消纳 1 前后差值 ＞875.5

---消纳 2 前后差值 ＞32　　消纳 3 前后差值 ＞32　　消纳 4 后差值 ＞32

---消纳 5 前后差值 ＞269.5　　消纳 6 前后差值 >1059.5

最终数据集

上述所有操作执行完毕后，重新添加F_half、F_day、F_week三列特征，生成了可用于训练和测试的DataSet7.0.csv文件。

对 2018 年的测试数据集做同样的处理，不同的是F_half、F_day、F_week三列设为空值，填补缺失值的方式为全部填充 0 值，且不处理任何离群点。

代码文件：addFeatures.py