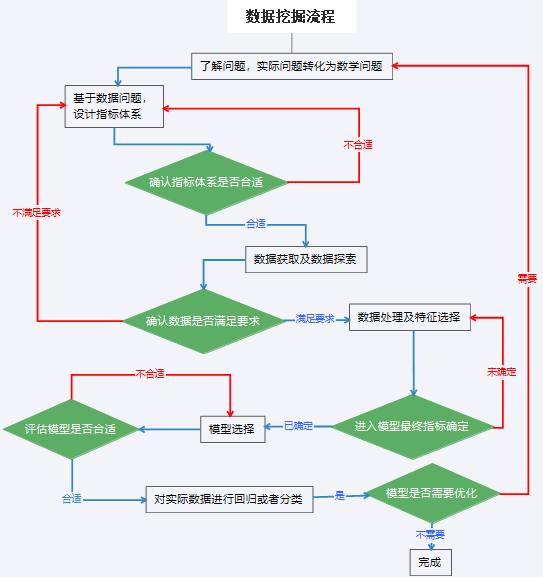
# 低能耗建筑用电量预测

## 项目介绍



低能耗建筑是指在围护结构、能源和设备系统、照明、智能控制、可再生能源利用等方面综合选用各项节能技术，能耗水平远低于常规建筑的建筑物，是一种不用或者尽量少用一次能源，而使用可再生能源的建筑物。而一个低能耗建筑的耗电量是多少呢，是否能准确预测出来？带着这些问题我们开始第一个数据挖掘项目：利用实验和天气等数据构建线性回归模型[（线性回归介绍）](https://zhuanlan.zhihu.com/p/23646095)预测比利时一个低能耗建筑的电量需求。

## 数据挖掘一般流程



数据挖掘的一般流程如上图，每一步的工作我们再详细介绍一下。第一步主要是结合实际要解决的问题，将实际问题转化为数学问题，也就是确认数据挖掘的目的，以后所有的工作都应该围绕着解决这个数学问题。

第二、三步就是设计指标体系及确认：指标体系的完善程度决定模型能否归纳出来正确的规律。所以指标体系要经过确认，特别是业务人员，指标体系的设计要符合业务流，如果指标体系确认有问题我们就要重新返回第一步，如果没有问题才进行第四步的数据获取、探索。

数据通常并不是集中在一个地方，很多的时候并没有现成的数据，要想办法获取这些数据，并且将数据整合在一起，进行数据探索，初步的了解数据。

在获取数据的时候很有可能并不能获得我们想要的数据，这时候我们就要回到设计指标体系这一步，设计可以获取数据的指标体系。当数据都获取完成，接着进行数据处理，数据在获取的时候可能会有缺失值、异常值或者文本型的数据，这时候我们就要对数据做相应的处理以适用后续模型的训练；在设计指标体系时的原则是越完善越好，这样就可能指标有冗余的情况出现，此时我们就要对特征进行选择或者转换，剔除多余的特征。这个步骤复杂且重要，需要反复确认调整，以保证进入模型的特征是合适的。

并不存在最好的模型，只有最合适的模型，在选择模型时需要综合考虑。就算综合考虑各种因素确认好了模型，也需要再对模型进行评估，评估模型的准确性、稳定性等等，最终评估通过的模型才能对实际的数据进行回归或者分类。现实问题是复杂多变的，这就要求模型要进行反复迭代优化以适应实际情况，所以数据挖掘过程是一个循环的过程。

由于此篇推文介绍的项目需要解决的问题及数据已经整理好，所以我们的前两步并没有涉及，我们只从数据获取开始介绍。

## 数据获取



本次项目的数据来源：实验人员使用m-bus电表每10分钟记录一次电量数据，持续约4.5个月，电量数据包括整个建筑的消耗的电量，也就是模型需要预测的数据，模型的自变量，还包括照明消耗的电量。实验人员还使用ZigBee无线传感器网络监测房屋温度和湿度条件，每个无线节点在3.3分钟内传输温度和湿度条件； 然后，将无线数据按照10分组平均，得到湿度和温度的数据。实验人员再从气象站的公共数据集下载天气数据，并根据日期和时间和实验数据集合并，这样就得到了我们此项目的数据。

## 数据探索

一般我们拿到一堆数据之后，并不知道数据有何规律，也不知道适合什么样的模型，为了了解数据特征的规律(概率分布)，我们先对数据进行分析探索。数据探索的目的是对数据有个大概的了解。弄清数据集大小、特征和样本数量，数据类型、数据的概率分布等。

### 数据规模

模型不同，运算量也是不同的，数据量更是影响到模型的准确性及稳定性。在选择模型的时候必须考虑数据量的大小。拿到数据我们首先要看我们数据的变量是多还是少，数据量是否足够。

本次数据共有19735行、27列，目前来看数据量不是特别少。

### 数据类型

通过了解数据类型来确定是否需要进一步编码或者转化。本次数据整个建筑物、照明的电量消耗为整型，时间为字符串，其他都为浮点型的数据。数据都是数值型的，不需要进行编码或者转换。

### 数据缺失情况

如果数据有缺失值则需要进行数据填补。通过查看本数据缺失情况，发现每列缺失值的数量都为0，也就是说，这个数据集是没有缺失值。

### 数据整体情况

通过查看数据分布，先了解数据整体分布情况。先通过了解各个变量常用描述性统计量的值让我们对数据有个整体的认识。

温度类数据：室内温度在14.9度和29.8度之间变化，变化差异不大；“建筑物外侧温度”在-6.1度到28.3度之间变化，这种变化差异较大的原因是传感器被放在房屋外面，外面的温度较室内温度波动大。

湿度类数据：除“浴室”和“建筑物外侧”外，房子内的湿度变化在20.5％到63.4％之间，而“浴室”和“建筑物外侧”由于所处的环境因素，分别在29.8％到96.3％、1％到99.9％之间变化。

“整个建筑消耗电量”：75％的设备消耗量低于100 Wh。最大功耗为1080 Wh时，初步判断此列中会有异常值，而且消耗量非常高的情况很少。

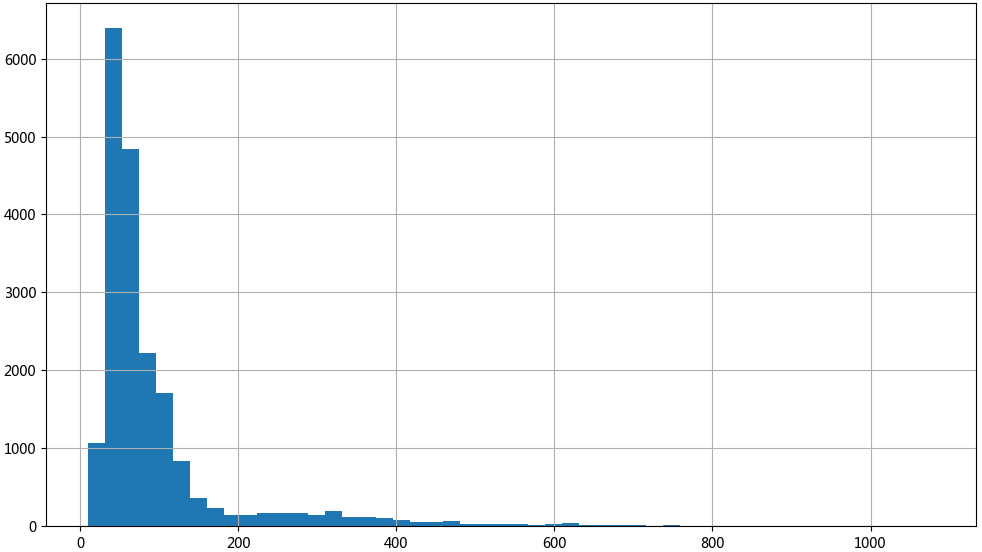
### 数据分布规律

模型的基线取决于数据的好坏，数据的好坏取决与你对数据的理解。所以为了更懂数据，就先看数据有哪些分布，同时查看是否有“异常值“。“异常值”（outlier）是指一组测定值中与平均值的偏差超过两倍标准差的测定值，与平均值的偏差超过三倍标准差的测定值，称为“高度异常值“。在一般的预测模型中，模型通常是对整体样本数据结构的一种表达方式，这种表达方式通常抓住的是整体样本一般性的性质，而那些在这些性质上表现完全与整体样本不一致的点，我们就称其为异常点，通常异常点在预测问题中是不受欢迎的，因为预测问题通产关注的是整体样本的性质，而异常点的生成机制与整体样本完全不一致，如果算法对异常点敏感，那么生成的模型并不能对整体样本有一个较好的表达，从而预测也会不准确，所以如果有异常值则需要对其进行处理。

通过上面的了解的数据类型我们知道我们的数据都是数值型的数据，对于数值型的数据我们画出变量的“直方图“。

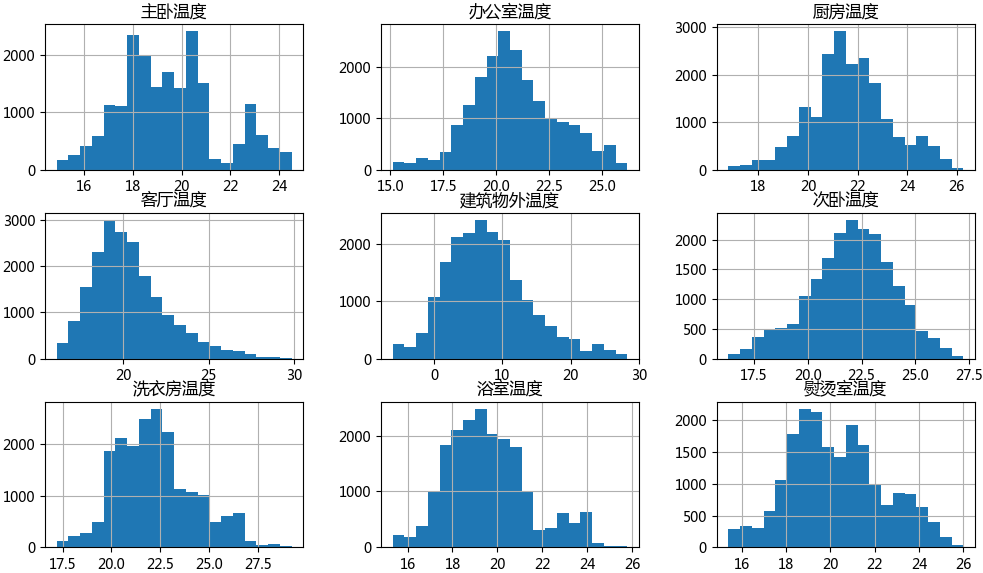
“直方图“是用来整理数值型的观测数据，用来分析数据是否服从正态分布，数据中心的偏移方向，同时判断数据有无异常。

为了好展示我们按照变量的类别来作图分析，首先目标变量“整个建筑物消耗的电量”：



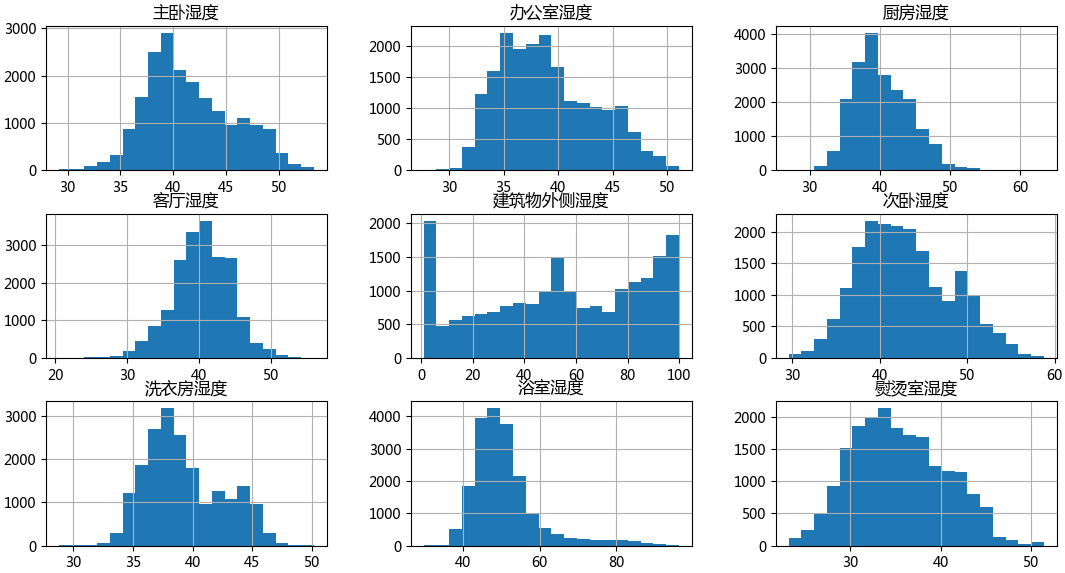
“整个建筑物消耗的电量”是正态分布的，而且是连续性的变量，那么我们选择模型是就要选择**回归模型**进行预测，而回归模型中最简单的就是线性回归模型，这里我们可以初步选择模型为线性回归模型；还可以看出图形是右偏的，大多数值均低于100 Wh。而且可以看出这个数据是存在“异常值”的。我看具体看一下异常值的数量。有874行的数据是高度异常值，占比4%，因为数据量并不是很多，我们进行删除。对于异常值的处理按照数据量及数据的不同有不同的处理办法，在这里我们就不进行详细的阐述了。

温度类数据：



数据不存在异常值，从图中可以看出除“主卧温度“外，所有列均近似遵循正态分布。

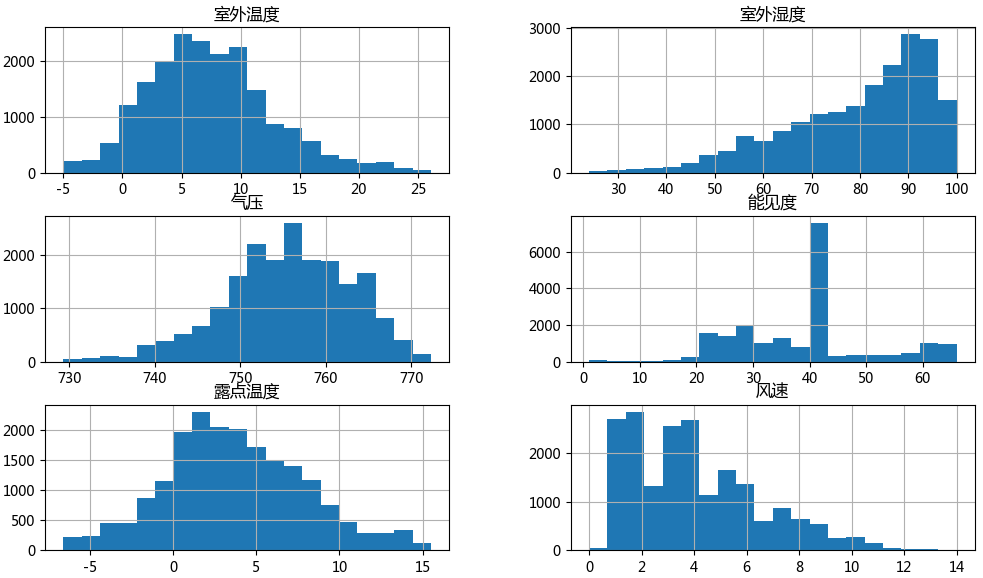
湿度数据：



除“建筑物外侧湿度“外，所有列均遵循正态分布，“建筑物外侧湿度“分布比较特殊，对数据进一步查看发现有1和99.9这两个值特别多，有可能这一列数据在采集不准确，在实际中需核查这一列的数据是否准确。

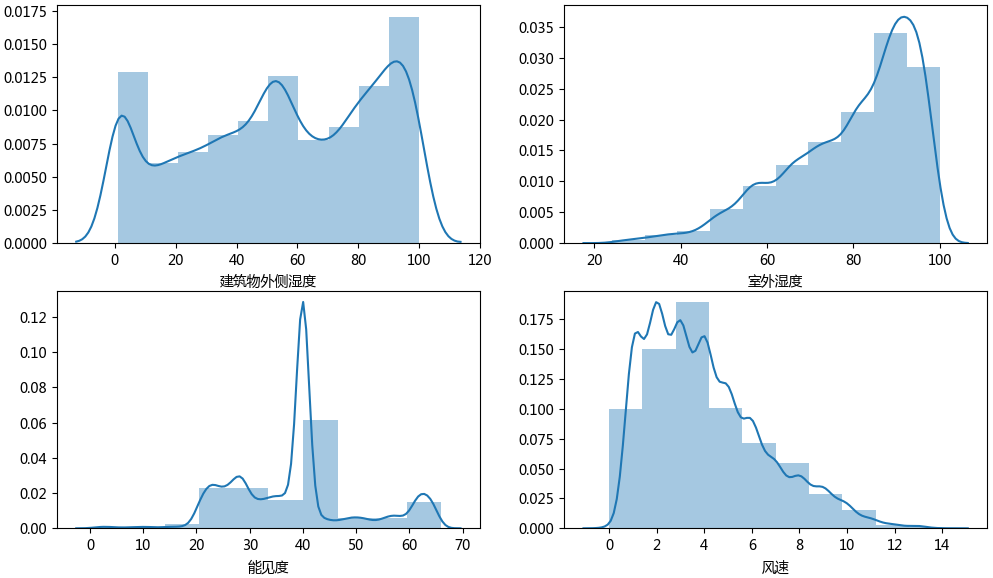


天气数据：



数据不存在异常值，能见度和风速不是正态分布。

“直方图”存在一个问题，组数不同数据分布显示也不同，而且“直方图”展示的分布曲线并不平滑，即在一个分组中的样本具有相等的概率密度，显然，这一点往往并不适合。解决这一问题的办法时增加组数，当组数增到到样本的最大值时，就能对样本的每一点都会有一个属于自己的概率，但同时会带来其他问题，样本中没出现的值的概率为0，概率密度函数不连续，这样的数据分布并不合理，而核密度估计可以解决问题，核密度估计完全利用数据本身信息，对数据分布不附加任何假定，避免人为主观带入得先验知识，就是采用平滑的峰值函数(“核”)来拟合观察到的数据点，从而对真实的概率分布曲线进行模拟，从数据样本本身出发研究数据分布特征的方法。所以，对于直方图不符合的正态分布的数据我们再画出核密度曲线图。



可以看出“建筑外测湿度“不符合正态分布，进一步怀疑这个数据采集存在问题。

数据探索在数据挖掘流程中起了很关键的作用，通过数据探索可以最大程度得到数据的直觉、发掘潜在的结构、提取重要的变量、删除异常值、检验潜在的假设、指导后续的数据处理、建立初步的模型、决定最优因子的设置。虽然在这个项目数据探索只让我们确认了初步的模型、删除了异常值。

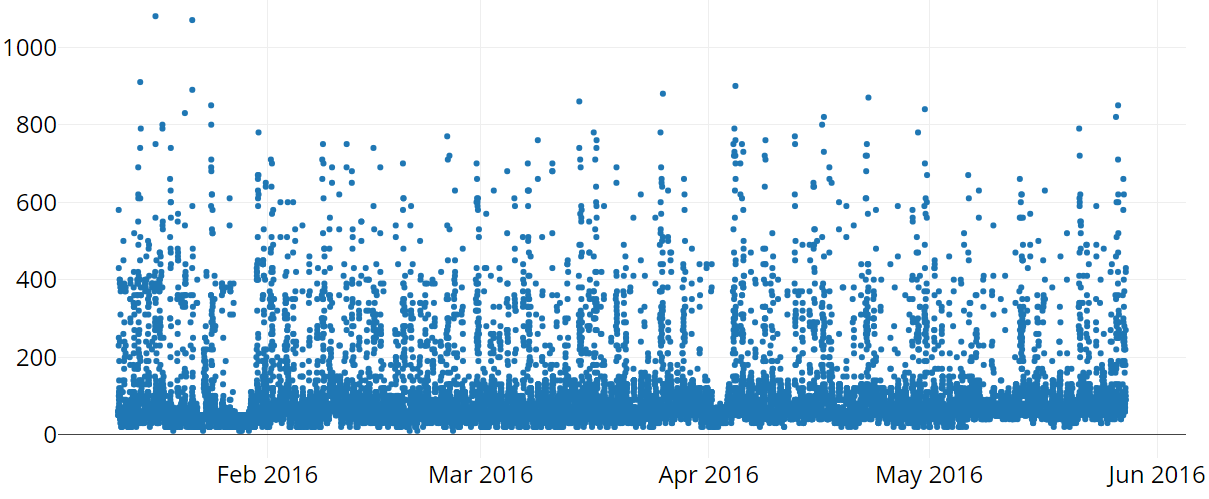
## 数据处理及特征选择

### 删除无关的变量

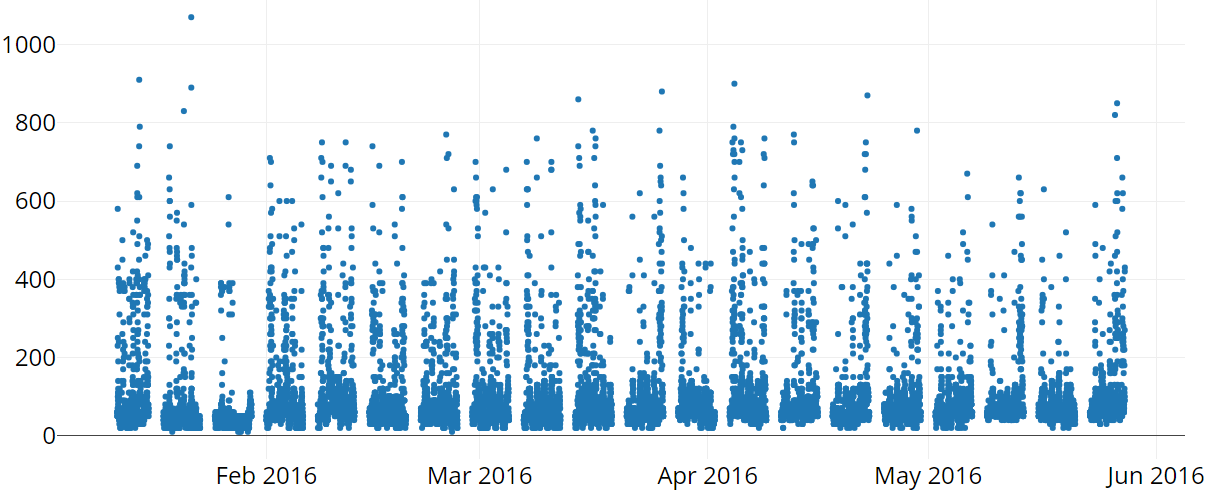
最初以为“灯具消耗的电量”将能够提供有用的信息，然而进一步查看这一列数据值时，发现有15252行都是0值，占比77%，这样的列不会对模型有贡献，所以我们忽略这一列。



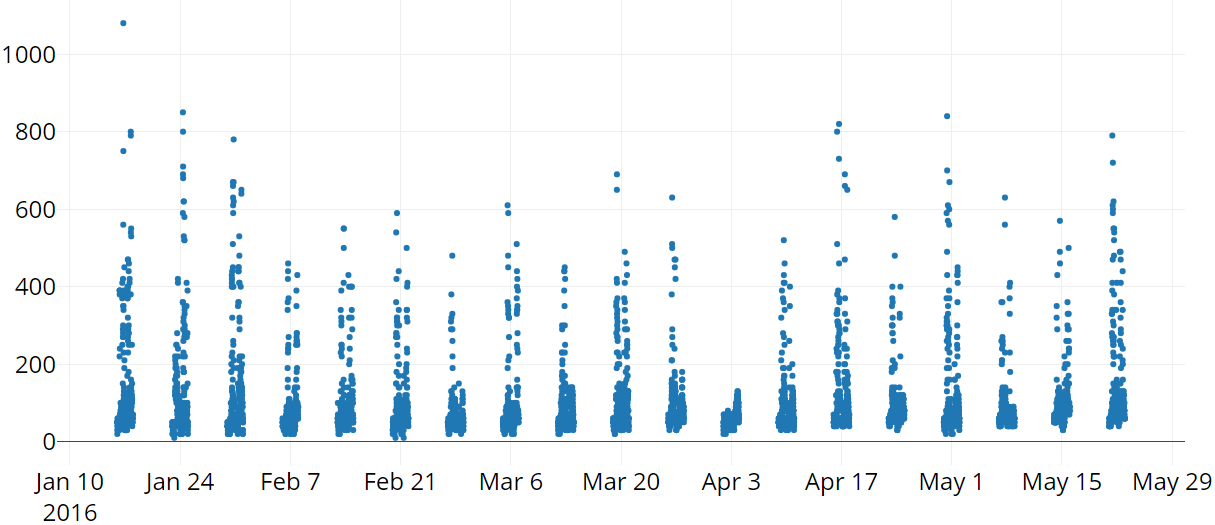
首先看一下目标变量也就是建筑的电量消耗的数据与时间序列的关系



再分别看一下工作日、周末建筑的电能消耗数据分布



最后看一下周末的



可以看到目标变量时间分布并没有什么关系，而且由于此次研究的是回归模型并不是时间序列模型，所以我们可以忽略“时间”列。

### 数据归一化处理

对于需要计算距离的模型，当数据有多个变量时，由于各变量的性质不同，通常具有不同的量纲和数量级。当各变量之间的水平相差很大时，如果直接用原始变量值进行距离的计算时，就会突出数值较高的变量对距离的影响，而消弱数值水平较低变量的作用。因此，为了保证结果的可靠性，需要对原始变量数据进行标准化处理，让各个变量对结果做出的贡献相同。而对于概率模型不需要归一化，因为这种模型不关心变量的取值，而是关心变量的分布和变量之间的条件概率。

在前面由于确认了此次项目运用线性回归模型，线性回归模型在用梯度下降法求解时需要计算距离，所以我们需要对数据进行标准化处理。

标准化的常用方法有最大最小标准化（Min-Max Normalization）、Z-score标准化、非线性归一化。我们先介绍一下各种标准化方法的适用场景，以便于选择合适的标准化方法。

（1）最大最小标准化（Min-Max Normalization）

最大最小标准化又称为离差标准化，使结果值映射到[0 ，1]之间，此方法比较适用在数值比较集中的情况。

缺陷：如果max和min不稳定，很容易使得归一化结果不稳定，使得后续使用效果也不稳定，实际使用中可以用经验常量来替代max和min。

 应用场景：在不涉及距离度量、协方差计算、数据不符合正态分布的时候，可以使用该方法。

（2）Z-score标准化

经过Z-score标准化处理后数据符合标准正态分布，即均值为0，标准差为1。 此缺陷：该方法要求原始数据的分布可以近似为高斯分布，否则归一化的效果会变得很糟糕。在前面对数据探索的时候可以发现我们的数据几乎都符合正态分布。

应用场景：在分类、聚类算法中，需要使用距离来度量相似性的时候、或者使用PCA技术进行降维的时候。

（3）非线性归一化

非线性归一化经常用在数据分化比较大的场景，有些数值很大，有些很小。通过一些数学函数，将原始值进行映射。包括log，正切转换等，需要根据数据分布的情况，决定非线性函数的曲线。

由于我们的模型需要计算距离且在前面数据探索的时候发现只有“建筑外测湿度“不符合正态分布，且这一数据很可能时存在问题的，综合考虑我们选择Z-score标准化对数据进行处理。查看进行处理后的数据情况如下（只选取了部分变量的数据）：

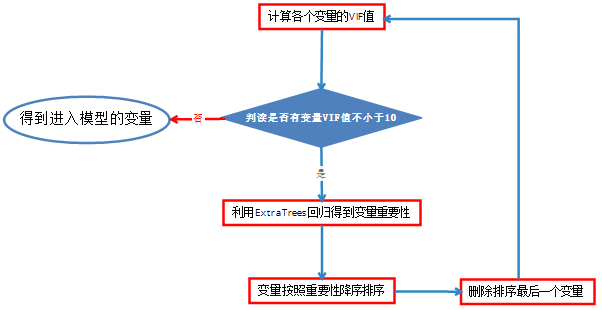


### 多重共线性检验

在前面我们确认了此次项目运用线性回归模型， 在回归模型中两个或两个以上的自变量彼此相关时，则称回归模型中存在多重共线性，也就是说共线性的自变量提供了重复的信息。这种多重共线性会有非常不好的影响：会造成回归系数、截距系数的估计非常不稳定，即整个模型不稳定。这种不稳定的具体表现是：很可能回归系数原来正，但因为共线性而变为负。这对于一些自变量的可解释性来讲可能是致命的，因为得到错误系数无法解释正常发生的现象。

所以，多重共线性检验就显得非常重要了。本次项目的数据我们通过判断就知道天气类数据中的室外温度、湿度分别与建筑物外侧的温度、建筑物外侧的湿度应该是相关性很高的。也就是我们在不进行检验的情况下已经知道数据的存在多重共线性。而如果在数据无法通过直接判断就能是否存在共线性的情况下，我们可以使用方差膨胀因子(VIF)来鉴定，**如果VIF大于10，则说明变量存在多重共线性**，通过计算各个变量的VIF发现我们这次的数据多重共线性很严重。出现这种情况，可以考虑删除变量和重新选择模型。由于此次只涉及介绍线性回归模型，我们解决共线性的办法只考虑删除变量。那我们就要对变量进行剔除了。

变量剔除的原则是通过ExtraTree回归[（Extra Tree回归原理介绍）](https://www.cnblogs.com/pinard/p/6156009.html)得到每个变量的重要性排序，按照重要性降序排序，从重要性最低的变量进行剔除，直到共线性检验通过。我们画出变量筛选的流程图如下：



经过筛选后符合共线性检验的数据如下：



“建筑物外侧湿度“VIF达到9.9，虽然没超过10，但是考虑到这个变量可能存在问题，我们剔除这个变量，所以最终选择进入模型的变量有8个：



## 构建模型及评估

### 构建模型

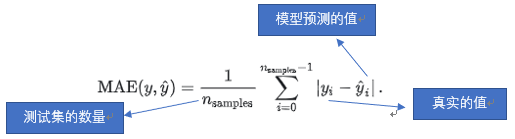
我们把经数据处理筛选的数据分为训练集和测试集，一般划分比例是7：3，不过也可以根据实际的项目进行调整。训练集用来训练模型得到各个变量的回归系数，测试集用来评估模型的好坏。经过训练变量的回归系数如下：



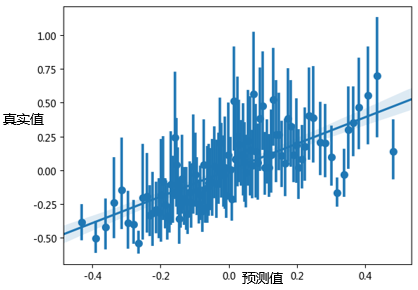
构建的线性回归模型为：

### 评估模型

那模型构建好了，怎么用具体数值来评估模型的好坏呢？我们用MAE（平均绝对误差）评估模型的好坏，也就是对所有测试集预测值与真实值之间差值绝对值的求平均，计算公式如下：



利用构建的模型在测试集上计算出来的MAE为0.6，由经过标准化后的数据整个建筑物的耗电量均值为0，方差为1，预测值与真实值平均相差0.6，说明模型的拟合程度并不好。



同时也可以画出真实值与预测值之间的差距，看模型的拟合情况，可以看出预测值偏离真实值较大。也就是说根据现有的数据利用线性回归模型并不能准确的预测整个建筑的用电量。出现这种情况可以重新获取数据或者重新选择模型，而通常获取新类型的数据并不是一件容易的事，很多时候更简捷的处理办法是重新选择模型，重新选择模型需要对数据及模型都很有很深刻的理解，在这篇推文中就不展开介绍了。