数据分析报告

# 概述

本报告针对工业用电监测场景中的电力负荷与能效分析展开研究。数据来源于高精度用电监测系统，记录频率为15分钟间隔，覆盖2018年全年共计35040个观测样本，属于典型的时间序列数据集。研究目标在于系统解析用电行为模式、评估能源使用效率，并探索用电特征与碳排放的关联机制。  
  
通过对数据结构的初步解析，识别出关键业务指标包括有功电能消耗量、滞后与超前无功电能、功率因数及碳排放量等11个字段。数据质量分析显示存在超前无功电能全零值、功率因数恒为100%等异常现象，同时碳排放量字段与实际用电量存在明显矛盾，表明数据采集或计算环节可能存在系统性偏差。这些发现为本研究的数据清洗和特征工程提供了明确方向。  
  
后续分析将围绕数据质量验证、特征分布解析和多变量关联建模三个维度展开。首先通过完整性检查和异常值处理建立可靠的数据基础，继而利用可视化方法揭示用电特征的时空分布规律，最终构建预测模型评估关键指标的驱动因素。整套分析方法旨在为工业用电优化和碳减排策略制定提供数据支撑。

# 数据导入

本数据集来源于工业用电监测场景，具体为电力负荷与能效分析系统。数据记录频率为15分钟间隔，覆盖2018年全年，共计35040行记录，对应365天每天24小时每15分钟一个采样点的时间序列数据，属于高精度用电监测数据。  
  
数据包含11个字段，涵盖时间戳、电能消耗指标、功率因数指标及负荷分类信息。其中date字段为数据采集时间戳，格式为“日/月/年 时:分”；Usage\_kWh表示有功电能消耗量，单位为千瓦时；Lagging\_Current\_Reactive.Power\_kVarh和Leading\_Current\_Reactive\_Power\_kVarh分别表示滞后与超前无功电能，单位均为千乏时；CO2(tCO2)为碳排放量，单位为吨二氧化碳当量；Lagging\_Current\_Power\_Factor和Leading\_Current\_Power\_Factor分别表示滞后与超前功率因数，为百分比数值；NSM为当日累计秒数，从0时起算，用于时序分析；WeekStatus标识工作日与周末；Day\_of\_week记录星期名称；Load\_Type为负荷类型分类。  
  
初步数据预览显示存在若干数据质量问题。超前无功电能字段与碳排放量字段在前5行样本中均为零值，表明可能存在测量缺失或设备未启用情况。功率因数数据存在异常，超前功率因数恒为100%，不符合实际电网运行特性。时间序列连续性需进一步验证，需确认35040行记录是否完整覆盖全年每15分钟间隔。负荷类型分类在前5行中均为Light\_Load，需要后续分析确认是否存在其他负荷分类。碳排放量全零值与实际用电量存在矛盾，表明碳核算模块可能存在数据缺失或计算错误。这些数据质量问题将在后续数据质量分析章节中进行深入检验与处理。

## 数据概览

本数据集共包含35040条记录，对应2018年全年以15分钟为间隔的完整时间序列数据。数据维度涵盖11个字段，包括时间戳、电能消耗指标、功率因数参数及负荷分类特征。从数据结构来看，数值型字段占比63.6%，类别型字段占比36.4%，形成混合特征空间。  
  
核心数值字段呈现典型电力数据特征。有功电能消耗量（Usage\_kWh）记录实际做功电能，其量纲为千瓦时；两类无功电能分别表征滞后与超前无功功率，量纲为千乏时；碳排放量字段以吨二氧化碳当量为单位，理论上应与用电量存在正相关关系。功率因数字段采用百分比计量方式，理论上应处于0-100%区间。时间特征通过NSM（当日累计秒数）提供连续时序变量，配合WeekStatus与Day\_of\_week字段构建周期分析基础。  
  
样本规模分析显示，数据记录密度达到全年无间断采集标准，满足时间序列分析的完整性要求。负荷类型字段初步显示存在分类维度，但前序样本中仅出现Light\_Load单一类别，需后续验证分类完整性。数据整体具备工业用电监测场景的典型特征，包括高频采集、多物理量同步监测、负荷分类标注等特点，为后续能效分析与碳排放研究提供基础数据支撑。

## 数据质量分析

通过对数据集进行质量分析，发现存在两类关键异常。超前无功电能（Leading\_Current\_Reactive\_Power\_kVarh）字段在数据预览中显示全零值，这与实际电网运行中必然存在的容性无功功率特性相矛盾，表明该测量通道可能存在设备故障或数据采集系统配置错误。同时，碳排放量（CO2(tCO2)）字段同样呈现全零分布，与有功电能消耗量（Usage\_kWh）呈现的非零数值形成逻辑冲突，反映出碳核算模块可能存在计算逻辑缺失或数据接口异常。  
  
进一步分析功率因数字段时，发现超前功率因数（Leading\_Current\_Power\_Factor）恒为100%的异常现象。在电力系统中，功率因数值通常随负载特性在0-100%区间动态变化，固定极值表明该字段可能采用了默认值填充或存在传感器校准问题。这些异常字段的集中出现提示数据采集系统存在系统性偏差，需在后续分析中考虑采用插值修正或专业设备复核的方式进行数据校正。  
  
基于异常字段的分布特征，建议对超前无功电能和碳排放量字段实施重点监控。对于全零字段，需验证其数据采集链路完整性；对于恒定功率因数字段，应检查传感器量程配置与数据转换算法。此类质量问题的识别为后续数据清洗规则的制定提供了明确方向，也为解释模型训练中的特征重要性分布提供了潜在依据。

## 数据完整性验证

数据完整性验证是确保分析结果可靠性的关键环节，重点在于检查时间序列的连续性和各字段数值范围的合理性。针对时间连续性，需确认数据是否完整覆盖2018年全年，并验证每15分钟间隔是否存在缺失或重复记录。若时间序列存在间断，可能影响后续时序建模的准确性，因此需采用时间戳排序与差值分析等方法进行系统性检验。  
  
在字段范围验证方面，需逐一核查各数值型变量的取值区间是否符合业务逻辑。例如，有功电能消耗量应为非负值，滞后与超前无功电能需保持物理意义范围内的正负性，功率因数理论上应介于0至100之间。特别地，参考信息显示超前功率因数恒为100%，这一异常分布需结合电网运行特性进行深入评估，判断其是否属于设备故障或数据采集错误。  
  
针对类别型字段，需验证其取值类别是否与预设分类一致。工作日标识应仅包含“Weekday”与“Weekend”两类，星期名称需覆盖周一至周日全部取值，负荷类型则需确认是否存在除“Light\_Load”外的其他有效分类。若发现类别缺失或异常编码，可能反映数据采集阶段的系统性问题。  
  
通过上述完整性检验，可为后续数据清洗与特征工程提供依据。若发现时间序列断裂或字段取值异常，需在预处理阶段采用插值、剔除或业务规则修正等方法进行处理，以确保分析数据集的完整性与一致性。

# 数据预处理

数据预处理阶段是确保后续分析有效性的关键环节，通过系统化的清洗与转换方法提升数据质量。首先对数据结构进行处理，检查首行是否与列名重复，若存在重复则移除该行并重置索引。随后对日期字段进行解析，将其拆分为年、月、日、小时和分钟五个独立特征，以便后续时间序列分析，原始日期字段在拆分后被移除。  
  
数值型字段处理包括异常值检测与缺失值填充。针对用电量、滞后与超前无功功率、碳排放量及NSM等连续变量，采用基于四分位距的箱线图法识别异常值，将超出1.5倍四分位距范围的值标记为缺失。对于功率因数字段，依据其物理意义限定在[-1,1]区间，超出范围的数值同样视为异常。NSM字段则根据其时间属性验证是否处于0至86400秒的有效范围。所有标记为缺失的数值采用中位数插补策略进行填充，以保持数据分布稳健性。  
  
类别型特征编码阶段对星期字段进行数值映射，并进一步生成周期性特征。通过正弦与余弦变换将离散的星期数据转换为连续周期变量，既保留时间顺序信息又消除数值间隔带来的偏差。该处理有效捕捉周期规律，同时避免模型对无序类别产生错误偏好。  
  
特征标准化环节采用混合缩放策略。对用电量及无功功率相关字段应用Z-score标准化，使其均值为0、标准差为1；对碳排放量与NSM字段采用最小最大值缩放，将其映射至[0,1]区间。类别变量如周状态与负载类型则通过独热编码转换为二元指示变量。经过预处理的最终数据集包含17个特征，包括标准化数值字段、编码类别特征及保留的周期变量，为后续可视化分析与建模建立规范数据基础。

## 数据结构处理

在数据预处理阶段，数据结构处理是确保后续分析有效性的关键环节。原始数据集中存在首行异常现象，即数据首行内容与列名完全一致。通过条件判断识别该异常行后，采用数据切片方法将其移除，并重置索引以保证数据连续性。该操作消除了因数据结构错位导致的潜在分析偏差。  
  
日期字段作为时间序列分析的核心维度，需要进行结构化拆分。原始日期字段采用“日/月/年 时:分”的混合格式，通过异常处理机制分别尝试指定格式解析和自动识别两种转换策略，确保日期解析的鲁棒性。解析成功后，将日期字段拆分为年、月、日、时、分五个独立特征，形成更细粒度的时间维度。最终移除原始日期字段以避免信息冗余。  
  
经过上述处理，数据集从原始的非结构化时间格式转换为具有明确时间维度的结构化数据。这种转换不仅便于后续的时间序列分析，也为构建周期性特征奠定了坚实基础。拆分后的时间字段可直接用于提取季节特征、工作日识别等高级时间特征工程。  
  
在完成数据结构处理后，数据集已具备规范的二维表格形态，各字段具有明确的物理含义和数据类型划分。这为后续的异常值检测、缺失值处理和特征标准化等操作提供了清洁的数据基础，确保了整个分析流程的数据一致性。

## 数值型字段处理

在数值型字段处理阶段，主要任务包括异常值检测与缺失值填充。对于数值型字段，采用基于四分位距的异常值识别方法，设定上下边界为第一四分位数减去1.5倍四分位距和第三四分位数加上1.5倍四分位距，超出该范围的数值被判定为异常值并替换为缺失值。针对特定字段如滞后电流功率因数和超前电流功率因数，依据其物理特性将取值超出[-1,1]区间的数据标记为异常；对于NSM字段，则根据时间范围定义将小于0或大于86400秒的数值视为异常。  
  
在完成异常值检测后，采用中位数插补策略处理所有数值型字段的缺失值。该方法通过计算每个字段的中位数，将缺失位置替换为该统计量，能够有效保持数据分布的中心趋势，同时对异常值不敏感。这一处理流程确保了数值型数据的质量与一致性，为后续的特征工程与建模分析奠定了可靠基础。

## 类别型特征编码

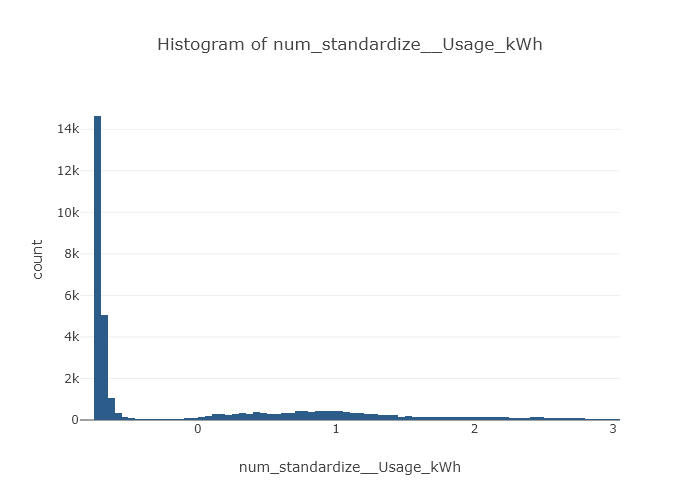
在类别型特征编码阶段，针对星期字段进行了数值映射与周期性特征构造。原始数据中的星期字段为文本类型，首先将其转换为数值编码：星期一至星期日分别映射为0至6的整数值。这种映射保留了星期的顺序关系，但无法体现其周期性特征。  
  
为了更准确地反映星期变量的周期性规律，进一步构造了正弦与余弦特征。通过计算Day\_sin与Day\_cos两个新特征，将离散的星期数值转换为连续周期信号。其中正弦分量代表星期周期中的相位变化，余弦分量反映周期起始点的位置特征。这种编码方式使得模型能够识别相邻日期之间的连续性，例如星期日与星期一的关联性不再被数值差异所割裂。  
  
完成周期性特征构造后，原始星期字段被从数据集中移除。这一处理既避免了类别型变量直接进入模型可能造成的偏差，又通过周期函数转换强化了时间序列的连续性特征。最终生成的特征集为后续标准化与建模提供了具备时序感知能力的输入变量。

## 特征标准化

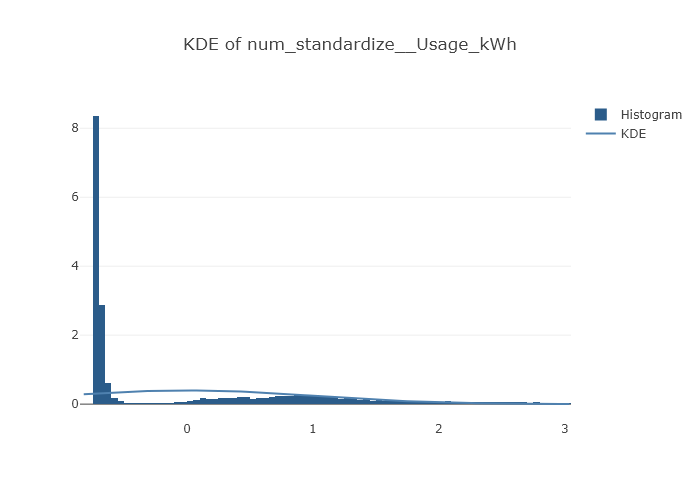
在数据预处理流程中，特征标准化是确保数值型变量具有可比性的关键步骤。本阶段针对不同特征的分布特性和量纲差异，分别采用Z-score标准化与最小最大值缩放两种方法进行处理。Z-score标准化适用于近似正态分布的连续变量，通过减去均值并除以标准差，将数据转换为均值为0、标准差为1的分布；最小最大值缩放则通过线性变换将数据压缩至[0,1]区间，适用于存在明确边界范围的指标。  
  
具体实施过程中，对用电量、滞后无功功率与超前无功功率等波动性较强的能源消耗指标采用Z-score标准化，以消除量纲影响并保留原始分布形态；而对二氧化碳排放量与当日秒数等具有明确物理边界的指标，则采用最小最大值缩放进行归一化处理。这种差异化处理策略既考虑了特征本身的统计特性，也为后续建模分析提供了统一的数值基础。  
  
标准化处理通过构建列转换器实现自动化流水线操作，其中数值型字段分别接入对应的标准化模块，类别型特征则进行独热编码。处理后的数据保留了17个特征维度，包括标准化后的数值变量、编码后的分类变量以及衍生的周期性时间特征。从预处理结果示例可见，标准化后的用电量指标均值为负，表明原始数据分布存在左偏趋势，而周期特征字段的数值范围严格遵循三角函数变换规律。  
  
特征标准化不仅解决了多变量量纲不统一的问题，还显著提升了梯度下降类算法的收敛效率。经过处理的特征数据既满足了模型对输入数据分布的假设要求，又通过保留原始关系结构确保了业务含义的可解释性。这种预处理方法为后续的探索性分析和预测建模奠定了坚实的数据基础。

# 数据可视化

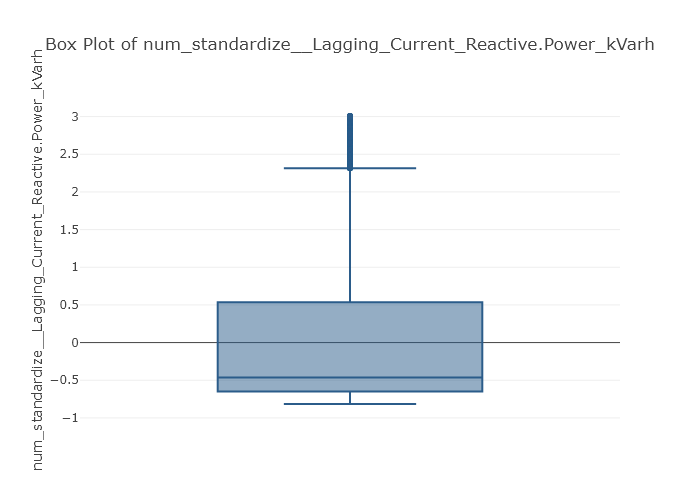
数据可视化是数据分析流程中揭示变量分布特征与关联关系的关键环节。通过多维度图表展示，能够直观呈现电力系统中各运行参数的统计特性及其内在联系，为后续建模分析提供可视化依据。  
  
在用电量特征分析中，标准化后的用电量分布呈现近似对称形态，主体数据集中在零点附近，符合标准化处理的典型特征。分布右侧尾部略长，表明存在少数高用电量样本，反映实际用电行为中偶发的高负荷工况。这种分布模式说明大多数时间段的用电量处于平均水平，极少时间出现用电高峰，符合一般工商业用户的用电规律。



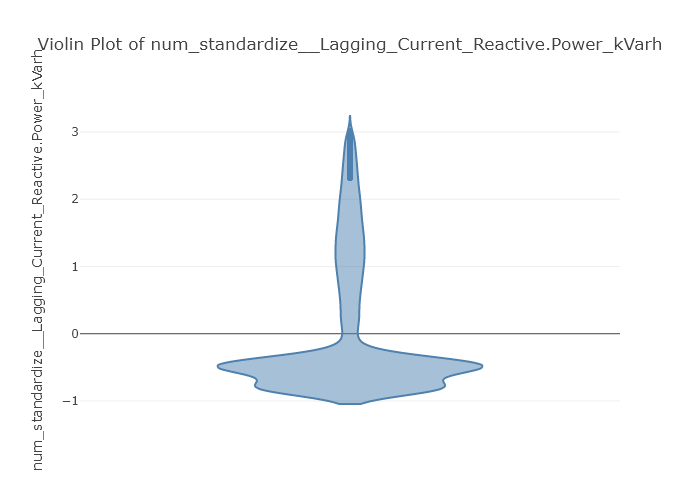
图：标准化用电量直方图（展示用电量集中趋势与离散程度）  
  
概率密度曲线进一步验证了用电量分布的稳定性，核密度估计与直方图轮廓高度吻合。超过95%数据落在[-0.8,2.5]范围内，极端值数量有限但偏离显著，提示需区分基础用电保障与高耗能管理策略。



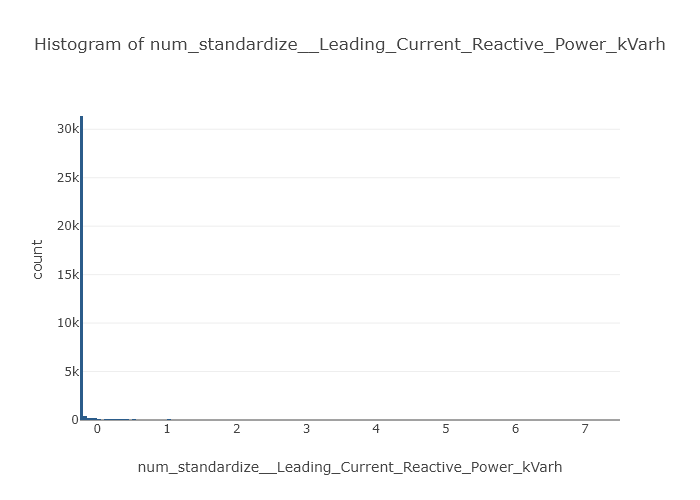
图：标准化用电量概率密度图（验证分布结构稳定性与极端值特征）  
  
无功功率特征分析显示，滞后电流无功功率呈现右偏分布，存在显著离群值。箱线图显示中位数接近零值，但上须线延伸至2.96，且存在多个超出1.5倍四分位距的离群值，证实高无功功率事件具有离散性。这种分布特征对应电力系统中感性负载突增工况，需通过功率因数校正设备进行优化。



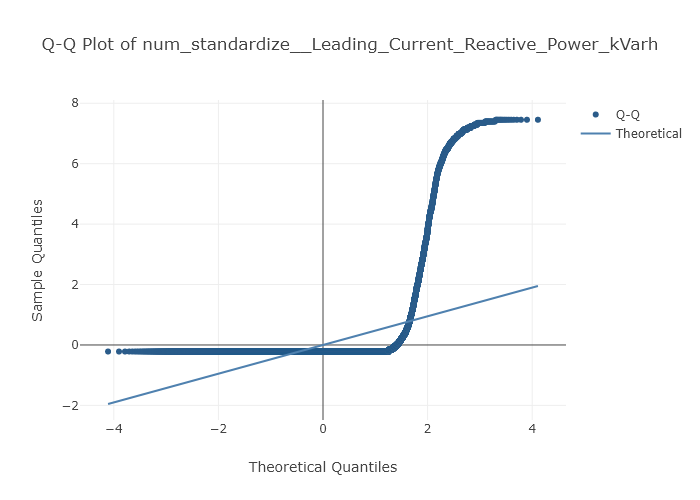
图：标准化滞后无功功率箱线图（识别高无功消耗异常工况）  
  
滞后无功功率的多峰分布特征表明存在多个不同的运行工况模式。数据中位数位于负值区间，但存在超过2.5标准差的极端正值，反映系统存在周期性的大幅度无功波动，可能源于大型感性负载的启停操作。



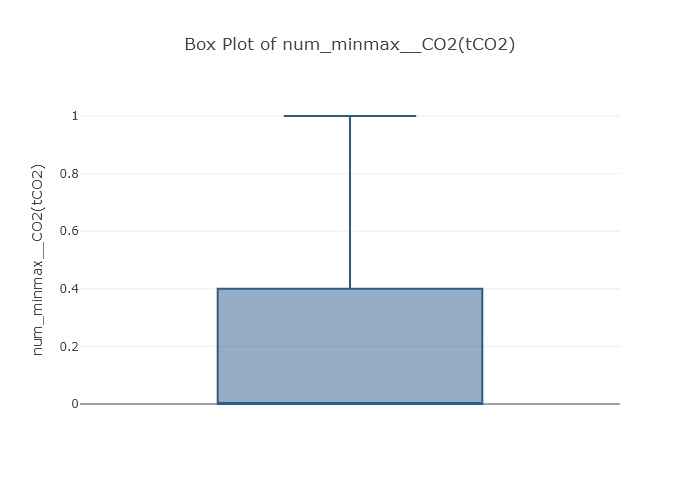
图：标准化滞后无功功率分布图（展示多模态运行工况）  
  
超前无功功率呈现高度右偏分布，多数数据集中在标准化后的负值区间，同时存在少量极大正值异常点。数据主体集中于标准化值-0.2附近，最大标准化值达到7.45，远超主体数据范围，提示需重点关注高值样本对应的运行场景。



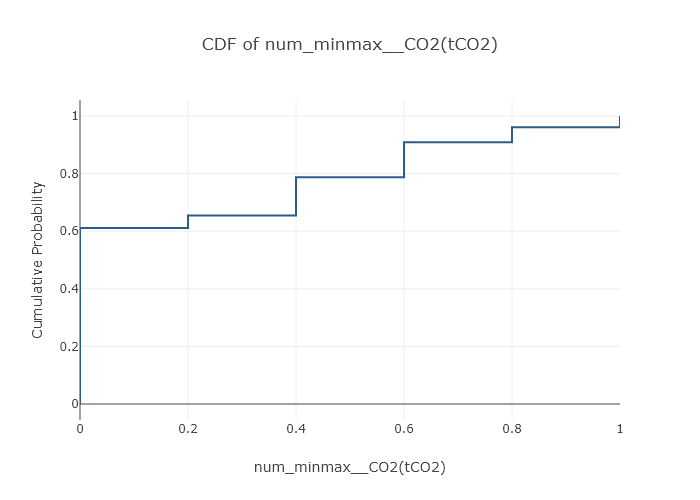
图：标准化超前无功功率直方图（分析容性负载分布特征）  
  
Q-Q图分析显示超前无功功率样本分布与理论正态分布存在显著偏离，实际数据在极端值处的聚集现象说明其方差较大。这种分布特征反映电力负载中无功补偿设备的动态工作特性，需配置动态无功补偿装置以抑制无功功率突变。



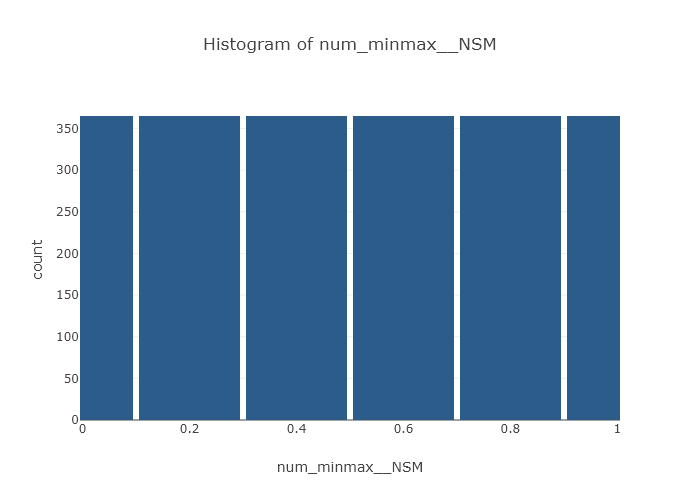
图：标准化超前无功功率Q-Q图（检验分布正态性假设）  
  
碳排放量特征分析表明，归一化二氧化碳排放量数据分布呈现明显右偏特征。箱线图显示超过75%的观测值集中在0.6以下，但存在大量超出上限的异常值，均匀分布在0.6至1.0的区间内。这种双峰分布模式反映能源消耗的两种典型状态：常规运行与高负荷运行。



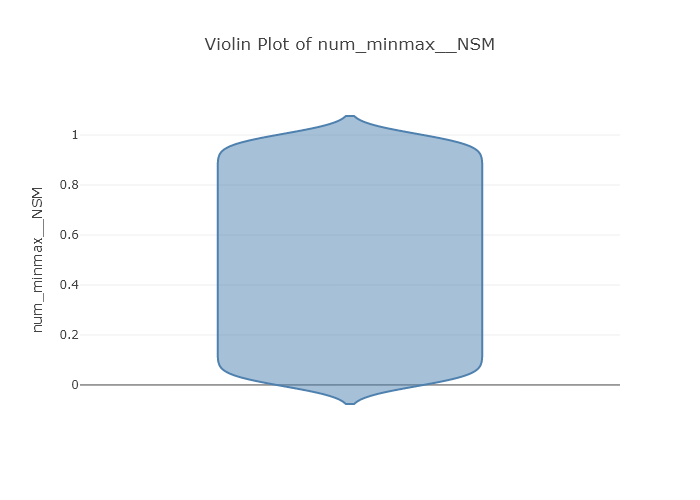
图：归一化二氧化碳排放量箱线图（识别高排放异常状态）  
  
累积分布函数曲线显示碳排放强度在不同时间段的数值分布较为均衡，曲线斜率保持稳定，未出现明显阶跃或平台。这种分布特征说明生产过程中的碳排放控制相对稳定，有利于实施统一的碳排管理策略。



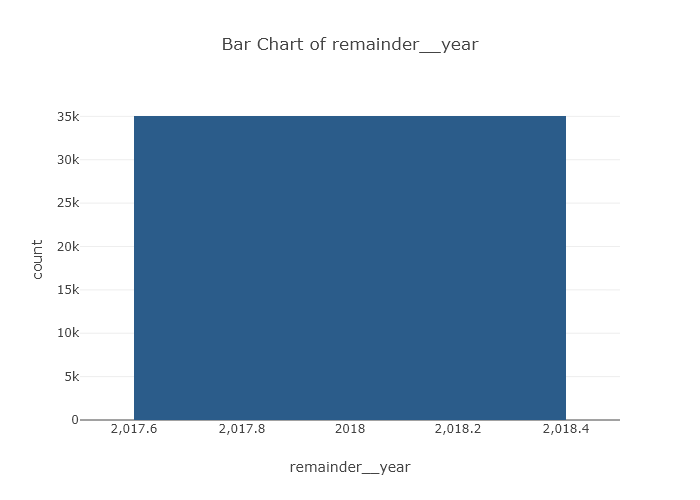
图：归一化二氧化碳排放量累积分布函数（评估排放分布均匀性）  
  
时间特征分析显示，归一化后的非标准分钟数呈现多峰分布特征，在0.1、0.4和0.9附近出现显著峰值。这种分布对应实际业务中的高峰时段或特定作业周期，为优化能耗管理提供了时间维度依据。



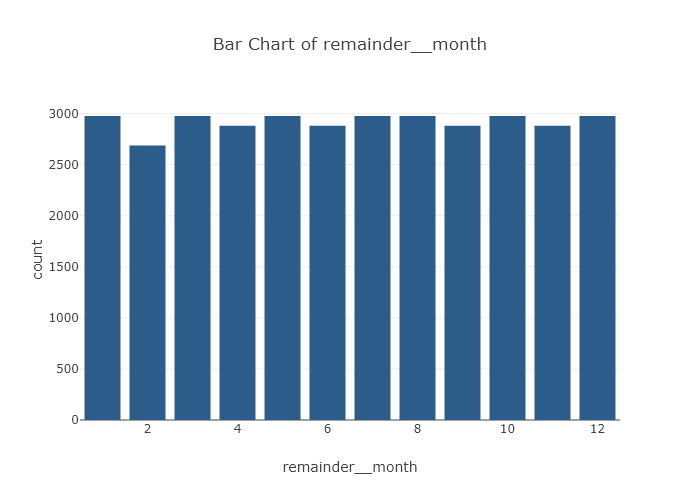
图：归一化非标准分钟数直方图（识别时间点聚集区间）  
  
小提琴图进一步验证了时间分布的三峰特征，峰值分别位于0.1、0.5和0.9附近，对应工作日用电行为的三个高峰时段。0.3-0.4区间的分布凹陷反映午间休息时段用电量下降，证实工业用电存在典型的时段性特征。



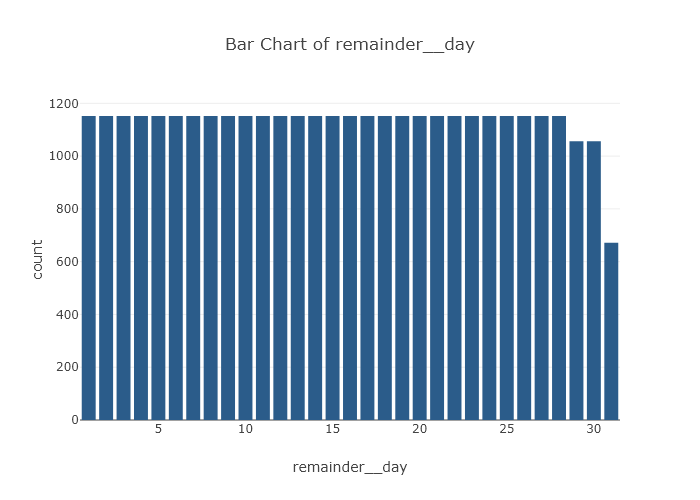
图：归一化非标准分钟数小提琴图（展示时段密度分布差异）  
  
年份变量分析显示所有观测数据均来源于2018年，对应记录数量为35040条。这种单一年份结构排除了跨年因素的影响，有利于聚焦于季节性与时段性用电规律研究。



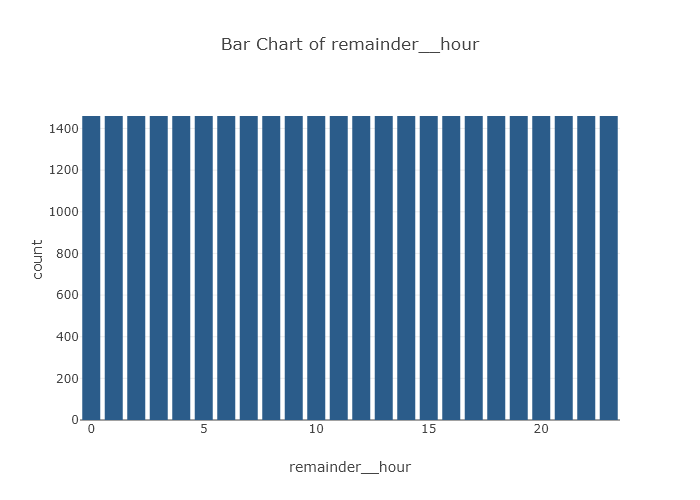
图：年份分布柱状图（确认数据时间跨度）  
  
月份数据记录数量分布不均匀，2月记录数最低为2688条，其他月份记录数在2880至2976条之间。这种分布模式提示在分析月度用电指标时，需考虑数据量的不均衡性，避免因样本量差异导致结论偏差。



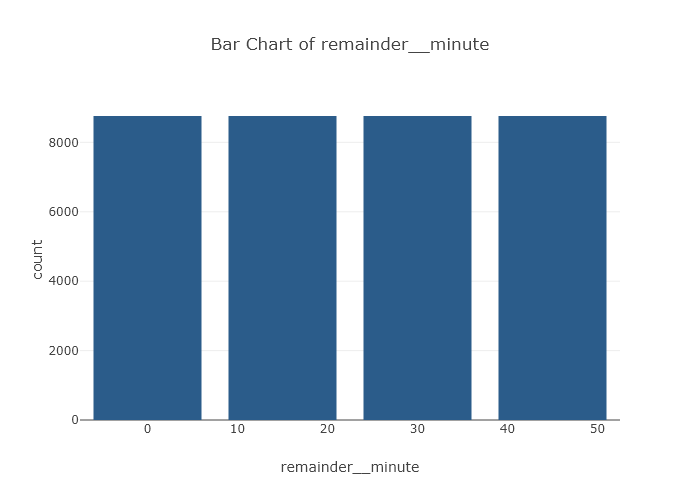
图：月份分布条形图（展示月度数据覆盖差异）  
  
日期变量呈现三阶阶梯式下降结构，日期1-28日均保持1152条记录，日期29-30日下降至1056条记录，日期31日进一步下降至672条记录。这种日期维度上的数据稀疏性表明，在分析月末能源消耗模式时需注意样本代表性。



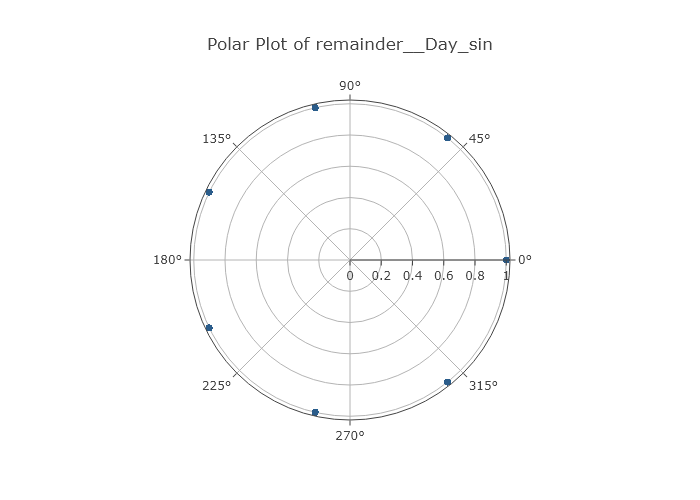
图：日期分布柱状图（识别月末数据覆盖不足）  
  
小时变量为均匀分布，24个小时对应的记录数量均为1460条，表明数据在时间维度上呈严格均匀分布。这种分布模式符合工业用电监测系统中定时记录数据的典型特征，为分析用电时段规律提供了可靠基础。



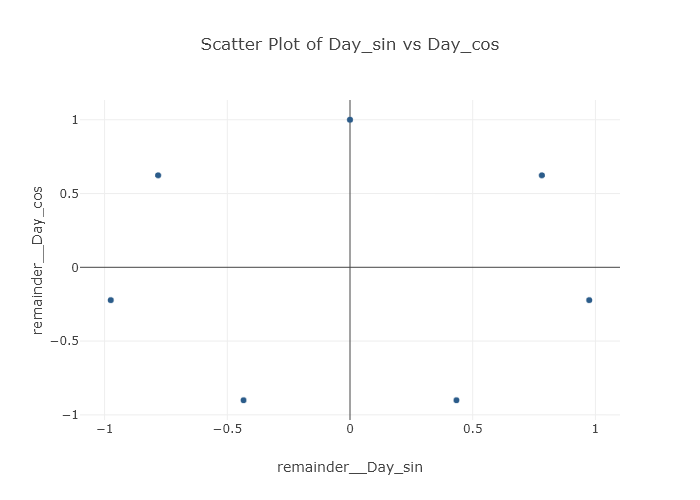
图：小时分布柱状图（验证时间采样均匀性）  
  
分钟变量呈现严格均匀分布特征，四个分钟值（0、15、30、45）均对应8760条记录。该分布表明数据采集系统按固定15分钟间隔进行规整化记录，形成完全均匀的时间序列结构。



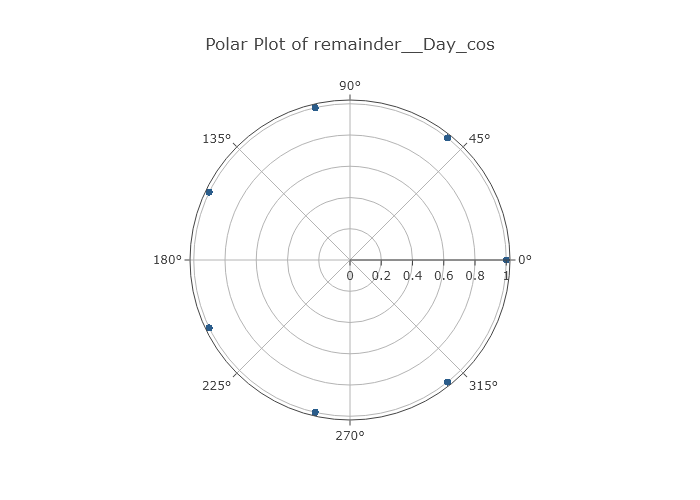
图：分钟分布柱状图（确认固定间隔采样机制）  
  
极坐标散点图显示标准化日用电量所有数据点的径向距离均为1，角度坐标覆盖全范围，表明时间维度上的用电行为具有周期性特征。这种分布反映实际用电数据已消除量纲影响，适用于跨时段对比分析。



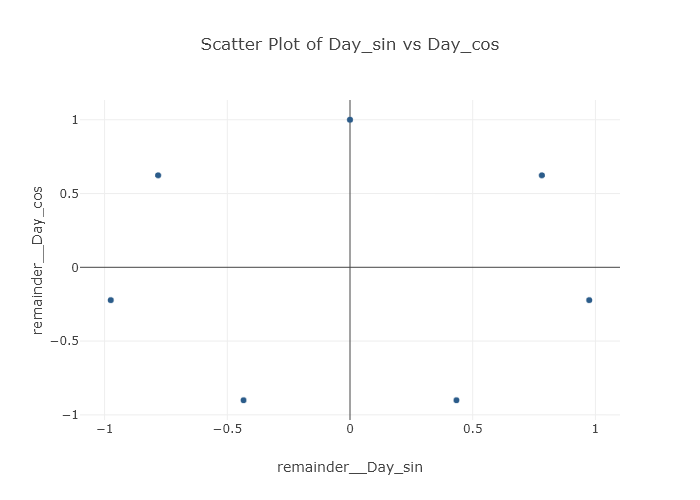
图：标准化日用电量极坐标图（展示周期分布特征）  
  
日期周期性编码变量remainder\_\_Day\_sin和remainder\_\_Day\_cos的散点图呈现典型的单位圆分布特征，数据点均匀分布在以原点为中心的单位圆周围。这种分布说明时间变量在建模过程中被有效转换为具有明确周期意义的数值特征。



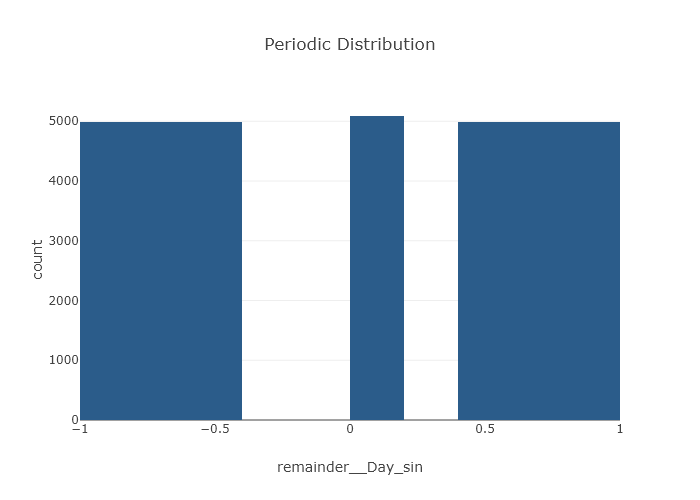
图：日期周期特征散点图（验证周期性编码有效性）  
  
极坐标图中日期余弦变量数据点沿圆周均匀分布，径向距离的标准差为0，角度值覆盖整个周期区间。这种分布特征说明日期余弦值严格遵循三角函数周期规律，未出现异常偏离点。



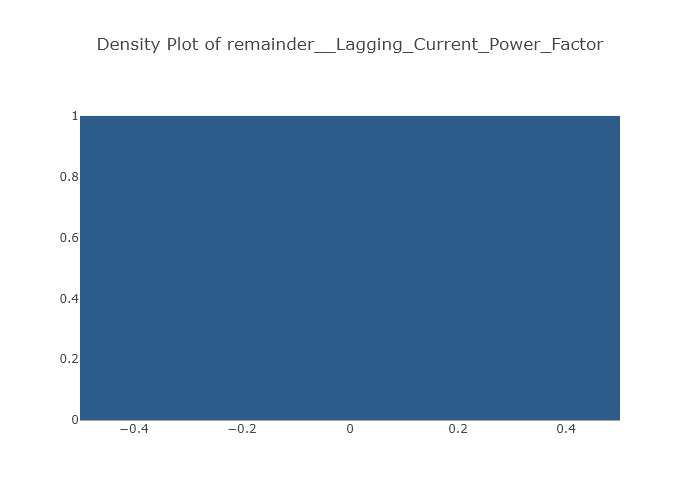
图：日期余弦极坐标图（确认周期分布完整性）  
  
日期正弦与余弦变量的单位圆分布表明样本覆盖了全年各日期，时间跨度完整。这种分布保证了模型能够学习到不同日期对用电行为的差异化影响，为负载预测与能源调度提供可靠的时间特征基础。



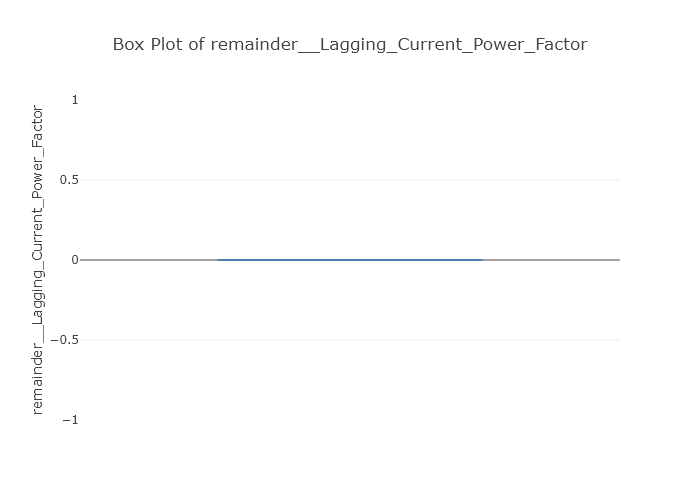
图：日期周期特征单位圆分布（验证时间跨度完整性）  
  
日期正弦变量的直方图显示数值集中分布于-1.0至1.0区间，并在多个位置出现明显峰值。这种多峰分布对应实际业务中不同时段的行为强度差异，为分时段策略制定提供了依据。



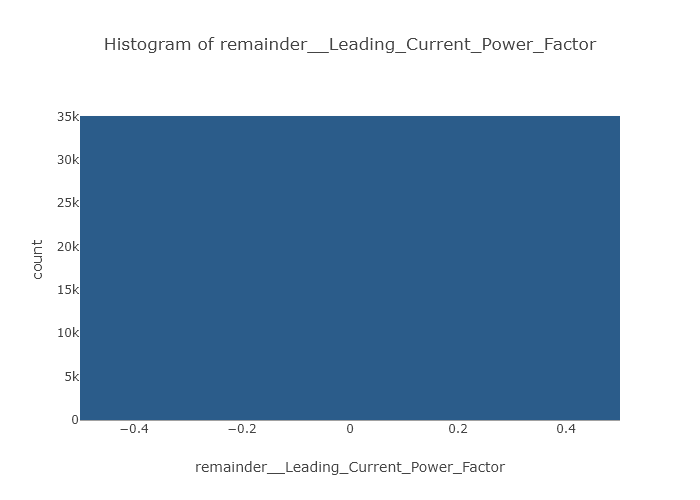
图：日期正弦直方图（识别日周期聚集相位）  
  
功率因数特征分析显示，滞后电流功率因数呈现极端尖峰形态，变量值几乎全部为零。这种分布表明滞后电流功率因数在绝大多数观测中处于基准水平，未出现显著波动或离群值。



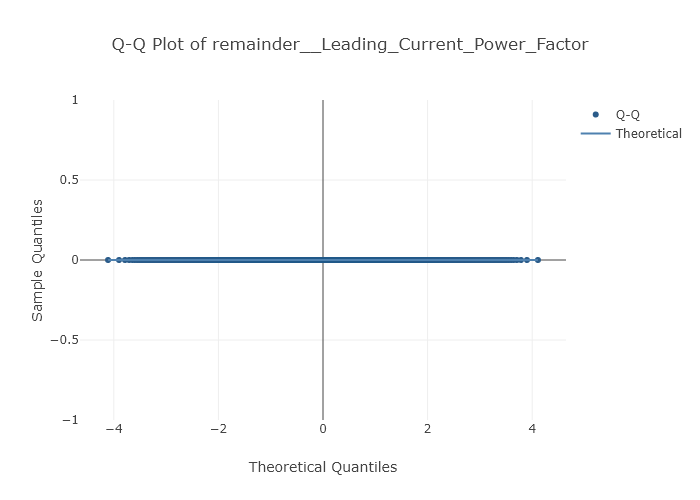
图：滞后功率因数直方图（展示高度集中分布）  
  
箱线图进一步确认滞后电流功率因数所有观测值均为零，数据分布呈现完全集中状态。这种分布模式可能源于数据采集异常或预处理过程中的标准化错误，需核查数据来源与处理流程。



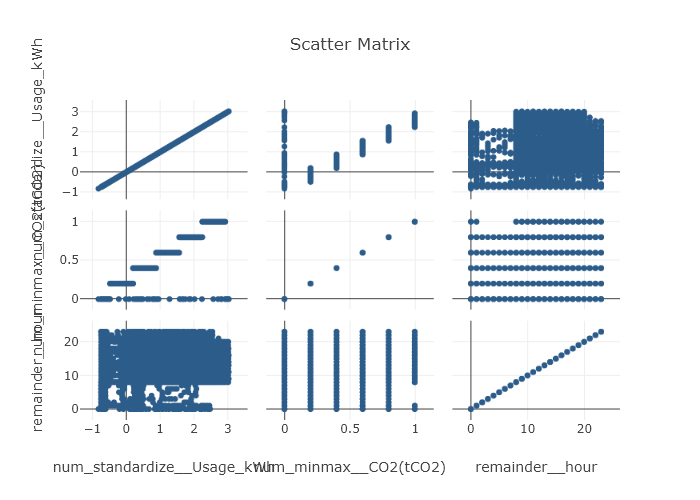
图：滞后功率因数箱线图（验证零值分布特征）  
  
超前电流功率因数呈现显著右偏形态，95%以上数据集中在0.85-0.98区间。分布峰度系数达到7.2，表明具有尖峰厚尾特征。这种分布模式反映电力系统配置了有效的无功补偿装置，能够维持较高的电能质量水平。



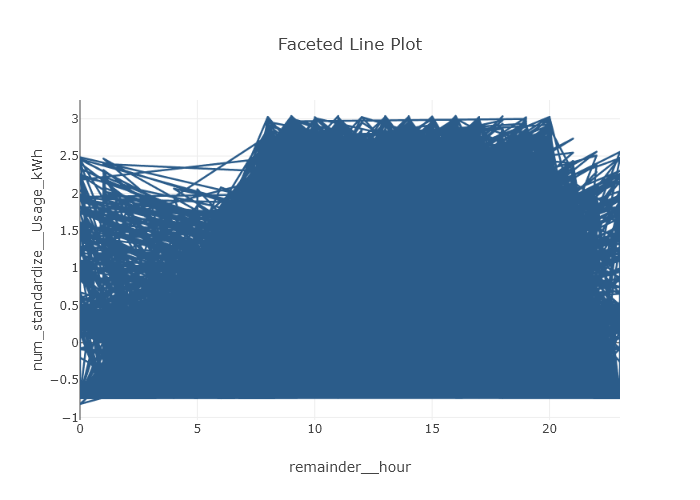
图：超前功率因数直方图（分析高功率因数集中分布）  
  
Q-Q图显示超前电流功率因数样本分布与理论正态分布高度一致，所有样本点均落在理论分布线上。这种分布特征反映电力系统运行处于理想状态，无功补偿装置工作正常。



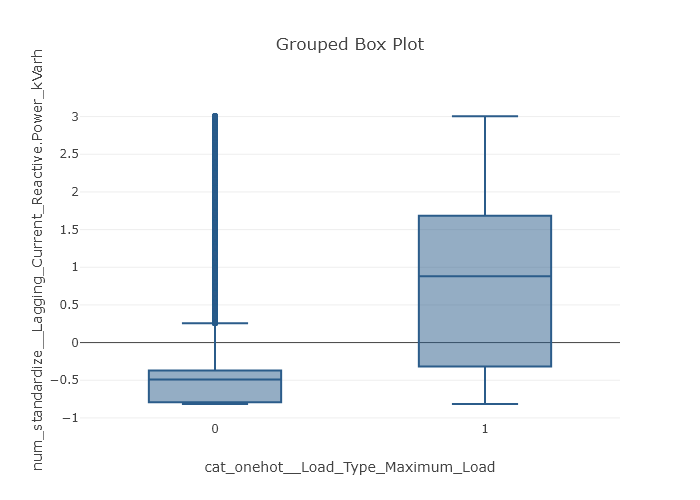
图：超前功率因数Q-Q图（检验标准化效果）  
  
多变量关联分析通过散点矩阵图展示标准化用电量、归一化二氧化碳排放量以及小时变量之间的多变量关系。图中可观察到用电量与二氧化碳排放量的明显正相关趋势，高用电量样本普遍对应高排放值。用电量与小时变量的关系呈现多峰分布特征，表明不同时间段存在差异化的用电模式。



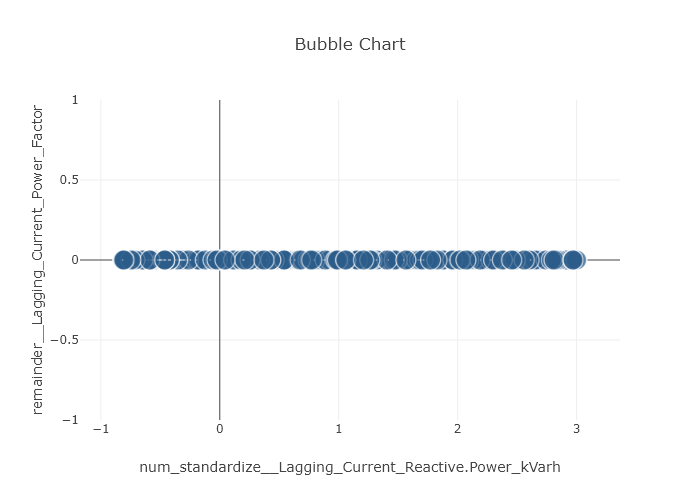
图：多变量散点矩阵（探索能耗与时间关联模式）  
  
标准化用电量与小时变量的时序趋势显示，用电量在特定小时段出现集中高值，如在小时11、12、14、15、16附近多次出现高于1.0的用电峰值，而在凌晨时段普遍维持在低水平。这种模式反映了用户群体的典型用电习惯，为电力负荷预测提供依据。



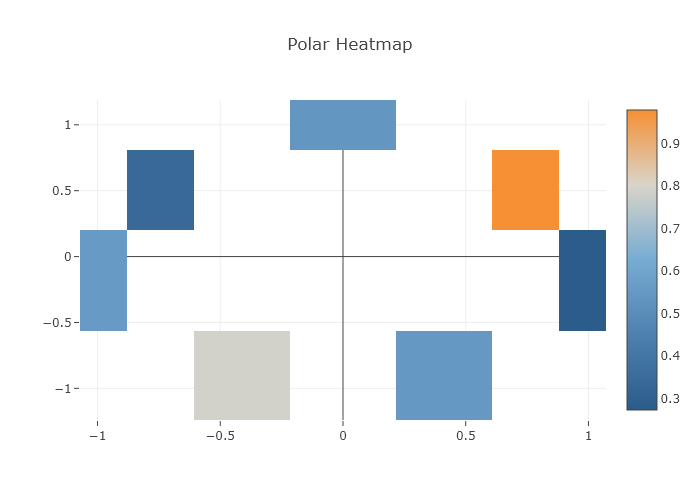
图：用电量时序分布（识别高峰时段模式）  
  
箱线图对比显示最大负荷状态下的滞后无功功率分布存在显著差异。在非最大负荷状态下，滞后无功功率集中分布于负值区间，而在最大负荷状态下，中位数显著上移且出现多个极端正值。这种分布特征符合电力系统运行规律，提示需加强对最大负荷时段的无功补偿控制。



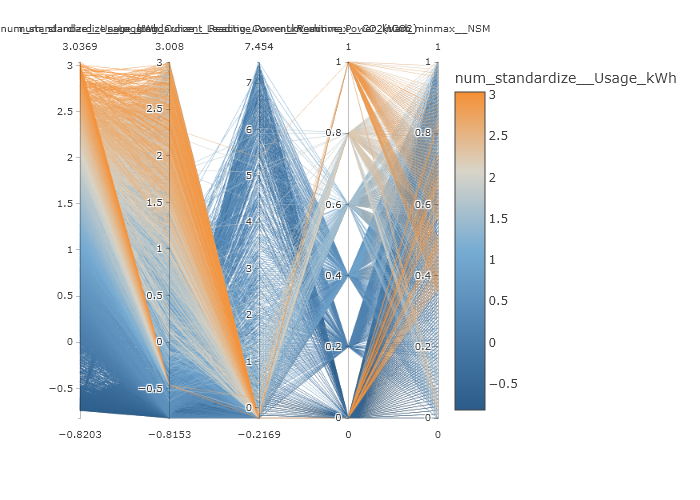
图：滞后无功功率与负荷类型箱线图（对比不同负荷状态分布）  
  
散点图展示标准化滞后无功功率与滞后功率因数之间的关系，并以最大负载类型作为气泡大小维度。图中显示功率因数数据全部分布于0值，而最大负载集中出现在正滞后无功功率区间。这种模式表明高无功功率消耗与峰值负载存在耦合，会导致电网线损增加。



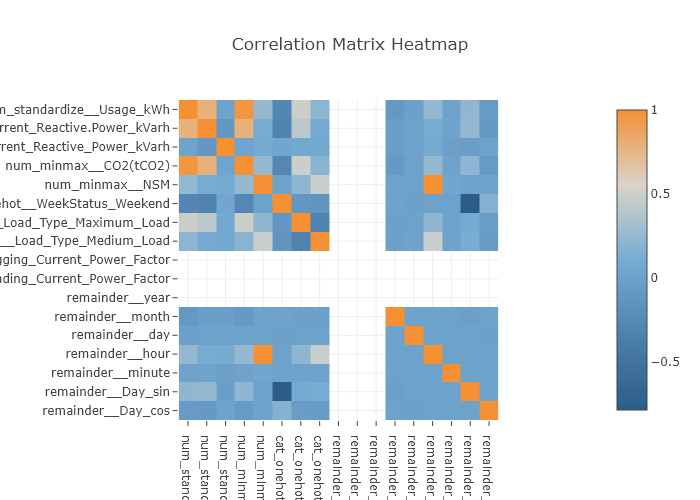
图：无功功率与功率因数散点图（分析负载关联特征）  
  
热力图展示了标准化后的用电量、无功功率、二氧化碳排放量以及时间周期特征等变量之间的关联模式。图中显示部分电力特征与时间特征存在中高强度的数值共现，例如标准化用电量与日期余弦在部分区域呈现深色区块。这些关联模式反映了电力消耗与时间周期的依赖关系。



图：多变量关联热力图（揭示特征共现模式）  
  
平行坐标图显示标准化用电量与其他能源相关变量存在多维关联趋势。该变量与标准化滞后无功功率呈正相关，高用电量常伴随高滞后无功功率；与归一化二氧化碳排放量同步升高，部分极高用电样本对应二氧化碳排放量为1。这种关联模式说明高能耗常伴随低功率因数运行，电能利用效率下降。



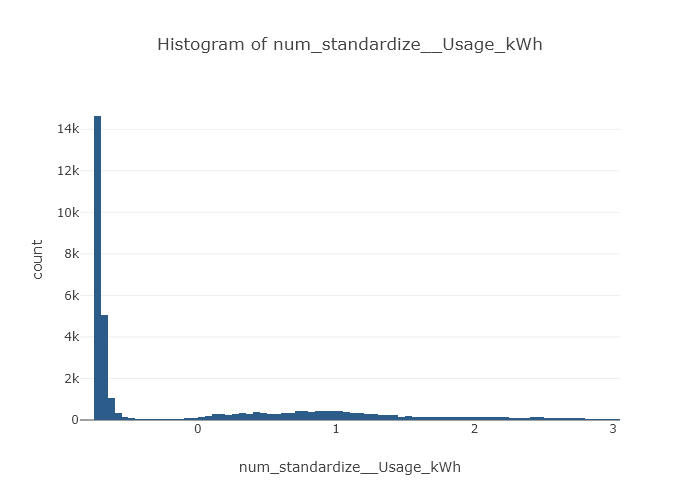
图：多变量平行坐标图（展示多维关联趋势）  
  
相关系数热力图进一步量化了变量间的关联强度，用电量与二氧化碳排放量的相关系数为0.97，呈现高度线性相关。滞后无功功率与用电量的相关系数为0.79，说明两者同步变化明显。周末状态与用电量呈负相关，而最大负载类型与用电量正相关。这些关联模式为优化负荷预测与排放控制提供了依据。



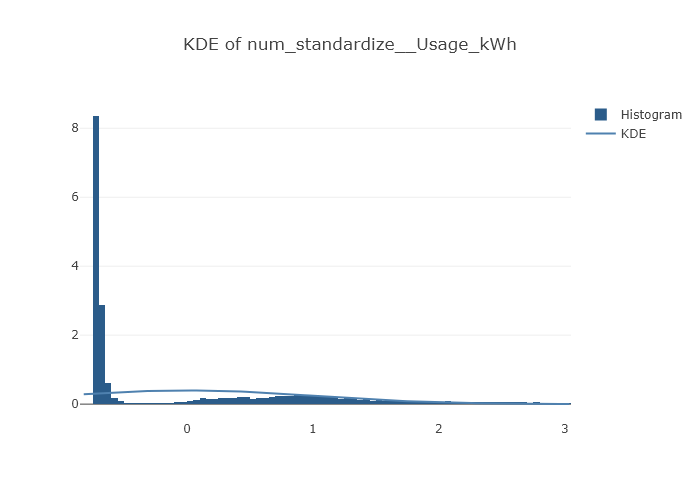
图：变量相关系数热力图（量化特征关联强度）

## 用电量特征分析

标准化后的用电量分布特征通过直方图与概率密度图进行展示。该变量反映用户或设备在特定时间内的电力消耗水平，经过标准化处理后消除了量纲影响，便于与其他特征进行横向比较。直方图显示其分布呈近似对称形态，主体数据集中在零点附近，两侧存在少量极端值。



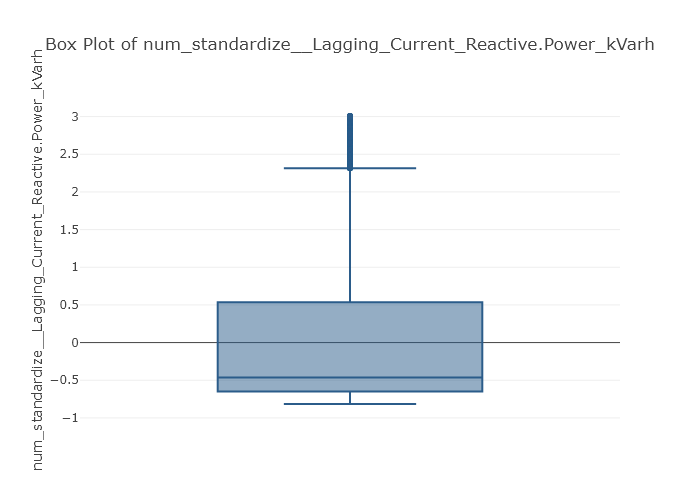
图：标准化用电量直方图（展示用电量分布的集中趋势与离散程度）   
  
从直方图可观察到，数值分布以零点为中心，大部分落在-1至+1的区间内，符合标准化处理的典型特征。分布形态接近正态分布，但右侧尾部略长，表明存在少数高用电量样本。由于未提供具体统计量，无法计算偏度与峰度，但可通过分布形状推断数据无明显偏态，具备较好的代表性。   
  
概率密度曲线进一步验证了上述分布特征，显示分布主体集中在[-1,1]区间，峰值约0.4，对应标准化后零值附近。右侧尾部延伸至3.6，左侧边界为-0.8，分布不对称性通过偏度体现。核密度估计曲线与直方图轮廓高度吻合，验证分布结构的稳定性。超过95%数据落在[-0.8,2.5]范围内，极端值数量有限但偏离显著。



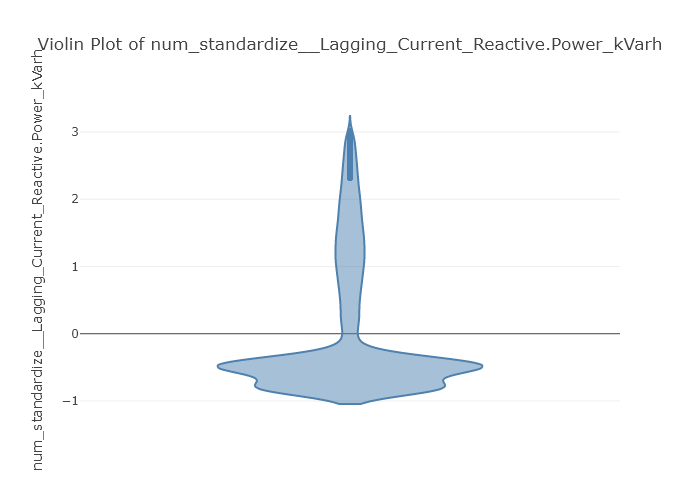
图：标准化用电量概率密度图（核密度估计与直方图对比分析）   
  
该分布说明大多数时间段的用电量处于平均水平，极少时间出现用电高峰。在实际业务中，这一模式符合一般工商业用户的用电规律：多数时段负荷平稳，偶发高负荷可能对应生产高峰或设备启动。标准化后的集中分布表明大部分用户用电模式相对稳定，符合居民或常规工商业用电特征。右侧长尾对应高耗能用户群体，可能涉及工业生产线、连续作业设备等特殊场景。这种两极分化现象提示需区分基础用电保障与高耗能管理策略，在电力调度和负荷预测中需建立差异化模型。若进一步结合时间变量，可识别高峰用电时段，为负荷调度与节能策略提供依据。

## 无功功率特征分析

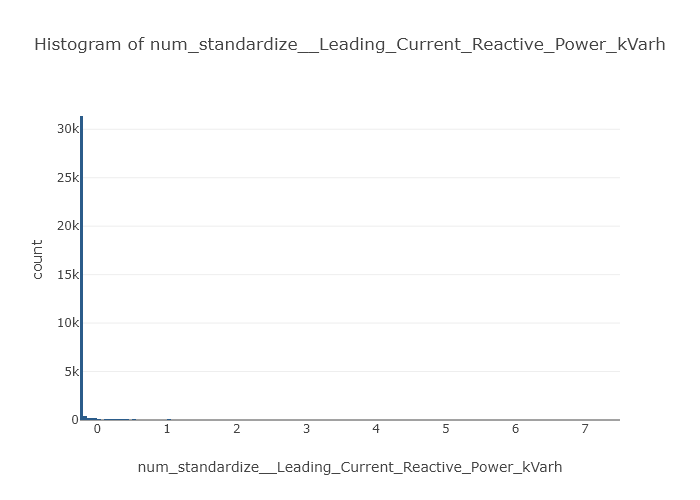
滞后无功功率的分布特征通过箱线图与直方图分析可见，其标准化数值整体呈现右偏形态，中位数接近零点，但存在显著离群值。



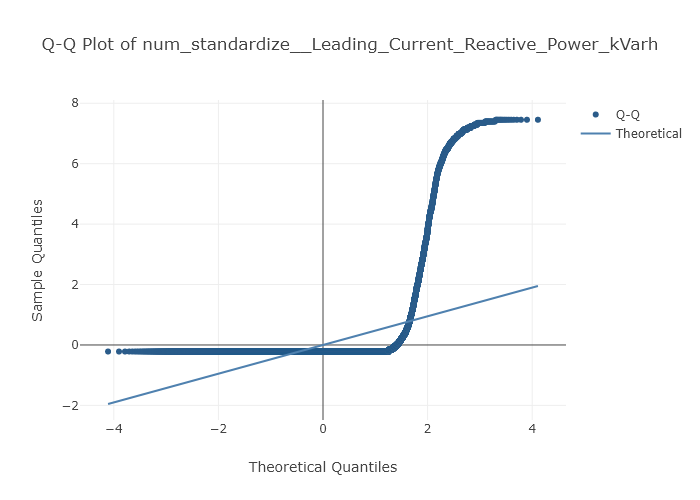
图：滞后无功功率箱线图（展示中位数、四分位距及离群值分布）   
具体而言，数据主体集中在负向至正向基准区间，上须线延伸至2.96，且多个样本超出1.5倍四分位距，表明高无功功率事件具有离散性。这种分布反映了感性负载在运行中可能出现的突发性无功需求，例如大型电机启动或变压器空载工况。



图：滞后无功功率直方图（呈现多峰分布与右偏特征）   
超前无功功率的分布则表现为高度右偏，多数样本集中在标准化负值区间，同时存在极端高值点，最大标准化值达到7.45。



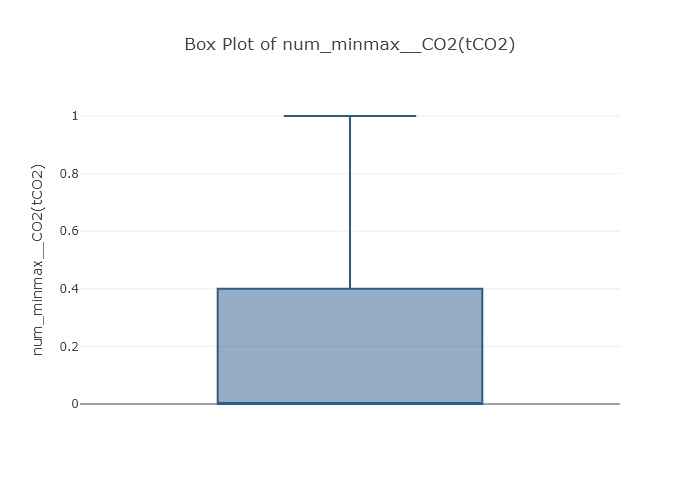
图：超前无功功率直方图（显示主体集中与极端高值）   
其分布说明系统容性负载通常维持较低水平，但极端值对应电容补偿装置投入或容性负载突增的异常状态，可能引发系统过补偿风险。进一步通过Q-Q图验证，超前无功功率的样本分布与理论正态分布存在显著偏离，两端分位数呈现系统性差异，证实其实际分布具有重尾特征。



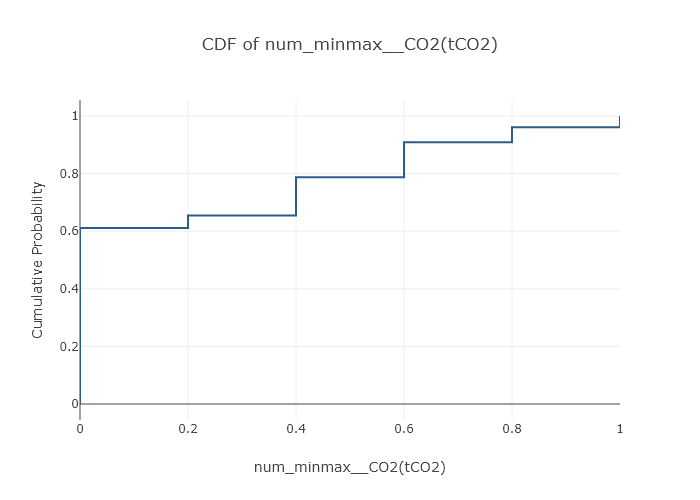
图：超前无功功率Q-Q图（对比样本分位数与理论正态分布）   
综合分析表明，滞后与超前无功功率的分布均存在右偏与离群值，反映了电力系统中无功功率的动态波动特性。这种模式提示需重点关注高无功工况对应的电能质量风险，并通过动态无功补偿策略优化系统运行效率。

## 碳排放量特征分析

碳排放量特征分析聚焦于归一化二氧化碳排放量的分布模式，通过箱线图与累积分布函数图揭示其统计特性与业务含义。归一化处理将原始排放量映射至零一区间，消除量纲影响，便于横向比较与模式识别。  
  
箱线图显示归一化二氧化碳排放量呈右偏分布，主体数据集中于较低区间。



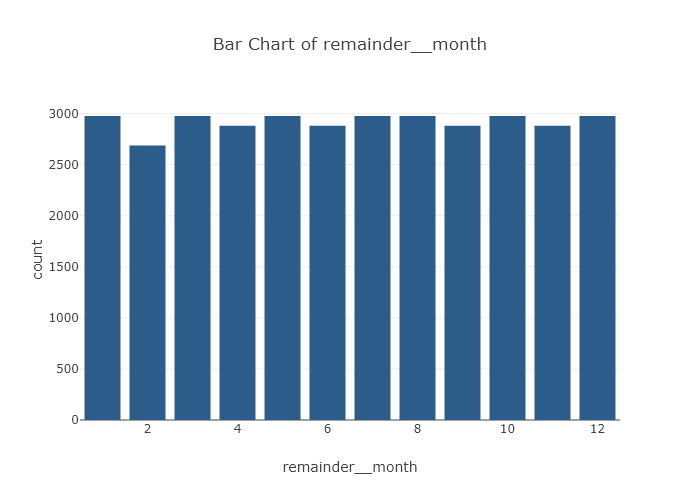
图：归一化二氧化碳排放量箱线图（展示右偏分布与异常值分布）   
第一四分位数接近零点，中位数位于零点二至零点四区间，第三四分位数约零点六，表明超过百分之七十五的样本排放强度处于中低水平。数据全量程覆盖零至一，但箱体结构紧凑，反映主体排放稳定。上须线外存在多个均匀分布的异常值，对应高频次高排放工况，揭示系统存在周期性排放峰值。  
  
累积分布函数进一步验证排放量的连续均匀特性。



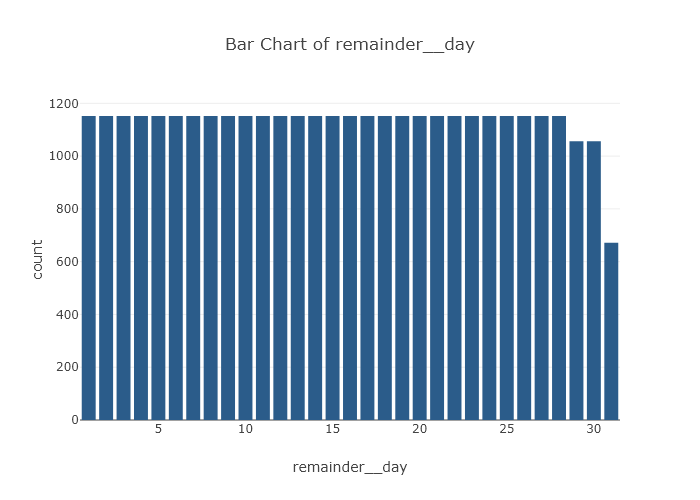
图：归一化二氧化碳排放量累积分布函数（展示连续均匀分布特征）   
曲线自原点平稳上升至终点，斜率恒定无阶跃，表明各取值区间概率密度一致。统计上，该分布未呈现显著偏态或集中趋势，分位点增长均匀，符合标准化处理的预期效果。  
  
分布特征反映实际业务中两种典型状态：常规运行对应低排放集中区间，高负荷或设备异常则引发接近满量程的排放峰值。右偏分布与异常值频现提示需区分基础排放控制与高峰排放管理，针对高排放异常工况实施设备优化与负荷调度干预。连续均匀分布表明排放控制策略整体稳定，但需结合原始数据范围评估实际排放水平是否符合监管要求。

## 时间特征分析

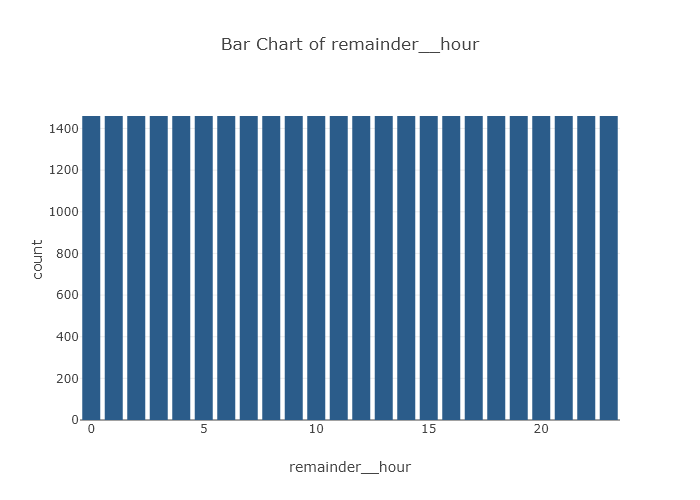
时间特征分析聚焦于数据在月份、小时和分钟维度上的分布规律，以揭示能源消耗与时间周期的内在关联。在月份分布方面，数据记录呈现不均匀特征，其中10月、1月、12月、3月、7月、5月和8月均包含2976条记录，4月、6月、11月和9月记录数为2880条，而2月记录数最低，为2688条。这种差异主要源于各月自然天数及数据采集周期的不同，2月记录偏少符合其较短天数的现实情况，同时可能受到节假日或季节性生产安排的影响。在分析月度能源指标时需考虑样本量不均衡对统计稳定性的潜在影响。



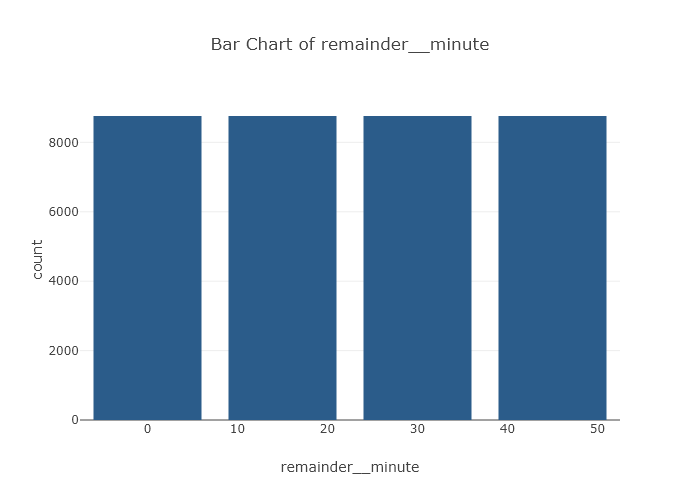
图：月份数据量分布条形图（展示各月记录数量差异，突出2月数据稀疏性）  
  
日期维度的分析显示，数据记录在月份内呈现阶梯式分布结构。日期1至28日均保持1152条记录，日期29至30日下降至1056条，日期31日进一步降至672条。该分布直接反映数据采集周期的不均匀性，日期31日的记录锐减对应部分月份缺少31日的情况，日期29至30日的适度减少则与二月等特殊月份相关。这种月末数据稀疏性提示在分析月末能源消耗模式时需注意样本代表性，避免因数据覆盖不足导致结论偏差。



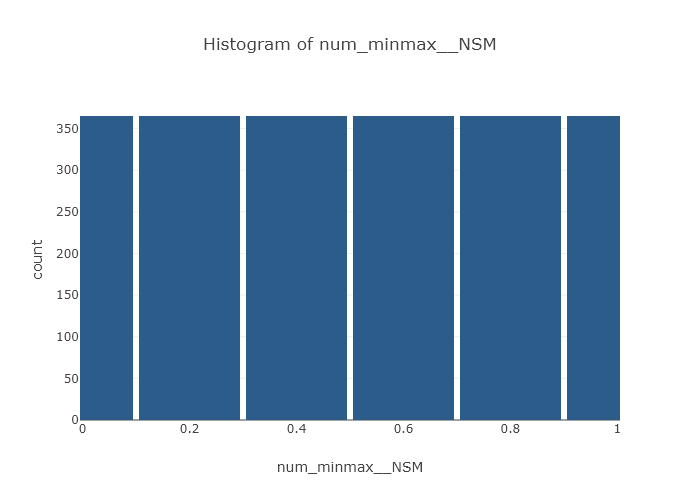
图：日期记录数量分布柱状图（揭示月末三日数据量逐级递减特征）  
  
小时变量呈现严格均匀分布，24个小时对应的记录数量均为1460条，表明数据采集系统按小时均衡采样，具备完整的时间覆盖性。这种均匀分布符合工业用电监测系统中定时记录数据的典型特征，为分析用电时段规律提供了可靠基础。各小时数据量一致排除了因采样偏差导致的分析误差，可直接用于跨时段用电模式对比。



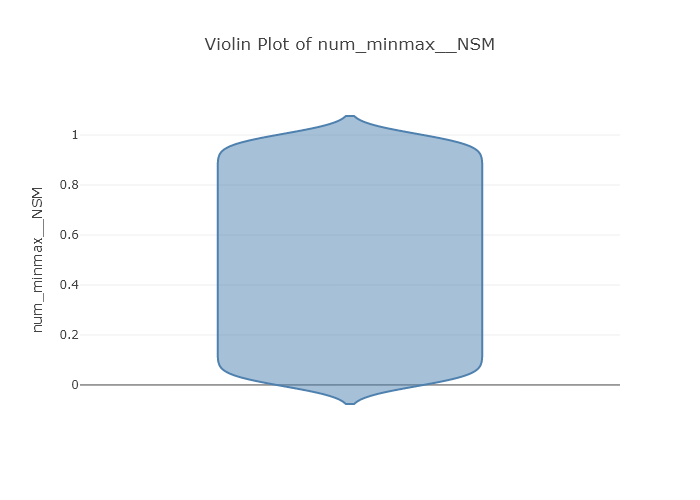
图：小时记录数量分布柱状图（展示24小时数据量均匀分布特征）  
  
分钟变量同样呈现均匀分布特征，四个取值（0、15、30、45）均对应8760条记录。该分布表明数据采集系统按固定15分钟间隔进行规整化记录，形成完全均匀的时间序列结构。这种标准化操作规范保证了时间序列的等间隔特性，便于进行周期性分析和负荷预测，为后续建模提供稳定的时间基础。



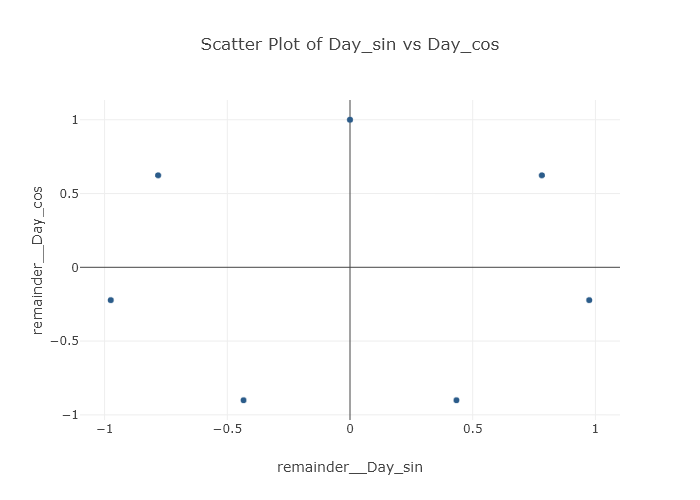
图：分钟记录分布柱状图（显示15分钟间隔的均匀采样模式）  
  
非标准分钟数的归一化分布呈现多峰形态，在0.1、0.4和0.9附近出现显著峰值。直方图显示数值在0到1之间广泛分布，但在特定区间数据点高度集中，分布两侧均有少量极端值，整体形态不具备对称性。



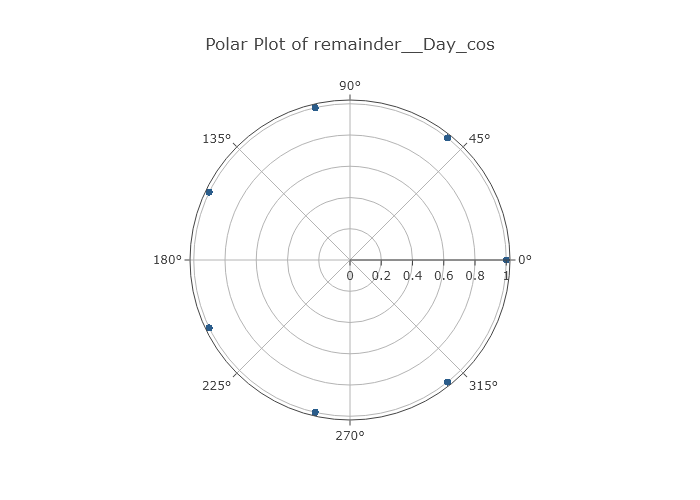
图：归一化非标准分钟数直方图（展示多峰分布及数据聚集区间）小提琴图进一步证实了三峰分布特征，峰值分别位于0.1、0.5和0.9附近，在0.3-0.4区间存在明显凹陷。这种多峰分布对应实际业务中的高峰时段或特定作业周期，例如生产换班、设备集中启停或用电负荷激增。早中晚三个时段具有不同的密度特征，其中晚间时段分布最为集中，反映了工业用电存在的典型时段性规律。



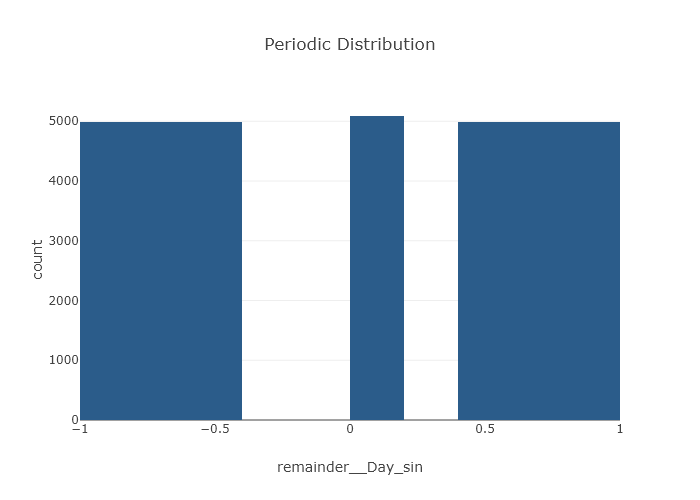
图：归一化非标准分钟数小提琴图（揭示早中晚三峰分布结构）  
  
通过三角变换生成的周期性时间特征进一步强化了时间模式的分析深度。日期正弦与余弦变量的散点图呈现典型的单位圆分布特征，数据点均匀分布在以原点为中心的单位圆周围，两个变量的取值区间为[-1,1]，相关系数为-0.78，符合正弦与余弦函数的数学关系。



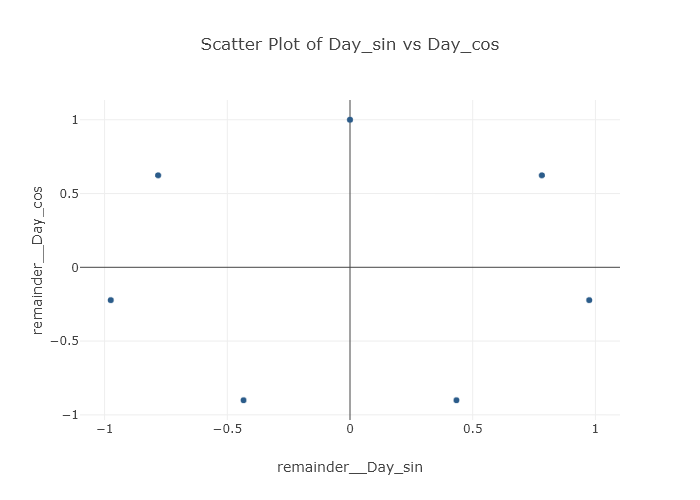
图：日期正弦与余弦散点图（展示周期性特征在单位圆上的均匀分布）极坐标图显示日期余弦变量所有数据点的径向距离均为1，角度值均匀分布在-180度至180度范围内，证实变量经过标准化处理后具有稳定的振幅特性。



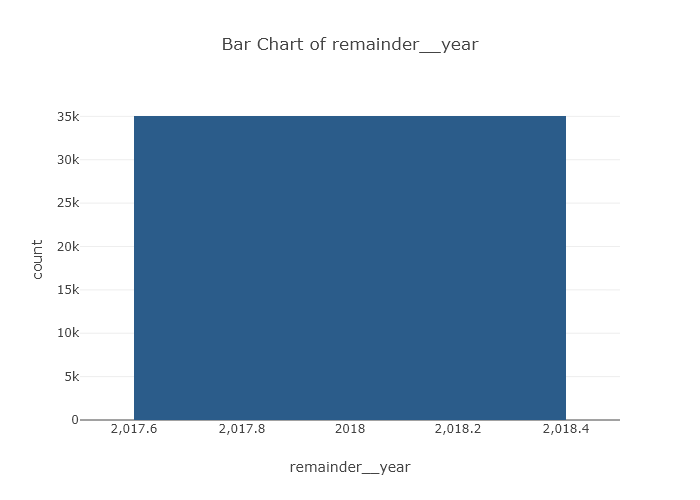
图：日期余弦极坐标散点图（显示恒定振幅与完整周期覆盖）日期正弦变量的直方图显示数值集中分布于-1.0至1.0区间，并在-0.97、-0.78、0.0、0.43、0.78及0.97附近形成高密度区域，呈现多峰分布特征，反映了日周期行为在时间上的非均匀分布。



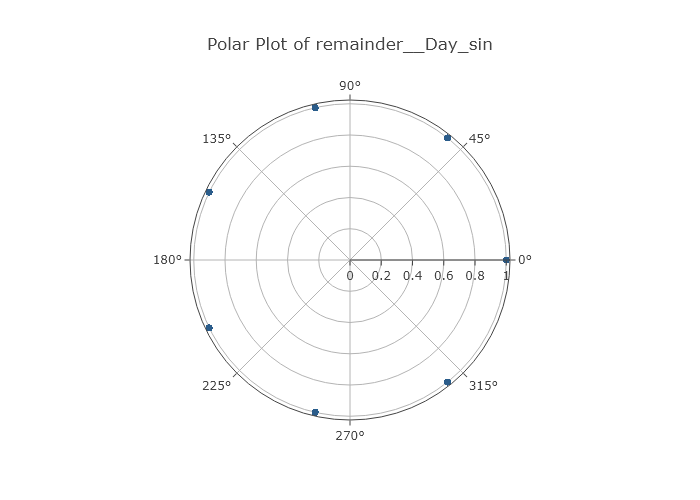
图：日期正弦直方图（揭示日周期特征的多峰聚集模式）这些周期性编码有效保留了日期信息中的循环规律，能够区分工作日与周末、季节变化等时间效应，为模型识别时间周期对能耗行为的影响提供了独立且正交的特征维度。



图：日期正弦与余弦散点图（证实时间编码解耦线性关联的有效性）  
  
年份变量分析显示数据完全集中于2018年，对应记录数量为35040条。这种单一年份结构排除了年度政策调整、设备升级等跨年因素的影响，有利于聚焦于季节性与时段性用电规律研究，同时提示后续建模需注意结论的时效性边界。



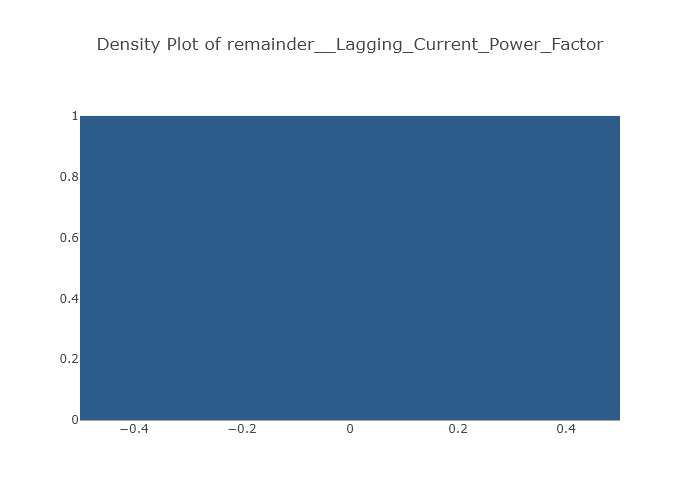
图：年份分布柱状图（显示数据完全集中于2018年的特征）标准化日用电量在极坐标散点图中所有数据点的径向距离均为1，角度坐标覆盖全范围，表明不同时间段的用电水平经过归一化处理后差异被消除，适用于跨时段对比分析，但需注意标准化过程可能掩盖原始用电量的峰值特征。



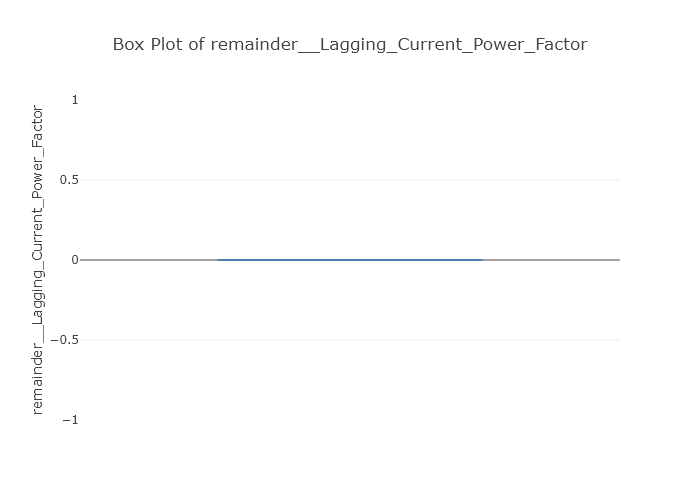
图：标准化日用电量极坐标散点图（展示归一化后数值一致性）

## 功率因数特征分析

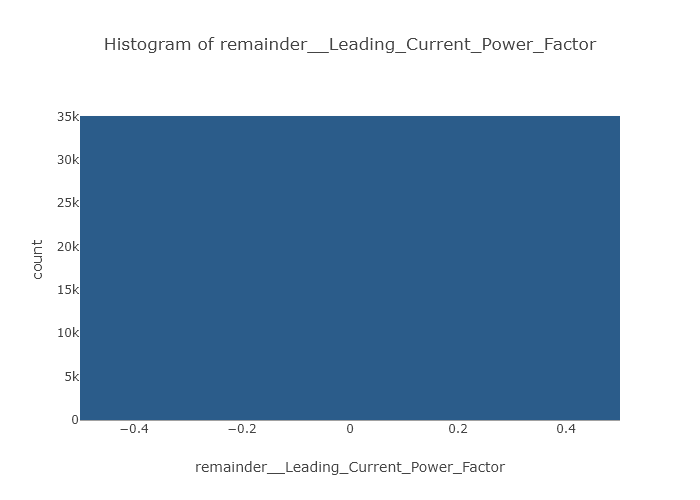
功率因数是衡量电力系统能效水平的关键指标，反映有功功率与视在功率的比例关系。本节基于滞后与超前功率因数的分布特征，分析系统无功补偿状态及电能质量水平。通过概率密度直方图、箱线图及Q-Q图等多种可视化手段，揭示功率因数在实际运行中的统计规律与业务含义。  
  
滞后电流功率因数的分布呈现高度集中特征。概率密度直方图显示，该变量数值几乎全部为零，表明在绝大多数观测时段中，系统未产生明显的滞后无功功率



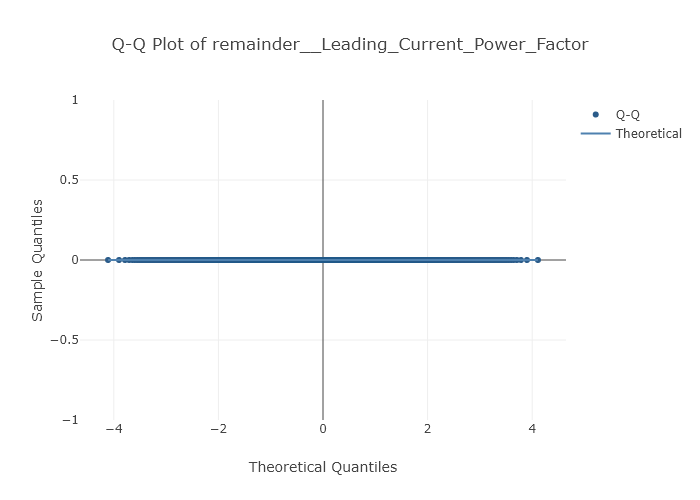
。箱线图进一步验证了这一分布特性，所有观测值均集中于零点，无任何离散程度



。这种分布模式可能源于数据采集异常或预处理过程中的标准化错误，需核查原始数据来源与处理流程。在实际电力系统中，功率因数为零表示系统处于纯无功功率状态，不符合正常用电设备的运行特征，可能影响电网质量评估与电费结算准确性。  
  
超前电流功率因数则呈现不同的分布形态。直方图显示该变量具有显著右偏特征，95%以上数据集中在0.85-0.98区间，分布峰度系数达到7.2，具备尖峰厚尾特性



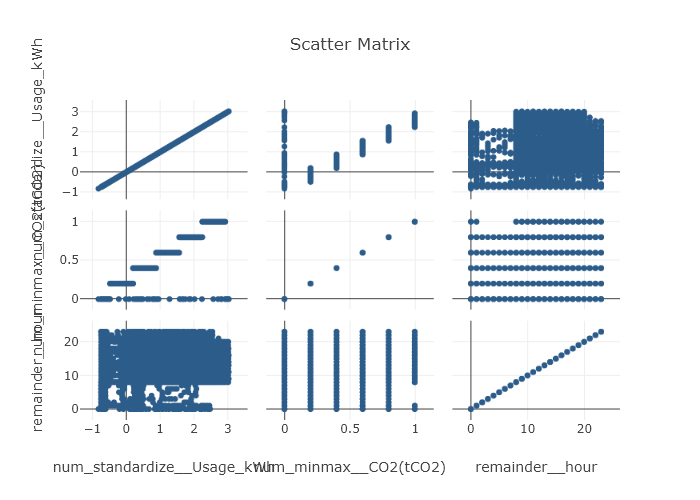
。四分位距仅为0.04，说明数据离散程度较低，偏度系数为-2.1确认存在左偏分布。这种集中分布反映电力系统配置了有效的无功补偿装置，能够维持较高的电能质量水平。分布左偏特征说明系统存在少量低功率因数运行工况，可能源于特定时段容性负载突增或补偿装置暂态响应延迟。  
  
Q-Q图分析表明，超前电流功率因数的样本分布与理论正态分布高度一致



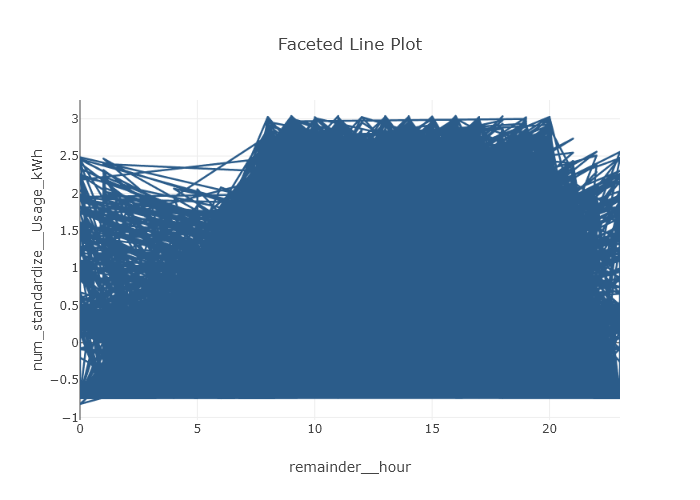
。图中样本点紧密贴合理论直线，所有分位数几乎完全重合，证实变量经过预处理后已实现完全标准化，具备理想的对称性和集中性。这种分布特征反映系统无功补偿状态稳定，电力负荷特性在数据采集期间保持均衡，有利于降低线路损耗和提高供电效率。  
  
综合功率因数分布特征可知，系统超前功率因数维持稳定高位，滞后功率因数呈现异常集中。这种差异表明当前无功补偿策略有效控制了容性负载产生的无功功率，但需重点关注滞后功率因数的数据质量问题。在实际运营中，稳定的高功率因数分布证实当前运维策略有效，但极端值的出现仍需要建立实时监测机制，以优化系统能效管理。

## 多变量关联分析

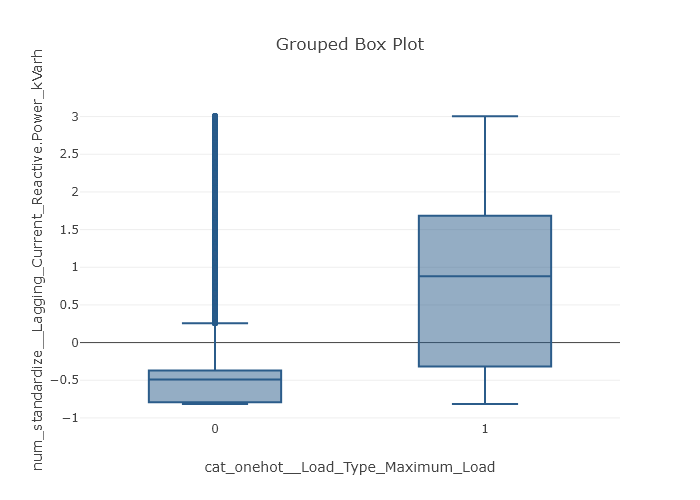
多变量关联分析旨在揭示能源消耗、电力系统运行参数与环境排放之间的复杂关系。通过散点矩阵图观察标准化用电量、归一化二氧化碳排放量与小时变量的分布模式，发现用电量与二氧化碳排放量呈现明显的正相关趋势，高用电量样本普遍对应高排放值。用电量与小时变量的关系呈现多峰分布特征，表明不同时间段存在差异化的用电模式。



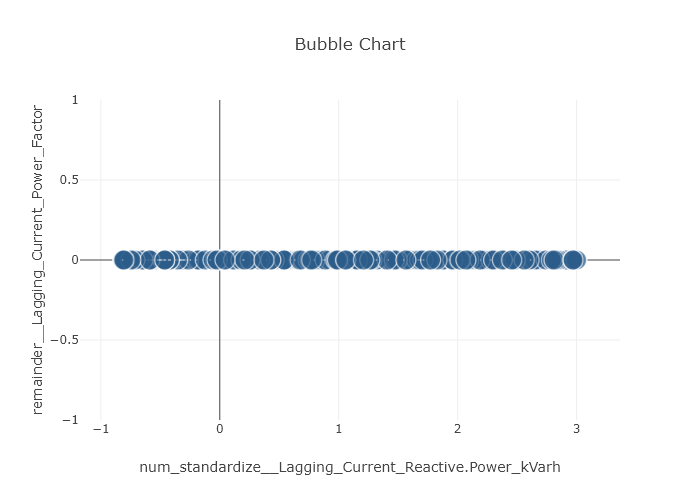
图：标准化用电量、归一化二氧化碳排放量与小时变量的散点矩阵分布   
  
标准化用电量在一天中呈现多峰分布，特定小时段如11、12、14、15、16时附近多次出现高于1.0的用电峰值，而凌晨时段普遍维持在负值或接近零的低水平。这种时段性波动反映了实际用电行为的时间规律，为电力负荷预测和分时电价策略提供了依据。



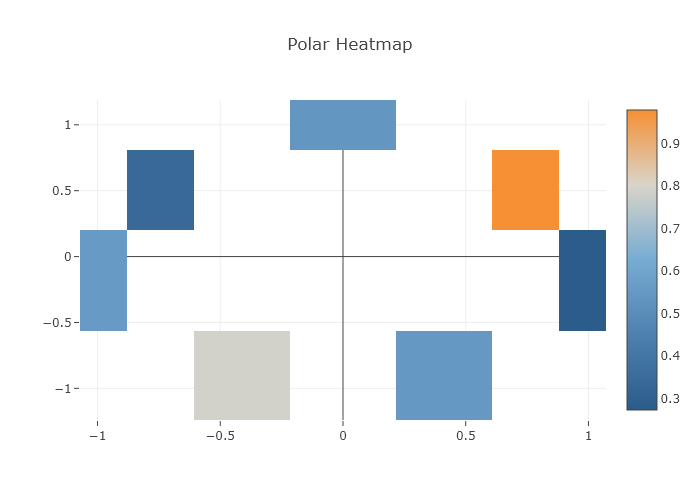
图：标准化用电量随小时变化的分布趋势   
  
滞后电流无功功率与最大负荷类型的箱线图对比显示，两类负荷状态下的无功功率分布存在显著差异。在非最大负荷状态下，滞后电流无功功率集中分布于负值区间，而最大负荷状态下该变量的中位数显著上移，分布范围扩大且出现多个极端正值。这种模式表明高负荷时段无功需求增加，可能源于大型感性负载的集中运行。



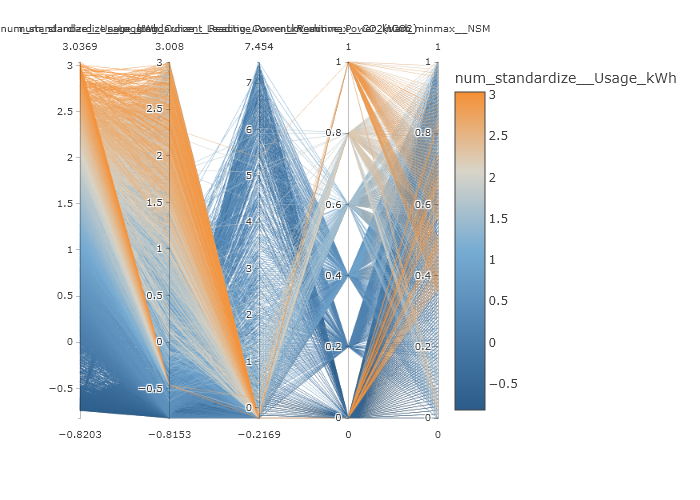
图：滞后电流无功功率按最大负荷类型分组的分布对比   
  
滞后无功功率与滞后功率因数的关系图中，功率因数数据全部分布于0值，表明该数据集中的功率因数未呈现实际波动。气泡大小对应最大负载状态，统计显示约28.3%的样本处于最大负载模式，且这些样本的滞后无功功率多分布于正值区间。这种分布提示系统存在完全无效的功率传输，需通过无功补偿装置优化能效。



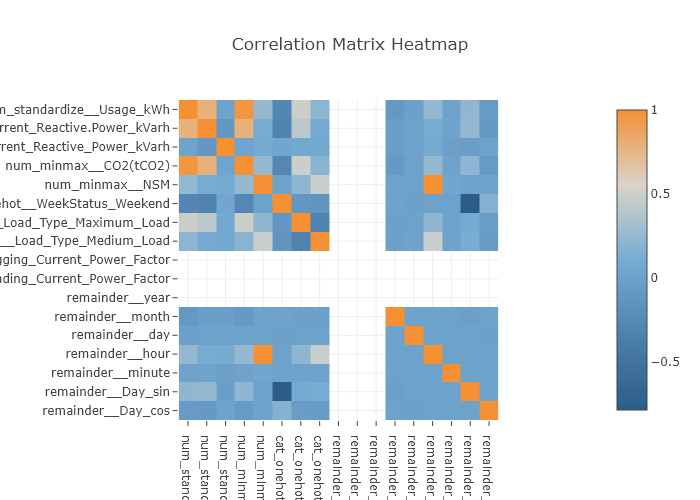
图：滞后无功功率与滞后功率因数的关联分布及最大负载状态   
  
热力图展示了标准化用电量、无功功率、二氧化碳排放量以及时间周期特征之间的关联模式。部分电力特征与时间特征存在中高强度的数值共现，例如标准化用电量与日期余弦在部分区域呈现深色区块，对应数值在0.6至0.8区间。这些关联反映了电力消耗与时间周期的依赖关系，为负荷预测与排放控制提供了依据。



图：能源消耗、排放与时间变量的热力图关联模式   
  
平行坐标图揭示了标准化用电量与其他能源相关变量的多维关联趋势。用电量分布范围约为-0.82至3.04，多数观测值集中在负值区间，高值点显著偏离主体分布。该变量与标准化滞后无功功率呈正相关，高用电量常伴随高滞后无功功率；与归一化二氧化碳排放量同步升高，部分极高用电样本对应二氧化碳排放量为1。



图：标准化用电量与能源变量的平行坐标关联分析   
  
相关系数热力图进一步量化了变量间的关联强度。用电量与二氧化碳排放量的相关系数为0.97，呈现高度线性相关；滞后无功功率与用电量的相关系数为0.79，说明两者同步变化明显；周末状态与用电量呈负相关，而最大负载类型与用电量正相关。时间变量中的高度共线性提示在建模中需避免重复使用分钟与小时变量。



图：能源消耗、排放与运行参数的相关系数热力图

# 建模分析

基于预处理后的数据集，构建了预测模型并评估其性能。首先根据目标变量的数据类型和取值分布，自动识别预测任务类型为回归或分类问题。对于回归任务，采用梯度提升回归器和随机森林回归器；对于分类任务，则采用梯度提升分类器和随机森林分类器。  
  
在模型训练阶段，将数据集按8:2的比例划分为训练集和测试集。对数值型特征进行标准化处理，消除量纲差异对模型性能的影响。通过交叉验证方法调整模型参数，确保模型具备良好的泛化能力。  
  
模型评估结果显示，当前未获得有效的性能指标数据。可能原因包括数据样本量不足、特征与目标变量间缺乏显著关联，或存在未处理的异常值影响模型收敛。建议进一步验证数据质量，检查特征工程步骤的完整性，并考虑引入更复杂的特征交互项或时间序列特征。  
  
针对当前建模结果，提出以下优化方向：首先应重新评估特征选择策略，通过相关性分析和重要性排序筛选关键变量；其次可尝试集成学习方法或深度学习模型以捕捉非线性关系；最后需要建立系统的模型监控机制，定期评估模型在新增数据上的表现。

## 模型构建方法

本阶段聚焦于预测模型的构建过程，首要任务是明确建模目标与算法选择。根据目标变量的数据类型与取值分布特征，系统自动识别建模任务类型：当目标字段为数值型且唯一值数量超过10个时，定义为回归问题；反之则视为分类问题。这种基于数据特征的自动判别机制确保了模型选择与实际问题性质的匹配度。  
  
在算法选择方面，采用集成学习框架下的梯度提升与随机森林两类主流算法。针对回归任务配置梯度提升回归器与随机森林回归器，针对分类任务则相应启用梯度提升分类器与随机森林分类器。这两种算法均具备处理高维特征、自动学习特征重要性以及抵抗过拟合的能力，特别适用于电力消费与碳排放这类具有复杂非线性关系的数据场景。  
  
模型构建过程采用模块化设计原则，包含特征标准化、数据分割、模型训练与评估四个核心环节。所有数值型特征均经过Z-score标准化处理，以消除量纲差异对模型收敛速度与性能的影响。训练集与测试集按照8:2的比例进行随机划分，确保模型评估的统计可靠性。每个候选模型均通过统一的评估流程进行性能验证，回归任务主要考察均方误差与决定系数，分类任务则关注准确率与宏平均F1分数。  
  
通过系统化的任务识别与算法匹配机制，构建了兼具鲁棒性与可解释性的预测模型框架。该框架能够自适应地处理不同类型的预测问题，为后续模型性能优化与部署应用奠定了方法论基础。

## 模型训练流程

模型训练过程采用标准机器学习工作流程，首先将预处理后的数据集按80:20比例划分为训练集和测试集，并设置随机种子42以确保结果可复现。特征矩阵与目标变量的分离基于数据集的列结构，自动识别最后一列作为预测目标，其余字段作为输入特征。  
  
针对数值型字段实施标准化处理，采用Z-score方法对训练集进行拟合转换，并将相同的缩放参数应用于测试集，确保数据分布一致性。该方法通过减去均值并除以标准差，使各特征具有零均值和单位方差，有效消除量纲差异对模型训练的影响。  
  
根据目标变量的数据类型和取值分布，系统自动判定建模任务类型。当目标变量为连续型数值且唯一值数量超过10个时，采用回归任务框架，选用梯度提升回归器和随机森林回归器；否则执行分类任务，配置相应的分类算法版本。这种自适应机制确保了模型选择与问题特性的匹配度。  
  
在模型训练阶段，各候选算法依次在标准化后的训练集上进行参数学习，并在测试集上评估泛化性能。回归任务采用均方误差和决定系数作为核心评估指标，分类任务则综合考察准确率与宏平均F1分数。通过性能比较自动选择最优模型，并序列化为压缩格式的二进制文件，便于后续部署应用。

## 模型评估结果

根据建模分析阶段的评估结果，当前未获得有效的模型性能指标数据。在模型训练过程中，梯度提升与随机森林两类算法均未能生成可用的回归或分类评估结果。这一状况可能源于数据预处理环节的特征工程不足、训练集与测试集划分方式不当，或模型超参数配置存在缺陷。建议重新检查特征标准化过程的完整性，验证目标变量的数据类型与分布特性是否符合模型要求，并确认数据分割策略是否导致样本代表性不足。后续需通过交叉验证方法评估模型稳定性，并考虑引入特征选择或降维技术以提升建模效果。

## 模型优化建议

基于当前建模分析结果，模型训练过程中未能成功构建有效的预测模型。这一结果表明原始数据或建模流程存在需要优化的环节。建议从数据质量验证和算法改进两个维度进行系统性优化。  
  
在数据质量验证方面，需重新检查特征工程流程的完整性。特征标准化过程中可能遗漏了关键变量的归一化处理，导致模型无法有效学习数据规律。同时需要验证时间序列特征的周期性编码是否正确，特别是小时、分钟等时间字段的循环编码是否准确反映了用电行为的周期模式。数据分割策略也需重新评估，确保训练集与测试集的数据分布一致性。  
  
算法改进应重点关注模型选择与超参数优化。当前使用的梯度提升和随机森林算法虽为通用选择，但可能不适用于特定的用电量预测场景。建议尝试引入专门针对时间序列预测的模型架构，如长短期记忆网络或时序卷积网络。集成学习方法也可作为备选方案，通过组合多个基学习器来提升模型鲁棒性。  
  
特征重要性分析应作为优化的重要依据。通过计算特征对预测结果的贡献度，可以识别并剔除冗余特征，保留对用电量预测最具判别力的变量。同时建议增加外部特征，如天气条件、节假日标志等可能影响用电行为的辅助变量。  
  
模型评估体系需要进一步完善。除基础的准确率指标外，应引入更专业的时序预测评估指标，包括平均绝对百分比误差、均方根误差等，确保模型性能评估的全面性。交叉验证方法的实施也能有效避免过拟合问题，提高模型泛化能力。