

# 电力需求数据集探索性数据分析与特征工程

[github.com/SakuraPuare/ElectricityDemand](https://github.com/SakuraPuare/ElectricityDemand)

廖嘉旺

Hubei University of Arts and Science

2025 年 4 月 23 日

# 目录

- 1 引言
- 2 数据集概览与质量
- 3 分析环境与方法
- 4 探索性数据分析 (EDA)
- 5 数据整合与频率匹配
- 6 特征工程
- 7 总结与下一步

- 电力需求预测的重要性
- 本报告目标：
  - 数据集理解与探索性分析 (EDA)
  - 数据质量评估
  - 数据整合与特征工程构建预测特征集

# 主要数据文件

该数据集包含三个主要文件：

- **电力需求数据 (data/demand.parquet):**
  - unique\_id: 仪表的唯一 ID
  - timestamp: 本地时间记录周期的开始时间戳
  - y: 当前时段的用电量 (kWh)
- **元数据 (data/metadata.parquet):**
  - 包含仪表信息: unique\_id, dataset, building\_id, location\_id
  - 地理信息: latitude, longitude, location, timezone
  - 建筑信息: freq, building\_class(住宅/商业), cluster\_size
- **天气数据 (data/weather.parquet):**
  - 基础信息: location\_id, timestamp
  - 主要天气变量: 温度、湿度、降水、风速、云量等

# 规模、缺失、重复与时间范围

## 1. 数据量:

- Demand: 2.38 亿条
- Metadata: 7572 条
- Weather: 60.5 万条

## 2. 缺失值:

- Demand: y ( 1.3%) 缺失
- Metadata: 位置信息 ( 3.1%) 缺失
- Weather: 无缺失

## 3. 重复值:

- Demand/Metadata: 未发现基于关键列的重复
- Weather: 极少量重复 (基于 location\_id, timestamp, 已处理)

## 4. 时间范围:

- Demand: 2011-01-01 ~ 2017-12-31
- Weather: 2011-01-01 ~ 2019-01-01 (覆盖需求数据)

## 技术栈:

- Apache Spark 3.5.0 & PySpark - 大规模数据处理
- Pandas/NumPy - 数据分析与处理
- Matplotlib/Seaborn - 可视化
- Jupyter Notebook - 交互式开发
- Loguru / Log Utils - 日志记录

## 计算环境:

- 96 核心 CPU, 196GB RAM
- 腾讯云服务器
- 运行时间: 约 10 小时

## 需求数据详细统计 (基于完整数据):

- 总记录数: 237,944,171 条
- y 非空记录: 234,857,893 条
- y 缺失记录: 3,086,278 条 (1.30%)
- y 非正值 ( $\leq 0$ ): 2,499,640 条 (占非缺失值的 1.06%)

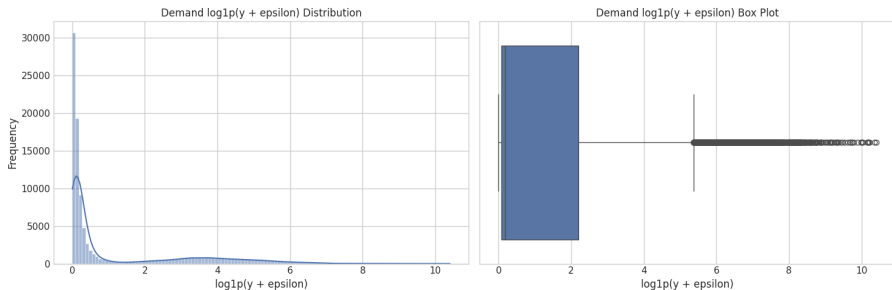
## 需求值分布统计:

- 均值 (Mean): 44.90 kWh
- 标准差 (Std): 394.25 kWh
- 最小值 (Min): 0.00 kWh
- 中位数 (Median): 0.20 kWh
- 75% 分位: 7.62 kWh
- 最大值 (Max): 221,228.00 kWh
- 多重周期性:
  - 日内周期 (Daily): 白天高, 夜间低。
  - 周内周期 (Weekly): 工作日与周末模式差异。
  - 年度周期 (Annual): 季节性波动 (如冬夏高峰)。
- 其他特性: 波动性、异常值、不同用户模式多样性。



# 分布形态

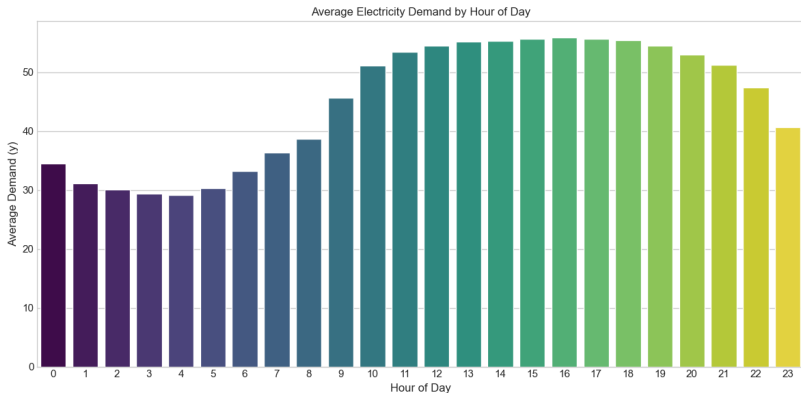
高度右偏，存在极端高值。Log1p 变换有助于改善对称性。



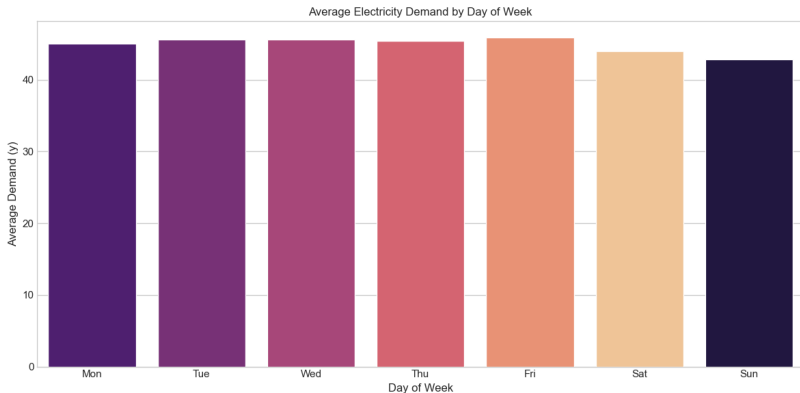
图：需求值  $\log_{1p}$  变换分布

# 时间序列样本示例

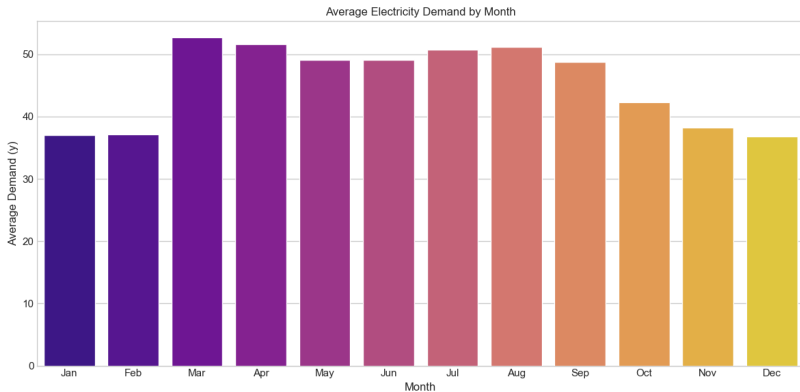
## 1. 小时周期性:



## 2. 日周期性:



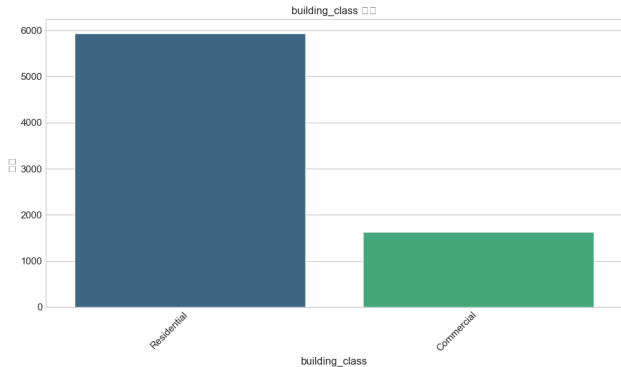
## 3. 月周期性:



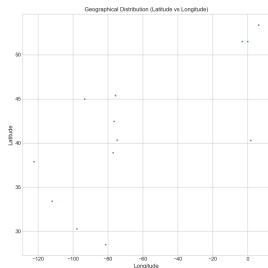
- **建筑类型 (Building Class):** 以 Residential (住宅) 为主 ( 78%), Commercial (商业) 占比较少 ( 22%)。
- **位置 (Location):** 大部分集中在 London, UK ( 74%), 少量分布在葡萄牙 (PT) 和华盛顿特区 (Washington DC)。存在缺失位置信息 ( 3.1%)。
- **采样频率 (Freq):** 主要频率包括 30 分钟 (30T, 73%), 1 小时 (1H, 21%) 和 15 分钟 (15T, 5%)。

# 主要天气特征分布与时间特性

## 主要分类特征分布:



图：建筑类型分布

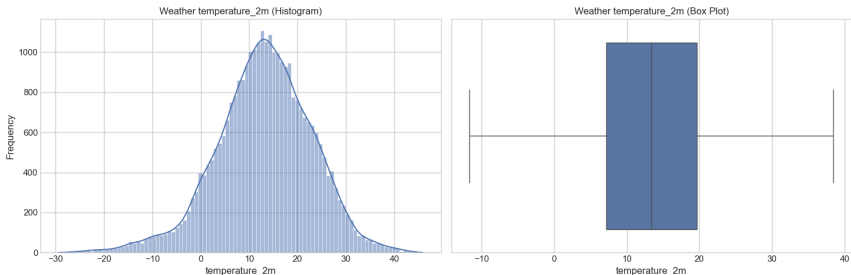


图：监测点地理位置

# 天气数据分析

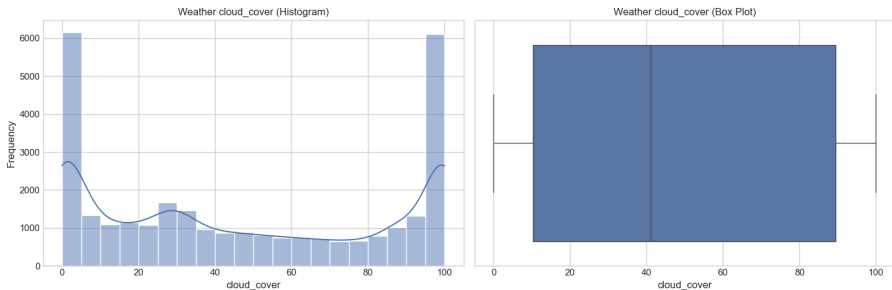
- **数值特征分布:** 温度 (temperature\_2m) 近似正态分布, 相对湿度 (relative\_humidity\_2m) 分布较广, 降水 (precipitation) 存在大量零值 (零膨胀), 风速 (wind\_speed\_10m) 呈右偏分布。
- **时间特性:** 天气数据主要以 1 小时为采样频率, 记录规整。

主要天气特征分布示例:



图：温度分布

## 主要天气特征分布示例:



图：总云量分布



# 建筑类型与天气的影响

## 1. 需求与元数据关系 (建筑类型):

- Commercial (商业) 建筑的电力需求量级通常显著高于 Residential (住宅), 且波动更大。

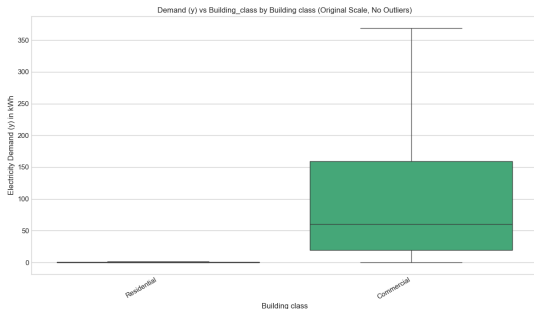


图: 需求 vs 建筑类型 (原始尺度)

## 2. 需求与天气关系 (基于抽样合并):

- 与 Temperature 呈弱正相关 ( 0.03)。
- 与 Relative Humidity 呈中度负相关 ( -0.20)。
- 天气特征内部存在相关性, 如温度与体感温度、云量特征之间。

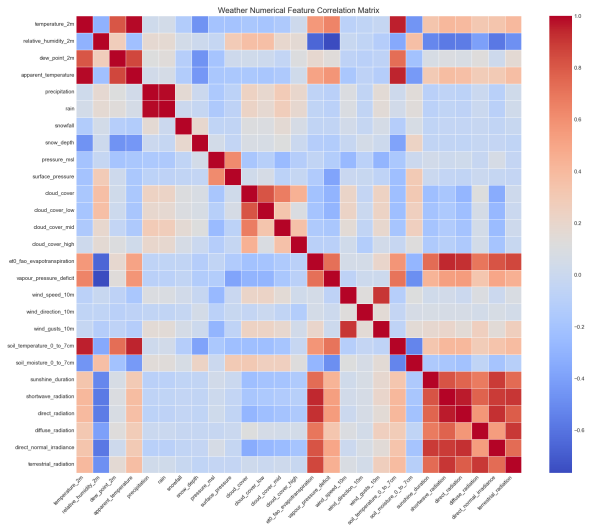
# 多源数据合并步骤

- ❶ **加载与初步处理:** 加载 Demand, Metadata, Weather 数据。对 Demand 进行初步清洗（处理缺失和非正值）。
- ❷ **需求数据频率匹配:** 将不同频率的 Demand 数据重采样并聚合到统一的 **小时 (1H)** 频率，以匹配 Weather 数据。对于  $\text{freq} < 1\text{H}$  的数据，进行求和或平均聚合；对于  $\text{freq} > 1\text{H}$  的数据，考虑插值或保持原样（本项目聚合到小时）。
- ❸ **Weather 数据处理:** 清理 Weather 数据中的少量重复记录，确保每个 `location_id` 在每个小时点只有一条记录。
- ❹ **数据合并:**
  - 将重采样后的 Demand 数据与 Metadata 数据通过 `unique_id` 进行左连接。
  - 将上一步的结果与处理后的 Weather 数据通过 `location_id` 和小时级 `timestamp` 进行左连接。

# 合并成功率与天气特征相关性

- **合并成功率:** 约 1.73% 的需求记录因元数据中缺少有效的 location\_id 或无法在天气数据中找到匹配的时间点, 未能成功关联天气信息。其余数据 ( 98.27%) 成功合并。
- **天气特征内部相关性:** 合并后的数据集中, 天气特征之间存在显著相关性, 如下图所示。这在特征选择或模型选择时需要注意 (如多重共线性)。

# 数据整合结果诊断



图：天气特征相关性矩阵

基于合并后的数据集，我们构建了以下类型的预测特征：

## 1. 时间特征：

- 从 timestamp 中提取：年 (year)，月 (month)，日 (day)，星期几 (dayofweek)，年内天 (dayofyear)，小时 (hour)。
- 考虑循环特征编码（如使用 sin/cos 转换小时和星期几）。

## 2. 滚动窗口统计特征：

- 基于历史电力需求 (y)。
- 在每个 unique\_id 的时间序列上计算。
- 考虑不同窗口大小（例如，过去 3H, 12H, 24H, 168H）。
- 计算统计量如：均值 (mean\_lag\_Xh)，标准差 (std\_lag\_Xh)，最小值 (min\_lag\_Xh)，最大值 (max\_lag\_Xh)。

## 3. 原始/合并特征:

- 来自 Metadata: building\_class, location\_id, freq 等 (需进行编码)。
- 来自 Weather: temperature\_2m, relative\_humidity\_2m, apparent\_temperature 等。

**缺失值处理:** 移除了目标变量 y 缺失的行; 滚动特征计算初期产生的缺失值也需处理 (例如, 移除或使用插补)。

**输出:** 最终特征集按年/月分区存储为 Parquet 文件 (data/features.parquet)。

# 主要发现回顾

- **数据特性:** 规模大, 包含需求、元数据、天气多源信息; 需求分布高度右偏, 存在非正值和异常值。
- **数据质量:** 存在少量 y 缺失和元数据位置信息缺失, 天气数据有少量重复 (已处理)。
- **关系:** 建筑类型 (商业需求显著高于住宅) 和天气 (温湿度) 与电力需求存在关联。天气特征内部有相关性。
- **时间模式:** 需求表现出清晰的日、周、年度周期性。不同数据源的时间频率不匹配已通过重采样到小时频率解决。
- **处理过程:** 利用 Spark 有效处理了大规模数据; 通过抽样和可视化进行了深入的 EDA; 成功整合了异构数据源并构建了初步的时间、滚动、原始特征集。



使用了基于小时频率聚合并进行特征工程的数据集 (data/features.parquet) 训练了两个 Spark MLlib 回归模型，并在测试集上进行了评估 (基于时间分割)。

## 1. MLlib 线性回归 (Linear Regression)

- 测试集 RMSE: 73.81
- 测试集 MAE: 5.86

## 2. MLlib GBT 回归 (Gradient Boosted Trees Regression)

- 测试集 RMSE: 175.40
- 测试集 MAE: 54.04

## GBT 模型特征重要性 (Top 10):

- y\_rolling\_max\_3h (近期最大需求): 0.1882
- y\_rolling\_stddev\_48h (过去 2 天需求波动): 0.1161
- y\_rolling\_stddev\_6h (近期需求波动): 0.0969
- hour (小时): 0.0946
- y\_rolling\_min\_48h (过去 2 天最小需求): 0.0908
- y\_rolling\_min\_168h (过去 1 周最小需求): 0.0888
- y\_rolling\_stddev\_3h (极近期需求波动): 0.0519
- soil\_temperature\_0\_to\_7cm (土壤温度): 0.0470
- y\_rolling\_stddev\_24h (过去 1 天需求波动): 0.0394
- year (年份): 0.0338

## 初步结论与解读:

- **模型性能对比:** 出乎意料地, 简单的线性回归模型在测试集上的表现 ( $RMSE=73.81$ ,  $MAE=5.86$ ) 显著优于梯度提升树回归模型 ( $RMSE=175.40$ ,  $MAE=54.04$ )。这表明当前的 GBT 模型可能存在调优不足、过拟合或特征处理等问题, 需要进一步检查。
- **特征重要性:** GBT 模型的结果显示, 近期历史用电量相关的滚动统计特征 (特别是过去 3 小时的最大值、不同时间窗口的标准差和最小值) 是最重要的预测因子。时间特征 (小时、年份) 也具有较高的重要性。土壤温度作为一个天气相关特征也进入了 Top 10。
- **下一步:** 需要深入分析 GBT 模型性能不佳的原因, 并考虑优化模型参数或尝试其他模型。线性回归的较好表现可能得益于其简单性或对当前特征集的适应性, 但也需警惕其可能无法捕捉复杂的非线性关系。

# 下一步：模型结果分析与改进

## ❶ 模型结果分析：

- 分析已训练模型的性能（RMSE, MAE）。
- 检查特征重要性，理解哪些特征对预测贡献最大。
- 进行误差分析，了解模型在哪些情况下表现不佳。

## ❷ 模型改进或探索其他模型：

- 尝试更复杂的模型架构或算法。
- 探索特征工程的其他方法，如交互特征或更高阶统计量。
- 考虑使用时间序列模型，如 ARIMA, Prophet, LSTM 等。

## ❸ 模型评估与选择：

- 在验证集上评估不同模型的性能。
- 选择最优模型并在测试集上进行最终评估。

## ❹ 预测与应用：使用最优模型对未来的电力需求进行预测。