电力需求数据集探索性数据分析与特征工程

SakuraPuare

github.com/SakuraPuare/ElectricityDemand

2025 年 4 月 22 日

目录

背景与目标

- 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
  - 特征工程构建预测特征集

数据集结构

数据集包含三个主要部分：

- **电力需求数据 (Demand Data)**: unique_id, timestamp, y (kWh)
- **元数据 (Metadata)**: unique_id, location_id, building_class, freq, etc.
- **天气数据 (Weather Data)**: location_id, timestamp, temperature_2m, humidity, etc.

核心信息速览

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: 无重复
   - Weather: 极少量重复 (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)

分布形态与异常值

- **原始尺度:** 高度右偏 (均值 >> 中位数)，标准差大，存在极端高值。存在少量非正值。
- **log1p 变换:** 改善对称性，更接近正态但仍有峰态。

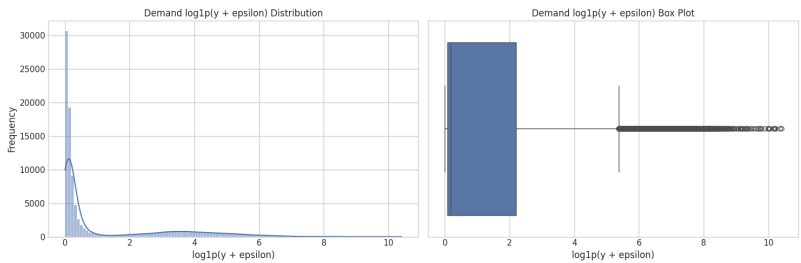


图: log1p 变换分布

典型模式与特性

- 多重周期性 (日内，周内，年度)
- 波动性，异常值，趋势性
- 不同用户模式多样性 (Residential vs Commercial)

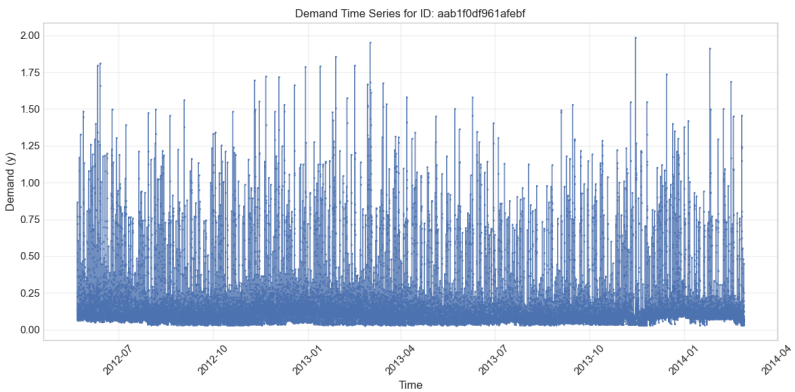


图: 样本 1 (日内/周内)

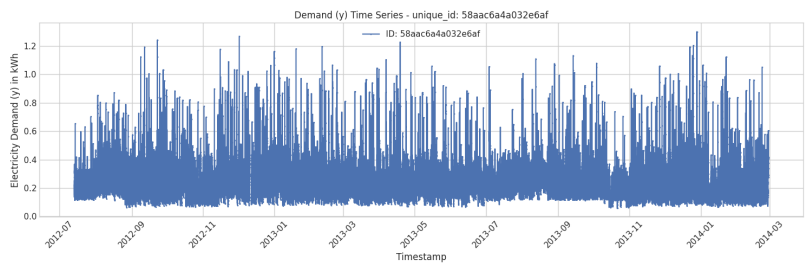
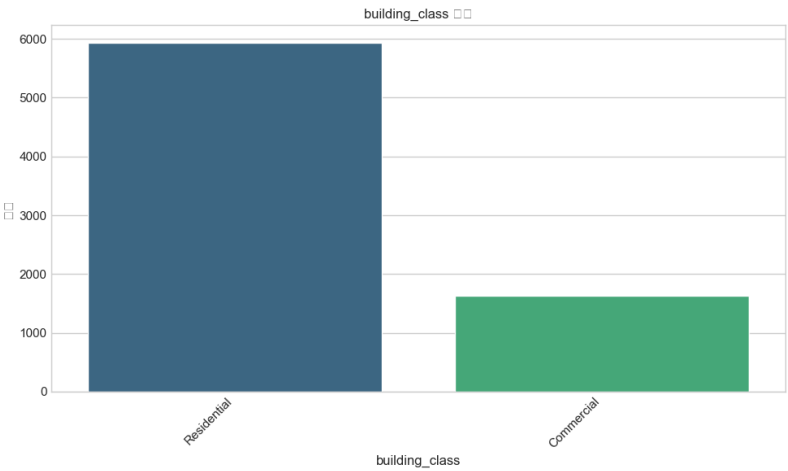


图: 样本 5 (工作日 vs 周末)

主要特征分布与地理位置

- **分类特征:** Building Class (住宅为主), Location (集中于伦敦), Freq (30T, 1H 为主), Timezone (Europe/London), Dataset Source。
- **数值/地理特征:** 经纬度集中分布，Cluster Size (多数为单个建筑)，地理位置有缺失记录。



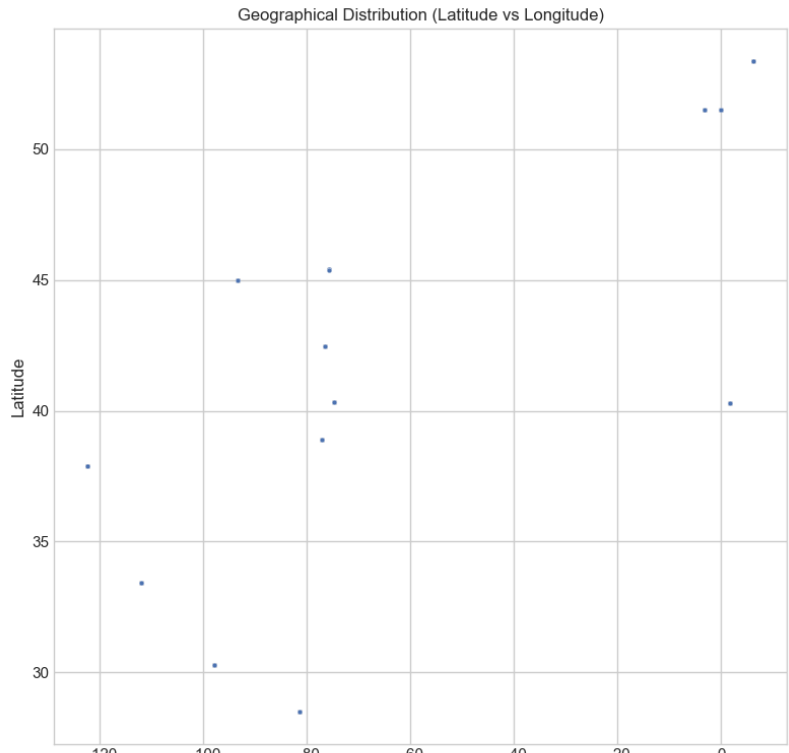


图: 建筑类型分布

主要天气特征分布

- **数值特征:** Temperature (近似正态), Relative Humidity (分布较广), Precipitation (零膨胀), Wind Speed (右偏) 等。
- **其他特征:** Cloud Cover (U 形), Sunshine Duration (两极), Weather Code (常见类型), Is Day (昼夜平衡)。

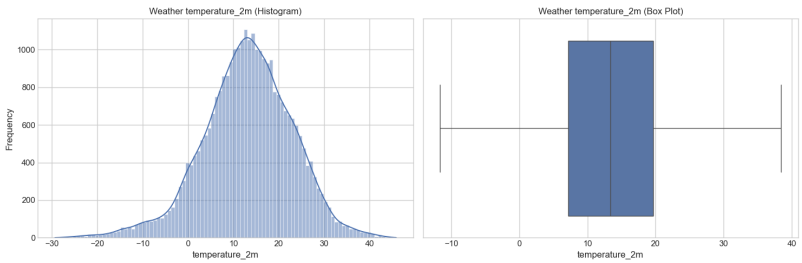


图: 温度分布

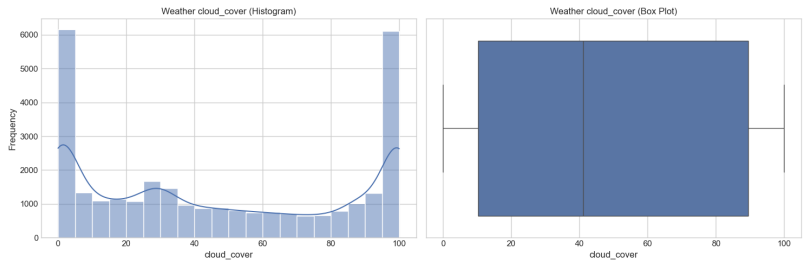


图: 总云量分布

建筑类型对需求的影响

- Commercial (商业) 建筑的电力需求通常显著高于 Residential (住宅)。
- 不同 Dataset/Location 的需求分布也存在差异。

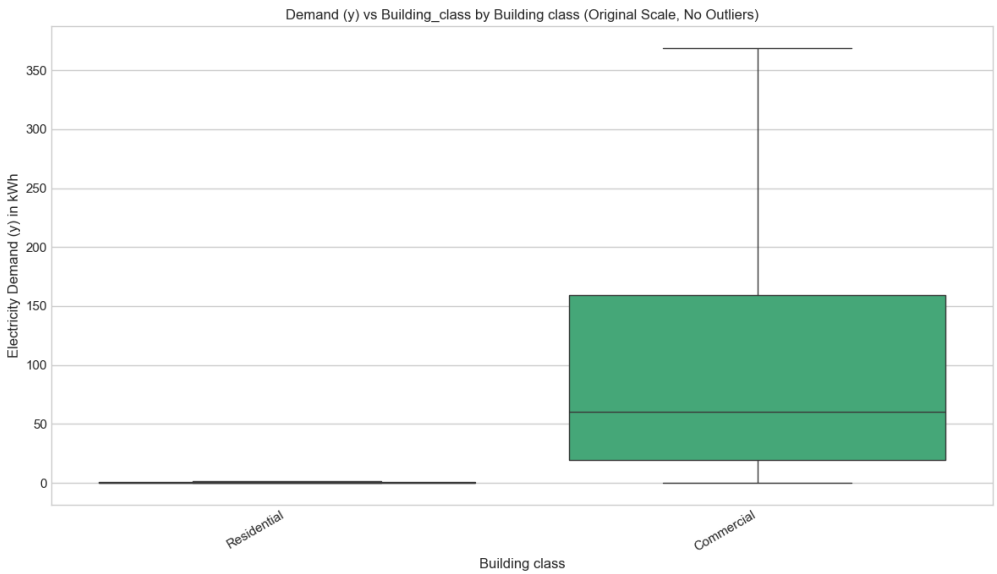


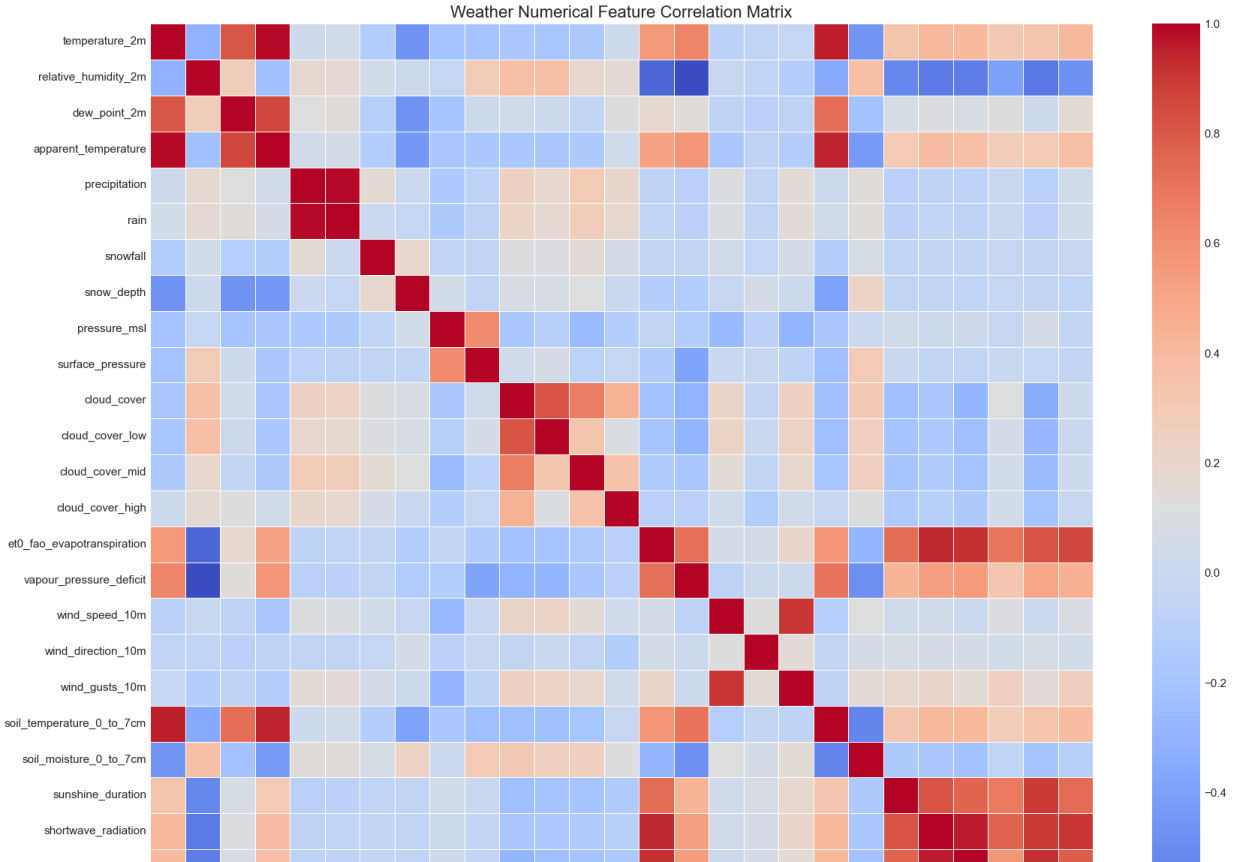
图: 需求 vs 建筑类型 (原始尺度箱线图)

相关性与合并流程

1. **初步相关性:**
   - Demand 与 Temperature 弱正相关 (0.028)
   - Demand 与 Apparent Temperature 弱正相关 (0.038)
   - Demand 与 Relative Humidity 中度负相关 (-0.202)
2. **数据合并流程:**
   1. 加载数据 & 需求重采样到小时。
   2. 需求 + 元数据合并 (基于 unique_id)。
   3. 天气数据去重 & 时间戳对齐 (到小时)。
   4. 需求/元数据数据 + 天气数据合并 (左连接 using location_id, hourly timestamp)。

合并诊断与天气特征关联

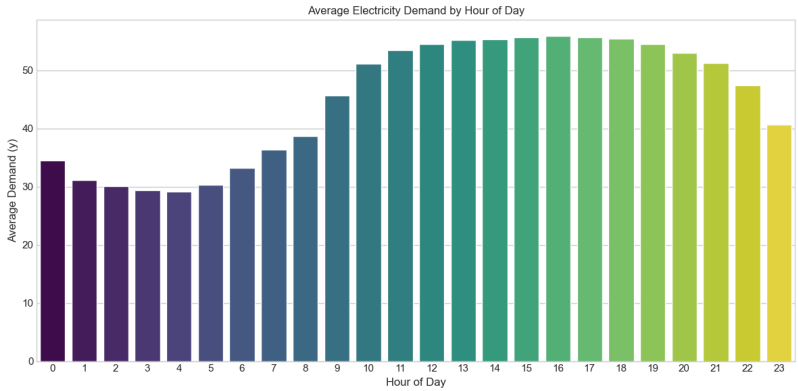
- **合并诊断:** 约 1.73% 记录因 Location ID 缺失/不匹配未能关联天气，其余成功关联。
- **天气特征相关性:** 如下图所示，天气特征内部存在相关性(例如温度与体感温度高度相关)。



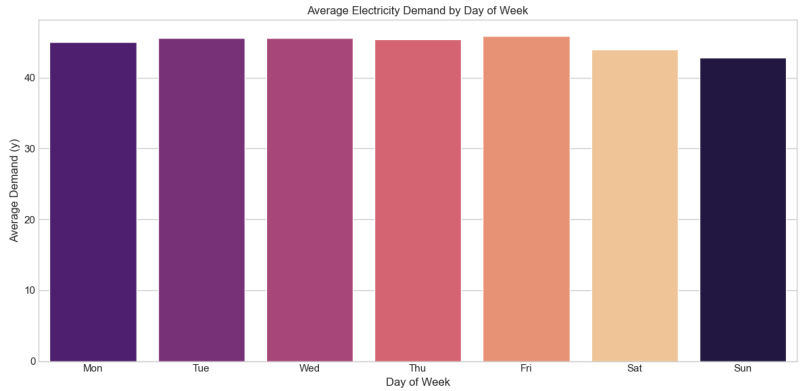
频率处理与时间模式

1. **时间频率匹配:**
   - Demand 数据频率多样 (15T, 30T, 1H 等)，Weather 主要为 1H。
   - **处理:** 将需求数据重采样并聚合到小时频率，以便与天气数据合并。
2. **周期性分析:** 需求数据表现出清晰的:
   - **日内周期:** 白天高峰，夜间低谷。
   - **周内周期:** 工作日与周末模式差异。
   - **年度周期:** 季节性波动 (通常冬夏高)。

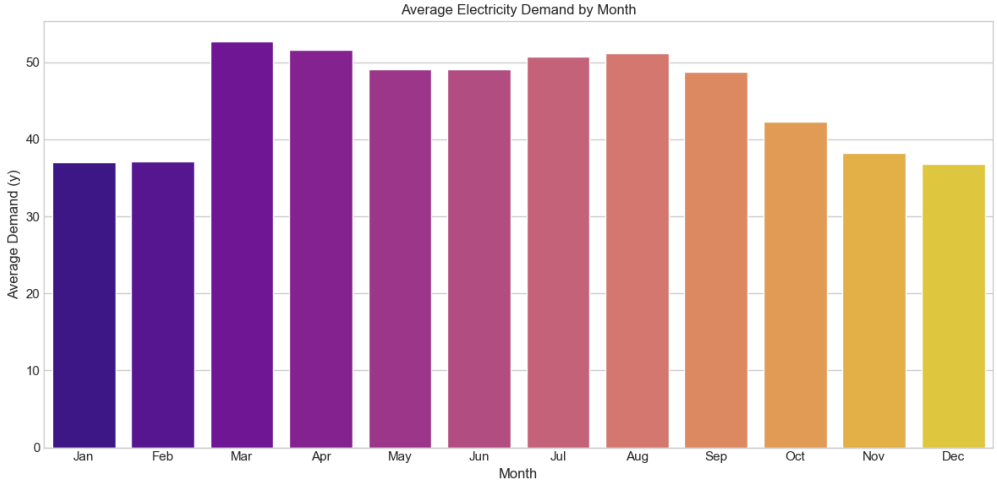
平均需求随时间变化



图：按小时平均



图：按星期平均



构建预测特征与处理

1. **特征类型:**
   - **时间特征:** 年，月，日，星期几，年内天，小时。
   - **滚动窗口统计特征:** 基于历史需求 (y)，窗口 (3H-168H)，统计量 (均值，标准差，最小值，最大值)。
   - **其他 (来自 Metadata/Weather):** Building Class, Location, Temperature, Humidity 等。
2. **缺失值处理:**
   - 移除目标 y 缺失的行。
   - 移除滚动特征计算初期的缺失值。
3. **特征集输出:** 按年/月分区存储包含原始数据、时间特征、滚动特征的数据集 (‘data/features.parquet‘)。

主要发现回顾

- **数据特性:** 规模大，异构多源，需求分布高右偏。
- **数据质量:** y 缺失，元数据位置缺失，天气重复 (已处理)。
- **关系:** 建筑类型、天气 (温湿度) 与需求相关。
- **时间模式:** 需求有清晰的日/周/年周期性，时间频率不匹配已通过重采样解决。

下一步：模型构建准备

1. **数据清洗/预处理:** 处理合并后天气缺失的约 1.73% 行 (插补或移除)。
2. **进一步特征工程 (可选):**
   - 滞后特征 (Lag Features)
   - 交互特征 (Interaction Features)
   - 分类特征编码 (独热，目标编码等)
   - 数值特征变换/归一化

下一步：模型选择与评估

3. **模型构建准备:** 合理划分训练/验证/测试集 (时序交叉验证)。
4. **模型选择与评估:**
   - 选择合适模型 (统计，ML, DL)。
   - 建立完整预测流程 (pipeline)。
   - 选择评估指标 (RMSE, MAE, MAPE)。
   - 在验证集上进行模型调优。
   - 在测试集上进行最终评估。