电力需求数据集探索性数据分析报告

SakuraPuare

April 22, 2025

Abstract

本文对 EDS-lab/electricity-demand 数据集进行了初步探索性分析。数据集包含电力需求、元数据和天气信息。分析涵盖了数据量、质量（缺失值、重复值）、关键变量（电力需求'y', building class, weather variables）的分布，以及变量间的初步关系。报告总结了数据集的主要特征和在进行电力需求预测建模前需要注意的关键问题，如数据分布的右偏、缺失值、天气与需求的时间频率不匹配等。

1 引言

电力需求预测是能源领域的重要任务。EDS-lab/electricity-demand 数据集提供了丰富的智能电表数据、元数据和天气数据，为电力需求分析和预测研究提供了基础。本报告旨在通过探索性数据分析 (EDA)，初步了解数据集的结构、内容、数据质量以及关键变量的特征和它们之间的关系，为后续的数据清洗、特征工程和模型构建提供指导。

2 数据集概览

该数据集主要包含三个部分：

- `demand.parquet`: 包含电力需求量 (`y`)、唯一 ID (`unique_id`) 和时间戳 (`timestamp`)。
- `metadata.parquet`: 包含每个唯一 ID 的元数据，如地理位置、建筑类型 (`building_class`)、数据频率 (`freq`) 等。
- `weather.parquet`: 包含各地理位置随时间变化的天气信息，如温度 (`temperature_2m`)、湿度 (`relative_humidity_2m`) 等。

3 数据量与数据质量分析

3.1 数据量

- Demand 数据：237,944,171 条记录。
- Metadata 数据：7,572 条记录 (对应 7,572 个 `unique_id`)。
- Weather 数据：604,848 条记录。

3.2 缺失值分析

- Demand 数据：`y` 列存在约 1.30% 的缺失值 (3,086,278 条)。其他列无缺失。
- Metadata 数据：与地理位置相关的列 (`location_id`, `latitude`, `longitude`, `location`) 存在约 3.13% 的缺失值 (237 行)。其他列无缺失。
- Weather 数据：无缺失值。

在后续处理中需要决定如何处理 Demand 和 Metadata 中的缺失值。

3.3 重复值分析

- Demand 数据：未发现基于 `unique_id` 和 `timestamp` 的重复行。
- Metadata 数据：未发现基于 `unique_id` 的重复行。
- Weather 数据：发现了 6 行基于 `location_id` 和 `timestamp` 的重复行 (涉及 12 条记录)。这些重复记录在后续处理中应予以移除。

3.4 时间范围分析

- Demand 数据：从 2011-01-01 00:30:00 到 2017-12-31 23:00:00。
- Weather 数据：从 2011-01-01 00:00:00 到 2019-01-01 06:00:00。Weather 数据覆盖了 Demand 数据的时间范围。

4 Demand (电力需求'y') 分析

通过对 Demand 数据进行 0.5% 的抽样分析，电力需求量 y 表现出高度的波动性和右偏分布特征。全量数据的描述性统计也证实了这一点 (均值：44.9, 标准差：394.3, 中位数：0.199)。抽样分析发现：

- 均值：约为 44.9 kWh。
- 标准差：很大，约 394.3 kWh。
- 中位数：仅为 0.199 kWh，远小于均值，表明存在大量小值。
- 极端高值：最大值达到 221,228 kWh。
- 非正值：全量数据中约 1.06% 的非缺失 y 值小于或等于 0。

图 1 和 2 展示了抽样数据的分布情况，包括原始尺度和对数变换 (log1p) 后的尺度。对数变换后，分布更接近正态分布，但仍存在一些峰值。这种高度右偏和存在极端值的分布特点需要注意，在建模时可能需要进行数据变换或采用对异常值鲁棒的模型。非正值可能表示零需求或数据采集问题，也需要进一步研究和处理。

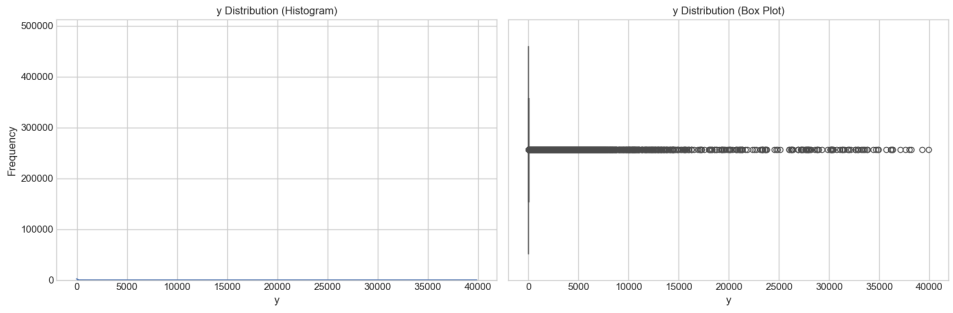

Figure 1: Demand (y) 值分布 (原始尺度，0.5% 抽样)

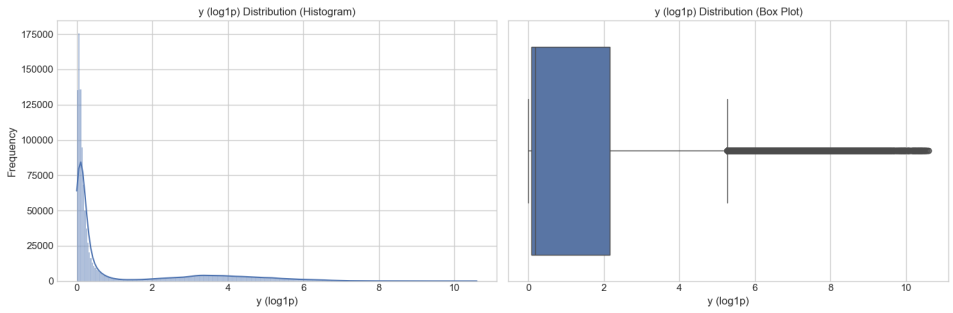

Figure 2: Demand (y) 值分布 (Log1p 尺度，0.5% 抽样)

5 Metadata (元数据) 分析

元数据提供了每个电表的重要属性信息（图 3 到 7）：

- Building Class: 主要为 `Residential` (5936 个)，`Commercial` (1636 个) 较少。
- Location: 主要集中在 `London, UK` (5634 个)，其他地点数量较少，存在 237 个缺失值。
- Frequency (`freq`): 最常见的采样频率是 `30T` (30 分钟，5566 个)，其次是 `1H` (1 小时，1636 个) 和 `15T` (15 分钟，370 个)。
- Timezone: 主要为 `Europe/London` (5781 个)。
- Dataset Source: 主要来源于 `London Smart Meter Data` (5566 个)。

建筑类型 (`building_class`) 是影响电力需求的重要因素。地理位置分布（图 8）显示数据主要集中在英国伦敦区域。

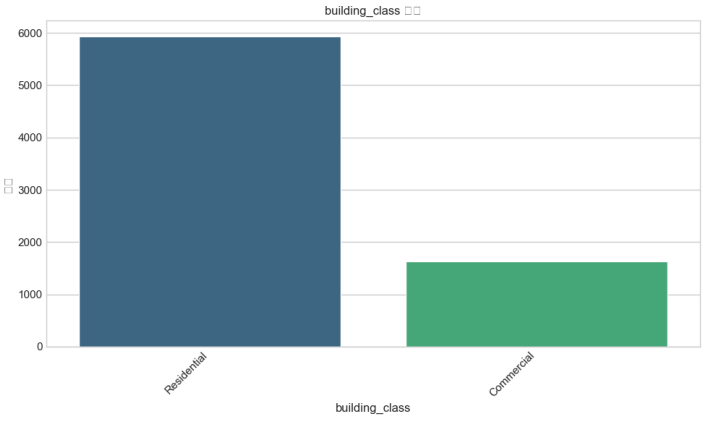

Figure 3: Metadata: Building Class 分布

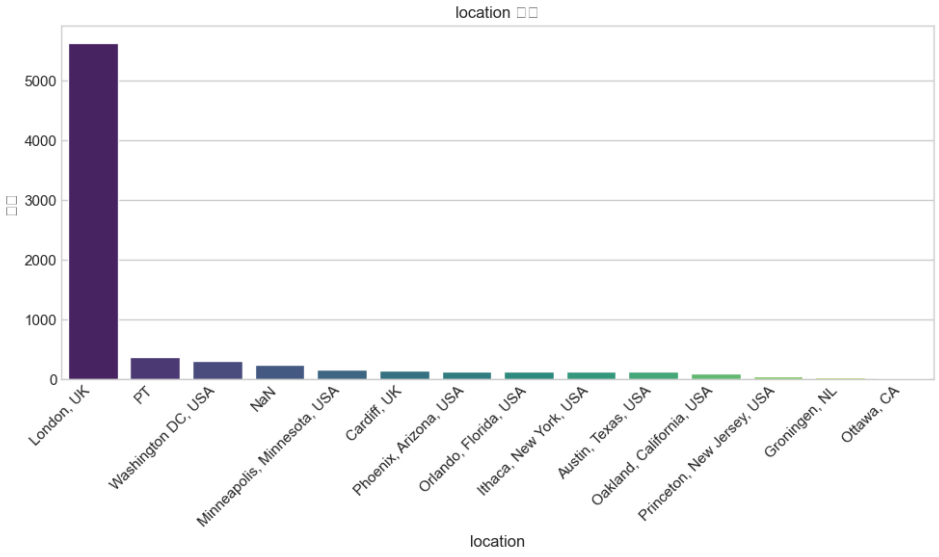

Figure 4: Metadata: Location 分布 (Top 10)

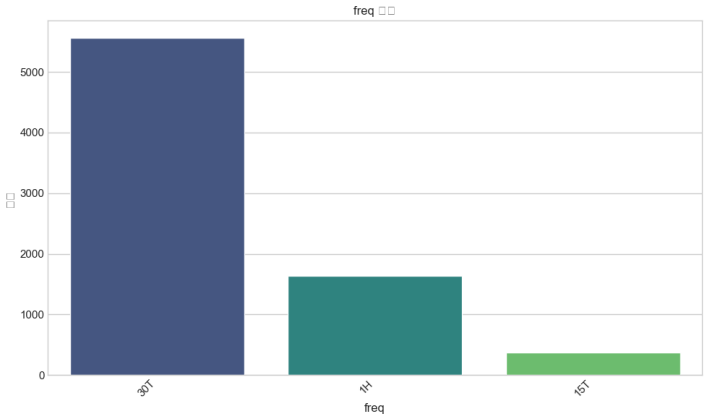

Figure 5: Metadata: Frequency (freq) 分布

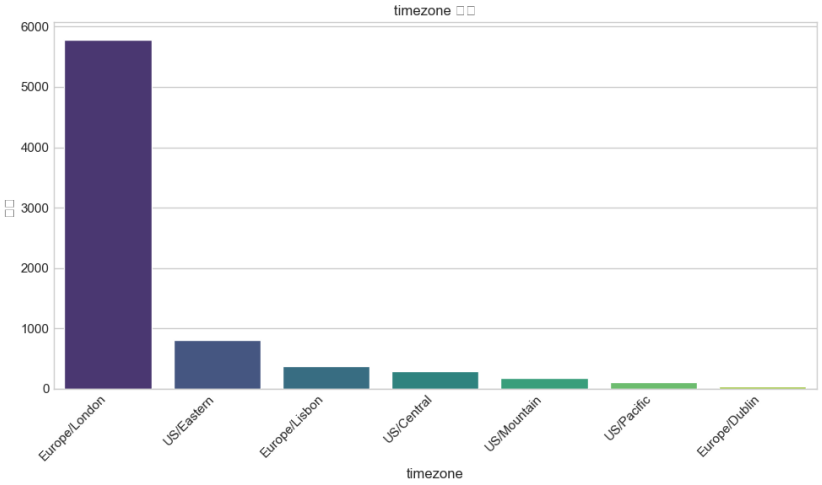

Figure 6: Metadata: Timezone 分布

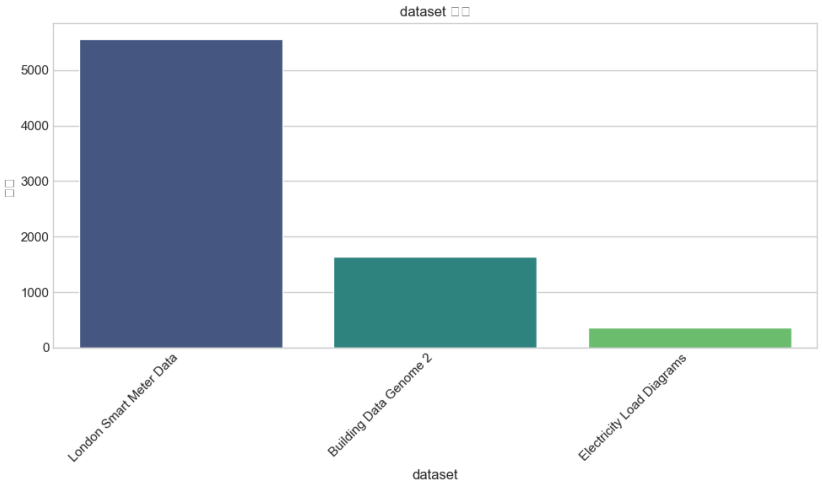

Figure 7: Metadata: Dataset Source 分布

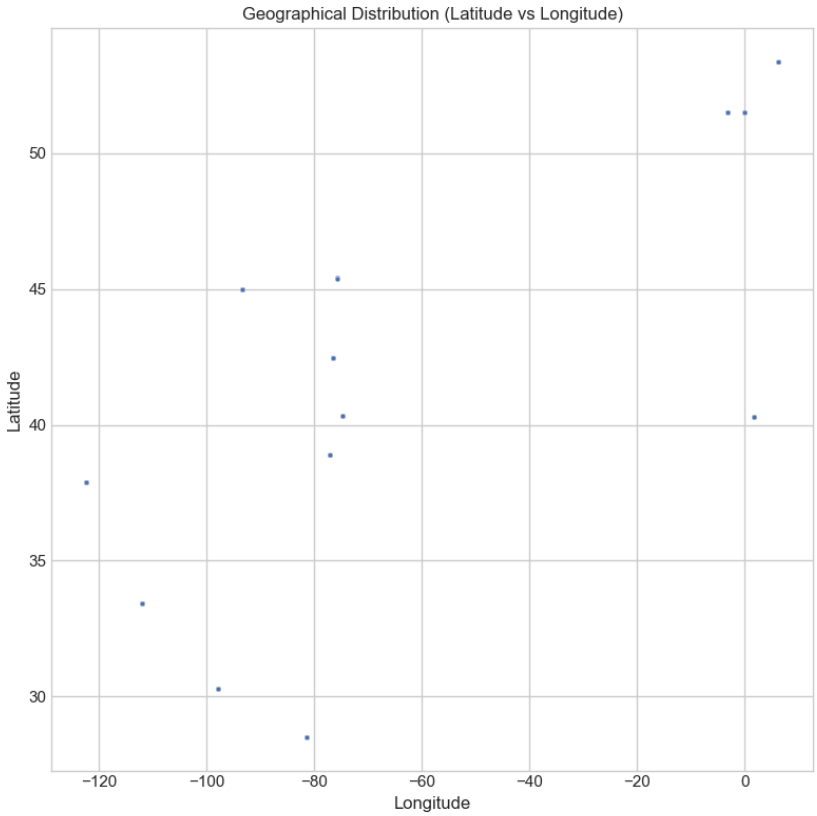

Figure 8: Metadata: 地理位置分布 (经纬度散点图)

6 Weather (天气) 分析

天气数据提供了可能影响电力需求的环境因素。关键数值特征（基于完整数据）的描述性统计如下：

- `temperature_2m` (温度): 均值约 13.0řC，标准差 9.9řC。
- `relative_humidity_2m` (相对湿度): 均值约 73.0%，标准差 19.8%。
- `precipitation` (降水): 均值为 0.10 mm/h，中位数为 0。
- `wind_speed_10m` (风速): 均值约 12.8 km/h，标准差 6.8 km/h。

图 9 和 10 展示了温度和相对湿度的分布（基于 5% 抽样）。未在 `precipitation`, `rain`, `snowfall` 列中发现负值，数据质量较好。天气代码 (`weather_code`) 和是否白天 (`is_day`) 的分布见图 11 和 12。

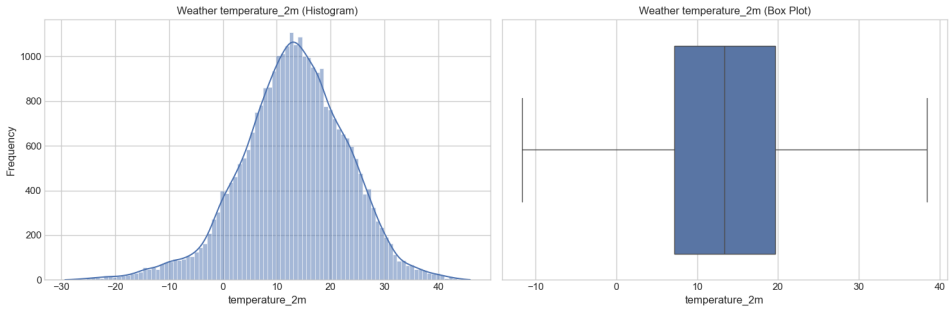

Figure 9: Weather: Temperature (2m) 分布 (5% 抽样)

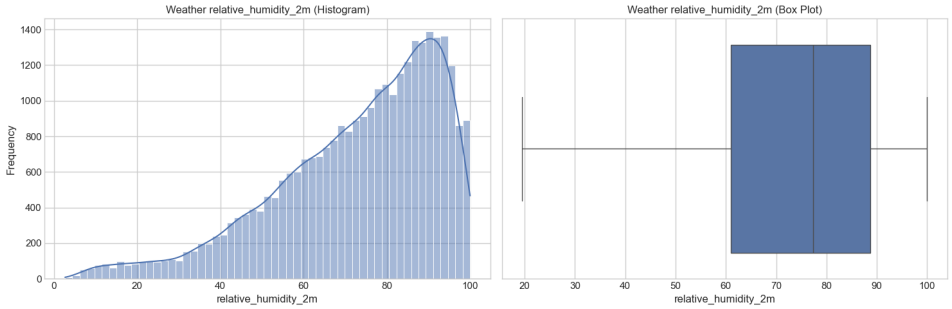

Figure 10: Weather: Relative Humidity (2m) 分布 (5% 抽样)

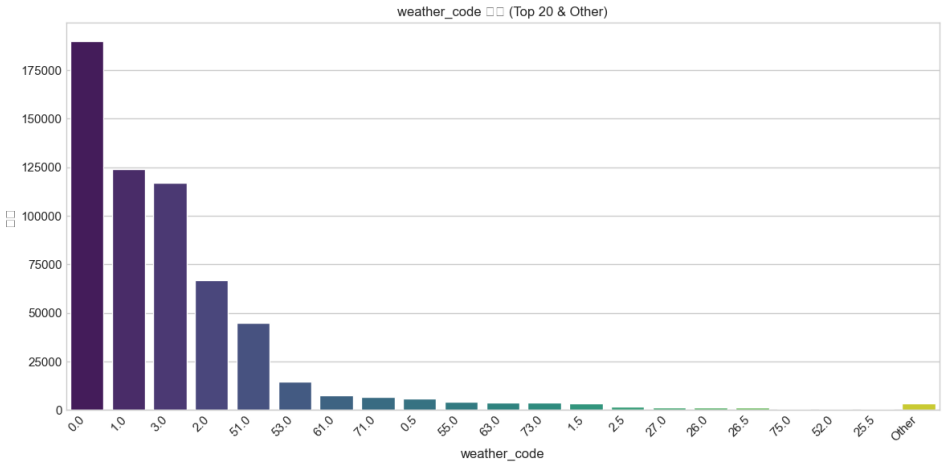

Figure 11: Weather: Weather Code 分布 (Top 20)

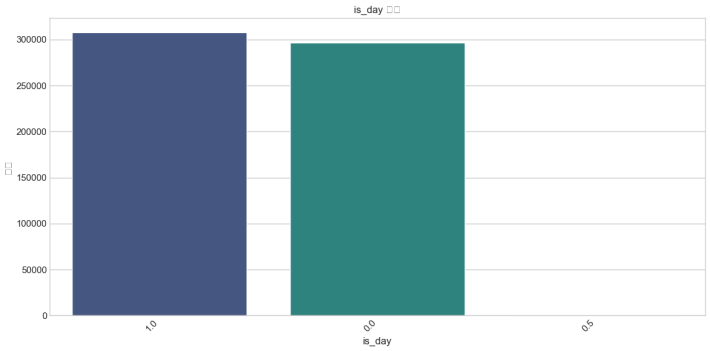

Figure 12: Weather: Is Day 分布

7 关系分析

7.1 Demand 与 Metadata

通过对 0.1% 抽样数据的箱线图分析发现（图 13），`Commercial` 建筑的电力需求 (`y` 值的中位数和分布范围) 通常显著高于 `Residential` 建筑。对数变换后的图也显示了类似的趋势，但更清晰地展示了低需求区域的分布。这表明建筑类型是预测电力需求的重要特征。类似地，不同数据集来源的电力需求分布也有显著差异（图 14），这可能与不同数据集采集的建筑类型或地区有关。

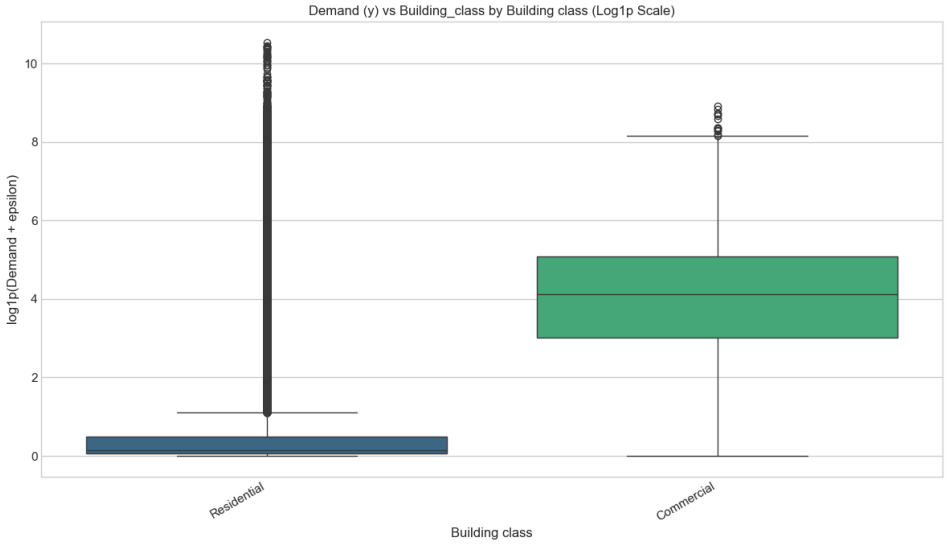

Figure 13: 电力需求 (`y`, Log1p 尺度) 与建筑类型 (`building_class`) 的关系 (0.1% 抽样)

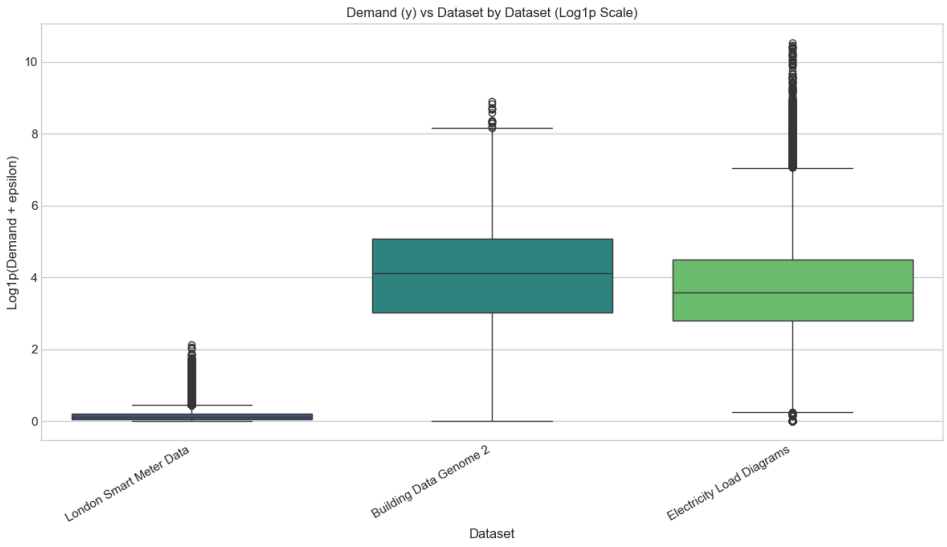

Figure 14: 电力需求 (y, Log1p 尺度) 与数据集来源 (dataset) 的关系 (0.1% 抽样)

7.2 Demand 与 Weather

基于对 50 个 unique_id 的抽样数据，并将其与对应的天气数据合并后，计算了电力需求 y 与部分关键天气特征的相关性（基于 Pandas 计算结果）：

- Demand 与 temperature_2m 呈正相关 (约 0.334)。
- Demand 与 relative_humidity_2m 呈负相关 (约 -0.165)。
- Demand 与 apparent_temperature 呈正相关 (约 0.328)。
- Demand 与 cloud_cover 呈负相关 (约 -0.225)。
- Demand 与 precipitation (约 0.013) 和 wind_speed_10m (约 0.028) 的相关性较弱。

这些相关性表明天气因素（尤其是温度、湿度和云量）对电力需求有一定影响。图 15 展示了天气特征之间的相关性，可见许多天气变量本身高度相关（如温度与体感温度、温度与露点温度等）。

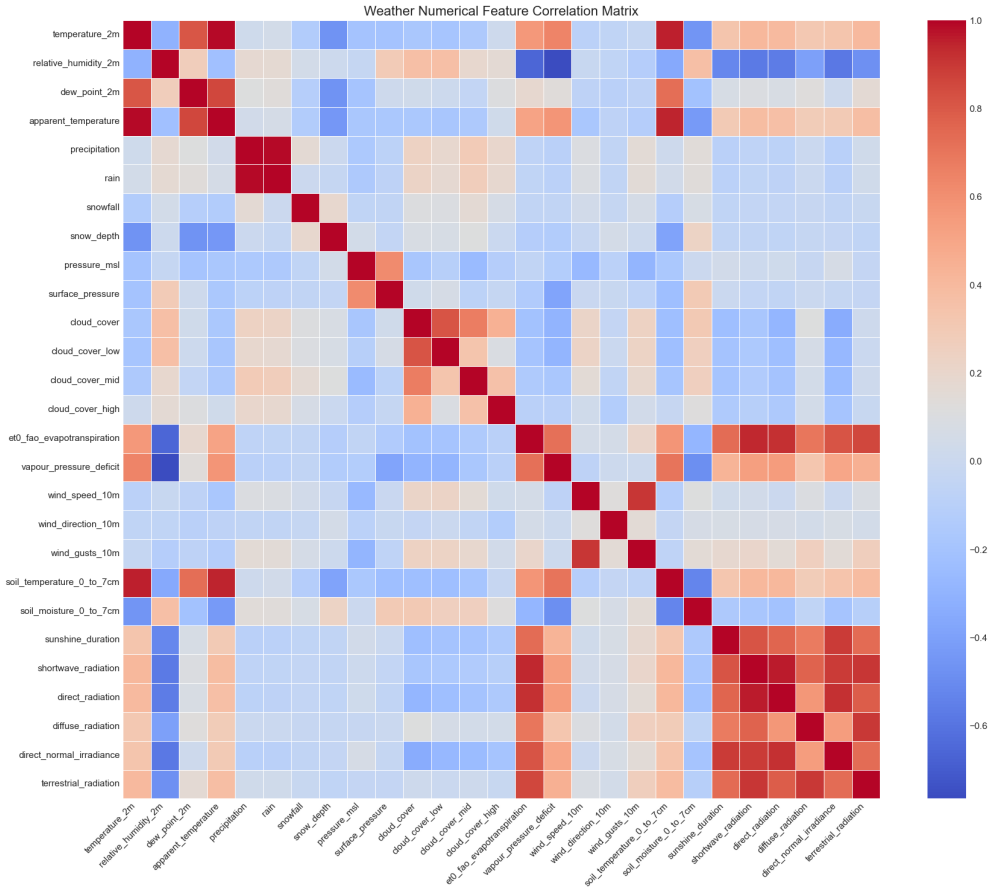

Figure 15: Weather 特征相关性热力图

8 时间特征分析

8.1 时间戳频率分析与匹配

抽样分析显示:

- Demand 数据存在多种采样频率，最常见的是 30 分钟 (约 16.8 万样本点)，其次是 15 分钟 (约 4.2 万样本点) 和 1 小时 (约 2.8 万样本点) (基于 0.1% 抽样)。
- Weather 数据主要为 1 小时频率 (抽样显示全部为 1 小时)。

在合并 Demand 和 Weather 数据进行建模时，需要处理这种时间频率不匹配的问题。常见的方法包括对数据进行重采样（例如，将 15 分钟和 30 分钟 Demand 数据聚合到 1 小时）或采用适当的插值方法。

8.2 周期性分析

通过聚合分析（图 16 至 18），可以观察到电力需求的周期性模式:

- ** 小时周期 **: 平均需求在白天较高，夜间较低，存在早晚高峰。
- ** 星期周期 **: 工作日 (周一至周五) 的平均需求通常高于周末。

- ** 月份周期 **: 平均需求在冬季和夏季较高，春秋季较低，呈现典型的季节性模式。

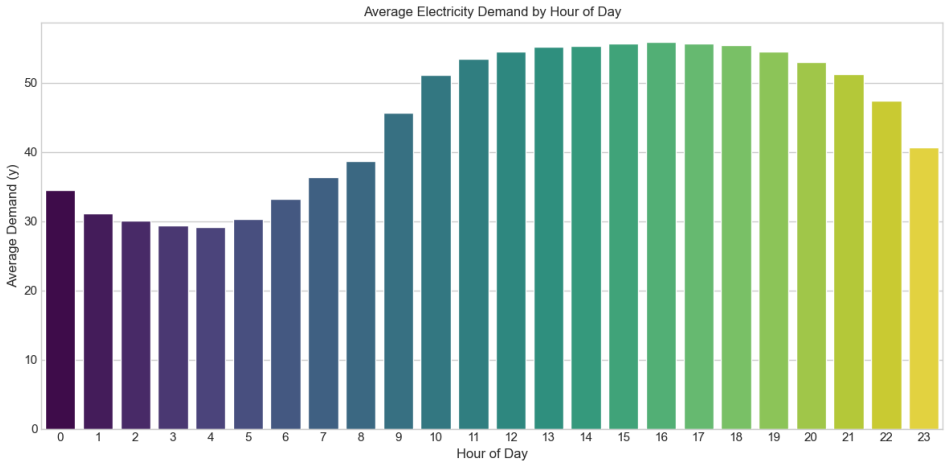

Figure 16: 按小时聚合的平均电力需求

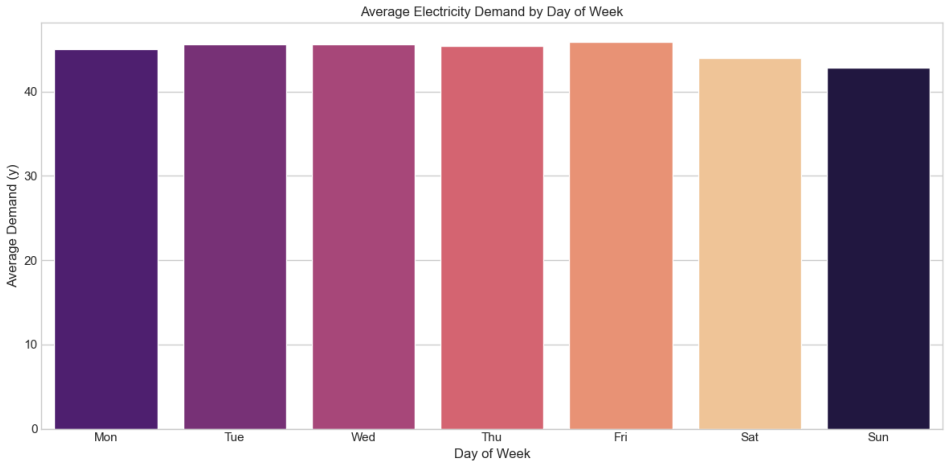

Figure 17: 按星期几聚合的平均电力需求

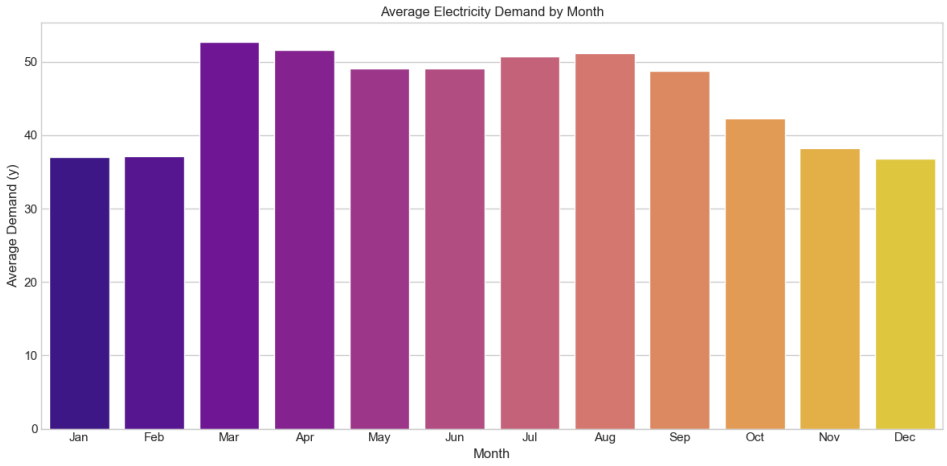

Figure 18: 按月份聚合的平均电力需求

9 总结与后续步骤

本次探索性数据分析揭示了电力需求数据集的关键特征和挑战：

- 数据量大，包含多源异构数据。
- 存在少量缺失值 (Demand: y, Metadata: location) 和重复值 (Weather)，需要进行清洗。
- 电力需求 (y) 具有高度右偏和极端值，且存在非正值。
- 建筑类型、数据集来源、地理位置是影响电力需求的重要类别特征。
- 天气因素，特别是温度、湿度和云量，与电力需求存在一定相关性。
- 电力需求存在明显的小时、星期、月份周期性。
- Demand 和 Weather 数据的时间频率不一致，需要进行对齐。

基于以上分析，后续的步骤将包括：

1. ** 数据清洗 **: 处理 Demand 和 Metadata 的缺失值, 移除 Weather 的重复行, 处理 Demand 的非正值。
2. ** 时间序列处理 **: 将不同频率的 Demand 数据统一（例如重采样到小时），并将 Demand 和 Weather 数据按时间戳和位置进行对齐。
3. ** 特征工程 **: 基于时间戳提取更详细的时间相关特征 (如小时、星期几、月份、年份、是否节假日、滞后需求等)，并合并处理后的天气数据，考虑对高度相关的天气特征进行筛选或组合。
4. ** 数据分割 **: 根据时间和 `unique_id` 构建合适的训练集、验证集和测试集，注意避免数据泄露。
5. ** 模型构建与评估 **: 选择合适的预测模型 (如 LightGBM, 时间序列模型如 ARIMA/Prophet, 或深度学习模型如 LSTM/Transformer) 进行训练、调优和评估，可能需要对 `y` 进行变换。

本报告为初步分析结果，详细的 EDA 过程和精确数值需要根据完整的分析脚本输出来填充。