ACENET
Microcredential in Advanced Computing
ISP Report

Project title: Harnessing Big Data and High Performance Compute to Understand Commercial Energy Usage and Utility Generation Dynamics

Participant name: Neil Nordeje

Date: 2024-07-11

Abstract:
This project analyzes electricity consumption to establish benchmark power usage for various industry sectors, uncovering trends to help consumers make informed energy decisions and contribute to climate change efforts.

1. Introduction

The focus of my Independent Study Project (ISP) is on analyzing electricity power consumption, primarily among commercial and retail users. The initial research question driving this project is: "How can analyzing electricity consumption data across different sectors reveal trends and insights that can lead to more informed energy usage decisions?" By exploring this question, the project seeks to uncover consumption patterns and benchmarks across various sectors, such as department stores, supermarkets, banks, and fast food chains, and analyze deviations from these benchmarks. The goal is to provide actionable insights to help consumers optimize their energy usage and contribute to a more sustainable future.

2. Background

The motivation for this project arises from the observation that many consumers simply pay their electricity bills without considering the trends and insights embedded in the data. Recognizing the wealth of information within electricity usage data, this project aims to empower consumers to make more informed energy decisions, ultimately contributing to climate change efforts.

3. Analysis

The approach for this project is detailed in the project outline, comprising the following steps:
1. Defining project objectives, formulating research questions, and acquiring the dataset.
2. Preprocessing and exploring the data, including initial data cleaning.
3. Conducting exploratory data analysis (EDA).
4. Performing in-depth analysis.
5. Interpreting results and drawing conclusions.
6. Preparing the final report.

Dataset
The dataset for this project was obtained from a previous employer who had access to extensive power consumption data. The dataset was open-sourced specifically for this purpose.

Dataset Description
The data was originally provided in a SQL database format (Figure 1) with three separate tables:

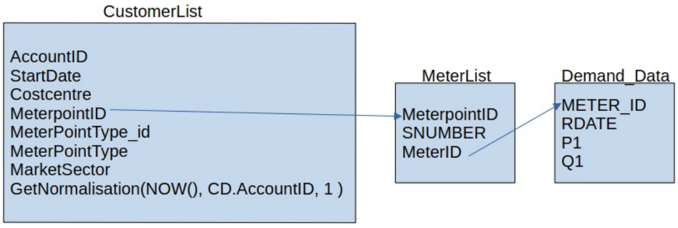

Figure 1: The tables from the relational database system (RDS), each arrow illustrates a foreign key.

CustomerList Table:
AccountID: 4301 unique customer accounts.
Costcentre: 15 unique categories for billing allocation according to power source or load.
MeterpointID: 10,848 unique values; represents virtual power meters, linked to physical meters.

MeterpointType: 25 unique values indicating energy consumption categories (e.g., Grid, Generator, Solar).
MarketSector: 79 unique market sectors in which the consumers operate.
GetNormalisation: Gross letting area, used to normalise the power consumption data.

MeterList Table:
Meterpoint_ID: 10,281 unique values, linking to multiple METER_IDs in the Demand_data table.
Meter_ID: 10,627 unique values, directly linked to the SNUMBER column.
SNUMBER: 10,627 unique values, linking to the METER_ID column.

Demand_data Table:
METER_ID: Unique power point identifier, linked to Meter_ID in Meterlist table
RDATE: Date and time of reading
P1: Real power consumption data in kWh
Q1: Reactive power consumption data in kVAR

Data Preparation

To prepare the data for analysis, two primary options were considered: establishing a direct SQL connection or exporting the data to a file format like CSV. The latter was chosen to build knowledge on handling large datasets with Dask Dataframes rather than pulling smaller, easily processable data from the SQL database. This method allowed for working with the complete dataset, despite the longer processing times compared to SQL queries.

The data was exported using an inner join SQL statement to combine all three tables, with "NaN" values (less than 0.01% of total data) removed. Only necessary columns were exported to minimize the file size, ensuring no missing values or incompatibilities. Knowing the SQL data types allowed for specifying the exact data type for each column when importing into the Dask Dataframe.

Analysis Method

The project aimed to establish industry benchmarks, relying heavily on statistical analysis. The scipy stats module with skewnorm was used to generate skewed distribution curves due to the non-normal distribution of the data. For plotting, matplotlib was used to generate a 1-week statistical benchmark plot as well as histograms with skewed distributions, including the mean, median, and standard deviation of power consumption trends per industry sector. The power consumption data was normalized to Wh/m^2.

Statistical methods were chosen to generate a clear picture of power consumption per sector, providing benchmarks for comparison. A comparison meter was randomly selected to plot against the benchmark, offering practical insights for retail or industrial power users.

HPC usage

The dataset, comprising almost 500 million rows with multiple columns, was too large to fit into the RAM (32GB) of a typical desktop PC. Therefore, the Dask Dataframe module was used for distributed computing, reducing the load on a single node. The code was designed to process data in batches by sector (79 sectors total), theoretically, 8 batches of 10 sectors each could've have been processed seperately. Dask partitions were sized at 0.1GB, allowing for efficient memory management with a total of 1886 partitions for the full dataset. Increasing the size of Dask partitions showed significantly slower performance.

Numerous issues were encountered while running the code on HPC, primarily due to discrepancies between the local and HPC environments. Trying to process the entire dataset on HPC with the modified code was not possible due to a 50GB disk quota limit. Initailly, I successfully transferred a compressed version of the full dataset using the xz utility to my HPC account. However, when attempting to extract it, the process would fail once the 50GB limit was reached. Attempting extraction in different locations, such as home (~) and scratch (/scratch/), did not solve the problem.

Table 1: Attempted and proposed configuration with sample and full dataset

	Dask Distribution			Siku Attempt with sample dataset				
Attempt #	Partitions size (GB)	Partitions	Total partition size (GB)	# CPU	Efficiency	#GB RAM	Efficiency	Total #GB RAM
1	0.1	69	6.9	8	24.89%	4	42.07%	32
2	0.1	69	6.9	4	44.97%	8	46.38%	32
3	0.1	69	6.9	2	74.78%	8	89.35%	16
Proposed #				Proposed Siku config with full dataset				
1	0.1	1886	188.6	8		48		384
2	0.1	1886	188.6	16		24		384

After several adjustments and using a smaller sample dataset (69 partitions of 0.1GB each), initial runs revealed low CPU efficiency (8 CPUs at 24.89%) but better memory utilization (8 CPUs with 4GB each, totaling 32GB, at 42.07% ≈ 16GB). Consequently, I reduced the number of CPUs allocated by half and doubled the RAM per CPU. As shown in Table 1, attempts 2 and 3 demonstrated improved efficiency. Note that the small sample dataset of 6.9GB required approximately more than double the RAM (16GB).

If I had processed the entire dataset, I would have scaled from the 69 Dask partitions of 0.1GB each (totaling 6.9GB and needing 16GB RAM) to 1886 partitions (totaling 188.6GB and requiring about 377GB RAM). Given that the small sample required 2 sets of 1 CPU with 8GB RAM each, I would have increased the allocation to either 8 CPUs with 48GB RAM each (totaling 384GB) or 16 CPUs with 24GB RAM each (totaling 384GB), assuming hardware limitations were not exceeded. Dividing the job into smaller batches would have reduced resource demands, as previously suggested.

Future HPC projects would benefit from creating an environment on Siku first and exporting it to the local machine to avoid compatibility issues.

4. Results

The primary objective was to develop a benchmark for specific sectors that businesses can compare their own data against. This objective has been largely achieved, as demonstrated in Figure 2 through Figure 10.

4.1 Bechmark data

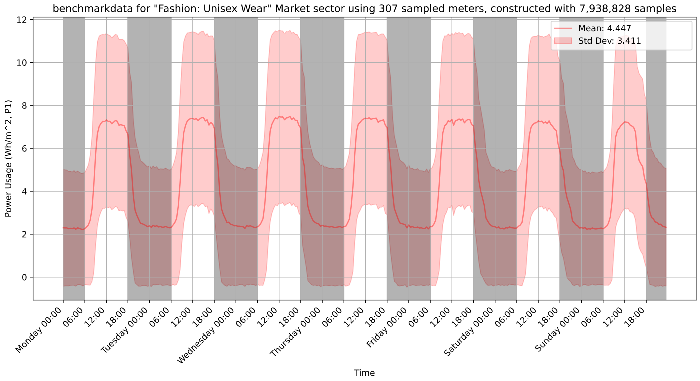

Figure 2: Statistical mean and standard deviation applied to timeseries power usage data. (Dark shaded areas indicate night time from 18:00 to 06:00)

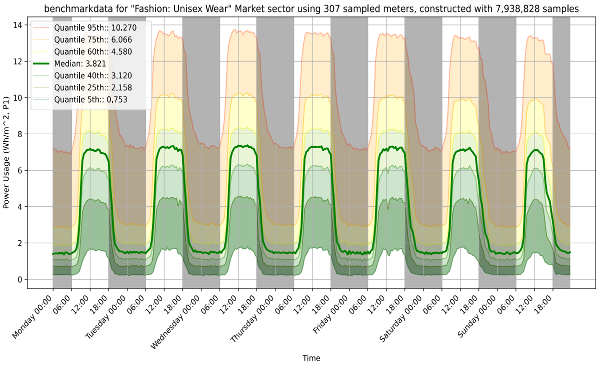

Figure 3: Statistical median and noteable percentile ranges applied to timeseries power usage data.

Figure 2 and Figure 3 display the statistical time series data for the benchmark of a specific sector, illustrating typical power usage over one week for that sector.

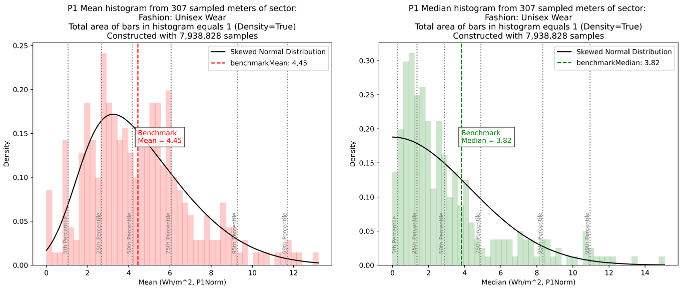

Figure 4: Skewed distribution fitted to the histogram of median and mean power usage data.

Figure 4 illustrates the skewed distribution fitted to a histogram. Each bin in the histogram represents the aggregated data from multiple years for a single sampled meter.

4.2 Bechmark comparison

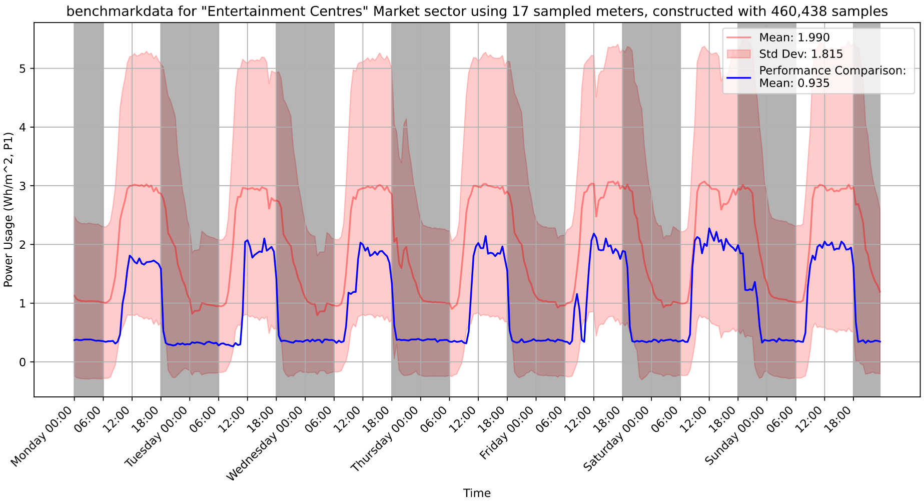

Figure 5: Mean and standard deviation time-series plot for a given sector, including performance comparison.

Figure 5 and Figure 6 overlays comparison data on the benchmark plot, depicting an almost ideal scenario where the comparison meter (blue line) exhibits consistent and predictable low consumption during nighttime and high consumption during daytime.

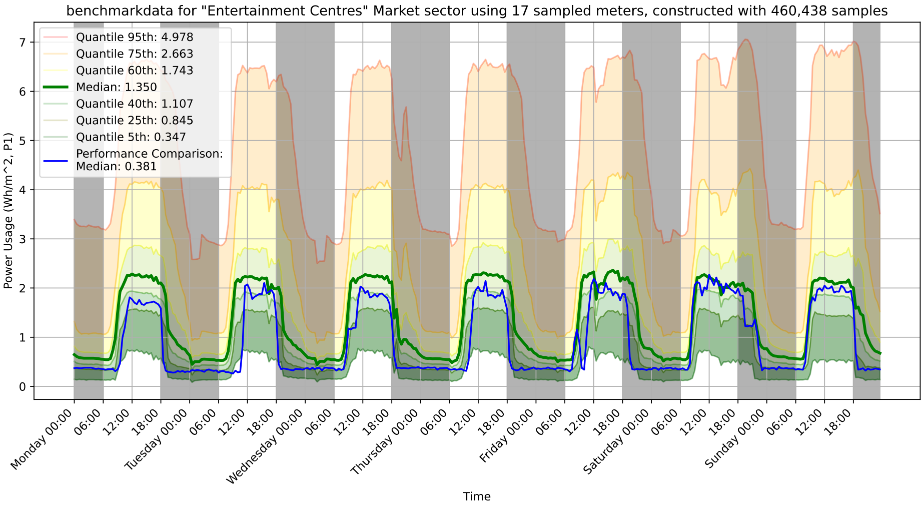

Figure 6: Median and percentile statistical time-series plot for a given sector with performance comparison (blue line)

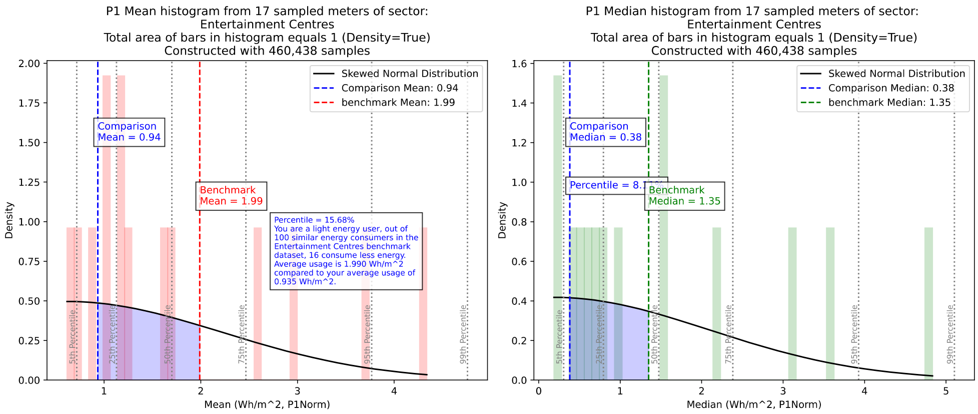

Figure 7: Histogram with performance comparison applied

Figure 7 indicates where the performance comparison falls within the skew distribution of mean and median values, highlighting a low energy density (Wh/m^2) footprint, an ideal situation.

4.3 Test Case #1

Figure 8 suggests a slightly above average energy density (Wh/m^2) footprint for the performance comparison. However, Figure 9 and Figure 10 reveals unfavorable high-highs during daytime and high-lows at night with no activity on Sundays, indicating a potential area for improving based no business operations on Sunday and higher than average energy density footprint.

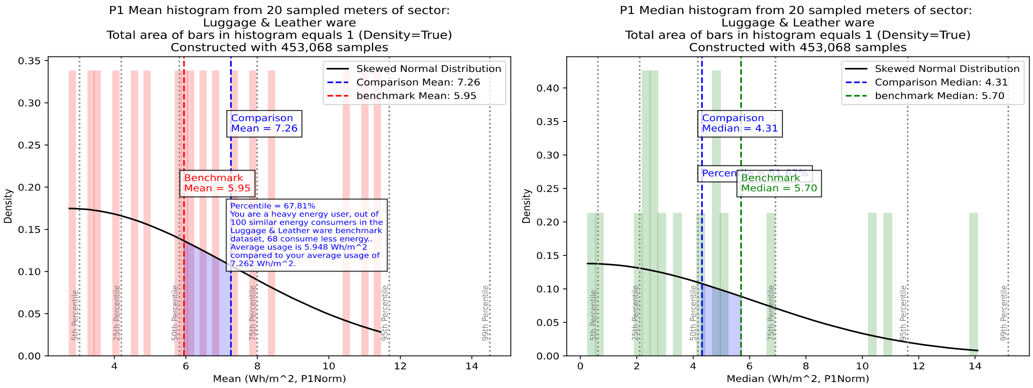

Figure 8: Test case #1, Midpoint mean and median on performance comparison

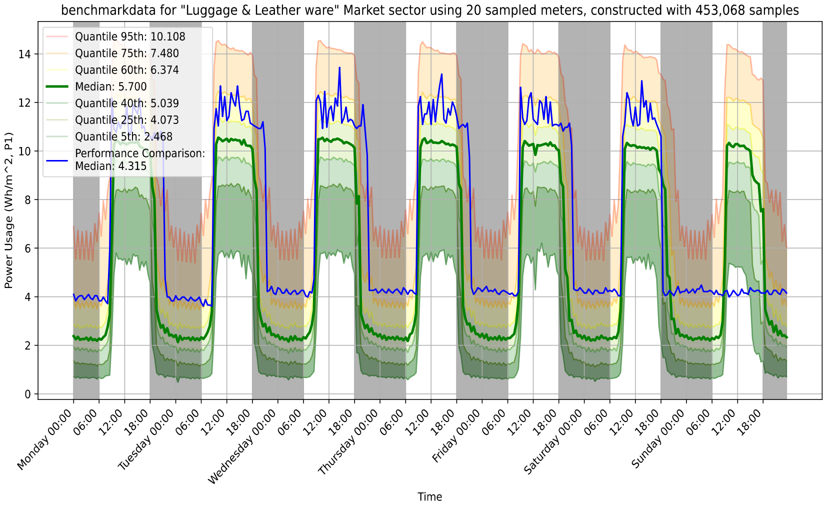

Figure 9: Test case #1, Median performance comparison with high-lows and high-highs

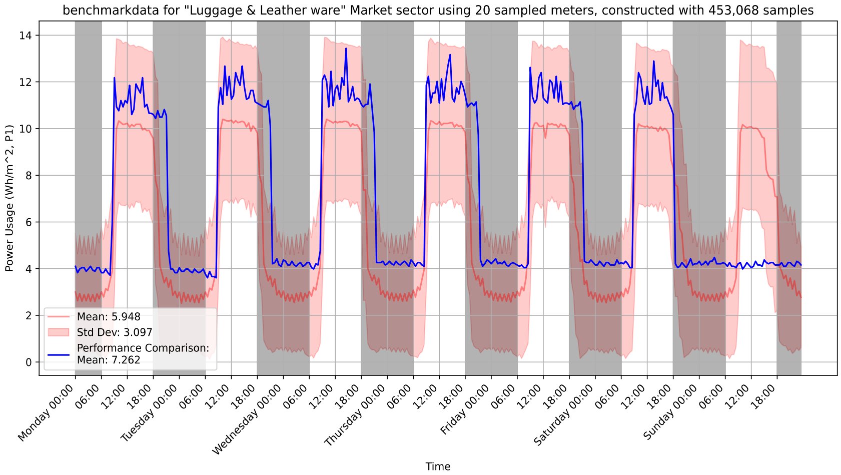

Figure 10: Test case #1, Mean performance comparison with high-lows and high-highs

Interpretation of Results

The initial research question was: *"How can analyzing electricity consumption data across different sectors uncover trends and insights that lead to more informed energy usage decisions?"*

The final product allows users within a specific industry to quickly compare their power usage against the sector average. If their consumption is above average, they can explore potential reasons. Additionally, users can analyze day/night patterns in their time series data to identify and investigate discrepancies. This tool supports making informed decisions to optimize energy usage and improve efficiency.

5. Discussion

Major adoption and shifting energy trends

If this model gains widespread adoption and leads to significant changes in behavior, such as reduced overall energy consumption within the industry, the current static model may become outdated. It may no longer accurately reflect true mean and median consumption within a given sector. Therefore, the model must be regularly updated to remain relevant.

Limited Range of Time Plot (1 Week vs. 1 Year)
One limitation of the benchmark analysis is that Figure 2 and Figure 3 only shows one week's worth of time series data. This limited time window does not account for seasonal variations in energy usage. Ideally, a yearly time series benchmark should be used to account for seasonality and holidays.

Normalization
Currently, normalization is applied using only the gross letting area (m^2). Ideally, power consumption should be normalized over one or more of the following variables: revenue generated, foot traffic, geographic location, and weather conditions.

Sample Sizes
Sample sizes varied considerably, ranging from 1,617 individual meter data points per sector to just one meter per sector. Sectors with a low meter count lacked sufficient data to generate statistically relevant results. Consequently, only the top 40 sectors with the most data out of 79 were processed

Siku HPC
As previously mentioned, there were difficulties working with the Siku HPC. These issues included unavailable or incompatible modules and a 50GB disk quota, leading to many hours of frustration.

6. Conclusion

This project aimed to develop a benchmark for specific sectors to enable users to compare their energy consumption against industry standards. The findings demonstrate that this goal has largely been achieved, as evidenced by the results presented in Figure 2 through Figure 10. These figures illustrate statistical time series data, histograms with skewed distributions, and comparison data over benchmark plots, respectively. Figure 5, Figure 6 and Figure 7 further highlights the comparison performance within the skew distribution of mean and median energy consumption, indicating an ideal scenario of low energy density.

Test Case #1 indicates that although the performance comparison reflects a slightly above average energy density footprint, there is room for improvement. Specifically, the data shows higher than usual energy consumption at night and no business operations on Sundays, as illustrated in Figure 10. This highlights the need for continuous analysis and investigation to optimize energy usage.

The significance of this project lies in its ability to provide a clear and actionable comparison for business owners within the industry. By enabling users to quickly see whether their energy consumption is above or below the sector average, they can identify areas for potential investigation and improvement. This tool empowers users to make informed decisions about their energy usage, contributing to overall energy efficiency.

However, there are limitations to this analysis. The static nature of the model may become outdated with significant changes in industry behavior, and the limited one-week time series data does not account for seasonal variations. Future research should focus on developing dynamic models that can be regularly updated and expanding the time series benchmark to a yearly scope. Additionally, incorporating multiple normalization factors, such as revenue generated, foot traffic, geographic location, and weather conditions, will provide a more comprehensive analysis.

Addressing the challenges faced with the Siku HPC, including module compatibility and data quota limitations, will also be crucial for future cohorts.

In conclusion, this project has provided a valuable tool for benchmarking energy consumption within specific sectors. With further enhancements and expanded research, this model can continue to aid in the drive towards more efficient and sustainable energy usage in the industry.

References
Sample dataset note that the complete dataset of 70GB was unable to fit onto github or other online hosting platforms

Supplementary Materials
Project link
Exploratory data analysis
Data processing and feature engineering