PART 2: SUMMARY REPORT

Xingchen (Estella) Ye
Department of Computer Science
Columbia University
New York, NY 10027, USA
xy2527@columbia.edu

1 PROBLEM FORMULATION

The objective of this task is to merge two data sources into a single dataset and evaluate the dataset for any anomalies. Specifically, we have two data files: one containing hourly electricity consumption data and another containing minute-by-minute appliance electricity consumption data. The goal is to merge these data sources, aggregate the new appliance data to an hourly level, and evaluate the merged dataset for anomalies.

- **USA_AL_Auburn-Opelika.AP.722284_TMY3_BASE.csv:** This file provides hourly electricity consumption data for a resident, with consumption values given in kilowatts (kW). Note that the hourly timestamps are in Hour Ending (HE) format.
- **new.app4.csv:** This file contains minute-by-minute electricity consumption data for a new appliance, with consumption values given in watts (W).

2 METHODOLOGY

1. **Standardize Column Name for Timestamp:**
   - Identify the column containing the time information in each dataset.
   - Rename the time columns to a consistent name to make the code reusable.
2. **Load and Process Data:**
   - Load the datasets and process them to prepare for merging.
   - For the appliance data:
     - Convert the timestamp to a uniform format.
     - Convert power units from W to kW.
   - For the hourly data:
     - Adjust the timestamp to ensure consistency across both datasets.
3. **Aggregate Appliance Data:**
   - Aggregate the minute-by-minute appliance data to an hourly level in HE formatting.
4. **Merge Data:**
   - Merge the two datasets based on their time columns.
   - Combine the datasets and calculate the total hourly electricity consumption by summing the consumption column for individual categories.

3 CODE IMPLEMENTATION

1. **Auxiliary Functions:**
   - `standarize_time_col(data)`
     - Identify the column which its name contains the word ”time”.
   - `load_and_process_new_app(file_path)`
     - Load the appliance data from the file path.
     - Standardize the time format to a uniform format.

  - Convert the power units.
  - Return the processed data.
- `load_and_process_all(file_path, base_year)`
  - Load the electricity consumption data from the file path.
  - Standardize the time format to a uniform format.
  - Adjust the timestamps to correct anomalies, such as replacing '24:00:00' with '00:00:00' and shifting the date.
  - Add the base year to the timestamps.
- `aggregate_app_data(data)`
  - Aggregate the data by summing the values within each hour.
  - Adjust the timestamps to be consistent with HE (Hour Ending) formatting. For example, the aggregation of all data entries in the time interval 11:00-11:59 of a day yield 11:00 in the timestamp column, which is supposed to be HE12/12:00 in HE formatting. So, instead of 11:00, the timestamp should be set to one hour later.
- `merge_data(data_all, data_app)`
  - Perform an inner join on the time columns, which finds the overlap period in the two datasets.
  - Fill any missing values in the appliance data with zeros to avoid inaccuracies in total consumption.
  - Calculate the total hourly electricity consumption by summing the relevant columns.

2. **Main Function:**
   - Load and process the appliance data and obtain the base year parameter. With the information of base, year, load and process the total hourly data.
   - Aggregate the appliance data to an hourly base.
   - Merge the two datasets.
   - Generate plots to visualize the data and identify anomalies.

4 Analysis and Observations

First, we print the start and end timestamps for the two input datasets and the merged dataset. The start and end timestamps for the appliance dataset are 2013-06-07 12:00:00 and 2013-09-18 00:00:00 respectively. The start and end timestamps for the household consumption dataset are 2013-01-01 01:00:00 and 2014-01-01 00:00:00 respectively. After the merging, the start timestamp is 2013-06-07 12:00:00, and the end timestamp is 2013-09-18 00:00:00, which are consistent with the overlap time interval.

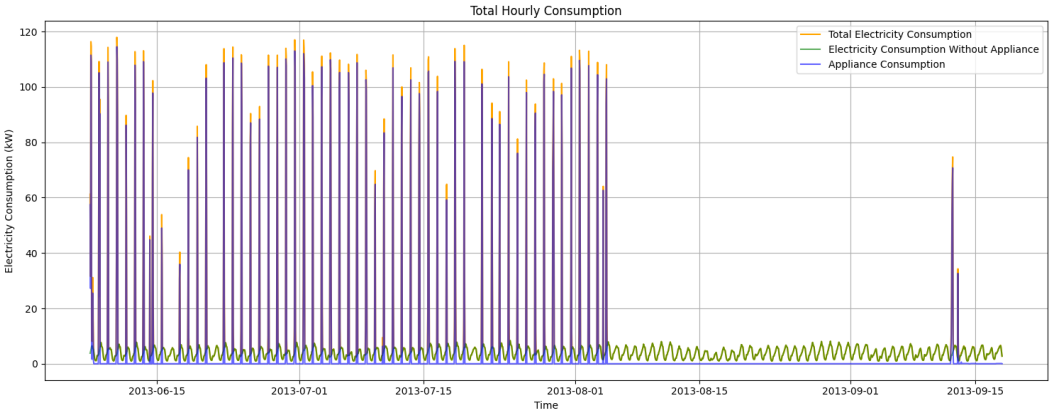

Figure 1: Visualization of Electricity Consumption

The total consumption with the new appliance exhibits a regular, high-frequency pattern, with many peaks from early June to mid-August and a few in September. The household consumption shows

a much lower and stable consumption pattern between 0kW and 10kW, remaining a seasonality and do not change much throughout the time period. The appliance consumption shows frequent and significant peaks from early June to mid-August, and a stable close-to-zero consumption since mid-August to September. These peaks align closely with the peaks in the total consumption, indicating that the appliance consumption significantly contributes to the pattern of the resulting total consumption.

The substantial drop in the appliance consumption starting in mid-August might suggest seasonal changes in appliance usage, and this significantly impact the total consumption of electricity for the household.

4.1 Consumption By Hour

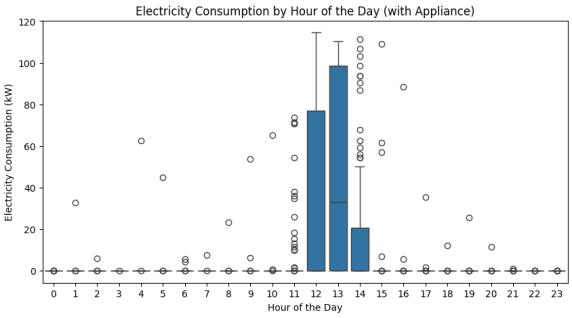

Figure 2: Electricity Consumption by Hour of the Day (Total)

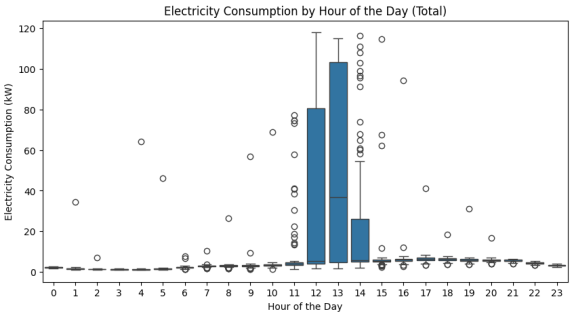

Figure 3: Electricity Consumption by Hour of the Day (Before Merge)

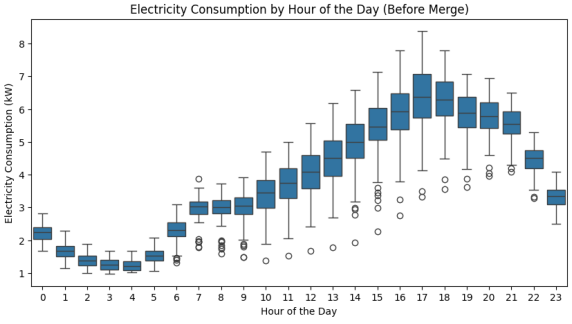

Figure 4: Electricity Consumption by Hour of the Day (with Appliance)

Without the appliance, the pattern gradually increases, peaking in the late afternoon and evening, followed by a gradual drop at night. With the appliance, there is a sharp, isolated peak around 11:00 AM to 12:00 PM, which is also reflected in the total consumption pattern.

4.2 Consumption By Day

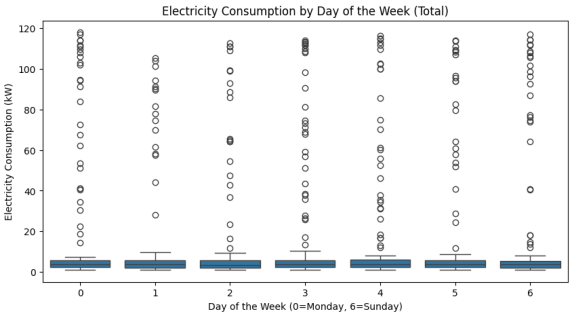

Figure 5: Electricity Consumption by Day of the Week (Total)

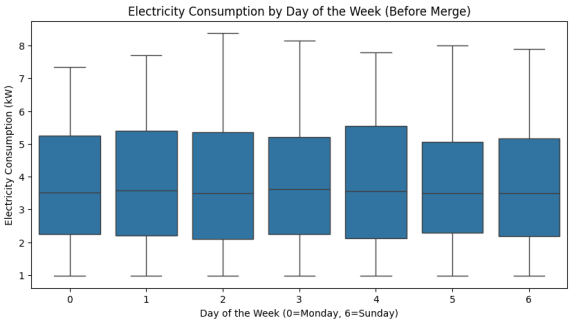

Figure 6: Electricity Consumption by Day of the Week (Before Merge)

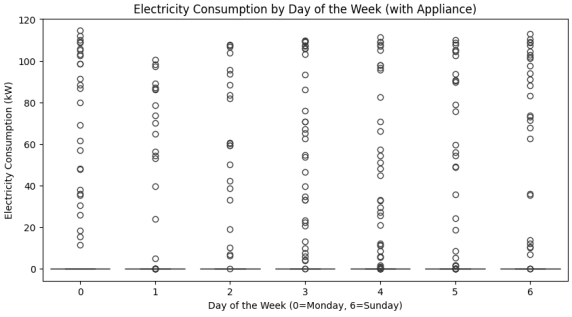

Figure 7: Electricity Consumption by Day of the Week (with Appliance)

Without the appliance, the pattern is stable, predictable, and relatively uniform throughout the week. With the appliance, many outliers indicate heavy but infrequent use. The pattern remains uniform across the week. The total consumption also shows a uniform pattern, with outliers and frequent peaks.

4.3 Consumption By Month

Without the appliance, the pattern is relatively uniform throughout the months. With the appliance, many outliers indicate heavy but infrequent use in June and July, while the number of outliers drops in August and September. The total consumption also shows a consistent pattern, with numerous outliers and frequent peaks in June and July, and fewer outliers and smaller peaks in August and September.

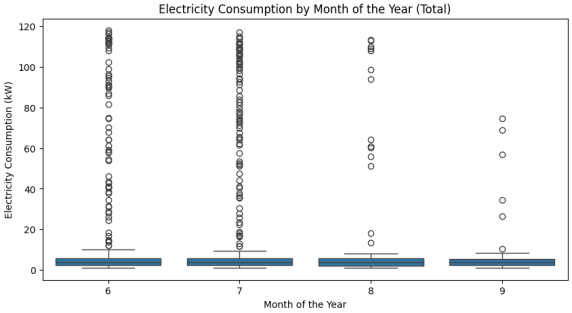

Figure 8: Electricity Consumption by Month (Total)

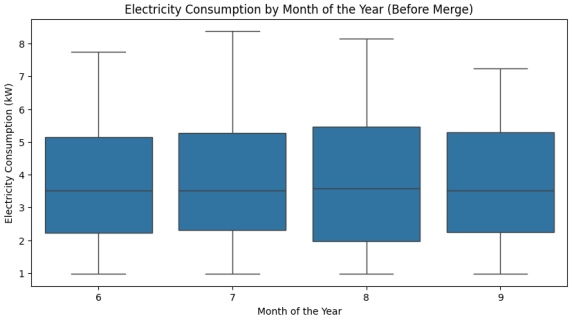

Figure 9: Electricity Consumption by Month (Before Merge)

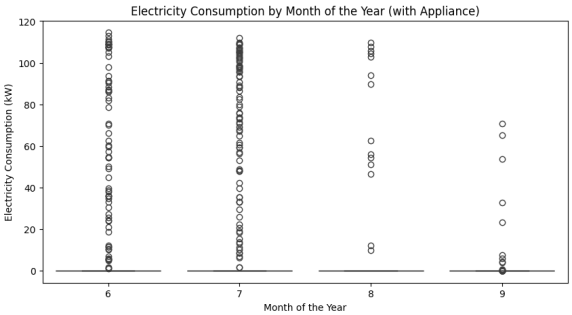

Figure 10: Electricity Consumption by Month (with Appliance)

4.4 Anomaly Detection

Among the three time period examinations, the hourly distribution of appliance consumption (Figure 3) reveals a clearer pattern and outliers. The appliance is used heavily between 11:00 AM and 12:00 PM, without much usage during other hours of the day. These frequent peaks during specific hours, along with the numerous outliers in other hours, indicate irregular use, which can be considered anomalies in the dataset.