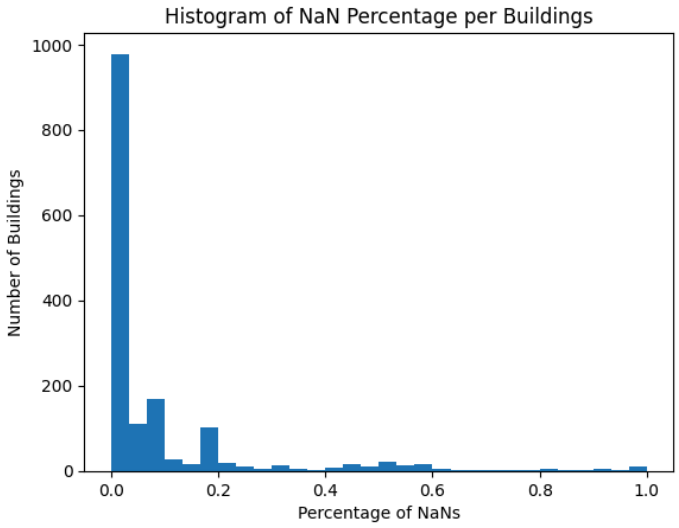
1. Data Cleaning Report

In order for us to use the data correctly, we must first begin with cleaning the data. The majority of data often comes with impurities such as missing values, anomalies, duplicates, and many more. It is crucial for a data scientist to identify these impurities and address them so that predictive models later on can be reliable and useful. The dataset was sourced from the GitHub repository linked in this project: https://github.com/buds-lab/building-data-genome-project-2/tree/master/data.

We are specifically working with the electricity_cleaned.csv file, which differs from the raw version only in how it handles errors from the meter reader. A meter reading of zero may indicate a faulty meter, so these values are converted to NaN rather than being treated as valid inputs. The code used for this section of the report can be found at https://github.com/whosphong/Data_Olympiad/tree/main/Data%20Cleaning.

The dataset contains 17544 hourly observations collected from meters located in various sites, buildings, and locations. Each observation represents a single hour of electricity usage data between the years 2016 and 2018, covering a two year time period. There are a total of 19 unique sites across the United States and Europe, and each of these sites contains buildings with 16 unique electricity usage.

One noticeable issue with this dataset is the volume of missing data originating from the meter readings. In total, across all sites and buildings, there are a staggering 2471853 missing values that need to be addressed. To better understand the distribution of missing values, it is helpful to visualize a histogram showing the percentage of missing values for buildings at each site.



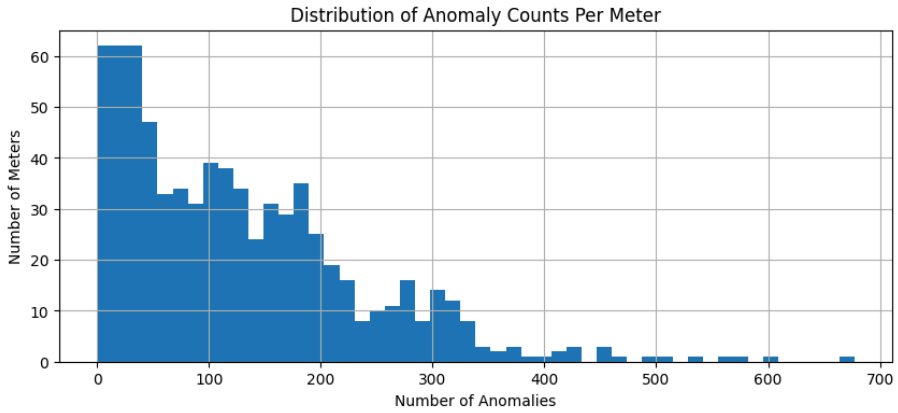
Although most missing values will be addressed through imputation, a problem arises when a building's meter readings consist of more than 40 percent missing values shown in the Figure. In such cases, it becomes necessary to remove those buildings from further analysis, as

imputing that much missing data could result in misrepresentation of the building's actual usage trends. We found that there are a total of 123 buildings that have missing values in 40 percent or more of their observations over the span of two years. These buildings were removed, leaving us with a remaining 1455 building locations.

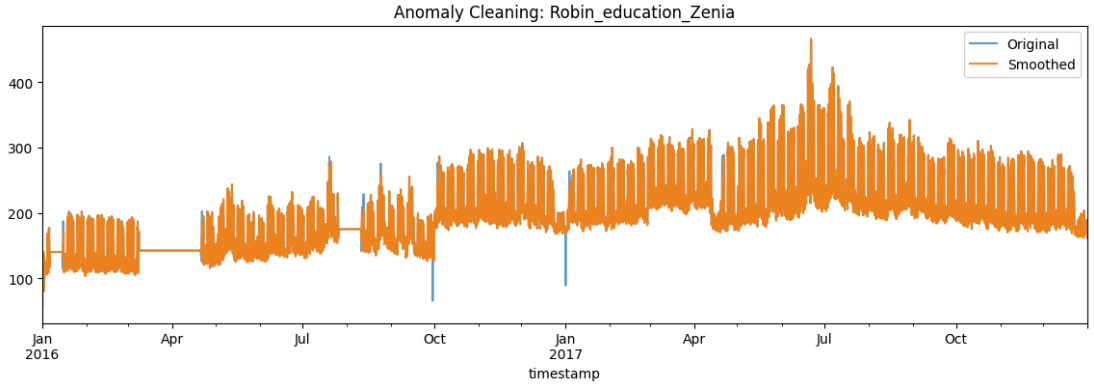
Beyond the total percentage of missing values, it is also important to consider the duration of consecutive missing periods. For example, if a building has a full week of missing data, it becomes difficult to recover the underlying usage pattern through imputation. To address this, we applied a weekly rolling window approach to further screen the data. We defined a threshold of 120 valid hourly readings per week as the minimum required for a meter to be considered reliable. This threshold allows for some gaps while still ensuring sufficient coverage. If a meter failed to meet this threshold in any weekly window, it was excluded from the analysis.

After applying this filter, we were left with a dataset containing 734 building meters, each with a complete and reliable time series over the full two year period. The final shape of the dataset is (17544, 734), representing 734 meters with 17544 hourly observations each. For the remaining buildings where some observations still contain missing values, we addressed them using a combination of backward fill and forward fill methods. Forward fill works by filling missing values with the last known valid observation, while backward fill uses the next valid observation to fill in the gaps. This allows us to maintain temporal consistency without drastically altering the underlying trends of electricity usage.

We also looked to identify any duplicated values within our dataset which came up and yielded as none. Secondly we want to look at anomalies where the meter reading seems to be faulty. We defined an anomaly as any meter reading that falls outside of three standard deviations from that building’s typical usage distribution. Based on this definition, a total of 94377 anomalies were identified across all meters. These values tend to fall in the extreme upper range, often beyond the 99.9th percentile, and are unlikely to reflect actual electricity usage. Instead, they are more indicative of short-term noise or faulty meter behavior, such as sudden spikes caused by data logging errors or sensor malfunctions. Including these values in the dataset would risk skewing the results of downstream analysis, especially in forecasting models where outliers can lead to poor generalization.



To address this, we replaced these anomalous values using a rolling median. A rolling median is calculated across 24 hours and provides an estimate of local behavior. By substituting the anomaly with the corresponding rolling median, we preserve the variability and trend in the data while effectively removing statistical noise. The figure below demonstrates an example of an original meter reading with a noticeable anomaly (Blue), followed by the cleaned version where the spike has been smoothed using the rolling median method (Orange).



Outliers such as values outside the Interquartile Range (IQR [25%-50%]) and are not anomalies were not removed, as these could represent significant or meaningful observations rather than errors. For instance, a sudden spike in electricity usage may occur when a large piece of equipment is turned on in a commercial building, or when a heating system activates during extreme cold in a residential site. Similarly, lower-than-expected values could reflect times when a building is unoccupied, such as during holidays, weekends, or maintenance shutdowns. These points could be addressed when making the predictive model via transformations, scaling, and/or winsorization.