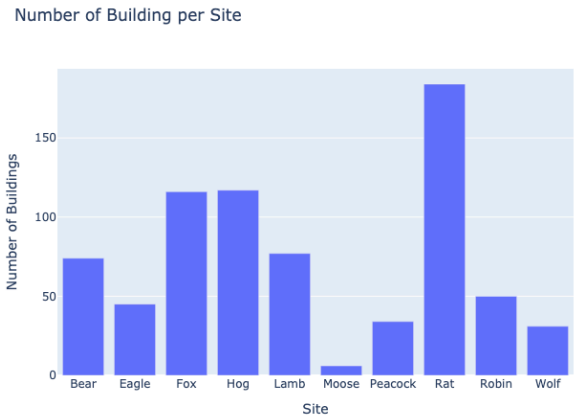
Exploratory Data Analysis (EDA)

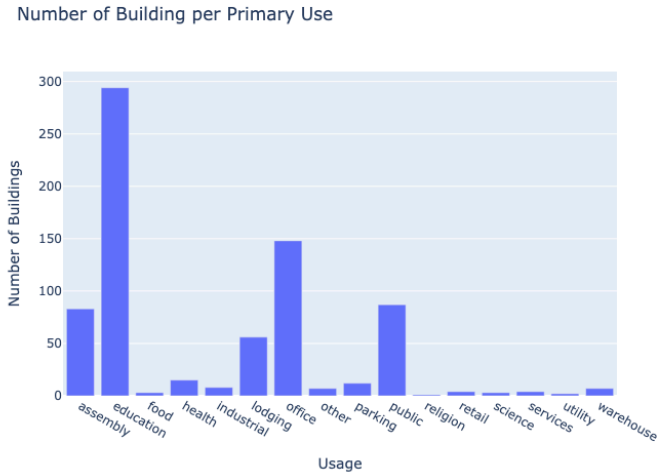
After data cleaning, we explored the dataset to understand the distributions and trends in energy consumption. We categorized the meter data by site and primary electricity usage.

Before Data Manipulation

Distribution

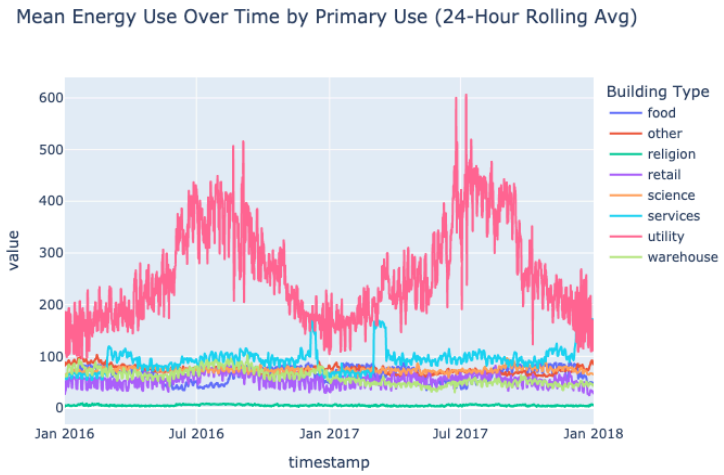
By plotting bar charts,





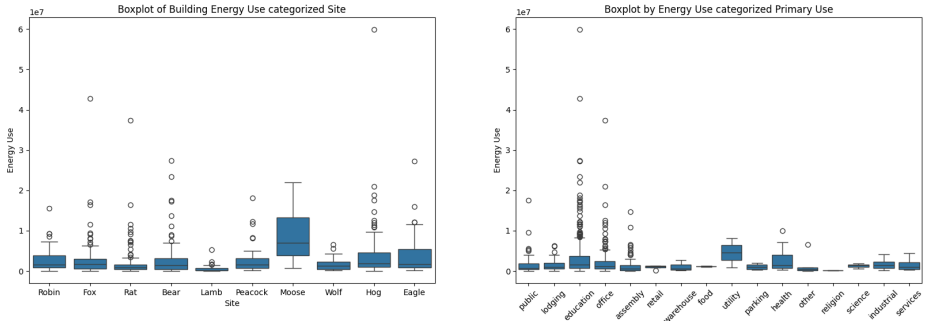
We observed that there are **Data imbalance** problems occur in both categorization by site and energy primary usage. For this site, “Rat” has the highest number of buildings (184, ~25% data) and Moose has the fewest (6, ~0.77% data). For primary use, “education” is the largest (294, ~40%) data and 8 categories of primary use have less than or equal to 8 buildings (~1.09% data).

Focused on primary usage categories with ≤8 buildings, we investigated whether categories have distinct energy trends to decide whether to combine them into a single group or not.

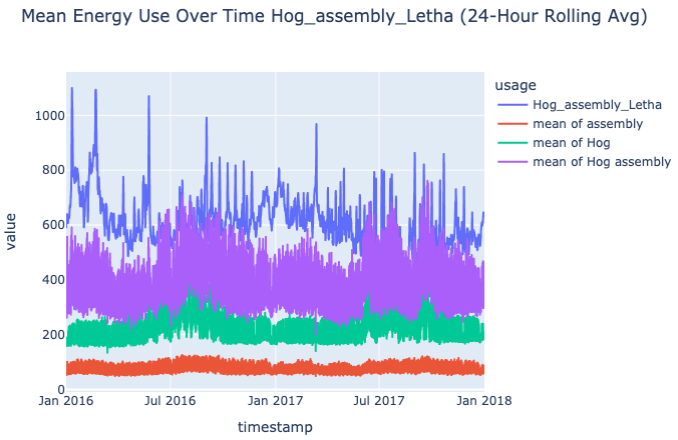


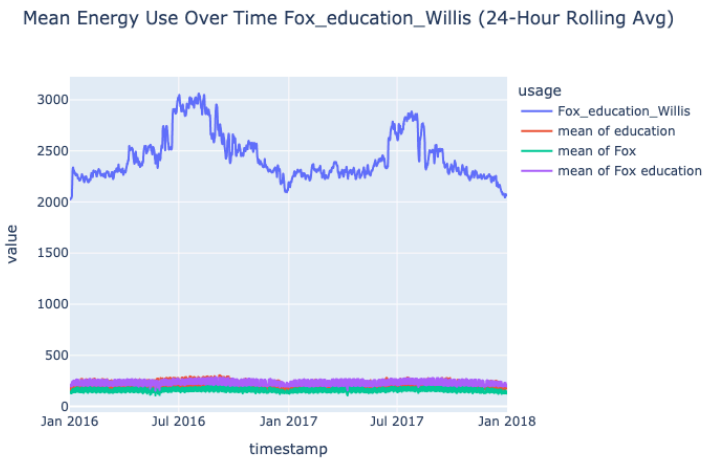
To assess this, we plotted line graphs of the primary use categories with few buildings. We observed that the “utility” category has consistently high energy consumption and decided to keep them as a single category. In contrast, “religion” has consistently low energy use (~ 5 kwh) at all times, making its contribution negligible compared to other categories. Moreover, data in category “services” have sharp increases in energy consumption during January and April, followed by decrease at different times, which does not show an obvious seasonal pattern. Therefore, we decided to (1) keep “utility” as a single category, (2) remove data in the “religion” category and (3) combine the remaining 6 categories, i.e. [], as category “other”.

By plotting box plots,



We observed some outliers and decided to look into the time series data of the outliers and compare it with the mean of energy use with the same site and/or primary use.





There are 2 kinds of outliers, the first kind of outlier (left side plot) is close to the mean of same primary use /same site and primary use and we classify these data as good to keep. The second kind of outlier is much higher than the means of the same site and/or primary use. As they have high energy use, they may have higher traffic of customers and are critical for us for predicting future energy usage and deriving business insights to reduce energy waste and cost. Also, these outliers with high energy use may have and As a result, we decided to keep these checked data.

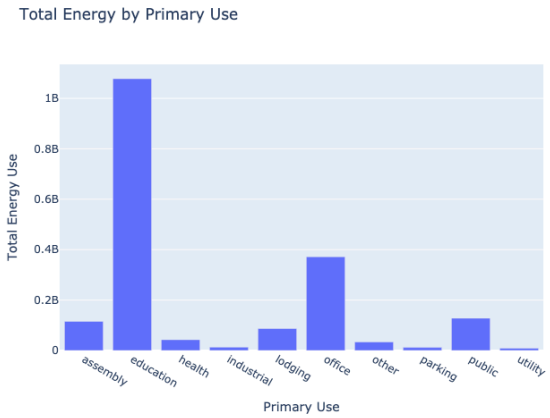
Data Manipulation

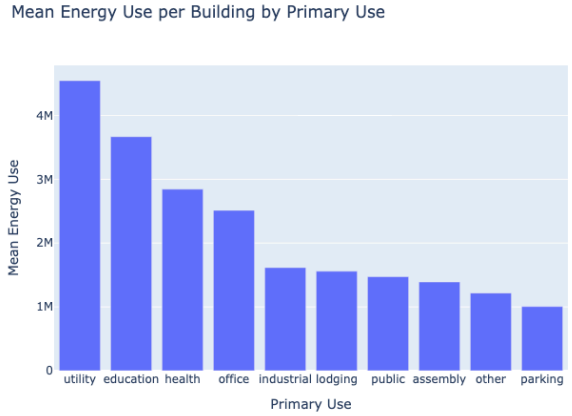
To conclude the above, we decided to remove data in the “religion” category and categories [] as category “other”.

Investigate Data after Data Manipulation

Data Distribution

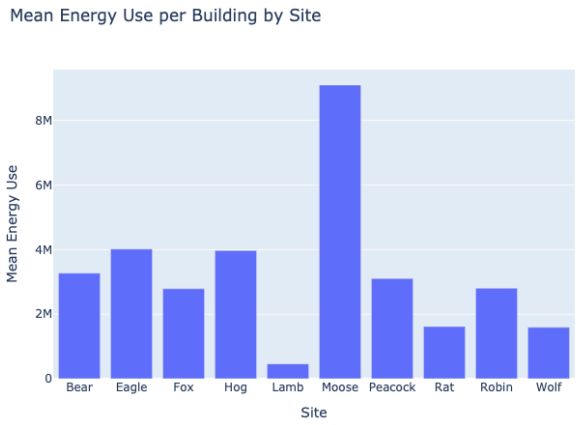
By plotting the bar charts in terms of primary use,

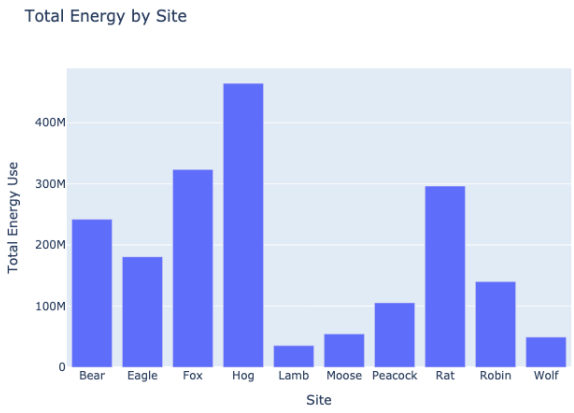




For total energy use, “education” type buildings have the highest total energy use and are followed by “office” and “public”. For mean energy use, “utility” type buildings have the highest mean energy use and are followed by “education”, “health” and “office” and their energy uses are significantly higher than the remaining.

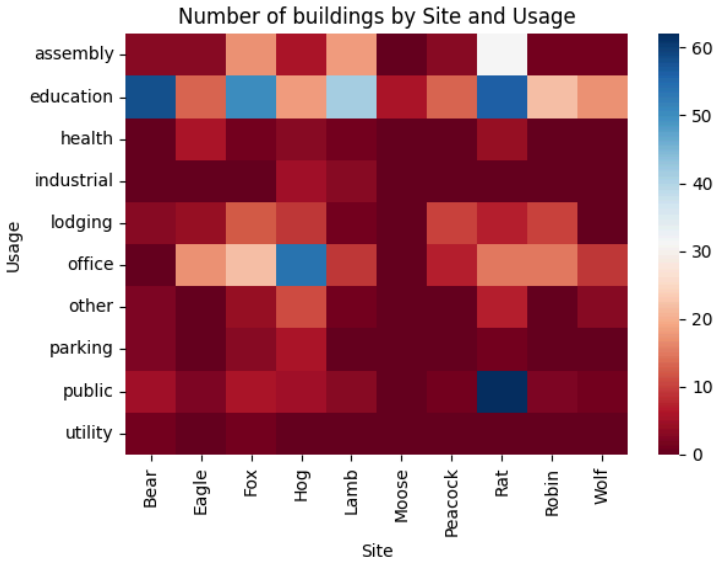
By plotting the bar charts in terms of site,

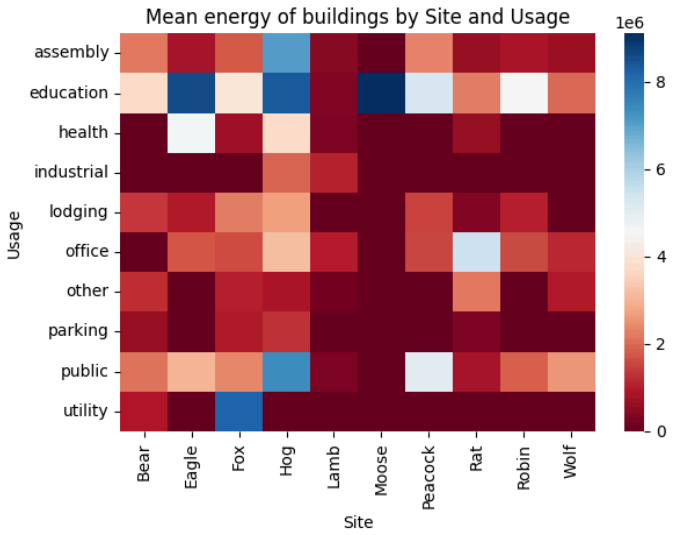


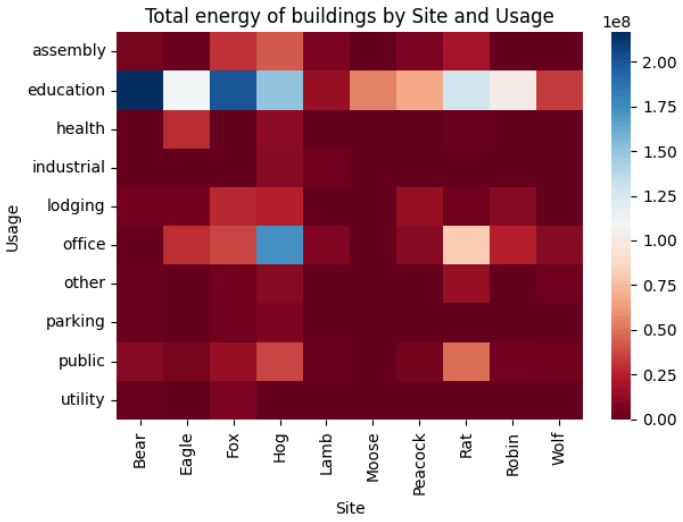


For total energy use, “Hog” site buildings have the highest total energy use and are followed by “Fox” and “Rat”. For mean energy use, “Moose” site buildings have the highest mean energy use and it is significantly higher than the others, and it is followed by “Hog” and “Eagle”.

By plotting heatmap,



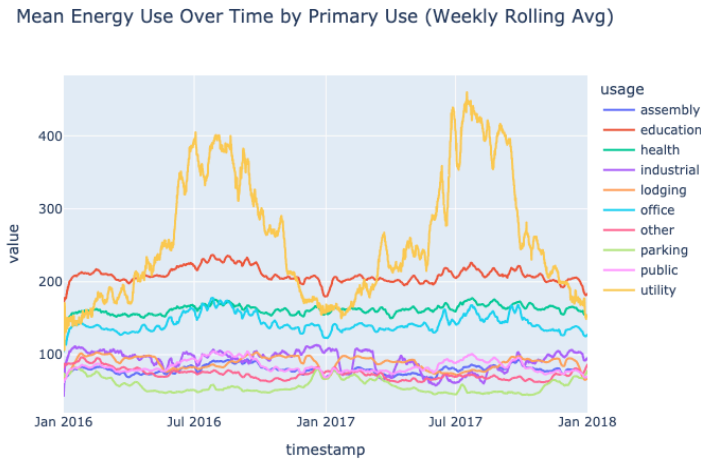


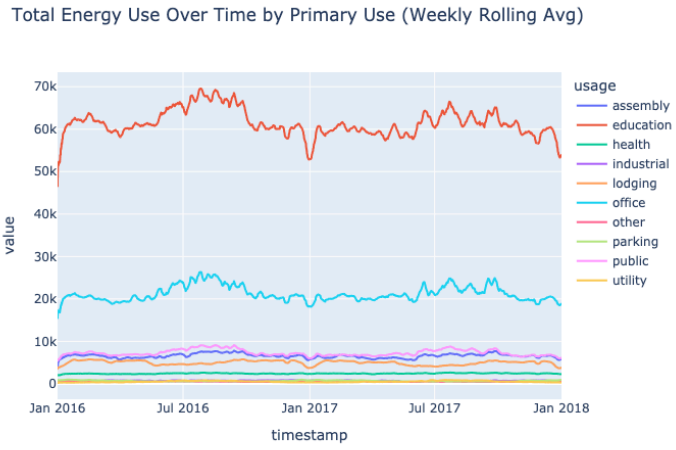


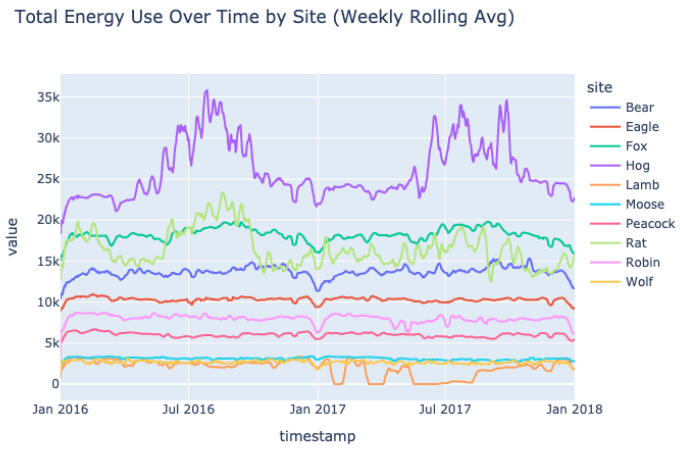
We observed that there are only education buildings on site “Moose”. Also, we observed that “education” buildings in all sites, “health” buildings in “Eagle”, “utility” buildings in “Fox” and “assembly”, “public” buildings in “Hog” havr. Moreover, there are sin

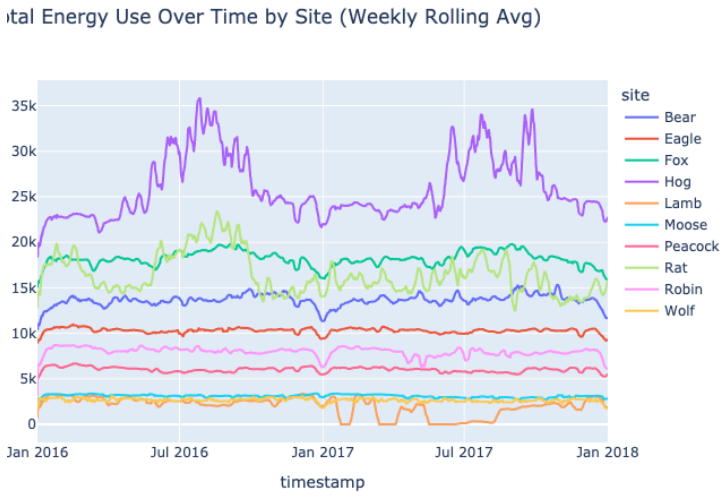
<u>Trend</u>

By plotting line graphs,









We observed the overall trends, changes of energy usages and We observed that the energy consumption of “utility” follows seasonal or temperature changes.

To conclude, we find seasonal changes in education and utility. Moreover, we observed high mean energy consumption in education, utility and public.