Data Analysis and Machine Learning Assignment

Name: Mohammed AL shuaili

ID: 20106181

Date: 26/11/2024

1. Question 1:
   1. Electricity and Gas usage 2023/2024: These datasets record daily gas and electricity consumption for the year 2023/2024, with usage measured at 15-minute intervals throughout each day and structured as follows:

Rows: 365 entries, one for each day of the year.

Columns (98 columns):

* Date: Represents the day (Interval attribute as the difference between dates has meaning and we can add days to a date and dates have meaningful distances).
* Values: Likely represents the total number of readings ((integer) Ratio because it has a meaningful zero and supports operations like addition, subtraction, multiplication, and division).
* Time Interval Columns (e.g., 00:00, 00:15, ...): Represent electricity usage in kilowatts (or similar units) at specific 15-minute intervals for 24 hours (Ratio).

The datasets have Some missing values at some intervals and date. This may indicate issues like incomplete measurements or errors during data collection.

* 1. If the dataset is recorded for urban households, we can apply analysis to understand how urban households can reduce energy and gas waste by identifying inefficiencies in daily consumption patterns and correcting abnormal usage behaviours. By analysing electricity and gas consumption at 60-minute intervals, we can identify periods of excessive usage that may indicate inefficiencies, such as:
* Excessive consumption during off-peak hours: Identifying when gas or electricity is being used unnecessarily during periods of low activity (e.g., late-night hours when households are inactive).
* Identifying equipment inefficiencies: Spotting unusual consumption patterns that could suggest faulty or inefficient appliances (e.g., malfunctioning heating systems or lights left on).
* Behavioural analysis: Understanding how human behaviours, such as leaving lights or heating systems running unnecessarily, contribute to higher and unnecessary energy consumption.

Correcting these inefficiencies can lead to significant energy savings, reduced overall consumption, and the analysis can also provide actionable insights and recommendations for households to optimize their energy usage, and reduce unnecessary costs.

* 1. This can be broken into multiple steps:

1. Data Cleansing:

Data cleansing ensures that the data is accurate, complete, and consistent before analysis. The steps involved in this process are:

1. Handling Missing Values:

* Identify Missing Data: Check for NaN, None, or empty values in the dataset.
  + Example: in the Electricity dataset 2023, to find the missing data the following code is used

missing\_values = Edf23.isnull().sum()

# Filter columns with missing values only

columns\_with\_missing = missing\_values[missing\_values > **0**]

columns\_with\_missing.sort\_values(ascending=False)

This shows that all columns include missing data. Then to get an overall idea about these columns we can run.

missing\_columns = Edf23.columns[Edf23.isnull().sum() > 0]

# Describe only those columns with missing data

missing\_data\_description = Edf23[missing\_columns].describe(include='all')

missing\_data\_description

Each column can then be analysed, and missing values can be imputed by identifying the mode for that specific time interval. This approach is most appropriate since these values represent units of electricity and replacing them with the mode (being the most frequently occurring unit) provides a practical solution.

The same can be applied to Gas usage 2023. However, Looking at Electricity/Gas Usage 2024, we can see that the datasets are incomplete. This is expected, as the datasets were recorded before the end of 2024. However, we must replace these values or remove them to prevent them from causing errors in the data summary and visualizations. In this case the best way is to remove these values as we can gain a rough idea of the dataset using the recorded values and just predicting these values won’t give us much information.

1. Identifying and Handling Outliers:

To identify outliers we can use boxplot, Jitter plot or histograms, but we must first reshape the dataset to prepare it for visualisation.

* Handling Outliers:
  + We can either remove them, cap them at a specific value, or transform the data.

1. Data Reshaping:

Data reshaping is the process of transforming the data into a format that suits the analysis or model you're working with. So, for the above issue the best suitable way is to merge gas with electricity. This is because the goal is to reduce overall energy consumption and combining both energy types into a single dataset allows for a more comprehensive view of the household's total energy usage.

1. A screenshot of a graph

   Description automatically generatedChage columns to represent hourly instead of every 15 minutes. Considering the below, since the values are similar across each hour, it makes sense to only consider hourly readings. (the standard deviation is small)
2. We apply the same process to the Gas dataset to ensure consistency, and then merge the gas data with the electricity data for analysis. Using this merged dataset, we can compare gas and electricity usage at any given time throughout 2023. Additionally, we can calculate the total consumption at each time point, providing an estimate of the overall units consumed, whether from gas or electricity
   1. The head of the dataset

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 02:00\_Gas | 05:00\_Gas | 08:00\_Gas | 11:00\_Gas | 14:00\_Gas | 17:00\_Gas | 20:00\_Gas | 23:00\_Gas | 02:00\_Electricity | 05:00\_Electricity | 08:00\_Electricity | 11:00\_Electricity | 14:00\_Electricity | 17:00\_Electricity | 20:00\_Electricity | 23:00\_Electricity |
| 1/1/2023 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 27 | 33 | 26 | 28 | 25 | 27 | 26 | 27 |
| 2/1/2023 | 0 | 0 | 22 | 22 | 11 | 11 | 22 | 0 | 28 | 28 | 30 | 30 | 32 | 32 | 29 | 28 |
| 3/1/2023 | 0 | 0 | 187 | 253 | 121 | 88 | 0 | 0 | 28 | 25 | 34 | 54 | 51 | 49 | 26 | 25 |
| 4/1/2023 | 0 | 0 | 110 | 187 | 154 | 110 | 0 | 0 | 26 | 27 | 33 | 66 | 64 | 55 | 28 | 26 |
| 5/1/2023 | 0 | 0 | 110 | 198 | 132 | 121 | 0 | 0 | 25 | 26 | 33 | 67 | 69 | 60 | 31 | 30 |
| 6/1/2023 | 0 | 0 | 187 | 33 | 33 | 121 | 0 | 0 | 30 | 31 | 38 | 64 | 65 | 57 | 28 | 27 |
| 7/1/2023 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 | 27 | 23 | 26 | 26 | 27 | 27 | 26 | 28 |
| 8/1/2023 | 0 | 0 | 0 | 0 | 22 | 0 | 0 | 0 | 26 | 25 | 25 | 26 | 27 | 28 | 31 | 29 |
| 9/1/2023 | 0 | 0 | 143 | 242 | 198 | 110 | 0 | 0 | 30 | 30 | 37 | 76 | 74 | 64 | 32 | 28 |
| 10/1/2023 | 0 | 0 | 209 | 132 | 110 | 88 | 0 | 0 | 29 | 27 | 37 | 69 | 72 | 65 | 32 | 28 |

General summary statistics with description (hourly):

1. Gas usage 2024:

* The provided dataset summarizes gas usage in 2024, measured at specific time intervals throughout the day (e.g., 00:00, 00:15, etc.). The statistics include key metrics such as the number of observations (count), mean (mean), standard deviation (std), and percentiles (min, 25%, 50%, 75%, max) for each time interval.
* Each interval has a consistent number of observations, with a count of **236** across all time periods (after cleaning). The mean gas usage varies significantly throughout the day, starting at **10.30 units** at midnight (00:00), peaking at **101.84 (102) units** at 07:30, and tapering off to lower values like **9.18 (9) units** by 22:30. This trend suggests higher consumption during the morning hours, likely coinciding with household activities, and reduced usage during nighttime.
* The variability in gas usage, represented by the standard deviation (std), also fluctuates. For instance, at 00:00, the standard deviation is **14 units**, increasing significantly to **86 units** at 08:45, indicating more inconsistent consumption patterns during peak hours. The minimum usage (min) across all intervals is consistently **0 units,** highlighting periods of no consumption.
* The dataset's percentiles provide further insights. For example, the 25th percentile (25%) remains **0 units** for many intervals, showing that at least a quarter of the readings indicate no gas usage. However, during active periods like 07:30, the 75th percentile (75%) rises to **154 units**, and the maximum (max) reaches **550 units** at 10:00, showing periods of exceptionally high consumption.
* This dataset captures both inactive and peak usage periods, making it valuable for identifying inefficiencies in gas consumption. However, the irregular intervals (e.g., 01:15, 02:30) may require standardisation for more comprehensive time-based analysis.

1. Electricity usage 2024:

* Standard deviation (std) shows variability in electricity usage. During the peak hours of 12:00 PM to 2:00 PM, variability is highest (e.g., 16.55 units at 12:00 PM), but it significantly decreases by late evening, reaching 2.87 units at 11:00 PM.
* Minimum (min) and Maximum (max) values show a wide range of usage, from as low as 8 units at noon to as high as 81 units at 11:00 AM.
* The 25% percentile consumption is steady at 36 units from 12:00 PM to 2:00 PM, dropping to 33 units by late evening (11:00 PM).
* The 50th percentile (median) peaks at 64 units at 12:00 PM, gradually declining through the afternoon and evening to 34 units by 11:00 PM.
* The 75th percentile, indicative of higher consumption, mirrors this pattern, peaking at 69 units from 12:00 PM to 2:00 PM and dropping to 36 units by 11:00 PM.

C) Electricity usage 2023:

* This dataset highlights consistent patterns in mean consumption, with a gradual increase from 30.28 (30) units at midnight to a peak of 58.40 (58) units by 12:00 PM. A slight dip follows, with the mean at 58.22 (58) units by 2:00 PM.

**Key Observations:**

* Standard deviation (std) reflects higher variability in late mornings and afternoons, peaking at 20.63 units at noon.
* Minimum (min) values range from 2 units (e.g., 12:00 PM) to 11 units (9:00 AM).
* Median (50%) values steadily climb, reaching 68 units at noon, indicating the central tendency aligns with peak usage times.
* Maximum (max) usage occurs at 12:00 PM (86 units), showcasing the busiest period.

This data indicates a similar daily cycle to 2024, with low usage overnight and peaks during late mornings and midday

D) Gas usage 2023:

* This data illustrates a marked rise in consumption during morning hours and a peak in late mornings, followed by a decline towards the afternoon and evening.

Key Observations:

* Mean Consumption:
* Low overnight: Starts at 2.50 (3) units (00:00), increasing to 5.39 (5) units (06:00).
* Peak usage: 71.03 units (08:00) and 86.28 units (09:00), with a gradual decline to 54.82 units (14:00) and 23.7 at (17:00).
* Standard Deviation (std): Significant variability observed during the peak hours (08:00–09:00, 68.17–93.81 units).
* Minimum (min) values: Gas usage is 0 units throughout all time intervals for at least some data points.
* Median (50%) values: Peaks at 55 units (08:00), then declines gradually.
* Maximum (max) values: Highest usage at 330 units (08:00), tapering off to 275 units by 14:00.

This dataset suggests high morning demand, likely reflecting heating or cooking activities, with usage tapering off as the day progresses.

Now, let’s examine each dataset.

A graph of colored lines

Description automatically generatedA graph of colored lines

Description automatically generatedPlotting the Electricity mean values (200 samples, first and last) (hourly)

Figure 1 Electricity usage 2023 first 100 samples

Figure 2 Electricity usage 2024 first 100 samples

A graph of a graph

Description automatically generated with medium confidenceA graph of a graph

Description automatically generated with medium confidence

Figure 3 Electricity usage 2023 last 100 samples

Figure 4 Electricity usage 2024 last 100 samples

Analysing the mean values for electricity consumption in 2023 and 2024, we observe that weekdays generally exhibit higher consumption compared to weekends. Additionally, several unexpected values are apparent, particularly in Figures 3, 2, and 1. These values could either be normal variations or potential outliers.

A graph of a graph

Description automatically generated with medium confidence

Figure 5 the mean values of electricity usage throughout a day

A graph of a gas usage

Description automatically generated with medium confidence

Figure 6 shows the mean values of Gas usage over the course of a day

Now let’s examine the Density:

A graph of distribution of electricity

Description automatically generated

Figure 7 shows Hourly mean distribution for electricity

A graph of gas prices

Description automatically generated with medium confidence

Figure 8 Shows hourly mean distribution for gas usage

Figure 7 and Figure 8 represent kernel density estimators for electricity and gas mean hourly values, respectively. From the figures, we can identify high density around 31 and 35 units of electricity in 2023 and 2024, reflecting dominant consumption patterns across these periods. For gas, notable peaks are observed near 10 and 18 units in 2023 and 2024, suggesting shifts in usage patterns. The differences in density curves indicate variations in distribution between the two years, highlighting potential changes in energy demand over time.

In order to create a boxplot to detect outliers and gain more information about the datasets, we can create a new data frame with the mean value of each as follows:

data\_for\_boxplot = pd.DataFrame({

'Year': ['2023'] \* len(hourly\_means\_2023\_electricity) + ['2024'] \* len(hourly\_means\_2024\_electricity),

'Electricity': hourly\_means\_2023\_electricity.tolist() + hourly\_means\_2024\_electricity.tolist(),

'Gas': hourly\_means\_2023\_gas.tolist() + hourly\_means\_2024\_gas.tolist()

})

The new DataFrame, `data\_for\_boxplot`, is created to organize the data for effective visualization. The `Year` column is included to indicate whether the data corresponds to 2023 or 2024, allowing for clear differentiation between the two years. The `Electricity` and `Gas` columns store the respective hourly mean values for each year, which are converted to lists using the `.tolist()` method and then concatenated to combine data from both years into a single structured format.

A diagram of a gas usage

Description automatically generated

Figure 9 boxplots of hourly mean gas usage in 2023/24

A diagram of a box diagram

Description automatically generated

Figure 10 boxplots of hourly mean Electricity usage in 2023/24

Figures 9 and 10 present boxplots of hourly mean electricity and gas usage, respectively. From Figure 9, electricity usage in 2023 shows a slightly broader interquartile range compared to 2024, suggesting greater variability in consumption during this period. The median values are relatively stable between the two years, with a small shift observed toward slightly higher usage in 2024. In Figure 10, gas usage shows that the median value in 2024 is higher than in 2023, alongside a noticeable narrowing of the interquartile range, reflecting a more uniform distribution of hourly gas consumption in 2024. These visualisations highlight potential improvements in energy efficiency or shifts in consumption behaviour over time.

* 1. To address the challenge of helping urban households reduce energy and gas waste a strategic use of visualization techniques can be highly effective. Visualisations can translate complex data into understandable insights for the general audience, prompting actionable changes in consumption patterns.