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 visualisation techniques can be highly effective. Visualisations can translate complex data into understandable insights for the general audience, prompting actionable changes in consumption patterns.
     1. Understand the Data Distribution using Fig 7 and 8:
* The kernel density plot for electricity usage reveals distinct patterns for 2023 and 2024. In both years, the highest density is observed around 31–35 units, indicating consistent demand in this range. However, in 2024, an additional peak is noticeable near 55 units, suggesting increased consumption during certain hours. This shift may be attributed to new appliances, behavioral changes, or increased energy usage during specific periods. The concentration of high-density values around certain ranges could indicate inefficiencies, such as simultaneous use of multiple high-energy appliances or unoptimized energy schedules.
* For gas usage, the density plot shows a more efficient and concentrated usage pattern for 2023, with values primarily below 10 units. In contrast, 2024 exhibits a broader distribution with higher peaks in the 10–20 range and irregular spikes up to 40–60 units. This suggests an increase in variability and possible inefficiencies in gas consumption, such as seasonal heating demand or system malfunctions. The spread in 2024 highlights the need to investigate abnormal patterns, optimize heating systems, or address potential leaks to reduce waste and improve efficiency.

B) Compare Energy Usage by Day

A graph of blue rectangular objects with white text

Description automatically generated

Figure 11Comparison of Daily Mean Gas and Electricity Usage Across the Week in 2023

This boxplot provides a detailed comparison of daily mean energy usage for gas (blue) and electricity (orange) across the days of the week (in 2023). Gas usage shows a clear pattern of higher consumption and variability on weekdays (Monday to Friday), with the median usage ranging from 20 to 40 units. The interquartile range (IQR) is also broader on weekdays, indicating a wider spread in consumption levels, while outliers suggest occasional spikes in usage. On weekends (Saturday and Sunday), gas usage drops significantly, with medians around 3 units, minimal variability, and tighter IQRs, reflecting reduced heating or cooking activities.

Electricity usage, on the other hand, remains consistent throughout the week, with median values around 30–50 units across all days. The IQR for electricity is much narrower than that for gas, indicating stable and predictable consumption patterns. A slight decrease in electricity usage at weekends suggests a shift in activities, possibly due to more time spent at home. This contrast between gas and electricity usage highlights opportunities to investigate and optimise weekday gas consumption while maintaining efficient and stable electricity use.

Figure12 illustrates the daily mean gas (blue) and electricity (orange) usage across the days of the week for 2024. Gas usage shows a consistent pattern of higher median values during weekdays (Monday to Friday), with medians ranging from 30 to 50 units and significant variability, as indicated by the wider interquartile ranges and presence of outliers. On weekends (Saturday and Sunday), gas usage drops sharply, with medians below 10 units and minimal variability. Electricity usage remains stable throughout the week, with median values consistently around 40–45 units, and a slight decrease on Weekends. The narrower interquartile ranges and fewer outliers for electricity indicate predictable and consistent consumption

A graph of a number of blue rectangular objects

Description automatically generated with medium confidence

Figure 12 Comparison of Daily Mean Gas and Electricity Usage Across the Week in 2024

The comparison shows that gas usage in 2024 had slightly higher weekend medians and more variability than 2023, while electricity usage remained stable across both years. Since 2023 provides clear patterns for analysis, it is sufficient for identifying trends and inefficiencies.

1. Joint Relationships Between Gas and Electricity

A diagram of gas consumption

Description automatically generated

Figure 13 Joint Relationship Between Daily Mean Gas and Electricity Usage

The scatter plot shows a moderate positive correlation between gas and electricity usage (Pearson: 0.60, Spearman: 0.67). Two dense clusters indicate consistent simultaneous usage at low and moderate levels, while higher gas usage aligns with increased electricity usage. Outliers may highlight abnormal patterns or isolated inefficiencies, suggesting opportunities to optimise combined energy consumption.

**Message for a General Audience**

Our energy consumption analysis highlights key patterns that can help households reduce waste and improve efficiency:

1. Electricity Usage: Electricity consumption remains consistent throughout the week, with slightly lower usage on weekends. This stability suggests well-regulated usage patterns but also an opportunity to further optimise weekday energy use by reducing non-essential appliances.
2. Gas Usage: Gas consumption is higher and more variable on weekdays, likely due to heating or cooking needs, and drops significantly on weekends. This pattern indicates potential inefficiencies during the week that can be addressed by better scheduling and system maintenance.
3. Combined Usage: A moderate positive correlation between gas and electricity usage suggests simultaneous usage of both energy types. This could indicate inefficiencies in heating or other systems that rely on both. Outliers point to unusual consumption that may require investigation.

Actionable Insights

* Reduce Weekday Gas Waste: Identify and fix inefficiencies in heating systems or cooking routines to lower gas usage during the week.
* Optimize Combined Usage: Inspect systems that rely on both gas and electricity (e.g., water heaters) to ensure they are running efficiently and only when needed.
* Weekend Electricity Management: Consider limiting the use of non-essential appliances during weekends to reduce unnecessary electricity consumption.

By adopting these practices, households can cut costs, improve energy efficiency, and contribute to a more sustainable future.

Q2

2.1) **Project Air View Dublin Dataset Description:**

1. Road Data File:

* Purpose: Records details about the streets and highways in Dublin where air quality measurements were collected.
* Rows: 24694 entries, each representing a unique road segment.
* Columns:
  + Road Metadata:
    - road\_id, osm\_id, osm\_code, osm\_fclass, osm\_name, osm\_ref, osm\_oneway, osm\_maxspeed, osm\_layer, osm\_bridge, osm\_tunnel: Represent OpenStreetMap (OSM) identifiers, road characteristics, and traffic rules.
    - Example: osm\_fclass (nominal).(road type, e.g., service or residential) and osm\_maxspeed (maximum allowed speed in km/h).
    - osm\_layer (Ordinal): Represents vertical positioning (e.g., ground level, bridge, tunnel).
    - osm\_name (Nominal): Name of the road (if available).
  + Geometric Data:
    - the\_geom: (Ratio) Encoded as LINESTRING, specifies the geographical coordinates of road segments.
  + Pollutant Data:
    - For each pollutant (e.g., NO2, NO, CO2, CO, O3, PM2.5), the file includes:
      * points: (Ratio) Number of measurement points.
      * drives: (Ratio) Number of times the road was driven during measurements.
      * ug/m3 or mg/m3: (Ratio) Average concentration values for each pollutant.
* Observations:
  + Some values, such as pollutant concentrations, are negative, possibly indicating calibration artifacts or errors.
  + Missing data for some pollutants or roads could suggest limited coverage or measurement challenges.

2. Measurements File:

* Purpose: Captures second-by-second air quality measurements recorded by the electric Street View car.
* Rows: Each row corresponds to a 1-second measurement.
* Columns:
  + Timestamp and Location:
    - gps\_timestamp: (Interval) Timestamp of the measurement in UTC.
    - latitude, longitude: (Ratio) Geographic coordinates of the measurement location.
  + Pollutant Concentrations:
    - NO\_ugm3, NO2\_ugm3, O3\_ugm3: (Ratio) Nitrogen Oxide, Nitrogen Dioxide, and Ozone levels in micrograms per cubic meter.
    - CO\_mgm3, CO2\_mgm3: (Ratio) Carbon Monoxide and Carbon Dioxide levels in milligrams per cubic meter.
    - PM25\_ugm3: Concentration of Particulate Matter (PM2.5) in micrograms per cubic meter.
  + Particle Counts:
    - PMch1\_perL to PMch6\_perL: Size-resolved particles count per Liter, representing particles of different diameters.
  + Observations:
    - Some measurements (e.g., O3\_ugm3, CO2\_mgm3) contain NaN, indicating missing data for certain pollutants.
    - Variability in measurements reflects dynamic air quality conditions influenced by location, time, and traffic.

General Observations:

* Coverage: Measurements are representative of daytime, weekday air quality in urban Dublin streets.
* Granularity: Data allows for both street-level spatial analysis and second-level temporal analysis.
* Data Quality:
  + Negative pollutant values and missing data points suggest the need for preprocessing and calibration.
  + High temporal and spatial resolution provide valuable insights for urban air quality studies and policymaking.

1.2) Real-World Problems Using the Air View Dublin Dataset

1. Regression Problem: Predicting Nitrogen Dioxide (NO₂) Concentrations

* Problem Statement: Develop a machine learning model to predict the average NO₂ concentration (NO2\_ugm3) for a given road segment based on road attributes and traffic conditions.
* Goal: Provide actionable insights for urban planners to identify high-risk areas and implement mitigation strategies.
* Features (Independent Variables):
  + Road metadata: osm\_fclass, osm\_maxspeed, osm\_oneway.
  + Traffic-related attributes: NO2points, NO2drives.
  + Geometric attributes: the\_geom.
  + Additional pollutant concentrations: CO2\_mgm3, PM25\_ugm3.
* The target for this regression problem is to predict the average NO₂ concentration, represented by the dependent variable NO2\_ugm3. The approach involves training regression models such as Linear Regression or Random Forest Regressor to estimate NO₂ concentrations accurately. Spatial data can also be incorporated using Geographic Information Systems (GIS) features if required for better predictive performance. This model has practical applications, including forecasting pollutant levels for roads not covered in the dataset and evaluating the potential impact of proposed traffic management changes on air quality.

2. Classification Problem: Categorizing Roads Based on Pollution Risk

* Problem Statement: Classify roads into pollution risk categories (e.g., "Low", "Moderate", "High") based on pollutant concentrations and traffic conditions.
* Goal: Assist policymakers in prioritizing roads for pollution mitigation measures.
* Features (Independent Variables):
  + Road metadata: osm\_fclass, osm\_maxspeed, osm\_oneway.
  + Pollutant attributes: NO2\_ugm3, CO\_mgm3, PM25\_ugm3, CO2\_mgm3.
  + Traffic-related attributes: NO2drives, COdrives, PM25drives.
* The target for this classification problem is to categorise roads based on pollution risk, represented as a categorical variable with labels such as "Low" (pollutants below a threshold), "Moderate" (pollutants within a mid-range), and "High" (pollutants exceeding safety limits). The approach involves applying classification algorithms such as Decision Trees, Logistic Regression, or Gradient Boosting to assign risk levels effectively. Domain knowledge will be crucial in defining appropriate risk thresholds to ensure accurate classification. This model can be applied to identify areas requiring air quality monitoring and intervention, as well as to support public health campaigns and inform urban policy decisions.

1.3)