Data Analysis Final Assignment Report

Team: Solo Warrior

Mohamed Elsayad & Mohamed Omar & Mohamed Abd wahed

1- Contributions:

*note it will be edited

- Mohamed Elsayad:
  – Dataset selection and acquisition
  – Data quality analysis and preprocessing pipeline
- Mohamad Omar:
  – Visualizations and EDA
  – Probability analysis tasks
- Mohamed Abd Elwahed:
  – Regression modeling and interpretation
  – Report writing and figure polishing

2-Dataset Description:

- Dataset name : zurich_electricity_consumption.csv
- source : https://www.kaggle.com/datasets/sainideeshk/zurich-electricity-consumption-dataset

This dataset is suitable for time-series analysis because it has regular 15-minute sampling, a long continuous time span, and clear temporal patterns, allowing trend, seasonality, resampling, and forecasting to be performed reliably.

- Time period covered and sampling frequency:

Time period covered:

From **January 1, 2015** to **August 31, 2022**.

Sampling frequency:

Data is recorded every **15 minutes** (15-minute intervals).

- Key variables analyzed :

Electricity consumption signals:

Value_NE5 and *Value_NE7* representing measured electrical load levels.

Meteorological variables (external sensors):

Temperature (°C) — ambient air temperature
Relative humidity (%) — atmospheric moisture content
Solar radiation (W/m²) — incoming global radiation
Wind speed (m/s) — horizontal and vertical wind velocity
Wind direction (°) — prevailing wind angle
Air pressure (hPa) — atmospheric pressure
Rain duration (min) — precipitation activity

•Size and structure:

Number of observations (rows):
268,705 time-stamped records.

Number of features (columns):
11 variables (including electricity consumption and weather measurements).

Target variable(s):

- *Value_NE5* (primary electricity consumption signal)

•Missing data summary:
The dataset contains **no missing values** across all variables. All time steps are fully populated, which indicates high data completeness and reliability

3 -Task 1. Data Preprocessing and Basic Analysis:

3.1 Basic statistical analysis

Descriptive statistics and quantiles were computed for key variables. Grouped summaries by hour, weekday, and month reveal clear daily and seasonal patterns.

3.2 Data quality analysis

No original missing values were found. Timestamp order and sampling regularity were verified. Outliers and suspicious values were identified using boxplots and the IQR rule. Physically impossible values were detected using valid ranges.

3.3 Data preprocessing

Invalid values, zero power readings, and outliers were replaced with NaN. Missing values were handled using limited linear interpolation followed by forward and backward filling. Data was resampled to hourly resolution and lag and rolling features were created.

3.4 Before vs after comparison

Visual comparison shows reduced noise and improved consistency while preserving overall temporal patterns. The main trade-off is reduced high-frequency detail due to resampling.

4-Task 2. Visualization and Exploratory Analysis:

4.1 Time-series visualizations

Electricity consumption (Value_NE5 and Value_NE7) was plotted over time. The time-series shows clear long-term trends and strong daily and seasonal variations. No abrupt structural breaks were observed, but recurring periodic patterns are clearly visible.

4.2 Distribution analysis with histograms

Histograms of key numeric variables show non-normal distributions. Electricity consumption exhibits right skewness with heavier upper tails, corresponding to peak demand periods. Weather variables display expected asymmetric and multi-modal behavior due to seasonal effects.

4.3 Correlation analysis and heatmaps

Pearson correlation was used to measure linear relationships between variables. The heatmap shows moderate correlations between electricity consumption and temperature, humidity, and solar radiation. Strong correlations are observed among meteorological variables themselves, confirming physical consistency.

4.4 Daily pattern analysis

Hourly aggregation was applied to analyze daily behavior. Plots reveal strong daily cycles, with lower consumption during nighttime and higher demand during daytime hours. Clear weekday–weekend differences are visible, while short-term fluctuations are treated as noise.

5-Task 3. Probability Analysis:

5.1 Threshold-based probability

A high-load event was defined using the upper quantile of electricity consumption. The probability of exceeding this threshold was estimated empirically and shows that high-demand events are relatively infrequent. Visualizations highlight periods above the threshold.

5.2 Cross tabulation analysis

Electricity consumption was categorized into normal and high levels, and time was grouped into weekday and weekend. The contingency table shows that high consumption occurs more often on weekdays.

5.3 Conditional probability analysis

Event A represents high electricity consumption and Event B represents weekdays. The conditional probability $P(A \mid B)$is higher than $P(A)$, indicating increased demand during weekdays.

5.4 Summary

High-load events are rare, more frequent on weekdays, and strongly influenced by time-related behavior.

6 -Task 4. Statistical Theory Applications:

6.1 Law of Large Numbers (LLN)

Using electricity consumption (Value_NE5), the sample mean converges to the true mean as sample size increases, confirming LLN behavior.

6.2 Central Limit Theorem (CLT)

Repeated sampling shows that the distribution of sample means becomes approximately normal as sample size increases, even though the original data is non-normal.

6.3 Result interpretation

LLN demonstrates mean stability with large samples, while CLT explains why sample means follow a normal distribution for sufficiently large n.

7.-Task 5 – Regression Analysis:

7.1 Model selection

Target and predictors:
The target variable is electricity consumption (Value_NE5). Predictors include weather variables and engineered features such as lagged consumption and rolling averages.

Linear vs polynomial motivation:
A linear regression model was selected as a baseline due to its interpretability and to avoid overfitting, while polynomial models were tested for capturing non-linear effects.

Train–test split rationale:
A time-aware train–test split was used to preserve temporal order and prevent data leakage.

7.2 Model fitting and validation

Fitting and preprocessing:
Features were scaled and relevant predictors were selected before model training.

Validation method:
A time-series holdout split was applied, training on earlier data and testing on later periods.

Evaluation metrics:
RMSE and MAE were used to measure prediction error magnitude, while R^2 was reported to assess explained variance.

Residual analysis:
Residual plots were examined to detect bias, heteroscedasticity, and systematic prediction errors.

7.3 Result interpretation

Main effects:
Electricity consumption is strongly influenced by past consumption (lag features) and temperature-related variables.

Failure cases:
The model performs poorly during extreme peak loads and sudden demand changes, indicating limitations of linear assumptions.