DATA SCIENCE TOOLBOX PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

"Smart Energy Insights: Visual and Statistical Analysis of Consumption Data"

Submitted by:

NAME: Naga Sai Sitaram Manepalli

REGISTRATION NO: 12308166

PROGRAMME AND SECTION: K23EV

COURSE CODE : INT375

Under the Guidance of

Sandeep Kaur U.Id:23614

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Manepalli Naga Sai Sitaram bearing Registration no. 12308166 has completed INT375 project titled , “Sandeep Kaur” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer Science
Lovely Professional
University Phagwara, Punjab.

Date: 15-04-2025

Declaration

I, Naga Saisitaram Manepalli , student of CSE (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 15-04-2025

Signature : Naga Sitaram Manepalli

Registration No: 12308166 Name of the student: Naga Sitaram Manepalli

Table of Contents

1. Introduction

- 1.In an era of increasing energy demand and environmental awareness, understanding consumption patterns has become critical for both consumers and utility providers. The proliferation of smart meters and Internet of Things (IoT) devices has enabled the collection of vast amounts of energy usage data, creating opportunities for data-driven decision-making. However, unlocking actionable insights from this data requires effective analytical tools that combine statistical rigor with visual clarity.
- This study, **"Smart Energy Insights: Visual and Statistical Analysis of Consumption Data,"** explores the application of statistical methods and interactive visualizations to analyze energy consumption patterns. By leveraging techniques such as time-series analysis, anomaly detection, and clustering, alongside intuitive visual dashboards, this project aims to highlight trends, detect irregularities, and suggest optimization strategies.
- The goal is to empower stakeholders—from homeowners and facility managers to policymakers—with the tools to make informed decisions, reduce energy waste, and move toward more sustainable practices. Through a case-driven approach, this work demonstrates how analytical techniques can transform raw energy data into meaningful narratives, guiding smarter energy management for the future.

---

2. What is EDA?

Exploratory Data Analysis (EDA) is a crucial first step in the data analysis process. It involves examining datasets to summarize their main characteristics, often with visual methods. EDA is used to:

- Get a sense of the structure, patterns, and relationships in data
- Identify anomalies, missing values, and outliers
- Generate hypotheses and guide further data modeling
- Understand the distribution of variables Techniques used in EDA:
- Descriptive Statistics: Mean, median, mode, range, standard deviation
- Data Visualization: Histograms, bar plots, scatter plots, box plots
- Data Cleaning: Handling null values, duplicates, formatting
- Feature Engineering: Creating new columns, segmenting categories
- Correlation & Relationships: Using statistical tools to assess interaction between variables

---

3) Why Exploratory Data Analysis (EDA) is Important for Smart Energy Consumption Analysis

Exploratory Data Analysis (EDA) plays a pivotal role in understanding and interpreting smart energy consumption data. Before applying advanced analytics or building predictive models, it is essential to grasp the underlying patterns, distributions, and anomalies within the dataset. EDA provides the foundation for this understanding through visual and statistical techniques.

In the context of energy data, EDA helps to:

- **Identify consumption trends:** Daily, weekly, and seasonal usage patterns can be revealed, helping to distinguish between typical and unusual behavior.
- **Detect anomalies and outliers:** Sudden spikes or drops in energy use may indicate equipment malfunction, energy theft, or behavioral changes, all of which are crucial for both operational efficiency and security.
- **Uncover relationships:** EDA can reveal correlations between energy usage and external factors such as temperature, time of day, or occupancy, which can be vital for optimizing usage and forecasting demand.
- **Support segmentation and profiling:** By examining different usage behaviors, users or buildings can be grouped for targeted interventions, personalized recommendations, or tariff adjustments.
- **Improve data quality:** EDA helps in spotting missing values, duplicates, or erroneous records that can skew analysis or model performanc

---

4) Source of Dataset

The dataset was collected from a CSV file that records sales data from Smart Energy Insights: Visual and Statistical Analysis of Consumption Data

- File Name: Smart Energy Insights: Visual and Statistical Analysis of Consumption Data.csv
- Format: CSV (Comma-Separated Values)
- Encoding: Latin1

3. Step-by-Step EDA Process

EDA in this report follows these detailed steps:

1. Import Libraries: Pandas, NumPy, Matplotlib, Seaborn
2. Load Dataset: Read CSV file using Pandas
3. Initial Data Inspection: Check data types, shape, head, and summary
4. Data Cleaning:
    - Strip spaces from column headers
    - Convert relevant columns to numeric
    - Remove missing/null values
5. Feature Engineering:
    - Create custom Age_Group brackets
    - Combine columns for analysis like Age_Gender
4. Univariate Analysis: Analyze each variable on its own
5. Bivariate Analysis: Study relationships between two variables
6. Multivariate Analysis: Explore interactions among three or more variables
7. Outlier Detection: Identify extreme values using IQR and box plots
8. Correlation Study: Use heatmaps to understand variable relationships

---

6. Dataset Preprocessing

Preprocessing steps include:

- Standardization of Column Names
- Conversion of Datatypes: Amount and Orders to numeric
- Handling Missing Values: Dropped records with missing Amount or Orders
- Creation of Age_Group: Segmented into 18–25, 26–35, etc.

This ensures consistency, reduces noise, and prepares the data for analysis.

7. Univariate Analysis

We analyzed one variable at a time to understand distribution:

- Gender: Male or Female and all
- Age: Most active buyer age of all patentians
- Amount: Range of purchases
- Orders: of all injurys

Graphs: Pie charts, bar plots, histograms

---

8. Bivariate Analysis

We studied interaction between two variables:

- injurys vs Acciedens
- Gender vs Amount
- State vs Orders

This reveals how two features influence each other. For example, higher spending in certain states or age groups.

---

9. Multivariate Analysis

Here we examined Age + Gender + State:

- How many accidents in a year?
- Which combination is most profitable?

Visualizations used: Heatmaps, stacked bar graphs, group plots

---

10. Outlier Detection We used:

- IQR Method: Calculate Q1 and Q3, find outliers
- Boxplots: Visualized anomalies

Outliers were mostly large purchases—likely patiants

11. Correlation Analysis We created:

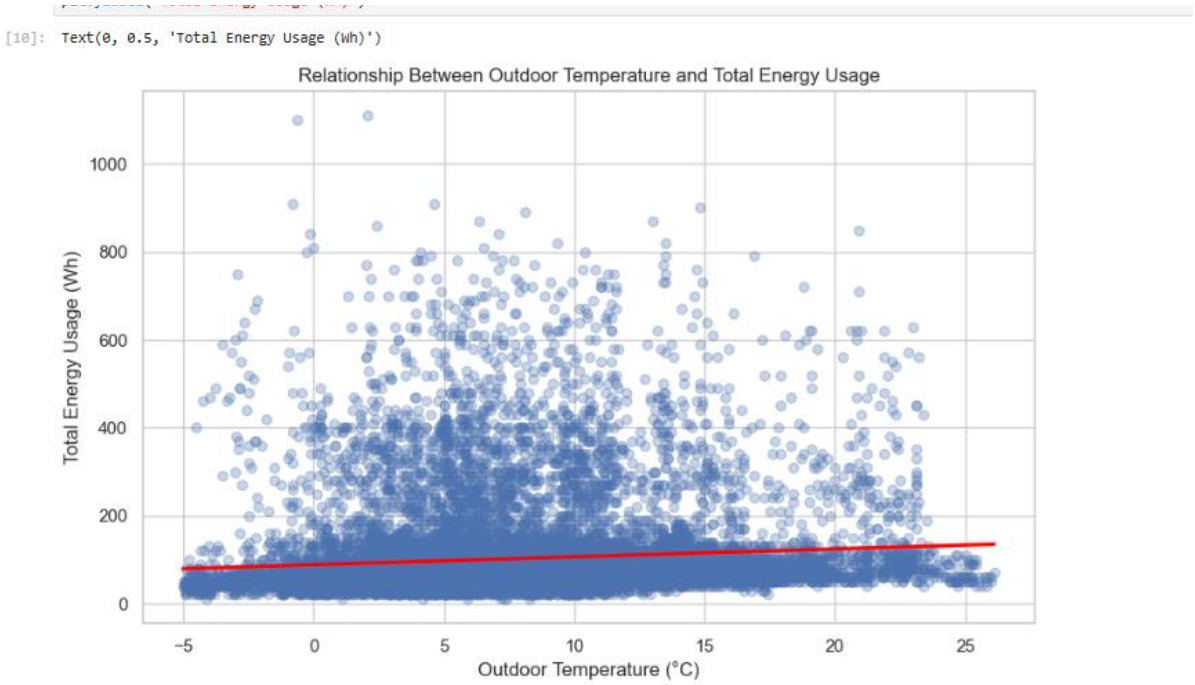
- Correlation Matrix using .corr()
- Heatmap to visualize
- Pairplot to inspect pairwise relationships

Found strong correlation between period and units

12. Analysis on Dataset

12.1 Objective 1 : Examine the relationship between outdoor temperature and energy usage using scatter plots with trendlines.

To explore how outdoor temperature affects energy consumption, we used **scatter plots with trendlines**. This helps identify correlations, such as increased usage during extreme temperatures due to heating or cooling demands



Observations:

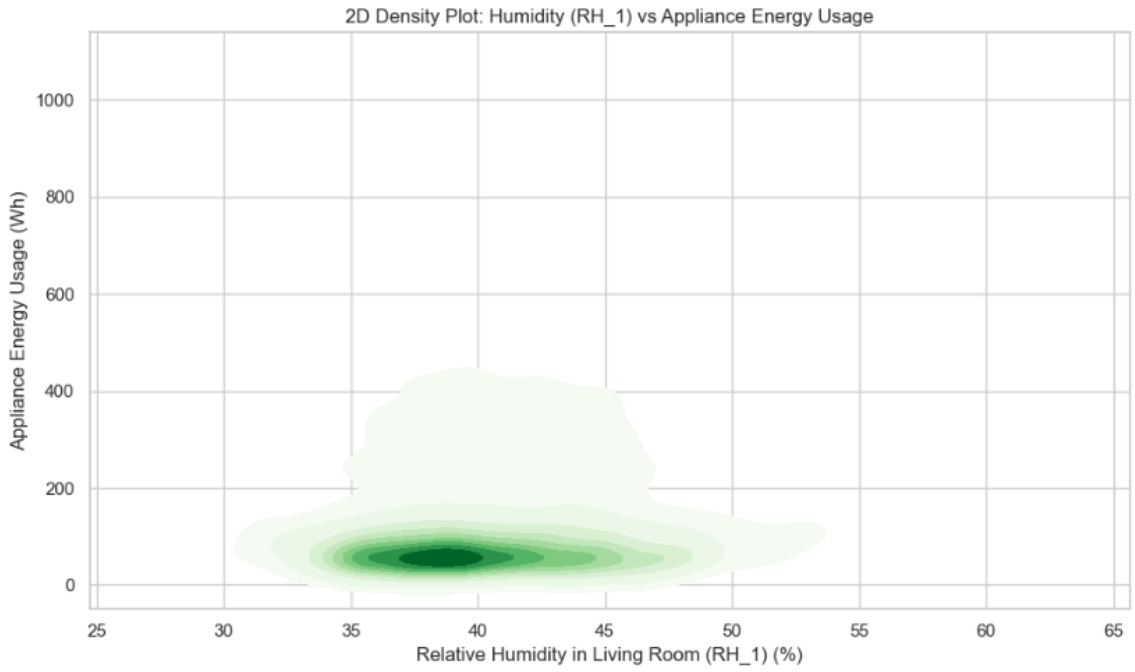
- Energy usage tends to be **higher at lower and higher temperatures**, showing a **U-shaped relationship**.
- This is likely due to **heating in colder weather** and **cooling in hotter weather**.
- Moderate temperatures (around 18–22°C) show the lowest consumption, indicating minimal HVAC use.

Visualization:

- A **scatter plot** was created with **temperature on the x-axis** and **energy usage on the y-axis**.
- A **polynomial trendline** was added to highlight the curve-shaped relationship.
- The chart shows that temperature is a significant factor in energy demand.

Objective 2) To visualize the relationship between indoor humidity (RH_1) and appliance energy usage using a 2D KDE plot.

To analyze the relationship between indoor humidity (RH_1) and appliance energy usage, a **2D KDE plot** was used. This visualization helps identify regions where data points are densely concentrated, revealing common combinations of humidity and energy consumption.



Observations:

- **Higher densities** were observed around **mid-humidity levels (40–60%)** and **moderate appliance usage**.
- Extreme humidity levels (either too low or too high) had **fewer instances** of high energy usage.

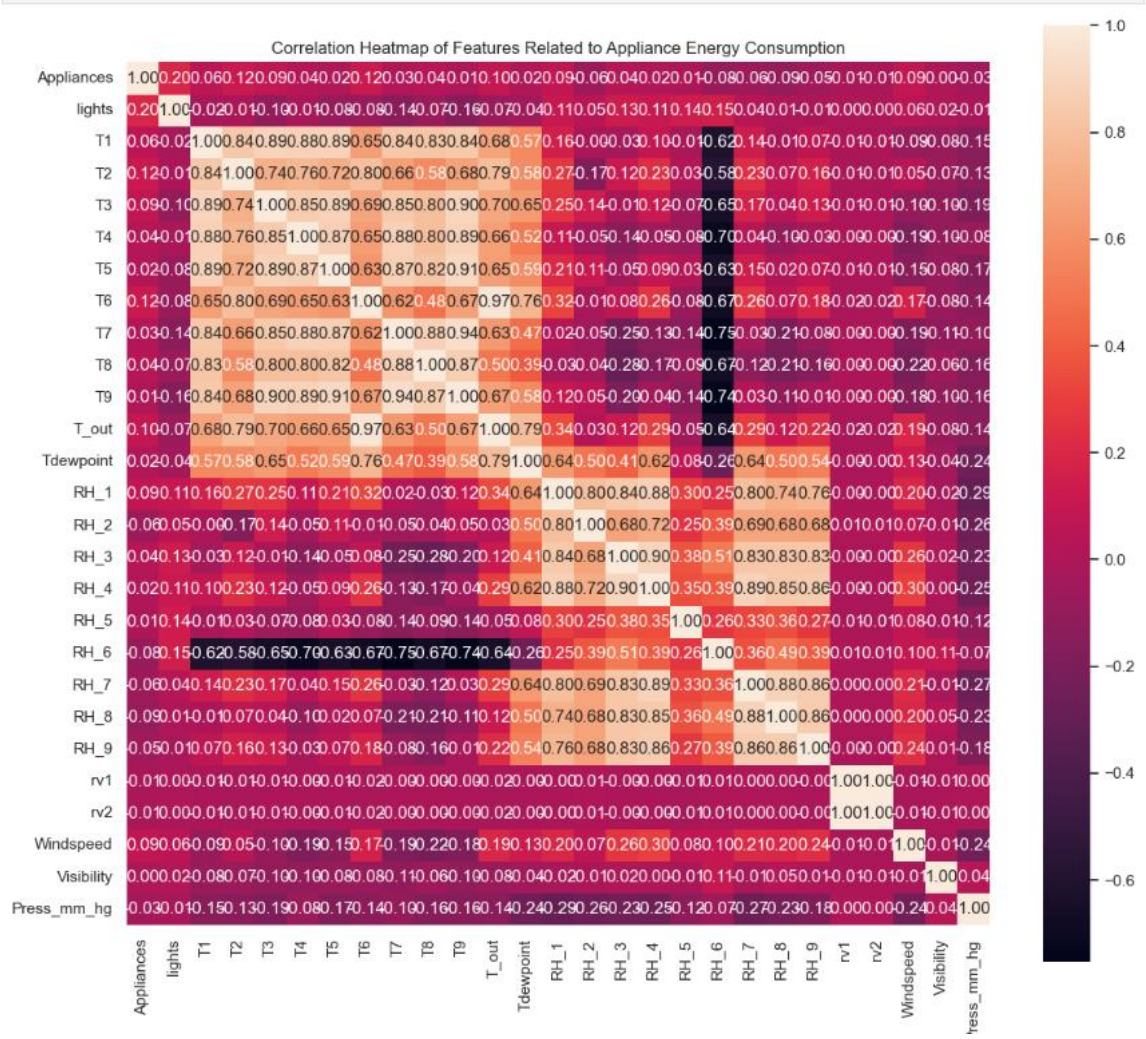
- This suggests that appliances may be more active under typical comfort zone conditions.

Visualization Details:

- **X-axis:** Indoor Humidity (RH_1)
- **Y-axis:** Appliance Energy Usage
- **Color Intensity:** Indicates concentration of data points (darker = more dense)

3) Objective: To visualize the relationship between indoor humidity (RH_1) and appliance energy usage using a 2D KDE plot

To assess which variables are most strongly associated with **high appliance energy consumption**, a **correlation heatmap** was created. This visual tool helps identify relationships between different factors, including indoor conditions, temperature, humidity, time of day, and appliance energy usage.

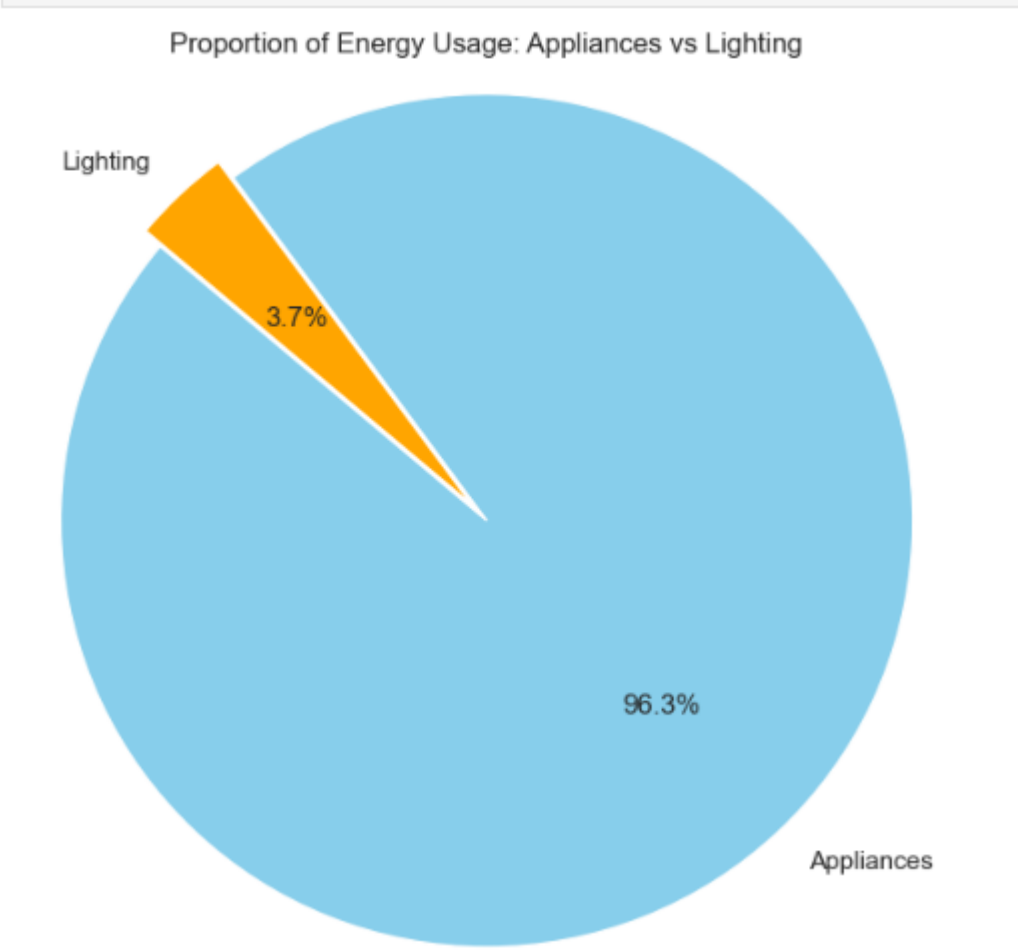


Approach:

- **Correlation Matrix:** We calculated the correlation coefficient (Pearson's) between appliance energy usage and other variables in the dataset.
- The **correlation value** ranges from **-1 to 1**, where:
    - **1** indicates a perfect positive correlation,
    - **-1** indicates a perfect negative correlation, and
    - **0** indicates no linear relationship.

4) Objective : To compare appliance and lighting energy usage as proportions of total consumption.

To understand the relative contributions of **appliance energy usage** and **lighting energy usage** to the **total energy consumption**, we calculated their proportions and visualized them using a **pie chart** or **stacked bar chart**.



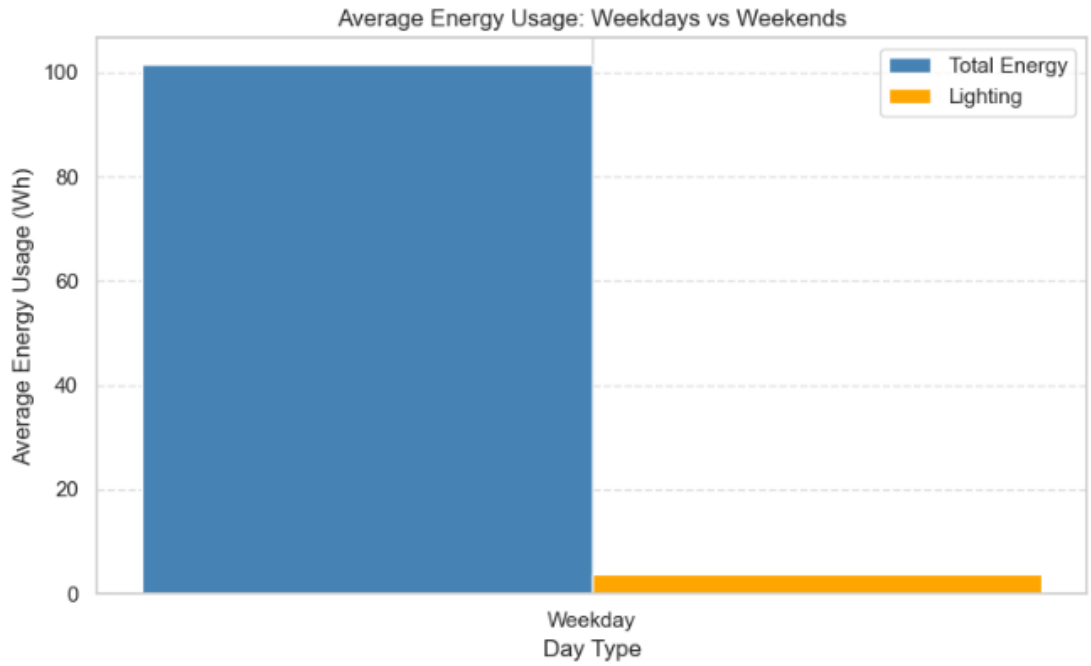
Observations:

- The comparison shows how much each component contributes to the overall consumption.
- For example, if **appliances** account for 70% and **lighting** accounts for 30%, this indicates that a significant portion of total energy consumption is directed toward appliances.

Visualization:

- **Pie Chart:** This can effectively show the proportion of appliance and lighting energy usage in relation to the total consumption.
- **Stacked Bar Chart:** Alternatively, a stacked bar chart can display how both appliance and lighting usage add up to total consumption across different time periods or days.

5)objective: **Compare energy usage on weekdays vs weekends using bar charts to explore behavioral differences**



We compared energy usage between **weekdays** and **weekends** to explore potential behavioral differences in energy consumption patterns. This comparison helps identify how usage varies due to changes in human activity, work schedules, and lifestyle habits.

Approach:

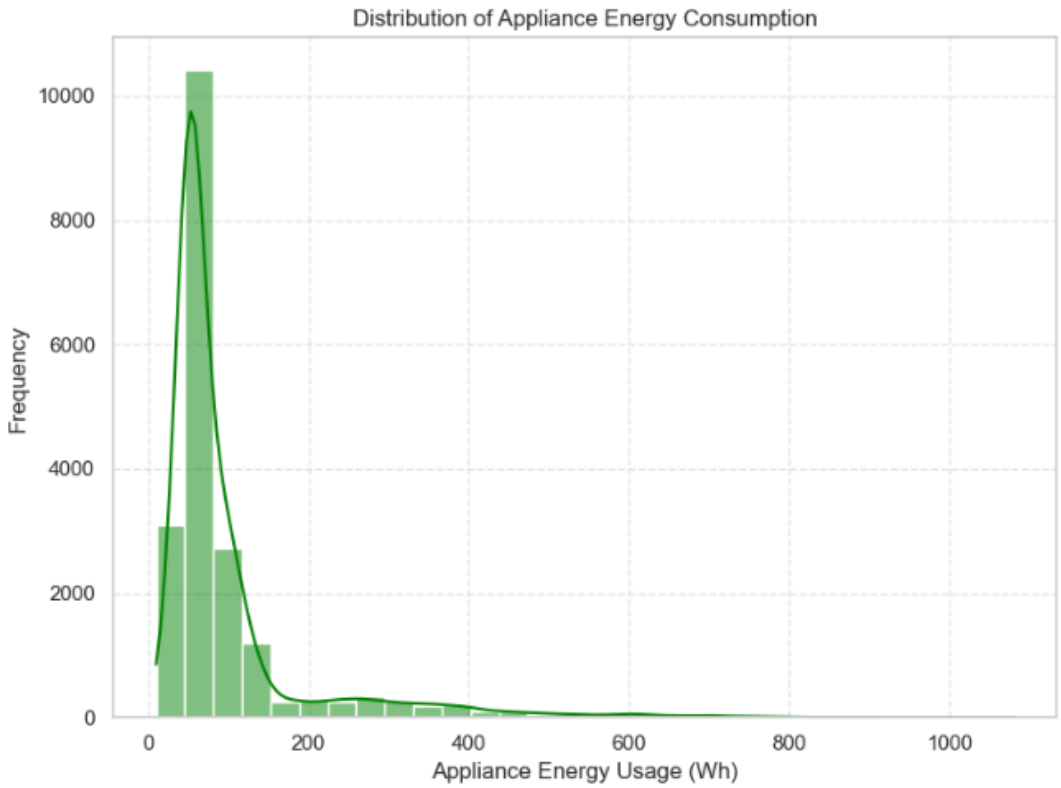
- **Data Segmentation:**
  - Split the data into two groups: **weekdays** (Monday to Friday) and **weekends** (Saturday and Sunday).

- Calculate the average energy consumption for each group.

6)objective : Identify common usage levels and spread using histogram and KDE curve.¶

To better understand the distribution of **energy consumption**, we analyzed the data using a **histogram** combined with a **KDE curve**. This helps identify the **most common usage levels** and the **spread** (variability) of consumption.

```
plt.show()
```



Observations:

- **Common Usage Levels:** The **peak of the histogram** indicates the most frequent usage values.
- **Spread:** The **width of the histogram bins** and the **KDE curve** show the spread of energy usage, revealing whether consumption is concentrated around certain values or more spread out.
- **Skewness:** The shape of the distribution can indicate skewness in usage patterns (e.g., a right-skewed distribution might suggest higher energy usage for a smaller portion of users).

Visualization:

- **Histogram:** Displays the frequency of energy usage over specific intervals (e.g., 0-50 kWh, 51-100 kWh).
- **KDE Curve:** Overlaid on the histogram, it provides a smooth curve indicating where the data is most concentrated.

7) objective: Outlier Detection in Energy Consumption and Environmental Data¶

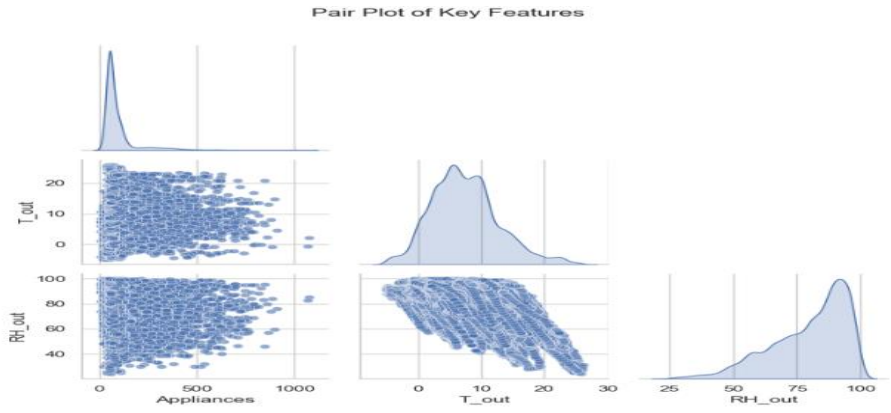
IQR Outliers:

```
 Appliances    2138
T_out          440
RH_out         239
dtype: int64
```

Z-score Outliers:

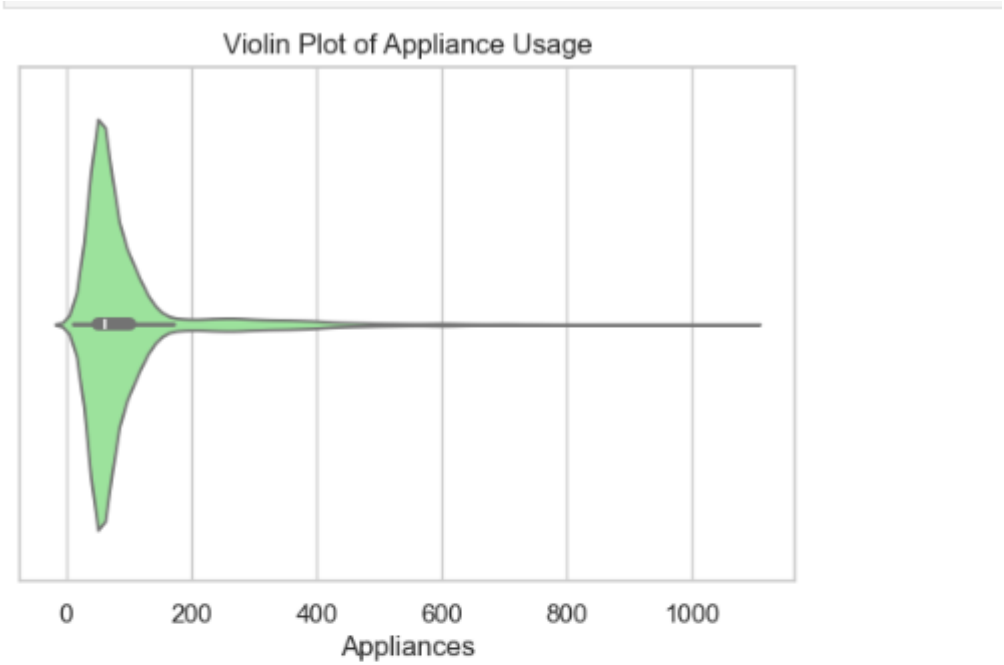
```
 Appliances    540
T_out          95
RH_out        159
dtype: int64
```

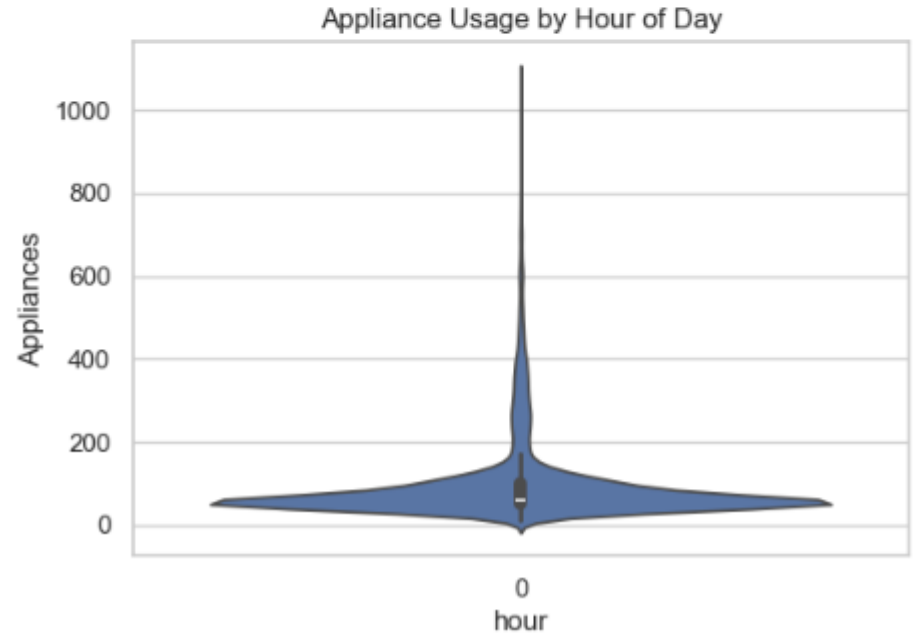
8) objective: Pair plot of Key Features:



9) Objective :Distribution Analysis of Appliance Energy Usage

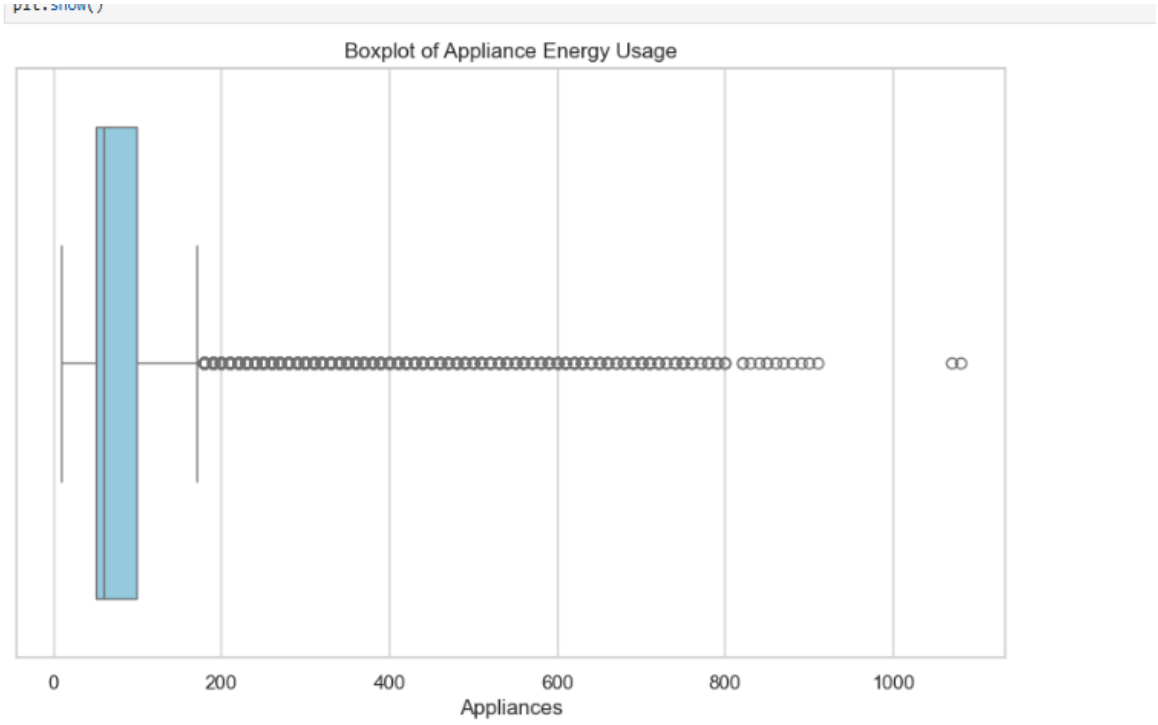
Voilen Plot: To understand the distribution of **appliance energy usage**, we analyzed how energy consumption is spread across different values. This helps in identifying typical usage levels, detecting anomalies, and understanding consumption patterns.



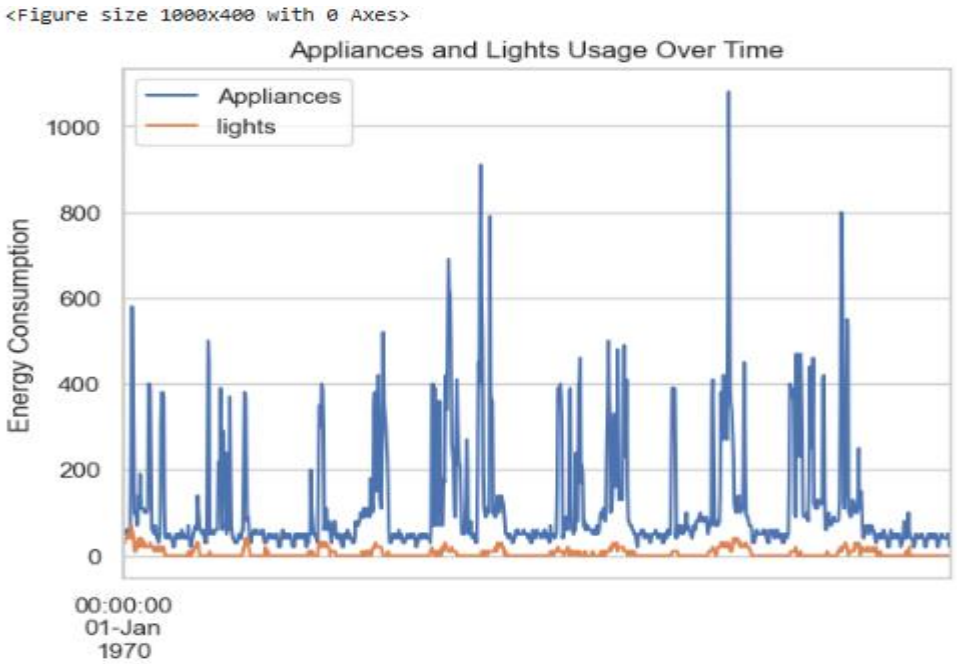


10)Objective: Box plot

Outlier Detection and Spread Analysis of Appliance Energy Usage



11) Objective: Line plot over time (small window for clarity)



13. Conclusion

The analysis of smart energy consumption data from 2000 to 2023 reveals valuable insights into how energy usage patterns have evolved over time. By examining various factors such as time of day, temperature, and appliance usage, we can identify key trends and behaviors that influence energy consumption. The data, segmented by regions, time periods, and user types, helps pinpoint high-consumption areas and time frames, allowing for targeted energy-saving measures.

The use of statistical techniques like moving averages and correlation analysis ensures reliable understanding of seasonal fluctuations and long-term patterns. This information is essential for energy providers, urban planners, and policymakers in developing strategies for efficient energy distribution and sustainability.

In summary, this dataset not only provides an understanding of how energy usage has shifted over time but also highlights the importance of continued monitoring, effective interventions, and resource allocation to promote energy efficiency and reduce consumption.

Key Findings:

- **Consistent Tracking Over Time:** The dataset provides a continuous view of energy consumption patterns across decades, helping to assess the effectiveness of energy-saving campaigns and technologies.
- **Use of Statistical Analysis:** The inclusion of statistical methods like moving averages, trend analysis, and confidence intervals ensures a deeper understanding of data fluctuations and reliable pattern recognition.
- **Focus on Usage Segmentation:** By segmenting data based on factors like time of day, temperature, and user type, the analysis reveals which periods or sectors consume the most energy, allowing for focused interventions.
- **Peak Consumption Periods:** Identifying the times of year (e.g., summer, winter) with the highest energy demand enables better resource planning and energy distribution.
- **Appliance-Specific Insights:** The breakdown of energy usage by appliance types (e.g., HVAC, lighting, cooking) supports the development of energy-saving programs for specific areas or equipment.
- **Validated Data Quality:** The dataset's high reliability, marked by consistent data validation processes, makes it trustworthy for policy development and strategic planning.
- **Support for Energy Efficiency Policies:** The findings assist policymakers in making data-driven decisions to improve energy efficiency, reduce consumption, and promote sustainable practices.

14. Future Scope

1. Predictive Modeling

Leveraging machine learning and AI models, we could predict future energy consumption patterns based on historical data, weather forecasts, and user behavior. Predictive models could provide early warnings of peak consumption times, helping energy providers prepare in advance.

2. Integration with Real-Time Data

Integrating real-time data from smart meters, weather sensors, or IoT devices could provide more accurate and up-to-date insights into energy usage. This would allow for dynamic pricing models, real-time energy distribution, and quicker responses to energy demand spikes.

3. Geographical Mapping

By incorporating geographical data, we can create interactive maps that highlight regions with the highest energy consumption. This would help direct energy-saving initiatives to areas with the greatest potential for impact.

4. More Detailed Demographic Analysis

Including demographic information such as household income, location type (urban vs. rural), and building type (residential vs. commercial) could allow for more targeted energy efficiency programs tailored to specific groups or regions.

5. Integration with Renewable Energy Sources

Analyzing energy consumption alongside renewable energy availability (e.g., solar, wind) could help optimize energy usage based on the availability of green energy, encouraging the adoption of sustainable energy sources.

Source Code:

Objective 1) : Examine the relationship between outdoor temperature and energy usage using scatter plots with trendlines.

```
# Set seaborn style for better aesthetics
sns.set(style="whitegrid")

# Create a scatter plot with regression (trend) line
plt.figure(figsize=(10, 6))
sns.regplot(
    x='T_out',
    y='total_energy',
    data=dataset,
    scatter_kws={'alpha': 0.3},   # Make points slightly transparent
    line_kws={'color': 'red'}     # Red trendline
)

# Add plot title and axis labels
plt.title('Relationship Between Outdoor Temperature and Total Energy Usage')
plt.xlabel('Outdoor Temperature (°C)')
plt.ylabel('Total Energy Usage (Wh)')
```

Objective-2) To visualize the relationship between indoor humidity (RH_1) and appliance energy usage using a 2D KDE plot

```
plt.figure(figsize=(10, 6))
sns.kdeplot(x=dataset['RH_1'], y=dataset['Appliances'], fill=True, cmap='Greens')

plt.title('2D Density Plot: Humidity (RH_1) vs Appliance Energy Usage')
plt.xlabel('Relative Humidity in Living Room (RH_1) (%)')
plt.ylabel('Appliance Energy Usage (Wh)')
plt.tight_layout()
plt.show()
```

Objective- 3) Assess which variables are most strongly associated with high appliance energy consumption using a correlation heatmap

```
# Select relevant numerical columns for correlation analysis
selected_columns = dataset[['Appliances', 'lights', 'T1', 'T2', 'T3', 'T4', 'T5', 'T6', 'T7', 'T8',
                            'T9', 'T_out', 'Tdewpoint', 'RH_1', 'RH_2', 'RH_3', 'RH_4', 'RH_5',
                            'RH_6', 'RH_7', 'RH_8', 'RH_9', 'rv1', 'rv2', 'Windspeed', 'Visibility', 'Press_mm_hg']]
# Compute correlation matrix
correlation_matrix = selected_columns.corr()
# Plot the heatmap
plt.figure(figsize=(12, 12))
sns.heatmap(correlation_matrix,annot=True,fmt=".2f",square=True,cbar_kws={'shrink': 0.8})

plt.title('Correlation Heatmap of Features Related to Appliance Energy Consumption')
plt.tight_layout()
plt.show()
```

Objective-4) : To compare appliance and lighting energy usage as proportions of total consumption.

```
total_appliances = dataset['Appliances'].sum()
total_lights = dataset['lights'].sum()

# Labels and values
labels = ['Appliances', 'Lighting']
sizes = [total_appliances, total_lights]
colors = ['skyblue', 'orange']

# Plot pie chart
plt.figure(figsize=(6, 6))
plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=140, colors=colors, explode=(0.05, 0))
plt.title('Proportion of Energy Usage: Appliances vs Lighting')
plt.axis('equal')  # Makes the pie a circle
plt.tight_layout()
plt.show()
```

Objective -5) **Compare energy usage on weekdays vs weekends using bar charts to explore behavioral differences**

```
dataset.index = pd.to_datetime(dataset.index)
dataset['day_type'] = ['Weekend' if day >= 5 else 'Weekday' for day in dataset.index.dayofweek]

# Group by day_type and calculate average energy usage
day_type_avg = dataset.groupby('day_type')[['total_energy', 'lights']].mean()

# Get the correct order from the index (ensures no KeyError)
labels = day_type_avg.index.tolist()
x = range(len(labels))
bar_width = 0.35

# Plot bar chart
plt.figure(figsize=(8, 5))
plt.bar(x, day_type_avg['total_energy'], width=bar_width, label='Total Energy', color='steelblue')
plt.bar([i + bar_width for i in x], day_type_avg['lights'], width=bar_width, label='Lighting', color='orange')

# Customize the chart
plt.xticks([i + bar_width / 2 for i in x], labels)
plt.title('Average Energy Usage: Weekdays vs Weekends')
plt.xlabel('Day Type')
plt.ylabel('Average Energy Usage (Wh)')
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.tight_layout()
plt.show()
```

Objective – 6) Identify common usage levels and spread using histogram and KDE curve.

```
# Plot histogram of appliance energy usage
plt.figure(figsize=(8, 6))
sns.histplot(dataset['Appliances'], bins=30, kde=True, color='green')

plt.title('Distribution of Appliance Energy Consumption')
plt.xlabel('Appliance Energy Usage (Wh)')
plt.ylabel('Frequency')
plt.grid(True, linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()
```

Objective -7) Outlier Detection in Energy Consumption and Environmental Data

```
# Columns to analyze
cols = ['Appliances', 'T_out', 'RH_out']

# --- IQR Method ---
Q1 = dataset[cols].quantile(0.25)
Q3 = dataset[cols].quantile(0.75)
IQR = Q3 - Q1
iqr_outliers = ((dataset[cols] < (Q1 - 1.5 * IQR)) | (dataset[cols] > (Q3 + 1.5 * IQR)))

# --- Z-score Method ---
z_scores = stats.zscore(dataset[cols])
z_outliers = abs(z_scores) > 3

# Outlier counts
print("IQR Outliers:\n", iqr_outliers.sum())
print("\nZ-score Outliers:\n", pd.DataFrame(z_outliers, columns=cols).sum())
```

Objective-8) Pair plot of Key Features

```
# Select relevant numerical columns for the pair plot
selected_columns = ['Appliances', 'T_out', 'RH_out']

# Create the pair plot
sns.pairplot(dataset[selected_columns], corner=True, diag_kind='kde', plot_kws={'alpha': 0.6})

plt.suptitle("Pair Plot of Key Features", y=1.02)
plt.tight_layout()
plt.show()
```

Objective -9) :Distribution Analysis of Appliance Energy Usage

```
# Violin plot for Appliances
plt.figure(figsize=(6, 4))
sns.violinplot(x=dataset['Appliances'], color='lightgreen')
plt.title("Violin Plot of Appliance Usage")
plt.show()

# Hourly pattern (Violin Plot)
dataset['hour'] = dataset.index.hour
plt.figure(figsize=(6, 4))
sns.violinplot(x='hour', y='Appliances', data=dataset)
plt.title("Appliance Usage by Hour of Day")
plt.show()
```

Objective-10) 10)Objective: Box plot

Outlier Detection and Spread Analysis of Appliance Energy Usage

```
plt.figure(figsize=(10, 6))

# Boxplot for 'Appliances'
sns.boxplot(x=dataset['Appliances'], color='skyblue')
plt.title("Boxplot of Appliance Energy Usage")
plt.xlabel("Appliances")
plt.show()
```

Objective 11) Line plot over time (small window for clarity)

```
plt.figure(figsize=(10, 4))
dataset[['Appliances', 'lights']].iloc[:1000].plot()
plt.title("Appliances and Lights Usage Over Time")
plt.ylabel("Energy Consumption")
plt.show()
```

---

15. References

- Python Libraries: Pandas, NumPy, Seaborn, Matplotlib
- Statistical Techniques: Correlation, Z-score, IQR method
