Progress Report I

1. Introduction

Energy consumption in residential buildings is influenced by various environmental factors such as temperature, humidity, and weather conditions. Understanding these relationships is crucial for improving energy efficiency and sustainability. This study explores the Appliance Energy Prediction dataset, which was collected to develop regression models for predicting energy consumption in a low-energy residential building. The dataset was collected using a ZigBee wireless sensor network to monitor temperature and humidity conditions. My motivation for selecting this dataset is that I have not extensively worked on regression problems. This project presents an opportunity to strengthen my skills in regression modeling.

2. Description of Data

The dataset is designed for experimental regression models analyzing energy consumption of appliances in a low-energy building. It also includes measurements from temperature and humidity sensors within a wireless network, weather data from a nearby airport station, and recorded energy usage of lighting fixtures. The data originates from the UC-Irvine Machine Learning Repository and consists of 19,735 observations with 29 features, with the target variable being appliance energy consumption.

3. Explanation of Features

The table below provides a concise overview of the features and columns in the dataset. The descriptions of these variables are primarily sourced from the attribute information section of the original dataset.

Variable Name	Role	Type	Description	Missing Values
Date	Feature	Date		No
Appliance	Target	Integer	Energy consumption (Wh) of appliances.	No
Lights	Feature	Integer	Energy consumption of lights (in Wh)	No
T1	Feature	Continuous	Temperature in living room	No
RH_1	Feature	Continuous	Relative Humidity (%) in living room	No
T2	Feature	Continuous	Temperature in kitchen	No
RH_2	Feature	Continuous	Relative Humidity (%) in kitchen	No

T3	Feature	Continuous	Temperature in laundry room	No
RH_3	Feature	Continuous	Relative Humidity (%) in laundry room	No
T4	Feature	Continuous	Temperature in office room	No
RH_4	Feature	Continuous	Relative Humidity (%) in office room	No
T5	Feature	Continuous	Temperature in bathroom	No
RH_5	Feature	Continuous	Relative Humidity (%) in bathroom	No
T6	Feature	Continuous	Temperature (°C) in north-facing rooms	No
RH_6	Feature	Continuous	Relative Humidity (%) in north-facing rooms	No
T7	Feature	Continuous	Temperature (°C) in ironing room	No
RH_7	Feature	Continuous	Relative Humidity (%) in ironing room	No
T8	Feature	Continuous	Temperature (°C) in teen room 1	No
RH_8	Feature	Continuous	Relative Humidity (%) in teen room 1	No
T9	Feature	Continuous	Temperature (°C) in parents' room	No
RH_9	Feature	Continuous	Relative Humidity (%) in parent's room	No
T_out	Feature	Continuous	Outdoor temperature (°C).	No
Press_mg	Feature	Continus	Atmospheric pressure (mm Hg)	No
RH_out	Feature	Continuous	Outdoor relative humidity (%)	No
Windspeed	Feature	Continuous	Wind speed (m/s)	No
Visibility	Feature	Continuous	Visibility (km)	No
Tdewpoint	Feature	Continuous	Dew point temperature (°C).	No
Rv1 & rv2	Feature	Continuous	Random variables with no specific meaning (used for testing)	No
Appliance	Feature	Continuous	Energy consumption (Wh) of appliances (target variable)	No

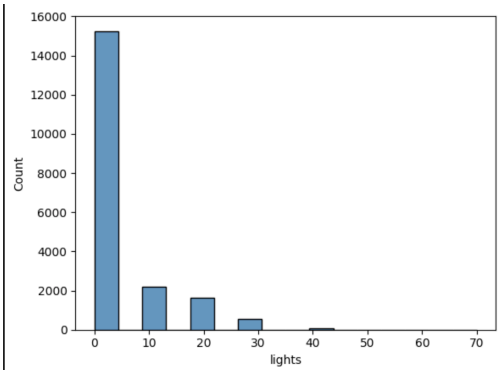
4. Data Preprocessing Steps

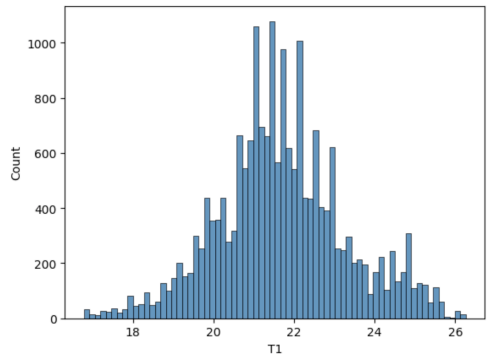
1. Load Dataset
2. Sanity Check of Data
3. **Exploratory Data Analysis (EDA)**

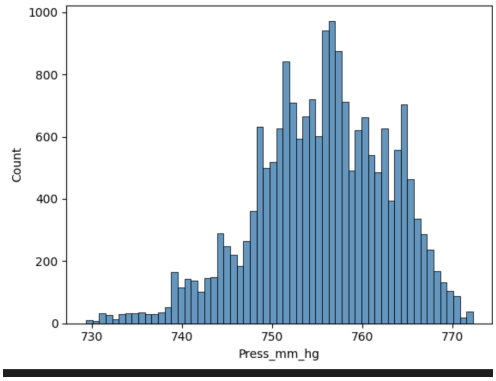
Describe function: The table presents a statistical summary of the dataset, including the count, mean, standard deviation (std), minimum (min), maximum (max), and quartiles (25%, 50% (median), 75%) for each feature.

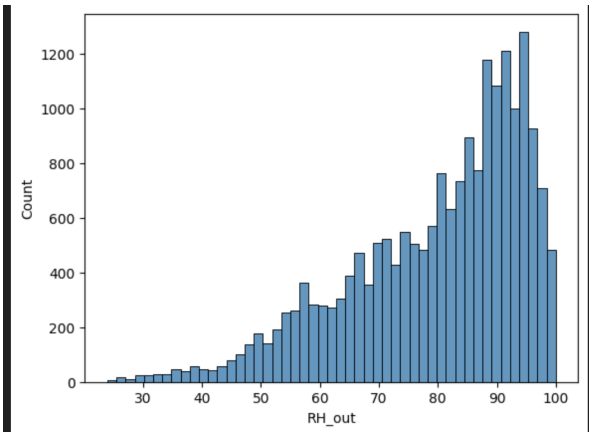
	count	mean	std	min	25%	50%	75%	max
lights	19735.0	3.801875	7.935988	0.000000	0.000000	0.000000	0.000000	70.000000
T1	19735.0	21.686571	1.606066	16.790000	20.760000	21.600000	22.600000	26.260000
RH_1	19735.0	40.259739	3.979299	27.023333	37.333333	39.656667	43.066667	63.360000
T2	19735.0	20.341219	2.192974	16.100000	18.790000	20.000000	21.500000	29.856667
RH_2	19735.0	40.420420	4.069813	20.463333	37.900000	40.500000	43.260000	56.026667
T3	19735.0	22.267611	2.006111	17.200000	20.790000	22.100000	23.290000	29.236000
RH_3	19735.0	39.242500	3.254576	28.766667	36.900000	38.530000	41.760000	50.163333
T4	19735.0	20.855335	2.042884	15.100000	19.530000	20.666667	22.100000	26.200000
RH_4	19735.0	39.026904	4.341321	27.660000	35.530000	38.400000	42.156667	51.090000
T5	19735.0	19.592106	1.844623	15.330000	18.277500	19.390000	20.619643	25.795000
RH_5	19735.0	50.949283	9.022034	29.815000	45.400000	49.090000	53.663333	96.321667
T6	19735.0	7.910939	6.090347	-6.065000	3.626667	7.300000	11.256000	28.290000
RH_6	19735.0	54.609083	31.149806	1.000000	30.025000	55.290000	83.226667	99.900000
T7	19735.0	20.267106	2.109993	15.390000	18.700000	20.033333	21.600000	26.000000
RH_7	19735.0	35.388200	5.114208	23.200000	31.500000	34.863333	39.000000	51.400000
T8	19735.0	22.029107	1.956162	16.306667	20.790000	22.100000	23.390000	27.230000
RH_8	19735.0	42.936165	5.224361	29.600000	39.066667	42.375000	46.536000	58.780000
T9	19735.0	19.485828	2.014712	14.890000	18.000000	19.390000	20.600000	24.500000
RH_9	19735.0	41.552401	4.151497	29.166667	38.500000	40.900000	44.338095	53.326667
T_out	19735.0	7.412580	5.318464	-5.000000	3.670000	6.920000	10.400000	26.100000
Press_mm_hg	19735.0	755.522602	7.399441	729.300000	750.933333	756.100000	760.933333	772.300000
RH_out	19735.0	79.750418	14.901088	24.000000	70.333333	83.666667	91.666667	100.000000
Windspeed	19735.0	4.039752	2.451221	0.000000	2.000000	3.666667	5.500000	14.000000
Visibility	19735.0	38.330834	11.794719	1.000000	29.000000	40.000000	40.000000	66.000000
Tdewpoint	19735.0	3.760995	4.195248	-6.600000	0.900000	3.430000	6.570000	15.500000
rv1	19735.0	24.988033	14.496634	0.005322	12.497889	24.897653	37.583769	49.996530
rv2	19735.0	24.988033	14.496634	0.005322	12.497889	24.897653	37.583769	49.996530
Appliances	19735.0	97.694958	102.524891	10.000000	50.000000	60.000000	100.000000	1080.000000

Histogram Analysis: To analyze the distribution of numerical features in the dataset, I used histograms. This visualization technique helped me identify the spread, central tendency, skewness, and presence of outliers in the data. I provided examples of histogram insights, such as different distribution shapes normal, skewed, or uniform. Additionally, I examined the distribution of each column individually to understand its behavior.



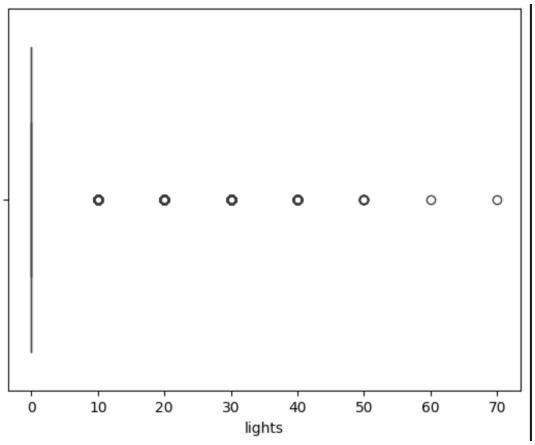


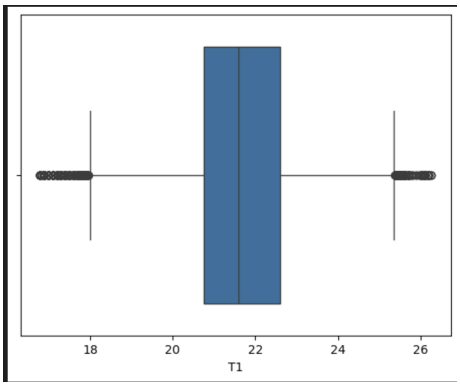


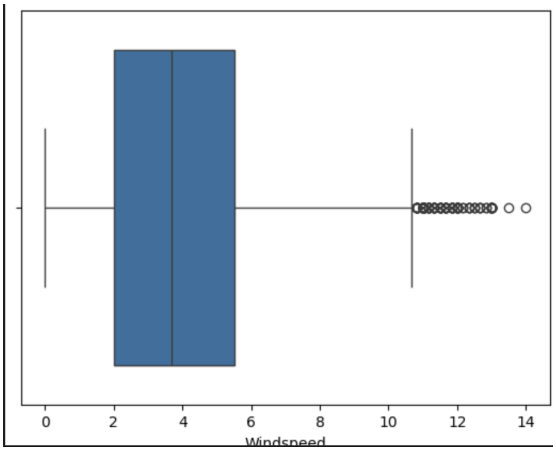


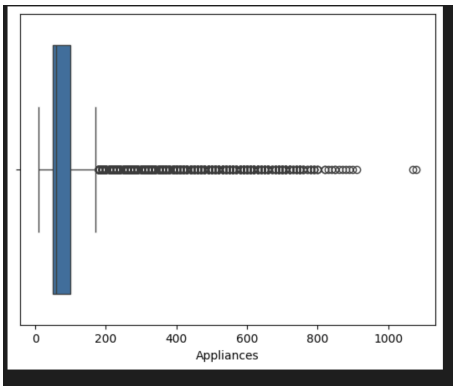
Boxplots: I used boxplots to detect potential errors in the dataset. Upon analysis, I observed that some columns contained extreme outliers. A

significant portion of the dataset exhibited outliers, and I provided examples to illustrate these findings.



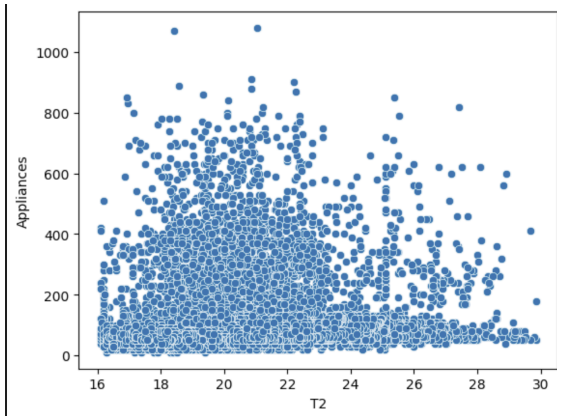


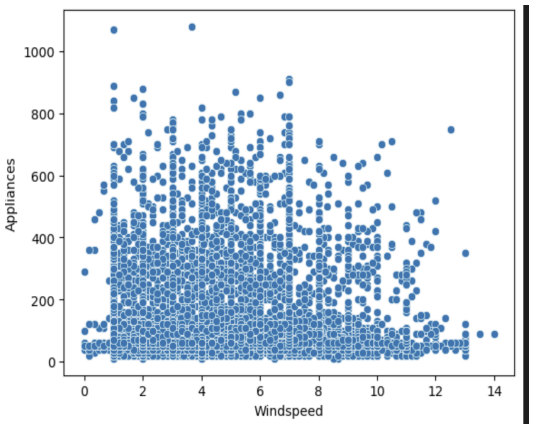


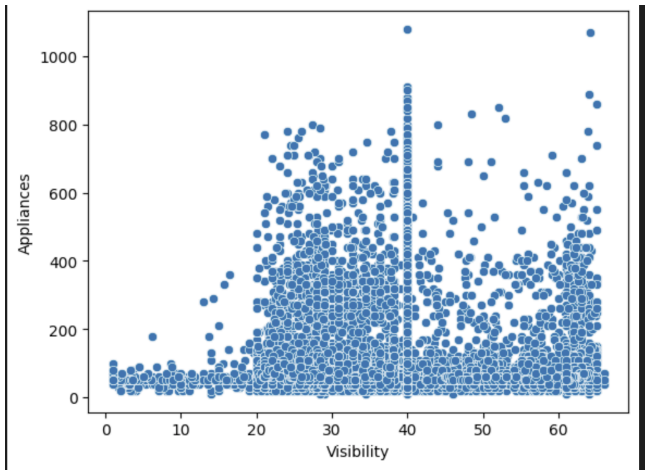


Scatter Plots:

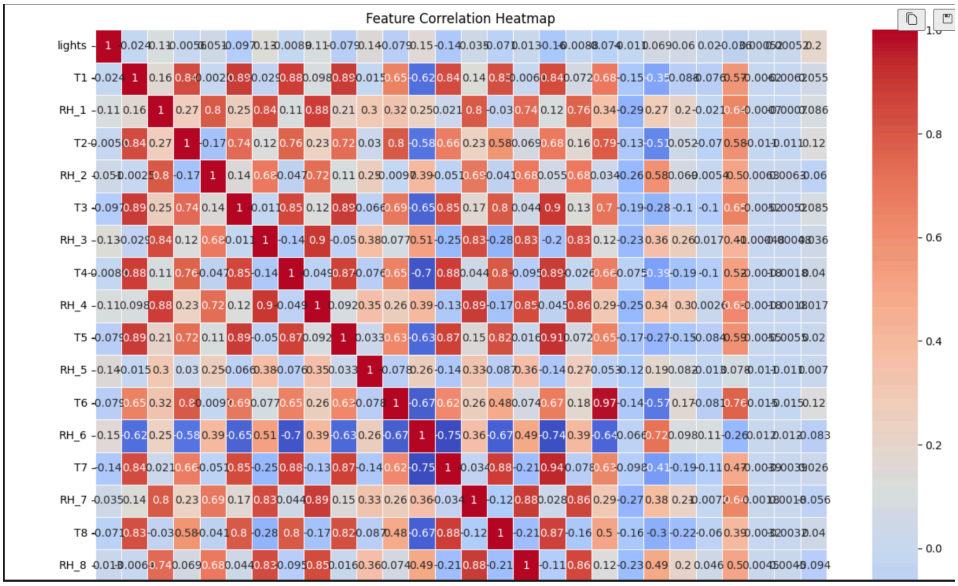
In this step, I aimed to understand how the data is distributed and explore the relationships between the target feature, Appliances, and the independent features. Upon analysis, I noticed that Appliances did not show a clear relationship with most features, which led me to investigate further. To gain deeper insights, I used a heatmap to examine the correlation between variables. This analysis revealed that the dataset is influenced by multiple factors rather than a direct relationship with any single feature.







HeatMap: In this step, I explored the correlations between features using a heatmap. The analysis revealed that while some features exhibit strong correlations, others do not. However, most features appear to have some level of correlation with each other, indicating possible relationships within the dataset. This insight helps in understanding feature dependencies and selecting relevant variables for further analysis.



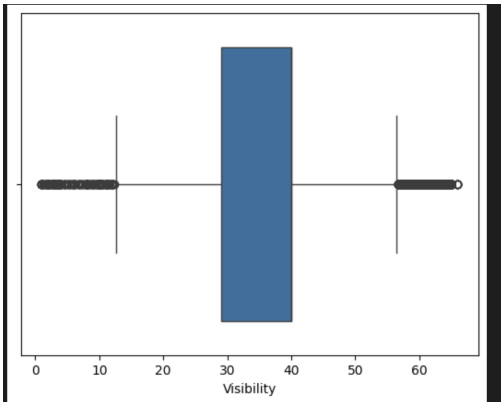
4. Handling Missing Values

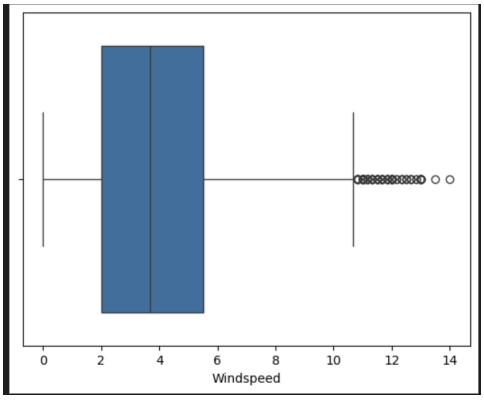
To ensure data quality, I conducted a missing values analysis by checking each column for missing or null entries. The results indicate that no missing values are present in the dataset across all features.

5. Outlier Detection and Treatment

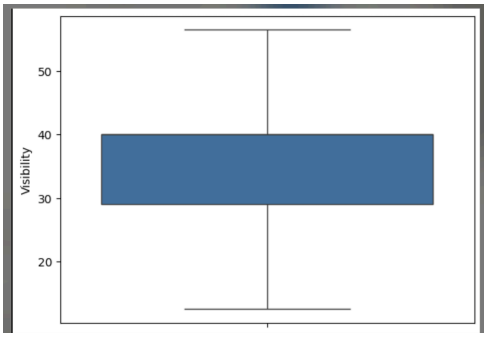
During my analysis, I observed that the dataset contains a significant number of outliers, with the majority of columns affected. To address this, I created a function that calculates whiskers (based on the IQR method) to determine the lower and upper bounds for detecting outliers.

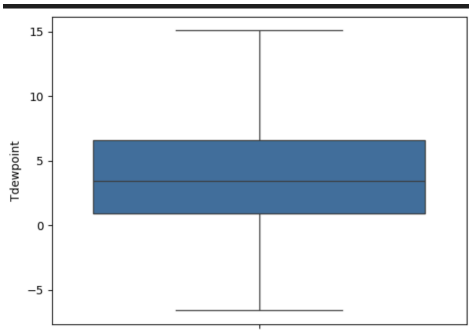
I have tested the function on a few columns, and it is working as expected. However, I have not yet decided whether to remove, transform, or retain the outliers, as I am still in the exploration phase of my analysis.





Columns with Outliers





After with no Outliers

5. Conclusion

In this report, I conducted data exploration and preprocessing to understand the structure, distribution, and relationships within the dataset. Key steps included:

- Checking for missing values, confirming that the dataset is complete.
- Using histograms and boxplots to analyze feature distributions and detect outliers.
- Examining scatter plots to assess relationships between the target variable (Appliances) and independent features.
- Applying a heatmap to visualize feature correlations, revealing dependencies among variables.

From this analysis, I observed that Appliances does not strongly correlate with most features, suggesting that multiple factors contribute to energy consumption. Additionally, many features contain outliers, which may require further treatment. I will decide later if I want to implement outlier treatments. For the future direction will be feature engineering, scaling, and model selection and other statistical techniques.

6. Appendix: GitHub Repo

For the full dataset, preprocessing, please refer to the **GitHub repository**:
🔗 **Statistic-Learning Repository**