Rudy Rajput(RXR210083), Shyam Devani(SJD210003)
CS 4372.501

Report

2.1 Pre-Processing

The dataset contained 52,416 records with nine attributes, including environmental variables (Temperature, Humidity, Wind Speed, General Diffuse Flows, Diffuse Flows) and three power consumption zones. The DateTime field was converted to a proper datetime format and then used to engineer additional features: Hour, Day, Month, and DayOfWeek.

Exploratory analysis was performed to check for data quality and attribute distributions. No missing values were detected, and all attributes were numeric, so no categorical encoding was necessary. Histograms revealed that most attributes were not normally distributed (Wind Speed was highly skewed, diffuse flows were concentrated near zero).

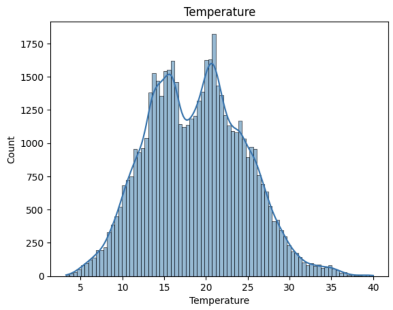
To ensure comparability, all features were standardized using StandardScaler.

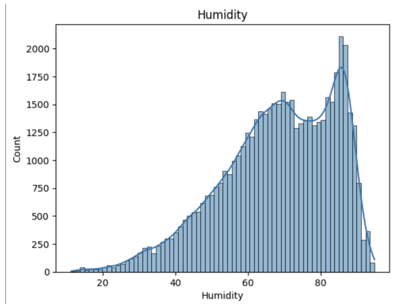
	DateTime	Temperature	Humidity	Wind Speed	general diffuse flows	diffuse flows	Zone 1 Power Consumption	Zone 2 Power Consumption	Zone 3 Power Consumption
count	52416	52416.000000	52416.000000	52416.000000	52416.000000	52416.000000	52416.000000	52416.000000	52416.000000
mean	2017-07-01 23:55:00	18.810024	68.259518	1.959489	182.696614	75.028022	32344.970564	21042.509082	17835.406218
min	2017-01-01 00:00:00	3.247000	11.340000	0.050000	0.004000	0.011000	13895.696200	8560.081466	5935.174070
25%	2017-04-01 23:57:30	14.410000	58.310000	0.078000	0.062000	0.122000	26310.668692	16980.766032	13129.326630
50%	2017-07-01 23:55:00	18.780000	69.860000	0.086000	5.035500	4.456000	32265.920340	20823.168405	16415.117470
75%	2017-09-30 23:52:30	22.890000	81.400000	4.915000	319.600000	101.000000	37309.018185	24713.717520	21624.100420
max	2017-12-30 23:50:00	40.010000	94.800000	6.483000	1163.000000	936.000000	52204.395120	37408.860760	47598.326360
std	NaN	5.815476	15.551177	2.348862	264.400960	124.210949	7130.562564	5201.465892	6622.165099

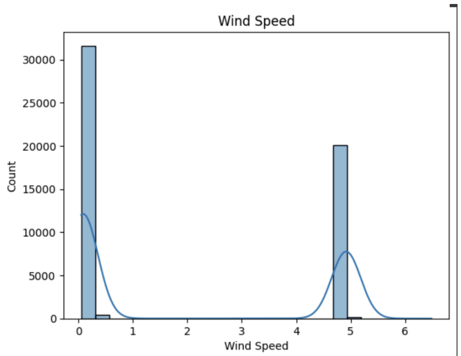
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52416 entries, 0 to 52415
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   DateTime                  52416 non-null  datetime64[ns]
 1   Temperature               52416 non-null  float64
 2   Humidity                  52416 non-null  float64
 3   Wind Speed                52416 non-null  float64
 4   general diffuse flows     52416 non-null  float64
 5   diffuse flows             52416 non-null  float64
 6   Zone 1 Power Consumption  52416 non-null  float64
 7   Zone 2  Power Consumption  52416 non-null  float64
 8   Zone 3  Power Consumption  52416 non-null  float64
dtypes: datetime64[ns](1), float64(8)
memory usage: 3.6 MB
```

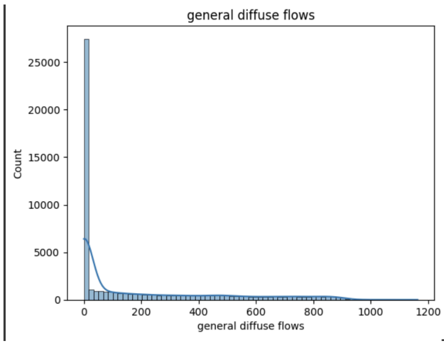
(52416, 9)

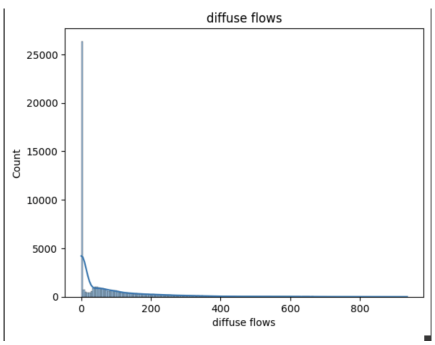
	DateTime	Temperature	Humidity	Wind Speed	general diffuse flows	diffuse flows	Zone 1 Power Consumption	Zone 2 Power Consumption	Zone 3 Power Consumption
0	2017-01-01 00:00:00	6.559	73.8	0.083	0.051	0.119	34055.69620	16128.87538	20240.96386
1	2017-01-01 00:10:00	6.414	74.5	0.083	0.070	0.085	29814.68354	19375.07599	20131.08434
2	2017-01-01 00:20:00	6.313	74.5	0.080	0.062	0.100	29128.10127	19006.68693	19668.43373
3	2017-01-01 00:30:00	6.121	75.0	0.083	0.091	0.096	28228.86076	18361.09422	18899.27711
4	2017-01-01 00:40:00	5.921	75.7	0.081	0.048	0.085	27335.69620	17872.34043	18442.40964

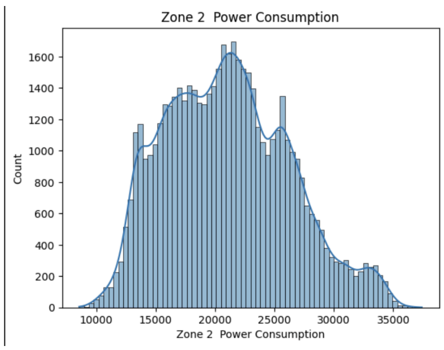

Temperature
Count
Temperature

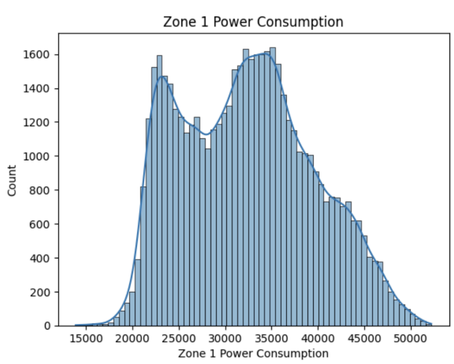

Humidity
Count
Humidity

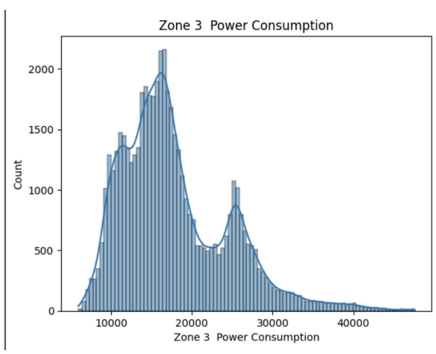

Wind Speed
Count
Wind Speed


general diffuse flows
Count
general diffuse flows


diffuse flows
Count
diffuse flows


Zone 2 Power Consumption
Count
Zone 2 Power Consumption


Zone 1 Power Consumption
Count
Zone 1 Power Consumption


Zone 3 Power Consumption
Count
Zone 3 Power Consumption

The distribution plots of the dataset reveal several important patterns:

- Temperature shows a roughly bell-shaped distribution with peaks around 15–25°C. This indicates seasonal variation, with most values concentrated in a comfortable range.
- Humidity has a right-skewed distribution with the majority of values between 60% and 85%. There are fewer low-humidity instances, suggesting the environment is generally humid.
- Wind Speed is highly skewed with two spikes: one near zero and another around 5 m/s. This suggests wind speeds are often calm or fall into a narrow operational range, with few moderate values in between.
- General Diffuse Flows and Diffuse Flows are extremely skewed toward zero, with most values clustered at very low levels. Only a small portion of the dataset contains higher values, which could weaken their predictive strength.
- Zone 1, Zone 2, and Zone 3 Power Consumption all display multi-modal distributions, with multiple peaks likely reflecting daily cycles of energy usage. Zone 1 has the highest average consumption, followed by Zone 2 and Zone 3.

Interpretation:
The variables are not normally distributed, which justifies standardization prior to modeling. In particular, skewed features such as wind speed and diffuse flows may have weaker predictive power, while cyclical effects are evident in the power consumption variables. These observations informed feature selection and the decision to include Hour as a derived predictor.

2.2 Model Construction

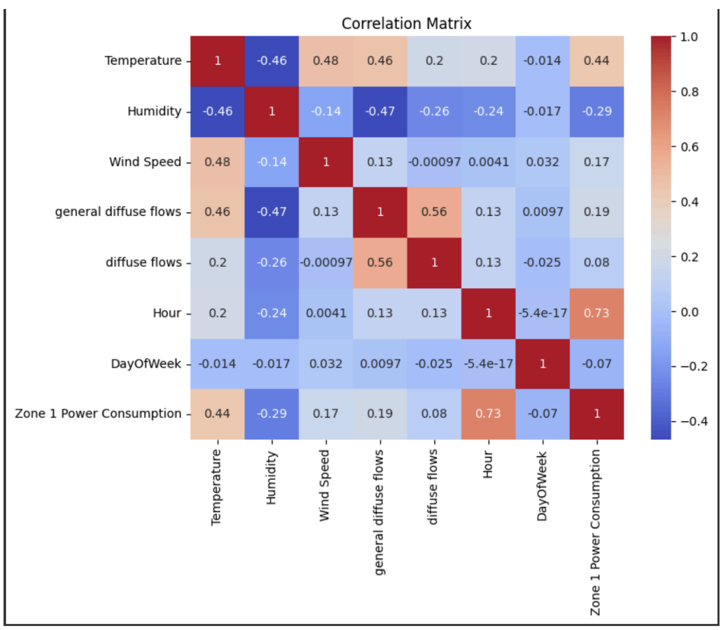
After preprocessing, the next step was to construct predictive models using both SGDRegressor from Scikit-Learn and Ordinary Least Squares (OLS) from the statsmodels library. This section describes the process of selecting features, scaling them, and building the two models.

Feature Selection

To avoid using all attributes blindly, I first examined correlations between predictors and the target variable. The correlation heatmap and the ranked correlation values showed that the most important predictors were:

- Hour (highest correlation, ~0.73)
- Temperature (~0.44)
- Humidity (~-0.29, negative correlation)
- Wind Speed (~0.17)
- General Diffuse Flows (~0.18)

These five features were selected for the final model.




```
Hour                        0.727953
Temperature                 0.440221
general diffuse flows       0.187965
Wind Speed                  0.167444
diffuse flows               0.080274
DayOfWeek                  -0.069708
Humidity                   -0.287421
Name: Zone 1 Power Consumption, dtype: float64
```

Standardization

Since SGDRegressor is sensitive to the scale of predictors, all five features were standardized using StandardScaler. The target variable was left unchanged. This ensured that no feature dominated the model purely because of its scale.

SGDRegressor with Hyperparameter Tuning

For the SGDRegressor, I performed hyperparameter tuning using a grid of reasonable parameter values. The parameters tuned included:

- loss: squared_error
- penalty: l2, elasticnet
- alpha: 1e-5, 1e-4, 1e-3
- learning_rate: adaptive
- eta0: 0.01
- l1_ratio: 0.15, 0.5

This produced 12 total combinations. Each configuration was trained and evaluated on a validation set, and the results were compared using R^2 and MSE. The top results are shown in the table below.

	loss	penalty	alpha	learning_rate	eta0	l1_ratio	train_R2	val_R2	train_MSE	val_MSE
0	squared_error	elasticnet	0.00010	adaptive	0.01	0.15	0.620820	0.631830	1.930737e+07	1.876592e+07
1	squared_error	l2	0.00010	adaptive	0.01	0.50	0.620820	0.631830	1.930737e+07	1.876593e+07
2	squared_error	l2	0.00010	adaptive	0.01	0.15	0.620820	0.631830	1.930737e+07	1.876593e+07
3	squared_error	elasticnet	0.00001	adaptive	0.01	0.50	0.620819	0.631828	1.930738e+07	1.876605e+07
4	squared_error	elasticnet	0.00001	adaptive	0.01	0.15	0.620819	0.631828	1.930738e+07	1.876605e+07
5	squared_error	l2	0.00001	adaptive	0.01	0.50	0.620819	0.631828	1.930738e+07	1.876605e+07
6	squared_error	l2	0.00001	adaptive	0.01	0.15	0.620819	0.631828	1.930738e+07	1.876605e+07
7	squared_error	elasticnet	0.00010	adaptive	0.01	0.50	0.620820	0.631828	1.930737e+07	1.876607e+07
8	squared_error	elasticnet	0.00100	adaptive	0.01	0.50	0.620820	0.631827	1.930734e+07	1.876612e+07
9	squared_error	elasticnet	0.00100	adaptive	0.01	0.15	0.620821	0.631824	1.930732e+07	1.876627e+07
10	squared_error	l2	0.00100	adaptive	0.01	0.15	0.620821	0.631822	1.930731e+07	1.876634e+07
11	squared_error	l2	0.00100	adaptive	0.01	0.50	0.620821	0.631822	1.930731e+07	1.876634e+07

From the grid search, the best configuration was:

- loss = squared_error
- penalty = l2
- alpha = 1e-4
- learning_rate = adaptive
- eta0 = 0.01

This configuration balanced bias and variance well and gave stable convergence.

OLS Regression

The OLS model was fit using the same five predictors. Unlike SGD, OLS does not require hyperparameters. Instead, it provides a full statistical summary, including coefficients, standard errors, t-statistics, p-values, R^2, adjusted R^2, and overall model significance (F-statistic). The OLS summary is included later in the results section, where it is compared directly with the tuned SGD model.

2.3 Result Analysis

This section presents the results of both the SGDRegressor (with hyperparameter tuning) and the OLS regression models, along with an interpretation of their performance and diagnostic statistics.

SGDRegressor Results

After hyperparameter tuning, the best SGDRegressor configuration achieved the following metrics:

- Training R^2: 0.623
- Testing R^2: 0.627
- Training MSE: ~1.92 × 10^7
- Testing MSE: ~1.88 × 10^7

```
SGD Train R²: 0.6230479611596695
SGD Test  R²: 0.6271400361417907
SGD Train MSE: 19198361.897367507
SGD Test  MSE: 18827335.09026342
```

These results indicate that the model generalizes well, with nearly identical train and test performance. The R^2 value around 0.62–0.63 suggests that the model explains about 62–63% of the variance in Zone 1 power consumption. While not perfect, this level of accuracy is reasonable given the variability in the data and relatively simple features.

OLS Results

The OLS regression using the same five predictors produced nearly identical performance metrics:

- Training R^2: 0.623
- Testing R^2: 0.627
- Training MSE: ~1.92 × 10^7
- Testing MSE: ~1.88 × 10^7

```
Training MSE: 19198361.154848322
Training R²: 0.6230479757387322
Testing MSE: 18827399.43699315
Testing R²: 0.6271387618074673
```

OLS Regression Results

Dep. Variable:	Zone 1 Power Consumption	**R-squared:**	0.623
Model:	OLS	**Adj. R-squared:**	0.623
Method:	Least Squares	**F-statistic:**	1.386e+04
Date:	Tue, 16 Sep 2025	**Prob (F-statistic):**	0.00
Time:	02:50:42	**Log-Likelihood:**	-4.1111e+05
No. Observations:	41932	**AIC:**	8.222e+05
Df Residuals:	41926	**BIC:**	8.223e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	3.234e+04	21.399	1511.111	0.000	3.23e+04	3.24e+04
x1	4799.8506	22.295	215.293	0.000	4756.153	4843.548
x2	2312.2341	29.026	79.660	0.000	2255.342	2369.126
x3	39.6047	25.927	1.528	0.127	-11.213	90.423
x4	125.5774	24.642	5.096	0.000	77.279	173.876
x5	-328.5833	25.534	-12.868	0.000	-378.631	-278.536

Omnibus:	778.616	**Durbin-Watson:**	2.016
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	868.030
Skew:	0.303	**Prob(JB):**	3.23e-189
Kurtosis:	3.361	**Cond. No.**	2.39

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The OLS summary also provided statistical insights into each predictor's significance:

- Hour and Temperature were strongly significant predictors (very low p-values).
- Humidity had a significant negative effect, consistent with the correlation analysis.
- Wind Speed showed a weak and statistically insignificant contribution ($p \approx 0.127$).
- General Diffuse Flows was significant but with a smaller effect size.

The F-statistic was highly significant ($p < 0.001$), confirming that the overall regression model is statistically valid.

Insights

- Both models (SGD and OLS) gave very similar results in terms of accuracy and error. This makes sense because SGD is basically a scalable optimization method for linear regression, and with proper tuning it converges to results close to OLS.
- The main takeaway is that time of day and temperature are the most influential predictors of Zone 1 energy use. Humidity also matters, but less so. Wind speed, on the other hand, doesn't contribute much.
- Although the models explained around 62–63% of the variation, there's still room for improvement. If we wanted to get better accuracy, we could try adding non-linear models (like Random Forests or Gradient Boosting), or include more features (such as occupancy data or appliance usage logs).