ELECTRICITY CONSUMPTION FORECASTING

USING MACHINE LEARNING: DATA ANALYSIS AND MODELING

BESHER KEBBE

Project Description

Project Objective: This project aims to predict the total electricity consumption of a city for one year.

Project Scope: This project involves forecasting electricity consumption using data analytics and machine learning techniques.

Project Goals:

The Energy Management Department of Chicago aims to predict the total electricity consumption for the next year to prevent issues like sudden consumption spikes in winter and peak air conditioning use in summer, thereby making energy management more efficient.

Data Used:

- Electricity consumption data (monthly/total)
- Building Characteristics (Building Type, Subtype, Average Number of Floors, Average Building Age, Average House Size)
- Population Data (Total Population)
- Geographical Location Data (Districts)

Expected Results:

- Develop a model that accurately predicts the total electricity consumption for the next year.
- Identify which factors affect electricity consumption and to what extent.
- Develop efficient strategies for energy planning and infrastructure investments by making accurate predictions.

Importance of Our Prediction Results

- **Energy Planning:** Accurately predicting the future energy demand of a city, region, or country is necessary for planning electricity supply and distribution.
- **Infrastructure Investments:** Decisions such as identifying regions for investment, constructing new plants or distribution lines, and strengthening existing infrastructure are based on predictions.
- **Pricing Strategies:** Electricity consumption forecasts are used to set energy prices and develop demand-based pricing strategies. Increasing prices during peak demand times can help balance demand.

Importance of Our Prediction Results

Renewable Energy Integration: It is possible to plan the electricity production of wind turbines or solar panels according to predicted demand levels.

Energy Efficiency: If an increase in electricity consumption is predicted for a certain area, projects that provide energy savings can be planned.

Environmental Quality and Carrying Capacity
Environmental Management and Development Strategy
Lifestyles of Sustainability
Sustainable Urban Development
Lowering Environmental Impact & Maintaining Carrying Capacity
A Sound Socio-economic Environment
Adequate Infrastructure
Development Strategy Considering Human and Natural Environment Concurrently
Sense of Wellbeing and Work-Life Balance
Cultural Diversity and Tolerance
Consumer Goods and Services
Sports, Leisure & Recreation
Environmental Quality Monitoring
Natural Resource Consumption

Dataset Overview

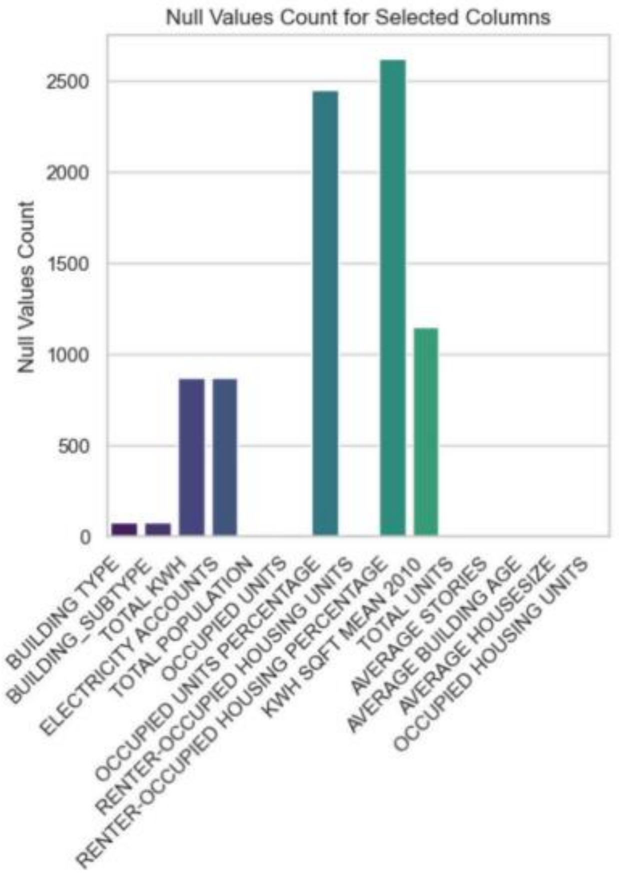
The dataset includes various metrics covering total electricity and gas consumption for 2010 in Chicago, residential unit occupancy, demographic information, building characteristics, and more.

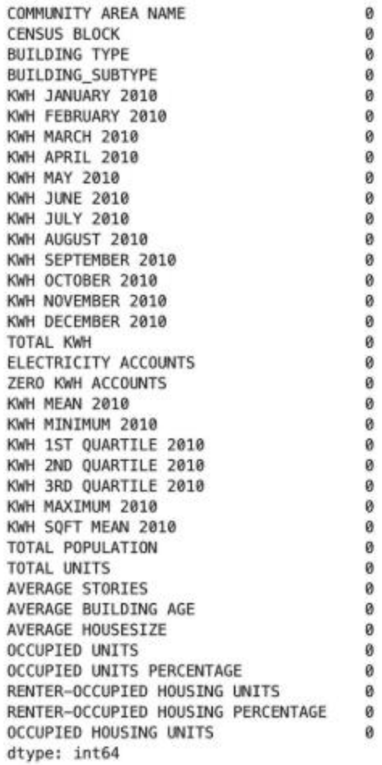
The dataset contains 67,051 rows and covers 73 features.

The total size of the dataset is 24.7MB.

https://data.cityofchicago.org/Environment-Sustainable-Development/Energy-Usage-2010/8yq3-m6wp/data

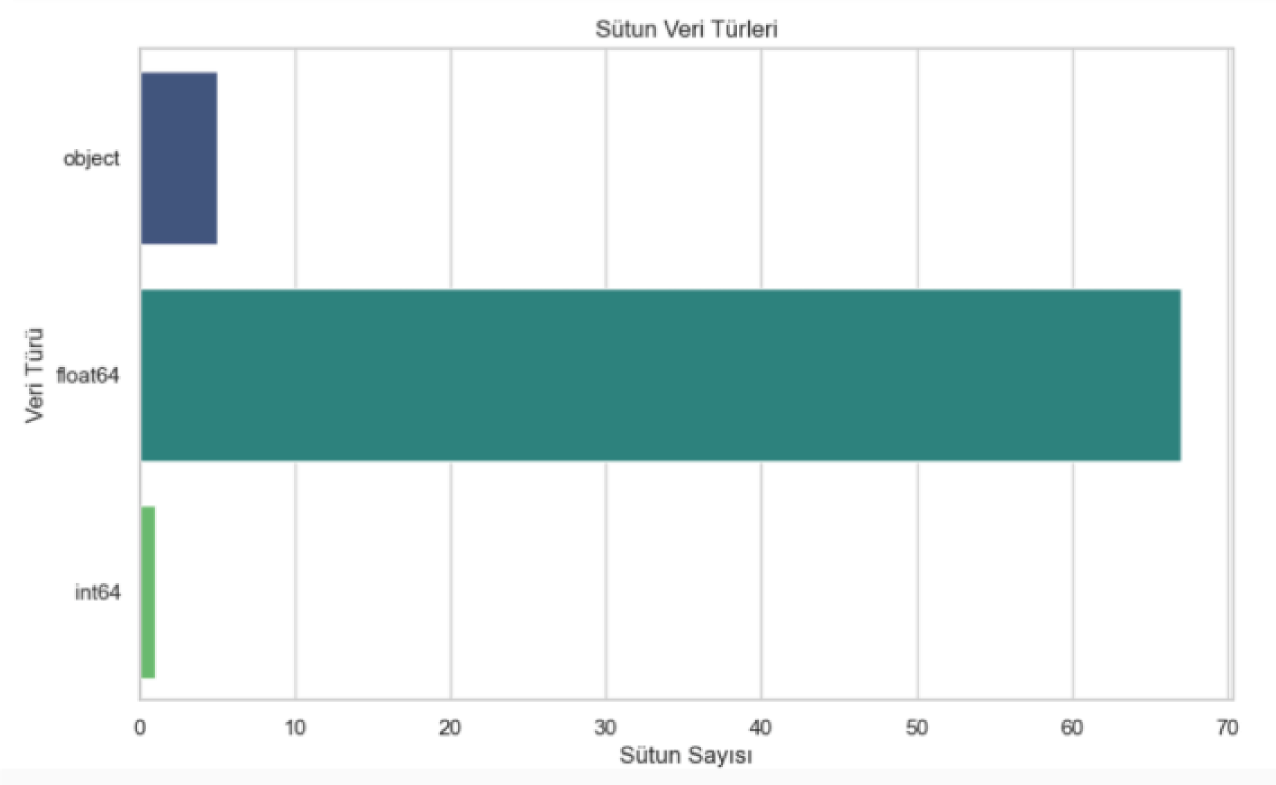
DATA ANALYSIS AND CLEANING: DETECTION AND CLEANING OF MISSING VALUES

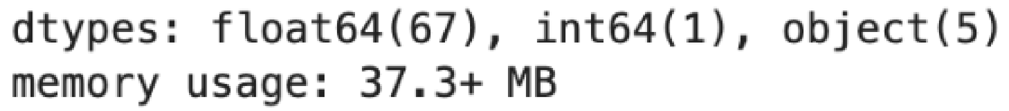




```
COMMUNITY AREA NAME                    0
CENSUS BLOCK                           0
BUILDING TYPE                          0
BUILDING_SUBTYPE                       0
KWH JANUARY 2010                       0
KWH FEBRUARY 2010                      0
KWH MARCH 2010                         0
KWH APRIL 2010                         0
KWH MAY 2010                           0
KWH JUNE 2010                          0
KWH JULY 2010                          0
KWH AUGUST 2010                        0
KWH SEPTEMBER 2010                     0
KWH OCTOBER 2010                       0
KWH NOVEMBER 2010                      0
KWH DECEMBER 2010                      0
TOTAL KWH                              0
ELECTRICITY ACCOUNTS                   0
ZERO KWH ACCOUNTS                      0
KWH MEAN 2010                          0
KWH MINIMUM 2010                       0
KWH 1ST QUARTILE 2010                  0
KWH 2ND QUARTILE 2010                  0
KWH 3RD QUARTILE 2010                  0
KWH MAXIMUM 2010                       0
KWH SQFT MEAN 2010                     0
TOTAL POPULATION                       0
TOTAL UNITS                            0
AVERAGE STORIES                        0
AVERAGE BUILDING AGE                   0
AVERAGE HOUSESIZE                      0
OCCUPIED UNITS                         0
OCCUPIED UNITS PERCENTAGE              0
RENTER-OCCUPIED HOUSING UNITS          0
RENTER-OCCUPIED HOUSING PERCENTAGE     0
OCCUPIED HOUSING UNITS                 0
dtype: int64
```

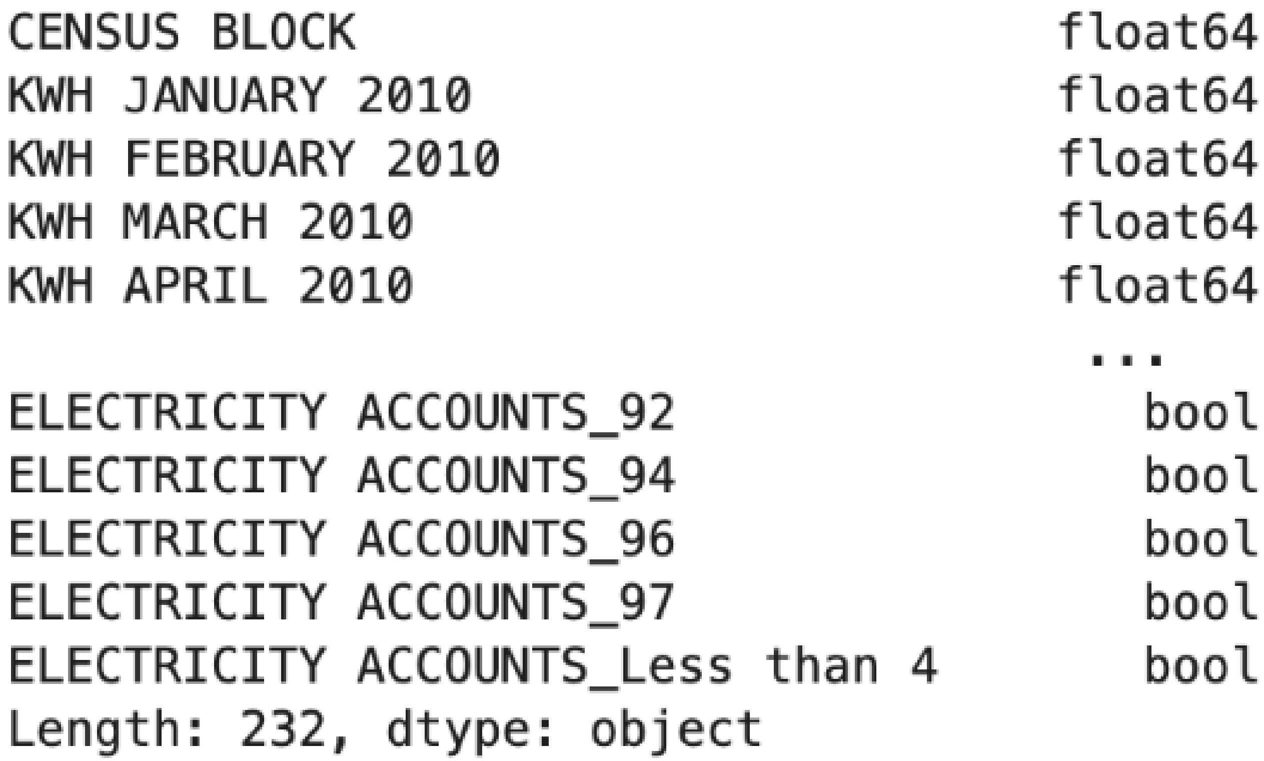
DATA ANALYSIS AND CLEANING: DETECTION OF CATEGORICAL DATA





```
dtypes: float64(67), int64(1), object(5)
memory usage: 37.3+ MB
```

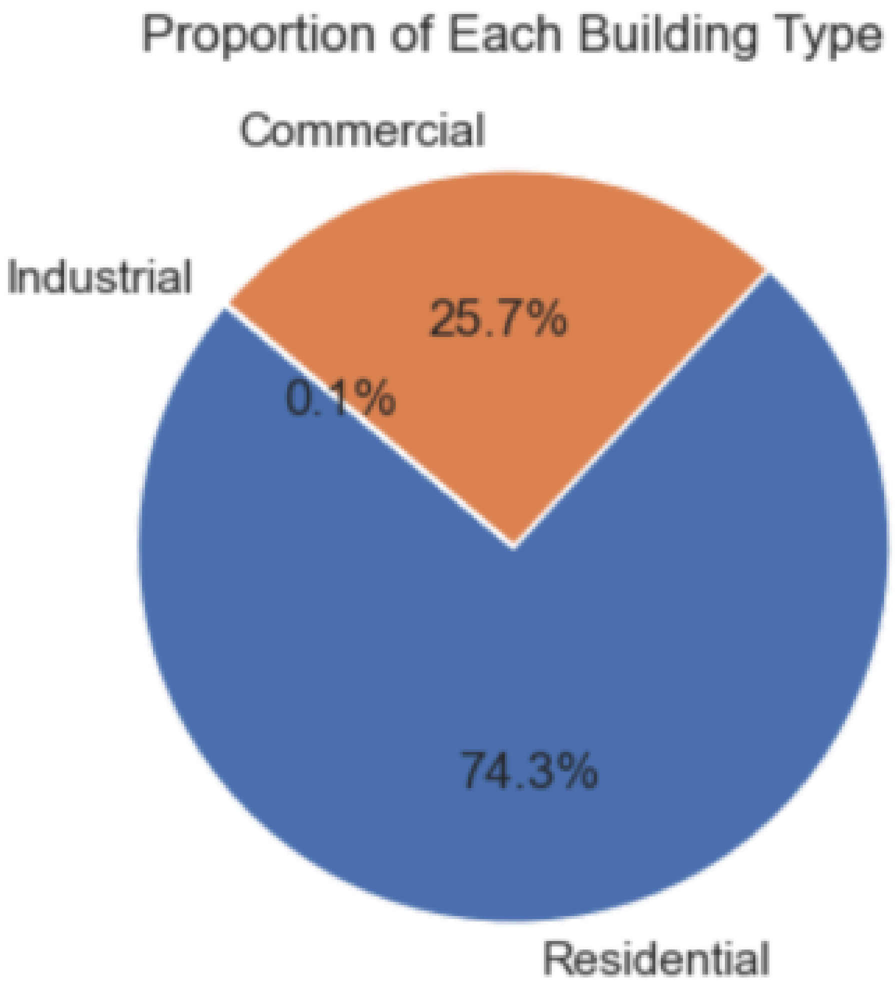
DATA ANALYSIS AND CLEANING: CORRECTION OF CATEGORICAL DATA

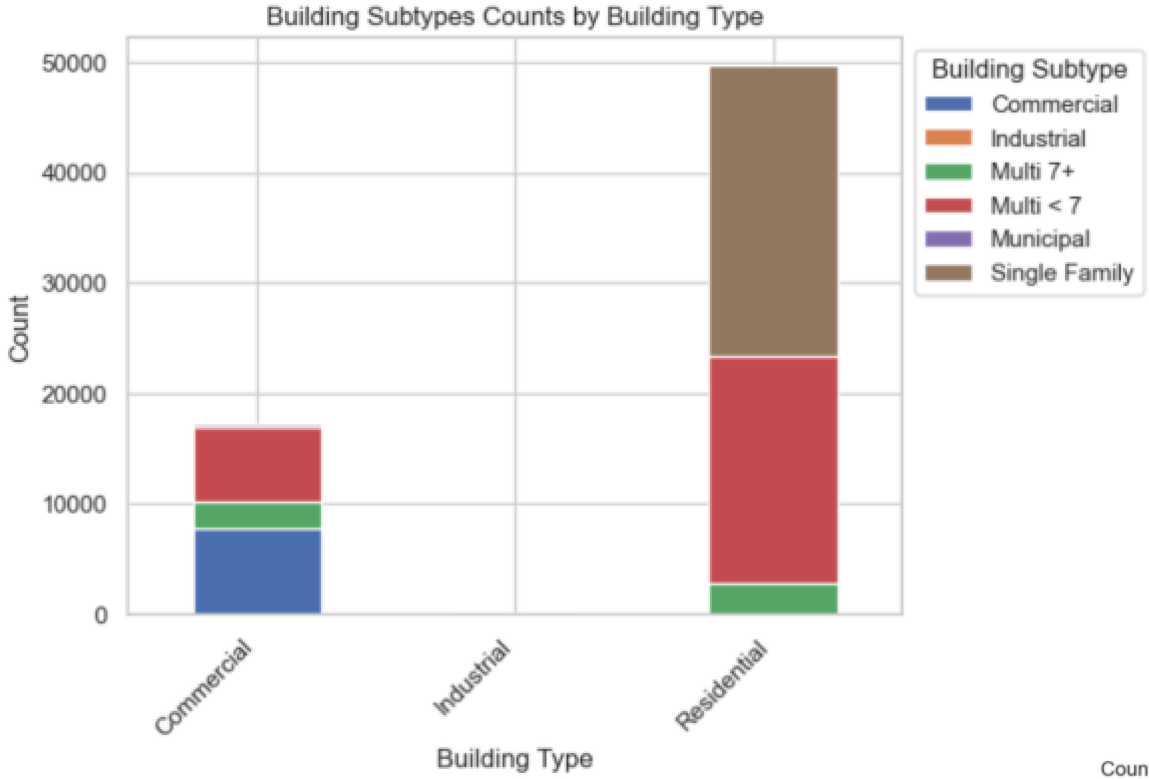


```
CENSUS BLOCK                        float64
KWH JANUARY 2010                    float64
KWH FEBRUARY 2010                   float64
KWH MARCH 2010                      float64
KWH APRIL 2010                      float64
                                     ...
ELECTRICITY ACCOUNTS_92                bool
ELECTRICITY ACCOUNTS_94                bool
ELECTRICITY ACCOUNTS_96                bool
ELECTRICITY ACCOUNTS_97                bool
ELECTRICITY ACCOUNTS_Less than 4       bool
Length: 232, dtype: object
```

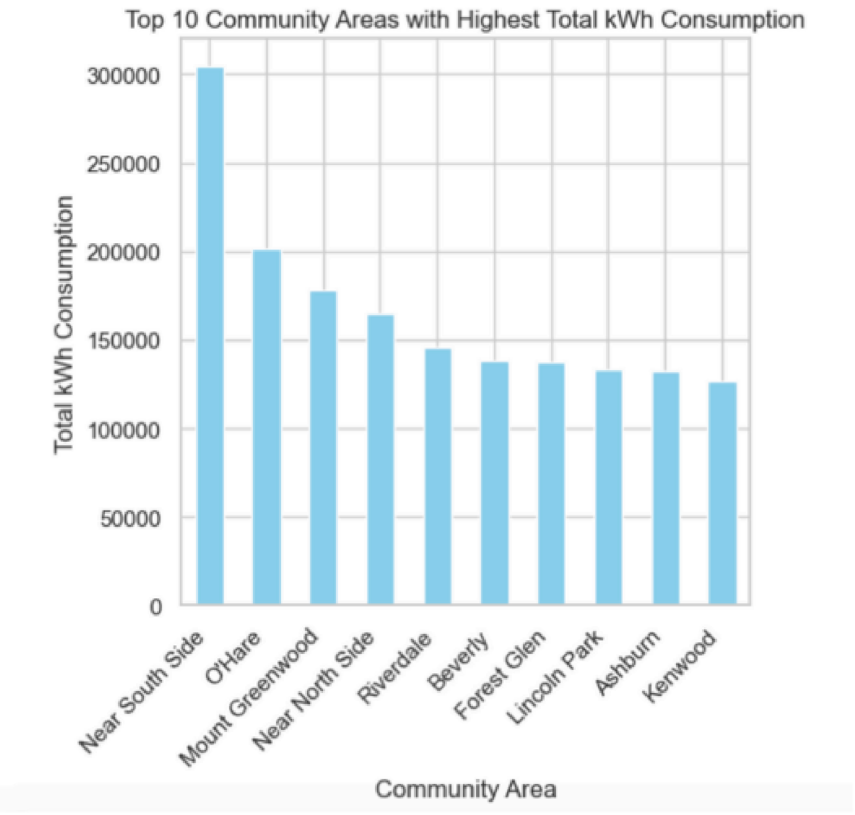
Data Analysis and Cleaning: Data Review

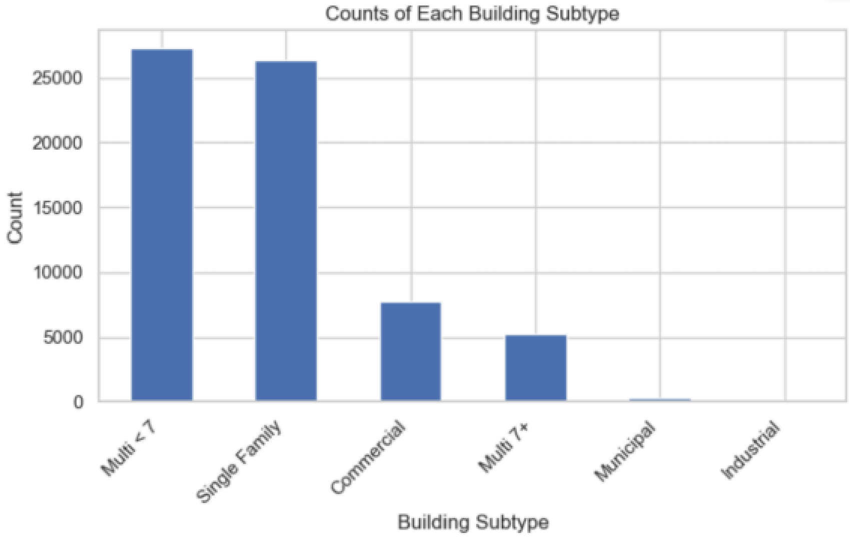
The dataset contains three building types: Residential, Commercial, Industrial. Out of 67,051 records, 49,747 are residential, 17,185 are commercial, and the remaining 47 records belong to industrial properties.



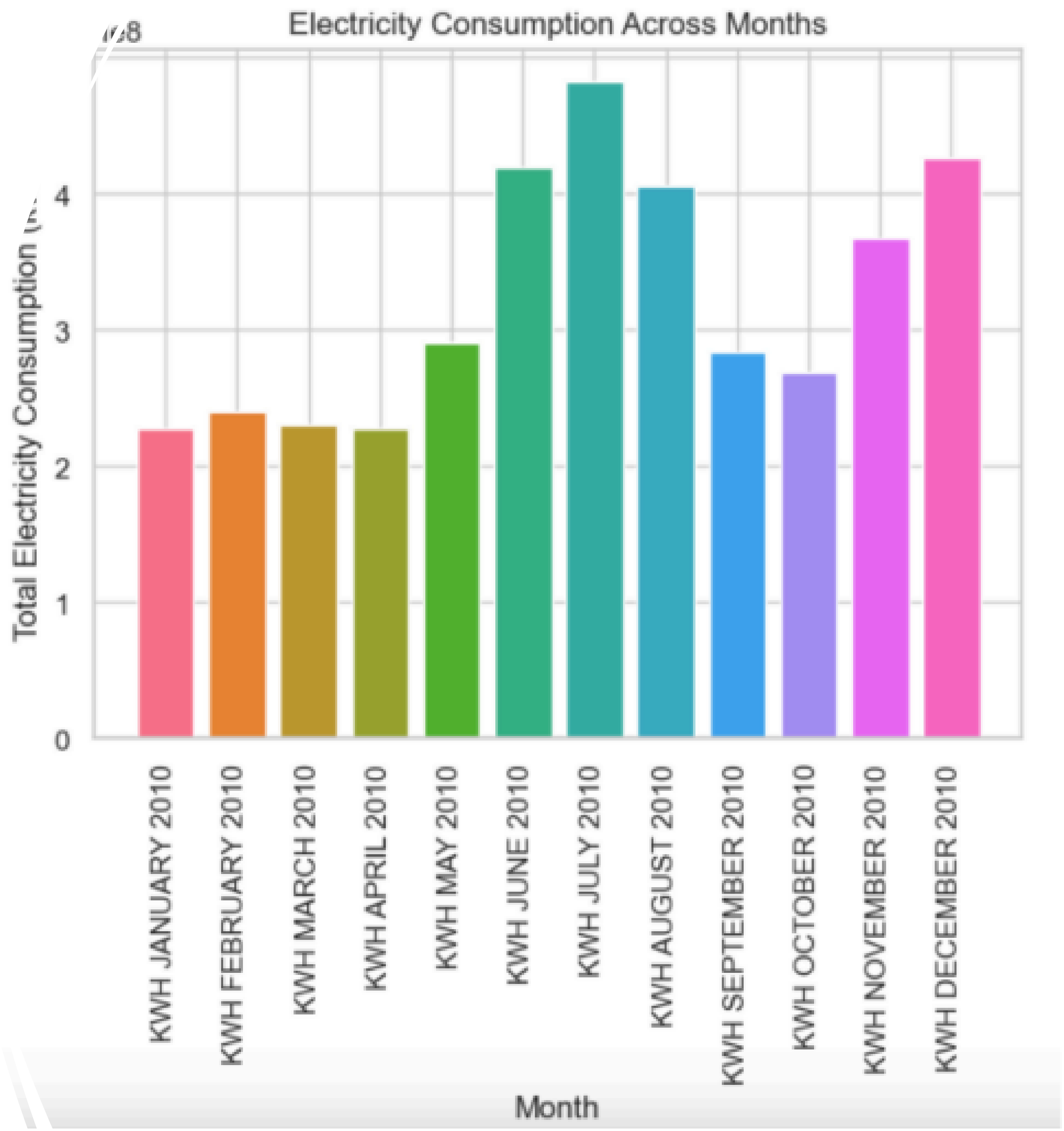

Building Subtypes Counts by Building Type
Count
50000
40000
30000
20000
10000
0
Commercial
Industrial
Residential
Building Type
Building Subtype
Commercial
Industrial
Multi 7+
Multi < 7
Municipal
Single Family

[redacted]

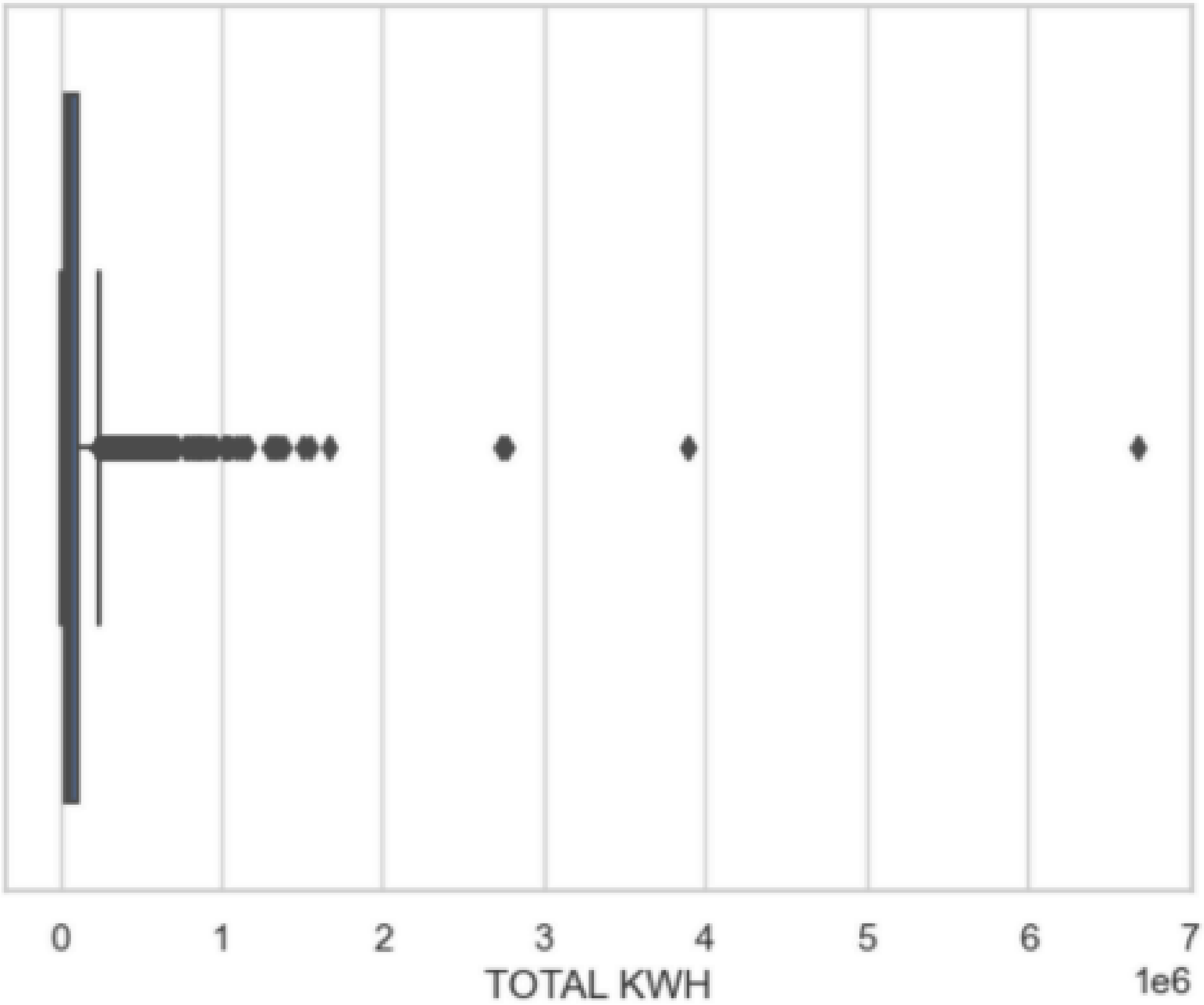

Top 10 Community Areas with Highest Total kWh Consumption
Total kWh Consumption
300000
250000
200000
150000
100000
50000
0
Near South Side
O'Hare
Mount Greenwood
Near North Side
Riverdale
Beverly
Forest Glen
Lincoln Park
Ashburn
Kenwood
Community Area


Counts of Each Building Subtype
Count
25000
20000
15000
10000
5000
0
Multi < 7
Single Family
Commercial
Multi 7+
Municipal
Industrial
Building Subtype

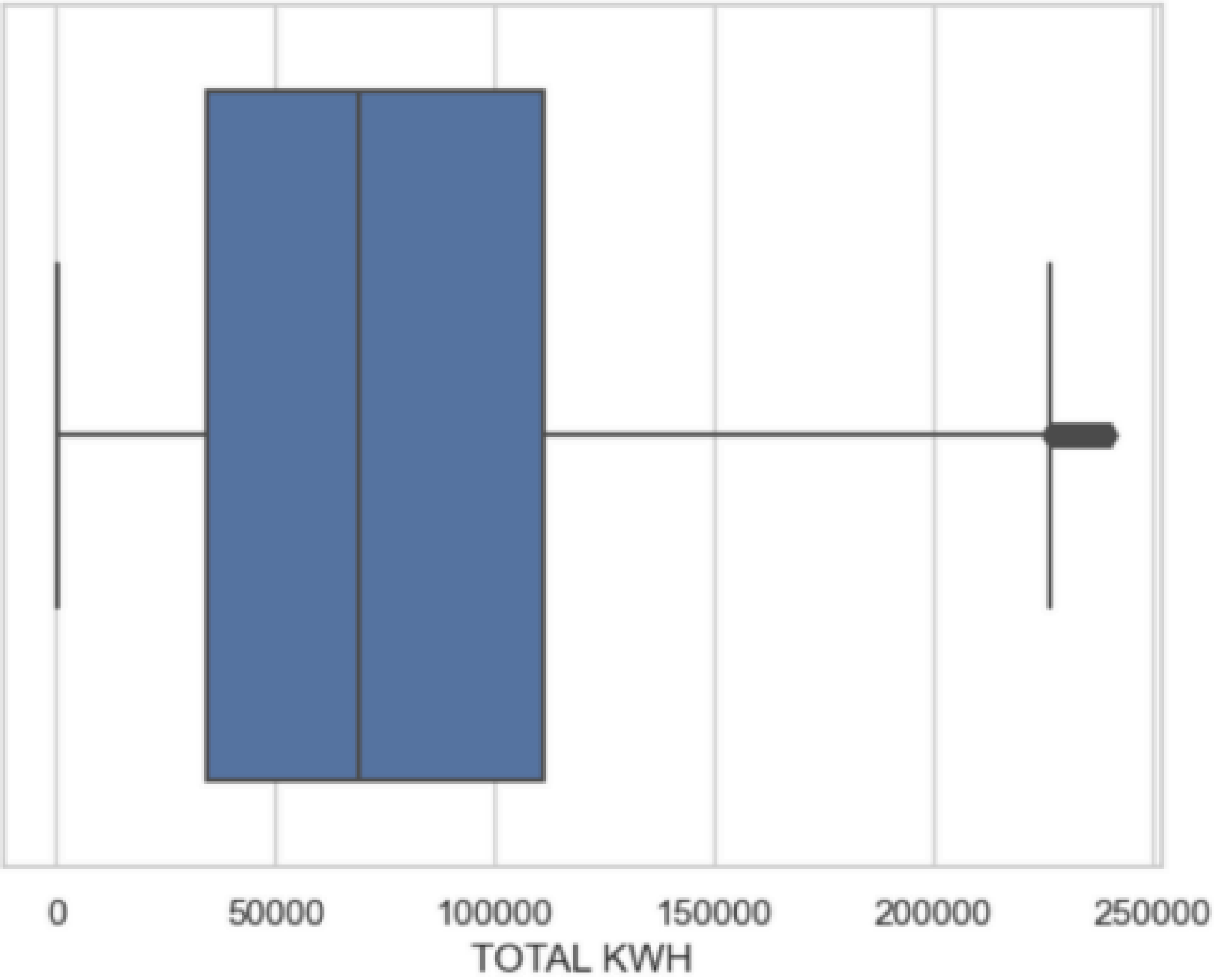
Data Analysis and Cleaning: Data Review



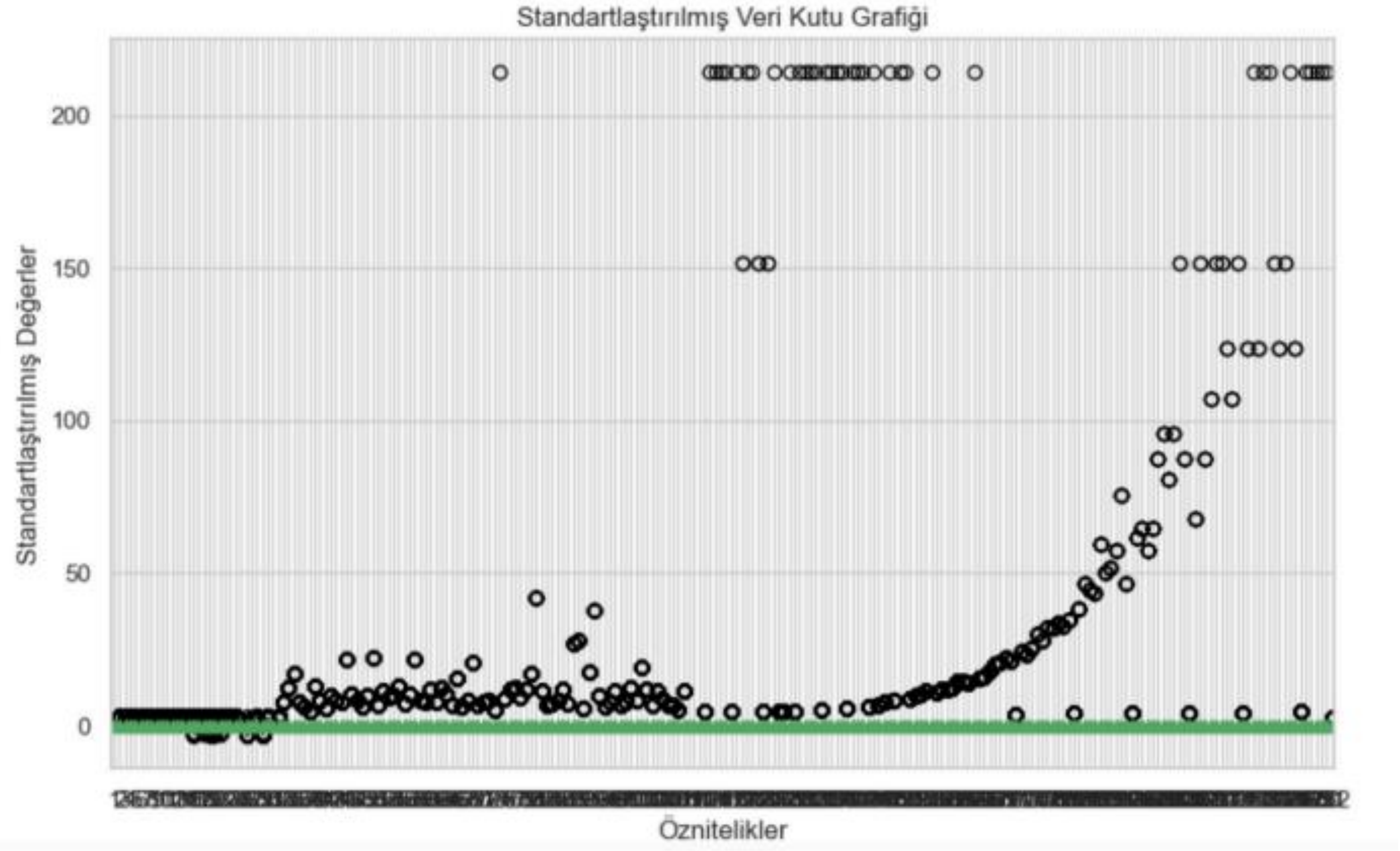
Data Analysis and Cleaning: Detection of Outliers



DATA ANALYSIS AND CLEANING: CORRECTION OF OUTLIERS

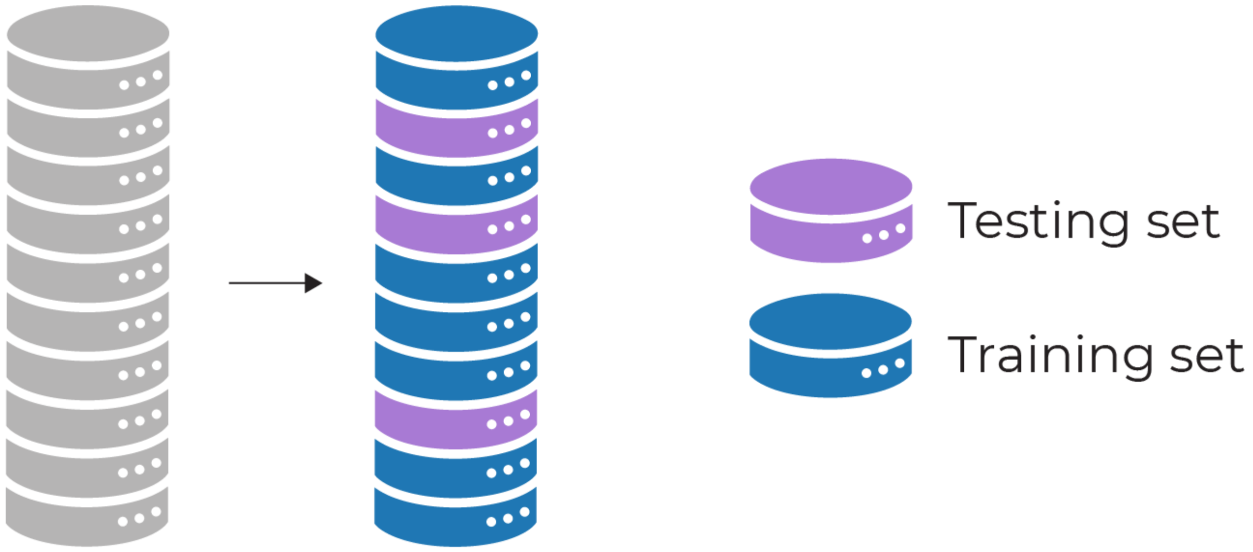


DATA ANALYSIS AND CLEANING: STANDARDIZATION



Data Split

- **Dependent Variables:** All features except "TOTAL KWH"
- **Independent Variable:** "TOTAL KWH"
- **Train-Test Split:** 25% test set, 75% training set



MACHINE LEARNING MODELS

LinearRegression

DecisionTreeRegressor

MultiOutputRegressor

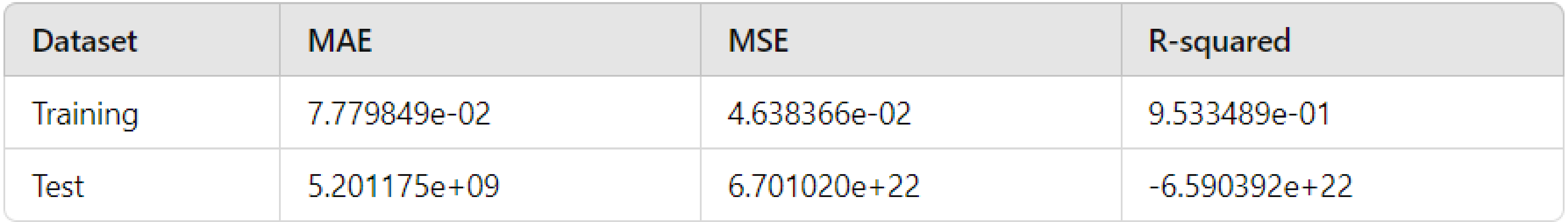
XGBRegressor

GradientBoostingRegressor

Linear Regression

The Linear Regression model was attempted, assuming a linear relationship between the dependent variable (total electricity consumption) and the independent variables (monthly electricity consumption, population, building characteristics, etc.).

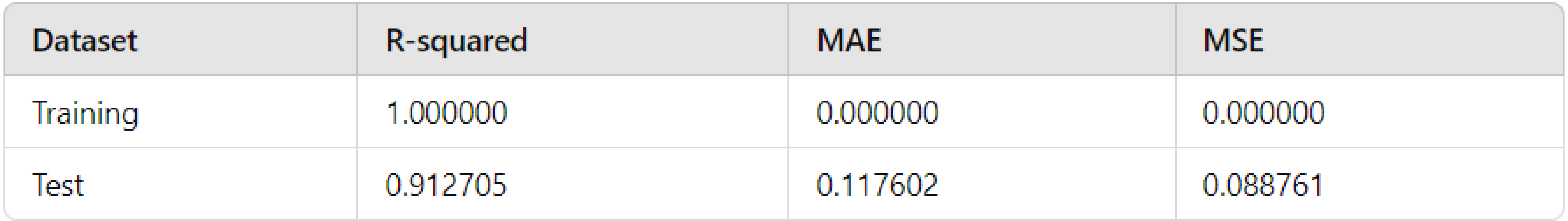
Training and test prediction results are provided below:



Dataset	MAE	MSE	R-squared
Training	7.779849e-02	4.638366e-02	9.533489e-01
Test	5.201175e+09	6.701020e+22	-6.590392e+22

Decision Tree Regressor

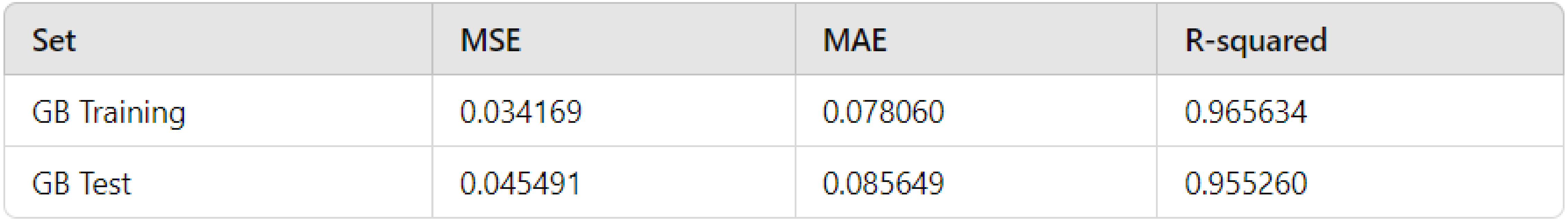
Using decision tree models known for their effectiveness in handling complex patterns aimed to improve prediction accuracy and robustness. The process revealed high variance issues, signaling potential overfitting and reduced generalization ability.



Dataset	R-squared	MAE	MSE
Training	1.000000	0.000000	0.000000
Test	0.912705	0.117602	0.088761

Gradient Boosting

The Gradient Boosting algorithm showed different performance on the training and test sets. Low error rates (MSE and MAE) and high R^2 value (0.9656) on the training set indicate the model fits the training data very well. The test set performance showed a negligible decrease.



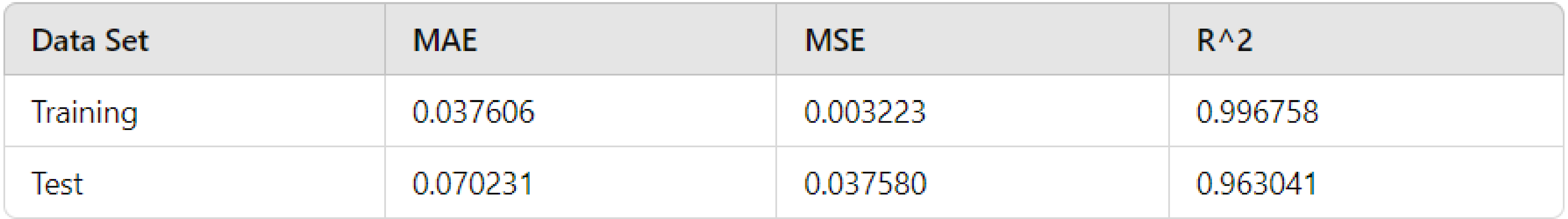
Set	MSE	MAE	R-squared
GB Training	0.034169	0.078060	0.965634
GB Test	0.045491	0.085649	0.955260

XGBoost

According to the prediction results using the XGBoost Regressor model, the model performed well on the training set.

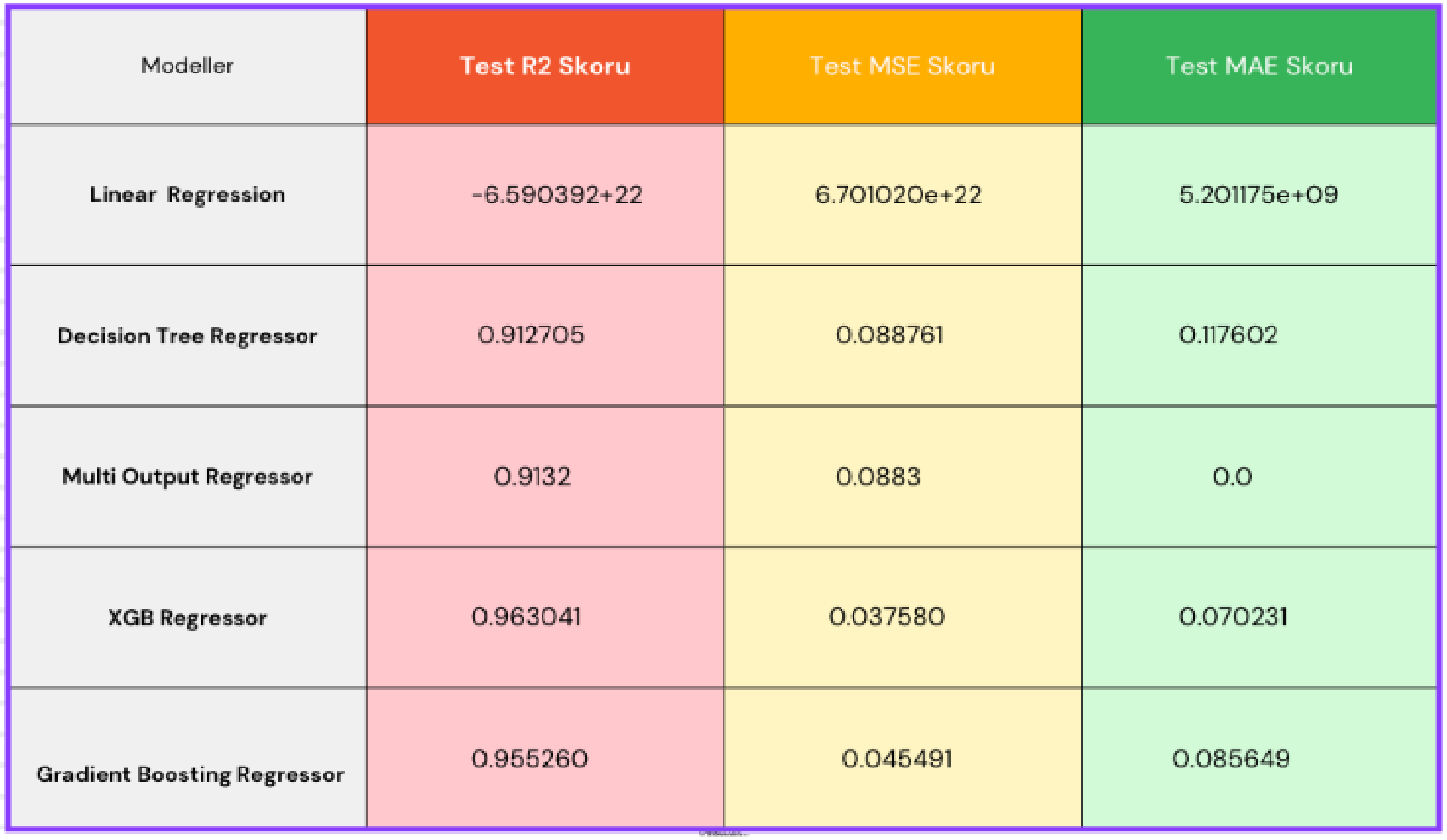
The Mean Absolute Error (MAE) value was 0.0376, and the Mean Squared Error (MSE) value was 0.0032. Additionally, the R-squared (R^2) value was quite high, explaining 99.68% of the variance. For the test set, the MAE value was 0.0702, and the MSE value was 0.0376. The test set R-squared value was found to be 96.30%.

These results indicate that the model maintained its performance in the test set, showing reliable and consistent predictions.



Data Set	MAE	MSE	R^2
Training	0.037606	0.003223	0.996758
Test	0.070231	0.037580	0.963041

GENERAL REVIEW



Modeller	Test R2 Skoru	Test MSE Skoru	Test MAE Skoru
Linear Regression	-6.590392+22	6.701020e+22	5.201175e+09
Decision Tree Regressor	0.912705	0.088761	0.117602
Multi Output Regressor	0.9132	0.0883	0.0
XGB Regressor	0.963041	0.037580	0.070231
Gradient Boosting Regressor	0.955260	0.045491	0.085649

Model Evaluation and Conclusion

The Linear Regression model showed results suggesting overfitting, becoming highly dependent on the training set. The model might memorize noise and fluctuations in the training data and fail when applied to new data.

The Decision Tree Regressor model appeared to fit perfectly on the training set, but its low R-squared value and high MSE on the test set indicate overfitting and poor generalization.

The XGBoost model performed very well on both training and test sets, with high R-squared values and low MSE, showing good data explanation and accurate predictions.

The Gradient Boosting Regressor model also performed very well on both sets, with high R-squared values and low MSE, indicating good data fit and accurate predictions.

Model Evaluation and Conclusion

In conclusion, the XGBoost model showed the best performance on this dataset, with lower MSE and higher R-squared values compared to other models.

The electricity consumption forecasts for the next year can be presented to the city's energy management department.

These forecasts will enable more efficient management of energy resources, for instance, preparing for increased demand in winter and developing strategic plans to prevent overloads in summer. The forecasting model produced will contribute to the development of efficient strategies for energy planning and resource management.

THANK YOU

BESHER KEBBE