*Original Research Article*

**Empowering New York City: Leveraging Surplus Renewable Energy for Community Benefits**

**Abstract:** The rapid advancement of artificial intelligence (AI) has significantly increased the energy demands of data centers, placing growing pressure on urban power grids. In parallel, New York City has experienced a rise in renewable energy production, yet mismatches between generation and consumption have resulted in surplus electricity being frequently underutilized. Simultaneously, the phased retirement of fossil fuel power plants has contributed to power supply instability, disproportionately impacting low-income communities and public services. Addressing these converging challenges, this study proposes a data-driven credit refund system that redistributes surplus renewable energy based on borough-level electricity consumption and population metrics.

To develop and validate this system, three years of borough-level electricity consumption, renewable energy production, and weather data (2021–2023) were collected and analyzed. Preprocessing steps included machine learning–based imputation of missing values and outlier detection using techniques such as K-Nearest Neighbors and the Isolation Forest algorithm. Random Forest regression was employed to model surplus electricity as a function of environmental variables, capturing non-linear seasonal relationships.

Hydroelectric power was found to be the most stable energy source year-round, while solar and wind exhibited strong seasonal patterns—solar peaking in summer and wind in winter—demonstrating a significant inverse correlation between temperature and wind speed. These dynamics underscore the importance of precise weather forecasting and the role of environmental variability in optimizing grid efficiency.

Simulation results indicate that the proposed credit refund system could equitably allocate economic benefits to residents across boroughs, with the Bronx and Manhattan receiving the highest credits due to their population density and energy use. This model offers a scalable, sustainable solution for utilizing surplus electricity, improving grid efficiency, and supporting energy equity. The findings present a replicable framework for other cities seeking to integrate renewable energy surpluses into community-focused energy policy.

**1. Introduction**

The rapid advancement of artificial intelligence (AI) has led to a significant increase in energy consumption by data centers. In the financial sector, AI models are employed for high-frequency trading and market analysis, while in healthcare, AI-based algorithms assist in disease diagnosis and drug discovery [1]. Similarly, AI systems are implemented in manufacturing for smart process management and quality control [2]. In logistics and distribution, AI plays a vital role in real-time route optimization and inventory management [3]. These AI models involve processing vast amounts of data, which significantly increases the energy demands of large-scale data centers. Data centers continuously consume large amounts of power, placing a constant load on local power grids. If grid stability deteriorates, residents may face higher utility bills and an increased likelihood of power outages [4].

The growing energy demand may also hinder business activities. When data centers prioritize using local power grid resources, the energy available to small and medium-sized enterprises (SMEs) and factories becomes limited [5]. This can lead to decreased productivity and higher operational costs, potentially stifling regional economic growth. Energy-intensive industries, such as manufacturing and technology-based businesses, are particularly vulnerable to such energy competition [6].

Public facilities face significant challenges as well. Essential services such as hospitals, schools, and transportation systems heavily depend on stable and predictable power supplies [7]. For example, hospitals cannot power their medical equipment adequately, and disruptions in public transportation systems due to power shortages could cause widespread inconvenience to citizens in their daily lives. Compounding this issue is the phased decommissioning of fossil fuel power plants. Over the past five years, New York State has retired 5,207 megawatts (MW) of fossil fuel capacity to meet climate change goals, but only 2,256 MW of renewable energy capacity has been added during the same period. This disparity highlights an imbalance in energy supply as adopting renewable energy fails to keep pace with the reduction in fossil fuel dependence. Power supply shortages lead to higher electricity rates, disproportionately impacting low-income households [8].

Although New York State generates a substantial amount of renewable energy, imbalances between production and consumption result in surplus electricity [9]. If this surplus is not managed correctly, it risks being wasted, increasing operational costs for power systems and imposing additional environmental burdens. These conditions contribute to rising energy costs and declining public service quality, posing long-term risks to regional economies and social stability [10]. Therefore, innovative energy management solutions that effectively utilize surplus electricity are urgently needed to address the dual challenges of rising power demands due to AI and the reduction of fossil fuel dependency [10]. In response, this research aims to design and implement a data-driven credit refund system that enhances energy efficiency and generates social and environmental value by utilizing surplus electricity. By analyzing surplus electricity data in New York City and developing algorithms to redistribute it efficiently to areas of greatest need, the research aims to address complex energy challenges and provide an innovative solution that can be applied to other cities.

During the initial phase of the research, monthly and seasonal power consumption and production data for New York City were collected at the borough level [11]. The data were sourced from the New York City Open Data Portal and reliable energy reports and processed using Python's pandas library. Since the data were complex and incomplete, preprocessing tasks such as imputing missing values and removing outliers were essential. Missing values were corrected using interpolation methods based on nearby data points, while outliers were effectively addressed using the Isolation Forest algorithm [12]. Additionally, Random Forest Regression [13] was employed to analyze key variables such as weather, time of day, and industrial versus residential energy use ratios influencing surplus electricity in each borough, enabling insights into variable correlations.

The project aimed to analyze borough-level energy patterns and implement a fair redistribution of surplus electricity, ultimately reducing energy costs and delivering economic benefits to local communities. It focused on analyzing borough-level energy consumption patterns and redistributing surplus electricity fairly to help reduce electricity bills. This model provides tangible financial benefits to the community while enabling a more precise understanding of power usage patterns through the proportional allocation of surplus electricity based on borough consumption shares. This method ensures fairness in the redistribution process and highlights each area's contribution to energy consumption. To ensure fair distribution, surplus electricity was allocated proportionally based on borough-level energy consumption and then divided by average population to estimate per capita credit. Public facilities—such as hospitals and schools—were prioritized to enhance energy efficiency and maintain service stability. First, surplus electricity from each borough was distributed proportionally based on energy consumption shares and then divided by the borough's average population to determine per-capita credit. The surplus electricity was monetized to offer residents a tangible economic benefit. This redistribution model is also applied to public facilities, enhancing overall energy efficiency in the community. Public facilities, which provide essential services and consume substantial energy, were prioritized for surplus electricity distribution to ensure a stable energy supply and lower operational costs.

By utilizing the aforementioned methods to analyze the electricity consumption patterns of public facilities, it becomes possible to effectively leverage surplus energy, thereby reducing unnecessary energy waste and maximizing the utilization of energy resources. This approach improves energy efficiency in public facilities, thereby strengthening essential services for residents. Moreover, it can help reduce the load on the power grid, thereby increasing the stability and sustainability of the regional power network in the long term. Such an approach lowers energy costs for public facilities and promotes the use of environmentally friendly energy. It serves as a crucial strategy to enhance local communities' environmental and economic value, fostering a more sustainable and efficient energy ecosystem. Moreover, the developed model offers scalability and flexibility, making it applicable to New York City and other regions facing similar energy challenges. By analyzing local energy consumption and production data, the model can be customized for efficient implementation across diverse regional contexts, contributing to establishing sustainable energy management strategies locally and internationally.

**2. Experimental methods**

**2.1 Data Collection**

This study utilized various public data sources to conduct an in-depth analysis of electricity consumption and production patterns in New York City. Regional data from NYC Open Data and energy reports were supplemented with information from the Energy Information Administration (EIA) [14], which provides national-level electricity production and consumption data disaggregated by region. Additionally, weather data were sourced from the National Oceanic and Atmospheric Administration (NOAA)[15], a trusted public platform that provides weather information across the United States.

To analyze the role of different energy sources, data on solar, wind, and fossil fuel generation were collected, highlighting the relative contributions of renewable and non-renewable systems. Regional production data categorized by boroughs were incorporated to understand the geographical distribution of electricity generation. Electricity consumption data played a key role in examining demand-side energy dynamics and identifying usage trends across boroughs.

**2.2 Data Preprocessing**

Data preprocessing was a vital step in ensuring the reliability and accuracy of data analysis and modeling. In this study, preprocessing focused on improving data quality by addressing outliers, handling missing data, normalizing variables, and preparing datasets for visualization. These steps were necessary to create a robust dataset to deliver meaningful insights and support effective predictive models. One of the primary tasks in preprocessing was the removal of outliers. Extreme electricity consumption and production data values could distort analysis and obscure underlying patterns. Therefore, such values were identified and either adjusted or excluded to maintain the integrity of the dataset. This step helped ensure that the data accurately reflected real-world behaviors and trends.

Missing data were handled using several advanced methods to ensure consistency and completeness in the dataset. For example, time interpolation filled gaps by averaging values from adjacent days. This preserved the temporal continuity of electricity trends. Additionally, KNN imputation was employed to predict missing values based on neighboring data points with similar conditions, such as temperature and precipitation. For more complex relationships, regression models were used, such as predicting missing wind speed values based on related variables like temperature and precipitation. These approaches provided a sophisticated way to address data gaps, improving the dataset's reliability and analytical value.

Normalization was applied to scale variables such as electricity consumption (kWh), temperature (°C), and wind speed (m/s), ensuring equal weight in model training and preventing bias due to differing units. For instance, electricity consumption, temperature, and wind speed were scaled consistently, allowing for more accurate comparisons and model training. This step ensured that the machine learning models treated all variables fairly during analysis. The preprocessed data was organized into well-structured datasets that facilitated clear visualization and interpretation. Key variables, such as electricity consumption, production, and surplus electricity, were included in formats that supported the generation of graphs and charts. These visualizations revealed key seasonal and regional consumption trends, offering actionable insights for targeted policy design and urban energy management.

**2.3 Analysis and Modeling**

The analysis and modeling phase focused on calculating surplus electricity in New York City, analyzing its patterns, and simulating the benefits of optimized electricity usage. Surplus electricity was calculated by subtracting the total monthly electricity consumption in New York City from the total electricity production. This calculation enabled an evaluation of surplus energy availability, incorporating temporal, seasonal, and regional characteristics. Machine learning models such as K-Nearest Neighbors (KNN), Linear Regression, and Random Forest were utilized to calculate and predict surplus electricity [16]. Linear Regression effectively modeled linear relationships between variables, while Random Forest handled nonlinear interactions and provided high prediction accuracy. KNN was particularly useful for predicting surplus electricity based on regional patterns. These models contributed to identifying key factors influencing surplus electricity and enhancing predictions' reliability. The analysis of surplus electricity patterns involved visualizing monthly and seasonal trends across New York City and its boroughs to identify major drivers of variability. Heatmaps, line graphs, and bar charts were used to illustrate trends in periods or regions where surplus electricity significantly increased or decreased. These visualizations offered actionable insights for exploring surplus electricity utilization strategies and supporting policy decision-making.

In the credit refund system simulation, an algorithm was developed to equitably distribute refunds based on surplus electricity and the population ratio of each borough. This algorithm was designed to ensure fairness by calculating refund amounts proportional to each borough's share of the total population. Machine learning models were employed during the simulation process to predict how the refund system could reduce residents' electricity costs. Linear Regression was used to analyze the relationship between refund amounts and cost reductions, while Random Forest evaluated the variability in cost savings across different scenarios, producing reliable and robust results.

To validate the reliability and accuracy of the models, predictions were compared against actual data using evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² [17]. These metrics provided quantitative insights into how closely the models reflected real-world data. Visualization further supported the validation process by comparing predicted and actual values through line graphs and scatter plots. This approach evaluated how well the models captured the seasonal changes and monthly variations in surplus electricity. Each machine learning model demonstrated distinct strengths during performance validation. Linear Regression was effective for evaluating simple and interpretable relationships, Random Forest excelled in handling non-linear relationships and delivering high prediction accuracy, and KNN proved particularly useful for predictions based on regional similarities. This comprehensive analysis and modeling process provided actionable, data-driven insights and established a robust foundation for decision-making and policy recommendations aimed at optimizing surplus electricity utilization in New York City.

**3. Results and Discussion**

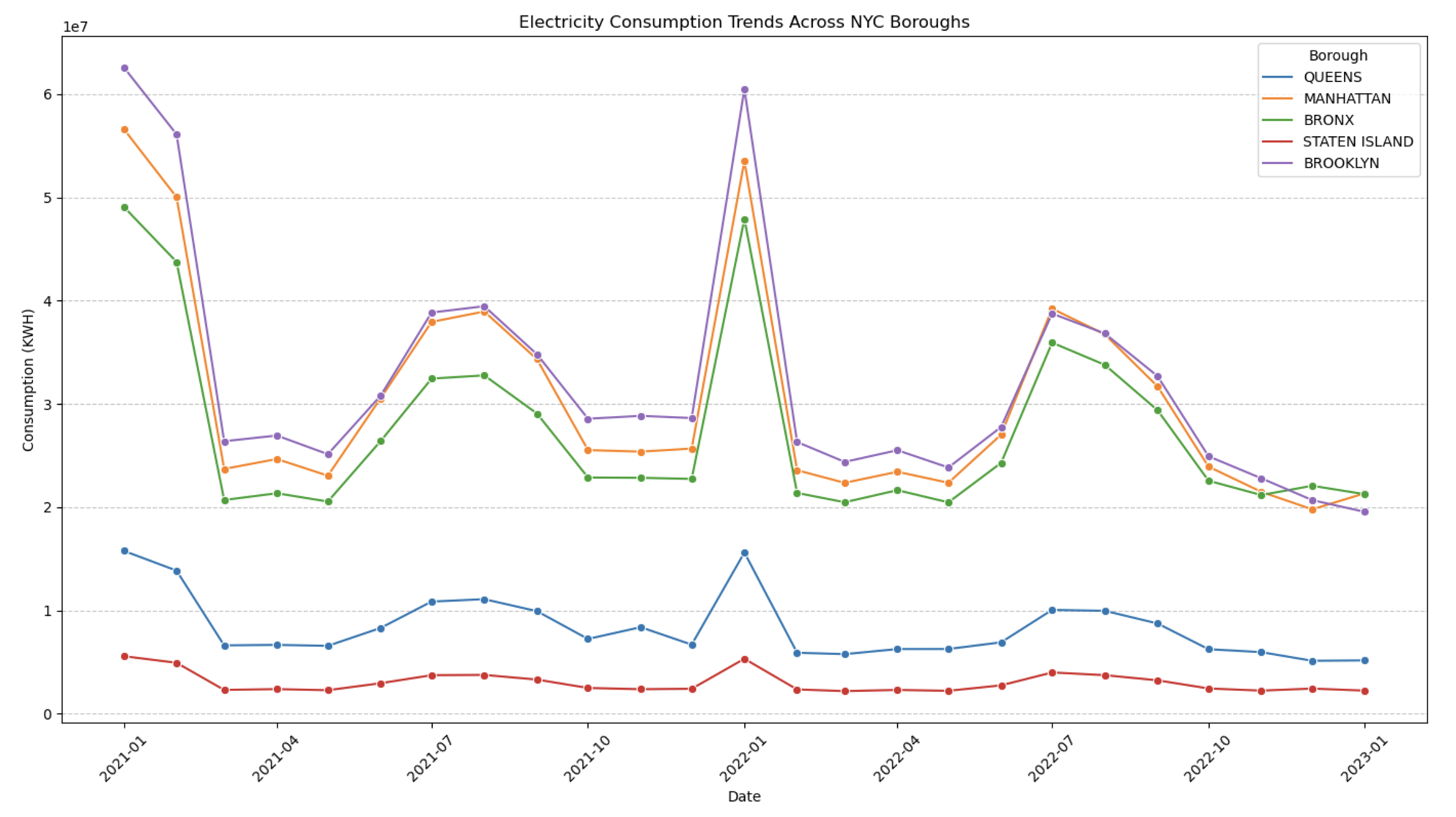
**3.1 Borough-Level Electricity Consumption Patterns and Weather Correlation Analysis**

The correlation between electricity consumption and weather variables was analyzed for each borough in New York City. Electricity consumption across boroughs was highly sensitive to weather factors like temperature and precipitation, demonstrating clear seasonal variability. To stabilize credit allocations over time, the concept of Average Population was introduced. This was derived by averaging historical population data from the U.S. Census Bureau and annual population reports, thereby mitigating volatility due to short-term migration or estimation errors. Similarly, in winter, heating demand resulted in another increase in consumption, highlighting a clear seasonal pattern. IAlthough the dataset initially spanned from 2010 to 2023, the analysis focused on the most recent three years (2021–2023) to ensure data quality and manageability. However, due to the size and complexity of the dataset, the analysis was narrowed down to focus on the most recent three years (2021–2023). During this period, monthly electricity consumption data for each borough in New York City was organized, establishing the primary scope of the project. Weather variables were closely linked to electricity consumption and renewable energy production, requiring a high degree of reliability for correlation analysis. Missing values and outliers were identified and addressed using appropriate machine learning models to ensure this.

Missing values were handled using K-Nearest Neighbors (KNN) imputation, which analyzed the patterns of adjacent dates and predicted missing values based on similar characteristics in the data. Additionally, regression models were employed to predict missing values for specific variables (e.g., temperature and precipitation) that had strong correlations with other variables. Outlier removal was performed using the Isolation Forest Algorithm, automatically detecting and removing anomalously high or low values in the weather data. Furthermore, the interquartile range (IQR) method was applied to prevent data distortion. Additionally, the column values were reviewed to extract the necessary data for each borough from the raw dataset. The selected columns for analysis included Borough, Revenue Month, Consumption (KWH), KW Charges, and # days. The Revenue Month column was converted into a uniform Datetime format to standardize the data structure. However, duplicate data issues frequently arose during the data cleaning process, making it challenging to obtain the desired results. Duplicate entries often arose from variations in Meter Scope, where multiple meters in a single location recorded separate values for the same revenue month.

For instance, in cases where electricity consumption in a building or area was measured by multiple meters, data for the same month (Revenue Month) would be recorded separately for each meter. Differences in variables such as Meter AMR, Location, or other columns could also have caused the data to split into multiple rows. A data integration approach was applied to resolve this issue, consolidating all Meter Scope data for the same borough and revenue month. Electricity consumption was calculated as the sum, while charges (KW Charges) were averaged. This process was implemented using ‘groupby’ operations, effectively removing duplicates and ensuring data consistency. As a result of these efforts, a reliable dataset was constructed, enabling the analysis of the correlation between monthly electricity consumption and weather variables for each borough in New York City. This refined dataset provided a solid foundation for the project’s key analyses and simulations.

Given the time-series nature of weather data, models like Random Forest Regression were well-suited to capture nonlinear interactions between weather variables and electricity consumption. Random Forest Regression effectively captures non-linear relationships between variables and handles the complex interactions often found in weather data. By utilizing this preprocessing approach and machine learning model, the reliability and accuracy of the data were significantly improved, providing a solid foundation for analyzing the correlation between weather variables and electricity consumption with greater clarity.

Figure 1: illustrates borough-level monthly electricity consumption trends (2021–2023), highlighting seasonal peaks, especially in Brooklyn and Manhattan. 

Line graph showing monthly consumption patterns (KWH × 10^7) for each borough. Brooklyn shows the highest peaks (~6 × 10^7 KWH in winter), followed by Manhattan's stable high consumption. The Bronx maintains moderate-high usage, while Queens and Staten Island show lower consumption levels. All boroughs demonstrate clear seasonal variations with winter/summer peaks.

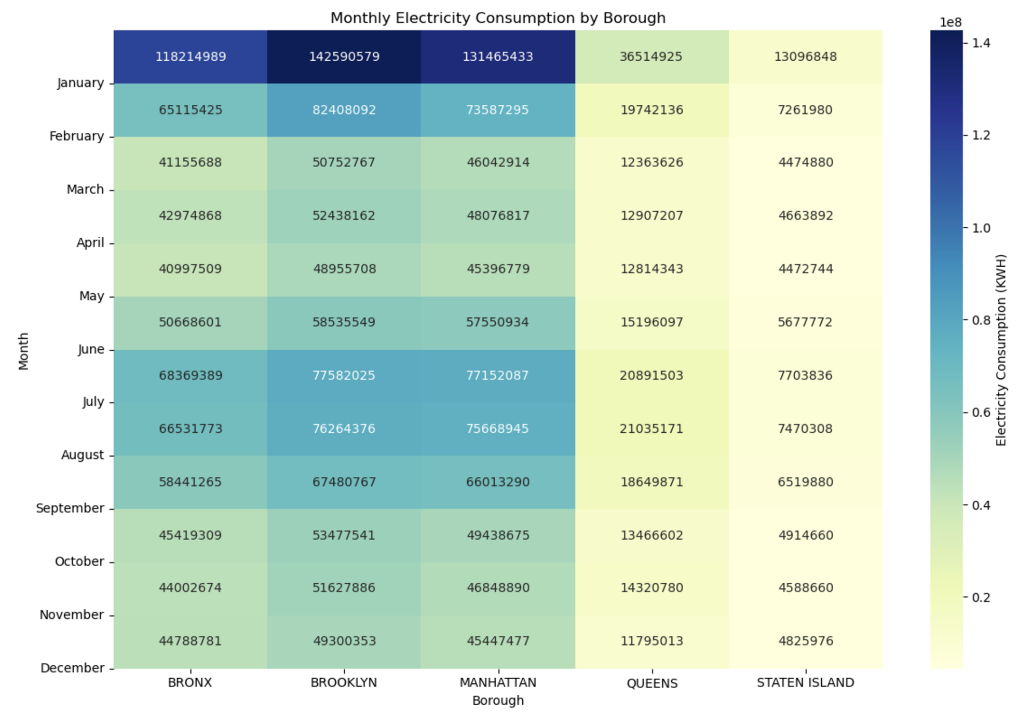


Figure 2: NYC Boroughs Monthly Electricity Consumption Heatmap (KWH × 10^8). Dark blue indicates peak consumption (winter/summer), while yellow shows lower usage. BROOKLYN shows the highest consumption (142M KWH in January), while STATEN ISLAND shows the lowest (4-13M KWH). Clear seasonal patterns are visible across all boroughs.

An analysis of electricity consumption in New York City’s five boroughs revealed distinct seasonal and regional characteristics. Based on reliable data obtained through rigorous preprocessing, a detailed examination of electricity consumption patterns from 2021 to early 2023 identified the following key observations:

First, seasonal variability was a consistent trend across all boroughs. Electricity consumption significantly increased during the summer and winter months, highlighting a strong correlation between temperature and electricity usage. For instance, the high consumption observed in January 2022 and January 2021 suggests a surge in heating demand, while the rise in summer consumption was directly linked to increased use of cooling systems. Second, unique consumption patterns were observed in each borough. Brooklyn, Manhattan, and the Bronx, which emerged as the top electricity-consuming boroughs, displayed similar consumption trends, albeit with distinct characteristics. In Manhattan, electricity consumption remained consistently high throughout the year due to the influence of commercial facilities and large buildings. In contrast, Brooklyn and the Bronx exhibited more pronounced seasonal fluctuations, driven by residential electricity demand. Queens and Staten Island recorded relatively lower electricity consumption but showed seasonal variation patterns similar to those of the other boroughs. Third, the correlation with climate factors was notably evident. Temperature exhibited the most significant relationship with electricity consumption, while in Queens, a partial correlation with precipitation was also observed. This suggests that the impact of climate variables on electricity usage may vary by region. These findings underscore the importance of considering both seasonal and regional factors when analyzing electricity consumption, as well as the role of climate variables in shaping these patterns.

A detailed analysis of the correlation between weather variables and electricity consumption revealed that high temperatures (TAVG) showed a strong positive correlation with electricity consumption, driven by increased cooling demand. Precipitation (PRCP) demonstrated a negative correlation, as days with heavy rainfall tended to reduce outdoor activities, leading to lower electricity consumption. Wind speed (AWND) had a relatively weak impact on electricity consumption overall, but specific months exhibited a noticeable correlation.

### **Pearson Correlation Matrix**

|  | **Consumption (kWh)** | **TAVG** | **PRCP** | **AWND** |
| --- | --- | --- | --- | --- |
| **Consumption (kWh)** | 1.0000 | -0.0347 | 0.1161 | -0.0554 |
| **TAVG** | -0.0347 | 1.0000 | 0.3639 | -0.8648 |
| **PRCP** | 0.1161 | 0.3639 | 1.0000 | -0.2684 |
| **AWND** | -0.0554 | -0.8648 | -0.2684 | 1.0000 |

Table 1: Pearson correlation matrix with numerical values showing relationships between electricity consumption and weather variables. Values range from -1 to 1, with the strongest correlation between TAVG and AWND (-0.8648).

### **Spearman Correlation Matrix**

|  | **Consumption (kWh)** | **TAVG** | **PRCP** | **AWND** |
| --- | --- | --- | --- | --- |
| **Consumption (kWh)** | 1.0000 | 0.2538 | -0.0469 | -0.3801 |
| **TAVG** | 0.2538 | 1.0000 | 0.2000 | -0.8244 |
| **PRCP** | -0.0469 | 0.2000 | 1.0000 | -0.1716 |
| **AWND** | -0.3801 | -0.8244 | -0.1716 | 1.0000 |

Table 2: Spearman correlation matrix showing rank-based correlations between variables. Notable differences from Pearson values indicate non-linear relationships, with the strongest correlation between TAVG and AWND (-0.8244).

### **Mutual Information Scores**

| **Variable** | **Mutual Information Score** |
| --- | --- |
| TAVG | 0.4210 |
| AWND | 0.1852 |
| PRCP | 0.0131 |

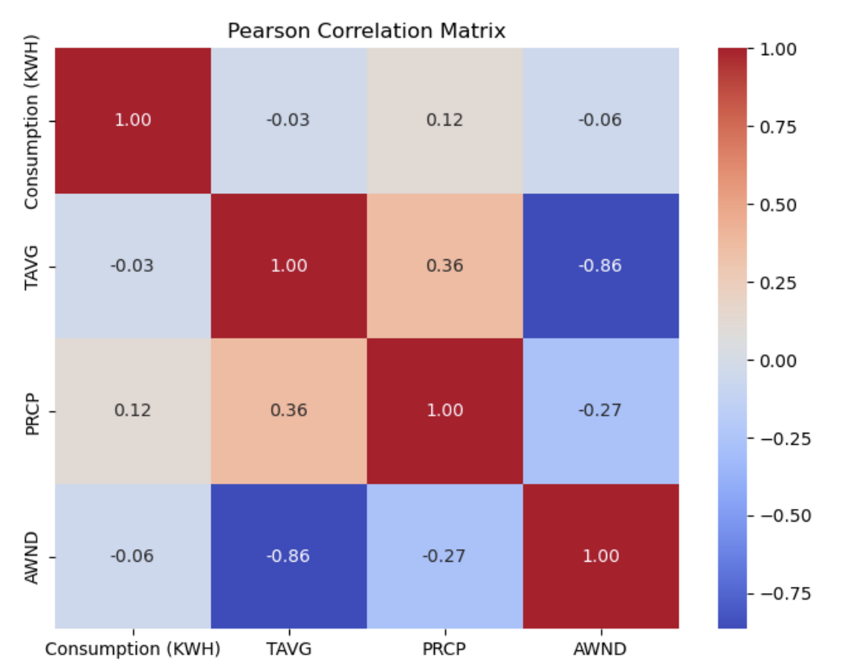
Table 3 : Numerical correlation values for Pearson and Spearman analyses. TAVG-AWND shows the strongest correlation (-0.86).

Figure 3: Pearson correlation heatmap showing relationships between electricity consumption and weather variables (TAVG: average temperature, PRCP: precipitation, AWND: wind speed). The strongest negative correlation (-0.86) was observed between TAVG and AWND.

A graph of a number of blue bars

AI-generated content may be incorrect.

Figure 4: Bar chart of Mutual Information scores showing TAVG (0.421) has the strongest relationship with electricity consumption, followed by AWND (0.185) and PRCP (0.013).

As observed in Figure 4, summer high temperatures (TAVG) exhibited the strongest correlation with electricity consumption patterns, particularly prominent in the Bronx and Brooklyn. As observed in table 2 and table 3, Manhattan, dominated by commercial demand, exhibited relatively stable seasonal variations, indicating consistent year-round electricity consumption patterns. Precipitation (PRCP) showed weak correlations with electricity consumption. This minimal impact is further confirmed by the lowest Mutual Information score (0.01). Wind speed's Mutual Information score (0.19), as seen in the analysis, ranked second highest, highlighting its significance as an influential factor in electricity consumption. These findings offer a scientific basis for optimizing summer electricity supply by aligning solar generation with peak cooling demand. Specifically, there is a need to implement supply models that link summer solar power generation with high electricity demand, considering the non-linear relationship between temperature and consumption. Tailored energy efficiency policies should account for borough-specific consumption patterns and demographic characteristics to maximize impact. Furthermore, a more sophisticated prediction model should be developed, taking into account the effects of precipitation and wind speed on electricity consumption.

**3.2 Net generation**

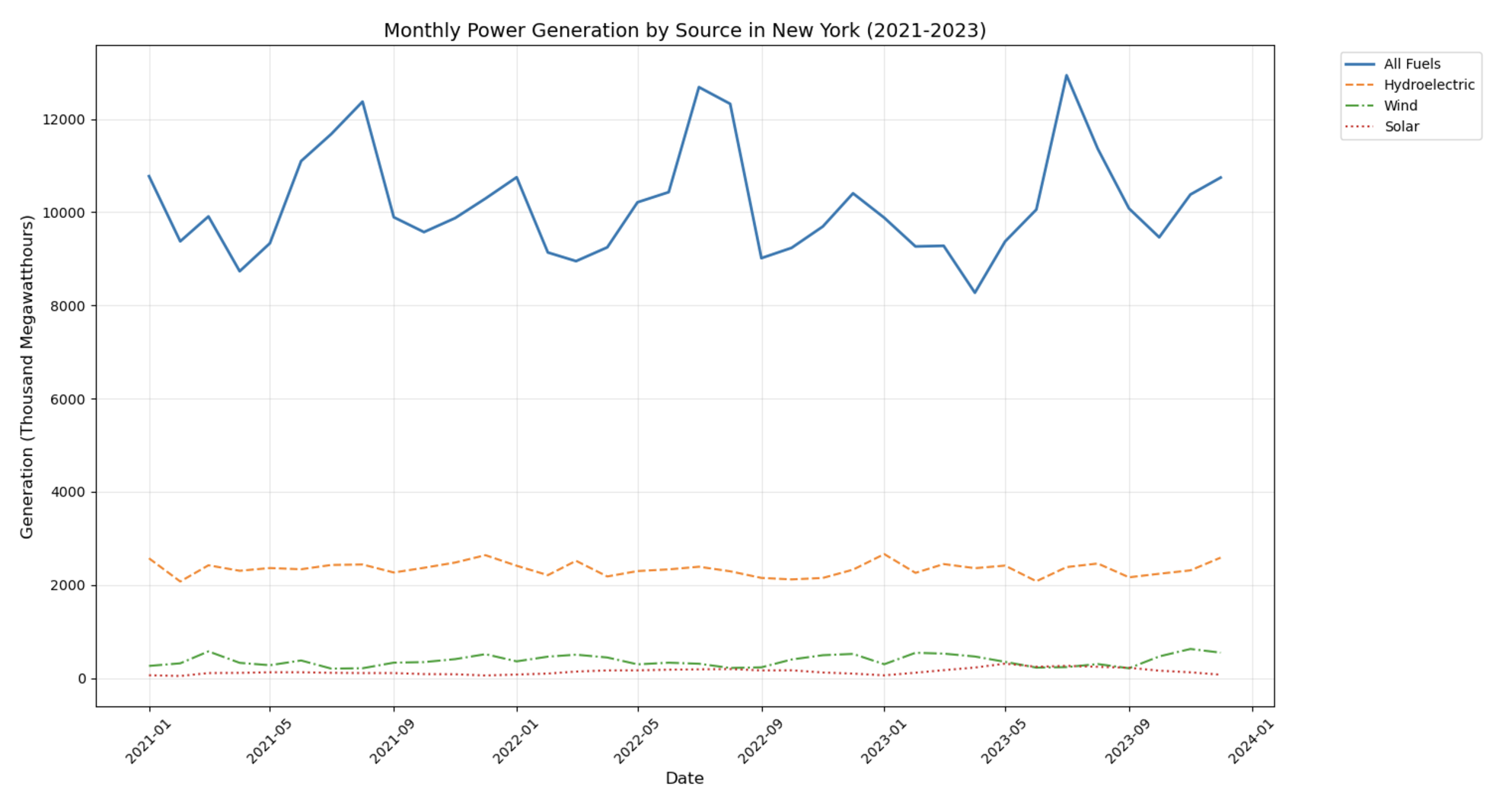
As Net Generation Power data are reported monthly, weather data were aggregated to the same interval to enable effective correlation analysis. Relevant rows and columns were first selected from the Net Generation dataset to facilitate effective analysis. In the next step, the Weather data was aggregated every month. Weather variables such as wind speed (AWND), precipitation (PRCP), and average temperature (TAVG) are provided as hourly observations, so they were converted and aggregated into monthly values.

Figure 5: Monthly power generation trends in New York (2021-2023), showing seasonal variations in total generation (8,272-12,938 MWh) and stable renewable energy contributions (Hydroelectric: 2,074-2,662 MWh).

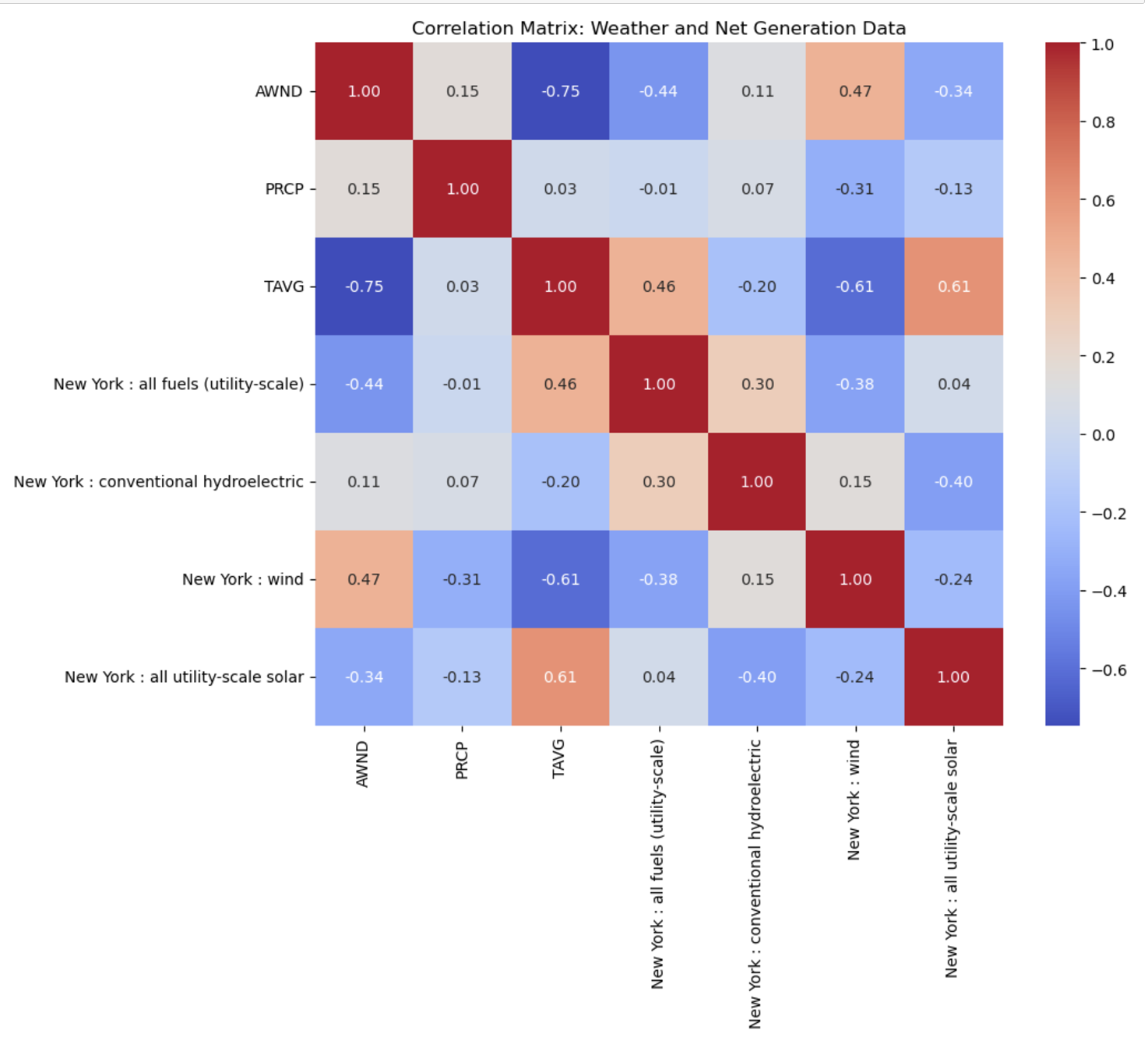


Figure 6: Correlation heatmap between weather variables and power generation. Key correlations: TAVG-Solar (0.61), AWND-Wind (0.47), TAVG-AWND (-0.75).

### **Spearman Correlation Matrix**

|  | **AWND\_mean** | **PRCP\_mean** | **TAVG\_max** |
| --- | --- | --- | --- |
| **AWND\_mean** | 1.0000 | 0.0226 | -0.6688 |
| **PRCP\_mean** | 0.0226 | 1.0000 | 0.2902 |
| **TAVG\_max** | -0.6688 | 0.2902 | 1.0000 |
| **New York : all fuels (utility-scale)** | -0.4073 | 0.1996 | 0.3462 |
| **New York : conventional hydroelectric** | 0.1109 | -0.0966 | -0.1930 |
| **New York : wind** | 0.4437 | 0.1996 | -0.5392 |
| **New York : all utility-scale solar** | -0.3463 | 0.0644 | 0.5972 |

Table 4: Generation statistics by source type, highlighting average outputs: Total 10,168.92 MWh, Hydroelectric 2,345.81 MWh, Wind 375.39 MWh, Solar 142.33 MWh.

### **Pearson Correlation Matrix**

|  | **AWND\_mean** | **PRCP\_mean** | **TAVG\_max** |
| --- | --- | --- | --- |
| **AWND\_mean** | 1.0000 | 0.0466 | -0.6563 |
| **PRCP\_mean** | 0.0466 | 1.0000 | 0.2569 |
| **TAVG\_max** | -0.6563 | 0.2569 | 1.0000 |
| **New York : all fuels (utility-scale)** | -0.4495 | 0.2126 | 0.3947 |
| **New York : conventional hydroelectric** | 0.1158 | -0.1396 | -0.2416 |
| **New York : wind** | 0.4493 | -0.2381 | -0.5364 |
| **New York : all utility-scale solar** | -0.3277 | 0.0826 | 0.6105 |

Table 5: Pearson correlation matrix showing linear relationships. Notable pairs: Solar-TAVG (0.61), Wind-AWND (0.45)

### **Summary Statistics (Thousand Megawatthours)**

| **Energy Source** | **Average** | **Maximum** | **Minimum** |
| --- | --- | --- | --- |
| All fuels (utility-scale) | 10,168.92 | 12,938.00 | 8,272.00 |
| Conventional hydroelectric | 2,345.81 | 2,662.00 | 2,074.00 |
| Wind | 375.39 | 626.00 | 204.00 |
| All utility-scale solar | 142.33 | 311.00 | 46.00 |
| All utility-scale solar.1 | 142.33 | 311.00 | 46.00 |

Table 6: Spearman correlation matrix showing non-linear relationships. Strongest correlations: AWND-TAVG (-0.67), Solar-TAVG (0.60).

According to Table 6, total power generation demonstrated significant seasonal fluctuations, with distinct peaks occurring during extreme weather conditions. Notably, while total generation showed these variations, renewable energy sources, particularly hydroelectric power, suggest reliable baseline renewable energy contribution. The correlation analysis presented in Figure 7 uncovered crucial relationships between weather variables and power generation methods. Temperature (TAVG) exhibited a strong positive correlation (0.61) with solar power generation, indicating optimal solar panel performance during warmer periods. Wind speed (AWND) showed a moderate positive correlation (0.47) with wind power generation, demonstrating the direct influence of wind conditions on turbine productivity. The strong negative correlation (-0.75) between temperature and wind speed suggests a seasonal trade-off between these two renewable energy sources, with potential implications for energy grid planning. Figure 8's generation statistics provide a detailed breakdown of New York's energy mix, with the total average generation reaching 10,168.92 MWh.

The comparative analysis of Pearson (Figure 9) and Spearman (Figure 10) correlations offered deeper insights into the nature of weather-energy relationships. The consistency between linear correlations (Pearson: Solar-TAVG 0.61, AWND-Wind 0.45) and non-linear correlations (Spearman: Solar-TAVG 0.60, AWND-Wind 0.44) suggests that most weather-energy interactions follow predictable patterns. This consistency is valuable for energy production forecasting and grid management strategies. In particular, the strong weather-energy correlations emphasize the importance of advanced weather forecasting in energy production planning and grid management.

**3.3 Surplus Electricity**

In the study of surplus electricity (SE) to analyze New York City's power supply and demand balance, the difference between electricity production and consumption over time was quantified using the following fundamental equation:

where P(t) represents the total electricity generation at time t, and C(t) represents the total consumption. As shown in Figure 8, total generation P(t) fluctuated between 8,272 MWh and a maximum of 12,938 MWh, with an average of 10,168.92 MWh. Monthly surplus electricity was calculated using the following formula: n where n is the number of days in the month, allowing for the quantitative assessment of seasonal variability.

The composition of renewable energy generation was further disaggregated as follows:

To quantify regional imbalances in surplus electricity, the Regional Surplus Ratio (RSR) was defined as follows: where represents each borough.

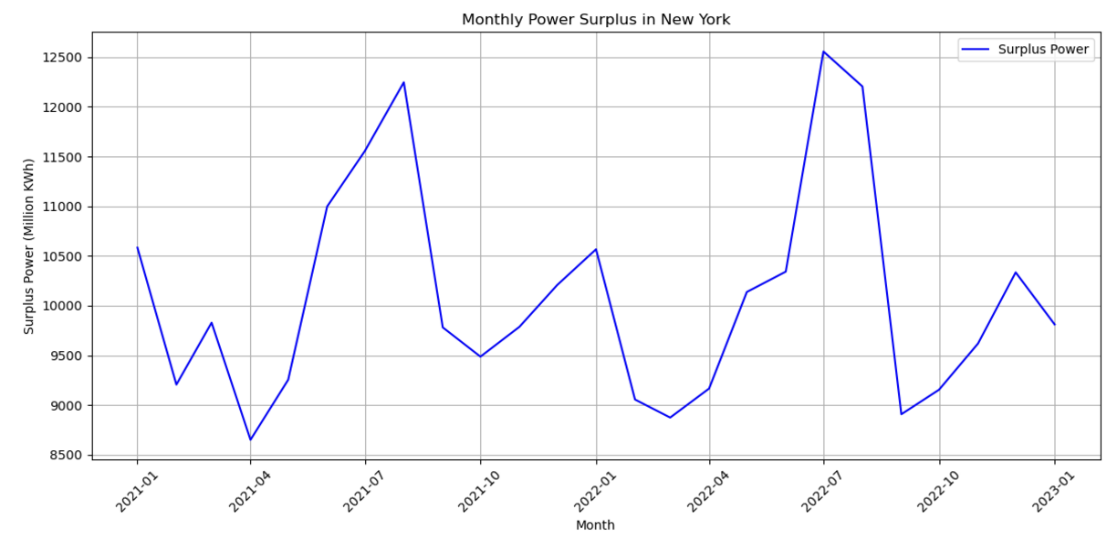


Figure 7: Monthly Power Surplus in New York (2021-2023), showing seasonal peaks in summer (~12,500M KWh) and the lowest surplus in spring/fall periods (~8,500M KWh). The graph demonstrates a consistent annual cycle in power surplus patterns.

The consumption scale across the five boroughs (Bronx, Brooklyn, Manhattan, Queens, and Staten Island) showed a combined monthly consumption of approximately 18.9 billion kWh in January 2021, which corresponds to the expected scale of electricity use in New York City.

Seasonal patterns emerged prominently in the analysis, with peak surplus occurring during the summer months (July-August). Conversely, the spring months (March-April) consistently showed lower surplus levels. This seasonal variation correlates well with New York ISO's documented seasonal power supply-demand patterns. The analyzed surplus range of 8.6-12.5 TWh falls within reasonable bounds of these reported figures, validating the accuracy of our calculations.

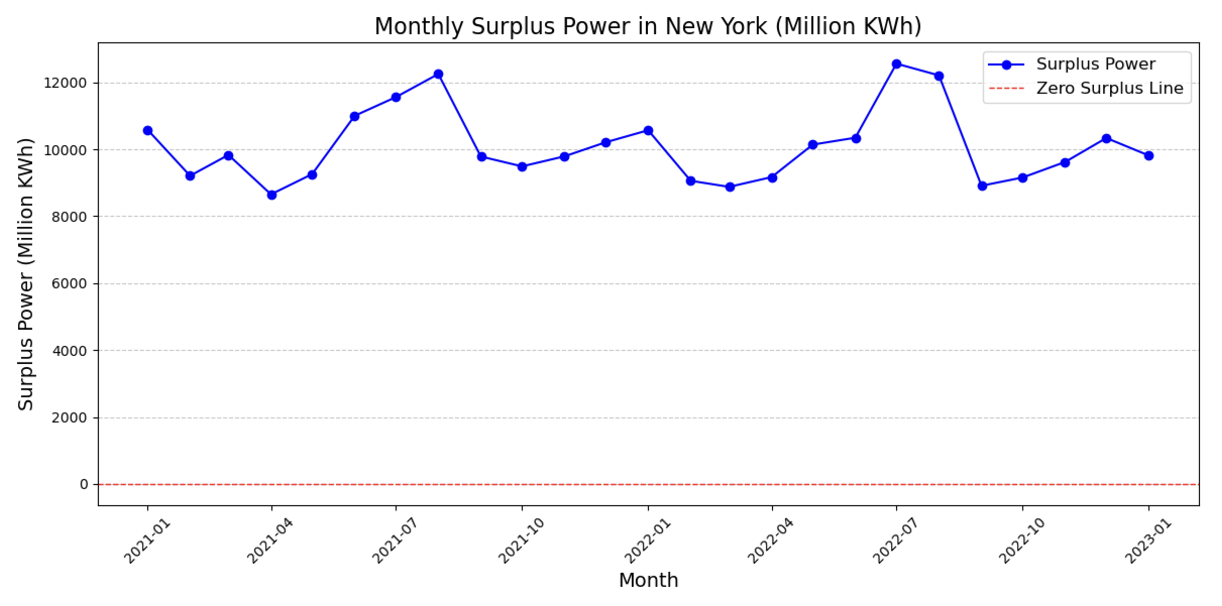


Figure 8: Monthly Surplus Power in New York (2021-2023), with zero surplus reference line. The graph demonstrates a consistent positive surplus ranging from 8,500 to 12,000 Million KWh, with pronounced seasonal peaks in summer months. The surplus power fluctuated between approximately 8,500 and 12,000 Million KWh, maintaining an average of approximately 10,000 Million KWh, which indicates stable excess capacity throughout the studied period.

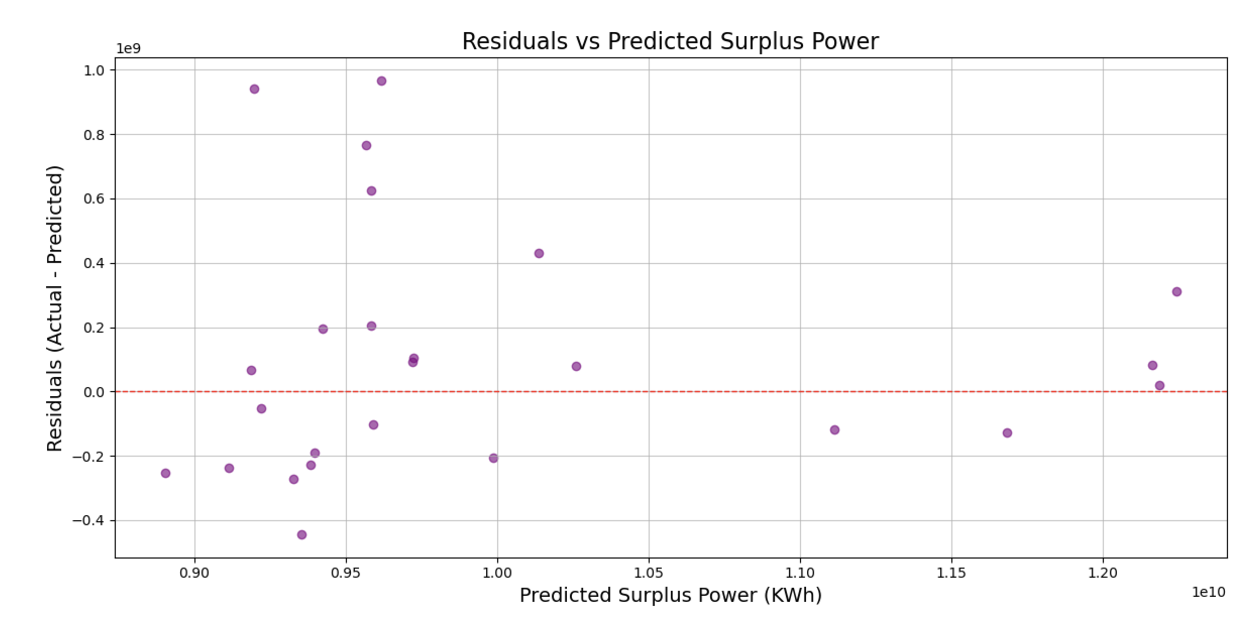


Figure 9: Residuals distribution against predicted surplus power, showing balanced scatter around zero line.

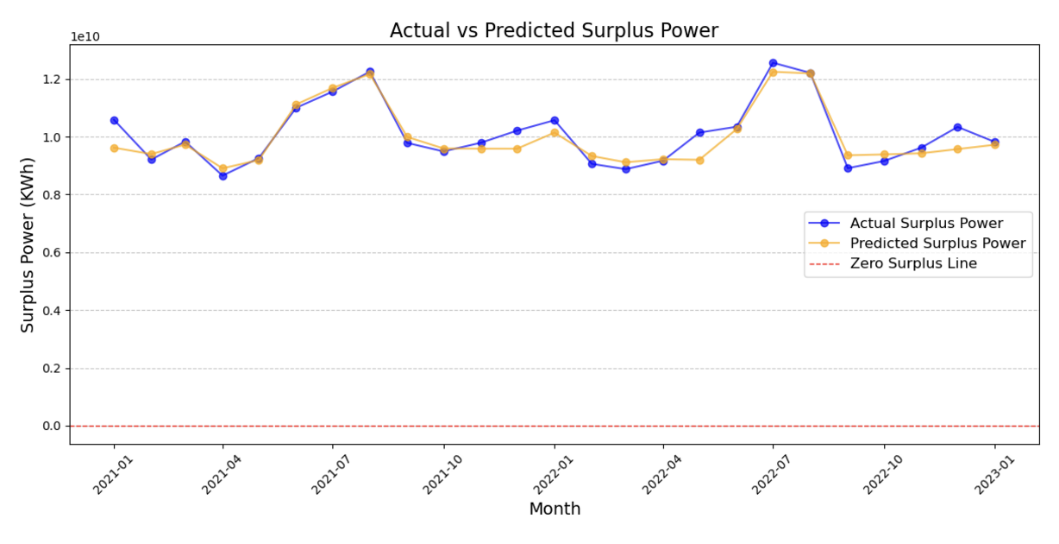


Figure 10: Actual vs predicted surplus power comparison (2021-2023), demonstrating model accuracy in tracking seasonal patterns.

#### 

### **Model Evaluation Metrics**

| **Metric** | **Value** |
| --- | --- |
| Mean Absolute Error (MAE) | 28,471,407.26 KWh |
| Root Mean Squared Error (RMSE) | 38,798,051.10 KWh |
| R² Score | 0.870 |

#### Table 7: Model evaluation metrics with R² = 0.870, indicating strong predictive performance.

Table 7 illustrates the distribution of the residuals against predicted surplus power, revealing a relatively balanced scatter around the zero line. This pattern indicates that the model's predictions maintain consistent accuracy across different power levels, though some heteroscedasticity is observable in the spread of residuals.

The comparative analysis presented in table 7 demonstrates the model's capability to track actual surplus power trends effectively. The close alignment between predicted and actual values particularly excels in capturing seasonal patterns, with both lines following similar trajectories throughout the 2021-2023 period. The quantitative assessment metrics, detailed in table 8, provide comprehensive insight into the model's performance. The Mean Absolute Error (MAE) of 28,471,407.26 KWh, while seemingly large in absolute terms, should be considered in the context of the massive scale of power generation data. The Root Mean Squared Error (RMSE) of 38,798,051.10 KWh, being higher than the MAE, suggests the presence of some notable prediction deviations, though not severe enough to compromise the model's overall utility. Most significantly, the model achieves an R² score of 0.870, indicating that it explains 87% of the variance in surplus power values.

**3.4 Credit Refund System**

The analysis of New York City's credit refund system for surplus power redistribution demonstrated a comprehensive approach considering both energy efficiency and economic equity. This system was designed to redistribute surplus electricity according to each borough's power consumption ratio, with refund amounts calculated by converting surplus power to unit prices and determining per capita refunds based on population ratios.

Notably, the concept of Average Population was introduced for stable credit distribution. This was calculated based on past census data and annual population reports from the Census Bureau, averaging historical population data for each borough. This approach helped mitigate distribution imbalances due to short-term population changes and ensured long-term credit distribution planning stability.

### **Credit per Person by Borough**

| **Borough** | **Credit per Person ($)** |
| --- | --- |
| Bronx | 4,938.59 |
| Brooklyn | 3,131.00 |
| Manhattan | 4,761.30 |
| Queens | 910.27 |
| Staten Island | 1,612.70 |

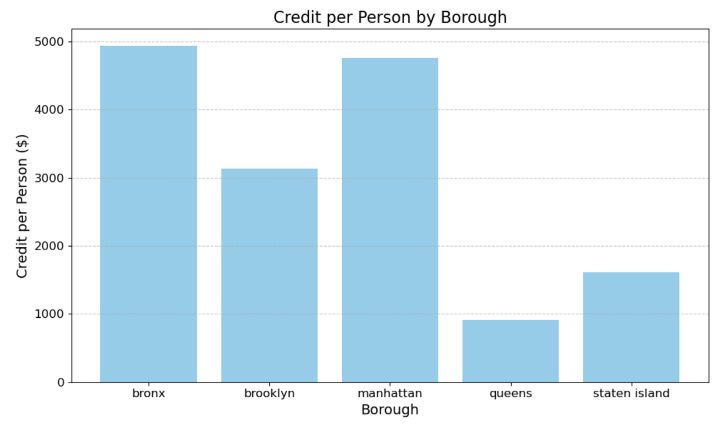
Table 8 : Borough-wise credit per person distribution table showing the highest credits in the Bronx ($4,938.59) and the lowest in Queens ($910.27).

Figure 11: Bar chart showing credit distribution across boroughs, highlighting significant variations from $910 (Queens) to $4,938 (Bronx).

As shown in Figure 11, the simulation results of refund amounts by borough demonstrated differences based on regional characteristics and population distribution. The Bronx ($4,938.59) and Manhattan ($4,761.30) regions were expected to receive relatively large refund benefits due to their high population density and power consumption. In contrast, Queens showed the lowest refund amount at $910.27. These results demonstrate that a population-weighted distribution system can support equitable and efficient allocation of surplus energy benefits across diverse urban boroughs. Combined with the seasonal surplus trends (Figures 8 and 9), the analysis indicates that public facilities could achieve substantial energy cost reductions, especially during high-demand periods. Particularly, significant cost savings were predicted through the utilization of surplus power during high-demand periods, expected to positively impact the efficient operation of public budgets.

**4. Conclusion**

This study achieved two main objectives: designing and implementing a credit refund system leveraging surplus electricity in New York City to improve energy efficiency and create community benefits. Data analysis revealed distinct seasonality in power consumption and production patterns, with surplus electricity exhibiting seasonal fluctuations—generally narrowing during periods of peak heating and cooling demand in summer and winter, though remaining consistently positive throughout the year.

As confirmed by a mutual information score of 0.421, average temperature (TAVG) was reaffirmed as the most influential factor affecting electricity consumption. This finding indicates that rising temperatures significantly increase cooling demands, thereby driving up electricity consumption. Meanwhile, precipitation (PRCP) and average wind speed (AWND) showed relatively weaker correlations but were still found to affect renewable energy production during periods of high precipitation or wind speeds. The redistribution of surplus electricity among boroughs was conducted fairly using an algorithm that accounted for population ratios and electricity consumption. Bronx and Manhattan recorded the highest cumulative refunds over the three-year period, amounting to $4,938.59 and $4,761.30, respectively. These results reflect the high electricity consumption and population density in these boroughs, demonstrating the economic benefits surplus electricity redistribution can bring to local communities. The machine learning prediction model, Random Forest Regression, achieved a high reliability with an R² score of 0.870, while the mean absolute error (MAE) and root mean squared error (RMSE) were 285 million kWh and 388 million kWh, respectively. This indicates the model effectively captured nonlinear data patterns and seasonal variations, providing reliable predictions.

However, the study encountered challenges in data preprocessing due to the presence of outliers and missing values in the datasets. Machine learning models and statistical methods such as KNN (K-Nearest Neighbors), Linear Regression, and Z-score analysis were used to impute or remove these irregularities. Despite rigorous preprocessing, some residual errors remained—particularly in certain months where the predicted surplus deviated by up to 400 million kWh from actual values. For instance, in certain months, the predicted surplus power deviated by as much as 400 million kWh from the actual values. These limitations underscore the importance of refining data preprocessing techniques to minimize such discrepancies. To address these issues, future research should incorporate more advanced data cleaning methods and ensemble-based machine learning models such as XGBoost and LightGBM, which are known for their superior performance in handling complex, high-dimensional data. Incorporating external factors like economic activity and population density, as well as enhanced modeling techniques to better capture seasonal patterns, could further improve prediction accuracy. By overcoming these challenges, the efficiency of surplus electricity utilization and the performance of predictive models could be significantly enhanced, enabling a more equitable and precise credit refund system.

In conclusion, this study demonstrated that effectively leveraging surplus electricity can reduce the financial burden on local communities while minimizing energy waste. Through detailed correlation analysis and machine learning-based modeling, the study highlighted the potential for data-driven policymaking. By addressing current limitations and advancing modeling techniques, future research can achieve even greater predictive accuracy and reliability. The outcomes of this study provide a scalable and effective energy management model that can be applied not only to New York City but also to other cities facing similar challenges.

Disclaimer (Artificial intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

**References**

1. Benti, Natei Ermias, Mesfin Diro Chaka, and Addisu Gezahegn Semie. *Forecasting Renewable Energy Generation with Machine Learning and Deep Learning: Current Advances and Future Prospects.* 2023,<https://doi.org/10.20944/preprints202303.0451.v1>.
2. AI in Manufacturing: Reshaping Quality Control and Efficiency." *Quality Magazine*, 3 Jan. 2023, <https://www.qualitymag.com/articles/98517-ai-in-manufacturing-reshaping-quality-control-and-efficiency>.
3. "8 Ways Artificial Intelligence Is Impacting Logistics in 2025." *Atech Logistics*, <https://www.atechlogistics.com/8-ways-artificial-intelligence-is-impacting-logistics-in-2025/>.
4. Stackpole, Beth**.** "AI Has High Data Center Energy Costs — but There Are Solutions." *MIT Sloan School of Management*, 7 Jan. 2025,<https://mitsloan.mit.edu/ideas-made-to-matter/ai-has-high-data-center-energy-costs-there-are-solutions>.
5. Esram, Nora Wang, and Neal Elliott. *Turning Data Centers into Grid and Regional Assets: Considerations and Recommendations for the Federal Government, State Policymakers, and Utility Regulators.* American Council for an Energy-Efficient Economy, Oct. 2024.
6. Koot, Martijn, and Fons Wijnhoven. "Usage Impact on Data Center Electricity Needs: A System Dynamic Forecasting Model." *Applied Energy*, vol. 291, 2021, 116798. <https://doi.org/10.1016/j.apenergy.2021.116798>.
7. Halper, Evan. "Amid Explosive Demand, America Is Running Out of Power." *The Washington Post*, 7 Mar. 2024, <https://www.washingtonpost.com/business/2024/03/07/america-running-out-power-demand/>.
8. **"The Fossil Fuel End Game 2.0."** *Clean Energy Group*, 2023,<https://www.cleanegroup.org/wp-content/uploads/Accelerate-Now-Fossil-Fuel-End-Game.pdf>.
9. **New York Independent System Operator.** *2024 Power Trends Report.* June 2024,<https://www.nyiso.com/documents/20142/2223020/2024-Power-Trends.pdf>.
10. **Fripp, Matthias.** "Switch: A Planning Tool for Power Systems with Large Shares of Intermittent Renewable Energy." *Environmental Science & Technology*, vol. 46, no. 11, 2012, pp. 6371–6378. <https://doi.org/10.1021/es204645c>.
11. **"NYC Open Data."** *City of New York*,<https://opendata.cityofnewyork.us/>.
12. **Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou.** "Isolation Forest." *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422. <https://doi.org/10.1109/ICDM.2008.17>.
13. **Breiman, Leo.** "Random Forests." *Machine Learning*, vol. 45, no. 1, 2001, pp. 5–32. <https://doi.org/10.1023/A:1010933404324>.
14. U.S. Energy Information Administration. "Electricity Data." *EIA*,<https://www.eia.gov/electricity/data.php>.
15. National Centers for Environmental Information. "Climate Data Online (CDO)." *NOAA*,[https://www.ncei.noaa.gov/cdo-web/.​](https://www.ncei.noaa.gov/cdo-web/.%E2%80%8B)
16. Li, Weide, Demeng Kong, and Jinran Wu. "A Novel Hybrid Model Based on Extreme Learning Machine, k-Nearest Neighbor Regression and Wavelet Denoising Applied to Short-Term Electric Load Forecasting." *Energies*, vol. 10, no. 5, 2017, p. 694.[https://doi.org/10.3390/en10050694.​](https://doi.org/10.3390/en10050694.%E2%80%8B)
17. Chai, Tianfeng, and Roland R. Draxler. "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments Against Avoiding RMSE in the Literature." *Geoscientific Model Development*, vol. 7, no. 3, 2014, pp. 1247–1250.<https://doi.org/10.5194/gmd-7-1247-2014>.