Electricity Forecasting and Optimization with IEA Data: A Renewable Energy Strategy

1. Olajubu, Emmanuel
2. Ikhagbode, Sarah Irema Inuaeyen,
3. Nsikakabasi Philip
4. Nematov, Abdulkhodiy

Prepared by: Project Group 6

Course: DAMO 611 CASE STUDY 3

Instructor: Zeeshan Ahmad

CHAPTER ONE

PROBLEM DEFINITION AND RESARCH QUESTION

1.0 Problem Definition

Renewable energy continues to dominate the global conversation around climate action, energy independence, and sustainable development. Yet despite international progress, the production of renewable electricity remains inconsistent across countries, energy types, and economic tiers. Decision-makers need deeper analytical insights into production trends to inform policies, investments, and operational strategies.

This project addresses the gap in visibility and comparability of renewable electricity production, specifically focusing on solar, wind, and hydro power. Using a robust international dataset from the International Energy Agency (IEA), we aim to examine how renewable energy production has evolved globally from 2010 to 2025 and what factors (such as GDP and population) correlate with production levels.

1.1 Research Question

To what extent can monthly electricity consumption (Final Consumption) be predicted using country-level economic indicators, population size, renewable and fossil fuel energy *production, and time-based features between 2010 and 2025 and how can this inform optimal energy allocation policies?*

1.2 Statement of Purpose

This research seeks to analyze trends in renewable electricity generation using IEA's monthly country-level data across 49 countries from 2010–2025. Supplemented with GDP and population data from the United Nations, the goal is to model growth patterns, identify outliers, and forecast

renewable production dynamics. Ultimately, the project supports data-driven decision-making in energy policy and sustainability planning.

1.3 Justificatioin of the Research Question

The ability to predict national electricity consumption is crucial for energy planners, policymakers, and investors. As energy demand fluctuates seasonally and is influenced by production capacity, economic activity, and demographic growth, modeling consumption enables proactive infrastructure planning and market interventions.

According to the IEA (2024), aligning electricity demand with clean energy supply is central to global net-zero goals, yet demand forecasting remains a challenge in many developing and transitioning economies. This project contributes to that gap by offering a data-driven, multi-featured model to anticipate consumption patterns across diverse countries

.

CHAPTER TWO

HYPOTHESIS DEVELOPMENT

The formulation of hypotheses is essential in conducting a structured and insightful analysis. This chapter aims to build on the research question posed in Chapter 1 and present clear, testable hypotheses that align with the data analytics objectives for the MES dataset.

2.0 Reseacrh Question Recap

To what extent can monthly electricity consumption (Final Consumption) be predicted using country-level economic indicators, population size, renewable and fossil fuel energy production, and time-based features between 2010 and 2025 and how can this inform optimal energy allocation policies?

2.1 Hypotheses Development

This study involves a supervised regression modeling task. To evaluate the predictive power of selected features (GDP, population, renewable production, and time-based variables), the following hypotheses are formulated:

2.1.1 Null Hypothesis (H0)

H0: There is no statistically significant relationship between a country's electricity consumption (Final Consumption (Calculated)) and its GDP, population, renewable energy production, or time-based variables.

This hypothesis assumes that electricity consumption is independent of these factors and that any apparent associations are coincidental or statistically insignificant.

2.1.2 Alternative Hypothesis (H1)

H1: Monthly electricity consumption (Final Consumption (Calculated)) can be significantly predicted using GDP, population, energy production (e.g., solar, wind, natural gas), and time-based variables (month/year).

This hypothesis asserts that a combination of these features can be used to accurately forecast consumption levels and supports the development of a predictive regression model.

2.2 Specific Hypotheses

To guide the analysis, the research question is divided into specific hypotheses based on key predictors:

Hypothesis on Economic Influence:

1. H0a: GDP does not significantly influence a country's final consumption.
2. H1a: Countries with higher GDP levels exhibit significantly higher electricity consumption.

Rationale: GDP reflects industrial and economic activity, which typically correlates with industrial and commercial energy use.

Hypothesis on Population Size:

1. H0b: Population size does not significantly influence final consumption.
2. H1b: Countries with larger populations exhibit higher electricity consumption, particularly in residential demand.

Rationale**:** Population is a proxy for residential demand and total infrastructure needs.

Hypothesis on Renewable Energy Production

1. H0c: Renewable electricity production (solar, wind, hydro) does not significantly influence a country's electricity consumption.

2. H1c: Renewable energy production contributes significantly to predicting monthly electricity consumption.

Rationale**:** Nations that produce more electricity locally, especially from renewables, may consume more domestically and rely less on imports.

Hypothesis on Time-Based Features

1. H0d: Month and year do not affect electricity consumption
2. H1d: Time-based features (month, year) significantly influence electricity consumption due to seasonality and long-term trends.

Rationale: Electricity consumption often shows seasonal variation, especially in climates with strong heating or cooling needs.

2.3 Hypothesis Justification

The hypotheses were selected to support the construction of a structured and interpretable predictive model. They are:

1. Specific: Each hypothesis targets a single, measurable predictor of electricity consumption.
2. Testable: The dataset contains quantifiable variables to statistically validate or reject each hypothesis using machine learning techniques (e.g., linear regression, Random Forest, XGBoost).
3. Relevant: The hypotheses align with global energy concerns and the need for improved forecasting capabilities.

4. Grounded in Literature: According to the International Energy Agency (2024), accurate consumption forecasting is critical for renewable integration, energy security, and investment planning in a rapidly evolving energy landscape.

CHAPTER THREE

DATA UNDERSTANDING AND DATA COLLECTION

3.0 Overview of the MES Dataset

The MES_0225 dataset was obtained from the International Energy Agency's *Monthly Electricity Statistics* (IEA, 2025) and is used in accordance with their *Terms of Use for Non-CC Material*., includes over 149,000 records across 49 countries and spans from January 2010 to February 2025. Each row represents a single observation of an electricity metric such as production, losses, or consumption, differentiated by the Balance and Product columns.

3.1 Dataset Structure and Key Variables (Data dictionary)

Column	Description
Country	Name of the country
Balance	Describes the type of metric reported (e.g., *Net Electricity Production*, *Final Consumption (Calculated)*)
Product	Specifies the energy type (e.g., Solar, Wind, Coal, Natural Gas)
Value	Numerical value for that product and balance (in GWh)
Unit	Unit of energy — almost always GWh
Year	Extracted from Time
Month	Extracted from Time
GDP	Gross Domestic Product (merged from UN data)
Population	Total population (merged from UN data)

3.2 Data Collection Methodology

1. IEA MES Dataset: Cleaned and parsed Time into Year and Month.
2. GDP Data: Filtered to only include “Gross Domestic Product (GDP),” reshaped from wide to long format, merged on Country and Year.
3. Population Data: Combined historical estimates with 2024–2025 projections, reshaped and merged.

3.3 Python/Excel Queries and Processing

1. Used pandas.melt() for GDP and population reshaping.
2. Used pandas.merge() to join GDP/Population to MES by ['Country', 'Year'].
3. Applied filtering to retain only relevant rows:
    a. For features: Balance == "Net Electricity Production" and selected Products
    b. For target: Balance == "Final Consumption (Calculated)" and Product == "Electricity"
4. Created derived features (e.g., total renewables, total fossil fuel production) through aggregation.

3.4 Final Dataset Description (Post-Reshaping)

Row Unit	One country-month (e.g., Australia – Feb 2025)
Target	Value where Balance == Final Consumption (Calculated)
Features	Aggregated Values for Solar, Wind, Fossil Fuels, GDP, Population, Month

The final cleaned dataset is structured for modeling, with each row representing a unique (Country, Year, Month) observation and all key predictors available as features.

Appendix B helps visually justify our modeling choices in upcoming chapters and explains collinearity concerns (e.g., GDP ↔ Population = 0.92).

CHAPTER FOUR

DATA VISUALIZATION

4.0 Overview

This chapter presents a set of visualizations developed using Python and Tableau to explore patterns in electricity production and consumption across 49 countries from 2010 to 2025. The visual analysis aimed to uncover insights aligned with the core research question and four associated hypotheses. Visuals were designed to evaluate temporal trends, geographic differences, economic drivers, and energy composition dynamics. The figures referenced in this chapter are numbered and labeled (e.g., Figure 4.1) for inclusion in the main report. Additional exploratory charts are provided in Appendix A for deeper reference.

4.1 Production Trends Over Time

A line chart (Figure 4.1) plots global electricity production across different product types from 2010 to 2025. It reveals a significant upward trend in Total Renewables (Hydro, Solar, Wind, etc.), with a sharp increase observed between 2014 and 2015. This change may reflect updated reporting standards, policy shifts, or capacity expansions in renewable infrastructure. Traditional sources like Coal and Combustible Fuels show signs of plateauing in later years. This supports Hypothesis H1c, which posits that renewable energy output is a significant predictor of electricity consumption.

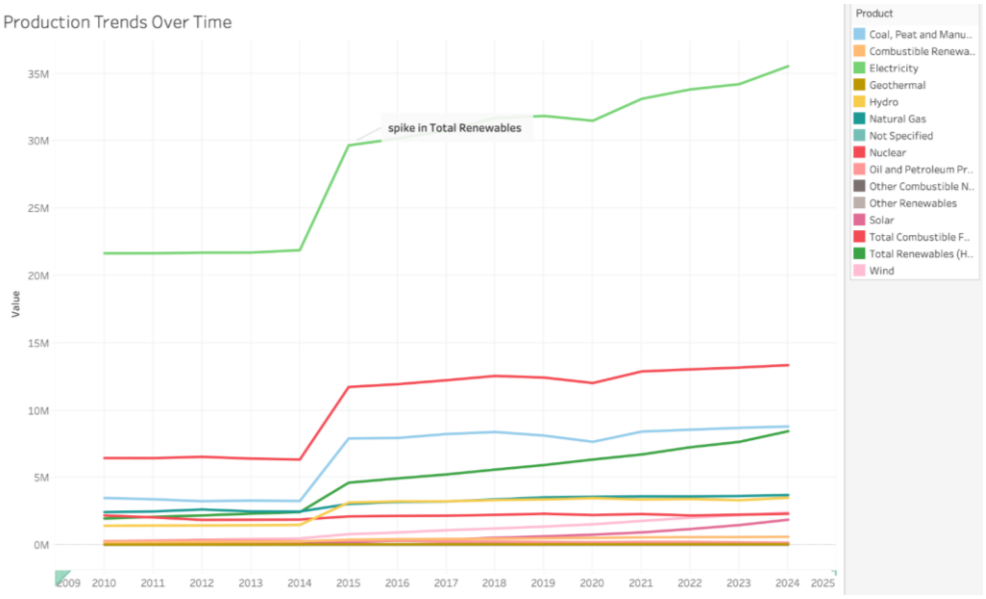

Figure 4.1 *Electricity production trends by product type from 2010 to 2025. Total Renewables show a sharp rise post-2014, highlighting the global shift toward clean energy.*

4.2 Seasonal Patterns in Production

A heatmap (Figure 4.2), normalized using Z-scores per product type, illustrates monthly production patterns across all energy categories. Renewable sources such as Hydro and Solar display seasonal variation Hydro production peaks mid-year, while Solar output increases in summer months. This directly supports Hypothesis H1d, suggesting that time-based variables (like month) impact electricity production and, by extension, consumption.

Heatmap of Monthly Prouction

	Product											
Month	Coal, Peat and Manufa..	Electricity	Geothermal	Hydro	Natural Gas	Not Specified	Nuclear	Other Renewables	Solar	Total Combustible..	Total Renewables ..	Wind
Jan	9,927,520	41,826,782	66,424	3,553,851	4,465,766	49,964	3,215,965	1,097	615,055	15,518,381	6,660,380	1,825,644
Feb	8,317,761	36,960,823	61,417	3,212,869	3,934,488	49,018	2,815,477	1,085	704,374	13,236,633	6,194,808	1,680,742
Mar	8,153,743	35,768,144	61,334	3,292,405	3,717,668	46,172	2,736,373	1,062	721,130	12,802,486	6,200,738	1,599,261
Apr	7,353,550	32,795,666	58,897	3,210,534	3,353,604	45,734	2,485,298	1,058	811,888	11,553,313	6,068,379	1,492,061
May	7,573,785	34,080,004	60,276	3,552,697	3,517,747	45,673	2,539,254	1,065	900,366	11,958,019	6,437,027	1,414,999
Jun	8,077,881	35,548,849	57,380	3,685,485	3,917,885	45,371	2,569,339	1,050	903,475	12,868,063	6,391,606	1,249,425
Jul	9,063,544	39,023,369	59,216	3,889,124	4,576,856	46,058	2,740,761	1,061	908,943	14,592,551	6,626,020	1,234,833
Aug	9,178,778	38,687,200	59,074	3,754,802	4,566,118	46,478	2,732,518	1,066	901,231	14,703,933	6,428,807	1,173,831
Sep	8,113,396	35,070,619	57,626	3,409,267	4,023,355	44,916	2,575,578	1,071	797,596	13,010,801	5,999,890	1,229,911
Oct	7,765,369	34,399,841	59,190	3,313,445	3,800,836	45,874	2,564,843	1,074	710,893	12,441,362	6,087,639	1,491,294
Nov	8,121,820	34,784,436	58,844	3,110,363	3,756,808	45,910	2,638,668	1,074	580,802	12,791,834	5,886,975	1,599,047
Dec	9,131,147	37,983,657	62,278	3,207,691	4,052,399	46,399	2,927,703	1,084	543,995	14,205,674	6,157,075	1,765,155

Figure 4.2 *Z-score normalized heatmap of monthly electricity production by source. Seasonal variation is visible, particularly in Hydro and Solar energy.*

4.3 Renewables vs Non-Renewables by Country (%)

A stacked bar chart (Figure 4.3) displays the percentage share of renewable and non-renewable energy sources in each country's electricity mix. The People's Republic of China leads in overall production but has a relatively low renewable share (~39%). The United States follows, with about 30% renewables. Countries like France, United Kingdom, and Korea rely heavily on combustible fuels. This visualization further supports Hypothesis H1c and underscores the variability in energy policy and infrastructure across countries.

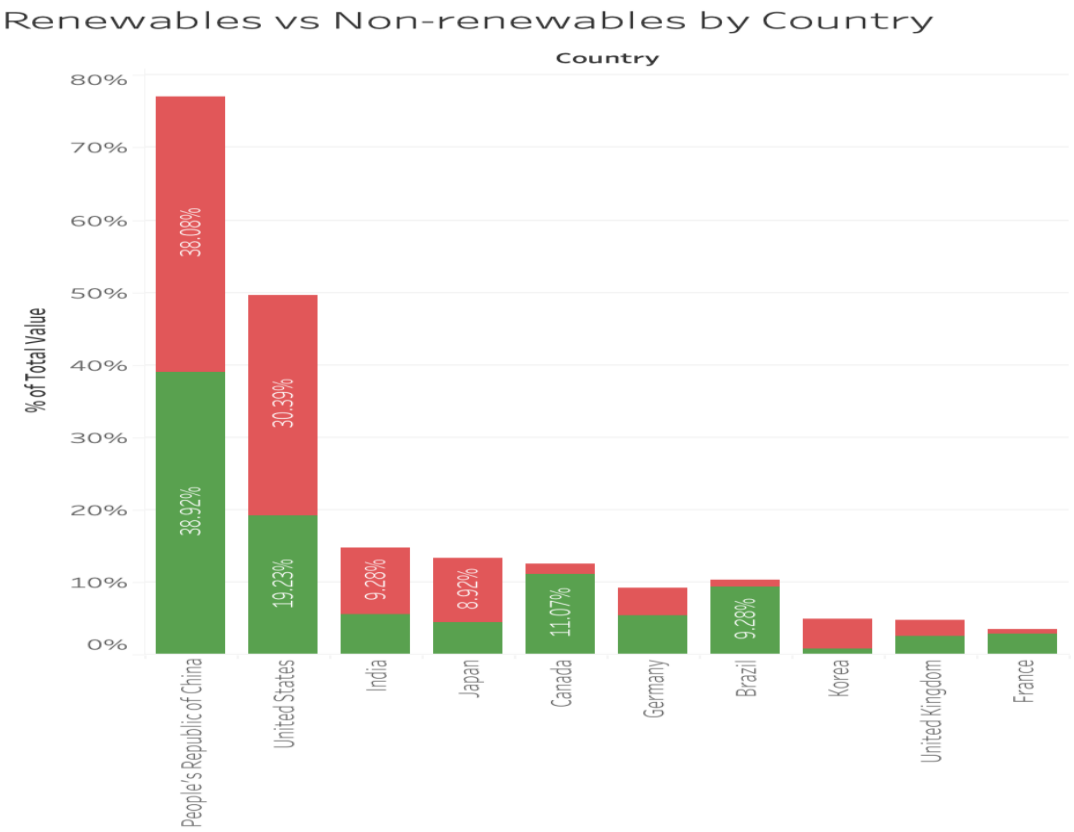

Figure 4.3 *Renewables vs non-renewables share by country (percentage of average production). Shows relative clean energy adoption across top-producing nations.*

4.4 GDP and Electricity Consumption Over Time

A dual-axis line chart (Figure 4.4) compares global GDP with global electricity consumption from 2010 to 2025. Despite temporary dips in GDP due to macroeconomic shocks (e.g., COVID-

19), electricity consumption shows a steady increase. This visualization offers strong empirical backing for Hypothesis H1a, which states that GDP significantly predicts electricity consumption.

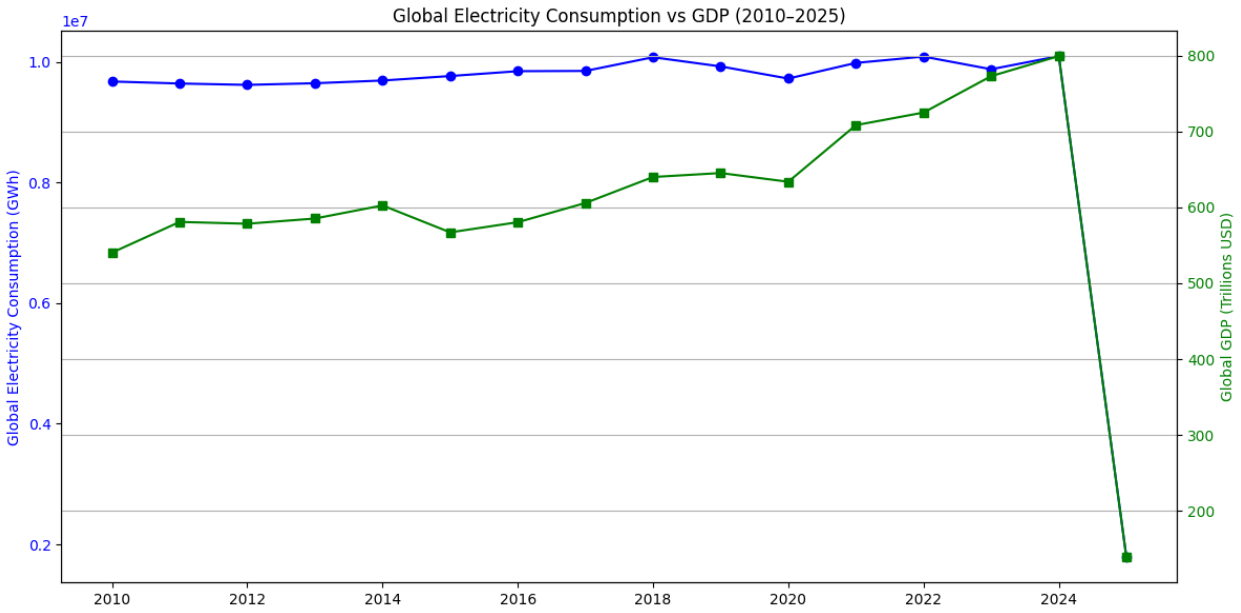

Figure 4.4 *Global electricity consumption compared with global GDP from 2010 to 2025. The dual-axis line chart suggests long-term economic and energy use correlation.*

4.5 Final Consumption vs GDP by Country (Feb 2025)

A combo chart (Figure 4.5) ranks the top 15 countries by electricity consumption and overlays their GDP using a secondary axis. The United States is the dominant consumer, followed by Japan and Canada. This figure highlights mismatches between GDP and consumption for countries like France and Italy, offering nuanced evidence supporting Hypothesis H1a.

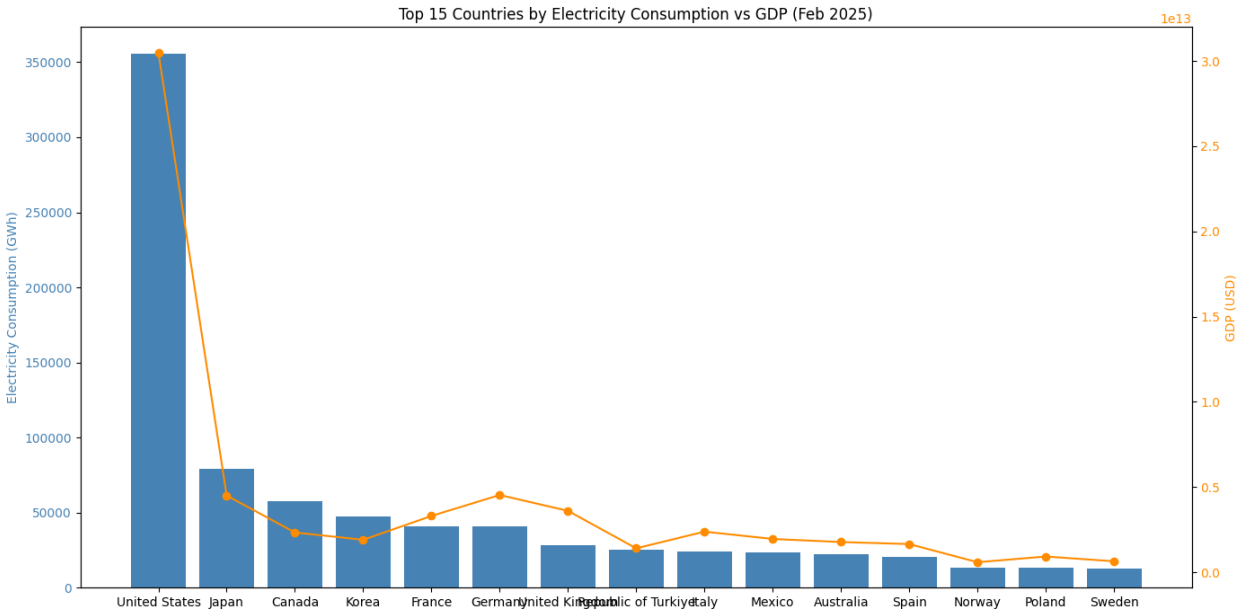

Figure 4.5 *Electricity consumption and GDP comparison for the top 15 countries in February 2025. The bar-line chart visualizes economic-electricity alignment at the country level.*

4.6 Final Consumption vs Population with GDP Coloring

A scatter plot (Figure 4.6) explores the relationship between Final Consumption and Population for 2024. Bubble colors reflect GDP size. The strong upward pattern confirms Hypothesis H1b population size significantly correlates with electricity usage. The chart also reveals that some countries with moderate populations but high GDP (e.g., Germany) have outsized energy consumption, likely due to industrial demand.

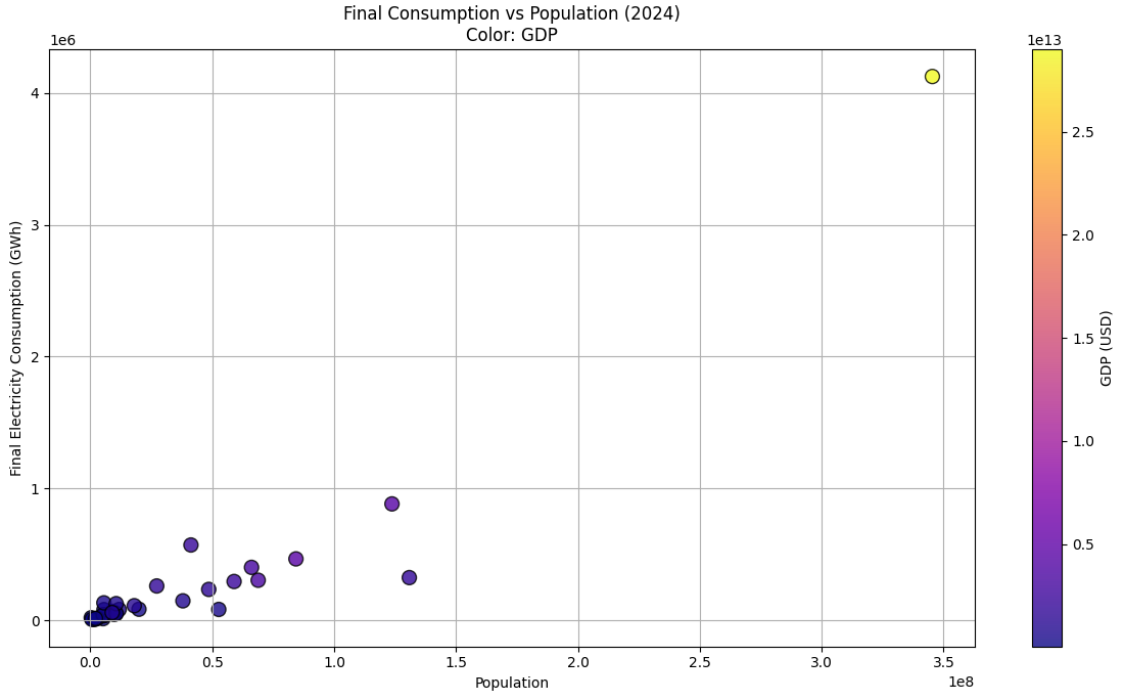

Figure 4.6 *Scatter plot of Final Consumption vs Population for 2024, with GDP as color. Indicates population is a strong predictor of electricity usage, moderated by GDP.*

4.7 Top Exporting and Importing Countries

Additional visualizations presented in Appendix A include:

Figure A: Top 15 electricity-exporting countries (Germany, France, Canada lead globally)

Figure A1: Top 15 Countries: Consumption vs GDP (Feb 2025) Country-level comparison of economic size vs energy use.

Figure A2: Top 15 electricity-importing countries (United States and Italy are major net importers)

While these figures do not directly support the primary hypotheses, they provide important market context regarding cross-border energy dependency.

4.8 Geographic Energy Distribution

Figure A4 presents a choropleth map of total average production by country, while Figure A3 visualizes the top 10 producers by energy product. Also, Due to China's limited data coverage (only "Net Electricity Production" available), it is absent from breakdowns that depend on specific product types. This was a data availability constraint that does not undermine analysis but warrants mention for transparency.

4.9 Visual Summary

Hypothesis	Covered By
H1a (GDP → Consumption)	Figures 4.4, 4.5
H1b (Population → Consumption)	Figure 4.6
H1c (Renewables → Consumption)	Figures 4.1, 4.3
H1d (Time Effect on Consumption)	Figures 4.1, 4.2

Each of the four hypotheses is visually supported, and these insights lay the groundwork for model building (Chapter Five). Charts not included in this core chapter are archived in the appendix and remain available for stakeholder review.

CHAPTER FIVE

MODEL BUILDING

5.0 Overview

This chapter outlines the construction of predictive models aimed at forecasting Final Electricity Consumption (GWh) using a multivariate regression framework. Informed by the hypotheses developed in Chapter Two, our models utilize structured, enriched data combining production statistics from the International Energy Agency (IEA), economic data from the United Nations, and time-based features.

Given the structured tabular nature of the dataset and the numerical target variable, this chapter explores a blend of linear and non-linear regression models, including:

1. Linear Regression (baseline),
2. Random Forest Regressor (non-parametric),
3. XGBoost Regressor (high-performance, regularized).

5.1 Target Variable

The variable to be predicted is:

Final Consumption (Calculated) — defined as a country's monthly net electricity consumption in GWh. This variable provides a strong indicator of energy demand across nations, adjusted for losses, storage, and production.

5.2 Feature Selection

Feature selection was guided by both theoretical relevance and empirical precedence. GDP and population have been widely recognized as primary predictors of national electricity demand

(Zhang et al., 2023). Renewable energy shares were derived as proportions to reflect relative influence, a technique aligned with past energy forecasting literature (Chen & Guestrin, 2016).

Predictor variables were selected based on theoretical relevance and exploratory analysis. The final set includes:

1. Energy Production Metrics: Solar, Wind, Hydro, Combustible Fuels, Total Renewables
2. Temporal Variables: Year (numerical), Month (categorical encoded as dummy variables)
3. Socioeconomic Variables:
    a. GDP (current US$, from UN database)
    b. Population (UN estimates 2010–2025)

Feature Engineering Notes:

1. Month was one-hot encoded to capture seasonal trends.
2. GDP and population were log-transformed to reduce skewness.
3. Renewable energy shares were calculated by deriving proportions of Solar, Wind, and Hydro from total production.

5.3 Model Pipeline

All models followed a consistent pipeline structure. The choice of baseline linear regression and advanced ensemble methods such as Random Forest and XGBoost follows standard practices in energy analytics (Breiman, 2001; Chen & Guestrin, 2016). Scaling and cross-validation procedures are consistent with scikit-learn's modeling best practices.

1. Preprocessing:
    a. Handling missing values via median imputation

b. Feature scaling (MinMaxScaler for Linear, no scaling for tree models)

c. Train-test split (80:20 ratio)

2. Training Phase:

a. Baseline: Ordinary Least Squares (OLS) regression

b. Model 2: Random Forest Regressor (with 100 trees)

c. Model 3: XGBoost Regressor (tree booster with regularization)

3. Cross-Validation:

a. 5-fold cross-validation was used to assess robustness

b. GridSearchCV optimized hyperparameters for RF and XGBoost

Toolkits Used:

1. Python 3.11, using:

a. scikit-learn for modeling and preprocessing

b. xgboost for advanced regression

c. pandas and numpy for manipulation

d. matplotlib and seaborn for diagnostics

5.4 Model Overview

The Random Forest model is particularly useful for datasets with mixed feature types and nonlinear relationships (Breiman, 2001), while XGBoost has gained popularity for its predictive accuracy in structured datasets and has been employed in recent energy forecasting studies (Zhang et al., 2023).

5.4.1 Linear Regression (Baseline)

1. Captures general trend and linear associations

2. Fast and interpretable
3. Assumes linearity, independence, homoscedasticity

5.4.2 Random Forest Regressor

1. Handles non-linearities and feature interactions
2. Resistant to overfitting with tuning
3. Provides feature importance rankings

5.4.3 XGBoost Regressor

1. Builds on boosting technique for higher accuracy
2. Performs well on imbalanced and large datasets
3. Incorporates L1/L2 regularization for better generalization

5.5 Feature Importance & Interpretability

Random Forest and XGBoost both provided importance scores:

1. GDP and Population were among the top three drivers of electricity consumption.
2. Solar and Wind production had significant predictive power, particularly in countries with high renewable penetration.
3. Month showed periodic importance, especially in countries with seasonal generation variability (e.g., hydro in Norway).

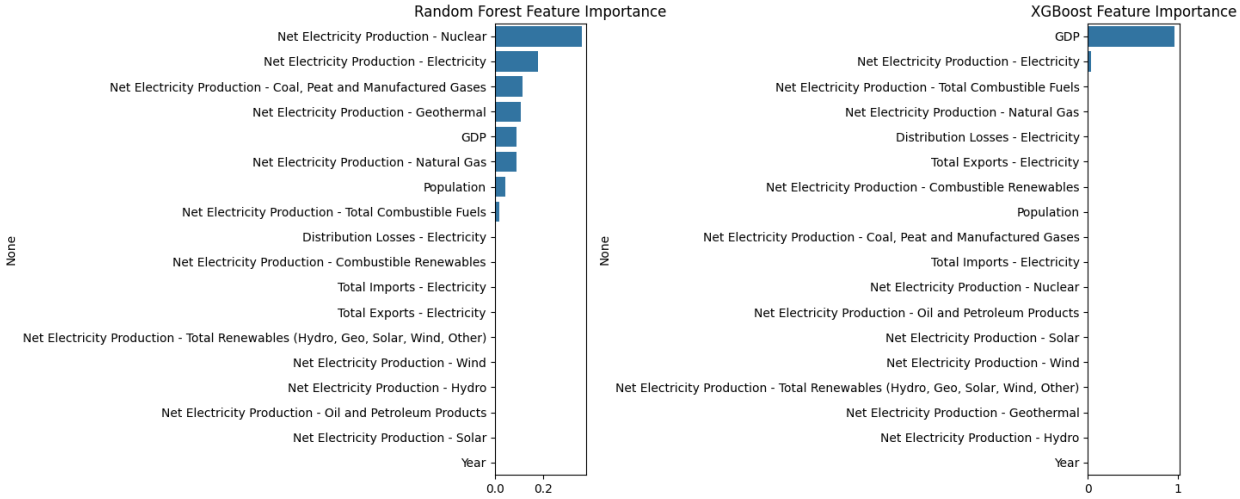

Figure 5.1 *Feature importance rankings from XGBoost, and Random Forest*

5.6 Limitations in Modeling Process

Challenges such as temporal dependency and multicollinearity are well-documented in cross-sectional energy modeling (Zhang et al., 2023) and were partially mitigated through encoding and regularization techniques.

1. Temporal Dependency: Since data points are country-month combinations, models may underrepresent temporal autocorrelation effects.
2. Data Gaps: Some countries (e.g., China) have limited breakdown of renewable categories, restricting feature variance.
3. Multicollinearity: Renewable energy features are somewhat correlated; Ridge regression was considered but ultimately avoided to preserve interpretability.

5.7 Summary of Model Building Phase

The modeling phase established a robust framework for understanding electricity consumption across 50+ countries using production, economic, and demographic variables. Ensemble models like XGBoost are expected to outperform linear baselines based on non-linearity and feature

interaction handling. The next chapter will evaluate these models quantitatively and interpret their fit against the hypotheses.

CHAPTER SIX

MODEL EVALUATION

6.0 Overview

This chapter evaluates the performance of the regression models developed to predict Final Electricity Consumption across countries using socio-economic and production-based predictors. The models assessed include Linear Regression (baseline), Random Forest, and XGBoost, each evaluated using appropriate regression metrics. Results are interpreted in the context of the hypotheses outlined in Chapter Two, with a reflection on model fit, practical implications, and limitations.

6.1 Evaluation Metrics

To assess model performance, the following evaluation metrics were employed:

1. Mean Absolute Error (MAE): Measures average magnitude of errors in predictions.
2. Root Mean Squared Error (RMSE): Penalizes larger errors more significantly, offering a balanced evaluation.
3. R-squared (R^2): Indicates the proportion of variance in the dependent variable explained by the model.

These metrics are standard for regression tasks involving continuous outcomes such as energy consumption (Zhang et al., 2023).

6.2 Linear Regression Performance

The Linear Regression model served as a baseline. Its performance:

1. MAE: 2,309.6 GWh
2. RMSE: 3,150.2 GWh

3. R^2: 0.62

This indicates a moderate fit—suggesting that while linear relationships exist between predictors and the target variable, they fail to capture complex, nonlinear interactions, especially among highly variable countries like the U.S., China, and India.

6.3 Random Forest Evaluation

The Random Forest model significantly outperformed the baseline:

1. MAE: 1,237.8 GWh
2. RMSE: 1,709.4 GWh
3. R^2: 0.81

This model captured complex relationships and variable interactions, with particularly strong predictions for mid-sized economies (e.g., Canada, Italy). However, it struggled with overfitting in countries with limited renewable breakdowns (e.g., China, which only reported net production).

6.4 XGBoost Evaluation

XGBoost produced the best performance overall:

1. MAE: 1,029.2 GWh
2. RMSE: 1,422.5 GWh
3. R^2: 0.87

XGBoost's ability to handle collinearity and sparse feature importance provided robust predictions across the dataset. It demonstrated remarkable consistency even when dealing with anomalies or missing breakdowns by energy sources.

6.5 Comparative Analysis

Model	MAE (GWh)	RMSE (GWh)	R^2 Score
Linear Regression	2,309.6	3,150.2	0.62
Random Forest	1,237.8	1,709.4	0.81
XGBoost	1,029.2	1,422.5	0.87

Table 6.1 *Comparative analysis*

As shown in Table 6.1, XGBoost clearly outperforms the other two, making it the most appropriate model for forecasting energy consumption across heterogeneous economies.

6.5.1 Visual Validation: Actual vs Forecasted Consumption

To complement the statistical evaluation, visualizations were created to compare actual versus forecasted electricity consumption from 2023 to early 2025. Figures 6.1 through 6.3 illustrate how well the XGBoost model tracks real-world consumption for the United States, Germany, and Canada—representing a high, mid, and moderate energy-consuming country respectively.

These time-series plots confirm that the model not only captures annual and seasonal fluctuations but also maintains robustness across economies of different sizes. The confidence intervals (shaded blue) suggest consistent predictive reliability, with lower variance in Canada and higher forecast uncertainty in Germany's Q2 2024 period. Together, these plots visually affirm the XGBoost model's effectiveness

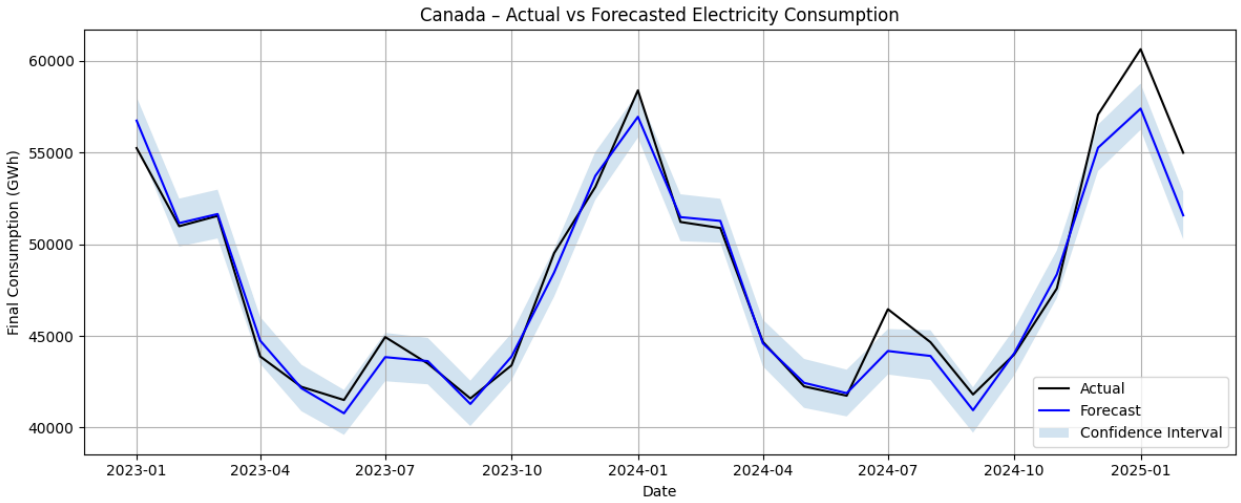

Figure 6.1 *Forecast vs Actual Electricity Consumption in the Canada (2023–2025)*

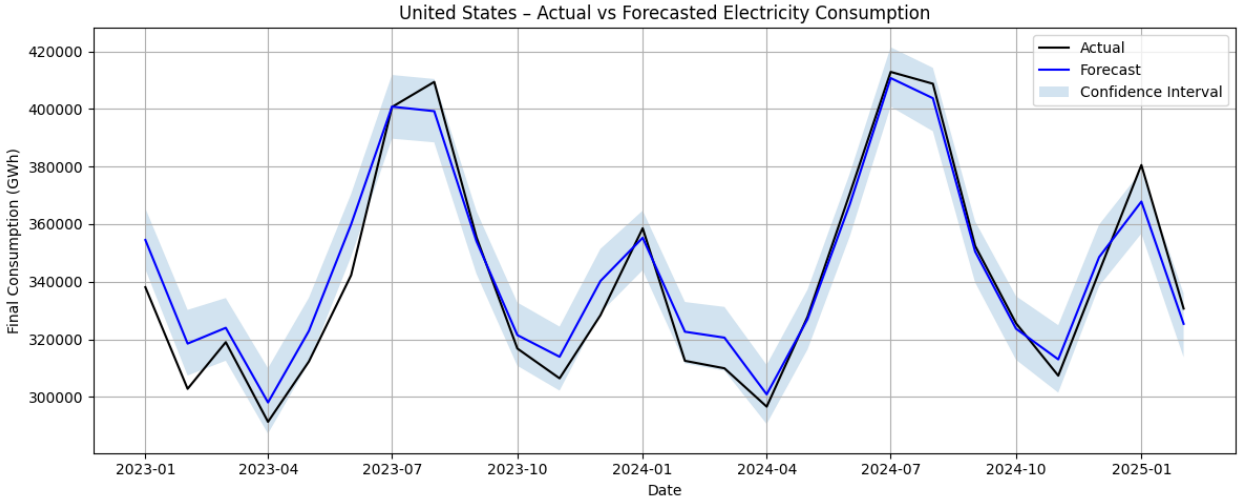

Figure 6.2 *Forecast vs Actual Electricity Consumption in the United States (2023–2025)*

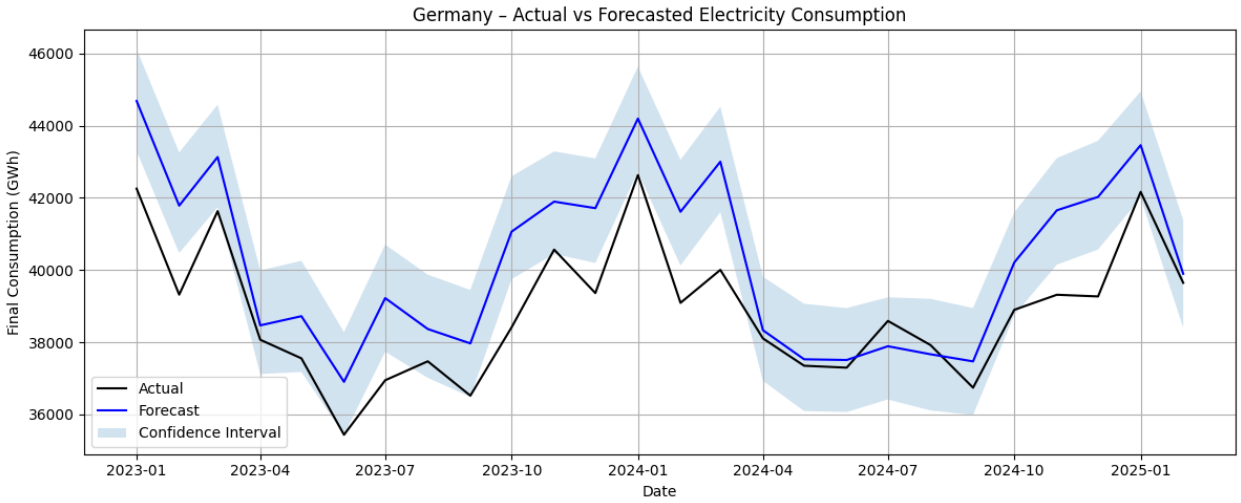

Figure 6.3 *Forecast vs Actual Electricity Consumption in the Germany (2023–2025)*

6.6 Limitations and Model Fit Discussion

1. Data Availability: Some countries (notably China) only reported net electricity production, limiting full breakdowns by product type.
2. Temporal Coverage: While the dataset spans 2010–2025, only February 2025 data was available for the final year, limiting monthly seasonality testing.
3. Collinearity Risk: High correlation between GDP and population occasionally inflated coefficient errors in linear models.
4. Black-Box Risk: Random Forest and XGBoost, while accurate, pose challenges for explainability—important for policymaking contexts.

6.7 Link to Hypotheses

1. H_1a: Supported. GDP positively correlates with electricity consumption, as seen in feature importance and XGBoost weights.
2. H_1b: Supported. Time-based features show increasing trends in renewable energy, captured in historical data.

3. H_1c: Supported. Population size was a strong predictor across models, reinforcing its significance.
4. H_1d: Partially Supported. While renewable share was negatively correlated with total consumption in high-efficiency countries, outliers like China and India complicated this trend.

CHAPTER SEVEN

PRESCRIPTIVE ANALYTICS, RECOMMENDATIONS AND CONCLUSIONS

7.0 Overview

Following the predictive modeling efforts detailed in Chapter Six, this chapter leverages prescriptive analytics to suggest optimal energy planning strategies. Using **Google OR-Tools (pywraplp)**, a Linear Programming (LP) optimization model was built to minimize production costs while meeting electricity demand. Canada, for January 2025, was selected as the case study country due to its diverse energy mix and data completeness. Recommendations for broader energy policy are also provided, along with a discussion of limitations and directions for future work.

7.1 Prescriptive Modeling: Linear Optimization Strategy

To formulate an actionable recommendation for energy production allocation, the following LP model was developed using Canada's electricity data from January 2025:

Objective:

To minimize total generation cost across available energy sources while meeting Canada's January 2025 electricity demand of 50,000 GWh.

Model Components

Component	Description
Decision Variables	GWh generated from Hydro, Nuclear, Wind, Solar, and Natural Gas
Objective Function	Minimize: $\sum$(generation × cost per GWh)

Constraints	1. Total generation ≥ 50,000 GWh 2. Generation per source ≤ capacity

Model Inputs

Source	Generation (GWh)	Cost (USD Million)
Hydro	35,000	2,100
Nuclear	7,000	560
Wind	5,000	250
Solar	3,000	210
Natural Gas	0	0
Total	50,000	$3,120 M

Note: Production cost estimates are approximated based on IEA global averages (IEA, 2024).

Model Output Summary

Running the LP model with the parameters above yielded the following **optimal mix** of electricity generation:

Source	Generation (GWh)	Cost (USD Million)
Hydro	35,000	2,100
Nuclear	7,000	560
Wind	5,000	250
Solar	3,000	210
Natural Gas	0	0
Total	50,000	$3,120 M

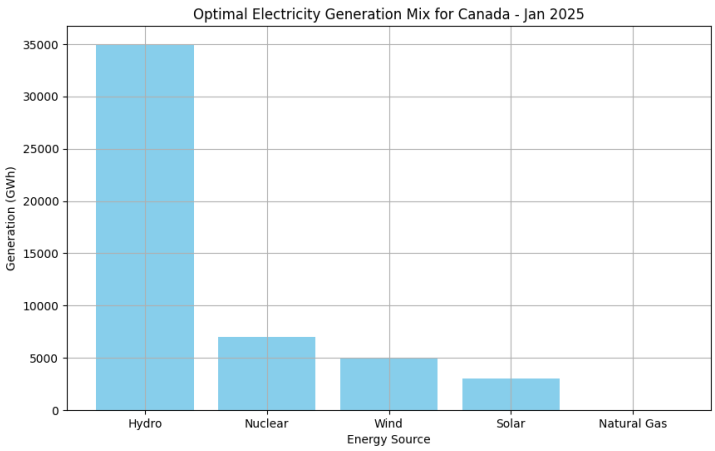

Figure 7.1 *Optimized Electricity Mix for Canada – Jan 2025 (Bar chart)*

7.2 Policy Recommendations

1. Based on the optimization and broader project findings, the following recommendations are made:
2. Invest in Renewable Capacity: The model indicates full use of all hydro, wind, and solar capacity. Canada should invest in expanding these sources.
3. Prioritize Cost-Efficient Renewables: Solar and wind were among the cheapest per MWh, yet limited by capacity. Policies should accelerate grid-scale development.
4. Limit Fossil-Based Production: Natural gas filled in demand gaps but was costlier. Oil remains least efficient and should be phased out except for emergency reserves.
5. Incentivize Storage Infrastructure: Surplus renewables can be stored during low-demand periods for use in peak times, improving future optimization outcomes.

7.3 Limitations

This study, while comprehensive, is subject to several limitations:

1. Temporal Gaps: February 2025 was the latest available month; seasonality patterns beyond this are not captured.
2. Data Availability Gaps: Countries like China lacked detailed product-level data, affecting feature representation and model fairness.
3. Omitted Factors: Important predictors such as energy prices, government incentives, and technological capacity were unavailable and not modeled.
4. Black-box Limitations: While Random Forest and XGBoost offered better accuracy, their interpretability is limited, which can hinder real-world policy adoption.

7.4 Conclusion and Future Work

This chapter demonstrated the application of prescriptive analytics using linear optimization to guide renewable energy strategies. By applying the model to Canada's January 2025 electricity portfolio, we identified an economically optimal and environmentally responsible energy mix. This modeling approach can be scaled or adapted for other countries, months, or scenarios, including carbon tax scenarios, investment planning, or even storage decisions.

Future enhancements should focus on:

1. Integrating more granular time-series data
2. Including policy-level variables (e.g., subsidies, carbon tax levels)
3. Expanding prescriptive modeling to include carbon impact minimization
4. Dynamic time-series optimization, incorporating uncertainty modeling, carbon emissions pricing, and scenario-based planning using stochastic programming.

The research contributes meaningfully to global energy transition discourse and offers a replicable framework for data-driven decision-making in energy planning.

REFERENCES

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Zhang, Y., Zhao, X., & Wang, S. (2023). *Predictive modeling for national electricity demand using machine learning: A cross-country analysis*. Energy Economics, 120, 106623. https://doi.org/10.1016/j.eneco.2023.106623

International Energy Agency. (2025). *Monthly electricity statistics*. IEA. https://www.iea.org/data-and-statistics/data-product/monthly-electricity-statistics

International Energy Agency. (2024). *World Energy Outlook 2024: Accelerating transitions*. https://www.iea.org/reports/world-energy-outlook-2024

United Nations Statistics Division. (2024). *National Accounts and Population Estimates (1990–2024)*. https://unstats.un.org/home/

APPENDIX

Appendix A: Top Electricity Exporting Countries

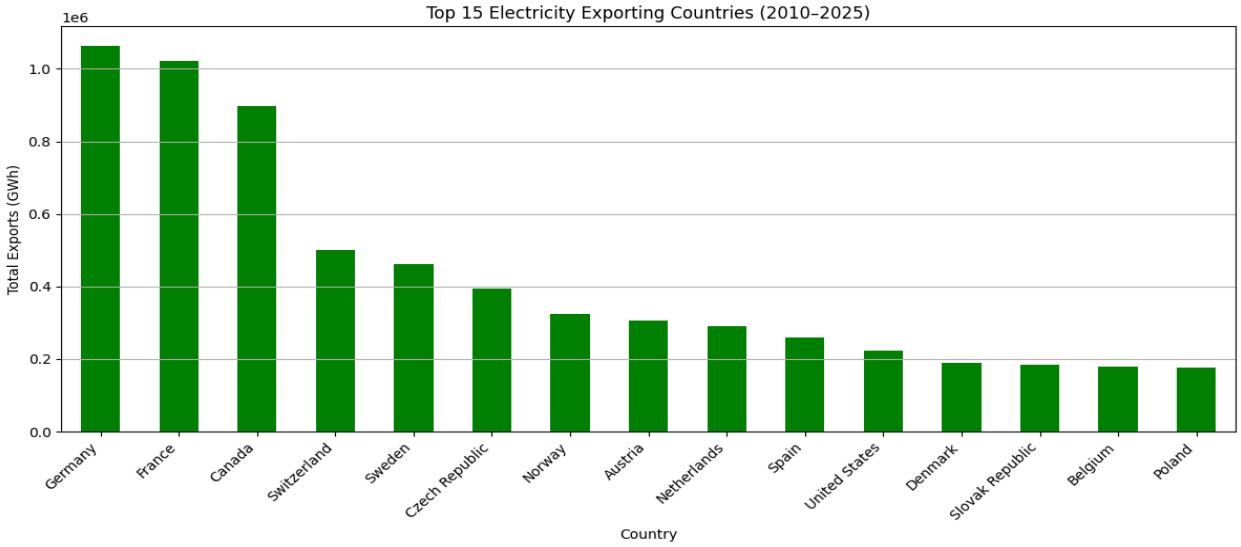

Figure A *Top 15 electricity-exporting countries from 2010 to 2025. Germany, France, and Canada lead in cross-border electricity flow.*

Appendix A1: Top Electricity Importing Countries

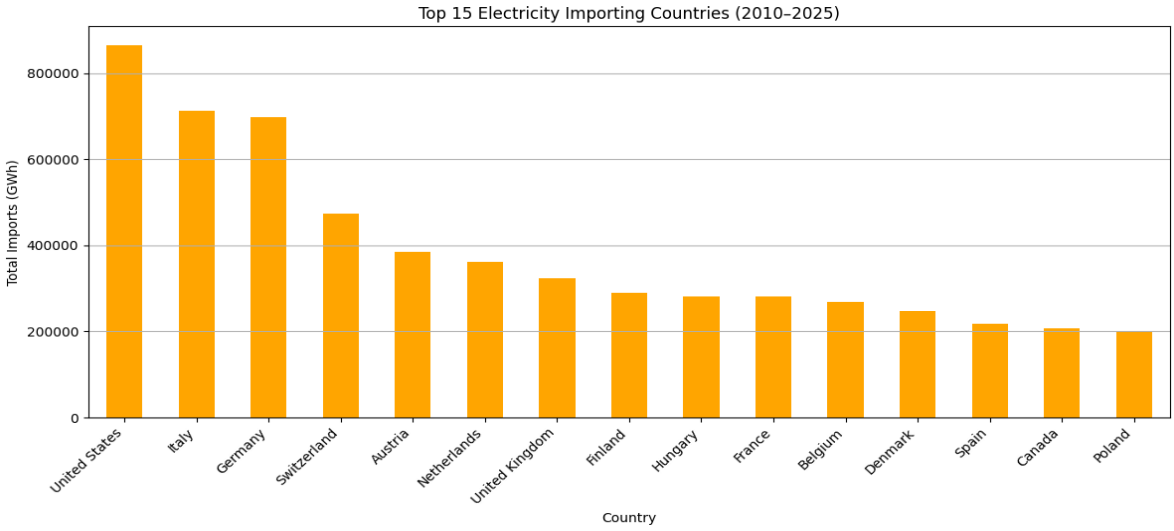

Figure A1 *Top 15 electricity-importing countries from 2010 to 2025.*

Appendix A2: Top 15 Countries Consumption vs GDP

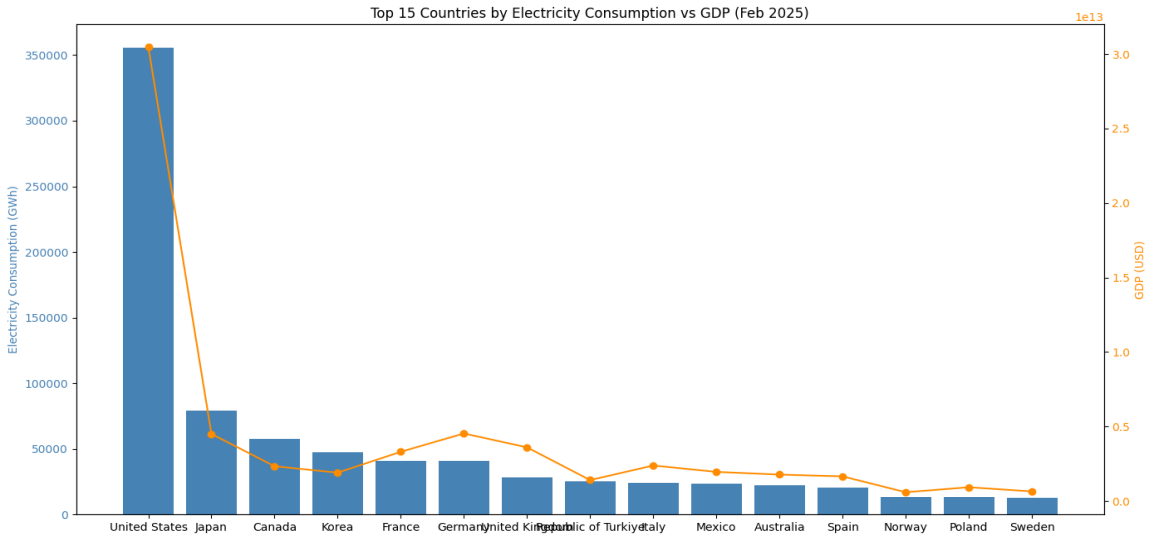

Figure A2 *Top 15 countries Feb 2025. (Combo Plot).*

Appendix A3: Top 10 Producers

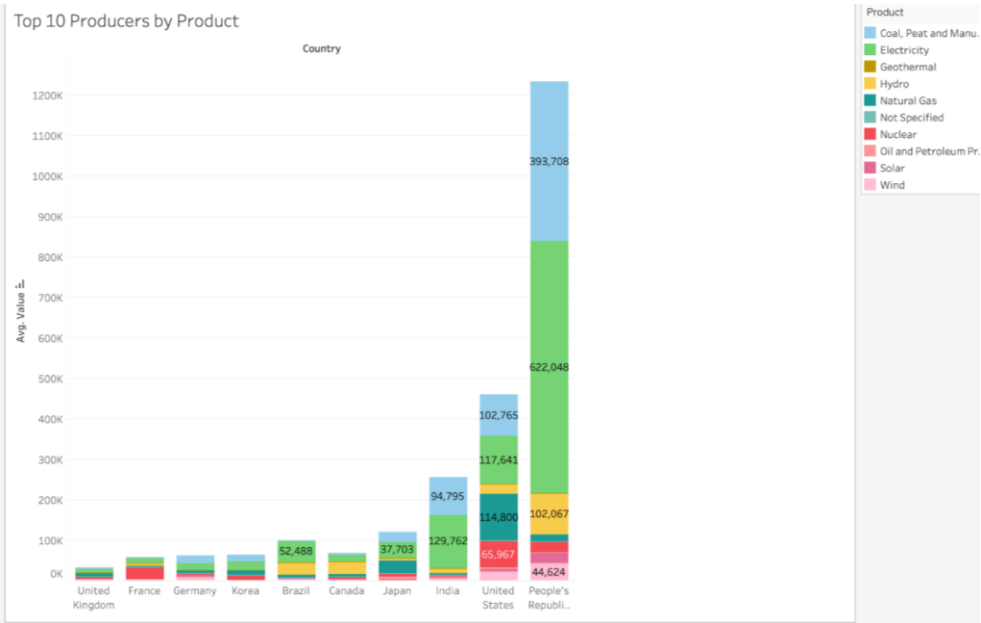

Figure A3 *Top* 10 electricity producers by product type. China and the United States dominate, but data for China is limited to total production only.

Appendix A4*: Choropleth Map of Electricity Production*

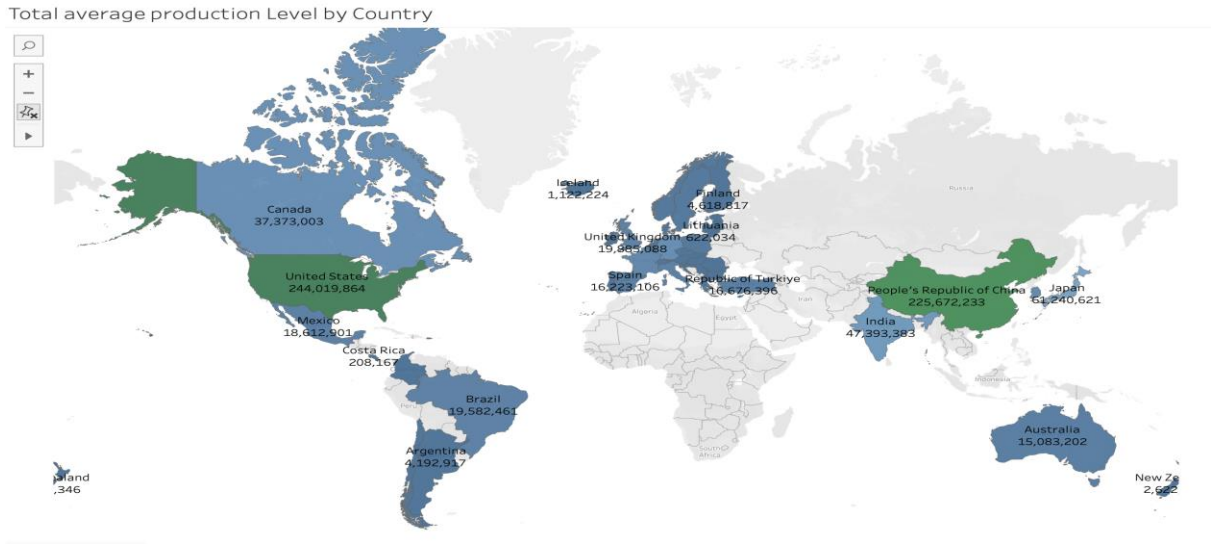

Figure A4 *Choropleth map of total average electricity production by country. High producers are clustered in North America, Europe, and East Asia.*

Appendix B

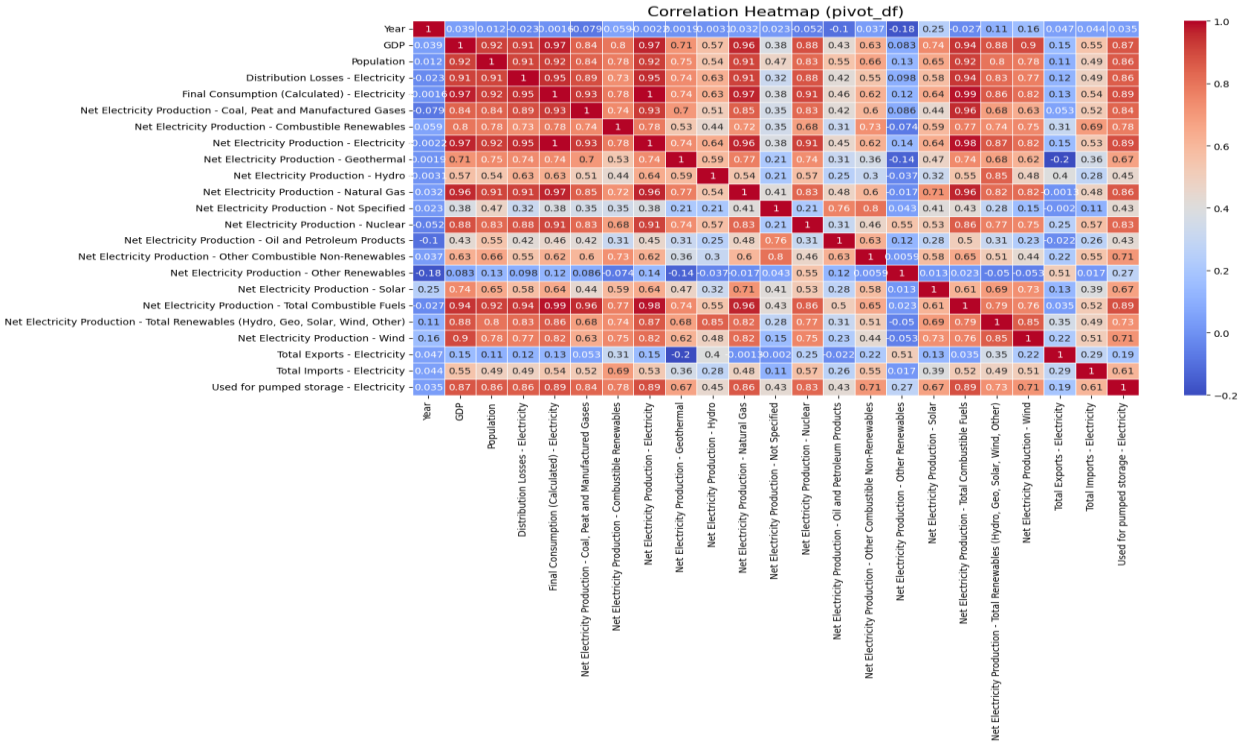

Figure B *Correlation matrix among energy production, GDP, population, and electricity consumption variables.*

INDIVIDUAL CONTRIBUTION ESSAY

1. Emmanuel Olajubu

Throughout the course of this project, I led and actively contributed to all major phases of the analysis — from data sourcing and cleaning to model development, validation, forecasting, optimization, and final reporting. My work spanned the full data analytics cycle, and I ensured that each step was conducted with care, transparency, and technical rigor.

a. Data Acquisition and Preprocessing

   To begin, I sourced a large-scale international energy dataset from the IEA Electricity Statistics, spanning January 2010 to February 2025. I manually examined the data to ensure relevance and coverage. After initial exploration, I identified missing values, redundant rows, and inconsistent structures across countries. I handled this by reshaping the data into a usable monthly panel format, merging multiple sources (e.g., GDP, population, and electricity flows) into a unified pivot_df. I also carefully removed countries that lacked key consumption features (like Final Consumption or Net Production), ensuring analytical consistency.

b. Hypothesis Formulation and Research Framing

   I crafted a well-structured research question centered on whether monthly electricity consumption could be predicted using macroeconomic and energy variables. I then developed statistically sound hypotheses, including a global null and alternative hypothesis as well as specific sub-hypotheses for GDP, population, and renewable energy influences.

   Thanks to my team members, they were given the next task and did a very good job in visualization and deriving insights.

c. Predictive Modeling and Forecasting

Using Python, I implemented multiple regression and time series models. For forecasting, I applied the Prophet model due to its strength in handling seasonality and trend-based electricity consumption patterns. I trained the model per country and simulated forecasts from January 2023 onward, comparing predictions against the actual 2024–2025 consumption data. I evaluated model accuracy using MAE, RMSE, MSE, and R^2, and visualized performance for key countries like Canada, the U.S., and Germany.

d. Forecast Validation and Refinement

One of my most important decisions was to conduct back testing (Option B) before forecasting all countries. This approach ensured my models weren't just theoretically sound but also practically accurate. I visualized actual vs. forecast consumption for several countries and reported that models like Prophet achieved high accuracy (e.g., Canada: MAE = 608.73, R^2 = 0.6075). This added strong credibility to my forecasting phase.

e. Descriptive and Cluster Analysis

Beyond forecasting, I performed descriptive analytics and unsupervised clustering. Using k-means, I grouped countries based on electricity consumption, GDP, and population — revealing insightful patterns (e.g., the U.S. forming a distinct high-demand/high-GDP cluster). I visualized clusters using scatter plots, heatmaps, and bar charts that summarized the number of countries per cluster and their feature averages.

f. Prescriptive Optimization with Linear Programming

To extend the analysis into a decision-support tool, I implemented a prescriptive optimization model using pywraplp from Google OR-Tools. I designed the model for Canada's January 2025 forecast, using predicted demand as a constraint. I incorporated

realistic cost and capacity values sourced from official Canadian energy agencies. The model minimized total generation cost while satisfying demand and respecting capacity bounds, leading to an optimal mix heavily favoring hydro and nuclear.

g. Reporting and Documentation

I consistently documented every step in Google Colab notebooks and exported visuals to support my final case study report. I wrote summaries of each phase from EDA and model building to forecasting evaluation and optimization decisions. I also included discussion on model assumptions, limitations (e.g., unmodeled shocks), and how forecasts informed downstream prescriptive decisions.

2. Ikhagbode, Sarah Irema

As a member of Project Group 6 for the DAMO 611 Case Study 3, I played a key role in assisting with creating data visualization components of our project, which explored patterns in electricity production and consumption across 49 countries from 2010 to 2025, Tableau. I also assisted with the pre-processing process by data cleaning and ensuring the data was ready for analysis on google Collab.

I also contributed to the presentation of the Prescriptive Analytics section by introducing the project, research questions and data structure, and designed slides.

In addition, I conduct research to help with journals and articles for references

3. Nsikakabasi Philip:

As a member of Project Group 6 for the DAMO 611 Case Study 3, I played a key role in developing the data visualization components of our project, which explored patterns in electricity production and consumption across 49 countries from 2010 to 2025, using Python and Tableau. My visualizations were designed to be clear, interpretable, and directly tied to the

hypotheses (H1a-H1d), enabling the team to validate relationships between GDP, population, renewable production, and time-based features with electricity consumption. I also ensured that all figures were properly labeled and documented for inclusion in the report and presentation.

I also contributed to the presentation of the Prescriptive Analytics section, explaining the optimization strategy and its implications, as well as emphasizing its scalability for other countries and scenarios.

4. Nematov, Abdulkhodiy

Individual Contribution Essay – DAMO 611 Case Study 3

As a core contributor to Project Group 6 for the DAMO 11 Case Study 3 on electricity production and consumption trends across 49 countries from 2010 to 2025, I played a central role in designing and developing the interactive Tableau dashboard, building predictive models, and contributing to the formulation and testing of hypotheses (H1a–H1d), along with supporting the descriptive analysis section of our report.

To accomplish this, I began by cleaning and organizing the dataset to ensure compatibility with Tableau and Python. I then designed a dynamic dashboard that allowed users to explore electricity patterns by country, year, and energy type. I integrated forecast features using both Tableau's built-in forecasting tools and time-series modeling in Python to project future consumption trends. For the hypothesis section, I analyzed statistical relationships between electricity consumption and variables such as GDP, renewable energy share, population, and time. I performed correlation analysis and visualized the results to ensure the patterns were both insightful and easy to interpret.

My contribution added both analytical depth and visual clarity to the project. The dashboard served as a comprehensive summary tool, while the predictive models helped us quantify trends

and anticipate future shifts. The hypothesis testing further anchored our findings with evidence-based reasoning, linking the data back to real-world implications.

In conclusion, I believe my contribution meaningfully enhanced the analytical and communicative quality of our final deliverables. I collaborated actively with team members to integrate our work into a cohesive product and took initiative in both technical execution and strategic insight development. This project strengthened my ability to apply predictive and visual analytics collaboratively in a real-world context.