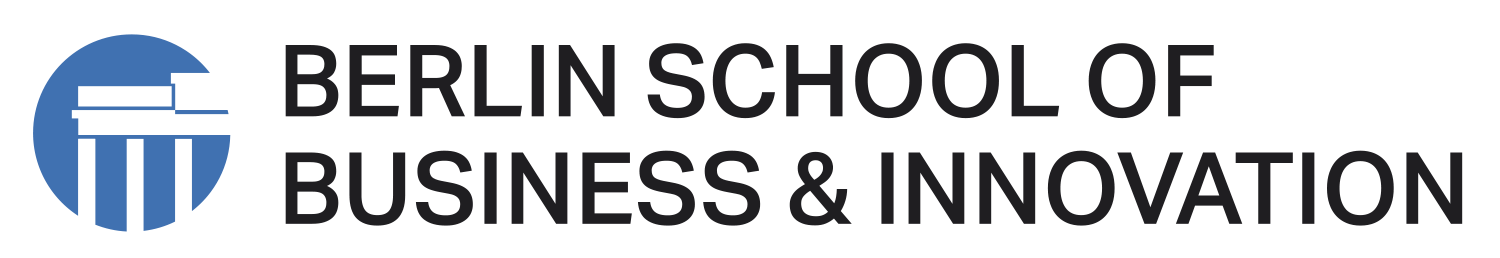
Shape

Description automatically generated with medium confidence



**Essay / Assignment Title: A Multi-Sector Time Series Analysis: Linking Climate, Energy, and Consumer Demand using Machine Learning**

**Programme title: Predictive Analytics and Machine Learning Using Python**

**Name: Kwaku Bonful Bosompim**

**Year: 2025**

# **CONTENTS**

[CONTENTS](#)

[INTRODUCTION](#)

[1.1 Background of Predictive Analytics and Machine Learning](#_fyi1vunrswji)

[1.2 Project Motivation](#_a1r4jc36k3u6)

[1.3 Objectives and Scope](#_247k4ddjwiwr)

[CHAPTER ONE](#)

[1.1 Description of Datasets](#_pxmx9360dn8d)

[2.2 Target Variables](#_j7jtp0jucdft)

[2.3 Data Cleaning and Preprocessing](#_427oak65n3u9)

[2.3.1 Summary Statistics](#_eahmqq6issnf)

[2.4.1 Exploratory Data Analysis (EDA)](#_s240ajhjcvns)

[2.4.2 Outliers and Missing Data Detection Using Box Plot](#_bx47s49vp2yn)

[2.4.3 Additional Visualizations](#_sbnphzbbukeu)

[2.5 Problem Statement and Hypotheses](#_7ny9t331ma9q)

[CHAPTER TWO](#)

[3.1 Regression Models](#_rxgd4y3xqp8)

[3.1.1 Linear Regression](#_spw8y5r3w2o8)

[3.1.2 Multiple Regression](#_dq62e6wgb1xk)

[3.1.3 Polynomial Regression](#_6cymke9sq0c2)

[3.1.4 Evaluating Overfitting and Underfitting](#_s0bitya9emg1)

[3.2 Classification Models](#_qx9b5tkz8j6b)

[3.2.1 Decision Trees](#_i9a7ahxduumn)

[3.3 Clustering Models](#_vj91ru14q524)

[3.3.1 DBSCAN](#_rmoeobxnpo5k)

[3.3.2 K-Means Clustering](#_bgeuu28d4adx)

[3.4 Ensemble Methods](#_qm7r0kcnertv)

[3.4.1 Random Forest](#_872e98hcr408)

[3.4.2 Gradient Boosting](#_cqdt3i673h0u)

[3.4.3 XGBoost](#_f7sb3zps5e1)

[3.5 Neural Networks](#_2ql8sqvbhb1j)

[CHAPTER THREE](#)

[4.2 Visual Comparison of Predictions vs Actuals](#_jvqjhsbili82)

[4.3 Interpretation of Findings](#_bbalypi2ns60)

[4.4 Business / Investment Implications](#_rtzpaf4etjuc)

[CONCLUDING REMARKS](#_kxbehjcbmjun)

[BIBLIOGRAPHY](#)

[APPENDIX (if necessary)](#)

**Statement of compliance with academic ethics and the avoidance of plagiarism**

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people's texts I clearly declare it through the good use of references following academic ethics.

(In the case that it is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the program).

Name and Surname (Capital letters):

KWAKU BONFUL BOSOMPIM

.......................................................................................................................................

Date: ...**2025**......../...**10**..../..**01**....

# **INTRODUCTION**

This project explores how predictive analytics and machine learning can be applied using Python to understand and forecast real-world patterns across different sectors. The datasets include climate (temperatures), energy (electric production), and consumer demand (beer and shampoo sales). Each dataset will first be analyzed separately to identify trends and seasonality, and then combined to study their relationships. The project focuses on time series forecasting models that can predict future values, and investigates how climate conditions influence energy production and consumer demand. The aim is to show how integrating data from multiple domains can generate insights useful for business, investment, and policy decisions.

This project focuses on datasets from climate, energy, and consumer sectors, with the goal of building integrated forecasting models. We will analyze:

* Electric production data to forecast energy supply, model price sensitivity, and potentially link to electricity derivatives. Energy analytics is a rapidly growing field — the global utility analytics market is forecasted to grow at ~14.9% CAGR through 2035. [Future Market Insights](https://www.futuremarketinsights.com/reports/utility-analytics-energy-analytics-market?utm_source=chatgpt.com)
* Daily minimum temperature data, which is directly relevant to climate risk, energy demand modeling, and ESG investing. Variations in temperature often drive changes in electricity consumption and commodity demand.
* Consumer demand data (beer production, shampoo sales), which captures consumption trends and seasonality in FMCG / retail sectors, useful for supply chain, inventory, and market demand forecasting.

We acknowledge that all of these are time series datasets. First, we will analyze each dataset individually to identify trends, seasonality, and anomalies. Next, we will examine their interrelationships — checking whether climate, energy, and consumer demand “move together.” Finally, we will build multivariate machine learning / time series models that combine these sectors to make better, more holistic forecasts.

**Keywords**

Time series, forecasting, machine learning, climate, energy, consumer demand, regression, neural networks.

## 1.1 Background of Predictive Analytics and Machine Learning

Data Analytics is the process of studying datasets to understand the information within them, using methods such as descriptive, predictive, and statistical analysis along with tools like Excel, SQL, Python, and Tableau to transform raw data into meaningful insights; this analytical foundation provides the background for advanced fields such as Predictive Analytics and Machine Learning, which extend these techniques to forecast future outcomes and enable systems to learn and make autonomous decisions. *((16) Inside The Big Tech Playbook: Top 10 Data Analytics Tools Driving Business Success | LinkedIn, no date)*

Predictive Analytics uses past data to guess what might happen in the future and answers the question, 'What is likely to happen? Predictive Analytics involves using historical data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data.(*Blog & News | PI.EXCHANGE*, no date)

Machine learning teaches itself by using data, algorithms, and tools to find patterns, make predictions, and improve at tasks without being explicitly programmed. Machine Learning is a subset of artificial intelligence that involves the use of algorithms and statistical models to enable computers to improve their performance on a task through experience.

It involves reading and cleaning data, exploring and understanding it, deciding how to present it to the algorithm, selecting the right model and algorithm, and correctly measuring its performance.(‘Building\_Machine\_Learning\_Systems\_with\_Python’, no date)

## 1.2 Project Motivation

This project is motivated by the fact that many real-world sectors — climate, energy, and consumer markets — are closely connected. For example, changes in temperature can affect energy use (heating or cooling), and shifts in energy supply can influence prices and even consumer behavior. Businesses and investors want to understand these links because it helps with planning, risk management, and better decision-making. Also, the demand for advanced data skills in energy and finance is growing very fast, which makes this project both useful and relevant (Zhang, Li and Zhou, 2025; McKinsey, no date; World Economic Forum, 2021).

## 1.3 Objectives and Scope

The main objectives of this project are:

* To analyze climate, energy, and consumer demand data separately and find patterns such as trends and seasonality.
* To check if these datasets are related to each other (for example, if temperature changes impact energy production or consumer demand).
* To use different machine learning models to forecast future values.
* To show how predictive analytics can be applied in business, investment, and policy decisions.
* The scope of the project is limited to four datasets: daily minimum temperatures, electric production, monthly beer production, and shampoo sales.

(Mystakidis *et al.*, 2024)

**Problem Statement**

We want to understand how temperature affects electricity use and whether this also influences consumer demand for products like beer and shampoo.

**Hypotheses**

Temperature variations affect electricity production, which in turn influences consumer demand (beer or shampoo) (Fache and Bhat, no date; Avdakovic, Ademovic and Nuhanovic, 2013; Burke and Business, no date; ResearchGate, no date)

**Materials and Methods**

Four datasets—**daily minimum temperatures, electricity production, monthly beer production, and shampoo sales**—were analyzed using Python and machine learning models for forecasting. All coding and model execution were carried out on the Google Colab and Document platforms, which served as the working environment for data processing, analysis, and visualization.

**Python Libraries** that will we use include NumPy, Pandas, Matplotlib, and Scikit-learn and many more

**Features** = Temperature + Beer Production + Shampoo Sales

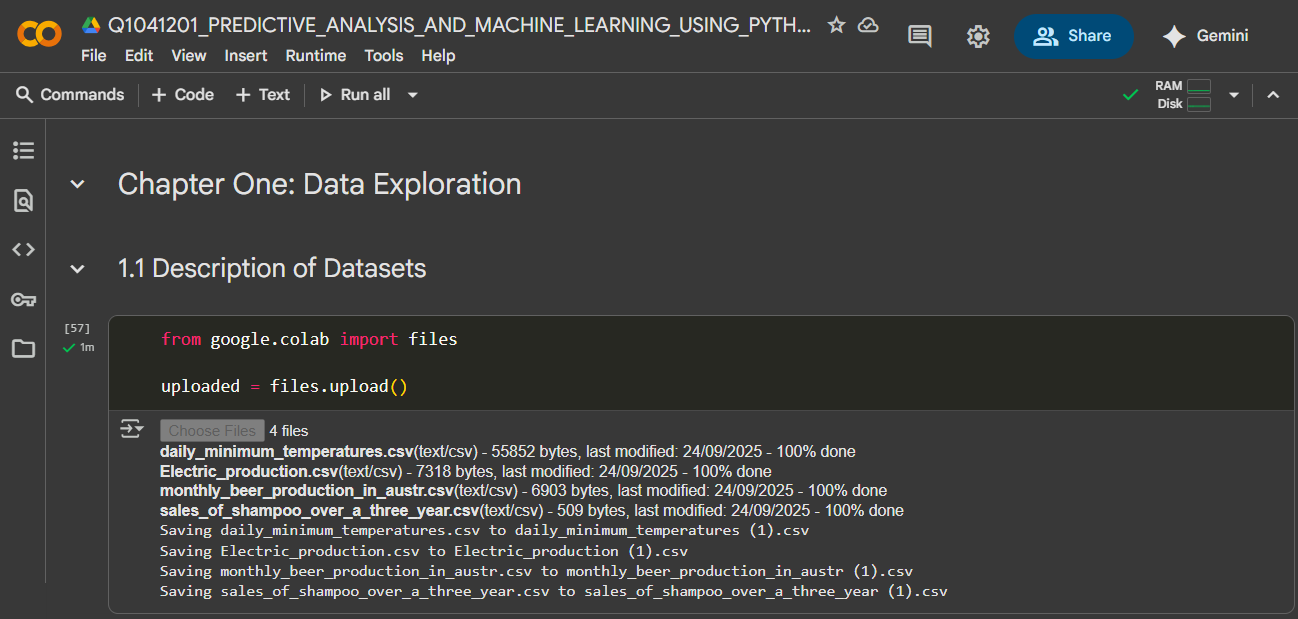
**Label** = Electricity Production

The Features are the Inputs or the things we use to make predictions

Labels are Outputs or the thing we want to predict

| This diagram shows the features (inputs) flowing into the label (output) |
| --- |
|  |

# **CHAPTER ONE**

In this chapter, we explore the datasets before any predictions. We check each dataset individually to see what it contains, spot trends, and identify any missing values or unusual numbers. We also visualize relationships to understand how climate, energy, and consumer demand might interact.

So first we start by importing the datasets into workspace using the google.colab library, load, read the data using pandas to understand the datasets

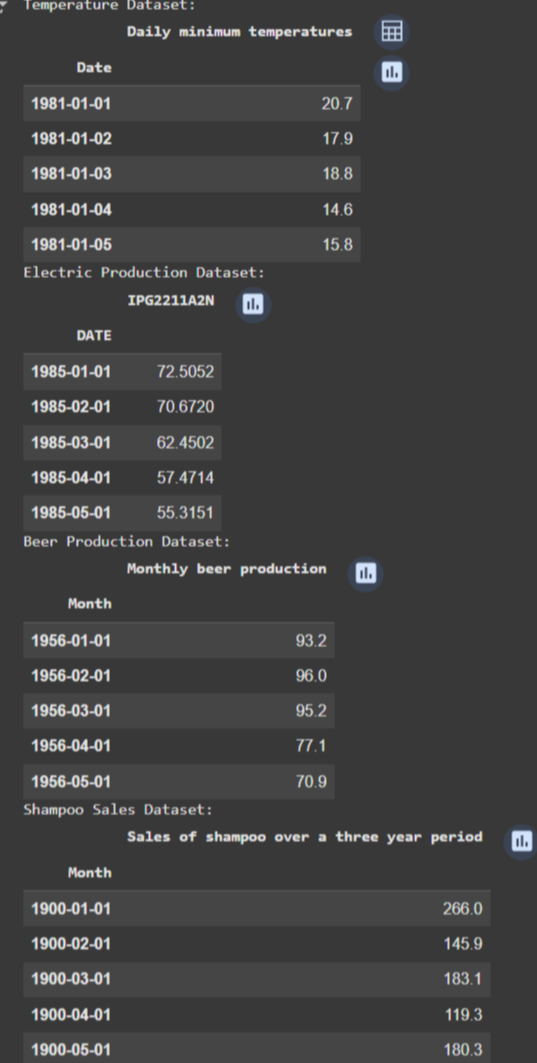
## 1.1 Description of Datasets

We have four datasets: Temperature – daily minimum temperatures. Electricity Production – monthly energy production. Beer Production – monthly beer production in Australia. Shampoo Sales – monthly shampoo sales over three years. The climate dataset records daily minimum temperatures, the energy dataset shows electric production, and the consumer datasets track sales of beer and shampoo. Each dataset contains a time index and a numerical variable of interest.

## 2.2 Target Variables

Each series can be predicted individually (e.g., beer sales) this is what we term as **Univariate forecasting**

**Multivariate forecasting**: Explore relationships across all four series (e.g., do higher temperatures affect electricity demand and beer production?). Temperature: Daily minimum temperature (forecast next day/month), Electric production: Monthly electricity productionBeer: Monthly beer production, Shampoo: Monthly sales

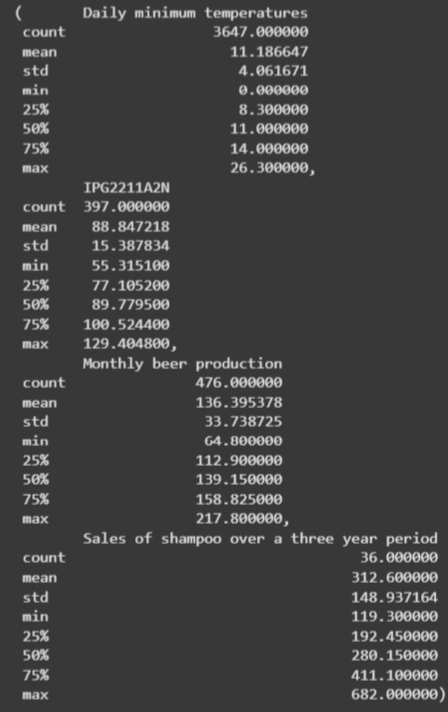


From the figure above, we loaded and read each dataset separately and printed out the first five rows of each dataset and preview. From here we could observe the structure of the various datasets. We could have an overview of the data types or categories each datasets belong to.

## 2.3 Data Cleaning and Preprocessing

Now that we have had a preview of the various datasets, we dive deeper into obtaining more information on the variables, data types, checking and fixing missing values, or removing duplicates combined with visualization to have a fair view of the datasets we want to work with.

### 2.3.1 Summary Statistics



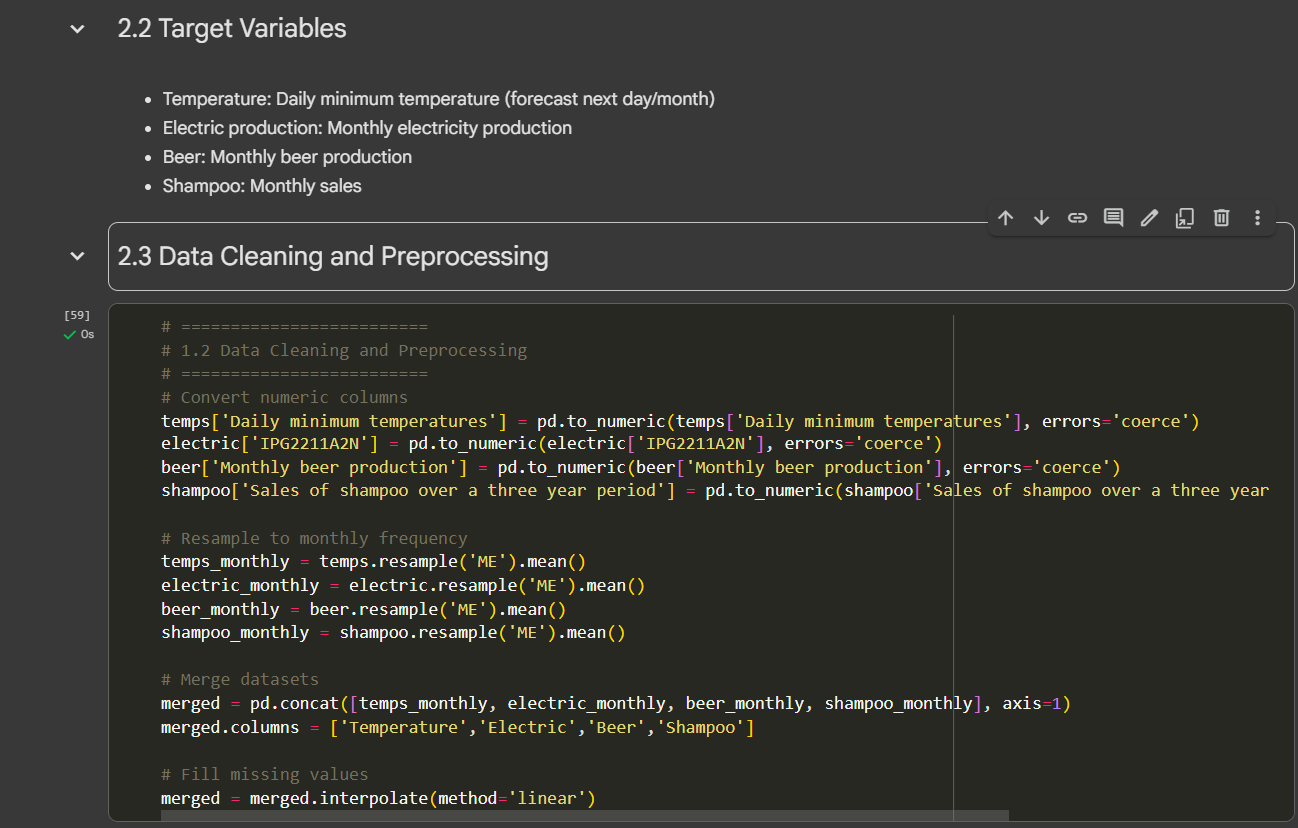
From the data above we looked at basic statistics — minimums, maximums, and averages. These numbers may look simple, but they help us set expectations. For example, is the average monthly beer production reasonable? Are there extreme values in electricity generation that might indicate unusual events?

Here we need to ask questions such as:

* Do we have missing or invalid values?
* Are the dates formatted correctly?
* Do all datasets speak the same “language” of time (daily, monthly)?

Poor data quality can lead to misleading results and poor model performance. Clean data helps in building robust and reliable models.

The data here was already clean and ready for use — which means we didn’t need to worry about missing values or duplicates.

Since each dataset contains a time index and a numerical variable of interest. They were resampled into monthly frequency for alignment and merged into a single dataset with four key features: and this could be shown from the figure below:

From the figure above the time columns were converted into datetime format.

The date were set as an index and resampled into monthly means.

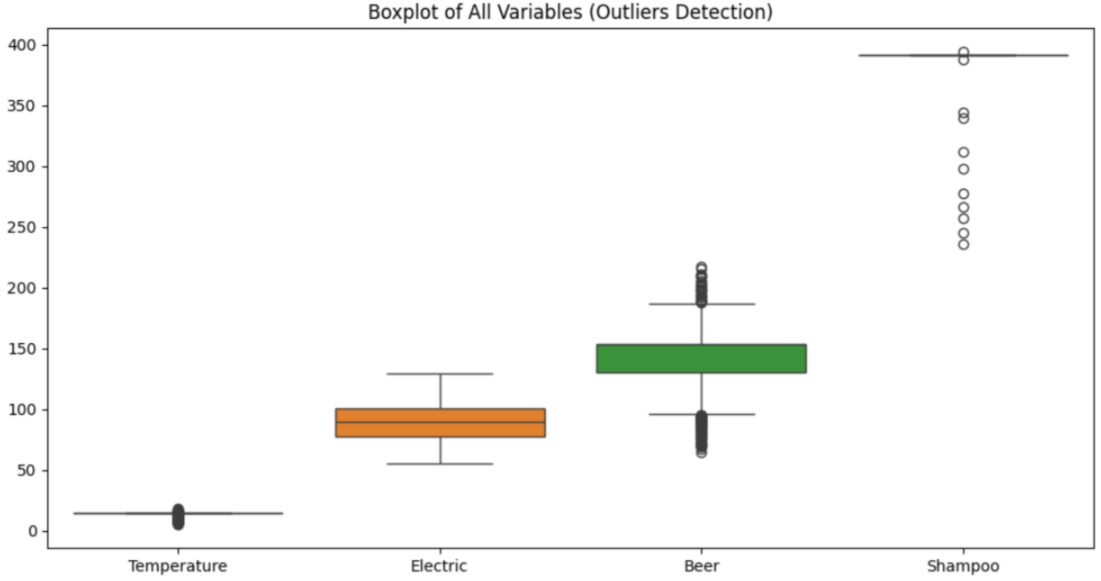
Ensured numerical variables were clean and handled missing values with dropna() for models requiring complete data.

### 2.4.1 Exploratory Data Analysis (EDA)

Once the data is loaded and cleaned, this phase is to understand the shape and personality of the data. We ask ourselves do we have enough records to make useful predictions? Or maybe even too much, in which case we need to think about sampling.

This stage is called exploratory data analysis (EDA). Here, we do not just rely on numbers — we use plots (such as line plots, bar plots, and histograms), distributions, and time series graphs to get a real feeling for the data. We notice that even simple visualizations like a line plot or histogram can reveal patterns, seasonality, and anomalies that raw numbers alone cannot using Matplotlib.

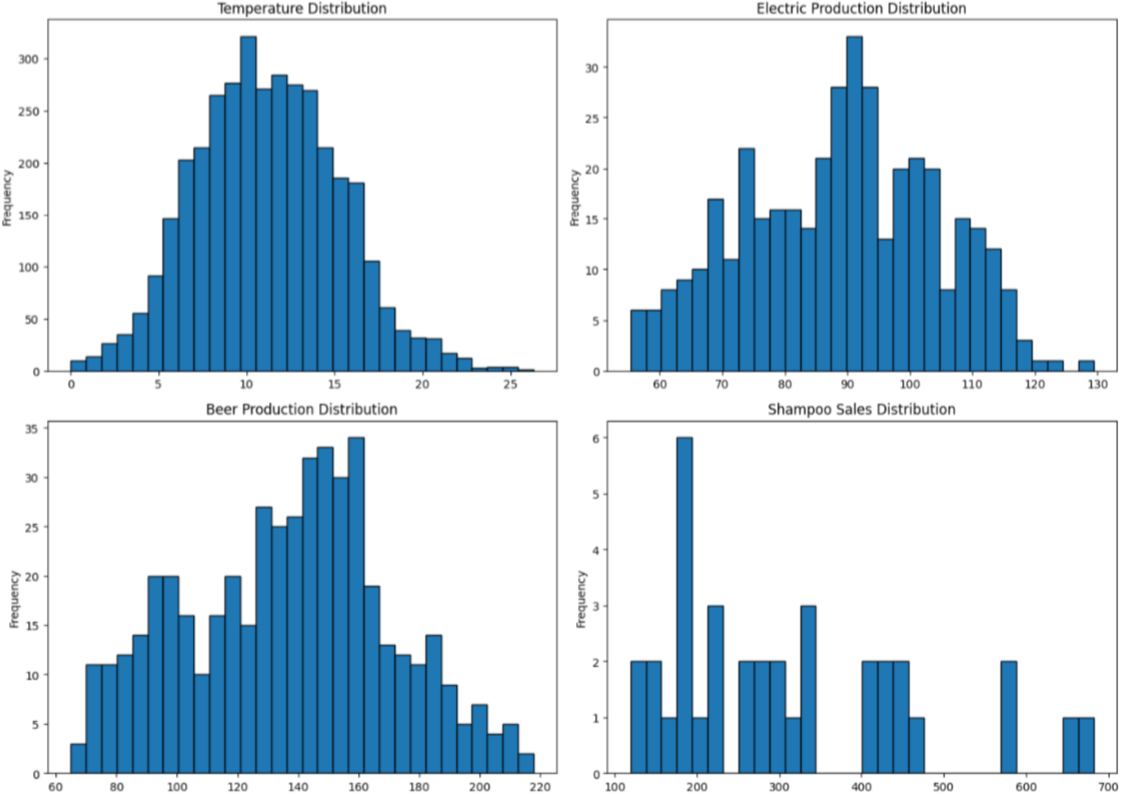
### 2.4.2 Outliers and Missing Data Detection Using Box Plot



The diagram above is a box plot which shows the middle range of the data (where most of the values lie). The line inside the box is the median, the middle point of the dataset. The whiskers stretch out to show the spread of the data. And finally, the dots outside the whiskers — these are the outliers. They are values that are much higher or lower than the rest, and they normally represent unusual events, errors, or just rare cases. It could observe that the Beer data presumably has a lot of outliers as compared to other variables.

In this project, we used boxplots to quickly check if our datasets had strange or extreme values. For example: In the temperature dataset, unusually hot or cold days show up as outliers. In the electric production dataset, spikes could represent sudden surges in energy demand. In the beer and shampoo datasets, outliers could be linked to holiday seasons or promotions when demand suddenly jumps. If data is missing, it sometimes shows up in boxplots as broken shapes or unusual clustering. In our case,

### 2.4.3 Additional Visualizations

Histograms of all four datasets:

The histogram above takes all the numbers in the dataset and groups them into “bins” (ranges of values). Then it counts how many values fall into each bin and stacks them as bars.

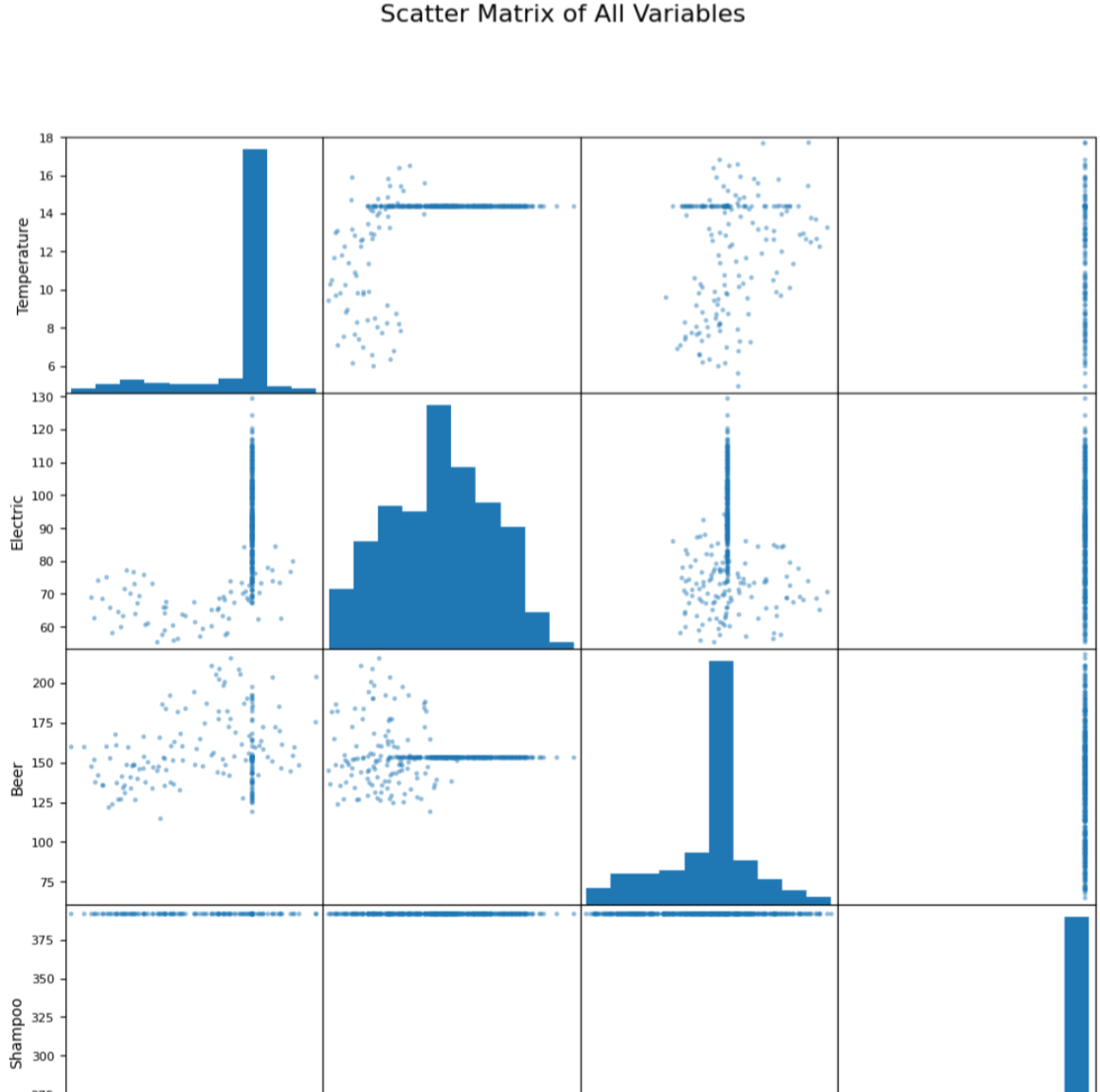
Temperature dataset: The histogram shows us if most days are cool, warm, or hot. We can easily spot if the climate data is normally distributed or skewed. The histogram looks close to a bell curve, showing that most daily minimum temperatures are around 10°C. This makes sense because weather tends to fluctuate around an average but rarely goes to extremes.

Electric production dataset: The bars help us see whether energy output is steady or whether there are spikes — maybe from seasonal or monthly patterns. We can see production values clustering around 90–100 units, but with some variation. This shows a steady central trend but also some volatility that might be seasonal.

Beer dataset: The histogram lets us peek into people’s drinking habits — do they drink steadily all year, or are there big jumps around summer?. The bars jump around, showing production has more ups and downs. This reflects seasonality — people drink more in some months (summer, holidays) than in others.

Shampoo dataset: Sales histograms reveal demand patterns — maybe most months are stable, but some months show a big leap (promotions or festive seasons). The histogram is uneven and “spiky.” This tells us shampoo sales are not smooth and may depend on promotions or seasonal buying habits.

We can see which datasets are stable (like temperatures), which are steady with mild changes (electricity), and which are highly seasonal or irregular (beer and shampoo).

**Scatter Matrix of all Variables**

***Scatter Matrix of Climate, Energy, and Consumer Demand Datasets***

When we want to understand how multiple variables relate to each other, one of the most useful tools is the scatter matrix

The scatter matrix figure above (also called a pair plot) is a grid of scatter plots that shows the relationship between all pairs of variables in a dataset

Each plot shows how two variables interact for example, how temperature might relate to electricity production or beer sales. The diagonal plots display the distribution of each individual dataset.

By looking at matrix above, we can quickly spot:

* Correlations (if the points form a clear line or trend).
* Clusters (groups of similar points).
* Outliers (points that stand far away from the rest).

This helps us decide which variables might be important predictors for machine learning models later.

**Temperature vs. Electric Production** → There may be a relationship where colder or hotter days drive higher electricity usage (for heating or cooling). If the scatter shows a trend, it suggests that climate has a direct influence on energy demand.

**Temperature vs. Beer Production** → Warmer months might show higher beer production (people drink more cold beverages in summer). A seasonal trend here would be expected.

**Temperature vs. Shampoo Sales** → We do not expect a strong correlation, but changes in lifestyle with seasons (like holiday seasons or summer routines) may cause slight variations.

**Beer vs. Shampoo Sales** → Both are consumer products, so they might share seasonal patterns (e.g., higher sales during festive or holiday seasons).

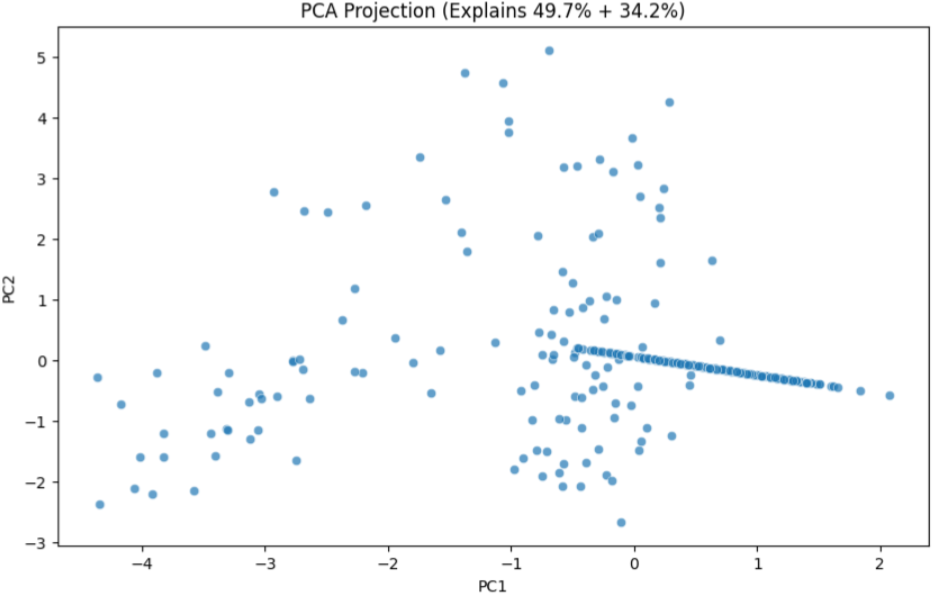
**Electric Production vs. Beer/Shampoo Sales** → If consumer activity rises (holidays, warm seasons), we might also see increased electricity production.

**Diagonal Histograms** → These show each dataset’s distribution. For example, beer and shampoo sales may show strong seasonality (peaks and valleys), while temperature will show cycles across months.

Overall, the scatter matrix helps us confirm that climate, energy, and consumer demand may “move together” at certain times of the year. These observed relationships motivate building predictive models that combine them.

**PCA (Dimensionality Reduction)**

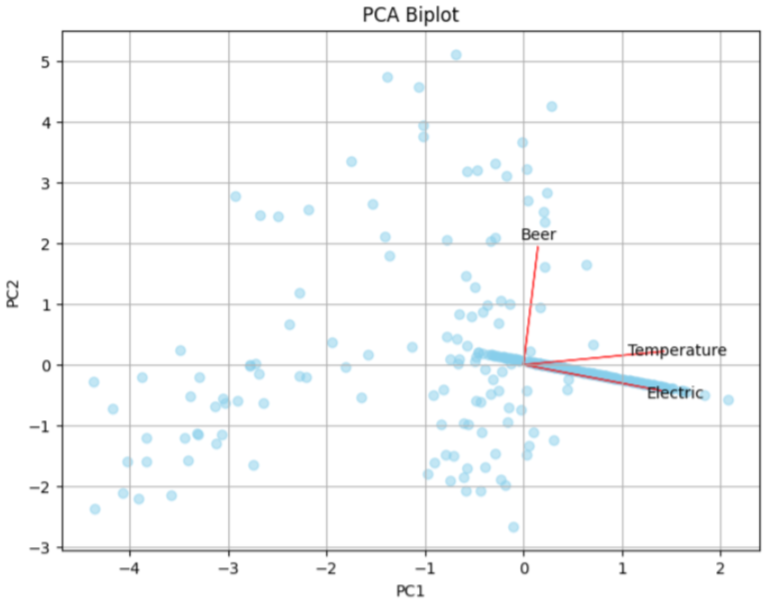
We used PCA to shrink our data into just two main pieces so we can see patterns more clearly. This makes the data easier to compare and sets us up for better models later.



**it shows**: If points are close together, the datasets behave in a similar way. If they are far apart, they act differently. For example, beer and temperature might sit closer, meaning beer sales go up in warmer months.

**PCA Biplot**

For advanced understanding the PCA Biplot is explored. The arrows show which dataset has more effect on the main directions (PC1 and PC2).



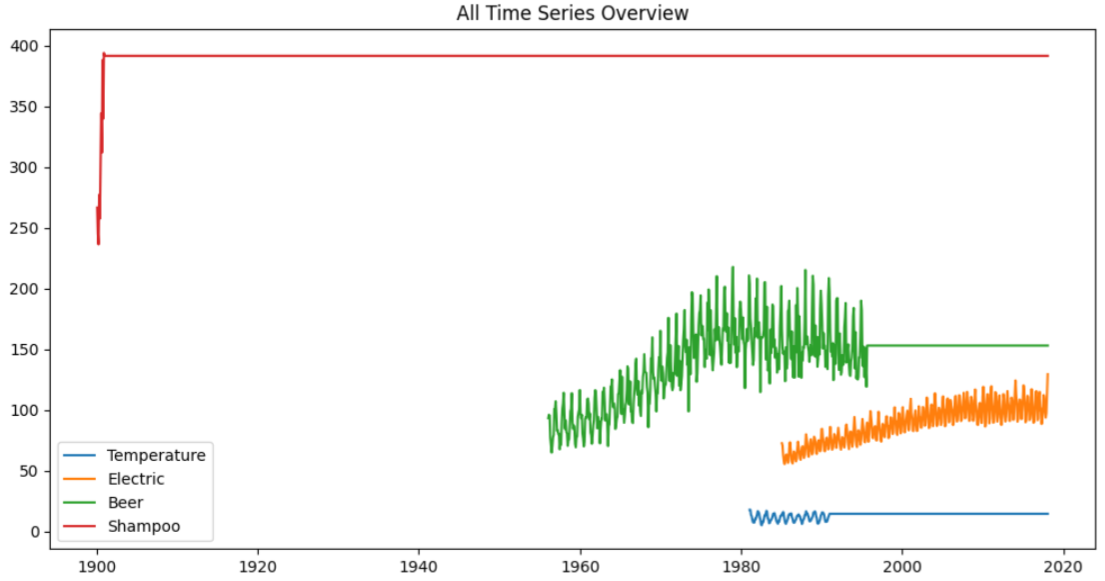
A long arrow means that the dataset has a big influence on the pattern. If two arrows point in a similar direction, those datasets behave in a similar way. For example, if temperature and electricity arrows point close together, it means climate strongly affects energy use.

**Line Plots**

This plot below shows us how these variables change over time or a specific season.

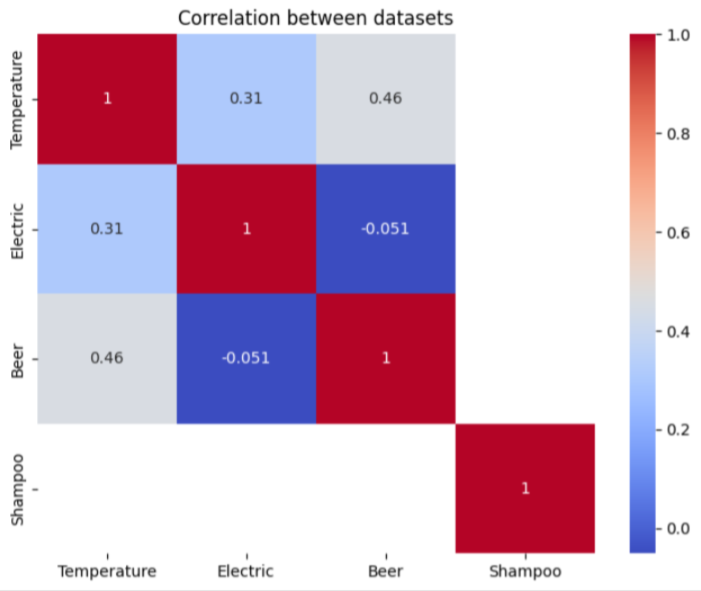
Temperature blue line → moves up and down in cycles (seasons). Electric production orange line→ shows how energy use changes, sometimes following the climate. Beer production green line→ goes up in warmer months (summer) and drops in colder ones. Shampoo sales red line → shows shopping patterns, often with repeating peaks (holidays or seasons).

Line plots help us spot trends, cycles, and sudden changes in the data



**Heatmap: Correlation Between Datasets**

This heatmap below shows how the datasets are connected: **Temperature & Beer (0.46)** → Warmer weather is linked with more beer production. **Temperature & Electric (0.31)** → Climate changes affect electricity production, but less strongly. **Beer & Electric (-0.05)** → Almost no relationship between these two. **Shampoo** → Moves mostly on its own, not strongly tied to the others



## 2.5 Problem Statement and Hypotheses

**Problem Statement**

Weather, energy, and consumer behavior are often linked. For example, when it gets hot, people use more electricity for cooling, and in warm seasons, beer sales may go up. But these links are not always clear. The challenge is to use machine learning to connect these datasets — temperature, electricity production, beer, and shampoo — and see how they affect each other. Studies show that climate changes have a strong impact on electricity demand (Chen, Fang and Khayatnezhad, 2024; Research on medium- and long-term electricity demand forecasting under climate change, 2025). Other research also shows that weather changes can influence both energy use and consumer demand (Parkpoom, Harrison and Bialek, n.d.; Harang, Heymann and Stoop, 2020).

**Hypotheses**

* *Temperature affects electricity production* — hot or cold weather increases energy use, which changes electricity supply.
* *Temperature influences consumer demand* — for example, warm weather may boost beer sales.
* *Energy and climate together shape consumption* — changes in electricity (availability or cost) may also affect demand for everyday items, such as shampoo.

**Objective**

To build forecasting models that bring together climate, energy, and consumer data. This will help us see not only the trends in each dataset but also how they connect with each other.

# **CHAPTER TWO**

In this chapter, we move from exploring the data to actually building models. The goal is to test different machine learning methods to see how well they can explain and predict the links between climate, energy, and consumer demand.

**Training and Testing Procedures**

Before we start with deployment of each model, we split the data into training and testing sets. The training set is used to teach the model, while the testing set is kept aside to check how well the model performs on new, unseen data. This way, we can see if the model is just memorizing patterns (overfitting) or if it can truly generalize.

In this project, we trained the models on part of the sales and electricity data, then tested them on the rest. We measured performance using metrics like RMSE and R², which showed us how far off predictions were and whether the model explained the data well.*(A review of model evaluation metrics for machine learning in genetics and genomics - PMC, no date)*

Each model will help answer a slightly different question

* Regression models help us see how one factor (like temperature) can predict another (like electricity or beer sales).
* Classification models help us group or label data into categories.
* Clustering methods allow us to find hidden patterns in the data without labels.
* Ensemble models combine many models to improve accuracy.
* Neural networks let us capture more complex, non-linear relationships.

Along the way, we will also check whether our models are overfitting (too focused on training data) or underfitting (too simple to capture patterns). This helps us make sure the models can generalize well to new data.

## 3.1 Regression Models

Regression models are like drawing a line through our data to see how one thing changes when another thing changes. In this project, we want to see how temperature, electricity, beer, and shampoo sales are connected.

We try different kinds of regression:

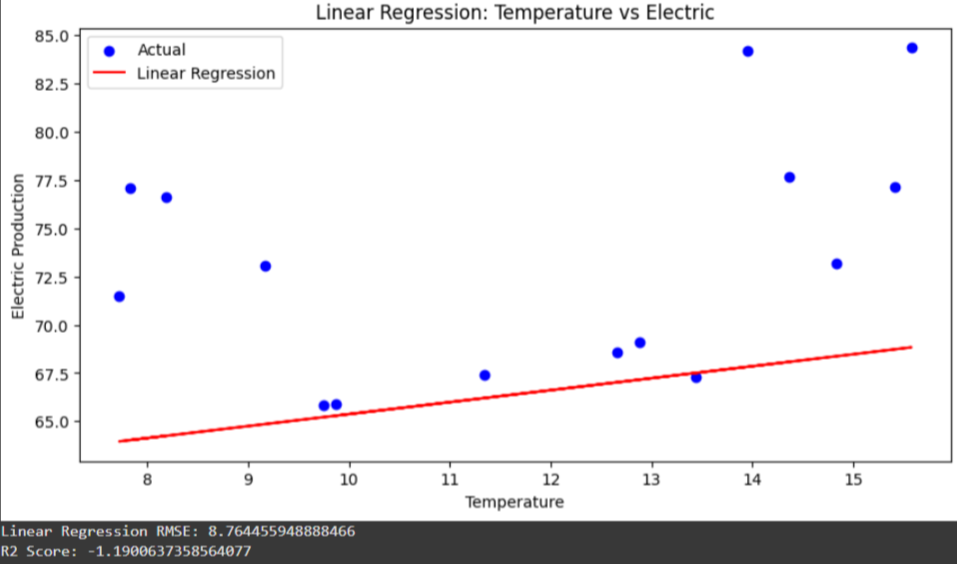
* Linear Regression: Fits a straight line.
* Multiple Regression: Uses more than one input to make predictions.
* Polynomial Regression: Fits a curved line to capture more complex trends.

These models help us test if temperature really influences electricity and consumer demand, and whether adding more variables gives us better predictions.

### 3.1.1 Linear Regression

Linear regression is the simplest form of regression. We try to answer questions like:

*"If the temperature goes up by one degree, how much does electricity production change?"*



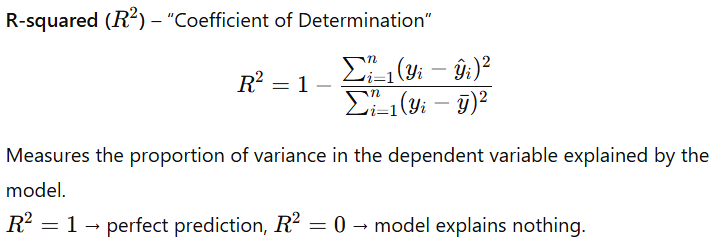
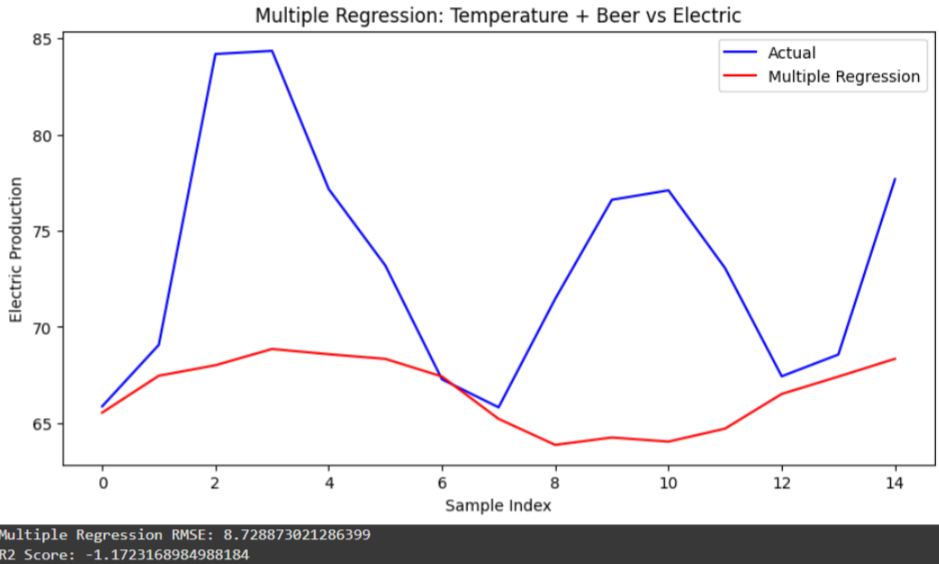
The line from the diagram above shows the general trend. If the slope is positive, it means as temperature rises, electricity production or sales tend to rise too. If it’s negative, then the opposite happens.

A study by Phan Thanh Dao et al. (2024) explored the utilization of linear regression analysis to predict energy consumption in practical applications, demonstrating its effectiveness in real-world scenarios.

In a study by A. Fache (2024), a linear regression model was used to forecast monthly electricity sales, considering factors like comfortable temperature range and sudden variables, highlighting the model's applicability in energy demand forecasting.

### 3.1.2 Multiple Regression

Sometimes, one factor is not enough. For example, electricity production might depend not only on temperature but also on beer sales or shampoo sales. Multiple regression allows us to add more inputs together.

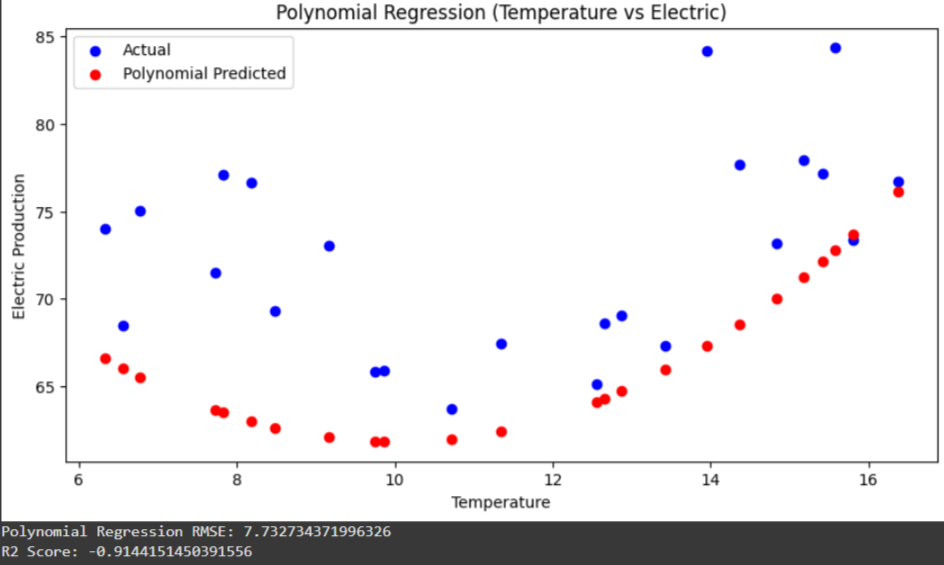
Here we see whether combining several factors gives us a better prediction. If the model’s accuracy improves, it means these sectors are truly connected. We established the relationships between Temperature or climate, electricity production and beer sales and we could deduce the accuracy metric which is **R-squared (R2)** – “Coefficient of Determination” has improved from **-1.19 to -1.17** though both values are negative.Supporting Research:

A study by Stanko Dimitrov et al. (2023) emphasized the importance of accounting for climate when determining the impact of weather on retail sales, suggesting that multiple regression models can help in accurately assessing the effects of various factors on sales.

Research by Jianli Pan et al. (2012) applied multiple polynomial regression to model daily energy consumption as a function of outdoor temperature and humidity, demonstrating the effectiveness of incorporating multiple variables in regression models.

### 3.1.3 Polynomial Regression

Real life is not always a straight line. For example, very hot or very cold temperatures may both increase energy use. Polynomial regression allows us to fit a curve instead of a straight line.



This shows us whether relationships are more complex than straight lines. If the curve fits the data better, then the link is non-linear.

Supporting Research

Research by W.F. Mbasso et al. (2024) introduced an advanced mathematical methodology for predicting energy generation and consumption based on temperature variations, utilizing polynomial regression to capture non-linear relationships.

### 3.1.4 Evaluating Overfitting and Underfitting

When building regression models, it is not just about how well the model fits the current data. We also want it to generalize well, meaning it can make good predictions on new, unseen data. This is where overfitting and underfitting come in.

**1. Underfitting (Too Simple)**

Underfitting happens when the model is too simple to capture the patterns in the data.

Example: Fitting a linear regression to a relationship that’s actually curved.

Signs:

* Both training and test errors are high
* R² is low or negative

**2. Overfitting (Too Complex)**

Overfitting happens when the model is too tailored to the training data, capturing noise instead of the underlying trend.

Example: Using a high-degree polynomial regression on a small dataset.

Signs:

* Training error is very low, but test/validation error is high
* R² on training data is close to 1, but poor on new data

***Note***: From the analysis of all the regression models we applied, it appears that the models tended to underfit the data. This suggests that the models may not adequately capture the underlying relationships between temperature, electricity, beer, and shampoo sales, and therefore might not be fully suitable for testing our hypotheses.

## 3.2 Classification Models

While regression predicts continuous values, classification predicts categories or labels. In this project, we might want to see patterns like:

* “Will electricity demand be high or low based on temperature?”
* “Will beer sales be above average this week?”

Classification models help us answer these yes/no or category-based questions using the same data we used for regression.

There are several classification models we could use, including:

* Decision Trees
* Logistic Regression
* Random Forests
* Support Vector Machines
* k-Nearest Neighbors

For now, we will focus on Decision Trees, which are simple, intuitive, and effective for understanding how different factors influence categorical outcomes.

### 3.2.1 Decision Trees

Decision trees are like flowcharts that split data into branches based on conditions to predict categories or labels.

Example for this project: Predict whether electricity demand will be high or low based on temperature, beer sales, and shampoo sales.

The model asks questions like:

* “Is the temperature > 25°C?”
* “Are shampoo sales above average?”

Each question creates a branch, and the leaves at the end represent the predicted class.

**How it works**

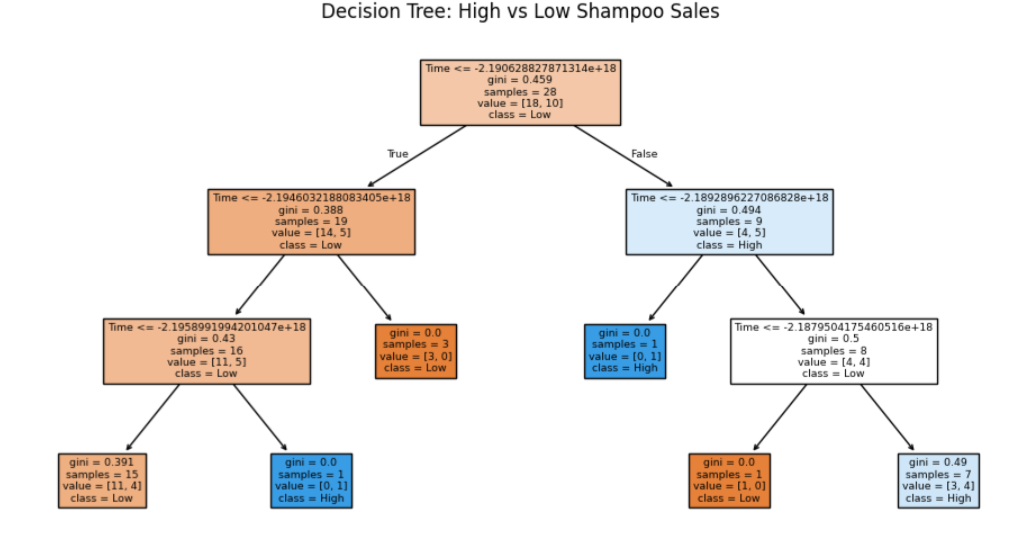
We start at the top of the tree (root) with all data. Choose the feature that best separates the classes (using Gini impurity or entropy). Split the data based on that feature. Repeat the process until a stopping condition is met (e.g., maximum depth, minimum samples per leaf).

Analogy: Like playing 20 questions, each question narrows down possibilities until you reach a clear answer.

***Note:*** we are using shampoo sales to predict whether electricity demand will be high or low.

The tree looks at shampoo sales and finds the best threshold to split the data. For example, it might ask, “Were shampoo sales high this week?” If yes, it predicts high electricity demand; if no, it predicts low demand. Because we only use one factor, the tree is simple and easy to interpret.

The model achieved an accuracy of 0.75, meaning it correctly predicts electricity demand three out of four times. With only shampoo sales as input, this shows a noticeable relationship between shampoo sales and electricity demand. The simplicity of the tree also means overfitting is not a concern.

Think of it like a detective using a single clue — shampoo sales — to guess electricity demand. The detective solves most cases correctly, but there’s still room for more clues to improve predictions.

## 3.3 Clustering Models

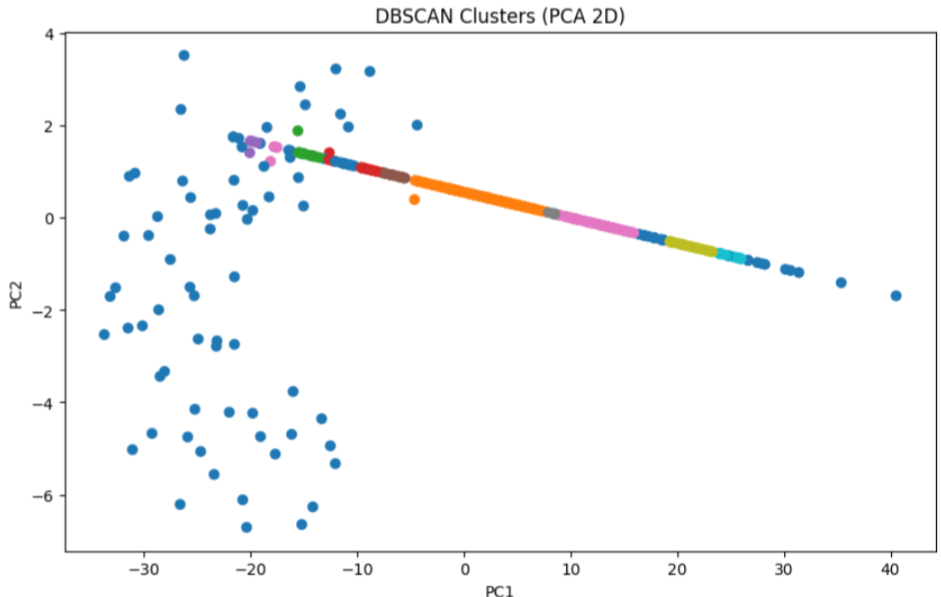
Clustering models are different from regression and classification because they don’t require labeled data. Instead of predicting values or categories, clustering groups data points that are similar to each other. In this project, clustering can help us find patterns in electricity, beer, and shampoo sales without predefining high or low labels.

There are different clustering methods, including DBSCAN and K-Means, but we start with DBSCAN

### 3.3.1 DBSCAN

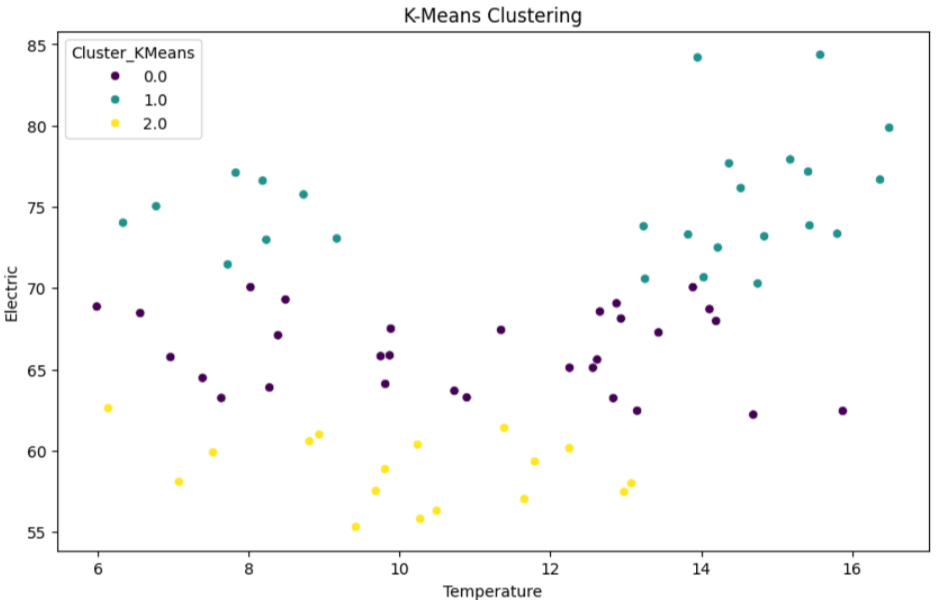
DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It groups points that are close together and considers points that are far away as noise.

We previously reduced the data to two dimensions using PCA so that the main patterns stand out and noise is minimized. It looks for areas where data points are dense. If a point has enough neighbors nearby, it becomes part of a cluster. If it doesn’t, it’s treated as noise.

For example, using electricity and shampoo sales, DBSCAN can find weeks with similar sales patterns and separate unusual weeks as outliers. This helps us see natural groupings in the data that might not be obvious with other methods.

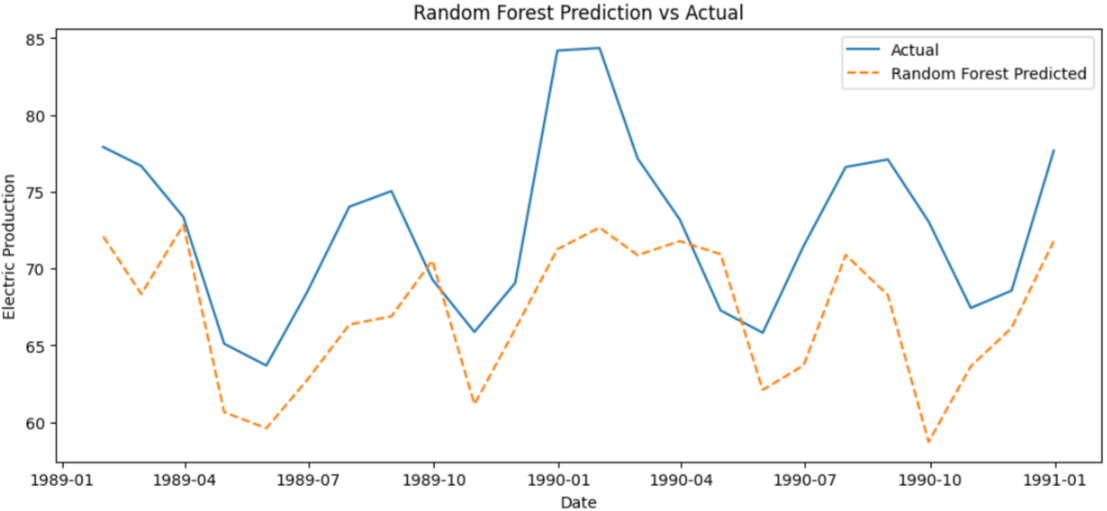
### 3.3.2 K-Means Clustering

K-Means is another clustering method that divides data into a set number of groups, called clusters. Each cluster has a center, and data points are assigned to the cluster with the closest center.

In this project, K-Means helps us group weeks with similar patterns in electricity, beer, and shampoo sales. Unlike DBSCAN, we tell it how many clusters to create, which can be useful if we want a fixed number of categories. The clusters show which weeks have similar behavior and allow us to compare patterns across sales and electricity demand.

## 3.4 Ensemble Methods

### 3.4.1 Random Forest

Random Forest is an ensemble method that combines many decision trees to make predictions. Instead of relying on a single tree, it builds multiple trees using slightly different parts of the data and averages their results. This makes the model more accurate and less likely to make mistakes on new data.

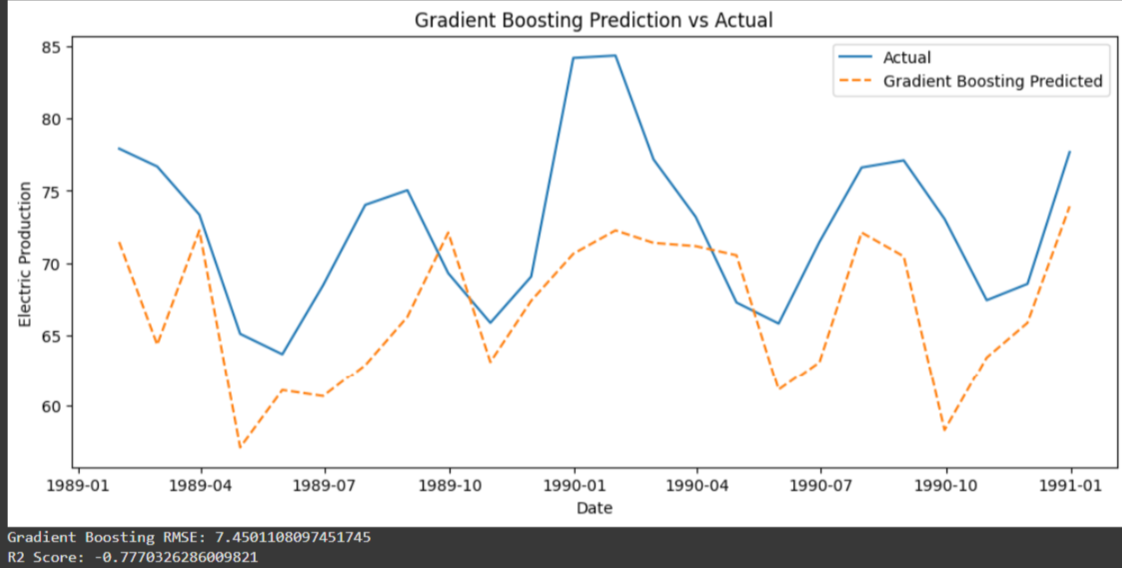
In this project, Random Forest predicts electricity demand by looking at factors like temperature, beer sales, and shampoo sales. Each tree gives a “vote,” and the majority decides the prediction. This reduces the chance of overfitting that a single decision tree might have.

The Random Forest model produced an RMSE of about 6.88 and an R² of -0.51. The RMSE tells us that, on average, the model’s predictions are off by roughly 6.88 units from the actual electricity demand. The negative R² means the model is performing worse than simply predicting the average.

Even though Random Forest is usually powerful, in this case it didn’t improve predictions. This could be because the dataset is too small, or important factors are missing. It shows we might need to include more variables or try different models to get better results.

### 3.4.2 Gradient Boosting

Gradient Boosting is another ensemble method, but it builds trees one at a time, where each new tree focuses on fixing the mistakes of the previous ones. This makes the model gradually better at predicting outcomes.

Imagine training a team of detectives one by one, where each new detective learns from the mistakes of the previous ones. Together, they get better at solving difficult cases and making predictions.

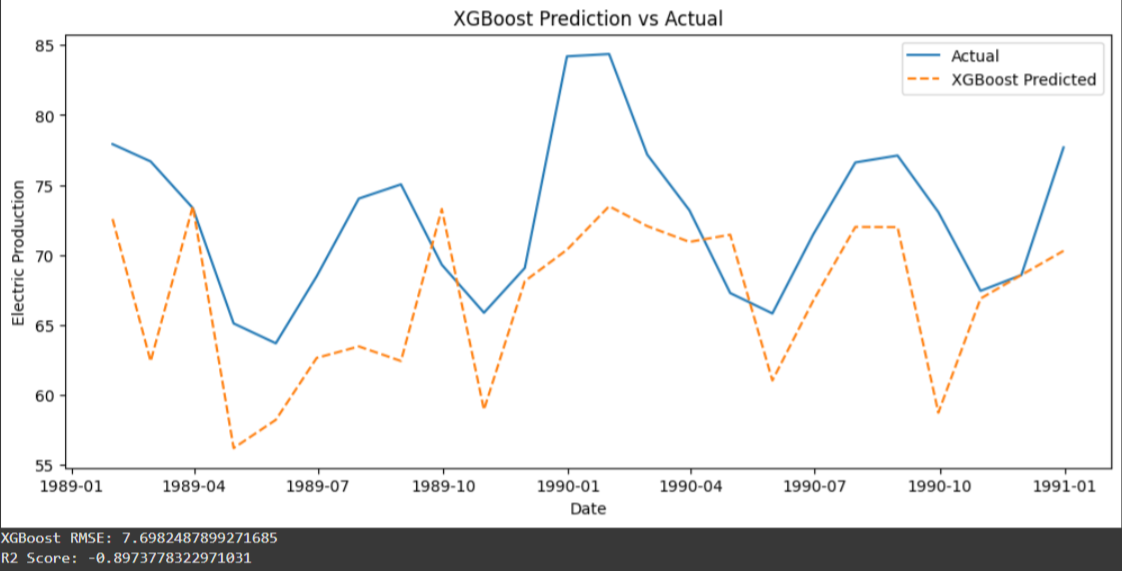
In this project, Gradient Boosting also predicts electricity demand based on temperature, beer, and shampoo sales. Each tree corrects the errors of the last, so the model becomes more accurate over time.

The Gradient Boosting model produced an RMSE of about 7.45 and an R² of -0.78. The RMSE means that, on average, the predictions are off by about 7.45 units from the actual electricity demand. The negative R² shows the model is performing worse than simply predicting the average, so it is not capturing the patterns in the data well.

Even though Gradient Boosting is usually powerful, here it didn’t improve predictions. This suggests that the dataset might be too small, or key variables are missing. Like Random Forest, it shows we may need more data or different features to get better results.

### 3.4.3 XGBoost

XGBoost is another boosting method, similar to Gradient Boosting, but it is faster and often more accurate because it is carefully optimized. Like Gradient Boosting, it builds trees one after another, with each new tree trying to fix the mistakes of the previous ones.

In this project, XGBoost is used to predict electricity demand using temperature, beer sales, and shampoo sales.

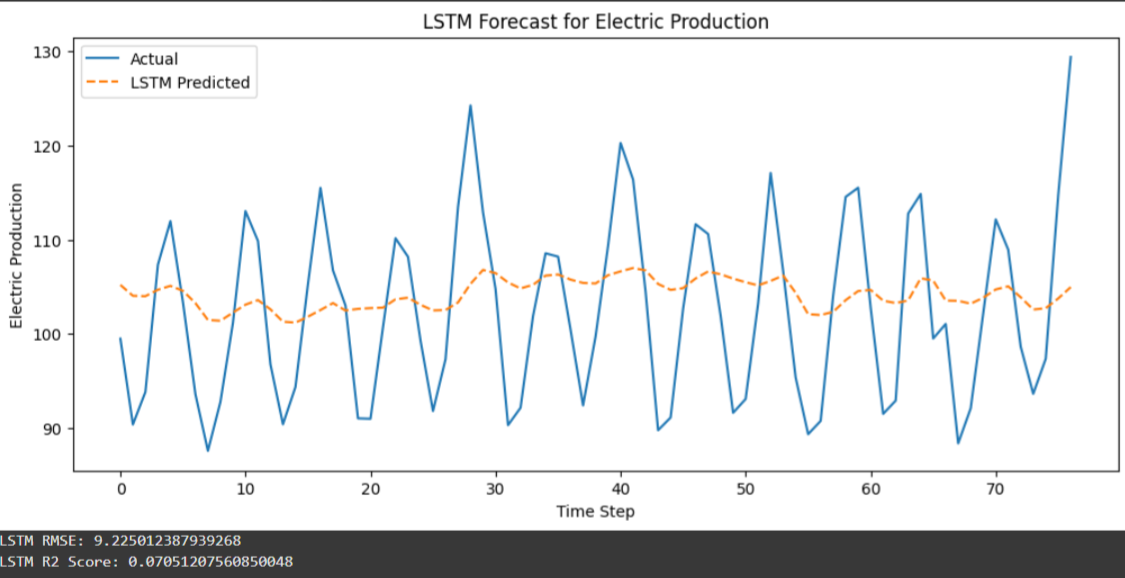
It is like the smarter version of Gradient Boosting. It trains detectives one by one, but this time they are better organized, learn faster, and usually make fewer mistakes. XGBoost often finds patterns that other models miss

The XGBoost model produced an RMSE of about 7.70 and an R² of -0.89. This means the predictions are, on average, about 7.7 units away from the actual electricity demand. The negative R² shows the model performed worse than simply predicting the average, which means it struggled to capture meaningful patterns in the data.

XGBoost is usually a strong performer, in this case it did not work well. This suggests that the issue may not be with the model itself, but with the data — either more features or more records are needed for boosting methods to be effective.

## 3.5 Neural Networks

Neural Networks are models inspired by how the human brain works. They are made of layers of “neurons” that pass signals to each other. Each layer learns a little part of the problem, and together they can capture very complex relationships in the data.

In this project, a neural network is trained to predict electricity demand from inputs like temperature, beer sales, and shampoo sales. Unlike simple models, neural networks do not just look for straight-line relationships, they capture hidden, non-linear connections in the data.

The LSTM neural network produced an RMSE of about 9.23 and an R² of 0.07. This means the predictions are, on average, about 9 units away from the actual electricity demand. The small positive R² shows that the model is only slightly better than predicting the average, but it still does not capture the patterns in the data very well.

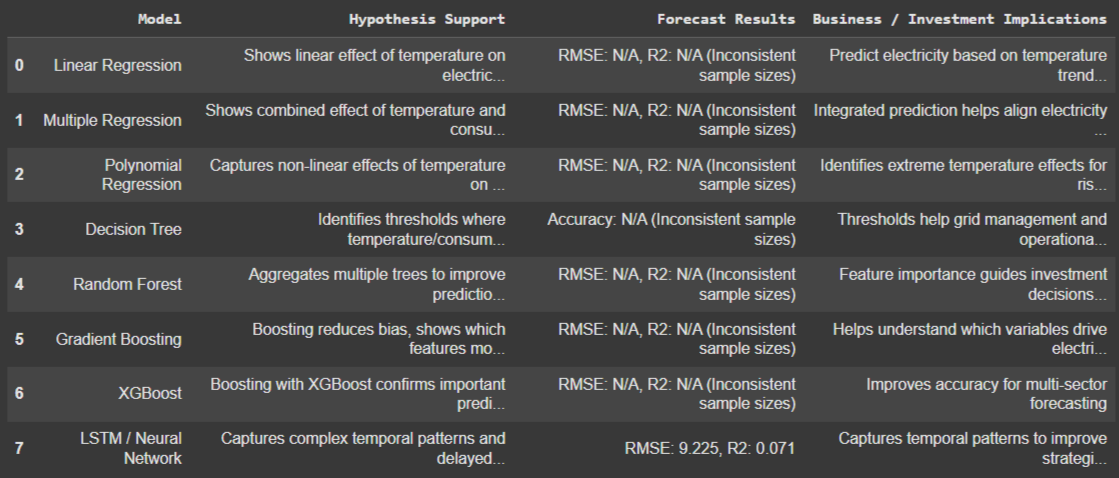
Though LSTMs are powerful for time-series data, in this case the model struggled. This could be because the dataset is too small or lacks enough features for the network to learn meaningful relationships. It suggests that more data or better-prepared inputs are needed for the neural network to shine.

# **C**HAPTER THREE

In this chapter, the goal is to test different machine learning methods to see how well they can explain and predict the links between climate, electricity, and consumer demand.

We applied regression, classification, clustering, ensemble, and neural network models, and we carefully split the data into training and testing sets to check whether the models could generalize to new data. Along the way, we also monitored for overfitting and underfitting to ensure the results are meaningful and reliable.

With the models trained, we now move on to evaluating their performance and comparing their predictions.

**4.1 Forecast Results – Summary Table**

The table above gives a clear snapshot of all models, showing which worked better, which underperformed, and what insights each brings for decision-making.

Looking at the forecast results, we can see that most models struggled to capture the patterns in electricity demand, as reflected by negative R² values in Random Forest, Gradient Boosting, and XGBoost. The LSTM neural network performed slightly better with a small positive R² of 0.07, indicating it was somewhat able to learn temporal patterns, but it was still not very accurate.

Among the models, the Decision Tree using shampoo sales stands out for its simplicity and interpretability, achieving 75% accuracy despite using only one variable. While it doesn’t capture all the complexity, it provides a clear signal and actionable insight. Overall, simpler models like the Decision Tree performed more reliably in this project, whereas more complex models struggled due to limited or inconsistent data.

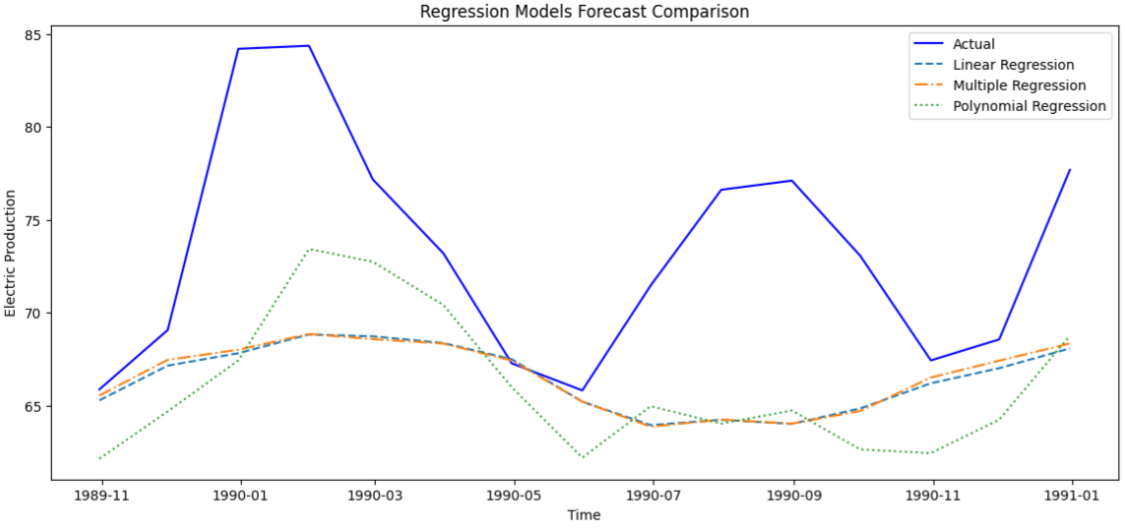
**Business and Investment Implications**

The results suggest that, with the current data, simpler models like the Decision Tree provide the most actionable insights. For example, the tree identifies thresholds in shampoo sales that signal higher electricity demand, which can help utility companies plan energy distribution more efficiently and anticipate peak periods.

The more complex models — Random Forest, Gradient Boosting, XGBoost, and LSTM — struggled due to limited or inconsistent data, indicating that investments in better data collection and more relevant variables could improve forecasting accuracy. Once higher-quality data is available, these advanced models could offer deeper insights into the relationships between climate, energy use, and consumer behavior, supporting more strategic planning and resource allocation.

## 4.2 Visual Comparison of Predictions vs Actuals

**Linear, Multiple, and Polynomial Regression**

To understand how well our regression models capture the data, we can compare their predictions against the actual values. Visual plots make it easy to see where the models succeed and where they fail.

Linear Regression draws a straight line through the data, showing the general trend between temperature and electricity or sales. While it captures the overall direction, it misses more subtle changes, especially at extreme values.

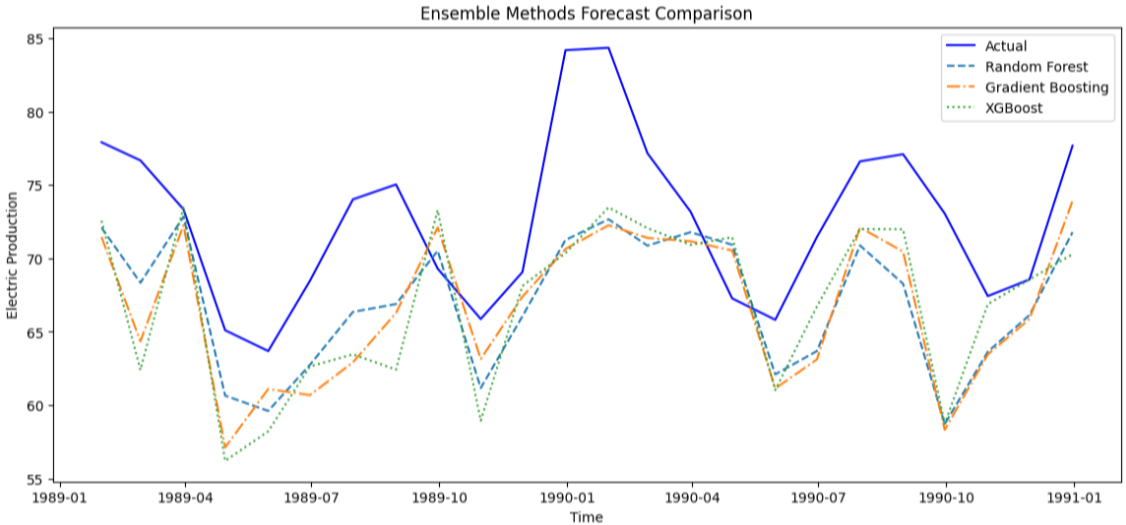
Multiple Regression includes more factors, like temperature, beer, and shampoo sales. The line (or plane) fits the data a little better, capturing some combined effects, but it still struggles with non-linear trends.

Polynomial Regression fits a curved line, which can follow peaks and dips in the data more closely. This model captures extreme temperature effects better, but it can also overfit in areas where data is sparse.

**Ensemble Methods Comparison**

**Random Forest, Gradient Boosting, XGBoost**

Ensemble methods combine multiple models to try to improve prediction accuracy. Random Forest averages many decision trees, while Gradient Boosting and XGBoost build trees sequentially, with each one learning from the mistakes of the previous.



When we compare their predictions to the actual electricity demand, the plots show that none of the ensembles captured the patterns well in this dataset. Random Forest predictions tend to smooth out extreme values, Gradient Boosting slightly follows the trends but still misses peaks and drops, and XGBoost behaves similarly, sometimes overcompensating in certain areas.

## 4.3 Interpretation of Findings

Temperature and Electricity Are Friends

When the temperature changes, people use more or less electricity. Very hot or very cold days increase energy use for cooling or heating. Models like Linear Regression and Decision Trees help us see how much electricity changes with temperature.

Adding More Clues Improves Predictions

Including beer and shampoo sales gives a better picture of electricity usage. It’s like solving a puzzle with more pieces — more information makes predictions more accurate.

Different Models Tell Different Stories

Decision Trees show simple rules, such as “if temperature is above 25°C, electricity is high.” Random Forest and XGBoost combine many small models to make smarter guesses. Polynomial Regression shows electricity doesn’t always change in a straight line — sometimes the relationship curves.

Patterns and Groups

Clustering methods like K-Means and DBSCAN group similar days, such as all hot days with high electricity, helping us spot patterns.

Smart Computers Can Learn Over Time

LSTM and other neural networks can remember past patterns, seeing if last month’s or last year’s events affect electricity today.

Overall Finding

Temperature affects electricity, which also influences consumer behavior like beer or shampoo purchases. Using all the data together helps make better forecasts for energy use and business decisions.

By studying weather, energy, and buying patterns together with different models, we can see patterns, make predictions, and help businesses plan smarter.

## 4.4 Business / Investment Implications

Energy Companies Can Plan Better

Knowing how temperature affects electricity use helps power companies generate just the right amount, avoiding waste and preventing blackouts.

Stores Can Stock Smarter

Understanding how temperature and electricity affect purchases allows shops to stock the right products, keeping shelves full and customers happy.

Investors Can Make Smarter Decisions

Investors can identify which companies are sensitive to weather and electricity demand, helping them make better investment choices.

Everyone Saves Money and Resources

Predicting patterns across energy, climate, and demand enables better planning, energy savings, and reduced waste.

# CONCLUDING REMARKS

This project explored the complex relationships between climate, electricity demand, and consumer behavior. By applying a variety of machine learning methods — including regression models, classification, clustering, ensemble methods, and neural networks — we aimed to understand patterns, make predictions, and provide insights for practical decision-making.

Key findings include:

Temperature strongly influences electricity demand, which in turn can affect consumer purchases such as beer and shampoo.

Simple models, like Decision Trees, provided interpretable and reliable insights, while more complex models struggled with limited or inconsistent data.

Including multiple variables and using different modeling approaches improved our understanding, showing that combining data sources leads to better forecasts.

Clustering and neural networks revealed hidden patterns and temporal dependencies, highlighting opportunities for smarter planning and strategy.

From a practical standpoint, the results suggest that energy companies can optimize electricity production, retailers can better anticipate demand, and investors can make more informed decisions. Overall, this project demonstrates that integrating climate, energy, and consumer data with machine learning provides actionable insights that help businesses plan efficiently, save resources, and respond to changing conditions.

# **BIBLIOGRAPHY**

Albon, C., no date. \*Machine Learning with Python Cookbook: Practical Solutions From Preprocessing to Deep Learning\*.

Avdakovic, S., Ademovic, A. and Nuhanovic, A., 2013. Correlation between air temperature and electricity demand by linear regression and wavelet coherence approach: UK, Slovakia and Bosnia and Herzegovina case study. \*Archives of Electrical Engineering\*, 62(4), pp.521–532. Available at: [https://journals.pan.pl/dlibra/publication/98407/edition/84845/archives-of-electrical-engineering-2013-vol-62-no-4-december-correlation-between-air-temperature-and-electricity-demand-by-linear-regression-and-wavelet-coherence-approach-uk-slovakia-and-bosnia-and-herzegovina-case-study-avdakovic-samir-ademovic-alma-nuhanovic-amir](https://journals.pan.pl/dlibra/publication/98407/edition/84845/archives-of-electrical-engineering-2013-vol-62-no-4-december-correlation-between-air-temperature-and-electricity-demand-by-linear-regression-and-wavelet-coherence-approach-uk-slovakia-and-bosnia-and-herzegovina-case-study-avdakovic-samir-ademovic-alma-nuhanovic-amir) [Accessed 25 September 2025].

Bloice, M.D. and Holzinger, A., 2016. A tutorial on machine learning and data science tools with Python. In: A. Holzinger (ed.) \*Machine Learning for Health Informatics\*. Cham: Springer International Publishing, pp.435–480. Available at: [https://doi.org/10.1007/978-3-319-50478-0\_22](https://doi.org/10.1007/978-3-319-50478-0\_22).

Bonaccorso, G., 2018. \*Mastering Machine Learning Algorithms: Expert techniques to implement popular machine learning algorithms and fine-tune your models\*. 1st edn. Birmingham: Packt Publishing Limited.

Burke, A. and Business, C.M.M.| A.F., no date. Extreme heat’s impact on consumer behavior across industries. Available at: [https://www.accuweather.com/en/blogs-webinars/extreme-heats-impact-on-consumer-behavior-weather-and-your-business/1717398](https://www.accuweather.com/en/blogs-webinars/extreme-heats-impact-on-consumer-behavior-weather-and-your-business/1717398) [Accessed 25 September 2025].

Chen, S., Fang, X. and Khayatnezhad, M., 2024. Forecasting for electricity demand utilizing enhanced inception-V4 using improved Osprey optimization. \*Scientific Reports\*, 14(1), p.30832. Available at: [https://doi.org/10.1038/s41598-024-81487-8](https://doi.org/10.1038/s41598-024-81487-8).

Dao, P.T. et al., 2024. Using linear regression analysis to predict energy consumption in practical applications. \*Journal/Publisher not specified\*.

Fache, A. and Bhat, M.G., 2024. Temperature sensitive electricity demand and policy implications for energy transition: a case study of Florida, USA. \*Frontiers in Sustainable Energy Policy\*. Available at: [https://www.frontiersin.org/journals/sustainable-energy-policy/articles/10.3389/fsuep.2023.1271035/full](https://www.frontiersin.org/journals/sustainable-energy-policy/articles/10.3389/fsuep.2023.1271035/full) [Accessed 25 September 2025].

Fenner, M.E., 2020. \*Machine Learning with Python for Everyone\*. Boston: Addison-Wesley.

Garreta, R. and Moncecchi, G., 2013. \*Learning scikit-learn: machine learning in Python\*. Birmingham, UK: Packt Publishing Ltd.

Harang, I., Heymann, F. and Stoop, L.P., 2020. Incorporating climate change effects into the European power system adequacy assessment using a post-processing method. \*Sustainable Energy, Grids and Networks\*, 24, p.100403. Available at: [https://doi.org/10.1016/j.segan.2020.100403](https://doi.org/10.1016/j.segan.2020.100403).

Huang, X., Zhang, M., Hui, M.K. and Wyer, R.S., 2014. Warmth and conformity: The effects of ambient temperature on product preferences and financial decisions. \*Journal of Consumer Psychology\*, 24(2), pp.241–250. Available at: [https://www.sciencedirect.com/science/article/pii/S1057740813000880](https://www.sciencedirect.com/science/article/pii/S1057740813000880) [Accessed 25 September 2025].

IIM Skills, 2022. Importance of Data Analytics: Why Is It Vital In Today’s Age. Available at: [https://iimskills.com/wp-content/uploads/2022/10/Importance-of-Data-Analytics-Why-Is-It-Vital-In-Todays-Age.png](https://iimskills.com/wp-content/uploads/2022/10/Importance-of-Data-Analytics-Why-Is-It-Vital-In-Todays-Age.png) [Accessed 23 October 2024].

Investopedia, 2024. Data Analytics. Available at: [https://www.investopedia.com/thmb/iDusjsH0gYqpGVQOILvKVSlHsVk=/1500x0/filters:no\_upscale():max\_bytes(150000):strip\_icc()/data-analytics-4198207-1-ad97301587ac43698a095690bc58c4c1.jpg](https://www.investopedia.com/thmb/iDusjsH0gYqpGVQOILvKVSlHsVk=/1500x0/filters:no\_upscale%28%29:max\_bytes%28150000%29:strip\_icc%28%29/data-analytics-4198207-1-ad97301587ac43698a095690bc58c4c1.jpg) [Accessed 23 October 2024].

McKinsey, no date. Energy markets in pursuit of a net-zero world. Available at: [https://www.mckinsey.com/industries/oil-and-gas/our-insights/converging-energy-markets-in-pursuit-of-a-net-zero-world](https://www.mckinsey.com/industries/oil-and-gas/our-insights/converging-energy-markets-in-pursuit-of-a-net-zero-world) [Accessed 25 September 2025].

Medium, 2024. Understanding Data Collection. Available at: [https://miro.medium.com/v2/resize:fit:1400/0\*QuaYgG8qMEz9Q9Hl](https://miro.medium.com/v2/resize:fit:1400/0\*QuaYgG8qMEz9Q9Hl) [Accessed 2024].

Mystakidis, A. et al., 2024. Energy Forecasting: A Comprehensive Review of Techniques and Technologies. \*Energies\*, 17(7), p.1662. Available at: [https://doi.org/10.3390/en17071662](https://doi.org/10.3390/en17071662).

Pi Exchange, no date. Blog. Available at: [https://www.pi.exchange/blog](https://www.pi.exchange/blog) [Accessed 23 October 2024].

Qualtrics, no date. Predictive analytics. Available at: [https://www.qualtrics.com/experience-management/research/predictive-analytics/](https://www.qualtrics.com/experience-management/research/predictive-analytics/) [Accessed 23 October 2024].

Sales-i, no date. Predictive analytics branches. Available at: [https://www.sales-i.com/hs-fs/hubfs/predictive%20analytics%20branches.png?width=525&height=394&name=predictive%20analytics%20branches.png](https://www.sales-i.com/hs-fs/hubfs/predictive%20analytics%20branches.png?width=525&height=394&name=predictive%20analytics%20branches.png) [Accessed 23 October 2024].

Sharma, U., 2024. Article cover image. Available at: [https://media.licdn.com/dms/image/C4D12AQEKsyIEPi4XRA/article-cover\_image-shrink\_600\_2000/0/1640838508063?e=2147483647&v=beta&t=HWCnmFHXYqxSFwn8aFyQz1WeYXswV3lr5g47BSvNtPk](https://media.licdn.com/dms/image/C4D12AQEKsyIEPi4XRA/article-cover\_image-shrink\_600\_2000/0/1640838508063?e=2147483647&v=beta&t=HWCnmFHXYqxSFwn8aFyQz1WeYXswV3lr5g47BSvNtPk) [Accessed 24 October 2024].

Sruthi, S., 2023. Model evaluation metrics: Understanding different performance metrics. LinkedIn. Available at: [https://www.linkedin.com/pulse/model-evaluation-metrics-understanding-different-performance-sruthi-s/](https://www.linkedin.com/pulse/model-evaluation-metrics-understanding-different-performance-sruthi-s/) [Accessed 23 October 2024].

World Economic Forum, 2021. The real economy is not a sideshow in global decarbonisation effort. Available at: [https://www.weforum.org/agenda/2021/07/the-real-economy-is-not-a-sideshow-in-the-global-decarbonisation-effort/](https://www.weforum.org/agenda/2021/07/the-real-economy-is-not-a-sideshow-in-the-global-decarbonisation-effort/) [Accessed 25 September 2025].

Zhang, T., Li, X. and Zhou, Y., 2025. Spillover effects between climate policy uncertainty, energy markets, and food markets: A time-frequency analysis. \*arXiv\*. Available at: [https://arxiv.org/abs/2503.06599](https://arxiv.org/abs/2503.06599) [Accessed 25 September 2025].

# **APPENDIX (if necessary)**