Tennessee Energy Insights: A Predictive Analytics Project for Forecasting Residential Energy Consumption

Hamza Salah

DSC630 – Predictive Analytics

Professor Andrew Hua

Spring 2025

Table of Contents

Executive Summary

Tennessee Energy Insights is a proposed initiative that will leverage data analytics to examine and optimize residential electricity consumption across the state of Tennessee. Given the growing energy demand driven by new infrastructure developments, such as Elon Musk's factory in the region and the impact of shifting climate patterns, understanding the drivers of electricity usage has become a vital priority for utility companies and policy makers alike.

In this project, time series forecasting will be a key technique used to predict future energy consumption based on historical data. Time series forecasting helps to model and predict the behavior of energy consumption over time by capturing underlying trends, seasonality, and other patterns that can influence demand. The primary focus will be on using historical temperature and electricity pricing data to develop a predictive model.

The anticipated outcomes of this project aim to provide actionable insights for utility providers, policymakers, and energy planners in Tennessee. By uncovering the factors that most significantly drive energy demand, stakeholders will be better equipped to adjust pricing strategies, improve demand forecasting, and implement targeted energy efficiency programs. Ultimately, this initiative seeks to support more sustainable energy practices and inform future infrastructure and policy decisions in the state of Tennessee.

Introduction

Project Redirection

Originally focused on forecasting e-commerce sales, the project was refocused in Milestone 3 to explore residential energy consumption in Tennessee. This pivot was driven by a desire to work with a more realistic and socially relevant scenario, one that aligns more closely with current energy concerns in the region. It also takes advantage of publicly available environmental and economic data. While the topic has shifted, the core regression modeling framework remains consistent and has been adapted to this new context.

Research Problem

This research addresses the rise in electricity consumption in the state of Tennessee, considering recent developments such as Elon Musk's new project, which is expected to drive increased energy demand through expanded industrial operations and technological infrastructure.

Purpose of the Study

This research aims to identify patterns in energy consumption across different sectors in the state of Tennessee, with a particular focus on the role of average temperature as a driving factor.

Significance of the Research

The findings from this project can help utility companies and energy planners better understand how consumption varies by sector and how it correlates with temperature trends. This insight can support more efficient energy distribution, policy planning, and sustainability efforts.

Origin

The project originated from a data science initiative focused on energy forecasting and climate-related consumption patterns. It centers on building accurate predictive models and uncovering sector-specific trends to inform future decision-making.

Stakeholders

Key stakeholders include city officials, utility companies, energy policy makers, and community planning agencies, all of whom have an interest in managing energy resources effectively.

Scope

The study covers energy consumption data from 2010 to 2024 across the Residential, Commercial, and Industrial sectors in Tennessee. Limitations include incomplete data for the Transportation sector, potential reporting inconsistencies, and the assumption that temperature is the primary external variable influencing consumption.

Data Collection and Structure

Data Sources

The project will use two key datasets to analyze residential energy consumption in Tennessee:

1. **Electricity Consumption Data (Sales):**
   This dataset retrieved from EIA.gov provides monthly electricity consumption data (in kWh) for residential customers. "Sales" refer to the total electricity sold to end-users, which serves as a proxy for energy consumption during the recorded period.
2. **Temperature Data :**
   Monthly temperature data (in degrees Fahrenheit) across Tennessee retrieved from NOAA will be used to examine how temperature fluctuations affect energy usage. This dataset will help assess the climatic impact on consumption.

The datasets will be merged to align monthly consumption data with the corresponding temperature and electricity price data, creating a comprehensive dataset for regression analysis.

Variables Description

The dataset comprises columns merged from multiple datasets, with additional variables retained specifically for usage determination during regression analysis.

- **Year**: Represents the year in which the data was recorded.
- **Month**: Represents the month in which the data was recorded.
- **Average_Temp**: Represents the average temperature during the recorded period.
- **Revenue**: Total revenue generated from electricity sales during the recorded period.
- **Sales (Consumption)**: Total amount of electricity sold during the recorded period, typically measured in kilowatt-hours (kWh) or similar units.

- **Customers**: Number of residential customers who purchased electricity during the recorded period.
- **Price**: Average price charged for electricity during the recorded period, typically measured in cents per kilowatt-hour (¢/kWh) or a similar unit.

Data Cleaning and Preparation

Describes the steps taken to handle missing or inconsistent data. The research questions will make more sense once the reader understands the data collection and cleaning process.

1. **Renaming Columns**:
   • Renamed columns to make them more meaningful and align them with the relevant sector names.
   o e.g., RESIDENTIAL → res_rev, COMMERCIAL → com_rev, etc.
2. **Dropping Rows and Columns**:
   • Dropped the first two rows (index=[0,1]) and the status column since they contained irrelevant information.
   • Removed the last row, which contained footer data not relevant to the analysis.
3. **Resetting Index**:
   • After dropping rows, reset the index of the DataFrame to ensure consistency.
4. **Converting Date Column**:
   • Combined the year and month columns to create a new date column, converting it into a datetime object with the day set to 1.
5. **Filtering Data for Temperature Data**:
   • Applied mini lambda functions to convert data into strings or integers and split the date column into year and month.
   • Renamed columns in the temperature DataFrame (df_temp) for clarity and restructured the date column into a proper datetime object.
6. **Filtering for Tennessee**:
   • Filtered the dataset to only include rows for Tennessee (df['state'] == 'TN'), as the project focuses on this state.

7. **Merging DataFrames**:
   • Merged the temperature data (df_temp) with the main data (df) on the date column to combine both datasets.

8. **Splitting by Sector**:
   • Created separate DataFrames for each sector: residential, commercial, industrial, transportation, and total, to analyze them individually.

9. **Renaming Columns in Sector DataFrames**:
   • Renamed the relevant columns in each sector DataFrame to general terms (e.g., revenue, consumption, customers, price), removing the sector-specific abbreviations.

10. **Standardizing and Transforming Sector Columns**:
    • Standardized the consumption and avg_temp columns in each sector DataFrame using a scaler (e.g., StandardScaler).
    • Created new columns consumption_scaled and avg_temp_scaled to store the transformed values.

Research Questions

Primary Research Question

How does average temperature and electricity pricing impact residential and sectoral electricity consumption patterns in Tennessee, and how can predictive analytics be used to forecast future consumption trends?

Secondary/Sub-Questions

- How does temperature impact electricity consumption in residential, commercial, industrial, and transportation sectors in Tennessee?

Methodology

Tools and Techniques

The following tools, programming languages, and frameworks were used throughout the analysis:

- **Python**: The primary programming language used for data analysis and modeling.
- **Pandas & NumPy**: For data wrangling, manipulation, and numerical computations.
- **Matplotlib & Seaborn**: For creating detailed visualizations and residual diagnostics.
- **Statsmodels**: To build ARIMA models, assess autocorrelation, and conduct time series diagnostics.
- **pmdarima**: Used for automatic ARIMA model selection (auto_arima).
- **Scikit-learn**: Employed for calculating evaluation metrics (e.g., RMSE, MAE, MAPE) and standardizing data where necessary.

Models Used

The primary model chosen for this analysis was **ARIMA (Autoregressive Integrated Moving Average)**.

- **Seasonality and Trends**: ARIMA is particularly effective for time series data where trends and seasonality are prominent, which is the case with energy consumption data.
- **Univariate Focus**: While MLR models multiple predictors, ARIMA is well-suited for univariate forecasting where the focus is on modeling the time-dependent structure of the data.
- **Accurate Forecasting**: ARIMA provides robust predictions for future consumption based on historical data, making it more appropriate for the forecast of energy consumption in this study.

Auto ARIMA (via pmdarima) was used to automate the selection of optimal ARIMA parameters (p, d, q), ensuring the best possible fit.

Diagnostic Checks: Residual analysis (histograms, Q-Q plots, ACF plots, and Ljung-Box test) was performed to ensure model validity and reliability.

Forecast Evaluation Metrics: RMSE, MAE, and MAPE were employed to assess the model's forecasting performance.

Evaluation Metrics

- **Mean Squared Error (MSE):** This will measure the average squared difference between the predicted and actual energy consumption values, providing insight into the accuracy of the regression model.
- **Mean Absolute Error (MAE):** This metric will provide a clearer interpretation of the average absolute difference between predicted and actual values in the same units as the target variable (kWh). It is useful for understanding the average magnitude of error in predictions.
- **Root Mean Squared Error (RMSE):** This metric, the square root of MSE, gives a more interpretable result in the same units as the target variable (kWh). RMSE is sensitive to large errors, making it useful for identifying significant prediction mistakes.
- **Mean Absolute Percentage Error (MAPE):** MAPE measures the average absolute percentage difference between predicted and actual values. It is helpful for understanding the relative error, especially when comparing models across different datasets or when absolute values are less important than proportional errors.
- **Coefficient of Determination (R Squared):** Represents the proportion of variance in the dependent variable that is predictable from the independent variables.

Justification of Methods

- ➢ **Auto ARIMA (via pmdarima)**

- **Reason for Use**: To ensure that the optimal parameters for the ARIMA model (p, d, q) were selected, the Auto ARIMA function was employed. This tool automates the model selection process based on statistical criteria such as AIC and BIC, which reduces the need for manual tuning and ensures the best fit for the data.

- **Diagnostic Checks (Residual Analysis, ACF, Q-Q Plot, Ljung-Box Test)**

- **Reason for Use**: These diagnostic checks were crucial for validating the assumptions of the ARIMA model and ensuring that it was appropriately capturing the underlying data structure. By examining the residuals, the model's predictive accuracy and adherence to assumptions (e.g., normality, no autocorrelation) could be assessed.

- **Evaluation Metrics (MSE, MAE, RMSE, MAPE, R2)**

- **Reason for Use**: These metrics were chosen because they provide both absolute and relative measures of model performance. MSE and RMSE give insights into the magnitude of prediction errors, while R^2 quantifies how well the model explains the variance in energy consumption. MAE and MAPE offer more intuitive, interpretable results for understanding prediction accuracy.

Data Analysis

Exploratory Data Analysis (EDA)

The dataset summarizes energy use across five sectors from 2010 to 2024. The Residential sector leads in revenue and consumption, followed by Commercial and Industrial, with Industrial offering the lowest prices. Transportation shows minimal activity. Combined sector data reflects overall averages: ~$803K revenue, ~8.3M units consumed, and 3.3M customers. Temperature remains steady (~59°F), while consumption and revenue vary widely across sectors.

Sector	Avg Temp (°F)	Revenue (Mean)	Consumption (Mean)	Customers (Mean)	Price (Mean)
Residential	59.10	$379,067	3,521,282	2,869,945	$10.80
Commercial	59.10	$300,732	2,811,132	492,098	$10.66
Industrial	59.10	$123,227	1,983,717	1,275	$6.16
Transportation	59.10	$5.08	45.66	0.33	$3.66
Combined	59.10	$803,031	8,316,176	3,363,320	$9.64

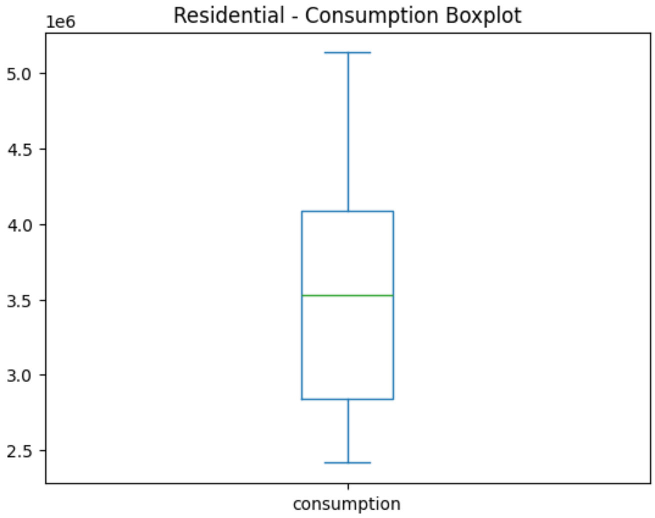
Techniques

In this project, I used a combination of **data cleaning**, **descriptive statistical analysis**, and **exploratory data analysis (EDA)** techniques. I summarized key metrics like mean, median, standard deviation, and ranges for variables such as revenue, consumption, customers, and price across different sectors. I also grouped data by sector and time to identify patterns, trends, and variability. These methods helped you compare energy usage behavior across sectors and uncover sector-specific insights.

The code I used for this project is listed in Appendix C (Code 1-17)

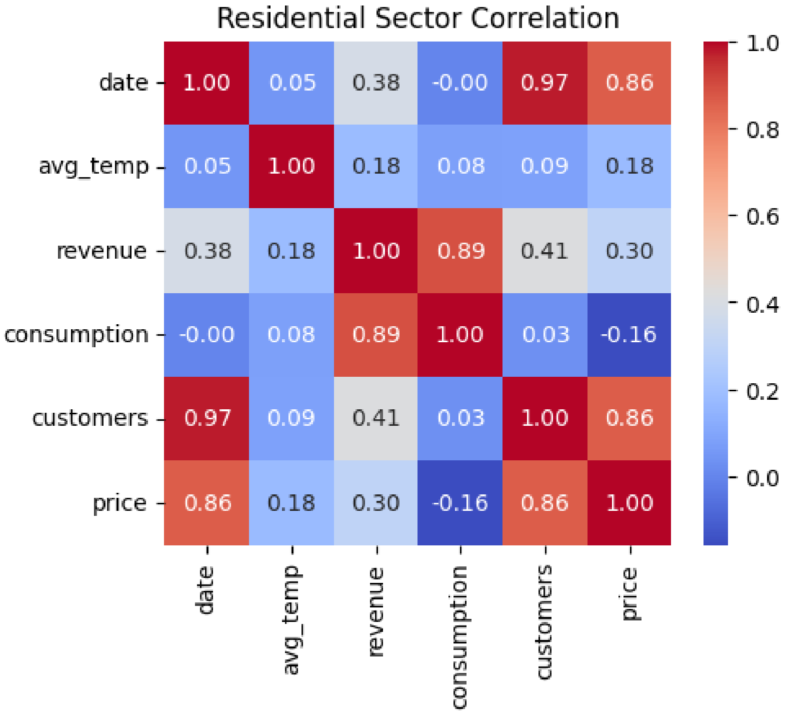
Visualizations

To check for outliers, I created a for loop that iterates over a list of DataFrames and outputs a boxplot. The plot below, which represents residential data, shows no outliers. The boxplots for the remaining sectors can be found in Appendix A (Figures 1-4).



The boxplots revealed four outliers in the commercial and industrial sectors. After reviewing the consumption data for these sectors, the values appeared consistent with typical usage patterns for the month of August. Since I planned to perform ARIMA modeling, I chose to retain these data points to capture the seasonal variation rather than remove them.

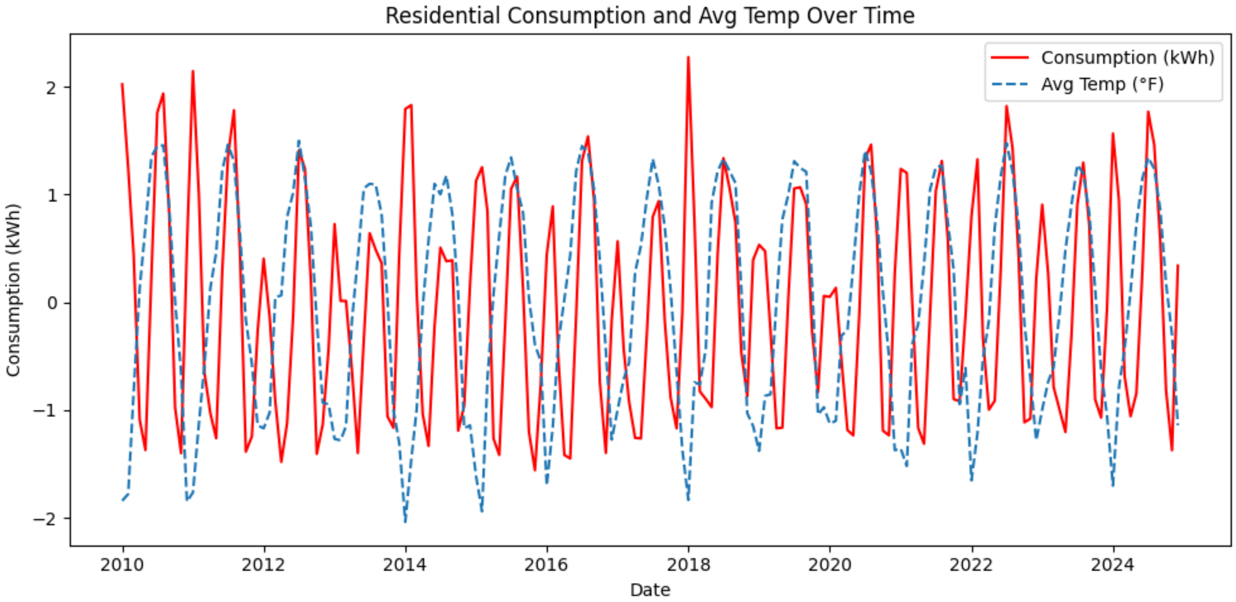
To analyze the relationship between electricity consumption and average temperature in the state of Tennessee, I created a correlation heatmap to visualize the strength of their correlation.



The correlation heatmap shows an 8% positive correlation which is very low, the other sectors in Appendix A (Figures 5-9) show a higher percentage, 60% for commercial and 26% for industrial and 9% for transportation. Overall, when combining all sectors, the correlation increases to 35%, highlighting that temperature influences energy consumption across sectors, but the strength of this relationship varies significantly by sector.

Further investigating the relationship between them, I graphed a time series dual axis chart of average temperature and residential energy consumption from 2010 to 2024. This visualization revealed seasonal patterns and potential correlations, such as increased consumption during periods of extreme temperatures. Additionally, I observed trends over the years that suggest how climate variability and long-term weather changes may be influencing residential energy demand.

Extending the analysis to other sectors, I created similar visualizations for the commercial and industrial sectors. These graphs highlighted distinct consumption behaviors based on sector-specific patterns that required further research.



Residential consumption maintains a consistent relationship with temperature throughout the entire period, suggesting that weather-dependent usage—particularly HVAC—could be the primary driver of residential energy demand.

The other sectors shown in Appendix A (Figures 9–12) follow a similar pattern, apart from the Industrial sector, which displays a steady decline in consumption regardless of temperature. From 2014 onward, the correlation weakens, indicating that temperature may not be the dominant factor influencing industrial energy use in later years.

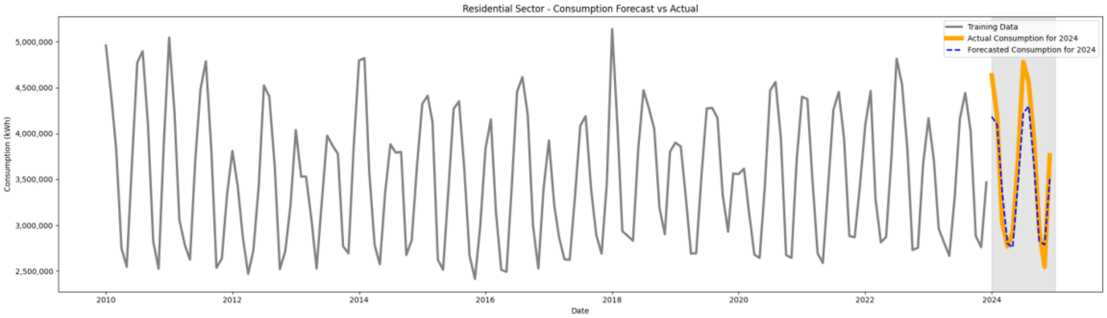
It is also worth noting that the transportation sector is missing a significant portion of data and will therefore be excluded from both analysis and model development.

Next, I built multiple regression models to understand the relationship between energy consumption and key predictors such as average temperature, price, and customer counts.

Number of Features	Features	R²	RMSE
1	['avg_temp']	0.0179	758,059
2	['avg_temp', 'price']	-0.211	841,776
3	['avg_temp', 'price', 'customers']	-0.1258	811,614

The regression analysis revealed that these variables, including temperature, alone had limited explanatory power, as evidenced by low R² values and high prediction errors, indicating that temporal dependencies and patterns were not adequately captured. Using this insight, I moved on to time series forecasting, where I successfully predicted energy consumption for the state of Tennessee in 2024 based on historical consumption data and temporal trends.

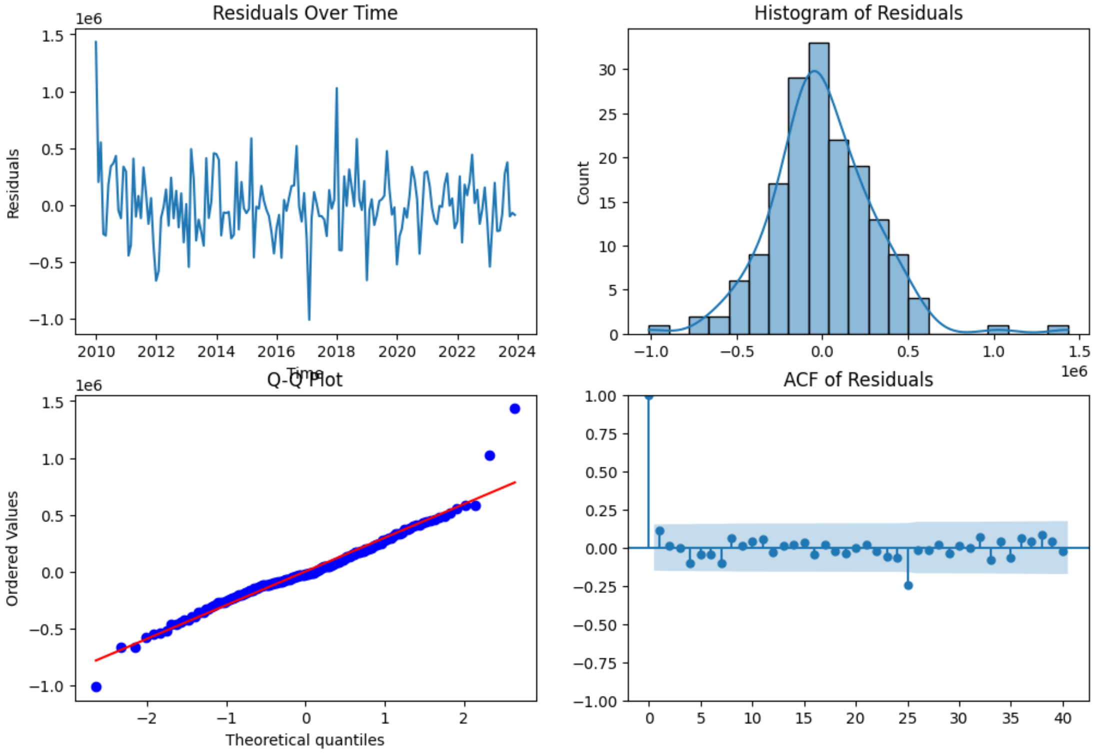
The next step in my analysis involved time series forecasting, where I successfully predicted energy consumption for the state of Tennessee in 2024 based on historical data. Forecasting results for each sector are presented in Appendix A (Figures 13-18), and the corresponding performance metrics tables can be found in Appendix B (Table 1-6).



The forecasted values (dashed blue line) closely follow the seasonal trend observed in previous years and align well with the actual 2024 consumption values, capturing both seasonality and trend.

Model Evaluation Metrics (Residential)			
MAE	MSE	RMSE	MAPE %
255290.6832	8.52E+10	291811.8391	0.06825

The Root Mean Squared Error (RMSE) is 291,812 kWh—a relatively low value given the overall consumption range of 2.5 to 5 million kWh—strong predictive performance.



- The residuals are centered around zero with no discernible patterns.
- The histogram of residuals appears roughly bell-shaped, suggesting normality.
- The Q-Q plot shows residuals falling mostly along the 45° line.
- The autocorrelation function (ACF) of the residuals displays no significant spikes, supporting the Ljung-Box test result and indicating no strong autocorrelation.

Key Findings

Summary of Results

The analysis revealed several key insights regarding energy consumption across different sectors in Tennessee:

- **Residential & Commercial Sectors**: Energy consumption shows a clear seasonal pattern, with temperature being a significant predictor of consumption. The ARIMA model successfully forecasted the 2024 consumption values with minimal deviations from the actual observed data.
- **Industrial Sector**: Unlike the residential sector, the industrial sector demonstrated a decline in consumption regardless of temperature variations, particularly after 2014. This suggests that factors other than temperature may influence industrial consumption in recent years.
- **Transportation Sector**: Most data for this sector were missing, and therefore, no analysis or forecasting was conducted for transportation energy consumption.
- **Model Performance**: The ARIMA model demonstrated strong predictive performance, with an RMSE of 291,812 kWh, indicating that the model is effectively capturing trends and seasonality in the data. The residuals analysis showed no strong patterns, confirming the model's reliability.

Interpretation of Findings

- **Residential & Commercial Sector Insights**: The strong correlation between residential and commercial energy consumption and temperature confirms the hypothesis that HVAC usage is a dominant driver of energy consumption in this sector. The seasonal fluctuations are well captured by the ARIMA model, providing a reliable forecasting tool for utility companies. This information is valuable for utility companies in predicting energy demand during peak seasons.
- **Industrial Sector Trends**: The steady decline in industrial energy consumption, especially post-2014, suggests a shift in industry practices or the adoption of energy-efficient technologies that are not influenced by weather. Understanding these trends is important for policy makers and utility

providers to account for changes in industrial demand and to adjust capacity planning accordingly.

- **Transportation Data Gaps**: The lack of sufficient data for the transportation sector highlights an area for improvement in data collection and monitoring. If data were more complete, this sector could provide further insights into its relationship with energy consumption and temperature.
- **Model Effectiveness**: The use of ARIMA for forecasting energy consumption proved effective in capturing the time-dependent patterns in the data. The low RMSE and well-behaved residuals indicate that the model has predictive value and can be used for short-term forecasting. This can aid in strategic planning for energy distribution and usage management.

Recommendations

Based on the findings from this analysis, I suggest the following practical actions for stakeholders in Tennessee's energy sector, particularly utility companies and policy makers:

1. **Customer Alerts for High Demand and Peak Hours**
    - **Recommendation**: Implement text message alerts to customers during high demand periods or peak hours. By notifying customers about the upcoming peak times, they can adjust their energy usage, potentially leading to a reduction in overall consumption.
2. **Incentive Programs for Energy-Efficient Practices**
    - **Recommendation**: Develop and promote programs that incentivize households to adopt energy-efficient appliances and practices. These programs could be tied to off-peak hours, encouraging consumers to use electricity more efficiently and reduce their overall consumption.
3. **Focus on Industrial Sector Efficiency**
    - **Recommendation**: Encourage industries to adopt more energy-efficient technologies through tax incentives or grants. Also, support the continued shift towards sustainability initiatives that reduce industrial energy consumption irrespective of temperature.
4. **Improve Data Collection for Transportation Sector**
    - **Recommendation**: Prioritize efforts to improve data collection for the transportation sector to better understand its relationship with energy consumption. This may include gathering more granular data on energy use across different transportation modes (e.g., electric vehicles, commercial fleets) and its dependence on temperature.

Practical Implications

These recommendations have several real-world applications:

- **Demand Management**: By alerting customers to peak consumption times, utilities can better manage demand, reduce the strain on energy infrastructure.

- **Environmental Impact**: Encouraging energy efficiency at both the residential and industrial levels could contribute to significant reductions in energy consumption and carbon emissions. This aligns with broader sustainability goals and helps meet regulatory requirements aimed at reducing the state's carbon footprint.
- **Cost Savings**: Both consumers and utilities stand to benefit from reducing energy usage during peak periods. For consumers, it may lead to lower electricity bills, while utilities can avoid the high costs associated with energy generation during peak demand.

Risks and Mitigation Strategies

Identified Risks

- **Data Quality:** One potential risk is the quality and completeness of the data. Missing or inconsistent data in either temperature or electricity consumption records could affect the model's accuracy. To mitigate this, we will implement comprehensive data cleaning techniques, including handling missing values through imputation and removing any outliers that could distort the model.
- **Model Overfitting/Underfitting:** With linear regression models, there's the risk of overfitting (if the model fits the training data too well) or underfitting (if the model does not capture the true patterns in the data). To address this, we will evaluate multiple models and use cross-validation techniques to ensure the selected model generalizes well to unseen data.
- **External Factors:** The model might not capture external factors such as holidays, special events, or regional economic shifts, which can influence energy consumption. These factors can lead to model bias or reduced accuracy.

Contingency Plan

If the ARIMA model does not yield satisfactory results, we will explore alternative non-linear models, such as Random Forest or Gradient Boosting Machines (GBM), which can better capture complex relationships between temperature, kWh cost, and other influencing factors. These models offer greater flexibility and may uncover interactions that ARIMA might not capture, especially if seasonal or non-linear trends are present. Additionally, we will consider enhancing the feature set by creating interaction terms between temperature and pricing, which could further improve model performance.

If temperature and price alone do not sufficiently explain energy consumption, we will expand the model by including additional features such as the number of customers, regional data, or other relevant demographic variables. This will help refine the model and improve its ability to explain variations in energy consumption across different time periods and regions.

Ethical Considerations

Data Privacy and Confidentiality

Since the dataset is aggregated at a state level and does not include personally identifiable information (PII), there are no data privacy concerns. However, care will be taken to ensure that the data used for the project is publicly available and adheres to ethical data usage standards.

Bias and Fairness

All assumptions made during the analysis, such as the choice of features or transformation methods, will be clearly communicated. Any conclusions derived from the model will be supported by data, and the limitations of the model (e.g., the inability to capture some external variables) will be openly discussed. This ensures that the findings are not misleading or overgeneralized.

Conclusion

Summary of Research

This study explored the patterns of energy consumption across various sectors in Tennessee, with a specific focus on the residential sector. By analyzing the relationship between temperature and energy consumption, the study revealed significant seasonal patterns, particularly in how weather drives residential energy demand through HVAC usage. The ARIMA model successfully forecasted energy consumption for 2024, demonstrating strong predictive performance. Based on relative RMSE, the commercial sector had the most accurate forecast (3.78%), followed by the combined (4.83%), residential (5.84%), and industrial (7.37%) sectors.

The study also identified that the industrial sector's energy consumption declined post-2014, suggesting other factors may now be more influential than temperature. The lack of sufficient data for the transportation sector was a key limitation, but also an area for future exploration. Overall, the research provides valuable insights into the dynamics of energy consumption and offers actionable recommendations for utility companies, policy makers, and other stakeholders to better manage energy demand, reduce consumption during peak periods, and lower carbon emissions.

Future Research Directions

Future work in this area could address several key areas for further investigation:

1. **Data Expansion for the Transportation Sector**: Collecting and analyzing more comprehensive data on transportation energy consumption, particularly as electric vehicles become more widespread, would provide valuable insights into how temperature impacts energy usage in this sector.

2. **Incorporating Economic and Technological Factors**: Future models could integrate additional variables, such as economic indicators (e.g., GDP growth, unemployment rates) and advancements in energy-efficient technologies, to better understand their impact on energy consumption across sectors.

3. **Long-Term Forecasting**: Extending the forecasting horizon beyond 2024 to assess long-term trends and the potential effects of climate change, policy changes, or energy infrastructure upgrades would help utilities better plan for the future.

References

U.S. Energy Information Administration (EIA). (2024). Monthly Energy Consumption Data. Link

National Centers for Environmental Information (NOAA). (2024). Monthly Average Temperature Data. Link

Appendices

Appendix A: Graphs

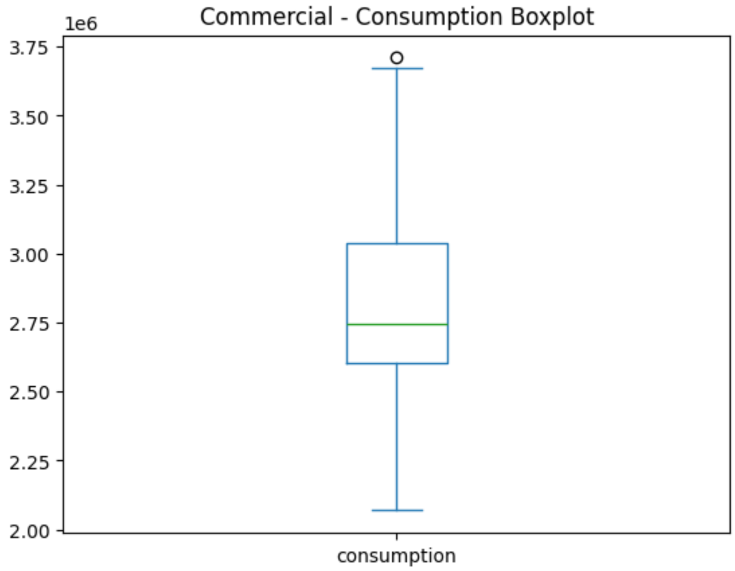


Figure 1: Box Plot (Commercial Sector)

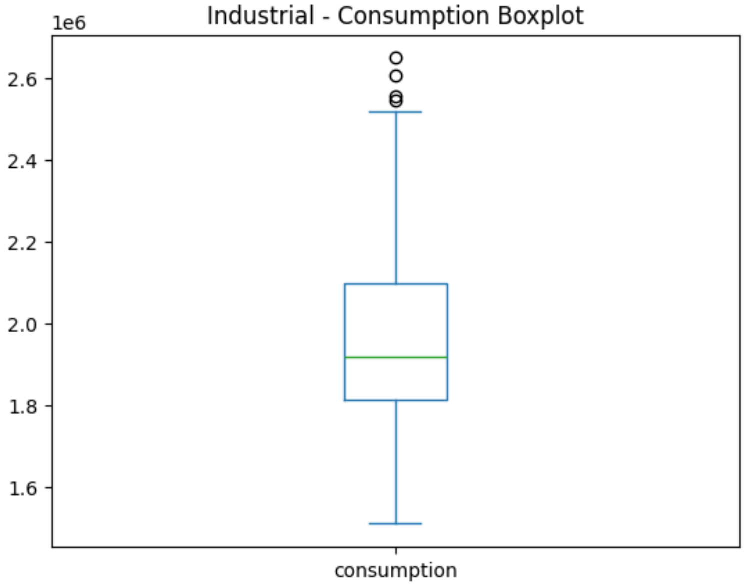


Figure 2: Box Plot (Industrial Sector)

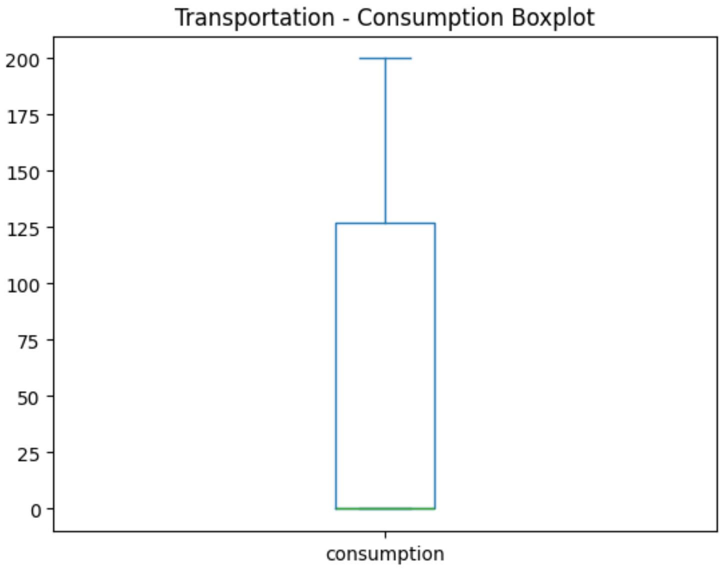


Figure 3: Box Plot (Transportation Sector)

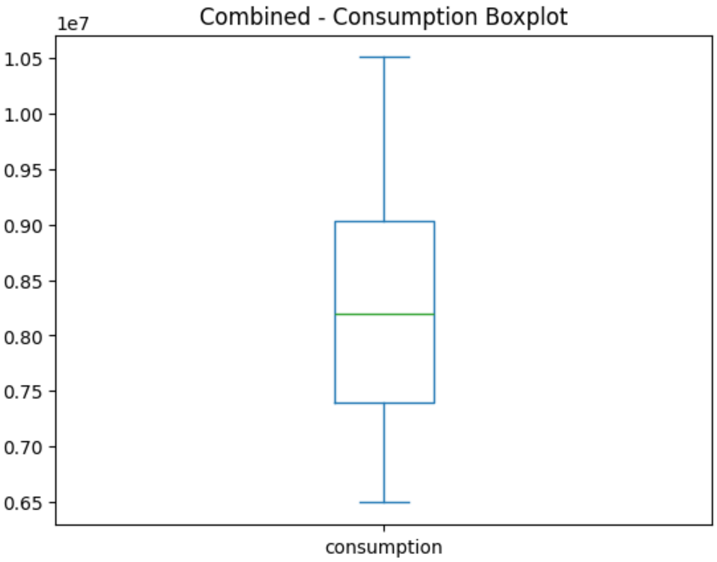


Figure 4: Correlation Matrix (All Sectors)

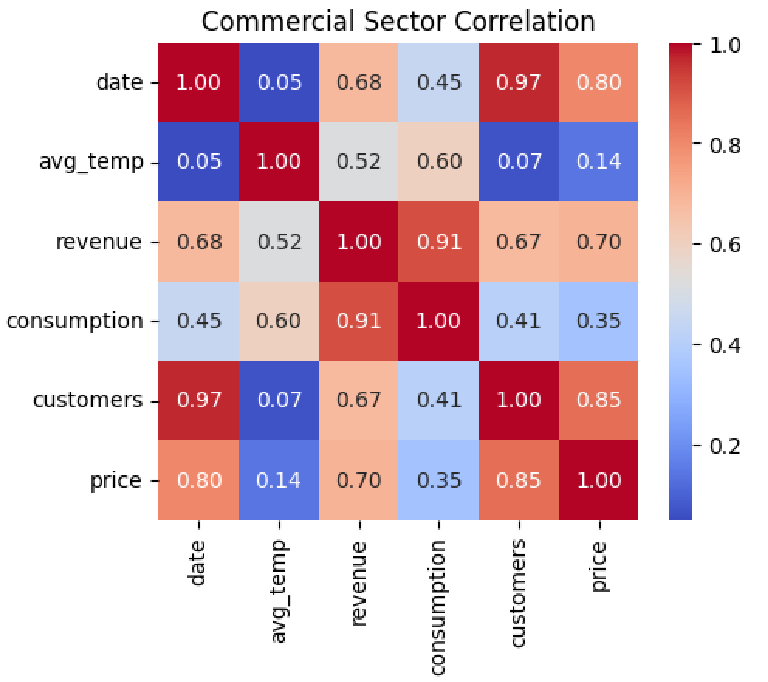


Figure 5: Correlation Matrix (Commercial Sector)

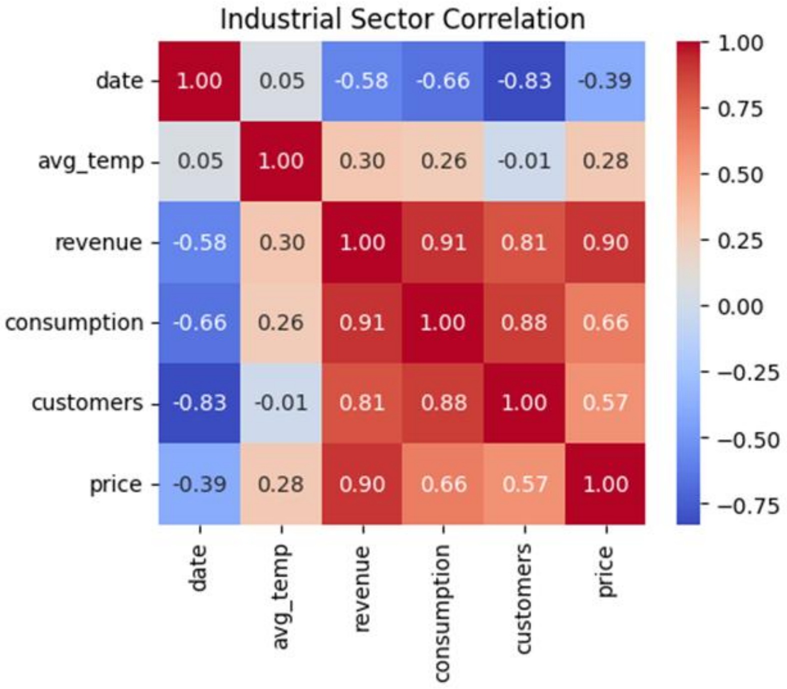


Figure 6: Correlation Matrix (Industrial Sector)

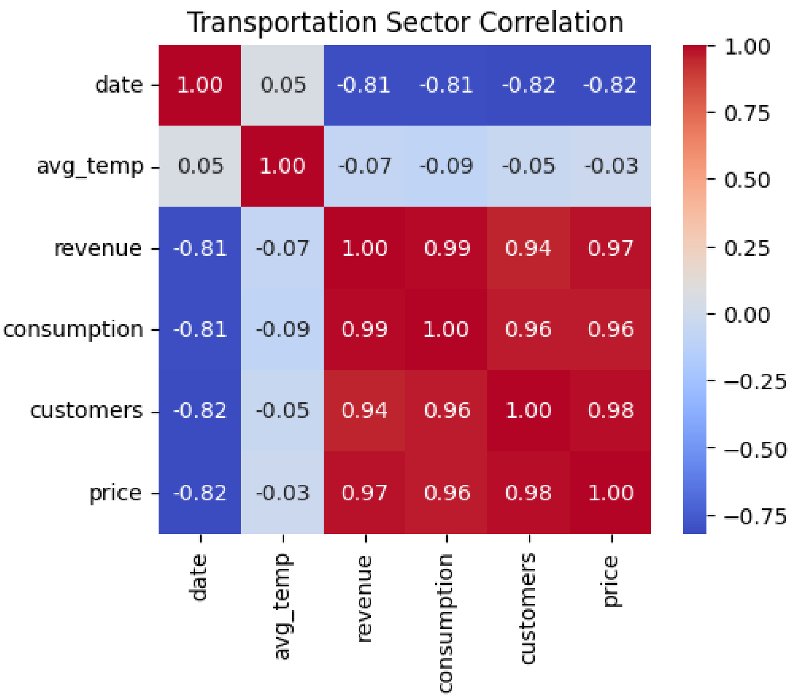


Figure 7: Correlation Matrix (Transportation Sector)

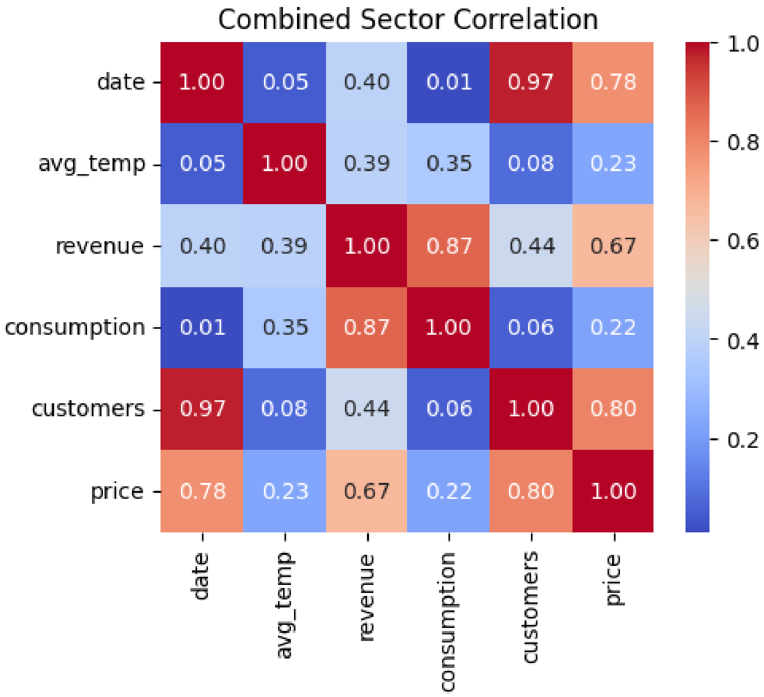


Figure 8: Correlation Matrix (All Sectors)

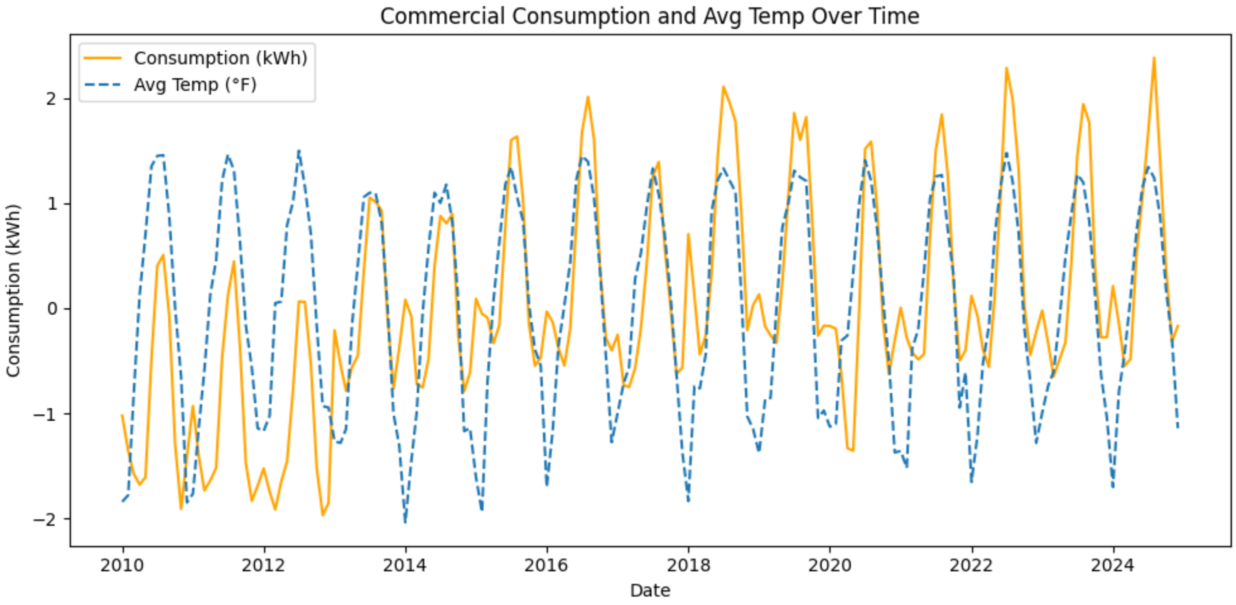


Figure 9: Time Series (Commercial Sector)

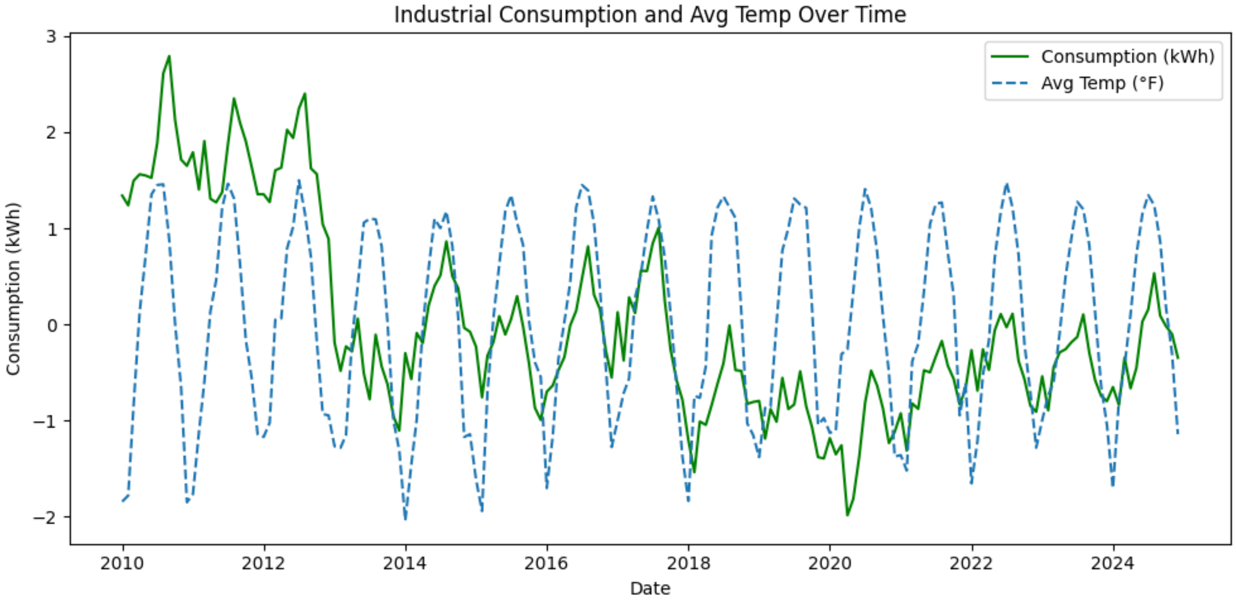


Figure 10: Time Series (Industrial Sector)

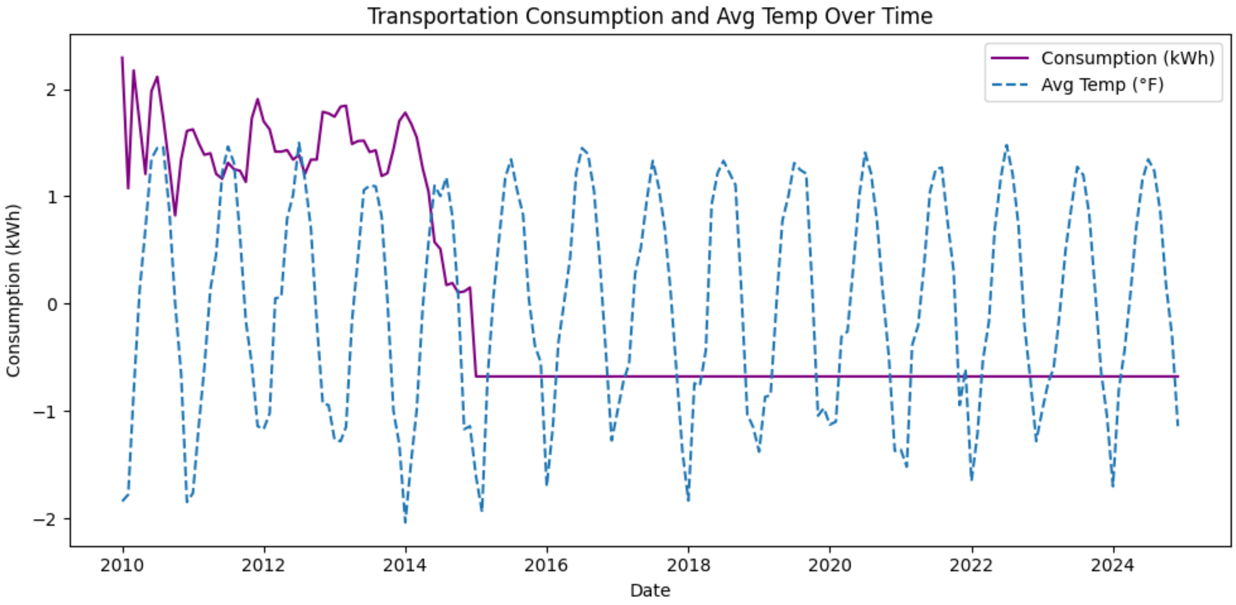


Figure 11: Time Series (Transportation Sector)

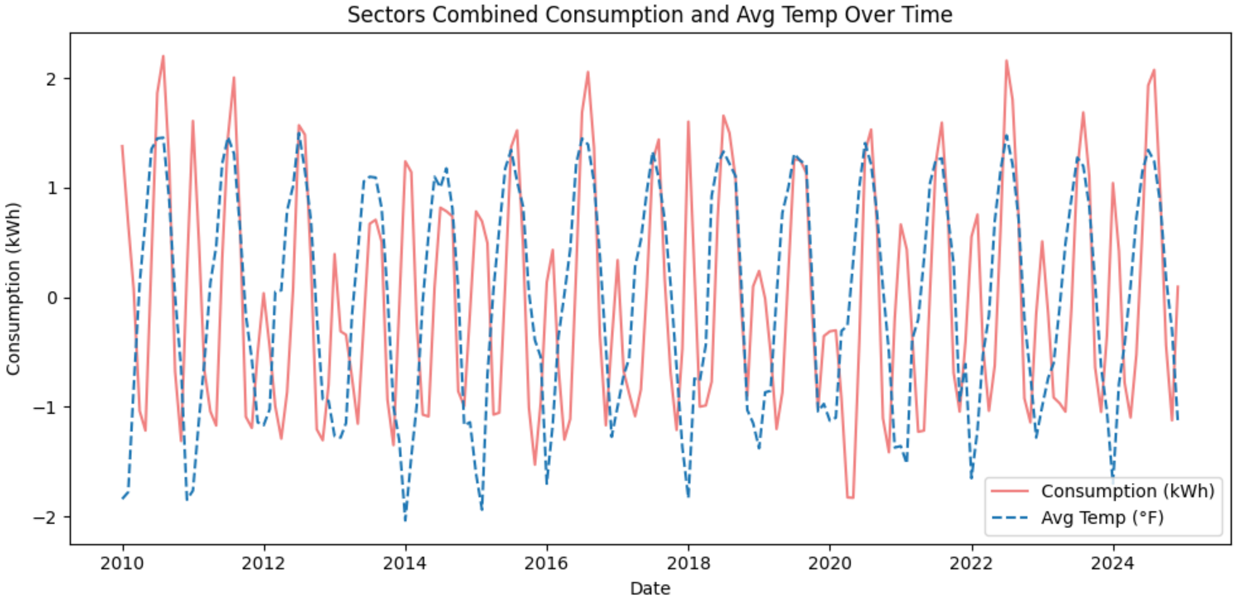


Figure 12: Time Series (All Sectors)

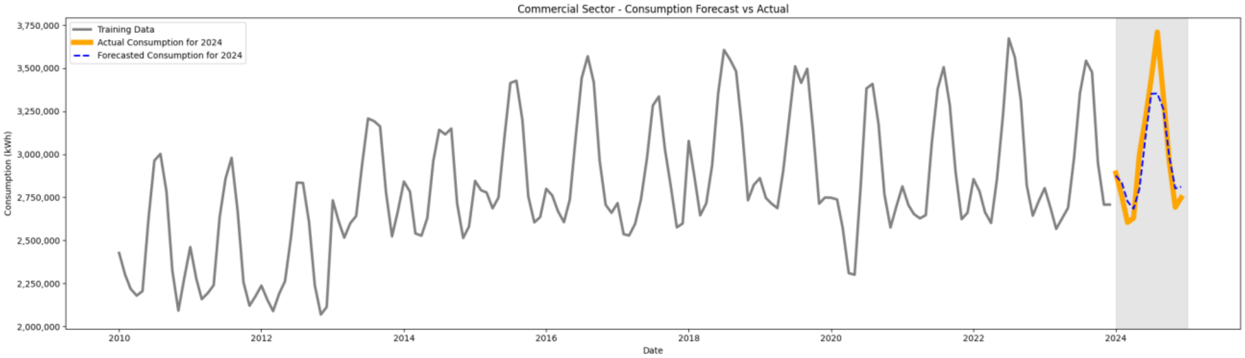


Figure 13: Forecast vs. Actual (Commercial Sector)

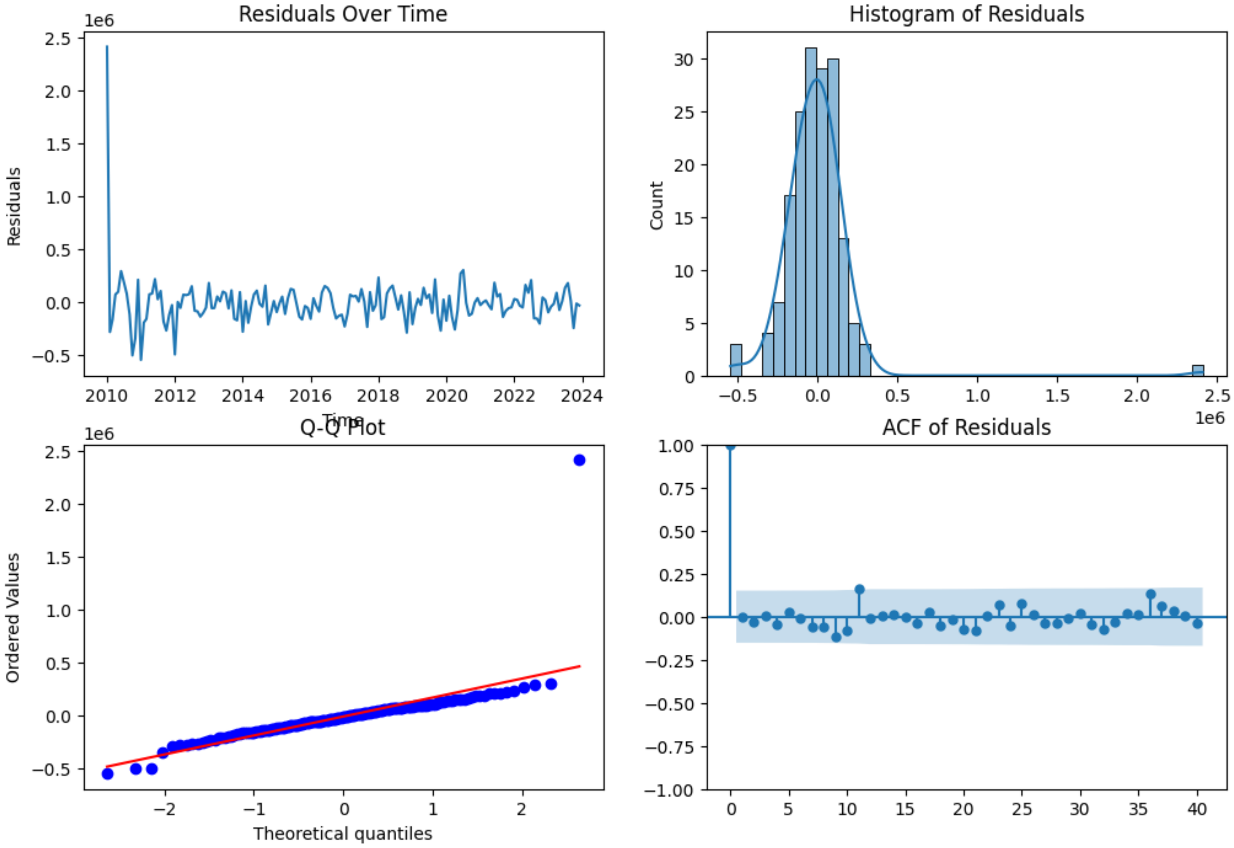


Figure 14: Residual Diagnostic Plots (Commercial Sector)

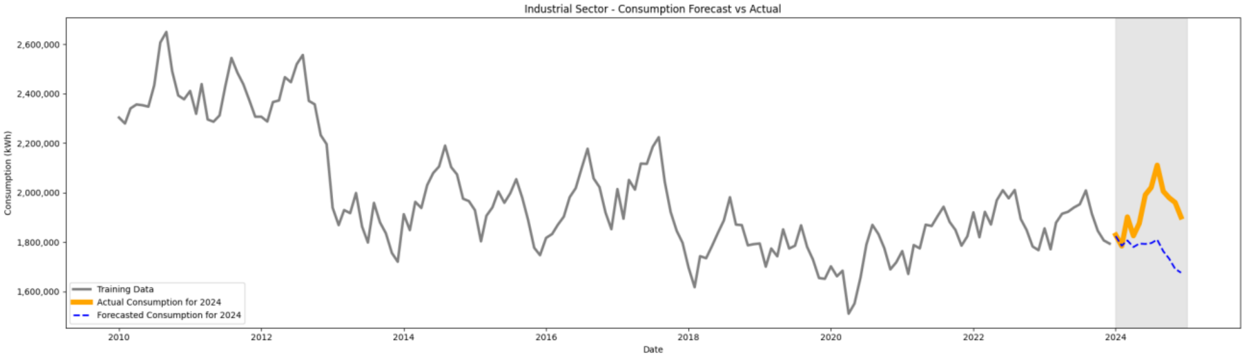


Figure 15: Forecast vs Actual (Industrial Sector)

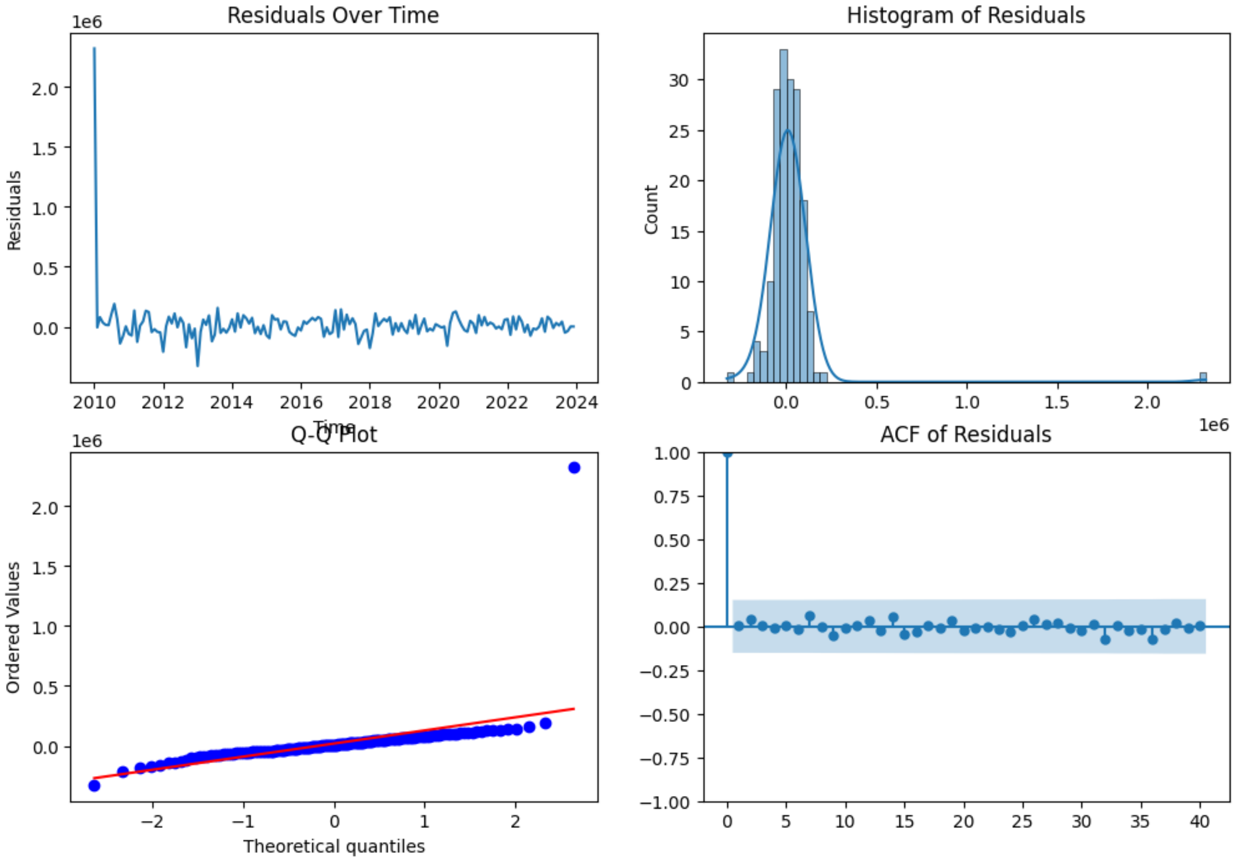


Figure 16: Residual Diagnostic Plots (Industrial Sector)

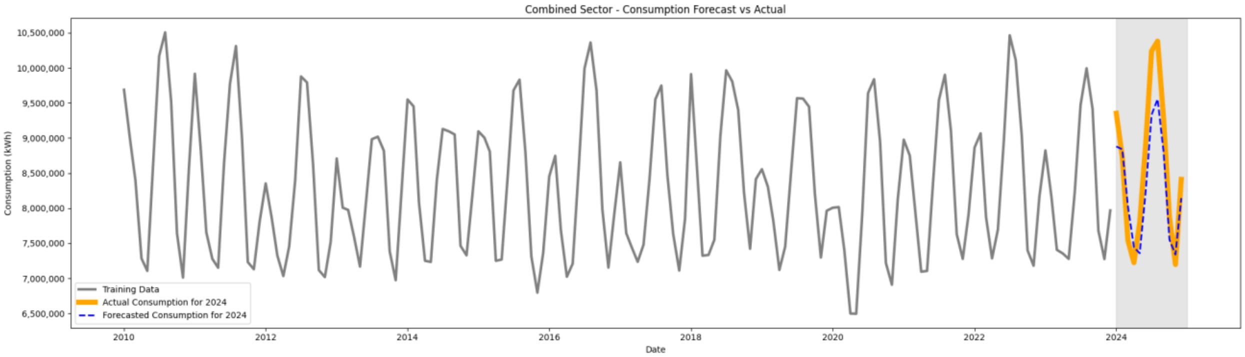


Figure 17: Forecast vs Actual (All Sectors)

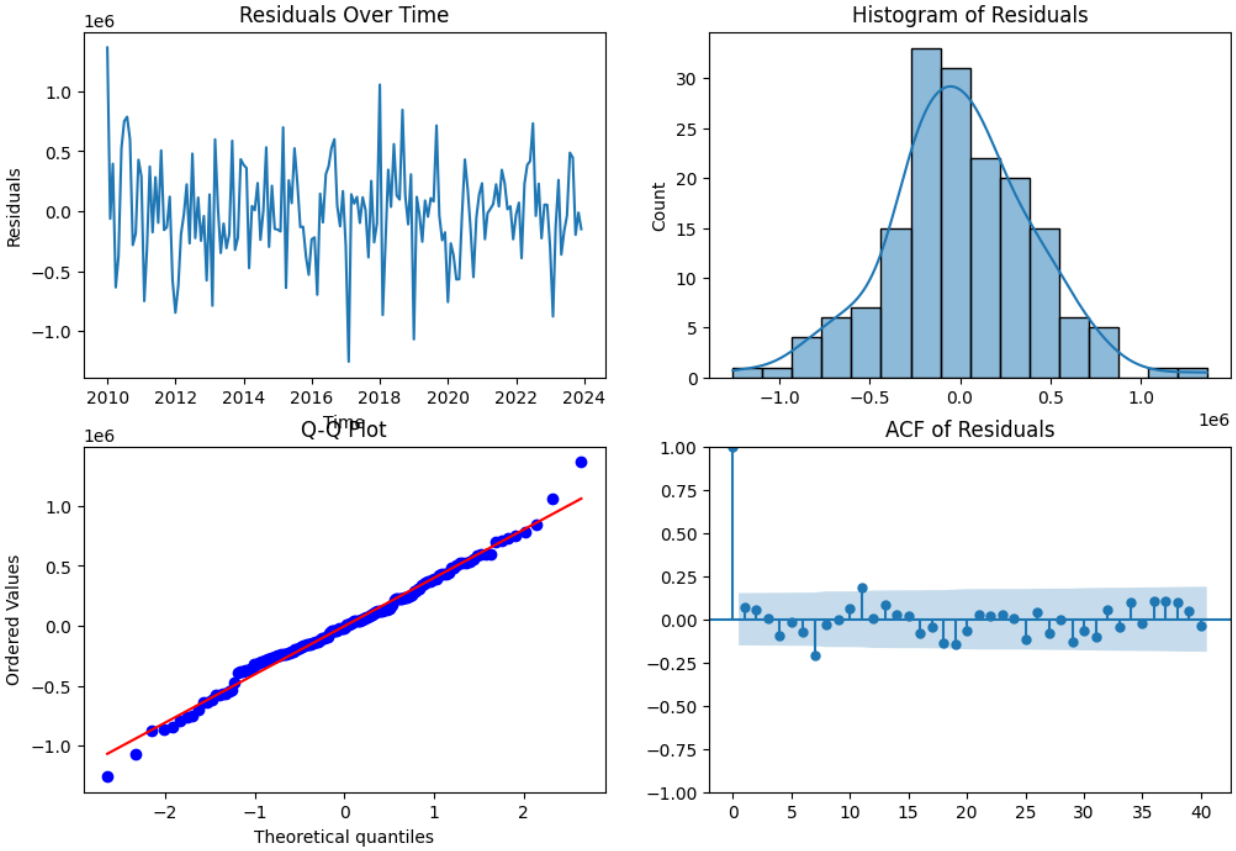


Figure 18: Residual Diagnostic Plots (All Sectors)

Appendix B: Tables

Model Evaluation Metrics			
MAE	MSE	RMSE	MAPE %
108490.616348	1.952588e+10	139735.033846	0.035004

Table 1: Model Evaluation Metrics (Commercial Sector)

Model Evaluation Metrics			
MAE	MSE	RMSE	MAPE %
161291.94	3.672569e+10	191639.473237	0.081365

Table 2: Model Evaluation Metrics (Industrial Sector)

Model Evaluation Metrics			
MAE	MSE	RMSE	MAPE %
444955.58	2.578050e+11	507744.97	0.050119

Table3: Model Evaluation Metrics (All Sectors)

Ljung-Box Test	
lb_stat	lb_pvalue
5.606136	0.847198

Table 4: Ljung-Box Test (Commercial Sector)

Ljung-Box Test	
lb_stat	lb_pvalue
1.506119	0.998916

Table 5: Ljung-Box Test (Industrial Sectors)

Ljung-Box Test	
lb_stat	lb_pvalue
12.649998	0.243904

Table 6: Ljung-Box Test (All Sectors)

Appendix C: Code

```
import numpy as np
import pandas as pd
import seaborn as sns
import scipy.stats as stats
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.stats.diagnostic import acorr_ljungbox
from pmdarima import auto_arima
from sklearn.metrics import mean_squared_error, mean_absolute_error, mean_absolute_percentage_error,
root_mean_squared_error
from sklearn.preprocessing import StandardScaler

import warnings
from pandas.errors import SettingWithCopyWarning
warnings.simplefilter(action='ignore', category=SettingWithCopyWarning)
```

Code 1: Import Libraries

```
url_1 = (r"C:\Users\Hamza\OneDrive\Documents\Data Science\Masters\6. Spring 2025\DSC630-T302 "
        r"Predictive Analytics (2255-1)\Weekly Assignments & Material\Week 9\HS861M 2010-.xlsx")

url_2 = (r"C:\Users\Hamza\OneDrive\Documents\Data Science\Masters\6. Spring 2025\DSC630-T302 "
        r"Predictive Analytics (2255-1)\Weekly Assignments & Material\Week 9\TN_AVG_TEMP_2010_2024.csv")

df_temp = pd.read_csv(url_2, skiprows=3)
df = pd.read_excel(url_1)
```

Code 2: Load Data

```
df.rename(columns={
    "Unnamed: 0":"year", "Unnamed: 1":"month", "Unnamed: 2":"state", "Unnamed: 3":"status",
    "RESIDENTIAL": 'res_rev', 'Unnamed: 5': 'res_sales', 'Unnamed: 6': 'res_cust', 'Unnamed: 7' : 'res_price',
    "COMMERCIAL": 'com_rev', 'Unnamed: 9': 'com_sales', 'Unnamed: 10': 'com_cust', 'Unnamed: 11' : 'com_price',
    "INDUSTRIAL": 'ind_rev', 'Unnamed: 13': 'ind_sales', 'Unnamed: 14': 'ind_cust', 'Unnamed: 15' : 'ind_price',
    "TRANSPORTATION": 'trans_rev', 'Unnamed: 17': 'trans_sales', 'Unnamed: 18': 'trans_cust', 'Unnamed: 19' : 'trans_price',
    "TOTAL": 'total_rev', 'Unnamed: 21': 'total_sales', 'Unnamed: 22': 'total_cust', 'Unnamed: 23' : 'total_price'
}, inplace=True)

df = df.drop(index=[0, 1], columns=['status'])
df.reset_index(drop=True, inplace=True)
df = df.convert_dtypes()
df = df.drop(df.index[-1])
df['date'] = pd.to_datetime(df[['year', 'month']].assign(day=1))
```

Code 3: Initial Data Cleaning

```
df_temp.rename(columns={"Date" : "date", 'Value': 'avg_temp'}, inplace=True)
df_temp['date']= df_temp['date'].astype(str)
df_temp['year'] = df_temp['date'].str[0:4]
df_temp['month'] = df_temp['date'].str[4:6]
df_temp['date'] = pd.to_datetime(df_temp[['year', 'month']].assign(day=1))
df_temp.drop(columns=['year', 'month'], inplace=True)
```

Code 4: Temperature Data Cleaning

```
df = df[df['state'] == 'TN']
df_merged = pd.merge(df_temp, df, on='date', how='inner')
```

Code 5: Filter and Merge

```
df_res = df_merged[['date', 'avg_temp', 'res_rev', 'res_sales', 'res_cust', 'res_price']]
df_com = df_merged[['date', 'avg_temp','com_rev', 'com_sales', 'com_cust', 'com_price']]
df_ind = df_merged[['date', 'avg_temp','ind_rev', 'ind_sales', 'ind_cust', 'ind_price']]
df_trans = df_merged[['date','avg_temp', 'trans_rev', 'trans_sales', 'trans_cust', 'trans_price']]
df_total = df_merged[['date', 'avg_temp','total_rev', 'total_sales', 'total_cust', 'total_price']]

sectors = [df_res, df_com, df_ind, df_trans, df_total]
for df in sectors:
    col_names = df.columns
    df.rename(columns={
        col_names[2]: 'revenue',
        col_names[3]: 'consumption',
        col_names[4]: 'customers',
        col_names[5]: 'price'
    }, inplace=True)
```

Code 6: Create Sector Subsets

```
sector_names = ["Residential", "Commercial", "Industrial", "Transportation", "Combined"]
for i, dfcor in enumerate(sectors):
    sector_naming = sector_names[i]
    print(f"{sector_naming} Sector")
    print(dfcor.describe())
    print("***"*30)
```

Code 7: Summary Statistics

```
for i, dfcor in enumerate(sectors):
    sector_naming = sector_names[i]
    print(f"{sector_naming} Sector")

    # Plot only the 'Consumption' column
    dfcor['consumption'].plot(kind='box')

    # Add a title for clarity
    plt.title(f"{sector_naming} - Consumption Boxplot")

    # Show each plot separately
    plt.show()

    q1 = dfcor['consumption'].quantile(0.25)

    q3 = dfcor['consumption'].quantile(0.75)

    iqr = q3 - q1
    outliers = dfcor[(dfcor['consumption'] < q1 - 1.5 * iqr) | (dfcor['consumption'] > q3 + 1.5 * iqr)]

    print(f"{sector_naming} Sector has {len(outliers)} outliers")

    print("***"*30)
```

Code 8: Box Plots

```
sector_names = ["Residential", "Commercial", "Industrial", "Transportation", "Combined"]

for i, dfcor in enumerate(sectors):
    sector_naming = sector_names[i]
    corr = dfcor.corr()

    plt.figure(figsize=(5, 4))
    sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
    plt.title(f"{sector_naming} Sector Correlation")
    plt.show()
```

Code 9: Correlation Heatmap

```
scaler = StandardScaler()

for df in sectors:
    df[['consumption_scaled', 'avg_temp_scaled']] = scaler.fit_transform(df[['consumption', 'avg_temp']])

colors = ["red", "orange", "green", "purple", "lightcoral"]

sector_names = ["Residential", "Commercial", "Industrial", "Transportation", "Sectors Combined"]

for df, sector_name, color in zip(sectors, sector_names, colors):
    plt.figure(figsize=(10, 4))
    plt.plot(df['date'], df['consumption_scaled'], label='Consumption (scaled)', color=color)
    plt.plot(df['date'], df['avg_temp_scaled'], label='Avg Temp (scaled)', linestyle='--', color='blue')
    plt.title(f"{sector_name} - Scaled Consumption vs. Avg Temp Over Time")
    plt.xlabel("Date")
    plt.ylabel("Scaled Values")
    plt.legend()
    plt.tight_layout()
    plt.show()
```

Code 10: Standardize and Plot Time Series

```
sectors = [df_res, df_com, df_ind, df_total]
for df in sectors:
    df.set_index("date", inplace=True)
```

Code 11: Prepare Sector DataFrames

```
x = df_total.drop(['revenue','customers', 'consumption', 'date', 'consumption_scaled', 'avg_temp_scaled'], axis = 1)
y = df_total['consumption']
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.1, random_state=101)
```

Code 12: Create Features & Target Variables & Training Split

```
mls = LinearRegression()
mls.fit(X_train,y_train) # fitting the model
predictions = mls.predict(X_test) # making predictions
```

Code 13: Regression Model Training & Prediction

```
# Calculate correlation of all features with 'consumption' target
# Exclude 'consumption', 'consumption_scaled', 'date', 'revenue', and 'avg_temp_scaled' from consideration
correlations = df_total.corr()['consumption'].drop([
    'consumption', 'consumption_scaled', 'date', 'revenue', 'avg_temp_scaled'])
# Sort features by absolute correlation in descending order to prioritize strongest relationships
sorted_features = correlations.abs().sort_values(ascending=False).index.tolist()
```

Code 14: Create Correlation Dataframe

```
results = []  # List to store performance metrics for different numbers of features
# Loop through feature subsets from 1 up to all sorted features
for i in range(1, len(sorted_features) + 1):
    selected_features = sorted_features[:i]  # Select top 'i' features based on correlation
    x = df_total[selected_features]          # Feature subset dataframe
    y = df_total['consumption']              # Target variable

    # Split data into train and test sets (90% train, 10% test), fixed random seed for reproducibility
    X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.1, random_state=101)
    model = LinearRegression()     # Initialize linear regression model
    model.fit(X_train, y_train)    # Train model on training data
    preds = model.predict(X_test)  # Predict on test data

    # Calculate R-squared and RMSE on test data to evaluate model performance
    r2 = r2_score(y_test, preds)
    rmse = np.sqrt(mean_squared_error(y_test, preds))
    # Store the number of features, selected features, and performance metrics
    results.append((i, selected_features, r2, rmse))

# Print the results for each subset of features tried
for count, features, r2, rmse in results:
    print(f"{count} features | R2: {r2:.4f} | RMSE: {rmse:,.0f} | Features: {features}")
```

Code 15: MLR on Selected Features

```
sectors_training_df_sets = []

sectors_testing_df_sets = []

for i, df in enumerate(sectors):
    train_df = df[df.index < '2024-01-01']

    test_df = df[df.index >= '2024-01-01']

    sectors_training_df_sets.append(train_df)

    sectors_testing_df_sets.append(test_df)
```

Code 16: Train-Test Split for ARIMA

```
# define sectors names so it can be added to the print out.
sector_names = ["Residential", "Commercial", "Industrial", "Combined"]


# define a for loop that goes over the training set and then fits and forecasts & evaluates
for i, dataf in enumerate(sectors_training_df_sets):

    # add sector name to identify instead of using index
    sector_name = sector_names[i]
```

```
print(f"Training ARIMA model for {sector_name} sector:")

# each test_df from the set to use for forecasting
test_df = sectors_testing_df_sets[i]



# model each sector
model = auto_arima(
    dataf['consumption'],
    seasonal=True,
    m=12,
    stepwise=True,
    suppress_warnings=True,
    error_action='ignore')


# forecast 2024 for each sector
forecast = model.predict(n_periods=12)

forecast_index = test_df.index

forecast_series = pd.Series(forecast, index=forecast_index)

print("Forecast vs. Actual Plot")


# Forecast vs. Actual Plot
plt.figure(figsize=(20, 6))

plt.plot(dataf.index, dataf['consumption'],
         label='Training Data', linewidth=3, color='grey')
plt.plot(test_df.index, test_df['consumption'],
         label='Actual Consumption for 2024', color='orange', linewidth=6)
plt.plot(forecast_series.index, forecast_series,
         label='Forecasted Consumption for 2024', color='blue', linestyle='--', linewidth=2)

# add gray color to background between 2024 and 2025
plt.axvspan(pd.Timestamp("2024-01-01"), pd.Timestamp("2025-01-01"),
            color='grey', alpha=0.2)

plt.title(f"{sector_name} Sector - Consumption Forecast vs Actual")
plt.xlabel("Date")
plt.ylabel("Consumption (kWh)")
plt.legend()
plt.tight_layout()

# change the yticks to show the full number in millions.
plt.gca().yaxis.set_major_formatter(ticker.FuncFormatter
                                    (lambda x, _: f'{x:,.0f}'))  # Adds commas to large numbers
plt.show()

print("--" * 20)

print("Model Evaluation Metrics")

mae = mean_absolute_error(test_df['consumption'], forecast)
mse = mean_squared_error(test_df['consumption'], forecast)
mape = mean_absolute_percentage_error(test_df['consumption'], forecast)
rmse = root_mean_squared_error(test_df['consumption'], forecast)

report = pd.DataFrame({
'MAE': [mae],
'MSE' : [mse],
'RMSE' : [rmse],
'MAPE %': [mape]
})

print(report.head())

print("--" * 20)


# Ljung-Box test
# Get in-sample residuals
residuals = pd.Series(model.resid(), index=dataf.index)

print("Ljung-Box Test")

ljung_box_result = acorr_ljungbox(residuals, lags=[10], return_df=True)

print(f"{ljung_box_result}")


print("--" * 20)


print("Residual Diagnostic Plots")


# Plot residuals
plt.figure(figsize=(12, 8))

plt.subplot(2, 2, 1)
plt.plot(residuals)
plt.title("Residuals Over Time")
plt.xlabel("Time")
plt.ylabel("Residuals")

plt.subplot(2, 2, 2)
sns.histplot(residuals, kde=True)
plt.title("Histogram of Residuals")

plt.subplot(2, 2, 3)
stats.probplot(residuals, dist="norm", plot=plt)
plt.title("Q-Q Plot")

plt.subplot(2, 2, 4)
plot_acf(residuals, lags=40, ax=plt.gca())
plt.title("ACF of Residuals")

plt.show()

# add a dash seperator
print("--" * 50)
print("--" * 50)
```

Code 17: Modeling, Forecasting, and Evaluation