Analyzing Energy Consumption Trends in Colorado: A Geospatial Approach

Nicholas S. Elias*

Northwest Missouri State University, Maryville MO 64468, USA
S576714@nwmissouri.edu,nicholas.elias95@gmail.com

Abstract. This paper presents a comprehensive analysis of energy consumption trends across Colorado, focusing on the interplay of geographic, climatic, and demographic factors to support smart grid optimization. Using a curated dataset from the National Renewable Energy Laboratory, we analyzed electricity and natural gas usage across residential, commercial, and industrial sectors. Exploratory data analysis revealed strong correlations between Department of Energy (DOE) climate zones and consumption patterns. We developed a predictive pipeline incorporating linear regression and decision tree models, though these yielded limited accuracy. In contrast, unsupervised methods such as K-Means clustering revealed distinct regional consumption clusters. The findings offer a foundation for localized policy interventions and energy efficiency strategies, highlighting the value of geospatial data analytics in shaping sustainable energy planning.

Keywords: Geospatial Analysis · Smart Grid Optimization · Machine Learning · Linear Regression · Decision Tree · Clustering Analysis · Climate Zones

1 Introduction

The global demand for energy continues to rise, presenting a pressing challenge as societies simultaneously strive to reduce greenhouse gas emissions and transition toward more sustainable systems. In this context, understanding the spatial and temporal patterns of energy consumption has become crucial for enabling smart grid technologies and promoting efficient resource distribution. Colorado, with its diverse geography, climate zones, and population densities, offers a unique setting to examine these patterns. This study investigates how location-specific factors—including climate classification, elevation, and population—impact energy usage across residential, commercial, and industrial sectors. By leveraging geospatial analysis and machine learning techniques, we aim to uncover actionable insights that can support data-driven policy making and foster more resilient energy infrastructure across the state.

* GitHub: https://github.com/NickElias01/Capstone$_N WMSU_D A$

2 Domain Selection

This project focuses on the **energy and sustainability** domain, specifically **smart grid optimization and renewable energy integration** in Colorado. This area is crucial because energy demand is increasing, and optimizing power distribution can reduce costs, improve grid reliability, and support the transition to renewable energy.

3 Data Problem and Its Importance

3.1 Problem Statement

This project analyzes energy consumption patterns across Colorado's cities and counties to understand the relationships between geographic, demographic, and climate factors on energy usage. The primary focus is on examining how different sectors (residential, commercial, and industrial) consume electricity and natural gas, and their resulting environmental impact through greenhouse gas (GHG) emissions.

3.2 Importance

Understanding energy consumption patterns is crucial for several reasons:

- **Resource Planning:** Analyzing consumption patterns helps utilities and municipalities plan for future energy needs and infrastructure development.
- **Environmental Impact:** By examining GHG emissions across different sectors, we can identify areas where environmental impact can be reduced most effectively.
- **Climate Considerations:** The relationship between DOE climate zones and energy usage provides insights into how climate affects consumption patterns.
- **Economic Implications:** Understanding sector-specific energy use (residential, commercial, industrial) helps in economic planning and development.

4 Project Implementation Phases

The project will be implemented in the following phases:

1. **Data Collection & Preparation**
   - Gather historical energy consumption data
2. **Data Storage & Processing**
   - Store the data using **SQLite** for structured analysis.
   - Process and clean the data using Python libraries (`pandas`, `NumPy`, SQL queries).

3. **Data Analysis & Insights**
   - Identify trends in energy consumption and peak demand hours.
   - Analyze correlations between weather conditions and energy usage.
4. **Visualization & Reporting**
   - Create graphs using **Matplotlib**
   - Present insights on energy consumption trends and their influencing factors.
5. **Conclusion & Future Work**
   - Summarize key findings and assess the significance of the observed trends.
   - Suggest improvements for grid efficiency and sustainability.
   - Explore potential extensions, such as integrating real-time data sources or expanding the analysis to multiple regions.

5 Data Source

The data was extracted from The National Renewable Energy Laboratory's Equitable Energy Investment Prioritization Data Set at:
`https://data.nrel.gov/submissions/175`.

Dataset Format: The dataset was downloaded in Excel Binary format: .XLSB and converted to CSV for easier manipulation.
Data Curation Process: The original dataset contained multiple rows for headers, which were consolidated into a single structured header. Irrelevant columns were removed, and missing values were analyzed. Null values and '-' symbols were retained for further evaluation in exploratory data analysis (EDA).
Major Data Attributes:

- **Electricity Consumption (MWh)**
  - Residential, Commercial, Industrial sectors
  - Measures total electrical energy usage per sector
- **Natural Gas Consumption (Mcf)**
  - Residential and Commercial usage
  - Used for heating, cooking, and power generation
- **Geographic Data**
  - Latitude, Longitude, County
  - DOE Climate Zone classification (1-8)
- **Demographic Data**
  - Population size to normalize energy consumption trends

Dataset Format: The dataset was downloaded in Excel Binary format: .XLSB
Tools Used:

- Python
- Pandas
- PyXLSB library (engine for reading .xlsb files)
- Visual Studio Code

Attributes and Records

- City Dataset: 273 Records, 20 Attributes
- County Dataset: 64 Records, 26 Attributes

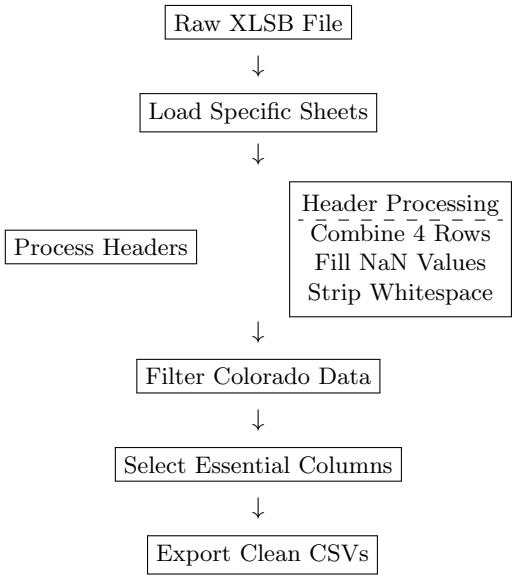

Fig. 1. Data Curation Process Flow

6 Exploratory Data Analysis

6.1 Overview and Importance

Exploratory Data Analysis (EDA) forms the foundation of our energy consumption study, serving as a systematic approach to understand patterns and relationships within Colorado's energy usage data. Through this initial investigation, we identified key relationships between climate zones and energy consumption, enabling informed decisions for subsequent detailed analyses.

6.2 Methodology and Techniques

Our EDA approach utilized three primary techniques:

- **Climate Zone Analysis:** Examination of energy consumption patterns across Colorado's climate zones (4-7)
- **Distribution Analysis:** Investigation of consumption patterns across different municipalities
- **Correlation Analysis:** Study of relationships between residential, commercial, and industrial energy usage

6.3 Analysis Results

The climate zone analysis (Figure 2) revealed distinct consumption patterns across Colorado's varied climate regions. Key findings include:

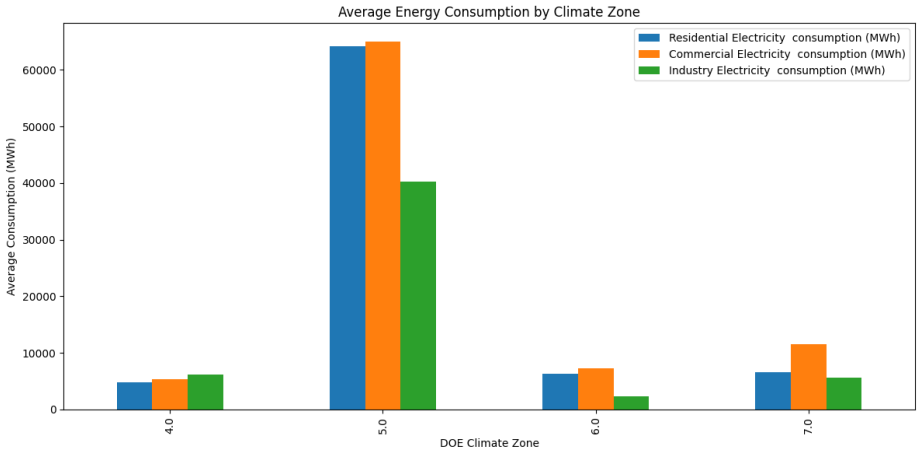

Fig. 2. Climate Zone Analysis

- Higher average consumption in specific climate zones
- Significant variation in residential electricity usage across zones
- Notable differences between commercial and industrial consumption patterns

6.4 Key Insights

This exploratory phase revealed several crucial insights:

1. Climate zones significantly influence energy consumption patterns
2. Strong correlations exist between residential and commercial energy use
3. Several municipalities exhibit unique consumption patterns warranting further investigation

These findings guide our subsequent detailed analysis of energy efficiency and consumption patterns across Colorado's diverse climate regions.

7 Mechanism for the Predictive Application

The pipeline built to accomplish the predictive application includes the following steps:

1. **Data Preprocessing:** Cleaning and transforming raw data into a usable format.
2. **Feature Engineering:** Selecting relevant attributes such as population, climate zone, and energy consumption metrics.
3. **Model Training:** Using machine learning algorithms to train predictive models.

4. **Model Evaluation:** Testing the model's performance using evaluation metrics.
5. **Prediction:** Generating predictions for energy consumption.

A diagram illustrating the pipeline is shown in Figure 3.

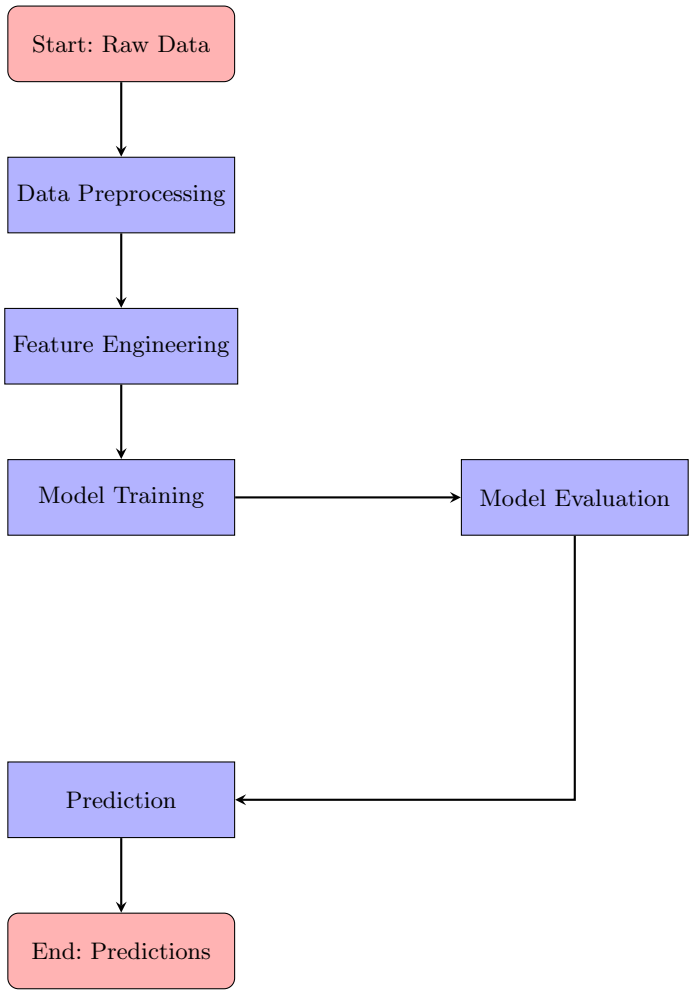

Fig. 3. Predictive Application Pipeline

8 Machine Learning Algorithms Used

The following machine learning algorithms were employed to analyze the data:

- **Linear Regression:** Predicts energy consumption based on population and climate zone.
- **Random Forest Regressor:** Captures non-linear relationships between features and energy consumption.
- **K-Means Clustering:** Groups cities/counties with similar energy consumption patterns.

9 Training and Testing Process

The training and testing process consisted of the following steps:

1. **Data Preparation:** Data was cleaned, features (e.g., population, `doe_climate_zone`) and target (Industry Electricity consumption) were selected, and missing values were removed.
2. **Splitting:** Data was split into 80% training and 20% testing sets.
3. **Scaling:** Features were standardized using `StandardScaler`.
4. **Models Trained:**
   - **Linear Regression:** Simple regression model.
   - **Random Forest Regressor:** Ensemble model for improved accuracy.
   - **K-Means Clustering:** Unsupervised clustering algorithm.

10 Implementation and Evaluation Process

The implementation and evaluation process is described below:

- **Evaluation:**
  - **Linear Regression:** High MAE/MSE and a negative R^2 (i.e., -1.17) indicate poor performance.
  - **Random Forest:** Performed better than Linear Regression but still underperformed ($R^2 = -0.09$).
  - **K-Means Clustering:** Achieved a Silhouette Score of 0.6187, indicating well-formed clusters.

11 Results of the Analysis

The results from the analysis are summarized in the table below.

Linear Regression Evaluation

- MAE: 17,274.51
- MSE: 3,238,976,130.99
- R^2: -1.17

Conclusion: The model performed poorly, with high errors and a negative R^2, indicating it failed to capture the relationship between the features and the target variable.

Random Forest Regressor Evaluation

- MAE: 12,797.84
- MSE: 1,627,876,745.22
- R^2: -0.09

Conclusion: Although Random Forest performed better than Linear Regression, it still struggled to explain the variance in the data.

K-Means Clustering Evaluation

- Silhouette Score: 0.6187

Conclusion: The clustering model performed well, forming distinct and well-separated clusters.

12 Findings and Conclusions

12.1 Results

The visualizations provided several key insights into the performance of the models and patterns within the dataset. Scatter plots for both Linear Regression and Random Forest models indicated poor predictive performance, with data points widely dispersed from the ideal diagonal line representing perfect predictions. However, Random Forest showed slightly improved performance, with predictions clustering closer to actual values compared to Linear Regression.

Bar plots effectively highlighted differences in energy consumption across sectors and climate zones, revealing where energy usage was most concentrated. Additionally, the K-Means clustering scatter plot displayed well-defined clusters with clearly separated centroids, corroborated by a Silhouette Score of 0.6187, indicating a good level of cluster separation and cohesion.

12.2 Interpretation and Inferences

The regression analysis suggests that the selected features—such as population and climate zone—were insufficient in explaining variations in industrial electricity consumption. This is reflected in the poor performance metrics observed in both models.

- **Linear Regression:** Exhibited high Mean Absolute Error (MAE) and Mean Squared Error (MSE), alongside a negative R^2 score of -1.17, indicating that the model performed worse than a simple mean predictor.
- **Random Forest:** Showed modest improvement with a less negative R^2 score of -0.09, though still indicative of inadequate model fit.
- **K-Means Clustering:** Demonstrated meaningful grouping of data points with a Silhouette Score of 0.6187, supporting the presence of underlying patterns in the data.

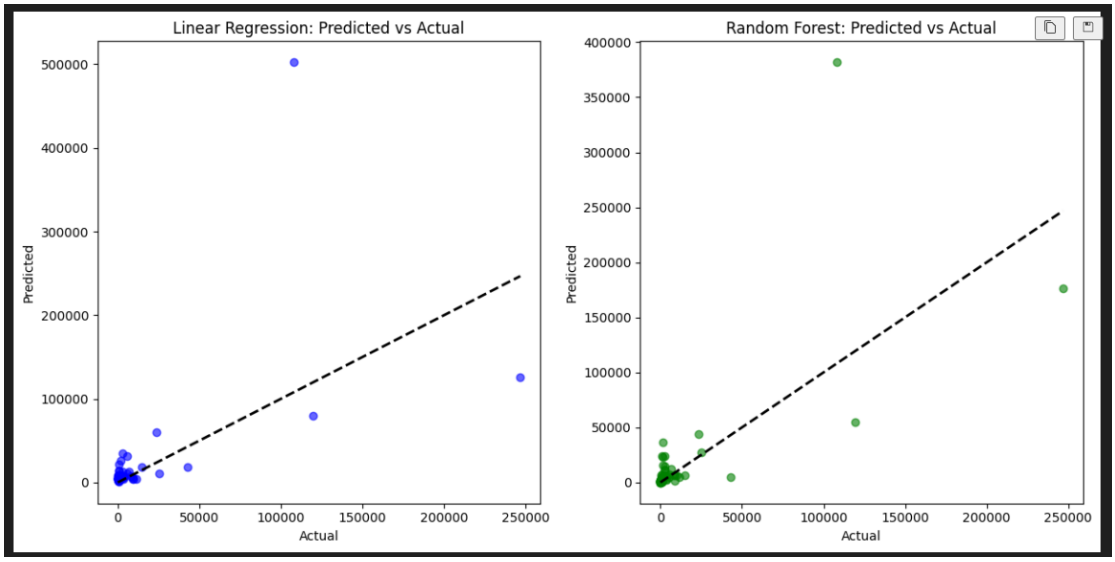

Fig. 4. Linear Regression and Random Forest Charts

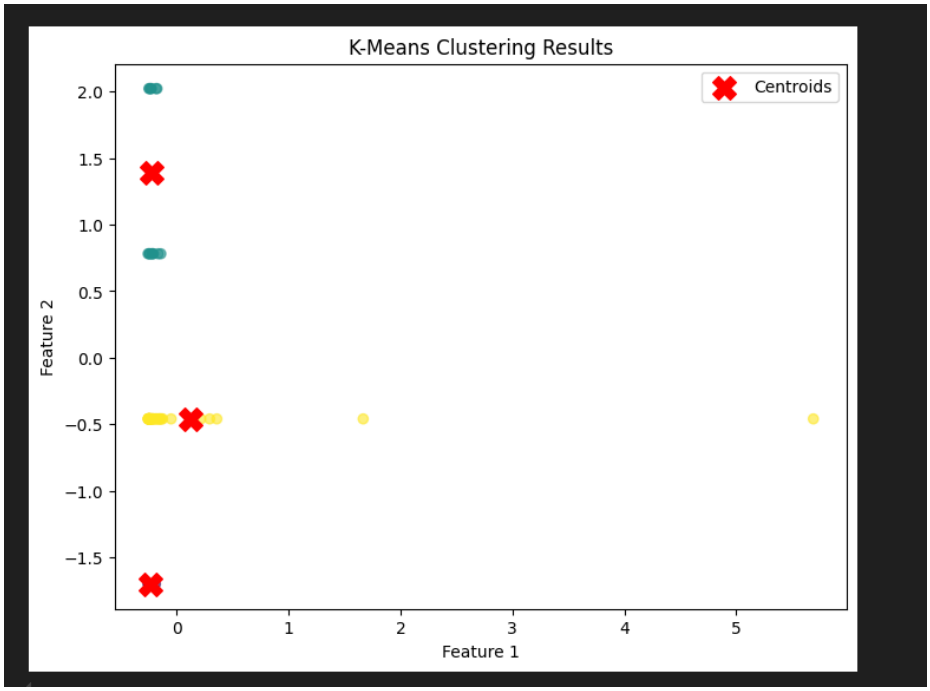

Fig. 5. K-Means Clustering Results

These results imply that while clustering revealed valuable structure within the data, the regression models were not successful in capturing meaningful relationships. This may point to the presence of non-linear interactions or missing explanatory variables in the dataset.

12.3 Conclusions and Recommendations

Overall, clustering analysis proved to be more insightful than regression modeling in this context. To improve predictive performance in future work, the following strategies are recommended:

- Incorporate additional and more relevant features (e.g., industrial output, energy policy data).
- Explore advanced machine learning models such as Gradient Boosting or XGBoost.
- Perform hyperparameter tuning for models like Random Forest to enhance accuracy.
- Conduct feature engineering to capture potential non-linear or interaction effects.

Additionally, addressing data limitations—such as missing variables or unaccounted variability across regions—could significantly enhance model performance. Further exploration into the identified clusters may also yield actionable insights into energy consumption behaviors across different industrial profiles.

13 Project Limitations and Future Improvements

13.1 Project Limitations

While this project aims to enhance energy distribution efficiency in Colorado's smart grid through data analytics and predictive modeling, several limitations were encountered:

1. **Data Availability and Quality**: Access to comprehensive and high-resolution datasets was limited. Some sources had gaps or inconsistencies, which could affect the accuracy of analyses.
2. **Model Complexity**: Developing predictive models that accurately capture the dynamic nature of energy consumption and the intermittency of renewable sources proved challenging. Simplified models may not fully represent real-world complexities.
3. **Integration with Existing Infrastructure**: Aligning the project's analytical models with Colorado's current energy infrastructure requires overcoming technical and regulatory hurdles, which were beyond the project's scope.
4. **Scalability**: The project focused on specific regions and may not be directly scalable to other areas without adjustments for local conditions and data availability.

13.2 Future Improvements

To address these limitations and enhance the project's impact, the following improvements are recommended:

1. **Enhanced Data Collection**: Collaborate with local utilities and government agencies to obtain more granular and accurate datasets. Implement data cleaning and preprocessing techniques to improve data quality.
2. **Advanced Modeling Techniques**: Explore machine learning algorithms capable of capturing complex patterns in energy consumption and renewable generation, such as deep learning models. Incorporate real-time data feeds to update models dynamically.
3. **Infrastructure Integration**: Engage with stakeholders to understand the technical and regulatory frameworks governing Colorado's energy infrastructure. Design models with interoperability in mind to facilitate smoother integration.
4. **Scalability Assessment**: Conduct pilot studies in different regions to assess the adaptability of models. Develop modular and flexible frameworks that can be customized for various locales.

References

1. U.S. Energy Information Administration (EIA). *Electricity Data Browser*. Available at: `https://www.eia.gov/electricity/`
2. National Renewable Energy Laboratory (NREL). *Renewable Energy Data*. Available at: `https://www.nrel.gov/research/re-data.html`
3. National Oceanic and Atmospheric Administration (NOAA). *Climate Data Online*. Available at: `https://www.ncdc.noaa.gov/cdo-web/`
4. Zhou, K., Liu, T., & Zhou, L. (2022). A Survey on IoT-Enabled Smart Grids: Emerging, Applications, and Challenges. *Energies*, 15(19), 6984. Available at: `https://www.mdpi.com/1996-1073/15/19/6984`