Electricity Prices Analysis and CO2 Emission Forecasting in the US Using EIA

Kedar Prashant Vichare Tiffany Li Leela Prasad Dammalapati Utkarsh Tripathi

Abstract

This project integrates energy consumption and environmental impact data from the U.S. Energy Information Administration (EIA) covering all the datasets on retail electric sales and CO2 emissions in the United States. The Retail Electric Sales Data provides smaller and deeper insights into electricity consumption, customer demographics, pricing, and revenue trends across different states and sectors. By performing data pre-processing and exploratory data analysis (EDA), going through the data we have identified significant patterns in energy usage and pricing from 2014 to 2024. Additionally, machine learning models include linear regression and random forest, are used to predict the electricity prices, comparatively Random Forest is giving the best accuracy. The CO2 Emissions Dataset shows how energy use and emissions are connected, which helps us understand the relationship between electricity demand and environment sustainability. The results are visualized through an interactive Power BI dashboard, highlighting state-wise

I. INTRODUCTION

This project focuses on analyzing electricity generation and CO2 emissions in the United States to derive actionable insights for policymakers and stakeholders. Using data visualization and predictive modeling, the project aims to explore the relationship between energy consumption trends, fuel types, the adoption of renewable energy, and their impacts on emissions and electricity costs. The study uses datasets from the US Energy Information Administration (EIA), including CO2 emissions data, retail electric sales data, operational electric power data, and state-specific emission data. These datasets provide a comprehensive view of energy consumption, emissions trends, pricing, and revenue dynamics across various sectors and states over multiple decades.

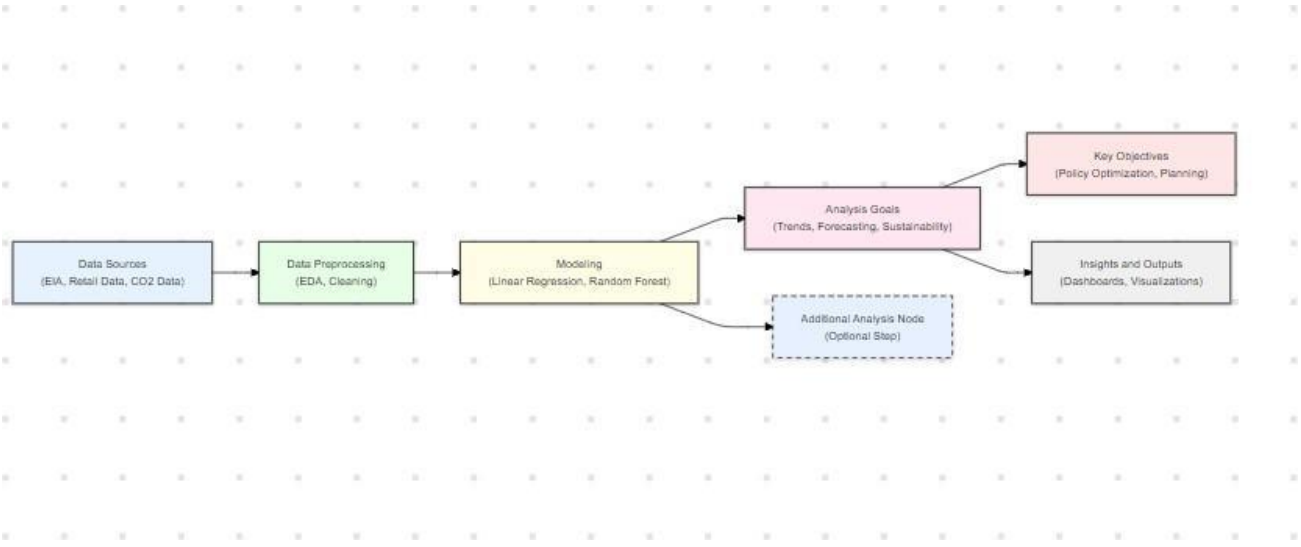


Fig. 1: System Diagram

Key objectives include:

1) Understanding electricity generation and fuel mix patterns across states and sectors.
2) Examining the influence of renewable energy on CO2 emissions and power costs.
3) Analyzing the relationship between energy prices, retail sales, and consumption trends.
4) Forecasting future trends in electricity consumption and emissions for better policy planning.
5) Through robust data pipelines and advanced analytics techniques, such as correlation analysis, univariate exploration, and machine learning models, this project enables stakeholders to optimize energy policies and develop sustainable solutions.

6) Interactive dashboards in Power BI further facilitate decision-making by visualizing key metrics, such as emissions by state, sector-wise trends, and renewable energy impacts.

By integrating advanced modeling with real-time data insights, this project aims to predict electricity prices and environmental sustainability in the United States.

The final stage of the project involves visualizing the results in Power BI. Interactive dashboards are developed to present key metrics such as current CO2 emissions, forecasted electricity prices depending on the various parameters such as state and sector in the United States, and correlations between electricity price and CO2 emissions. These visualizations equip energy providers with the tools to better analyze electricity consumption trends, improve grid efficiency, and develop strategies to manage peak demand or reduce the impact of renewable energy variability and fluctuation in electricity prices.

By combining robust data pipelines, advanced modeling, and interactive visualizations, this project provides a comprehensive solution for energy consumption analysis and demand forecasting. It empowers decision-makers to optimize resource allocation, improve energy efficiency, and ensure a reliable energy supply.

II. DATA DESCRIPTION

This section provides a comprehensive description of the datasets used in this analysis. The first dataset contains CO2 emission data from different types of sectors and states, the second dataset provides electricity prices and third dataset provides emission by fuel. All the datasets are essential for analyzing CO2 emissions and predicting the electricity prices.

A. CO2 Emission Dataset

The energy dataset contains hourly energy consumption data from multiple utility companies in California, including Pacific Gas and Electric (PGAE), Southern California Edison (SCE), San Diego Gas and Electric (SDGE), and others. The data includes energy consumption values measured in megawatt-hours (MWh) and is used to assess energy usage trends across different regions.

1) **Period** (datetime):
   Description: The year when the CO2 emission data was recorded. This field represent the calendar year of the measurement.
2) **SectorId** (String):
   Description: It is an abbreviation for the sector contributing to CO2 emissions. For examples RC stands for Residential, EC stands for Electric Power, TC stands for Transportation.
3) **Sector Name** (String):
   Description: The full name of the sector contributing to CO2 emissions. For instance, ”Residential carbon dioxide emissions,” ”Electric Power carbon dioxide emissions,” etc.
4) **Fuel Id** (String):
   Description: It is an abbreviation for the fuel type associated with CO2 emissions. For examples CO stands for Coal, PE stands for Petroleum, NG stands for Natural Gas.
5) **Fuel Name** (String):
   Description: The full name of the fuel type associated with the CO2 emissions. For instance, ”Coal,” ”Petroleum,” ”Natural Gas,” etc.
6) **StateId** (String):
   Description: The full name of the state associated with the CO2 emissions were recorded. For instance, ”OH”, ”Ohio”, ”WY”, ”Wyoming”, etc.
7) **State Name** (String):
   Description: The full name of the State Name associated with the CO2 emissions. For instance, ”Ohio”, ”Wyoming”.
8) **Value** (Float):
   Description: The total CO2 emissions for the given sector, fuel type, and state in the specified year. The values are measured in million metric tons of CO2.
9) **Value Units** (String):
   Description: The unit of measurement used for the CO2 emission values. In this dataset, the unit is ”million metric tons of CO2,” representing the total amount of carbon dioxide emissions.

B. Electric Sales Dataset

This dataset provides detailed information about electricity sales, customer counts, prices, revenues, and consumption across different states, sectors, and time periods. Below are the variable descriptions:

1) **Period** (String):
   Description: The date or time period when the data was recorded. Typically formatted as YYYY-MM to represent a specific month and year.

2) **StateId** (String):
   Description: A unique identifier or abbreviation for the state. For example, ”MA” (Massachusetts), ”WI” (Wisconsin), etc.
3) **State Description** (String):
   Description: The full name of the state corresponding to the StateId. For instance, ”Massachusetts,” ”Wisconsin,” etc.
4) **Sector Id** (String):
   Description: A unique identifier or abbreviation for the sector. Examples include ”IND” (Industrial), ”RES” (Residential), ”TRA” (Transportation), and ”ALL” (All Sectors)
5) **Sector Name** (String):
   Description: The full name of the sector. For instance, ”industrial,” ”residential,” ”transportation,” etc.
6) **Customers** (Float):
   Description: The number of customers in the specified state and sector during the given period. Measured as the count of customers.
7) **Price** (Float):
   Description: The average price of electricity during the period. Measured in cents per kilowatt-hour (c/kWh).
8) **Revenue** (Float):
   Description: A categorical description of the weather conditions on the recorded date, such as ”Clear”, ”Rainy”, ”Cloudy”, or ”Partly Cloudy”. This helps understand the general weather context, which can drive energy usage trends.
9) **Sales** (Float):
   Description: The total electricity sold during the period. Measured in million kilowatt-hours (kWh).
10) **Customer Units** (String):
    Description: The unit of measurement for the Customers variable. In this dataset, the unit is ”number of customers.
11) **Price Units** (String):
    Description: The unit of measurement for the Price variable. In this dataset, the unit is ”cents per kilowatt-hour.”
12) **Revenue Units** (String):
    Description: Description: The unit of measurement for the Revenue variable. In this dataset, the unit is ”million dollars.”
13) **Sales Units** (String):
    Description: The unit of measurement for the Sales variable. In this dataset, the unit is ”million kilowatt-hours.”

C. Emission by Fuel Dataset

This dataset provides detailed emissions data categorized by state, fuel type, and emission metrics. Below are the variable descriptions:

1) **Period** (String):
   Description: The year when the emission data was recorded. This field represents the calendar year of the measurement.
2) **StateId** (String):
   Description: A unique identifier or abbreviation for the state. Examples include ”AZ” (Arizona), ”WY” (Wyoming), etc.
3) **State Description** (String):
   Description: The full name of the state corresponding to the StateId. For instance, ”Arizona,” ”Wyoming,” etc.
4) **FuelId** (String):
   Description: A unique identifier or abbreviation for the fuel type. Examples include ”COL” (Coal), ”PET” (Petroleum), ”NG” (Natural Gas), and ”ALL” (Total for all fuels).
5) **Fuel Description** (String):
   Description: The full name of the fuel type. For instance, ”Coal,” ”Petroleum,” ”Natural Gas,” etc.
6) **CO2 Rate (lbs/MWh)** (Float):
   Description: The emission rate of carbon dioxide in pounds per megawatt-hour (lbs/MWh) for the specified fuel type.
7) **CO2 Thousand Metric Tons** (Float):
   Description: The total emissions of carbon dioxide measured in thousand metric tons.
8) **NOx Rate (lbs/MWh)** (Float):
   Description: The emission rate of nitrogen oxides in pounds per megawatt-hour (lbs/MWh) for the specified fuel type.
9) **NOx Short Tons** (int):
   Description: The total emissions of nitrogen oxides measured in short tons.
10) **SO2 Rate (lbs/MWh)** (Float):
    Description: The emission rate of sulfur dioxide in pounds per megawatt-hour (lbs/MWh) for the specified fuel type.
11) **SO2 Short Tons** (int):
    Description: The total emissions of sulfur dioxide measured in short tons.

12) **CO2 Rate Units** (String):
Description: The unit of measurement for the CO2 Rate variable. In this dataset, it is ”pounds per megawatthour.”
13) **CO2 Thousand Metric Tons Units** (String):
Description: The unit of measurement for the CO2 Thousand Metric Tons variable. In this dataset, it is ”thousand metric tons.”
14) **NOx Rate Units** (String):
Description: The unit of measurement for the NOx Rate variable. In this dataset, it is ”pounds per megawatthour.”
15) **NOx Short Tons Units** (String):
Description: The unit of measurement for the NOx Short Tons variable. In this dataset, it is ”short tons.”
16) **SO2 Rate Units** (String):
Description: The unit of measurement for the SO2 Rate variable. In this dataset, it is ”pounds per megawatthour.”
17) **SO2 Short Tons Units** (String):
Description: The unit of measurement for the SO2 Short Tons variable. In this dataset, it is ”short tons.”

III. DATA EXPLORATION

Exploratory Data Analysis (EDA)

- CO2 Emissions Dataset: Univariate analysis of emissions patterns over time.
- State-Level Insights: Analysis of emissions by sector and state.
- Pricing Data: Examination of average prices and revenues across states.

CO2 Emissions Dataset: Univariate Analysis of Emissions Patterns Over Time

The analysis tracks CO2 emissions trends from 2014 to 2022 across various sectors. It highlights key patterns such as peaks, emission state and sector wise.

State-Level Insights: Analysis of Emissions by Sector and State:

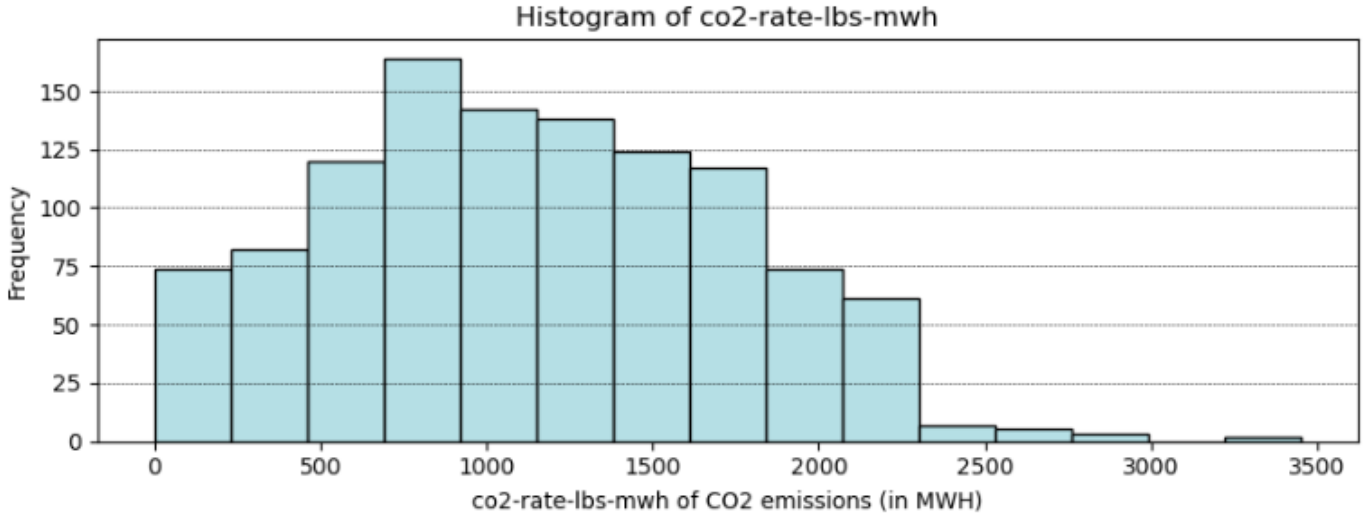
This section breaks down state and sector wise, revealing disparities across region and industries.

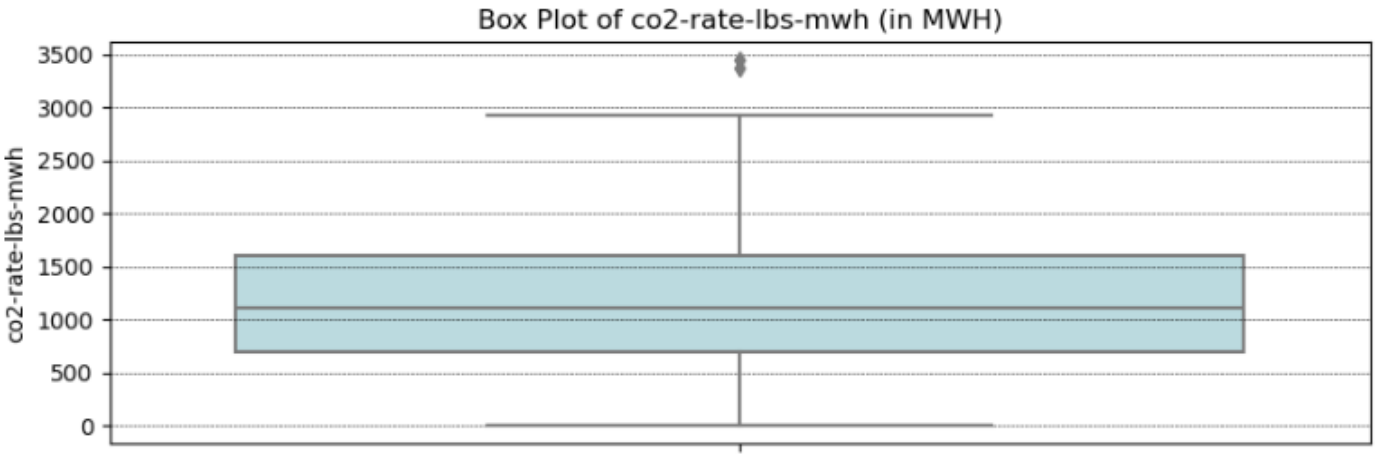
Pricing Data: Examination of Average Prices and Revenues Across States:

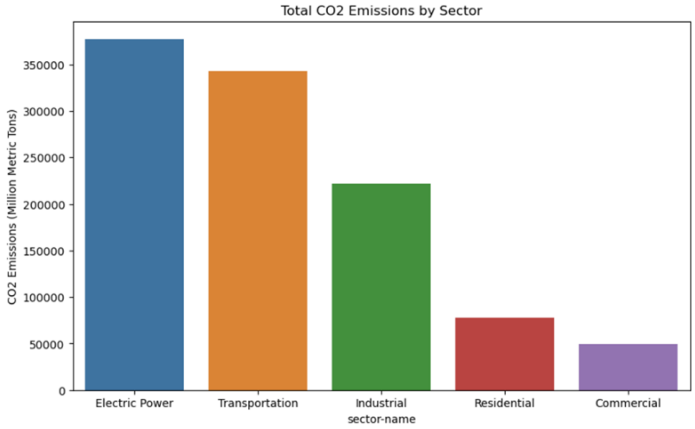
It identifies the relation between energy costs and consumption. Also, it examines state-specific electricity prices and revenue trends.

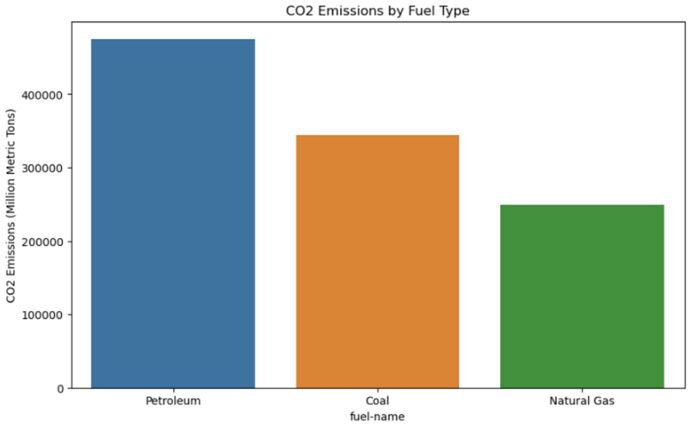
Correlation Analysis

Investigates relationships between key variables like emissions, fuel consumption, and electricity pricing to identify influential factors.

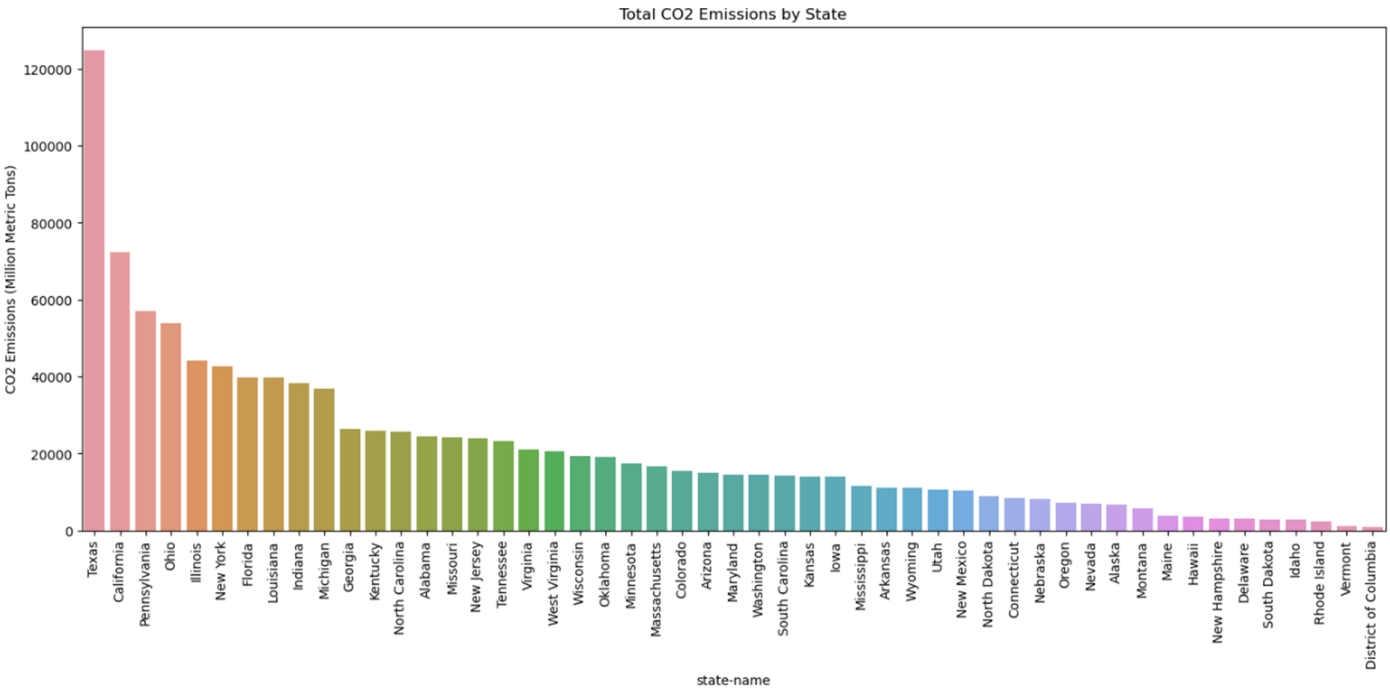




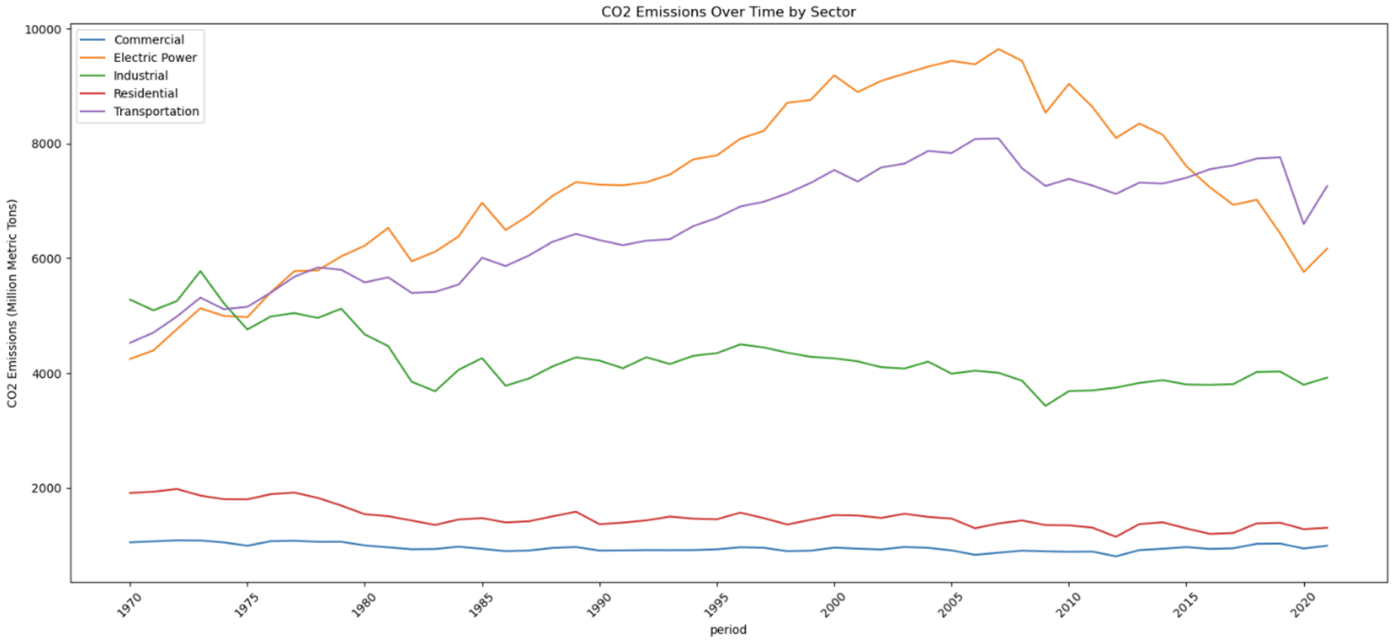




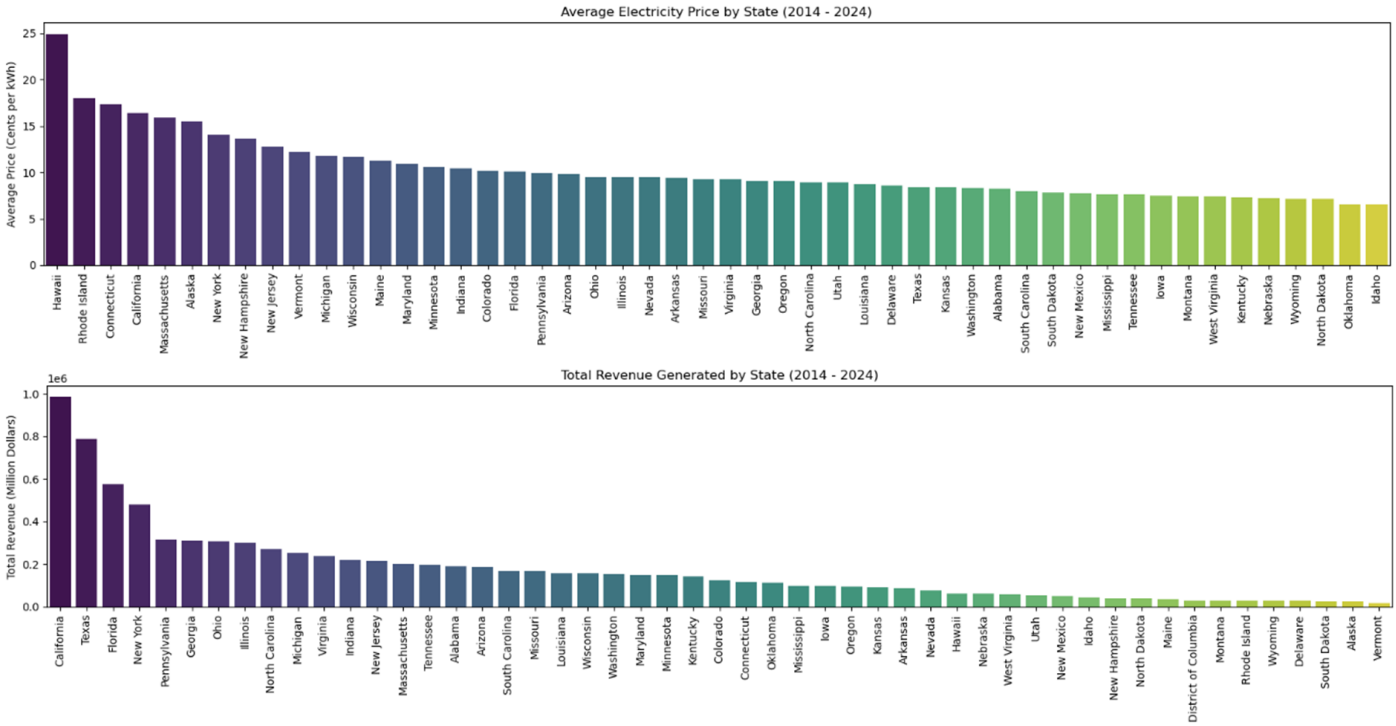
CO2 Emission by State for All Sectors



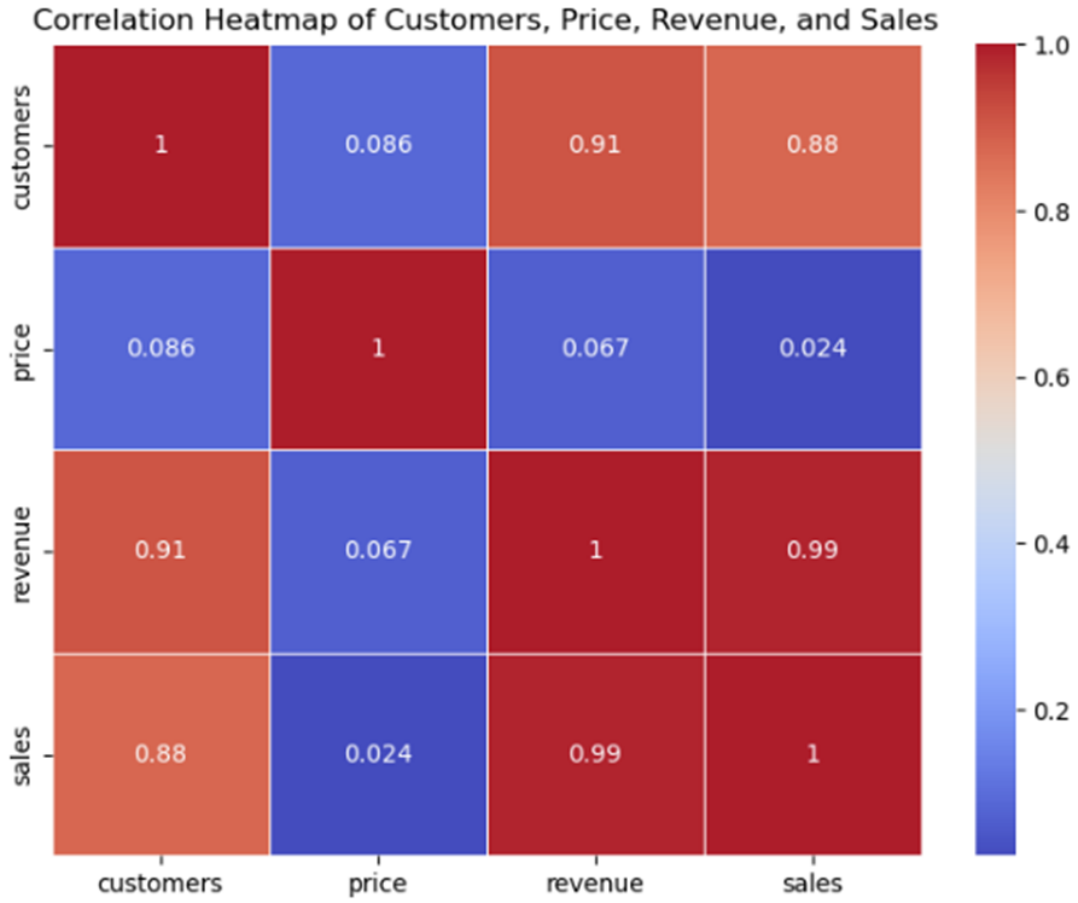
CO2 Emissions Over Time by Sector



Average Price & Total Revenue by State



Correlation Analysis



IV. MACHINE LEARNING ANALYSIS

Model Selection

Predicting electricity prices is a challenging task in energy markets due to the multidimensional market dynamics. To develop an effective machine learning model, this project evaluates two machine learning methods: Linear Regression and Random Forest algorithms. Linear Regression was selected for its simplicity and transparent parameter interpretation, enabling direct analysis of feature importance and marginal effects on price formation. The Random Forest model, an ensemble learning method that builds multiple decision trees, was selected to capture potential non-linear relationships and interaction effects among predictor variables. The comparative analysis of these models' predictive performance aims to identify the optimal model for accurate electricity price forecasting.

Feature selection

The purpose of feature selection is to identify the most relevant features in the dataset. There were seven features selected for model training. Input features included the number of customers, the primary demand indicator, and state-level descriptions to capture regional variations in electricity pricing. Sector-specific variables were represented through one-hot encoding of five distinct sectors, allowing the models to account for differences across economic activities. Temporal features included year and month. The target feature was price. For data splitting, 60% of the data was used for training, 20% for validation, and the remaining 20% for testing the model performance. This structured approach ensured the models were trained on diverse patterns and evaluated for accuracy.

Sample Dataset After Feature Engineering

period	stateid	stateDescription	customers	price	stateDescription_mapping	sectorid_ALL	sectorid_COM	sectorid_IND	sectorid_RES	sectorid_TRA	Year	Month
2024-06-01	MA	Massachusetts	9850.0	18.05	23	0	0	1	0	0	2024	6
2024-06-01	WI	Wisconsin	3222461.0	13.06	60	1	0	0	0	0	2024	6
2024-06-01	WV	West Virginia	0.0	0.00	59	0	0	0	0	1	2024	6
2024-06-01	WV	West Virginia	864734.0	15.48	59	0	0	0	1	0	2024	6
2024-06-01	WV	West Virginia	10953.0	7.84	59	0	0	1	0	0	2024	6

Results

The following image shows sample predictions from both models. A notable example is row 7, where Linear Regression predicted 14.86 for an actual value of 32.39, which is underestimated. In contrast, Random Forest predicted 32.14, which is much closer to the actual value.

Sample Predictions from Linear Regression and Random Forest Models

	Actual	Predicted
0	12.29	13.42
1	10.56	11.46
2	9.07	11.86
3	10.61	15.20
4	0.00	7.21
5	10.67	11.70
6	9.37	12.39
7	32.39	14.86
8	18.87	14.56
9	5.92	5.75

	Actual	Predicted
0	12.29	11.71
1	10.56	10.61
2	9.07	8.59
3	10.61	11.50
4	0.00	0.00
5	10.67	10.43
6	9.37	9.39
7	32.39	32.14
8	18.87	17.57
9	5.92	6.87

For evaluation, four metrics were selected to assess model performance:

MAE (Mean Absolute Error): Represents the average absolute difference between predictions and actual values. Lower is better.

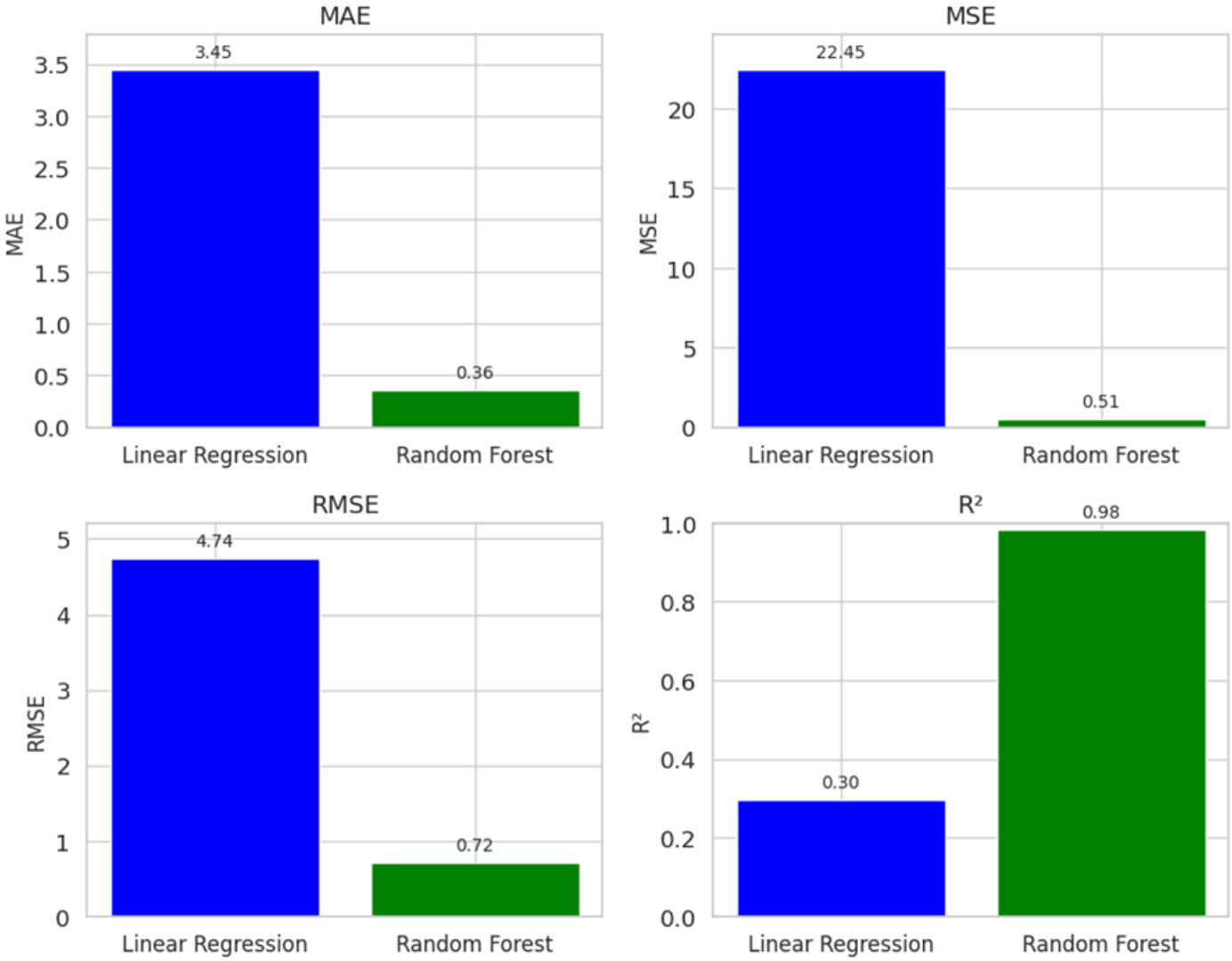
MSE (Mean Squared Error): Measures the mean of squared differences, giving more weight to larger errors. Lower is better.

RMSE (Root Mean Squared Error): The square root of MSE, expressed in the same units as the target variable for better interpretability. Lower is better.

R^2 (R-squared): Indicates the proportion of variance in the target variable explained by the model. Higher is better.

The image below shows a comparison of these metrics across the models.

Performance Metrics Comparison Between Linear Regression and Random Forest



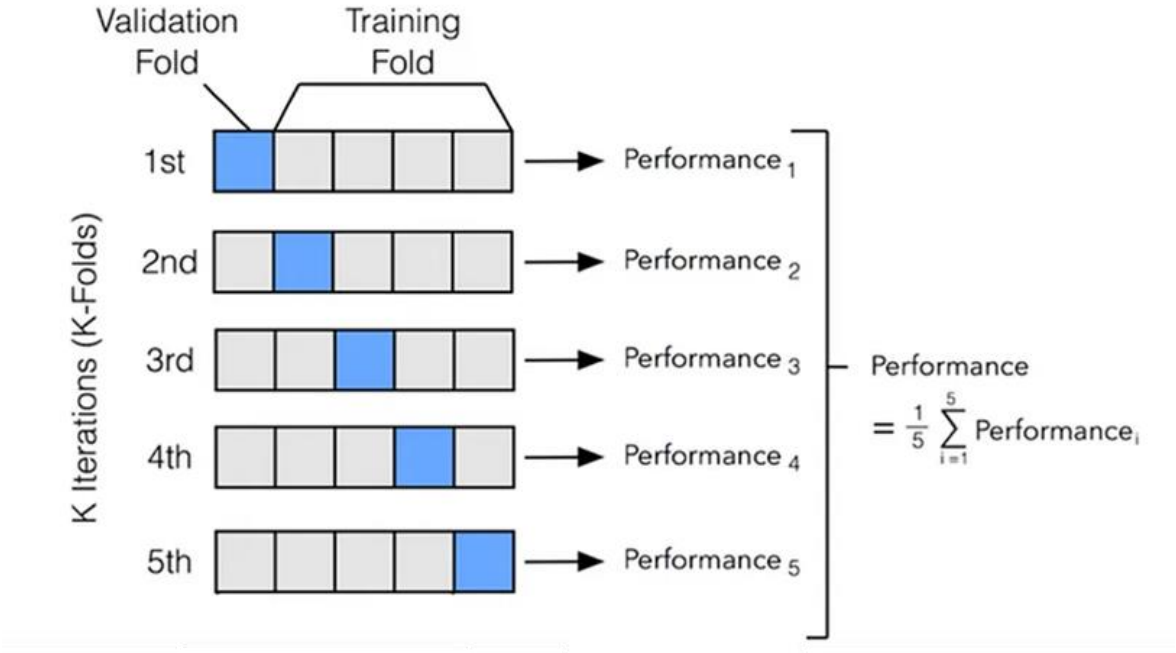
Based on the performance comparison, the Random Forest model performed better than Linear Regression in predicting electricity prices. Random Forest had a lower Mean Absolute Error (MAE) of 0.36, meaning it made smaller mistakes on average, compared to Linear Regression's 3.45. Its Mean Squared Error (MSE) was also much lower at 0.51, showing it handled larger errors better than Linear Regression, which had an MSE of 22.45. The Root Mean Squared Error (RMSE) was 0.72 for Random Forest, much smaller than Linear Regression's 4.74. Finally, Random Forest had an R^2 value of 0.98, meaning it explained 98% of the variation in electricity prices, while Linear Regression only explained 30% ($R^2 = 0.30$).

Evaluation

Cross validation

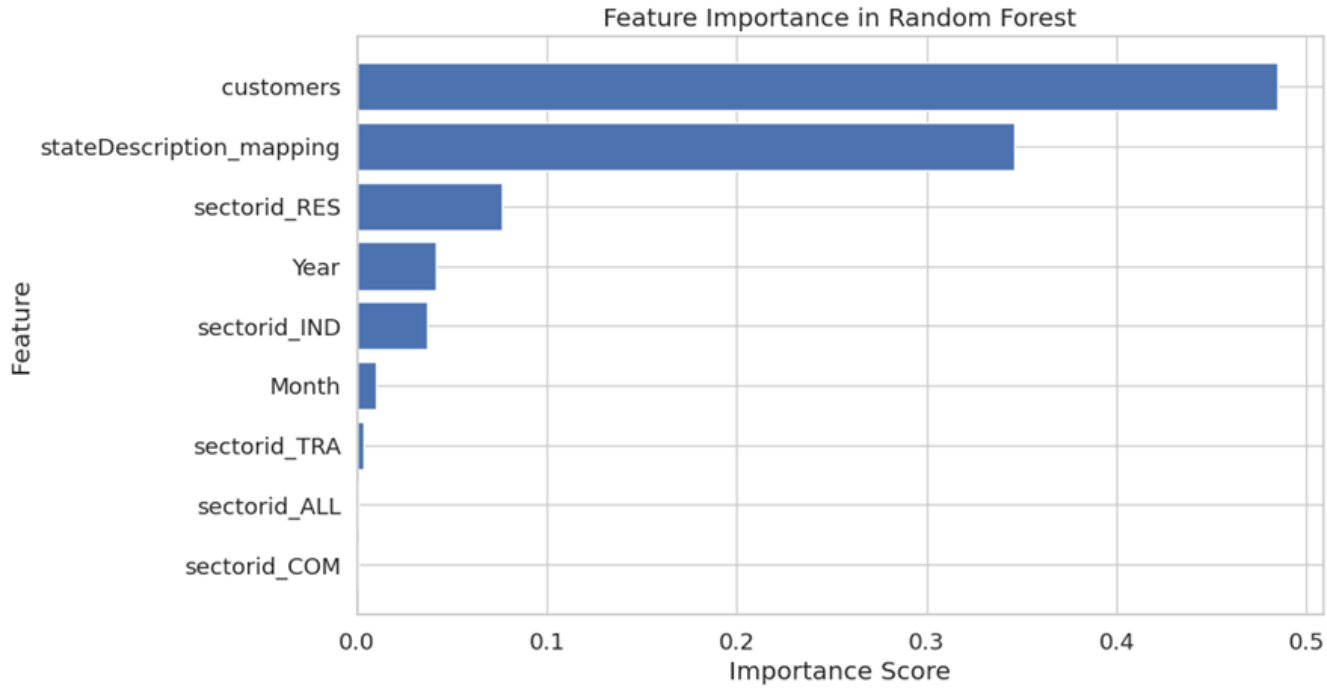
To validate the model's ability to generalize to new, unseen data, a 5-fold cross-validation technique was implemented. The dataset was divided into five folds, with each fold taking turns as the validation set while the remaining four folds were used for training. This process is to make sure that every data point was used for both training and validation, providing comprehensive evaluation. The results showed that the Random Forest model generalized better compared to Linear Regression, confirming its reliability for electricity price prediction.

Illustration of 5-Fold Cross-Validation Process



Feature importance analysis was conducted to understand the factors influencing electricity price predictions. The analysis showed that the number of customers and geographic location (stateDescription_mapping) were the most significant drivers of price variations, with the highest importance scores. Sector-specific variables, such as sectorid_RES (residential) and temporal features like Year, also contributed meaningfully, though to a lesser extent. Other features, such as Month and additional sector identifiers, had minimal impact. These insights highlight the key variables shaping electricity prices and offer valuable understanding for decision-making in energy markets.

Feature Importance in Random Forest Model



After testing both models, Random Forest is as the optimal model due to its higher accuracy and ability to capture complex relationships in the data. Key factors affecting electricity prices were identified as the number of customers, geographic location, and demand from the residential sector.

This model can assist utility companies in forecasting prices more accurately, optimizing resource planning, and improving pricing strategies. Additionally, it provides insights for policymakers and businesses to understand electricity demand trends and make

informed decisions for better energy management.

Power BI Report

Purpose of the dashboard

The Electricity Prices and CO2 Emissions Dashboard is a user-friendly tool that brings together two crucial aspects of the energy sector - electricity pricing and environmental impact.

Navigation Made Simple

At the top of the screen, you'll find a straightforward navigation bar that lets you easily switch between different views. Whether you're interested in sales figures, price patterns, or emission data, everything is just a click away.

Two Main Entry Points

The dashboard welcomes you with two clear choices:

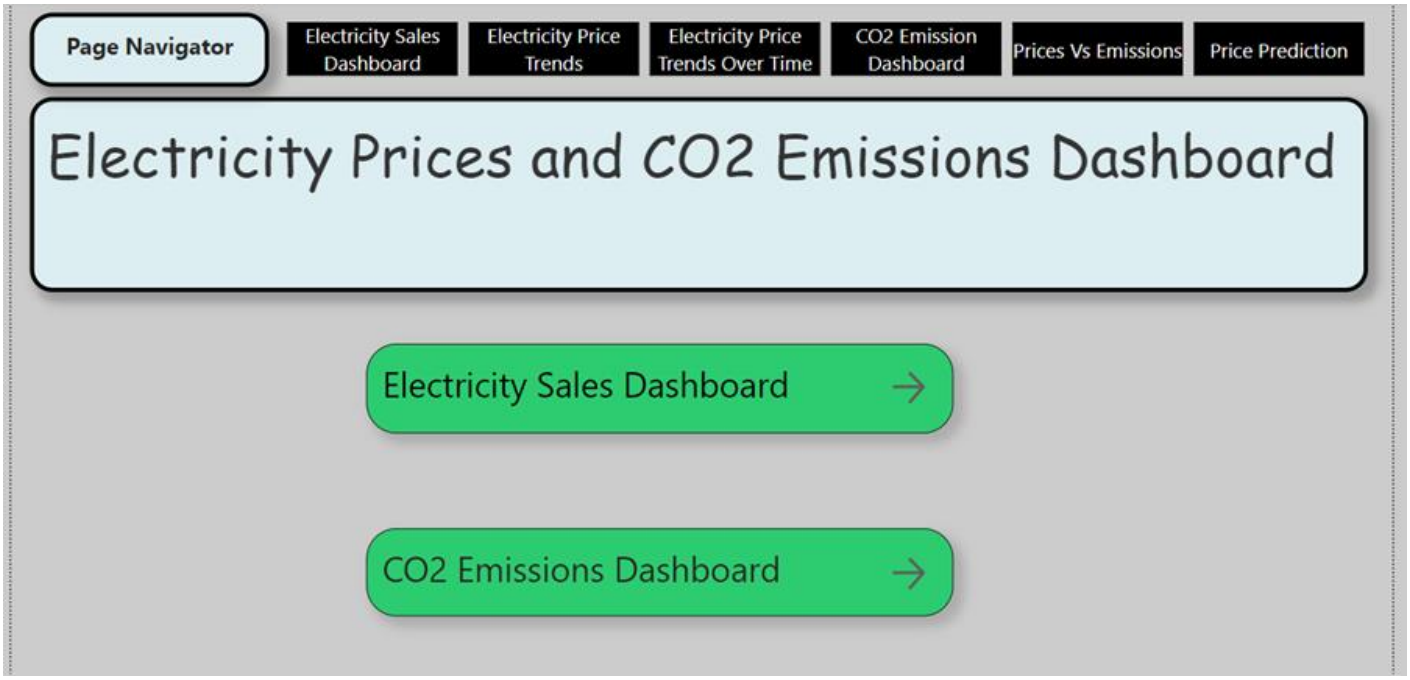
- A green button leading to the Electricity Sales Dashboard, where you can explore all things related to power sales and pricing
- Another green button directing you to the CO2 Emissions Dashboard, focusing on environmental impact data

What Makes It Special

This dashboard takes complex data and presents it in a way that makes sense to everyone. You can:

- Track how electricity prices change over time
- See how prices and emissions relate to each other
- Look at sales trends and patterns
- Get insights into future price movements through predictions

It's like having an energy market expert and environmental analyst at your fingertips, helping you understand both the business and environmental sides of electricity consumption in one place.



Extract:

After performing Exploratory Data Analysis in python, data is saved in csv files. We are having total of five flat files:

- CO2 Emission by Electricity Generation.
- CO2 Emission by Sector and Fuels
- Electricity Sales Data
- Linear Regression Predicted Prices Data
- Random Forest Predicted Prices Data

Below are the attached files:

- CO2 Emission by Electricity Generation.
- CO2 Emission by Sector and Fuels
- Electricity Sales Data
- Linear Regression Predicted Prices Data
- Random Forest Predicted Prices Data

Above files are being generated by Python code and getting saved in local folder. From there, these files have been loaded into Power BI through Power Query Editor.

Transform:

First table **co2_emission_by_SectorsAndFuels** is being created using **CO2 Emission by Sector and Fuels** flat file. This source file is having data from 1970 to 2022. We have removed redundant data from this file and filter data for period ranging from 2014 to 2022.

The period column in this file was not getting recognized by Power BI engine, so, we have made new column named Period Time using Column from Examples option and dummy dates are created based on the criteria of first day of each year.

Additionally, this data source is also having rows aggregated based on sector id and fuel which we have filtered out to remove unexpected data.

Then, we have created Primary Key named CO2-emission-id, merging Period Time, State Id, Sector Id and one foreign key CO2_Emission_Fuel_ID to merge it with another data source named CO2 emissions by Electricity Generation.

Second table we have created from electricity_data.xlsx file for getting electricity price, revenue, sales and customers data. In this source, we have filtered the rows based on All sectors to remove unwanted values.

We have modified the Sector ID to align it with the data in other table as source were having different Sector IDs for sector names which were consistent.

Added country column with hardcoded value to have geospatial data visualized perfectly on map along with creating Primary key to join other tables.

co2_emission_by_Electricity_Generation:

This table is created for having data of CO2 emissions data based on particular source of electricity generation. CO2 Emission by Electricity Generation file is used to populate this table.

This table is filtered to have data from 2014 – 2023 along with creation of Primary key Fuel_Emission_ID to link it with CO2 emissions table already created.

Electricity Sales Predicted Data & Electricity Sales RF table is created to bring Predicted values from ML model implemented in Python and mentioned above.

These tables are then structured similarly to have both data appended into one table which is further being joined with Electricity Sales Data.

After creating these tables, data model is created based on the suitable and needed relationship.

UI/ Dashboard Layer:

In our report, we are using below KPIs:

1. CO2 Emissions Change Percentage
2. CO2 Emissions Trend
3. Electricity Price Trend
4. Forecast Accuracy (%)
5. Price Change Percentage

These KPIs have been created using DAX language.

Price Change Percentage =

```
VAR PriceDiff = [Price Last Year] - [Price First Year]

RETURN

DIVIDE(PriceDiff, [Price First Year], 0) * 100
```

Forecast Accuracy (%) =

```
IF (

   SUM('Electricity Sales Data'[price]) <> 0,

   (1 - ABS(SUM('Electricity Sales Data'[price]) - SUM('Electricity Sales Predicted Consolidated'[Predicted_price_aggregated])) /
SUM('Electricity Sales Data'[price])) * 100,

   BLANK()

)
```

Electricity Price Trend =

```
"Electricity prices have " &

IF([Price Change Percentage] > 0, "increased", "decreased") &

" by " & FORMAT(ABS([Price Change Percentage]), "0.00") &

"% over the last 9 years."
```

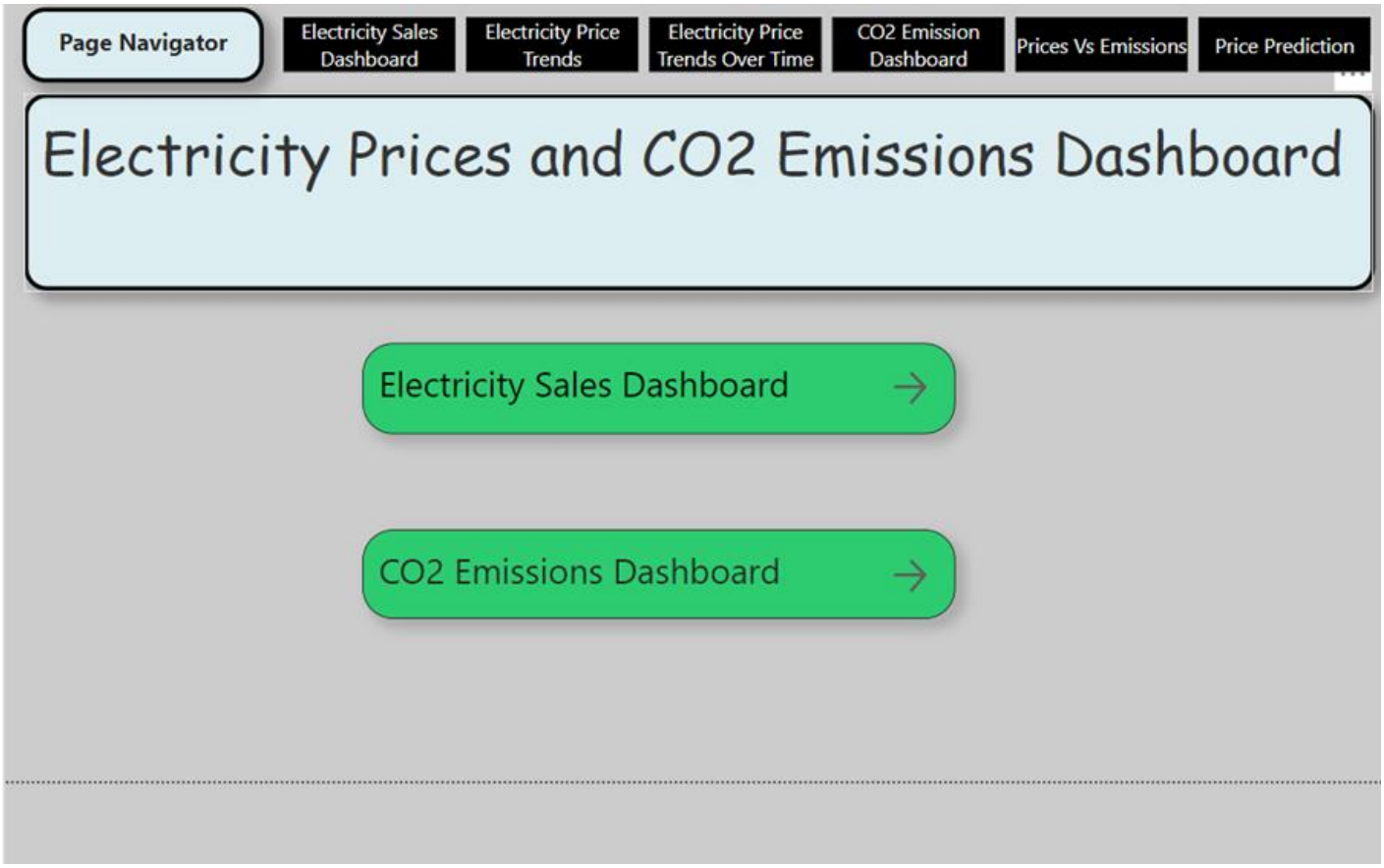
```
CO2 Emissions Trend =

"CO2 emissions have " &

IF([CO2 Emissions Change Percentage] < 0, "decreased", "increased") &

" by " & FORMAT(ABS([CO2 Emissions Change Percentage]), "0.00") & "% over the last 8 years."
```

```
CO2 Emissions Change Percentage =

VAR EmDiff = [CO2 Emissions Last Year]-[CO2 Emissions First Year]

RETURN

DIVIDE(EmDiff, [CO2 Emissions First Year], 0) * 100
```

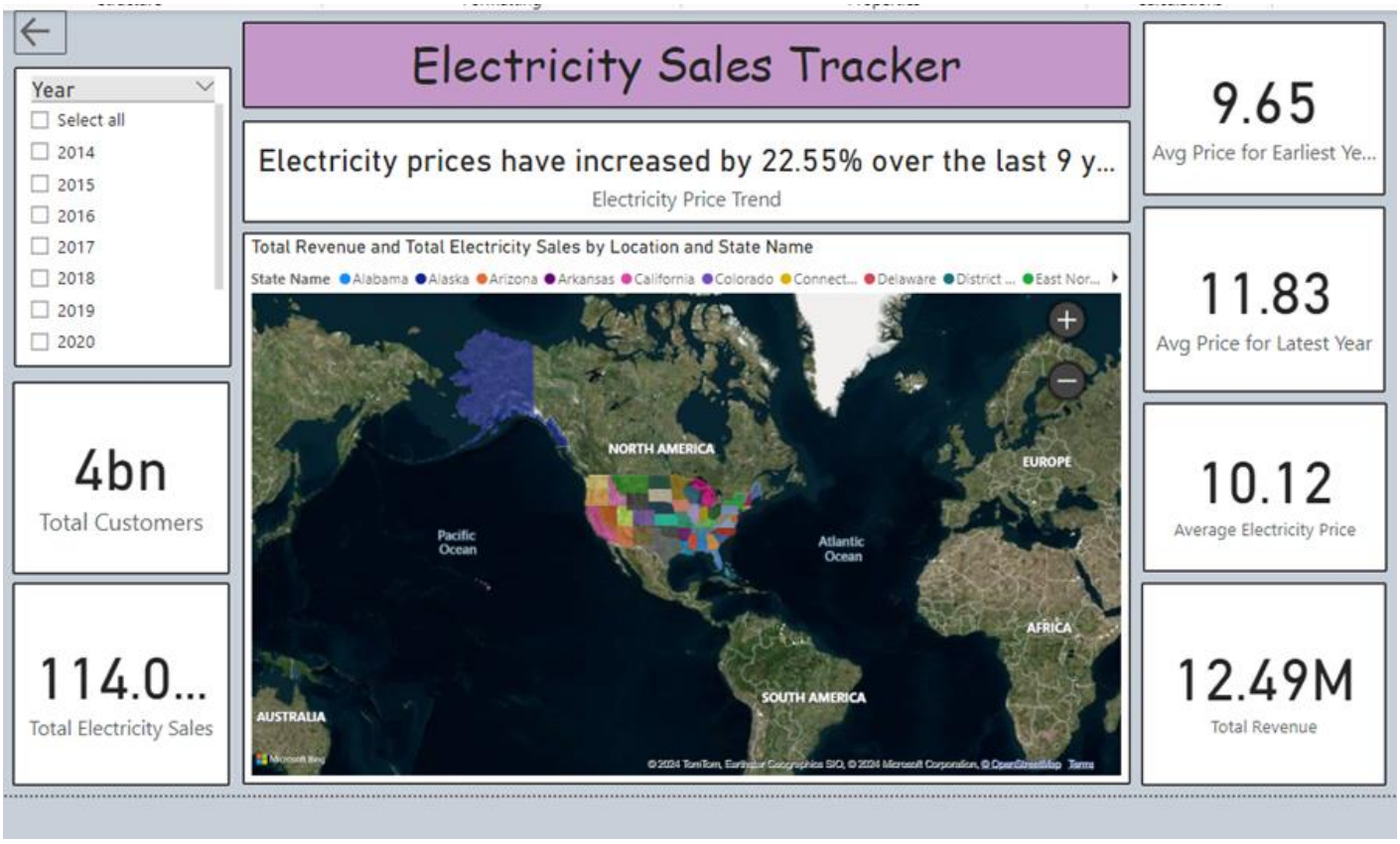
UI layer consists of below sheets:

- Home Page
- Electricity Sales Dashboard
- Electricity Price Trend
- Electricity Price Trends Over Time
- CO2 Emission Dashboard
- Prices Vs Emissions
- Price Prediction.

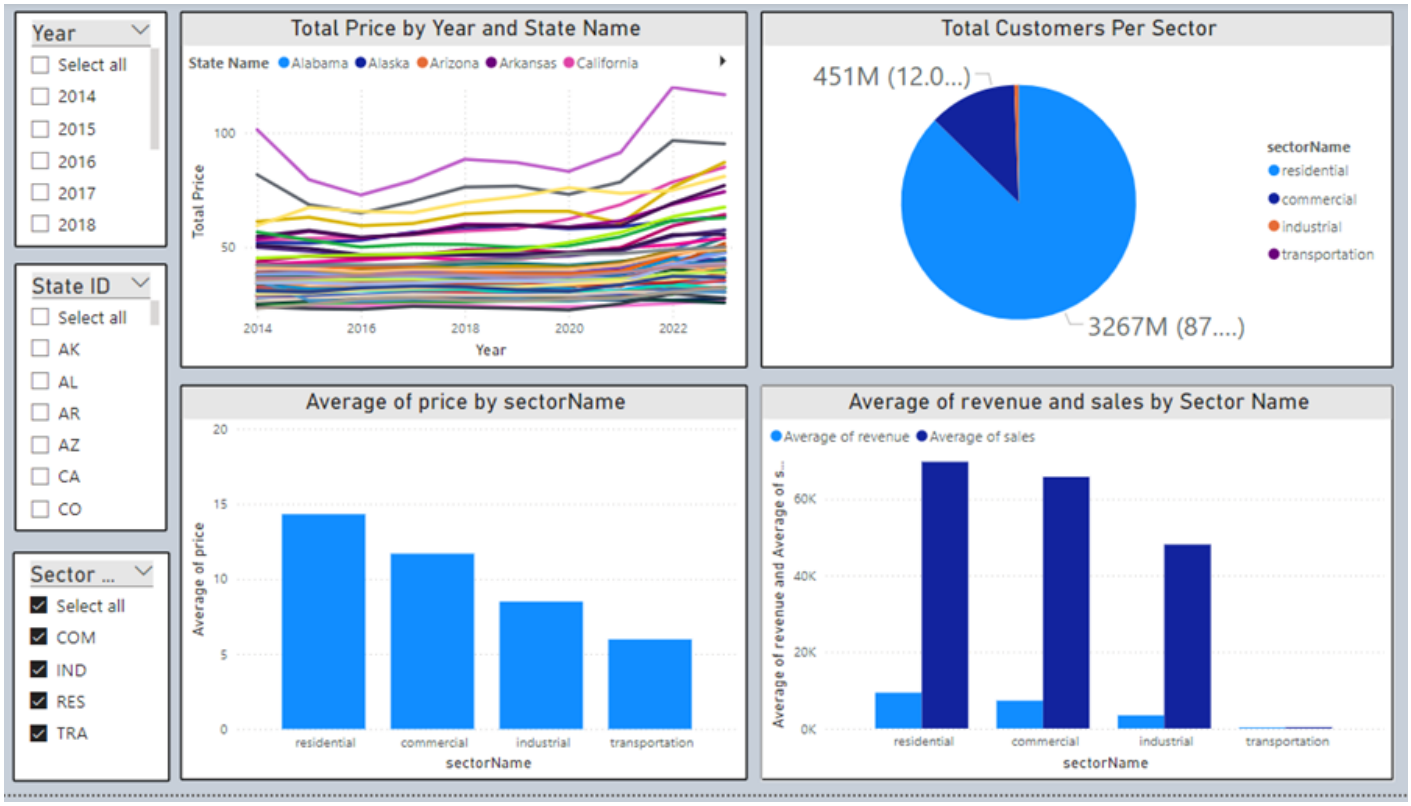
1. Home Page: This Page consists of Page Navigator at top for all the pages along with link for Electricity Sales Dashboard and CO2 Emissions Dashboard.



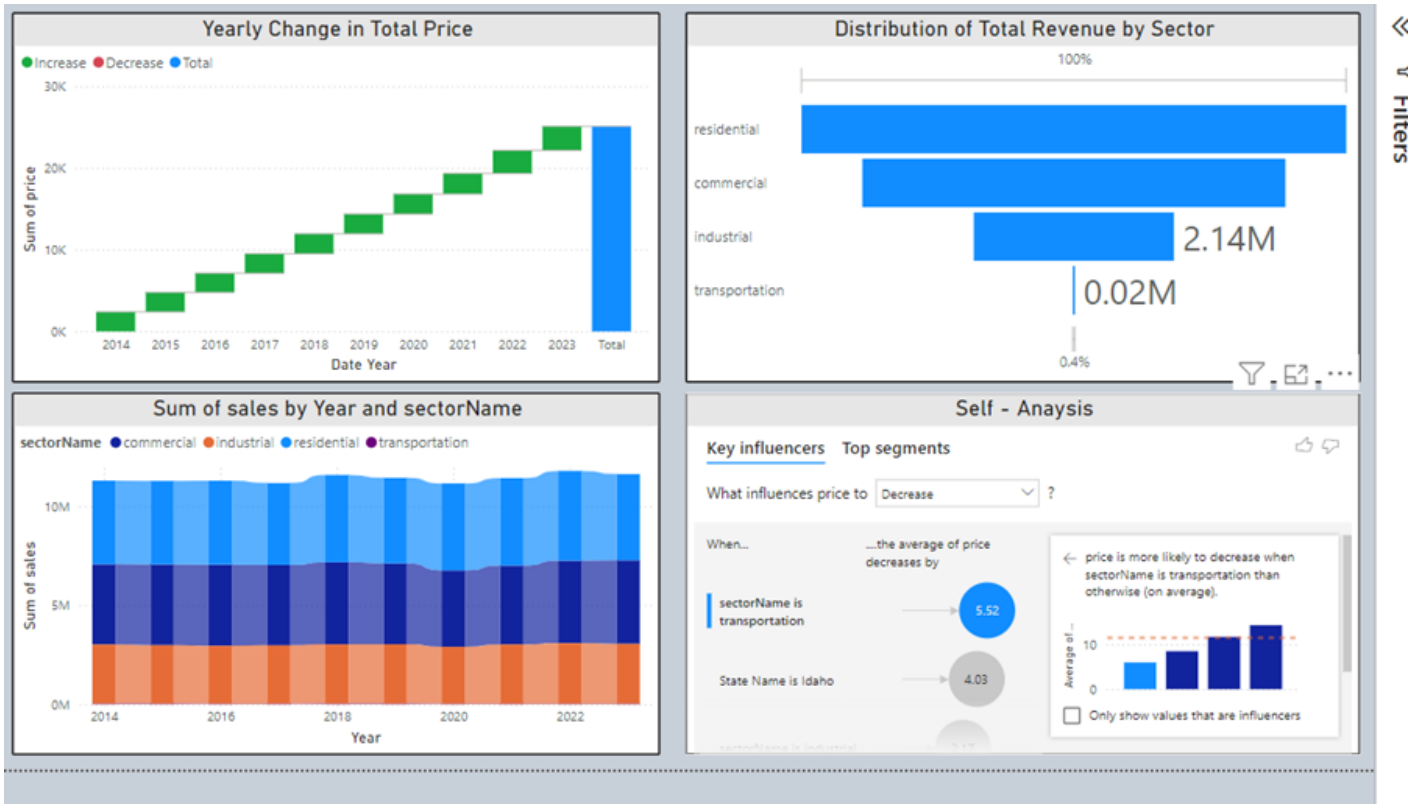
2. Electricity Sales Dashboard: This page reveals Electric Prices Trend along with KPI cards Average Electricity Price, total customers, total electricity sales, average price for first year (2014) and latest year (2023) and map focusing on Total Revenue and Total Electric Sales.



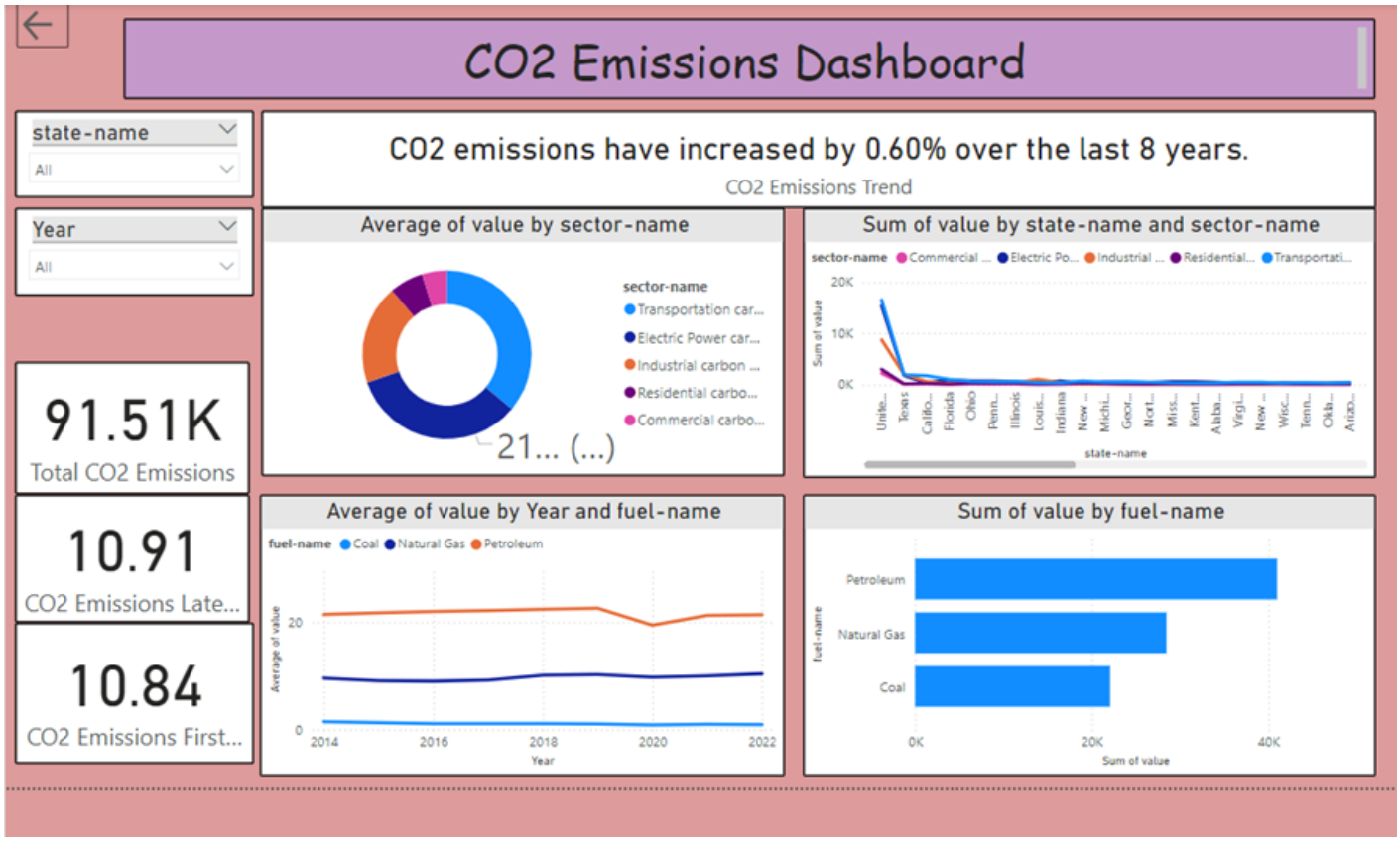
3. Electricity Price Trends: This page focuses on Total Price by year and State Name and sector.



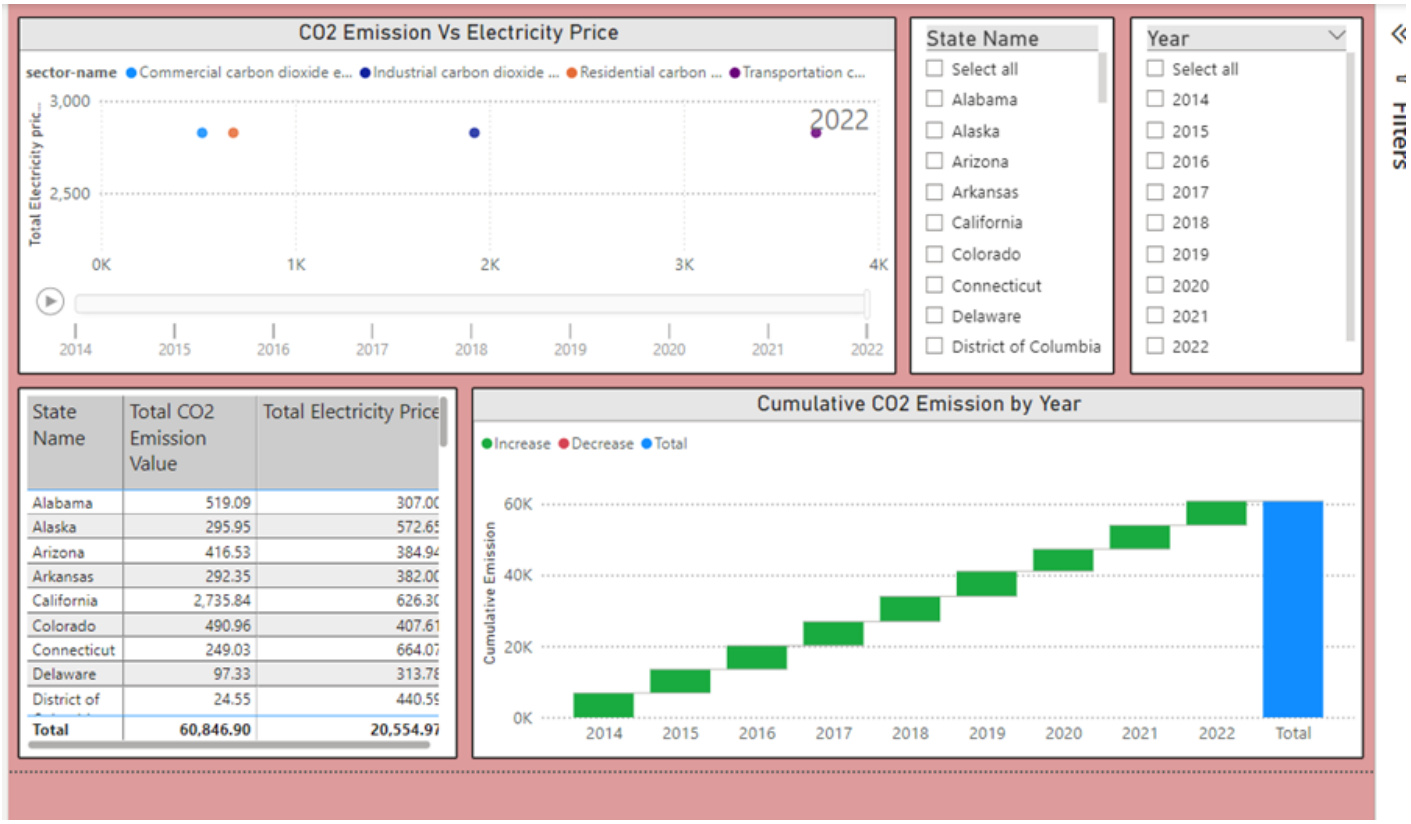
4. Electricity Price Trends over Time: This page reveals Total price change over years in the form of waterfall report. Funnel chart displays total revenue by sector. Ribbon chart tells about cumulative Sales by Sector and special self- analysis visualization which is inbuilt feature of Power BI that tells Price to either increase / decrease based on affecting parameters and tells at what data point's combination, price is going to increase / decrease.



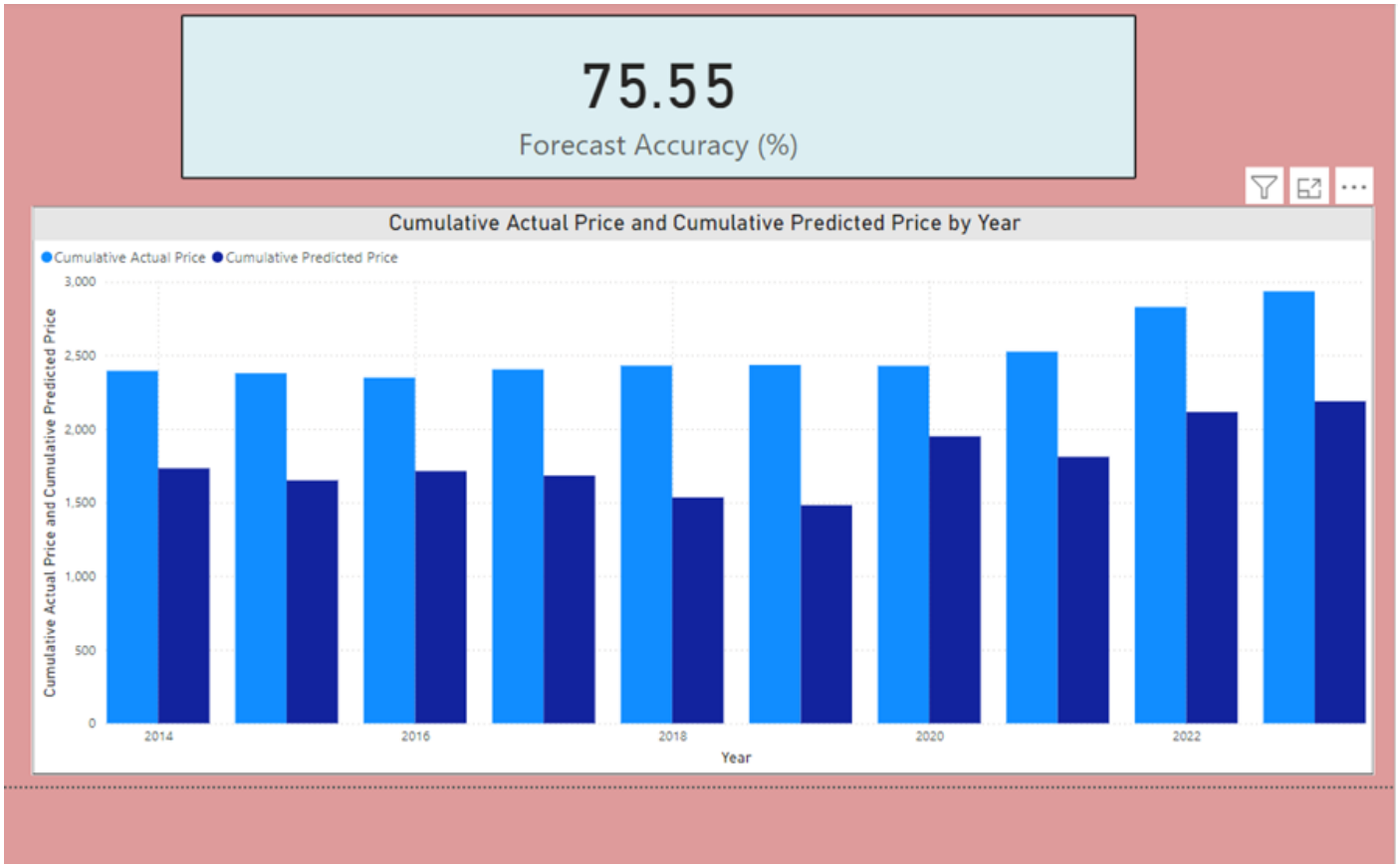
5. CO2 Emission Dashboard: This page shows trend in CO2 emissions over last 8 years from 2014 to 2022, along with KPI cards showing CO2 emissions measures and visualization charts showing average of CO2 emissions against sector, fuel type and year.



6. Prices Vs Emissions: This page mainly features CO2 emissions increase pattern over the years from 2014-2022 accounting for change per each year. CO2 emission be Electricity price through scatter plot.



7. Price Prediction: This page compares Actual Price Vs Predicted Price data using ML model along with KPI card computing Forecast Accuracy based on year.



Power BI Report

Purpose of the dashboard

The Electricity Prices and CO2 Emissions Dashboard is a user-friendly tool that brings together two crucial aspects of the energy sector - electricity pricing and environmental impact.

Navigation Made Simple

At the top of the screen, you'll find a straightforward navigation bar that lets you easily switch between different views. Whether you're interested in sales figures, price patterns, or emission data, everything is just a click away.

Two Main Entry Points

The dashboard welcomes you with two clear choices:

- A green button leading to the Electricity Sales Dashboard, where you can explore all things related to power sales and pricing
- Another green button directing you to the CO2 Emissions Dashboard, focusing on environmental impact data

What Makes It Special

This dashboard takes complex data and presents it in a way that makes sense to everyone. You can:

- Track how electricity prices change over time
- See how prices and emissions relate to each other

- Look at sales trends and patterns
- Get insights into future price movements through predictions

It's like having an energy market expert and environmental analyst at your fingertips, helping you understand both the business and environmental sides of electricity consumption in one place.

Extract:

After performing Exploratory Data Analysis in python, data is saved in csv files. We are having total of five flat files:

- CO2 Emission by Electricity Generation.
- CO2 Emission by Sector and Fuels
- Electricity Sales Data
- Linear Regression Predicted Prices Data
- Random Forest Predicted Prices Data

Below are the attached files:

- CO2 Emission by Electricity Generation.
- CO2 Emission by Sector and Fuels
- Electricity Sales Data
- Linear Regression Predicted Prices Data
- Random Forest Predicted Prices Data

Above files are being generated by Python code and getting saved in local folder. From there, these files have been loaded into Power BI through Power Query Editor.

Transform:

First table co2_emission_by_SectorsAndFuels is being created using CO2 Emission by Sector and Fuels flat file. This source file is having data from 1970 to 2022. We have removed redundant data from this file and filter data for period ranging from 2014 to 2022.

The period column in this file was not getting recognized by Power BI engine, so, we have made new column named Period Time using Column from Examples option and dummy dates are created based on the criteria of first day of each year.

Additionally, this data source is also having rows aggregated based on sector id and fuel which we have filtered out to remove unexpected data.

Then, we have created Primary Key named CO2-emission-id, merging Period Time, State Id, Sector Id and one foreign key CO2_Emission_Fuel_ID to merge it with another data source named CO2 emissions by Electricity Generation.

Second table we have created from electricity_data.xlsx file for getting electricity price, revenue, sales and customers data. In this source, we have filtered the rows based on All sectors to remove unwanted values.

We have modified the Sector ID to align it with the data in other table as source were having different Sector IDs for sector names which were consistent.

Added country column with hardcoded value to have geospatial data visualized perfectly on map along with creating Primary key to join other tables.

co2_emission_by_Electricity_Generation:

This table is created for having data of CO2 emissions data based on particular source of electricity generation. CO2 Emission by Electricity Generation file is used to populate this table.

This table is filtered to have data from 2014 – 2023 along with creation of Primary key Fuel_Emission_ID to link it with CO2 emissions table already created.

Electricity Sales Predicted Data & Electricity Sales RF table is created to bring Predicted values from ML model implemented in Python and mentioned above.

These tables are then structured similarly to have both data appended into one table which is further being joined with Electricity Sales Data.

After creating these tables, data model is created based on the suitable and needed relationship.

UI/ Dashboard Layer:

In our report, we are using below KPIs:

1. CO2 Emissions Change Percentage
2. CO2 Emissions Trend
3. Electricity Price Trend
4. Forecast Accuracy (%)
5. Price Change Percentage

These KPIs have been created using DAX language.

```
Price Change Percentage =

VAR PriceDiff = [Price Last Year] - [Price First Year]

RETURN

DIVIDE(PriceDiff, [Price First Year], 0) * 100
```

```
Forecast Accuracy (%) =

IF (

    SUM('Electricity Sales Data'[price]) <> 0,

    (1 - ABS(SUM('Electricity Sales Data'[price]) - SUM('Electricity Sales Predicted Consolidated'[Predicted_price_aggregated])) /
SUM('Electricity Sales Data'[price])) * 100,

    BLANK()

)
```

```
Electricity Price Trend =

"Electricity prices have " &

IF([Price Change Percentage] > 0, "increased", "decreased") &

" by " & FORMAT(ABS([Price Change Percentage]), "0.00") &

"% over the last 9 years."
```

```
CO2 Emissions Trend =

"CO2 emissions have " &

IF([CO2 Emissions Change Percentage] < 0, "decreased", "increased") &

" by " & FORMAT(ABS([CO2 Emissions Change Percentage]), "0.00") & "% over the last 8 years."
```

```
CO2 Emissions Change Percentage =

VAR EmDiff = [CO2 Emissions Last Year]-[CO2 Emissions First Year]

RETURN

DIVIDE(EmDiff, [CO2 Emissions First Year], 0) * 100
```

UI layer consists of below sheets:

- Home Page
- Electricity Sales Dashboard
- Electricity Price Trend
- Electricity Price Trends Over Time
- CO2 Emission Dashboard
- Prices Vs Emissions
- Price Prediction.

1. Home Page: This Page consists of Page Navigator at top for all the pages along with link for Electricity Sales Dashboard and CO2 Emissions Dashboard.

2. Electricity Sales Dashboard: This page reveals Electric Prices Trend along with KPI cards Average Electricity Price, total customers, total electricity sales, average price for first year (2014) and latest year (2023) and map focusing on Total Revenue and Total Electric Sales.

3. Electricity Price Trends: This page focuses on Total Price by year and State Name and sector.

4. Electricity Price Trends over Time: This page reveals Total price change over years in the form of waterfall report. Funnel chart displays total revenue by sector. Ribbon chart tells about cumulative Sales by Sector and special self- analysis visualization which is inbuilt feature of Power BI that tells Price to either increase / decrease based on affecting parameters and tells at what data point's combination, price is going to increase / decrease.

5. CO2 Emission Dashboard: This page shows trend in CO2 emissions over last 8 years from 2014 to 2022, along with KPI cards showing CO2 emissions measures and visualization charts showing average of CO2 emissions against sector, fuel type and year.

6. Prices Vs Emissions: This page mainly features CO2 emissions increase pattern over the years from 2014-2022 accounting for change per each year. CO2 emission be Electricity price through scatter plot.

7. Price Prediction: This page compares Actual Price Vs Predicted Price data using ML model along with KPI card computing Forecast Accuracy based on year.

Below is the PBIx file for our project.

V. Conclusion

The analysis of energy consumption, sales, and emissions datasets reveals significant variability across states and sectors, highlighting seasonal demand and the impact of pricing strategies on consumer behavior. Emissions data emphasize the need for cleaner energy transitions, with coal and petroleum as major contributors. Data preprocessing addressed challenges like missing values and outliers, ensuring reliability. These insights enable policymakers to target high-emission regions and prioritize renewable initiatives, while utility companies can optimize pricing and efficiency. This comprehensive analysis supports data-driven decisions in energy, economics, and sustainability.