Final Project Report

DS 2500 Intermediate Programming with Data

Team Members: Harshini Dinesh (dinesh.h@northeastern.edu) , Ryan Jiang (jiang.ry@northeastern.edu), Hersh Joshi (joshi.her@northeastern.edu), Hang Hang (hang.h@northeastern.edu)

Date: 8/16/2024

Problem Statement and Background

In recent years, the U.S. has seen a surge in the deployment of renewable energy sources, driven both by environmental concerns and economic incentives. While various news outlets (i.e. Bloomberg, Wall Street Journal) and databases (e.g. Census Bureau, Energy Information Adminstration) provide data on the current state of renewable energy and its economic implications, a deeper analysis is needed to understand the economic impact of different renewable energy sources across U.S. states and sectors. Our project aim is to bridge the gap by examining the relationship between renewable energy production, consumption, and economic indicators such as GDP. These analyses can help policymakers, environmental researchers, data scientists, and large organizations like the Environmental Protection Agency (EPA) and the United Nations (U.N.) develop targeted strategies for optimizing renewable energy deployment, reducing reliance on traditional energy sources, and mitigating climate change. For example, by identifying trends in the adoption of Alternative Fuel Stations (AFS) and their focus on specific energy sources, agencies like EPA and U.N. can allocate resources more effectively such as focusing more on gathering and distributing funds to create more AFSs to accelerate the transition to renewable energy, thus supporting both the environment and the economy.

Introducing and Explaining the Data

Our datasets are mainly sourced from government institutions which focus on state-level datasets with no personal identifiable information. However, our energy consumption and generation data are based off on 2020 Residential Energy Consumption Survey (RECS), which is a national survey (subjecting to sampling errors or high RSE) that runs every four years (missing important trends) only on single families, apartments, and mobile homes, which may exclude other family types and is not a good fit for all demographics and geographies. Datasets used in our report: state consumption, production, and GDP data over time, net generation for all sectors, coastal vs. inland states (manually compiled based on the state listing with coastal residents)), U.S. state shapefile, levelized cost of charging EV, data about

total renewable energy consumed by sector (electric power, industrial, residential, commercial), alternative fuel stations (AFS) data, and comparing diesel vs electric vehicles.

Data Science Approaches

Loading and Merging Data: Data from government institutions, both Excel and CSV files, were first loaded into Pandas DataFrame, including net generation for all sectors, coastal vs. inland state classification, and levelized cost of charging (LCOC) electric vehicles. A U.S. state-level shapefile was loaded using the Geopandas library. We merged these DataFrames on common columns of state names to consolidate all relevant data for analysis.

Handling missing values: We addressed various forms of missing values. In one of our datasets "net generation for all sectors", for the specific year column, we replaced entries marked as 'NM' (not meaningful due to large relative standard error) and '--' with '0', avoiding skewed results due to missing values.

Basic statistics: We calculated the basic statistics for the workplace vs. residential cost of charging EVs to understand the overall change in cost pattern and states with extreme values.

Correlation analysis: We calculate the correlation between various variables, including energy consumption/production, state locations (inland vs. coastal), GDP, and EV usage to find out if these variables are related and to determine the strength and direction of their relationships, if any.

Regression analysis: We performed linear regression analysis on Vermont's wind energy generation, which has consistently increased from 2001 to 2023. The analysis aimed to predict future wind energy generation based on historical data, with the model's RMSE at 72.36. The model's accuracy could be improved by incorporating more data points over a longer period.

Visualizations: To display relationships between quantitative variables, we used scatter plots and regression plots to visualize these relationships. We also used boxplots to display the median, range, and any outliers in our data.

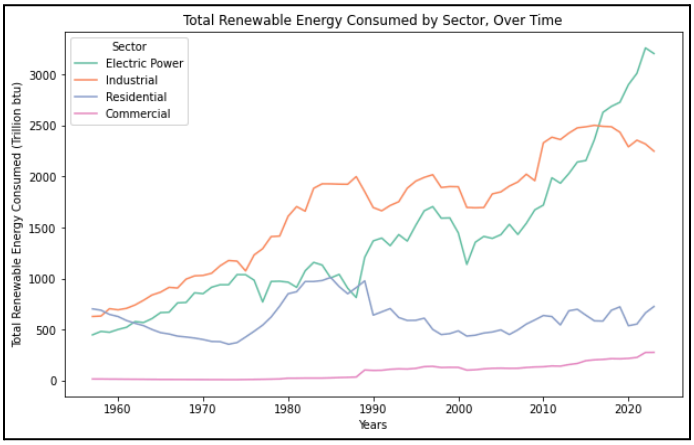
- **Geospatial heatmaps**: We used Geopandas to visualize the state-by-state energy generation where darker shades indicate higher percentages.
- **Box plot:** Box plot is used to display the spread of renewable energy generation data, highlighting the median, quartiles, and potential outliers.
- **Bar chart**: A bar plot compares baseline residential and workplace LCOC for Battery EVs (BEVs)
- **Residual plot**: We used residual plot to assess the linear regression model's fit and accuracy.
- **Confusion matrix**: To determine the dominant renewable energy source across different states, we employed a KNN classifier, evaluating the model with a confusion matrix. The input features (x) represent the amounts of various renewable energy sources, such as conventional hydroelectric, all solar, wind, and biomass, generated by each state. The output labels (y) were categorical values, like "High wind" or "High biomass," based on the dominant energy source in each state. Cross-validation was used to identify the optimal k-values for both recall (k=4) and precision (k=9).

Results and Conclusions

Our analysis of renewable energy consumption and production over time revealed significant differences among states. The top 5 states with the greatest increases in consumption since 1970 were Texas, California, Iowa, Florida, and Georgia, while the top 5 states with the greatest increases in production were Iowa, Texas, Nebraska, California, and Illinois. These trends can be attributed to various factors, including government policies, natural resource availability, and population size. For instance, Texas and California appear in the top 5 for both consumption and production, likely due to their large populations and proactive state policies promoting greater consumption and production of renewable

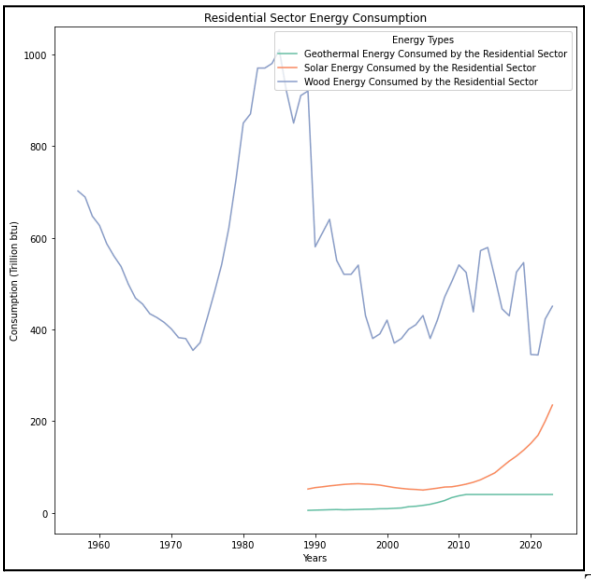
energy. Additionally, our analysis found strong positive correlations between a state's renewable energy consumption and production with its real GDP, with correlation coefficients of 0.8333 and 0.5530, respectively. To gain further insights, we identified the top 5 states with the highest renewable energy consumption and production relative to their GDP. Surprisingly, the same states—Washington D.C., Delaware, New Jersey, Maryland, and Massachusetts—topped the list for both metrics. This may be because states with high renewable energy production often have correspondingly high consumption, as they tend to use much of the energy they produce.

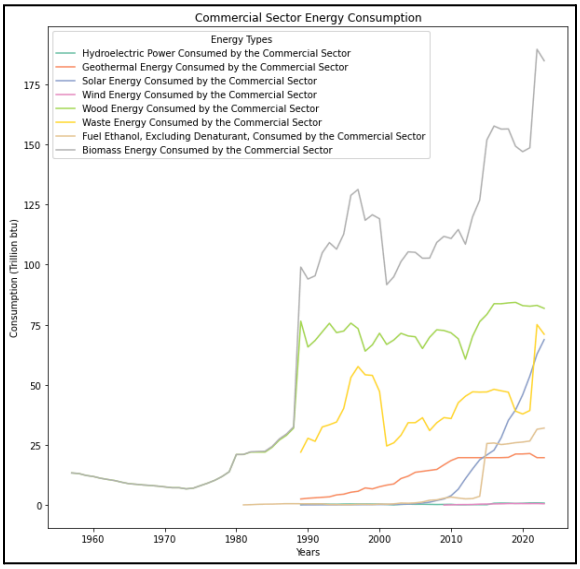
To further delve into how clean energy impacts our economy, we looked at four different sectors' (electric power, industrial, residential, commercial) energy consumption level data from 1949 to 2023.. Initially, we hypothesized that given the growth of the EV industry, the electric power sector had consumed more renewable energy than the industrial, commercial, and residential sectors. When we plotted a line plot showing how much total energy was consumed by each sector, the hypothesis seemed to hold true because we saw consistent large increases in the total energy consumed by the electric power sector. However, when we performed the calculations, we found that the sector with the highest total renewable energy consumption was the Industrial sector at 111415.74 Trillion btu, not the Electric Power sector at 91597.39 Trillion btu. Since both values were really close, it makes sense why we weren't able to clearly see the difference in our line chart. We think that the industrial sector has consumed the most energy between 1949 and 2023 because it takes up vast amounts of energy to support its production and manufacturing industries.



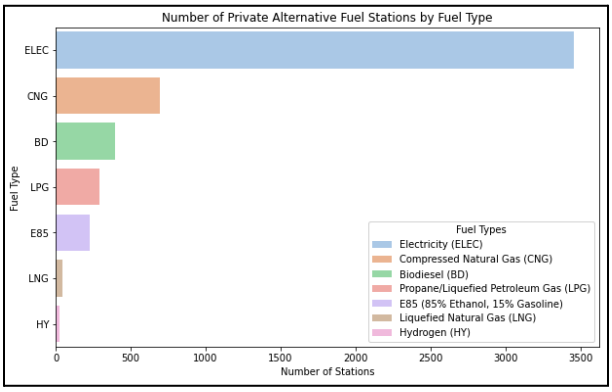
To gain more insight about each sector's most consumed renewable energy source, we conducted more line plots and calculations. We found that the Industrial Sector consumes biomass energy the most at 110,480.94 Trillion btu, the Electric Power Sector consumes conventional hydroelectric power at

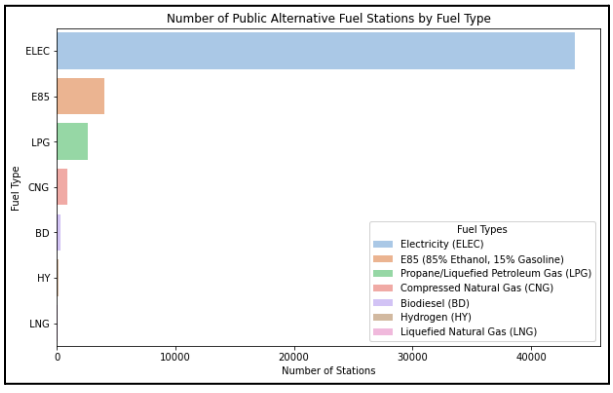
58,757.67 Trillion btu, the Residential Sector consumes wood energy at 36,997.09 Trillion btu, and the Commercial Sector consumes biomass energy at 4,732.62 Trillion btu. It was not surprising to see that the residential sector consumes mostly wood energy because people use wood consistently to power and heat their homes. Plus, biomass energy appears to be becoming the next most popular renewable energy source in industrial and commercial sectors, moving away from fossil fuels and other harmful energy sources.



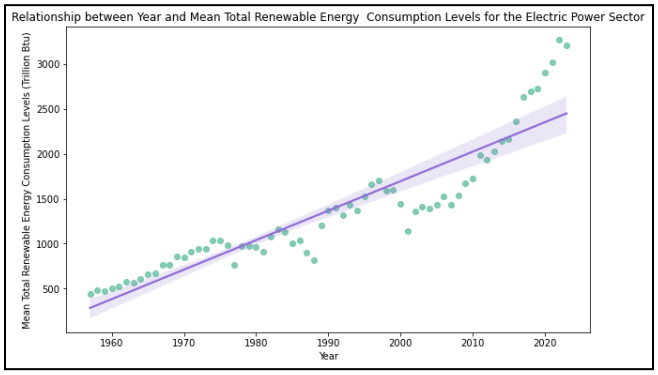


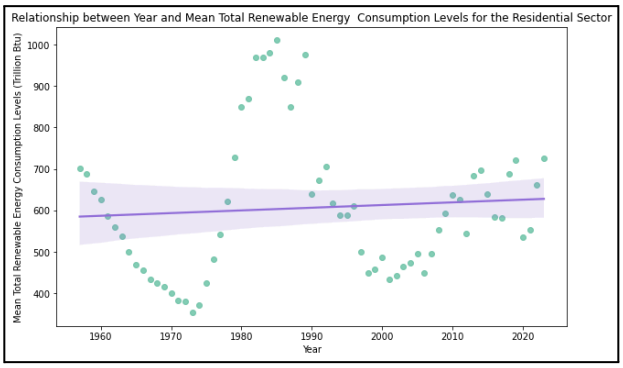
Then we analyzed the alternative fuel stations data. We created bar charts to see how much of each fuel type public versus private alternative fuel stations provide to see if EV adoption has increased stations providing electricity more than other fuels. We found that there are more public alternative fuel stations (43,632) which provide electricity than private alternative fuel stations (3,454).



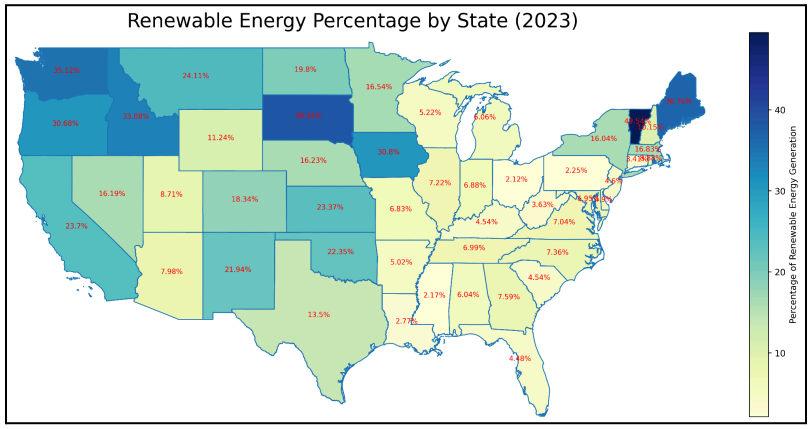


Finally, we conducted correlation and regression analyses to show the relationship between the year and mean total renewable energy consumption levels for the different sectors. While the correlations between year and electric power, industrial, and residential sectors were strong at 0.914, 0.943, and 0.93 respectively, the residential sector had a really weak correlation to year at 0.074. Its regression plots showed likewise with the residential sector having a sinusoidal curve, suggesting no relationship, while all other sectors had an increasing linear relationship. This suggests that all other sectors are focusing on increasing their total renewable energy consumption over the years and leaving climate-damaging energy sources behind (oil, fossil fuels etc.) while residential sectors have years where they increase and decrease their renewable energy consumption levels.



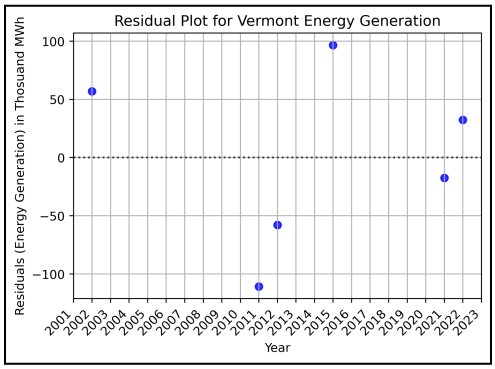


Additionally, the Pearson correlation coefficient between renewable energy percentage and state location was found to be **r = −0.2434**. This indicates a weak negative correlation, suggesting that inland states generally produce slightly less renewable energy than coastal states. However, this correlation is not strong enough to draw definitive conclusions about the impact of geographic location on energy production.

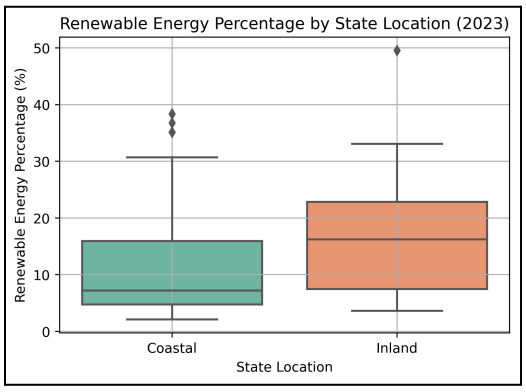


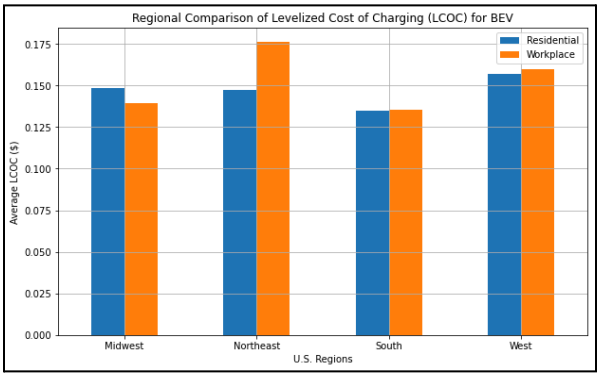
With our geospatial map, we noticed that Vermont (Inland) leads in renewable energy generation percentage, contributing 49% of its total energy from renewable sources. Following Vermont are South Dakota (Coastal): 38%, Maine (Coastal): 37%, Washington (Coastal): 35%, Idaho (Inland): 33%. This finding challenged our initial hypothesis that coastal states produce more renewable energy, suggesting that factors like natural resources, climate, policies, and infrastructures may have greater and more complex impact on energy production than location alone.

In the linear regression analysis of Vermont's wind energy production from 2001 to 2023, the model showed a relatively high RMSE of 72.36, indicating that the model struggles to precisely predict wind energy generation, especially considering that the majority of data points range between 10 and 400 thousand megawatt-hours. The residual plot further underscores this variability, highlighting discrepancies between the actual and predicted values. To improve the model's accuracy, incorporating a broader dataset or introducing additional variables - government incentives, technological advancements, or policy changes could offer a more nuanced understanding and lead to more reliable predictions.



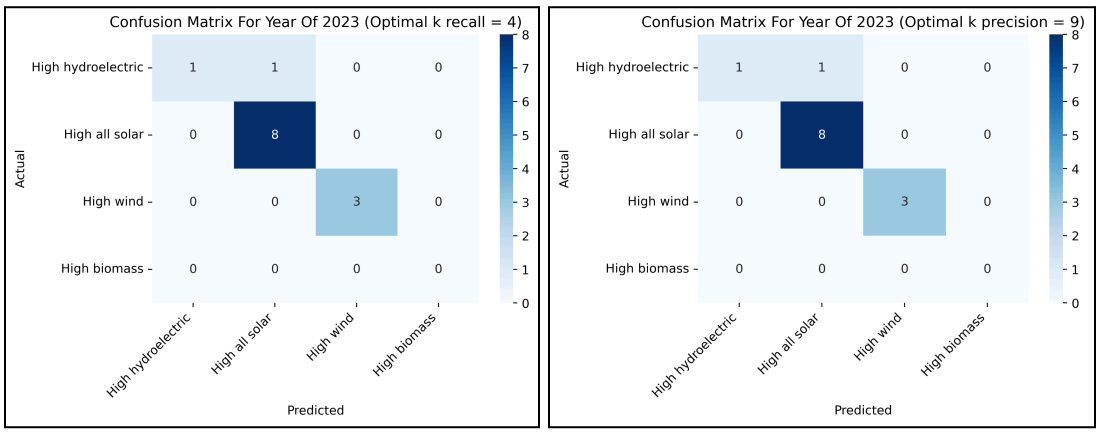
Our boxplots highlighted that inland states tend to have higher median and mean energy generation, along with greater variability (wider IQR), suggesting more diverse energy output likely due to different climates and other influencing factors. Further, the bar plot revealed significant variability between actual and predicted values and RMSE is relatively high which is 72.37 considering the context that most of Vermont's wind energy ranges between 10 and 400 thousand megawatt hours. Having more data points





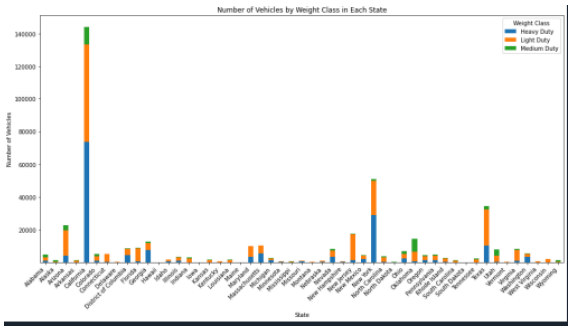
over a broader year range or including more variables such as government incentives could be beneficial.

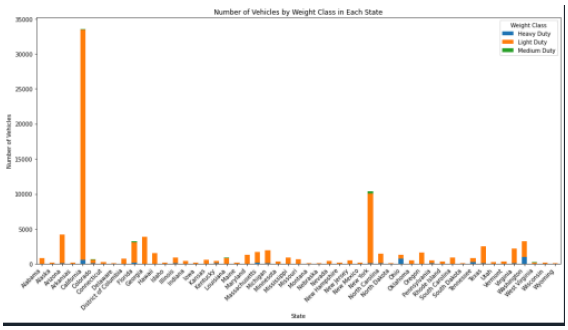
Regarding the K-Nearest Neighbors (KNN) performance, despite different optimal k-values, the confusion matrix results were identical, likely due to the model's high accuracy score (0.9231) and the clear and distinct nature of the dataset's labels. This clarity in the data led to minimal misclassification, with only one state, categorized as "High hydroelectric," being incorrectly predicted as "High solar." The result indicates that the KNN classifier effectively distinguished between states' dominant renewable energy sources, making it relatively reliable for our prediction.



Finally we decided to hone into the transportation sector and measured sustainability in vehicles across the 50 states over the last 15 years in both, electric and diesel dependent vehicles for various weight classes, heavy, light, and medium. To compute this numerically, we had initially sought out to compose of a correlation function between the years passed and the fuel consumed for different weight classes but duen to unequal dataframes of length this was not possible, and if I reduced the length of one of these or extended the other it either overlook or underlook important data points. Henceforth, I decided to go for a scatterplot and bar plot illustrating how these data points would look in comparison to another, when comparing the amount of fuel consumed per different weights of vehicles (light, medium, or heavy) with electric and diesel vehicles. These graphs were meant to go back to our hypothesis 1: do climate, population, and geographical location have an impact on the amount and type of fuel consumed in that area.

These graphs did indeed show our datasets for this matter would not be an accurate point to help us prove or negate our hypothesis for numerous reasons. Firstly, correlation does not equal causation. Secondly, the states with two very drastically different climates, California and New York, are shown to have the highest amounts of fuel/energy consumed by their vehicles, so much so that they outweigh all other 48 states on this graph, and another graph excluding these two states would be a better visualization to show the more accurate results for the rest of the states. Nonetheless, this did help suggest that there could be truth to one of our initial beliefs: that legislation does indeed play a role in dictating how much is consumed by some vehicles, as both of these states are highly left leaning politically.





Future Work

Our findings could be used to further research how renewable energy sources can bolster the U.S. economy through state-level incentives, infrastructure expansion, cost-benefit analysis, increased electric vehicle (EV) adoption, efficient energy consumption, and increased AFS accessibility. To further explore this, we would need to conduct more data collection and data analysis. Possible analyses include: analyzing the effectiveness of specific state incentives, like MassEVIP, in reducing LCOC and their impact on EV adoption rates; analyzing how expanding advanced charging infrastructure (e.g., Level 2, DC fast charging) influences both residential and workplace charging costs; analyzing the costs and benefits of EV charging infrastructure investments, considering long-term savings for the charging station users; analyzing the relationships between EV incentives and its adoption rate; analyzing and forecasting future energy consumption levels by sector and future AFS accessibility; analyzing how consumption of each energy source and how increases in AFS impact climate change. We could perform all of the

aforementioned analyses using correlations, linear regression models, Monte Carlo simulations, and hypothesis testing.

In the future we could also broaden our horizons regarding our mediums for finding data to include more climate specific data. Upon attempting to incorporate feedback from our presentation, we came across a website called NOAA (National Oceanic Atmospheric Administration), that provides more comprehensive datasets for region specific and climate specific data to better support our claims, especially regarding renewable energy generation in various sectors of our country. Should we be working on a similar topic in the future, this would be another site where we could find a great deal of resources from.